Estatística e Probabilidade

Bacharelado em Sistemas de Informação

Aula 3: Resumindo dados: Medidas de Dispersão

Prof. Dr. Samuel Sanches

MEDIDAS DE DISPERSÃO: AMPLITUDE

★ Os conjuntos de dados muitas vezes possuem certa **variabilidade**: dois pacientes tiveram os pulsos medidos, A, 72, 76 e 74 e o B, 72, 91 e 59, para os dois a média é 74, porém o B possui uma **variação** muito grande!

★ A **amplitude** nos ajuda a verificar essa **variação**, sendo definida como: **diferença entre o maior e o menor valor**.

★ Paciente A: Amplitude = 76 - 72 = 4; Paciente B: Amplitude = 91 - 59 = 32.

MEDIDAS DE DISPERSÃO: AMPLITUDE

★ Veja os dados:

Conjunto A:	5	18	18	18	18	18	18	18	18	18
Conjunto B:	5	5	5	5	5	18	18	18	18	18
Conjunto C:	5	6	8	9	10	12	14	15	17	18

Todos possuem amplitude de 18 - 5 = 13, porém veja a dispersão (espaçamento entre valores), que é totalmente diferente!

★ **Amplitude interquartil**: Q3 - Q1. Para abranger aproximadamente 50% dos valores centrais.

★ **Amplitude semi-interquartil** ou **desvio quartil**: (Q3 - Q1)/2

MEDIDAS DE DISPERSÃO: DESVIO-PADRÃO E VARIÂNCIA

★ Uma das medidas de dispersão mais úteis!

★ Valores **concentrados** ao redor da **média**: dispersão **pequena**.

★ Valores **espalhados** ao redor da **média**: dispersão **alta**.

★ Quanto o valor está afastado da média (**desvio da média**): $x_n - \bar{x}$

★ Então vamos utilizar o quadrado desse valor (para evitar valores negativos) somando todos resultados, para depois dividir pela quantidade n de dados (média) e calcular a raiz quadrada, para compensar o fato de ter elevado ao quadrado, temos o chamado **desvio da raiz dos quadrados médios**.

MEDIDAS DE DISPERSÃO: DESVIO-PADRÃO E VARIÂNCIA

★ Temos o **desvio-padrão amostral**:

$$s = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}}$$

$$s = \sqrt{\frac{S_{xx}}{n-1}} \quad \text{onde} \quad S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$$

★ Elevando ao quadrado, $\mathbf{s^2}$, temos a chamada **variância amostral**.

★ Para populações, **desvio-padrão populacional**, temos a mesma expressão com modificações nas notações, $s \rightarrow \sigma$, $n \rightarrow N$, média $\rightarrow \mu$. E a **variância população** dada por σ^2

MEDIDAS DE DISPERSÃO: DESVIO-PADRÃO E VARIÂNCIA

★ **Exemplo**: Um laboratório encontrou 8, 11, 7, 13, 10, 11, 7 e 9 bactérias em oito pessoas saudáveis. Calcule s.

$$\bar{x} = \frac{8+11+7+13+10+11+7+9}{8} = 9{,}5$$

x	$x - \bar{x}$	$(x - \bar{x})^2$
8	−1,5	2,25
11	1,5	2,25
7	−2,5	6,25
13	3,5	12,25
10	0,5	0,25
11	1,5	2,25
7	−2,5	6,25
9	−0,5	0,25
	0,0	32,00

$$s = \sqrt{\frac{32{,}00}{7}} = \sqrt{4{,}57} = 2{,}14$$

MEDIDAS DE DISPERSÃO: DESVIO-PADRÃO E VARIÂNCIA

★ **Exemplo**: Um laboratório encontrou 8, 11, 7, 13, 10, 11, 7 e 9 bactérias em oito pessoas saudáveis. Calcule s (com a outra expressão).

$$\sum x = 8 + 11 + 7 + \cdots + 7 + 9$$
$$= 76$$

$$\sum x^2 = 64 + 121 + 49 + 169 + 100 + 121 + 49 + 81$$
$$= 754$$

$$S_{xx} = 754 - \frac{(76)^2}{8} = 32$$

$$\boxed{s = \sqrt{\frac{32}{7}} = 2{,}14}$$

Planilha Aula 3:
https://docs.google.com/spreadsheets/d/1HoKhlBw2n47HyA9HRg76fRTJrKQ8srW8ST9198OS91A/edit?usp=share_link

