114 學年度第 1 學期

統 計 學 (一)

Anderson's Statistics for Business & Economics (14/E)

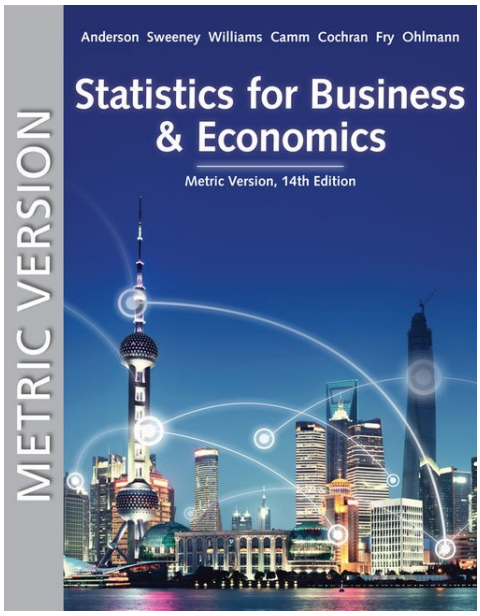
授課教師: 吳漢銘 國立政治大學統計學系

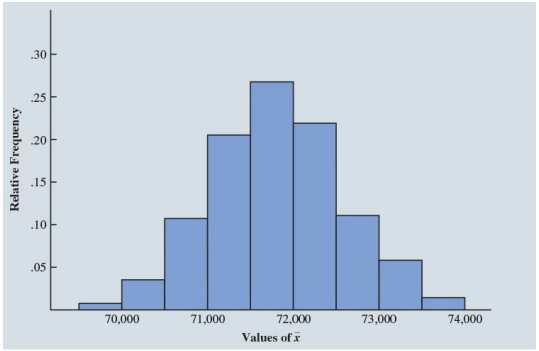
開課單位: 統計系

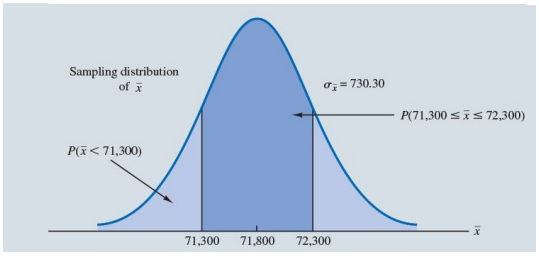
科目代碼: 000359041

教學網站: https://hmwu.idv.tw

系級: _____ 學號: _____ 姓名: _____







目錄

Ch 1. Data and Statistics

Ch 2. Descriptive Statistics: Tabular and Graphical Displays

Ch 3. Descriptive Statistics: Numerical Measures

Ch 4. Introduction to Probability

Ch 5. Discrete Probability Distributions

Ch 6. Continuous Probability Distributions

Ch 7. Sampling and Sampling Distributions

Ch 8. Interval Estimation

Ch 9. Hypothesis Tests

Ch 10. Inference about Means and Proportions with Two Populations (Optional)

附錄：113-1 學年第 1 學期小考題、期中考題、期末考題。

叮嚀

A. 平常就要唸書，做習題。

B. 考過的題目，要主動訂正。

C. 上課以「互相尊重」為最高原則並盡到「告知老師」的義務。

D. 上課可小聲討論、上廁所安靜去回、不鼓勵飲食。(請一定要維護教室整潔)

E. 四不一要: 「上課不聊天，睡覺不趴著，手機不要滑，考試不作弊，要認真。」

統計學 (一)

Anderson's Statistics for Business & Economics (14/E)

Chapter 1: Data and Statistics

上課時間地點: 二 D56, 商館 260206
授課教師: 吳漢銘 (國立政治大學統計學系副教授)
教學網站: https://hmwu.idv.tw

Overview: What is Statistics?

1. The term statistics can refer to _____ such as averages, medians, percentages, and maximums that help us _____ a variety of business and economic situations.
2. Wikipedia: Statistics is the discipline (art and science) that concerns the _____, _____, _____, _____, and _____ of data.

1.1 Applications in Business and Economics

Accounting

Public accounting firms use statistical _____ procedures when conducting audits for their clients.

Economics

Economists use statistical information in making _____ about the future of the economy or some aspect of it.

Finance

Financial advisors use price-earnings _____ (本益比) and dividend yields (現金殖利率) to guide their investment advice.

Marketing

Electronic point-of-sale scanners at retail checkout counters are used to _____ data for a variety of marketing research applications.

Production

A variety of statistical _____ are used to monitor the output of a production process.

Information Systems

A variety of statistical information helps administrators assess the _____ of computer networks.

1.2 Data

1. Data are the _____ collected, analyzed, and summarized for presentation and interpretation.
2. All the data collected in a particular study are referred to as the _____ for the study.

Elements, Variables, and Observations

1. Elements are the _____ on which data are collected.
2. A variable is a _____ of interest for the elements.
3. The set of measurements obtained for a particular element is called an _____.
4. A data set with _____ elements contains _____ observations.

5. The total number of _____ in a complete data set is the number of elements multiplied by the number of variables.

6. Example WTO, 惠譽國際 (Fitch Group) (Per Capita 人均)

TABLE 1.1 Data Set for 60 Nations in the World Trade Organization

Nation	WTO Status	Per Capita GDP ($)	Fitch Rating	Fitch Outlook
Armenia	Member	3,615	BB−	Stable
Australia	Member	49,755	AAA	Stable
Austria	Member	44,758	AAA	Stable
Azerbaijan	Observer	3,879	BBB−	Stable
Bahrain	Member	22,579	BBB	Stable
Belgium	Member	41,271	AA	Stable
Brazil	Member	8,650	BBB	Stable
Bulgaria	Member	7,469	BBB−	Stable
Canada	Member	42,349	AAA	Stable
Cape Verde	Member	2,998	B+	Stable
Chile	Member	13,793	A+	Stable
China	Member	8,123	A+	Stable
Colombia	Member	5,806	BBB−	Stable
Costa Rica	Member	11,825	BB+	Stable
Croatia	Member	12,149	BBB−	Negative

Scales of Measurement

1. Scales of measurement include: _____.

2. The scale determines the _____ contained in the data.

3. The scale indicates the _____ (污名化) and statistical analyses that are most appropriate.

4. **Nominal scale**

   (a) Data are _____ used to identify an attribute of the element.

   (b) A _____ or _____ may be used.

   (c) Example: Students of a university are classified by the school in which they are enrolled using a nonnumeric label such as Business, Humanities, Education, and so on. Alternatively, a numeric code could be used for the school variable (e.g., 1 denotes Business, 2 denotes Humanities, 3 denotes Education, and so on).

5. **Ordinal scale**

   (a) The data have the properties of nominal data and the _____ of the data is meaningful.

   (b) A _____ or _____ may be used.

   (c) Example: Students of a university are classified by their class standing using a non-numeric label such as Freshman, Sophomore, Junior, or Senior. Alternatively, a numeric code could be used for the class standing variable (e.g., 1 denotes Freshman, 2 denotes Sophomore, and so on).

6. **Interval scale**

   (a) The data have the properties of ordinal data, and the_____ between observations is expressed in terms of a _____.

   (b) Interval data are always _____.

   (c) Example: Melissa has an SAT score of 1985, while Kevin has an SAT score of 1880. Melissa scored 105 points more than Kevin.

7. **Ratio scale**

   (a) Data have all the properties of interval data and the _____ of two values is meaningful.

   (b) Ratio data are always _____.

   (c) _____ is included in the scale.

   (d) Example: Price of a book at a retail store is $200, while the price of the same book sold online is $100. The ratio property shows that retail stores charge twice the online price.

Categorical and Quantitative Data

1. Data can be further classified as being _____ or _____.

2. The statistical analysis that is appropriate depends on whether the data for the variable are categorical or quantitative.

3. **Categorical Data**

(a) Labels or names are used to identify an _____ of each element.

(b) Often referred to as _____.

(c) Use either the _____ or _____ scale of measurement.

(d) Can be either _____ or nonnumeric.

(e) Appropriate statistical analyses are rather _____.

4. **Quantitative Data**

(a) Quantitative data indicate _____ or _____.

(b) Quantitative data are always _____.

(c) _____ operations are meaningful for quantitative data.

Cross-Sectional and Time Series Data

1. **Cross-sectional data**

(a) Cross-sectional data are collected _____ or _____ point in time.

(b) Example: The data in Table 1.1 are cross-sectional because they describe the five variables for the 60 World Trade Organization nations at the same point in time.

2. **Time Series Data**

(a) Time series data are collected _____.

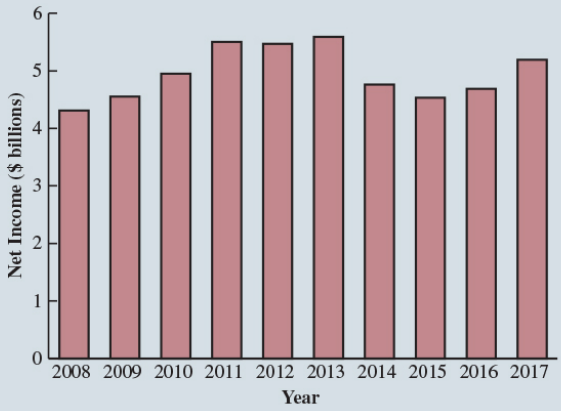
(b) Example: Data detailing the number of building permits issued in Lucas County, Ohio in each of the last 36 months.

(c) Graphs of time series data help analysts understand: (i) what happened in the _____, (ii) identify any _____ over time, and (iii) _____ future levels for the time series.

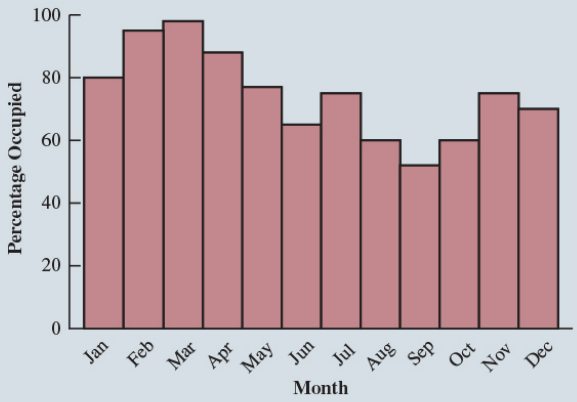
FIGURE 1.2 A Variety of Graphs of Time Series Data



(A) Dow Jones Industrial Average Index



(B) Net Income for McDonald's Inc.



(C) Occupancy Rate of South Florida Hotels

1.3 Data Sources

Existing Sources

1. Internal company records – almost any department
2. Business database services – Dow Jones & Co.
3. Government agencies – U.S. Department of Labor
4. Industry associations – Travel Industry Association of America
5. Special-interest organizations – Graduate Management Admission Council (GMAT)
6. Internet –e.g., 政府資料開放平臺: `https://data.gov.tw`

Observational Study

1. In observational (nonexperimental) studies _____ is made to _____ the variables of interest.
2. Example: Survey and public opinion, e.g., studies of smokers and nonsmokers are observational studies because researchers do not determine or control who will smoke and who will not smoke.

Experiment

1. In experimental studies the variable of interest is first _____. Then one or more other variables are _____ so that data can be obtained about how they influence the variable of interest.
2. Example: The largest experimental study ever conducted is believed to be the 1954 Public Health Service experiment for the Salk polio vaccine (小兒痲痺疫苗). Nearly two million U.S. children (grades 1- 3) were selected.

Time and Cost Issues

1. Searching for information can be _____.
2. Information may _____ by the time it is available.

Data Acquisition Errors

1. Organizations often charge for information even when it is not their primary business activity.

2. Using any data that happen to be available or were acquired with little care can lead to _____ information.

1.4 Descriptive Statistics

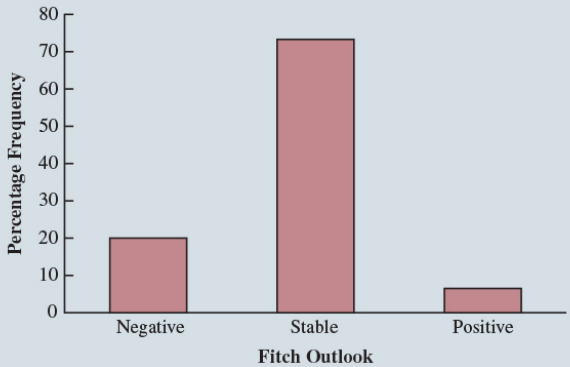
1. Most of the statistical information in the media, company reports, and other publications consists of data that are summarized and presented in a form that is easy for the reader to understand. Such _____, which may be _____, _____, or _____, are referred to as _____.

2. Example A tabular summary of the data showing the number of nations with each of the Fitch Outlook ratings (Table 1.4). A graphical summary of the same data, called a _____ (Figure 1.4). We can see that the majority of Fitch Outlook credit ratings are stable, with 73.3% of the nations having this rating. More nations have a negative outlook (20%) than a positive outlook (6.7%).

TABLE 1.4 Frequencies and Percent Frequencies for the Fitch Credit Rating Outlook of 60 Nations

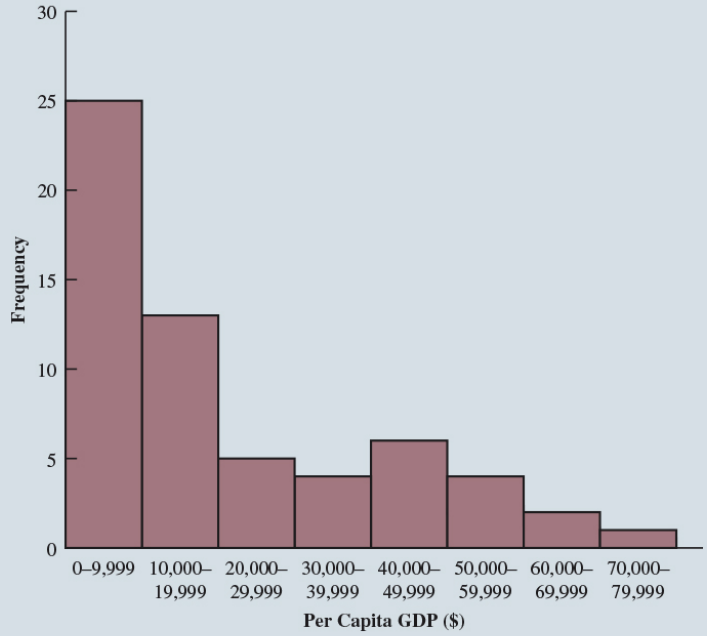
Fitch Outlook	Frequency	Percent Frequency (%)
Positive	4	6.7
Stable	44	73.2
Negative	12	20.0

FIGURE 1.4 Bar Chart for the Fitch Credit Rating Outlook for 60 Nations



3. Example A graphical summary of the data for the quantitative variable Per Capita GDP in Table 1.1, called a _____ (Figure 1.5). Using the histogram, it is easy to see that Per Capita GDP for the 60 nations ranges from \$0 to \$80,000, with the highest concentration between \$0 and \$10,000. Only one nation had a Per Capita GDP exceeding \$70,000.

FIGURE 1.5 Histogram of Per Capita GDP for 60 Nations



4. The most common numerical descriptive statistic is the _____.

5. The mean demonstrates a measure of the _____, or _____ of the data for a variable.

1.5 Statistical Inference

1. **Population**: The set of _____ in a particular study.

2. **Sample**: A _____ of the population.

3. **Statistical inference**: The process of using data obtained from a sample to _____ about the characteristics of a population.

4. **Census**: Collecting data for the _____ population.

5. **Sample survey**: Collecting data for a _____.

6. Example Rogers Industries manufactures lithium batteries used in rechargeable electronics such as laptop computers and tablets. In an attempt to increase battery life for its products, Rogers has developed a new solid-state lithium battery that should last longer and be safer to use.

   (a) population: all lithium batteries that could be produced using the new solid-state technology.

   (b) sample (data): 200 batteries manufactured with the new solid-state technology were tested. (Table 1.5) the number of hours each battery lasted before needing to be recharged under controlled conditions.

TABLE 1.5 Hours Until Recharge for a Sample of 200 Batteries for the Rogers Industries Example

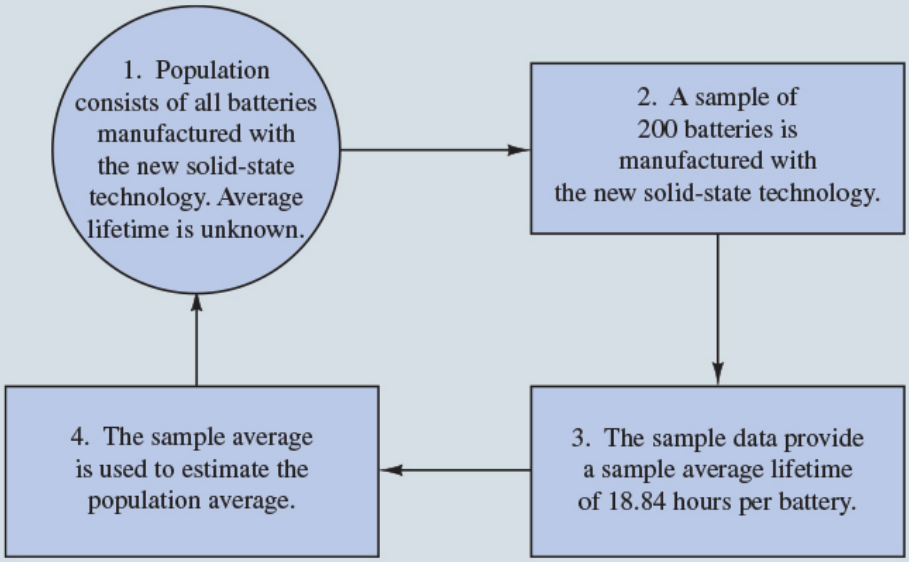
Battery Life (hours)									
19.49	18.18	18.65	19.45	19.89	18.94	17.72	18.35	18.66	18.23
19.08	19.92	19.01	18.84	17.73	19.70	18.37	18.69	19.98	18.80
19.11	18.26	19.05	17.89	19.61	18.52	18.10	19.08	18.27	18.29
19.55	18.81	18.68	17.43	20.34	17.73	17.66	18.52	19.90	19.33
18.81	19.12	18.39	19.27	19.43	19.29	19.11	18.96	19.65	18.20
19.18	20.07	18.54	18.37	18.13	18.29	19.11	20.22	18.07	18.91

   (c) Suppose Rogers wants to use the sample data to make an inference about the average hours of battery life for the population of all batteries that could be produced with the new solid-state technology.

   (d) The sample average battery life: _____. We can use this sample result to estimate that the average lifetime for the batteries in the population is _____ hours.

   (e) (Figure 1.6) a graphical summary of the statistical inference process for Rogers Industries.

FIGURE 1.6 The Process of Statistical Inference for the Norris Electronics Example



1.6 Analytics

1. Analytics is the scientific process of transforming data into _____ for making better _____.

2. Three categories of techniques:

   (a) **Descriptive analytics**: This describes what has happened in the _____.

   (b) **Predictive analytics**: Use models constructed from past data to _____ or to assess the _____ of one variable on another.

   (c) **Prescriptive analytics**: The set of analytical techniques that yield a _____

1.7 Big Data and Data Mining

1. **Big data**: Large and complex data set. Three V′ s of Big data:

   (a) _____: Amount of available data.

   (b) _____: Speed at which data is collected and processed.

(c) _____: Different data types.

2. **Data Mining**

   (a) Methods for developing useful _____ information from _____.

   (b) Using a combination of procedures from statistics, mathematics, and computer science, analysts _____ to convert it into useful information.

   (c) The most effective data mining systems use automated procedures to discover relationships in the data and predict future outcomes prompted by general and even vague queries by the user.

   (d) _____ such as multiple regression, logistic regression, and correlation are heavily used.

   (e) Also needed are computer science technologies involving _____ and _____.

   (f) With the enormous amount of data available, the data set can be partitioned into a _____ (for _____ development) and a _____ (for _____ the model).

   (g) Careful interpretation of results and extensive testing is important.

1.8 Computers and Statistical Analysis

1. Computer: laptop, PC, Server, ...
2. OS: Windows, Mac OS, Linux, ...
3. Software for statistical analysis: Microsoft Excel, JMP, SPSS, SAS, R, Python

1.9 Ethical Guidelines for Statistical Practice

1. In a statistical study, unethical behavior can take a variety of forms including: improper _____, inappropriate _____ of the data, development of _____, use of inappropriate _____, _____ of the statistical results.

2. One should strive to be _____, _____, _____, and _____(中性的) as you collect, analyze, and present data.

3. As a consumer of statistics, one should also be aware of the possibility of _____ by others.

4. The American Statistical Association (ASA) developed the report "Ethical Guidelines for Statistical Practice". It contains 67 guidelines organized into 8 topic areas:

   (a) Professionalism.

   (b) Responsibilities to Funders, Clients, and Employers.

   (c) Responsibilities in Publications and Testimony.

   (d) Responsibilities to Research Subjects.

   (e) Responsibilities to Research Team Colleagues.

   (f) Responsibilities to Other Statisticians/Practitioners.

   (g) Responsibilities Regarding Allegations of Misconduct.

   (h) Responsibilities of Employers Including Organizations, Individuals, Attorneys, or Other Clients.

😍 SUPPLEMENTARY EXERCISES

3, 4, 5, 9, 13, 14, 19, 25

名言佳句

“不要努力想成為一個成功者，要努力成為一個有價值的人”
“Try not to become a man of success, but rather try to become a man of value”
— *Albert Einstein (March 14, 1879 – April 18, 1955)*

統計小故事：南丁格爾與戰爭醫院的數據革命

十九世紀中葉，克里米亞戰爭爆發。對於英國社會而言，戰爭的死亡被視為「無可避免」的命運。但有一位年輕的女性卻不願接受這種說法，她就是佛羅倫斯．南丁格爾 (Florence Nightingale)。

南丁格爾出身於富裕家庭，本可安逸地過一生，但她選擇成為護士。1854 年，她被派往戰場醫院，迎接她的卻是震撼的景象：骯髒的病房、腐敗的食物、缺乏基本的衛生措施。士兵們大批死去，但並非死於槍傷，而是死於霍亂、痢疾和壞血病。

她並沒有只憑直覺抱怨，而是開始系統地「收集資料」。她每天記錄士兵入院人數、死亡人數與死因，逐月整理成龐大的數據表。當她回到倫敦後，將這些資料彙整分析，結果顯示：超過七成的死亡其實來自傳染病，而非戰傷。

然而，單純的數字往往乏善可陳，難以說服軍方。於是南丁格爾採用了當時極具創新的方法——「圖像化統計」。她設計了著名的「玫瑰圖」(coxcomb diagram)，利用色塊面積來呈現不同死亡原因的比例。這些圖表清楚地顯示，因衛生問題而死亡的士兵數遠超過戰場本身的損失。

這份報告震撼了英國國會與軍方。最終，醫院開始改革：建立排水系統、清潔飲水、改善食物品質。幾年後，士兵的死亡率降至戰前的一半。

南丁格爾的故事提醒我們：統計不只是數字，而是能轉化為制度與政策的武器。透過資料的收集、整理與分析，人類得以將「感覺」變成「證據」，將「證據」化為「行動」。

對學生而言，這正是第一章的核心：統計學的起點就是「資料」。我們先收集，再描述，最後利用它來支持決策。若沒有南丁格爾的數據，死亡可能依舊被視為命運；但有了統計，命運就可以被改寫。

統計學 (一)

Anderson's Statistics for Business & Economics (14/E)

Chapter 2: Descriptive Statistics: Tabular and Graphical Displays

上課時間地點: 二 D56, 商館 260206
授課教師: 吳漢銘 (國立政治大學統計學系副教授)
教學網站: https://hmwu.idv.tw

2.1 Summarizing Data for a Categorical Variable

Frequency Distribution

1. **Frequency distribution**: A frequency distribution is a _____ of data showing the number (frequency) of observations in each of several _____ categories or classes.

2. Example Soft Drinks Data (Table 2.1)

   (a) Coca-Cola, Diet Coke, Dr. Pepper, Pepsi, and Sprite are five popular soft drinks. The data show below the soft drink selected in a sample of 50 soft drink purchases.

TABLE 2.1 Data from a Sample of 50 Soft Drink Purchases

Coca-Cola	Coca-Cola	Coca-Cola	Sprite	Coca-Cola
Diet Coke	Dr. Pepper	Diet Coke	Dr. Pepper	Diet Coke
Pepsi	Sprite	Coca-Cola	Pepsi	Pepsi
Diet Coke	Coca-Cola	Sprite	Diet Coke	Pepsi
Coca-Cola	Diet Coke	Pepsi	Pepsi	Pepsi
Coca-Cola	Coca-Cola	Coca-Cola	Coca-Cola	Pepsi
Dr. Pepper	Coca-Cola	Coca-Cola	Coca-Cola	Coca-Cola
Diet Coke	Sprite	Coca-Cola	Coca-Cola	Dr. Pepper
Pepsi	Coca-Cola	Pepsi	Pepsi	Pepsi
Pepsi	Diet Coke	Coca-Cola	Dr. Pepper	Sprite

(b) The number of times each soft drink (_____) summarizes information about the popularity of the five soft drinks.

TABLE 2.2 Frequency Distribution of Soft Drink Purchases

Soft Drink	Frequency
Coca-Cola	19
Diet Coke	8
Dr. Pepper	5
Pepsi	13
Sprite	5
Total	50

Relative Frequency and Percent Frequency Distribution

1. The relative frequency of a class equals the _____ of observations belonging to a class.

2. For a data set with n observations, the relative frequency of each class is:

   Relative frequency of a class = _____

3. A relative frequency distribution gives a tabular summary of data showing the relative frequency for each class.

4. The percent frequency of a class is the relative frequency _____. A percent frequency distribution summarizes the percent frequency of the data for each class.

5. Example Soft Drinks Data (Table 2.3)

   (a) The relative frequency for Coca-Cola is $19/50 = 0.38$, the relative frequency for Diet Coke is $8/50 = 0.16$, and so on.

   (b) From the percent frequency distribution, we see that 38% of the purchases were Coca-Cola, 16% of the purchases were Diet Coke, and so on.

   (c) We can also note that $38\% + 26\% + 16\% = 80\%$ of the purchases were for the top three soft drinks.

TABLE 2.3 Relative Frequency and Percent Frequency Distributions of Soft Drink Purchases

Soft Drink	Relative Frequency	Percent Frequency
Coca-Cola	.38	38
Diet Coke	.16	16
Dr. Pepper	.10	10
Pepsi	.26	26
Sprite	.10	10
Total	1.00	100

Bar Chart

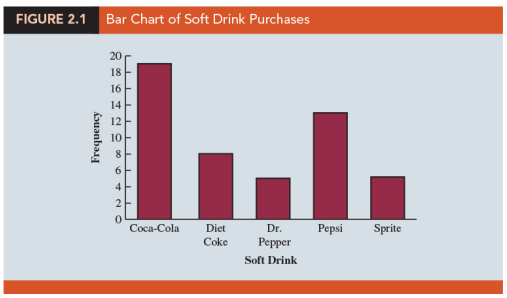
1. A bar chart is a graphical display for depicting _____ data summarized in a frequency, relative frequency, or percent frequency distribution.

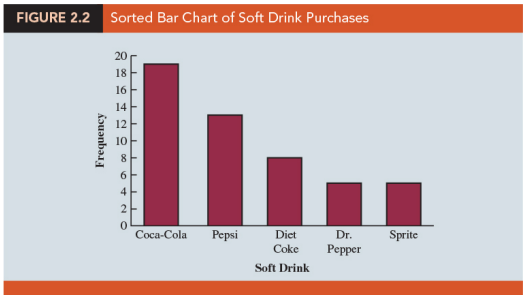
2. On one axis of the chart (usually the _____ axis), we specify the _____ that are used for the classes (categories). A frequency, relative frequency, or percent frequency scale can be used for the other axis of the chart (usually the _____ axis).

3. Using a bar of _____ drawn above each class label, we extend the length of the bar until we reach the frequency, relative frequency, or percent frequency of the class.

4. For categorical data, the bars should be separated to emphasize the fact that each category is _____.

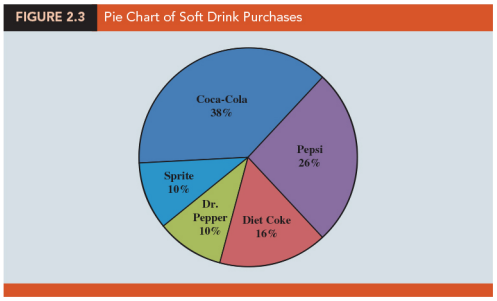
5. Example Soft Drinks Data (Figure 2.1)(Figure 2.2)

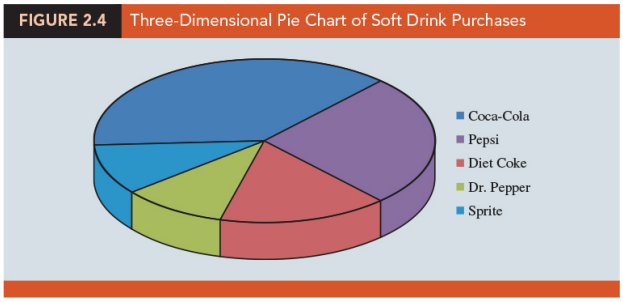




Pie Chart

1. The pie chart is a commonly used graphical display for presenting _____ and _____ for categorical data.
2. First draw a _____, then use the relative frequencies to subdivide the circle into _____ that correspond to the relative frequency for each class.
3. Because there are 360 degrees in a circle, a class with a relative frequency of 0.25 would consume _____ degrees of the circle.
4. In most cases, a bar chart is superior to a pie chart for displaying _____ data.
5. Numerous options involving the use of colors, shading, legends, text font, and three-dimensional perspectives are available to enhance the _____ of bar and pie charts.
6. In general, pie charts are not the best way to present percentages for _____.
7. Example Soft Drinks Data (Figure 2.3)(Figure 2.4)





2.2 Summarizing Data for a Quantitative Variable

Frequency Distribution

1. The three steps necessary to define the classes for a frequency distribution with _____ data are
   (a) **Determine the number of nonoverlapping classes**: Classes are formed by specifying ranges that will be used to group the data. The goal is to use

enough classes to show the _____ in the data, but not so many classes that some contain only a few data items.

(b) **Determine the width of each class**: As a general guideline, we recommend that the width be the same for each class:

Approximate class width = _____

The approximate class width can be _____ to a more convenient value based on the preference of the person developing the frequency distribution.

(c) **Determine the class limits**: Class limits must be chosen so that each data item belongs to _____ class.

2. With the number of classes, class width, and class limits determined, a frequency distribution can be obtained by _____ the number of data values belonging to each class.

3. Example Audit Time Data (Table 2.4.)(Table 2.5.)

TABLE 2.4 Year-End Audit Times (In Days)

12	14	19	18
15	15	18	17
20	27	22	23
22	21	33	28
14	18	16	13

(a) **Number of Classes**: The number of data items relatively small _____, we chose to develop a frequency distribution with _____.

(b) **Width of the Classes**: After deciding to use five classes, the width for each class is _____.

(c) **Class limits**: A total of five classes: _____, and _____.

TABLE 2.5 Frequency Distribution for the Audit Time Data

Audit Time (days)	Frequency
10–14	4
15–19	8
20–24	5
25–29	2
30–34	1
Total	20

(d) The _____ is the value halfway between the lower and upper class limits. For the audit time data, the five class midpoints are 12, 17, 22, 27, and 32.

Relative Frequency and Percent Frequency Distributions

1. The relative frequency is the _____ of the observations belonging to a class. With n observations,

   Relative frequency of the class = _____

2. The percent frequency of a class is the relative frequency _____.

3. Example Audit Time Data (Table 2.6.)

TABLE 2.6 Relative Frequency and Percent Frequency Distributions for the Audit Time Data

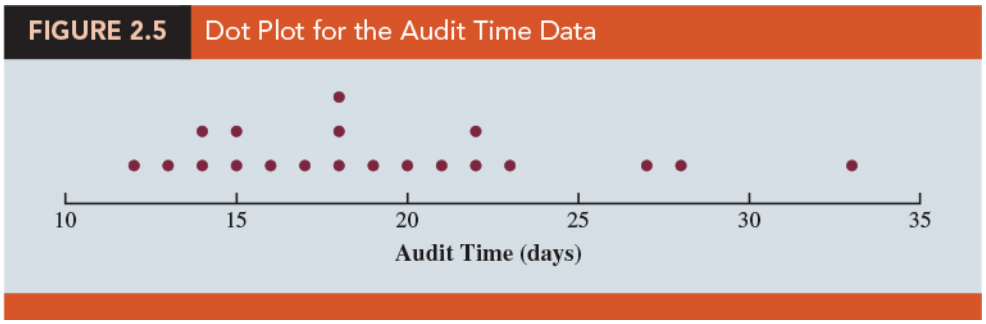
Audit Time (days)	Relative Frequency	Percent Frequency
10–14	.20	20
15–19	.40	40
20–24	.25	25
25–29	.10	10
30–34	.05	5
Total	1.00	100

Dot Plot

1. A horizontal axis of a a dot plot shows the _____ of data values.

2. Then each data value is represented by a dot placed above the axis. A vertical axis shows the _____ of data values.

3. Example Audit Time Data (Figure 2.5.)

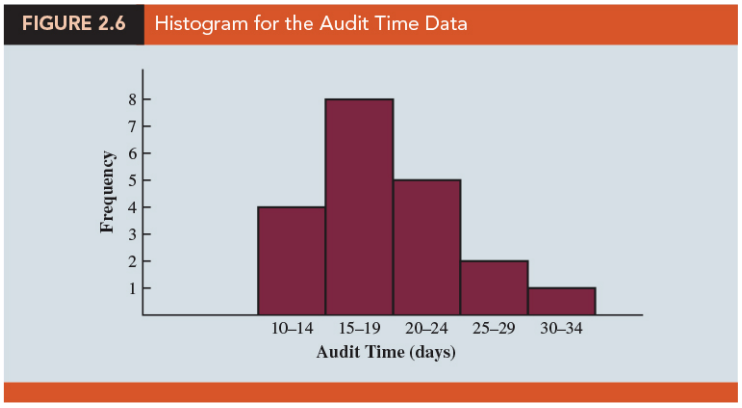


Histogram

1. A histogram is constructed by placing the variable of interest on the _____ axis and the frequency, relative frequency, or percent frequency on the _____ axis.

2. The frequency, relative frequency, or percent frequency of each class is shown by drawing a _____ whose base is determined by the _____ on the horizontal axis and whose _____ is the corresponding frequency, relative frequency, or percent frequency.

3. Example Audit Time Data (Figure 2.6.)
Note that the class with the greatest frequency is shown by the rectangle appearing above the class of _____ days. The height of the rectangle shows that the frequency of this class is _____.



4. The _____ rectangles of a histogram touch one another. Unlike a bar graph, a histogram contains no natural _____ between the rectangles of adjacent classes.

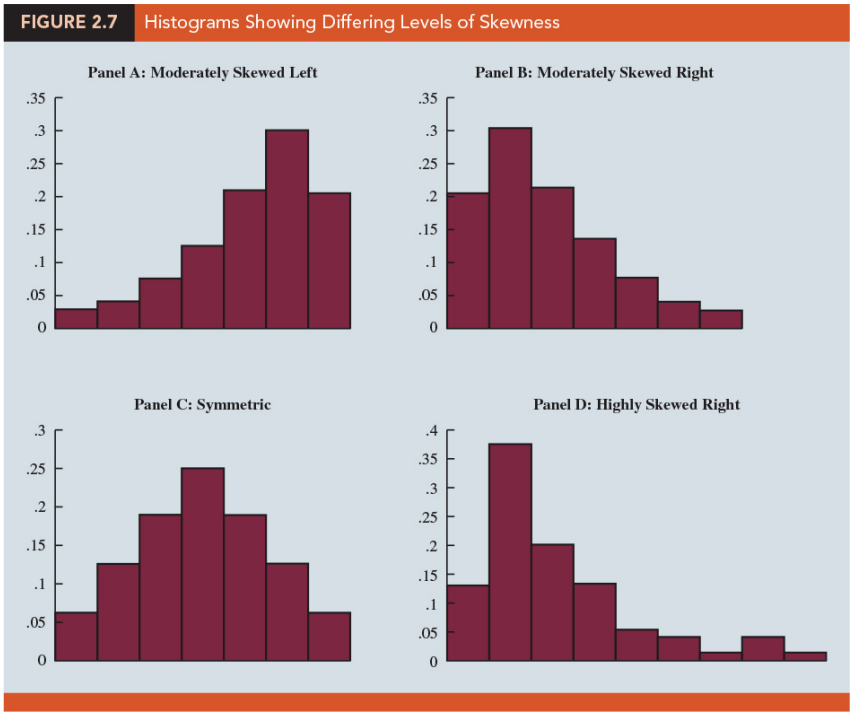
5. (Figure 2.7.) A histogram provides information about the _____, or _____, of a _____.

   (a) **Panel A**: A histogram is said to be _____ if its tail extends farther to the left. E.g., exam scores; with no scores above 100%, most of the scores above 70%, and only a few really low scores.

   (b) **Panel B**: A histogram is said to be _____ if its tail extends farther to the right. E.g., housing prices; a few expensive houses create the skewness in the right tail.

   (c) **Panel C**: In a _____ histogram, the left tail mirrors the shape of the right tail. Data for SAT scores, heights and weights of people, and so on lead to histograms that are _____ symmetric.

   (d) **Panel D**: a histogram _____. E.g., data on housing prices, salaries, purchase amounts, and so on often result in histograms skewed to the right.



Cumulative Distributions

1. The cumulative frequency distribution shows the number of data items with values _____ the upper class limit of each class.

2. Example Audit Time Data (Table 2.7.)
   The cumulative frequency for the class ("less than or equal to 24") is simply the sum of the frequencies for all classes with data values less than or equal to 24.

TABLE 2.7 Cumulative Frequency, Cumulative Relative Frequency, and Cumulative Percent Frequency Distributions for the Audit Time Data

Audit Time (days)	Cumulative Frequency	Cumulative Relative Frequency	Cumulative Percent Frequency
Less than or equal to 14	4	.20	20
Less than or equal to 19	12	.60	60
Less than or equal to 24	17	.85	85
Less than or equal to 29	19	.95	95
Less than or equal to 34	20	1.00	100

Stem-and-Leaf Display

1. A stem-and-leaf display is a graphical display used to show simultaneously the _____ order and _____ of a distribution of data.

2. Example Aptitude Test Data (Table 2.8.)
   These data result from a 150-question aptitude test given to 50 individuals recently interviewed for a position at Haskens Manufacturing. The data indicate the number of questions answered correctly.

TABLE 2.8 Number of Questions Answered Correctly on an Aptitude Test

112	72	69	97	107
73	92	76	86	73
126	128	118	127	124
82	104	132	134	83
92	108	96	100	92
115	76	91	102	81
95	141	81	80	106
84	119	113	98	75
68	98	115	106	95
100	85	94	106	119

6	8 9
7	2 3 3 5 6 6
8	0 1 1 2 3 4 5 6
9	1 2 2 2 4 5 5 6 7 8 8
10	0 0 2 4 6 6 6 7 8
11	2 3 5 5 8 9 9
12	4 6 7 8
13	2 4
14	1

3. **Stretched Stem-and-Leaf Display**: If we believe the original stem-and-leaf display has _____ the data too much, we can stretch the display vertically by using _____ for each leading digit(s).

6	8 9
7	2 3 3
7	5 6 6
8	0 1 1 2 3 4
8	5 6
9	1 2 2 2 4
9	5 5 6 7 8 8
10	0 0 2 4
10	6 6 6 7 8
11	2 3
11	5 5 8 9 9
12	4
12	6 7 8
13	2 4
13	
14	1

2.3 Summarizing Data for Two Variables Using Tables

Crosstabulation

1. A crosstabulation is a _____ of data for two variables. Both variables can be either _____ or _____. One variable is categorical and the other variable is quantitative are just as common.

2. Example Quality Rating and Meal Price Data for 300 Los Angeles Restaurants (Table 2.9.) (the first 10 restaurants)

TABLE 2.9 Quality Rating and Meal Price Data for 300 Los Angeles Restaurants

Restaurant	Quality Rating	Meal Price ($)
1	Good	18
2	Very Good	22
3	Good	28
4	Excellent	38
5	Very Good	33
6	Good	28
7	Very Good	19
8	Very Good	11
9	Very Good	23
10	Good	13
.	.	.
.	.	.
.	.	.

   (a) Quality rating is a _____ variable with rating categories of good, very good, and excellent. Meal price is a _____ variable that ranges from $10 to $49.

TABLE 2.10 Crosstabulation of Quality Rating and Meal Price Data for 300 Los Angeles Restaurants

	Meal Price				
Quality Rating	$10–19	$20–29	$30–39	$40–49	Total
Good	42	40	2	0	84
Very Good	34	64	46	6	150
Excellent	2	14	28	22	66
Total	78	118	76	28	300

(b) The right and bottom margins of the crosstabulation provide the _____ for quality rating and meal price separately.

(c) Note that the values in the relative frequency column do not add exactly to 1.00 and the values in the percent frequency distribution do not add exactly to 100; the reason is that the values being summed are _____.

(d) Restaurants with _____ meal prices received _____ quality ratings than restaurants with lower meal prices.

3. When quantitative variables are used, however, we must first _____ for the values of the variable.

4. Example Quality Rating and Meal Price Data (Table 2.11.)
For row percentages, the results of dividing each frequency in Table 2.10 by its corresponding row total are shown in Table 2.11. Each row of Table 2.11 is a percent frequency distribution of meal price for one of the quality rating categories.

TABLE 2.11 Row Percentages for Each Quality Rating Category

	Meal Price				
Quality Rating	$10–19	$20–29	$30–39	$40–49	Total
Good	50.0	47.6	2.4	.0	100
Very Good	22.7	42.7	30.6	4.0	100
Excellent	3.0	21.2	42.4	33.4	100

Simpson's Paradox

1. The data in two or more crosstabulations are often combined or aggregated to produce a summary crosstabulation showing how two variables are related. In such

cases, conclusions drawn from two or more separate crosstabulations can be reversed when the data are aggregated into a single crosstabulation. The _____ based on aggregate and unaggregated data is called _____.

2. Example Two Judges in Two different Courts

   (a) Judges Ron Luckett and Dennis Kendall presided over cases in Common Pleas Court (民事訴訟法庭) and Municipal Court (市政法庭) during the past three years. Some of the verdicts they rendered were appealed. In most of these cases the appeals (上訴) court upheld (堅持) the original verdicts, but in some cases those verdicts were reversed.

   (b) For each judge a crosstabulation was developed based upon two variables: Verdict (upheld or reversed) and Type of Court (Common Pleas and Municipal).

	Judge		
Verdict	Luckett	Kendall	Total
Upheld	129 (86%)	110 (88%)	239
Reversed	21 (14%)	15 (12%)	36
Total (%)	150 (100%)	125 (100%)	275

   (c) This crosstabulation shows the number of appeals in which the verdict was upheld and the number in which the verdict was reversed for both judges. A review of the column percentages shows that 86% of the verdicts were upheld for Judge Luckett, while 88% of the verdicts were upheld for Judge Kendall.

   (d) From this aggregated crosstabulation, we conclude that Judge Kendall is doing the better job because a greater percentage of Judge Kendall′ s verdicts are being upheld.

   (e) When we unaggregate the data, we see that Judge Luckett has a better record because a greater percentage of Judge Luckett's verdicts are being upheld in both courts.

Judge Luckett

Verdict	Common Pleas	Municipal Court	Total
Upheld	29 (91%)	100 (85%)	129
Reversed	3 (9%)	18 (15%)	21
Total (%)	32 (100%)	118 (100%)	150

Judge Kendall

Verdict	Common Pleas	Municipal Court	Total
Upheld	90 (90%)	20 (80%)	110
Reversed	10 (10%)	5 (20%)	15
Total (%)	100 (100%)	25 (100%)	125

(f) Note that for both judges the percentage of appeals that resulted in reversals was much higher in Municipal Court than in Common Pleas Court. Because Judge Luckett tried a much higher percentage of his cases in Municipal Court, the aggregated data favored Judge Kendall. When we look at the crosstabulations for the two courts separately, however, Judge Luckett shows the better record.

(g) Thus, for the original crosstabulation, we see that _____ is a hidden variable that cannot be ignored when evaluating the records of the two judges.

3. Before drawing a conclusion, you may want to investigate whether the _____ or _____ form of the crosstabulation provides the better insight and conclusion.

2.4 Summarizing Data for Two Variables Using Graphical Displays

Scatter Diagram and Trendline

1. A scatter diagram (scatter plot) is a graphical presentation of the overall relationship between two _____ variables.

2. One variable is shown on the horizontal axis and the other variable is shown on the vertical axis.

3. A trendline provides an _____ of the relationship.

4. Example Electronics Store in San Francisco (Table 2.14.)

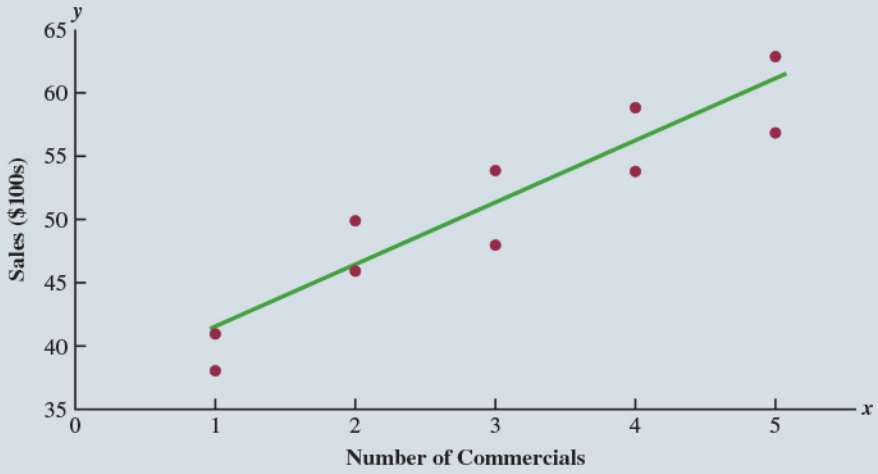
   (a) Consider the advertising/sales relationship for an electronics store in San Francisco. On 10 occasions during the past three months, the store used weekend television commercials to promote sales at its stores. The managers want to investigate whether a relationship exists between the number of commercials shown and sales at the store during the following week. Sample data for the 10 weeks with sales in hundreds of dollars are shown in Table 2.14.

TABLE 2.14 Sample Data for the San Francisco Electronics Store

Week	Number of Commercials x	Sales ($100s) y
1	2	50
2	5	57
3	1	41
4	3	54
5	4	54
6	1	38
7	5	63
8	3	48
9	4	59
10	2	46

   (b) The scatter diagram in Figure 2.8 indicates a _____ relationship between the number of commercials and sales. Higher sales are associated with a higher number of commercials.

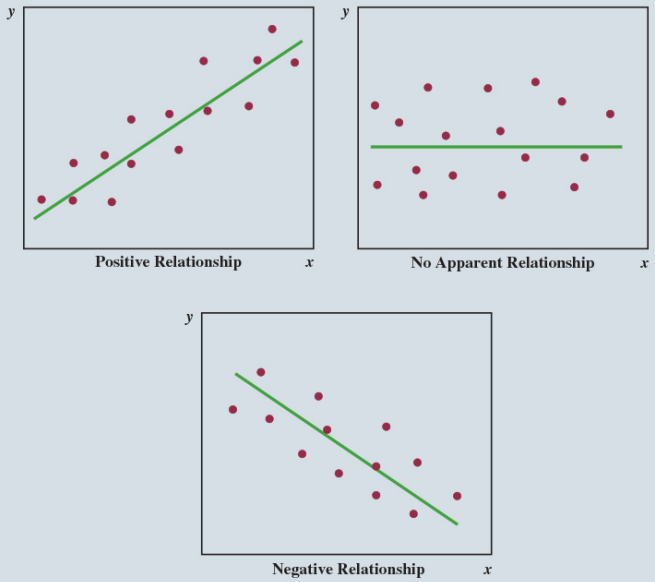
FIGURE 2.8 Scatter Diagram and Trendline for the San Francisco Electronics Store



(c) The relationship is not perfect in that all points are not on a _____. However, the general pattern of the points and the trendline suggest that the overall relationship is positive.

5. Some general scatter diagram patterns and the types of relationships they suggest are shown in Figure 2.9.

FIGURE 2.9 Types of Relationships Depicted by Scatter Diagrams



Side-by-Side Bar Chart

1. A side-by-side bar chart is a graphical display for depicting _____ bar charts on the same display.

2. Each _____ represents one value of the first variable.

3. Each _____ represents one value of the second variable.

4. Example Quality Rating and Meal Price Data (Figure 2.10.)

   (a) We see that the lowest meal price category ($10–$19) received mostly good and very good ratings, but very few excellent ratings. The highest price category ($40–$49) received mostly excellent ratings, some very good ratings, but no good ratings.

   (b) Notice that as the price increases (left to right), the height of the light blue bars decreases and the height of the dark blue bars generally increases. This indicates that as price _____, the quality rating tends to be _____.

   (c) The very good rating, as expected, tends to be more prominent in the middle price categories as indicated by the dominance of the middle bar in the moderate price ranges of the chart.

FIGURE 2.10 Side-By-Side Bar Chart for the Quality and Meal Price Data



Stacked Bar Chart

1. A stacked bar chart is a bar chart in which each bar is broken into _____ of a different color.

2. If percentage frequencies are displayed, all bars will be of the same height (or length), extending to the 100% mark.

3. Example Quality Rating and Meal Price Data (Table 2.15.)(Figure 2.11.)

   (a) Table 2.15 shows the column percentages for each meal price category.

   (b) Figure 2.11 shows even more clearly than Figure 2.10 the relationship between the variables. As we move from the low price category ($10–19) to the high price category ($40–49), the length of the light blue bars decreases and the length of the dark blue bars increases.

TABLE 2.15 Column Percentages for each Meal Price Category

	Meal Price			
Quality Rating	$10–19	$20–29	$30–39	$40–49
Good	53.8%	33.9%	2.6%	.0%
Very Good	43.6	54.2	60.5	21.4
Excellent	2.6	11.9	36.8	78.6
Total	100.0%	100.0%	100.0%	100.0%

FIGURE 2.11 Stacked Bar Chart for Quality Rating and Meal Price Data



2.5 Data Visualization: Best Practices in Creating Effective Graphical Displays

1. Data visualization is a term used to describe the use of graphical displays to _____ and _____ information about a data set.
2. The goal of data visualization is to _____ as effectively and clearly as possible, the key information about the data.

Creating Effective Graphical Displays

1. Give the display a clear and concise _____.
2. Keep the display _____. Do not use three dimensions when two dimensions are sufficient.
3. Clearly label each _____ and provide the units of measure.
4. If color is used to distinguish categories, make sure the colors are _____.
5. If multiple colors or line types are used, use a _____ to define how they are used and place the legend close to the representation of the data.

