Unidade 2 - Estatística descritiva

Distribuições de frequência

Há muitas maneiras de se organizar e descrever um conjunto de dados, algumas características importantes são centro, variabilidade, forma.

Quando um conjunto de dados é extenso, pode ser difícil de observar padrões. É conveniente agrupar as informações em intervalos chamados classes, formando uma distribuição de frequência.

Distribuição de frequência:

É uma tabela que mostra classes ou intervalos dos valores com a contagem do numero de ocorrencias em cada classe ou intervalo.

A frequência F de uma classe é o número de ocorrencias de dados na classe

Amplitude da classe:

Distância entre os limites inferiores (ou superiores) de classes consecutivas

Amplitude da distribuição:

Diferença entre os valores maximos o mínimos

Construção de uma Distribuição de Frequência com base em um conjunto de dados:

1. Decida o número de classes para serem incluídas na distribuição de frequência.
2. Encontre a amplitude de classe: determine a amplitude dos dados, divida a amplitude pelo numero de classes e arredonde para um número próximo conveniente
3. Encontre os limites de classe. Estabeleça um limite inferior usando o menor valor de todos os dados, apartir desse adicione a amplitude de classe ao limite inferior da classe procedente e então encontre os limites superiores.
4. Faça uma marca de contagem para cada registro na linha da classe apropriada
5. Conte as marcas para encontrar a frequência total F para cada classe.

É possível incluir novas colunas com características adicionais das classes que auxiliam na interpretação dos dados (ponto médio, frequência relativa e aumulada)

Ponto médio:

É a soma dos limies de uma classe diividida por dois, pode se chamado de marca da classe

Frequência relativa:

É a fração, ou poporção de dados que está nessa classe. Para calcula-la basta dividir a frequência F pelo tamanho N de uma amostra, para representa-la em percentagem basta multiplica-la por 100.

Frequência acumulada:

Trata da soma das frenquências dessa classe com todas as anteriores. A frequência acumulada da última classe é igual ao tamanho N da amostra

Gráficos de distribuições de frequência

Histograma de frequência

É um diagrama de barras que representa a distribuição de frequência de um conjunto de dados. Ele possui as seguintes propriedades:

1. A escala horizontal é quantitativa e indica os valores dos dados
2. A escala vertical indica as frequências da das classes
3. Barras consecutivas devem estar encostadas umas nas outras

As fronteiras das classes (limites reais) são os números que separam as classes sem formar lacunas entre elas

É possível guardar o eixo horizontal co os pontos médios ou com fronteiras estabelecidas para as classes

Outra maneira de representar graficamente uma distribuição de frequência é usar um polígono de frequência.Um polígono de frequência é um gráfico de linha que enfatiza as mudanças contínuas nas frequências.

São utilizadas as mesmas escalas horizontals e vertical do histograma com os pontos médios e conecte-os em uma única linha.

Preço (em dólares) 
Histograma de frequência 

Histograma de frequência relativa:

Tem a mesma forma e escala horizontal do correspondente histograma de frequência. A diferença é que a escala vertical indica as frequências relaitivas, e não as absolutas.

0.30 
0,25 
0,20 
0,10 
0,05 
64,5 
Histograma de frequência relativa 
30% dos 
navegadores GPS 
têm preços entre 
US$ 104,50 e 
US$ 144,50 
104,5 
144,5 
184,5 
224,5 
264,5 
304,5 
344,5 

Gráfico de frequênia acumulada (ogiva):

GRáfico de linhas que mostra a frequência acumulada até cada classe em sua fronteira superior. As fronteiras superiores são marcadas no eixo horizontal e as frequências acumuladas são marcadas no eixo vertical.

Usadas para descrever o número de registros de dados que são menores que ou iguais a um determinado limite, utiliza-se um gráfico de frequência acumulada

Construindo uma ogiva:

1. Construa uma distribuição de frequência que inclua uma coluna com as frequências acumuladas
2. Especifique as escalas verticais e horizontais. A escala horizontal consiste nas fronteiras superiores das classes e a escala vertical indica as frequências acumuladas
3. Assinale os pontos que representam as fronteiras superiores das classses e as frequências acumuladas correspondentes
4. Conecte os pontos em ordem da esquerda para a direita com segmentos lineares
5. Ográfico deve começar na fronteira inferior da primeir classe (a frequência acumulada é igual ao tamanho da amostra)

Preço de navegadores GPS 
30 
645 
104,5 
144,5 1845 224,5 264,5 
Preso (em dólares) 
304 
344,5 

Conjuntos de dados qualitativos

Gráficos de pizza (setoriais):

Fornecem uma maneira conveniente de apresentar graficamente dados qualitativos como percentagens de um todo

É um círculo dividido em setores que representam categorias. Aárea de cada setor é proporcional à frequência de cada categoria

Montando uma gráfico setorial:

1. Calcule o valor, ou percentagem de cada categoria
2. Construa um gráfico usando o angulo central que corresponda a cada categoria (para determinar o angulo central, multiplique afrequência relativa da categoria por 360)

