

Module 2

Statistical tools and probability theory

Upon completion of this module the student will be able to:



Outcomes

- **display** data in the form of graphs and charts;
- **identify** appropriate statistical tools in analysing data (mean, median, and so on);
- **interpret** results correctly to inform decisions;
- **understand** probabilities of events;
- **handle** probability distributions and understand their role in modelling; and
- **apply** expected values, variance and standard deviation to risk and volatility.



Unit 5

Collecting and presenting data

The basis of decision-making lies in data and information. Data have to be analysed and interpreted before we can present it as information and use it to make management decisions.

Upon completion of this unit the student will be able to:



Outcomes

- **classify** different types of data;
- **represent** data in tables;
- **display** data in charts and graphs; and
- **draw** and **display** frequency tables.

Data collection and tables for data

Activity 2.1



Activity

Interpret graphs

What will you do?

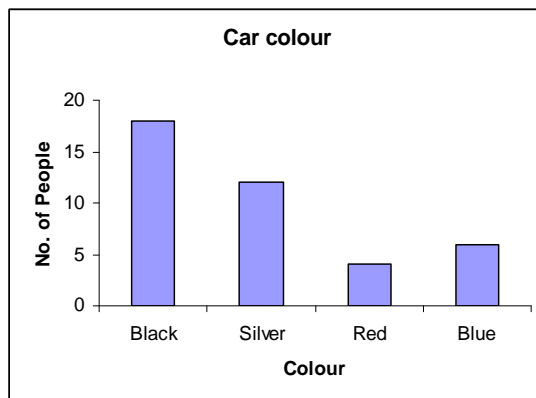
1. Calculate the size of slice each colour makes in the pie chart in Figure 6 and show that the slices add up to 360° .
2. Represent the car colour data as a column chart.

Solutions

1. Black: 45% of $360^\circ = 162^\circ$
Silver: 30% of $360^\circ = 108^\circ$
Red: 10% of $360^\circ = 36^\circ$
Blue: 15% of $360^\circ = 54^\circ$

The percentages add up to 100% and the angles add up to 360° .

2.



Frequency distributions: tables and histograms

Activity 2.2



Activity

Make graphs and tables

What will you do?

1. You want to conduct a survey to find out which new flavour of coffee will sell well in your country. The available flavours are chocolate, hazelnut, cinnamon and rum. Describe how you would go about gathering data.
2. Your supplier in another country has this data for her own country:
 - 36% of people preferred hazelnut,
 - 28% preferred chocolate,
 - 22% preferred rum
 - 8% preferred cinnamon, and
 - the remainder didn't like any of the flavours.

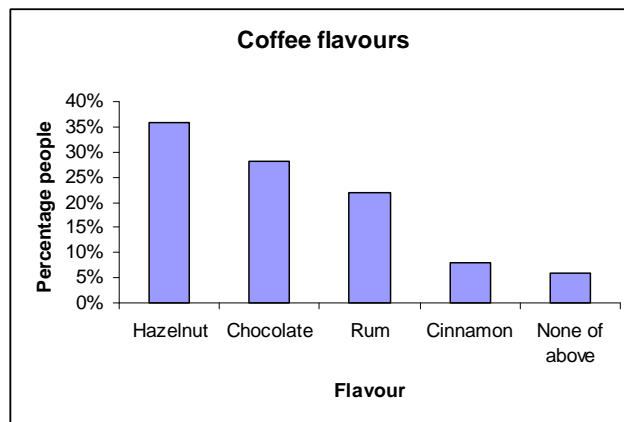
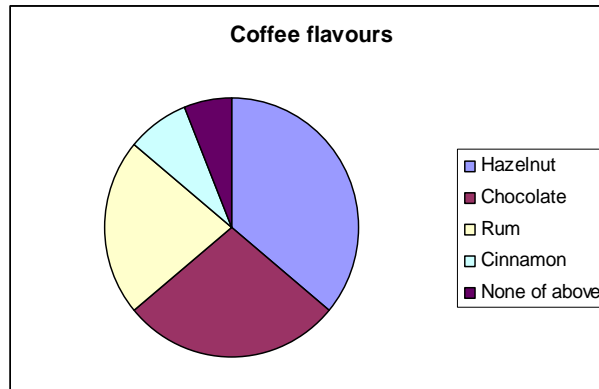
Present the supplier's findings graphically in two ways. What conclusion can you draw?

3. Draw a frequency table, histogram and ogive for the following data. Of 2,000 adults surveyed:
 - 8% had a mass between 120 kg and 140 kg,
 - 20% had a mass between 100 kg and 120 kg,
 - 36% had a mass between 80 kg and 100 kg,
 - 28% had a mass between 60 kg and 80 kg, and
 - the remainder had a mass between 40 kg and 60 kg.