Choosing the Type of Graphical Display

1. **Displays used to show the distribution of data**:
   (a) _____ to show the frequency distribution and relative frequency distribution for categorical data
   (b) _____ to show the relative frequency and percent frequency for categorical data
   (c) _____ to show the distribution for quantitative data over the entire range of the data
   (d) _____ to show the frequency distribution for quantitative data over a set of class intervals

(e) _____ to show both the rank order and shape of the distribution for quantitative data

2. **Displays used to make comparisons**:

(a) _____ to compare two variables

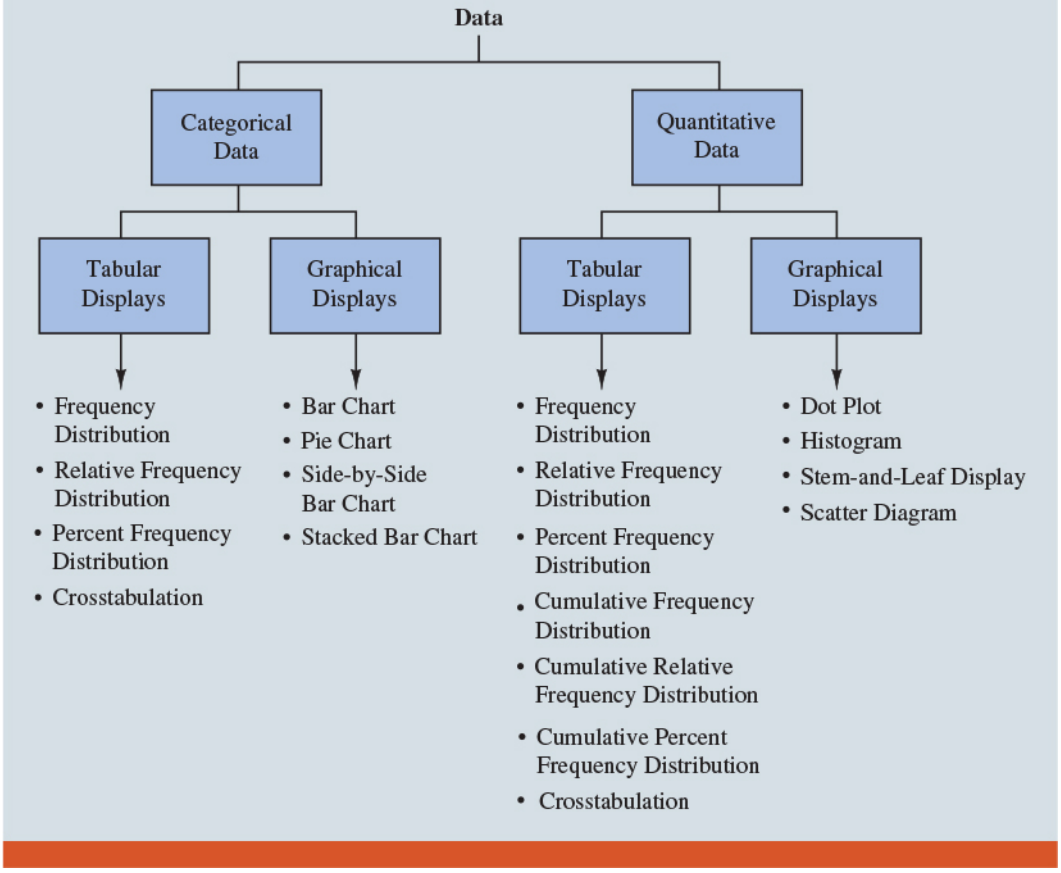
(b) _____ to compare the relative frequency or percent frequency of two categorical variables

3. **Displays used to show relationships**:

(a) _____ to show the relationship between two quantitative variables

(b) _____ to approximate the relationship of data in a scatter diagram

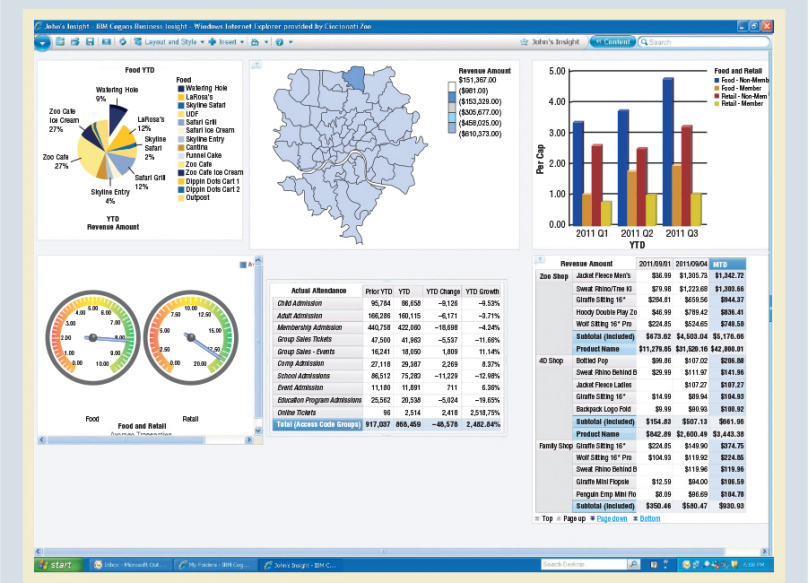
FIGURE 2.16 Tabular and Graphical Displays for Summarizing Data



Data Dashboards

1. A data dashboard is a widely used data visualization tool.
2. It organizes and presents _____ (KPIs) used to monitor an organization or process.
3. It provides timely summary information that is easy to read, understand, and interpret.
4. Some additional guidelines include:
   (a) Minimize the need for screen scrolling.
   (b) Avoid unnecessary use of color or 3D displays.
   (c) Use borders between charts to improve readability.

FIGURE 2.14 Data Dashboard for the Cincinnati Zoo



NOTE : use "misleading graphs" as the search keyword on Google or YouTube.

😍 EXERCISES

2.1 : 3, 4, 5, 10

2.2 : 11, 17, 20, 25

2.3 : 31, 33, 34

2.4 : 39, 41, 43

2.5 : 49, 54, 55, 56

名言佳句

“因為每個人在某些方面都會發生失敗，所以我能夠接受失敗，但我不能接受沒有嘗試”
“I can accept failure, everyone fails at something. But I can’t accept not trying”
— *Michael Jordan (February 17, 1963 –)*

統計小故事：孟加拉救災中的表格與圖表

1970 年代，孟加拉國遭遇毀滅性洪水，數百萬人受災。國際社會雖然紛紛伸出援手，但問題很快浮現：救援資源有限，如何決定哪些村莊應該優先獲得糧食和藥品？

當時，一位年輕的經濟學家——阿姆亞特．森（Amartya Sen）——參與了分析工作。他後來因「貧窮與饑荒研究」獲得諾貝爾經濟學獎，但這次救災就是他實踐理論的起點。

在災後的混亂中，政府官員習慣以「死亡人數」來衡量災情。但森敏銳地指出，死亡數字往往滯後，無法反映「未來即將爆發的危機」。他建議建立一套系統化的資料表格，收集每個村莊的人口數、糧食存量、兒童營養不良比例、家庭受災比例。

這些資料一旦以「表格」形式整理出來，災情就不再只是零碎的消息，而是可比對、可排序的資訊。森進一步利用條狀圖與熱點地圖，將數據轉換成直觀的圖像。決策者不再需要翻閱厚重的文件，只需一眼就能看出哪些村莊最危險。

例如，一個村莊可能死亡數字還不多，但表格顯示它的糧食存量只夠三天，兒童營養不良比例高達 60%。若沒有這些資料，它可能會被忽略；但透過描述統計的整理，決策者能立刻判斷，這個地區應該優先獲得糧食投放。

結果，援助資源被更有效地分配，許多可能的饑荒得以避免。森後來在研究中強調：描述統計的表格與圖像化展示，不僅僅是學術技巧，而是影響生死的工具。

對學生來說，這正好對應第二章的主題。表格能讓我們看見資料的「結構」，圖表能讓我們理解資料的「趨勢」。當資料以清楚的方式呈現時，它的力量遠比單純的文字描述來得巨大。

回頭看，我們或許會發現：統計的價值從來不是停留在紙上，而是幫助人們「看見現實，做出選擇」。在孟加拉救災的故事裡，描述統計拯救了無數生命；在我們的課堂裡，它將是學生理解數據世界的第一步。

統計學 (一)

Anderson's Statistics for Business & Economics (14/E)

Chapter 3: Descriptive Statistics: Numerical Measures

上課時間地點: 二 D56, 商館 260206

授課教師: 吳漢銘 (國立政治大學統計學系副教授)

教學網站: `https://hmwu.idv.tw`

Overview

1. Develop numerical _____ for data sets consisting of a single variable: location, variability, and distribution shape.
2. Other topics: relative location, detecting outliers, five-number summaries and box plots, the measures of association between two variables,
3. _____: if the measures are computed for data from a sample.
4. _____: if the measures are computed for data from a population.
5. A sample statistic is referred to as the _____ of the corresponding population parameter.

3.1 Measures of Location

Mean

1. The mean of a data set is the _____ of all the data values _____.
2. The sample mean _____ is the point estimator of the population mean _____.
3. The mean provides a measure of _____ location.

Weighted Mean

1. In some instances, the mean is computed by giving each observation a weight _____ that reflects its relative importance.
2. The choice of weights depends on the application.
3. The weighted sample mean
$$\bar{x} = _____$$
4. Example The computation of a grade point average (GPA) for college students.
   (a) The data values generally used are 4 for an A grade, 3 for a B grade, 2 for a C grade, 1 for a D grade, and 0 for an F grade.
   (b) The weights are the number of credit hours earned for each grade.

Median

1. The median of a data set is the value in the _____ when the data items are arranged in _____ order (smallest value to largest value).
   (a) For an odd number of observations, the median is the _____ value.
   (b) For an even number of observations, the median is the _____ of the two middle values.
2. Whenever a data set has _____, the median is the preferred measure of central location than mean.
3. The median is the measure of location most often reported for annual income and property value data.

Geometric Mean

1. The geometric mean is calculated by finding the nth root of the product of n values.
$$\bar{x}_g = _____$$
2. It is often used in analyzing _____ in financial data (where using the _____ will provide misleading results).

3. The sample mean is only appropriate for an _____ process. For a _____ process, such as applications involving growth rates, the geometric mean is the appropriate measure of location.

4. While the applications of the geometric mean to problems in finance, investments, and banking are particularly common, the geometric mean should be applied any time you want to determine the _____ over several successive periods.

✎ Question ..... (p113)

Consider Table 3.2, which shows the percentage annual returns, or growth rates, for a mutual fund over the past 10 years. (a) Compute how much $100 invested in the fund at the beginning of year 1 would be worth at the end of year 10. (b) What was the mean percentage annual return or mean rate of growth for this investment over the 10-year period?

TABLE 3.2 Percentage Annual Returns and Growth Factors for the Mutual Fund Data

Year	Return (%)	Growth Factor
1	−22.1	.779
2	28.7	1.287
3	10.9	1.109
4	4.9	1.049
5	15.8	1.158
6	5.5	1.055
7	−37.0	.630
8	26.5	1.265
9	15.1	1.151
10	2.1	1.021

sol:

Mode

1. The mode of a data set is the value that occurs with _____.
2. The greatest frequency can occur at two or more different values.
3. If the data have exactly two modes, the data are _____.
4. If the data have more than two modes, the data are _____.

Percentiles

1. A percentile provides information about how the data are _____ from the smallest value to the largest value.
2. The _____ percentile of a data set is a value such that _____ percent of the items take on this value or _____ and _____ of the items take on this value or _____.
3. To calculate the pth percentile for a data set containing n observations,:
   (a) Arrange the data in _____ order.
   (b) Compute L_p, the location of the pth percentile

$$L_p = _____$$

4. The 50th percentile is also the _____.
5. Admission test scores for colleges and universities are frequently reported in terms of _____.

Question ..... (p116)

Compute and interpret the 80th percentile for the starting salary data in Table 3.1.

TABLE 3.1 Monthly Starting Salaries for a Sample of 12 Business School Graduates

Graduate	Monthly Starting Salary ($)	Graduate	Monthly Starting Salary ($)
1	5850	7	5890
2	5950	8	6130
3	6050	9	5940
4	5880	10	6325
5	5755	11	5920
6	5710	12	5880

sol:

Quartiles

1. Quartiles are specific _____.
2. Q_1: _____ Quartile = _____ Percentile
3. Q_2: _____ Quartile = _____ Percentile = Median
4. Q_3: _____ Quartile = _____ Percentile

Question ..... (p116)

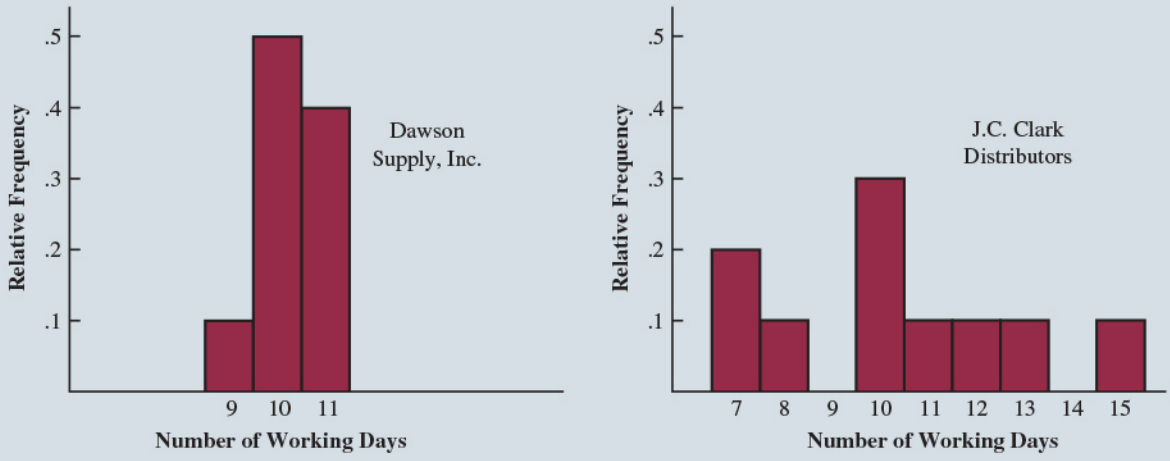
Compute the first and third quartiles for the starting salary data in Table 3.1.

sol:

3.2 Measures of Variability

1. It is often desirable to consider measures of variability (_____), as well as measures of location.
2. For example, in choosing supplier A or supplier B we might consider not only the average delivery time for each, but also the variability in delivery time for each.

FIGURE 3.2 Historical Data Showing the Number of Days Required to Fill Orders



Range

1. The range of a data set is the _____ between the largest and smallest data value.

   Range = _____

2. It is very sensitive to the smallest and largest data values.

Interquartile Range (IQR)

1. The interquartile range of a data set is the difference between the _____ quartile and the _____ quartile: _____.
2. It is the _____ for the middle 50% of the data.
3. It overcomes the sensitivity to _____ data values.

Variance

1. The variance is a measure of variability that utilizes _____ the data.
2. _____: the difference between the value of each observation (x_i) and the mean (_____ for a sample, _____ for a population): ($x_i - \bar{x}$, $x_i - \mu$).
3. For any data set, the sum of the deviations about the mean will always equal _____: $\sum(x_i - \bar{x}) = 0$.
4. The variance is the _____ between each data value and the mean.
5. The variance for a population is: $\sigma^2 =$ _____.
6. The variance of a sample is: $s^2 =$ _____.

Standard Deviation

1. The standard deviation of a data set is the _____ of the variance.
2. It is measured in the _____ as the original data, making it more easily interpreted than the variance.
3. The standard deviation of a sample is: _____.
4. The standard deviation of a population is: _____.

Coefficient of Variation

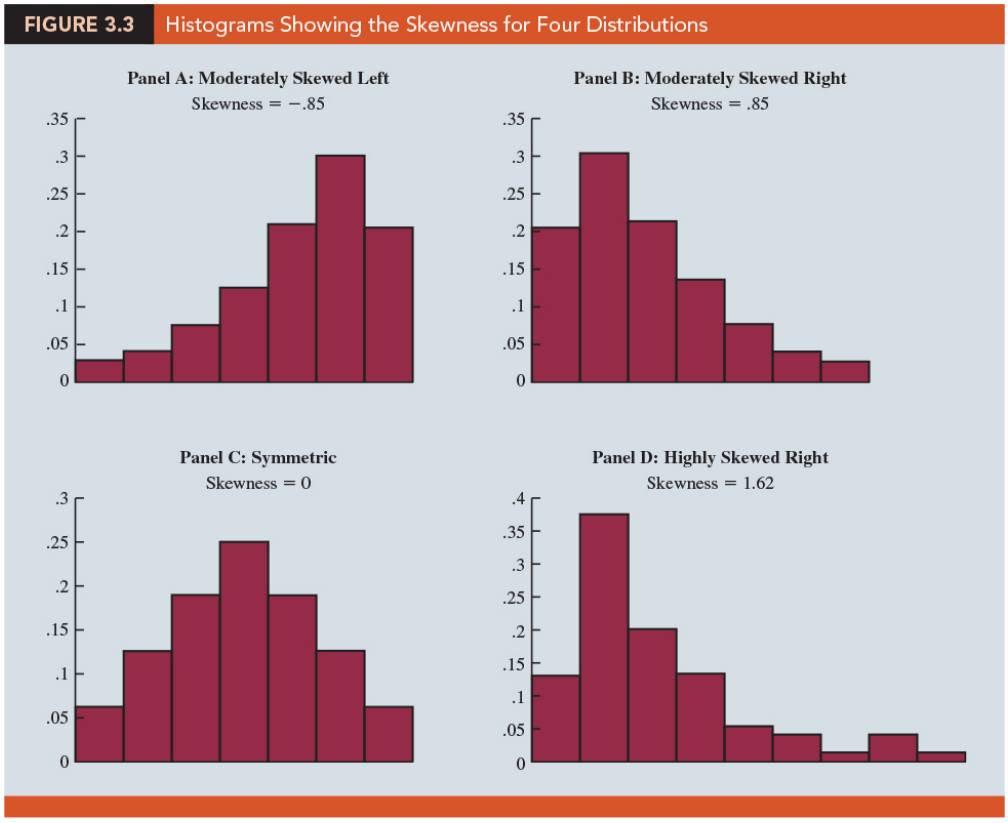
1. The coefficient of variation indicates how large the _____ is relative to the _____.
2. The coefficient of variation = _____.
3. In general, the coefficient of variation is a useful statistic for comparing the _____ of variables that have different standard deviations and different means.

3.3 Measures of Distribution Shape, Relative Location, and Detecting Outliers

Distribution Shape

1. A _____ provides a graphical display showing the shape of a distribution.
2. An important numerical measure of the shape of a distribution is called _____.
3. The formula for the skewness of sample data is

$$\text{Skewness} = \frac{\qquad\qquad\qquad\qquad}{\qquad\qquad\qquad\qquad}.$$



4. For data skewed to the left, the skewness is _____; for data skewed to the right, the skewness is _____. If the data are symmetric, the skewness is _____.

5. For a symmetric distribution, the mean and the median are _____. When the data are _____ skewed, the mean will usually be _____ than the median.

z-Scores

1. The z-score is often called the _____ value: $z_i =$ _____.
2. It denotes the _____ of standard deviations a data value x_i is from the _____.
3. An observation's z-score is a measure of the _____ of the observation in a data set.
4. A data value equal to the sample mean will have a z-score of _____.
5. A data value greater (less) than the sample mean will have a z-score greater (less) than zero.

Chebyshev's Theorem

1. At least _____ of the data values must be within _____ standard deviations of the mean, where z is any value greater than 1.
2. Chebyshev's theorem enables us to make statements about the _____ of data values that must be within a specified number of standard deviations of the mean.
3. Chebyshev's theorem requires _____; but z need not be an _____.
4. At least _____ of the data values must be within _____ standard deviations of the mean.

✎ Question ..... (p132)

Suppose that the midterm test scores for 100 students in a college business statistics course had a mean of 70 and a standard deviation of 5. How many students had test scores between 60 and 80? How many students had test scores between 58 and 82? (Using Chebyshev's theorem)

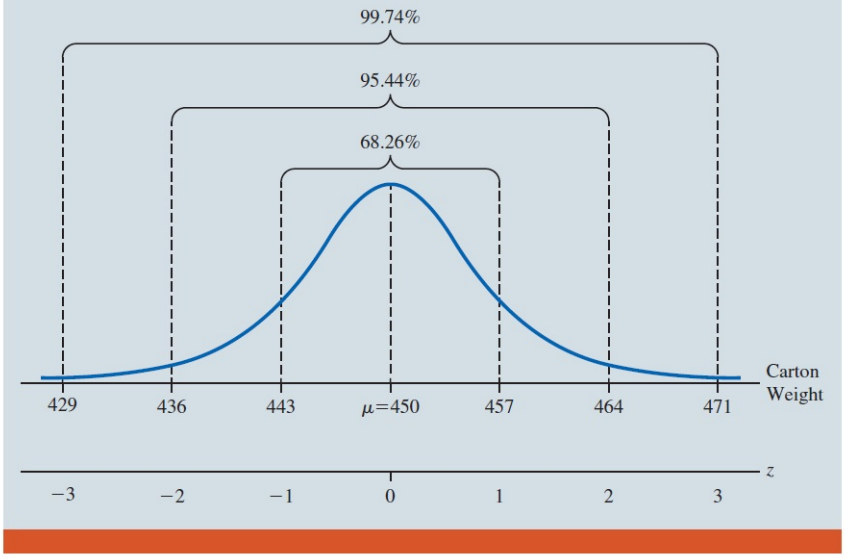
sol:

Empirical Rule

1. When the data are believed to approximate a _____ distribution (e.g., _____ distribution), the empirical rule can be used to determine the percentage of data values that must be within a specified number of standard deviations of the mean.

2. Approximately _____ of the data values will be within _____ standard deviation of the mean.

FIGURE 3.5 A Bell-Shaped Distribution of Detergent Carton Weights with Percentage of Data Values Within 1, 2, and 3 Standard Deviations



✎ Question ..... (p134)

The liquid detergent cartons are filled automatically on a production line. Filling weights frequently have a bell-shaped distribution. If the mean filling weight is 16 ounces and the standard deviation is .25 ounces, how many filled cartons will (a) weights between 16 and 16.25 ounces? (b) weights between 15.50 and 16 ounces? (c) weights less than 15.50 ounces? (d) weights between 15.50 and 16.25 ounces?

sol:

Detecting Outliers

1. An _____ is an unusually small or unusually large value in a data set.
2. A data value with a _____ less than _____ or greater than _____ might be considered an outlier.
3. Another approach to identifying outliers is based upon the values of the first and third quartiles (Q_1 and Q_3) and the interquartile range (IQR).

   Lower Limit = _____, Upper Limit = _____
4. An observation is classified as an outlier if its value is _____ than the Lower limit or _____ than the Upper limit.

3.4 Five-Number Summaries and Box Plots

Five-Number Summary

1. Five-Number Summary: smallest value, _____, _____, third quartile (Q_3), and largest value.

   ✎ Question ..... (p138)

   The monthly starting salary data is shown in ascending order below,

   5710 5755 5850 5880 5880 5890 5920 5940 5950 6050 6130 6325

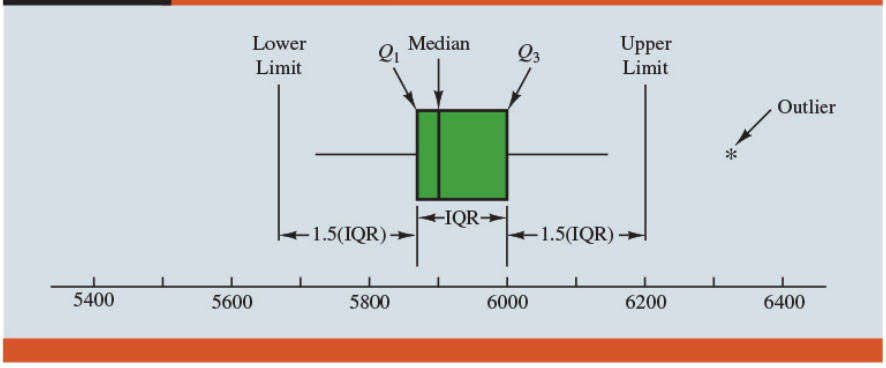
   Find the five-number summary.

   sol:

Box Plot

1. A _____ is a graphical display of data that is based on a five-number summary.

FIGURE 3.6 Boxplot of the Monthly Starting Salary Data with Lines Showing the Lower and Upper Limits



2. Box plots provide another way to identify _____.

   (a) A _____ is drawn with its ends located at the first and third quartiles.

   (b) A vertical line is drawn in the box at the location of the _____ (second quartile).

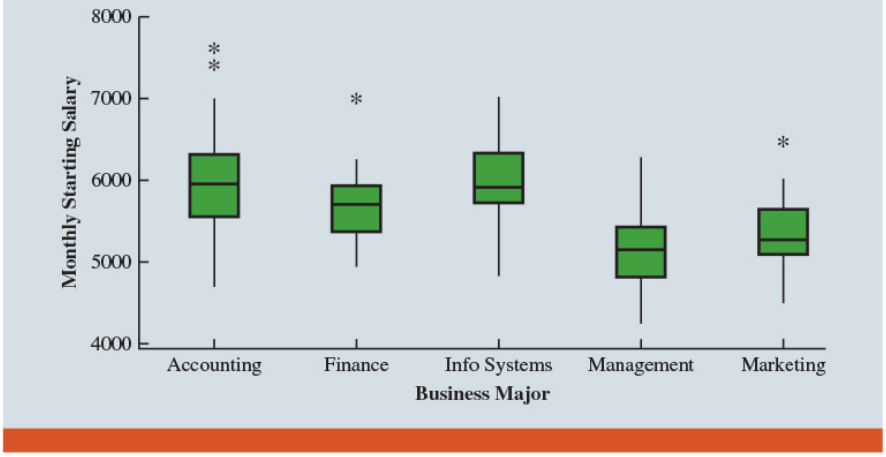
   (c) Box limits are located using the _____.

   (d) Data outside the box _____ are considered _____ (shown with the symbol _____).

   (e) The horizontal lines extending from each end of the box in Figure 3.6 are called _____. The whiskers are drawn from the ends of the box to the smallest and largest values inside the limits.

3. Boxplots can also be used to provide a graphical summary of two or more groups and facilitate _____ among the _____.

FIGURE 3.8 Comparative Boxplots of Monthly Starting Salary by Major



3.5 Measures of Association Between Two Variables

Covariance

1. The covariance is a measure of the _____ between two variables. For a sample of size n with the observations $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$, the sample covariance is defined as follows:

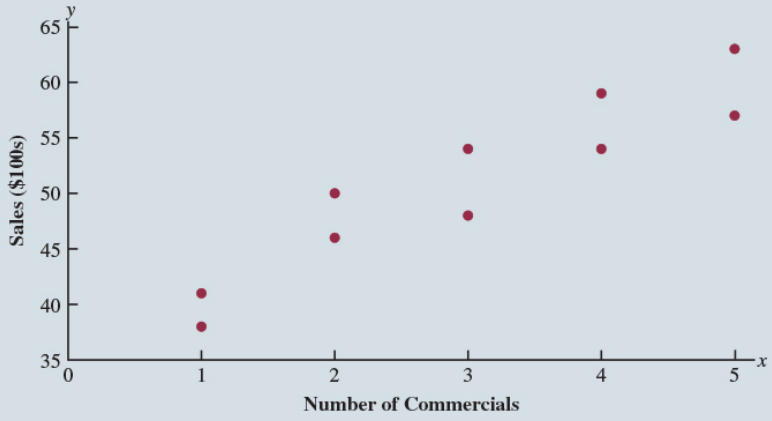
$$S_{xy} = _____$$

2. Let _____ be the population mean of the variable x and _____ be the population mean of the variable y. The population covariance σ_{xy} is defined for a population of size N:

$$\sigma_{xy} = _____$$

3. Positive _____ values of covariance indicate a positive _____ relationship.

FIGURE 3.9 Scatter Diagram for the San Francisco Electronics Store



Question ..... (p143)

Compute the sample covariance to measure the strength of the linear relationship between the number of commercials x and the sales volume y in the San Francisco electronics store problem.

sol:

TABLE 3.7 Calculations for the Sample Covariance

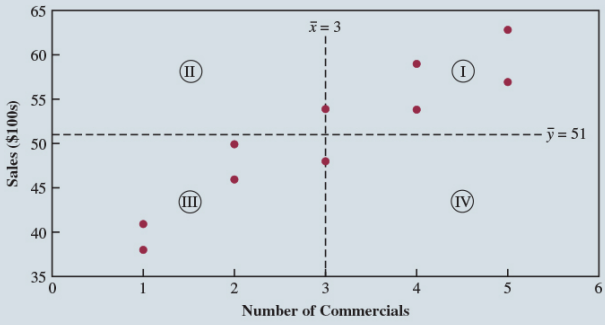
	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
	2	50	−1	−1	1
	5	57	2	6	12
	1	41	−2	−10	20
	3	54	0	3	0
	4	54	1	3	3
	1	38	−2	−13	26
	5	63	2	12	24
	3	48	0	−3	0
	4	59	1	8	8
	2	46	−1	−5	5
Totals	30	510	0	0	99

$$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{99}{10-1} = 11$$

Interpretation of the Covariance

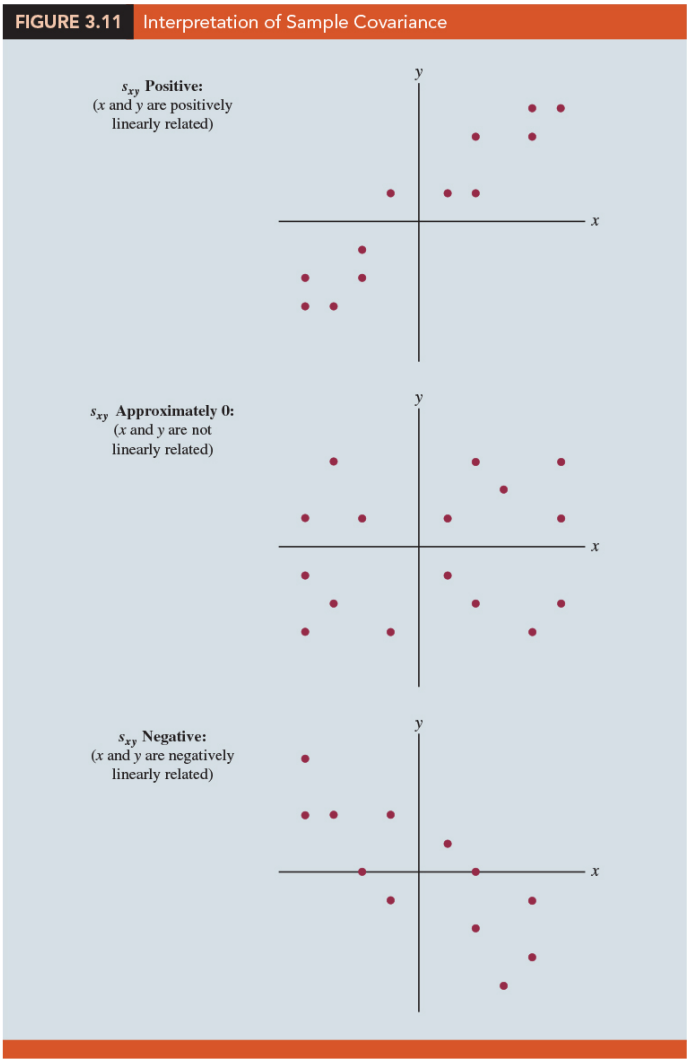
1. (Figure 3.10) Points in quadrant I correspond to x_i _____ than $\bar{x}$ and y_i _____ than $\bar{y}$.
2. If the value of s_{xy} is positive (_____), the points with the greatest influence on s_{xy} must be in quadrants I and III (_____). Hence, a positive (_____) value for s_{xy} indicates a positive (_____) linear association between x and y; that is, as the value of x increases (_____), the value of y increases (_____).
3. The points are evenly distributed across all four quadrants, the value of s_{xy} will be close to zero, indicating _____ between x and y.

FIGURE 3.10 Partitioned Scatter Diagram for the San Francisco Electronics Store



4. (Figure 3.11) the values of s_{xy} can be expected with three different types of scatter diagrams.

FIGURE 3.11 Interpretation of Sample Covariance



5. However, one problem with using covariance as a measure of the strength of the linear relationship is that the value of the covariance depends on the _____ for x and y.

6. A measure of the relationship between two variables that is not affected by the units of measurement for x and y is the _____.

Correlation Coefficient

1. Correlation is a measure of linear association and not necessarily _____.

2. The sample Pearson product moment correlation coefficient:

$$r_{xy} = \frac{\quad}{\quad},$$

where s_x (s_y) is the sample standard deviation of x (y).

3. The population Pearson product moment correlation coefficient:

$$\rho_{xy} = \frac{\quad}{\quad},$$

where σ_x (σ_y) is the population standard deviation of x (y).

4. The coefficient can take on values between _____.

5. Values near _____ (+1) indicate a _____ (positive) _____ relationship.

6. The closer the correlation is to _____, the weaker the relationship.

7. NOTE: A high correlation between two variables does not mean that changes in one variable will cause changes in the other variable.

✎ Question ..... (p147)

Let $(x_i, y_i), i = 1, 2, 3$ be the $(5, 10), (10, 30), (15, 50)$. Compute the sample correlation between x and y.

sol:

TABLE 3.8 Computations Used in Calculating the Sample Correlation Coefficient

	x_i	y_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
	5	10	−5	25	−20	400	100
	10	30	0	0	0	0	0
	15	50	5	25	20	400	100
Totals	30	90	0	50	0	800	200
	$\bar{x} = 10$	$\bar{y} = 30$					

3.6 Data Dashboards: Adding Numerical Measures to Improve Effectiveness*

😍 EXERCISES

3.1 : 5, 6, 10

3.2 : 25, 32, 34

3.3 : 38, 41, 43, 45

3.4 : 46, 47, 50

3.5 : 55, 59, 60

SUP : 64, 67, 71

名言佳句

“學習知識要善於思考，思考，再思考，我就是這個方法成為科學家的”
“Learning knowledge must be good at thinking, thinking, and thinking, I become a scientist by this method.”

— *Albert Einstein (March 14, 1879 – April 18, 1955)*

統計小故事：保險業的平均數與變異數

十九世紀的英國，工業革命帶來了社會與經濟的劇烈轉變。人口迅速膨脹，城市中貧富差距擴大，工人階級的生活環境惡劣，疾病與死亡率居高不下。此時，壽險與年金制度開始逐漸興起。人們渴望透過保險來對抗人生的不確定性，但一個問題始終存在：保險公司該如何設計合理的保費？

這並不是簡單的商業問題，而是一個統計問題。若保費過低，公司會因為賠付過多而倒閉；若保費過高，民眾就不願投保。要在兩者之間找到平衡，必須依靠「數據」與「數值衡量」。

在這樣的背景下，出現了一位重要的統計學家——威廉．法爾 (William Farr)。他是英國國家統計局的核心人物，被視為現代公共衛生統計的先驅之一。法爾的貢獻之一，就是將「數字」與「人類壽命」結合，建立了系統性的生命表 (life table)。

法爾收集了大量死亡紀錄，根據不同年齡層整理出生存與死亡的機率。他發現，單純計算「平均壽命」遠遠不夠。舉例來說，若某一群人的平均壽命是 50 歲，並不代表每個人都活到 50 歲。有些人可能在 30 歲早逝，有些人卻能活到 80 歲。若只看平均數，我們會忽略掉群體內部的巨大差異。

因此，法爾特別強調變異數與標準差的意義。變異數告訴我們數據分散的程度，而標準差則讓我們更直觀地理解這種波動。這些概念在保險計算中尤其重要。因為保險公司需要知道的不是「平均壽命」，而是「可能的壽命分布」。只有掌握分布的範圍與變異，他們才能合理設計保費，既不至於承擔過大的風險，也不會讓顧客覺得不公平。

法爾的生命表後來成為現代精算學的基礎。今天的保險公司依然依靠同樣的邏輯：收集數據 $\rightarrow$ 計算平均值 $\rightarrow$ 衡量變異 $\rightarrow$ 建立壽命分布模型 $\rightarrow$ 制定保費方案。甚至退休金制度、社會保險的設計，也都離不開這套方法。

除了保險業，數值衡量的影響還擴展到其他領域。例如，在教育中，平均成績告訴我們全班的學習水準，但標準差則揭示學生成績是否集中或分散。在投資中，平均報酬率表示可能的收益，而標準差則代表投資的風險程度。若僅看平均值，投資者可能誤以為報酬穩定，但若變異數極高，就意味著風險巨大。

這些例子都指向同一個道理：平均數是描述「中心」，變異數則是理解「不確定性」的關鍵。缺一不可。

教材呼應上，第三章要學生掌握的正是這些核心數字：平均數、中位數、眾數、變異數、標準差。它們看似簡單，但背後支撐著整個保險業、金融業、教育與公共政策。

當法爾在十九世紀設計出第一份生命表時，他其實不只是提供了一張表格，而是建立了一種「用數字看世界」的方式。

對學生而言，這一章的重點不在於死記公式，而在於理解：為什麼人類需要這些數值衡量？它們幫助我們化解不確定性，讓決策更有依據。換句話說，當我們計算平均數與標準差時，我們正在延續法爾的工作，學習用數字去理解人類的生命與風險。

統計學 (一)

Anderson's Statistics for Business & Economics (14/E)

Chapter 4: Introduction to Probability

上課時間地點: 二 D56, 商館 260206
授課教師: 吳漢銘 (國立政治大學統計學系副教授)
教學網站: https://hmwu.idv.tw

Overview

1. Example Managers often base their decisions on an analysis of _____ such as the following:
   (a) What are the _____ that the sales will decrease if we increase prices?
   (b) What is the _____ a new assembly method will increase productivity?
   (c) How _____ is it that the project will be finished on time?
   (d) What is the _____ that a new investment will be profitable?
2. Probability:
   (a) Probability is a numerical measure of the _____ that an event will occur.
   (b) Probability values are always assigned on a scale from _____.
   (c) A probability near zero indicates an event is _____ to occur.
   (d) A probability near one indicates an event is _____ to occur.

4.1 Random Experiments, Counting Rules, and Assigning Probabilities

1. **Random experiments**: a random experiment is a process that generates _____

experimental outcomes. On any single repetition or trial, the _____ that occurs is determined completely _____.

2. **Sample space** S: the sample space for a random experiment is the set of _____ experimental outcomes.

3. An experimental outcome is also called a _____ to identify it as an element of the sample space.

4. Example Tossing a Coin

   (a) Tossing the coin the upward face will be either a _____ or a _____. There are two possible experimental outcomes: head or tail (sample points).

   (b) Each time we toss the coin we will either observe a head or a tail. And, the outcome that occurs on any trial is determined solely _____ or _____ variability.

   (c) By specifying all the possible experimental outcomes, we identify the sample space _____ or _____ for a random experiment.

5. Example Rolling a Die

   (a) The experimental outcomes, defined as the number of dots appearing on the face of the die, are the six sample points in the sample space _____ for this random experiment.

Counting Rules, Combinations, and Permutations

1. Being able to identify and _____ the experimental outcomes is a necessary step in assigning probabilities.

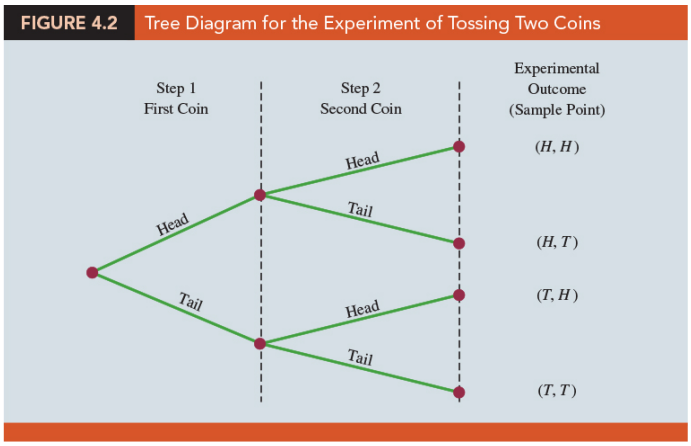
2. **Counting Rule for Multiple-Step Experiments**: If an experiment can be described as a sequence of _____ steps with n_1 possible outcomes on the first step, n_2 possible outcomes on the second step, and so on, then the total number of experimental outcomes is given by _____.

3. **Tree Diagram**: a graphical representation that helps in visualizing a multiple-step experiment.

4. Example Tossing two Coins

   (a) The experiment of tossing two coins can be thought of as a two-step experiment in which step 1 is the tossing of the first coin _____ and step 2 is the tossing of the second coin _____.

   (b) The sample space for this coin-tossing experiment _____, 4 distinct experimental outcomes are possible.



Question (p182)

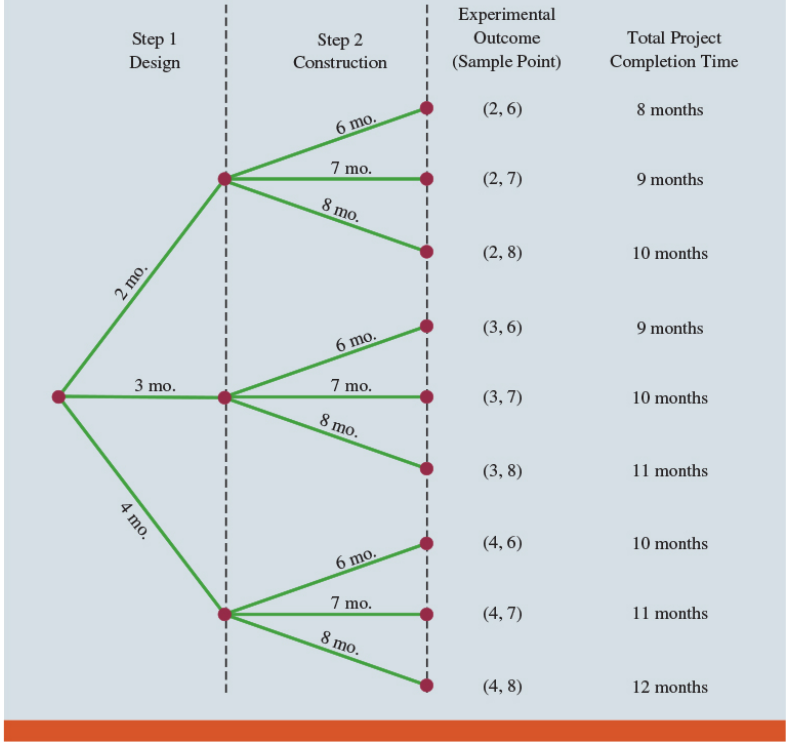
The Kentucky Power & Light Company (KP&L) is starting a project designed to increase the generating capacity of one of its plants in northern Kentucky. The project is divided into two sequential stages or steps: stage 1 (design) and stage 2 (construction). An analysis of similar construction projects revealed possible completion times for the design stage of 2, 3, or 4 months and possible completion times for the construction stage of 6, 7, or 8 months. Management sets a goal of 10 months for the completion of the entire project. Summarizes the experimental outcomes to complete the entire project for the KP&L problem. Draw the tree diagram to show how the outcomes (sample points) occur.

sol:

TABLE 4.1 Experimental Outcomes (Sample Points) for the KP&L Project

Completion Time (months)		Notation for Experimental Outcome	Total Project Completion Time (months)
Stage 1 Design	Stage 2 Construction		
2	6	(2, 6)	8
2	7	(2, 7)	9
2	8	(2, 8)	10
3	6	(3, 6)	9
3	7	(3, 7)	10
3	8	(3, 8)	11
4	6	(4, 6)	10
4	7	(4, 7)	11
4	8	(4, 8)	12

FIGURE 4.3 Tree Diagram for the KP&L Project



5. **Counting Rule for Combinations**: count the number of experimental outcomes when the experiment involves selecting _____ objects from a set of _____ objects.

6. The number of combinations of N objects taken n at a time is

   _____,

   where $N! = N(N-1)(N-2)\cdots(2)(1)$, $n! = n(n-1)(n-2)\cdots(2)(1)$, and $0! = 1$.

7. Example Consider a quality control procedure in which an inspector randomly selects two ($n = 2$) of five parts ($N = 5$) to test for defects. In a group of five parts, there are 10 combinations of two parts can be selected:

   $$C_2^5 = \binom{5}{2} = \frac{5!}{2!(5-2)!} = 10.$$

   If we label the five parts as A, B, C, D, and E, the 10 combinations or experimental outcomes can be identified as _____.

8. **Counting Rule for Permutations**: the number of experimental outcomes when n objects are to be selected from a set of N objects where the _____ of selection is important. The same n objects selected in a different order are considered a different experimental outcome.

9. The number of permutations of N objects taken n at a time is given by

   _____

10. Example Consider again the quality control process in which an inspector selects two ($n = 2$) of five parts ($N = 5$) to inspect for defects. There are 20 permutations may be selected:

    $$P_2^5 = \frac{5!}{(5-2)!} = \frac{120}{6} = 20.$$

    If we label the parts A, B, C, D, and E, the 20 permutations are AB, BA, AC, CA, AD, DA, AE, EA, BC, CB, BD, DB, BE, EB, CD, DC, CE, EC, DE, and ED.

Assigning Probabilities

1. **Basic Requirements for Assigning Probabilities**: the probability assigned to each experimental outcome must be between 0 and 1, inclusively.

(a) Let E_i is the ith experimental outcome and $P(E_i)$ is its probability

_____, for all i

(b) The sum of the probabilities for all experimental outcomes must equal 1.

_____

where n is the number of experimental outcomes.

2. **Classical Method**: assigning probabilities based on the assumption of _____ outcomes.

3. Example Consider the experiment of tossing a fair coin; the two experimental outcomes—head and tail—are equally likely. Because one of the two equally likely outcomes is a head, the probability of observing a head is _____. Similarly, the probability of observing a tail is also _____.

4. **Relative Frequency Method**: assigning probabilities based on _____ or historical data. That is when data are available to estimate the proportion of the time the experimental outcome will occur if the experiment is _____.

5. Example Consider a study of waiting times in the X-ray department for a local hospital. A clerk recorded the number of patients waiting for service at 9:00 a.m. on 20 successive days and obtained the following results.

```
Number Waiting: 0, 1, 2, 3, 4
Number of Days Outcome Occured: 2, 5, 6, 4, 3
```

These data show that on 2 of the 20 days, zero patients were waiting for service; on 5 of the days, one patient was waiting for service; and so on. Using the relative frequency method, we would assign a probability of $2/20 = 0.10$ ($5/20 = 0.25$) ($6/20 = 0.30$) ($4/20 = 0.20$) ($3/20 = 0.15$) to the experimental outcome of zero (one) (two) (three) (four) patients waiting for service. As with the classical method, using the relative frequency method automatically satisfies the two basic requirements.

6. **Subjective Method**: assigning probability based on judgment. That is, a probability value that expresses our _____ (on a scale from 0 to 1) that the experimental outcome will occur is specified.

7. Example Consider the case in which Tom and Judy make an offer to purchase a house. Two outcomes are possible:

   ```
   E_1 = their offer is accepted
   E_2 = their offer is rejected
   ```

   Judy believes that the probability their offer will be accepted is 0.8; thus, Judy would set $P(E_1) = 0.8$ and $P(E_2) = 0.2$. Tom, however, believes that the probability that their offer will be accepted is 0.6; hence, Tom would set $P(E_1) = 0.6$ and $P(E_2) = 0.4$. Note that Tom′ s probability estimate for E_1 reflects a greater pessimism that their offer will be accepted. Both Judy and Tom assigned probabilities that satisfy the two basic requirements. The fact that their probability estimates are different emphasizes the personal nature of the subjective method.

8. The best probability estimates often are obtained by combining the estimates from the _____ or _____ approach with the _____ estimate.

✎ Question ..... (p113)

To perform further analysis on the KP&L project, we must develop probabilities for each of the nine experimental outcomes listed in Table 4.1. On the basis of experience and judgment, management concluded that the experimental outcomes were not equally likely. Hence, the classical method of assigning probabilities could not be used. Management then decided to conduct a study of the completion times for similar projects undertaken by KP&L over the past three years. The results of a study of 40 similar projects are summarized in Table 4.2. After reviewing the results of the study, management decided to employ the relative frequency method of assigning probabilities. Management could have provided subjective probability estimates but felt that the current project was quite similar to the 40 previous projects. Thus, the relative frequency method was judged best.

TABLE 4.2 Completion Results for 40 KP&L Projects

Completion Time (months)			
Stage 1 Design	Stage 2 Construction	Sample Point	Number of Past Projects Having These Completion Times
2	6	(2, 6)	6
2	7	(2, 7)	6
2	8	(2, 8)	2
3	6	(3, 6)	4
3	7	(3, 7)	8
3	8	(3, 8)	2
4	6	(4, 6)	2
4	7	(4, 7)	4
4	8	(4, 8)	6
		Total	40

In using the data in Table 4.2 to compute probabilities for each outcome.

sol:

TABLE 4.3 Probability Assignments for the KP&L Project Based on the Relative Frequency Method

Sample Point	Project Completion Time	Probability of Sample Point
(2, 6)	8 months	$P(2, 6) = 6/40 = .15$
(2, 7)	9 months	$P(2, 7) = 6/40 = .15$
(2, 8)	10 months	$P(2, 8) = 2/40 = .05$
(3, 6)	9 months	$P(3, 6) = 4/40 = .10$
(3, 7)	10 months	$P(3, 7) = 8/40 = .20$
(3, 8)	11 months	$P(3, 8) = 2/40 = .05$
(4, 6)	10 months	$P(4, 6) = 2/40 = .05$
(4, 7)	11 months	$P(4, 7) = 4/40 = .10$
(4, 8)	12 months	$P(4, 8) = 6/40 = .15$
		Total 1.00

4.2 Events and Their Probabilities

1. **Event**: an event is a _____ of sample points.

2. **Probability of an Event**: the probability of any event is equal to the sum of the probabilities of the sample points in the event.

3. If we can identify all the sample points of an experiment and assign a probability to each, we can compute the _____.

4. Example KP&L project

   (a) Assume that the project manager is interested in the event that the entire project can be completed in 10 months or less. Referring to Table 4.3, we see that six sample points - (2, 6), (2, 7), (2, 8), (3, 6), (3, 7), and (4, 6) - provide a project completion time of 10 months or less.

   (b) Let C denote the event that the project is completed in 10 months or less:

   $$C = _____$$

   Event C is said to occur if any one of these six sample points appears as the experimental outcome.

   (c) Other events that might be of interest:

   L = The event that the project is completed in less than 10 months

   M = The event that the project is completed in more than 10 months

   Using the information in Table 4.3, we see that these events consist of the following sample points.

   $$L = _____$$
   $$M = _____$$

✎ Question ..... (p190)

In the KP&L project problem, compute the probability that the project will take 10 months or less to complete. Compute the probability of the event that the project is completed in more than 10 months.

sol: Using these probability results, we can now tell KP&L management that:

The probability that the project will be completed in 10 months or less:

_____

The probability that the project will be completed in less than 10 months:

_____

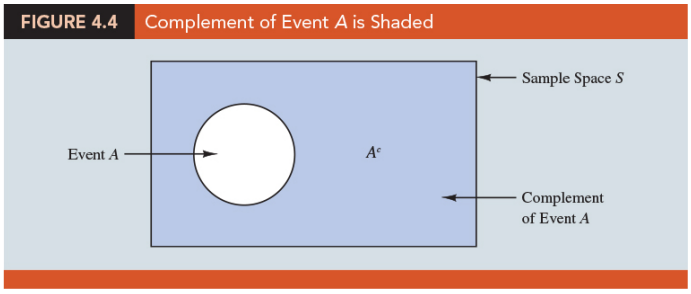
The probability that the project will be completed in more than 10 months:

_____

4.3 Some Basic Relationships of Probability

Complement of an Event

1. The _____ of event A is defined to be the event consisting of all sample points that are _____.
2. The complement of A is denoted by _____.
3. Computing probability using the complement: _____.
4. (Figure 4.4): Venn diagram illustrates the concept of a complement.



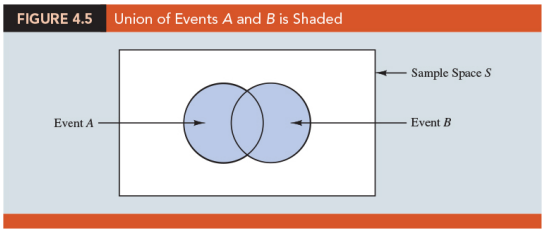
Addition Law

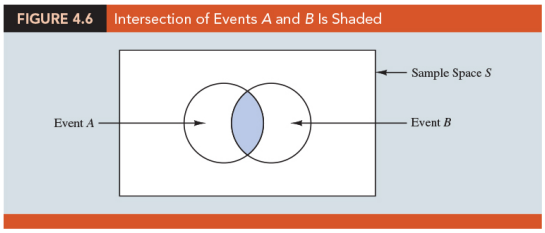
1. **Union**: given two events A and B, the union of A and B (denoted by _____) is the event containing all sample points belonging _____.

2. **Intersection**: The intersection of events A and B (denoted by _____) is the set of all sample points that are in _____ A and B.

3. **Addition Law**: a way to compute the probability of event A, or B, or both A and B occurring:

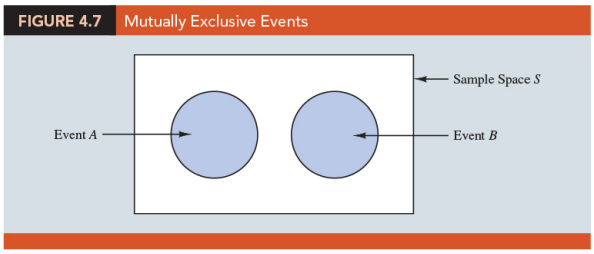
_____





4. Two events are said to be _____ if the events have no sample points in common. Two events are mutually exclusive if, when one event occurs, the other cannot occur.

5. **Addition Law for Mutually Exclusive Events**: _____



Question ..... (p195)

Consider the case of a small assembly plant with 50 employees. Each worker is expected to complete work assignments on time and in such a way that the assembled product will pass a final inspection. On occasion, some of the workers fail to meet the performance standards by completing work late or assembling a defective product. At the end of a performance evaluation period, the production manager

found that 5 of the 50 workers completed work late, 6 of the 50 workers assembled a defective product, and 2 of the 50 workers both completed work late and assembled a defective product. Let

```
L = the event that the work is completed late
D = the event that the assembled product is defective
```

After reviewing the performance data, the production manager decided to assign a poor performance rating to any employee whose work was either late or defective; thus the event of interest is $L \cup D$. What is the probability that the production manager assigned an employee a poor performance rating?

sol:

The relative frequency information:

_____.

To compute $P(L \cup D)$:

$$\begin{aligned} P(L \cup D) &= \underline{\qquad\qquad\qquad\qquad} \\ &= \underline{\qquad\qquad\qquad\qquad} \end{aligned}$$

There is a 0.18 probability that a randomly selected employee received a poor performance rating.

✎ Question ..... (p196)

Consider a recent study conducted by the personnel manager of a major computer software company. The study showed that 30% of the employees who left the firm within two years did so primarily because they were dissatisfied with their salary, 20% left because they were dissatisfied with their work assignments, and 12% of the former employees indicated dissatisfaction with both their salary and their work assignments. What is the probability that an employee who leaves within two years does so because of dissatisfaction with salary, dissatisfaction with the work assignment, or both?

sol:

Let

S = _____

W = _____

We have $P(S) = 0.30$, $P(W) = 0.20$, and $P(S \cap W) = 0.12$.

By the addition law, we have

_____.

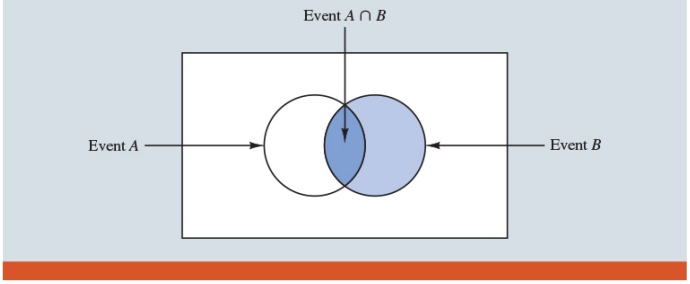
We find a 0.38 probability that an employee leaves for salary or work assignment reasons.

