Contents

1 Permutations and Combinations **3**
- 1 Basic Counting Principles . . . 3
  - 1.1 Addition Principle for Counting . . . 3
  - 1.2 Multiplication Principle for Counting . . . 3
- 2 Permutations . . . 4
  - 2.1 Permutations Involving Identical Objects . . . 5
  - 2.2 Permutations Involving Non-Identical Objects . . . 5
  - 2.3 Permutations in a Circle . . . 7
- 3 Combinations . . . 8
- 4 Grouping Method, Inserting Method, Complementary Method . . . 10
- 5 Combinations, Pascal's Triangle and The Binomial Distribution . . . 12

2 Probability **15**
- 1 Gambling and Probability . . . 15
- 2 Introduction to Probability . . . 16
- 3 Venn Diagrams . . . 17
- 4 Independent Events . . . 18
- 5 Dependent Events . . . 18
- 6 Simple Combined Events . . . 19
  - 6.1 Possibility Diagrams . . . 19
  - 6.2 Tree Diagrams . . . 20
- 7 Sampling With and Without Replacement . . . 20
- 8 Laws of Probability . . . 21
  - 8.1 Addition Law of Probability . . . 21
  - 8.2 Mutually Exclusive Events . . . 21
  - 8.3 Conditional Probability . . . 22
- 9 Probabilities Using Permutations and Combinations . . . 23
- 10 De Meres Problem . . . 25
- 11 Newton-Pepys Problem . . . 25
- 12 The St. Petersberg Paradox . . . 26
- 13 The Monty Hall Problem . . . 27

3 Discrete Random Variables **29**
- 1 Discrete Random Variables . . . 29
- 2 Probability Distribution . . . 30
- 3 Expected Value of a Random Variable . . . 32
- 4 Variance and Standard Deviation or a Random Variable . . . 35

4 Binomial Distribution **36**
- 1 Binomial Experiments . . . 36
- 2 The Binomial Distribution . . . 37
- 3 Mean and Varianace of a Binomial Distribution . . . 40
- 4 Binomial Distribution Problems Involving Probability Trees . . . 41

5 Normal Distribution **43**
1 The Normal Distribution . . . 43
1.1 Properties of The Normal Distribution . . . 44
1.2 Finding Probabilities Using Graphic Calculator . . . 44
2 The Standard Normal Distribution (Z-Distribution) . . . 46
3 Linear Combinations of Normal Random Variables . . . 48

6 Sampling Distribution (Central Limit Theorem) **50**
1 Population and Sample . . . 50
2 Random Sampling . . . 50
3 Population Parameters and Sample Statistics . . . 51
4 The Sample Mean, $\overline{X}$ as a Random Variable . . . 52
5 Distribution of $\overline{X}$ . . . 52
5.1 When X Follows a Normal Distribution . . . 52
5.2 When X Follows Any Distribution (Central Limit Theorem) . . . 53

7 Hypothesis Testing **55**
1 Null and Alternative Hypotheses . . . 55
2 Hypothesis Testing . . . 56
2.1 The Test Statistic . . . 56
2.2 Level of Significance . . . 57
2.3 Critical Region and Critical Values . . . 58
2.4 The p−value . . . 60
3 The Z−Test . . . 62
4 Unbiased Estimates of Population Parameters . . . 65
4.1 Unbiased Estimate for Population Mean μ . . . 65
4.2 Unbiased Estimate for Population Variance s^2 . . . 65
4.3 Hypothesis Test With Unknown Variance and Large Sample Size n . . . 68

8 Correlation and Regression **73**
1 Bivariate Data . . . 73
2 Scatter Diagrams . . . 73
3 Correlation . . . 74
3.1 Direction . . . 74
3.2 Linearity . . . 74
3.3 Strength . . . 75
4 Causality . . . 76
5 Pearson's Product-Moment Correlation Coefficient, r . . . 77
5.1 Properties of Pearson's Product-Moment Correlation Coefficient . . . 78
5.2 Limitations of Pearson's Product-Moment Correlation Coefficient . . . 79
6 Linear Regression . . . 82
6.1 Independent and Dependent Variables . . . 82
6.2 Least Squares Regression Line of y on x . . . 82
6.3 Interpolation and Extrapolation . . . 83
6.4 Least Squares Regression Line of x on y . . . 85
7 Transformations to Linearize Bivariate Data . . . 90

CHAPTER

1

Permutations and Combinations

1 Basic Counting Principles

Is a password with at least one uppercase letter really more secure than one without? What are my odds of winning the lottery? How many ways could I arrange my friends around the dining table at my dinner party? In everyday life, we often need to "count" the number of ways to arrange objects. However listing out all the possible arrangements is not always easy. In this chapter we will be learning some fundamental counting techniques, to help us answer some of life's burning questions.

1.1 Addition Principle for Counting

The **Addition Principle for Counting** states that if we have a_1 ways of doing something and a_2 ways of doing another thing and we cannot do both at the same time, then the number of ways we can choose a_1 **or** a_2 is

$$a_1 + a_2$$

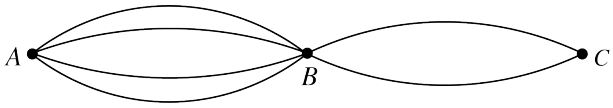
Zoe is trying to decide what to wear to the prom. She has 2 ball gowns, 3 blouses and 1 sundress.
How many different outfits can Zoe choose from?
Zoe has $2+3+1=6$ outfits to choose from.

1.2 Multiplication Principle for Counting

The **Multiplication Principle for Counting** states that if there is a sequence of independent events that can occur $a_1, a_2, a_3, \ldots a_n$ ways, then the number of ways all the events can occur is

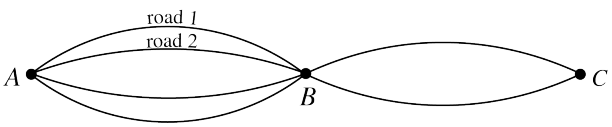
$$a_1 \times a_2 \times a_3 \times \ldots \times a_n$$

Suppose there are three towns A, B and C and that four different roads can be taken from A to B and two different roads from B to C.



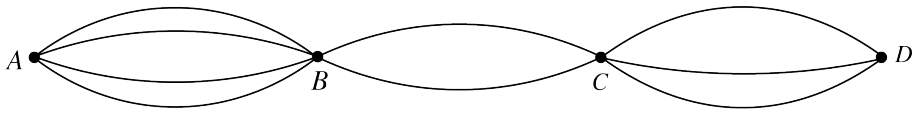
This begs the question: "How many different pathways are there from A to C, going through B?"
If we take road 1, there are two different roads to complete our trip.

If we take road 2, there are two different roads to complete our trip ... etc.



So there are $2+2+2+2=4\times 2$ different pathways we can take from A to C.

Similarly, for



There would be $4 \times 2 \times 3 = 24$ different pathways from A to D, passing through B and C.

Lisa's Birthday Party

Lisa invites six guests to her birthday party: Raj, Alice, Jerry, Michelle, Ryan and Heather. When they arrive, they shake hands with each other. Jerry asks: "How many handshakes happened in total?'"
"I shook 6 hands altogether" says Ryan, "and I guess, so did everybody else."
"Since there are seven of us, this should mean $7 \cdot 6 = 42$ handshakes in total!" ventures Michelle.
"This seems too many" says Heather. "The same logic gives 2 handshakes if two persons meet, which is clearly wrong."
"That is because every handshake was counted twice. We have to divide 42 by 2, to get the right number: 21." settles Lisa the issue.

2 Permutations

A **permutation** is an **ordered arrangement** of a number of objects.

Lisa and her guests make their way to the dining table to eat. Since it is Lisa's birthday, she sits at the head of the table. How many different seatings are there (with Lisa's place fixed)?
Let us fill the seats one by one, starting with the chair on Lisa's right. There are 6 choices for the guest who will sit down first. How many choices are there for the guest who goes second? There are only 5 choices as the person who went first is already seated. Therefore there are a total of $6 \cdot 5$ ways to seat the first two guests.
We can then proceed in a similar manner: we have 4 choices for the third guest to be seated, 3 choices for the fourth guest to be seated, and so on. Therefore, the number of ways in which the guests can be seated is

$$6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 6! = 720$$

The number of permutations of n different objects is $n!$

Example 1.1 Seven Layer Cake

Alice made a seven layer rainbow cake for Lisa. How many ways could she have arranged the layers of the cake with the 7 distinct colours of the rainbow?
Solution

The answer is $7! = 5040$.

The simplicity of the answer to the this question was due to several factors: we used each of our objects exactly once, the order of the objects mattered, and the objects were all different. In the rest of this section we will study problems without one or more of these simplifying factors.

2.1 Permutations Involving Identical Objects

Example 1.2 Flowers in a Row (Repeated Colours)

A gardener has five red flowers, three yellow flowers and two white flowers to plant in a row. In how many different ways can she do that?

Solution

This problem differs from the previous one in only one aspect: the objects are not all different. We are going to answer this question by reducing it to the previous one, in which all objects were different. Assume our gardener plants her flowers in a row, then sticks labels (say numbers 1 through 5 for the red flowers, 1 through 3 for the yellow ones, and 1 through 2 for the white ones) to her flowers so that she can distinguish them. Now she has ten different flowers, and therefore the row of flowers she has just finished working on can look 10! different ways. We have to tell how many of these arrangements differ only because of these labels.

The five red flowers could be given five different labels in 5! different ways. The three yellow flowers could be given three different labels in 3! different ways. The two white flowers could be given two different labels in 2! different ways. Therefore, the labeling of all ten flowers can be done in $5! \cdot 3! \cdot 2!$ different ways. Therefore, the total number of arrangements when the labels are removed is

$$\frac{10!}{5! \cdot 3! \cdot 2!} = 2520$$

The number of permutations of n objects with n_1 identical objects of the first type, n_2 identical objects of the second type, ..., and n_k identical objects of the k type is

$$\frac{n!}{n_1! n_2! \dots n_k!}$$

2.2 Permutations Involving Non-Identical Objects

Non-Identical Objects Taken Without Repetition

Example 1.3 20 Politicians Filling 5 Position

A president must choose five politicians from a pool of 20 candidates to fill five different cabinet positions. In how many different ways can she do that?

Solution

Indeed, we have 20 choices for the first candidate, 19 choices for the second, and so on, just as we did the case of factorials. The only difference is that here we do not have 20 slots to fill. We stop after choosing 5 of them.

The number $20 \cdot 19 \cdot 18 \cdot 17 \cdot 16$ is denoted by ${}^{20}P_5$. Thus, there are ${}^{20}P_5 = 1860480$ ways of filling this cabinet. If the candidates are all equally qualified, it may take a while.

The number of permutations of n different objects taken r at a time **without repetition** is

$${}^{n}P_r = \frac{n!}{(n-r)!}$$

Non-Identical Objects Taken With Repetition

Example 1.4 Building 10 Intersections

A city has recently built ten intersections. Some of these will get traffic lights, and some of those that get traffic lights will also get a gas station. In how many different ways can this happen?

Solution

For each intersection, there are three possible scenarios:

1. No traffic light and no gas station.
2. Traffic light but no gas station.
3. Traffic light and a gas station.

Since we have three possible arrangements for each intersection, there will be $3 \cdot 3$ possible arrangements for two intersections, and a total of 3^{10} arrangements for the ten intersections to be built.

The number of permutations of n different objects taken r at a time **with repetition** is n^r.

Example 1.5 Computer Password

A certain computer access password consists of 3 through 5 lowercase letters chosen from the 26 letters in the Roman alphabet, with repetitions allowed. How many different passwords are possible?
The set of all passwords can be split into three subsets consisting of passwords with lengths 3, 4, and 5.
By the addition rule, the total number of passwords equals the number with length 3, plus the number with length 4, plus the number with length 5.
Total number of passwords with length $3 = 26^3$
Total number of passwords with length $4 = 26^4$
Total number of passwords with length $5 = 26^5$
Hence the total number of passwords is

$$26^3 + 26^4 + 26^5 = 12355928$$

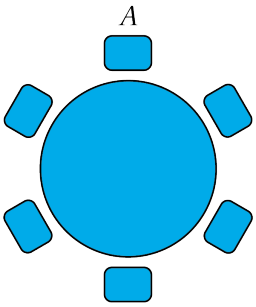
(Approximately 12 million)
How many different passwords are possible if the password contains both uppercase and lowercase letters?
Now instead of having 26 characters to choose from, we have 52.
Total number of passwords with length $3 = 52^3$
Total number of passwords with length $4 = 52^4$
Total number of passwords with length $5 = 52^5$
Hence the total number of passwords is

$$52^3 + 52^4 + 52^5 = 387656256$$

(Approximately 388 million)

2.3 Permutations in a Circle

At a meeting of diplomats, the six participants are to be seated around a circular table. Since the table has no ends to confer particular status, it doesn't matter who sits in which chair. But it does matter how the diplomats are seated relative to each other. In other words, two seatings are considered the same if one is a rotation of the other. How many different ways can the diplomats be seated?
Call the diplomats by the letters A, B, C, D, E, and F. Since only relative position matters, you can start with any diplomat (say, A), place that diplomat anywhere. There is only one way to do this. Then consider all arrangements of the other diplomats around that one. The five diplomats B through F can be arranged in the seats around diplomat A in all possible orders. So there are $5! = 120$ ways to seat the group.



Five other diplomats to be seated: B, C, D, E, F

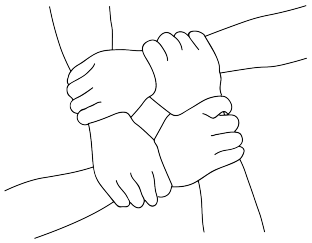
If the seats where numbered/labelled/distinct from each other in some way, then it would matter what seat they were sitting in and there would be $n!$ ways of arranging the diplomats.

> The number of ways of arranging n different objects in a circle is $(n-1)!$

Unless specified otherwise, all seats are assumed to be not numbered when circular arrangement is involved.

Example 1.6 Circular Permutations

How many ways can four friends use their left hands to hold another friend as shown in the diagram?



Solution

Number of ways = $(4-1)! = 3! = 6$

3 Combinations

A **combination** is a selection of objects **without regard to order or arrangement**.

How many combinations of 3 letter strings can be taken from the set of 4 letters $\{A, B, C, D\}$?
There are 4 combinations: ABC, BCD, ABD, ACD.

The number of combinations of n different objects taken r at a time is

$$ {}^nC_r = \binom{n}{r} = \frac{n!}{r!\,(n-r)!} $$

How many ways are there of choosing 5 cards out of a standard 52−card deck?
The answer is that there are ${}^{52}C_5 = 2598960$ ways.

Example 1.7 Probability of Winning The Lottery

In a lottery 5 numbers are selected from 90. If George buys 100 tickets, what is the probability that he wins?

Solution

Total number of combinations $= {}^{90}C_5 = 43,949,268$

$$ \begin{aligned} \mathrm{P}(\text{George wins}) &= \frac{100}{43949268} \\ &= 2.28 \times 10^{-6} \end{aligned} $$

Example 1.8 Forming a Team

The tennis team comprises 8 boys and 12 girls. Three boys and two girls are to be chosen at random to enter a competition. In how many ways can this be done?

Solution

Number of ways to choose 3 boys $= {}^8C_3$
Number of ways to choose 2 girls $= {}^{12}C_2$
Therefore, the total number of ways to choose 3 boys and 2 girls is ${}^8C_3 \times {}^{12}C_2 = 3696$.

Example 1.9 Filling Up Seats

In how many ways can 5 people be chosen to occupy 3 seats in a row?

Solution

Number of ways $= {}^5C_3 \times 3! = 60$
Alternatively, Number of ways $= {}^5P_3 = 60$

Note:

- ${}^nP_r = {}^nC_r \times r!$
- ${}^nC_r = {}^nC_{n-r}$
- ${}^nC_0 = {}^nC_n = 1$

Example 1.10 Forming a Team With Restrictions

Suppose a group of twelve consists of five men and seven women.

(a) How many five-person teams can be chosen that consist of three men and two women?

(b) How many five-person teams contain at least one man?

(c) How many five-person teams contain at most one man?

Solution

(a) There are ${}^5C_3$ ways to choose 3 men and ${}^7C_2$ ways to choose two women.
Hence by the multiplication rule,
Number of teams $= {}^5C_3 \times {}^7C_2$
$= 210$

(b) Number of teams with at least one man
= Total number of teams – Number of teams with no men
$= {}^{12}C_5 - {}^7C_5$
$= 771$

(c) Number of teams with at most one man
= Number of teams with no men + Number of teams with one man
$= {}^5C_0 \times {}^7C_5 + {}^5C_1 \times {}^7C_4$
$= 196$

4 Grouping Method, Inserting Method, Complementary Method

Example 1.11 Sitting in a Row With Restrictions

In how many ways can 4 boys and 3 girls be arranged in a row,

(a) so that the 3 girls are always together?

(b) so that the first and last place are occupied by boys?

(c) if all the girls are to be separated?

(d) if the 3 girls are not all together together?

Solution

(a) We can put the 3 girls together as 1 group.
The number of ways to arrange the 4 boys and 1 group of girls is 5!.
The number of ways the 3 girls can arrange each other among themselves is 3!.
Thus, the total number of ways that the 3 girls are always together is $5! \times 3! = 720$.
(We call this the **grouping method**)

(b) $\underline{4}\_\_\_\_\_\underline{3}$
The 1st and last place can be filled in $3 \times 4 = 12$ ways.
The remaining 5 people can be arranged in 5! ways.
Thus, the total number of ways the first and last place are occupied by boys is $12 \cdot 5! = 1440$.
(Alternatively, we could use $\left[{}^4C_2 \times 2!\right] \times 5! = 1440$)

(c) Here we use the **slotting method**.
$\_B_1\_B_2\_B_3\_B_4\_$
Boys can arrange themselves in 4! ways.
There are 5 slots for the 3 girls. The girls can arrange themselves ${}^5C_3 \times 3! = 60$ ways.
Thus, the total number of ways that the girls are seperated is $4! \times 60 = 1440$.

(d) Since for any arrangement, EITHER all the 3 girls are together, OR not all the 3 girls are together,

$$\begin{aligned}\text{Number of ways 3 girls not all together} &= \text{Total} - \text{Complement} \\ &= 7! - 720 \quad \text{(From part a)} \\ &= 4320\end{aligned}$$

(This is known as the **complementary method**)

Example 1.12 Sitting Around a Round Table With Restrictions

4 boys and 3 girls are to be seated at a round table. Andrew, Jane and Bob are 3 particular people among the 7. How many arrangements are there if

(a) there is no restriction?

(b) Andrew and Jane must be together?

(c) Andrew and Jane must be separated?

(d) Jane must sit between Andrew and Bob?

How many arrangements are there if there is no restriction, but the seats are numbered?

Solution

(a) Number of arrangements = $(7-1)! = 6! = 720$

(b) Group Andrew and Jane together as 1 unit.

Number of ways to arrange 6 units = $5!$

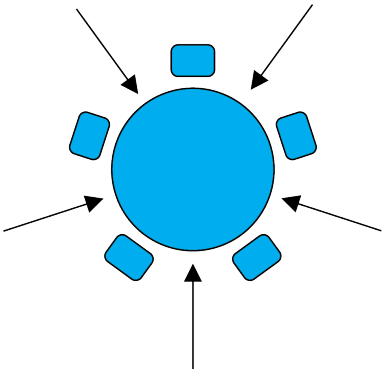
Number of internal arrangements for Andrew and Jane = $2!$

Number of arrangements = $5! \times 2!$

$= 240$

(c) Use the slotting method.
Number of ways to arrange other 5 people = $(5-1)! = 4!$
Andrew and Jane can arrange themselves in ${}^5C_2 \times 2!$ ways.
Number of arrangements = $4! \times {}^5C_2 \times 2!$

$= 480$



(d) Consider Andrew Jane and Bob as 1 unit.

Number of ways to arrange 5 units = $4!$

Andrew and Bob can permute among themselves in $2!$ ways.

Number of arrangements = $4! \times 2!$

$= 48$

If seats are numbered, it is equivalent to arranging them in a row.
Therefore, number of arrangements = $7! = 5040$.

Example 1.13 Boys and Girls Alternating

In how many ways can 5 boys and 5 girls be arranged in a row such that the boys and girls alternate?
Solution

Case 1: The first person is a boy. $BGBGBGBGBG$
Number of ways = $5! \times 5!$

Case 2: The first person is a girl. $GBGBGBGBGB$
Number of ways = $5! \times 5!$

Hence, total number of ways = $5! \times 5! + 5! \times 5! = 28800$

5 Combinations, Pascal's Triangle and The Binomial Distribution

Mathematics is the art of giving the same name to different things.

-Henri Poincaré

Pascal's triangle consists of a triangle of numbers where each term is equal to the sum of the two terms above it. We adopt the convention that the topmost row is row 0 and the leftmost term of each row is the 0th term. It turns out that the rth term in the nth row is given by ${}^nC_r$. Try to verify this for yourself for a few of the terms!

$n = 0$: 1
$n = 1$: 1 1
$n = 2$: 1 2 1
$n = 3$: 1 3 3 1
$n = 4$: 1 4 6 4 1
$n = 5$: 1 5 10 10 5 1
$n = 6$: 1 6 15 20 15 6 1
$n = 7$: 1 7 21 35 35 21 7 1
$n = 8$: 1 8 28 56 70 56 28 8 1

The following equations are the binomial expansion of $(1+x)^n$, where $n = 0, 1, 2, 3, 4, 5, 6$.

$$(1+x)^0 = 1$$
$$(1+x)^1 = 1 + x$$
$$(1+x)^2 = 1 + 2x + x^2$$
$$(1+x)^3 = 1 + 3x + 3x^2 + x^3$$
$$(1+x)^4 = 1 + 4x + 6x^2 + 4x^3 + x^4$$
$$(1+x)^5 = 1 + 5x + 10x^2 + 10x^3 + 5x^4 + x^5$$
$$(1+x)^6 = 1 + 6x + 15x^2 + 20x^3 + 15x^4 + 6x^5 + x^6$$

Notice that the coefficients of the binomial expansion of $(1+x)^n$ are simply $\binom{n}{0}, \binom{n}{1}, \binom{n}{2}, \ldots \binom{n}{n}$.

TEST YOURSELF

(a) A permutation is an _____ of a number of objects.

(b) The number of permutations of a set of n elements equals _____.

(c) How many ways can the letters in the word COMPUTER be arranged in a row?

(d) The number of permutations of n different objects taken r at a time without repetition is _____.

(e) A father has 5 books and wishes to give one to each of his 3 children. In how many ways can he do this?

(f) The number of permutations of n different objects taken r at a time with repetition is _____.

(g) How many 3−digit numbers can be formed from the set $\{1,2,3,4,5\}$ if the digits may be used more than once?

(h) How many ways can we arrange 2 identical white balls and 3 identical red balls in a row?

(i) How many ways can n people sit at a round table?

(j) How many ways can n people sit at a round table if the seats are numbered?

(k) A combination is a selection of objects without regard to _____.

(l) The formula for ${}^nC_r$ is _____.

(m) How many ways can we select 5 people for a team from a group of 12.

ANSWER

(a) A permutation is an **ordered arrangement** of a number of objects.

(b) The number of permutations of a set of n elements equals $\boldsymbol{n!}$.

(c) COMPUTER can be arranged $\mathbf{8!} = \mathbf{40320}$ ways.

(d) The number of permutations of n different objects taken r at a time without repetition is ${}^{n}\boldsymbol{P}_{r} = \frac{\boldsymbol{n!}}{\boldsymbol{(n-r)!}}$.

(e) The father can distribute the books ${}^{5}\boldsymbol{P}_{3} = \mathbf{60}$ ways.

(f) The number of permutations of n different objects taken r at a time with repetition is $\boldsymbol{n^{r}}$.

(g) From the set $\{1,2,3,4,5\}$, we can form $\mathbf{5^{3}} = \mathbf{125}$ 3−digit numbers.

(h) We can arrange 2 identical white balls and 3 identical red balls $\frac{\mathbf{5!}}{\mathbf{2!\times 3!}} = \mathbf{10}$ ways.

(i) n people can sit around a table in $\boldsymbol{(n-1)!}$ ways.

(j) If the seats are numbered, they can sit in $\boldsymbol{n!}$ ways.

(k) A combination is a selection of objects without regard to **order or arrangement**.

