ECON 3818: Introduction to Statistics with Computer Applications

---

Kyle Butts

November 16, 2020

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

Sex \ Degree	Associate	Bachelor's	Master's	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:

Sex \ Degree	Associate	Bachelor's	Master's	Doctorate
Women	673	1050	481	95
Joint %	**0.17**	**0.27**	**0.12**	**0.024**
Men	401	780	342	89
Joint %	**0.103**	**0.2**	**0.09**	**0.023**

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:

Sex \ Degree	Associate's	Bachelor's	Master's	Doctorate	Row Marginal
Women	673	1050	481	95	
Joint %	**0.17**	**0.27**	**0.12**	**0.024**	**0.584**
Men	401	780	342	89	
Joint %	**0.103**	**0.2**	**0.09**	**0.023**	**0.416**

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:

Sex \ Degree	Associate's	Bachelor's	Master's	Doctorate
Women	673	1050	481	95
Joint %	**0.17**	**0.27**	**0.12**	**0.024**
Men	401	780	342	89
Joint %	**0.103**	**0.2**	**0.09**	**0.023**
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 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:

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

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

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

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

Conditional Distribution

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

P(Male | Associate's) =
P(Male and Associate's)/P(Associate's)=.103/0.273 = 37.7%

P(Female | Associate's) =
P(Female and Associate's)/P(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

- P(work in tech job $\cap$ live in Boulder) vs.
  P(work in tech job | live in Boulder)
  - P(work in tech) = work in tech/entire US population = relatively small, let's say 7%
  - P(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 | 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

	Class of Person on Titanic			
Survival	First	Second	Third	Crew
Alive	203	118	178	212
Dead	122	167	528	673

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%

Simpson's Paradox

https://ed.ted.com/lessons/
how-statistics-can-be-misleading-mark-liddell

Simpson's Paradox

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

Counts	**Helicopter**	**Road**
Victim Died	64	260
Victim Survived	136	840
Total	200	1100

Percents	**Died**	**Survived**
Helicopter	32%	68%
Road	24%	76%

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

Simpson's Paradox

The idea is there a confounding variable, the severity of the accident, behind the Statistics

Serious Accidents

Counts	Helicopter	Road
Victim Died	48	60
Victim Survived	52	40
Total	100	100

Percents	Died	Survived
Helicopter	48%	52%
Road	60%	40%

Less Serious Accidents

Counts	Helicopter	Road
Victim Died	16	200
Victim Survived	84	800
Total	200	1100

Percents	Died	Survived
Helicopter	16%	84%
Road	20%	80%

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