1 - Amostragem

$\theta =$ parâmetro, o valor real da característica na população. μ (média), σ^2 (variância) e ρ (coeficiente de correlação) são parâmetros.

$\hat{\theta} =$ estimador, estatística de parâmetro. Função dos elementos da amostra. É uma variável aleatória (pois os valores são diferentes em cada amostra). $\overline{X}$ (média), S^2 (variância) e r (coeficiente de correlação) são estimadores.

$\hat{\theta}_0 =$ estimativa, o valor obtido pelo estimador.

$\epsilon = \theta - \hat{\theta}$, o erro amostral. Soma da parte casual com a parte viés/desvio, em que casual é a diferença entre a estimativa e a esperança do estimador ($\hat{\theta} - E(\hat{\theta})$); e viés é a esperança menos o parâmetro na população ($E(\hat{\theta}) - \theta$).

Como o erro casual é natural, se tira que sempre há uma esperança diferente da estimativa. O viés é a esperança não estar alinhada ao parâmetro, mas se supõe que isto possa ser eliminado.

$E(\hat{\theta}) =$ "esperança matemática" da variável aleatória.

Viés de seleção: algum elemento ter chance de não pertencer a uma amostra ("eliminação"). Se houver *alguma* chance para todos os elementos (probabilística), não há viés de seleção.

"The expected value of a random variable, intuitively, is the long-run average value of repetitions of the same experiment it represents. (...) For example, the expected value in rolling a six-sided die is 3.5, because the average of all the numbers that come up is 3.5 as the number of rolls approaches infinity. In other words, the law of large numbers states that the arithmetic mean of the values almost surely converges to the expected value as the number of repetitions approaches infinity. (...) More practically, the expected value of a discrete random variable is the probability-weighted average of all possible values. In other words, each possible value the random variable can assume is multiplied by its probability of occurring, and the resulting products are summed to produce the expected value. The same principle applies to an absolutely continuous random variable, except that an integral of the variable with respect to its probability density replaces the sum. (...) The expected value is a key aspect of how one characterizes a probability distribution; it is one type of location parameter. By contrast, the variance is a measure of dispersion of the possible values of the random variable around the expected value. The variance itself is defined in terms of two expectations: it is the expected value of the squared deviation of the variable's value from the variable's expected value."[1]

"Let X be a random variable with a finite number of finite outcomes $x_1, x_2, ..., x_k$ occurring with probabilities $p_1, p_2, ..., p_k$. The expectation of X is defined as $E(X) = \sum_{i=1}^{k} x_i\, p_i$. Since all probabilities add up to 1, the expected value is the *weighted average*, with the probabilities being the weights."

"The weighted arithmetic mean is similar to an ordinary arithmetic mean (the most common type of average), except that instead of each of the data points contributing equally to the final average, some data points contribute more than others. If all the weights are equal, then the weighted mean is the same as the arithmetic mean."[2]

If all outcomes are equiprobable, then the weighted average turns into the simple average. This is intuitive: the expected value of a random variable is the average of all values it can take; thus the expected value is what one expects to happen on average.
If the outcomes are not equiprobable, then the simple average must be replaced with the weighted average, which takes into account the fact that some outcomes are more likely than the others. The intuition however remains the same: the expected value of X is what one expects to happen on average."[3]

Ou seja, estamos lidando com a *distribuição* da *variável aleatória*.
O que significa a distribuição dos valores que ela gera em cada execução, o que é um valor para cada amostra.

2 - Análise exploratória dos dados de uma amostra

Não conheçamos a média e variância de uma população X. Retiramos uma amostra de n elementos e estimamos estes parâmetros.

Estimador média $\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$, estimador variância $S^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}$.

Estimador média não viciado/enviesado pois $E(\overline{x}) = \mu$. (A esperança da média amostral é igual à média populacional.)

Estimador variância não viciado/enviesado pois $E(s^2) = \sigma^2$. (A esperança da variância amostral é igual à variância populacional.)

O estimador de variância $S^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n}$ é viciado/enviesado: $E(s^2) \neq \sigma^2$. (A esperança da variância amostral é diferente da variância populacional.)

In[1]:=
```
Clear[log]
log=SemanticImport[
    NotebookDirectory[]<>"LogTest1.txt",
    <|"PIM"→Automatic|>,
    "NamedColumns",HeaderLines→1,Delimiters→";"];
```

In[∘]:=
```
Length[log["PIM"]]
```

Out[∘]= 29 225

In[∘]:= $N\left[\left\{\frac{\sum_{i=1}^{Length[log["PIM"]]} log["PIM"][[i]]}{Length[log["PIM"]]}, Mean[log["PIM"]]\right\}\right]$

Out[∘]= {3195.73, 3195.73}

In[∘]:= $N\left[\left\{\frac{\sum_{i=1}^{Length[log["PIM"]]} (log["PIM"][[i]] - Mean[log["PIM"]])^2}{Length[log["PIM"]] - 1}, Variance[log["PIM"]]\right\}\right]$

Out[∘]= $\{2.27091 \times 10^7, 2.27091 \times 10^7\}$

In[∘]:= $N\left[\frac{\sum_{i=1}^{Length[log["PIM"]]} (log["PIM"][[i]] - Mean[log["PIM"]])^2}{Length[log["PIM"]]}\right]$

Out[∘]= 2.27084×10^7

Viciado valor um pouco diferente.

