The summarizing and visualizing variables and relationships between variables is known as *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_* statistics.

The type of summary statistics and visualization methods we use depends on whether we are analyzing *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_* variables, *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_* variables, or both.

Let’s consider quantitative variables first, and let’s deal with only one categorical variable.

The first way we can summarize categorical variables is using a *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_* table.

A frequency table shows the number of cases in each category as well as the total.

Remember – all cases must fall in one and ONLY *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_* category.

|  |  |
| --- | --- |
| Android | 458 |
| iPhone | 437 |
| Blackberry | 141 |
| Non Smartphone | 924 |
| No cell phone | 293 |
| Total | 2,253 |

If this frequency table represents the type of cell phones owned by students, you can see 458 use an Android, 437 use an iPhone, etc. In total, 2,253 students were surveyed.

Another way to describe data is by providing proportions. The *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_* in a category is found by dividing the number in a category by the total sample size.

Depending on whether we are talking about a population or a sample, we have specific notation for proportions.

For *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*, we denote proportion as p “hat”.

For *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*, we denote proportion as a lowercase p.

Going back to the type of cell phone used by students, this is a sample, so we use p hat.

p-hat android = 458/2253 = .203

p-hat iphone = 437/2253 = .194

p-hat blackberry = 141/2253 = .063

p-hat smartphone = 924/2253 = .410

p-hat no cellphone = 293/2253 = .130

|  |  |
| --- | --- |
| Android | 0.203 |
| iPhone | 0.194 |
| Blackberry | 0.063 |
| Non Smartphone | 0.410 |
| No cell phone | 0.130 |

If we add the proportion from each category, it should sum to one, and it does.

We display proportions using a *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_* frequency table. Notice that it looks like a regular frequency table with two exceptions.

1 – It displays frequencies rather than counts.

2 – There is no total, but you should always make sure the frequencies sum to one.

Next, let’s look at a couple of ways to graphically present categorical variables.

First up is a bar chart.

In a bar chart, the height of the bar represents the *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_* of cases in that category. In a bar chart, the bars *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_* touch.

Next, we can use a pie chart. Pie charts are a good way to present *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_* because the relative area of each slice of the pie corresponds to the proportion in each category.

Example –

In a survey, students in a class identified themselves as the following:

14 First-Year

2 Sophomores

2 Juniors

3 Seniors

1. From what population was this sample drawn? *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*
2. Is the sample a random sample? *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*
3. Create a frequency table from the sample data.
4. What proportion of the sample are first-year? *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*
5. Create a relative frequency table from the sample data.
6. Create a bar chart using the sample data.

Now let’s look at two quantitative variables.

Let’s say we collected data from college students in two areas:

1. Relationship status

2. Gender

We can display these data in a two-way table.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Female | Male | Total |
| In a relationship | 32 | 10 | 42 |
| It’s complicated | 12 | 7 | 19 |
| Single | 63 | 45 | 108 |
| Total | 107 | 62 | 169 |

Notice we have “Gender” in columns and “Relationship status” in rows. It doesn’t matter which variable is displayed in rows and which is displayed in columns.

We can use a two-way table to determine proportions.

1. What proportion of students is in a relationship? *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

2. What proportion of females is in a relationship? *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

3. What proportion of students in a relationship is female? *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

4. What proportion of males is in a relationship? *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

Be sure to accurately interpret the question. “The proportion of females in a relationship” is NOT the same as “The proportion of people in a relationship who are female”!

One is 30% and the other is 76%

30% of females in the sample say they are in a relationship.

16% of males in the sample say they are in a relationship.

Why the difference?

A difference in proportions is a difference in proportion for one categorical variable calculated for different levels of the other categorical variable.

Example - What is the difference in proportion of females in a relationship compared to the proportion of males in a relationship?

p-hat F – p-hat M = .30 - .16 =.14

What is the difference in proportion of students in a relationship who are female compared to the proportion of single students who are female?

p-hat R – p-hat S = (32/42) – (63/108) = .76 - .58 = .18

To display data in a two-way table in a graph, the height of each bar is the *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_* in the corresponding cell in the two-way table.

A *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_* bar graph is like a side-by-side bar graph, but the bars are stacked instead of side-by-side.

A certain class identified themselves as:

7 Female First-Years

1 Female Sophomore

2 Female Junior

1 Female Senior

6 Male First-Years

1 Male Sophomore

0 Male Junior

2 Male Senior

1. Create a two-way table.

2. What proportion of the sample responded as female freshmen? *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

3. What proportion of females responded as freshmen? *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

4. What proportion of freshmen responded as female? *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

5. What proportion of freshmen responded as male? *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

6. Must #4 and #5 total to 1? *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_* Why/Why not? *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

An *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*  is an observed value that is notably distinct from the other values in a dataset.

Outliers can be informally identified by looking at a plot, but one general rule of thumb for identifying outliers is data values more than 1.5 Outliers can be informally identified by looking at a plot, but one general rule of thumb for identifying outliers is data values more than 1.5 IQRs beyond the quartiles beyond the quartiles

A *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*  association means that values of one variable tend to be higher when values of the other variable are higher.

A n*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*  association means that values of one variable tend to be lower when values of the other variable are higher.

Two variables are not *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*  if knowing the value of one variable does not give you any information about the value of the other variable.

A *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_* is the graph of the relationship between two variables.

*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_* is a measure of the strength and direction of linear association between two quantitative variables.

Symbols for correlation:

Sample correlation: *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

Population correlation: *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*  ≤ *r* ≤ *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

The sign indicates the direction of association

positive association: *r* > *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

negative association: *r* < *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

no linear association: *r* ≈ *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

The closer *r* is to ±1, the *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*  the linear association