Subject Syllabus

CC3101 Business Statistics

<table>
<thead>
<tr>
<th>Level</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Credits</td>
<td>3</td>
</tr>
<tr>
<td>Nature</td>
<td>Science</td>
</tr>
<tr>
<td>Mode of Study</td>
<td>28 hours of Lecture</td>
</tr>
<tr>
<td></td>
<td>14 hours of Tutorial</td>
</tr>
<tr>
<td>Prerequisites</td>
<td>Nil</td>
</tr>
<tr>
<td>Assessment</td>
<td>40% Coursework</td>
</tr>
<tr>
<td></td>
<td>60% Examination</td>
</tr>
</tbody>
</table>

Aims

This subject equips students with various statistical skills that are necessary for solving business decision problem under uncertainty. It provides students with an understanding on how business data are collected, summarised, presented, analysed, and interpreted. Studying the subject will also help develop students’ critical thinking and analytic skills for their life-long learning.

Learning Outcomes

On successfully completing this subject, students will be able to:

- Apply statistical techniques and make rigorous statistical analysis in decision making problem.
- Decide which methods can be used to collect, describe and present business data.
- Analyse business data and interpret the results for making recommendations.
- Relate probability theories to solve problems involving uncertainty.
- Use some common statistical packages such as Excel or SPSS.

Indicative Contents

- **Descriptive Statistics**
  Basic statistical terms; Sampling methods; Sampling and non-sampling errors; Types of data; Use of tables and charts to summarise categorical and numerical data; Measures of central tendency; Measures of variation; Interpretation of the numerical descriptive measures.

- **Basic Probability Law**
  Permutations and combinations; Probability Laws: addition law, multiplicative law and complement law; Conditional probability; Statistical independence; Bayes Theorem.

- **Probability Distribution**
  Discrete and continuous random variables; Expected value and standard deviation of random variables; Covariance analysis; Binomial and Poisson probability distribution; Normal, exponential and uniform distributions; Test for normality.

- **Sampling Distribution**
  Sampling distribution of mean and proportion; Central limit theorem; Probability problems involving sampling distribution.

- **Statistical Inference**
Estimator; Confidence interval; Sample size determination; Hypothesis testing; Normal test; t test; Inference on proportions.

- **Simple Linear Regression**
  Causal relationship in business data; Scatter diagram; Least squares method; Assumptions of the regression analysis; Coefficient of determination; Coefficient of correlation; Standard error of estimate; Prediction.

- **Time Series Analysis**
  Time series model; Time series trend; Seasonal variation and forecasting; Index relatives; Composite index numbers; Other business indices from publications.

**Teaching/Learning Approach**

Lectures focus on the introduction and explanation of key statistical concepts, with specific reference to current business decision problems. Occasional group discussions will be conducted.

Tutorials provide students with the opportunity to deepen their understanding of the concepts taught in lectures and to apply the theories to the analysis of real-life business decision problems. The activities in tutorials normally include discussions of problems sets.

**Assessment Approach**

A variety of assessment tools will be used, including interactions between teacher and students, group discussions, assignments, tests and examination designed to develop and assess critical thinking as well as analytical and interpersonal skills.

**Indicative Readings**

Recommended Textbook

References
Dillman, Don A., Salant, Priscilla, *How to Conduct Your Own Survey*, Wiley (latest ed.).


Hong Kong Community College  
CC3101 Business Statistics  
Tentative Teaching Plan  
Semester Two 2009/2010

Subject Leader

Chan Chun Man (Office: WK S1332, Tel: 3746-0120, email: ccchancm@hkcc-polyu.edu.hk)

Subject Lecturer

<table>
<thead>
<tr>
<th>Lecturer</th>
<th>Office</th>
<th>Tel</th>
<th>email</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chan Chun Man</td>
<td>WK S1332</td>
<td>3746-0120</td>
<td><a href="mailto:ccchancm@hkcc-polyu.edu.hk">ccchancm@hkcc-polyu.edu.hk</a></td>
</tr>
<tr>
<td>Francis Tong</td>
<td>WK S1310</td>
<td>3746-0486</td>
<td><a href="mailto:ccftong@hkcc-polyu.edu.hk">ccftong@hkcc-polyu.edu.hk</a></td>
</tr>
<tr>
<td>Jodie Lee</td>
<td>HHB 1551</td>
<td>3746-0291</td>
<td><a href="mailto:ccjodie@hkcc-polyu.edu.hk">ccjodie@hkcc-polyu.edu.hk</a></td>
</tr>
<tr>
<td>Peter Lui</td>
<td>HHB 1607</td>
<td>3746-0896</td>
<td><a href="mailto:ccwslui@hkcc-polyu.edu.hk">ccwslui@hkcc-polyu.edu.hk</a></td>
</tr>
<tr>
<td>Wilson Kwan</td>
<td>WK S1312</td>
<td>3746-0258</td>
<td><a href="mailto:ccwilson@hkcc-polyu.edu.hk">ccwilson@hkcc-polyu.edu.hk</a></td>
</tr>
</tbody>
</table>

Learning Outcomes:

On successfully completing this subject, students will be able to:

- Apply statistical techniques and make rigorous statistical analysis in decision making problem.
- Decide which methods can be used to collect, describe and present business data.
- Analyse business data and interpret the results for making recommendations.
- Relate probability theories to solve problems involving uncertainty.
- Use some common statistical packages such as Excel or SPSS.

Tentative Teaching Schedule

<table>
<thead>
<tr>
<th>No</th>
<th>Lecture</th>
<th>Remarks</th>
<th>No</th>
<th>Content (Selected text book exercises)</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Introduction / Data Collection</td>
<td>Sections 1.1 to 1.5</td>
<td>1</td>
<td>Ch1: 2, 3, 5, 11, 12, 22</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Tabular and Graphical Presentation of data</td>
<td>Sections 2.1 to 2.5</td>
<td>2</td>
<td>Ch2: 2, 4, 12, 13, 28, 30, 51</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Numerical Descriptive Measures</td>
<td>Sections 3.1 to 3.6</td>
<td>3</td>
<td>Ch3: 9, 16, 18, 26, 29, 30, 41, 53, 54</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Basic Probability</td>
<td>Sections 4.1 to 4.5</td>
<td>4</td>
<td>Ch4: 2, 4, 9, 10, 15, 18, 23, 28, 33, 36, 39, 42</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Discrete Probability Distributions</td>
<td>Sections 5.1 to 5.6</td>
<td>5</td>
<td>Ch5: 1, 3, 7, 8, 20, 30, 33, 40, 42, 50, 52</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Continuous Probability Distributions</td>
<td>Sections 6.1, 6.2, 6.4</td>
<td>6</td>
<td>Ch6: 2, 4, 6, 13, 15, 18, 23, 35, 54</td>
<td>Submission of Assignment 1</td>
</tr>
<tr>
<td>---</td>
<td>-------------------------------------</td>
<td>-----------------------</td>
<td>---</td>
<td>--------------------------------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td>7</td>
<td>Sampling Distributions</td>
<td>Sections 7.1 to 7.7</td>
<td>7</td>
<td>Ch7: 1, 11, 16, 19, 24, 26, 32, 35, 36</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Interval Estimation</td>
<td>Sections 8.1 to 8.4</td>
<td>8</td>
<td></td>
<td>Review for test</td>
</tr>
<tr>
<td>9</td>
<td>Hypothesis testing</td>
<td>Sections 9.1 to 9.3</td>
<td>9</td>
<td>Ch8: 2, 5, 8, 13, 15, 20, 24, 26, 35, 38</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Hypothesis testing</td>
<td>Sections 9.4 to 9.5</td>
<td>10</td>
<td>Ch9: 2, 5, 10, 11, 15, 18, 20</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Linear Regression</td>
<td>Sections 12.1 to 12.2</td>
<td>11</td>
<td>Ch9: 27, 28, 34, 36, 38, 40</td>
<td>Submission of Assignment 2</td>
</tr>
<tr>
<td>12</td>
<td>Linear Regression</td>
<td>Sections 12.3 to 12.4; and Part of 12.5</td>
<td>12</td>
<td>Ch12: 1, 4, 6,</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Introduction to Time series Analysis</td>
<td></td>
<td>13</td>
<td>Ch12: 15, 16, 18, 22</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Revision</td>
<td></td>
<td>14</td>
<td></td>
<td>Review for final exam</td>
</tr>
<tr>
<td></td>
<td>Make up lectures (if necessary)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Assessment Weighting**

| Coursework: | 40% |
| Examination: | 60% |
| **100%** | |

Students must obtain pass grades in both coursework and examination components in order to pass this subject.

**Assessment Methods for Coursework**

- **Test:** 50% All CC3101 students will write a common Mid-term test. **Tentative schedule: 27 March (Saturday), week 9.**
- **Assignment 1:** 25% (Individual) Due on week 6
- **Assignment 2:** 25% (Individual) Due on week 11

100%

**Assignment:** Every student is required to hand-in 2 individual assignments. Late assignment will receive penalty of marks.

**Attendance and other Rules/Regulations**

The attendance requirement and all other rules and regulations in the HKCC Student Handbook and in the respective Programme Definitive Document apply. Please refer to these documents for details.
Lecture/Tutorial Notes and Assignments

Students are required to download lecture/tutorial notes and assignments from the MOODLES.

