ISTA 116: Statistical Foundations for the Information Age

Multivariate Categorical Data

14 September 2011

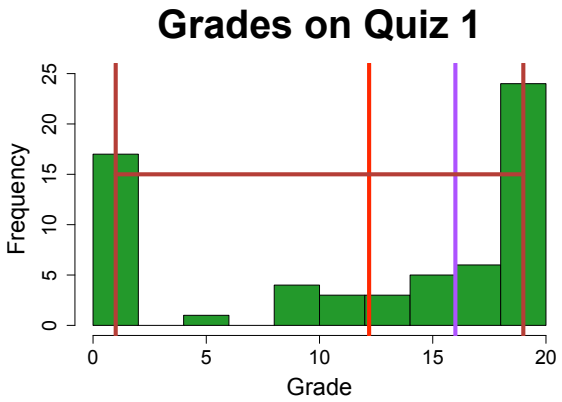
Outline

Reminders/Announcements

- Web Quiz 3 (up now) due Friday. Download the quiz from Content, and then submit your answers via Quizzes.
  - Reminder: You should not be using R for web quizzes
  - Please make your first attempt without using reference material
- Web Quiz 4 (up soon) due next Wednesday
- Lab 3 (up soon) due a week from Friday
- Please keep track of the Announcements page on d2l; we will sometimes tell you about useful references there.

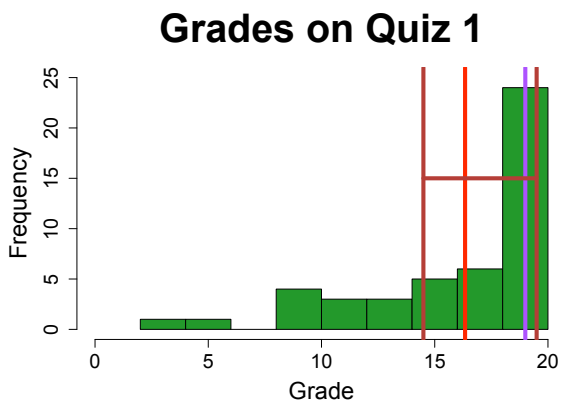
Quiz 1

- What can we say about this distribution?



Quiz 1

- The distribution for those who actually took the quiz:



Some Latin

- Recall: if a data set has one variable, we say that it is **univariate**
- Two variables: **bivariate**
- More than two: **multivariate**
- Verzani covers bivariate data in Ch. 3 and multivariate in Ch. 4. We'll cover them together, since the types of the variables makes a bigger difference in what you can do than the number.

Multiple Univariate vs. One Multivariate

- What's the difference between this...

Person	Sex
1	M
2	F
3	F
4	M
5	F
6	F
7	M
8	M

Person	Height (in.)
A	64
B	74
C	72
D	68
E	61
F	70
G	68
H	69

Multiple Univariate vs. One Multivariate

- and this?

Sex	Height (in.)
M	74
F	64
F	61
M	68
F	70
F	69
M	72
M	68

Multiple Univariate vs. One Multivariate

- When we observe more than one characteristic from the *same person*, we can look at the *relationship* between the variables.
  - Are males taller than females on average?
  - Are males more variable than females?
- With two isolated variables, and no correspondence between the values, there's no way to examine relationships.

Three Cases

- The kinds of relationships we can identify depend on the types of variables we have
- Three Cases:
  - All categorical variables
  - A mix of categorical and numeric
  - All numeric

Contingency Tables

- Recall: with a univariate categorical data set, we summarized by *counting* the observations in each category
- With more than one variable, we do the same thing, but we keep track of *combinations*.
- With two variables, we can store the counts in a two-dimensional grid called a **contingency table**

A Simple Contingency Table

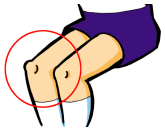
Sex	Computer
M	PC
F	Mac
F	PC
M	PC
F	PC
F	Mac
M	Mac
M	PC

$\Longrightarrow$

		Computer	
		PC	**Mac**
Sex	**M**	3	1
	F	2	2

Joint Frequencies

- With one categorical variable, we had a **frequency table**: the entries represent how many times (how *frequently*) each category appears in the data
- With two or more variables, we can count how often two categories (say, male and PC) appear together (jointly).
- The counts of the combinations are called **joint frequencies**.



Marginal Frequencies

- We may be interested in how often one category appears, *regardless* of what it's combined with.
- To find the total number of females (say), sum across the row.
- We could write the sum in the "margin" of the contingency table.
- This sum is called the **marginal frequency** for females (say).

