02

Entendendo melhor a distribuição: Variância e Desvio Padrão

Olá! Na aula passada, começamos a discutir porque que só a **média** não é suficiente quando eu quero comparar duas distribuições. Porque a média por si só não me fala o quanto a distribuição está espalhada. Ela me dá a **tendência central**, mas não me fala quanto os dados estão espalhados a partir dessa média.

E eu preciso saber isso para conseguir comparar o **João** e o **José**. Qual desses dois será que tem um ritmo melhor de trabalho? Qual tem uma frequência melhor?

Na aula passada começamos a tentar resolver esse desafio e chegamos naquela fórmula da **amplitude**, em que eu pegava o **limite superior** subtraia o **limite inferior** que resultava em um número, só que vimos que os *outliers* atrapalham. O 30 do José, em particular, atrapalhou bastante a gente.

Começamos a pensar em como eliminar esses *outliers* e chegamos na ideia da **distribuição por quartis**. Joga os *outliers* da esquerda para fora, joga os *outliers* da direita para fora e uso os dados que estão no meio.

Ótimo! Agora, vamos lá.

Vamos tentar pensar em uma maneira "elegante" de achar essa diferença, esse desvio que os números tem da média.

Para isso eu não vou usar nem o João, nem o José. eu vou usar uma distribuição menor para facilitar os nossos cálculos.

Imagine que eu tenha uma distribuição (1, 2, 9); e a média é 4.

Eu quero pensar, então, como eu posso saber a **variância média** entre os números. Uma boa ideia seria ver, por exemplo, a distância que o 1 tem da média, que é 4. Então o 1 está a 3 pontos - irei chamar de pontos - da média. O 2 para 4 faltam 2. O 9 para 4 faltam 5.

Então, a minha ideia vai ser essa: vou calcular a distância de cada ponto da média e aí, calcular a média desse número. É a **média da dispersão dos dados**.

Para essa distribuição fica fácil: `(4-1)+(4-2)+(4-9)/3`

Divido por 3 porque eu tenho 3 elementos na minha distribuição.

Isso vai me dar: `3+2-5/3`

O resultado será 0.

Está aí o primeiro problema da nossa fórmula, porque eu tenho números à esquerda, números à direita e quando eu subtraiu eles da média, eu acabo tendo **números negativos**. E o número negativo pode cancelar o número positivo. E eu não quero isso. Eu quero que cada ponto do meu gráfico tenha influência nesse valor final.

Vamos tentar melhorar nossa ideia. **Como que eu posso fazer para eliminar números negativos?**

Uma maneira fácil para isso é pôr esses números **ao quadrado**. Qualquer número ao quadrado é sempre positivo!

```
(4-1)^2+(4-2)^2+(4-9)^2/3
```

```
3^2+2^2+(-5)^2/3
```

```
9+4+25/3
```

```
38/3 ~= 13
```

(Você faça a conta direito, com 2, 3, 4 casas decimais. Aqui, eu estou arredondando para facilitar.)

Eu cheguei no número 13. Esse 13 é minha primeira tentativa de dizer o quanto os números estão dispersos dessa média, na média.

Só que o problema é que como eu elevei esses números ao quadrado, eu não tenho um número em uma dimensão que eu quero. Então, se eu elevei ao quadrado, agora eu vou resolver isso calculando a **raiz**.

Se eu pegar esse 13 e calcular a raiz dele, eu vou ter aproximadamente 3,5.

```
raiz de 13 = 3,5
```

Agora, esse 3,5 me é útil, porque ele está me dizendo que os números dessa distribuição tem a média 4 ±3,5. Então, na média, nessa distribuição, os números estão espalhados por 3,5 para lá e para cá.

Esse número 3,5 tem um nome bonito na Estatística que você já conhece: **desvio padrão**.

O 13, esse número intermediário que nós usamos para o cálculo do desvio padrão também possui um nome em Estatística: **variância**.

Veja só que usamos a **variância** para chegar no **desvio padrão**, e o desvio padrão dá para a gente o quanto aqueles números estão dispersos.

Mais importante do que a fórmula - porque a fórmula é mecânica e você decora - o importante é a ideia de como o **desvio padrão** funciona. Ele tenta descobrir a distância daquele ponto em relação à média. Ele faz isso para todos os pontos e calcula a **média** disso aí. E, assim, temos o **desvio padrão**. Isso dá para a gente a **dispersão** da nossa distribuição. E, agora, eu consigo comparar.

> Vai ficar de lição de casa para você, calcular o desvio padrão e a variância do João e do José.

Mas você vai ver que na hora de calcular, você vai chegar nos seguintes números:

```
João 1,73
```

```
José 7,02
```

O que esses números me dizem?

Eles me dizem que o João é mais consistente, porque o **desvio padrão** dele é MENOR. Então a **média** é 10±1. Então, quando eu contratar o João a chance que eu tenho dele consertar de 10, -1 para a esquerda, -1 para direita é grande, porque esse é o desvio.

Agora, o José é um cara mais complicado. A **média** dele é 10, mas ele pode ter um dia muito bom e ele consertar 17; mas ele pode ter um dia muito ruim, e consertar só 3.

Então, talvez, contratar o João te dê mais segurança, porque o João vai fazer 10±1 e o José 10±7.

Nesse caso em que eu quero maximizar o número de consertos, talvez o JOÃO seja uma aposta melhor.

Veja só, que nós tínhamos duas distribuições na aula passada - **João** e **José** -, a **média** era igual - então, para mim, eles pareciam iguais -, mas olha só como o **desvio padrão** de cada um deles é diferente.

Perceba isso: quando for comparar duas distribuições, não olhe só a **tendência central**, olhe também o **desvio padrão** desses dados e você vai entender como aquela distribuição está se comportando.

Nessa aula foi isso: **DESVIO PADRÃO** e **VARIÂNCIA**.

Te vejo no próximo capítulo. Obrigado!