Estatística Descritiva

PNV-3421 – Processos Estocásticos

Prof. Dr. André Bergsten Mendes

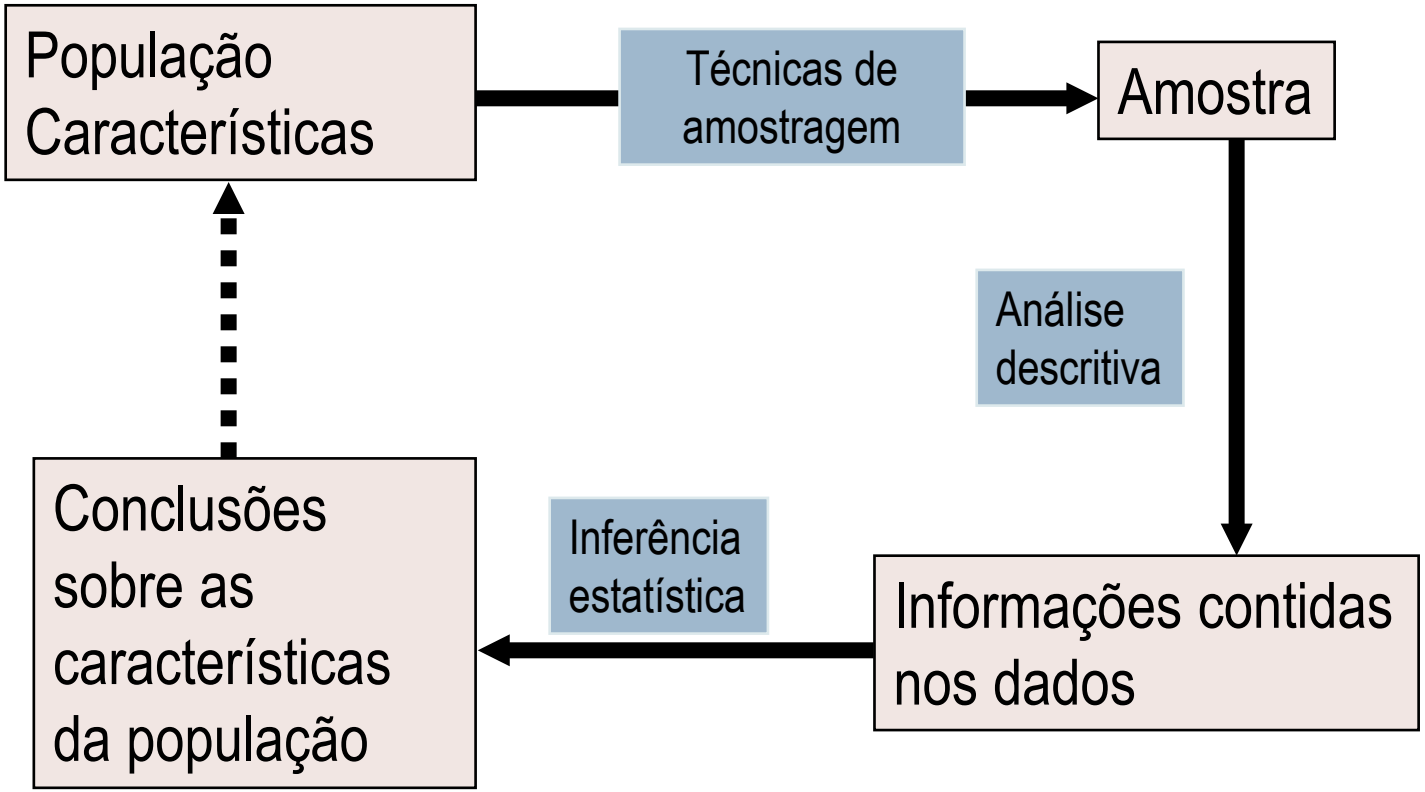
Estatística

- Conjunto de técnicas que permite, de forma sistemática, coletar, organizar, descrever, analisar e interpretar dados oriundos de experimentos, realizados em qualquer área do conhecimento.

Exemplos

- Programas de Previdência Social e de Saúde Pública - previsões sobre a longevidade da população;
- Que fatores aumentam o risco de um indivíduo desenvolver uma doença cardíaca coronariana ?
- Um novo produto deve ser lançado no mercado? O produto encontrará seu nicho?
- Previsão de demanda de produtos;
- Níveis mínimos, médios e máximos de estoques.

Estatística



Fonte dos dados

1. Consulta a banco de dados existentes (dados históricos).
2. Especificação do fabricante.
3. Estimativa de gestores.
4. Dados coletados em tempo real.
5. Observação direta.

Dados Determinísticos x Probabilísticos

- Dados determinísticos
  - *Exemplo*: máquina de controle numérico computadorizado; intervalos para realização de manutenção preventiva; velocidade de esteiras (constante).
- Dados probabilísticos
  - *Exemplo*: intervalo entre chegadas; tempo de atendimento de um cliente; tempo de reparo.

Alguns Problemas na Coleta de Dados

- Os dados são antigos (desatualizados);
- Amostra pequena;
- Formato errado (discreto x contínuo);
- Dados estão registrados em classes (não existe o dado bruto);
- Os dados não são representativos;
- Os *outliers* foram omitidos;
- Dúvidas e critérios distintos de preenchimento;
- Os dados agrupam processos complexos;
- Não há documentação.

Estatística Descritiva

O que fazer com as observações que coletamos?

Resumo dos dados = Estatística Descritiva

Estatística Descritiva

- Variável: Qualquer característica associada a uma população.
- Classificação das variáveis:
  - Qualitativa
    - **Nominal** – Ex: nacionalidade, cor dos olhos
    - **Ordinal** – Ex: grau de instrução, ranking esportivo
  - Quantitativo
    - **Discretos** – Ex: número de pessoas que chegam a um sistema; número de tarefas que uma máquina processa antes de uma falha.
    - **Contínuos** – Ex: intervalo de tempo entre chegadas; tempos de serviço; tempos de viagem.

Estatística Descritiva

Estatística Descritiva

- **Medidas de Posição**: Média, Média Aparada, Média Ponderada, Mediana, Moda.
- **Medidas de Dispersão**: Variância, Desvio Padrão, Coeficiente de Variação, Desvio Absoluto Médio, Amplitude, Percentis, Intervalo-Interquartil, Identificação de Outliers.

Medidas de Posição
1. Média

- ***Média:*** é a soma de todos os valores dividido pelo número de valores.
- $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
- **Exemplo**:

  Dados: 2, 3, 5, 7, 8

  Média: 5

Medidas de Posição
2. Média Aparada

- ***Média aparada:*** estando o conjunto de dados ordenados, elimina-se uma quantidade de fixa de dados, em cada ponta, e a média é calculada para os demais valores. Isto elimina a influência de valores extremos.
- $\bar{x} = \frac{\sum_{i=p+1}^{n-p} x_{(i)}}{n-2p}$
- **Exemplo**:

  Dados: 1, 1, 2, 2, 3, 5, 7, 8, 9, 11, 13, 80

  Média: 11,8

  Média aparada (*p*=1): 6,1

Medidas de Posição
3. Média Ponderada

- ***Média ponderada:*** se cada dado possuir um peso distinto, então a média pode ser calculada como:
- $\bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$
- Exemplo: supor que os dados de foram coletados por sensores. Sabendo que alguns sensores estão defeituosos, pesos menores serão atribuídos a estes dados.

Medidas de Posição
4. Mediana

- ***Mediana:*** é o valor da variável que ocupa a posição central de um conjunto de *n* dados ordenados.
- **Exemplo**:

1) Dados: 2, 5, 3, 7, 8

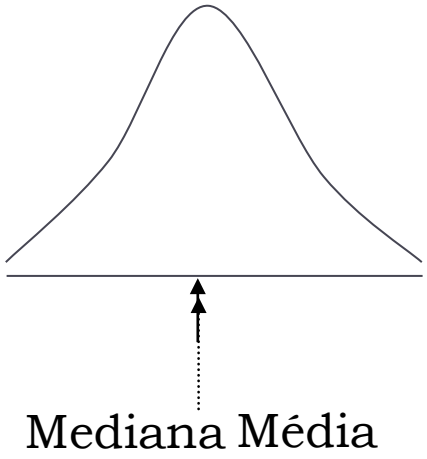
Dados ordenados: 2, 3, 5, 7, 8 => (5+1)/2 = 3 => Md = 5

2) Dados: 3, 5, 2, 1, 8, 6

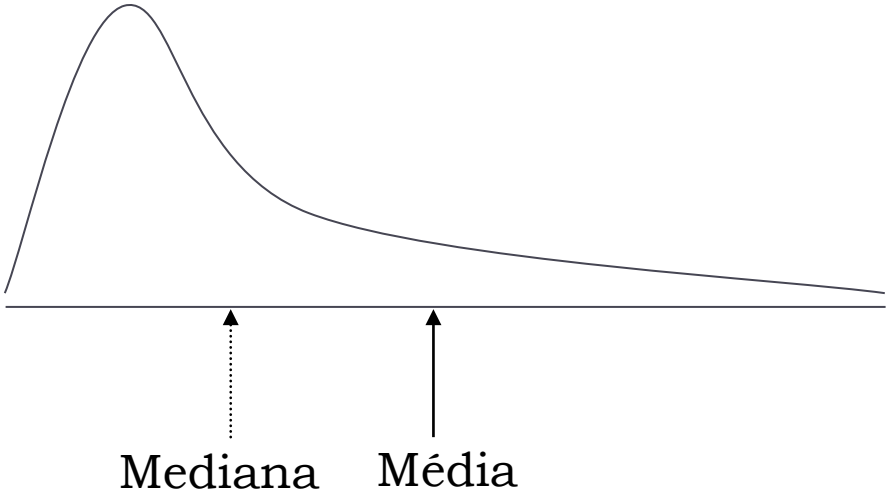
Dados ordenados: 1, 2, 3, 5, 6, 8 => (6+1)/2=3,5 => Md = (3+5)/2=4