In[3]:=
```
Clear[data1]
data1={5,9,2,8.8,6.3,10.1,3.4,2.9,4.1};
```

Se o viés é a esperança ser diferente do parâmetro, a esperança vem da própria amostra (e não da população). O cálculo da esperança é a multiplicação... temos algumas variáveis diferentes:

- A média da amostra
- O valor dos elementos
- A probabilidade dos elementos

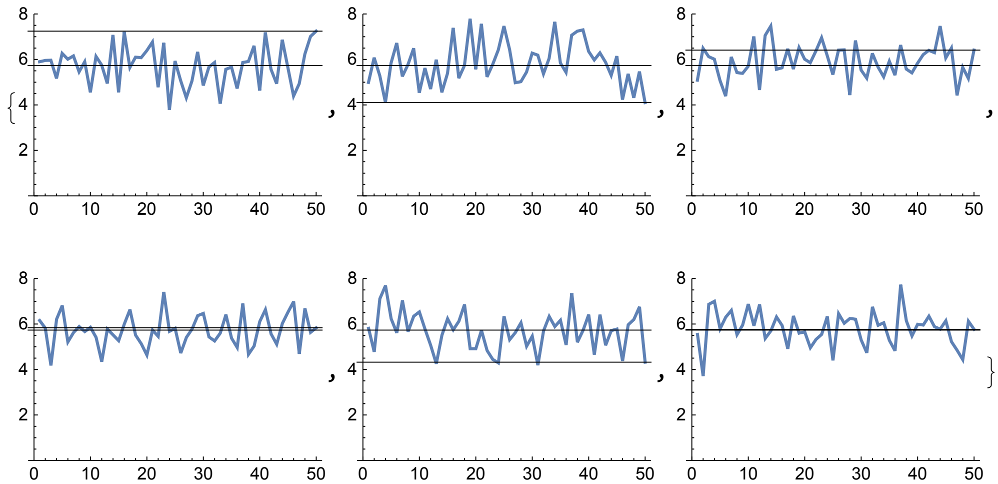
Apenas algumas amostras.

In[∘]:=
```
Module[{samples, means},
 Print[Mean[data1]];
 Print[samples = Table[Flatten[Subsets[RandomSample[data1], {3}, 1]], 3]];
 Print[means = Table[N[Mean[sample]], {sample, samples}]];
 Print[Mean[means]];
]
```

```
5.73333
{{4.1, 6.3, 5}, {5, 3.4, 2.9}, {10.1, 4.1, 3.4}}
{5.13333, 3.76667, 5.86667}
4.92222
```

In[ ]:=
```
Module[{samples, means, meansMean},
 Table[
  meansMean = Table[
    samples = Table[Flatten[Subsets[RandomSample[data1], {3}, 1]], 3];
    means = Table[N[Mean[sample]], {sample, samples}];
    Mean[means]
, 50];
  ListLinePlot[meansMean, ImageSize → 160, PlotRange → {0, 8},
   Epilog → {, {InfiniteLine[{{0, Mean[data1]}, {2, Mean[data1]}}],
      InfiniteLine[{{0, Mean[means]}, {2, Mean[means]}}]}}]
  , 6]
]
```

Out[ ]=



Média das médias diferente, mas sempre variando em uma faixa.

Todas as amostras.

```
Module[{samples, means},
 Table[
  Print[Mean[data1]];
  Print[samples = Subsets[RandomSample[data1], {3}]];
  Print[means = Table[N[Mean[sample]], {sample, samples}]];
  Print[Mean[means]];
  ListLinePlot[means,
   Epilog → {, InfiniteLine[{{0, Mean[data1]}, {2, Mean[data1]}}]}, ImageSize → 160]
  , 3]
]
```

```
5.73333
```

```
{{8.8, 4.1, 6.3}, {8.8, 4.1, 2.9}, {8.8, 4.1, 2}, {8.8, 4.1, 3.4}, {8.8, 4.1, 10.1}, {8.8, 4.1, 5},
 {8.8, 4.1, 9}, {8.8, 6.3, 2.9}, {8.8, 6.3, 2}, {8.8, 6.3, 3.4}, {8.8, 6.3, 10.1}, {8.8, 6.3, 5},
 {8.8, 6.3, 9}, {8.8, 2.9, 2}, {8.8, 2.9, 3.4}, {8.8, 2.9, 10.1}, {8.8, 2.9, 5}, {8.8, 2.9, 9},
 {8.8, 2, 3.4}, {8.8, 2, 10.1}, {8.8, 2, 5}, {8.8, 2, 9}, {8.8, 3.4, 10.1}, {8.8, 3.4, 5},
 {8.8, 3.4, 9}, {8.8, 10.1, 5}, {8.8, 10.1, 9}, {8.8, 5, 9}, {4.1, 6.3, 2.9}, {4.1, 6.3, 2},
 {4.1, 6.3, 3.4}, {4.1, 6.3, 10.1}, {4.1, 6.3, 5}, {4.1, 6.3, 9}, {4.1, 2.9, 2}, {4.1, 2.9, 3.4},
 {4.1, 2.9, 10.1}, {4.1, 2.9, 5}, {4.1, 2.9, 9}, {4.1, 2, 3.4}, {4.1, 2, 10.1}, {4.1, 2, 5},
 {4.1, 2, 9}, {4.1, 3.4, 10.1}, {4.1, 3.4, 5}, {4.1, 3.4, 9}, {4.1, 10.1, 5}, {4.1, 10.1, 9},
 {4.1, 5, 9}, {6.3, 2.9, 2}, {6.3, 2.9, 3.4}, {6.3, 2.9, 10.1}, {6.3, 2.9, 5}, {6.3, 2.9, 9},
 {6.3, 2, 3.4}, {6.3, 2, 10.1}, {6.3, 2, 5}, {6.3, 2, 9}, {6.3, 3.4, 10.1}, {6.3, 3.4, 5},
 {6.3, 3.4, 9}, {6.3, 10.1, 5}, {6.3, 10.1, 9}, {6.3, 5, 9}, {2.9, 2, 3.4}, {2.9, 2, 10.1},
 {2.9, 2, 5}, {2.9, 2, 9}, {2.9, 3.4, 10.1}, {2.9, 3.4, 5}, {2.9, 3.4, 9}, {2.9, 10.1, 5},
 {2.9, 10.1, 9}, {2.9, 5, 9}, {2, 3.4, 10.1}, {2, 3.4, 5}, {2, 3.4, 9}, {2, 10.1, 5},
 {2, 10.1, 9}, {2, 5, 9}, {3.4, 10.1, 5}, {3.4, 10.1, 9}, {3.4, 5, 9}, {10.1, 5, 9}}
```

```
{6.4, 5.26667, 4.96667, 5.43333, 7.66667, 5.96667, 7.3, 6., 5.7, 6.16667, 8.4, 6.7, 8.03333,
 4.56667, 5.03333, 7.26667, 5.56667, 6.9, 4.73333, 6.96667, 5.26667, 6.6, 7.43333, 5.73333,
 7.06667, 7.96667, 9.3, 7.6, 4.43333, 4.13333, 4.6, 6.83333, 5.13333, 6.46667, 3., 3.46667,
 5.7, 4., 5.33333, 3.16667, 5.4, 3.7, 5.03333, 5.86667, 4.16667, 5.5, 6.4, 7.73333, 6.03333,
 3.73333, 4.2, 6.43333, 4.73333, 6.06667, 3.9, 6.13333, 4.43333, 5.76667, 6.6, 4.9, 6.23333,
 7.13333, 8.46667, 6.76667, 2.76667, 5., 3.3, 4.63333, 5.46667, 3.76667, 5.1, 6., 7.33333,
 5.63333, 5.16667, 3.46667, 4.8, 5.7, 7.03333, 5.33333, 6.16667, 7.5, 5.8, 8.03333}
```

