# IMD0033 - Probabilidade
## Aula 19 - Medidas de Variabilidade

Ivanovitch Silva
Maio, 2019

# Agenda

- Faixa
- Distância Média
- Desvio médio absoluto
- Variância e desvio padrão
- Desvio padrão da amostra
- Correção de Bessel

# Atualizar o repositório

```
git clone https://github.com/ivanovitchm/imd0033_2019_1.git
```

Ou ....

```
git pull
```

# The range

So far we've focused entirely on summarizing distributions using the **mean**, the **weighted mean**, the **median**, and the **mode**. An interesting distribution property we haven't yet discussed is **variability**.

The values of the distribution A don't vary

$$A = [4, 4, 4, 4]$$

$$B = [0, 8, 0, 8]$$

What variability value should we assign to distribution B?

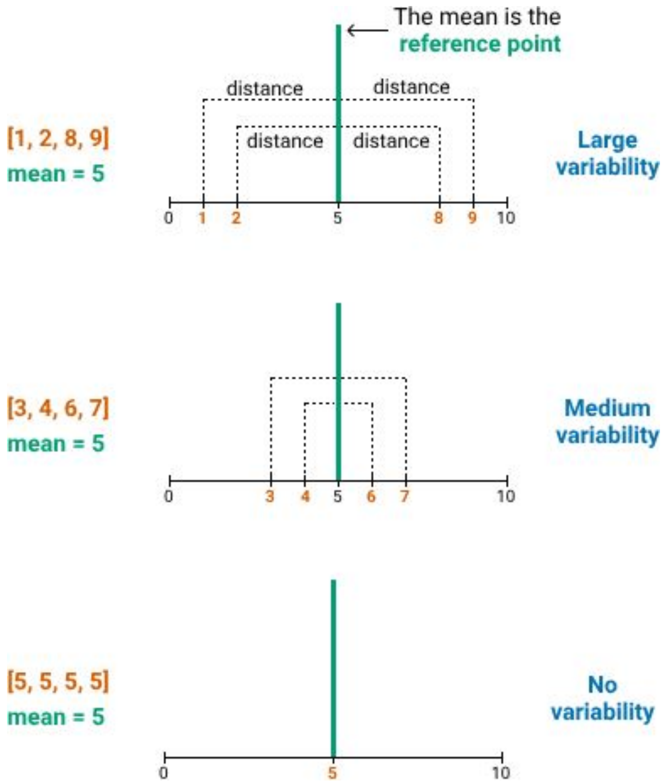
# The range

$$A = [4, 4, 4, 4] \qquad max(A) - min(A) = 4 - 4 = 0$$

$$B = [0, 8, 0, 8] \qquad max(B) - min(B) = 8 - 0 = 8$$

$$range(X) = max(X) - min(X)$$

# The average distance




The problem with the range is that it considers only two values in the distribution — the minimum and the maximum value.

$$C = [1, 1, 1, 1, 1, 1, 1, 1, 1, 21]$$

$$max(C) - min(C) = 21 - 1 = 20$$

$$average\ distance = \frac{\overbrace{(x_1 - \mu)}^{distance} + \overbrace{(x_2 - \mu)}^{distance} + \ldots + \overbrace{(x_N - \mu)}^{distance}}{N} = \frac{\overbrace{\sum_{i=1}^{n} \overbrace{(x_i - \mu)}^{distance}}^{total\ distance}}{N}$$

# Mean absolute deviation

Values that are **below** the mean

| $x_i - \mu$ | Distance |
|---|---|
| 1 - 3 | - 2 |
| 1 - 3 | - 2 |
| 1 - 3 | - 2 |
| 1 - 3 | - 2 |
| 1 - 3 | - 2 |
| 1 - 3 | - 2 |
| 1 - 3 | - 2 |
| 1 - 3 | - 2 |
| 1 - 3 | - 2 |

