

MATHEMATICS

A SHORT COURSE IN STATISTICS

FREQUENCY DISTRIBUTIONS

HISTOGRAMS.

It is the area of a bar which represents the frequency. Thus if an interval is wider than "standard" the height must be reduced and vice versa. (See example.)

NB. The horizontal scale must be linear.

CUMULATIVE FREQUENCY GRAPHS

Points must be plotted against upper boundaries of intervals.

Median = $\frac{n+1}{2}$ th reading. (Unless % cumulative frequencies used in which case median = 50th percentile)

Lower quartile, $Q_1 = \frac{n+1}{4}$ th reading.

Upper quartile, $Q_3 = \frac{3(n+1)}{4}$ th reading.

Semi interquartile range = $\frac{Q_3 - Q_1}{2}$

MEAN AND VARIANCE

Sample mean, $\bar{x} = \frac{\sum x}{n}$, or , for a frequency distribution $\bar{x} = \frac{\sum fx}{\sum f}$

Sample Variance, $s^2 = \frac{\sum (x - \bar{x})^2}{n}$, or , $s^2 = \frac{\sum f(x - \bar{x})^2}{\sum f}$.

For ease of calculation: $s^2 = \frac{\sum x^2}{n} - \bar{x}^2$ or $s^2 = \frac{\sum f(x)^2}{\sum f} - \bar{x}^2$

Standard deviation $s = \sqrt{\text{variance}}$.

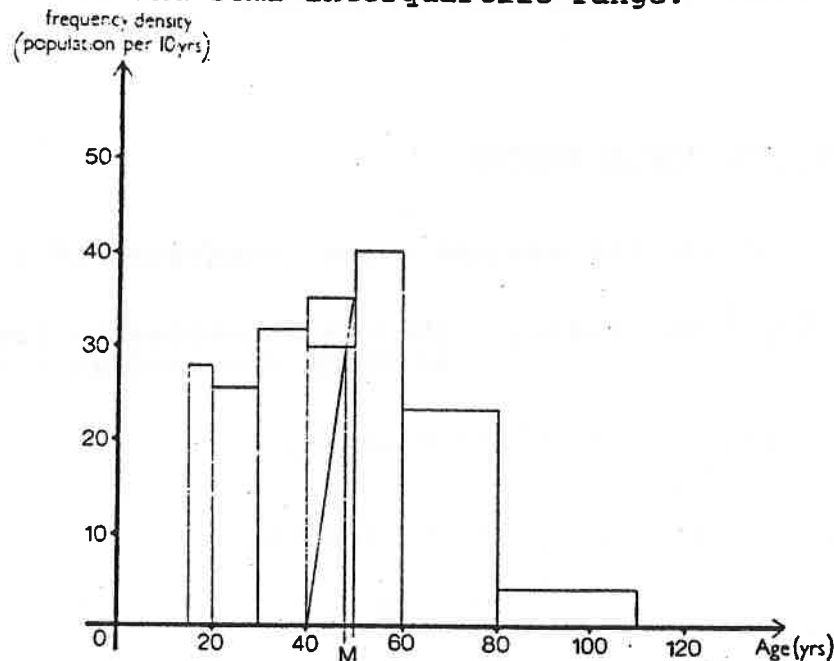
REVISION IN STATISTICS : Chapter 1

EXAMPLE 1

The table shows the ages in completed years of the adult population

AGE	15-19	20-29	30-39	40-49	50-59	60-79	80-109
No. OF PEOPLE	14	26	32	35	40	47	10

Plot a histogram to illustrate the data and comment on the shape.
Calculate the median and semi interquartile range.



SOLUTION

The histogram shows two features : -

1. Lower frequencies in the older age ranges caused by deaths reducing the population,
2. Lower frequencies in the younger age range suggesting that the village population is likely to be declining, possibly as people leave the village to find employment elsewhere.

There is a total frequency of 204, therefore the median has 102 on each side. Therefore it lies in the 40-49 age group.

There are 72 people to the left of this group, so we require a further 30.

$$\begin{aligned}
 \text{Hence Median} &= 40 + \frac{30}{35} \times 10 \\
 &= 48.6 \text{ years} \quad (\text{or by similar triangles, see diagram}) \\
 &= \underline{48 \text{ years } 7 \text{ months.}}
 \end{aligned}$$

There are 51 people to the left of the lower quartile.

$$\text{Therefore } Q_1 = 30 + \frac{11}{32} \times 10$$

$$= 33.4 \text{ years}$$

$$= \underline{33 \text{ years } 5 \text{ months.}}$$

REVISION IN STATISTICS : Chapter 1

There are 153 people to the left of the upper quartile .

$$\text{Therefore } Q_3 = 60 + \frac{6}{47} \times 20$$

$$= 62.55 \text{ years}$$

$$= \underline{62 \text{ years } 7 \text{ months}}.$$

$$\text{Semi interquartile range} = \frac{\underline{62 \text{ years } 7 \text{ months}} - \underline{33 \text{ years } 5 \text{ months}}}{2}$$

$$= \underline{14 \text{ years } 7 \text{ months}}$$

REVISION IN STATISTICS : Chapter 1

EXAMPLE 2

Draw up a cumulative frequency table for the data of question 1 and hence draw a cumulative frequency graph.

Obtain the median and semi interquartile range from this graph. What proportion of this population would you estimate to be senior citizens (65 and over)?

If 73% of the population are too old to qualify for the local football team, what is the qualifying age?

SOLUTION

AGE	CUMULATIVE FREQUENCY
15	0
20	14
30	40
40	72
50	107
60	147
80	194
110	204

Median = 48.5 years
= 48 years 6 months.

Q_1 = 33.5 years
= 33 years 6 months.

Q_3 = 62 years.

Semi interquartile range = $\frac{62 - 33.5}{2}$

= 14 years 3 months.

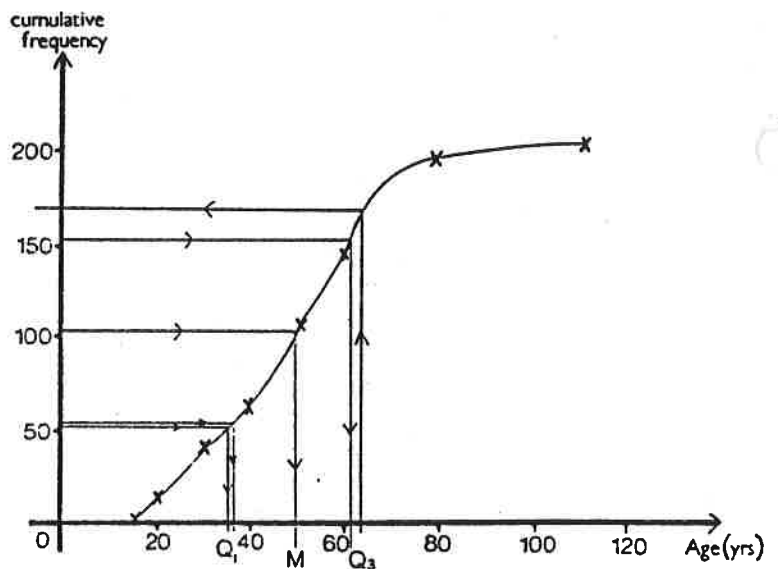
There are 167 people under 65. Therefore there are 37 senior citizens.

Percentage of senior citizens = $\frac{37}{204} \times 100$

= 18.1%

73% do not qualify, so 27% of 204 do, i.e. 55 people.

Therefore those qualifying are under 35 years.



REVISION IN STATISTICS : Chapter 1

EXAMPLE 3

The table shows the temperatures recorded, to the nearest degree, at a holiday resort for the fifteen weeks holiday period.

TEMPERATURE °C	16	17	18	19	20	21	22	23	24	25	26	27	28
NO. OF DAYS	2	0	7	7	12	15	18	18	12	6	4	3	1

Calculate the mean and standard deviation for this sample.

SOLUTION

x	f	y=x-22	fy	fy ²
16	2	-6	-12	72
17	0	-5	0	0
18	7	-4	-28	112
19	7	-3	-21	63
20	12	-2	-24	48
21	15	-1	-15	15
22	18	0	0	0
23	18	1	18	18
24	12	2	24	48
25	6	3	18	54
26	4	4	16	64
27	3	5	15	75
28	1	6	6	36
	<u>105</u>		<u>-3</u>	<u>605</u>

$$\bar{y} = \frac{-3}{105}$$

$$= -0.029$$

$$\bar{x} = \bar{y} + 22$$

$$= \underline{21.971}$$

$$s_x = s_y$$

$$= \sqrt{\left\{ \frac{605}{105} - (-0.029)^2 \right\}}$$

$$= \underline{2.4}$$

REVISION IN STATISTICS : Chapter 1

EXAMPLE 4

The marks in a GCSE Mathematics examination for a sample of 80 pupils were as follows: -

MARK	1-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80	81-90	91-100
FREQUENCY	0	3	15	25	20	10	5	0	2	0

Calculate the mean and standard deviation for this sample.

SOLUTION

MARKS	f	mid-point x	$y = \frac{x - 35.5}{10}$	f_y	f_y^2
1-10	0	5.5	-3	0	0
11-20	3	15.5	-2	-6	12
21-30	15	25.5	-1	-15	15
31-40	25	35.5	0	0	0
41-50	20	45.5	1	20	20
51-60	10	55.5	2	20	40
61-70	5	65.5	3	15	45
71-80	0	75.5	4	0	0
81-90	2	85.5	5	10	50
91-100	0	95.5	6	0	0
	<u>80</u>			<u>44</u>	<u>182</u>

$$\bar{y} = \frac{44}{80}$$

$$= 0.55$$

$$\bar{x} = 10\bar{y} + 35.5$$

$$= \underline{41}$$

$$s_y = \sqrt{\left(\frac{182}{80} - 0.55^2\right)}$$

$$= \underline{1.404}$$

$$s_x = 10s_y$$

$$= \underline{14.04}$$

REVISION IN STATISTICS : Chapter 1

EXERCISES

1. The table shows the times for the 400 metres, to the nearest second, recorded over a season by a member of an athletics club.

TIME	50	51	52	53	54	55	56	57	58	59	60
FREQUENCY	2	4	9	12	7	6	8	11	5	0	2

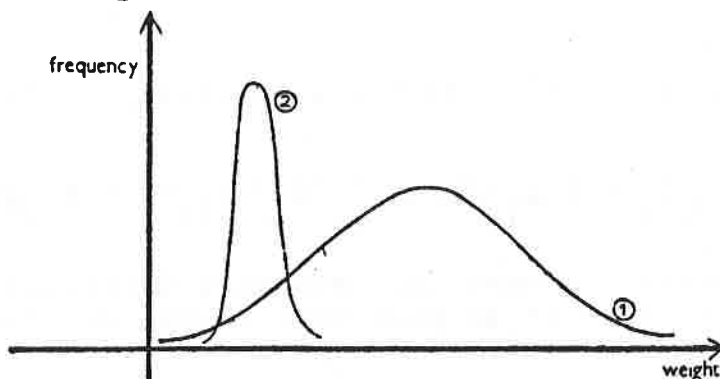
Draw a histogram to illustrate the data and suggest a reason for the bimodal nature of the distribution.

2. The graph shows the distribution of weights of a sample of dogs and a sample of cats.

- (a) Which graph represents dogs and which cats? Explain your answer with reference to:-

- (i) The modal weights
(ii) the dispersions of the weights.

- (b) Which was the larger sample, cats or dogs? Give a reason for your answer.



3. A Basic Numeracy test was given to two first form mathematics sets in two different secondary schools, (A and B). The results were as shown.

MARK	0	1	2	3	4	5	6	7	8	9	10
FREQUENCY (A)	0	0	2	3	5	7	6	3	2	1	1
FREQUENCY (B)	0	0	0	0	0	3	5	9	10	2	1

Draw two frequency polygons on the same pair of axes. Calculate the mean and standard deviation for each group. Comment on the different achievement levels of the two groups and suggest a possible reason for this.

4. The table shows the lengths of pea-pods to the nearest 10mm.

LENGTH (mm)	30	40	50	60	70	80	90	100	110
NO. OF PODS	2	0	4	16	38	26	10	3	1

Represent the data by a histogram. Calculate the mean and standard deviation for the sample.

REVISION IN STATISTICS : Chapter 1

5. The table shows the heights, to the nearest 5cm. of the police constables in a particular force.

HEIGHT	175	180	185	190	195	200	205	210
FREQUENCY	7	15	32	40	44	37	13	2

Draw a cumulative frequency graph to represent the data and find: -

- the median
- the semi interquartile range
- What proportion of the officers were under 6 feet tall?
(1ft. = 30.5cm.)
- If the tallest 20% of them are required for special duties what will be the qualifying height?

6. The table shows the ages, in completed years, of the residents in a holiday camp during a week in summer.

AGE	0-4	5-9	10-19	20-29	30-39	40-49	50-69	70-
FREQUENCY	165	205	260	320	350	242	168	52

Draw a histogram to illustrate the data.

