Estatística

Estatística Descritiva

Felipe Álvares
felipealvares@decom.cefetmg.br

Cefet - MG

2017

Sumário

Variáveis

- Variável é a característica de interesse que é medida em cada elemento da amostra ou população.
- Seus valores variam de elemento para elemento.
- Podem ser característicos numéricos sob os quais operações aritméticas podem ser realizadas como salário, altura, ou tempo de vida; ou podem ser atributos como sexo, zona de moradia ou classe social.

Tipos de Variáveis

A grosso modo, variáveis podem ser classificadas como quantitativas ou qualitativas de acordo a característica representada.

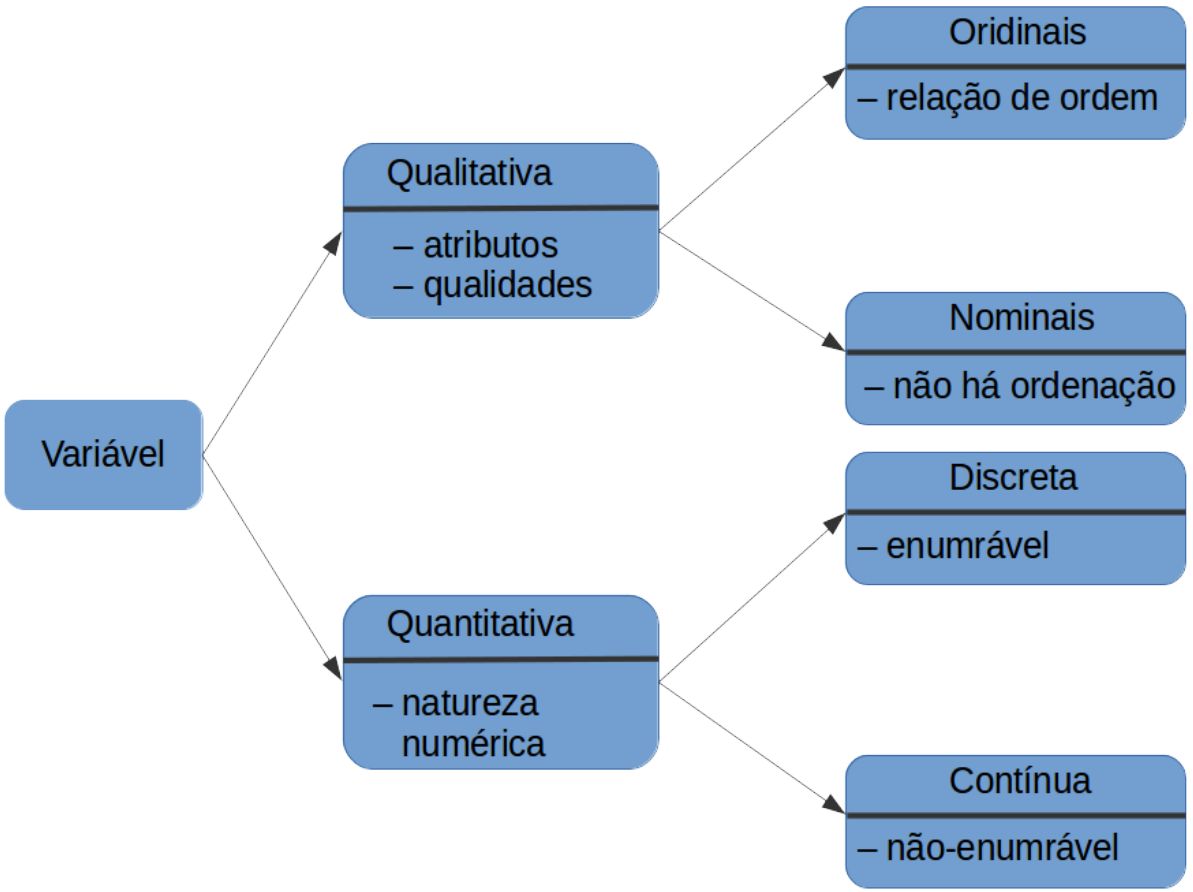
- **Quantitativas** – são características que podem ser medidas em uma escala quantitativa, ou seja, apresentam valores numéricos que nos permite efetuar operações aritméticas diretamente. Podem ser contínuas ou discretas.
  1. **Variáveis discretas:** características mensuráveis que podem assumir apenas um número finito ou contável de valores. Geralmente são o resultado de contagens: números de filhos, número de bactérias por litro de água, número de peças defeituosas por lote, etc.
  2. **Variáveis contínuas:** características mensuráveis que assumem valores em uma escala contínua (na reta real). Usualmente devem ser medidas através de algum instrumento: peso, altura, tempo, etc.