(l) The formula for ${}^{n}C_{r}$ is $\frac{\boldsymbol{n!}}{\boldsymbol{r!(n-r)!}}$.

(m) We can select 5 people from a group of 12 in ${}^{12}\boldsymbol{C}_{5} = \mathbf{792}$ ways.

CHAPTER 2

Probability

1 Gambling and Probability

Gambling is the wagering of money or something of value on an event with an uncertain outcome, with the primary intent of winning money or material goods. The passion for gambling is as old as humanity itself. In places such as China, Egypt, Greece, Rome, there are evidences that date back to thousands of years ago. Dice games date back to 500BC in ancient rome. Playing cards were found in China as early as the 9th Century during the Tang Dynasty. The first casinos opened in Italy in the 17th century, the first ever being the Casino di Venezia in 1638.
Mans love for gambling is what led the way for the early developments in probability theory. We wanted to make more money when gambling, so we searched for optimal gambling strategies. Many developments in probability theory were stimulated by solving gambling problems such as

- De Mere's problem
- Newton - Pepys problem
- The St. Petersberg Paradox

These days, the magic of gambling has somewhat dissipated. There is no more uncertainty in the games of chance we play. Whether it is blackjack, roulette, or slot machines; one thing is for sure: in the long run, the house always wins. The gambler will inevitably fall victim of the law of large numbers.

2 Introduction to Probability

Probability is a measure of chance. In the case where an experiment has finitely many outcomes and all outcomes are equally likely to occur, the probability of an event (set of outcomes) is just the ratio of the number of outcomes in the event to the total number of outcomes (which we call the sample space).

> A **sample space** is the set of all possible outcomes of a random process or experiment.
>
> An **event** is a subset of a sample space.

For example, a fair six-sided die has a sample of $S = \{1,2,3,4,5,6\}$.
$\therefore n(S) = 6$
Suppose E is the event that an even number is rolled. Since the even numbers in S are $2,4,6$, event E can be written as $E = \{2,4,6\}$.
$\therefore n(E) = 3$

> If S is a finite sample space in which all outcomes are equally likely and E is an event in S, then the probability of E, denoted $\mathrm{P}(E)$, is
>
> $$\mathrm{P}(E) = \frac{\text{the number of outcomes in } E}{\text{the total number of ountcomes in } S}$$

We can thus find that the probability of event E, rolling an even number on a die, is

$$\mathrm{P}(E) = \frac{n(E)}{n(S)} = \frac{3}{6} = \frac{1}{2}$$

- For any event E, $0 \leq \mathrm{P}(E) \leq 1$.
- If E is an impossible event, then $\mathrm{P}(E) = 0$, i.e. it will never occur.
- If E is an certain event, then $\mathrm{P}(E) = 1$, i.e. it will definitely occur.
- If E is any event, then $\mathrm{P}(E) = 1 - \mathrm{P}(E')$, where $\mathrm{P}(E')$ is the probability that E does not occur.

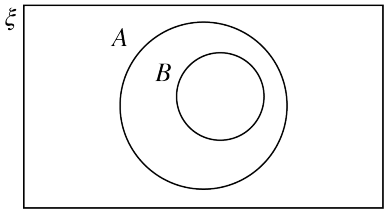
Notations	What it means
$A \cup B$	Union of A and B
$A \cap B$	Intersection of A and B
$\in$	"... is an element of ..."
$\notin$	"... is not an element of ..."
$n(A)$	Number of elements in set A
A'	Complement of set A
$\emptyset$	Empty set or null set
ξ	Universal set
$A \subseteq B$	A is a subset of B
$A \subset B$	A is a proper subset of B
$A \not\subseteq B$	A is not a subset of B
$A \not\subset B$	A is not a proper subset of B

3 Venn Diagrams

A Venn diagram consists of a universal set ξ represented by a rectangle. Sets within the universal set are represented by circles.

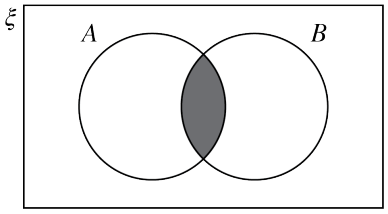
Subset

If $B \subseteq A$ then every element of B is also in A. The circle representing B is placed within the circle representing A and does not leave its boundaries.



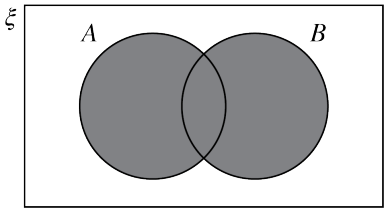
Intersection

$A \cap B$ consists of all elements common to both A and B. It is the shaded region where the circles representing A and B overlap.



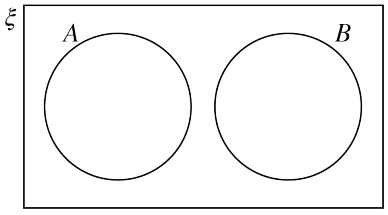
Union

$A \cup B$ consists of all elements which are in A or B. It is the shaded region which includes both circles.



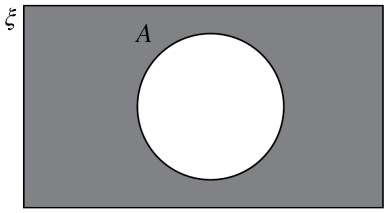
Disjoint or Mutually Exclusive

Disjoint sets do not have common elements. They are represented by non-overlapping circles.



Complement

A' is called the complement of A. It is the shaded region which includes everything except the set A.



4 Independent Events

Independent events are events where the occurrence of one of the events **does not** affect the occurrence of the other event.

For example, if a coin and a die are tossed simultaneously, there is no way that the outcome of one will affect the outcome of the other. Thus, the two events "getting a heads" and "rolling a six" are independent events.

If A and B are **independent events** then $\mathrm{P}(A \textbf{ and } B) = \mathrm{P}(A) \times \mathrm{P}(B)$

Example 2.1 Determining Whether Events are Independent

Two ordinary fair dice, one red and one blue, are thrown. Events A, B and C are defined as follows:
Event A: the number showing on the red die is 5 or 6
Event B: the total of the numbers showing on the two dice is 7
Determine whether events A and B are independent.

Solution

$$\mathrm{P}(A) = \frac{2}{6} = \frac{1}{3} \qquad \mathrm{P}(B) = \frac{6}{36} = \frac{1}{6} \qquad \mathrm{P}(A \cap B) = \frac{2}{36} = \frac{1}{18}$$

Since $\mathrm{P}(A \cap B) = \mathrm{P}(A) \times \mathrm{P}(B) = \frac{1}{18}$, therefore A and B are independent.

5 Dependent Events

Suppose a hat contains 4 red and 2 blue tickets. One ticket is randomly chosen, its colour is noted and it is thrown in a bin. A second ticket is randomly selected. What is the chance that it is red?

If the first ticket was red, P(second is red) $= \frac{3}{5}$

If the first ticket was blue, P(second is red) $= \frac{4}{5}$

So, the probability of the second ticket being red **depends** on what colour the first ticket was. Here we have **dependent events**.

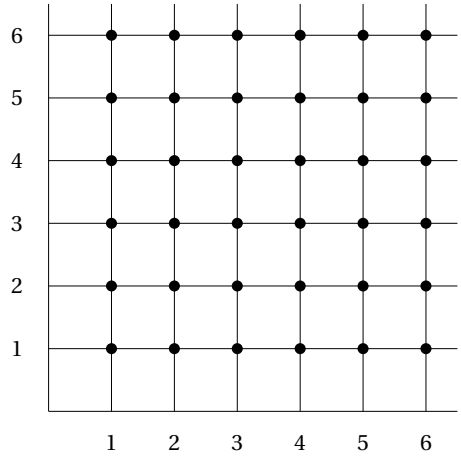
Dependent events are events where the occurrence of one of the events **does** affect the occurrence of the other event.

6 Simple Combined Events

In this section, we will learn two ways of illustrating the sample space of an experiment involving two or more objects and calculate probabilities for simple combined events.

6.1 Possibility Diagrams

A fair die has the numbers $1,2,3,4,5,6$. It is rolled *twice*. We can illustrate the sameple space using a 2D grid, known as a **possibility diagram**.



From the above possibility diagram, we can see that that the total number of possible outcomes is $6 \times 6 = 36$. Using this diagram will make it easier for us to find the probability of certain events.

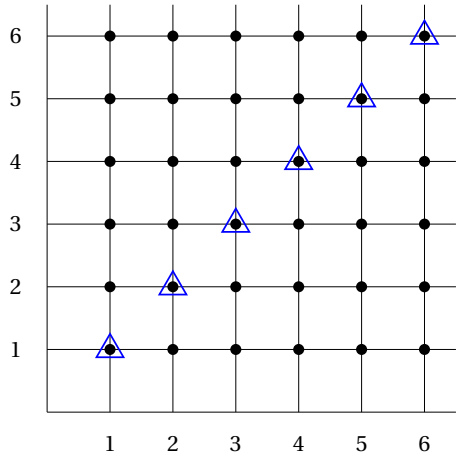
Example 2.2 Possibility Diagram for Dice

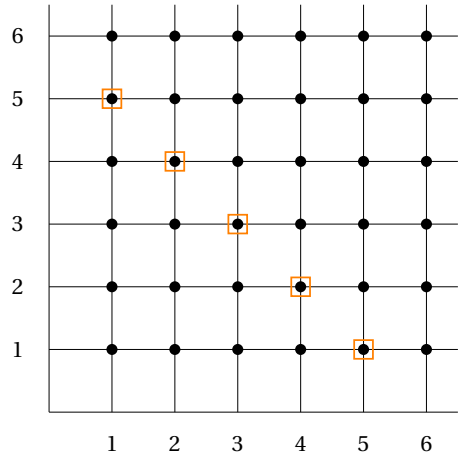
Find the probability of rolling:

(a) the same number on both die

(b) a total of 6

Solution



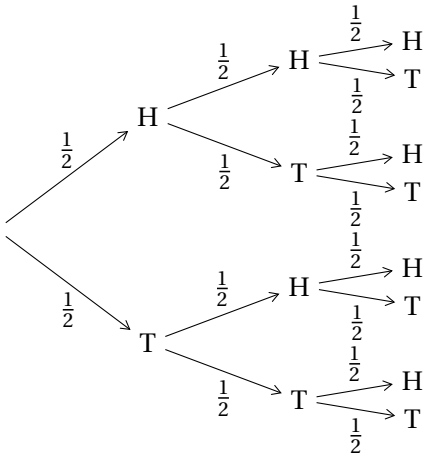


(a) P(same number on both dice) $= \frac{6}{36} = \frac{1}{6}$

(b) P(total of 6) $= \frac{5}{36}$

6.2 Tree Diagrams

Tree diagrams can be used to illustrate sample spaces, provided that the alternatives are not too numerous. Below we see the tree diagram for tossing three fair coins.



From the above tree diagram, we can see that that the total number of possible outcomes is $2 \times 2 \times 2 = 8$. Tree diagrams are useful for illustrating the sample space when we have *more than two* events occurring in sequence.

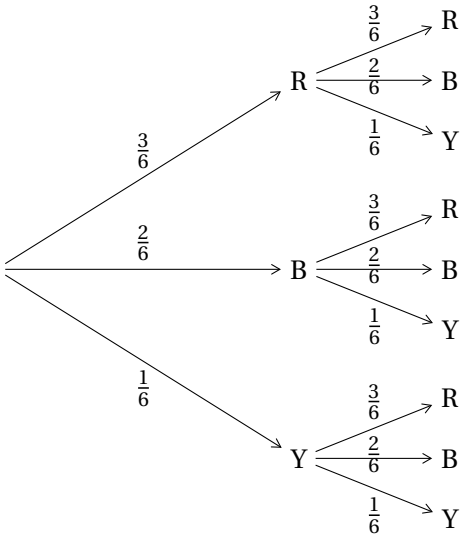
7 Sampling With and Without Replacement

Sampling is the process of selecting an object from a large group of objects and inspecting it, noting some feature(s). The object is then either **put back** (sampling **with replacement**) or **put to one side** (sampling **without replacement**).
Consider a box containing 3 red, 2 blue and 1 yellow marble. Suppose we wish to sample two marbles:

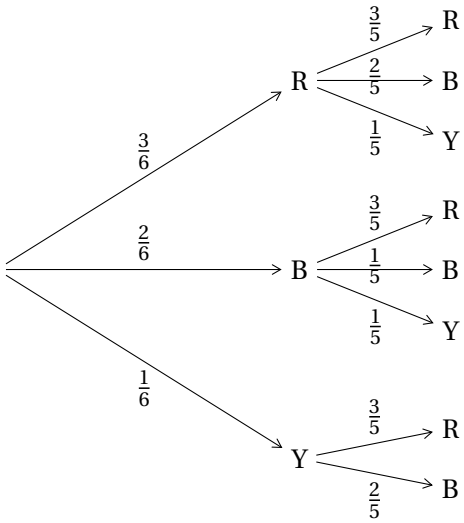
- with replacement of the first before the second is drawn
- without replacement of the first before the second is drawn

Examine how the tree diagrams differ:

With relplacement



Without Replacement



8 Laws of Probability

8.1 Addition Law of Probability

The **addition law of probability** states that for two events A and B,

$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$$

This can be also written as

$$\mathrm{P}(A \textbf{ or } B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \textbf{ and } B)$$

Sometimes this law is referred to as the **Inclusion/Exclusion Rule**.

8.2 Mutually Exclusive Events

Two events are said to be **mutually exclusive** if they **cannot happen at the same time**.

For example, consider the sample space of drawing one card from an ordinary deck of 52 cards.

- A is the event that a spade is drawn,
- B is the event that a heart is drawn and
- C is the event that an ace is drawn.

Events A and B are mutually exclusive because both events cannot happen at the same time.
Events A and C are not mutually exclusive because we can draw the Ace of Spades.
Events B and C are also not mutually exclusive.

If A and B are mutually exclusive, then $\mathrm{P}(A \cap B) = 0$. Thus the addition law for mutually exclusive events becomes

$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B)$$

Example 2.3 Conditions for Mutually Exclusive Events and Independent Events

The events A and B are such that $\mathrm{P}(A) = 0.43$, $\mathrm{P}(B) = 0.48$, and $\mathrm{P}(A \cup B) = 0.78$. Show that A and B are neither mutually exclusive nor independent.

Solution

$$\begin{aligned}\mathrm{P}(A \cup B) &= \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)\\ \mathrm{P}(A \cap B) &= \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cup B)\\ &= 0.43 + 0.48 - 0.78\\ &= 0.13 \neq 0\end{aligned}$$

Hence, A and B are not mutually exclusive.

$$\begin{aligned}\mathrm{P}(A) \times \mathrm{P}(B) &= 0.43 \times 0.48\\ &= 0.2064 \neq 0.13 = \mathrm{P}(A \cap B)\end{aligned}$$

Hence , A and B are not independent.

8.3 Conditional Probability

Suppose we have two events A and B, then

$A \mid B$ is used to represent that "A occurs **given** B has occurred".

The **conditional probability** of event A occurring, given that event B has occurerd is given by

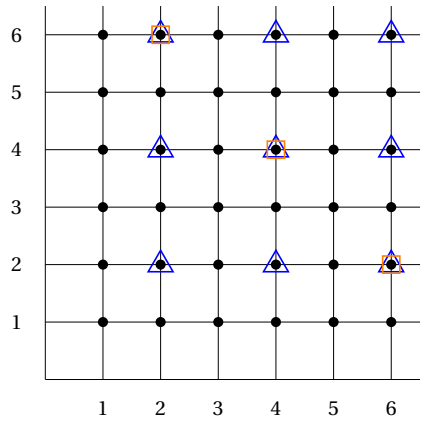
$$\mathrm{P}(A \mid B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}$$

Example 2.4 Conditional Probability With Dice

A pair of fair dice are rolled. What is the probability that the sum of the numbers showing face up is 8, given that both of the numbers are even?

Solution

$$\begin{aligned}\mathrm{P}(\text{sum of 8} \mid \text{both are even}) &= \frac{\mathrm{P}(\text{sum of 8 } \textbf{and} \text{ both are even})}{\mathrm{P}(\text{both are even})} \\ &= \frac{\frac{3}{36}}{\frac{9}{36}} \\ &= \frac{1}{3}\end{aligned}$$



It follows that

$$\mathrm{P}(A \cap B) = \mathrm{P}(B) \times \mathrm{P}(A \mid B) = \mathrm{P}(A) \times \mathrm{P}(B \mid A)$$

A and B are **independent events** if the occurrence (or non-occurrence) of one event does not affect the occurrence of the other,

$$\text{i.e.} \quad \mathrm{P}(A \mid B) = \mathrm{P}(A) \quad \text{and} \quad \mathrm{P}(B \mid A) = \mathrm{P}(B)$$

So,

A and B are **independent events** $\Leftrightarrow \mathrm{P}(A \cap B) = \mathrm{P}(A) \times \mathrm{P}(B)$

Example 2.5 Using Tree Diagrams

In an archery competition, Bill is allowed up to three attempts to hit the target. If he succeeds on any attempt, he does not make any more attempts. The probability that he will hit the target on the first attempt is 0.6. If he misses, the probability that he will hit the target on his second attempt is 0.7. If he misses the second attempt, the probability that he will hit the target on his third attempt is 0.8.

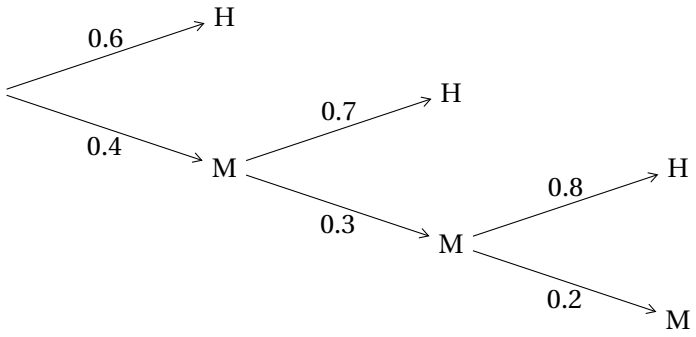
(a) Draw a fully labelled tree diagram.

(b) Find the probability that Bill hits the target.

(c) Given that Bill hits the target, find the probability that he made at least two attempts.

Solution

(a)



H: Hit, M: Miss

(b) $\mathrm{P}(\text{Bill hits target}) = 0.6 + 0.4 \times 0.7 + 0.4 \times 0.3 \times 0.8$

$= 0.976$

(c) $\mathrm{P}(\text{at least 2 attempts} \mid \text{hits target}) = \dfrac{\mathrm{P}(\text{at least 2 attempts and hits target})}{\mathrm{P}(\text{hits target})}$

$= \dfrac{0.976 - 0.6}{0.976}$

$= \dfrac{47}{122}$

9 Probabilities Using Permutations and Combinations

Permutations and combinations can sometimes be used to find probabilities of various events particularly when large sample sizes occur. It is useful to remember that

$$\mathrm{P}(A) = \frac{\text{Total number of ways for event } A \text{ to occur}}{\text{Total number of ways to perform the experiment}}$$

For example, if we select at random a team of 4 boys and 3 girls from a squad of 8 boys and 7 girls, the total number of unrestricted possibilities is ${}^{15}C_7$. The number of combinations with the restriction of "4 boys and 3 girls" is ${}^{8}C_4 \times {}^{7}C_3$.

$$\therefore \mathrm{P}(\text{4 boys and 3 girls}) = \frac{{}^{8}C_4 \times {}^{7}C_3}{{}^{15}C_7}$$

The biggest hurdle in probability problems involving permutations or combinations seems to be in sorting out which to use. Remember

- **permutations** involve **arrangements** of letters/people/things, whereas
- **combinations** involve **selections** such as committees/teams/delegations.

Example 2.6 Finding Probabilities Using PnC

5 girls and 7 boys are to be seated in a row. What is the probability that

(a) not all the girls are seated next to one another?

(b) all the girls are not seated next to one another?

(c) 3 particular boys are to be seated together?

(d) either all the girls are not seated next to one another or the 3 particular boys are to be seated together or both?

Solution

(a) Group the 5 girls as 1 unit.

$$\begin{aligned}\mathrm{P}(\text{not all the girls are seated next to one another}) &= 1-\mathrm{P}(\text{all girls are seated next to each other})\\ &= 1-\frac{8!\times 5!}{12!}\\ &= \frac{98}{99}\end{aligned}$$

(b) _ B _ B _ B _ B _ B _ B _

$$\begin{aligned}\mathrm{P}(\text{all the girls are not seated next to one another}) &= \frac{7!\times {}^{8}P_{5}}{12!}\\ &= \frac{7}{99}\end{aligned}$$

(c) Group the 3 boys as 1 unit.

$$\begin{aligned}\mathrm{P}(\text{3 particular boys are seated together}) &= \frac{10!\times 3!}{12!}\\ &= \frac{1}{22}\end{aligned}$$

(d) Let A and B denote the events that all the girls are not seated next to one another and that the 3 particular boys are seated together respectively. Then

$$\begin{aligned}\mathrm{P}(A\cap B) &= \frac{5!\times 3!\times {}^{6}P_{5}}{12!}\\ &= \frac{1}{924}\end{aligned}$$

Hence,

$$\begin{aligned}\mathrm{P}(A\cup B) &= \mathrm{P}(A)+\mathrm{P}(B)-\mathrm{P}(A\cap B)\\ &= \frac{7}{99}+\frac{1}{22}-\frac{1}{924}\\ &= \frac{29}{252}\end{aligned}$$

10 De Meres Problem

Gamblers in 1654 France used to bet on the event of getting at least one six in four rolls of a dice. As a more trying variation, two die were rolled 24 times with a bet on having at least one double six. De Méré's problem is as follows:

Example 2.7 De Meres Problem

Is it more likely to get at least one double six in 24 rolls of a pair of dice or to get at least one six in four rolls of a single die?

Solution

$$\begin{aligned} \text{P (at least one six in four rolls)} &= 1 - \text{P (no six in four rolls)} \\ &= 1 - \left(\frac{5}{6}\right)^4 \\ &= 0.518 \text{ (3 s.f.)} \end{aligned}$$

$$\begin{aligned} \text{P (at least one double six in 24 rolls)} &= 1 - \text{P (no double six in 24 rolls)} \\ &= 1 - \left(\frac{35}{36}\right)^{24} \\ &= 0.491 \text{ (3 s.f.)} \end{aligned}$$

Thus is it more likely that we roll at least one six in four rolls.

11 Newton-Pepys Problem

Example 2.8 Newton-Pepys Problem

In 1693 Samuel Pepys and Isaac Newton corresponded over a problem posed by Pepys in relation to a wager he planned to make. Pepys asked which was more likely,

A: At least one six when six dice are rolled,

B: At least two sixes when 12 dice are rolled, or

C: At least three sixes when 18 dice are rolled.

Pepys initially thought that outcome C had the highest probability. Check to see if he was right or wrong by finding each of the corresponding probabilities.

Solution

$$\begin{aligned} \text{P (at least one six in six rolls)} &= 1 - \text{P (no six in six rolls)} \\ &= 1 - \left(\frac{5}{6}\right)^6 \\ &= 0.665 \text{ (3 s.f.)} \end{aligned}$$

$$\begin{aligned}
\text{P(at least two sixes in 12 rolls)} &= 1 - \text{P(less than two sixes in 12 rolls)} \\
&= 1 - [\text{P(no six in 12 rolls)} + \text{P(one six in 12 rolls)}] \\
&= 1 - \left[\left(\frac{5}{6}\right)^{12} + \left(\frac{1}{6}\right)\left(\frac{5}{6}\right)^{11} \times {}^{12}C_1\right] \\
&= 0.619 \text{ (3 s.f.)}
\end{aligned}$$

$$\begin{aligned}
\text{P(at least three sixes in 18 rolls)} &= 1 - \text{P(less than three sixes in 18 rolls)} \\
&= 1 - [\text{P(no six in 18 rolls)} + \text{P(one six in 18 rolls)} + \text{P(two sixes in 18 rolls)}] \\
&= 1 - \left[\left(\frac{5}{6}\right)^{18} + \left(\frac{1}{6}\right)\left(\frac{5}{6}\right)^{17} \times {}^{18}C_1 + \left(\frac{1}{6}\right)^{2} \times \left(\frac{5}{6}\right)^{16} \times {}^{18}C_2\right] \\
&= 0.597 \text{ (3 s.f.)}
\end{aligned}$$

12 The St. Petersberg Paradox

A casino offers a game of chance for a single player in which a fair coin is tossed at each stage. The initial stake begins at 2 dollars and is doubled every time heads appears. The first time tails appears, the game ends and the player wins whatever is in the pot. Thus the player wins 2 dollars if tails appears on the first toss, 4 dollars if heads appears on the first toss and tails on the second, 8 dollars if heads appears on the first two tosses and tails on the third, and so on.

