Inference on Proportions

Statistics 149

Spring 2006

Example: Aspirin to Prevent Strokes

This was a study of the use of aspirin to prevent future strokes in patients with an earlier stroke.

155 people were studied, 78 receiving aspirin, 77 placebo. The results were as follows

Treatment	No Stroke	Stroke	Total
Aspirin	63	15	78
Placebo	43	34	77
Total	106	49	155

From a quick look at this table, it appears that Aspirin helps prevent strokes. Lets look at the theory to justify this statement

Bernoulli and Binomial Sampling

Bernoulli: This distribution is useful for describing the results of a single trial that is either a success (Prob $= \pi$) or a failure (Prob $= 1 - \pi = \varphi$).

$$Pr[Y = y] = \begin{cases} \pi & y = 1 \\ 1 - \pi & u = 0 \\ 0 & \text{Otherwise} \end{cases}$$

This is the model for flipping a biased coin. Its also the distribution of an indicator RV. (Denoted by $Ber(\pi)$)

$$E[Y] = \pi; \quad \text{Var}(Y) = \pi(1 - \pi); \quad \text{SD}(Y) = \sqrt{\pi(1 - \pi)}$$

Binomial: Let $Y_1, Y_2, \ldots, Y_n$ be independent $Ber(\pi)$ RVs. Then

$$S = \sum_{i=1}^{n} Y_i$$

is a binomial RV (Denoted $Bin(n, \pi)$). Note that $Ber(\pi) = Bin(1, \pi)$

S is the number of successes in n independent, identical (same π) trials.

So in the example, we have two binomial samples (at least that's the usual assumption).

Let S_i be the number of people that don't have a stroke.

1. Aspirin: $S_1 \sim Bin(78, \pi_1)$

2. Placebo: $S_2 \sim Bin(77, \pi_2)$

The question about whether aspirin works or not involves investigating the relationship between π_1 and π_2.

Let $F_i = n_i - X_i$ be the number of strokes under treatment i. Then its easy to see that

1. Aspirin: $F_1 \sim Bin(78, \varphi_1) = Bin(78, 1 - \pi_1)$

2. Placebo: $F_2 \sim Bin(77, \varphi_2) = Bin(77, 1 - \pi_2)$

Note in this second case, what we are calling a success (what we are counting) is actually a failure (i.e. stroke) for a practical purposes.

The PMF for the binomial distribution is

$$Pr[S = k] = \binom{n}{k} \pi^k (1 - \pi)^{n-k}; \quad k = 0, 1, \ldots, n$$

The moments of the distribution are

$$E[S] = np; \quad \text{Var}(S) = np(1 - p); \quad \text{SD}(S) = \sqrt{np(1 - p)}$$

For inference purposes

$$\hat{\pi} = \frac{S}{n}$$

the sample proportion is usually more useful to work with. It has the following properties

- $E[\hat{\pi}] = \pi$
- $\mathrm{Var}(\hat{\pi}) = \frac{\pi(1-\pi)}{n}$
- If n is large enough,

$$\hat{\pi} \overset{approx.}{\sim} N\left(\pi, \frac{\pi(1-\pi)}{n}\right)$$

The common rule of thumb for n being large enough is

$$n\pi > 5 \quad \text{and} \quad n(1-\pi) > 5$$

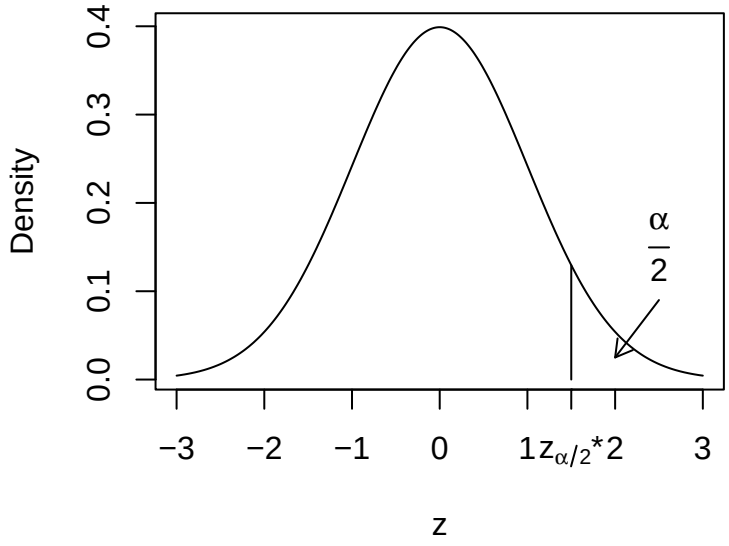
In general, this approximation works better the closer π is to 0.5.