APLICAÇÕES

- ★ **Teorema de Tchebichev**: Para qualquer conjunto de dados (população ou amostra) e qualquer constante k maior do que 1, a proporção dos dados que devem estar a menos de k desvios-padrão de qualquer um dos dois lados da média é pelo menos **$1 - 1/k^2$**

- ★ Isso nos permite afirmar coisas como 75% dos valores de qualquer conjunto de dados devem estar a pelo menos 2 desvios-padrão de qualquer um dos dois lados da média ($1 - 1/2^2 = ¾ = 0,75$). 96% dos valores estão a 5 desvios-padrão ($1 - 1/5^2 = 24/25 = 0,96$) e 99% a menos de 10 desvios-padrão ($1 - 1/10^2 = 99/100 = 0,99$).

APLICAÇÕES

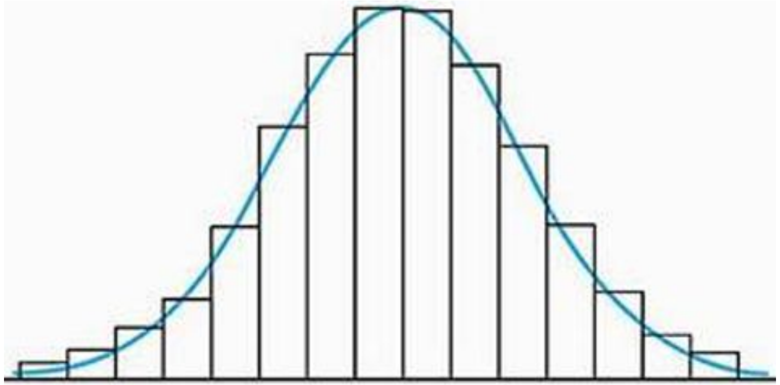
★ **Exemplo**: Um estudo sobre queijos mostrou que, em média, uma fatia de 30 gramas contém 3,50 gramas de gordura com um desvio-padrão de 0,04 gramas de gordura. **a)** De acordo com o teorema de Tchebichev, pelo menos qual percentagem de uma fatia de 30 gramas desse tipo de queijo deve ter um conteúdo de gordura entre 3,38 e 3,62 gramas de gordura? **b)** De acordo com o teorema, entre quais valores deve estar o conteúdo de gordura de pelo menos 93,75% das fatias de 30 gramas desse tipo queijo?

a) 3,62 - 3,50 = 0,12, então k*0,04 = 0,12 (para descobrir quantos desvios-padrão temos), k = 3, assim, $1 - 1/3^2 = 8/9 = 0{,}889$, ou seja, 88,9% possuem o indicado.

b) $1 - 1/k^2 = 0{,}9375$, $1/k^2 = 1 - 0{,}9375 = 0{,}0625$, $k^2 = 1/0{,}0625 = 16$, k = 4, então, 93,75% possuem entre 3,50 - 4*(0,04) = 3,34 e 3,50 + 4*(0,04) = 3,66 gramas de gordura.

APLICAÇÕES

★ Possui problemas, temos uma regra **empírica** (registrada a partir de dados) que caso a distribuição siga o formato de sino (**distribuição normal**), podemos afirmar:



Cerca de 68% dos valores estão a menos de um desvio-padrão da média, isto é, entre $\bar{x} - s$ e $\bar{x} + s$.

Cerca de 95% dos valores estão a menos de dois desvios-padrão da média, isto é, entre $\bar{x} - 2s$ e $\bar{x} + 2s$.

Cerca de 99,7% dos valores estão a menos de três desvios-padrão da média, isto é, entre $\bar{x} - 3s$ e $\bar{x} + 3s$.

APLICAÇÕES

★ **Exemplo**: Usamos os dados de tempo de espera de erupções, onde a média encontrada foi de 78,59 e aqueles dados possuem desvio-padrão de 14,35. Utilizando esses valores, determine qual percentagem está a menos de <u>três</u> desvios-padrão da média.

