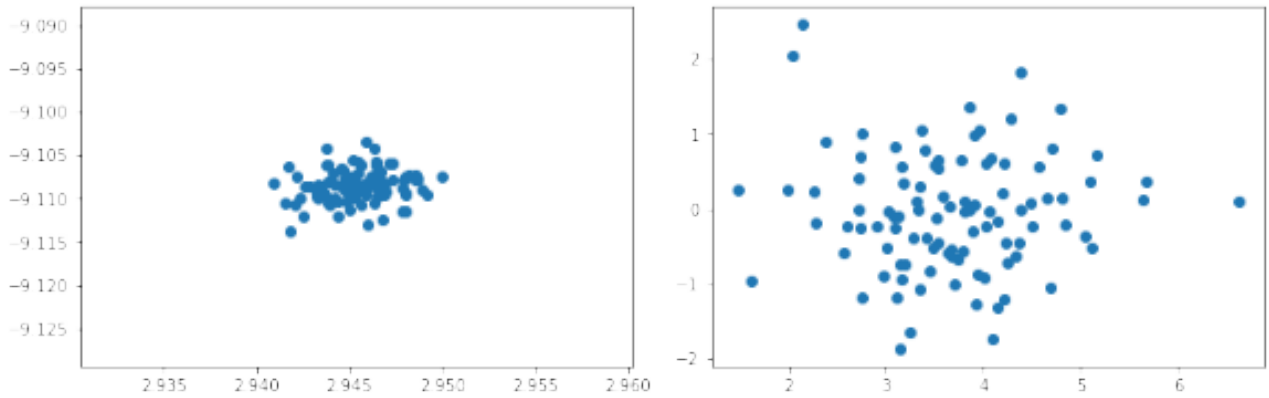
Introdução à datascience com R

Cleuton Sampaio

Lição 3: Estatística básica: Medidas de dispersão

Medidas de dispersão

Medem a dispersão dos valores com relação à média. Vamos entender o que é isso de modo visual.



Na figura, vemos dois gráficos de datasets. O da esquerda parece mais concentrado em torno de um ponto central, e o da direita, parece mais "espalhado", ou seja, sua dispersão é maior.

Lembrando: ***Outlier*** ou "valor espúrio" é um valor aberrante ou valor atípico, é uma observação que apresenta um grande afastamento das demais da série (que está "fora" dela), ou que é inconsistente. A existência de **outliers** implica, tipicamente, em prejuízos a interpretação dos resultados dos testes estatísticos aplicados às amostras (wikipedia).

Olhando os gráficos, você consegue identificar outliers?

Vamos voltar ao exemplo dos pesos. Vamos supor que temos uma turma de alunos com estes pesos:

```
# Este código é em R… Você já sabe isso, não?

turma1 <- c(75.02786847, 56.51450656, 55.57517955, 62.00893933,
            82.82022277, 91.78076684, 71.53028442, 82.22315417,
            71.14621041, 76.27644453)
```

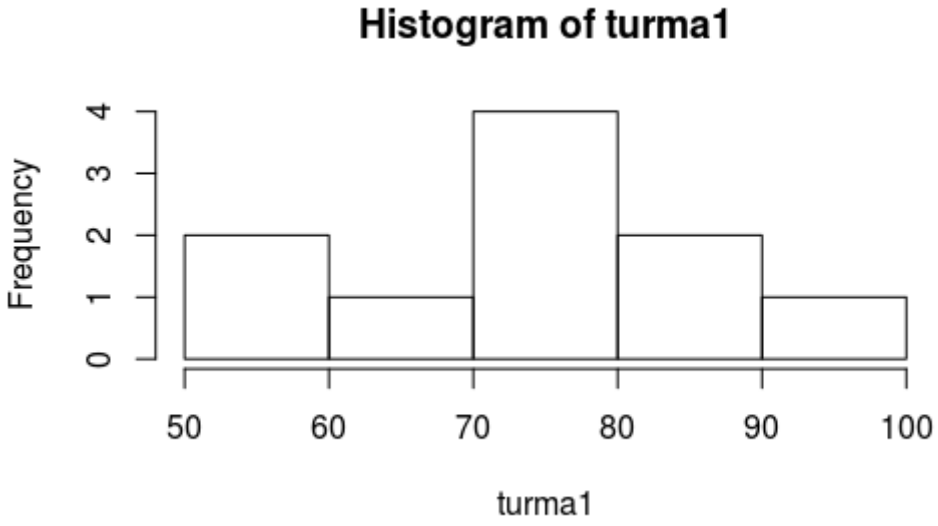
- Média: 72,490357705
- Mediana: 73,2790764451

