Chapter 4

Hypothesis testing for categorical data

4.1 Objectives

1. To know how to perform hypothesis tests for population proportions and differences between population proportions.

2. To be able to use R to resolve hypothesis tests for population proportions and differences between population proportions.

4.2 Introduction

David Salsburg's book *The lady tasting tea: How statistics revolutionized science in the 20th century* describes an incident at a summer tea party in Cambridge, England, when a lady stated that a cup of tea prepared by pouring the tea into milk tastes different from a cup of tea prepared by pouring the milk into tea. Her notion was disputed by the scientific minds of the group. But one guest, by the name Ronald Fisher, proposed to scientifically test the lady's hypothesis. She was given eight cups of tea, some of which were made milk first, and others tea first, and was asked to give her opinion on which of these two ways had been used.

Would you be convinced that her theory was valid if she got all 8 correct? 7 correct? 6 correct?

4.3 Hypothesis Testing

We can calculate a P-value to help us judge the "conviction"[1] available to us from data. A P-value is defined as the probability of observing a value of the test statistic as or more extreme than the one actually observed, assuming that the null hypothesis is true.

Here is how we would calculate the P-value for the outcome: "6 out of 8 correct." The test statistic is X, the number of correct guesses out of $n = 8$ trials. The null hypothesis, usually denoted H_0, would be that there is no difference between the two ways of making tea. If we assumed H_0, the lady's choices would be merely guesses and she would expect to get about half of them right. She wouldn't always get exactly half of them right, but X would be a binomial random variable with $n = 8$ and $p = 0.5$ (assuming that she made 8 independent choices, and was not told how many of each type would be presented to her).

The alternative hypothesis, usually denoted H_1, is that she could tell with a probability greater than 0.5.

If she got 6 out of 8 right, then $\hat{p} = \frac{6}{8} = 0.75$, which is somewhat higher than 0.5. But is it *convincingly* higher than 0.5?

The P-value is $\Pr(X \geq 6)$ ("as or more extreme"). If we look up binomial tables (partly shown in the table below), we find $\Pr(X \geq 6) = 0.1094 + 0.0313 + 0.0039 = 0.1446$.

[1] We have avoided using the word *evidence*, here, as numerous authors have argued convincingly that P-values are flawed measures of evidence. See, for example, Richard Royall's "Statistical Evidence".

		p	
	x	0.50	
$n = 8$	0	.0039	8
	1	.0313	7
	2	.1094	6
	3	.2188	5
	4	.2734	4
	5	.2188	3
	6	.1094	2
	7	.0313	1
	8	.0039	0
		0.50	x

Probability mass function for $X \sim \text{Bi}(8, p)$.

This P-value is not very impressive. Conventionally, the P-value would need to be < 0.05 before we would be convinced. We would not reject the null hypothesis in this case. If the lady could only identify 6 out of 8 correctly, there is insufficient evidence to indicate that the two teamaking techniques can be distinguished.[2]

R can help us with this test, as it provides the cumulative distribution function for the binomial distribution: `pbinom`. Recall that the cdf is the probability that a random variable of a given distribution takes a value less than or equal to the value at which the cdf is computed. So, the P-value is

```
> 1 - pbinom(5, size = 8, prob = 0.5)

[1] 0.1445312
```

Had the test been two-tailed, that is, had the alternative hypothesis been that the probability was not 0.5 (being either higher or lower), then we would double this P-value (unless it is greater than 0.5, in which case we would take $P = 1$).

For much larger values of n, we can use the normal approximation to the binomial distribution.

Example In 1965, the U.S. Supreme Court decided the case of Swain *versus* Alabama.[3] Swain, an African American man, was convicted in Talladega County, Alabama, of raping a white woman and was sentenced to death. The case was appealed to the Supreme Court on the grounds that there were no African Americans on the jury.

The Supreme Court denied the appeal, on the following grounds. As provided by Alabama law, the jury was selected from a panel of about 100 persons. There were 8 African Americans on the panel. (They did not serve on the jury because they were "struck", through challenges by the prosecution. Such challenges are constitutionally protected.) The presence of only 8 African Americans on the panel of prospective jurors showed, so ruled the Court, "the overall percentage disparity has been small and reflects no studied attempt to include or exclude a specified number of Negroes."

