Summarizing Data: One Variable

Categorical Variables

So far we have been working with quantitative variables. The variables are represented by numbers. For quantitative variables we can look at

- minimum, maximum
- median, quartiles
- boxplots
- mean
- standard deviation
- histogram
- skewed vs symmetric
- extreme values

Example 1

Let's look back at the life expectancy data. Each country in the dataset is located in one of six regions. The geographic region is a **categorical variable** since it classifies observations. We cannot perform mathematical operations such as calculating the mean for a categorical variable. One simple thing we can do is count how many of the countries are located in each of the six regions as shown in Table 1.

Region	Count
The Americas	39
Europe & Central Asia	50
East Asia & Pacific	30
South Asia	8
Middle East & N. Africa	21
Sub-Saharan Africa	49
Total	**197**

Table 1: Six geographical regions part of the life expectancy dataset

We can represent the numbers in Table 1 by the **bar chart** in Figure 1. In a bar chart there is a separate bar for each of the different categories. The height of the bar corresponds to how many countries are in that region. For example, the red bar indicates that in the Americas there are close to 40 countries.

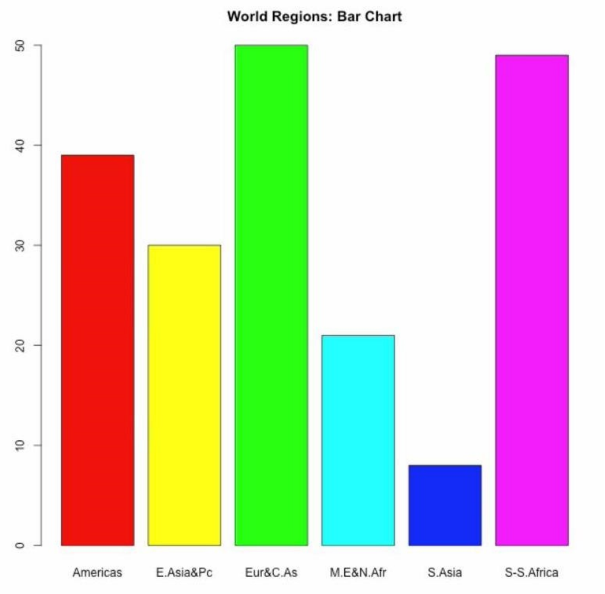

Figure 1: World regions bar chart

Instead of representing the counts, we can represent the relative frequencies, which is the fraction of the countries in each of the six different regions:

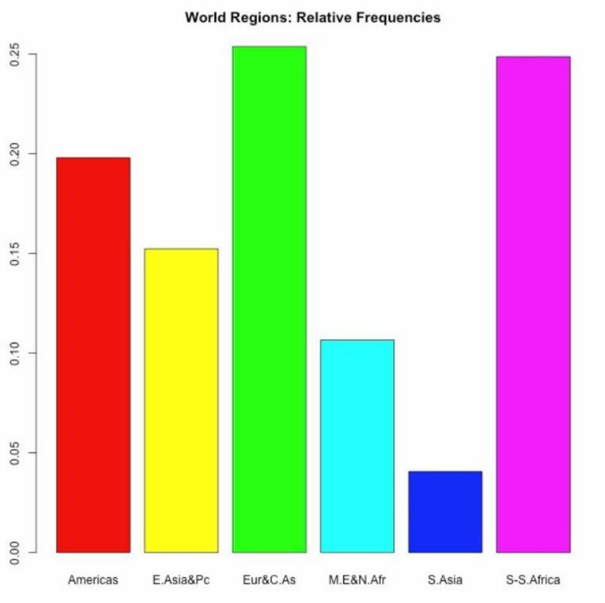

Figure 2: World regions relative frequency bar chart

The bar chart in Figure 2 looks exactly like the bar chart in Figure 1. The difference is that the scale on the y axis is changed since we have divided by the total number of countries.

For example, the relative frequency of the countries in the Americas is $39/197 = 0.198$ and the height of the red bar is just under 0.2 in the bar chart in Figure 2. Note that now the heights of the bars add up to 1.

Another way that we can illustrate these data is with a pie chart:

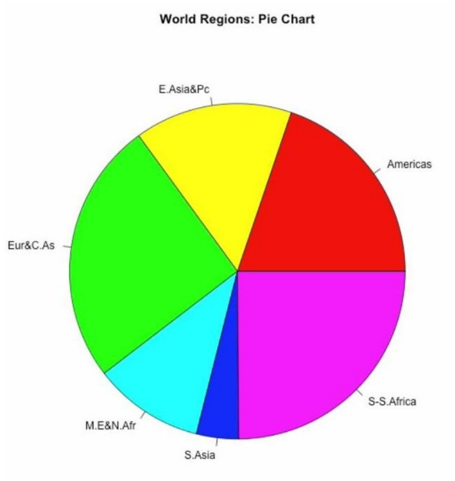

Figure 3: World regions pie chart

A pie chart consists of a circle divided into pieces. The size of each piece is in proportion to how many of the data fit into each of the different categories. The pie charts provides similar information to the bar chart.