Tipos de Variáveis

- **Qualitativas (ou categóricas)** – são características definidas por categorias, ou seja, representam uma classificação dos indivíduos. Podem ser nominais ou ordinais.
  1. **Variáveis nominais:** não existe ordenação entre as categorias em estudo: sexo (F ou M), cor dos olhos (castanhos, azuis, verdes), fumante (sim ou não), turma (diurno ou noturno), etc.
  2. **Variáveis ordinais:** existe uma ordenação entre as categorias em estudo: escolaridade (ensino médio, técnico, superior), tamanho (pequeno, médio, grande), classe social (baixa, média, alta).

Obs.: variáveis qualitativas **podem** ser tratadas como variáveis quantitativas discretas.

Tipos de variáveis



Organização dos dados

- Em um primeiro estágio, conjunto de dados são normalmente apresntados em forma de tabelas brutas ou listas.
- Pouco pode ser inferido de maneira imediata nestes casos.
- Algumas formas de organização podem ajudar a evidenciar características relevantes.
- Dentre as principais metodologias destacam-se a construção de *diagramas de ramos e folhas* e *tabelas de frequências*.

Diagramas de ramos e folhas

- Busca oferecer uma apresentação simultaneamente visual e informativa dos dados.
- Aplica-se a variáveis quantitativas cujos valores têm ao menos dois dígitos.
- **Construção**
  1. seja $x_1, x_2, ..., x_n$ um conjunto qualquer de valores observados;
  2. divida cada número x_i em duas partes:
     - 2.1 um **ramo** consistindo de um ou mais dígitos;
     - 2.2 uma **folha** consistindo dos dígitos restantes;
  3. liste os valores do ramo em uma coluna vertical;
  4. ao lado do ramo, registre a folha de cada observação;
  5. descreva, em uma coluna mais à direita, o número de folhas contidas em cada ramo.

Exemplo

Os dados a seguir correspondem à resistência a compressão (medida em libra-força por polegada quadrada) de 80 corpos de prova de uma certa liga de alumínio em estudo.

105	221	183	186	121	181	180	143
97	154	153	174	120	168	167	141
245	228	174	199	181	158	176	110
163	131	154	115	160	208	158	133
207	180	190	193	194	133	156	123
134	178	76	167	184	135	229	146
218	157	101	171	165	172	158	169
199	151	142	163	145	171	148	158
160	175	149	87	160	237	150	135
196	201	200	176	150	170	118	149

Exemplo (continuação)

ramo	folha	frequência
7	6	1
8	7	1
9	7	1
10	5 1	2
11	5 8 0	3
12	1 0 3	3
13	4 1 3 5 3 5	6
14	2 9 5 8 3 1 6 9	8
15	4 7 1 3 4 0 8 8 6 8 0 8	12
16	3 0 7 3 0 5 0 8 7 9	10
17	8 5 4 4 1 6 2 1 0 6	10
18	0 3 6 1 4 1 0	7
19	9 6 0 9 3 4	6
20	7 1 0 8	4
21	8	1
22	1 8 9	3
23	7	1
24	5	1

A partir do diagrama observa-se de imediato que:

1. a maioria das resistências está entre 120 e 200;
2. o valor central está entre 150 e 160;
3. os valores são distribuídos de forma aproximadamente simétrica.

Diagramas de ramos e folhas

- Em algumas aplicações, pode ser interessante prover mais ramos.
- Consideremos, por exemplo, o seguinte diagrama obtido de um segundo conjunto de dados:

ramo	**folha**
6	1 3 4 5 5 6
7	0 1 1 3 5 7 8 8 9
8	1 3 4 4 7 8 8
9	2 3 5

Exemplo

- Cada ramo poderia ser subdividido em dois outros ramos:
  - um contendo apenas folhas inferiores ou iguais a 4;
  - outro contendo apenas folhas superiores ou iguais a 5;

ramo	**folha**
6L	1 3 4
6U	5 5 6
7L	0 1 1 3
7U	5 7 8 8 9
8L	1 3 4 4
8U	7 8 8
9L	2 3
9U	5

Exemplo (continuação)

- Outra alternativa seria subdividir cada ramo em cinco outros ramos:
  - ramo A contendo as folhas 0 e 1;
  - ramo B contendo as folhas 2 e 3;
  - ramo C contendo as folhas 4 e 5;
  - ramo D contendo as folhas 6 e 7;
  - ramo E contendo as folhas 8 e 9;

Exemplo (continuação)

ramo	folha
6A	1
6B	3
6C	4 5 5
6D	6
6E	
7A	0 1 1
7B	3
7C	5
7D	7
7E	8 8 9
8A	1
8B	3
8C	4 4
8D	7
8E	8 8
9A	
9B	2 3
9C	5
9D	
9E	

Diagramas de ramos e folhas

- Em geral, devemos escolher um número de ramos relativamente pequeno em comparação com o número de observações.
- Muitos ramos podem gerar uma perda significativa de informação em uma primeira análise visual.
- É comum considerarmos algum valor entre 5 e 20 ramos.