Solutions

1. Surveys can be done by post, email, telephone, interviewing people in coffee shops, and so on. Data can be presented in tables, column charts, pie charts, and so on.
2. Column and pie charts can be used.

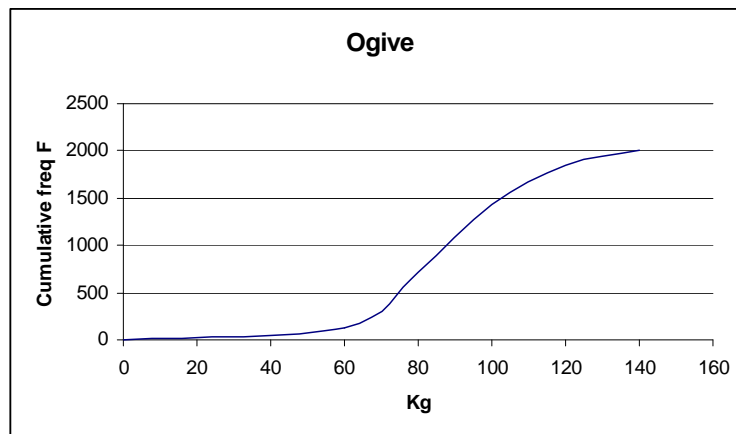
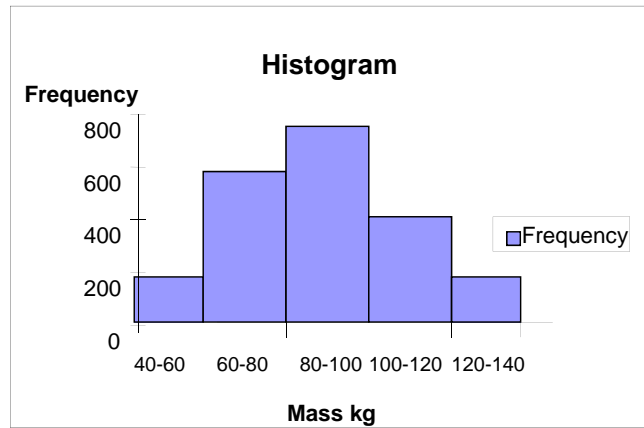


The disadvantage of the pie chart is that the percentages are not stated. It may be difficult to distinguish between the popularity of chocolate versus rum flavours. Percentage data also does not indicate how large the sample was.

- 3.

Mass in kilograms (x)	Number of people (frequency f(x))	Cumulative frequency F(x)
40–60	160	160
60–80	560	720
80–100	720	1440
100–120	400	1840
120–140	160	2000

Histogram





Unit 6

Measures for analysing data

Upon completion of this unit students will be able to:



Outcomes

- **calculate** the mean value of discrete data;
- **calculate** the median and mode of discrete data;
- **determine** the mean and median for grouped data with frequency distributions;
- **apply** different measures of spread or dispersion of data;
- **compare** different data sets using co-variance or correlation; and
- **critique** some uses of statistics.

Mean, median and mode

Activity 2.5



Activity

Practise some more

1. Find the mean, median and mode of this set of data:
5, 2, 6, 4, 10, 7, 9, 11, 4, 7, 8, 3
Represent the data in a scatter graph.
2. Study the frequency table for continuous data:

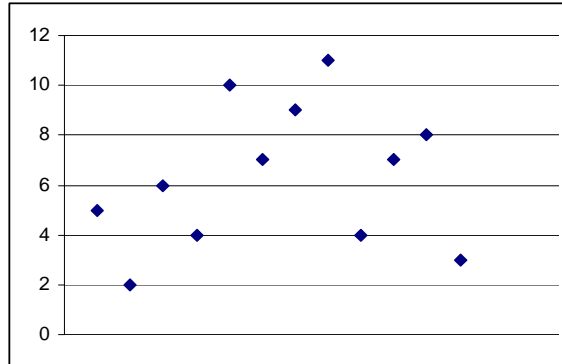
Class of values (x)	Frequency $f(x)$
2.0–4	4
4.0–6	7
6.0–8	15
8.0–10	10
10.0–12	2

Figure 2

- a) Expand the table to include cumulative and relative frequencies.
- b) Find the mean and median for the distribution.
- c) Draw a histogram and ogive.
- d) Use the ogive to confirm the value of the median.