Text and References

Textbook:

References:
Subject Learning Outcomes

(a) Apply statistical techniques and make rigorous statistical analysis in decision-making problem.
(b) Decide which methods can be used to collect, describe and present business data.
(c) Analyze business data and interpret the results for making recommendations.
(d) Relate probability theories to solve problems involving uncertainty.
(e) Use some common statistical packages such as Excel or SPSS

Topics to be included in this Subject

Topic 1: Introduction and Data Collection
Topic 2: Tabular and Graphical Presentation of Data
Topic 3: Numerical Descriptive Measures
Topic 4: Basic Probability
Topic 5: Discrete Probability Distributions
Topic 6: Continuous Probability Distributions
Topic 7: Sampling Distribution
Topic 8: Interval Estimation
Topic 9: Hypothesis Testing
Topic 10: Regression Analysis
Topic 11: Introduction to Time Series Analysis
Additional Topic: Examples of Computer Application

<table>
<thead>
<tr>
<th>Topic</th>
<th>(a)</th>
<th>(b)</th>
<th>(c)</th>
<th>(d)</th>
<th>(e)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic 1</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Topic 2</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Topic 3</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Topic 4</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Topic 5</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Topic 6</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Topic 7</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Topic 8</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Topic 9</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Topic 10</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Topic 11</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Additional Topic</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>
## Assessment Criteria: Quality of Assignments

<table>
<thead>
<tr>
<th>Standard</th>
<th>Quality of calculations</th>
<th>Quality of reasoning</th>
<th>Understanding of tools and concepts</th>
<th>Formulation of solution and conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Unsatisfactory</strong></td>
<td>Inaccurate and incomplete</td>
<td>There is no argument. Isolated steps are made but not connected.</td>
<td>Unable to use the tools selected. Unsound approach and analysis.</td>
<td>Fail to formula a solution. Conclusion is illogical or not supported by evidence and critical argument.</td>
</tr>
<tr>
<td><strong>Satisfactory</strong></td>
<td>Slightly inaccurate and incomplete</td>
<td>There is an argument but it is logically flawed and unclear.</td>
<td>Some faults in the use of tools and concepts, but the overall approach is sound.</td>
<td>Attempt to formulate a solution to the problem and draw conclusion. Argument is flawed.</td>
</tr>
<tr>
<td><strong>Good</strong></td>
<td>Accurate and slightly incomplete</td>
<td>The argument seems logically correct but is unclear in some areas.</td>
<td>Make effective use of a range of relevant tools and concepts in the course of analysis.</td>
<td>Attempt to formulate a solution to the problem and draw sensible conclusion with logical argument.</td>
</tr>
<tr>
<td><strong>Excellent</strong></td>
<td>Accurate and complete</td>
<td>The argument is logically correct and clear in all important aspects.</td>
<td>Clearly explain and effectively use the tools and concepts employed in the assignment.</td>
<td>Argument leads to a clear solution and conclusion. Show evidence of critical and creative thinking and originality.</td>
</tr>
</tbody>
</table>
Chapter 1  Basic Definitions of Probability

Definitions:

**Experiment:** A process leading to at least two possible outcomes with uncertainty as to which will occur.

**Sample Space S:** The set of all possible outcomes of an experiment.

**Event:** A subset of a sample space which consists of one or more outcomes with a common characteristic.

1. The **intersection** of two events A and B denoted by \((A \cap B)\) is the event containing all outcomes that are common to A and B.

2. Two or more events are **mutually exclusive** if they have no elements in common (i.e. they cannot occur together).

3. The union of two events A and B denoted by \((A \cup B)\) is the event containing all the elements that belong to A or to B or to both.

4. Events are **collectively exhaustive** if no other outcome is possible for a given experiment.

5. The symbol \(P\) is used to designate the probability of an event. Thus \(P(A)\) denotes the probability that event A will occur in a single observation or experiment.

6. The possible elementary outcomes favorable to event A are defined as \(n(A)\). The possible outcomes included in the sample space are defined as \(n(S)\). If all the elementary outcomes are equally likely and mutually exclusive, then the probability that event A will occur is \(P(A) = \frac{n(A)}{n(S)}\).

7. The smallest value of a probability statement is zero, i.e. \(P(A) = 0\) (indicating the event is impossible). The largest value of a probability statement is one (indicating the event is certain to occur). Thus, in general, \(0 \leq P(A) \leq 1\) for any event A.

8. The **complement** of an event A with respect to S is the set of outcomes of S that is not in A denoted by \(\overline{A}\). Thus, we have \(P(A) + P(\overline{A}) = 1\).
9. The rule of addition for not mutually exclusive events is

\[ P(A \cup B) = P(A) + P(B) - P(A \cap B) \]

Hence, the rule of addition for mutually exclusive (no intersection) events is

\[ P(A \cup B) = P(A) + P(B) \]

10. Let A and B be two events. The conditional probability of event A, given event B, denoted by \( P(A \mid B) \) is defined as \( P(A \mid B) = \frac{P(A \cap B)}{P(B)} \) where \( P(B) > 0 \).

Similarly, we have \( P(B \mid A) = \frac{P(A \cap B)}{P(A)} \) where \( P(A) > 0 \).

11. Two events are independent when the occurrence or non-occurrence of one event has no effect on the probability of occurrence of the other event. Thus, we have

\[ P(A \mid B) = P(A) \]

The rule of multiplication for independent events is \( P(A \cap B) = P(A)P(B) \).

12. **(Law of Total Probability)** Assume that \( B_1, B_2, \ldots, B_n \) are collectively exhaustive events where \( P(B_i) > 0 \), for \( i = 1, 2, \ldots, n \). Events \( B_i \) and \( B_j \) are mutually exclusive for \( i \neq j \). Then for any event \( A \),

\[ P(A) = P(B_1)P(A \mid B_1) + P(B_2)P(A \mid B_2) + \cdots + P(B_n)P(A \mid B_n) \]

13. **(Bayes’ Theorem)** Suppose that \( B_1, B_2, \ldots, B_n \) are \( n \) mutually exclusive events, then we have

\[ P(B_k \mid A) = \frac{P(B_k)P(A \mid B_k)}{P(B_1)P(A \mid B_1) + P(B_2)P(A \mid B_2) + \cdots + P(B_n)P(A \mid B_n)} \]

14. **(Permutations)** The number of permutations of \( n \) objects is the number of ways in which the objects can be arranged in terms of order:

Permutation of \( n \) objects = \( n! = n \times (n - 1) \times \cdots \times (3) \times (2) \times (1) \).

The symbol \( n! \) is read as ‘\( n \) factorial’. It is noted by definition \( 0! = 1 \).

The number of permutation of \( n \) objects taken \( r \) objects at a time, where \( r < n \):

\[ ^nP_r = \frac{n!}{(n-r)!} \]

15. **(Combination)** The combination is a collection of \( n \) objects taken any selections of \( r \) objects at a time where order does not count. The number of combination is

\[ ^nC_r = \binom{n}{r} = \frac{n!}{r!(n-r)!} \]
EXAMPLES

Example 1.1
A manager of a hamburger chain found that 65% of all customers order French Fries, 78% order soft drink, and 55% order both. What is the probability that a customer will order at least one of these?

Solution:
Let \( A \) be the event that a customer orders French Fries. Let \( B \) be the event that the customer orders soft drink. By the given information, we have

\[
P(A) = 0.65; \quad P(B) = 0.78; \quad \text{and} \quad P(A \cap B) = 0.55.
\]

The required probability is

\[
P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.65 + 0.78 - 0.55 = 0.88.
\]

Example 1.2
An urn contains 10 white, 5 yellow, and 10 black marbles. A marble is chosen at random from the urn, and it is not a black marble. What is the probability that it is yellow?

Solution:
Let \( Y \) denote the event that the selected marble is yellow. Let \( B \) denote the event that the selected marble is black. By the conditional probability,

\[
P(Y \mid \overline{B}) = \frac{P(Y \cap \overline{B})}{P(\overline{B})}
\]

However, \( P(Y \cap \overline{B}) = P(Y) \). Since the marble is yellow (i.e. \( Y \)) and not black (i.e. \( \overline{B} \)) if and only if it is yellow. Hence, assuming that each of the 25 marbles is equally likely to be chosen, we obtain that

\[
P(Y \mid \overline{B}) = \frac{P(Y \cap \overline{B})}{P(\overline{B})} = \frac{\frac{5}{25}}{\frac{15}{25}} = \frac{1}{3}.
\]
Example 1.3
Box I contains 2 white and 4 red balls, whereas box II contains 1 white and 1 red ball. A ball is randomly chosen from box I and put into box II, and a ball is then randomly selected from box II.
(a) What is the probability that the ball selected from box II is white?
(b) What is the probability that the transferred ball was white, given that a white ball is selected from box II?

Solution:
Let \( W_1 \) be the event that the ball selected from box I is white. Let \( W_2 \) be the event that the ball selected from box II is white.

From the given information, we obtain
\[
P(W_1) = \frac{2}{6}, \quad P(W_1') = \frac{4}{6}, \quad P(W_2 | W_1) = \frac{2}{3}, \quad P(W_2 | W_1') = \frac{1}{3}
\]

(a) By the law of total probability, we have
\[
P(W_2) = P(W_1 \cap W_2) + P(W_1' \cap W_2)
= P(W_1)P(W_2 | W_1) + P(W_1')P(W_2 | W_1')
= \frac{2}{6} \times \frac{2}{3} + \frac{4}{6} \times \frac{1}{3}
= \frac{4}{9} = 0.4444
\]

(b) The required conditional probability is
\[
P(W_1 | W_2) = \frac{P(W_1 \cap W_2)}{P(W_2)} = \frac{P(W_1)P(W_2 | W_1)}{P(W_2)} = \frac{\frac{2}{6} \times \frac{2}{3}}{\frac{4}{9}} = \frac{1}{2} = 0.5
\]

Examples 1.4
A company received a total of eight applicants for six different jobs. In how many different ways can the jobs be filled with the eight applicants?

Solution:
As the jobs are different, the order of selection is considered (i.e. Permutation). The number of ways that six different jobs can be filled by eight applicants is
\[
P_6 = \frac{8!}{(8-6)!} = 20160.
\]
Example 1.5
How many ways can an executive committee of 5 be chosen from a board of directors consisting of 15 members?

Solution:
Since the order is irrelevant and we use combinations, the answer is

$$\binom{15}{5} = \frac{15!}{5!(15-5)!} = 3003.$$
EXERCISE 1

Question 1.1
How many ways can three items be selected from a group of six items? Use the letters A, B, C, D, E and F to identify the items, and list each of the different combinations of three items.

Question 1.2
An experiment with three outcomes has been repeated 50 times, and it was learned that $E_1$ occurred 20 times, $E_2$ occurred 13 times, and $E_3$ occurred 17 times. Assign probabilities to the outcomes. What method did you use?

Question 1.3
Simple random sampling uses a sample of size n from a population of size N to obtain data that can be used to make inferences about the characteristics of a population. Suppose that a random sample of four accounts is taken from a population of 50 bank accounts. How many different random samples of four accounts are possible?