Tabelas de frequências

- Oferecem um resumo mais compactado dos dados na comparação com os diagramas de ramos e folhas.
- Ajudam a explicar a distribuição dos dados empiricamente.
- A idéia é agrupar os valores observados em classes e em seguida indicar suas frequências (absolutas ou relativas).

Tabelas de frequências – dados discretos

- No caso discreto, cada observação constitui naturalmente uma classe (uma linha da tabela).
- Basta então contar o número de repetições de cada um dos diferentes valores observados e agrupá-los em uma tabela.

Exemplo: Em um questionário estudantil, avaliou-se sexo, idade, altura, peso, número de filhos, hábito de fumar (sim ou não) e tolerância ao cigarro (indiferente, incomoda pouco e incomoda muito) dos 25 indivíduos de uma turma.

Identificação	Sexo	Idade	Altura	Peso	Filhos	Fuma	Toler
1	F	17	1,60	60,5	2	NAO	P
2	F	18	1,69	55,0	1	NAO	M
3	M	18	1,85	72,8	2	NAO	P
4	M	25	1,85	80,9	2	NAO	P
5	F	19	1,58	55,0	1	NAO	M
6	M	19	1,76	60,0	3	NAO	M
7	F	20	1,60	58,0	1	NAO	P
8	F	18	1,64	47,0	1	SIM	I
9	F	18	1,62	57,8	3	NAO	M
10	F	17	1,64	58,0	2	NAO	M
11	F	18	1,72	70,0	1	SIM	I
12	F	18	1,66	54,0	3	NAO	M
13	F	21	1,70	58,0	2	NAO	M
14	M	19	1,78	68,5	1	SIM	I
15	F	18	1,65	63,5	1	NAO	I
16	F	19	1,63	47,4	3	NAO	P
17	F	17	1,82	66,0	1	NAO	P
18	M	18	1,80	85,2	2	NAO	P
19	F	20	1,60	54,5	1	NAO	P
20	F	18	1,68	52,5	3	NAO	M
21	F	21	1,70	60,0	2	NAO	P
22	F	18	1,65	58,5	1	NAO	M
23	F	18	1,57	49,2	1	SIM	I
24	F	20	1,55	48,0	2	SIM	I
25	F	20	1,69	51,6	2	NAO	P

Exemplo (continuação)

Tabelas de frequência absoluta para as variáveis sexo, idade, número de filhos, hábito de fumar e tolerância ao cigarro:

Sexo	**freq.**
F	20
M	5
Total	25

Idade	**freq.**
17	3
18	11
19	4
20	4
21	2
25	1
Total	25

Filhos	**freq.**
1	12
2	8
3	5
Total	25

Fuma	**freq.**
S	5
N	20
Total	25

Toler	**freq.**
I	6
P	10
M	9
Total	25

Tabelas de frequências – caso contínuo

- Variáveis contínuas assumem valores em conjuntos não-enumeráveis; logo, é pouco provável, mas não necessariamente impossível, observarmos valores repetidos.
- Neste caso, uma tabela de frequências pode ser obtida através de um agrupamento em classes (intervalos).
- O número de classes é escolhido de acordo com a quantidade de observações.
- Uma prática usual consiste de escolher o número de classes como aproximadamente $\sqrt{n}$, onde n é o número total de observações.
- Uma vez especificado o número de classes, o limite de cada classe é escolhido de modo a garantir homogeneidade nas amplitudes.

Exemplo

- Consideremos o questionário estudantil do exemplo anterior.
- A fim de construir uma tabela de frequências podemos considerar $k = 5$ classes já que $\sqrt{25} = 5$.
- A amplitude das classes, com respeito às variáveis altura e peso, serã0 dadas respectivamente por:

$$\text{amplitude}_{alt} = \frac{\text{amplitude total}}{k} = \frac{1,85 - 1,55}{5} = 0,06;$$

$$\text{amplitude}_{peso} = \frac{\text{amplitude total}}{k} = \frac{85,2 - 47,0}{5} = 7,64.$$

- Intervalos resultantes:

$$\text{altura} - [1,55; 1,61),\ [1,61; 1,67),\ [1,67; 1,73),\ [1,73; 1,79),\ [1,79; 1,85];$$

$$\text{peso} - [47,0; 54,64),\ [54,64; 62,28),\ [62,28; 69,92),\ [69,92; 77,56),\ [77,56; 85,2].$$

- Agora basta contar o número de ocorrências em cada intervalo.

