2 CONCEITOS DE ESTATÍSTICA

Quando se trata de simulação, uma das grandes vantagens em utilizá-la é a possibilidade de se trabalhar com variáveis aleatórias, diferentemente de programação linear, por exemplo, que trabalha com parâmetros fixos.

O objetivo deste capítulo é apresentar e rever alguns conceitos básicos de estatística importantes para a modelagem de sistemas e para a interpretação dos resultados das iterações.

Alguns conceitos que não são neste capítulo abordados, são apresentados em outros capítulos, quando houver a necessidade de aprofundamento para aplicações e análises específicas.

2.1 Distribuições de frequências

Moore et al. (2006) definem estatística como sendo a ciência dos dados. Mas, o que são dados? Anderson, Sweeney e Williams (2002) apresentam a definição de dados como sendo “os fatos e números coletados, analisados e sintetizados para a apresentação e interpretação. Juntos, os dados coletados em um estudo particular são denominados conjunto de dados”. Assim, começa-se este estudo pela distribuição de frequências. Frequência é a quantidade de vezes que um dado se repete em um conjunto de dados. Como exemplo, parte-se da Tabela 2.1 que apresenta os volumes de vendas mensais de diversas obras em uma livraria:

Tabela 2.1: Vendas mensais de uma livraria.

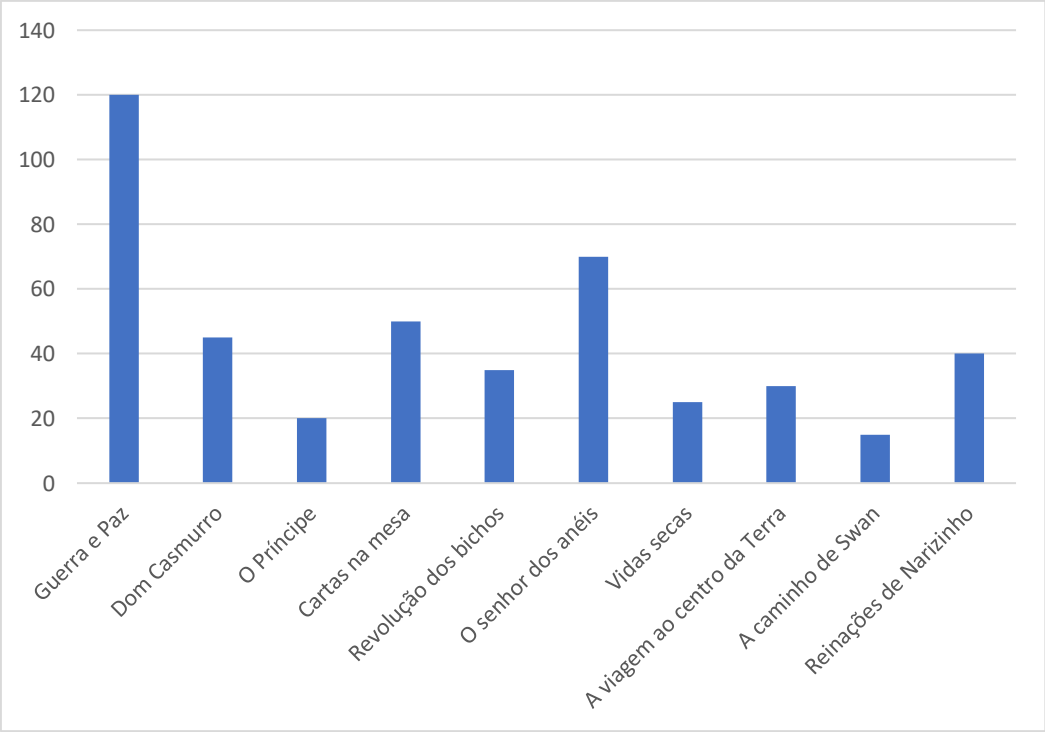
Título	Frequência (unidades vendidas)	Percentual
Guerra e Paz	120	26,7%
Dom Casmurro	45	10,0%
O Príncipe	20	4,4%
Cartas na mesa	50	11,1%
Revolução dos bichos	35	7,8%
O senhor dos anéis	70	15,6%
Vidas secas	25	5,6%
A viagem ao centro da Terra	30	6,7%
A caminho de Swan	15	3,3%
Reinações de Narizinho	40	8,9%

Fonte: o autor

O Título, na primeira coluna, representa a categoria do dado, isto é, o tipo. Neste caso, são dados qualitativos, não representam quantidades. Mas, há dados quantitativos ligados aos títulos. Na segunda coluna tem-se as quantidades de exemplares vendidos de cada título, ou seja, a frequência de vendas. Na terceira coluna apresenta-se o quanto cada frequência representa em relação ao total das vendas, expresso em percentagem. Assim, Guerra e Paz, que vendeu 120 unidades durante o mês, representa 26,7% dos livros vendidos no período (450 exemplares).

Pode-se, ainda, representar estes dados de maneira gráfica. Em primeiro por um gráfico de colunas, no qual a altura da coluna representa a frequência de vendas de cada título, em valores absolutos.

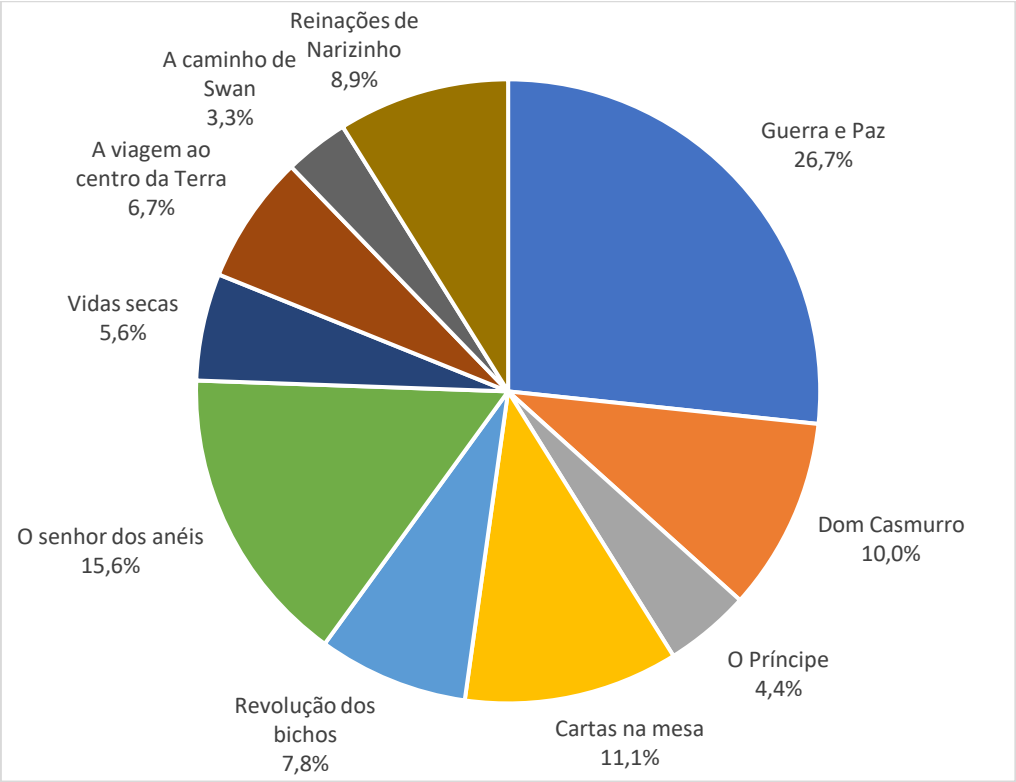
Figura 2.1 Gráfico de colunas



Fonte: o autor (utilizando Excel).

