Medidas de Posição

As medidas de posição são aquelas que tem por objetivo representar um determinado conjunto de dados, a partir de uma única informação.

“Ué, Bonel....mas, é a mesma definição das medidas de tendência central?”

“Mesma definição” é uma expressão muito forte! É semelhante. A diferença é que a medida de tendência central, tende ao centro, ou seja, é uma informação que tem objetivo representar um conjunto de dados, a partir de uma única informação que tende ao centro.

Já, as medidas de posição, nem sempre tendem ao centro. São medidas que tem por objetivo representar 1 determinada parte do conjunto de dados, que em alguns casos pode, inclusive ser uma medida de tendência central (como veremos a seguir), porém nem sempre será.

Alguns teóricos entendem que as medidas de posição devem ser consideradas como medidas de tendência central.

Vamos entender o conceito de Quartil, para que você também seja capaz de compreender, analisar e também tomar sua própria decisão, acerca das medidas de posição.

Quartis

Nessa disciplina, tomaremos como base os quartis, que são medidas de posição que dividem o conjunto de dados em 4 partes (daí o nome, quartil)

Primeiro, é importante que você seja capaz de compreender qual a necessidade/problema, dado que na grande maioria das vezes o problema não será explícito no que tange a sua necessidade.

Sendo assim, a primeira coisa a se fazer é dialogar com seu interlocutor – conforme estudamos em aulas anteriores - para entender a fundo qual o objetivo.

Dessa forma, se a o problema for:

“Gostaria de saber quais os municípios que mais registraram roubo de veículos e também aqueles que menos registraram”

Perceba que não está explícito que se trata de um cálculo de quartil. É possível, inclusive, que seja outra coisa, diferente de quartil. Por isso é importante a dialética/alinhamento de expectativas e uma compreensão profunda das necessidades apresentadas.

Dito isso, algumas perguntas poderiam de ajudar nesse desafio, como:

- Quando você diz municípios com mais e com menos, você se refere a um ranking?
    - Se for um ranking, talvez os quartis não sejam a melhor medida. Bastando apenas realizar um somatório e ordenar pela variável quantitativa, para identificar as cidades com mais e menos registros.
- Você quer identificar o top 10 (10 maiores) e o botton 10 (10 menores)?
    - Também se trata de um ranking.
- Por favor, qual seu objetivo final com essa análise?
    - Essa é uma pergunta-chave, dado que é possibilita um esclarecimento mais detalhado acerca do que precisa ser analisado. E a próxima pergunta poderá contribuir ainda mais com a compreensão do que precisa ser feito.
- O seu objetivo é analisar as cidades que mais registraram e que menos registraram, independentemente de um ranqueamento específico, como um top 10 ou botton 10?
    - Nesse ponto, pode ser que os quartis contribuam (veremos mais adiante)
- Atualmente você já faz essa análise de alguma forma, talvez em uma planilha? Se sim, pode me mostrar como é feita?
    - Entender se o que a pessoa deseja, já é feito de alguma forma, é um bom caminho para uma compreensão mais profunda, na qual você poderá inclusive propor melhorias.

Procure convidar o seu interlocutor a uma reflexão mais profunda sobre a sua solicitação. Lembre-se que é obrigação NOSSA obter as informações necessárias para o desenvolvimento, então sejamos facilitadores/as.

Para essa hipótese apresentada, vamos pressupor que o objetivo é analisar as cidades que mais registraram e que menos registraram, independentemente de um ranqueamento específico, como um top 10 ou botton 10.