Marginal Frequencies

- We start with joint frequencies:

		Computer	
		PC	**Mac**
Sex	**M**	3	1
	F	2	2

Marginal Frequencies

- and compute the marginal frequencies for sex:

		Computer		
		PC	**Mac**	**Marginal**
Sex	**M**	3	1	4
	F	2	2	4

Marginal Frequencies

- We can also compute the marginal frequencies for Computer:

		Computer		
		PC	**Mac**	**Marginal**
Sex	**M**	3	1	4
	F	2	2	4
	Marginal	5	3	

- Notice that we can start with a set of marginal frequencies, and "marginalize" in the other direction to get the total number of observations $(= n)$

Marginal Frequencies

- Call our two variables X and Y, and let x and y represent particular values they can take on.
- If we use Freq$(\cdot)$ to represent the frequency of whatever we put inside the parentheses, then we have:

$$\begin{aligned}
\text{Freq}(X = x) &= \sum_y \text{Freq}(X = x \text{ and } Y = y) \\
\text{Freq}(Y = y) &= \sum_x \text{Freq}(X = x \text{ and } Y = y) \\
\sum_x Freq(X = x) &= \sum_y Freq(Y = y) = n
\end{aligned}$$

Marginal Frequencies

- For example:

$$\begin{aligned}
&\text{Freq}(Sex = M) \\
&= \textstyle\sum_y \text{Freq}(Sex = M \text{ and } Computer = y)
\end{aligned}$$

$$\begin{aligned}
&\text{Freq}(Computer = PC) \\
&= \textstyle\sum_x \text{Freq}(Sex = x \text{ and } Computer = PC)
\end{aligned}$$

Joint and Marginal Proportions

- With univariate categorical data, we computed **relative frequencies** for each category: the *proportion* of the time that category appeared.
- We can do the same thing for the frequencies in a contingency table, simply by dividing each frequency by the grand total.
- When applied to joint frequencies, we get **joint proportions**; when applied to marginal frequencies, we get **marginal proportions**

Joint and Marginal Proportions

- Starting with the frequencies...

		Computer PC	Mac	Marginal
Sex	**M**	3	1	4
	F	2	2	4
	Marginal	5	3	$n = 8$

Joint and Marginal Proportions

- Divide each frequency by n (here $n = 8$):

		Computer PC	Mac	Marginal
Sex	**M**	$3/n$	$1/n$	$4/n$
	F	$2/n$	$2/n$	$4/n$
	Marginal	$5/n$	$3/n$	n/n

Joint and Marginal Proportions

- The result is a table of joint and marginal proportions.

		Computer PC	Mac	Marginal
Sex	**M**	0.375	0.125	0.500
	F	0.250	0.250	0.500
	Marginal	0.625	0.375	1.000

- Notice that the joints still sum to the marginals, and all the joints, as well as any particular set of marginals, sum to 1.000.

Joint and Marginal Proportions

- If we define $\text{Prop}(\cdot)$ to be the proportion of whatever is inside, then using our previous notation:

$$\text{Prop}(\cdot) = \frac{1}{n}\text{Freq}(\cdot)$$

Joint and Marginal Proportions

- In particular:

$$\begin{aligned}
\text{Prop}(X = x, Y = y) &= \frac{1}{n}\text{Freq}(X = x, Y = y) \\
\text{Prop}(X = x) &= \frac{1}{n}\text{Freq}(X = x) \\
&= \frac{1}{n}\sum_y \text{Freq}(X = x, Y = y) \\
&= \sum_y \frac{1}{n}\text{Freq}(X = x, Y = y) \\
&= \sum_y \text{Prop}(X = x, Y = y)
\end{aligned}$$

Joint and Marginal Distributions

- If we take the set of all the joint proportions, we define a **joint distribution** over the variables.
- Similarly, each set of marginal proportions defines a **marginal distribution** for that variable.

Joint and Marginal Distributions

- Each color represents a different distribution.
  - Lavender: Joint distribution of Sex and Computer
  - Pink: Marginal distribution of Sex
  - Lime Green: Marginal distribution of Computer

		Computer		
		PC	**Mac**	**Marginal**
Sex	**M**	0.375	0.125	0.500
	F	0.250	0.250	0.500
	Marginal	0.625	0.375	1.000

- Notice that each distribution sums to 1.

