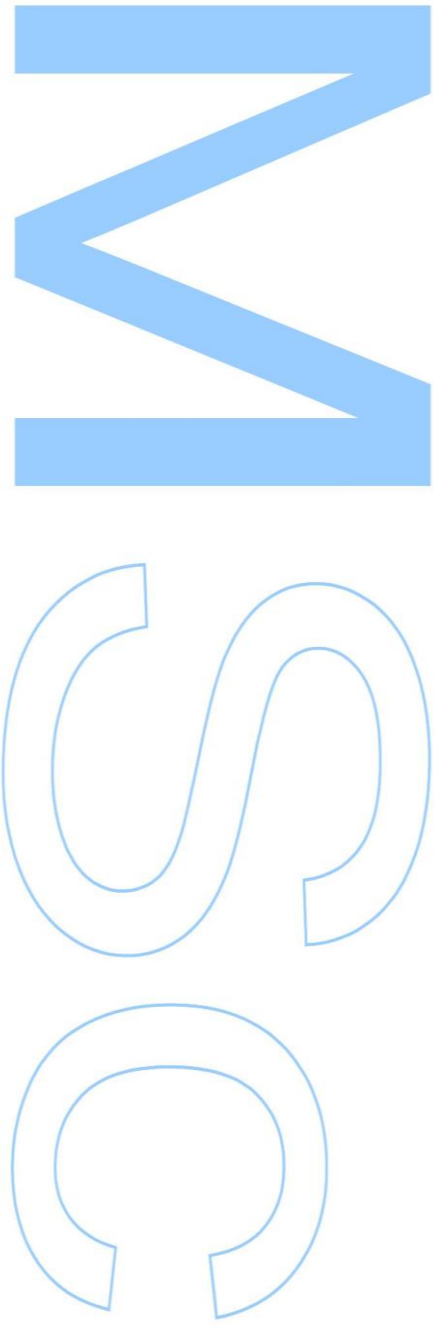
SPSS e R como Ferramentas no Ensino de Estatística no 12º Ano de Escolaridade em Timor-Leste

Januário Gomes
Mestrado em Matemática para Professores
Departamento de Matemática
2016

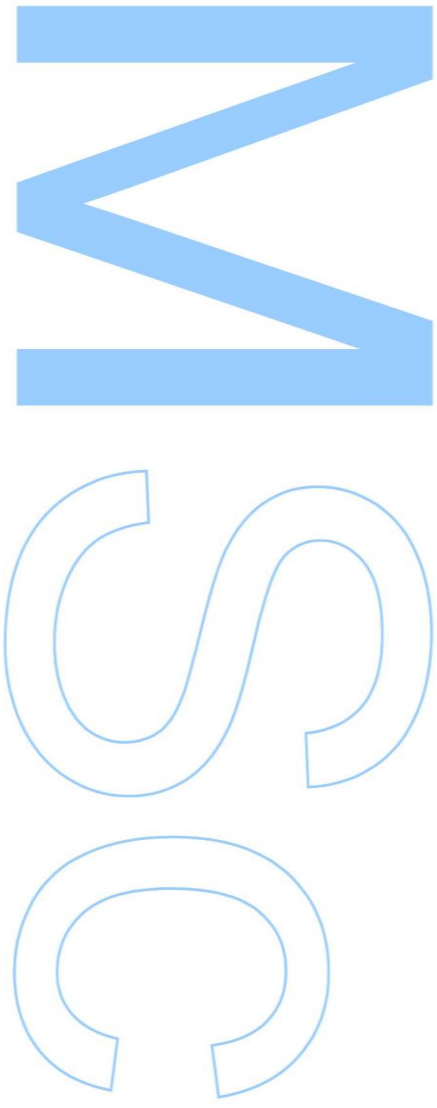
Orientador
Óscar António Louro Felgueiras, Professor Auxiliar
Faculdade de Ciências da Universidade do Porto



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

O Presidente do Júri,

Porto, _____/_____/_____



À minha esposa (Celsa da Costa), à minha filha (Angelina Maria Indira Gomes) e aos meus pais (Raimundo Gomes e Felismina da Costa)

Agradecimento

Agradeço pelo apoio e pela graça de Deus, que me acompanha ao longo da minha vida. Considero ter enfrentado muitas dificuldades durante a elaboração deste trabalho, mas todas foram superadas pelo apoio tanto moral quanto material recebido. Ao mesmo tempo não me esqueço de agradecer:

Ao meu orientador Professor Doutor Óscar António Louro Felgueiras, pela disponibilidade manifestada, pela dedicação durante as orientações e pelas valiosas sugestões fornecidas para este trabalho.

Aos professores do Curso de Mestrado em Matemática para Professores da Faculdade de Ciências da Universidade do Porto por me terem preparado com conhecimentos importantes para a minha vida profissional.

A todos os colegas do Curso de Mestrado em Matemática para Professores da Faculdade de Ciências da Universidade do Porto do ano letivo de 2013/2014 pelo apoio ao longo do curso.

Resumo

Actualmente, a tecnologia dos computadores e os programas de estatística já fazem parte do ensino de estatística em todos os níveis da educação e possibilitam a análise dos dados e a representação gráfica. Por isso, este trabalho tem como finalidade analisar e entender as funcionalidades do programa SPSS e da linguagem **R** como ferramentas pedagógicas para professores no Ensino de Estatística, em Timor-Leste.

O SPSS e **R** vão ser utilizados para resolver exercícios do manual de matemática do 12º ano de escolaridade e alguns exercícios com outras referências. Para facilitar a análise dos exercícios e descrições das resoluções, cada exercício será resolvido, ao mesmo tempo, com os dois programas de estatística e serão apresentadas algumas análises comparativas entre resoluções feitas nos programas e no manual do aluno.

