Perguntas e respostas frequentes sobre a atividade das aulas 2 e 3

Neste documento encontram-se as perguntas frequentes realizadas por estudantes da disciplina Estatística I da Escola de Artes, Ciências e Humanidades da USP ao longo do segundo semestre de 2021, agrupadas por tema abordado.

Elaboração: Letícia Figueiredo Collado

Supervisão: Alexandre Ribeiro Leichsenring

SUMÁRIO

Frequência

1. *O que significa calcular a frequência acumulada?*

A frequência absoluta indica quantas vezes aquele valor ou observação ocorre em um conjunto de dados. Quando calculamos a frequência acumulada, somamos os valores de frequência absoluta, o que nos mostra quanto daquele conjunto de dados ocorre até determinado valor. Por exemplo, no caso da tabela a seguir, sabemos que 55 das 75 pessoas do grupo têm até 22 anos de idade.

Idade	**f_i**	**F_i**
19	10	10
21	20	30
22	25	55
25	15	70
27	5	75
Total	**75**	

Média e Mediana

2. *Por que em determinados casos a mediana é uma medida mais adequada para representar a amostra do que a média?*

A **média** é afetada pela ocorrência de valores extremos, uma vez que é calculada a partir dos **valores** de um conjunto de dados. A **mediana**, por sua vez, corresponde ao valor que ocupa a **posição central** em um conjunto de dados. Por exemplo: em um grupo de 5 pessoas com idades {5; 7; 12; 15; 85}, a presença de uma única pessoa mais velha interfere na média (24,8), enquanto a mediana (12) indica que ao menos metade das pessoas deste grupo tem até 12 anos de idade.

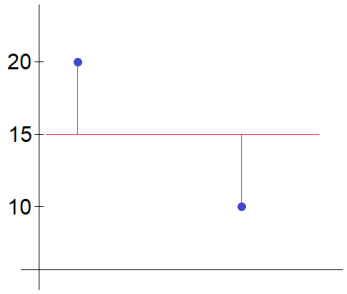
3. *Quando intervalo de classes por idade tem o valor 1 deve mesmo assim fazer a média entre as classes e deixar número quebrado? Gostaria de saber, quando recebemos um conjunto de dados com valores organizados em intervalos, se é possível utilizar outra medida além do ponto médio para os cálculos da média*

Como desconhecemos como os valores estão distribuídos dentro de cada faixa é uma boa estratégia adotar o ponto médio porque ele é equidistante dos limites inferior e superior. Só faria sentido usar outro ponto do interior do intervalo se tivéssemos alguma informação sobre a distribuição dos valores dentro de cada faixa.

Variância e Desvio Padrão

4. *Qual a diferença entre variância e desvio padrão? Por que calcular um ou outro?*

Ambas são medidas de dispersão, e indicam como estão distribuídos os valores de um conjunto de dados ao redor da média - ou seja, representam o padrão da distância de cada valor em relação à média do grupo. Entretanto, justamente por estarmos tratando de distâncias em relação à média, a variância eleva os valores ao quadrado resultando em valores não negativos:



$$(x_1 - \bar{x})^2 = (20 - 15)^2 = (5)^2 = 25$$

$$(x_2 - \bar{x})^2 = (10 - 15)^2 = (-5)^2 = 25$$

> Os dois pontos estão igualmente distantes da média!

Por outro lado, o desvio padrão, como raiz quadrada da variância, nos oferece um valor que está na mesma unidade de medida que o conjunto de dados originais, facilitando a interpretação.

5. *Visto que os valores do cálculo da variância são elevados ao quadrado, existe a possibilidade da variância ser negativa?*

Não! Os valores de variância e desvio padrão sempre serão não negativos (positivos sempre que não forem constantes).

6. *Como interpretar os resultados de variância e o desvio padrão? Existe um valor ideal?*

Os valores de variância e desvio padrão são melhor interpretados quando há mais de um conjunto de dados que podem ser comparados. Não existem valores ideais de variância e DP que possam ser sempre usados como referência, pois a avaliação da distribuição em torno da média varia com a natureza do dado que está sendo analisado.

