How do we analyze qualitative data for associations?

Statistics
Class Notes
Contingency Tables and Association (Section 4.4)

Is there an association between your level of education and whether or not you're employed? Can we analyze data to see if women are being discriminated against in college admissions? Does your age make you more or less likely to purchase products that are labeled "Made in America"? We cannot use regression to find associations between categorical (qualitative) data, but we have some tools at our disposal.

We will often see data for two qualitative variables displayed in a table. This will be called a **contingency table** or a **two-way table**, because it relates two categories of data. For the one shown below, the **row variable** is employment status, because each row in the table describes the employment status of a group. The **column variable** is level of education. Each box inside the table is referred to as a **cell**.

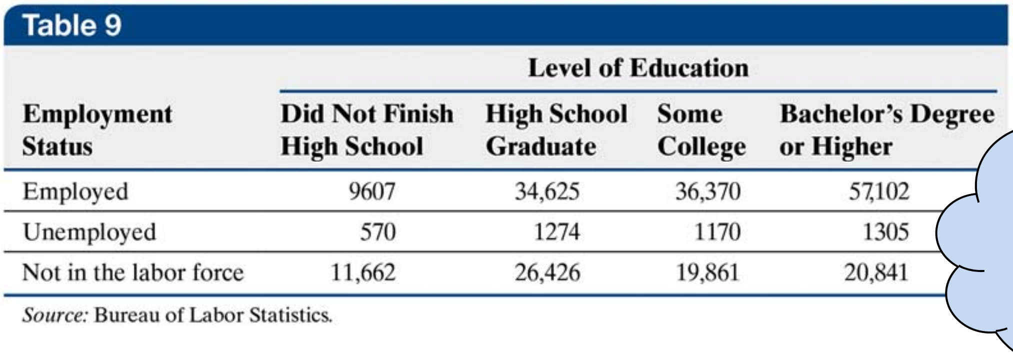


Table 9

	Level of Education			
Employment Status	**Did Not Finish High School**	**High School Graduate**	**Some College**	**Bachelor's Degree or Higher**
Employed	9607	34,625	36,370	57,102
Unemployed	570	1274	1170	1305
Not in the labor force	11,662	26,426	19,861	20,841

Source: Bureau of Labor Statistics.

We use the margins of the table for the totals. Hence the name.

This data (measured in thousands) is the number of all US residents aged 25 years and older in November 2018 for each category.

The first thing we will do with such a table is calculate the **marginal distribution**. This simply means we will add each row and each column, adding a **Total** row and column.

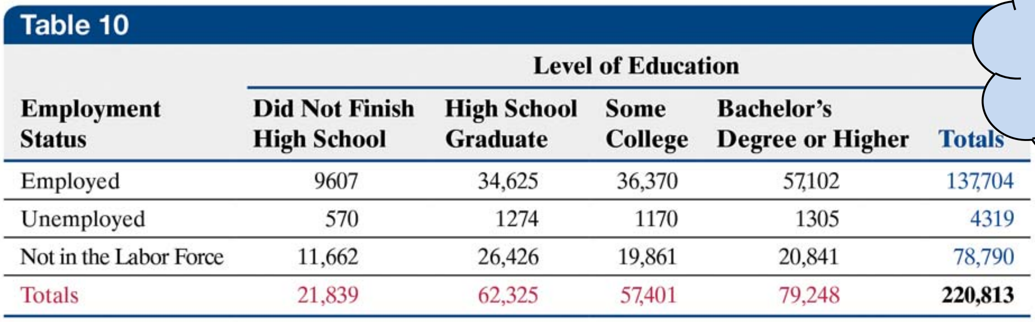


Table 10

	Level of Education				
Employment Status	**Did Not Finish High School**	**High School Graduate**	**Some College**	**Bachelor's Degree or Higher**	**Totals**
Employed	9607	34,625	36,370	57,102	137,704
Unemployed	570	1274	1170	1305	4319
Not in the Labor Force	11,662	26,426	19,861	20,841	78,790
Totals	21,839	62,325	57,401	79,248	**220,813**

Percent of total who are employed = 137,704/220,813 ≈ 0.624 or 62.4 %

Recognize that these totals are **frequency** counts. You will want to find **relative frequencies** by dividing each total by the overall total (in this case, 220,813 found in the lower right corner). These relative frequencies should add to 1 (or 100%), except possible for rounding.

So, does obtaining more education relate to being employed more? If level of education makes no difference, then we would expect the relative frequencies along the “Employed” row (for each level of education) to be close to 62.4 %. Let’s analyze the data.

Conditional Distributions:

Definition: A **conditional distribution** lists the relative frequency of each category of the response variable (employment status for our example), given a specific value of the explanatory variable (level of education) in the contingency table.

expl 1a: Use the marginal frequency table (reproduced here) to complete the conditional distribution for this data.