Na pretensão de representar diagramas de extremos e quartis, tanto o **R** como o SPSS partem da representação de diagramas de caixa e bigodes. O **R** possui o comando *boxplot()* com opção *range = 0* para excluir os valores atípicos, enquanto que o SPSS não possui essa opção. Além disso, os quartis produzidos por ambos os programas seguem o método inclusivo ao passo que no manual do aluno é seguido o método exclusivo. No caso de o número de observações ser par, os dois métodos coincidem. No caso de ser ímpar, os quartis do método exclusivo podem ser obtidos no **R** com o comando qboxplot(...,type=6).

A variância e o desvio são apresentados, pelos programas estatísticos, também de forma diferente do conteúdo do 12º ano. Ambos os programas calculam a variância e o desvio padrão amostral, ou seja, ao calcular estas duas medidas, o SPSS determina os mesmos valores obtidos pelos comandos *var*() e *sd*() do **R**, enquanto no livro do aluno se está a calcular a variância e o desvio padrão populacionais. Para calcular o valor da variância populacional, exatamente igual à resolução do manual, é necessário executar os comandos do **R**: varp=function(x){sum((x - mean (x))^2)/(length (x))} e sdp=function(x){sqrt(sum((x-mean(x))^2)/(length(x)))}, onde varp(x) dá a variância populacional e sdp(x) dá o desvio padrão populacional.

A pesquisa constatou que os programas de estatística são ferramentas muito úteis para os professores, pois possibilitam a análise dos dados e a construção de gráficos. Contribuem também para o desenvolvimento de conceitos estudados por professores e alunos na sala de aula e sua aplicação em trabalhos profissionais.

Palavras-Chave: ENSINO DE ESTATÍSTICA, 12º ANO DE ESCOLARIDADE DE TIMOR-LESTE, SPSS E **R**.

Abstract

Nowadays, computer technology and statistical software already play an important role in teaching statistics at all education levels and allow users to perform data analysis and graphical representations. For those reasons, this work has the goal of analysing and understanding the capabilities of the SPSS program and the **R** language as pedagogical tools for teachers in Statistics Education, in East-Timor.

SPSS and **R** will be used for solving exercises from the 12th grade textbook of mathematics and some exercises from other books. In order to make it easier the analysis of the exercises and the description of the resolutions, each excercise will be solved simultaneously by both statistical software programs and comparisons between resolutions obtained with or without using software will be shown.

Intending to represent boxplots, both **R** and SPSS show outliers by default. In **R**, the command *boxplot()* with the option *range=0* excludes outliers, while SPSS does not have that feature. Moreover, quartiles produced by both software programs follow the inclusive method whereas the student textbook follows the exclusive method. In case the number of observations is even, both methods coincide. In case it is odd, quartiles from the exclusive method may be obtained in **R** with the command *qboxplot(. . . ,type=6)*.

Variance and standard deviation are also computed by both statistical software programs differently from how it is done in the 12th grade textbook. Both software programs compute sample variance and sample standard deviation, meaning that SPSS determines the same values obtained by the commands *var()* and *sd()* from **R**, whereas the student textbook shows population variance and population standard deviation. To compute population variance using the same formula from the student textbook, one can define in **R** the functions varp=function(x){sum((x - mean (x))^2)/(length (x))} and sdp=function(x){sqrt(sum((x-mean(x))^2)/(length(x)))}, where varp(x) gives the population variance and sdp(x) gives the population standard deviation.

Our research has found that statistical software programs are very useful tools for teachers, because they allow them to perform data analysis and graphical constructions. They also contribute for the development of concepts studied by teachers and students in the classroom and their application in professional projects.

Keywords: STATISTICS EDUCATION, 12TH GRADE CURRICULUM IN EAST-TIMOR, SPSS AND **R**.

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

O presente capítulo começa por uma breve apresentação dos motivos fundamentais que fazem surgir este estudo. Serão também apresentados os objetivos do estudo proposto neste trabalho.

1.1 Motivos fundamentais deste estudo

Todos os dias encontramos dados numéricos, tanto na escola como fora da escola, o que nos obriga a fazer análise e a tirar conclusões estatísticas sobre esses dados. Para facilitar a análise dos dados necessitamos de apoio de programas estatísticos. O SPSS e **R** são ferramentas que fornecem uma ampla variedade de técnicas estatísticas (manipular, analisar os dados, classificar, agrupar) e que podem trazer muitas vantagens para os professores.

Atualmente, o contributo dos computadores no ensino da matemática aumentou, pois são considerados como ferramenta muito útil no ensino e aprendizagem. A utilização de meios informáticos no ensino de estatística é particularmente importante no ensino superior devido às numerosas aplicações disponíveis. Os programas SPSS e **R** destacam-se aí devido à sua grande versatilidade. No ensino secundário, os professores poderão beneficiar com um melhor conhecimento destes programas, de modo a aplicá-los na lecionação. Estes programas podem ser utilizados pelos professores nas elaborações de enunciados. Em vez de apresentar perguntas a partir de dados numéricos, torna-se possível fazê-lo com o recurso a gráficos produzidos com estes programas. Além disso, os professores podem usar o software para mostrar tabelas e gráficos aos alunos.

Existem muitos professores que não possuem suficientes conhecimentos teóricos nem práticos para operar com estes programas e em muitas escolas nem sequer estão instaladas redes de computadores ou laboratório de computadores. Por esta razão, muitos professores ainda continuam a ensinar estatística aos seus alunos dando mais relevância às fórmulas

e processo de cálculo do que à apresentação gráfica. Os gráficos são feitos manualmente com grande dificuldade na interpretação e compreensão dos resultados.

Além dos motivos apresentados em cima, a motivação pessoal do pesquisador para fazer este estudo surgiu depois de frequentar a parte curricular do curso de Mestrado, em Matemática, para professores da Universidade do Porto ano lectivo de 2014/2015. Conheci pela primeira vez o programa **R** na disciplina de Tópicos de Matemática Aplicada e o programa SPSS ao longo da orientação da dissertação. Estes programas suscitaram-me interesse pois apercebi-me que eles poderiam constituir ferramentas de modo a ajudar os meus alunos e os meus amigos professores a ultrapassarem dificuldades na aprendizagem da estatística.