Confidence Interval for π

Standard Interval:

The usual $100(1-\alpha)\%$ (approximate) confidence interval for π is given by

$$\hat{\pi} \pm z^*_{\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$



where

$$P[N(0,1) \geq z^*_{\alpha/2}] = \frac{\alpha}{2}$$

So a 95% CI for π_1 is given by the calculations

$$\hat{\pi} = \frac{63}{78} = 0.808$$

$$SE(\tilde{\pi}) = \sqrt{\frac{0.808(1-0.808)}{78}} = 0.0446$$

$$
\begin{aligned}
CI &= 0.808 \pm 1.96 \times 0.0446 \\
&= 0.808 \pm 0.087 \\
&= (0.720, 0.895)
\end{aligned}
$$

Treatment	$\hat{\pi}$	95% CI for π
Aspirin	0.808	0.808 ± 0.087
Placebo	0.558	0.558 ± 0.111

With this interval, you need to worry about the sample size considerations more.

One way to check it is to look at the quantities $nL, nU, n(1-L)$, and $n(1-U)$ (based on a interval of the form (L, U)) and check to see if they are all bigger than 5.

Now, lets consider the case when $S = 1$ and n=10. The 95% interval based on this data set is (-0.0859, 0.2859).

So for small sample sizes and S near 0 or n, the standard interval will contain a significant part of the interval less than 0 or greater than 1.

So for small sample sizes, we need an alternative procedure.

Agresti-Coull "Add Two Success and Two Failures" Interval:

This interval is based on an idea of EB Wilson (1927)

Note: This procedure is a bit different than the standard approach. For large n both procedures give similar answers. However for smaller n, this procedure tends to work better.

Instead of basing inference on $\hat{\pi}$, the quantity

$$\tilde{\pi} = \frac{S+2}{n+4}$$

is used. One way to think of this is to take the original sample and add 2 successes and 2 failures.

Then an approximate $100(1-\alpha)\%$ confidence interval for π is given by

$$\tilde{\pi} \pm z^*_{\alpha/2}\sqrt{\frac{\tilde{\pi}(1-\tilde{\pi})}{n+4}}$$

So this is the standard interval with modified data.

So a 95% CI for π_1 given by the calculations

$$\tilde{\pi} = \frac{63+2}{78+4} = 0.792$$

$$SE(\tilde{\pi}) = \sqrt{\frac{0.792(1-0.792)}{82}} = 0.0448$$

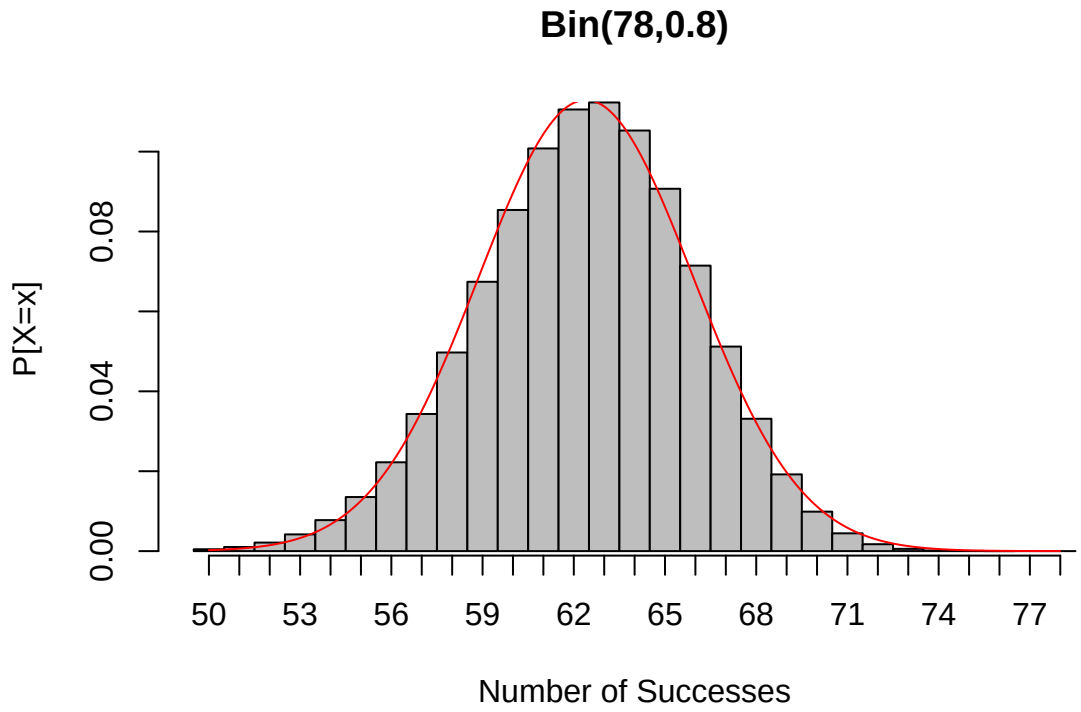
$$
\begin{aligned}
CI &= 0.793 \pm 1.96 \times 0.0448 \\
&= 0.793 \pm 0.088 \\
&= (0.705, 0.880)
\end{aligned}
$$