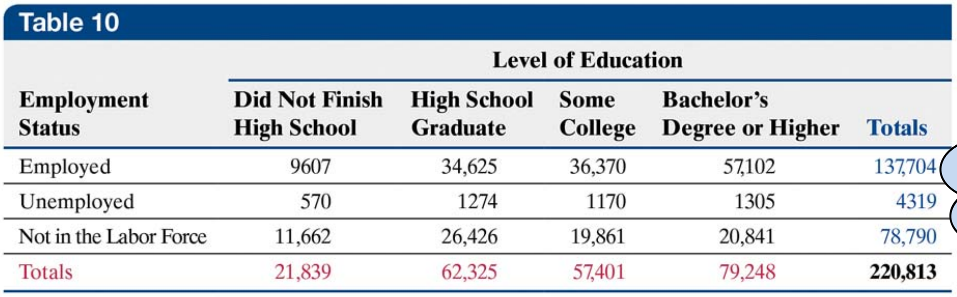


Table 10

Employment Status	Level of Education Did Not Finish High School	High School Graduate	Some College	Bachelor’s Degree or Higher	Totals
Employed	9607	34,625	36,370	57,102	137,704
Unemployed	570	1274	1170	1305	4319
Not in the Labor Force	11,662	26,426	19,861	20,841	78,790
Totals	21,839	62,325	57,401	79,248	**220,813**

Divide each number by its column total.

	Level of Education			
Employment Status	**Did not finish HS**	**HS graduate**	**Some college**	**Bachelor’s degree or higher**
Employed				
Unemployed				
Not in Labor Force				

expl 1b: Read the top row from left to right. What happens to the percent of those employed as the level of education increases? Does this happen for each category of employment status?

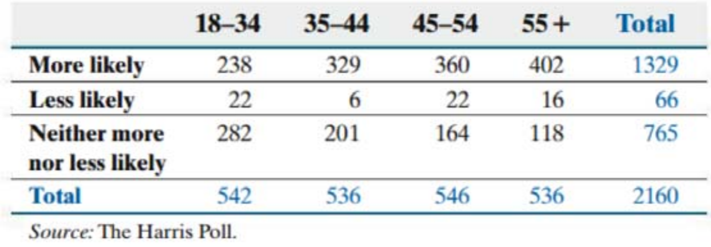
expl 1c: What is the percent of US residents (who are 25 years or older) who are *unemployed*?

expl 1d: What is the percent of US residents with Bachelor's degrees or higher (who are 25 years or older) who are employed?

We could have swapped the roles of the variables, dividing each frequency by the row total.

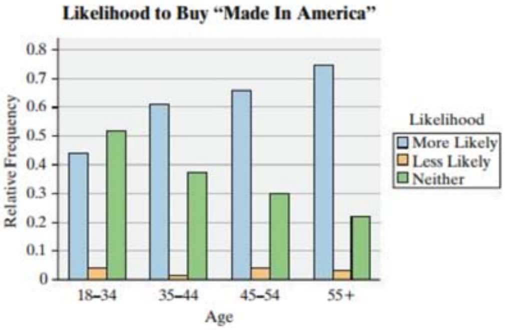
Bar Graphs of Conditional Distributions:
A side-by-side bar graph can help show off the differences we see in data.

expl 2: Here is data from a random sample of Americans (18 years and older) who were asked if they are more or less likely to buy a product that is labeled "Made in America". The data is split up by age groups.



	18–34	35–44	45–54	55+	Total
More likely	238	329	360	402	1329
Less likely	22	6	22	16	66
Neither more nor less likely	282	201	164	118	765
Total	542	536	546	536	2160

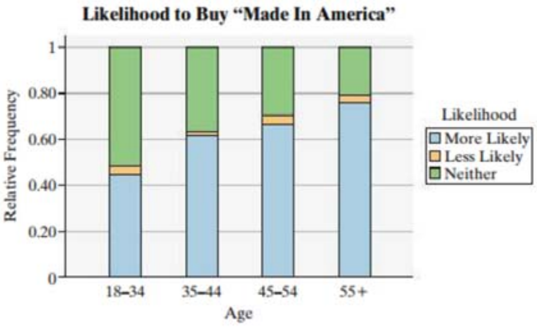
Source: The Harris Poll.



Comment on the connection between likelihood to buy and age.

An alternative to the bar graph you see above is the stacked bar graph here. We see the same trend as age increases.

We also are able to see each segment as a percentage of their age group. Notice each "bar" has a total height of 100%.



