02

Resumindo dados

Transcrição

Existe uma maneira muito usada para análise, além dos histogramas, e que acaba resumindo os dados. Existirão casos nos quais teremos que trabalhar com mais de 10 mil dados e não fará sentindo lermos todos eles. Precisamos resumi-los em poucos valores, que ajudem a explicar determinados comportamentos. Números que serão usados como guias para se ter uma visão rápido do que está acontecendo. Basicamente, eles terão os seguintes objetivos.

Valor Mínimo
1o Quartil
2o Quartil
3o Quartil
Valor Máximo
Média
Moda
Variância
Desvio Padrão

Nós iremos começar com um pequeno conjunto de dados que contem 15 informações, ordenadas em ordem crescente.

1	2
2	8
3	12
4	25
5	31
6	31
7	41
8	46
9	49
10	56
11	63
12	70
13	71
14	84
15	95

Iremos preencher os nossos objetivos com os dados. É fácil identificar quais são os valores mínimos e máximos. Nós iremos explicar o que aconteceu no "meio" da lista de dados, com três números que chamaremos de **quartil**, que é referente a quarta fração. Para encontrá-los, iremos dividir os números em quatro blocos. De `1` até `15`, o número posicionado no meio é `48` e será o `2o quartil`. O `1º quartil` será o que está no primeiro quarto da lista, assim como o `3º quartil` será o que estará no terceiro quarto.

1	2	
2	8	
3	12	
4	25	1 Quartil
5	31	
6	31	
7	41	
8	46	2 Quartil
9	49	
10	56	
11	63	
12	70	3 Quartil
13	71	
14	84	
15	95	

Já conseguimos preencher o resumo com alguns dados.

Valor Mínimo	2
1o Quartil	25
2o Quartil	46
3o Quartil	70
Valor Máximo	95
Média	
Moda	
Variância	
Desvio Padrão	

Percebemos que 25% dos dados estão abaixo de `25`, metade estão abaixo de `46`, 75% dos dados estão abaixo de `70`. E o maior valor é `95`. O valor mínimo também recebe o nome de `0º quartil` e o máximo também é chamado de `4º quartil`. O valor que está no meio também recebe o nome de `mediana`.

Iremos preencher os valores que faltaram no resumo: a `média` se refere à média aritmética. Neste caso ela irá coincidir com a mediana. O conceito de `moda` trata dos valores que aparecem mais vezes listados, no nosso caso, será `31`. Na prática, a `moda` é pouco utilizada.

O números chamados de `Média`, `Mediana` e `Moda` são chamados **números de tendência central**. Eles indicam que os números da lista estão distribuídos entorno de um ponto. A `variância` padrão mostrará o quão dispersos os números estão. Para calculá-la iremos usar a função `VAR` do Spreedsheet e selecionar os dados da lista. O resultado será `778,114`. Também podemos calcular o desvio padrão usando o `STD` (*Standard deviation*, em inglês) ou encontrando a raiz quadrada da variância, usando a função `SQRT` (*square root*).

No fim, ficamos com os dados preenchidos da seguinte forma:

Valor Mínimo	2
1o Quartil	25
2o Quartil	46
3o Quartil	70
Valor Máximo	95
Média	46
Moda	31
Variância	778.1142857
Desvio Padrão	27.89469996

Ele servem para descrever os valores com que trabalhamos.

1	2	0 Quaril	Valor Mínimo
2	8		
3	12		
4	25	1 Quartil	
5	31		
6	31		
7	41		
8	46	2 Quartil	Mediana
9	49		
10	56		
11	63		
12	70	3 Quartil	
13	71		
14	84		
15	95	4 Quartil	Valor Máximo

Do resumo, geralmente é levado tão em consideração a `moda` e a `variância`. Já o `desvio padrão`, sim, é relevante. No caso, encontramos o valor `27,89`, o que significa que a dispersão do centro tem este valor. Este dado pode ser usado para compararmos esta distribuição com outra.