Estatística Descritiva

1. Introdução

A **Estatística Descritiva** é a área da estatística que busca conhecer e sintetizar as informações a partir de um conjunto de dados quaisquer. Existem diversas ferramentas estatísticas que podem ser utilizadas na interpretação dos dados, entre elas métricas, tabelas e gráficos que irão auxiliar no entendimento e ajudar a resumir as informações deste conjunto de dados. Um ponto importante é que para que essa análise estatística seja feita de forma clara e direta, deve-se entender os **tipos de variáveis** e suas características para escolher as melhores abordagens.

2. Tipos de Variáveis

As variáveis são valores, numéricos ou não, que representam características de interesse a respeito do conjunto de dados. Na Estatística Descritiva, as variáveis mais utilizadas são separadas em duas categorias, sendo elas **qualitativas** e **quantitativas**, onde, dentro destas categorias, pode-se dividir essas variáveis em dois grupos cada:

- **qualitativa nominal**: as variáveis do tipo *qualitativas* não apresentam valores mensuráveis. No caso das variáveis **qualitativas** e **nominais**, as variáveis **não apresentam uma ordenação ou hierarquia** entre as categorias. **Exemplo:** Sexo, País, estado civil e etc;
- **qualitativa ordinal**: Já para as variáveis **qualitativas** e **ordinais**, as variáveis **apresentam uma ordenação ou hierarquia** entre as categorias. **Exemplo:** escolaridade, faixa salarial, período do dia e etc;
- **quantitativa discreta**: as variáveis do tipo *quantitativas* apresentam valores mensuráveis, e para as variáveis **quantitativas** e **discretas**, as variáveis são representadas por **quantidades enumeráveis** (isto é, que podemos contar). Exemplos: Quantidade de filhos, quantidade de TVs, número de carros que trafegam por dia em determinada rua, entre outros;
- **quantitativa contínua:** as variáveis **quantitativas** e **contínuas** podem apresentar valores contínuos dentro da escala real, podendo apresentar valores fracionários, decimais, etc. Exemplo: Salário, fração de *Bitcoin*, altura e etc.

Como exemplo de fixação, utilizando os dados da tabela abaixo que foram retirados do site do IBGE (Instituto brasileiro de Geografia e Estatística) a respeito sobre algumas características de cidades brasileiras:

Cidade	População	Densidade Demográfica	Ranking Pop. Residente	Região

São Paulo	11.253.503	7.398,26	1	Sudeste
Curitiba	1.751.907	4.027,04	8	Sul
Brasília	2.570.160	444,66	4	Centro-Oeste
Manaus	1.802.014	158,06	7	Norte
Fortaleza	2.452.185	7.786,44	5	Nordeste

Classificando os dados apresentados na tabela, com os tipos de variáveis apresentados anteriormente, tem-se que:

- **Cidade**: qualitativa e nominal;
- **População**: quantitativa e discreta;
- **Densidade Demográfica** (habitantes/km^2): quantitativa e contínua;
- **Ranking População Residente**: qualitativa e ordinal;
- **Região**: qualitativa e nominal.

3. Métricas de Posição e Dispersão

Como mencionamos anteriormente, o principal objetivo da estatística descritiva é gerar **medidas que resumem** o conjunto de dados a serem analisados, ou seja, medidas que descrevem a **distribuição** dos dados, de forma quantitativa. Temos dois tipos de métricas disponíveis: as métricas de **posição** e **dispersão**.

3.1 Métricas de Posição

As **métricas de posição** são medidas que resumem algumas propriedades dos dados, indicando tendências e descrevendo como se comporta determinada distribuição de dados, de forma quantitativa. As principais métricas de posição utilizadas na estatística descritiva são:

Média Aritmética

Dado X uma amostra aleatória e $x_1, x_2, x_3, ..., x_n$ os dados observados por esta amostra.

Define-se **média aritmética** de $\overline{X}$ como sendo:

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + x_3 + ... + x_n}{n}$$

Mediana

A **mediana** é a métrica que indica a tendência central dos dados, representando o valor central que separa os **dados ordenados** de uma determinada distribuição. O valor da mediana varia de acordo com o número de elementos que têm na amostra, portanto pode-se definir a mediana como:

- Dado que o número de elementos da amostra é **ímpar**, a mediana será $X = X_{\frac{n+1}{2}}$
- Dado que o número de elementos da amostra é **par**, a mediana será $X = \frac{X_{\frac{n}{2}} + X_{\frac{n+2}{2}}}{2}$

Moda

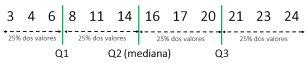
A **moda** é a métrica de posição que indica o valor de **maior ocorrência** em um conjunto de dados. Para o caso da moda, dependendo do conjunto de dados, ele pode ser definida das seguintes formas:

- **Sem Moda**: Todos os valores da amostra são distintos, ou seja nenhum valor se repete;
- **Unimodal**: Apenas um valores se repete com maior frequência no conjunto de dados;
- **Multimodal**: 2 ou mais valores se repetem com maior frequência no conjunto de dados.