Solutions

1. Mean = 6.33 (rounded), median = 6.5, modes = 4 and 7



2. a)

Class of values (x)	Frequency $f(x)$	Cumulative frequency F	Relative frequency
2.0–4	4	4	0.1053
4.0–6	7	11	0.1842
6.0–8	15	26	0.3947
8.0–10	10	36	0.2632
10.0–12	2	38	0.0526

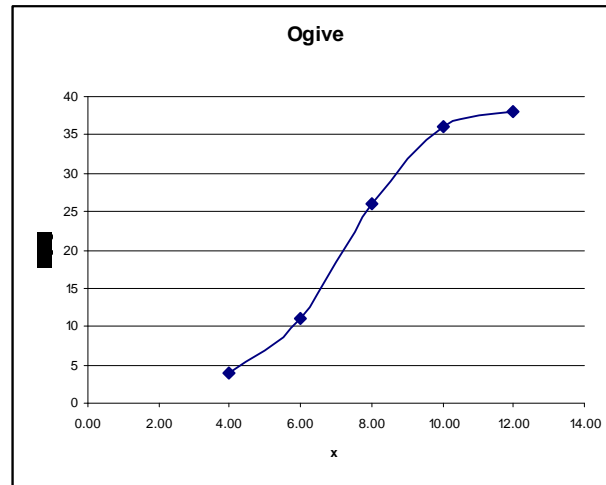
- a) Midpoints x_i of classes: 3, 5, 7, 9, 11

$$\begin{aligned} \text{Mean} &= \frac{1}{N} \sum_{i=1}^N x_i f(x_i) = \frac{1}{38} [3(4) + 5(7) + 7(15) + 9(10) + \\ & 11(2)] \\ &= 6.95 \text{ (rounded)} \end{aligned}$$

The middle values are observations 19 and 20, which fall in class 6–8. Since $19 - 11 = 8$ and $20 - 11 = 9$, the middle value takes position between 8 and 9 of the 15 values in class 6–8. Now $6 + (8/15)2 = 7.07$ and $6 + (9/15)2 = 7.2$ so the median is $\frac{1}{2}(7.07 + 7.2) = 7.135$.

- c) Straightforward

d)



The midpoint of values on the vertical axis is 17. The corresponding x -value is about 7.

Skewness and spread of data



Case study: Investment decisions

You are analysing the daily price movements of two shares, with a view to making a recommendation about buying one, or both, of the shares.

These numbers are the share prices (in dollars) of shares S_1 and S_2 collected daily at the close of the trading day for 21 days in February 2009.

Day	1	2	3	4	5	6
S_1	12,00	12,15	12,85	14,00	13,80	13,80
S_2	37,00	36,20	35,00	34,80	38,00	44,50

Day	7	8	9	10	11
S_1	11,00	10,40	10,00	13,00	15,50
S_2	46,00	47,80	45,00	42,00	39,80

Day	12	13	14	15	16
S_1	16,50	16,00	15,00	14,50	14,20
S_2	35,00	48,00	52,40	54,00	54,20

Day	17	18	19	20	21
S_1	13,00	12,20	13,00	14,00	14,20
S_2	56,00	58,20	52,40	51,00	50,00

The daily rate of return for S_1 at the end of day two is calculated as:

$$\frac{12,15 - 12,00}{12,00} = 0,0125 = 1,25\%$$

Similarly, the daily rate of return for S_2 for day 10 is:

$$\frac{42,00 - 45,00}{45,00} = -0,067 = -6,7\%$$

1. Draw a table to show the daily prices and the 20 daily rates of return for each share.
2. Use graphs to present your data in an attractive way.
3. Find the mean daily rates of return over the month for each share.
4. Calculate the one-day standard deviation of rates of return for each share over the month. Use this to estimate the monthly volatility of each share.
5. Calculate the correlation between the two share price processes, as well as the correlation between the returns for the month.
6. Do some research on the topic of “diversification in investments” and write a short report on this. Show how diversification is linked with correlation of share prices.
7. Write a short paragraph comparing the two shares in terms of expected returns and volatility. What recommendations will you make for investment in these shares?

Solutions

1.

ReturnR1	ReturnR2
0.0125	-0.02162
0.057613	-0.03315
0.089494	-0.00571
-0.01429	0.091954
0	0.171053
-0.2029	0.033708
-0.05455	0.03913
-0.03846	-0.05858
0.3	-0.06667
0.192308	-0.05238
0.064516	-0.1206
-0.0303	0.371429



ReturnR1	ReturnR2
-0.0625	0.091667
-0.03333	0.030534
-0.02069	0.003704
-0.08451	0.03321
-0.06154	0.039286
0.065574	-0.09966
0.076923	-0.02672
0.014286	-0.01961

2. A variety of graphs can be drawn. Students can graph daily share prices or daily rates of return.
3.

Mean return	0.013508	0.020049
-------------	----------	----------
4.

Std dev return	0.10576	0.107628
cor(S1,S2)	-0.04361	
cor(R1,R2)	-0.39843	
5. The shares are negatively correlated. The combination of S1 and S2 gives some diversification in a portfolio and protection. Information on diversification is available on the Internet.
6. Negative correlation means as one variable tends to rise above its mean, the other tends to fall below its mean. In portfolio analysis, this provides protection by not having all prices rise and fall together (positive correlation). This is what is meant by diversification. One should try to include a number of shares in a portfolio, some of which should be negatively correlated.
7. Share S2 has a higher daily rate of return and a slightly higher risk (volatility or standard deviation of return) than Share S1.