Por exemplo: ao estudar a evolução das notas do ENEM de um determinado grupo em um intervalo de 10 anos, podemos analisar os valores de variância e desvio padrão para cada ano.

Assim teremos uma perspectiva da distribuição das notas dos estudantes em torno da média ao longo da década. Se a variância em um determinado ano é mais baixa, isso indica um desempenho mais homogêneo do grupo, com notas majoritariamente próximas da média. Por outro lado, se a variância é alta, isso indica desempenho mais heterogêneo, com notas distantes da média (para cima e para baixo).

Os valores de variância e desvio padrão são apenas uma forma de estudar um conjunto de dados e idealmente devem estar acompanhados de outras medidas, que complementam a caracterização do conjunto. Com uma caracterização mais detalhada, criamos subsídio para melhor compreender o comportamento de uma variável em uma parcela da população e podemos repensar diretrizes, metas, objetivos e instrumentos de políticas públicas existentes ou novas que sejam direcionadas a esta população.

7. *Quando temos dois grupos distintos com o mesmo tipo de dado, a variância deve ser calculada a partir da média geral dos grupos ou devemos calcular a média de cada grupo? (ex: as alturas de alunos de duas salas diferentes, como no slide 20 da aula 3)*

Se os grupos forem tratados de maneira distinta, e você tiver a intenção de descrevê-los separadamente (e eventualmente compará-los), então no cálculo da variância de um grupo se ignora qualquer informação do outro (como a média, por exemplo). Se você tiver a intenção de calcular a variância do universo inteiro, indiferentemente de que grupo pertence cada indivíduo, então se calcula a variância com os dados da união dos grupos, e a média utilizada deverá ser aquela resultante do universo composto pelos dois grupos. Nesse caso, você estaria interessado na descrição global do universo, e não na comparação entre os grupos.

8. *Em que situação o cálculo da amplitude dos dados pode ser mais útil do que o desvio padrão quando se trata de um cálculo de dispersão?*

Cada medida fornece um tipo de característica da distribuição dos dados. Elas são complementares e devem ser usadas em conjunto para uma melhor compreensão da distribuição dos dados.

9. *Para que serve e como interpretar o coeficiente de variação?*

A variância e o desvio padrão medem a dispersão em relação à média de forma absoluta. Para que seja possível comparar conjuntos de dados com médias diferentes, ou até com unidades de medida diferentes, calculamos o coeficiente de variação, que permite analisarmos a dispersão em termos relativos. O valor do CV é uma porcentagem em relação à média; quando o CV for alto, a dispersão é alta e, portanto, os dados são mais heterogêneos; quando o CV for baixo, a dispersão é baixa e, portanto, os dados são mais homogêneos.

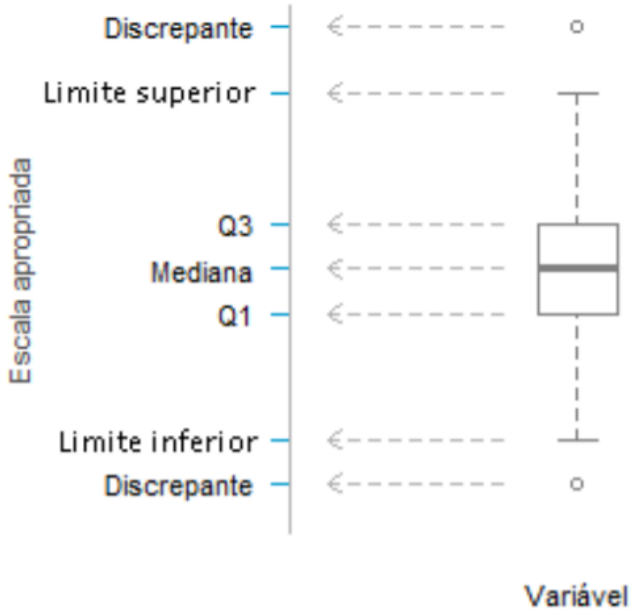
Boxplot

10. Como o boxplot representa graficamente a distribuição de um conjunto de dados?

Com o *boxplot* podemos observar onde está concentrada a grande massa de valores do conjunto e se temos (e onde estão) valores discrepantes, que se distanciam muito do resto do grupo.