Treatment	$\hat{\pi}$	$\tilde{\pi}$	95% CI for π
Aspirin	0.808	0.793	0.793 ± 0.088
Placebo	0.558	0.556	0.556 ± 0.108

Note that the AC CI for π is not symmetric around $\hat{\pi}$, the standard estimate of π. Instead it is pulled towards $\frac{1}{2}$.

This is desirable, since the binomial distribution is not symmetric about its mean, unless $\pi = \frac{1}{2}$.

The asymmetry matches the skewness of the binomial distribution.



This adjustment from the standard interval tends to give better coverage (closer to desired confidence).

For larger sample sizes, the widths of the AC and the standard intervals are similar, as can be seen in this example. However the AC interval is shifted at bit towards 0.5.

Treatment	Standard Interval	AC Interval
Aspirin	0.808 ± 0.087	0.793 ± 0.088
Placebo	0.558 ± 0.111	0.556 ± 0.108

A Motivation for this interval

One approach to getting this interval is to approximate a Bayesian credibility interval.

- Prior: $\pi \sim Beta(2, 2)$
- Likelihood: $S|\pi \sim Bin(n, \pi)$
- Posterior: $\pi|S \sim Beta(S + 2, n - S + 2)$

  The posterior mean and variance are

$$E[\pi|S] = \frac{S+2}{n+4} = \tilde{\pi}$$

$$\text{Var}(\pi|S) = \frac{\tilde{\pi}(1-\tilde{\pi})}{n+4+1} \approx \frac{\tilde{\pi}(1-\tilde{\pi})}{n+4}$$

In addition, as n gets large, the posterior distribution of $\pi|S$ is approximately normal.

The original motivation for this interval was to approximate to Wilson's Score confidence interval which was shown to have better properties than the standard interval.

Find all π_0 s.t.

$$\left| \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \right| \leq z^*_{\alpha/2}$$

This is the acceptance region for the usual normal based test of

$$H_0 : \pi = \pi_0 \quad \text{vs} \quad H_A : \pi \neq \pi_0$$

Aside: Note that the AC interval and the standard interval have the same basic approach. You can think AC interval as an adjustment to the standard interval to fix the coverage properties of the standard interval in addition to its Bayesian and Score interval interpretations.

Clopper-Pearson Interval:

This is an exact confidence interval based on inverting the hypothesis test

$$H_0 : \pi = \pi_0 \quad \text{vs} \quad H_A : \pi \neq \pi_0$$

using exact binomial probabilities. This interval is conservative (i.e. the coverage level $\geq 1 - \alpha$ for every π). However it tends to be overly conservative in that it can give much wider intervals than needed.

One way of calculating the interval is

$$\left(F^{-1}\left(\tfrac{\alpha}{2}, S, n - S + 1\right), F^{-1}\left(1 - \tfrac{\alpha}{2}, S + 1, n - S\right)\right)$$

where $F^{-1}(q, a, b)$ is the quantile function of a $Beta(a, b)$ distribution.

This interval has the property of never going outside (0,1).

This interval is available in **R** via the functions `binom.test()` and `binom.exact()`. The second function is part of the library `epitools`.

