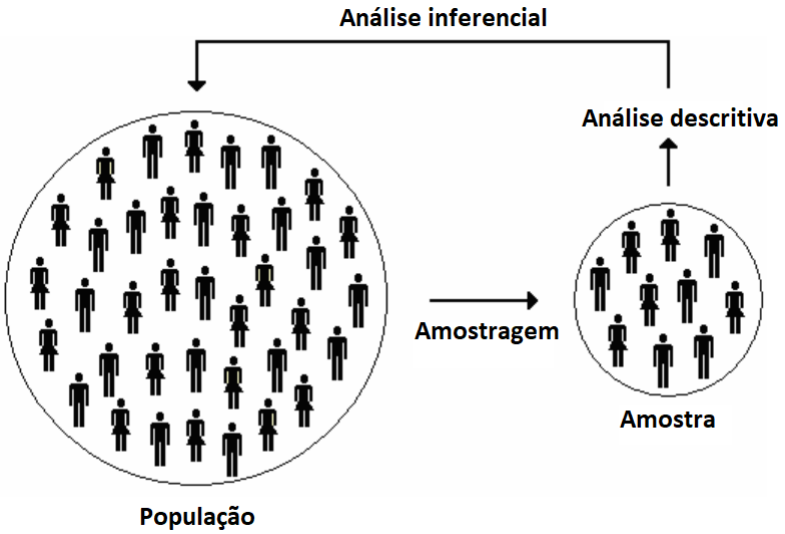
1 Revisão

1.a Introdução

A estatística fornece métodos para coleta, resumo, organização, descrição, apresentação, análise e interpretação de dados, para a utilização dos mesmos na tomada de decisões.

Figura 1: A estatística busca responder uma pergunta de interesse sobre a população com base em uma amostra.



Fonte: Adaptada de https://www.google.com/imghp?hl=pt-BR

Inferência: a análise e a interpretação dos dados — obtidos a partir de uma amostra — para a tomada de decisões sobre a população constitui o problema central da inferência estatística.

Probabilidade: as inferências estatísticas são feitas utilizando-se alguns resultados da teoria de probabilidades. Embora intimamente associada à estatística, tem suas características próprias e busca quantificar a incerteza existente em determinada situação.

1.b Análise Descritiva

A estatística descritiva se ocupa da organização, apresentação e sintetização de dados.

Variáveis:

- são características das unidades de análise;
- podem ser qualitativas (nominais ou ordinais);
  - a escala de mensuração nominal classifica as variáveis em termos de atributos. O caso mais simples é formado pela divisão em classes classes que indicam a presença ou não de determinada característica (as classes são mutuamente excludentes);
  - a escala ordinal é utilizada quando os fenômenos ou observações podem ser arranjados segundo uma ordenação(grandeza, hierarquia, preferência, importância, distância, etc.).
- podem ser quantitativas (discretas ou contínuas).

Tabelas de Frequências:

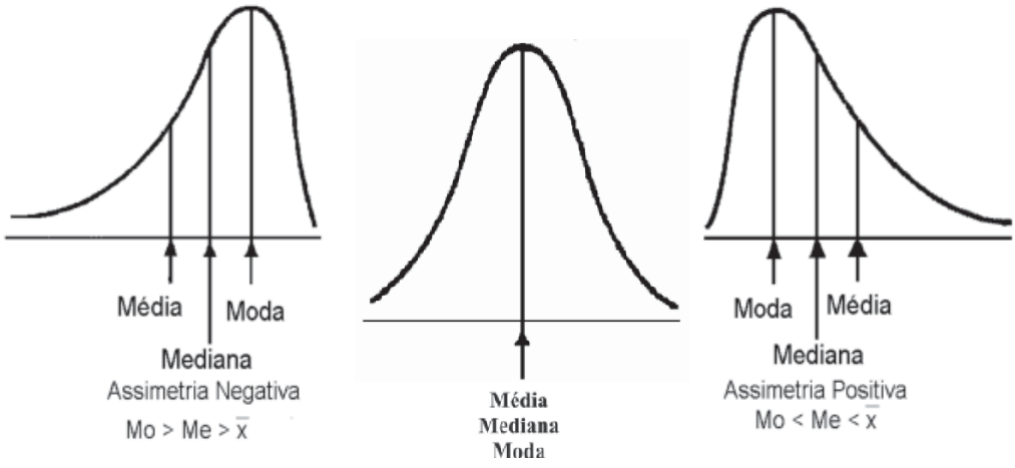
- Uma distribuição de frequência (ou tabela de frequência) lista as categorias ou os valores observados individualmente ou por intervalos (classes), juntamente com sua contagem (frequência);
  - se a variável em estudo for contínua, é conveniente agrupar os valores observados em classes;
  - se a variável é quantitativa discreta e o número de observações da variável for muito grande, recomenda-se o agrupamento dos dados em classes de valores;
  - no caso da representação das frequências por classes, há uma simplificação da realidade, pois perde-se informação com relação aos verdadeiros valores observados;
  - adota-se a hipótese de que todos os valores de uma classe são iguais ao valor que se encontra no centro da classe:

$$\text{Ponto médio da classe} = \frac{\text{limite inferior da classe} + \text{limite superior da classe}}{2}.$$