5.73333

5.73333

```
{{9, 4.1, 8.8}, {9, 4.1, 5}, {9, 4.1, 2.9}, {9, 4.1, 6.3}, {9, 4.1, 10.1}, {9, 4.1, 3.4},
 {9, 4.1, 2}, {9, 8.8, 5}, {9, 8.8, 2.9}, {9, 8.8, 6.3}, {9, 8.8, 10.1}, {9, 8.8, 3.4},
 {9, 8.8, 2}, {9, 5, 2.9}, {9, 5, 6.3}, {9, 5, 10.1}, {9, 5, 3.4}, {9, 5, 2}, {9, 2.9, 6.3},
 {9, 2.9, 10.1}, {9, 2.9, 3.4}, {9, 2.9, 2}, {9, 6.3, 10.1}, {9, 6.3, 3.4}, {9, 6.3, 2},
 {9, 10.1, 3.4}, {9, 10.1, 2}, {9, 3.4, 2}, {4.1, 8.8, 5}, {4.1, 8.8, 2.9}, {4.1, 8.8, 6.3},
 {4.1, 8.8, 10.1}, {4.1, 8.8, 3.4}, {4.1, 8.8, 2}, {4.1, 5, 2.9}, {4.1, 5, 6.3}, {4.1, 5, 10.1},
 {4.1, 5, 3.4}, {4.1, 5, 2}, {4.1, 2.9, 6.3}, {4.1, 2.9, 10.1}, {4.1, 2.9, 3.4}, {4.1, 2.9, 2},
 {4.1, 6.3, 10.1}, {4.1, 6.3, 3.4}, {4.1, 6.3, 2}, {4.1, 10.1, 3.4}, {4.1, 10.1, 2}, {4.1, 3.4, 2},
 {8.8, 5, 2.9}, {8.8, 5, 6.3}, {8.8, 5, 10.1}, {8.8, 5, 3.4}, {8.8, 5, 2}, {8.8, 2.9, 6.3},
 {8.8, 2.9, 10.1}, {8.8, 2.9, 3.4}, {8.8, 2.9, 2}, {8.8, 6.3, 10.1}, {8.8, 6.3, 3.4},
 {8.8, 6.3, 2}, {8.8, 10.1, 3.4}, {8.8, 10.1, 2}, {8.8, 3.4, 2}, {5, 2.9, 6.3}, {5, 2.9, 10.1},
 {5, 2.9, 3.4}, {5, 2.9, 2}, {5, 6.3, 10.1}, {5, 6.3, 3.4}, {5, 6.3, 2}, {5, 10.1, 3.4},
 {5, 10.1, 2}, {5, 3.4, 2}, {2.9, 6.3, 10.1}, {2.9, 6.3, 3.4}, {2.9, 6.3, 2}, {2.9, 10.1, 3.4},
 {2.9, 10.1, 2}, {2.9, 3.4, 2}, {6.3, 10.1, 3.4}, {6.3, 10.1, 2}, {6.3, 3.4, 2}, {10.1, 3.4, 2}}
```

```
{7.3, 6.03333, 5.33333, 6.46667, 7.73333, 5.5, 5.03333, 7.6, 6.9, 8.03333, 9.3, 7.06667, 6.6,
 5.63333, 6.76667, 8.03333, 5.8, 5.33333, 6.06667, 7.33333, 5.1, 4.63333, 8.46667, 6.23333,
 5.76667, 7.5, 7.03333, 4.8, 5.96667, 5.26667, 6.4, 7.66667, 5.43333, 4.96667, 4., 5.13333,
 6.4, 4.16667, 3.7, 4.43333, 5.7, 3.46667, 3., 6.83333, 4.6, 4.13333, 5.86667, 5.4, 3.16667,
 5.56667, 6.7, 7.96667, 5.73333, 5.26667, 6., 7.26667, 5.03333, 4.56667, 8.4, 6.16667, 5.7,
 7.43333, 6.96667, 4.73333, 4.73333, 6., 3.76667, 3.3, 7.13333, 4.9, 4.43333, 6.16667,
 5.7, 3.46667, 6.43333, 4.2, 3.73333, 5.46667, 5., 2.76667, 6.6, 6.13333, 3.9, 5.16667}
```