4.4 Conditional Probability

1. **Conditional probability**: the probability of an event A given that another event B has occurred is called a _____ (denoted by _____):

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

FIGURE 4.8 Conditional Probability $P(A \mid B) = P(A \cap B)/P(B)$



2. **Joint probability**: The probability of the intersection of two events is called _____.

3. **Marginal probabilities**: The values in the _____ of the _____ provide the probabilities of each event separately. The marginal probabilities are found by summing the joint probabilities in the corresponding row or column of the joint probability table.

4. Example Promotion status of male and female officers

   (a) Consider the situation of the promotion status of male and female officers of a major metropolitan police force in the eastern United States. The police force consists of 1200 officers, 960 men and 240 women. Over the past two years, 324 officers on the police force received promotions. The specific breakdown of promotions for male and female officers is shown in Table 4.4.

TABLE 4.4 Promotion Status of Police Officers Over the Past Two Years

	Men	Women	Total
Promoted	288	36	324
Not Promoted	672	204	876
Total	960	240	1200

   (b) After reviewing the promotion record, a committee of female officers raised a discrimination case on the basis that 288 male officers had received promotions, but only 36 female officers had received promotions. The police administration argued that the relatively low number of promotions for female officers was due not to discrimination, but to the fact that relatively few females are members of the police force.

(c) Let us show how conditional probability could be used to analyze the discrimination charge. Let

M = event an officer is a man, W = event an officer is a woman

A = event an officer is promoted, A^c = event an officer is not promoted

Obtain the following joint probability table:

TABLE 4.5 Joint Probability Table for Promotions

Joint probabilities appear in the body of the table.

	Men (*M*)	Women (*W*)	Total
Promoted (*A*)	.24	.03	.27
Not Promoted (A^c)	.56	.17	.73
Total	.80	.20	1.00

Marginal probabilities appear in the margins of the table.

- probability that a randomly selected officer is a man and is promoted: _____.
- probability that a randomly selected officer is a man and is not promoted: _____.
- probability that a randomly selected officer is a woman and is promoted: _____.
- probability that a randomly selected officer is a woman and is not promoted: _____.

(d) The marginal probabilities: _____, and _____. We see that 80% of the force is male, 20% of the force is female, 27% of all officers received promotions, and 73% were not promoted.

(e) The conditional probability of being promoted given that the officer is a mam, $P(A|M)$:

_____.

(f) Similarly, the probability that an officer is promoted given that the officer is a woman; that is, $P(A|W)$:

_____.

(g) The marginal probability in row 1 of Table 4.5 shows that the probability of promotion of an officer is $P(A) = 0.27$ (regardless of whether that officer is male or female). However, the critical issue in the discrimination case involves the two conditional probabilities $P(A|M)$ and $P(A|W)$. That is, what is the probability of a promotion given that the officer is a man, and what is the probability of a promotion given that the officer is a woman? If these two probabilities are equal, a discrimination argument has no basis because the chances of a promotion are the same for male and female officers. However, a difference in the two conditional probabilities will support the position that male and female officers are treated differently in promotion decisions.

(h) What conclusion do you draw?

i. The probability of a promotion given that the officer is a man is 0.30, twice the 0.15 probability of a promotion given that the officer is a woman.

ii. Although the use of conditional probability does not in itself prove that discrimination exists in this case, the conditional probability values support the argument presented by the female officers.

Independent Events

1. Two events A and B are _____ if

   _____ or _____

   Otherwise, the events are _____.

2. Example Promotion status of male and female officers

   (a) $P(A) = 0.27$, $P(A|M) = 0.30$, and $P(A|W) = 0.15$. We see that the probability of a promotion (event A) is affected or influenced by whether the officer is a man or a woman.

   (b) Particularly, because $P(A|M) \neq P(A)$, we would say that events A and M are dependent events. That is, the probability of event A (promotion) is altered or affected by knowing that event M (the officer is a man) exists.

   (c) Similarly, with $P(A|W) \neq P(A)$, we would say that events A and W are dependent events.

Multiplication Law

1. **Multiplication law**: a way to compute the probability of the intersection of two events:

   _____ and _____

2. Example Telecommunications company

   (a) Consider a telecommunications company that offers services such as high-speed Internet, cable television, and telephone services. For a particular city, it is known that 84% of the households subscribe to high-speed Internet service. If we let H denote the event that a household subscribes to high-speed Internet service, $P(H) = 0.84$. In addition, it is known that the probability that a household that already subscribes to high-speed Internet service also subscribes to cable television service (event C) is 0.75; that is, $P(C|H) = 0.75$. What is the probability that a household subscribes to both high-speed Internet and cable television services?

   (b) Using the multiplication law, we compute the desired $P(C \cap H)$ as

   $$P(C \cap H) = _____$$

   We now know that 63% of the households subscribe to both high-speed Internet and cable television services.

3. **Multiplication law for two independent evens**: events A and B are independent whenever $P(A|B) = P(A)$ or $P(B|A) = P(B)$, the multiplication law is

   _____

4. If $P(A \cap B) = P(A)P(B)$, then A and B are independent; if _____, then A and B are dependent.

5. Example Service station manager

   (a) Consider the situation of a service station manager who knows from past experience that 80% of the customers use a credit card when they purchase gasoline. What is the probability that the next two customers purchasing gasoline will each use a credit card?

(b) Let

$$A = \text{the event that the first customer uses a credit card}$$
$$B = \text{the event that the second customer uses a credit card}$$

then the event of interest is $A \cap B$. Assume that A and B are independent events. Thus,

_____

4.5 Bayes' Theorem

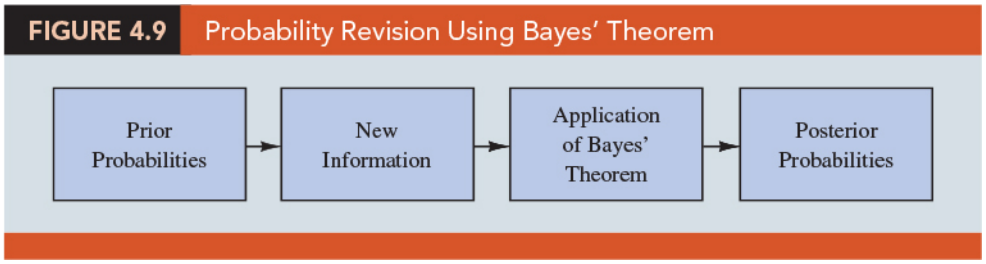
1. Motivation:

   (a) Often we begin probability analysis with _____ or _____ probabilities.

   (b) Then, from a sample, special report, or a product test, we obtain some additional information.

   (c) Given this information, we calculate _____ or _____ probabilities.

   (d) Bayes' theorem provides the means for _____ the prior probabilities.



2. The Bayes' theorem for the case of two events:

$$P(A_1|B) = \frac{\qquad\qquad}{\qquad\qquad}.$$

$$P(A_2|B) = \frac{\qquad\qquad}{\qquad\qquad}.$$

3. Example Manufacturing firm receives shipments from two suppliers

   (a) Consider a manufacturing firm that receives shipments of parts from two different suppliers. Let A_1 (A_2) denote the event that a part is from supplier 1 (2).

   (b) Currently, 65% (35%) of the parts purchased by the company are from supplier 1 (2). Hence, if a part is selected at random, we would assign the prior probabilities _____ and _____.

   (c) The quality of the purchased parts varies with the source of supply. Historical data suggest that the quality ratings of the two suppliers are as shown in Table 4.6.

TABLE 4.6 Historical Quality Levels of Two Suppliers

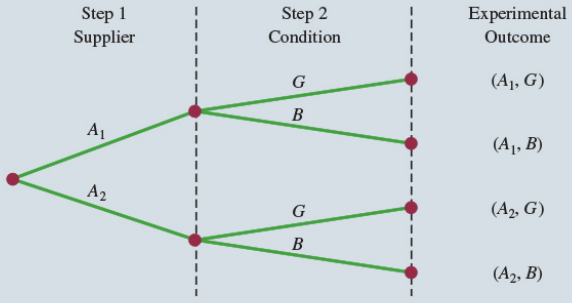
	Percentage Good Parts	Percentage Bad Parts
Supplier 1	98	2
Supplier 2	95	5

   (d) Let G (B) denote the event that a part is good (bad), the information in Table 4.6 provides the following conditional probability values.

$$P(G|A_1) = _____, \quad P(B|A_1) = _____$$
$$P(G|A_2) = _____, \quad P(B|A_2) = _____$$

   (e) (Figure 4.10) The tree diagram depicts the process of the firm receiving a part from one of the two suppliers and then discovering that the part is good or bad as a two-step experiment. We see that four experimental outcomes are possible; two correspond to the part being good and two correspond to the part being bad.

FIGURE 4.10 Tree Diagram for Two-Supplier Example



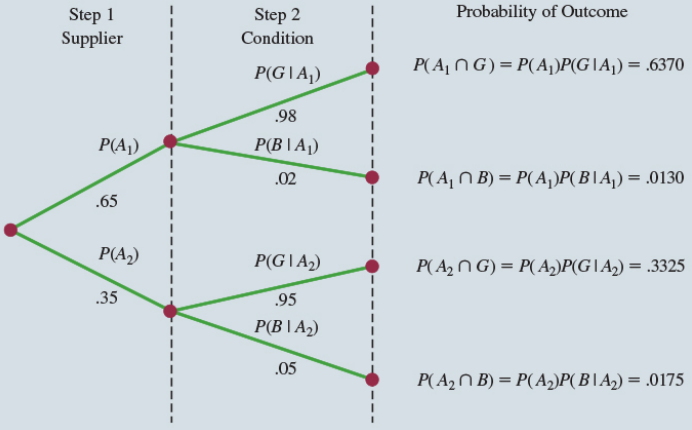
Note: Step 1 shows that the part comes from one of two suppliers, and step 2 shows whether the part is good or bad.

(f) Each of the experimental outcomes is the intersection of two events, so we can use the multiplication rule to compute the probabilities. For instance,

---

(g) **Probability tree** (Figure 4.11): From left to right through the tree, the probabilities for each branch at step 1 are prior probabilities and the probabilities for each branch at step 2 are conditional probabilities.

FIGURE 4.11 Probability Tree for Two-Supplier Example



(h) To find the probabilities of each experimental outcome, we simply multiply the probabilities on the branches leading to the outcome. Each of these joint probabilities is shown in Figure 4.11 along with the known probabilities for each branch.

Question (p209)

Suppose now that the parts from the two suppliers are used in the firm's manufacturing process and that a machine breaks down because it attempts to process a bad part. Given the information that the part is bad, what is the probability that it came from supplier 1 and what is the probability that it came from supplier 2?

sol:

- Letting B denote the event that the part is bad, we are looking for the posterior probabilities $P(A_1|B)$ and $P(A_2|B)$. We know that

$$P(A_1|B) = \frac{P(A_1 \cap B)}{P(B)} \quad \Rightarrow \quad \underline{\qquad\qquad\qquad\qquad}$$

- Find $P(B)$:

$$P(B) = \underline{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad}$$

- We have

$$\begin{aligned} P(A_1|B) &= \\ &= \frac{(0.65)(0.02)}{(0.65)(0.02) + (0.35)(0.05)} = \frac{0.0130}{0.0130 + 0.0175} \\ &= \frac{0.0130}{0.0305} = .4262 \\ P(A_2|B) &= \frac{(0.35)(0.05)}{(0.65)(0.02) + (0.35)(0.05)} \\ &= \frac{0.0175}{0.0130 + 0.0175} = \frac{0.0175}{0.0305} = 0.5738 \end{aligned}$$

- Note that in this application we started with a probability of _____ that a part selected at random was from supplier 1. However, given information that the part is bad, the probability that the part is from supplier 1 drops to _____. In fact, if the part is bad, it has better than a 50–50 chance that it came from supplier 2; that is, _____.

3. Bayes' theorem is applicable when the events for which we want to compute posterior probabilities are mutually exclusive and their union is the entire sample space. For the case of n mutually exclusive events _____, whose union is the entire sample space, Bayes' theorem can be used to compute any posterior probability $P(A_i|B)$:

$$P(A_i|B) = \frac{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad}{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad}$$

Tabular Approach

1. A tabular approach is helpful in conducting the Bayes' theorem calculations.

TABLE 4.7 Tabular Approach to Bayes' Theorem Calculations for the Two-Supplier Problem

(1) Events A_i	(2) Prior Probabilities $P(A_i)$	(3) Conditional Probabilities $P(B \mid A_i)$	(4) Joint Probabilities $P(A_i \cap B)$	(5) Posterior Probabilities $P(A_i \mid B)$
A_1	.65	.02	.0130	.0130/.0305 = .4262
A_2	.35	.05	.0175	.0175/.0305 = .5738
	1.00		$P(B)$ = .0305	1.0000

Step 1. Prepare the following three columns:

Column 1 : The _____ events A_i for which posterior probabilities are desired.

Column 2 : The _____ probabilities $P(A_i)$ for the events.

Column 3 : The _____ probabilities $P(B|A_i)$ of the new information B given each event.

Step 2. In column 4, compute the joint probabilities $P(A_i \cap B)$ for each event and the new information B by using the multiplication law: _____.

Step 3. Sum the joint probabilities in column 4. The sum is the probability of the new information, $P(B)$.

Example Thus we see in Table 4.7 that there is a 0.0130 probability that the part came from supplier 1 and is bad and a 0.0175 probability that the part came from supplier 2 and is bad. Because these are the only two ways in which a bad part can be obtained, the sum 0.0130 + 0.0175 shows an overall probability of 0.0305 of finding a bad part from the combined shipments of the two suppliers.

Step 4. In column 5, compute the posterior probabilities using the basic relationship of conditional probability.

$$P(A_i|B) = \frac{\qquad\qquad}{\qquad\qquad}$$

Note that the joint probabilities $P(A_i \cap B)$ are in column 4 and the probability $P(B)$ is the sum of column 4.

😍 EXERCISES

4.1 : 4, 6, 11, 13

4.2 : 16, 17, 19

4.3 : 23, 26, 29

4.4 : 31, 32, 35, 36

4.5 : 40, 42, 45

SUP : 49, 51, 56, 59

名言佳句

“你必須找到你的熱情所在”
“You’ve got to find what you love”

— *Steve Jobs (February 24, 1955 –October 5, 2011)*

統計小故事：帕斯卡與費馬的賭博難題

十七世紀的歐洲，是理性思潮逐漸興起的時代。雖然數學已經在幾何與代數領域取得重要進展，但「機率」這一概念仍停留在直覺層次。人們在日常生活中常接觸骰子、紙牌與賭博遊戲，卻缺乏一套嚴謹的數學框架去理解「隨機事件」的規律。而正是一次關於賭博的爭論，催生了現代機率論的起點。

故事的主角是一位法國貴族——德梅雷（Chevalier de Méré）。他酷愛賭博，經常在貴族沙龍中與人玩骰子遊戲。作為一位頗有數學興趣的人，他觀察到有些下注方式似乎「不划算」，卻無法精確解釋其中原因。某次，他與朋友們在玩「多次擲骰子」的遊戲時，提出了一個困惑已久的問題：如果一場遊戲尚未結束就被迫中斷，那麼賭金該如何公平分配？

這個問題聽似瑣碎，但其背後卻觸及到「公平」與「機率」的核心。過去人們往往以「目前的得分」來分配賭金，但這樣常導致爭議，因為它忽略了「未來的勝算」。於是，德梅雷把這個難題請教給數學家布萊士．帕斯卡（Blaise Pascal）。

帕斯卡並沒有獨自解答，而是與另一位數學巨擘皮埃爾．德．費馬（Pierre de Fermat）展開了通信。這場通信被譽為機率論的誕生契機。在往返的書信中，兩人逐步建立了一種全新的思考方式：與其爭論該如何分配，不如計算若遊戲繼續下去，雙方各自的「勝率」是多少。

舉例來說，如果遊戲規則是「先贏四局者獲勝」，而此刻甲已贏三局、乙贏兩局，那麼遊戲若繼續進行，甲贏的可能性遠大於乙。帕斯卡與費馬提出，應該按照這些「未來勝率」來分配賭金。這個概念，就是期望值（expected value）的雛形。

帕斯卡與費馬的討論不僅解決了德梅雷的困惑，更帶來一場觀念上的革命：隨機事件並非不可捉摸，而是可以透過數學計算來理解。這是人類第一次將「機率」從直覺提升為科學。

他們的成果迅速在歐洲知識界引發迴響。之後，惠更斯（Christiaan Huygens）在 1657 年出版了第一本機率論專著《De Ratiociniis in Ludo Aleae》（賭博遊戲中的推理），進一步系統化這些思想。從此，機率成為數學的一個獨立分支。

這一突破影響深遠。保險業開始利用機率計算死亡風險，設計合理的保費；天文學家運用機率推估觀測誤差；商人與投資者則利用機率來衡量風險與收益。甚至在現代社會，從醫學臨床試驗到人工智慧演算法，都能追溯到這場十七世紀的「賭博難題」。

對學生而言，這段歷史故事傳達了一個重要訊息：機率論的起點並不是抽象的數學，而是人類生活中的公平與決策需求。帕斯卡與費馬的通信讓我們看到，當數字與

邏輯介入，人類能夠從「感覺」轉向「計算」，從「爭論」走向「科學」。

在教材呼應上，本章的重點是機率的基本概念與規則，包括加法規則、乘法規則與條件機率。這些看似枯燥的規則，其實正是帕斯卡與費馬為了解決現實問題而建立的數學語言。當學生學會計算「一枚硬幣連擲三次出現兩次正面」的機率時，他們其實正延續著這段數學史上的重要傳統。

因此，本章的學習不僅是掌握計算技巧，更是理解一種思維方式：面對不確定性時，我們不必依靠直覺或運氣，而可以依靠數學與邏輯來做出合理判斷。這就是機率的力量，也是統計學的根基。

統計學 (一)

Anderson's Statistics for Business & Economics (14/E)

Chapter 5: Discrete Probability Distributions

上課時間地點: 二 D56, 商館 260206
授課教師: 吳漢銘 (國立政治大學統計學系副教授)
教學網站: https://hmwu.idv.tw

5.1 Random Variables

1. A random variable is a _____ description of the _____ of an experiment.
2. A random variable can be classified as being either _____ or _____ depending on the numerical values it assumes.
3. A random variable, often denoted by X (or some other capital letter), is a _____ that maps _____ to _____ on the real line.

Discrete Random Variables

1. A discrete random variable assumes either a _____ of values or an _____ of values such as $0, 1, 2, \cdots$.

TABLE 5.1 Examples of Discrete Random Variables

Random Experiment	Random Variable (x)	Possible Values for the Random Variable
Flip a coin	Face of coin showing	1 if heads; 0 if tails
Roll a die	Number of dots showing on top of die	1, 2, 3, 4, 5, 6
Contact five customers	Number of customers who place an order	0, 1, 2, 3, 4, 5
Operate a health care clinic for one day	Number of patients who arrive	0, 1, 2, 3, . . .
Offer a customer the choice of two products	Product chosen by customer	0 if none; 1 if choose product A; 2 if choose product B

Continuous Random Variables

1. A continuous random variable assumes any _____ value in an _____ or collection of intervals.

2. Experimental outcomes based on measurement scales such as time, weight, distance, and temperature can be described by _____ random variables.

TABLE 5.2 Examples of Continuous Random Variables

Random Experiment	Random Variable (x)	Possible Values for the Random Variable
Customer visits a web page	Time customer spends on web page in minutes	$x \geq 0$
Fill a soft drink can (max capacity = 360 milliliters)	Number of milliliters	$0 \leq x \leq 360$
Test a new chemical process	Temperature when the desired reaction takes place (min temperature = 65°C; max temperature = 100°C)	$65 \leq x \leq 100$
Invest $10,000 in the stock market	Value of investment after one year	$x \geq 0$

5.2 Developing Discrete Probability Distributions

1. The **probability distribution** for a random variable describes how _____ are distributed over the _____ of the random variable.

2. For a discrete random variable X, a probability function, denoted by _____, provides the probability for each value of the random variable.

3. The _____, _____, and _____ methods of assigning probabilities can be used to develop discrete probability distributions.

4. *The classical method*:

   (a) The classical method of assigning probabilities to values of a random variable is applicable when the experimental outcomes generate values of the random variable that are _____.

   (b) Example The experiment of rolling a die.

      i. Each of th outcomes (the numbers 1, 2, 3, 4, 5, or 6) is equally likely.

      ii. Let _____.

      iii. (Table 5.3) the probability distribution function $f(x)$ of X.

TABLE 5.3 Probability Distribution for Number Obtained on One Roll of a Die

Number Obtained x	Probability of x $f(x)$
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

5. *The subjective method*:

   (a) Each probability is assigned by user's best _____. Different people can be expected to obtain different probability distributions.

6. *The relative frequency method*:

   (a) We treat the (large amounts of) data as if they were the _____ and use the _____ method to assign probabilities to the experimental outcomes.

   (b) The use of the relative frequency method to develop discrete probability distributions leads to what is called an _____ discrete distribution. (more widely used in practice)

Question (p229)

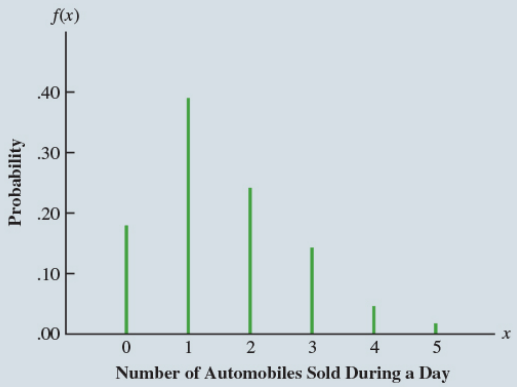
Over the past 300 days, DiCarlo Motors in Saratoga, New York, has experienced 54 (117, 72, 42, 12, 3) days with no (1, 2, 3, 4, 5) automobiles sold. Suppose we consider the experiment of observing a day of operations at DiCarlo Motors and define the random variable of interest as X = the number of automobiles sold during a day. (a) Use the relative frequency method to develop a probability distribution for the number of cars sold per day. (b) Graph the probability distribution such that the values of the random variable X for DiCarlo Motors are shown on the horizontal axis and the probability associated with these values is shown on the vertical axis.

sol:

TABLE 5.4 Probability Distribution for the Number of Automobiles Sold During a Day at Dicarlo Motors

x	$f(x)$
0	.18
1	.39
2	.24
3	.14
4	.04
5	.01
Total	1.00

FIGURE 5.1 Graphical Representation of the Probability Distribution for the Number of Automobiles Sold During a Day at Dicarlo Motors



7. A primary advantage of defining a random variable and its probability distribution is that once the probability distribution is known, it is relatively easy to determine the probability of a variety of _____ that may be of interest to a decision maker.

8. *Required Conditions for a Discrete Probability Function*:

   _____ and _____

9. **Discrete Uniform Probability Distribution**: _____ where n is the number of values the random variable may assume.

✎ Question ..... (p230)

For the experiment of rolling a fair die, we define the random variable X to be the number of dots on the upward face. Find the probability function for this discrete random variable.

sol:

5.3 Expected Value and Variance

Expected Value

1. The expected value, or _____, of a random variable X is a measure of the _____ location for the discrete random variable:

$$E(X) = _____ = _____.$$

2. The expected value is a _____ of the values of the random variable where the weights are the _____.

Variance

1. Use variance to summarize the _____ in the values of a discrete random variable:

$$Var(X) = _____ = _____$$

2. The variance is a weighted average of the _____ of a random variable from its mean. The weights are the probabilities.

Question ..... (p234)

Let the random variable X be the number of auto mobiles sold during a day in the DiCarlo Motors automobile sales example from Section 5.2, find the expected value and the variance of this discrete random variable.

sol:

TABLE 5.5 Calculation of the Expected Value for the Number of Automobiles Sold During a Day at Dicarlo Motors

x	f(x)	xf(x)
0	.18	0(.18) = .00
1	.39	1(.39) = .39
2	.24	2(.24) = .48
3	.14	3(.14) = .42
4	.04	4(.04) = .16
5	.01	5(.01) = .05
		1.50

$E(x) = \mu = \Sigma xf(x)$

TABLE 5.6 Calculation of the Variance for the Number of Automobiles Sold During a Day at Dicarlo Motors

x	$x-\mu$	$(x-\mu)^2$	$f(x)$	$(x-\mu)^2 f(x)$
0	$0-1.50=-1.50$	2.25	.18	$2.25(.18)=$ .4050
1	$1-1.50=-.50$	.25	.39	$.25(.39)=$ .0975
2	$2-1.50=.50$	.25	.24	$.25(.24)=$ .0600
3	$3-1.50=1.50$	2.25	.14	$2.25(.14)=$ .3150
4	$4-1.50=2.50$	6.25	.04	$6.25(.04)=$ .2500
5	$5-1.50=3.50$	12.25	.01	$12.25(.01)=$ .1225
				1.2500

$\sigma^2 = \Sigma(x-\mu)^2 f(x)$

5.4 Bivariate Distributions, Covariance, and Financial Portfolios

1. A probability distribution involving two random variables is called a _____ probability distribution.

2. **Bivariate experiment**: Each outcome for a bivariate experiment consists of _____, one for each random variable.

3. Example Consider the bivariate experiment of rolling a pair of dice. The outcome consists of two values, the number obtained with the first die and the number obtained with the second die.

4. Example Consider the experiment of observing the financial markets for a year and recording the percentage gain for a stock fund and a bond fund. Again, the experimental outcome provides a value for two random variables, the percent gain in the stock fund and the percent gain in the bond fund.

A Bivariate Empirical Discrete Probability Distribution

1. Example (Section 5.2) DiCarlo Motors automobile dealership in Saratoga, New York.

(a) (Table 5.7) DiCarlo has another dealership in Geneva, New York. Table 5.7 shows the number of cars sold at each of the dealerships over a 300-day period.

(b) The numbers in the bottom (total) row are the frequencies we used to develop an empirical probability distribution for daily sales at DiCarlo's Saratoga dealership. The numbers in the rightmost (total) column are the frequencies of daily sales for the Geneva dealership.

(c) Entries in the body of the table give the number of days the Geneva dealership had a level of sales indicated by the row, when the Saratoga dealership had the level of sales indicated by the column.

TABLE 5.7 Number of Automobiles Sold at DiCarlo's Saratoga and Geneva Dealerships Over 300 Days

	Saratoga Dealership						
Geneva Dealership	0	1	2	3	4	5	Total
0	21	30	24	9	2	0	86
1	21	36	33	18	2	1	111
2	9	42	9	12	3	2	77
3	3	9	6	3	5	0	26
Total	54	117	72	42	12	3	300

(d) Consider the bivariate experiment of observing a day of operations at DiCarlo Motors and recording the number of cars sold. Define

$$X = \text{number of cars sold at the Geneva dealership}$$

$$Y = \text{the number of cars sold at the Saratoga dealership}$$

(e) (Table 5.8) divide all of the frequencies in Table 5.7 by the number of observations (300) to develop a _____ probability distribution for automobile sales at the two DiCarlo dealerships.

TABLE 5.8 Bivariate Empirical Discrete Probability Distribution for Daily Sales at DiCarlo Dealerships in Saratoga and Geneva, New York

	Saratoga Dealership						
Geneva Dealership	0	1	2	3	4	5	Total
0	.0700	.1000	.0800	.0300	.0067	.0000	.2867
1	.0700	.1200	.1100	.0600	.0067	.0033	.3700
2	.0300	.1400	.0300	.0400	.0100	.0067	.2567
3	.0100	.0300	.0200	.0100	.0167	.0000	.0867
Total	.18	.39	.24	.14	.04	.01	1.0000

(f) The probabilities in the lower (right) margin provide the _____ distribution for the DiCarlo Motors Saratoga (Geneva) dealership.

(g) The probabilities in the body of the table provide the _____ probability distribution for sales at both dealerships.

2. Bivariate probabilities are often called _____.

3. Note that there is one bivariate probability for each experimental outcome.

4. With _____ possible values for X and _____ possible values for Y, there are _____ experimental outcomes and bivariate probabilities.

Question ..... (p240)

Refer to the DiCarlo Motors automobile dealership data, find the probability distribution for total sales at both DiCarlo dealerships and the expected value and variance of total sales.

sol:

TABLE 5.9 Calculation of the Expected Value and Variance for Total Daily Sales at DiCarlo Motors

s	$f(s)$	$sf(s)$	$s - E(s)$	$(s - E(s))^2$	$(s - E(s))^2 f(s)$
0	.0700	.0000	−2.6433	6.9872	.4891
1	.1700	.1700	−1.6433	2.7005	.4591
2	.2300	.4600	−.6433	.4139	.0952
3	.2900	.8700	.3567	.1272	.0369
4	.1267	.5067	1.3567	1.8405	.2331
5	.0667	.3333	2.3567	5.5539	.3703
6	.0233	.1400	3.3567	11.2672	.2629
7	.0233	.1633	4.3567	18.9805	.4429
8	.0000	.0000	5.3567	28.6939	.0000
		$E(s) = 2.6433$			$Var(s) = 2.3895$

5. The covariance between two random variables X and Y is:

$$\sigma_{xy} = Cov(X,Y) = _____ \text{ or}$$

$$\sigma_{xy} = Cov(X,Y) = _____.$$

6. The correlation between two random variables X and Y is the covariance divided by the product of the standard deviations for the two random variables:

$$\rho_{xy} = \underline{\qquad\qquad}.$$

7. *Expected Value of a Linear Combination of Random Variables X and Y*:

$$E(aX + bY) = \underline{\qquad\qquad\qquad\qquad}.$$

8. *Variance of a Linear Combination of Two Random Variables*:

$$Var(aX + bY) = \underline{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad}$$

where $Cov(X, Y) = \sigma_{xy}$ is the covariance of X and Y.

Question ..... (p241)

Refer to the DiCarlo Motors automobile dealership data, compute the covariance and the correlation coefficient between daily sales at the two DiCarlo dealerships.

sol:

TABLE 5.10 Calculation of the Expected Value and Variance of Daily Automobile Sales at DiCarlo Motors' Geneva Dealership

x	$f(x)$	$xf(x)$	$x - E(x)$	$[(x - E(x)]^2$	$[x - E(x)]^2 f(x)$
0	.2867	.0000	−1.1435	1.3076	.3749
1	.3700	.3700	−.1435	.0206	.0076
2	.2567	.5134	.8565	.7336	.1883
3	.0867	.2601	1.8565	3.447	.2988
		$E(x) = 1.1435$			$Var(x) = .8696$

Financial Applications

1. To see how what we have learned can be useful in constructing financial portfolios that provide a good balance of _____ and _____.
2. The _____ of percent return is often used as a measure of risk associated with an investment.

✎ Question ..... (p242)

(Table 5.11) A financial advisor is considering four possible economic scenarios for the coming year and has developed a probability distribution showing the percent return, X, for investing in a largecapstock fund and the percent return, Y, for investing in a long-term government bond fund given each of the scenarios. The bivariate probability distribution for X and Y is shown in Table 5.11.

TABLE 5.11 Probability Distribution of Percent Returns for Investing in a Large-Cap Stock Fund, x, and Investing in a Long-Term Government Bond Fund, y

Economic Scenario	Probability $f(x, y)$	Large-Cap Stock Fund (x)	Long-Term Government Bond Fund (y)
Recession	.10	−40	30
Weak Growth	.25	5	5
Stable Growth	.50	15	4
Strong Growth	.15	30	2

Compute the expected percent return for investing in the stock fund, $E(X)$, and the expected percent return for investing in the bond fund, $E(Y)$. Draw the conclusion.

sol:

✎ Question ..... (p242)

Refer to the financial advisor example. Compute the standard deviation of the percent returns for the stock and bond fund investments.

sol:

3. We have already seen that the stock fund offers a greater expected return, so if we want to choose between investing in either the _____ fund or the _____ fund it depends on our attitude toward _____.

4. An _____ investor might choose the stock fund because of the higher expected return; a _____ investor might choose the bond fund because of the lower risk. But, there are other options.

5. What about the possibility of investing in a portfolio consisting of both an investment in the stock fund and an investment in the bond fund?

✎ Question ..... (p242)

Refer to the financial advisor example. Suppose we would like to consider a portfolio by investing equal amounts in the largecap stock fund and in the longterm government bond fund. Evaluate this portfolio by computing its expected return and risk.

sol:

1.

2.

3.

Question (p242)

Refer to the financial advisor example. Compare the three investment alternatives according to their expected returns, variances, and standard deviations: investing solely in the stock fund, investing solely in the bond fund, or creating a portfolio by dividing our investment amount equally between the stock and bond funds. Which of these alternatives would you prefer?

TABLE 5.12 Expected Values, Variances, and Standard Deviations for Three Investment Alternatives

Investment Alternative	Expected Return (%)	Variance of Return	Standard Deviation of Return (%)
100% in Stock Fund	9.25	328.1875	18.1159
100% in Bond Fund	6.55	61.9475	7.8707
Portfolio (50% in stock fund, 50% in bond fund)	7.90	29.865	5.4650

sol:

1. The expected return is highest for investing 100% in the stock fund, but the risk is also highest. The standard deviation is 18.1159%.
2. Investing 100% in the bond fund has a lower expected return, but a significantly smaller risk.
3. Investing 50% in the stock fund and 50% in the bond fund (the portfolio) has an expected return that is halfway between that of the stock fund alone and the bond fund alone. But note that it has less risk than investing 100% in either of the individual funds. It has both a higher return and less risk (smaller standard deviation) than investing solely in the bond fund.
4. Whether you would choose to invest in the stock fund or the portfolio depends on your attitude toward risk.
5. The stock fund has a higher expected return. But the portfolio has significantly less risk and also provides a fairly good return. Many would choose it.
6. It is the negative covariance between the stock and bond funds that has caused the portfolio risk to be so much smaller than the risk of investing solely in either of the individual funds.

5.5 Binomial Probability Distribution

A Binomial Experiment

1. A binomial experiment exhibits the following four properties.
   (a) The experiment consists of a sequence of _____.
   (b) Two outcomes are possible on each trial. We refer to one outcome as a _____ and the other outcome as a _____.
   (c) The probability of a success, denoted by _____, does not change from trial to trial. Consequently, the probability of a failure, denoted by _____, does not change from trial to trial. (stationarity assumption)
   (d) The trials are _____.
2. If properties (b), (c), and (d) are present, we say the trials are generated by a _____ process.
3. (Figure 5.2) depicts one possible sequence of successes and failures for a binomial experiment involving eight trials.

FIGURE 5.2 One Possible Sequence of Successes and Failures for an Eight-Trial Binomial Experiment

Property 1: The experiment consists of $n = 8$ identical trials.

Property 2: Each trial results in either success (S) or failure (F).

Trials →	1	2	3	4	5	6	7	8
Outcomes →	S	F	F	S	S	F	S	S

4. In a binomial experiment, our interest is in the _____ occurring in the _____. If we let X denote the number of successes occurring in the n trials, we see that X can assume the values of _____.

5. Because the number of values is _____, X is a discrete random variable. The probability distribution associated with this random variable is called the _____ probability distribution. Denote by _____.

✎ Question ..... (p248)

Consider the experiment of tossing a coin five times and on each toss observing whether the coin lands with a head or a tail on its upward face. Suppose we want to count the number of heads appearing over the five tosses. Does this experiment show the properties of a binomial experiment? What is the random variable of interest?

sol:

Four properties of a binomial experiment are satisfied:

1. The experiment consists of five identical trials; each trial involves the tossing of one coin.
2. Two outcomes are possible for each trial: a head or a tail. We can designate head a success and tail a failure.
3. The probability of a head and the probability of a tail are the same for each trial, with $p = 0.5$ and $1 - p = 0.5$.
4. The trials or tosses are independent because the outcome on any one trial is not affected by what happens on other trials or tosses.

The random variable of interest is $X =$ the number of heads appearing in the five trials. In this case, X can assume the values of 0, 1, 2, 3, 4, or 5.

✎ Question ..... (p249)

Consider an insurance salesperson who visits 10 randomly selected families. The outcome associated with each visit is classified as a success if the family purchases an insurance policy and a failure if the family does not. From past experience, the salesperson knows the probability that a randomly selected family will purchase an insurance policy is 0.10. Checking the properties of a binomial experiment.

sol:

Four assumptions are satisfied, this example is a binomial experiment:

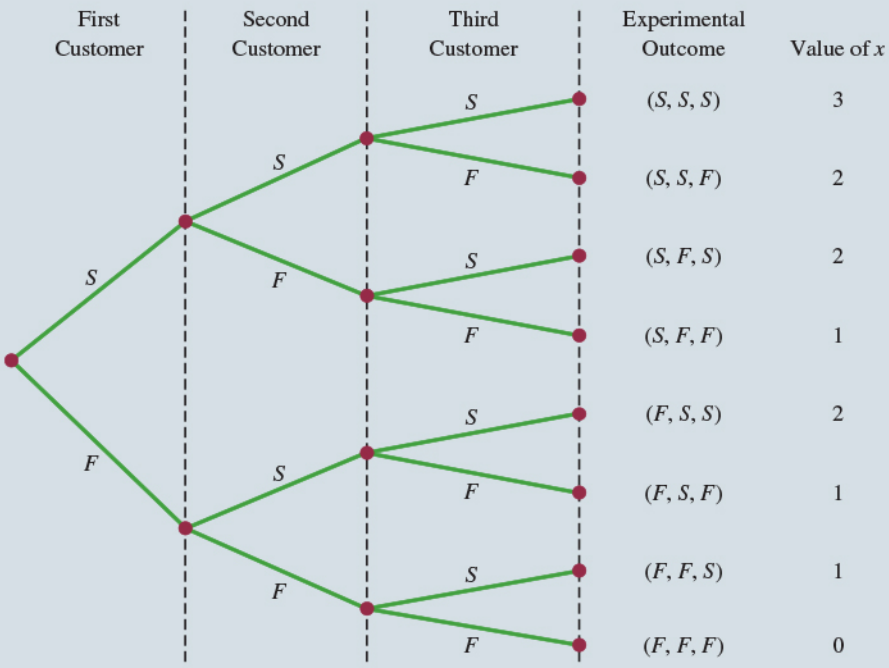
1. The experiment consists of 10 identical trials; each trial involves contacting one family.
2. Two outcomes are possible on each trial: the family purchases a policy (success) or the family does not purchase a policy (failure).
3. The probabilities of a purchase and a nonpurchase are assumed to be the same for each sales call, with $p = 0.10$ and $1-p = 0.90$.
4. The trials are independent because the families are randomly selected.

The random variable of interest is the number of sales obtained in contacting the 10 families. In this case, X can assume the values of $0, 1, \cdots, 10$.

Martin Clothing Store Problem

1. Let us consider the purchase decisions of the next three customers who enter the Martin Clothing Store. On the basis of past experience, the store manager estimates the probability that any one customer will make a purchase is 0.30. What is the probability that two of the next three customers will make a purchase?
2. (Figure 5.3) tree diagram shows that the experiment of observing the three customers each making a purchase decision has eight possible outcomes.

FIGURE 5.3 Tree Diagram for the Martin Clothing Store Problem



3. Using S to denote success (_____) and F to denote failure (_____), we are interested in experimental outcomes involving two successes in the three trials (purchase decisions).

4. Verify that the experiment involving the sequence of three purchase decisions can be viewed as a binomial experiment:

   (a) The experiment can be described as a sequence of three identical trials, one trial for each of the three customers who will enter the store.

   (b) Two outcomes - the customer makes a purchase (success) or the customer does not make a purchase (failure) - are possible for each trial.

   (c) The probability that the customer will make a purchase (0.30) or will not make a purchase (0.70) is assumed to be the same for all customers.

   (d) The purchase decision of each customer is independent of the decisions of the other customers.

5. The number of experimental outcomes resulting in exactly x successes in n trials can be computed using:

   _____,

   where $n! = n(n-1)(n-2)\cdots(2)(1)$ and $0! = 1$.

6. Because the trials of a binomial experiment are independent, we can simply _____ the probabilities associated with each trial outcome to find the probability of a particular sequence of successes and failures.

7. The probability of purchases by the first two customers and no purchase by the third customer, denoted (S, S, F), is given by _____. With a 0.30 probability of a purchase on any one trial, the probability of a purchase on the first two trials and no purchase on the third is given by _____.

8. In any binomial experiment, all sequences of trial outcomes yielding x successes in n trials have the same probability of occurrence. The probability of each sequence of trials yielding x successes in n trials is _____.

9. **The binomial probability function**

   $$f(x) = \text{_____},$$

   where

   $$\begin{aligned} x &= \text{the number of successes}, x = 0, 1, 2, \cdots, n \\ p &= \text{the probability of a success on one trial} \\ n &= \text{the number of trials} \\ f(x) &= \text{the probability of } x \text{ successes in } n \text{ trials} \end{aligned}$$

10. For the binomial probability distribution, X is a discrete random variable with the probability function _____ applicable for values of $x = 0, 1, 2, \cdots, n$. (_____)

Question ..... (p113)

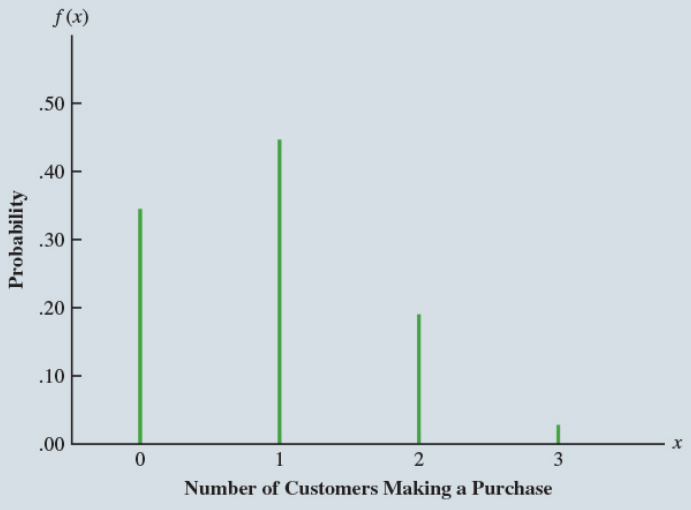
Refer to the Martin Clothing Store example, compute the probability that no customer makes a purchase, exactly one customer makes a purchase, exactly two customers make a purchase, and all three customers make a purchase.

sol: The calculations are summarized in Table 5.13, which gives the probability distribution of the number of customers making a purchase. Figure 5.4 is a graph of this probability distribution.

TABLE 5.13 Probability Distribution for the Number of Customers Making a Purchase

x	$f(x)$
0	$\frac{3!}{0!3!}(.30)^0(.70)^3 = .343$
1	$\frac{3!}{1!2!}(.30)^1(.70)^2 = .441$
2	$\frac{3!}{2!1!}(.30)^2(.70)^1 = .189$
3	$\frac{3!}{3!0!}(.30)^3(.70)^0 = .027$
	1.000

FIGURE 5.4 Graphical Representation of the Probability Distribution for the Number of Customers Making a Purchase



Using Tables of Binomial Probabilities

1. (Table 5.14) Table 5 of Appendix B provides a table of binomial probabilities. To use this table, we must specify the values of n, p, and x for the binomial experiment of interest.

TABLE 5.14 Selected Values from the Binomial Probability Table
Example: $n = 10$, $x = 3$, $P = .40$; $f(3) = .2150$

		p									
n	x	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50
9	0	.6302	.3874	.2316	.1342	.0751	.0404	.0207	.0101	.0046	.0020
	1	.2985	.3874	.3679	.3020	.2253	.1556	.1004	.0605	.0339	.0176
	2	.0629	.1722	.2597	.3020	.3003	.2668	.2162	.1612	.1110	.0703
	3	.0077	.0446	.1069	.1762	.2336	.2668	.2716	.2508	.2119	.1641
	4	.0006	.0074	.0283	.0661	.1168	.1715	.2194	.2508	.2600	.2461
	5	.0000	.0008	.0050	.0165	.0389	.0735	.1181	.1672	.2128	.2461
	6	.0000	.0001	.0006	.0028	.0087	.0210	.0424	.0743	.1160	.1641
	7	.0000	.0000	.0000	.0003	.0012	.0039	.0098	.0212	.0407	.0703
	8	.0000	.0000	.0000	.0000	.0001	.0004	.0013	.0035	.0083	.0176
	9	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0008	.0020
10	0	.5987	.3487	.1969	.1074	.0563	.0282	.0135	.0060	.0025	.0010
	1	.3151	.3874	.3474	.2684	.1877	.1211	.0725	.0403	.0207	.0098
	2	.0746	.1937	.2759	.3020	.2816	.2335	.1757	.1209	.0763	.0439
	3	.0105	.0574	.1298	.2013	.2503	.2668	.2522	**.2150**	.1665	.1172
	4	.0010	.0112	.0401	.0881	.1460	.2001	.2377	.2508	.2384	.2051
	5	.0001	.0015	.0085	.0264	.0584	.1029	.1536	.2007	.2340	.2461
	6	.0000	.0001	.0012	.0055	.0162	.0368	.0689	.1115	.1596	.2051
	7	.0000	.0000	.0001	.0008	.0031	.0090	.0212	.0425	.0746	.1172
	8	.0000	.0000	.0000	.0001	.0004	.0014	.0043	.0106	.0229	.0439
	9	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0016	.0042	.0098
	10	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0010

Expected Value and Variance for the Binomial Distribution

1. Expected Value and Variance for the Binomial Distribution:

$$E(X) = \mu = _____$$
$$Var(X) = \sigma^2 = _____$$

✎ Question ..... (p255)

Refer to the Martin Clothing Store problem with three customers, (a) compute the expected number, the variance and standard deviation of customers who will make a purchase. (b) Suppose that for the next month the Martin Clothing Store forecasts 1000 customers will enter the store. What is the expected number, the variance and standard deviation of customers who will make a purchase?

sol:

5.6 Poisson Probability Distribution

1. The probability distribution of _____ random variable is called a Poisson distribution. It is a discrete random variable that is often useful in estimating the _____ over a specified interval of time or space.

2. Example The random variable of interest might be the number of arrivals at a car wash in one hour, the number of repairs needed in 10 miles of highway, or the number of leaks in 100 miles of pipeline.

3. Two properties must be satisfied, the number of occurrences is a random variable described by the Poisson probability distribution:

   (a) The probability of an occurrence is the _____ for any two intervals of equal length.

(b) The occurrence or nonoccurrence in any interval is _____ of the occurrence or nonoccurrence in any other interval.

4. The **Poisson probability mass function** is defined by

$$f(x) = _____,$$

where

$$\begin{aligned} x &= \text{the number of occurrences in an interval}, x = 0, 1, 2, \cdots \\ f(x) &= \text{the probability of } X \text{ occurrences in an interval} \\ \mu &= \text{expected value or mean number of occurrences} \\ e &= 2.71828. \end{aligned}$$

5. For the Poisson probability distribution, X is a discrete random variable indicating the number of occurrences in the interval. (Denote _____ or _____)

6. Since there is no stated upper limit for the number of occurrences, the probability function $f(x)$ is applicable for values _____ without limit.

7. In practical applications, X will eventually become large enough so that $f(x)$ is approximately _____ and the probability of any larger values of X becomes _____.

An Example Involving Time Intervals

1. Suppose that we are interested in the number of patients who arrive at the emergency room of a large hospital during a 15-minute period on weekday mornings. If we can assume that the probability of a patient arriving is the same for any two time periods of equal length and that the arrival or nonarrival of a patient in any time period is independent of the arrival or nonarrival in any other time period, the Poisson probability function is applicable.

2. Suppose these assumptions are satisfied and an analysis of historical data shows that the average number of patients arriving in a 15-minute period of time is 10; in this case, the following probability function applies:

$$f(x) = _____$$

The random variable here is $X =$ _____ in any 15-minute period.

3. If management wanted to know the probability of exactly five arrivals in 15 minutes, we would set $X = 5$ and thus obtain

   Probability of exactly 5 arrivals in 15-minutes $= f(5) =$ _____

4. (Table 5.15) Table 7 of Appendix B provides probabilities for specific values of x and μ.

5. In the preceding example, the mean of the Poisson distribution is $\mu = 10$ arrivals per 15-minute period.

6. A property of the Poisson distribution is that the mean of the distribution and the variance of the distribution are _____. Thus, the variance for the number of arrivals during 15-minute periods is $\sigma^2 = 10$. The standard deviation is $\sigma = \sqrt{10} = 3.16$.

7. When computing a Poisson probability for a different time interval, we must first _____ the _____ to the time period of interest and then compute the probability.

✎ Question ..... (p260)

Suppose the number of patients who arrive at the emergency room of a large hospital during a 15minute period on weekday morning is the Poisson distribution with the average number of patients arriving in a 15-minute period of time is 10, compute the probability of one arrival in a 3-minute period.

sol:

An Example Involving Length or Distance Intervals

✎ Question ..... (p260)

Suppose we are concerned with the occurrence of major defects in a highway one month after resurfacing. We will assume that the probability of a defect is the same for any two highway intervals of equal length and that the occurrence or nonoccurrence of a defect in any one interval is independent of the occurrence or nonoccurrence of a defect in any other interval. Hence, the Poisson distribution can be applied. Suppose we learn that major defects one month after resurfacing occur at the average rate of two per mile. Find the probability of no major defects in a particular three mile section of the highway.

sol:

5.7 Hypergeometric Probability Distribution

1. The hypergeometric probability distribution is closely related to the _____ distribution.

2. The two probability distributions differ in two key ways. With the hypergeometric distribution, the trials are not _____; and the probability of success _____ from trial to trial.

3. Notation for the hypergeometric distribution, _____ denotes the number of elements in the population of size _____ labeled _____, and _____ denotes the number of elements in the population labeled _____.

4. The hypergeometric probability function is used to compute the probability that in a random selection of n elements, selected _____, we obtain x elements labeled success and $n-x$ elements labeled failure.

5. For this outcome to occur, we must obtain x successes from the r successes in the population and $n-x$ failures from the $N-r$ failures.

6. **Hypergeometric Probability Function** (Denote by _____):

$$f(x) = \frac{\quad}{\quad}$$

where:

$$\begin{aligned} x &= \text{number of successes}, x = 0, 1, 2, \cdots, n \\ n &= \text{number of trials} \\ f(x) &= \text{probability of } x \text{ successes in } n \text{ trials} \\ N &= \text{number of elements in the population} \\ r &= \text{number of elements in the population labeled success}, x \leq r \end{aligned}$$

$\binom{N}{n}$: the number of ways n elements can be selected from a population of size N

$\binom{r}{x}$: the number of ways that x successes can be selected from a total of r successes in the

$\binom{N-r}{n-x}$: the number of ways that $n-x$ failures can be selected from a total of $N-r$ failures in

7. The mean and variance of a hypergeometric distribution are

$$\begin{aligned} E(X) &= \mu = \frac{\quad}{\quad} \\ Var(X) &= \sigma^2 = \frac{\quad}{\quad} \end{aligned}$$

✎ Question ..... (p263)

Electric fuses produced by Ontario Electric are packaged in boxes of 12 units each. Suppose an inspector randomly selects three of the 12 fuses in a box for testing. (a) If the box contains exactly five defective fuses, what is the probability that the inspector will find exactly one of the three fuses defective? (b) What is the probability of finding at least one defective fuse. (c) Find the mean and variance for the number of defective fuses.

sol:

😍 EXERCISES

5.1 : 1, 3, 5

5.2 : 9, 10

5.3 : 16, 21, 23

5.4 : 26, 28, 29

5.5 : 32, 35, 39, 42

5.6 : 45, 49, 50

5.7 : 55, 57, 58

SUP : 59, 62, 66, 73

😍 See Also: SUMMARY, GLOSSARY, KEY FORMULAS

名言佳句

“我們每天做什麼，就會成為什麼樣的人。因此，卓越不是一種行為，而是一個習慣”
“We are what we repeatedly do. Excellence, therefore, is not an act, but a habit.”
— *Aristotle (384–322 BC)*

統計小故事：拉普拉斯與二項分配的誕生

十八世紀的法國，正處於啟蒙運動的高峰期。人們開始以理性與科學的方法思考世界，而不再完全依賴傳統與宗教。數學家皮埃爾．西蒙．拉普拉斯 (Pierre-Simon Laplace) 正是在這樣的時代中嶄露頭角。他的一生，幾乎橫跨了整個法國大革命與拿破崙時代，他的數學工作深深影響了科學、天文學，甚至政治與社會。

拉普拉斯對「隨機性」特別感興趣。當時，賭博遊戲風行歐洲，擲骰子、紙牌、抽籤成為貴族與平民共同的娛樂。拉普拉斯注意到，這些遊戲雖然帶有偶然性，但長期觀察下卻展現出某種規律。例如：擲一枚公平的硬幣，單次結果難以預測，但若擲上一百次，正反面大致接近各一半。這種「在多次試驗中趨於穩定的比例」，引起了他的數學直覺。

於是，拉普拉斯開始思考：如何用數學公式，去描述「某事件在重複試驗中出現的次數分布」? 他的研究最終導向了二項分配 (binomial distribution) 的發展。

二項分配的精髓在於：若一個事件只有「成功」與「失敗」兩種結果，且每次試驗的成功機率固定，那麼在 n 次試驗中，成功 k 次的機率，就能用二項式展開公式計算。雖然這套公式最早可追溯至雅各布．伯努利 (Jacob Bernoulli)，但拉普拉斯將它系統化並推廣，使其成為統計學與應用數學的基礎工具之一。

舉個例子：假設某工廠生產的螺絲有 5% 的瑕疵率。如果我們隨機抽取 10 顆螺絲，恰好發現 2 顆有瑕疵的機率是多少？拉普拉斯會告訴我們，這就是一個二項分配問題：$n = 10$，$p = 0.05$，$k = 2$。利用二項分配公式，我們可以精確算出答案。

這樣的數學工具，對於工業革命時期的品質管理極為重要。過去工匠作坊生產規模小，產品品質全憑師傅經驗。但在大規模工廠中，必須依靠統計方法監控品質。二項分配使得企業能夠透過抽樣檢驗，合理推估整體產品的合格率，這大大提高了效率與信任度。

不僅如此，二項分配也被應用在醫學與社會科學中。例如，若一項新藥有 80% 的成功率，醫生想知道在治療 20 位病人時，至少有 15 人康復的機率是多少？這同樣可以用二項分配來回答。在政治民調中，若支持率為 60%，隨機抽樣 100 人，觀察到支持者人數的分布也遵循二項分配。

拉普拉斯的貢獻不僅是公式的建立，更在於觀念的轉變。他讓人們意識到：離散的隨機事件並非完全不可測，而是可以透過數學語言來理解與預測。這使得人類第一次能夠用嚴謹的方式掌握「不確定性」。

在教材呼應上，本章的核心是「離散機率分配」，其中二項分配是最重要的一種。學生需要理解二項分配的條件：每次試驗獨立、結果只有兩種、成功機率固定。透過學

習這些概念，學生能將理論與實務連結，從擲硬幣的遊戲，到工業抽樣檢驗，乃至於醫學實驗與社會調查，都能找到二項分配的身影。

更重要的是，二項分配為後續的「常態近似」與「中心極限定理」鋪路。當試驗次數很大時，二項分配的形狀會趨近於鐘形曲線，這為第六章的連續機率分布做好了橋接。

因此，當學生在計算「十次擲硬幣，恰好出現六次正面」的機率時，請記住，他們其實正在重現拉普拉斯的思考軌跡。那不僅是課本上的練習，而是數學史上一場重要的知識革命：人類第一次用公式，捕捉到隨機性的脈動。

統計學 (一)

Anderson's Statistics for Business & Economics (14/E)

Chapter 6: Continuous Probability Distributions

上課時間地點: 二 D56, 商館 260206
授課教師: 吳漢銘 (國立政治大學統計學系副教授)
教學網站: https://hmwu.idv.tw

系級: _____ 學號: _____ 姓名: _____

Overview

1. For a discrete random variable, the probability function $f(x)$ provides the probability that the random variable assumes a _____.
2. With continuous random variables, the counterpart of the probability function is the _____ (pdf), also denoted by _____.
3. The _____ under the graph of $f(x)$ corresponding to a given _____ does provide the probability that the continuous random variable X assumes a value in that interval.
4. The _____ distribution is of major importance because of its wide applicability and its extensive use in _____.
5. The _____ distribution is useful in applications involving such factors as waiting times and service times.

