Categorical Data, Simpson's Paradox

A major concern in modern health care is that many patients who go to the hospital for treatment for one condition become infected with another disease while there. A study conducted in Norway investigated whether prescribing antibiotics as soon as a patient entered the hospital could reduce the chances that a patient would develop a urinary tract infections (UTI) (Reintjes, 2000). They recruited patients from 8 hospitals in Norway and assigned each to either take preventative antibiotics or not. They then recorded whether or not each patient developed a UTI. Also recorded in the full data set is whether or not the patient was in a hospital with high incidence of UTI (depending on whether the UTI rate is less than or greater than 2.5%).

The following R code reads these data in:

```r
library(readr)
library(dplyr)
library(tidyr)

uti_prevention <- read_csv(
  "http://www.evanlray.com/data/norton_simpsons_paradox/uti_infections.csv")

uti_prevention <- uti_prevention %>%
  mutate(
    hospital_class = factor(hospital_class,
      levels = c("low incidence", "high incidence"),
      ordered = TRUE),
    antibiotics_used = factor(antibiotics_used,
      levels = c("no", "yes"),
      ordered = TRUE),
    uti = factor(uti, levels = c("no", "yes"), ordered = TRUE))
```

It's always good to take a quick look at the data with the `head`, `str`, and `dim` functions:

```r
str(uti_prevention)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    3519 obs. of  4 variables:
##  $ patient_id      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ hospital_class  : Ord.factor w/ 2 levels "low incidence"<..: 1 1 1 1 2
##  $ antibiotics_used: Ord.factor w/ 2 levels "no"<"yes": 2 1 2 2 1 2 1 2 2
##  $ uti             : Ord.factor w/ 2 levels "no"<"yes": 1 1 1 1 2 1 1 1 1
```

```
head(uti_prevention)
```

```
## # A tibble: 6 x 4
##   patient_id hospital_class antibiotics_used uti
##        <int> <ord>          <ord>            <ord>
## 1          1 low incidence  yes              no
## 2          2 low incidence  no               no
## 3          3 low incidence  yes              no
## 4          4 low incidence  yes              no
## 5          5 high incidence no               yes
## 6          6 low incidence  yes              no
```

```
dim(uti_prevention)
```

```
## [1] 3519    4
```

I. 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 antibiotics_used and uti

Here are the same data, summarized by counting the number of patients in each combination of levels of the antibiotics_used and uti variables.

```
uti_prevention %>%
  count(antibiotics_used, uti)
```

```
## # A tibble: 4 x 3
##   antibiotics_used uti       n
##   <ord>            <ord> <int>
## 1 no               no     2136
## 2 no               yes     104
## 3 yes              no     1237
## 4 yes              yes      42
```

It can be helpful to put this in the format of a contingency table:

```
uti_prevention %>%
  count(antibiotics_used, uti) %>%
  spread(uti, n)
```

```
## # A tibble: 2 x 3
##   antibiotics_used    no   yes
##   <ord>            <int> <int>
## 1 no                2136   104
## 2 yes               1237    42
```

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 antibiotics_used and uti variables?

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

(b) What proportion of the observational units fall into each level of the `uti` variable (aggregating across all values of `antibiotics_used`)?

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

(c) Among those cases where the patient took preventative antibiotics, what proportion of the observational units fall in each level of the `uti` variable?

This is the **conditional distribution** of `uti` given that the patient was taking antibiotics.

(d) Among those cases where the patient didn't take preventative antibiotics, what proportion of the observational units fall in each level of the `uti` variable?

This is the **conditional distribution** of `uti` given that the patient was not taking antibiotics.

(e) Is the conditional distribution of `uti` the same for patients taking antibiotics as it is for patients not taking antibiotics?

We say that those two variables are **independent** if the conditional distribution of `uti` is the same for all values of the `antibiotics_used` variable.

III. Looking a little deeper

We've just examined the connection between using antibiotics and developing UTI's in some detail – but the data set also included another variable, an indicator of whether there was overall low or high prevalence of UTIs at the hospital where the patient was treated. In groups of about 4, let's break these results down by the groups of hospitals. Within each group, one pair will work through the calculations using just the low-incidence hospitals, and another pair will work through these calculations using just the high-incidence hospitals. Then you will share your results with each other and see what the data have to say.

