Advanced Institute for Artificial Intelligence

Introdução a Estatística

Advanced Institute for Artificial Intelligence

https://advancedinstitute.ai

Sumário

Introdução a Estatística

Estatística

- Ciência de aprendizagem a partir de dados.
- Métodos que auxiliam o processo de tomada de decisão.
- A Estatística está presente em todas as áreas da ciência que envolvam a coleta e análise de dados.

Introdução a Estatística

- Dados:
  - Observações de uma ou mais variáveis.
  - Variável é aquilo que se deseja observar para obter algum conhecimento, por ex., idade, sexo, peso e outras.

Introdução a Estatística

- Estatística: envolve coletar, classificar, resumir, organizar, analisar e interpretar dados.
- Envolve:
  - Descrever Conjuntos de Dados
  - Tirar conclusões (fazer estimativas, decisões, previsões, etc. a cerca de conjuntos de dados

Introdução a Estatística

- Estatística descritiva: descrever os dados
  - Coletar, Apresentar e caracterizar
- Estatística inferencial: tomar decisão baseado nas características da população
  - Estimativa, teste de hipótese

Introdução a Estatística

- Unidade experimental: objeto sobre o qual coletamos dados
- População: todos os itens de interesse
- Variável: um conjunto de valores registrados para uma unidade experimental individual
- Amostra: subconjunto de uma população

Introdução a Estatística

- Inferência estatística: estimativa ou previsão ou generalização sobre uma população com base nas informações contidas em uma amostra
- Medida de Confiabilidade: declaração (geralmente qualificada) sobre o grau de incerteza associado a uma inferência estatística

Introdução a Estatística

- As variáveis podem ser categóricas (qualitativas) ou numéricas (quantitativas)
  - Variáveis qualitativas: São características de uma população que não pode ser medidas.
    - Ordinais: Grau de gravidade de uma doença
    - Nominais: Presença de um sintoma
  - Variáveis quantitativas: São características de uma população que pode ser quantificadas.
    - Discretas: Número de cirurgias
    - Contínuas: Idade, Pressão Arterial

Introdução a Estatística

Obtendo dados para análises estatísticas:

- Fonte publicada: livro, jornal, jornal, site
- Experiência projetada: pesquisador exerce controle rígido sobre as unidades
- Pesquisa: um grupo de pessoas é pesquisado e suas respostas são registradas
- Estudo de observação: unidades são observadas em ambiente natural e variáveis de interesse são registradas

Introdução a Estatística

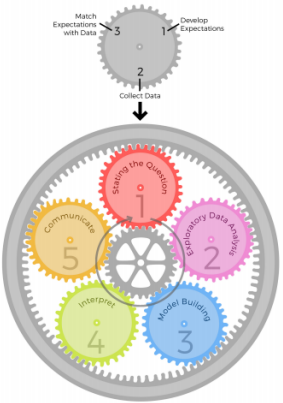
- Estatística é uma técnica essencial para responder questões científicas
- Para isso são realizados estudos experimentais ou observacionais
- O padrão de variação nos dados faz com que a resposta não seja óbvia

Amostragem dos dados:

- A amostra deve ser representativa do problema em estudo
- O tamanho da amostra deve ser suficiente para representar o problema
- Aleatoriedade da amostra: garantir que todos os elementos da população tenham chance de pertencer à amostra

Introdução a Estatística

Epíciclos de ciência dos dados [1]



[1] "The Art of Data Science A Guide for Anyone Who Works with Data". Roger D. Peng and Elizabeth Matsui. 2016

- Os epíciclos de ciência dos dados são etapas genéricas que podem ter complexidade variável
- Um projeto pode ser feito seguindo cada etapa em apenas um dia
- Projetos maiores podem demandar a execução de cada etapa por dias ou meses

Etapas do epicíclo:

- ☐ Definir a questão
- ☐ Análise de Dados Exploratória
- ☐ Construção de um modelo
- ☐ Interpretação
- ☐ Comunicação

Cada etapa segue um processo de alinhamento de expectativa

- Estabelecendo Expectativas
- Coletando informações (dados), comparando os dados com suas expectativas e, se as expectativas não corresponderem,
- Revise suas expectativas ou ajuste os dados para que seu dados e suas expectativas correspondam.

Forma de representação da freqüência de cada valor distinto da variável em estudo.

