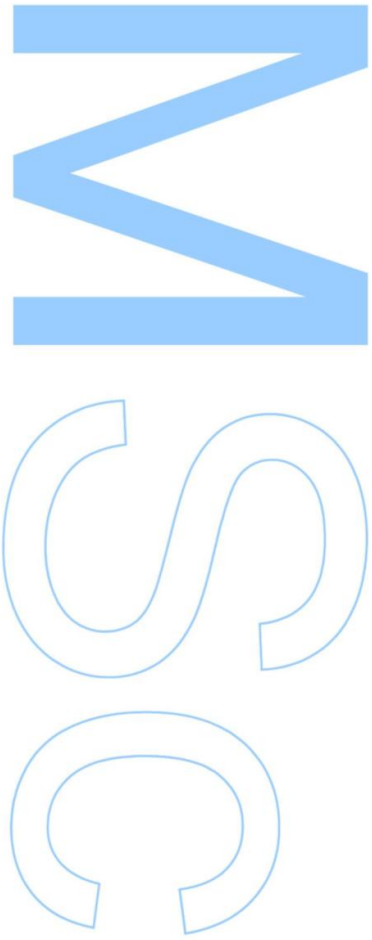
Excel e R como ferramentas no ensino de Estatística

Kumbo João
Mestrado em Matemática para Professores
Departamento de Matemática
Ano 2017

Orientador
Óscar António Louro Felgueiras, Professor Auxiliar
Faculdade de Ciências da Universidade do Porto



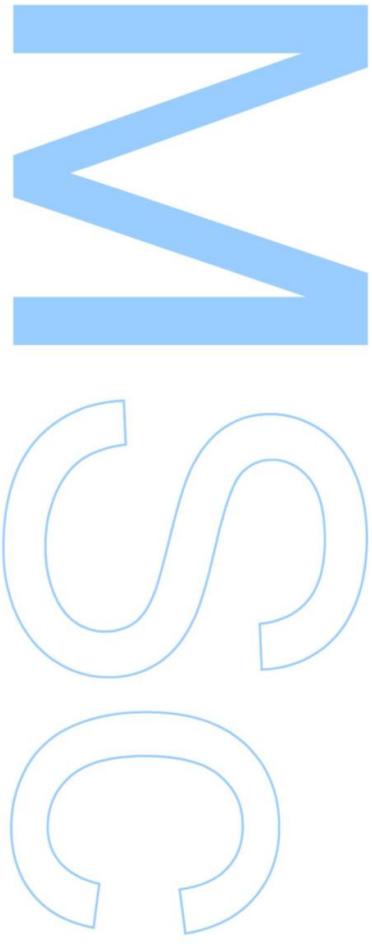
U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, _____/_____/_____



Dedicatória

Dedico este trabalho aos meus pais João Ndonga e Maria Lengula, pelo seu amor multiplicado por 100, apoio sem limite e pelas orações.

À minha querida esposa Maria Regina Capitão João Ndonga, pelo seu amor, carinho, confiança, lealdade e responsabilidade. Ao meu querido filho Cléo, que muitas vezes sentiu o seu desejo tão forte de estar comigo interrompido por causa desta formação.

Agradecimentos

Os meus agradecimentos vão:

- A Deus pai todo poderoso, pela saúde mental e disposição física concedida durante a formação;
- Ao Diretor Geral do ISCED/Uíge Dr. Kimpolo Nzau, por ter acreditado em mim concedendo-me a autorização para esta formação;
- Ao meu irmão Diassonama João, por ter me alojado na sua casa durante o periodo da formação;
- Ao meu orientador Dr. Óscar António Louro Fergueiras, por ter aceitado em trabalhar comigo, mesmo não conhecendo as minhas limitações e potencialidades, agradeço ainda pela sua dedicação, paciência e sobretudo a forma sábia como foi orientado o estudo;
- Ao Diretor do Curso de Mestrado em Matemática para Professores, Dr. Paulo Maurício, pela sua disponibilidade e simplicidade em ajudar sempre que é solicitado;
- A todos professores deste Mestrado, pela preparação concedida para a vida profissional;
- Aos meus amigos José C. Chibaca, Kananito Esteves e a todos colegas deste Mestrado, particularmente, o Victor, a Claúdia, a Dulce, a Cristina e o Zé que sempre colaboraram comigo nesta empreitada;
- A toda a minha família e a todos que direta ou indiretamente contribuiram para que esta formação fosse um facto.

Resumo

Este estudo, tem como finalidade mostrar por meio de exemplos as diversas potencialidades que fazem do Excel e R ferramentas adequadas para apoiar o trabalho do professor no ensino de Estatística. A não utilização de ferramentas computacionais no ensino de Estatística no ISCED/Uíge e as preocupações dos alunos, na aprendizagem desta disciplina, motivaram este estudo. O mesmo privilegia uma pesquisa teórica ou simplesmente bibliográfica, cobrindo os tópicos de Estatística que estão no Manual de Matemática do 11º ano do Ensino Secundário de Angola e no programa de Estatística que é ensinado no ISCED/Uíge.

Em termos formais, o estudo encontra-se dividido em cinco capítulos. O primeiro é introdutório, abordando a motivação e objetivos do estudo. O segundo refere-se a tabelas de distribuições de frequências e gráficos. O terceiro é dedicado a vários tipos de medidas de estatística descritiva. O quarto refere-se a distribuições de probabilidades e testes de hipóteses. Finalmente o quinto, trata de conclusões finais. Porém, em cada um dos capítulos foi feita uma introdução de forma a clarificar os conceitos estatísticos a estudar e posteriormente os procedimentos informáticos necessários à obtenção dos resultados esperados. São resolvidos 30 exemplos e em cada um deles explicamos de forma detalhada os procedimentos de resolução e especificando as ferramentas de análise estatística utilizadas. Esta integração torna a aprendizagem da Estatística eficaz e mostra que há várias alternativas metodológicas para alcançar o mesmo resultado.

Palavras-chaves: Ensino de Estatística, ISCED/Uíge, Excel e R.

Abstract

This study aims to show through the use of examples the various potentialities that make Excel and R adequate tools to support the work of the teacher in the teaching of Statistics. The non-use of computational tools in the teaching of Statistics in ISCED/Uíge and the students' concerns in the learning of this discipline heve motivated this study. The same privileges a theoretical or simply bibliographic research, covering the topics of Statistics that are in the Manual of Mathematics of the eleventh year of Secondary Education of Angola and in the program of Statistics that is taught in the ISCED/Uíge.

In formal terms, the study is divided into five chapters. The first is introductory, addressing the motivation and objectives of the study. The second refers to tables of frequency distributions and graphs. The third is devoted to various types of descriptive statistics measures. The fourth refers to probability distributions and hypothesis tests. Finally, the fifth, deals with final conclusions. However, in each of the chapters an introduction was made in order to clarify the statistical concepts to be studied and later the computer procedures necessary to obtain the expected results. 30 examples are solved and in each of them we explain in detail the resolution procedures and specifying the statistical analysis tools used. This integration makes the learning of statistics effective and shows that there are several methodological alternatives to achieve the same result.

Keywords: Statistics Teaching, ISCED/Uige, Excel and R.

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Cada vez mais, o ensino de Estatística vem sendo o foco de professores universitários e investigadores desta área no que se refere a quais seriam as melhores metodologias a serem utilizadas na sala de aula de modo a desenvolver nos alunos a capacidade e compreensão de análise de dados, a resolução de problemas e casos de estudo. É nesta perspetiva que pretendemos contribuir com este estudo, cobrindo os tópicos de Estatística que estão no Manual de Matemática do 11º ano do Ensino Secundário de Angola e no programa de Estatística que é ensinado no ISCED/Uíge.

1.1 Motivação do estudo

Hoje em dia, praticamente todas as esferas da atividade humana estão influenciadas pela Estatística. Todos os dias, os jornais, a televisão, as revistas apresentam-nos estimativas, previsões, gráficos, tabelas, sondagens e resultados de inquéritos, aproximando assim a Estatística à vida quotidiana. Portanto, torna-se imprescindível o seu estudo em todos níveis do ensino.

No Instituto Superior de Ciências da Educação ISCED/Uíge, a disciplina de Estatística é semestral e enquadra-se no 2º ano de graduação em todos os cursos. Através de uma experiência vivida como Assistente Estagiário na mesma instituição, constatei que nos cursos de Ensino de Ciências Sociais e Pedagógicas, esta disciplina constitui uma preocupação por parte dos alunos, uma vez que muitos deles provêm de cursos de Ciências Sociais e Humanas no Ensino Secundário; para estes a Matematica não é aprofundada, pois deixam de entrar em contacto com ela muito cedo. E quando se deparam com a Estatística nos níveis subsequentes a preocupação torna-se maior.

Além disso, o ensino formal dos conteúdos desta disciplina muitas das vezes limita-se à reprodução de fórmulas matemáticas, as construções gráficas são feitas manualmente não fazendo recurso a ferramentas computacionais que bem deveriam contribuir na forma rápida e eficiente de resolver vários problemas morosos e complexos. É precisamente para esta componente da formação que o presente estudo visa contribuir, propondo o Excel e R, como ferramentas no ensino de Estatística.

De acordo com FLORENTIN(2003, p.163), parte importante do conhecimento profissional dos professores diz respeito ao uso das TIC's como ferramentas cada vez mais presentes na atividade dos professores de matemática, constituindo: um meio educacional auxiliar para apoiar a aprendizagem dos alunos; um instrumento de produtividade pessoal, para preparar material para as aulas, para realizar tarefas administrativas e para procurar informações e materiais; um meio interativo para interagir e colaborar com outros professores e parceiros educacionais.

Com base nos aporte teóricos apresentados acima, importa realçar que atualmente há muitas aplicações informáticas para Estatística. A opção por Excel e R reside no seguinte.

O Excel embora não seja exclusivamente um software para fins de operações de estatística, possui funções que permitem realizar cálculos de medidas de estatística descritiva, as distribuições de probabilidades, as principais representações gráficas e mediante recurso a outras funções predefinidas permite ainda efetuar procedimentos não imediatos como selecionar aleatoriamente uma amostra, construir histogramas com classes de diferente amplitude, organizar os dados em tabelas de contingência, estimar inferências, etc. Por outro lado, o Excel tem ainda a seu favor pelo facto de estar instalado em quase todos os computadores e de ser extensamente utilizado.

O R é um conjunto integrado de ferramentas computacionais que permite a manipulação e análise de dados, o cálculo numérico e a produção de gráficos, PIAIRO e PEREIRA(2012, p.13). Este programa utiliza uma linguagem de programção simples que é interpretada, na medida em que os comandos são imediatamente executados. Uma das suas principais carateristicas é o seu caráter gratuito e a sua disponibilidade para uma gama bastente variada de sistemas operativos, disponível em: `(https://www.r-project.org/)`.

O R trouxe novas formas de explorar a Estatística, proporcionando maior rapidez na resolução de problemas e permitindo a comparação expedita de soluções. Além disso, abriu caminho a um conjunto de utilizadores nos meios académico, empresarial e administrativo que desta forma puderam a passar a utilizar Estatística como ferramenta eficaz na resposta aos seus problemas [1].

Apesar do seu caráter gratuito o R é uma ferramenta bastante poderosa com boas capacidades ao nível da programação e contém um conjunto bastante vasto (e em constante crescimento) de packages que acrescentam bastantes potencialidades à já poderosa versão base do R, PIAIRO e PEREIRA(2012, p.13).

1.2 Objetivos do estudo

O conhecimento e a aplicação da Estatística na prática requerem não só o estudo formal dos conteúdos envolvidos, mas também um indispensável trabalho de resolução de problemas e casos de estudo, usando diferentes meios e técnicas. É comum encontrarmos nos dias atuais escolas com salas equipadas com diversos materiais que podem constituir uma mais valia para a prática pedagógica.

De acordo com MENDES(2009, p.113) a informática, atualmente, é considerada uma das componentes tecnológicas mais importantes para a efetivação da aprendizagem matemática no mundo moderno. A sua relação com o Ensino da Matemática se estabelece a partir das perspetivas metodológicas atribuídas à informática como meio de superação de alguns obstáculos encontrados por professores e alunos no processo de ensino e aprendizagem.

A informática desde o seu advento, tem coadjuvado a Estatística, oferecendo ferramentas que efetuam cálculos morosos e complexos e facilitam a construção de cenários alternativos em prol da melhor compreensão dos dados e das estruturas em que se associam, CARVALHO(2015).

Com o crescente desenvolvimento tecnológico muitos materiais didáticos foram aprimorados e modernizados. É preciso, portanto, que a escola incorpore tais tecnologias, pois elas são importantes para a melhoria da qualidade do ensino aliada às inovações da prática pedagógica.

[1] Disponível em: `http://www.alea.pt/html/statofic/html/dossier/doc/publicacao_2009_web.pdf`(consultado a 28/03/2017)

Segundo FLORENTIN(2003, p.227), a exploração das possibilidades tecnológicas implica a construção de um saber matemático significativo; no âmbito do contexto educativo, tal exploração deveria constituir necessariamente uma obrigação para a política educacional, um desafio para professores e um incentivo para os alunos, no sentido de descobrir, ao menos, o necessário para a sua formação básica como ser integrante de uma sociedade que se transforma a cada dia.

Fundamentado nestas perspetivas, este estudo tem como principais objetivos:

1. Mostrar por meio de exemplos as diversas potencialidades que fazem do Excel e R, ferramentas adequadas para apoiar o trabalho do professor no ensino de Estatística;
2. Elaborar um material que sirva de apoio para o trabalho do professor e aprendizagem dos alunos.

O mesmo encontra-se dividido em cinco capítulos. Após o primeiro capítulo introdutório, no segundo, desenvolvem-se em detalhe diferentes modos de apresentação de dados através de tabelas e gráficos (distribuições de frequências, gráficos de barras, de setores, histogramas e polígono de frequências); o capítulo 3 refere-se a vários tipos de medidas de Estatística descritiva (medidas de localização, de dispersão, de assimetria, de curtose) e distribuições bidimensionais (diagrama de dispersão, reta de regressão simples e coeficiente de correlação linear); o quarto capítulo, trata-se de cálculos de distribuições de probabilidades para variáveis aleatórias discretas (Distribuição Binomial, Distribuição de Poisson, Distribuição Hipergeométrica), e para variáveis aleatórias contínuas (Distribuição Normal e Exponencial) e testes de hipóteses para igualdade e diferença de médias; o último capítulo é reservado para as conclusões finais.

Na elaboração deste trabalho, foi utilizada a versão 2016 do Excel da Microsoft e a versão `R-3.3.2`. Todavia, está garantida a compatibilidade de funções descritas com versões anteriores (pelo menos a partir da versão 2007 do Excel).

Capítulo 2

Distribuições de frequências e construções gráficas

2.1 Introdução

Um dos objetivos da Estatística é sintetizar os valores que uma ou mais variáveis podem assumir de modo a obter uma visão global da variação dessas variáveis. A Estatística consegue isto, inicialmente apresentando os valores em tabelas e gráficos que fornecem rápidas e seguras informações a respeito das variáveis em estudo. Por essa razão se fará um estudo mais completo sobre estas tabelas, mas antes é necessário definir alguns conceitos importantes.

2.1.1 População e amostra

Ao conjunto de indivíduos ou objetos que apresentam pelo menos uma característica em comum, denomina-se população. Ela pode ser finita ou infinita. Na maioria das vezes, por impossibilidade ou inviabilidade económica ou temporal, limita-se as observações referentes a uma determinada pesquisa apenas a uma parte da população. A essa parte proveniente da população em estudo denomina-se amostra. Ou seja, uma amostra é um subconjunto finito de uma população, FONSECA e MARTINS(1996, p.111).

2.1.2 Variáveis discretas e contínuas

Convencionalmente, o conjunto de resultados possíveis de um fenómeno denomina-se variável. Por exemplo, para o fenómeno **"Sexo"** são dois os resultados possíveis, masculino e femenino. Para o fenómeno **"Número de alunos"**, há um número de resultados possíveis expressos através dos números naturais (0, 1, 2, etc). Para o fenómeno **"Estatura"**, temos uma situação diferente, pois os resultados podem tomar número infinito de valores numéricos dentro de um determinado intervalo.

Os exemlos acima, mostram claramente que uma variável pode ser qualitativa ou quantitativa. Uma variável é dita qualitativa quando os seus valores são expressos por atributos (sexo, nacionalidade, cor da pele, etc). É quantitativa quando os seus valores são expressos em números (salário dos trabalhadores, idades dos alunos, etc). A variável quantitativa, atendendo os valores que pode tomar, pode classificar-se em variável contínua ou variável discreta. Uma variável é designada por contínua se, entre dois quaisquer valores, pode teoricamente assumir qualquer valor intermédio, caso contrário é designada por variável discreta, SPIEGEL(2001, p.2).

2.2 Distribuição de frequências

Uma distribuição de frequências é o arranjo dos valores x_i e as suas respetivas frequências fi_i. Quando se trata de variável discreta de variação relativamente pequena, é habitual apresentar-se uma tabela na seguinte forma:

x_i	fi_i
x_1	fi_1
x_2	fi_2
$\vdots$	$\vdots$
x_n	fi_n

Tabela 2.1: Tabela de distribuição de frequências

Quando o conjunto de dados em estudo é muito grande, muita das vezes é útil distribuir os dados em classes, determinando o número de indivíduos pertecentes a cada classe, designados por frequências. Para isto são fundamentais os seguintes elementos:

1. **Dados brutos**

   São dados recolhidos que ainda não foram organizados numericamente.

2. **Rol**

   É a organização dos dados brutos em ordem crescente ou decrescente.

3. **Amplitude total (At)**

   É a diferença entre o maior e o menor valor observados.

4. **Número de classes (k)**

   Designa a variação das variáveis num determinado conjunto. De modo geral, a sua determinação é feita utilizando a fórmula de Sturges, isto é, $k = 1 + 3,22 * log(n)$.

5. **Limites das classes**

   Os extremos de cada classe, denominam-se limites. O menor valor designa-se limite inferior (Li) e o maior valor limite superior (Ls).

6. **Amplitude das classes (h)**

   É o quociente entre amplitude total (At) e o número de classes.

7. **Frequência absoluta (fi)**

   É o número de vezes que o elemento aparece na mostra, ou o número de elementos pertecentes a uma classe. O somatório de todas frequências absolutas é igual ao tamanho da amostra, isto é, $\sum fi = n$

8. **Frequência relativa (fr)**

   É quociente entre o número de indivíduos que verificam um acontecimento (frequência absoluta) e o número total de indivíduos (amostra). E representa-se deste modo: $fr_i = \frac{fi_i}{n}$.

9. **Pontos médios das classes (pm)**

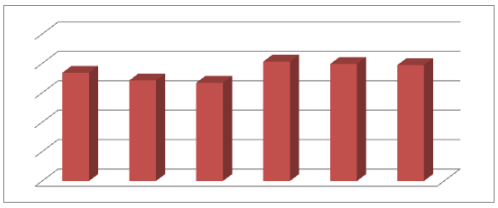
   É a média aritmética entre o limite superior e o limite inferior da classe.

2.3 Construções gráficas

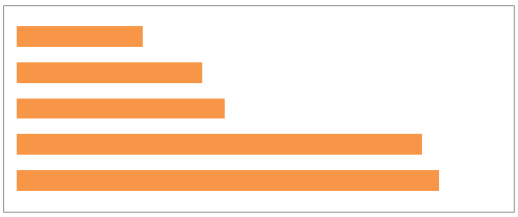
A representação gráfica dos dados estatísticos tem por finalidade, representar os resultados obtidos, permitindo chegar-se a conclusões sobre a evolução do fenómeno em estudo ou sobre como se relacionam os valores apresentados, REIS(2012, p.26). No entanto, para que isso seja conseguido a representação gráfica deve obedecer a certos requisitos: simplicidade, clareza e veracidade.

2.3.1 Gráfico de barras

É a representação de uma série por meio de retângulos, dispostos verticalmente, ou horizontalmente, FONSECA e MARTINS(1996, p.106). Constroi-se colocando os valores da variável em observação num dos eixos (horizontal) e as respetivas frequências (absolutas ou relativas) no outro eixo (vertical). Para cada valor da variável desenha-se, em seguida, um retângulo cuja altura deverá ser proporcional às frequências observadas. Nas figuras a seguir estão representados os exemplos genéricos de gráficos de barras.



(a) Gráfico de barras



(b) Gráfico de barras

2.3.2 Gráfico de setores

É a representação gráfica dos resultados num círculo, dividido em setores, REIS(2012, p.31). É utilizado principalmente quando se pretende comparar cada valor da parte com o total. Para construi-lo, divide-se o círculo em setores cujas áreas são proporcionais aos valores das partes. Na figura 2.1 apresenta-se um exemplo genérico de gráfico de setores.

Figura 2.1: Gráfico de setores

2.3.3 Histograma

Um histograma, ou histograma de frequências, consiste num conjunto de retângulos com as bases assentes no eixo horizontal e centrados na marca da classe e com largura igual à amplitude da classe, e às áreas proporcionais as frequências das respetivas classes, SPIEGEL(2001, p.8). Na figura 2.2 apresenta-se um exemplo genérico de um histograma.

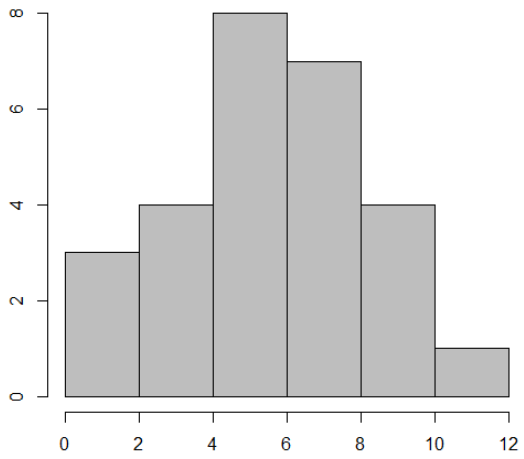


Figura 2.2: Histograma de frequências

No caso de termos sempre a mesma amplitude para todas as classes a altura do retângulo passa a ser proporcional à frequência. Se as amplitudes forem diferentes entre si convém normalizar todas as frequências para que a proporcionalidade das áreas se verifique. Isto poderá ser feito dividindo as frequências das classes pelas respetivas amplitudes e construindo o histograma a apartir destas frequências.

2.3.4 Polígono de frequências

Um polígono de frequências é um gráfico de linha das frequências das classes a passar pela marca de cada classe, SPIEGEL(2001, p.9). Pode ser construido ligando os pontos médios do topo de cada retângulo do histograma. Na figura 2.3 apresenta-se um exemplo de polígono de frequências.

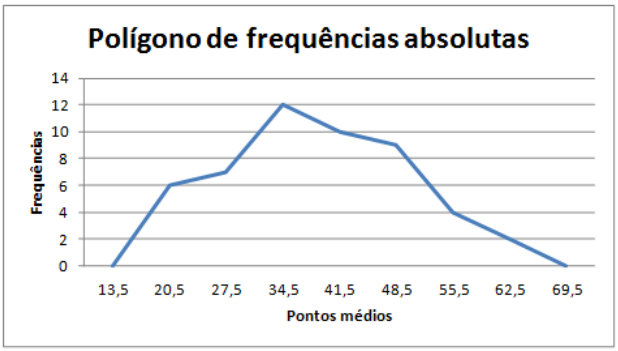


Figura 2.3: Polígono de frequências

2.3.5 Histograma e polígono de frequências

A partir do histograma é usual construir-se o polígono de frequências, uma forma visual alternativa ao histograma que se obtém unindo os pontos médios dos topos de retângulos com segmentos de reta (no caso das classes terem todas igual amplitude). Para o polígono, é necessário criar uma classe adicional em cada um dos extremos do histograma, com amplitude idêntica à das classes adjacentes e com frequência nula. Na figura 2.4 apresenta-se um histograma com polígono de frequências simultaneamente.

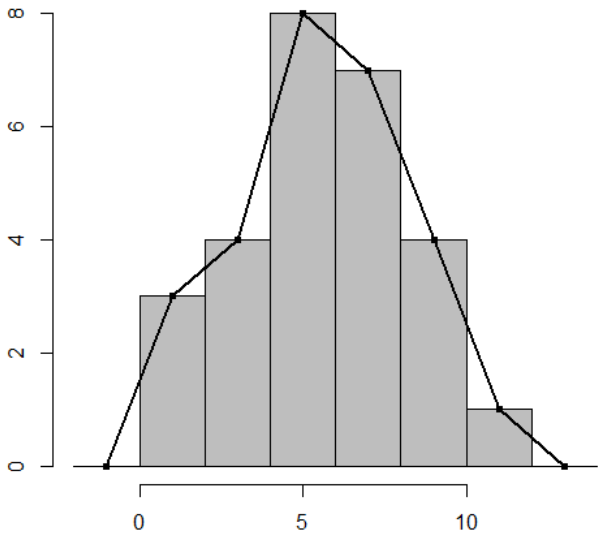


Figura 2.4: Histograma e polígono de frequências

2.4 Exemplos de distribuições de frequências e construções gráficas

Nesta secção são apresentados os procedimentos a seguir nas construções de tabelas de distribuições de frequências e gráficos, explorando a folha de cálculo do Excel e R.