7. The marks out of 80 for a chemistry paper were as follows:-

MARK	0-10	11-20	21-30	31-40	41-50	51-60	61-80
FREQUENCY	8	15	20	13	7	5	2

Calculate the mean and standard deviation for these marks. The marks were so poor that it was decided to increase them to be out of 100.

Two schemes were suggested:-

- to add 20 to each mark,
- Multiply each mark by 1.25

State the new mean and standard deviation under each scheme.

8. In a 6th form Statistics experiment, 300 3rd formers were asked to estimate the length of a wooden rod to the nearest mm. The table shows the cumulative frequency table obtained.

LENGTH	700-800	≤900	≤1000	≤1100	≤1200	≤1300	≤1400
No. of PUPILS	4	15	86	230	286	298	300

Represent the data graphically and estimate the median length.

Obtain the original frequency distribution for the data and estimate the sample mean and standard deviation.

PROBABILITY

ADDITION RULE

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

or, using sets, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

If A and B are mutually exclusive events then $P(A \text{ and } B) = 0$, i.e. $A \cap B$ is empty, then the addition rule becomes

$$P(A \text{ or } B) = P(A) + P(B)$$

MULTIPLICATION RULE

$$P(A \text{ and } B) = P(A).P(B/A)$$

or $P(A \cap B) = P(A).P(B/A)$

If A and B are independent events then $P(B/A) = P(B)$ and the multiplication rule becomes

$$P(A \text{ and } B) = P(A).P(B)$$

or $P(A \cap B) = P(A).P(B)$

REVISION IN STATISTICS : Chapter 2

EXAMPLE 1

A bag contains 3 yellow, 4 green and 5 red marbles. Three are drawn at random (i.e. without replacement).

Calculate the probability that

- (a) all are green
- (b) all are the same colour
- (c) two are red and one is green
- (d) they are all different colours
- (e) at least two are green.

SOLUTION

- (a) $P(G \text{ and } G/G \text{ and } G/G/G) = 4/12 \times 3/11 \times 2/10$
 $= \underline{1/55}.$
- (b) $P(GGG \text{ or } RRR \text{ or } YYY) = 1/55 + 5/12 \times 4/11 \times 3/10 + 3/12$
 $\times 2/11 \times 1/10$
 $= \underline{3/44}.$
- (c) $P(RRG \text{ or } RGR \text{ or } GRR) = 5/12 \times 4/11 \times 4/10 \times 3 \text{ (order)}$
 $= \underline{2/11}.$
- (d) $P(\text{all different}) = P(RGY) \times 3! \text{ (order)}$
 $= 5/12 \times 4/11 \times 3/10 \times 6$
 $= \underline{3/11}.$
- (e) $P(\text{at least two green}) = P(2 \text{ green and } 1 \text{ not or } 3 \text{ green})$
 $= 4/12 \times 3/11 \times 8/10 \times 3 + 1/55$
 $= \underline{13/55}.$

EXAMPLE 2

Two events A and B are such that $P(A) = 1/3$, $P(B) = 1/4$,
 $P(A \cap B) = 1/24$.

Evaluate

- (a) $P(A/B)$
- (b) $P(A \cup B)$
- (c) $P(A' \cap B')$.

Another event C is such that A and C are mutually exclusive and B and C are independent.

If $P(C) = 1/2$, calculate $P(A \cup B \cup C)$.

SOLUTION

- (a) By the multiplication rule $P(A \cap B) = P(B).P(A/B)$ therefore

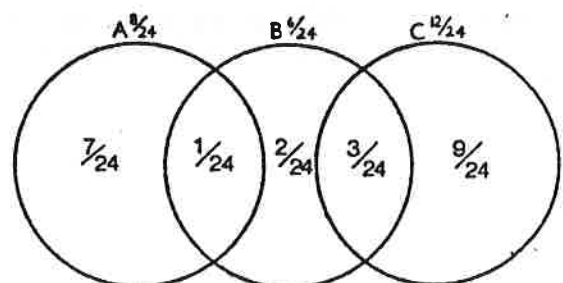
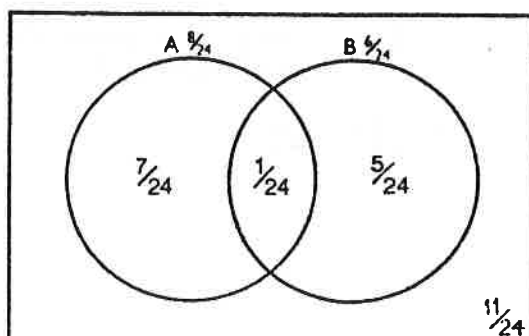
$$\begin{aligned} P(A/B) &= \frac{P(A \cap B)}{P(B)} \\ &= \frac{1/24}{1/4} \\ &= \underline{1/6} \end{aligned}$$

- (b) We now have $P(B \cap A') = 5/24$
 and therefore $P(A \cup B) = 1/3 + 5/24$
 $= \underline{13/24}$

- (c) $P(A' \cap B') = 1 - P(A \cup B)$
 $= \underline{11/24}$

A and C being mutually exclusive means that $A \cap C = \emptyset$, see diagram.

Since B and C are independent $P(B \cap C) = P(B).P(C) = 1/8$.
 $P(A \cup B \cup C) = P(A) + P(B \cap A') + P(C \cap B') = 22/24$
 or $\underline{11/12}$, see diagram.



REVISION IN STATISTICS : Chapter 2

EXAMPLE 3

The 28 children in a class are selected at random to play either soccer (10 pupils), badminton (4 pupils) or netball (14 pupils).

Three of the pupils are called Anne, Bert and Charles.

Calculate the probabilities that

- (a) Anne plays netball,
- (b) Bert plays netball,
- (c) both Anne and Bert play netball,
- (d) all three play netball,
- (e) they all play the same sport,
- (f) they all play different sports,
- (g) two of them play the same sport and the third a different one.

SOLUTION

- (a) $P(\text{Anne plays netball}) = 14/28$
 $= \underline{1/2}$
- (b) $P(\text{Bert plays netball}) = \underline{1/2}$
- (c) $P(\text{A and B both play netball}) = 14/28 \times 13/27$
 $= \underline{13/54}$
- (d) $P(\text{all three play netball}) = 14/28 \times 13/27 \times 12/26$
 $= \underline{1/9}$
- (e) $P(\text{all the same}) = 1/9 + 4/28 \times 3/27 \times 2/26 + 10/28 \times 9/27 \times 8/26$
 $= \underline{122/819}$
- (f) $P(\text{all different}) = P(\text{A soccer, B badminton, C netball}) \cdot 3!$
 $= 10/28 \times 4/27 \times 14/26 \times 6$
 $= \underline{20/117}$
- (g) $P(\text{2 play one sport, the third another}) = 1 - P(\text{all the same or all different})$
 $= \underline{557/819}$

EXAMPLE 4

Two players, A and B, take turns to throw a fair tetrahedral die numbered 1, 2, 3, 4. The player who first scores a 4 on the base wins.

Calculate the probability that A wins and the probability that B wins.

SOLUTION

$P(A \text{ loses}) \times P(B \text{ wins})$.

	A	B
1st round	$1/4$	$3/4 \times 1/4$
2nd round	$(3/4)^2 \cdot 1/4$	$(3/4)^3 \cdot 1/4$
3rd round	$(3/4)^4 \cdot 1/4$	$(3/4)^5 \cdot 1/4$

$P(A \text{ loses}) P(B \text{ loses}) \times P(A \text{ wins})$ etc.

$$P(A \text{ wins}) = 1/4 + (3/4)^2 \cdot 1/4 + (3/4)^4 \cdot 1/4 + \dots \text{ (G.P. } a = 1/4, r = (3/4)^2 \text{)}$$

$$= \frac{1/4}{1 - (3/4)^2}$$

$$= \frac{4}{7}$$

$$P(B \text{ wins}) = \frac{3}{7}$$

EXERCISES

1. A box contains 9 discs of which 4 are red, 3 are white and 2 are blue. Three discs are to be drawn at random, without replacement.

Calculate the probabilities that: -

- (a) the discs, in order, will be coloured red, white and blue respectively
- (b) one disc of each colour will be drawn
- (c) the third disc will be red
- (d) no red disc will be drawn.

P(A)

P(B)

n P(A)
beco.

2. The probability that a hen lays an egg on a given day is 0.7 if it laid the previous day and 0.9 if it did not.

Calculate the probability that a hen lays on Tuesday given that it did lay on the previous Sunday.

3. In a class of 30 pupils, 5 have red hair and 12 study Russian. If a quarter of those who study Russian have red hair, find the probabilities that: -

- (a) a member of the class has red hair and studies Russian,
- (b) a child who studies Russian does not have red hair,
- (c) a member of the class neither studies Russian nor has red hair.

B/.

4. Broad bean seeds are planted in pairs because, on average only 70% of them germinate. A bed is planted with 100 pairs of seeds.

Calculate the expected number of pairs in which: -

- (a) both seeds germinate,
- (b) only one seed germinates,
- (c) no seeds germinate,
- (d) at least one seed germinates.

5. A driving instructor assesses that three of his pupils have probabilities of $\frac{5}{6}$, $\frac{4}{5}$ and $\frac{3}{4}$ of passing their tests.

Calculate the probabilities that: -

- (a) exactly one pupil passes,
- (b) not more than one passes,
- (c) at least one passes.

REVISION IN STATISTICS : Chapter 2

6. A bag contains 6 black balls and 4 white balls. Balls are drawn at random without replacement.

Calculate the probability that the balls alternate BWBW etc. until only black balls remain.

Repeat to obtain a general formula if there are b black balls and w white ones initially.

7. In a class of 24 pupils, what is the probability of at least one repeated birthday. (Ignore leap years.)

8. A woman has three possible routes to drive to work, A, B and C, which she uses with respective probabilities $1/2$, $1/4$ and $1/4$. Using these routes gives her probabilities of being late of $1/12$, $1/4$ and $1/3$ respectively.

Calculate:-

- (a) the probability that she is late on any particular morning,
 - (b) the probability that she travelled by route B on a morning she was late.
9. Two independent events A and B have associated probabilities of $3p$ and $4p$ of occurring.

If $P(A \cap B) = p$, calculate the value of p and find: -

- (a) $P(A/B)$,
- (b) $P(B'/A)$,
- (c) $P(A' \cap B')$.

A third event C is such that $P(A \cup C) = 5/8$ and $P(A' \cap B \cap C') = 1/8$. If A and C are mutually exclusive, show that B and C are independent.

10. Three fair dice are thrown. Calculate the probabilities that: -
- (a) the total score is 17,
 - (b) the total score is less than 17,
 - (c) none of the individual scores is less than 2,
 - (d) the lowest individual score is 2.

DISCRETE PROBABILITY DISTRIBUTIONS

DEFINITION

X is a discrete random variable if it can take discrete values $x_1, x_2, x_3, \dots, x_n$ with associated probabilities $p_1, p_2, p_3, \dots, p_n$

such that
$$\sum_{i=1}^n p_i = 1$$

MEAN

The theoretical mean value of X, or the EXPECTATION of X, is given by

$$E(X) = \sum x \cdot P(X=x)$$

VARIANCE

$$\begin{aligned} \text{Var}(X) &= \sum x^2 \cdot P(X=x) - \text{mean}^2 \\ &= \sum x^2 P(X=x) - \left[\sum x \cdot P(X=x) \right]^2 \\ &= E(X^2) - [E(X)]^2 \end{aligned}$$

EXPECTATION AND VARIANCE OF OTHER FUNCTIONS

- | | | |
|---|------------------------------------|---|
| 1 | E{f(X)} = $\sum f(x) \cdot P(X=x)$ | where f(X) is any function of X. |
| 2 | E(X + c) = E(X) + c | and Var(X+c) = Var(X)
where c is a constant. |
| 3 | E(cX) = c.E(X) | and Var(cX) = $c^2 \cdot \text{Var}(X)$
where c is a constant. |
| 4 | E(X + Y) = E(X) + E(Y) | and Var(X+Y) = Var(X) + Var(Y)
X, Y independent. |
| 5 | E(X - Y) = E(X) - E(Y) | BUT Var(X-Y) = Var(X) + Var(Y)
X, Y independent. |

6. $E(ax \pm b) = a E(x) \pm b$

7. $\text{Var}(ax \pm b) = a^2 \text{Var}(x)$

REVISION IN STATISTICS : Chapter 3

EXAMPLE 1

For the following probability distribution calculate

(a) $E(X)$, (b) $E(X^2)$, (c) $\text{Var}(X)$.

X	4	6	8	10	12
$P(X)$	$1/8$	$1/6$	$3/8$	$1/4$	$1/12$

SOLUTION

$$(a) \quad E(X) = 4 \times 1/8 + 6 \times 1/6 + 8 \times 3/8 + 10 \times 1/4 + 12 \times 1/12 = \underline{8}$$

$$(b) \quad E(X^2) = 4^2 \times 1/8 + 6^2 \times 1/6 + 8^2 \times 3/8 + 10^2 \times 1/4 + 12^2 \times 1/12 \\ = \underline{69}$$

$$(c) \quad \text{Var}(X) = E(X^2) - E^2(X)$$

$$= 69 - 8^2$$

$$= \underline{5}$$

EXAMPLE 2

The number of eggs X laid by a lesser spotted willow warbler can take the values 0, 1, 2, 3, 4 with probabilities proportional to X . Find these probabilities. Calculate the mean and variance for this probability distribution.