Question 1.4
Venture capital can provide a big boost in funds available to companies. According to Venture Economics (Investors Business Daily, April 28, 2000), of 2374 venture capital disbursements, 1434 were to companies in California, 390 were to companies in Massachusetts, 217 were to companies in New York, and 112 were to companies in Colorado. Twenty-two percent of the companies receiving funds were in the early stages of development and 55% of the companies were in an expansion stage. Suppose you want to randomly choose one of these companies to learn about how venture capital funds are used,

(a) What is the probability the company chosen will be from California?
(b) What is the probability the company chosen will not be from one of the four states mentioned?
(c) What is the probability the company will not be in the early stages of development?
(d) Assuming the companies in the early stages of development were evenly distributed across the country, how many Massachusetts companies receiving venture capital funds were in their early stages of development?
(e) The total amount of funds invested was $32.4 billion. Estimate the amount that went to Colorado.
Question 1.5
If there is an experiment of selecting a playing card from a deck of 52 playing cards, each card corresponds to a sample point with a 1/52 probability.
(a) List the sample points in the event an ace is selected.
(b) List the sample points in the event a club is selected.
(c) List the sample points in the event a face card (jack, queen or king) is selected.
(d) Find the probabilities associated with each of the events in parts (a), (b) and (c).

Question 1.6
Suppose a manager of a large apartment complex provides the following subjective probability estimates about the number of vacancies that will exist next month,

<table>
<thead>
<tr>
<th>Vacancies</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability</td>
<td>0.05</td>
<td>0.15</td>
<td>0.35</td>
<td>0.25</td>
<td>0.10</td>
<td>0.10</td>
</tr>
</tbody>
</table>

Provide the probability of each of the following events.
(a) No vacancies
(b) At least four vacancies
(c) Two or fewer vacancies

Question 1.7
Suppose we have a sample space S = \{E_1, E_2, E_3, E_4, E_5, E_6, E_7\} where E_1, E_2, E_3, E_4, E_5, E_6 and E_7 denote the sample points, the following probability assignments apply: \( P(E_1) = 0.05 \), \( P(E_2) = 0.20 \), \( P(E_3) = 0.20 \), \( P(E_4) = 0.25 \), \( P(E_5) = 0.15 \), \( P(E_6) = 0.10 \), and \( P(E_7) = 0.05 \). Let A = \{E_1, E_4, E_6\}; B = \{E_2, E_4, E_7\}; and C = \{E_2, E_3, E_5, E_7\}.
(a) Find \( P(A) \), \( P(B) \), and \( P(C) \).
(b) Find \( A \cup B \) and \( P(A \cup B) \).
(c) Find \( A \cap B \) and \( P(A \cap B) \).
(d) Are events A and C mutually exclusive?
(e) Find \( \overline{B} \) and \( P(\overline{B}) \).

Question 1.8
A survey of magazine subscribers showed that 45.8% rented a car during the past 12 months for business reasons, 54% rented a car during the past 12 months for personal reasons, and 30% rented a car during the past 12 months for both business and personal reasons.
(a) What is the probability that a subscriber rented a car during the past 12 months
for business or personal reasons?

(b) What is the probability that a subscriber did not rent a car during the past 12 months for either business or personal reasons?

**Question 1.9**
In a survey of MBA students, the following data were obtained on “students’ first reason for application to the school in which they matriculated”.

<table>
<thead>
<tr>
<th>Enrollment Status</th>
<th>Reason for Application</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>School Quality</td>
<td></td>
</tr>
<tr>
<td>Full Time</td>
<td>421</td>
<td>890</td>
</tr>
<tr>
<td>Part Time</td>
<td>400</td>
<td>1039</td>
</tr>
<tr>
<td>Total</td>
<td>821</td>
<td>1929</td>
</tr>
<tr>
<td></td>
<td>School Cost or Convenience</td>
<td></td>
</tr>
<tr>
<td>Full Time</td>
<td>393</td>
<td></td>
</tr>
<tr>
<td>Part Time</td>
<td>593</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>986</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Others</td>
<td></td>
</tr>
<tr>
<td>Full Time</td>
<td>76</td>
<td></td>
</tr>
<tr>
<td>Part Time</td>
<td>46</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>122</td>
<td></td>
</tr>
</tbody>
</table>

(a) Develop a joint probability table for these data.
(b) Find the marginal probabilities of school quality, school cost or convenience, and other comment.
(c) If a full time MBA student is selected, what is the probability that school quality is the first reason for choosing a school?
(d) If a part time MBA student is selected, what is the probability that school quality is the first reason for choosing a school?
(e) Let A denote the event that a full time MBA student is selected. Let B denote the event that the student lists school quality as the first reason for application. Are events A and B independent? Justify your answer.

**Question 1.10**
Reggie Miller of the Indiana Pacers is the National Basketball Association’s best career free throw shooter, making 89% of his shots (USA Today, January 22, 2004). Assume that late in a basketball game, Reggie Miller is fouled and is awarded two shots.

(a) What is the probability that he will make both shots?
(b) What is the probability that he will make at least one shot?
(c) What is the probability that he will miss both shots?
(d) Late in a basketball game, a team often intentionally fouls an opposing player in order to stop the game clock. The usual strategy is to intentionally foul the other’s team worst free throw shooter. Assume that the Indiana Pacers’ center makes 58% of his free throw shots. Calculate the probabilities for the center as
shown in parts (a), (b) and (c). Show that intentionally fouling the Indiana Pacers’ center is a better strategy than intentionally fouling Reggie Miller.

**Question 1.11**
The probabilities for events $A_1$ and $A_2$ are $P(A_1) = 0.40$ and $P(A_2) = 0.60$. It is also known that $P(A_1 \cap A_2) = 0$. Suppose $P(B \mid A_1) = 0.20$ and $P(B \mid A_2) = 0.05$.

(a) Are $A_1$ and $A_2$ mutually exclusive? Explain.

(b) Compute $P(A_1 \cap B)$ and $P(A_2 \cap B)$.

(c) Compute $P(B)$.

(d) Apply Bayes’ theorem to compute $P(A_1 \mid B)$ and $P(A_2 \mid B)$.

**Question 1.12**
A local bank reviewed its credit card policy with the intention of recalling some of its credit cards. In the past, approximately 5% of cardholders defaulted, leaving the bank unable to collect the outstanding balance. Hence, management established a prior probability of 0.05 that any particular cardholder will default. The bank also found that the probability of missing a monthly payment is 0.20 for customers who do not default. Of course, the probability of missing a monthly payment for those who default is 1.

(a) Given that a customer missed one or more monthly payments, compute the posterior probability that the customer will default.

(b) The bank would like to recall its card if the probability that a customer will default is greater than 0.20. Should the bank recall its card if the customer misses a monthly payment? Why or why not?
Answers to EXERCISE 1:

1.1 20 ways: ABC, ABD, ABE, AEF, ACD, ACE, ACF, ADE, ADF, AEF, BCD, BCE, BCF, BDE, BDF, BEF, CDE, CDF, CEF, DEF.

1.2 \( P(E_1) = 0.40; P(E_2) = 0.26; P(E_3) = 0.34 \) and Relative Frequency.

1.3 230,300.

1.4 a) 0.60; b) 0.09; c) 0.78; d) 86; e) $1.53b.

1.5 a) \( \spadesuit A \heartsuit A \spadesuit A \); b) \( \spadesuit \spadesuit \spadesuit \heartsuit \heartsuit \heartsuit \); c) \( \spadesuit \spadesuit \heartsuit \heartsuit \spadesuit \spadesuit \spadesuit \ ); d) \( P(\text{Ace}) = 0.08, P(\text{Club}) = 0.25, P(\text{J, Q, K}) = 0.23 \).

1.6 a) 0.05; b) 0.20; c) 0.55.

1.7 a) \( P(A) = 0.40, P(B) = 0.50, P(C) = 0.60 \); b) \( \{E_1, E_2, E_4, E_6, E_7\} = A \cup B = \{E_4\} = A \cap B \); c) \( P(A \cup B) = 0.65 \); d) Yes; e) \( B = \{E_1, E_3, E_5, E_6\} \).

1.8 a) 0.698; b) 0.302.

1.9 a) Joint Probability Table

<table>
<thead>
<tr>
<th>Enrollment Status</th>
<th>School Quality</th>
<th>School Cost or Convenience</th>
<th>Others</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full Time</td>
<td>0.218</td>
<td>0.204</td>
<td>0.039</td>
<td>0.461</td>
</tr>
<tr>
<td>Part Time</td>
<td>0.208</td>
<td>0.307</td>
<td>0.024</td>
<td>0.539</td>
</tr>
<tr>
<td>Total</td>
<td>0.426</td>
<td>0.511</td>
<td>0.063</td>
<td>1.000</td>
</tr>
</tbody>
</table>

c) 0.473; d) 0.386;
e) No: \( P(A \cap B) = 0.218 \neq P(A)P(B) = 0.461 \times 0.426 = 0.196 \).

1.10 a) 0.7921; b) 0.9879; c) 0.0121;
d) 0.3364, 0.8236, 0.1764.

1.11 a) Yes: \( P(A_1 \cap A_2) = 0 \); b) 0.08, 0.03; c) 0.11; d) 0.7273, 0.2727.

1.12 a) 0.21; b) Yes.
Chapter 2  Discrete Probability Distributions

1. **Random Variable** A random variable (R.V.) is a variable that takes on different numerical values determined by the outcome of a random experiment. For a discrete random variable, the observed values can occur only at isolated points along a scale of values.

2. **Probability distribution** The probability distribution of a random variable is a representation of the probabilities for all the possible outcomes. This representation might be algebraic, graphical or tabular. A table or a formula listing all possible values that a discrete variable can take on, together with the associated probability is called a discrete probability distribution.

3. **Probability function** The probability function, \( f(x) \), of a discrete random variable \( X \) expresses the probability that \( X \) takes the value \( x \), as a function of \( x \). Thus, \( f(x) = P(X = x) \) where the function is evaluated at all possible values of \( x \).

4. **Expectation** The expected value is the mean of a random variable. The expected value, \( E(X) \), of a discrete random variable \( X \) is defined as 
\[
E(X) = \mu_X = \sum x P(X = x) .
\]

5. **Variance** The variance of a random variable \( X \) is denoted by \( \text{Var}(X) \) or \( \sigma_X^2 \).

