Contingency Tables

Ryan Miller

Introduction

- *Univariate* summaries are the first step in a statistical analysis, but most analyses involve establishing relationships between *multiple variables*
  - These slides focus on methods for expressing relationships between *two categorical variables*

Association

Two variables, X and Y, are **associated** if the distribution of X depends upon the distribution of Y

- Usually, we designate an **explanatory variable** (suspected cause) and a **response variable** (suspected outcome)
  - This is done using prior knowledge (ie: Exam #1 score could cause final grade, but not vice versa)

Association

Two variables, X and Y, are **associated** if the distribution of X depends upon the distribution of Y

- Usually, we designate an **explanatory variable** (suspected cause) and a **response variable** (suspected outcome)
  - This is done using prior knowledge (ie: Exam #1 score could cause final grade, but not vice versa)

Note:

1. Association is general term, we'll soon cover specific types of association (ie: linear, non-linear, etc.)

Association

Two variables, X and Y, are **associated** if the distribution of X depends upon the distribution of Y

- Usually, we designate an **explanatory variable** (suspected cause) and a **response variable** (suspected outcome)
  - This is done using prior knowledge (ie: Exam #1 score could cause final grade, but not vice versa)

Note:

1. Association is general term, we'll soon cover specific types of association (ie: linear, non-linear, etc.)
2. Observing an association between X and Y doesn't mean that X causes Y, or that Y causes X, *causation* is a complex topic that we'll discuss soon

Contingency Tables

- For *two categorical variables*, we can display frequencies for *each combination* of the variables in a **contingency table** (also called a two-way frequency table)
- Below is a two-way frequency table describing the historic 2015-16 Golden State Warriors season:

	Win	Loss
Home	39	2
Away	34	7

Practice

What do you think the raw data that was used to construct this table looks like? Try writing out a few rows.

	Win	Loss
Home	39	2
Away	34	7

Practice (solution)

Recognize you're only able to discern the last two columns from the contingency table on the prior slide

Date	Opp	Location	Win
10/27/2015	NOP	Home	W
10/30/2015	HOU	Away	W
10/31/2015	NOP	Away	W
11/2/2015	MEM	Home	W
11/4/2015	LAC	Home	W
11/6/2015	DEN	Home	W
11/7/2015	SAC	Away	W
11/9/2015	DET	Home	W
11/11/2015	MEM	Away	W
11/12/2015	MIN	Away	W
11/14/2015	BRK	Home	W
11/17/2015	TOR	Home	W
11/19/2015	LAC	Away	W
11/20/2015	CHI	Home	W
11/22/2015	DEN	Away	W
11/24/2015	LAL	Home	W
11/27/2015	PHO	Away	W
11/28/2015	SAC	Home	W
11/30/2015	UTA	Away	W
12/2/2015	CHO	Away	W
12/5/2015	TOR	Away	W
12/6/2015	BRK	Away	W
12/8/2015	IND	Away	W
12/11/2015	BOS	Away	W
12/12/2015	MIL	Away	L
12/16/2015	PHO	Home	W
12/18/2015	MIL	Home	W

Margins

A useful step when working with contingency tables is to add *table margins*:

	Win	Loss	Row Total
Home	39	2	41
Away	34	7	41
Column Total	73	9	82

Margins

A useful step when working with contingency tables is to add *table margins*:

	Win	Loss	Row Total
Home	39	2	41
Away	34	7	41
Column Total	73	9	82

- Row and column totals are sometimes called **marginal distributions**
  - The marginal distribution of the "win" variable (win/loss) is characterized by the frequencies $\{73, 9\}$ and the proportions $\{0.89, 0.11\}$
  - The marginal distribution of the "location" variable (home/away) is characterized by the frequencies $\{41, 41\}$ and the proportions $\{0.5, 0.5\}$

Conditional Proportions

- From a contingency table, **conditional proportions** allow us to determine whether the two variables displayed are *associated*
- There are two types of conditional proportions: **row proportions** are calculated using each row's total, the bottom table show how to calculate these

	Win	Loss	Row Total
Home	39	2	41
Away	34	7	41
Column Total	73	9	82

	Win	Loss	Row Total
Home	$39/41 = 0.95$	$2/41 = 0.05$	1
Away	$34/41 = 0.83$	$7/41 = 0.17$	1
Column Total	$73/82 = 0.89$	$9/82 = 0.11$	1

Conditional Proportions

Column proportions are calculated in a similar way:

	Win	Loss	Row Total
Home	39	2	41
Away	34	7	41
Column Total	73	9	82

	Win	Loss	Row Total
Home	$39/73 = 0.53$	$2/9 = 0.22$	$41/82 = 0.5$
Away	$34/73 = 0.47$	$7/41 = 0.78$	$41/82 = 0.5$
Column Total	1	1	1

Conditional Distributions and Association

- Two variables are **associated** if the distribution of one variable depends upon that of the other variable
- So, we might compare the distribution of win/loss proportions *conditional upon a game being at home* with the distribution of win/loss proportions *conditional upon a game being away*
  - If these distributions differ, the variables "location" and "win" are associated

Practice #1

1. Using the row proportions given below, do you think there is an association between whether the Warriors were home/away and winning?
2. How would you explain this association?

