**Section 1.2 – Two binary variables**

This section extends the methods from Section 1.1 to a heterogeneous setting where individual items come from one of two groups. Because there is still a binary response, we can summarize the sample in a 2×2 contingency table. Below are a few examples.

Example: Larry Bird (data source: Wardrop, *American Statistician*, 1995)

Free throws are typically shot in pairs. Below is a contingency table summarizing Larry Bird’s first and second free throw attempts during the 1980-1 and 1981-2 NBA seasons:

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Second |   |
|   |   | Made | Missed | Total |
| First | Made | 251 | 34 | 285 |
| Missed | 48 | 5 | 53 |
|   | Total | 299 | 39 | 338 |

Interpreting the table:

* 251 first and second free throw attempts were both made
* 34 first free throw attempts were made and the second were missed
* 48 first throw attempts were missed and the second free throw were made
* 5 first and second free throw attempts were both missed
* 285 first free throws were made regardless what happened on the second attempt
* 299 second free throws were made regardless what happened on the first attempt
* 338 free throw pairs were shot during these seasons

Note that the “total” rows and columns are not included when calling this is a 2×2 table. The total rows and columns are often included to help with some calculations that we will discuss later.

What types of questions would be of interest for this data?

Example: Field goals (data source: Bilder and Loughin, *Chance*, 1998)

Below is a two-way table summarizing field goals from the 1995 NFL season. The data can be considered a representative sample from the population. The two categorical variables in the table are stadium type (dome or outdoors) and field goal result (success or failure).

|  |  |  |
| --- | --- | --- |
|  | Field goal result |   |
| Success | Failure | Total |
| Stadium type | Dome | 335 | 52 | 387 |
| Outdoors | 927 | 111 | 1038 |
|   | Total | 1262 | 163 | 1425 |

What types of questions would be of interest for this data?

Example: Polio vaccine clinical trials (data source: Francis et al., *American Journal of Public Health*, 1955)

In the clinical trials for the polio vaccine developed by Jonas Salk, two large groups were involved in the placebo-control phase of the study. The first group, which received the vaccination, consisted of 200,745 individuals. The second group, which received a placebo, consisted of 201,229 individuals. There were 57 cases of polio in the first group and 142 cases of polio in the second group.

|  |  |  |  |
| --- | --- | --- | --- |
|   | Polio | Polio free | Total |
| Vaccine | 57 | 200,688 | 200,745 |
| Placebo | 142 | 201,087 | 201,229 |
| Total | 199 | 401,775 | 401,974 |

What types of questions would be of interest for this data?

Example: HIV vaccine clinical trials (data source: Rerks-Ngram et al., *New England Journal of Medicine*, 2009)

Clinical trials have been performed to evaluate the effectiveness of a number of HIV vaccines. The results of one trial in particular were discussed a lot in the media in 2009 (see book’s website for links to news stories). Below is a contingency table summarizing the data used in the “modified intent-to-treat analysis” of the paper.

|  |  |  |  |
| --- | --- | --- | --- |
|  | HIV | HIV free |  |
| Vaccine | 51 | 8,146 | 8,197 |
| Placebo | 74 | 8,124 | 8,198 |
|  | 125 | 16,270 | 16,395 |

Exercises 19-21 provide additional details. What types of questions would be of interest for this data?

Contingency tables can be larger than 2×2 table! These will be discussed in Chapter 3.

**Section 1.2.1 – Notation and model**

* Let Y11, …,  be Bernoulli random variables for group 1 (row 1 of the contingency table).
* Let Y12, …,  be Bernoulli random variables for group 2 (row 2 of the contingency table).
* The number of “successes” for a group is represented by .
* Wj has a binomial distribution with success probability πj and number of trials of nj
* W1 is independent of W2; thus, we have an “independent binomial model”. Some people refer to this as “independent binomial sampling” as a way to describe how the contingency table counts come about.
* The MLE of πj is 
* A “+” in a subscript is used to denote indices in a subscript that are being summed over. For example, W+ = W1 + W2 is the total number of successes and n+ = n1 + n2 is the total sample size. In fact, .