The general deviations form of the formula for the variance of a discrete random variable is 
\[
\sigma_X^2 = E[(X - \mu_X)^2] = \sum x^2 P(X = x) - \mu_X^2 .
\]

**Binomial Distribution**

A binomial experiment consists of a sequence of \( n \) identical trials. There are two possible outcomes, success and failure, in each trial. The probability of success in a trial, denoted by \( p \), does not change from trial to trial. The trials are independent. The probability distribution of a binomial random variable \( X \), the number of successes in \( n \) independent trials, is 
\[
P(X = x) = \frac{n!}{x!(n-x)!} p^x (1 - p)^{n-x} \quad \text{where} \quad x = 0,1,2,\ldots,n
\]

The expected number of successes: \( E(X) = np \).

The variance of the number of successes: \( \text{Var}(X) = np(1-p) \).
Hypergeometric Distribution

When sampling is done without replacement of each sampled item taken from a finite population of items, there is a systematic change in the probability of success as items are removed from the population. The hypergeometric distribution is the appropriate discrete probability distribution. Given that $X$ is the designated number of successes, $N$ is the total number of items in the population, $r$ is the total number of successes included in the population, and $n$ is the number of items in the sample, the formula for determining hypergeometric probabilities is

$$P(X = x) = \binom{r}{x} \binom{N-r}{n-x} \binom{N}{n}, \quad \text{for } 0 \leq x \leq r$$

The expected number of successes: $E(X) = \frac{nr}{N}$.

The variance of the number of successes: $Var(X) = n\left(\frac{r}{N}\right)\left(1-\frac{r}{N}\right)\left(\frac{N-n}{N-1}\right)$.

Poisson Distribution

A Poisson process is used to determine the probability of a designated number of events occurring when the events occur in a specified interval of time or space. It is a discrete random variable that may assume an infinite sequence of values $x = 0, 1, 2, \ldots$.

Two Properties of a Poisson Experiment:

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

(ii) The occurrence or nonoccurrence in any interval is independent of the occurrence or nonoccurrence in any other non-overlapping interval.

Denote $\mu$ be the long-run mean number of events for the specific time. The probability of a designated number of successes $X$ in a Poisson distribution is

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

The expected value of $X$: $E(X) = \mu$.

The variance of $X$: $Var(X) = \mu$. 

20
EXAMPLES

Example 2.1
The probability that a randomly chosen sales prospect will make a purchase is 0.2. If a sales representative calls on six prospects, what is the probability that exactly four sales will be made?

Solution:
The required probability is

\[ P(X = 4) = \frac{6!}{4!(6 - 4)!} (0.2)^4 (1 - 0.2)^{6-4} = 0.015. \]

Example 2.2
Of six employees, three have been with the company for five or more years. If four employees are chosen randomly from the group of six, what is the probability that exactly two will have five or more years’ seniority?

Solution:
The required probability is

\[ P(X = 2) = \binom{3}{2} \binom{6-3}{4-2} = 0.6. \]

Example 2.3
On average five calls for service per hour are received by a machine repair department. What is the probability that exactly three calls for service will be received in a randomly selected hour?

Solution:
The required probability is

\[ P(X = 3) = \frac{5^3 e^{-5}}{3!} = 0.1404. \]
EXERCISE 2

**Question 2.1**
Consider the experiment of tossing a coin twice.
(a) List the experimental outcomes.
(b) Define a random variable that represents the number of heads occurring on the two tosses.
(c) Show what value the random variable would assume for each of the experimental outcomes.
(d) Is this random variable discrete or continuous?

**Question 2.2**
Three students scheduled interviews for summer employment at the Brookwood Institute. In each case, the interview results are either an offer for a position or no offer. Experimental outcomes are defined in terms of the results of the interviews.
(a) List the experimental outcomes.
(b) Define a random variable that represents the number of offers made. Is the random variable continuous?
(c) Show the value of the random variable for each of the experimental outcomes.

**Question 2.3**
The probability distribution for the random variable $x$ is shown as follows:

<table>
<thead>
<tr>
<th>$x$</th>
<th>$f(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>.20</td>
</tr>
<tr>
<td>25</td>
<td>.15</td>
</tr>
<tr>
<td>30</td>
<td>.25</td>
</tr>
<tr>
<td>35</td>
<td>.40</td>
</tr>
</tbody>
</table>
(a) Is this probability distribution valid? Explain.
(b) What is the probability that $x = 30$?
(c) What is the probability that $x$ is less than or equal to 25?
(d) What is the probability that $x$ is greater than 30?

**Question 2.4**
The following data were collected by counting the number of operating rooms in use at Tampa General Hospital over a 20-day period: On three of the days only one operating room was used, on five of the days two were used, on eight of the days three were used, and on four days all four of the hospital’s operating rooms were used.
(a) Use the relative frequency approach to construct a probability distribution for
the number of operating rooms in use on any given day.

(b) Draw a graph of the probability distribution.

**Question 2.5**

The following table provides a probability distribution for the random variable \( y \).

<table>
<thead>
<tr>
<th>( y )</th>
<th>( P(Y) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>.20</td>
</tr>
<tr>
<td>4</td>
<td>.30</td>
</tr>
<tr>
<td>7</td>
<td>.40</td>
</tr>
<tr>
<td>8</td>
<td>.10</td>
</tr>
</tbody>
</table>

(a) Compute \( E(y) \).

(b) Compute \( \text{Var}(y) \) and \( \sigma \).

**Question 2.6**

The American Housing Survey reported the following data on the number of bedrooms in owner-occupied and renter-occupied houses in central cities (http://www.census.gov, March 31, 2003).

<table>
<thead>
<tr>
<th>Number of Houses (1000s)</th>
<th>No. of Bedrooms</th>
<th>Renter-Occupied</th>
<th>Owner-Occupied</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>547</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>5012</td>
<td>541</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>6100</td>
<td>3832</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2644</td>
<td>8690</td>
<td></td>
</tr>
<tr>
<td>4 or more</td>
<td>557</td>
<td>3783</td>
<td></td>
</tr>
</tbody>
</table>

(a) Define a random variable \( x \) = number of bedrooms in renter-occupied houses and develop a probability distribution for the random variable. (Let \( x = 4 \) represent 4 or more bedrooms.)

(b) Compute the expected value and variance for the number of bedrooms in renter-occupied houses.

(c) Define a random variable \( y \) = number of bedrooms in owner-occupied houses and develop a probability distribution for the random variable. (Let \( y = 4 \) represent 4 or more bedrooms.)

(d) Compute the expected value and variance for the number of bedrooms in owner-occupied houses.

**Question 2.7**

The probability distribution for damage claims paid by the Newton Automobile Insurance Company on collision insurance is shown as follows:
<table>
<thead>
<tr>
<th>Payment ($)</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>.90</td>
</tr>
<tr>
<td>400</td>
<td>.04</td>
</tr>
<tr>
<td>1000</td>
<td>.03</td>
</tr>
<tr>
<td>2000</td>
<td>.01</td>
</tr>
<tr>
<td>4000</td>
<td>.01</td>
</tr>
<tr>
<td>6000</td>
<td>.01</td>
</tr>
</tbody>
</table>

(a) Use the expected collision payment to determine the collision insurance premium that would enable the company to break even.

(b) The insurance company charges an annual rate of $260 for the collision coverage. What is the expected value of the collision policy for a policyholder? (Hint: It is the expected payments from the company minus the cost of coverage.) Why does the policyholder purchase a collision policy with this expected value?

**Question 2.8**

When a new machine is functioning properly, only 3% of the items produced are defective. Assume that two parts produced by the machine are randomly selected and we are interested in the number of defective parts found,
(a) How many experimental outcomes result in exactly one defect being found?
(b) Compute the probabilities associated with finding no defects, exactly one defect, and two defects.

**Question 2.9**

Forty percent of business travelers carry either a cell phone or a laptop (USA Today, September 12, 2000). A sample of 15 business travelers is randomly selected.
(a) Compute the probability that three of the travelers carry a cell phone or laptop.
(b) Compute the probability that 12 of the travelers carry neither a cell phone nor a laptop.
(c) Compute the probability that at least three of the travelers carry a cell phone or a laptop.

**Question 2.10**

Phone calls arrive at the rate of 48 per hour at the reservation desk for Regional Airways.
(a) Compute the probability of receiving three calls in a five-minute interval of time.
(b) Compute the probability of receiving exactly 10 calls in 15 minutes.
(c) Suppose no calls are currently on hold, if the agent takes five minutes to complete the current call, how many callers do you expect to be waiting by that time? What is the probability that none will be waiting?
(d) If no calls are currently being processed, what is the probability that the agent can take three minutes for personal time without being interrupted by a call?

**Question 2.11**
More than 50 million guests stayed at Bed and Breakfasts (B&Bs) last year. The Web site for the Bed and Breakfast Inns of North America (www.bestinns.net), which averages approximately seven visitors per minute, enables B&Bs to attract guests without waiting years to be mentioned in guidebooks (Time, September 2001).

(a) Compute the probability of no Web site visitors in a one-minute period.
(b) Compute the probability of two or more Web site visitors in a one-minute period.
(c) Compute the probability of one or more Web site visitors in a 30-second period.
(d) Compute the probability of five or more Web site visitors in a one-minute period.

**Question 2.12**
Axline Computers manufactures personal computers at two plants, one in Texas and the other in Hawaii. The Texas plant has 40 employees; the Hawaii plant has 20. A random sample of 10 employees is to be asked to fill out a benefits questionnaire.

(a) What is the probability that none of the employees in the sample work at the plant in Hawaii?
(b) What is the probability that one of the employees in the sample works at the plant in Hawaii?
(c) What is the probability that two or more of the employees in the sample work at the plant in Hawaii?
(d) What is the probability that nine of the employees in the sample work at the plant in Texas?

**Question 2.13**
A shipment of ID items has two defective and eight non-defective items. In the inspection of the shipment, a sample of items will be selected and tested. If a defective item is found, the shipment of 10 items will be rejected.

(a) If a sample of three items is selected, what is the probability that the shipment will be rejected?
(b) If a sample of four items is selected, what is the probability that the shipment
will be rejected?
(c) If a sample of five items is selected, what is the probability that the shipment will be rejected?
(d) If management would like a .90 probability of rejecting a shipment with two defective and eight non-defective items, how large a sample would you recommend?