Conditional Proportions

- Both joint and marginal proportions represent proportions of the *whole data set*.
- Sometimes, however, we want to ask about proportions within a *subset* of the data
  - Example: What proportion of males use a PC?
  - What proportion of Mac users are female?
- We can use these values to make interesting comparisons
  - Are women more likely than men to buy a Mac?
  - Are urban residents more likely to get lung cancer than rural or suburban residents?
  - Other examples?

Conditional Proportions

- Quantities like "what proportion *of males* use a PC?" are called **conditional proportions**: we "condition" on being male (i.e., restrict attention to males) before calculating the proportion.
- We write the above as $\text{Prop}_{Sex=M}(Computer = PC)$, which we can read as "the conditional proportion of PC users, *given* that Sex = M" (later we will connect this to probability: "What's the probability someone will buy a PC, *given* that they are male?").

Conditioning on Sex

- Again, we start with the frequencies...

		Computer PC	Mac	Marginal
Sex	**M**	3	1	4
	F	2	2	4
	Marginal	5	3	8

Conditioning on Sex

- But instead of dividing by n, we divide by the joints by the corresponding marginal for Sex.

		Computer PC	Mac	Marginal
Sex	**M**	3/4	1/4	4/4
	F	2/4	2/4	4/4
	Marginal	5/8	3/8	8/8

- The lime green does not represent conditional proportions: it is the same marginal (unconditional) distribution from before.

Conditioning on Sex

- The resulting conditional proportions make up *two different* conditional distributions: one for each sex.

		Computer PC	Mac	Marginal
Sex	**M**	0.75	0.25	1.00
	F	0.50	0.50	1.00
	Marginal	0.625	0.375	1.00

- Notice that, as before, *each distribution* sums to one (but now each row is its own distribution: we consider males and females separately).

Conditioning on Computer

- Suppose instead we want to know "what proportion of Mac users are female?". Now we are conditioning on Computer: restrict attention to Mac users.
- Begin with frequencies as before, but group based on Computer:

		Computer PC	Mac	Marginal
Sex	**M**	3	1	4
	F	2	2	4
	Marginal	5	3	8

Conditioning on Computer

- Divide each frequency by the total *for that computer*:

		Computer PC	Mac	Marginal
Sex	**M**	3/5	1/3	4/8
	F	2/5	2/3	4/8
	Marginal	5/5	3/3	8/8

Conditioning on Computer

- Again we get two different conditional distributions: one for each computer.

		Computer PC	Mac	Marginal
Sex	**M**	0.60	0.33	0.50
	F	0.40	0.67	0.50
	Marginal	1.00	1.00	1.00

- Notice that the rows do *not* form distributions: the marginal distribution of sex is a *weighted average* of the conditional distributions (what are the weights?)

Conditional Proportions

- We calculate conditional proportions from frequencies in the obvious way: restrict attention to ("condition on") a particular category, and divide the frequency by the total in that category.
- In our notation:

$$\begin{aligned} \text{Prop}_{X=x}(Y=y) &= \frac{1}{n_{X=x}}\text{Freq}_{X=x}(Y=y) \\ &= \frac{\text{Freq}(X=x, Y=y)}{\text{Freq}(X=x)} \end{aligned}$$

Debt Ceiling Vote

Figure: Display of House votes by party on the recent debt-ceiling deal

Exercise

- Compute the distribution of votes, both unconditionally and conditioned on party.

		Vote			
		Yea	**Nay**	**No Vote**	**Marginal**
Party	**GOP**	171	59	10	
	Dem	68	77	48	
	Marginal				

Association

- If two variables are unrelated to each other, the conditional distributions for one variable should be similar across levels of the conditioning variable.
- For example, we would probably not expect a relationship between, say, the last digit of someone's phone number, and their marital status
  - Similar percentages of "4"s are married as of "6"s.
  - That is, the *conditional distributions* of marital status given phone digit are the same.
- When this does not happen, two variables are **associated**.

Parallel Categories

- One form of association can occur when the two variables have the same levels (set of values).

		Child		
		Buck.	**Unbuck.**	**Marginal**
Parent	**Buck.**	56	8	64
	Unbuck.	2	16	18
	Marginal	58	24	82

- Notice that children of unbuckled parents tend to be buckled; similarly, children of unbuckled parents tend to be unbuckled.

Parallel Categories

- This is even clearer if we condition on parent's status and compute proportions.

		Child		
		Buck.	**Unbuck.**	**Marginal**
Parent	**Buck.**	0.875	0.125	1.000
	Unbuck.	0.1111	0.8889	1.000
	Marginal	0.7073	0.2927	1.000

- The cells where the two variables have the same value are called the **main diagonal** of the contingency table.
- Large numbers on the main diagonal show an association.

