Chapter 3

Point and interval estimation for categorical data

3.1 Objectives

1. To understand the binomial probability distribution, and to check and interpret the necessary assumptions.
2. To form point estimates of proportions for categorical data and to choose and compute appropriate interval estimates.
3. To use R to calculate point and interval estimates of proportions.

3.2 Introduction

Statistical inference is the process by which we draw conclusions about a *population* or a *process* on the basis of data — a process that relies on probability theory. This chapter assumes basic knowledge about probability (e.g. finding the joint probability of independent events, such as throwing a 6 with a die and obtaining a head with a coin toss).

There are two main kinds of inference, the first of which has two subclasses:

- Estimation
    1. *Point estimation:* what is the best estimate of a parameter?
    2. *Interval estimation:* an interval within which we're quite confident that the parameter lies,; referred to as a **confidence interval**.
- Hypothesis testing. A hypothesis about a parameter is claimed: how much evidence is there against this hypothesis?

In this chapter we focus on point and interval estimation for categorical data. We will tackle hypothesis testing in the next chapter. As usual, we will use R largely as a calculator, albeit one that happens to know a lot about probability.

The population under study is not always easy to define. We can think of taking a sample of trees from a forest; here the population is well defined (it's all the trees), and we can take a sample using one of a number of well-established sampling approaches. On the other hand, we might be sampling the outcome of a manufacturing process – for example, the construction of vehicles on an assembly line. The population is harder to pin down: it may be all the cars made on the line or all the cars made in the line (and lines like it) or all *possible* cars made on the line, and so on.

What we choose to believe about the population from which our sample came will affect the degree to which our statements can be generalised to represent the population.

3.3 Random variables

A simple case to consider is when the outcome being studied is categorical, and has two possible states, such as the result of the tossing of a coin. We can use a *random variable* to represent this outcome. Here it is a *binary* outcome. Formally, a random variable is a function that maps from a sample space to the real line. Informally, we can think of a random variable as a measurement taken on one or more sampling units. The height of a tree is a random variable, as is the mean height of several trees. The outcome of a coin toss is a random variable, as is the count of tails from a sequence of tosses.

In general, we will symbolize a random variable by X, and measurements (or realisations) of it, for example on the ith unit, by x_i, $i = 1, \ldots, n$. For n identical binary processes (such as a coin toss), we specify a distribution for the random variable X (in this case, a binomial distribution) by

$$X \sim \text{Bi}(n, p). \tag{3.1}$$

Here we are claiming that X, which represents the process from which the sample came, is distributed according to the binomial distribution, and that X represents the number of successes of n independent trials, each with probability of success p. Success can be defined as either of the two outcomes.

Example **Salk vaccine I**

In Salk's *randomised controlled double-blind* study, 57 of the 200,745 vaccinated children developed polio. What is an estimate of the probability (p) that a randomly-selected vaccinated child will develop polio?

We need a model, to provide us with a framework for interpreting the parameters. We use X to denote the random variable: the number of children that develop polio. Here, x, the realisation of X, is 57. We will assume that each child has the same probability of contracting polio, and that the contraction of polio for a child is independent of whether any other child contracts polio. Neither assumption is necessarily true, but without further information we might have to proceed regardless.

Note that failure of the assumptions to be true may not invalidate the estimate as a sound estimate of the population probability, but it may render it irrelevant. For example, there may be two sub-populations of children, one with $p = 0.0001$, and the other with $p = 0.01$. Then it would probably be of greater interest to try to determine which subpopulation a child belongs to, rather than estimate the overall proportion. Also, if the assumption of independence is untrue, and (for example) the probability of contracting polio is much higher if someone that you know has contracted polio, then again the average rate is less informative than a rate that conditions on these things.

3.3.1 Probabilistic definitions

When we select a probability distribution model for our situation, we implicitly select the parameters of interest, and provide ourselves with tools that we can use to manipulate the model.

Probability models can be represented a number of different ways. Each has its value depending on circumstances.

pmf The probability mass function $f_X(x)$ provides the probability that a realization of the random variable X takes on the value x. In R, the *** pmf is written `d***()`, e.g. `dbinom()` for the binomial pmf.

cdf The cumulative distribution function $F_X(x)$ provides the probability that a realisation of the random variable X takes on the value x or some value less than x. In R, the *** cdf is written `p***()`, e.g. `pbinom()` for the binomial cdf.