*Total = - 18*

Values that are **above** the mean

| $x_i - \mu$ | Distance |
|---|---|
| 21 - 3 | + 18 |

*Total = + 18*

$$\text{mean absolute distance} = \frac{|x_1 - \mu| + |x_2 - \mu| + \ldots + |x_N - \mu|}{N} = \frac{\sum_{i=1}^{N} |x_i - \mu|}{N}$$

$$C = [1, 1, 1, 1, 1, 1, 1, 1, 1, 21]$$

$$\text{average distance} = \frac{-18 + 18}{10} = \frac{0}{10}$$

# Variance and Standard Deviation

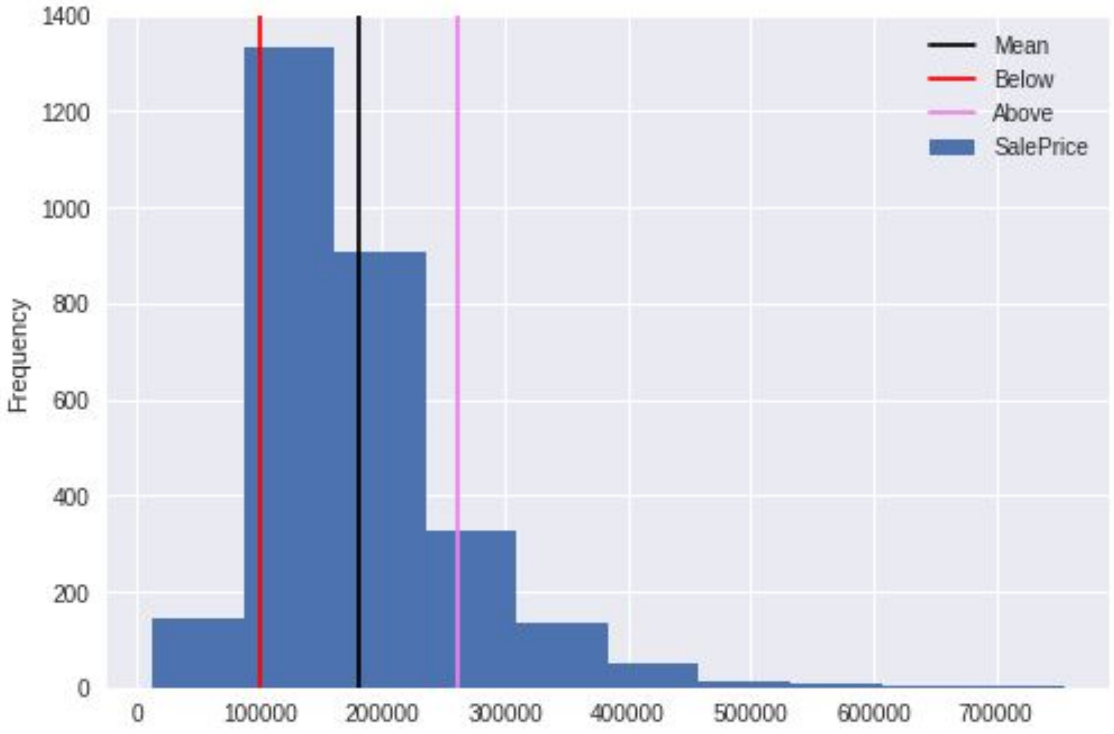
$$\text{mean squared distance} = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \ldots + (x_N - \mu)^2}{N} = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

$$\text{standard deviation} = \sqrt{\frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \ldots + (x_N - \mu)^2}{N}} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