Pontos médios de classe x	*Freqüência* f	$x \cdot f$
34,5	2	69,0
44,5	2	89,0
54,5	4	218,0
64,5	19	1.225,5
74,5	24	1.788,0
84,5	39	3.295,5
94,5	15	1.417,5
104,5	3	313,5
114,5	2	229,0
	110	8.645,0

Então: 78,59 - 3*(13,35) = 35,54 (limite inferior), 78,59 + 3*(13,35) = 121,64 (limite superior), temos 2 valores menores que 35,54 e nenhum acima de 121, então 110 - 2 = 108 dos valores estão compreendidos no intervalo, ou seja, (108/110)*100 = 98,2% dos tempos caem a menos de 3 desvios-padrão da média (próximo do valor de 99,7% do slide anterior, vale lembrar que essa distribuição não tem forma de sino).

APLICAÇÕES

- ★ Para compararmos dados aparentemente diferentes, usamos **unidades padronizadas** (ou **escores**), z.

$$z = \frac{x - \bar{x}}{s} \quad ou \quad z = \frac{x - \mu}{\sigma}$$

- ★ Dois exames, vocabulário obteve 66 e gramática 80, aparentemente melhor em gramática, média turma 51 com desvio 12 em vocabulário e média 72 com desvio 16 em gramática.

  Vocabulário: (66 - 51)/12 = 1,25 desvios-padrão acima da média.
  Gramática: (80 - 72)/16 = 0,50 desvios-padrão acima da média.

  Então, veja que ele está melhor do que a turma em vocabulário (mais longe da média).

APLICAÇÕES

- **Exemplo**: Uma faixa etária possui massa média de 56 kg, com desvio-padrão de 6 kg, outra faixa etária possui massa média 82 kg com desvio-padrão de 9 kg. Se uma pessoa da 1º faixa possui 66 kg e outra da 2º faixa possui 96 kg, qual dos dois, relativamente a massa média de sua faixa etária, está com maior excesso de massa?

  1º faixa: (66 - 56)/6 = 1,66
  2º faixa: (96 - 82)/9 = 1,55

  Então, a pessoa da 1º faixa está com excesso relativo a sua faixa etária.

APLICAÇÕES

- ★ O desvio-padrão depende das unidades de medida, por exemplo um objeto possui desvio-padrão de 0,1 grama, não temos como saber se isso é muito ou se é pouco, assim podemos usar uma **medida de dispersão relativa**, como o **coeficiente de dispersão**:

$$V = \frac{s}{\bar{x}} \cdot 100\% \quad \text{ou} \quad V = \frac{\sigma}{\mu} \cdot 100\%$$

- ★ Ele nos informa o desvio-padrão como uma porcentagem da medida em média.

APLICAÇÕES

★ **Exemplo**: Várias medições efetuadas com um micrômetro, acusaram média de 2,49 mm e desvio-padrão de 0,012 mm, com um outro micrômetro e um outro objeto teve 0,75 cm de média e desvio-padrão de 0,002 cm. Qual dos dois micrômetros é mais preciso?

Coeficiente de dispersão:

$$\frac{0{,}012}{2{,}49} \cdot 100 \approx 0{,}48\% \quad \text{e} \quad \frac{0{,}002}{0{,}75} \cdot 100 \approx 0{,}27\%$$

Observe que o 2º possui menos variação, então é mais preciso.

DADOS AGRUPADOS

- ★ Como temos uma perda de informação para cada dado, os itens dentro de uma mesma classe corresponderão ao ponto médio da classe. Então soma de todas medidas:

$$\sum x \cdot f = x_1 f_1 + x_2 f_2 + \cdots + x_k f_k$$

- ★ Soma de seus quadrados:

$$\sum x^2 \cdot f = x_1^2 f_1 + x_2^2 f_2 + \cdots + x_k^2 f_k$$

- ★ Desvio-padrão para dados amostrais agrupados:

$$s = \sqrt{\frac{S_{xx}}{n-1}} \quad \text{onde} \quad S_{xx} = \sum x^2 \cdot f - \frac{\left(\sum x \cdot f\right)^2}{n}$$

DADOS AGRUPADOS

★ **Exemplo**: Com os dados do tempo de espera entre erupções (tabela ao lado), calcule o desvio-padrão.

