# Measures of Center

Stat 131A, Fall 2018, Prof. Sanchez

**Learning Objectives**

- Compute the mean
- Become familiar with the function `mean()`
- Interpret the mean as the balancing point

## Intro to Descriptive Statistics

As we mentioned in the previous script, the first part of the course has to do with **Descriptive Statistics**. The main idea is to make a "large" or "complicated" dataset more compact and easier to understand by using three major tools:

- summary and frequency tables
- charts and graphics
- key numeric summaries

In this script we will focus on various numeric summaries that are typically used to condense information of quantitative variables.

One common way to classify numeric summaries is in 1) measures of center, and 2) measures of spread or variability. The idea of both types of measures is to obtain one or more numeric values that reflect a "central" value, and the amount of "spread".

- Measures of Center
  - average or mean
  - median
- Measures of Spread
  - range
  - interquartile range
  - standard deviation (and variance)

## The Average

Perhaps the most common type of measure of center is the average or mean. Consider a list of numbers formed by: 0, 1, 2, 3, 5, and 7. The average is calculated as the sum of all values divided the number of values:

$$average = \frac{0 + 1 + 2 + 3 + 5 + 7}{6} = 3$$

You can use R to compute the previous average:

```
(0 + 1 + 2 + 3 + 5 + 7) / 6
```

```
## [1] 3
```

Algebraically, we typically denote a set of values by $x_1, x_2, \dots, x_n$, in which the index $n$ represents the total number of values. Using this notation, the formula of the average is expressed as:

$$average = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Using summation notation, the average can be compactly expressed as:

$$average = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Summation notation uses the uppercase Greek letter $\Sigma$ (sigma), is used as an abbreviation for the phrase "the sum of". So, in place of $x_1 + x_2 + \cdots + x_n$, we can use summation notation as "the sum of the observations of the variable $x$."

In R, you can create a vector `x` to store the previous numbers:

```
x = c(0, 1, 2, 3, 5, 7)
```

Then, you can use the function `sum()` to add all the values in `x`, and compute the average as:

```
sum(x) / length(x)
```

```
## [1] 3
```

An alternative way to compute the average in R is using the `mean()` function:

```
mean(x)
```

```
## [1] 3
```

### The Average is the balancing point

Usually, the average of a set of $n$ values $x_1, x_2, \dots, x_n$ is expressed as $\bar{x}$ (pronounced *x-bar*).

To understand how the average is a type of central or mid-value, we need to talk about **deviations**. A deviation is the difference between an observed value $x_i$ and another value of reference $ref$, that is, $(x_i - ref)$.

Taking the average value $\bar{x}$ as a reference value, we can calculate the deviations of all observations from the average: $(x_i - \bar{x})$

Given a reference value $ref$, we can also compute the sum of all deviations around such value:

$$\sum_{i=1}^{n} (x_i - ref)$$

It turns out that the average is the ONLY reference value such that the sum of deviations around it becomes zero:

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

Let's verify that in R

```
avg = mean(x)
deviations = x - avg
deviations
```

```
## [1] -3 -2 -1  0  2  4
```

The sum of the deviations around the mean should be zero:

```
sum(deviations)
```

```
## [1] 0
```

This is the reason why we say that the average is one type of center or mid-value. In simpler terms, you can think of the average as the balance point of a distribution. The average is that point that cancels out the sum of deviations around it.

### Your turn

We know that the average of `x` is 3. What happens to this average if:

- you add a constant $b$ to all values in `x`?
- you multiply the values in `x` times a constant $a$?

For instance, let's add 2 to all vaues in `x`?

```
mean(x + 2)
```

```
## [1] 5
```

Now, let's multiply by 2 all values in `x`:

```
mean(x * 2)
```

```
## [1] 6
```

Spend some time in R to examine what happens to the average of $x + k$ and $k \times x$ with several choices of $k$, e.g. -2, 5, 100.

Now, let's see what happens to the average when you add a constant $b$ to all values in `x`, and multiply them times some constant $a$?

```r
mean(x)
```

```
## [1] 3
```

```r
a = 2
b = 3
mean(a*x + b)
```

```
## [1] 9
```

Again, spend some time in R trying different values for `a` and `b`. What's your conclusion?

## The Median

Another common type of measure of center is the **median**. The median is the literal middle value of an ordered distribution. By *middle value* we mean that half of observations are below the median, and the other half of observations are above it.

The easiest way to calculate the median in R is with the homonym function `median()`. Consider again the numbers in the vector `x`, the median of this set of values is:

```r
x = c(0, 1, 2, 3, 5, 7)

median(x)
```

```
## [1] 2.5
```

The median depends on the number of values. If you have a variable with an even number of values, then the median is the average of the two middle-values. If you have a variable with an odd number of values, then the median is the middle-value.