Example 2

Let's consider the skeleton data one more time. In the past we talked about quantitative variables like the estimated age at the time of death, or the difference between the estimated and the true age at the time of death. In this dataset we also have some categorical variables such as the sex of the skeleton. Figure 4 shows that there are 281 male and 119 female skeletons.

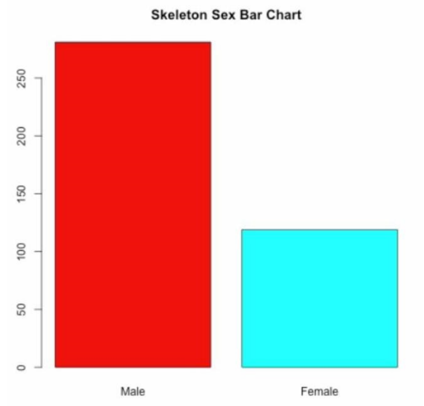

Figure 4: Skeleton sex bar chart

Here is a bar chart showing the relative frequencies of sex

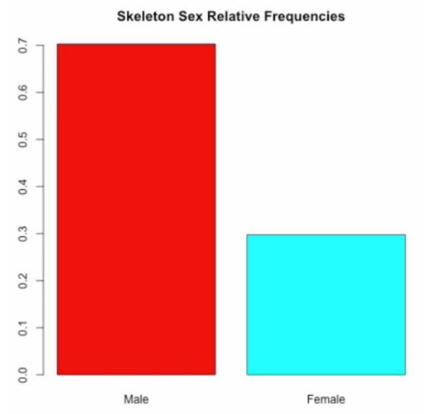

Figure 5: Skeleton sex relative frequency bar chart

and here is the same information displayed in a pie chart.

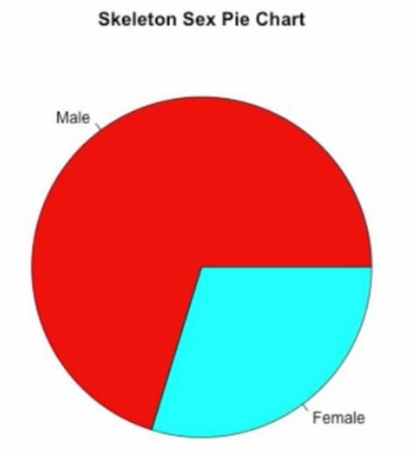

Figure 6: Skeleton sex pie chart

Example 3
Another example of a categorical variable for these skeletons is the mass category. Each person at the time of death was categorized as being of normal weight (225 skeletons) or underweight (74 skeletons) or overweight (82 skeletons) or obese (20 skeletons). Table 2 displays this information about the mass category variable.

Mass Category	Count
Normal	225
Underweight	74
Overweight	81
Obese	20
Total	**400**

Table 2: Four BMI mass categories part of the skeletons dataset.

Figures 7, 8 and 9 show the distribution of the mass category variable in a bar chart with counts, a relative frequency bar chart, and a pie chart, respectively.

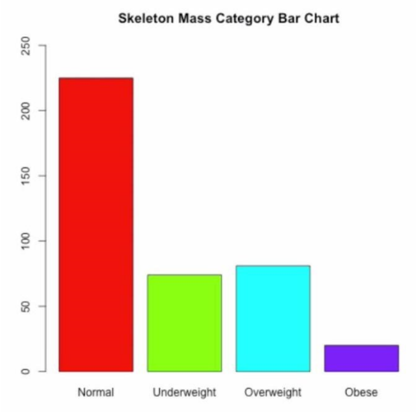

Figure 7: Skeleton mass category bar chart

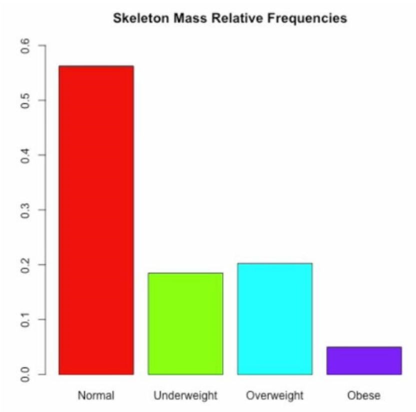

Figure 8: Skeleton mass category relative frequency bar chart

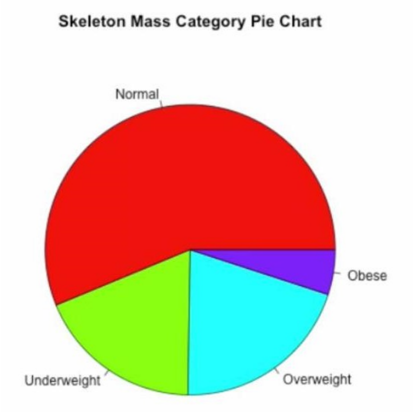

Figure 9: Skeleton mass category pie chart