Eggs are collected from three nests. Calculate the probability that only four eggs are collected altogether.

SOLUTION

Since the probabilities are proportional to X and their sum is unity, the probability distribution is as follows: -

X	0	1	2	3	4
$P(X)$	0	1/10	2/10	3/10	4/10

$$E(X) = 1 \times 1/10 + 2 \times 2/10 + 3 \times 3/10 + 4 \times 4/10 = 3$$

$$\text{Var}(X) = 1^2 \times 1/10 + 2^2 \times 2/10 + 3^2 \times 3/10 + 4^2 \times 4/10 - 3^2 = 1$$

Since the probability that a nest contains 0 eggs is 0, the only possibilities giving a total of 4 eggs are 1,1,2; 1,2,1; 2,1,1.

$$\text{Therefore } P(4 \text{ eggs}) = 1/10 \times 1/10 \times 2/10 \times 3 = \underline{3/500}$$

EXAMPLE 3

p is the constant probability that an event will occur in a single trial of an experiment and $q (= 1 - p)$ is the constant probability of failure.

If the experiment is repeated, successive trials being independent, until the first success occurs, derive an expression for the probability function $P(X)$, the probability that the first success occurs on the X^{th} trial.

Show that $\Sigma P(X) = 1$ and show that the mean value of X is $1/p$.

A single card is drawn from a well shuffled pack of 52. What is the probability that it is a spade?

This is repeated, with replacement, until a spade is drawn. How many trials, on average, will be needed until a spade is drawn?

What is the probability that this number of trials is necessary?

SOLUTION

$$P(\text{1st success on } x^{\text{th}} \text{ trial}) = q \times q \times q \dots \times q \times q \times p = q^{x-1} \cdot p$$

$$\begin{aligned} EP(X) &= p + q \cdot p + q^2 \cdot p + q^3 \cdot p + \dots \\ &= p(1 + q + q^2 + q^3 + \dots) \\ &= p(1 - q)^{-1} \\ &= p/p \\ &= \underline{1} \end{aligned}$$

$$\begin{aligned} E(X) &= 1 \times p + 2 \times q \cdot p + 3 \times q^2 \cdot p + 4 \times q^3 \cdot p + \dots \\ &= p(1 + 2q + 3q^2 + 4q^3 + \dots) \\ &= p(1 - q)^{-2} \\ &= p \times 1/p^2 \\ &= \underline{1/p} \end{aligned}$$

$$P(\text{Spade}) = \underline{1/4}$$

So we have the above distribution with $p = 1/4$ and the mean number of trials is 4.

And the probability of 4 trials being needed is $(3/4)^3 \times 1/4$
i.e. 27/256

EXAMPLE 4

Two tetrahedral dice are thrown together and the total score, X is recorded. Draw up a table for the probability distribution for X .

Calculate the mean and variance of the following functions: -

- (a) X , (b) $2X$, (c) $2X + 4$.

SOLUTION

X	2	3	4	5	6	7	8
$P(X)$	1/16	1/8	3/16	1/4	3/16	1/8	1/16

$$E(X) = \underline{5} \text{ by symmetry.}$$

$$\text{Var}(X) = \underline{2.5} \text{ see previous examples.}$$

$$E(2X) = 2 \times E(X) = \underline{10}$$

$$\text{Var}(2X) = 2^2 \times \text{Var}(X) = \underline{10}$$

$$E(2X + 4) = E(2X) + 4 = \underline{14}$$

$$\text{Var}(2X + 4) = \text{Var}(2X) = \underline{10}$$

EXERCISES

1. The number of eggs laid by giant tortoises have the following distribution: -

no. of eggs X	1	2	3	4	5
P(X)	1/10	2/10	3/10	3/10	1/10

Calculate (a) $E(X)$, (b) $E(X^2)$ (c) $\text{Var}(X)$.

2. The random variable X has a Binomial distribution with $n = 16$ and $p = 1/4$ and the random variable Y has a Poisson distribution with mean 1.

Find the mean and standard deviation of

(a) $X - Y$ (b) $X + 3$ (c) $3X + 4Y$.

(Assume X and Y are independent. See later for Binomial and Poisson distributions.)

3. The random variable X can take values 1, 2, 3, 4, 5 with equal probabilities (rectangular distribution). Find $E(X)$ and $\text{Var}(X)$.

Repeat to obtain formulae for the mean and variance of the rectangular distribution when X can take values 1, 2, 3, ... n.

4. A fair six sided die has its faces numbered as follows; one 1, two 2's, three 3's. If it is thrown once, calculate the mean and variance of the score. If X is the total score when the die is thrown twice, draw up a table to show the probability distribution for X and calculate the mean and variance of X.

5. A tetrahedral die is such that the face on the base, X, can take values 1, 2, 3, 4. If $P(X = 4)$ is twice as likely as each of the other scores, find $E(1/X)$.

REVISION IN STATISTICS : Chapter 3

6. X is the number of heads when three fair coins are thrown together and Y is the number on the base of a fair tetrahedral die with faces 1, 2, 3, 4. Draw up tables to show the probability distributions of X and Y. Calculate
 (a) $P(X > Y)$, (b) $P(Y > X)$.

7. A fair die is thrown repeatedly until a six appears. If X denotes the number of throws before the first six appears, obtain the probability distribution for X and find $E(X)$ and $\text{Var}(X)$.

8. Rebecca has a bunch of 3 keys, one of which fits a lock. If she tries the keys at random (i.e. forgetting which she has tried previously) obtain the probability distribution for X, the number of attempts made until the lock opens. Julian has more sense and tries the keys in sequence. Obtain the distribution for Y, the number of attempts Julian makes. Find
 (a) $P(X < Y)$, (B) $P(Y < X)$.

9. Adam and Betty each have a fair six sided die numbered 1, 2, 3, 4, 5, 6. Adam throws his die once and records X, his score multiplied by 3. Betty throws her die 3 times and records Y, the sum of the 3 scores. Calculate the mean and variance for a single throw of a die and hence deduce the mean and variance for both X and Y.

10. At a roll-a-ball stall at the funfair the stake is 10p for 4 balls. Each roll of a ball is independent of any others and the balls can roll into slots with scores as follows: -

Score X	1	2	3	4	5
P(X)	1/8	2/8	2/8	2/8	1/8

Prizes are awarded as follows: -

Total score	18 or more	8	less than 8
Prize	£1	20p	50p

Calculate the expected prize money from a roll of 4 balls.

THE BINOMIAL DISTRIBUTION

DEFINITION

If p is the constant probability that an event occurs in a single trial and q ($= 1 - p$) is the probability that it does not occur then the probability of x successes in n independent trials is given by: -

$$P(X = x) = {}^nC_x \cdot p^x \cdot q^{n-x}$$

where x covers the sample space $0, 1, 2, 3, \dots, n$

MEAN AND VARIANCE

If $x \sim B(n, p)$ then $E(X) = np$ and $\text{Var}(X) = npq$.

EXAMPLE 1

If 20% of garments produced in a clothing factory are defective, find the probability that if 5 garments are chosen at random: -

- (a) 1 is defective,
- (b) 3 are defective,
- (c) at most 3 are defective.

If 100 such samples of 5 were drawn, how many of them would be expected to have no defective items?

SOLUTION

Let X be the number of defectives in a sample of 5.

Then $X \sim B(5, 1/5)$.

$$\begin{aligned} (a) P(X = 1) &= 5 \times (1/5) \times (4/5)^4 \\ &= \underline{0.410} \quad (\text{or using cumulative Binomial tables, } 0.737 - 0.328) \end{aligned}$$

$$\begin{aligned} (b) P(X = 3) &= 5! / 3! 2! \times (1/5)^3 \times (4/5)^2 \\ &= \underline{0.051} \quad (0.993 - 0.942) \end{aligned}$$

$$\begin{aligned} (c) P(X = 0, 1, 2, \text{ or } 3) &= 1 - P(X = 4 \text{ or } 5) \\ &= 1 - 5 \times (1/5)^4 \times (4/5) - (1/5)^5 \\ &= \underline{0.993} \quad (\text{or directly from tables}) \end{aligned}$$

$$\begin{aligned} P(X = 0) &= (4/5)^5 \\ &= \underline{0.328} \quad (\text{or directly from tables}) \end{aligned}$$

$$\begin{aligned} \text{So } 100 \times 0.328 &= \underline{33} \text{ (say)} \quad \text{samples would be expected to have no defectives.} \end{aligned}$$

EXAMPLE 2

Derive the mean and variance of the Binomial distribution.

If X has a Binomial distribution with mean 9 and variance 4.95 find the probability that X is greater than 10.

SOLUTION

$$E(X) = \sum x.P(X = x)$$

$$= 0.q^n + 1.nq^{n-1}p + 2.\frac{n(n-1)}{2!}q^{n-2}p^2 + 3.\frac{n(n-1)(n-2)}{3!}q^{n-3}p^3 + \dots + n.p^n$$

$$= np(q^{n-1} + (n-1)q^{n-2}p + \frac{(n-1)(n-2)}{2!}q^{n-3}p^2 + \dots + p^{n-1})$$

$$= np(q + p)^{n-1}$$

$$= np \text{ since } p + q = 1.$$

$$E(X^2) = \sum X^2.P(X = x)$$

$$= 0^2q^n + 1^2nq^{n-1}p + 2^2\frac{n(n-1)}{2!}q^{n-2}p^2 + 3^2\frac{n(n-1)(n-2)}{3!}q^{n-3}p^3 + \dots + n^2p^n$$

$$= np\{q^{n-1} + 2(n-1)q^{n-2}p + 3\frac{(n-1)(n-2)}{2!}q^{n-3}p^2 + \dots + np^{n-1}\}$$

$$= np\{[q^{n-1} + (n-1)q^{n-2}p + \frac{(n-1)(n-2)}{2!}q^{n-3}p^2 + \dots + p^{n-1}]\}$$

$$+ [(n-1)q^{n-2}p + 2\frac{(n-1)(n-2)}{2!}q^{n-3}p^2 + \dots + (n-1)p^{n-1}]\}$$

$$= np\{[q + p]^{n-1} + (n-1)p[q^{n-2} + (n-2)q^{n-3}p^2 + \dots + p^{n-2}]\}$$

$$= np\{1 + (n-1)p[q + p]^{n-2}\}$$

$$= np\{1 + (n-1)p\}$$

$$= np(1 - p) + n^2p^2$$

$$\text{Now Var}(x) = E(X^2) - E^2(X)$$

$$= np(1 - p) + n^2p^2 - (np)^2$$

$$= npq$$

$$np = 9$$

$$npq = 4.95$$

$$\text{thus } q = 0.55$$

$$\text{and } p = 0.45$$

$$\text{and } n = 20$$

$$\text{i.e. } X \sim B(20, 0.45)$$

$$P(X > 10) = 1 - P(X \leq 10)$$

$$= 1 - 0.751 \quad (\text{tables})$$

$$= \underline{0.249}$$

(b) Let X be the number of damaged tiles, then $X \sim B(12, p)$.

Since $1/3$ of the tables are seconds, the ratio

seconds : perfect = 1 : 2

$P(X = 1) : P(X = 0) = 1 : 2$

so we use the above formula with $n = 12$, $x = 1$: -

$$12p/q = 1/2$$

$$24p = 1 - p$$

$$25p = 1$$

$$p = 1/25$$

$$\text{Proportion scrapped} = P(X > 1)$$

$$= 1 - P(X = 0 \text{ or } 1)$$

$$= 1 - \{(24/25)^{12} + 12 \cdot (1/25) \cdot (24/25)^{11}\}$$

$$= 1 - 0.919$$

$$= (\text{say}) \underline{8\% \text{ scrapped.}}$$

Expected profit per table	£30	£10	-£20
With probability that $X =$	0	1	>1

expected profit from 100 tables

$$= 100[\text{£}30 \times (24/25)^{12} + \text{£}10 \times 12 \times (1/25) \times (24/25)^{11} - \text{£}20 \times \{(1 - P(X=0,1))\}]$$

$$= 100[\text{£}30 \times 0.6127 + \text{£}10 \times 0.3064 - \text{£}20 \times 0.0809]$$

$$= \underline{\text{£}1980} \quad 3 \text{ significant figures.}$$

EXAMPLE 3

Show that if $X \sim B(n, p)$ then $\frac{P(X = x)}{P(X = x - 1)} = \frac{(n - x + 1)p}{xq}$

(a) If $X \sim B(8, p)$ and $P(X = 5) = 6P(X = 6)$ find the value of p .

(b) Coffee tables have tiled tops consisting of 12 tiles.

The probability that a single tile is damaged is p .

If no tiles on a table top are damaged it is sold as perfect for £50.

If only one tile is damaged it is sold as a "second" for £30.

If more than one tile is damaged it is scrapped.

It is found that, of all tables sold, $1/3$ of them are seconds.

Calculate the value of p .

Calculate the proportion of tables which are scrapped.

