Achievement Standard 91584

Evaluate statistically based reports

MATHEMATICS AND STATISTICS

3.12

Externally assessed 4 credits

Statistics is the study of the collecting, organising, analysing, and interpreting of numerical information or data.

A **statistical report** is written to describe the findings of a **statistical investigation**. Most good-quality formal reports follow this format:

- Title – should fairly reflect the nature of the study
- Contents – generally only included if report is lengthy
- Summary – so nature and relevance of research is briefly explained
- Introduction – includes the statistical question(s) being investigated
- Aim – usually to answer the question that has been posed
- Data and methods
- Results and conclusions
- Appendices (usually present only if there is a lot of data)

Some reports are shorter, e.g. computer-generated reports, which often include graphs and tables of data that have been produced automatically, but omit many of the above aspects of a full report.

There are many organisations in New Zealand that produce statistical reports, e.g. market research companies such as Colmar Brunton. These reports are freely available in the media.

Questions
Reading reports

1. Roy Morgan produced the following very simple report arising from a poll.

Ans. p. 91

Roy Morgan poll late May 2012

June 4, 2012 at 10:01 pm · Filed under NZ Political Party Polls · Tagged Country Direction, NZ Political Party Polls, Roy Morgan

Polling Company: Roy Morgan Research

Poll Method: Random Phone

Poll Size: 944, of whom 906 have a party preference

Undecideds: 4.0%

Dates: 14 May 2012 to 27 May 2012

Client: Self Published

Report: Roy Morgan Website

Party Support

- National 44.0% (–0.5%)
- Labour 30.5% (+0.5%)
- Green 13.5% (–1.5%)
- NZ First 5.0% (–0.5%)
- Maori 2.0% (+1.0%)
- United Future 0.5% (–0.5%)
- ACT 1.0% (+1.0%)
- Mana 1.0% (+0.5%)

Public Poll Average »

Curia's Time & Size Weighted Public Polls Average

Sun 22/04/2012

	Vote	Seats
National	48.8%	62
Labour	28.6%	37
Green	13.6%	17
Act	0.3%	1
Maori	1.5%	3
United	0.4%	1
Mana	0.9%	1
NZ First	4.2%	0
Total	98.4%	122

Pages »

About
Public Poll Average Calculations

a. What is the title of this report?

b. What was the name of the producer of this report?

c. Who published this report initially?

d. What was the aim of the report?

e. Why do you think the report has no 'contents' as many reports do?

f. What statistical question is the report answering?

g. What was the method used by the producers of this report?

h. What type of data is shown?

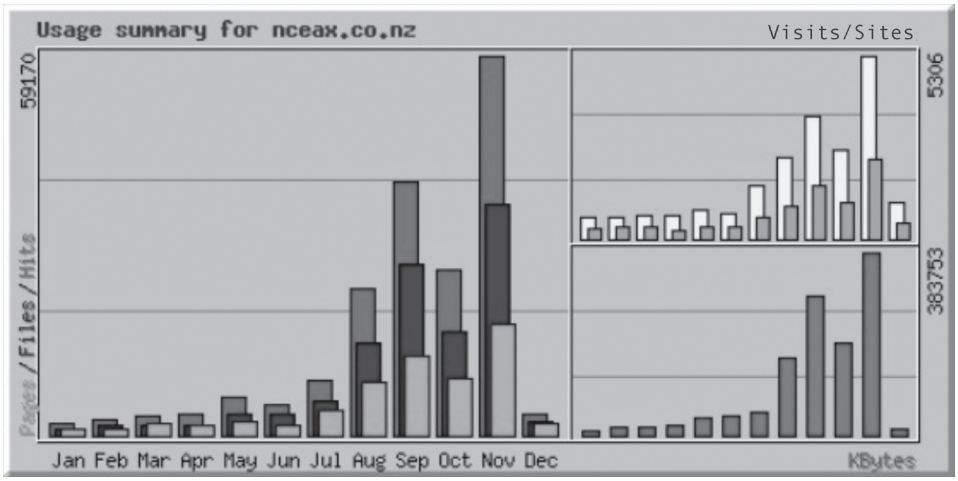
i. Why is there no written summary?

j. Is the report very clear or are there confusing parts to it?