Mathematically, the player wins 2^k dollars, where k is the number of consecutive head tosses. What would be a fair price to pay the casino for entering the game?

To answer this, one needs to consider what would be the expected payout at each stage: with probability $\frac{1}{2}$ the player wins 2 dollars; with probability $\frac{1}{4}$ the player wins 4 dollars; with probability $\frac{1}{8}$ the player wins 8 dollars, and so on. Assuming the game can continue as long as the coin toss results in heads and, in particular, that the casino has unlimited resources, the expected value is thus

$$\begin{aligned}
\text{E}(X) &= \frac{1}{2} \times 2 + \frac{1}{4} \times 4 + \frac{1}{8} \times 8 + \frac{1}{16} \times 16 + \ldots \\
&= 1 + 1 + 1 + 1 + \ldots \\
&= \infty
\end{aligned}$$

Considering nothing but the expected value of the net change in one's monetary wealth, one should therefore play the game at any price if offered the opportunity. In a strict logical sense, the St. Petersburg paradox is not a paradox because no formal contradiction is derived. However, to claim that a rational agent should pay millions, or even billions, for playing this game seems absurd. Few of us would be willing to pay even $20 to enter such a game.

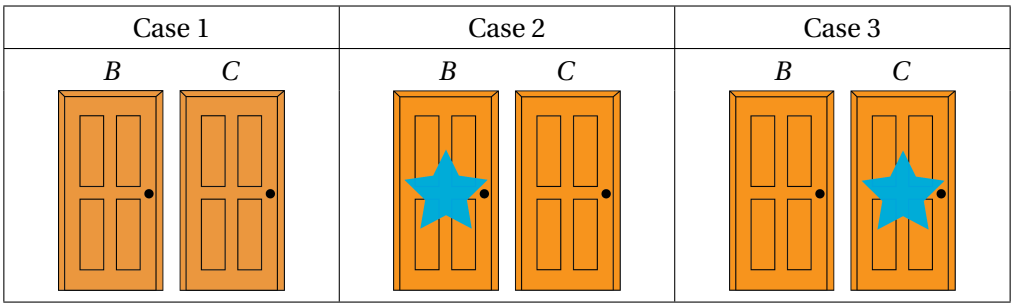
13 The Monty Hall Problem

The next example is called the Monty Hall problem, named for the first host of the game show "Let's Make A Deal." When it was originally publicized in a newspaper column and on a radio show, it created tremendous controversy. Many highly educated people, even some with Ph.D.'s, submitted incorrect solutions or argued vociferously against the correct solution. Before you read the answer, think about what your own response to the situation would be.

Example 2.9 The Monty Hall Problem

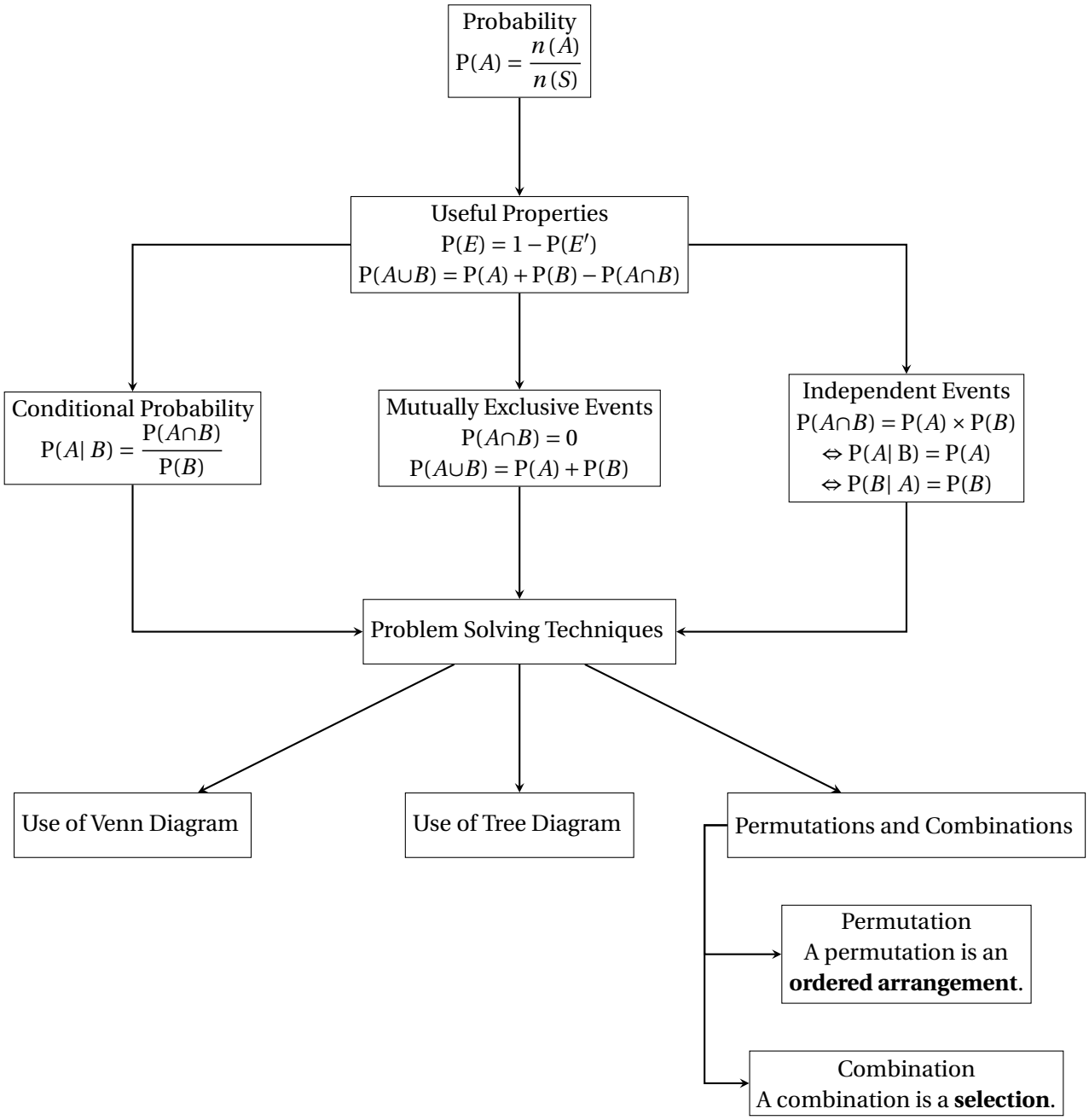
There are three doors on the set for a game show. Let's call them A, B, and C. Behind one of the doors, there is a car. If you pick the correct door, you win the car. You pick door A. The host of the show then opens one of the other doors and reveals that there is no car behind it. Keeping the remaining two doors closed, he asks you whether you want to switch your choice to the other closed door or stay with your original choice of door A. What should you do if you want to maximize your chance of winning the car: stay with door A or switch—or would the likelihood of winning be the same either way?

Solution



At the point just before the host opens one of the closed doors, there is no information about the location of the prize. Thus there are three equally likely possibilities for what lies behind the doors: (Case 1) the prize is behind A; (Case 2) the prize is behind B; or (Case 3) the prize is behind C. Since there is no prize behind the door the host opens, in Case 1 the host could open either door and you would win by staying with your original choice: door A. In Case 2 the host must open door C, and so you would win by switching to door B. In Case 3 the host must open door B, and so you would win by switching to door C. Thus, in two of the three equally likely cases, you would win by switching from A to the other closed door. In only one of the three equally likely cases would you win by staying with your original choice. Therefore, **you should switch**.

Probability Summary



CHAPTER 3

Discrete Random Variables

1 Discrete Random Variables

A **random variable** is a numerical quantity that is generated by a random experiment, we denote random variables by capital letters such as X.

A random variable can be **discrete** or **continuous**.

A random variable X is called **discrete** if it has a **countable** number of possible values.

For example, X could be:

- the number of rotten apples in a basket of fuits,
- the number of males in a group of 5 students,
- the number of new bicycles sold each year by a bicycle store.

A random variable X is called **continuous** if it takes on a **range** of values.

For example, X could be:

- the height of a randomly selected student,
- the weight of a randomly selected bag of sugar,
- the time taken for a randomly selected student to complete a 2.4 km run.

Consider the experiment of tossing a fair coin 3 times and observing the outcome.
The sample space may be represented by

$$\{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$$

Let X denote the number of heads obtained when tossing a fair coin 3 times.

- The possible values (variable) X can take are 0, 1, 2, 3 (discrete). The value it assumes is subject to chance (random). Thus X is a discrete random variable.
- $\mathrm{P}(X = x)$ refers to the probability of the discrete random variable X assuming a specific value x.
- E.g. $\mathrm{P}(X = 2)$ refers to the probabiltiy of obtaining two heads in 3 tosses. In this case, $\mathrm{P}(X = 2) = \frac{3}{8}$

2 Probability Distribution

The **probability distribution** is a table that lists down each probability $\mathrm{P}(X = x)$ for all possible values of x.

x			
$\mathrm{P}(X = x)$			

For the experiment of tossing a fair coin 3 times where X is the number of heads obtained, the probability distribution of X is as follows

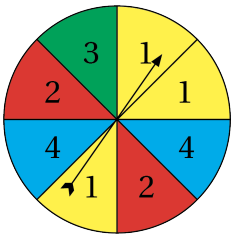
x	0	1	2	3
$\mathrm{P}(X = x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

For a discrete random variable X, the sum of probabilities is **1**:

$$\sum_{\text{all } x} \mathrm{P}(X = x) = 1$$

Example 3.1 Probability Distribution For a Spinner

Let X be the result when the spinner alongside is spun.



(a) Display the probability distribution of X in a table.

(b) Find $\mathrm{P}(X \leq 3)$.

(c) Find $\mathrm{P}(1 < X < 4)$.

Solution

(a)

x	1	2	3	4
$\mathrm{P}(X = x)$	$\frac{3}{8}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{4}$

(b) $\mathrm{P}(X \leq 3) = 1 - \mathrm{P}(X = 4)$

$$= 1 - \frac{1}{4}$$
$$= \frac{3}{4}$$

(c) $\mathrm{P}(1 < X < 4) = \mathrm{P}(X = 2) + \mathrm{P}(X = 3)$

$$= \frac{1}{4} + \frac{1}{8}$$
$$= \frac{3}{8}$$

Example 3.2 Probability Distribution Function

The probabilities that a discrete random variable X takes are given by $\mathrm{P}(X = x) = cx^2$ where $x \in \{1,2,3,4\}$. Given that c is a constant, find the value of c. Hence display the probability distribution of X.

Solution

$$\mathrm{P}(X = 1) = c(1)^2$$
$$\mathrm{P}(X = 2) = c(2)^2$$
$$\mathrm{P}(X = 3) = c(3)^2$$
$$\mathrm{P}(X = 4) = c(4)^2$$

Since the sum of probabilities is equal to 1,

$$c(1)^2 + c(2)^2 + c(3)^2 + c(4)^2 = 1$$
$$30c = 1$$
$$c = \frac{1}{30}$$

x	1	2	3	4
$\mathrm{P}(X = x)$	$\frac{1}{30}$	$\frac{2}{15}$	$\frac{3}{10}$	$\frac{8}{15}$

3 Expected Value of a Random Variable

The **expected value** or **mean** of a discrete random variable of X, $E(X)$, is given by:

$$\mathrm{E}(X) = \sum_{\text{all } x} x\mathrm{P}(X = x)$$

$\mathrm{E}(X)$ is also denoted by the symbol μ.

Example 3.3 Finding Expected Value Using a Probability Distribution

Find $\mathrm{E}(X)$ for the probability distribution in Example 2.2.

Solution

x	1	2	3	4
$\mathrm{P}(X = x)$	$\frac{1}{30}$	$\frac{2}{15}$	$\frac{3}{10}$	$\frac{8}{15}$

$$\begin{aligned}\mathrm{E}(X) &= 1 \times \frac{1}{30} + 2 \times \frac{2}{15} + 3 \times \frac{3}{10} + 4 \times \frac{8}{15} \\ &= \frac{10}{3}\end{aligned}$$

Example 3.4 Finding Unknowns in a Probability Distribution

Given that $\mathrm{E}(X) = 2.5$, find a and b.

x	1	2	3	4
$\mathrm{P}(X = x)$	0.3	a	b	0.2

Solution

Since the sum of probabilities is equal to 1,

$$\begin{aligned}0.3 + a + b + 0.2 &= 1 \\ a + b &= 0.5 \qquad (1)\end{aligned}$$

Since $\mathrm{E}(X) = 2.5$,

$$\begin{aligned}1 \times 0.3 + 2 \times a + 3 \times b + 4 \times 0.2 &= 2.5 \\ 2a + 3b &= 1.4 \qquad (2)\end{aligned}$$

Solving (1) and (2), $a = 0.1$ and $b = 0.4$.

Example 3.5 Game of Chance

In a game of chance, a player spins a square spinner labelled $1, 2, 3, 4$. The player wins the amount of money according to the table below.

Number	1	2	3	4
Winnings	\$1	\$2	\$5	\$8

(a) Find the expected payout for one spin of the spinner.

(b) Find the expected gain for one spin if it costs \$5 to play each game.

(c) Discuss whether you would reccomend playing this game.

Solution

(a) Let X denote the payout from one spin.

Each outcome is equally likely, so the probability of each outcome is $\frac{1}{4}$.

$$\begin{aligned} \mathrm{E}(X) &= \frac{1}{4} \times 1 + \frac{1}{4} \times 2 + \frac{1}{4} \times 5 + \frac{1}{4} \times 8 \\ &= \$4 \end{aligned}$$

(b) Let Y denote the gain of the player from each game.

Since it costs \$5 to play the game, the expected gain is

$$\begin{aligned} \mathrm{E}(Y) &= \mathrm{E}(X) - 5 \\ &= 4 - 5 \\ &= -\$1 \end{aligned}$$

(c) Since $\mathrm{E}(Y) = -\$1$, we expect the player to lose \$1 on average for each spin. We would not recommend a person play the game (unless the fun of each spin is worth more than \$1 to them).

Example 3.6 Expectation of X

Find c and $\mathrm{E}(X)$.

x	-2	-1	0	1	2
$\mathrm{P}(X = x)$	0.3	c	0.15	0.4	0.05

Solution

Since the sum of probabilities is equal to 1,

$$\begin{aligned} 0.3 + c + 0.15 + 0.4 + 0.05 &= 1 \\ c &= 0.1 \end{aligned}$$

$$\begin{aligned} \mathrm{E}(X) &= (-2 \times 0.3) + (-1 \times 0.1) + (0 \times 0.15) + (1 \times 0.4) + (2 \times 0.05) \\ &= -0.2 \end{aligned}$$

Let X be a discrete random variable which takes values from a set S.
If $g(X)$ is a function of X, then $g(X)$ is also a discrete random variable and

$$\mathrm{E}\left[g(x)\right] = \sum_{\text{all } x} g(x)\,\mathrm{P}(X = x)$$

where the summation is over all elements x in S.
For any random variables X and Y,

(a) $\mathrm{E}(a) = a$

(b) $\mathrm{E}(aX) = a\mathrm{E}(X)$

(c) $\mathrm{E}(aX \pm b) = a\mathrm{E}(X) \pm b$

(d) $\mathrm{E}(aX \pm bY) = a\mathrm{E}(X) \pm b\mathrm{E}(Y)$

where a and b are constants.

Example 3.7 Expectation of $g(X)$

With X as defined in Example 2.6, find $\mathrm{E}\left(X^2\right)$ and $\mathrm{E}(|X+1|)$.
Hence, find the value of $\mathrm{E}\left(3X^2 - 2|X+1|\right)$.

Solution

x	-2	-1	0	1	2
x^2	4	1	0	1	4
$\lvert x+1\rvert$	1	0	1	2	3
$\mathrm{P}(X = x)$	0.3	0.1	0.15	0.4	0.05

$$\begin{aligned}\mathrm{E}\left(X^2\right) &= \sum_{\text{all } x} x^2\mathrm{P}(X = x)\\ &= (4 \times 0.3) + (1 \times 0.1) + (0 \times 0.15) + (1 \times 0.4) + (4 \times 0.05)\\ &= 1.9\end{aligned}$$

$$\begin{aligned}\mathrm{E}(|X+1|) &= \sum_{\text{all } x} |x+1|\,\mathrm{P}(X = x)\\ &= (1 \times 0.3) + (0 \times 0.1) + (1 \times 0.15) + (2 \times 0.4) + (3 \times 0.05)\\ &= 1.4\end{aligned}$$

$$\begin{aligned}\mathrm{E}\left(3X^2 - 2|X+1|\right) &= 3\mathrm{E}\left(X^2\right) - 2\mathrm{E}(X+1)\\ &= 3(1.9) - 2(1.4)\\ &= 2.9\end{aligned}$$

4 Variance and Standard Deviation or a Random Variable

For any random variable X with $E(X) = \mu$, the variance of X, $\text{Var}(X)$, is given by:

$$\text{Var}(X) = \text{E}\left(X - \mu\right)^2$$

Alternatively,

$$\text{Var}(X) = \text{E}\left(X^2\right) - \mu^2$$

$\text{Var}(X)$ is also denoted by the symbol σ^2.
Note that $\text{Var}(X)$ is a non-negative value.
The standard deviation, σ, of X is given by

$$\sigma = \sqrt{\text{Var}(X)}$$

For any random variable X and Y,

(a) $\text{Var}(a) = 0$

(b) $\text{Var}(aX) = a^2\text{Var}(X)$

(c) $\text{Var}(aX \pm b) = a^2\text{Var}(X)$

(d) If X and Y are **independent**, then
$\text{Var}(aX \pm bY) = a^2\text{Var}(X) + b^2\text{Var}(Y)$

where a and b are constants.
Consider a random variable X. Take n observations $X_1, X_2, \ldots, X_n$ from X, then

$$\text{E}(X_1 + X_2 + \ldots + X_n) = n\text{E}(X)$$

If $X_1, X_2, \ldots, X_n$ are independent, then

$$\text{Var}(X_1 + X_2 + \ldots + X_n) = n\text{Var}(X)$$

Example 3.8 Variance and Standard Deviation of X

With X as defined in Example 2.6, find the variance and standard deviation of X.
Solution

$$\begin{aligned}\text{Var}(X) &= \text{E}\left(X^2\right) - \mu^2 \\ &= 1.9 - (-0.2)^2 \\ &= 1.86\end{aligned}$$

$$\begin{aligned}\sigma &= \sqrt{\text{Var}(X)} \\ &= \sqrt{1.86} \\ &= 1.36 \text{ (3 s.f.)}\end{aligned}$$

CHAPTER 4

Binomial Distribution

1 Binomial Experiments

Experiments consisting of n independent trials, each with two possible outcomes that may be regarded as either success or failure, are very common in the study of probability. If, in addition, the probability of getting a success (or a failure) at each trial remains constant and the trials are independent, then we call such experiments **binomial experiments**.
For example, when we toss a coin 10 times, the outcome of each toss may be a head or a tail. Let us regard getting a head as a success. Since we are using the same coin, the probability of success will remain constant. Obviously, the tosses are independent of each other. This coin tossing experiment is then a binomial experiment.

The **characteristics of a binomial distribution** are as follows:

1. The experiment has n repeated and independent trials.
2. Each trial has only two possible outcomes, "success" and "failure".
3. The probability of success for each trial, p, remains constant.

Example 4.1 Binomial Experiments

For which of these probability experiments does the binomial distribution apply? Explain your answers.

(a) A coin is thrown 100 times. The variable is the number of heads.

(b) 5 cards are drawn from a deck of cards one at a time, replacing the card before the next one is drawn. The variable is the number of Queens drawn.

(c) 3 marbles are drawn from a bag of 5 blue marbles and 4 red marbles, one at a time, without replacing the marble before the next is drawn. The variable is the number of red marbles drawn.

(d) A large bin contains ten thousand bolts, 1% of which are faulty. A sample of 10 bolts is drawn from the bin. The variable is the number of faulty bolts.

Solution

(a) Has 100 repeated and independent trials. ✓

Has only two possible outcomes (H or T). ✓

The probability of getting a heads remains constant. ✓

This is a binomial experiment.

(b) Has 5 repeated and independent trials. ✓

Has only two possible outcomes (queen or not queen). ✓

The probability of getting a queen remains constant. ✓

This is a binomial experiment.

(c) The binomial distribution does not apply as the result after the first draw is dependent on the results of previous draws.

(d) The binomial distribution does not apply as the 10 bolts are drawn without replacement. We do not have a repetition of independent trials. However, since we have such a large number of bolts in the bin, the trials are approximately independent, so the distribution is approximately binomial.

2 The Binomial Distribution

If $X \sim \mathrm{B}(n, p)$, then the probability distribution function of X is given by

$$\mathrm{P}(X = r) = \binom{n}{r} p^r (1-p)^{n-r}$$

where

- n is the number of trials,
- p is the probability of success
- $r = 0, 1, 2, \ldots, n$

If $X \sim \mathrm{B}(n, p)$, using the graphic calculator to find:

- $\mathrm{P}(X = r)$

  Step 1: Press 2ND VARS

  Step 2: Select A:binompdf and key the values in the format n, p, r.

- $\mathrm{P}(X \leq r)$

  Step 1: Press 2ND VARS

  Step 2: Select B:binomcdf and key the values in the format n, p, r.

Example 4.2 Probabilities of a Fair Coin

A fair coin is flipped 5 times. If X represents the number of heads obtained, find $\mathrm{P}(X = x)$ for $x = 0, 1, 2, 3, 4, 5$.

Solution

$X \sim \mathrm{B}\left(5, \frac{1}{2}\right)$

$$\mathrm{P}(X=0) = \binom{5}{0}\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^5 = \frac{1}{32} \qquad \mathrm{P}(X=3) = \binom{5}{3}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^2 = \frac{10}{32}$$

$$\mathrm{P}(X=1) = \binom{5}{1}\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^4 = \frac{5}{32} \qquad \mathrm{P}(X=4) = \binom{5}{4}\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right)^1 = \frac{5}{32}$$

$$\mathrm{P}(X=2) = \binom{5}{2}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^3 = \frac{10}{32} \qquad \mathrm{P}(X=5) = \binom{5}{5}\left(\frac{1}{2}\right)^5\left(\frac{1}{2}\right)^0 = \frac{1}{32}$$

Example 4.3 Probabilities of a Fair Die

A fair die is tossed 5 times. Find the probability

(a) of obtaining 2 sixes,

(b) that not more than 3 even numbers are obtained,

(c) that at least 1 prime number is obtained.

Solution

(a) $X \sim$ no. of sixes out of 5 tosses of the die

$X \sim \mathrm{B}\left(5, \frac{1}{6}\right)$

$$\mathrm{P}(X=2) = \binom{5}{2}\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^3$$
$$= 0.161 \text{ (3 s.f.)}$$

(We can also find this using our GC)

(b) $Y \sim$ no. of even numbers out of 5 tosses of the die

$Y \sim \mathrm{B}\left(5, \frac{1}{2}\right)$

From GC,

$$\mathrm{P}(Y \leq 3) = 0.8125$$

(c) $W \sim$ no. of prime numbers out of 5 tosses of the die

$W \sim \mathrm{B}\left(5, \frac{1}{2}\right)$

$$\mathrm{P}(W \geq 1) = 1 - \mathrm{P}(W = 0)$$
$$= 1 - 0.03125$$
$$= 0.96875$$

Example 4.4 Finding p

The probability that a shooter hits his target is p. The shots he makes are independent of each other and the probability of him hitting or missing the target remains constant. If the probability that he makes 4 out of 5 shots is 0.2, find p, given that $p > 0.7$.