Turma 2 =

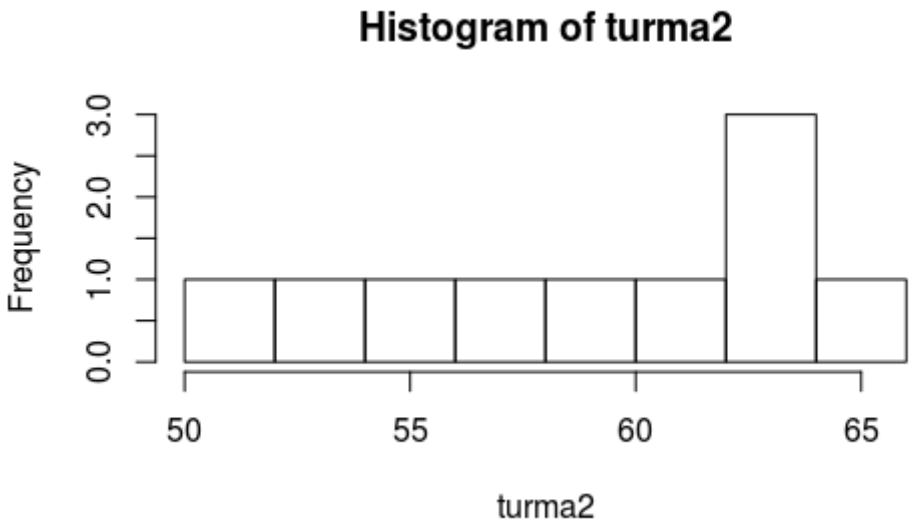
```
turma2 <- c(63.96213546,  51.00946728,  54.48449137,  53.62955058,
            61.62138863,  59.99119596,  57.61297576,  62.52220793,
            64.54041384,  63.95477107)
```

- Média: 59.3328597878
- Mediana: 60.8062922919

O que podemos deduzir dessas duas turmas? A média parece próxima da mediana… Mas, visualmente, o que podemos dizer sobre o comportamento dos pesos, com relação à média? O quanto confiamos nessa média, como uma estimativa dos pesos dos alunos de cada turma?



Os alunos da turma 1 parecem ter seus pesos distribuídos em torno da média, que é cerca de 72 Kg, com alguns "*outliers*" abaixo de 60 e acima de 90 kg. O que isto lhe diz? Cerca de 4 alunos estão em torno da média, mas, se somarmos os valores extremos, vemos que 3 alunos etão muito longe dela. Se tomarmos 72 kg como estimativa de peso, vamos errar no máximo por 28 kg, e no mínimo por 18 kg.



Já a turma 2 é bem engraçada… A média é cerca de 60 kg, e a variação entre os pesos dos alunos é menor. A maior variação é de 10 Kg, e os pesos parecem uniformemente distribuídos. Se tomarmos 60 kg como estimativa de peso, vamos errar no máximo por 10 kg de diferença.

Para saber algo sobre a dispersão dos dados é preciso conhecer algumas medidas de dispersão.

Amplitude

É a diferença entre o maior e o menor valor de um dataset. Vejamos as amplitudes das duas turmas:

- Turma 1: 36,21;
- Turma 2: 13,53.

Já ficou óbvio que na turma 2 os alunos possuem pesos mais próximos, logo a dispersão é menor. Só tem um problema: A amplitude é sensível aos "*outliers*", ou valores espúrios.

Variância

Como podemos medir mais precisamente os desvios? A amplitude apenas considera o maior e o menor valor… Se fizermos um somatório de todos os desvios? A soma simples dos desvios tenderia a zero.

Porém, se elevarmos cada diferença ao quadrado, poderemos ter uma medida mais significativa, e podemos dividir pela quantidade de elementos, obtendo assim, a variância.

$$variância = \frac{\sum_{i=1}^{n} (x_i - média)^2}{tamanho}$$

No caso da turma 1 é aproximadamente **124,84** (variância da população) e na turma 2 é **21,40** (variância da população), o que confirma nossa suspeita de que os valores dos pesos da turma 2 estão mais juntos em torno da média.