Exemplo (continuação)

Tabelas de frequência relativa diferem do caso anterior pelo fato de dividirem as frequências absolutas pelo número de observaçõs, oferecendo assim uma noção de proporção. Tabelas para as variáveis altura e peso:

Altura	**freq.**
$1,55 \vdash 1,61$	0,2
$1,61 \vdash 1,67$	0,28
$1,67 \vdash 1,73$	0,28
$1,73 \vdash 1,79$	0,08
$1,79 \vdash\dashv 1,85$	0,16
Total	1,00

Peso	**freq.**
$47,00 \vdash 54,64$	0,32
$54,64 \vdash 62,28$	0,40
$62,28 \vdash 69,92$	0,12
$69,92 \vdash 77,56$	0,08
$77,56 \vdash\dashv 85,2$	0,08
Total	1,00

Resumo dos dados

- Objetivo: descrever numericamente, e de forma sucinta, características importantes dos dados.
- Em algumas situações, o simples armazenamento de medidas resumo é suficiente para o desenvolvimento da análise estatística de um problema.
- Resumos incluem basicamente medidas de posição, separação e dispersão.

Medidas de posição

- Medidas de posição indicam alguma tendência central ou comportamento esperado com respeito extraídos de uma amostra qualquer.
- Dentre as principais medidas, destacam-se: media, mediana e moda.

Média aritmética

- Sejam $x_1, x_2, ..., x_n$ n observações de um fenômeno aleatório qualquer. A quantidade

$$\overline{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i, \tag{1}$$

é denominada a média aritmética da amostra.

Média aritmética

- Pode ser interpretada como o centro de massa (baricentro das observações).

- Caso os dados estejam organizados em uma tabela de frequências, isto é, se para cada x_i foi verificado um número f_i relativo à sua frequência, então

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i}. \tag{2}$$

Exemplo 1

Em um estudo informal, desejamos consultar a opinião de 30 colegas a respeito do perfil social da população de uma dada cidade. Cada colega deve responder se acredita que a maioria da população é de classe baixa (B), média (M) ou alta (A). As respostas foram sumarizadas na seguinte tabela de frequências:

classe	**freq**
B	5
M	10
A	15
Total	30

Baseando apenas na opinião dos colegas, qual deve ser a tendência do resultado?

Exemplo 1 (continuação)

- Variáveis descritivas podem, em alguns casos, ser associadas a valores discretos a fim de tornar plausível a realização de operações aritméticas.
- Considerando $B = 0$, $M = 1$ e $A = 2$, temos que

$$\overline{x} = \frac{5 \cdot 0 + 1 \cdot 10 + 15 \cdot 2}{30} = 1,\overline{3}.$$

- Neste caso, temos uma tendência mais favorável à classe M.

Observações

- Não há um concenso dentre os estatísticos quanto à utilização de médias em estudos envolvendo variáveis ordinais.
- Práticas mais usuais envolvem a associação de intervalos a cada uma das categorias. Poderíamos considerar, por exemplo,

Classe	Renda *per capta* ($\times R\$1000$)
B	$0 \vdash 1$
M	$1 \vdash 2$
A	$2 \vdash$

  Neste caso, teríamos que $1,\overline{3}$ de fato é favorável à classe M.
- Medidas como mediana e moda (ver na sequência) oferecem interpretações mais razoáveis nestes casos.

Exemplo 2

Uma pesquisa avaliou a idade dos 25 alunos de uma turma de engenharia. Os dados foram organizados na seguinte tabela de frequências:

Idade	**freq.**
19	3
20	5
21	1
22	8
23	4
24	1
25	0
26	3
Total	25

A média aritmética pode ser obtida diretamente do dispositivo (1):

$$\overline{x} = \frac{\sum_{i=1}^{8} x_i f_i}{\sum_{i=1}^{8} f_i} = \frac{548}{25} = 21,92.$$

Exemplo 3

Os dados a seguir correspondem ao tempo de vida, em anos, de cada elemento de uma amostra de lâmpadas produzidas por uma fábrica qualquer.

t	freq.
$0 \vdash 0,5$	3
$0,5 \vdash 1,0$	12
$1,0 \vdash 1,5$	20
$1,5 \vdash 2,0$	5
$2,0 \vdash$	0
Total	40

Quando buscamos uma tendência central e os dados estão agrupados em intervalos, devemos selecionar um elemento específico como representante de cada classe a fim de efetuarmos os cálculos necessários. Algumas das escolhas comuns são: centro ou extremos do intervalo. Neste caso, dizemos que a média foi **estimada** e não calculada.

Exemplo 3 (Continuação)

Tomando o ponto médio dos quatro primeiros intervalos e o extremo esquerdo do último intervalo, obtemos:

$$\overline{x} = \frac{\sum_{i=1}^{5} x_i f_i}{\sum_{i=1}^{5} f_i} = \frac{43,5}{40} = 1,0875 \text{ anos.}$$

Sensibilidade a valores extremos

- Médias aritméticas são bastante sensíveis a valores extremos.
- Pode levar a conclusões equivocadas.
- Suponhamos, por exemplo, que os salários dos funcionários de uma empresa de pequeno porte são: R$2500,00, R$2500,00, R$1000,00, R$1800,00 e R$20000. A média salarial da empresa é

$$\overline{x} = \frac{2500 + 2500 + 1000 + 1800 + 20000}{5} = 5560$$

e remete a uma conclusão equivocada a respeito da tendência salarial da empresa.