1.2 Objetivos do estudo

Os programas de estatística trouxeram benefícios para os professores, porque oferecem vários comandos (menus) que podem ser utilizados para analisar e visualizar os dados, tabelas e gráficos. A utilização destes programas no ensino e aprendizagem de estatística já é uma realidade e são ferramentas usadas por professores ao modernizar a sua metodologia de ensinar na sala de aula. Ou seja, a integração de programas como o SPSS e o **R** no ensino é fundamental. Para o NCTM(1991):

> "A tecnologia de computação permite que os alunos representem, de forma rápida, a informação por meio de gráficos (com ajustamento de curvas já executado) e calculem valores estatísticos com uma precisão considerável, utilizando apenas as teclas do computador. Aquilo que falta - e que o estudo da estatística deve possibilitar - é o entendimento de quais as medidas apropriadas para um dado problema e o que é que medidas como a média, a variância e o coeficiente de correlação lhes podem dizer acerca desse problema. Além disso, é essencial que os alunos aprendam a interpretar os resultados de um modo inteligente."

Os professores devem utilizar a tecnologia de computação como um recurso para facilitar a compreensão dos conceitos através da apresentação visual. O uso desta tecnologia para interpretar os dados estatísticos torna as aulas mais atrativas. Como já referido, a estatística oferece métodos para recolher, agrupar e analisar dados. Permite também representar gráficos e tirar conclusões sobre os esses dados. O computador como ferramenta de apoio possibilita rapidamente a execução destas tarefas estatísticas e estabelecer a respetiva correspondência entre conceito estudado e resultado apresentado.

Atualmente em Timor-Leste, a tecnologia de computação para professores vive uma realidade muito particular. A maior parte dos professores continuam a mostrar dificuldades

em operar programas de estatística. O material de estatística utilizado pelos professores e distribuído aos alunos, também é feito exclusivamente em papel, utilizando a álgebra, cálculo numérico e gráficos. A maioria dos alunos mostram dificuldades de aprendizagem através desta metodologia. Por outro lado, os professores não dão devida importância ao uso de programas de estatística.

Baseando-se nesta particularidade, este estudo tem por objectivos gerais:

1. Convidar os professores a conhecerem o software de estatística.

2. Sugerir uma reformulação no currículo para considerar a importância do computador e programas de estatística no ensino de estatística em Timor-Leste.

Neste trabalho far-se-á o estudo do programa SPSS e linguagem **R** como ferramentas no ensino de estatística. Embora se utilize dois programas de estatística, não será objetivo desta pesquisa identificar qual deles tem melhor capacidade como programa, na linguagem, no ambiente e nas funções para análises estatísticas no ensino/aprendizagem. Este estudo tem por objetivo específico analisar e entender a funcionalidade da aplicação dos programas de estatística como ferramentas para professores de matemática na resolução de problemas, tanto no ensino como em outras situações, como por exemplo profissionais.

Capítulo 2

Construções de tabelas de frequências

2.1 Introdução

Inicialmente, mostrar-se-ão temáticas e demonstrações das fórmulas matemáticas de Estatística que estão relacionadas com os conteúdos do programa do décimo segundo ano de escolaridade, em Timor-Leste. Em seguida, serão apresentadas as componentes ou ferramentas básicas do programa de SPSS e linguagem **R** para o ensino de Estatística que servem como meio ou fundamento para operar e interpretar os resultados, produzidos por esses programas, ao longo destes três capítulos (capítulo 2, 3 e 4)

A versão de SPSS que será utilizada na análise dos dados é IBM SPSS Statistics 22, enquanto a de **R** vai ser R-3.1.3 for Windows. Ao longo destes capítulos apenas serão utilizadas duas Janelas de **R**, consideradas muito simples e fáceis de operar, que são *Console* e *Script*.

Pretende-se propor aos professores para terem em consideração a importância dos programas de estatística, sobretudo o SPSS e o **R**, como recurso mais adequado nos processos de organização, análise e interpretação de dados. Para facilitar a análise dos exercícios e descrições das resoluções, cada exercício será resolvido ao mesmo tempo com os dois programas de estatística, ou seja, aplicar-se-ão tanto o SPSS como o **R** ao mesmo exercício. Vão sendo mostradas as fases da resolução do problema através dos menus ou comandos e serão apresentadas ao mesmo tempo as janelas da aplicação.

Assim, os presentes capítulos (capítulo 2, 3, e 4) irão mostrar como é que os programas de estatística podem ser considerados ferramentas muito úteis para a resolução dos problemas ou exercícios contidos no livro do décimo segundo ano de Timor-Leste ou em outras situações consideradas relevantes. Além disso, dão uma noção da linguagem utilizada por cada um dos programas e respetiva capacidade computacional, sobretudo

na manipulação e visualização dos dados, tanto nas tabelas como nos gráficos.

Para desenvolver estas temáticas será utilizada principalmente uma referência tida como simples, adequadas e importante: Afonso e Nunes (2001), *Estatística e Probabilidade: Aplicações e Soluções em SPSS*. Este livro será utilizado para descrever as temáticas da Estatística Descritiva e Indutiva. Além destas referências será acrescentado o livro de matemática do décimo segundo ano de escolaridade que é utilizado em Timor-Leste (2014). As temáticas de Estatística estudadas no décimo segundo ano de escolaridade, em Timor-Leste, são apresentadas na tabela 2.1.