Outra forma é apresentar um gráfico de setores (vulgarmente chamado de gráfico de pizza) que mostra, a partir de setores de um círculo a proporção de frequência de cada dado em relação ao total. Para o exemplo, o gráfico de setores correspondente é apresentado na Figura 2.2:

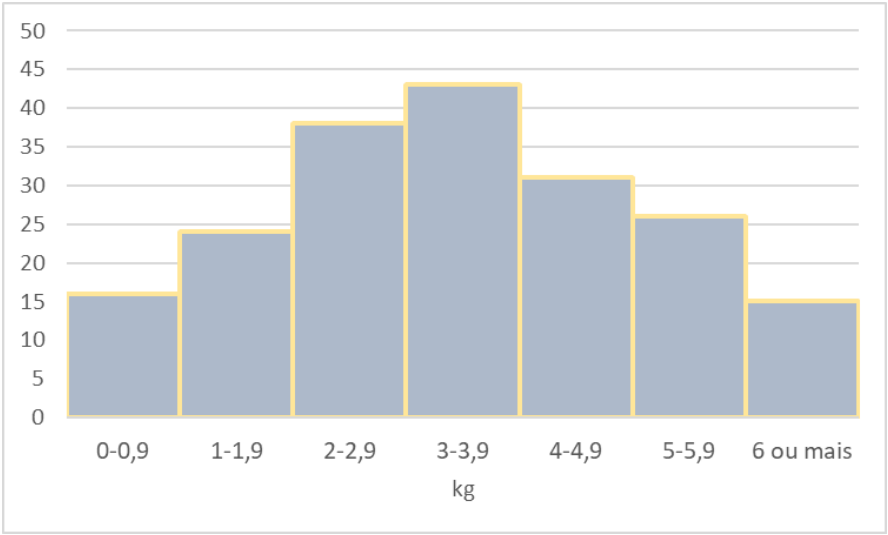
Figura 2.2: Gráfico de setores



Fonte: o autor (utilizando Excel)

Já o histograma é um gráfico de colunas que representa a frequência em que dados quantitativos aparecem em um determinado intervalo ou classe. Dessa forma, trata-se de variáveis contínuas, como, por exemplo, peso de caixas conforme a Figura 2.3:

Figura 2.3: Pesos de caixas em um depósito (em kg).



Fonte: o autor (utilizando Excel).

Percebe-se que no histograma não há espaço entre as colunas devido a natureza contínua dos dados, diferentemente do gráfico de colunas, no qual os dados são discretos.

Para construir um histograma, é necessário agrupar os dados em classes, conforme mostra a Figura 2.3. Virgillito (2006) apresenta o método da raiz para calcular a quantidade de classes. Este método define a quantidade de classes (k) como sendo a raiz quadrada do tamanho da amostra (n) (Expressão 2.1):

$$k = \sqrt{n} \qquad \text{Expressão 2.1}$$

Por exemplo, tome-se uma amostra com 100 elementos, cujo menor valor é 10 e o maior é 30. A quantidade de classes por esse método é:

$$k = \sqrt{100} = 10\ classes$$

Outro método apresentado por Virgillito (2006) é o método de Sturges:

$$k = 1 + 3{,}22 \log n \qquad \text{Expressão 2.2}$$

Para o exemplo apresentado, pelo método de Sturges é:

$$k = 1 + 3{,}22 \log 100 \cong 7\ classes$$

Com a definição da quantidade de classes, calcula-se a amplitude da classe (h), ou seja, a faixa de valores de cada classe. Tomando o resultado do método da raiz, tem-se:

$$h = \frac{Maior\ valor - Menor\ valor}{k} = \frac{30-10}{10} = 2 \qquad \text{Expressão 2.3}$$

Cada classe terá, então, o valor de 2. Assim:

1ª classe: $10 \vdash 12$
2ª classe: $12 \vdash 14$
3ª classe: $14 \vdash 16$
4ª classe: $16 \vdash 18$
5ª classe: $18 \vdash 20$
6ª classe: $20 \vdash 22$
7ª classe: $22 \vdash 24$
8ª classe: $24 \vdash 26$
9ª classe: $26 \vdash 28$
10ª classe: $28 - 30$

O símbolo "⊢" significa que o primeiro valor está incluso na classe, mas o segundo não. Por exemplo, na primeira classe, o valor 10 está incluso, mas o valor 12 não. Na décima classe, o símbolo "–" indica que ambos os valores estão inclusos.

Em grande parte das vezes, este conjunto de dados é uma parte de um todo. O todo é chamado de população, ou seja, o conjunto de todos os possíveis elementos de interesse de um estudo (concebíveis ou hipotéticos), enquanto que o conjunto de dados retirado desse todo (uma parte dos elementos) é chamado de amostra (ANDERSON; SWEENEY; WILLIAMS, 2002; FREUND; SIMON, 2000).

Esta amostra deve ter uma quantidade tal que possa representar a população por ter características semelhantes a esta. Dessa forma, é possível utilizar a amostra para fazer estimativas e testar hipóteses sobre as características da população. A este processo chama-se inferência estatística, que considera que a amostra tem comportamento semelhante à da população (ANDERSON; SWEENEY; WILLIAMS, 2002; FREUND; SIMON, 2000).

2.2 Medidas estatísticas

Há dois grupos de medidas estatísticas para um conjunto de dados: medidas de centro e medidas de dispersão. Medidas de centro são aquelas que buscam o ponto central dos dados. Como medidas de centro são apresentadas a média e a mediana. Medidas de dispersão são aquelas que descrevem como os dados estão distribuídos ao redor da medida de centro. Como medidas de dispersão, o desvio padrão e os quartis.

2.2.1 Média

Média é o valor correspondente à soma de todos os valores (x_i) da amostra ($\bar{x}$) ou da população (μ) dividida pela quantidade de dados (n) (Expressão 2.4):

$$\bar{x} \; ou \; \mu = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \text{Expressão 2.4}$$

Como exemplo, tem-se a Tabela 2.2:

Tabela 2.2: Conjunto de dados

Dado 1	Dado 2	Dado 3	Dado 4	Dado 5
1	3	4	5	6

Fonte: o autor

Assim, o cálculo da média é:

$$\bar{x} = \frac{1+3+4+5+6}{5} = 3{,}8$$

2.2.2 Mediana

Mediana é o elemento central ou médio de uma distribuição. Moore et al (2006) apresentam o algoritmo para a identificação da mediana:

1. Colocar os dados em ordem crescente;
2. Caso a quantidade de dados for ímpar, a mediana será o elemento central da sequência. Para saber a localização, conta-se $\frac{n+1}{2}$ observações a partir do início. Exemplo:

Retomando o exemplo de uma amostra da Tabela 2.2 (já em ordem crescente):

Dado 1	Dado 2	Dado 3	Dado 4	Dado 5
1	3	4	5	6

Como são 5 dados, tem-se:

$$Posição\ da\ mediana = \frac{n+1}{2} = \frac{5+1}{2} = 3$$

Dessa forma, o terceiro dado é a mediana. Assim, a mediana é igual a 4

3. Caso a quantidade de dados for par, a mediana será a média entre os dois elementos centrais. Exemplo:

   Supondo uma amostra com 6 dados, conforme a Tabela 2.3 (já em ordem crescente):

   Tabela 2.3: conjunto de dados

Dado 1	Dado 2	Dado 3	Dado 4	Dado 5	Dado 6
1	3	4	5	6	7

Fonte: o autor

Os dados centrais são os de posições 3 e 4, cujos valores são, respectivamente, 4 e 5. A mediana será:

$$Mediana = \frac{4+5}{2} = 4{,}5$$

Anderson, Sweeney e Williams (2002) indicam que quando há uma variação muito grande entre os menores e maiores valores de uma amostra, isto influência a média, mas tem pouca influência na mediana por indicar o elemento central, que depende de sua posição e não dos valores mínimos e máximos da amostra.

