

Avaliação de Desempenho de Redes e Sistemas Computacionais

Estatística e Sumarização de Dados

UFPB / CI / DSC

Josilene Aires Moreira

Sumarizando dados em um único numero

- Índice de tendência central
 - Média
 - Mediana
 - Moda
- Quando usar?
- (*) Estatística descritiva

Sumarizando dados em um único numero

- $\{20, 30, 30, 40, 50, 50, 50\}$
- Média = ? $(20 + 30 + 30 + 40 + 50 + 50 + 50)/7 = 38,57$
- Mediana = ? Valor do meio dos valores ordenados
= 40
- Moda = ? Valor que ocorre com maior frequência
= 50

Mediana

É o valor que separa a metade maior e a metade menor de uma amostra, uma população ou uma distribuição de probabilidade.

- {12, 23, 23, 25, 27, 34, 41} Mediana = 25
- {11, 12, 15, 17, 21, 32} Mediana = ?
- Mediana = $(15 + 17) / 2 = 16$ (está a meio caminho entre a terceira e a quarta observação em uma sequencia ordenada)
- Pouco sensível a valores extremos **70**
- Notas de Tom = 20, 40, 70, 75, 80 media = 57, mediana = ? **70**
- Notas de Maria = 60, 65, 70, 90, 95 media = 76, mediana = ? **70**
- Notas de Jose = 50, 65, 70, 75, 90 media = 70, mediana = ?

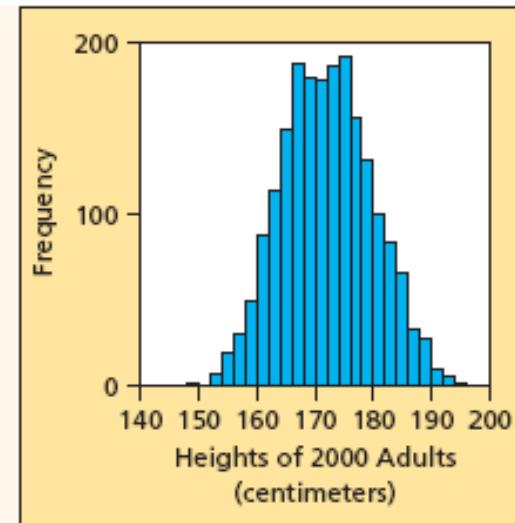
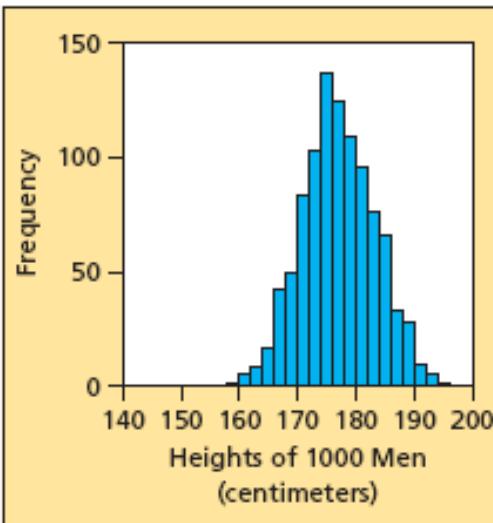
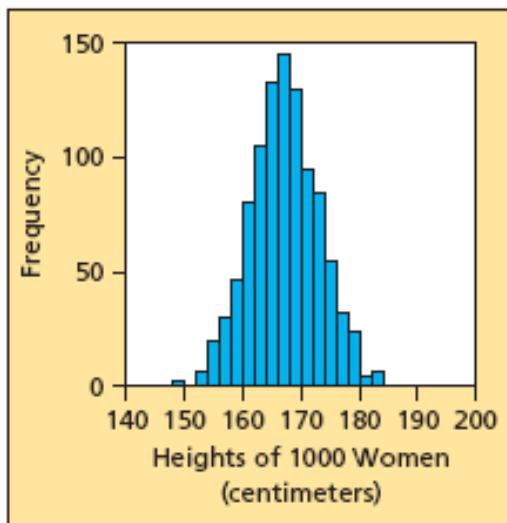
Moda

Valor com maior frequência de ocorrência nos dados

- {60, 70, 70, 70, 80} • media = 70, mediana=70, moda = 70
- {45, 45, 70, 90, 100} • media = 70, mediana = 70, moda = 45
- {50, 60, 70, 80, 90} • media = 70, mediana = 70, moda = nenhuma
- {50, 50, 70, 90, 90} • media = 70, mediana = 70, modas = 50, 90
 (bimodal)

Moda

- Alturas de mulheres e homens



- Não é bem visualizada no terceiro gráfico
- Quando existe heterogeneidade, é melhor criar histogramas separados

Média

- Exemplos de uso inadequado
- Tempo de resposta a uma consulta
 - $T_1 = 10 \text{ ms}$
 - $T_2 = 1000 \text{ ms}$
- Embora a média seja 505 ms, não é adequado considerar esta média, pois induz a pensar em valores muitos distantes dos valores reais

Média

	Sistema A	Sistema B
	10	5
	9	5
	11	5
	10	4
	10	31
Soma	50	50
Media	10	10
Valor típico (moda)	10	5

- A média não é adequada para representar o sistema B

Parte II – Medidas de dispersão



Dispersão

- A dispersão mostra como os dados estão “espalhados” em relação ao centro de uma distribuição

* *Medidas de dispersão*

Estatistica	Formula	Excel	Pro	Con
amplitude	$x_{\max} - x_{\min}$	=MAX(Data)-MIN(Data)	Fácil de calcular	Sensível a valores extremos