5.73333

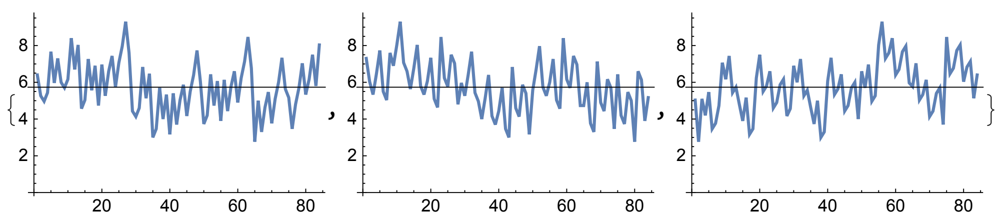
5.73333

```
{{3.4, 2.9, 8.8}, {3.4, 2.9, 2}, {3.4, 2.9, 9}, {3.4, 2.9, 6.3}, {3.4, 2.9, 10.1}, {3.4, 2.9, 4.1},
 {3.4, 2.9, 5}, {3.4, 8.8, 2}, {3.4, 8.8, 9}, {3.4, 8.8, 6.3}, {3.4, 8.8, 10.1}, {3.4, 8.8, 4.1},
 {3.4, 8.8, 5}, {3.4, 2, 9}, {3.4, 2, 6.3}, {3.4, 2, 10.1}, {3.4, 2, 4.1}, {3.4, 2, 5},
 {3.4, 9, 6.3}, {3.4, 9, 10.1}, {3.4, 9, 4.1}, {3.4, 9, 5}, {3.4, 6.3, 10.1}, {3.4, 6.3, 4.1},
 {3.4, 6.3, 5}, {3.4, 10.1, 4.1}, {3.4, 10.1, 5}, {3.4, 4.1, 5}, {2.9, 8.8, 2}, {2.9, 8.8, 9},
 {2.9, 8.8, 6.3}, {2.9, 8.8, 10.1}, {2.9, 8.8, 4.1}, {2.9, 8.8, 5}, {2.9, 2, 9}, {2.9, 2, 6.3},
 {2.9, 2, 10.1}, {2.9, 2, 4.1}, {2.9, 2, 5}, {2.9, 9, 6.3}, {2.9, 9, 10.1}, {2.9, 9, 4.1},
 {2.9, 9, 5}, {2.9, 6.3, 10.1}, {2.9, 6.3, 4.1}, {2.9, 6.3, 5}, {2.9, 10.1, 4.1}, {2.9, 10.1, 5},
 {2.9, 4.1, 5}, {8.8, 2, 9}, {8.8, 2, 6.3}, {8.8, 2, 10.1}, {8.8, 2, 4.1}, {8.8, 2, 5},
 {8.8, 9, 6.3}, {8.8, 9, 10.1}, {8.8, 9, 4.1}, {8.8, 9, 5}, {8.8, 6.3, 10.1}, {8.8, 6.3, 4.1},
 {8.8, 6.3, 5}, {8.8, 10.1, 4.1}, {8.8, 10.1, 5}, {8.8, 4.1, 5}, {2, 9, 6.3}, {2, 9, 10.1},
 {2, 9, 4.1}, {2, 9, 5}, {2, 6.3, 10.1}, {2, 6.3, 4.1}, {2, 6.3, 5}, {2, 10.1, 4.1},
 {2, 10.1, 5}, {2, 4.1, 5}, {9, 6.3, 10.1}, {9, 6.3, 4.1}, {9, 6.3, 5}, {9, 10.1, 4.1},
 {9, 10.1, 5}, {9, 4.1, 5}, {6.3, 10.1, 4.1}, {6.3, 10.1, 5}, {6.3, 4.1, 5}, {10.1, 4.1, 5}}
```

```
{5.03333, 2.76667, 5.1, 4.2, 5.46667, 3.46667, 3.76667, 4.73333, 7.06667, 6.16667, 7.43333,
 5.43333, 5.73333, 4.8, 3.9, 5.16667, 3.16667, 3.46667, 6.23333, 7.5, 5.5, 5.8, 6.6, 4.6,
 4.9, 5.86667, 6.16667, 4.16667, 4.56667, 6.9, 6., 7.26667, 5.26667, 5.56667, 4.63333,
 3.73333, 5., 3., 3.3, 6.06667, 7.33333, 5.33333, 5.63333, 6.43333, 4.43333, 4.73333, 5.7,
 6., 4., 6.6, 5.7, 6.96667, 4.96667, 5.26667, 8.03333, 9.3, 7.3, 7.6, 8.4, 6.4, 6.7, 7.66667,
 7.96667, 5.96667, 5.76667, 7.03333, 5.03333, 5.33333, 6.13333, 4.13333, 4.43333, 5.4, 5.7,
 3.7, 8.46667, 6.46667, 6.76667, 7.73333, 8.03333, 6.03333, 6.83333, 7.13333, 5.13333, 6.4}
```

```
5.73333
```

Out[ ]=



As diferentes linhas só refletem as diferentes ordens das amostras. A média das médias das amostras entre TODAS as amostras de um tamanho é igual à média da população.

"The expectation value of a function $f(x)$ in a variable x is denoted $\langle f(x)\rangle$ or $E\{f(x)\}$. For a single discrete variable, it is defined by $\langle f(x)\rangle = \sum_x f(x)\, P(x)$, where $P(x)$ is the *probability density function*. For a single continuous variable it is defined by $\langle f(x)\rangle = \int f(x)\, P(x)\, dx$."[4]

Weighted average.

In[ ]:=
```
{data1, Length[data1]}
```

Out[ ]=
```
{{5, 9, 2, 8.8, 6.3, 10.1, 3.4, 2.9, 4.1}, 9}
```

In[ ]:=
$$\left\{\sum_{i=1}^{\text{Length[data1]}} \text{data1[[i]]},\ \sum_{i=1}^{\text{Length[data1]}} \text{data1[[i]]} * (1/\text{Length[data1]})\right\}$$

Out[ ]=
```
{51.6, 5.73333}
```

A média é a soma dos valores cada um multiplicado por seu “peso”.

Normalmente, o peso é $1/n$ igual para todos.

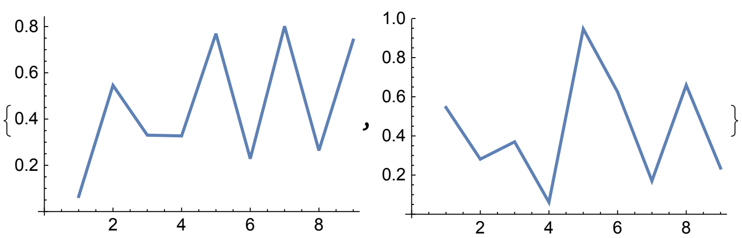
Quando há pesos diferentes de $1/n$, a média muda.