The cdf function is invertible, which can be very useful. In R, the *** inverse cdf is written `q***()`,e.g. `qbinom()` for the binomial inverse cdf.

The binomial distribution is characterised by two parameters: the number of trials n, which is assumed to be fixed and known, and the probability of success p. Having chosen the binomial distribution, the task is then to estimate p.

Example We can obtain values for the pmf and cdf for $X \sim \text{Bi}(10, 0.5)$ and the cdf for $X \sim \text{N}(0, 1)$ in R. In the `plot` command, `type = "h"` results in "histogram-like vertical lines".

```
> x1 <- 0:10
> pmf <- dbinom(x1, 10, 0.5)
> pmf

 [1] 0.0009765625 0.0097656250 0.0439453125 0.1171875000 0.2050781250
 [6] 0.2460937500 0.2050781250 0.1171875000 0.0439453125 0.0097656250
[11] 0.0009765625

> plot(pmf ~ x1, type = "h")
```

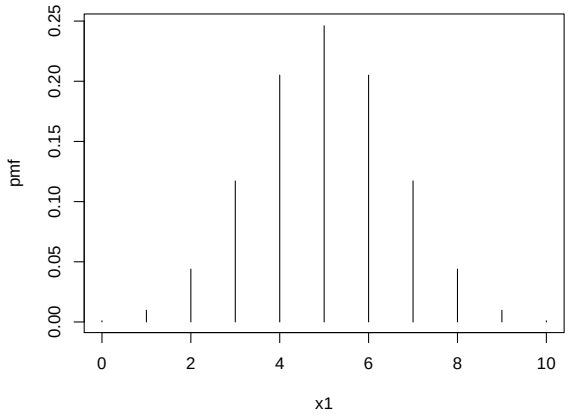

Figure 3.1: Binomial pmf

```
> cdf <- pbinom(x1, 10, 0.5)
> cdf

 [1] 0.0009765625 0.0107421875 0.0546875000 0.1718750000 0.3769531250
 [6] 0.6230468750 0.8281250000 0.9453125000 0.9892578125 0.9990234375
[11] 1.0000000000

> plot(cdf ~ x1, type = "h")
```

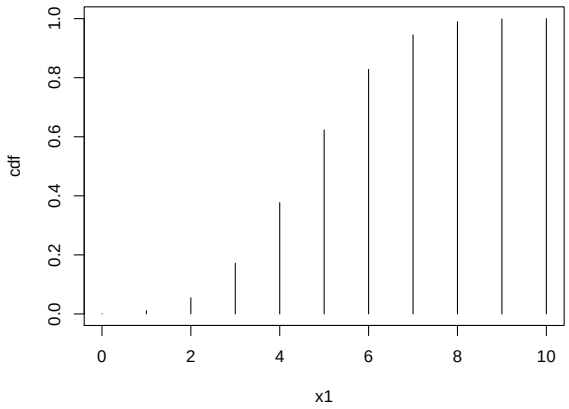

Figure 3.2: Binomial cdf

```
> x2 <- seq(-3, 3, by = 0.1)
> x2

 [1] -3.0 -2.9 -2.8 -2.7 -2.6 -2.5 -2.4 -2.3 -2.2 -2.1 -2.0 -1.9 -1.8 -1.7 -1.6
[16] -1.5 -1.4 -1.3 -1.2 -1.1 -1.0 -0.9 -0.8 -0.7 -0.6 -0.5 -0.4 -0.3 -0.2 -0.1
[31]  0.0  0.1  0.2  0.3  0.4  0.5  0.6  0.7  0.8  0.9  1.0  1.1  1.2  1.3  1.4
[46]  1.5  1.6  1.7  1.8  1.9  2.0  2.1  2.2  2.3  2.4  2.5  2.6  2.7  2.8  2.9
[61]  3.0

> cdf <- pnorm(x2, 0, 1)

> plot(cdf ~ x2, type = "l")
```