Below is a contingency table summarizing the notation:

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Response |  |
|  |  | 1 | 2 |  |
| Group | 1 | w1 | n1 – w1 | n1 |
| 2 | w2 | n2 – w2 | n2 |
|  |  | w+ | n+ – w+ | n+ |

Below is the table again, but with probabilities rather than counts within cells:

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Response |  |
|  |  | 1 | 2 |  |
| Group | 1 | π1 | 1 – π1 | 1 |
| 2 | π2 | 1 – π2 | 1 |

Comments:

* Please see the example in the book for how to simulate data under this model structure.
* Larger contingency tables can occur too! I am postponing discussions of these cases until later chapters when different types of model structures are more useful.

Example: Larry Bird (Bird.R)

The purpose of this example is to show how to work with a contingency table structure in R.

> c.table<-array(data = c(251, 48, 34, 5), dim = c(2,2),

 dimnames = list(First = c("made", "missed"), Second =

 c("made", "missed")))

> c.table

 Second

First made missed

made 251 34

missed 48 5

> list(First = c("made", "missed"), Second = c("made",

 "missed"))

$First

[1] "made" "missed"

$Second

[1] "made" "missed"

To access parts of the table, we can refer to the cells by their rows and columns:

> c.table[1,1] #w1

[1] 251

> c.table[1,] #w1 and n1-w1

 made missed

 251 34

> sum(c.table[1,]) #n1

[1] 285

To calculate the probability of success for each group:

> rowSums(c.table) #n1 and n2

 made missed

 285 53

> pi.hat.table<-c.table/rowSums(c.table)

> pi.hat.table

 Second

First made missed

 made 0.8807018 0.11929825

 missed 0.9056604 0.09433962

The estimated probability that Larry Bird makes his second free throw attempt is  = 0.8807 given that he makes the first and  = 0.9057 given he misses the first.

This is somewhat counterintuitive to most basketball fans perceptions that a missed first free throw should lower the probability of success on the second free throw. However, this is only for one sample. We would like to generalize to the population of all free throw attempts by Larry Bird. In order to make this generalization, we need to use statistical inference procedures.

Suppose the data was in a “raw” form of a data frame where each row represented a (group, response) pair. This is the form that you would expect the data to be in first. The table() and xtabs() function can be used to create a contingency table:

> #Suppose data is in all.data2 – see program

> head(all.data2)

 first second

1 made made

2 missed made

3 made missed

4 missed missed

5 made made

6 missed made

> bird.table1<-table(all.data2$first, all.data2$second)

> bird.table1

 made missed

made 251 34

missed 48 5

> bird.table1[1,1] #w1

[1] 251

> bird.table2<-xtabs(formula = ~ first + second, data =

 all.data2)

> bird.table2

 second

first made missed

 made 251 34

 missed 48 5

> bird.table2[1,1] #w1

[1] 251

**Section 1.2.2 – Confidence intervals for the difference of two probabilities**

Remember from Section 1.1 that the estimated probability of success  can be treated as an approximate normal random variable with mean π and variance  for a large sample. Using the notation in this chapter, this means that

 ~ N(1, 1(1 – 1)/n1) and

 ~ N(2, 2(1 – 2)/n2)

approximately for large n1 and n2. Note that  and  are treated as random variables here, not the observed values as in the last example.

The statistic that estimates  is . The distribution can be approximated by

N(1 – 2, 1(1 – 1)/n1 + 2(1 – 2)/n2)

for large n1 and n2.

Note:  because  and  are independent random variables. Some of you may have seen the following: Let X and Y be independent random variables and let a and b be constants. Then Var(aX+bY) = a2Var(X) + b2Var(Y).

The estimate of the variance is then



A (1 – α)100% Wald confidence interval for 1 – 2 is then

 ± Z1-α/2

Do you remember the problems with the Wald confidence interval in Section 1.1? Similar problems occur here ☹.

Agresti and Caffo (2000) recommend adding two successes and two failures to the data for an interval of ANY level of confidence.

Let  and . The Agresti-Caffo confidence interval is



Agresti and Caffo do not change the adjustment for different confidence levels!