At that time in Alabama, only men over 21 were eligible for jury duty. There were about 16,000 such men in Taladega county, of whom 26% were African Americans. If the 100 people in the panel had indeed been chosen at random from this population, what would the chance be of 8 or fewer being African American?

```
> pbinom(8, 100, 0.26)

[1] 4.734795e-06
```

The probability is 0.0000047 or about one chance in 200,000. Note that in an R function we only need to include the value of an argument (i.e. '100'), not its name ('size=100'), though we may sometimes choose to do so for record-keeping purposes.

For hypothesis testing, we can also use the normal approximation arguments as with confidence intervals. The only difference is that, since we are computing probabilities assuming that the null hypothesis $\text{H}_0 : p = p_0$ is true, we **don't** need to make the approximation we made with confidence intervals that $\frac{p(1-p)}{n} \approx \frac{\hat{p}(1-\hat{p}}{n}$), or the Agresti-Coull correction. Instead we just use p_0 in computing the standard error and P-value.

[2] David Salsburg's book reveals on good authority (from a colleague of Fisher's) that she actually got 8 out of 8 correct.
[3] From *Freedman, Pisani and Purves, p338.*

The estimate is $\hat{p}$; its standard error (estimated standard deviation) is $\sqrt{\frac{p_0(1-p_0)}{n}}$; the test statistic

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

is approximated by a standard normal distribution under the null hypothesis.

Note the different standard error formulae used for confidence intervals and hypothesis testing. The standard error for confidence intervals uses $\hat{p}$ whereas the standard error for hypothesis testing uses p_0.

Example **Swain versus Alabama** (continued)
The question being asked here can be expressed as $H_0 : p = 0.26$ versus $H_1 : p < 0.26$. Having (observed) 8 African Americans in a panel of (about) 100, the required P-value is given by $Pr\left(Z \leq \frac{0.08-0.26}{\sqrt{0.26\times 0.74/100}}\right) = Pr(Z \leq -4.104) = 0.000020$ (about one chance in 50,000).

Hence we reject H_0 and conclude that the Court was (most likely) wrong in its ruling.

4.3.1 The sign test

The sign test is a hypothesis test which uses the binomial distribution. The test is most commonly used in the context of paired samples when the assumption of normality of the differences is not reasonable. In this context it is a test for the median of the differences, usually with the null hypothesis that the median of the differences is zero. The test can be based on either the number or the proportion of the differences that are positive (or negative); the number is more commonly used in practice.

Example **Airborne pollution**
Airborne particles such as dust and smoke are an important part of air pollution. To measure particulate pollution, a vacuum motor is used to draw air through a filter for 24 hours. The filter is weighed at the beginning and end of the period and the weight gained is a measure of the concentration of particles in the air. A study of pollution made measurements every six days with identical instruments in the centre of a small city and at a rural location 10 miles southwest of the city. Because the prevailing winds blow from the west, it was suspected that the rural readings may be generally lower than the city readings. The data below are the readings (particulate levels, in grams) taken (roughly) every six days over a four month period.

Day	Rural	City	Diff	Day	Rural	City	Diff
1	67	68	1	11	43	42	-1
2	42	42	0	12	39	38	-1
3	33	34	1	13	52	57	5
4	46	48	2	14	48	50	2
5	43	45	2	15	56	58	2
6	38	39	1	16	44	45	1
7	108	123	15	17	51	69	18
8	57	59	2	18	21	23	2
9	70	71	1	19	74	72	-2
10	42	41	-1	20	48	49	1

The hypothesis we are interested in here is:

H_0 : There is no difference in particulate levels between the rural and city locations;

H_1 : There is a difference (H_0 is simply not true – a two-sided test), or city readings are generally higher (a one-sided test).

If H_0 is true, then the median of the differences should be zero. Days with a difference of zero do not provide any useful information for the sign test, and we ignore them. To test the null hypothesis we use the number of negative differences (4) among the non-zero differences (19). Assuming that the differences for different days are independent, the number of negative differences, among the non-zero differences, should follow a binomial $(19, p)$ distribution; when H_0 is true $p = 0.5$.

Let $X \sim \text{Bi}(19, 0.5)$. Using the binomial distribution, we find that $\Pr(X \geq 15) = 1 - \Pr(X < 15) = 1 - \Pr(X \leq 14) = \Pr(X \leq 4) = 0.0096$. Hence the $P-$value for the one-sided alternative is 0.0096, and for the two-sided alternative it is $2 \times 0.0096 = 0.0192$.