For the $S = 1$ and n=10 case, lets look at the three intervals

```
> library(epitools)  # only need to run once per session

> binom.approx(1,10)  # standard interval
  x  n proportion       lower     upper conf.level
1 1 10        0.1 -0.08593851 0.2859385       0.95

> binom.ac(1,10)   # a function I wrote
  x  n proportion    ptilde         lower     upper conf.level
1 1 10        0.1 0.2142857 -0.0006521887 0.4292236       0.95

> binom.exact(1,10)
  x  n proportion       lower     upper conf.level
1 1 10        0.1 0.002528579 0.4450161       0.95
```

This exhibits what tends to happen for smaller sample sizes, the standard interval has problems but the AC interval and the Clopper-Pearson interval are similar.

For larger n, all three intervals act similarly, with the exact interval tending to be a bit wider than the other 2.

```
> binom.approx(63,78)
   x  n proportion     lower     upper conf.level
1 63 78  0.8076923 0.7202298 0.8951548       0.95

> binom.ac(63,78)
   x  n proportion   ptilde     lower     upper conf.level
1 63 78  0.8076923 0.792683 0.7049407 0.8804251       0.95

> binom.exact(63,78)
   x  n proportion     lower     upper conf.level
1 63 78  0.8076923 0.7027294 0.8881803       0.95
```

Sampling Distribution for the Difference of Two Proportions

One way of comparing the difference between aspirin and placebo in the example is to look at $\pi_1 - \pi_2$. (We will look at two other measures as well). The usual estimate of this quantity is $\hat{\pi}_1 - \hat{\pi}_2$. It has the following properties

- $E[\hat{\pi}_1 - \hat{\pi}_2] = \pi_1 - \pi_2$
- $\mathrm{Var}(\hat{\pi}_1 - \hat{\pi}_2) = \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}$
- If n_1 and n_2 are large, $\hat{\pi}_1 - \hat{\pi}_2$ is approximately normal distributed with the above mean and variance

  The rules of thumb for n_i being large enough depends on whether you are constructing a confidence interval for $\pi_1 - \pi_2$ or testing

$$H_0 : \pi_1 = \pi_2 \quad \text{vs} \quad H_A : \pi_1 \neq \pi_2$$

A simple rule that works in both cases is if the observed number of successes and the observed number of failures in each group is at least 5.

In either case, the normal approximation works better the closer the π_is are to 0.5.

Confidence Interval for $\pi_1 - \pi_2$

Standard Interval:

The usual $100(1-\alpha)\%$ (approximate) confidence interval for $\pi_1 - \pi_2$ is given by

$$(\hat{\pi}_1 - \hat{\pi}_2) \pm z^*_{\alpha/2}\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}$$

This interval does not work well when n_1 or n_2 (or both) are small. The rule of thumb mentioned earlier (all observed numbers of successes and failures > 5) works well here.

So a 95% CI for $\hat{\pi}_1 - \hat{\pi}_2$ is given by the calculations

$$\hat{\pi}_1 - \hat{\pi}_2 = \frac{63}{78} - \frac{43}{77} = 0.808 - 0.558 = 0.249$$

$$SE(\tilde{\pi}) = \sqrt{\frac{0.808(1-0.808)}{78} + \frac{0.558(1-0.558)}{77}} = 0.0721$$

$$
\begin{aligned}
CI &= 0.249 \pm 1.96 \times 0.0721 \\
&= 0.249 \pm 0.141 \\
&= (0.109, 0.391)
\end{aligned}
$$

So we have fairly good evidence that aspirin is better than placebo in preventing strokes.

Risk of Urethritis in Seminal Super Shedding (SSS) in HIV-I

A sample of 72 men infected by HIV-I were classified on whether they have had problems with urethritis. Within each group, the rate SSS was then examined. The question of interest was whether men with the urethritis tended to higher rates of the SSE form HIV-I.

Urethritis	SSS	No SSS	Total
No	6	60	66
Yes	3	3	6
Total	9	63	72

This dataset does not meet the rule of thumb for sample size. However we can still calculate the interval

```
> binom.approx2(6,66,3,6)  # again a function of mine
    prop1 prop2     diff      se   lower     upper conf.level
1 0.09091   0.5 -0.40909 0.20717 -0.8151 -0.003048       0.95
```

While this doesn't appear to give impossible values, this interval procedure has poor coverage properties with sample sizes like this.

So we need another procedure.