Quartis

Os **quartis** são valores dados a partir do conjunto de observações ordenado em ordem crescente, que dividem os dados **em quatros partes iguais**. Dessa forma define-se 3 métricas:

- O primeiro quartil (Q_1), sendo o número que separa 25% das observações abaixo deste valor e 75% acima;
- O segundo quartil (Q_2) equivale a **mediana**, ou seja é o número que separa as observações em duas partes iguais (50%);
- O terceiro quartil (Q_3), sendo o número que separa 75% das observações abaixo deste valor e 25% acima.



Fonte: Aprendendo Gestão

Intervalo Interquartil (IQR)

O **intervalo interquartil** é uma métrica auxiliar para a construção do gráfico de caixas (*box-plot*), onde consiste em identificar 50% das observações ao redor da mediana (25% para cada lado) e avaliar o espalhamento destes dados. O cálculo do intervalo interquartil é dado por:

$$IQR = Q_3 - Q_1$$

3.2 Métricas de Dispersão

As **métricas de dispersão** são medidas de **variabilidade**, que indicam o quanto as observações de um conjunto de dados variam ao redor de alguma medida de centralidade (média, mediana, etc.). Dessa forma indicam o quão afastada determinada observação pode estar em relação a uma métrica de posição. Algumas das principais métricas de dispersão utilizadas na estatística descritiva, estão descritas a seguir:

Amplitude

A **amplitude** identifica justamente a diferença entre o valor **máximo** e **mínimo** de uma determinada amostra aleatória X, indicando o tamanho da sequência de valores possíveis que a amostra de dados possa assumir:

$$A = max(X) - min(X) = max(x_1, \ldots, x_n) - min(x_1, \ldots, x_n)$$

Importante salientar que por trabalhar com os valores extremos de uma distribuição, a amplitude é uma métrica muito sensível a **valores discrepantes** (*outliers*).

Variância

A variância representa o quão distantes os dados estão em relação a média das observações, definida pela fórmula a seguir:

$$Var(X) = \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n-1} = \frac{(x_1 - \overline{x})^2}{n-1} + \frac{(x_2 - \overline{x})^2}{n-1} + \cdots + \frac{(x_n - \overline{x})^2}{n-1}$$

Note que quanto mais próximo as observações estão do valor médio, menor vai ser o valor da variância; analogamente quanto maior o valor da variância mais afastados estão as observações da média.

Desvio Padrão

O **desvio padrão**, de forma análoga a variância, mede o quão distantes os dados estão em relação a média. Mas como a variância trabalha com os **valores quadráticos** (as unidades serão quadráticas, como m^2, $R\$^2$), no desvio padrão isto é corrigido aplicando a **raiz quadrada** da variância (assim, as unidades serão as mesmas dos dados, como m, y):

$$\sigma(X) \equiv \sqrt{Var(X)} = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n-1}}$$

4. Associação entre Variáveis Quantitativas

Dados X e Y amostras aleatórias e suas respectivas observações $x_1, \ldots, x_n$ e $y_1, \ldots, y_n$, respectivamente, pode se calcular algumas métricas de maneira análoga desenvolvido para as métricas de dispersão, mas que avalia interação entre diferentes amostras. Algumas dessas métricas são descritas a seguir:

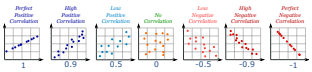
4.1 Covariância

A **Covariância** seria o caso geral para a variância, onde faz-se o comparativo de quão distantes duas amostras aleatórias X e Y estão das suas respectivas médias:

$$S_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

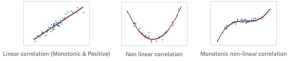
4.2 Correlação

A **correlação** é uma métrica utilizada para avaliar a dependência entre duas variáveis, onde é possível quantificar como diferentes variáveis interagem entre si. O valor da correlação varia entre $-1 \leq r \leq 1$, ou seja para valores mais próximos dos extremos, as variáveis apresentam **maior correlação** entre si e quanto mais proxímo de zero a correlação, diminui cada vez mais a dependência dessas variáveis. Essa progressão entre os valores extremos da correlação podem ser representados na figura a seguir:



Fonte : DataDeck

Importante também salientar que as correlações **não necessariamente são lineares**, podendo também apresentar correlação não-lineares entre as variáveis:



Fonte : Medium

Existem alguns testes específicos para calcular a correlação, onde o principal utilizado é a correlação de __Pearson__.