Tabela 2.1: Temáticas da organização e tratamento de dados

Nº	Estatística descritiva e indutiva	Nº	Estatística descritiva e indutiva
1	Introdução	8.4	Mediana
2	Recenseamento e sondagem	8.5	Quartis
3	Estatística descritiva e indutiva	8.6	Diagrama de extremos e quartis
4	Atributos estáticos	9	Medidas de dispersão
5	Organização de dados	9.1	Amplitude total e interquartis
5.1	Tabelas de frequências	9.2	Variância
5.2	Distribuições estáticas	9.3	Desvio padrão
5.3	Frequências absolutas	9.4	Propriedades
5.4	Frequências acumuladas	10	Distribuições bidimensionais
5.5	Função cumulativa	10.1	Recta de regressão
6	Dados agrupados em classes	10.2	Coeficiente de correlação
7	Representações gráficas	11	Distribuições de probabilidade
7.1	Diagrama de barras	11.1	Valor médio de uma variável aleatória
7.2	Diagrama circular	11.2	Desvio padrão de uma variável aleatória
7.3	Pictogramas	12	Variáveis aleatórias discretas
7.4	Histograma	12.1	Distribuição binominal/modelo binominal
7.5	Polígono de frequências	12.2	Modelo de poisson
8	Medidas de localização	13	Variáveis aleatórias contínuas
8.1	Média	13.1.	Distribuição normal
8.2	Propriedades da média	13.2	Características da curva normal
8.3	Moda	13.2	-

Fonte:Manual do Aluno (2014)

De modo geral, a teoria da Estatística Descritiva e Indutiva é considerada uma componente muito importante dos conteúdos deste currículo, os quais passaram a ter temáticas mais avançadas e ambiciosas a partir de 2014.

2.2 Tabela de dados univariados

Recolher dados numéricos em maior ou menor quantidade, analisar e interpretar são actividades de rotina da sociedade moderna. Analisar grandes quantidades de dados,

sobretudo quando estão desorganizados, não é uma tarefa fácil. Para que a análise se torne mais simples é necessário um melhor conhecimento na tabulação destes dados. Distribuir o número de observações na tabela de distribuição de frequências é um processo muito útil para interpretação e obtenção rápida dos valores desejados pelas pessoas.

Esta tabela contém várias classes ou categorias que, por sua vez, contêm os dados ou números que se chamam frequências. Segundo Afonso e Nunes (2011, p.11), uma tabela de frequências relaciona as categorias ou classes de valores com o número de ocorrências, ou frequência, de observações que pertencem a cada categoria ou classe. Salienta-se também que as categorias ou classes de valores devem ser:

1. Mutuamente exclusivas, ou seja, cada valor observado só poderá pertencer a uma das categorias ou classes;

2. Exaustivas, ou seja, as categorias ou classes devem compreender todos os valores observados.

Notação: A notação utilizada nas tabelas de frequências é:

k	Número de categorias ou valores distintos ou classes de valores que os dados assumem;
n_i	Frequência absoluta de categoria ou valor ou classes de valores i;
$f_i = \frac{n_i}{n}$	Frequência relativa de categoria ou valor ou classe de valores i;
$N_i = \sum_{h=1}^{i} n_i$	Frequência absoluta acumulada de categoria/valor ou classe de valores i;
$F_i = \sum_{h=1}^{i} f_i$	Frequência relativa acumulada de categoria/valor ou classe de valores i.

2.3 Tabela de frequências para dados qualitativos ou quantitativos discretos

A construção de uma tabela de frequência para dados qualitativos ou quantitativos discretos (Tabela 2.2) depende da definição das seguintes colunas (Afonso e Nunes, 2011, p.12):

1ª Coluna	Todas as k categorias ou valores distintos, x_i que os dados assumem.

2ª Coluna As frequências absolutas, n_i, ou seja, o número de vezes que cada categoria (valor) foi observada (o).

3ª Coluna As frequências relativas, f_i, ou seja, a proporção de vezes que cada categoria (valor) foi observada (o).

4ª Coluna As frequências absolutas acumuladas, N_i , ou seja, o número de ocorrências das categorias (valores) inferiores ou iguais à categoria (valor) actual.

5ª Coluna As frequências relativas acumuladas, F_i , ou seja, a proporção de ocorrências das categorias (valores) inferiores ou iguais à categoria (valor) actual.

Observação: Para dados qualitativos na escala nominal, não se calculam as frequências absolutas e relativas acumuladas (4ª e 5ª colunas).

x_i	n_i	f_i	N_i	F_i
x_1	n_1	f_1	N_1	F_1
x_2	n_2	f_2	N_2	F_2
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
x_k	n_k	f_k	$N_k = n$	$F_k = 1$
Total	n	1		

Tabela 2.2: Tabela de frequência para dados qualitativos ou quantitativos

Exercício 1: Num estudo para analisar a ocorrência de acidentes de trabalho num determinado hospital, em 397 profissionais de saúde verificou-se que 16 não sofreram qualquer acidente, 32 tiveram 1 acidente, 89 reportaram 2 acidentes, 137 sofreram 3 acidentes, 98 sofreram 4 acidentes e 25 profissionais reportaram 5 acidentes (Ver Tabela 2.3). (Afonso e Nunes. 2011, p.12-13)

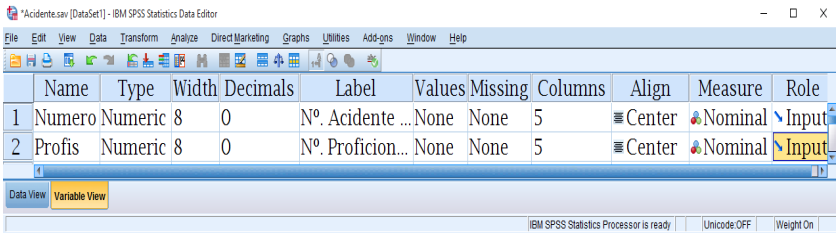
x_i'	n_i	f_i	N_i	F_i
0	16	0.0403	16	0.0403
1	32	0.0806	48	0.1209
2	89	0.2242	137	0.3451
3	137	0.3451	274	0.6902
4	98	0.2469	372	0.9370
5	25	0.0630	397	1.0000
Total	397	1.0000		

Tabela 2.3: Tabela de frequências relativas ao número de acidentes por profissional.

Resolução com SPSS

Os dados da tabela 2.3 podem ser obtidos através do SPSS. Antes de começar, deve ter-se em atenção que, se os dados ainda não estiverem inseridos no SPSS, é necessário introduzi-los na área de inserção.

