Advanced
Institute for
Artificial
Intelligence

Introdução à Estatística

https://advancedinstitute.ai

Introdução à Estatística

Referências

Referências e Fontes das Imagens

- Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython (Book)
- Learning the Pandas Library: Python Tools for Data Munging, Analysis, and Visualization (Book)

Introdução

- Ciência de **aprender com dados**
  - Ajuda a usar os métodos adequados para coletar os dados, empregar a análise correta e apresentar os resultados de forma eficaz
- Os dados que estudamos são observações (amostras) de uma ou mais variáveis.
- Uma **variável** (aleatória) é aquilo que é observado para estudar um determinado fenômeno (idade, sexo, peso, etc.)
- A Estatística provê meios para classificar, resumir, organizar, analisar e interpretar dados.
- Envolve: descrever Conjuntos de Dados e tirar conclusões (fazer estimativas, decisões, previsões, etc. a cerca de conjuntos de dados

Introdução

- Estatística descritiva vs inferencial:
  - Estatísticas descritivas se concentram na descrição das características visíveis de um conjunto de dados (uma população ou amostra)
  - Estatísticas inferenciais se concentram em fazer previsões ou generalizações sobre um conjunto de dados maior, com base em uma amostra desses dados.

Variáveis

- Os dados são as informações que você coleta para aprender, tirar conclusões e testar hipóteses;
- Diferentes tipos de variáveis que registram vários tipos de informações
- O tipo de informação determina o que você pode aprender com ela;
- As variáveis podem ser categóricas (qualitativas) ou numéricas (quantitativas)

Variáveis Qualitativas vs Quantitativas

- **Variáveis qualitativas**: As informações representam características que você não mede com números. Em vez disso, as observações caem dentro de um número finito de grupos.
  - **Ordinais**: Grau de gravidade de uma doença
  - **Nominais**: Presença de um sintoma
  - Em alguns casos podemos realizar o mapeamento para valores discretos e depois proceder a análise como quantitativa
- **Variáveis quantitativas**: As informações são registradas como números e representam uma medição objetiva ou uma contagem
  - Discretas: Número de cirurgias
  - Contínuas: Idade, Pressão Arterial

Variáveis Qualitativas vs Quantitativas

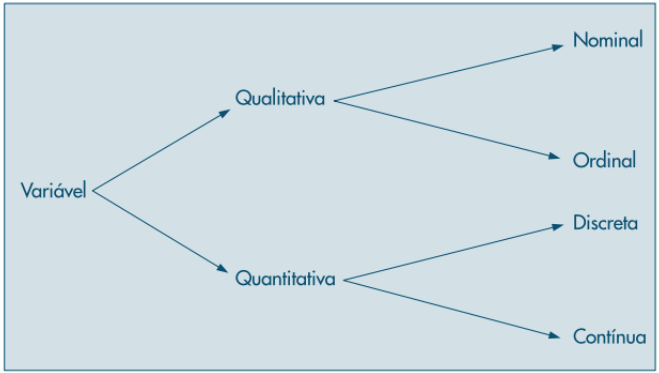


Figure: Classificação de Variáveis

Medidas Resumo Tabela de Frequências

- Forma de representação da frequência de cada valor distinto da variável em estudo
- Frequências relativas: percentagem relativa à frequência
- Frequências acumuladas: número de vezes que uma variável assume um valor inferior ou igual a esse valor.
- Frequências relativas acumuladas: percentagem relativa à frequência acumulada

Medidas Resumo Tabela de Frequências

Modo de contratação	Frequência absoluta	Frequência relativa
PJ	6	6 -> 10 60%
CLT	4	4 -> 10 40%
Total	10	100%

Medidas Resumo Medidas de Posição

- Valores representativos de uma série completa
- **Redução drástica dos dados**
- **Moda**: valor de maior frequência no conjunto de valores observados (podem existir mais que uma moda);
- **Mediana**: Valor que ocupa a posição centrar numa série de observações ordenada em ordem crescente.
  - Quando quantidade de observações for par, utilizar média aritmética entre valores centrais
- **Média Aritmética:** Soma de todos os valores dividido pelo número de observações

Medidas Resumo

Considerando a série de valores:

$$Z = \{17.5, 16.8, 15.4, 15.1, 14.4, 13.9, 13.8, 13.3, 13.0, 13.0\}$$

- **Média** ($\overline{Z}$): 14.62
- **Moda** (Mod(Z): 13.0
- **Mediana** (Med(Z): 14.15

