DATA SCIENCE & ANALYTICS

MBA

STATS

00000000

Table of Contents

1. INTRODUÇÃO E CONTEXTO

1.1. IMPORTÂNCIA DAS ESTATÍSTICAS DESCRITIVAS E INFERENCIAIS

A estatística é uma ciência que se dedica à coleta, análise, interpretação e apresentação de dados. Dentro desse campo, as estatísticas descritivas e inferenciais desempenham papéis cruciais.

- Estatísticas Descritivas: Referem-se aos métodos utilizados para descrever e resumir os dados coletados. Elas fornecem uma visão geral das características dos dados por meio de medidas de tendência central, como a média, mediana e moda, e medidas de dispersão, como a amplitude, variância e desvio padrão. O principal objetivo das estatísticas descritivas é simplificar grandes quantidades de dados em resumos concisos e informativos, permitindo uma compreensão rápida e eficiente das informações principais.

Exemplo: Ao analisar os dados de salários de uma empresa, as estatísticas descritivas podem revelar a média salarial, a faixa de variação dos salários e a distribuição dos salários entre os funcionários.

- Estatísticas Inferenciais: Referem-se aos métodos utilizados para fazer previsões ou inferências sobre uma população com base em uma amostra de dados. Isso envolve a utilização de técnicas de amostragem e a aplicação de testes de hipóteses, intervalos de confiança e modelos de regressão para generalizar os resultados da amostra para a população inteira. O principal objetivo das estatísticas inferenciais é tirar conclusões válidas e confiáveis sobre a população a partir da análise dos dados amostrais.

Exemplo: Ao realizar uma pesquisa de opinião pública, as estatísticas inferenciais permitem estimar a porcentagem da população que apoia uma determinada política com base em uma amostra representativa.

A integração dessas duas abordagens é fundamental para a tomada de decisões informadas e baseadas em dados, seja em pesquisa científica, negócios, políticas públicas ou outras áreas.

1.2. Contextualização das Medidas de Tendência Central, Dispersão e Associação

Dentro do campo das estatísticas descritivas, as medidas de tendência central, dispersão e Associação são ferramentas essenciais para a análise e interpretação dos dados.

- ### Medidas de Tendência Central:

Estas medidas são utilizadas para identificar o ponto central ou valor típico de um conjunto de dados. As principais medidas de tendência central são a média, mediana e moda. Cada uma dessas medidas oferece uma perspectiva diferente sobre os dados:

- Média: Representa a soma de todos os valores dividida pelo número de observações. É a medida mais comum e intuitiva, mas pode ser influenciada por valores extremos.
- Mediana: O valor central de um conjunto de dados ordenado. É mais resistente a outliers e fornece uma visão melhor do centro dos dados quando a distribuição é assimétrica.
- Moda: O valor mais frequente em um conjunto de dados. É especialmente útil para dados categóricos.

Exemplo: Em um estudo sobre os rendimentos mensais de uma população, a média fornece uma visão geral dos rendimentos, enquanto a mediana pode oferecer uma visão mais precisa em caso de grandes desigualdades.

Medidas de Dispersão:

Estas medidas indicam o grau de variação ou espalhamento dos dados em torno da medida de tendência central. As principais medidas de dispersão são a amplitude, variância, desvio padrão e coeficiente de variação:

- Amplitude: A diferença entre o valor máximo e o valor mínimo. Fornece uma noção básica da extensão dos dados.
- Variância: A média dos quadrados das diferenças entre cada valor e a média. Mede a dispersão dos dados em torno da média.
- Desvio Padrão: A raiz quadrada da variância. Expressa a dispersão dos dados na mesma unidade dos valores originais.
- Coeficiente de Variação (CV): A razão entre o desvio padrão e a média, expressa como uma porcentagem. Permite a comparação da dispersão entre diferentes conjuntos de dados.

Exemplo: Ao analisar o desempenho dos alunos em um exame, a variância e o desvio padrão podem indicar a consistência das notas, enquanto a amplitude mostra a faixa completa das notas obtidas.

▪ Medidas de Associação

As medidas de associação são essenciais para entender a relação entre variáveis, ajudando a quantificar a força e a direção dessas relações.

A escolha da medida adequada depende do tipo de dados e da natureza da relação entre as variáveis.

❖ *Coeficiente de Correlação de Pearson*

- Definição: Mede a força e a direção da relação linear entre duas variáveis quantitativas.
- Intervalo: -1 a +1
- Aplicabilidade: Relações lineares entre variáveis quantitativas.
- Exemplo: Horas de estudo e notas em um exame.