2.2.3 Amplitude

Amplitude (A), também chamada de intervalo total, é simplesmente a diferença entre o maior e o menor valor de uma amostra ou de uma população (ANDERSON; SWEENEY; WILLIAMS, 2002; FREUND; SIMON, 2000). Tendo como exemplo os dados da Tabela 2.3, o cálculo da amplitude é:

$$A = 7 - 1 = 6$$

Assim, o valor da amplitude é 6. Este conceito já foi utilizado nesse capítulo na construção de histogramas, para a definição da amplitude de classe.

2.2.3 Desvio padrão

Conforme Moore et al (2006), desvio padrão mede a dispersão considerando a distância que cada observação individual se encontra da média da distribuição. Matematicamente, o desvio padrão (s) é a raiz quadrada da variância (s^2). Variância é a média dos quadrados das diferenças (desvios) entre o valor de cada observação e a média da amostra. Como neste trabalho os dados são amostrais, os cálculos são efetuados considerando um elemento a menos da quantidade de dados (n-1), para absorver o erro assumido para fazer a inferência da população.

Assim, o cálculo da variância é expresso por (Expressão 2.5):

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$ Expressão 2.5

E o desvio padrão (Expressão 2.6):

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$ Expressão 2.6

Novamente, utilizando o exemplo com cinco elementos, calculando a variância e o desvio padrão:

a. Primeiro, calcula-se a diferença (desvio) entre o valor do dado e a média da amostra (Tabela 2.4).

Tabela 2.4: Desvios

	Valor	x-xi
Dado 1	1	-3
Dado 2	3	-1
Dado 3	4	0
Dado 4	5	1
Dado 5	7	3

Fonte: o autor

b. Em seguida, eleva-se ao quadrado os resultados encontrados (Tabela 2.5).

Tabela 2.5: cálculo dos quadrados dos desvios

	Valor	x-xi	$(x\text{-}xi)^2$
Dado 1	1	-3	9
Dado 2	3	-1	1
Dado 3	4	0	0
Dado 4	5	1	1
Dado 5	7	3	9

Fonte: o autor

c. Procede-se o cálculo da variância.

$$s^2 = \frac{9 + 1 + 0 + 1 + 9}{5 - 1} = 5$$

d. O desvio padrão é:

$$s = \sqrt{5} = 2{,}24$$

Quanto menor for o valor do desvio padrão, menor a dispersão dos dados em relação à média. Uma forma de avaliar esta dispersão é calcular o coeficiente de variação (Expressão 2.7), que expressa o desvio padrão em termos de percentagem em relação à média (FREUND; SIMON, 2000):

$$V = \frac{s}{\bar{x}}.100 \qquad \text{Expressão 2.7}$$

Para o nosso exemplo, o coeficiente de variação é:

$$V = \frac{2{,}24}{4}.100 = 56\%$$

2.2.4 Quartis

Os quartis estão relacionados à dispersão em relação à mediana. Os quartis são a divisão dos dados ordenados de forma crescente em quatro subconjuntos de mesma quantidade de dados. Ou seja, cada quartil representa 25% da amostra.

Conforme Moore et al. (2006), a sequência para definir os quartis é:

1. Dispor as observações em ordem crescente e identificar a mediana (que é o segundo quartil), conforme descrito no item 2.2.2. Com base no exemplo anterior, com 6 elementos:

Dado 1	Dado 2	Dado 3	Dado 4	Dado 5	Dado 6
1	3	4	5	6	7

A mediana, já calculada (4,5) é o segundo quartil (Q2).

2. O primeiro quartil é a mediana da metade à esquerda da mediana do conjunto de dados. Pelo exemplo:

Dado 1	Dado 2	Dado 3
1	3	4

Assim, o primeiro quartil (Q1) corresponde ao Dado 2 (3).

3. O terceiro quartil é a mediana da metade à direita da mediana do conjunto de dados. Portanto:

Dado 4	Dado 5	Dado 6
5	6	7

O terceiro quartil (Q3) é o Dado 5 (6).

2.3 Probabilidade

Probabilidade é a medida numérica da possibilidade de ocorrer um evento futuro ou de uma variável aleatória assumir um determinado valor. A probabilidade é calculada a partir do espaço amostral, ou seja, o conjunto de todos os resultados possíveis. O valor da probabilidade varia de 0 a 1, ou de 0 a 100%. Cálculo da probabilidade, conforme esse conceito, é a relação entre a frequência de um determinado dado (f) pelo tamanho (n) da amostra ou população (Expressão 2.8) (ANDERSON; SWEENEY; WILLIAMS, 2002; MOORE et al., 2006; FREUND; SIMON, 2000):

$$P = \frac{f}{n}.100 \qquad \text{Expressão 2.8}$$

Como exemplo, em uma amostra com 100 elementos de tempos de atendimentos de clientes, 15 clientes foram atendidos em 5 minutos. Dessa forma a probabilidade de, aleatoriamente, se escolher um cliente que foi atendido em 5 minutos:

$$P = \frac{15}{100}.100 = 15\%$$

A probabilidade, portanto, é de 15%.

As variáveis aleatórias podem ser discretas ou contínuas. Variáveis aleatórias discretas são aquelas que podem assumir tanto um número finito quanto infinito de valores dentre uma sequência (ANDERSON; SWEENEY; WILLIAMS, 2002). Estes valores são números inteiros. Por exemplo, a quantidade de caminhões que chegou a um centro de distribuição em um determinado dia, que foi de 5 caminhões.

As variáveis aleatórias contínuas podem assumir qualquer valor dentro de um intervalo (ANDERSON; SWEENEY; WILLIAMS, 2002). Por exemplo, o volume de água em copos descartáveis pode apresentar o volume de 249,2 ml, ou seja, valores fracionados ou decimais são aceitos.

No item 2.1 foi apresentado o conceito da distribuição de frequências, que é o registro das quantidades de ocorrência dos diversos valores da variável aleatória. A partir da distribuição de frequência, tem-se a distribuição de

probabilidade que é a função que descreve o comportamento da probabilidade de uma variável aleatória assumir determinado valor (ANDERSON; SWEENEY; WILLIAMS, 2002; MOORE et al., 2006). Por exemplo, a distribuição normal.

No próximo tópico são apresentados alguns tipos funções de distribuição de probabilidade que são mais usuais no processo de simulação, na construção de modelos probabilísticos que permitam inferir as propriedades de um fenômeno aleatório. Por exemplo, qual o intervalo médio e desvio padrão do intervalo de tempo entre as chegadas de clientes em uma agência bancária.

2.2 Distribuições contínuas e discretas

A distribuição das variáveis pode ser aproximada por um modelo de distribuição de probabilidades conhecido (normal, beta, Erlang, Poisson, etc.) por meio de um teste de aderência (visto em um capítulo mais adiante).