Answers to EXERCISE 2
2.1 a) \{H, H\}, \{H, T\}, \{T, H\}, \{T, T\};
    b) Number of heads; c) 0, 1, 2; d) discrete.
2.2 a) \{N, N, N\}, \{N, N, O\}, \{N, O, N\}, \{O, N, N\}, \{N, O, O\}, \{O, O, N\}, \{O, O, O\}; b) Number of offers made, No; c) 0, 1, 2, 3.
2.3 a) Yes; b) 0.25; c) 0.35; d) 0.4.
2.4 a) Probability Distribution

<table>
<thead>
<tr>
<th>No. of operating rooms in use, x</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(x)</td>
<td>0.15</td>
<td>0.25</td>
<td>0.40</td>
<td>0.20</td>
</tr>
</tbody>
</table>

2.5 a) 5.20; b) 4.56, 2.14.
2.6 a) Probability Distribution for Renter-Occupied Houses

<table>
<thead>
<tr>
<th>No. of bedrooms in renter-occupied houses, x</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(x)</td>
<td>0.037</td>
<td>0.337</td>
<td>0.411</td>
<td>0.178</td>
<td>0.037</td>
</tr>
</tbody>
</table>

b) 1.84, 0.79;

c) Probability Distribution for Owner-Occupied Houses

<table>
<thead>
<tr>
<th>No. of bedrooms in owner-occupied houses, x</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(x)</td>
<td>0.002</td>
<td>0.032</td>
<td>0.227</td>
<td>0.515</td>
<td>0.224</td>
</tr>
</tbody>
</table>

d) 2.93, 0.59.
2.7 a) 166; b) -94.
2.8  a) two; b) 0.9409, 0.0582, 0.009.
2.9  a) 0.0634; b) 0.0634; c) 0.9729.
2.10 a) 0.1952; b) 0.1048; c) 3, 0.0183; d) 0.0907.
2.11 a) 0.0009; b) 0.9927; c) 0.9698; d) 0.8271.
2.12 a) 0.01; b) 0.07; c) 0.92; d) 0.92.
2.13 a) 0.5333; b) 0.6667; c) 0.7778; d) 0.9333.
Chapter 3  Confidence Interval

An interval is constructed around the point estimate, and it is stated that this interval is likely to contain the corresponding population parameter. Interval estimates indicate the precision, or accuracy, of an estimate and are therefore preferable.

1. Confidence Interval for population mean $\mu$ with Known variances

In the case of sample mean, the sample size is considered to be large when the sample size is 30 or larger. According to central limit theorem, for a large sample the sampling distribution of the sample mean is (approximately) normal irrespective of the shape of the population from which the sample is drawn.

When the sample size is 30 or larger, we will use the normal distribution to construct a confidence interval for $\mu$.

$$\bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Where

- $\bar{X}$ = sample mean
- $\sigma$ = population standard deviation
- $N$ = the sample size
- $Z_{\alpha/2}$ = read from the standard normal distribution

2. Confidence Interval for population mean with Unknown variance

If the sample size is small, the normal distribution can still be used to construct a confidence interval for $\mu$ if (1) the population from which the sample is drawn is normally distributed, and (2) the value of $\sigma$ is known. But more often we do not know $\sigma$ and, consequently, we have to use the sample standard deviation $s$ as an estimator of $\sigma$. In such case, the normal distribution cannot be used to make confidence intervals about $\mu$. When (1) the population from which the sample is selected is (approximately) normally distributed, and (2) the population standard deviation $\sigma$ is not known, the normal distribution is replaced by the $t$ distribution to construct confidence interval about $\mu$. 
\[ \bar{X} \pm t_{\alpha/2, n-1} \frac{s}{\sqrt{n}} \]

Where

- \( \bar{X} \) = sample mean
- \( s \) = population standard deviation
- \( n \) = the sample size
- \( t_{\alpha/2, n-1} \) = read from the \( t \) distribution table for \( n - 1 \) degrees of freedom

3. Interval Estimation of a population proportion with Large Sample

Recall that the population proportion is denoted by \( p \) and the sample proportion is denoted by \( \bar{p} \). The sample proportion \( \bar{p} \) is a sample statistic, and it possesses a sampling distribution. For large samples:

1. The sampling distribution of the sample proportion \( \bar{p} \) is (approximately) normal.
2. The mean of the sampling distribution of \( \bar{p} \) is equal to the population proportion \( p \).
3. The standard deviation of the sampling distribution of the sample proportion \( \bar{p} \) is
   \[ \sqrt{\frac{p(1 - p)}{n}} . \]

When estimating the value of a population proportion, we do not know the values of \( p \) and \( 1 - p \). In this case, we will use
\[ \sqrt{\frac{\bar{p}(1 - \bar{p})}{n}} \] to estimate the population proportion.

\[ \bar{p} \pm Z_{\alpha/2} \sqrt{\frac{\bar{p}(1 - \bar{p})}{n}} \]

Where

- \( \bar{p} \) = the sample proportion
- \( N \) = the sample size
- \( Z_{\alpha/2} \) = read from the standard normal distribution
4. Sample Size Determination

One reason why we usually conduct a sample survey and not a census is that almost always we have limited resources at our disposal. In light of this, if a smaller sample can serve our purpose, then we will be wasting our resources by taking a larger sample.

a) For the estimation of population mean

\[ n = \frac{Z_{\alpha/2}^2 \sigma^2}{E^2} \]

\( \sigma \) is population standard deviation
\( Z_{\alpha/2} \) is read from the standard normal distribution
\( \bar{p} \) is the sample proportion
E is the maximum error

b) For the estimation of proportion

\[ n = \frac{Z_{\alpha/2}^2 \bar{p}(1 - \bar{p})}{E^2} \]

EXAMPLES

Example 3.1
A simple random sample of 9 items from a population with \( \sigma = 2.5 \) resulted in a sample mean of 30.11. Find the 90% confidence interval for the population mean.

Solution

\[ \bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \]
\[ = 30.11 \pm 2.33 \times \frac{2.5}{\sqrt{9}} \]
\[ = 30.11 \pm 1.9417 \]
\[ = 28.168 \text{ to } 32.052 \]
Example 3.2
An educational organization conducted a piece of research on the weekly pocket money of 9 different primary school students in Wan Chai.

| 28 | 32 | 32 | 27 | 31 | 32 | 29 | 26 | 34 |

What is the 90% confidence interval for the mean pocket money of primary school students in Hong Kong? What assumption about the distribution should be made?

Solution

\[ \bar{X} = 30.11, \ s = 2.7131, \ t_{0.05,8} = 1.8595 \]

\[ \bar{X} \pm t_{\alpha/2, n-1} \frac{s}{\sqrt{n}} \]
\[ = 30.11 \pm 1.8595 \times \frac{2.7131}{\sqrt{9}} \]
\[ = 30.11 \pm 1.6817 \]
\[ = 28.428 \text{ to } 31.792 \]

Example 3.3
Suppose 18% of the sampled students are smokers, make a 99% confidence interval to estimate the true proportion of students who are smokers if the sample size of the survey is 200.

Solution

\[ \bar{p} \pm Z_{\alpha/2} \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \]
\[ = 0.18 \pm 2.58 \times \sqrt{\frac{0.18(1-0.18)}{200}} \]
\[ = 0.18 \pm 0.07 \]
\[ = 0.11 \text{ to } 0.25 \]
EXERCISE 3

Question 3.1
In a survey, the researcher wants to estimate the mean amount per customer spent in a supermarket; data were collected from a sample of 50 customers. Assume a population standard deviation of $6,
a) Find the margin of error.
b) If the sample mean is $32, what is the 95% confidence interval for the population mean?

Question 3.2
Referring to Question 3.1, will the width of the confidence interval be increased if 99% confidence interval is used?

Question 3.3
The number of books in the school bags of 8 primary schools students is shown below.

| 5 | 6 | 11 | 13 | 12 | 15 | 8 | 10 |

What is the 95% confidence interval for the population mean?

Question 3.4
A final year student conducted a survey on traffic counts on a main road in every minute. Given the mean of 65 minutes is 19.5 and the standard deviation is 5.2. What is the 90% for the population mean traffic counts for the population?

Question 3.5
A survey showing that 46% of construction workers from a total of 611 workers indicated that they are suffered from respiratory diseases. What is the 90% confidence interval for the proportion of the population of construction workers who are suffering from the disease?

Question 3.6
Suppose 64.2% of 162 children play TV games every day, what is the margin of error for the 95% confidence interval in this claim?

Question 3.7
Suppose the range for a set of data is estimated to be 12, how large a sample would provide a margin of error of 4 at 95% confidence?
Question 3.8
Referring to Question 3.6, how large a sample is needed if the desired margin of error is 0.1?

Answers to EXERCISE 3
3.1  a) 1.66  b) 30.34 to 33.66
3.2  Increased. 29.81 to 34.19
3.3  7.1 to 12.9
3.4  18.21 to 20.79
3.5  0.4268 to 0.4932
3.6  0.0738
3.7  2.16 or 3
3.8  88.29 or 89
Chapter 4  Hypothesis Testing

A hypothesis is an assumption – a statement made to explain a set of facts and to form a basis for further investigation. It is understood that the statement is subject to proof or checking. The testing of hypothesis is the second major part of statistical inference. It is of great importance because it is used as the basis for decision-making in industry, business and government.

Basic Concept of Hypothesis Testing:

Statistical testing begins with a hypothesis – an assumption about the value of a population parameter. A sample is chosen from the population, and the value of the sample statistics is calculated. A decision then has to be made. If there is no significant difference between the values, the hypothesis may be accepted; if there is a difference, it may be rejected. These decisions are made on the significance size of the difference.

1. Hypothesis Test about a Population Mean with Known Population Variance

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>One-Tailed Test</th>
<th>One-Tailed Test</th>
<th>Two-Tailed Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null</td>
<td>( H_0 : \mu \geq \mu_0 )</td>
<td>( H_0 : \mu \leq \mu_0 )</td>
<td>( H_0 : \mu = \mu_0 )</td>
</tr>
<tr>
<td>Alternative</td>
<td>( H_0 : \mu &lt; \mu_0 )</td>
<td>( H_0 : \mu &gt; \mu_0 )</td>
<td>( H_0 : \mu \neq \mu_0 )</td>
</tr>
</tbody>
</table>

Rejection Rule

Reject \( H_0 \) if \( Z < -Z_\alpha \)

Reject \( H_0 \) if \( Z > Z_\alpha \)

Reject \( H_0 \) if \( Z > Z_\alpha \) or \( Z < -Z_\alpha \)

Test statistics:

\[
Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}
\]

2. Concept of \( p \)-value

The \( p \)-value can be used to make the decision for hypothesis test by noting that if the \( p \)-value is less than the level of significance, \( \alpha \), the value of the test statistics must be in the rejection region. Similarly, if the \( p \)-value is greater than or equal to \( \alpha \), the value of the test statistic is not in the rejection region.