Medidas de Posição
4. Mediana

Distribuição Simétrica



Distribuição Assimétrica



Medidas de Posição
5. Moda (p/dados discretos)

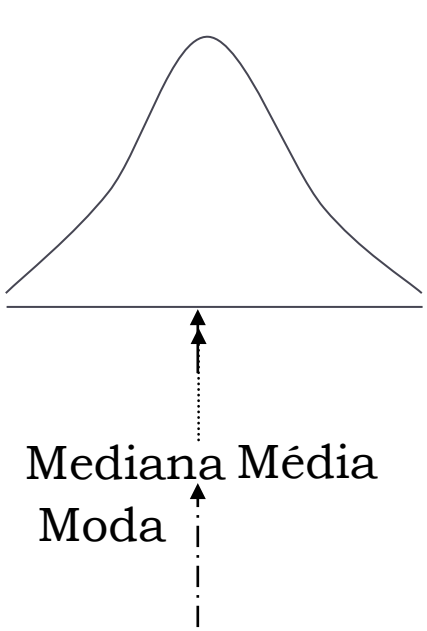
- ***Moda***: É o valor (ou atributo) que ocorre com maior frequência
- **Exemplo**:

  Dados: 3, 4, 5, 4, 6, 5, 8, 4, 4, 9, 10

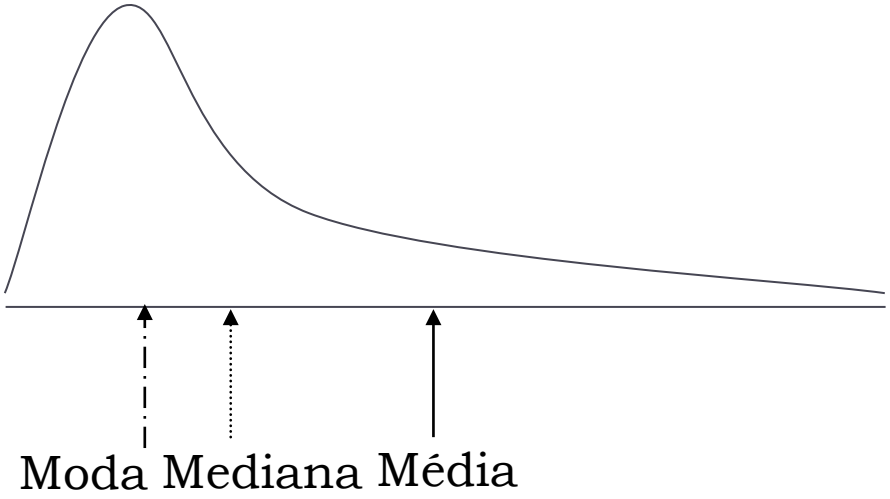
  Moda: 4

Medidas de Posição
5. Moda

Distribuição Simétrica



Distribuição Assimétrica



Medidas de Dispersão
1. Variância

- **Variância (s^2):** É a média da diferença quadrática entre cada valor de uma série de dados e a respectiva média.
- $$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Medidas de Dispersão
2. Desvio Padrão

- **Desvio Padrão (s)**: É a raiz quadrada da média.
- $s = \sqrt[2]{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$

Medidas de Dispersão
3. Coeficiente de Variação

- **Coeficiente de Variação**: É a divisão do desvio padrão pela média. É adimensional, e permite comparar fenômenos distintos quanto à variabilidade.
- $cv = s/\bar{x}$

Medidas de Dispersão
4. Desvio Absoluto Médio

- **Desvio absoluto médio**: É a média dos desvios em relação à média.
- $MAD = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$

Medidas de Dispersão
Exemplo

i	x_i	$x_i - \bar{x}$	$\lvert x_i - \bar{x} \rvert$	$(x_i - \bar{x})^2$
1	1	-9	9	81
2	3	-7	7	49
3	5	-5	5	25
4	7	-3	3	9
5	9	-1	1	1
6	11	1	1	1
7	13	3	3	9
8	15	5	5	25
9	17	7	7	49
10	19	9	9	81
Soma	100	0	50	330
Média	10	0	5	33

- $\bar{x} = 10 \quad s^2 = \frac{330}{9} = 36.7 \quad s = 6.1 \quad cv = 0.6 \quad MAD = 5$

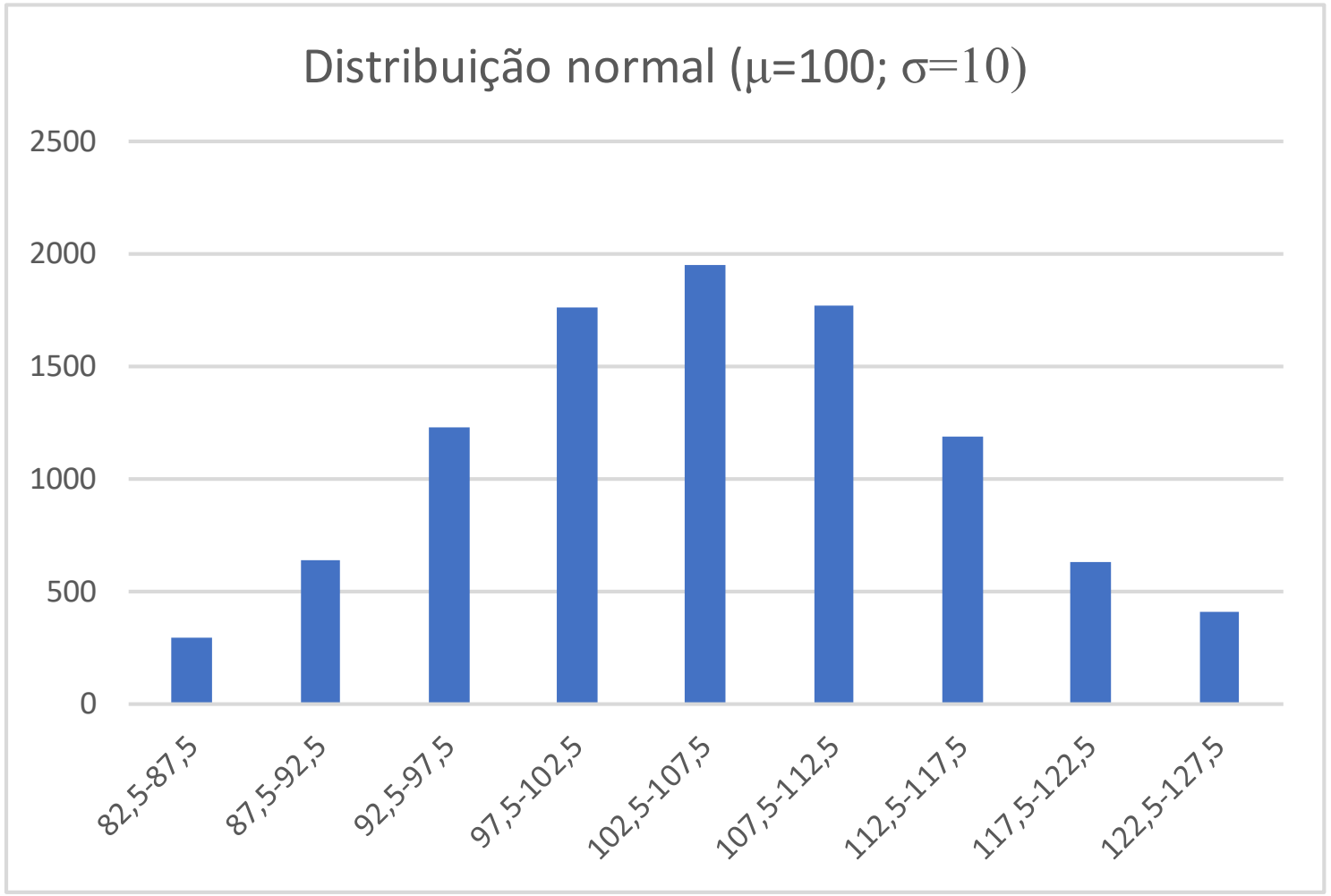
Medidas de Dispersão
5. Amplitude

- ***Amplitude(A)***: máximo – mínimo.