6.1 Uniform Probability Distribution

1. **Uniform Probability Density Function**: the _____ probability density function for a random variable X (denoted by _____):

$$f(x) = \begin{cases} \dfrac{\quad}{\quad}, & \dfrac{\quad}{\quad} \\ 0, & \text{elsewhere} \end{cases}$$

2. For a continuous random variable, we consider probability only in terms of the likelihood that a random variable assumes a value _____ a specified _____.

3. Example Flight Time Example:

   (a) Consider the random variable X representing the _____ in the interval from 120 minutes to 140 minutes of an airplane traveling from Chicago to New York.

   (b) Because the random variable X can assume any value in that interval, X is a _____ rather than a discrete random variable.

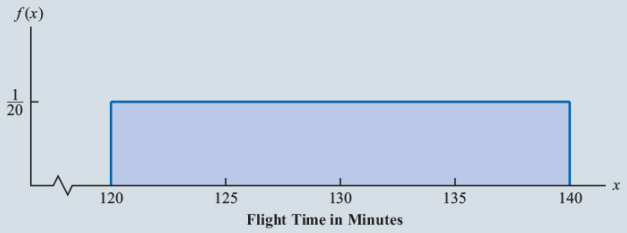
   (c) Let us assume that sufficient actual flight data are available to conclude that the probability of a flight time within any 1-minute interval is the same as the probability of a flight time within any other 1-minute interval contained in the larger interval from 120 to 140 minutes.

   (d) With every 1-minute interval being equally likely, the random variable X is said to have a _____.

   (e) (Figure 6.1) The probability density function, which defines the uniform distribution for the flight-time random variable, is

$$f(x) = _____.$$

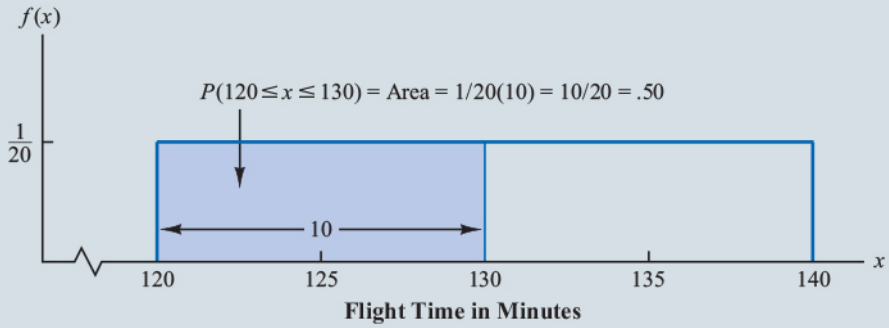
FIGURE 6.1 Uniform Probability Distribution for Flight Time



Area as a Measure of Probability

1. (Figure 6.2) The area under the graph of $f(x)$ in the interval from 120 to 130 provides probability of a flight time between 120 and 130 minutes: _____.

FIGURE 6.2 Area Provides Probability of a Flight Time between 120 and 130 Minutes



2. Once a probability density function $f(x)$ is identified, the probability that X takes a value between some lower value a and some higher value b can be found by computing the area under the graph of $f(x)$ over the interval from a to b.

3. Note that _____; that is, the total area under the graph of $f(x)$ is equal to 1. This property holds for all continuous probability distributions.

4. For a continuous probability density function, we must also require that _____ for all values of X.

5. Given the uniform distribution for flight time, what is the probability of a flight time between 128 and 136 minutes?

   _____.

   sol:

6. Two major differences stand out between the treatment of continuous random variables and the treatment of their discrete counterparts.

(a) We no longer talk about the probability of the random variable assuming a _____. Instead, we talk about the probability of the random variable assuming a value within some _____.

(b) The _____ of a continuous random variable assuming a value within some given interval from a to b is defined to be the _____ under the graph of the probability density function between a and b.

7. Because a single point is an interval of _____, this implies that the probability of a continuous random variable assuming any particular value exactly is _____.

8. The expected value and variance for a continuous uniform random variable is

$$E(X) = _____, \quad Var(X) = _____.$$

9. Example Applying $E(X)$ and $Var(X)$ to the uniform distribution for flight times from Chicago to New York, we obtain

$$E(X) = _____, \quad Var(X) = _____$$

The standard deviation of flight times can be found by taking the square root of the variance. Thus, $\sigma = 5.77$ minutes.

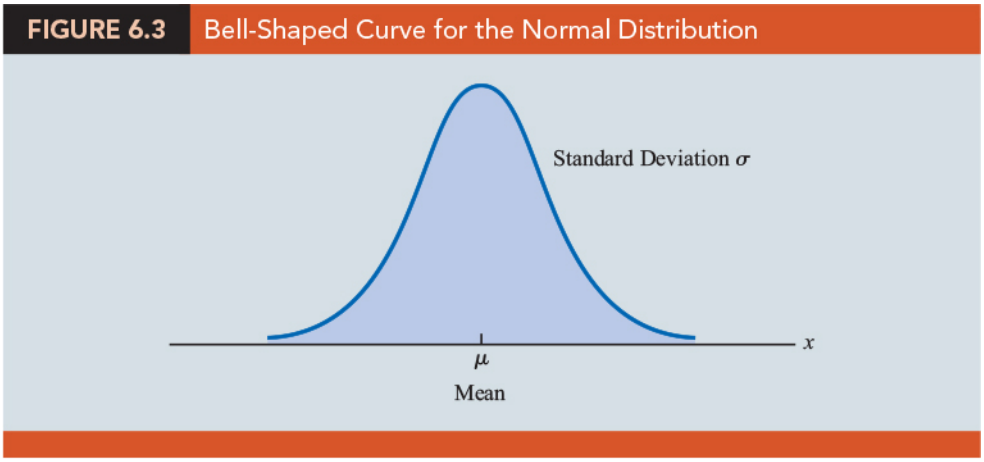
6.2 Normal Probability Distribution

1. The normal distribution (_____ distribution) has been used in a wide variety of practical applications in which the _____ are heights and weights of people, test scores, scientific measurements, amounts of rainfall, and other similar values.

2. It is also widely used in _____.

Normal Curve

1. (Figure 6.3) The form, or shape, of the normal distribution is illustrated by the _____.



2. **Normal Probability Density Function**:

$$f(x) = _____$$

where

μ = mean
σ = standard deviation
π = 3.14159
e = _____

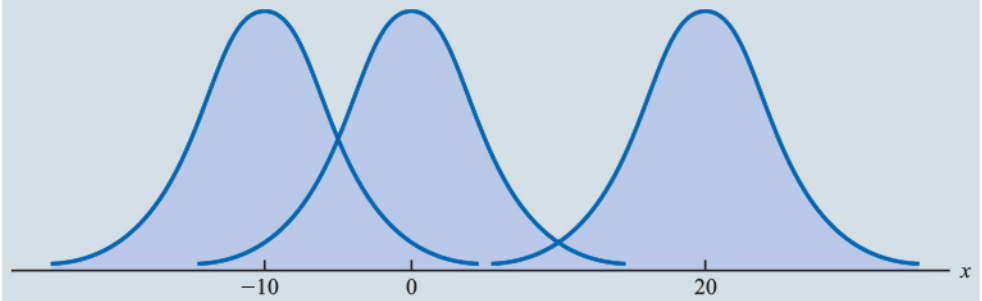
3. Several observations about the characteristics of the normal distribution.

   (a) The entire family of normal distributions is differentiated by two parameters: the mean _____ and the standard deviation _____.

   (b) The highest point on the normal curve is at the _____, which is also the _____ and _____ of the distribution.

   (c) (Figure 6.4) The mean of the distribution can be any numerical value: negative, zero, or positive.

FIGURE 6.4 Normal Distributions with Same Standard Deviation and Different Means

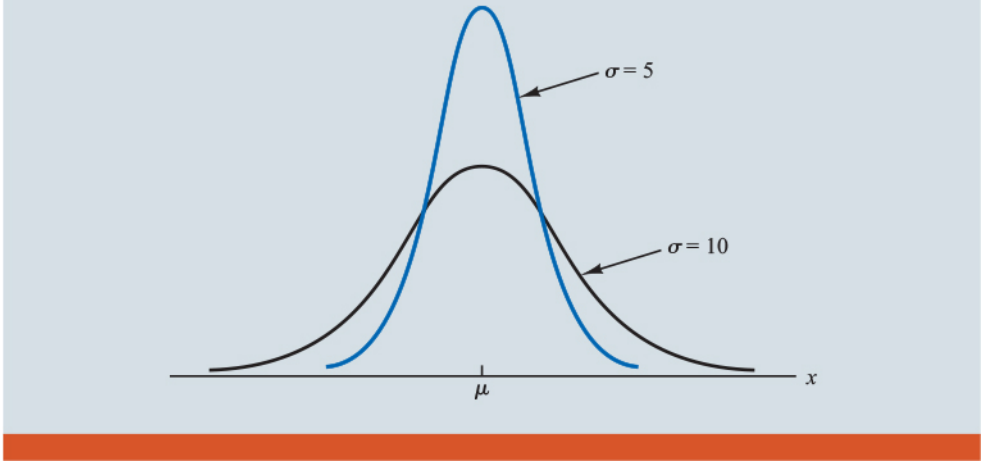


(d) The normal distribution is _____, with the shape of the normal curve to the left of the mean a _____ of the shape of the normal curve to the right of the mean.

(e) The tails of the normal curve extend to _____ in both directions and theoretically never touch the _____. Because it is symmetric, the normal distribution is _____; its skewness measure is _____.

(f) (Figure 6.5) The standard deviation determines how _____ and _____ the normal curve is. Larger values of the standard deviation result in wider, flatter curves, showing more _____ in the data.

FIGURE 6.5 Normal Distributions with Same Mean and Different Standard Deviations

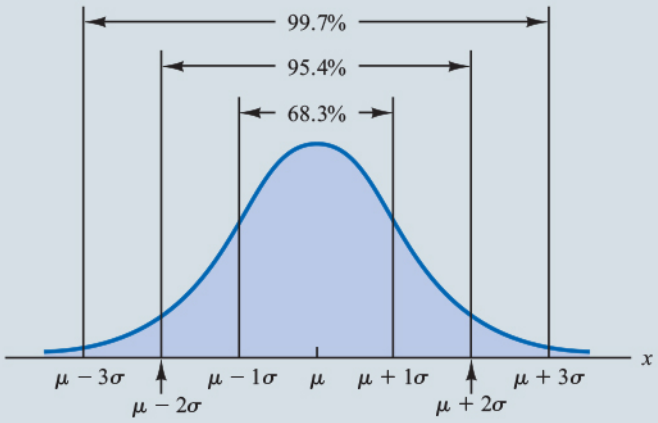


(g) Probabilities for the normal random variable are given by areas under the normal curve. The total area under the curve for the normal distribution is

_____. Because the distribution is symmetric, the area under the curve to the left (right) of the mean is _____.

(h) (Figure 6.6) The percentage of values in some commonly used intervals: _____ _____ of the values of a normal random variable are within plus or minus _____ standard deviation of its mean.

FIGURE 6.6 Areas Under the Curve for Any Normal Distribution

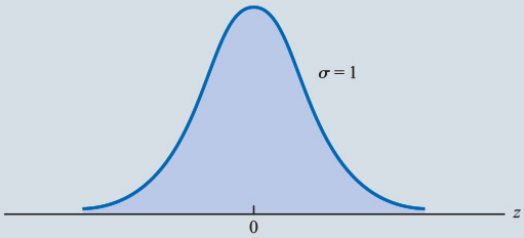


Standard Normal Probability Distribution

1. (Figure 6.7) A random variable, _____, that has a normal distribution with a mean of _____ and a standard deviation of _____ is said to have a _____ normal probability distribution.

$$f(z) = _____$$

FIGURE 6.7 The Standard Normal Distribution



2. The three types of probabilities we need to compute:

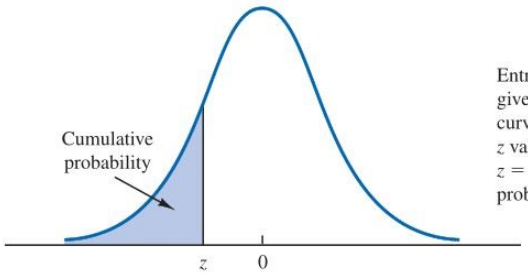
(a) the probability that the standard normal random variable Z will be less than or equal to a given value z; _____

(b) the probability that Z will be between two given values z_1, z_2; _____

(c) the probability that Z will be greater than or equal to a given value z; _____.

3. (Appendix B, Table 1) Cumulative probabilities for the standard normal probability:

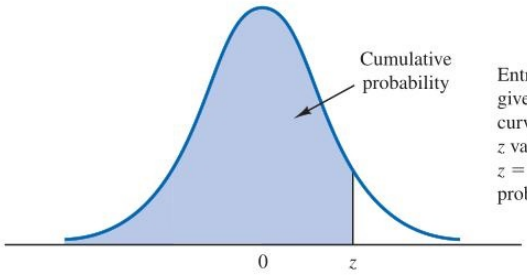
TABLE 1 Cumulative Probabilities for the standard Normal Distribution



Entries in the table give the area under the curve to the left of the z value. For example, for $z = -.85$, the cumulative probability is .1977.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064

TABLE 1 Cumulative Probabilities for the standard Normal Distribution (*Continued*)



Entries in the table give the area under the curve to the left of the z value. For example, for $z = 1.25$, the cumulative probability is .8944.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319

4. The table of cumulative probabilities for the standard normal probability distribution can be used to find _____ associated with values of the standard normal random variable Z.

5. Two types of questions can be asked.

   (a) The first type of question specifies a _____, for Z and asks us to use the table to determine the corresponding areas or probabilities.

   (b) The second type of question provides an area, or _____, and asks us to use the table to determine the corresponding Z value.

✎ Question ..... (p290)

Using the standard normal probability Table to compute

1. (Figure 6.8) the probability that the standard normal random variable Z is less than or equal to 1.00;
2. (Figure 6.9) the probability that Z is in the interval between -0.50 and 1.25;
3. (Figure 6.10) the probability that Z is within one standard deviation of the mean;
4. (Figure 6.11) the probability of obtaining a Z value of at least 1.58.

sol:

1. $P(Z \leq 1.00) =$
2. $P(-0.50 \leq Z \leq 1.25) =$
3. $P(-1.00 \leq Z \leq 1.00) =$
4. $P(Z \geq 1.58) =$

FIGURE 6.8 Cumulative Probability for Normal Distribution Corresponding to $P(z \le 1.00)$

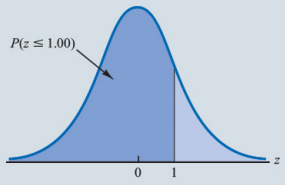


FIGURE 6.9 Cumulative Probability for Normal Distribution Corresponding to $P(-.50 \le z \le 1.25)$

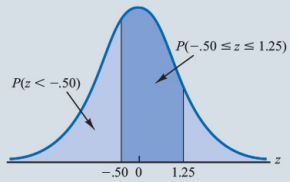


FIGURE 6.10 Cumulative Probability for Normal Distribution Corresponding to $P(-1.00 \le z \le 1.00)$

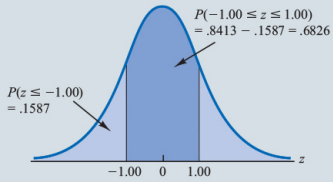
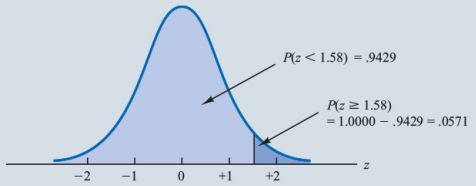


FIGURE 6.11 Cumulative Probability for Normal Distribution Corresponding to $P(z \ge 1.58)$

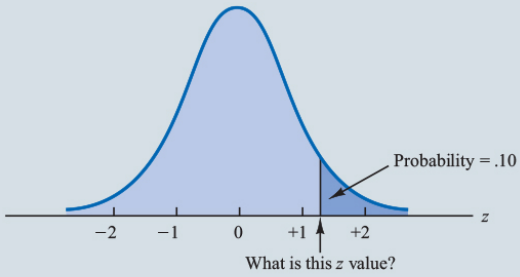


Question (p293)

(Figure 6.12) Find a Z value such that the probability of obtaining a larger Z value is 0.10.

sol:

FIGURE 6.12 Finding z Value such that Probability of Obtaining a Larger z Value is .10



Computing Probabilities for Any Normal Probability Distribution

1. Probabilities for all normal distributions can be computed using the _____ normal distribution.

2. Convert any normal random variable X with mean μ and standard deviation σ to the standard normal random variable Z:

   _____

3. We can interpret Z as the _____ that the normal random variable X is from its _____.

✎ Question ..... (p294)

Suppose we have a normal random variable X with $\mu = 10$ and $\sigma = 2$. What is the probability that the random variable X is between 10 and 14?

sol:

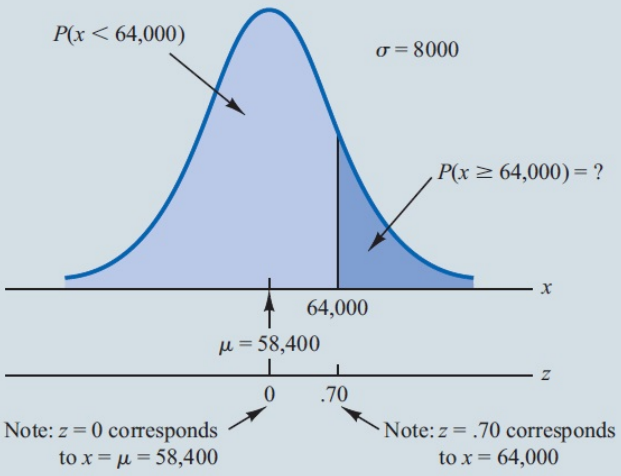
Grear Tire Company Problem

Question ..... (p294)

(**Grear Tire Company Problem**) (Figure 6.13) Suppose the Grear Tire Company developed a new steelbelted radial tire to be sold through a national chain of discount stores. Because the tire is a new product, Grear's managers believe that the mileage guarantee offered with the tire will be an important factor in the acceptance of the product. Before finalizing the tire mileage guarantee policy, Grear's managers want probability information about X = number of miles the tires will last. From actual road tests with the tires, Grear's engineering group estimated that the mean tire mileage is $\mu = 36,500$ miles and that the standard deviation is $\sigma = 5000$. In addition, the data collected indicate that a normal distribution is a reasonable assumption. What percentage of the tires can be expected to last more than 40,000 miles? In other words, what is the probability that the tire mileage, X, will exceed 40,000?

sol:

FIGURE 6.13 Grear Tire Company Tire Lifetime Distribution

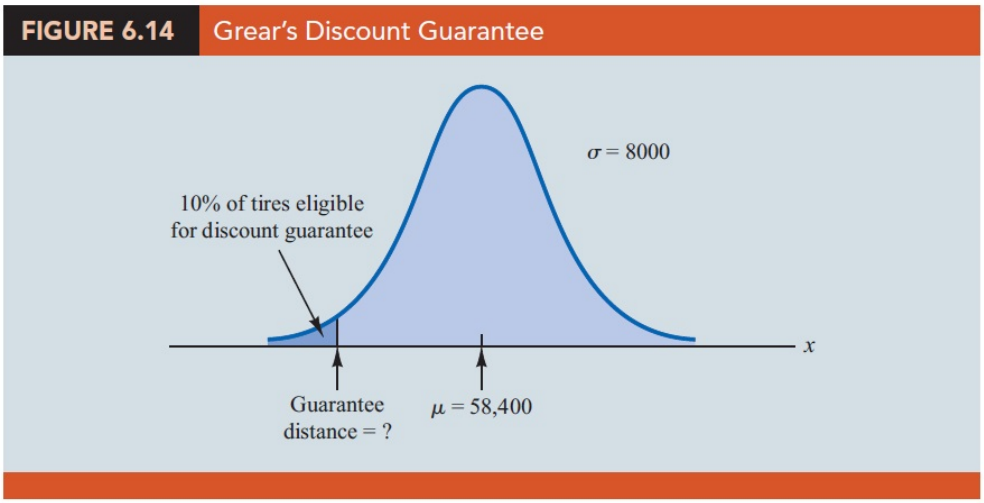


Question ..... (p294)

(**Grear Tire Company Problem**) (Figure 6.14) Let us now assume that Grear is considering a guarantee that will provide a discount on replacement tires if the original tires do not provide the guaranteed mileage. What should the guarantee mileage be if Grear wants no more than 10% of the tires to be eligible for the discount guarantee?

sol:

FIGURE 6.14 Grear's Discount Guarantee



1. **The important role of the probability distributions**: providing decision - making information. Namely, once a probability distribution is established for a particular application, it can be used to obtain _____ information about the problem.
2. Probability does not make a decision recommendation directly, but it provides information that helps the decision maker better understand the _____ and _____ associated with the problem.

6.3 Normal Approximation of Binomial Probabilities

1. Recall that a binomial experiment:
   (a) consists of a sequence of _____ with each trial having two possible outcomes, a success or a failure.
   (b) The probability of a _____ on a trial is the same for all trials and is denoted by _____.
   (c) The binomial random variable is the _____ in the n trials, and probability questions pertain to the _____ of x successes in the n trials.
2. In cases where _____, and _____, the normal distribution provides an _____ of binomial probabilities.
3. When using the normal approximation to the binomial, we set _____ and _____ in the definition of the normal curve.
4. **Continuity correction**: Recall that, with a continuous probability distribution, probabilities are computed as areas under the probability density function. As a result, the probability of any single value for the random variable is _____. Thus to approximate the binomial probability of x successes, we compute the area under the corresponding normal curve between _____ and _____. The 0.5 that we add and subtract from x is called a _____.
5. 補充: 理論上以常態機率逼近二項式機率，可以使用的情境是: 「n 要很大且 p 不要近於 0 或 1」。但即使「n 有點小且 p 近於 1/2」，逼近效果也是可在接受範圍。依據上面兩點，不同教科書裡有一些不儘相同的經驗法則 (rules of thumb):
   (a) $np \geq 5$ 且 $n(1-p) \geq 5$。[我們的教科書]
   (b) $n > 5$, 且 $|skewness| < 1/3$。
   (c) $(np) \pm \sqrt{np(1-p)} \in (0, n)$。(等同 $np > 9(1-p)$ 且 $n(1-p) > 9p$)。
   (d) $np(1-p) \geq 10$
   (e) n 很大且 $np >= 10$，且 $n(1-p) \geq 10$

在下面篇 1945 年的論文裡，証明中的確有提到「$\sqrt{np(1-p)} > 5$」的條件。W. Feller, (1945) On the Normal Approximation to the Binomial Distribution, Ann. Math. Statist. 16(4): 319-329 (December, 1945). DOI: 10.1214/aoms/1177731058。

6. 補充: 使用「連續修正項」需注意「有沒有等號」:

Binomial	Normal
If $P(X = n)$	use $P(n - 0.5 < X < n + 0.5)$
If $P(X > n)$	use $P(X > n + 0.5)$
If $P(X \leq n)$	use $P(X < n + 0.5)$
If $P(X < n)$	use $P(X < n - 0.5)$
If $P(X \geq n)$	use $P(X > n - 0.5)$

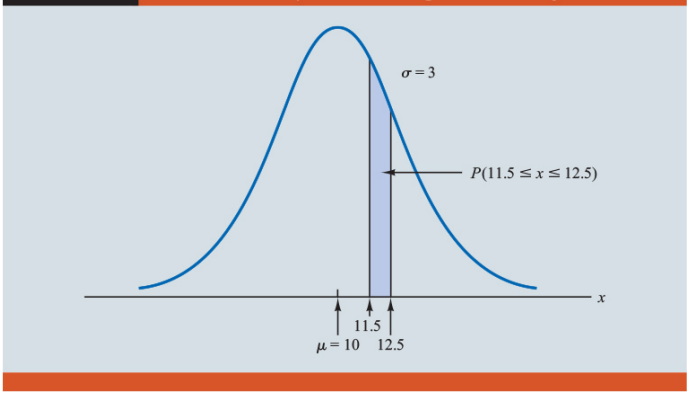
7. 補充: 「Normal Approximation + continuity correction」R 程式範例：
`https://www.macmillanlearning.com/studentresources/college/statistics/introstats3e/optional_sections/00_KokoskaIntroStat3e_04962_ch06.5_online_001_010_4PP.pdf`

✎ Question ..... (p299)

(Figure 6.15) Suppose that a particular company has a history of making errors in 10% of its invoices. A sample of 100 invoices has been taken, and we want to compute the probability that 12 invoices contain errors. That is, find the binomial probability of 12 successes in 100 trials by applying the normal approximation.

sol:

FIGURE 6.15 Normal Approximation to a Binomial Probability Distribution with $n = 100$ and $p = .10$ Showing the Probability of 12 Errors

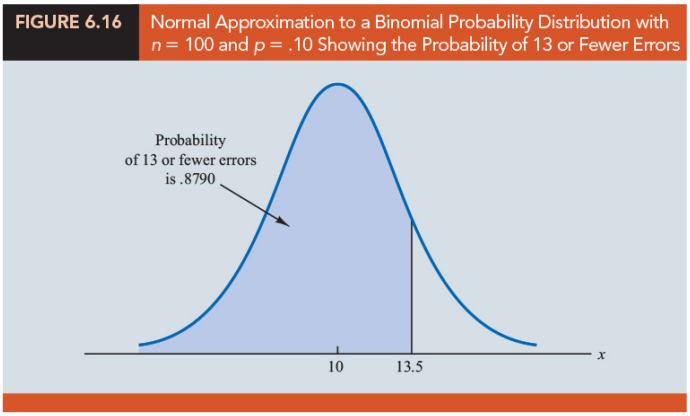


Question (p300)

(Figure 6.16) Following the previous question, compute the probability of 13 or fewer errors in the sample of 100 invoices.

sol:

FIGURE 6.16 Normal Approximation to a Binomial Probability Distribution with $n = 100$ and $p = .10$ Showing the Probability of 13 or Fewer Errors



6.4 Exponential Probability Distribution

1. The exponential probability distribution may be used for random variables such as the _____ between arrivals at a hospital emergency room, the _____ required to load a truck, the _____ between major defects in a highway, and so on.

2. The exponential probability density function (denoted by _____):

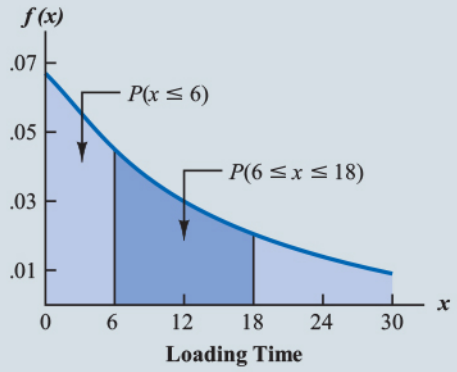
$$f(x) = _____, \quad x \geq 0,$$

where μ = expected value or mean.

3. (Figure 6.17) (Schips loading dock example) Suppose that X represents the loading time for a truck at the Schips loading dock and follows a exponential distribution. If the mean, or average, loading time is 15 minutes (_____), the appropriate probability density function for X is

$$f(x) = _____$$

FIGURE 6.17 Exponential Distribution for the Schips Loading Dock Example



4. The mean and the standard deviation of the exponential distribution are _____.

5. (Figure 6.17) The probability that loading a truck will take 6 minutes or less, _____, is defined to be the area under the curve in Figure 6.17 from $x = 0$ to $x = 6$. The probability that the loading time will be between 6 minutes and 18 minutes, _____, is given by the area under the curve from $x = 6$ to $x = 18$.

Computing Probabilities for the Exponential Distribution

1. The cumulative probability of obtaining a value for the exponential random variable of less than or equal to some specific value denoted by x:

$$F(x) = P(X \leq x) = _____.$$

✎ Question ..... (p303)

In the Schips loading dock example, Find (a) the probability that loading a truck will take 6 minutes or less; (b) the probability that the loading time will be 18 minutes or less; (c) the probability that the loading time will be between 6 minutes and 18 minutes.

sol:

Relationship Between the Poisson and Exponential Distributions

1. (Section 5.6) The Poisson distribution is a discrete probability distribution that is often useful in examining the _____ of an event over a specified interval of time or space.

$$f(x) = _____,$$

where μ = expected value or mean number of occurrences over a specified interval.

2. If the Poisson distribution provides an appropriate description of the _____ _____, the exponential distribution provides a description of the _____ _____.

3. Example Suppose the number of patients who arrive at a hospital emergency room during one hour is described by a Poisson probability distribution with a mean of 10 patients per hour. The Poisson probability function that gives the probability of x arrivals per hour is

$$f(x) = _____$$

Because the average number of arrivals is 10 patients per hour, the average time between patients arriving is

$$_____ \text{ hour/patient.}$$

Thus, the corresponding exponential distribution that describes the time between the arrivals has a mean of _____ hour per patient; as a result, the appropriate exponential probability density function is

$$f(x) = _____.$$

☺ R **functions**:

1. `dnorm, pnorm, qnorm, rnorm`:
   `https://stat.ethz.ch/R-manual/R-devel/library/stats/html/Normal.html`
2. `dexp, pexp, qexp, rexp`:
   `https://stat.ethz.ch/R-manual/R-devel/library/stats/html/Exponential.html`

😍 EXERCISES

6.1 : 2, 6

6.2 : 11, 14, 19, 23

6.3 : 27, 30

6.4 : 33, 37, 38

SUP : 39, 42, 49, 52, 53

名言佳句

“樹林裡有兩條岔路，而我走了人跡較少的那一條，因此有了完全不同的人生”
“Two roads diverged in a wood, and I took the one less traveled by, And that has made all the difference.”

— Robert Frost (March 26, 1874 – January 29, 1963)

統計小故事：高斯與鐘形曲線

十九世紀初的德國，正值科學與數學迅速發展的黃金時期。在哥廷根大學，一位年輕數學家卡爾．弗里德里希．高斯 (Carl Friedrich Gauss) 正在觀測行星的運動軌跡。當時的天文學家面臨一個難題：如何在充滿誤差的觀測數據中，推算出天體的真實位置？

望遠鏡雖已進步，但觀測結果仍充滿偏差。有些數據比真實位置偏高，有些偏低。問題在於：這些誤差並非完全隨機，似乎遵循著某種規律。高斯敏銳地察覺到，如果能找到誤差的分布模式，就能用數學方法修正觀測，進而更精準地描繪宇宙。

經過長時間的計算與推理，高斯提出了一個驚人的發現：誤差的出現並非均勻隨機，而是呈現「中間多、兩端少」的形態。也就是說，大部分觀測值集中在接近真實值的位置，而極端偏差則相對罕見。為了描述這種現象，他提出了數學函數，後來被稱為常態分配 (normal distribution)，或俗稱「鐘形曲線」。

鐘形曲線具有幾個重要特徵：呈對稱形狀，中心是平均數。大多數數值集中在平均數附近。偏離平均數越遠，機率越小。這一模型不僅解釋了天文觀測的誤差，也揭示了自然界的普遍規律。人的身高、考試成績、血壓值，甚至金融市場的日報酬率，都常接近常態分布。這使得鐘形曲線成為統計學中最重要的連續機率分布之一。

高斯的發現意義非凡。他不僅解決了行星軌跡的推算問題，也奠定了現代統計推論的基礎。許多統計方法，如 t 檢定、迴迴歸分析與信賴區間的計算，都建立在常態分布的假設之上。可以說，若沒有高斯的鐘形曲線，現代統計學將失去一塊最重要的基石。

更有趣的是，常態分布與我們日常生活息息相關。當學生拿到考試成績時，往往會看到「平均數」與「標準差」。這其實就是利用常態分布來理解學生成績的分布狀況。如果一門課的平均分是 70 分，標準差是 10，那麼大約有 68% 的學生介於 60 到 80 分之間，約 95% 的學生介於 50 到 90 分之間。這就是常態分布的「68-95-99.7 法則」。

此外，在醫學臨床上，常態分布被廣泛用於描述生理指標。例如血壓、膽固醇濃度、體重指數，通常都呈常態分布。醫生判斷病人是否異常，就是根據是否落在分布的極端尾部。這樣的應用直接影響診斷與治療。

金融市場中，投資者也會用常態分布來建模資產報酬。雖然現實市場存在「厚尾現象」，不完全符合常態，但鐘形曲線依然是投資風險管理的重要起點。它讓投資人能用數學方式量化風險，而不是只憑直覺。

在教材呼應上，本章的重點是「連續型機率分布」。相較於離散分布只處理「個別次

數」，連續分布處理的是「範圍內的機率」。例如，離散分布回答「十次擲硬幣，恰好出現六次正面的機率是多少？」；而常態分布則回答「一位學生的分數落在 60 到 80 之間的機率是多少？」這種從點到區間的轉變，正是連續分布的精髓。

因此，當學生在課堂上畫出鐘形曲線，並計算某個範圍下的機率時，他們其實正延續高斯兩百年前的發現。那不僅是數學上的演練，更是人類在理解世界過程中的一個重大突破。

總結來說，高斯的鐘形曲線提醒我們：即使是誤差，也有其數學規律。而這種規律，讓人類從混亂中看見秩序，從不確定性中找到穩定。這就是連續機率分布的力量，也是統計學真正的美麗所在。

統計學 (一)

Anderson's Statistics for Business & Economics (14/E)

Chapter 7: Sampling and Sampling Distributions

上課時間地點: 二 D56, 商館 260206
授課教師: 吳漢銘 (國立政治大學統計學系副教授)
教學網站: https://hmwu.idv.tw

系級: _____ 學號: _____ 姓名: _____

Overview

1. Definitions:
   (a) An _____ is the entity on which data are collected.
   (b) A _____ is the collection of all the elements of interest.
   (c) A _____ is a subset of the population.
2. Why select a sample? Answer: collect data to make an _____ and answer research _____ about a population.
3. The sample contains only a _____ of the population. Sample results provide only _____ of the values of the corresponding population characteristics.
4. Some _____ is to be expected. With proper sampling methods, the sample results will provide _____ of the population parameters.
5. Numerical characteristics of a population are called _____.
6. This chapter:

(a) show how simple random sampling can be used to select a sample from a _____.

(b) describe how a random sample can be taken from an _____ population that is generated by an ongoing process.

(c) how to compute _____ of a population mean, a population standard deviation, and a population proportion.

(d) introduce the important concept of a _____. It enables us to make statements about how close the _____ are to the corresponding _____.

(e) discuss some alternatives to simple random sampling that are often employed in practice and the ramifications of _____ on sampling distributions.

7.1 The Electronics Associates Sampling Problem

1. *Background*: The director of personnel for Electronics Associates, Inc. (EAI), has been assigned the task of developing a profile of the company's 2500 managers (population, $N = 2500$).

2. *Interested*: The characteristics to be identified include the _____ annual salary for the managers and the _____ of managers having completed the company's management training program.

3. *Population Data*: The population contains the annual salary and the training program status for all 2500 managers (referring to the firm's personnel records) (Dataset: `EAI`).

4. *Population Parameters*: The population _____ annual salary ($\mu = \$71,800$), the population _____ of annual salary ($\sigma = \$4000$), and the population _____ that completed the training program ($p = 1500/2500 = 0.60$) are parameters of the population of EAI managers.

5. *Question*: Suppose that the necessary information on all the EAI managers was not readily available in the company's database. How the firm's director of personnel can obtain _____ of the population parameters by using a sample of managers, say a sample of 30 managers ($n = 30$).

6. *Sample Data*: If the personnel director could be assured that a sample of 30 managers would provide adequate information about the population of 2500 managers, the _____ and the _____ of developing a profile would be substantially less.

7.2 Selecting a Sample

Sampling from a Finite Population

1. **A simple random sample**: A _____ from a finite population of size N is a sample selected such that each possible sample of size n has the same probability of being selected.

2. One procedure for selecting a simple random sample from a finite population is to use a _____ to choose the elements for the _____ at a time in such a way that, at each step, each of the elements remaining in the population has the _____ of being selected.

3. Example EAI managers

   (a) First _____ 1 to 2500 in the order that their names appear in the EAI personnel file.

   (b) (Table 7.1) Select random numbers from _____ in the table and _____ in a direction of our choice in sets or groups of _____ (largest number in the population list).

   (c) E.g., Use the first row of Table 7.1 and move from left to right. The first 7 four-digit random numbers are

6327 1599 8671 7445 1102 1514 1807

TABLE 7.1 Random Numbers

63271	59986	71744	51102	15141	80714	58683	93108	13554	79945
88547	09896	95436	79115	08303	01041	20030	63754	08459	28364
55957	57243	83865	09911	19761	66535	40102	26646	60147	15702
46276	87453	44790	67122	45573	84358	21625	16999	13385	22782
55363	07449	34835	15290	76616	67191	12777	21861	68689	03263
69393	92785	49902	58447	42048	30378	87618	26933	40640	16281
13186	29431	88190	04588	38733	81290	89541	70290	40113	08243
17726	28652	56836	78351	47327	18518	92222	55201	27340	10493
36520	64465	05550	30157	82242	29520	69753	72602	23756	54935
81628	36100	39254	56835	37636	02421	98063	89641	64953	99337
84649	48968	75215	75498	49539	74240	03466	49292	36401	45525
63291	11618	12613	75055	43915	26488	41116	64531	56827	30825
70502	53225	03655	05915	37140	57051	48393	91322	25653	06543
06426	24771	59935	49801	11082	66762	94477	02494	88215	27191
20711	55609	29430	70165	45406	78484	31639	52009	18873	96927
41990	70538	77191	25860	55204	73417	83920	69468	74972	38712
72452	36618	76298	26678	89334	33938	95567	29380	75906	91807
37042	40318	57099	10528	09925	89773	41335	96244	29002	46453
53766	52875	15987	46962	67342	77592	57651	95508	80033	69828
90585	58955	53122	16025	84299	53310	67380	84249	25348	04332
32001	96293	37203	64516	51530	37069	40261	61374	05815	06714
62606	64324	46354	72157	67248	20135	49804	09226	64419	29457
10078	28073	85389	50324	14500	15562	64165	06125	71353	77669
91561	46145	24177	15294	10061	98124	75732	00815	83452	97355
13091	98112	53959	79607	52244	63303	10413	63839	74762	50289

(d) The first number, 6327, is greater than 2500: _____. The first manager selected for the random sample is the 2nd number 1599 on the list of EAI managers.

(e) This process continues until the simple random sample of 30 EAI managers has been obtained.

(f) Any previously used random numbers are _____ because the corresponding manager is already included in the sample.

4. If we selected a sample such that previously used random numbers are acceptable and specific managers could be included in the sample two or more times, we would be _____. Otherwise, it is referred to as _____ _____.

5. Sampling without replacement is the sampling procedure used most often in practice. When we refer to simple random sampling, we will assume the sampling is

_____ replacement.

Sampling from an Infinite Population

1. *The infinite population case*: The population is _____ or the elements of the population are being generated by an ongoing process for which there is _____ on the number of elements that can be generated. Thus, it is not possible to develop a _____ of all the elements in the population.
2. A random sample of size n from an infinite population is a sample selected such that two conditions are satisfied.
   (a) Each element selected comes from the _____ population.
   (b) Each element is selected _____.
3. The purpose of the second condition of the random sample selection procedure (each element is selected independently) is to prevent _____.
4. Example 1: Quality Control Inspection
   (a) Consider a production line designed to fill boxes of a breakfast cereal with a mean weight of 24 ounces of breakfast cereal per box.
   (b) Samples of 12 boxes filled by this process are _____ by a quality control inspector to determine if the process is operating properly or if, perhaps, a machine malfunction has caused the process to begin underfilling or overfilling the boxes.
   (c) *Condition 1*: The boxes must be selected at approximately the _____. This way the inspector avoids the possibility of selecting some boxes when the process is operating properly and other boxes when the process is not operating properly and is underfilling or overfilling the boxes.
   (d) *Condition 2*: designing the production process so that each box of cereal is filled _____.
5. Example 2: Customers Arrival
   (a) Suppose an employee is asked to select and interview a sample of customers in order to develop a profile of customers who visit the fast-food restaurant.

(b) The customer arrival process is ongoing and there is no way to obtain a list of all customers in the population. So, for practical purposes, the population for this ongoing process is considered _____.

(c) *Condition 1*: The employee collects the sample from people who come into the restaurant and make a purchase to ensure that the same population condition is satisfied. the person came into the restaurant to use the restroom would not be a customer.

(d) *Condition 2*: Selection bias (a particular age group, a group of customers would be likely to exhibit similar characteristics) can be avoided by ensuring that the selection of a particular customer does not influence the selection of any other customer.

6. Situations involving sampling from an infinite population are usually associated with a process that operates _____. Examples include parts being manufactured on a production line, repeated experimental trials in a laboratory, transactions occurring at a bank, telephone calls arriving at a technical support center, and customers entering a retail store.

7. As long as the sampled elements are selected from the _____ and are selected _____, the sample is considered a random sample from an infinite population.

7.3 Point Estimation

1. Glossary:

(a) **Parameter**: A _____ characteristic of a _____, such as a population mean μ, a population standard deviation σ, a population proportion p, and so on.

(b) **Random sample**: A random sample from an _____ population is a sample selected such that each element selected comes from the _____ population and each element is selected _____.

(c) **Sampling without replacement**: Once an element has been included in the sample, it is removed from the population and cannot be selected a second time.

(d) **Sampling with replacement**: Once an element has been included in the sample, it is returned to the population. A previously selected element can be selected again and therefore may appear in the sample more than once.

(e) **Simple random sample**: A simple random sample of size n from a _____ population of size N is a sample selected such that each possible sample of size n has the _____ of being selected.

(f) **Sample statistic**: A sample _____, such as a sample mean $\bar{x}$, a sample standard deviation s, a sample proportion $\bar{p}$, and so on. The _____ of the sample statistic is used to _____ the value of the corresponding population parameter.

2. Example The EAI problem.

(a) To estimate the value of a population parameter (e.g., the population mean μ and the population standard deviation σ for the annual salary of EAI managers), we compute a corresponding characteristic of the sample, referred to as a _____.

(b) (Table 7.2) A simple random sample of 30 managers and the corresponding data on annual salary _____ and management training program participation _____.

TABLE 7.2 Annual Salary and Training Program Status for a Simple Random Sample of 30 EAI Managers

Annual Salary ($)	Management Training Program	Annual Salary ($)	Management Training Program
$x_1 = 69{,}094.30$	Yes	$x_{16} = 71{,}766.00$	Yes
$x_2 = 73{,}263.90$	Yes	$x_{17} = 72{,}541.30$	No
$x_3 = 69{,}643.50$	Yes	$x_{18} = 64{,}980.00$	Yes
$x_4 = 69{,}894.90$	Yes	$x_{19} = 71{,}932.60$	Yes
$x_5 = 67{,}621.60$	No	$x_{20} = 72{,}973.00$	Yes
$x_6 = 75{,}924.00$	Yes	$x_{21} = 65{,}120.90$	Yes
$x_7 = 69{,}092.30$	Yes	$x_{22} = 71{,}753.00$	Yes
$x_8 = 71{,}404.40$	Yes	$x_{23} = 74{,}391.80$	No
$x_9 = 70{,}957.70$	Yes	$x_{24} = 70{,}164.20$	No
$x_{10} = 75{,}109.70$	Yes	$x_{25} = 72{,}973.60$	No
$x_{11} = 65{,}922.60$	Yes	$x_{26} = 70{,}241.30$	No
$x_{12} = 77{,}268.40$	No	$x_{27} = 72{,}793.90$	No
$x_{13} = 75{,}688.80$	Yes	$x_{28} = 70{,}979.40$	Yes
$x_{14} = 71{,}564.70$	No	$x_{29} = 75{,}860.90$	Yes
$x_{15} = 76{,}188.20$	No	$x_{30} = 77{,}309.10$	No

(c) We use the data in Table 7.2 to calculate the corresponding sample statistics: the sample mean $\bar{x}$ and the sample standard deviation s:

$$\bar{x} = \underline{\qquad\qquad\qquad\qquad}\,,$$

$$s = \underline{\qquad\qquad\qquad\qquad\qquad\qquad}\,,$$

(d) The corresponding sample proportion $\bar{p}$, who completed the management training program:

$$\bar{p} = \underline{\qquad\qquad}\,.$$

3. **Point estimation**: the preceding computations, we perform the statistical procedure called point estimation.

4. We refer to the sample mean $\bar{x}$ as the _____ of the population mean μ, the sample standard deviation s as the point estimator of the population standard deviation σ, and the sample proportion $\bar{p}$ as the point estimator of the population proportion p.

5. **Point estimator**: The _____, such as $\bar{x}$, s, or $\bar{p}$, that provides the _____ of the population parameter.

6. **Point estimate**: The _____ obtained for $\bar{x}$, s, or $\bar{p}$ is called the point estimate.

7. (Table 7.3) The point estimates differ somewhat from the corresponding population parameters. This difference is to be expected because a sample, and not a _____ of the entire population, is being used to develop the point estimates.

TABLE 7.3 Summary of Point Estimates Obtained from a Simple Random Sample of 30 EAI Managers

Population Parameter	Parameter Value	Point Estimator	Point Estimate
μ = Population mean annual salary	$71,800	$\bar{x}$ = Sample mean annual salary	$71,814
σ = Population standard deviation for annual salary	$4000	s = Sample standard deviation for annual salary	$3348
p = Population proportion having completed the management training program	.60	$\bar{p}$ = Sample proportion having completed the management training program	.63

Practical Advice

1. Point estimation is a form of _____. We use a sample statistic to make an inference about a population parameter.

2. When making inferences about a population based on a sample, it is important to have a close correspondence between the _____ population and the _____ population.

3. In some cases, it is not as easy to obtain a close correspondence between the sampled and target populations.

4. Example Consider the case of an amusement park selecting a sample of its customers to learn about characteristics such as age and time spent at the park.

(a) Suppose all the sample elements were selected on a day when park attendance was restricted to employees of a single company.

(b) The sampled population would be composed of employees of that company and members of their families.

(c) If the target population we wanted to make inferences about were typical park customers over a typical summer, then we might encounter a significant difference between the sampled population and the target population.

(d) In such a case, we would question the validity of the point estimates being made.

(e) Park management would be in the best position to know whether a sample taken on a particular day was likely to be representative of the target population.

5. In summary, whenever a sample is used to make inferences about a population, we should make sure that the study is designed so that the sampled population and the target population are in _____.

6. Good judgment is a necessary _____ of sound statistical practice.

7.4 Introduction to Sampling Distributions

1. (Table 7.4) For the simple random sample of 30 EAI managers, the point estimate of μ is $\bar{x} = \$71,814$ and the point estimate of p is $\bar{p} = 0.63$. Suppose we _____ the process of selecting a simple random sample of 30 EAI managers over and over again (e.g., 500 simple random samples), each time computing the values of $\bar{x}$ and $\bar{p}$.

TABLE 7.4 Values of $\bar{x}$ and $\bar{p}$ from 500 Simple Random Samples of 30 EAI Managers

Sample Number	Sample Mean ($\bar{x}$)	Sample Proportion ($\bar{p}$)
1	71,814	.63
2	72,670	.70
3	71,780	.67
4	71,588	.53
.	.	.
.	.	.
.	.	.
500	71,752	.50

2. (Table 7.5) The frequency and relative frequency distributions for the 500 $\bar{x}$ values.

TABLE 7.5 Frequency and Relative Frequency Distributions of $\bar{x}$ from 500 Simple Random Samples of 30 EAI Managers

Mean Annual Salary ($)	Frequency	Relative Frequency
69,500.00–69,999.99	2	.004
70,000.00–70,499.99	16	.032
70,500.00–70,999.99	52	.104
71,000.00–71,499.99	101	.202
71,500.00–71,999.99	133	.266
72,000.00–72,499.99	110	.220
72,500.00–72,999.99	54	.108
73,000.00–73,499.99	26	.052
73,500.00–73,999.99	6	.012
Totals	500	1.000

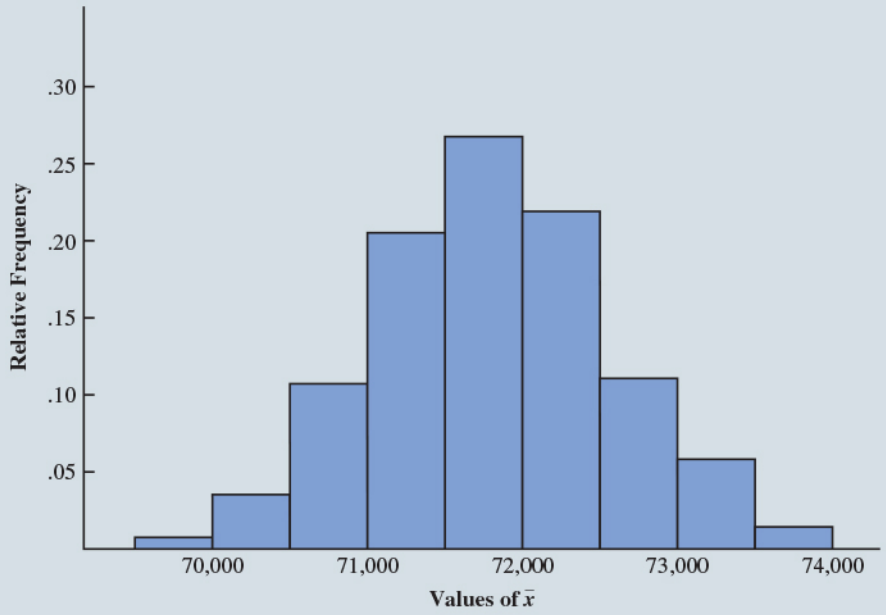
3. Recall: a random variable is a _____ of the outcome of an experiment. If we consider the process of selecting a simple random sample as an experiment, the sample mean $\bar{x}$ is a random variable. $\bar{x}$ has a _____ or expected value, a _____, and a _____.

4. **The Sampling Distribution**: The various possible values of $\bar{x}$ are the result of different simple random samples, the probability distribution of $\bar{x}$ is called the _____ of $\bar{x}$.

5. (Figure 7.1) The histogram of 500 $\bar{x}$ values gives an approximation of this sampling distribution. From the approximation we observe the _____ appearance

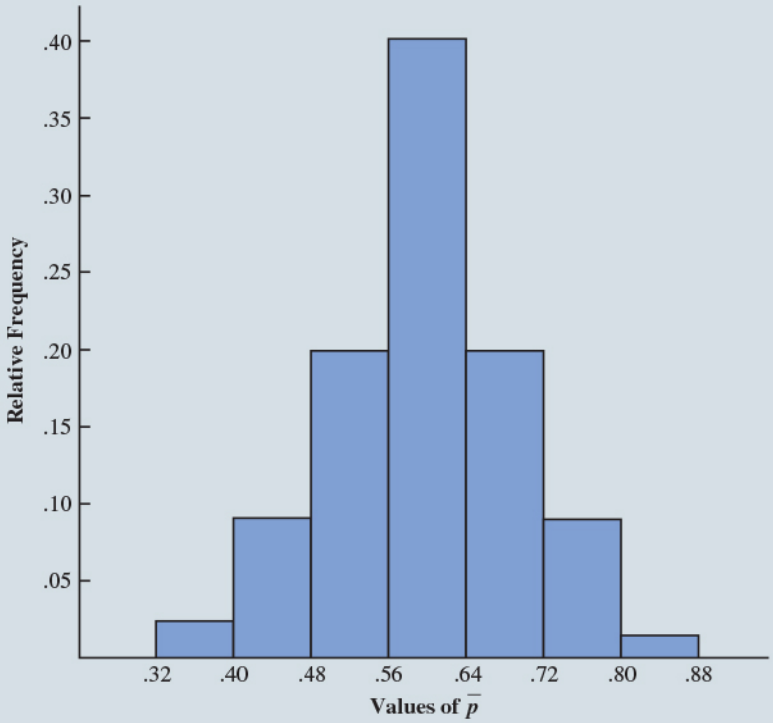
of the distribution. The largest concentration of the $\bar{x}$ values and the mean of the 500 $\bar{x}$ values are near the _____ $\mu = \$71,800$.