Juntamente com as freqüências simples, a tabela poderá ainda incluir:

- Frequências relativas: percentagem relativa à freqüência
- Frequências acumuladas: número de vezes que uma variável assume um valor inferior ou igual a esse valor.
- Frequências relativas acumuladas: percentagem relativa à freqüência acumulada

Listagem de modo de contração por colaborador

Nome	Modo de contratação
Bernita Lawhon	CLT
Caron Abernathy	PJ
Rita Figgins	PJ
Bulah Zackery	PJ
Ja Berber	CLT
Tennille Verrill	CLT
Francina Samaniego	PJ
Elinor Sowder	PJ
Shantay Crane	CLT
Suzan Coldwell	PJ

Tabela de frequência de contagem de tipo de contração

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

Elementos essenciais uma tabela

- Título: uma indicação que antecede a tabela e explique tudo referente a tabela.
- Cabeçalho: colocado na parte superior da tabela, especificando o conteúdo das colunas.
- Corpo: corresponde ao conjunto de colunas e de linhas que contêm informações sobre o fenômeno estudado.

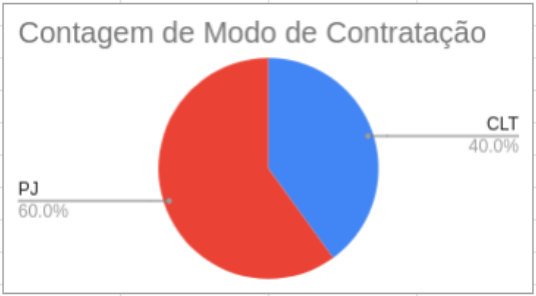
Recomendações quando a construção de gráficos

- Todo gráfico deve ter título, escala e fonte de dados
- As escalas devem crescer da esquerda para a direita e de baixo para cima.
- As distâncias que indicam as unidades devem ser uniformes.

Introdução a Estatística

Sumarização de variável qualitativa:

- Tabelas usando contagens ou porcentagens
- Gráfico de Barras ou Gráfico de Setores


Contagem de Modo de Contratação
CLT
40.0%
PJ
60.0%

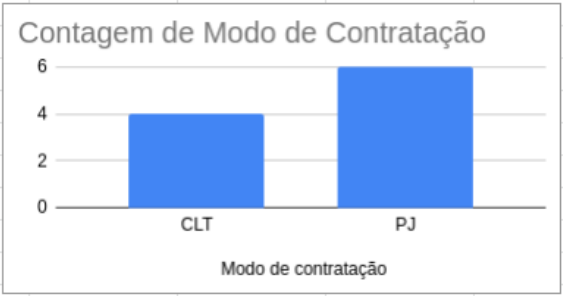

Contagem de Modo de Contratação
6
4
2
0
CLT
PJ
Modo de contratação

Tabela de variáveis quantitativas - Nível de consumo de alcool.

local	nível
Belarus	17.5
Moldova	16.8
Lithuania	15.4
Russia	15.1
Romania	14.4
Ukraine	13.9
Andorra	13.8
Hungary	13.3
Czech Republic	13.0
Slovakia	13.0

Introdução a Estatística

Sumarização de variável quantitativas:

- Tabelas de Freqüências
- Histograma

Critério para determinar a quantidade de classes:

$$k = 1 + 3.3 + log(n)$$

Amplitude das classes

$$a = (maiorValor - menorValor/numeroDeClasses)$$

no pandas a quantidade de classes é definida por série usando o parâmetro bins do método value_counts()

Histograma: Representação gráfica da distribuição das frequências absolutas ou relativas
Normalmente utilizado para variáveis contínuas.

