ECON 3818

Chapter 6

Kyle Butts

26 September 2021

Chapter 6: Two-Way Tables

Introduction

This chapter discusses the relationship between two categorical variables

- To analyze categorical data, we use the *counts* or *percentages* of individuals that fall into various categories

Two-Way Table

Typically, published data is grouped in order to save space. This chapter will talk about **Two-Way Tables**

	ASSOCIATE	BACHELORS	MASTERS	DOCTORATE
WOMEN	673	1050	481	95
MEN	401	780	342	89

This two-way table describes two categorical variables. One is the sex of an individual, this is the **row variable** and the other is the degree attained, which is the **column variable**

Joint Distribution

The joint distribution is found by dividing each cell by the total count.

$$P(x, y) = \frac{\text{number of times x and y occurs}}{\text{total number of occurrences}}$$

Since we have 3911 people in total, we get the following joint probabilities:

	ASSOCIATE	BACHELORS	MASTERS	DOCTORATE
WOMEN	673	1,050	481	95
JOINT	17%	27%	12%	2%
MEN	401	780	342	89
JOINT	10%	20%	9%	2%

This means the probability of being a woman **and** having a Master's degree is 12%

Marginal Distribution

A **marginal distribution** is the probability distribution associated with only one of the random variables

In order to calculate, we need to look at the distribution of each variable *separately*. We do this by looking at the "Total" column and "Total" row.

We will have two different marginal distributions, the "row" marginal and the "column" marginal

Row Marginal

In this scenario, the row marginal is the distribution of sex alone:

	ASSOCIATE	BACHELORS	MASTERS	DOCTORATE	ROW MARGINAL
WOMEN	673	1,050	481	95	2,299
JOINT	17%	27%	12%	2%	58%
MEN	401	780	342	89	1,612
JOINT	10%	20%	9%	2%	42%

58.4% of individuals in this sample of degree holders are women.

Column Marginal

In this scenario, the column marginal is the distribution of degrees alone:

	ASSOCIATE	BACHELORS	MASTERS	DOCTORATE
WOMEN	673	1,050	481	95
JOINT	17%	27%	12%	2%
MEN	401	780	342	89
JOINT	10%	20%	9%	2%
COLUMN MARGINAL	0.273	0.47	0.21	0.047

The probability of an individual having a Bachelor's degree is 47%.

Conditional Distribution

We can use these tables to back out conditional distributions.

Remember:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

This is the same expression (just different notation) as:

$$P(x|y) = \frac{P(x, y)}{P(y)}$$

Conditional Distribution

For example, we can calculate the probability of holding each degree, given the individual is a woman:

$$P(\text{Associate's}|\text{Woman}) = \frac{P(\text{Associate's and Woman})}{P(\text{Woman})} = 0.17/0.584 = 29.1\%$$

$$P(\text{Bachelor's}|\text{Woman}) = \frac{P(\text{Bachelor's and Woman})}{P(\text{Woman})} = 0.27/0.584 = 46.2\%$$

$$P(\text{Master's}|\text{Woman}) = \frac{P(\text{Master's and Woman})}{P(\text{Woman})} = 0.12/0.584 = 20.5\%$$

$$P(\text{Doctorate}|\text{Woman}) = \frac{P(\text{Doctorate and Woman})}{P(\text{Woman})} = 0.024/0.584 = 4.1\%$$

Conditional Distribution

We could also calculate the probability an individual is a particular sex, based of holding an Associate's degree

$$P(\text{Male}|\text{Associate's}) = \frac{P(\text{Male and Associate's})}{P(\text{Associate's})} = .103/0.273 = 37.7\%$$

$$P(\text{Female}|\text{Associate's}) = \frac{P(\text{Female and Associate's})}{P(\text{Associate's})} = .17/.273 = 62.3\%$$

Joint Probabilities vs. Conditional Probabilities

- Joint probabilities take into account the probability that each event happens on its own
- Conditional probabilities assume that one event has already happened

Joint Probability vs. Conditional Probability{Example}

$$P(\text{work in tech job} \cap \text{live in Boulder}) \text{ vs. } P(\text{work in tech job} \mid \text{live in Boulder})$$

- $P(\text{work in tech})$ = work in tech/entire US population = relatively small, let's say 7%
- $P(\text{live in Boulder})$ = Boulder population/entire US population = also small, $< 1\%$

This means the probability of BOTH happening is small, because both events are unlikely compared to the state space of the entire US population

- But the P(work in tech $\mid$ live in Boulder) will be higher because now the state space is Boulder population, which has a greater concentration of high-tech employees

Marginal and Conditional Distributions

Here are some formal definitions:

The **marginal distribution** of one of the categorical variables in a two-way table of counts is the **distribution of that variable among all individuals described by the table**

- distribution of sex or degrees alone

A **conditional distribution** of a variable is the **distribution of values of that variable among only individuals who have a given value of the other variable**. There is a separate conditional distribution for each value of the other variable.

- there are two sets of conditional distributions for any two-way table, probability of having a degree based on sex, probability of being a particular sex based on degree held

Clicker Question

	FIRST	SECOND	THIRD	CREW
ALIVE	203	118	178	212
DEAD	122	167	528	673
TOTAL	325	285	706	885

Given this joint distribution, what is the probability of survival, given you are a first class passenger?

a. 9%

b. 62%

c. 1.3%

d. 30%

Source

Source

Simpson's Paradox

Consider the survival rates for the following groups of victims who were taken to the hospital, either by helicopter or by car:

	HELICOPTER	CAR
VICTIM DIED	64	260
VICTIM SURVIVED	136	840
TOTAL	200	1100

The probability of died conditional on helicopter, is higher than the probability of died conditional on car. Does this mean that this (more costly) mode of transportation isn't helping?