1. Low-incidence hospitals

We can use the `filter` function to select just those cases where the patient was in a hospital that had low incidence of UTIs. This command creates a new data frame called `low_incidence` with just those cases. We then use `count` and `spread` functions to look at the break down of `antibiotics_used` and `uti` among just those patients who were treated in a low-incidence hospital.

```
low_incidence_cases <- uti_prevention %>%
  filter(hospital_class == "low incidence")

low_incidence_cases %>%
  count(antibiotics_used, uti) %>%
  spread(uti, n)
```

```
## # A tibble: 2 x 3
##   antibiotics_used    no   yes
##   <ord>            <int> <int>
## 1 no                 715     5
## 2 yes               1093    20
```

(a) What is the joint distribution of `antibiotics_used` and `uti`, among patients treated in low-incidence hospitals?

(b) What is the marginal distribution of `uti`, among patients treated in low-incidence hospitals?

Note that this could also be framed as the conditional distribution of `uti` given that the patient was treated in a low-incidence hospital.

(c) What is the conditional distribution of `uti`, given that the patient was treated in a low incidence hospital and they were taking antibiotics?

(d) What is the conditional distribution of `uti`, given that the patient was treated in a low incidence hospital and they were not taking antibiotics?

(e) In cases where the patient was treated in a low incidence hospital, is whether a patient develops a UTI independent of whether they are taking antibiotics? If not, does taking antibiotics seem to be helpful or harmful?

2. High-incidence hospitals

We can use the `filter` function to select just those cases where the patient was in a hospital that had high incidence of UTIs. This command creates a new data frame called `high_incidence` with just those cases. We then use `count` and `spread` functions to look at the break down of `antibiotics_used` and `uti` among just those patients who were treated in a high-incidence hospital.

```
high_incidence_cases <- uti_prevention %>%
  filter(hospital_class == "high incidence")

high_incidence_cases %>%
  count(antibiotics_used, uti) %>%
  spread(uti, n)
```

```
## # A tibble: 2 x 3
##   antibiotics_used    no   yes
##   <ord>            <int> <int>
## 1 no                1421    99
## 2 yes                144    22
```

(a) What is the joint distribution of `antibiotics_used` and `uti`, among patients treated in high-incidence hospitals?

(b) What is the marginal distribution of `uti`, among patients treated in high-incidence hospitals?

Note that this could also be framed as the conditional distribution of `uti` given that the patient was treated in a high-incidence hospital.

(c) What is the conditional distribution of `uti`, given that the patient was treated in a high incidence hospital and they were taking antibiotics?

(d) What is the conditional distribution of `uti`, given that the patient was treated in a high incidence hospital and they were not taking antibiotics?

(e) In cases where the patient was treated in a high incidence hospital, is whether a patient develops a UTI independent of whether they are taking antibiotics? If not, does taking antibiotics seem to be helpful or harmful?

3. Tying it all together

(a) The effects of breaking results down by hospital type.

Within your group of 4, compare your answers to:

- part II. (e) (where we looked at the relationship between antibiotics use and whether a patient develops a UTI, across all patient)
- part III. 1 (e) (where we looked at the relationship between antibiotics use and whether a patient develops a UTI, among just those patients treated in a low-incidence hospital)
- part III. 2 (e) (where we looked at the relationship between antibiotics use and whether a patient develops a UTI, among just those patients treated in a high-incidence hospital).

In each of those three scenarios, were patients who took antibiotics more or less likely to develop a UTI than patients who did not take antibiotics?

Does this relationship stay the same or change when we break the results down by the hospital class?

(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?

In part II, we found that in the overall data, aggregating across all hospitals, patients who were prescribed antibiotics as soon as they entered the hospitals were less likely to develop a UTI. However, in part III you found that looking just at low-incidence hospitals, patients who were prescribed antibiotics as soon as they entered the hospital were more likely to develop a UTI, and similar for patients at high incidence hospitals.

This happens because of two things put together:

- Overall, a much higher proportion of patients developed UTIs at the high incidence hospitals than at the low incidence hospitals
- In this study, most of the patients who were prescribed antibiotics were being treated at low incidence hospitals

In the aggregated data, it looked like antibiotics were helpful - but that just showed up in the data because antibiotics were prescribed more often in low incidence hospitals. When we look at the results within low incidence hospitals and within high incidence hospitals, we can see that in reality, use of antibiotics immediately after entering the hospital is associated with higher chances of developing a UTI.

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:

Reintjes, R., de Boer, A., van Pelt, W. and Mintjes-de Groot, J. (2000) Simpson's paradox: An example from hospital epidemiology. Epidemiology, 11(1), 81–83.

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

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