Data Collection, Probability and Statistics

A Form 6 Summary, with Applications to Past Examinations

Brad Bachu and Arjun Isa Mohammed
Alumni of Presentation College, Chaguanas

Nicholas Sammy and Kerry Shastri Singh
Alumni of Presentation College, San Fernando

November 2016
Version 1

Disclaimer

The views and opinions expressed in this document are those of the authors and do not reflect the official position of any examination board or academic body. A portion of this document is simply meant to provide commentary on examination past papers. We do not own the past papers and claim no authorship thereof. They can be found in 'CAPE ® Applied Mathematics Past Papers, Macmillian Education, December 2015' and 'CAPE ® Applied Mathematics Past Papers, Macmillian Education, December 2016'. We do not take responsibility for any outcomes that might occur as a result of the information provided in this document. You use this information at your own risk.

Acknowledgments

We express our thanks to Dipchand Bahall, Hassina Ramasra and Kieron Scott Singh for their guidance while pursuing this subject matter as secondary school students. Furthermore, for providing assistance with this document, we thank Vikash Maraj, Arshard Hosein, Mahendra Singh, Kieran Ramlogan and Marc Bhopalsingh. Lastly, we recognize Deepak Lall, Leon Kalicharan and Raphael Ramlal for support and encouragement in this undertaking.

About the Authors

Brad Bachu is currently a senior at MIT, double majoring in Physics and Mathematics with a minor in Philosophy. He won the President's Gold Medal in 2012 and was the CXC top performer in Technical Studies. At MIT, he has conducted research in experimental and theoretical high energy particle physics, focusing on dark matter and the use of effective field theories at the LHC and future colliders. His research has taken him to CERN as well as the Cluster of Excellence PRISMA.

Arjun Isa Mohammed attained an open scholarship in Mathematics in 2016. He was the top performer in Unit one Applied Mathematics. Arjun has represented Trinidad twice at the International Mathematics Olympiad, Thailand (2015) and Hong Kong (2016). Arjun is currently pursuing Computer Engineering and Physics.

Nicholas Sammy was awarded the Dennis Irvine Award for the Most Outstanding Candidate Overall for CAPE 2010 from CXC. He took up his Open Scholarship at Imperial College London, graduating with a BSc in Chemistry, having gained research experience into nanoparticulate catalysts for methane synthesis and homogeneous catalysts for hydrogenation of biomass-derived compounds. Currently a researcher at UWI St Augustine, his research area is in the field of heterogeneous catalysts for selective oxidation of methane.

Kerry Shastri Singh won the President's Gold Medal in 2011. He was the CXC regional top performer in Mathematics and Natural Sciences, as well as the winner of the Dennis Irvine award for the Most Outstanding Candidate Overall. He also won Trinidad and Tobago's first ever Silver Medal at the International Mathematical Olympiad in Kazakhstan in 2010. Kerry graduated from MIT in 2016 with a major in Biological Engineering and a minor in Theater Arts. He is currently a Research Fellow in the Department of Genetics at Harvard Medical. In his spare time, Kerry enjoys playing FIFA, going to the movies and playing football.

Preface

There were many motivating factors which led to the authors creating this document. The main factor was that there was a lack of supporting literature for this subject matter, specifically targeted to Form Six Caribbean students. Furthermore, since a guide of this type has not yet been provided, we decided that by making this document freely available, it would have a greater capability to help any student who wants to succeed. Combining these factors with our passion for teaching and desire to help other students, we were able to work together to build a resource that would be easily accessible to all students.

This document is intended to be more than just a 'solutions' manual. Our goal was to provide a 'walk-through', where appropriate, for questions and present a thought process to solve the problems. The hope is that it can serve the needs of students who simply want a numeric check, the students who are still unsure on how to approach some topics, and the students who are open to learning more.

The the document was typeset using LaTeXunder the format of a book. Consequently, we recommend two ways to use this document. You can either print it and take advantage of the sectioning and layout, or use the PDF to take advantage of cross referencing of definitions and equations in the document. Whichever you decide, we hope that the structure will prove well suited to your needs.

Our desire is that readers of this document will gain a greater appreciation of the subject matter and/ or become motivated to advance their studies in a related field. If you feel like this could be improved in anyway, whether it be fixing errors, or expanding or improving ideas, feel free to reach out to us and let us know. We hope that you will enjoy using this document as much as we enjoyed creating it.

Contents

Part I

Introduction

Chapter 1

Why you should be interested in this topic and how to approach it

1.0.1 A High Probability You Will Encounter Probability

Data collection, probability and statistics are some of the most important tools in the arsenal of a student pursuing a career in science, technology, engineering or mathematics. You will probably encounter at least one of these topics again in a university course or in your career. For some of us, these topics can become basis for which we seek to pursue a deeper theoretical understanding. For others, it will become the machinery you use everyday to understand the world around you.

Many of the topics you explore in lower six have deeper philosophical implications and mathematical meaning you may realize. Take for example, the most fundamental thing you know about probability and try to think about it for sets of infinite size. Now consider the fact that the rational numbers are 'countably infinite' and the irrational numbers are 'uncountably infinite'. A philosopher might propose that if you throw a dart at a number line, it will never hit a rational number because the infinity of the irrational numbers is much larger than the infinity of the rational numbers. A deeper search in the theoretical basis of probability might lead one to stumble upon measure theory and the Lebesgue measure to make sense of the philosopher's statement.

Furthermore, with the advent of quantum mechanics, physicists discovered that we lived in a probabilistic world. To illustrate, the most accurate way to describe the position of microscopic objects, such as the electron, is with a sort of 'probability function'. Immediately, quantum mechanics imposed all the mathematical restrictions of probability theory to the real world. A physcists might even (rightly) go as far as trying to convince you that if you try to run through a wall, according to quantum mechanics, there is a non-zero probability that you will 'tunnel through' it, or even worse, if you try to run off a cliff, there is a non-zero probability that you will 'reflect', instead of fall off (do not try this at home, in fact, anywhere).

Probability also has major significance in the field of computational biology. Drug design, for example, would not be possible without a fundamental understanding of statistical thermodynamics. Macromolecules, like proteins, do not have fixed geometric conformations and are modeled as having multiple conformations with assigned probabilities. These probabilities help inform the entropy (think spatial chaos) of the system which may then be used to predict how a particular reaction or interaction is going to proceed. Equipped with this information, biologists can better model and understand the interactions between small drug molecules and the proteins or other biomolecules which they target. This is the sort of analysis which goes into the design of drugs such as acetaminophen (commonly known as Tylenol or Panadol) whose mode of action is the inhibition of the enzyme cyclooxygenase to alleviate pain and inflammation.

In sum, probability can be found anywhere you search for it. In fact, even while you search 'probability' in Google, it will use a Markov chain to determine what websites you want to see. So if you are intrigued by these ideas and concepts, then pursuing this topic in lower six is the right place to be. Having a strong foundation in the topics we explore will go a long way in helping you understand more advanced topics in the future.

1.0.2 How to Master the Material

To begin with, for whatever reason you find yourself pursuing this subject, it is important to have a positive attitude towards it and try your best. Although the term ‘Applied Mathematics’ might sound intimidating, you need not worry. You have already been exposed to many fundamental concepts such as sets, data collection, probability and statistics, in lower forms. As these techniques have proved beneficial to us, we present them in the hope that you can make use if it as you see fit:

1. **Be patient with the material**: Some people catch on to these concepts and get comfortable applying them quickly. For others, it takes time to convince themselves what they are told is true before they can proceed confidently. Whatever your approach, it is important to not associate your learning pace with how ‘smart’ you are.

2. **Ask Questions**: Engage your teachers on the aspects you do not understand. It is a teacher’s duty to assist students in understanding the material of the course and they are often willing to help. It is not sensible to sit in class and be lost/ unable to follow the material because you do not feel courageous enough to ask for clarifications. It becomes more difficult to ‘catch up’ as time goes on.

3. **Get a second opinion**: If you feel that the explanation you are presented with does not seem to be getting through to you, or you want more, then approach some other teacher, a friend or the internet.

4. **Try to teach it to a friend**: If your classmate is having trouble understanding a topic that you understand, explain it to him/ her. It reinforces your understanding of the topic and can help your classmate to finally get it in simpler terms than a teacher would use.

5. **Work with a friend**: Many of the problems you solve will have multiple approaches. It is important to not stick to one method as exploring multiple methods is the best way to double check your results.

6. **SHARE!**: Cooperation and sharing of information helps EVERYONE to succeed in their exams. Withholding information and help is selfish and wrong.

Part II

Important Definitions and Notes

Chapter 2

Module 1: Collecting and Describing Data

2.1 Sources of Data

Definition 2.1.1. **Discrete data** refers to a set of data points which can only take exact values.

Definition 2.1.2. **Continuous data** cannot take exact values but can be given only within a specified range or measured to a specified degree of accuracy.

Definition 2.1.3. **Quantitative data** is anything that can be expressed as a number that is countable. This type of data manifests through ordinal, internal or ratio scales and lend themselves to statistical manipulations.

Definition 2.1.4. **Qualitative data** refers to research on variables which cannot be numerically measured. This type of data is usually called explanatory data and provides in depth analyses of a problem.

Definition 2.1.5. The **population** refers to all the elements or individuals that meet the selection criteria for a group to be studied, and from which a **sample** is usually chosen from to be examined in detail.

Definition 2.1.6. A **parameter** is a numerical measure calculated from a population.

Definition 2.1.7. A **sample** is a subset of the population which is studied in detail so as to find numerical data, central tendencies or any other patterns. These statistics can usually be extended to the population but depend on the degree of representation of the sample to the population.

Definition 2.1.8. A **statistic** refers to a numerical measure calculated from the sample.

2.2 Data Collection

Definition 2.2.1. A **sampling frame** is an extensive list containing all members of the group under examination from which the sample is chosen.

Definition 2.2.2. A **simple random sample** is a subset of items of size n, selected from a larger set (the population) of size N, such that the probability of each item being chosen is the same.

Note 2.2.1. Simple random sampling has the advantage that is that it is relatively cheap. On the other hand, it has the disadvantage that members within subgroups of population may not have a proportional representation in the sample.

Definition 2.2.3.

A **stratified random sample** is a subset of items of size n, selected from a larger set (the population) of size N, such that

(i) The population is divided up in to M mutually exclusive, and collectively exhaustive strata of size m_i, such that $m_1 + ... + m_i + ... + m_M = N$.

(ii) A simple random sample is done within each strata to obtain a subset of size s_i such that $s_1 + ... + s_i + ... + s_n = n$

(iii) The proportions of each s_i in the sample is the same as the proportions of each m_i in the population. i.e $\frac{n_i}{n} = \frac{m_i}{N}$

Note 2.2.2. Stratified sampling has the advantage that it represents the population proportionately according to the strata, and as a result, can sometimes lead to a smaller sample. On the other hand, it has the disadvantage that it may require more time and effort than a simple random sample, and that it may be difficult to establish strata.

Definition 2.2.4. In **systematic sampling** the population is listed in some order (e.g. alphabetically) and then every k^{th} member is chosen from the list starting from some random point.

$$k = \frac{N}{n}, \tag{2.2.1}$$

where N is the population size and n is the sample size.

Note 2.2.3. Systematic sampling offers the advantage that it is quick to execute and easy to check for errors. On the other hand, it has the disadvantage that there may be a periodic cycle within the frame itself.

Definition 2.2.5. In a **cluster sample**, the population is divided into non-overlapping (non-intersecting or mutually exclusive) clusters or groups, and a random sample of these groups is chosen. Cluster sampling may be further classified as either one-stage or two-stage sampling. One-stage sampling describes the scenario in which all members of the randomly selected clusters are chosen. On the other hand, if all of the members are not chosen, and another random sample is needed to select the members of a cluster, this is referred to as two-stage sampling.

Note 2.2.4. Cluster sampling offers the advantage that there is no need to have the complete sample frame and it is less costly than random sampling. On the other hand, the disadvantage is that it is non-random.

Note 2.2.5. *Cluster versus Strata*: Clusters are similar to other clusters, and the population within a cluster should be as heterogeneous as possible. Strata are different from other strata, and the population within a strata should be as homogeneous as possible. In stratified random sampling, a random sample is drawn from each of the strata, whereas in cluster sampling, the cluster is treated as the sampling unit and only the selected clusters are studied.

Definition 2.2.6. In **quota** sampling, the population is segmented into mutually exclusive subgroups. Then, judgment is used to select the subjects or units from each segment based on specified proportions. The selection of the sample is non-random and unreliable.

Note 2.2.6. Quota sampling offers the advantage that it is quick to use, has low complications, any member of the sample can be replaced by another member with the same characteristics and it is practical if no sample frame exists. On the other hand, it has the disadvantage that it is non-random, there is possibility of bias in the selection process, and it may exclude a substantial part of the population.

2.3 Data Analysis

Given a collection of data points of size N, let x_i be the value of the ith element in the collection.

Definition 2.3.1. The **mean** $\bar{x}$ of the population,

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{2.3.1}$$

For *grouped data*, the mean is If x_j occurs with frequency f_j, then

$$\bar{x} = \frac{\sum x_j f_j}{\sum f_j} \tag{2.3.2}$$

Definition 2.3.2. The **variance** of the *entire population*, σ^2,

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(\bar{x} - x_i)^2 \tag{2.3.3}$$

$$= \frac{\sum x^2}{N} - \bar{x}^2 \tag{2.3.4}$$

See Appendix [15.4.2]

If x_j occurs with frequency f_j, then

$$\sigma^2 = \frac{\sum f_j x_j^2}{\sum f_j} - (\bar{x})^2 \tag{2.3.5}$$

Definition 2.3.3. The $X\%$ **trimmed mean** is a calculation of the mean after discarding $X\%$ of the upper and lower end of ordered data points.

Definition 2.3.4. The $a^{\textbf{th}}$ **percentile** is the value $a\%$ of the way through the distribution. Some famous percentiles are called the **quartiles**.

$$a^{\text{th}} \text{ percentile} = \frac{a}{100}(n+1)^{\text{th}} \text{ term} \tag{2.3.6}$$

Sometimes, this takes a non-integer value. Suppose it took a value of $b.25$, $b.50$ or $b.75$ and c was the next integer value. Then we do the following:

$$b.25^{\text{th}}\text{term} = 0.25 \times (c^{\text{th}}\text{term} - b^{\text{th}}\text{term}) \tag{2.3.7}$$

$$b.50^{\text{th}}\text{term} = 0.50 \times (c^{\text{th}}\text{term} - b^{\text{th}}\text{term}) \tag{2.3.8}$$

$$b.75^{\text{th}}\text{term} = 0.75 \times (c^{\text{th}}\text{term} - b^{\text{th}}\text{term}) \tag{2.3.9}$$

For a more general formalism see Appendix [15.4.3]

Definition 2.3.5. In descriptive statistics, **quartiles** divide the data into four equal groups.

(i) Q_1: The **lower quartile** or 25th percentile.

(ii) Q_2: The **median** or 50th percentile. When n is odd, the median is the middle value. When n is even, there are two middle values a and b and the median is $\frac{a+b}{2}$

(iii) Q_3: The **upper quartile** or 75th percentile.

(iv) IQR: The **inter quartile** range. $IQR = Q_3 - Q_1$

Definition 2.3.6. The asymmetry of the distribtion can be measured by its **skewness**. There are three general skew states:

1) **Normal Distribution**

   (i) Mode = Median

   (ii) Median = Mean

   (iii) $(Q_3 - Q_2) = (Q_2 - Q_1)$

2) **Positive Skew**

   (i) Mean > Median > Mode

   (ii) $(Q_3 - Q_2) > (Q_2 - Q_1)$

3) **Negative Skew**

   (i) Mean < Median < Mode

   (ii) $(Q_3 - Q_2) < (Q_2 - Q_1)$

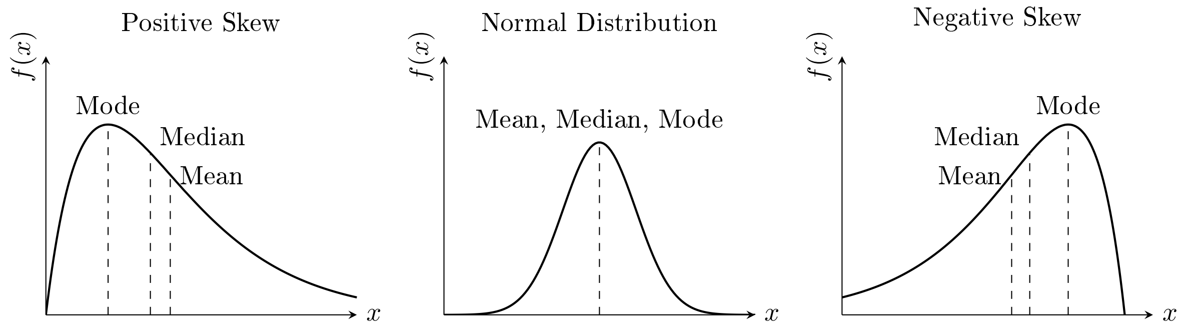


Figure 2.1: The general shapes and relative positions of the mean, mode and median in positively skewed, normal, and negatively skewed distributions.

Chapter 3

Module 2: Probability and Random Variables

3.1 Probability Theory

Definition 3.1.1. The **sample space** refers to the set of all possible outcomes of an experiment and is usually denoted by S.

Definition 3.1.2. An **event** is any subset of the sample space i.e. a set consisting of possible outcomes of an experiment.

Law 3.1.1. ***DeMorgan's Law*** *provides a useful relationship between unions, intersections and complements:*

$$(A \cup B)^c = A^c \cap B^c \tag{3.1.1}$$
$$(A \cap B)^c = A^c \cup B^c \tag{3.1.2}$$

Axiom. *Consider an experiment whose sample space is S. The* ***probability of the event*** *E, denoted as $P(E)$, is a number that satisfies the following three axioms*

1 $0 \leq P(E) \leq 1$

2 $P(S) = 1$

3 *For any sequence of mutually exclusive events $E_1, E_2, ..., E_n$.*
$P\left(\cup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} P(E_i)$
Eg. $P(E_1 \cup E_2) = P(E_1) + P(E_2)$

Furthermore, we define,

$$P(\textit{event E occurs}) = \frac{\textit{no. of times A can occur}}{\textit{total number of outcomes}} \tag{3.1.3}$$

Definition 3.1.3. The probability of the **complement** of event A, denoted as $P(A^c)$ or $P(\bar{A})$ or $P(A')$, is defined as

$$P(\bar{A}) = 1 - P(A) \tag{3.1.4}$$

Proposition 3.1.2. *Some useful propositions can be easily derived using the above and drawing your own diagrams*

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{3.1.5}$$
$$P(A^c \cap B) = P(B) - P(A \cap B) \tag{3.1.6}$$
$$P(A^c \cap B^c) = P(A \cup B)^c \tag{3.1.7}$$
$$P(A^c \cup B^c) = P(A \cap B)^c \tag{3.1.8}$$

Definition 3.1.4. Two events A and B are said to be **mutually exclusive** if and only if

$$P(A \cap B) \quad = \quad 0 \tag{3.1.9}$$

Applying Eq [3.1.5], this implies

$$P(A \cup B) \quad = \quad P(A) + P(B) \tag{3.1.10}$$

Definition 3.1.5. Two events A and B are said to be **independent** if and only if

$$P(A \cap B) \quad = \quad P(A) \times P(B) \tag{3.1.11}$$

Definition 3.1.6. For events A and B, the **conditional probability** of A given B has occurred, denoted by $P(A|B)$, is defined by

$$P(A|B) \quad = \quad \frac{P(A \cap B)}{P(B)} \tag{3.1.12}$$

Note 3.1.1. If A and B are mutually exclusive,

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} && (3.1.13) \\ &= \frac{0}{P(B)} = 0 && (3.1.14) \end{aligned}$$

Note 3.1.2. If A and B are independent,

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} && (3.1.15) \\ &= \frac{P(A) \times P(B)}{P(B)} && (3.1.16) \\ &= P(A) && (3.1.17) \end{aligned}$$

In words, if B does not influence A, then the probability that A occurs given that B has occurred is simply the probability of A.

Note 3.1.3. Two events can be both independent and mutually exclusive if at least one of the events has a probability of zero.

$$\begin{aligned} P(A \cap B) &= P(A) \times P(B) \text{ by independence} && (3.1.18) \\ P(A \cap B) &= 0 \text{ by mutually exclusive} && (3.1.19) \\ \therefore P(A) \times P(B) &= 0 && (3.1.20) \\ \Rightarrow P(A) = 0 \quad &\text{or} \quad P(B) = 0 && (3.1.21) \end{aligned}$$

3.2 Random Variables

Definition 3.2.1. A **discrete random variable**, X, can take on at most a countable number of possible values. We define its *probability mass* function, $P(x)$ by

$$P(x) \quad = \quad P(X = x) \tag{3.2.1}$$

Proposition 3.2.1. *If X is a discrete random variable with probability mass function $P(x)$, then we know from Axiom [1] that*

$$0 \leq P(X = x) \leq 1 \, \forall x \tag{3.2.2}$$

and from Axiom [2] and Axiom [3] that

$$\sum_{\forall x} P(X = x) = 1 \tag{3.2.3}$$

Definition 3.2.2. If X is a discrete random variable, the **expectation value**, **mean** or **first moment of** X, $E[X]$, is defined by

$$E[X] = \sum_{\forall i} x_i P(X = x_i) \tag{3.2.4}$$

Definition 3.2.3. If X is a discrete random variable, the **second moment of** X is defined by

$$E[X^2] = \sum_{\forall i} x_i^2 P(X = x_i) \tag{3.2.5}$$

Definition 3.2.4. If X is a random variable with mean μ, then the **variance**, $Var[X]$, is defined by

$$\begin{aligned} Var[X] &= E[(X-\mu)^2] && (3.2.6)\\ &= E[X^2] - (E[X])^2 = E[X^2] - \mu^2 && (3.2.7) \end{aligned}$$

Definition 3.2.5. if X is a discrete random variable, the **cumulative distribution function** $F(x)$ is defined as

$$\begin{aligned} F(x) &= P(X \leq x) && (3.2.8)\\ &= \sum_{-\infty}^{x} P(X \leq x) && (3.2.9) \end{aligned}$$

Definition 3.2.6. We say that X is a **continuous random variable**, if there exists a non-negative function f, defined for all real $x \in (-\infty, \infty)$, having the property that, for any set of real numbers B,

$$P(X \in B) = \int_B f(x).dx \tag{3.2.10}$$

The function f is called the *probability density function* of the random variable X.

Proposition 3.2.2. *Since X must assume some value, f must satisfy,*

$$1 = P(X \in (-\infty, \infty)) = \int_{-\infty}^{\infty} f(x).dx \tag{3.2.11}$$

Note 3.2.1. If $B = [a, b]$, then

$$\begin{aligned} P(X \in B) &= P(a \leq X \leq b) && (3.2.12)\\ &= \int_a^b f(x).dx && (3.2.13) \end{aligned}$$

Proposition 3.2.3. *For any continuous random variable X and real number a,*

$$P(X = a) = 0 \tag{3.2.14}$$

Note 3.2.2. This implies

$$P(X \leq a) = P(X < a) \tag{3.2.15}$$

Definition 3.2.7. If X is a continuous random variable, the **expectation value** of X is given by,

$$E[X] = \int_{-\infty}^{\infty} x f(x).dx \tag{3.2.16}$$

3.3 Binomial Distribution

Definition 3.3.1. A discrete random variable X is said to follow a **binomial distribution** with parameters n and p, $X \sim \text{Bin}(n, p)$, if its probability mass function, P, is given by

$$P(X = x) \quad = \quad \binom{n}{x} p^x (1-p)^{n-x} \quad x \in (0, 1, 2, ..., n) \tag{3.3.1}$$

Note 3.3.1. The *expected value (mean)* and *variance* of X are given by:

$$E[X] \quad = \quad np \tag{3.3.2}$$

$$Var[X] \quad = \quad np(1-p) \tag{3.3.3}$$

$$= \quad npq \text{ where } q = (1-p) \tag{3.3.4}$$

Note 3.3.2. There are four conditions that describe a binomial distribution

(i) The experiment consist of a fixed number of trialS n.

(ii) The trials are independent.

(iii) Each trail can be classified as a success or failure.

(iv) The probability of success, p, is constant

3.4 Normal Distribution

Definition 3.4.1. A continuous random variable X is said to follow a **normal distribution** with parameters μ and σ^2, $X \sim \text{N}(\mu, \sigma^2)$, if its probability density function, f, is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} \quad -\infty < x < \infty \tag{3.4.1}$$

The parameters μ and σ^2 represent the expected value (mean) and variance respectively.

Note 3.4.1. Normal random variables have important properties

(i) If X is normally distributed with parameters μ and σ^2, then $Y = aX + b$ is normally distributed with parameters $a\mu + b$ and $a^2\sigma^2$ respectively.

(ii) The previous property implies that if X is normally distributed with parameters μ and σ^2, then $Z = (X - \mu)\sigma$ is normally distributed with parameters 0 and 1. We call Z a *standard* or *unit* normal random variable.

Theorem 3.4.1. *The **DeMoivre-Laplace limit theorem** or the **Normal Approximation to the Binomial Distribution**. If $X \sim Bin(n, p)$, for large n, we can make the approximation that $X \sim N(np, np(1-p))$.*

Note 3.4.2. This approximation is good for values of n satisfying $np > 5$ and $np(1-p) = npq > 5$.

Note 3.4.3. To use this approximation, we must note that because the binomial is a discrete integer-valued random variable and the normal is a continuous random variable, we must write

$$P(X = i) \quad = \quad P\left(i - \frac{1}{2} < X < i + \frac{1}{2}\right) \tag{3.4.2}$$

$$P(a \leq X < b) \quad = \quad P\left(a - \frac{1}{2} < X < b - \frac{1}{2}\right) \tag{3.4.3}$$

$$P(a < X \leq b) \quad = \quad P\left(a + \frac{1}{2} < X < b + \frac{1}{2}\right) \tag{3.4.4}$$

$$P(a \leq X \leq b) \quad = \quad P\left(a - \frac{1}{2} < X < b + \frac{1}{2}\right) \tag{3.4.5}$$

This is called the **continuity correction**.

Chapter 4

Module 3 : Analyzing and Interpreting data

4.1 Sampling Distribution and Estimations

Theorem 4.1.1. *The **Central Limit Theorem** states that if $X_1, X_2, ..., X_n$ is a random sample, that is, a sequence of independent and identically distributed random variables, taken from any population with mean μ and variance σ^2, then as n grows ($n \to \infty$), $\bar{X}$ converges in distribution to a normal distribution with mean μ and variance $\frac{\sigma^2}{n}$, where μ and σ^2 are parameters of the sample distribution . We usually say for $n \geq 30$, $\bar{X}$ can be approximately normally distributed with mean μ and variance $\frac{\sigma^2}{n}$ i.e. $\bar{X} \sim \mathbf{N}(\mu, \frac{\sigma^2}{n})$. More on this statement is given in Appendix [15.6.1]*

Note 4.1.1. If the population is normally distributed, then the Central Limit Theorem holds for any sample size when $n \geq 2$.

Definition 4.1.1. If $E[u] = \theta$, then u is an **unbiased estimator** for θ. Otherwise u is a **biased estimator**.

Note 4.1.2. Whenever we collect a sample of data, we can use this sample to compute statistics. However, the goal of sampling is not only to determine the sample statistics but the population parameters. Because we can not always experiment on the entire population, we use unbiased estimators. This allows us to use our data in the sample to estimate population parameters.

Definition 4.1.2. An unbiased estimate $\hat{\mu}$ for the population mean μ with the sample of size n with mean $\bar{x}$ is given by

$$\hat{\mu} \quad = \quad \bar{x} = \frac{\sum x}{n} = \frac{\sum(x-a)}{n} + a, \quad \forall a \tag{4.1.1}$$

See Appendix [15.6.2].

Definition 4.1.3. An unbiased estimate $\hat{\sigma^2}$ for the population variance σ^2 from a sample of size n with variance s^2 is given by

$$\begin{aligned} \hat{\sigma^2} &= \frac{n}{n-1}s^2 = \frac{1}{n-1}\left[\sum x^2 - \frac{(\sum x)^2}{n}\right] && (4.1.2) \\ &= \frac{1}{n-1}\left[\sum(x-a)^2 - \frac{(\sum(x-a))^2}{n}\right] = \frac{\sum(x-\bar{x})^2}{n-1} && (4.1.3) \end{aligned}$$

Definition 4.1.4. An unbiased estimate $\hat{p}$ for the population proportion p from a sample of size n with sample proportion p is given by:

$$\hat{p} \quad = \quad p_s = \frac{x}{n} \tag{4.1.4}$$

Definition 4.1.5. A $(1-\alpha).100\%$ confidence interval for μ means that we will find with probability $\frac{1-\alpha}{100}$ a confidence interval in which the actual value of the parameter μ will lie within. We can compute a $(1-\alpha).100\%$ confidence interval for μ under three cases

(i) σ^2 known.

$$\bar{x} \pm z_{\alpha/2} \times \sqrt{\frac{\sigma^2}{n}} \tag{4.1.5}$$

(ii) σ^2 unknown, $n \geq 30$

$$\bar{x} \pm z_{\alpha/2} \times \sqrt{\frac{\hat{\sigma^2}}{n}} \tag{4.1.6}$$

(iii) σ^2 unknown, $n < 30$

$$\bar{x} \pm t_{\alpha/2}^{(n-1)} \times \sqrt{\frac{\hat{\sigma}^2}{n}} \tag{4.1.7}$$

A $(1-\alpha)\%$ confidence interval for the population proportion p as

$$p_s \pm z_{\alpha/2}\sqrt{\frac{p_s q_s}{n}} \tag{4.1.8}$$

where $p_s = \frac{x}{n}$ and $q_s = 1 - p_s$.

4.2 Hypothesis Testing

For intuition on how the algorithm is constructed see Appendix [15.6.4].

4.2.1 Algorithm

One can conduct a test for the population mean μ by following simple procedure:

Step 1. State the null hypothesis

$$H_0 : \mu = \mu_0 \tag{4.2.1}$$

Step 2. State the alternative hypothesis

$$\begin{array}{lll} H_1: & \mu > \mu_0 & \text{(one-tailed test, upper tial)} \\ & \mu < \mu_0 & \text{(one-tailed test, lower tial)} \\ & \mu \neq \mu_0 & \text{(two-tailed test, upper and lower tail)} \end{array}$$

Step 3. Look at conditions provided by problem. Assuming H_0 is true, calculate the test statistic.

Step 4. State the type of test (one or two tailed) and consider the significance level of the test. Determine the critical region for hypothesis test.

Step 5. Consider test statistic and formulate conclusion for the test.

4.2.2 Step 3: Calculating the test statistic

The procedure for calculating the test statistic varies based on the distribution, whether σ^2 is known or unknown, and the size of the sample n.

(i) Normally distributed, σ^2 unknown, $\forall n$,

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \tag{4.2.2}$$

(ii) σ^2 known, $n \geq 30$ (uses Central Limit Theorem [4.1.1]), non-normal population,

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \tag{4.2.3}$$

(iii) σ^2 unknown, $n \geq 30$, any distribution (uses Central Limit Theorem [4.1.1]) for non-normal distributions),

$$z = \frac{\bar{x} - \mu_0}{\hat{\sigma}/\sqrt{n}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n-1}} \tag{4.2.4}$$

(iv) $n < 30$, σ^2 unknown, normally distributed,

$$t = \frac{\bar{x} - \mu_0}{\hat{\sigma}/\sqrt{n}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n-1}} \tag{4.2.5}$$

4.2.3 Step 4: Finding the Critical Region

If the test statistic lies within the critical region, *reject the null hypothesis in favor of the alternative hypothesis*. Else, we *fail to reject the null hypothesis*. Notice that we did not say ‘accept the null hypothesis’.

The critical region depends on what our alternative hypothesis is, the size of the distribution and whether or not the population variance is known.

In general,

(i) If σ^2 is known or $n \geq 30$ use z-distribution and z_{calc}

(ii) If σ^2 is unknown and $n < 30$ use $T \sim t(n-1)$ distribution and t_{calc}

(A) If $\mu > \mu_0$

α% level of significance.

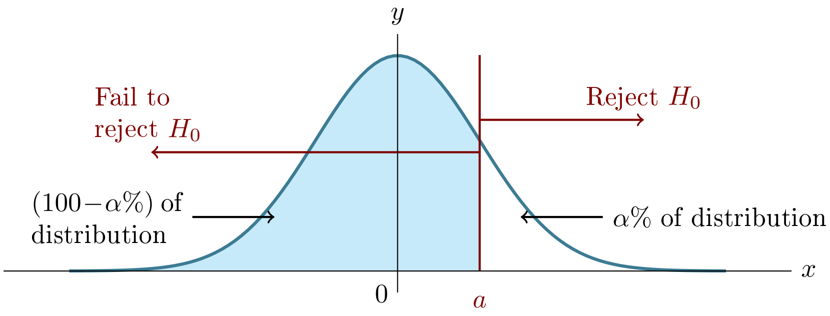


Figure 4.1: We reject H_0 if $z_{\text{calc}} > a$ or $t_{\text{calc}} > a$

(B) If $\mu < \mu_0$,

α% level of significance.

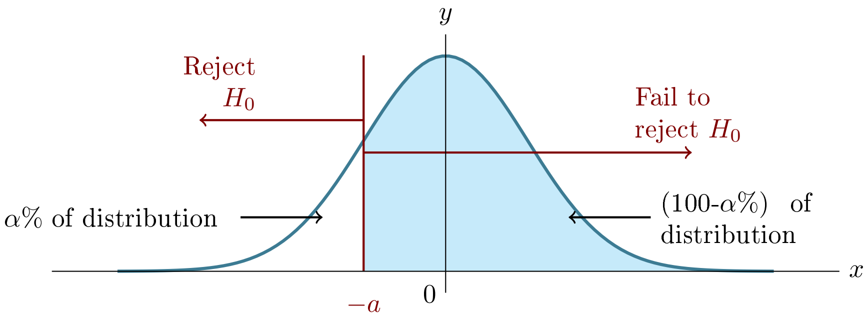


Figure 4.2: Reject H_0 if $z_{\text{calc}} < -a$ or $t_{\text{calc}} < -a$

(C) $\mu \neq \mu_0$

α% level of significance.

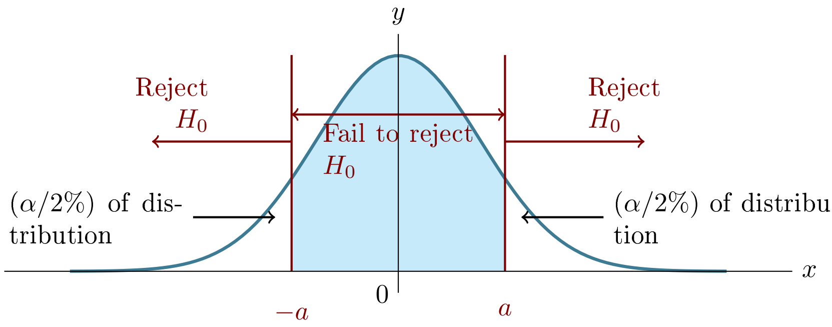


Figure 4.3: Reject H_0 if $z_{\text{calc}} > a$ or $z_{\text{calc}} < -a$; if $t_{\text{calc}} > a$ or $t_{\text{calc}} < -a$

4.3 χ^2 test

This test is used to determine if there is a significant association between two categorical variables from a single population. One can conduct a χ^2 test following the simple procedure:

1) Define the null hypothesis, H_0, and alternative hypothesis, H_1:

$$H_0 \quad : \quad \text{The attributes are independent.} \tag{4.3.1}$$

$$H_1 \quad : \quad \text{The attributes depend on each other.} \tag{4.3.2}$$

2) Pick a level of significance α

3) Identify critical region: Reject H_0 if $\chi^2_{\text{calc}} > \chi^2_{\alpha}(\nu)$

4) $\chi^2_{\text{calc}} = \sum \frac{(O-E)^2}{E}$ where O is the observed data, and E is the expected frequency. We can compute the expected frequency using $E_{ij} = \frac{n_i \times n_j}{N}$, where E_{ij} represents the expected frequency of the value in the i^{th} row and j^{th} column, n_i represents the sum of values in the i^{th} row, n_j represents the sum of values in the j^{th} column, and N represents the total sample size.

5) Decision

Note 4.3.1. χ^2_{calc} is not valid if $E < 5$. When this occurs we combine classes to make all values of $E > 5$.

4.4 Correlation and Linear Regression

Definition 4.4.1. The **Pearson product-moment correlation coefficient**, r, measures the degree of the linear correlation, or linear relationship, between two variables, X and Y.

Given a dataset containing n values of X and Y respectively, we compute r as

$$\begin{aligned} r &= \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}} && (4.4.1)\\ &= \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}} && (4.4.2)\end{aligned}$$

Note 4.4.1. r can only take a value such that $-1 \leq r \leq 1$, where

(i) $r = -1$ means total negative linear correlation.

(ii) $r = 0$ mean no linear correlation.

(iii) $r = 1$ means total positive linear correlation

Note 4.4.2. We can make stronger statements on r as follows

(i) $0 - 0.5$: low positive linear correlation

(ii) $0.5 - 0.8$: relatively high positive linear correlation

(iii) > 0.8: high positive linear correlation

Replace 'positive' with 'negative' if r is negative.

Definition 4.4.2. A line of **linear regression** of Y on X has the form

$$\begin{aligned} y &= a + bx && (4.4.3)\\ \text{where } b &= \frac{n\sum xy - \sum x . \sum y}{n\sum x^2 - (\sum x)^2} && (4.4.4)\\ \text{and } a &= \bar{y} - b\bar{x} && (4.4.5)\\ \text{where } \bar{y} &= \frac{\sum y}{n} \text{ and } \bar{x} = \frac{\sum x}{n} && (4.4.6)\end{aligned}$$

Note 4.4.3. $(\bar{x}, \bar{y})$ is on the line.

Part III

Solutions

Chapter 5

2005

5.1 Module 1: Collecting and Describing Data

1. **(a)** We define the terms ‘a simple random’ and ‘a stratified random’ sample as follows:

(i) From Def [2.2.2], a simple random sample is a subset of items of size n, selected from a larger set (the population) of size N, such that all possible combinations of n are equally likely to occur.

(ii) From Def [2.2.3], a stratified random sample is a subset of individuals of size n, the sample, selected from a larger set of size N, the population, such that

i. The population is divided up into M mutually exclusive, and collectively exhaustive strata of size m_i, such that $m_1 + ... + m_i + ... + m_M = N$.

ii. A simple random sample is done within each strata to obtain a subset of size s_i such that $s_1 + ... + s_i + ... + s_n = n$

iii. The proportions of each s_i in the sample is the same as the proportions of each m_i in the population. i.e $\frac{n_i}{n} = \frac{m_i}{N}$

(b) We omit this question.

(c) A specified sample is required.

(i) We can assign each of the 600 persons a number between 1 and 600. Randomly select 30 times from these 600 numbers, for example, by writing them on a piece of paper and blindly drawing from a container. The people whose number corresponds to one of the 30 numbers are the sample.

(ii) There are disadvantages with other sample methods.

a The people who all leave within the same time period; people sitting closer to the door; and a group of friends who came together are all likely to be sampled in this method. These can all be considered specific stratum. Hence a sample of this sort will be biased towards people in this stratum.

b It may not proportionally represent all the levels or strata.

c It may be difficult to establish mutually exclusive strata and execute the procedure within the specified time period.