❖ Coeficiente de Correlação de Spearman

- Definição: Mede a força e a direção da associação monotônica entre duas variáveis usando classificações.
- Intervalo: -1 a +1
- Aplicabilidade: Dados ordinais ou relações não lineares.
- Exemplo: Classificação de satisfação do cliente e lealdade.

❖ Coeficiente de Correlação de Kendall

- Definição: Mede a força da associação monotônica entre duas variáveis ordinais.
- Intervalo: -1 a +1
- Aplicabilidade: Dados ordinais com muitos empates.
- Exemplo: Preferência de produtos e qualidade percebida.

❖ Coeficiente Phi

- Definição: Mede a associação entre duas variáveis categóricas binárias.
- Intervalo: -1 a +1
- Aplicabilidade: Tabelas de contingência 2x2.
- Exemplo: Fumar (sim/não) e doença cardíaca (sim/não).

❖ Coeficiente de Contingência de Cramér

- Definição: Mede a associação entre duas variáveis categóricas em tabelas de contingência maiores que 2x2.
- Intervalo: 0 a 1
- Aplicabilidade: Tabelas de contingência de dimensões maiores.
- Exemplo: Nível de educação e preferência política.

2. MEDIDAS DE TENDÊNCIA CENTRAL

As medidas de tendência central são estatísticas que descrevem o ponto central ou típico de um conjunto de dados.

Medidas de tendência central são fundamentais na análise de dados, pois resumem um conjunto de observações com um único valor representativo, facilitando a compreensão e comparação de diferentes conjuntos de dados.

As principais medidas de tendência central são a média, a mediana e a moda.

2.1. MÉDIA

A média, também conhecida como média aritmética, é a soma de todos os valores de um conjunto de dados dividida pelo número total de valores. É a medida de tendência central mais comum e amplamente utilizada.

- Média: É a soma de todos os valores dividida pelo número de observações. É sensível a valores extremos.

2.1.1. VANTAGENS E DESVANTAGENS DA MÉDIA:

- Utiliza todos os valores do conjunto de dados.
- Fácil de calcular e interpretar.

    - Sensível a valores extremos (outliers), que podem distorcer a média.

2.2. Mediana

A mediana é uma medida estatística que indica o valor central de um conjunto de dados, quando esses dados são organizados em ordem crescente (ou decrescente).

Mediana é uma medida de tendência central, utilizada para entender a distribuição dos dados. A mediana é especialmente útil em distribuições assimétricas, pois não é afetada por valores extremos (outliers) da mesma forma que a média.

> Mediana: É o valor que separa a metade superior da metade inferior de um conjunto de dados ordenados.

Para calcular a mediana:

- Ordena-se o conjunto de dados.
- Se o número de observações (n) for ímpar, a mediana é o valor no meio da lista ordenada.
- Se o número de observações (n) for par, a mediana é a média dos dois valores centrais.

A mediana é o valor central de um conjunto de dados ordenado. Se o número de observações for ímpar, a mediana é o valor do meio. Se for par, a mediana é a média dos dois valores centrais.

2.2.1. Vantagens e desvantagens da mediana:

- Não é influenciada por valores extremos.
- Útil para dados ordinais e distribuições assimétricas.

- Não utiliza todas as informações dos dados.
- Pode ser menos intuitiva em dados grandes ou complexos.

2.3. Moda

A moda é o valor que aparece com maior frequência em um conjunto de dados.

Um conjunto de dados pode ter uma única moda (unimodal), mais de uma moda (bimodal, multimodal) ou nenhuma moda.

- Moda: É o valor que aparece com mais frequência em um conjunto de dados. Pode haver mais de uma moda ou nenhuma moda.

2.3.1.Vantagens e Desvantagens da Moda:

- Simples de entender e calcular.
- Aplicável a dados qualitativos e categóricos.

- Pode não ser única ou não existir.
- Não usa todas as informações dos dados.

3. MEDIDAS DE DISPERSÃO

Após entender a centralidade, é lógico seguir com a variabilidade dos dados para ter uma visão mais completa.

As medidas de dispersão são estatísticas que descrevem o grau de variação ou espalhamento dos valores em um conjunto de dados.

Medidas de Dispersão complementam as medidas de tendência central, oferecendo uma visão mais completa sobre a distribuição dos dados.

As principais medidas de dispersão incluem a amplitude, a variância, o desvio padrão e o coeficiente de variação.

3.1. AMPLITUDE

A amplitude é a diferença entre o valor máximo e o valor mínimo de um conjunto de dados.