# Average variability around the mean

Applied to Computer Simulation

Confidence Interval



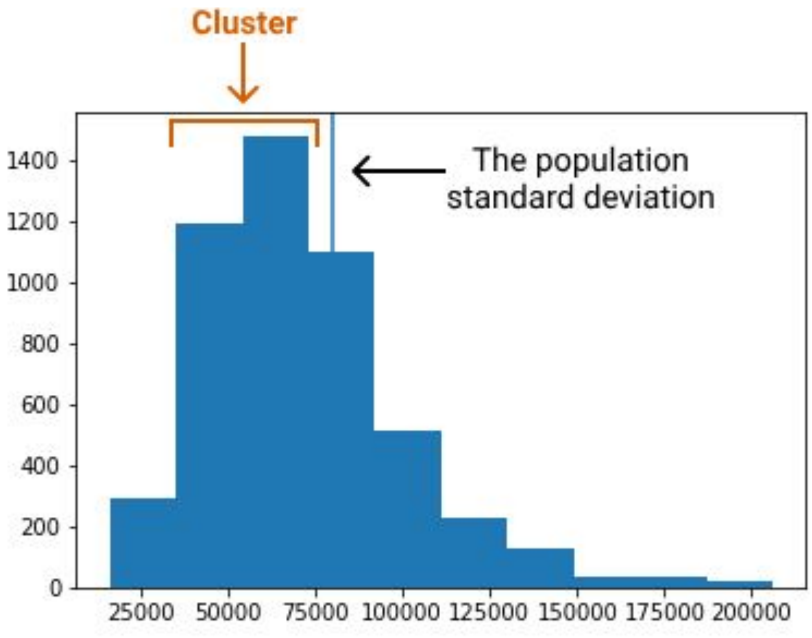
# The sample standard deviation

- Population vs Sample
  - In practice, we generally work with samples
  - But most of the time we're not actually interested in describing the samples
- Rather, we want to use the samples to make inferences about their corresponding populations.
- **Let's find out whether the standard deviation of a sample is a good estimate for the standard deviation in the corresponding population.**

# The sample standard deviation

$$SD = \sqrt{\frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \ldots + (x_N - \mu)^2}{N}} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