2. We are given a frequency table.

(a) We can determine the following statistics.

(i) Recall that the mode is the value with the highest frequency. Hence

$$\text{Mode } = 5 \tag{5.1.1}$$

(ii) Recall from Def [2.3.5] that the median is given by:

$$Q_2 = \frac{n+1}{2}^{\text{th}} \text{ term} \tag{5.1.2}$$

$$= \frac{30+1}{2}^{\text{th}} \text{ term} \tag{5.1.3}$$

$$= 15.5^{\text{th}} \text{ term} \tag{5.1.4}$$

Because the data is grouped, we need to pay attention to the frequencies to see where this term lies. Starting from the lowest mark, we can add the frequencies and see that the 15.5th term at the border the 3 and 4 mark. Hence

$$Q_2 = 15^{\text{th}} \text{ term} + 0.5 \times (16^{\text{th}} \text{ term} - 15^{\text{th}} \text{ term}) \tag{5.1.5}$$

$$= 3 + 0.5(16 - 15) \tag{5.1.6}$$

$$= 3.5 \tag{5.1.7}$$

(iii) Recall from Eq [2.3.2], that when we have grouped data, we can calculate the mean as,

$$\bar{x} = \frac{\sum x_i f_i}{\sum f_i} \tag{5.1.8}$$

$$= \frac{0(2) + 1(4) + 2(3) + 3(6) + 4(7) + 5(8)}{2 + 4 + 3 + 6 + 7 + 8} \tag{5.1.9}$$

$$= 3.2 \tag{5.1.10}$$

(iv) Recall from Def [2.3.2] that when we have grouped data, we can calculate the variance as,

$$\sigma^2 = \frac{\sum f_i x_i^2}{\sum f_i} - (\bar{x})^2 \tag{5.1.11}$$

$$= \frac{0^2(2) + 1^2(4) + 2^2(3) + 3^2(6) + 4^2(7) + 5^2(8)}{2 + 4 + 3 + 6 + 7 + 8} - 3.2^2 \tag{5.1.12}$$

$$= 2.49 \tag{5.1.13}$$

(v) Recall from Def [2.3.3], that for a 10% trimmed mean, we need to calculate the mean after discarding $0.1 \times 30 = 3$ data points from the upper and lower end of the distribution. Dropping the lower 3 means we have no students with 0 marks and one less student with 1 mark. Dropping the upper 3 means we have 3 less students with 5 marks. So the trimmed mean, $\bar{x}_{\text{trim}}$ can be computed as

$$\bar{x}_{\text{trim}} = \frac{0(2-2) + 1(4-1) + 2(3) + 3(6) + 4(7) + 5(8-3)}{2 + 4 + 3 + 6 + 7 + 8 - 3 - 3} \tag{5.1.14}$$

$$= 3.33 \tag{5.1.15}$$

(vi) Recall from Def [2.3.4] that the 30th percentile is the value 30% of the way through the distribution. This corresponds to the $0.3 \times 30 = 9^{\text{th}}$ term.

$$9^{\text{th}} \text{ term} = 2 \tag{5.1.16}$$

(vii) Recall that we can compute the lower quartile Q_1 and upper quartile Q_3 from Def [2.3.5],

$$\begin{aligned} Q_1 &= \frac{n+1}{4}^{\text{th}} \text{ term} && (5.1.17)\\ &= 7\frac{3}{4}^{\text{th}} \text{ term} && (5.1.18)\\ &= 7^{\text{th}} \text{ term} + \frac{3}{4}(8^{\text{th}} \text{ term} - 7^{\text{rd}} \text{ term}) && (5.1.19)\\ &= 2 + \frac{3}{4}(2-2) && (5.1.20)\\ &= 1 && (5.1.21)\\ Q_3 &= \frac{3(n+1)}{4}^{\text{th}} \text{ term} && (5.1.22)\\ &= 23\frac{1}{4}^{\text{th}} \text{ term} && (5.1.23)\\ &= 23^{\text{rd}} \text{ term} + \frac{1}{4}(24^{\text{th}} \text{ term} - 23^{\text{rd}} \text{ term}) && (5.1.24)\\ &= 5 + \frac{1}{4}(5-5) && (5.1.25)\\ &= 5 && (5.1.26) \end{aligned}$$

(b) **(i)** If a distribution is strongly skewed, the listed measures would not be able to represent the central value of the distribution accurately.

(ii) If we add a certain value to every data point, we expect that the variance should remain unchanged. We explore why in more detail in Appendix [15.4.1].

5.2 Module 2: Managing Uncertainty

3. **a)** We are given a probability density function $f(x)$ of a continuous random variable X.

(i) We know from Proposition [3.2.2] that for any probability density function $f(x)$, $\int_{-\infty}^{\infty} f(x).dx = 1$. This is simply the total area under the curve. In this case, our curve takes the form of a trapezoid. So we have,

$$\begin{aligned} \int_{-\infty}^{\infty} f(x).dx &= 1 && (5.2.1)\\ \text{Area of trapezoid} &= \frac{a+b}{2} \times h && (5.2.2)\\ &= \frac{10+4}{2} \times k = 1 && (5.2.3)\\ \Rightarrow k &= \frac{1}{7} && (5.2.4) \end{aligned}$$

(ii) For a continuous random variable X with a probability density function $f(x)$, we know from Note [3.2.1],

$$P(b \le X \le a) = \int_b^a f(x).dx \qquad (5.2.5)$$

Thus, we have that

$$\begin{aligned} P(4 \le X \le 6) &= \int_4^6 f(x).dx && (5.2.6)\\ &= (6-4) \times \frac{1}{7} && (5.2.7)\\ &= \frac{2}{7} && (5.2.8) \end{aligned}$$

where the integral was interpreted as the area of the rectangle of height $\frac{1}{7}$ and width $(6-4)$.

b) We are given a discrete random variable Y and its probability mass function $p(y)$

(i) Since Y is a discrete random variable, we know from Proposition [3.2.1] that the sum of all the probabilities in the range must equal to 1,

$$\sum_{\forall y} P(Y = y) = 1 \tag{5.2.9}$$

and that the probabilities of the random variable taking a value y must be between 0 and 1

$$0 \leq P(Y = y) \leq 1 \quad \forall y \in (0, 2, 4, 8) \tag{5.2.10}$$

(ii) With the probability mass function, it is possible to make further

a) Recall from Def [3.2.5] that if $F(y)$ is the cumulative distribution function for the random variable Y, then

$$\begin{aligned} F(y) &= P(Y \leq y) && (5.2.11) \\ &= \sum_{-\infty}^{y} P(Y = y) && (5.2.12) \end{aligned}$$

Hence, we can easily compute $F(6)$ as

$$\begin{aligned} F(6) &= \sum_{y=0}^{6} P(Y = y) && (5.2.13) \\ &= P(Y = 0) + P(Y = 2) + P(Y = 4) && (5.2.14) \\ &= 0.12 + 0.32 + 0.26 && (5.2.15) \\ &= 0.7 && (5.2.16) \end{aligned}$$

b) We know from Def [3.2.4], that

$$Var(Y) = E[Y^2] - (E[Y])^2 \tag{5.2.17}$$

Thus we need to compute $E[Y^2]$ and $E[Y]$. We can do this by Def [3.2.2] and Def [3.2.3] respectively

$$\begin{aligned} E[Y] &= \sum yP(Y = y) && (5.2.18) \\ &= 0.(0.12) + 2.(0.32) + 4.(0.26) + 8.(0.3) && (5.2.19) \\ &= 4.08 && (5.2.20) \\ E[Y^2] &= \sum y^2 P(Y = y) && (5.2.21) \\ &= 0^2.(0.12) + 2^2.(0.32) + 4^2.(0.26) + 8^2.(0.3) && (5.2.22) \\ &= 24.64 && (5.2.23) \\ \therefore Var(Y) &= 24.64 - (4.08)^2 = 7.99 && (5.2.24) \end{aligned}$$

(iii) Thus, we want to determine $P(Y_2 - Y_1 = 4)$. Before we proceed, we should note from Def [3.1.5] that the independence of Y_1 and Y_2 tell us that

$$P(Y_1 \cap Y_2) = P(Y_1) \times P(Y_2) \tag{5.2.25}$$

Now, for the given values of y, the desired condition is only satisfied by two pairs of Y_1, Y_2,

$$\begin{aligned} P(Y_2 - Y_1 = 4) &= P(Y_2 = 4 \cap Y_1 = 0)) + P(Y_2 = 8 \cap Y_1 = 4) && (5.2.26) \\ &= P(Y_2 = 4).P(Y_1 = 0) + P(Y_2 = 8).P(Y_1 = 4) && (5.2.27) \\ &= (0.26)(0.12) + (0.3)(0.26) && (5.2.28) \\ &= 0.1092 && (5.2.29) \end{aligned}$$

4. We are given information on the probabilities of choices.

a) We can represent the information on a probability tree diagram. Before we do this, let us define a few events:

Let A be the event A was chosen.

Let M be the event M was chosen.

Let O be the event O was chosen.

Let H be the event H was added.

Now, we can construct the tree diagram as follows:

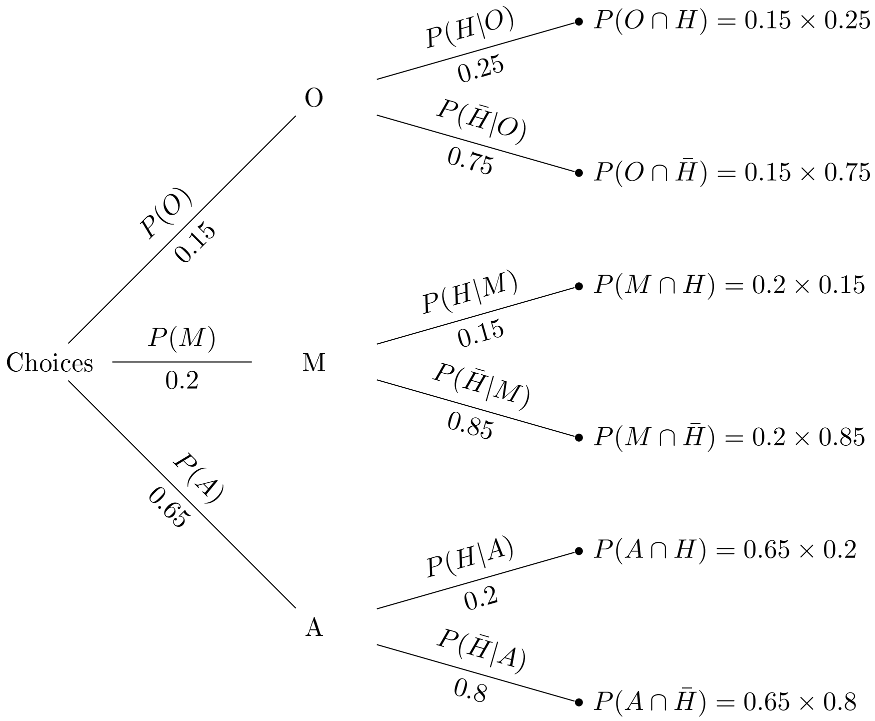

Choices
P(O)
0.15
O
$P(M)$
0.2
M
P(A)
0.65
A
P(H\O)
0.25
$P(\bar{H}|O)$
0.75
P(H\M)
0.15
$P(\bar{H}|M)$
0.85
P(H\A)
0.2
$P(\bar{H}|A)$
0.8
$P(O \cap H) = 0.15 \times 0.25$
$P(O \cap \bar{H}) = 0.15 \times 0.75$
$P(M \cap H) = 0.2 \times 0.15$
$P(M \cap \bar{H}) = 0.2 \times 0.85$
$P(A \cap H) = 0.65 \times 0.2$
$P(A \cap \bar{H}) = 0.65 \times 0.8$

Figure 5.1: A probability tree diagram of the information given

Where the probabilities at the end of the tree were computed in anticipation of the second part of the problem.

b) Using the tree diagram, we can determine many probabilities.

(i) We want to find $P(M \cap H)$. From the tree diagram, we see that

$$P(M \cap H) = 0.2 \times 0.15 = 0.03 \tag{5.2.30}$$

(ii) We want to find $P(H)$. To do this, we need to use conditional probability from Def [3.1.6],

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{5.2.31}$$

Furthermore, since the individual can add H to any of the three choices, we need to sum over each of the probabilities as follows,

$$\begin{aligned} P(H) &= P(H|O).P(O) + P(H|M).P(M) + P(H|A).P(A) && (5.2.32) \\ &= (0.25)(0.15) + (0.15)(0.2) + (0.2)(0.65) && (5.2.33) \\ &= 0.1975 && (5.2.34) \end{aligned}$$

(iii) We want to find $P(A \cup H)$. We must to be careful since these sets intersect

$$\begin{aligned} P(A \cup H) &= P(A) + P(H) - P(A \cap H) && (5.2.35) \\ &= 0.65 + 0.1975 - (0.65)(0.2) && (5.2.36) \\ &= 0.7175 && (5.2.37) \end{aligned}$$

(iv) We want to find $P(O|H)$. Once again, we need to use Def [3.1.6] for conditional probability,

$$P(O|H) = \frac{P(O \cap H)}{P(H)} = \frac{0.15 \times 0.25}{0.1975} = \frac{15}{79} \tag{5.2.38}$$

5.3 Module 3: Analyzing and Interpreting Data

5. **(a)** Since we are looking for a change in the mean i.e. both $\mu < \mu_0$ or $\mu > \mu_0$, a two-tailed test must be done.

(b) We can write the null and alternative hypothesis as follows:

$$\begin{aligned} H_0 &: \mu = 40 \text{ minutes} && (5.3.1) \\ H_1 &: \mu \neq 40 \text{ minutes} && (5.3.2) \end{aligned}$$

(c) Since, $n \geq 30$, we see that the Central Limit Theorem [4.1.1] is applicable.

(d) Testing at the 4% level of significance, we reject H_0 if

$$\begin{aligned} z_{\text{calc}} < -z_{\alpha/2} &\quad \text{or} \quad z_{\text{calc}} > z_{\alpha/2} && (5.3.3) \\ z_{\text{calc}} < -z_{0.02} &\quad \text{or} \quad z_{\text{calc}} > z_{0.02} && (5.3.4) \\ z_{\text{calc}} < -2.054 &\quad \text{or} \quad z_{\text{calc}} > -2.054 && (5.3.5) \end{aligned}$$

(e) If we increase the level of significance, we expect the critical region to increase. We see this by looking at Fig [4.3]. If we increase α, the region in which we reject H_0, the critical region, increases.

(f) We can calculate the test statistic following Section [4.2.2]. Since $n \geq 30$ and σ^2 is known

$$\begin{aligned} z_{\text{calc}} &= \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} && (5.3.6) \\ &= \frac{35 - 40}{9/\sqrt{50}} && (5.3.7) \\ &= -3.928 && (5.3.8) \end{aligned}$$

(g) Since $z_{\text{calc}} = -3.928 < -2,054$, we reject the null hypothesis, that the mean is equal to 40 minutes, in favor of the alternative hypothsis, that the mean is not equal to 40 minutes.

6. We are given 12 pairs of (x, y) data points.

(a) We can represent the information on a scatter plot.

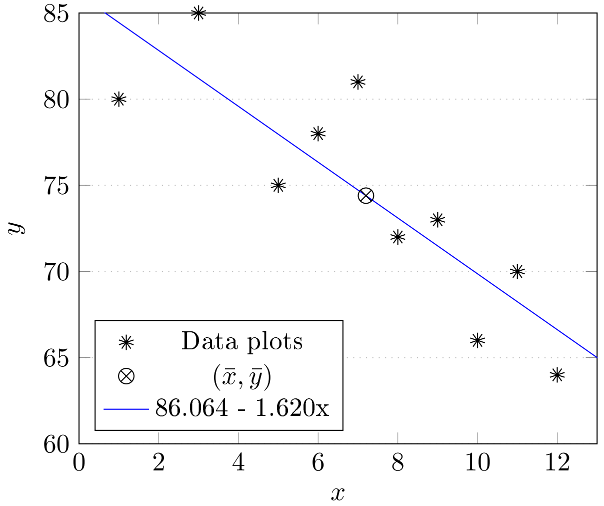


Figure 5.2: A scatter diagram of the data given for 9 data points.

(b) Using Def [4.4.1], we can calculate the product moment correlation coefficient to be:

$$\begin{aligned} r &= \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}} && (5.3.9)\\ &= \frac{10(5176) - 72(744)}{\sqrt{10(630) - 72^2}\sqrt{10(55760) - 744^2}} && (5.3.10)\\ &= -0.849 && (5.3.11) \end{aligned}$$

From Note [4.4.2], we can interpret this as a high negative linear correlation.

(c) We can calculate the equation of linear regression, $y = a + bx$, according to Def [4.4.2].

$$\begin{aligned} b &= \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} && (5.3.12)\\ &= \frac{10(5176) - 72(744)}{10(630) - 72^2} && (5.3.13)\\ &= -1.620 && (5.3.14) \end{aligned}$$

Using this, we can now calculate a,

$$\begin{aligned} a &= \bar{y} - b\bar{x} && (5.3.15)\\ &= \frac{\sum y}{n} - b\frac{\sum x}{n} && (5.3.16)\\ &= \frac{744}{10} + 1.620 \times \frac{72}{10} && (5.3.17)\\ &= 74.4 + 1.620(7.2) && (5.3.18)\\ &= 86.064 && (5.3.19) \end{aligned}$$

Thus, the equation of regression takes the form

$$y = 86.064 - 1.620x \qquad (5.3.20)$$

(d) Given that $x = 13$, we want to determine the corresponding value of y. This amounts to simply substituting $x = 13$ in our equation above,

$$\begin{aligned} y &= 86.064 - 1.620(13) && (5.3.21)\\ &= 65.004 \approx 65 \text{ trees} && (5.3.22) \end{aligned}$$

We note that this is not reliable as we are extrapolating beyond the range of the scatter diagram. The graph may not follow a linear scale as we have assumed.

Chapter 6

2006

6.1 Module 1: Collecting and Describing Data

1. We are given unordered data on durations of time to the nearest minute for 30 people.

(a) **(i)** We can illustrate this data on a stem and leaf diagram as follows

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

Key: 1|8 means 18

Figure 6.1: A stem and leaf diagram of the data given.

(ii) We see by direct comparison with Fig [2.1], that the distribution is positively skewed.

(b) We can determine the following statistics from the given data.

(i) Recall from Def [2.3.5], that the median, Q_2, is given by

$$\begin{aligned}
Q_2 &= \frac{n+1}{2}^{\text{th}} \text{ term} && (6.1.1)\\
&= \frac{31}{2}^{\text{th}} \text{ term} && (6.1.2)\\
&= 15.5^{\text{th}} \text{ term} && (6.1.3)\\
&= 15^{\text{th}} \text{ term} + 0.5 \times (16^{\text{th}} \text{ term} - 15^{\text{th}} \text{ term}) && (6.1.4)\\
&= 30 + 0.5 \times (31 - 30) && (6.1.5)\\
&= 30 + 0.5(1) && (6.1.6)\\
&= 30.5 && (6.1.7)
\end{aligned}$$

(ii) We can compute the lower quartile and the upper quartile, Q_3, from Def [2.3.5]. Hence,

$$\begin{aligned}
Q_1 &= \frac{n+1}{4}^{\text{th}} \text{ term} && (6.1.8)\\
&= \frac{31}{4}^{\text{th}} \text{ term} && (6.1.9)\\
&= 7.75^{\text{th}} \text{ term} && (6.1.10)\\
&= 7^{\text{th}} \text{ term} + 0.75 \times (8^{\text{th}} \text{ term} - 7^{\text{th}} \text{ term}) && (6.1.11)\\
&= 25 + 0.75(26 - 25) && (6.1.12)\\
&= 25.75 && (6.1.13)\\
Q_3 &= \frac{3(n+1)}{4}^{\text{th}} \text{ term} && (6.1.14)\\
&= 23.25^{\text{th}} \text{ term} && (6.1.15)\\
&= 23^{\text{th}} \text{ term} + 0.25 \times (24^{\text{th}} \text{ term} - 23^{\text{th}} \text{ term}) && (6.1.16)\\
&= 37 + 0.25 \times (39 - 37) && (6.1.17)\\
&= 37.5 && (6.1.18)
\end{aligned}$$

(iii) Recall that the range is simply the difference between the highest and lowest value. We take these from the first and last entry of our stem and leaf diagram.

$$\begin{aligned}
\text{Range} &= 51 - 18 && (6.1.19)\\
&= 33 && (6.1.20)
\end{aligned}$$

(c) We can represent this information on a box and whisker plot as follows

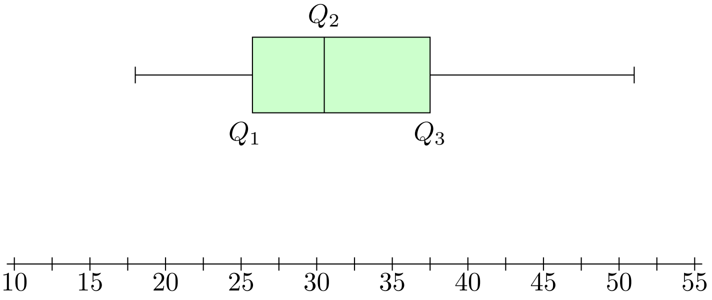


Figure 6.2: A Box and Whisker plot of the data given

2. **(a)** From Note [2.2.1], one advantage is that it is relatively cheap. On the other hand, one disadvantage is that members within subgroups of population may not have a proportional representation in the sample.

(b) First assign a unique number from 1 to 100 to every individual in your population. Next we use a random method to select 25 of these numbers. The simplest example is to put all the numbers on individual pieces of paper and place them in a bowl. Next, have a blind-folded researcher select 25 of these papers. The names corresponding to the numbers on the 25 selected pieces of paper repepresents your simple random sample of size 25.

(c) From Note [2.2.2], one advantage is that it gives a proportional representation of the population. One disadvantage is that it is time consuming to establish strata and execute.

(d) We first determine the size of the population

$$\begin{aligned}
N &= |X| + |Y| + |Z| && (6.1.21)\\
&= 200 + 125 + 175 && (6.1.22)\\
&= 500 && (6.1.23)
\end{aligned}$$

Now, in order to make sure that the sample represents these strata proportionately, we ensure that the ratio of the strata in the population is equal to the ratio of the strata in the sample:

$$\begin{aligned}
\frac{200}{500} &= \frac{x}{60} &(6.1.24)\\
\Rightarrow x &= \frac{200}{500} \times 60 &(6.1.25)\\
&= 24 \text{ teachers} &(6.1.26)\\
\frac{125}{500} &= \frac{y}{60} &(6.1.27)\\
\Rightarrow y &= \frac{125}{500} \times 60 &(6.1.28)\\
&= 15 \text{ teachers} &(6.1.29)\\
\frac{175}{500} &= \frac{z}{60} &(6.1.30)\\
\Rightarrow z &= \frac{175}{500} \times 60 &(6.1.31)\\
&= 21 \text{ teachers} &(6.1.32)
\end{aligned}$$

6.2 Module 2: Managing Uncertainty

3. We are given a table organizing data in terms of classes.

a) We can use this table to determine many properties.

(i) Let H be the event that the item is type H. Thus, we want to find $P(H)$. We can do this by refering to the definition of probability in Eq [3.1.3]

$$\begin{aligned}
P(H) &= \frac{|H|}{|S|} &(6.2.1)\\
&= \frac{218}{400} &(6.2.2)\\
&= 0.54 &(6.2.3)
\end{aligned}$$

Where S is the sample space.

(ii) Let R be the event that the item is for the class Pr. Thus, we want to find $P(\bar{R})$. From Def [3.1.3], we know that we can find the complement by

$$\begin{aligned}
P(\bar{R}) &= 1 - P(R) &(6.2.4)\\
&= 1 - \frac{|R|}{|S|} &(6.2.5)\\
&= 1 - \frac{167}{400} &(6.2.6)\\
&= \frac{223}{400} &(6.2.7)\\
&= 0.5825 &(6.2.8)
\end{aligned}$$

(iii) Let A be the event that the book is for class SL. Let B be the event that the book type S. Thus, we want to find $P(A \cap B)$,

$$\begin{aligned}
P(A \cap B) &= \frac{|A \cap B|}{|S|} &(6.2.9)\\
&= \frac{64}{400} &(6.2.10)\\
&= 0.16 &(6.2.11)
\end{aligned}$$

(iv) Let T be the event that a book is for class T. Thus, we want to find $P(T \cup H)$,

$$\begin{aligned} P(T \cup H) &= \frac{|T \cup H|}{|S|} && (6.2.12)\\ &= \frac{100+83+35+51}{400} && (6.2.13)\\ &= 0.6725 && (6.2.14) \end{aligned}$$

b) We can reword this question as follows: *Find the probability that a randomly chosen item is for the class SL, given that it is of type H.* Keeping the notation established in (a), we want to find $P(A|H)$. Using Def [3.1.6], we can determine the conditional probability as,

$$\begin{aligned} P(A|H) &= \frac{P(A \cap H)}{P(H)} && (6.2.15)\\ &= \frac{|A \cap H|}{|H|} && (6.2.16)\\ &= \frac{83}{218} && (6.2.17)\\ &= 0.381 && (6.2.18) \end{aligned}$$

c) This can only happen if the first item is of type S and the second is of type H, or vice-versa. Thus, we want to find

$$\begin{aligned} P((1^{st} \in H \cap 2^{nd} \in B) \cup (1^{st} \in B \cap 2^{nd} \in H)) &= P(1^{st} \in H \cap 2^{nd} \in B) && (6.2.19)\\ &+ P(1^{st} \in B \cap 2^{nd} \in H) && (6.2.20)\\ &= \frac{218}{400}.\frac{182}{399} + \frac{182}{400}.\frac{218}{399} && (6.2.21)\\ &= 0.497 && (6.2.22) \end{aligned}$$

Where we used Def [3.1.4] to write $P(A \cup B) = P(A) + P(B)$, if A and B are mutually exclusive. Alternatively, we can also use a combinatoric approach.

$$\begin{aligned} P &= \frac{\#\text{ ways to CHOOSE 1 of type H} \times \#\text{ ways to CHOOSE 1 of type S}}{\#\text{ ways to CHOOSE 2 of any type}} && (6.2.23)\\ &= \frac{{}^{182}C_1 \times {}^{218}C_1}{{}^{400}C_2} && (6.2.24)\\ &= 0.497 && (6.2.25) \end{aligned}$$

4. **a)** We are given that $n = 12$ and the probability of success is $p = 0.7$.

(i) We know from Note [3.3.2] that there are four conditions that describe a binomial distribution:

1. The experiment consist of a fixed number of trails n.
2. The trials are independent.
3. Each trail can be classified as a success or failure.
4. The probability of success, p, is constant

(ii) Let X be the number of successes in 12 trials. Thus, we want to find $P(X < 12)$. Since X satisfies the conditions above, we can say that is is a binomial random variable, $X \sim Bin(12, 0.7)$. Then, according to Def [3.3.1]

$$P(X = x) = \binom{12}{x} 0.7^x 0.3^{12-x}\,, \quad x \in (0, 1, ..., 12) \qquad (6.2.26)$$

The long way to do this would be,

$$P(X < 12) = P(X = 0) + P(X = 1) + ... + P(X = 10) + P(X = 11) \qquad (6.2.27)$$

However, it would be much quicker to recognize that,

$$\begin{aligned} P(X<12) &= 1-P(X\geq 12) && (6.2.28)\\ &= 1-P(X=12) && (6.2.29)\\ &= 1-\binom{12}{12}0.7^{12}0.3^{0} && (6.2.30)\\ &= 0.986 && (6.2.31)\end{aligned}$$

b) From Note [3.4.2], since

$$\begin{aligned} n_2p &= 200(0.7) && (6.2.32)\\ &= 140>5 && (6.2.33)\\ n_2pq &= 200(0.7)(0.3) && (6.2.34)\\ &= 42>5 && (6.2.35)\end{aligned}$$

we can use the normal approximation to the binomial distribution. Let Y be the number of success in 200 trials. Then, by the normal approximation to the binomial distribution, Theorem [3.4.1], we can say that X is approximately normally distributed with mean and variance

$$\begin{aligned} \mu &= np && (6.2.36)\\ &= 200(0.7) && (6.2.37)\\ &= 140 && (6.2.38)\\ \sigma^2 &= npq && (6.2.39)\\ &= 200(0.7)(0.3) && (6.2.40)\\ &= 42 && (6.2.41)\end{aligned}$$

So $X\sim N(140,42)$. Now, we want to find $P(X<145)$, but since we have approximated our discrete random variable with a continuous random variable, we must apply a continuity correction according to Note [3.4.3]. Hence,

$$P(X<145)\to P(X<144.5) \qquad (6.2.42)$$

We want to find $P(X<144.5)$. We can proceed by standardizing,

$$\begin{aligned} P(X<144.5) &= P\left(\frac{X-\mu}{\sigma}<\frac{144.5-\mu}{\sigma}\right) && (6.2.43)\\ &= P\left(Z<\frac{144.5-140}{\sqrt{42}}\right) && (6.2.44)\\ &= P(Z<0.6944) && (6.2.45)\\ &= \Phi(0.6944) && (6.2.46)\\ &= 0.7549 && (6.2.47)\end{aligned}$$

c) For large n, the conditions of the experiment may change due to internal (tiredness) and external factors (wind) and hence change the probability of success p. In fact, even for small n, there may be variability in internal factors (position) and as a result change p. If p is not constant then this breaks the criteria for X to be modeled by a binomial distribution.

Additionally, the trials may not be independent. Succeeding or failing on the first few tries may affect the ability to succeed in the later trials.

6.3 Module 3: Analyzing and Interpreting Data

5. (a) We can define the null and alternative hypotheses as follows:

H_0 : The mean mathematics score of Form 1 students at a particular high school is 85.

H_1 : The mean mathematics score of Form 1 students at a particular high school is less than 85.

(b) Recall from Section [4.2.3] that the t-distribution should be used since the variance, σ^2, is not known and the sample size, n, is less than 30.

(c) If we decrease the level of significance, we expect the critical region to decrease. We see this by looking at Fig [4.3]. If we decrease α, the region in which we reject H_0, the critical region, decreases.

(d) **(i)** The number of degrees of freedom, d.o.f., is given by,

$$\begin{aligned} d.o.f &= n-1 && (6.3.1)\\ &= 10-1 && (6.3.2)\\ &= 9 && (6.3.3) \end{aligned}$$

(ii) The critical value is given by

$$\begin{aligned} t_\alpha^{n-1} &= t_{0.025}^{9} && (6.3.4)\\ &= 2.262 && (6.3.5) \end{aligned}$$

(iii) We reject the null hypothesis, H_0, if

$$\begin{aligned} & t_{test} < -t_\alpha^{n-1} && (6.3.6)\\ \Rightarrow\ & t_{test} < -2.262 && (6.3.7) \end{aligned}$$

(e) Given that $\sum x = 824$ and $\sum x^2 = 68510$

The test statistic for a t-test is given by,

$$t_{test} = \frac{\bar{x}-\mu}{\sqrt{\frac{\hat{\sigma}^2}{n}}} \qquad (6.3.8)$$

However, we must first calculate the unbiased estimate for the sample variance. We do this according to Def [4.1.3]

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n-1}\left[\sum x^2 - \frac{(\sum x)^2}{n}\right] && (6.3.9)\\ &= \frac{1}{9}\left[68510 - \frac{824^2}{10}\right] && (6.3.10)\\ &= 68.044 && (6.3.11) \end{aligned}$$

Thus, we can now calculate the test statistic as,

$$\begin{aligned} t_{test} &= \frac{82.4-85}{\sqrt{\frac{68.044}{10}}} && (6.3.12)\\ &= -0.997 && (6.3.13) \end{aligned}$$

(f) Since our test statistic $t_{\text{calc}} = -0.997 > -2.262$, it does not lie within the critical region. Hence, we fail to reject the null hypothesis. Thus, the principal may conclude that the mean mathematics score of Form 1 students at a particular high school is 85.

6. We are given 6 pairs of (x, y) data.

(a) We can represent the information given in a scatter diagram

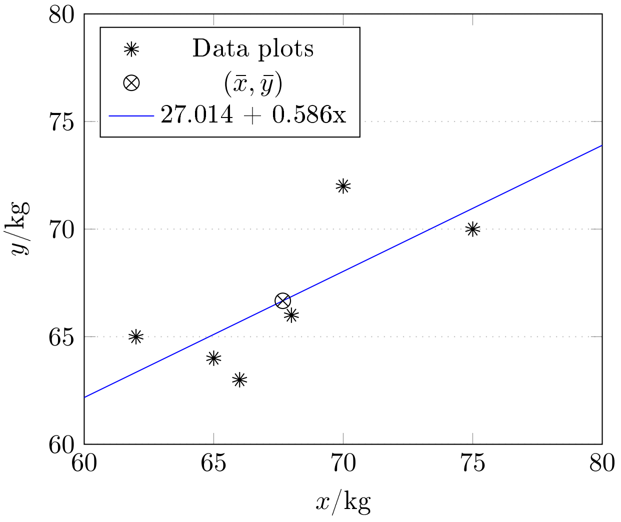


Figure 6.3: A scatter diagram for the information given.

(b) We can calculate the product moment correlation coefficient for this data according to Def [4.4.1]

$$\begin{align} r &= \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}} \tag{6.3.14}\\ &= \frac{6(27126) - (406)(400)}{\sqrt{6(27574) - 406^2}\sqrt{6(26730) - 400^2}} \tag{6.3.15}\\ &= 0.741 \tag{6.3.16}\end{align}$$

From Note [4.4.2], we can interpret this as a high positive linear correlation.

(c) **(i)** We know from Def [4.4.2], that the line of linear regression of X on Y takes the form,

$$y = a + bx \tag{6.3.17}$$

where b is given by

$$\begin{align} b &= \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} \tag{6.3.18}\\ &= \frac{6(27126) - (406)(400)}{6(27574) - 406^2} \tag{6.3.19}\\ &= 0.586 \tag{6.3.20}\end{align}$$

and a,

$$\begin{align} a &= \bar{y} - b\bar{x} \tag{6.3.21}\\ &= \frac{\sum y}{n} - b\frac{\sum x}{n} \tag{6.3.22}\\ &= \frac{400}{6} - 0.586 \times \frac{406}{6} \tag{6.3.23}\\ &= 27.014 \tag{6.3.24}\end{align}$$

Therefore, the equation of the regression of y on x is given by:

$$y = 27.014 + 0.586x \tag{6.3.25}$$

We note that the coordinate $(\bar{x}, \bar{y}) = (67.667, 66.667)$ is on this line.

(ii) We can plot this line on our scatter diagram.

(d) **(i)** We want to find what value of y that corresponds to an x value of 62.

$$\begin{align} y &= 27.014 + 0.586x \tag{6.3.26}\\ &= 27.014 + 0.586(62) \tag{6.3.27}\\ &= 63.346 \tag{6.3.28}\end{align}$$

(ii) Using this, we want to calculate the difference between the actual and predicted values for the mass of a daughter at age 30 from a mother whose mass at age 30 is 62kg.

$$\begin{aligned}
\text{Actual mass} &= 65 && (6.3.29)\\
\text{Predicted mass} &= 63.346 && (6.3.30)\\
\text{Difference} &= 65 - 63.346 && (6.3.31)\\
&= 1.645 && (6.3.32)
\end{aligned}$$

(e) It would be more appropriate to make predictions if we wanted to predict a value of x for a given value of y. For this data set, this would mean predicting the mass of a mother at age 30 given the mass of their eldest daughter at age 30.

Chapter 7

2007

7.1 Module 1: Collecting and Describing Data

1. **(a)** We distinguish between a population and sample according to Def [2.1.5] and Def [2.1.7]

The population refers to all the elements or individuals that meet the selection criteria for a group to be studied, and from which a sample is usually chosen from to be examined in detail.

A sample is a subset of the population which is studied in detail so as to find numerical data, central tendencies or any other patterns. These statistics can usually be extended to the population but depend on the degree of representation of the sample to the population.

(b) Similarity: Both quota and stratified random sampling involve grouping the individuals of the population into mutually exclusive subsets

Difference: Quota sampling is non-probabilistic i.e. it is based on judgment, proximity, etc. Stratified random sampling is a probabilistic method as each member of a stratum has an equal chance of being selected.

(c) **(i)**

- P: Quota
- Q: Stratified random sampling.
- R: Simple random sampling
- S: Systematic sampling

(ii) The method of stratified random sampling, Q, usually gives the most representative sample.

(iii) If we apply Q, 5 of the 16 people selected are over 50. We know that method Q ensures that the proportions of the people over 50 in the sample is the same as in the proportion in the population. Thus, if x is the number of people over 50 in the population, then

$$\frac{5}{16} = \frac{x}{656} \tag{7.1.1}$$

$$x = 656 \times \frac{5}{16} \tag{7.1.2}$$

$$= 205 \tag{7.1.3}$$

(iv) This question in intentionally omitted.

(v) For this to work, k must be such that it equally partitions 656 people 16 times. So,

$$k = \frac{656}{16} \tag{7.1.4}$$

$$= 41 \tag{7.1.5}$$

2. **(a)** We are given a stem and leaf diagram.

(i)

- The smallest values are at the top of the stem. So $p = 47$.
- The highest values are at the bottom of the stem. So $q = 92$.

- Recall from Def [2.3.5], that we can compute Q_1 as

$$\begin{aligned} Q_1 &= \frac{n+1}{4}^{\text{th}} \text{ term} && (7.1.6)\\ &= \frac{21}{4}^{\text{th}} \text{ term} && (7.1.7)\\ &= 5.25^{\text{th}} \text{ term} && (7.1.8)\\ &= 5^{\text{th}} \text{ term} + 0.25 \times (6^{\text{th}} \text{ term} - 5^{th} \text{ term}) && (7.1.9)\\ &= 64 + 0.25 \times (65 - 64) && (7.1.10)\\ r &= 64.25 && (7.1.11) \end{aligned}$$

- Recall from Def [2.3.5], that we can compute Q_2 as

$$\begin{aligned} Q_2 &= \frac{n+1}{2}^{\text{th}} \text{ term} && (7.1.12)\\ &= \frac{21}{2}^{\text{th}} \text{ term} && (7.1.13)\\ &= 10.5^{\text{th}} \text{ term} && (7.1.14)\\ &= 10^{\text{th}} \text{ term} + 0.5 \times (11^{\text{th}} \text{ term} - 10^{\text{th}} \text{ term}) && (7.1.15)\\ &= 64 + 0.5 \times (68 - 64) && (7.1.16)\\ s &= 66 && (7.1.17) \end{aligned}$$

- Recall from Def [2.3.5], that we can compute Q_3 as

$$\begin{aligned} Q_3 &= \frac{3(n+1)}{4}^{\text{th}} \text{ term} && (7.1.18)\\ &= 15.75^{\text{th}} \text{ term} && (7.1.19)\\ &= 15^{\text{th}} \text{ term} + 0.75 \times (16^{\text{th}} \text{ term} - 15^{th} \text{ term}) && (7.1.20)\\ &= 86 + 0.75 \times (87 - 86) && (7.1.21)\\ t &= 86.75 && (7.1.22) \end{aligned}$$

(ii) **a)** The range is defined to be the difference between the highest and lowest value. Thus, the range of the English marks is,

$$\begin{aligned} \text{Range} &= 92 - 42 && (7.1.23)\\ &= 50 && (7.1.24) \end{aligned}$$

b) Recall from Def [2.3.5], that we can compute the interquartile range as,

$$\begin{aligned} IQR &= Q_3 - Q1 && (7.1.25)\\ &= 86.75 - 64.25 && (7.1.26)\\ &= 22.5 && (7.1.27) \end{aligned}$$

c) One observation from data with the table is that students perform better in Mathematics than in English. This is indicated from the fact that Q_1, Q_2 and Q_3 is higher for Mathematics than for English.