Solution

$X \sim$ no. of shots that hit the target, out of 5
$X \sim \mathrm{B}(5, p)$

$$\mathrm{P}(X = 4) = 0.2$$
$$\binom{5}{4} p^4 (1-p) = 0.2$$
$$5(p^4 - p^5) = 0.2$$
$$p^5 - p^4 - 0.04 = 0$$
$$p = 0.951 \text{ or } p = 0.544 \text{ (rej. } \because p > 0.7)$$

Example 4.5 Probability That Shelly is Late For Work

Shelly must pass through 15 traffic lights on her way to work. She has probability 0.6 of being stopped at any given traffic light. If she is stopped at more than 11 traffic lights, she will be late.

(a) Find the probability that Shelly will be late for work on a given day.

(b) Find the probability that Shelly is on time for work each day of a 5 day work week.

(c) Shelly wants to increase the probability in (b) to at least 80%. She decides to leave a little earlier, so she must now be stopped at more than 12 traffic lights in order to be late. Has Shelly achieved her goal? Explain your answer.

Solution

(a) $X \sim$ no. of traffic lights Shelly is stopped at, out of 15

$X \sim \mathrm{B}(15, 0.6)$

$\mathrm{P}(X > 11) = 1 - \mathrm{P}(X \leq 11)$
$= 0.090502$ (5 s.f.)
$= 0.0905$ (3 s.f.)

(b) $Y \sim$ no. of days Shelly is late, out of 5

$Y \sim \mathrm{B}(5, 0.090502)$

$\mathrm{P}(Y = 0) = 0.622$ (3 s.f.)

(c) $\mathrm{P}(X > 12) = 1 - \mathrm{P}(X \leq 12)$
$= 0.027114$ (5 s.f.)

$Y \sim \mathrm{B}(5, 0.027114)$

$\mathrm{P}(Y = 0) = 0.872$ (3 s.f.)

Shelly has achieved her goal. The probability that Shelly is on time for work everyday in the week is now 87.2%.

3 Mean and Varianace of a Binomial Distribution

For a binomial distribution $X \sim \mathrm{B}(n, p)$,

$$\text{Expected value or mean of } X = \mathrm{E}(X) = \mu = np$$

$$\text{Variance of } X = \mathrm{Var}(X) = \sigma^2 = np(1-p)$$

$$\text{Standard Deviation of } X = \sqrt{\mathrm{Var}(X)} = \sigma = \sqrt{np(1-p)}$$

Example 4.6 Finding n and p

A binomial random variable X has mean 1.2 and variance 1.08. Evaluate the parameters of the binomial distribution.

Solution

$X \sim \mathrm{B}(n, p)$
Given $\mathrm{E}(X) = 1.2$,
$np = 1.2$
Given $\mathrm{Var}(X) = 1.08$,
$np(1-p) = 1.08$

$$(1-p) = \frac{1.08}{1.2}$$
$$p = 0.1$$

$$n(0.1) = 1.2$$
$$n = 12$$

$\therefore X \sim \mathrm{B}(12, 0.1)$

4 Binomial Distribution Problems Involving Probability Trees

Example 4.7 Pastry Chef

At a resturaunt, the pastries made by the junior chef have to be inspected by the senior chef. 4% of the pastries made by the junior chef do not meet the passing standard. The senior chef randomly selects a sample of 12 pastries from a batch made by the junior chef. If more than 2 pastries do not meet the passing standard, the entire batch of pastries is rejected. If less than 2 pastries do not meet the passing standard, the batch is accepted. If exactly 2 pastries do not meet the passing standard, a 2nd sample of 8 pastries from the batch is selected. The batch is accepted if the 2nd sample does not contain any pastry that does not meet the passing standard.

(a) Calculate the probability that the batch of pastries is accepted as a result of inspection of the 1st sample.

(b) Find the probability that a 2nd sample is selected for inspection and is rejected.

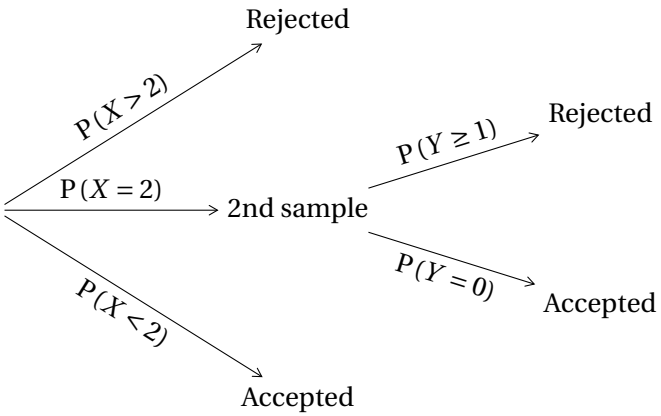
(c) Calculate the probability that the batch will be rejected.

(d) The junior chef and the senior chef work independently. 2.5% of the pastries made by the senior chef do not meet the passing standard.

(i) 8 pastries made by the junior chef and 6 pastries made by the senior chef are randomly selected. Calculate the probability that exactly 2 pastries do not meet the passing standard.

(ii) The number of pastries made by the junior chef and senior chef make up 40% and 60% of the pastries in the resturaunt respectively. If a customer purchases a box of 20 randomly chosen pastries, calculate the probability that there are more than 3 pastries in the box that do not meet the passing standard.

Solution



(a) $X \sim$ no. of pastries by junior chef that do not meet the passing standard, out of 12

$X \sim \mathrm{B}(12, 0.04)$

$$\begin{aligned}\mathrm{P}(\text{accepted from the inspection of the 1st sample}) &= \mathrm{P}(X < 2)\\ &= \mathrm{P}(X \leq 1)\\ &= 0.919 \text{ (3 s.f.)}\end{aligned}$$

(b) $Y \sim$ no. of pastries by junior chef that do not meet the passing standard, out of 8

$Y \sim \mathrm{B}(8, 0.04)$

$$\begin{aligned}\mathrm{P}(\text{2nd sample is selected for inspection and rejected}) &= \mathrm{P}(X = 2) \times \mathrm{P}(Y \geq 1) \\ &= \mathrm{P}(X = 2) \times [1 - \mathrm{P}(Y = 0)] \\ &= 0.0196 \text{ (3 s.f.)}\end{aligned}$$

(c) $$\begin{aligned}\mathrm{P}(\text{batch of pastries is rejected}) &= \mathrm{P}(X > 2) + \mathrm{P}(X = 2) \times \mathrm{P}(Y \geq 1) \\ &= [1 - \mathrm{P}(X \leq 2)] + \mathrm{P}(X = 2) \times [1 - \mathrm{P}(Y = 0)] \\ &= 0.0303 \text{ (3 s.f.)}\end{aligned}$$

(d) (i) $W \sim$ no. of pastries by senior chef that do not meet passing standard, out of 6

$W \sim \mathrm{B}(6, 0.025)$

$\mathrm{P}(\text{exactly 2 pastries do not meet passing standard})$

$= \mathrm{P}(Y = 2) \times \mathrm{P}(W = 0) + \mathrm{P}(Y = 0) \times \mathrm{P}(W = 2) + \mathrm{P}(Y = 1) \times \mathrm{P}(W = 1)$

$= 0.0680$ (3 s.f.)

(ii) $$\begin{aligned}\mathrm{P}(\text{a pastry does not meet the passing standard}) &= 0.4\,(0.04) + 0.6\,(0.025) \\ &= 0.31\end{aligned}$$

$A \sim$ no. of pastries that do not meet the passing standard, out of 20

$A \sim \mathrm{B}(20, 0.031)$

$$\begin{aligned}\mathrm{P}(\text{more than 3 do not meet the passing standard}) &= \mathrm{P}(A > 3) \\ &= 1 - \mathrm{P}(A \leq 3) \\ &= 0.00300 \text{ (3 s.f.)}\end{aligned}$$

CHAPTER

5

Normal Distribution

1 The Normal Distribution

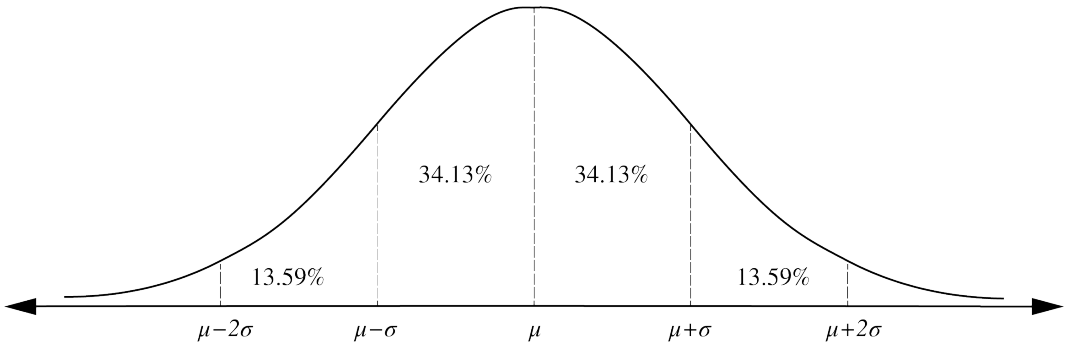
The normal distribution is the **most important** distribution for a continuous random variable. Many naturally occurring phenomena have a distribution that is normal, or approximately normal. Some examples are:

- the height of an adult male,
- the systolic blood pressure in healthy adults,
- the IQ scores of a large population.

While a discrete random variable is defined by its probability distribution, a continuous random variable is defined by its **probability density function**.

If X is normally distributed then its probability density function is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad \text{for } -\infty < x < \infty$$



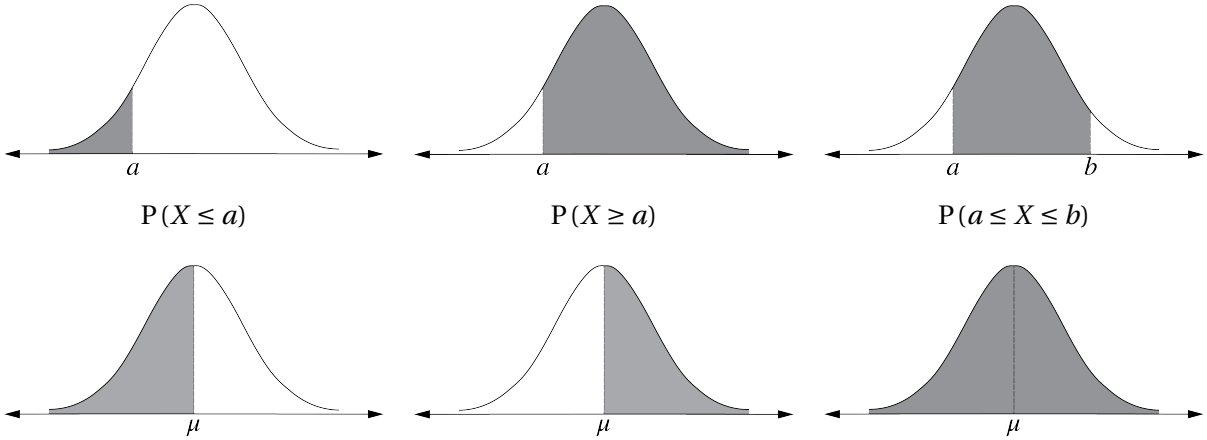
If X has a normal distribution with mean μ and variance σ^2 , we write

$$X \sim \mathrm{N}(\mu, \sigma^2)$$

For a normal curve, the standard deviation is uniquely determined as the horizontal distance from the line of symmetry $x = \mu$ to a point of inflection.

1.1 Properties of The Normal Distribution

1. The curve is bell shaped and **symmetrical about the vertical line** $x = \mu$.
2. The probability of any range of values is given by the **area under the pdf** within that interval.
3. The **total area** under the pdf curve is **1**.



$\mathrm{P}(X \le a)$ $\quad$ $\mathrm{P}(X \ge a)$ $\quad$ $\mathrm{P}(a \le X \le b)$

$\mathrm{P}(X \le \mu) = \frac{1}{2}$ $\quad$ $\mathrm{P}(X \ge \mu) = \frac{1}{2}$ $\quad$ Total Area = 1

Note: $\mathrm{P}(X = a) = 0$ since the corresponding area under the curve is zero. Hence for a normal random variable, we have

$$\mathrm{P}(X \le a) = \mathrm{P}(X < a) + \mathrm{P}(X = a) = \mathrm{P}(X < a) \text{ and similarly,}$$
$$\mathrm{P}(X \ge a) = \mathrm{P}(X > a) + \mathrm{P}(X = a) = \mathrm{P}(X \ge a)$$

1.2 Finding Probabilities Using Graphic Calculator

For $X \sim \mathrm{N}(\mu, \sigma^2)$, using the graphic calculator: Press **2ND** **VARS**

- To find $\mathrm{P}(X \le b)$: Select 2:normalcdf and key the values in the format 10^{-99}, b, μ, σ.
- To find $\mathrm{P}(X \ge a)$: Select 2:normalcdf and key the values in the format a, 10^{99}, μ, σ.
- To find $\mathrm{P}(a \le X \le b)$: Select 2:normalcdf and key the values in the format a, b, μ, σ.
- To find a such that $\mathrm{P}(X \le a) = p$: Select 3:invNorm and key the values in the format p, μ, σ, tail.

Note: Remember to enter σ (standard deviation) and not σ^2 (variance) when using the GC.

Example 5.1 Finding Probabilities of Normal Distributions (Area Under The Curve)

The weight of a randomly chosen student from a population may be assumed to have a normal distribution with mean 65 kg and standard deviation 3 kg. A student was randomly chosen from this population. Find the proabability that

(a) his weight exceeds 70 kg,

(b) his weight is below 62 kg,

(c) his weight is between 60 kg and 75 kg,

(d) his weight is within one standard deviation from the mean weight.

Solution

Let X be the weight of a randomly chosen student
$X \sim \mathrm{N}(65, 9)$

(a) $\mathrm{P}(X > 70) = 0.478$ (3 s.f.)

(b) $\mathrm{P}(X < 62) = 0.159$ (3 s.f.)

(c) $\mathrm{P}(60 < X < 75) = 0.952$ (3 s.f.)

(d) $\mathrm{P}(|X - 65| < 3) = \mathrm{P}(62 < X < 68)$
$= 0.683$ (3 s.f.)

Example 5.2 Inverse Normal Values

Given $X \sim \mathrm{N}(10, 4)$, find, correct to 3 significant figures, the values of a, b and c such that

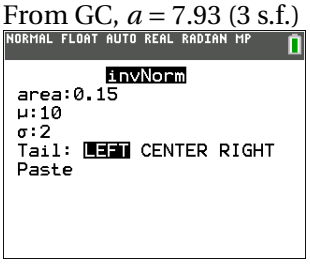
(a) $\mathrm{P}(X \leq a) = 0.15$

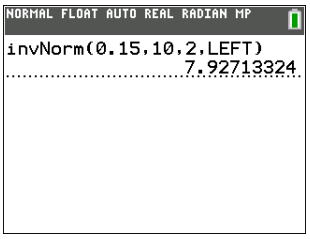
(b) $\mathrm{P}(X \geq b) = 0.23$

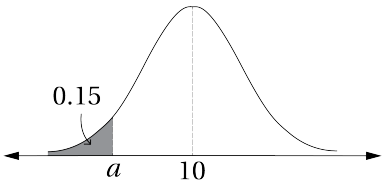
(c) $\mathrm{P}(9 < X < c) = 0.64$

Solution

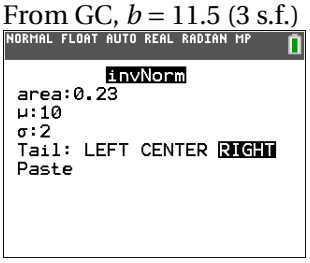
(a) From GC, $a = 7.93$ (3 s.f.)

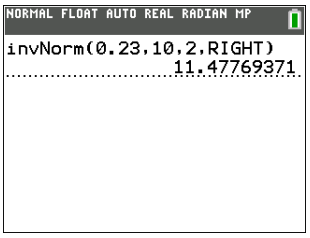


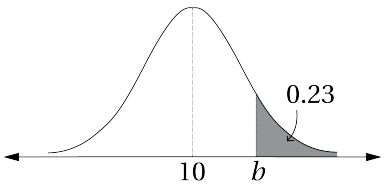




(b) From GC, $b = 11.5$ (3 s.f.)





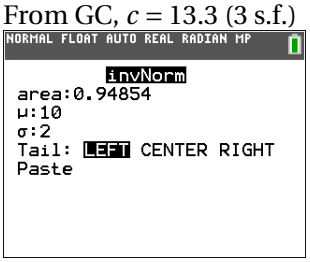


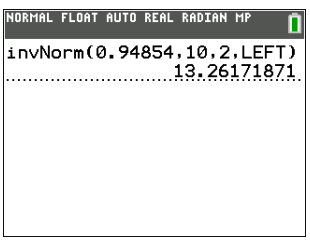
(c) $\mathrm{P}(X < c) - \mathrm{P}(X < 9) = 0.64$

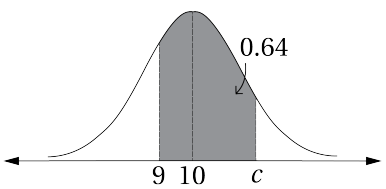
$\mathrm{P}(X < c) = 0.64 + \mathrm{P}(X < 9)$

$= 0.94854$ (5 s.f.)

From GC, $c = 13.3$ (3 s.f.)







2 The Standard Normal Distribution (Z-Distribution)

The standard normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$ is denoted by Z, i.e. $Z \sim \mathrm{N}(0,1)$.

Every normal distribution can be transformed into the standard normal distribution or $Z-$distribution using the transformation $Z = \dfrac{X-\mu}{\sigma}$.

This process of converting $X \sim \mathrm{N}\left(\mu, \sigma^2\right)$ into $Z \sim \mathrm{N}(0,1)$ is known as standardisation.

We usually perform standardisation when μ and/or σ are unknown.

Example 5.3 Standardisation to Find μ

Given that $X \sim \mathrm{N}\left(\mu, 0.04\right)$ and $\mathrm{P}(X < 5) = 0.3$, find μ.

Solution

$$\mathrm{P}(X < 5) = 0.3$$
$$\mathrm{P}\left(Z < \frac{5-\mu}{\sqrt{0.04}}\right) = 0.3$$

From GC,

$$\frac{5-\mu}{0.2} = -0.52440 \text{ (5 s.f.)}$$
$$\mu = 5.10 \text{ (3 s.f.)}$$

Example 5.4 Standardisation to Find σ

Given that $X \sim \mathrm{N}\left(12, \sigma^2\right)$ and $\mathrm{P}(X \geq 15) = 0.0668$, find the standard deviation of X.

Solution

$$\mathrm{P}(X \geq 15) = 0.0668$$
$$\mathrm{P}\left(Z < \frac{15-12}{\sigma}\right) = 0.0668$$

From GC,

$$\frac{3}{\sigma} = 1.5001 \text{ (5 s.f.)}$$
$$\sigma = 2.00 \text{ (3 s.f.)}$$

Example 5.5 Standardisation to Find μ and σ

In a particular town, 34 out of 100 people have heights exceeding 1.65 m and 4 out of 5 people have heights below 1.8 m. Assuming that their heights follow a normal distribution, find the mean and standard deviation of the distribution.

Solution

Let X be the height of a randomly chosen person from the town.
$X \sim \mathrm{N}(\mu, \sigma^2)$
Given that $\mathrm{P}(X > 1.65) = 0.34$,

$$\begin{aligned} \mathrm{P}(X > 1.65) &= 0.34 \\ \mathrm{P}\left(Z > \frac{1.65-\mu}{\sigma}\right) &= 0.34 \\ \mathrm{P}\left(Z \leq \frac{1.65-\mu}{\sigma}\right) &= 0.66 \\ \frac{1.65-\mu}{\sigma} &= 0.41246 \text{ (5 s.f.)} \end{aligned}$$

and given that $\mathrm{P}(X < 1.8) = 0.8$,

$$\begin{aligned} \mathrm{P}(X < 1.8) &= 0.8 \\ \mathrm{P}\left(Z < \frac{1.8-\mu}{\sigma}\right) &= 0.8 \\ \frac{1.8-\mu}{\sigma} &= 0.84162 \text{ (5 s.f.)} \end{aligned}$$

Rearranging the equations, we have the following simultaneous equations:

$$\begin{aligned} \mu + 0.41246\sigma &= 1.65 \cdots\cdots (1) \\ \mu + 0.84162 &= 1.8 \cdots\cdots (2) \end{aligned}$$

Using GC to solve (1) and (2), $\mu = 1.51$ and $\sigma = 0.350$.

3 Linear Combinations of Normal Random Variables

Recall that we have the following properties for the expectation and variance of two random variables.

For any random variables X and Y, and constants a and b,

(a) $\mathrm{E}(a) = a$

(b) $\mathrm{E}(aX) = a\mathrm{E}(X)$

(c) $\mathrm{E}(aX \pm b) = a\mathrm{E}(X) \pm b$

(d) $\mathrm{E}(aX \pm bY) = a\mathrm{E}(X) \pm b\mathrm{E}(Y)$

(a) $\mathrm{Var}(a) = 0$

(b) $\mathrm{Var}(aX) = a^2\mathrm{Var}(X)$

(c) $\mathrm{Var}(aX \pm b) = a^2\mathrm{Var}(X)$

(d) If X and Y are **independent**, then $\mathrm{Var}(aX \pm bY) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y)$

If X and Y are two **independent normal** random variables, and a and b are non zero constants, then

$$aX \pm bY \text{ are also normal random variables.}$$

Thus, if $X \sim \mathrm{N}(\mu_1, \sigma_1^2)$ and $Y \sim \mathrm{N}(\mu_2, \sigma_2^2)$ are independent random variables, and a and b are non zero constants, then

(a) $aX \sim \mathrm{N}(a\mu_1, a^2\sigma_1^2)$

(b) $aX \pm b \sim \mathrm{N}(a\mu_1 \pm b, a^2\sigma_1^2)$

(c) $aX \pm bY \sim \mathrm{N}(a\mu_1 \pm b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2)$

Example 5.6 Finding Distribution of Linear Combinations of Normal Random Variables

Given X and Y are independent with $X \sim \mathrm{N}(3,9)$ and $Y \sim \mathrm{N}(1,4)$
State the distribution of:

(a) $2X$

(b) $X+3$

(c) $X+2Y$

(d) $X-2Y$

(e) $X_1 + X_2 + X_3 + Y_1 + Y_2$, where X_1, X_2, X_3 are independent observations of X and Y_1, Y_2 independent observations of Y.

Solution

(a) $2X \sim \mathrm{N}(6,36)$

(b) $X+3 \sim \mathrm{N}(6,9)$

(c) $X+2Y \sim \mathrm{N}(5,25)$

(d) $X-2Y \sim \mathrm{N}(1,25)$

(e) $X_1 + X_2 + X_3 + Y_1 + Y_2 \sim \mathrm{N}(11,35)$

Example 5.7 Heights of Males vs Females

In a particular town, the heights, in centimetres, of the males and females are normally distributed as follows:

	Mean	Standard Deviation
Male	177	18.4
Female	162	13.2

(a) Find the probability that a randomly chosen female is taller than a randomly chosen male.

(b) Find the probability that the total height of three randomly chosen females exceeds that of 2 randomly chosen males by more than 100 cm.

Solution

Let X be the height of a randomly chosen male from the town.
$X \sim \mathrm{N}\left(177, 18.4^2\right)$
Let Y be the height of a randomly chosen female from the town.
$Y \sim \mathrm{N}\left(162, 13.2^2\right)$

(a) $Y - X \sim \mathrm{N}(-15, 512.8)$

$$\begin{aligned} \mathrm{P}(Y > X) &= \mathrm{P}(Y - X > 0) \\ &= 0.254 \text{ (3 s.f.)} \end{aligned}$$

(b) $(Y_1 + Y_2 + Y_3) - (X_1 + X_2) \sim \mathrm{N}(132, 1199.84)$

$$\mathrm{P}(Y_1 + Y_2 + Y_3 - X_1 + X_2 > 100) = 0.822 \text{ (3 s.f.)}$$

Example 5.8 Mass of Papayas

Papayas are sold by mass. The masses, in kg, of papayas follow a normal distribution with mean 0.7 and standard deviation 6.1.
Find the probability that the total mass of 2 randomly chosen papayas is heavier than twice the mass of a randomly chosen papaya by at most 80 g.