Ans. p. 91

2. The computer-generated report below was prepared automatically for the owner of the New Zealand mathematics revision site www.nceax.co.nz

Reports (computer generated)



Summary by Month										
Month	Daily Avg				Monthly Totals					
	Hits	Files	Pages	Visits	Sites	KBytes	Visits	Pages	Files	Hits
Dec 2012	156	99	88	47	440	14691	1049	1941	2197	3447
Nov 2012	1972	1202	581	176	2309	383753	5306	17459	36084	59170
Oct 2012	830	522	288	83	1065	194425	2594	8949	16184	25732
Sep 2012	1315	890	414	118	1557	293560	3544	12423	26719	39456
Aug 2012	739	465	267	75	930	164276	2333	8307	14431	22911
Jul 2012	279	170	127	50	625	50124	1558	3937	5300	8671
Jun 2012	163	107	54	24	344	39438	725	1623	3213	4904
May 2012	188	108	66	26	358	35244	815	2057	3351	5841
Apr 2012	107	52	58	23	229	20409	696	1741	1578	3214
Mar 2012	101	54	58	21	341	16998	666	1823	1694	3139
Feb 2012	88	55	39	22	379	18796	645	1159	1607	2557
Jan 2012	58	29	31	20	285	10245	626	971	908	1825
Totals						**1241959**	**20557**	**62390**	**113266**	**180867**

a. For whom was this computer-generated report produced?

b. Which features of a statistical report seem to be missing?

c. Why did the creator of this statistical report omit these features?

d. **i.** At a glance what does this report tell you about the usage of the site?

ii. Suggest a reason for this pattern of usage.

e. What graphs and tables were utilised in this report?

f. The following comment was made about this report: 'Many of the features of a statistical report are missing; however, the creator of this report has borne in mind who the report was for and what would be looked for and has presented this information superbly.' Comment on this opinion.

Ans. p. 91

3. Popularity of *Best* wines

The *Best Wine Company* has been advertising on *Export61*, a website devoted to the liquor industry.

The purpose of the report was to show that the number of visitors to the *Best Company* site was increasing at a greater rate than the number of visitors to *Export61*.

It was also hoped to find out information about the visitors and the popularity of the various products offered by *Best*.

Period: from 1st January to 30th April 2001

Best Wine Company		
Summary statistics	**This quarter**	**Previous quarter**
Visitors to *Export61*	24 578	22 987
Visitors to *Best* website	784	611
Visitors jumped to *Best* website from *Export61*	128	144
Number of printed price lists	171	98
Time spent at *Best Company* section	**This quarter**	**Previous quarter**
Less than 1 min	216	179
1–5 min	398	294
5–10 min	133	98
More than 10 min	37	40
Products statistics (online sales)	**This quarter**	**Previous quarter**
Best Company Shiraz 1998	563	376
Best Company Chardonnay	475	324
Best Company Tawny Port	415	418
Best Company Cabernet Savignon 1997	326	207
Best Company Riesling	256	200
Best Company Liqueur Muscat	118	65
Geographical statistics – visitors from:	**This quarter**	**Previous quarter**
Germany	97	57
Japan	72	68
USA	70	46
Canada	64	26
Korea	61	54
Sweden	51	32
South Africa	43	15
Great Britain	43	49
Russia	42	53
New Zealand	37	12
Indonesia	36	61
Malaysia	24	14

Achievement Standard 91586

Apply probability distributions in solving problems

MATHEMATICS AND STATISTICS

3.14

Externally assessed 4 credits

Probability distribution of discrete random variables

In a **probability distribution**, probabilities lie between 0 and 1 and the sum of the probabilities is 1.

Discrete random variables

A **discrete random variable** is a measurement taken on the outcome of a discrete random experiment (an experiment whose outcomes can be listed or counted). For example, for the sample space generated by flipping two coins, i.e. {HH, HT, TH, TT}, the random variable X may be defined as the number of heads.

