

Análise exploratória I

Motivação e análise descritiva univariada para variáveis qualitativas e quantitativas.

Prof. Me. Lineu Alberto Cavazani de Freitas

CE003 – Estatística II

Departamento de Estatística
Laboratório de Estatística e Geoinformação



Análise exploratória

- ▶ Parte primordial de qualquer análise estatística é chamada análise descritiva ou exploratória.
- ▶ Consiste basicamente de tabelas, resumos numéricos e análises gráficas das variáveis disponíveis em um conjunto de dados.
- ▶ Trata-se de uma etapa de extrema importância e deve preceder qualquer análise mais sofisticada.
- ▶ As técnicas de análise exploratória visam resumir e apresentar as informações de um conjunto de dados brutos.

Análise exploratória



- ▶ Tentar compreender um conjunto de dados sem algum método que permita resumir as informações é inviável.
- ▶ A análise exploratória é a primeira forma de tentarmos entender o que acontece nos nossos dados.
- ▶ Compreende também a etapa de consistência dos dados, isto é, verificar se os dados coletados são condizentes com a realidade.

Análise exploratória

- ▶ O conjunto de técnicas aplicáveis está diretamente associado ao tipo das variáveis de interesse (quantitativas x qualitativas) e suas ramificações.
- ▶ Podemos conduzir análises focadas nas variáveis uma a uma (análises univariadas).
- ▶ Bem como conduzir análises focadas em avaliar a relação entre as variáveis (análises multivariadas).

Análise exploratória

Podemos fazer uso diversas técnicas, tais como

- ▶ Medidas de posição central.
- ▶ Medidas de posição relativa.
- ▶ Medidas de forma.
- ▶ Medidas de dispersão.
- ▶ Medidas de associação.
- ▶ Tabelas de frequência absolutas.
- ▶ Tabelas de frequência relativas.
- ▶ Tabelas de frequência acumuladas.
- ▶ Tabelas para múltiplas variáveis.
- ▶ Gráficos (para análise uni e multivariada).



Análise descritiva univariada para variáveis qualitativas

Análise descritiva univariada para variáveis qualitativas

- ▶ Uma variável qualitativa representa um atributo que pode ser expresso por meio de rótulos ou categorias.
- ▶ Podem ser classificadas em nominais (sem ordenação natural entre os níveis) ou ordinais (com ordenação natural entre os níveis).
- ▶ As categorias também são chamadas de classes ou níveis.
- ▶ Na análise descritiva de uma variável qualitativa estamos interessados em avaliar as frequências das classes.

Tipos de frequência

- ▶ Frequência absoluta: número de observações no conjunto de dados que pertence a uma determinada classe.
- ▶ Frequência relativa: frequência de classe dividida pelo número total de observações no conjunto de dados.
 - ▶ Pode ser apresentada em forma de percentual, quando multiplicada por 100.
- ▶ Frequência acumulada: frequência absoluta ou relativa acumulada conforme disposição das classes.
 - ▶ Não faz muito sentido para variáveis qualitativas nominais.

Exemplos



Tabelas de frequência para uma variável qualitativa

- ▶ Utilizando apenas os dados brutos é difícil responder questões de interesse.
- ▶ Para reduzir os dados originais de forma que fique mais claro o entendimento dos mesmos são utilizadas as tabelas de frequência.
- ▶ No caso de variáveis qualitativas consiste em listar os possíveis níveis da variável e fazer a contagem de quantas vezes cada nível aparece nos dados brutos.

Tabelas de frequência para uma variável qualitativa

- ▶ Cada linha da tabela diz respeito a um nível da variável categórica.
- ▶ As colunas podem apresentar diferentes tipos de frequência (absoluta, relativa, acumulada).
- ▶ Alguns cuidados para a apresentação dos resultados dizem respeito ao tipo de variável categórica em questão: nominal ou ordinal.

Tabelas de frequência para uma variável qualitativa

- ▶ Os níveis de variáveis nominais não apresentam uma ordenação natural, portanto, na apresentação dos resultados pode ser interessante ordenar os níveis por frequência ou por ordem alfabética.
- ▶ Esta estratégia não é recomendada para variáveis ordinais, pois estas apresentam uma ordenação natural e esta ordenação deve ser preferencialmente mantida na exposição dos resultados.

Tabelas de frequência para uma variável qualitativa

EXEMPLO QUALITATIVA NOMINAL

Tabelas de frequência para uma variável qualitativa

EXEMPLO QUALITATIVA ORDINAL

Gráficos para representação de frequências de uma variável qualitativa

- ▶ A representação por meio de tabelas é útil mas nem sempre eficiente.
- ▶ Em diversos casos pode ser mais conveniente utilizar um gráfico.
- ▶ “Uma imagem vale mais que mil palavras”.
- ▶ Os cuidados com a ordenação dos níveis de acordo com o tipo da variável se mantém.