Para iniciar, deve abrir-se a janela **Variable View**, localizada na parte inferior, à esquerda, da **Data View**, clicando em **Variable View**, apenas uma vez. Em seguida, cria-se a variável com o nome adequado. Esse nome deve ser escrito de forma a ser fácil a sua compreensão *(Numeros e profis)*. Entretanto, guarda-se o ficheiro com o nome escolhido, neste caso *Acidente.sav*, como mostra a janela seguinte:



A janela de **Variable View** é um espaço para definir ou inserir os nomes das variáveis de estatística. Esta janela inclui:

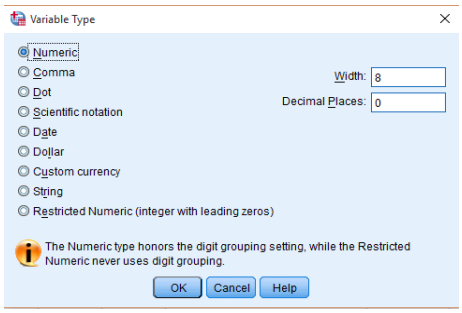
1. Name (nome): Nesta coluna pode-se inserir ou modificar os nomes das variáveis. Neste caso, existem duas variáveis; a primeira variável é *Numero* e a segunda é *profis*. Para escrever os nomes devem ser observados alguns requisitos:

   (a) O nome da variável não pode começar por números;

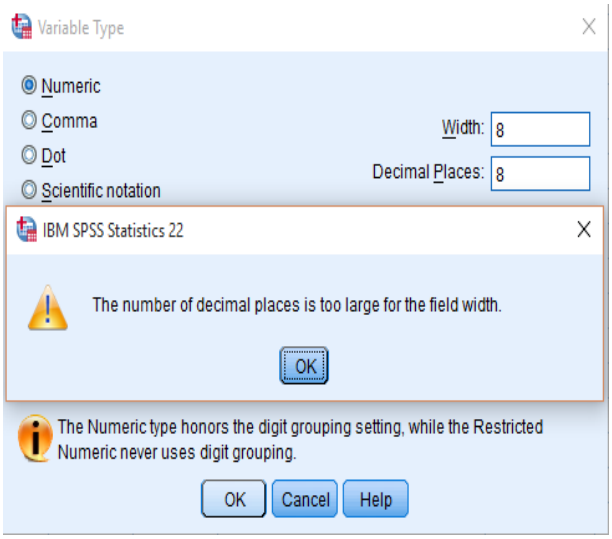
   (b) O nome da variável não pode conter mais de oito caracteres;

   (c) Não pode conter espaços; se se quiser pôr o espaço ou separar algumas palavras deve utilizar-se o hífen "__", por exemplo (dados__ numeros) e alguns símbolos aritméticos$(+, -, *, \backslash)$

2. Type (tipo) - O SPSS oferece vários tipos de dados para cada variável estatística que podem ser String, Numeric e Date, como mostra a janela seguinte:

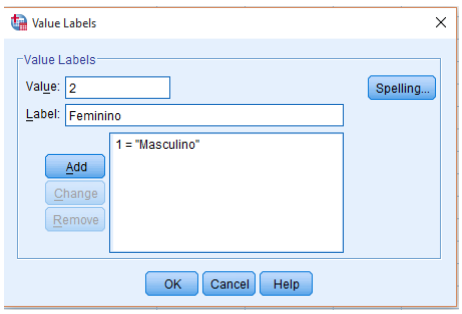


Neste caso, as *variáveis Numero e profis* possuem os mesmos tipos de dados numeric, porque os elementos de cada variável são números. **Width** compreende oito colunas ou seja, cada variável deve ser escrita com oito (ou menos) caracteres e **Decimal Places** compreende ao numero de casas decimais. O número de casas decimais deve ser menor que o número de caracteres de **width (8)**, caso contrário (igual ou maior) o SPSS vai apresentar a informação de que o número de casas décimais (decimal places) é maior ou igual a **width**, como mostra a figura seguinte (janela sobreposta da Variable Type)



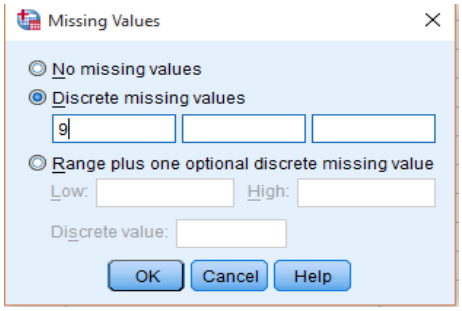
3. Label: É uma coluna que descreve de forma muita clara o nome da variável estatística na coluna de **Name**. Por exemplo, o nome da primeira variável na coluna de **Name** é Numero. Portanto, na coluna de **label**, este nome pode ser escrito por *Nº de acidente por profissional* enquanto *profis* por *Nº de profissionais*.

4. Values: É um código dado a cada elemento de uma variável categórica, por exemplo, *variável género*, composto por *Masculino*, codificado por 1 e *Feminino* codificado por 2, como mostra a seguinte janela:



Nesse caso, os dados da tabela 2.3 não estão em categorias, por isso, não é necessário codificar os elementos da variável.

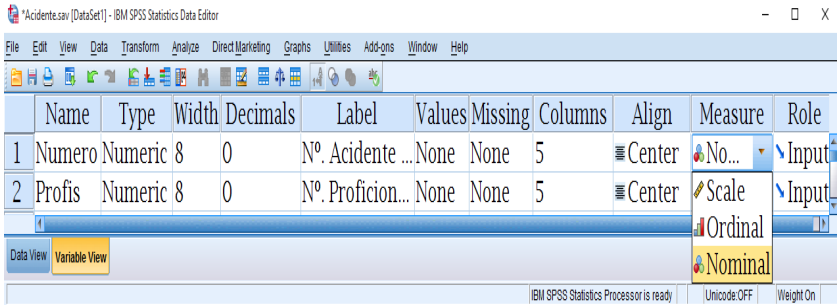
5. Missing: Dados em falta ou seja, são os dados que não serão incluídos na análise. Em missing há três opções: No missing values, Discrete missing values e Range plus one optional discrete missing values.