$X(\text{HH}) = 2$, $X(\text{HT}) = 1$, $X(\text{TH}) = 1$, $X(\text{TT}) = 0$ i.e. X can take on the values 0, 1, 2.

Discrete probability distributions

A table which gives all possible values of a discrete random variable, along with the associated probabilities, is called a **probability distribution**.

The probability distribution for X = the number of heads (in the two-coin example above) is shown below.

x (number of heads)	0	1	2
$P(X = x)$ (probability)	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

Note: 1. Each probability lies between 0 and 1.

2. The sum of the probabilities is always 1.

'The probability a random variable, X, takes on the value x' is written $P(X = x)$ or $p(x)$.

Example

The probability distribution of a random variable X is shown in the table alongside. X can take the values 0, 1, 2 or 3 with probabilities 0.3, 0.2, 0.2 and 0.3 respectively (note that 0.3 + 0.2 + 0.2 + 0.3 = 1).

x	0	1	2	3
$P(X = x)$	0.3	0.2	0.2	0.3

From the table, $P(X = 2) = 0.2$, i.e. the probability X is 2 is 0.2. Similarly $P(X \geq 2) = 0.2 + 0.3 = 0.5$

Means and variances of discrete random variables

The **mean of a discrete random variable** X is often called the **expected value of** X (written $E(X)$).

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

The **variance of a discrete random variable** X is often given the symbol $\text{Var}(X)$.

$$\text{Var}(X) = \sum x^2 P(X = x) - [E(X)]^2 = E(X^2) - [E(X)]^2$$

Example

Q. Find the expected value and variance of the random variable X whose probability distribution is shown in the table.

x	0	1	2	3
$P(X = x)$	0.3	0.2	0.2	0.3

A. The expected value, $E(X) = \sum x.P(X = x)$
$= 0 \times 0.3 + 1 \times 0.2 + 2 \times 0.2 + 3 \times 0.3$
$= 1.5$

The variance of X, $Var(X) = \sum x^2P(X = x) - [E(X)]^2$
$= 0^2 \times 0.3 + 1^2 \times 0.2 + 2^2 \times 0.2 + 3^2 \times 0.3 - 1.5^2$
$= 1.45$

The **standard deviation** of a random variable X is the square root of the variance of X.

$$SD(X) = \sqrt{Var\,(X)}$$

In the above example, $SD(X) = \sqrt{1.45} = 1.20$ (2 dp).

3.14

Linear combinations of independent random variables

Random variables may be combined in various ways to give other random variables, e.g. by adding or subtracting two random variables, or by multiplying a random variable by a constant.

If X and Y are two random variables, then it can be proved that the mean of the sum of X and Y equals the sum of their means. A similar result holds true for the variances if the variables are independent. Thus:

$$E(X + Y) = E(X) + E(Y)$$
$$E(X - Y) = E(X) - E(Y)$$
$$Var(X + Y) = Var(X) + Var(Y), \text{ provided } X \text{ and } Y \text{ are independent}$$
$$Var(X - Y) = Var(X) + Var(Y), \text{ provided } X \text{ and } Y \text{ are independent}$$

Note: It is important to note that the standard deviation of $(X + Y)$ is not equal to the sum of the standard deviations of X and Y, i.e. $\sigma_{X+Y} \neq \sigma_X + \sigma_Y$. In fact, $\sigma_{X+Y} = \sqrt{\sigma_X^2 + \sigma_Y^2}$, provided X and Y are independent.

Similarly, if the values of a random variable X are multiplied by 3 then the mean and spread of the random variable $Y = 3X$ would be three times as great. The variance of Y, being the square of the standard deviation (a measure of spread) would be $3^2 = 9$ times as great as the variance of X. Thus:

$$E(kX) = kE(X) \qquad Var(kX) = k^2Var(X)$$
where X is a random variable and k is a constant

Example

Q. Fred and Fiona compete in a tournament. Fred plays a game in which his mean score is 10 with a standard deviation of 2. Fiona's game score has a mean of 11 and a standard deviation of 3. Fred and Fiona each play independently and their scores are added to give their total score, S, for the tournament.