Visualizing Bivariate Categorical Data

- Various contingency tables can tell us all we want to know, but as with univariate data, it's often easier to have a picture.
- Can always show conditional distributions with pie charts, but direct comparison is easier with **stacked** or **grouped bar plots**.

Stacked and Grouped Bar Plots

Here the bars represent joint frequencies

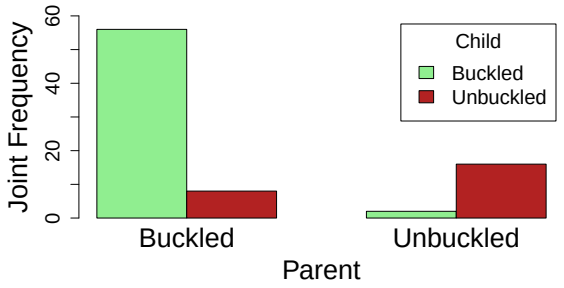


Figure: Parent and Child Seatbelt Status

Stacked and Grouped Bar Plots

To compare the two parent groups, condition on parent status:

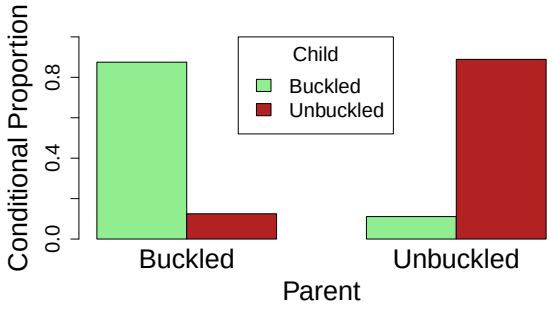


Figure: Parent and Child Seatbelt Status

Stacked and Grouped Bar Plots

When conditioning (and without too many levels) we can also stack (sort of like "linear pie charts")

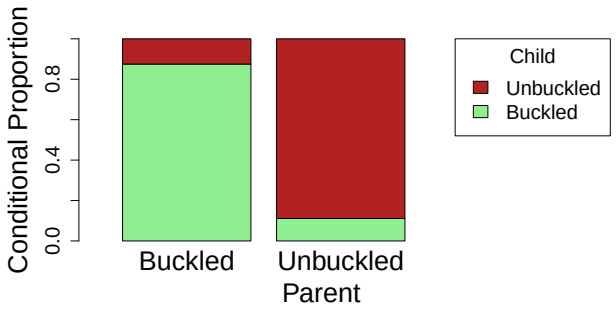


Figure: Parent and Child Seatbelt Status

Stacked and Grouped Bar Plots

Occasionally we may condition on the "color", rather than the group, to compare distributions across colors.

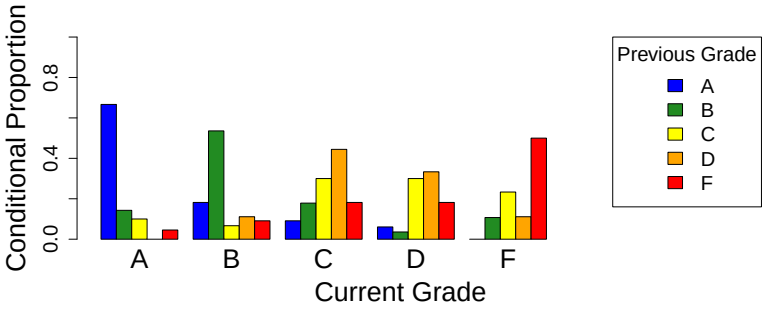


Figure: Grades on an Exam, Conditioned on Previous Grade

Vietnam War Opinions

January 1971 Gallup Poll

"A proposal has been made in Congress to require the U.S. government to bring home all U.S. troops before the end of this year. Would you like to have your congressman vote for or against this proposal?"

Vietnam War Opinions

- Conditional distributions of opinion about Vietnam withdrawal, given education level, based on a January 1971 Gallup poll.

		Education Level			
		Grade	**High**	**College**	**Marginal**
Opinion	**"Dove"**				0.73
	"Hawk"				0.27
	Marginal	1.00	1.00	1.00	

Vietnam War Opinions

From an October 2001 article in *The Economist* entitled "Treason of the Intellectuals?"

"Back in Vietnam days, the anti-war movement spread from the intelligentsia into the rest of the population, eventually paralysing the country's will to fight."