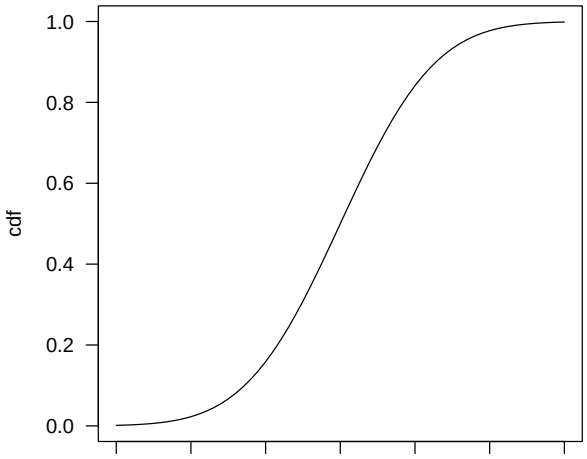

Figure 3.3: Normal cdf

3.4 Point estimation

Two overarching, and sometimes competing, principles guide how we perform estimation: method of moments (MM) and maximum likelihood (ML). Often the techniques that represent the two principles will yield identical estimates, but not always. ML is more complicated, and we only describe it briefly.

3.4.1 Method of moments

Consider the random variable X with a binomial distribution:

$$X \sim \text{Bi}(n, p). \tag{3.2}$$

The Method of Moments equates the "moments" of the population (location, spread, skewness, kurtosis, loosely speaking) with the moments of the sample. That is, if the sample mean is 12, then estimate the population mean as 12; likewise if the sample variance is 4, then estimate the population variance by 4. In the binomial case, the proportion p represents the mean number of successes. Therefore to estimate p for the population, the method of moments estimate is the sample value for p. Sometimes the moments may be a function of the parameter we are interested in estimating, rather than the parameter itself. In this instance, we equate this function to the sample moment and solve for the parameter to get our estimate.

Example **Salk vaccine II**

The estimate is $\hat{p} = \dfrac{57}{200,745} = 0.000284$, or 0.0284%.

```
> 57/200745

[1] 0.0002839423
```

3.4.2 Maximum likelihood

The Maximum Likelihood method finds the values of the parameters (here, just p) that make the observed data the most probable. That is, we have to find the values that maximize the joint probability mass function of the sample. For practical reasons, we usually prefer to maximize the log of the joint probability mass function, called the log-likelihood.

This maximization can be performed in one of two ways: either by writing down the log-likelihood and using differential calculus to maximise it, or using numerical methods. The former often provides us with simple and straightforward functions of the data. However, it's not always so simple to perform this operation and numerical methods in packages such as R are needed.

3.5 Interval estimation

Having found a point estimate, it is good to know the degree of confidence that we can place in our estimate as representative of the population from which it was selected. We do this by means of an *interval estimate*.

We will again start with the model: $X \sim Bi(n, p)$. Let $\hat{P} = \frac{X}{n}$ denote the random variable used to estimate p, $\mathbb{E}$ the expectation or mean of a random variable and `var` the variance. Then, from the properties of the binomial distribution,

$$\mathbb{E}(\hat{P}) = \frac{1}{n}\mathbb{E}(X) = \frac{1}{n} \times np = p$$

and

$$\text{var}(\hat{P}) = \frac{1}{n^2}\,\text{var}(X) = \frac{1}{n^2} \times np(1-p) = \frac{p(1-p)}{n};\ \text{SD}(\hat{P}) = \sqrt{\frac{p(1-p)}{n}}.$$

This provides us enough information to make interval estimates.

Some of you will be familiar with the large-sample theory "Wald" confidence intervals, even if you don't know them by that name; they are based on the normal distribution, and we provide these below for completeness. However, they do not produce the best interval estimates. The best interval estimation method for a binomial parameter is the subject of some debate, and an area of active statistical research.

In general, the main criterion used to compare interval estimate methods is *coverage*, which concerns the percentage of intervals which contain the true parameter value. For example, a 95% confidence interval should contain the true value 95% of the time in the long run. Three confidence interval methods for p are presented here.