Solution

Let X be the mass of a randomly chosen papaya
$X \sim \mathrm{N}\left(0.7, 6.1^2\right)$
$(X_1 + X_2) - 2X \sim \mathrm{N}(0, 223.26)$

$$\mathrm{P}(X_1 + X_2 - 2X < 0.08) = 0.502 \text{ (3 s.f.)}$$

CHAPTER

6

Sampling Distribution (Central Limit Theorem)

1 Population and Sample

When we need to gather information about a **population**, very rarely in practice can we afford the luxury of examining the complete population. The two obvious reasons would be that the cost is too high and the population is dynamic in that the individuals making up the population may change over time.
Commonly, **sample** observations and analyses are used to make inferences or predictions about a population. That is, we take the results of an analysis using a sample and generalize it to the larger population that the sample represents. This is known as *statistical inference*.

> A **population** is the whole set of items that we want to study.
> A **sample** is a subset of the population.

In order for the statistical inference to be as accurate as possible, however, it is imperative that the sample is representative of the population (i.e. accurately reflect the characteristics of the population) to which it is being generalized. As such, the sample must be carefully gathered through **random sampling**.

2 Random Sampling

> In **random sampling**, every element in the population must have an equal chance of selection, i.e. it is free from bias.

In non-random sampling, each element in the population does not have an equal chance of being selected, resulting in certain segments of the population being over represented, as some members are systematically or deliberately excluded from study.

Example 6.1 Random Sampling Methods

In a small school there are 3 classes, each with 30 students, and 2 classes, each with 20 students. To find the students perception of the food sold in a canteen, a sample of 10 students is taken by the following three methods.

Method 1: 2 students are randomly chosen from each of the 5 classes.

Method 2: Assign every student in the school with a number from 1 to 130 by arranging their names in alphabetical order. Use a computer to generate 10 random numbers. The 10 students with the corresponding numbers would be the ones chosen to be in the sample.

Method 3: The first 10 students who patrionize the chicken rice stall are selected as the sample.

Solution

In choosing a random sample, each member of the population must have an equal chance of being selected.

Method 1
Let A be the event that a particular student from the class of 20 is chosen.

$$\mathrm{P}(A) = \frac{1}{10}$$

Let B be the event that a particular student from the class of 30 is chosen.

$$\mathrm{P}(B) = \frac{1}{15}$$

Since the probability of choosing any one student is not equal, this method does not involve random sampling.
Method 2
Since each student has a probability of $\frac{1}{13}$ of being chosen, this method does involve random sampling.
Method 3
This method does not involve random sampling because students who do not eat chicken rice will not patrionize the chicken rice stall and thus will not be selected.

3 Population Parameters and Sample Statistics

A **population parameter** is a number that is characteristic of the population under study.

A population parameter is a number that is characteristic of the population under study.
In many cases, we are concerned with two population parameters,

- **population mean**
- **population variance**

These are constants of a population and they are often unknown. The study of a population often involves finding estimates of these parameters.

A **sample statistic** is a number that is characteristic of a sample drawn from the population.

In general, a sample statistic is a **random variable** which contains information about the sample while any corresponding population parameter is a (possibly unknown) constant.
We use the following notation for parameters and statistics.

Quantity	Population	Sample	
		As a random variable	As a possible value
Mean	μ	$\overline{X}$	$\overline{x}$
Variance	σ^2	σ_X^2	σ_x^2

4 The Sample Mean, $\overline{X}$ as a Random Variable

Let $X_1, X_2, X_3, ..., X_n$ be n independent observations taken from a population X, with mean μ and variance σ^2.
Then the sample mean $\overline{X}$, defined by

$$\overline{X} = \frac{X_1 + X_2 + X_3 + \ldots + X_n}{n}$$

is a random variable with $\mathrm{E}\left(\overline{X}\right) = \mu$ and $\mathrm{Var}\left(\overline{X}\right) = \frac{\sigma^2}{n}$.

Proof:

$$\begin{aligned}\mathrm{E}\left(\overline{X}\right) &= \mathrm{E}\left(\frac{X_1 + X_2 + X_3 + \ldots + X_n}{n}\right)\\ &= \frac{1}{n}\mathrm{E}\left(X_1 + X_2 + X_3 + \ldots + X_n\right)\\ &= \frac{1}{n}\left(n\mathrm{E}\left(X\right)\right)\\ &= \mu\end{aligned}$$

$$\begin{aligned}\mathrm{Var}\left(\overline{X}\right) &= \mathrm{Var}\left(\frac{X_1 + X_2 + X_3 + \ldots + X_n}{n}\right)\\ &= \frac{1}{n^2}\mathrm{Var}\left(X_1 + X_2 + X_3 + \ldots + X_n\right)\\ &= \frac{1}{n^2}\left(n\mathrm{Var}\left(X\right)\right)\\ &= \frac{\sigma^2}{n}\end{aligned}$$

5 Distribution of $\overline{X}$

5.1 When X Follows a Normal Distribution

If $X_1, X_2, X_3, ..., X_n$ are n independent observations from a **normal distribution** with mean μ and variance σ^2, then the sample mean $\overline{X}$ is also normally distributed such that

$$\overline{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right) \textbf{ exactly}$$

Example 6.2 Sample Mean of a Normal Distribution

A random sample of size 15 is taken from a population which is normally distributed with mean 60 and standard deviation 4. Find the probability that the mean of the sample is less than 58.

Solution

Let $X \sim \mathrm{N}\left(60, 4^2\right)$

$\overline{X} \sim \mathrm{N}\left(60, \frac{4^2}{15}\right)$ exactly

Using GC,

$$\mathrm{P}\left(\overline{X} < 58\right) = 0.0264$$

Example 6.3 Finding n By Standardization

A large number of random samples of size n are taken from a normal population with mean 74 and variance 36 and the sample means are calculated. If at most 85% of the sample means exceed 72, find the largest possible value of n.

Solution

Let $X \sim \mathrm{N}(74, 36)$

$\overline{X} \sim \mathrm{N}\left(74, \frac{36}{n}\right)$ exactly

Given $\mathrm{P}\left(\overline{X} > 72\right) \leq 0.85$,

$$\mathrm{P}\left(Z > \frac{72-74}{\sqrt{\frac{36}{n}}}\right) \leq 0.85$$

$$\mathrm{P}\left(Z > -\frac{\sqrt{n}}{3}\right) \leq 0.85$$

$$-\frac{\sqrt{n}}{3} \leq -1.0364 \text{(5 s.f.)}$$

$$n \leq 9.67 \text{(3 s.f.)}$$

Thus the largest possible value of n is 9.

5.2 When X Follows Any Distribution (Central Limit Theorem)

If $X_1, X_2, X_3, ..., X_n$ are n independent observations from **ANY distribution**, X, with mean μ and variance σ^2, then when n is sufficiently large ($n \geq 30$), by the **Central Limit Theorem**,

$$\overline{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right) \textbf{ approximately}$$

Example 6.4 Sample Mean of a Binomial Distribution Using CLT

Ten fair die are thrown and the number of sixes is recorded. If the experiment is repeated 50 times and the number of sixes are recorded each time, estimate the probability that the mean number of sixes in each experiment will be less than 1.6.

Solution

Let X be the no. of sixes out of 10 fair die.

$X \sim \mathrm{B}\left(10, \frac{1}{6}\right)$

$$\mathrm{E}(X) = np = \frac{5}{3} \qquad \mathrm{Var}(X) = np(1-p) = \frac{25}{18}$$

Since the sample size $n = 50$ is sufficiently large (≥ 30), by the Central Limit Theorem,

$$\overline{X} \sim \mathrm{N}\left(\frac{5}{3}, \frac{1}{36}\right) \text{ approximately}$$

Thus,

$$\mathrm{P}\left(\overline{X} < 1.6\right) = 0.345 \text{(3 s.f.)}$$

Example 6.5 Sample Mean of a Probability Distribution Using CLT

A random variable X has the probability distribution shown in the table below. The sample mean for 50 independent observations of X is denoted by $\overline{X}$. Using a suitable approximation, find $\mathrm{P}\left(\overline{X} > 0\right)$.

x	-2	-1	0	1	2
$\mathrm{P}(X = x)$	0.3	0.1	0.15	0.4	0.05

Solution

$$\begin{aligned}\mathrm{E}(X) &= -2(0.3) - 1(0.1) + 0 + 1(0.4) + 2(0.05)\\ &= -0.2\end{aligned}$$

$$\begin{aligned}\mathrm{E}\left(X^2\right) &= 4(0.3) + 1(0.1) + 0 + 1(0.4) + 4(0.05)\\ &= 1.9\end{aligned}$$

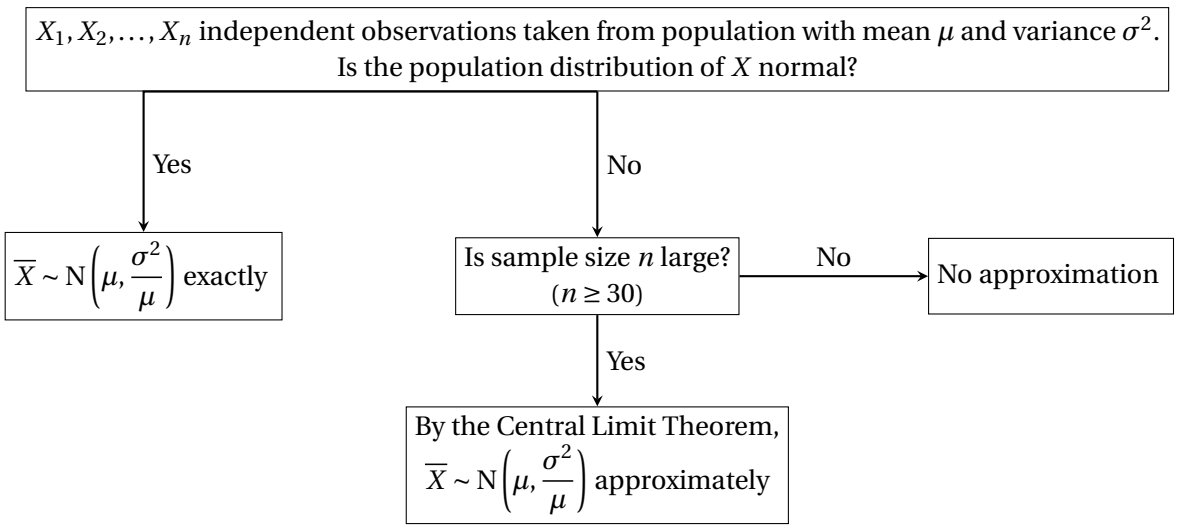
$$\begin{aligned}\mathrm{Var}(X) &= \mathrm{E}\left(X^2\right) - [\mathrm{E}(X)]^2\\ &= 1.86\end{aligned}$$

Since the sample size $n = 50$ is sufficiently large (≥ 30), by the Central Limit Theorem,

$$\overline{X} \sim \mathrm{N}\left(-0.2, \frac{1.86}{50}\right) \text{ approximately}$$

$$\mathrm{P}\left(\overline{X} > 0\right) = 0.150 (3 \text{ s.f.})$$

Central Limit Theorem Summary



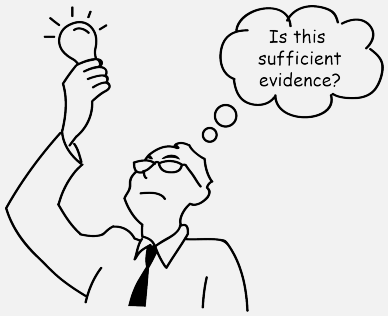
CHAPTER

7

Hypothesis Testing

Opening Problem

A manufacturer claims that the light bulbs he produces have a mean lifespan of 600 hours, and a standard deviation of 60 hours. In statistics, we call this a **statistical hypothesis**.
A retailer, having received numerous complaints from his customers, suspects that they do not last as long. He contacts the manufacturer and they decide to test a random sample of 50 bulbs. It turns out that the average lifespan of this sample, $\overline{x}$, is 580 hours.
Is this proof that the average lifespan of their light bulbs is below 600 hours? It is possible that the average lifespan of the population is actually 600 and that the sampled light bulbs just had a lower than average lifespan. Our sample cannot tell us with certainty the exact population mean μ. How do we then decide between the validity of manufacturer's claim and that of the angry customers?



Statisticians have devised a procedure to determine whether a statistical hypothesis is reasonable. We call it a **hypothesis test**.

A **statistical hypothesis** is an assertion or conjecture[1] concerning one or more populations.

To prove that a hypothesis is true, or false, with absolute certainty, we would need absolute knowledge. That is, we would have to examine the entire population. In most cases, it would not make sense to examine the entire population. For example if the light bulb manufacturer were to test the lifespan every light bulb he produces, he would have no light bulbs left to sell. Instead, hypothesis testing concerns on how to use a random sample to judge if it is evidence that supports or not the hypothesis.

1 Null and Alternative Hypotheses

Suppose a claim is made that a population mean μ has a value μ_0. We call this the **null hypothesis** H_0, and we write

$$H_0 : \mu = \mu_0$$

This statement is assumed to be true unless we have enough evidence to reject it.
The alternative hypothesis, denoted by H_1, is used to contradict the null hypothesis.

Given the null hypothesis $H_0 : \mu = \mu_0$, there are three ways to set up the alternative hypothesis:

- $H_1 : \mu > \mu_0$
- $H_1 : \mu < \mu_0$

} **One-tail Test**

- $H_1 : \mu \neq \mu_0$ **Two-tail Test**

[1] An opinion or conclusion formed on the basis of incomplete information.

Consider the **Opening Problem**. The manufacturer claims (null hypothesis) that his light bulbs have a lifespan of 600 hours. We write, $H_0 : \mu = 600$.
The retailer's suspicion that this claim is over-estimated (alternative hypothesis) is written formally as $H_1 : \mu < 600$.

Example 7.1 Null and Alternative Hypothesis

For each of the following scenarios, write down the null and alternative hypotheses. State whether a one-tail or two tail test applies.

(a) The top speed of submarines currently being produced by a manufacturer is currently 26.3 knots. When their engineers modify the design to reduce drag, they believe the maximum speed will be increased.

(b) The average peak-hour travel time along a particular stretch of road is currently 27 minutes. To help reduce travel times, electronic signs displaying real-time information are erected. If the travel times improve, the signs will be widely implemented.

(c) Whitex produces copy paper, and the weight of the copy paper is given as 80 g per m^2. The company wants to determine whether this information is correct.

Solution

(a) $H_0 : \mu = 26.3$, $H_1 : \mu > 26.3$
One-tail test

(b) $H_0 : \mu = 27$, $H_1 : \mu < 27$
One-tail test

(c) $H_0 : \mu = 80$, $H_1 : \mu \neq 80$
Two-tail test

2 Hypothesis Testing

2.1 The Test Statistic

To carry out the test, our focus moves from X (e.g. the lifespan of light bulbs) to the distribution of $\overline{X}$ (e.g. the mean lifespan from a sample of lightbulbs). $\overline{X}$ is called the **test statistic** for the population mean μ. The decision to reject or not to reject H_0 depends on how far the observed sample mean, $\overline{x}$, is from the claimed mean.

A **test statistic** is a random variable that is calculated from sample data (e.g. sample mean).

The distribution of the test statistic under the assumptions of H_0 is called the **null distirbution**.

We have seen that for a population which is normally distributed with mean μ and standard deviation σ, the sample mean

$$\overline{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right) \text{ exactly}$$

If the population is not normally distributed but the sample size is sufficiently large, we can apply the Central Limit Theorem

$$\overline{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right) \text{ approximately}$$

Under the assumptions of H_0, $\mu = \mu_0$, so our null distribution is $\overline{X} \sim \mathrm{N}\left(\mu_0, \frac{\sigma^2}{n}\right)$.

In our light bulb example, $H_0 : \mu = 600$. Under the assumption that the null hypothesis is true, the population mean and standard deviation are $\mu = 600$ and $\sigma = 60$ (these are the values given by the manufacturer). Our sample size n is 50. So, by the Central Limit Theorem, we have the null distribution

$$\overline{X} \sim \mathrm{N}\left(600, \frac{60^2}{50}\right) \text{ approximately, since } n = 50 \text{ is sufficiently large}$$

However in hypothesis testing, we will not be using the test statistic $\overline{X}$. Instead, the standardised test statistic is used,

$$Z = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}, \text{ where } Z \sim \mathrm{N}(0,1)$$

Consider a statistical hypothesis test of $H_0 : \mu = \mu_0$ for a normally distributed population with known standard deviation σ. Given a sample of size n with observed sample mean $\overline{x}$:

The standardised test statistic is $Z = \dfrac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$, where $Z \sim \mathrm{N}(0,1)$

which has observed value $z = \dfrac{\overline{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$.

Back to the light bulb example:
The population standard deviation is 60 hours.
When they take a random sample of 50 light bulbs, they find that the mean lifespan is $\bar{x} = 580$ hours.
The observed value of the test statistic is $z = \dfrac{580 - 600}{\frac{60}{\sqrt{50}}} \approx -2.36$.

2.2 Level of Significance

In hypothesis testing, the null hypothesis may be rejected (or not rejected) not with certainty but with confidence that the likelihood of error in making the decision is small. We control the chance of wrongly rejecting the null hypothesis, that is, reject null hypothesis when it is actually true. This is known as the significance level of the test.

The level of significance of a hypothesis test, denoted by α, is defined as the probability of rejecting H_0 when H_0 is in fact true.

For example, if the level of significance is set at 5%, we are saying that there is a 5% chance (or probability of 0.05) that one chooses to reject H_0 when H_0 is actually correct.
Appropriate values for α depends on which area of study we are engaged in. For social sciences, it might be as high as $\alpha = 0.3$, the biological and medical fields mostly use $\alpha = 0.05$ or smaller.
The lower the level of significance, the stronger the evidence needed to reject H_0.

2.3 Critical Region and Critical Values

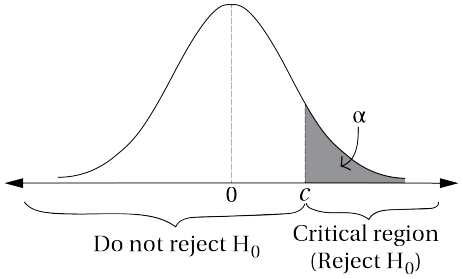
A decision needs to be made about the cut-off point which indicates the boundary of the region where values of z would be considered too far away from the claimed mean and therefore too unlikely to occur. The **critical region** is the set of values of the test statistic which result in H_0 being rejected. We determine the critical region from the level of significance.

- If the test statistic lies within the **critical region** then we will reject H_0.
- The boundary values of the critical region are called **critical values**.

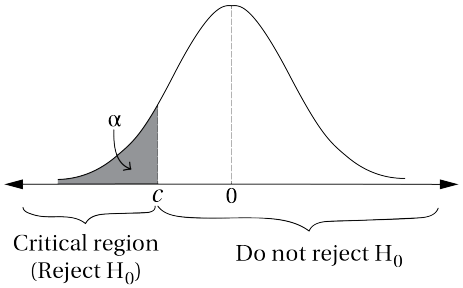
In order to decide where the critical region is, we need to know whether the hypothesis test is a one-tail or two-tail test.
The shaded areas are indicated by the level of significance.

One-Tail Test

In a one-tail test, the alternative hypothesis H_1 looks for an **increase** or **decrease** in the value of the claimed mean.
For an increase, $H_1 : \mu > \mu_0$, the critical region is the **upper tail** of the standard normal distribution curve.

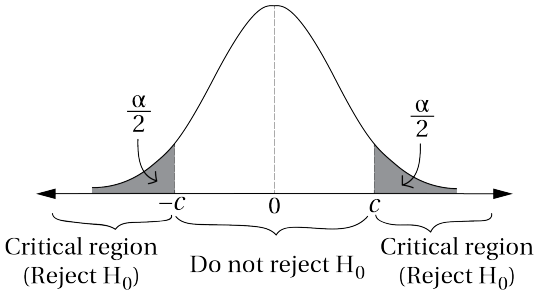


For a decrease, $H_1 : \mu < \mu_0$, the critical region is the **lower tail** of the standard normal distribution curve.



Two-Tail Test

In a two-tail test, the alternative hypothesis H_1 looks for a change in the claimed value without specifying whether it is an increase or decrease.
If we have a two-tailed alternative hypothesis, $H_1 : \mu \neq \mu_0$, then there are two critical values. However because of symmetry, we only need to perform one calculation.



Lets say that the light bulb manufacturer chose a 5% (or 0.05) level of significance before he started this experiment.
So far he has determined the following:

1. Null and alternative hypotheses:

$$H_0 : \mu = 600 \quad \text{and} \quad H_1 : \mu < 600$$

2. The distribution of the sample mean is

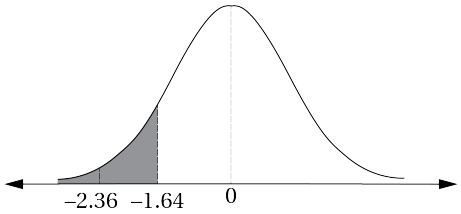
$$\overline{X} \sim \text{N}\left(600, \frac{60^2}{50}\right) \text{ approximately}$$

3. The observed test statistic is

$$z = \frac{580 - 600}{\frac{60}{\sqrt{50}}} = -2.36 \text{ (3 s.f.)}$$

Since the level of significance is 5%, the area under the curve in our critical region is 0.05. We can find the critical value by using the **invNorm** function.

$$\text{Critical value} = -1.64 \text{ (3 s.f.)}$$
$$\text{Critical region} : z < -1.64$$



Since our test statistic $z = -2.36$ falls within the critical region, **we reject the null hypothesis** H_0. There is **sufficient evidence**, at the 5% level, to conclude that $\mu < 600$.

2.4 The $p-$value

Our light bulb macufacturer wants to calculate the probability of obtaining a sample with mean as low as 580 by chance under the assumption of the null hypothesis H_0. We call this probability the ***p*−value**.

> The ***p*−value** of a test statistic is the probability of that result being observed if H_0 is true.

Instead of comparing the observed or calculated value of the test statistic with the critical values to determine whether or not to reject H_0, we can also consider the $p-$value and compare it with the level of significance.

- We will reject H_0 if the $p-$value is less than the level of significance.
- We will not reject H_0 if the $p-$value is greater than the level of significance.

We can find the $p-$value by using our GC.

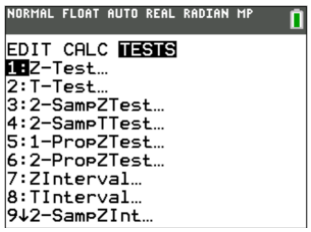
Step 1: Press **stat** and arrow right to "TESTS".

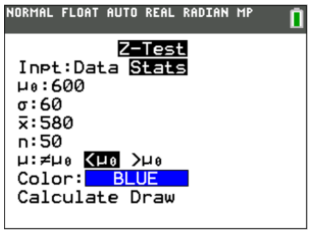
Step 2: Select "1:$Z-$Test" and select "Inpt:Stats". (We choose Stats because we are given the sample mean rather than the raw data)

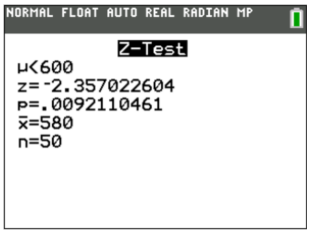
Step 3: Key in the values as required.

Step 4: With the cursor on "Calculate" press **enter** to obtain the $p-$value.

Lets try to find the $p-$value for the light bulb manufacturer's test statistic (i.e. what is the probability we observe $\overline{x} = 580$ if H_0 is true).



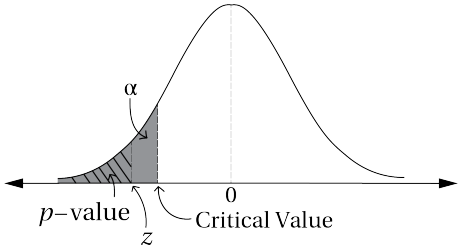




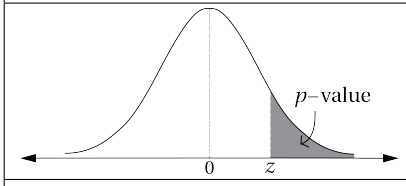
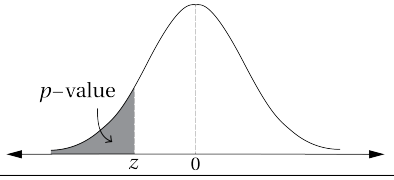
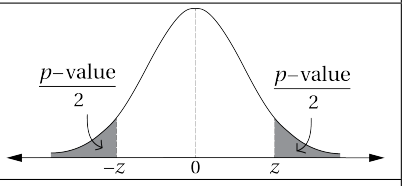
From GC, $p-$value = 0.00921 (which is less than our level of significance 0.05).
What this means is that if our null hypothesis is true, then there is only a 0.00921 probability of getting a sample mean of 580 or less.