	HELICOPTER	CAR
VICTIM DIED	32%	68%
VICTIM SURVIVED	24%	76%

Simpson's Paradox

The idea is there a confounding variable, the severity of the accident, whose proportion differs between planes and car crashes

Serious Accidents

	HELICOPTER	CAR
VICTIM DIED	48	60
VICTIM SURVIVED	52	40
TOTAL	100	100

Less Serious Accidents

	HELICOPTER	CAR
VICTIM DIED	16	200
VICTIM SURVIVED	84	800
TOTAL	100	1000

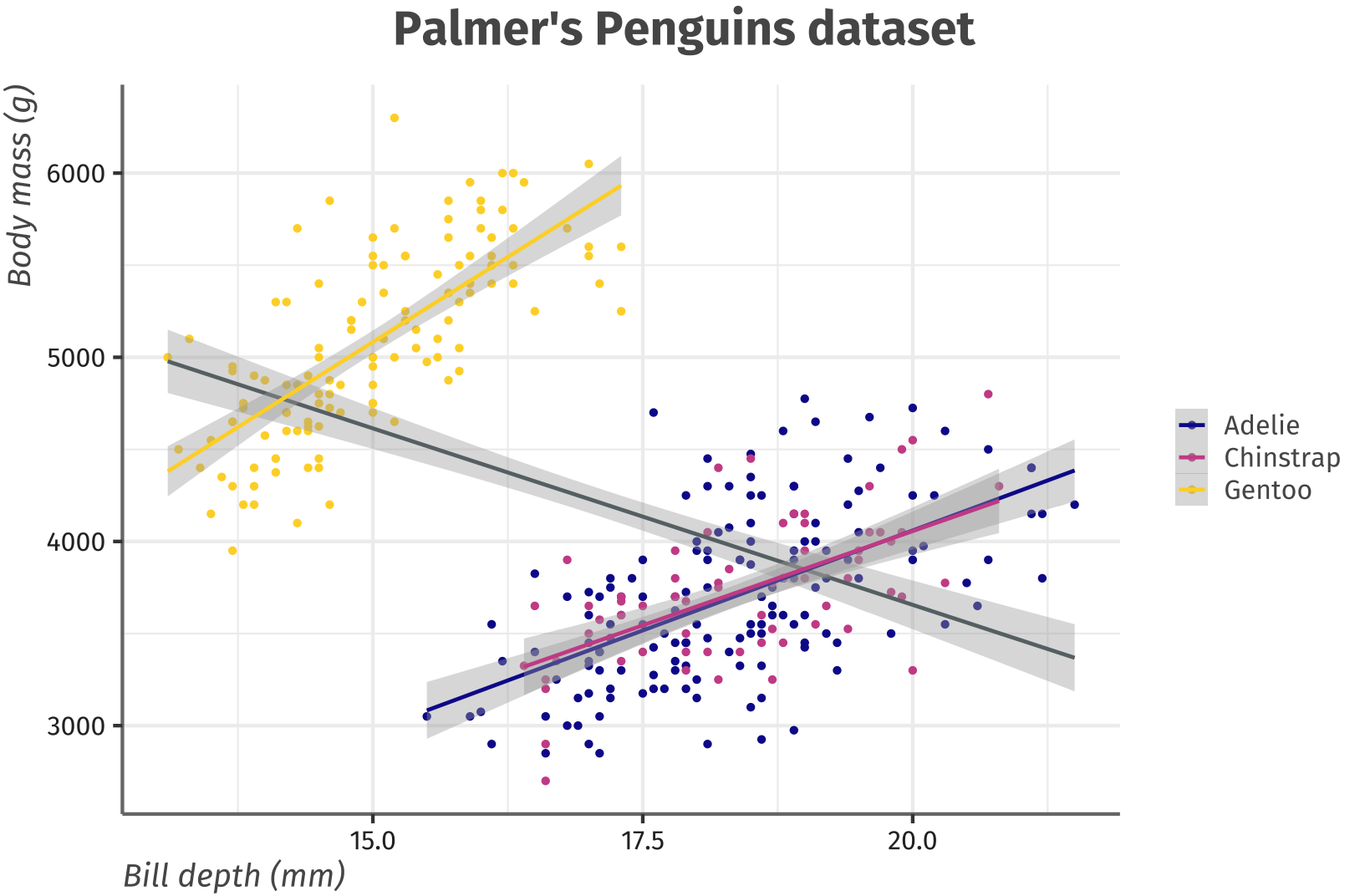
Serious Accidents

	HELICOPTER	CAR
VICTIM DIED	48%	52%
VICTIM SURVIVED	60%	40%

Less Serious Accidents

	HELICOPTER	CAR
VICTIM DIED	16%	84%
VICTIM SURVIVED	20%	80%

Simpson's Paradox



Clicker Question

From 2010-2013 the US median wage increased 1%, however over the same time period the median wage has decreased within each education subgroup (high school drop outs, high school graduates, some college, bachelor's or more).

Which of the following explanations is consistent with Simpson's paradox?

a. The BLS didn't control for inflation
b. There are more people with bachelor's degrees (high income people)
c. The wage of the highest income earner went up even more
d. There are less unemployed people now