3.5.1 Wald

If n is sufficiently large, the sampling distribution of p is approximately normal, so using the mean and standard deviation of a binomial distribution, the approximate distribution of $\hat{P}$ is

$$\hat{P} \overset{d}{\simeq} N\left(p,\ \sqrt{\frac{p(1-p)}{n}}\right).$$

We can use this to derive a *Wald* confidence interval. The 0.975 quantile of the standard normal distribution is 1.96, so

$$\Pr\left(-1.96 < \frac{\hat{P}-p}{\sqrt{p(1-p)/n}} < 1.96\right) \approx 0.95$$

$$\Rightarrow \Pr\left(\hat{P} - 1.96\sqrt{\frac{p(1-p)}{n}} < p < \hat{P} + 1.96\sqrt{\frac{p(1-p)}{n}}\right) \approx 0.95.$$

Don't worry if this is challenging – we only need the result! These inequalities lead to a quadratic in p which can readily be solved, but an accepted simpler method is to use an additional approximation, which replaces p under the $\surd$ by the observed proportion $\hat{p}$:

$$\frac{p(1-p)}{n} \approx \frac{\hat{p}(1-\hat{p})}{n}.$$

Using this, a 95% confidence interval for p is

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Note that $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ is the standard error of $\hat{P}$, $se(\hat{P})$, so that the confidence interval is of the familiar form: estimate $\pm\ k \times se$(estimate).

Example **Salk vaccine III**

In the *randomised controlled double-blind* study, 57 of the 200,745 vaccinated children developed polio. What is the probability (p) that a vaccinated child will develop polio?

The obvious estimate is $\hat{p} = \dfrac{57}{200,745} = 0.000284$, but how precise is this estimate?

Or, in statistical terms, what is a confidence interval for p?

Let X denote the number of children in a sample of size n that develop polio. Then, assuming that each child has the same probability p of developing polio, independent of any other child,

$\hat{p} = 0.000284$, hence $se(\hat{P}) = \sqrt{\dfrac{0.000284 \times 0.999716}{200,745}} = 0.0000376.$

Hence a 95% Wald confidence interval for p is:

$$\begin{aligned} 0.000284 \pm 1.96 \times 0.0000376 &= 0.000284 \pm 0.000074 \\ &= (0.000210,\ 0.000358) \end{aligned}$$

```
> n <- 200745
> (p.hat <- 57/n)
[1] 0.0002839423
> p.hat + c(-1.96, 1.96) * sqrt(p.hat * (1 - p.hat)/n)
[1] 0.0002102390 0.0003576456
```

Consider another study in which 2 out of 1000 vaccinated children developed polio. Calculate a 95% confidence interval for p. Any problems?

3.5.2 Agresti-Coull

An interval estimation method was proposed by Agresti and Coull (1998) in the paper "Approximate is better than 'exact' for interval estimation of binomial proportions", *The American Statistician* 52:119–126. The resulting confidence intervals are now referred to as Agresti-Coull intervals. In brief, these are easily-computed interval estimates that have substantially better coverage than the Wald intervals. For 95% confidence intervals, compute

$$\tilde{p} \pm 1.96\sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}}$$

where $\tilde{p} = \frac{x+2}{n+4}$ and $\tilde{n} = n + 4$.

In short, add two successes and two failures, and compute the usual Wald 95% interval. As a further approximation, 2 can be used as the multiplier instead of 1.96. Note that for intervals of different coverage, different constants must be used.

These intervals are ad-hoc, and are not claimed to be optimum. However, they are simple, and have been shown to be reliable (i.e. provide coverage close to the claimed level) under most circumstances. Note that the point estimate of p should still be $\hat{p} = \frac{x}{n}$; adding two successes and two failures only applies to the interval estimate.

Example **Salk vaccine IV**

```
> (n.tilde <- 200745 + 4)

[1] 200749

> (p.tilde <- (57 + 2)/n.tilde)

[1] 0.0002938993

> p.tilde + c(-1.96, 1.96) * sqrt(p.tilde * (1 - p.tilde)/n.tilde)

[1] 0.0002189160 0.0003688827
```

3.5.3 Jeffreys prior

One disadvantage of the Agresti-Coull intervals is that when the observed proportion is close to 0 or 1, it is possible for them to exit the $[0, 1]$ interval. Under such circumstances, it is acceptable to use intervals that are usually referred to as Jeffreys prior intervals. These are Bayesian confidence intervals[1] that have good frequentist properties, meaning that they behave well in terms of coverage. A formal definition is as follows: for a binomial random variable X, with x successes from n independent trials, the $1 - \alpha$ confidence interval is computed using the inverse of the Beta cumulative distribution function, evaluated with parameters $a = x + 0.5$ and $b = n - x + 0.5$. Even if that does not make a lot of sense, the intervals are easy to compute in R:

Example **Salk vaccine V**

The Jeffreys prior 95% confidence interval is:

```
> qbeta(c(0.025, 0.975), 57 + 0.5, 200745 - 57 + 0.5)

[1] 0.0002172298 0.0003650525
```

Jeffreys prior intervals are not generally symmetric, and will always be within 0 and 1. These characteristics are an advantage. This interval estimation method should be used when $\hat{p}$ is close to 0 or 1 (say, $\hat{p} < 0.05$, or $\hat{p} > 0.95$), especially when n is small.

3.6 Assumptions

There are several assumptions being made in these estimation methods, and they do need to be checked or the results can be misleading. The main assumptions are as follows:

- n is large enough. A rough guide is that if there are at least 5 cases with the attribute of interest (e.g. disease) and at least 5 without it, the normal approximation (on which the Wald and Agresti-Coull methods rely) is okay (some guides suggest 10 and 10). So, more items are needed to use the normal approximation when studying rare (or extremely common) events.
- The items are independent. This can fail if, for instance, we randomly selected 10 companies and from each company selected 13 or 14 employees. Or, if the observations are taken in sequence, and there is a tendency for repeated outcomes. We can only check this assumption by observing the data collection process, or reviewing a summary of it.
- The sample is representative of the population. Good experimental or sample design is the key. We will cover the principles of design in a future chapter.

[1] Don't worry – you don't need to know what this means for this subject.

3.7 More than two categories

Point estimation of proportions or probabilities when there are more than two categories remains straightforward. Consider the `ufc` data containing the tree measurements from a forest inventory, which we met in week 1. Previously we used the `table` function in R to find the *number* of trees in each of the four species. We now find the *proportion* of trees in each of the four species, by dividing by the number of observations (or number of rows).

```
> ufc <- read.csv("../data/ufc.csv")
> table(ufc$species)

 DF  GF  WC  WL
 57 118 139  22

> table(ufc$species)/dim(ufc)[1]

        DF         GF         WC         WL
0.16964286 0.35119048 0.41369048 0.06547619
```

The `dim()` function in R gives the dimensions of the data frame; the first element, `dim()[1]`, is the number of rows, which is the appropriate denominator in calculating the proportion in each category. Another way of obtaining this denominator is `length(ufc$species)`, which gives the length of the species vector. Note that `table` can accept more than one categorizing argument.

Now, we find the proportion of trees in each combination of plot and species. There are 132 plots, so we print out the first few plots only.

```
> head(table(ufc$plot, ufc$species))/dim(ufc)[1]

             DF          GF          WC          WL
  2 0.002976190 0.000000000 0.000000000 0.002976190
  3 0.000000000 0.002976190 0.005952381 0.000000000
  4 0.002976190 0.000000000 0.002976190 0.000000000
  5 0.002976190 0.002976190 0.000000000 0.000000000
  6 0.002976190 0.000000000 0.011904762 0.000000000
  7 0.000000000 0.000000000 0.005952381 0.000000000
```

3.8 Difference between two proportions

A substantial advantage to assuming that estimates of random variables are normally distributed is that we can make the same assumption about linear combinations of those random variables.

If $X_1 \sim \mathrm{N}(\mu_1, \sigma_1)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2)$, then

$$aX_1 + bX_2 \sim \mathrm{N}\left(a\mu_1 + b\mu_2, \sqrt{a^2\sigma_1^2 + b^2\sigma_2^2}\right) \tag{3.3}$$

That is, normality is preserved under linear transformations. The mean of the transformed data is just the transformation of the means, and variances are additive but scaled by the square of the constants.