1. Find their mean total score, $E(S)$.
2. Find the standard deviation of S.

A. If X is Fred's score and Y is Fiona's score, then $S = X + Y$

1. $E(S) = E(X + Y)$
$= E(X) + E(Y)$
$= 21$ [since $E(X) = 10$ and $E(Y) = 11$]

2. Standard deviation of S is $\sqrt{Var\,(S)}$

$Var(S) = Var\,(X + Y)$
$= Var(X) + Var(Y)$ [as X and Y are independent]
$= 2^2 + 3^2$ [standard deviation of $X = 2 \therefore Var(X) = 2^2$, similarly for Y]
$= 13$

Hence standard deviation of S is $\sqrt{13} = 3.606$

If all members of a population have a constant added to them, the mean will change by the same amount. The spread, however, remains the same (variance and standard deviation are unaltered). Thus if X is a random variable and c is a constant:

$$E(X + c) = E(X) + c \qquad Var(X + c) = Var(X)$$

Combining these results with the previous table gives

$$E(aX + b) = aE(X) + b \qquad Var(aX + b) = a^2Var(X)$$

where a and b are constants

A **linear combination** of two random variables X and Y is any expression of the form $aX + bY$ where a and b are constants. Expected values and variances of linear combinations obey the following rules:

$$E(aX + bY) = aE(X) + bE(Y)$$

$$Var(aX + bY) = a^2Var(X) + b^2Var(Y), \text{ if } X \text{ and } Y \text{ are independent}$$

Note that a and b may be negative, e.g. $Var\,(2X - 3Y) = 2^2\,Var\,(X) + (-3)^2\,Var(Y)$

$$= 4Var(X) + 9Var(Y)$$

3.14

Questions
Probability distribution of discrete random variables

1. A car sales yard is running a promotion.

Year 2007 Ans. p. 101

- Each buyer spins a wheel, which is divided into 36 equally sized sections.
- Each section on the wheel represents either a \$500, \$1 000, or \$5 000 discount on a car purchase.
- Twenty-four randomly allocated sections have \$500 marked on them.
- Ten randomly allocated sections have \$1 000 marked on them.
- The remaining two sections have \$5 000 marked on them.

The table below shows the probability distribution of the random variable D, the amount of discount won.

d (\$)	500	1 000	5 000
$P(D = d)$	$\frac{24}{36}$	$\frac{10}{36}$	$\frac{2}{36}$

Suppose they sell 15 cars at the yard over the weekend that the promotion runs.

Calculate the total amount of money the company would expect to discount.

Year 2011 Ans. p. 101

2. In 2006, a survey was conducted on households in Hamilton, Canada.

a. Let the random variable X represent the number of cars in a randomly chosen household at the time of the survey. The survey gave the following probability distribution for X.

x	0	1	2	3	4
$P(X = x)$	0.059	0.383	0.377	0.153	0.028

Find the expected value of X.

b. A similar survey conducted in Hamilton, New Zealand, gave the following probability distribution for Y the number of cars in a randomly chosen household in Hamilton, New Zealand, at the time of the survey.

y	0	1	2	3	4
$P(Y = y)$	0.09	0.298	0.423	0.119	0.07

Find the probability that a randomly selected household in Hamilton, New Zealand has more cars than a randomly selected household in Hamilton, Canada.

Answers and explanations

Achievement Standard 91584 (Mathematics and Statistics 3.12): Evaluate statistically based reports

3.12 Reading reports

p. 1

1. a. Roy Morgan poll late May 2012. (A)
   b. Roy Morgan Research. (A)
   c. Roy Morgan Research. (A)
   d. To provide an indication of political party support in New Zealand in late May 2012. (A)
   e. Probably because the report is so brief that one is unnecessary. (A)
   f. What are the political party preferences of the New Zealand public in late May 2012? (A)
   g. Random phone. (A)
   h. The raw data has been processed to give percentages. (A)
   i. There was none required. (A)
   j. *Answers will vary; an example is given.*

   Presumably this report is part of a series of polls whose format is generally understood. As a stand-alone document, some details should be more fully explained, e.g. the small positive or negative percentages alongside the party support figures (assumed to indicate the percentage change in support from a previous poll, but how long ago was the previous poll conducted?). Without this explanation, a figure such as 1.0% (+1.0%) for ACT is confusing.