Exemplo 1 *Um jornalista procura saber a idade de um grupo de alunos na paragem de autocarros do 1º de Maio, em Luanda e registou-se o seguinte:*

14, 18, 19, 15
15, 17, 15, 15
16, 16, 15, 15
14, 17, 14, 16
16, 14, 15, 16

1. *Construa*
   (a) *Tabela de distribuição de frequências para este conjunto;*
   (b) *Gráfico de barras e de setores.*

Resolução com Excel

Esta resolução obedece aos seguintes passos:

1. Digitar as idades dos alunos na folha de cálculo de $(A1 : D5)$.
2. Determinar a menor e a maior idade dos alunos, digitando:

   $=$ MÁXIMO$(A1 : D5)$;

   $=$ MÍNIMO$(A1 : D5)$.
3. Enumerar os valores inteiros entre a menor e a maior idade, de $(A8 : A13)$:
4. Calcular as frequências absolutas, selecionando primeiro a coluna onde serão inseridas as mesma, neste caso é de $(B8 : B13)$ e digitando:

   $=$ FREQUÊNCIA$(A1 : D5 \ ; \ A8 : A13)$, premindo: SHIFT + CTRL + ENTER.
5. Calcular o somatório das frequências absolutas, digitando em $B14$:

   $=$ SOMA$(B8 : B13)$.
6. Calcular as frequências relativas, digitando em $C8$:

   $=$ $(B8/ \ \$B\$14)$. Copiar o resultado de $C8$ e arrastá-lo até $C13$.
7. Calcular as frequências relativas em percentagem (%), digitando em $D8$:

   $=$ 100*$C8$, onde $C8$ corresponde ao primeiro valor de frequências relativas. Copiar o resultado de $D8$ e arrastá-lo até $D13$.
8. Calcular as frequências absolutas acumuladas, digitando em $E8$:

   $=$ SOMA$(\$B\$8 : \ B8)$, copiar o resultado de $E8$ e arrastá-lo até $E13$.
9. Calcular as frequências relativas acumuladas, digitando em $F8$:

   $=$ SOMA$(\$C\$8 : \ C8)$, copiar o resultado de $F8$ e arrastá-lo até $F13$.
10. Calcular as frequências relativas acumuladas em percentgem (%), digitando em $G8$:

    $=$ 100*$F8$, onde $F8$ corresponde ao primeiro valor de frequência relativa acumulada. Copiar o resultado de $G8$ e arrastá-lo até $G13$.

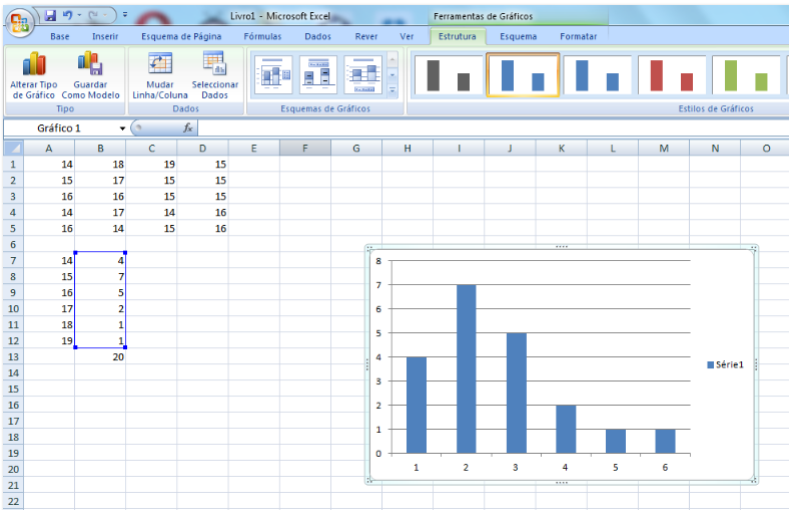
    O resultado final será semelhante conforme a tabela abaixo:

6	A	B	C	D	E	F	G
7	Xi	fi	fr	fr%	Fi	Fr	Fr%
8	14	4	0,20	20	4	0,20	20
9	15	7	0,35	35	11	0,55	55
10	16	5	0,25	25	16	0,80	80
11	17	2	0,10	10	18	0,90	90
12	18	1	0,05	5	19	0,95	95
13	19	1	0,05	5	20	1,00	100
14	Total	20	1,00	100			

Construção do gráfico de barras com Excel

Esta construção obedece aos seguintes passos:

1. Selecionar a coluna de frequências absolutas.
2. Inserir, gráfico de colunas ou barras, colunas empilhadas:



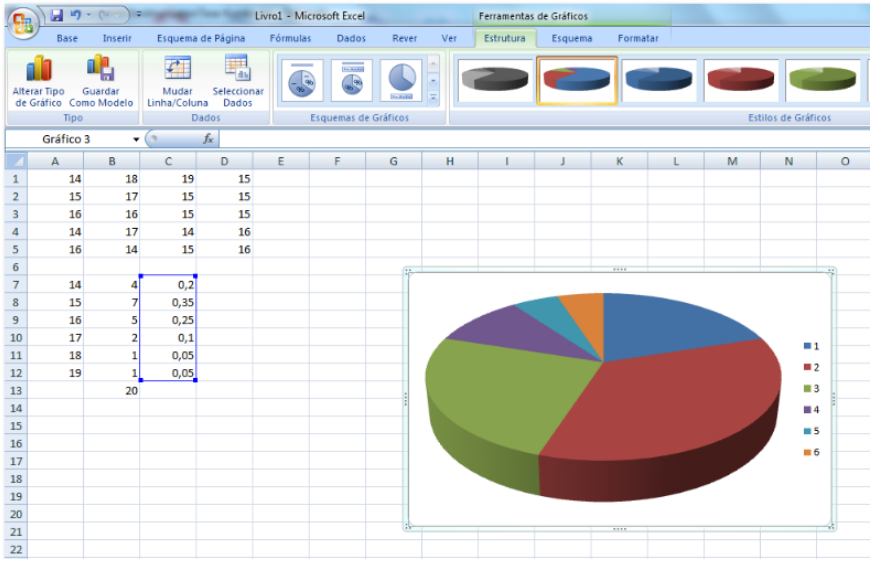
3. Substituir os valores apresentados no eixo horizontal do gráfico, pelos respetivos valores enumerados entre o mínimo e o máximo da amostra, selecionando:
   - Estrutura, selecionar dados;
   - Editar rótulos do eixo (categoria) horizontal;
   - Selecionar a coluna de valores de $(A7 : A12)$, ok.
4. Altere outros formatos do gráfico, selecionando os respetivos elementos, mudar o fundo do gráfico, adicionar o título do gráfico e dos eixos.

Figura 2.5: Gráfico de barras

Construção do gráfico de setores com Excel

Esta construção obedece aos seguintes passos:

1. Selecionar a coluna de frequências relativas;
2. Inserir, gráfico de pie em 3D, obtendo o seguinte resultado:



3. Substituir os valores apresentados na legenda, pelos respetivos valores enumerados entre o mínimo e o máximo da amostra, selecionando:
   - Legenda;
   - Estrutura, selecionar dados;
   - Editar rótulos do eixo (categoria) horizontal;
   - Selecionar a coluna de valores de $(A7 : A12)$, ok.

Figura 2.6: Gráfico de setores

Resolução com R

Esta construção obedece aos seguintes passos:

1. Digitar as idades dos alunos em new script:
   x_i =c(14, 18, 19, 15, 15, 17, 15, 15, 16, 16, 15, 15, 14, 17, 14, 16, 16, 14, 15, 16).
2. Calcular as frequências, digitando:
   - fi = table(x_i);
   - fr = $fi/sum(fi)$;
   - frp = 100*fr;
   - Fi = cumsum (fi);
   - Fr = cumsum (fr);
   - Frp = cumsum (frp);
3. Construir a tabela, digitando:
   tabela = cbind(fi, fr, frp, Fi, Fr, Frp).
   tabela.

O resultado final será semelhante conforme apresentado abaixo:

	fi	fr	frp	Fi	Fr	Frp
14	4	0.20	20	4	0.20	20
15	7	0.35	35	11	0.55	55
16	5	0.25	25	16	0.80	80
17	2	0.10	10	18	0.90	90
18	1	0.05	5	19	0.95	95
19	1	0.05	5	20	1.00	100

Construção do gráfico de barras com R

Esta construção obedece aos seguintes passos:

1. Digitar os seguintes dados em new script:
   - $x = c(14, 18, 19, 15, 15, 17, 15, 15, 16, 16, 15, 15, 14, 17, 14, 16, 16, 14, 15, 16)$;
   - $fi = table(x)$;
   - tabela =cbind(fi).
2. Construir o gráfico de barras digitando a seguinte síntaxe de comando:
   barplot(fi, col = "red", main = "Gráfico de colunas", xlab = "xi(idade em anos)", ylab = "fi(nº de alunos)").

O resultado esperado:

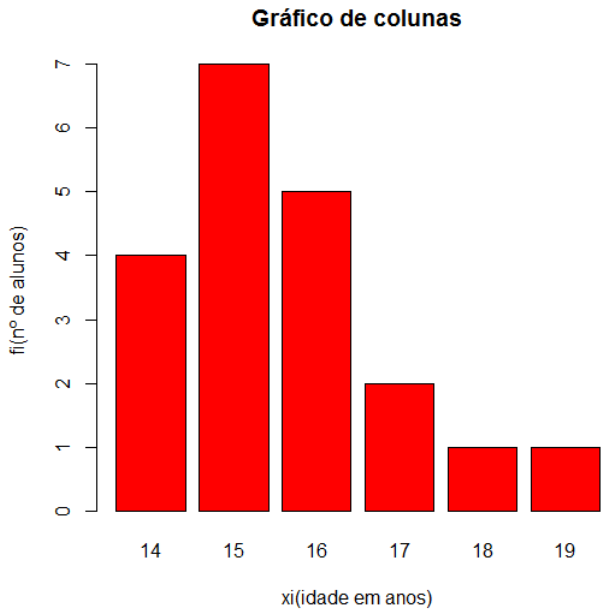


Figura 2.7: Gráfico de barras

Construção do gráfico de setores com R

Esta construção, obedece aos seguintes passos:

1. Digitar os dados da tabela em new script:
   - $x = c(14, 18, 19, 15, 15, 17, 15, 15, 16, 16, 15, 15, 14, 17, 14, 16, 16, 14, 15, 16)$;
   - $fi = table(x)$;
   - $fr = fi/sum(fi)$;
   - idades< $-$c(14,15,16,17,18,19).
2. Carregar o pacote plotrix.
3. Construir o gráfico digitando a seguinte síntaxe de comando:
   pie3D(fr, labels=idades, explode=0.0).

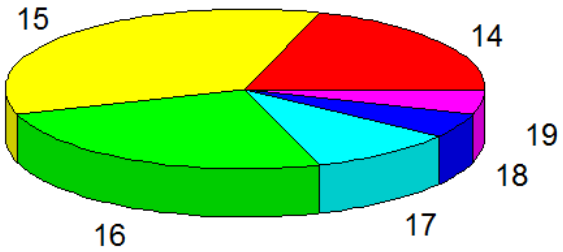


Figura 2.8: Gráfico de setores

Exemplo 2 *Com objetivo de divulgar um dos seus produtos, determinada Industria entrevistou 600 pessoas, para saber que veículo de informação (Jornal, Rádio, Revista e Televisão), era mais utilizado por elas. Dentre os entrevistados, 72 preferiram Jornal, 276 Rádio, 42 Revista e 210 Televisão.*

1. *Construa:*
   (a) *Uma tabela de distribuições de frequências relacionando os 4 veículos de informação;*
   (b) *Construir os gráficos de barras e de setores.*

Resolução com Excel

Esta construção, obedece aos seguintes passos:

1. Digitar os veículos de informação na folha de cálculo de A2 : A5 e as preferências de cada veículo de B2: B5.
2. Calcular o somatório das preferências de cada veículo, digitando em B6:
   =SOMA(B2: B5).
3. Calcular as frequências relativas, digitando em C2:
   $= B2/\$B\$6$, copiar o resultado de C2 e arrastá-lo até C5.
4. Calcular as frequências absolutas acumuladas, digitando em D2:
   $= SOMA(\$B\$2 : B2)$, copiar o resultado de D2 e arrastá-lo até D5.
5. Calcular as frequências relativas acumuladas, digitando em E2:
   $= SOMA(\$C\$2 : C2)$, copiar o resultado de E2 e arrastá-lo até E5.
6. Calcular as frequências relativas acumuladas em percentagens, digitando em F2:
   $= E2 * 100$, copiar o resultado de F2 e arrastá-lo até F5.

O resultado esperado:

	A	B	C	D	E	F
1	Veículos	fi	fr	fac	frac	frac%
2	Jornal	72	0,12	72	0,12	12
3	Rádio	276	0,46	348	0,58	58
4	Revista	42	0,07	390	0,65	65
5	Televisão	210	0,35	600	1	100
6		600				

Construção do gráfico de barras com Excel

Com a tabela de distribuição de frequências já elaborada, esta construção obedece aos seguintes passos:

1. Selecionar a coluna de frequências absolutas;
2. Inserir, gráfico de colunas ou barras, colunas empilhadas;
3. Substituir os valores apresentados no eixo horizontal do gráfico, pelos respetivos veículos de informação, selecionando:
   - Estrutura, selecionar dados;
   - Editar rótulos do eixo (categoria) horizontal;
   - Selecionar a lista de veículos de informação ou seja (A2 : A5), ok.
4. Altere outros formatos do gráfico, selecionando os respetivos elementos, mudar o fundo do gráfico, adicionar o título do gráfico e dos eixos.
5. O Resultado esperado:

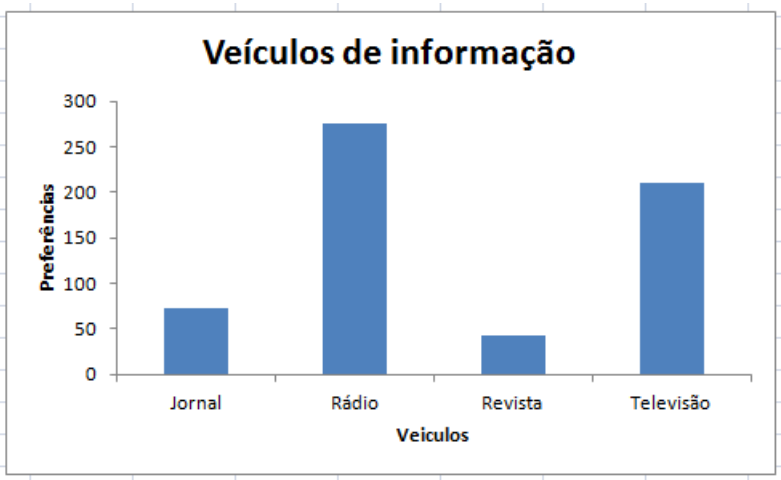


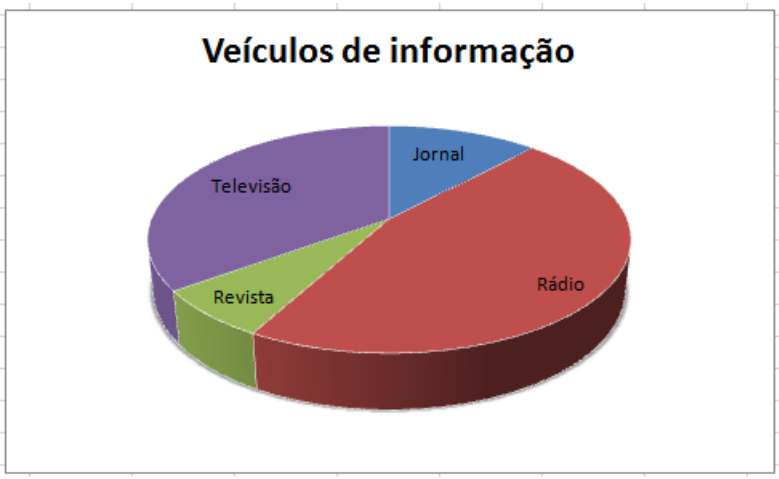
Figura 2.9: Gráfico de barras

Construção do gráfico de setores com Excel

Esta construção obedece aos seguintes passos:

1. Selecionar a coluna de frequências relativas, inserir gráfico de pie em 3D.
2. Substituir os valores apresentados na legenda, pelos respetivos veículos de informação, selecionando:
   - Legenda, estrutura, selecionar dados;
   - Editar rótulos do eixo (categoria) horizontal;
   - Selecionar a coluna de veículos ou seja(A2 : A5), ok.

3. Altere outros formatos do gráfico, selecionando os respetivos elementos, mudar o fundo do gráfico, adicionar o título do gráfico.



Resolução com R

Esta construção, obedece aos seguintes passos:

1. Digitar os seguintes dados em new script:
   - Veiculos=c("Jornal","Rádio","Revista","Televisão");
   - fi=c(72, 276, 42, 210).
2. Construir a tabela de distribuição de frequências, digitando:
   tabela=cbind(Veiculos,fi,fr=fi/sum(fi),fac=cumsum(fi),frac=cumsum(fi/sum(fi)), frap=cumsum(fi/sum(fi)*100)).

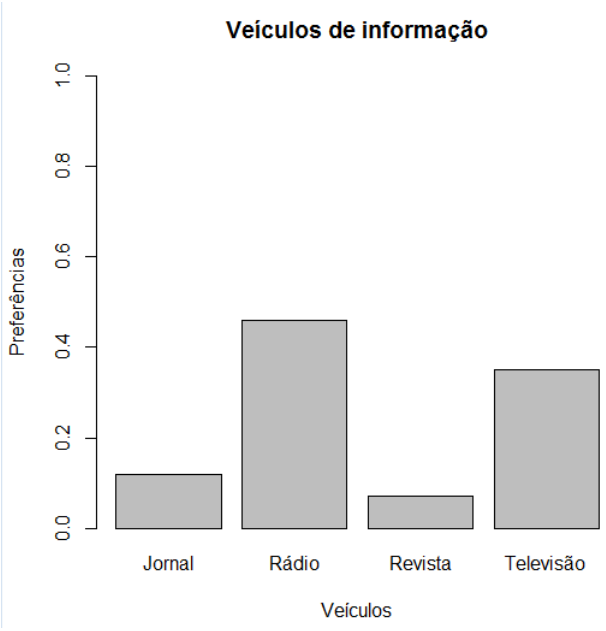
Veiculos	fi	fr	fac	frac	frap
Jornal	72	0.12	72	0.12	12
Rádio	276	0.46	348	0.58	58
Revista	42	0.07	390	0.65	65
Televisão	210	0.35	600	1	100

Construção do gráfico de barras com R

Esta construção, obedece aos seguintes passos:

1. Digitar os seguintes dados em new script:
   - veiculos< −c(rep(1, 72),rep(2, 276),rep(3, 42),rep(4, 210));
   - veiculos.nom< −c("Jornal","Rádio","Revista","Televisão").
2. Construir o gráfico de barras, digitando o seguinte comando:
   barplot(table(veiculos)/sum(table(veiculos)),ylim=c(0,1), names=veiculos.nom, xlab="Veículos", ylab="Preferências", main="Veículos de informação").

O resultado esperado:



Construção do gráfico de setores com R

Esta construção, obedece aos seguintes passos:

1. Digitar os seguinte dados em new script:
   - veiculos< −c(rep(1, 72),rep(2, 276),rep(3, 42),rep(4, 210));
   - veiculos.nom< −c("Jornal","Rádio","Revista","Televisão").
2. Carregar o pacote plotrix, digitando:
   - install.packages("plotrix")
   - library(plotrix)
3. Construir o gráfico de setores, digitando o seguinte comando: pie3D(table(veiculos)/sum(table(veiculos)),labels=veiculos.nom,explode=0.0).

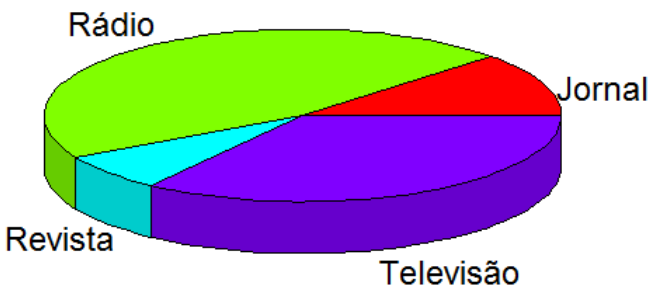


Figura 2.10: Gráfico de setores R

Os gráficos de barras e de setores são representações mais habituais para dados quantitativos discretos e qualitativos, podendo ambos ser feitos com as frequências absolutas ou com as relativas. Para este efeito, construimos primeiro a tabela de frequências onde fi e fr representam respetivamente a frequência absoluta e a relativa da i-ésima variável ou categoria.

O gráfico de barras é de representação imediata tanto no Excel como no R. No eixo das abcissas, representa-se as variáveis ou categorias e no eixo das ordenadas, as respetivas frequências. Para a representação do gráfico de setores, é necessário dividir o círculo em setores e afetá-los às diferentes variáveis ou categorias. No Excel, estas construções são feitas com base nas funções gráficas pré-definidas na folha de cálculo, conforme descrito nos exemplos anteriores. O R, utiliza os comandos barplot() e pie() para representar estes gráficos. No caso do gráfico de setores em 3D, recorre-se ao comando pie3D() do pacote plotrix, definindo devidamente as partes (labels).

Exemplo 3 *A procura diária de determinados produtos num certo estabelecimento comercial é dada por:*

B	B	C	B	C	C	C	D	D	A
A	C	A	C	D	D	A	B	A	B
A	A	C	A	D	D	D	C	A	B

Construa a tabela de distribuição de frequências: absolutas, relativas, absolutas acumuladas e relativas acumuladas.

Resolução com Excel

Esta resolução obedece aos seguintes passos:

1. Inserir os produtos na folha de cálculo de $(A1 : F5)$.
2. Digitar cada tipo de produto procurado de $(A8 : A11)$.
3. Calcular as frequências absolutas, selecionando primeiro a coluna onde serão inseridas, neste caso é, $(B8 : B11)$, digitando:
   = FREQUÊNCIA(CÓDIGO$(A1 : F5)$;CÓDIGO$(A8 : A11)$).
4. Calcular o somatório de frequências absolutas, digitando em $B12$:
   = SOMA$(B8 : B11)$.
5. Calcular as frequências relativas, digitando em $C8$:
   = $B8$/ \$$B$\$12. Copiar o resultado de $C8$ e arrastá-lo até $C11$.
6. Calcular o somatório de frequências relativas, digitando em $C12$:
   = SOMA$(C8 : C11)$.
7. Calcular as frequências absolutas acumuladas, digitando em $D8$:
   = SOMA(\$$B$\$8 : $B8$), copiar o resultado de $D8$ e arrastá-lo até $D11$.

8. Calcular as frequências relativas acumuladas, digitando em $F8$:

   $=$ SOMA($\$C\$8 : C8$), copiar o resultado de $F8$ e arrastá-lo até $F11$.

   O resultado esperado:

6	A	B	C	D	E
7	Produtos	fi	fr	fac	frac
8	A	9	0,3	9	0,3
9	B	6	0,2	15	0,5
10	C	8	0,266667	23	0,8
11	D	7	0,233333	30	1,0
12		30	1		

Resolução com R

Esta resolução obedece os seguintes passos:

1. Digitar os produtos procurados em new script:

   $x =$ c(”B”, ”B”, ”C”, ”BC”, ”C”, ”C”, ”D”, ”D”, ”A”, ”A”, ”A”, ”C”, ”A”, ”D”, ”D”, ”D”, ”C”, ”A”, ”B”, ”A”, ”C”, ”A”, ”C”, ”D”, ”D”, ”A”, ”B”, ”A”, ”B”).

2. Calcular as frequências, digitando:

   - fi $=$ table(x);
   - fr $= fi/sum(fi)$;
   - Fac $=$ cumsum (fi);
   - Frac $=$ cumsum (fr).

3. Construir a tabela, digitando:

   tabela $=$ cbind(fi, fr, Fac, Frac).

	fi	fr	fac	frac
A	9	0.30	9	0.30
B	6	0.20	15	0.50
C	8	0.27	23	0.77
D	7	0.23	30	1.00

Exemplo 4 *Fizeram-se 40 lançamentos de um dado e apuraram-se os seguintes resultados: o ás apareceu 5 vezes, o duque 8 vezes, o terno 5 vezes, a quadra 6 vezes, a quina 10 vezes, e a sena 6 vezes.*

Construa uma tabela de distribuição de frequências: absolutas, relativas, absolutas acumuladas, relativas acumuladas e reletiva acumulada em percentagem.

Resolução com Excel

Esta resolução obedece aos seguintes passos:

1. Digitar na folha de cálculo de (A3 : A8), os valores correspondentes a cada lado de um dado e os respetivos números de vezes que cada lado aparece de (B3 : B8);
2. Calcular o somatório de frequências absolutas, digitando em $B9$:
   $=$ SOMA$(B3 : B8)$.
3. Calcular as frequências relativas, digitando em $C3$:
   $= B3/\$B\$9$. Copiar o resultado de $C3$ e arrastá-lo até $C8$.
4. Calcular o somatório de frequências relativas, digitando em $C9$:
   $=$ SOMA$(C3 : C8)$.
5. Calcular as frequências absolutas acumuladas, digitando em $D3$:
   $=$ SOMA$(\$B\$3 : B3)$, copiar o resultado de $D3$ e arrastá-lo até $D8$.
6. Calcular as frequências relativas acumuladas, digitando em $E3$:
   $=$ SOMA$(\$C\$3 : C3\ )$, copiar o resultado de $E3$ e arrastá-lo até $E8$.
7. Calcular as frequências relativas acumuladas em percentgem (%), digitando em $F3$:
   $= 100{*}C3$, onde $C3$ corresponde ao primeiro valor de frequência relativa. Copiar o resultado de $F3$ e arrastá-lo até $F8$.

O resultado final será semelhante conforme apresentado abaixo:

1	A	B	C	D	E	F
2	xi	fi	fr	fac	frac	fr%
3	1	5	0,125	5	0,125	12,5
4	2	8	0,2	13	0,325	20
5	3	5	0,125	18	0,45	12,5
6	4	6	0,15	24	0,6	15
7	5	10	0,25	34	0,85	25
8	6	6	0,15	40	1	15
9		40	1			

Resolução com R

Esta resolução obedece aos seguintes passos:

1. Digitar os valores correspondentes a cada lado de um dado como sendo x e os respetivos número de vezes que cada lado aparece, como sendo f_i, isto é:
   $x =$ c(1, 2, 3, 4, 5, 6);
   fi $=$ c(5, 8, 5, 6, 10, 6).

2. Construir a tabela de distribuição de frequências, digitando:

   tabela = cbind(x, fi, fr = fi/sum(fi), fac = cumsum(fi), frac= cumsum(fi/sum(fi)), frp =100*fi/sum(fi));

   tabela.