In[5]:=
```
Clear[data1Weights,data1Weights2]
data1Weights=Table[RandomReal[],Length[data1]]
data1Weights2=Table[RandomReal[],Length[data1]]
{ListLinePlot[data1Weights,ImageSize→Small],ListLinePlot[data1Weights2,ImageSize→Small]}
```

Out[6]= {0.0651111, 0.545815, 0.330387, 0.327167, 0.769338, 0.228513, 0.801367, 0.264045, 0.741207}

Out[7]= {0.545511, 0.281314, 0.369909, 0.0612241, 0.946807, 0.624723, 0.170142, 0.659397, 0.235823}

Out[8]=



In[ ]:=
```
{Total[data1Weights], Total[data1Weights2]}
```

Out[ ]= {4.77995, 3.46624}

In[ ]:= $\left\{\sum_{i=1}^{Length[data1]} data1[[i]] * data1Weights[[i]], \sum_{i=1}^{Length[data1]} data1[[i]] * data1Weights2[[i]]\right\}$

Out[ ]= {27.8658, 18.116}

E se for a soma dos pesos dividida pelos pesos?

In[ ]:= $\left\{\sum_{i=1}^{Length[data1]} data1[[i]] * (Total[data1Weights] / data1Weights[[i]]), \sum_{i=1}^{Length[data1]} data1[[i]] * (Total[data1Weights] / data1Weights2[[i]])\right\}$

Out[ ]= {632.608, 1430.01}

Nada a ver, porque a média aritmética é a soma dos valores multiplicados pelos seus pesos (que no caso de peso uniforme pode ser calculado como $1/n$); igualmente a média ponderada é a soma dos valores multiplicados pelos seus pesos, que por não serem uniformes não podem ser calculados, e são informados.

A expectativa da variável randômica é a média ponderada dos valores com as probabilidades como pesos.

Então com cada distribuição de probabilidades diferente, a expectativa é diferente.
Como a expectativa é uma função da distribuição, o objetivo (para alinhar ao parâmetro) é encontrar a mesma distribuição da variável na população.

A diferença da média ponderada para a esperança é que os pesos da média ponderada não somam 1.
Vamos criar um pesos como uma distribuição.

```
In[]:= data1
Table[RandomVariate[EmpiricalDistribution[data1]], 30]
Mean[EmpiricalDistribution[data1]]
Table[RandomVariate[EmpiricalDistribution[data1Weights → data1]], 30]
Mean[EmpiricalDistribution[data1Weights → data1]]

Out[]= {5, 9, 2, 8.8, 6.3, 10.1, 3.4, 2.9, 4.1}

Out[]= {6.3, 10.1, 10.1, 3.4, 2, 8.8, 10.1, 4.1, 6.3, 10.1, 3.4, 9, 2, 4.1,
 4.1, 2.9, 8.8, 4.1, 9, 10.1, 2.9, 9, 10.1, 2.9, 3.4, 6.3, 4.1, 6.3, 6.3, 2}

Out[]= 5.73333

Out[]= {6.3, 3.4, 3.4, 6.3, 2, 9, 9, 2.9, 2, 9, 4.1, 10.1, 10.1,
 9, 5, 3.4, 9, 3.4, 3.4, 10.1, 2, 5, 9, 9, 9, 3.4, 10.1, 2, 8.8, 2.9}

Out[]= 5.82973
```

Parece que os pesos foram divididos para somar 1.

```
In[]:= data1Weights
data1Weights / Total[data1Weights]
Total[data1Weights / Total[data1Weights]]

Out[]= {0.312963, 0.629985, 0.622442, 0.188865, 0.765437, 0.789048, 0.84873, 0.42102, 0.201459}

Out[]= {0.0654742, 0.131797, 0.130219, 0.0395118,
 0.160135, 0.165075, 0.177561, 0.0880805, 0.0421467}

Out[]= 1.
```

Neste caso, a média ponderada seria

```
In[]:= Sum[data1[[i]] * (data1Weights[[i]] / Total[data1Weights]), {i, 1, Length[data1]}]

Out[]= 5.82973
```

$$\sum_{i=1}^{\text{Length[data1]}} \text{data1[[i]]} * \left(\text{data1Weights[[i]]} / \text{Total[data1Weights]}\right)$$

☺ Exatamente.
Que deveria ser a expectativa.

```
In[]:= Expectation[x, x ≈ EmpiricalDistribution[data1Weights → data1]]

Out[]= 5.82973
```

Exato.
Então a média aritmética é só a expectativa de um conjunto equiprovável.
E então sempre que se fala na esperança, é apenas a média ponderada sobre as probabilidades que

somam 1.

Primeiro verificamos que a média das médias de todas as amostras (ou a *expectativa* do estimador da média) é igual à média da população, ou seja, o parâmetro.
Agora verificar se o mesmo ocorre para a variância. Agora, a esperança da variância **não** é a variância das variâncias das amostras, é a **média (**ponderada sobre as probabilidades?) das variâncias das amostras.

```
In[◦]:= Sum[data1[[i]] - Mean[data1], {i, 1, Length[data1]}] / (Length[data1] - 1)
```

$$\frac{\sum_{i=1}^{Length[data1]} data1[[i]] - Mean[data1]}{Length[data1] - 1}$$

```
Out[◦]= 5.73333
```