Pontos médios de classe x	Freqüência f	$x \cdot f$
34,5	2	69,0
44,5	2	89,0
54,5	4	218,0
64,5	19	1.225,5
74,5	24	1.788,0
84,5	39	3.295,5
94,5	15	1.417,5
104,5	3	313,5
114,5	2	229,0
	110	8.645,0

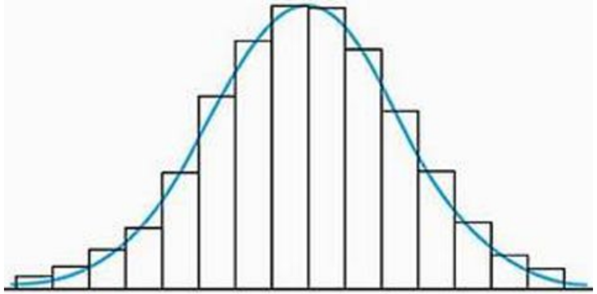
x	f	$x \cdot f$	$x^2 \cdot f$
34,5	2	69	2.380,5
44,5	2	89	3.960,5
54,5	4	218	11.881
64,5	19	1.225,5	79.044,75
74,5	24	1.788	133.206
84,5	39	3.295,5	278.469,75
94,5	15	1.417,5	133.953,75
104,5	3	313,5	32.760,75
114,5	2	229	26.220,5
	110	8.645	701.877,5

$$S_{xx} = 701.877,5 - \frac{(8.645)^2}{110} \approx 22.459,1$$

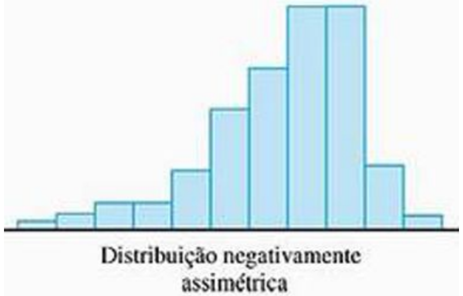
$$s = \sqrt{\frac{22.459,1}{109}} \approx 14,35$$

DESCRIÇÕES ESTATÍSTICAS

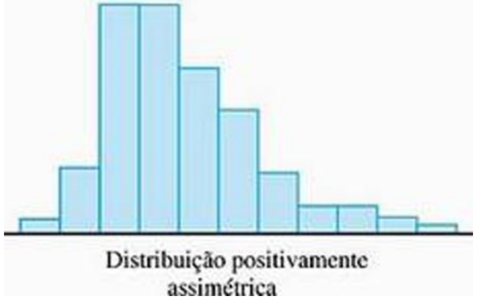
★ Distribuição **perfeitamente** simétrica: Distribuição em forma de sino.



★ Distribuição **negativamente** assimétrica: “cauda” à esquerda.

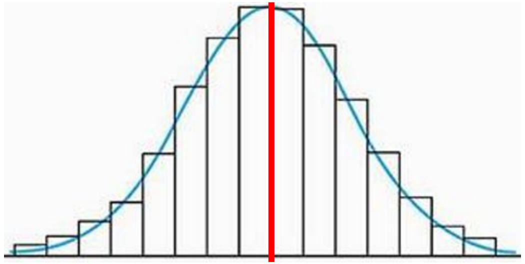


★ Distribuição **positivamente** assimétrica: “cauda” à direita

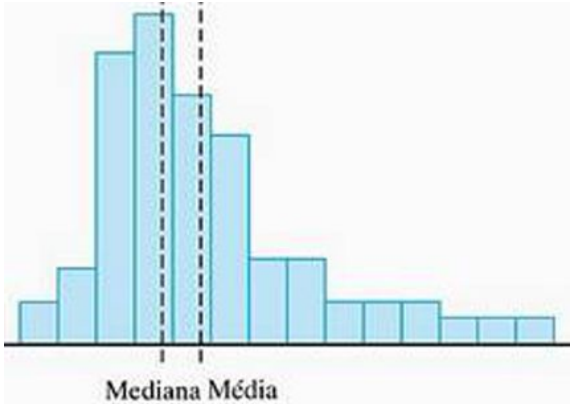


MEDIDAS DE ASSIMETRIA

★ Média e mediana coincidem: simetria perfeita.



★ Assimetria positiva: média é maior que a mediana e Assimetria negativa: média é menor que a mediana.



