Categorical Data, Simpson's Paradox

Evan L. Ray

January 26, 2018

I. Murder Cases in Indiana, 1977-1998 =(

Blume et {al.} (2004) assembled data about every murder case that went to trial and where the offender was found guilty in the state of Indiana between 1977 and 1998. For each court case, we have information about the race of the offender ("white" or "black"), the race of the victim ("white" or "black"), and the sentence that was handed down in that case ("jail" or "death"). During this time period in Indiana, there were not an appreciable number of murder cases involving offenders or victims of races other than white or black; any such cases have been dropped from this analysis.

The following R code reads this data in:

```r
library(readr)
library(dplyr)
murder_cases <-
  read_csv("http://www.evanlray.com/stat140_s2018/lecture/20180126_categorical/murder_cases.csv")
murder_cases <- mutate(murder_cases,
  offender_race = factor(offender_race),
  victim_race = factor(victim_race),
  sentence = factor(sentence, levels = c("jail", "death"), ordered = TRUE))
head(murder_cases)
```

```
## # A tibble: 6 x 4
##   case_id offender_race victim_race sentence
##     <int>        <fctr>      <fctr>    <ord>
## 1       1         white       white     jail
## 2       2         black       black     jail
## 3       3         black       black     jail
## 4       4         black       black     jail
## 5       5         white       black     jail
## 6       6         black       black     jail
```

```r
glimpse(murder_cases)
```

```
## Observations: 4,898
## Variables: 4
## $ case_id       <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 1...
## $ offender_race <fctr> white, black, black, black, white, black, black...
## $ victim_race   <fctr> white, black, black, black, black, black, black...
## $ sentence      <ord> jail, jail, jail, jail, jail, jail, jail, jail, ...
```

```r
dim(murder_cases)
```

```
## [1] 4898    4
```

Warm Up:

(a) What are the observational units in this data set? How many observational units are there?

(b) What are the variables? Is each variable an identifier variable, a categorical variable, or a quantitative variable? Are the categorical variables nominal or ordinal?

II. Relationship between `sentence` and `offender_race`

Here are the same data, summarized in a contingency table.

```
library(mosaic)
tally(sentence ~ offender_race, data = murder_cases, margins = TRUE)
```

```
##         offender_race
## sentence black white
##    jail   2498  2323
##    death    28    49
##    Total  2526  2372
```

This code used the `tally` function, which is in the `mosaic` package. The `tally` function calculates the number of cases that had each combination of the possible values for `sentence` and `offender_race`, where those variables are found in the data frame named `murder_cases`. The general format of tally is like this:

```
tally(<response variable name> ~ <explanatory variable name>, data = <data frame name>,
  margins = TRUE)
```

In this example, our working hypothesis is that the `offender_race` for a case might tell us some information about (or **explain** something about) the `sentence` for that case. That's why I have used `sentence` for the **response** variable and `offender_race` for the **explanatory** variable. In the output from this command, which variable is used as the explanatory variable and which is used as the response doesn't really matter – but the format of putting response variables first and explanatory variables second is consistent across many R functions, so it's good to get used to the idea of it now. The `margins = TRUE` argument tells R to calculate the row and column totals.

There are a few types of questions we might want to answer based on these numbers:

(a) What proportion of the data fall in each combination of levels of the `offender_race` and `sentence` variables?

This is the **joint distribution** of the offender's race and the sentence.

(b) Looking at just the `sentence` variable (aggregating across all offender races), what proportion of the observational units fall in each level of that variable?

This is the **marginal distribution** of the sentence.

(c) Among those cases where the offender's race was white, what proportion of the observational units fall in each level of the `sentence` variable?

This is the **conditional distribution** of the sentence given that the offender's race was white.

(d) We can ask the same question again for cases where the offender's race was black: Among those cases where the offender's race was black, what proportion of the observational units fall in each level of the `sentence` variable?

This is the **conditional distribution** of the sentence given that the offender's race was black.

(e) Is the conditional distribution of the `sentence` the same for white offenders as it is for black offenders? If not, do offenders of one race appear to be sentenced to death more often than the other?

If the conditional distribution of `sentence` is the same for all values of the `offender_race`, we say that those two variables are **independent**.

III. Looking a little deeper

We've just examined the connection between the offender's race and the sentence in some detail – but the data set also included another variable, the victim's race. In groups of about 4, let's break these results down by the victim's race as well. Within each group, one pair will work through the calculations using just the cases where the victim's race was white, and another pair will work through these calculations using just the cases where the victim's race was black. Then you will share your results with each other and see what the data have to say.