$$SD_{sample} = \sqrt{\frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}}$$




Notice that most sample standard deviations are clustered below the population standard deviation

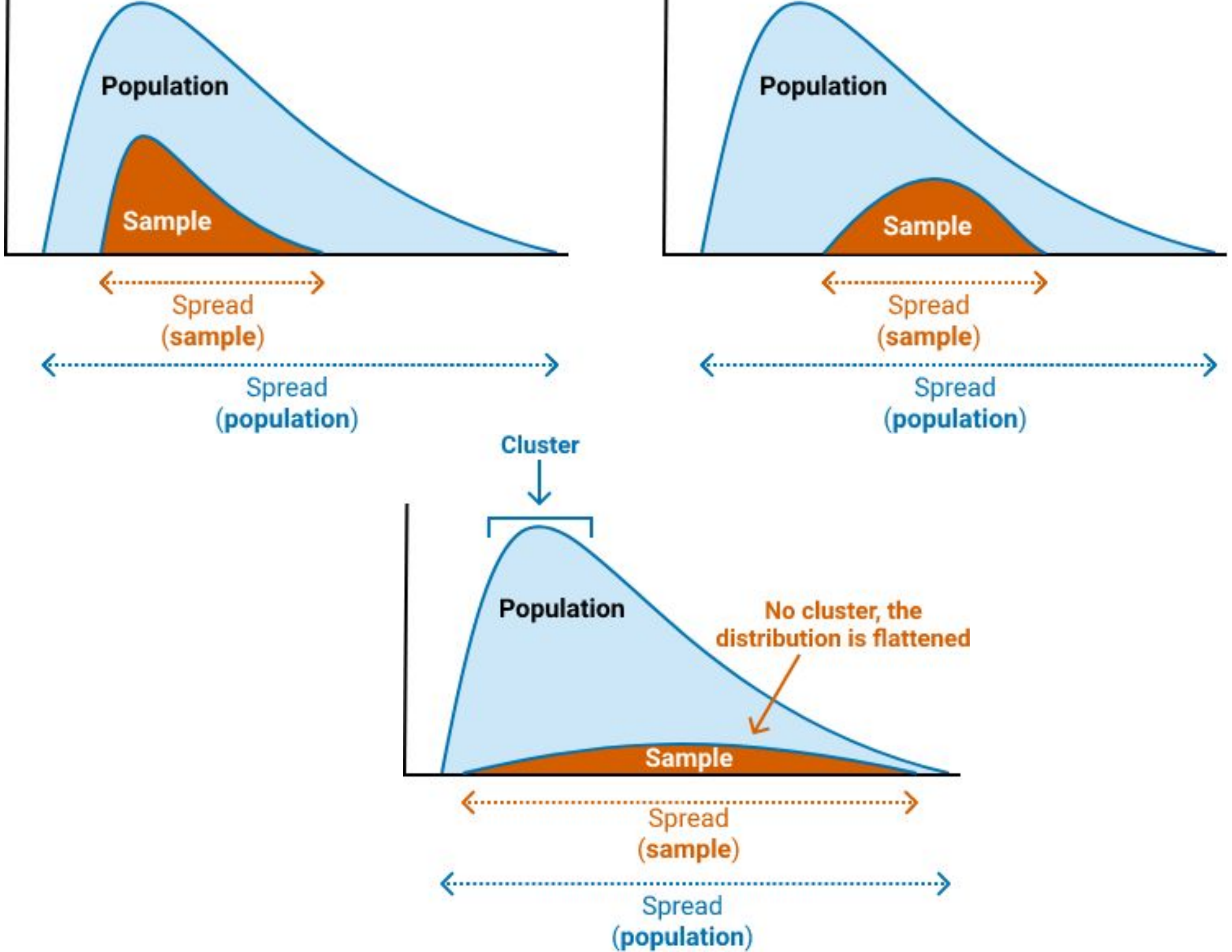
This suggests that the sample standard deviation usually underestimates the population standard deviation.

```
for i in range(5000):
    sample = houses['SalePrice'].sample(10, random_state = i)
    st_dev = standard_deviation(sample)
    st_devs.append(st_dev)

plt.hist(st_devs)
plt.axvline(standard_deviation(houses['SalePrice']))
```