FIGURE 7.1 Relative Frequency Histogram of $\bar{x}$ Values from 500 Simple Random Samples of Size 30 Each



6. (Figure 7.2) The 500 values of the sample proportion $\bar{p}$ are summarized by the _____. The relative frequency histogram of the 500 sample values provides a general idea of the appearance of the sampling distribution of $\bar{p}$.

FIGURE 7.2 Relative Frequency Histogram of $\bar{p}$ Values from 500 Simple Random Samples of Size 30 Each



7. In practice, we select only one simple random sample from the population. We repeated the sampling process many times and that the different samples generate a variety of values for the _____ $\bar{x}$ and $\bar{p}$.

8. The _____ of any particular sample statistic is called the _____ of the statistic.

7.5 Sampling Distribution of $\bar{X}$

1. **Sampling Distribution of $\bar{x}$**: The sampling distribution of $\bar{x}$ is the _____ distribution of _____ of the sample mean $\bar{x}$.

2. The sampling distribution of $\bar{x}$ has an expected value or mean, a standard deviation, and a characteristic shape or form.

Expected Value of $\bar{x}$

1. Different simple random samples result in a variety of values for the sample mean $\bar{x}$.

2. Because many different values of the random variable $\bar{x}$ are possible, we are often interested in the _____ of all possible values of $\bar{x}$ that can be generated by the various simple random samples.

3. The mean of the $\bar{x}$ random variable is the _____ of $\bar{x}$.

4. Let $E(\bar{x})$ represent the expected value of $\bar{x}$ and μ represent the mean of the population from which we are selecting a simple random sample. It can be shown that with simple random sampling, the expected value of $\bar{x}$ and the population mean are equal. (_____).

5. When the expected value of a point estimator equals the population parameter, we say the point estimator is _____. The sample mean $\bar{x}$ is an unbiased estimator of the population mean μ.

Standard Deviation of $\bar{x}$

1. Notations:

   (a) _____: the standard deviation of $\bar{x}$. _____: the standard deviation of the population.

   (b) n: the sample size. N: the population size.

2. The two formulas for the standard deviation of $\bar{x}$:

   (a) Finite Population: _____.

   (b) Infinite Population: _____.

3. **The finite population correction factor**: $\sqrt{(N-n)/(N-1)}$.

4. In many practical sampling situations, we find that the population involved, although finite, is _____ whereas the sample size is relatively _____ In such cases the finite population correction factor $\sqrt{(N-n)/(N-1)}$ is close to _____. As a result, the difference between the values of the standard deviation of $\bar{x}$ for the finite and infinite population cases becomes _____. Then, _____ becomes a good approximation to the standard deviation of $\bar{x}$ even though the population is _____.

5. *General guideline*, or rule of thumb, for computing the standard deviation of $\bar{x}$: Use _____ whenever

   (a) The population is _____; or

   (b) The population is finite and the sample size is less than or equal to _____ of the population size; that is, _____.

6. In cases where _____, the finite population version of formula should be used in the computation of $\sigma_{\bar{x}}$. Unless otherwise noted, throughout the text we will assume that the population size is "large," n/N 0.05, and expression _____ can be used to compute $\sigma_{\bar{x}}$.

7. We refer to the standard deviation of $\bar{x}$, $\sigma_{\bar{x}}$, as the _____.

8. In general, the term standard error refers to the standard deviation of a _____.

Question ..... (p335)

(**EAI managers Problem**) In Section 7.1 we saw that the standard deviation of annual salary for the population of 2500 EAI managers is $\sigma = 4000$. Compute the standard error.

sol: The population is finite, with $N = 2500$. However, with a sample size of 30, we have $n/N = 30/2500 = 0.012$. Because the sample size is less than 5% of the population size, we can ignore the finite population correction factor and use equation (7.3) to compute the standard error:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{4000}{\sqrt{30}} = 730.3$$

Form of the Sampling Distribution of $\bar{X}$

1. Consider two cases to determine the form or shape of $\bar{x}$: (1) The population has a normal distribution; and (2) the population does not have a normal distribution.

2. *Population has a Normal Distribution*: When the population has a normal distribution, the sampling distribution of $\bar{x}$ is _____ distributed for _____ sample size.

3. *Population does not have a Normal Distribution*: When the population from which we are selecting a random sample does not have a normal distribution, the _____ _____ is helpful in identifying the shape of the sampling distribution of $\bar{x}$.

4. **Central Limit Theorem (CLT)**: In selecting random samples of size n from a population, the sampling distribution of the _____ can be approximated by a normal distribution as the sample size becomes large.

5. (Figure 7.3) Illustration of the Central Limit Theorem for Three Populations

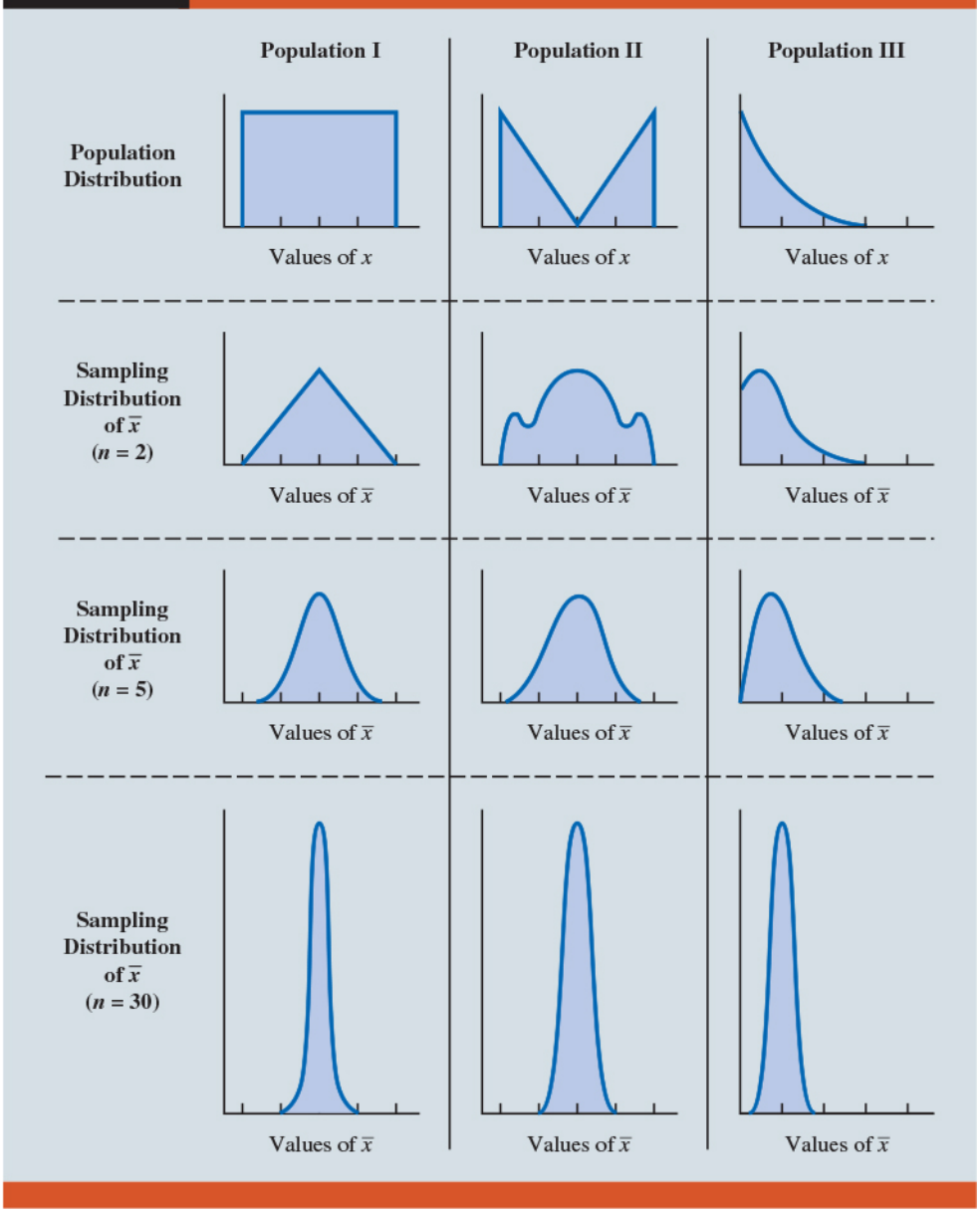
   (a) Population I follows a _____ distribution. Population II is called the _____ distribution. It is symmetric, but the more likely values fall in the tails of the distribution. Population III is shaped like the _____ distribution; it is skewed to the right.

   (b) When the sample size is $n = 2$, the shape of each sampling distribution is _____ from the shape of the corresponding population distribution.

   (c) For samples of size $n = 5$, the shapes of the sampling distributions for populations I and II begin to look similar to the shape of a _____ distribution. Even though the shape of the sampling distribution for population III begins to look similar to the shape of a normal distribution, some _____ to the right is still present.

   (d) For samples of size 30, the shapes of each of the three sampling distributions are approximately _____.

FIGURE 7.3 Illustration of the Central Limit Theorem for Three Populations



6. General statistical practice is to assume that, for most applications, the sampling distribution of $\bar{x}$ can be approximated by a normal distribution whenever the sample is _____.

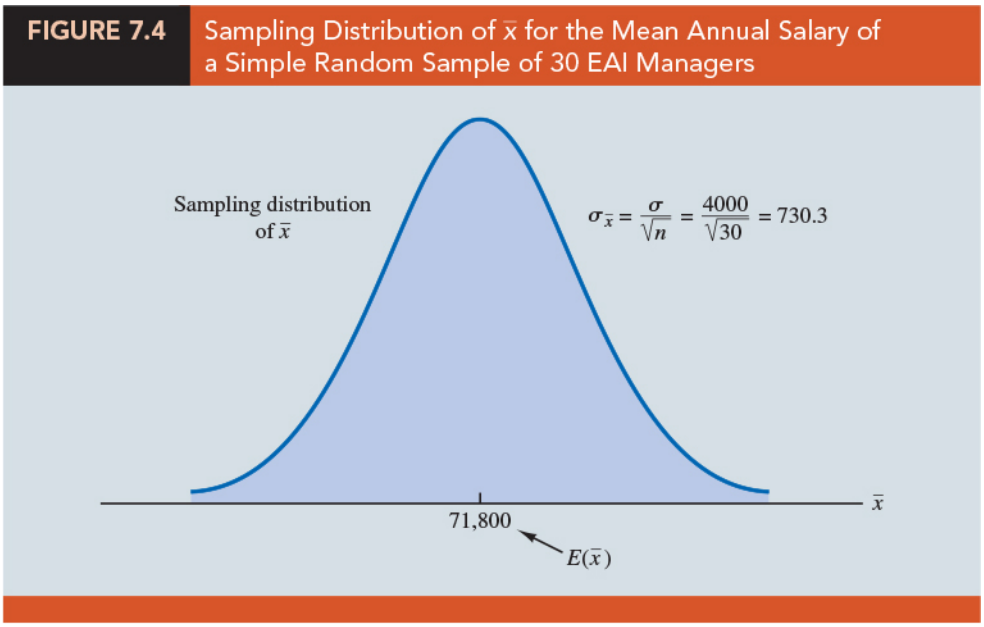
7. In cases where the population is _____ or _____ are present, samples of size _____ may be needed. Finally, if the population is _____, the sample size needed for a normal approximation often depends on the population _____.

Sampling Distribution of $\bar{x}$ for the EAI Problem

1. The expected value and the standard deviation of the annual salary of EAI managers are $E(x) = \$71,800$ and $\sigma_{\bar{x}} = 730.3$. We do not have any information about the population distribution

2. If the population has a normal distribution, the sampling distribution of $\bar{x}$ is normally distributed. If the population does not have a normal distribution, the simple random sample of 30 managers and the central limit theorem enable us to conclude that the sampling distribution of $\bar{x}$ can be approximated by a normal distribution.

3. (Figure 7.4) In either case, we are comfortable proceeding with the conclusion that the sampling distribution of $\bar{x}$ can be described by the normal distribution.



Practical Value of the Sampling Distribution of $\bar{x}$

1. Whenever a simple random sample is selected, we cannot expect the sample mean to _____ the population mean.

2. The practical reason we are interested in the sampling distribution of $\bar{x}$ is that it can be used to provide _____ information about the _____ between the sample mean and the population mean.

3. Example EAI problem:

   (a) Suppose the personnel director believes the sample mean will be an acceptable estimate of the population mean if the sample mean is within \$500 of the population mean. However, it is not possible to guarantee that the sample mean will be _____ \$500 of the population mean.

   (b) Indeed, Table 7.5 and Figure 7.1 show that some of the 500 sample means differed by more than \$2000 from the population mean. So we must think of the personnel director′ s request in _____.

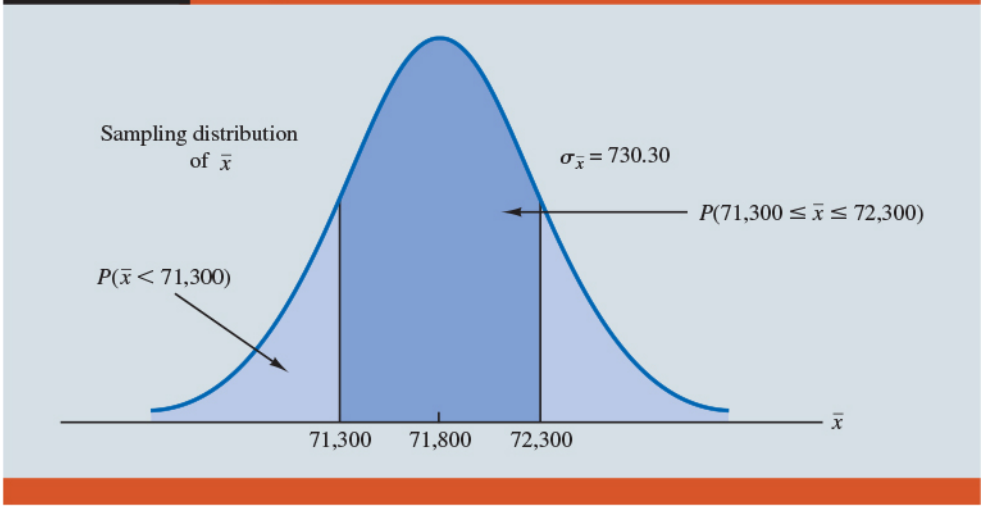
   (c) That is, the personnel director is concerned with the following question: What is the probability that the sample mean computed using a simple random sample of 30 EAI managers will be within \$500 of the population mean?

   (d) ✎ Question ..... (p338)

   Consider the EAI problem, the mean and the standard deviation of annual salary for the population of $N = 2500$ EAI managers is $\mu = \$71,800$ and $\sigma = 4000$. With a sample size of $n = 30$, the personnel director wants to know the probability that $\bar{x}$ is between \$71,300 and \$72,300.

   sol: (Figure 7.5) Because the sampling distribution is normally distributed, with mean \$71,800 and standard error of the mean 730.3, we can use the standard normal probability table to find the area or probability.

FIGURE 7.5 Probability of a Sample Mean Being within \$500 of the Population Mean for a Simple Random Sample of 30 EAI Managers



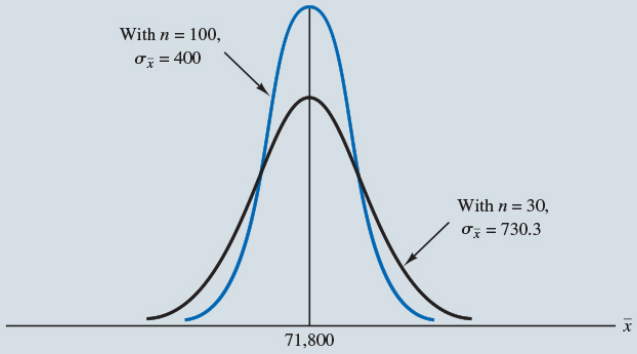
(e) The preceding computations show that a simple random sample of 30 EAI managers has a _____ probability of providing a sample mean $\bar{x}$ that is within \$500 of the population mean.

(f) Thus, there is a _____ probability that the difference between $\bar{x}$ and $\mu = \$71,800$ will be more than \$500. In other words, a simple random sample of 30 EAI managers has roughly a _____ of providing a sample mean within the allowable \$500. Perhaps a _____ should be considered.

Relationship Between the Sample Size and the Sampling Distribution of $\bar{X}$

1. (Figure 7.6) The sampling distributions of $\bar{x}$ with $n = 30$ and $n = 100$. Because the sampling distribution with $n = 100$ has a smaller standard error, the values of $\bar{x}$ have _____ and tend to be _____ to the population mean than the values of $\bar{x}$ with $n = 30$.

FIGURE 7.6 A Comparison of the Sampling Distributions of $\bar{x}$ for Simple Random Samples of $n = 30$ and $n = 100$ EAI Managers

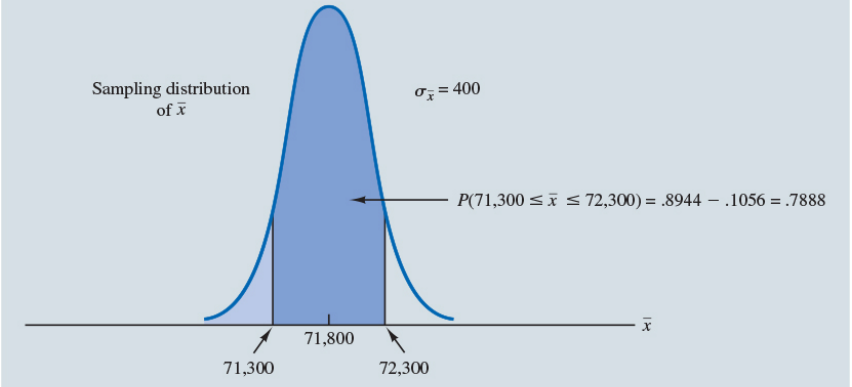


2. ✎ Question ..... (p339)

Consider the EAI problem, With a sample size of $n = 100$ EAI managers, compute the probability that $\bar{x}$ is within \$500 of the population mean $\mu = \$71,800$.

sol:

FIGURE 7.7 Probability of a Sample Mean being within \$500 of the Population Mean for a Simple Random Sample of 100 EAI Managers



3. Thus, by increasing the sample size from 30 to 100 EAI managers, we increase the probability of obtaining a sample mean within \$500 of the population mean from 0.5034 to 0.7888. The important point in this discussion is that as the sample size is increased, the _____ decreases.

4. As a result, the _____ provides a higher probability that the sample mean is within a specified distance of the population mean.

7.6 Sampling Distribution of $\bar{p}$

1. The sample proportion $\bar{p}$ is the point estimator of the population proportion p:

   where x is the number of elements in the sample that _____ the characteristic of interest and n is the sample size.

2. The sample proportion $\bar{p}$ is a _____ and its probability distribution is called the _____ of $\bar{p}$.

3. **Sampling Distribution of $\bar{p}$** The sampling distribution of $\bar{p}$ is the _____ of all possible values of the sample proportion $\bar{p}$.

Expected Value of $\bar{p}$

1. The expected value of $\bar{p}$, the mean of _____ of $\bar{p}$, is equal to the population proportion p:

   _____.

2. Because $E(\bar{p}) = p$, $\bar{p}$ is an _____ estimator of p.

3. Example (Section 7.1) we noted that $p = 0.60$ for the EAI population, where p is the proportion of the population of managers who participated in the company's management training program. Thus, the expected value of $\bar{p}$ for the EAI sampling problem is 0.60.

Standard Deviation of $\bar{p}$

1. The standard deviation of $\bar{p}$ depends on whether the population is finite or infinite:
   (a) Finite Population: _____.
   (b) Infinite Population: _____.
2. The difference between the expressions for the finite population and the infinite population becomes _____ if the size of the finite population is large in _____ to the sample size.
3. Rule of thumb:
   (a) if the population is finite with _____, use _____.
   (b) if the population is finite with _____, the finite population _____ _____ should be used.
   (c) Unless specifically noted, throughout the text we will assume that the population size is large in relation to the sample size and thus the finite population correction factor is _____.
4. We stated that in general the term standard error refers to the standard deviation of a _____. Thus, for proportions we use standard error of the proportion to refer to the _____.
5. Example (The EAI example) With _____, we can ignore the finite population correction factor when we compute the standard error of the proportion. For the simple random sample of 30 managers, the standard error of the proportion associated with simple random samples of 30 EAI managers:

$$\sigma_{\bar{p}} = _____.$$

Form of the Sampling Distribution of $\bar{p}$

1. For a simple random sample from a large population, the value of x is a _____ _____ indicating the _____ of elements in the sample with the characteristic of interest.

2. Because n is a constant, the probability of x/n is the same as the _____ of x, which means that the sampling distribution of $\bar{p}$ is also a _____ probability distribution and that the probability for each value of _____ is the same as the probability of _____.

3. (Chapter 6) we showed that a binomial distribution can be approximated by a normal distribution whenever the sample size is large enough to satisfy two conditions: (1) _____ and _____.

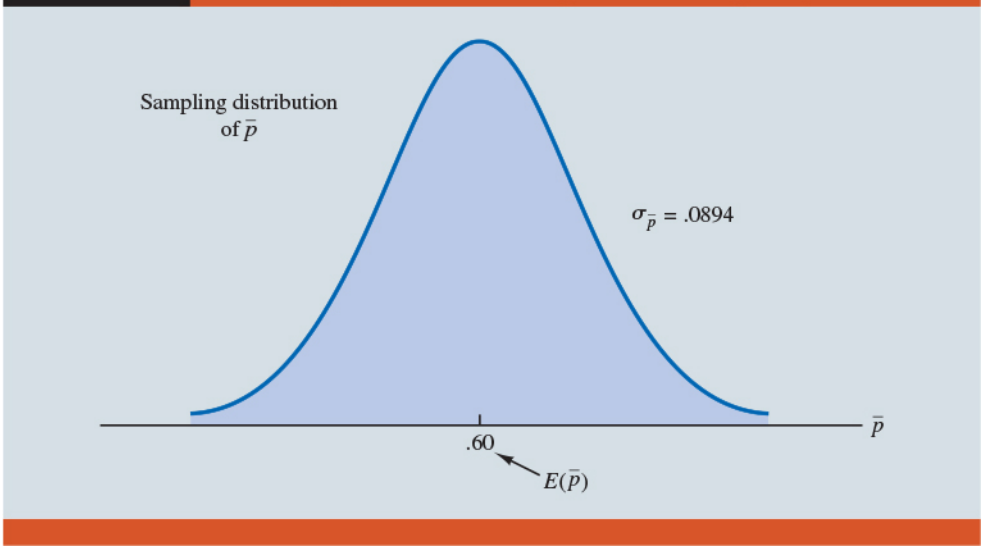
4. Assuming these two conditions are satisfied, the probability distribution of x in the sample proportion, _____, can be approximated by a normal distribution. And because _____ is a constant, the sampling distribution of _____ can also be approximated by a normal distribution.

5. The sampling distribution of $\bar{p}$ can be approximated by a normal distribution whenever _____ and _____.

6. In practical applications, when an estimate of a population proportion is desired, we find that sample sizes are almost always _____ to permit the use of a normal approximation for the sampling distribution of $\bar{p}$.

7. Example (Figure 7.8) (The EAI example) Recall that we know that the population proportion of managers who participated in the training program is $p = 0.60$. With a simple random sample of size 30, we have _____ and _____. Thus, the sampling distribution of $\bar{p}$ can be approximated by a normal distribution.

FIGURE 7.8 Sampling Distribution of $\bar{p}$ for the Proportion of EAI Managers who Participated in the Management Training Program



Practical Value of the Sampling Distribution of $\bar{p}$

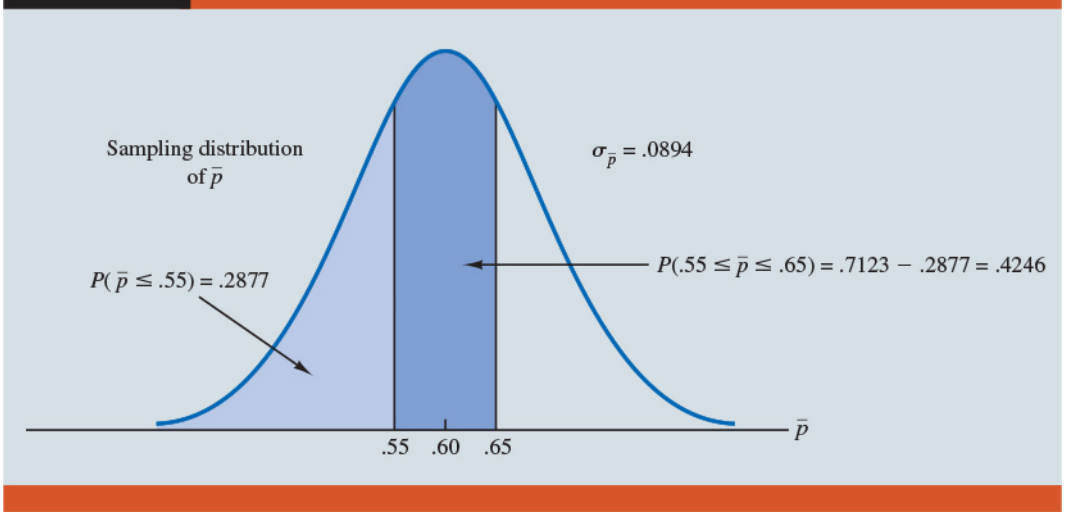
1. The _____ of the sampling distribution of $\bar{p}$ is that it can be used to provide probability information about the _____ between the sample proportion and the population proportion.

Question ..... (p346)

Consider the EAI problem, suppose that the personnel director wants to know the probability of obtaining a value of $\bar{p}$ that is within 0.05 of the population proportion of EAI managers who participated in the training program. That is, what is the probability of obtaining a sample with a sample proportion $\bar{p}$ between 0.55 and 0.65? Consider two cases: $n = 30$ and $n = 100$ using a normal approximation.

sol:

FIGURE 7.9 Probability of Obtaining $\bar{p}$ Between .55 and .65



7.7 Properties of Point Estimators

1. Before using a sample statistic as a point estimator, statisticians check to see whether the _____ demonstrates certain properties associated with _____ point estimators.
2. We discuss three properties of good point estimators: _____, _____, and _____.
3. General notation:
   (a) _____: the population parameter of interest,
   (b) _____: the sample statistic or point estimator of θ.

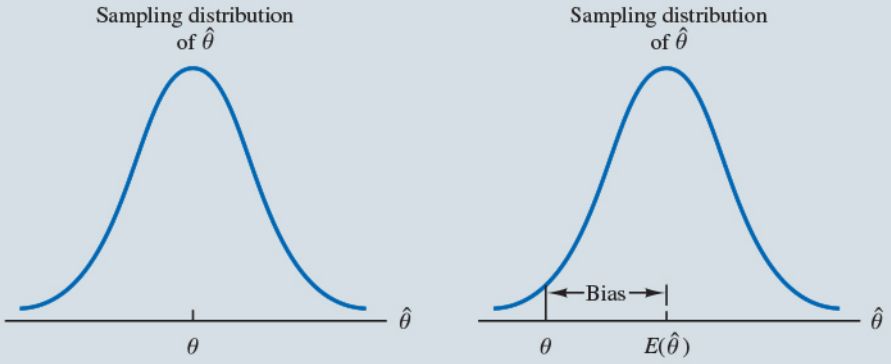
Unbiased

1. If the expected value of the sample statistic is equal to the population parameter being estimated, the sample statistic is said to be an _____ of the population parameter.
2. **Unbiased estimator**: The sample statistic $\hat{\theta}$ is an unbiased estimator of the population parameter θ if

   _____,

   where $E(\hat{\theta})$ is the _____ of the sample statistic $\hat{\theta}$.
3. (Figure 7.10) the cases of unbiased and biased point estimators.
   (a) The unbiased estimator: the _____ is equal to the value of the population parameter.
   (b) A biased estimator: the mean of the sampling distribution is less than or greater than the value of the population parameter. The amount of the _____ is shown.

FIGURE 7.10 Examples of Unbiased and Biased Point Estimators



4. We stated that _____ and _____. Both $\bar{x}$ and $\bar{p}$ are unbiased estimators of their corresponding population parameters μ and p.

5. It can be shown that _____, the sample variance s^2 is an unbiased estimator of the population variance σ^2.

6. The reason for using _____ rather than n in sample variance (and the sample standard deviation) is to make the sample variance an _____ estimator of the population variance.

Efficiency

1. Assume that a simple random sample of n elements can be used to provide _____ _____ of the same population parameter. In this situation, we would prefer to use the point estimator with the _____, because it tends to provide estimates closer to the population parameter.

2. The point estimator with the smaller standard error is said to have greater _____ than the other.

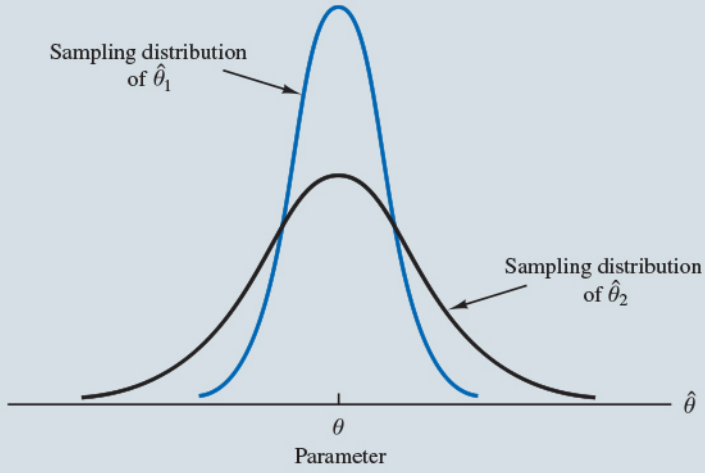
3. (Figure 7.11) the sampling distributions of two unbiased point estimators, $\hat{\theta}_1$ and $\hat{\theta}_2$.

(a) The standard error of $\hat{\theta}_1$ is less than the standard error of $\hat{\theta}_2$;

(b) Values of $\hat{\theta}_1$ have a _____ of being close to the parameter θ than do values of $\hat{\theta}_2$.

(c) $\hat{\theta}_1$ is relatively more efficient than $\hat{\theta}_2$ and is the preferred point estimator.

FIGURE 7.11 Sampling Distributions of Two Unbiased Point Estimators



Consistency

1. A point estimator is _____ if the values of the point estimator tend to become closer to the population parameter as the _____ becomes _____.

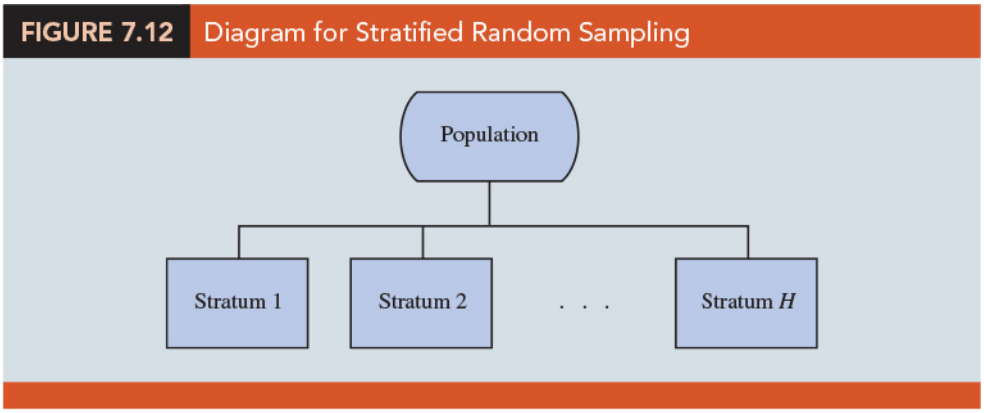
2. A _____ sample size tends to provide a better point estimate than a _____ sample size.

3. The standard error of $\bar{x}$, _____, is related to the sample size such that larger sample sizes provide smaller values for $\sigma_{\bar{x}}$, we conclude the sample mean $\bar{x}$ is a _____ estimator of the population mean μ.

4. The sample proportion $\bar{p}$ is a _____ estimator of the population proportion p.

7.8 Other Sampling Methods[1]

Stratified Random Sampling



1. (Figure 7.12) In stratified random sampling, the elements in the population are first divided into _____ called _____, such that each element in the population belongs to one and only one stratum. A _____ is taken from each stratum.

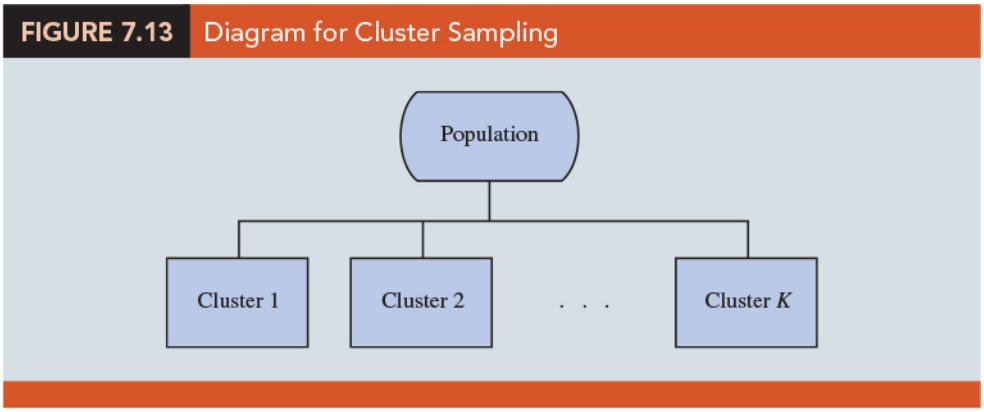
2. The basis for forming the strata, such as department, location, age, industry type, and so on, is at the _____ of the designer of the sample.

3. However, the best results are obtained when the elements _____ each stratum are as _____ as possible.

4. Formulas are available for combining the results for the individual stratum samples into one estimate of the population parameter of interest.

[1] A more in-depth treatment is provided in Chapter 22.

Cluster Sampling



1. (Figure 7.13) In cluster sampling, the elements in the population are first divided into separate groups called _____.

2. Each element of the population belongs to one and only one cluster. A simple random sample of the _____ is then taken. All elements within each sampled cluster form the sample.

3. Cluster sampling tends to provide the best results when the elements within the clusters are _____.

4. In the ideal case, each cluster is a representative _____ of the entire population. The value of cluster sampling depends on how representative each cluster is of the entire population.

5. If all clusters are _____ in this regard, sampling a small number of clusters will provide good estimates of the population parameters.

6. One of the primary applications of cluster sampling is area sampling, where clusters are _____ or other well-defined areas.

Systematic Sampling

1. Systematic sampling is a type of probability sampling method in which sample members from a _____ are selected according to a _____ but with a fixed, periodic interval.

2. Example If a sample size of 50 is desired from a population containing 5000 elements, we will sample one element for every $5000/50 = 100$ elements in the population.

3. A systematic sample for this case involves selecting randomly one of the _____ from the population list.

4. Other sample elements are identified by starting with the first sampled element and then selecting every $100th$ element that follows in the population list.

5. In effect, the sample of 50 is identified by _____ through the population and identifying every $100th$ element after the first randomly selected element.

6. Because the first element selected is a random choice, a systematic sample is usually assumed to have the properties of a simple random sample.

7. This assumption is especially applicable when the list of elements in the population is a _____ of the elements.

Convenience Sampling

1. **Probability sampling techniques**: elements selected from the population have a known _____ of being included in the sample.

2. Convenience sampling is a _____ sampling technique. The sample is identified primarily by _____. Elements are included in the sample without prespecified or known probabilities of being selected.

3. Example A professor conducting research at a university may use student volunteers to constitute a sample simply because they are readily available and will participate as subjects for little or no cost.

4. Convenience samples have the advantage of relatively easy sample selection and data collection; however, it is impossible to evaluate the "goodness" of the sample in terms of its _____ of the population.

5. A convenience sample may provide good results or it may not; no statistically justified procedure allows a probability analysis and inference about the quality of the sample results.

Judgment Sampling

1. Judgment sampling is a _____ sampling technique.
2. In this approach, the person most knowledgeable on the subject of the study selects elements of the population that he or she feels are _____ of the population.
3. The quality of the sample results depends on the _____ of the person selecting the sample.

7.9 Big Data and Standard Errors of Sampling Distributions

1. The purpose of statistical inference is to use _____ data to quickly and inexpensively gain insight into some characteristic of a _____.
2. There are two general reasons a sample may fail to be representative of the population of interest: _____ and _____.

Sampling Error

1. One reason a sample may fail to represent the population from which it has been taken is _____, or _____ of the sample from the population that results from random sampling.
2. When collecting a sample randomly, the data in the sample cannot be expected to be perfectly _____ of the population from which it has been taken.

3. Sampling error is _____ when collecting a random sample; this is a risk we must accept when we chose to collect a random sample rather than incur the costs associated with taking a census of the population.

4. The _____ of the sampling distributions of the sample mean $\bar{x}$ and the sample proportion of $\bar{p}$ reflect the potential for sampling error when using sample data to estimate the population mean μ and the population proportion p, respectively.

5. For an extremely _____ there may be little potential for sampling error.

Nonsampling Error

1. We can not conclude that an extremely large sample will always provide reliable information about the population of interest.

2. Deviations of the sample from the population that occur for reasons other than random sampling are referred to as _____.

3. Nonsampling error can occur for a variety of reasons.

   (a) **Coverage error**: Nonsampling error that results when the research objective and the population from which the sample is to be drawn are not aligned.

   (b) **Nonresponse error**: Nonsampling error that results when some segments of the population are either more or less likely to respond to the survey mechanism.

   (c) **Measurement error**: Nonsampling error that results from the incorrect measurement of the population characteristic of interest.

4. Understanding these limitations of sampling will enable us to be more _____ and _____ when interpreting sample data and using sample data to draw conclusions about the target population.

Understanding What Big Data Is

1. The processes that generate big data can be described by four attributes or dimensions that are referred to as the four V's:

(a) **Volume**: the amount of data generated

(b) **Variety**: the diversity in types and structures of data generated

(c) **Veracity**: the reliability of the data generated

(d) **Velocity**: the speed at which the data are generated

2. **Tall data**: A data set that has so _____ that traditional statistical inference has little meaning.

3. **Wide data**: A data set that has so _____ that simultaneous consideration of all variables is infeasible.

4. Statistics are useful tools for understanding the information embedded in a big data set, but we must be careful when using statistics to analyze big data.

5. It is important that we understand the limitations of statistics when applied to big data and we temper our interpretations accordingly.

Implications of Big Data for Sampling Error

1. (Table 7.7) the standard error of the sampling distribution of the sample mean time spent by individual customers when they visit PDT's website decreases as the sample size increases.

2. (Table 7.8) the standard error of the sampling distribution of the proportion of the sample that clicked on any of the ads featured on PenningtonDailyTimes.com decreases as the sample size increases.

TABLE 7.7 Standard Error of the Sample Mean $\bar{x}$ When $s = 20$ at Various Sample Sizes n

Sample Size n	Standard Error $s_{\bar{x}} = s/\sqrt{n}$
10	6.32456
100	2.00000
1,000	.63246
10,000	.20000
100,000	.06325
1,000,000	.02000
10,000,000	.00632
100,000,000	.00200
1,000,000,000	.00063

TABLE 7.8 Standard Error of the Sample Proportion $\bar{p}$ When $p = .51$ at Various Sample Sizes n

Sample Size n	Standard Error $\sigma_{\bar{p}} = \sqrt{\bar{p}(1-\bar{p})/n}$
10	.15808
100	.04999
1,000	.01581
10,000	.00500
100,000	.00158
1,000,000	.00050
10,000,000	.00016
100,000,000	.00005
1,000,000,000	.00002

😍 EXERCISES

7.2 : 2, 10

7.3 : 15, 17

7.5 : 18, 24, 25, 30

7.6 : 34, 35, 39

7.9 : 42, 44, 47, 49

SUP : 52, 57, 62, 65

名言佳句

“我哭泣是因為我沒有鞋子，直到我遇到了一個人 — 他沒有腳”
“I cried because I had no shoes until I met a man who had no feet”
— *Helen Keller (June 27, 1880 – June 1, 1968)*

統計小故事：1936 年美國大選與蓋洛普民調的啟示

二十世紀三　年代，美國正從經濟大蕭條中逐步復甦。1936 年的總統大選，是現任總統富蘭克林．羅斯福（Franklin D. Roosevelt）尋求連任的重要時刻。當時的媒體與民意調查機構，無不想預測選舉的結果，於是展開了一場規模空前的民調實驗。

當時最具影響力的雜誌之一《文摘》（Literary Digest），決定進行一場龐大的民意調查。他們寄出了一千萬份問卷，並收回了大約兩百三十萬份回覆，成為史上最大規模的調查之一。如此龐大的樣本數，理應帶來極高的準確度。然而，調查結果卻讓人大跌眼鏡——他們預測共和黨候選人阿爾夫．蘭登（Alf Landon）將以壓倒性優勢擊敗羅斯福。

然而，選舉的真正結果卻是羅斯福大勝，成功連任。這個錯誤的預測不僅讓《文摘》顏面盡失，也成為統計學史上的經典案例。問題究竟出在哪裡？

答案在於抽樣方法。雖然《文摘》的樣本數驚人，但他們的名單來自電話簿與汽車登記名冊。要知道，1930 年代的美國，電話與汽車並非普及品，通常只有較富裕的中上階層家庭才擁有。這樣的樣本偏向於支持共和黨的選民，而忽略了廣大工薪階層的聲音。換句話說，他們的樣本數雖然龐大，但並不具代表性。這就是典型的抽樣偏差（sampling bias）。

與此同時，一位名不見經傳的年輕學者——喬治．蓋洛普（George Gallup）——進行了自己的調查。他只抽樣了五萬人，樣本數遠小於《文摘》。但他強調樣本的「代表性」，而非單純的數量。他採用隨機抽樣與分層抽樣的方法，確保不同地區、不同階層、不同背景的選民都能被納入樣本。最終，蓋洛普的預測正確地指出羅斯福將會大勝。這場對比讓世人驚嘆：原來調查的準確性，不取決於樣本數的大小，而在於抽樣方法是否科學。

這個故事的意義，遠超越一次選舉。它讓人們深刻理解到，抽樣與抽樣分布的科學性，才是統計推論的核心。一個設計不良的樣本，即便有上百萬筆數據，也只會放大偏差；而一個設計良好的樣本，即便數量較少，也能帶來準確的結論。

在統計學中，這正是「抽樣分布（sampling distribution）」的重要性。抽樣分布告訴我們：當我們從母體中不斷抽樣，樣本統計量（例如樣本平均數）的分布會呈現出一定的規律。雖然每個樣本的平均數會有所不同，但隨著樣本數的增加，這些平均數會集中在母體平均數附近，並遵循可預測的分布形態。這個概念為後續的統計推論——例如信賴區間與假設檢定——奠定了基礎。

回到 1936 年的美國大選，我們可以這樣理解：《文摘》的樣本雖然大，但偏差嚴重，導致樣本分布偏離了真實的母體分布。蓋洛普的樣本雖小，但設計得當，使得樣本分

布能夠合理反映母體特徵。因此，統計學家後來常引用一句話：「一個小而代表性的樣本，勝過一個大而有偏差的樣本。」

在教材呼應上，本章的重點是讓學生理解「抽樣」不僅僅是隨意挑選，而是需要嚴謹的設計。抽樣分布讓我們有信心地說：「雖然我們無法觀察整個母體，但只要抽樣得當，就能用部分推論整體。」這正是統計學之所以強大而實用的原因。

因此，當學生在課堂上計算「樣本平均數的標準誤」時，請記住，他們不僅僅是在做一個數學練習，而是在學習一場歷史教訓：數據量不是一切，方法與設計才是關鍵。蓋洛普的成功與《文摘》的失敗，正是這一章最生動的註腳。

統計學 (一)

Anderson's Statistics for Business & Economics (14/E)

Chapter 8: Interval Estimation

上課時間地點: 二 D56, 商館 260206
授課教師: 吳漢銘 (國立政治大學統計學系副教授)
教學網站: https://hmwu.idv.tw

系級: _____ 學號: _____ 姓名: _____

Overview

1. A point estimator is a _____ used to estimate a _____.
   (a) The sample mean _____ is a point estimator of the population mean _____.
   (b) The sample proportion _____ is a point estimator of the population proportion _____.
2. A point estimator cannot be expected to provide the _____ of the population parameter, an _____ is often computed by adding and subtracting a value, called the _____, to the point estimate.
3. The general form of an interval estimate:

   _____
4. The purpose of an interval estimate is to provide information about _____ the point estimate, provided by the _____, is to the _____ of the population parameter.

5. This chapter shows how to compute _____ of a population mean μ and a population proportion p.

6. The _____ of $\bar{x}$ and $\bar{p}$ play key roles in computing these interval estimates.

8.1 Population Mean: σ Known

1. To develop an interval estimate of a population mean, either the population standard deviation_____ or the sample standard deviation _____ must be used to compute the _____.

2. The cases that the σ is known:

   (a) Large amounts of relevant _____ are available and can be used to estimate the population standard deviation prior to sampling.

   (b) In _____ applications where a _____ is assumed to be operating correctly, or "in control," it is appropriate to treat the population standard deviation as known.

3. Example Lloyd's example

   (a) Each week Lloyd's Department Store selects a simple random sample of 100 customers in order to learn about _____ per shopping trip.

   (b) With x representing the amount spent per shopping trip, the sample mean $\bar{x}$ provides a point estimate of μ, _____ per shopping trip for the population of all Lloyd's customers.

   (c) Based on the historical data, Lloyd's now assumes a known value of _____ for the population standard deviation and the population follows a _____ distribution.

(d) During the most recent week, Lloyd′ s surveyed 100 customers ($n = 100$) and obtained a sample mean of _____.

4. How to compute the margin of error for the sample mean and develop an interval estimate of the population mean.

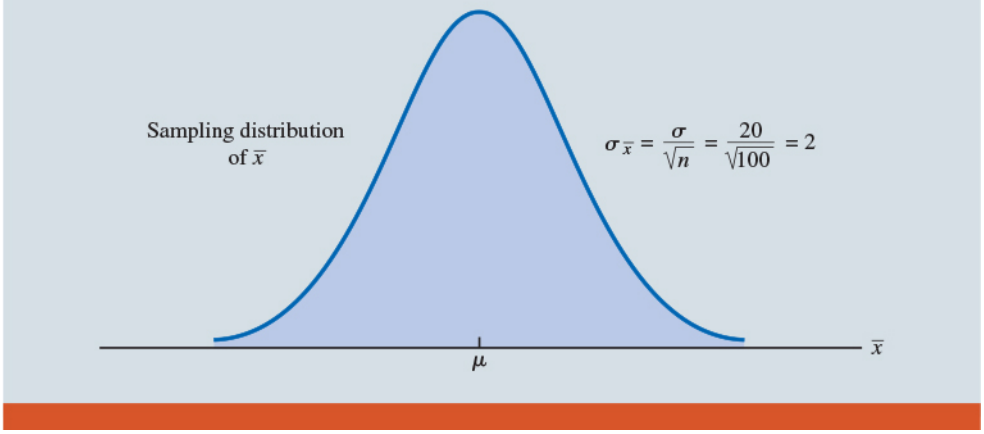
Margin of Error and the Interval Estimate

1. (Chapter 7) The _____ of $\bar{x}$ can be used to compute the probability that $\bar{x}$ will be _____ a given distance of μ.

2. Example Lloyd's example

   (a) (Figure 8.1) The population of amounts spent is normally distributed with a standard deviation of $\sigma = 20$. The sampling distribution of _____ follows a _____ distribution with a standard error of _____.

FIGURE 8.1 Sampling Distribution of the Sample Mean Amount Spent from Simple Random Samples of 100 Customers



   (b) The sampling distribution shows how _____ around the population mean μ,

   (c) The sampling distribution of $\bar{x}$ provides information about the possible _____ between $\bar{x}$ and μ.

3. (The standard normal probability table) The _____ of the values of any normally distributed random variable are _____.

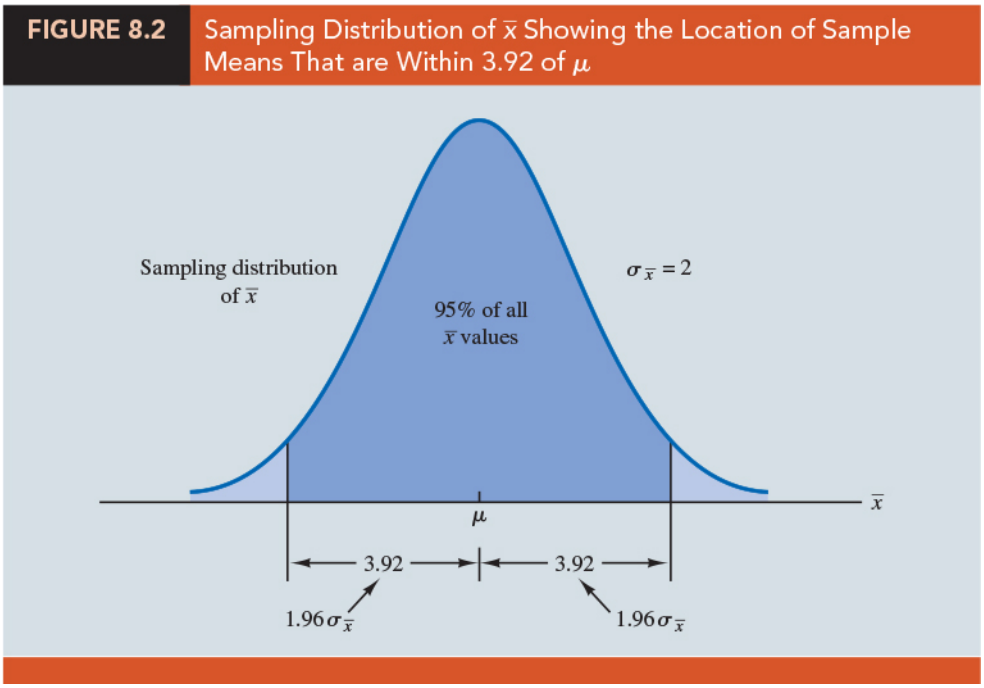
4. When the sampling distribution of $\bar{x}$ is normally distributed, 95% of the $\bar{x}$ values must be within _____ of the mean μ.

5. Example Lloyd's example:

   (a) (Figure 8.2) The sampling distribution of $\bar{x}$ is normally distributed with a standard error of $\sigma_{\bar{x}} = 2$.

   (b) Because _____, we can conclude that 95% of all $\bar{x}$ values obtained using a sample size of $n = 100$ will be within _____ of the population mean μ.

FIGURE 8.2 Sampling Distribution of $\bar{x}$ Showing the Location of Sample Means That are Within 3.92 of μ

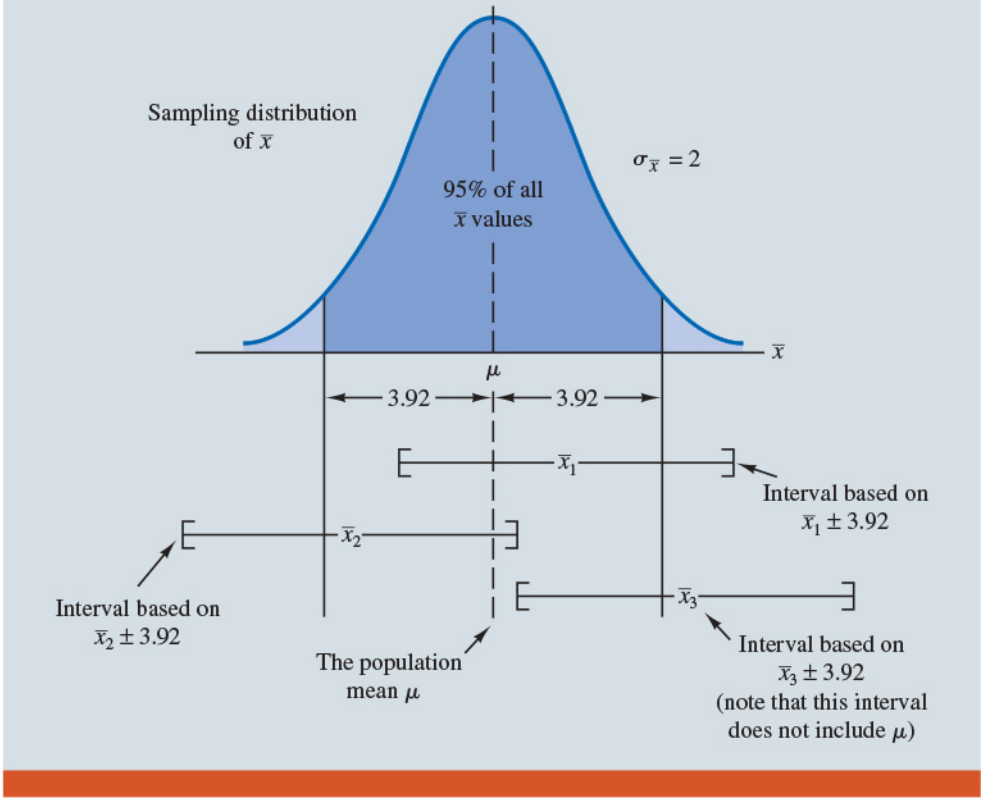


   (c) Suppose we set the _____ equal to 3.92 and compute the interval estimate of _____.