6. Columns: **Columns** é semelhante a **width**, pois possui a função de oferecer a largura da coluna para inserir os dados de uma variável.

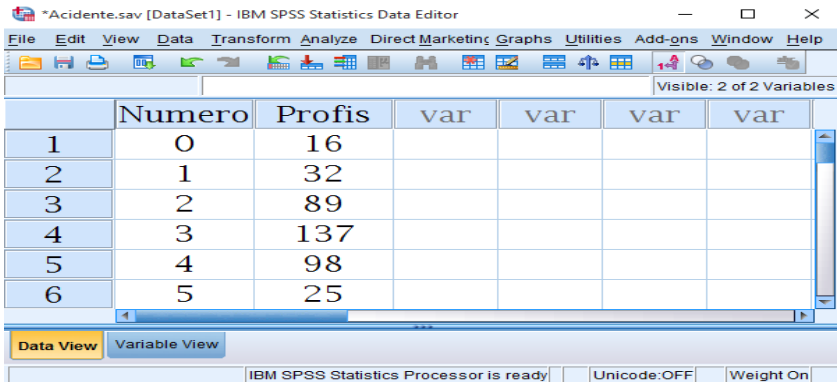
7. Align: É a posição dos dados; podem ser alinhados à direita, à esquerda ou ao centro.

8. Measure: É o tipo da variável que determina os modelos de análise. Na janela seguinte podem ver-se três tipos de **Measure** :



A variável *Numero* e a *Profis (Nº.Profissionais)* serão classificadas como **Nominal**. Porque variável *Numero* é do tipo Qualitativo de *escala nominal*

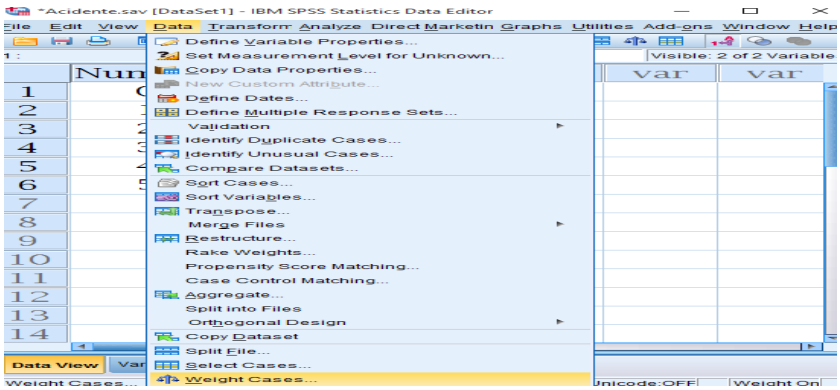
Depois de definir as variáveis, abre-se a janela **Data View**, que está ao lado de **Variable View**, para introduzir os dados da tabela 2.3.



Acede-se ao menu da janela **Data Editor** com o seguinte comando para ponderar os dados:

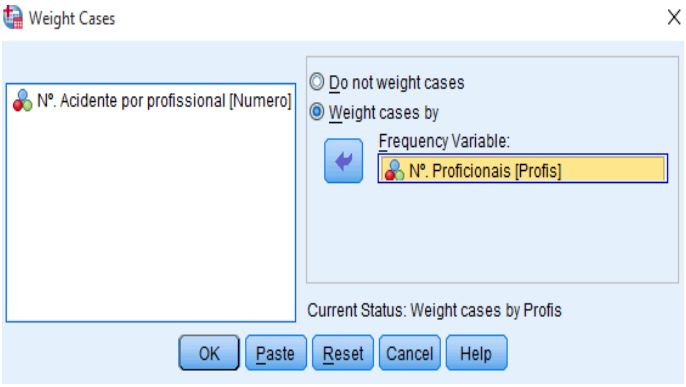
1. Primeira fase:

   ▶ Data/Weigth Cases



   ▶ Frequency Variable: Número Profissionais

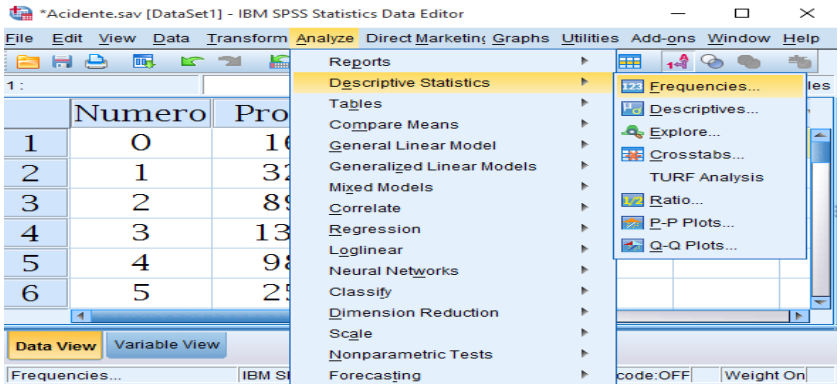
   ▶ OK



   As considerações apresentadas acima (Codificação da variável, introdução dos dados e opção Weight Cases), servem de modelo para todos os exercícios que virão a ser apresentados e não serão repetidas.