The \( p \)-value is the smallest level of significance \( \alpha \) for which the sample data indicate that the null hypothesis should be rejected.
If $H_0 : \mu \leq \mu_0$ and $H_a : \mu > \mu_0$ Then $p-value = P\left( Z > \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \right)$

If $H_0 : \mu \geq \mu_0$ and $H_a : \mu < \mu_0$ Then $p-value = P\left( Z < \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \right)$

**p-value Criterion for Hypothesis Testing:**

Reject $H_0$ if the $p-value < \alpha$

---

3. **Errors involved in Hypothesis Testing**

Ideally, the hypothesis procedure would always lead us to reject $H_0$ when it is true and reject $H_0$ when it is false. This is not always the case, for hypothesis testing errors can occur.

Type I error occurs if we reject a null hypothesis when it is true. The probability of committing of Type I error is usually denoted by $\alpha$. This is called the level of significance for a hypothesis test.

Type II error occurs if we accept a null hypothesis when it is false. The probability of making Type II error is usually denoted by $\beta$.

<table>
<thead>
<tr>
<th>Conclusion</th>
<th>State of Nature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Do not reject $H_0$</td>
<td>$H_0$ true</td>
</tr>
<tr>
<td>Reject $H_0$</td>
<td>Correct Conclusion</td>
</tr>
<tr>
<td></td>
<td>Type I Error</td>
</tr>
</tbody>
</table>

4. **Hypothesis Test about a Population Mean with Unknown Population Variance**

Suppose the population standard deviation is unknown, if we can assume that the population has a normal distribution, the $t$ distribution can be used to make inference about the value of a population mean.
### Hypothesis Test

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>One-Tailed Test</th>
<th>One-Tailed Test</th>
<th>Two-Tailed Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null</td>
<td>$H_0: \mu \geq \mu_0$</td>
<td>$H_0: \mu \leq \mu_0$</td>
<td>$H_0: \mu = \mu_0$</td>
</tr>
<tr>
<td>Alternative</td>
<td>$H_0: \mu &lt; \mu_0$</td>
<td>$H_0: \mu &gt; \mu_0$</td>
<td>$H_0: \mu \neq \mu_0$</td>
</tr>
<tr>
<td>Rejection Rule</td>
<td>Reject $H_0$ if $t &lt; -t_\alpha$</td>
<td>Reject $H_0$ if $t &gt; t_\alpha$</td>
<td>Reject $H_0$ if $t &gt; t_\alpha$ or $t &lt; -t_\alpha$</td>
</tr>
<tr>
<td>Degree of freedom</td>
<td>$v = n-1$</td>
<td>$v = n-1$</td>
<td>$v = n-1$</td>
</tr>
<tr>
<td>Test statistics</td>
<td>$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$</td>
<td>$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$</td>
<td>$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$</td>
</tr>
</tbody>
</table>

### 5. Small Sample Hypothesis Test about a Population Proportion: Large Samples

We can use the following hypothesis test if $np \geq 5$ and $n(1-p) \geq 5$. In this case, we assume the distribution of $\bar{p}$ is a normal distribution.

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>One-Tailed Test</th>
<th>One-Tailed Test</th>
<th>Two-Tailed Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null</td>
<td>$H_0: p \geq p_0$</td>
<td>$H_0: p \leq p_0$</td>
<td>$H_0: p = p_0$</td>
</tr>
<tr>
<td>Alternative</td>
<td>$H_0: p &lt; p_0$</td>
<td>$H_0: p &gt; p_0$</td>
<td>$H_0: p \neq p_0$</td>
</tr>
<tr>
<td>Rejection Rule</td>
<td>Reject $H_0$ if $Z &lt; -Z_\alpha$</td>
<td>Reject $H_0$ if $Z &gt; Z_\alpha$</td>
<td>Reject $H_0$ if $Z &gt; Z_\alpha$ or $Z &lt; -Z_\alpha$</td>
</tr>
<tr>
<td>Test statistics</td>
<td>$Z = \frac{\bar{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$</td>
<td>$Z = \frac{\bar{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$</td>
<td>$Z = \frac{\bar{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$</td>
</tr>
</tbody>
</table>

### EXAMPLES

**Example 4.1**

The supervisor of a supermarket has recently surveyed a random sample of 250 customers. He likes to determine whether or not the mean spending of his customers is over 60. Suppose he found that the sample mean was 60.45 and the population standard deviation was 5, at 2.5% significance level, test the hypothesis that the mean spending of the customers is over 60.

**Solution**

$H_0: \mu = 60$

$H_1: \mu > 60$

Since $\alpha = 0.025$, thus the critical $z$ value for this one tail test is 1.96.

We reject $H_0$ if $z \geq 1.96$
\[ z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} = \frac{60.45 - 60}{5 / \sqrt{250}} = 1.423 \]

Since the calculated \( z \) value (1.423) is less than the critical \( z \) value (1.96), thus we do not reject \( H_0 \).

**Example 4.2**  
Given that the sample mean of 36 random samples is 23.39 and the sample standard deviation is 12.64, test the following hypothesis:

\[ H_0 : \mu \leq 25 \]
\[ H_1 : \mu > 25 \]

**Solution**

\[ t = \frac{\bar{X} - \mu}{s / \sqrt{n}} = \frac{23.39 - 25}{2.11} = -0.7630 \]

And \( t_{0.05,35} = 1.6896 \)

Reject rule: Reject \( H_0 \) if \( t \leq -1.6896 \)

Since -0.7630 is not smaller than -1.6896, we cannot reject \( H_0 \).

So the data do not support the claim.

**Example 4.3**  
The supervisor of a supermarket has recently surveyed a random sample of 250 customers. He found that 120 of the 250 customers like to drink coke every day. At 10% significance level, test the hypothesis that the true proportion of customers who like to drink coke every day is different from 40%.

**Solution**

\[ H_0 : p = 0.4 \]
\[ H_1 : p \neq 0.4 \]

Since \( \alpha \) = 0.1, thus the critical \( z \) values for this two tail tests are \pm 1.645.

We reject \( H_0 \) if \( z \geq 1.645 \) or \( z \leq -1.645 \)

\[ z = \frac{\bar{p} - p}{\sqrt{p(1-p)} / \sqrt{n}} = \frac{120}{250} - 0.4 = 2.582 \]

Since the calculated \( z \) value (2.5820) is greater than the critical \( z \) value (1.645), thus we reject \( H_0 \).
EXERCISE 4

Question 4.1
Consider the following hypothesis test:

\[ H_0 : \mu \leq 45 \]
\[ H_1 : \mu > 45 \]

A sample of 40 provided a sample mean of 46.4. The population standard deviation is 6.
(a) Compute the value of test statistics.
(b) At \( \alpha = 0.01 \), what is the conclusion using critical value approach?

Question 4.2
Consider the following hypothesis test:

\[ H_0 : \mu \leq 95 \]
\[ H_1 : \mu > 95 \]

A sample of 40 provided a sample mean of 96.4. The population standard deviation is 6.
(a) Find the \( p \)-value.
(b) At \( \alpha = 0.01 \), what is the conclusion using \( p \)-value approach?

Question 4.3
Suppose the average mark of 10 short questions in a particular test is 6.1, the mean of a sample of 40 is found to be 5.4. Assume the population standard deviation is 2, a teacher would like to conduct a hypothesis test to see whether the population mean is significantly different from the average mark of a test. At \( \alpha = 0.05 \), test the hypothesis and state your conclusion using critical value approach.

Question 4.4
Referring to the information given in question 3, what is the conclusion if \( p \)-value is used?

Question 4.5
A random sample of 14 data is selected as follows.

<table>
<thead>
<tr>
<th>20.9</th>
<th>16.7</th>
<th>18.3</th>
<th>21.0</th>
<th>15.4</th>
<th>17.8</th>
<th>15.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>13.2</td>
<td>14.5</td>
<td>20.6</td>
<td>12.8</td>
<td>24.3</td>
<td>16.9</td>
<td>18.8</td>
</tr>
</tbody>
</table>

(a) Test, at 10% significance level, the claim that the mean value is different from 18.
(b) Explain in the context of the above scenario the meaning of Type II error.
Question 4.6
Suppose a primary school student drinks at least the following amount of water (in L) every week.

<table>
<thead>
<tr>
<th>Amount (L)</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.48</td>
</tr>
<tr>
<td>7.74</td>
</tr>
<tr>
<td>9.93</td>
</tr>
<tr>
<td>10.88</td>
</tr>
<tr>
<td>10.84</td>
</tr>
<tr>
<td>13.95</td>
</tr>
<tr>
<td>15.98</td>
</tr>
<tr>
<td>10.63</td>
</tr>
<tr>
<td>10.01</td>
</tr>
<tr>
<td>8.89</td>
</tr>
<tr>
<td>12.78</td>
</tr>
</tbody>
</table>

(a) Use the sample results and 5% significance level to test the claim that the mean amount of water drunk is different from 12 L.

(b) Explain in the context of the above scenario the meaning of Type I error.

Question 4.7
A recent survey of 400,000 randomly selected smokers in China showed that 200 of them suffered from respiratory disease. For those non-smokers, it was found that the rate of such disease is 0.052%. At 5% significance level, test the claim that the rate for smokers suffering from respiratory disease is greater than the rate for non-smokers who suffer from respiratory disease.

Question 4.8
Education researchers conduct a survey on whether primary school students enjoy learning English language. A total of 26 out of 36 respondents claimed that they enjoyed. Is it reasonable to conclude that a majority of the students agree that they enjoy learning English language? Support your answer by a hypothesis test with 5% significance level.

Answers to EXERCISE 4

4.1 a) 1.48
   b) Since 1.48 < 2.33, we do not reject $H_0$. The population mean is not significantly larger than 45.

4.2 a) $p$-value is 0.0694
   b) Do not reject $H_0$.

4.3 We reject $H_0$ and conclude that the true mean is significantly different from the average mark of a test.

4.4 $p$-value = 0.0272. Since $p$-value = 0.0272 < 0.05, same conclusion as question 3.

4.5 $t = -0.45974$, do not reject $H_0$.
   We conclude that the data show no evidence that the mean value is different from 18 but indeed it is different from 18.