Medidas de Tendência Central:

- média aritmética: encontrada adicionando-se todos os valores e dividindo-se o resultado pelo número total de ocorrências $\left(\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}\right)$;
- média ponderada: calculada multiplicando-se cada valor possível do conjunto de dados por um peso a ele atribuído; depois encontra-se a soma desses valores e divide-se pela soma dos pesos $\left(\bar{X} = \frac{\sum_{i=1}^{k} X_i * W_i}{\sum_{i=1}^{k} W_i}\right)$;
- mediana: valor que divide uma distribuição exatamente em duas metades;
- moda: valor mais frequente do conjunto de dados observados;
- se o conjunto de dados tem distribuição simétrica, média, mediana e moda são idênticas;
- se um conjunto de dados tem uma distribuição assimétrica positiva, o valor da média é superior ao da mediana, que por sua vez é maior que a moda;
- se um conjunto de dados apresenta uma distribuição assimétrica negativa, o valor da média é menor do que o da mediana, que por sua vez é menor que a moda.

Figura 2: Simetria e assimetria em uma amostra.



Fonte: Adaptada de https://www.cesadufs.com.br/ORBI/public/uploadCatalago/14440324022014Bioestatistica_Aula_03.pdf

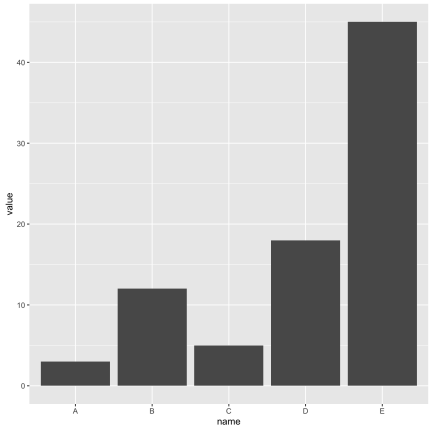
Medidas de Dispersão ou de Variabilidade:

- variância: mostra o quão distante cada valor do conjunto de dados está do valor central, e quanto maior a variância, mais heterogêneo é o conjunto de dados $\left(\hat{\sigma^2} = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n} \text{ ou } \hat{S^2} = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}\right)$;
- desvio-padrão: é dado pela raiz quadrada da variância, e maiores valores do desvio padrão indicam maior variação ($\sigma = \sqrt{\sigma^2}$ ou $S = \sqrt{S^2}$);
- coeficiente de variação: é a razão entre o desvio-padrão e a média, e quanto maior o coeficiente de variação, maior é a variabilidade dos dados $\left(CV = \frac{S}{\bar{X}} * 100\%\right)$.

Representações Gráficas:

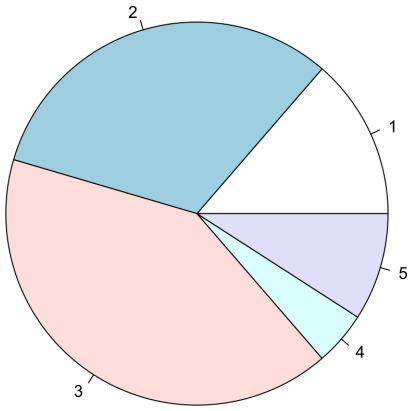
- gráfico de barras: usa barras de igual largura para mostrar as frequências (eixo vertical) das categorias dos dados (eixo horizontal);
- gráfico de setores: retrata dados como setores de um círculo, e cada setor é proporcional à frequência para a categoria representada;

Figura 3: Exemplo de gráfico de barras.