   O resultado esperado será, semelhante conforme a tabela abaixo:

xi	fi	fr	fac	frac	frp
1	5	0.125	5	0.125	12.5
2	8	0.200	13	0.325	20.0
3	5	0.125	18	0.450	12.5
4	6	0.150	24	0.600	15.0
5	10	0.250	34	0.850	25.0
6	6	0.150	40	1.000	15.0

Exemplo 5 *Os dados abaixo indicados referem-se aos resultados de Matemática dos alunos do 3º ano do PUNIV do Cazenga, no ano letivo 2005.*

12, 12, 10, 11, 10, 8, 9, 18
13, 15, 11, 18, 16, 14, 10, 17
17, 11, 12, 12, 13, 14, 13, 11
10, 9, 8, 11, 12, 15, 16, 11
15, 10, 11, 14, 13, 12, 14, 10

1. *Construa:*

   (a) *Tabela de distribuição de frequências, tendo em conta em: fi, fr, fac, frac e frac%;*

   (b) *Apresente os dados na forma de um histograma.*

Resolução com Excel

Esta resolução obedece aos seguintes passos:

1. Digitar as notas na folha de cálculo de (A1 : G5);
2. Determinar os valores dos seguintes elementos:
   - Dimensão da amostra(n), digitando:
     = CONTAR. VAL(A1 : G5).
   - Máximo e mínimo da amostra, digitando:
     = MÁXIMO(A1 : G5);
     = MÍNIMO (A1 : G5).

- Amplitude total (At), digitando:
  = valor máximo−valor mínimo.
- Número de classes(k), digitando:
  = $ARRED.PARA.CIMA(1 + 3,22 * log(n);0)$, onde n é a dimensão da amostra.
- Intervalo de classe (h), digitando:
  =ARRED.PARA.CIMA(At/k;0).

3. Inserir as classes na folha de cálculo de (A8 : A13), isto é, partindo do valor mínimo de amostra adicionando amplitude de classe (h) até completar o número de classes determinado.

4. Determinar os possíveis valores de limite superior fechado de cada classe, para este caso, subtraindo uma unidade em cada limite superior da amostra. E, encontrando o primeiro valor em (B8), para obter outros valores correspondentes, basta digitar em B9:
   = (B8+h), onde h é o intervalo de classe, copiar o B9 e arrastá-lo até B13.

5. Calcular as frequências absolutas, selecionando primeiro a coluna onde serão inseridas as mesmas, neste caso é (C8 : C13) e digitar:
   = FREQUÊNCIA(A1 : G5; B8 : B13), premindo SHIFT+CTRL+ENTER.

6. Calcular o somatório das frequências absolutas, digitando em C14:
   =SOMA(C8 : C13).

7. Calcular as frequências relativas, digitando em D8:
   =C8/$C$14, copiar o resultado de D8 e arrastá-lo até D13.

8. Calcular as frequências absolutas acumuladas, digitando em E8:
   =SOMA($C$8 : C8), copiar o resultado de E8 e arrastá-lo até E13.

9. Calcular as frequências relativas acumuladas, digitando em F8:
   =SOMA($D$8 : D8), copiar o resultado do F8 e arrastá-lo até F13.

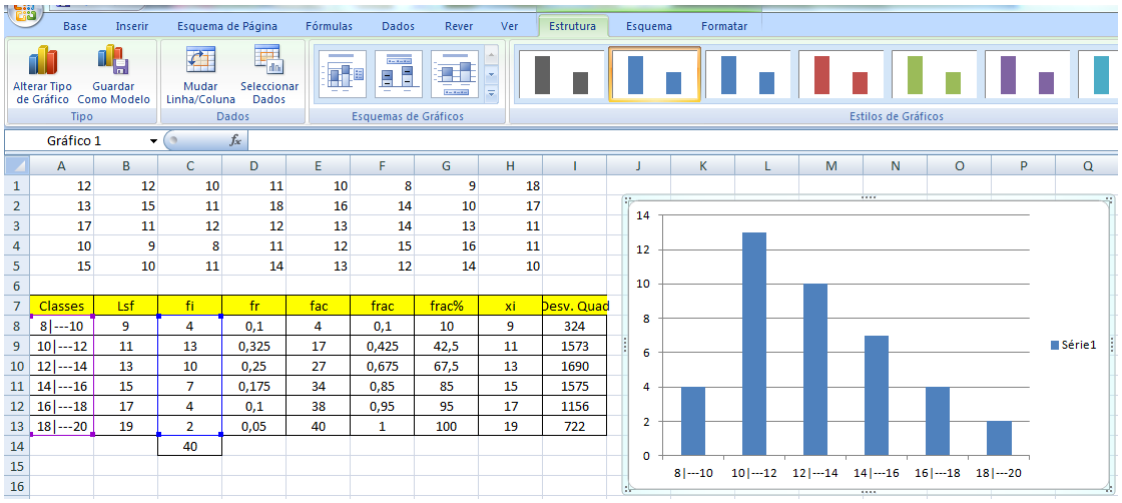
10. Calcular as frequências relativas acumuladas em percentagem(%), digitando em G8:
    = 100∗ F8, onde F8 corresponde ao primeiro valor de frequência relativa acumulada, copiar o resutado de G8 arrastá-lo até G13.

A	B	C	D	E	F	G
xi	Lsf	fi	fr	fac	frac	frac%
8|---10	9	4	0,10	4	0,10	10
10|---12	11	13	0,33	17	0,43	42,5
12|---14	13	10	0,25	27	0,68	67,5
14|---16	15	7	0,18	34	0,85	85
16|---18	17	4	0,10	38	0,95	95
18|---20	19	2	0,05	40	1,00	100
		40	1,00			

Construção de histograma com Excel

Com a tabela de distribuição de frequências já elaborada, para construir um histograma de frequências, pode-se proceder da seguinte forma:

1. Selecionar as classes e frequências absolutas respetivamente;
2. Inserir, gráfico de colunas ou barras;
3. Colunas empilhadas, obtendo o seguinte resultado:



	A	B	C	D	E	F	G	H	I
1	12	12	10	11	10	8	9	18	
2	13	15	11	18	16	14	10	17	
3	17	11	12	12	13	14	13	11	
4	10	9	8	11	12	15	16	11	
5	15	10	11	14	13	12	14	10	
6									
7	Classes	Lsf	fi	fr	fac	frac	frac%	xi	Desv. Quad
8	8|---10	9	4	0,1	4	0,1	10	9	324
9	10|---12	11	13	0,325	17	0,425	42,5	11	1573
10	12|---14	13	10	0,25	27	0,675	67,5	13	1690
11	14|---16	15	7	0,175	34	0,85	85	15	1575
12	16|---18	17	4	0,1	38	0,95	95	17	1156
13	18|---20	19	2	0,05	40	1	100	19	722
14			40						
15									
16									

4. Altere o formato do gráfico, selecionando:
   - As barras do gráfico construido;
   - Formatar seleção;
   - Largura do intervalo (deve ficar 0%);
   - Preenchimento, variação de cores por ponto;
   - Fechar.
5. Altere outros formatos do gráfico, selecionando os respetivos elementos, mudar o fundo, adicionar título do gráfico e dos eixos.

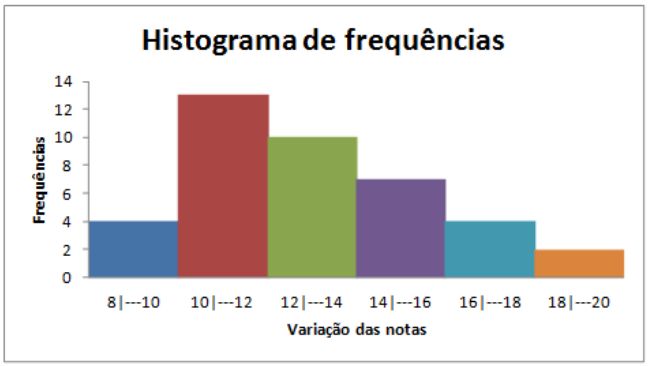


Figura 2.11: Histograma de frequências Excel

Resolução com R

Esta resolução obedece aos seguintes passos:

1. Digitar as notas dos alunos em new script:
   notas = c(12, 12, 10, 11, 10, 8, 9, 18, 13, 15, 11, 18, 16, 14, 10, 17, 17, 11, 12, 12, 13, 14, 13, 11, 10, 9, 8, 11, 12, 15, 16, 11, 15, 10, 11, 14, 13, 12, 14, 10).
2. Determinar os valores de:
   - Dimensão da amostra (n), digitando:
     lenght(notas).
   - A maior e a mínima nota, digitando:
     max(notas);
     min(notas).
   - Amplitude total (At), digitando:
     diff(range(notas)).
   - Número de classes (k), digitando:
     ceiling(log(n,2))+1, sendo n a dimensão da amostra.
   - Intervalo de classes(h), digitando:
     ceiling(At/k).
3. Carregar o pacote "fdth"(frequency, distribution table histogram and polygon), digitando:
   - Install packages("fdth");
   - library(fdth).
4. Construir a tabela, digitando:
   tabela = fdt(notas, start = 8, end = 20, h = 2).

Classes	fi	fr	fr(%)	Fac	Fr(%)
[8,10)	4	0.10	10.0	4	10.0
[10,12)	13	0.32	32.5	17	42.5
[12,14)	10	0.25	25.0	27	67.5
[14,16)	7	0.18	17.5	34	85.0
[16,18)	4	0.10	10.0	38	95.0
[18,20)	2	0.05	5.0	40	100.0

Construção do histograma com R

Esta construção, obedece aos seguintes passos:

1. Inserir as notas em new script:
   notas=c(12, 12, 10 ,11, 10, 8, 9, 18, 13, 15, 11, 18, 16, 14, 10, 17, 17, 11, 12, 12, 13, 14, 13, 11, 10, 9, 8, 11, 12, 15, 16, 11, 15, 10, 11, 14, 13, 12, 14, 10).

2. Construir o histograma, digitando:

   hist(notas, main="Histograma de frequências", xlab="notas", ylab="Frequências", col=rainbow(5), v=TRUE, cex=.8).

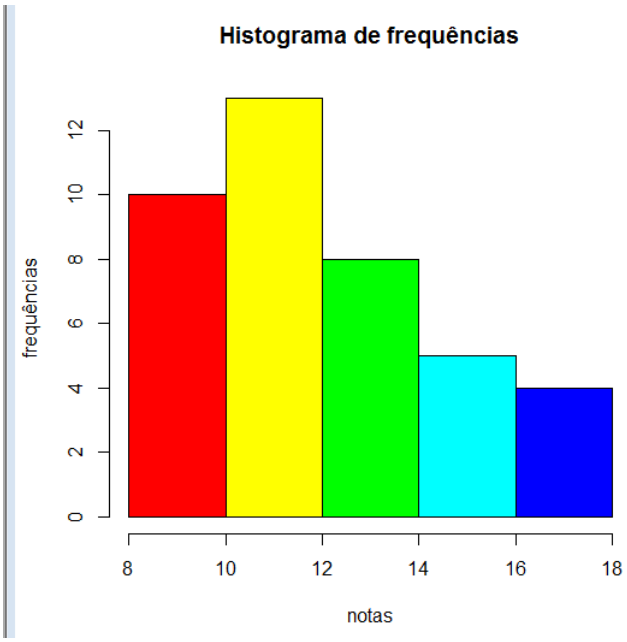


Figura 2.12: Histograma de frequências R

No caso de dados agrupados, o processo de construção de tabelas de distribuições de frequências no Excel é um pouco mais trabalhoso, já que a definição das classes não é tão imediata. É necessário determinar a dimensão da amostra, o máximo e o mínimo das obsrvações, o número de classes, a amplitude total, a amplitude de classe e definir ainda os possíveis limites superiores fechados de cada classe de modo a obter as frequências de cada classe de forma automática. No R, apesar de ser necessário determinar a dimensão da amostra, a amplitude de classe, o máximo e o mínimo das obsrvações, o processo é bem mais facilitado pelo simples uso do pacote "fdth".

A construção do histograma é relativamente simples, tanto no Excel como no R. Ambos os programas possuem funções e comandos que facilitam esta construção. No Excel parte-se de um gráfico de barras, fazendo algumas transformações no formato do gráfico, até obter o histograma. No R, destaca-se o comando hist() que constrói o histograma de forma rápida e eficiente. No caso de dados agrupados, pode recorrer-se ainda ao comando plot() que também facilita esta construção.

Exemplo 6 *Abaixo temos a distribuição de frequência de pesos de uma amostra de 100 alunos:*

Peso em kg	30|--- 40	40|---50	50|---60	60|---70	70|---80
Nº de alunos	10	20	35	25	10

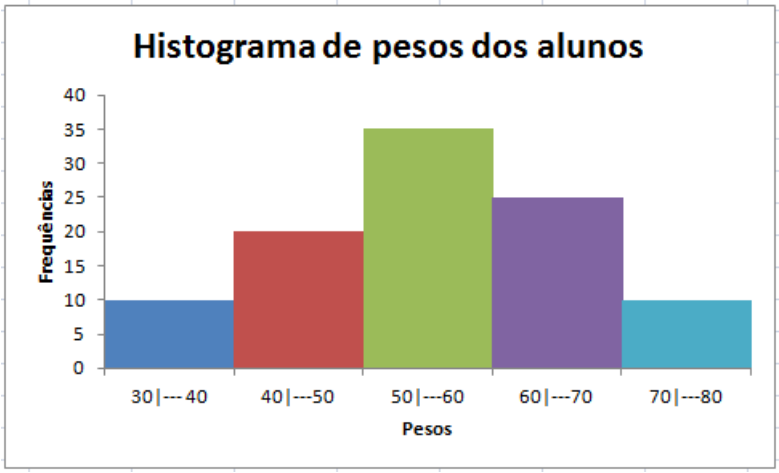
Apresente os dados na forma de um histograma.

Resolução com Excel

Esta resolução obedece aos seguintes passos:

1. Inserir os dados da tabela na folha de cálculo de A1: F2;
2. Selecione os dados da tabela ou seja A1: F2;
3. Inserir, gráfico de colunas ou barras, colunas empilhadas;
4. Altere o formato do gráfico, selecionando:
   - As barras do gráfico construido;
   - Formatar seleção, largura do intervalo (deve ficar 0%);
   - Preenchimento, variação de cores por ponto, fechar.
5. Altere outros formatos do gráfico, selecionando os respetivos elementos, adicionar o título do gráfico, mudar o fundo.

O resultado esperado:



Resolução com R

Esta resolução obedece aos seguintes passos:

1. Carregar o pacote (fdth), digitando:
   - Install packages("fdth");
   - library(fdth).
2. Digitar os seguintes dados em new script:
   tab=make.fdt(f=c(10,20,35,25,10),start=30,end=80)
3. Construir o histograma, digitando:
   plot(tab,main="Histograma de pesos dos alunos",xlab="Pesos",ylab="frequências", col=rainbow(5),v=TRUE,cex=.8).

O resultado final será semelhante, conforme o gráfico apresentado a baixo:

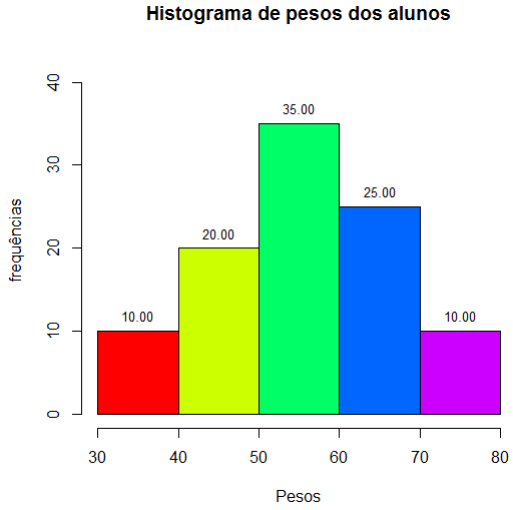


Figura 2.13: Histograma R

Exemplo 7 *Na Maternidade Lucrécia Paim, em Luanda, registou-se durante uma semana o nascimento de 34 bebés com os seguintes pesos (em kgs):*

3,45	3,50	2,85	3,00	3,40	3,70	3,05
3,55	3,15	3,80	4,00	3,15	4,20	3,40
3,40	3,50	2,90	2,80	4,10	4,30	3,20
3,35	3,40	2,73	2,75	4,78	4,50	4,00
3,00	3,30	2,95	4,15	4,40	4,60	

Tabela 2.2: Pesos de bebés em kgs

1. *Construa:*
   (a) *Tabela de distribuições de frequências tendo em conta em: fi, fr, fr%, fac e frac%;*
   (b) *Histograma e polígono de frequências.*

Resolução com Excel

Esta resolução obedece aos seguintes passos:

1. Digitar o peso de cada bebé na folha de cálculo de (A3 : G6)
2. Determinar os valores dos seguintes elementos:
   - Dimensão da amostra(n), digitando:
     = CONTAR. VAL(A3 : G6).
   - Máximo e mínimo da amostra, digitando:
     = MÁXIMO(A3 : G6);
     = MÍNIMO (A3 : G6).

- Amplitude total (At), digitando:
  = valor máximo−valor mínimo.
- Número de classes(k), digitando:
  = $ARRED.PARA.CIMA(1 + 3,22 * log(n);0)$, onde n corresponde a dimensão da amostra.
- Intervalo de classe (h), digitando:
  =ARRED.PARA.CIMA(At/k;0).

3. Inserir as classes na folha de cálculo de (A11 : A16), isto é, partindo do valor mínimo de amostra adicionando o intervalo de classe (h) até completar o número de classes determinado.

4. Determinar os possíveis valores de limite superior fechado de cada classe, para este caso, subtraindo (0,01) em cada limite superior da amostra. E, encontrando o primeiro valor em (B11), para obter outros valores correspondentes, basta digitar em B12:

   = (B11+h), onde h é o intervalo de classe, copiar o resultado do B12 e arrastá-lo até B16.

5. Calcular as frequências absolutas, selecionando primeiro a coluna onde serão inseridas as mesmas, neste caso é (C11 : C16) e digitando:

   = FREQUÊNCIA(A3 : G6; B11 : B16), premindo SHIFT+CTRL+ENTER.

6. Calcular o somatório das frequências absolutas, digitando em C17:

   =SOMA(C11 : C16).

7. Calcular as frequências relativas, digitando em D11:

   = C11/$C$17, copiar o resultado de D11 e arrastá-lo até D16.

8. Calcular as frequências relativas em percentagem, digitando em E11:

   = 100*D11, onde D11 corresponde ao primeiro valor de frequências relativas. Copiar o resultado de E11 e arrastá-lo até E16.

9. Calcular as frequências absolutas acumuladas, digitando em F11:

   =SOMA($C$11 : C11), copiar o resultado de F11 e arrastá-lo até F16.

10. Calcular as frequências relativas acumuladas em percentagem(%), digitando em G11:

    = SOMA($E$11 : E11), onde E11 corresponde ao primeiro valor de frequência relativa em percentagem copiar o resutado de G11 arrastá-lo até G16.

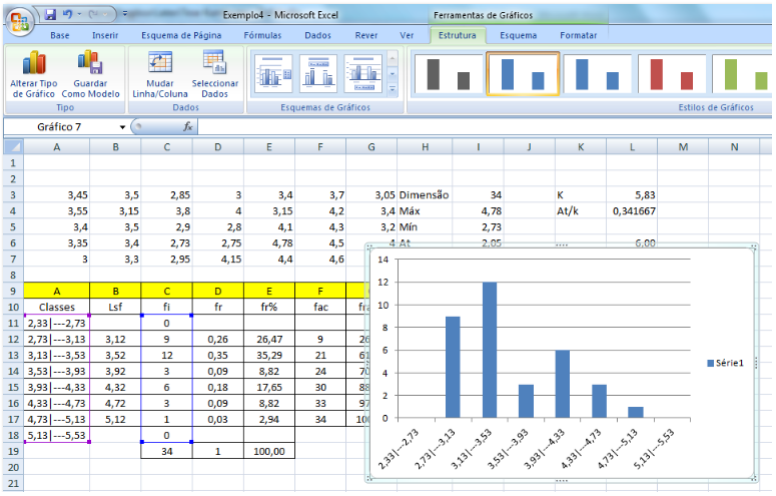
    O resultado final será semelhante conforme apresentado a seguir:

9	A	B	C	D	E	F	G
10	Classes	Lsf	fi	fr	fr%	fac	frac%
11	2,73|---3,13	3,12	9	0,26	26,47	9	26,47
12	3,13|---3,53	3,52	12	0,35	35,29	21	61,76
13	3,53|---3,93	3,92	3	0,09	8,82	24	70,59
14	3,93|---4,33	4,32	6	0,18	17,65	30	88,24
15	4,33|---4,73	4,72	3	0,09	8,82	33	97,06
16	4,73|---5,13	5,12	1	0,03	2,94	34	100,00
17			34	1	100,00		

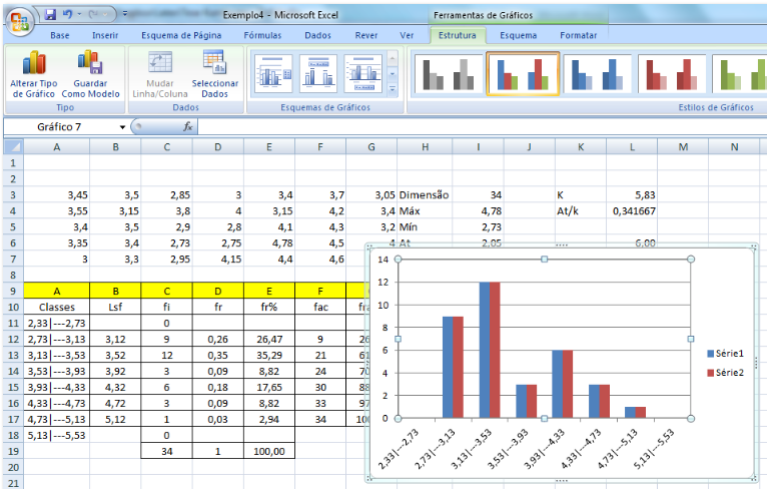
Construção de histograma e polígono de frequências com Excel

Para construir um histograma simultaneamente com o polígono de frequências usando Excel, é necessário antes de tudo criar uma classe adicional em cada um dos extremos, com amplitude idêntica à das classe adjacentes com frequência nula.

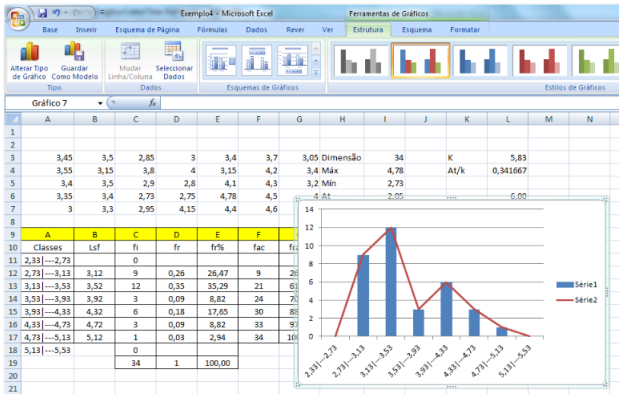
1. Construir o histograma e polígono de frequências, selecionando:
   - As classes e as respetivas frequências absolutas ou seja (A11 : A18) e (C11 : C18);
   - Inserir, gráfico de coluna ou de barras;
   - Colunas agrupadas, obtendo o seguinte resultado:



2. Adicionar nava série, selecionando:
   - As barras do gráfico construido;
   - Estrutura, selecionar dados;
   - Adicionar, valores da série, apagando = {1};
   - Selecionar as frequências, ok.

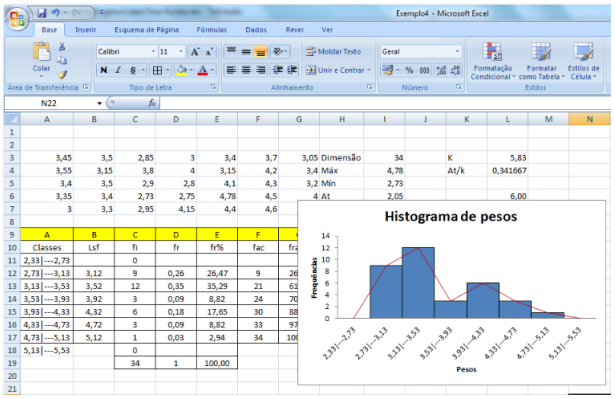


3. Selecionar as barras da nova série, alterar tipo do gráfico, linha com marcadores, obtendo o seguite resultado:



4. Alterar o formato do gráfico, selecionando:

   - As barras, formatar seleção;
   - Largura do intervalo (deve ficar 0%);
   - Cor do limite, linha contínua, definir a cor, fechar, obtendo o seguinte resultado:



5. Altere outros formatos do gráfico, selecionando os respetivos elementos, mudar o fundo, adicionar o título do gráfico e dos eixos.

Resolução com R

Para resolver este problema no R, é fundamental substituir a virgula por ponto e obedece aos seguintes passos:

1. Digitar o peso de cada bebé em new script:

   pesos = c(3.55, 3.15, 3.80, 4.00, 3.15, 4.20, 3.40, 3.40, 3.50, 2.90, 2.80, 4.10, 4.30, 3.20, 3.35, 3.40, 2.73, 2.75, 4.78, 4.50, 4.00, 3.00, 3.30, 2.95, 4.15, 4.40, 4.60).

2. Determinar os valores de:

   - Dimensão da amostra (n), digitando:
     lenght(pesos).
   - O valor máximo e mínimo observado, digitando:
     max(pesos);
     min(pesos).
   - Amplitude total (At), digitando:
     diff(range(pesos)).
   - Número de classes (k), digitando:
     ceiling(log(n,2))+1, sendo n a dimensão da amostra.
   - Intervalo de classes(h), digitando:
     ceiling(At/k).

3. Com o pacote (fdth) carregado, pode-se construir a tabela, digitando:

   tabela = fdt(pesos, start = 2.73, end = 5.13, h = 0.4).

   O resultado final será semelhante conforme a tabela abaixo:

Classes	fi	fr	fr(%)	Fac	Fr(%)
[2.73,3.13)	9	0.26	26.47	9	26.47
[3.13,3.53)	12	0.35	35.29	21	61.76
[3.53,3.93)	3	0.09	8.82	24	70.59
[3.93,4.33)	6	0.18	17.65	30	88.24
[4.33,4.73)	3	0.09	8.82	33	97.06
[4.73,5.13)	1	0.03	2.94	34	100.00

Construção de histograma e polígono de fequências com R

Esta construção no R obedece aos seguintes passos:

1. Digitar os dados em new script:

   pesos = c(3.55, 3.15, 3.80, 4.00, 3.15, 4.20, 3.40, 3.40, 3.50, 2.90, 2.80, 4.10, 4.30, 3.20, 3.35, 3.40, 2.73, 2.75, 4.78, 4.50, 4.00, 3.00, 3.30, 2.95, 4.15, 4.40, 4.60).