> Since the $p-$value is less than the level of significance, we reject H_0. There is **sufficient evidence**, at the 5% level, to conclude that $\mu < 600$.

Note: The level of significance is the area bounded by the critical value while the $p-$value is the area bounded by the test statistic.



The $p-$value is calculated as shown in the diagram

$H_1 : \mu > \mu_0$	$H_1 : \mu < \mu_0$	$H_1 : \mu \neq \mu_0$
		
$p-\text{value} = \text{P}(Z > z)$	$p-\text{value} = \text{P}(Z < z)$	$p-\text{value} = 2\text{P}(Z > z)$

Now, at this point you might be wondering: didn't we just conclude the exact same thing when we determined that z lies within our critical region? And you would be correct. In fact we only need to find and compare

- z and the critical region , **or**
- the $p-$value and the level of significance.

However, I recommend doing both as it is not too much extra work and it gives us extra confidence in our answer. When carrying out a hypothesis test, we will use this general procedure:

Step 1: State the **null hypothesis** $H_0 : \mu = \mu_0$ and **alternative hypothesis** H_1.

Step 2: State the **distribution** of X (if known) and $\overline{X}$.

Step 3: State the **level of significance**, α (usually given in the question) and determine the **critical region** depending on the nature of the alternative hypothesis.

Step 4: Using data from a sample, calculate the **observed standardised test statistic**:

$$z = \frac{\overline{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

Step 5: Using your GC, calculate the ***p*−value** for the test statistic.

Step 6: **Make your conclusion** about the hypotheses.

Template for writing your conclusion:

> Since the $p-$value is (less/more) than the level of significance, we (reject/do not reject) H_0. There is (sufficient/insufficient) evidence, at the (level of significance) level, to conclude that (H_1 is true, with context if applicable).

3 The $Z-$Test

The $Z-$**test** is used to test hypotheses when:

- test statistic follows a **normal distribution**, and
- the **population variance σ^2 is known**.

Because of the Central Limit Theorem, many test statistics are approximately normally distributed for large samples. Therefore, many statistical tests can be conveniently performed as $Z-$tests if the sample size is large and the population variance is known.

Example 7.2 Hypothesis Test Given a Normal Distribution

The lengths of metal bars produced by a particular machine are normally distributed with mean 420 cm and standard deviation 12 cm. The machine is serviced, after which a sample of 100 metal bars is taken and the length of each is measured. The result shows that the sample mean is 422 cm. Is there evidence, at the 3% level of significance, that there is a change in the mean length of the metal bars produced by this machine?

Solution

Step 1: $H_0 : \mu = 420$

$H_1 : \mu \neq 420$

Step 2: $X \sim \mathrm{N}\left(420, 12^2\right)$

$\overline{X} \sim \mathrm{N}\left(420, \dfrac{12^2}{100}\right)$

Step 3: Level of significance: 3%

Critical Region: $z < -2.17$ or $z > 2.17$

Step 4: $z = \dfrac{422 - 420}{\frac{12}{\sqrt{100}}} = 1.67 < 2.17$

Step 5: Using GC, $p-\text{value} = 0.0956 > 0.03$

Step 6: Since the $p-$value is **more** than the level of significance, we **do not reject** H_0. There is **insufficient** evidence, at the 3% level, to conclude that there is a change in the mean length of metal bars produced.

Example 7.3 Hypothesis Test of Any Distribution With Large Sample Size (CLT)

Bags of salted cashew nuts state that their net contents is 100 g. The manufacturer knows that the standard deviation of the population is 1.6 g. A customer claims that the bag have been lighter in recent purchases, so the factory quality control manager decides to investigate. He samples 40 bags and finds that their mean weight is 99.4 g. Perform a hypothesis test at the 5% level of significance to determine whether the customers claim is valid.

Solution

Step 1: $H_0 : \mu = 100$

$H_1 : \mu < 100$

Step 2: Since $n = 40$ is large, by the Central Limit Theorem,

$$\overline{X} \sim \mathrm{N}\left(100, \frac{1.6^2}{40}\right) \text{ approximately}$$

Step 3: Level of significance: 5%

Critical Region: $z < -1.64$

Step 4: $z = \dfrac{99.4 - 100}{\frac{1.6}{\sqrt{40}}} = -2.37$ (3 s.f.) < -1.64

Step 5: Using GC, $p-$value $= 0.00885 < 0.05$

Step 6: Since the $p-$value is **less** than the level of significance, we **reject** H_0. There is **sufficient** evidence, at the 5% level, to conclude that the customers claim is valid.

Example 7.4 Hypothesis Test With Different Levels of Significance

The random variable X is thought to have a mean of 50 but it is known that the standard deviation is 14.5. A random sample of 100 gives a mean of 52.6.
Is there evidence that the population mean has increased

(a) at the 5% level of significance?

(b) at the 1% level of significance?

Find the least value of the sample mean such that there is sufficient evidence at the 1% level of significance that the population mean has increased, giving your answer correct to 1 decimal place.
State, giving a reason, if any assumption needs to be made about the distribution of X.

Solution

$H_0 : \mu = 50$, $H_1 : \mu > 50$
Since $n = 100$ is large, by the Central Limit Theorem,

$$\overline{X} \sim \mathrm{N}\left(50, \frac{14.5^2}{100}\right) \text{ approximately}$$

$$z = \frac{52.6 - 50}{\frac{14.5}{\sqrt{100}}} = 1.79$$

(a) At the 5% level of significance,

Critical Region: $z > 1.64$

$z = 1.79 > 1.64$

Using GC, $p-\text{value} = 0.0365 < 0.05$

> Since the $p-$value is **less** than the level of significance, we **reject** H_0. There is **sufficient** evidence, at the 5% level, to conclude that the population mean has increased.

(b) At the 1% level of significance,

Critical Region: $z > 2.33$

$z = 1.79 < 2.33$

$p-\text{value} = 0.0365 > 0.01$

> Since the $p-$value is **more** than the level of significance, we **do not reject** H_0. There is **insufficient** evidence, at the 1% level, to conclude that the population mean has increased.

To reject H_0 at the 1% level of significance, the test statistic should fall inside the critical region: $z > 2.33$.

$$\begin{aligned} z &> 2.33 \\ \frac{\overline{x} - 50}{\frac{14.5}{\sqrt{100}}} &> 2.33 \\ \overline{x} - 50 &> 3.3785 \\ \overline{x} &> 53.4 \text{ (1 d.p.)} \end{aligned}$$

No assumption about the distribution of X is needed as $n = 100$ is large and hence by the Central Limit Theorem, the sample mean will be approximated by a normal distribution.

4 Unbiased Estimates of Population Parameters

In the previous chapter, we studied the distribution of the sample mean, assuming complete knowledge of the population parameters. However in most situations, we **do not know** the population parameters. In this section, we will look at the ways in which population parameters (i.e. mean and variance) can be **estimated** based on information from random samples.

4.1 Unbiased Estimate for Population Mean μ

There are several ways to obtain an estimate for a population parameter. In general, we obtain a sample and compute a value based on the sample. Hopefully this value (the estimate) is close to the actual value of the parameter to be estimated. It can be shown that the sample mean is the preferred unbiased estimator for the population mean.

Let $x_1, x_2, x_3, \ldots, x_n$ be observed values from a random sample of size n taken from a population with **unknown population mean** μ. Then

$$\text{The sample mean, } \overline{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{\sum x}{n}$$

is an **unbiased estimate** of μ.

4.2 Unbiased Estimate for Population Variance s^2

Let $x_1, x_2, x_3, \ldots, x_n$ be observed values from a random sample of size n taken from a population with unknown population variance σ^2. Then

$$s^2 = \frac{1}{n-1}\sum (x-\overline{x})^2 = \frac{1}{n-1}\left[\sum x^2 - \frac{(\sum x)^2}{n}\right] \text{ (in MF26)}$$

is an **unbiased estimate** of σ^2.

Note: $s^2 = \dfrac{n}{n-1} \times$ (sample variance), so sample variance is **not** an unbiased estimator of the population variance σ^2.

Example 7.5 Finding Unbiased Estimates of Population Parameters

(a) A random sample of size 50 is taken from a population with mean μ and variance σ^2. The sample data are summarized by

$$\sum x = 134 \qquad \sum x^2 = 1032$$

Calculate the unbiased estimates of μ and σ^2.

(b) The data from a random sample of size 50 are summarized by

$$\sum (x-40) = -27 \qquad \sum (x-40)^2 = 167$$

Find unbiased estimates of the population mean and population variance.

Solution

(a)

$$\text{Unbiased estimate of } \mu \text{ is } \bar{x} = \frac{\sum x}{n} = \frac{134}{50} = 2.68 \text{ (3 s.f.)}$$

$$\text{Unbiased estimate of } \sigma^2 \text{ is } s^2 = \frac{1}{n-1}\left[\sum x^2 - \frac{(\sum x)^2}{n}\right] = \frac{1}{49}\left[1032 - \frac{134^2}{50}\right] = 13.7 \text{ (3 s.f.)}$$

(b)

$$\mathrm{E}(X-40) = \mathrm{E}(X) - 40$$
$$\mathrm{E}(X) = \mathrm{E}(X-40) + 40$$

$$\text{Unbiased estimate of } \mu \text{ is } \bar{x} = \frac{\sum (x-40)}{n} + 40 = \frac{-27}{50} + 40 = 39.5 \text{ (3 s.f.)}$$

$$\mathrm{Var}(X-40) = \mathrm{Var}(X)$$

$$\text{Unbiased estimate of } \sigma^2 \text{ is } s^2 = \frac{1}{n-1}\left[\sum (x-40)^2 - \frac{(\sum (x-40))^2}{n}\right] = \frac{1}{49}\left[167 - \frac{(-27)^2}{50}\right] = 3.11 \text{ (3 s.f.)}$$

Example 7.6 Finding Unbiased Estimates of Population Parameters

The speeds of 120 randomly selected cars are measured as they pass a camera on a motorway. Denoting the speed by x km/h, the results are summarised by

$$\sum (x-100) = -221 \qquad \sum (x-100)^2 = 4708$$

Find unbiased estimates of the population mean and population variance, giving your answers correct to 2 decimal places.

Solution

$$\mathrm{E}(X-100) = \mathrm{E}(X) - 100$$
$$\mathrm{E}(X) = \mathrm{E}(X-100) + 100$$

$$\text{Unbiased estimate of } \mu \text{ is } \bar{x} = \frac{\sum (x-100)}{n} + 100 = \frac{-221}{120} + 100 = 98.16 \text{ (2 d.p.)}$$

$$\text{Var}(X-100) = \text{Var}(X)$$

$$\text{Unbiased estimate of } \sigma^2 \text{ is } s^2 = \frac{1}{n-1}\left[\sum (x-100)^2 - \frac{(\sum (x-100))^2}{n}\right]$$
$$= \frac{1}{119}\left[4708 - \frac{(-221)^2}{120}\right]$$
$$= 36.14 \text{ (2 d.p.)}$$

Individual Data

If we are given the individual data points instead of the total sum, we can use our GC to find our unbiased estimates $\overline{x}$ and s^2.

Step 1: Press **stat** and select "1.Edit".

Step 2: Enter the data points in the list L_1. (If there is already data in the list, move cursor to cover L_1 and press **clear** **enter**.

Step 3: Press **stat**, right arrow to "1−Var Stats" and press **enter**.

Step 4: Enter L_1 under "List" and with the cursor on "Calculate" press **enter**.

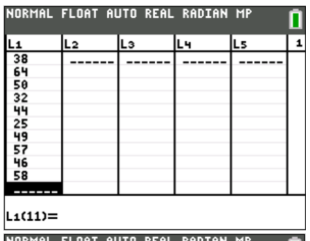
Example 7.7 Unbiased Estimates For Individual Data Using GC

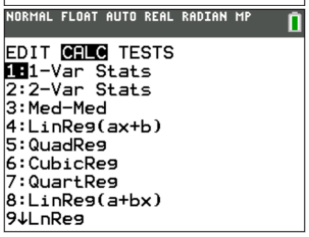
Changi Airport handles thousands of pieces of luggage per day. A random sample of 10 pieces of luggage is taken, and the masses (in kg) of the pieces are as follows:

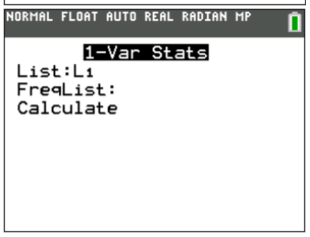
38 64 50 32 44 25 49 57 46 58

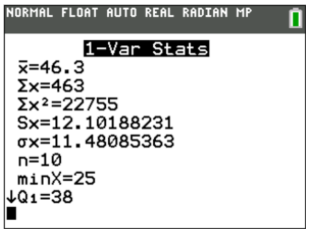
Calculate the unbiased estimates for the population mean and variance.

Solution









From GC,
$\overline{x} = 46.3$
$s^2 = (12.102)^2 = 146$ (3 s.f.)

Grouped Data

If we are given the data in a frequency table, we can also use our GC to find our unbiased estimates $\overline{x}$ and s^2.

Step 1: Press **stat** and select "1.Edit".

Step 2: Enter the data points in the list L_1. Enter the frequencies into L_2.

Step 3: Press **stat**, right arrow to "1−Var Stats" and press **enter**.

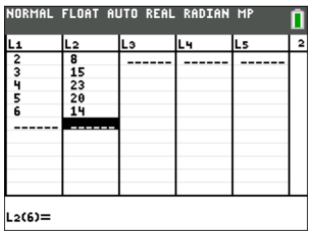
Step 4: Enter L_1 under "List" and L_2 under "FreqList". With the cursor on "Calculate" press **enter**.

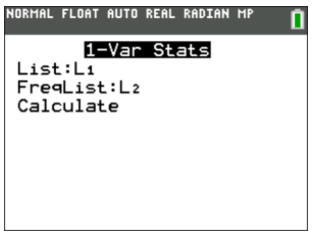
Example 7.8 Unbiased Estimates For Grouped Data Using GC

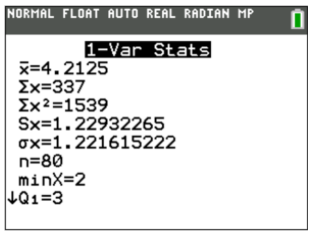
A sample of 80 customers at McDonald's were asked how many hamburgers each could eat for a meal and the results were tabulated. Calculate unbiased estimates for the population mean and variance.

Number of Hamburgers	2	3	4	5	6
Frequency	8	15	23	20	14

Solution







From GC,
$\overline{x} = 4.2125$
$s^2 = (1.2293)^2 = 1.51$ (3 s.f.)

4.3 Hypothesis Test With Unknown Variance and Large Sample Size n

Since σ^2 is unknown, an unbiased estimator, s^2, is used instead, where

$$s^2 = \frac{1}{n-1}\left(\sum x^2 - \frac{(\sum x)^2}{n}\right) = \frac{1}{n-1}\sum (x-\overline{x})^2$$

If X follows a normal distirbution, then

$$\overline{X} \sim \mathrm{N}\left(\mu_0, \frac{s^2}{n}\right) \text{ exactly}$$

If X does not follow a normal distirbution, but n is large, then by the Central Limit Theorem,

$$\overline{X} \sim \mathrm{N}\left(\mu_0, \frac{s^2}{n}\right) \text{ approximately}$$

Example 7.9 Hypothesis Test Using Unbiased Estimates of Population Parameters

A teacher sets an examination paper which she thinks a typical student should take 50 minutes to complete. She gave he paper to 60 randomly chosen students.
Let X be the time, in minutes, taken by a student to complete the examination paper.
The results are summarised by

$$\sum x = 3048 \qquad \sum (x-\overline{x})^2 = 465$$

Calculate unbiased estimates for the population mean and population variance.
Hence, test at at 2% significance level, whether the population mean time for a student to complete the examination differs from 50 minutes.

Solution

$$\begin{aligned}\text{Unbiased estimate of } \mu \text{ is } \overline{x} &= \frac{\sum x}{n} \\ &= \frac{3048}{60} \\ &= 50.8\end{aligned}$$

$$\begin{aligned}\text{Unbiased estimate of } \sigma^2 \text{ is } s^2 &= \frac{1}{n-1}\sum (x-\overline{x})^2 \\ &= \frac{465}{59}\end{aligned}$$

$H_0 : \mu = 50$, $H_1 : \mu \neq 50$
Since $n = 60$ is large, by the Central Limit Theorem,

$$\overline{X} \sim \text{N}\left(50, \frac{\frac{465}{59}}{60}\right) \text{ approximately}$$

Level of significance: 2%
Critical Region: $z < -2.33$ or $z > 2.33$
$z = \dfrac{50.8-50}{\sqrt{\frac{465}{3540}}} = 2.21$ (3 s.f.) < 2.33
Using GC, $p-$value $= 0.0273$ (3 s.f.) > 0.02

Since the $p-$value is **more** than the level of significance, we **do not reject** H_0. There is **insufficient** evidence, at the 2% level, to conclude that the mean time for a student to complete the examination differs from 50 minutes.

Example 7.10 Hypothesis Test Using Unbiased Estimates of Population Parameters

An electronic device is advertised as being able to retain information stored in it for 80 hours after power has been switched off. In experiments carried out to test this claim, the retention time in hours, X, was measured on 250 occasions, and the data obtained is summarized by

$$\sum (x-76) = 683 \qquad \sum (x-76)^2 = 26132$$

The population mean and variance of X are denoted by μ and σ^2 respectively.

(a) Show that, correct to one decimal place, an unbiased estimate of σ^2 is 97.5.

(b) Test the hypothesis that $\mu = 80$ against the alternative hypothesis that $\mu < 80$, at the 5% significance level.

Solution

(a)

$$\text{Var}(X-76) = \text{Var}(X)$$

$$\begin{aligned}\text{Unbiased estimate of } \sigma^2 \text{ is } s^2 &= \frac{1}{n-1}\left[\sum (x-76)^2 - \frac{(\sum (x-76))^2}{n}\right] \\ &= \frac{1}{249}\left[26132 - \frac{(683)^2}{250}\right] \\ &= 97.454 \text{ (5 s.f.)} \\ &= 97.5 \text{ (1 d.p.)}\end{aligned}$$

(b)

$$\begin{aligned}\text{E}(X-76) &= \text{E}(X) - 76 \\ \text{E}(X) &= \text{E}(X-76) + 76\end{aligned}$$

$$\begin{aligned}\text{Unbiased estimate of } \mu \text{ is } \overline{x} &= \frac{\sum (x-76)}{n} + 76 \\ &= \frac{683}{250} + 76 \\ &= 78.732\end{aligned}$$

$H_0 : \mu = 80$, $H_1 : \mu < 80$

Since $n = 250$ is large, by the Central Limit Theorem,

$$\overline{X} \sim \text{N}\left(78.732, \frac{97.454}{250}\right) \text{ approximately}$$

Level of significance: 5%

Critical Region: $z < -1.64$

$$z = \frac{78.732 - 80}{\sqrt{\frac{97.454}{250}}} = -2.03 \text{ (3 s.f.)} < -1.64$$

Using GC, $p-\text{value} = 0.0211$ (3 s.f.) < 0.05

Since the $p-$value is **less** than the level of significance, we **reject** H_0. There is **sufficient** evidence, at the 5% level, to conclude that the mean retention time is less than 80 hours.

In some contexts, the value of the sample variance may be given instead of $\sum x^2$. The sample variance, denoted by σ_x^2, is a **biased estimator** of the population variance.
If we are given σ_x^2 in the question, we will first need to compute s^2 using the relation

$$s^2 = \frac{n}{n-1}\sigma_x^2$$

Example 7.11 Using Sample Variance σ_x^2 to find s^2

The average starting salary of a university graduate is claimed to be \$2700. A random sample of 50 graduates has a mean starting salary of \$2640 with a standard deviation of \$145. Determine whether there is sufficient evidence that average starting salary differs from \$2700 at the 5% level of significance.

Solution

$$\begin{aligned} s^2 &= \frac{n}{n-1}\sigma_x^2 \\ &= \frac{50}{49}(145)^2 \\ &= 21454 \text{ (5 s.f.)} \end{aligned}$$

$\overline{x} = 2640$
$H_0 : \mu = 2700,\ H_1 : \mu \neq 2700$
$X \sim$ average starting salary of a university graduate
Since $n = 50$ is large, by the Central Limit Theorem,

$$\overline{X} \sim \text{N}\left(2700, \frac{21454}{50}\right) \text{ approximately}$$

Level of significance: 5%
Critical region: $z < -1.96$ or $z > 1.96$

$$z = \frac{2640 - 2700}{\sqrt{\frac{21454}{50}}} = -2.90 < -1.96$$

Using GC, $p-\text{value} = 0.00337$ (3 s.f.) < 0.05

Since the $p-$value is **less** than the level of significance, we **reject** H_0. There is **sufficient** evidence, at the 5% level, to conclude that the average starting salary of a university graduate differs from \$2700.

Case	Distribution Under H_0	Test Statistic
• **Population variance known** • **Normally distributed**	$\overline{X} \sim \mathrm{N}\left(\mu_0, \frac{\sigma^2}{n}\right)$	$Z = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$
• **Population variance known** • **Not normally distributed** • **Sample size is large**	$\overline{X} \sim \mathrm{N}\left(\mu_0, \frac{\sigma^2}{n}\right)$ approximately by CLT	$Z = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$
• **Population variance unknown** • **Normally distributed**	$\overline{X} \sim \mathrm{N}\left(\mu_0, \frac{s^2}{n}\right)$	$Z = \frac{\overline{X} - \mu_0}{\frac{s}{\sqrt{n}}}$
• **Population variance unknown** • **Not normally distributed** • **Sample size is large**	$\overline{X} \sim \mathrm{N}\left(\mu_0, \frac{s^2}{n}\right)$ approximately by CLT	$Z = \frac{\overline{X} - \mu_0}{\frac{s}{\sqrt{n}}}$
• **Population variance unknown** • **Not normally distributed** • **Sample size is small**	Not in syllabus	

CHAPTER

8

Correlation and Regression

Opening Problem

At a junior tournament, some young atheles each throw a discus. The age and distance thrown are recorded for each athlete.

Athlete	A	B	C	D	E	F	G	H	I	J	K	L
Age (years)	12	16	16	18	13	19	11	10	20	17	15	13
Distance thrown (m)	20	35	23	38	27	47	18	15	50	33	22	20

Things to think about:

(a) Do you think the distance an athlete can throw is related to the persons age?

(b) What happens to the distance thrown as the age of the athlete increases?

(c) How could you graph the data to more clearly see the relationship between the variables?

(d) How can we measure the relationship between the variables?

1 Bivariate Data

In the **Opening Problem**, each athelete has had two variables (age and distance thrown) recorded about them. This type of data is called **bivariate data**. We study it to understand the relationship between two variables.

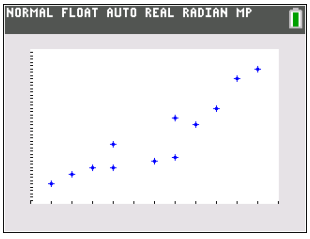
For example, we expect the distance thrown will depend on the athlete's age, so the age is the **indepdendent variable** and distance thrown is the **dependent variable**.

> **Bivariate data** is data that consists of the values of two variables obtained from the same sample expressed as ordered pairs.

2 Scatter Diagrams

We can observe the relationship between two variables using a **scatter diagram**. We usually place the independent vairable on the horizontal axis, and the dependent variable on the vertical axis.

In the **Opening Problem**, the indepdent variable *age* is placed on the horizontal axis, and the dependent variable *distance thrown* is place on the vertical axis.



We can the graph each data value as a point on the scatter diagram.

From the general shape formed from the dots, we can see that as the *age* increases, so does the *distance thrown*.