Medidas de Dispersão
6. Medida de Assimetria

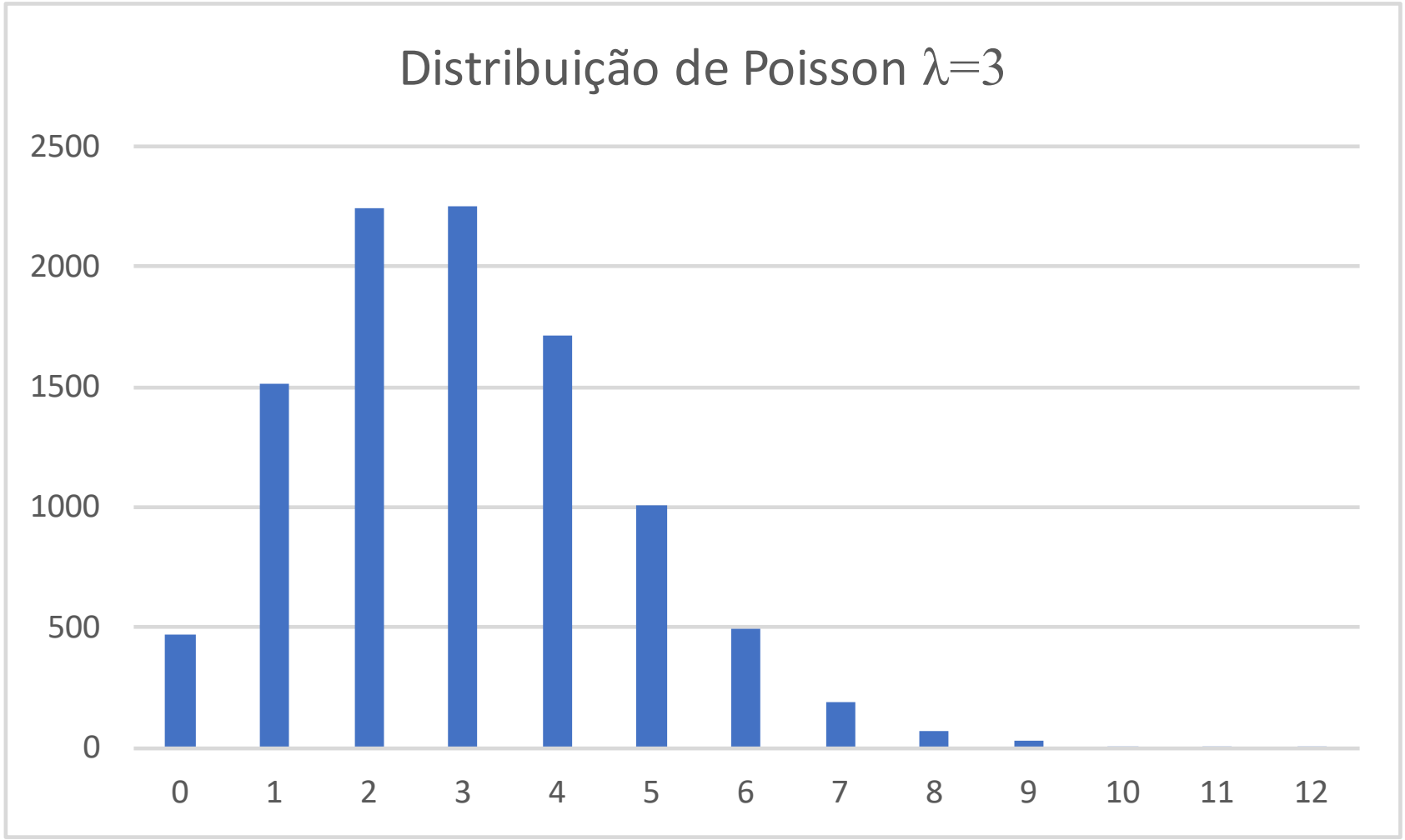
- ***Assimetria***: É a medida de falta de simetria. Um distribuição, ou um conjunto de dados, será simétrica, se tiver o mesmo perfil à esquerda e à direita do ponto central.
- $g_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3/n}{s^3}$
- OBS: i) para o cálculo da medida de assimetria, o desvio padrão deve ser calculado tendo no denominador *n*, ao invés de *n-1*; ii) g_1 é chamado de coeficiente de assimetria de Fisher-Pearson.