2. Construir o histograma, digitando:

   - h=hist(pesos, plot=FALSE);
   - int=h$breaks[2]−h$breaks[1];
   - end=length(h$breaks);
   - h1=hist(pesos, col="mediumseagreen", main="Histograma e polígono de frequências", breaks=c(h$breaks[1]−int, h$breaks, h$breaks[end]+int)).

3. Construir o polígono de frequências, digitando:

   lines(h1$mids, h1$counts, type="o", pch=20,lwd=2).

   O resultado esperado:

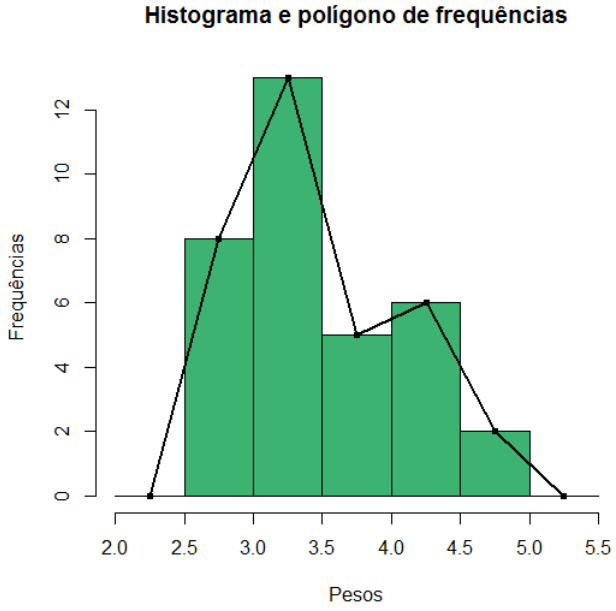


Figura 2.14: Histograma e polígono R

Exemplo 8 *Considere o seguinte conjunto de 50 notas dos alunos (dadas em percentagem):*

33, 35, 35, 39, 41, 41, 42, 45, 47, 48
50, 52, 53, 54, 55, 57, 59, 60, 60, 55
61, 64, 65, 66, 66, 66, 67, 68, 65, 65
69, 71, 73, 73, 74, 74, 77, 77, 78, 76
80, 81, 84, 85, 85, 88, 89, 91, 94, 97

Definindo previamente as classes para este conjunto, construa um polígono de frequências.

Resolução com Excel

Esta resolução obedece aos seguintes passos:

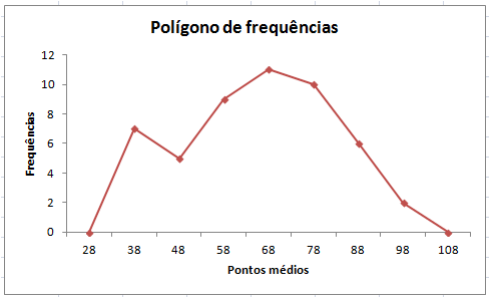
1. Digitar as notas na folha de cálculo de A1: J5;
2. Agrupar os dodos em classes, procedendo como no caso anterior;
3. Construir a tabela de distribuição de frequências, procedendo da seguinte forma:
   - Criar uma classe adicional em cada um dos extremos, com amplitude idêntica à das classe adjacentes com frequência nula;
   - Inserir as classes na folha de cálculo de (A8 : A16);
   - Determinar os possíveis valores de limite superior fechado de cada classe, para este caso, subtraindo uma unidade em cada limite superior da amostra. E, encontrando o primeiro valor em (B8), para obter outros valores correspondentes, basta digitar em B9:
     = (B8+h), onde h é o intervalo de classe, copiar o B9 e arrastá-lo até B16.
   - Calcular as frequências absolutas, selecionando primeiro a coluna onde serão inseridas as mesmas, neste caso (C8 : C16) e digitar:
     = FREQUÊNCIA(A1 : J5; B8 : B16), premindo SHIFT+CTRL+ENTER.
   - Calcular os pontos médios, digitando em D8 e D9:
     $= (23 + 33)/2$;
     $= D8 + h$, onde h corresponde ao intervalo de classe, copiar o resultado de D9 e arrastá-lo até D16.

	A	B	C	D
1	Classes	LSF	fi	xi
2	23|---33	32	0	28
3	33|---43	42	7	38
4	43|---53	52	5	48
5	53|---63	62	9	58
6	63|---73	72	11	68
7	73|---83	82	10	78
8	83|---93	92	6	88
9	93|---103	102	2	98
10	103!---113	112	0	108

4. Construir o polígono de frequências, procedendo da seguinte forma:
   - Selecionar as classes e as frequências;
   - Inserir, linha com marcadores;
   - Estrutura, selecionar dados;
   - Editar rótulo do eixo (categoria) horizontal;
   - Selecionar a coluna dos pontos médios, ok.

5. Altere outros formatos do gráfico, selecionando os respetivos elementos, mudar o fundo do gráfico, adicionar o título do gráfico e dos eixos.

O resultado esperado será semelhante conforme o gráfico apresentado a seguir:



Resolução com R

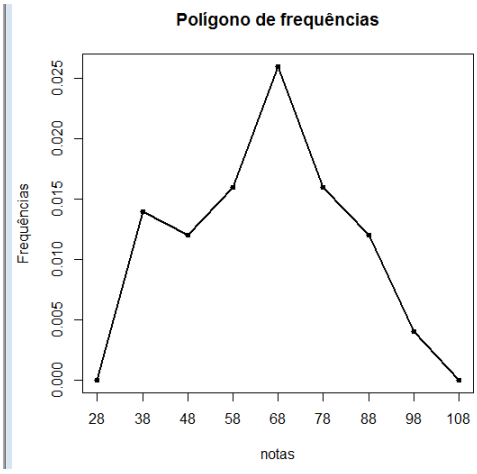
Esta resolução obedece aos seguintes passos:

1. Digitar as notas em new script:

   notas=c(33, 50, 61, 69, 80, 35, 52, 64, 71, 81, 35, 53, 65, 73, 84, 39, 54, 65, 73, 85, 41, 55, 66, 74, 85, 41, 55, 66, 74, 88, 42, 57, 66, 76, 89, 45, 59, 66, 77, 91, 47, 60, 67, 77, 94, 48, 60, 68, 78, 97).

2. Construir o polígono de frequências, digitando:

   - $h1 <-$hist(notas,plot=F,breaks=c(33,43,53,63,73,83,93,103));
   - $x <-$c(28,h1$mids,108);
   - $y <-$c(0,h1$counts/(50*10),0);
   - plot(x,y,type="o",xlab="notas",ylab="Frequências",main="Polígono de frequências",pch=20,xlim=c(28,108),xaxt="n",lwd=2);
   - axis(1,at=x).



Exemplo 9 *Os resultados do 1º teste de Matemática nas quatro turmas da 9ª classe do IMEL foram:*

Turma A	Turma B
11 10 0 10 10	10 10 7 10 14
14 14 16 14 11	18 10 12 20 2
10 11 12 15 6	10 10 12 11 15
12 10 11 11 10	12 19 11 2 0
11 14 10 10 11	0 8 10 11 15
Turma C	Turma D
14 10 12 14 2	6 7 10 0 10
10 16 10 14 6	14 6 15 3 10
10 0 11 11 10	11 14 10 16 0
10 11 10 11 10	12 13 11 13 11
10 12 12 12 10	12 12 14 6 11

Pretende-se construir a respetiva tabela de distribuição de frequências, mas por ser elevado o número de resultados pensou-se na forma de agrupar os dados em classes, classificando-os em Mau de 0 −4, Medíocre de 5−9, suficiente de 10−13, Bom de 14−17 e Muito Bom de 18−20. Assim sendo, agrupou-se os dados em cinco classes, conforme a tabela apresentado a seguir:

Classes	fi	fac	fr	fr%	prac
Mau 0 − 4	10	10	0,1	10	0,1
Med. 5 − 9	8	18	0,08	8	0,18
Suf. 10 − 13	61	79	0,61	61	0,79
Bom 14 − 17	18	97	0,18	18	0,97
M. Bom 18 − 20	3	100	0,03	3,	1
Total	100		1	100	

Tabela 2.3: Tabela de distribuição de frequências das classificações

Neste exemplo, constatou-se que a amplitude varia em cada uma das classes. No entanto, serão apresentados os procedimentos a seguir para resolver este tipo de problema usando Excel e R:

Resolução com Excel

Esta resolução obedece aos seguintes passos:

1. Digitar as classificações dos alunos na folha de cálculo de (A1 : J10).
2. Determinar os valores dos seguintes elementos:
   - Dimensão da amostra(n), digitando:
     = CONTAR. VAL(A1 : J10).
   - Máximo e mínimo da amostra, digitando:
     = MÁXIMO(A1 : J10);
     = MÍNIMO(A1 : J10).

- Amplitude total (At), digitando:
  = valor máximo−valor mínimo.
- Número de classes(k), digitando:
  $= ARRED.PARA.CIMA(1 + 3,22 * log(n);0)$, onde n corresponde a dimensão da amostra.
- Intervalo de classe (h), digitando:
  =ARRED.PARA.CIMA (At/k;0).

3. Inserir as classes na folha de cálculo de (A13 : A17);

4. Inserir na folha de cálculo de (B13 : B17) os limites superiores de cada classe ou seja $(4, 9, 13, 17, 20)$;

5. Calcular as frequências absolutas, selecionando primeiro a coluna onde serão inseridas as mesmas, neste caso é (C13 : C17) e digitando:
   = FREQUÊNCIA(A1 : J10; B13 : B17), premindo SHIFT+CTRL+ENTER.

6. Calcular o somatório das frequências absolutas, digitando em C18:
   =SOMA(C13 : C17).

7. Calcular as frequências absolutas acumuladas, digitando em D13:
   =SOMA($C$13 : C13), copiar o resultado de D13 e arrastá-lo até D17.

8. Calcular as frequências relativas, digitando em E13:
   =C13/$C$18, copiar E13 e arrastá-lo até E18.

9. Calcular as frequências relativas em percentagem, digitando em F13:
   = 100*E13, onde E13 corresponde ao primeiro valor de frequências relativas. Copiar o resultado de F13 e arrastá-lo até F18.

10. Calcular as frequências relativas acumuladas, digitando em G13:
    = SOMA($E$13 : E13), onde E13 corresponde ao primeiro valor de frequência relativa, copiar o resutado de G13 arrastá-lo até G17.

    O resultado final será semelhante conforme a tabela abaixo:

A	B	C	D	E	F	G
Classes	Ls	fi	fac	fr	fr%	prac
Mau 0 -- 4	4	10	10	0,1	10	10
Med. 5 -- 9	9	8	18	0,08	8	18
Suf. 10 -- 13	13	61	79	0,61	61	79
Bom 14 -- 17	17	18	97	0,18	18	97
M. Bom 18 -- 20	20	3	100	0,03	3	100
		100		1		

Resolução com R

Esta resolução obedece aos seguintes passos:

1. Digitar as classificações em new script:

   notas=c(11, 10, 0, 10, 10, 14, 14, 16, 14, 11, 10, 11, 12, 15, 6, 12, 10, 10, 11, 11, 10, 11, 14, 10, 10, 11, 10, 10, 7, 10, 14, 18, 10, 12, 20, 2, 10, 10, 12, 11, 15, 12, 19, 11, 2, 0, 0, 8, 10, 11, 15, 14, 10, 12, 14, 2, 10, 16, 10, 14, 6, 10, 0, 11, 11, 10, 10, 11, 10, 11, 10, 10, 12, 12, 12, 10, 6, 7, 10, 0, 10, 14, 6, 15, 3, 10, 11, 14, 10, 16, 0, 12, 13, 11, 13, 11, 12, 12, 14, 6, 11).

2. Determinar as frequências de cada classe, digitando:

   - h=hist(notas, plot=FALSE, breaks=c(0, 4, 9, 13, 17, 20));
   - h$counts.

3. Carregar o pacote (fdth):

4. Construir a tabela, digitando:

   - classif< − c(rep(1, 10),rep(2, 8),rep(3, 61),rep(4, 18),rep(5, 3));
   - classif.nom< − c("Mau(0–4)", "Med.(5–9)", "Suf.(10–13)", "Bom(14–17)", "M. Bom(18–20)");
   - tb=fdt(classif, start=1, end=6);
   - tb2=tb$table;
   - tb2[,1]=classif.nom;
   - tb2.

O resultado esperado será semelhante conforma a tabela abaixo:

Class limits	f	rf	rf(%)	cf	cf(%)
Mau(0--4)	10	0.10	10	10	10
Med.(5--9)	8	0.08	8	18	18
Suf.(10--13)	61	0.61	61	79	79
Bom(14--17)	18	0.18	18	97	97
M. Bom(18--20)	3	0.03	3	100	100

Capítulo 3

Medidas de Estatística descritiva

3.1 Introdução

No capítulo anterior, vimos como é possível sintetizar os dados sob forma de tabelas, gráficos e distribuições de frequências, usando as ferramentas Excel e R. Neste, são apresentados as funções e comandos para o cálculo de medidas de Estatística descritiva usando as mesmas ferramentas.

3.2 Medidas de localização

Muita das vezes torna-se útil descrever um conjunto de dados estatísticos por meio de um valor apenas. São as medidas de localização ou medidas de tendência central, REIS(2012, p.64). São chamadas medidas de tendência central, pois representam o fenómeno pelos seus valores médios, em torno dos quais tendem a concentrar-se os dados. As principais medidas de localização são: a média, a mediana e a moda.

3.2.1 Média aritmética amostral

A média aritmética é a soma de todos valores observados dividida pelo número de observações. As fórmulas de cálculo da média aritmética varia consoante o tipo de distribuição. Sejam $x_1, x_2, ..., x_n$, portanto n valores da variável X. A média aritmética amostral de x representada por $\bar{x}$ é definida por:

$$\boxed{\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{\sum x_i}{n}} \tag{3.1}$$

Quando os dados estiverem agrupados numa distribuição de frequências, a média aritmética dos valores $x_1, x_2, ..., x_n$ é ponderada pelas respetivas frequências absolutas: $f_1, f_2, ..., f_n$.

$$\boxed{\bar{x} = \frac{\sum x_i.f_i}{n}} \tag{3.2}$$

Existem casos em que para efetuar o cálculo da média aritmética, é fundamental considerar as várias classes ou intervalos em que a variável se encontra dividida. Neste caso, a média calcula-se substituindo os valores da amostra pelos centros ou pontos médios de cada classe.

3.2.2 Mediana

A mediana de um conjunto de valores colocados em ordem crescente ou decrescente é o valor que divide a população ou amostra em duas partes iguais, REIS(2012, p.78). Existe um método prático para o cálculo da mediana:

- Se n for ímpar, a mediana será o elemento central (de ordem $\frac{n+1}{2}$);
- Se n for par, a mediana será a média entre os elementos centrais (de ordem $\frac{n}{2}$ e $\frac{n}{2}+1$).

3.2.3 Moda

A moda é o valor mais frequente de uma distribuição. Para variáveis discretas, a identificação da moda é facilitada pela simples observação do elemento que apresenta maior frequência. Pode existir mais do que uma moda. Se houver uma só moda, a distribuição diz-se unimodal, se houver duas modas, diz-se bimodal, se houver três ou mais modas, multimodal.

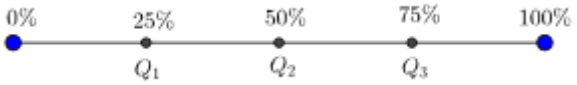
3.3 Medidas de dispersão

São medidas estatísticas utilizadas para avaliar o grau de variabilidade, ou dispersão, dos valores em torno da média, REIS(2012, p.98).

As medidas de dispersão servem para verficarmos a representatividade das medidas de localização, pois é muito comum encontrarmos variáveis que, apesar de terem a mesma média, são compostas de valores bem distintos.

3.3.1 Quartis

Os quartis dividem um conjunto de dados em quatro partes iguais.



O primeiro quartil (Q_1), deixa 25% dos elementos, o segundo (Q_2) coincide com a mediana, deixando 50% dos elementos, enquanto que o terceiro quartil (Q_3), deixa 75% dos elementos.

Existem várias formas de cálcular os quartis, dentre elas são úteis:

1. Método exclusivo:

   Quando n é ímpar, este método funciona da seguinte forma:

   - Ordenar os dados por ordem crescente e calcular a mediana;
   - O 1º quartil (Q_1), é o valor central de dados que ficam a esquerda da mediana sem a incluir;
   - O 3º quartil (Q_3), é o valor central de dados que ficam a direita da mediana sem a incluir.

   Seja o conjunto: $7, 15, 36, 39, 40, 41, 50$:

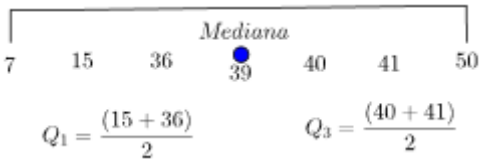
Para este conjunto, temos:

2. Método inclusivo:

   Como no caso anterior, quando n é ímpar, este método funciona da seguinte forma:

   - Ordenar os dados por ordem crescente e calcular a mediana;
   - O 1º quartil (Q_1), é a média aritmética de valores centrais do conjunto de dados que ficam a esquerda da mediana incluindo-a;
   - O 3º quartil (Q_3), é a média aritmética de valores centrais do conjunto de dados que ficam a direita da mediana incluindo-a.

   Assim sendo, do conjunto anterior, temos:



No caso do conjunto ser par 2n com $n \in \mathbb{N}$, não há ambiguidade na definição dos quartis que serão sempre as medianas de conjuntos de dados com n elementos.

Um método mais elaborado para o cálculo dos quartis (Muteira, 2002), será descrito a seguir, por ser também o método utilizado na folha de cálculo de Excel da Microsoft. Neste processo, considera-se na determinação do 1º e 3º quartis, respetivamente as posições $\frac{n+3}{4}$ e $\frac{3n+1}{4}$. No que respeita o 1º quartil, se $\frac{n+3}{4}$ for inteiro, então Q_1 será a observação de ordem $\frac{n+3}{4} = k$ na amostra ordenada. Se $\frac{n+3}{4}$ não for inteiro, seja $\frac{n+3}{4} = k + \epsilon$, em que representamos por ϵ a parte decimal (0,25, 0,5 ou 0,75). Para obter o 1º quartil considera-se uma interpolação linear, fazendo uma média ponderada entre a observação de ordem k e a observação de ordem $(k+1)$, [1]. Assim, temos:

$$\boxed{Q_1 = \text{Observação de ordem} \quad k + \epsilon \times (\text{observação de ordem}(k+1) - \text{observação de ordem} \quad k)} \tag{3.3}$$

O raciocínio utilizado para obter o 3º quartil é idêntico ao descrito para a determinação do 1º quartil, considerando agora a posição $\frac{3n+1}{4}$.

Seja o conjunto 109, 125, 126, 130, 130, 130, 131, 132, 133, 133, 133, 133, 134, 134, 140, 140, 142, 144, 145, 152.

De acordo com a metodologia descrita anteriormente o 1º quartil e o 3º quartil estão respetivamente nas posições 5.75 e 15.25 pelo que realizando as interpolações lineares, tem-se:

$Q_1 = 130 + 0,75 \times (130 - 130) = 130 \quad$ e $\quad Q_3 = 140 + 0,25 \times (140 - 140) = 140.$

[1] Disponível em: `http://wikiciencias.casadasciencias.org/wiki/index.php/Quartis` (consultado a 27/04/2017)

Quando a dimensão da amostra for ímpar, este processo conduz aos mesmos resultados com os do método inclusivo.

Os decis são os valores da variável que dividem a distribuição em dez partes iguais, enquanto que os percentis a dividem em 100 partes iguais. Assim sendo, temos o número de decis é nove e o de percentis é 99.

3.3.2 Variância amostral

Para dados simples, a variância é a soma do quadrado das diferenças entre os valores da amostra e a média, dividida pelo número total das observações subtraido de uma unidade:

$$\boxed{s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}} \tag{3.4}$$

Para dados agrupados, a variância amostral é ponderada pelas respetivas frequências.

$$\boxed{s^2 = \frac{\sum(x_i - \bar{x})^2 . f_i}{n-1}} \tag{3.5}$$

3.3.3 Desvio-padrão

O desvio-padrão é a raiz quadrada da variância. No entanto, para calcular o desvio-padrão deve-se primeiramente determinar o valor da variância e, em seguida, extrair a raiz quadrada desse resultado.

$$\boxed{s = \sqrt{s^2}} \tag{3.6}$$

3.3.4 Coeficiente de variação

Trata-se de uma medida relativa de dispersão útil para a comparação em termos relativos do grau de concentração em torno da média de séries distintas. É dado por:

$$\boxed{CV = \frac{S}{\bar{x}} \times 100} \tag{3.7}$$

O coeficiente de variação é expresso em percentegem.

3.4 Medidas de assimetria

O grau de afastamento de uma distribuição da unidade de simetria, denomina-se assimetria, MARTINS e FONSECA (1996, p.148).

O método mais simples para se medir o grau de assimetria de uma distribuição consiste na comparação de três medidas de tendência central: a média, a mediana e a moda.

Em uma distribuição simétrica tem-se igualdade dos valores da média, mediana, e moda. Apresenta-se a seguir um exemplo gráfico de distribuição simétrica.

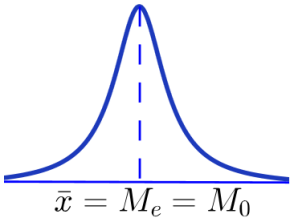


Figura 3.1: Distribuição simétrica

Em uma distribuição assimétrica positiva ou assimétrica à direita, tem-se:

$M_0 < M_e < \bar{x}$. Eis um exemplo gráfico de distribuição assimétrica positiva

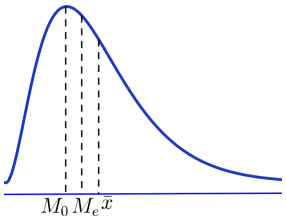


Figura 3.2: Distribuição simétrica positiva

Em uma distribuição assimétrica negativa, ou assimétrica à esquerda, tem-se:

$\bar{x} < M_e < M_0$. Eis um exemplo gráfico de distribuição assimétrica negativa:

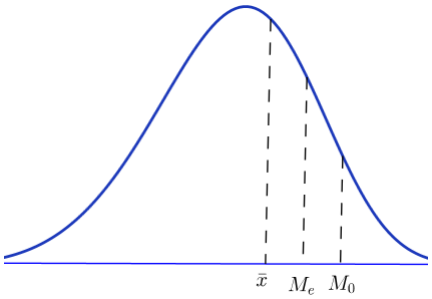


Figura 3.3: Distribuição simétrica negativa

Existem várias fórmulas para o cálculo do coeficiente de assimetria, dentre elas são úteis:

$$\boxed{g_1 = \frac{3(\bar{x} - M_e)}{s}} \tag{3.8}$$

Pearson definiu o segundo método de calcular o grau de assimetria de uma distribuição quando não dispomos da média e do desvio padrão, utilizando apenas os quartis da distribuição.

$$\boxed{g_2 = \frac{Q_3 + Q_1 - 2M_e}{Q_3 - Q_1}} \tag{3.9}$$

O Coeficiente de assimetria de Fisher é dado por:

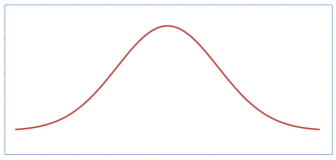
$$\boxed{g = \frac{m_3}{\sqrt{m_2^3}}} \tag{3.10}$$

Para identificar o tipo de assimetria, analisa-se as seguintes condições:

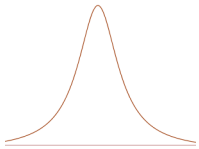
- Se g= 0 a distribuição é simétrica;
- Se g> 0 a distribuição é assimétrica positiva;
- Se g< 0 a distribuição é assimétrica negativa.

3.5 Medidas de curtose

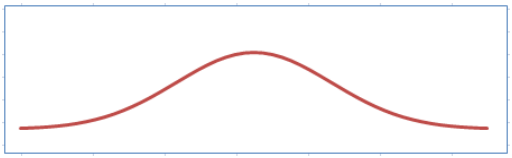
O grau de achatamento de uma distribuição, denomina-se curtose REIS(2012). E pode ser: mesocúrtica, leptocúrtica e platicúrtica. Nas figuras a baixos apresenta-se exemplos de gráficos de curtoses.



(a) Mesocúrtica



(b) Leptocúrtica



(c) Platicúrtica

Para medir o achatamento (curtose), usa-se o coeficiente de Fisher ou coeficiente de curtose:

$$\boxed{b_2 = \frac{m_4}{m_2^2}} \tag{3.11}$$

onde m_2 e m_4 são o segundo e o quarto momentos centrais respetivamente, dados por:

$m_4 = M[(x - \bar{x})^4]; \quad m_2 = M[(x - \bar{x})^2] \quad$ e $g_2 = b_2 - 3$

Classificação das distribuições quanto à kurtosis e achatamento:

- Se $b_2 = 3 \quad e \quad g_2 = 0$ a distribuição de frequências é mesocúrtica;
- Se $b_2 < 3 \quad e \quad g_2 < 0$ a distribuição diz-se platicúrtica;
- Se $b_2 > 3 \quad e \quad g_2 > 0$ a distribuição é leptocúrtica.

3.6 Exemplos de cálculo de medidas de Estatística descritiva

Nesta secção, são apresentados os comandos e os procedimentos a seguir para calcular medidas de Estatística descritiva, usando as ferramentas Excel e R.

Exemplo 10 *Usando os dados do exemplo 1, calcule a média das idades dos alunos, a mediana, a moda, a variância e o desvio-padrão.*

Resolução manual do problema

Como podemos observar, temos 20 dados, mas apenas 6 variáveis (xi) e o número de alunos para cada idade é a frequência (fi).

A média é dada por: $\bar{x} = \frac{\sum xifi}{n} = \frac{312}{20} = 15,6$.

