Chapter 4

Probabilities & Proportions

Aims

The aim of this chapter is to introduce us to basic ideas about probabilities and how to work with probabilities and proportions.

By the end of this chapter you should know how to calculate probabilities and proportions:

- through tables of counts
  - simple probabilities/proportions
  - joint probabilities/proportions
  - conditional probabilities/proportions
- for independent events

What are probabilities?

A probability is a **number between 0 and 1** that quantifies uncertainty.

The **probability that an event *A* occurs** is written as **pr(*A*)**.

Examples:

I toss a fair coin (where ‘fair’ means ‘equally likely outcomes’)

- What are the possible outcomes? **H & T**
- What is the probability it will turn up heads? **1/2**

I choose a person at random and check which eye she/he winks with

- What are the possible outcomes? **L & R**
- What is the probability they wink with their left eye? **?**

For **equally likely outcomes**:

$$\text{pr}(A) = \frac{\textbf{number of outcomes in } \boldsymbol{A}}{\textbf{total number of outcomes}}$$

The probability of getting a four when a fair die is rolled is **1/6**

Probabilities and proportions are numerically equivalent.

- The proportion of New Zealanders who are left handed is 0.1.
- A randomly selected New Zealander is left handed with a probability of 0.1.

Tables of Counts

House Sales

A sample of 343 houses in Auckland city sold in April 2005 is shown in the table below.

	Weeks on the market			
Sale price	Less than 3	3 – 6	More than 6	**Total**
Under $300,000	12	5	11	28
$300,000 - $600,000	70	69	47	186
Over $600,000	52	47	30	129
Total	134	121	88	343

Source: *Real Estate Institute of New Zealand 2005*

Let *A* be the event that a sale is made within 3 weeks
B be the event that a sale is over $600,000

What proportion of these sales were:

a. over $600,000?

pr(B) = 129/343 = 0.38

b. not over $600,000?

pr(B) = (28+186)/343 = 0.62

c. made in 3 or more weeks?

pr(A) = 1 – 134/343 = 0.61

d. made within 3 weeks and sold for over $600,000?

pr(A and B) = 52/343 = 0.15

e. made within 3 weeks or sold for over $600,000?

pr(A or B) = (134+129-52)/343 = 0.62

f. on the market for less than 3 weeks given that they sold for over $600,000?

52/129 = 0.40

Conditional Probabilities

The sample space is reduced.

Key words that indicate conditional probability are: **given that, of those, if, assuming that**

"The probability of event *A* occurring given that event *B* has already occurred" is written in shorthand as: **pr(A|B)**

g. What proportion of the houses that sold in less than 3 weeks, sold for more than $600,000?

pr(B|A) = 52/134 = 0.39

Filled jobs by industry and type

Filled job numbers are in thousands.

	Type			
Industry	Working Owner	Part time	Full time	**Total**
Manufacturing	15	19	212	246
Retail Trade	30	90	112	232
Property & Business Services	22	67	122	211
Health & Community Services	8	86	88	182
Education	2	37	87	125
Construction	20	7	84	111
Wholesale Trade	10	15	86	111
Transport, Storage & Communication	6	28	66	100
Accommodation, Cafes & Restaurants	10	57	33	99
Personal & Other Services	5	18	38	61
Government Admin. & Defence	0	7	48	55
Finance & Insurance	3	9	38	49
Cultural & Recreational Services	2	16	25	43
Forestry & Mining	2	1	10	13
Electricity, Gas & Water	0	1	7	8
Total	132	458	1056	1646

Source: *Statistics New Zealand March 2005*

Let *PT* be the event of working as a part time employee
FT be working as a full time employee
R be in the retail trade
A be in the accommodation, cafes & restaurants industry
E be in education

a. What proportion of workers were part time employees?

pr(*PT*) = 458/1646 = 0.28

b. The industry with the highest proportion of part time workers in March 2005 was accommodation, cafes & restaurants. What was this proportion?

pr(*PT* | *A*) = 57/99 = 0.58

c. What proportion of workers were in the retail trade?

pr(*R*) = 232/1646 = 0.14

d. What proportion of workers were full time employees working in education?

pr(*E* and *FT*) = 87/1646 = 0.05

Response Rates by Survey Format

A study was conducted to compare response rates for online and paper survey methods.

First year students were randomly selected to receive a copy of the survey as a paper or web-based form.