Example: Larry Bird (Bird.R)

The code below continues the code from earlier:

> alpha<-0.05

> pi.hat1<-pi.hat.table[1,1]

> pi.hat2<-pi.hat.table[2,1]

> #Wald

> var.wald<-pi.hat1\*(1-pi.hat1) / sum(c.table[1,]) +

 pi.hat2\*(1-pi.hat2) / sum(c.table[2,])

> pi.hat1 - pi.hat2 + qnorm(p = c(alpha/2, 1-alpha/2)) \*

 sqrt(var.wald)

[1] -0.11218742 0.06227017

> #Agresti-Caffo

> pi.tilde1<-(c.table[1,1] + 1) / (sum(c.table[1,]) + 2)

> pi.tilde2<-(c.table[2,1] + 1) / (sum(c.table[2,]) + 2)

> var.AC<-pi.tilde1\*(1-pi.tilde1) / (sum(c.table[1,]) + 2)

 + pi.tilde2\*(1-pi.tilde2) / (sum(c.table[2,]) + 2)

> pi.tilde1 - pi.tilde2 + qnorm(p = c(alpha/2, 1-alpha/2))

 \* sqrt(var.AC)

[1] -0.10353254 0.07781192

Therefore, the 95% Wald confidence interval is

-0.1122 < π1 - π2 < 0.0623

and the 95% Agresti-Caffo confidence interval is

-0.1035 < π1 - π2 < 0.0778

There is not sufficient evidence to indicate a difference in the proportions. What does this mean in terms of the original problem?

Other ways to perform these calculations:

> w1<-251

> n1<-285

> w2<-48

> n2<-53

> alpha<-0.05

> pi.hat1<-w1/n1

> pi.hat2<-w2/n2

> var.wald<-pi.hat1\*(1-pi.hat1) / n1 + pi.hat2\*(1-pi.hat2)

 / n2

> pi.hat1 - pi.hat2 + qnorm(p = c(alpha/2, 1-alpha/2)) \*

 sqrt(var.wald)

[1] -0.11218742 0.06227017

> #Calculations using the PropCIs package

> library(package = PropCIs)

> #Wald

> wald2ci(x1 = c.table[1,1], n1 = sum(c.table[1,]), x2 =

 c.table[2,1], n2 = sum(c.table[2,]), conf.level = 0.95,

 adjust = "Wald")

data:

95 percent confidence interval:

 -0.11218742 0.06227017

sample estimates:

[1] -0.02495862

> #Agresti-Caffo

> wald2ci(x1 = c.table[1,1], n1 = sum(c.table[1,]), x2

 = c.table[2,1], n2 = sum(c.table[2,]), conf.level =

 0.95, adjust = "AC")

data:

95 percent confidence interval:

 -0.10353254 0.07781192

sample estimates:

[1] -0.01286031

True confidence level

Below are two plots from the Agresti and Caffo (2000) comparing the estimated true confidence levels of the Agresti-Caffo interval to the Wald interval. The solid line denotes the Agresti-Caffo interval. The y-axis shows the true confidence level (coverage) of the confidence intervals. The x-axis shows various values of π1 where π2 is fixed at 0.3.



For the plots below, the value of π1 was no longer fixed.



The Agresti and Caffo interval tends to be much better than the Wald interval.

Note that other confidence intervals can be calculated. Agresti and Caffo’s (2000) objective was to present an interval that was “better” than the Wald which could be used in elementary statistics courses. See Newcombe (*Statistics in Medicine*, 1998, p. 857-872) for a review of other intervals. In particular, a good interval is the score interval. This interval is discussed in Exercise 24.

How can we calculate these true confidence levels?

**Section 1.2.3 – Test for the difference of two probabilities**

Typically, confidence intervals are preferred over hypothesis tests when a simple set of parameters, like π1 – π2, are of interest. This is because a confidence interval gives a range of possible parameter values, which a hypothesis test cannot.

Still, if you want to perform a hypothesis test of H0:π1 – π2 = 0 vs. Ha: :π1 – π2 ≠ 0 using a test statistic and p-value, one way is to use the test statistic of



where . This test statistic has a standard normal distribution for a large sample. Therefore, you can reject H0 if |Z0| > Z1-α/2.

Question: Why is  used in the test statistic?