- Normal
- Beta
- Uniforme
- Triangular
- Exponencial
- Erlang
- Gamma
- LogNormal
- Weibull
- Poisson

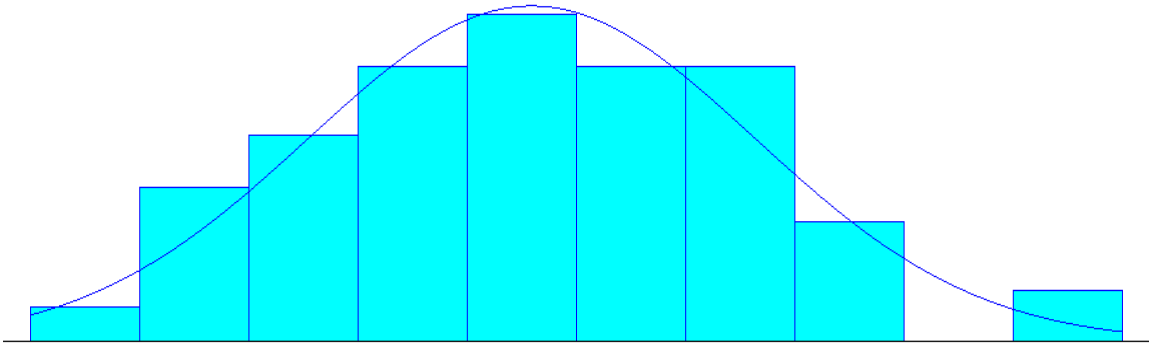
Em seguida, são apresentadas as características dessas distribuições de probabilidade.

2.2.1 Distribuição normal (Galton, 1889)

Também chamada de Gaussiana, é a distribuição contínua que descreve fenômenos regidos por variáveis aleatórias que possuem variação simétrica acima e abaixo da média (assimetria=0). Apresenta o formato de sino, conforme mostrado na Expressão 2.9 e na Figura 2.4.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad \text{Expressão 2.9}$$

Figura 2.4: Distribuição normal (μ=47,3; σ=18,4)



Fonte: o autor, utilizando o Input Analyser do ARENA®.

Os parâmetros dessa distribuição são a média, representada por μ (Expressão 2.10) e desvio padrão, representado por σ (Expressão 2.11).

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \text{Expressão 2.10}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n}} \qquad \text{Expressão 2.11}$$

Os exemplos típicos de aplicação dessa distribuição são tempos de processo, tempos de máquinas, medidas de peças, volume de recipientes, etc.

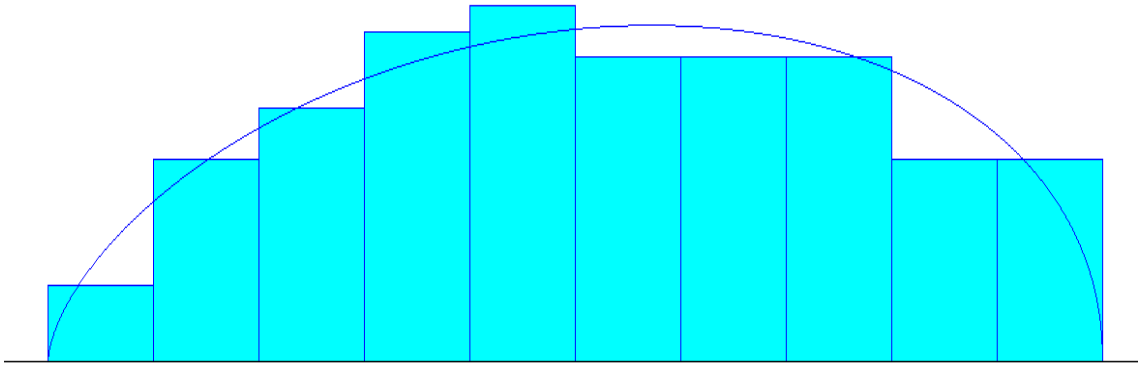
2.2.2 Distribuição Beta

Criada por Gini, em 1911, a distribuição beta é a mais flexível, capaz de assumir diversas formas, sendo utilizada para fazer aproximação quando houver ausência de dados. Seus parâmetros são coeficiente de forma (α_1) e coeficiente de escala (α_2), conforme apresentado na Expressão 2.12 (HARREL et al., 2002).

$$\boldsymbol{f(x) = \frac{x^{\alpha_1 - 1} . (1-x)^{\alpha_2 - 1}}{B(\alpha_1, \alpha_2)}} \qquad \text{Expressão 2.12}$$

O formato ARENA para essa distribuição de probabilidade é BETA(α_1, α_2).

Figura 2.5: Distribuição beta (1.62,1.46)



Fonte: o autor, utilizando o Input Analyser do ARENA®.

2.2.3 Distribuição uniforme

É uma distribuição contínua, criada por Uspensky em 1937. O formato no ARENA é UNIF(a, b). Cada valor entre um valor mínimo (a) e um máximo (b) especificados têm igual probabilidade de ocorrer (Figura 3). Sua representação gráfica, portanto, é uma linha reta horizontal. A função de densidade de probabilidade (Expressão 2.13) e a função de distribuição (Expressão 2.14) são apresentadas a seguir:

$$f(x) = \frac{1}{b-a} \qquad \text{Expressão 2.13}$$

$$F(x) = \frac{(x-a)}{b-a} \qquad \text{Expressão 2.14}$$

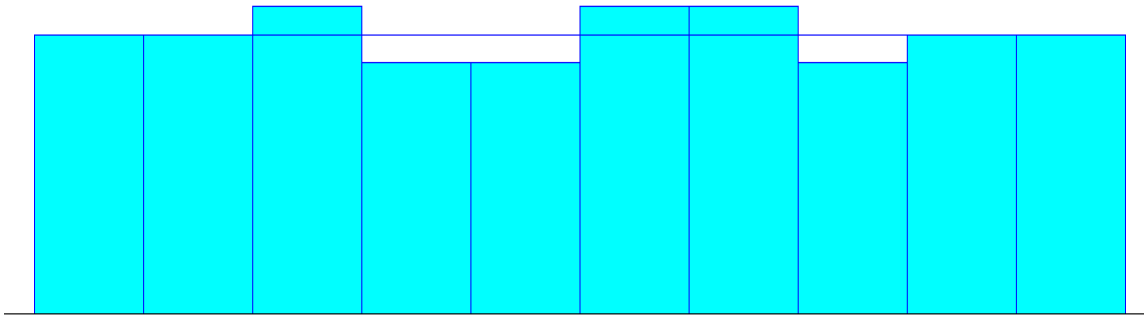
A média (μ) e a variância (σ^2) são calculadas por (Expressões 2.15 e 2.16) (ANDERSON; SWEENEY; WILLIAMS, 2002; MOORE et al., 2006):

$$\mu = \frac{a+b}{2} \qquad \text{Expressão 2.15}$$

$$\sigma^2 = \frac{(b-a)^2}{12} \qquad \text{Expressão 2.16}$$

A forma da curva é apresenta na Figura 2.6.

Figura 2.6: Distribuição uniforme (a=20; b=99)



Fonte: o autor utilizando o Input Analyser do ARENA®.