## More numeric summaries

Another interesting function in R that you can use to obtain descriptive information about a variable is `summary()`. When you use this function on a numeric vector (i.e. quantitative variable), the returned output includes:

- `Min.`: minimum
- `1st Qu.`: first quartile
- `Median`: median
- `Mean`: average
- `3rd Qu.`: third quartile
- `Max.`: maximum

```
x = c(0, 1, 2, 3, 5, 7)

summary(x)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    1.25    2.50    3.00    4.50    7.00
```

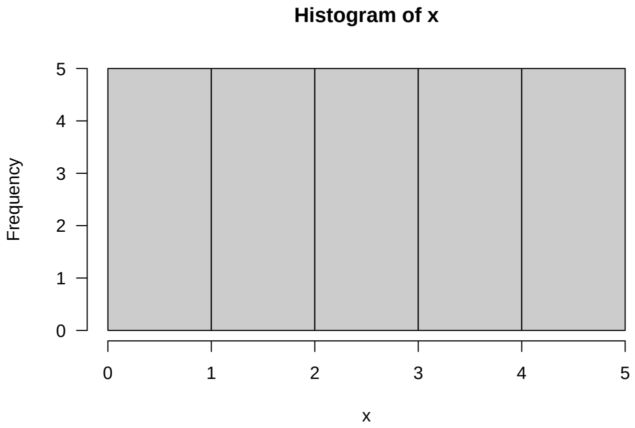
## Average -vs- Median

Consider a new vector `x` that contains 25 numbers: five 1's, five 2's, five 3's, five 4's, and five 5's:

```
x = rep(1:5, each = 5)
```

As you can tell, all values in `x` occur with the same frequency. And if you get a histogram, R will plot all bars with the same height:

```
hist(x, breaks = c(0, 1, 2, 3, 4, 5), las = 1, col = 'gray80')
```



In this data, the average and the median are the same. In fact, this happens all the time you have a perfect symmetric distribution:

```
mean(x)
```

```
## [1] 3
```

```
median(x)
```

```
## [1] 3
```

Now let's add one more observation to `x` with a value of 10, and obtain the average and the median:

```
y = c(x, 10)
mean(y)
```

```
## [1] 3.269231
```

```
median(y)
```

```
## [1] 3
```

Note that the average increased from 3 to 3.27, while the median remained unchanged.

Let's make it more extreme and instead of adding a value of 10 let's add a value of 100 to `x`. The average and the median are:

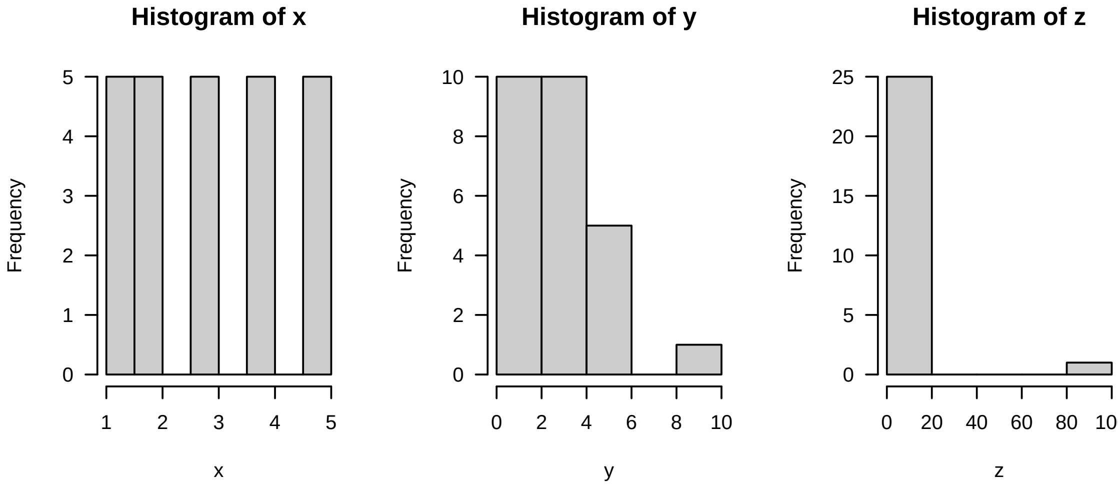
```
z = c(x, 100)
mean(z)
```

```
## [1] 6.730769
```

```
median(z)
```

```
## [1] 3
```

You can look at the distributions of `x`, `y`, and `z` using the default plots produced by `hist()`:



This is a toy example that illustrates one difference between the median and the average. The median is more resistant (or robust) to extreme values, but not the average. Small and large values affect the average of a distribution.

### Example

Here's one more example that shows you how to use R to solve a typical textbook exercise. The average and median of the first 99 values of a data set of 198 values are all equal to 120. If the average and median of the final 99 values are all equal to 100, what can you say about the average of the entire data set. What can you say about the median?

You can solve theis type of questions analytically, or you can use R. Here's how. The problem deals with a data set of 198 values formed by two sets of numbers: the first 99 values are all equal to 120, the final 99 values are all equal to 100. You can create two R vectors to build the two sets of 99 values. This is achieved with the function `rep()` that allows you to **repeat** one or more numeric values given a number of times:

```
# first 99 values equal to 120
first_values = rep(120, times = 99)

# final 99 values equal to 100
final_values = rep(100, times = 99)

# all values
all_values = c(first_values, final_values)
```

Having defined `first_values` and `final_values`, we build the entire list of 198 values by combining them in the vector `all_values`. The next step involves finding the average and the median:

```
# average
mean(all_values)
```

```
## [1] 110
```

```
# median
median(all_values)
```

```
## [1] 110
```