Amplitude fornece uma medida simples do intervalo em que os dados estão distribuídos.

3.1.1. VANTAGENS E DESVANTAGENS DA AMPLITUDE:

- Fácil de calcular e interpretar.
- Sensível a valores extremos (outliers).
- Não considera a distribuição de todos os valores.

3.2. VARIÂNCIA

A variância é uma medida de dispersão que indica o quão distante, em média, os valores de um conjunto de dados estão da média desses valores.

É uma das principais ferramentas em estatística para quantificar a variabilidade dos dados.

- A variância mede a dispersão dos valores em relação à média;
- Variância é a média dos quadrados das diferenças entre cada valor e a média do conjunto de dados.

3.2.1.VANTAGENS E DESVANTAGENS DA VARIÂNCIA:

- Utiliza todas as observações do conjunto de dados.
- Fornece uma base para outras medidas de dispersão, como o desvio padrão.

- Difícil de interpretar, pois está em unidades ao quadrado.

3.3. Desvio Padrão

O desvio padrão é a raiz quadrada da variância. Ele expressa a dispersão dos dados na mesma unidade dos valores originais, tornando-o mais intuitivo.

3.3.1. Vantagens e desvantagens do desvio padrão:

- Mais fácil de interpretar do que a variância.
- Utiliza todas as observações do conjunto de dados.

- Ainda pode ser influenciado por outliers.

3.4. Coeficiente de Variação (CV)

O coeficiente de variação é a razão entre o desvio padrão e a média, expressa como uma porcentagem.

O coeficiente de variação permite a comparação da dispersão entre conjuntos de dados com diferentes unidades ou médias.

3.4.1.Vantagens e desvantagens do coeficiente de variação:

- Permite comparações de variabilidade entre diferentes conjuntos de dados.
- Normaliza a dispersão em relação à média.

- Não é definido para médias iguais a zero.

4. Medidas de Associação

Depois de cobrir a centralidade e a dispersão, é apropriado discutir como as variáveis se relacionam entre si.

4.1. Coeficiente de Correlação de Pearson

O coeficiente de correlação de Pearson, também conhecido como correlação produto-momento de Pearson, é uma medida estatística que quantifica a força e a direção da relação linear entre duas variáveis quantitativas.

Este coeficiente é amplamente utilizado em estatística e análise de dados para determinar como duas variáveis se movem juntas.

4.1.1. Aplicabilidade do Coeficiente de Correlação de Pearson:

O coeficiente de correlação de Pearson é aplicável quando:

- As variáveis são quantitativas e contínuas.
- A relação entre as variáveis é linear.
- Os dados são normalmente distribuídos.

4.1.2. Interpretação de Resultados

4.1.2.1. Correlação Positiva

1. **Correlação Positiva:**
   Considere os seguintes pares de valores para as variáveis X e Y:

$$X = [1, 2, 3, 4, 5]$$

$$Y = [2, 4, 6, 8, 10]$$

   O cálculo do coeficiente de correlação de Pearson resultaria em $r = 1$, indicando uma correlação positiva perfeita.

4.1.2.2. Correlação Negativa:

2. **Correlação Negativa:**
   Considere os seguintes pares de valores para as variáveis X e Y:

$$X = [1, 2, 3, 4, 5]$$

$$Y = [10, 8, 6, 4, 2]$$

   O cálculo do coeficiente de correlação de Pearson resultaria em $r = -1$, indicando uma correlação negativa perfeita.

4.1.2.3. Nenhuma Correlação:

3. **Nenhuma Correlação:**

   Considere os seguintes pares de valores para as variáveis X e Y:

$$X = [1, 2, 3, 4, 5]$$

$$Y = [2, 3, 1, 5, 4]$$

   O cálculo do coeficiente de correlação de Pearson resultaria em r próximo de 0, indicando nenhuma correlação linear significativa.

5. Assimetria e Curtose

A assimetria e a curtose são medidas de forma utilizadas para descrever a distribuição de variáveis quantitativas.

- A Assimetria indica a direção e o grau de inclinação da distribuição;
- A Curtose fornece informações sobre a altura e a largura das caudas da distribuição.

Ambas são fundamentais para a análise estatística, pois ajudam a entender melhor a distribuição dos dados e a presença de outliers.

5.1. Aplicação às Variáveis Quantitativas:

Tanto a assimetria quanto a curtose são aplicadas a variáveis quantitativas, pois essas medidas dependem de cálculos que envolvem valores numéricos contínuos ou discretos. Variáveis qualitativas, por sua natureza categórica, não possuem uma distribuição numérica que permita a análise de forma, simetria ou caudas.