Recommendation: Based on this risk-return analysis only, it is worthwhile investing in Share S2. The increase in risk is very small (0.2%) for an increase in return of about 0.7% as compared to Share S1.

Unit 7

Basic probability theory

Upon completion of this unit students will be able to:



Outcomes

- **describe** outcomes and events for experiments;
- **calculate** probabilities for outcomes and events;
- **understand** conditional probability and Bayes' formula; and
- **identify** mutually exclusive and independent events.

Dependent, independent and mutually exclusive events

Activity 2.9



Activity

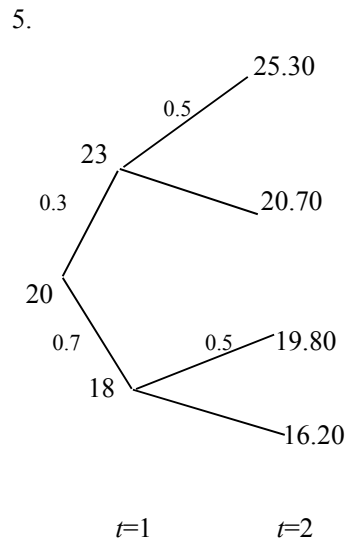
Calculate more
probabilities

1. Event E has probability 0.65 and event F has probability 0.3.
 - a) If the sample space Ω has 200 outcomes, how many observations are there in each event?
 - b) 1.2. What is $P(E^c)$? Calculate $P(E \cup E^c)$ and discuss your answer. Calculate $P(E \cap E^c)$ and discuss your answer.
 - c) If E and F are independent, determine $P(EF)$.
2. You have 120 staff members: 86 female and 34 male. There are 5% of staff members who are disabled. What is the probability that a staff member is female and disabled?
3. What is the total number of outcomes in having to choose any three letters from the English alphabet? What is the total number of outcomes in choosing three letters from the English alphabet if you may only choose each letter once?
4. A lottery requires you to pick five different numbers, each from the set of numbers from 1 to 20. What is the probability of picking the correct number? How does the probability change when you can choose from numbers 1 to 50?
5. The share price today is USD 20. At the end of the week it will go either up by 15% with probability 0.3 or down by 10% with probability 0.7 respectively. At the end of the following week the then price will go either up or down by 10% with probabilities 0.5 and 0.5, respectively. Draw a probability tree. What is the probability that the share price will be USD 19.80 at the end of the second week?



Solutions

- E: 130 observations F: 60 observations
 - $P(E^c) = 0.35$ $P(E \cup E^c) = 1$ $P(E \cap E^c) = 0$
 $E \cup E^c = \Omega$ and $P(\Omega) = 1$ $E \cap E^c = \text{the empty set}$
 - $P(EF) = P(E)P(F) = 0.65 \times 0.3 = 0.195$
- Let F be the event staff member is female. Let M be the event staff member is male. Let D be the event staff member who is disabled.
 $P(F) = 86/120 = 0.7167$ $P(M) = 0.2833$ $P(D) = 0.05$ (given)
 $P(F \text{ and } D) = P(FD) = P(F)P(D)$ (female and disabled are independent)
 $= 0.7176(0.05) = 0.03588$
- Total number of outcomes = $26 \times 26 \times 26 = 17,576$ (each slot can contain any of 26 letters). If each letter can be chosen only once, the number of outcomes is $26 \times 25 \times 24 = 15,600$ (once a letter has been picked it is no longer available).
- The number of possible choices is: $20 \times 19 \times 18 \times 17 \times 16 = 1,860,480$
The probability of choosing the correct number is $1/1,860,480$ or 0.00000054
If you choose from numbers 1 to 50, the probability is $0.000000004!$



The probability that the share price will be 19.80 at the end of two weeks is:

$$P(\text{price goes down first and then up}) = 0.7 \times 0.5 = 0.35$$

Conditional probability and Bayes' formula

Knowledge of the occurrence of one event can affect the probability of the other event.

Activity 2.10



Activity

Consolidate probability

What will you do?

- Suppose events E and F are independent, with $P(E) = 0.3$ and $P(F) = 0.55$. Calculate:
 - $P(E \cup F)$
 - $P(E|F)$
- You have two managers in your division. What is the conditional probability that both are female (f), given that at least one of them is female?

Hint: The sample space is $\Omega = \{(f, f), (f, m), (m, f), (m, m)\}$. Let E be the event that both are female and F the event that at least one of them is female.

- Your staff are given a competency test in the form of a multiple-choice test. Each question has four answers to choose from. You want to evaluate their performance on a specific question. A person either knows the answer to this question or guesses. Assume that the probability of a person knowing the answer is 0.5 and the probability of them guessing is 0.5. Assume also that the probability that a person who guesses gets the answer correct is $\frac{1}{4}$.