	Win	Loss	Row Total
Home	0.95	0.05	1
Away	0.83	0.17	1
Column Total	0.89	0.11	1

Practice #1 (solution)

1. Yes, there is an association between "location" and "win"
2. The warriors look to be *more likely* to win when playing at home. In other words, the distribution of wins/losses for home games differs from the distribution of wins/losses for away games.

Remarks

- Recognize that row and column proportions tell you fundamentally different things about your data
  - In our example, *row proportions* can describe the *proportion of wins conditional on the game being at home*
  - Contrast that with *column proportions*, which can describe the *proportion of home games conditional on that game being a win*

Remarks

- Recognize that row and column proportions tell you fundamentally different things about your data
  - In our example, *row proportions* can describe the *proportion of wins conditional on the game being at home*
  - Contrast that with *column proportions*, which can describe the *proportion of home games conditional on that game being a win*
- The row proportions suggest how often home games were won, while the column proportions suggest how often wins were home games
  - This distinction doesn't seem to matter much here, but let's look at another example

Practice #2

Were crew members on the Titanic more likely to survive than 1st class passengers?

1) Download the “Titanic” dataset from this link or our course website.
2) Upload the data into the “Two Categorical Variables” section of StatKey.
3) Analyze the contingency table (comparing 1st class and crew).

Practice #2 (solution)

- No, using *row proportions* we see that $\frac{212}{623+212} = 0.24$, or 24% of the crew survived; while $\frac{203}{122+203} = 0.62$, or 62% of first class passengers survived

	Survived	Died
Crew	212	673
1st Class	203	122

Practice #2 (solution)

- No, using *row proportions* we see that $\frac{212}{623+212} = 0.24$, or 24% of the crew survived; while $\frac{203}{122+203} = 0.62$, or 62% of first class passengers survived

	Survived	Died
Crew	212	673
1st Class	203	122

- Notice that this particular question *cannot be answered* using column proportions
  - The proportion of survivors who were crew is $\frac{212}{212+203} = 0.51$, while the proportion of survivors who were first class passengers is $\frac{203}{212+203} = 0.49$

Practice #2 (solution)

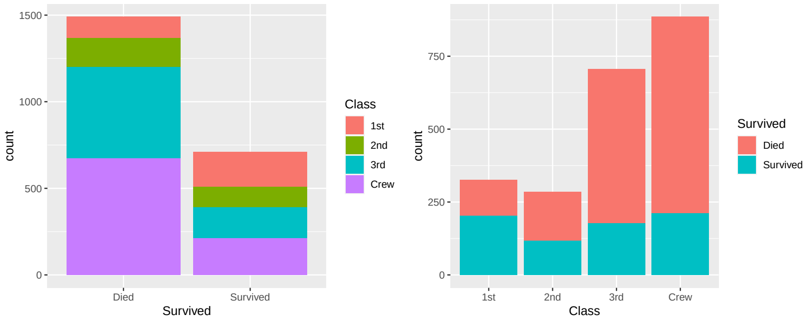
- No, using *row proportions* we see that $\frac{212}{623+212} = 0.24$, or 24% of the crew survived; while $\frac{203}{122+203} = 0.62$, or 62% of first class passengers survived

	Survived	Died
Crew	212	673
1st Class	203	122

- Notice that this particular question *cannot be answered* using column proportions
  - The proportion of survivors who were crew is $\frac{212}{212+203} = 0.51$, while the proportion of survivors who were first class passengers is $\frac{203}{212+203} = 0.49$
  - Conditioning on the column variable is problematic here because the *marginal distribution* of 1st class/crew is *skewed towards crew*
  - In other words, most of the survivors were crew members because there were so many more crew members, not because individual crew members were more likely to survive

Data visualizations

Finally, it's important to recognize that we can use barcharts to graph the information contained in a contingency table:



Both of the above graphs convey the same information, but which do you find more effective?

Conclusions

1) Contingency tables display the possible combinations of two categorical variables
2) Row proportions or column proportions within a contingency are used to find and describe associations
3) Just because an association exists does not mean that one variable caused changes in the other