5.2. Assimetria (Skewness):

A assimetria é uma medida que descreve a simetria ou a falta de simetria na distribuição de uma variável. Em uma distribuição simétrica, os dados são distribuídos igualmente em torno da média. A assimetria pode ser positiva (distribuição inclinada à direita) ou negativa (distribuição inclinada à esquerda).

- Assimetria Positiva: A cauda da distribuição se estende mais à direita.
- Assimetria Negativa: A cauda da distribuição se estende mais à esquerda.
- Assimetria Nula: A distribuição é perfeitamente simétrica.

5.2.1. Exemplo: Assimetria

```python
import numpy as np
from scipy.stats import skew

salarios = [30000, 35000, 40000, 45000, 50000, 60000, 150000]
assimetria = skew(salarios)
print(f"Assimetria: {assimetria}")
```

Neste exemplo, a assimetria será positiva, indicando que há uma cauda à direita mais longa, com um valor extremo alto (150000).

5.3. Curtose (Kurtosis):

A curtose é uma medida que descreve a "altura" e a "largura" das caudas de uma distribuição em comparação com uma distribuição normal. Ela quantifica a extremidade dos valores na distribuição, ou seja, a frequência de valores extremos (outliers).

- Leptocúrtica: Distribuições com curtose positiva (>3), que têm caudas mais longas e picos mais altos que a normal.
- Mesocúrtica: Distribuições com curtose próxima de zero (aproximadamente 3), como a distribuição normal.
- Platicúrtica: Distribuições com curtose negativa (<3), que têm caudas mais curtas e picos mais baixos que a normal.

5.3.1.Exemplo: Curtose

Considere um conjunto de dados sobre alturas:

```python
from scipy.stats import kurtosis

alturas = [160, 165, 170, 175, 180, 185, 190, 200, 210]
curtose_valor = kurtosis(alturas)
print(f"Curtose: {curtose_valor}")
```

Se a curtose for positiva, a distribuição terá caudas mais longas e um pico mais alto em comparação com a normal. Se for negativa, terá caudas mais curtas e um pico mais achatado.

6. Teste Qui-Quadrado

O teste qui-quadrado é um teste estatístico utilizado para examinar se há uma associação significativa entre duas variáveis categóricas.

Ele compara a distribuição observada dos dados com a distribuição esperada sob a hipótese nula de que as variáveis são independentes.

O valor da estatística qui-quadrado e seu p-valor são utilizados para analisar se existe uma associação estatisticamente significativa entre duas variáveis categóricas em uma tabela de contingência.

- Se o p-valor for menor que o nível de significância, rejeitamos a hipótese nula de independência, concluindo que há uma associação significativa entre as variáveis.

6.1. Tabela de Contingência:

Uma tabela de contingência é uma matriz que mostra a frequência de observações para combinações de categorias de duas variáveis categóricas.

Cada célula na tabela contém a contagem de casos que correspondem a uma combinação específica de categorias das duas variáveis.

6.2. P-Valor:

6.2.1. Hipótese Nula (H0) e Hipótese Alternativa (HA):

- Hipótese Nula (H0): É a afirmação de que não há efeito, associação ou diferença significativa entre as variáveis em estudo. No caso de testes de independência, como o teste qui-quadrado, a hipótese nula afirma que as variáveis são independentes. Ou seja, o valor de uma variável não fornece informação sobre o valor da outra variável.

- Hipótese Alternativa (HA): É a afirmação de que há um efeito, associação ou diferença significativa entre as variáveis. No caso de testes de independência, a

hipótese alternativa afirma que as variáveis não são independentes, ou seja, existe uma associação entre elas.

O p-valor é a probabilidade de observar, por acaso, uma estatística de teste que é tão extrema ou mais extrema do que a estatística observada, sob a suposição de que a hipótese nula (H0) é verdadeira.

Em outras palavras, o p-valor ajuda a determinar a evidência contra a hipótese nula.

6.3. Interpretação do P-Valor:

- P-Valor Baixo (geralmente ≤ 0.05): Indica evidência forte contra a hipótese nula, levando à sua rejeição. Isso sugere que os resultados observados são improváveis de ocorrer se a hipótese nula fosse verdadeira.

    - Um p-valor baixo indica que é improvável que a associação observada seja devida ao acaso, sugerindo uma associação significativa entre as variáveis.

- P-Valor Alto (geralmente > 0.05): Indica evidência insuficiente contra a hipótese nula, não levando à sua rejeição. Isso sugere que os resultados observados não são improváveis de ocorrer sob a hipótese nula.