Mediana

- A mediana corresponde ao valor que ocupa a posição central dos dados após sua ordenação.
- Desta forma, a mediana desconsidera os valores observados em si já que basta analisar a posição das observações perante sua ordenação.
- Informalmente, esta medida trata os valores das observações de forma indireta já que estes são úteis apenas para o desenvolvimento da ordenação.
- Pode apresentar uma medida de tendência mais honesta na presença de valores atípicos.

Mediana – dados discretos ou não agrupados

- Caso os dados de interesse sejam discretos ou, sejam contínuos e ainda não tenham passado por um agrupamento, então o cálculo da mediana é mais simples.
- Basta observar a paridade do número n de observações:
  - *i*. se n é ímpar, então a mediana é exatamente o valor central das observações
  $$x^{\left(\frac{n+1}{2}\right)};$$
  - *ii*. caso contrário, a mediana é dada pela média aritmética dos dois valores centrais
  $$\frac{x^{\left(\frac{n}{2}\right)} + x^{\left(\frac{n}{2}+1\right)}}{2}.$$

Notaçõ: a expressão $x^{(i)}$ corresponde ao i-ésimo elemento da amostra ordenada. Logo, $x^{(i)} \neq x_i$.

- No exemplo dos salários, temos a seguinte ordenação: R$1000,00, R$1800,00, R$2500,00, R$2500,00, R$20000. Como o número de observações é ímpar, temos que a mediana é dada pelo elemento central $x^{(3)} = 2500$.
- No exemplo das classes sociais, temos a seguinte ordenação: B, B, B, B, B, M, M, M, M, M, M, M, M, M, M, A, A, A, A, A, A, A, A, A, A, A, A, A, A, A. Como o número de observações é par, segue que a mediana é dada pelo ponto médio de $x^{(15)}$ e $x^{(16)}$. Uma vez que ambos são iguais a A, segue que a mediana corresponde à classe A.

Mediana – dados agrupados em intervalos

- Caso os dados estejam agrupados em intervalos, então toda a extensão contínua de tais intervalos deve ser considerada no cálculo da mediana.
- Suponha que o conjunto de observações em estudo foi agrupado em k intervalos de mesmo tamanho:

$$[a_1, b_1), [a_2, b_2), ..., [a_k, b_k),$$

onde $b_i = a_{i+1}$, $\forall i = 1, 2, ..., k-1$.

- Considere ainda que o valor central, aquele que deixa 50% das observações à sua esquerda e 50% das observações à sua direita, pertence ao intervalo $[a_j, b_j)$ de frequência relativa f_j.
- Denotando por F_j a frequência relativa acumulada à esquerda de $[a_j, b_j)$, isto é, $F_j = \sum_{i=1}^{j-1} f_i$, temos que a mediana pode ser **estimada** pela relação:

$$\frac{mediana - a_j}{0,5 - F_j} = \frac{b_j - a_j}{f_j}. \quad (3)$$

Exemplo 3 (continuação)

Consideremos novamente o exemplo dos tempos de vida.

t	freq. abs.	freq. rel.
$0 \vdash 0,5$	3	0,075
$0,5 \vdash 1,0$	12	0,300
$1,0 \vdash 1,5$	20	0,500
$1,5 \vdash 2,0$	5	0,125
$2,0 \vdash$	0	0,000
Total	40	1,000

A mediana ocorre no terceiro intervalo já que até o valor 1,5 temos acumadas 87,5% das observações e até o valor 1,0 temos acumaladas apenas 37,5% das mesmas. Assim,

$$\frac{mediana - 1}{0,500 - 0,375} = \frac{1,5 - 1,0}{0,500}$$

$$\therefore mediana = 1 + (0,500 - 0,375)\frac{1,5 - 1,0}{0,500} = 1,125.$$

Moda

- A moda é o valor que mais se repete na amostra.
- Assim como a mediana, é pouco sensível a valores atípicos. Contudo, pode não ser única.
- Pode ser aplicada a variáveis nominais.

Exemplo 1 (continuação)

- Voltando à situação do exemplo 1, temos uma moda evidente observada na classe A.

classe	**freq**
B	5
M	10
A	15
Total	30

- Resultado coincide com o obtido em termos da mediana.
- Tanto a mediana quanto a moda apresentam interpretações mais razoáveis, em relação à média, quando desejamos estudar variáveis ordinais.

Exemplo 4

Suponha que os resumos a seguir correspodem aos salários (em milhares de reais) pagos por duas companhias, digamos A e B.

Companhia	A	B
média	2,5	2,0
mediana	1,7	1,9
moda	1,5	1,9

Considerando apenas os dados indicados no resumo acima, qual empresa é mais atrativa?