Medidas de Dispersão

- Uma única medida representativa de posição central **esconde a informação sobre a variabilidade** do conjunto de observações;
- Considere o exemplo: suponhamos que cinco grupos de alunos submeteram-se a um teste, obtendo-se as seguintes notas:

grupo A (variável X): 3, 4, 5, 6, 7
grupo B (variável Y): 1, 3, 5, 7, 9
grupo C (variável Z): 5, 5, 5, 5, 5
grupo D (variável W): 3, 5, 5, 7
grupo E (variável V): 3, 5, 5, 6, 6

- Fácil verificar: $\overline{X} = \overline{Y} = \overline{Z} = \overline{W} = \overline{V} = 5.0$

Medidas de Dispersão

- A identificação de cada uma destas séries por sua média (5, em todos os casos) **nada informa sobre suas diferentes variabilidades**.
- Necessidade de serem criadas medidas que sumarizem a variabilidade de um conjunto de observações
  - Permite comparar conjuntos diferentes de valores, segundo algum critério estabelecido
    - E.g., dispersão em torno da média
- Análise de desvios em relação a média: $x_i - \overline{x}$
  - Para o grupo A: $\{-2, -1, 0, 1, 2\}$
  - soma igual a zero para qualquer conjunto de dados;

Medidas de Dispersão

- Alternativas ao Desvio da Média:
  - Considerar o total dos desvios em valor absoluto
  - Considerar o total dos quadrados dos desvios
  - Utilizar a média para poder comparar conjuntos com escalas diferentes

$$dm(X) = \frac{\sum_{i=1}^{n} |x_i - \overline{x}|}{n}$$

$$var(X) = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n}$$

Medidas de Dispersão

- Variância é uma medida de unidade igual ao quadrado da dimensão dos dados (e.g., se os dados são expressos em *cm*, a variância será expressa em cm^2);
  - Problemas de interpretação
- **Desvio padrão**: raiz quadrada positiva da variância:

$$dp(X) = \sqrt{var(X)}$$

- Ambas as medidas de dispersão (dm e dp) indicam:
  - Em média qual será **o "erro" (desvio) ao tentar substituir cada observação pela medida resumo do conjunto de dados** (no caso, a média)

Medidas de Dispersão Exercício

Considerando a série de valores:

$$\mathcal{X} = \{2,59; 2,64; 2,60; 2,62; 2,57; 2,55; 2,61; 2,50; 2,63; 2,64\}$$

Calcular:

1. Média;
2. Mediana;
3. Desvio Médio;
4. Desvio Padrão;

Quantis Empíricos

- Tanto a média como o desvio padrão **podem não ser medidas adequadas para representar um conjunto de dados**, pois:
  - São afetados, de forma exagerada, por **valores extremos**;
  - Apenas com estes dois valores não temos **ideia da simetria ou assimetria da distribuição dos dados**.
- A mediana é um valor que deixa metade dos dados abaixo dela e metade acima
- Quantil de Ordem p ($q(p)$), onde p é uma proporção qualquer, $0 < p < 1$, tal que 100% das observações sejam menores do que $q(p)$

Quantis Empíricos

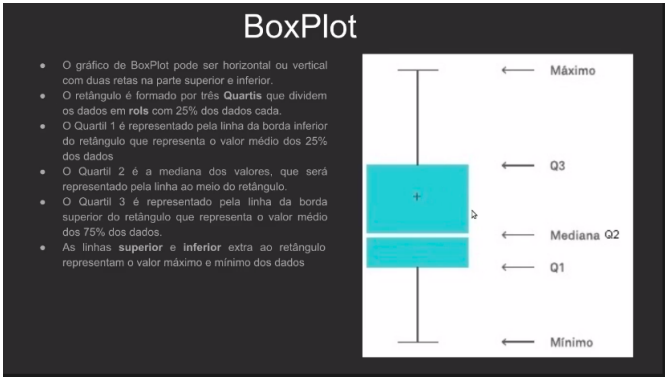


Figure: Diferentes Quantis

Quantis Empíricos

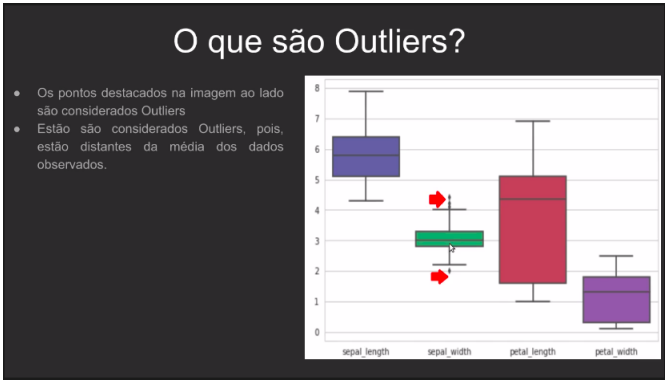
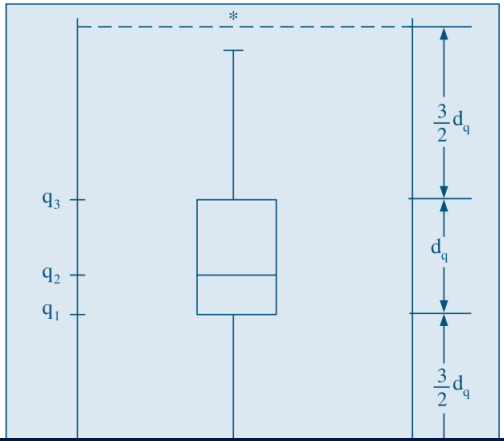