Note that Z0 is another example of a score test statistic. In fact, a confidence interval can be constructed using this statistic in much the same way as we saw in Section 1.1 when computing a confidence interval for π. Unfortunately, there is not a closed form expression for the confidence interval that can be presented; i.e., there is not one equation that can be written out to compute the interval. Again, Exercise 24 in my book provides details on its computation (the diffscoreci() function from the PropCIs package computes the interval).

Another way to perform the hypothesis test is through using a Pearson chi-square test. This is a general procedure that can be used for a large number of problems including the test of interest here.

The general form of a test statistic of this type is



which is formed across all cells of the contingency table. The estimated expected count is what we would expect a cell count to be if the null hypothesis was true. For our test, the estimated expected count is  for the success column and  for the failure column. The test statistic is



Through using some algebra, we can simplify the test statistic to be



For large samples, this test statistic has an approximate  probability distribution. Large values indicate evidence against the null hypothesis so that we reject H0 if 

Question: Why do large values indicate evidence against the null hypothesis?

One can show that ! Also, because the square of a standard normal random variable is the same as chi-square random variable (e.g., ), the Pearson chi-square test is the same as the score test here!

Example: Larry Bird (Bird.R)

The purpose of this problem is to test H0: π1 – π2 = 0 vs. Ha: π1 – π2 ≠ 0 (1 = First FT made, 2 = First FT missed)

Note that . Then



Because -1.96 < -0.5222 < 1.96, do not reject H0 when α = 0.05. There is not sufficient evidence to conclude that the probability of success on the second free throw differs based on what happened for the first free throw.

The code below continues the code from earlier:

> prop.test(x = c.table, conf.level = 0.95, correct =

 FALSE)

 2-sample test for equality of proportions without continuity correction

data: c.table

X-squared = 0.2727, df = 1, p-value = 0.6015

alternative hypothesis: two.sided

95 percent confidence interval:

-0.11218742 0.06227017

sample estimates:

 prop 1 prop 2

0.8807018 0.9056604

Note that R gives . Also, R gives the Wald confidence interval in its output.

There are other ways to perform this test in R. For example,

> chisq.test(x = c.table, correct = FALSE)

 Pearson's Chi-squared test

data: c.table

X-squared = 0.2727, df = 1, p-value = 0.6015

Another way to perform a test for the difference of two probabilities is through a likelihood ratio test (LRT). The test statistic is



Notice the similar format to the statistic compared to what we saw in Section 1.1. The null hypothesis is rejected if .

Larntz (1978) showed that the Score (Pearson) test is better to use than the LRT in these situations. Please see the book for further details on the LRT.

**Section 1.2.4 – Relative risks**

Frequently, the news media will report the results of a study and say something like “the researchers showed that an individual was <number> times more likely to get <disease> if they did <one behavior choice> than <second behavior choice>. IF the reporter understands basic statistics, they are reporting about a numerical measure called the *relative risk*.

One problem with basing inference on π1 – π2 is that it measures a quantity whose meaning changes depending on the sizes of π1 – π2. For example, consider the following two scenarios:

|  |  |  |
| --- | --- | --- |
|  | Adverse reactions |   |
|   | Yes | No | Total |
| Drug | π1 = 0.510 | 1 – π1 = 0.490 | 1 |
| Placebo | π2 = 0.501 | 1 – π2 = 0.499 | 1 |

π1 – π2 = 0.510 – 0.501 = 0.009

|  |  |  |
| --- | --- | --- |
|  | Adverse reactions |   |
|   | Yes | No | Total |
| Drug | π1 = 0.010 | 1 – π1 = 0.990 | 1 |
| Placebo | π2 = 0.001 | 1 – π2 = 0.999 | 1 |

π1 – π2 = 0.010 – 0.001 = 0.009

In the first scenario, an increase of 0.009 is rather small **relative** to the already sizable probabilities given for the two groups. On the other hand, the second scenario has a much larger adverse reaction probability for the drug group **relative** to the placebo group.