We often wish to know about the distribution of *differences* between pairs of random variables. This is given by

$$X_1 - X_2 \sim \mathrm{N}\left(\mu_1 - \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2}\right) \tag{3.4}$$

For proportions (which are really just means of a variable taking the values 0 or 1), we obtain the following approximate result:

$$\hat{P}_1 - \hat{P}_2 \sim \mathrm{N}\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right),$$

which leads to the following approximate 95% confidence interval for $p_1 - p_2$:

$$\hat{p}_1 - \hat{p}_2 \pm 1.96\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

3.9 Exercises

1. On January 30, 1995, Time magazine reported the results of a poll of adult Americans in which it asked "Have you ever driven a car when you probably had too much alcohol to drive safely?" The exact results were not given, but from the information given we can guess at what they were. Of the 300 men who answered, 189 (63%) said yes and 108 (36%) said no while the remaining 3 men weren't sure. Compute an approximate 95% confidence interval for the proportion of men who would say yes to this question, using the Agresti-Coull method.

2. According to Mendel's theory of inheritance, the progeny produced when a certain type of sweet pea is crossed should be:

$$\begin{array}{ccc} \text{round } (R) & \text{or} & \text{wrinkled } (W) \\ & \text{and} & \\ \text{yellow (Y)} & \text{or} & \text{green } (G) \end{array}$$

in the following proportions:

$$RY : WY : RG : WG = 9 : 3 : 3 : 1$$

One set of data reported by Mendel was as follows:

Type	Frequency
RY	315
WY	101
RG	108
WG	32
Total	556

Enter the data into R, and make point estimates of the population proportions. Think carefully about the assumptions. In what important ways could they be wrong? What do these data say about Mendel's theory?

3. According to the Red Cross, the four major blood groups occur in the population with the following proportions:

Blood group	O	A	B	AB
Proportion	0.45	0.42	0.10	0.03

However it is known that these proportions vary throughout the world. The following numbers were found in a sample of 100 students from the University of Melbourne:

Blood group	O	A	B	AB
Number	51	38	9	2

Use R to compute point estimates of the population proportion for blood group AB and find a 95% confidence interval using the Wald and Jeffreys prior methods. Which method is more appropriate here? Think carefully about the assumptions. In what important ways could they be wrong? What do these data say about the claim from the Red Cross?

4. K & P Electrics sell, among other things, kitchen appliances. Recent sales figures for white enamel and stainless steel refrigerators and dishwashers are given below. An article in *Consumer Choice* magazine claims that 70% of customers prefer a stainless steel finish for a kitchen appliance over a white enamel finish.

	Finish	
	white enamel	stainless steel
Refrigerators	70	130
Dish washers	48	52
Total	118	182

   (a) Enter the data into R and produce the table of frequencies.

   (b) Use R to compute point and interval estimates of the population proportion of customers preferring a stainless steel finish. Think carefully about the assumptions. In what important ways could they be wrong? What do these data say about the claim made in the article?

   (c) Use R to estimate the difference between the proportion of customers that prefer stainless steel for refrigerators as opposed to the proportion of customers that prefer stainless steel for dish washers. Compute both a point and an interval estimate of the difference, and comment on them. Use the approximate result derived in Section 3.8.

5. A sample of college students was asked if they would return the money if they found a wallet on the street. Of the 93 females, 84 said "yes", and of the 75 males, 53 said "yes". Assume these students represent all college students.

   (a) Find separate approximate 95% confidence intervals for the proportions of college females and college males who would say "yes" to this question.

   (b) Find an approximate 95% confidence interval for the difference in the proportions of college males and females who would say "yes" to this question.

   (c) Write a few sentences interpreting the interval in part (b).

6. A clinical trial examined the effectiveness of aspirin in the treatment of cerebral ischemia (stroke). Patients were randomised into treatment and control groups. The study was double-blind in the sense that neither the patients nor the physicians who evaluated the patients knew which patients received aspirin and which the placebo tablet. After 6 months of treatment, the attending physicians evaluated each patient's progress as either favorable or unfavorable. Of the 78 patients in the aspirin group, 63 had favorable outcomes; 43 of the 77 control patients had favorable outcomes. [2]

   (a) Comment on the design of this study.

   (b) Compute the sample proportions of patients having favourable outcomes in the two groups.

   (c) Give a 95% confidence interval for the difference between the favourable proportions in the treatment and control groups.

[2] From William S. Fields et al., "Controlled trial of aspirin in cerebral ischemia," *Stroke*, 8 (1977), pp 301–315.