"Add Two Success and Two Failures" Interval:

This is a analogue to the Agresti-Coull procedure for a single proportion. (I'm not sure who's idea it was. It was not mentioned in the Agresti and Coull paper. My source for it is Moore and McCabe (2006), Introduction to the Practice of Statistics.) However similar motivations should give this interval.

This idea again is to add observations to each group. Instead in this case, we'll add one success and one failure in each group. The confidence interval is based on the estimates

$$\tilde{\pi}_1 = \frac{S_1 + 1}{n_1 + 2} \quad \text{and} \quad \tilde{\pi}_2 = \frac{S_2 + 1}{n_2 + 2}$$

The interval has the form

$$(\tilde{\pi}_1 - \tilde{\pi}_2) \pm z^*_{\alpha/2} \sqrt{\frac{\tilde{\pi}_1(1 - \tilde{\pi}_1)}{n_1 + 2} + \frac{\tilde{\pi}_2(1 - \tilde{\pi}_2)}{n_2 + 2}}$$

Again it uses the standard interval procedure, plugging in modified data.

So a 95% CI for $\hat{\pi}_1 - \hat{\pi}_2$ by this procedure for the HIV data is given by the calculations

$$\tilde{\pi}_1 - \tilde{\pi}_2 = \frac{6+1}{66+2} - \frac{3+1}{6+2} = 0.103 - 0.5 = -0.397$$

$$SE(\tilde{\pi}) = \sqrt{\frac{0.103(1-0.103)}{66+2} + \frac{0.5(1-0.5)}{6+2}} = 0.1806$$

$$\begin{aligned} CI &= -0.397 \pm 1.96 \times 0.1806 \\ &= -0.397 \pm 0.354 \\ &= (-0.751, -0.043) \end{aligned}$$

```
> binom.approx2(6,66,3,6)  # again a function of mine
    prop1 prop2     diff      se   lower     upper conf.level
1 0.09091   0.5 -0.40909 0.20717 -0.8151 -0.003048       0.95

> binom.ac2(6,66,3,6)
    prop1 prop2     diff      se    lower   upper conf.level
1 0.09091   0.5 -0.39706 0.18058 -0.7510 -0.04313       0.95
```

So in this example, it appears that there is a difference in SSS rates between the two groups.

The AC interval will tend to give a narrower interval. Also it will tend to be centered closer to 0, though it doesn't have to. Even though it is a narrower interval, it tends to have better coverage properties than the standard interval.

For large samples, the two procedure give similar intervals as can be seen in the stroke example

```
> binom.approx2(63,78,43,77)
   prop1  prop2   diff     se  lower  upper conf.level
1 0.8077 0.5584 0.2493 0.0721 0.1080 0.3905       0.95

> binom.ac2(63,78,43,77)
   prop1  prop2   diff     se  lower  upper conf.level
1 0.8077 0.5584 0.2430 0.0716 0.1027 0.3833       0.95
```

Tests on $\pi_1 - \pi_2$

While tests of the form

$$H_0 : \pi_1 - \pi_2 = c \quad \text{vs} \quad H_A : \pi_1 - \pi_2 \neq c$$

can be examined by the test statistic

$$z = \frac{(\hat{\pi}_1 - \hat{\pi}_2) - c}{SE(\hat{\pi}_1 - \hat{\pi}_2)}$$

I want to focus on the special case

$$H_0 : \pi_1 - \pi_2 = 0 \quad \text{vs} \quad H_A : \pi_1 - \pi_2 \neq 0$$

or equivalently

$$H_0 : \pi_1 = \pi_2 \quad \text{vs} \quad H_A : \pi_1 \neq \pi_2$$

(Note can also consider one-sided hypotheses.)

The case for a general c usually is not interesting. For example, researchers usually won't ask the question of whether a treatment will decrease a proportion by 0.1. It may not even make sense (i.e. what if $\pi = 0.05$). Instead they usually ask questions like whether a treatment will half the current proportion.

However the question of whether 2 proportions are the same or different is of interest.

For example, is the proportion of people having strokes different in the aspirin and placebo groups.

We want to base of inference on the statistic $\hat{\pi}_1 - \hat{\pi}_2$. What is the distribution of this under $H_0 : \pi_1 = \pi_2$, where π_c is the common, but unknown value of π_1 and π_2.