What is the conditional probability that a person really knew the answer to the question, given that she or he answered it correctly?

Hint: Use Bayes' formula, with E being the event that they actually knew the answer and F the event that they gave the correct answer.

Solutions

- $$\begin{aligned}
 P(E \cup F) &= P(E) + P(F) - P(EF) \\
 &= P(E) + P(F) - P(E)P(F) \text{ (independence)} \\
 &= 0.3 + 0.55 - (0.3)(0.55) \\
 &= 0.685
 \end{aligned}$$
 - $$P(E|F) = \frac{P(EF)}{P(F)} = \frac{0.165}{0.55} = 0.3 = P(E)$$



2. $E = \{(f, f)\}$ and $F = \{(f, f), (f, m), (m, f)\}$

$$P(E|F) = \frac{P(EF)}{P(F)} \text{ where } P(E) = \frac{1}{4} \text{ and } P(F) = \frac{3}{4}$$

Now $P(EF) = \frac{1}{4}$ since the intersection $E \cap F = \{(f, f)\}$. Therefore

$$P(E|F) = \frac{0.25}{0.75} = \frac{1}{3}$$

3. F is the event that a person gives the correct answer to the question. E is the event that the person actually knows the answer to the question and E^c is the mutually exclusive event that they guess. We know $P(E) = 0.5$, $P(E^c) = 0.5$, $P(F|E^c) = 0.25$ and so $P(F|E) = 0.75$

We must find $P(E|F)$.

Bayes' formula states that:

$$\begin{aligned} P(E|F) &= \frac{P(F|E)P(E)}{P(F|E)P(E) + P(F|E^c)P(E^c)} \\ &= \frac{0.75(0.5)}{0.75(0.5) + 0.25(0.5)} = 0.75 \end{aligned}$$

Unit 8

Probability distributions and applications

Upon completion of this unit the student will be able to:



Outcomes

- **identify** discrete random variables and their probability mass functions;
- **apply** the binomial distribution and Poisson distribution;
- **identify** continuous random variables;
- **understand** the normal and uniform distributions and their density functions; and
- **apply** the normal distribution in market research by **calculating** confidence intervals and **applying** hypothesis tests.

Probability distributions and applications

Activity 2.14



Activity

Revisit probability

What will you do?

1. A machine produces items with a 10 per cent probability of being defective. What is the probability that in a sample of three items, at most one will be defective?
2. The number of accidents occurring on a road each week has a Poisson distribution with mean 6. Determine the probability that there will be 10 accidents the following week.
3. A continuous random variable X has probability density function $f(x) = 0.5$ for $0 < x < 2$, and $f(x) = 0$ for all other x .

Calculate \bar{X} and $\sigma^2(X)$.

4. Returns on assets are assumed to be normally distributed, with mean value 20 per cent and standard deviation $\sigma = 40$ per cent.

What is the probability that the return will be:

- a) greater than 20 per cent?
- b) between -20 per cent and 60 per cent?

Solutions

1. X is the number of defective items per sample of three. Use the binomial distribution with $p = 0.10$ and $N = 3$. The probability that at



most one item is defective is: $P(X = 0 \text{ or } X = 1) = P(X = 0) + P(X = 1)$
 $= p(0) + p(1)$

$$\text{Now } p(0) = \binom{3}{0} (0.1)^0 (1 - 0.1)^{3-0} = 1 \times 1 \times (0.9)^3 = 0.729$$

$$p(1) = \binom{3}{1} (0.1)^1 (1 - 0.1)^{3-1} = 3 \times (0.1) \times (0.9)^2 = 0.243$$

Therefore $P(X = 0 \text{ or } X = 1) = 0.972 = 97.2\%$

Additional questions: Calculate the mean number of defective items in a sample of 20. Answer: $20 \times (0.1) = 2$

2. X is the number of accidents per week.

$$p(10) = P(X = 10) = \frac{e^{-6} 6^{10}}{10(9)(8)\dots 1} = 0.0413$$

The standard deviation is $\sqrt{6} = 2.45 = 2$ accidents per week (rounded down).

3. Expected value of $X = E[X] = \bar{X} = \int_0^2 0.5x dx = 0.25x^2 \Big|_{x=0}^{x=2} = 1$

$$\begin{aligned} \text{Variance of } X = \sigma_X^2 &= \overline{X^2} - (\bar{X})^2 \\ &= \int x^2 (0.5) dx \Big|_0^2 - 1 \\ &= \frac{1}{6} x^3 \Big|_{x=0}^{x=2} - 1 = 0.33 \end{aligned}$$

4. Let X denote return (assume all values are annualised).

a) $P(X > 0.20) = P(X > \mu) = 0.5 = 50\%$

b) $P(-0.20 < X < 0.60) = P(\mu - \sigma < X < \mu + \sigma) = 0.68 = 68\%$

Application of the normal distribution to market research

Activity 2.15



Activity

Calculate confidence interval and average

What will you do?