Como $n = 20$ é par, então a ordem dos elementos que compõem a mediana é dada por: $\frac{n}{2}$ e $\frac{n}{2} + 1$. O valor da mediana pode ser obtido, interpolando os valores que estão na posição 10 e 11 do conjunto ordenado das observações. Isto é, $Me = x_{10} + 0.5(x_{11} - x_{10})$. Assim: $Me = 15$.

A identificação da moda é facilitada pela simples observação do elemento que apresenta maior frequência. Assim para esta distribuição a moda é $M_0 = 15$.

Para o cálculo da variância o processo é facilitado por conhecermos já a média. Resta encontrar o valor de $\sum \frac{x_i^2 f_i}{n}$ Assim:

A variância é dada por: $s^2 = \frac{\sum x_i^2 f_i}{n} - \bar{x}^2 = \frac{4902}{20} - (15,6)^2 = 1,74$.

O desvio padrão é dado por: $s = \sqrt{s^2} = \sqrt{1,74} = 1,32$.

Resolução com Excel

Esta resolução, obedece aos seguintes passos:

1. Digitar os dados na folha de cálculo de (A1 : D5)
2. Calcular a média, digitando:
   =MÉDIA(A1 : D5).
3. Calcular a mediana, digitando
   = MED(A1 : D5).
4. Calcular a moda, digitando
   = MODA(A1 : D5).
5. Calcular a variância, digitando
   = VAR.S(A1 : A5).
6. Calcular o desvio-padrão, digitando:
   = DESVPAD(A : A5).

Resolução com R

Esta resolução obedece aos seguintes passos:

1. Digitar os dados em new script:
   idades= c(14, 18, 19, 15, 15, 17, 15, 15, 16, 16, 15, 15, 14, 17, 14, 16, 16, 14, 15, 16).
2. Calcular a média, digitando:
   mean(idades).

3. Calcular a mediana, digitando:
   median(idades).
4. Carregar o pacote (modeest), para calcular a moda:
   mfv(idades).
5. Calcular a variância, digitando:
   var(idades).
6. Calcular o desvio-padrão, digitando:
   sd(idades).

Exemplo 11 *Fez-se um estudo sobre o número de filhos a 100 famílias residentes no Huambo, tendo-se obtido os seguintes resultados:*

Nº de filhos	0	1	2	3	4	5
Nº de famílias	12	26	21	20	16	5

1. *Determine:*
   (a) *A média da distribuição;*
   (b) *A mediana da distribuição;*
   (c) *A moda da distribuição;*
   (d) *A variância;*
   (e) *O desvio médio e padrão da distribuição.*

Resolução manual do problema

Sendo xi a variável representativa de número de filhos e fi número de famílias, temos:

- A média é dada por: $\bar{x} = \frac{\sum xifi}{n} = \frac{217}{100} = 2,17$.

Como $n = 100$ é par então, a mediana é a média entre os elementos de ordem $\frac{n}{2}$ e $\frac{n}{2} + 1$ ou seja $\frac{100}{2} = 50$ e $\frac{100}{2} + 1 = 51$. Identifica-se os elementos de 50 e 51 pelas frequências acumuladas.

- Assim sendo, para esta distribuição a mediana é: $Me = 2$;
- A moda é o valor mais frequente. Para esta distribuição é $M_0 = 1$.

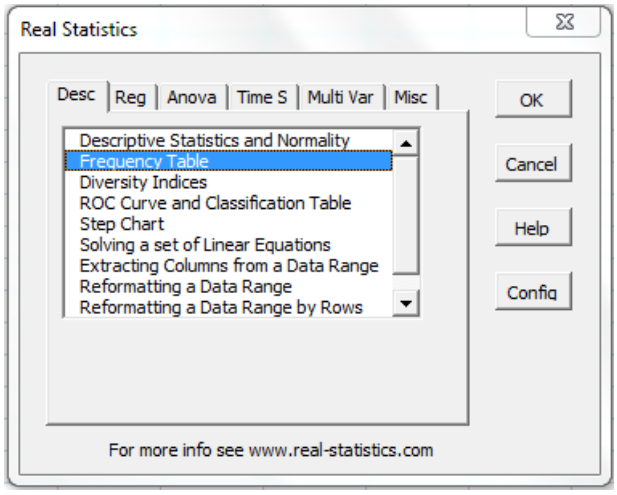
Como no caso anterior, o cálculo da variância é facilitado por conhecermos já a média. Resta encontrar o valor de $\sum \frac{x_i^2 f_i}{n}$ Assim:

- A variância é dada por: $s^2 = \frac{\sum x_i^2 f_i}{n} - \bar{x}^2 = \frac{671}{100} - (2,17)^2 = 2,01$;
- O desvio médio é dado por: $\frac{\sum |x_i - \bar{x}| f_i}{n} = \frac{120,06}{100} = 1,20$;
- O desvio padrão é dado por: $s = \sqrt{s^2} = \sqrt{2,01} = 1,42$.

Resolução com Excel

O Excel dispõe do pacote "Real Statistics"[2] que devolve medidas de Estatística descritiva. Usando, este pacote a resolução do problema obedece os seguintes passos:

1. Digitar o número de filhos na folha de cálculo de A2 : A7, e o número de famílias de B2 : B7;
2. Carregar o pacote "Real Statistics", premindo CTRL + M.



3. Converter os dados numa lista, selecionando:
   - Frequency table, ok;
   - Input range (A2: B7);
   - Frequency table;
   - Convert to Raw data;
   - Output range (D1). Obtendo uma lista de valores que varia de (D3: D102).
4. Calcular a média, digitando:
   =MÉDIA (D3: D102).
5. Calcular a mediana, digitando:
   =MED (D3: D102).
6. Calcular a moda, digitando:
   =MODA (D3: D102).
7. Calcular a variância, digitando:
   = VAR.S (D3: D102).
8. Calcular o desvio médio, digitando:
   = DESV.MÉDIO (D3: D102).
9. Calcular o desvio padrão, digitando:
   =DESVPAD (D3: D102).

[2]Disponível em: http://www.real-statistics.com/, (consultado a 29/04/2017)

Resolução com R

Seja x a variável representativa de número de filhos e y o número de famílias, esta resolução obedece aos seguintes passos:

1. Digitar os dados em new script:
   - x =c(0, 1, 2, 3, 4, 5);
   - y =c(12, 26, 21, 20, 16, 5);
   - tb< −c(rep(0, 12), rep(1, 26), rep(2, 21), rep(3, 20), rep(4, 16), rep(5, 5)).
2. Calcular a média digitando:
   mean(tb).
3. Calcular a mediana digitando:
   median(tb).
4. Calcular a moda, digitando:
   mfv(tb).
5. Calcular a variância digitando:
   var(tb).
6. Carregar o pacote (lsr), para calcular o desvio médio:
   aad(tb).
7. Calcular o desvio-padrão, digitando:
   sd(tb).

Exemplo 12 *Numa empresa de construção civil com 300 funcionários, fez-se a sua separação em grupos etários nos termos da tabela abaixo indicada.*

Idades	Nº de funcionários
$15-25$	56
$25-35$	42
$35-45$	62
$45-55$	70
$55-65$	48
$65-75$	22

1. *Determine:*
   (a) *A média;*
   (b) *Mediana;*
   (c) *A variância;*
   (d) *O desvio-padrão e o coeficiente de variação.*

Resolução manual do problema

Neste caso, para calcular a média as classes são representadas pelos seus pontos médios:

- $\bar{x} = \frac{\sum x_i f_i}{n} = \frac{12780}{300} = 42,6.$

Para a mediana, calcula-se $\frac{n}{2}$, como $n = 300$, temos $\frac{300}{2} = 150$. Identifica-se a classe mediana pelas frequências acumuladas, neste caso é a 3ª, isto é, 35 à 45. Aplicando a fórmula:

- $Md = li_{Md} + \frac{(n/2 - \sum f)}{f_{Md}} \times h = 35 + \frac{(150-98)}{62} \times 10 = 43,387.$
- A variância é dada por: $s^2 = \frac{\sum x_i^2 f_i}{n} - \bar{x}^2 = \frac{615000}{300} - (42,6)^2 = 235,24.$
- O desvio padrão é dado por: $s = \sqrt{s^2} = \sqrt{235,24} = 15,33.$
- O coeficiente de variação é dado por: $Cv = \frac{s}{\bar{x}} \times 100 = \frac{15,33}{42,6} \times 100 = 35,9\%.$

Resolução com Excel

Esta resolução, obedece aos seguintes passos:

1. Digitar as idades na planilha de A2: A7 e o número de funcionários em cada grupo etário de B2: B7.
2. Calcular o somatório do número de funcionários, digitando em B8:
   =SOMA(B2: B7).
3. Calcular as frequências relativas, digitando em C2:
   $= B2/\$B\$8$, copiar o resultado de C2 e arrastá-lo até C7.
4. Calcular as frequências absolutas acumuladas, digitando em D2:
   =SOMA($B$2: B2), copiar o resultado de D2 e arrastá-lo até D7.
5. Calcular os pontos médios, digitando em E2 e E3:
   =(15+25)/2;
   =E2+10, copiar o resultado de D3 e arrastá-lo até D7.
6. Calcular a média, digitando em B10:
   =SOMARPRODUTO(E2 : E7; C2 : C7).
7. Calcular a mediana, digitando em B11 e B12:
   =ÍNDICE(A2: A7; CORRESP(B8/2; D2: D7;1)+1; 1);
   $=35+\frac{(B8/2-D3)}{B4} * C10.$
8. Calcular a variância, digitando em F2 e B13:
   $= (E2 - \$B\$10)^2 * B2$, copiar o resultado de F2 e arrastá-lo até F7;
   =SOMA(F2: F7)/(B8-1).

9. Calcular o desvio padrão, digitando em B14:
   =RAIZQ(B13).
10. Calcular o coeficiente de variação, digitando em B15:
    $= \frac{B14}{B10} * 100$.

O resultado esperado:

	A	B	C	D	E	F	G
1	Idades	fi	fr	fac	xi	Desv. Quad	frac
2	15|---25	56	0,187	56	20	28602,56	0,187
3	25|---35	42	0,140	98	30	6667,92	0,327
4	35|---45	62	0,207	160	40	419,12	0,533
5	45|---55	70	0,233	230	50	3833,2	0,767
6	55|---65	48	0,160	278	60	14532,48	0,927
7	65|---75	22	0,073	300	70	16516,72	1,000
8		300					
9	Resultados						
10	Média	42,6	10				
11		35|---45					
12	Mediana	43,3871					
13	Var	236,0268					
14	Desv-pad	15,36316					
15	Cv	36,06376					

Resolução com R

Esta resolução obedece aos seguintes passos:

1. Carregar o pacote fdth (frequency distribution table histogram end polygon).
2. Construir a tabela de distribuição de frequências, digitando:
   tb=make.fdt(f=c(56, 42, 62, 70, 48, 22), start= 15, end= 75).
3. Calcular a média, digitando:
   mean(tb).
4. Calcular a mediana, digitando:
   median(tb).
5. Calcular a variância, digitando:
   var(tb).
6. Calcular o desvio padrão, digitando:
   sd(tb).

7. Calcular o coeficiente de variação, digitando:
   $\frac{sd(tb)}{mean(tb)} * 100$.

No Excel, o cálculo de medidas de Estatística descritiva para dados agrupados, requer um conhecimento de diversos passos manuais de forma a manusear os dados para obter os resultados desejados. Por exemplo, para obter o valor aproximado para a média no exemplo apresentado acima, procedeu-se da seguinte forma:

- Construimos uma tabela de distribuição de frequências;
- Adicionou-se na mesma uma nova coluna com os pontos médios das amplitudes das classes que se obtêm fazendo a semi-soma dos limites das amplitudes;
- Adicionou-se ainda outra coluna com os produtos dos pontos médios das amplitudes das classes pelas frequências relativas;
- Determinou-se o somatório dos produtos dos pontos médios das amplitudes das classes pelas frequências relativas respetivas, que resultou na média procurada.

Para obter a mediana, adicionou-se na tabela de frequências uma nova coluna com as frequências relativas acumuladas de modo a identificar a classe mediana que geralmente encontra-se onde o conjunto das observações atinge os 50%. Com a classe mediana identificada calculou-se a mediana obedecendo à seguinte fórmula:

$$Li + \frac{(n/2 - \sum f)}{F_{Md}} * h$$

onde:

- Li é o limite inferior da classe mediana;
- n/2 Posição da classe mediana;
- $\sum f$ Soma das frequências anteriores a classe mediana;
- F_{Md} Frequência absoluta da classe mediana;
- h Amplitude da classe mediana.

A variância é calculada, adicionando uma nova coluna na tabela de frequências com os desvios quadráticos, somando os resultados desta nova coluna e determinando o quociente entre o somatório dos desvios quadráticos com o número total das observações subtraido de uma unidade.

Por outro lado, no R o processo é bem mais simplificado, bastando carregar o pacote "fdth", construir a tabela de distribuição de frequências e calculando-se de forma rápida e eficiente as medidas solicitadas com os comandos mean(), median(), var(), sd(). O coeficiente de variação é calculado de forma similar em ambos os programas, dividindo o desvio padrão pela média e multiplicando por 100.

Exemplo 13 *Dada a amostra de 60 rendas (em milhares) de uma dada região geográfica:*

10, 7, 8, 5, 4, 3, 2, 9, 9, 6,
3, 15, 1, 13, 14, 4, 3, 6, 6, 8,
10, 11, 12, 13, 14, 2, 15, 5, 4, 10,
2, 1, 3, 8, 10, 11, 13, 14, 15, 16,
8, 9, 5, 3, 2, 3, 3, 4, 4, 4,
5, 6, 7, 8, 9, 1, 12, 13, 14, 16

1. *Calcular:*
   (a) *A renda média, mediana e modal;*
   (b) *A variância, o desvio médio e desvio padrão;*
   (c) *O 1º e o 3º quartis;*
   (d) *Coeficiente de assimetria;*
   (e) *A curtose da distribuição.*
2. *Represente a caixa-de-bogodes.*

Resolução manual do problema

1. Seja x_i a variável representativa da renda (em milhares):
   (a) Cálculo da média, mediana e moda:
   - A média é dada por: $\bar{x} = \frac{\sum x_i}{n} = \frac{461}{60} = 7,68$.
   - Como $n = 60$ é par a mediana será a média entre os elementos de ordem $\frac{n}{2} = \frac{60}{2} = 30$ e $\frac{n}{1} + 1 = 30 + 1 = 31$, colocando os valores em ordem crescente, identifica-se os elementos da posição 30^o e 31^o, neste caso é 7 e 8. Assim: $M_e = 7,5$.
   - A moda é o valor mais frequente, para esta distribuição é $M_0 = 3$.

   (b) Cálculo da variância, desvio médio e desvio padrão:
   - A variância é dada por: $S^2 = \frac{(\sum x_i - \bar{x})^2}{n-1} = \frac{1198,983}{60-1} = 20,32175$.
   - O desvio médio é dado por: $DM = \frac{\sum |x_i - \bar{x}|}{n} = \frac{233}{60} = 3,883$.
   - O desvio padrão é dado por:$S = \sqrt{S^2} = \sqrt{20,32175} = 4,508$.

   (c) Cálculo de quartis:
   Com base na metodologia descrita anteriormente o 1^o quartil e o 3^o quartil estão respetivamente nas posições $\frac{n+3}{4} = 15,75$ e $\frac{3n+1}{4} = 45,25$. Assim, fazendo as interpolações lineares, tem-se:
   - $Q_1 = 4 + 0,75 \times (4 - 4) = 4$ e $Q_3 = 11 + 0,25 \times (12 - 11) = 11,25$.

   (d) A assimetria é dada por:
   - $a = \frac{m_3}{\sqrt{m_2^3}} = \frac{\sum (x_i - \bar{x})^3 f_i}{\sqrt{\sum [(x_i - \bar{x})^2 f_i]^3}} = \frac{\frac{1493,929}{60}}{\sqrt{\left(\frac{1198,984}{60}\right)^3}} = \frac{24,898}{89,323} = 0.27$.

(e) Coeficiente de curtose é dado por:

- $b_2 = \frac{m_4}{m_2^2} = \frac{\sum(x_i - \bar{x})^4 f_i}{\sum[(x_i - \bar{x})^2 f_i]^2} = \frac{\frac{43683,97}{60}}{\left(\frac{1198,984}{60}\right)^2} = \frac{728,066}{399,320} = 1.823.$
- $g_2 = b2 - 3 = 1,823 - 3 = -1,17.$

O diagrama de extremos e quartis ou boxplot é uma ferramenta gráfica para representar a variação de dados observados de uma variável numérica por meio de quartis. Para a sua construção é fundamental determinar:

- A amplitude interquartil é:
  $AIQ = Q_3 - Q_1 = 11,25 - 4 = 7,25.$
- A barreira interna superior:
  $Q_3 + 1,5 \times AIQ = 11,25 + 1,5 \times 7,25 = 22,125$ que por ser superior ao valor máximo da amostra será ignorada.
- A barreira interna inferior:
  $Q_1 + 1,5 \times AIQ = 4 + 1,5 \times 7,25 = -6,875$ que por ser inferior ao valor mínimo da amostra será ignorada. Conclui-se portanto, que não existem possíveis outliers.

Resolução com Excel

Esta resolução obedece aos seguintes passos:

1. Digitar os dados na folha de cálculo de A1: A60.
2. Calcular a renda média, digitando:
   = MÉDIA (A1: A60).
3. Calcular a mediana, digitando:
   = MED (A1: A60).
4. Calcular a moda, digitando:
   = MODA (A1: A60).
5. Calcular a variância, digitando:
   = VAR.S (A1: A60).
6. Calcular o 1º e 3º quartis, digitando:
   = QUARTILE.INC (A1: A60; 1);
   = QUARTILE.INC (A1: A60; 3).
7. Calcular o grau de assimetria, digitando:
   = DISTORÇÃO (A1: A60).
8. Calcular o grau de curtose, digitando:
   =KURT (A1: A60).
9. Representar a caixa de bigodes, selecionando:
   - Os dados (A1: A60);
   - Inserir gráfico estatístico;
   - Caixa & bigodes.

10. Alterar outros formatos do gráfico, selecionando os respetivos elementos, mudar o fundo, o título, etc.

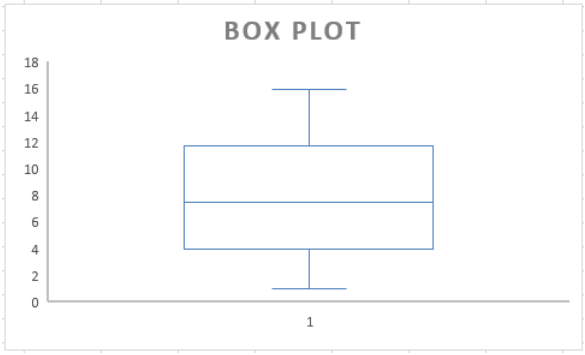


Figura 3.4: Caixa-de-bigodes Excel

Resolução com R

Esta resolução obedece os seguintes passos:

1. Digitar os dados em new script:
   rendas< −c(10, 7, 8, 5, 4, 3, 2, 9, 9, 6, 3, 15, 1, 13, 14, 4, 3, 6, 6, 8, 10, 11, 12, 13, 14, 2, 15, 5, 4, 10, 2, 1, 3, 8, 10, 11, 13, 14, 15, 16, 8, 9, 5, 3, 2, 3, 3, 4, 4, 4, 5, 6, 7, 8, 9, 1, 12, 13, 14, 16).
2. Calcular a média, digitando:
   mean(rendas).
3. Calcular a mediana, digitando:
   median(rendas).
4. Carregar o pacote (Modeest) para calcular a moda:
   mfv(rendas).
5. Calcular a variância, digitando:
   var(rendas).
6. Carregar o pacote (lsr) para calcular o desvio médio:
   aad(rendas).
7. Calcular o desvio padrão, digitando:
   sd(rendas).
8. Determinar o 1º e 3º quartis, digitando:
   quantile(rendas).
9. Carregar o pacote (e1071) para calcular a assimetria e curtose:
   skewness(rendas,type= 2).
   kurtosis(rendas,type=2).

10. Representar a caixa-de-bigodes, digitando:
    boxplot(rendas).

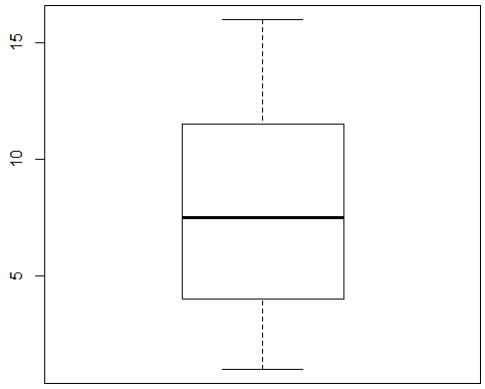


Figura 3.5: Caixa-de-bigodes R

Exemplo 14 *Considere a seguinte distribuição de frequência:* $x_i(1,2,3,4,5)$ *e* $f_i(2,2,1,4,1)$ *analise assimetria, a curtose e represente o box plot da distribuição.*

Resolução manual do problema

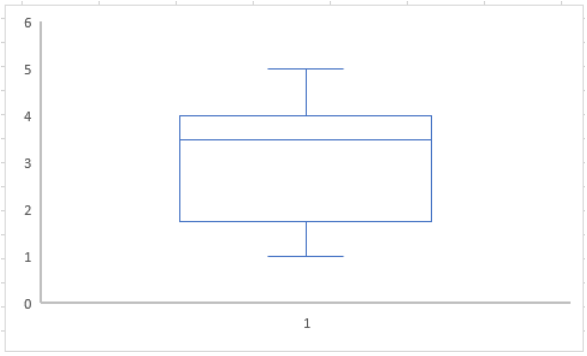
1. A assimetria é dada por:
   $a = \frac{m_3}{\sqrt{m_2^3}} = \frac{\sum(x_i-\bar{x})^3 f_i}{\sqrt{\sum[(x_i-\bar{x})^2 f_i]^3}} = -\frac{\frac{6}{10}}{\sqrt{\left(\frac{18}{10}\right)^3}} = -\frac{0.6}{2.414} = -0.25$. A distribuição é assimétrica à esquerda.
2. A curtose é dada por:
   - $b_2 = \frac{m_4}{m_2^2} = \frac{\sum(x_i-\bar{x})^4 f_i}{\sum[(x_i-\bar{x})^2 f_i]^2} = \frac{\frac{54}{10}}{\left(\frac{18}{10}\right)^2} = \frac{5{,}4}{3{,}24} = 1.67.$
   - $g_2 = b2 - 3 = 1,67 - 3 = -1.33$. A distribuição é platicúrtica.
3. A caixa de bigodes está representada na figura 3.6.

Resolução com Excel

Esta resolução obedece aos seguintes passos:

1. Digitar os valore de (x_i) na folha de cálculo de A1: A5 e os de (f_i) de B1: B5.
2. Converter os dados numa lista variando de C1: C10, isto é repetir cada valor de (x_i) em função da respetiva frequência.
3. Calcular o grau de assimetria, digitando:
   = DISTORÇÃO(C1: C10).
4. Calcular o grau de curtose, digitando:
   = KURT(C1: C10).

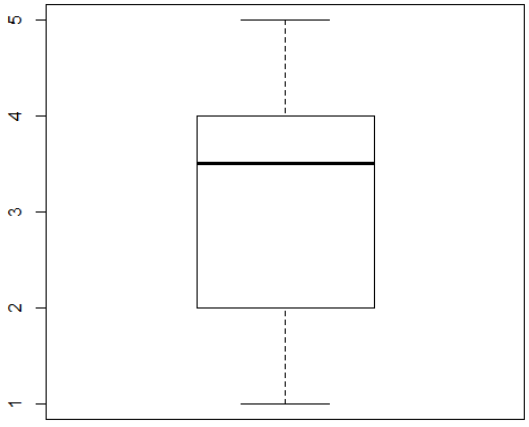
5. Representar o box plot, selecionando:
    - Os dados (C1: C10);
    - Inserir gráfico estatísticos;
    - Caixa & bigodes.



Resolução com R

Esta resolução obedece aos seguintes passos:

1. Digitar os dados em new script:
    - $x_i <-$c(1,2,3,4,5);
    - $f_i <-$c(2,2,1,4,1);
    - tb$<-$c(rep(x_i,f_i)).
2. Carregar o pacote (e1071) para calcular a assimetria e curtose:
   skewness(tb,type= 2);
   kurtosis(tb,type= 2).
3. Representar a caixa-de-bigodes, digitando:
   boxplot(tb).



No Excel, a determinação dos quartis faz-se com a função QUARTILE.INC(Matriz;Quarto), onde o primeiro argumento indica o endereço das células de que queremos calcular o quartil e o segundo argumento pode tomar diferentes valores conforme a medida de localização a determinar nomeadamente: 0 = Mínimo, 1 = 1º Quartil, 2 = Mediana, 3 = 3º Quartil e 4 = Máximo. O Excel por defeito constrói o boxplot, usando o método exclusivo. Para usar o método inclusivo, basta clicar com botão direito do mouse sobre a caixa e selecionar *formatar série de dados*, na janela que se abre selecionar incluir mediana. O pacote "Real statistic"do Excel, também constrói o boxplot e determina à parte a amplitude interquartil, a barreira interna inferior e superior, o primeiro e terceiro quartis, a mediana, o mínimo e o máximo. O R permite calcular quartis de onze maneiras diferentes: Além do método inclusivo usado no boxplot () e fivenum (), existem ainda nove tipos diferentes de quantis usados no quantile e qboxplot() do pacote MKmisc por meio da opção type. No caso do número das observações ser par o método exclusivo coincide com o de boxplot() e qboxplot(..., type= 2) e type= 5, caso contrário coincide com o de qboxplot(..., type= 6).

3.7 Distribuições bidimensionais

No processo de tomada de decisões, muita das vezes é necessário fazer previsões, para tal é necessário que exista uma relação de causa-efeito entre duas variáveis. É nesta conformidade que pretendemos abordar as distribuições bidimensionais, destacando o diagrama de dispersão, a reta de regressão e coeficiente de correlação linear.