The tables cost £20 each to make, calculate the expected profit from 100 tables.

SOLUTION

$$\begin{aligned} \frac{P(X = x)}{P(X = x - 1)} &= \frac{{}^nC_x p^x q^{n-x}}{{}^nC_{x-1} p^{x-1} q^{n-(x-1)}} \\ &= \frac{n! / x! (n-x)! \cdot p^x q^{n-x}}{n! / (x-1)! (n-x+1)! \cdot p^{x-1} q^{n-x+1}} \\ &= \frac{n!}{x! (n-x)!} \cdot \frac{(x-1)! (n-x+1)!}{n!} \cdot \frac{p}{q} \\ &= \frac{(n-x+1) \cdot p}{x \cdot q} \end{aligned}$$

(a) Using the above formula with $n = 8$, $x = 6$ and $P(X = 6)/P(X = 5) = 1/6$ gives

$$3p/6q = 1/6$$

$$3p = q$$

$$3p = 1 - p$$

$$4p = 1$$

$$p = 1/4$$

EXAMPLE 4

An employee of a mail order club receives £10 commission for each new customer he obtains.

The probability that any prospective client is persuaded to join the club is 0.05.

- (a) Calculate the probability that he will obtain 3 or more members if he knocks on 20 doors.
- (b) State the commission he could expect from knocking on 20 doors.
- (c) State how many doors he would have to knock on per week if he were to obtain 15 new customers per week.
- (d) Calculate the number of doors he must knock on to be 90% sure of getting at least one new customer.

SOLUTION

Let X be the number of new customers, then $X \sim B(20, 1/20)$

$$\begin{aligned}
 (a) \quad P(X \geq 3) &= 1 - P(X = 0, 1 \text{ or } 2) \\
 &= 1 - 0.925 \\
 &= \underline{0.075}
 \end{aligned}$$

$$\begin{aligned}
 (b) \quad \text{Expected number of new customers} &= \text{mean of the Binomial distribution} \\
 &= 20 \times 1/20 \\
 &= 1 \text{ new customer.}
 \end{aligned}$$

$$\text{Therefore expected commission} = \underline{\text{£10.}}$$

$$(c) \quad \text{Now } X \sim B(n, 1/20) \text{ and } E(X) = 15$$

$$\text{Therefore } n \times 1/20 = 15$$

$$\underline{n = 300}$$

$$(d) \quad \text{Again } X \sim B(N, 1/20) \text{ and we require } P(X \geq 1) \geq 0.9$$

$$\text{i.e. } 1 - P(X = 0) \geq 0.9$$

$$1 - (19/20)^N \geq 0.9$$

$$(19/20)^N \leq 0.1$$

$$\begin{aligned}
 \underline{N = 45.} \quad & \text{(using trial \& error, by} \\
 & \text{calculator or} \\
 & \text{by logs.)}
 \end{aligned}$$

EXERCISES

1. A darts player has a probability of $1/10$ of hitting the bull.
Calculate the probability that, with 6 darts he scores
(a) 1, (b) at least 1 bull.

2. In the manufacture of carpets it was found that 4% were below standard. Find the probability that, in a sample of 10: -
(a) 2, (b) not more than 2 were not up to standard.

3. If $X \sim B(n, p)$ and $E(X) = 4$ and $\text{Var}(X) = 2.4$, calculate n and p . Find $P(X < 4)$

4. The probability that an "A" level maths candidate will pass is 0.7. Calculate the probabilities that, out of 20 candidates:-
(a) all pass
(b) 14 pass
(c) at least 14 pass
(d) not more than 10 fail

5. Show that if $X \sim B(n, p)$ then the ratio $x + 1$ successes to x successes is

$$\frac{(n - x)p}{(x + 1)(1 - p)}$$

Lettuce plants are planted in batches of 12. The number of plants not maturing, X , in 200 such batches is tabulated below.

Fit a Binomial distribution to the data.

Number of plants X	0	1	2	3	4
frequency f	73	78	36	10	3

REVISION IN STATISTICS : Chapter 4

6. A multiple choice test consists of 60 questions each with 5 possible answers. If X is the number of correct solutions from a candidate who guesses the answers at random, state the mean and standard deviation for X .
Calculate the probability that such a candidate scores this mean score.
7. Define the Binomial distribution and derive its mean and variance.
8. Playing cards are drawn at random, with replacement from a normal pack of 52.
If the probability that at least one spade is drawn is to be greater than 0.95, how many cards need to be drawn?
9. Fireworks leave the factory in batches of 10,000. If, on average, 0.5% are duds find the mean and standard deviation of the number of duds per batch.
10. A pupil is early for school nine mornings out of ten.
(a) Find the expected value and standard deviation of the number of times he is early in a week. (5 days)
(b) Find the probability that he will be early at least 4 times in a week.
(c) Given that he is early on Monday, find the probability that he will be early at least 4 times that week.

THE POISSON DISTRIBUTION

DEFINITION

If the random variable X represents the number of occurrences of some "rare event" (i.e. the probability of the event occurring is low) such that

$$P(X = x) = e^{-a} \cdot \frac{a^x}{x!} \quad \text{for } x = 0, 1, 2, 3, \dots$$

then X has a Poisson distribution with mean a .
Or, in short $X \sim Pn(a)$

MEAN AND VARIANCE

$$E(X) = a$$

$$\text{and } \text{Var}(X) = a.$$

POISSON APPROXIMATION TO BINOMIAL

If $X \sim B(n, p)$ then, provided n is large (≥ 30 ish) and p is small. (such that $np \leq 6$ ish), then $X \div Pn(a = np)$.

REVISION IN STATISTICS : Chapter 5

EXAMPLE 1

200 samples, each of 50 screws, are to be drawn from a production line. If 5% of the whole production is defective, calculate the number of samples in which you would expect

(a) 0, (b) 1, (c) 2, (d) 3 or more defectives

SOLUTION

Let X be the number of defective screws per sample. Then

$X \sim B(50, 0.05)$. Since n is large and p is small, the Poisson approximation is $X \sim Pn(2.5)$

$$\begin{aligned} \text{(a) } P(X = 0) &= e^{-2.5} \cdot \frac{2.5^0}{0!} \\ &= e^{-2.5} \\ &= \underline{0.082} \end{aligned}$$

$$\text{No. of samples} = 200 \times 0.082 = \underline{16}$$

$$\begin{aligned} \text{(b) } P(X = 1) &= e^{-2.5} \cdot \frac{2.5^1}{1!} \\ &= 2.5 \times p_0 \end{aligned}$$

$$\begin{aligned} &= \underline{0.205} \\ \text{No. of samples} &= 200 \times 0.205 = \underline{41} \end{aligned}$$

$$\begin{aligned} \text{(c) } P(X = 2) &= e^{-2.5} \cdot \frac{2.5^2}{2!} \\ &= \frac{2.5}{2} \times p_1 \end{aligned}$$

$$\begin{aligned} &= \underline{0.257} \\ \text{No. of samples} &= 200 \times 0.257 = \underline{51} \end{aligned}$$

$$\begin{aligned} \text{(d) } P(X \geq 3) &= 1 - P(X = 0, 1, 2) \\ &= 1 - (0.082 + 0.205 + 0.257) \\ &= \underline{0.456} \end{aligned}$$

$$\text{No. of samples} = 200 \times 0.456 = \underline{92}$$

(Cumulative Poisson probability tables could have been used throughout this question)

EXAMPLE 2

Show that, if $X \sim P(a)$, then $\frac{P(X = x + 1)}{P(X = x)} = \frac{a}{x+1}$

Past experience has shown that, on average, a motorist has 2.8 punctures per year. Find the probabilities that in any year he has: -

- (a) 0, (b) 1, (c) 2, (d) 3, (e) 4, (f) 5 or more punctures.

If he owns a car for 18 months, state the expected number of punctures and find (g) the probability that he has at least 2 punctures in this period.

SOLUTION

$$\begin{aligned} \frac{P(X = x + 1)}{P(X = x)} &= \frac{e^{-a} \cdot a^{x+1}/(x+1)!}{e^{-a} \cdot a^x/x!} \\ &= \frac{a^{x+1} \cdot x!}{a^x \cdot (x+1)!} \\ &= \frac{a}{x+1} \end{aligned}$$

Let X be the number of punctures per year, then $X \sim Pn(2.8)$

$$\begin{aligned} \text{(a)} \quad P(X = 0) &= e^{-2.8} \\ &= \underline{0.0608} \\ \text{(b)} \quad P(X = 1) &= 0.0608 \times 2.8 \\ &= \underline{0.1703} \\ \text{(c)} \quad P(X = 2) &= 0.1703 \times 2.8/2 \\ &= \underline{0.2384} \\ \text{(d)} \quad P(X = 3) &= 0.2384 \times 2.8/3 \\ &= \underline{0.2225} \\ \text{(e)} \quad P(X = 4) &= 0.2225 \times 2.8/4 \\ &= \underline{0.1557} \\ \text{(f)} \quad P(X \geq 5) &= 1 - P(X = 0, 1, 2, 3 \text{ or } 4) \\ &= \underline{0.1523} \end{aligned}$$

If Y is the number of punctures in 1.5 years, then $Y \sim Pn(1.5 \times 2.8)$, $Y \sim Pn(4.2)$

$$\begin{aligned} \text{(g)} \quad P(Y \geq 2) &= 1 - P(Y = 0 \text{ or } 1) \\ &= 1 - e^{-4.2}(1 + 4.2) \\ &= \underline{0.922} \end{aligned}$$

EXAMPLE 3

If the random variables X and Y are independently distributed as Poisson distributions with means a and b respectively, state the distribution of $X + Y$.

Verify this result for $X + Y = 0, 1$ and 2 .

Fred has, on average, a road accident every 4 years.

His wife Betty averages one every 8 years independent of Fred.

Calculate the probabilities that in a given year

- (a) Fred has one accident,
- (b) Betty has one accident,
- (c) They have a total of 2 accidents between them.

SOLUTION

If $X \sim Pn(a)$ and $Y \sim Pn(b)$ then $X + Y \sim Pn(a + b)$

$$\begin{aligned} P(X + Y = 0) &= P(X = 0 \text{ and } Y = 0) \\ &= e^{-a} \cdot e^{-b} \\ &= \underline{e^{-(a+b)}} \end{aligned}$$

$$\begin{aligned} P(X + Y = 1) &= P(X = 0 \text{ and } Y = 1 \text{ or } X = 1 \text{ and } Y = 0) \\ &= e^{-a} \cdot be^{-b} + ae^{-a} \cdot e^{-b} \\ &= \underline{(a + b) \cdot e^{-(a+b)}} \end{aligned}$$

$$\begin{aligned} P(X + Y = 2) &= P(X = 2 \text{ and } Y = 0 \text{ or } X = 1 \text{ and } Y = 1 \text{ or } X = 0 \text{ and } Y = 2) \\ &= a^2/2 \cdot e^{-a} \cdot e^{-b} + a \cdot e^{-a} \cdot b \cdot e^{-b} + e^{-a} \cdot b^2/2 \cdot e^{-b} \\ &= e^{-(a+b)}/2 \cdot (a^2 + 2ab + b^2) \\ &= \underline{e^{-(a+b)} \cdot (a + b)^2/2} \end{aligned}$$

and each of these probabilities is consistent with $X + Y$ having a Poisson distribution with mean $a + b$.

Let X be the number of accidents Fred has per year, then $X \sim Pn(1/4)$.

If Y is the number of accidents that Betty has per year, $Y \sim Pn(1/8)$.

If $X + Y$ is the total number of accidents they have per year,
 $X + Y \sim Pn(3/8)$.

$$\begin{aligned} (a) \quad P(X = 1) &= e^{-0.25} \times 0.25 \\ &= \underline{0.1947} \end{aligned}$$

$$\begin{aligned} (b) \quad P(Y = 1) &= e^{-0.125} \times 0.125 \\ &= \underline{0.1103} \end{aligned}$$

$$\begin{aligned} (c) \quad P(X + Y = 2) &= e^{-0.375} \times 0.375^2/2 \\ &= \underline{0.0483} \end{aligned}$$

EXAMPLE 4

If on average, 2% of rats are albinos, find the probability that there are no albinos in a sample of

- (a) 10, (b) 100. (c) How large must the sample be to be 99% certain that it contains at least 1 albino.

SOLUTION

- (a) Let X be number of albinos in a sample of 10. Then $X \sim B(10, 0.02)$

$$\begin{aligned} P(X = 0) &= 0.98^{10} \\ &= \underline{0.817} \quad (\text{Poisson approximation gives } 0.819) \end{aligned}$$

- (b) If Y is the number of albinos in a sample of 100, then $Y \sim B(100, 0.02)$

Now since n is large and np is small, the Poisson approximation can be used

$$Y \sim Pn(2)$$

$$\begin{aligned} P(Y = 0) &= e^{-2} \\ &= \underline{0.135} \end{aligned}$$

- (c) If Z is the number of albinos in a sample of n , then, $Z \sim B(n, 0.02)$ and, providing n is large, $Z \sim Pn(0.02n)$

$$\text{We require } P(Z \geq 1) \geq 0.99$$

$$1 - P(Z = 0) \geq 0.99$$

$$1 - e^{-0.02n} \geq 0.99$$

$$e^{-0.02n} \leq 0.01$$

$$e^{0.02n} \geq 100$$

$$\underline{n \geq 231}$$

(calculator trial and error or logs)

EXERCISES

1. A typist makes, on average, 1.2 errors per page. Find the probabilities that, on a given page he makes
(a) 0, (b) 1, (c) 2, (d) at least one mistake.