(iii) By comparison with Fig [2.1], one can immediately tell that the Mathematics distibution is negatively skewed. However, we can also tell this by the fact that it satisfies the conditions in Def [2.3.6]

$$\begin{aligned} (Q_3 - Q_2) &< (Q_2 - Q_1) && (7.1.28)\\ 86.75 - 77.5 &< 77.5 - 64.25 && (7.1.29)\\ 9.25 &< 13.25 && (7.1.30) \end{aligned}$$

(iv) Recall from Def [2.3.3] that a 10% trimmed mean of 20 data points means we discard $0.1 \times 20 = 2$ data points from the top and bottom of the distribution. So we discard 47 and 52 from the bottom and 94 and 93 from the top.

Thus, we calculate the trimmed mean to be,

$$\begin{aligned}\bar{y} &= \frac{\sum y}{n} &(7.1.31)\\ &= \frac{53 + ... + (80 + x) + ...88}{20 - 4} &(7.1.32)\\ &= \frac{1127 + (80 + x)}{16} &(7.1.33)\end{aligned}$$

But we are told that the trimmed mean $\bar{y} = 75.5$. So

$$\begin{aligned}75.5 &= \frac{1127 + (80 + x)}{16} &(7.1.34)\\ (80 + x) &= 75.5 \times 16 - 1127 &(7.1.35)\\ &= 81 &(7.1.36)\\ x &= 1 &(7.1.37)\end{aligned}$$

7.2 Module 2: Managing Uncertainty

3. A random sample with 300 individuals was taken and the results are given in a table.

a) Using the table we can determine the probability of some events. To do this we refer to Eq [3.1.3],

(i) Let M be the event that a individual is a M. Thus, we want to find $P(M)$. According to Eq [3.1.3],

$$\begin{aligned}P(M) &= \frac{|M|}{|S|} &(7.2.1)\\ &= \frac{140}{300} &(7.2.2)\\ &= \frac{7}{15} \approx 0.467 &(7.2.3)\end{aligned}$$

Where S denotes the sample space.

(ii) Let C be the event that an individual prefers C. Thus, we want to find $P(M \cap C)$,

$$\begin{aligned}P(M \cap C) &= \frac{|M \cap C|}{|S|} &(7.2.4)\\ &= \frac{106}{300} \approx 0.353 &(7.2.5)\end{aligned}$$

(iii) Let F be the event that a an individual is F. Let N be the event that a person prefers WM. Thus, we want to find $P(N \cup F)$,

$$\begin{aligned}P(N \cup F) &= \frac{|N \cup F|}{|S|} &(7.2.6)\\ &= \frac{64 + 96 + 34}{300} = \frac{194}{300} &(7.2.7)\\ &= \frac{97}{150} \approx 0.647 &(7.2.8)\end{aligned}$$

(iv) Let C be the event that an individual prefers WC. Thus, we want to find $P(C|M)$. To do this, we use conditional probability in Def [3.1.6],

$$\begin{aligned}P(C|M) &= \frac{P(C \cap M)}{P(M)} &(7.2.9)\\ &= \frac{|C \cap M|}{|S|} \div \frac{|M|}{|S|} &(7.2.10)\\ &= \frac{|M \cap C|}{|M|} &(7.2.11)\\ &= \frac{106}{140} = \frac{53}{70} \approx 0.757 &(7.2.12)\end{aligned}$$

b) We are given two events A and B. [1]

(i) Recall from Def [3.1.5] that the events A and B are independent if and only if

$$P(A \cap B) = P(A) \times P(B) \tag{7.2.13}$$

Thus, to determine if A and B are independent we need to compute both sides of the equation and see if they are equal.

$$P(A \cap B) \equiv P(M \cap C) = \frac{106}{300} \tag{7.2.14}$$

$$P(A) \equiv P(M) = \frac{7}{15} \tag{7.2.15}$$

$$P(B) \equiv P(C) = \frac{|C|}{|S|} = \frac{170}{300} = \frac{17}{30} = 0.567 \tag{7.2.16}$$

$$P(M \cap C) \overset{?}{=} P(M) \times P(C) \tag{7.2.17}$$

$$\frac{106}{300} \neq \frac{7}{15} \times \frac{17}{30} \tag{7.2.18}$$

Therefore the events A and B are not independent.

(ii) **a)** $A' \cap B'$ translates directly to 'the complement of A intersect the complement of B'. This corresponds to the event that 'An individual chosen at random is not M and does not prefer WC'. Alternatively, we can use De Morgan's Law [3.1.1] to write

$$A^c \cap B^c = (A \cup B)^c \tag{7.2.19}$$

This now translates directly as neither A nor B, corresponding to 'an individual chosen at random is neither M nor prefers WC'. Another option is to consider what this means in terms of the rest of the sample space. What the complement does is leave us with the event 'the individual is F and prefers WM'.

b) $B|A$ translates directly to 'B given A'. This corresponds to 'Given that an M chosen, the M prefers WC'.

4. **(a)** We are given $|R| = 4$, $|W| = 5$.

(i) I can construct the tree diagram as follows:

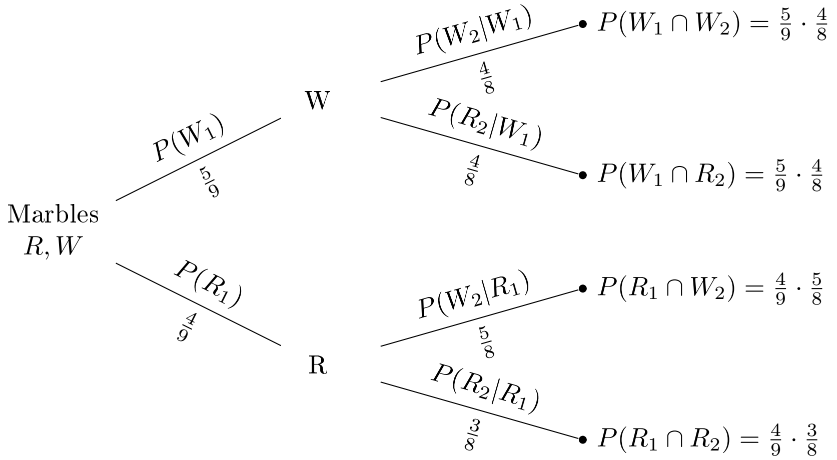


Figure 7.1: A tree diagram illustrating the probability of the possible combinations. The subscript was introduced to establish the order of the events.

(ii) We want to find $P(R_1 \cap W_2)$. From our tree diagram, we see that,

$$P(R_1 \cap W_2) = \frac{4}{9} \times \frac{5}{8} = \frac{5}{18} \tag{7.2.20}$$

[1] Did they just try to define my gender?...Triggered!

(iii) We want to find the probability of W_1W_2 or R_1R_2 i.e. $P(W_1W_2 \cup R_1R_2)$. Since these events are mutually exclusive, we use Def [3.1.4]

$$\begin{aligned} P(W_1W_2 \cup R_1R_2) &= P(W_1W_2) + P(R_1R_2) && (7.2.21)\\ &= P(W_1 \cap W_2) + P(R_1 \cap R_2) && (7.2.22)\\ &= \left(\frac{5}{9} \times \frac{4}{8}\right) + \left(\frac{4}{9} \times \frac{3}{8}\right) = \frac{4}{9} && (7.2.23) \end{aligned}$$

(b) We are given a probability density function $f(x)$ for a continuous random variable x. We know from Proposition [3.2.2] that the probability density function must satisfy the condition that

$$\int_{-\infty}^{\infty} f(x)dx = 1 \tag{7.2.24}$$

This is simply the area under the curve. Applying this to our function

$$\int_{-\infty}^{\infty} f(x)dx = \int_{0}^{4} f(x)dx = 1 \tag{7.2.25}$$

To compute the area, we can split our object into two parts A and B, where A is a trapezoid with vertexes $(0,k),(0,0),(3,0)$ and $(3,0.1)$ and B is a rectangle with vertexes $(3,0),(3,0.1),(4,0.1)$ and $(4,0)$. Now we compute the area under the curve as the sum areas of A and B.

$$\begin{aligned} \text{Area of } A &= \frac{a+b}{2} \times h && (7.2.26)\\ &= \frac{1}{2}(0.1+k)(3) && (7.2.27)\\ \text{Area of } B &= c \times d && (7.2.28)\\ &= 1 \times 0.1 && (7.2.29) \end{aligned}$$

Putting this all together:

$$\begin{aligned} \text{Area under curve} &= \text{Area of } A + \text{Area of } B && (7.2.30)\\ 1 &= \frac{1}{2}(0.1+k)(3) + 1(0.1) && (7.2.31)\\ \therefore k &= 0.5 && (7.2.32) \end{aligned}$$

7.3 Module 3: Analyzing and Interpreting Data

5. Recall from Section [4.3] the procedure to calculate a χ^2 test at the 5% level of significance.

(a) H_0: There is no association between the surgeon performing the operation and the patient being transferred.

H_1: There is an association between the surgeon performing the operation and the patient being transferred.

(b) We can use the information given to compute the values of $\frac{(O-E)^2}{E}$.

O	E	$\frac{(O-E)^2}{E}$
7	8.519	0.271
8	13.05	1.954
14	7.431	5.807
40	38.481	0.060
64	58.95	0.433
27	33.569	1.285
	Total	9.810

Table 7.1: A summary of the observed values O, expected values E, and the value of $\frac{(O-E)^2}{E}$.

(c) To finish the χ^2 test, we need to determine other values.

(i) We can calculate the number of degrees of freedom as follows

$$\begin{aligned} D.o.f &= (r-1)(c-1) && (7.3.1)\\ &= (3-1)(2-1) && (7.3.2)\\ &= 2 && (7.3.3) \end{aligned}$$

Where r and c represent the number of rows and columns respectively.

(ii) To calculate the critical region, we need to find the χ^2 value in the table that corresponds to 2 degrees of freedom and a 5% level of significance. Thus, the critical region is

$$\chi^2_{calc} > 5.991 \qquad (7.3.4)$$

(d) We can determine χ^2_{calc} from our table

$$\chi^2_{\text{calc}} = 9.810 \qquad (7.3.5)$$

Since our value of $\chi^2_{calc} = 9.810 > 5.991$, it lies within the critical region. Hence, we must reject the null hypothesis in favor of the alternative hypothesis. Thus, we conclude that there is an association between the surgeon performing the operation and the patient being transferred.

6. We are given that $\mu = 45$ minutes and $\sigma = 5$ minutes.

(a) **(i)** Let X be the continuous random variable representing the time.
Recall the Central Limit Theorem [4.1.1]. We can apply this to our X,

$$\bar{X} \sim N\left(45, \frac{5^2}{49}\right) \qquad (7.3.6)$$

Thus, we want to find $P(\bar{X} > 46)$. We can proceed by standardizing,

$$\begin{aligned} P(\bar{X} > 46) &= P\left(\frac{\bar{X}-\mu}{\sigma^2} > \frac{46-45}{\sqrt{\frac{25}{49}}}\right) && (7.3.7)\\ &= P(Z > 1.4) && (7.3.8)\\ &= 1 - \Phi(1.4) && (7.3.9)\\ &= 0.0808 && (7.3.10) \end{aligned}$$

(ii) We know that Central Limit Theorem [4.1.1] is valid when $n \geq 30$.

(b) **(i)** A two-tailed test will be more appropriate for this situation. This is because our null and alternative hypothesis is given by

$$\begin{aligned} H_0 &: \mu = 45 && (7.3.11)\\ H_1 &: \mu \neq 45 && (7.3.12) \end{aligned}$$

(ii) Recall from Section [4.2.3] that we use the z-distribution since σ^2 is known, and that for two-tailed test, we reject H_0 if

$$\begin{aligned} Z_{calc} > Z_{\alpha/2} &\quad \text{or} \quad Z_{calc} < -Z_{\alpha/2} && (7.3.13)\\ Z_{calc} > Z_{0.05/2} &\quad \text{or} \quad Z_{calc} < -Z_{0.05/2} && (7.3.14)\\ Z_{calc} > 1.96 &\quad \text{or} \quad Z_{calc} < -1.96 && (7.3.15) \end{aligned}$$

(iii) We are given that $z_{\text{calc}} = 4.24 > 1.96$. Since our test statistic lies within the critical region, we must reject the null hypothesis in favor of the alternative hypothesis. Thus, we conclude that the mean waiting time of patients for a particular doctor is not 45 minutes.

Chapter 8

2008

8.1 Module 1: Collecting and Describing Data

1. (a) (i) There are many reasons why in some cases it is more appropriate to examine a sample rather than a population

1. A well chosen sample should be representative of the population so errors detected within the sample can be used to infer errors in printing the population
2. If one waits for a census of the printing of the newspapers, the company may not have time to rectify an error in time for paper distribution
3. It is not practical in this case to examine thousands of newspapers, as it may require greater manpower which is not necessarily available.

(ii) Clusters are groups of similar members whereas strata are groups of different members. So in this case, a cluster sample may look at papers printed by 3 out of 5 printing machines whereas a stratified sample will look at papers from all 5 machines in proportion to how many papers are printed by each machine.
Cluster sampling is non-random whereas stratified sampling is random. Hence, as cluster sample may be chosen from the 3 machines that are functioning correctly while the other 2 machines are malfunctioned. While the choice of which machines to check may be random, cluster sampling could miss a proportion of the population with a hidden common trait, i.e. misprints. Stratified sampling would easily eliminate this hidden common characteristic across all machines.

(iii) Recall that in stratified random sampling, the proportions of the strata in the population are the same as the proportions of the strata in the sample. In order to keep the proportions the same, we must first determine the size of the population, N,

$$N = 35 + 38 + 41 + 42 + 46 + 39 + 59 \tag{8.1.1}$$

$$= 300 \tag{8.1.2}$$

Let x be the amount of our strata in our sample. In order to keep the proportions the same, we require that

$$\frac{x}{5} = \frac{35}{300} \tag{8.1.3}$$

$$\Rightarrow x = \frac{35}{300} \times 5 \tag{8.1.4}$$

$$= \frac{7}{12} \tag{8.1.5}$$

Therefore we require to have 583 papers from Monday's production.

(b) In drawing a pie chart we must ensure that the proportion of the circle is equivalent to the proportion of the stratum we wish to represent.

$$\begin{aligned}
\text{M} &= \frac{35}{300} \times 360 = 42^\circ && (8.1.6)\\
\text{Tu} &= \frac{38}{300} \times 360 = 45.6^\circ \approx 46^\circ && (8.1.7)\\
\text{W} &= \frac{41}{300} \times 360 = 49.2^\circ \approx 49^\circ && (8.1.8)\\
\text{Tr} &= \frac{42}{300} \times 360 = 50.4^\circ \approx 50^\circ && (8.1.9)\\
\text{F} &= \frac{46}{300} \times 360 = 55.2^\circ \approx 55^\circ && (8.1.10)\\
\text{Sa} &= \frac{39}{300} \times 360 = 46.8 \approx 47^\circ && (8.1.11)\\
\text{Su} &= \frac{59}{300} \times 360 = 70.8 \approx 71^\circ && (8.1.12)\\
& && (8.1.13)
\end{aligned}$$

We can now use the information to construct the following pie chart

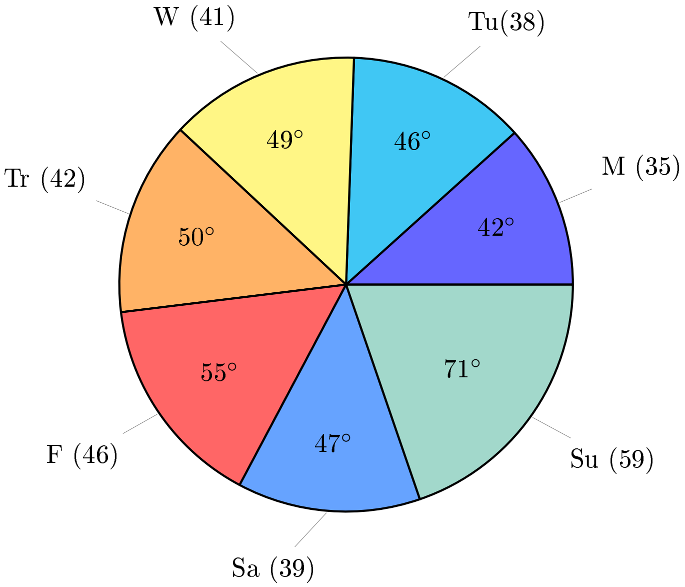


Figure 8.1: A pie chart representing the relative sizes of the strata given.

(c) **(i)** Recall that we can calculate the mean by Def [2.3.1]

$$\begin{aligned}
\bar{x} &= \frac{\sum x}{N} && (8.1.14)\\
&= \frac{35+38+41+42+46+39+59}{7} && (8.1.15)\\
&= 42.8571 && (8.1.16)
\end{aligned}$$

Therefore the mean daily production is 42857 newspapers.

(ii) Recall that we can compute the variance by Def [2.3.2]

$$
\begin{aligned}
\sigma^2 &= \frac{1}{N}\sum(\bar{x}-x)^2 && (8.1.17)\\
&= \frac{1}{7}((35-42.8571)^2+(38-42.8571)^2+(41-42.8571)^2 && (8.1.18)\\
&+ (42-42.8571)^2+(46-42.8571)^2+(39-42.8571)^2+(59-42.8571)^2) && (8.1.19)\\
&= 53.551 && (8.1.20)\\
& && (8.1.21)
\end{aligned}
$$

Alternatively, we can use the other equation in Def [2.3.2

$$
\begin{aligned}
\sigma^2 &= \frac{\sum x^2}{n}-(\bar{x})^2 && (8.1.22)\\
&= \frac{\sum 35^2+38^2+41^2+42^2+46^2+39^2+59^2}{7}-(42.8571)^2 && (8.1.23)\\
&\approx 53.551 && (8.1.24)
\end{aligned}
$$

[1]Hence, we can now compute the standard deviation,

$$\sigma = \sqrt{53.551} = 7.318 \qquad (8.1.25)$$

2. **(a)** We can construct the cumulative frequency table as follows

Class Limits	Cumulative Frequency
10 - 30	15
30 - 45	50
45 - 60	104
60 - 90	146
90 - 120	156

Table 8.1: A cumulative frequency table representing the data provided

(b) We can use Table [8.1] to construct a cumulative frequency curve as follows:

[1] To get identical results for both you need to keep the exact values throughout the calculation

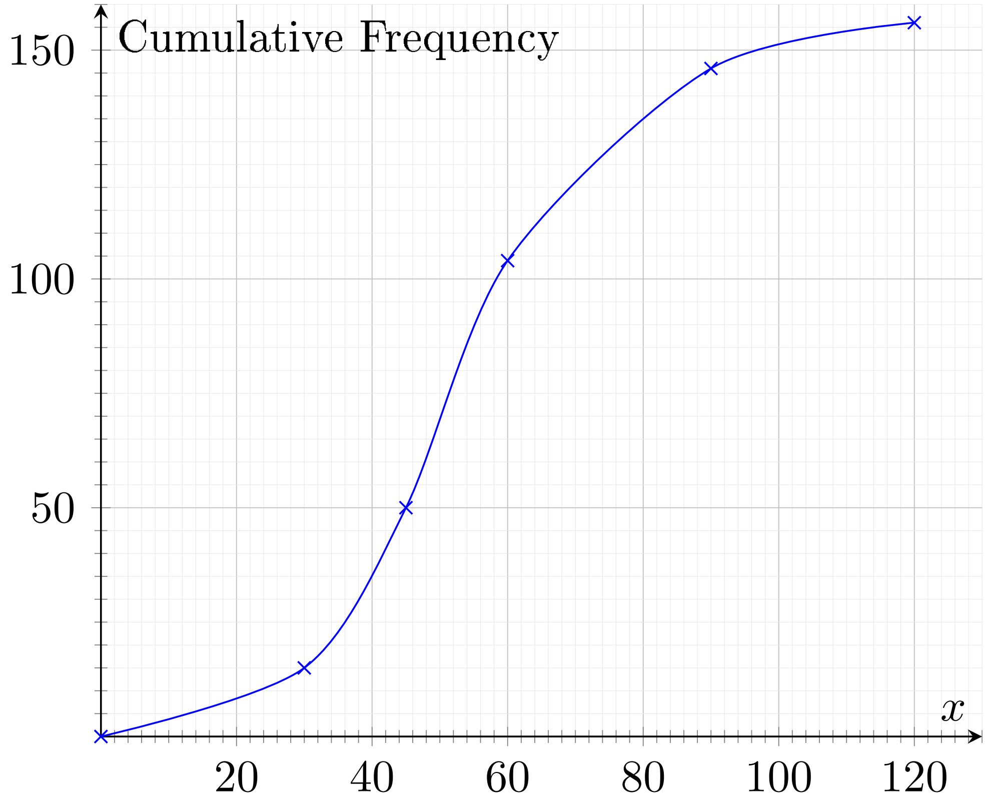


Figure 8.2: A cumulative frequency curve of the information provided above

(c) A cumulative frequency curve can be used to derive further statistics.

(i) We want to find the value for x such that $F(x) = 35$. Reading this off the graph, we see that

$$F(40) = 35 \tag{8.1.26}$$

(ii) Recall from Def [2.3.5] the median of a distribution. Thus, we want to find Q_2 such that

$$\begin{aligned} F(Q_2) &= 50\% \times 156 && (8.1.27)\\ &= 0.5 \times 156 && (8.1.28)\\ &= 78 && (8.1.29)\\ \Rightarrow Q_2 &= 52 && (8.1.30) \end{aligned}$$

(iii) Recall from Def [2.3.5] the lower and upper quartiles of a distribution. Thus, we want to find the values of Q_1 and Q_3

$$\begin{aligned} F(Q_1) &= 25\% \times 156 && (8.1.31)\\ F(Q_1) &= 0.25 \times 156 && (8.1.32)\\ &= 42 && (8.1.33)\\ \Rightarrow Q_1 &= 41 && (8.1.34)\\ F(Q_3) &= 75\% \times 150 && (8.1.35)\\ F(Q_3) &= 0.75 \times 150 && (8.1.36)\\ &= 117 && (8.1.37)\\ \Rightarrow Q_3 &= 67 && (8.1.38)\\ IQR &= Q_3 - Q_1 && (8.1.39)\\ &= 67 - 42 && (8.1.40)\\ &= 25 && (8.1.41) \end{aligned}$$

(d) We can represent this information on a box and whisker as follows

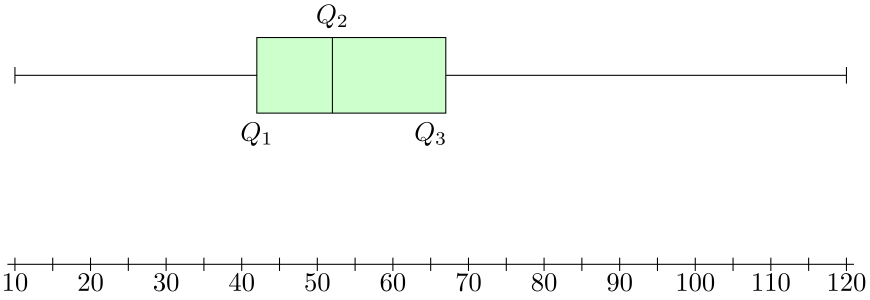


Figure 8.3: A Box and Whisker plot of the data given

(e) To calculate the mean, we take the midpoint of the classes in Eq [2.3.2],

$$\begin{aligned} \bar{x} &= \frac{\sum f_j x_j}{\sum f_j} && (8.1.42) \\ &= \frac{15(15) + 37.5(35) + 52.5(54) + 75(42) + 105(10)}{15 + 35 + 54 + 42 + 10} && (8.1.43) \\ &= 54.952 \approx 55 \text{ people} && (8.1.44) \end{aligned}$$

(f) From Def [2.3.6], we see that since,

$$\begin{aligned} Q_3 - Q_2 &> Q_2 - Q_1 && (8.1.45) \\ 67 - 52 &> 52 - 42 && (8.1.46) \\ 15 &> 10 && (8.1.47) \end{aligned}$$

And so, since $Q_3 - Q_2 > Q_2 - Q_1$, we conclude that the data is positively skewed.

8.2 Module 2: Managing Uncertainty

3. (a) We are given that $|R| = 8$ and $|G| = 12$

(i) Let G be the event that the candle is g and S be the sample space. Thus, we want to find $P(G)$. We can use Eq [3.1.3],

$$\begin{aligned} P(G) &= \frac{|G|}{|S|} && (8.2.1) \\ &= \frac{12}{12+8} = 0.6 && (8.2.2) \end{aligned}$$

(ii) We are given that three items are chosen at random.

a) We want to list all the possible combinations of 3 items such that any one can be r or g. First, we should note how many possible combinations we are looking for to make sure we do not miss any. Since each draw can either be r or g, there are 2^3 possible elements in our possibility space. Now, we can proceed to list all the elements.
The most efficient way to accomplish this (although it might seem difficult the first time) is to add in binary, since we are only dealing with two possible values for each draw. Let us represent an element of the possibility space with a 3 digit binary number, where a 1 and 0 can be used to represent R and G respectively.
eg. the digit 100 means RGG.

Starting with 000 we will add 1 to the right most digit until we have 8 elements which corresponds to the binary number 7 (since we started counting from 0).

$$\begin{aligned} 000 &\rightarrow GGG \quad (8.2.3)\\ 001 &\rightarrow GGR \quad (8.2.4)\\ 010 &\rightarrow GRG \quad (8.2.5)\\ 011 &\rightarrow GRR \quad (8.2.6)\\ 100 &\rightarrow RGG \quad (8.2.7)\\ 101 &\rightarrow RGR \quad (8.2.8)\\ 110 &\rightarrow RRG \quad (8.2.9)\\ 111 &\rightarrow RRR \quad (8.2.10) \end{aligned}$$

b) We can look at our possibility space to determine probabilities. We see that it is possible to get exactly 2 r items if the event GRR, RRG or RGR occurs. Thus, we must calculate $P((GRR) \cup (RGR) \cup (RRG))$,

$$\begin{aligned} P((GRR) \cup (RGR) \cup (RRG)) &= P(GRR) + P(RGR) + P(RRG) \quad (8.2.11)\\ &= \frac{12}{20}.\frac{8}{19}.\frac{7}{18} + \frac{8}{20}.\frac{12}{19}.\frac{7}{18} + \frac{8}{20}.\frac{7}{19}.\frac{12}{18} \quad (8.2.12)\\ &= \frac{28}{95} = 0.295 \quad (8.2.13) \end{aligned}$$

Alternatively, we can use a combinatoric approach as a cross check. The probability that this occurs corresponds to

$$\begin{aligned} P &= \frac{\text{\# ways to CHOOSE 2 r} \times \text{\# ways to CHOOSE 1 g}}{\text{\#ways to CHOOSE 3 items}} \quad (8.2.14)\\ &= \frac{{}^{8}C_2 \times {}^{12} C_1}{{}^{20}C_3} \quad (8.2.15)\\ &= \frac{28}{95} = 0.295 \quad (8.2.16) \end{aligned}$$

(b) We are given $P(A) = 0.5$, $P(B) = 0.40$ and $P(A \cap B) = 0.12$.

(i) Recall from Def [3.1.5] that two events A and B are mutually exclusive if $P(A \cap B) = 0$. But we are given that $P(A \cap B) = 0.12 \neq 0$. Thus, the events A and B are not mutually exclusive.

(ii) Recall from Def [3.1.5] that events A and B are independent if and only if

$$P(A \cap B) = P(A) \times P(B) \quad (8.2.17)$$

Using the data given in the problem, we can proceed as follows,

$$\begin{aligned} P(A) \times P(B) &= 0.5 \times 0.4 = 0.02 \quad (8.2.18)\\ \therefore P(A) \times P(B) &\neq P(A \cap B) \quad (8.2.19) \end{aligned}$$

Thus, the events A and B are not independent.

(c) Let F be the event an individual is F.

Let S be the event an individual is S.

Thus, we are given that

$$\begin{aligned} P(F) &= 0.45 \quad (8.2.20)\\ P(S) &= 0.25 \quad (8.2.21)\\ P(F \cap S) &= 0.12 \quad (8.2.22) \end{aligned}$$

Using this information, we can determine the probability of many events.

(i) We want to find $P(\overline{F \cup S})$. We can use a few identities to reduce this into something we have given values for. We apply the complement according to Eq [3.1.4] and then Proposition [3.1.5] to get

$$\begin{aligned} P(\overline{F \cup S}) &= 1 - P(F \cup S) && (8.2.23) \\ &= 1 - (P(F) + P(S) - P(F \cap S)) && (8.2.24) \\ &= 1 - (0.45 + 0.25 - 0.12) = 0.42 && (8.2.25) \end{aligned}$$

(ii) We want to find $P(F \cap \bar{S})$. To do this, we can use a series of identities as in part (i).

$$\begin{aligned} P(F \cap \bar{S}) &= P(F) - P(F \cap S) && (8.2.26) \\ &= 0.45 - 0.12 = 0.33 && (8.2.27) \end{aligned}$$

(iii) We want to find $P(F|S)$. To do this, we can use conditional probability, Def [3.1.6], which states

$$\begin{aligned} P(F|S) &= \frac{P(F \cap S)}{P(S)} && (8.2.28) \\ &= \frac{0.12}{0.25} = 0.48 && (8.2.29) \end{aligned}$$

(d) We are given a probability distribution function for a discrete random variable X.

(i) X is a random variable since it satisfies our conditions in Proposition [3.2.1]

$$\forall x, \quad 0 \leq P(X = x) \leq 1 \qquad (8.2.30)$$

$$\sum_{\forall x} P(X = x) = 1 \qquad (8.2.31)$$

(ii) We can determine $E[X]$ by Def [3.2.2]

$$\begin{aligned} E[X] &= \sum_{\forall x} xP(X = x) && (8.2.32) \\ &= 7(0.35) + 8(0.2) + 9(0.3) + 10(0.08) + 11(0.07) && (8.2.33) \\ &= 8.32 && (8.2.34) \end{aligned}$$

To find $\sigma(X)$, we can start by finding $Var(X)$ and then make use of the fact that $Var(X) = \sigma^2(X)$. From Eq [3.2.6] we have

$$Var(X) = E[X^2] - (E[X])^2 \qquad (8.2.35)$$

We compute the second moment of X by Def [3.2.3],

$$\begin{aligned} E[X^2] &= \sum_{\forall x} x^2 P(X = x) && (8.2.36) \\ &= 7^2(0.35) + 8^2(0.2) + 9^2(0.3) + 10^2(0.08) + 11^2(0.07) && (8.2.37) \\ &= 70.72 && (8.2.38) \end{aligned}$$

Putting this all together,

$$\begin{aligned} \sigma^2(X) &= E[X^2] - (E[X])^2 = 1.4976 && (8.2.39) \\ \sigma(X) &= 1.22 && (8.2.40) \end{aligned}$$

4. **(a)** We know from Note [3.3.2], that a random variable X is said to follow a binomial distribution if

(i) The experiment consist of a fixed number of trails n.

(ii) The trials are independent.

(iii) Each trail can be classified as a success or failure.

(iv) The probability of success, p, is constant

(b) We should check if each of the conditions mentioned in (a) are satisfied.

(i) Since there are three different colors, we cannot classify each event as a success or failure, so condition (iii) is not satisfied.

(ii) The fact that a disc is chosen without replacement tells us that the probability of success will change with time. So condition (iv) is clearly not satisfied.

(iii) The fact that there are 3 trials tell us condition (i) is satisfied. Condition (ii) and (iii) is clearly satisfied. Lastly, since the disc is replaced, the probability will remain constant with time, and hence condition (iv) is also satisfied.

(c) We are given that $p = 0.4$.

(i) Let X be the discrete random variable representing the number of success in 40 trials. We see that X satisfies the conditions to be modeled by a binomial distribution with parameters $n = 40$ and $p = 0.4$, $X \sim Bin(40, 0.4)$. Consequently, from Def [3.3.1],

$$P(X = x) = \binom{40}{x} 0.4^x 0.6^{40-x} \quad n \in (0, 1, 2, ..., 40) \tag{8.2.41}$$

We can determine $E[X]$ from Eq [3.3.2]

$$E[X] = np = 40 \times 0.4 = 16 \tag{8.2.42}$$

(ii) We are given that $n = 7$ now. Let Y be the number of success in $n = 7$ trials, $Y \sim Bin(7, 0.4)$. From Def [3.3.1], we have

$$P(Y = y) = \binom{7}{y} 0.4^7 0.6^{7-y} \quad \forall y \in \{0, ..., 7\} \tag{8.2.43}$$

a) We want to find $P(Y = 0)$. Using the definition above,

$$P(Y = 0) = \binom{7}{0} 0.4^0 0.6^7 \tag{8.2.44}$$
$$= 0.0279 \tag{8.2.45}$$

b) We want to find $P(Y = 3)$. Again, using the definition above,

$$P(Y = 3) = \binom{7}{3} 0.4^3 0.6^{37} \tag{8.2.46}$$
$$= 0.2903 \tag{8.2.47}$$

c) We want to determine $P(Y \geq 1)$. The long way (but still correct) would be to do this

$$P(Y \geq 1) = P(Y = 1) + P(Y = 2) + ... + P(Y = 7) \tag{8.2.48}$$

However, it is much quicker to recognize the following,

$$P(Y \geq 1) = 1 - P(Y = 0) \tag{8.2.49}$$
$$= 1 - \binom{7}{0} 0.4^0 0.6^7 \tag{8.2.50}$$
$$= 1 - 0.0279 \tag{8.2.51}$$
$$= 0.9721 \tag{8.2.52}$$

(d) Let M be the continuous random variable representing the mass of the bread, $X \sim N(480, 20^2)$ We want to determine $P(465 < M < 500)$. We can proceed by standardizing X,

$$P(465 < M < 500) = P\left(\frac{465 - 480}{20} < \frac{X - \mu}{\sigma} < \frac{500 - 480}{20}\right) \tag{8.2.53}$$
$$= P(-0.75 < Z < 1) \tag{8.2.54}$$
$$= \Phi(1) - \Phi(-0.75) \tag{8.2.55}$$
$$= \Phi(1) - (1 - \Phi(0.75)) \tag{8.2.56}$$
$$= 0.8431 - (1 - 0.7734) = 0.6165 \tag{8.2.57}$$

8.3 Module 3: Analyzing and Interpreting Data

5. **(a)** We are given 6 data points drawn at random a normal distribution.

(i) Recall from Def [4.1.1] that whenever we have a sample, we can use it to compute unbiased estimates of population parameters.

a) Recall from Def [4.1.2] that the unbiased estimate for the mean is given by

$$\hat{\mu} = \frac{\sum x}{n} \tag{8.3.1}$$
$$= \frac{40 + 44 + 50 + 30 + 46 + 48}{6} \tag{8.3.2}$$
$$= 43 \tag{8.3.3}$$

b) Recall from Def [4.1.3] that the unbiased estimate for the variance is given by

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n-1}\left(\sum x^2 - \frac{(\sum x)^2}{n}\right) && (8.3.4)\\ &= \frac{1}{6-1}\left(11356 - \frac{258^2}{6}\right) && (8.3.5)\\ &= 52.4 && (8.3.6)\\ \sqrt{\hat{\sigma}^2} &= 7.239 && (8.3.7)\end{aligned}$$

(ii) We are given that $\alpha = 5\%$ and that

$$\begin{aligned}H_0 &: \mu = 45 && (8.3.8)\\ H_1 &: \mu < 45 && (8.3.9)\end{aligned}$$

a) Recall from Section [4.2.3] that since σ^2 is unknown and $n < 30$, the critical region for the given hypothesis test is given by:

$$\begin{aligned}t_{\text{calc}} &< -t_{\alpha}^{(n-1)} && (8.3.10)\\ &< -t_{0.05}^{(6-1)} && (8.3.11)\\ &< -2.015 && (8.3.12)\end{aligned}$$

b) Recall from Section [4.2.2] that we can determine the appropriate test statistic by,

$$\begin{aligned}t_{\text{calc}} &= \frac{\bar{x} - \mu}{\sqrt{\hat{\sigma}^2/n}} && (8.3.13)\\ &= \frac{43 - 45}{\sqrt{\frac{52.4}{6}}} && (8.3.14)\\ &= -5.910 && (8.3.15)\end{aligned}$$

c) Since $t_{\text{calc}} = -5.910 < -2.015$, out test static lies within the critical region and we must reject the null hypothesis in favor of the alternative hypothesis.

(b) (i) Let (a, b) represent the value of dice A and B respectively when rolled together. We know that there are $6 \times 6 = 36$ possible outcomes for this process. The combinations that give us a sum $a + b = 9$ include $(6, 3)$, $(5, 4)$,$(4, 5)$, and $(3, 6)$.
Let S represent the sample space of all possible outcomes of the die. Let E be the event that the sum of the faces of the dice is 9. The by Eq. [3.1.3],

$$\begin{aligned}P(E) &= \frac{|A|}{|S|} && (8.3.16)\\ &= \frac{4}{36} && (8.3.17)\\ &= \frac{1}{12} && (8.3.18)\end{aligned}$$

(ii) We can state our null and alternative hypothesis as

$$\begin{aligned}H_0 &: p = \frac{1}{12} && (8.3.19)\\ H_1 &: p < \frac{1}{12} && (8.3.20)\end{aligned}$$

(iii) For the given test, we reject H_0 if

$$\begin{aligned}z_{\text{calc}} &< -z_{\alpha} && (8.3.21)\\ &< -z_{0.05} && (8.3.22)\\ &< -1.645 && (8.3.23)\end{aligned}$$

(iv) We calculate the test statistic as follows:

$$z_{\text{calc}} = \frac{p_s - p + \frac{1}{2n}}{\sqrt{\frac{p(1-p)}{n}}} \tag{8.3.24}$$

$$= \frac{\frac{12}{180} - \frac{1}{12} + \frac{1}{2\times180}}{\sqrt{\frac{\frac{1}{12}(1-\frac{1}{12})}{180}}} \tag{8.3.25}$$

$$= -0.052 \tag{8.3.26}$$

(v) As our test statistic $z_{\text{calc}} = -0.052 > -1.645$, it does not lie within in the critical region. Hence we fail to reject the null hypothesis at the 5% level of significance.

6. We are given 9 pairs of (x, y) coordinates.

(a) We can represent the information on a scatter plot.

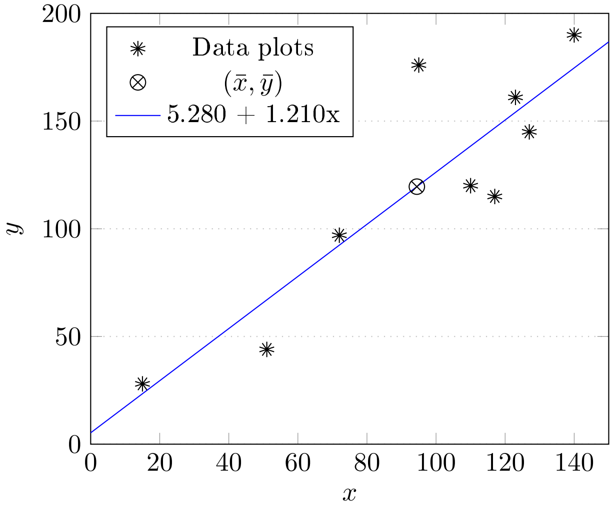


Figure 8.4: A scatter diagram of the data given for 9 data points.