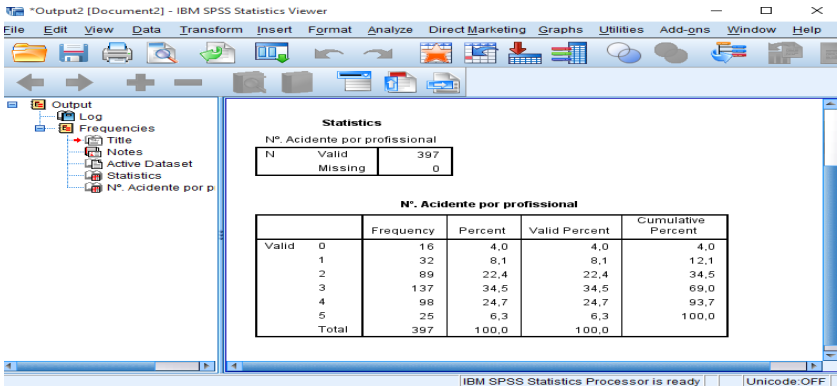
2. Segunda fase: Por último, pode executar os seguintes comandos

   ▶ Analize/Descriptive Statistics/Frequencies ...



- ▶ Variable(s): Nº. Acidentes por profissional
- ▶ ☑ Display frequency tables
- ▶ OK

O resultado é o seguinte:



Resolução com R

O trabalho da linguagem **R** baseia-se no tipo de estrutura dos dados. A estrutura mais simples é o vetor. O vetor será formado pela função Combine, c(), e o resultado da análise representado por objecto.

Neste caso, o *Nº.Acidentes por profissional* será designado por objecto **x** e o *Nº. Profissionais* por objecto **y**. Será utilizado o comando **cbind** para apresentar estes dois objetos e será gravada no *Script* com o nome de *Acidente*.

```
> x = c(0, 1, 2, 3, 4, 5)
> x
> [1] 0 1 2 3 4 5
> y = c(16, 32, 89, 137, 98, 25)
> y
> [1] 16 32 89 137 98 25
```

> tabela = cbind(x, y,fi = y/sum (y), Ni = cumsum(y), Fi = cumsum(y/sum(y)))

> tabela

	x	y	fi	Ni	Fi
[1,]	0	16	0.04030227	16	0.04030227
[2,]	1	32	0.08060453	48	0.12090680
[3,]	2	89	0.22418136	137	0.34508816
[4,]	3	137	0.34508816	274	0.69017632
[5,]	4	98	0.24685139	372	0.93702771
[6,]	5	25	0.06297229	397	1.00000000

Observando as duas tabelas de distribuição de frequências para dados univariados apresentadas pelos SPSS e **R**, percebe-se que os resultados, apresentados por estes programas, na construção da tabela de distribuição de frequências, são muito simples e estruturados. Os valores 0, 1, 2, 3, 4 e 5 são os elementos da *Variável Numero*; 16, 32, 89, 137, 98 e 25 são valores que pertencem à *Frequência ou Profissional* e 397 é a *Soma das Frequências*. Também possuem *Frequência Relativa*, *Relativa Acumulada*, e *Absoluta Acumulada*.

2.4 Tabela de dados quantitativos contínuos

Quando os dados são do tipo quantitativo contínuo é necessário definir k classes de valores que constituem as categorias dos dados em estudo Para construir esses classes existem vários métodos possíveis. Por exemplo, Se interessa comparar os resultados de um estudo com os resultados do outro estudo, é fundamental que se utilizem as mesmas classes para ser possível efectuar as comparações. A forma como se definem as classes condiciona os resultados que, apenas, são válidos para a classificação efectuada. Seja qual for o método utilizado é aconselhável obter um número muito elevado nem muito reduzido de classes (habitualmente $5 \leq k \leq 20$) (Afonso e Nunes. p.13, 2011). Salienta-se também que o método de construção de classes devem ser:

1. Determinar o número k de classes a construir, com base nas n observações, fazendo (regra de Sturges):
   $k = \left[\frac{\ln n}{\ln 2}\right] + 1 \Leftrightarrow \text{N}^{\text{o}} \text{ de classe} = \left[\frac{\ln (\text{N}^{\text{o}} \text{ de observações})}{\ln 2}\right] + 1.$
   Onde [Número] representa a parte inteira do número obtido;

2. Determinar a amplitude a do conjunto de dados fazendo:

   a = amplitude = máximo - mínimo.

3. Determinar a amplitude ac de cada uma das classes fazendo:
   $ac = \frac{a}{k} \Leftrightarrow$ amplitude das classes $= \frac{\text{amplitude}}{\text{Nº de classes}}$.
4. Construir as classes c_i da seguinte forma:

   $c_1 = [\text{mínimo; mínimo} + ac[$,

   $c_2 = [\text{mínimo} + ac; \text{mínimo} + 2 \cdot ac[$,

   ...

   $c_k = [\text{mínimo} + (k-1) \cdot ac; \text{mínimo} + k \cdot ac]$.

Exercício 2: O Sr. Nobre decidiu dedicar-se à criação de leitões, que vende quando atingem os dois meses de idade e pesam mais de 9 kg. Pretendendo fazer um estudo sobre os lucros obtidos com essa actividade, resolveu pesar 60 leitões com dois meses de idade, tendo obtido os seguintes resultados: (Afonso e Nunes. 2011, p.14)

4,1	5,8	5,8	6,1	6,7	7,0	7,0	7,5	7,5	7,5
7,7	8,2	8,8	9,0	9,0	9,1	9,1	9,1	9,2	9,2
9,2	9,2	9,4	9,4	9,7	9,8	10,0	10,0	10,2	10,2
10,3	10,6	10,6	10,8	10,9	10,9	11,6	11,7	11,8	11,8
11,8	11,8	12,0	12,2	12,3	12,5	12,6	12,7	8,3	9,4
11,0	14,0	8,5	9,5	11,1	14,2	8,7	9,5	11,1	14,8

A construção de uma tabela de distribuição de frequência pode ser feita, por cálculo manual, utilizando a regra de Sturges:

1. Determinar o número k de classes:

   $$k = \left[\frac{\ln n}{\ln 2}\right] + 1 = \left[\frac{4.0943}{0.6931}\right] + 1 = 6.909 = [5.907] + 1 = 5 + 1 = 6$$

   Logo, k será 6 classes.