2. State how, and under what conditions, the Poisson distribution may be used as an approximation to a Binomial distribution.

A farmer finds that, over a long period of time, one egg in 200 has a double yoke.

Find the probability that in a tray of 30 eggs there are
(a) 0, (b) 1, (c) more than one double yokes.

3. The number of calls arriving at a police station switchboard is one every 30 seconds. Calculate the probability that, in a period of 5 minutes, only 8 calls are received.

4. It is thought that the number of children per family follows a Poisson distribution. The following sample was drawn to test this hypothesis.

Number of children	0	1	2	3	4	5	6	7
Number of families	12	12	17	7	2	7	1	2

 Calculate the mean and standard deviation for the number of children per family and thus calculate the theoretical Poisson frequencies.

 Plot, on the same axes, frequency polygons to compare the realised and theoretical frequencies.

5. If $X \sim Pn(a)$ show that the ratio

$$P(X = x) : P(X = x - 1) = a : x, \quad (x > 0)$$

 If $X \sim Pn(3.2)$ obtain successive Poisson probabilities until
 $P(X = x) < 0.1, \quad (X > 0)$

6. A small supermarket has 2 checkouts. It is found that the first checkout uses, on average 1.2 till rolls per week, the second 1.3 per week.

Calculate the probabilities that 4 or more rolls are used in a particular week by: -

- (a) the first checkout,
- (b) the second checkout,
- (c) the shop.

If both tills run out of rolls at the end of a week, calculate the number of rolls the shopkeeper must have in stock if he is to be at least 95% sure that he does not run out during the following week. (You may assume that the rolls can be inter-changed, if necessary.)

7. Derive the mean and variance of the Poisson distribution.

8. If $X \sim B(n, 0.003)$ and $E(X) = 3$, calculate $P(X > 3)$.

9. Misprints in the Guardain Newspepar occur on average at a rate of 1 every 2 columns. Calculate the probabilities that a column has

- (a) 1, (b) 2, (c) at least 1 misprint.

A page has 8 columns. Calculate the probabilities that, on page 3, there will be

- (d) 5, (e) less than 5 misprints.

Saturday's paper contains 24 pages. Calculate the probability that there will be not more than 100 misprints in that edition. (see chapter 7 for Normal approximation)

10. Two golfers A and B find that on average they score 2 and 3 "eagles" respectively per year. Calculate the probabilities that, in a given year,

- (a) neither A nor B scores an eagle,
- (b) they both score the same number of eagles,
- (c) B scores at least 2 eagles more than A.

Give answers correct to 3 decimal places.

The first part of the report deals with the general situation of the country and the progress of the work during the year.

The second part of the report deals with the results of the work during the year.

The third part of the report deals with the conclusions of the work during the year.

The fourth part of the report deals with the recommendations of the work during the year.

The fifth part of the report deals with the summary of the work during the year.

The sixth part of the report deals with the appendix of the work during the year.

The seventh part of the report deals with the bibliography of the work during the year.

The eighth part of the report deals with the index of the work during the year.

The ninth part of the report deals with the list of the work during the year.

The tenth part of the report deals with the conclusion of the work during the year.

The eleventh part of the report deals with the summary of the work during the year.

The twelfth part of the report deals with the appendix of the work during the year.

The thirteenth part of the report deals with the bibliography of the work during the year.

The fourteenth part of the report deals with the index of the work during the year.

The fifteenth part of the report deals with the list of the work during the year.

The sixteenth part of the report deals with the conclusion of the work during the year.

STATISTICAL TESTS FOR

WILCOXON SIGN TEST - page 2

SIGN TEST - page 3

CHI SQUARE - page 7

MANN WHITNEY UTEST - page 4

PEARSON'S PRODUCT MOMENT - page 11

SPEARMAN'S RHO - page 10

Operation schedule 7. The Wilcoxon matched-pairs signed ranks test

Data requirements

Scores must be paired off in some way, and of at least ordinal status.

Example

The Wilcoxon test can be carried out on the rating scores obtained for the two brands of washing up liquid, Gresego and Kwiklene (see pages 68-70). The null hypothesis (H_0) is that the two sets of scores do not differ, and the experimental hypothesis (H_1 or H_A) is that Gresego gives better results than Kwiklene.

Subject	Rating		Subject	Rating	
	Gresego	Kwiklene		Gresego	Kwiklene
1	8	5	6	7	6
2	7	5	7	9	5
3	9	2	8	6	5
4	7	6	9	5	6
5	8	9			

Step 1: The data will be paired off, but in doing so, it does not matter which column comes first. Subtract each number in the first column from its partner in the second. Record the differences, being sure to note the sign.

$$8 - 5 = +3, 7 - 5 = +2, \dots, 5 - 6 = -1$$

The differences are: +3, +2, +7, +1, -1, +1, +4, +1, -1

Step 2: Rank the differences according to size, and giving the smallest difference rank 1. Do not rank any values of 0 which occur. Ignore the signs during the ranking procedure. (See schedule 6 for ranking instructions.)

Differences +3 +2 +7 +1 -1 +1 +4 +1 -1
Ranks 7 6 9 3 3 3 8 3 3

Step 3: Add the ranks for all the differences which are positive, and all the negative differences. Keep the two totals separate.

$$\text{Positive difference ranks } 7 + 6 + 9 + 3 + 3 + 8 + 3 = 39$$

$$\text{Negative difference ranks } 3 + 3 = 6$$

Step 4: Whichever total of the two obtained in Step 3 is the smaller is the value of the test statistic T .

$$T = 6$$

Step 4: To determine N , in order to decide whether a given value of T is significant using table S2, count the number of pairs of scores used, but subtract any pairs whose difference was found to be 0 in Step 1.

There are 9 pairs of scores, and no values of 0. $N = 9$.

Operation schedule 7a. Evaluation of Wilcoxon's T statistic

Example

The data analysed using Wilcoxon's test, in schedule 7 will be evaluated for significance, using table S2. From the data, $T = 6$, and $N = 9$.

Step 1: Decide whether the experimental hypothesis (H_1) was a one-tailed or two-tailed hypothesis.

In the experiment on washing up liquid, it had been hypothesised that scores from Gresego

Table S2. Critical values of T for the Wilcoxon matched-pairs signed ranks test. T must be equal to or less than the stated value to be significant

N	Level of significance for one-tailed test				Level of significance for one-tailed test				
	0.05	0.025	0.01	0.005	0.05	0.025	0.01	0.005	
	Level of significance for two-tailed test				Level of significance for two-tailed test				
N	0.10	0.05	0.02	0.01	N	0.10	0.05	0.02	0.01
5	1	—	—	—	28	130	117	101	92
6	2	1	—	—	29	141	127	111	100
7	4	2	0	—	30	152	137	120	109
8	6	4	2	0	31	163	148	130	118
9	8	6	3	2	32	175	159	141	128
10	11	8	5	3	33	188	171	151	138
11	14	11	7	5	34	201	183	162	149
12	17	14	10	7	35	214	195	174	160
13	21	17	13	10	36	228	208	186	171
14	26	21	16	13	37	242	222	198	183
15	30	25	20	16	38	256	235	211	195
16	36	30	24	19	39	271	250	224	208
17	41	35	28	23	40	287	264	238	221
18	47	40	33	28	41	303	279	252	234
19	54	46	38	32	42	319	295	267	248
20	60	52	43	37	43	336	311	281	262
21	68	59	49	43	44	353	327	297	277
22	75	66	56	49	45	371	343	313	292
23	83	73	62	55	46	389	361	329	307
24	92	81	69	61	47	408	379	345	323
25	101	90	77	68	48	427	397	362	339
26	110	98	85	76	49	446	415	380	356
27	120	107	93	84	50	466	434	398	373

Source: F. Wilcoxon and R. A. Wilcox, *Some rapid approximate statistical procedures*, page 28, table 2, 1964. New York: Lederle Laboratories. Reproduced with the permission of American Cyanamid Company

Step 2: Locate the appropriate value for N in table S2, using the column headed N . Opposite $N = 9$, you should read the values 8 6 3 2.

Step 3: Move along the row until you arrive at the tabled value which is equal to or just larger than the value of T which you have just obtained.

The second value of 6 is equal to the T just obtained.

Step 4: Read the appropriate significance level off from the top of the column reached in Step 3. Take the one- or two-tailed probability level stated, according to the decision made in Step 1.

The obtained value of 6 is significant at the 0.025 level for a one-tailed hypothesis.

Step 5: State the conclusion.

The results were significant at the $p = 0.025$ level, using Wilcoxon's test ($N = 9$, $T = 6$; one-tailed hypothesis). Therefore the null hypothesis can be rejected, and it is concluded that the performance of Gresego was superior to that of Kwiklene.

Note

If the obtained value of T falls between tabled values, both of which are significant, then the most conservative (i.e. the least significant, and with the highest value of p)

8 The sign test

Step 2: Count the number of times the less frequent sign occurs, to obtain the statistic S .

The minus signs are less frequent, and occur twice. So $S = 2$.

Step 3: Count the total number of pluses and minuses, to obtain N . (Note that any differences of 0 will not be included in N .)

There are 9 signs, and so $N = 9$.

Step 4: Decide whether your test comprises a one- or two-tailed test.

The hypothesis under consideration is a one-tailed hypothesis.

Table S3. Critical values of S for the sign test. S must be equal to or less than the stated value to be significant

N	Level of significance for one-tailed test			Level of significance for two-tailed test		
	0.05	0.025	0.01	0.05	0.02	0.01
5	0	—	—	—	—	—
6	0	0	—	—	—	—
7	0	0	—	—	—	—
8	1	0	0	0	0	—
9	1	1	0	0	0	—
10	1	1	0	0	0	—
11	2	1	1	0	0	0
12	2	2	1	1	1	0
13	3	2	1	1	1	0
14	3	2	2	1	1	0
15	3	3	2	2	2	1
16	4	3	2	2	2	1
17	4	4	3	2	2	1
18	5	4	3	3	3	1
19	5	4	4	3	3	2
20	5	5	4	4	4	2
25	7	7	6	5	5	4
30	10	9	8	7	7	5
35	12	11	10	9	9	7

Step 5: Use table S3 to evaluate the statistic S (from Step 2), in conjunction with N (determined in Step 3). The table is used in just the same manner as table S2 for the Wilcoxon test (see schedule 7a).

Opposite the value 9, we read 1, 1, 0, 0. — Our obtained value of $S = 2$ exceeds all these tabulated values, and so we conclude that as the probability exceeds that given for the 0.05 level, one-tailed test, our results are not significant.

Step 6: State the conclusion.

The results of the analysis were nonsignificant ($S = 2$, $N = 9$, sign test), and so the null hypothesis cannot be rejected. It is concluded that the two washing up liquids do not differ in their performance.



Operation schedule 8: The sign test

Data requirements

The scores must be paired off in some way, but can be the simplest type of data (i.e. nominal).

Example

The data derived from the washing up experiment and analysed in schedule 7 will be used once more.

Subject	Rating		Subject	Rating	
	Gresego	Kwiklenc		Gresego	Kwiklenc
1	8	5	6	7	6
2	7	5	7	9	5
3	9	2	8	6	5
4	7	6	9	5	6
5	8	9			

Step 1: Put the data into a table, as above. It does not matter which column of scores comes first. Mentally subtract each value in the second column from its partner in the first. If the answer is positive, give the pairs a plus (+) sign, and if negative, a minus (−) sign. Note 0 if the two values are equal.

<i>S</i>	Gresgo	Kwiklone	Sign	<i>S</i>	Gresgo	Kwiklone	Sign
1	8	5	+	6	7	6	+
2	7	5	+	7	9	5	+
3	9	2	+	8	6	5	+
4	7	6	+	9	5	6	-
5	8	9	-				

Operation schedule 9. The Mann-Whitney U test

Data requirements

Scores need to be of at least ordinal status.

Example

The results gathered whilst investigating a new memorising technique, described in chapter 1, will be used for analysis. They are:

Set 1: 30, 35, 45, 50, 75, 80 (the unaided group)

Set 2: 45, 50, 58, 62, 69, 70 (the aided group)

The experimental hypothesis was two-tailed, and the null hypothesis states that there will be no difference between the two sets of scores.

Step 1: Put the two sets of scores into two vertical columns, with a good gap between them. If one of the lists is smaller than the other, put it first, and call it 'list A'. If the lists are the same size, then it doesn't matter which comes first. Call the second list 'list B'.

List A	List B
30	45
35	50
45	58
50	62
75	69
80	70

Step 2: Count the number of scores in each list, to obtain N_A and N_B . Then multiply the two values together. This gives $N_A N_B$.

$$N_A = 6, N_B = 6; N_A N_B = 6 \times 6 = 36$$

Step 3: Take the number of scores in list A, and add the value of 1.