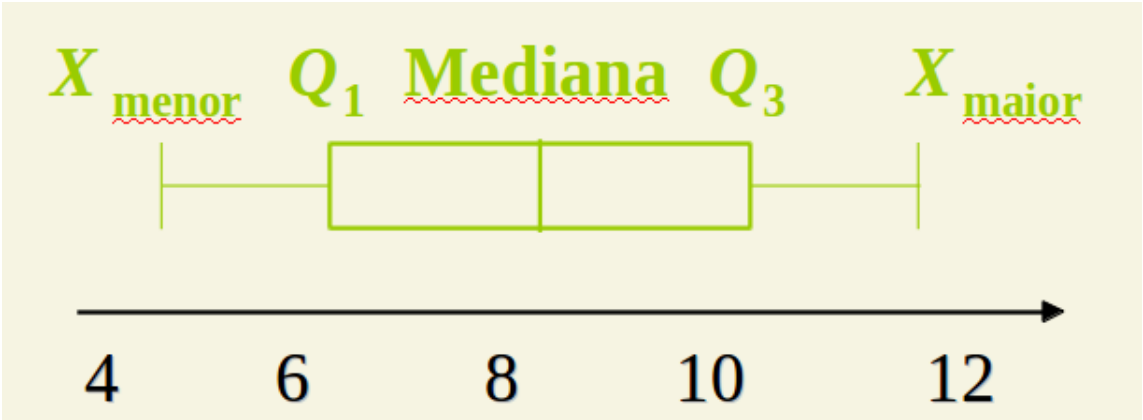
- As barras devem estar todas juntas
- Cada barra representa a freqüência de um intervalo de valores
- Os intervalos devem ter todos a mesma amplitude

Medidas de Tendência Central:

- Servem para termos uma idéia acerca dos valores médios da variável em estudo
- São usados para sintetizar em um único número os dados observados
- São exemplos de medidas de tendência central: Média, Moda e Mediana

Medidas Separatrizes separam a distribuição em partes iguais (depois de ordenadas).

- Quartis (quatro partes)
- Decis (dez partes)
- Percentis (cem partes)


X_{menor}
Q_1
Mediana
Q_3
X_{maior}
4
6
8
10
12

Gráficos do tipo boxplot representam gráficamente cinco medidas separatrizes:

- mínimo
- quartil inferior
- mediana
- quartil superior
- máximo

Esse gráfico é útil para identificar valores chamados *outliers*, que são valores muito altos ou muito baixos em relação ao restante da população

- Medidas de Variabilidade reflitem a variação dentro de um conjunto de dados
- Essas medidas serão pequenas se os dados forem próximos e grandes se eles estiverem muito espalhados.
- permitem comparar amostras de diferentes tamanhos e determinar se uma amostra é mais variável (ou heterogênea) que outra.

Introdução a Estatística

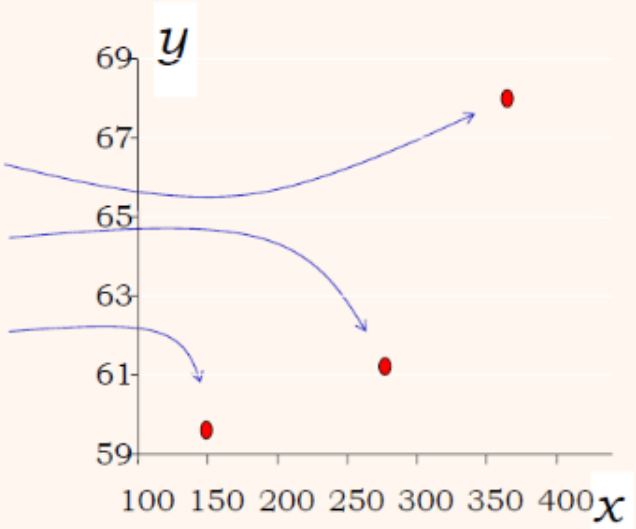
- Variância: é um indicativo da dispersão de um conjunto de dados em relação à média
- Desvio Padrão: Corresponde à raiz quadrada da variância
- A medida mais usada na comparação de diferenças entre grupos
- Quanto maior o desvio-padrão, maior a variabilidade dos dados

Análise MultiVariada

- Investigar relação entre variáveis
- Variáveis qualitativas
  - tabelas de freqüência com dupla entrada. Essas tabelas de dados cruzados são conhecidas por tabelas de contingência
- Variáveis quantitativas
  - Gráficos de Dispersão

Origem/grau de intrução	fundamental	médio	superior	total
Capital	11 %	14 %	6 %	31 %
Interior	8 %	19 %	6 %	33 %
Outra	14 %	17 %	6 %	36 %
Slovakia	33 %	50 %	17 %	100 %

x	y
365	67,99
278	61,19
150	59,58



Métodos para medir correlação entre variáveis disponíveis no Pandas

- pearson
- kendall
- spearman