If we knew π_c,

$$\mathrm{Var}(\hat{\pi}_1 - \hat{\pi}_2) = \frac{\pi_c(1-\pi_c)}{n_1} + \frac{\pi_c(1-\pi_c)}{n_2} = \pi_c(1-\pi_c)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

and

$$z = \frac{\hat{\pi}_1 - \hat{\pi}_2}{SD_0(\hat{\pi}_1 - \hat{\pi}_2)}$$

is approximately $N(0,1)$ distributed where

$$SD_0(\hat{\pi}_1 - \hat{\pi}_2) = \sqrt{\pi_c(1-\pi_c)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

While we don't know π_c we can estimate it, assuming H_0 is true. In this case $\hat{\pi}_1$ and $\hat{\pi}_2$ are both estimates of this quantity. So some combination of these should give a better estimate (more data, better estimate).

In addition,

$$S_1 + S_2 \sim Bin(n_1 + n_2, \pi_c)$$

For this combined data, we get

$$\hat{\pi}_c = \frac{S_1 + S_2}{n_1 + n_2}$$

This can also be thought of as a combination of $\hat{\pi}_1$ and $\hat{\pi}_2$ as

$$\hat{\pi}_c = \frac{n_1}{n_1 + n_2}\hat{\pi}_1 + \frac{n_2}{n_1 + n_2}\hat{\pi}_2$$

(a weighted combination of $\hat{\pi}_1$ and $\hat{\pi}_2$ with the sample sizes as the weights).

This gives us an estimated standard error of

$$SE_0(\hat{\pi}_1 - \hat{\pi}_2) = \sqrt{\hat{\pi}_c(1 - \hat{\pi}_c)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

and a test statistic of

and

$$z = \frac{\hat{\pi}_1 - \hat{\pi}_2}{SE_0(\hat{\pi}_1 - \hat{\pi}_2)}$$

which asymptotically has a $N(0, 1)$ distribution in n_1 and n_2 are big enough.

A common rule of thumb is to check whether $n_s\hat{\pi}_c$ and $n_s(1-\hat{\pi}_c)$ are both at least 5, where $n_s = \min(n_1, n_2)$. This is equivalent to checking whether the expected number of successes and failure is at least 5 for both groups under H_0.

The p-values for this test are given by

Alternative Hypothesis	p-value
$H_A : \pi_1 - \pi_2 < 0$	$P[N(0,1) \leq z]$
$H_A : \pi_1 - \pi_2 > 0$	$P[N(0,1) \geq z]$
$H_A : \pi_1 - \pi_2 \neq 0$	$P[\lvert N(0,1)\rvert \geq \lvert z\rvert] = 2P[N(0,1) \geq \lvert z\rvert]$

So for the stroke example

$$\hat{\pi}_1 - \hat{\pi}_2 = 0.808 - 0.558 = 0.249$$

$$\hat{\pi}_c = \frac{63 + 43}{78 + 77} = 0.684$$

$$SE_0(\hat{\pi}_1 - \hat{\pi}_2) = \sqrt{0.684(1-0.684)\left(\frac{1}{78}+\frac{1}{77}\right)}$$
$$= 0.0747$$

$$z = \frac{0.249}{0.0747} = 3.337$$

$$p\text{-value} = 2P[N(0,1) \geq 3.337] = 0.00084 \qquad \text{(2-sided alternative)}$$

So as we have seen before, there is strong evidence in this data set the aspirin appears to drop the rate of strokes.

1-sided Versus 2-sided Tests

In most situations, you want to do 2-sided test, not 1-sided tests. In many cases, there is an alternative you want to see. For example, in the stroke example, you want to see $\pi_1 > \pi_2$ (aspirin better than placebo). However the other possibility is usually also possible (aspirin is actually harmful). Due to this you want to do the 2-sided test.

In some cases it is actually required. FDA regulations involving tests on treatment effects of candidate treatment require 2-sided tests. So if you don't specify you are doing 2-sided tests, the reviewers immediately double all your p-values. (Actually if the submission was this badly written, they probably would reject it.)

In the example in the text looking at CVD deaths and obesity, they are looking at the alternative hypothesis $H_A : \pi_1 - \pi_2 > 0$ and they get a p-valueof 0.37. I feel that in this case, the 2-sided test is more appropriate. Are there strong apriori reasons saying that obesity must increase the rate of deaths? I say no and in fact the writeup in the text suggests it could go the other way.

So in this case, the p-value should be doubled to 0.74. Note that it doesn't change the conclusion about statistical significance in this example, but there are situations where it can.