**Topic 6**

**Two-Way Tables**

Recall that the first thing we do in the data analysis process (that is, after compiling and cleaning the data – recall the 31 letters I removed, this was a bit of hubris), we look at it.

Two way tables are an important and effective way of displaying

* categorical variables
* ordinal variables
* quantitative variable that have been categorized

Why do them?

1. They are ubiquitous. One of the most common ways to display data, widely understood
2. The are not only very common, but very useful

Usually, after we collect data for a study with M questions an 2024 participants, the data looks like this

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

OU q1 q2 q3 q4 … qM

--------------------------------------------------------------------------

0001 M 4 3 A … 4

0002 F 4 7 S … 3

… … … … … … …

2024 M 4 4 N … 4

---------------------------------------------------------------------------

q1 is Gender (Male,Female), and

q4 is “Do you like surveys” (Always, Sometimes, Never)

If we aggregate the information, we might get

q1: Female Male

 n 1062 962

 p .52 .48

q4: n p

 Always 120 .06

Sometimes 1586 .78

Never 318 .16

OK, that might be interesting in itself, and we might not want more, but say we are interested how the two variables might be associated.

First, let’s rewrite the tables above where the two questions are intermingled.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Females | Males | Total |
| Always |  |  | 120 |
| Sometimes |  |  | 1586 |
| Never |  |  | 318 |
| Total | 1062 | 962 | 2024 |

The colored areas are the **marginal distributions.**  They occur in the margins of the two way table

|  |  |  |  |
| --- | --- | --- | --- |
|  | Females | Males | Total |
| Always |  |  | 120 |
| Sometimes |  |  | 1586 |
| Never |  |  | 318 |
| Total | 1062 | 962 | 2024 |

Note that the sum of the different responses add to the total number of surveys in the sample.

This chart has a dual in terms of proportions.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Females | Males | Total |
| Always |  |  | 120 / 2024 |
| Sometimes |  |  | 1586 / 2024 |
| Never |  |  | 318 / 2024 |
| Total | 1062 / 2024 | 962 / 2024 | 2024 / 2024 |

So In terms of **proportions** (rounding to 2 digits)

|  |  |  |  |
| --- | --- | --- | --- |
|  | Females | Males | Total |
| Always |  |  | .06 |
| Sometimes |  |  | .78 |
| Never |  |  | .16 |
| Total | .52 | .48 | 1 |

Or in terms of **percentages** (% = $\hat{p}$ x 100)

|  |  |  |  |
| --- | --- | --- | --- |
|  | Females | Males | Total |
| Always |  |  | 6% |
| Sometimes |  |  | 78% |
| Never |  |  | 16% |
| Total | 52% | 48% | 100% |

Marginal distributions give the proportion (or percentage) of the response relative to the total population. Note that all n’s, $\hat{p}$’s, or %’s in a margin must add to N, 1, or 100, respectively

We can examine the survey and find values for the empty cells:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Females | Males | Total |
| Always | 70 | 50 | 120 |
| Sometimes | 950 | 636 | 1586 |
| Never | 42 | 276 | 318 |
| Total | 1062 | 962 | 2024 |

The counts filling the cells are called frequency counts

And once again, this table has a dual

|  |  |  |  |
| --- | --- | --- | --- |
|  | Females | Males | Total |
| Always | .03 | .03 | .06 |
| Sometimes | .47 | .31 | .78 |
| Never | .02 | .14 | .16 |
| Total | .52 | .48 | 1 |

With the values called relative frequencies.

Notice that the rows and columns add to their marginal values.

Sometimes one only gets a table of ***joint distributions***

|  |  |  |
| --- | --- | --- |
|  | Females | Males |
| Always | .03 | .03 |
| Sometimes | .47 | .31 |
| Never | .02 | .14 |

Join distribution occurs when two or more variables are measured from one OU.

The 3 components of a two-way table are the

* Two Variables and their values
* Marginal Distribution of data
* Joint Distribution of data

Marginal distributions *can* be calculated from the joint distributions.

Joint distributions cannot be calculated from marginal distributions unless the variables are independent.

**CONDITIONAL DISTRIBUTIONS**

Sometimes one wants to know the proportion of responses given a specific value in the other variable.

This is called a **Conditional Distribution.** In this case, the proportional values (or relative frequencies) can change, because the population is being restricted to specific value of the other variable.

The conditional distribution of responses, given the responder was Female, would look like

|  |  |
| --- | --- |
|  | Females |
| Always | 70 |
| Sometimes | 950 |
| Never | 42 |
| Total | 1062 |

|  |  |
| --- | --- |
|  | Females |
| Always | 70 / 1062 |
| Sometimes | 950 / 1062 |
| Never | 42 / 1062 |
| Total | 1062 / 1062 |

The *conditional* relative frequencies

|  |  |
| --- | --- |
|  | Females |
| Always | .07 |
| Sometimes | .89 |
| Never | .04 |
| Total | 1 |

These values can be calculated for each column, and we get the **Relative Frequencies for Columns**

|  |  |  |
| --- | --- | --- |
|  | Females | Males |
| Always | .07 | .05 |
| Sometimes | .89 | .66 |
| Never | .04 | .29 |
| Total | 1 | 1 |

And likewise we can build a table of **Relative Frequencies for Rows**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Females | Males | Total |
| Always | .58 | .42 | 1 |
| Sometimes | .60 | .40 | 1 |
| Never | .13 | .87 | 1 |

**Random distributions**

Two way tables are prone to extreme values just like any other sampling distribution. Extreme values are highly improbable, but not impossible