**Chapter 4.2: Relationships Between Categorical Variables**

* A large sample survey interviewed a random sample of young adults in 2000 and 2001. One question asked was “Where do you live now? That is, where do you stay most often?”
* The table below displays the results for ages 19-22 in a **two-way table.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **19** | **20** | **21** | **22** | TOTAL |
| **Parent’s Home** | 324 | 378 | 337 | 318 | 1357 |
| **Another person’s home** | 37 | 47 | 40 | 38 | 162 |
| **Your own place** | 116 | 279 | 372 | 487 | 1254 |
| **Group Quarters** | 58 | 60 | 49 | 25 | 192 |
| **Other** | 5 | 2 | 3 | 9 | 19 |
| TOTAL | 540 | 766 | 801 | 877 | 2984 |

* To *interpret the table*, first look at the **marginal distribution** of where young people live.





* Let’s look at the **marginal distribution** of the ages in this study.
* Find the **conditional distribution** of living arrangements among 19 year olds.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

* Find the **conditional distributions** of living arrangements among the remaining ages.
* We could also find the conditional distributions of ages, given a specific living arrangement.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |

* There is no standard graph to picture relationships between categorical variables (although bar-graphs can be helpful)
* Also, there is no single standard numerical measure of association.
	+ If there is clear explanatory/response relationship, compare the \_\_\_\_\_\_\_\_\_\_\_\_ distributions of the \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ variable for the separate values of the \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ variable.

**SIMPSON’S PARADOX**

* Just like relationships between quantitative variables can be influenced by “\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_”, the relationships between categorical variables are susceptible.
	+ In fact, some lurking variables \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

As some of you might know, Bart Simpson is a bright little lad, but not the most academically inclined little kid.  However his sister, Lisa, is both bright and academically inclined.  It was on a Thanksgiving Day at the Simpson’s dinner table far off in the future, when Lisa had come home for vacation from college, where Bart made such an outlandish claim that it is now known as, “Simpson’s Paradox”.

   Right after a large belch and a giggle from his dad, Bart said out of the blue, “Did you know that I’m a better overall math student then Ms. College Girl smarty pants!”  The family gasped, and Grandpa began to choke on a turkey bone.

After Bart calculated an overall average of all Junior High and High School math test scores, he found his average to be an **83.6%** and Lisa’s to be an **81.4%.**  So how did Bart come to such an outlandish claim?  Let’s analyze.  Lisa being smart and all decided to look at the grades broken down by Junior High and High School.  Here’s what she found.

So for both sets of scores Lisa score much better.  However Lisa skipped most of Junior High and went right to High School, consequently she only took 5 tests.  Bart enjoyed Junior High math, but played hooky most of the time in High School and consequently took only 5 tests.

**Simpson’s Paradox** can happen when two things occur

 (1)

 (2)