Medidas de Dispersão
6. Medida de Assimetria



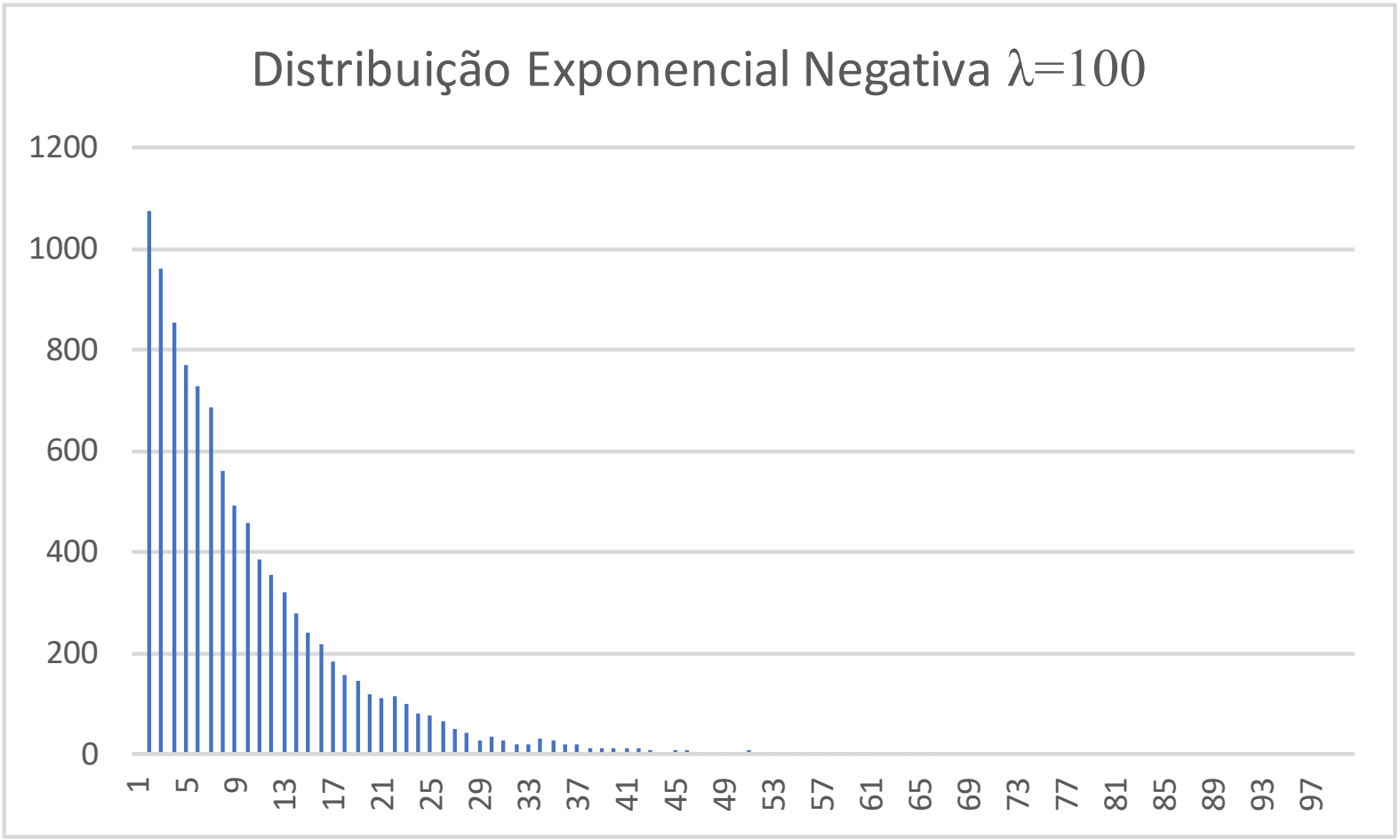
$$g_1 = 0{,}023$$

Medidas de Dispersão
6. Medida de Assimetria



$$g_1 = 0{,}5876$$

Medidas de Dispersão
6. Medida de Assimetria

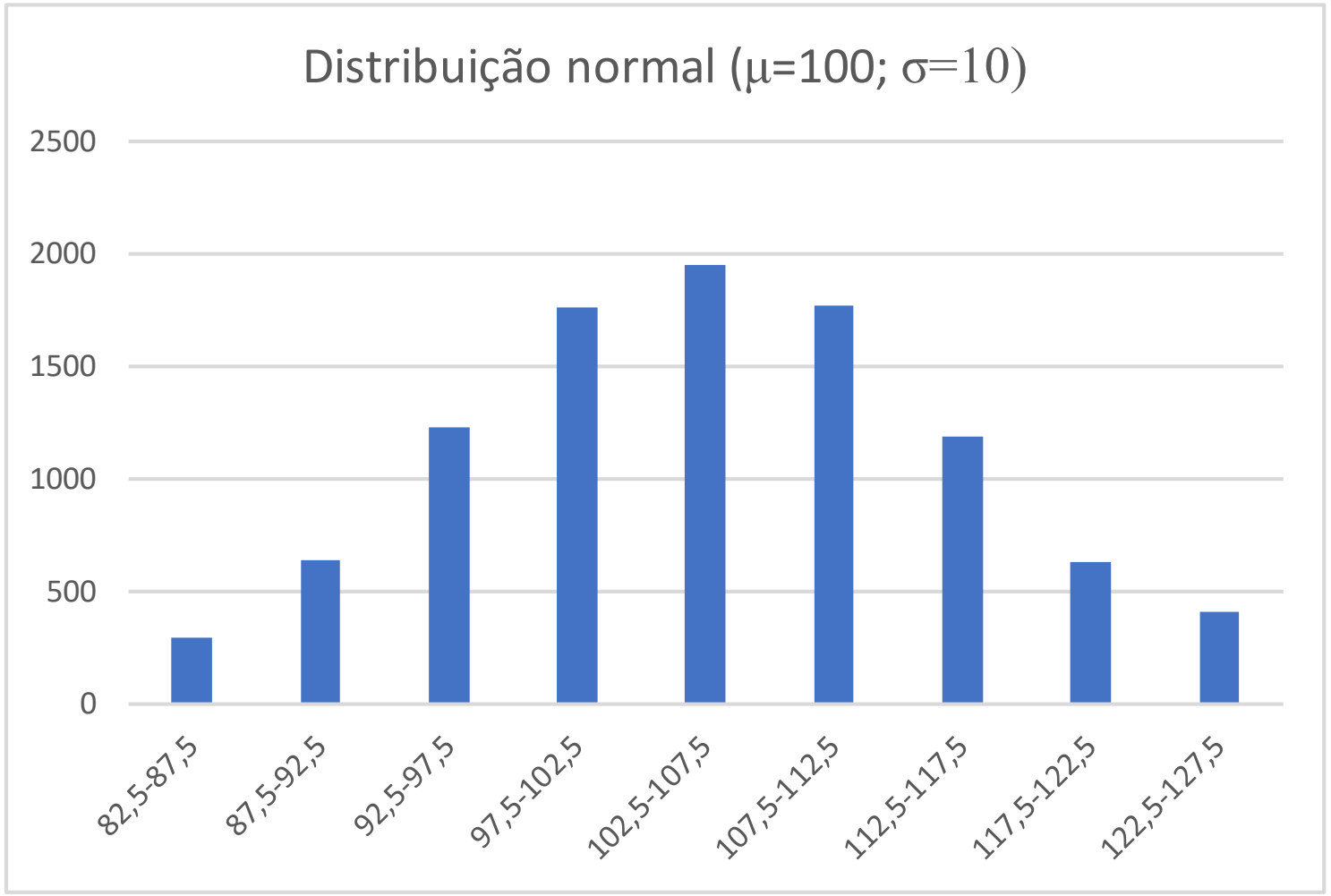


$$g_1 = 1{,}90$$

Medidas de Dispersão
7. Curtose

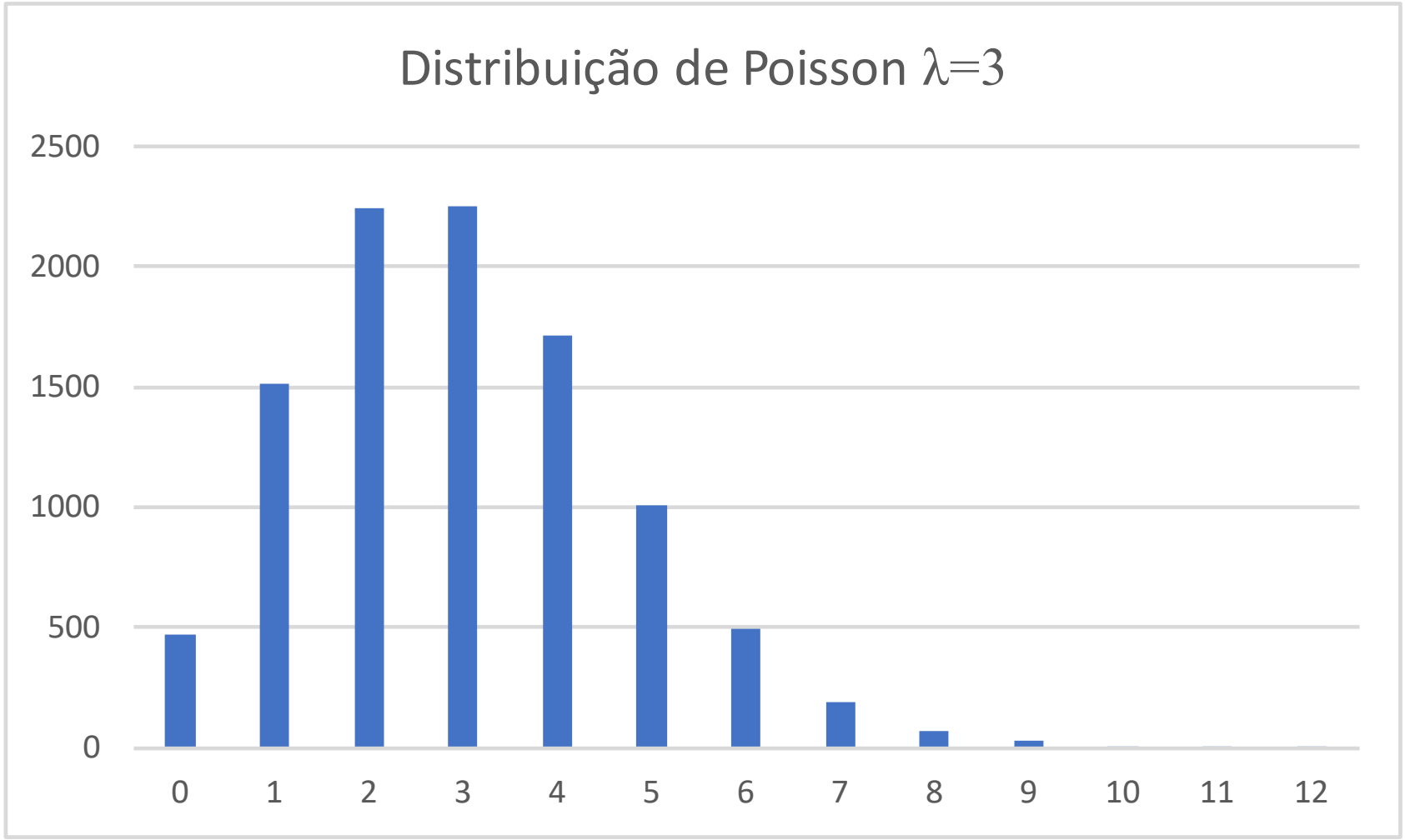
- ***Curtose***: Curtose é uma medida de quão achata é a distribuição dos dados, em relação à distribuição normal. Conjunto de dados com alta curtose tende a ter cauda longa ou outliers. Conjunto de dados com baixa curtose tende a ser uma cauda curta, ou falta de outliers.
- $curtose = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4/n}{s^4}$ ou $\frac{\sum_{i=1}^{n}(x_i - \bar{x})^4/n}{s^4} - 3$
- OBS: para o cálculo da curtose, o desvio padrão deve ser calculado tendo no denominador *n*, ao invés de *n-1*.