We need to be able to convey the relative magnitudes of these changes better than differences allow. The relative risk allows us to make these comparisons by taking the ratio of the two success probabilities:



For scenario #2 above, the relative risk is RR = 0.010/0.001 = 10. The interpretation for this case is:

An adverse reaction is 10 *times as likely* for those individuals taking the drug than those individuals taking the placebo.

or

The probability of an adverse reaction is 10 *times as large* for those individuals taking the drug than those individuals taking the placebo.

where “likely” and “probability” are synonymous. Alternatively, we could also say,

An adverse reaction is 9 *times* *more* *likely* for individuals taking the drug than those individuals taking the placebo.

or

The probability of an adverse reaction is 9 *times* *larger* for individuals taking the drug than those individuals taking the placebo.

Please see the top of p. 38 of my book for additional wording.

Why does one interpretation have 10 (RR) and another have 9 (RR – 1)? It can be helpful to look at some examples.

 “1 times as likely” is equivalent to π1/π2 = 1. In other words, they are equal. “2 times as likely” is equivalent to π1/π2 = 2. In other words, π1 is twice the size of π2; π1 = 2×π2.

“2 times more likely” is equivalent to π1/π2 = 3. The “more” there is what causes the difference from the previous interpretation.

As another example, π1/π2 = 1.5 means that a success is 50% more likely for group 1 than for group 2. Alternatively, a success is 1.5 times as likely for group 1 than for group 2.

Other interpretations are possible depending on the context of the application. Please see the upcoming example for a specific case.

Questions:

* What does a relative risk of 1 mean?
* What is the numerical range of the relative risk?

The MLE of RR can be found by substituting MLEs of π1 and π2 into the equation for RR:



As with other maximum likelihood estimators, we could use a normal approximation with the statistic here and form a Wald confidence interval. However, the large sample normal approximation can be improved upon by working with the log transformation of the relative risk first. Thus, a (1 – α)100% Wald confidence interval for is

 

where



is derived through using a delta method approximation (see Appendix B).

Of course, we are still interested in RR itself, so we can apply the exponential function to find the (1 – α)100% Wald confidence interval for RR:



What if w1 or w2 is equal to 0? You will have difficulty calculating the variance used in the interval! An ad-hoc solution is to add a small constant, such as 0.5 to the wj and its corresponding nj count.

Example: Polio vaccine clinical trials (Salk.R)

|  |  |  |  |
| --- | --- | --- | --- |
|   | Polio | Polio free | Total |
| Vaccine | 57 | 200,688 | 200,745 |
| Placebo | 142 | 201,087 | 201,229 |
| Total | 199 | 401,775 | 401,974 |

> c.table<-array(data = c(57, 142, 200688, 201087), dim =

 c(2,2), dimnames = list(Treatment = c("vaccine",

 "placebo"), Result = c("polio", "polio free")))

> c.table

 Result

Treatment polio polio free

 vaccine 57 200688

 placebo 142 201087

> pi.hat.table<-c.table/rowSums(c.table)

> pi.hat.table

 Result

Treatment polio polio free

 vaccine 0.0002839423 0.9997161

 placebo 0.0007056637 0.9992943

> pi.hat1<-pi.hat.table[1,1]

> pi.hat2<-pi.hat.table[2,1]

> round(pi.hat1/pi.hat2, 4)

[1] 0.4024

> round(1/(pi.hat1/pi.hat2), 4) #inverted

[1] 2.4852

> alpha<-0.05

> n1<-sum(c.table[1,])

> n2<-sum(c.table[2,])

> var.log.rr<-(1-pi.hat1) / (n1\*pi.hat1) + (1-pi.hat2) /

 (n2\*pi.hat2)

> ci<-exp(log(pi.hat1/pi.hat2) + qnorm(p = c(alpha/2, 1-

 alpha/2)) \* sqrt(var.log.rr))

> round(ci, 4)

[1] 0.2959 0.5471

> rev(round(1/ci, 4)) #inverted

[1] 1.8278 3.3792

> #Could also calculate the variance like this:

> 1/c.table[1,1] - 1/sum(c.table[1,]) + 1/c.table[2,1] –

 1/sum(c.table[2,])

[1] 0.02457616

> var.log.rr

[1] 0.02457616

Defining index 1 to represent the vaccine group and 2 the placebo group, we find  = 0.40. The **estimated** probability of contracting polio is 0.4 times as large for those individuals receiving the vaccine than those receiving the placebo.