Medidas separatrizes

- Medidas separatrizes dividem a sequência ordenada dos dados em partes que contêm a mesma quantidade de elementos.
- A mediana trata de um caso particular pois separa a amostra em duas porções as quais contêm 50% da informação cada.
- Considerando, por exemplo, o conjunto de dados 1,2,5,5,5,8,10,11,12,12,13,15, temos:

$$\underbrace{1, 2, 5, 5, 5, 8}_{\text{50\% das observações}} \quad | \quad \underbrace{10, 11, 12, 12, 13, 15}_{\text{50\% das observações}}$$

$$\downarrow$$

$$\text{mediana: } \frac{8 + 10}{2} = 9$$

- As regras de cálculo são análogas às adotadas no caso da mediana; basta adaptá-las para a porção desejada dos dados.

Quartil

- Ao dividir os dados ordenados em 4 partes de mesmo tamanho, cada um resumirá 25% da informação.
- Os elementos Q_1, Q_2 e Q_3 que separam tais grupos são denominados os quartis amostrais.
- Em particular, o segundo quartil Q_2 corresponde à mediana.

$$\underbrace{1,2,5}_{\text{25\% dos dados}} \;|\; \underbrace{5,5,8}_{\text{25\% dos dados}} \;|\; \underbrace{10,11,12}_{\text{25\% dos dados}} \;|\; \underbrace{12,13,15}_{\text{25\% dos dados}}$$

$$Q_1 = \frac{5+5}{2} = 5 \quad Q_2 = \frac{8+10}{2} = 9 \quad Q_3 = \frac{12+12}{2} = 12$$

Generalização

- Os quatro elementos K_1, K_2, K_3 e K_4 que separam os dados em cinco partes iguais são denominados quintis amostrais.
- Analogamente:
  i. 10 partes iguais – decis amostrais;
  ii. 100 partes iguais – percentis amostrais;
  iii. q partes iguais – quantis amostrais (q é um inteiro não-negativo qualquer)

Dispositivo prático – dados discretos ou não-agrupados

- Para a **estimativa** de uma separatriz qualquer, basta identificar o percentil p_i correspondente:

$$p_i = \frac{i \times (n+1)}{100},$$

onde n é o número de observações.

- Se p_i for um número inteiro, então a separatriz desejada é dada por um dos elementos da série:

$$x^{(p_i)}.$$

- Caso contrário, a separatriz é especificada proporcionalmente (interpolação linear) dentro do intervalo:

$$\left(x^{(\lfloor p_i \rfloor)}, x^{(\lfloor p_i \rfloor + 1)}\right).$$

Exemplo 5

Considere novamente a sequência de valores 1, 2, 5, 5, 5, 8, 10, 11, 12, 12, 13, 15.

Obtenha Q_1, Q_2, K_2 e o oitavo decil dos dados.

- Para o cálculo de Q_1, devemos buscar o percentil 25%: $\frac{25\times(12+1)}{100} = 3,25$. Logo,

$$Q_1 = x^{(3)} + 0,25 \times (x^{(4)} - x^{(3)}) = 5 + 0,25(5-5) = 5.$$

- Para o cálculo de Q_2, devemos buscar o percentil 50%: $\frac{50\times(12+1)}{100} = 6,5$. Logo,

$$Q_2 = x^{(6)} + 0,5 \times (x^{(7)} - x^{(6)}) = 8 + 0,5(10-8) = 9.$$

- Para o cálculo de K_2, devemos buscar o percentil 40%: $\frac{40\times(12+1)}{100} = 5,2$. Logo,

$$K_2 = x^{(5)} + 0,2 \times (x^{(6)} - x^{(5)}) = 5 + 0,2(8-5) = 5,6.$$

- Para o cálculo do oitavo decil, devemos buscar o percentil 80%: $\frac{80\times(12+1)}{100} = 10,4$. Logo,

$$D_8 = x^{(10)} + 0,4 \times (x^{(11)} - x^{(10)}) = 12 + 0,4(13-12) = 12,4.$$

Dispositivo prático – dados agrupados em intervalos

- O cálculo de separatrizes quando os dados estão agrupados em intervalos é similar ao adotado no caso da mediana.
- Basta substituir o nível de corte 50% (0,5) pelo quantil (q) desejado.
- Neste caso, a separatriz S_q pode ser estimada por

$$\frac{S_q - a_j}{q - F_j} = \frac{b_j - a_j}{f_j}, \tag{4}$$

onde F_j, f_j, a_j e b_j são definidos de forma análoga à utilizada na equação (3).

Exemplo 6

Uma empresa desenvolveu um estudo com respeito a um de seus produtos e obteve a seguinte tabela de frequências para o tempo (em meses) até a verificação da primeira falha de funcionamento.

t	**freq.**
$0 \vdash 6$	9
$6 \vdash 12$	11
$12 \vdash 18$	25
$18 \vdash 24$	21
$24 \vdash 30$	40
$30 \vdash 36$	38
$36 \vdash 42$	30
$42 \vdash 48$	20
$48 \vdash$	7
Total	201

Se tal empresa deseja criar um plano de garantia que lhe obrigue a trocar, em média, 25% dos produtos vendidos, qual deve ser o prazo de garantia oferecido?