Figure: Diferentes Quantis

Quantis Empíricos

- Boxplot do nosso exemplo:



Quantis Empíricos

$q(0{,}25) = q_1$:	1º Quartil = 25º Percentil
$q(0{,}50) = q_2$:	Mediana = 2º Quartil = 50º Percentil
$q(0{,}75) = q_3$:	3º Quartil = 75º Percentil
$q(0{,}40)$:	4º Decil
$q(0{,}95)$:	95º Percentil

Figure: Diferentes Quantis

- Considerando a variável $\mathcal{X} = \{15, 5, 3, 8, 10, 2, 7, 11, 12\}$
  - Mediana $md = q(0,5) = q_2 = x_{(5)} = 8$
  - $q_1 = \dfrac{3+5}{2} = 4$, $q_3 = \dfrac{11+12}{2} = 11.5$

Quantis Empíricos

Acrescentemos, agora, o valor **67** à lista de nove valores do exemplo anterior, obtendo-se $n = 10$ valores.

Question

Qual o efeito na média e na mediana nesse caso?

Quantis Empíricos

Acrescentemos, agora, o valor **67** à lista de nove valores do exemplo anterior, obtendo-se $n = 10$ valores.

Question

Qual o efeito na média e na mediana nesse caso?

- $q_2 = \dfrac{x_{(5)} + x_{(6)}}{2} = 9$
- $2 < 3 < \mathbf{5} < 7 < 8 < 10 < 11 < \mathbf{12} < 15 < 67$
- Distância inter-quartil $d_q = q_3 - q_1 = 12 - 5 = 7$

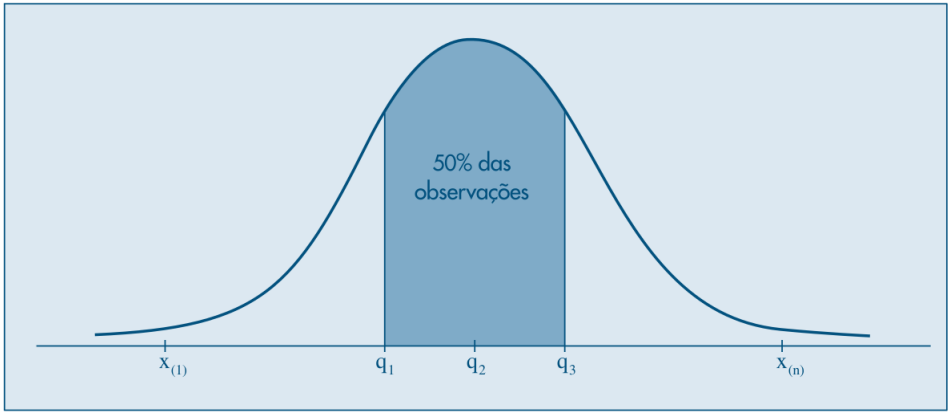
Quantis Empíricos

- Os valores, x(1), q1, q2, q3 e x(n) são importantes para se ter uma ideia da assimetria da distribuição dos dados
- Para uma distribuição simétrica ou aproximadamente simétrica, deveríamos ter:

(a) $q_2 - x_{(1)} \simeq x_{(n)} - q_2$;
(b) $q_2 - q_1 \simeq q_3 - q_2$;
(c) $q_1 - x_{(1)} \simeq x_{(n)} - q_3$;
(d) distâncias entre mediana e q_1, q_3 menores do que distâncias entre os extremos e q_1, q_3.