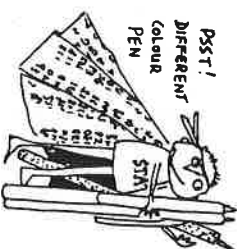
$$6 + 1 = 7$$

Step 4: Multiply the value found in Step 3 by the number of scores in list A, and divide the answer by 2.

$$\frac{7 \times 6}{2} = \frac{42}{2} = 21$$

Step 5: Rank all the numbers in both groups (see schedule 6 for ranking procedure), taking both sets of numbers together, and giving the smallest score rank 1. Write the ranks out in two more columns, headed R_A and R_B , and put each to the immediate right of the original lists of scores A and B.

List A	R_A	List B	R_B
30	1	45	31
35	2	50	51
45	31	58	7
50	51	62	18
75	11	69	9
80	12	70	10



Step 6: Add the ranks which were given to the items on list A (i.e. R_A)

$$1 + 2 + 31 + 51 + 11 + 12 = 35$$

Step 7: Add the value found in Step 2 to that found in Step 4, and subtract the value of Step 6.

$$36 + 21 - 35 = 22$$

Step 8: Subtract the value of Step 7 from the result of Step 2.

$$36 - 22 = 14$$

Step 9: The values of Steps 7 and 8 give two values for the statistics U and U' (pronounced 'U prime'). The smaller value will be U , and the larger U' .

$$U = 14 \text{ and } U' = 22$$

Step 10: Use the two values of N_A and N_B , found in Step 2, in conjunction with table S4 (overleaf), to determine whether the obtained value of U is significant. If U is equal to or smaller than the critical values listed, then the null hypothesis can be rejected. Two significance levels are given in the table, for the 0.05 and 0.01 probabilities (two-tailed), and the 0.025 and 0.005 probabilities (one-tailed tests).

Looking in the table at the point where $N_A = 6$ and $N_B = 5$ intersect, we read off the two critical values of 5 and 2. Our value of U (14) exceeds both numbers, and so we cannot reject the null hypothesis.

Step 11: State the conclusion.

The results of the analysis were nonsignificant ($U = 14$, $N_A = N_B = 6$; Mann-Whitney U test), and so the null hypothesis cannot be rejected. It is concluded that the scores obtained from subjects who were using the memory aid do not differ from those obtained by subjects without the aid.

Note

A useful check is to rework all the calculations, but call the original list A, list B. You should obtain precisely the same values for U and U' . Again, the smaller of the two values is labelled U , and evaluated.

Formulae

The formula for U is either

$$N_A N_B + \frac{N_A(N_A + 1)}{2} - \Sigma R_A$$

$$\text{or } N_A N_B + \frac{N_B(N_B + 1)}{2} - \Sigma R_B$$

where N_A = the number of items in list A

N_B = the number of items in list B

R_A = the sum of the ranks given to items in list A

R_B = the sum of the ranks given to items in list B

Table 5d Critical values of U for a one-tailed test at 0.05; two-tailed test at 0.10^* (Mann-Whitney)

n_2	n_1																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
2	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
3	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
4	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
5	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
6	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
7	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
8	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
9	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
10	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
11	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
12	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
13	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
14	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
15	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
16	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
17	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
18	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
19	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
20	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—

* Dashes in the body of the table indicate that no decision is possible at the stated level of significance.

For any n_1 and n_2 the observed value of U is significant at a given level of significance if it is equal to or less than the critical values shown.
Source: R. Runyon and A. Haber (1976) Fundamentals of Behavioural Statistics (3rd ed.) Reading, Mass.: McGraw Hill, Inc. with kind permission of the publisher.

Table 5c Critical values of U for a one-tailed test at 0.025; two-tailed test at 0.05* (Mann-Whitney)

n_2	n_1																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
2	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
3	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
4	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
5	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
6	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
7	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
8	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
9	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
10	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
11	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
12	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
13	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
14	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
15	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
16	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
17	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
18	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
19	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
20	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—

* Dashes in the body of the table indicate that no decision is possible at the stated level of significance.

For any n_1 and n_2 , the observed value of U is significant at a given level of significance if it is equal to or less than the critical values shown.

Source: R. Runyon and A. Haber (1976) Fundamentals of Behavioural Statistics (3rd ed.) Reading, Mass.: McGraw Hill, Inc. with kind permission of the publisher.

Operation schedule 12. Simple chi-square

Data requirements

- 1 Data must be in frequency form, i.e. counted, rather than scores.
- 2 Entries in each cell must be independent.
- 3 The *expected* number for each cell must not be less than 5.

Example

One hundred and fifty-six boys and two hundred and four girls were asked whether they liked vanilla slices. Ninety-four boys said that they did, and one hundred and seventy-five girls. The remainder did not like them. Is there any evidence that there is a sex difference in preference for vanilla slices?

Step 1: Plot the data into a large 2×2 table. The number in each cell is the *observed* frequency. Obtain row and column totals, and a grand total (N).

	Column 1 Like slices	Column 2 Dislike slices	Totals
Row 1 Boys	Cell A 94	Cell B 62	156
Row 2 Girls	Cell C 175	Cell D 29	204
Totals	269	91	360

Step 2: Multiply the row 1 total with the column 1 total, and divide the answer by the overall total, N . This gives the *expected* frequency for cell A.

By multiplying the row and column totals relating to each cell together, and dividing the answer by the overall total, N , you obtain the *expected* frequencies for each cell. Write the value obtained for each cell below the *obtained* frequency (with a different coloured pen).

$$\text{Cell A: } \frac{156 \times 269}{360} = 116.6$$

$$\text{Cell B: } \frac{156 \times 91}{360} = 39.4$$

$$\text{Cell C: } \frac{204 \times 269}{360} = 152.4$$

$$\text{Cell D: } \frac{204 \times 91}{360} = 51.6$$

Step 3: Find the difference between each *observed* and *expected* frequency for each cell, always taking the smaller from the larger. Then, for each cell value, subtract 0.5.

$$\text{Cell A: } 116.6 - 94 - 0.5 = 22.1$$

$$\text{Cell B: } 62 - 39.4 - 0.5 = 22.1$$

$$\text{Cell C: } 175 - 152.4 - 0.5 = 22.1$$

$$\text{Cell D: } 51.6 - 29 - 0.5 = 22.1$$

Step 4: Square each of the cell values obtained in Step 3, and divide the answer by the *expected* frequency for that particular cell.

$$\text{Cell A: } \frac{(22.1)^2}{116.6} = 4.189 \quad \text{Cell B: } \frac{(22.1)^2}{39.4} = 12.396$$

$$\text{Cell C: } \frac{(22.1)^2}{152.4} = 3.205 \quad \text{Cell D: } \frac{(22.1)^2}{51.6} = 9.465$$

Step 5: Add together the four values obtained in Step 4 to get χ^2 .
 $4.189 + 12.396 + 3.205 + 9.465 = 29.255$

Step 6: Evaluate the value of χ^2 , using table S6. In a simple *chi*-square test the *df* will always be 1.

The obtained value of 29.25 exceeds the tabulated value of 10.83 which is significant at the 0.001 level (two-tailed test). The null hypothesis can be rejected.

Step 7: State the conclusion.

A *chi*-square test carried out on the data obtained was significant at the 0.001 level ($\chi^2 = 29.254$, $df = 1$), and so it is concluded that there is a difference between the sexes in liking for vanilla slices.

Formula

The formula for *chi*-square, in which Yates' correction is incorporated, is

$$\chi^2 = \sum \frac{(O - E - 0.5)^2}{E}$$

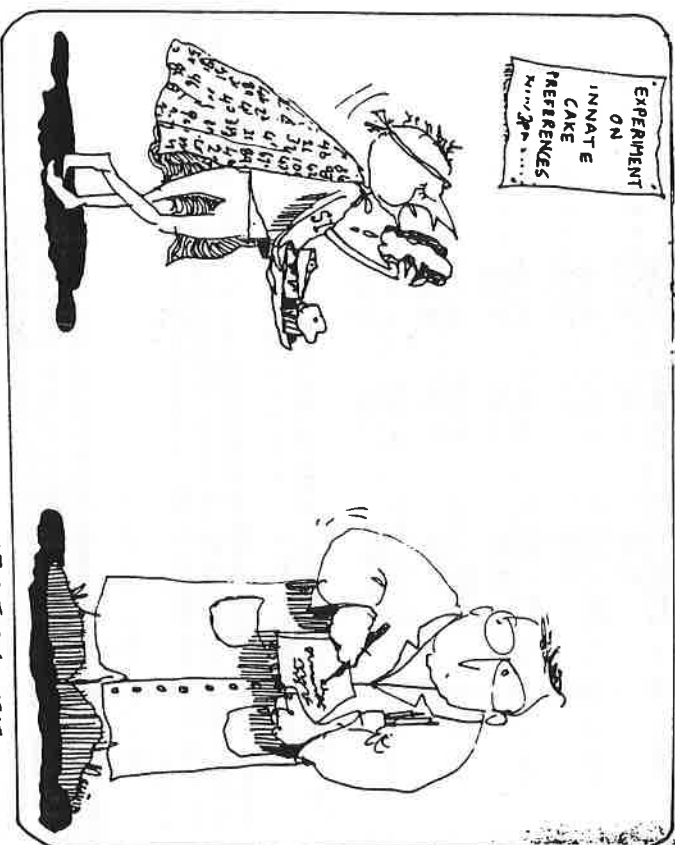
where E = the *expected* frequencies

O = the *observed* frequencies

Σ = the sum of

Table S6. Critical values of χ^2 . χ^2 must be equal to or more than the stated value to be significant

df	Level of significance for one-tailed test			Level of significance for two-tailed test		
	0.05	0.025	0.005	0.01	0.001	
1	2.706	3.841	6.635	10.83		
2	4.605	5.991	9.210	13.82		
3	6.251	7.815	11.34	16.27		
4	7.779	9.488	13.28	18.47		
5	9.236	11.07	15.09	20.52		
6	10.64	12.59	16.81	22.46		
7	12.02	14.07	18.48	24.32		
8	13.36	15.51	20.09	26.12		
9	14.68	16.92	21.67	27.88		
10	15.99	18.31	23.21	29.59		
11	17.28	19.68	24.73	31.26		
12	18.55	21.03	26.22	32.91		
13	19.81	22.36	27.69	34.53		



Operation schedule 13. Complex chi-square

Data requirements

Data must be in frequency form, i.e. counted, rather than actual scores. Entries in each cell must be independent. The expected number for each cell must not be less than 5.

Example

The different degree classifications obtained by finalists at Wetwang University (page 99) provide material for a complex chi-square analysis. The null hypothesis states that there is no difference between the three groups in terms of the proportions obtaining the different classifications.

Step 1: Put the data for the three groups of students and five degree classes into a 3×5 contingency table. The numbers in each cell stand for the numbers of students obtaining the particular degree and degree class, and are the *observed* frequencies. Add together the numbers across the rows to get row totals, and down the columns to give column totals. The row and column totals added together should agree, and that total will be the overall total, N . Leave plenty of space in the table, for you will later be writing the *expected* values in each cell.

Subjects	Degree classification					Totals
	I	Ii	Iii	III	Pass	
Languages	5	10	20	35	30	100
Maths	35	40	80	25	20	200
Economics	0	10	10	20	20	60
Totals	40	60	110	80	70	360

Step 2: To compute the *expected* frequency for individual cells, follow this procedure. Take the row sum and the column sum which intersect at the cell, and multiply the two values together. Then divide the answer by the grand total, N .

For the top left-hand cell, Languages, and a class I degree, it will be

$$\text{Row 1, Col. 1: } \frac{100 \times 40}{360} = \frac{4000}{360} = 11.111$$

Don't round off the decimal places at less than three.

$$\text{Row 2, Col. 1: } \frac{200 \times 40}{360} = \frac{8000}{360} = 22.222$$

$$\text{Row 3, Col. 1: } \frac{60 \times 40}{360} = \frac{2400}{360} = 6.667$$

$$\text{Row 1, Col. 2: } \frac{100 \times 60}{360} = \frac{6000}{360} = 16.667$$

$$\text{Row 2, Col. 2: } \frac{200 \times 60}{360} = \frac{12000}{360} = 33.333$$

$$\text{Row 3, Col. 2: } \frac{60 \times 60}{360} = \frac{3600}{360} = 10.000$$

$$\text{Row 1, Col. 3: } \frac{100 \times 110}{360} = \frac{11000}{360} = 30.556$$

$$\text{Row 2, Col. 3: } \frac{200 \times 110}{360} = \frac{22000}{360} = 61.111$$

$$\text{Row 3, Col. 3: } \frac{60 \times 110}{360} = \frac{6600}{360} = 18.333$$

$$\text{Row 1, Col. 4: } \frac{100 \times 80}{360} = \frac{8000}{360} = 22.222$$

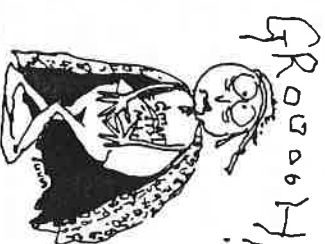
$$\text{Row 2, Col. 4: } \frac{200 \times 80}{360} = \frac{16000}{360} = 44.444$$

$$\text{Row 3, Col. 4: } \frac{60 \times 80}{360} = \frac{4800}{360} = 13.333$$

$$\text{Row 1, Col. 5: } \frac{100 \times 70}{360} = \frac{7000}{360} = 19.444$$

$$\text{Row 2, Col. 5: } \frac{200 \times 70}{360} = \frac{14000}{360} = 38.889$$

$$\text{Row 3, Col. 5: } \frac{60 \times 70}{360} = \frac{4200}{360} = 11.667$$



... WELL, NOT AT THE TIME OF PRACTICATION

Step 3: When the expected values have been calculated, put each one into the table below the appropriate obtained frequency. Check that they all exceed 4.