We should remind ourselves *not* to conclude that level of education *causes* the difference in employment status or that age *causes* the difference we see in likelihood to buy products labeled "Made in America". We see an association but we do *not* yet have evidence that one thing is *caused* by the other (causality). There are certainly other factors at work.

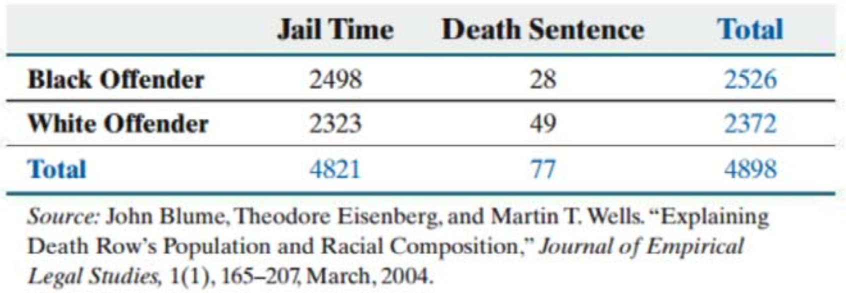
Hey Lisa, what's that?

Simpson's Paradox:

Simpson's paradox is named after Edward H. Simpson, a British statistician who first described the phenomenon in a 1951 paper, detailing the statistical paradox where a trend appears in separate groups of data but reverses when the groups are combined; although similar effects were noted by other statisticians like Karl Pearson earlier.

Got this tidbit from the AI overview in Google. So, you know, hope it's true…

expl 3: Let's check out this data concerning the rates of death sentences handed down for black and white offenders convicted of murder.

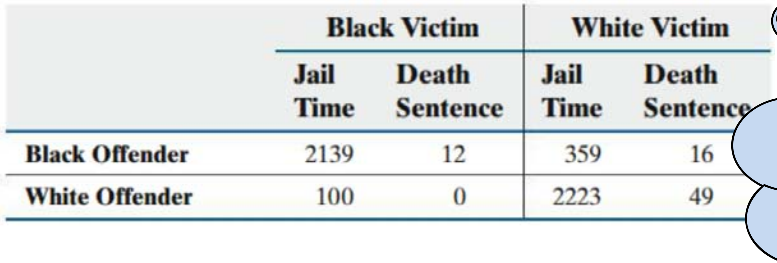


	Jail Time	Death Sentence	Total
Black Offender	2498	28	2526
White Offender	2323	49	2372
Total	4821	77	4898

Source: John Blume, Theodore Eisenberg, and Martin T. Wells. "Explaining Death Row's Population and Racial Composition," *Journal of Empirical Legal Studies*, 1(1), 165–207, March, 2004.

a.) Find the percentage of black offenders who were sentenced to death. Do the same for white offenders. Which race appears to receive the death sentence more frequently?

b.) Here is the same data but this time we break it up by race of the victim.



	Black Victim		White Victim	
	Jail Time	Death Sentence	Jail Time	Death Sentence
Black Offender	2139	12	359	16
White Offender	100	0	2223	49

Notice how 15% of the victims of black offenders were white. In contrast, 96% of the victims of white offenders were white.

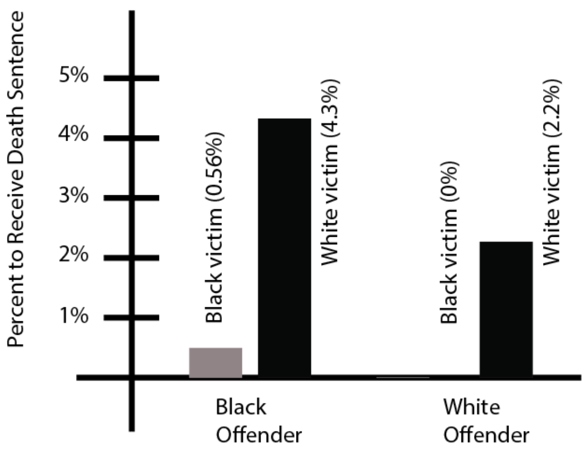
Let's make a conditional distribution by race of the victim.

	Black Victim		**White Victim**	
	Jail Time	**Death Sentence**	**Jail Time**	**Death Sentence**
Black Offender				
White Offender				

For black and white offenders separately, divide each number by the total for that race's victims.

Here is a bar graph that helps illustrate the point that even though it is true that white offenders are (overall) sentenced to death at a higher percentage than black offenders, the reason is because they murder more white people and that is what gets you a death sentence.

Sentenced to Death, by Victim’s Race



Which race offender now appears to receive the higher percentage of death sentences?

Worksheet: Simpson’s Paradox:

Here we explore data that appears to show smoking contributes to higher survival rates. Can it be true? Delving deeper into the data by age reveals something else.