Fonte: r-graph-gallery.com/218-basic-barplots-with-ggplot2.html

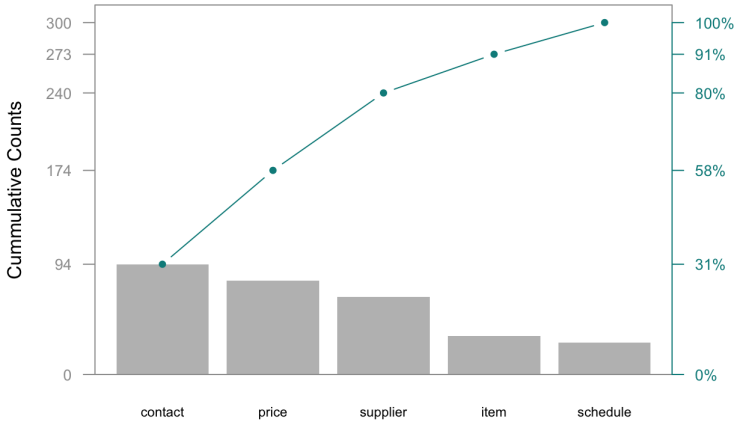
Figura 4: Exemplo de gráfico de setores.



Fonte: https://www.r-graph-gallery.com/131-pie-plot-with-r.html

- gráfico de Pareto: composto por um gráfico de barras que ordena as frequências das ocorrências em ordem decrescente, permite uma fácil visualização e identificação das categorias mais importantes (Princípio de Pareto $80\% - 20\%$);

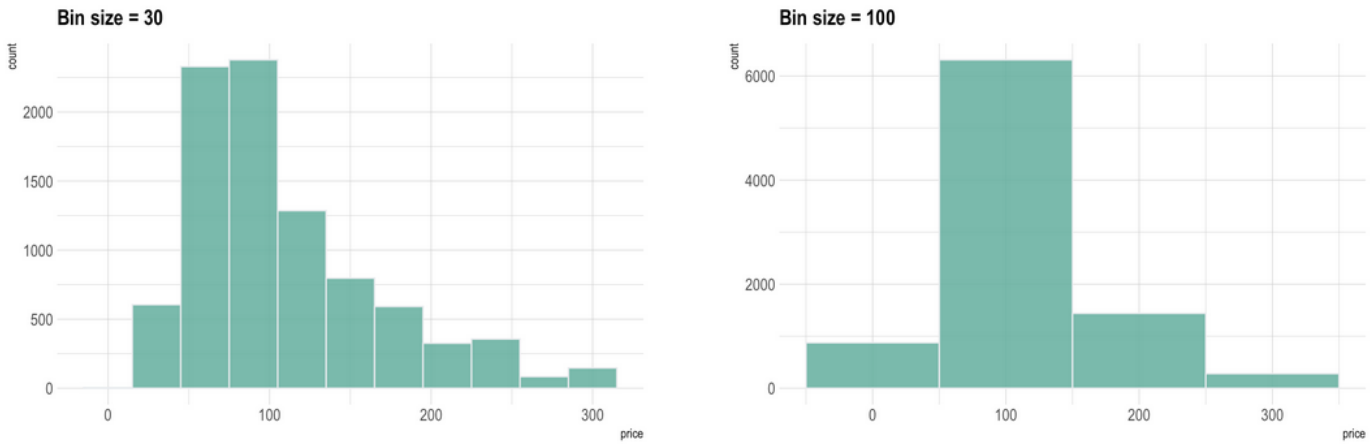
Figura 5: Exemplo de gráfico de Pareto.



Fonte: https://rstudio-pubs-static.s3.amazonaws.com/72023_670962b57f444c04999fd1a0a393e113.html

- histograma: versão gráfica da distribuição de frequência (escala vertical) por classes (escala horizontal), cujas barras são de mesma largura e desenhadas e adjacentes umas às outras;

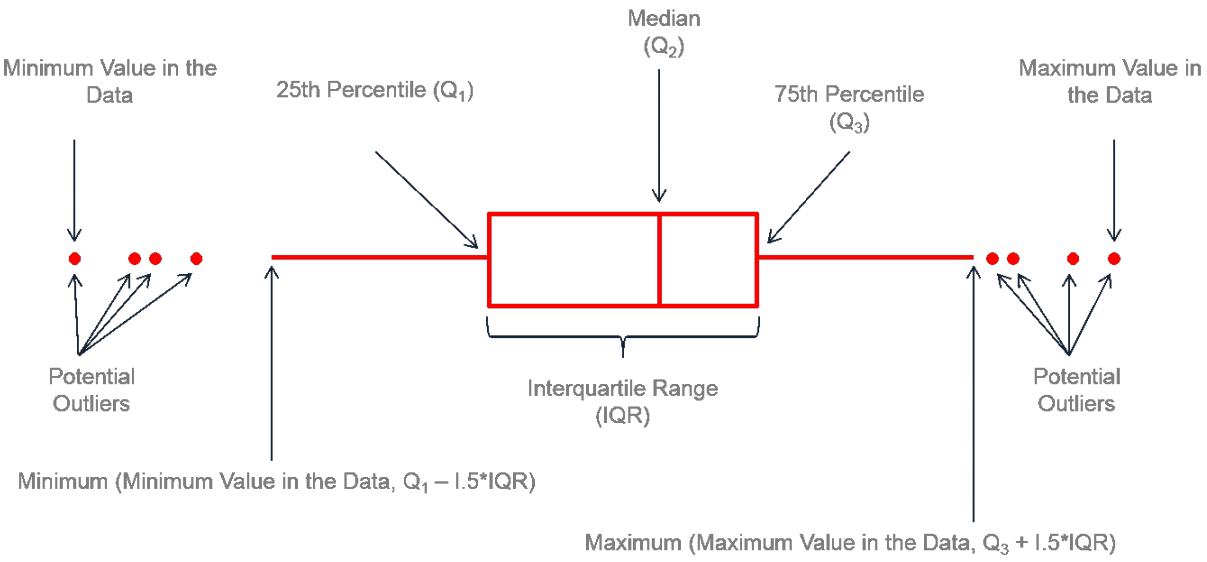
Figura 6: Exemplo de histograma.