This test can be carried out in R as follows:

```
> pbinom(4, 19, 0.5) * 2

[1] 0.01921082
```

We would reject H_0, and conclude that city readings are generally higher.

One or two-sided alternative hypotheses

There is a strong and conservative convention in science to prefer two-sided tests. This is because with a one-sided alternative hypothesis you are asserting that you are *sure* that there cannot be any effect in the opposite direction. Even if you think or hope that the effect will be in one direction, it is another thing altogether to have such conviction about this that you refuse to contemplate the possibility of an effect in the opposite direction. The convention to prefer two-sided tests is a good default. One-sided tests have their place (we started this lab with an example of one), but the onus is on the researcher to show that a one-sided alternative is appropriate.

4.4 Contingency Tables

4.4.1 Introduction

If the response variable is categorical with more than two categories, or we want to compare more than two proportions, or we simply want to investigate the association between two categorical variables, then the **chi-squared** test (based on the χ^2 distribution) is often appropriate. The chi-squared test enables us to compare a set of observed frequencies with the frequencies that would be expected if some null hypothesis were true. We will use the following two examples to illustrate the procedure.

Example **Tossing a die**

A (6-sided) die was tossed 120 times with the following outcomes:

	1	2	3	4	5	6
frequency	25	18	28	20	16	13

Is the die biased?

Example **Alcohol and nicotine consumption**

Alcohol and nicotine consumption during pregnancy may harm children. Because drinking and smoking behaviours may be related, it is important to understand the nature of this relationship when assessing the possible effects on children. One study classified 452 mothers according to their alcohol intake prior to pregnancy recognition and their nicotine intake during pregnancy.[4]

	Nicotine (milligrams/day			
Alcohol (ounces/day)	None	1 – 15	16 or more	total
None	105	7	11	123
0.01 – 0.10	58	5	13	76
0.11 – 0.99	84	37	42	163
1.00 or more	57	16	17	90
total	304	65	83	452

Is there an association between alcohol consumption and nicotine consumption?

Note that the χ^2 test assesses overall association, but does not indicate the nature or size of any effects; it should be regarded as a preliminary test to be followed by more careful sorting out of effects that are found to be significant.

[4] Data from Ann P. Streissguth et al., 'Intrauterine alcohol and nicotine exposure: attention and reaction time in 4-year-old children,' *Developmental Psychology,* 20 (1984), pp. 533-541.

4.4.2 The chi-squared (χ^2) test

The test statistic takes the form

$$X^2 = \sum \frac{(observed - expected)^2}{expected}$$

Where:

- *observed* refers to the observed frequency of a category, such as the number of mothers who consume at least 1 ounce of alcohol and at least 16 milligrams of nicotine per day.
- *expected* refers to the frequency that would be expected if the hypothesis being tested (H_0) is true. For the die tossing example, all of the expected frequencies are 20 – if the die is unbiased.

Like the t and F tests, the χ^2 test depends on *degrees of freedom* (*df*), which here are given by

$$\text{number of categories (cells)} - \text{number of constraints.}$$

For the die tossing example the number of categories is 6 and the number of constraints is 1 (sum of the expected frequencies = sum of observed frequencies = 120), so the test has 5 df.

Example **Tossing a die** (continued)

	1	2	3	4	5	6
observed frequency	25	18	28	20	16	13
expected frequency	20	20	20	20	20	20

H_0 : the die is unbiased (all outcomes are equally likely)

H_1 : H_0 is not true (some outcomes are more likely than others)

$$\begin{aligned} X^2 &= \frac{(25-20)^2}{20} + \frac{(18-20)^2}{20} + \frac{(28-20)^2}{20} + \frac{(20-20)^2}{20} + \frac{(16-20)^2}{20} + \frac{(13-20)^2}{20} \\ &= 1.25 + 0.20 + 3.20 + 0 + 0.80 + 2.45 \\ &= 7.90 \end{aligned}$$

and hence the P-value is $\Pr(X^2 \geq 7.90)$ where $X^2 \sim \chi^2_5$. From statistical tables, P is between 0.1 and 0.25 and hence we do not reject H_0 (at the 5% level).