Medidas de Dispersão
7. Curtose



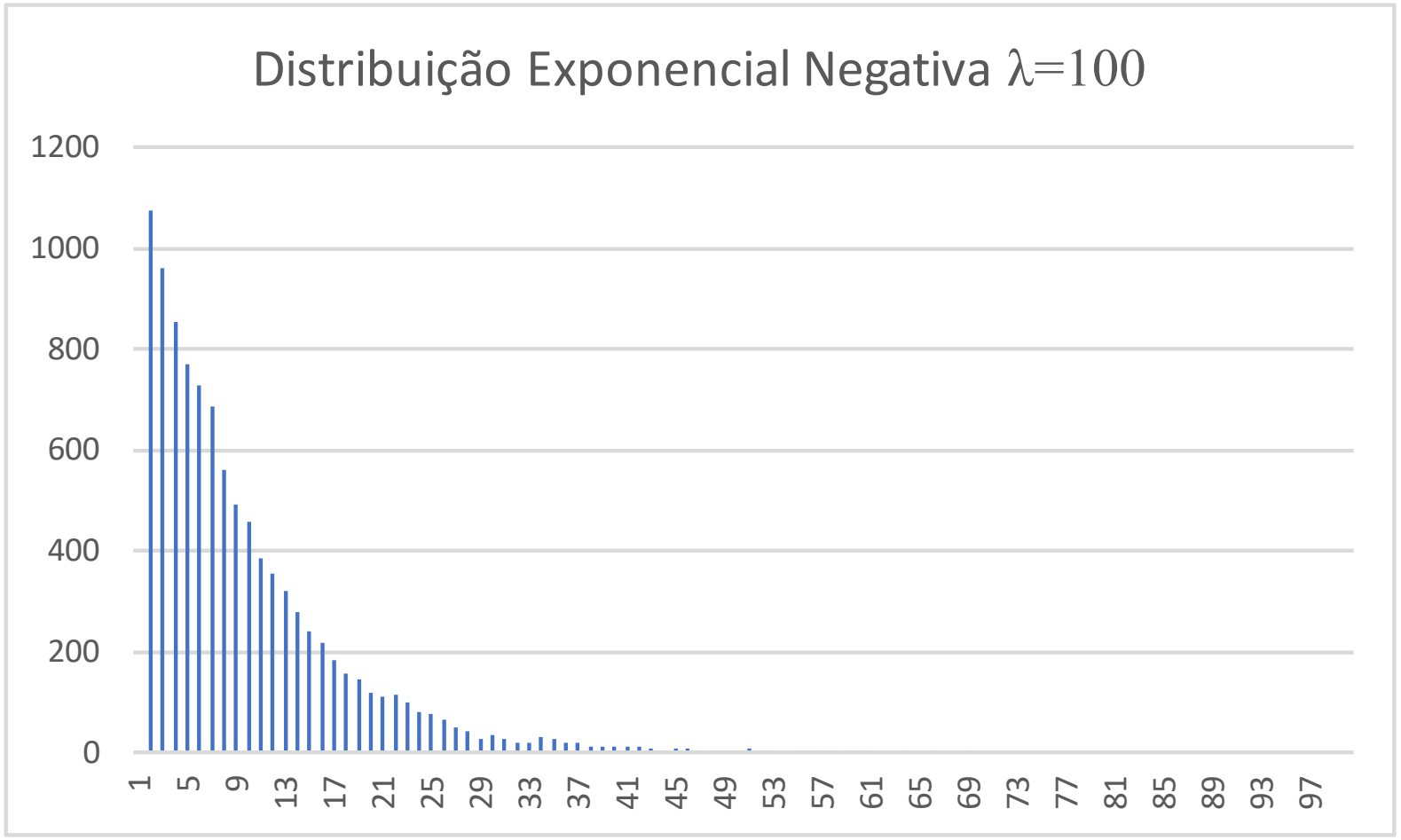
$$curtose = -0{,}03$$

Medidas de Dispersão
7. Curtose



$$curtose = 0{,}447$$

Medidas de Dispersão
7. Curtose



$$curtose = 5{,}14$$

Medidas de Dispersão
8. Percentis

- ***Percentis*:** O percentil de ordem $100*p$ ($0<p<1$), em um conjunto de dados de tamanho n, é o valor da variável que ocupa a posição $(n+1)*p$ do conjunto de dados ordenados.

  O percentil de ordem p (ou p-quantil) deixa $100\%p$ das observações abaixo dele na amostra ordenada.

Medidas de Dispersão
8. Percentis

- ***Casos Particulares***

  Percentil 25%: primeiro quartil (Q1)

  Percentil 50%: segundo quartil (Q2) ou mediana

  Percentil 75%: terceiro quartil (Q3)

Medidas de Dispersão
8. Percentis

- **Exemplo 1**: {1,9 2,0 2,1 2,5 3,0 3,1 3,3 3,7 6,1 7,7} → n=10

  Q1: 0,25(11)=2,75 → Q1 = (2+2,1)/2 = 2,05

  Q2: 0,50(11)=5,50 → Q2 = (3+3,1)/2 = 3,05

  Q3: 0,75(11)=8,25 → Q3 = (3,7+6,1)/2 = 4,9

- **Exemplo 2**: {0,9 1,0 1,7 2,9 3,1 5,3 5,5 12,2 12,9 14,0 33,6} → n=11

  Q1=1,7 Q2=5,3 Q3=12,9

Medidas de Dispersão
9. Intervalo Interquantil

- ***Intervalo Interquartil*** (*d*) - É a diferença entre o terceiro e o primeiro quartil, *d*=Q3-Q1
- **Exemplo**:

Dados: 15, 5, 3, 8, 10, 2, 7, 11, 12

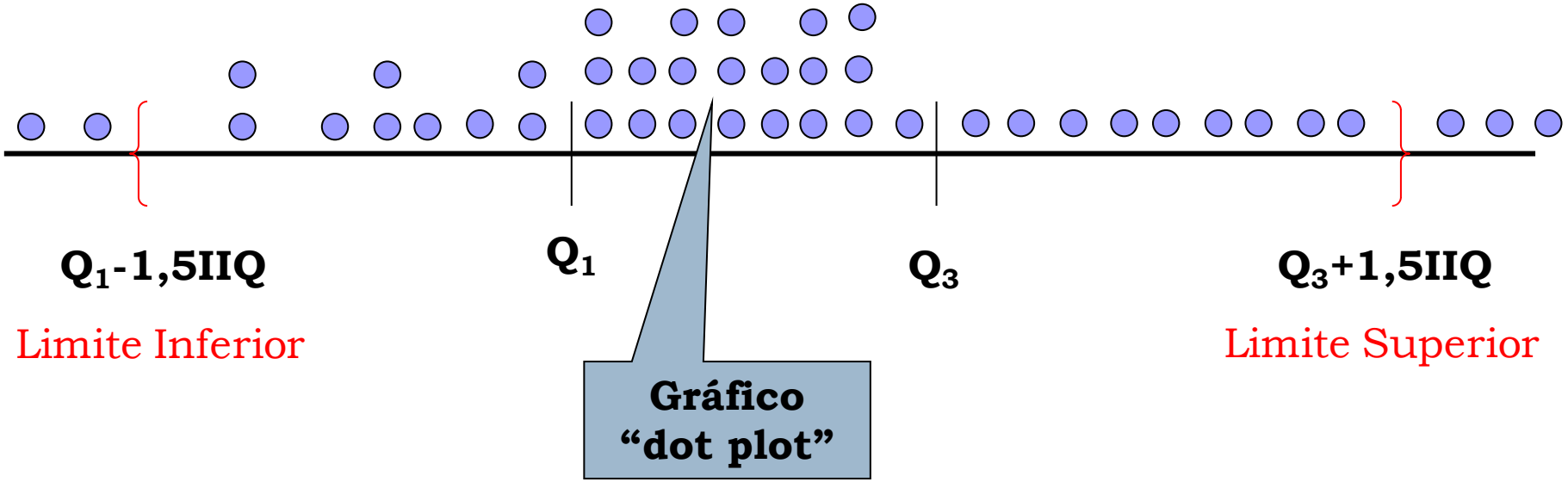
Dados ordenados: 2, 3, 5, 7, 8, 10, 11, 12, 15

Q1=4; Q3=11.5; d=11.5 – 4,0 = 7.5

Medidas de Dispersão
9. Identificação de *outliers*

- **Identificação de *outliers***

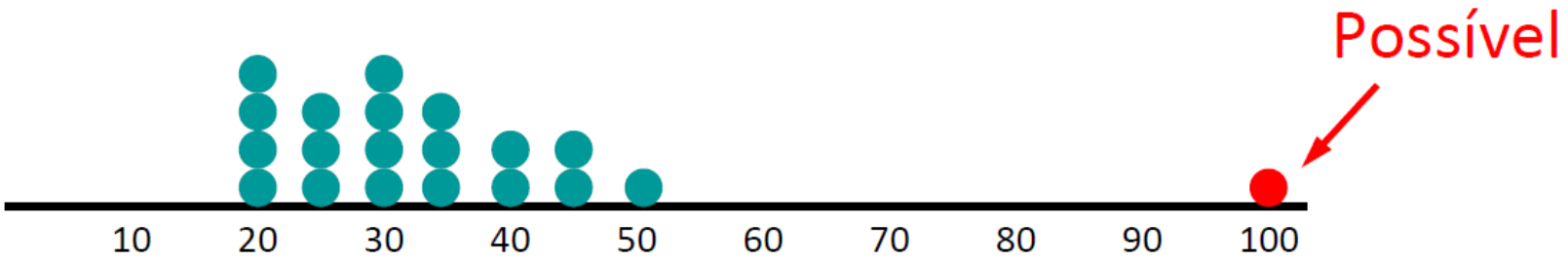
$$Intervalo\ Inter\ Quartil(IIQ) = Q_3 - Q_1$$



Medidas de Dispersão
9. Identificação de *outliers*

Idade

Notas de um exame



Medidas de Dispersão
9. Identificação de *outliers*

- **Exemplo** - Considere os tempos de viagem de veículos entre uma fábrica em São Paulo e um cliente localizado no sul do estado de MG. Faça uma análise estatística dos tempos de viagem, que permita dar subsídios a uma renegociação dos prazos de entrega. Para isso, o setor responsável listou o apontamento feito das últimas 187 viagens, que estão listados no arquivo Exemplo 01 - Tempos de viagens.xlsx.

Medidas de Dispersão
9. Identificação de *outliers*

Tempos de viagem

13	15	5	5	20	0	10	20	0	20
0	44	35	7	0	5	10	660	5	20
5	10	70	5	5	5	0	5	0	10
5	0	0	10	9	50	0	10	0	10
0	5	0	0	5	0	0	15	5	28
15	5	0	5	5	0	8	10	10	57
5	6	5	10	5	10	15	70	5	0
5	10	10	0	6	10	60	0	15	
10	0	0	0	35	0	2	0	0	
20	5	15	5	95	5	10	5	0	
0	5	0	10	0	10	0	0	5	
0	5	75	1245	0	0	0	10	5	
10	8	0	0	0	0	5	0	5	
5	0	10	5	0	5	20	10	0	
0	5	5	13	30	7	10	10	15	
0	0	5	0	0	75	15	5	5	
5	5	5	0	5	0	0	5	7	
5	5	5	5	5	0	5	5	10	
10	0	10	5	5	4	5	40	30	
15	5	5	15	5	5	10	0	5	

Medidas de Dispersão
9. Identificação de *outliers*

Valor	Freq	Valor	Freq
0	55	30	2
2	1	35	2
4	1	40	1
5	59	44	1
6	2	50	1
7	3	57	1
8	2	60	1
9	1	70	2
10	28	75	2
13	2	95	1
15	10	660	1
20	6	1245	1
28	1		

Erro de Apontamento!