Format	Responses	Nonresponses	**Total**
Paper only	325	1153	1478
Paper with web option	352	1116	1468
Web-only with response incentive	125	608	733
Web-only without response incentive	146	591	737
Total	948	3468	4416

Source: *Research in Higher Education*

a. What proportion of the students received an incentive and responded?

125/4416 = 0.03

b. What was the overall response rate to the survey?

948/4416 = 0.21

c. Which format had the highest response rate?

325/1478 = 0.22

352/1468 = 0.24

125/733 = 0.17

146/737 = 0.20

Paper with web option

Building a Table from a Story

HIV Transmission

A European study on the transmission of the HIV virus involved 305 heterosexual couples. Originally only one of the partners in each couple was infected with the virus. There were 171 couples that always used condoms. From this group, 3 of the non-infected partners became infected with the virus. Of the 134 couples who did not always use a condom, 16 of the non-infected partners became infected with the virus.

Let C be the event that **the couple always used condoms**
I be the event that **the non-infected partner became infected**

	C	C	Total
I	3	16	19
I	168	118	286
Total	171	134	305

Source: *Journal of Acquired Immune Deficiency Syndromes 1993*

a. What proportion of the couples always used condoms?

pr(C) = 171/305 = 0.56

b. Of the couples who always used condoms, what proportion had a non-infected partner who became infected?

pr(I|C) = 3/171 = 0.02

i.e., an estimate of the *risk* of transmission given that a condom was always used was 2%.

c. Of the couples who did not always use condoms, what proportion had a non-infected partner who became infected?

pr(I|C) = 16/134 = 0.12

i.e., an estimate of the *risk* of transmission given that a condom was not always used was 12%.

d. For a couple who did NOT always use a condom, how does the risk of the non-infected partner becoming infected compare with that for a couple who always used a condom?

pr(I|C) / pr(I|C) = 0.12/0.02 = 6 times

i.e., for a couple who do NOT always use a condom, an estimate of the risk of transmission was 6 times the risk for a couple who always use a condom. (i.e., the *relative risk* is 6.)

Chances of Getting the Death Penalty

University of Florida sociologist, Michael Radelet, believed that, in Florida, the chance of getting the death penalty if you had killed a white person was three times the chance of getting the death penalty if you had killed a black person.

In a study Radelet classified 326 murderers by race of the victim and type of sentence given to the murderer. 36 of the convicted murderers received the death sentence. Of this group, 30 had murdered a white person whereas 184 of the group that did not receive the death sentence had murdered a white person.

Let W be the event that **the victim is white;**

D be the event that **the sentence is death.**

	W	W	Total
D	30	6	36
D	184	106	290
Total	214	112	326

Source: *American Sociological Review 1981*

a. What is the probability of a murderer receiving the death sentence?

pr(D) = 36/326 = 0.11

b. What is the probability of a murderer receiving the death penalty given that the victim was white?

pr(D | W) = 30/214 = 0.14

i.e, an estimate of the *risk* of getting the death sentence, given that the victim was white, was 0.14.

c. What is the probability of a murderer receiving the death penalty given that the victim was black?

pr(D | W) = 6/112 = 0.05

i.e, an estimate of the *risk* of getting the death sentence, given that the victim was black, was 0.05.

Radelet's "*three times*" comes from comparing the answers in **b** and **c**. This is *relative risk.*

Raising EEO Issues

As part of its responsibilities to EEO (Equal Employment Opportunities), a NZ Government Department surveyed its employees. One question in the survey asked: "If an EEO issue was concerning you, would you raise it with your manager?"

52.5% of those surveyed were males. Of the males, 62% replied "Yes" and 13% replied "No". Of the females, 55% replied "Yes" and 17% replied "No". The remainder of both groups replied "Don't know".

	M	*F*	**Total**
Y	62% of 52500 = 32550	55% of 47500 = 26125	58675
N	13% of 52500 = 6825	17% of 47500 = 8075	14900
?	13125	13300	26425
Total	52500	47500	100000

Of those who replied "No", what proportion were female?

pr(*F* | *N*) = 8075/14900= 0.54

Imperfect Testing

ELISA HIV Test

For people who are HIV positive: 99.7% test positive

pr(Test +ve|HIV+) = 0.997

For people who are HIV negative: 0.3% test positive (false positive)

pr(Test +ve|HIV-) = 0.003

It is estimated that 0.1% of the New Zealand population are HIV positive.
Suppose that a person is picked at random from New Zealand.

Let HIV+ be the event of having HIV;
Test +ve be the event of the ELISA test being positive.