   More explanation could also be provided for the Curia average – how many polls were averaged, and how recently? How valid is the time weighted method? (M/E)

2. a. The owner of the nceax website. (A)
   b. Contents, summary, introduction, aim, methods, conclusions, appendices. (A)
   c. The report was prepared regularly for the owner using automatic computer methods. The graphs would be considered to 'speak' for themselves and extra reporting would be unnecessary for the owner, and add costs. (M)
   d. i. Usage was quite low until June when the usage increased very rapidly reaching a peak in September. It fell back in October but greatly increased in November. It was very low in December. (A)
      ii. Students were not revising until June when they would be having school exams. However, after that the exam season would have been in full swing. The MCAT exam for Year 11 students took place in September and many schools would be also be having exams then. The site caters for CIE students who have their examinations before NCEA students in October then in November the site would have been massively used by NCEA students. These reasons seem plausible; however, they may not be true. (M)
   e. Bar graphs showing important monthly information, tables showing this information as averages and totals. (A)
   f. *Answers will vary; an example is given.*

   It is assumed that this information is for the owner of the site who will want to see the monthly usage and the year-long profile of usage. It is assumed that the owner commissioned the report, so it contains all details necessary for his/her use.

   The graphs provide an excellent summary of a lot of information, affording easy comparisons from month to month, although a clear vertical scale is not provided (to ensure scale starts at zero, etc.), so that it is not clear whether the graphs show daily averages or monthly totals. The table is clearly set out, with colour coding adding to the ease of reading.

   All useful information appears easily accessible and attractively presented, so I would tend to agree with the comment. (M/E)

3. a. The *Best Wine Company*. (A)
   b. This information was not given (although possibly by *MarkRes* who will be engaged to do further research). (A)
   c. The *Export61* site. (A)
   d. Tables attractively set out and showing important information comparing this quarter's statistics with the previous quarters; bar graphs allowing quarterly comparisons in visitor numbers for the two sites (these could be misleading as the scales on each bar graph will be different). (A/M)
   e. Quarter spelt 'quater'. Spelling errors give an unprofessional look to a report as it appears that the report has not been thoroughly checked. (Alternatively this could be a 6-character limitation in the software.) (A/M)
   f. Visitor numbers from some countries fell, e.g. Indonesia, Great Britain and Russia. There was a fall in sales of Tawny port and there was fall in the number of people who jumped to the *Best Wine* site from the *Export61* site (although this could be due to people accessing the site directly). (A/M)

3.12 Statistical questions

p. 6

1. 'Are pigs in the Waikato kept in humane conditions ?'; pigs in the Waikato. (A)
2. 'Should expelled list MPs have to resign?'; New Zealanders (above a set age, i.e. not young children). (A)
3. 'At what age of starting the consumption of marijuana to the point that a user is addicted is there a permanent loss in intelligence?'; marijuana users who started before 18 years and were addicted by the age of 38. (A)

4. a. There is no stated statistical question but one is implied by the objective: 'Does chocolate eating have an effect on the risk of developing cardiometabolic disorders?'
   b. Adult human chocolate consumers. (A)
5. 'Are dog owners more likely than cat owners to be single?'; New Zealand cat and dog owners. (A)

   An initial reading suggests that the report is only concerned with cat and dog owners; however, reading the above extract shows that people had to identify themselves as a cat or dog owner so there were other questions in the study as well as the above such as: 'What percentage of New Zealanders own a cat?' and 'What percentage of New Zealanders own a dog?' If this is true then the target population would be all New Zealanders. (A/M)