Erro ? Pode Acontecer?

Medidas de Dispersão
9. Identificação de *outliers*

Medida Descritiva	**Amostra\ {0}**	**Amostra\ {0, 1245}**	**Amostra\{0, 660, 1245}**
Média	27,5	18,2	13,3
Média Aparada 5%	11,1		
Mediana	7,5	7,0	7,0
Q1	5,0	5,0	5,0
Q3	*13,5*	13,0	12,3
d (IIQ)	8,5	8,0	7,3
Limite Inferior	-7,75	-7	-5,875
Limite Superior	26,25	25	23,125
Amplitude	1243,0	658,0	93,0
Desvio Padrão	121,8	58,8	16,4
Variância	14834,2	3459,5	269,1
Coeficiente Var.	4,4	3,2	1,2
N	132	131	130
Curtose	82,5	111,2	8,9
Coeficiente Assimetria	7,6	10,0	1,9

Medidas de Dispersão
9. Identificação de *outliers*

Medida Descritiva	Amostra\{0}
Média	=MÉDIA($H$3:$H$134)
Média Aparada 5%	=MÉDIA(H9:H127)
Mediana	=MED($H$3:$H$134)
Q1	=QUARTIL($H$3:$H$134;1)
Q3	=QUARTIL($H$3:$H$134;3)
d (IIQ)	=M7-M6
Limite Inferior	=M6-1,5*M8
Limite Superior	=M7+1,5*M8
Amplitude	=$H$134-$H$3
Desvio Padrão	=DESVPADA($H$3:$H$134)
Variância	=VARA($H$3:$H$134)
Coeficiente Var.	=M12/M3
N	=CONT.NÚM(H3:H134)
Curtose	=CURT($H$3:$H$134)
Coeficiente Assimetria	=M31/POTÊNCIA(M12;3)

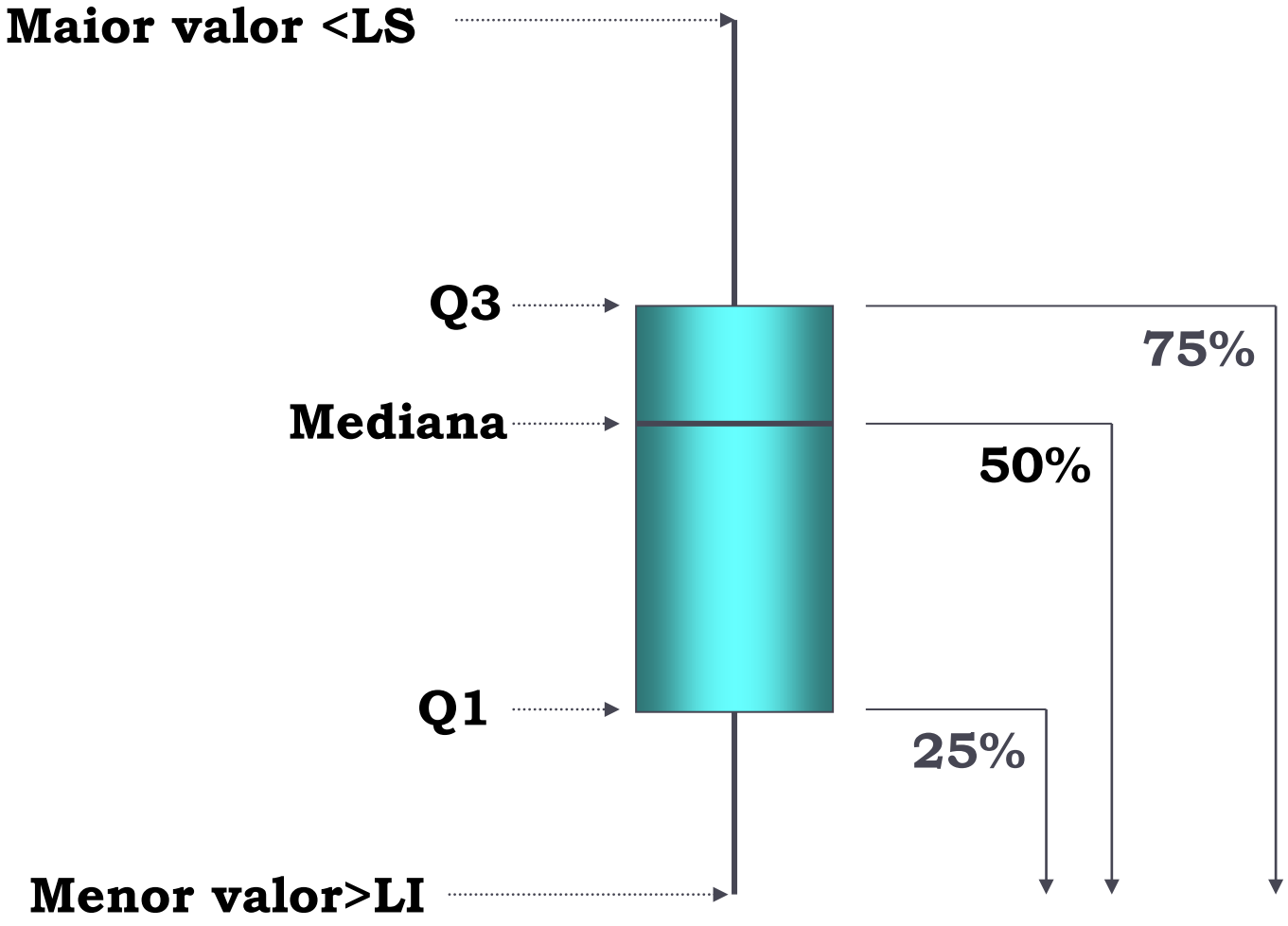
Medidas de Dispersão
9. Identificação de *outliers*

Medida Descritiva	**Amostra\ {0}**	**Amostra\ {0, 1245}**	**Amostra\{0, 660, 1245}**
Limite Inferior	-7,75	-7	-5,875
Limite Superior	26,25	25	23,125
"Menor Valor"	2	2	2
"Maior Valor"	20	20	20
# dados outliers	17	18	15

OBS: É razoável considerar o tempo = 2h como correto? Se não... eliminar, à semelhança do 0.

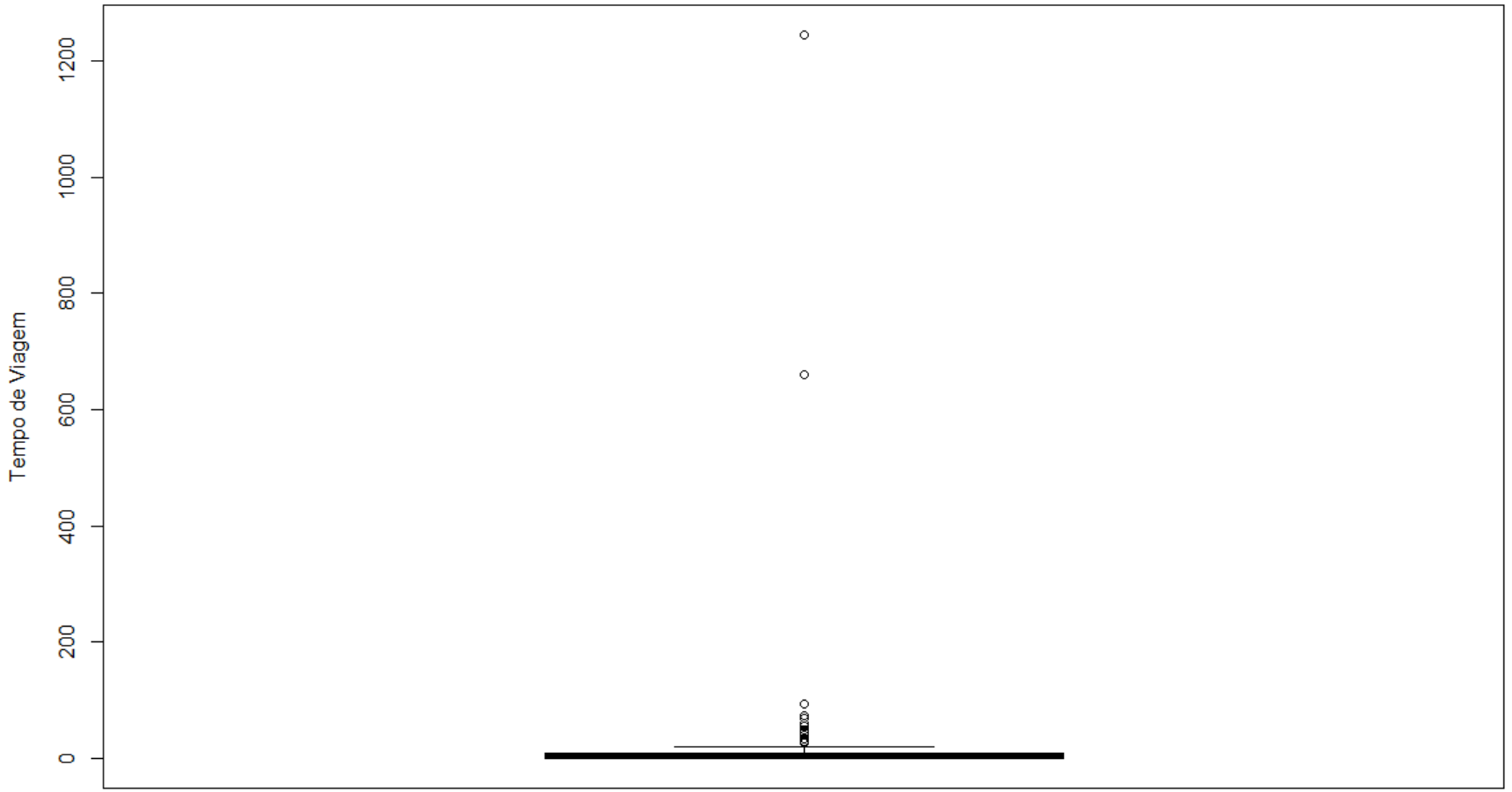
Medidas de Dispersão
10. Gráfico BoxPlot