6. *An interpretation for this interval estimate*:

   (a) (Figure 8.3) Consider the values of $\bar{x}$ that could be obtained if we took three different simple random samples, each consisting of 100 Lloyd's customers.

FIGURE 8.3 Intervals Formed from Selected Sample Means at Locations $\bar{x}_1$, $\bar{x}_2$, and $\bar{x}_3$



(b) The first sample mean might turn out to have the value shown as $\bar{x}_1$ in Figure 8.3. The interval formed by subtracting 3.92 from $\bar{x}_1$ and adding 3.92 to $\bar{x}_1$ _____ the population mean μ.

(c) The interval $\bar{x}_2 \pm 3.92$ obtained by the second sample mean $\bar{x}_2$ also _____ the population mean μ.

(d) The interval $\bar{x}_3 \pm 3.92$ formed by the third sample mean _____ the population mean μ.

(e) Any sample mean $\bar{x}$ that is within the darkly shaded region of Figure 8.3 will provide an interval that _____ the population mean μ.

(f) Because _____ are in the darkly shaded region, 95% of all intervals formed by subtracting 3.92 from $\bar{x}$ and adding 3.92 to $\bar{x}$ will include the population mean μ.

(g) The quality assurance team at Lloyd's surveyed 100 customers and obtained

a sample mean amount spent of _____. Using $\bar{x} \pm 3.92$ to construct the interval estimate, we obtain _____.

(h) Because 95% of all the intervals constructed using $\bar{x} \pm 3.92$ will contain the population mean, we say that _____ _____.

7. We say that this interval has been established at the _____. The value 0.95 is referred to as the _____, and the interval 78.08 to 85.92 is called the 95% _____.

8. **Interval Estimate of a Population Mean: σ Known**:

_____ (8.1)

where _____ is the confidence coefficient and $z_{\alpha/2}$ is the z value providing an area of $\alpha/2$ in the upper tail of the standard normal probability distribution.

9. Comparing the results for the 90%, 95%, and 99% confidence levels, we see that in order to have a higher degree of confidence, the margin of error and thus the width of the confidence interval must be _____.

Question ..... (p113)

Construct the 90%, 95% and 99% confidence intervals for the Lloyd's example.

sol:

TABLE 8.1 Values of $z_{\alpha/2}$ for the Most Commonly Used Confidence Levels

Confidence Level	α	$\alpha/2$	$z_{\alpha/2}$
90%	.10	.05	1.645
95%	.05	.025	1.960
99%	.01	.005	2.576

Practical Advice

1. If the population follows a normal distribution, the confidence interval provided by expression (8.1) is _____.
2. If expression (8.1) were used _____ to generate 95% confidence intervals, _____ of the intervals generated would _____ the population mean.
3. If the population does not follow a normal distribution, the confidence interval provided by expression (8.1) will be _____. In this case, the quality of the approximation depends on both the _____ and the _____.
4. In most applications, a sample size of _____ is adequate when using expression (8.1) to develop an interval estimate of a population mean.
5. If the population is not normally distributed but is _____, sample sizes _____ can be expected to provide good approximate confidence intervals.
6. With _____, expression (8.1) should only be used if the analyst believes, or is willing to assume, that the population distribution is at least _____.

8.2 Population Mean: σ Unknown

1. When the population standard deviation σ is _____, we must use the _____ to estimate both _____ and _____ to obtain an interval estimate of a population mean μ.
2. When s is used to estimate σ, the _____ and the _____ for the population mean are based on a probability distribution known as the _____.

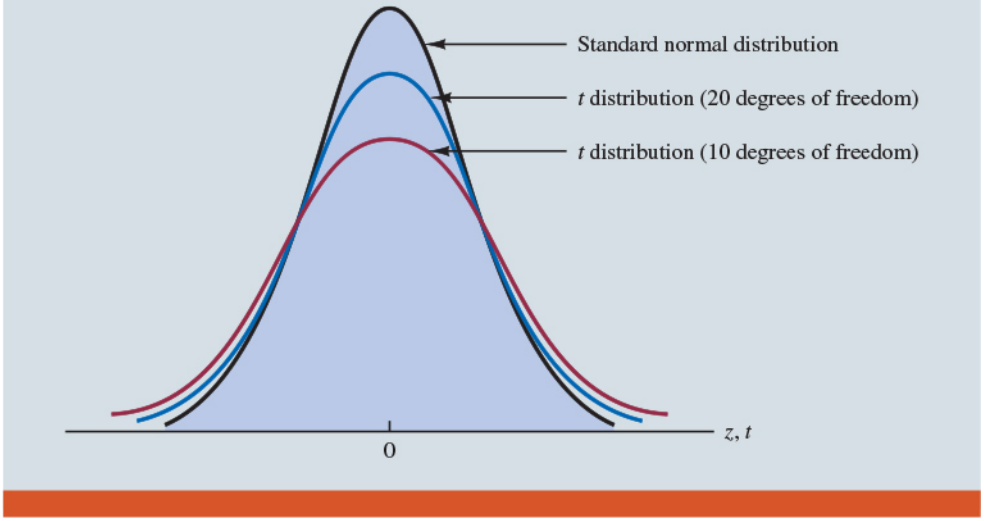
3. Although the mathematical development of the t distribution is based on the assumption of a _____ for the population we are sampling from, research shows that the t distribution can be successfully applied in many situations where the _____ deviates significantly from normal.

4. The t distribution is a family of similar probability distributions, with a specific t distribution depending on a parameter known as the _____.

5. (Figure 8.4) The t distribution with one degree of freedom is unique, as is the t distribution with two degrees of freedom, with three degrees of freedom, and so on. As the number of degrees of freedom _____, the difference between the t distribution and the _____ distribution becomes smaller and smaller.

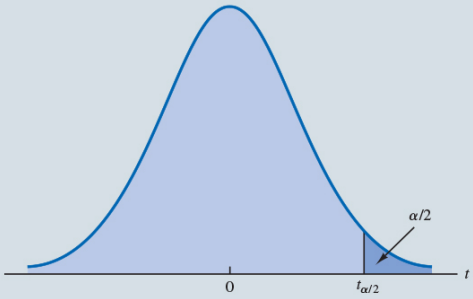
6. A t distribution with more degrees of freedom exhibits _____ and more closely resembles the _____ distribution.

FIGURE 8.4 Comparison of the Standard Normal Distribution with t Distributions Having 10 and 20 Degrees of Freedom



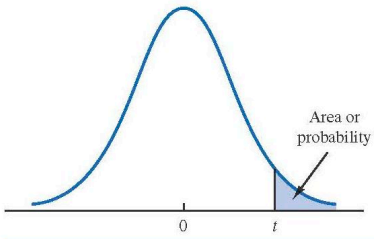
7. (Figure 8.5) The mean of the t distribution is _____. In general, we will use the notation _____ to represent a _____ with an area of _____ in the _____ of the t distribution.

FIGURE 8.5 *t* Distribution with α/2 Area or Probability in the Upper Tail



8. (Table 2, Appendix B) Each row in the table corresponds to a separate t distribution with the degrees of freedom. (e.g., _____). As the degrees of freedom continue to increase, $t_{0.025}$ approaches _____. The standard normal distribution z values can be found in the _____ degrees of freedom (labeled _____) of the t distribution table. For more than _____ degrees of freedom, the standard normal z value provides a good approximation to the t value.

TABLE 2 *t* Distribution



Entries in the table give t values for an area or probability in the upper tail of the t distribution. For example, with 10 degrees of freedom and a .05 area in the upper tail, $t_{.05} = 1.812$.

Degrees of Freedom	Area in Upper Tail					
	.20	.10	.05	.025	.01	.005
1	1.376	3.078	6.314	12.706	31.821	63.656
2	1.061	1.886	2.920	4.303	6.965	9.925
3	.978	1.638	2.353	3.182	4.541	5.841
4	.941	1.533	2.132	2.776	3.747	4.604
5	.920	1.476	2.015	2.571	3.365	4.032
6	.906	1.440	1.943	2.447	3.143	3.707
7	.896	1.415	1.895	2.365	2.998	3.499
8	.889	1.397	1.860	2.306	2.896	3.355
9	.883	1.383	1.833	2.262	2.821	3.250
10	.879	1.372	1.812	2.228	2.764	3.169
11	.876	1.363	1.796	2.201	2.718	3.106
12	.873	1.356	1.782	2.179	2.681	3.055
13	.870	1.350	1.771	2.160	2.650	3.012
14	.868	1.345	1.761	2.145	2.624	2.977

Margin of Error and the Interval Estimate

1. An interval estimate of a population mean for the σ known case: _____

2. To compute an interval estimate of μ for the σ unknown case, the sample standard deviation s is used to estimate σ, and _____ is replaced by the t distribution value _____.

3. The margin of error is then given by _____.

4. **Interval Estimate of a Population Mean: σ Unknown**:

$$\text{_____} \qquad (8.2)$$

where s is the sample standard deviation, _____ is the _____, and $t_{\alpha/2}$ is the t value providing an area of $\alpha/2$ in the upper tail of the t distribution with _____ degrees of freedom.

5. The reason the number of degrees of freedom associated with the t value in expression (8.2) is $n-1$ concerns the use of _____ as an estimate of the population standard deviation _____.

   (a) The expression for the sample standard deviation is

   $$s = \text{_____}.$$

   (b) Degrees of freedom refer to the _____ pieces of information that go into the computation of _____.

   (c) The _____ pieces of information involved in computing $\sum(x_i-\bar{x})^2$ are as follows: _____.

   (d) Note that _____ for any data set.

   (e) Thus, only _____ of the _____ values are independent; that is, if we know $n-1$ of the values, the remaining value can be determined exactly by using the condition that _____ values must be _____.

   (f) Thus, _____ is the number of degrees of freedom associated with _____ and hence the number of degrees of freedom for the t distribution in expression (8.2).

Question (p384)

(Table 8.3) To illustrate the interval estimation procedure for the σ unknown case, we will consider a study designed to estimate the mean credit card debt for the population of U.S. households. A sample of $n = 70$ households provided the credit card balances shown in Table 8.3. For this situation, no previous estimate of the population standard deviation σ is available. Thus, the sample data must be used to estimate both the population mean and the population standard deviation. Compute an interval estimate of the population mean credit card balance.

sol:

Compute: $\bar{x} = \$9312$, $s = 4007$.

With 95% confidence and $n-1 = 69$ degrees of freedom: $t_{69,0.025} = 1.995$.

An interval estimate of the population mean credit card balance:

$$9312 \pm 1.995\frac{4007}{\sqrt{70}} = 9312 \pm 955.$$

The margin of error is \$955, and the 95% confidence interval is $9312-955 = \$8357$ to $9312 + 955 = \$10,267$. Thus, we are 95% confident that the mean credit card balance for the population of all households is between \$8357 and \$10,267.

TABLE 8.3 Credit Card Balances for a Sample of 70 Households

9430	14661	7159	9071	9691	11032
7535	12195	8137	3603	11448	6525
4078	10544	9467	16804	8279	5239
5604	13659	12595	13479	5649	6195
5179	7061	7917	14044	11298	12584
4416	6245	11346	6817	4353	15415
10676	13021	12806	6845	3467	15917
1627	9719	4972	10493	6191	12591
10112	2200	11356	615	12851	9743
6567	10746	7117	13627	5337	10324
13627	12744	9465	12557	8372	
18719	5742	19263	6232	7445	

Practical Advice

1. If the population follows a _____ distribution, the confidence interval provided by expression (8.2) is _____ and can be used for _____ sample size.

$$\underline{\qquad\qquad} \tag{8.2}$$

2. If the population does not follow a normal distribution, the confidence interval provided by expression (8.2) will be _____. The quality of the approximation depends on both the _____ and the _____.

   (a) In most applications, a sample size of _____ is adequate when using expression (8.2) to develop an interval estimate of a population mean.

   (b) If the population distribution is _____ or contains _____, most statisticians would recommend increasing the sample size to _____.

   (c) If the population is not normally distributed but is _____, sample sizes as small as _____ can be expected to provide good approximate confidence intervals.

   (d) With _____, expression (8.2) should only be used if the analyst _____, or is willing to _____, that the population distribution is at least approximately normal.

Using a Small Sample

1. Using a histogram of the sample data to learn about the distribution of a population is not always conclusive, but in many cases it provides the only information available. The histogram, along with judgment on the part of the analyst, can often be used to decide whether expression (8.2) can be used to develop the interval estimate.

✎ Question ..... (p385)

Scheer Industries is considering a new computer-assisted program to train maintenance employees to do machine repairs. In order to fully evaluate the program, the director of manufacturing requested an estimate of the population mean time required for maintenance employees to complete the computer-assisted training. A

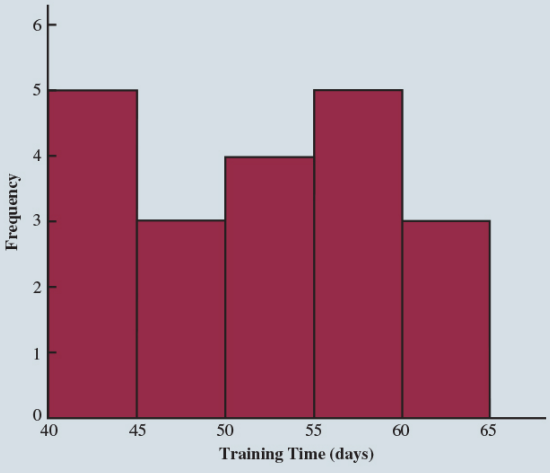
sample of 20 employees is selected, with each employee in the sample completing the training program. Data on the training time in days for the 20 employees are shown in Table 8.4.

TABLE 8.4 Training Time In Days for a Sample of 20 Scheer Industries Employees

52	59	54	42
44	50	42	48
55	54	60	55
44	62	62	57
45	46	43	56

1. A histogram of the sample data is shown in Figure 8.6. What can we say about the distribution of the population based on this histogram?

FIGURE 8.6 Histogram of Training Times for the Scheer Industries Sample



2. With a 95% confidence, find the interval estimate for the population mean.

sol:

1. The sample data do not support the conclusion that the distribution of the population is normal, yet we do not see any evidence of skewness or outliers. Therefore, we conclude that an interval estimate based on the t distribution appears acceptable for the sample of 20 employees.

2. Compute the sample mean and sample standard deviation as

$$\bar{x} = \frac{\sum x_i}{n} = \frac{1030}{20} = 51.5 \text{ days} \qquad s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{889}{20-1}} = 6.84 \text{ days}$$

For a 95% confidence interval, $t_{19,0.025} = 2.093$. The interval estimate of the population mean is:

$$51.56 \pm 2.093\left(\frac{6.84}{\sqrt{20}}\right) = 51.5 \pm 3.2$$

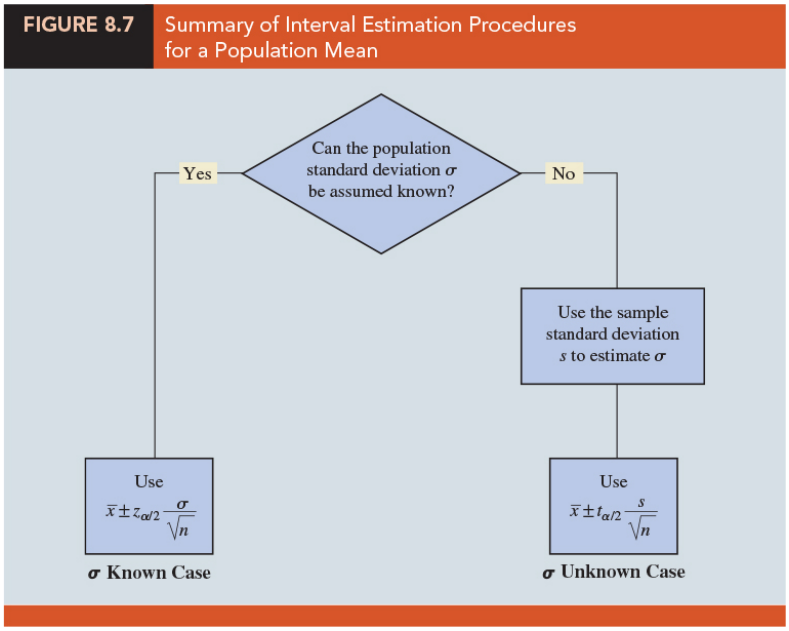
The 95% confidence interval is $51.5 - 3.2 = 48.3$ days to $51.5 + 3.2 = 54.7$ days.

Summary of Interval Estimation Procedures

1. Two approaches to develop an interval estimate of a population mean.

   (a) For the σ known case, _____ and the _____ distribution are used in expression (8.1) to compute the margin of error and to develop the interval estimate.

   (b) For the σ unknown case, the _____ and the _____ distribution are used in expression (8.2) to compute the margin of error and to develop the interval estimate.



2. (Figure 8.7) In most applications, a sample size of _____ is adequate. If the population has a _____ or _____ normal distribution, however, smaller sample sizes may be used. For the σ unknown case a sample size of _____ is recommended if the population distribution is _____ to be _____ or has _____.

8.3 Determining the Sample Size

1. How to choose a _____ large enough to provide a desired margin of error?
2. For the σ known case, the interval estimate is

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}},$$

   The quantity _____ is the margin of error.
3. Once we select a confidence coefficient _____, $z_{\alpha/2}$ can be determined. Then, if we have a value for σ, we can determine the sample size n needed to provide any desired margin of error.

✎ Question ..... (p390)

Suppose σ is known, using the interval estimate of a population mean at the chosen confidence level $1-\alpha$ to develop the formula used to compute the required sample size n.

sol:

4. **Sample Size for an Interval Estimate of a Population Mean**:

$$n = \frac{\quad}{\quad}, \ E \text{ is the desired margin of error} \qquad (8.3)$$

   This sample size provides the desired margin of error at the chosen confidence level.
5. Although user preference must be considered, _____ confidence is the most frequently chosen value (_____).

6. However, even if _____ is unknown, we can use equation (8.3) provided we have a preliminary or planning value for σ.

7. In practice, one of the following procedures can be chosen.

   (a) Use the estimate of the population standard deviation computed from _____ as the planning value for σ.

   (b) Use a _____ to select a _____ sample. The sample standard deviation from the preliminary sample can be used as the planning value for σ.

   (c) Use _____ or a _____ for the value of σ. For example, we might begin by estimating the _____ and _____ data values in the population. The difference between the largest and smallest values provides an estimate of the _____ for the data. Finally, the _____ is often suggested as a rough approximation of the _____ and thus an acceptable planning value for σ.

✎ Question ..... (p391)

A previous study that investigated the cost of renting automobiles in the United States found a mean cost of approximately $55 per day for renting a midsize automobile. Suppose that the organization that conducted this study would like to conduct a new study in order to estimate the population mean daily rental cost for a midsize automobile in the United States. In designing the new study, the project director specifies that the population mean daily rental cost be estimated with a margin of error of $2 and a 95% level of confidence. The project director specified a desired margin of error of $E = 2$, and the 95% level of confidence indicates $z_{0.025} = 1.96$. Thus, we only need a planning value for the population standard deviation σ in order to compute the required sample size. At this point, an analyst reviewed the sample data from the previous study and found that the sample standard deviation for the daily rental cost was $9.65.

sol:

Using 9.65 as the planning value for σ, we obtain

$$n = \frac{(z_{\alpha/2})^2\sigma^2}{E^2} = \frac{(1.96)^2(9.65)^2}{2^2} = 89.43$$

Thus, the sample size for the new study needs to be at least 89.43 midsize automobile rentals in order to satisfy the project director's \$2 margin of error requirement. In cases where the computed n is not an integer, we round up to the next integer value; hence, the recommended sample size is 90 midsize automobile rentals.

8.4 Population Proportion

1. The general form of an interval estimate of a population proportion p is

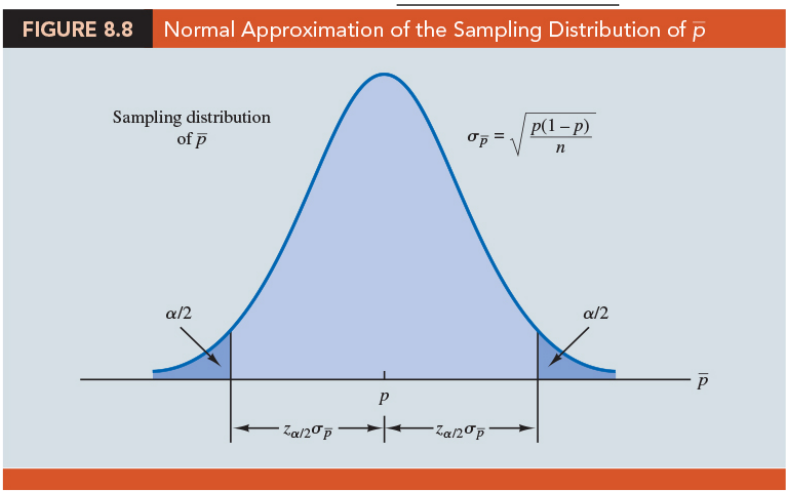
   _____

2. (Chapter 7) Recall: the sampling distribution of $\bar{p}$ can be approximated by a _____ distribution whenever _____ and _____.

3. (Figure 8.8) shows the _____ of the sampling distribution of $\bar{p}$. The mean of the sampling distribution of $\bar{p}$ is the population proportion _____, and the standard error of $\bar{p}$ is

   $$\sigma_{\bar{p}} = _____ \qquad (8.4)$$

FIGURE 8.8 Normal Approximation of the Sampling Distribution of $\bar{p}$



4. Because the sampling distribution of $\bar{p}$ is normally distributed, if we choose _____ as the margin of error in an interval estimate of a population proportion, we know that _____ of the intervals generated will contain the true population proportion.

5. But _____ cannot be used directly in the computation of the margin of error because p will not be known; p is what we are trying to estimate. So $\bar{p}$ is substituted for p and the margin of error for an interval estimate of a population proportion is given by

$$\text{Margin of error} = _____ \tag{8.5}$$

6. **Interval Estimate of a Population Proportion**:

$$_____ \tag{8.6}$$

where $1-\alpha$ is the confidence coefficient and $z_{\alpha/2}$ is the z value providing an area of $\alpha/2$ in the _____ of the standard normal distribution.

✎ Question ..... (p394)

A national survey of 900 women golfers was conducted to learn how women golfers view their treatment at golf courses in the United States. The survey found that 396 of the women golfers were satisfied with the availability of tee times. Compute and interpret the 95% confidence interval estimate of the population proportion.

sol:

Interpretation: The margin of error is .0324 and the 95% confidence interval estimate of the population proportion is 0.4076 to 0.4724. Using percentages, the survey results enable us to state with 95% confidence that between 40.76% and 47.24% of all women golfers are satisfied with the availability of tee times.

Determining the Sample Size

1. How large the _____ should be to obtain an estimate of a population proportion at a specified level of precision.

2. Let E denote the desired margin of error:

$$E = _____$$

3. Solving this equation for n provides a formula for the sample size that will provide a margin of error of size E:

$$n = _____$$

4. We cannot use this formula to compute the sample size that will provide the desired margin of error because _____ will not be known until after we select the _____.

5. What we need is a planning value for $\bar{p}$ (denoted by _____) that can be used to make the computation.

6. **Sample Size for an Interval Estimate of a Population Proportion**:

$$n = _____$$

7. In practice, the planning value p^* can be chosen by one of the following procedures:

   (a) Use the sample proportion from a _____ of the same or similar units.

   (b) Use a _____ to select a preliminary sample. The sample proportion from this sample can be used as the planning value, p^*.

   (c) Use _____ or a _____ for the value of p^*.

   (d) If none of the preceding alternatives applies, use a planning value of _____.

✎ Question ..... (p395)

Consider the survey of women golfers and assume that the company is interested in conducting a new survey to estimate the current proportion of the population of women golfers who are satisfied with the availability of tee times. How large should the sample be if the survey director wants to estimate the population proportion with a margin of error of 0.025 at 95% confidence? Using the previous survey result of $p = 0.44$ as the planning value p^*.

sol:

8. The fourth alternative suggested for selecting a planning value _____ is frequently used when no other information is available.

9. The numerator of equation (8.7) shows that the sample size is _____ to the quantity _____. A larger value for the quantity $p^*(1-p^*)$ will result in a larger sample size.

10. (Table 8.5) in case of any uncertainty about an appropriate planning value, we know that _____ will provide the _____ recommendation. In effect, we play it _____ by recommending the largest necessary sample size.

TABLE 8.5 Some Possible Values for $p^*(1 - p^*)$

p^*	$p^*(1 - p^*)$	
.10	(.10)(.90) = .09	
.30	(.30)(.70) = .21	
.40	(.40)(.60) = .24	
.50	(.50)(.50) = .25	← Largest value for $p^*(1 - p^*)$
.60	(.60)(.40) = .24	
.70	(.70)(.30) = .21	
.90	(.90)(.10) = .09	

11. If the sample proportion turns out to be different from the 0.50 planning value, the margin of error will be _____ than anticipated. Thus, in using $p^* = 0.50$, we guarantee that the sample size will be _____ to obtain the desired margin of error.

12. In the survey of women golfers example, a planning value of $p^* = 0.50$ would have provided the sample size

$$n = \frac{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad}{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad}.$$

Thus, a slightly larger sample size of _____ women golfers would be recommended.

8.5 Big Data and Confidence Intervals

1. The _____ are powerful tools for making inferences about population parameters.

2. We now consider the ramifications of _____ on confidences intervals for means and proportions.

Big Data and the Precision of Confidence Intervals

1. The confidence intervals for the population mean μ and population proportion p become more _____ as the sample size _____. Therefore, the potential _____ also decreases as the sample size increases.

2. Example Consider the online news service PenningtonDailyTimes.com (PDT). Prospective advertisers are willing to pay a premium to advertise on websites that have long visit times, so the time customers spend during their visits to PDT's website has a substantial impact on PDT's advertising revenues. Suppose PDT's management wants to develop a 95% confidence interval estimate of the mean amount of time customers spend during their visits to PDT's website.

3. (Table 8.6) shows how the margin of error at the 95% confidence level decreases as the sample size increases when $s = 20$.

TABLE 8.6 Margin of Error for Interval Estimates of the Population Mean at the 95% Confidence Level for Various Sample Sizes n

Sample Size n	Margin of Error $t_{\alpha/2}s_{\bar{x}}$
10	14.30714
100	3.96843
1,000	1.24109
10,000	.39204
100,000	.12396
1,000,000	.03920
10,000,000	.01240
100,000,000	.00392
1,000,000,000	.00124

4. (Table 8.7) Suppose PDT would like to develop a 95% confidence interval estimate of the proportion of its website visitors that click on an ad. Table 8.7 shows how the margin of error for a 95% confidence interval estimate of the population proportion decreases as the sample size increases when the sample proportion is $\bar{p} = 0.51$.

TABLE 8.7 Margin of Error for Interval Estimates of the Population Proportion at the 95% Confidence Level for Various Sample Sizes n

Sample Size n	Margin of Error $z_{\alpha/2}\sigma_{\bar{p}}$
10	.30984
100	.09798
1,000	.03098
10,000	.00980
100,000	.00310
1,000,000	.00098
10,000,000	.00031
100,000,000	.00010
1,000,000,000	.00003

5. We see in Tables 8.6 and 8.7 that at a given confidence level, the margins of error decrease as the sample sizes increase.

   (a) If the sample mean time spent by customers when they visit PDT's website is 84.1 seconds, the 95% confidence interval estimate of the population mean time spent by customers when they visit PDT's website decreases from (69.79286, 98.40714) for a sample of $n = 10$ to (83.97604, 84.22396) for a sample of $n = 100,000$ to (84.09876, 84.10124) for a sample of $n = 1,000,000,000$.

   (b) Similarly, if the sample proportion of its website visitors who clicked on an ad is 0.51, the 95% confidence interval estimate of the population proportion of its website visitors who clicked on an ad decreases from (0.20016, 0.81984) for a sample of $n = 10$ to (0.50690, 0.51310) for a sample of $n = 100,000$ to (0.50997, 0.51003) for a sample of $n = 1,000,000,000$.

6. In both instances, as the sample size becomes extremely _____, the margin of error becomes extremely _____ and the resulting confidence intervals become extremely _____.

Implications of Big Data for Confidence Intervals

1. Example Last year the mean time spent by all visitors to PenningtonDailyTimes.com was 84 seconds. Suppose that PDT wants to assess whether the population mean time has changed since last year. PDT now collects a new sample of 1,000,000 visitors to its website and calculates the sample mean time spent by these visitors

to the PDT website to be $\bar{x} = 84.1$ seconds. The estimated population standard deviation is $s = 20$ seconds, so the standard error is _____.

2. Furthermore, the sample is sufficiently large to ensure that the sampling distribution of the sample mean will be normally distributed. Thus, the 95% confidence interval estimate of the population mean is

   _____

3. What could PDT conclude from these results? There are three possible reasons that PDT's sample mean of 84.1 seconds differs from last year's population mean of 84 seconds: (1) _____, (2) _____, or (3) the population mean has _____ since last year.

   (a) The 95% confidence interval estimate of the population mean does not include the value for the mean time spent by all visitors to the PDT website for last year (84 seconds), suggesting that the difference between PDT's sample mean for the new sample (84.1 seconds) and the mean from last year (84 seconds) is not likely to be exclusively a consequence of _____.

   (b) Nonsampling error is a possible explanation and should be investigated as the results of statistical inference become _____ as _____ is introduced into the sample data.

   (c) If PDT determines that it introduced little or no nonsampling error into its sample data, the only remaining plausible explanation for a difference of this magnitude is that the population mean _____ since last year.

      i. If PDT concludes that the sample has provided reliable evidence and the population mean has changed since last year, management must still consider the potential impact of the difference between the sample mean and the mean from last year.

      ii. If a 0.1 second difference in the time spent by visitors to PenningtonDailyTimes.com has a consequential effect on what PDT can charge for advertising on its site, this result could have _____ for PDT.

iii. Otherwise, there may be _____ of the 0.1 second difference in the time spent by visitors to PenningtonDailyTimes.com.

4. Confidence intervals are extremely useful, but as with any other statistical tool, they are only effective when _____.

5. Because interval estimates become increasingly precise as the sample size increases, extremely _____ samples will yield extremely _____ estimates.

6. However, no interval estimate, no matter how precise, will accurately reflect the parameter being estimated unless the sample is relatively _____. Therefore, when using interval estimation, it is always important to carefully consider whether a _____ of the population of interest has been taken.

😍 EXERCISES

8.1 : 3, 8, 9

8.2 : 13, 16, 18, 20

8.3 : 24, 25, 29

8.4 : 31, 33, 37, 39, 43

8.5 : 47

SUP : 50, 51, 54, 56, 62

名言佳句

“你要搞清楚自己人生的劇本 – 不是你父母的續集，不是你子女的前傳，更不是你朋友的外篇。對待生命你不妨大膽冒險一點，因為好歹你要失去它。生命中最難的階段不是沒有人懂你，而是你不懂你自己。”
“You want to figure out your life’s script – not your parent’s sequel, not your child’s prequel, but not just a chapter in your friend’s book. Treat life like you want adventure and to be bold; because whatever the outcome, you can’t live forever. The hardest part of life is not that no one understands you, but you do not understand yourself.”
— Friedrich Wilhelm Nietzsche (October 15, 1844 – August 25, 1900)

統計小故事：內曼與信賴區間的誕生

二十世紀初，統計學逐漸從描述性工具走向推論性科學。研究者不再滿足於僅僅整理數據，而是希望透過樣本來推斷母體，並能在不確定性中作出合理判斷。在這個轉型的關鍵時刻，一位波蘭統計學家——耶日．內曼 (Jerzy Neyman) ——提出了一個改變統計推論的概念：信賴區間 (confidence interval)。

在內曼之前，研究者常常用「點估計」來表達結論。舉例來說，若抽樣計算得出某地區居民的平均收入為 3000 美元，人們就會將 3000 視為母體的代表。然而，內曼認為這樣的做法過於冒險，因為單一數字並不能真實反映不確定性。樣本畢竟只是母體的一部分，每次抽樣的結果可能不同，那麼為什麼我們要假裝只有一個確定答案呢？

於是，內曼提出了「區間估計」的概念。他主張，與其只報告一個點估計，不如提供一個範圍，並附上「信心水準」。例如，在 95% 的信心水準下，居民的平均收入介於 2800 到 3200 美元之間。這樣一來，我們就能以數學方式表達不確定性，並讓讀者明白推論的可信程度。

信賴區間的價值不僅在於精確性，更在於透明性。它誠實地告訴我們：結論不是絕對的，而是有一定的不確定範圍。這樣的觀念徹底改變了統計報告的方式，也讓統計學更符合科學精神。

為了說明信賴區間的重要性，我們可以舉一個醫學臨床試驗的例子。假設一項新藥的試驗結果顯示：平均能降低病人的血壓 5 毫米汞柱。如果我們僅報告「5」，讀者可能會以為藥效是穩定不變的。但若報告「在 95% 的信心水準下，降壓效果介於 3 到 7 毫米汞柱」，讀者就能更清楚地理解：藥效存在波動，但大致落在這個範圍內。這樣的表述更科學，也更能避免誤導。

信賴區間的誕生，也與另一位統計學家費雪 (Ronald A. Fisher) 的工作形成對比。費雪發展了「最大概似估計」與「假設檢定」，偏重於用數據檢驗假說是否成立。而內曼則更關心如何在不確定性下給出合理的範圍估計。兩人的方法看似不同，卻共同推動了現代統計推論的成熟。

除了醫學，信賴區間在民意調查中也有廣泛應用。當媒體報導「某候選人支持率為 52%」時，如果不附上信賴區間，數字很容易被過度解讀。正確的表述應該是：「在 95% 信心水準下，支持率介於 49% 到 55% 之間。」這樣的區間提醒讀者：調查有抽樣

誤差，選情可能比單一數字顯示的更接近。

在教材呼應上，本章的重點是「區間估計」的計算與解釋。學生需要理解，信賴區間的寬度取決於三個因素：樣本大小、變異程度、以及信心水準。樣本越大、變異越小，區間就越窄，推論也越精確；而信心水準越高，區間則越寬，表示我們更保守但也更有把握。

因此，當學生在課堂上計算「母體平均數的 95% 信賴區間」時，他們其實正延續內曼的理念：在不確定性中尋找合理的答案。這不僅是數學公式的練習，更是一種科學態度的養成。

總結來說，內曼的信賴區間教會我們：統計推論不是追求絕對正確，而是學會誠實地表達不確定性。在一個充滿隨機性的世界裡，這種思維方式或許比確定答案更重要。

統計學 (一)

Anderson's Statistics for Business & Economics (14/E)

Chapter 9: Hypothesis Tests

上課時間地點: 二 D56, 商館 260206

授課教師: 吳漢銘 (國立政治大學統計學系副教授)

教學網站: `https://hmwu.idv.tw`

系級: _____ 學號: _____ 姓名: _____

Overview

1. Statistical inference: how hypothesis testing can be used to determine whether a _____ about the value of a _____ should or should not be _____.
2. The _____ hypothesis (_____): making a _____ about a population parameter.
3. The _____ hypothesis (_____): the opposite of what is stated in H_0.
4. The hypothesis testing procedure uses _____ to test the two competing statements indicated by H_0 and H_a.
5. This chapter shows how hypothesis tests can be conducted about a _____ and a _____.

9.1 Developing Null and Alternative Hypotheses

1. It is _____ how the null and alternative hypotheses should be formulated.
2. All hypothesis testing applications involve collecting a _____ and using the sample results to provide _____ for drawing a _____.
3. In some situations it is easier to identify _____ first and then develop _____.
4. In other situations it is easier to identify _____ first and then develop _____.

The Alternative Hypothesis as a Research Hypothesis

1. Many applications of hypothesis testing involve an attempt to gather evidence in _____. In these situations, it is often best to begin with the _____ hypothesis and make it the conclusion that the researcher _____.
2. Example Consider a particular automobile that currently attains a fuel efficiency of 24 miles per gallon in city driving.
   (a) *Goal*: A product research group has developed a _____ fuel injection system (燃料噴射系統) designed to _____ the miles-per-gallon rating. The group will run controlled tests with the new fuel injection system looking for statistical support for the conclusion that the new fuel injection system provides more miles per gallon than the current system.
   (b) Several new fuel injection units will be manufactured, installed in test automobiles, and subjected to research-controlled driving conditions.
   (c) The _____ for these automobiles will be computed and used in a hypothesis test to determine if it can be concluded that the new system provides _____.
   (d) In terms of the population mean miles per gallon _____, the research hypothesis _____ becomes the alternative hypothesis.

(e) Since the current system provides an average or mean of 24 miles per gallon, we will make the tentative assumption that the new system is not any better than the current system and choose _____ as the null hypothesis.

_____

(f) If the sample results lead to the conclusion to reject H_0, the inference can be made that _____ is true.

(g) The researchers have the _____ to state that the new fuel injection system increases the mean number of miles per gallon.

(h) If the sample results lead to the conclusion that H_0 cannot be rejected, the researchers cannot conclude that the new fuel injection system is better than the current system. Production of automobiles with the new fuel injection system on the basis of better gas mileage cannot be justified. Perhaps _____ _____ can be conducted.

3. Before adopting something _____ (e.g., products, methods, systems), it is desirable to conduct research to determine if there is _____ for the conclusion that the new approach is indeed better. In such cases, the research hypothesis is stated as the _____.

(a) Example A new teaching method is developed that is believed to be better than the current method.

i. H_0: the new method is no better than the old method.

ii. H_a: the new method is _____.

(b) Example A new sales force bonus plan is developed in an attempt to increase sales.

i. H_0: the new bonus plan does not increase sales.

ii. H_a: the new bonus plan _____.

(c) Example A new drug is developed with the goal of lowering blood pressure more than an existing drug.

i. H_0: the new drug does not provide lower blood pressure than the existing drug.

ii. H_a: the new drug _____ blood pressure _____ the existing drug.

4. In each case, _____ of the null hypothesis H_0 provides _____ for the research hypothesis.

The Null Hypothesis as an Assumption to Be Challenged

1. The situations below that it is helpful to develop the null hypothesis first.
   (a) Consider applications of hypothesis testing where we begin with a _____ or an _____ that a statement about the value of a _____ _____ is _____.
   (b) We will then use a hypothesis test to _____ and determine if there is statistical evidence to conclude that the assumption is _____.
2. The null hypothesis H_0 expresses the _____ about the value of the population parameter. The alternative hypothesis H_a is that the belief or assumption is _____.
3. Example Consider the situation of a manufacturer of soft drink products.
   (a) The label on a soft drink bottle states that it contains 67.6 fluid ounces. We consider the label correct provided the _____ filling weight for the bottles is _____ 67.6 fluid ounces.
   (b) We would begin with the _____ that the label is correct and state the null hypothesis as _____.
   (c) The challenge to this assumption would imply that the label is incorrect and the bottles are being under-filled. This challenge would be stated as the alternative hypothesis _____.

       _____.
   (d) A government agency with the responsibility for validating manufacturing labels could select a sample of soft drinks bottles, compute the _____ filling weight, and use the sample results to test the preceding hypotheses.

(e) If the sample results lead to the conclusion to _____, the inference that _____ can be made. With this statistical support, the agency is justified in concluding that the _____ and _____ of the bottles is occurring.

(f) If the sample results indicate _____, the assumption that the manufacturer's labeling is correct cannot be rejected. With this conclusion, _____ would be taken.

4. Example Consider the soft drink bottle filling example from the manufacturer′ s point of view.

(a) The bottle-filling operation has been designed to fill soft drink bottles with 67.6 fluid ounces as stated on the label.

i. The company does not want to _____ the containers because that could result in an underfilling complaint from customers or, perhaps, a government agency.

ii. However, the company does not want to _____ containers either because putting more soft drink than necessary into the containers would be an unnecessary cost.

(b) The company's goal would be to adjust the bottle-filling operation so that the population mean filling weight per bottle is 67.6 fluid ounces as specified on the label.

(c) In a hypothesis testing application, we would begin with the assumption that the production process is operating correctly and state the null hypothesis as _____ fluid ounces.

(d) The alternative hypothesis that challenges this assumption is that _____, which indicates either overfilling or underfilling is occurring.

(e) Suppose that the soft drink manufacturer uses a quality control procedure to periodically select a sample of bottles from the filling operation and computes the _____ filling weight per bottle.

i. If the sample results lead to the conclusion to _____, the inference is made that _____ is true. We conclude that the bottles are not being filled properly and the _____ to restore the population mean to 67.6 fluid ounces per bottle.

ii. If the sample results indicate _____, the assumption that the manufacturer's bottle filling operation is functioning properly cannot be rejected. In this case, _____ would be taken and the production operation would continue to run.

5. The two preceding forms of the soft drink manufacturing hypothesis test show that the null and alternative hypotheses may _____ of the researcher or decision maker.

6. To correctly _____ it is important to understand the context of the situation and structure the hypotheses to provide the information the researcher or decision maker wants.

Summary Of Forms for Null and Alternative Hypotheses

1. Depending on the situation, hypothesis tests about a population parameter (the population mean and the population proportion) may take one of three forms:

2. The first two forms are called _____. The third form is called a _____.

3. The _____ part of the expression (either $\geq$, $\leq$, or $=$) always appears in the _____ hypothesis.

4. In selecting the proper form of H_0 and H_a, keep in mind that the _____ hypothesis is often what _____. Hence, asking whether the user is looking for evidence to support _____ will help determine H_a.

9.2 Type I and Type II Errors

1. Ideally the hypothesis testing procedure should lead to the _____ when _____ and the rejection of H_0 when H_a is true.

2. (Table 9.1) The correct conclusions are not always possible.

TABLE 9.1 Errors and Correct Conclusions in Hypothesis Testing

		Population Condition	
		H_0 True	H_a True
Conclusion	Accept H_0	Correct Conclusion	Type II Error
	Reject H_0	Type I Error	Correct Conclusion

   (a) We reject H_0 if H_0 is true, we make a _____.

   (b) If H_a is true, the conclusion is correct when we reject H_0.

   (c) If H_0 is true, the conclusion is correct when we accept H_0.

   (d) If H_0 is false (H_a is true), we make a _____ when we accept H_0.

3. Example An automobile product research group developed a new fuel injection system designed to increase the miles-per-gallon rating of a particular automobile.

   (a) With the current model obtaining an average of 24 miles per gallon, the hypothesis test was formulated as follows.

   _____

   (b) The alternative hypothesis, $H_a : \mu > 24$, indicates that the researchers are looking for _____ the conclusion that the population mean miles per gallon with the new fuel injection system is _____ 24.

   (c) *Type I error*: _____ corresponds to the researchers _____ that the new system improves the miles-per-gallon rating ($\mu > 24$) when _____ the new system is not any better than the current system.

(d) *Type II error*: _____ corresponds to the researchers _____ that the new system is not any better than the current system ($\mu \leq 24$) when _____ the new system improves miles-per-gallon performance.

4. **Level of Significance**: The level of significance is the probability of making a Type I error when the null hypothesis is true as an _____.

   (a) Example For the miles-per-gallon rating hypothesis test, the null hypothesis is $H_0 : \mu \leq 24$. Suppose the null hypothesis is true as an _____; that is, _____. The level of significance is the probability of _____ when _____.

   (b) The Greek symbol _____ (alpha) is used to denote the level of significance, and common choices for α are _____.

   (c) In practice, the person responsible for the hypothesis test specifies the level of significance. By selecting α, that person is _____ of making a Type I error.

   (d) If the cost of making a Type I error is _____, _____ values of α are preferred.

5. **The significance tests**: Applications of hypothesis testing that only control for the Type I error are called _____.

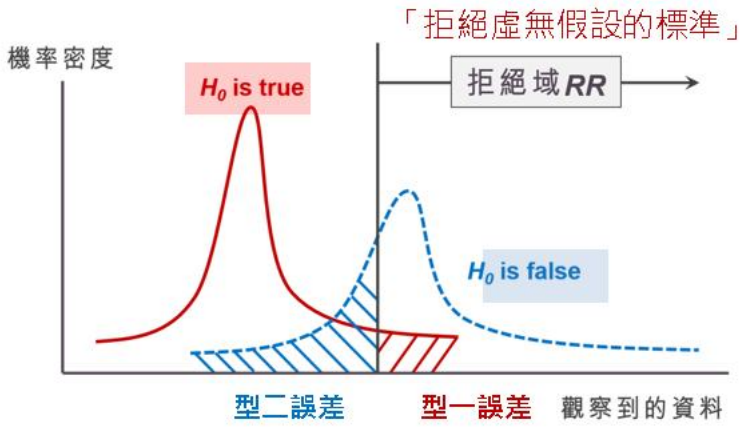
6. Although most applications of hypothesis testing control for the probability of making a Type I error, they do not always control for the probability of making a _____.

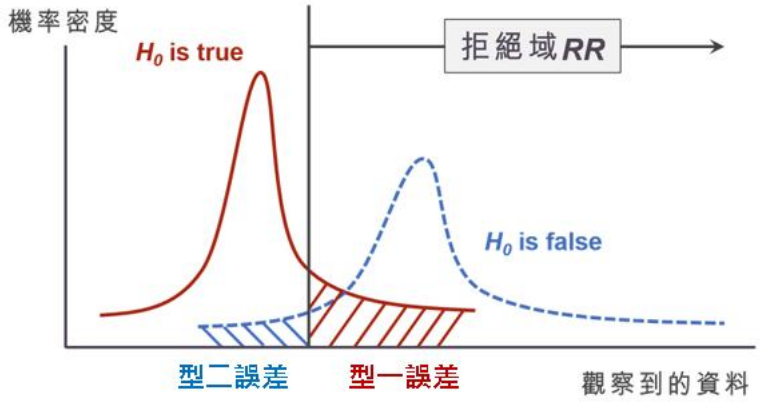
   (a) Hence, if we decide to accept H_0, we cannot determine _____ we can be with that decision. Because of the _____ associated with making a Type II error when conducting significance tests, statisticians usually recommend that we use the statement _____ instead of _____

   (b) Using the statement "do not reject H_0" carries the recommendation to _____ _____. In effect, by not directly accepting H_0, the statistician _____ of making a Type II error.

(c) Whenever the probability of making a Type II error has not been determined and controlled, we will not make the statement _____ In such cases, only two conclusions are possible: _____ or _____.

(d) Although controlling for a Type II error in hypothesis testing is _____, it can be done. In Sections 9.7 and 9.8 we will illustrate procedures for determining and controlling the probability of making a Type II error. If proper controls have been established for this error, _____ based on the _____ conclusion can be appropriate.





9.3 Population Mean: σ Known

One-tailed Test

1. One-tailed tests about a population mean take one of the following two forms:

---

2. Example The Federal Trade Commission (FTC) periodically conducts statistical studies designed to test the claims that manufacturers make about their products. For example, the label on a large can of Hilltop Coffee states that the can contains 3 pounds of coffee. The FTC knows that Hilltop's production process cannot place exactly 3 pounds of coffee in each can, even if the mean filling weight for the population of all cans filled is 3 pounds per can ($\mu_0 = 3$). However, as long as the population mean filling weight is at least 3 pounds per can, the rights of consumers will be protected. Thus, the FTC interprets the label information on a large can of coffee as a claim by Hilltop that the population mean filling weight is at least 3 pounds per can. We will show how the FTC can check Hilltop's claim by conducting a _____.

   (a) *Develop the null and alternative hypotheses for the test.* If the population mean filling weight is at least 3 pounds per can, Hilltop's claim is correct:

   $$H_0 : _____ \quad H_a : _____$$

      i. If the sample data indicate that _____, no action should be taken against Hilltop.

      ii. If the sample data indicate _____, $H_a : \mu < 3$, is true. A conclusion of _____ and a charge of a label violation against Hilltop would be justified.

      iii. Suppose a sample of $n = 36$ cans of coffee is selected and the sample mean _____ is computed as an estimate of the population mean _____. If the value of the sample mean $\bar{x}$ is _____ 3 pounds, the sample results will _____ on the null hypothesis.

(b) *Specifying the level of significance, α:*

i. (Recall) The level of significance is the probability of making a Type I error by _____ when H_0 is true as an _____.

ii. If the _____ of making a Type I error is _____, a _____ _____ should be chosen for the level of significance.

iii. In the Hilltop Coffee study, the director of the FTC's testing program made the following statement: "If the company is meeting its weight specifications at $\mu = 3$, I do not want to take action against them. But, I am willing to risk a 1% chance of making such an error." From the director's statement, we set the level of significance for the hypothesis test at _____.

iv. Thus, we must design the hypothesis test so that the probability of making a _____ when $\mu = 3$ is 0.01.

3. By developing the null and alternative _____ and specifying the _____ _____ for the test, we carry out the first two steps required in conducting every hypothesis test. We are now ready to perform the third step of hypothesis testing: _____ data and compute the value of what is called a _____.

Test statistic

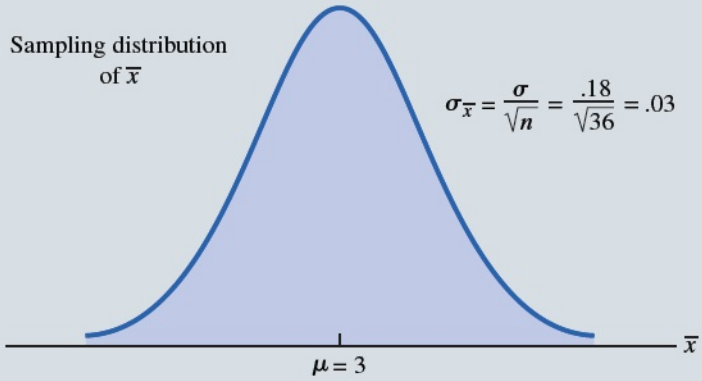
1. Example For the Hilltop Coffee study, previous FTC tests show that the population standard deviation can be assumed _____ with a value of $\sigma = 0.18$. These tests also show that the population of filling weights can be assumed to have a _____ distribution.

2. The sampling distribution of $\bar{x}$ is _____ distributed with a known value of $\sigma = 0.18$ and a sample size of $n = 36$.

3. (Figure 9.1) the sampling distribution of $\bar{x}$ when the null hypothesis is true as an equality ($\mu = \mu_0 = 3$). The standard error of $\bar{x}$ is given by $\sigma_{\bar{x}} =$ _____. The sampling distribution of

$$z = \text{_____} \quad \text{is a standard normal distribution.}$$

FIGURE 9.1 Sampling Distribution of $\bar{x}$ for the Hilltop Coffee Study When the Null Hypothesis Is True as an Equality ($\mu = 3$)



4. A value of _____ means that the value of $\bar{x}$ is _____ _____ below the hypothesized value of the mean.

5. The lower tail area at $z = -3.00$ is _____. Hence, the probability of obtaining a value of z that is three or more standard errors _____ is 0.0013.

6. The probability of obtaining a value of $\bar{x}$ that is 3 or more standard errors below the hypothesized population mean $\mu_0 = 3$ is also 0.0013. Such a result is _____ if the null hypothesis is true.

7. For hypothesis tests about a population mean in the σ known case, we use the standard normal random variable _____ as a _____ to determine whether $\bar{x}$ _____ the hypothesized value of μ enough to justify rejecting the null hypothesis.

8. **Test Statistic for Hypothesis Tests About a Population Mean: σ Known**

   _____.

9. The key question for a lower tail test is, _____ must the test statistic z be before we choose to _____? Two approaches: the _____ approach and the _____ approach.

p-value approach

1. **p-value**: A p-value is a probability that provides a _____ of the evidence _____ provided by the _____.
   (a) The p-value is used to determine whether H_0 should be _____.
   (b) A _____ indicates the value of the test statistic is _____ given the assumption that _____.
   (c) _____ p-values indicate _____ against H_0.
2. The value of the test statistic is used to compute the p-value.
   (a) For a _____ test, the p-value is the probability of obtaining a value for the test statistic _____ or _____ than that provided by the sample.
   (b) To compute the p-value for the lower tail test in the σ known case, we use the standard normal distribution to find the probability that _____ is _____ the value of the test statistic.
   (c) After computing the p-value, we must then decide whether it is _____ to reject the null hypothesis; as we will show, this decision involves comparing the p-value to the level of significance.

✎ Question ..... (p427)

Suppose the sample of 36 Hilltop coffee cans provides a sample mean of $\bar{x} = 2.92$ pounds. Is $\bar{x} = 2.92$ small enough to cause us to reject H_0? Compute the p-value for the Hilltop Coffee lower tail test.

sol:

– Because this is a lower tail test, the p-value is the area under the standard normal curve for values of $z \leq$ the value of the test statistic.

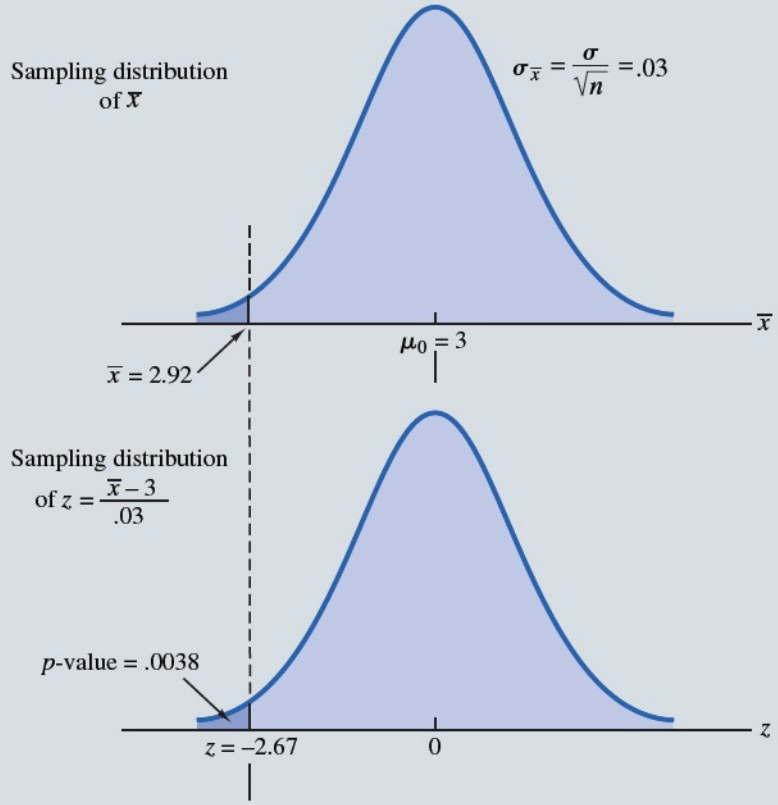
– Using $\bar{x} = 2.92$, $\sigma = 0.18$, and $n = 36$, the value of the test statistic z:

$$z = \underline{\qquad\qquad\qquad\qquad}.$$

– p-value _____.