(b) Recall from Eq [4.4.1] that we can calculate the product moment correlation coefficient as

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}} \tag{8.3.27}$$

$$= \frac{9(117841) - (850)(1076)}{\sqrt{9(93682) - 850^2}\sqrt{9(153776) - 1076^2}} \tag{8.3.28}$$

$$= 0.884 \tag{8.3.29}$$

From Note [4.4.2], we see that this value of r can be interpreted as the data having a high positive linear correlation.

(c) **(i)** Recall from Def [4.4.2] that the equation of regression the line y on x is given by

$$y = a + bx \tag{8.3.30}$$

where b can be calculated by

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} \tag{8.3.31}$$

$$= \frac{9(117841) - (850)(1076)}{9(93682) - (850)^2} \tag{8.3.32}$$

$$= 1.210 \tag{8.3.33}$$

and a by

$$\begin{aligned} a &= \bar{y} - b\bar{x} && (8.3.34)\\ &= \frac{\sum y}{n} - 1.210\left(\frac{\sum x}{n}\right) && (8.3.35)\\ &= \frac{1076}{9} - 1.20\left(\frac{850}{9}\right) && (8.3.36)\\ &= 5.280 && (8.3.37)\end{aligned}$$

So the regression line of y on x takes the form $y = 5.280 + 1.210x$.

(d) (i) The coefficient b tells us that for every 1 km traveled a time period of 1 hour elapses.

(ii) The coefficient a tells us that 5.280 hours elapses before any distance has been traveled.

(e) (i) Now that we have established a linear relation between the data points, we can make a guess as to what the corresponding y value is for a given x. Suppose $x = 145$, then

$$\begin{aligned} y &= 5.280 + 1.210(145) && (8.3.38)\\ &= 180.73 && (8.3.39)\end{aligned}$$

(ii) In general, extrapolation is not allowed in regression. Since $x = 145$ and this lies outside the range of x values we used to determine the equation of regression, we can conclude that this is not a reliable value.

Chapter 9

2009

9.1 Module 1: Collecting and Describing Data

1. **(a)** For this question we number the variables given from A to G.

   (i) The qualitative variables are A, E, indicated by the key words 'place' and 'type'.

   (ii) The quantitative variables are B, C, D, F and G. indicated by the word 'number' and also the fact that some are measurable quantities.

   (iii) The discrete variables are B, C, indicated by the key word 'number'. Although F has this word we must be careful since the quantity it is describing has a continuous value.

   (iv) One variable that is continuous is F, because this can take any value within a given range.

   (b) **(i)** To sample 10 persons from 65 using the method of random numbers, we first assign a unique number from $1 - 65$ to every person in our population. Then, using the table of random numbers, we start at any point and write out the numbers in the table in a certain order, going horizontally for example. Lastly, we select the first 10 numbers that are within the range of $1 - 65$. These numbers correspond to the 10 persons randomly selected.

   (ii) We omit this question.

   (c) **(i)** **(a)** From Def [2.1.5], the population refers to all the elements or individuals that meet the selection criteria for a group to be studied, and from which a **sample** is usually chosen from to be examined in detail.

   (b) From Def [2.1.7], a sample is a subset of the population which is studied in detail so as to find numerical data, central tendencies or any other patterns. These statistics can usually be extended to the population but depend on the degree of representation of the sample to the population.

   (ii) The population in this survey refers to all the students in the school.

   (iii) Stratified random sampling was employed as the proportions in the sample reflected the proportions in the survey.

   (iv) Let x be the number of fourth year students in the sample. This must ensure that the proportions are the same so,

$$\frac{x}{50} = \frac{90}{500} \tag{9.1.1}$$

$$x = 9 \tag{9.1.2}$$

   (v) To determine the angle for each response category, we simply look at the ratio of each category in the sample and determine the same ratio in a circle.

$$360 \times \frac{10}{50} = 72.0^\circ \tag{9.1.3}$$

$$360 \times \frac{15}{50} = 108.0^\circ \tag{9.1.4}$$

$$360 \times \frac{8}{50} = 57.6^\circ \tag{9.1.5}$$

$$360 \times \frac{6}{50} = 43.2^\circ \tag{9.1.6}$$

$$360 \times \frac{11}{50} = 79.2^\circ \tag{9.1.7}$$

2. (a) (i) Histograms. This is usually used for continuous data as you can bin the data as necessary

(ii) Pie chart. Since it is only 4 persons, this will easily represent the data and allow comparison.

(iii) Bar chart. Better suited for comparison of 6 persons.

(b) We are provided with responses to a survey.

(i) A frequency table to show the data is given below.

Response	M	F	B	S
Frequency	9	4	6	3

Table 9.1: A frequency distribution table for the responses of the survey

(ii) We can draw a bar chart to show the information.

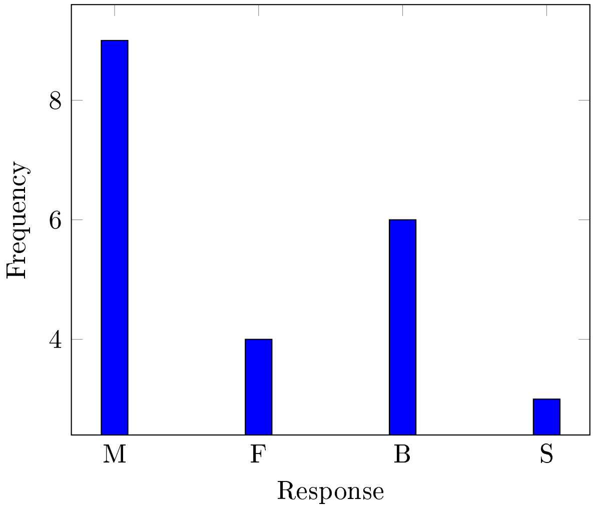


Figure 9.1: A bar chart representing the results of the survey

(c) (i) A stratified random sample ensures that the proportions of the strata in the sample are representativeness of the proportions of the strata in the population. This is beneficial for the given situation as in such cases, it is not the total number of employees, but the proportion of specialized labor that is important for the effective running of a new establishment.

Simple random samples pay no attention to proportions and hence it can result in a disproportionate number of people in one department.

(ii) Using stratified random sampling, we first determine the number of persons in each stratum to maintain the ratios. The total, T, is

$$T = 12 + 24 + 8 + 6 \tag{9.1.8}$$

$$= 50 \tag{9.1.9}$$

If x_c, x_a, x_p and x_s are the number of people from each stratum in our department, we keep the ratios constant to obtain:

$$\frac{x_c}{25} = \frac{12}{50} \tag{9.1.10}$$
$$\Rightarrow x_c = 6 \tag{9.1.11}$$
$$\frac{x_a}{25} = \frac{24}{50} \tag{9.1.12}$$
$$\Rightarrow x_a = 12 \tag{9.1.13}$$
$$\frac{x_p}{25} = \frac{8}{50} \tag{9.1.14}$$
$$\Rightarrow x_p = 4 \tag{9.1.15}$$
$$\frac{x_s}{25} = \frac{6}{50} \tag{9.1.16}$$
$$\Rightarrow x_s = 3 \tag{9.1.17}$$

The calculation was presented for clarity. However, it would have been quicker to notice that the sample size was half the population size, so we simply needed to halve the size of all our strata to keep the proportions the same.

(d) We are given ordered data.

(i) **a)** To determine the mode, we should summarize the data given in a frequency table

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

Table 9.2: A frequency distribution table for the data given

So the mode is 7.
We note that for practicality, the table is not necessary. By inspection one can determine that the most frequent numbers were either 6 or 7 and just pay attention to those two.

b) Recall from Def [2.3.5] the median, Q_2, of the data, is given by

$$Q_2 = \frac{n+1}{2} \text{ th term} \tag{9.1.18}$$
$$= \frac{25+1}{2} \text{ th term} \tag{9.1.19}$$
$$= 13 \text{ th term} \tag{9.1.20}$$
$$= 7 \tag{9.1.21}$$

(ii) **a)** Recall from Eq [2.3.2], that we can compute the mean when given grouped data as follows,

$$\bar{x} = \frac{\sum f_j x_j}{\sum f_j} \tag{9.1.22}$$
$$= \frac{3(1)+4(4)+5(2)+6(6)+7(7)+8(4)+9(1)+10(2)+...+13(1)}{1+1+2+6+7+4+1+2+...+1} \tag{9.1.23}$$
$$= \frac{188}{25} \tag{9.1.24}$$
$$= 7.52 \tag{9.1.25}$$

b) Recall from Def [2.3.5] the interquartile range. First we compute the lower quartile, Q_1, and the upper quartile, Q_3,

$$\begin{aligned}
Q_1 &= \frac{n+1}{4} \text{ th term} && (9.1.26)\\
&= 6.5 \text{ th term} && (9.1.27)\\
&= 6 + 0.5(6-6) && (9.1.28)\\
&= 6 && (9.1.29)\\
Q_3 &= \frac{3(n+1)}{4} \text{ th term} && (9.1.30)\\
&= 19.5 \text{ th term} && (9.1.31)\\
&= 8 + 0.5(8-8) && (9.1.32)\\
&= 8 && (9.1.33)\\
IQR &= Q_3 - Q_1 && (9.1.34)\\
&= 8-6 && (9.1.35)\\
&= 2 && (9.1.36)
\end{aligned}$$

(iii) We can represent the information on a box and whisker plot as follows:

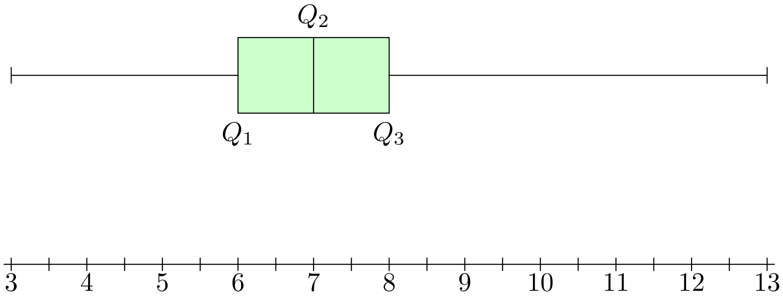


Figure 9.2: A Box and Whisker plot of the data given

(iv) From Def [2.3.6], we see that the conditions are satisfied to describe the data as a normal i.e.

$$\begin{aligned}
\text{Mode} &= \text{Median} && (9.1.37)\\
Q_2 - Q_1 &= Q_3 - Q_2 && (9.1.38)\\
7-6 &= 8-7 && (9.1.39)
\end{aligned}$$

9.2 Module 2: Managing Uncertainty

3. (a) Given two events A and B, such that $P(A) = 0.3$, $P(B) = 0.4$ and $P(A \cap B) = 0.2$, we want to calculate:

(i) We can determine $P(A \cup B)$ by Eq [3.1.5]

$$\begin{aligned}
P(A \cup B) &= P(A) + P(B) - P(A \cap B) && (9.2.1)\\
&= 0.3 + 0.4 - 0.2 && (9.2.2)\\
&= 0.5 && (9.2.3)
\end{aligned}$$

(ii) We can determine $P(A|B)$ with conditional probability according to Def [3.1.6],

$$\begin{aligned}
P(A|B) &= \frac{P(A \cap B)}{P(B)} && (9.2.4)\\
&= \frac{0.2}{0.4} = 0.5 && (9.2.5)
\end{aligned}$$

(b) We are given that G and H are independent events. $P(G) = 0.7$ and $P(H) = 0.4$. Before we continue, we should recall from Def [3.1.5] that the independence of G and H imply $P(G \cap H) = P(G) \times P(H)$.

(i) We want to calculate $P(G \cap H)$
As noted above, the fact that G and H are independent tells us that

$$\begin{align} P(G \cap H) &= P(G) \times P(H) \tag{9.2.6} \\ &= 0.7 \times 0.4 = 0.28 \tag{9.2.7} \end{align}$$

(ii) We want to calculate $P(G \cup H)$.
Now, the information we have about H and G are $P(G)$, $P(H)$ and $P(G \cap H)$. We can use Eq [3.1.5] that relates these pieces of information to the one required,

$$\begin{align} P(G \cup H) &= P(G) + P(H) - P(G \cap H) \tag{9.2.8} \\ &= 0.7 + 0.4 - 0.28 \tag{9.2.9} \\ &= 0.82 \tag{9.2.10} \end{align}$$

(c) Let B be the event that an item was from B.
Let A be the event that an item was from machine A.
Let D be the event that an item is defective.

(i) We want to show this information on a well-labeled tree diagram.

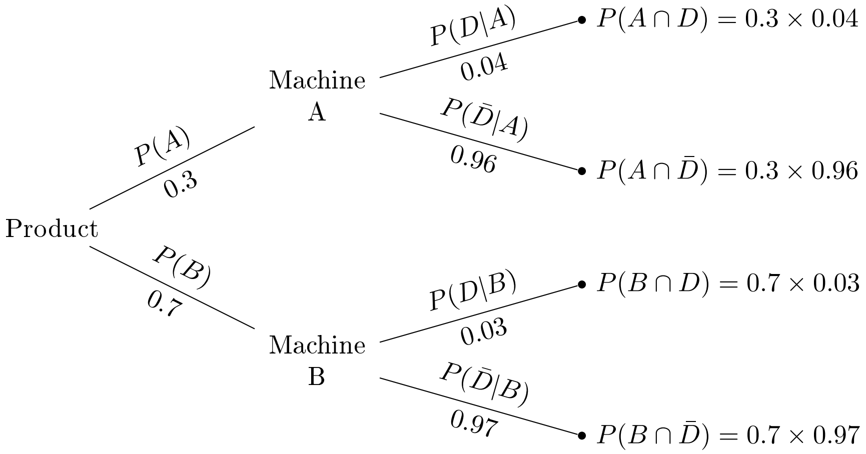

Product
P(A)
0.3
P(B)
0.7
Machine A
Machine B
P(D\A)
0.04
$P(\bar{D}|A)$
0.96
P(D\B)
0.03
$P(\bar{D}|B)$
0.97
$P(A \cap D) = 0.3 \times 0.04$
$P(A \cap \bar{D}) = 0.3 \times 0.96$
$P(B \cap D) = 0.7 \times 0.03$
$P(B \cap \bar{D}) = 0.7 \times 0.97$

Figure 9.3: A probability tree diagram representing the probabilities related to output

(ii) We want to find $P(D)$. A item can be defective if it came from either A or B. Thus,

$$\begin{align} P(D) &= P(D_A) + P(D_B) \tag{9.2.11} \\ &= P(A \cap D) + P(B \cap D) \tag{9.2.12} \end{align}$$

Now, to find those probabilities we can conditional probability, Def [3.1.6], $P(A|B) = \frac{P(A \cap B)}{P(B)}$

$$\begin{align} P(A \cap D) &= P(D|A) \times P(A) \tag{9.2.13} \\ &= 0.04 \times 0.3 \tag{9.2.14} \\ &= 0.012 \tag{9.2.15} \\ P(B \cap D) &= P(D|B) \times P(B) \tag{9.2.16} \\ &= 0.03 \times 0.7 \tag{9.2.17} \\ &= 0.021 \tag{9.2.18} \\ \therefore P(D) &= 0.012 + 0.021 \tag{9.2.19} \\ &= 0.033 \tag{9.2.20} \end{align}$$

(iii) We want to find $P(A|D)$. Applying Def [3.1.6] for conditional probability directly to this,

$$\begin{align} P(A|D) &= \frac{P(D \cap A)}{P(D)} \tag{9.2.21} \\ &= \frac{0.012}{0.033} = \frac{12}{33} \tag{9.2.22} \end{align}$$

(d) Let X be the discrete random variable representing the amount of hours worked overtime. We should represent this information in a clearer way before we proceed with the problem.

X	4	6	8	10
$P(X = x)$	0.4	0.3	k	0.1

Table 9.3: A table representing the probability mass function of X

(i) We want to calculate the value of k. Since X is a discrete random variable, we know from Proposition [3.2.1] that

$$\sum_{\forall x} P(X = x) = 1 \tag{9.2.23}$$

$$\therefore 0.4 + 0.3 + k + 0.1 = 1 \tag{9.2.24}$$

$$\Rightarrow k = 0.2 \text{ hrs} \tag{9.2.25}$$

(ii) We want to find $E[X]$. We can solve for this by Def [3.2.2]

$$E[X] = \sum_{\forall x} xP(X = x) \tag{9.2.26}$$

$$= 4(0.4) + 6(0.3) + 8(0.2) + 10(0.1) \tag{9.2.27}$$

$$= 6 \text{ hrs} \tag{9.2.28}$$

(iii) We want to find $P(X < 8)$. We can easily see this from our table,

$$P(X < 8) = P(X = 4) + P(X = 6) \tag{9.2.29}$$

$$= 0.4 + 0.3 \tag{9.2.30}$$

$$= 0.7 \tag{9.2.31}$$

4. (a) Before we continue, we should recall the Note [3.3.2] for the conditions necessary for a distribution to be modeled by a binomial distribution.

(i) The experiment consist of a fixed number of trails n.

(ii) The trials are independent.

(iii) Each trail can be classified as a success or failure.

(iv) The probability of success, p, is constant

Now we can easily check if a distribution can be modeled by a binomial distribution.

(i) All conditions are satisfied, so X can be modeled by a binomial distribution.

(ii) We do not have a fixed number of trails. Since (i) is not satisfied, we cannot model X with a binomial distribution.

(iii) All conditions are satisfied, thus we can model X with a binomial distribution.

(iv) In this case, the probability of drawing a red changes since the marbles are not being replaced. So the probability of success is not constant, (iv), and we cannot model X with a binomial distribution.

(b) We are given that $p = 0.3$ and $n = 15$.

(i) Let X be the discrete random variable representing the number of people who do not return in a group of 15. We see that X satisfies the conditions in Note [3.3.2] to be modeled by a binomial distribution with parameters $n = 15$ and $p = 0.3$, $X \sim Bin(15, 0.3)$. We want to find $E[X]$. From Eq [3.3.2], we have

$$E[X] = np \tag{9.2.32}$$

$$= 0.3 \times 15 \tag{9.2.33}$$

$$= 4.5 \approx 5 \tag{9.2.34}$$

(ii) We can use the distribution to determine some further probabilities.

a) We want to find $P(X = 4)$. Since X follows a binomial distribution, we know from Def [3.3.1]

$$P(X = x) = \binom{15}{x} 0.3^x 0.7^{15-x} \quad x \in (0, 1, ..., 15) \tag{9.2.35}$$

$$P(X = 4) = \binom{15}{4} 0.3^4 0.7^{11} \tag{9.2.36}$$

$$= 0.219 \tag{9.2.37}$$

b) We want to find $P(X \leq 2)$,

$$P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2) \tag{9.2.38}$$

$$= \binom{15}{0} 0.3^0 0.7^{15} + \binom{15}{1} 0.3^1 0.7^{14} + \binom{15}{2} 0.3^2 0.7^{13} \tag{9.2.39}$$

$$= 0.127 \tag{9.2.40}$$

(c) We are given that $X \sim N(12, 4)$. We want to find the constant c such that $P(X < c) = 0.9332$. We can proceed by standardizing c and then determining what value it should correspond to.

$$P(X < c) = P\left(\frac{X - \mu}{\sigma} < \frac{c - 12}{2}\right) \tag{9.2.41}$$

$$= P\left(Z < \frac{c - 12}{2}\right) \tag{9.2.42}$$

$$= 0.9332 \tag{9.2.43}$$

$$\therefore \quad \Phi\left(\frac{c - 12}{2}\right) = 0.9332 \tag{9.2.44}$$

$$\frac{c - 12}{2} = 1.5 \tag{9.2.45}$$

$$c = 15 \tag{9.2.46}$$

(d) Let X be the continuous random variable representing the length of a metal curtain rod, $X \sim N(60, 4^2)$.

(i) We want to find $P(X > 65)$. We can proceed by standardizing.

$$P(X > 65) = P\left(\frac{X - \mu}{\sigma} > \frac{65 - 60}{4}\right) \tag{9.2.47}$$

$$= P(Z > 1.25) \tag{9.2.48}$$

$$= 1 - P(Z < 1.25) \tag{9.2.49}$$

$$= 1 - \Phi(1.25) \tag{9.2.50}$$

$$= 1 - 0.8944 \tag{9.2.51}$$

$$= 0.1056 \tag{9.2.52}$$

(ii) We want to find $P(X < 54)$. We can proceed by standardizing,

$$P(X < 54) = P\left(\frac{X - \mu}{\sigma} < \frac{54 - 60}{4}\right) \tag{9.2.53}$$

$$= P(Z < -1.5) \tag{9.2.54}$$

$$= 1 - \Phi(1.5) \tag{9.2.55}$$

$$= 0.0668 \tag{9.2.56}$$

9.3 Module 3: Analyzing and Interpreting Data

5. **(a)** Recall from Def [4.1.5], that a 95% confidence interval for μ means that we will find with probability 0.95 a confidence interval in which the actual value of the parameter μ will lie within.

(b) We are given that $n = 45$, $\bar{x} = 950$ and $s = 135$.

(i) We want to find $\sqrt{\hat{\sigma^2}}$ that corresponds to the given s. Recall from Def [4.1.3] the unbiased estimate for variance,

$$\hat{\sigma^2} = \frac{n}{n-1}s^2 \tag{9.3.1}$$

$$\sqrt{\hat{\sigma^2}} = \sqrt{\frac{n}{n-1}}s \tag{9.3.2}$$

$$= \sqrt{\frac{45}{44}} \times 135 = 136.5 \tag{9.3.3}$$

(ii) Since $n \geq 30$ and σ^2 is unknown, we see from Def [4.1.6] our 95% confidence interval will take the form

$$\bar{x} \pm Z_{\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{n}} \tag{9.3.4}$$

$$\bar{x} \pm Z_{2.5}\sqrt{\frac{\hat{\sigma}^2}{n}} \tag{9.3.5}$$

$$950 \pm 1.96\sqrt{\frac{136.5^2}{45}} \tag{9.3.6}$$

$$\Rightarrow (910, 990) \tag{9.3.7}$$

(iii) **a)** To determine this, we should look at our general expression for a confidence interval: $\bar{x} \pm Z_{\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{n}}$. From this, we see that the width of the interval takes the form $2Z_{\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{n}}$. Thus, to decrease the width we can increase n or decrease $Z_{\alpha/2}$. Further, decreasing $Z_{\alpha/2}$ means we should decrease our confidence level (you should convince yourself that this is true).

b) Increasing n is better as it serves to lower the variance of the sampling distribution while allowing for a conservative confidence level to be maintained. To elaborate, larger samples result in smaller standard errors and hence distributions that are more clustered around the population mean. Alternatively, decreasing the confidence level is less favorable since despite narrowing the confidence interval, we become less confident that the calculated interval captures the population parameter.

(iv) Looking at our general expression for confidence intervals in Eq [4.1.6], we see that the width can be expressed as $2Z_{\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{n}}$. Therefore we want to find the value of n such that $2Z_{\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{n}} < 40$. For a 99% confidence interval, $Z_{1/2}$ corresponds to 2.58, thus

$$2Z_{\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{n}} < 40 \tag{9.3.8}$$

$$n > 136.5^2\left(\frac{2 \times 2.58}{40}\right)^2 \tag{9.3.9}$$

$$> 310.06 \tag{9.3.10}$$

$$\therefore n = 311 \text{ farmers} \tag{9.3.11}$$

(c) To answer this, we must look back to Def [4.1.5] of what it means to construct a confidence interval. Thus, with a 90% confidence interval we can expect to find, with probability 0.1, a confidence interval within which the actual value of μ does not lie within. So this means that we can expect $60 \times 0.1 = 6$ intervals to NOT contain μ.

(d) We are given $X \sim \text{Bin}(100, 0.05)$

(i) To determine the distribution of $\bar{X}$, we apply the Central Limit Theorem [4.1.1]. Since X

follows a Binomial distribution with parameters $n = 100$ and $p = 0.05$, we have,

$$\begin{aligned} E[X] &= \mu = np &(9.3.12)\\ &= 100 \times 0.05 = 5 &(9.3.13)\\ \sigma^2 &= np(1-p) &(9.3.14)\\ &= 100 \times 0.05 \times 0.95 = 4.75 &(9.3.15)\\ \Rightarrow \bar{X} &\sim \mathrm{N}\left(5, \frac{4.75}{100}\right) &(9.3.16)\\ \bar{X} &\sim \mathrm{N}(5, 0.0475) &(9.3.17) \end{aligned}$$

(ii) We can calculate $P(\bar{X} > 4.5)$ by standardizing as follows,

$$\begin{aligned} P(\bar{X} > 4.5) &= P\left(\frac{\bar{X}-\mu}{\sigma} > \frac{4.5-5}{\sqrt{0.0475}}\right) &(9.3.18)\\ &= P(Z > -2.294) &(9.3.19)\\ &= P(Z < 2.294) &(9.3.20)\\ &= \Phi(2.294) = 0.989 &(9.3.21) \end{aligned}$$

6. (a) Recall from Section [4.3] how to compute a χ^2 test at the 5% significance level.

(i) We can state the null and alternative hypothesis as:
H_0: There is no association between a child's early upbringing in a nursery or home and his behavior in the early period of primary school.
H_1: There is an association between a child's early upbringing in a nursery or home and his behavior in the early period of primary school.

(ii) To determine the missing values, we need to use the relation between E and the contingency table. We know that $E_{ij} = \frac{n_i \times n_j}{N}$, where E_{ij} represents the expected frequency of the value in the i^{th} row and j^{th} column, n_i represents the sum of values in the i^{th} row, n_j represents the sum of values in the j^{th} column, and N represents the total sample size.
An observed value of 40 corresponds to E_{31}, so

$$\begin{aligned} c &= E_{31} = \frac{170 \times 210}{400} &(9.3.22)\\ &= 89.25 &(9.3.23) \end{aligned}$$

This also gives us the information to compute g:

$$\begin{aligned} g &= \frac{O^2}{E} &(9.3.24)\\ &= \frac{40^2}{89.25} = 17.93 &(9.3.25) \end{aligned}$$

f and h follow the same procedure:

$$\begin{aligned} f &= \frac{O^2}{E} &(9.3.26)\\ &= \frac{100^2}{68.25} = 146.52 &(9.3.27)\\ h &= \frac{30^2}{61.75} = 14.57 &(9.3.28) \end{aligned}$$

To compute e, we rearrange the relationship

$$\begin{aligned} \frac{O^2}{E} &= 209.29 &(9.3.29)\\ \therefore \frac{130^2}{e} &= 209.29 &(9.3.30)\\ \Rightarrow e &= \frac{130^2}{209.29} = 80.75 &(9.3.31) \end{aligned}$$

We do the same for a

$$\frac{O^2}{E} = 93.33 \tag{9.3.32}$$

$$\therefore \frac{a^2}{52.50} = 93.33 \tag{9.3.33}$$

$$\Rightarrow a = \sqrt{52.50 \times 93.33} = 70.00 \tag{9.3.34}$$

To compute b, we simply use the fact that all sum of observed values is given.

$$\sum O = 400 \tag{9.3.35}$$

$$\Rightarrow 400 = 100 + 70 + 40 + 30 + b + 130 \tag{9.3.36}$$

$$b = 30 \tag{9.3.37}$$

Now that we know b, we can easily compute d using the same procedure we used for e,

$$\frac{30^2}{d} = 18.95 \tag{9.3.38}$$

$$\Rightarrow d = \frac{30^2}{18.95} = 47.49 \tag{9.3.39}$$

Correct to two decimal places, the value of χ^2 test statistic $\left(\sum \frac{O^2}{E}\right) - N = 100.59$

(iii) **a)** The number of degrees of freedom

We can calculate the number of degrees of freedom as follows

$$D.o.f = (r-1)(c-1) = (3-1)(2-1) = 2 \tag{9.3.40}$$

Where r and c represent the number of rows and columns, in the contingency table, respectively.

b) To determine the critical region, we must look in the table to see what value corresponds to 2 degrees of freedom and a level of significance of 5%. Thus, the critical region is $\chi_{test} > 5.991$.

(iv) Since $\chi_{test} = 100.59 > 5.991$, we fail to reject the null hypothesis. Therefore there is sufficient evidence at the 5% level of significance to reject the null hypothesis and conclude that there is an association between a child's early upbringing in a nursery or home and his behavior in the early period of primary school.

(b) We know from Note [4.4.3] that the line of regression always passes through the point $(\bar{x}, \bar{y})$. Thus, we have two points that the line passes through. This is enough information to determine the equation of the line. Using our formula from coordinate geometry for the equation of a line given two points,

$$\frac{y - y_1}{y_2 - y_1} = \frac{x - x_1}{x_2 - x_1} \tag{9.3.41}$$

$$\frac{y - 2}{8 - 2} = \frac{x - 0}{6 - 0} \tag{9.3.42}$$

$$y = x + 2 \tag{9.3.43}$$

Which is in the form $y = a + bx$ with $a = 2$ and $b = 1$.

Chapter 10

2010

10.1 Module 1: Collecting and Describing Data

1. (a) (i) C and E use words such as ‘types’ which indicate its qualitative nature

(ii) A, B, D and F can all be measured or counted.

(iii) A, B and F are all discrete, as the word ‘number’ indicates they can be counted.

(iv) D is continuous as ‘area’ can take any value within a given range.

(b) (i) Recall from Def [2.1.6] that the parameter is calculated from the population. We have two given options for parameters.

$$\mu = 72 \tag{10.1.1}$$

$$\sigma^2 = 15 \tag{10.1.2}$$

(ii) Recall from Def [2.1.8] that the statistic is calculated from sample data. We have one option for a statistic

$$\bar{x} = 67 \tag{10.1.3}$$

(c) (i) This is a sample since there was a selection from a larger population of people.

(ii) This is a population, indicated by the word ‘total’.

(d) (i) Cluster

(ii) Stratified

(iii) Systematic

(iv) Quota

(v) Simple random

(e) We are given a table showing data.

(i) The third class is $25 - 29$. So the boundaries are 24.5 and 29.5.

(ii) The fifth class is $40 - 49$. So the width is

$$\text{UCB} - \text{LCB} = 49.5 - 39.5 \tag{10.1.4}$$

$$= 10 \tag{10.1.5}$$

(iii) First we determine the frequency and then the class width

$$\text{Frequency density} = \frac{15}{24.5 - 19.5} \tag{10.1.6}$$

$$= 3 \tag{10.1.7}$$

(iv) From the table, we see that this should contain part of class 6 and all of class 7.

$$\text{Estimate} = \left(\frac{64.5 - 60}{64.5 - 49.5} \times 9\right) + 5 \tag{10.1.8}$$

$$= 7.7 \approx 8 \tag{10.1.9}$$

(v) The individual data is lost and further statistics have to be estimated.

2. **(a)** We are given data for the ages of 25 students.

(i) We want to illustrate this data in a stem and leaf diagram

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

Key: 1|3 means 13

Figure 10.1: A stem and leaf diagram of the data given.

(ii) One advantage is that the individual data points are not lost.

(b) **(i)** Recall from Def [2.3.5] that the median value is given by

$$
\begin{aligned}
Q_2 &= \frac{n+1}{2}^{\text{th}} \text{ term} && (10.1.10)\\
&= 13^{\text{th}}\text{term} && (10.1.11)\\
&= 24 && (10.1.12)
\end{aligned}
$$

(ii) The age with the higest frequency, the mode, is 19.

(c) Recall from Def [2.3.1], that we can find the mean by,

$$
\begin{aligned}
\bar{x} &= \frac{17+17+...+42+59}{25} && (10.1.13)\\
&= 27.08 \approx 27 \text{ years} && (10.1.14)
\end{aligned}
$$

(d) One disadvantage of the mean is that it can be affected by outliers.

(e) Recall from Def [2.3.3], an 8% trimmed mean means we discard $0.08 * 25 = 2$ data points from the upper and lower end of the data.
So the new mean $\bar{x}'$ is

$$
\begin{aligned}
\bar{x}' &= \frac{18+19+...+40+41}{25-4} && (10.1.15)\\
&= 25.81 \approx 26 \text{ years} && (10.1.16)
\end{aligned}
$$

(f) **(i)** Recall from Def [2.3.5] that we can determine the

$$
\begin{aligned}
Q_1 &= \frac{n+1}{4} \text{ th term} && (10.1.17)\\
&= 6.5 \text{ th term} && (10.1.18)\\
&= 19 + 0.5(21-19) && (10.1.19)\\
&= 20 && (10.1.20)\\
Q_3 &= \frac{3(n+1)}{4} \text{ th term} && (10.1.21)\\
&= 19.5 \text{ th term} && (10.1.22)\\
&= 32 + 0.5(32-32) && (10.1.23)\\
&= 32 && (10.1.24)
\end{aligned}
$$

(ii) Recall Def [2.3.5] that we can determine the interquartile, and hence semi-interquartile range, as follows

$$
\begin{aligned}
\frac{IQR}{2} &= \frac{Q_3 - Q_1}{2} && (10.1.25)\\
&= \frac{32-20}{2} && (10.1.26)\\
&= 6 && (10.1.27)
\end{aligned}
$$

(g) We have computed the mode, $\bar{x}$, Q_1, Q_2 and Q_3. Recalling from Def [2.3.5], we see that our distribution satisfies the requirements to be described with a positive skew, i.e

$$\begin{aligned} Q_3 - Q_2 &> Q_2 - Q_1 && (10.1.28)\\ 32 - 24 &> 24 - 20 && (10.1.29)\\ 8 &> 4 && (10.1.30) \end{aligned}$$

10.2 Module 2: Managing Uncertainty

3. (a) We are given that $P(A \cup B) = 0.86$, $P(B) = 0.72$ and $P(A) = 0.64$.

(i) We want to find $P(A \cap B)$. To do this we can use an identity from Eq [3.1.5]

$$\begin{aligned} P(A \cap B) &= P(A) + P(B) - P(A \cap B) && (10.2.1)\\ &= 0.72 + 0.64 - 0.86 && (10.2.2)\\ &= 0.5 && (10.2.3) \end{aligned}$$

(ii) We want to find $P(A|B)$. To do this, we can use Def [3.1.6] of conditional probability

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} && (10.2.4)\\ &= \frac{0.5}{0.72} = \frac{25}{36} && (10.2.5)\\ &= 0.694 && (10.2.6) \end{aligned}$$

(iii) We want to find $P(A' \cap B')$. Drawing a Venn diagram will make it clear that what we want to find is in fact,

$$\begin{aligned} P(A' \cap B') &= 1 - P(A \cup B) && (10.2.7)\\ &= 1 - 0.86 && (10.2.8)\\ &= 0.14 && (10.2.9) \end{aligned}$$

(b) Let H be the event that an individual practices hip-hop.

Let J be the event that an individual practices jazz.

Let S be the sample space.

(i) We want to find $P(J)$. To do this we can go back to the definition of probability and use Eq [3.1.3]

$$P(J) = \frac{|J|}{|S|} = \frac{8}{15} \qquad (10.2.10)$$

(ii) We want to find $P(J \cap H)$. To do this, we can use the identity from Eq [3.1.5]

$$\begin{aligned} P(J \cap H) &= P(J) + P(H) - P(J \cup H) && (10.2.11)\\ &= \frac{8}{15} + \frac{10}{15} - \frac{15}{15} && (10.2.12)\\ &= \frac{3}{15} = 0.2 && (10.2.13) \end{aligned}$$

(iii) Since the choices are independent, we know from Def [3.1.5] that

$$\begin{aligned} P(1st \text{ is Hip-hop and } 2nd \text{ is Hip-hop }) &= P(1st \text{ is Hip-hop }) && (10.2.14)\\ &\times P(2nd \text{ is Hip-hop }) && (10.2.15) \end{aligned}$$

Now lets find these probabilities. Finding the first probability is identical to what we did in (i) so we get ,

$$P(1st \text{ is Hip-hop }) = \frac{10}{15} \qquad (10.2.16)$$

Also, if the first one practiced hip-hop, that means there are 9 more students that practice hip-hop left in a group of 14. So,

$$P(2nd \text{ is Hip-hop }) = \frac{9}{14} \tag{10.2.17}$$

$$\therefore P(1st \text{ is Hip-hop and } 2nd \text{ is Hip-hop }) = \frac{10}{15} \times \frac{9}{14} \tag{10.2.18}$$

$$= \frac{3}{7} \tag{10.2.19}$$

Alternatively, we could have used a combinatorial approach to this problem. The probability that 2 students chosen at random practice hip-hop is equivalent to determining

$$P(\text{ 2 students practice hip-hop }) \tag{10.2.20}$$

$$= \frac{\text{no. of ways I can CHOOSE two hip-hop students}}{\text{no. of ways I can CHOOSE two students}} \tag{10.2.21}$$

$$= \frac{\binom{10}{2}}{\binom{15}{2}} = \frac{3}{7} \tag{10.2.22}$$

(c) Let C be the event that a patron chooses a chicken patty. $P(C) = 0.45$

Let F be the event that an individual choses fish. $P(F) = 0.35$

Let B be the event that an individual chooses a beef. $P(B) = 0.2$

We assume that EACH individual chooses independently.

(i) We want to find $P(CCC)$, where I have used CCC as short for $C \cap C \cap C$. Since the patrons do not influence the choices of the other, we use the independence, see Def [3.1.5], of the choices to write,

$$P(CCC) = P(C)P(C)P(C) \tag{10.2.23}$$

$$= 0.45 \times 0.45 \times 0.45 \tag{10.2.24}$$

$$= 0.091 \tag{10.2.25}$$

(ii) This means that they can either all choose chicken , fish or beef. Hence, we want to find,

$$P(CCC \cup FFF \cup BBB) = P(CCC) + P(FFF) + P(BBB) \tag{10.2.26}$$

$$= P(C)^3 + P(F)^3 + P(B)^3 \tag{10.2.27}$$

$$= 0.45^3 + 0.35^3 + 0.2^3 \tag{10.2.28}$$

$$= 0.142 \tag{10.2.29}$$

where we again used the independence of the choices to reduce $P(FFF)$ to $P(F)^3$ and $P(BBB)$ to $P(B)^3$.

(iii) We want to find the probability that they are all different. We can express this as $P(CFB \cup CBF \cup BFC \cup BCF \cup FCB \cup FBC)$.

$$P(CFB \cup CBF \cup BFC \cup BCF \cup FCB \cup FBC) \tag{10.2.30}$$

$$= P(CFB) + P(CBF) + P(BFC) + P(BFC) + P(FCB) + P(FBC) \tag{10.2.31}$$

$$= 6 \times (0.45)(0.35)(0.2) \tag{10.2.32}$$

$$= 0.189 \tag{10.2.33}$$

This was quite cumbersome and one can easily make a mistake if they do not interpret what just happened. We know that the independence of the choices ensures that any arrangement of C, F, B gives the same probability. i.e.

$$P(CFB) = P(C) \times P(F) \times P(B) \tag{10.2.34}$$

$$= P(F) \times P(B) \times P(C) \tag{10.2.35}$$

$$= P(FBC) \tag{10.2.36}$$

Now we just need a quicker way to get the factor of 6. Since we showed that all the permutations contribute an equal amount to the probability, we need to determine the number of ways, n, we can permute 3 distinct objects in 3 spaces.

$$\begin{aligned} n &= {}^3P_3 \tag{10.2.37}\\ &= \frac{3!}{(3-3)!} \tag{10.2.38}\\ &= 6 \tag{10.2.39}\end{aligned}$$

(iv) We want to find $P(CCC|\mathit{all\ same})$. To do this, we can use Def [3.1.6] for conditional probability and also the probability we calculated in (ii),

$$P(CCC|\mathit{all\ same}) = \frac{P(CCC \cap \mathit{all\ same})}{P(\mathit{all\ same})} \tag{10.2.40}$$

But $(CCC \cap \mathit{all\ same}) = P(CCC)$. So,

$$\begin{aligned} P(CCC|\mathit{all\ same}) &= \frac{P(CCC)}{P(\mathit{all\ same})} \tag{10.2.41}\\ &= \frac{0.45^3}{0.142} \tag{10.2.42}\\ &= \frac{729}{1136} = 0.642 \tag{10.2.43}\end{aligned}$$

4. (a) We are given $p = 0.55$ where p is the probability of success. Assume the events are independent of each other.