***LS*=Q3+1,5(Q3-Q1)**

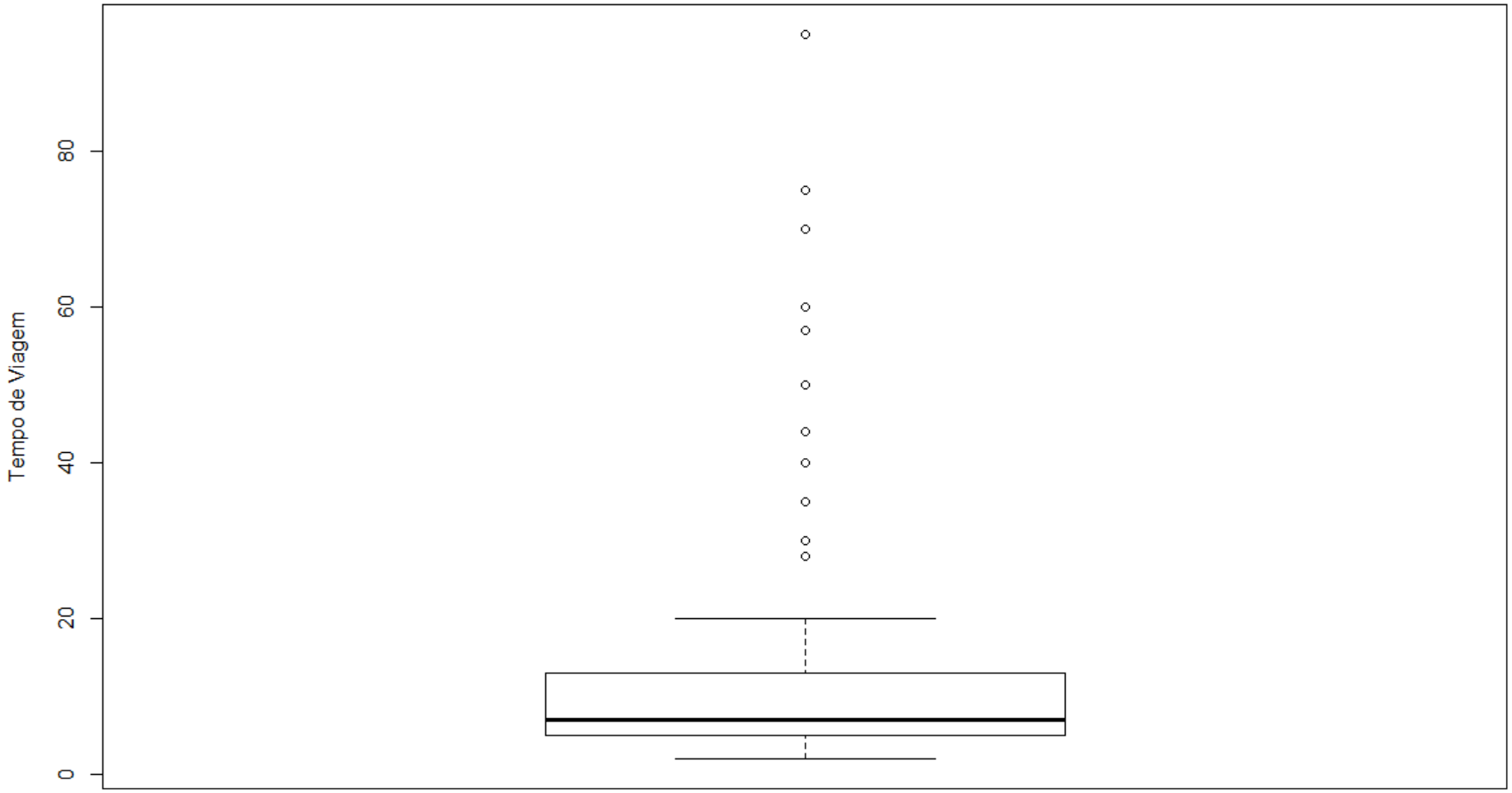


***LI*=Q1-1,5(Q3-Q1)**

Medidas de Dispersão
10. Gráfico BoxPlot



Medidas de Dispersão
10. Gráfico BoxPlot (excluindo 0s, 660, 1245)



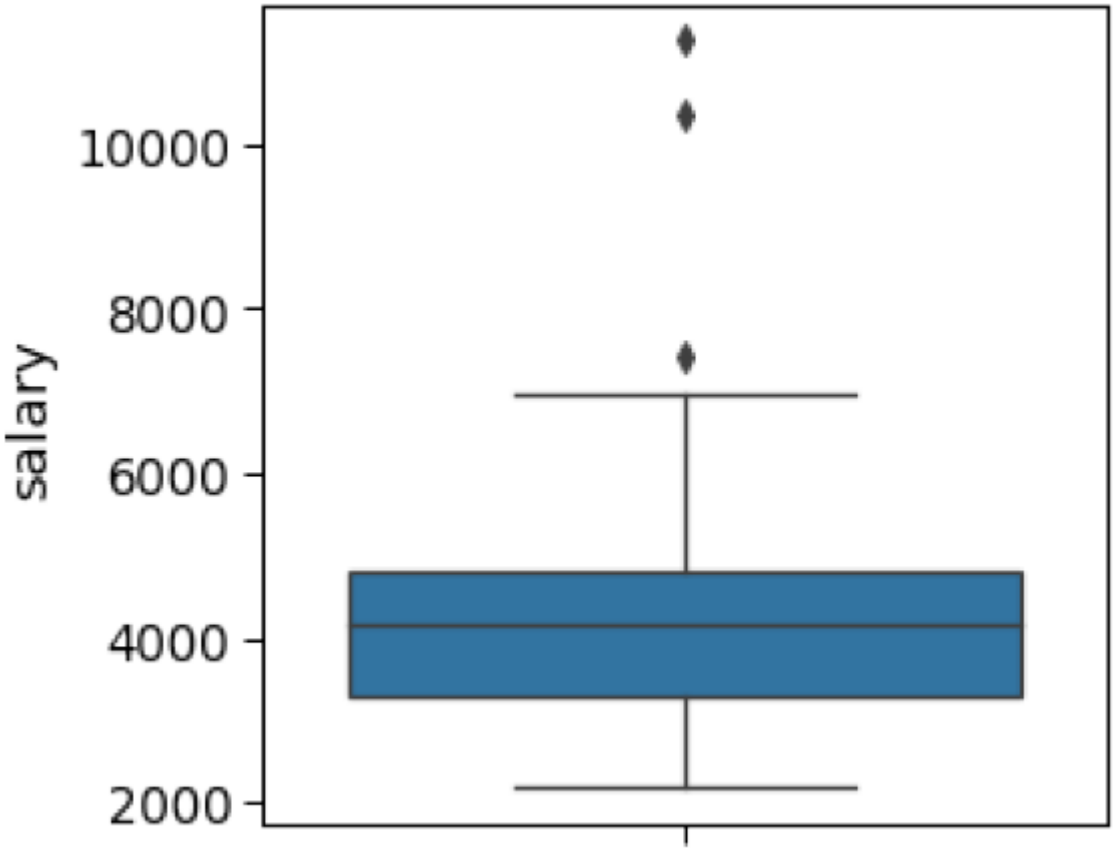
Medidas de Dispersão
10. Gráfico BoxPlot no Python

```
# Import the required libraries
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline

# load the dataset
df = pd.read_csv("StatFun2_4_1.csv")
```

Medidas de Dispersão
10. Gráfico BoxPlot no Python

```
# Set the plot size
plt.figure(figsize=(4,4))

# Create and show the box plot
sns.boxplot(y=df['salary'])
plt.show()
```