Samples are taken annually to calculate the average number of hours worked per month by agricultural labourers, and their average monthly wages. Standard deviations (std dev) are also calculated.

The data for years 2005–2008 are shown in this table:

	2005	2006	2007	2008
Average monthly hours worked (and std dev)	178 (10)	162 (14)	148 (7.2)	140 (12.1)
Average monthly wages in pounds (and std dev)	370.00 (10.56)	450.20 (5.53)	500.00 (12.54)	480.60 (14.32)
Sample size n (number of workers)	200	200	138	112

1. Calculate the 95 per cent confidence interval for each year. Discuss the meaning of the values.
2. Determine the average wages per hour for each year.

Solution

1. The 95% confidence interval is: $(\bar{x} - 1.96s/\sqrt{n}, \bar{x} + 1.96s/\sqrt{n})$

Year 2005:

Confidence interval for hours worked

$$(178 - 1.96(10)/\sqrt{200}, 178 + 1.96(10)/\sqrt{200}) = (176.61, 179.39)$$

The average number of hours worked per month in the population of all agricultural labourers lies in this interval with 95% certainty.

Monthly wages interval

$$(370 - 1.96(10.56)/\sqrt{200}, 370 + 1.96(10.56)/\sqrt{200}) = (368.54, 371.46)$$

The average monthly wage in the population of all agricultural labourers lies in this interval with 95% certainty.

This calculation is repeated for the other years.

2. Average monthly wages per hour for 2005: $370/178 = 2.08$ pounds per hour for the sample.



Case study:
Market research

Attempt this case study in market research and hypothesis testing before doing the assessment for this module.

Companies are awarded a special rating for their social commitment, which gives them certain tax benefits. The rating system is based on assessments by the companies themselves as to how much they spend on social projects, improving the living conditions of their workers, combating pollution, and so on. They send an annual report to the Internal Revenue Service (IRS).

The IRS awards the ratings, but it cannot check each company's self-assessment every year. It was initially thought that at least 90 per cent of the self-assessments would be correct and fair. However, the possibility exists of mistakes or outright fraud.

After two years, the IRS has thoroughly checked a sample of the reports of 2,100 companies. It found that 12.01 per cent of reports were not quite correct.

Discuss and test the initial view (hypothesis) of the IRS that at least 90 per cent of reports would be correct.

Solution

Step 1: The null hypothesis H_0 is that at least 90% of reports will be correct. Let us denote the population percentage by π .

$$H_0: \pi \geq 90\%$$

The alternative hypothesis H_1 is that less than 90% of reports will be correct.

$$H_1: \pi < 90\%$$

We must choose between these two hypotheses. We either reject H_0 and accept H_1 or we do not reject H_0 . It is important to note that not rejecting H_0 does not mean we accept H_0 . There is merely not enough evidence to reject it.

Step 2: The actual, but unknown, situation is that H_0 is either true or false. If H_0 is true we may, based on the sample data, reject H_0 . In this case, we will have made an error (Type I). In our example, we want the probability of a Type I error to be as small as possible and choose $\alpha = 0.01$.

Step 3: Determine the critical Z values. $Z_\alpha = 2.33$ for $\alpha = 0.01$

and $Z_{calc} (= \frac{\bar{x} - \mu}{s / \sqrt{n}})$ is calculated from the sampling distribution. The

population mean is $\pi = 90\%$. The observed mean value was 87.99%. The standard deviation s/\sqrt{n} must be adapted for percentages. It becomes:

$$\text{becomes: } \sqrt{\frac{\pi(100 - \pi)}{n}} = \sqrt{\frac{90(100 - 90)}{2100}} = 0.655$$

$$\text{and so } Z_{calc} = \frac{87.99 - 90}{0.655} = -3.07 \text{ or } |Z_{calc}| = 3.07$$

We see that $|Z_{calc}| > |Z_{\alpha}|$ ($3.07 > 2.33$)

Step 4: Since $|Z_{calc}| > |Z_{\alpha}|$ we reject H_0 . The sample result of 87.99% is 3.07 standard deviations away from the hypothesised population mean of 90%. This is outside the region determined by 2.33 standard deviations. We accept the alternative hypothesis H_1 .

Assessment



Assessment — Module 2

- Find the mean, median and mode of this set of data. Represent the data in a scatter graph:
15, 6, 14, 7, 12, 11, 7, 10, 8, 13
- Study the frequency table for the consumption of diesel by a group of trucks. (For example, there are six trucks using between 25 and 30 litres of diesel per 100 km.)