Fonte: https://www.r-graph-gallery.com/220-basic-ggplot2-histogram.html

- box-plot: representa a variação de dados observados por meio de quartis; a caixa tem suas linhas traçadas no primeiro quartil (25%), na mediana (50%) e no terceiro quartil (75%), retas (ou bigodes) estendem–se a partir da caixa, indicando a variabilidade fora do quartil superior e do quartil inferior, e valores atípicos ou *outliers* são plotados como pontos individuais.

Figura 7: Box-plot.

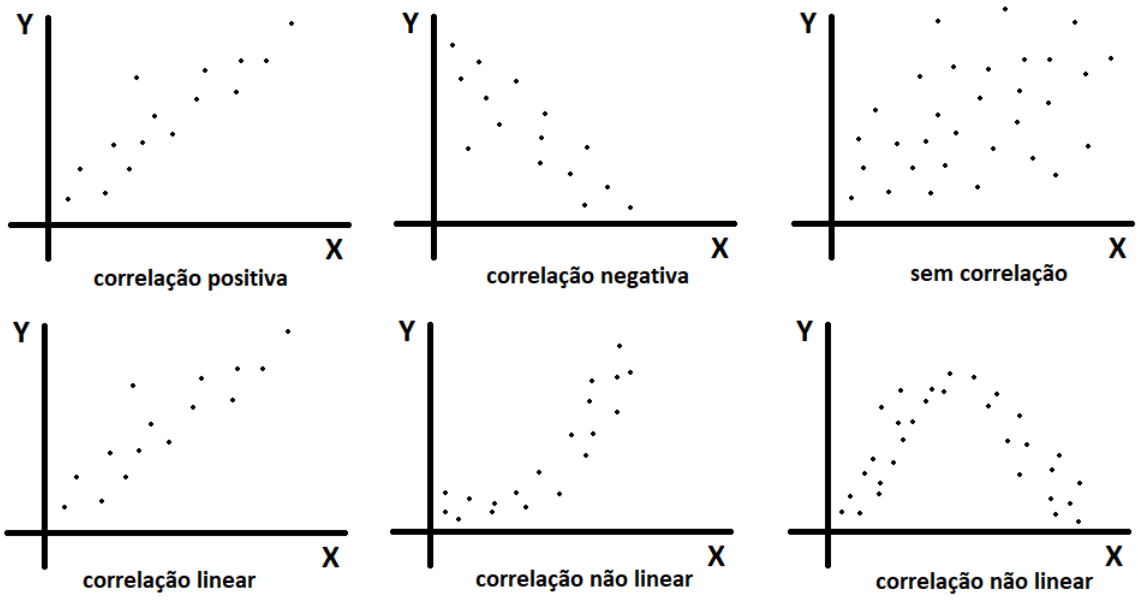


Fonte: https://www.r-graph-gallery.com/boxplot.html

Associação de Variáveis:

- diagrama de dispersão: é um gráfico de pares de dados quantitativos, e o padrão dos pontos marcados é usado para se determinar a existência, ou não, de alguma relação entre o par de variáveis;

Figura 8: Box-plot.



Fonte: Elaborada pelo autor

- coeficiente de correlação linear: medida de relação linear entre as variáveis quantitativas, calculada como a razão entre a covariância entre as duas variáveis e a raiz quadrada do produto das variâncias $\left(\rho = \frac{cov(X,Y)}{\sqrt{\mathbb{V}(X)\mathbb{V}(Y)}},\ \text{em que } cov(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \hat{x})(y_i - \hat{y})}{n-1}\right)$.