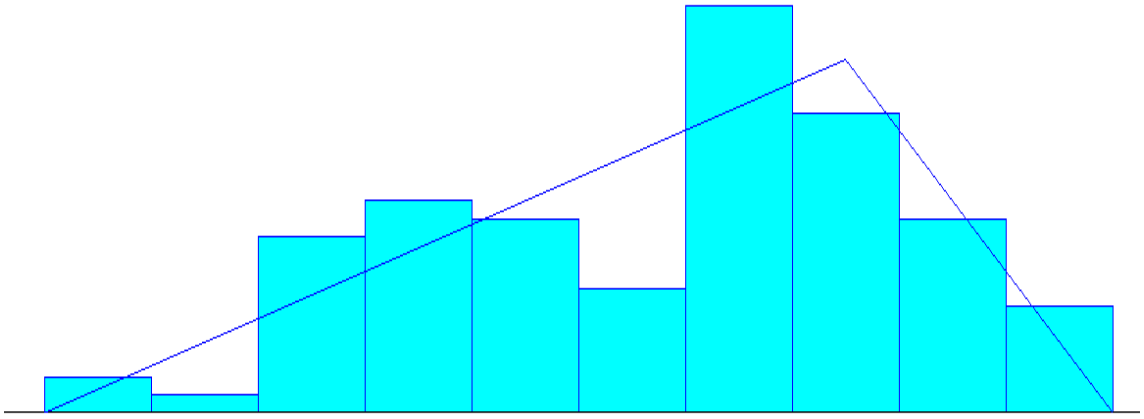
A distribuição uniforme é utilizada quando há pouco conhecimento sobre o comportamento da variável aleatória tratada. Para simular essa distribuição é necessário ser capaz de gerar um conjunto um conjunto de números que sejam equiprováveis, independentes e reproduzíveis. Os dois primeiros critérios são requisitos teóricos e o último é prático no que é importante ser capaz de replicar resultados como uma verificação no programa de simulação (MOONEY, 1997).

2.2.4 Distribuição triangular

É uma distribuição útil para uma primeira aproximação quando há falta de dados mais claros. Apresenta, como indica o nome, uma fora triangular (Figura 2.7), e os parâmetros dessa distribuição são o valor mínimo (a), o valor máximo (b) e o valor mais provável (c, a moda) (HARREL et al., 2002). As expressões (Expressão 2.16) para o cálculo das probabilidades dependem se o valor procurado (x) está abaixo da moda ou acima dela.

$$f(x) = \begin{cases} \frac{2(x-a)}{(b-a)(c-a)} \; para \; a \leq x \leq c \\ \frac{2(b-x)}{(b-a)(b-c)} \; para \; c \leq x \leq b \end{cases}$$ Expressão 2.16

Figura 2.7: Distribuição triangular (a=38, c=47, b=50).



Fonte: o autor utilizando o Input Analyser do ARENA®.

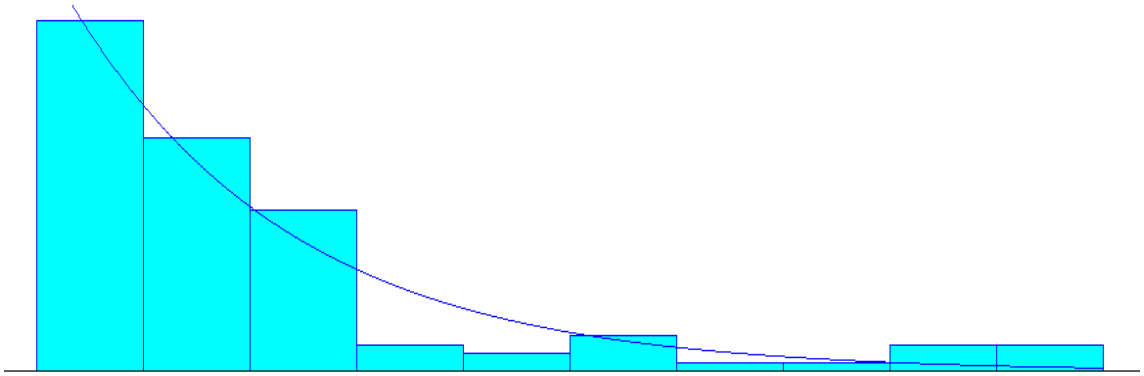
2.2.5 Distribuição exponencial

É uma família de distribuições de viés à direita e possui grande variabilidade**.** O seu parâmetro é o inverso da média dos dados ($\lambda = {}^{1}/_{\mu}$). A Expressão 2.7 é a função de densidade da distribuição exponencial:

$$f(x) = \lambda . e^{-\lambda . x} \qquad \text{Expressão 2.17}$$

A distribuição exponencial (Figura 2.8) é muito utilizada para a distribuição dos períodos entre dois eventos, como o intervalo de tempo entre duas chegadas consecutivas de clientes em um estabelecimento.

Figura 2.8: Distribuição exponencial (λ=6,27).



Fonte: o autor utilizando o Input Analyser do ARENA®.

.

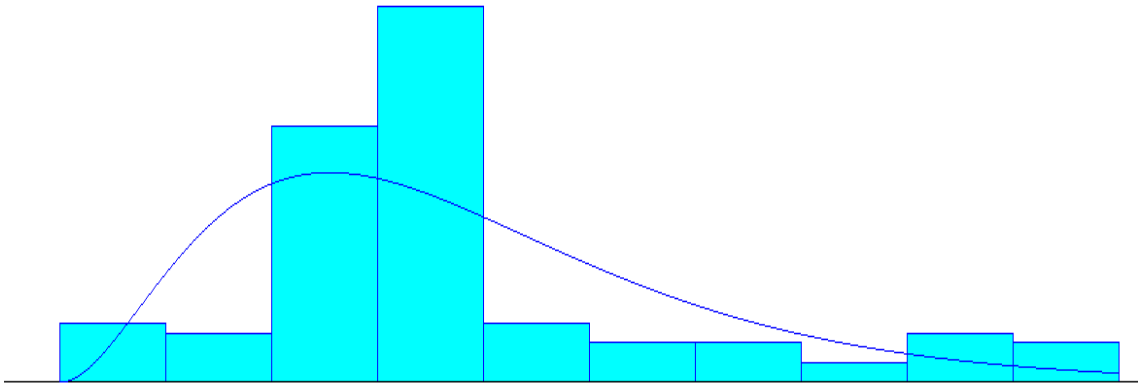
2.2.6 Distribuição Erlang

Criada por Erlang em seus estudos sobre filas (vide Capítulo 3: Teoria das Filas), é utilizada em situações em que a entidade entra em uma estação para ser servida por uma série de postos (k). A Expressão 2.19 representa essa distribuição:

$$f(x) = \frac{(\lambda.k)^k}{(k-1)!} x^{k-1}.e^{-k\lambda x} \qquad \text{Expressão 2.19}$$

Seus parâmetros são a média exponencial (λ) e quantidade de postos (k, coeficiente de forma inteiro) e sua forma é apresenta na Figura 2.10.

Figura 2.10: distribuição Erlang (λ=9,94; k=3)



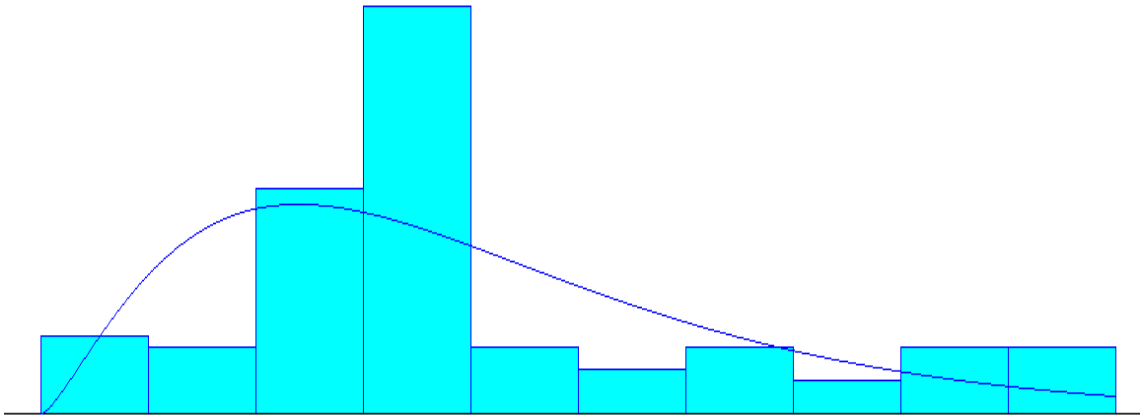
Fonte: o autor utilizando o Input Analyser do ARENA®.