Correlação de Pearson

A **correlação de Pearson** é definida como a taxa de relação linear entre duas variáveis numéricas. Quanto mais próximos dos extremos -1 ou 1, mais **linearmente correlacionadas** estão as variáveis analisadas. O cálculo para a correlação de Pearson é definido como:

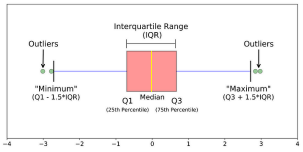
$$r = \frac{s_{XY}}{\sqrt{s_X s_Y}}$$

5. *Boxplot* e *Outliers*

O *Boxplot* é uma representação utilizada para avaliar a distribuição de um determinado conjunto de dados. Para a construção do *Boxplot* são utilizadas as métricas de primeiro quartil (Q_1) mediana ou segundo quartil (Q_2) e terceiro quartil (Q_3). As hastes, também conhecidas como bigodes (**whiskers**), representam os limites dessa representação. Os valores dos limites superiores e inferiores são definidos da seguinte forma:

$$Lim_{inferior} = Q_1 - 1,5IQR = Q_1 - 1,5(Q_3 - Q_1) \;\; Lim_{superior} = Q_3 + 1,5IQR = Q_3 + 1,5(Q_3 - Q_1)$$

A representação gráfica do *Boxplot* pode ser visualizada na figura a seguir:

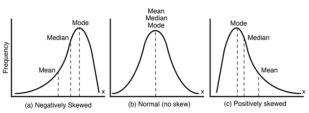


Fonte: Ichi.Pro

Os pontos que ultrapassam os limites do *Boxplot* são chamdos de **valores discrepantes** (*Outliers*). Estes valores são identificados como observações que **destoam do padrão da distribuição** dos dados. Os *Outliers* podem existir por diversas razões, sejam elas erro humano ou experimental, ou mesmo valores discrepantes mas que façam sentido dentro do contexto da distribuição.Sempre que identificar valores discrepantes em um conjunto de dados, é importante entender qual é a natureza destes valores e se foram oriundos de algum erro, por fim fazer o tratamento adequado destes dados.

6. Assimetria

A **assimetria** é definida como o grau de desvio ou afastamento da simetria de uma distribuição. Quando determinada distribuição é simétrica, as métricas de posição como média, mediana e moda localizam-se em um mesmo ponto, havendo um perfeito equilíbrio na distribuição, mas normalmente este equilíbrio dentro de uma distribuição não acontece, a média, a mediana e a moda recaem em pontos diferentes da distribuição, e a distribuição é considerada **assimétrica**. Abaixo tem-se a representação das formas que uma distribuição qualquer pode apresentar com relação a simetria:



Fonte: Research Gate

Uma distribuição assimétrica pode ser classificada em 2 tipos de assimetrias:

- Distribuição assimétrica **Negativa** ou **"à esquerda"**: as características das métricas para esta assimetria são que média < mediana < moda;
- Distribuição assimétrica **Positiva** ou **à direita**: já para o caso da assimetria positiva tem-se que moda < mediana < média.

A métrica utilizada para avaliar o grau de assimetria de uma determinada distribuição é a **distorção** (*skewness*), definido da seguinte forma:

$$s(X) = \frac{1}{\sigma^3} \sum_{i=1}^{n} \frac{(x_i - \overline{x})^3}{n}$$

Para identificar o tipo de assimetria a partir do *skewness*, deve-se seguir a referência de valores abaixo:

- s = 0: distribuição simétrica
- s > 0: assimetria à direita
- s < 0: assimetria à esquerda

7. Tabelas de Frequências

A distribuição ou tabela de frequências é um agrupamento de dados em classes ou categorias, de tal forma que contabiliza-se o número de ocorrências em cada uma delas. O objetivo é apresentar os dados de uma maneira sumarizada e que permita analisar os dados a partir destes grupos formados. Existem 3 tipos de frequências para ser utilizadas em agrupamentos, sendo elas indicadas a seguir:

- **Frequência absoluta (fi):** É o número de observações correspondente a cada classe, comumente chamada apenas de frequência:

Nível	Alunos
Ensino Infantil	87
Ensino Fundamental I	122
Ensino Fundamental II	117
Ensino Médio	154

- **Frequência relativa ($f$ri):** É o quociente entre a frequência absoluta da classe correspondente e a soma das frequências (total observado), isto é, $f_{ri} = \frac{f_i}{\sum_j^n f_j}$ onde n representa o número total de observações.

Nível	Relativa
Ensino Infantil	0,181
Ensino Fundamental I	0,254
Ensino Fundamental II	0,244
Ensino Médio	0,321

- **Frequência percentual (pi):** É obtida multiplicando a frequência relativa a 100%.

Nível	Percentual (%)
Ensino Infantil	18,1
Ensino Fundamental I	25,4
Ensino Fundamental II	24,4
Ensino Médio	32,1

Materiais Complementares

Documentação do NumPy;

Artigo sobre Métricaa de Tendência Central publicado pela Jéssica Temporal;

Referências

Pedro A. Morettin, Wilton O. Bussab, Estatística Básica, 8ª edição

Peter Bruce, Andrew Bruce & Peter Gedeck, Practical Statistics for Data Scientists, 50+ Essential Concepts Using R and Python, 2ª edition

Ron Larson & Betsy Farber, Estatística Aplicada, 6ª edição.