– (Figure 9.2) $\bar{x} = 2.92$ corresponds to $z = -2.67$ and a p-value $= 0.0038$.

FIGURE 9.2 p-Value for the Hilltop Coffee Study When $\bar{x} = 2.92$ and $z = -2.67$



3. This p-value (0.0038) indicates a small probability of _____ of $\bar{x} = 2.92$ (and a test statistic of -2.67) or smaller when sampling from a population with _____.

4. This p-value does not provide much support for the null hypothesis, but _____ enough to cause us to reject H_0? The answer depends upon _____ for the test.

5. As noted previously, the director of the FTC's testing program selected a value of 0.01 for the level of significance means that the director is _____ a probability of 0.01 of rejecting H_0 when it is true as an _____.

6. The sample of 36 coffee cans in the Hilltop Coffee study resulted in a p-value = 0.0038, which means that the _____ of obtaining a value of $\bar{x} = 2.92$ or less when H_0 is true as an equality is _____.

7. Because _____, we _____. Therefore, we find _____ to reject the null hypothesis at the 0.01 level of significance.

8. **Rejection Rule Using p-value**. For a level of significance α, the rejection rule using the p-value approach is:

   _____

9. Example In the Hilltop Coffee test, the p-value of 0.0038 resulted in the rejection of H_0. The observed p-value of 0.0038 means that we would reject H_0 for any value of _____. For this reason, the p-value is also called _____ _____.

10. Different decision makers may express different opinions concerning the _____ _____ and may choose a _____.

Critical value approach

1. The critical value is the value of the test statistic that corresponds to an _____ in the lower tail of the _____ of the test statistic.

2. The critical value is the _____ of the test statistic that will result in the rejection of the null hypothesis.

3. For a lower tail test, the critical value serves as a _____ for determining whether the value of the test statistic is _____ to reject the null hypothesis.

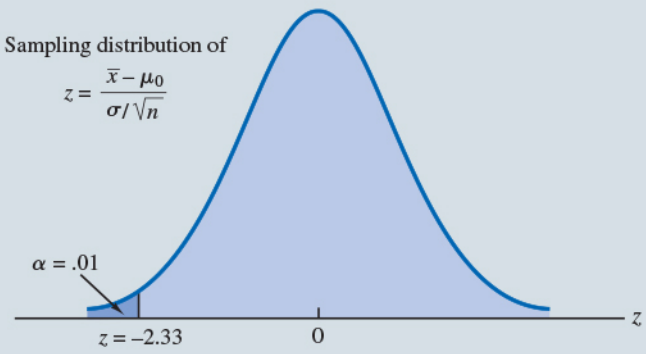
4. Example the Hilltop Coffee example.

   (a) In the σ known case, the sampling distribution for the test statistic z is a _____ distribution. Therefore, the critical value is the value of the test statistic that corresponds to an area of _____ in the lower tail of a standard normal distribution.

   (b) (Figure 9.3) We find that _____ provides an area of 0.01 in the lower tail, _____.

   (c) If the sample results in a value of the test statistic that is less than or equal to -2.33, the corresponding p-value will be less than or equal to 0.01; in this case, we should reject H_0.

FIGURE 9.3 Critical Value = −2.33 for the Hilltop Coffee Hypothesis Test



   (d) Hence, for the Hilltop Coffee study the critical value _____ for a level of significance of 0.01 is

   Reject H_0 if _____

   (e) In the Hilltop Coffee example, $\bar{x} = 2.92$ and the test statistic is $z = -2.67$. Because _____, we can reject H_0 and conclude that Hilltop Coffee is _____ cans.

5. **Rejection Rule for a Lower Tail Test: Critical Value Approach**. We can generalize the rejection rule for the critical value approach to handle any level of significance. The rejection rule for a lower tail test follows.

   Reject H_0 if _____

where $-z_\alpha$ is the _____; that is, the z value that provides an area of α in the lower tail of the standard normal distribution.

Summary

1. The p-value approach to hypothesis testing and the critical value approach will always lead to _____ rejection decision.
2. The advantage of the p-value approach is that the p-value tells us _____ the results are (the observed level of significance).
3. If we use the critical value approach, we only know that the results are significant _____.
4. We can use the same general approach to conduct an _____. The test statistic z is still computed using equation (9.1). But, for an upper tail test, the p-value is the probability of obtaining a value for the test statistic _____ _____ that provided by the sample.
5. To compute the p-value for the upper tail test in the σ known case, we must use the standard normal distribution to find the probability that z is _____ the value of the test statistic.
6. **Computation of p-Values for One-Tailed Tests**
   (a) Compute the value of the test statistic using equation (9.1):

   $$z = _____$$

   (b) *Lower tail test*: Using the standard normal distribution, compute the probability that z is _____ the value of the test statistic (area in the _____ tail).
   (c) *Upper tail test*: Using the standard normal distribution, compute the probability that z is _____ the value of the test statistic (area in the _____ tail).

Two-tailed Test

1. The general form for a two-tailed test about a population mean:

   _____

2. Example The U.S. Golf Association (USGA) establishes rules that manufacturers of golf equipment must meet if their products are to be acceptable for use in USGA events. MaxFlight Inc. uses a high-technology manufacturing process to produce golf balls with a mean driving distance of 295 yards. Sometimes, however, the process gets out of adjustment and produces golf balls with a mean driving distance different from 295 yards. When the mean distance falls below 295 yards, the company worries about losing sales because the golf balls do not provide as much distance as advertised. When the mean distance passes 295 yards, MaxFlight's golf balls may be rejected by the USGA for exceeding the overall distance standard concerning carry and roll. MaxFlight's quality control program involves taking periodic samples of 50 golf balls to monitor the manufacturing process. For each sample, a hypothesis test is conducted to determine whether the process has fallen out of adjustment.

   (a) We begin by assuming that the process is functioning correctly; that is, the golf balls being produced have a mean distance of 295 yards. This assumption establishes the null hypothesis. The alternative hypothesis is that the mean distance is not equal to 295 yards.

   _____.

   (b) If the sample mean $\bar{x}$ is _____ than 295 yards or _____ than 295 yards, we will reject H_0. In this case, corrective action will be taken to adjust the manufacturing process.

   (c) If $\bar{x}$ does not deviate from the hypothesized mean $\mu_0 = 295$ by a significant amount, H_0 will not be rejected and _____ will be taken to adjust the manufacturing process.

   (d) The quality control team selected $\alpha = 0.05$ as the level of significance for the test. Data from previous tests conducted when the process was known to be

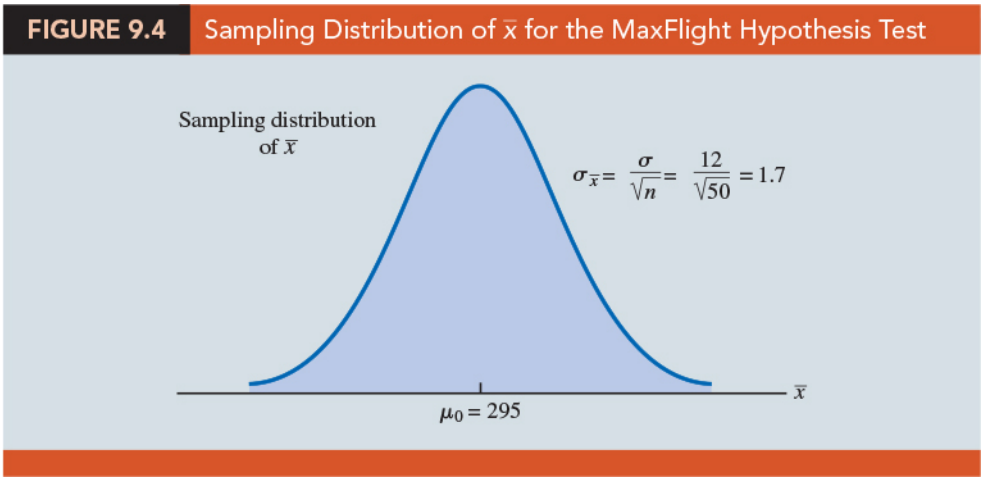
in adjustment show that the population standard deviation can be assumed known with a value of _____. Thus, with a sample size of $n = 50$, the standard error of $\bar{x}$ is

$$\sigma_{\bar{x}} = _____$$

(e) Because the sample size is large, the _____ allows us to conclude that the sampling distribution of $\bar{x}$ can be approximated by a _____.

(f) (Figure 9.4) the sampling distribution of $\bar{x}$ for the MaxFlight hypothesis test with a hypothesized population mean of $\mu_0 = 295$.

FIGURE 9.4 Sampling Distribution of $\bar{x}$ for the MaxFlight Hypothesis Test



3. Suppose that a sample of 50 golf balls is selected and that the sample mean is $\bar{x} = 297.6$ yards. This sample mean provides support for the conclusion that the population mean is larger than 295 yards. Is this value of $\bar{x}$ _____ 295 to cause us to reject H_0 at the 0.05 level of significance?

p-value approach

1. (Recall) the p-value is a probability used to determine whether the null hypothesis should be rejected.

2. For a two-tailed test, values of the test statistic _____ provide evidence against the null hypothesis.

3. For a two-tailed test, the p-value is the probability of obtaining a value for the test statistic _____ or _____ that provided by the sample.

4. Example the MaxFlight hypothesis test example.

   (a) *Compute the value of the test statistic.* For the σ known case, the test statistic z is a standard normal random variable. Using equation (9.1) with $\bar{x} = 297.6$, the value of the test statistic is

   $$z = \underline{\qquad\qquad}.$$

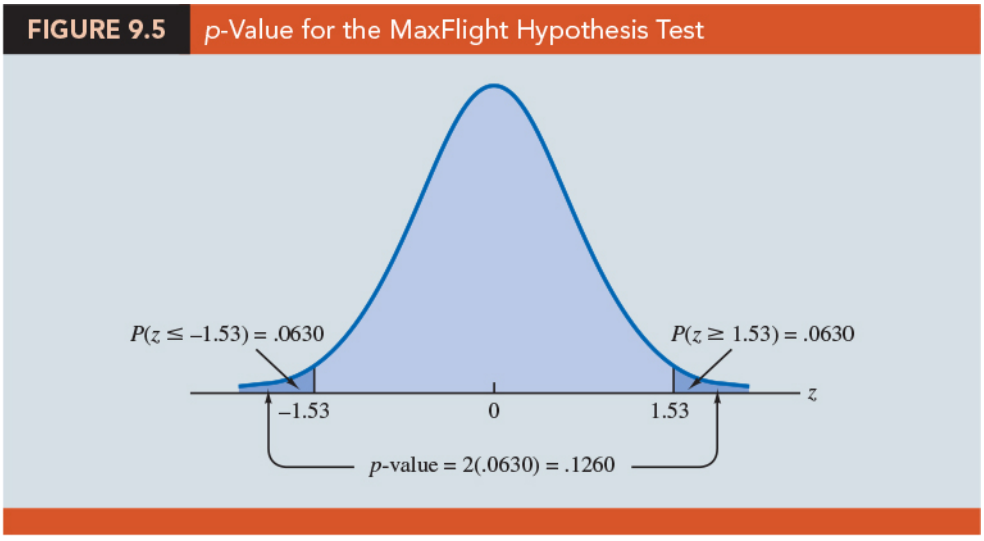
   (b) *Compute the p-value.* Find the probability of obtaining a value for the test statistic _____. Clearly values of _____ are at least as unlikely.

   (c) But, because this is a _____ test, values of _____ are also at least as unlikely as the value of the test statistic provided by the sample.

   (d) (Figure 9.5) the two-tailed p-value: _____.

   (e) Because the normal curve is symmetric, _____. Thus, the upper tail area is $P(z \geq 1.53) =$ _____.

FIGURE 9.5 *p*-Value for the MaxFlight Hypothesis Test



   (f) The p-value for the MaxFlight two-tailed hypothesis test is

   $$p\text{-value} = \underline{\qquad\qquad}.$$

(g) With a level of significance of $\alpha = 0.05$, we _____ because the _____. Because the null hypothesis is not rejected, no action will be taken to adjust the MaxFlight manufacturing process.

5. **Computation of p-Values for Two-Tailed Tests.**

   (a) Compute the value of the test statistic using equation (9.1).

   (b) If the value of the test statistic is in the _____, compute the probability that z is _____ the value of the test statistic (the upper tail area).

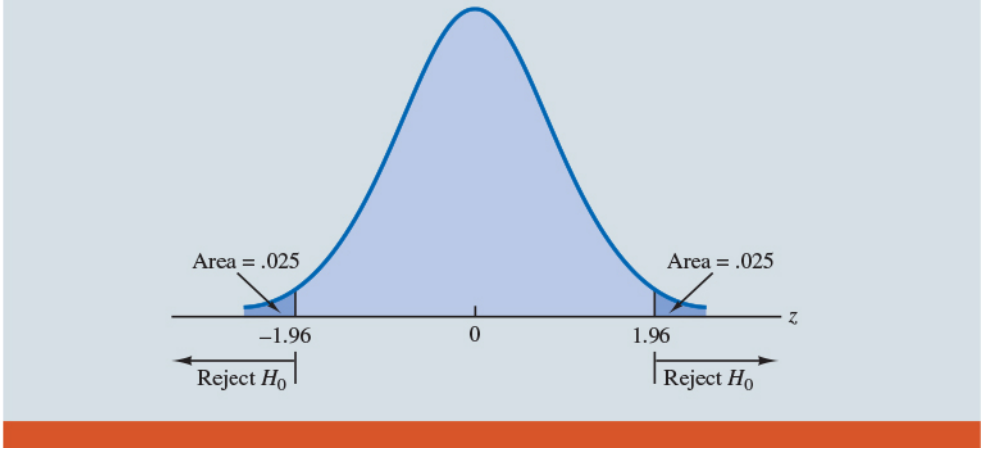
   (c) If the value of the test statistic is in the _____, compute the probability that z is _____ the value of the test statistic (the lower tail area).

   (d) Double the probability (or tail area) from step (b) or (c) to obtain the p-value.

Critical value approach

1. (Figure 9.6) the critical values for the test will occur in both the lower and upper tails of the standard normal distribution. With a level of significance of $\alpha = 0.05$, the area in each tail corresponding to the critical values is _____.

FIGURE 9.6 Critical Values for the MaxFlight Hypothesis Test



2. The critical values for the test statistic are _____ and _____.

3. The two-tailed rejection rule is

---

4. Because the value of the test statistic for the MaxFlight study is $z = 1.53$, the statistical evidence will not permit us to reject the null hypothesis at the 0.05 level of significance.

Summary and Practical Advice

1. Summary of the hypothesis testing procedures about a population mean for the σ known case. Note that μ_0 is the hypothesized value of the population mean.

TABLE 9.2 Summary of Hypothesis Tests About a Population Mean: σ Known Case

	Lower Tail Test	Upper Tail Test	Two-Tailed Test
Hypotheses	$H_0: \mu \geq \mu_0$ $H_a: \mu < \mu_0$	$H_0: \mu \leq \mu_0$ $H_a: \mu > \mu_0$	$H_0: \mu = \mu_0$ $H_a: \mu \neq \mu_0$
Test Statistic	$z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$	$z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$	$z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$
Rejection Rule: *p*-Value Approach	Reject H_0 if p-value $\leq \alpha$	Reject H_0 if p-value $\leq \alpha$	Reject H_0 if p-value $\leq \alpha$
Rejection Rule: Critical Value Approach	Reject H_0 if $z \leq -z_\alpha$	Reject H_0 if $z \geq z_\alpha$	Reject H_0 if $z \leq -z_{\alpha/2}$ or if $z \geq z_{\alpha/2}$

2. **Steps of Hypothesis Testing**

Step 1. Develop the null and alternative hypotheses (_____).

Step 2. Specify the level of significance (_____).

Step 3. Collect the sample data (_____) and compute the value of the test statistic (_____).

p-value Approach

Step 4. Use the value of the test statistic to compute the p-value.

Step 5. Reject H_0 if the _____.

Step 6. Interpret the _____ in the context of the application.

Critical Value Approach

Step 4. Use α to determine the critical value (_____ or _____) and the rejection rule.

Step 5. Use the value of the test statistic and the rejection rule to determine whether to reject H_0.

Step 6. Interpret the statistical conclusion in the context of the application.

3. Practical advice about the sample size for hypothesis tests is similar to the advice we provided about the sample size for interval estimation in Chapter 8.

   (a) In most applications, a sample size of _____ is adequate when using the hypothesis testing procedure described in this section.

   (b) If the population is _____ distributed, the hypothesis testing procedure that we described is _____ and can be used for any sample size.

   (c) If the population is not normally distributed but is at least _____, sample sizes _____ can be expected to provide acceptable results.

Relationship Between Interval Estimation and Hypothesis Testing

1. (Recall, Chapter 8) For the σ known case, the $(1-\alpha)\%$ confidence interval estimate of a population mean is given by

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

2. (Recall, Chapter 9) a two-tailed hypothesis test about a population mean:

$$H_0 : \mu = \mu_0, \quad H_a : \mu \neq \mu_0$$

   where μ_0 is the hypothesized value for the population mean.

3. Constructing a $100(1-\alpha)\%$ confidence interval for the population mean: _____ of the confidence intervals generated _____ the population mean and _____ of the confidence intervals generated _____ the population mean.

4. If we reject H_0 whenever the confidence interval does not contain μ_0, we will be rejecting H_0 when it is true ($\mu = \mu_0$) with probability α.

5. Recall that α is the probability of rejecting the null hypothesis when it is true.

6. So constructing a _____ confidence interval and rejecting H_0 whenever the interval does not contain μ_0 is _____ to conducting a _____ hypothesis test with _____ as the level of significance.

7. A Confidence Interval Approach to Testing a Hypothesis of the Form:

$$H_0 : \mu = \mu_0, \quad H_a : \mu \neq \mu_0$$

   (a) Select a simple random sample from the population and use the value of the sample mean $\bar{x}$ to develop the confidence interval for the population mean μ.

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

   (b) If the confidence interval contains the hypothesized value_____, do not reject H_0. Otherwise, reject H_0

8. Note that this discussion and example pertain to two-tailed hypothesis tests about a population mean. However, the same confidence interval and two-tailed hypothesis testing relationship exists for other population parameters.

9. The relationship can also be extended to one-tailed tests about population parameters. Doing so, however, requires the development of _____, which are rarely used in practice.

✎ Question ..... (p435)

The MaxFlight hypothesis test takes the following form:

$$H_0 : \mu = 295, \quad H_a : \mu \neq 295.$$

Conducting the MaxFlight hypothesis test with a level of significance of $\alpha = 0.05$ using the confidence interval approach.

sol:

- We sampled _____ golf balls and found a sample mean distance of _____ yards. Recall that the population standard deviation is _____ .
- The 95% confidence interval estimate of the population mean is

$$297.6 \pm 1.96\frac{12}{\sqrt{50}}$$
$$297.6 \pm 3.3 \quad or \quad (294.3, 300.9).$$

- With 95% confidence, the mean distance for the population of golf balls is between 294.3 and 300.9 yards.
- Because the interval contains the hypothesized value for the population mean, $\mu_0 = 295$, the hypothesis testing conclusion is that the null hypothesis, $H_0 : \mu = 295$, cannot be rejected.

9.4 Population Mean: σ Unknown

1. To conduct a hypothesis test about a population mean for the σ unknown case, the sample mean _____ is used as an estimate of _____ and the sample standard deviation _____ is used as an estimate of _____.

2. (Recall) For the σ known case, the sampling distribution of the test statistic has a _____ distribution. For the σ unknown case, however, the sampling distribution of the test statistic follows the _____; it has slightly more variability because the sample is used to develop estimates of both μ and σ.

3. **Test Statistic for Hypothesis Tests about a Population Mean: σ Unknown** the test statistic has a t distribution with $n-1$ degrees of freedom:

$$\frac{\qquad\qquad}{\qquad\qquad} \tag{9.2}$$

4. The t distribution is based on an assumption that the _____ from which we are sampling has a _____ distribution. However, research shows that this assumption can be relaxed considerably when the sample size is _____.

One-tailed Test

1. Example A business travel magazine wants to classify transatlantic gateway airports according to the mean rating for the population of business travelers. A rating scale with a low score of 0 and a high score of 10 will be used, and airports with a population mean rating greater than 7 will be designated as superior service airports. The magazine staff surveyed a sample of 60 business travelers at each airport to obtain the ratings data. The sample for London's Heathrow Airport provided a sample mean rating of $\bar{x} = 7.25$ and a sample standard deviation of $s = 1.052$. Do the data indicate that Heathrow should be designated as a superior service airport?

   (a) We want to develop a hypothesis test for which the decision to reject H_0 will lead to the conclusion that the population mean rating for the Heathrow Airport is greater than 7.

   (b) The null and alternative hypotheses for this upper tail test:

   _____

   (c) Use $\alpha = 0.05$, with $\bar{x} = 7.25$, $\mu_0 = 7$, $s = 1.052$, and $n = 60$, the value of the test statistic:

   $$t = \frac{\qquad\qquad}{\qquad\qquad}$$

(d) The sampling distribution of t has _____ degrees of freedom. Because the test is an upper tail test, the p-value is _____, that is, the upper tail area corresponding to the value of the test statistic.

(e) (Table 2 in Appendix B) the t distribution with 59 degrees of freedom provides the following information.

Area in Upper Tail	0.20	0.10	0.05	0.025	0.01	0.005
t-Value (59 df)	0.848	1.296	1.671	2.001	2.391	2.662

i. We see that $t = 1.84$ is between _____. The values in the " Area in Upper Tail" row show that the p-value must be less than _____ and greater than _____.

ii. With a level of significance of $\alpha = 0.05$, this placement is all we need to know to make the decision to reject the null hypothesis and conclude that Heathrow should be classified as a superior service airport.

(f) (Using software) $t = 1.84$ provides the upper tail p-value of _____ for the Heathrow Airport hypothesis test. With _____, we reject the null hypothesis and conclude that Heathrow should be classified as a superior service airport.

(g) The critical value corresponding to an area of $\alpha = 0.05$ in the upper tail of a t distribution with 59 degrees of freedom is _____.

(h) The rejection rule using the critical value approach is to reject H_0 if $t \geq 1.671$. Because _____, H_0 is rejected. Heathrow should be classified as a superior service airport.

Two-tailed Test

1. Example Consider the hypothesis testing situation facing Holiday Toys. The company manufactures and distributes its products through more than 1000 retail outlets. In planning production levels for the coming winter season, Holiday must decide how many units of each product to produce prior to knowing the actual demand at the retail level. For this year's most important new toy, Holiday′ s marketing director is expecting demand to average 40 units per retail outlet. Prior to making the final production decision based upon this estimate, Holiday decided

to survey a sample of 25 retailers in order to develop more information about the demand for the new product. Each retailer was provided with information about the features of the new toy along with the cost and the suggested selling price. Then each retailer was asked to specify an anticipated order quantity. With μ denoting the population mean order quantity per retail outlet, the sample data will be used to conduct the following two-tailed hypothesis test:

(a) If H_0 cannot be rejected, Holiday will continue its production planning based on the marketing director's estimate that the population mean order quantity per retail outlet will be $\mu = 40$ units.

(b) If H_0 is rejected, Holiday will immediately reevaluate its production plan for the product.

(c) A two-tailed hypothesis test is used because Holiday wants to reevaluate the production plan if the population mean quantity per retail outlet is less than anticipated or greater than anticipated.

(d) Because no historical data are available (it's a new product), the population mean μ and the population standard deviation must both be estimated using _____ from the sample data.

(e) The sample of 25 retailers provided a mean of $\bar{x} = 37.4$ and a standard deviation of $s = 11.79$ units.

(f) (*Check on the form of the population distribution*). The histogram of the sample data showed no evidence of skewness or any extreme outliers, so the analyst concluded that the use of the _____ with $n-1 = 24$ degrees of freedom was appropriate.

(g) Using equation (9.2) with $\bar{x} = 37.4$, $\mu_0 = 40$, $s = 11.79$, and $n = 25$, the value of the test statistic is

$$t = \underline{\qquad\qquad\qquad\qquad}.$$

(h) The t distribution table only contains positive t values. Because the t distribution is _____, however, the upper tail area at _____ is the same as the lower tail area at _____.

(i) (Table 2 in Appendix B)

Area in Upper Tail	0.20	0.10	0.05	0.025	0.01	0.005
t-Value (24 *df*)	0.857	1.318	1.711	2.064	2.492	2.797

(j) We see that $t = 1.10$ is between _____. From the "Area in Upper Tail" row, we see that the area in the upper tail at $t = 1.10$ is between _____.

(k) When we double these amounts, we see that the p-value must be between _____ and _____. With a level of significance of $\alpha = 0.05$, we now know that the p-value is greater than α. Therefore, H_0 cannot be rejected. Sufficient evidence is not available to conclude that Holiday should change its production plan for the coming season.

(l) (Software) The p-value obtained is _____. With a level of significance of $\alpha = 0.05$, we cannot reject H_0 because $0.2822 > 0.05$.

(m) With $\alpha = 0.05$ and the t distribution with 24 degrees of freedom, _____ and _____ are the critical values for the two-tailed test. The rejection rule using the test statistic is

_____.

(n) Based on the test statistic $t = -1.10$, H_0 cannot be rejected. This result indicates that Holiday should continue its production planning for the coming season based on the expectation that _____.

Summary and Practical Advice

1. (Table 9.3) A summary of the hypothesis testing procedures about a population mean for the σ unknown case.

TABLE 9.3 Summary of Hypothesis Tests About a Population Mean: σ Unknown Case

	Lower Tail Test	Upper Tail Test	Two-Tailed Test
Hypotheses	$H_0: \mu \geq \mu_0$ $H_a: \mu < \mu_0$	$H_0: \mu \leq \mu_0$ $H_a: \mu > \mu_0$	$H_0: \mu = \mu_0$ $H_a: \mu \neq \mu_0$
Test Statistic	$t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$	$t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$	$t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$
Rejection Rule: *p*-Value Approach	Reject H_0 if p-value $\leq \alpha$	Reject H_0 if p-value $\leq \alpha$	Reject H_0 if p-value $\leq \alpha$
Rejection Rule: Critical Value Approach	Reject H_0 if $t \leq -t_\alpha$	Reject H_0 if $t \geq t_\alpha$	Reject H_0 if $t \leq -t_{\alpha/2}$ or if $t \geq t_{\alpha/2}$

2. The applicability of the hypothesis testing procedures of this section is dependent on the _____ being sampled from and the _____.

9.5 Population Proportion

1. Using p_0 to denote the hypothesized value for the population proportion, the three forms for a hypothesis test about a population proportion:

_____

2. Hypothesis tests about a population proportion are based on the _____ between the sample proportion _____ and the hypothesized population proportion _____.

3. The _____ of $\bar{p}$, the point estimator of the population parameter p, is the basis for developing the test statistic.

4. When the null hypothesis is true as an equality, the expected value of p equals the hypothesized value p_0; that is, _____. The standard error of $\bar{p}$ is given by

_____

5. (Recall, Chapter 7) we said that if _____ and _____, the sampling distribution of $\bar{p}$ can be approximated by a _____ distribution. Under these conditions, which usually apply in practice, the quantity

$$z = \frac{\qquad\qquad}{\qquad\qquad} \tag{9.3}$$

has a standard normal probability distribution.

6. **Test Statistic for Hypothesis Tests About a Population Proportion**
With _____, the standard normal random variable z is the test statistic used to conduct hypothesis tests about a population proportion.

$$z = \frac{\qquad\qquad}{\qquad\qquad}$$

7. Example (Pine Creek golf course example). Over the past year, 20% of the players at Pine Creek were women. In an effort to increase the proportion of women players, Pine Creek implemented a special promotion designed to attract women golfers. One month after the promotion was implemented, the course manager requested a statistical study to determine whether the proportion of women players at Pine Creek had increased.

   (a) Because the objective of the study is to determine whether the proportion of women golfers increased, an upper tail test with _____ is appropriate:

   _____

   (b) If H_0 can be rejected, the test results will give statistical support for the conclusion that the proportion of women golfers increased and the promotion was beneficial.

(c) The course manager specified that a level of significance of _____ be used in carrying out this hypothesis test.

(d) The next step of the hypothesis testing procedure is to select a sample and compute the value of an appropriate test statistic. Suppose a random sample of $n = 400$ players was selected, and that $x = 100$ of the players were women. The proportion of women golfers in the sample is

_____

Using equation (9.4), the value of the test statistic is

$z =$

_____

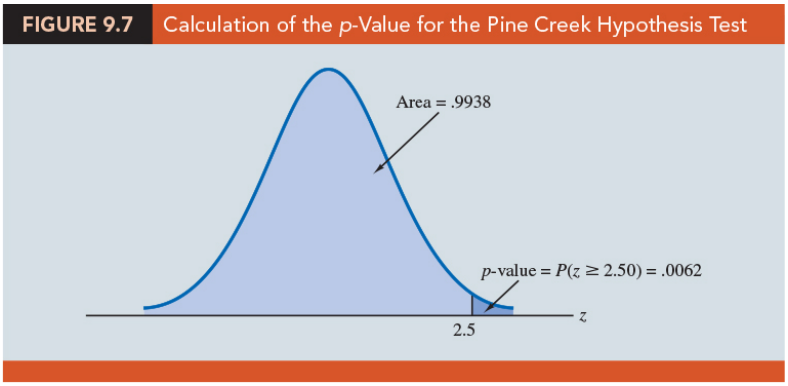
(e) *The p-value approach.*

i. The p-value is the probability that z is greater than or equal to _____. _____, the p-value for the Pine Creek test is _____

ii. (Figure 9.7) Recall that the course manager specified a level of significance of $\alpha = 0.05$. A _____ gives sufficient statistical evidence to reject H_0 at the 0.05 level of significance.

iii. The test provides statistical support for the conclusion that the special promotion increased the proportion of women players at the Pine Creek golf course.

FIGURE 9.7 Calculation of the p-Value for the Pine Creek Hypothesis Test



(f) *The critical value approach.* The critical value corresponding to an area of 0.05 in the upper tail of a normal probability distribution is _____.

Thus, the rejection rule using the critical value approach is to reject H_0 if _____. Because _____, H_0 is rejected.

(g) The p-value approach provides more information. With a p-value $= 0.0062$, the null hypothesis would be rejected for any level of significance _____ _____.

Summary

1. The procedure used to conduct a hypothesis test about a _____ is similar to the procedure used to conduct a hypothesis test about a _____.

2. Although we only illustrated how to conduct a hypothesis test about a population proportion for an upper tail test, similar procedures can be used for _____ and _____.

3. (Table 9.4) A summary of the hypothesis tests about a population proportion. We assume that $np \geq 5$ and $n(1-p) \geq 5$; thus the _____ probability distribution can be used to approximate the sampling distribution of $\bar{p}$.

TABLE 9.4 Summary of Hypothesis Tests About a Population Proportion

	Lower Tail Test	Upper Tail Test	Two-Tailed Test
Hypotheses	$H_0: p \geq p_0$ $H_a: p < p_0$	$H_0: p \leq p_0$ $H_a: p > p_0$	$H_0: p = p_0$ $H_a: p \neq p_0$
Test Statistic	$z = \dfrac{\bar{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$	$z = \dfrac{\bar{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$	$z = \dfrac{\bar{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$
Rejection Rule: *p*-Value Approach	Reject H_0 if p-value $\leq \alpha$	Reject H_0 if p-value $\leq \alpha$	Reject H_0 if p-value $\leq \alpha$
Rejection Rule: Critical Value Approach	Reject H_0 if $z \leq -z_\alpha$	Reject H_0 if $z \geq z_\alpha$	Reject H_0 if $z \leq -z_{\alpha/2}$ or if $z \geq z_{\alpha/2}$

9.6 Hypothesis Testing and Decision Making

1. The hypothesis testing applications are considered as _____:
   (a) formulate the null and alternative hypotheses, H_0, H_a.
   (b) specify the level of significance, α.
   (c) select a sample, $\mathbf{x} = \{x_1, x_2, \cdots, x_n\}$.
   (d) compute the value of a test statistic, $T(\mathbf{x})$.
   (e) compute the associated p-value.
   (f) compare the p-value to α.
   (g) conclude "reject H_0" and declare the results significant if p-value $\leq \alpha$; otherwise, we made the conclusion "do not reject H_0."

2. With a significance test, we control the probability of making the Type I error, but not the Type II error. Thus, we recommended the conclusion _____ rather than _____ because the latter puts us at _____ of making the Type II error of accepting H_0 when it is false.

3. With the conclusion "do not reject H_0," the statistical evidence is considered _____ and is usually an indication to _____ until further research and testing can be undertaken.

4. If the purpose of a hypothesis test is to _____ when H_0 is true and a _____ when H_a is true, the decision maker may want to, and in some cases be forced to, take action with both the conclusion do not reject H_0 and the conclusion reject H_0. If this situation occurs, statisticians generally recommend controlling the probability of making a _____.

5. With the probabilities of both the Type I and Type II error controlled, the conclusion from the hypothesis test is _____.

6. Example (lot-acceptance example) A quality control manager must decide to accept a shipment of batteries from a supplier or to return the shipment because of poor quality.

(a) Assume that design specifications require batteries from the supplier to have a mean useful life of at least 120 hours. To evaluate the quality of an incoming shipment, a sample of 36 batteries will be selected and tested.

(b) On the basis of the sample, a decision must be made to accept the shipment of batteries or to return it to the supplier because of poor quality.

(c) Let μ denote the mean number of hours of useful life for batteries in the shipment. The null and alternative hypotheses about the population mean:

---

i. If H_0 is rejected, the alternative hypothesis is concluded to be true. This conclusion indicates that the appropriate action is to _____ the shipment to the supplier.

ii. If H_0 is not rejected, the decision maker must still determine what action should be taken. Thus, without directly concluding that H_0 is true, but merely by not rejecting it, the decision maker will have made the decision to _____ the shipment as being of satisfactory quality.

(d) In such decision-making situations, it is recommended that the hypothesis testing procedure be extended to control the probability of making a Type II error.

7. Because a decision will be made and action taken when we do not reject H_0, knowledge of the probability of making a _____ will be helpful.

9.7 Calculating The Probability of Type II Errors

1. Example (lot-acceptance example) The null and alternative hypotheses about the mean number of hours of useful life for a shipment of batteries:

$$H_0 : \mu \geq 120, \quad H_a : \mu < 120.$$

(a) If H_0 is rejected, the decision will be to _____ the shipment to the supplier because the mean hours of useful life are less than the specified 120 hours.

(b) If H_0 is not rejected, the decision will be to _____ the shipment.

2. Suppose a level of significance of $\alpha = 0.05$, the test statistic in the σ known case:

$$z = \underline{\qquad\qquad\qquad\qquad}$$

3. The rejection rule for the lower tail test:

_____

4. Suppose a sample of $n = 36$ batteries will be selected and based upon previous testing the population standard deviation can be assumed known with a value of $\sigma = 12$ hours.

5. The rejection rule indicates that we will reject H_0 if

$$z = \underline{\qquad\qquad\qquad\qquad}$$

6. Solving for $\bar{x}$ in the preceding expression indicates that we will reject H_0 if

_____

7. Rejecting H_0 when $\bar{x} \leq 116.71$ means that we will make the decision to accept the shipment whenever $\bar{x} > 116.71$.

8. Compute probabilities associated with making a Type II error.

(a) (Recall) we make a Type II error whenever the true shipment mean is less than 120 hours and we make the decision to accept $H_0 : \mu \geq 120$.

(b) Hence, to compute the probability of making a Type II error, we must select a value of _____.

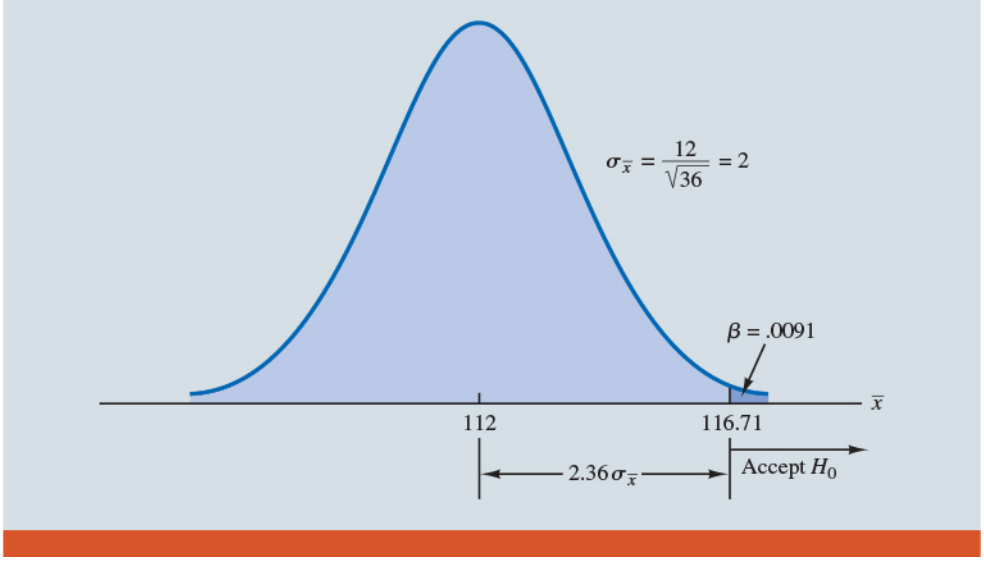
(c) For example, suppose the shipment is considered to be of poor quality if the batteries have a mean life of $\mu = 112$ hours.

(d) If $\mu = 112$ is really true, what is the probability of accepting $H_0 : \mu \geq 120$ and hence committing a Type II error?

_____.

(e) (Figure 9.8) the sampling distribution of $\bar{x}$ when the mean is $\mu = 112$. The shaded area in the upper tail gives the probability of obtaining _____.

FIGURE 9.8 Probability of a Type II Error When $\mu = 112$



(f) Using the standard normal distribution, we see that at $\bar{x} = 116.71$

$z =$ _____

(g) The probability of making a Type II error when $\mu = 112$ is _____.

(h) Therefore, we can conclude that if the mean of the population is 112 hours, the probability of making a Type II error is only 0.0091.

(i) We can repeat these calculations for other values of μ less than 120.

9. (Table 9.5) we show the probability of making a Type II error for a variety of values of μ less than 120. Note that as μ increases toward 120, the probability of making a Type II error increases toward an upper bound of 0.95. However, as μ decreases to values farther below 120, the probability of making a Type II error diminishes.

TABLE 9.5 Probability of Making a Type II Error for the Lot-Acceptance Hypothesis Test

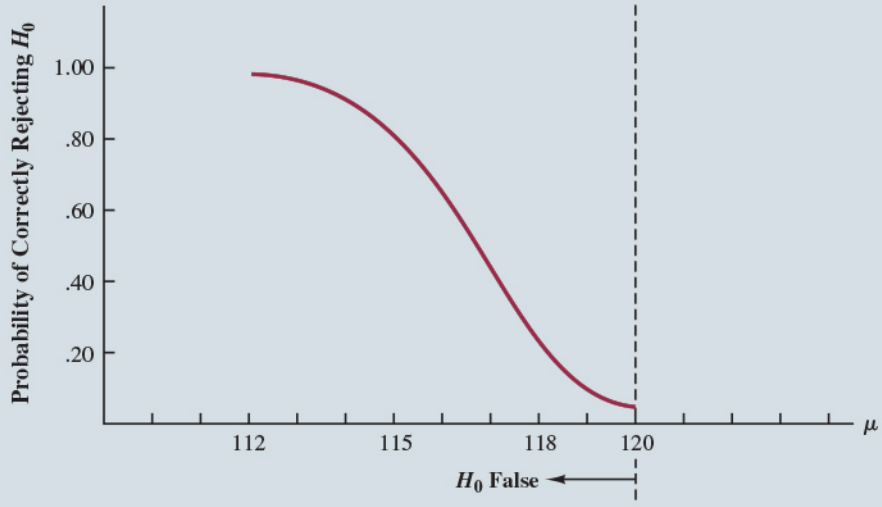
Value of μ	$z = \frac{116.71 - \mu}{12/\sqrt{36}}$	Probability of a Type II Error (β)	Power ($1 - \beta$)
112	2.36	.0091	.9909
114	1.36	.0869	.9131
115	.86	.1949	.8051
116.71	.00	.5000	.5000
117	−.15	.5596	.4404
118	−.65	.7422	.2578
119.999	−1.645	.9500	.0500

10. When the true population mean μ is _____ the null hypothesis value of $\mu = 120$, the probability is _____ that we will make a Type II error.

11. For any particular value of μ, the _____; that is, the probability of _____ is 1 minus the probability of making a Type II error.

12. (Figure 9.9) *Power curve*: the power associated with each value of μ:

FIGURE 9.9 Power Curve for the Lot-Acceptance Hypothesis Test



(a) Note that the power curve extends over the values of μ for which the _____.

(b) The _____ of the power curve at any value of μ indicates the probability of correctly rejecting H_0 when H_0 is false.

13. **The step-by-step procedure to compute the probability of making a Type II error in hypothesis tests about a population mean**

(a) Formulate the null and alternative hypotheses.

(b) Use the level of significance α and the critical value approach to determine the critical value and the _____ for the test.

(c) Use the rejection rule to solve for the value of the _____ corresponding to the critical value of the test statistic.

(d) Use the results from step (c) to state the values of the sample mean that lead to the acceptance of H_0. These values define the _____ for the test.

(e) Use the sampling distribution of $\bar{x}$ for a value of μ satisfying H_a, and the acceptance region from step (d), to compute the probability that the sample mean will be in the acceptance region.

(f) This probability is the probability of making a Type II error at the chosen value of μ.

9.8 Determining The Sample Size for a Hypothesis Test about a Population Mean

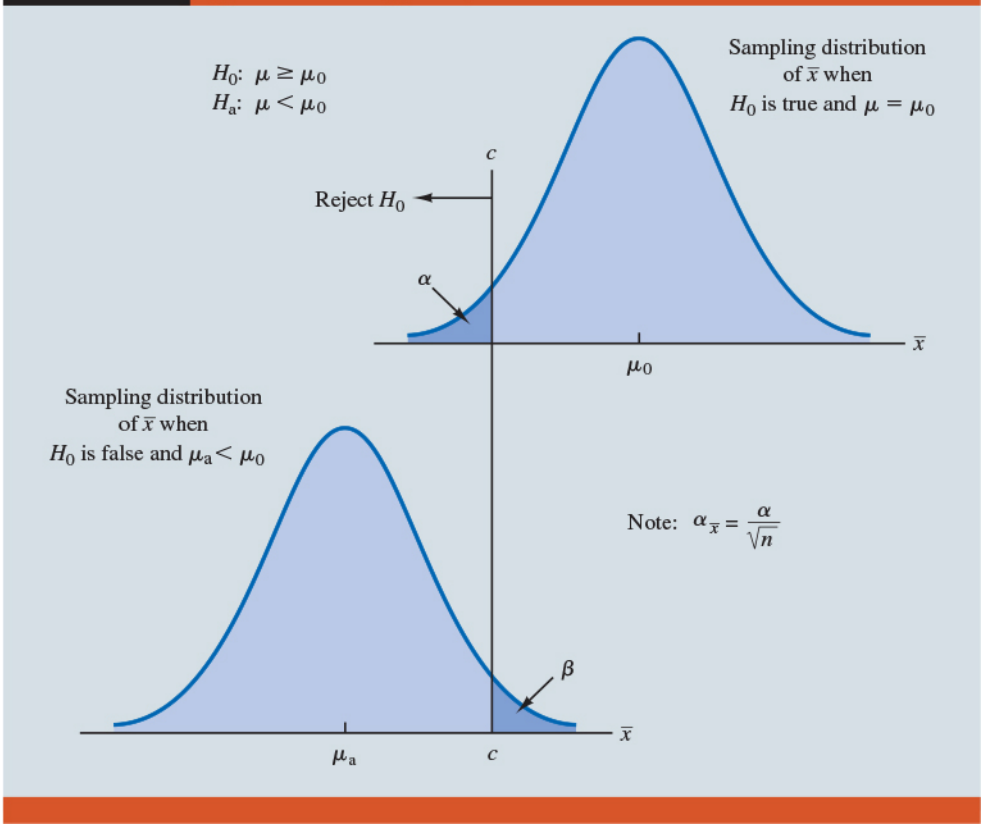
1. Assume that a hypothesis test is to be conducted about the value of a population mean. The level of significance specified by the user determines the probability of making a Type I error for the test. By controlling the _____, the user can also control the probability of making a _____.

2. Let us show how a sample size can be determined for the following lower tail test about a population mean.

_____

3. (Figure 9.10) The upper panel is the sampling distribution of $\bar{x}$ when H_0 is true with $\mu = \mu_0$.

FIGURE 9.10 Determining the Sample Size for Specified Levels of the Type I (α) and Type II (β) Errors



4. For a lower tail test, the critical value of the test statistic is denoted _____. In the upper panel of the figure the vertical line, _____, is the corresponding value of _____.

5. If we reject H_0 when $\bar{x} \leq c$, the probability of a Type I error will be α: _____.

6. With z_α representing the z value corresponding to an area of α in the _____ of the standard normal distribution, we compute c using the following formula:

$$c = _____$$

7. (Figure 9.10) The lower panel is the sampling distribution of $\bar{x}$ when the alternative hypothesis is true with _____. The shaded region shows _____, the probability of a Type II error that the decision maker will be exposed to if the null hypothesis is accepted when $\bar{x} > c$.

8. With z_β representing the z value corresponding to an area of β in the upper tail of the standard normal distribution, we compute c using the following formula:

$$c = \qquad (9.6)$$

9. Now what we want to do is to select a value for c so that when we _____ and _____, the probability of a Type I error is equal to the chosen value of _____ and the probability of a Type II error is equal to the chosen value of _____. Therefore, both equations (9.5) and (9.6) must provide the same value for c:

$$\begin{aligned} \mu_0 - z_\alpha \frac{\sigma}{\sqrt{n}} &= \mu_a + z_\beta \frac{\sigma}{\sqrt{n}} \\ \mu_0 - \mu_a &= z_\alpha \frac{\sigma}{\sqrt{n}} + z_\beta \frac{\sigma}{\sqrt{n}} \\ \mu_0 - \mu_a &= \frac{(z_\alpha + z_\beta)\sigma}{\sqrt{n}} \\ \sqrt{n} &= \frac{(z_\alpha + z_\beta)\sigma}{(\mu_0 - \mu_a)} \end{aligned}$$

10. **Sample Size for a One-Tailed Hypothesis Test About a Population Mean**

$$n = \qquad (9.7)$$

where

- $z_\alpha = z$ value providing an area of α in the upper tail of a standard normal distribution.
- $z_\beta = z$ value providing an area of β in the upper tail of a standard normal distribution.
- $\sigma =$ the population standard deviation.
- $\mu_0 =$ the value of the population mean in the null hypothesis.
- $\mu_\alpha =$ the value of the population mean used for the Type II error.

Note: In a two-tailed hypothesis test, use (9.7) with _____ replacing _____.

11. Example lot-acceptance example

    (a) The design specification for the shipment of batteries indicated a mean useful life of at least 120 hours for the batteries. Shipments were rejected if $H_0 : \mu \geq 120$ was rejected.

    (b) Let us assume that the quality control manager makes the following statements about the allowable probabilities for the Type I and Type II errors.

        i. *Type I error statement*: If the mean life of the batteries in the shipment is $\mu = 120$, I am willing to risk an $\alpha = 0.05$ probability of rejecting the shipment.

        ii. *Type II error statement*: If the mean life of the batteries in the shipment is 5 hours under the specification (i.e., $\mu = 115$), I am willing to risk a $\beta = 0.10$ probability of accepting the shipment.

        iii. These statements are based on the judgment of the manager. Someone else might specify different restrictions on the probabilities. However, statements about the allowable probabilities of both errors must be made before the sample size can be determined.

    (c) In the example, _____ and _____. Using the standard normal probability distribution, we have _____ and _____. From the statements about the error probabilities, we note that _____ and _____. Finally, the population standard deviation was assumed known at _____.

    (d) The recommended sample size for the lot-acceptance example is

    $$n = \frac{\qquad\qquad}{\qquad\qquad} = \frac{\qquad\qquad\qquad\qquad\qquad\qquad}{\qquad\qquad\qquad\qquad\qquad\qquad}$$

    Rounding up, we recommend a sample size of 50.

    (e) Because both the Type I and Type II error probabilities have been _____ at allowable levels with $n = 50$, the quality control manager is now justified in using the accept H_0 and reject H_0 statements for the hypothesis test.

12. We can make three observations about the relationship among α, β, and the sample size n.

(a) Once two of the three values are known, the other can be computed.

(b) For a given α, increasing _____ will reduce _____.

(c) For a given n, decreasing _____ will increase _____, whereas increasing _____ will decrease _____.

13. The third observation should be kept in mind when the probability of a Type II error is not being controlled. It suggests that one should not choose _____ for the level of significance _____.

14. For a given sample size, choosing a smaller level of significance means more exposure to a Type II error. Inexperienced users of hypothesis testing often think that smaller values of α are always better. They are better if we are concerned only about making a Type I error. However, smaller values of α have the disadvantage of increasing the probability of making a _____.

9.9 Big Data And Hypothesis Testing*

😍 EXERCISES

9.1 : 2, 3

9.2 : 6, 7

9.3 : 9, 11, 15, 18, 22

9.4 : 23, 24, 27, 33

9.5 : 35, 36, 43, 44

9.7 : 46, 50, 53

9.8 : 55, 59

SUP : 69, 79, 83

名言佳句

“成功是歷經一個又一個的失敗卻不失去熱情。”
“Success is the ability to go from failure to failure without losing your enthusiasm.”
— *Winston Churchill (November 30, 1874 – January 24, 1965)*

統計小故事：費雪與品茶實驗

二十世紀初，統計學正處於快速發展的階段。從農業實驗到醫學研究，越來越多的科學家意識到，他們需要一種方法來回答這樣的問題：某種觀察結果，究竟是因為真實效應，還是僅僅由於隨機偶然？這便是假設檢定（hypothesis testing）的起點。

在這段歷史中，最著名的故事之一，便是英國統計學家羅納德．費雪（Ronald A. Fisher）設計的「品茶實驗」。

據說，某個午後，在劍橋大學的花園聚會上，一位女士聲稱她能分辨出一杯茶究竟是「先加牛奶」還是「先加紅茶」。這個說法引來了旁人的懷疑與調侃，但費雪卻將此視為一個絕佳的科學問題。他並沒有直接質疑女士的說法，而是提出要透過一個嚴謹的實驗來檢驗。

費雪設計的實驗非常精巧：他準備了八杯茶，其中四杯先加牛奶、四杯先加紅茶，隨機排列，然後請這位女士分辨。費雪的邏輯是，如果女士只是隨機猜測，那麼正確辨認八杯的機率極低；若她能顯著超越隨機水準，就可以認為她的判斷有真實依據。

這個設計引出了統計學中的核心概念——虛無假設（null hypothesis）與對立假設（alternative hypothesis）。虛無假設代表「女士無法分辨，只是隨機猜測」；對立假設則是「女士具有分辨能力」。透過計算在虛無假設下出現觀察結果的機率，費雪能判斷是否該拒絕虛無假設。

這裡涉及另一個關鍵概念——顯著水準（significance level）。假設我們設定顯著水準為 5%，那麼只要女士的表現達到在隨機情況下少於 5% 的可能性，我們就有理由拒絕虛無假設。這種方法提供了一個客觀標準，避免了「憑感覺判斷」的爭議。

實驗結果顯示，這位女士確實能正確分辨所有茶杯。這不僅是一次花園趣事，更是統計史上的里程碑。費雪藉由這個小實驗，展示了假設檢定的力量，並在他的重要著作《The Design of Experiments》（1935）中，將這套方法系統化，成為科學研究的標準。

假設檢定的意義遠遠超越「喝茶」。在農業領域，費雪利用假設檢定設計田間實驗，檢驗不同肥料對作物產量的影響；在醫學領域，研究者利用假設檢定判斷新藥是否優於舊藥或安慰劑；在社會科學中，學者利用假設檢定來檢驗政策是否對不同族群造成差異影響。

然而，假設檢定也不是完美無缺。它帶來了第一類錯誤（Type I error）與第二類錯誤（Type II error）的概念。第一類錯誤指的是「錯誤地拒絕了真實的虛無假設」，也就是把一個偶然當作真實效應；第二類錯誤則是「錯誤地接受了虛無假設」，也就是錯

過了真實的效應。這提醒我們，假設檢定不是絕對的真理，而是一種在不確定性下作出理性判斷的方法。

在教材呼應上，本章的重點是讓學生理解假設檢定的基本流程：建立虛無假設與對立假設。選擇顯著水準 (如 5%)。計算統計量並找出 p 值。根據結果判斷是否拒絕虛無假設。這一流程看似簡單，但其背後正是費雪的智慧：將「懷疑與驗證」轉化為一種數學化的決策框架。

因此，當學生在課堂上計算「某新藥是否顯著優於舊藥」時，請記住，他們所做的並不只是代入公式，而是在延續一場發生於劍橋花園的歷史實驗。那場小小的品茶挑戰，讓統計學從描述數據的工具，轉變為檢驗真理的科學。

總結來說，假設檢定教會我們：科學不是為了證明假設為真，而是透過數據檢驗，看看證據是否足以拒絕「無效假設」。在這個過程中，我們接受不確定性，但依然能做出理性判斷。這正是假設檢定的價值所在。

統計學 (一)

Anderson's Statistics for Business & Economics (14/E)

Chapter 10: Inference About Means and Proportions with Two Populations

上課時間地點: 二 D56, 商館 260206
授課教師: 吳漢銘 (國立政治大學統計學系副教授)
教學網站: https://hmwu.idv.tw

系級: _____ 學號: _____ 姓名: _____

Overview

1. Discuss the statistical inference (_____ and _____) for two population means (three situations: population standard deviations known, unknown; match samples) and the two population proportions.
2. *Examples*:
   (a) Develop an interval estimate of the difference between the mean starting salary for a population of men and the mean starting salary for a population of women.
   (b) Conduct a hypothesis test to determine whether any difference is present between the proportion of defective parts in a population of parts produced by supplier A and the proportion of defective parts in a population of parts produced by supplier B.