Class: Litres per 100 km	Frequency: Number of trucks
20–25	2
25–30	6
30–35	12
35–40	8
40–45	5

- Expand the table to include cumulative and relative frequencies.
 - Find the mean and median for the distribution.
 - Draw a histogram and ogive.
 - Use the ogive to confirm the value of the median.
 - Is the distribution positively or negatively skewed?
 - Calculate the standard deviation for the data.
- Sales figures S_1 and S_2 (income in a local currency) for a week for two branches of a coffee shop are:

Day	Sales for shop 1	Sales for shop 2
1	1,200	1,300
2	2,300	2,450
3	2,400	3,500
4	1,800	2,000
5	1,100	1,500
6	2,000	2,200
7	2,200	1,800

- Determine the mean and standard deviation of sales for each shop.
- Calculate the co-variance and correlation co-efficient for sales figures S_1 and S_2 .
- Present the data in column or scatter graphs.
- Interpret and discuss your results for 3a) – 3c).

4. What is the probability of getting a sum greater than 10 when throwing two dice?
5. Let E and F be mutually exclusive and independent events. What can you say about the probabilities $P(E \cup F)$ and $P(E \cap F)$?
6. A fair coin is tossed three times. What is the probability of throwing two tails and one head? Use a probability tree for your calculations.
7. The conditional probability $P(E|F)$ is 0.4 and the probability $P(E \cap F)$ is 0.2. What is the value of $P(F)$?
8. Random variable X is binomially distributed. There are 20 trials and the probability of success is 0.15. What is the probability of 10 successes?
9. Random variable Y has the Poisson distribution with mean value 4.2.
 - a) What is the variance of the distribution?
 - b) What is the probability of 14 successes?
10. Random variable X is normally distributed, with mean value 30 and a variance of 14.
 - a) What is the probability that X will take on values greater than 44? (Hint: No tables are necessary.)
 - b) What is the probability that X will take on values less than 20? (Hint: Transform X to a standard normal variable Z and use tables.)
11. A company wants to find out how long it takes a worker to assemble a certain component of a machine. A sample of 40 workers shows an average time of 76.4 seconds, and a standard deviation of 17.2 seconds. Find the 95 per cent confidence interval for the population mean.
12. A sample of 500 patients is given a new treatment for AIDS. Testing the patients after one year shows that 67 per cent of these patients have responded very well, with increased CD4- or T-cell counts. Determine a 95 per cent confidence interval for the percentage of patients in the population who will have increased CD4 counts after the treatment.



Solutions

1. Mean: 10.3 Median: 10.5 Mode: 7
 2. a)

Class: Litres per 100 km	Frequency f : Number of trucks	Cumulative frequency F	Relative frequency
20–25	2	2	0.061
25–30	6	8	0.182
30–35	12	20	0.364
35–40	8	28	0.242
40–45	5	33	0.152

- b) Mean: Midpoints x_i of classes: 22.5, 27.5, 32.5, 37.5, 42.5

$$\begin{aligned} \text{Mean} &= \frac{1}{N} \sum_{i=1}^N x_i f(x_i) \\ &= \frac{1}{33} [22.5(2) + 27.5(6) + 32.5(12) + 37.5(8) + 42.5(5)] \\ &= 33.71 \text{ (rounded) litres per 100 km} \end{aligned}$$

The middle value is observation 17, which falls in class 30–35. Since $17 - 8 = 9$, the middle value takes position 9 of the 12 values in class 30–35. Now $30 + (9/12)5 = 33.75$ and this is the median.

- c) and d) are straightforward. See Activity 2.5.

- e) Skewness is proportional to the difference (mean – median). Skewness is negative here: more than half the values lie to the right of the mean value. The distribution is not symmetrical.

$$\text{f) Variance } (x) = s^2 = \frac{\sum_{i=1}^N (x_i - \bar{x})^2 f(x_i)}{\sum_{i=1}^N f(x_i)} \text{ for grouped data values}$$

with midpoint values x_i of classes and frequencies $f(x_i)$.

We have $\bar{x} = 33.71$

$$s^2 = 29.039 \text{ and } s = 5.389$$

3. a) Shop 1: Mean daily sales income $\overline{S1} = 1,857.14$
 Std dev. = 483.60 (rounded)

$$\text{Shop 2: Mean daily sales income } \overline{S2} = 2,107.14$$

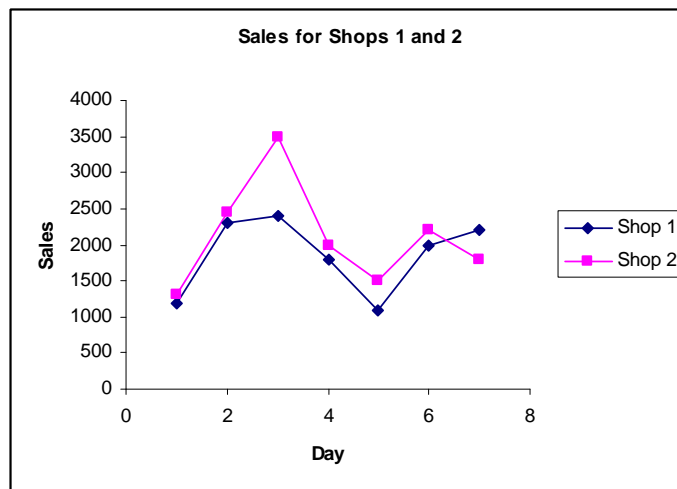
Std dev = 675.29 (rounded)