3.7.1 Diagrama de dispersão

Obter-se-á uma melhor visão de uma distribuição bidimensional de variáveis X e Y, efetuando uma representação gráfica num sistema de eixo coordenados, marcando uma das variáveis no eixo das abcissas e a outra no eixo das ordenadas. O conjunto dos pontos obtidos constitui uma nuvem de pontos. A esta representação designam-se diagrama de dispersão, TOMÁS (2006, p.185). Na figura a seguir apresenta-se um exemplo genérico de um diagrama de dispersão.

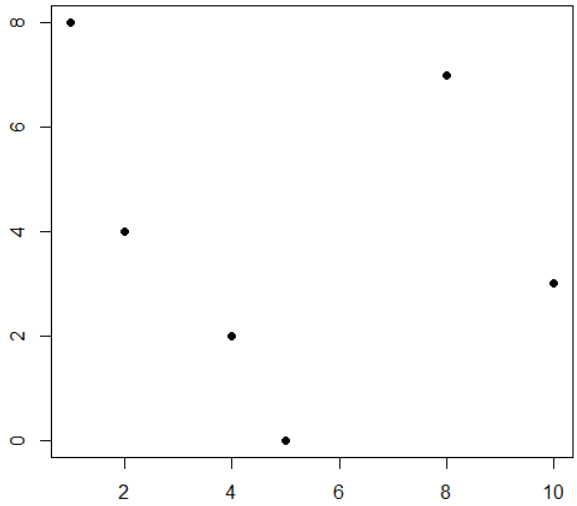


Figura 3.6: Diagrama de dispersão

3.7.2 Reta de regressão

A utilização mais importante e mais comum da reta de regressão é feita com o objetivo de prever o comportamento da variável explicada com base em valores conhecidos da variável explicativa, REIS (2012, p.179).

Se todos os pontos de um diagrama se situam nas proximidades de uma reta, esta reta chama-se reta de regressão e a correlação diz-se linear TOMAS (2006, p.186).

A equação dessa reta é dada por $y = a + bx$, onde a é conhecida como o intercepto e b é a inclinação. O método padrão para obter a melhor reta ajustada é chamado mínimos quadrados o qual literalmente minimiza a soma dos quadrados das distâncias de y_i a reta ajustada. Em princípio isto requer traçar retas possíveis, calculando a soma dos quadrados das distâncias.

$$s = \sum_{i=1}^{n} (y_i - y)^2 \tag{3.12}$$

E encontrar os valores de a e b que fornecem o menor valor de s. É possível mostrar a equação que melhor se ajuste a reta que é dada por:

$$b = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$$

$$a = \bar{y} - b\bar{x}$$

3.7.3 Coeficiente de correlação linear

O coeficiente de correlação linear é uma medida do grau de associação linear entre variáveis. Esta medida toma valores entre -1 e $+1$.

Quando se mede a correlação entre variáveis, $+1$ significa uma relação linear perfeita e positiva, enquanto que -1 é também uma relação linear perfeita mas negativa. Valores próximos de zero para o coeficiente de correlação linear indicam uma associação linear pobre entre as variáveis.

Dadas duas variáveis X e Y, quantitativas, o coeficiente de correlação linear entre X e Y é definido como:

$$\boxed{r = \frac{n \sum X_i Y_i - \sum X_i \sum Y_i}{\sqrt{[\sum X_i^2 - (\sum X_i)^2][n \sum Y_i^2 - (\sum Y_i)^2]}}} \tag{3.13}$$

A partir desta fórmula, facilmente se demonstra que o coeficiente de Pearson corresponde a um quociente entre indicadores importantes: no numerador encontra-se a covariância entre as duas variáveis e no denominador o produto dos desvios padrões de X e Y.

3.8 Exemplos de distribuições bidimensionais

Nesta secção são apresentados os procedimentos para se obter, um diagrama de dispersão, a reta de regressão e o coeficiente de correlação linear de Pearson, com auxílio do Excel e R.

Exemplo 15 *Numa turma do 3º ano do PUNIV do Maculusso, em Luanda, registaram-se a altura e o pesso de cada aluno.*

Altura(m)	1,27	1,41	1,32	1,40	1,43	1,39	1,46	1,41
Peso(kg)	27	38	24	30	34	30,5	35	33
Altura(m)	1,36	1,33	1,40	1,39	1,31	1,40	1,39	
Peso(kg)	29	29	31,5	28	32,5	33,5	32	

1. *Determine:*
   (a) *O coefiente de correlação linear de Pearson;*
   (b) *O diagrama de dispersão.*

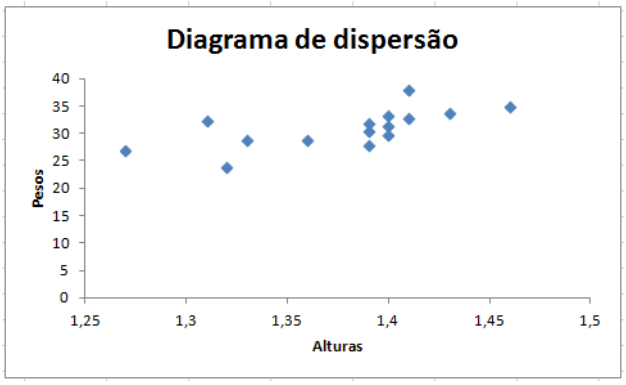
Resolução com Excel

Esta resolução obedece aos seguintes passos:

1. Digitar a altura e o peso de cada aluno na folha de cálculo de A2: A7 e de B2: B7, respetivamente.
2. Calcular o coeficiente de correlação linear, digitando em C2:
   =CORREL(A2: A7; B2: B7).
3. Construir o diagrama de dispersão, selecionando:
   - Os dados A2: A7 e B2: B7;
   - Inserir, gráfico de dispersão;
   - Dispersão apenas com marcadores.
4. Altere outros formatos do gráfico, selecionando os respetivos elementos, legenda, eixos, título, fundo, etc.

	A	B	C
1	Altura	Peso	R
2	1,27	27	0,670513
3	1,41	38	
4	1,32	24	
5	1,4	30	
6	1,43	34	
7	1,39	30,5	
8	1,46	35	
9	1,36	29	
10	1,33	29	
11	1,4	31,5	
12	1,39	28	
13	1,31	32,5	
14	1,4	33,5	
15	1,39	32	
16	1,41	33	

(a) Dados



(b) Diagrama

Resolução com R

Esta resolução obedece os seguintes passos:

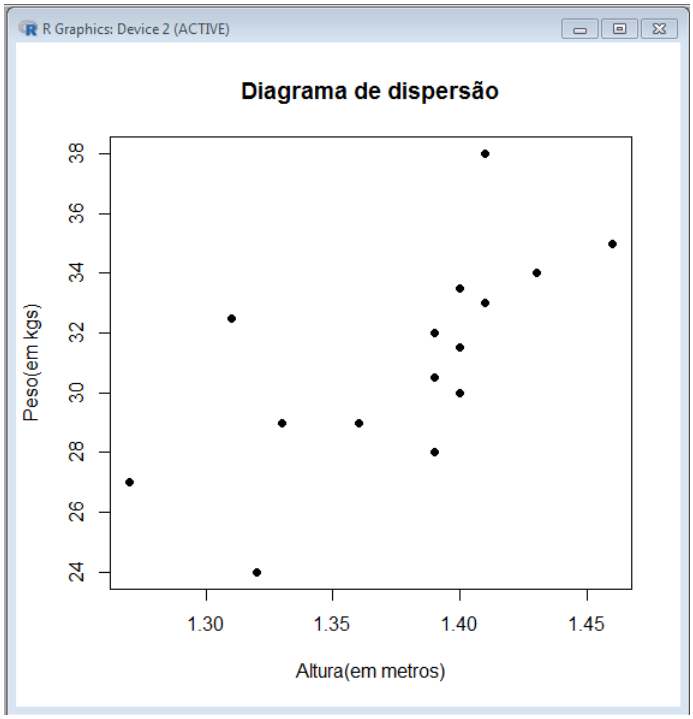
1. Digitar os dados em new script:
   - $x < -c(1.27, 1.41, 1.32, 1.40, 1.43, 1.39, 1.46, 1.36, 1.33, 1.40, 1.39, 1.31, 1.40, 1.39, 1.41)$.
   - $y < -c(27, 38, 24, 30, 34, 30.5, 35, 29, 29, 31.5, 28, 32.5, 33.5, 32, 33)$.
2. Calcular o coeficiente de correlação, digitando:

   r=cor(x, y);

   [1] 0.6705135.
3. Construir o diagrama de dispersão, digitando:

   Plot(x, y, xlab="Altura (em metros), ylab="Peso (em kgs),main="Diagrama de dispersão").

O resultado final será semelhante conforme o gráfico apresentado a abaixo:



3.8.1 Centro de gravidade de uma nuvem de pontos

Chama-se ponto médio ou centro de gravidade de uma nuvem de pontos, ao ponto, cujas coordenadas são respetivamente, a média das abcissas e a média das ordenadas dos pontos da nuvem, TOMÁS(2006, p.188).

O centro de gravidade de uma nuvem de pontos permite-nos a utilização de um processo gráfico, para decidir se existe correlação entre as duas variáveis. Para tal consideramos o centro de gravidade de uma nuvem de pontos como origem de novos eixos coordenados para os anteriores.

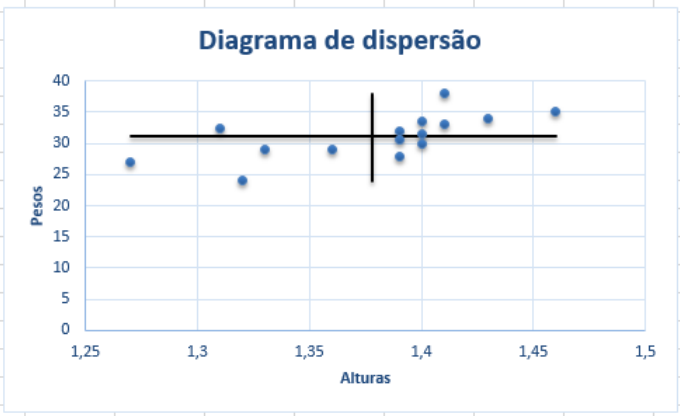
Exemplo 16 *Usando os dados do exemplo 15 determine o centro de gravidade da distribuição bidimensional das alturas e dos pesos dos alunos do 3º Ano do PUNIV, marcando este ponto na respetiva nuvem e considerando-o como origem de novos eixos coordenados.*

Resolução com Excel

Esta resolução obedece aos seguintes passos:

1. Digitar as alturas e pesos dos alunos na folha de cálculo de A2: A16 e de B2: B16 respetivamente.
2. Representar o diagrama de dispersão, procedendo como no caso anterior.
3. Determinar o mínimo e o máximo das alturas, digitando em A18 e A19:
   =MÍN(A2: A16);
   =MÁX(A2: A16).
4. Determinar a média dos pesos dos alunos, digitando em B18 e B19:
   =MÉDIA(B2: B16);
   =MÉDIA(B2: B16).
5. Determinar a reta horizontal das alturas e pesos dos alunos, selecionando:
   - O diagrama de dispersão, estrutura, selecionar dados;
   - Adicionar valor da série, nome da série (Alturas), valores da série X (A18: A19), valores da série Y (B18: B19), ok.
6. Clique com botão direito do mouse sobre um dos pontos criados no passo anterior e selecione `formatar série de dados`, na janela que se abre selecione cor da linha, linha contínua, opções de marcador, nenhum, fechar.
7. Determinar o mínimo e o máximo dos pesos dos alunos, digitando em B21 e B22:
   =MÍN(B2: B16);
   =MÁX(B2: B16).
8. Determinar a média dos pesos dos alunos, digitando em A21 e A22:
   =MÉDIA(A2: A16);
   =MÉDIA(A2: A16).
9. Determinar a reta vertical das alturas e pesos dos alunos, selecionando:
   - O diagrama de dispersão, estrutura, selecionar dados;
   - Adicionar valor da série, nome da série (Pesos), valores da série X (B21: B22), valores da série Y (A21: A22), ok.
10. Clique com botão direito do mouse sobre um dos pontos criados no passo anterior e selecione `formatar série de dados`, na janela que se abre selecione cor da linha, linha contínua, opções de marcador, nenhum, fechar.

O resultado esperado:



Resolução com R

Seja x a variável representativa das alturas e y os pesos dos alunos, esta resolução, obedece os seguintes passos:

1. Digitar as alturas e pesos dos alunos em new script:

   $x <-$ c(1.27, 1.41, 1.32, 1.40, 1.43, 1.39, 1.46, 1.36, 1.33, 1.40, 1.39, 1.31, 1.40, 1.39, 1.41).

   $y <-$ c(27, 38, 24, 30, 34, 30.5, 35, 29, 29, 31.5, 28, 32.5, 33.5, 32, 33).

2. Representar o diagrama de dispersão, digitando:

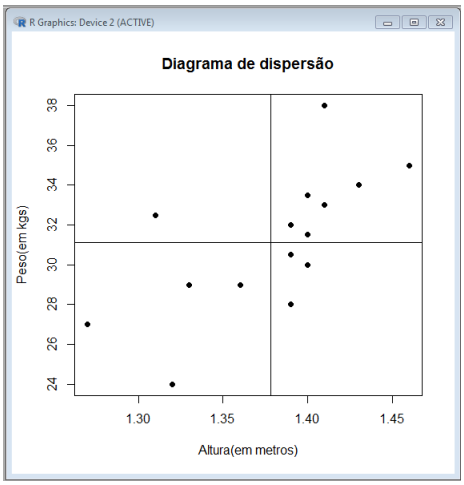
   plot(x,y,xlab="Altura(em metros)",ylab="Peso(em kgs)",pch=19,main="Diagrama de dispersão").

3. Determinar a reta horizontal e vertical respetivamente, usando os comandos:

   abline(h=mean(y));

   abline(v=mean(x)).

O resultado esperado:



A representação gráfica do ponto médio ou centro de gravidade da nuvem de pontos, no Excel é um pouco mais trabalhosa, requer várias instruções para obter este ponto, tal como foi descrito acima. Ao passo que no R, o processo é simplificado pelo simples uso dos comandos plot() e abline(). Conforme afirmamos anteriormente, o centro de gravidade de uma nuvem de pontos permite-nos a utilização de um processo gráfico, para decidir se existe correlação entre as duas variáveis. Portanto, neste exemplo relativamente aos novos eixos, verificamos que a maioria dos pontos se situam no primeiro e terceiro quadrantes, isto significa que temos uma correlação positiva.

Exemplo 17 *Considere duas séries estatísticas X e Y, tais que:*

X	1	3	4	6	8	9	11	14
Y	1	2	4	4	5	7	8	9

Determine graficamente a nuvem de pontos e a reta de regressão.

Resolução com Excel

Esta resolução obedece os seguintes passos:

1. Digitar os dados na planilha de A1: B9.
2. Representar graficamente a nuvem de pontos, selecionando:
   - Os dados A1: B9;
   - Inserir, diagrama de dispersão, dispersão apenas com marcadores.
3. Representar a reta de regressão, selecionando:
   - O gráfico de dispersão, estrutura, esquema de gráficos (Layout 9).
4. Altere outros formatos do gráfico, selecionando os respetivos elementos, legenda, eixos, título,fundo, etc.

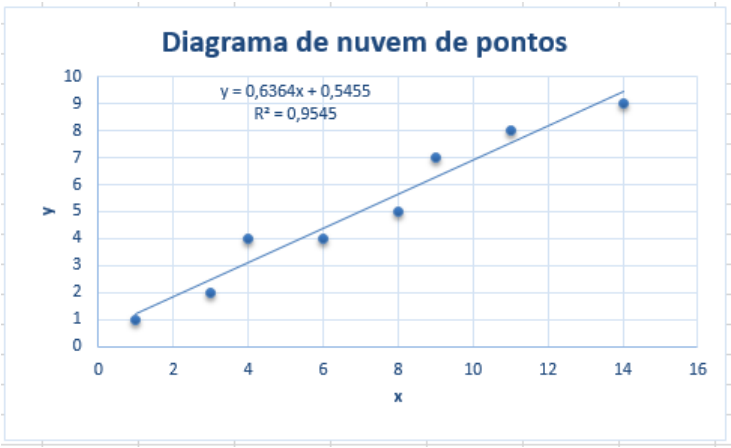


Figura 3.7: Diagrama de nuvem de pontos Excel

Resolução com R

Esta resolução obedece os seguintes passos:

1. Digitar os dados da tabela em new script:
   - $X < -c(1, 3, 4, 6, 8, 9, 11, 14)$;
   - $Y < -c(1, 2, 4, 4, 5, 7, 8, 9)$.
2. Representar graficamente a nuvem de pontos, digitando:
   - plot(X, Y, xlab="x", ylab="y", main="Digrama de nuvem de pontos", pch= 20).
3. Representar a reta de regressão, digitando:
   - R=lm(Y X);
   - abline(R).

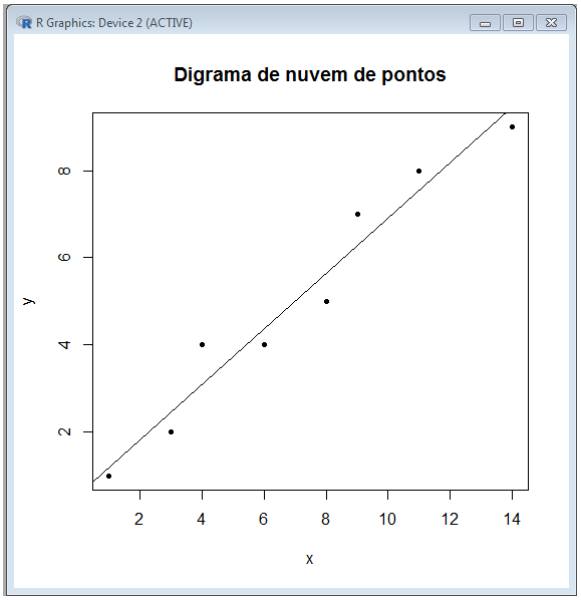


Figura 3.8: Diagrama de nuvem de pontos R

A representação gráfica dos pares das observações para X e Y, é o primeiro passo para detetar o tipo de associação existente entre elas. No eixo horizontal representa-se os valores da variável independente X, enquanto que no eixo vertical se representa os valores da variável dependente Y. Cada um dos pontos representados refere-se a um par ordenados de valores de (X, Y), constituindo o diagrama de dispersão. Este diagrama tem uma dupla função: ajuda a determinar se existe alguma relação entre as variáveis e permite identificar qual é a equação mais apropriada para descrever essa relação.

No Excel, ao representarmos a reta de regressão, usando "Layout 9", num diagrama de dispersão, identifica-se facilmente a equação e o coeficiente de determinação que descreve a respetiva relação.

O R, utiliza os comandos plot(), lm() e abline(), para representar graficamente o diagrama de dispersão e a reta de regressão. O comando lm(), também fornece os coeficientes da equação que descreve a relação entre as variáveis. No R, para obter o coeficiente de determinação, é necessário elevar ao quadrado o coeficiente de correlação que é calculado, com o comando cor().

Capítulo 4

Distribuições de probabilidades

4.1 Introdução

Neste capítulo, são apresentados os comandos e os procedimentos para calcular as principais distribuições de probabilidades de variáveis aleatórias discretas e contínuas, usando Excel e R.

4.1.1 Variável aleatória

Uma variável aleatória X é uma função que associa um número real a cada elemento do espaço de resultados. Com a, b e c $\in \mathbb{R}$, podemos calcular $P(X = a)$; $P(a < X < b)$; $P(X \leq c)$; $P(X > c)$.

4.1.2 Distribuição binomial

Se p é a probabilidade de ocorrência de um acontecimento numa única tentativa (chamada probabilidade de sucesso) e $q = 1 - p$ a probabilidade da sua não ocorrência (chamada probabilidade de insucesso), então a probabilidade de esse acontecimento ocorrer exatamente X vezes em n tentativas é dada por:

$$\boxed{p(X = x) = C_x^n p^x q^{n-x} = \frac{n!}{x!(n-x)!} p^x q^{n-x}} \tag{4.1}$$

Neste caso dizemos que a variável aleatória X tem uma distribuição binomial com parâmetros n e p, e escreve-se: $X \sim Bi(n, p)$. O valor esperado é $\mu = np$, e $\sigma^2 = npq$, é a variância da distribuição.

4.1.3 Distribuição de Poisson

Define-se a distribuição de Poisson com média λ, como:

$$\boxed{p(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}} \tag{4.2}$$

Onde:

- $X = 0, 1, 2, ...$
- $e = 2,71828...$
- λ uma constante dada.

4.1.4 Distribuição hipergeométrica

Esta distribuição está relacionada com amostragens sem reposição. Suponha-se que se tem uma população com N elementos, dos quais d pertence a uma categoria A e $N - d$ a uma outra categoria B.

Seja X a variável aleatória representativa do número de elementos da categoria A numa amostra sem reposição de n elementos da população, então:

$$\boxed{f(x) = P(X = x) = \frac{C_x^d C_{n-x}^{N-d}}{C_n^N}} \quad X \in \{0, 1, 2, ..., min(d, n)\} \tag{4.3}$$

X tem distribuição hipergeométrica com parâmetros n, d, e N e escreve-se:
$X \sim H(n, d, N)$ e $p = \frac{d}{n}$.

- Valor esperado $E(X) = np$;
- Variância $V(X) = pn(1 - p)\frac{N-n}{N-1}$.

4.1.5 Distribuição normal

Um exemplo muito importante de distribuições contínuas de probabilidade é a distribuição normal, curva normal ou distribuição gaussiana. Uma v.a contínua X tem uma distribuição normal com parâmetros μ e σ se a sua função densidade de probabilidade é:

$$\boxed{P(a < X < b) = \int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx} \tag{4.4}$$

Quando a variável X está em unidades padronizadas ($z = \frac{X-\mu}{\sigma}$) a equação 4.4, transforma-se na sua forma padronizada.

$$\boxed{f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}} \tag{4.5}$$

Neste caso diz-se que z tem uma distribuição normal com média 0 e variância 1.

4.1.6 Distribuição exponencial

Uma variável aleatória X tem distribuição exponencial, com parâmetro $\lambda > 0$, se a sua função densidade de probabilidade é definida por:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

E escreve-se $X \sim Exp(\lambda)$. Neste caso o valor esperado e o desvio padrão da distribuição coincidem.

$E(X) = \frac{1}{\lambda}$.
$V(X) = \frac{1}{\lambda^2}$.

4.1.7 Exemplos de cálculos de distribuições de probabilidades

Nesta secção são apresentados os comandos e os procedimentos de cálculo de distribuições de probabilidades, com auxílio das ferramentas Excel e R.

Exemplo 18 *Lançou-se uma moeda 6 vezes ao ar. Seja X o acontecimento saída de caras.*

1. *Determine:*
   (a) *As probabilidades;*
   (b) *Represente graficamente a lei de probabilidade correspondente;*
   (c) *A função de distribuição e o gráfico correspondente.*

Resolução com Excel

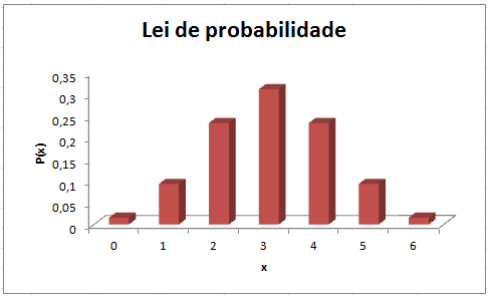
Seja X a variável aleatória discreta que representa o número de caras que ocorre em 6 lançamentos: $X \sim B(n, p)$. Esta resolução obedece os seguintes passos:

1. Digitar em C2 e D2 os valores correspondentes a n e p, isto é, $X \sim B(6, 0.5)$.
2. Digitar os possíveis valores de X de A2: A8.
3. Calcular a probabilidade de cada valor de X, digitando em B2:
   =DISTRBINOM(A2; $C$2; $D$2; 0), copiar o resultado de B2 e arrastá-lo até B8. Obtendo os seguintes resultados:

	A	B	C	D
1	x	P(x)	n	p
2	0	0,015625	6	0,5
3	1	0,09375		
4	2	0,234375		
5	3	0,3125		
6	4	0,234375		
7	5	0,09375		
8	6	0,015625		

4. Construir o gráfico de barras verticais selecionando:
   - Os valores de A2 : A8 e de B2 : B8 repetivamente;
   - Inserir gráfico de colunas ou barras;
   - Colunas agrupadas;
   - Estrutura, selecionar dados, clique em X na entradas de legenda (Série), premindo Remover;
   - Editar em Rótulos do Eixo (categoria) Horizontal, selecionando A2: A8.
5. Altere outros formatos do gráfico, selecionando os respetivos elementos, legenda, eixos, título, fundo, etc.

O resultado esperado:



6. Calcular a função de distribuição, procedendo da seguinte forma:

   - Copiar e colar os valores de X e P(X) de A12: B18 e Digitar em C12: =SOMA($B$12: B12), copiar o resultado de C12 e arrastá-lo até C18.

11	X	P(X)	F(X)
12	0	0,015625	0,015625
13	1	0,09375	0,109375
14	2	0,234375	0,34375
15	3	0,3125	0,65625
16	4	0,234375	0,890625
17	5	0,09375	0,984375
18	6	0,015625	1

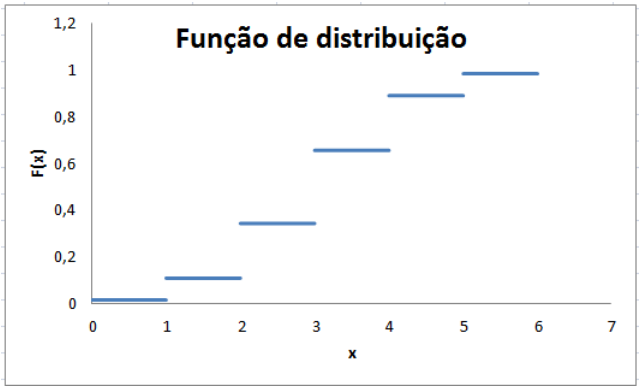
7. Para construir o gráfico da função de distribuição pode se proceder da seguinte forma:

   - Reescrever os valores de X e F(x) na folha de cálculo, conforme apresentado a seguir:

	A	B
1	X	F(X)
2	0	0
3		
4	0	0,015625
5	1	0,015625
6		
7	1	0,109375
8	2	0,109375
9		
10	2	0,34375
11	3	0,34375
12		
13	3	0,65625
14	4	0,65625
15		
16	4	0,890625
17	5	0,890625
18		
19	5	0,984375
20	6	0,984375
21		
22	6	1

- Selecionar os valores de X e $F(X)$ inseridos no passo anterior;
- Inserir, gráfico de dispersão, dispersão com linhas suaves.