(i) Let X be the number of success in a 30 day time period. We see that X satisfies conditions in Note [3.3.2] to be modeled by a binomial distribution i.e $X \sim Bin(30, 0.55)$ We want to find $E[X]$. Recall from Eq [3.3.2],

$$\begin{aligned} E[X] &= np \tag{10.2.44}\\ &= 30 \times 0.55 \tag{10.2.45}\\ &= 16.5 \approx 17 \tag{10.2.46}\end{aligned}$$

(ii) Let Y be the number of success in a 7 day time period. We see that Y satisfies conditions in Note [3.3.2] to be modeled by a binomial distribution, i.e. $Y \sim Bin(7, 0.55)$. We want to compute $P(Y = 3)$. Using Def [3.3.1], we know

$$\begin{aligned} P(Y = y) &= \binom{7}{y} 0.55^y (1 - 0.55)^{7-y} \quad y \in (0, 1, 2, ..., 7) \tag{10.2.47}\\ P(Y = 3) &= \binom{7}{3} 0.55^3 (1 - 0.55)^{7-3} \tag{10.2.48}\\ &= 0.239 \tag{10.2.49}\end{aligned}$$

(iii) We want to determine the probability of success on the 4th, 5th and 6th day of the 30 day time period.

We do not care about what happens on the other days, so we just want to find the probability of 3 consecutive suceesses. Since the events are independent we have

$$\begin{aligned} P(3 \text{ successes}) &= p^3 \tag{10.2.50}\\ &= 0.55^3 \tag{10.2.51}\\ &= 0.166 \tag{10.2.52}\end{aligned}$$

(b) We are given the random varaiable Y such that $Y \sim Bin(n, 0.7)$ and $Var[Y] = 3.15$.

(i) We want to show that $n = 15$. We can do this by using Def [3.3.3]

$$\begin{aligned} Var[Y] &= np(1-p) \tag{10.2.53}\\ 3.15 &= n \times 0.7 \times (1 - 0.7) \tag{10.2.54}\\ n &= 15 \tag{10.2.55}\end{aligned}$$

(ii) We want to calculate $E[Y]$ from Def [3.3.2]

$$\begin{aligned} E[Y] &= np \tag{10.2.56}\\ &= 15 \times 0.7 \tag{10.2.57}\\ &= 10.5 \tag{10.2.58}\end{aligned}$$

(iii) We want to calculate $P(Y = 10)$. We can do this from Def [3.3.1],

$$P(Y = y) = \binom{15}{y} 0.7^y (1 - 0.7)^{15-y} \quad y \in (0, 1, 2, ..., 15) \tag{10.2.59}$$

$$P(Y = 10) = \binom{15}{10} 0.7^1 0(1 - 0.7)^{15-10} \tag{10.2.60}$$

$$= 0.206 \tag{10.2.61}$$

(c) **(i)** We know from Note [3.4.2], that we should have $np(1 - p) \geq 5$, $np > 5$ and $n > 30$.

(ii) We are told that the probability of success is $p = 0.06$. Let X be the number of successes in $n = 100$ trials. We want to find $P(X \leq 5)$.

We see that X satisfies the conditions in Note [3.3.2] to be modeled by a binomial distribution according to Def [3.3.1. We can take the lengthy procedure to this question and find $P(X \leq 5)$ by,

$$P(X \leq 5) = P(X = 0) + ... + P(X = 5) \tag{10.2.62}$$

$$= \binom{100}{0}(0.06)^0(1 - 0.06)^{100} + ... + \binom{100}{5} 0.06^5(1 - 0.06)^{100-5} \tag{10.2.63}$$

$$= 0.441 \tag{10.2.64}$$

However, it would be easier if X could be approximated by a normal distribution. We see that

$$np = 100(0.06) = 6 > 5 \tag{10.2.65}$$

$$n(1 - p) = 100(0.94) = 94 > 5 \tag{10.2.66}$$

So it satisfies the conditions to be modeled by such. Thus we have from Def [3.4.1],

$$\mu = np = 6 \tag{10.2.67}$$

$$\sigma^2 = np(1 - p) = 5.46 \tag{10.2.68}$$

So $X \sim N(6, 5.46)$. Now to find $P(X \leq 5)$, we apply the continuity correction Note [3.4.3] to get

$$P(X \leq 5) \rightarrow P(X < 5.5) \tag{10.2.69}$$

We determine the probability by standardizing,

$$P(X < 5.5) = P(X - \mu < 5.5 - 6) \tag{10.2.70}$$

$$= P\left(\frac{X - \mu}{\sigma} < \frac{5.5 - 6}{\sqrt{5.46}}\right) \tag{10.2.71}$$

$$= P(Z < -0.21398) \tag{10.2.72}$$

$$= \Phi(-0.21398) \tag{10.2.73}$$

$$= 1 - \Phi(0.21398) \tag{10.2.74}$$

$$= 0.417 \tag{10.2.75}$$

10.3 Module 3: Analyzing and Interpreting Data

5. We are given that $X \sim N(\mu, 0.19^2)$

(a) We can apply the Central Limit Theorem [4.1.1] directly. Since X is normally distributed, then for any n, we can apply the theorem and get:

$$E[\bar{X}] = \mu \tag{10.3.1}$$

$$Var(\bar{X}) = \frac{\sigma^2}{n} = \frac{0.19^2}{n} \tag{10.3.2}$$

$$\therefore \quad \bar{X} \sim \mathbf{N}\left(\mu, \frac{0.19^2}{n}\right) \tag{10.3.3}$$

(b) We are given a 90% confidence interval $(0.54, 0.61)$.

(i) Recall from Def [4.1.5] that the limits of a confidence interval are symmetric about $\bar{X}$. i.e. it takes the form $(\bar{X}-\beta, \bar{X}+\beta)$. Thus, we can find $\bar{X}$ by simply taking the average of the limits

$$\bar{X} = \frac{0.54+0.61}{2} = 0.575 \tag{10.3.4}$$

(ii) We want to find the value of n such that the limits of the confidence interval are $(0.54, 0.61)$. We know from Eq [4.1.5] that for σ^2 known and $n \geq 30$, the confidence interval takes the form $\bar{X} \pm Z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}}$, for a $(1-\alpha)$.100% confidence interval. Thus we can compare this to the given limits, and then solve for n

$$\bar{X} + Z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}} = 0.61 \tag{10.3.5}$$

$$\sqrt{\frac{\sigma^2}{n}} = \frac{0.61-\bar{X}}{Z_{\frac{\alpha}{2}}} \tag{10.3.6}$$

$$\frac{\sigma^2}{n} = \frac{0.19^2}{n} = \left(\frac{0.61-\bar{X}}{Z_{\frac{10}{2}}}\right)^2 \tag{10.3.7}$$

$$\therefore \quad n = \left(\frac{0.19 \times Z_5}{0.61-\bar{X}}\right)^2 \tag{10.3.8}$$

$$= \left(\frac{0.19 \times 1.645}{0.61-0.575}\right)^2 \tag{10.3.9}$$

$$= 79.74 \approx 80 \tag{10.3.10}$$

Alternatively, we could have used the lower limit instead and again solved for n

$$\bar{X} - Z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}} = 0.54 \tag{10.3.11}$$

$$\sqrt{\frac{\sigma^2}{n}} = \frac{\bar{X}-0.54}{Z_{\frac{\alpha}{2}}} \tag{10.3.12}$$

$$n \approx 80 \tag{10.3.13}$$

Again, we can solve this in one other way, which provides the advantage that you can get this right even if you make a mistake with $\bar{X}$. We know that the width of the confidence interval should be $2 \times Z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}$. So equating this to the width of our interval $(0.54, 0.61)$,

$$2 \times Z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}} = 0.61 - 0.54 \tag{10.3.14}$$

$$n = \left(\frac{\sigma \times 2 \times z_5}{0.61-0.54}\right)^2 \tag{10.3.15}$$

$$= \left(\frac{0.19 \times 2 \times 1.65}{0.07}\right)^2 \tag{10.3.16}$$

$$\approx 80 \tag{10.3.17}$$

(c) We should recall what it means to construct a 90% confidence interval for μ from Def [4.1.5]. Thus, a 90% confidence interval for μ means that we will find, with probability 0.90 a confidence interval in which the actual value of the parameter μ will be between the stochastic endpoints of the confidence interval. This means that with probability 0.10 our confidence interval will not contain μ, so if we construct 20 confidence intervals, we can expect 2 of them to not contain the population mean μ.

(d) We are given that $\sum x = 49.5$ and $\sum x^2 = 348.57$.

(i) **a)** Recall from Def [4.1.2] that we can compute the unbiased estimate for the mean by:

$$\begin{aligned} \bar{X} &= \frac{\sum X}{n} && (10.3.18)\\ &= \frac{49.5}{50} && (10.3.19)\\ &= 0.99 \quad \text{thousand dollars} && (10.3.20)\\ &= \$990.00 && (10.3.21) \end{aligned}$$

b) Recall from Def [4.1.3] that we can determine an unbiased population variance by,

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n-1}\left[\sum x^2 - \frac{(\sum x)^2}{n}\right] && (10.3.22)\\ &= \frac{1}{50-1}\left[348.57 - \frac{49.5^2}{50}\right] && (10.3.23)\\ &= 6.1136 && (10.3.24)\\ \therefore \quad \sqrt{\hat{\sigma}^2} &= \sqrt{6.1136} = 2.473 \quad \textit{thousand dollars} && (10.3.25) \end{aligned}$$

(ii) Recall from Eq [4.1.6], since σ^2 is unknown and $n \geq 30$, the confidence interval for μ takes the form $\bar{x} \pm Z_{\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{n}}$.

$$\begin{aligned} \bar{x} &\pm Z_{\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{n}} && (10.3.26)\\ 0.99 &\pm Z_{0.06/2}\sqrt{\frac{6.1136}{50}} && (10.3.27)\\ 0.99 &\pm Z_{0.03}\sqrt{\frac{6.1136}{50}} && (10.3.28)\\ 0.99 &\pm 1.88\sqrt{\frac{6.1136}{50}} && (10.3.29)\\ 0.99 &\pm 0.657 && (10.3.30)\\ &\Rightarrow (0.333, 1.647) && (10.3.31) \end{aligned}$$

6. **(a)** Recall from Section [4.3] how to conduct a χ^2 test at the 5% level of significance.

(i) We can state the null and alternative hypothesis as
H_0: There is no association between the type of book sold and the type of cover it has
H_1: There is an association between the type of book sold and the type of cover it has.

(ii) We can compute the expected frequency using $E_{ij} = \frac{n_i \times n_j}{N}$, where E_{ij} represents the expected frequency of the value in the i^{th} row and j^{th} column, n_i represents the sum of values in the i^{th} row, n_j represents the sum of values in the j^{th} column, and N represents the total sample size.

a) Textbooks with hard back covers correspond to E_{12},

$$E_{12} = \frac{56 \times 20}{150} = 7.467 \approx 7 \qquad (10.3.32)$$

b) Novels with paper back covers corresponds to E_{21},

$$E_{21} = \frac{90 \times 94}{150} = 56.4 \approx 56 \qquad (10.3.33)$$

(iii) To identify the critical region for this test, we must determine the number of degrees of freedom and then find the corresponding value in the χ^2 table at the 5% significant level. We can calculate the number of degrees of freedom as follows

$$D.o.f = (r-1)(c-1) = (2-1)(3-1) = 2 \qquad (10.3.34)$$

Where r and c represent the number of rows and columns respectively. This corresponds to a value of 5.991. Therefore, the critical region for this test is $\chi > 5.991$.

(iv) Since $\chi_{calc} = 9.2649 > 5.991$, it is inside the critical region and hence we must reject H_0. Thus, at the 5% level of significance, there is enough evidence to suggest that there is an association between the type of book and its cover.

(b) Recall from Section [4.2.2] how to execute a t-test.

- **(i)** We know from Section [4.2.2] that the t-test will be valid if we assume that the mass follows a normal distribution.
- **(ii)** We can state the null and alternative hypothesis as

$$\begin{aligned} H_0 &: \mu = \mu_0 = 150 \text{ g} && (10.3.35) \\ H_1 &: \mu > \mu_0 && (10.3.36) \\ &: \mu > 150 \text{ g} && (10.3.37) \end{aligned}$$

- **(iii)** Recall from Section [4.2.3] that for a one-tailed test at the 5% level of significance, we reject H_0 if

$$\begin{aligned} t_{\text{calc}} &> t_{0.05}^{(15-1)} && (10.3.38) \\ t_{\text{calc}} &> 1.761 && (10.3.39) \end{aligned}$$

- **(iv)** Recall from Section [4.2.2] that we can calculate the test statistic as

$$\begin{aligned} t_{\text{calc}} &= \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n-1}}} && (10.3.40) \\ &= \frac{158 - 150}{43/\sqrt{14}} && (10.3.41) \\ &= 0.696 && (10.3.42) \end{aligned}$$

- **(v)** Since $t_{\text{calc}} = 0.696 < 1.761$, our test statistic does not lie within the critical region. Hence, we fail to reject H_0.

Chapter 11

2011

11.1 Module 1: Collecting and Describing Data

1. **(a)** **(i)** A, D and E can be obtained through some quantifiable measurable process.

(ii) B and C are descriptive, subjective or difficult to measure.

(iii) A, E can all be counted.

(iv) D can take any value within a range.

(v) B and C because it is qualitative data

(b) A sample of 20 may be selected using different sampling methods.

(i) We can identify the following

- G: Simple random, Def [2.2.2]
- H: Systematic, Def [2.2.4]
- I: Quota, Def [2.2.6]
- J: Stratified, Def [2.2.3]
- K: Cluster, Def [2.2.5]

(ii) The question fails to inform us of the population size. Assuming that the population size mentioned in a subsection is 300, we know that n must divide the 300 cans in 20 sections, so

$$\begin{aligned} n &= \frac{300}{20} && (11.1.1)\\ &= 15 && (11.1.2)\end{aligned}$$

(iii) We need to keep the ratio of blue in the sample representative of the ratio of blue in the entire population. Then, the number of blue in the sample, n_b, must satisfy the following equation:

$$\begin{aligned} \frac{120}{300} &= \frac{n_b}{20} && (11.1.3)\\ \Rightarrow n_b &= 8 && (11.1.4)\end{aligned}$$

(iv) (1) Stratified sampling is more representative of the entire population. Simple random sampling sometimes leads to over or under representation of a stratum

(2) In stratified sampling, every member within a strata has an equal chance of selection. In simple random sampling, every member is the population has an equal chance of selection

(c) From Def [2.1.6] and Def [2.1.8], we see that both quantities are numerical measures. However, the parameter is calculated from the population whereas the statistic is calculated from the sample data.

2. **(a)** Let $F(X)$ be the number of people who waited $x \leq X$ minutes in the clinic.

(i) From Def [2.3.5], we see that we want to find the value of X such that $F(X) = 50$. Reading this off from the graph:

$$Q_2 = 25 \text{ minutes} \qquad (11.1.5)$$

(ii) From Def [2.3.5], we first need to compute Q_1 and Q_3. Reading off the graph,

$$Q_1 = 20 \tag{11.1.6}$$
$$Q_3 = 29 \tag{11.1.7}$$
$$IQR = Q_3 - Q_2 \tag{11.1.8}$$
$$= 29 - 20 \tag{11.1.9}$$
$$= 9 \tag{11.1.10}$$

(iii) To determine how many individuals waited more than 20 minutes, we must find

$$100 - F(20) = 100 - 26 \tag{11.1.11}$$
$$= 74 \tag{11.1.12}$$

(b) **(i)** We are given an ordered list of numbers.

a) We want to draw a stem and leaf diagram to represent this data,

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

Key: 1|3 means 13

Figure 11.1: A stem and leaf diagram of the data given.

b) From Figure [11.1], we see that the values with the highest frequencies are 10 and 13.

c) Recall from Def [2.3.5] that we can determine the median by

$$Q_2 = \frac{n+1}{2}^{\text{th}} \text{ term} \tag{11.1.13}$$
$$= \frac{25+1}{2}^{\text{th}} \text{ term} \tag{11.1.14}$$
$$= 13^{\text{th}} \text{ term} \tag{11.1.15}$$
$$= 12 \tag{11.1.16}$$

d) Recall Def [2.3.5] how to determine the upper and lower quartiles, and hence, the interquartile range,

$$\begin{aligned} Q_1 &= \frac{n+1}{4}^{\text{th}} \text{ term} && (11.1.17)\\ &= \frac{25+1}{4}^{\text{th}} \text{ term} && (11.1.18)\\ &= 6.25^{\text{th}} \text{ term} && (11.1.19)\\ &= 6^{\text{th}} \text{ term} + 0.25 \times (7^{\text{th}} \text{ term} - 6^{\text{th}} \text{ term}) && (11.1.20)\\ &= 9 + 0.25(9-9) && (11.1.21)\\ &= 9 && (11.1.22)\\ Q_3 &= \frac{3(n+1)}{4}^{\text{th}} \text{ term} && (11.1.23)\\ &= 18.75^{\text{th}} \text{ term} && (11.1.24)\\ &= 18^{\text{th}} \text{ term} + 0.75(19^{\text{th}} \text{ term} - 18^{\text{th}} \text{ term}) && (11.1.25)\\ &= 14 + 0.75(16-14) && (11.1.26)\\ &= 15.5 && (11.1.27)\\ IQR &= Q_3 - Q_2 && (11.1.28)\\ &= 15.5 - 9 && (11.1.29)\\ &= 6.5 \approx 7 && (11.1.30) \end{aligned}$$

e) Recall from Def [2.3.3], that an 8% trimmed mean means we discard $0.08 * 25 = 2$ data points from the upper and lower end of the data. So we discard the $\{4,4\}$ and $\{25,30\}$ from the lower and upper ends respectively. Recalculating the mean,

$$\begin{aligned} \bar{x} &= \frac{sumx}{n} && (11.1.31)\\ &= \frac{7+8+...18+20}{25-4} && (11.1.32)\\ &= 12.4 && (11.1.33) \end{aligned}$$

f) Recall from Def [2.3.6] how to determine skewness. Computing the differences in the quartiles, we see

$$\begin{aligned} Q_3 - Q_2 &> Q_2 - Q_1 && (11.1.34)\\ 15.5 - 12 &> 12 - 9 && (11.1.35) \end{aligned}$$

and we conclude that it is positively skewed.

(ii) **a)** Mode: Not affected. If the value occurs with the highest frequency it would not be considered an outlier.

b) Mean: Affected by definition.

c) Median: Not affected. Having an outliers has the same effect on the median as having a data point that is not an outlier.

d) Range: Affected by definition of range.

11.2 Module 2: Managing Uncertainty

3. **(a)** Let X be the number of success in 10 trials. From the conditions in Note [3.3.2], we see that X can be modeled by a binomial distribution, $X \sim \text{Bin}(10, 2)$. So from Def [3.3.1], we have

$$P(X = x) = \binom{10}{x}(0.2)^x(1-0.2)^{10-x} \quad x \in 0, 1, 2, ..., 10 \qquad (11.2.1)$$

(i) We want to find $P(X = 2)$. Using the formula above we see that,

$$\begin{aligned} P(X=2) &= \binom{10}{2}0.2^2 0.8^8 && (11.2.2)\\ &= 0.302 && (11.2.3) \end{aligned}$$

(ii) We want to find, $P(X \geq 4)$. We can calculate this using the definition above as follows:

$$P(X \geq 4) = P(X = 4) + P(X = 5) + P(X = 6)... + P(X = 10) \tag{11.2.4}$$

However, it would be much quicker to recognize that $P(X \geq 4) = 1 - P(X < 4)$, which significantly reduces the calculation time.

$$\begin{aligned}
P(X \geq 4) &= 1 - P(X < 4) && (11.2.5)\\
&= 1 - [P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)] && (11.2.6)\\
&= 1 - \binom{10}{0}(0.2)^0(0.8)^{10} - \binom{10}{1}(0.2)(0.8)^9 && (11.2.7)\\
&- \binom{10}{2}(0.2)^2(0.8)^8 - \binom{10}{3}(0.2)^3(0.8)^7 && (11.2.8)\\
&= 0.121 && (11.2.9)
\end{aligned}$$

(b) We want to find $E[X]$. For a binomial distribution, the expectation is given by Eq [3.3.2],

$$\begin{aligned}
E[X] &= np && (11.2.10)\\
&= (10)(0.2) && (11.2.11)\\
&= 2 && (11.2.12)
\end{aligned}$$

(c) From Note [3.4.2] we see that since

$$\begin{aligned}
np &= 95(0.2) && (11.2.13)\\
&= 19 > 5 && (11.2.14)\\
npq &= 95(0.2)(0,8) && (11.2.15)\\
&= 15.2 > 5 && (11.2.16)
\end{aligned}$$

we can use the normal approximation to the binomial distribution according to Theorem [3.4.1] Let Y be the number of successes in 95 trials. Using the normal approximation, we can say that Y is normally distributed with parameters $\mu = 19$ and $\sigma^2 = 15.2$ i.e $Y \sim \mathrm{N}(19, 15.2)$. We want to find $P(14 \leq Y \leq 21)$. Applying the continuity correction, Note [3.4.3],

$$P(14 \leq Y \leq 21) \rightarrow P(13.5 < Y << 21.5) \tag{11.2.17}$$

We can proceed by standardizing Y,

$$\begin{aligned}
P(13.5 < Y < 21.5) &= P\left(\frac{13.5-\mu}{\sigma} < \frac{Y-\mu}{\sigma} < \frac{21.5-\mu}{\sigma}\right) && (11.2.18)\\
&= P\left(\frac{13.5-19}{\sqrt{15.2}} < Z < \frac{21.5-19}{\sqrt{15.2}}\right) && (11.2.19)\\
&= P\left(-1.411 < Z < 0.641\right) && (11.2.20)\\
&= \Phi(0.641) - \Phi(-1.411) && (11.2.21)\\
&= \Phi(0.641) - (1 - \Phi(1.411)) && (11.2.22)\\
&= 0.7389 - (1 - 0.9207) && (11.2.23)\\
&= 0.660 && (11.2.24)
\end{aligned}$$

(d) We are given a cumulative distribution function $F(X)$ for the discrete random variable X.

(i) To do this, we should first recall from Def [3.2.5] that for a discrete random variable, the cumulative distribution function is given by

$$\begin{aligned}
F(x) &= P(X \leq X) && (11.2.25)\\
& && (11.2.26)
\end{aligned}$$

. Using this definition, we can calculate the required probabilities using the fact that

$$\begin{aligned}
f(x) = P(X = x) & && (11.2.27)\\
&= P(X \leq x) - P(X \leq x - 1) && (11.2.28)\\
&= F(x) - F(x - 1) && (11.2.29)
\end{aligned}$$

x	2	3	4	5	6
f(x)	0.01	0.22	0.41	0.22	0.14

Table 11.1: The probability distribution table for the discrete random variable X

(ii) We want to determine $P(X > 3)$. Looking at Table 11.1, we see that

$$\begin{aligned} P(X > 3) &= P(X = 4) + P(X = 5) + P(X = 6) && (11.2.30)\\ &= f(4) + f(5) + f(6) = 0.41 + 0.22 + 0.14 && (11.2.31)\\ &= 0.77 && (11.2.32) \end{aligned}$$

Alternatively, it would be easier to see that

$$\begin{aligned} P(X > 3) &= 1 - P(X \leq 3) && (11.2.33)\\ &= 1 - F(3) && (11.2.34)\\ &= 1 - 0.23 && (11.2.35)\\ &= 0.77 && (11.2.36) \end{aligned}$$

(iii) We can compute $E[X]$ by Eq [3.3.2] and then use the table to get the required probabilities.

$$\begin{aligned} E[X] &= \sum_{\forall x} xP(X = x) && (11.2.37)\\ &= 2(0.01) + 3(0.22) + 4(0.41) + 5(0.22) + 6(0.14) && (11.2.38)\\ &= 4.26 && (11.2.39) \end{aligned}$$

We also know by Eq [3.3.3] that

$$Var(X) = E[X^2] - (E[X])^2 \qquad (11.2.40)$$

But first, we need to find the second moment of X, $E[X^2]$ from Def [3.2.3],

$$\begin{aligned} E[X^2] &= \sum_{\forall x} x^2 P(X = x) && (11.2.41)\\ &= 2^2(0.01) + 3^2(0.22) + 4^2(0.41) + 5^2(0.22) + 6^2(0.14) && (11.2.42)\\ &= 19.12 && (11.2.43)\\ \therefore Var(X) &= E[X^2] - (E[X])^2 && (11.2.44)\\ &= 19.12 - 4.26^2 && (11.2.45)\\ &= 0.9724 && (11.2.46) \end{aligned}$$

4. (a) We are given $P(A) = 0.6$, $P(B) = 0.4$ and $P(A|B) = 0.2$. Before, we continue, we should note from Def [3.1.6] what the conditional probability means since this is in the question.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad (11.2.47)$$

(i) We want to find $P(A \cap B)$ To do this, we can use equation above and rearrange to find the desired term.

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} && (11.2.48)\\ P(A \cap B) &= P(A|B) \times P(B) && (11.2.49)\\ &= 0.2 \times 0.4 && (11.2.50)\\ &= 0.08 && (11.2.51) \end{aligned}$$

(ii) We want to find $P(A \cup B)$. To do this, we can use the identity in Eq [3.1.5],

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) && (11.2.52)\\ &= 0.6 + 0.4 - 0.08 && (11.2.53)\\ &= 0.92 && (11.2.54) \end{aligned}$$

(iii) We want to find $P(B|A)$. To do this, we can return to Def [3.1.6],

$$\begin{aligned} P(B|A) &= \frac{P(B \cap A)}{P(A)} && (11.2.55) \\ &= \frac{0.08}{0.6} = \frac{2}{15} && (11.2.56) \end{aligned}$$

(iv) We want to find $P(\bar{A} \cap B)$. To do this, a Venn Diagram would be useful to determine how we can reduce this into quantities we know.

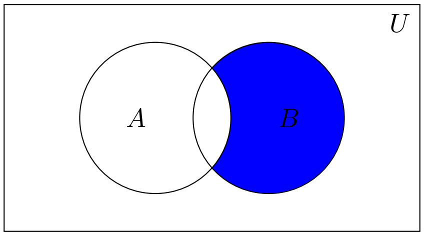


Figure 11.2: A Venn Diagram representing $P(\bar{A} \cap B)$

Hence, from Fig 11.2, we see that

$$\begin{aligned} P(\bar{A} \cap B) &= P(B) - P(A \cap B) && (11.2.57) \\ &= 0.4 - 0.08 && (11.2.58) \\ &= 0.32 && (11.2.59) \end{aligned}$$

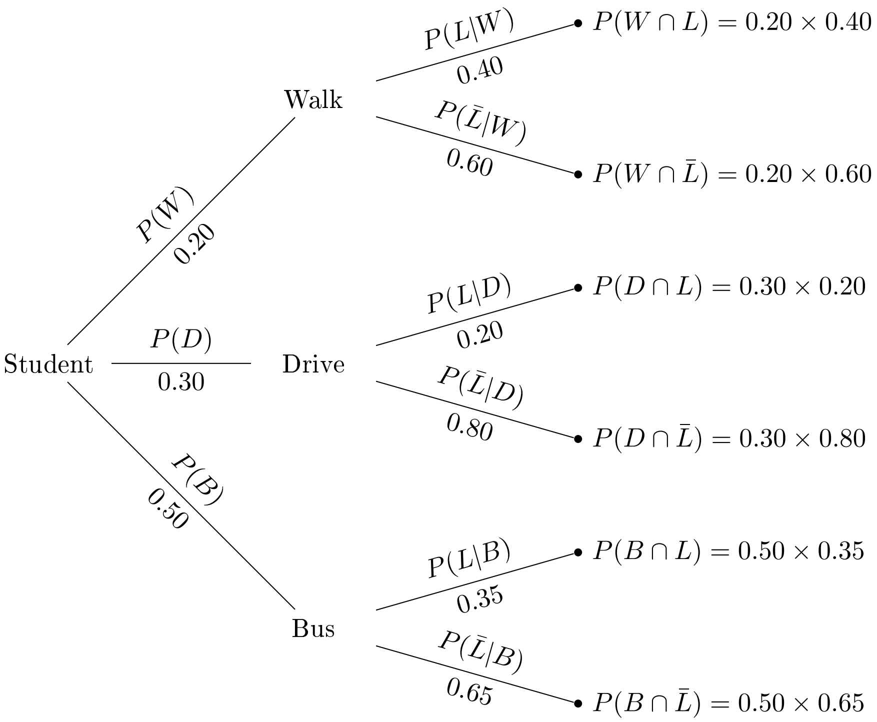
We could have also done this from Eq [3.1.6], however it is advisable to do this from diagram instead.

(v) Recall from Def [3.1.5], we can determine if two events are independent if:

$$\begin{aligned} P(A \cap B) &\overset{?}{=} P(A) \times P(B) && (11.2.60) \\ 0.08 &\overset{?}{=} 0.6 \times 0.4 && (11.2.61) \\ \therefore P(A \cap B) &\neq P(A) \times P(B) && (11.2.62) \end{aligned}$$

Hence, the events A and B are not independent.

(b) (i) Let W be the event that an individual walks.
Let D be the event that an individual drives.
Let B be the event that an individual takes the bus.
Let L be the event that an individual is late.



Student
P(W)
0.20
P(D)
0.30
P(B)
0.50
Walk
Drive
Bus
P(L|W)
0.40
P(L̄|W)
0.60
P(L|D)
0.20
P(L̄|D)
0.80
P(L|B)
0.35
P(L̄|B)
0.65
$P(W \cap L) = 0.20 \times 0.40$
$P(W \cap \bar{L}) = 0.20 \times 0.60$
$P(D \cap L) = 0.30 \times 0.20$
$P(D \cap \bar{L}) = 0.30 \times 0.80$
$P(B \cap L) = 0.50 \times 0.35$
$P(B \cap \bar{L}) = 0.50 \times 0.65$

Figure 11.3: A tree diagram illustrating the information given for a student's walk to school.

(ii) Using the tree diagram we can compute numerous probabilities.

a) We want to find $P(L)$. An individual can be late if they walk, drive or take the bus. Hence, we must take these three scenarios into consideration. Using Def [3.1.6] for conditional probability, we compute the probability of being late in each scenario,

$$\begin{aligned} P(L) &= P(L_{walk}) + P(L_{drive}) + P(L_{Bus}) && (11.2.63)\\ &= P(L|W)P(W) + P(L|D)P(D) + P(L|B)P(B) && (11.2.64)\\ &= (0.4)(0.2) + (0.2)(0.3) + (0.35)(0.5) && (11.2.65)\\ &= 0.315 && (11.2.66) \end{aligned}$$

b) We want to find $P(B|\bar{L})$. Once again, we can use Def [3.1.6]

$$P(B|\bar{L}) = \frac{P(B \cap \bar{L})}{P(\bar{L})} \qquad (11.2.67)$$

We can read off the numerator from the value given in the tree diagram. The denominator can be easily calculated using Def [3.1.3]. Hence,

$$\begin{aligned} P(B|\bar{L}) &= \frac{P(B \cap \bar{L})}{P(\bar{L})} && (11.2.68)\\ &= \frac{0.5 \times 0.65}{1 - 0.315} && (11.2.69)\\ &= \frac{65}{137} = 0.474 && (11.2.70) \end{aligned}$$

11.3 Module 3: Analysing and Interpreting Data

5. We are given a table and summarized information.

(a) We can represent the information given the table with a scatter diagram.

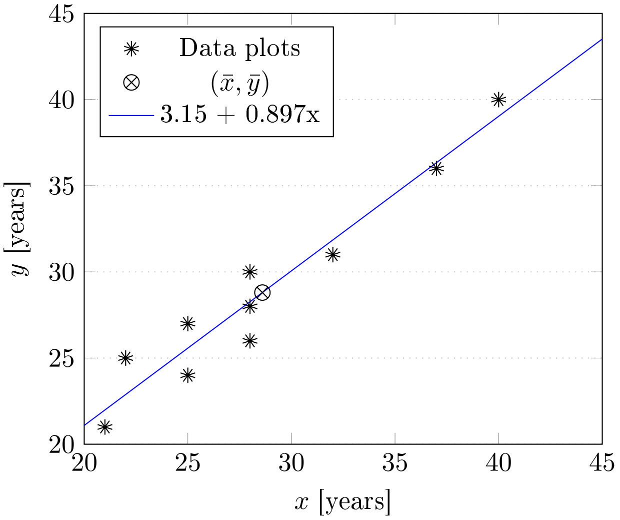


Figure 11.4: A scatter diagram of the actual age (y years) versus the predicted age (x years) for 10 persons.

(b) **(i)** We can calculate the mean of x by Def [4.4.2] of $\bar{x}$,

$$\bar{x} = \frac{\sum x}{n} \tag{11.3.1}$$

$$= \frac{286}{10} = 28.6 \tag{11.3.2}$$

(ii) Similarly, we can do $\bar{y}$ by Def [4.4.2]

$$\bar{y} = \frac{\sum y}{n} \tag{11.3.3}$$

$$= \frac{288}{10} = 28.8 \tag{11.3.4}$$

(iii) We plot the point $(\bar{x}, \bar{y})$ on Figure [11.4]. From Note [4.4.3], we see that this point should lie on the line of regression.

(c) Recall the equation of linear regression from Def [4.4.2]. Thus we want to find the values of a and b such that $y = a + bx$. Using the given formulas and provided data,

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} \tag{11.3.5}$$

$$= \frac{10(8542) - (286)(288)}{10(8520) - 286^2} \tag{11.3.6}$$

$$= 0.897 \tag{11.3.7}$$

$$a = \bar{y} - b\bar{x} \tag{11.3.8}$$

$$= 28.8 - 0.897(28.6) \tag{11.3.9}$$

$$= 3.15 \tag{11.3.10}$$

$$\therefore \quad y = 3.15 + 0.897x \tag{11.3.11}$$

(d) Recall from Def [4.4.1], that we can calculate the product moment correlation coefficient by,

$$r = \frac{\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}} \tag{11.3.12}$$

$$= \frac{10(8542) - (286)(288)}{\sqrt{10(8520) - 286^2}\sqrt{10(8588) - 288^2}} \tag{11.3.13}$$

$$= 0.965 \tag{11.3.14}$$

From Note [4.4.2], we can comment that this is indicates high positive linear correlation.

(e) We want to find the y that would correspond to an x value of 45. We can simply plug in $x = 45$ in Eq [11.3.4]

$$\begin{aligned} y &= 3.15 + 0.897(45) \quad (11.3.15) \\ &= 43.5 \approx 44 years \quad (11.3.16) \end{aligned}$$

(f) With regression, we usually do not allow extrapolation. Since $x = 45$ is outside the range of x in our data set, our estimate for y is not reliable.

6. **(a)** We are given that $\mu = 100$, $\sigma^2 = 16$ and $n = 50$.
This is a direct application of the Central Limit Theorem [4.1.1]. Applying this to our X,

$$\begin{aligned} E[\bar{X}] &= \mu = 100 \quad (11.3.17) \\ Var(\bar{X}) &= \frac{\sigma^2}{n} = \frac{16}{50} \quad (11.3.18) \\ \therefore \quad \bar{X} &\sim \mathbf{N}\left(100, \frac{16}{50}\right) \quad (11.3.19) \end{aligned}$$

(b) Using the Chi-Square Distribution Table, we see that a value of 9.488 corresponds to 4 degrees of freedom and a significance level of 0.05.

(c) Recall from Section [4.3] how to conduct a χ^2 test at the 5% level of significance.

(i) We can state the null and alternative hypothesis as: H_0: no association between gender and response of individuals.
H_1 : there is an association between gender and response of individuals.

(ii) **a)** We can calculate the number of degrees of freedom by definition using the given contingency table.

$$D.o.f = (r-1)(c-1) = (2-1)(3-1) = 2 \quad (11.3.20)$$

Where r and c represent the number of rows and columns respectively.

b) We calculate the critical by reading off the value in the table that corresponds to 2 degrees of freedom and 5% level of significance. From the table we get 5.991. Therefore, the critical region is $\chi^2 > 5.991$.

c) We want to find $\chi^2_{\text{test}} = \sum \frac{(O-E)^2}{E}$, where O is the observed data and E is the expected frequency. First we calculate the expected frequencies and summarize it in Table [11.2]

	In Favor	Against	No Opinion	Total
Male	$\frac{175\times175}{300} = 102.083$	$\frac{175\times107}{300} = 62.417$	$\frac{175\times18}{300} = 10.500$	175
Female	$\frac{125\times175}{300} = 72.917$	$\frac{125\times107}{300} = 44.583$	$\frac{125\times18}{300} = 7.500$	125
Total	175	107	18	300

Table 11.2: A table of the expected frequencies

Next, we calculate χ^2_{test} by taking $\sum \frac{(O-E)^2}{E}$. The results are summarized in Table 11.3.

Observed	Expected	$\frac{(O-E)^2}{E}$
90	102.083	$\frac{(90-102.083)^2}{102.083} = 1.430$
75	62.417	$\frac{(75-62.417)^2}{62.417} = 2.537$
10	10.500	$\frac{(10-10.500)^2}{10.500} = 0.024$
85	72.917	$\frac{(85-72.917)^2}{72.917} = 2.00$
32	44.583	$\frac{(32-44.384)^2}{44.384} = 3.556$
8	7.500	$\frac{(8-7.500)^2}{7.500} = 0.033$
		9.580

Table 11.3: A table used to compute the test statistic

(iii) The value of $\chi^2_{test} = 9.580 > 5.991$, it lies within the critical region, and so we must reject the null hypothesis H_0. Thus, at the 5% level of significance, there is enough evidence to suggest that there is an association between gender and response.

Chapter 12

2012

12.1 Module 1: Collecting and Describing Data

1. **(a)** A body has decided to conduct a survey among a sample of 60 of the individuals

(i) A. Cluster sampling, Def [2.2.5]

B. Stratified random sampling, Def [2.2.3]

C. Systematic sampling, Def [2.2.4]

(ii) From Note [2.2.4], one advantage is that this method is very cost and time efficient as extensive preparation of a sampling frame is not necessary.

(iii) From Note [2.2.3], one disadvantage is that the sample obtained may be biased if hidden periodicity in population coincides with that of selection.

(iv) To do this we can multiply the ratio of people at this location to the total population by the size of the sample.

$$\frac{90}{350} \times 60 = 15.429 \approx 15 \text{ persons} \tag{12.1.1}$$

(b) **(i)** From Def [2.2.1], a sampling frame is an extensive list of all elements or individuals in the population from which the sample is selected.

(ii) The telephone directory may not be a representative sampling frame as not every person will be listed there.

(iii) This survey will clearly be biased as the sample of callers would be part of only the listeners of this station and not of all stations.

(c) Given data on a sample of 25 students.