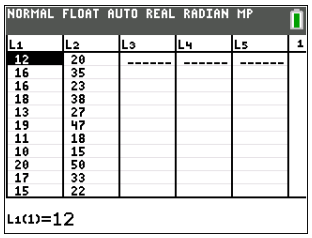
To plot a scatter diagram using our GC,

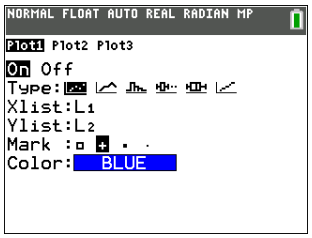
Step 1: Press **stat** and select "1.Edit".

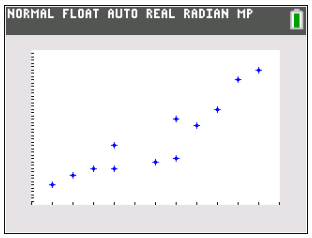
Step 2: Enter the values of the indepdent variable in the list L_1. Enter the values of the dependent variable into L_2.

Step 3: Press 2nd Y=, select "1: Plot 1" and turn on the Stat Plot function. Ensure that the axes are set to correspond to the correct lists L_1 and L_2.

Step 4: Press ZOOM 9 to view the scatter diagram.







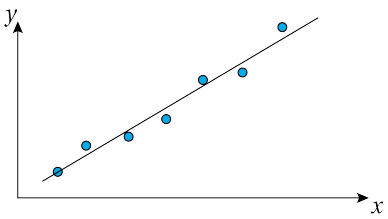
3 Correlation

Correlation is a measure of the degree of association between two vairables.

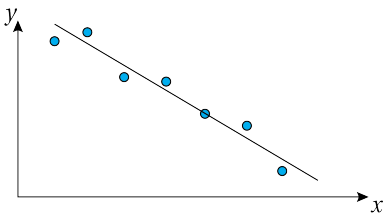
There are several characteristics we consider when describing the correlation between two variables:

- **Direction**
- **Linearity**
- **Strength**

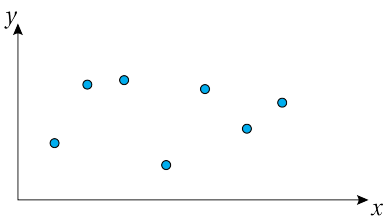
3.1 Direction



For a generally upward trend, we say that there is a **positive correlation**. An increase in the dependent variable generally results in an increase in the dependent variable.



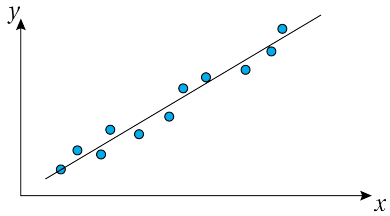
For a generally downward trend, we say that there is a **negative correlation**. An increase in the dependent variable generally results in a decrease in the dependent variable.



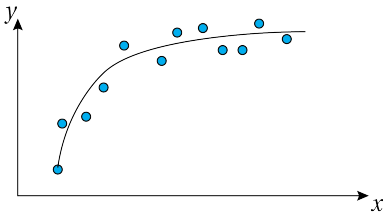
For randomly scattered points, with no upward or downward trend, we say that there is **no correlation**.

3.2 Linearity

When a trend exists, if the points approximately from a straight line, we say the trend is **linear**.



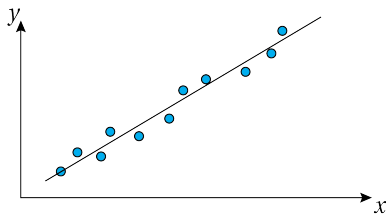
Linear Correlation



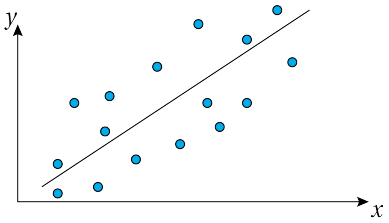
Non-Linear Correlation

3.3 Strength

To describe how closely the data follows a trend, we talk about the **strength** of the correlation. It is usually described as either **strong** or **weak**.



Strong Correlation

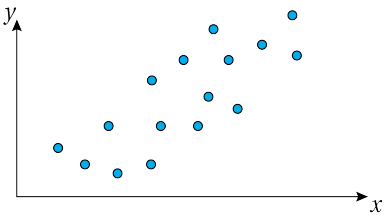


Weak Correlation

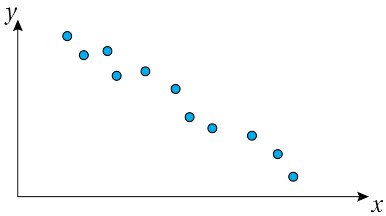
Example 8.1 Describing The Relationship Between Variables

For each scatter diagram, describe the relationship between the variables. Consider the direction, linearity and strength of the relationship.

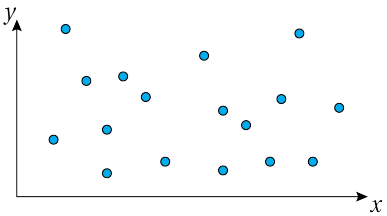
(a)



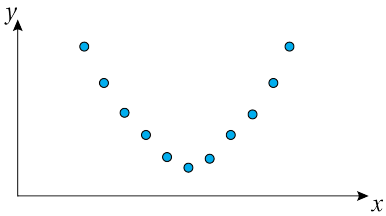
(b)



(c)



(d)



Solution

(a) Positive linear correlation.

(b) Strong negative linear correlation.

(c) No correlation.

(d) Non-linear correlation.

4 Causality

Correlation between two variables does not necessarily mean that one variable causes the other.
For example,

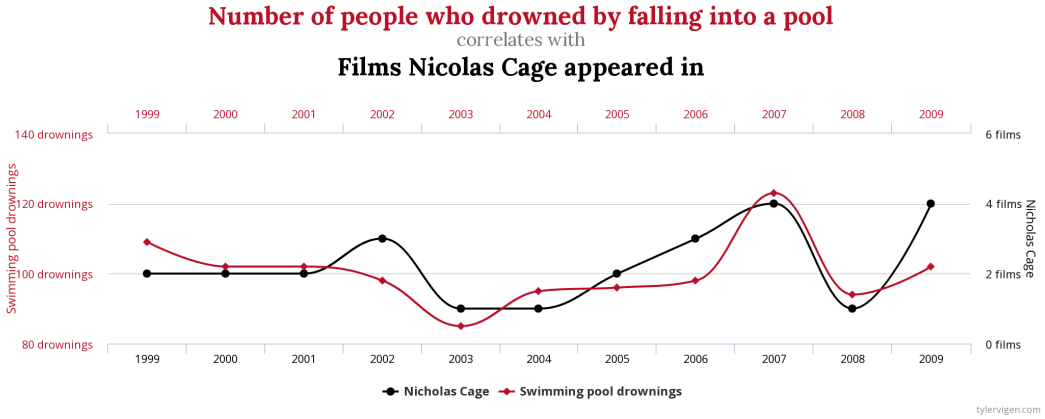
- The *arm length* and *running speed* of a sample of young children were measured, and a strong, positive correlation was found between the variables.

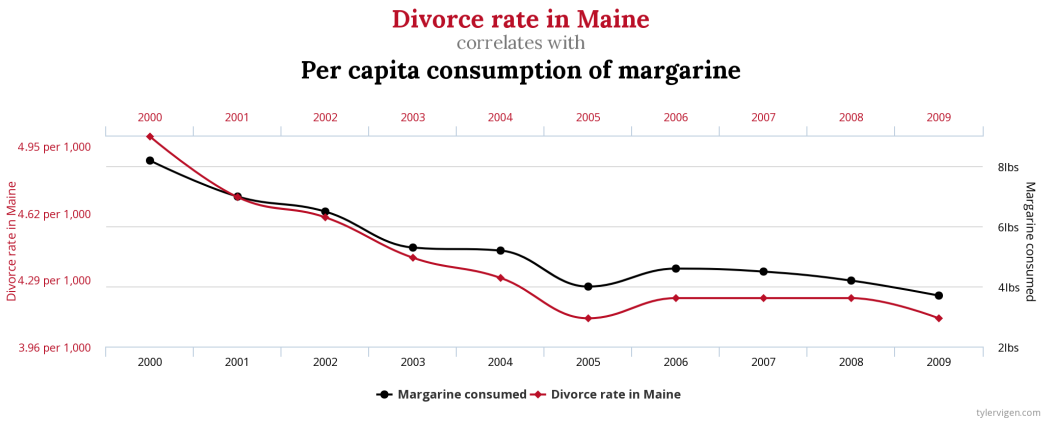
  This does not mean that short arms cause a reduction in running speed, or that a high running speed causes your arms to grow long.

  Rather, there is a strong, positive correlation between the variables because both *arm length* and *running speed* are closely related to a third variable, *age*. Up to a certain age, both *arm length* and *running speed* increase with *age*.

- Data is collected on the *total cancer incidence* and the *number of cell phone users*. A strong, positive correlation was found between the variables. Does this mean that cell phones cause cancer? Probably not. It is coincidental that both variables increased over this period of time.

If a change in one variable causes a change in the other variable then we say a **causal relationship** exists. In these cases, we can say that the independent variable explains the dependent variable. It may be more natural to use the terminology **explanatory variable** and **response variable** (instead of independent variable and dependent variable respectively).
In cases where a causal relationship is not apparent, we cannot conclude a causal relationship exists based on high correlation alone.
Here are some examples of spurious correlations. Source: https://tylervigen.com/spurious-correlations





5 Pearson's Product-Moment Correlation Coefficient, *r*

A scatter diagram visually shows the relation between two variables. In the previous section, we observed the points on a scatter diagram, and judged how strongly the points formed a linear relationship. In this section, we wil measure the strength of the linear relation between the variables by quantifying it.
The Pearson's Product-Moment Correlation Coefficient, denoted by r, is a numerical measure of the linear relation between two variables.
There are two equivalent formulae for r that can be found in your MF26.

$$r = \frac{\sum (x-\overline{x})(y-\overline{y})}{\sqrt{\sum (x-\overline{x})^2 \sum (y-\overline{y})^2}} \qquad \text{and} \qquad r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left(\sum x^2 - \frac{(\sum x)^2}{n}\right)\left(\sum y^2 - \frac{(\sum y)^2}{n}\right)}}$$

You are not required to use this formula, but you should be able to calculate r using your GC.
To find Pearson's Product-Moment Correlation Coefficient, r with your GC,

Step 1: Press **stat** and select "1.Edit".

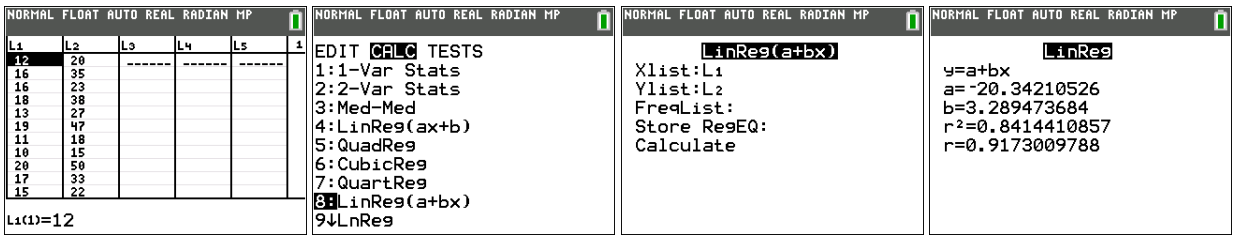
Step 2: Enter the values of the indepdent variable in the list L_1. Enter the values of the dependent variable into L_2.

Step 3: Press **stat** and right arrow to select "CALC".

Step 4: Select 8:LinReg(a+bx) and ensure that the Xlist and Ylist are set to the correct lists L_1 and L_2.

Step 5: Press "Calculate".

If you didn't get r, press **mode** and turn on "STAT DIAGNOSTICS".
Lets find the value of r for our **Opening Problem**.



Since $r = 0.917$ (3 s.f.) is close to 1, there is a strong positive linear correlation between the *age* of the athlete and *distance thrown*.

5.1 Properties of Pearson's Product-Moment Correlation Coefficient

- The values of r range from -1 to $+1$
- The sign of r indicates the direction of the correlation.
  - ▷ A positive value of r indicates the variables are positively correlated.
  - ▷ A negative value of r indicates the variables are negatively correlated.
- The size of r indicates the strength of the correlation.
  - ▷ A value of r close to 1 or -1 indicates a **strong** linear correlation.
  - ▷ A value of r close to 0 indicates no linear correlation.

Example 8.2 Finding The Strength of Correlation Using r

Jill does her washing every Saturday and hangs her clothes out to dry. She notices her clothes dry faster on some days than others. She investigates the relationship between temperature and the time her clothes take to dry.

Temperature (x°C)	25	32	27	39	35	24	30	36	29	35
Drying time (y minutes)	100	70	95	25	38	105	70	35	70	40

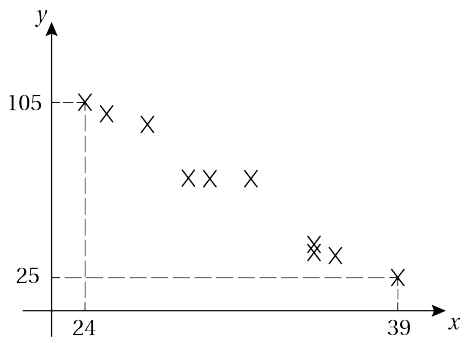
(a) Draw a scatter diagram for the data. (Indicate the maximum and minimum value on each axis)

(b) Calculate the product moment coefficient between x and y.

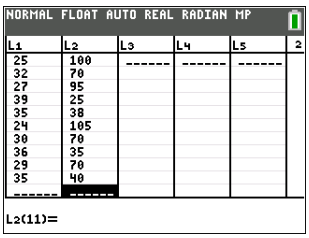
(c) Describe the correlation between *temperature* and *drying time*.

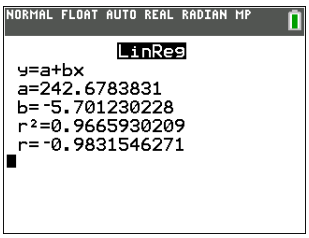
Solution

(a)



(b)





From GC, $r = -0.983$ (3 s.f.).

(c) There is a strong negative linear correlation between temperature and drying time.

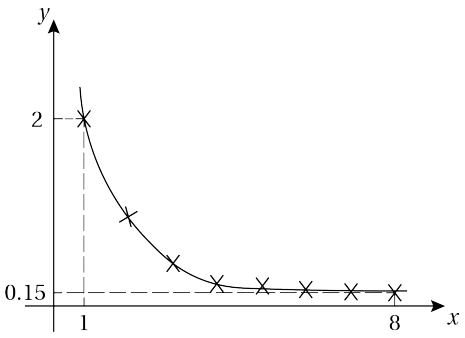
5.2 Limitations of Pearson's Product-Moment Correlation Coefficient

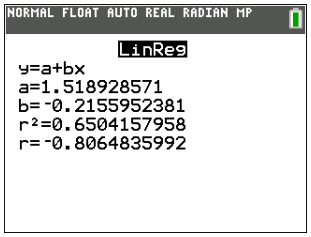
It is important to exercise logical thinking when analysing bivariate data. We need to plot the scatter diagram **and** find Pearson's Product-Moment Correlation Coefficient if we want to get the full picture of the relationship between the variables.

Non-Linear Relationships

Consider the following set of data

x	1	2	3	4	5	6	7	8
y	2	1	0.5	0.25	0.2	0.18	0.16	0.15





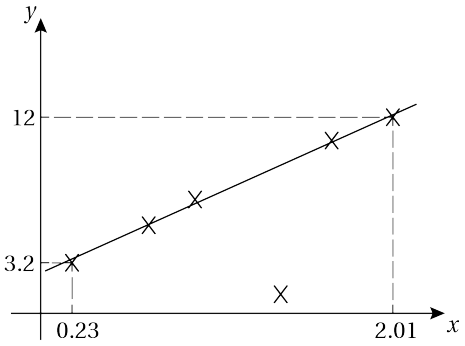
Even though $r = -0.806$ indicates a strong negative linear correlation, the scatter diagram shows that there is a non-linear relationship between the variables.

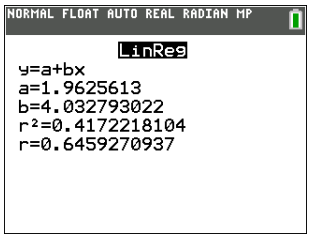
Outliers

Outliers are isolated points which do not follow the trend formed by the main body of the data.

Consider the following set of data

x	0.23	0.65	0.90	1.39	1.67	2.01
y	3.2	5.4	7.1	1.2	10.5	12





Even though $r = 0.646$ does not indicate a strong positive linear correlation, the scatter diagram shows that most of the points follow a linear relationship, except for the outlier $(1.39, 1.2)$. If that point is excluded, $r \approx 1.00$. There is actually a strong linear correlation between x and y.

Example 8.3 Effect of Outliers in Data

The table shows the number of supermarkets in 10 towns, and the number of car accidents that have occurred in these towns in the last month.

Number of Supermarkets, x	5	8	12	7	6	2	15	10	7	3
Number of car accidents, y	10	13	27	19	10	6	40	30	22	37

(a) Calculate the product moment coefficient between x and y. What does this indicate about the relationship between x and y?

(b) Draw a scatter diagram for the data.

(c) Identify the outlier in the data. What effect will this point have on the product moment coefficient?

(d) If was found that the outlier was due to an error in the data collection process.

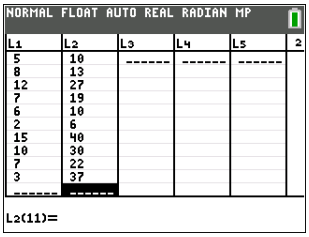
(i) Recalculate r with the outlier removed.

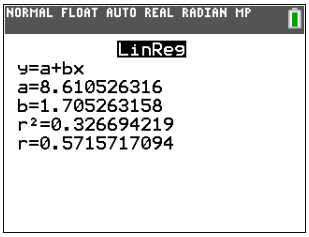
(ii) Describe the relationship between the variables.

(e) Do you think there is a causal relationship between the variables? If not, propose a possible cause for the trend in the data.

Solution

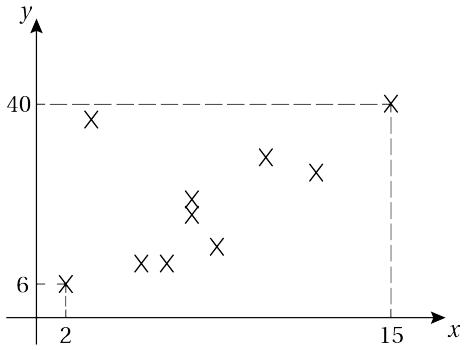
(a)





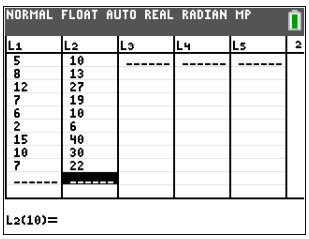
From GC, $r = 0.571$ (3 s.f.). This indicates that there is a weak positive linear correlation between x and y.

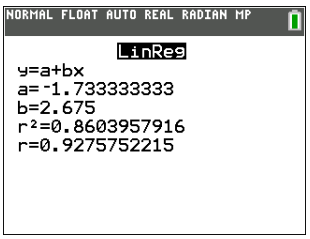
(b)



(c) The point (3,37) is an outlier. It will cause r to be closer to 0.

(d) (i)





From GC, $r = 0.928$ (3 s.f.).

(ii) There is a strong positive linear correlation between the number of supermarkets and the number of car accidents in a town.

(e) No, it is not a causal relationship. A more plausible explnation is that both variables depend on the number of people in each town. A larger population would result in more supermarkets as well as more car accidents.

6 Linear Regression

6.1 Independent and Dependent Variables

Recall that

- The **independent** (or **explanatory**) variable is the **cause**. Its value is independent of the other variable.
- The **dependent** (or **response**) variable is the **effect**. Its value depends on changes in the independent variable.

Example 8.4 Identifying Independent and Dependent Variables

For each of the following, state which is the independent variable and dependent variable respectively.

(a) An experiment was carried out to examine the relationship between the temperature, x (in °C) and the yield of tomatoes, y (in kg) on a farm.

(b) An experiment was carried out to examine the relationship between the volume of water, V (in cm^3) given to plants and the plants height, h (in cm).

(c) An experiment was carried out to examine the relationship between the scores on a Physics test, x and the scores on a Math test, y.

Solution

(a) x is the independent variable and y is the dependent variable.

(b) V is the independent variable and h is the dependent variable.

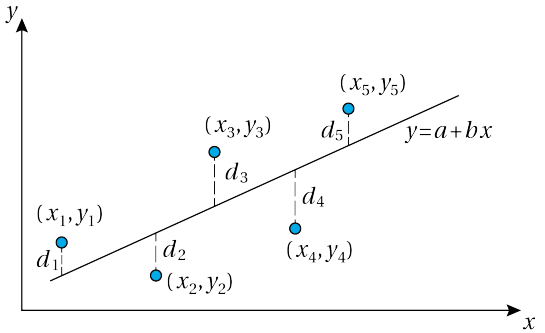
(c) There is insufficient information to determine which is the independent variable. We can use either depending on whether we are trying to estimate x or y.

6.2 Least Squares Regression Line of y on x

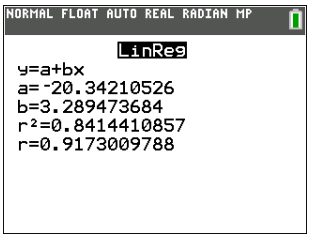
Linear regression attempts to model the relationship between two variables by fitting a linear equation to the set of observed data. For bivariate data that has a linear relationship, we can obtain the line of best fit by the method of "**least squares**".

In least squares linear regression, we minimise the sum of the squares of the vertical distances between each data point and the regression line.

In other words, we need to find the straight line $y = a + bx$, where a and b are chosen to minimse $D = \sum_{i=1}^{n} d_i^2$.



We can find the equation of linear regression of y on x with our GC on the same tab we found r, using the "LinReg(ax+b)" function. Recall in our **Opening Problem**:



From our GC, we can see that our least squares regression line of y on x is $y = -20.3 + 3.29x$ (3 s.f.).

Interpreting Regression Line of y on x

Given the equation of the regression line is $y = a + bx$,

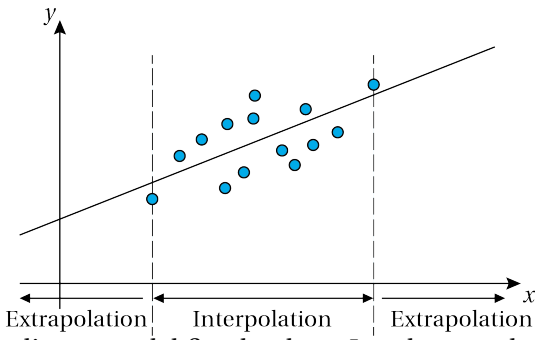
- b represents the slope of the regression line and can be interpreted as the change in y per unit change in x.
- a represents the intercept of the regression line with the $y-$axis and can be interpreted as the estimated value of y when $x = 0$

(Try to give your answer in the context of the question)

6.3 Interpolation and Extrapolation

Consider the data in the scatter diagram alongside. The data with the highest and lowest values are called the poles.

The line of best fit for the data is also drawn on the scatter diagram. We can use this line to predict the value of one variable given the value of the other.



- If we predict a y value for an x value **in between the poles**, we say we are **interpolating**.
- If we predict a y value for an x value **outside the poles**, we say we are **extrapolating**.

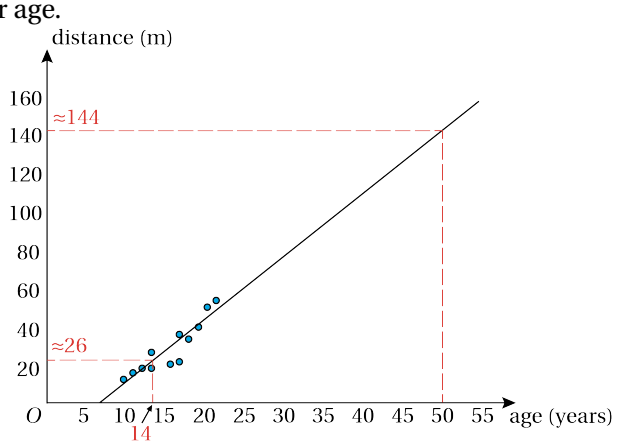
The accuracy of an interpolation depends on how well the linear model fits the data. In other words, our estimate is only reliable if coupled with a strong linear correlation.