Tomando novamente só algumas e todas as amostras.

```
In[◦]:= Module[{samples, variances},
  Print["Alguns:"];
  Print[samples = Table[RandomSample[data1, 3], 3]];
  Print[
   variances = N[Table[(Sum_{i=1}^{Length[sample]} sample[[i]] - Mean[sample]) / (Length[sample] - 1), {sample, samples}]]];
  Print["Média aritmética das variâncias:"];
  Print[Mean[variances]];
  Print["Todos:"];
  Print[samples = Subsets[RandomSample[data1], {3}]];
  Print[
   variances = N[Table[(Sum_{i=1}^{Length[sample]} sample[[i]] - Mean[sample]) / (Length[sample] - 1), {sample, samples}]]];
  Print["Média aritmética das variâncias:"];
  Print[Mean[variances]];
 ];
```

```
Alguns:

{{9, 8.8, 3.4}, {2, 9, 5}, {10.1, 6.3, 3.4}}

{7.06667, 5.33333, 6.6}

Média aritmética das variâncias:

6.33333

Todos:

{{3.4, 6.3, 2}, {3.4, 6.3, 2.9}, {3.4, 6.3, 8.8}, {3.4, 6.3, 10.1}, {3.4, 6.3, 4.1}, {3.4, 6.3, 9},
 {3.4, 6.3, 5}, {3.4, 2, 2.9}, {3.4, 2, 8.8}, {3.4, 2, 10.1}, {3.4, 2, 4.1}, {3.4, 2, 9},
 {3.4, 2, 5}, {3.4, 2.9, 8.8}, {3.4, 2.9, 10.1}, {3.4, 2.9, 4.1}, {3.4, 2.9, 9}, {3.4, 2.9, 5},
 {3.4, 8.8, 10.1}, {3.4, 8.8, 4.1}, {3.4, 8.8, 9}, {3.4, 8.8, 5}, {3.4, 10.1, 4.1}, {3.4, 10.1, 9},
 {3.4, 10.1, 5}, {3.4, 4.1, 9}, {3.4, 4.1, 5}, {3.4, 9, 5}, {6.3, 2, 2.9}, {6.3, 2, 8.8},
 {6.3, 2, 10.1}, {6.3, 2, 4.1}, {6.3, 2, 9}, {6.3, 2, 5}, {6.3, 2.9, 8.8}, {6.3, 2.9, 10.1},
 {6.3, 2.9, 4.1}, {6.3, 2.9, 9}, {6.3, 2.9, 5}, {6.3, 8.8, 10.1}, {6.3, 8.8, 4.1}, {6.3, 8.8, 9},
 {6.3, 8.8, 5}, {6.3, 10.1, 4.1}, {6.3, 10.1, 9}, {6.3, 10.1, 5}, {6.3, 4.1, 9}, {6.3, 4.1, 5},
 {6.3, 9, 5}, {2, 2.9, 8.8}, {2, 2.9, 10.1}, {2, 2.9, 4.1}, {2, 2.9, 9}, {2, 2.9, 5},
 {2, 8.8, 10.1}, {2, 8.8, 4.1}, {2, 8.8, 9}, {2, 8.8, 5}, {2, 10.1, 4.1}, {2, 10.1, 9},
 {2, 10.1, 5}, {2, 4.1, 9}, {2, 4.1, 5}, {2, 9, 5}, {2.9, 8.8, 10.1}, {2.9, 8.8, 4.1},
 {2.9, 8.8, 9}, {2.9, 8.8, 5}, {2.9, 10.1, 4.1}, {2.9, 10.1, 9}, {2.9, 10.1, 5}, {2.9, 4.1, 9},
 {2.9, 4.1, 5}, {2.9, 9, 5}, {8.8, 10.1, 4.1}, {8.8, 10.1, 9}, {8.8, 10.1, 5}, {8.8, 4.1, 9},
 {8.8, 4.1, 5}, {8.8, 9, 5}, {10.1, 4.1, 9}, {10.1, 4.1, 5}, {10.1, 9, 5}, {4.1, 9, 5}}

{3.9, 4.2, 6.16667, 6.6, 4.6, 6.23333, 4.9, 2.76667, 4.73333, 5.16667, 3.16667, 4.8, 3.46667,
 5.03333, 5.46667, 3.46667, 5.1, 3.76667, 7.43333, 5.43333, 7.06667, 5.73333, 5.86667, 7.5,
 6.16667, 5.5, 4.16667, 5.8, 3.73333, 5.7, 6.13333, 4.13333, 5.76667, 4.43333, 6., 6.43333,
 4.43333, 6.06667, 4.73333, 8.4, 6.4, 8.03333, 6.7, 6.83333, 8.46667, 7.13333, 6.46667,
 5.13333, 6.76667, 4.56667, 5., 3., 4.63333, 3.3, 6.96667, 4.96667, 6.6, 5.26667, 5.4, 7.03333,
 5.7, 5.03333, 3.7, 5.33333, 7.26667, 5.26667, 6.9, 5.56667, 5.7, 7.33333, 6., 5.33333,
 4., 5.63333, 7.66667, 9.3, 7.96667, 7.3, 5.96667, 7.6, 7.73333, 6.4, 8.03333, 6.03333}

Média aritmética das variâncias:

5.73333
```

Que bate com a variância da população. Mas a probabilidade não entra no estimador da variância?

Do início. Cada lista de números tem uma média ponderada, que é a soma cada número × seu peso ou $\sum_{i=1}^{n} x_i \cdot p_i$. Se a soma dos pesos não for 1, a média ponderada, mesmo de pesos 1, será ≠ da média aritmética. Então vamos descartar estes casos. Se a soma dos pesos for 1, eles podem ser igualitários ou não. Se forem, a p de cada x será $1/n$ e a média ponderada, agora podendo chamar de esperança, $\sum_{i=1}^{n} \frac{x_i}{n}$, que é a fórmula da média aritmética, que é a esperança de números sem peso.

Melhor. Cada lista de números tem uma média ponderada. A média ponderada é um número qualquer, porque os pesos podem ser quaisquer. No caso especial dos pesos serem iguais e 1, a média ponderada reduz a $\sum_{i=1}^{n} x_i \cdot \frac{1}{n} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$, que é a média aritmética. Mas este é um caso especial, devemos esquecer média aritmética e pensar em ponderada (multiplicação e não divisão).

Quando os pesos são probabilidades (esperança), a soma é 1. **Por isso**, cada $x_i \cdot p_i$ será $< x_i$ porque $p_i < 1$.

Todo *conjunto* X tem pesos, uma média ponderada e variância.

