**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 (a.k.a., two-way table). Below are a few examples.

Example: Larry Bird; data from Wardrop (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 from Bilder and Loughin (1998)

Below is a summarizing field goals from one NFL season. The two categorical variables in the table are stadium type and field goal result.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | 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: COVID-19 vaccine clinical trial; data from Polack et al. (2020)

Clinical trials are needed to examine the effectiveness of of vaccines. Below is a contingency table summarizing the data used for the Pfizer-BioNTech COVID-19 vaccine.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Positive | Negative |  |
| Vaccine | 8 | 17,403 | 17,411 |
| Placebo | 162 | 17,349 | 17,511 |
|  | 170 | 34,752 | 34,922 |

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

Contingency tables can be larger than a 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 |

Please see the example in the book for how to simulate data under this model structure.

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. 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 made missed

3 made missed

4 made made

5 made made

6 made 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