2.2.7 Distribuição Gamma

Desenvolvida por Weatherburn, em 1946, é utilizada para representar o tempo para completar uma tarefa (tempo de reparo, por exemplo). Seus parâmetros são a média (λ) e o coeficiente de forma (α, não inteiro). O formato da curva varia de acordo com o valor de α, diferentemente das distribuições normal e exponencial (Expressão 2.20) (HARREL et al., 2002; FISHMAN, 1995).

$$f(x) = \frac{x^{(\alpha-1)} e^{-x/\lambda}}{\lambda^{\alpha} \Gamma(\alpha)}$$ Expressão 2.20

Figura 2.11: Distribuição Gamma (λ=13,5; α=2,4)



Fonte: o autor utilizando o Input Analyser do ARENA®.

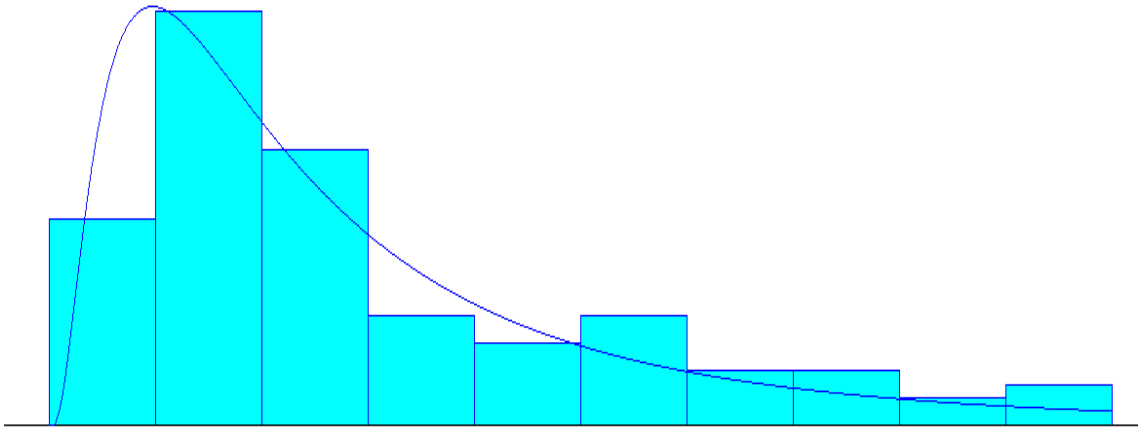
2.2.8 Distribuição Lognormal

Desenvolvida por Gaddum, em 1945, é utilizada para representar tempos de atividades com distribuição não simétrica (assimétrica). Na expressão 2.7, apresenta-se a diferença com a distribuição normal, por se utilizar o logaritmo natural de x (Expressão 2.21) (MOONEY, 1997, HARREL et al., 2002).

$$f(x) = \frac{1}{2\pi\sigma^2} . e^{-\frac{(\ln x - \mu)^2}{2.\sigma^2}}$$ Expressão 2.21

Seus parâmetros também são, como na distribuição normal, a média μ e desvio padrão σ. Sua curva característica é mostrada na Figura 2.12:

Figura 2.12: Distribuição lognormal (μ= 11,2; σ=12,5).



Fonte: o autor utilizando o Input Analyser do ARENA®.

.

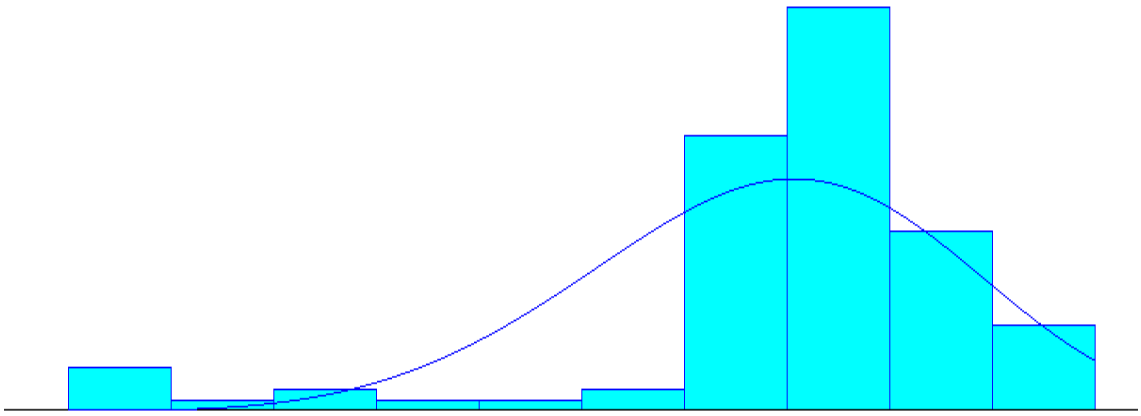
2.2.9 Distribuição Weibull

Desenvolvida por Weibull, em 1939, representa o tempo de vida de equipamentos, sendo utilizada em cálculos de expectativa de vida (MOONEY, 1997; HARREL et al., 2002).

Seus parâmetros são coeficiente de forma (α) e coeficiente de escala (β),conforme a Expressão 2.22.

$$f(x) = (\alpha.\beta^{-\alpha}.x^{\alpha-1})e^{-\left(\frac{x}{\beta}\right)^{\alpha}} \qquad \text{Expressão 2.22}$$

Figura 2.13: Distribuição de Weibull (α=48,1; β=4,34).



Fonte: o autor utilizando o Input Analyser do ARENA®.

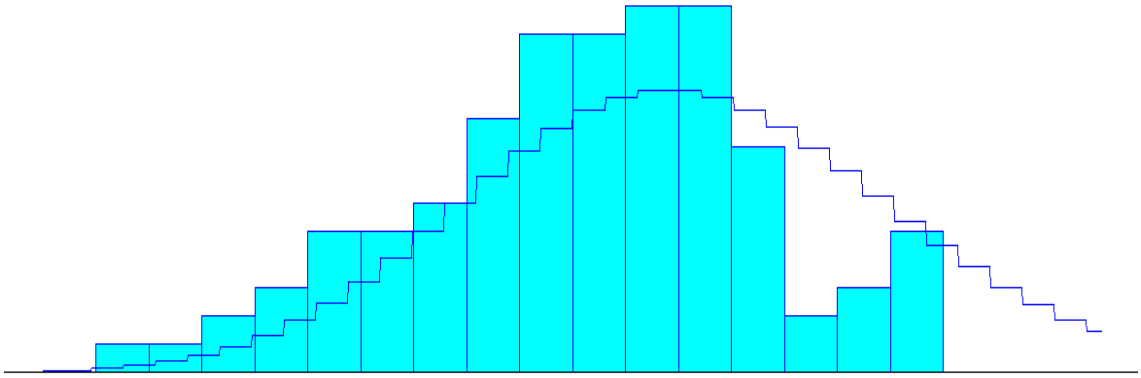
2.2.10 Distribuição de Poisson

É uma distribuição discreta, desenvolvida por Soper em 1914, voltada a resultados de eventos dicotômicos ocorridos em um dado período de tempo. O parâmetro dessa distribuição é a média (λ), que expressa a taxa de sucesso, ou seja, a quantidade de eventos ocorridos (MOONEY, 1997, HARREL et al., 2002).