Um passo é estabelecer *quais são os pesos para cada elemento*.
Ao se fazer, obtém-se a *distribuição* e média ponderada do conjunto.
A variância também depende dos pesos (indiretamente) porque ela é a diferença para a média ponderada:

$$\mathrm{Var}(X) = \sum_{i=1}^{n} x_i - \overline{x} =$$
$$\sum_{i=1}^{n} x_i - x_i \cdot p_i$$

Portanto, se mudar algum peso, muda a média ponderada, e, aí sim, muda a variância.

Outro passo é executar a distribuição e obter randomicamente, de acordo com ela, um novo conjunto X. (Não diferenciamos x de X; chamamos X do conjunto e x dos elementos.)

Este conjunto tem novamente pesos, média ponderada e variância. A questão é *quais* são os pesos, e é a primeira questão, anterior a qual é a média ponderada e variância.

A resposta é este conjunto não tem pesos, ou, tem pesos 1, porque ele não é uma variável aleatória.

Então podemos usar X (maiúsculas) para denominar variáveis aleatórias...

Variáveis aleatórias então são conjuntos *com* pesos diferentes de $1/n$. E variáveis fixas ou comuns, conjuntos *com apenas* pesos iguais $1/n$.
A diferença entre variáveis aleatórias e não aleatórias está em se são **equiprováveis ou não**.
Poderíamos chamar variáveis aleatórias de variáveis inequiprováveis.
Variáveis aleatórias e não-aleatórias podem ser tratadas como funções que "puxam" valores de um conjunto. A diferença é que a não-aleatória manda puxar "qualquer um" (ou "aceita qualquer um que vier"), e a aleatória "consulta a tabela de pesos".

Se esse novo conjunto (vamos chamar de subconjunto) não tem pesos, ou são equiprováveis, ou não é variável aleatória, a média poderada é média aritmética e a variância é a "simples" (vamos chamar de aritmética), "divisora" e não "multiplicativa". O subconjunto não tem distribuição (diferente de equiprovável). Portanto encerra aqui a análise sobre a **distribuição do subconjunto** ("amostra").

Dito isso, a média e variância aritméticas são úteis pois têm relações com a média e variância ponderadas do conjunto-pai.

O conjunto-pai produz os subconjuntos randômicos. Cada um pode ter elementos distintos um do outro, sendo a probabilidade de serem iguais ou não determinada pela distribuição do conjunto-pai e pelo “tipo de geração do subconjunto” ou “tipo de amostragem”.

Independente de quais são os elementos, eles *serão* por via de regra (pois a variável é aleatória) distintos em cada subconjunto, e o importante é que isso **gera uma média e variância aritméticas diferentes para cada subconjunto**.

In[12]:=
```
{data1, data1Weights, data1Weights2}
```

Out[12]=
```
{{5, 9, 2, 8.8, 6.3, 10.1, 3.4, 2.9, 4.1}, {0.0651111, 0.545815, 0.330387,
  0.327167, 0.769338, 0.228513, 0.801367, 0.264045, 0.741207}, {0.545511,
  0.281314, 0.369909, 0.0612241, 0.946807, 0.624723, 0.170142, 0.659397, 0.235823}}
```

In[29]:=
```
Module[{rss, mns, vars},
  rss = Table[RandomSample[data1Weights → data1, 4], 10];
  Table[Print[rs → "mean " <> ToString[Mean[rs]] → "variance " <> ToString[Variance[rs]]],
   {rs, rss}];
  mns = Table[Mean[rs], {rs, rss}];
  vars = Table[Variance[rs], {rs, rss}];
  Print["means mean ", Mean[mns], ", variances mean ", Mean[vars]];
 ];
```

```
{3.4, 6.3, 4.1, 8.8} → mean 5.65 → variance 5.93667
{5, 2.9, 4.1, 6.3} → mean 4.575 → variance 2.0625
{3.4, 2.9, 6.3, 8.8} → mean 5.35 → variance 7.53667
{6.3, 4.1, 3.4, 8.8} → mean 5.65 → variance 5.93667
{9, 6.3, 4.1, 8.8} → mean 7.05 → variance 5.37667
{9, 6.3, 3.4, 4.1} → mean 5.7 → variance 6.36667
{2.9, 2, 10.1, 8.8} → mean 5.95 → variance 16.75
{2, 3.4, 6.3, 8.8} → mean 5.125 → variance 9.20917
{8.8, 2, 9, 3.4} → mean 5.8 → variance 13.1467
{4.1, 8.8, 3.4, 2.9} → mean 4.8 → variance 7.35333
means mean 5.565, variances mean 7.9675
```

Porém, quanto maior a amostra, mais...

In[32]:=
```
Module[{rss, mns, vars},
  rss = Table[RandomSample[data1Weights → data1, 2], 10];
  Table[Print[rs → "mean " <> ToString[Mean[rs]] → "variance " <> ToString[Variance[rs]]],
   {rs, rss}];
  mns = Table[Mean[rs], {rs, rss}];
  vars = Table[Variance[rs], {rs, rss}];
  Print["means mean ", Mean[mns], ", variances mean ", Mean[vars]];
 ];
```

```
{6.3, 2} → mean 4.15 → variance 9.245

{6.3, 2} → mean 4.15 → variance 9.245

{8.8, 4.1} → mean 6.45 → variance 11.045

{10.1, 2.9} → mean 6.5 → variance 25.92

{6.3, 9} → mean 7.65 → variance 3.645

{3.4, 4.1} → mean 3.75 → variance 0.245

{6.3, 3.4} → mean 4.85 → variance 4.205

{4.1, 2} → mean 3.05 → variance 2.205

{6.3, 4.1} → mean 5.2 → variance 2.42

{4.1, 10.1} → mean 7.1 → variance 18.

means mean 5.285, variances mean 8.6175
```