Gráficos para representação de frequências de uma variável qualitativa

Algumas possibilidades são:

- ▶ Gráfico de barras verticais.
- ▶ Gráfico de barras horizontais.
- ▶ Gráfico de barras empilhadas.
- ▶ Gráfico de setores.
- ▶ Gráfico de rosca.

Gráfico de barras verticais ou horizontais

- ▶ Utiliza os possíveis níveis das variáveis em um eixo.
- ▶ As frequências ou porcentagens ficam no outro eixo.
- ▶ O tamanho da barra correspondente à frequência.

Gráfico de barras verticais

EXEMPLO



Gráfico de barras horizontais

EXEMPLO



Gráfico de barras empilhadas

- ▶ No caso de uma barra empilhada representa-se a frequência relativa ou o percentual.
- ▶ Existe uma única barra que representa 100%.
- ▶ Esta barra é dividida de acordo com a contribuição relativa de cada nível da variável.

Gráfico de barras empilhadas

EXEMPLO



Gráfico de setores

- ▶ Consiste em repartir um círculo em setores de tamanhos proporcionais às frequências relativas ou às porcentagens de cada valor.
- ▶ Podem ser usados para representar variáveis com poucos níveis.
- ▶ Para variáveis com muitos níveis, o gráfico tende a ficar visualmente poluído e pouco informativo.
- ▶ Além disso, o cérebro humano tem dificuldade em relacionar frequências com áreas relativas.
- ▶ Apesar de muito usado e preferido em diversas áreas, deve ser evitado.
- ▶ Uma variação do gráfico de setores é o gráfico de rosca.

Gráfico de setores

EXEMPLO





Análise descritiva univariada para variáveis quantitativas

Análise descritiva univariada para variáveis quantitativas

- ▶ Uma variável quantitativa é uma característica que pode ser representada numericamente.
- ▶ Podem ser classificadas em discretas (finitos valores em um dado intervalo) ou contínuas (infinitos valores em um dado intervalo).
- ▶ Quando estamos lidando com variáveis quantitativas discretas com poucos possíveis valores, as técnicas apresentadas para variáveis qualitativas se aplicam.

Tabelas de frequência

EXEMPLO TABELA PARA QUANTITATIVA DISCRETA COM POUCOS NÍVEIS

Gráficos de frequência

EXEMPLOS GRÁFICOS PARA QUANTITATIVA DISCRETA COM POUCOS NIVEIS

Análise descritiva univariada para variáveis quantitativas

- ▶ Para variáveis quantitativas contínuas ou discretas com muitos possíveis valores, precisamos de técnicas específicas.
- ▶ Uma estratégia comum é o agrupamento em faixas de valores, e avaliação das frequências nestas faixas.
- ▶ Cuidados devem ser tomados quanto às notações e tipos de faixas (aberto e fechado à esquerda ou direita).
- ▶ Podem ser usadas tabelas de frequências absolutas, relativas e acumuladas para as faixas de valores.
- ▶ Utilizando a razão entre frequência relativa e a amplitude das faixas de valores, geramos a densidade.

Análise descritiva univariada para variáveis quantitativas

- ▶ Como agrupar em classes?
- ▶ Qual o tamanho ideal das faixas de valores?
- ▶ Classes definidas com a mesma amplitude é o procedimento mais usual.
- ▶ Existem procedimentos que podem ser usados para obter a amplitude, como Sturges.

Tabelas de frequência para uma variável quantitativa

EXEMPLO



Gráficos para representação de frequências de uma variável quantitativa

- ▶ Assim como no caso de variáveis qualitativas ou quantitativas discretas com poucos possíveis valores, a representação por meio de gráficos pode ser bastante benéfica para análise de variáveis quantitativas.

Algumas possibilidades são

- ▶ Histograma da frequência absoluta (afetado pelo número de classes).
- ▶ Histograma com amplitude de classe variável.
- ▶ Histograma da densidade.
- ▶ Histograma da densidade para amplitude de classe variável.
- ▶ Gráfico de frequências acumuladas.
- ▶ Gráfico de densidade empírica.
- ▶ Box-plot

Histograma

- ▶ Consiste em retângulos contíguos de base dada pelas faixas de valores definidas para uma variável.
- ▶ A área igual é igual à frequência da respectiva faixa.
- ▶ Em uma possível representação a altura pode representar a frequência absoluta na faixa de valores.
- ▶ Outra possibilidade é a altura de cada retângulo representar o quociente da área pela amplitude da faixa: a densidade.

Histograma

EXEMPLO



Gráfico de densidade empírica



Assimetria



Gráfico de frequências acumuladas



Box-plot

- ▶ Outra importante visualização é o box-plot.
- ▶ É possível analisar a distribuição dos dados, aspectos quanto a posição, variabilidade, assimetria e também a presença de valores atípicos.
- ▶ Retomaremos o box-plot após estudar quartis, em medidas descritivas.

EXEMPLO BOX-PLOT