Exemplo 6 (continuação)

- Basta buscarmos o tempo em meses responsável por acumular 25% das observações.
- Em outras palavras, devemos buscar o primeiro quartil amostral (Q_1).
- Para facilitar os cálculos, podemos transformar as frequências absolutas em relativas:

t	freq. abs.	freq. rel.
$0 \vdash 6$	9	0,0448
$6 \vdash 12$	11	0,0547
$12 \vdash 18$	25	0,1244
$18 \vdash 24$	21	0,1045
$24 \vdash 30$	40	0,1990
$30 \vdash 36$	38	0,1891
$36 \vdash 42$	30	0,1493
$42 \vdash 48$	20	0,0995
$48 \vdash$	7	0,0347
Total	201	1,000

Exemplo 6 (continuação)

- Neste caso, Q_1 ocorre no intervalo $[18, 24)$.
- Além disso, temos que $F_j = 0,2239$ e $f_j = 0,1045$.
- Logo,

$$\frac{Q_1 - 18}{0,2500 - 0,2239} = \frac{24 - 18}{0,1045}$$

$$\therefore Q_1 = 18 + 0,0267 \times \frac{6}{0,1045} = 19,53301435 \approx 20 \text{ meses.}$$

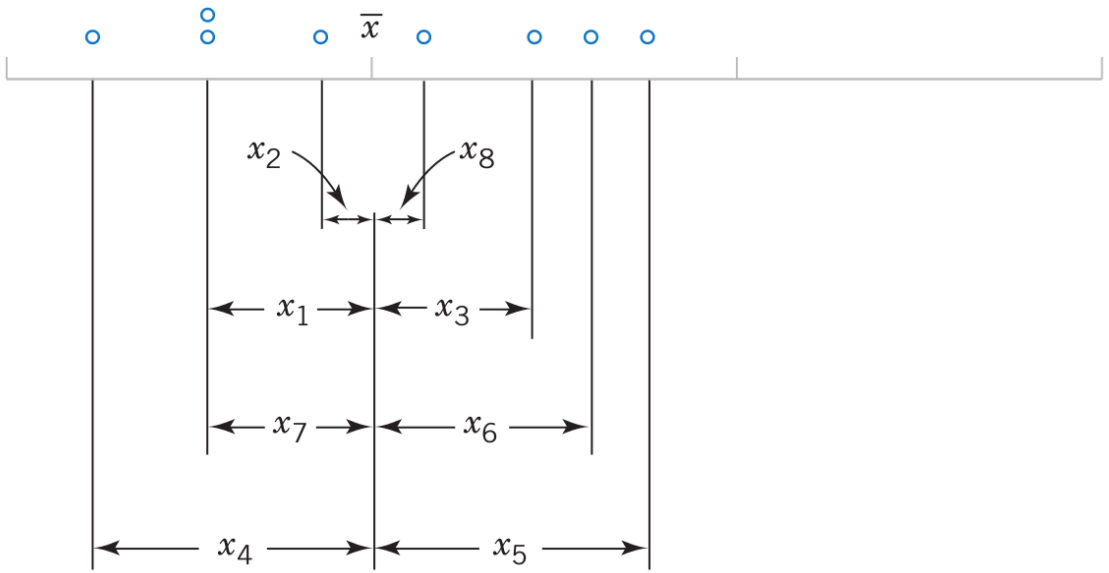
- Ao estabelecer uma garantia de 20 meses, temos que, em média, apenas 25% dos produtos vendidos serão trocados pela fábrica.

Medidas de dispersão

- Medidas de dispersão descrevem a variabilidade dos dados em torno de uma tendência central.
- Podem ser interpretadas como uma medida de precisão associada às medidas de posição.
- Principais medidas: variância, desvio padrão, amplitude, intervalo interquartil.

Variância

- A variância amostral, usualmente denotada por S^2, fornece uma maneira de medir o desvio médio das observações com respeito ao valor central $\overline{x}$.



- Seja $x_1, x_2, ..., x_n$ um conjunto qualquer de n observações. O i-ésimo desvio d_i é definido como

$$d_i = x_i - \overline{x}. \tag{5}$$

Variância

- Formalmente, a variância amostral corresponde à quantidade

$$S^2 = \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n}. \tag{6}$$

- Isto é, a variância corresponde à média dos quadrados dos desvios amostrais.
- O cálculo é baseado nos quadrados dos desvios pois

$$\frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x}) = 0$$

para qualquer conjunto de dados.

Exemplo 7

Calcular a variância amostral da sequência: 12,6; 12,9; 13,4; 12,3; 13,6; 13,5; 12,6; 13,1.