$$\bar{x}_{SD} = 71303$$

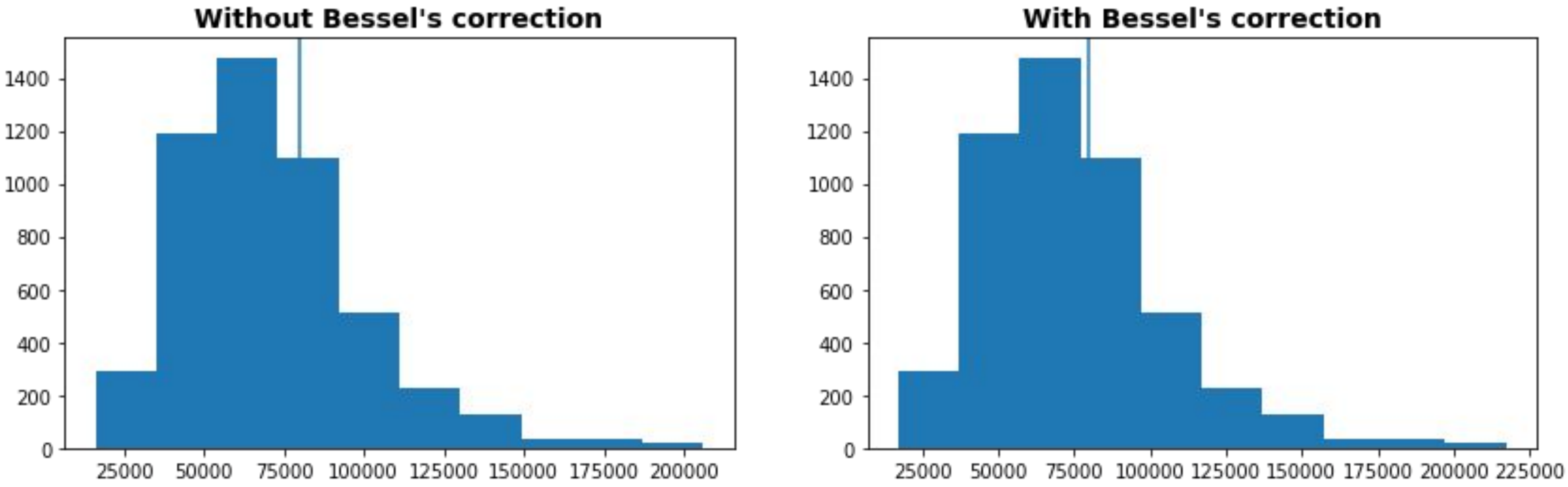
$$\mu_{SD} = 79873$$


Population
Sample
Spread
(sample)
Spread
(population)
Population
Sample
Spread
(sample)
Spread
(population)
Cluster
Population
No cluster, the
distribution is flattened
Sample
Spread
(sample)
Spread
(population)

# Bessel's Correction

$$SD_{sample} = \sqrt{\frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$$

# Bessel's Correction

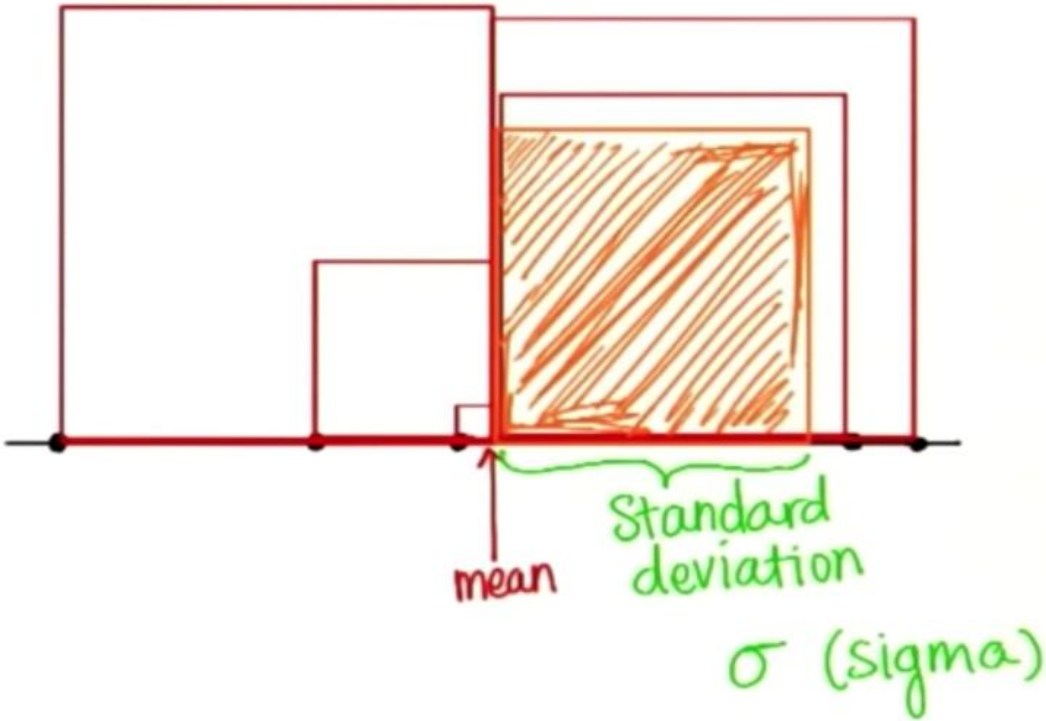


# Standard Notation

$$\sigma = \sqrt{\frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \ldots + (x_N - \mu)^2}{N}} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

$$\sigma^2 = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \ldots + (x_N - \mu)^2}{N} = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

# Standard Notation



# Standard Notation

$$s = \sqrt{\frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \ldots + (x_n - \mu)^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n-1}}$$

$$s^2 = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \ldots + (x_n - \mu)^2}{n-1} = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n-1}$$


MacBook Pro