4.6 Since the value of the test statistics does not fall into the rejection region, thus we
do not reject $H_0$.

Type I error occurs when you conclude that the mean water drunk is different from 12 when in fact the mean water drunk is equal to 12.

4.7 The value of the test statistics falls into the rejection region, thus we reject $H_0$.

4.8 Since $z = 2.6667 > 1.645$. Reject $H_0$ and conclude that the students enjoy learning English language.
Chapter 5  Continuous Distribution

A continuous random variable is defined as a random variable whose values are not countable. A continuous random variable can assume any value over an interval or intervals. Because the number of values contained in any interval is infinite, the possible number of values that a continuous random variable can assume is also infinite.

1. Normal Distribution

The normal distribution is the most important and most widely used of all the probability distributions. A large number of phenomena in the real world are normally distributed either exactly or approximately.

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

**Standardization:**

For a normal random variable $X$, a particular value of $X$ can be converted to a $Z$ value by using the formula

$$Z = \frac{X - \mu}{\sigma}$$

2. Uniform Distribution

A continuous uniform probability distribution is a simple distribution with a rectangular shape, and it is useful in a diverse number of applications. For example, the time that a commuter waits to board a MTR train from Central to Wan Chai has a uniform distribution.
A continuous random variable $X$ is said to have a continuous uniform probability distribution on the interval $(a,b)$ if and only if the probability density function is

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{elsewhere} \end{cases}$$

The mean and variance of a continuous uniform probability for each distribution $X$ are

$$E(X) = \frac{a + b}{2} \quad \text{and} \quad Var(x) = \frac{(b-a)^2}{12}$$

3. The Exponential Distribution.

A continuous probability distribution that is often useful in describing the time it takes to complete a task is the exponential probability distribution. The exponential random variable can be used to describe such things as the time between arrivals at a car wash.

The exponential random variable can be used to describe such things as the time between arrivals. The exponential probability density function is as follows:

$$f(x) = \frac{1}{\mu} e^{-x/\mu} \quad \text{for } x \geq 0, \mu > 0.$$ 

Five additional formulas
(a) $P(X \leq c) = 1 - e^{-c/\mu}$
(b) $P(X \geq c) = e^{-c/\mu}$
(c) $P(c \leq X \leq d) = e^{-c/\mu} - e^{-d/\mu}$
(d) $E(x) = \mu$
(e) $Var(x) = \mu^2$

EXAMPLES

Example 5.1
The life of newly developed light bulbs is normally distributed with a mean of 18
months and a standard deviation of 5 months. The company that produces the light bulb is considering a warranty for the light bulbs.
(a) What proportion of the light bulbs that have a life of more than 28 months?
(b) What proportion of the light bulbs last between 12 and 27 months?
(c) If the manager of the company wants to replace less than 7% of the light bulbs under a warranty, how many months should the warranty of the light bulb company cover?

Solution
(a) 
\[ X \sim N(18,25) \]
\[ P(X > 28) = P\left( \frac{X - \mu}{\sigma} > \frac{28 - 18}{5} \right) = P(Z > 2) = 0.0228 \]
(b) 
\[ P(12 < X < 27) = P\left( \frac{12 - 18}{5} < \frac{X - \mu}{\sigma} < \frac{27 - 18}{5} \right) = P(-1.2 < Z < 1.8) = 0.849 \]
(c) 
\[ A = \mu + z\sigma = 18 + (-1.48)(5) = 10.6 \]

Example 5.2

The random variable \( x \) is known to be uniformly distributed between 10 and 20.
(a) Compute \( P(x < 15) \)
(b) Compute \( P(12 \leq x \leq 18) \).
(c) Compute \( E(x) \)
(d) Compute \( Var(x) \)

Solution
(a) \( P(x < 15) = 0.1(5) = 0.5 \)
(b) \( P(12 \leq x \leq 18) = 0.1(6) = 0.6 \)
(c) \( E(x) = \frac{10 + 20}{2} = 15 \)
(d) \( Var(x) = \frac{(20 - 10)^2}{12} = 8.33 \)

Example 5.3

Suppose the time between car wash can be modeled by an exponential distribution
with a mean of 8 minutes, if a car has just arrived,
(a) Find the probability that no car wash within 7 minutes.
(b) Find the probability that at least one car washes within 10 minutes.
(c) Determine, \( k \), such that the probability that at least one car washes before time \( k \) minutes is 0.95.

**Solution**

(a) Given \( E(X) = \mu = 8 \)
\[
P(X > 7) = e^{-7/8} = 0.41686
\]
(b) \( P(X < 10) = 1 - e^{-10/8} = 0.7135 \)
(c) \( P(X < k) = 0.95 \)
\[
1 - e^{-k/8} = 0.95 \\
e^{-k/8} = 0.05 \\
-k/8 = \ln 0.05 \\
k = 23.966
\]

**Exercise 5**

**Question 5.1**
Given that \( z \) is a standard normal random variable, compute the following:

a) \( P(-1.98 \leq z \leq 0.49) \)

b) \( P(0.52 \leq z \leq 1.22) \)

c) \( P(-1.75 \leq z \leq -1.04) \)

**Question 5.2**
Given that \( z \) is a standard normal random variable, find \( z \) for each of the following situations.

a) The area to the left of \( z \) is 0.02119.

b) The area between \(-z\) and \( z \) is 0.9030.

c) The area between \(-z\) and \( z \) is 0.2052.

d) The area to the left of \( z \) is 0.9948.

**Question 5.3**
The average selling price for a new designed drink is $30, and the standard deviation is $8.2. Assume the price of the drink is normally distributed,

(a) What is the probability that the selling price of a drink is at least $40?
(b) What is the probability that the selling price of a drink is no higher than $20?
(c) How high does the selling price of a drink have to be put the drink in the top 10%?

**Question 5.4**
Consider a random sample under normal distribution with the mean of 2.17 and standard deviation of 0.21, find the value of $K$ if the probability that one randomly selected component has length greater than $K$ is 0.8888.

**Question 5.5**
Suppose
\[
f(x) = \begin{cases} 
0 & \text{for} \quad 0 \leq x \leq 1 \\
1 & \text{elsewhere}
\end{cases}
\]
What is the probability of generating a random number between 0.25 and 0.75?

**Question 5.6**
A continuous random variable $x$ is uniformly distributed between $a$ and 60.
(a) If $P(X > 30) = 0.75$, find the value of $a$.
(b) Find $E(x)$.

**Question 5.7**
Suppose the length of time spent in finishing a project is uniformly distributed between 6 and 15 days, what is the probability that a project could be finished within 12 days?

**Question 5.8**
The time between arrivals of customers at a supermarket follows an exponential probability distribution with a mean of 12 seconds.
(a) What is the probability that the arrival time between customers is 12 seconds or less?
(b) What is the probability of 30 or more seconds between customers’ arrivals?

**Question 5.9**
The time (in minutes) between patients’ arrivals at a clinic has the following exponential probability distribution.
\[
f(x) = 0.5e^{-0.5x} \quad \text{for} \quad x \geq 0
\]
(a) What is the probability of having 1 minute or less between customers’ arrival?
(b) What is the probability of having 5 or more minutes without a customer’s arrival?
Answers to EXERCISE 5

5.1  a) 0.6640 b) 0.1903 c) 0.1091
5.2  a) -0.80 b) 1.66 c) 0.26 d) 2.56 e) -0.50
5.3  a) 0.1112 b) 0.1112 c) 40.5
5.4  1.9138
5.5  0.50
5.6  a) \(a = 20\) b) \(E(x) = 40\)
5.7  0.6667
5.8  a) 0.6321 b) 0.0821
5.9  a) 0.3935 b) 0.3935
Chapter 6  Regression Analysis

Regression Analysis

It is used to explain the impact of changes in an independent variable (X) on the dependent variable (Y). It predicts the value of a dependent variable based on the value of at least one independent variable. That is to find the regression equation:

\[ \hat{y}_i = b_0 + b_1 x_i \]

**Dependent variable Y:** The variable we wish to predict or explain.

**Independent variable X:** The variable used to explain the dependent variable.

The estimated simple linear regression equation (i.e. \( \hat{y}_i = b_0 + b_1 x_i \)) provides an estimate of the simple linear regression equation.

**Interpretation of the regression parameters \((b_0 \text{ and } b_1)\):**

\( b_0 \) is the estimated average value of \( y \) when the value of \( x \) is 0.

\( b_1 \) measures the estimated change in the average value of \( y \) as a result of a 1-unit change in \( x \).

**Coefficient of Correlation \((r)\)**

Coefficient of Correlation, denoted by \( r \), measures the relative strength of the linear relationship between two variables. The value of \( r \) ranges between \(-1.0 \) and \(+1.0\). Values closer to \(-1\) imply stronger positive linear association between the two variables. Values closer to \(+1\) imply stronger positive linear association. Values closer to 0 imply weaker linear association. However, strong correlation does not necessarily imply a causation effect.

**Coefficient of Determination \((r^2)\)**

The coefficient of determination, denoted by \( r^2 \), is the “portion of the total variation in the dependent variable \((y)\) that is explained by the variation in the independent variable \((x)\)”. Unlike the coefficient of correlation, the value of \( r^2 \) ranges between 0 and 1.
The coefficient of determination is one of the most important values we have to pay special attention. Usually, researchers will look for an $r^2$ close to 1, which indicates a strong relationship between the dependent and independent variables.

**EXERCISE 6**

**Question 6.1**
Estimate the regression line of $Y$ on $X$ and the associated coefficient of determination for the following sets of data:

<table>
<thead>
<tr>
<th>X</th>
<th>3</th>
<th>6</th>
<th>2</th>
<th>8</th>
<th>4</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>9</td>
<td>11</td>
<td>3</td>
<td>16</td>
<td>10</td>
<td>17</td>
</tr>
</tbody>
</table>

**Question 6.2**
Estimate the regression line of $Y$ on $X$ and the associated coefficient of determination for the following sets of data:

<table>
<thead>
<tr>
<th>X</th>
<th>2</th>
<th>7</th>
<th>13</th>
<th>18</th>
<th>26</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>7</td>
<td>10</td>
<td>9</td>
<td>24</td>
<td>22</td>
<td>28</td>
</tr>
</tbody>
</table>

Interpret the coefficient of determination and the corresponding coefficient of correlation.