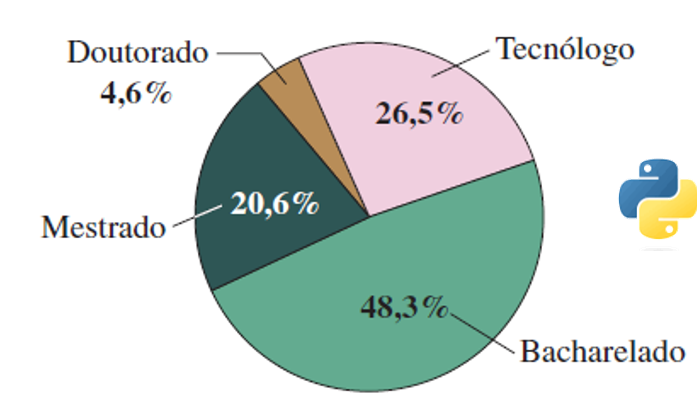


Gráfico de pareto:

É um gráfico de barras verticais no qual a altura de cada barra representa a frequência ou a frequência relativa

As barras são posicionadas em ordem decrescente de altura, a mais altafica á direita

Principais causas de redução de estoque. 
16 
8 
2 
Roubo por 
funcionários 
Fraude Desconhecida 
em lojas administrativo nas vendas 
E comum acrescentar um eixo secundário à direita, 
representando a frequência acumulada 

Conjuntos de dados emparelhados

Quando cada valor de um conjunto de dados corresponde a um valor em um segundo conjunto de dados, tais conjuntos são chamados de conjuntos de dados emparelhados

Uma maneira de reprsentar esses conjuntos é usando um gráfico de dispersão, onde os pares ordenados são representados como pontos em um plano coordenado

Série temporal:

Um conjunto de dados quantitativos cujos valores são obtidos em intervalos regulares, durante um período de tempo

Um gráfico de série temporal tem em seu eixo horizontal o período de tempo considerado e no eixo vertical a grandeza observada

Assinantes de telefonia celular 
325 
— 275 
175 
150 
125 
100 
75 
2002 
2003 
2004 
2005 
2006 
2007 
2008 
2009 
2010 
2011 
2012 

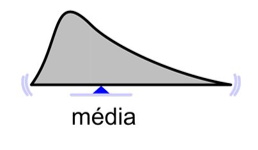
Medidas de tendência central

É um valor que representa uma observação típica ou central de um conjunto de dados

Média:

Conjunto de dados é a soma dosvalores dos dados dividida pelo número de observações. Para determinar a média de um conjunto de dados, use uma das formulas a seguir

É afetada por outliers

Ех 
M6dia popuIacionaI: — 
M6dia amostra[: Х 

Mediana:

A mediana de uma conjunto de dados éum valor que está no meio dos dados quando o conjunto está ordenado. A mediana indica o centro de um conjunto de dados ordenado, dividindo-o em duas partes com quantidades iguais de valores.

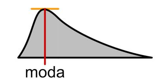
Quando o númerode elementos éimpar a mediana é o elemento central, já quando é par ela é a média dos dois valores centrais.

Nao é afetada por outliers

% 50% 
mediana 

Moda:

É o valor que ocorre com a maior frequência com a maior frequência. Um conjunto de dados pode ter uma moda, ou não ter moda ou ter mais de um moda.Caso nenhum valor se repita, o conjunto naotem moda. Quando dois valores ocorrem com a mesma frequência, cadauma é uma moda e o conjunto é bimodal



Outlier:

É um valor que está muito afastado dos demais valores do conjunto de dados

A média pode ser muito afetadapor esses valores

Alguns outliers são válidos, mas outros podem ser resultados de falhas nos registros de dados

Um conjunto de dados pode ter umou mais outliers e as conclusões obtidas a partir de conjuntos com outliers podem ser falhas

Média ponderada:

É a medida de um conjunto de dados cujos valores tem pesos variados



Média de distribuição de cada frequência:

ΣΙΧ. η 
MOta = 
e de Cade 
respecttvamer-.to_ 

Formas das distribuições

Distribuição de frequência simétrica:

Quando uma linha vertical pode ser desenhada pelo meio do gráfico da distribuição e as metades resultantes são imagens espelhadas

* Média mediana e moda são iguais

(a) Distribuição simétrica. 
35 
30 
9 11 
Média 
Mediana 
Moda 
13 
15 

Distribuição de frequência uniforme/retangular:

Quando todos os valores ou classes na distribuição têm frequências iguais ou aproximadamente iguais, também é uma distribuição simétrica.