(i) We can construct a stem and leaf diagram as follows:

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

Key: 5|3 means 53

Figure 12.1: A stem and leaf diagram of the marks obtained by a sample of 25 students in a class test.

(ii) One advantage of a stem of leaf diagram is that all the data is presented as each individual value is shown.

(iii) **a)** Recall from Def [2.3.5] that we can determine the median as

$$Q_2 = \frac{n+1}{2} \tag{12.1.2}$$

$$= \frac{25+1}{2} \text{ th term} \tag{12.1.3}$$

$$= 13 \text{ th term} \tag{12.1.4}$$

$$= 62 \tag{12.1.5}$$

b) Recall from Def [2.3.5] that we must first compute Q_1 and Q_3 to determine the interquartile range,

$$IQR = Q_3 - Q_1 \tag{12.1.6}$$

$$= \frac{3(n+1)}{4} \text{ th term} - \frac{n+1}{4} \text{ th term} \tag{12.1.7}$$

$$= 19.5 \text{ th term} - 6.5 \text{ th term} \tag{12.1.8}$$

$$19.5 \text{ th term} = \frac{19 \text{ th term} + 20 \text{ th term}}{2} \tag{12.1.9}$$

$$= \frac{71+74}{2} \tag{12.1.10}$$

$$= 72.5 \tag{12.1.11}$$

$$6.5 \text{ th term} = \frac{6 \text{ th term} + 7 \text{ th term}}{2} \tag{12.1.12}$$

$$= \frac{57+58}{2} \tag{12.1.13}$$

$$= 57.5 \tag{12.1.14}$$

$$IQR = 72.5 - 57.5 \tag{12.1.15}$$

$$= 15 \tag{12.1.16}$$

2. **(a)** **(i)** *Population*: since it refers to all of the individuals

(ii) *Census*: since it is a survey on the entire population

(iii) *Sample*: since it is a subset of the population

(iv) *Statistic*: since it is a quantity derived from the sample

(v) *Parameter*: since it is a quantity derived from the population

(b) **(i)** **a)** Recall that the mode is simply the number that occurs with the highest frequency

$$\text{Mode} = 3 \text{ times} \tag{12.1.17}$$

b) Recall from Def [2.3.5] that the median of the distribution can be found as,

$$\text{Median} = \frac{n+1}{2}^{\text{th}} \text{ term} \tag{12.1.18}$$

$$= \frac{32+1}{2}^{\text{th}} \text{ term} \tag{12.1.19}$$

$$= 16.5^{\text{th}} \text{ term} \tag{12.1.20}$$

$$= 16^{\text{th}} \text{ term} + 0.5 \times (17^{\text{th}} \text{ term} - 16^{\text{th}} \text{ term}) \tag{12.1.21}$$

$$= 3 + 0.5 \times (3 - 3) \tag{12.1.22}$$

$$= 3 \tag{12.1.23}$$

(ii) **a)** Recall from Def [2.3.1] that we can determine the mean as

$$\bar{x} = \frac{\sum xf(x)}{\sum f(x)} \tag{12.1.24}$$

$$= \frac{2(0)+1(4)+2(5)+3(8)+4(7)+5(6)}{2+4+5+8+7+6} \tag{12.1.25}$$

$$= \frac{96}{32} \tag{12.1.26}$$

$$= 3 \tag{12.1.27}$$

b) Recall from Def [2.3.2] that we can determine the variance when we have grouped data as,

$$\sigma^2 = \frac{\sum fx^2}{\sum f} - \bar{x}^2 \tag{12.1.28}$$

$$= \frac{2(0^2)+4(1^2)+5(2^2)+8(3^2)+7(4^2)+6(5^2)}{2+4+5+8+7+6} - 3^2 \tag{12.1.29}$$

$$= 2.1875 \tag{12.1.30}$$

(iii) Since we have already determined its mode, mean and median, we see from Def [2.3.6] that it follows a normal distribution since

$$\text{Mode} = \text{Mean} = \text{Median} \tag{12.1.31}$$

(c) Let $F(x)$ be the number of people that took x minutes or less to finish.

(i) The number of people that took part is the highest value on the y-axis. So 350 people took part.

(ii) We want to find $F(25)$. We read off from the graph $\sim$ 35 walkers

(iii) We want to find $350 - F(55)$,

$$\text{Percentage} = \frac{350 - F(55)}{350} \times 100\% \tag{12.1.32}$$

$$= \frac{350 - 325}{350} \times 100\% = 7.14\% \tag{12.1.33}$$

(iv) Recall from Def [2.3.5] that we can determine the interquartile range as

$$IQR = Q_3 - Q_1 \tag{12.1.34}$$

$$= \frac{3n}{4}\text{ th value} - \frac{n}{4}\text{ th value} \tag{12.1.35}$$

$$= 44.5 - 31.0 \tag{12.1.36}$$

$$= 13.5 \text{ mins} \tag{12.1.37}$$

(v) It is easier to think about the value of x for which 40% of the people took less than or equal to x minutes to finish. This is simply the value of x such that

$$F(x) = 0.4 \times 350 = 140 \tag{12.1.38}$$

$$\Rightarrow x = 36.5 \text{ mins} \tag{12.1.39}$$

12.2 Module 2: Managing Uncertainty

3. (a) We are given that $P(R) = 0.4$, $P(Q) = 0.6$ and $P(R \cap Q) = 0.12$.

(i) a) We want to calculate $P(R \cup Q)$
To do this we can use Eq [3.1.5],

$$P(R \cup Q) = P(R) + P(Q) - P(R \cap Q) \tag{12.2.1}$$

$$= 0.4 + 0.6 - 0.12 \tag{12.2.2}$$

$$= 0.88 \tag{12.2.3}$$

b) We want to calculate $P(R|Q)$
To do this, we can use Def [3.1.6] of conditional probability

$$P(R|Q) = \frac{P(R \cap Q)}{P(Q)} \tag{12.2.4}$$

$$= \frac{0.12}{0.6} = 0.2 \tag{12.2.5}$$

(ii) a) Recall Def [3.1.5]. Hence, we must determine:

$$P(R \cap Q) \overset{?}{=} P(R) \times P(Q) \tag{12.2.6}$$

$$0.12 \neq 0.4 \times 0.6 \tag{12.2.7}$$

Therefore, the events R and Q are not independent.

b) Recall, Def [3.1.9] that the events R and Q are mutually exclusive if

$$\begin{aligned} P(R \cap Q) &\overset{?}{=} 0 && (12.2.8) \\ P(R \cap Q) &= 0.12 \neq 0 && (12.2.9) \end{aligned}$$

Therefore the events R and Q are not mutually exclusive.

(b) First, let us set up some notation.

Let E be the event that a student is taking Economics.

Let F be the event that a student is taking Finance.

Hence, the information given tells us that $P(E) = 0.63$, $P(F) = 0.58$ and $P(E \cap F) = 0.28$.

(i) a) We want to find $P(F' \cap E')$. To do this, we can apply some identities to simplify it into terms we do know. Let us apply Eq [3.1.1], Def [3.1.3] and Eq [3.1.5], in that order to the following.

$$\begin{aligned} P(F' \cap E') &= P((F \cup E)') = 1 - P(F \cup E) && (12.2.10) \\ &= 1 - \langle P(F) + P(E) - P(F \cap E) \rangle && (12.2.11) \\ &= 1 - [0.58 + 0.63 - 0.28] && (12.2.12) \\ &= 0.07 && (12.2.13) \end{aligned}$$

b) The probability of taking Economics only is $P(E \cap F')$ and the probability of taking Finance only is $P(F \cap E')$. Thus, we want to find $P((E \cap F') \cup (E' \cap F))$. Since the events $E' \cap F$ and $F' \cap E$ are mutually exclusive, we have,

$$\begin{aligned} P((E \cap F') \cup (E' \cap F)) &= P(E \cap F') + P(E' \cap F) && (12.2.14) \\ &= P(E) - P(E \cap F) + P(F) - P(E \cap F) && (12.2.15) \\ &= P(E) + P(F) - 2P(E \cap F) && (12.2.16) \\ &= 0.63 + 0.58 - 2(0.28) && (12.2.17) \\ &= 0.65 && (12.2.18) \end{aligned}$$

c) We want to find $P(E|F)$. To do this, we can use Def [3.1.6] for conditional probability,

$$\begin{aligned} P(E|F) &= \frac{P(E \cap F)}{P(F)} && (12.2.19) \\ &= \frac{0.28}{0.58} = 0.483 && (12.2.20) \end{aligned}$$

(ii) Let $\hat{E}$ and $\hat{F}$ be the events that a student studies Economics only and Finance only respectively. So,

$$\begin{aligned} P(\hat{E}) &= P(E \cap F') && (12.2.21) \\ &= P(E) - P(E \cap F) && (12.2.22) \\ &= 0.35 && (12.2.23) \\ P(\hat{F}) &= P(F \cap E') && (12.2.24) \\ &= P(F) - P(E \cap F) && (12.2.25) \\ &= 0.3 && (12.2.26) \end{aligned}$$

Thus, we want to find $P((\hat{E}_1 \cap \hat{F}_2) \cup (\hat{F}_1 \cap \hat{E}_2))$, where the subscript indicates the first or second student. So this reads 'the first does Economics only and the second does Finance only OR the first does Finance only and the second does Economics only'.

$$\begin{aligned} P((\hat{E} \cap \hat{F}) \cup (\hat{F} \cap \hat{E})) &= P(\hat{E} \cap \hat{F}) + P(\hat{F} \cap \hat{E}) && (12.2.27) \\ &= P(\hat{E}) \times P(\hat{F}) + P(\hat{F}) \times P(\hat{E}) && (12.2.28) \\ &= (0.35)(0.3) + (0.3)(0.35) && (12.2.29) \\ &= 0.21 && (12.2.30) \end{aligned}$$

(c) We are given a probability distribution table for a discrete random variable X.

(i) We want to find $P(X > 6)$,

$$\begin{aligned} P(X > 6) &= P(X = 7) + P(X = 8) + P(X = 9) && (12.2.31)\\ &= 0.25 + 0.12 + 0.08 && (12.2.32)\\ &= 0.45 && (12.2.33) \end{aligned}$$

(ii) We want to find $E[X]$. Using Def [3.3.2],

$$\begin{aligned} E[X] &= \sum_{\forall x} xP(X = x) && (12.2.34)\\ &= 4(0.05) + 5(0.15) + 6(0.35) + 7(0.25) + 8(0.12) + 9(0.08) && (12.2.35)\\ &= 6.48 && (12.2.36) \end{aligned}$$

4. (a) We are given that $p = 0.45$ and $n = 12$. Let X be the number of success in $n = 12$ trials.

(i) We see that X satisfies the conditions in Note [3.3.2] to be modeled by a Binomial distribution with $n = 12$ and $p = 0.45$, ie $X \sim Bin(12, 0.45)$. From Def [3.3.1], we have

$$P(X = x) = \binom{12}{x} 0.45^x 0.55^{12-x}, \quad x \in (0, 1, ..., 12) \qquad (12.2.37)$$

(ii) **a)** We want to find $P(X = 3)$. We do this by the definition above

$$\begin{aligned} P(X = 3) &= \binom{12}{3} 0.45^3 0.55^9 && (12.2.38)\\ &= 0.092 && (12.2.39) \end{aligned}$$

b) We want to find $P(X \geq 1)$. The long way would be to calculate the following,

$$P(X \geq 1) = P(X = 1) + P(X = 2) + ... + P(X = 12) \qquad (12.2.40)$$

However, it would be much quicker to recognize that

$$\begin{aligned} P(X \geq 1) &= 1 - P(X < 1) && (12.2.41)\\ &= 1 - P(X = 0) && (12.2.42)\\ &= 1 - \binom{12}{0} 0.45^0 0.55^{12} && (12.2.43)\\ &= 0.999 && (12.2.44) \end{aligned}$$

(iii) Now suppose $n = 40$. Let Y be the discrete random variable representing the number of successes 40 trials. So we have $Y \sim Bin(40, 0.45)$. Thus, we want to find $E[Y]$. We know from Eq [3.3.2]

$$E[Y] = np = 40 \times 0.45 = 18 \qquad (12.2.45)$$

(b) We are given that $X \sim N(630, 14.5^2)$

(i) We want to find $P(X > 650)$. We can proceed by standardizing

$$\begin{aligned} P(X > 650) &= P\left(\frac{X - \mu}{\sigma} > \frac{650 - 630}{14.5}\right) && (12.2.46)\\ &= P(Z > 1.379) && (12.2.47)\\ &= 1 - \Phi(1.379) && (12.2.48)\\ &= 0.084 && (12.2.49) \end{aligned}$$

(ii) Let l be the minimum weight the they must exceed in order to be considered large.[1] Thus, we want to find the l such that $P(X > l) = 0.2$. Once again, we can proceed by standardizing,

$$\begin{aligned} P(X > l) &= P\left(\frac{X - \mu}{\sigma} > \frac{l - \mu}{14.5}\right) && (12.2.50)\\ &= P\left(Z > \frac{l - \mu}{\sigma}\right) && (12.2.51)\\ &= 0.2 && (12.2.52) \end{aligned}$$

[1] Do it to cabbages and nobody blinks an eye. Do it to humans and everyone loses their mind.

Since we know $l > \mu$

$$\begin{aligned}
P\left(Z > \frac{l-\mu}{\sigma}\right) &= 1 - \Phi\left(\frac{l-\mu}{\sigma}\right) = 0.2 \tag{12.2.53}\\
\frac{l-\mu}{\sigma} &= \Phi^{-1}(0.8) \tag{12.2.54}\\
l &= \mu + 0.842\sigma \tag{12.2.55}\\
&= 630 + 0.45(14.5) \tag{12.2.56}\\
&= 642.2 \text{ grams} \tag{12.2.57}
\end{aligned}$$

(iii) We want to find $P(610 < X < 650)$. We can proceed by standardizing,

$$\begin{aligned}
P(610 < X < 650) &= P\left(\frac{610-630}{14.5} < \frac{X-\mu}{\sigma} < \frac{650-630}{14.5}\right) \tag{12.2.58}\\
&= P(-1.379 < Z < 1.379) \tag{12.2.59}\\
&= \Phi(1.379) - \Phi(-1.379) \tag{12.2.60}\\
&= 2\Phi(1.379) - 1 \tag{12.2.61}\\
&= 0.8324 \tag{12.2.62}
\end{aligned}$$

(iv) Let Y be the number of cabbages which are between 610 and 650 grams in a crop of 65 cabbages. We see that Y satisfies the conditions in Note [3.3.2] to be modeled by a binomial distribution we parameters $n = 65$ and $p = 0.8324$. Thus, we can find $E[Y]$ according to Eq [3.3.2]

$$\begin{aligned}
E[Y] &= np \tag{12.2.63}\\
= &= 65 \times 0.8324 \tag{12.2.64}\\
&= 54.08 \approx 54 \text{ cabbages} \tag{12.2.65}
\end{aligned}$$

12.3 Module 3: Analysing and Interpreting Data

5. (a) We are given that $X \sim N(420, 12.7^2)$. A sample of $n = 49$ was taken.

(i) We can comment on the distribution of $\bar{X}$ by applying the Central Limit Theorem [4.1.1].

$$\begin{aligned}
\bar{X} &\sim N\left(420, \frac{12.7^2}{49}\right) \tag{12.3.1}\\
\bar{X} &\sim N(420, 3.2916) \tag{12.3.2}
\end{aligned}$$

(ii) We want to find $P(\bar{X} < 417)$. We can proceed by standardizing,

$$\begin{aligned}
P(\bar{X} < 417) &= P\left(\frac{\bar{X}-\mu}{\sigma} < \frac{417-420}{\frac{12.7}{\sqrt{49}}}\right) \tag{12.3.3}\\
&= P(Z < -1.654) \tag{12.3.4}\\
&= 1 - \Phi(1.654) \tag{12.3.5}\\
&= 0.0495 \tag{12.3.6}
\end{aligned}$$

(b) We are given that $\sigma^2 = 0.25$, $n = 45$ and $\bar{x} = 75.9$mm.

Recall from Eq [4.1.5], that since σ^2 is known, $(1-\alpha).100\%$ confidence interval for μ is given by

$$\bar{x} \pm Z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}} \tag{12.3.7}$$

So a 98% confidence interval for μ can be computed as,

$$\begin{aligned} 75.9 &\pm z_{0.02/2}\sqrt{\frac{0.25}{45}} && (12.3.8)\\ 75.9 &\pm z_{0.01}\sqrt{\frac{0.25}{45}} && (12.3.9)\\ 75.9 &\pm 2.33\sqrt{\frac{0.25}{45}} && (12.3.10)\\ 75.9 &\pm 0.174 && (12.3.11) \end{aligned}$$

Thus, our confidence interval takes the form $(75.726, 76.074)$.

(c) We are given $\sum x = 27.2, \sum x^2 = 92.76$

(i) We can calculate an unbiased estimate for the population variance by Def [4.1.3],

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n-1}\left[\sum x^2 - \frac{(\sum x)^2}{n}\right] && (12.3.12)\\ &= \frac{1}{8-1}\left[92.76 - \frac{27.2^2}{8}\right] && (12.3.13)\\ &= 0.04 && (12.3.14) \end{aligned}$$

(ii) From Section [4.2.2], we know that when distribution is normal and the sample was randomly selected a t-test can be used.

(iii) Recall from Section [4.2.2] how to calculate a t-test at the 5% level of significance.

a) We can state the null and alternative hypothesis as
H_0: $\mu = 3.5$
H_1: $\mu < 3.5$

b) From Section [4.2.2], we write the test statistic as

$$\begin{aligned} t_{calc} &= \frac{\bar{x} - \mu}{\sqrt{\frac{\hat{\sigma}^2}{n}}} && (12.3.15)\\ &= \frac{\frac{\sum x}{n} - \mu}{\sqrt{\frac{\hat{\sigma}^2}{n}}} && (12.3.16)\\ &= \frac{\frac{27.2}{8} - 3.5}{\sqrt{\frac{0.04}{8}}} && (12.3.17)\\ &= -1.414 && (12.3.18) \end{aligned}$$

c) From Section [4.2.3], we reject H_0 if,

$$t_{calc} < -t_{\alpha}^{n-1} = -t_{0.05}^{7} = -1.895 \qquad (12.3.19)$$

d) Since $t_{calc} = -1.414 > -1.895$, it is outside the critical region. Hence, we fail to reject the null hypothesis. We conclude that the mean battery life for the laptop model is 3.5 hours.

6. (a) Recall from Section [4.3] how to conduct a χ^2 test at the 5% level of significance.

(i) We state appropriate null and alternative hypothesis as
H_0: There is no association between the sex of a student and their opinion.
H_1: There is an association between the sex of a student and their opinion.

(ii) We determine the number of degrees of freedom by definition using the given contingency table.

$$D.o.f = (r-1)(c-1) = (2-1)(3-1) = 2 \qquad (12.3.20)$$

Where r and c represent the number of rows and columns respectively.

(iii) We reject H_0 if,

$$\begin{aligned} \chi^2_{calc} &> \chi^2_{\alpha}(\nu) = \chi^2_{0.05}(2) && (12.3.21)\\ \chi^2_{calc} &> 5.991 && (12.3.22) \end{aligned}$$

(iv) We can compute the expected frequency by using the following formula,

$$E_{ij} = \frac{n_i \times n_j}{N} \tag{12.3.23}$$

Where E_{ij} represents the expected frequency of the value in the i^{th} row and j^{th} column, n_i represents the sum of values in the i^{th} row, n_j represents the sum of values in the j^{th} column and N represents the total sample size.

	In Favor	Opposed	No Opinion	Total
Female	99	70	31	200
Male	99	70	31	200
Total	198	140	62	400

Table 12.1: A table representing the expected frequencies for each cell

(v) Given that $\chi^2_{\text{calc}} = 9.534$, we see that χ_{calc} lies inside the critical region. Hence, we must reject the null hypothesis in favor of the alternative hypothesis. We conclude that there is an association between the sex of a student and their opinion.

(b) We are given 9 pairs of data points of the form (x, y)

(i) We can plot these values as a scatter diagram.

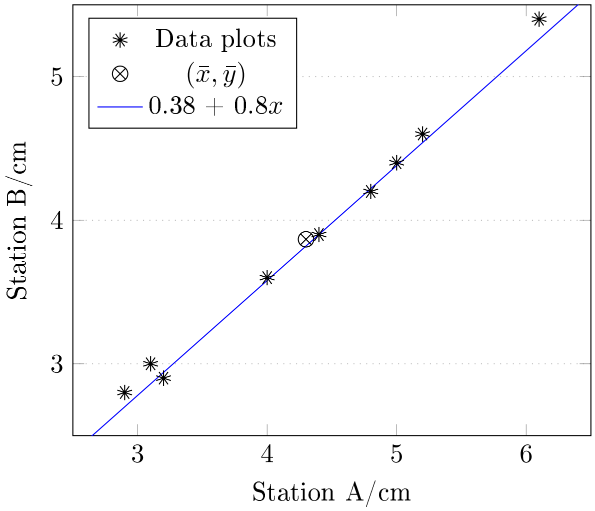


Figure 12.2: A scatter plot of the data given

(ii) We are given the regression equation as $y = 0.8x + 0.38$

a) In general, we read this as 'for every unit increase in x, y increases by 0.8 units' In context, this reads, 'for every 1cm of rainfall at Station A, the amount of rainfall at Station B to increases by 0.8cm'.

b) We want to find $\bar{x}$ and $\bar{y}$. We do this from Def [4.4.2]

$$\begin{aligned}
\bar{x} &= \frac{\sum x}{n} && (12.3.24)\\
&= \frac{5.2+4.8+6.1+5.0+3.2+2.9+4.4+4.0+3.1}{9} && (12.3.25)\\
&= 4.3 && (12.3.26)\\
\bar{y} &= \frac{\sum y}{n} && (12.3.27)\\
&= \frac{4.6+4.2+5.4+4.4+2.9+2.8+3.9+3.6+3.0}{9} && (12.3.28)\\
&= 3.867 && (12.3.29)
\end{aligned}$$

c) We draw the regression line $y = 0.8x + 0.38$ on Figure [12.2]

d) We want to find the value of y that corresponds to an x value of 4.5cm.

$$\begin{aligned}
y &= 0.8(4.5) + 0.38 && (12.3.30)\\
&= 3.98 \text{ cm} && (12.3.31)
\end{aligned}$$

Chapter 13

2013

13.1 Module 1:Collecting and Describing Data

1. **(a)** **(i)** Qualitative: descriptive

(ii) Quantitative and Discrete: can be counted

(iii) Quantitative and Continuous: can be measured and takes any value in a given range

(iv) Qualitative: descriptive

(b) **(i)** Population

(ii) Sample

(c) **(i)** Census as the entire population was asked the question.

(ii) Sample Survey as only a part of the entire population was asked.

(d) **(i)** Sample surveys are usually more cost effective and time efficient in comparison to a census.

(ii) In some experiments a census may destroy an entire population so a sample is made necessary. Consider the study 'The Effects of Pesticide on Earthworms' , where the entire population of earthworms would be a risk if a census is undertaken.

(e) **(i)** In order to use stratified random sampling, the population must be able to be grouped into various mutually exclusive strata or groups but since each student can sign up for more than one subject, stratified sampling is made inappropriate. It is also clear that there exists students doing more than one subject since the sum of students in each subject is greater than 90.

(ii) To use a random number table to select the sample of 15 students:

1. Using a role sheet or a list of all the students in the program, label the students from 10-99.

2. Start at a random number on the number table and begin choosing two digit numbers. Ignoring repetitions and ineligible numbers (01,02,03,04,05,06,07,08,09), continue choosing two digit numbers until 15 distinct numbers are chosen.

3. Now using the initial list, determine the name of the students corresponding to the numbers obtained. The names found will be the students selected as the sample.

2. **(a)** **(i)** The modal class refers to the class with the highest frequency. In the data given this corresponds to the class $66 - 68$. Hence, the boundaries of the modal class are 65.5 to 68.5.

(ii) We can assume that the heights in the class $66 - 68$ are uniformly distributed. Hence, in this class, the number of people who have more than 67 is

$$n = \left(\frac{68.5 - 67}{68.5 - 65.5} \times 10\right) + 2 \tag{13.1.1}$$

$$= 5 + 2 \tag{13.1.2}$$

$$= 7 \tag{13.1.3}$$

This corresponds to a percentage of $\frac{7}{20} \times 100 = 35\%$.

(iii) We know from Eq [2.3.2] that we can compute the mean when given grouped data as

$$\begin{aligned} \bar{x} &= \frac{\sum fx}{\sum f} && (13.1.4) \\ &= \frac{3(61)+5(64)+10(67)+2(70)}{3+5+10+2} && (13.1.5) \\ &= 65.65 && (13.1.6) \end{aligned}$$

(iv) We know from Eq [2.3.5] that we can compute the variance when given grouped data as

$$\begin{aligned} \sigma^2 &= \frac{\sum fx^2}{\sum f} - \bar{x}^2 && (13.1.7) \\ &= \frac{3(61^2)+5(64^2)+10(67^2)+2(70^2)}{3+5+10+2} - 65.65^2 && (13.1.8) \\ &= 6.7275 && (13.1.9) \\ \Rightarrow \sigma &= 2.59 && (13.1.10) \end{aligned}$$

(b) (i) The advantage is that you do not lose the individual data points.

(ii) a) We determine the size of the sample by determining how many entries there are in the stem and leaf diagram. This corresponds to $N = 26$.

b) Recall from Def [2.3.5] that we can determine the median as

$$\begin{aligned} Q_2 &= \frac{n+1}{2}^{\text{th}} \text{ term} && (13.1.11) \\ &= \frac{26+1}{2}^{\text{th}} \text{ term} && (13.1.12) \\ &= 13.5^{\text{th}} \text{ term} && (13.1.13) \\ &= 13^{\text{th}} \text{ term} + 0.5 \times (14^{\text{th}} \text{ term} - 13^{\text{th}} \text{ term}) && (13.1.14) \\ &= 29 + 0.5 \times (29 - 29) && (13.1.15) \\ &= 29 && (13.1.16) \end{aligned}$$

c) Recall from Def [2.3.5] that we can determine the lower quartile, upper quartile, and hence interquartile range as follows,

$$\begin{aligned} Q_1 &= \frac{n+1}{4}^{\text{th}} \text{ term} && (13.1.17) \\ &= \frac{27}{4}^{\text{th}} \text{ term} && (13.1.18) \\ &= 6.75^{\text{th}} \text{ term} && (13.1.19) \\ &= 6^{\text{th}} \text{ term} + 0.75 \times (7^{\text{th}} \text{ term} - 6^{\text{th}} \text{ term}) && (13.1.20) \\ &= 17 + 0.75 \times (19 - 17) && (13.1.21) \\ &= 17 + 1.5 && (13.1.22) \\ &= 18.5 && (13.1.23) \\ Q_3 &= \frac{3(n+1)}{4}^{\text{th}} \text{ term} && (13.1.24) \\ &= 20.25^{\text{th}} \text{ term} && (13.1.25) \\ &= 20^{\text{th}} \text{ term} + 0.25 \times (21^{\text{st}} \text{ term} - 20^{\text{th}} \text{ term}) && (13.1.26) \\ &= 37 + 0.25 \times (40 - 37) && (13.1.27) \\ &= 37.75 && (13.1.28) \\ IQR &= Q_3 - Q_1 && (13.1.29) \\ &= 37.5 - 18.5 && (13.1.30) \\ &= 19 && (13.1.31) \end{aligned}$$

d) We can represent the data given in a box and whisker diagram as follows

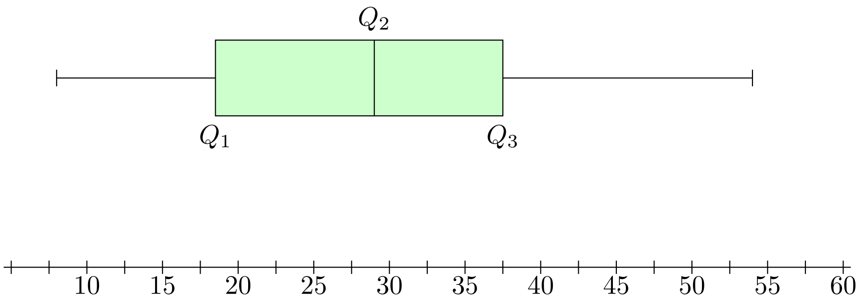


Figure 13.1: A Box and Whisker plot of the data given

e) From Def [2.3.6], we see that since

$$\begin{aligned} Q_3 - Q_2 &< Q_2 - Q_1 && (13.1.32)\\ 37.5 - 29 &< 29 - 18.5 && (13.1.33)\\ 8.5 &< 10.5 && (13.1.34) \end{aligned}$$

the distribution is negatively skewed.

13.2 Module 2: Managing Uncertainty

3. (a) We are given that $P(R) = 0.3$ and $P(T) = 0.5$ and $P(R|T) = 0.25$,

(i) a) We want to find $P(R \cap T)$. To do this, we use Def [3.1.6] of conditional probability to note that $P(A|B) = \frac{P(A \cap B)}{P(B)}$, and rearrange to get the desired term,

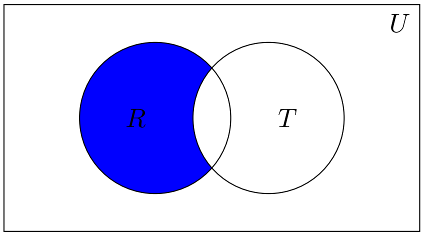
$$\begin{aligned} P(R|T) &= \frac{P(R \cap T)}{P(T)} && (13.2.1)\\ P(R \cap T) &= P(R|T) \times P(T) && (13.2.2)\\ &= 0.25 \times 0.5 = 0.125 && (13.2.3) \end{aligned}$$

b) We want to find $P(R \cup T)$. To do this, we can use Eq [3.1.5]

$$\begin{aligned} P(R \cup T) &= P(R) + P(T) - P(R \cap T) && (13.2.4)\\ &= 0.3 + 0.5 - 0.125 && (13.2.5)\\ &= 0.675 && (13.2.6) \end{aligned}$$

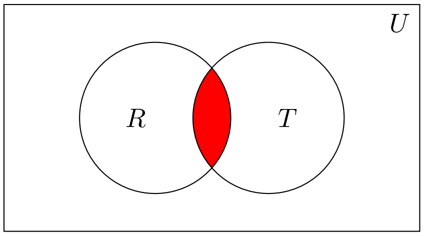
(ii) We want to draw a Venn diagram for some probabilities.

- $(R \cap T')$



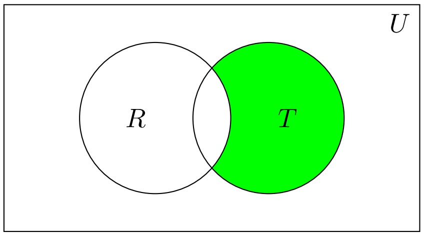
$$\begin{aligned} P(R \cap T') &= P(R) - P(R \cap T) && (13.2.7)\\ &= 0.3 - 0.125 = 0.175 && (13.2.8) \end{aligned}$$

- $(R \cap T)$



$$P(R \cap T) = 0.125 \tag{13.2.9}$$

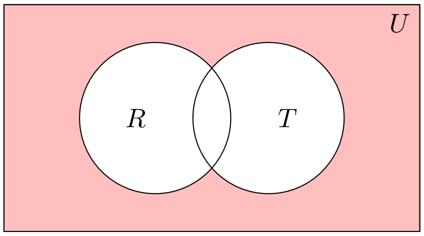
- $(R' \cap T)$



$$P(R' \cap T) = P(T) - P(R \cap T) \tag{13.2.10}$$

$$= 0.5 - 0.125 = 0.375 \tag{13.2.11}$$

- $(R \cup T)'$



$$P(R \cup T)' = P(U) - P(R \cup T) \tag{13.2.12}$$

$$= 1 - 0.675 = 0.375 \tag{13.2.13}$$

(iii) We are given that Q and T are independent, $P(T \cap Q) = 0.2$ and $P(R \cup Q) = 0.7$. Before we proceed, we should recall what independence means from Def [3.1.5]. So the independence of Q and T tell us $P(Q \cap T) = P(Q) \times P(T)$.

a) We want to find $P(Q)$. We can do this by rearranging the information we got by looking at the independence,

$$P(Q) = \frac{P(Q \cap T)}{P(T)} \tag{13.2.14}$$

$$= \frac{0.2}{0.5} = 0.4 \tag{13.2.15}$$

b) We want to show that R and Q are mutually exclusive. Recall from Def [3.1.4] that the events R and Q are mutually exclusive if

$$P(R \cap Q) = 0 \tag{13.2.16}$$

We are given $P(R \cup Q)$, which we can relate to the desired term by the following identity,

$$P(R \cup Q) = P(R) + P(Q) + P(R \cap Q) \tag{13.2.17}$$

$$P(R \cap Q) = 0.7 - 0.3 - 0.4 = 0 \tag{13.2.18}$$

(b) We are given a table that shows the results of a sample survey.

(i) We want to determine how many people were in the survey. We simply add up all the entries in the table.

$$\text{Total number of people} = 20 + 15 + 5 + 15 + 13 + 12 \tag{13.2.19}$$

$$= 80 \tag{13.2.20}$$

(ii) a) Let I be the event 'a person thought that there was an improvement in the service'. Let S be the sample space. Note that we have to consider contributions to I from both male and female participants.
Thus,

$$P(I) = \frac{|I|}{|S|} = \frac{20+15}{80} = \frac{35}{80} \tag{13.2.21}$$

b) Let F be the event 'the patron was female'. We want to find $P(F \cap \bar{I})$.

$$P(F \cap \bar{I}) = \frac{|F \cap \bar{I}|}{|S|} \tag{13.2.22}$$

$$= \frac{13}{40} = 0.4375 \tag{13.2.23}$$

c) Let M be the event 'the patron was a male'. Thus, we want to find $P(I|M)$. To do this we can use our Def [3.1.6] of conditional probability, $P(A|B) = \frac{P(A \cap B)}{P(B)}$,

$$P(I|M) = \frac{P(I \cap M)}{P(M)} \tag{13.2.24}$$

But we need to find the probabilities of the numerator and denominator first from the table. We get,

$$P(M \cap I) = \frac{20}{80} \tag{13.2.25}$$

$$P(M) = \frac{40}{80} \tag{13.2.26}$$

Plugging these equations in,

$$P(I|M) = \frac{20}{80} \div \frac{40}{80} = \frac{20}{40} = 0.5 \tag{13.2.27}$$

Alternatively, we could have noted that $P(I|M)$ tells us that the new sample space, S' is the set of all males. Now we can see directly that using this new sample space $P(I) = \frac{|I|}{|S'|} = \frac{20}{40}$.

4. (a) We are given a probability distribution for a discrete random variable X.

(i) We want to determine the value of a. We can use Proposition [3.2.1]

$$\sum_x P(X=x) = 1 \tag{13.2.28}$$

$$= P(X=0) + P(X=1) + P(X=2) + P(X=3) + P(X=4) \tag{13.2.29}$$

$$1 = 0.2 + 0.1 + a + 0.3 + 0.1 \tag{13.2.30}$$

$$= 0.7 + a \tag{13.2.31}$$

$$a = 0.3 \tag{13.2.32}$$

(ii) a) We can calculate $E[X]$ using Eq [3.2.4]

$$E[X] = \sum_x xP(X=x) \tag{13.2.33}$$

$$= (0)(0.2) + (1)(0.1) + (2)(0.3) + (3)(0.3) + (4)(0.1) \tag{13.2.34}$$

$$= 2 \tag{13.2.35}$$

b) We can calculate $Var(X)$ according to Eq [3.2.7]

$$Var(X) = E[X^2] - (E[X])^2 \tag{13.2.36}$$

To complete this we need to find $E[X^2]$. This can also be found by Def [3.2.3,]

$$E[X^2] = \sum_x x^2 P(X=x) \tag{13.2.37}$$

$$= (0^2)(0.2) + (1^2)(0.1) + (2^2)(0.3) + (3^2)(0.3) + (4^2)(0.1) \tag{13.2.38}$$

$$= 5.6 \tag{13.2.39}$$

$$\therefore Var(X) = 5.6 - 2^2 = 1.6 \tag{13.2.40}$$

c) We can determine $P(X \geq 2)$ in two ways:

$$\begin{aligned} P(X \geq 2) &= P(X=2)+P(X=3)+P(X=4) && (13.2.41)\\ &= 0.3+0.3+0.1 && (13.2.42)\\ &= 0.7 && (13.2.43) \end{aligned}$$

Or we can use the fact that $P(X \geq a) = 1 - P(X < a)$

$$\begin{aligned} P(X \geq 2) &= 1-P(X<2) && (13.2.44)\\ &= 1-(P(X=0)+P(X=1)) && (13.2.45)\\ &= 1-0.3 && (13.2.46)\\ &= 0.7 && (13.2.47) \end{aligned}$$

(b) We can note that the continuous random variable X has uniform distribution over the interval $[1, k]$.

(i) We want to find the value of A. To do this we can use Proposition [3.2.2] i.e the area under the graph is equal to one. Applying this to the uniform distribution:

$$\begin{aligned} \int_{-\infty}^{\infty} f_X.dx &= \int_1^k \frac{1}{5}.dx && (13.2.48)\\ 1 &= \frac{1}{5}(k-1) && (13.2.49)\\ k &= 6 && (13.2.50) \end{aligned}$$

Note: Less formally, but still accurate, we can recognize that we needed to make the area of the rectangle 1. Since one of the lengths was given as $\frac{1}{5}$, the other has to be 5. Further, since it starts at $x=1$, the rectangle should end at $x=6$, for its other side to have length 5.

(ii) We want to show that $P(X>2)=\frac{4}{5}$. We can do this by use of Note [3.2.1], integrating the area under f_X in the region $X>2$.

$$\begin{aligned} P(X>2) &= \int_2^{\infty} f_X.dx && (13.2.51)\\ &= \int_2^6 \frac{1}{5}.dx && (13.2.52)\\ &= \frac{1}{5}(6-2)=\frac{4}{5} && (13.2.53) \end{aligned}$$

Note: This can also be viewed as finding the area of the rectangle between $x=2$ to $x=6$. i.e $Area = 0.2 \times (6-2)$.

(c) We are given that $p=0.2$.