**Question 6.3**
Suppose an engineer’s salary level (in $1,000) is related to the corresponding years of experience as shown in the following table:

<table>
<thead>
<tr>
<th>X</th>
<th>1</th>
<th>3</th>
<th>4</th>
<th>4</th>
<th>6</th>
<th>8</th>
<th>10</th>
<th>10</th>
<th>11</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>40</td>
<td>49</td>
<td>46</td>
<td>51</td>
<td>52</td>
<td>56</td>
<td>60</td>
<td>62</td>
<td>59</td>
<td>68</td>
</tr>
</tbody>
</table>

Develop an estimated regression equation that can be used to predict the salary of an engineer given the years of experience. Use the estimated regression equation to predict the salary of an engineer with 9 years of experience.
**Question 6.4**

The following data show the quality rating and the price of a certain product:

<table>
<thead>
<tr>
<th>Rating</th>
<th>4</th>
<th>4</th>
<th>4</th>
<th>3</th>
<th>2.5</th>
<th>4</th>
<th>3</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Price ($)</td>
<td>531</td>
<td>466</td>
<td>630</td>
<td>265</td>
<td>200</td>
<td>665</td>
<td>402</td>
<td>194</td>
<td>349</td>
</tr>
</tbody>
</table>

Use the least squares method to develop the estimated regression equation. Provide an interpretation for the slope of the estimated regression. Predict the price for a certain product with a quality rating of 3.5.

**Question 6.5**

The following data show the late arrival rate and late departure rates in a certain airport:

<table>
<thead>
<tr>
<th>Arrival (%)</th>
<th>24</th>
<th>21</th>
<th>30</th>
<th>20</th>
<th>16</th>
<th>23</th>
<th>18</th>
<th>20</th>
<th>18</th>
</tr>
</thead>
<tbody>
<tr>
<td>Departures (%)</td>
<td>22</td>
<td>22</td>
<td>29</td>
<td>19</td>
<td>16</td>
<td>23</td>
<td>19</td>
<td>16</td>
<td>18</td>
</tr>
</tbody>
</table>

Use the least squares method to develop the estimated regression equation. Provide an interpretation for the slope of the estimated regression. Suppose the percentage of late arrivals was 22%, what is an estimate of the percentage of late departures?

**Question 6.6**

The following data show the production volumes and total cost data for a manufacturing operation:

<table>
<thead>
<tr>
<th>Volume</th>
<th>400</th>
<th>450</th>
<th>550</th>
<th>600</th>
<th>700</th>
<th>750</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cost</td>
<td>4000</td>
<td>5000</td>
<td>5400</td>
<td>5900</td>
<td>6400</td>
<td>7000</td>
</tr>
</tbody>
</table>

Use the data to develop the estimated regression equation that could be used to predict the total cost for a given production volume. What is the cost per unit produced? Compute the coefficient of determination. What percentage of the variation in total cost can be explained by production volume?
Answers to EXERCISE 6

6.1 \( b_0 = 1, b_1 = 2, r^2 = 0.8615 \)
6.2 \( b_0 = 4.516, b_1 = 0.759, r^2 = 0.8296, r = 0.9108 \)
   Interpretation of \( r^2 \): We see that 82.96% of the variability in Y has been explained by the estimated regression equation.
6.3 \( b_0 = 40.152, b_1 = 2.021, r^2 = 0.9324, \text{ Salary } = 58,342 \)
6.4 \( b_0 = -286.35, b_1 = 212.85, r^2 = 0.8365, \text{ Predicted price } = 458.63 \)
   Interpretation of \( b_1 \): A one point increase in the quality rating will increase the price by approximately $212.85.
6.5 \( b_0 = 1.208, b_1 = 0.9112, r^2 = 0.8592, \text{ Predicted late departure } = 21.25\% \)
   Interpretation of \( b_1 \): A one percent increase in the percentage of late arrivals will increase the percentage of late arrivals by 0.9112 or slightly less than one percent.
6.6 \( b_0 = 1246.67, b_1 = 7.6, r^2 = 0.9587, \text{ Cost per unit produced } = 7.6 \)
   Interpretation of \( r^2 \): We see that 95.87% of the variability in Cost has been explained by the estimated regression equation.
Chapter 7  Examples of Computer Application

Example 7.1: Hypothesis Testing

To test whether the average test score of the student population is different from 140 out of 250 by using the following sample test scores obtained from 22 students.

<table>
<thead>
<tr>
<th>Test Score</th>
<th>138</th>
<th>119</th>
<th>106</th>
<th>135</th>
<th>180</th>
<th>108</th>
<th>128</th>
<th>160</th>
<th>143</th>
<th>175</th>
<th>170</th>
<th>205</th>
<th>195</th>
<th>185</th>
<th>182</th>
<th>150</th>
<th>175</th>
<th>190</th>
<th>180</th>
<th>195</th>
<th>230</th>
<th>235</th>
</tr>
</thead>
</table>

Perform the One Sample t-Test: 

1. Click through Analyze/Compare Means/One Sample T-Test…
2. Select the variable “test score” into the Test Variable box, and enter the Test Value which the average value to be tested (i.e. 140 in this example), and then click “Continue”
3. Click “OK” to perform the test and estimation.

Interpretation of Output:
The one sample t-test statistic is 3.594 and the p-value from this statistic is 0.002 and that is less than 0.05 (the level of significance usually used for the test). This p-value indicates that the average test score of the sampled population is statistically significantly different from 140. The 95% confidence interval estimate for the difference between the population mean test score and 140 is (11.27, 43.34).
Example 7.2: Regression Analysis

To examine whether there is a linear relationship between the polishing time and the product size by linear regression, a model checking exercise should be conducted to make sure that the model is valid. Finally, the developed regression model can be used to predict the polishing time (Time) by the product size (Diam).
(Data set: SPSS sample data file “polishing.sav”)

<table>
<thead>
<tr>
<th>Time</th>
<th>Diam</th>
<th>Time</th>
<th>Diam</th>
<th>Time</th>
<th>Diam</th>
<th>Time</th>
<th>Diam</th>
<th>Time</th>
<th>Diam</th>
</tr>
</thead>
<tbody>
<tr>
<td>47.65</td>
<td>10.7</td>
<td>16.41</td>
<td>7.4</td>
<td>29.48</td>
<td>12</td>
<td>86.42</td>
<td>13</td>
<td>20.83</td>
<td>7.5</td>
</tr>
<tr>
<td>63.13</td>
<td>14</td>
<td>12.02</td>
<td>5.4</td>
<td>15.61</td>
<td>5.5</td>
<td>39.71</td>
<td>13</td>
<td>20.59</td>
<td>9</td>
</tr>
<tr>
<td>58.76</td>
<td>9</td>
<td>49.48</td>
<td>15.4</td>
<td>13.25</td>
<td>6</td>
<td>26.52</td>
<td>11.7</td>
<td>33.7</td>
<td>14</td>
</tr>
<tr>
<td>34.88</td>
<td>8</td>
<td>48.74</td>
<td>12.4</td>
<td>45.78</td>
<td>12</td>
<td>33.89</td>
<td>12.3</td>
<td>32.9</td>
<td>12.4</td>
</tr>
<tr>
<td>55.53</td>
<td>10</td>
<td>23.21</td>
<td>6</td>
<td>26.53</td>
<td>5.5</td>
<td>64.3</td>
<td>19.5</td>
<td>27.76</td>
<td>8.8</td>
</tr>
<tr>
<td>43.14</td>
<td>10.5</td>
<td>28.64</td>
<td>9</td>
<td>37.11</td>
<td>14.2</td>
<td>22.55</td>
<td>15.2</td>
<td>30.2</td>
<td>8.5</td>
</tr>
<tr>
<td>54.86</td>
<td>16</td>
<td>44.95</td>
<td>9</td>
<td>45.12</td>
<td>11</td>
<td>31.86</td>
<td>10</td>
<td>20.85</td>
<td>6</td>
</tr>
<tr>
<td>44.14</td>
<td>15</td>
<td>23.77</td>
<td>12.4</td>
<td>26.09</td>
<td>16</td>
<td>53.18</td>
<td>11</td>
<td>26.25</td>
<td>11</td>
</tr>
<tr>
<td>17.46</td>
<td>6.5</td>
<td>20.21</td>
<td>7.5</td>
<td>68.63</td>
<td>13.5</td>
<td>74.48</td>
<td>17.8</td>
<td>21.87</td>
<td>11.1</td>
</tr>
<tr>
<td>21.04</td>
<td>5</td>
<td>32.62</td>
<td>14</td>
<td>33.71</td>
<td>11.1</td>
<td>34.16</td>
<td>11.5</td>
<td>23.88</td>
<td>14.5</td>
</tr>
<tr>
<td>109.38</td>
<td>25</td>
<td>17.84</td>
<td>7</td>
<td>44.45</td>
<td>9.8</td>
<td>31.46</td>
<td>12.7</td>
<td>16.66</td>
<td>5</td>
</tr>
<tr>
<td>17.67</td>
<td>10.4</td>
<td>22.82</td>
<td>9</td>
<td>23.74</td>
<td>10</td>
<td>21.34</td>
<td>8</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

I. View the Data with a Scatter Plot:

1. Click through Graphs>Scatter>Simple>Define…
2. Select “time” to the Y-axis box, and “diam” to the X-axis box.
3. Click “OK” to generate the scatter plot.
II: Perform Regression Analysis

1. Click through Analyze\Regression\Linear…
2. Select “time” to the Dependent box, and “diam” to the Independent box.
3. Click on the “SAVE” button at the bottom of the dialogue box.
4. Check the Unstandardized boxes under the groups labeled Predicted Values and Residuals.
5. Click “Continue” and then “OK” to perform the regression analysis.

The resultant scatter plot appears to be suitable for conducting linear regression analysis.

The above Coefficients table shows the coefficient of the regression line. It states that the expected polishing time is given by: \( \text{Time} = -1.955 + 3.457 \times \text{Diam} \).
The Model Summary table reports the strength of the relationship between the model and the dependent variable. “R”, the correlation coefficient (which is the $r$ mentioned above), is the linear correlation between the observed and model-predicted values of the dependent variable “time”. Its large value (0.700) indicates a strong positive relationship. R Square, the coefficient of determination (which is the $r^2$ mentioned above), is the squared value of the multiple correlation coefficient. It shows that about half of the variation in time is explained by the variation in Diam.
References
