Modeling Binary Data and the Concept of Odds

Learning Objectives

In this lecture, you will learn the following:

- What binary data are and how to summarize them
- The concept of **odds** and how to interpret them

Odds of an Event

The Odds of an Event are defined as

$$Odds(\ Event\ ) = \frac{\text{Probability that an event occurs}}{\text{Probability that an event does not occur}}$$

Let P be the probability that an event occurs

So, the probability that an event does not occur is by given 1 - p

$$Odds(\ Event\ ) = \frac{P}{1 - P}$$

The odds of an event measure how likely the event is to occur compared to it not occurring.

Introduction

In many situations, we want to examine how the probability of an event depends on other factors. For example, spending more time studying should increase the probability of earning a passing grade.

Example

The probability that a student **drives to school** is 0.8

Question: Calculate the odds of **driving** and interpret it.

$$Odds(\boldsymbol{Drive}) = \frac{\text{Probability of driving}}{\text{Probability of not driving}}$$

Probability of driving ⟵ 0.8

Probability of not driving ⟵ 0.2 (= 1- 0.8)

$$= \frac{0.8}{0.2} = 4$$

Interpretation:

Driving to school is **4 times as likely as** not driving

Odds of an Event – Alternative Interpretation

Let's say we randomly select **n** students (e.g. 100)

$$\underbrace{Odds(\boldsymbol{Drive})}_{\downarrow} = \frac{\text{Probability of driving}}{\text{Probability of not driving}} \times \frac{n}{n}$$

$$= \frac{\text{Expected Number of students driving}}{\text{Expected Number of students not driving}}$$

$$4 = \frac{4}{1} = \frac{\text{Expected Number of students driving}}{\text{Expected Number of students not driving}}$$

Interpretation:

For 4 students who drive to school, we expect 1 does not drive.

Exercise

The probability that a student **bikes to school** is 0.05.

Question: Calculate the odds of biking and interpret it.

Binary Data

A random sample of students is selected from a large statistics class.

We record the follow variable:

- the number of hours they spent studying
- Their exam grade, **pass (P)** or **fail (F)**

Hours	Grade
0	F
0	F
0.5	F
1.5	F
1.5	F
1.5	P
2	F
2.5	F
2.5	F
⋮	⋮
10.5	P
11	P
11	P

*The full dataset '**Hours-and-Grades**' can be downloaded from Brightspace*

Binary Data

Definition Binary Data

Binary data is a type of **categorical data** with exactly **two categories** - commonly labeled as **success** or **failure.**

Success represents the outcome of interest, and

failure represents the outcome not of interest.

- In our example, we record students' exam grades.
- Each exam grade has only two possible outcomes: **Pass** or **Fail**.
- Since the outcome of interest is whether a student **passes**, **Pass** is considered the **success**.
- Therefore, the **final grade data** is an example of **binary data**.

Modeling / Summarizing Binary Data

To model Binary data, all we need to know is

- the probability of **success** (or **failure**) or
- the odds of **success** (or fail)

So, to summarize Binary data, we can

- calculate the **Proportion** of students who **pass** (or **fail**) the exam and
- use it to estimate
  - the probability that a student **passes** or **fails** the exam
  - the **odds** of **passing** (or **failing**) to determine which **passing** (or **failing**) is more common on the exam.

I use the following R commands to:

- construct a frequency table for grades, and
- calculate the proportions of passing and failing.

```
#Assume the datafile (Hours-and-Grades.csv)
#has been saved on your computer.
#Use the file.choose() function to locate the file
fileLocation = file.choose()

#Read the data file using read.csv()
mydata = read.csv( fileLocation )

#Use the names(…) function to display the names of all variables
names( mydata )
# [1] "Hours" "Grade"

#Create a frequency table for the Grade and
#convert the frequency table to proportions using prop.table(…)
frequency.table = table( mydata$Grade )
grade.proportions = prop.table( frequency.table)
```

R Console

```
>
> #Create a frequency table for the Grade and
> #convert the frequency table to proportions using prop.table(…)
> frequency.table = table( mydata$Grade )
> grade.proportions = prop.table( frequency.table)
>
> grade.proportions

   F    P
0.34 0.66
```

Summary