(i) Let X be the discrete random variable that represents the number of success in 10 trials. We see that the conditions in Note [3.3.2] are satisfied to model X by a binomial distribution. i.e $X \sim Bin(10, 0.2)$, and so, from Def [3.3.1], we have

$$P(X=x)=\binom{10}{x}0.2^x(1-0.2)^{10-x}, \quad x \in (0,1,...,10) \qquad (13.2.54)$$

Now we can easily find $P(X=3)$,

$$\begin{aligned} P(X=3) &= \binom{10}{3}0.2^3 0.8^7 && (13.2.55)\\ &= 0.201 && (13.2.56) \end{aligned}$$

(ii) Let Y be the discrete random variable that represents the number of successes in 500 trials. We see that the conditions are satisfied to model Y by a binomial distribution. Thus, we want to find $E[Y]$. According to Eq [3.3.2]

$$\begin{aligned} E[Y] &= np && (13.2.57)\\ &= 500 \times 0.2 = 100 && (13.2.58) \end{aligned}$$

(iii) We want to find $P(Y > 90)$. Since

$$\begin{aligned} np(1-p) &= (500)(0.2)(0.8) && (13.2.59)\\ &= 80 \geq 5 && (13.2.60)\\ np &= 500(0.2) && (13.2.61)\\ &= 100 > 5 && (13.2.62)\end{aligned}$$

then according to Note [3.4.2], we can use the normal approximation to the binomial distribution. Therefore, we can say that Y is normally distributed with $\mu = 100$ and $\sigma^2 = 80$, $Y \sim N(100, 80)$. We apply the continuity correction according to Note [3.4.3] as follows

$$P(Y > 90) \rightarrow P(Y > 90.5) \qquad (13.2.63)$$

We can proceed by standardizing Y as follows:

$$\begin{aligned} P(Y > 90.5) &= P\left(\frac{Y-100}{\sqrt{80}} > \frac{90.5-100}{\sqrt{80}}\right) && (13.2.64)\\ &= P\left(Z > -1.06213\right) && (13.2.65)\\ &= P(Z < 1.06213) && (13.2.66)\\ &= \Phi(1.06213) && (13.2.67)\\ &= 0.8554 && (13.2.68)\end{aligned}$$

13.3 Module 3: Analyzing and Interpreting Data

5. (a) We are given that $X \sim N(35, 10^2)$

(i) To determine the distribution of the sample mean $\bar{X}$ we can apply the Central Limit Theorem [4.1.1].

$$\bar{X} \sim \mathbf{N}\left(35, \frac{10^2}{8}\right) \qquad (13.3.1)$$

(ii) We can calculate $P(\bar{X} \geq 38.7)$ by standardizing,

$$\begin{aligned} P(\bar{X} \geq 38.7) &= P\left(\frac{\bar{X}-\mu}{\sigma} \geq \frac{38.7-35}{\sqrt{\frac{10^2}{8}}}\right) && (13.3.2)\\ &= P(Z \geq 1.047) && (13.3.3)\\ &= 1-\Phi(1.047) && (13.3.4)\\ &= 0.1469 && (13.3.5)\end{aligned}$$

(b) We are given that $\bar{x} = 4.2$ and $s = 1.5$ when $n = 85$.

(i) Recall from Eq [4.1.6], that when σ^2 is unknown and $n \geq 30$, a 94% confidence interval for the true mean mass, μ, of the packages takes the form

$$\begin{aligned} \bar{x} &\pm Z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}} && (13.3.6)\\ 4.2 &\pm Z_{0.03}\sqrt{\frac{1.5^2}{85}} && (13.3.7)\\ 4.2 &\pm 1.88\sqrt{\frac{1.5^2}{85}} && (13.3.8)\end{aligned}$$

Thus, the confidence interval for μ is $(3.894, 4.506)$

(ii) Since, by definition, a 94% confidence interval means that if confidence intervals are constructed across many separate data analyses of replicated experiments, the proportion of such intervals that contain the true value of the parameter will be 0.94, then we expect $40 \times 0.94 = 37.6 \approx 38$ intervals to contain μ.

(c) (i) A t-test would be appropriate since the sample size n is less than 30 and the variance of the population σ^2 is unknown.

(ii) We write the appropriate null and alternative hypothesis as

$$H_0 \quad : \quad \mu = 10 \tag{13.3.9}$$

$$H_1 \quad : \quad \mu < 10 \tag{13.3.10}$$

(iii) The next step in conducting the t-test is to determine the critical region. Recall from Section [4.2.3] that at the 5% level of significance, with $n-1$ or 9 degrees of freedom, we reject H_0 if

$$t_{\text{calc}} \quad < \quad -t_{\alpha}^{(n-1)} = t_{0.05}^{9} \tag{13.3.11}$$

$$t_{\text{calc}} \quad < \quad -1.833 \tag{13.3.12}$$

Next, we need to calculate the test statistic. From Section [4.2.2], we see that

$$t_{\text{calc}} \quad = \quad \frac{\bar{x}-\mu}{\sqrt{\frac{\hat{\sigma}^2}{n}}} \tag{13.3.13}$$

$$= \quad \frac{9.6-10}{\sqrt{\frac{6^2}{8}}} \tag{13.3.14}$$

$$= \quad -0.189 \tag{13.3.15}$$

Since our test statistic, $t_{\text{calc}} = -0.189 < -1.833$, it lies within the critical region. Hence, we must reject the null hypothesis in favor of the alternative hypothesis. We conclude that at the 5% level of significance, mean amount of coffee dispensed is less than 10 ounces.

6. **(a)** We are given the regression equation $y = 6.6 + 0.09x$

(i) The value of 0.09 tells us that for every year of the foreman assigned to each project, we can expect to see an increase of 0.09 dollars per sales dollar earned.

(ii) The value of 6.6 in this equation tells us that if the foreman had 0 years, we would still expect to see a profit of 6.6 dollars per sales dollar earned.

(iii) We want to estimate the value of y when $x = 15$.

$$y \quad = \quad 6.6 + 0.09(15) = 7.95 \tag{13.3.16}$$

(iv) We know from Note [4.4.1], that when $r = 0.32$, this means that there is a low positive linear correlation.

(b) Recall from Section [4.3] how to compute a χ^2 test at the 5% level of significance.

(i) We state the appropriate null and alternative hypothesis as:
H_0: There is no association between the grades obtained by the students and the teachers who taught the course.
H_1: There is an association between the grades obtained by the students and the teachers who taught the course.

(ii) We want to complete the following table which shows the observed (O) and the expected (E) frequencies of the grades obtained in the examinations.

	O	E	O	E	O	E	Total
T 1	18	18.333	12	10	20	21.667	50
T 2	26	25.667	12	14	32	30.333	70
Total	44	44	24	24	52	52	120

(iii) a) We can calculate this by definition using the given contingency table.

$$D.o.f = (r-1)(c-1) = (2-1)(3-1) = 2 \tag{13.3.17}$$

Where r and c represent the number of rows and columns respectively.

b) the critical region of the test.
We reject the null hypothesis if $\chi^2_{calc} > \chi^2_{\alpha}(\nu) = \chi^2_{0.05}(2)$,

$$\chi^2_{calc} > 5.991 \tag{13.3.18}$$

c) the χ^2 test statistic.
The test statistic is given by $\sum \frac{(O-E)^2}{E}$. It is easiest to represent this in a table,

O	E	$\frac{(O-E)^2}{E}$
18	18.333	0.006
26	25.667	0.004
12	10.000	0.400
12	14.000	0.286
20	21.668	0.128
32	30.333	0.092
	Total	0.916

(iv) Since our test statistic lies outside our critical region, $\chi^2_{\text{calc}} = 0.916 < 5.991$, we fail to reject the null hypothesis. Thus, we conclude that there is no association between the grades obtained b the students and the teachers who taught the course.

Chapter 14

2014

14.1 Module 1: Collecting and Describing Data

1. We are given information in a tabular format.

(a) **(a)** The 15 indicates the number of employees in the children's clothing department.

(b) The 3 indicates the average weekly sales of 30000 for the 'Accessories' department.

(b) **(i)** Recall from Def [2.2.2] that this describes simple random sampling.

(ii) Recall from Def [2.2.4] that this describes systematic sampling.

(iii) Recall from Def [2.2.3] that this describes stratified sampling.

(c) The method in (ii), stratified sampling, is most representative of the employees from EACH department.

(d) Using the method in 1(b)(iii), we want to calculate the number of employees in the sample that will be drawn from the Jewellery and Perfume department.

The total number of employees, N_{tot}, is given by:

$$\begin{aligned} N_{tot} &= 25+30+15+12+8+10 && (14.1.1)\\ &= 100 && (14.1.2) \end{aligned}$$

We are required to randomly select x employees from the department based on the proportion size of the department. Thus the following ratio needs to be satisfied:

$$\begin{aligned} \frac{10}{100} &= \frac{x}{20} && (14.1.3)\\ \Rightarrow x &= 2 && (14.1.4) \end{aligned}$$

(e) **(i)** Let x_i be the average weekly sales of the ith department [1]. We use Def [2.3.1] to determine the mean, $\bar{x}$,

$$\begin{aligned} \bar{x} &= \frac{12+15+8+4+3+7}{6} \times 10000 && (14.1.5)\\ &= 81666.7 \approx 82000 && (14.1.6) \end{aligned}$$

(ii) Let σ^2 be the variance of the average sales of the departments in the entire store. From

[1] Technically speaking this is the average weekly sales PER DEPARTMENT and not for the ENTIRE store. If we wanted to compute the average weekly sales for the entire store we would need to add up what they gave us per department and not divide. Furthermore we would require more information than what was provided

Def [2.3.2], we can compute σ^2 as follows

$$\begin{align*}
\sigma^2 &= \frac{\sum(\bar{x}-x)^2}{n} \tag{14.1.7}\\
&= \frac{1}{6}\times((12-8.16)^2+(15-8.16)^2+(8-8.16)^2+(4-8.16)^2 \tag{14.1.8}\\
&+ (3-8.16)^2+(7-8.16)^2 \tag{14.1.9}\\
&= 178056.00 \tag{14.1.10}\\
\sigma &\approx 422.00 \tag{14.1.11}
\end{align*}$$

(f) Let t be the time in minutes. Let $F(t)$ be the number of shoppers that spent time $\leq t$.

(i) Let $N_{30<t\leq 60}$ be the number of shoppers who spent between 30 and 60 minutes. Thus,

$$\begin{align*}
N_{30<t\leq 60} &= F(60)-F(30) \tag{14.1.12}\\
&= 41-7=34 \tag{14.1.13}
\end{align*}$$

(ii) We want to find the value of t such that $F(t)=0.6\times 50=30$. Reading this value off,

$$t = 53 \text{ minutes} \tag{14.1.14}$$

(iii) Recall the median from Def [2.3.5]. We want to find the value Q_2 such that $F(Q_2)=25$. Reading this off,

$$Q_2 = 50 \text{ minutes} \tag{14.1.15}$$

(iv) We want to draw a box-and-whisker diagram to represent the information given in the cumulative frequency graph. We first need to determine Q_1 and Q_3 according to Def [2.3.5],

$$\begin{align*}
F(Q_1) &= 0.25\times 50 \tag{14.1.16}\\
&= 12.5 \tag{14.1.17}\\
\Rightarrow Q_1 &= 40 \text{ minutes} \tag{14.1.18}\\
F(Q_3) &= 0.75\times 50 \tag{14.1.19}\\
&= 37.5 \tag{14.1.20}\\
\Rightarrow Q_3 &= 58 \text{ minutes} \tag{14.1.21}
\end{align*}$$

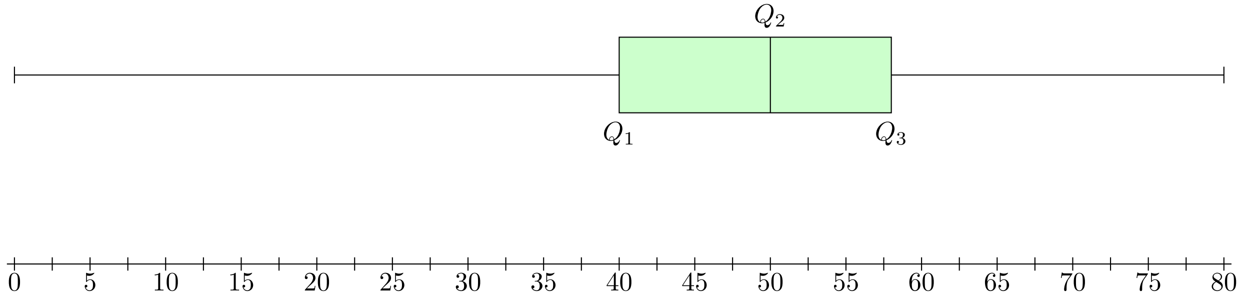


Figure 14.1: A Box and Whisker plot of the data given

(v) Recall from Def [2.3.6] that since,

$$\begin{align*}
Q_2-Q_1 &> Q_3-Q_2 \tag{14.1.22}\\
50-40 &> 58-50 \tag{14.1.23}\\
10 &> 8 \tag{14.1.24}
\end{align*}$$

we can conclude that the distribution is negatively skewed.

2. We are given a table with grouped data.

(a) **(i)** The third class is $30-39$. Thus, the boundaries are given by

$$29.5-39.5 \tag{14.1.25}$$

(ii) We can calculate the size of the class by looking at its boundaries.

$$\begin{aligned} \text{Size} &= 39.5 - 29.5 && (14.1.26) \\ &= 10 && (14.1.27) \end{aligned}$$

(iii) Presenting data as a grouped frequency presents the disadvantage that the exact values are lost, and hence further statistics have to be estimated.

(b) **(i)** Let x_i be the midpoint value for the time in each class. We can use Eq [2.3.2 to determine the mean from grouped data,

$$\begin{aligned} \bar{x} &= \frac{\sum f_i x_i}{\sum f_i} && (14.1.28) \\ &= \frac{1}{50} \times [3(14.5) + 14(24.5) + 22(34.5) + 10(44.5) + 1(54.5)] && (14.1.29) \\ &= 32.9 \text{ minutes} && (14.1.30) \end{aligned}$$

(ii) We use Def [2.3.2] to determine the variance and again, remember to multiply by f_i for each class

$$\begin{aligned} \sigma^2 &= \frac{1}{N} \sum_{i=1}^{N} f_i (x_i - \bar{x})^2 && (14.1.31) \\ &= \frac{1}{50} \times (3(14.5 - 32.9)^2 + 14(24.5 - 32.9)^2 + 22(34.5 - 32.9)^2 && (14.1.32) \\ &+ 10(44.5 - 32.9)^2 + 1(54.5 - 32.9)^2) && (14.1.33) \\ &= 77.4 && (14.1.34) \end{aligned}$$

(iii) We can now easily compute the standard deviation as

$$\sigma = \sqrt{77.4} = 8.8 \qquad (14.1.35)$$

(c) We want to draw a histogram to represent the information given in the table.

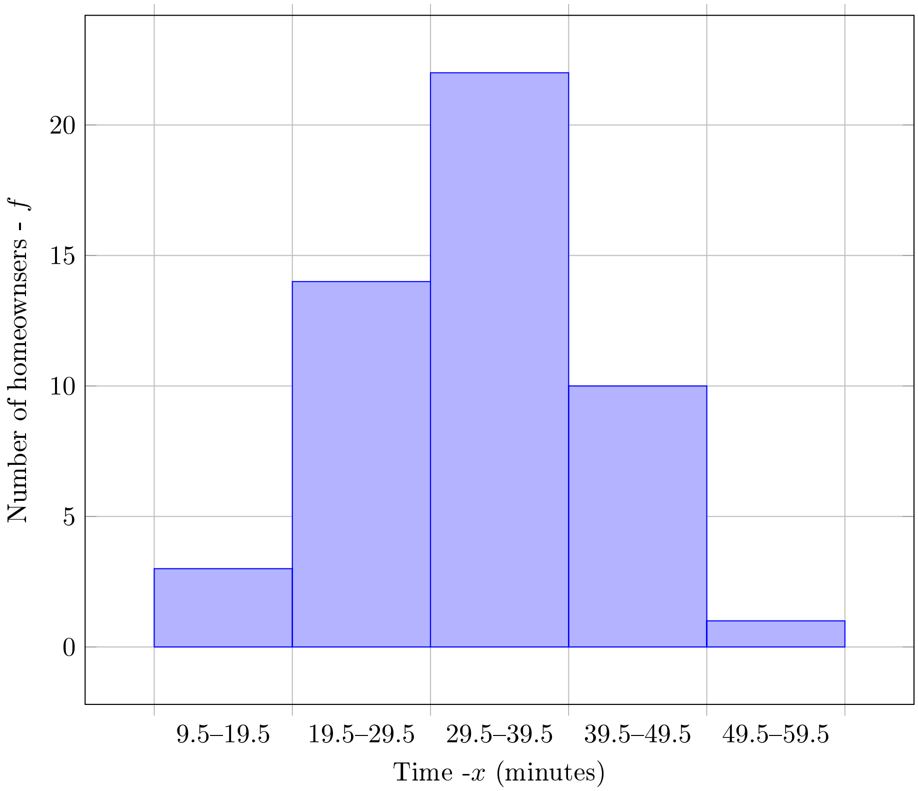


Figure 14.2: A histogram of the grouped data provided

(d) **(i)** The mode can be estimated by the midpoint value of the class with the largest f. We can read this from the table or histogram.

$$\text{Mode} = 34.5 \tag{14.1.36}$$

(ii) Since this is an estimate, we can average the number of people within the class $40 - 49$ that lasted 45 minutes or more as

$$10 \times \frac{49.5 - 45}{49.5 - 39.5} = 4.5 \tag{14.1.37}$$

Since there is 1 person in the class $50 - 59$, the number of people that lasted 45 minutes or more is

$$4.5 + 1 = 5.5 \approx 6 \tag{14.1.38}$$

14.2 Module 2: Managing Uncertainty

3. **(a)** We are given information on probabilities.

(i) We want to draw a tree diagram to show this information.
Let A be the event 'machine A was used'.
Let B be the event 'machine B was used'.
Let D be the event 'an item was defective'.
We can construct the tree diagram as follows:

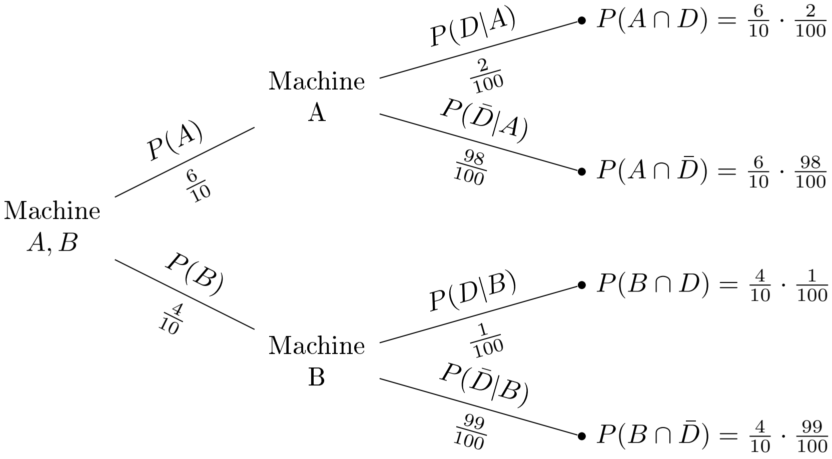

Machine A, B
P(A)
$\frac{6}{10}$
P(B)
$\frac{4}{10}$
Machine A
Machine B
P(D\A)
$\frac{2}{100}$
$P(\bar{D}|A)$
$\frac{98}{100}$
P(D\B)
$\frac{1}{100}$
$P(\bar{D}|B)$
$\frac{99}{100}$
$P(A \cap D) = \frac{6}{10} \cdot \frac{2}{100}$
$P(A \cap \bar{D}) = \frac{6}{10} \cdot \frac{98}{100}$
$P(B \cap D) = \frac{4}{10} \cdot \frac{1}{100}$
$P(B \cap \bar{D}) = \frac{4}{10} \cdot \frac{99}{100}$

Figure 14.3: A tree diagram of the events and associated probabilities.

(ii) We want to find $P(A \cap D)$
From the tree diagram, we see that:

$$P(A \cap D) = \frac{6}{10}.\frac{2}{100} \tag{14.2.1}$$

We can also see this from Def [3.1.6] of conditional probability

$$P(D|A) = \frac{P(D \cap A)}{P(A)} \tag{14.2.2}$$

$$P(D \cap A) = P(A).P(D|A) \tag{14.2.3}$$

$$P(A \cap D) = \frac{6}{10}.\frac{2}{100} = 0.012 \tag{14.2.4}$$

(iii) We want to find $P(D)$.
Since a defective item can either be produced by machine A or B, we need to use Def [3.1.6] for conditional probability.

$$\begin{aligned} P(D) &= P(A \cap D) + P(B \cap D) && (14.2.5)\\ &= P(A).P(D|A) + P(B).P(D|B) && (14.2.6)\\ &= \frac{6}{10}.\frac{2}{100} + \frac{4}{10}.\frac{1}{100} && (14.2.7)\\ &= 0.016 && (14.2.8) \end{aligned}$$

(iv) We want to find $P(A|D)$
To do this we can use Def [3.1.6] for conditional probability,

$$\begin{aligned} P(A|D) &= \frac{P(A \cap D)}{P(D)} && (14.2.9)\\ &= \frac{0.012}{0.016} = 0.75 && (14.2.10) \end{aligned}$$

(v) We want to find the probability that EXACTLY ONE of item is defective if two are chosen at random.
We note that if the random variable X represents the number of defective items in two randomly selected items, we can say X has a binomial distribution with parameters $n = 2$ and $p = 0.016$ (where p was calculated in part (iii)). i.e $X \sim Bin(2, 0.016)$ Thus we need to find $P(X = 1)$,

$$\begin{aligned} P(X = x) &= \binom{2}{x}(0.016)^x(1 - 0.016)^{2-x}, \quad x \in (0, 1, 2) && (14.2.11)\\ P(X = 1) &= \binom{2}{1}(0.016)(1 - 0.016) = 0.0315 && (14.2.12) \end{aligned}$$

(b) We are given that $P(M) = 0.6$, $P(M \cap N) = 0.2$, $P(M \cup N) = 0.85$.

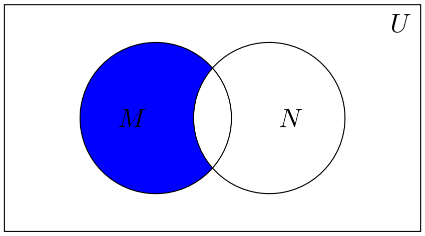
(i) We want to calculate $P(N)$. To do this we can use the identity Eq [3.1.5] and rearrange to get the term we need.

$$\begin{aligned} P(M \cup N) &= P(M) + P(N) - P(M \cap N) && (14.2.13)\\ P(N) &= P(M \cup N) + P(M \cap N) - P(M) && (14.2.14)\\ &= 0.85 + 0.2 - 0.6 && (14.2.15)\\ &= 0.45 && (14.2.16) \end{aligned}$$

(ii) We want to find $P(N|M)$. To do this we use Def [3.1.6] of conditional probability,

$$\begin{aligned} P(N|M) &= \frac{P(N \cap M)}{P(M)} && (14.2.17)\\ &= \frac{0.2}{0.6} && (14.2.18)\\ &= \frac{1}{3} && (14.2.19) \end{aligned}$$

(iii) We want to find $P(M \cap N')$.
First we should visualize what this is on a Venn diagram. The shaded region below represents the region $P(M \cap N')$



Hence we can calculate the shaded region, as:

$$\begin{aligned} P(M \cap N') &= P(M) - P(M \cap N) && (14.2.20)\\ &= 0.6 - 0.2 && (14.2.21)\\ &= 0.4 && (14.2.22) \end{aligned}$$

(c) **(i)** To determine if M and N are mutually exclusive we must determine if $P(M \cap N) = 0$ according to Def [3.1.4]. But the data given in the problem tells us $P(M \cap N) = 0.2 \neq 0$. Hence M and N are not mutually exclusive.

(ii) To determine if M and N are independent, we need to determine if $P(M \cap N) = P(M) \times P(N)$, according to Def [3.1.5],

$$\begin{aligned} P(M \cap N) &\overset{?}{=} P(M) \times P(N) && (14.2.23)\\ 0.2 &\overset{?}{=} 0.6 \times 0.45 && (14.2.24)\\ 0.2 &\neq 0.27 && (14.2.25)\end{aligned}$$

Thus M and N are not independent events.

4. **(a)** Recall from Note [3.3.2] that are three conditions necessary to model a binomial distribution:

(i) The experiment consist of a fixed number of trails n.
(ii) The trials are independent.
(iii) Each trail can be classified as a success or failure.
(iv) The probability of success, p, is constant

(b) Let X be a binomial random variable with $n = 12$ and $p = 0.6$. We summarize this as $X \sim \text{Bin}(12, 0.6)$.

(i) We want to calculate $P(X = 3)$. From Def [3.3.1], we know

$$P(X = 3) = \binom{12}{3}(0.6)^3(0.4)^9 = 0.0125 \quad (14.2.26)$$

(ii) WE want to calculate $P(X \geq 2)$. The lengthy calculation would be to do the following:

$$\begin{aligned} P(X \geq 2) &= P(X = 2) + P(X = 3) + P(X = 4) + ... && (14.2.27)\\ &+ P(X = 11) + P(x = 12) && (14.2.28)\end{aligned}$$

However it will be much quicker if you recognize

$$\begin{aligned} P(X \geq 2) &= 1 - P(X < 2) && (14.2.29)\\ &= 1 - (P(X = 0) + P(X = 1)) && (14.2.30)\\ &= 1 - \binom{12}{0}(0.6)^0(0.4)^{12} - \binom{12}{1}(0.6)^1(0.4)^{11} = 0.9997 \approx 1 && (14.2.31)\end{aligned}$$

(c) Let X be the number of days that an insect lives. We summarize its distribution as $X \sim \text{N}(72, 8^2)$ Thus, we want to find $P(X > 84)$. Standardizing, we have

$$\begin{aligned} P(X > 84) &= P\left(\frac{X - 72}{8} > \frac{84 - 72}{8}\right) && (14.2.32)\\ &= P(Z > 1.5) && (14.2.33)\\ &= 1 - P(Z \leq 1.5) && (14.2.34)\\ &= 1 - \Phi(1.5) = 1 - 0.9332 && (14.2.35)\\ &= 0.0668 && (14.2.36)\end{aligned}$$

(d) Let X be the number of success in 200 trials.

We can see that X satisfies the conditions to be modeled by a binomial distribution, with parameters $n = 200$ and $p = 0.82$. We summarize its distribution as $X \sim \text{Bin}(200, 0.82)$.

(i) For a binomial distribution, we can determine $E[X]$ according to Eq [3.3.2]

$$\begin{aligned} E[X] &= (200)(0.82) && (14.2.37)\\ &= 164 && (14.2.38)\end{aligned}$$

(ii) For a binomial distribution, we can compute the $Var(X)$ from Eq [3.3.3],

$$\begin{aligned} Var(X) &= (200)(0.82)(1.8) && (14.2.39)\\ &= \frac{738}{25} = 29.52 && (14.2.40)\\ \sigma &\approx 5.43 && (14.2.41)\end{aligned}$$

(iii) We can use the normal approximation to the binomial distribution since the conditions in Note [3.4.2] are satisfied:

$$np = 164 > 5 \tag{14.2.42}$$

$$npq = 5.43 > 5 \tag{14.2.43}$$

Thus, X is normally distributed with parameters $\mu = 164$ and $\sigma = 5.43$, or $X \sim \mathrm{N}(164, 5.43^2)$. We first apply a continuity correction according to Note [3.4.3]

$$P\,(X < 175) \to (X < 174.5) \tag{14.2.44}$$

and then proceed by standardizing,

$$P\,(X < 174.5) = P\left(\frac{X - 164}{5.43} < \frac{174.5 - 164}{5.43}\right) \tag{14.2.45}$$

$$= P\,(Z < 1.9337) \tag{14.2.46}$$

$$= \Phi(1.9337) \tag{14.2.47}$$

$$= 0.9732 \tag{14.2.48}$$

14.3 Module 3: Analyzing and Interpreting Data

5. (a) We are given a sample of size $n = 12$.

(i) Recall from Def [4.1.1] that whenever we have a sample it is useful to calculate unbiased estimates.

a) We want to calculate the unbiased estimate $\hat{\mu}$ according to Def [4.1.2],

$$\bar{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{14.3.1}$$

$$= \frac{3 + 13 + 5 + 8 + 12 + 5 + 6 + 4 + 8 + 7 + 10 + 8}{12} \tag{14.3.2}$$

$$= 7\frac{5}{12} \tag{14.3.3}$$

b) Recall from Def [4.1.3] that we can determine $\hat{\sigma^2}$ by,

$$\hat{\sigma^2} = \frac{1}{n-1}\left[\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right] \tag{14.3.4}$$

$$\sum x^2 = 3^2 + 13^2 + 5^2 + 8^2 + 12^2 + 5^2 + 6^2 + 4^2 \tag{14.3.5}$$

$$+ \; 8^2 + 7^2 + 10^2 + 8^2 = 765 \tag{14.3.6}$$

$$\left(\sum x\right)^2 = 89^2 \tag{14.3.7}$$

$$\therefore \hat{\sigma^2} = \frac{1}{11}\left[765 - \frac{89^2}{12}\right] \tag{14.3.8}$$

$$= 9.54 \tag{14.3.9}$$

(ii) a) We can state the null and alternative hypothesis as follows,

$$H_0 : \mu = 7 \tag{14.3.10}$$

$$H_1 : \mu > 7 \tag{14.3.11}$$

b) Recall from Section [4.2.3], that for a one-tailed t-test at the 5% level of significance, we reject H_0 if

$$t_{\text{calc}} > t_{\alpha}^{(n-1)} = t_{0.05}^{11} \tag{14.3.12}$$

$$\Rightarrow t_{\text{calc}} > 1.796 \tag{14.3.13}$$

c) Recall from Section [4.2.2], that we can determine the test statistic by

$$t = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \quad (14.3.14)$$

$$= \frac{7\frac{5}{12}-7}{\sqrt{\frac{9.54}{12}}} \quad (14.3.15)$$

$$= 0.467 \quad (14.3.16)$$

d) Since $t_{\text{calc}} = 0.467 < 1.796$, it lies outside of the critical region and hence we fail to reject H_0.

e) We assumed that the random variable followed a normal distribution and that the population variance was unknown.

(b) We know from Eq [4.1.5] that we can compute a $(1-\alpha)\%$ confidence interval for the population proportion p as

$$p_s \pm z_{\alpha/2}\sqrt{\frac{p_s q_s}{n}} \quad (14.3.17)$$

$$\frac{18}{52} \pm z_{0.05/2}\sqrt{\frac{\frac{18}{52}\frac{34}{52}}{52}} \quad (14.3.18)$$

$$\frac{18}{52} \pm 0.196(0.06597) \quad (14.3.19)$$

$$\Rightarrow (0.217, 0.475) \quad (14.3.20)$$

6. (a) Recall from Section [4.3] how to conduct a χ^2 test at the 5% level of significance.

(i) We can state the appropriate null and alternative hypothesis as
H_0: The teacher's prediction and the actual results are independent.
H_1: The teacher's prediction and the actual results are not independent.

(ii) a) We can calculate the number of degrees of freedom by definition using the given contingency table.

$$D.o.f = (r-1)(c-1) \quad (14.3.21)$$

$$= (3-1)(3-1) = 4 \quad (14.3.22)$$

Where r and c represent the number of rows and columns respectively.

b) We determine the critical region this by reading off the value in the table that corresponds to 4 degrees of freedom and 5% level of significance. From the table we get 9.488. Therefore, we can reject H_0 in favor of H_1 if $\chi^2 > 9.488$.

(iii) We can compute the expected frequency by using the following formula,

$$E_{ij} = \frac{n_i \times n_j}{N} \quad (14.3.23)$$

Where E_{ij} represents the expected frequency of the value in the i^{th} row and j^{th} column, n_i represents the sum of values in the i^{th} row, n_j represents the sum of values in the j^{th} column and N represents the total sample size. Thus, we want to find E_{32},

$$E_{32} = \frac{n_3 \times n_2}{N} \quad (14.3.24)$$

$$= \frac{53 \times 65}{150} \quad (14.3.25)$$

$$= 22.97 \approx 23.0 \quad (14.3.26)$$

(iv) Since $\chi^2 = 9.1625 < 9.488$, we fail to reject the null hypothesis. Thus, at the 5% level of significance, we can conclude that he teachers prediction and the actual results are independent.

(b) We are given that $\sum x = 377$, $\sum y = 297$, $\sum xy = 11305$, $\sum x^2 = 14397$, $\sum y^2 = 9145$.

(i) Recall from Def [4.4.2], that a line of linear regression of Y on X has the form

$$y = a + bx \quad (14.3.27)$$

where b is computed as

$$b = \frac{n\sum xy - \sum x.\sum y}{n\sum x^2 - (\sum x)^2} \quad (14.3.28)$$

$$= \frac{11305 - \frac{1}{10}(337)(297)}{14397 - \frac{1}{10}(337)^2} \quad (14.3.29)$$

$$= 0.587 \quad (14.3.30)$$

$$(14.3.31)$$

Now that we have b, we can compute a as,

$$a = \bar{y} - b\bar{x} \quad (14.3.32)$$

$$\text{where } \bar{y} = \frac{\sum y}{n} \text{ and } \bar{x} = \frac{\sum x}{n} \quad (14.3.33)$$

$$\Rightarrow a = \frac{297}{10} - 0.587\frac{377}{10} \quad (14.3.34)$$

$$= 7.56 \quad (14.3.35)$$

So we have the following regression equation:

$$y = 7.56 + 0.587x \quad (14.3.36)$$

(ii) We want to find the approximate y that would correspond to a x of 37. We can do this using the regression equation above,

$$y = 7.56 + 0.587(37) \quad (14.3.37)$$

$$= 29.3 \quad (14.3.38)$$

(iii) b is the gradient of the regression line. Thus an increase in x by one unit corresponds to an increase in y by 0.587. In terms of the information, an increase of 1 mark in the aptitude test corresponds to an increase of 0.587 on the productivity score.

(iv) Recall from Def [4.4.1] that we can calculate the product-moment correlation coefficient r as

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}} \quad (14.3.39)$$

$$= \frac{10(11305) - (377)(297)}{\sqrt{10(14397) - (377^2)}\sqrt{10(9145) - (297^2)}} \quad (14.3.40)$$

$$= 0.443 \quad (14.3.41)$$

Thus, we can state that the mark on the aptitude test has a moderate to weak and positive correlation with the productivity score.

Chapter 15

2015

15.1 Module 1: Collecting and Describing Data

1. **(a)** **(i)** The key word is “type”, indicating that it is a description and hence the data is *qualitative.*

   (ii) The key word is “weight”, indicating that the data is an ascribed number and hence it is *quantitative.*

   (iii) The key word is “seats”. Since you can only have an integer number of seats, the data is *discrete.*

   (iv) The key word is “kilowatt-hours”. Since you need to measure this, and the value can be any number of hours can be any number within a range eg. 2.001, 45.035 etc, the data is *continuous.*

   (b) **(i)** A ‘census’ collects data from the entire population. In contrast, a ‘sample survey’ does not and usually collects data from a representative group of the population.

   (ii) A ‘sample’ is usually used when it is not possible or practical to conduct a ‘census’. For example, when a company is testing the lifetime of light bulbs it would be impractical to use all their light bulbs. Additionally, a sample has the advantage that it is usually cheaper to execute and requires less resources.

   (iii) **a)** We see that only a fraction of the entire population of customers, 15 out of 40, is used. Hence a survey was conducted.

   b) We see that the entire population, all of the 5 persons, was used to collect data. Hence a census was conducted.

   (c) **(i)** Direct observation, since it is more time efficient to simply wait by the entrance and collect the data

   (ii) Personal survey, as it is impractical to do a direct observation.

   (iii) Personal survey, as it is impractical to do a direct observation.

   (d) **(i)** We need to keep the ratio of masons in this sample representative of the ratio of masons in the entire population. So first let us calculate the latter. The total number of persons in the company is

$$\text{Total} = 5 + 10 + 45 + 30 + 10 + 25 \qquad (15.1.1)$$

$$= 125 \qquad (15.1.2)$$

Hence the fraction of masons in the entire population is $\frac{45}{125}$. To maintain this ratio in the sample, we need the number of masons, m, to be such that the following equation is satisfied.

$$\frac{m}{25} = \frac{45}{125} \qquad (15.1.3)$$

$$\Rightarrow m = \frac{45}{125} \times 25 \qquad (15.1.4)$$

$$= 9 \qquad (15.1.5)$$

(ii) We avoid losing data from an under-represented group, such as the engineers, had a simple random sample been taken.

(iii) (a) The question uses the word 'or', indicating that you can either answer Q1: *'how often do you buy lunch at the work site'* or Q2: *'do you bring lunch from home every day'.* Answering one of them does not provide information about the other in this case. To illustrate, if you chose Q1 and answered '0', you do not know if this is because the person brings lunch from home everyday or buys food off the work site. Similarly, if you chose Q2 and said 'no', you have no idea on how many times this person buys lunch at the work site.

One can also say that an open ended question was combined with a closed ended question which can lead to respondents only partially answering either question.

(b) Since the only reason you would need to know how often the person buys lunch at the work site is if they did not bring their own lunch everyday, a suitable question might be: *"Do you bring lunch from home every day? If not, how often do you buy lunch at the work site?"*

2. (a) We are given grouped data in a tabular form.

(i) When displaying data in groups, we lose the exact values of the data points and hence further statistics have to be estimated.

(ii) The third class represents the range $60 - 74$. So the size of this class is

$$\begin{aligned} \text{Upper Class Boundary} - \text{Lower Class Boundary} &= 74.5 - 59.5 && (15.1.6)\\ &= 15 && (15.1.7)\end{aligned}$$

(iii) We can draw a histogram to represent the information given.

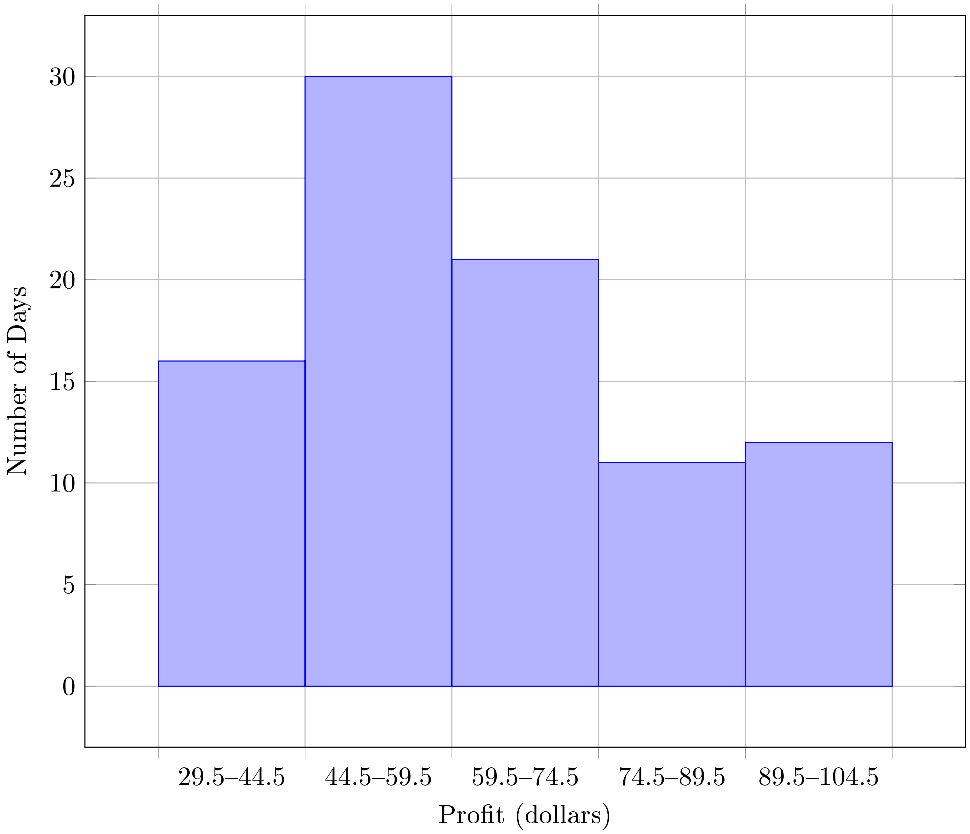


Figure 15.1: A histogram representing the grouped data provided

(iv) Since the mode is the value that occurs most frequently, we see that the mode of the above distribution is $45 - 59$.