The 95% confidence interval is 0.30 < RR < 0.55. Therefore, with 95% confidence, probability of contracting polio is between 0.30 and 0.55 times as large for those individuals receiving the vaccine than those receiving the placebo. We could also say:

Comments:

1. For cases like this, we are interested in “risk reduction” provided by one of the groups, so it is natural to interpret the relative risk in this manner. For example, we could say

With 95% confidence, the vaccine reduces the risk of contracting polio by 45% to 70%.

1. In other situations, we will want to invert the relative risk so that we are always interpreting a quantity greater than 1. For example, the estimated probability of contracting polio is 2.49 times as large for the placebo group than the vaccine group. This interpretations is probably not as desirable.
2. Notice that “estimated” is not used for the confidence interval interpretation because it is stated for a particular parameter not the statistic (estimate).
3. Would you rather be in the placebo or vaccine group? Why?
4. You may have expected to see a more dramatic reduction in polio with the vaccine group. Perhaps the smaller reduction than expected is due to people who had polio going into the trial that were unknown to the researchers.

Example: Larry Bird (Bird.R)

Examining the relative risk may not be as important for this problem, but we can still calculate it.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Second |   |
|   |   | Made | Missed | Total |
| First | Made | =0.8807 | =0.1193 | 1 |
| Missed | =0.9057 | =0.0943 | 1 |

Continuing the R code from before

> cat("The sample relative risk is", round(pi.hat1/pi.hat2,

 4), "\n \n")

 The sample relative risk is 0.9724

> alpha<-0.05

> n1<-sum(c.table[1,])

> n2<-sum(c.table[2,])

> ci<-exp(log(pi.hat1/pi.hat2) + qnorm(p = c(alpha/2, 1-

 alpha/2)) \* sqrt((1-pi.hat1)/(n1\*pi.hat1) + (1-

 pi.hat2)/(n2\*pi.hat2)))

> round(ci, 4)

[1] 0.8827 1.0713

> rev(round(1/ci, 4)) #inverted

[1] 0.9334 1.1329

/= 0.8807/0.9057 = 0.9724

If the relative risk is inverted: / = 0.9057/0.8807 = 1.0284. Thus, a successful second free throw is estimated to be 2.84% times more likely to occur when the first free throw is missed rather than made. Alternatively, a successful second free throw attempt is estimated to be 1.0284 times as likely when the first free throw is missed rather than made.

With 95% confidence, a second free throw success is between 0.9334 and 1.1329 times as likely when the first free throw is missed rather than made.

Questions:

1. What else could be said here if one wanted to do a hypothesis of H0: π1/π2 = 1 vs. Ha: π1/π2 ≠ 1.
2. What if the interval was 2 < π1/π2 < 4?

Perhaps it may be better to examine a relative risk with respect to a missed second free throw attempt:

> (1-pi.hat1)/(1-pi.hat2)

[1] 1.264561

> exp(log((1-pi.hat1)/(1-pi.hat2)) + qnorm(p = c(alpha/2,

 1-alpha/2)) \* sqrt((pi.hat1)/(n1\*(1-pi.hat1)) +

 (pi.hat2)/(n2\*(1-pi.hat2))))

[1] 0.5183628 3.0849352

A missed second free throw is estimated to be 26.5% times more likely to occur when the first free throw is made than when it is missed. While this may sound to be an important finding, notice the confidence interval contains 1.

**Section 1.2.5 – Odds ratios**

*Odds* are the probability of success divided by the probability of a failure. With respect to a 2×2 contingency table, we have

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Response |  |
|  |  | 1 = success | 2 = failure |  |
| Group | 1 | π1 | 1 – π1 | 1 |
| 2 | π2 | 1 – π2 | 1 |

* For row 1, the “odds of a success” are
odds1 = π1/(1-π1).
* For row 2, the “odds of a success” are
odds2 = π2/(1-π2).

Notice that the odds are just a rescaling of the probability of success! For example, if P(success) = 0.75, then the odds are 3 or “3 to 1 odds”. The probability of a success are three times as large as the probability of a failure.