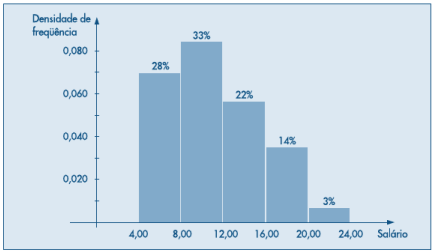
Quantis Empíricos

- Distribuição normal ou gaussiana



Histogramas

- Gráfico de barras contíguas, com as **bases proporcionais aos intervalos das classes** e a **área de cada retângulo proporcional à respectiva frequência**
- **Quanto mais dados** tivermos em cada classe, **mais alto deve ser o retângulo**
- Densidade de Frequência: sendo Δ_i o tamanho do intervalo de agrupamento, a altura deve ser proporcional a n_i/Δ_i



Conhecendo as chances

- Perceber o comportamento de eventos aleatórios é de grande importância para a nossa sociedade
- a área de estudo conhecida como probabilidade faz a análise desses eventos para entender quais são as chances reais de eles ocorrerem.
- A probabilidade conta com conceitos importantes, como experimento aleatório, evento, espaço amostral, e eventos equiprováveis.
- Para compreender o cálculo da probabilidade, precisamos dominar alguns conceitos, como espaço amostral, evento e experimento aleatório.

Observações

- A probabilidade pode ser representada como fração, como porcentagem ou como número decimal.
- A probabilidade é sempre um número decimal entre 0 e 1, ou uma porcentagem entre 0% e 100%.
- Se $P(A) = 0$ então A é um evento impossível.
- Se $P(A) = 1$ então A é um evento certo.

Espaço amostral

- É o conjunto de todos os resultados possíveis de um evento aleatório.
- Em um experimento aleatório, conhecer o espaço amostral é essencial para que a gente consiga calcular a probabilidade desse evento acontecer.
- Por exemplo, em um lançamento de um dado normal, o espaço amostral será Ω: 1,2,3,4,5,6

Ponto amostral

- É um elemento que pertence ao espaço amostral, ou seja, um entre os vários resultados possíveis do experimento aleatório.
- Por exemplo, ao lançar-se uma moeda para o alto, o resultado coroa é um ponto amostral assim como o resultado cara, a depender de qual dos lados aparece após a queda do objeto.

Evento

- É qualquer subconjunto do espaço amostral.
- O evento pode ser representado utilizando-se notação de conjuntos, ou seja, por letras maiúscula.
- Em um experimento aleatório, será sorteado ao acaso um estado brasileiro. Nesse experimento podemos tirar vários possíveis eventos, por exemplo, podemos pensar no resultado ser um estado do Sul, logo, meu evento pode ser representado pelo conjunto A: Rio Grande do Sul, Paraná, Santa Catarina.
- Outro possível evento é o conjunto de estados cujos nomes comecem com a letra s, nesse caso o evento será o conjunto B: Santa Catarina, Sergipe, São Paulo.

Cálculo da probabilidade

- Todos os conceitos vistos são essenciais para compreender-se o cálculo da probabilidade.
- Dado um experimento aleatório, calculamos a chance de um determinado evento ocorrer, essa probabilidade é dada pela razão entre o número de elementos do meu conjunto evento, ou seja, o número de casos favoráveis sobre o número de elementos no meu espaço amostral, ou seja, o número de casos possíveis.

$$P(A) = \frac{n(A)}{n(\Omega)}$$

P(A) → probabilidade do evento A

n(A) → número de elementos no conjunto A

n(Ω) → número de elementos no conjunto

Exemplo

- É qualquer subconjunto do espaço amostral.Uma urna contém bolas brancas, vermelhas e verdes. Sabendo-se que nela há 12 bolas brancas, 8 vermelhas e que as 5 restantes são brancas, se uma bola for retirada ao acaso, qual é a probabilidade de que ela seja:
- **Branca:**

Exemplo

- Nosso evento A é $\rightarrow$ sair uma bola branca. Sabemos que n(A) $= 12$, ou seja, há 12 casos favoráveis.
- Nosso espaço amostral possui um total de $12 + 8 + 5 = 25$, então n(Ω) $= 25$.
- Dessa forma, a probabilidade de o evento A ocorrer pode ser representada por:

$$P(A) = \frac{n(A)}{n(\Omega)}$$

$$P(A) = \frac{12}{25} = 0,48 \; ou \; 48\%$$

Exemplo

- É qualquer subconjunto do espaço amostral.Uma urna contém bolas brancas, vermelhas e verdes. Sabendo-se que nela há 12 bolas brancas, 8 vermelhas e que as 5 restantes são brancas, se uma bola for retirada ao acaso, qual é a probabilidade de que ela seja:
- **Não é Branca:**

Exemplo

- Nosso evento B é $\rightarrow$ sair uma bola não branca. Sabemos que $n(B) = 13$.
- Como o espaço amostral continua o mesmo, então $n(\Omega) = 25$.

$$P(A) = \frac{n(A)}{n(\Omega)}$$

$$P(A) = \frac{13}{25} = 0,52 \; ou \; 52\%$$

Dúvidas?