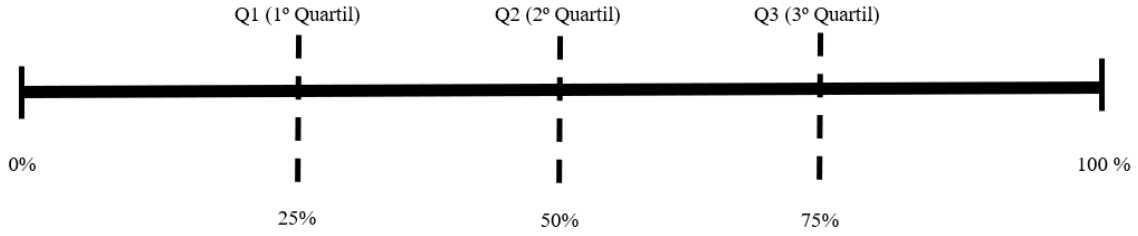
Para esse tipo de problema, certamente um bom conjunto de medidas a ser utilizado é o “quartil”, no qual nos confere 3 medidas de posição.

Medidas de posição: Quartis, limites e outliers

É importante saber que o conjunto de dados da variável qualitativa, assim como a mediana, deve estar ordenado de forma crescente.

Os quartis dividem a distribuição em 4 partes iguais, cada uma com 25%, da seguinte forma:

Figura 1 - Quartis



Fonte: Autoria própria

Vamos a fundamentação de cada uma dessas medidas:

- Q1 (1º Quartil): Essa medida vai descrever o primeiro quarto do conjunto de dados. Por exemplo, se a medida Q1 = 90, pode-se dizer que:
    - 25% dos dados desse conjunto são menores que 90
    - 75% dos dados desse conjunto são maiores que 90
- Q2 (2º Quartil): Essa medida tem uma particularidade, visto que divide o conjunto de dados em 2 partes iguais (50% menores e 50% maiores), assim como a mediana, ou seja, Q2 = Mediana. Por exemplo, se Q2 = 300, pode-se dizer que:
    - A mediana = 300
    - 50% dos dados desse conjunto são menores que 300
    - 50% dos dados desse conjunto são maiores que 300
- Q3 (3º Quartil): Essa medida vai descrever o terceiro quarto do conjunto de dados. Por exemplo, se Q3 = 800, pode-se dizer que:
    - 25% dos dados desse conjunto são maiores que 800
    - 75% dos dados desse conjunto são menores que 800

Contextualizado os quartis, retorna-se a necessidade em questão que é identificar os municípios com menos e com mais roubos de veículos. Para isso, um bom caminho para solucionar essa necessidade e calcular o Q1 para observar os municípios com menos e o Q3 para aqueles com mais.

Outliers (Intervalor interquartil, limites inferior e superior)

Uma outra possibilidade que se apresenta e que poderia complementar o problema apresentado anteriormente é a identificação de *outliers.*

Os *outliers*, em português "valores atípicos", são aqueles dados que são muito discrepantes, fogem ao comportamento de um determinado conjunto de conjunto de dados de uma variável quantitativa.

Imagine, que ao calcular os municípios com mais roubos de veículos, obteve-se o Q3 = 5.000, ou seja, todos os municípios que registraram mais de 20.000 veículos representam os 25% dos municípios que mais registraram veículos. Veja a tabela exemplo a seguir:

Tabela 1 – Exemplo munícipios acima de Q3

Município	**Qtde de roubos de veículos**
Rio de Janeiro	350.00
Duque de Caxias	280.000
São Gonçalo	230.000
Campos	80.000
Macaé	70.000
Cabo Frio	55.000
Itaboraí	50.000
Arraial do cabo	30.000
Araruama	25.000
Búzios	21.000

Fonte: Autoria própria

Observando a tabela de exemplo anterior, observa-se os 25% dos Municípios que mais registraram roubos de veículos. Agora, note que Rio de Janeiro, Duque de Caxias e São Gonçalo, além de serem os 3 Município com mais roubos de veículos, também possuem uma quantidade bem discrepante em relação aos demais Munícipios.

Observou isso?

Ou seja, isso quer dizer que além desses 3 Municípios representarem os 25% daqueles que mais registraram ocorrências, também podem ser considerados *outliers*, dado que possuem uma quantidade de registros de ocorrências bem discrepantes aos demais, possui um comportamento diferente.