(v) We use Eq [2.3.2] to determine the mean from grouped data, with x_i taken to be the midpoint value in each range,

$$\begin{aligned} \bar{x} &= \frac{\sum f_i x_i}{\sum f_i} && (15.1.8)\\ &= \frac{16(37) + 30(52) + 21(67) + 11(82) + 12(97)}{16 + 30 + 21 + 11 + 12} && (15.1.9)\\ &= 62.5 && (15.1.10) \end{aligned}$$

(vi) We use Eq [2.3.5] to determine the variance, σ^2. First we need to determine $E[X^2]$

$$\begin{aligned} \sigma^2 &= \frac{\sum f_i x_i^2}{\sum f_i} - \bar{x}^2 && (15.1.11)\\ &= \frac{16(37^2) + 30(52^2) + 21(67^2) + 11(82^2) + 12(97^2)}{16 + 30 + 21 + 11 + 12} - 62.5^2 && (15.1.12)\\ &= 362.25 && (15.1.13)\\ \sigma &= 19.033 && (15.1.14) \end{aligned}$$

(b) We are given a pie chart.

(i) To determine the total number in the sample, we need to know both the amount of students in a given sector, and the angle of that sector. This is most obvious in the sector 'Bus'. We now use the fact that the following ratio must be true:

$$\begin{aligned} \frac{36}{\text{Total}} &= \frac{162}{360} && (15.1.15)\\ \Rightarrow \text{Total} &= 36 \times \frac{360}{162} && (15.1.16)\\ &= 80 && (15.1.17) \end{aligned}$$

(ii) We know that the sum of angles in a circle is 360°. Thus,

$$\begin{aligned} \theta_B &= 360 - (162 + 90 + 72) && (15.1.18)\\ &= 36^\circ && (15.1.19) \end{aligned}$$

We can also see this if we look at the proportions

$$\frac{8}{80} \times 360^\circ = 36^\circ \qquad (15.1.20)$$

(iii) Similar to part (i), we use the fact that the following ratio must be true:

$$\begin{aligned} \frac{n_w}{80} &= \frac{72}{360} && (15.1.21)\\ n_w &= \frac{72}{360} \times 80 && (15.1.22)\\ &= 16 && (15.1.23) \end{aligned}$$

15.2 Module 2: Managing Uncertainty

3. (a) Let F be the event that an individual buys French.

Let M be the event that an individual buys macaroni.

Thus, we can express the given information concisely as,

$$\begin{aligned} P(F) &= 0.5 && (15.2.1)\\ P(M) &= 0.45 && (15.2.2)\\ P(M|F) &= 0.32 && (15.2.3) \end{aligned}$$

(i) **a)** We want to find $P(F \cap M)$. To do this we can use the equation above and the Def [3.1.6] of conditional probability,

$$P(M|F) = \frac{P(M \cap F)}{P(F)} \tag{15.2.4}$$

$$P(F \cap M) = P(M|F) \times P(F) \tag{15.2.5}$$

$$= 0.32 \times 0.5 \tag{15.2.6}$$

$$= 0.16 \tag{15.2.7}$$

b) We want to find the probability of French only.

$$P(F \cap \bar{M}) = P(F) - P(F \cap M) \tag{15.2.8}$$

$$= 0.5 - 0.16 \tag{15.2.9}$$

$$= 0.34 \tag{15.2.10}$$

(ii) We can display this information in a Venn diagram

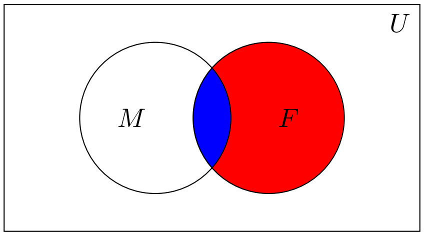


Figure 15.2: The blue region represents $P(F \cap M) = 0.16$ and the red region represents $P(F \cap \bar{M})$

(iii) We want to find $P(\overline{F \cup M})$. We can first apply the complement according to Def [3.1.3]

$$P(\overline{F \cup M}) = 1 - P(F \cup M) \tag{15.2.11}$$

Next, we apply Eq [3.1.5],

$$1 - P(F \cup M) = 1 - [P(F) + P(M) - P(F \cap M)] \tag{15.2.12}$$

$$= 1 - [0.5 + 0.45 - 0.16] \tag{15.2.13}$$

$$= 0.21 \tag{15.2.14}$$

(b) **(i)** Let F be the event that a person is female. Thus, we want to find $P(F)$.
Let S be the sample space i.e. all the persons who were asked. We can calculate $P(F)$ by Eq [3.1.3],

$$P(F) = \frac{|F|}{|S|} \tag{15.2.15}$$

$$= \frac{10 + 16 + 14}{90} \tag{15.2.16}$$

$$= \frac{40}{90} \approx 0.444 \tag{15.2.17}$$

(ii) Let Q be the event that a person prefers Pepsi. Following the same procedure,

$$P(Q) = \frac{|Q|}{|S|} = \frac{20 + 16}{90} \tag{15.2.18}$$

$$= \frac{2}{5} = 0.4 \tag{15.2.19}$$

(iii) Let G be the event that a person likes Ginger ale.
Let M be the event that a person is male.
Thus, we want to find $P(M \cap G)$. Following the same procedure as above,

$$\frac{|M \cap G|}{|S|} = \frac{12}{90} \tag{15.2.20}$$

$$= \frac{2}{15} \approx 0.133 \tag{15.2.21}$$

(iv) Let C be the event that the person likes Coca-Cola. Thus, we want to find $P(C|F)$. We can proceed using Def [3.1.6] for conditional probability,

$$P(C|F) = \frac{P(C\cap F)}{P(F)} \tag{15.2.22}$$

$$= \frac{|C\cap F|}{|S|} \div \frac{|F|}{|S|} \tag{15.2.23}$$

$$= \frac{|C\cap F|}{|F|} \tag{15.2.24}$$

$$= \frac{10}{40} = 0.25 \tag{15.2.25}$$

(c) **(i)** We want to list all the possible selections
Let a 0 be associated with a vanilla, V, and 1 be associated with a coffee C. We now add the left hand side in binary to ensure that we cover all possible combinations.

$$000 \rightarrow VVV \tag{15.2.26}$$

$$001 \rightarrow VVC \tag{15.2.27}$$

$$010 \rightarrow VCV \tag{15.2.28}$$

$$011 \rightarrow VCC \tag{15.2.29}$$

$$100 \rightarrow CVV \tag{15.2.30}$$

$$101 \rightarrow CVC \tag{15.2.31}$$

$$110 \rightarrow CCV \tag{15.2.32}$$

$$111 \rightarrow CCC \tag{15.2.33}$$

(ii) Thus, we want to find $P(2nd\,C|1st\,C)$. We can use Def [3.1.6] of conditional probability,

$$P(2nd\,C|1st\,C) = \frac{P(2nd\,C\cap 1st\,C)}{P(1st\,C)} \tag{15.2.34}$$

$$= \frac{\frac{7}{13}\cdot\frac{8}{14}}{\frac{8}{14}} \tag{15.2.35}$$

$$= \frac{7}{13} \approx 0.538 \tag{15.2.36}$$

It would have been quicker and more intuitive in this case to just look at the new sample space after the first drawn was coffee . i.e 7 coffee and 6 vanilla. So the probability of getting a coffee on the next draw is simply

$$P(C) = \frac{|C|}{|S'|} \tag{15.2.37}$$

$$= \frac{7}{13} \approx 0.538 \tag{15.2.38}$$

where S' is the new sample space.

(iii) This can only happen if we get the events CVV, VCV or VVC. So, we want to find $P(CVV\cup VCV\cup VVC)$. Since, the events are mutually exclusive,

$$P(CVV\cup VCV\cup VVC) = P(CVV)+P(VCV)+P(VVC) \tag{15.2.39}$$

$$= \frac{8}{14}.\frac{6}{13}.\frac{5}{12}+\frac{6}{14}.\frac{8}{13}\frac{5}{12}+\frac{6}{14}.\frac{5}{13}.\frac{8}{12} \tag{15.2.40}$$

$$= \frac{30}{91} \approx 0.330 \tag{15.2.41}$$

Alternatively, we could have also approached this problem with combinatorics.

$$P(EXACTLY\ 2\,V) = \frac{|EXACTLY\ 2V|}{|S|} \quad (15.2.42)$$

$$|EXACTLY\ 2\,V| = \binom{6}{2} \times \binom{8}{1} \quad (15.2.43)$$

$$|S| = \binom{14}{3} \quad (15.2.44)$$

$$P(EXACTLY\ 2\,V) = \frac{\binom{6}{2}\binom{8}{1}}{\binom{14}{3}} = \frac{30}{91} \approx 0.330 \quad (15.2.45)$$

4. **a)** Let X denote the number of errors found on a page.

(i) We can calculate the probabilities for every X by following Eq [3.1.3]

$$P(X = x) = \frac{|X = x|}{\sum |X = x|} \quad (15.2.46)$$

$$= \frac{|X = x|}{350} \quad (15.2.47)$$

x	0	1	2	3	4	5
$P(X = x)$	0.166	0.211	0.280	0.186	0.100	0.057

Table 15.1: A probability distribution table for the discrete random variable X

(ii) We want to calculate $P(2 \le X \le 4)$. Since X is a discrete random variable,

$$P(2 \le X \le 4) = \sum_{x=2}^{4} P(X = x) \quad (15.2.48)$$

$$= P(X = 2) + P(X = 3) + P(X = 4) \quad (15.2.49)$$

$$= 0.280 + 0.186 + 0.100 \quad (15.2.50)$$

$$= 0.566 \quad (15.2.51)$$

(iii) We want to find $E[X]$. We can do this by Def [3.2.2]

$$E[X] = \sum_{\forall x} xP(X = x) \quad (15.2.52)$$

$$= 0(0.166) + 1(0.211) + 2(0.280) + 3(0.186) + 4(0.100) + 5(0.057) \quad (15.2.53)$$

$$= 2.014 \quad (15.2.54)$$

b) Given that Y is the number of success in 10 trials, and $p = 0.3$.

(i) We see that Y satisfies the conditions in Note [3.3.2] to be modeled by a binomial distribution The parameters of this binomial distribution are $n = 10$ and $p = 0.3$, i.e. $X \sim Bin(10, 0.3)$. Furthermore, we know from Def [3.3.1] that,

$$P(Y = y) = \binom{10}{y} 0.3^y (1 - 0.3)^{n-y}, \quad y \in (0, 1, ..., 10) \quad (15.2.55)$$

$$(15.2.56)$$

(ii) **a)** We want to find $P(Y = 4)$. Using the definition above,

$$P(Y = 4) = \binom{10}{4} 0.3^4 0.7^6 = 0.2 \quad (15.2.57)$$

b) We want to find $P(Y \ge 1)$. The long way to do this would be

$$P(Y \ge 1) = P(X = 1) + P(X = 2) + ... + P(X = 10) \quad (15.2.58)$$

$$= \binom{10}{1} 0.3^1 0.7^9 + \binom{10}{2} 0.3^2 0.7^8 + ... + \binom{10}{10} 0.3^{10} 0.7^0 \quad (15.2.59)$$

However, it would be quicker to see that,

$$\begin{aligned} P(Y \geq 1) &= 1 - P(Y < 1) && (15.2.60)\\ &= 1 - P(Y = 0) && (15.2.61)\\ &= 1 - \binom{10}{0} 0.3^0 0.7^{10} && (15.2.62)\\ &= 0.972 && (15.2.63) \end{aligned}$$

c) Let X be the random variable representing the results of the statistics examination, $X \sim N(68, 8)$. Thus, we want to find $P(X > 82) \times 100$. We can proceed by standardizing X,

$$\begin{aligned} P(X > 82) &= P\left(\frac{X-\mu}{\sigma} > \frac{82-68}{8}\right) && (15.2.64)\\ &= P(Z > 1.75) && (15.2.65)\\ &= 1 - P(Z \leq 1.75) && (15.2.66)\\ &= 1 - \Phi(1.75) && (15.2.67)\\ &= 1 - 0.9599 && (15.2.68)\\ &= 0.0401 && (15.2.69) \end{aligned}$$

Thus, the percent of students who will get a Grade A is 4.01%

15.3 Module 3: Analyzing and Interpreting Data

5. (a) We are given that a 95% confidence interval for μ of a normal distribution with σ^2 known is $14.06 \leq \mu \leq 19.54$ and that the

(i) Let w be the width of the interval. Thus,

$$\begin{aligned} w &= 19.54 - 14.06 && (15.3.1)\\ &= 5.48 && (15.3.2) \end{aligned}$$

(ii) Recall from Eq [4.1.5] that the form the mean takes in the confidence interval is given by:

$$\bar{x} \pm z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}} \qquad (15.3.3)$$

$$\Rightarrow \left(\bar{x} - z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}, \bar{x} + z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}\right) \equiv (a, b) \qquad (15.3.4)$$

$$\Rightarrow \bar{x} = \frac{a+b}{2} \qquad (15.3.5)$$

We see that $\bar{x}$ is simply the midpoint of the limits,

$$\begin{aligned} \bar{x} &= \frac{14.06 + 19.54}{2} && (15.3.6)\\ &= 16.80 && (15.3.7) \end{aligned}$$

(iii) We know from Eq [4.1.5] that a $(1-\alpha).100\%$ confidence interval for a population mean when σ^s is known is given by

$$\bar{x} \pm z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}} \qquad (15.3.8)$$

From this, we can see that knowing the width of the interval is enough to determine σ^2 since the width, w, is in fact $2z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}$, and for a 95% confidence interval, $z_{\alpha/2} = 1.96$. Thus,

$$\begin{aligned} 2 \times 1.96 \times \sqrt{\frac{\sigma^2}{60}} &= 5.48 && (15.3.9)\\ \sigma &= \sqrt{\left(\frac{5.48}{2 \times 1.96}\right)^2 \times 60} && (15.3.10)\\ &= 10.829 && (15.3.11) \end{aligned}$$

(b) We are given that $E[X] = 25$ and $Var[X] = 144$ and $n = 81$.

(i) To determine the distribution of $\bar{X}$ we can directly application of the Central Limit Theorem [4.1.1]. We see that $\bar{X}$ follows normal distribution with parameters $\mu = 25$ and $\sigma^2 = \frac{144}{81}$, i.e. $\bar{X} \sim N(25, \frac{144}{81})$

(ii) We can calculate $P(\bar{X} < 28)$. by standardizing,

$$P(\bar{X} < 28) = P\left(\frac{\bar{X} - \mu}{\sigma} < \frac{28 - 25}{\sqrt{\frac{144}{81}}}\right) \tag{15.3.12}$$

$$P(\bar{X} < 28) = P\left(\frac{\bar{X} - \mu}{\sigma} < \frac{3}{\sqrt{\frac{16}{9}}}\right) \tag{15.3.13}$$

$$= P\left(Z < \frac{9}{4}\right) \tag{15.3.14}$$

$$= \Phi(2.25) \tag{15.3.15}$$

$$= 0.9878 \tag{15.3.16}$$

(c) We are given that $\mu_0 = 48$g and $\sigma = 2.3$g before. After, from a sample of $n = 49$, it was found that $\bar{x} = 47.2$g with the same σ.

(i) We state the null and alternative hypothesis as

$$H_0 \ : \ \mu = 48 \tag{15.3.17}$$

$$H_1 \ : \ \mu \neq 48 \tag{15.3.18}$$

(ii) Recall from Section [4.2.3] how to determine the critical region at a 4% level of significance. Since σ^2 is known, we use a Z test. Also, since this is a two-tailed test, we reject H_0 if

$$Z_{calc} > z_{\alpha/2} \quad \text{or} \quad Z_{calc} < -z_{\alpha/2} \tag{15.3.19}$$

$$Z_{calc} > z_{0.04/2} \quad \text{or} \quad Z_{calc} < -z_{0.04/2} \tag{15.3.20}$$

$$Z_{calc} > z_{0.02} \quad \text{or} \quad Z_{calc} < -z_{0.02} \tag{15.3.21}$$

$$Z_{calc} > 2.06 \quad \text{or} \quad Z_{calc} < -2.06 \tag{15.3.22}$$

$$\tag{15.3.23}$$

(iii) Recall from Section [4.2.2] that we can determine the test statistic as

$$Z_{calc} = \frac{\bar{x} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \tag{15.3.24}$$

$$= \frac{47.2 - 48}{\sqrt{\frac{2.3^2}{49}}} \tag{15.3.25}$$

$$= -2.435 \tag{15.3.26}$$

(iv) Since $Z_{calc} = -2.435 < -2.06$, our test statistic lies within the critical region, and we must reject the null hypothesis in favor of the alternative hypothesis. Thus, at the 4% level of significance, there is enough evidence to suggest that the mean changed.

6. **(a)** In general, the dependent variable is usually influenced by one more more independent variables

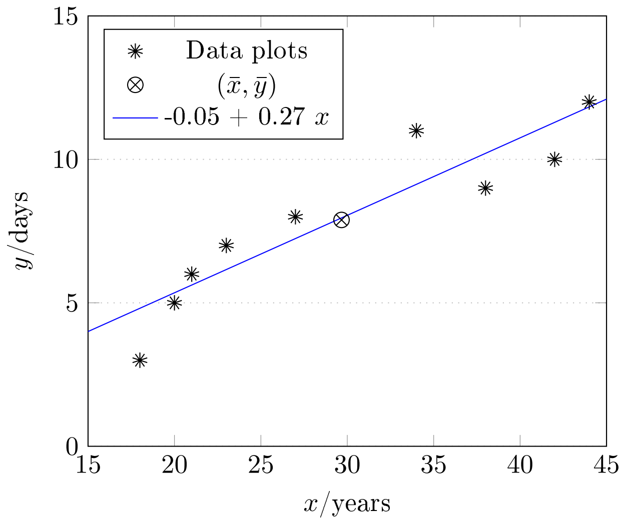
(i) 1: Independent
2: Dependent

(ii) 1: Independent
2: Dependent

(iii) 1: Independent
2: Dependent

(b) We are given 9 pairs of (x, y) data points.

a) We can plot these in a scatter diagram as follows,



b) We can calculate mean values for x and y by Def [2.3.1]

$$\bar{x} = \frac{\sum x}{n} \quad (15.3.27)$$
$$= \frac{18+20+21+27+23+34+42+38+44}{9} \quad (15.3.28)$$
$$= 29.667 \approx 30 \text{ days} \quad (15.3.29)$$
$$\bar{y} = \frac{\sum y}{n} \quad (15.3.30)$$
$$= \frac{3+5+6+8+7+11+10+9+12}{9} \quad (15.3.31)$$
$$= 7.889 \approx 8 \text{ days} \quad (15.3.32)$$

c) We can calculate point $(\bar{x}, \bar{y})$ by definition,

d) We can draw the line of regression given on the scatter diagram.

e) We want to find the y that corresponds to an $x = 22$. To do this we simply plug in our value for x in the line of regression.

$$y = -0.05 + 0.27(22) \quad (15.3.33)$$
$$= 5.89 \approx 6 \ years \quad (15.3.34)$$

(c) Recall from Section [4.3] how to conduct a χ^2 test at the 5% level of significance.

(i) We can state the appropriate null and alternative hypothesis as
H_0 : There is no association between the grade and the school.
H_1 : There is an association between the grade and the school.

(ii) To determine the critical region, we must determine the number of degrees of freedom and then find the corresponding value in the χ^2 table at the 5% significant level. We can calculate the number of degrees of freedom as follows

$$D.o.f = (r-1)(c-1) \quad (15.3.35)$$
$$= (3-1)(4-1) \quad (15.3.36)$$
$$= 6 \quad (15.3.37)$$

Where r and c represent the number of rows and columns respectively. This corresponds to a value of 12.592. Therefore, the critical region for this test is $\chi > 12.592$.

(iii) Calculate the expected number of students who attended school B and received a Grade II. We can compute the expected frequency using $E_{ij} = \frac{n_i \times n_j}{N}$, where E_{ij} represents the expected frequency of the value in the i^{th} row and j^{th} column, n_i represents the sum of values in the i^{th} row, n_j represents the sum of values in the j^{th} column, and N represents the total sample size. Using this notation, we want to find E_{22},

$$\begin{aligned} E_{22} &= \frac{(16+14+12)\times(10+14+11+15)}{165} && (15.3.38) \\ &= 12.727 \approx 12 \text{ students} && (15.3.39) \end{aligned}$$

(iv) We are given that $\chi_{\text{calc}} = 9.1625$. Since $\chi_{calc} = 9.1625 < 12.592$, it is outside the critical region and we fail to reject the null hypothesis. Thus, at the 5% level of significance, there is enough evidence to suggest that there is no association between the grade and the school.

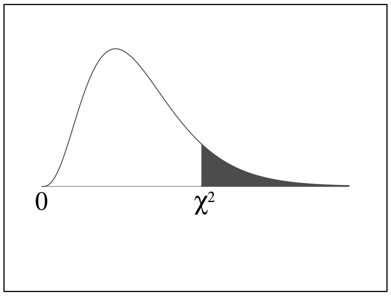
Part IV

Tables

Cumulative Normal Distribution Table

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7703	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Chi-Square Distribution Table



The shaded area is equal to α for $\chi^2 = \chi^2_\alpha$.

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Table entry for p and C is the critical value t^* with probability p lying to its right and probability C lying between $-t^*$ and t^*.

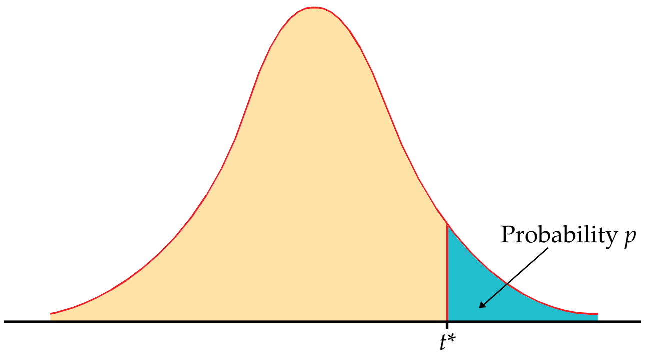


TABLE D

t distribution critical values

	Upper-tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	0.678	0.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	0.675	0.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
z^*	0.674	0.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level C											

Chi-Square from http://sites.stat.psu.edu/~mga/401/tables/Chi-square-table.pdf

Part V

Appendix

15.4 Module 1

15.4.1 The Variance is Unaffected by an Overall Constant

We explore in slightly more detail why the variance is unaffected if a constant value is added to every value in your dataset.

The intuition comes from the fact that the variance quantifies how much spread there is in the data. If we add a the same value to every point, we essentially shift all our points up. We can illustrate this graphically. Suppose the grades of 8 students were distributed as shown below. We then add 3 points to the grade of every student.

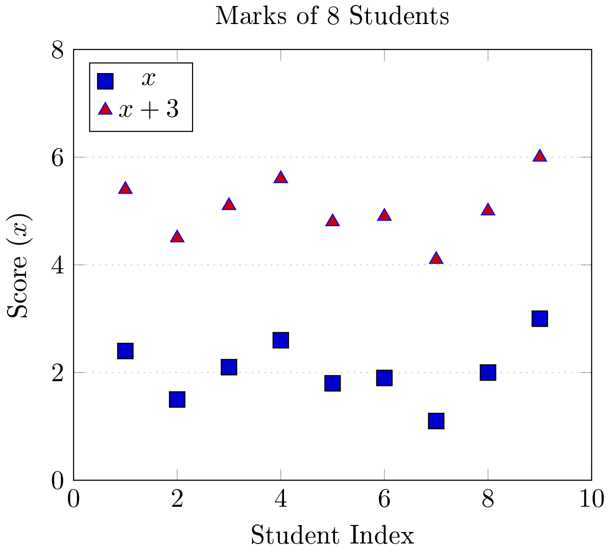


Figure 15.3: The 8 students are labelled $1-8$ on the x-axis and the score of the i^{th} student is given on the y-axis

Although this changes the mean of the distribution, it does not affect the relative location of the data points among themselves.

We see this in the equations if we look at what happens to shifted mean, $\bar{x}_{\text{shift}}$,

$$\begin{aligned}
\bar{x}_{\text{shift}} &= \frac{1}{N}\sum(x_i+a) && (15.4.1)\\
&= \frac{1}{N}\sum x_i + \frac{1}{N}\sum a && (15.4.2)\\
&= \bar{x}+a && (15.4.3)
\end{aligned}$$

So it is as our intuition predicted, the mean increases by a. Now we compute the shifted variance σ^2_{shift}

$$\begin{aligned}
\sigma^2_{\text{shift}} &= \frac{1}{N}\sum(\bar{x}_{\text{shift}}-(x_i+a))^2 && (15.4.4)\\
&= \frac{1}{N}\sum(\bar{x}+a-x_i-a)^2 && (15.4.5)\\
&= \frac{1}{N}\sum(\bar{x}-x_i)^2 && (15.4.6)\\
&= \sigma^2 && (15.4.7)
\end{aligned}$$

15.4.2 Computing Variance

The core equation for the variance, σ^2, of a finite population with size n and mean $\bar{x}$ is given by Def [2.3.2]

$$\sigma^2 = \frac{1}{n}\sum(x-\bar{x})^2 \tag{15.4.8}$$

However, due to limited time in an exam, this formula can be quite burdensome. Thus, we can simplify it to a more convinent expression, Eq [2.3.4]

$$\begin{aligned}
\sigma^2 &= \frac{1}{n}\sum(x-\bar{x})^2 && (15.4.9)\\
&= \frac{1}{n}\sum(x^2-2\bar{x}x-\bar{x}^2) && (15.4.10)\\
&= \frac{\sum x^2}{n} - 2\bar{x}\frac{\sum x}{n} + \frac{n\bar{x}}{n} && (15.4.11)\\
&= \frac{\sum x^2}{n} - 2\bar{x}(\bar{x}) + \mu^2 \text{ as } \frac{\sum x}{n} = \bar{x} && (15.4.12)\\
&= \frac{\sum x^2}{n} - 2\bar{x}^2 + \bar{x}^2 && (15.4.13)\\
&= \frac{\sum x^2}{n} - \bar{x}^2 && (15.4.14)
\end{aligned}$$

This formula is much easier to work with and sometimes necessary if only $\sum x$ and $\sum x^2$ are given.

15.4.3 The a^{th} Percentile

Suppose you calculate the term for the a^{th} percentile following Def [2.3.4],

$$a^{\text{th}} \text{ percentile} = \frac{a}{100}(n+1)^{\text{th}} \text{ term} \tag{15.4.15}$$

If this number is not an integer, lets say b, then we take the

$$\lfloor b \rfloor^{\text{th}} \text{ term} + (b - \lfloor b \rfloor) \times (\lceil b \rceil^{\text{th}} \text{ term} - \lfloor b \rfloor^{\text{th}} \text{ term}) \tag{15.4.16}$$

Where we have introduced the floor function $\lfloor x \rfloor$ and ceiling function $\lceil x \rceil$ for notational convenience. The floor function maps a real number to the largest previous integer. eg

$$\lfloor 3.25 \rfloor = 3 \tag{15.4.17}$$

The ceiling function maps a real number to the smallest next integer. eg

$$\lceil 3.25 \rceil = 4 \tag{15.4.18}$$

15.5 Module 2

15.5.1 Standardizing and the Φ Tables

You are frequently required *standardize* a random variable X and then look up some values in a Φ table. It is important to be able to do this with ease, but it is also important to remember why you do this. In this section we will motivate the procedure of standardizing a normal distribution and using Φ tables.

To begin with, recall from Def [3.4.1] that a continuous random variable X is normally distributed with parameters μ and σ^2, i.e. $X \sim N(\mu, \sigma^2)$, if the density of X is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/2\sigma^2}, \quad -\infty < x < \infty \tag{15.5.1}$$

Next, we define a special case of the above called the **standard or unit normal variable** Z. This is a continuous random variable that is normally distributed with parameters $\mu = 0$ and $\sigma^2 = 1$ i.e. $Z \sim N(0,1)$. So the density of Z, lets call it $g(z)$ to avoid confusion, can be found from above by setting $\mu = 0$ and $\sigma^2 = 1$,

$$g(z) \quad = \quad \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad -\infty < z < \infty \tag{15.5.2}$$

Now suppose you wanted to determine $P(Z < b)$. We know from Prop. [3.2.2] that we need to integrate the density function of Z,

$$P(Z < b) \quad = \quad \int_{-\infty}^{b} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \tag{15.5.3}$$

It is customary to denote this integral, which is also the cumulative distribution function, by $\Phi(b)$,

$$\Phi(b) \quad = \quad P(Z < b) = \int_{-\infty}^{b} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \tag{15.5.4}$$

This is what the Φ table does for you. It is a list of results of integrals for various values of b. For example, if you wanted to find $P(Z < 0.5)$, this can be read from the table, and it represents the result of the following integral,

$$\begin{aligned} \Phi(0.5) &= P(Z < 0.5) && (15.5.5)\\ &= \int_{-\infty}^{0.5} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz && (15.5.6)\\ \Phi(0.5) &= 0.6915 && (15.5.7) \end{aligned}$$

The next big question is, what is unique about this table and why is it on the back of so many textbooks and exams? We claim that you can use this table to compute probabilities for any normal variable, and not just the standard normal variable. We can motivate this in two ways.

Recall from Note [3.2.1], that if X is normally distributed with parameters μ and σ^2, i.e. $X \sim N(\mu, \sigma^2)$, then $Y = aX + b$ is normally distributed with parameters $a\mu + b$ and $a^2\sigma^2$, i.e. $Y \sim N(a\mu + b, a^2\sigma^2)$. So if we consider the random variable $Z = \frac{X-\mu}{\sigma}$,

$$\begin{aligned} Z &= \frac{X - \mu}{\sigma} && (15.5.8)\\ &= \frac{1}{\sigma} X - \frac{\sigma}{\mu} && (15.5.9) \end{aligned}$$

then Z is normally distributed as

$$\begin{aligned} Z &\sim N\left(\frac{1}{\sigma}\mu - \frac{\mu}{\sigma}, \frac{1}{\sigma^2}\sigma^2\right) && (15.5.10)\\ &\sim N(0,1) && (15.5.11) \end{aligned}$$

So if you want to find $P(X < a)$, the procedure of standardizing,

$$P(X < a) \to P\left(\frac{X - \mu}{\sigma} < \frac{a - \mu}{\sigma}\right) \tag{15.5.12}$$

allows you to reduce the problem to one that you have answers for already,

$$\begin{aligned} P\left(\frac{X - \mu}{\sigma} < \frac{a - \mu}{\sigma}\right) &= P\left(Z < \frac{a - \mu}{\sigma}\right) && (15.5.13)\\ &= \Phi\left(\frac{a - \mu}{\sigma}\right) && (15.5.14) \end{aligned}$$

Although this is sufficient to motivate the procedure, we present one more view, since it does not depend on Note [3.2.1].

Again, suppose we wanted to find $P(X < a)$ for $X \sim N(\mu, \sigma^2)$. We know that we must compute the following integral,

$$\begin{aligned} P(X < a) &= \int_{-\infty}^{a} f(x)dx && (15.5.15) \\ &= \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} dx && (15.5.16) \end{aligned}$$

We change variables for simplification. Let $z = \frac{x-\mu}{\sigma}$,

$$\begin{aligned} \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} dx &= \int_{-\infty}^{(a-\mu)/\sigma} \frac{1}{\sqrt{2\pi}\sigma} e^{-z^2/2} \sigma dz && (15.5.17) \\ &= \int_{-\infty}^{(a-\mu)/\sigma} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz && (15.5.18) \\ &= \Phi\left(\frac{a-\mu}{\sigma}\right) && (15.5.19) \end{aligned}$$

So we again, get the same result.

15.6 Module 3

15.6.1 Central Limit Theorem

For those wanting a deeper intuition on how we got the parameters in Central Limit Theorem (4.1.1), we could have easily derived this result directly by applying some properties of expectations and variance. These properties can be derived with all the knowledge you already have.

Since $\bar{X}$ represents the sum of independent, identically random variables with mean μ and variance σ^2,

$$\begin{aligned} E[\bar{X}] &= E\left[\frac{\sum X_i}{n}\right] && (15.6.1) \\ &= \frac{1}{n} E\left[X_1 + X_2 + ... + X_n\right] && (15.6.2) \\ &= \frac{1}{n}\left(E[X_1] + E[X_2] + ... + E[X_n]\right) && (15.6.3) \\ &= \frac{1}{n}\left(nE[X]\right) = E[X] = \mu && (15.6.4) \end{aligned}$$

The variance is a bit more challenging. To show this we need to use the fact that $Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var\left(X_i\right)$ if $X_1, X_2, ...X_n$ are pairwise independent, which is a statement that is beyond the scope of the syllabus. Using this, we can start with the sample variance s^2 and apply it to $Var(\bar{X})$,

$$\begin{aligned} s^2 &= \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1} && (15.6.5) \\ Var(\bar{X}) &= Var\left(\frac{\sum X_i}{n}\right) = \sum Var\left(\frac{X_i}{n}\right) && (15.6.6) \\ &= \sum \frac{\left(\frac{X_i}{n} - \frac{\bar{X}}{n}\right)^2}{n-1} = \sum \frac{1}{n^2} \frac{(X_i - \bar{X})^2}{n-1} && (15.6.7) \\ &= \sum \frac{1}{n^2} Var(X_i) = \frac{1}{n^2} \sum Var(X_i) && (15.6.8) \\ &= \frac{1}{n^2}(nVar(X)) = \frac{Var(X)}{n} = \frac{\sigma^2}{n} && (15.6.9) \end{aligned}$$

15.6.2 Alternate Formula for the Mean

We provide an additional formula for the mean that can help reduce the complexity of certain computations. We claim that the mean $\bar{x}$ can be found by the following formula

$$\bar{x} = \frac{\sum x - a}{n} + a \tag{15.6.10}$$

Proof. Starting from Def [2.3.1], we can do the following manipulations

$$\begin{aligned}
\bar{x} &= \frac{\sum x}{n} && (15.6.11)\\
\Rightarrow n\bar{x} &= \sum x && (15.6.12)\\
&= x_1 + ... + x_n && (15.6.13)\\
&= (x_1 - a) + a + ... + (x_n - a) + a \quad \forall a && (15.6.14)\\
&= \sum (x - a) + na && (15.6.15)\\
\bar{x} &= \frac{\sum (x - a)}{n} + a && (15.6.16)
\end{aligned}$$

□

This formula also shows that if all the terms are changed by ±a, the mean is also changed to ±a.

15.6.3 Confidence with Confidence Interval

Computing a confidence interval for a population mean μ is a common task that has been simplified into some basic formulas in Def [4.1.5]. With such a simple formula, it is easy not to think about where it comes from. Here we provide some motivation using the simplest case in the hope that you can take this idea and understand the others on your own.

Suppose we wanted to determine a $(1 - \alpha).100\%$ confidence interval for the mean of a random variable X. To do this, we collect a sample of size n. We know from the Central Limit Theorem, that if the variance is known, then the mean, $\bar{X}$, is normally distributed as $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ (if the variance is not known then we leave it as an exercise to adjust the procedure accordingly).

Before we continue, let us standardize our parameter $\bar{X}$ to make it easier to deal with probabilities in the future. We define Z such that

$$Z = \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \tag{15.6.17}$$

Note that we standardized not knowing what the actual value of μ is. So $Z \sim N(0, 1)$

Next, to construct the confidence interval, we should recall from Def [4.1.5] that an $(1 - \alpha)\%$ confidence interval for μ means that we will find with probability $\frac{1-\alpha}{100}$ a confidence interval in which the actual value of the parameter μ will lie within. Since we have constructed Z, and we also know that Z is symmetric about $z = 0$, we can write this desire as an equation. We pick a convention (that will become clear in retrospect) to define $z_{\alpha/2}$ as the value of z such that $\Phi(z) = 1 - \frac{\alpha}{2}$. So we want to find the value of z such that

$$\begin{aligned}
P(-z < Z < z) &= 1 - \alpha && (15.6.18)\\
P(Z < z) - P(Z < -z) &= 1 - \alpha && (15.6.19)\\
2P(Z < z) - 1 &= 1 - \alpha && (15.6.20)\\
P(Z < z) &= 1 - \frac{\alpha}{2} && (15.6.21)\\
\Phi(z) &= 1 - \frac{\alpha}{2} && (15.6.22)\\
\Rightarrow z &= z_{\alpha/2} && (15.6.23)\\
\Rightarrow z_{\alpha/2} &= \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) && (15.6.24)
\end{aligned}$$

Now that we have determined that $z = z_{\alpha/2}$, we need to stop thinking about Z and remember that it is really $\bar{X}$ that we care about! Remember, standardizing is simply a trick to reduce problems into ones that we have answers for. In this case, we knew the probability, but we didn't know what value was needed to give that probability. So, in terms of $\bar{X}$,

$$\begin{aligned}
P(-z_{\alpha/2} < Z < z_{\alpha/2}) &= 1-\alpha && (15.6.25)\\
P\left(-z_{\alpha/2} < \frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} < z_{\alpha/2}\right) &= 1-\alpha && (15.6.26)\\
P\left(-z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}} < \bar{X}-\mu < z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}\right) &= 1-\alpha && (15.6.27)\\
P\left(-\bar{X}-z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}} < -\mu < -\bar{X}+z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}\right) &= 1-\alpha && (15.6.28)\\
P\left(\bar{X}+z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}} > \mu > \bar{X}-z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}\right) &= 1-\alpha && (15.6.29)\\
P\left(\bar{X}-z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}} < \mu < \bar{X}+z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}\right) &= 1-\alpha && (15.6.30)
\end{aligned}$$

And so, we see that the interval $\bar{x} \pm -z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}$ gives us the probability that we wanted i.e. the probability that the actual mean μ will lie in this interval is $(1-\alpha)$.

15.6.4 Hypothesis Testing Intuition

This is a term used to describe the process by which we make statistical inferences about certain features of a population, based on clues we obtain from a sample. The hypothesis that is tested is the assumption that some characteristic of the population is true. This is called the null hypothesis. We also state an alternative hypothesis which is a statement that opposes the null hypothesis. Once the null and alternative hypotheses are stated, we can then use the information from sample data to derive a quantity known as a test statistic. This is simply a number which we calculate based on knowledge of the sampling distribution.

Once we have simplified the data set to a singular number, our test statistic, we can proceed to identifying the critical region. This region depends on the distribution in question and is determined by the level of significance that we define (sometimes referred to as the alpha level). This refers to the probability that we reject the null hypothesis given that it is true; in other words, a false positive. This means that, regardless of what the hypotheses state or the test statistic we obtain, there will always be a 5% chance that we incorrectly conclude that the statement given in the null hypothesis is untrue.

The clever student might then ask, so if the alpha level is bad for our test, why don't we make it as small as possible? Well, incidentally, the chance of failing to reject an incorrect null hypothesis increases as the alpha level decreases. This is because, as the alpha level decreases, the test becomes stricter and only very extreme observations can be used to conclude that the null hypothesis is untrue. Once we establish the 5% region/s of the distribution, we can compare our test statistic to the boundaries of the critical region to determine if it falls within the critical region or not. If it does not, we would fail to reject the null hypothesis. On the other hand, if it does fall within the critical region, we would then reject the null hypothesis in favor of the alternative hypothesis.

The importance of the terminology here is a bit subtle. We say 'fail to reject' rather than 'accept' since we assume that the null hypothesis is true from the outset and we try to provide evidence in support of the alternative hypothesis. [1]

[1] All of the prior considerations highlight the importance of knowing and understanding the different types of sampling methods in order to acquire data that is representative of the population; in other words, try not to take Module 1 for granted.