O conjunto de dados pode ser dividido em quartis, que correspondem aos valores que dividem o conjunto em quatro partes iguais. O primeiro quartil indica o limite dos valores que correspondem aos primeiros 25% do total de observações ordenadas; o segundo quartil indica o limite dos valores que correspondem aos primeiros 50% do total de observações ordenadas - ou seja, a mediana, valor que ocupa a posição central no conjunto; o terceiro quartil indica o limite dos valores que correspondem aos primeiros 75% do total de observações ordenadas.

A amplitude do intervalo interquartil (AIQ), que é dada pela distância entre o primeiro e o terceiro quartis (AIQ = $Q_3 - Q_1$), indica onde está concentrada a metade central dos valores observados daquele conjunto - valores que estão entre os primeiros 25% e 75% do total.



11. Por que os outliers não são considerados o ponto máximo e/ou mínimo de um box plot?

"Ponto máximo" e "ponto mínimo" de fato não são os melhores nomes para o que eles significam nesse contexto. No *boxplot*, eles determinam valores máximos e mínimos *típicos* da maioria dos conjuntos de dados. Ele se vale de características de uma distribuição teórica conhecida (a distribuição Normal) para indicar um intervalo onde se situam valores *comuns*

da distribuição. Valores situados para além desses limites (ou seja, do "máximo" e do "mínimo"), são considerados discrepantes (*outliers).*

Os limites são calculados subtraindo-se de do primeiro quartil (Q_1) uma vez e meia a amplitude do intervalo interquartil e somando esse mesmo valor ao terceiro quartil (Q_3).

$$LI = Q_1 - 1{,}5 \cdot (Q_3 - Q_1)$$

$$LS = Q_3 + 1{,}5 \cdot (Q_3 - Q_1)$$

Valores do conjunto que sejam menores do que o limite inferior ou maiores do que o limite superior são considerados *outliers*.

12. Como evitar que outliers nos levem a leituras incorretas do conjunto de dados?

Estudar quais são e onde estão os valores discrepantes é o primeiro passo para pensar em como lidar com eles. *Outliers* podem indicar um problema na coleta dos dados (respostas discrepantes que não foram previstas quando a pergunta foi formulada, um registro errado, etc) e podem causar distorções na interpretação de valores de outras medidas sensíveis a eles (como a média). Por outro lado, também podem apontar para casos específicos a serem estudados dentro do conjunto maior de dados e contribuir com a caracterização da amostra em questão.

Geral

13. Qual a aplicação prática destes conceitos?

As medidas de tendência central e medidas de dispersão são utilizadas na etapa de estatística descritiva para descrever e representar um conjunto de dados. Com isso, nos auxiliam na compreensão das características de um determinado grupo e podem contribuir com o direcionamento da análise e interpretação dos dados a partir do uso de diversas técnicas estatísticas. Importantes pesquisas públicas, como a Pesquisa Nacional por Amostra de Domicílios (PNAD) e os Censos Demográficos do IBGE, o Relatório Anual de Informações Sociais (RAIS) do Ministério do Trabalho, a Pesquisa Origem Destino do Metrô, entre outras, oferecem dados diversos sobre a população brasileira sobre os quais podemos nos debruçar para avaliar a pertinência e os efeitos das políticas públicas existentes e formular novas políticas.

14. Como escolher quais variáveis utilizar no meu estudo?

A escolha das variáveis a serem utilizadas depende, em grande parte, do escopo da pesquisa; um levantamento bibliográfico inicial é essencial para reunir informações sobre o que outros autores têm colocado como fatores relevantes para a discussão até o momento, quais são as lacunas em estudos sobre o tema, se há discordâncias entre pesquisadores da área, etc, ajudando a direcionar nosso olhar e processo de tomada de decisão. Além disso, grande parte das pesquisas utiliza fontes secundárias de informação - dados já coletados por outros pesquisadores e/ou instituições, como as pesquisas do IBGE, dados disponíveis em portais de transparência, informações requeridas por meio da Lei de Acesso à Informação, etc, - restringindo também a escolha das variáveis à sua disponibilidade, uma vez que nem sempre tudo que seria ideal para a realização de uma pesquisa está disponível ou acessível em tempo hábil para o pesquisador.