Dispersão

* *Amplitude*

- A diferença entre o maior e o menor valor entre as observações

$$\text{Amplitude} = x_{\max} - x_{\min}$$

7	8	8	10	10	10	10	12	13	13	13	13	13	13	13	14	14
14	15	15	15	15	15	16	16	16	17	18	18	18	18	19	19	19
19	19	20	20	20	21	21	21	22	22	23	23	23	24	25	26	26
26	26	27	29	29	30	31	34	36	37	40	41	45	48	55	68	91

$$\text{Amplitude} = 91 - 7 = 84$$



Dispersão

* Variância

Statistic	Formula	Excel	Pro	Con	
Variância (s^2)	$\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$	=VAR(Data)	Muito importante na Estatística	Não é intuitiva	

- A dispersão para uma população é medida pela variância populacional (N)
- A fórmula mostra a variância amostral

Dispersão

* Desvio padrão

Statistic	Formula	Excel	Pros	Cons
Desvio padrão Standard deviation (s)	$\sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$	=STDEV(Data)	Muito comum	Não intuitivo

- O desvio padrão ajuda a entender como os valores individuais em um conjunto de dados variam ao redor da média

Dispersão

* *Desvio padrão*

- A raiz-quadrada da variância
- Valor não-negativo
- Só pode ser usado para comparar valores de dados medidos nas mesmas unidades
 - Não pode ser usado para comparar dólares com Euros, por exemplo; ou Kg com Libras

Dispersão

* *Coeficiente de Variação*

Coeficiente de variacao (CV)	$100 \times \frac{s}{\bar{x}}$	-	Permite comparar dados com unidades de medidas diferentes	Dados tem que ser positivos
-------------------------------------	--------------------------------	---	---	-----------------------------

- s = desvio padrão, \bar{x} = média
- O CV é o desvio padrão expresso como uma percentagem da média
- Dá uma ideia razoável da dispersão, de forma intuitiva

Dispersão

* *Coeficiente de Variação*

$$CV = 100 \times \frac{s}{\bar{x}}$$

MEDIDAS DE DISPERSÃO

Coeficiente de Variação (CV)

$$CV = \frac{s}{\bar{x}}$$

Amostra X apresenta média de 24 metros e desvio padrão de 6 metros. A amostra Y, média igual, porém desvio padrão de 8 metros. Qual tem a maior variação?

$$CV(x) = \frac{s}{\bar{x}} = \frac{6}{24} = 25\%$$

$$CV(y) = \frac{s}{\bar{y}} = \frac{8}{24} = 33\%$$



Dispersão

* Desvio médio absoluto

Estatística	Formula	Excel	Pro	Con
Desvio médio absoluto (MAD)	$\frac{\sum_{i=1}^n x_i - \bar{x} }{n}$	=AVEDEV(Data)	Distancia média dos valores ao centro – fácil de entender	.

Dispersão

* *Medida de tendência central x Dispersão*

- Notas atribuidas a professores por alunos
- Escala de 10-pontos

Attribute	Same Mean, Different Variance		Different Mean, Same Variance	
	Prof. Wu	Prof. Jones	Prof. Smith	Prof. Gopal
1. Challenging	6.1	5.1	5.8	6.3
2. Approachable	6.7	5.5	6.4	6.9
3. Enthusiastic	6.9	5.9	6.6	7.1
4. Helps students	7.0	6.4	6.8	7.3
5. Fair exams	7.4	7.8	7.3	7.8
6. Knowledge	7.5	8.3	7.3	7.8
7. Lecture ability	7.6	9.2	7.5	8.0
8. Organized	8.4	9.4	8.3	8.8
Mean	7.20	7.20	7.00	7.50
Std Dev	0.69	1.69	0.77	0.77
CV	9.6%	23.5%	11.0%	10.2%

Dispersão

- Jones and Wu tem médias idênticas mas desvios-padrão diferentes.

Same Mean, Different Variance		
Attribute	Prof. Wu	Prof. Jones
1. Challenging	6.1	5.1
2. Approachable	6.7	5.5
3. Enthusiastic	6.9	5.9
4. Helps students	7.0	6.4
5. Fair exams	7.4	7.8
6. Knowledge	7.5	8.3
7. Lecture ability	7.6	9.2
8. Organized	8.4	9.4
Mean	7.20	7.20
Std Dev	0.69	1.69
CV	9.6%	23.5%

Dispersão

- Smith and Gopal apresentam médias diferentes mas desvios-padrão identicos

Attribute	<i>Different Mean, Same Variance</i>	
	Prof. Smith	Prof. Gopal
1. Challenging	5.8	6.3
2. Approachable	6.4	6.9
3. Enthusiastic	6.6	7.1
4. Helps students	6.8	7.3
5. Fair exams	7.3	7.8
6. Knowledge	7.3	7.8
7. Lecture ability	7.5	8.0
8. Organized	8.3	8.8
Mean	7.00	7.50
Std Dev	0.77	0.77
CV	11.0%	10.2%

Dispersão

- Uma média alta (melhor colocado no ranking) e baixo desvio-padrão (maior consistência) é o resultado desejado.
- Qual professor é o melhor na sua opinião?

Attribute	Same Mean, Different Variance		Different Mean, Same Variance	
	Prof. Wu	Prof. Jones	Prof. Smith	Prof. Gopal
Mean	7.20	7.20	7.00	7.50
Std Dev	0.69	1.69	0.77	0.77
CV	9.6%	23.5%	11.0%	10.2%

Tarefa da semana

- Extrair um conjunto de dados
- Estudar estes dados
 - Medidas de sumarização
 - Medidas de dispersão
- Estudo: Desempenho das escolas públicas x Escolas privadas, baseado nos dados do INEP
 - Escolher medida de desempenho
 - Escolher um estado
 - Ver instruções detalhadas no SIGAA