From R, we provide the observed counts for each category to the `chisq.test` function. The default test is for the null hypothesis of equal proportions.

```
> chisq.test(c(25, 18, 28, 20, 16, 13))

        Chi-squared test for given probabilities

data:  c(25, 18, 28, 20, 16, 13)
X-squared = 7.9, df = 5, p-value = 0.1618
```

Hence the P-value = 0.1618. So, we would not conclude that the die is biased. Note that if the null hypothesis is something other than equal proportions then you should use the `p` argument to provide the probabilities that represent the null hypothesis. See `?chisq.test` for details.

4.4.3 Analysis of contingency tables

A table like that in the alcohol-nicotine example is known as a two-way) **contingency table**. There are two factors, which determine the 'rows' and 'columns' of the table, with r and c levels, respectively.

The hypothesis tested is

H_0 : no association (or independence) between the two factors, versus
H_1 : H_0 is not true.

Note that, whereas for the die tossing example H_0 determined the (exact) probability of each outcome, here H_0 merely specifies relationships between probabilities. In general, events A and B are independent if and only if $\Pr(AB) = \Pr(A) \times \Pr(B)$. Hence, if H_0 is true, then, for example,

Pr(a mother consumes ≥ 1 ounce of alcohol **and** ≥ 16 milligrams of nicotine per day)
$= \Pr(\geq 1$ ounce of alcohol$) \times \Pr(\geq 16$ milligrams of nicotine$)$

It follows that the expected values for the contingency table are given by

$$\text{expected} = \frac{\text{row total} \times \text{column total}}{\text{table total}}$$

For example, of the 90 mothers who consumed ≥ 1 ounce of alcohol per day, we would expect a fraction of $\frac{83}{452}$ of them to consume ≥ 16 mg of nicotine per day – if there is no association between alcohol and nicotine consumption. Hence the expected frequency is $\frac{83}{452} \times 90$ or $\frac{83 \times 90}{452} = 16.53$.

The degrees of freedom for the test are given by

$$(\#\ rows - 1) \times (\#\ columns - 1) = (r-1)(c-1) = rc - (r + c - 1)$$

There are rc 'cells' or categories, while the constraints on the expected frequencies are:

row totals = those of the observed frequencies ($\Rightarrow r$ constraints)
column totals = those of the observed frequencies ($\Rightarrow c$ constraints)

But the sum of the row totals = sum of the column totals = table total, so that we have counted one constraint twice. Hence the number of (independent) constraints is $r + c - 1$. However, it is easier to remember $(r-1)(c-1)$ as the df for the test.

Example **Alcohol and nicotine consumption** (continued)

Under H_0, the expected frequencies are

	Nicotine (milligrams/day			
Alcohol (ounces/day)	None	1 – 15	16 or more	total
None	82.73	17.69	22.59	123
0.01 – 0.10	51.12	10.93	13.96	76
0.11 – 0.99	109.63	23.44	29.93	163
1.00 or more	60.53	12.94	16.53	90
total	304	65	83	452

Hence

$$X^2 = \frac{(105 - 82.73)^2}{82.73} + \frac{(58 - 51.12)^2}{51.12} + \ldots + \frac{(17 - 16.53)^2}{16.53} = 42.252$$

From the χ^2_6 distribution, we find that $P < 0.001$.

This test can readily be carried out in R as follows.

```
> alcohol.by.nicotine <- matrix(c(105, 7, 11,
+                                 58,  5, 13,
+                                 84, 37, 42,
+                                 57, 16, 17), nrow = 4, byrow = TRUE)
> chisq.test(alcohol.by.nicotine)
```

The null hypothesis is decisively rejected and we would conclude that there is an association between alcohol and nicotine consumption.

4.4.4 Requirements

Using the χ^2 distribution (to obtain P-values) for the X^2 test statistic, involves an approximation analogous to the normal approximation to the binomial distribution. And just as we had guidelines for valid use of the normal approximation, there are guidelines for the valid use of the χ^2 test, namely:

1. Cell contents must be actual counts (not proportions).
2. Categories must not overlap.
3. All expected frequencies should be ≥ 1.
4. At least 80% of expected frequencies should be ≥ 5.