The estimated odds are:

 and 

Notice what cells these correspond to in the contingency table:

|  |  |  |
| --- | --- | --- |
|  |  | Response |
|  |  | 1 = success | 2 = failure |
| Group | 1 | w1 | n1 – w1 |
| 2 | w2 | n2 – w2 |

Questions:

* What is the numerical range of an odds?
* What does it mean for an odds to be 1?

To incorporate information from both rows 1 and 2 into a single number, the ratio of the two odds is found. This is called an “odds ratio”. Formally, it is defined as:



**Odds ratios are VERY useful in categorical data analysis and will be used throughout this course!**

Questions:

* What is the numerical range of OR?
* What does it mean for OR to be 1?
* What does it mean for OR > 1?
* What does it mean for OR < 1?

The estimated odds ratio is



This is the MLE of OR. Notice that the estimate is a product of the counts on the “diagonal” (top-left to bottom-right) of the contingency table divided by a product of the counts on the off diagonal.

Interpretation

The interpretation of OR can be difficult for some students. I recommend that you always go back to the basic property that an odds ratio is the ratio of two odds. Below are some example interpretations:

1. The estimated odds of a success are  times as large as in group 1 than in group 2.
2. The estimated odds of a success are  times as large as in group 2 than in group 1.

Notice that case 2) inverts the odds ratio so that row 2 is divided by row 1.

Consider the case now of interpreting the odds of a failure (1-π1)/ π1. Now, the ratio of group 1 to group 2 becomes:



This leads to the following additional interpretations:

1. The estimated odds of a failure are  times as large as in group 1 than in group 2.
2. The estimated odds of a failure are  times as large as in group 2 than in group 1.

Confidence interval

Because  is a maximum likelihood estimate, we can use the “usual” properties of them to find the confidence interval. However, using the log() often works better (i.e., its distribution is closer to being a normal distribution). It can be shown that:

* log() has an approximate normal distribution with mean log(OR) for large n.
*  = . This is derived through using a delta method approximation (see Appendix B).

The (1 – α)100% Wald confidence interval for log(OR) is



The (1 – α)100% Wald confidence interval for OR is



Lui and Lin (*Biometrical Journal*, 2003, p. 231) show this interval is conservative. What does “conservative” mean here?

Problems with small cell counts

### What happens to  if a cell count is 0?

When there is a 0 or small cell count, the estimator is changed a little to prevent problems. The estimator becomes



Thus, 0.5 is added to each cell count. The estimate of  becomes



and the confidence interval for OR can be found using the same form as before.

Sometimes, a small number is just added to a cell with a 0 count instead.

Example: Larry Bird (Bird.R)

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Second |   |
|   |   | Made | Missed | Total |
| First | Made | 251 | 34 | 285 |
| Missed | 48 | 5 | 53 |
|   | Total | 299 | 39 | 338 |

> OR.hat<-c.table[1,1] \* c.table[2,2] / (c.table[2,1] \*

 c.table[1,2])

> round(OR.hat, 2)

[1] 0.77

> round(1/OR.hat, 2)

[1] 1.3

.

Interpretation:

* The estimated odds of a made second free throw attempt are 0.7690 times as large when the first free throw is made than when the first free throw is missed.
* The estimated odds of a made second free throw attempt are 1/0.7690 = 1.3 times as large when the first free throw is missed than when the first free throw is made.

In actual application, you would only present one of these interpretations. My preference is for the second interpretation. Also, one could also phrase the interpretation as

The estimated odds of a made second free throw attempt are 30% times larger when the first free throw is missed than when the first free throw is made.

Below are examples of how to INCORRECTLY interpret odds ratios:

* “The estimated odds of a made second free throw attempt are 1.3 times as **likely** … ” is incorrect because “likely” means probabilities are being compared.
* Replacing “odds” with “probability” in any correct interpretation.
* “The estimated odds are 1.3 times higher …” is incorrect because 1.3 means 30% times higher not 130%. Please see the relative risk interpretation discussion when using “times more likely” and “times larger”.