A Expressão 2.23 representa a função de distribuição de probabilidades de Poisson e a Figura 2.14 a sua forma:

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!} \qquad \text{Expressão 2.23}$$

Figura 2.14: distribuição de Poisson (λ=43).



Fonte: o autor utilizando o Input Analyser do ARENA®.

2.3 Limpeza de dados

A limpeza de dados é uma técnica útil para se eliminar dados que estejam muito discrepantes em relação ao restante da amostra, pois podem levar a resultados distorcidos. Estes pontos são chamados de pontos fora da curva ou *outliers*. Uma forma de se identificar estes pontos é a regra 1,5xIIQ (MOORE et al, 2006).

Esta regra baseia-se no intervalo interquartil (IIQ). Quartil já foi definido no item 2.2.4. A ideia da regra é calcular os limites inferior (abaixo do qual se despreza os dados) e superior (acima do qual também se despreza os dados). As Expressões 2.24 e 2.25 para os limites são:

$$LI = Q1 - 1{,}5.(Q3 - Q1) \qquad \text{Expressão 2.24}$$

$$LS = Q3 + 1{,}5.(Q3 - Q1)$$ Expressão 2.25

O limite inferior é representado por LI e o limite superior por LS. Q1 e Q3 são, respectivamente, o primeiro e o terceiro quartis, conforme já visto anteriormente. A diferença (Q3-Q1) é o IIQ.

Supondo, como exemplo, os pesos (em kg) de caixas movimentadas de um dia em uma agência do correio, chegou-se a seguinte amostra com 150 elementos, apresentada na Tabela 2.6:

Tabela 2.6: Pesos de caixas (kg)

10,00	11,00	7,00	12,97	8,18	15,81	7,87	18,81	7,76	21,80
8,33	2,40	10,36	2,00	13,09	3,50	12,05	2,01	11,00	2,00
13,73	4,75	10,73	1,86	13,61	4,45	10,71	1,57	13,71	4,30
10,96	1,30	13,86	4,18	11,27	1,29	14,15	4,29	11,41	1,55
14,41	4,44	11,52	1,84	14,41	4,73	8,00	7,70	10,96	4,70
7,96	7,68	10,92	4,69	7,92	7,68	10,88	4,71	7,88	7,70
10,85	4,74	7,85	7,74	10,83	4,78	7,83	7,78	10,83	4,82
7,84	7,82	10,84	4,86	7,87	7,86	10,87	2,70	12,16	2,00
14,88	4,20	12,27	3,32	11,74	1,12	14,44	3,98	12,21	2,94
12,82	3,69	11,25	2,40	13,27	2,00	16,00	2,00	13,00	2,30
15,24	3,66	13,75	6,44	14,21	9,43	14,19	12,42	13,77	15,22
15,16	16,78	12,53	16,68	10,05	14,91	12,20	13,83	15,05	15,66
15,20	17,11	12,25	16,17	8,00	19,13	3,00	19,56	6,00	18,72
2,00	21,45	2,00	24,18	29,00	22,18	28,42	21,74	5,00	18,87
3,00	19,29	7,00	21,26	9,00	22,49	7,56	25,36	8,24	28,15

Fonte: o autor

Procedendo conforme o algoritmo já apresentado, a mediana desta amostra, ou segundo quartil é:

Q2= 10,535 kg

Dessa forma, identificando as medianas da metade abaixo (Q1) e acima (Q3) de Q2, tem-se:

Q1= 4,7375 kg e Q3=13,9325 kg

Calculando-se os limites:

$$LI = 4{,}7375 - 1{,}5.(13{,}9325 - 4{,}7325) = -9{,}055\ kg$$

$$LS = 13{,}9325 + 1{,}5.(13{,}9325 - 4{,}7325) = 27{,}725\ kg$$

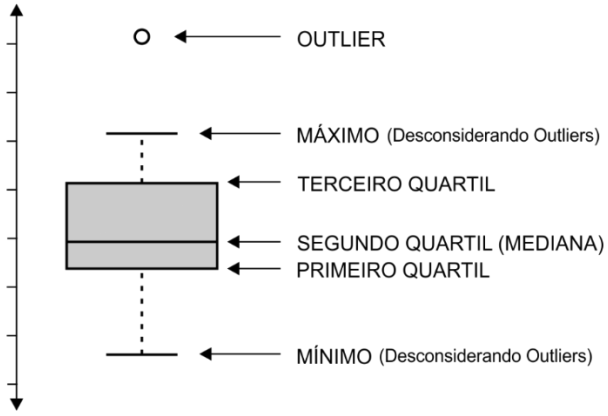
O resultado do limite inferior (LI) é negativo. Como não faz sentido o peso ser negativo, considera-se que o valor de LI é zero (0). Assim, não há nenhum outlier abaixo de LI.

Por outro lado, o limite superior (LS) é de 27,725 kg e há dois valores acima do LS: 28,15 kg e 29 kg. Estes dois dados serão desconsiderados nas análises estatísticas, pois distorceriam os resultados.

2.3.1 Representação gráfica de outliers

Uma forma de representar os outliers, calculados no item anterior, é o diagrama de caixa ou *boxplot*, apresentado na Figura 2.15.

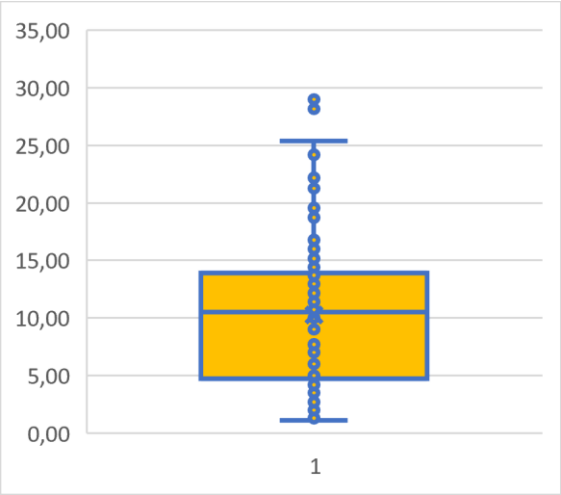
Figura 2.15: Diagrama de caixa (boxplot).



Fonte: ABG Consultoria (2018)

Retomando o exemplo do item anterior, por meio do Excel, constrói-se o *boxplot*, conforme a Figura 2.16:

Figura 2.16: boxplot do exemplo



Fonte: o autor utilizando o Excel.

Percebe-se na Figura os dois pontos acima do LS, correspondentes aos valores 28,15 kg e 29 kg.

2.4 Teorema central do limite

Conforme Fernandes (2005), o teorema central do limite (ou teorema do limite central) estabelece que a função de distribuição acumulada de uma soma de uma variável aleatória independente é aproximadamente idêntica à função de distribuição acumulada de uma variável aleatória gaussiana (normal).