If these requirements are not satisfied then (some) cells could be combined to ensure that they are satisfied. The way any cells are combined should be *sensible* within the context of the study. If cells are combined, then the degrees of freedom are reduced and the power of the test is reduced.

Example **Aluminium and Alzheimer's disease**[5]

A study compared a group of Alzheimer's patients with a carefully selected control group of people who did not have Alzheimer's but were similar in other ways. (This is an example of a case–control study.) The focus of the study was on the use of antacids that contain aluminium.

	Aluminum-containing antacid use			
	None	Low	Medium	High
Alzheimer's patients	112	3	5	8
Control group	114	9	3	2

```
> aluminium.by.alzheimers <- matrix(c(112, 3, 5, 8,
+                                     114, 9, 3, 2), nrow=2, byrow = TRUE)
> (a.by.a.test <- chisq.test(aluminium.by.alzheimers))

        Pearsons Chi-squared test

data:  aluminium.by.alzheimers
X-squared = 7.1177, df = 3, p-value = 0.06824
```

Note that by assigning the function output to an object (here, `a.by.a.test`), *and* wrapping the assignment in parentheses, we save the object and print out the result. In line with our earlier concern, we should check the expected cell counts.

```
> a.by.a.test$expected

     [,1] [,2] [,3] [,4]
[1,]  113    6    4    5
[2,]  113    6    4    5
```

2 of the 8 cells have expected frequencies < 5, hence the P-value (0.069, which is reasonably small, but not quite significant at the 5% level), may not be too reliable.

We can address this problem in a few ways. One way is by using simulation to obtain a more accurate P-value, for example:

```
> chisq.test(aluminium.by.alzheimers, simulate.p.value = TRUE)

        Pearsons Chi-squared test with simulated p-value (based on 2000
        replicates)

data:  aluminium.by.alzheimers
X-squared = 7.1177, df = NA, p-value = 0.06047
```

[5] From *Moore and McCabe*, based on a paper by Amy Borenstein Graves et al., 'The association between aluminum-containing products and Alzheimer's disease' *Journal of Clinical Epidemiology,* 43 (1990), pp 35–44.

This simulation process produces 2000 random tables of independent data to assess how extreme the observed table is under the null hypothesis, without assuming the Chi-squared distribution.

Another way of addressing the problem is by reclassifying the data. The more problematic cells are those for Medium antacid use, and it would be reasonable to combine the Medium and High use categories to form the following 'reduced' table:

	Aluminum-containing antacid use		
	None	Low	Med–High
Alzheimer's patients	112	3	13
Control group	114	9	5

```
> aluminium.by.alzheimers <- matrix(c(112, 3, 13,
+                                     114, 9, 5), nrow=2, byrow = TRUE)
> (a.by.a.test <- chisq.test(aluminium.by.alzheimers))

        Pearsons Chi-squared test

data:  aluminium.by.alzheimers
X-squared = 6.5733, df = 2, p-value = 0.03738

> a.by.a.test$expected

     [,1] [,2] [,3]
[1,]  113    6    9
[2,]  113    6    9
```

Now all of the expected frequencies are ≥ 5 and the test gives $P = 0.038$, so we might conclude that there is evidence of (some) association between antacid use and Alzheimer's disease.

However, it would also be reasonable to combine the Low and Medium categories into a single Low-Med category. Try it, and see whether it would affect the conclusion.

4.5 Fisher's exact test for a 2×2 table

There is another approach to testing for association in a 2×2 contingency table, which does not need any large sample approximation. It was first proposed by R.A. Fisher in the 1920s, but for a long time the test could only be used in very simple cases because of its computational intensiveness. This is no longer a prohibitive factor, and so Fisher's exact test is now widely used.

Example **Gender and left-handedness**

Suppose that 50 arts students are randomly selected for a survey in which they are asked questions about motor characteristics such as left- and right-handedness. Forty-seven students respond, and the researcher wants to examine the effect of gender on each characteristic. Here is a cross-table of gender by handedness:

	Left	Right	Total
Female	1	30	31
Male	4	12	16
Total	5	42	47

The proportion of females who are left-handed ($1/31 = 0.032$) is less than the proportion of males ($4/16 = 0.25$). To examine whether these proportions are significantly different, we could possibly use a χ^2 test, but two of the four cells have expected values less than 5.