Subjects	Degree classification				Totals
	I	II	III	Pass	
Languages	5	10	20	35	30
	11.111	16.667	30.556	22.222	19.444
Maths	35	40	80	25	20
	22.222	33.333	61.111	44.444	38.889
Economics	0	10	10	20	20
	6.667	10.000	18.333	13.333	11.667
Totals	40	60	110	80	70
					360

Step 4: To obtain the values which will be added together to get χ^2 , the following procedure must be undertaken for each cell. The difference between the obtained and expected frequency for every cell is found, i.e. by subtracting the smaller value from the larger. This is then squared, and the value divided by the expected frequency for that particular cell.

Row 1, Col. 1:	$\frac{(11.111 - 5)^2}{11.111} = \frac{(6.111)^2}{11.111} = \frac{37.344}{11.111} = 3.361$
Row 2, Col. 1:	$\frac{(35 - 22.222)^2}{22.222} = \frac{(12.778)^2}{22.222} = \frac{163.277}{22.222} = 7.348$
Row 3, Col. 1:	$\frac{(6.667 - 0)^2}{6.667} = \frac{(6.667)^2}{6.667} = \frac{44.449}{6.667} = 6.667$
Row 1, Col. 2:	$\frac{(16.667 - 10)^2}{16.667} = \frac{(6.667)^2}{16.667} = \frac{44.449}{16.667} = 2.667$
Row 2, Col. 2:	$\frac{(40 - 33.333)^2}{33.333} = \frac{(6.667)^2}{33.333} = \frac{44.449}{33.333} = 1.333$
Row 3, Col. 2:	$\frac{(10 - 10)^2}{10} = \frac{(0)^2}{10} = \frac{0}{10} = 0$
Row 1, Col. 3:	$\frac{(30.556 - 20)^2}{30.556} = \frac{(10.556)^2}{30.556} = \frac{111.429}{30.556} = 3.647$
Row 2, Col. 3:	$\frac{(80 - 61.111)^2}{61.111} = \frac{(18.889)^2}{61.111} = \frac{356.794}{61.111} = 5.838$
Row 3, Col. 3:	$\frac{(18.333 - 10)^2}{18.333} = \frac{(8.333)^2}{18.333} = \frac{69.439}{18.333} = 3.788$
Row 1, Col. 4:	$\frac{(35 - 22.222)^2}{22.222} = \frac{(12.778)^2}{22.222} = \frac{163.277}{22.222} = 7.348$

13 Complex chi-square

Row 2, Col. 4:	$\frac{(44.444 - 25)^2}{44.444} = \frac{(19.444)^2}{44.444} = \frac{378.069}{44.444} = 8.507$
Row 3, Col. 4:	$\frac{(20 - 13.333)^2}{13.333} = \frac{(6.667)^2}{13.333} = \frac{44.449}{13.333} = 3.334$
Row 1, Col. 5:	$\frac{(30 - 19.444)^2}{19.444} = \frac{(10.556)^2}{19.444} = \frac{111.429}{19.444} = 5.731$
Row 2, Col. 5:	$\frac{(38.889 - 20)^2}{38.889} = \frac{(18.889)^2}{38.889} = \frac{356.794}{38.889} = 9.175$
Row 3, Col. 5:	$\frac{(20 - 11.667)^2}{11.667} = \frac{(8.333)^2}{11.667} = \frac{69.439}{11.667} = 5.952$

Step 5: Obtain the value of χ^2 by adding all the values obtained in Step 4.
 $3.361 + 7.348 + 6.667 + \dots + 9.175 + 5.952 = 74.696$

Step 6: Obtain the df which will be needed to evaluate χ^2 . It will always be the number of rows minus 1 multiplied by the number of columns minus 1, unless you are using the steps for a one-sample chi-square, when the df is always the number of cells minus 1.

$$df = (3 - 1)(5 - 1) = 2 \times 4 = 8$$

Step 7: Use table S6 to evaluate the significance of the χ^2 value obtained.

When $df = 8$, we read off the values 13.36, 15.51, 20.09 and 26.12. Our value exceeds all these, and so we conclude that it is significant at the 0.001 level for a two-tailed test.

Step 8: State the conclusion.

The different degree classifications for finalists in languages, maths and economics were analysed, using chi-square. The value of χ^2 was found to be significant at the 0.001 level ($\chi^2 = 74.696$, $df = 8$), and so it was concluded that the proportions of students obtaining the five classes of degree vary substantially between subjects.

Note

The test only tells us that the numbers come from distributions having different shapes. We have to decide how the shapes differ, and where the most substantial discrepancies lie by inspection of the data. In addition, take care over the interpretation of this significant result. We cannot say that 'brighter' students take maths etc., for the results may be dependent upon factors other than students' abilities, such as departmental marking policies. Chi-square is only a measure of association.

Formula

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

where E = the expected frequencies

O = the observed frequencies

Σ = the sum of



*Maud has a garden of roses
And blue fair on a lawn;*

Operation schedule 14. Spearman's rho

Data requirements

- 1 Scores must be of at least ordinal status.
- 2 Scores for comparison must be paired off in some manner. Usually they will have been obtained from one source, but this may not always be the case.
- 3 When plotted on a scattergram, the points can be in a linear or slightly curved pattern.

Example

It was noted that residents at the Heyshott Asylum frequently referred to Tennyson's poem *Maud*. Enquiries revealed that those patients most seriously afflicted with insanity (as assessed by the Heyshott Lunacy Scale, HLS) appeared to have read the poem rather often. All inmates were given a rating on the HLS, and asked how many times they had read *Maud*. The results were then used to obtain Spearman's rho.

Subject	HLS score	Maud readings
1	10	0
2	100	4
3	125	4
4	45	15
5	90	23
6	170	45
7	165	61

Because the staff anticipated that degree of insanity would be positively associated with number of *Maud* readings, they have established a one-tailed prediction.

Step 1: Draw a scattergram for the two sets of data; the vertical axis measuring one variable, and the horizontal axis the other.

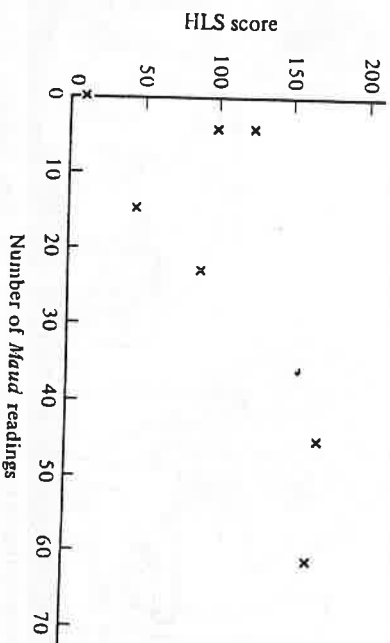


Figure 1. Scattergram showing the relationship between HLS scores and number of *Maud* readings

Step 2: Now cast the data into a table with seven columns. These will be:

- S The label for each subject, animal or source of paired scores, using numbers or letters of the alphabet.
- A The scores on one variable.
- B The scores on the second variable.
- R_A The scores from A, given ranks (see schedule 6).
- R_B The scores from B, given ranks.
- D Each value in list R_B subtracted from its partner in R_A .
- D^2 Each value in D squared.

S	A	B	R_A	R_B	D	D^2
1	10	0	1	1	0	0
2	100	4	4	2 $\frac{1}{2}$	1 $\frac{1}{2}$	2 $\frac{1}{4}$
3	125	4	5	2 $\frac{1}{2}$	2 $\frac{1}{2}$	6 $\frac{1}{4}$
4	45	15	2	4	-2	4
5	90	23	3	5	-2	4
6	170	45	7	6	1	1
7	165	61	6	7	-1	1

Step 3: Count the number of paired scores in the sample, to obtain N.
There are 7 pairs of scores; $N = 7$.

Step 4: Multiply N by its own value, twice, and then subtract its own value.
 $(7 \times 7 \times 7) - 7 = 343 - 7 = 336$

Step 5: Total the values in column D^2 .
 $0 + 2\frac{1}{4} + 6\frac{1}{4} + 4 + 4 + 1 + 1 = 18\frac{1}{2}$

Step 6: Multiply the value obtained in Step 5 by the number 6; then divide the result by the value found in Step 4.
 $\frac{18\frac{1}{2} \times 6}{336} = \frac{111}{336} = 0.3304$

Step 7: To find rho, subtract the value of Step 6 from the number 1. Retain the sign.
The result should always lie between -1 and +1.
 $1 - 0.3304 = +0.6696$

Step 8: Evaluate the significance of rho, using table S7. Follow the numbers along the row opposite the appropriate N, to find the value of rho which must be exceeded for the various significance levels given.

When $N = 7$, we read 0.714, 0.786, 0.893 and 0.929 for the one-tailed significance levels. Our obtained value of rho, 0.6696, does not exceed any of these, and so the null hypothesis cannot be rejected.

Step 9: State the conclusion.

From the results of a correlation carried out on the number of times a patient had read the poem *Maud*, and their rating on the HLS, it was found that there did not appear to be a significant association between the two variables ($\rho = +0.67, N = 7$). However, the association was positive, and the relationship may be found to be a significant one if a larger sample were used.

Formula and abbreviations

Spearman's ρ is sometimes written ρ or r_s . The formula used for calculations is
$$r_s = 1 - \frac{6\sum D^2}{(N^3 - N)}$$
 where $\sum D^2$ = the squared values of the differences between the ranked scores, totalled N = the number of paired scores

Table S7: Critical values of Spearman's ρ . ρ must be equal to or more than the stated value to be significant

N	Level of significance for one-tailed test				Level of significance for a two-tailed test			
	0.05	0.025	0.01	0.005	0.1	0.05	0.02	0.01
5	0.900	1.000	1.000	—	0.564	0.648	0.746	0.794
6	0.829	0.886	0.943	1.000	0.506	0.591	0.712	0.777
7	0.714	0.786	0.893	0.929	0.456	0.544	0.645	0.715
8	0.643	0.738	0.833	0.881	0.425	0.506	0.601	0.665
9	0.600	0.683	0.783	0.833	0.399	0.475	0.564	0.625
10	0.564	0.648	0.746	0.794	0.377	0.450	0.534	0.591
12	0.506	0.591	0.712	0.777	0.359	0.428	0.508	0.562
14	0.456	0.544	0.645	0.715	0.343	0.409	0.485	0.537
16	0.425	0.506	0.601	0.665	0.329	0.392	0.465	0.515
18	0.399	0.475	0.564	0.625	0.317	0.377	0.448	0.496
20	0.377	0.450	0.534	0.591	0.306	0.364	0.432	0.478
22	0.359	0.428	0.508	0.562				
24	0.343	0.409	0.485	0.537				
26	0.329	0.392	0.465	0.515				
28	0.317	0.377	0.448	0.496				
30	0.306	0.364	0.432	0.478				

N = the number of paired scores used.
Treat a negative value of ρ as if it were positive, when using the table, but when interpreting it don't forget that it will indicate an inverse relationship. Source: Snodgrass, table C.6

Operation schedule 15. The Pearson product-moment correlation (r)

Data requirements

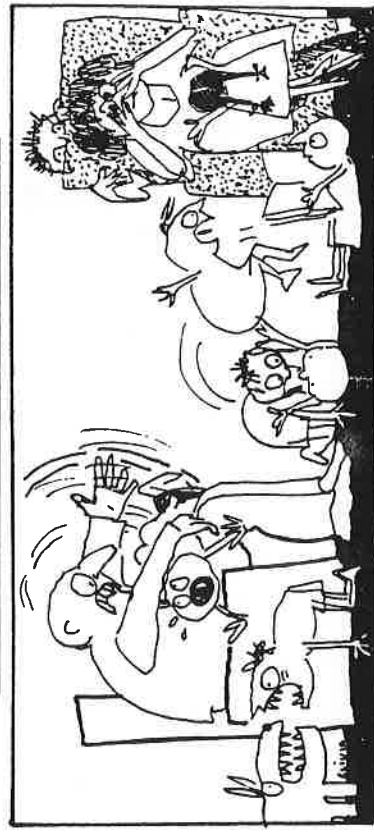
- 1 Scores for comparison must be paired off in some manner. Usually they will have been obtained from the same source, but this may not always be the case.
- 2 The relationship between the two variables must be a linear one. This can be assessed from a scattergram.
- 3 The scores must be of at least interval status.
- 4 The scores must be normally distributed.
- 5 The two sets of scores must have similar variances.