1. Victim's race is white

We can use the `filter` function to select just those cases where the victim's race was white. This command creates a new data frame called `murder_cases_victim_white` with just those cases. We then use the `tally` function to look at the break down of `sentence` and `offender_race` among just those cases with white victims.

```
murder_cases_victim_white <- filter(murder_cases, victim_race == "white")
tally(sentence ~ offender_race, data = murder_cases_victim_white, margins = TRUE)
```

```
##          offender_race
## sentence black white
##     jail   359  2223
##     death   16    49
##     Total  375  2272
```

(a) What the joint distribution of the offender's race and the sentence, among those cases with white victims?

(b) What is the marginal distribution of the sentence, among those cases with white victims?

(c) What is the conditional distribution of the sentence, given that the offender's race was white and the victim's race was white?

(d) What is the conditional distribution of the sentence, given that the offender's race was black and the victim's race was white?

(e) In cases where the victim's race was white, is the `sentence` independent of the offender's race? If not, do offenders of one race appear to be sentenced to death more often than the other?

2. Victim's race is black

We can use the `filter` function to select just those cases where the victim's race was black. This command creates a new data frame called `murder_cases_victim_black` with just those cases. We then use the `tally` function to look at the break down of `sentence` and `offender_race` among just those cases with black victims.

```
murder_cases_victim_black <- filter(murder_cases, victim_race == "black")
tally(sentence ~ offender_race, data = murder_cases_victim_black, margins = TRUE)
```

```
##         offender_race
## sentence black white
##    jail   2139   100
##    death    12     0
##    Total  2151   100
```

(a) What the joint distribution of the offender's race and the sentence, among those cases with black victims?

(b) What is the marginal distribution of the sentence, among those cases with black victims?

(c) What is the conditional distribution of the sentence, given that the offender's race was white and the victim's race was black?

(d) What is the conditional distribution of the sentence, given that the offender's race was black and the victim's race was black?

(e) In cases where the victim's race was black, is the `sentence` independent of the offender's race? If not, do offenders of one race appear to be sentenced to death more often than the other?

3. Tying it all together

(a) The effects of breaking results down by the victim's race.

Within your group of 4, compare your answers to parts II. (e) (where we looked at the relationship between sentence and offender's race, across all cases), III. 1 (e) (where we looked at the relationship between sentence and offender's race among just those cases where the victim was white), and III. 2 (e) (where we looked at the relationship between sentence and offender's race among just those cases where the victim was black).

In each of those three scenarios, were white or black offenders more likely to receive a death sentence? Does this relationship stay the same or change when we break the results down by the victim's race?

(b) Can you figure out what's going on? A description of the answer is on the next page, but see if you can figure it out before you look! All of the information you need is in the tables and your calculations above.

What's going on?

You should have found that in the overall data, aggregating across all values of victim's race, white offenders were more likely to be sentenced to death – but looking at just cases where the victim was black, or just cases where the victim was white, black offenders were more likely to be sentened to death. This surprising finding happens because of two facts put together:

- A death sentence is more likely when the victim is white than when the victim is black, no matter the race of the offender (you can check your answers to parts (b) above to verify this).
- More black offenders are involved in cases where the victim was black, and more white offenders are involved in cases where the victim was white.

Putting those facts together, we see that black offenders are more likely to be involved in cases with black victims, and therefore are sentenced to death less often that white offenders overall – even though black offenders are more likely to be sentenced to death in cases with white victims and are more likely to be sentenced to death in cases with black victims.

Summary

There are a few things I want you to get out of this example:

1. The definitions of **joint distributions**, **marginal distributions**, and **conditional distributions**, and how these distributions are calculated.
2. The definition of **independence**, and how independence of two variables can be verified. We will return to this in more detail in a few weeks.
3. The idea that the relationships you observe in data can change when you break the data down by additional variables. This is called **Simpson's Paradox**.

References

The original data were published in:

Blume et al., Explaining Death Row's Population and Racial Composition. Journal of Empirical Legal Studies, Vol. 1, Issue 1, p. 165-207, 2004.

The issue of Simpson's paradox in relation to these data was discussed further in:

Norton et al., Simpson's Paradox and How to Avoid It. Significance, p. 40-43, August, 2015.