(b) Distribuiqäo uniforme (simétrica). 
35 
30 
20 
15 
10 
13 15 
Média 
Mediana 

Distribuição de frequência assimetrica:

Quando a cauda do gráfico se alonga mais para um dos lados

Assimetria negativa: alongamento para a esquerda

* A média é menor que a mediana,que é menor que a moda

(с) Distribuisio assimbtrica esquerda. 
15 
Мое1а 

Assimetria positiva: alongamento para a direita

* A média é maior que a mediana que é maior que a moda

Medidas de variação

Medida de variação:

É um valor que representa a disperção de um conjunto de dados

Amplitudae:

A amplitude de um conjuto de dados é definida pela diferença entre o maior e menor valor (quantitativos) daquele conjunto.

Desvio:

O desvio de um valor X em uma população é a diferença entre o valor e a média do conjunto de dados.



A soma dos desvios de qualquer conjunto de dados é zero. Por não fazer sentido encontrar a média dos desvios, calculamos a soma dos quadrados dos desvios SSx

Variância populacional:

É a média dos quadrados dos desvios de uma população. Uma desvantagem é que sua unidade de medida é diferente da unidade do conjunto de dados, é na verdade o quadrado dessa medida.



Desvio padrão populacional:

É a raiz quadrada da variância pupolacional.

Ele é usado para medir a variação em média, do quanto os valores se desviam da média desse conjunto. Tem a mesma unidade de medida que o conjunto de dados.

É sempre maior ou igual a zero, quando ela é igual a zero significa que o conjunto não apresenta variação.

À medida que os valores se afastam da média o valor da disperção aumenta.

Pode ser util para comparar a variação em conjuntos de dados que tem a mesma unidade de medida e médias aproximadamente iguais.

Desvios incomuns, cujo desvio é maior que 3, e eles tem influência maior no desvio padrão.



Variância amostral e desvio padrão amostral:

Para calcular usa-se x ao invés de µ além disso SSx é dividido por n-1 ao invés de N.





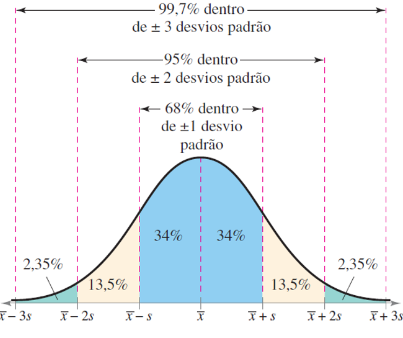
Empírico:

Que se baseia somente na experiência ou observação, se guia sem levar em consideração teorias ou métodos científicos.

Regra Empírica:

Para conjuntos de ddos que são aproximadamente simétricas e com forma de sino, o desvio padrão tem essas características:

* Cerca de 68% dos dados encontram-se dentro do intervalo de +-1 desvio padrão em relação á medida
* Cerca de 95% dos dados se encontram dentro do intervalo de +- dois desvios padrão em relação á média
* Cerca de 99,7% dos dados se encontram dentro do intervalo de +- 3 desvios padrão em relção á média



Quando um distribuição de frequência tem classes intervalores, é possível estimar a média e o desvio padrão amostral usando o ponto de cada classe

Coeficiente de variação:

Conjunto de dados descreve o desvio padrão como porcentagem na média

População: CV= 
• 100% 
Amostra: CV 
S 
• 100% 
x 

Mede a variação de um conjunto de dados com relação à média.

Usado para comparar conjuntos de dados com inudades de medida diferentes ou médias diferentes.

Medidas de Posição

Separtrizes:

Números que partilham, dividem, conjunto de dados ordenado em partes iguais (com mesmo número de elementos)

A mediana por exemplo é uma separatriz pois divide um conjunto de dados ordenado em duas partes com quantidades iguais de elementos

Ajuda a especificar a posição de um elemento dentro de um conjunto de dados

Quartis:

Os tres quartis dividem um cojunto de dados ordenado em quatro partes iguais. É um ponto específico

¼ recai sobre/abaixo de primeiro quartil Q1

½ recai sobre/abaixo do segundo quartil Q2

¾ recai sobre ou abaixo do terceiro quartil Q3

25% 25% 25% 25% 
(minimo) 
(màximo) 
(mediana) 

Mediana:

Medida de tendência central baseada na posição.

A medida de variação que é baseada na posição é a amplitude interquartil, indica a dispersão dos 50% centrais

Amplitude interquartil:

Medida de variação que fornece a amplitude da porção central dos dados.

A AIQ é a diferença entre o terceiro e o primeio quartil

AIQ = Q3- Q1

Outlier:

Um valor que está muito distante dos deamis uma maneira de identificar um outlier é usar a amplitude inter quartil.

Apliacações de quartis:

Representar conjuntos de ddos usando diagramas de boxplot

É um ferramenta de análise exploratória que destaca características importantes de um conjunto de dados

Percentis:

Divide o conjunto de dados em 100 partes iguais

O percentil de um valor específico X:

Escorre padrão:

Representa o número de desvios padrão em que um valor x encontra-se

Se: z < 0 <=> x < u

z > 0 <=> x > u

z = 0 <=> x = u

Pode ser usado para identificar valores incomuns de um conjunto de dados em formato de sino