Exemplo 7

Calcular a variância amostral da sequência: 12,6; 12,9; 13,4; 12,3; 13,6; 13,5; 12,6; 13,1.

i	x_i	$d_i^2 = (x_i - \overline{x})^2$
1	12,6	0,16
2	12,9	0,01
3	13,4	0,16
4	12,3	0,49
5	13,6	0,36
6	13,5	0,25
7	12,6	0,16
8	13,1	0,01
$\sum$	104,0	1,6
$\frac{1}{n}\sum$	$\overline{x} = 13,0$	$S^2 = 0,2$

Variância (amostras pequenas)

- É comum considerarmos

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{7}$$

  para amostras pequenas ($n < 30$).
- Tal adaptação pode ser associada aos $n-1$ **graus de liberdade** da amostra:

$$\sum_{i=1}^{n} d_i = 0 \Longrightarrow d_i = \sum_{j=1, j\neq i}^{n} d_j, \qquad \forall i = 1, 2, ..., n.$$

Variância (cálculo alternativo)

- Em muitos casos pode ser útil trabalhar com a variância em função da média dos quadrados das observações:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2. \quad (8)$$

Variância (cálculo alternativo)

- Em muitos casos pode ser útil trabalhar com a variância em função da média dos quadrados das observações:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \overline{x}^2. \tag{8}$$

- No caso da fórmula alternativa:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2\right). \tag{9}$$

Exemplo 7 (continuação)

- Calcular a variância amostral da sequência: 12,6; 12,9; 13,4; 12,3; 13,6; 13,5; 12,6; 13,1.

Exemplo 7 (continuação)

- Calcular a variância amostral da sequência: 12,6; 12,9; 13,4; 12,3; 13,6; 13,5; 12,6; 13,1.

i	x_i	$d_i^2 = (x_i - \overline{x})^2$	x_i^2
1	12,6	0,16	158,76
2	12,9	0,01	166,41
3	13,4	0,16	179,56
4	12,3	0,49	151,29
5	13,6	0,36	184,96
6	13,5	0,25	182,25
7	12,6	0,16	158,76
8	13,1	0,01	171,61
$\sum$	104,0	1,6	1353,6
$\frac{1}{n}\sum$	$\overline{x} = 13,0$	$S^2 = 0,2$	169,2

- Pelo cálculo alternativo:

$$S^2 = 169,2 - 13,0^2 = 0,2.$$

Desvio padrão

- O conceito de variância é bastante rico, contudo, deve ser utilizado com cautela já que trata do problema original em escala quadrática.
- O desvio padrão surge como uma alternativa para corrigir este detalhe e assim facilitar a análise dos resultados.
- Tal medida é dada pela raiz quadrada da variância amostral:

$$S = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i^2 - \overline{x}^2}. \qquad (10)$$

Desvio padrão

- O conceito de variância é bastante rico, contudo, deve ser utilizado com cautela já que trata do problema original em escala quadrática.
- O desvio padrão surge como uma alternativa para corrigir este detalhe e assim facilitar a análise dos resultados.
- Tal medida é dada pela raiz quadrada da variância amostral:

$$S = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i^2 - \overline{x}^2}. \qquad (10)$$

- Em geral, a variância amostral possui propriedades matemáticas melhores enquanto o desvio padrão oferece interpretações mais razoáveis.

Intervalo interquartil

- Variância e desvio padrão também são sensíveis a valores discrepantes por considerar os valores observados diretamente.
- Uma maneira alternativa de contornar tal problema é considerar a amplitude interquartil:

$$A_{IQ} = Q_3 - Q_2. \tag{11}$$

- Tal quantidade indica a faixa de variação dos 50% centrais das observações.
- A escala original dos dados é preservada neste caso.

Exemplo 7 (continuação)

- Calcular a variância amostral da sequência: 12,6; 12,9; 13,4; 12,3; 13,6; 13,5; 12,6; 13,1.

Exemplo 7 (continuação)

- Calcular a variância amostral da sequência: 12,6; 12,9; 13,4; 12,3; 13,6; 13,5; 12,6; 13,1.
- Ordenando os dados, obtemos: 12,3; 12,6; 12,6; 12,9; 13,1; 13,4; 13,5; 13,6;
- Logo:

$$Q_3 = \frac{13,4 + 13,5}{2} = 13,45 \qquad \text{e} \qquad Q_1 = \frac{12,6 + 12,6}{2} = 12,6$$

e,

$$A_{IQ} = 13,45 - 12,6 = 0,85.$$

Amplitude

- Quando a amostra é muito pequena ($n < 5$, por exemplo) a utilização das medidas anteriores resulta em análises muito pobres.
- Nestes casos, a amplitude amostral

$$x^{(n)} - x^{(1)}$$

pode ser uma alternativa interessante de avaliação da dispersão.

- Não deve ser utilizada em amostras maiores.