★ **Coeficiente de assimetria de Pearson**:

$$SK = \frac{3\,(média - mediana)}{desvio\text{-}padrão}$$

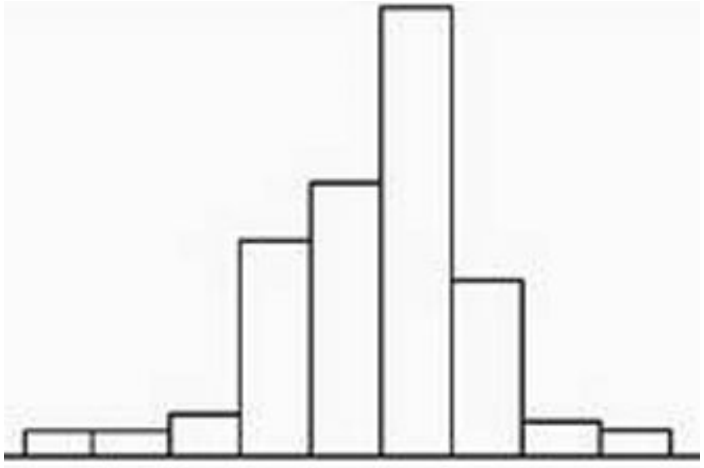
SK = 0 → simetria perfeita, SK deve cair entre -3 e 3.

MEDIDAS DE ASSIMETRIA

- **Exemplo**: Calcule SK para a distribuição de tempos de espera entre as erupções, onde temos média de 78,59, mediana de 80,53 e desvio-padrão de 14,35.

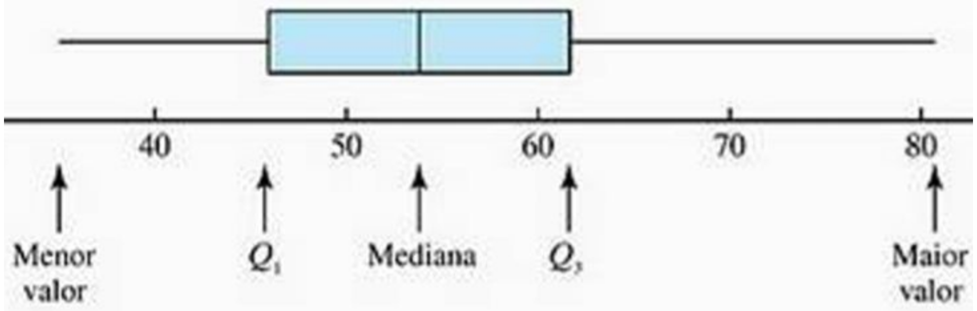
$$SK = \frac{3(78{,}59 - 80{,}53)}{14{,}35} \approx -0{,}41$$

Simetria negativa (fraca/pequena), veja o histograma.



MEDIDAS DE ASSIMETRIA

- ★ Conjunto pequeno de dados, histograma não é boa escolha, gráfico de caixa nos trás mais informações, a partir da posição da mediana em relação aos dois quartis Q1 e Q3.



- ★ Reta da mediana está no centro ou próximo, temos uma boa simetria, muito à esquerda, simetria positiva, muito à direita, simetria negativa. Bem como as retas de menor e maior valor, quanto mais diferentes, mais assimétrico.