10.1 Inferences About the Difference Between Two Population Means: σ_1 and σ_2 Known

1. _____ denote the mean of population 1 (2), we will focus on inferences about the difference between the means: _____.
2. A simple random sample of _____ units from population 1 (2). The two samples, taken separately and independently, are referred to as _____ simple random samples.
3. Assume the two population standard deviations, _____, can be assumed _____ to collecting the samples.
4. Question: how to compute a _____ and develop an _____ of the difference between the two population means when σ_1 and σ_2 are known.

Interval Estimation of $\mu_1 - \mu_2$

1. Example **Greystone Department Stores**, Inc., operates two stores in Buffalo, New York: One is in the inner city and the other is in a suburban shopping center. The regional manager noticed that products that sell well in one store do not always sell well in the other. The manager believes this situation may be attributable to differences in customer demographics at the two locations. Customers may differ in age, education, income, and so on. (對觀察事物提出問題)
2. Suppose the manager asks us to investigate the difference between the _____ of the customers who shop at the two stores. (針對問題收集資料)
3. Let us define population 1 as all customers who shop at the _____ and population 2 as all customers who shop at the _____.
   (a) _____: mean of population 1 (i.e., the mean age of all customers who shop at the inner-city store)
   (b) _____: 5 mean of population 2 (i.e., the mean age of all customers who shop at the suburban store)

4. The difference between the two population means is _____. To estimate $\mu_1-\mu_2$, we will select a simple random sample of _____ customers from population 1 and a simple random sample of _____ customers from population 2.

5. We then compute the two sample means.

   (a) _____: sample mean age for the simple random sample of n_1 inner-city customers

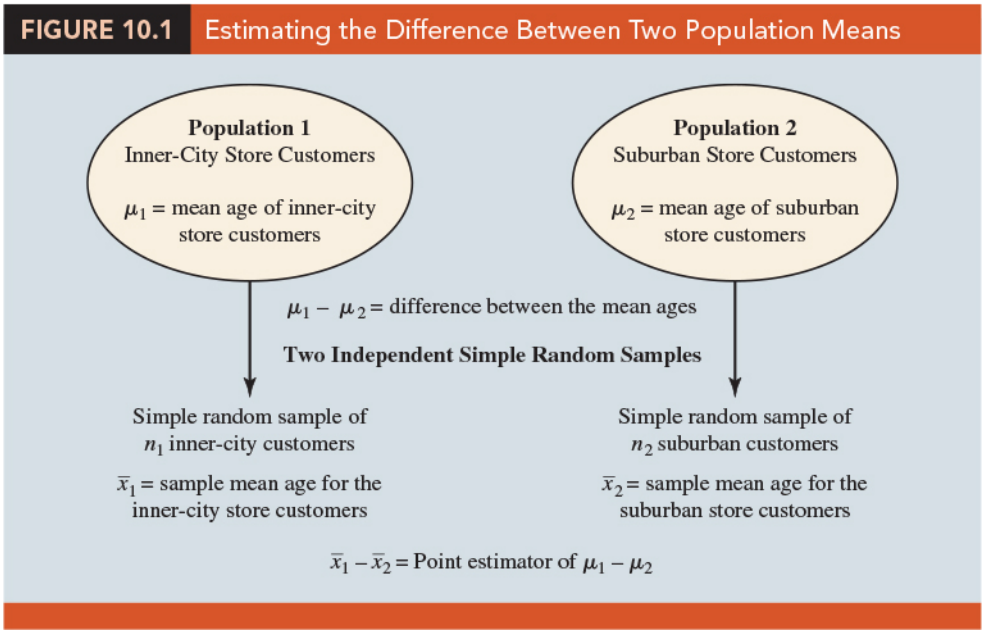
   (b) _____: sample mean age for the simple random sample of n_2 suburban customers

6. The point estimator of the difference between the two _____ is the difference between the two _____.

7. **Point Estimator of the Difference Between Two Population Means**

   _____ (10.1)

8. (Figure 10.1) the process used to estimate the difference between two population means based on two independent simple random samples.



9. The point estimator $\bar{x}_1-\bar{x}_2$ has a standard error that describes the _____ in the sampling distribution of the estimator.

10. **Standard Error of $\bar{x}_1 - \bar{x}_2$** With two independent simple random samples, the standard error of $\bar{x}_1 - \bar{x}_2$ is as follows:

(10.2)

(證明如下:) (Hint: $Var(aX + bY) = a^2Var(X) + b^2Var(Y) + 2abCov(X, Y)$).

11. If both populations have a _____ distribution, or if the sample sizes are large enough that the _____ enables us to conclude that the sampling distributions of $\bar{x}_1$ and $\bar{x}_2$ can be approximated by a normal distribution, the sampling distribution of $\bar{x}_1 - \bar{x}_2$ will have a _____ distribution with mean given by _____. (Denoted by _____)

12. In general, an interval estimate is given by a point estimate ± a margin of error. In the case of estimation of the difference between two population means, an interval estimate will take the following form:

13. With the sampling distribution of $\bar{x}_1 - \bar{x}_2$ having a normal distribution, we can write the margin of error as follows:

$$\text{Margin of error} = _____ = _____ \quad (10.3)$$

14. **Interval Estimate of the Difference Between Two Population Means: σ_1 and σ_2 Known**

(10.4)

where $1-\alpha$ is the confidence coefficient.

(公式說明如下:)

✎ Question ..... (p485)

Example Greystone example. Based on data from previous customer demographic studies, the two population standard deviations are known with $\sigma_1 = 9$ years and $\sigma_2 = 10$ years. The data collected from the two independent simple random samples of Greystone customers provided the following results.

	Inner City Store	Suburban Store
Sample Size	$n_1 = 36$	$n_2 = 49$
Sample Mean	$\bar{x}_1 = 40$ years	$\bar{x}_2 = 35$ years

Find the margin of error and the 95% confidence interval estimate of the difference between the two population means.

sol:

Hypothesis Tests About $\mu_1 - \mu_2$

1. Let us consider hypothesis tests about the difference between two population means. Using _____ to denote the hypothesized difference between μ_1 and μ_2, the three

forms for a hypothesis test are as follows:

Left-tailed test	Right-tailed test	Two-tailed test
H_0 : _____	H_0 : _____	H_0 : _____
H_a : _____	H_a : _____	H_a : _____

2. In many applications, _____. Using the two-tailed test as an example, when $D_0 = 0$ the null hypothesis is $H_0 : \mu_1 - \mu_2 = 0$.

3. In this case, the null hypothesis is that μ_1 and μ_2 are equal. Rejection of H_0 leads to the conclusion that $H_a : \mu_1 - \mu_2 \neq 0$ is true; that is, μ_1 and μ_2 are not equal.

4. The general steps for conducting hypothesis tests: choose a _____, compute the value of the _____, and find the _____ to _____ whether the null hypothesis should be rejected.

5. With two independent simple random samples, we showed that the point estimator $\bar{x}_1 - \bar{x}_2$ has a standard error $\sigma_{\bar{x}_1 - \bar{x}_2}$ given by expression (10.2) and, when the sample sizes are large enough, the distribution of $\bar{x}_1 - \bar{x}_2$ can be described by a _____ distribution.

6. **Test Statistic for Hypothesis Tests About $\mu_1 - \mu_2$: σ_1 and σ_2 Known**

(10.5)

_____

7. We demonstrated a two-tailed hypothesis test about the difference between two population means. Lower tail and upper tail tests can also be considered. These tests use the _____ as given in equation (10.5). The procedure for computing the p-value and the rejection rules for these one-tailed tests are the same as those for hypothesis tests involving a single population mean and single population proportion.

Question (p486)

As part of a study to evaluate differences in education quality between two training centers, a standardized examination is given to individuals who are trained at the centers. The difference between the mean examination scores is used to assess quality differences between the centers. The population means for the two centers are as follows. μ_1 is the mean examination score for the population of individuals trained at center A, μ_2 is the mean examination score for the population of individuals trained at center B. We begin with the tentative assumption that no difference exists between the training quality provided at the two centers. The standardized examination given previously in a variety of settings always resulted in an examination score standard deviation near 10 points. Thus, we will use this information to assume that the population standard deviations are known with $\sigma_1 = 10$ and $\sigma_2 = 10$. An $\alpha = 0.05$ level of significance is specified for the study. Independent simple random samples of $n_1 = 30$ individuals from training center A and $n_2 = 40$ individuals from training center B are taken. The respective sample means are $\bar{x}_1 = 82$ and $\bar{x}_2 = 78$. Do these data suggest a significant difference between the population means at the two training centers? State the null and alternative hypotheses for this two-tailed test, compute the test statistic, and state the decision rules based on the p-value approach and the critical value approach and make the decision.

sol:

Practical Advice

1. In most applications of the interval estimation and hypothesis testing procedures presented in this section, random samples with _____ and _____ are adequate.

2. In cases where either or both sample sizes are less than 30, the _____ of the populations become important considerations.

3. In general, with smaller sample sizes, it is more important for the analyst to be satisfied that it is reasonable to assume that the distributions of the two populations are at least _____.

10.2 Inferences About The Difference Between Two Population Means: σ_1 and σ_2 Unknown

1. Extend the discussion of inferences about the difference between two population means to the case when the two population standard deviations, _____ and _____, are _____.

2. In this case, we will use the sample standard deviations, _____ and _____, to estimate the unknown population standard deviations.

3. When we use the sample standard deviations, the interval estimation and hypothesis testing procedures will be based on the _____ rather than the standard normal distribution.

Interval Estimation of $\mu_1 - \mu_2$

1. Let us develop the margin of error and an interval estimate of the difference between these two population means. (Recall) The interval estimate for the case when the

population standard deviations, σ_1 and σ_2, are known.

2. With σ_1 and σ_2 unknown, we will use the sample standard deviations s_1 and s_2 to estimate _____ and replace $z_{\alpha/2}$ with _____.

3. **Interval Estimate of the Difference Between Two Population Means: σ_1 and σ_2 Unknown**

(10.6)

where $1 - \alpha$ is the confidence coefficient.

4. In this expression, the use of the t distribution is an _____, but it provides excellent results and is relatively easy to use. The only difficulty that we encounter in using expression (10.6) is determining the appropriate _____ for $t_{\alpha/2}$.

5. Statistical software packages compute the appropriate degrees of freedom automatically. The formula used is as follows:
Degrees of Freedom: t Distribution With Two Independent Random Samples

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2} \qquad (10.7)$$

✎ Question ..... (p490)

(**Clearwater National Bank example**) Clearwater National Bank is conducting a study designed to identify differences between checking account practices by customers at two of its branch banks. A simple random sample of 28 checking accounts is selected from the Cherry Grove Branch and an in dependent simple random sample of 22 checking accounts is selected from the Beechmont Branch. The current checking account balance is recorded for each of the checking accounts. A summary of the account balances follows:

	Cherry Grove	Beechmont
Sample Size	$n_1 = 28$	$n_2 = 22$
Sample Mean	$\bar{x}_1 = \$1025$	$\bar{x}_2 = \$910$
Sample Standard Deviation	$s_1 = \$150$	$s_2 = \$125$

Clearwater National Bank would like to estimate the difference between the mean checking account balance maintained by the population of Cherry Grove customers and the population of Beechmont customers. Compute a 95% confidence interval estimate of the difference between the population mean checking account balances at the two branch banks.

sol:

Hypothesis Tests About $\mu_1 - \mu_2$

1. (Recall) Letting D_0 denote the hypothesized difference between μ_1 and μ_2, the test statistic used for the case where σ_1 and σ_2 are known is as follows.

   The test statistic, z, follows the standard normal distribution.

2. When σ_1 and σ_2 are unknown, we use s_1 as an estimator of σ_1 and s_2 as an estimator of σ_2. Substituting these sample standard deviations for σ_1 and σ_2 provides the following test statistic when σ_1 and σ_2 are unknown.

3. **Test Statistic for Hypothesis Tests About $\mu_1 - \mu_2$: σ_1 and σ_2 Unknown**

(10.8)

The degrees of freedom for t are given by equation (10.7).

Question (p491)

Consider a new computer software package developed to help systems analysts reduce the time required to design, develop, and implement an information system. To evaluate the benefits of the new software package, a random sample of 24 systems analysts is selected. Each analyst is given specifications for a hypothetical information system. Then 12 of the analysts are instructed to produce the information system by using current technology. The other 12 analysts are trained in the use of the new software package and then instructed to use it to produce the information system. This study involves two populations: a population of systems analysts using the current technology and a population of systems analysts using the new software package. In terms of the time required to complete the information system design project, the population means are as follows. μ_1 is the mean project completion time for systems analysts using the current technology and μ_2 is the mean project completion time for systems analysts using the new software package The researcher in charge of the new software evaluation project hopes to show that the new software package will provide a shorter mean project completion time. Thus, the researcher is looking for evidence to conclude that μ_2 is less than μ_1; in this case, the difference between the two population means, $\mu_1 - \mu_2$, will be greater than zero. Suppose that the 24 analysts complete the study with the results shown in Table 10.1.

TABLE 10.1 Completion Time Data and Summary Statistics for the Software Testing Study

	Current Technology	New Software
	300	274
	280	220
	344	308
	385	336
	372	198
	360	300
	288	315
	321	258
	376	318
	290	310
	301	332
	283	263
Summary Statistics		
Sample size	$n_1 = 12$	$n_2 = 12$
Sample mean	$\bar{x}_1 = 325$ hours	$\bar{x}_2 = 286$ hours
Sample standard deviation	$s_1 = 40$	$s_2 = 44$

Let the level of significance be $\alpha = 0.05$. State the null and the alternative hypothesis, the test statistic, p-value, the rejection rule, make a decision and conclusion.

sol:

(Software Output)

TABLE 10.2 Output for the Hypothesis Test on the Difference Between the Current and New Software Technology

	Current	New
Mean	325	286
Variance	1600	1936
Observations	12	12
Hypothesized Mean Difference	0	
Degrees of Freedom	21	
Test Statistic	2.272	
One-Tail *p*-value	0.017	
One-Tail Critical Value	1.717	

Practical Advice

1. The interval estimation and hypothesis testing procedures presented in this section are _____ and can be used with _____ sample sizes.
2. In most applications, equal or nearly equal sample sizes such that the total sample size _____ can be expected to provide very good results even if the populations are not normal.
3. Larger sample sizes are recommended if the distributions of the populations are _____ or contain _____.
4. Smaller sample sizes should only be used if the analyst is satisfied that the distributions of the populations are at least _____.

Notes + Comments

1. How to make inferences about the difference between two population means when σ_1 and σ_2 are equal and unknown (_____)?
2. Based on above assumption, the two sample standard deviations are combined to provide the following pooled sample variance:

$$_____$$

3. The t test statistic becomes

   and has _____ degrees of freedom. At this point, the computation of the p-value and the interpretation of the sample results are identical to the procedures discussed earlier in this section.

4. A difficulty with this procedure is that the assumption that the two population standard deviations are equal is usually difficult to _____. Unequal population standard deviations are frequently encountered.

5. Using the pooled procedure may not provide satisfactory results, especially if the sample sizes n_1 and n_2 are _____.

6. The t procedure that we presented in this section does not require the assumption of equal population standard deviations and can be applied whether the population standard deviations are equal or not. It is a more general procedure and is recommended for most applications.

10.3 Inferences About The Difference Between Two Population Means: Matched Samples

1. Example **Matched**.

   (a) Suppose employees at a manufacturing company can use two different methods to perform a production task. To maximize production output, the company wants to identify the method with the smaller population mean completion time.

   (b) Let _____ denote the population mean completion time for production method 1 and _____ denote the population mean completion time for production method 2.

(c) With no preliminary indication of the preferred production method, we begin by tentatively assuming that the two production methods have the same population mean completion time. Thus, the null hypothesis is _____.

(d) If this hypothesis is rejected, we can conclude that the population mean completion times differ. In this case, the method providing the smaller mean completion time would be recommended.

(e) The null and alternative hypotheses are written as follows.

---

2. In choosing the sampling procedure that will be used to collect production time data and test the hypotheses, we consider two alternative designs. One is based on _____ and the other is based on _____.

   (a) *Independent sample design*: A simple random sample of workers is selected and each worker in the sample uses method 1. A second independent simple random sample of workers is selected and each worker in this sample uses method 2. The test of the difference between population means is based on the procedures in Section 10.2.

   (b) *Matched sample design*: One simple random sample of workers is selected. Each worker first uses one method and then uses the other method. The order of the two methods is assigned randomly to the workers, with some workers performing method 1 first and others performing method 2 first. Each worker provides _____, one value for method 1 and another value for method 2.

3. In the matched sample design the two production methods are tested under similar conditions (i.e., with the same workers); hence this design often leads to a _____ than the independent sample design. The primary reason is that in a matched sample design, _____ is eliminated because the same workers are used for both production methods.

4. Assuming the analysis of a matched sample design is the method used to test the difference between population means for the two production methods. The key

to the analysis of the matched sample design is to realize that we consider only _____.

5. Therefore, we have six data values (0.6, −0.2, 0.5, 0.3, 0.0, and 0.6) that will be used to analyze the difference between population means of the two production methods.

6. Let _____ is the mean of the difference in values for the population of workers. With this notation, the null and alternative hypotheses are rewritten as follows.

_____

7. Assume the population of _____ has a _____ distribution. This assumption is necessary so that we may use the _____ for hypothesis testing and interval estimation procedures. Based on this assumption, the following test statistic has a t distribution with _____ degrees of freedom.

8. **Test Statistic for Hypothesis Tests Involving Matched Samples**

_____

where

_____, and _____ (10.9)

✎ Question ..... (p498)

(Table 10.3) (Matched Example). A random sample of six workers is used. The data on completion times for the six workers are given in Table 10.3. Note that each worker provides a pair of data values, one for each production method. Also note that the last column contains the difference in completion times d_i for each worker in the sample. Assume that the population of differences has a normal distribution. Test the hypotheses $H_0 : \mu_d = 0$ and $H_a : \mu_d \neq 0$, using $\alpha = 0.05$. Compute the test statistic, the p-value and draw a conclusion. Compute the 95% confidence interval for the difference between the population means of the two production methods. If H_0 is rejected, we can conclude that the population mean completion times differ.

TABLE 10.3 Task Completion Times for a Matched Sample Design

Worker	Completion Time for Method 1 (minutes)	Completion Time for Method 2 (minutes)	Difference in Completion Times (d_i)
1	6.0	5.4	.6
2	5.0	5.2	−.2
3	7.0	6.5	.5
4	6.2	5.9	.3
5	6.0	6.0	.0
6	6.4	5.8	.6

sol:

Area in Upper Tail	0.20	0.10	0.05	0.025	0.01	0.005
t-Value (5 df)	0.920	1.476	2.015	2.571	3.365	4.032

10.4 Inferences About The Difference Between Two Population Proportions

1. Letting _____ denote the proportion for population 1 and _____ denote the proportion for population 2.

2. Consider inferences about the difference between the two population proportions: _____.

3. To make an inference about this difference, we will select two independent random samples consisting of n_1 units from population 1 and n_2 units from population 2.

Interval Estimation of $p_1 - p_2$

1. Example **Tax Preparation Firm**
   A tax preparation firm is interested in comparing the quality of work at two of its regional offices. By randomly selecting samples of tax returns prepared at each office and verifying the sample returns' accuracy, the firm will be able to estimate the proportion of erroneous returns prepared at each office. Of particular interest is the difference between these proportions.

   (a) p_1: proportion of erroneous returns for population 1 (office 1)

   (b) p_2: proportion of erroneous returns for population 2 (office 2)

   (c) _____: sample proportion for a simple random sample from population 1

   (d) _____: sample proportion for a simple random sample from population 2

2. **Point Estimator of the Difference Between Two Population Proportions**

   _____ (10.10)

3. Thus, the point estimator of the difference between two _____ proportions is the difference between the _____ proportions of two independent simple random samples.

4. As with other point estimators, the point estimator $\bar{p}_1 - \bar{p}_2$ has a sampling distribution that reflects the possible values of $\bar{p}_1 - \bar{p}_2$ if we repeatedly took two independent

random samples. The mean of this sampling distribution is _____ and the standard error of $\bar{p}_1 - \bar{p}_2$ is:

Standard Error of $\bar{p}_1 - \bar{p}_2$

(10.11)

_____

5. If the sample sizes are large enough that _____, _____, _____, and _____ are all greater than or equal to _____, the sampling distribution of $\bar{p}_1 - \bar{p}_2$ can be approximated by a _____ distribution.

6. With the sampling distribution of $\bar{p}_1 - \bar{p}_2$ approximated by a normal distribution, we would like to use _____ as the margin of error.

7. However, $\sigma_{\bar{p}_1 - \bar{p}_2}$ given by equation (10.11) cannot be used directly because the two population proportions, p_1 and p_2, are unknown. Using the sample proportion $\bar{p}_1$ to estimate $\bar{p}_1$ and the sample proportion $\bar{p}_2$ to estimate p_2, the margin of error is:

Margin of error = (10.12)

_____

8. **Interval Estimate of the Difference Between Two Population Proportions**

(10.13)

_____

where $1-\alpha$ is the confidence coefficient.

Question (p504)

(**Tax Preparation Example**) We find that independent simple random samples from the two offices provide the following information.

Office	1	2
n_i	250	300
Number of returns with errors	35	27

Find a margin of error and interval estimate of the difference between the two population proportions. and 90% confidence interval.

sol:

Hypothesis Tests About $p_1 - p_2$

1. Let us now consider hypothesis tests about no difference between the proportions of two populations. In this case, the three forms for a hypothesis test are as follows:

$$H_0 : p_1 - p_2 \geq 0, \qquad H_0 : p_1 - p_2 \leq 0, \qquad H_0 : p_1 - p_2 = 0$$
$$H_a : p_1 - p_2 < 0 \qquad H_a : p_1 - p_2 > 0 \qquad H_a : p_1 - p_2 \neq 0$$

2. When we assume _____, we have $p_1 - p_2 = 0$, which is the same as saying that the population proportions are equal, $p_1 = p_2$.

3. Under the assumption H_0 is true as an equality, the population proportions are equal and _____. In this case, $\sigma_{\bar{p}_1 - \bar{p}_2}$ becomes **Standard Error of** $\bar{p}_1 - \bar{p}_2$ **when** $p_1 = p_2 = p$

$$\sigma_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} = \underline{\qquad\qquad} \tag{10.14}$$

4. With p unknown, we pool, or combine, the point estimators from the two samples ($\bar{p}_1$ and $\bar{p}_2$) to obtain a single point estimator of p as follows:
Pooled Estimator of p **When** $p_1 = p_2 = p$

$$\underline{\qquad\qquad} \tag{10.15}$$

This pooled estimator of p is a weighted average of $\bar{p}_1$ and $\bar{p}_2$.

5. Substituting $\bar{p}$ for p in equation (10.14), we obtain an estimate of the standard error of $\bar{p}_1 - \bar{p}_2$. This estimate of the standard error is used in the test statistic.

6. The general form of the test statistic for hypothesis tests about the difference between two population proportions is the point estimator divided by the estimate of $\sigma_{\bar{p}_1 - \bar{p}_2}$.

7. **Test Statistic for Hypothesis Tests About $p_1 - p_2$**

(10.16)

This test statistic applies to large sample situations where $n_1 p_1$, $n_1(1-p_1)$, $n_2 p_2$, and $n_2(1-p_2)$ are all greater than or equal to 5.

✎ Question ..... (p506)

(**Tax Preparation Firm Example**) Assume that the firm wants to use a hypothesis test to determine whether the error proportions differ between the two offices. A two-tailed test is required. Use $\alpha = 0.10$ as the level of significance. State the null and alternative hypotheses. Compute the test statistic, and the p-value for this two-tailed test. State the decison rule and draw a conclusion.

sol:

😍 EXERCISES

10.1 : 1, 2, 4, 6

10.2 : 9, 10, 13, 14, 15

10.3 : 19, 23, 24

10.4 : 28, 29, 31, 34

SUP : 38, 39, 44

名言佳句

“成功人士與其他人的區別，不是在於他們的力量或知識，而是在於他們有多堅持”
“The difference between a successful person and others is not a lack of strength, not a lack of knowledge, but rather a lack of will.”

— *Vince Lombardi (June 11, 1913 – September 3, 1970)*

統計小故事：隨機對照試驗與雙組比較的力量

在醫學史上，二十世紀被視為一個重大轉折點。過去的醫療實踐，往往依靠醫師的經驗與直覺，或根據少數病人的個案報告來判斷藥物是否有效。然而，這樣的方式存在巨大的風險。藥效可能只是偶然，或是來自心理作用。如何用科學方法來比較新藥與舊藥、治療組與對照組的差異，成為醫學進步的核心問題。

這場革命的推手之一，是英國醫師與統計學家奧斯汀．布拉德福德．希爾 (Austin Bradford Hill)。他在二十世紀五　年代推廣並系統化了隨機對照試驗 (Randomized Controlled Trial, RCT) 的方法，這種設計依賴於嚴格的統計推論，尤其是「兩個母體之間的平均數與比例比較」。

在傳統的臨床觀察中，醫師可能會將某一批病人使用新藥，然後觀察康復率。但這種方法容易受到偏差影響。例如，若新藥組剛好包含較年輕或體質較好的病人，那麼康復率自然會較高，但這並不一定代表新藥有效。希爾意識到，必須用「隨機分組」來確保兩組病人具有可比性。

在 RCT 中，病人會被隨機分配到「實驗組」與「對照組」。實驗組接受新藥，對照組則接受安慰劑或舊藥。這樣一來，兩組的差異可以視為隨機的，消除了潛在偏差的影響。接著，研究者就能比較兩組的平均康復率或比例，並用統計方法檢驗其差異是否顯著。

例如，假設有 200 名病人被隨機分成兩組，每組 100 人。新藥組有 70 人康復，安慰劑組有 55 人康復。這樣的差異 (70% 對 55%) 是否足以證明新藥有效？這就需要用到統計學中「兩母體平均數與比例的比較」的方法。透過檢定與信賴區間，我們可以判斷差異是否顯著，並推斷這個差異是否可能存在於整個病人群體。

希爾的工作帶來了醫學研究的一場革命。此後，所有新藥上市前，必須經過多階段的臨床試驗，其中雙組比較是最核心的一環。這不僅保障了病人的安全，也確保了醫療決策的科學性。換言之，現代人能夠安心服用藥物，背後正是統計學在發揮作用。

雙組比較的力量並不限於醫學。在社會科學研究中，學者也會透過「實驗組」與「對照組」來檢驗政策的效果。例如，一個城市若要評估新的教育方案是否能提升學生表現，可以隨機挑選部分學校實施新方案，其他學校則維持原狀。最後比較兩組的平均成績，就能推論新政策是否有效。

同樣地，在商業領域，A/B 測試也是雙組比較的一種應用。科技公司會隨機將用戶分為兩組，一組看到新版本介面，另一組維持舊版本，然後比較兩組的點擊率或轉換率差異。這種方法已成為數位產品開發的標準流程。

然而，雙組比較也有需要謹慎之處。樣本數過小可能導致檢定力不足，難以發現真實差異；樣本分配若不隨機，可能導致結果偏誤。因此，設計良好的實驗與合理的統計分析，缺一不可。

在教材呼應上，本章的核心是「兩個母體的平均數與比例比較」。學生需要掌握如何建立假設、計算檢定統計量，並利用 p 值或信賴區間來做出推論。這不僅是一個數學技巧，更是一種科學推理框架。

因此，當學生在課堂上計算「兩家公司的顧客滿意度是否有顯著差異」時，他們其實正延續著希爾與無數科學家的努力。這些數學方法不僅改變了研究方式，更改變了我們的生活。從藥品到政策，從教育到商業，雙組比較是我們理解世界、驗證假設的重要工具。

總結來說，比較兩個群體的差異，不僅是數字運算，而是科學決策的核心。希爾的隨機對照試驗讓醫學更可靠，而今天的我們，則透過同樣的方法，在各領域持續追求更準確的答案。

國立政治大學 113 學年度第 1 學期 Quiz (1) 考試命題紙

Subject： 統計學 (一) 開課班別： 統計學整合開課 Teacher: Han-Ming Wu

Date：24 Oct. (Thur) 15:00-16:00	※Allowed:「O」· Prohibited:「×」				1. 需加發計算紙或答案紙請在試題內封袋備註。 2. 為環保節能減碳，試題一律採雙面印刷，如有特殊印製需求，請註記：
Pages: 3 · Copies: 50	Calculator	Textbook	Class notes	3C product	
Scope: §ch1-3	O	×	×	×	

Notes:

(1) Fill in the student ID number and name on the answer sheet。

(2) Answer all questions in English (ignore the grammar and spelling)。

(3) Answer each question in the order it appears。The total score is 120.

(4) It is recommended to use a dark ballpoint pen。(pencil is allowed)

(5) The calculation process (for parts **IV** and **V**) is required (calculate to 4 decimal places)。

(6) Return both the answer sheet and the question sheet.

(-) Declaration (0%): Please transcribe the following oath onto the first page of the answer sheet in either Chinese or English. (複寫下列宣誓詞至答案卷的第一頁上)。(10 points will be deducted if not written.)(不寫扣 10 分)

0. " 本人姓名 恪遵各項考試規則，若如違反，願受校方最嚴厲處罰，謹誓。"

"I (your name here) will strictly adhere to all examination rules. If I break this oath, I am willing to accept the most severe punishment imposed by the school. Solemnly sworn."

(I) Multiple choice (20%, 5% each); select one correct answer.

1. A sample of 15 children shows their favorite kind of pet:
`Dog, Gerbil, Cat, Fish, Dog, Dog, Gerbil, Cat, Dog, Cat, Lizard, Fish, Dog, Fish, Dog.`
Which of the following displays is most appropriate for this data? (A). Side-by-side bar chart (B). Histogram (C). Stacked bar chart (D). Pie chart

2. Histograms based on data on housing prices and salaries typically are (A). skewed to the left. (B). skewed to the right. (C). stacked. (D). symmetric.

3. The measure of location which is the most likely to be influenced by extreme values in the data set is the (A). range. (B). median. (C). mode. (D). mean.

4. Statements about the proportion of data values that must be within a specified number of standard deviations of the mean can be made using (A). Chebyshev's theorem. (B). The empirical rule. (C). A five-number summary. (D). Percentiles.

國立政治大學 113 學年度第 1 學期 Quiz (1) 考試命題紙

考試科目： 統計學 (一)　　開課班別： 統計學整合開課　　命題教授: Han-Ming Wu

考試日期：24 Oct. (Thur) 15:00-16:00	※准帶項目打「O」，否則打「×」					1. 需加發計算紙或答案紙請在試題內封袋備註。 2. 為環保節能減碳，試題一律採雙面印刷，如有特殊印製需求，請註記：
本試題共3頁，印刷份數: 50 份	計算機	課本	筆記	字典	手機平板筆電	
備註：注意事項要看!! (Scope: §ch1-3)	O	×	×	×	×	

(II) Fill-in-the-blank (Correct spelling should be used if possible.) (20%, 10% each)

5. The growth factors in the population of Dalton for the past three years are shown below.

   Year 1: 1.0313; Year 2: 1.0429; Year 3: 1.0343.

   What has been the percentage growth in the population of Dalton? _____.

6. Let the ith data value of the variable x denotes by x_i. The z-score is often called the _____ value of x_i. The formula is: _____.

(III) Short answer (20%, 10% each)(write down the statement (or definition), formula if any, interpretation)

7. What is the Chebyshev's theorem?

8. What is the Covariance?

(IV) Calculation (40%, 20% each)

9. **Apple iPads in Schools.** The New York Times reported that Apple has unveiled a new iPad marketed specifically to school districts for use by students (The New York Times website). The 9.7-inch iPads will have faster processors and a cheaper price point in an effort to take market share away from Google Chromebooks in public school districts. Suppose that the following data represent the percentages of students currently using Apple iPads for a sample of 18 U.S. public school districts. ($\bar{x} = 24.67, s = 11.87$)

   `15, 22, 12, 21, 26, 18, 42, 29, 64, 20, 15, 22, 18, 24, 27, 24, 26, 19`

   (a) Compute the first and third quartiles for these data.

   (b) Identify the outliers in these data if any. (The calculation process is required)

國立政治大學 113 學年度第 1 學期 Quiz (1) 考試命題紙

考試科目： 統計學 (一) 開課班別： 統計學整合開課 命題教授: Han-Ming Wu

考試日期：24 Oct. (Thur) 15:00-16:00	※准帶項目打「O」，否則打「×」					1. 需加發計算紙或答案紙請在試題內封袋備註。 2. 為環保節能減碳，試題一律採雙面印刷，如有特殊印製需求，請註記：
本試題共3頁，印刷份數: 50 份	計算機	課本	筆記	字典	手機平板筆電	
備註：注意事項要看!! (Scope: §ch1-3)	O	×	×	×	×	

(V) Bonus (20%)

10. Jason just completed 15 credit hours at a local college. His grade report is presented below.

Course:	Calculus	Biology	English	Music	P.E.
Credit Hours:	5	4	3	2	1
Grades:	C	A	D	B	A

The local university uses a 4 point grading system, i.e., $A = 4, B = 3, C = 2, D = 1, F = 0$. Compute Jason's semester grade point average.

<The blank pages at the back can be used as scratch paper. (後面空白頁可當計算紙)>

注意： 1、考試求公平及公正，請同學務必自律，維護學校與學生之榮譽。

2、考試時不得有交談、窺視、夾帶、抄襲、傳遞、代考或其它作弊等舞弊行為，考畢務必交卷，不得攜卷出場，違者依考場規則議處。

國立政治大學 113 學年度第 1 學期 Quiz (2) 考試命題紙

Subject： 統計學 (一) 開課班別： 統計學整合開課 Teacher: Han-Ming Wu

Date：26 Dec. (Thur) 15:00-16:00	※Allowed:「O」· Prohibited:「×」				1. 需加發計算紙或答案紙請在試題內封袋備註。 2. 為環保節能減碳，試題一律採雙面印刷，如有特殊印製需求，請註記：
Pages: 3 · Copies: 50	Calculator	Textbook	Class notes	3C product	
Scope: §ch5-6	O	×	×	×	

Notes:

(1) Fill in the student ID number and name on the answer sheet。

(2) Answer all questions in English (ignore the grammar and spelling)。

(3) Answer each question in the order it appears。The total score is 120.

(4) It is recommended to use a dark ballpoint pen。(pencil is allowed)

(5) The calculation process (for parts **IV** and **V**) is required (calculate to 4 decimal places)。

(6) Return both the answer sheet and the question sheet.

(-) Declaration (0%): Please transcribe the following oath onto the first page of the answer sheet in either Chinese or English. (複寫下列宣誓詞至答案卷的第一頁上)。(10 points will be deducted if not written.)(不寫扣 10 分)

0. ”本人姓名 恪遵各項考試規則，若如違反，願受校方最嚴厲處罰，謹誓。”

”I (your name here) will strictly adhere to all examination rules. If I break this oath, I am willing to accept the most severe punishment imposed by the school. Solemnly sworn.”

(I) Multiple choice (20%, 5% each); select one correct answer.

1. For a continuous random variable X, the height of the function at $X = x$ is (A). the probability at a given value of x. (B). named the probability density function $f(x)$. (C). 0.50, since it is the middle value. (D). a value less than zero.

2. Random variable X has the probability function $f(x) = x/6$, for $x = 1, 2$ or 3. The expected value of X is (A). 0.500. (B). 2.000. (C). 2.333. (D). 2.5

3. 20. For any continuous random variable, the probability that the random variable takes on exactly a specific value is (A). 1.00. (B). 0.50. (C). any value between 0 to 1. (D). 0.

4. Which of the following is not a characteristic of an experiment where the binomial probability distribution is applicable? (A). The experiment has a sequence of n identical trials. (B). Exactly two outcomes are possible on each trial. (C). The experiment is independent. (D). The probabilities of the outcomes do not change from one trial to another.

國立政治大學 113 學年度第 1 學期 Quiz (2) 考試命題紙

考試科目： 統計學 (一) 開課班別： 統計學整合開課 命題教授: Han-Ming Wu

考試日期：26 Dec. (Thur) 15:00-16:00	※准帶項目打「O」，否則打「×」					1. 需加發計算紙或答案紙請在試題內封袋備註。 2. 為環保節能減碳，試題一律採雙面印刷，如有特殊印製需求，請註記：
本試題共3頁，印刷份數: 50 份	計算機	課本	筆記	字典	手機平板筆電	
備註：注意事項要看!! (Scope: §ch5-6)	O	×	×	×	×	

(II) Fill-in-the-blank (Correct spelling should be used if possible.) (20%, 10% each)

5. The expected value is a _____ of the values of the discrete random variable where the weights are the probabilities.

6. In "Normal Approximation of Binomial Probabilities", the 0.5 that we add and subtract from x is called a _____.

(III) Short answer (20%, 10% each)(write down the statement (or definition), formula if any, interpretation)

7. What is the "probability distribution" for a random variable? (in discrete case)

8. What is the "Normal Approximation of Binomial Probabilities?" (You SHOULD NOT just answer: " use normal to approximate binomial." nor "use normal probabilities to approximate binomial probabilities.")

(IV) Calculation (40%)

9. **Visitors to Rocky Mountain National Park.** Rocky Mountain National Park is a popular park for outdoor recreation activities in Colorado. According to U.S. National Park Service statistics, 46.7% of visitors to Rocky Mountain National Park in 2018 entered through the Beaver Meadows park entrance, 24.3% of visitors entered through the Fall River park entrance, 6.3% of visitors entered through the Grand Lake park entrance, and 22.7% of visitors had no recorded point of entry to the park. Consider a random sample of 175 Rocky Mountain National Park visitors.

   (a) What is the probability that at least 70 but less than 80 visitors had a recorded entry through the Beaver Meadows park entrance?

   (b) What is the probability that more than 45 visitors have no recorded point of entry?

考試日期：26 Dec. (Thur) 15:00-16:00	※准帶項目打「O」，否則打「×」					1. 需加發計算紙或答案紙請在試題內封袋備註。 2. 為環保節能減碳，試題一律採雙面印刷，如有特殊印製需求，請註記：
本試題共3頁，印刷份數: 50 份	計算機	課本	筆記	字典	手機平板筆電	
備註：注意事項要看!! (Scope: §ch5-6)	O	×	×	×	×	

(V) Bonus (20%)

10. (20%) **Canadian Restaurant Ratings.** The Chamber of Commerce in a Canadian city has conducted an evaluation of 300 restaurants in its metropolitan area. Each restaurant received a rating on a 3point scale on typical meal price (1 least expensive to 3 most expensive) and quality (1 lowest quality to 3 greatest quality). A crosstabulation of the rating data is shown below. Fortytwo of the restaurants received a rating of 1 on quality and 1 on meal price, 39 of the restaurants received a rating of 1 on quality and 2 on meal price, and so on. Fortyeight of the restaurants received the highest rating of 3 on both quality and meal price.

	Meal Price (*y*)			
Quality (*x*)	1	2	3	Total
1	42	39	3	84
2	33	63	54	150
3	3	15	48	66
Total	78	117	105	300

Compute the correlation coefficient between quality and meal price.

機率表

Lower tail probability: $Z \sim N(0,1), P(Z \le z) = p.$

z	-1.85	-1.78	-1.7	-0.34	-0.26	-0.19	0.86	0.95	1.04
p	0.0322	0.0375	0.0446	0.3669	0.3974	0.4247	0.8051	0.8289	0.8508

<The blank pages at the back can be used as scratch paper. (後面空白頁可當計算紙)>

注意： 1、考試求公平及公正，請同學務必自律，維護學校與學生之榮譽。
2、考試時不得有交談、窺視、夾帶、抄襲、傳遞、代考或其它作弊等舞弊行為，考畢務必交卷，不得攜卷出場，違者依考場規則議處。

國立政治大學 113 學年度第 1 學期 Midterm Exam 考試命題紙

Subject： 統計學 (一) 開課班別： 統計學整合開課 Teacher: Han-Ming Wu

Date：14 Nov. (Thur) 13:10-14:50	※Allowed:「O」, Prohibited:「×」				1. 需加發計算紙或答案紙請在試題內封袋備註。 2. 為環保節能減碳，試題一律採雙面印刷，如有特殊印製需求，請註記：
Pages: 3，Copies: 50	Calculator	Textbook	Class notes	3C product	
Scope: §ch1-4	O	×	×	×	

Notes:

(1) Fill in the student ID number and name on the answer sheet。

(2) Answer all questions in English (ignore the grammar and spelling)。

(3) Answer each question in the order it appears。The total score is 120.

(4) It is recommended to use a dark ballpoint pen。(pencil is allowed)

(5) The calculation process (for parts **IV** and **V**) is required (calculate to 4 decimal places)。

(6) Return both the answer sheet and the question sheet.

(-) Declaration (0%): Please transcribe the following oath onto the first page of the answer sheet in either Chinese or English. (複寫下列宣誓詞至答案卷的第一頁上)。(10 points will be deducted if not written.)(不寫扣 10 分)

0. " 本人姓名 恪遵各項考試規則，若如違反，願受校方最嚴厲處罰，謹誓。"

"I (your name here) will strictly adhere to all examination rules. If I break this oath, I am willing to accept the most severe punishment imposed by the school. Solemnly sworn."

(I) Multiple choice (20%, 5% each); select one correct answer.

1. The flashlight batteries produced by one of the northern manufacturers are known to have an average life of 60 hours with a standard deviation of 4 hours. At least what percentage of flashlights will have a life of 54 to 66 hours? (A). 56%. (B). 60%. (C). 64%. (D). 68%.

2. Consider a sample with the following data values: 62, 90, 50, 94, 74. Compute the Z scores for the above five observations. What is the mean of these Z scores? (A). -1. (B). 0. (C). 0.5. (D). 1 .

3. If X and Y are mutually exclusive events with $P(A) = 0.295$, $P(B) = 0.32$, then $P(A|B) =$
(A). 0.0944. (B). 0.6150. (C). 1.0000. (D). 0.0000.

4. If $P(A \cap B) = 0$, (A). $P(A) + P(B) = 1$. (B). either $P(A) = 0$ or $P(B) = 0$. (C). A and B are mutually exclusive events. (D). A and B are independent events.

考試日期：14 Nov. (Thur) 13:10-14:50	※准帶項目打「O」, 否則打「×」					1. 需加發計算紙或答案紙請在試題內封袋備註。 2. 為環保節能減碳，試題一律採雙面印刷，如有特殊印製需求，請註記：
本試題共3頁，印刷份數: 50 份	計算機	課本	筆記	字典	手機平板筆電	
備註：注意事項要看!! (Scope: §ch1-4)	O	×	×	×	×	

(II) Fill-in-the-blank (Correct spelling should be used if possible.) (20%, 10% each)

5. The following frequency distribution shows the time (in minutes) that a sample of students uses the computer terminals per day. Compute the mean. _____.

Time:	20 – 39	40 – 59	60 – 79	80 – 99	100 – 119
Frequency:	2	4	6	4	2

6. The results of a survey of 800 married couples and the number of children they had is shown below.

Number of Children:	0	1	2	3	4	5
Probability:	0.050	0.125	0.600	0.150	0.050	0.025

If a couple is selected at random, what is the probability that the couple will have Either 2 or 3 children? _____.

(III) Short answer (20%, 10% each)(write down the statement (or definition), formula if any, interpretation)

7. What is the so-called "joint probability"?

8. What is the "Bayes' theorem"? (formula is required.)

(IV) Calculation (40%, 20% each)

9. **Intent to Pursue MBA.** Students taking the Graduate Management Admissions Test (GMAT) were asked about their undergraduate major and intent to pursue their MBA as a full-time or part-time student. A summary of their responses follows.

		Undergraduate Major			
		Business	Engineering	Other	Totals
Intended Enrollment Status	Full-Time	352	197	251	800
	Part-Time	150	161	194	505
	Totals	502	358	445	1305

Are the event that the student intends to attend classes full-time in pursuit of an MBA degree, and the event that the student was an undergraduate business major independent? Justify your answer.

考試日期：14 Nov. (Thur) 13:10-14:50	※准帶項目打「O」，否則打「×」					1. 需加發計算紙或答案紙請在試題內封袋備註。 2. 為環保節能減碳，試題一律採雙面印刷，如有特殊印製需求，請註記：
本試題共3頁，印刷份數: 50 份	計算機	課本	筆記	字典	手機平板筆電	
備註：注意事項要看!! (Scope: §ch1-4)	O	×	×	×	×	

10. **Consulting Firm Bids.** A consulting firm submitted a bid for a large research project. The firm's management initially felt they had a 50–50 chance of getting the project. However, the agency to which the bid was submitted subsequently requested additional information on the bid. Past experience indicates that for 75% of the successful bids and 40% of the unsuccessful bids the agency requested additional information. Compute the posterior probability that the bid will be successful given a request for additional information.

(V) Bonus (20%, 5%each)

11. (a) If $P(A) = 0.4$, $P(B|A) = 0.35$, $P(A \cup B) = 0.69$, then $P(B) =$ _____.
    (b) If $P(A) = 0.50$, $P(B) = 0.40$ and $P(A \cup B) = 0.88$, then $P(B|A) =$ _____.
    (c) If $P(A_1) = 0.40$, $P(A_2) = 0.60$ and $P(A_1 \cap A_2) = 0$. Suppose $P(B|A_1) = 0.20$ and $P(B|A_2) = 0.05$, Then $P(A_1 \cap B) =$ _____.
    (d) (same as above) $P(A_2|B) =$ _____.

<The blank pages at the back can be used as scratch paper. (後面空白頁可當計算紙)>

注意： 1、考試求公平及公正，請同學務必自律，維護學校與學生之榮譽。
2、考試時不得有交談、窺視、夾帶、抄襲、傳遞、代考或其它作弊等舞弊行為，考畢務必交卷，不得攜卷出場，違者依考場規則議處。

國立政治大學 113 學年度第 1 學期 Final Exam 考試命題紙

Subject： 統計學 (一) 開課班別： 統計學整合開課 Teacher: Han-Ming Wu

Date：09 Jan. (Thur) 13:10-14:50	※Allowed:「O」· Prohibited:「×」				1. 需加發計算紙或答案紙請在試題內封袋備註。 2. 為環保節能減碳，試題一律採雙面印刷，如有特殊印製需求，請註記：
Pages: 3 · Copies: 50	Calculator	Textbook	Class notes	3C product	
Scope: §ch5-8	O	×	×	×	

Notes:

(1) Fill in the student ID number and name on the answer sheet。

(2) Answer all questions in English (ignore the grammar and spelling)。

(3) Answer each question in the order it appears。The total score is 120.

(4) It is recommended to use a dark ballpoint pen。(pencil is allowed)

(5) The calculation process (for parts **IV** and **V**) is required (calculate to 4 decimal places)。

(6) Return both the answer sheet and the question sheet.

(-) Declaration (0%): Please transcribe the following oath onto the first page of the answer sheet in either Chinese or English. (複寫下列宣誓詞至答案卷的第一頁上)。(10 points will be deducted if not written.)(不寫扣 10 分)

0. " 本人姓名 恪遵各項考試規則，若如違反，願受校方最嚴厲處罰，謹誓。"

"I (your name here) will strictly adhere to all examination rules. If I break this oath, I am willing to accept the most severe punishment imposed by the school. Solemnly sworn."

(I) Multiple choice (20%, 5% each); select one correct answer.

1. The closer the sample mean is to the population mean, (A). the larger the sampling error. (B). the smaller the sampling error. (C). the sampling error equals 1. (D). none of these alternatives is correct.

2. The purpose of statistical inference is to provide information about the (A). population based upon information contained in the sample. (B). sample based upon information contained in the population. (C). population based upon information contained in the population. (D). mean of the sample based upon the mean of the population.

3. If we change a 95% confidence interval estimate to a 99% confidence interval estimate, we can expect the (A). width of the confidence interval to increase. (B). width of the confidence interval to decrease. (C). width of the confidence interval to remain the same. (D). sample size to increase.

4. For a given confidence level and when is known, the margin of error in a confidence interval estimate (A). is independent of sample size. (B). varies from sample to sample of the same size. (C). increases as the sample size increases. (D). is the same for all samples of the same size.

國立政治大學 113 學年度第 1 學期 Final Exam 考試命題紙

考試科目： 統計學 (一) 開課班別： 統計學整合開課 命題教授: Han-Ming Wu

考試日期：09 Jan. (Thur) 13:10-14:50	※准帶項目打「O」，否則打「×」					1. 需加發計算紙或答案紙請在試題內封袋備註。 2. 為環保節能減碳，試題一律採雙面印刷，如有特殊印製需求，請註記：
本試題共3頁，印刷份數: 50 份	計算機	課本	筆記	字典	手機平板筆電	
備註：注意事項要看!! (Scope: §ch5-8)	O	×	×	×	×	

(II) Fill-in-the-blank (Correct spelling should be used if possible.) (20%, 5% each)

5. A random variable, often denoted by X (or some other capital letter), is a _____ that maps outcomes to numbers on the real line.

6. What is the normal probability density function (let μ denotes the population mean, and σ the population standard deviation): _____.

7. A sample characteristic, such as a sample mean $\bar{X}$, a sample standard deviation s, a sample proportion $\bar{p}$, and so on, is called the _____.

8. The t distribution is a family of similar probability distributions, with a specific t distribution depending on a parameter known as the _____.

(III) Short answer (20%, 10% each)(write down the statement (or definition), formula if any, interpretation)

9. What is the so-called "sampling distribution"?

10. How to interpret a 95% confidence interval for a population parameter θ?

(IV) Calculation (40%, 20% each)

11. **Employee Ages**. To estimate the mean age for a population of 4000 employees, a simple random sample of 40 employees is selected.

    (a) If the population standard deviation is $\sigma = 8.2$ years, compute the standard error both with and without the finite population correction factor. What is the rationale for ignoring the finite population correction factor whenever $n/N \leq 0.05$?

    (b) What is the probability that the sample mean age of the employees will be within ± 2 years of the population mean age?

考試日期：09 Jan. (Thur) 13:10-14:50	※准帶項目打「O」，否則打「×」					1. 需加發計算紙或答案紙請在試題內封袋備註。 2. 為環保節能減碳，試題一律採雙面印刷，如有特殊印製需求，請註記：
本試題共3頁，印刷份數: 50 份	計算機	課本	筆記	字典	手機平板筆電	
備註：注意事項要看!! (Scope: §ch5-8)	O	×	×	×	×	

12. **Meal Cost in Hong Kong**. The mean cost of a meal for two in a midrange restaurant in Tokyo is \$40 (Numbeo.com website). How do prices for comparable meals in Hong Kong compare? (The data file HongKongMeals contains the costs for a sample of 42 recent meals for two in Hong Kong midrange restaurants (the sample mean is \$32.66, and the sample variance is 46.60793)).

    (a) What is the 95% confidence interval estimate of the population mean in Hong Kong?

    (b) How do prices for meals for two in midrange restaurants in Hong Kong compare to prices for comparable meals in Tokyo restaurants? (Hint: Whether the mean cost for a mid-range meal for two in Tokyo is in or not in the 95% confidence interval for comparable meals in Hong Kong? if yes, what does that mean? Or if not, what does that mean?)

(V) Bonus (20%)

13. Let $X_1, X_2, \cdots, X_n$ be i.i.d. random sample from a Normal population with a unknown mean μ and a known standard deviation σ. Derive the formula of the 95% confidence interval for μ.

機率表

Lower tail probability of a standard normal distribution: $Z \sim N(0,1), P(Z \le z) = p$.

z	-1.29	-0.26	-0.19	0.86	0.95	1.54	1.78
p	0.0985	0.3974	0.4247	0.8051	0.8289	0.9382	0.9625

Lower tail probability of a t distribution: $T \sim t_{(df)}, P(T \le t) = p$.

df	41	41	41	42	42	42	43	43	43
t	-2.4208	-2.0195	-1.6829	-2.4185	-2.0181	-1.6819	-2.4163	-2.0166	-1.6811
p	0.01	0.025	0.05	0.01	0.025	0.05	0.01	0.025	0.05

<The blank pages at the back can be used as scratch paper. (後面空白頁可當計算紙)>

注意： 1、考試求公平及公正，請同學務必自律，維護學校與學生之榮譽。
2、考試時不得有交談、窺視、夾帶、抄襲、傳遞、代考或其它作弊等舞弊行為，考畢務必交卷，不得攜卷出場，違者依考場規則議處。

“人生路上遇到的挫折只是為了要塑造你、讓你達成目標”
“The struggles along the way are only meant to shape you for your purpose.”
— *Chadwick Boseman (November 29, 1976 – August 28, 2020)*