3.12 Planning and data

p. 10

1. a. Yes, as there is variability in the results of the investigation. (A)
   b. Answers will vary, for example 'What are the literacy levels of the incoming Year 9 students at this school?' (A)
   c. The incoming Year 9 pupils at the school. (A)
   d. Since the literacy levels of all Year 9 pupils will be determined this is a census. (A)
   e. The independent variables are the questions on the PAT tests, the response variables are the student answers. (A)
   f. Independent data are the questions on the PAT tests. These are nominal and are not measured. The measurement of the responses is done by the marking of the PAT tests. The marks create a set of discrete data. (A)
   g. If the difficulty level of the tests was able to be manipulated then this would be an experiment; however, if no manipulation of the tests was possible then it would be an observation. (A/M)
   h. The incoming Year 9 pupils. (A)
   i. Primary. (A)
2. a. There is variability in the results of the investigation. (A)
   b. 'What are the reported and actual tooth-brushing habits of 30-month-old children and how much fluoride is ingested from toothpaste?' (A)
   c. 30-month-old children in the North West region of England. (A)
   d. Sampling, as only 50 children were observed. (A)
   e. Independent: Questions about tooth-brushing habits, fluoride levels in toothpaste; Response: Answers to questions, amount of fluoride retained. (A)
   f. Independent: Questions (nominal) and the amounts of toothpaste (continuous) Response: Answers to the question (nominal); weight of fluoride retained (continuous). Measurement of answers to questions would either by the investigator writing them in or they are generated automatically. The amounts of toothpaste could be done by weight, the method for measuring the amount of fluoride retained is not specified. It probably involves a complicated chemical procedure. (A)
   g. Experiment. (A)
   h. 50 selected children. (A)
   i. Primary. (A)
3. a. There is variability in the results of the investigation. (A)
   b. Answers will vary, e.g.'What is the best type of exercise for losing weight?' (A)
   c. Obese, middle-aged people without diabetes. (A)
   d. Sampling required (200 people). (A)
   e. Independent : type of exercise; Response: weight loss. (A)
   f. Independent data is nominal and requires no measurement, the response variable 'weight loss' can be calculated by scale and calculation. (A)
   g. Experiment. (A)
   h. The 200 people involved. (A)
   i. Primary. (A)
4. a. There is variability in the results of the investigation. (A)
   b. Answers will vary, e.g. 'Does chocolate have negative effects on the risk of cardiometabolic disorders?' (A)
   c. Human adults. (A)
   d. Sampling. (A)
   e. Independent: chocolate consumption; Response: disease rate. (A)
   f. Although the full report may specify the types of data and the measurement of them, this is not clear from the extract. Any definitive answer would be speculation. (A/M)
   g. Observation (although a case can be made for experiment). (A)
   h. Adults in the investigation. (A)
   i. Secondary. (A)
5. a. There is variability in the results of the investigation. (A)
   b. Answers will vary, e.g. 'Is there an association between sleep in childhood and obesity?' (A)
   c. People in New Zealand. (A)
   d. Sampling. (A)
   e. Independent: Amount of childhood sleep; Response: Adult body mass index. (A)
   f. Independent is continuous and would have been measured by a clock then calculation. Response is also continuous and is measured by calculation of body mass index. (A)
   g. Observation, as the amount of sleep cannot be controlled by investigator. (A)
   h. 1 000 Dunedin people born between 1972 and 1973. (A)
   i. Probably primary if the calculating of sleep times and body mass index was done by the researchers. (A/M)

3.12 Sampling and sampling errors

p. 16

1. a. 131 (A)
   b. There is no indication of the process used. (A)
   c. No. 131 is less than ideal for a population of 250 million. (A)
   d. *Answers will vary; examples follow.*

      Non-response, where participants drop out / get lost from the study over the 25 years.

      People involved will not form a cross-section of divorced families, as only cooperative and willing families will participate, whose outcomes would possibly be different from other families going through divorce. (M)
2. a. No – although the sample sizes would be assumed to be large as the report combines data from several studies. (A)
   b. Randomised trials and cohort, case-control, and cross sectional studies carried out in human adults. (A)
   c. *Answers will vary; an example is given.*

      Recent data is not included as this is a meta-analysis of older studies. (M)