The accuracy of an extrapolation depends not only on how well the model fits, but also on the assumption that the linear trend will continue past the poles. The validity of this assumption greatly depends on the situation we are looking at.

Let us consider the line of best fit for the data in the **Opening Problem**. It can be used to predict the distance a discus will be thrown by an athelete at a particular age.

The age 14 is within the range of ages in the original data, so it is reasonable to predict that a 14 year old will be able to throw the discus 26 m.

However, it is unlikely that the linear trend shown in the data will continue far beyond the poles. For example, the line predicts that a 50 year old would throw the discus 144 m. This is almost twice the current world record, so it would clearly be an unreasonable prediction.



Example 8.5 Estimating By Interpolation and Extrapolation

The annual income and average weekly grocery bill for a selection of families is shown below:

Income (x thousand dollars)	55	36	25	47	60	64	42	50
Grocery bill (y dollars)	120	90	60	160	190	250	110	150

(a) Sketch a scatter diagram to illustrate the data.

(b) Find the product moment correlation coefficient. Describe the relationship between the variables.

(c) Find the equation of the regression line. State and interpret its gradient.

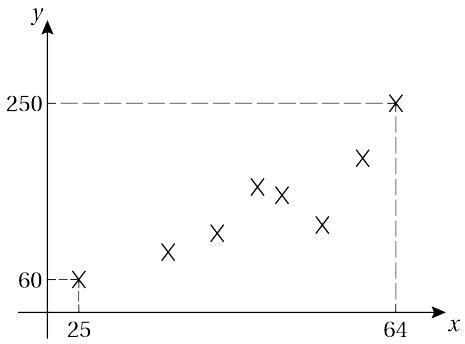
(d) Estimate the weekly grocery bill for a family with an annual income of \$95 000.

(e) Estimate the annual income of a family whose weekly grocery bill is \$100.

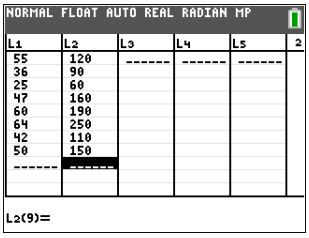
(f) Comment on whether the estimates in c and d are likely to be reliable.

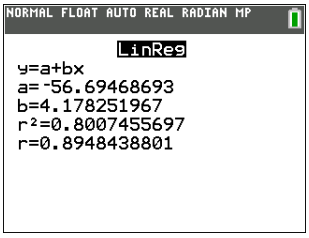
Solution

(a)



(b)





From GC, $r = 0.895$ (3 s.f.). There is a strong positive linear correlation between the annual income and the weekly grocery bill.

(c) $y = -56.7 + 4.18x$ (3 s.f.)

The gradient of the line of regression is 4.18. This means that for every additional \$1000 of income, a family's weekly grocery bill will increase by an average of \$4.18.

(d) When $x = 95$,

$$\begin{aligned} y &= -56.695 + 4.1783\,(95) \\ &= 340 \text{ (3 s.f.)} \end{aligned}$$

So, we expect a family of income \$95 000 to have a weekly grocery bill of \$340.

(e) When $y = 100$,

$$100 = -56.695 + 4.1783x$$
$$x = 37.5 \text{ (3 s.f.)}$$

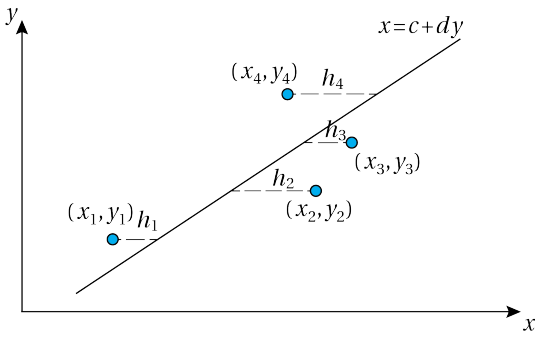
So, we expect a family with a weekly grocery bill of \$100 to have an annual income of approximately \$37 500.

(f) The estimate in (c) is an extrapolation, so the estimate may not be reliable as it cannot be assumed to be valid for values of x way beyond the given range.

The estimate in (d) is an interpolation and since there is a strong linear correlation between the variables, we can expect this estimate to be reliable.

6.4 Least Squares Regression Line of x on y

When y is the independent variable and x is the dependent variable, we will use the least squares regression of x on y. In this case, we minimise the horizontal distances of points from the line.
We consider a line of the form $x = c + dy$, and choose the constants m and c to minimise $H = \sum_{i=1}^{n} h_i^2$.



In general, the regression line of x on y is not the same as the regression line of y on x. (Unless $r = \pm 1$)

Scenario	To estimate the value of y for a given value of x	To estimate the value of x for a given value of y
x is the independent variable and y is the dependent variable	y on x	y on x
y is the independent variable and x is the dependent variable	x on y	x on y
No dependence of x or y on each other	y on x	x on y

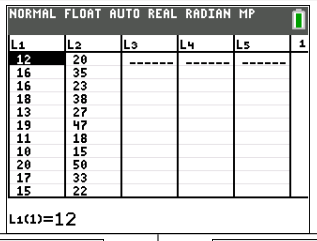
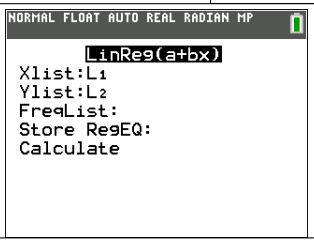
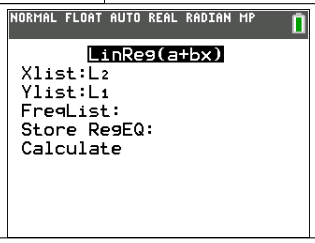
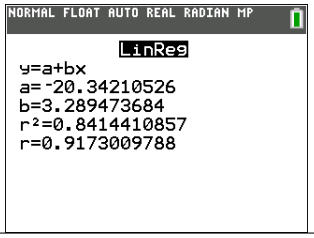
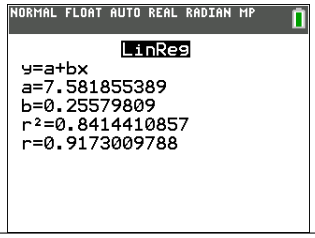
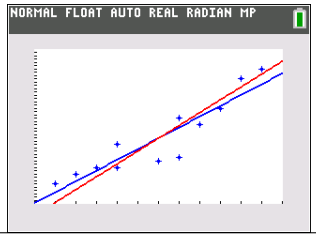
Note:

(a) $\overline{x} = \dfrac{\Sigma x}{n}$ and $\overline{y} = \dfrac{\Sigma y}{n}$

(b) Both regression lines pass through $(\overline{x}, \overline{y})$.

(c) The point of intersection of the least squares regression lines of y on x and x on y is $(\overline{x}, \overline{y})$.

As a last exercise for our **Opening Problem**, lets try to compare the regression line of y on x and the regression line of x on y.
To find our regression line of x on y, we can use the same data stored in the GC, just swap L_1 and L_2 in the "LinReg(ax+b)" tab.

Regression line of y on x	Regression line of x on y
	
	
	
$y = -20.3 + 3.29x$	$x = 7.58 + 0.256y$
	

- r is the same for both lines of regression because x is as correlated with y as y is with x.
- The lines meet at the point $(\overline{x}, \overline{y})$, which in this case is $(29, 15)$. (Can be found using "2−Var Stats")

Example 8.6 Estimation Using Regression Lines

The following data were collected during a study, under experimental conditions, of the effect of temperature x°C, on the pH, y of skimmed milk.

x	4	9	17	24	32	40	46	57	63	69	72	78
y	6.85	6.75	6.74	6.73	6.68	6.52	6.54	6.48	6.36	6.33	6.35	6.29

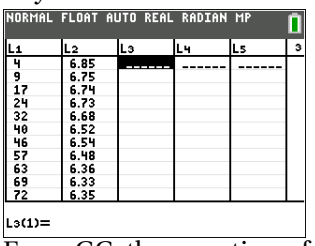
Use the appropriate least squares regression line(s) to estimate

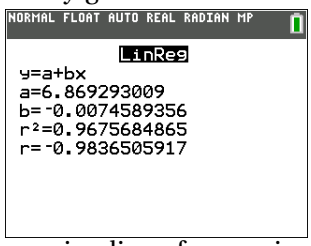
(a) the pH of skimmed milk when the temperature is 27°C,

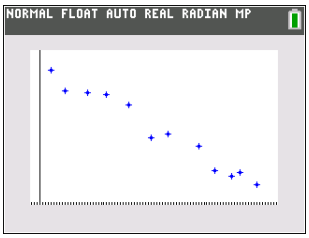
(b) the temperature (to the nearest °C) of the skimmed milk when the pH is measured to be 6.64.

Solution

Since x is the independent variable and y is the dependent variable, we will use the regression line of y on x to estimate values of x and y given the other.







From GC, the equation of the regression line of y on x is

$$y = 6.8693 - 0.0074589x \text{ (3 s.f.)}$$

(a) When $x = 27$,

$$\begin{aligned} y &= 6.8693 - 0.0074589(27) \\ &= 6.67 \text{ (3 s.f.)} \end{aligned}$$

Thus, the pH of skimmed milk when the temperature is 27°C is estimated to be 6.67.

(b) When $y = 6.64$,

$$\begin{aligned} 6.64 &= 6.8693 - 0.0074589x \\ x &= 31 \text{ (to nearest °C)} \end{aligned}$$

Thus when the pH is 6.64, the temperature of the milk is estimated to be 31°C.

Example 8.7 Estimation Using Regression Lines

The following table shows the marks scored by 10 randomly selected students in a French test and in an English test.

French score, x	20	43	33	56	50	67	73	68	77	43
English Sscore, y	19	42	44	52	51	53	66	56	60	37

(a) Sketch a scatter diagram to illustrate the data.

(b) Find the product moment correlation coefficient.

(c) Find the equation of the regression line y on x in the form $y = a + bx$, giving the values of a and b correct to 3 significant figures.

(d) Find the equation of the regression line x on y in the form $x = c + dy$, giving the values of c and d correct to 3 significant figures.

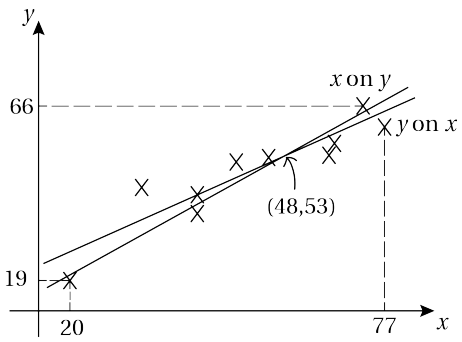
(e) Sketch both lines on the scatter diagram and find the point of intersection between the two regression lines.

(f) A student, Tom, was absent for the French test. Predict a suitable mark for him if he scored 65 in the English test.

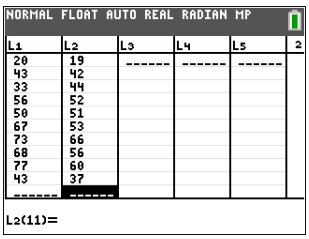
(g) Another student, Jerry, was absent for the English test. Predict a suitable mark for him if he scored 70 in his French test.

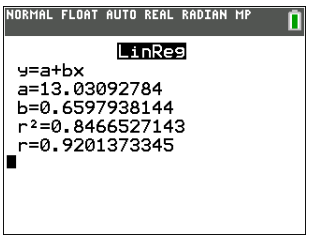
Solution

(a)



(b)



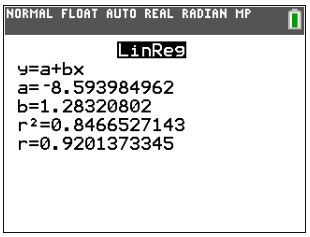


From GC, $r = 0.920$ (3 s.f.).

(c) The regression line of y on x is

$$y = 13.031 + 0.65979x \text{ (5 s.f.)}$$
$$= 13.0 + 0.660x \text{ (3 s.f.)}$$

(d)



The regression line of x on y is

$$\begin{aligned} x &= -8.5940 + 1.2832y \text{ (5 s.f.)} \\ &= -8.59 + 1.28y \text{ (3 s.f.)} \end{aligned}$$

(e) From GC, the point of intersection between the two lines is $(48, 53)$.

(f) Since neither variable is dependent on another, the choice of regression line depends on which value is being estimated.

Here, we need to predict the value of x given y, so we use the regression line of x on y.

When $y = 65$,

$$\begin{aligned} x &= -8.5940 + 1.2832\,(65) \\ &= 75 \text{ (to nearest integer)} \end{aligned}$$

The predicted mark for his French test is 75.

(g) Here, we need to predict the value of y given x, so we use the regression line of y on x.

When $x = 70$,

$$\begin{aligned} y &= 13.031 + 0.65979\,(70) \text{ (5 s.f.)} \\ &= 59 \text{ (to nearest integer)} \end{aligned}$$

The predicted mark for his English test is 59.

7 Transformations to Linearize Bivariate Data

Often a straight-line pattern is not the best model for depicting a relationship between two variables. A clear indication of this problem is when the scatterplot shows a distinctive curved pattern. In such a case, the nonlinear model can sometimes be revealed by transforming one or both of the variables and then noting a linear relationship. Below we see some transformations for common non-linear equations.

Non-Linear Equations	Transformed Variables
$y = ax^2 + b$	$y = aw + b$, where $w = x^2$
$y = a\sqrt{x} + b$	$y = aw + b$, where $w = \sqrt{x}$
$y = \frac{a}{x} + b$	$y = aw + b$, where $w = \frac{1}{x}$
$y = ab^x$, where $a > 0, b > 1$	Consider $\ln y = \ln(ab^x) \Rightarrow \ln y = \ln a + x \ln b$ $w = \ln a + x \ln b$, where $w = \ln y$.

Example 8.8 Linearizing Bivariate Data

A stationary retailer supplies pens to offices. In order to encourage customers to buy in bulk, the stationary retailer comes up with the following price scheme.

Number of pens ordered, x	5	10	20	40	80	160
Unit price in dollar, y	1.50	1.20	1.10	0.99	0.90	0.80

(a) Calculate the value of the product moment correlation coefficient.

(b) Sketch a scatter diagram to illustrate the data and hence comment on the value found in (a).

(c) State with reason, which model of the form $y = a + bw$ is appropriate, where a and b are constants, $b > 0$ and w is as follows.

A: $w = x^2$
B: $w = \frac{1}{x}$
C: $w = \ln x$

Using the appropriate model selected above, find

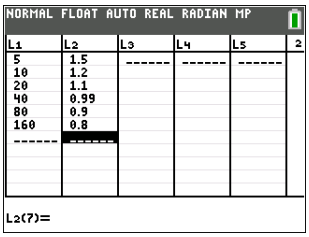
(d) the value of the product moment correlation coefficient,

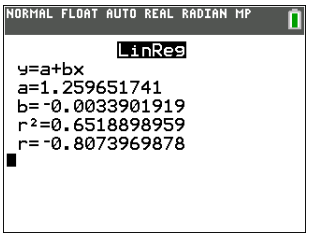
(e) the equation of the regression line of y on w, and hence obtain an estimate of the number of pens ordered (to the nearest whole number) when a unit price of \$1.40 is quoted.

Comment on the reliability of your answer.

Solution

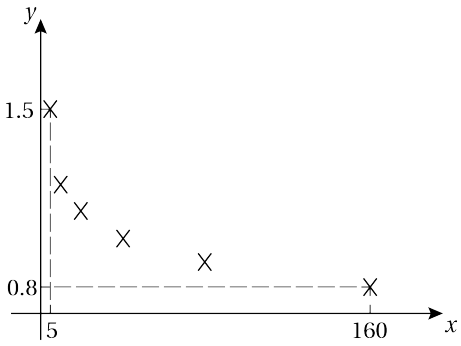
(a)





From GC, $r = -0.807$ (3 s.f.).

(b)

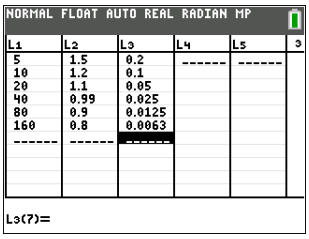


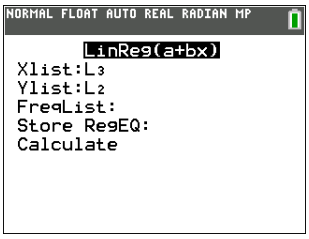
The value found in (a) suggests there is a strong negative linear correlation between x and y, but the scatter diagram shows that the relationship between x and y is non-linear. Thus, the value of r is not a good indication of the linearity between x and y.

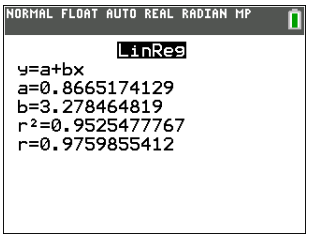
(c) For models **A** and **C**, y increases as x increases.

From the scatter diagram, we see that y decreases as x increases, which suggests that model **B** would be most appropriate.

(d) With our x and y values stored in L_1 and L_2 respectively, we can enter $L_3 = \dfrac{1}{L_1}$ to obtain the list for $\dfrac{1}{x}$.







From GC, the product moment correlation coefficient between y and w is $r = 0.976$ (3 s.f.).

(e) The equation of the regression line of y onto w is

$$y = 0.86652 + 3.2785w \text{ (5 s.f.)}$$
$$= 0.867 + 3.28w \text{ (5 s.f.)}$$

When $y = 1.40$,

$$1.4 = 0.86652 + 3.2785w$$
$$w = 1.6272$$
$$\frac{1}{x} = 1.6272$$
$$x = 6 \text{ (to the nearest whole number)}$$

This estimate is highly reliable since $y = 1.40$ is within the poles, which is an interpolation, and the product moment correlation coefficient between y and w is 0.976 (3 s.f.), which suggests that there is a strong positive linear correlation between y and $\dfrac{1}{x}$.

Example 8.9 Linearizing Bivariate Data

The number of computers sold by a shop in six successive years is given below.

Year, x	1	2	3	4	5	6
Sales, y	11	29	68	138	215	560

The company believes that the sales, y, and the year, x, are related by the equation $y = A(B)^x$, where A and B are constants.

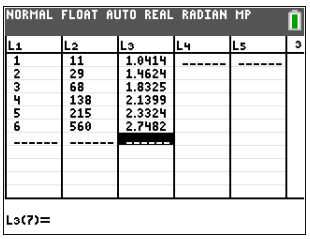
(a) Using the transformation involving $w = \log y$, give a sketch of the data and explain whether the relation is a resonable model. Find the least squares estimate of A and B, and draw the estimated regression line on your sketch.

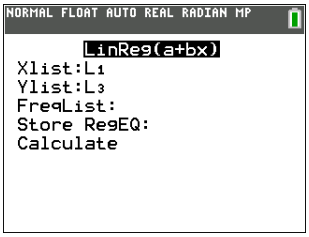
(b) The owner of the shop uses this relation to predict the sales in year 12. Find the predicted sales to the nearest whole number and comment on the prediction.

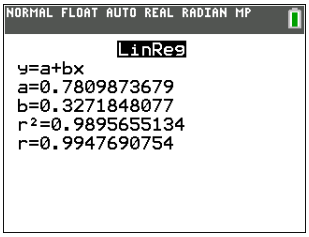
Solution

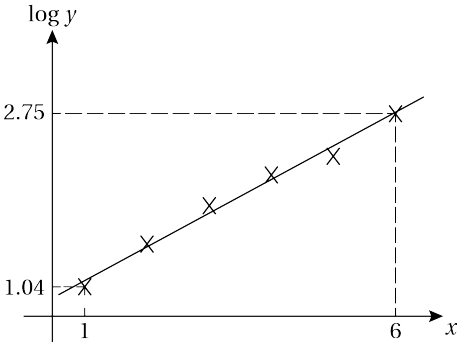
(a) $y = A(B)^n$

$\log y = \log A + n \log B$









After the transformation, there is a strong linear correlation between x and $\log y$. Hence $y = A(B)^x$ is a reasonable model.

From GC, the equation of the line of regression of $\log y$ on x is

$$\log y = 0.79099 + 0.32718x \text{ (5 s.f.)}$$

$$\log A = 0.79099 \qquad \log B = 0.32718$$
$$A = 10^{0.79099} \qquad B = 10^{0.32718}$$
$$= 6.04 \text{ (3 s.f.)} \qquad = 2.12 \text{ (3 s.f.)}$$

(b) When $x = 12$,

$$\begin{aligned} \log y &= 0.79099 + 0.32718\,(12) \\ &= 4.7072 \\ y &= 10^{4.7072} \\ &= 50957 \text{ (to the nearest whole number)} \end{aligned}$$

Since $x = 12$ is outside the poles, it is an extrapolation. Even though there is a strong linear correlation within the poles, we cannot be certain if the relationship between $\log y$ and x is still linear beyond the poles. Thus, this prediction is not very reliable.

Example 8.10 Linearizing Bivariate Data

In an agricultural experiment, a certain fertilizer is applied at different rates to ten identical plots of land. Seeds of a type of grass are then sown and several weeks later, the mean height of the grass of each plot is measured. The results are shown in the table.

Rate of application of fertilizer (x g/m^2)	10	20	30	40	50	60	70	80	90	100
Mean height of grass (y cm)	6.2	11.4	13.2	14.8	15.8	17.0	19.4	19.4	20.6	20.8

(a) Draw the scatter diagram for these values, labelling the axes clearly.

It is thought that the mean height, y cm, can be modelled by one of the formulae

$$y = a + bx \quad \text{or} \quad y = c + d\ln x$$

where a, b, c and d are constants.

(b) Find, correct to 4 decimal places, the value of the product moment correlation coefficient between

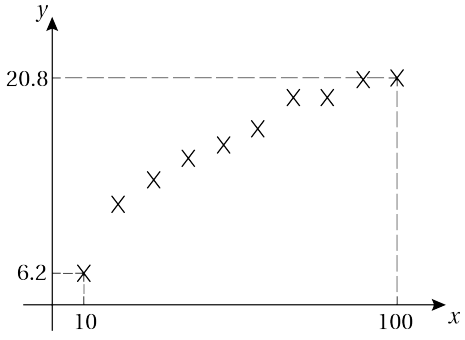
(i) x and y (ii) $\ln x$ and y

(c) Use your answers in parts (a) and (b) to explain which of $y = a + bx$ or $y = c + d\ln x$ is the better model.

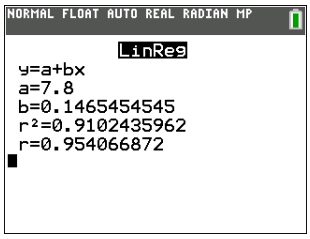
(d) It is required to estimate the value of x for which $y = 17.2$. Find the equation of a suitable regression line and use it to find the required estimate.

Solution

(a)

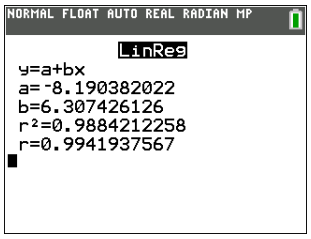


(b) (i)



From GC, $r = 0.9541$ (4 d.p.).

(ii)



From GC, $r = 0.9942$ (4 d.p.).

(c) The scatter diagram shows a non-linear correlation between x and y. In addition, the product moment correlation coefficient between $\ln x$ and y, 0.9942, is closer to 1 as compared to that between x and y, 0.9541. Hence, $y = c + d\ln x$ is the better model.

(d) Equation of regression line of y on $\ln x$ is

$$\begin{aligned} y &= -8.1904 + 6.3074\ln x \text{ (5 s.f.)} \\ &= -8.19 + 6.31\ln x \text{ (3 s.f.)} \end{aligned}$$

When $y = 17.2$,

$$\begin{aligned} 17.2 &= -8.1904 + 6.3074\ln x \\ x &= 56.0 \text{ (3 s.f.)} \end{aligned}$$