To find the confidence interval:

> alpha<-0.05

> var.log.or<-1/c.table[1,1] + 1/c.table[1,2] +

 1/c.table[2,1] + 1/c.table[2,2]

> OR.CI<-exp(log(OR.hat) + qnorm(p = c(alpha/2, 1-alpha/2))

 \* sqrt(var.log.or))

> round(OR.CI, 2)

[1] 0.29 2.07

> rev(round(1/OR.CI, 2))

[1] 0.48 3.49

The 95% confidence interval for OR is 0.29 < OR < 2.07. If the interval is inverted, the 95% confidence interval for 1/OR is 0.48 < 1/OR < 3.49.

The interpretation can be extended to:

With 95% confidence, the odds of a made second free throw attempt are between 0.48 and 3.49 times as large when the first free throw is missed than when the first free throw is made.

Notice that I do not include the word “estimated” here.

Because 1 is in the interval, there is not sufficient evidence to indicate that the first free throw result has an effect on the second free throw result.

Example: Polio vaccine clinical trials (Salk.R)

|  |  |  |  |
| --- | --- | --- | --- |
|   | Polio | Polio free | Total |
| Vaccine |  57  |  200,688  |  200,745  |
| Placebo |  142  |  201,087  |  201,229  |

R code and output:

> OR.hat<-c.table[1,1] \* c.table[2,2] / (c.table[2,1] \*

 c.table[1,2])

> round(OR.hat, 4)

[1] 0.4024

> round(1/OR.hat, 4)

[1] 2.4863

> alpha<-0.05

> var.log.or<-1/c.table[1,1] + 1/c.table[1,2] +

 1/c.table[2,1] + 1/c.table[2,2]

> OR.CI<-exp(log(OR.hat) + qnorm(p = c(alpha/2, 1-alpha/2))

 \* sqrt(var.log.or))

> round(OR.CI, 2)

[1] 0.30 0.55

> rev(round(1/OR.CI, 2))

[1] 1.83 3.38

Interpretation:

* The estimated odds of contracting polio are 0.40 times as large when the vaccine is given instead of a placebo.
* The estimated odds of being polio free are 2.49 times as large as when the vaccine is given than when the placebo is given. Notice that I used the “The estimated odds of a failure …” interpretation here.
* One could also phrase the interpretation as

The estimated odds of contracting polio are reduced by 60% when the vaccine is given instead of a placebo.

The 95% confidence interval is 0.30 < OR < 0.55. If the interval is inverted, the 95% confidence interval is 1.83 < 1/OR < 3.38. With 95% confidence, the odds of being polio free are between 1.83 and 3.38 times as large when the vaccine is given instead of the placebo.

Would you want to receive the vaccine?

Please see the discussion in the book about when the RR and OR are similar in numerical value.

**Section 1.2.6 – Matched pairs data**

When comparing two probabilities, we have focused on comparing π1 and π2 in

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Response |  |
|  |  | 1 = success | 2 = failure |  |
| Group | 1 | π1 | 1 – π1 | 1 |
| 2 | π2 | 1 – π2 | 1 |

For this situation, there is one response variable denoted by the columns of the table (success and failure). We treat the number of successes as being observed from two binomial distributions with probabilities of success π1 or π2 depending on the group (row).

There are other situations where both the row and column designations can be thought of as response variables. For example, consider the situation where a new disease diagnosis test is being used and compared

to a “gold standard” diagnosis test:

|  |  |  |
| --- | --- | --- |
|  |  | Standard test |
|  |  | 1 = + | 2 = – |
| New test | 1 = + |  |  |
| 2 = – |  |  |

In this situation, we would like to compare P(Standard test = +) to the P(New test = +) as a way to evaluate the new test. Thus, we are comparing marginal probabilities for both tests.

Please see my book for details about how to compare these two marginal probabilities.

**Section 1.2.7 – Larger contingency tables**

While Section 1.2 focuses on two groups with binary responses, there are many instances where more than two groups (i.e., there are more than two rows within a contingency table) exist. These types of probabilities are better handled with the analysis methods of Chapter 2. Plus, more complex scenarios can be considered.

There may also be more than two response categories for the column variable as well. Combined with more than two rows, a contingency table with I rows and J columns can be formed. In this case, we can easily extend the use of the multinomial distribution that we saw in the last section. Alternatively, we can also use the Poisson distribution to model the counts in the table. I will postpone discussions of these situations until Chapters 3 (multinomial) and 4 (Poisson).