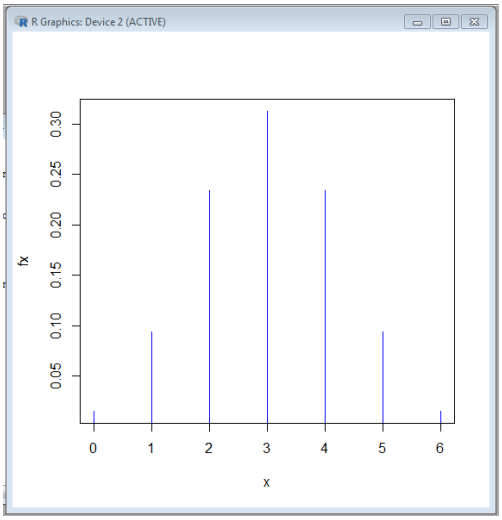
8. Altera outros formatos do gráfico selecionando os respetivos elementos eixos, título do gráfico, fundo, etc.



Resolução com R

Esta resolução obedece os seguintes passos:

1. Seja x os valores da variável aleatória correspondente, digitar em new script:
   $x < -0 : 6$;
   $n < -6$;
   $p < -0.5$.
2. Determinar as probabilidades, digitando:
   $fx <$ −dbinom(x, n, p).
3. Representar graficamente a lei de probabilidade, digitando:
   plot(x, fx, type="h", col="blue").



4. Calcular a função de distribuição, digitando:
   $Fx <-$ pbinom(x, n, p).

X	0	1	2	3	4	5	6
F(X)	0.015625	0.109375	0.343750	0.656250	0.890625	0.984375	1.000.000

5. Construir o gráfico correspondente, digitando:
   cdf=c(0, Fx);
   cdfp=stepfun(x, cdf, f=0);
   cdfp=stepfun(x, cdf, f=0);
   plot.stepfun(cdfp,verticals=F, pch=16, xaxt="n", main="Função de distribuição");
   axis(side=1,at=x).
   O resultado esperado:

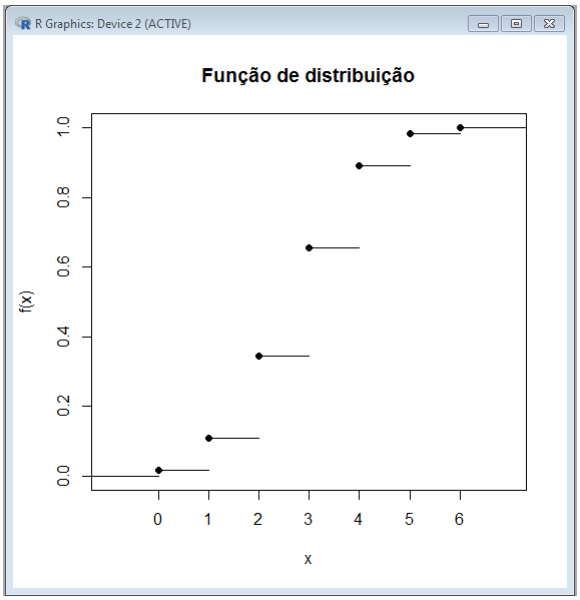


Figura 4.1: Função de distribuição

A função de distribuição é uma função definida para todo o valor real x, e que para cada x dá a soma das frequências dos valores da amostra menores ou iguais a x. Quando temos uma variável do tipo discreto, a função de distribuição é uma função em escada, isto é, uma função que cresce por degraus, mudando de degrau nos pontos em que a frequência é diferente de zero e em que a altura do degrau é igual à frequência respetiva.[1] O Excel não dispõe de uma representação imediata para esta função, portanto, é necessário utilizar um pequeno artifício, conforme descrito acima. Por sua vez, o R utiliza o comando plot(), para representar graficamente esta função.

[1]Disponível em: `http://www.dren.min-edu.pt/Alea/images/recursos/DossiesDidacticos/pdf/dossier13_cap2.pdf` consultado à 12/06/2017

Exemplo 19 *Considere um jogo no qual uma moeda não viciada é lançada três vezes, detrmine a probabilidade de saírem duas caras, pelo menos uma cara, no máximo duas caras.*

Resolução do problema

Seja X a variável aleatória discreta que representa o número de caras que ocorre em 3 lançamentos: $X \sim B(n,p)$ ou seja $X \sim B(3, 0.5)$.

1. A probabilidade de saírem duas caras é dada por:
   - $P(X=2) = C_2^3 \times \left(\frac{1}{2}\right)^2 \times \left(\frac{1}{2}\right)^1 = 0,375.$
2. Função do Excel para o cálculo da probabilidade pedida:
   $\boxed{= DISTRBINOM(2;3;0,5;FALSO)}$
3. Comando do R para o cálculo da probabilidade pedida:
   $\boxed{dbinom(2,3,0.5)}$
4. A probabilidade de saírem pelo menos uma cara é dada por:
   - $P(X \geq 1) = 1 - P(X < 1) = 1 - P(X = 0) = 0,875.$
5. Função do Excel para o cálculo da probabilidade pedida:
   $\boxed{= 1 - DISTRBINOM(0;3;0,5;VERDADEIRO)}$
6. Comando do R para o cálculo da probabilidade pedida:
   $\boxed{1 - pbinom(0,3,0.5)}$
7. A probabilidade de saírem no máximo duas caras é dada por:
   - $P(X \leq 2) = P(1) + P(2) + P(3) = 0,875.$
8. Função do Excel para o cálculo da probabilidade pedida:
   $\boxed{= DISTRBINOM(2;3;0,5;VERDADEIRO)}$
9. Comando do R para o cálculo da probabilidade pedida:
   $\boxed{pbinom(2,3,0.5)}$

Exemplo 20 *Um assistente recebe chamadas telefónicas em média de 120 por hora. Supondo que o número de chamadas é uma variável aleatória com distribuição de Poisson.*

1. *Determine a probabilidade de que:*
   - *(a) Seja recebida uma chamada durante um dado periodo;*
   - *(b) Seja recebida pelo menos duas chamadas durante esse periodo;*
   - *(c) Seja recebida mais de 6 e menos de 10 chamadas em 8 minutos;*
   - *(d) Não chegue qualquer chamada durante três minutos.*

Resolução do problema

X Segue uma distribuição de Poisson de parâmetro λ, isto é, $X \sim P(\lambda)$ onde $X \sim P(2)$.

1. A probabilidade de que seja recebida uma chamada durante um dado periodo é dada por:
   - $P(X = 1) = \frac{2^1}{1!}e^{-2} = 0.2706$.
2. Função do Excel para o cálculo da propabilidade pedida:
   $\boxed{= POISSON(1; 2; FALSO)}$
3. Comando do R para o cálculo da propabilidade pedida:
   $\boxed{dpois(1, 2)}$
4. A probabilidade de que seja recebida pelo menos duas chamadas durante esse periodo é dada por:
   - $P(X \geq 2) = 1 - P(X \leq 2) = 1 - [P(X = 0) + P(X = 1)] = 0.593$.
5. Função do Excel para o cálculo da probabilidade pedida:
   $\boxed{= 1 - POISSON(1; 2; VERDADEIRO)}$
6. Comando do R para o cálculo da probabilidade pedida:
   $\boxed{1 - ppois(1, 2)}$
7. A probabilidade de que seja recebida mais de 6 e menos de 10 chamadas em 8 minutos é dada por:
   - $P(6 < X < 10) = P(X \leq 9) - P(X \leq 6) = 0.039$ onde $\lambda = 2 \times 8$.
8. Função do Excel para o cálculo da probabilidade pedida:
   $\boxed{= POISSON(9; 16; VERDADEIRO) - POISSON(6; 16; VERDADEIRO)}$
9. Comando do R para o cálculo da probabilidade pedida:
   $\boxed{ppois(9, 16) - ppois(6, 16)}$
10. A probabilidade de que não chegue qualquer chamada durante três minutos é dada por:
    - $P(X = 0) = \frac{6^0}{0!}e^{-6} = 0.0024$ onde $\lambda = 2 \times 3$.
11. Função do Excel para o cálculo da probabilidade pedida:
    $\boxed{= POISSON(0; 6; FALSO)}$
12. Comando do R para o cálculo da probabilidade pedida:
    $\boxed{dpois(0, 6)}$

Exemplo 21 *Considere-se uma população de 100 computadores dos quais 10 sofrem de determinada a varia. Escolhida aleatoriamente, sem reposição uma amostra de 5 computadores, qual a probabilidade de nenhum deles estar avariado?*

Resolução do problema

Seja X a variável aleatória discreta que representa o número de computadores avariados existentes na amostra, $X \sim H(5, 10, 100)$.

1. A probabilidade de que nenhum deles esteja avriado é dada por:
   - $P(X = 0) = \frac{C_0^{10} * C_5^{90}}{C_5^{100}} = 0,584.$
2. Função do Excel para o cálculo da probabilidade pedida, digitando:
   $\boxed{= DIST.HIPRGEOM(0; 5; 10; 100; FALSO)}$
3. Comando do R para o cálculo da probabilidade pedida:
   $\boxed{dhyper(0, 10, 90, 5)}$

Exemplo 22 *Numa agência de viagens e turismo com 100 clientes, dos quais 15 estão insatisfeitos com os serviços da agência, seleciona-se aleatoriamente e sem reposição uma amostra de 30 clientes.*

1. *Calcular a probabilidade de se encontrarem nessa amostra:*
   (a) *Seis clientes insatisfeitos;*
   (b) *No máximo três clientes insatisfeitos;*
   (c) *Pelo menos dois clientes insatisfeitos.*

Resolução do problema

Seja X a variável aleatória discreta que representa o número de clientes insatisfeitos que existem numa amostra $X \sim H(30, 15, 100)$.

1. A probabilidade de encontrarem 6 clientes insatisfeitos numa amostra de 30 clientes é dada por:
   - $P(X = 6) = \frac{C_6^{15} * C_{24}^{85}}{C_{30}^{100}} = 0.1524.$
2. Função do Excel para o cálculo da probabilidade pedida, digitando:
   $\boxed{= DIST.HIPRGEOM(6; 30; 15; 100; FALSO)}$
3. Comando do R para o cálculo da probabilidade pedida:
   $\boxed{dhyper(6, 15, 85, 30)}$
4. A probabilidade de se encontrarem no máximo três clientes insatisfeitos nesta amostra é dada por:
   - $P(X \leq 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(x = 3) = 0.2777.$
5. Função do Excel para o cálculo da probabilidade pedida:
   $\boxed{= DIST.HIPGEOM(3; 30; 15; 100; VERDADEIRO)}$

6. Comando do R para o cálculo da probabilidade pedida:

   $\boxed{phyper(3, 15, 85, 30)}$

7. A probabilidade de se encontrarem pelo menos dois clientes insatisfeitos é dada por:

   - $P(X \geq 2) = 1 - P(X \geq 2) = 1 - [P(X = 0) + P(X = 1)] = 0.9742.$

8. Função do Excel para o cálculo da probabilidade pedida:

   $\boxed{= 1 - DIST.HIPGEOM(1; 30; 15; 100; VERDADEIRO)}$

9. Comando do R para o cálculo da probabilidade pedida:

   $\boxed{1 - phyper(1, 15, 85, 30)}$

Exemplo 23 *As alturas dos alunos de determinada escola normalmente distribuidas com média 1,60 m e desvio-padrão 0,30 m.*

1. *Encontre a probabilidade de um aluno medir:*
   - *(a) Entre 1,50 e 1,80 m;*
   - *(b) Mais de 1,75 m;*
   - *(c) Menos de 1,48 m.*

Resolução do problema

Seja X a variável aleatória representativa das alturas dos alunos, $X \sim N(1,60, 0,30^2)$ onde $\mu = 1,60$ e $\sigma = 0,30$, então:

1. A probabilidade de um aluno medir entre 1,50 e 1,80 m é dada por:

   - $P(1,50 \leq X \leq 1,80) = P(Z_1 \leq Z \leq Z_2)$, onde:

     $z_1 = \frac{X-\mu}{\sigma} = \frac{1,50-1,60}{0,30} = -0,33$ e $z_2 = \frac{1,80-1,60}{0,30} = 0,67.$

     Para se obter a probabilidade, basta entrar com abscissa 0,3 (na primeira coluna) e 0,03 (na primeira linha) da tabela, na interseção encontra-se o valor (0.1293). Analogamente, entrando com abscissa 0,6 (na primeira coluna) e 0,07 (na primeira linha), encontrando (0.2486). Assim:

     $P(-0,33 \leq Z \leq 0,67) = 0,1293 + 0,2486 = 0,3779.$

2. Função do Excel para o cálculo da probabilidade pedida:

   $\boxed{= DIST.NORM(1,80; 1,60; 0,30; 1) - DIST.NORM(1,50; 1,60; 0,30; 1)}$ ou
   $\boxed{= DIST.NORM(0,67; 0; 1; 1) - DIST.NORM(-0,33; 0; 1; 1)}$

3. Comando do R para o cálculo da probabilidade pedida:

   $\boxed{pnorm(1.80, 1.60, 0.30) - pnorm(1.50, 1.60, 0.30)}$ ou
   $\boxed{pnorm(0.67, 0, 1) - pnorm(-0.33, 0, 1)}$

4. A probabilidade de um aluno medir mais de 1,75 m é dada por:

   - $P(X > 1,75) = P(Z > Z_1) = P(Z > 0,5) = 0,5000 - 0,1915 = 0,3085$, onde:

     $Z_1 = \frac{1,75-1,60}{0,30} = 0,50$.

5. Função do Excel para o cálculo da probabilidade pedida:

   $\boxed{= 1 - DIST.NORM(1,75; 1,60; 0,30; 1) \quad ou \quad = 1 - DIST.NORM(0,5; 0; 1; 1)}$

6. Comando do R para o cálculo da probabilidade pedida:

   $\boxed{1 - pnorm(1.75, 1.60, 0.30) \quad ou \quad 1 - pnorm(0.5, 0, 1)}$

7. A probabilidade de um aluno medir menos de 1,48 m é dada por:

   - $P(X < 1,48) = P(Z < Z_1) = P(Z < -0,4) = 0,5000 - 0,1554 = 0,3446$, onde: $Z_1 = \frac{1,48-1,60}{0,30} = -0,4$.

8. Função do Excel para o cálculo da probabilidade pedida:

   $\boxed{= DIST.NORM(1,48; 1,60; 0,30; 1) \quad ou \quad = 1 - DIST.NORM(0,4; 0; 1; 1)}$

9. Comando do R para o cálculo da probabilidade pedida:

   $\boxed{pnorm(1.48, 1.60, 0.30) \quad ou \quad pnorm(-0.4, 0, 1)}$

Exemplo 24 *A duração de um certo componente eletrônico segue uma distribuição normal de média 10 anos e desvio-padrão 3,5 anos. Calcular a probabilidade desse componente não durar mais de 7 anos. E qual é a probabilidade desse componente durar pelo menos 12 anos?*

Resolução do problema

Seja X a variável aleatória representativa da duração de um certo componente eletrônico, $X \sim N(10, 3, 5^2) \Rightarrow Z \sim N(0, 1)$.

1. A probabilidade desse componente não durar mais de 7 anos é dada por:

   - $P(X \leq 7) = P(Z \leq Z_1) = P(Z \leq -0,85)$, onde $Z_1 = \frac{X-\mu}{\sigma} = \frac{7-10}{3,5} = -0,85$

     $P(X \leq 7) = P(Z \leq -0,85) = P(X \geq 0,85) = 0,5 - 0,3023 = 0,1977$.

2. Função do Excel para o cálculo da probabilidade pedida:

   $\boxed{= DIST.NORM(7; 10; 3,5; 1) \quad ou \quad = 1 - DIST.NORM(0,85; 0; 1; 1)}$

3. Comando do R para o cálculo da probabilidade pedida:

   $\boxed{pnorm(7, 10, 3.5) \quad ou \quad 1 - pnorm(0.85, 0, 1)}$

4. A probabilidade desse componente durar pelo menos 12 anos é dada por:

   - $P(X \geq 12) = P(Z \geq Z_1) = P(Z \geq 0,57)$, onde $Z_1 = \frac{X-\mu}{\sigma} = \frac{12-10}{3,5} = 0,57$

     $P(X \geq 12) = P(Z \geq 0,57) = P(Z \leq 0,57) = 0,5 - 0,2157 = 0,2843$.

5. Função do Excel para o cálculo da probabilidade pedida:

$$\boxed{= 1 - DIST.NORM(12; 10; 3,5; 1) \quad ou \quad = 1 - DIST.NORM(0,57; 0; 1; 1)}$$

6. Comando do R para o cálculo da probabilidade pedida:

$$\boxed{1 - pnorm(12, 10, 3.5) \quad ou \quad 1 - penorm(0.57, 0, 1)}$$

Exemplo 25 *A resistência à compreensão de amostras de cimento de um certo tipo é uma variável aleatória que pode ser modelada por uma distribuição normal com média* $6000kg/cm^2$ *e desvio padrão* $100kg/cm^2$*. Qual é a probabilidade para que uma amostra de cimento tenha resistência superior à* $6150kg/cm^2$*?*

Resolução do problema

Seja X a variável aleatória representativa da amostra do cimento, $X \sim N(600, 100^2) \Rightarrow$
$Z \sim N(0, 1)$.

1. A probabilidade para que uma amostra de cimento tenha resistência superior à $6150kg/cm^2$ é dada por:
   - $P(X > 6150) = P\left(\frac{X-6000}{100} > \frac{6150-6000}{100}\right) =;$
   - $P(Z > 1,5) = 1 - P(Z > 1,5) = 1 - 0,9332 = 0,0668.$
2. Função do Excel para o cálculo da probabilidade pedida:

$$\boxed{= 1 - DIST.NORM(6150; 6000; 100; 1) \quad ou \quad = 1 - DIST.NORM(1,5; 0; 1; 1)}$$

3. Comando do R para o cálculo da probabilidade pedida:

$$\boxed{1 - pnorm(6150, 6000, 100) \quad ou \quad 1 - penorm(1.5, 0, 1)}$$

Exemplo 26 *Represente graficamente uma variável aleatória Z com uma distribuição normal padronizada ou seja, com média igual a* 0 *e variância igual a* 1*.*

Resolução com Excel

Esta resolução, obedece aos seguintes passos:

1. Digitar em A2 e B2 os valores da média e variância.
2. Digitar o possível valor mínimo de Z em A4.
3. Determinar o intervalo de variação dos valores de Z, digitando em C2:
   $= (A2 - A4)/3.$
4. Definir os valores de Z, entre -3 e 3, digitando em A5:
   $= A4 + \$C\$2$, copiar o resultado de A5 e arrastá-lo até A10.
5. Calcular as probabilidades de valores de Z, digitando em B4:
   =DIST.NORM(A4; $\$A\$2$; $\$B\$2$; FALSO), copiar o B4 e arrastá-lo até B10.

Obtendo os seguintes resultados:

	A	B	C
1	Média	Variância	Intervalo
2	0	1	1
3	X	P(X)	
4	-3	0,004432	
5	-2	0,053991	
6	-1	0,241971	
7	0	0,398942	
8	1	0,241971	
9	2	0,053991	
10	3	0,004432	

6. Construir o gráfico de distribuição normal, selecionando:
   - Os valores de Z e as respetivas probabilidades;
   - Inserir linha em 2D, remover a linha da primeira série.
7. Substituir os valores apresentados no eixo horizontal do gráfico, pelos respetivos valores de Z, selecionando:
   - Estrutura, selecionar dados;
   - Editar rótulos do eixo (categoria) horizontal;
   - Selecionar a coluna de valores de Z ou seja (A4 : A10), ok.
8. Alterar o formato do gráfico, selecionando:
   - A curva, farmater seleção;
   - Estilo da linha, Suavizada.
9. Altere outros formatos do gráfico, selecionando os respetivos elementos, legenda, fundo, etc.

O resultado esperado:

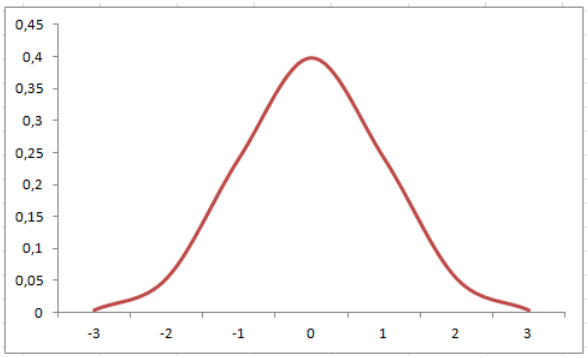


Figura 4.2: Distribuição normal Excel

Resolução com R

Esta resolução, obedece:

1. Digitar aos seguintes dados em new script:
   - $x < -seq(-3, 3, .01)$;
   - $hx < -dnorm(x, 0, 1)$.
2. Construir o gráfico de distribuição normal, digitando:
   plot(x, hx, type="l", lwd= 2).

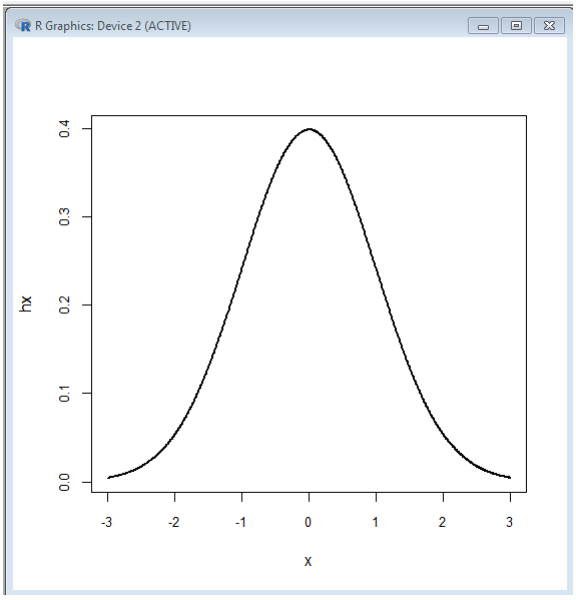


Figura 4.3: Distribuição normal R

Exemplo 27 *Os defeitos de um tecido seguem a distribuição de Poisson com média de um defeito a cada 400 m.*

1. *Qual a probabilidade de que o intervalo entre dois defeitos consecutivos seja:*
   (a) No mínimo de 1000 m;
   (b) Entre 800 e 1000 m.

Resolução do problema

Sabe-se que numa distribuição de Poisson, $X \sim Exp(\lambda)$, o valor esperado é $E(X) = \frac{1}{\lambda}$, então, $\lambda = \frac{1}{400}$.

1. A probabilidade de que o intervalo entre dois defeitos seja no mínimo de $1000m$ é dada por:
   - $P(X \geq 1000) = e^{-\lambda x} = e^{-\frac{1000}{400}} = e^{-2,5} = 0,0821$.

2. Função do Excel para o cáculo da probabilidade pedida:

   $\boxed{= 1 - DISTEXPON(1000; 1/400; VERDADEIRO)}$

3. Comando do R para o cáculo da probabilidade pedida:

   $\boxed{1 - pexp(1000, 1/400)}$

4. A probabilidade de que o intervalo entre dois defeitos consecutivos seja entre 800 e $1000m$ é dada por:

   - $P(800 \leq X \leq 1000) = P(X \geq 800) - P(X \geq 1000) = e^{-\frac{800}{400}} - e^{-\frac{1000}{400}} = 0,0532.$

5. Função do Excel para o cálculo da probabilidade pedida:

   $\boxed{= DISTEXPON(1000; 1/400; 1) - DISTEXPON(800; 1/400; 1)}$

6. Comando do R para o cálculo da probabilidade pedida:

   $\boxed{pexp(1000, 1/400) - pexp(800, 1/400)}$

Exemplo 28 *Suponha-se que o tempo de vida útil de um componente eletrônico é uma variável aleatória X, com distribuição exponencial com média igual à* 600 *horas. Qual é a probabilidade do componente durar mais de* 700 *horas?*

Resolução do problema

Seja X a variável aleatória representativa do tempo de vida útil do componente eletrônico, $X \sim Exp(\lambda)$.

1. A probabilidade do componente durar mais de 700 horas é, portanto:

   $P(X > 700) = \int_{700}^{+\infty} \frac{1}{600} e^{-\frac{x}{600}} dx = e^{-7/6} = 0,31.$

2. Função do Excel para o cálculo da probabilidade pedida:

   $\boxed{= 1 - DISTEXPON(700; 1/600; VERDADEIRO)}$

3. Comando do R para o cáculo da probabilidade pedida:

   $\boxed{1 - pexp(700, 1/600)}$

As distribuições de probabilidades para variáveis aleatórias discretas e contínuas são calculadas de forma bastante simples, tanto no Excel como no R. Ambos os programas possuem funções e comandos que facilitam estes cálculos. No Excel, as funções como: DISTBINOM(), POISSON(), DIST.HIPRGEOM(), DIST.NORM() e DISTEXPON(), destacam-se no cálculo destas distribuições. O R, utiliza os comandos dbinom(), pbinom(), dpois(), ppois(), dhyper(), phyper(), dnorm(), pnorm(), dexp() e pexp(), para calcular estas distribuições.

A representação gráfica de uma distribuição normal não é tão imediata no Excel, requer várias instruções para obter este gráfico, tal como foi descrito no exemplo 26. Ao passo que, o R utiliza o comando plot(..., type="l") para representar graficamente esta distribuição de forma rápida e eficiente.

4.2 Testes de hipóteses

O procedimento para realização dos testes de hipóteses é resumido nos seguintes passos:

1. Enunciar as hipóteses H_0 e H_1;
2. Fixar o nível de significância α, e identificar a estatística do teste;
3. Com auxílio das tabelas estatísticas, considerando α e a estat istica do teste, determinar RC (região crítica) e RA (região de não rejeição) para H_0;
4. Com os elementos amostrais, calcular o valor da variável do teste;
5. Concluir pela rejeição ou não rejeição de H_0 pela comparação do valor obtido no 4º passo com RA e RC.