Pela definição apresentada por Moore et al (2006): "se uma população tiver distribuição N(μ, σ) então a média amostral $\bar{x}$ de n observações independentes terá distribuição N(μ, $\sigma/\sqrt{n}$)". Anderson, Sweeney e Williams (2002) ressaltam que a distribuição da média amostral $\bar{x}$ se aproxima da distribuição normal quanto maior for o tamanho da amostra.

2.5 Intervalo de confiança

Na maioria das vezes, não se tem acesso a todos os dados de uma população. Devido a isto, utiliza-se uma amostra significativa, ou seja, que represente adequadamente a população. Tomando como exemplo o peso de caixas, retira-se aleatoriamente uma certa quantidade de caixas do total de caixas prontas para serem expedidas. Assumindo que foram retiradas 100 caixas do total de caixas como amostra. Estas 100 caixas foram pesadas e os resultados registrados. Neste exemplo, o peso médio ($\bar{x}$) foi de 20 kg. Mas este peso médio refere-se às caixas da amostra e não das caixas da população. Para se fazer a inferência para a população temos que calcular o intervalo de confiança.

Conforme Moore et al (2006), o intervalo de confiança é composto de dois componentes: a média e o erro, conforme a expressão:

$$\bar{x} \pm z\frac{s}{\sqrt{n}} \qquad \text{Expressão 2.26}$$

Onde:

z: coeficiente da distribuição normal;

s: desvio padrão da amostra

n: Tamanho da amostra

O coeficiente z está ligado à confiança que se tem no resultado. Por exemplo, se há a probabilidade de 5% de que a média da população esteja fora do intervalo de confiança, então a confiança é de 95%. Na Tabela 2.7 são apresentados alguns valores de z em função da confiança desejada:

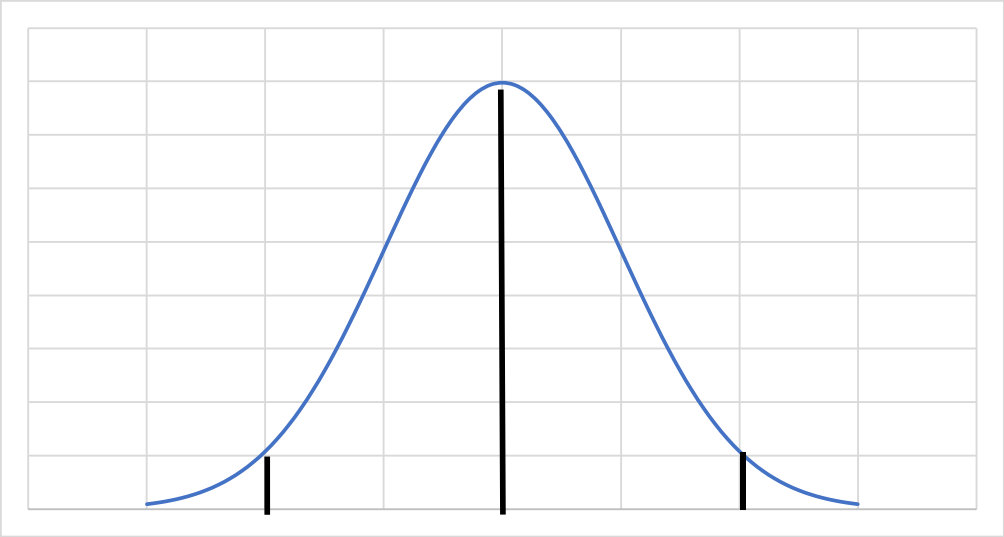
Tabela 2.7: Coeficiente z

Confiança (%)	90	95	99
z	1,645	1,960	2,576

Fonte: o autor

Na Figura 2.17, tem-se a representação do intervalo de confiança em um gráfico de distribuição normal:

Figura 2.17: intervalo de confiança



Fonte: o autor

O erro é subtraído da média para identificar o limite inferior do intervalo de confiança e somado para se identificar o limite superior. O cálculo do desvio padrão s já foi visto anteriormente. E o tamanho de amostra n refere-se à quantidade de elementos da amostra.

Voltando ao exemplo no qual o peso médio da amostra de 100 caixas é 20 kg, considerando desvio padrão de 1,2 kg e confiança de 95%, tem-se:

$$\bar{x} \pm z\frac{s}{\sqrt{n}} = 20 \pm 1{,}96\frac{1{,}2}{\sqrt{100}} = 20 \pm 0{,}24\ kg$$

Desta forma, o erro cometido é de 0,24 kg, ou seja, a média da população, com 95% de confiança, está entre 19,76 kg e 20,24 kg.

2.6 Exercício

Foi retirada uma amostra de 200 clientes de um restaurante fast food e tabelou-se a idade deles, conforme a Tabela 2.7:

Tabela 2.7: amostra de idade de clientes.

39	40	35	36	33	32	30	32	39	40
35	32	37	32	39	33	40	40	32	37
35	34	40	34	38	34	40	38	34	38
32	35	39	33	30	34	30	30	34	31
39	37	31	36	35	36	31	31	30	38
40	30	32	33	39	40	40	34	36	40
35	32	33	35	33	40	39	40	30	34
36	31	34	39	40	32	30	36	31	31
36	37	38	40	39	32	32	34	34	31
36	35	35	34	40	31	32	39	32	31
36	33	35	38	33	37	30	34	40	34
36	36	37	33	39	36	30	36	33	36
30	30	32	30	38	35	34	40	33	33
31	35	35	32	40	38	32	30	31	40
32	50	36	30	31	36	35	34	36	38
37	38	40	40	39	35	38	38	34	35
35	31	32	37	30	40	38	40	39	32
36	33	38	37	38	38	37	32	40	31
32	38	32	38	35	35	37	37	36	37
39	40	31	37	37	35	39	36	35	35

Para essa amostra:

a. Construir histogramas pelo método da raiz e pelo método de Sturges;
b. Calcular a média, a variância e o desvio padrão;
c. Identificar a mediana, a amplitude e os quartis;
d. Calcular o intervalo de confiança da média com 95% de confiança;
e. Calcular a probabilidade de se escolher um cliente, aleatoriamente, com 33 anos;
f. Verificar se há algum elemento fora da curva (outlier).

Referências bibliográficas

ABG CONSULTORIA. **Boxplot – Como interpretar?** Disponível em: http://www.abgconsultoria.com.br/blog/boxplot-como-interpretar/ . Acessado em: 22/01/2019. 2018.

ANDERSON, D. R.; SWEENEY, D. J.; WILLIAMS, T. A. **Estatística aplicada à administração e economia. Pioneira Thomson Learning.** São Paulo. 2002

FISHMAN, G. S. **Monte Carlo:** concepts, algorithms and applications. Springer. New York. 1995.

FERNANDES, C. A. B. A. **Gerenciamento de riscos em projetos:** como usar o Microsoft Excel para realizar a simulação Monte Carlo. Disponível em: 2005

FREUND, J. E.; SIMMON, G. A. **Estatística aplicada**: economia, administração e contabilidade. 9. ed. Bookman. Porto Alegre. 2000.

HARREL, C. R., et al. **Simulação:** otimizando sistemas. 2. ed. IMAM. São Paulo. 2002.

MOONEY, C. Z. **Monte Carlo Simulation**. Sage Publications. Thousand Oaks. California.USA. 1997.

MOORE, D. S.; MCCABE, G. P.; DUCKWORTH, W. M.; SCLOVE, S. L. **Estatística empresarial:** como usar dados para tomar decisões. LTC. Rio de Janeiro. 2006.

VIRGILLITO, S. B. **Estatística aplicada**. 3. ed. Edicon. São Paulo. 2006.