Há diferenças entre a variância da população e da amostra. No caso das turmas, cada turma é uma população, pois é o nosso alvo de estudo. Porém, se extraíssemos um subconjunto de cada turma, criaremos amostras. Quando lidamos com amostras, perdemos um "grau de liberdade", devendo dividir pelo tamanho da amostra subtraído de 1.

***Grau de liberdade** é, em estatística, o número de determinações independentes (dimensão da amostra) menos o número de parâmetros estatísticos a serem avaliados na população.*

É um estimador do número de categorias independentes num teste particular ou experiência estatística. Encontram-se mediante a fórmula n-1, onde n é o número de elementos na amostra (também podem ser representados por k-1 onde k é o número de grupos, quando se realizam operações com grupos e não com sujeitos individuais).

(Wikipedia)

Variância da população

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

Conhecida por sigma ao quadrado, é o somatório dos quadrados das diferenças entre cada elemento da população e a média populacional, dividido pela quantidade de elementos da população.

Variância da amostra

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{(n-1)}$$

Se considerarmos os dados da turma 1 e da turma 2 como amostras da população de alunos (e não como as próprias populações) então o resultado da variância é:

- Turma 1: 138,70;
- Turma 2: 23,78;

Desvio padrão

Esta é a medida mais utilizada quando analisamos erros em estimativas. O desvio padrão é a raiz quadrada da variância, simples assim! Porém:

- Desvio padrão da população: $\sigma = \sqrt{\sigma^2}$
- Desvio padrão da amostra: $s = \sqrt{s^2}$

Estes valores são os desvios amostrais das turmas 1 e 2. Mais uma vez, quando se tratar de desvio padrão da população, deve ser calculado com base na variância da população (tendo "n" como divisor), e, quando se tratar de desvio padrão da amostra, deve ser calculado com base na variância da amostra (tendo "n – 1" como dividor).

Usando o desvio padrão

Como podemos usar o desvio padrão? Uma das maneiras é para calcular a margem de erro de um conjunto de dados:

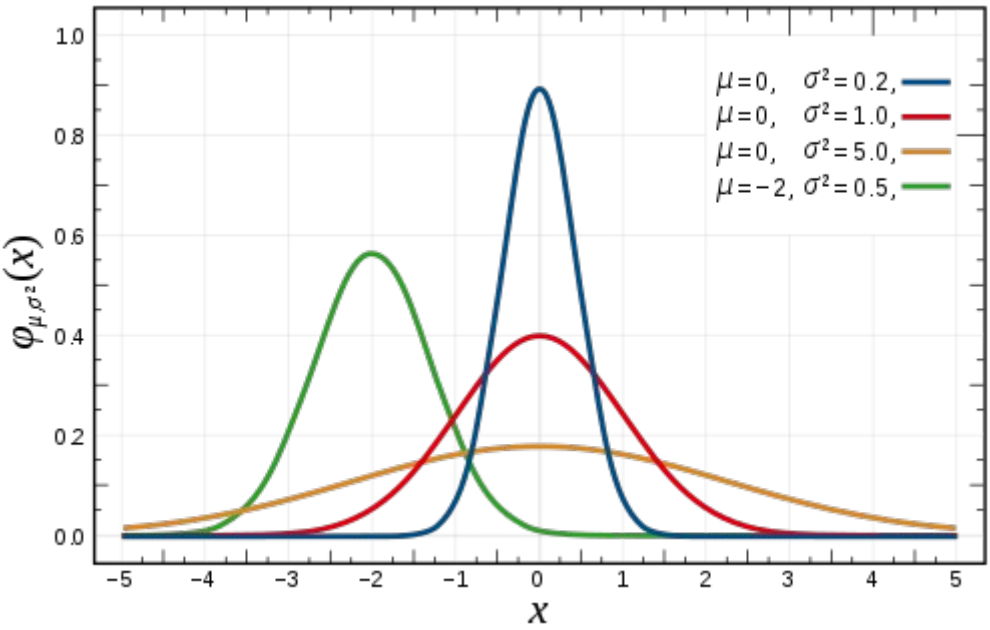
O erro padrão da média é calculado a partir do desvio padrão das médias, as quais poderiam ser computadas a partir de uma população se um número infinito de amostras e uma média para cada amostra fossem considerados. A margem de erro de uma pesquisa é calculada a partir do erro padrão da média (produto do desvio padrão populacional e do inverso da raiz quadrada do tamanho da amostra), e cerca do dobro do erro padrão da média é a metade da largura de 95% do intervalo de confiança para a média (populacional) (Wikipedia)

Estatística Z

Outro uso importante é para analisar se os valores de uma amostra se aproximam da "Distribuição Normal", de acordo com o teorema do limite central (veremos em outra sessão):

***Teste Z** é qualquer teste estatístico no qual a distribuição do teste estatístico sob a hipótese nula pode ser aproximada por uma distribuição normal. É um teste estatístico usado para inferência, capaz de determinar se a diferença entre a média da amostra e da população é grande o suficiente para ser significativa estatisticamente.[1]*

Por conta do teorema central do limite, muitos testes estatísticos são normalmente distribuídos para grandes amostras.