In[33]:=
```
Module[{rss, mns, vars},
  rss = Table[RandomSample[data1Weights → data1, 8], 10];
  Table[Print[rs → "mean " <> ToString[Mean[rs]] → "variance " <> ToString[Variance[rs]]],
   {rs, rss}];
  mns = Table[Mean[rs], {rs, rss}];
  vars = Table[Variance[rs], {rs, rss}];
  Print["means mean ", Mean[mns], ", variances mean ", Mean[vars]];
 ];
```

```
{3.4, 8.8, 2, 2.9, 4.1, 6.3, 9, 5} → mean 5.1875 → variance 6.94696

{4.1, 10.1, 6.3, 9, 3.4, 2.9, 2, 8.8} → mean 5.825 → variance 9.925

{9, 3.4, 6.3, 2.9, 5, 10.1, 4.1, 2} → mean 5.35 → variance 8.5

{4.1, 3.4, 9, 8.8, 6.3, 2, 5, 2.9} → mean 5.1875 → variance 6.94696

{3.4, 2.9, 2, 10.1, 8.8, 4.1, 6.3, 9} → mean 5.825 → variance 9.925

{9, 4.1, 6.3, 8.8, 10.1, 2.9, 3.4, 2} → mean 5.825 → variance 9.925

{2.9, 6.3, 2, 8.8, 5, 4.1, 3.4, 9} → mean 5.1875 → variance 6.94696

{3.4, 8.8, 9, 4.1, 6.3, 2, 2.9, 10.1} → mean 5.825 → variance 9.925

{4.1, 6.3, 3.4, 2.9, 10.1, 5, 8.8, 2} → mean 5.325 → variance 8.29643

{2.9, 3.4, 9, 8.8, 10.1, 4.1, 5, 6.3} → mean 6.2 → variance 7.77143

means mean 5.57375, variances mean 8.51088
```

In[57]:=
```
Module[{subsets, mns, vars},
  Print["set with ", Length[data1],
   " elements, mean ", Mean[data1], ", variance ", Variance[data1]];
  Table[
   subsets = Subsets[data1, {n}];
   (*Print[n," elements: ",Length[subsets]," subsets"];*)
   mns = Table[Mean[subsets[[i]]], {i, Length[subsets]}];
   vars = Table[Variance[subsets[[i]]], {i, Length[subsets]}];
   Print[n, " elements: ", Length[subsets],
    " subsets, means mean ", Mean[mns], ", variances mean ", Mean[vars]];
   , {n, 2, Length[data1]}
  ];
 ];
```

```
set with 9 elements, mean 5.73333, variance 8.76

2 elements: 36 subsets, means mean 5.73333, variances mean 8.76

3 elements: 84 subsets, means mean 5.73333, variances mean 8.76

4 elements: 126 subsets, means mean 5.73333, variances mean 8.76

5 elements: 126 subsets, means mean 5.73333, variances mean 8.76

6 elements: 84 subsets, means mean 5.73333, variances mean 8.76

7 elements: 36 subsets, means mean 5.73333, variances mean 8.76

8 elements: 9 subsets, means mean 5.73333, variances mean 8.76

9 elements: 1 subsets, means mean 5.73333, variances mean 8.76
```

In[54]:= $\text{Table}\left[9^{n}, \{n, 0, \text{Length}[\text{data1}]\}\right]$

Out[54]= {1, 9, 81, 729, 6561, 59 049, 531 441, 4 782 969, 43 046 721, 387 420 489}

Este é o set equiprovável.

In[72]:=
```
Module[{x}, {Expectation[x, x ≈ data1Weights], Expectation[x, x ≈ data1Weights2]}]
```

Out[72]= {0.45255, 0.432761}

In[71]:=
```
RandomVariate[EmpiricalDistribution[data1Weights]]
```

Out[71]= 0.327167

Esperança e valor randômico são tirados quando de uma *amostra*, após a execução de uma amostragem da população com a distribuição informada.
Aqui estou tirando somente da distribuição...
Primeiro, preciso tirar amostras com a distribuição.
Mas lembrando que as amostras não serão variáveis aleatórias/não terão probabilidades/distribuição.
Portanto não terão expectativa e valor randômico.

Lembrando que a esperança é a *multiplicação* do valor X_i pela probabilidade... Idem para a variância esperada. Não apenas a probabilidade.

`Expectation[]` é para funções, não conjuntos. Vou calcular manualmente.

In[73]:=
```
RandomSample[data1Weights → data1, 4]
```
Out[73]=
```
{2.9, 9, 10.1, 4.1}
```

A questão é... como tirar *todos* os random samples de um tamanho de um conjunto?
Na verdade, aí já não serão mais random... Estou tentando tirar todas as permutações.

Mas esta é a questão... se tirar as permutações, não estarei mais “usando” a distribuição. Estarei tirando “todos os resultados possíveis do processo randômico”. (Que é o que eu estava fazendo acima.)
“Usar” a distribuição é tirar *apenas algumas* amostras, e *quais* amostras saem é regido pela distribuição.
Isso quer dizer que a esperança (do que for) só faz sentido, ou é diferente de “um valor absoluto”, quando estamos considerando *algumas amostras* apenas, ou informação incompleta.

Então para comprovar as tendências da média e variância das amostras em comparação às da população, ou tiro muitas (mas não todas) amostras e verifico empiricamente, ou vou para o Central Limit Theorem para comprovar.

3 - Distribuição amostral dos estimadores

4 - Estimação

5 - Intervalos de confiança para médias e proporções

6 - Testes de hipóteses para médias e proporções

7 - Erros de decisão

8 - Distribuição de t de Student IC E TH para a média de população normal com variância desconhecida

9 - Comparação de duas médias: TH para a diferença de duas médias

10 - Distribuição de x^2 Qui-Quadrado IC e TH para a variância de populações normais

11 - Testes de aderência e tabelas de contingência

12 - Distribuição de F de Fisher-Snedecor IC e TH para quociente de variâncias

1 https://en.wikipedia.org/wiki/Expected_value

2 https://en.wikipedia.org/wiki/Weighted_arithmetic_mean

3 https://en.wikipedia.org/wiki/Expected_value

4 http://mathworld.wolfram.com/ExpectationValue.html