Isso também acontece sob a ótica dos 25% menores, porém a quantidade é tão pequena que destoa dos demais.

É importante destacar que pode ser que um conjunto de dados, de uma determinada variável quantitativa, não possua *outliers*, isso é um indicativo de que há uma tendência do conjunto de dados está com um comportamento mais padronizado. Também pode acontecer de só existirem *outliers* sob a ótica dos maiores valores e não sob a ótica dos menores e, o contrário também é verdade.

Mas, como saber se existem ou não *outliers*?

Primeiro é interessante que você saiba que, para encontrar os *outliers* é necessário calcularmos o os primeiro e terceiro quartis. Os quartis não somente nos auxiliam a observar e descrever a distribuição, eles também são a base para se calcular os limites que separam os *outliers* do restante do conjunto de dados

Sabendo disso, vamos calcular as seguintes medidas para nos auxiliar, tanto na observação e descrição da distribuição, quanto na identificação dos outliers:

- Intervalo interquartil (IQR) = Q3 – Q1
    - O IQR tem por objetivo aferir a amplitude dentro dos quartis, observando a diferença entre o Q3 e Q1. Note que de Q1(25%) até Q3(75%) estão os 50% (75%-25%=50%) dos dados mais concentrados da nossa distribuição, sendo assim é menos sensível a influência dos valores extremos (menores e maiores)
    - Caso o resultado desse cálculo seja um valor mais próximo de Q3, a distribuição desses 50% dos dados mais concentrados, tende a ter uma alta variabilidade, ou seja, tendem a uma dispersão.
        - Exemplo:
            - Q1: 100
            - Q3: 2000
            - IQR = Q3 – Q1 = 2000 – 100 = 1900
    - Caso o resultado desse cálculo esteja tendendo a zero, ou seja, o valor de Q1 é muito próximo ao de Q3, a distribuição desses 50% dos dados mais concentrados, tendem uma baixa variabilidade.
        - Exemplo:
        - Q1: 1900

        - Q3: 2000
        - IQR = Q3 – Q1 = 2000 – 1900 = 100
    - Além dessas observações, o IQR também é utilizado como base de cálculo para se encontrar os limites que "separam" os *outliers* da distribuição
- Limite superior = Q3 + (1.5 * IQR)
    - O limite superior é a medida que "separa" os *outliers* superiores, ou seja, aqueles que possuem os maiores valores discrepantes. Ressalta-se que, por ser um cálculo matemático, o limite superior não necessariamente será um valor que exista na distribuição. Ele é apenas um cálculo que, a partir do valor do Q3 e do IQR, sugere um valor que possa delimitar os outliers.
    - Uma vez encontrado o limite superior, deve-se analisar junto com o maior valor da distribuição, da seguinte forma:
        - Se o limite superior for menor que o maior valor, significa que existem outliers. Do contrário, não existem *outliers* superiores (acima de Q3)
- Limite inferior = Q1 - (1.5 * IQR)
    - O limite inferior é a medida que "separa" os *outliers* inferiores, ou seja, aqueles que possuem os menores valores discrepantes. Ressalta-se que, por ser um cálculo matemático, o limite inferior não necessariamente será um valor que exista na distribuição. Ele é apenas um cálculo que, a partir do valor do Q1 e do IQR, sugere um valor que possa delimitar os outliers.
    - Uma vez encontrado o limite inferior, deve-se analisar junto com o menor valor da distribuição, da seguinte forma:
        - Se o limite inferior for maior que o menor valor, significa que existem outliers. Do contrário, não existem outliers inferiores (abaixo de Q1)

Referências:

CARVALHO, André C. P. L. F. de; MENEZES, Angelo Garangau; BONIDIA, Robson Parmezan. **Ciência de dados: fundamentos e aplicações**. 1. ed. 2. reimp. Rio de Janeiro: LTC, 2024.

MORETTIN, Pedro Alberto; SINGER, Júlio da Motta. **Estatística e ciência de dados**. Rio de Janeiro: LTC, 2022.