$$\text{b) Co-variance} = \frac{\sum_{i=1}^N (S1_i - \bar{S1})(S2_i - \bar{S2})}{7}$$

$$= 263,265.27$$

$$\text{Correlation co-efficient} = \frac{263265.27}{(483.6)(675.29)} = 0.806$$

c) We show a scatter graph.



d) Shop 2 has a greater average income, but also a greater standard deviation and therefore uncertainty in income. The two incomes from the shops are positively correlated and close to 1: they tend to rise and fall together.

4. There are 36 pairs of outcomes from throwing two dice. Let the sum be random variable S .

$$P(S > 10) = P(S = 11) + P(S = 12) = (2/36) + (2/36) = 1/9$$

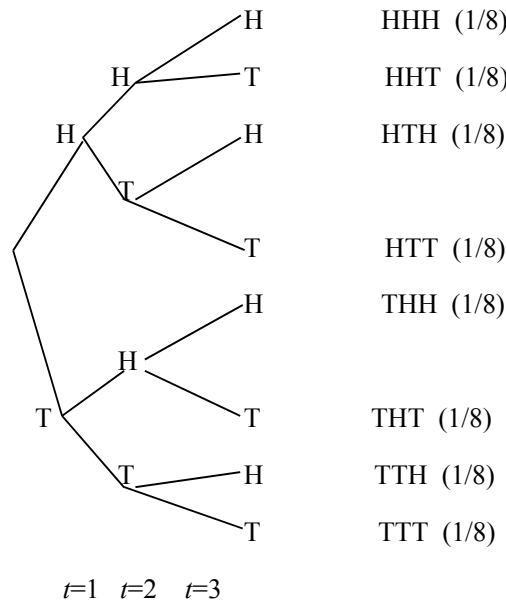
5. $P(E \text{ or } F) = P(E \cup F) = P(E) + P(F)$

$$P(E \text{ and } F) = P(E \cap F) = P(E) P(F) = 0.$$

This means at least one of $P(E)$ or $P(F)$ equals zero, so that at least one of E or F is an impossible event.



6. At the end of each branch we write the string of outcomes and probability. Outcomes with two tails are highlighted:



There are three independent outcomes with two tails and the probability is $3/8$.

7.
$$P(E | F) = \frac{P(EF)}{P(F)}$$

Therefore
$$P(F) = \frac{P(EF)}{P(E|F)} = \frac{0.2}{0.4} = 0.5$$

8.
$$P(X = 10) = \binom{20}{10} 0.15^{10} (1 - 0.15)^{20-10}$$

$$= 0.00021$$

9. a) Variance of $X = \sigma^2(X) = Np = \mu = 4.2$

b)
$$P(X = 14) = \frac{e^{-4.2} (4.2)^{14}}{14(13)(12)...1} = 0.00009$$

10. a)
$$P(X > 44) = P(X > \mu + \sigma) = \frac{1}{2}(1 - P(\mu - \sigma \leq X \leq \mu + \sigma))$$

$$= \frac{1}{2}(1 - 0.68) \text{ (from symmetry)}$$

$$= 0.16$$

b)
$$P(X < 20) = P\left(\frac{X - \mu}{\sigma} < \frac{20 - \mu}{\sigma}\right)$$

$$= P(Z < -2.67) = P(Z > 2.67) \text{ (symmetry)}$$

$$= 0.5 - P(0 < Z < 2.67) \text{ (see the Appendix)}$$

$$= 0.5 - 0.496 = 0.004$$

11. A 95% confidence interval around sample mean \bar{x} is:
 $(\bar{x} - 1.96s/\sqrt{n}, \bar{x} + 1.96s/\sqrt{n})$

$$\begin{aligned} &= (\bar{x} - 1.96s/\sqrt{n}, \bar{x} + 1.96s/\sqrt{n}) \\ &= (71.07, 81.17) \end{aligned}$$

12. A 95% confidence interval around sample mean \bar{x} is:

$$\begin{aligned} &(\bar{x} - 1.96\sqrt{\frac{\pi(1-\pi)}{n}}, \bar{x} + 1.96\sqrt{\frac{\pi(1-\pi)}{n}}) \\ &= (0.67 - 1.96(0.00054), 0.67 + 1.96(0.00054)) \\ &= (0.669, 0.6711) \end{aligned}$$