Source `http://www.economist.com/node/806289`

Vietnam War Opinions

- The actual results:

		Education Level			
		Grade	**High**	**College**	**Marginal**
Opinion	**"Dove"**	0.80	0.75	0.60	0.73
	"Hawk"	0.20	0.25	0.40	0.27
	Marginal	1.00	1.00	1.00	

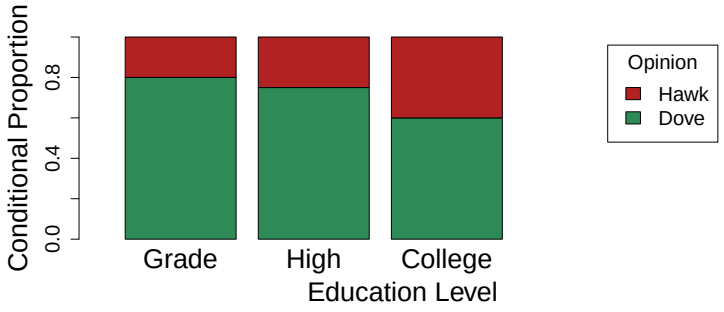


Figure: Opinions about Vietnam Withdrawal, Conditioned on Education Level

More Than Two Variables

- With more than two categorical variables, the ideas are the same; things are just a bit more complicated.
- Can't draw contingency tables in 3D very easily, so we have a separate table for each level of the third variable.
- Think of a "cube" that we slice into layers.

Example: Race and the Death Penalty

- Example: Florida homocide cases

Victim=White	**Death Penalty**	
	Yes	No
Defendant		
White	19	132
Black	11	52

Victim=Black	**Death Penalty**	
	Yes	No
Defendant		
White	0	9
Black	6	97

Example: Race and the Death Penalty

- Now, "marginalizing" (summing) over one variable gives us back a 2×2 table:

	Death Penalty	
	Yes	No
Defendant		
White	19	141
Black	17	149

Example: Race and the Death Penalty

- Conditioning on Defendant's race, the rate of death penalty is similar between races (slightly higher for white defendants):

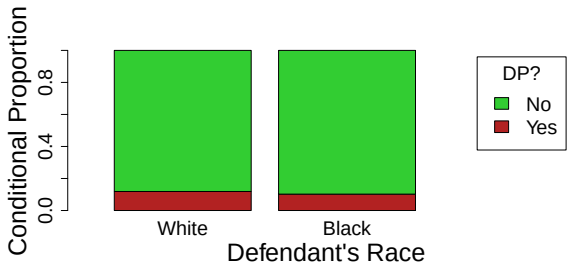


Figure: Proportions of Death Sentences Conditioned on Defendant's Race

Example: Race and the Death Penalty

- However, notice what happens when we condition on *both* defendant's and victim's race (i.e., on all combinations):

Figure: Proportions of Death Sentences Conditioned on Both Victim's and Defendant's Race

Example: Race and the Death Penalty

- What's happening?

Victim=White	Death Penalty	
	Yes	No
Defendant		
White	19	132
Black	11	52

Victim=Black	Death Penalty	
	Yes	No
Defendant		
White	0	9
Black	6	97

Example: Race and the Death Penalty

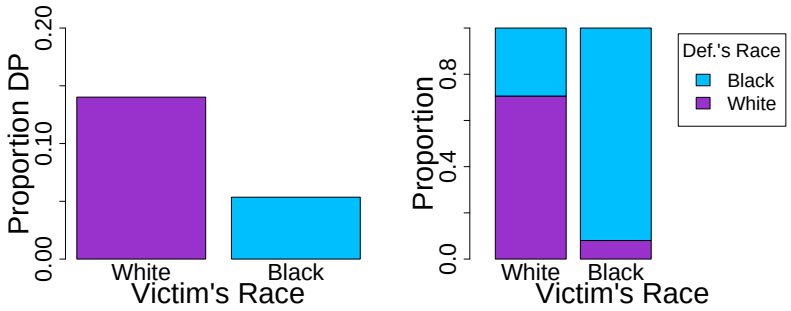
- What's happening?

DP = Yes	Defendant	
	White	Black
Victim		
White	19	11
Black	0	6

DP = No	Defendant	
	White	Black
Victim		
White	132	52
Black	9	97

Example: Race and the Death Penalty

- What's happening?



- The DP is applied more often for white victims; and most homocides involve same-race individuals.

Simpson's "Paradox"

- This phenomenon, of a (marginal) association changing direction when conditioning on a third variable, is called **Simpson's Paradox**.