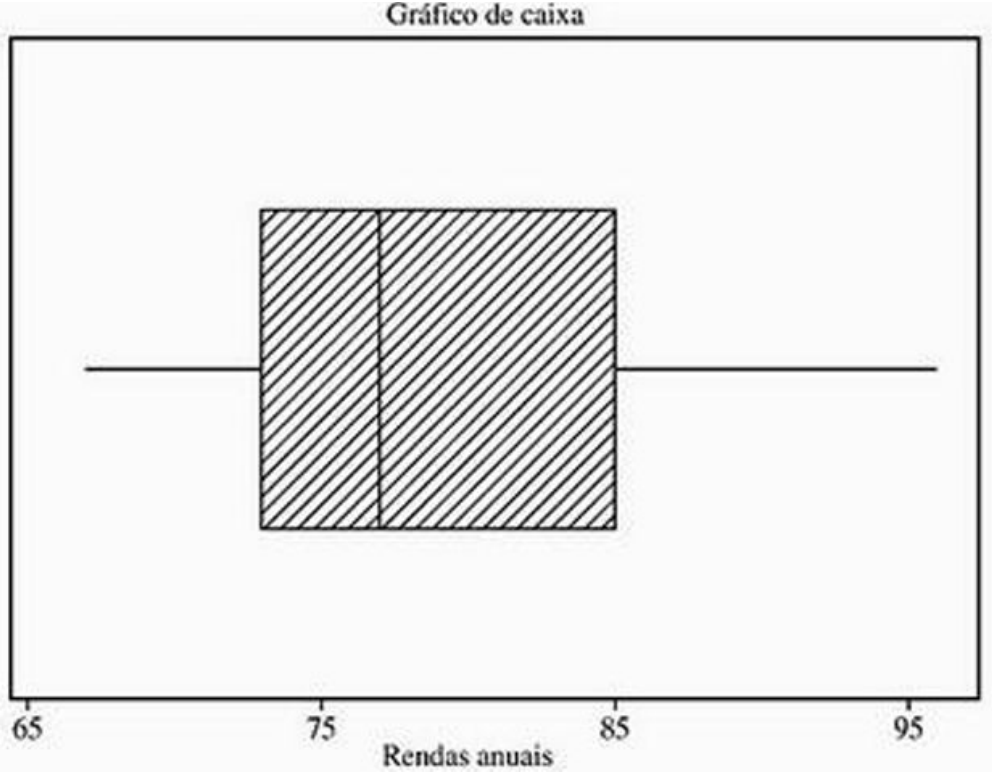
MEDIDAS DE ASSIMETRIA

★ **Exemplo**: Os dados representam as rendas anuais de quinze contadores em milhares de unidades monetárias: 88, 77, 70, 80, 74, 82, 85, 96, 76, 67, 80, 75, 73, 93 e 72. Esboce um gráfico de caixa e use-o para determinar a simetria ou a falta de simetria dos dados.

Ordem: 67, 70, 72, 73, 74, 75, 76, 77, 80, 80, 82, 85, 88, 93, 96, menor 67, maior 96, mediana 77 (8º valor de qualquer lado), Q1 é 73 (4º valor a partir da esquerda), Q3 é 85 (4º valor a partir da direita).

MEDIDAS DE ASSIMETRIA

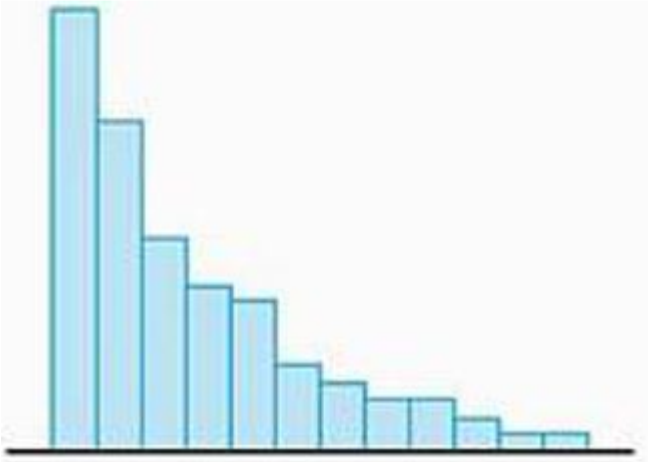
- **Exemplo**: Temos tendência que os dados são negativamente assimétricos, a reta mediana está à esquerda e a reta da direita é maior.



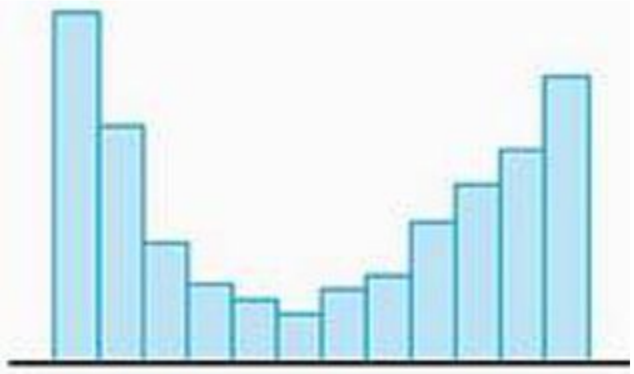
Planilha Aula 3:
https://docs.google.com/spreadsheets/d/1HoKhlBw2n47HyA9HRg76fRTJrKQ8srW8ST9198OS91A/edit?usp=share_link

MEDIDAS DE ASSIMETRIA

★ Distribuição em forma de J invertido:



★ Distribuição em forma de U:



EXERCÍCIOS

★ **Lista 1 de Exercícios → Parte 2**

https://drive.google.com/file/d/1f755nMDqpEIkcHnFF49CjbAfw8CvWlCS/view?usp=share_link

- ★ Muito obrigado pela atenção!