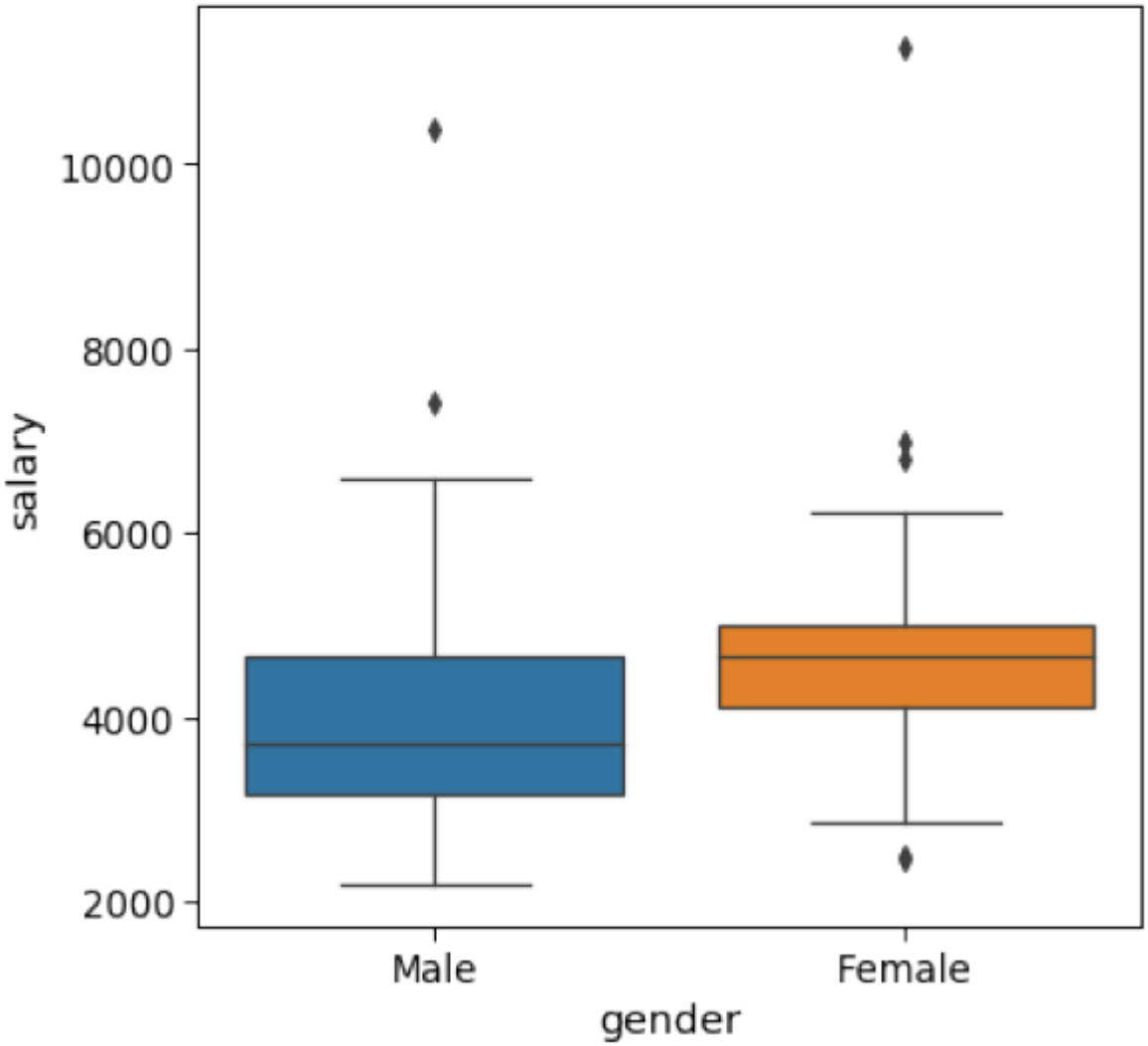
Medidas de Dispersão
10. Gráfico BoxPlot no Python



Medidas de Dispersão
10. Gráfico BoxPlot no Python

```
# Set the plot size
plt.figure(figsize=(6,6))

# Create box plots for each gender
sns.boxplot(x='gender', y='salary', data=df)
plt.show()
```

Medidas de Dispersão
10. Gráfico BoxPlot no Python



Medidas de Dispersão
10. Gráfico BoxPlot no Python

```
# Do not display outliers
## Do it by setting sym to ""
plt.figure(figsize=(6,6))
sns.boxplot(x='gender', y='salary', data=df,
                    sym="")
plt.show()
```

Medidas de Dispersão
10. Gráfico BoxPlot no Python

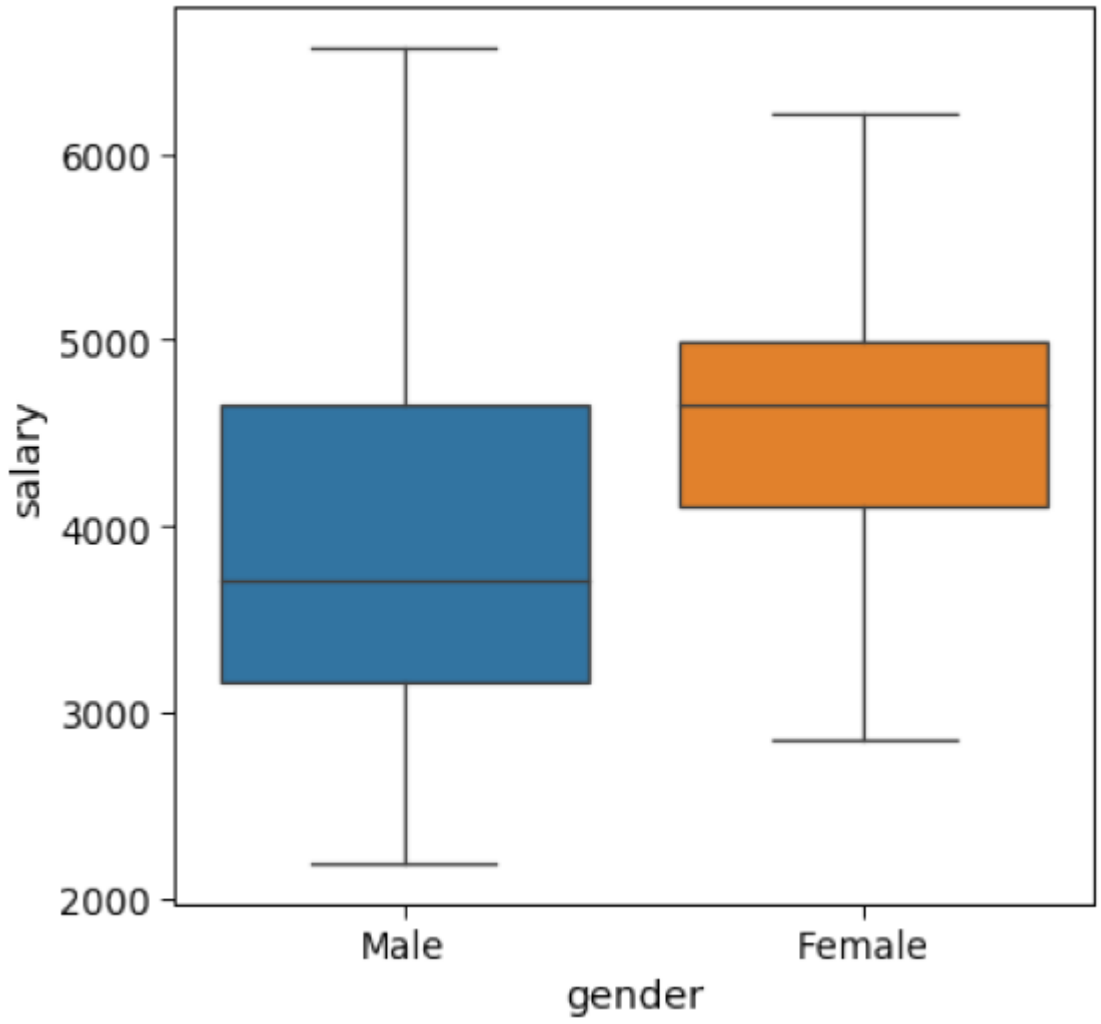


Tabela de Frequências

- Distribuição de frequências de uma variável é uma lista dos valores individuais ou dos intervalos de valores que a variável pode assumir, com as respectivas frequências de ocorrência.
- Histograma é uma forma de visualização de uma tabela de frequências.

Estatística Descritiva

Representação Gráfica

Intervalo entre Chegadas (tempos em segundos)

5	41	3	3	68	1	38	82	82	20
7	43	24	51	6	12	9	86	90	120
50	7	40	42	40	13	8	19	27	28
26	0	4	37	45	11	21	19	55	8
54	39	44	34	2	1	32	22	2	1
5	20	23	35	31	106	3	2	62	71
29	25	30	3	3	24	27	33	66	3
3	68	6	3	33	33	81	9	15	14
16	50	49	50	49	100	13	17	110	3
0	39	15	14	16	40	9	13	17	5

Mínimo: 0; **Máximo**: 120; **# Dados**: 100; **Média**: 30; **Número de classes** ~ 10

Estatística Descritiva

- **Representação Gráfica**

Bloco	Freq.	Freq. Acum	%	% cumul.
< 12	32	32	32,0%	32,0%
12 a 24	20	52	20,0%	52,0%
24 a 36	14	66	14,0%	66,0%
36 a 48	12	78	12,0%	78,0%
48 a 60	8	86	8,0%	86,0%
60 a 72	5	91	5,0%	91,0%
72 a 84	3	94	3,0%	94,0%
84 a 96	2	96	2,0%	96,0%
96 a 108	2	98	2,0%	98,0%
108 a 120	2	100	2,0%	100,0%
> 120	0	100	0%	100,0%

Estatística Descritiva

- **Histograma**