4.2.1 Teste de hipóteses para médias

O teste de hipóteses para médias, obedece aos seguintes passos:

1. $H_0 : \mu = \mu_0$.
   H_1: uma das alternativas:
   (a) $\mu \neq \mu_0$;
   (b) $\mu > \mu_0$;
   (c) $\mu < \mu_0$.
2. Fixar α, admitindo-se que σ^2 é desconhecida, a estatística do teste será "t"de student, como $\phi = (n-1)$;
3. Com auxílio da tabela "t"determina-se RA e RC;
4. Cálculo do valor da variável:
   $t_{cal} = \frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}}$, onde:
   - $\bar{x}$ = média amostral;
   - μ_0 = valor da hipótese nula;
   - S = desvio padrão amostral;
   - n = tamanho da amostra;
5. Conclusões:
   (a) Se $-t\frac{\alpha}{2} \leq t_{cal} \leq t\frac{\alpha}{2}$ não se pode rejeitar H_0;
   Se $t_{cal} > t\frac{\alpha}{2}$ ou $t_{cal} < -t\frac{\alpha}{2}$, rejeita-se H_0;
   (b) Se $t_{cal} < t\alpha$, não se pode rejeitar H_0;
   Se $t_{cal} > t\alpha$, rejeita-se H_0;
   (c) Se $t_{cal} > t\alpha$, não se pode rejeitar H_0;
   Se $t_{cal} < t\alpha$, rejeita-se H_0.

Exemplo 29 *Os dois registos dos últimos aos de um colégio, atestam para os caloiros admitidos uma nota média* 115 *(teste vocacional). Para testar a hipótese de que a média de uma nova turma é a mesma, tirou-se ao acaso, uma amostra de* 20 *notas, obtendo-se média* 118 *e desvio-padrão* 20. *Admitir que* $\alpha = 0,05$*, para efetuar o teste.*

Resolução manual do problema

Assumindo que as notas seguem uma distribuição normal, a resolução obedece aos seguintes passos:

1. Enunciar as hipóteses:
   - $H_0 : \mu = 115$;
   - $H_1 : \mu \neq 115$.
2. Fixar $\alpha = 0,05$, Variável "t"com 19 graus de liberdade;
3. Com auxílio da tabela de distribuição t de student, considerando α e a estatística do teste determinou-se RC e RA para H_0;
4. Cálculo do valor da estatística: $t_{cal} = \frac{115-118}{\frac{20}{\sqrt{20}}} = -0,67$.
5. Conclusão:

   Como $-2,0930 \leq t_{cal} \leq 2,0930$, não se pode rejeitar $H_0 : \mu = 115$ com esse nível de significância.

Resolução com Excel

Esta resolução obdece aos seguintes passos:

1. Digitar os registos na folha de cálculo de A1: E2
2. Determinar o grau de liberdade, digitando em $F2$:

   $= C2 - 1$.
3. Determinar RA e RC, digitando em $G2$:

   =INVT(E2; F2).
4. Calcular o valor da estatística t, digitando em $H2$:

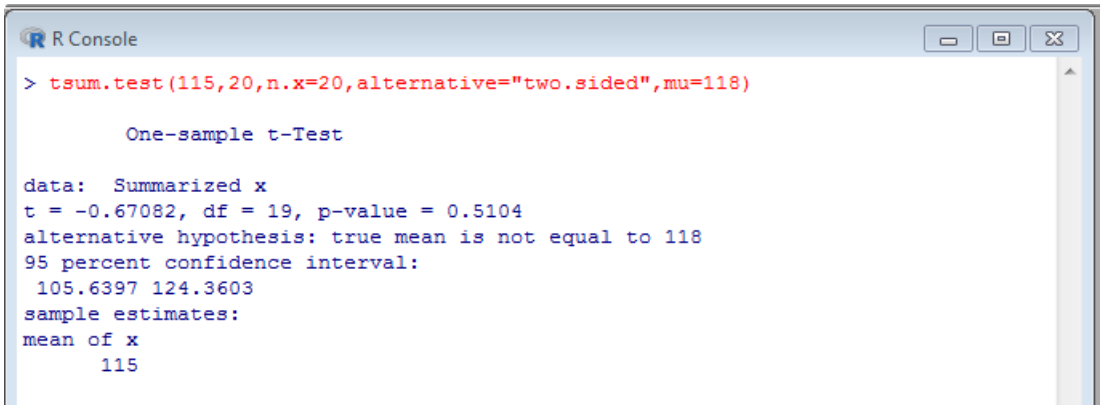
   =(A2−B2)/(D2/RAIZQ(C2)).

	A	B	C	D	E	F	G	H
1	Média1	Média2	Amostra	Desvio padrão	Significância	GL	RA e RC	tcal
2	115	118	20	20	0,05	19	2,093024	-0,67082

Resolução com R

Esta resolução, obedece aos seguintes passos:

1. Carregar o pacote BSDA.
2. Calcular o valor de t, digitando:
   tsum.test(115, 20, n.x= 20, alternative="two.sided",mu=118).



4.2.2 Teste de hipóteses para igualdade de duas médias

As variâncias populacionais são desconhecidas e variáveis admitidas iguais, independentes e normais.

1. H_0:$\mu_1 = \mu_2$ ou $\mu_1 - \mu_2 = d$
   H_1:$\mu_1 \neq \mu_2$ ou $\mu_1 - \mu_2 \neq d$;
2. Fixar α, escolher a variável "t"com $\phi = (n_1 + n_2 - 2)$;
3. Com auxílio da tabela da distribuição "t", determinam-se RA e RC.
4. Cálculo do valor de estatística t:
   - $t_{cal} = \frac{(\bar{x_1} - \bar{x_2} - d)}{Sc \times \sqrt{\frac{n_1+n_2}{n_1 \times n_2}}}$, onde:
   - S_c = desvio padrão comum que é dado por:
     $S_c = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}$.
5. Conclusões:
   - Se $-t\frac{\alpha}{2} \leq t_{cal} \leq t\frac{\alpha}{2}$ não se pode rejeitar H_0;
   - Se $t_{cal} > t\frac{\alpha}{2}$ ou $t_{cal} < -t\frac{\alpha}{2}$, rejeita-se H_0.

Exemplo 30 *Dois tipos de tinta foram testados sob as mesmas condições meteorológicas. O tipo A registou uma média de 80 com um desvio de 5 em uma dimensão de 5 partes. O tipo B, uma média de 83 com um desvio de 4 em uma dimensão de 6 partes. Adotando* $\alpha = 0,05$ *testar a hipótese da igualdade das médias.*

Resolução manual do problema

Esta resolução, obedece aos seguintes passos:

1. Enunciar as hipóteses:
   - $H_0 : \mu_1 = \mu_2$;
   - $H_1 : \mu_1 \neq \mu_2$.
2. Fixar $\alpha = 0,05$ e a estatística t com $5 + 6 - 2 = 9$ graus de liberdade;
3. Com auxílio da tabela calculou-se RA e RC (-2,2622 e 2,2622);
4. Calcular o desvio padrão comum.
   - $Sc = \sqrt{\frac{(5-1)5^2+(6-1)4^2}{5+6-2}} = 4,47$.
5. Calcular o valor da estatística do test:
   - $t_{cal} = \frac{80-83}{4,47\sqrt{\frac{5+6}{30}}} = -1,11$.
6. Conclusões:
   - Como $-2,2622 \leq t_{cal} \leq 2,2622$, não se pode rejeitar H_0 com esse nível de significância.

Resolução com Excel

Esta resolução, obedece aos seguintes passos:

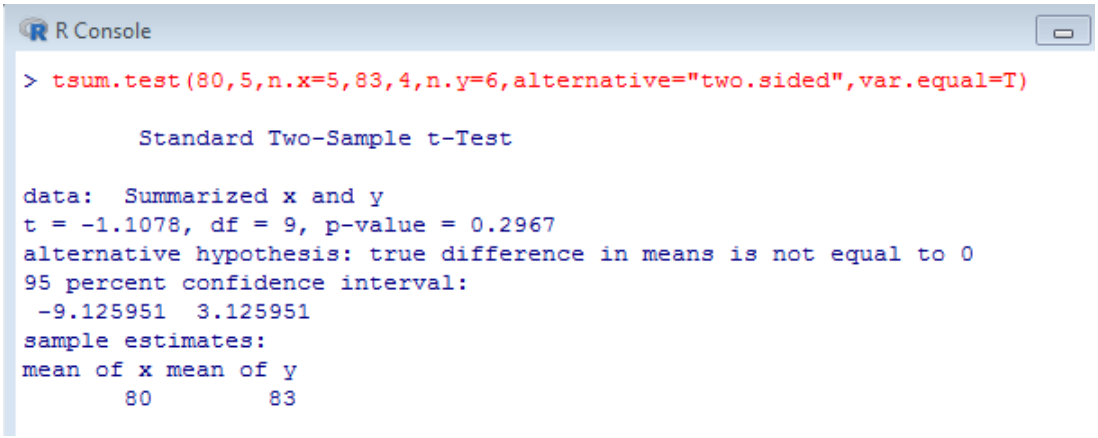
1. Digitar os registos de cada tipo na folha de cálculo de A2: G2.
2. Determinar o grau de liberdade, digitando em $B3$:
   $= C2 + F2 - 2$.
3. Determinar RA e RC, digitando em $D3$:
   $= INVET(G2; B3)$.
4. Calcular o desvio padrão comum de A e B, digitando em $F3$:
   $= RAIZQ(((C2 - 1) * B2^2 + (F2 - 1) * E2^2)/B3)$.
5. Calcular o valor da variável t, digitando em $H3$:
   $= ((A2 - D2)/(F3 * RAIZQ((C2 + F2)/(C2 * F2))))$.

	A	B	C	D	E	F	G	H
1	Média A	Desvio padrão A	Amostra A	Média B	Desvio padrão B	Amostra B	Nível de significância	
2	80	5	5	83	4	6	0,05	
3	Gl	9	RA e RC	2,2622	SC(AB)	4,472	Valor da variável t	-1,10782

Resolução com R

Esta resolução, obedece os seguintes passos:

1. Carregar o pacote "BSDA";
2. Calcular o valor da variável, digitando:
   tsum.test(80,5,n.x=5,83,4,n.y=6,alternative="two.sided",var.equal=T)



Exemplo 31 *Na tabela abaixo estão registados os índices de vendas em 6 supermercados para os produtos concorrentes da marca A e marca B. Testar a hipótese de que a diferença das médias no índice de vendas entre as marcas é zero. Sendo* $\alpha = 5\%$.

Supermercado	Marca A	Marca B
1	14	4
2	20	16
3	2	28
4	11	9
5	5	31
6	12	10

Resolução com Excel

Esta resolução, obedece aos seguintes passos:

1. Digitar os produtos concorrentes da marca A e marca B, de cada supermercado na folha de cálculo de A1: B8;
2. Calcular o valor da variável t_{cal}, selecionando:
   - Dados, Análise de dados;
   - Test T: Duas amostras com variâncias iguais, ok;
   - Intervalo da variável 1 (A1: A7);
   - Intervalo da variável 2 (B1: B7);
   - Rótulos, alfa 0, 05;
   - Opções de saídas;
   - Nova folha de cálculo, ok.

O resultado esperado:

	A	B	C
1	Teste T: duas amostras com variâncias iguais		
2			
3		*Marca A*	*Marca B*
4	Média	10,6666667	16,33333333
5	Variância	41,4666667	119,4666667
6	Observações	6	6
7	Variância agrupada	80,4666667	
8	Hipótese de diferença de média	0	
9	gl	10	
10	Stat t	-1,0941586	
11	P(T<=t) uni-caudal	0,14976994	
12	t crítico uni-caudal	1,8124611	
13	P(T<=t) bi-caudal	0,29953988	
14	t crítico bi-caudal	2,22813884	

3. Conclusão: Como $-2,2281 \leq t_{cal} \leq 2,2281$, não se pode rejeitar a hipótese de igualdade das médias ao nível de significância de 5%.

Resolução com R

Esta resolução, obedece os seguintes passos:

1. Digitar os seguintes dados em new script:
   - x=c(14, 20, 2, 11, 5, 12);
   - y=c(4, 16, 28, 9, 31, 10).
2. Calcular o valor da variável t_{cal}, digitando o seguinte comando:
   - t.test(x, y, alternative="two.sided", val.equal=T).

```
R Console
> x=c(14,20,2,11,5,12)
> y=c(4,16,28,9,31,10)
> t.test(x,y,alternative="two.sided",val.equal=T)

        Welch Two Sample t-test

data:  x and y
t = -1.0942, df = 8.0978, p-value = 0.3054
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -17.584450   6.251116
sample estimates:
mean of x mean of y
 10.66667  16.33333
```

Neste estudo, trabalhamos apenas com o test t de student (para variância desconhecida). O Excel possui a função test t, que realiza o teste de hipótese para duas amostras com opções variâncias iguais, variâncias desiguais e duas amostras emparelhadas. Esta função retorna a média, a variância, o número total das observações, variância agrupada, hipótese de diferença de média, os graus de liberdade, stat t (valor de t), o $P(t <= t)$ uni-caudal, o t crítico uni-caudal, o $P(t <= t)$ bi-caudal e t crítico bi-caudal. A decisão do teste pode ser tomada com base na região crítica, comparando o valor de stat t com t crítico bi-caudal.

- Se $-t$ crítico bi-caudal $\leq$ stat t $\leq$ t crítico bi-caudal, não se rejeita H_0
- Se stat t $>$ t crítico bi-caudal ou stat t $< -t$ crítico bi-caudal, rejeita-se H_0

Existem várias formas de realizar os testes no R. No presente estudo, usamos os comandos t.test() e tsum.test() do pacote "BSDA", que retornam a hipótese alternativa, o intervalo de confiança, a média amostral, o valor de t, os graus de liberdade da distribuição e o valor P (P-value). O valor P (P-value), indica uma estatística de teste mais extrema que aquela observada na amostra, assumindo que a hipótese nula é verdadeira. A decisão do teste pode ser tomada com base na região crítica para $\bar{x}$ ou para t ou com base no valor do P (P-value), ou seja, a decisão pode ser tomada de uma das seguintes formas equivalentes:

- Se $\bar{x} \in RC$, rejeita-se H_0 a um nível de significância de α
- Se $t \in RC$, rejeita-se H_0 a um nível de significância de α
- Se P-value $> \alpha$, não se rejeita H_0
- Se P-value $< \alpha$, rejeita-se H_0

De modo geral, os testes foram efetuados de forma bastante rápida e eficiente no R do que no Excel, principalmente naqueles casos em que os conjuntos das observações não foram dados na forma de uma matriz.

Capítulo 5

Conclusões finais

O Computador exerce um papel decisivo no Ensino da Matemática, nos dias atuais, em virtude das possibilidades de construção de modelos virtuais para a matemática imaginária MENDES(2009, p.113). Todavia, apresenta uma série de vantagens e riscos, conforme os modos de uso e com base em cada proposta pedagógica em que está apoiado.

Existem muitos ambientes computacionais que podem ser utilizados na educação. No presente estudo, foram abordados alguns ambientes que podem ser utilizados na sala de aula, baseados na resolução de exemplos de tópicos de Estatística que estão no Manual de Matemática do 11º ano do Ensino Secundário de Angola e alguns exemplos com outras referências, explorando a folha de cálculo do Excel da Microsoft e R. Tem como finalidade, mostrar por meio de exemplos as diversas potencialidades que fazem do Excel e R ferramentas adequadas para apoiar o trabalho do professor no ensino de Estatística.

Para cumprir este objetivo apresentamos no segundo, terceiro e quarto capítulos, as resoluções dos exemplos explicando de forma detalhada os procedimentos de resolução e especificando as ferramentas de análise estatística utilizadas. Essa integração torna a aprendizagem da Estatística eficaz e mostra que há várias alternativas metodológicas para alcançar o mesmo resultado.

A nível do 2º capítulo foram resolvidos nove exemplos com a finalidade de mostrar as diferentes formas de construções de tabelas de distribuições de frequências e gráficos. Concretamente, mostrando os passos a seguir quando estamos perante um conjunto de dados discretos, contínuos ou ainda perante o conjunto de dados agrupados. No Excel, as construções foram feitas com base nas funções pré-definidas na folha de cálculo e algumas com procedimentos não tão imediatos precisou-se de um conhecimento de diversos passos manuais na forma de manusear os dados para obter os resultados desejados de forma automática. No R, destacou-se os pacotes "fdth", "cbind", os comandos "plot()"e "hist()"que de certa maneira facilitaram estas construções.

No terceiro capítulo, mostraram-se os procedimentos para o cálculo de medidas de Estatística descritiva, determinação do diagrama de dispersão, o coeficiente de correlação linear e a reta de regressão simples. No Excel além de funções pré-definidas na folha de cálculo, destacou-se ainda o pacote "Real statistics"que também retorna estas medidas. O R possui comandos que facilitam o cálculo das medidas de Estatística descritiva, mas em alguns casos é necessário carregar alguns pacotes, como é o caso da moda que usa "mfv()"do pacote "modeeste", o desvio médio que usa "aad()"do pacote "lsr", a assimetria "skewness(..., type=)"e curtose "kurtosis (...,type=)"que utilizam respetivamente o pacote "e1070", onde o "type"permite obter resultados diferentes de graus de assimetria e curtose, por exemplo, para obter resultado semelhante com o obtido no Excel usa-se "type= 2".

No quarto capítulo, mostraram-se as diferentes formas de calcular as principais distribuições de probabilidades para variáveis aleatórias discretas e contínuas. Também abordaram-se neste capítulo, os testes de hipóteses para as diferenças e igualdades de médias. Os cálculos foram efetuados de forma bastante equivalentes, tanto no Excel como no R, onde em cada um dos programas os cálculos baseou-se nas diversas funções e comandos que possuem. Relativamente aos testes efetuados, notou-se ligeiras diferenças na execução das funções onde o R, revelou ser mais rápido e eficiente, principalmente nos casos em que o conjunto de dados vem especificado para cada medida.

No decorrer desta pesquisa, constatamos que em algumas situações, essas aplicações exigem um conhecimento profundo tanto da matemática como do próprio software, razão pela qual, concluimos que o papel do professor como mediador é fundamental na escolha e aplicação de ambiente computacional no processo de ensino e aprendizagem. Podemos ter noção da dimensão desta situação recorrendo a FLORENTIN(2003, p.246):

> A escolha de um ambiente computacional, a ser utilizado no processo de ensino e aprendizagem da matemática, relaciona-se com diversos aspetos, tanto teóricos quanto metodológicos; entretanto, um dos aspetos fundamentais consiste na mediação do professor, uma vez que o ambiente, por mais construtivo que seja, não é suficiente para promover, por si só, contextos propícios à construção do conhecimento.

Conforme o mesmo autor, a mediação do professor desempenha um papel determinante, na medida em que ele cria situações desafiantes, recortando-as em vários problemas intermediários que possibilitam aos alunos deslocarem-se muitas vezes do problema principal, olhando-o e percebendo-o de uma outra perspetiva, possibilitando-lhes a busca de novos caminhos, a constante reavaliação de suas estratégias de seus objetivos, enfim, o seu envolvimento cada vez maior no processo de construção do conhecimento.

De modo geral, a utilização de ferramentas computacionais no processo do ensino e aprendizagem é sempre uma mais valia, para o trabalho do professor e aprendizagem dos alunos, pois incentiva a curiosidade, o aumento de confiança e o gosto pela matemática, ajustando a criar ambientes de trabalhos em que os alunos são encorajados a fazer e testar conjeturas, tal como destaca ainda o FLORENTIN(2003, p.226):

> A tecnologia não existe apenas em um recurso a mais para os professores motivarem suas aulas; consiste sobretudo, em um meio poderoso que pode propiciar aos alunos novas formas de gerar e disseminar conhecimento, e, consequentemente, propiciar uma formação condizente com os anseios da sociedade. Assim sendo, os professores de matemática devem refletir sobre sua utilização, trabalhando em pesquisas que implemente projetos nas escolas, design de ambientes interativos de aprendizagem colaborativa que possam oferecer oportunidades para que seus alunos aprendam a matemática e, ao mesmo tempo, utilizem a tecnologia de forma que a matemática, no contexto tecnológico, torne-se um caminho para a superação das desigualdades sociais para a formação e a inserção adequada do sujeito a uma sociedade permeada pela tecnologia.

A Estatística tem um campo vasto e fértil de aplicação na ciência e na tecnologia. Portanto, espera-se que as informações obtidas através desta pesquisa possam gerar reflexões acerca das dificuldades encontradas no ensino e aprendizagem desta disciplina, assim como criar meios que possam levar à sua superação e principalmente à busca de alternativas pedagógicas preventivas.

Bibliografia

[1] CARVALHO, Adelaide, Exercícios de Excel para Estatística, FCA-Editora, Lisboa 2015;

[2] Dossiês-Alea, Acedido à 28 de Março 2017, de `http://www.alea.pt/html/statofic/html/dossier/doc/publicacao_2009_web.pdf`

[3] FLORENTINI, Dário e outros "Formação de professores de matemática, explorando novos caminhos com outros olhares", campinas, SP: mercado de letras 2003.

[4] FONSECA, Jairro Simon da e Martins Andrade Gilberto de, Curso de Estatística, 6ª Edição, S. Paulo 1996;

[5] GOMES, Januário, SPSS e R como ferramentas no ensino de Estatística no 12º ano de escolaridade em Tímor-Leste, tese de mestrado em Matemática para Professores FCUP, 2016;

[6] MENDES, Iran Abreu, "Matemática e investigação em sala de aula" 2ª Edição revisada e ampliada, Editora Livraria de Física, S. Paulo 2009;

[7] PIAIRO, Helena e Pereira Miguel, Introdução à Estatística em R e SPSS, Chiado Editora, Lisboa 2012;

[8] REIS, Elisabeth, Estatística Descretiva, 7ª Edição, Lisboa 2012;

[9] SPIEGEL, Murray R. Estatística curso intensivo, editora McGraw-Hill, Lisboa 2001;

[10] TELES, Paulo et all, Estatística Descritiva e Probabilidades, Problemas resolvidos e propostos com aplicações em R, 2ª ed, Escolar Editora, Lisboa 2009;

[11] TELES, Paulo et all, Inferência Estatística, Problemas resolvidos e propostos com aplicações em R, Escolar Editora, Lisboa 2017;

[12] TOMÁS, Marta Teresinha, Matemática 11ª Classe, 2º Ciclo do Ensino Secundário, texto editora, Lda-Angola, 2006.

Anexo 1

PROGRAMA DE ESTATÍSTICA APLICADA (ISCED/UÍGE)

2º- ANO –SEMESTRAL

II SEMESTRE -3 HORAS/S (1HTP+2HP – 3UC)

TOTAL: 45 H

OBJECTIVOS:

Compreender os conceitos básicos, regras de Estatística; - Desenvolver a capacidade de raciocínio indutivo e dedutivo; Proporcionar a nível científico uma sólida formação em Estatística Aplicada; Analisar, interpretar e avaliar estudos de natureza estatística; Aprofundar a objetividade dos conhecimentos de Estatística em Educação.

A-PLANO SINTÉTICO:

UNIDADE I: Parte teórico-prática

UNIDADE II: Parte prática

B-PLANO ANALÍTICO

SISTEMA DE CONHECIMENTOS

UNIDADE I: Parte teórico-prática

- Problemas gerais

- Estatística Descritiva: essência, distribuição de frequências, medidas de localização, de dispersão, de assimetria e achatamento; - Correcção e Regressão simples;

- Probabilidades

- Problemas gerais e estimação de um parâmetro – Distribuições discretas de Probabilidades (Binomial, Hipergeométrica e de Poisson);

- Distribuições contínuas de Probabilidades (Normal e Exponencial)

- Distribuição de Amostragem - Comparação do valor de duas amostras

- Testes de Hipóteses - Teste T de Stundent

UNIDADE II: Parte prática

Trabalho Prático: - Aulas práticas – **Laboratório de Matemática:** determinar as formas de aplicabilidade da estatística; elaboração de frequências, medidas; determinar as probabilidades; distribuição de Amostragens e elaboração de testes;

- Exercitar todos os itens do programa e resolução de exercícios práticos. Estão reservadas: uma hora teórico-prática e duas horas práticas.

METODOLOGIA

- Conferências e exposições, método de elaboração conjunta, método de trabalho independente, método demonstrativo, ilustrativo e resolução de problemas, trabalhos práticos em grupo e individuais, ou seja:

- As temáticas serão desenvolvidas em aulas teórico-práticas e práticas

SISTEMA DE AVALIAÇÂO:

Provas de frequência e exames teórico-práticos; Classificação das aulas práticas.

BIBLIOGRAFIA:

- Bento Murteira e Outros, Introdução à Estatística, McGraw-Hill, Portugal, 2002.

- Guimarães, Rui Campos Cabral, José Sarsfield, Estatística, Editora McGraw-Hill, Portugal, 1997.

- Mário Barroso e Outros, Exercícios de estatística Descritiva para as Ciências Sociais, Ed, Sílabo, Lisboa, 2003.

- Murteira, B. Blac, GHJ, Estatística Descritiva, Ed. McGraw Hill, 1983;

- Reis, E. Estatística Descritiva, Editora Sílaba, Lisboa, 3ª- edição

- Reis, E., Estatística, Ed. Sílaba, Lisboa, 1996

- Wonnacott, T.H., Wonnacott, R. J., Fundamentos de estatística, Livro Técnico e Científicos Editora AS, Rio de Janeiro, 1997

Anexo 2

Instalação do R

Esta instalação, obedece as seguintes instruções:

1. Aceder a pagina oficial do programa, disponível em: `https://cran.r-project.org/bin/windows/base/`;
2. Duplo clique no ficheiro de execução e a seguir clique em Run;
3. Selecione o idioma que se pretende e clique em ok;
4. Clique no botão seguinte;
5. Ler as instruções e clique no botão seguinte;
6. Selecione o local onde pretendes que o R seja instalado e clique no botão seguinte;
7. Clique no botão seguinte;
8. Selecione No (accept defaults) e clique no botão seguinte;
9. Escolher onde colocar um atalho do programa e clique no botão seguinte;
10. Selecione tarefas adicionais e clique no botão seguinte, aguardando a instalação;
11. Clicar em concluir.