Example

A schoolgirl, recently annoyed by the antics of her small sisters, vowed that she would never inflict siblings on any child she might have when she was older. Curious to know her classmates' opinions, she asked six of them how many children they thought was a good number to have, and compared this with the number of children under the age of fifteen actually in each one's family.

Bearing in mind her own experiences, she imagined that the more siblings a person had, the fewer they would think was desirable! Thus she has predicted a negative correlation: this directional prediction constitutes a one-tailed hypothesis. The data she collected are given below:

Subject	No. of children in own family	No. of children thought ideal
Christine	1	5
David	2	3
Eric	3	4
Lynne	4	2
Margaret	4	1
Tom	5	1



MOST PARENTS HOLD THE SAME VIEW ABOUT THE IDEAL NUMBER OF CHILDREN

Operation schedules

Step 1: Draw a scattergram for the two sets of data, the vertical axis measuring one variable, the horizontal the other. At this stage, check for linearity.

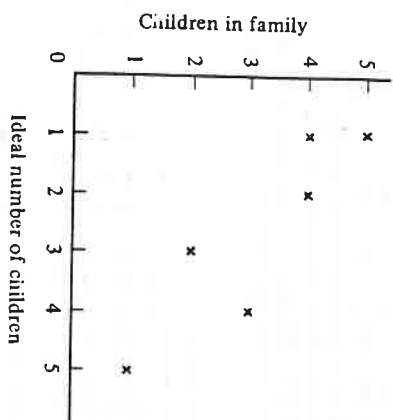


Figure 2. Scattergram showing the relationship between the number of children in a young person's family and the number that they would consider to be 'ideal'.

Step 2: Now cast the data into a table, with six columns. These will be:

- 5 The label for each subject, animal or source of paired scores, using numbers or letters of the alphabet.
 - 1 The scores on one variable (and which are added to give ΣA).
 - A^2 Each A score squared (and which added gives ΣA^2).
 - B The scores on the second variable (and which are added to give ΣB).
 - B^2 Each B score squared (and which added gives ΣB^2).
 - AB Each A score multiplied by its matching B score. (Added gives ΣAB)
- Add all the columns, apart from the first one. This is a good stage to check that the data meets all the requirements other than linearity.

Subject	A	A^2	B	B^2	AB
C	1	1	5	25	5
D	2	4	3	9	6
E	3	9	4	16	12
L	4	16	2	4	8
M	4	16	1	1	4
T	5	25	1	1	5
	19	71	16	56	40

15 The Pearson product-moment correlation

Step 3: Count the number of scores which are paired off, to get N .

There are 6 pairs of scores; $N = 6$.

Step 4: Multiply the total of the column for A^2 by N to obtain $N\Sigma A^2$.

$$71 \times 6 = 426$$

Step 5: Obtain $(\Sigma A)^2$ by squaring the total of the column for A .

$$19 \times 19 = 361$$

Step 6: Subtract the value of Step 5 from that found in Step 4.

$$426 - 361 = 65$$

Step 7: Multiply the total of the column for B^2 by N to obtain $N\Sigma B^2$.

$$56 \times 6 = 336$$

Step 8: Obtain $(\Sigma B)^2$ by squaring the total of the column for B .

$$16 \times 16 = 256$$

Step 9: Subtract the value of Step 8 from that found in Step 7.

$$336 - 256 = 80$$

Step 10: Multiply together the values obtained in Steps 6 and 9.

$$65 \times 80 = 5200$$

Step 11: Take the square root of the value found in Step 10.

$$\sqrt{5200} = 72.111$$

Step 12: Multiply the total of the column for AB by the value of N (Step 3).

$$40 \times 6 = 240$$

Step 13: Multiply together the totals of the columns for A and B .

$$19 \times 16 = 304$$

Step 14: Subtract the value of Step 13 from that of Step 12, making sure you keep the appropriate sign.

$$240 - 304 = -64$$

Step 15: Divide the value of Step 14 by the value of Step 11, to get r .

$$r = \frac{-64}{72.111} = -0.8875$$

Step 16: Evaluate the significance of r , using table S8. First obtain the value of $N - 2$, which will be used for the table, by subtracting 2 from the value of N found in Step 2.

$$6 - 2 = 4$$

Opposite $N - 2 = 4$, we read the values 0.729, 0.811, 0.9172 and 0.9741 for the 0.05, 0.025, 0.01 and 0.005 probability levels (one tailed hypothesis). Our value of 0.8875 exceeds that of 0.811 given for the 0.025 level, but not the figure of 0.9172 for the 0.01 level. Therefore we conclude that we have a negative correlation, which is significant at the 0.025 level.

Step 17: State the conclusion

From the results of the correlation carried out on the number of children in a young person's family, and the number of children they considered ideal, a significant negative association was found ($r = -0.887$, $N = 6$, $p = 0.025$). It was concluded that the more siblings there are in a family, the less likely a child of school-age is to appreciate their numbers, whilst only children appeared to consider brothers or sisters more of an asset!

Table S8 Critical values of Pearson's r. r must be equal to or more than the stated value to be significant

N - 2	Level of significance for one-tailed test	
	0.05	0.005
Level of significance for two-tailed test		
N - 2	0.10	0.01
2	0.9000	0.9900
3	0.805	0.9587
4	0.729	0.9172
5	0.669	0.875
6	0.621	0.834
7	0.582	0.798
8	0.549	0.765
9	0.521	0.735
10	0.497	0.708
11	0.476	0.684
12	0.457	0.661
13	0.441	0.641
14	0.426	0.623
15	0.412	0.606
16	0.400	0.590
17	0.389	0.575
18	0.378	0.561
19	0.369	0.549
20	0.360	0.537
25	0.323	0.487
30	0.296	0.449
35	0.275	0.418
40	0.257	0.393
45	0.243	0.372
50	0.231	0.354
60	0.211	0.325
70	0.195	0.302
80	0.183	0.283
90	0.173	0.267
100	0.164	0.254

Source: Powell, page 69

15 The Pearson product-moment correlation

Formula

The formula for Pearson's r is:

$$r = \frac{N\sum AB - (\sum A)(\sum B)}{\sqrt{[N\sum A^2 - (\sum A)^2][N\sum B^2 - (\sum B)^2]}}$$

where N = the number of pairs of scores

$\sum AB$ = the sum of the products of the paired scores.

$\sum A$ = the sum of the scores on one variable

$\sum B$ = the sum of the scores on the second variable

$\sum A^2$ = the sum of the squared scores on the A variable

$\sum B^2$ = the sum of the squared scores on the B variable

Table S9 Random numbers

19	90	70	99	00	20	21	14	68	86	14	52	41	52	48	87	63	93	95	17	11	29	01	95	80
65	97	38	20	46	85	43	01	72	73	03	37	18	39	11	08	61	74	51	69	74	39	82	15	
51	67	11	52	49	59	50	99	52	18	16	36	78	86	08	52	85	08	40	87	80	61	65	31	
17	95	70	45	80	72	68	49	39	31	56	80	30	19	44	89	84	46	06	59	73	19	85	23	
63	52	01	41	88	02	84	27	83	78	35	34	08	72	42	29	72	23	19	66	56	45	65	79	
60	61	97	22	61	49	64	92	85	44	01	64	18	39	96	16	40	12	89	88	50	14	49	81	
98	99	46	30	67	12	83	11	41	16	61	14	53	32	52	25	58	19	68	70	77	02	54	00	
76	38	03	59	63	79	44	61	40	15	86	63	59	80	02	14	53	40	65	39	27	31	58	00	
53	05	70	53	30	36	30	06	38	21	01	47	59	38	00	14	47	07	26	54	96	87	53	32	
02	87	40	41	45	47	24	49	57	74	22	13	88	83	34	32	25	43	62	17	10	97	11	69	
35	14	87	35	33	68	95	23	92	35	56	54	29	56	93	87	02	22	57	51	61	09	43	95	
84	51	33	67	13	79	33	37	55	14	44	99	81	57	39	77	32	77	09	85	52	05	30	62	
91	51	88	32	46	09	61	87	55	21	13	80	55	62	51	13	28	06	24	25	93	16	71	13	
65	09	29	75	63	20	44	90	32	64	53	89	74	60	41	27	67	63	99	61	40	38	03	93	
20	71	53	20	25	73	37	34	04	05	56	07	93	30	69	30	16	09	05	88	69	58	28	99	
01	82	77	45	12	07	10	63	76	35	19	48	27	44	87	03	04	79	88	08	13	13	85	51	
53	43	37	15	26	92	38	70	96	92	82	11	08	95	97	52	06	79	45	82	63	18	27	44	
11	39	03	34	25	99	53	93	61	28	88	12	57	21	77	52	70	05	48	34	56	65	05	61	
40	36	40	96	76	93	86	54	77	65	99	82	93	24	98	15	33	59	05	28	32	26	07	47	
99	63	22	98	18	46	23	34	27	43	11	71	99	31	85	13	99	24	44	49	18	09	79	49	
58	24	82	03	47	24	53	63	09	74	54	13	26	84	51	10	76	47	91	44	04	95	49	66	
47	83	51	62	74	22	06	34	72	52	04	32	92	09	82	21	15	65	20	33	29	94	71	11	
23	05	47	47	25	07	16	39	33	66	18	55	61	77	09	86	56	10	56	79	20	30	17	12	
69	81	21	99	31	29	70	81	63	51	70	47	14	54	31	99	74	10	52	36	87	09	11	09	
35	07	44	75	47	57	90	12	02	07	54	96	09	11	06	23	47	37	17	31	54	08	01	88	
55	34	37	72	69	33	35	72	67	47	82	80	84	25	39	77	34	55	45	70	08	18	27	38	
09	66	92	19	09	49	41	06	70	05	98	90	07	35	42	38	06	45	18	64	84	73	31	65	
86	92	48	29	00	94	09	84	38	70	63	49	30	21	30	22	00	27	69	85	29	81	94	78	
74	16	31	23	02	98	77	87	68	07	66	39	67	98	60	91	51	67	62	44	40	98	05	93	
39	60	04	59	81	00	41	86	79	79	47	53	33	38	05	68	47	22	00	20	35	55	31	51	
15	91	29	12	03	57	99	90	37	75	91	12	81	19	36	63	31	08	38	37	40	13	68	97	
90	49	22	23	62	12	59	52	57	02	55	65	79	78	07	22	07	90	47	03	28	14	11	30	
98	60	16	03	03	31	51	10	96	46	54	34	81	85	35	92	06	88	07	77	56	11	50	81	
39	41	88	92	10	96	11	83	44	80	03	92	18	66	75	34	68	35	48	77	33	42	40	90	
16	95	86	70	75	85	47	04	66	08	00	83	36	91	03	34	72	57	59	11	82	43	80	46	
52	53	37	97	15	72	82	32	99	90	06	66	24	12	27	63	95	71	76	63	89	74	44	99	
56	61	87	39	12	91	36	74	43	53	13	29	54	19	38	30	82	13	54	00	78	45	63	98	
21	94	47	90	12	77	53	84	46	47	85	72	13	49	21	31	91	18	95	58	24	16	74	11	
31	65	41	18	37	27	47	39	19	65	80	39	07	84	83	70	07	48	53	21	40	06	71		
00	83	61	22	55	34	18	04	52	35	99	01	30	98	64	56	27	09	24	86	61	85	53	45	
87	64	81	07	83	11	20	99	45	18	45	76	08	64	27	48	13	93	55	34	18	37	79	49	
20	69	22	40	98	27	37	83	28	71	69	62	03	42	73	00	06	41	74	45	89	09	39	84	
40	23	52	31	39	10	65	81	92	59	73	42	37	11	61	58	76	17	14	97	04	76	62	16	
70	76	53	97	80	59	71	74	17	32	64	63	91	08	25	75	10	24	19	23	71	82	13	74	
38	26	61	70	04	33	73	99	19	87	95	60	78	26	72	39	27	67	53	77	67	68	93		
48	67	26	43	18	77	14	77	43	96	99	17	43	76	43	00	65	98	50	45	60	33	01	07	
55	03	36	67	68	72	87	08	62	04	62	01	61	16	06	10	89	20	23	21	34	74	97		
44	10	13	85	57	73	96	07	94	32	19	59	50	88	92	09	65	90	77	47	25	76	19	33	
95	06	79	88	54	79	96	23	53	10	48	03	45	15	22	65	39	07	16	29	45	33	02	43	
68	15	54	35	02	42	35	48	96	31	95	33	95	33	22	00	18	74	72	00	18	22	85	61	
58	42	16	72	24	58	37	52	18	51	90	82	60	79	30	24	16	59	87	38	67	80	43	79	
95	67	29	83	94	69	40	06	07	46	62	08	82	54	97	80	56	95	27	62	50	90	72		
98	57	07	23	69	65	95	19	69	58	20	31	89	03	43	38	46	82	68	72	33	78	80	77	
56	69	47	07	41	90	22	91	07	12	71	59	73	05	50	08	22	23	71	77	13	13	92	66	99

Source: Powell, page 55