	HIV+	HIV-	Total
Test +ve	997	2997	3994
Test -ve	3	996003	996006
Total	1000	999000	1000000

a. Of those who test positive, what proportion are actually HIV+ ?

pr(HIV+|Test +ve) = 997/3994 = 0.250

b. Of those who test positive, why do so few actually have HIV?

Because of the very small number of **HIV+** people overall. A very high percentage of a very small number (1000) gives a **small** number (997)!

c. In 1988 it was reported that an estimated 80% of drug addicts in New York City were HIV positive. What is the probability that, in 1988, if we had randomly selected a New York drug addict had HIV given that he/she tested positive?

pr(HIV+|Test +ve) = 7976/7982 = 0.999

	HIV+	HIV-	Total
Test +ve	7976	6	7982
Test -ve	24	1994	2018
Total	8000	2000	10000

Note: in this case a small percentage of a small number incorrectly test positive.

- If the prevalence of HIV is very low then the majority who test positive would be false positives.
- If the prevalence of the HIV high, the majority who test positive would be true positives.

Tax Audits

Suppose that the incidence of tax evasion is 1 in 100 firms, that 90% of all cases of tax evasion are detected by an automated system and of those firms that are not evading tax, the system indicates that 5% are possibly evading tax.

Find the probability that a firm has actually evaded tax when the system indicates tax evasion.

Let T be the event the firm evades tax
D be the event evasion is indicated by the system

	T	T	Total
D	90	495	585
D	10	9405	9415
Total	100	9900	10000

pr(T|D) = 90/585 = 0.15

Statistical Independence

Events A and B are statistically **independent** if

$$pr(A \mid B) = pr(A)$$

From this we are able to show that:

If A and B are statistically **independent**, then

$$pr(A \text{ and } B) = pr(A) \times pr(B)$$

If the n events $A_1, A_2, \ldots A_n$ are **mutually independent** then

$$pr(A_1 \text{ and } A_2 \text{ and } \ldots \text{ and } A_n) = pr(A_1) \times pr(A_2) \times \ldots \times pr(A_n)$$

Challenger Space Shuttle

The probability of a single field joint succeeding was given as 97.7%. For the whole system of field joints to succeed, all 6 field joints had to succeed. The success of one field joint was independent of the success of any other field joint.

Find the probability of the system failing due to field joint failure.

pr(one field joint ok) = 0.977

pr(all 6 field joints ok) = **pr(1st ok)** × … × **pr(6th ok)** **(by indep)**

= 0.977 × … × **0.977**

= 0.87

pr(system fails) = pr(at least 1 field joint fails)

= 1 – pr(all 6 field joints ok)

= 0.13

White Toyotas

According to one New Zealand survey, 26% of cars are white and 27% of cars are made by Toyota. Now if these characteristics appear independently, and there is strong evidence that they do, the percentage of cars in New Zealand which are white Toyotas is:

pr(White and Toyota) = pr(White) × pr(Toyota) (by indep)

= 0.26 × 0.27

= 7%

Probabilities and Proportions

Probabilities and proportions are numerically equivalent

- Proportions come from data
- Proportions can be estimates of future probabilities
- Probabilities involve a random event

For equally likely outcomes:

$$\text{pr}(A) = \frac{\text{number of outcomes in } A}{\text{total number of outcomes}}$$

Conditional probabilities & proportions:

- Key words that indicate conditional situations are: given that, of those, if, assuming that
- "The probability of event A occurring given that event B has already occurred" is written in shorthand as: $\text{pr}(A \mid B)$

Risk / Relative risk:

- See 'HIV Transmission' and 'Chances of Getting the Death Penalty' exercises on page 4.

Statistical independence:

- Events A and B are statistically independent if $\text{pr}(A \mid B) = \text{pr}(A)$
- If A and B are statistically independent, then $\text{pr}(A \text{ and } B) = \text{pr}(A) \times \text{pr}(B)$
- If the n events $A_1, A_2, \ldots A_n$ are mutually independent then
  $\text{pr}(A_1 \text{ and } A_2 \text{ and } \ldots \text{ and } A_n) = \text{pr}(A_1) \times \text{pr}(A_2) \times \ldots \times \text{pr}(A_n)$

Sample Exam / Test Questions

The following data comes from a study into child obesity in New Zealand. 871 children were selected in 1989 and another 894 children in 2000. All the children were from the Hawkes Bay region. Questions **1** to **3** refer to the 1989 sample data, seen in the table below.