Fisher's exact test works as follows: suppose we knew the *marginal totals*, but not the entries within the table. The incomplete table would look like this:

	Left	Right	Total
Female			31
Male			16
Total	5	42	47

With these marginal totals, there are in fact only 6 possible outcomes, because the top left-hand entry can only take the values 0, 1, 2, 3, 4, or 5; and once one of the table's entries is specified, the others are determined by subtraction. The possible outcomes are as follows, with their probabilities of occurring listed underneath:

	L	R	L	R	L	R	L	R	L	R	L	R
Female	0	31	1	30	2	29	3	28	4	27	5	26
Male	5	11	4	12	3	13	2	14	1	15	0	16
			observed									
Probability	0.0028		0.0368		0.1698		0.3516		0.3282		0.1108	

The probabilities are calculated using the *hypergeometric probability distribution*, which is similar to the binomial distribution, but assumes sampling *without replacement* from a population (the binomial distribution assumes sampling *with replacement*).

Because of the asymmetry of the probabilities, a two-sided P-value (which is generally more appropriate than a one-sided P-value) is not uniquely defined. One method is to calculate both one-sided P-values, and then double the smaller of the two. Another method, which R uses, is to sum the probability of all outcomes equal in probability or less probable than the observed table. With this method, the less probable outcomes are taken as "more extreme" than the observed outcome in their departure from the null hypothesis. In this example, the P-value is

$$P = 0.0368 + 0.0028 = 0.040$$

Using the 0.05 significance level, we would reject the hypothesis of no association, and conclude that the proportion of females who are left-handed is lower than the proportion of males. In R the test is performed as follows:

```
> handedness <- matrix(c(1, 30, 4, 12), nrow = 2, byrow = TRUE)
> fisher.test(handedness)

        Fishers Exact Test for Count Data

data:  handedness
p-value = 0.03963
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.001973399 1.206146041
sample estimates:
odds ratio
 0.1055741
```

Fisher's test is called "exact" because it is based on computation of probabilities rather than an approximation. In R, the Fisher's exact test function also estimates the odds ratio, which we will consider later in the course. More details on the χ^2 test and Fisher's exact test can be found on many websites and statistical texts. A good text on analysis involving frequencies and proportions is Fleiss et al. (2003), *Statistical Methods for Rates and Proportions.*

4.6 Exercises

Many of the exercises in this lab use the same data as the previous lab exercises.

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

round (R)	or	wrinkled (W)
	and	
yellow (Y)	or	green (G)

   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

   Use R to formally test: what do these data say about Mendel's theory? You will need to create a variable with four expected proportions under the null hypothesis. This variable becomes the second argument in the `chisq.test()` function.

2. 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

   Perform a test of whether the preference of finish is associated with the type of appliance. Perform a test to examine the claim made by the article.

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

   The following numbers were found in a sample of 100 Melbourne University students:

Blood group	O	A	B	AB
Number	51	38	9	2

   Use R to formally test whether the sample is consistent with the Red Cross proportions. Comment whether this analysis satisfies the usual requirements of the χ^2 test. If it doesn't, how would you address the issue?

4. The two key experiments for testing the use of the Salk vaccine against polio were a "field trial" and a randomized control trial (RCT).[6] The data are presented below.

Study	Group	n	polio cases
Field trial	Vaccinated	221,998	56
	Controls	725,173	391
RCT	Vaccinated	200,745	57
	Placebo	201,229	142

   For the field trial, perform a test on whether there is a difference between the vaccinated and control proportions. Think about the assumptions. In what ways could they be wrong?

5. 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.[7]

   Test whether the favourable outcomes were in the same proportion in the treatment and control groups, using a χ^2 test. Perform the test using Fisher's exact test.

6. The following data come from a study comparing the health of juvenile delinquent boys and a non-delinquent control group. They relate to the subset of the boys who failed a vision test, and show the numbers who did and did not wear glasses.

		Juvenile delinquent	Non-delinquent
Spectacle wearer	Yes	1	5
	No	8	2

   Using a suitable procedure, test whether there is an association between delinquency and spectacle wearing.

[6] Source: *An Evaluation of the 1954 Poliomyelitis Vaccine Trials*, American Journal of Public Health, 1955.
[7] From William S. Fields et al., "Controlled trial of aspirin in cerebral ischemia," *Stroke*, 8 (1977), pp 301–315.