Measures of Spread

Stat 131A, Fall 2018, Prof. Sanchez

Learning Objectives

- Becoming familiar with various measures of spread
- Intro to the functions `range()`, `IQR()`, and `sd()`
- Understand the concept of r.m.s. size of a list of numbers
- Be aware of the difference between SD and SD+

Introduction

Quantitative variables can be summarized using two groups of measures: 1) center, and 2) spread. Just like there are various measures of center (e.g. average, median, mode), we also have several measures of spread or variability:

- range
- interquartile range
- standard deviation (and variance)

Range

The most basic type of measure of spread is the **overall range** or "range". The range is obtained as the difference between the maximum value and the minimum value.

For example, let's consider the values 0, 5, -8, 7, and -3 used in the classic *Statistics* textbook by Freedman, Pisani, and Purves (page 66). To find the range, you need to determine the smallest and largest values which in this case are 7 and -8, respectively. And then obtain the difference:

$$range = 7 - (-8) = 15$$

For illustration purposes, let's implement this minimalist example in R. First we create a vector `x` with the five values. You can use the functions `max()` and `min()` to get the largest and smallest values in `x`:

```
x = c(0, 5, -8, 7, -3)
maximum = max(x)
minimum = min(x)

# range
maximum - minimum
```

```
## [1] 15
```

Actually, there is a `range()` function in R, which gives you the maximum and minimum value (but not the subtraction):

```
# range: max value, and min value
range(x)
```

```
## [1] -8  7
```

```
# range value
range(x)[2] - range(x)[1]
```

```
## [1] 15
```

The range is one type of measure of variability. It tells you the *length* of the scatter in the data. The issue with the range is that extreme values may have a considerable effect on it. For example, if you add a value of 20 to `x` the new range becomes:

```
y = c(x, 20)

# range
max(y) - min(y)
```

```
## [1] 28
```

As you can tell, the presence of outliers will affect the magnitude of the range.

Interquartile Range (IQR)

To overcome the limitations of the range we can use a different type of range called the **interquartile range** or *IQR*. This is a range based not on the minimum and maximum values but on the first and third quartiles.

One way to compute quartiles in R is with the function `quantile()`. There are slightly different formulas to compute quartiles. To find the quartiles—as discussed in most introductory statistics books—you need to use the argument `type = 2` inside the `quantile()` function:

```
x = c(0, 5, -8, 7, -3)

# 1st quartile
Q1 = quantile(x, probs = 0.25, type = 2)
Q1
```

```
## 25%
##  -3
```

```r
# 3rd quartile
Q3 = quantile(x, probs = 0.75, type = 2)
Q3
```

```
## 75%
##   5
```

```r
# IQR
Q3 - Q1
```

```
## 75%
##   8
```

You can also use the dedicated function `IQR()` to compute the interquartile range:

```r
IQR(x, type = 2)
```

```
## [1] 8
```

Compared to the classic range, the IQR is more resistant to outliers because it does not consider the entire set of values, just those between the first and third quartile. If we add a extreme negative value -50, and an extreme positive value of 40 to `x`, the IQR should not be affected:

```r
y = c(x, -50, 40)

# IRQ
IQR(y, type = 2)
```

```
## [1] 15
```

The Root Mean Square (RMS)

Another measure of spread is the Standard Deviation (SD). However, in order to talk about the SD, I will follow the same approach of the SticiGui book and I will first talk about the **Root Mean Square** or RMS.

The values in our toy example are 0, 5, -8, 7, and -3. To find a central value we can use either the average or the median:

```r
x = c(0, 5, -8, 7, -3)

mean(x)
median(x)
```

What about a measure of *size*? In other words, how would you find a measure of how small or how big the values in `x` are? Is it possible to obtain a quantity that tells you something about the representative *magnitude* of values in `x`?

To answer this question about a typical magnitude of values we need to ignore the signs. One way to do this is by looking at the absolute values, and then compute the average:

```
abs(x)
```

```
## [1] 0 5 8 7 3
```

```
mean(abs(x))
```

```
## [1] 4.6
```

For convenience reasons (e.g. algebraic manipulation and nice mathematical properties), statisticians prefer to square the values instead of using the absolute values. And then compute the average of such squares:

```
# square value
x^2
```

```
## [1]  0 25 64 49  9
```

```
# average of square values
sum(x^2) / length(x)
```

```
## [1] 29.4
```

The issue with using square values is that now you end up working with square units, and with a larger number that has little to do with a typical magnitude of the original values. To tackle this problem, we take the square root:

```
# root-mean-square (r.m.s)
sqrt(sum(x^2) / length(x))
```

```
## [1] 5.422177
```

```
# equivalent to
sqrt(mean(x^2))
```

```
## [1] 5.422177
```

The value 5.42 is referred to as the *r.m.s. size* of the numbers in `x`. The RMS size provides a numeric summary for the magnitude of the data. It is not really the average magnitude, but you can think of it as such.

Standard Deviation (SD)

Now that we have introduced the concept of r.m.s. size of a list of numbers, we can talk about a third measure of spread known as the **Standard Deviation** (SD). Simply put, the Standard Deviation is a measure of spread that quantifies the amount of variation around the average.

A keyword is the term **deviation**. In the previous script tutorial—about measures of center—I introduced the concept of *deviations*. If we denote a set of n values with $x_1, x_2, \ldots, x_n$, and a reference value by ref, a deviation is the difference between an observed value x_i and the value of reference ref, that is, $(x_i - ref)$.

A special type of deviation is when the reference value becomes the average. If avg represents the average of $x_1, x_2, \ldots, x_n$, we can calculate the deviations of all observations from the average: $(x_i - avg)$.

The Standard Deviation is based on these deviations. To be more precise, it is based on the R.M.S. size of deviations from the average:

$$SD = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - avg)^2}$$

The SD says how far away numbers $x_1, x_2, \ldots x_n$ are from their average. In this sense, you can think of the SD as the typical magnitude of scatter around the average.

The `sd()` function

All statistical packages come with a function that allows you to calculate the Standard Deviation. In R, there is the function `sd()`. However, the way `sd()` works is by using a slightly different formula:

$$SD^{+} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - avg)^2}$$

Note that SD^{+} divides by $n-1$ instead of n. When the number of values n is big, $\sqrt{n-1}$ is very close to $\sqrt{n}$. However, for relatively small values of n, there diference between $\sqrt{n-1}$ and $\sqrt{n}$ can be considerable.

If you want to use `sd()` to obtain SD, you need to multiply the output by a correction factor of $\frac{\sqrt{n-1}}{n}$:

```
x = c(0, 5, -8, 7, -3)
n = length(x)

# SD
sqrt((n-1)/n) * sd(x)
```

```
## [1] 5.418487
```

```
# SD+
sd(x)
```

```
## [1] 6.058052
```