	Weight Classification			
Ethnicity	Normal	Overweight	Obese	**Total**
European	535	43	4	582
Maori	168	42	15	225
Pacific Island	20	6	1	27
Other	31	5	1	37
Total	754	96	21	871

Source: *Journal of Paediatrics and Child Health, January 2004*

1. The percentage of children who were Maori is approximately:

- **a.** 86.6%
- **b.** 19.3%
- **c.** 30.1%
- **d.** 25.8%
- **e.** 29.8%

2. The percentage of children who were overweight Europeans is approximately:

- **a.** 4.9%
- **b.** 66.8%
- **c.** 7.4%
- **d.** 11.0%
- **e.** 44.8%

3. The percentage of Europeans who were overweight is approximately:

- **a.** 4.9%
- **b.** 66.8%
- **c.** 7.4%
- **d.** 11.0%
- **e.** 44.8%

From the data collected from 894 children in 2000, 313 of the 633 children within the normal weight range were females. Of the 450 females, 99 were overweight. 33 males were obese.

4. The percentage of obese children in the 2000 sample who were female is approximately:

a. 7.9%

b. 50.3%

c. 53.5%

d. 4.3%

e. 8.4%

5. An investment company randomly surveyed 500 potential investors, and asked them to categorise themselves as either a “High Risk” taker or a “Low Risk” taker. 68% of the potential investors categorised themselves as a “Low Risk” taker. The company also asked if the person surveyed was less than 35 years of age or not. 200 people answered that they were less than 35 years of age and would categorise themselves as a “Low Risk” taker. 65 people answered that they were aged 35 years or more and would categorise themselves as a “High Risk” taker.

The probability that a randomly selected potential investor is under 35 years of age, given that they would categorise themselves as a “High Risk” taker, is approximately:

a. 0.5938

b. 0.3170

c. 0.1900

d. 0.3220

e. 0.4063

6. According to a recent survey into cellphone advertising, it was found that response to the advertising was independent of gender. 44% of people said they bought the advertised brand after receiving the advertising message. Assume that half the population are females.

If a person who receives one of these messages is chosen at random, what is the probability that they are a female who buys the brand afterwards?

pr(Female and buys) = pr(Female) × pr(Buys) (by indep)

= 0.5 × 0.44

= 0.22

Answers: (See Section D: Lecture and Tutorial Answers or the fill-ins on Cecil)

Tutorial

1. In 1995 there were 2011 students enrolled in either 528.181 or 528.188 (Stage I Statistics) at the City campus. The numbers of female and male students are given in the following table.

	Females	Males	**Total**
528.181	604	593	1197
528.188	387	427	814
Total	991	1020	2011

(a) One of the 2011 students is chosen at random. What is the probability that the student chosen is:

(i) a male taking 528.181?

(ii) a female?

(iii) a female taking 528.188?

(b) Given that a student is taking 528.188, what is the probability that they are male?

(c) What is the probability that a randomly chosen male student is taking 528.188?

2. The medical records of a group of diabetic patients presenting at a clinic showed that 50 presented as serious cases, while 36 presented as mild cases. Of the 31 patients aged under 40, 16 presented as mild cases.

(a) Present this information in the table below.

(b) A patient is chosen at random. Find the probabilities that:

(i) the patient is under 40 and has a mild case.

(ii) the patient is at least 40 years old or has a serious case.

(iii) the patient has a serious case and is at least 40 years old.

(c) Of those presenting with serious cases, what proportion are aged under 40?

(d) Of those aged at least 40, what proportion present with mild cases?

3. A bank classifies borrowers as high-risk or low-risk. Of all its loans, 5% are in default. Forty percent (40%) of those loans in default are to high-risk borrowers, while 77% of loans not in default are to low-risk borrowers.

(a) Complete the table.

		10000

(b) What percentage of loans is made to borrowers in the high-risk category?

(c) What is the probability that a high-risk borrower will default on his or her loan?

4. A drinking pattern found by a survey is that 19% of male drinkers and 10% of female drinkers drink alcohol daily. Also, 51% of all drinkers are male (a ‘drinker’ was defined as someone who had consumed alcohol in the previous 12 months).

The probability that a randomly selected drinker from this survey who drinks alcohol daily is female is:

(1) 0.3448

(2) 0.3358

(3) 0.0490

(4) 0.1459

(5) 0.2041

Answers: (See Section D: Lecture and Tutorial Answers or the fill-ins on Cecil)