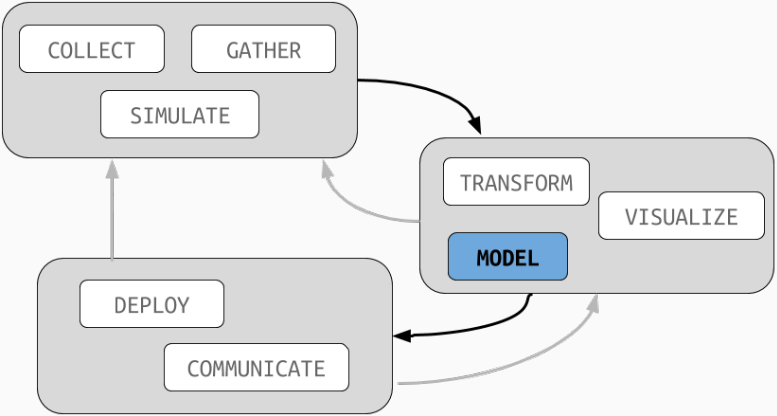
Lecture 05: Numeric Summaries

Taylor Arnold


COLLECT
GATHER
SIMULATE
TRANSFORM
VISUALIZE
MODEL
DEPLOY
COMMUNICATE

Numeric Summaries

Graphics are an excellent way of summarizing and presenting information contained in a dataset. In many cases it can be useful to combine these with purely numeric summaries. These summaries are something colloquially called *statistics*, though I prefer to avoid this terminology.

Mean

The first statistical summary that most people learn about is the **mean**, also commonly known as an average. It is calculated by adding all of a variables values together and dividing by the total number of values. If we have a dataset of n points with a variable x (denoting x_1 as the first value, x_2 and the second, and so forth), the mean can be formally defined as:

$$\begin{aligned} \text{mean}(x) &= \overline{x} \\ &= \frac{x_1 + x_2 + \cdots x_n}{n} \\ &= \frac{1}{n} \cdot \sum_i x_i \end{aligned}$$

Mean

The notation of using $\overline{x}$ to represent the mean is very common throughout the sciences and social sciences. It is often used in textbooks and papers without even being defined. To calculate means in R, as we have already, seen we can use the `mean` function. Here is an illustration that `mean` behaves as expected using the `sum` and `nrow` functions for comparison.

```r
mean(msleep$awake)
```

```
## [1] 13.56747
```

```r
sum(msleep$awake) / nrow(msleep)
```

```
## [1] 13.56747
```

Quantiles

Deciles

There are a number of functions that allow us to compute quantiles, a generalization of percentiles. For example, the `deciles` function splits the dataset into 10 equally sized buckets:

```r
deciles(msleep$awake)
```

```
##    0%   10%   20%   30%   40%   50%   60%   70%   80%   90%  100%
##  4.10  8.12  9.60 11.20 12.86 13.90 14.52 15.48 17.76 20.08 22.10
```

Deciles

```
##    0%   10%   20%   30%   40%   50%   60%   70%   80%   90%  100%
##  4.10  8.12  9.60 11.20 12.86 13.90 14.52 15.48 17.76 20.08 22.10
```

- about 1/2 of the mammals are awake less than 13.90 hours and about 1/2 are awake more than 13.90. I use the word "about" here due to subtitles regarding ties and repeated values; for all practical purposes this is generally not important.
- the 50% percentile has a special name that you have probably heard before: the *median*.
- roughly 1/10 of the mammals are awake less than 8.12 hours and 1/10 are awake more than 20.8 hours.
- the sleepiest mammal is awake for only 4.1 hours and that one mammal is awake 22.1 hours of the day.

Quartiles

We can similarly calculate what are called quartiles, splitting the data into four equally using the `quartiles` function:

```r
quartiles(msleep$awake)
```

```
##    0%   25%   50%   75%  100%
##  4.10 10.25 13.90 16.15 22.10
```

Notice that four buckets requires 5 numbers, and that three of these line up with the deciles above.

Ventiles and Percentiles

There are also functions `ventiles` (20) and `percentiles` that can be quite useful:

```r
ventiles(msleep$awake)
percentiles(msleep$awake)
```

Ventiles and Percentiles

Ventiles are a bit esoteric, but I have found in my work that they can be very useful in practice. Percentiles are often useful when we want to look at the extreme values, such as the 97th, 98th and 99th percentiles.

```
percentiles(msleep$awake)[95:100]
```

```
##    94%    95%    96%    97%    98%    99%
## 20.716 20.880 20.972 21.054 21.190 21.485
```

Deviation

Deviation

Once we have defined the mean, we can then define the **deviation** of a data value as the difference between the value and its mean:

$$\begin{aligned} d_1 &= x_1 - \overline{x} \\ d_2 &= x_2 - \overline{x} \\ &\vdots \\ d_n &= x_n - \overline{x} \end{aligned}$$

Deviation

There is not a special R function for deviances because they are very easy to calculate using the `mean` function. As an example, here is how to create them:

```r
msleep$awake - mean(msleep$awake)
```

Sum of Squares

We can use deviations to measure how spread of a variable by computing the sum of their squares. These are calculated by the following formula:

$$\begin{aligned}\text{sum of squares} &= (x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots (x_n - \overline{x})^2 \\ &= \sum_i (x_i - \overline{x})^2\end{aligned}$$

Sum of Squares

The sum of squares can be computed in R as:

```r
sum((msleep$awake - mean(msleep$awake))^2)
```

```
## [1] 1625.327
```

Variance

The sum of squares cannot be used directly to compare datasets of different sizes as it grows with the number of points. In order to compare sums of squares across datasets, we use a measurement called **variance** which is simply the average of the sums of squares:

$$\begin{aligned}\text{variance} &= s^2 \\ &= \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots (x_n - \overline{x})^2}{n-1} \\ &= \frac{1}{n-1} \cdot \sum_i (x_i - \overline{x})^2\end{aligned}$$

Variance

- the notation of using s^2 to represent the variance of a dataset is quite common.
- why do we use $n - 1$ rather than n to take the average? The technical reason is that if we want to measure the variance of a population using a sample from that population, we need to use $n - 1$ in order to have an unbiased estimate of the population value.

Variance

The variance can be computed using the var function, or manually as follows:

```
sum((msleep$awake - mean(msleep$awake))^2) /
    (nrow(msleep) - 1)
```

```
## [1] 19.82106
```

```
var(msleep$awake)
```

```
## [1] 19.82106
```

Standard deviation

We often with a quantity called the **standard deviation**, defined as simply the square root of the variance. Why bother taking the square root? For one thing, it is a matter of units. In our example, the variance is given in “squared people” (a nearly meaningless quantity), but the standard deviation is given in “people” just like the variable itself. We can calculate the standard deviation using the function `sd`:

```r
sd(msleep$awake)
```

```
## [1] 4.452085
```