Gráficos de distribuições normais, com médias e desvios variados

Usando R para calcular dispersão

O script desta aula é "dispersao.R" está no repositório:

https://github.com/cleuton/datascience/tree/master/R-course/lession3

Criar os vetores das duas turmas é fácil, e você já viu isso:

```
turma1 <- c(75.02786847, 56.51450656, 55.57517955, 62.00893933,
            82.82022277, 91.78076684, 71.53028442, 82.22315417,
            71.14621041, 76.27644453)
print(paste('Média da turma 1:',mean(turma1)))
print(paste('Mediana da turma 1:',median(turma1)))

turma2 <- c(63.96213546,  51.00946728,  54.48449137,  53.62955058,
            61.62138863,  59.99119596,  57.61297576,  62.52220793,
            64.54041384,  63.95477107)
print(paste('Média da turma 2:',mean(turma2)))
print(paste('Mediana da turma 2:',median(turma2)))
```

Temos que ver esta função "paste()", que serve para converter os parâmetros em caracteres e concatená-los, retornando um texto único. É assim que criamos uma mensagem composta pelo rótulo e o valor numérico.

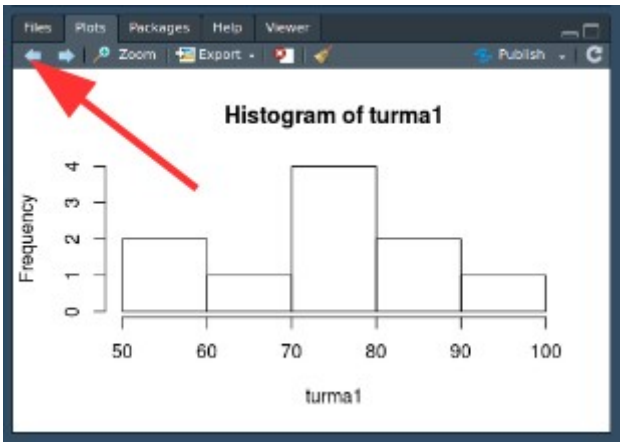
Se quisermos exibir uma mensagem composta, a função "paste()" como parâmetro do comando "print()" é uma boa opção.

O que é esta função "c" (turma1 <-- c(...)) que usamos para criar os vetores? É a função "combine", que combina os parâmetros informados em um Vetor. Os parâmetros devem ter o mesmo tipo de dados.

Geramos os histogramas das turmas 1 e 2 com estes comandos:

```
hist(turma1)
hist(turma2)
```

Os dois gráficos aparecerão na janela "plots", e você pode usar as setas azuis para alternar entre eles:



A amplitude é calculada utilizando-se as funções "max()" (pega o maior valor) e "min()" (pega o menor valor):

```
ampTurma1 <- max(turma1) - min(turma1)
ampTurma2 <- max(turma2) – min(turma2)
```

A variância amostral é calculada utilizando-se a função "var()":

```
varTurma1 <- var(turma1) # amostra
varTurma2 <- var(turma2) # amostra
```

Lembre-se que a função "var()" calcula a variância amostral (dividindo por "n – 1"). Ah, e o caracter "#" significa que o resto da linha é comentário.

A variância populacional é calculada pela equação:

```
vPopT1 <- mean((turma1-mean(turma1))^2)
vPopT2 <- mean((turma2-mean(turma2))^2)
```

Para elevar ao quadrado, usamos o operador matemático "^".

Para calcular o desvio padrão amostral, usamos a função "sd()":

```
dTurma1 <- sd(turma1) # amostral
dTurma2 <- sd(turma2) # amostral
```

E, finalmente, para calcular o desvio padrão populacional, usamos esta equação:

```
dPopT1 <- sqrt(mean((turma1-mean(turma1))^2)) # populacional
dPopT2 <- sqrt(mean((turma1-mean(turma2))^2)) # populacional
```