2. Determinar a amplitude a dos dados:

   $a =$ máximo - mínimo $= 14.8 - 4.1 = 10.7$

3. Definir a amplitude ac de cada classe

   $$ac = \frac{a}{k} = \frac{10.7}{6} = 1.7833 = 1.8$$

A distribuição de frequência é a seguinte:

Tabela 2.4: Distribuição de frequência dos pesos dos leitões

Classe	Frequências
$[4.1 - 5.9[$	3
$[5.9 - 7.7[$	7
$[7.7 - 9.5[$	18
$[9.5 - 11.3[$	17
$[11.3 - 13.1[$	12
$[13.1 - 14.9[$	3
Total	60

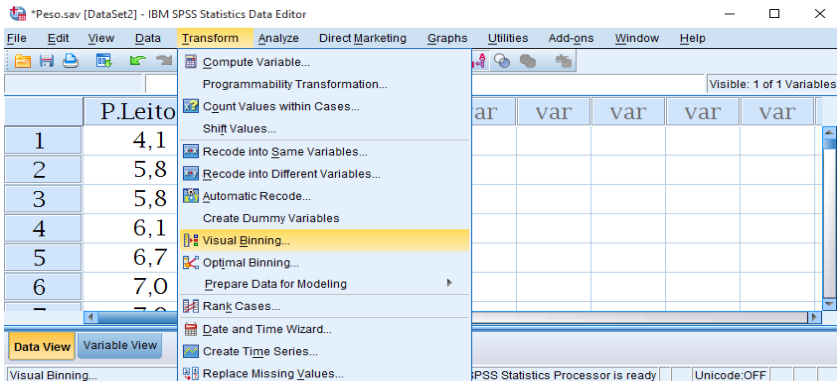
Resoluções com o SPSS e o R

Os dados que constam do exercício 2 podem ser apresentados na tabela de distribuição de frequências ou agrupados em classes com o SPSS e o **R**. A seguir mostra-se o resultado do processamento de dados.

Resolução com SPSS

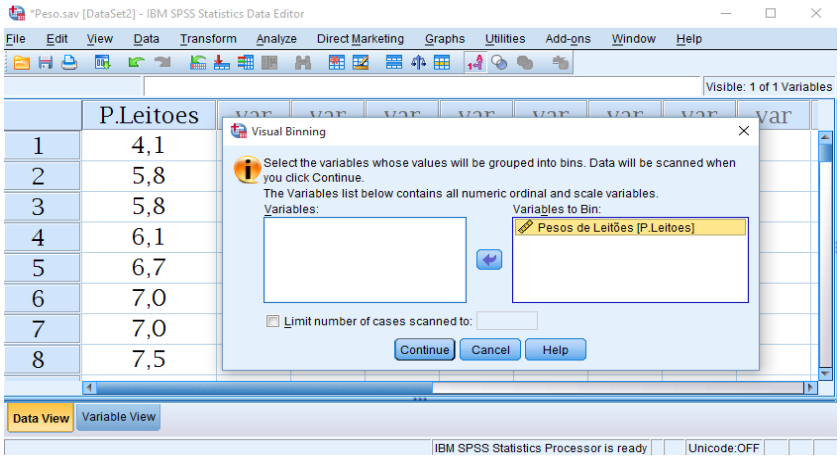
Para iniciar, deve abrir-se a janela **Variable View**, Em seguida, cria-se a variável com o nome adequado, por exemplo, *P.Leitoes*. Entretanto, guarda-se o ficheiro com o nome escolhido (*Peso.sav)*. Depois de definir a variável, abre-se a janela **Data View**, para introduzir os dados dos 60 leitões. Depois, basta seguir os seguintes processos de execução:

- Transform /Visual Binnig ...



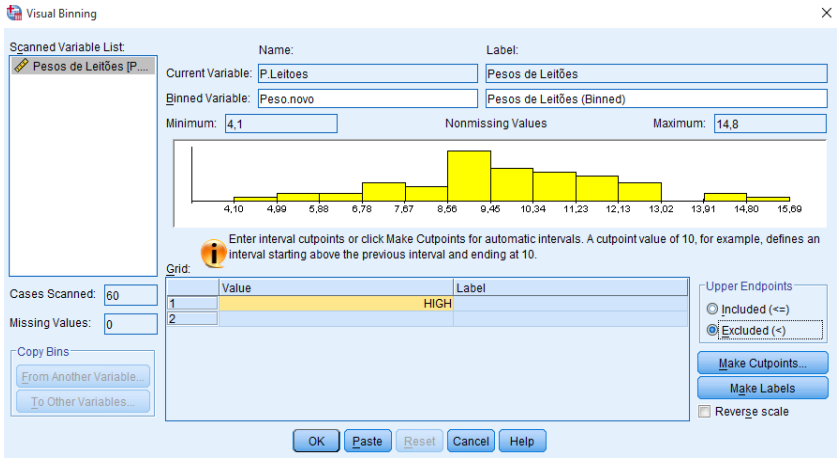
- Variables to Bin: Pesos de Leitões

  Depois de clicar em **visual Binning** aparece a janela **visual Binning**, mover a variável *Pesos de Leitões (P.Leitoes)*, a ser analisada, da coluna da esquerda **Variables:** para a da direita **Variables to Bin:**, como mostra a imagem abaixo.



- Em seguida, clicar no botão Continue

  Depois de executar **Continue** aparece a seguinte janela que contém a variável *P.Leitoes*, o valor mínimo e máximo dos dados, o seu histograma, o número total dos dados **Cases Scanned**, neste caso são sessenta dados, e **Missing Values** igual a zero ou seja, todos os dados estão totalmente incluídos para serem analisados por este programa.



Na mesma janela executar:

- Binned Variable: Peso.novo Deve escrever-se o nome da nova variável que é diferente do nome da primeira *P.Leitoes*

- ⊙ Excluded (<)

- Make Cutpoints···

  Depois de clicar **Make Cutpoints...**, vai aparecer a janela de **Make Cutpoints**. Nesta janela, executa-se o seguinte procedimento:

- ⊙ Equal Width Intervals

  Interval-fill in at least two fields

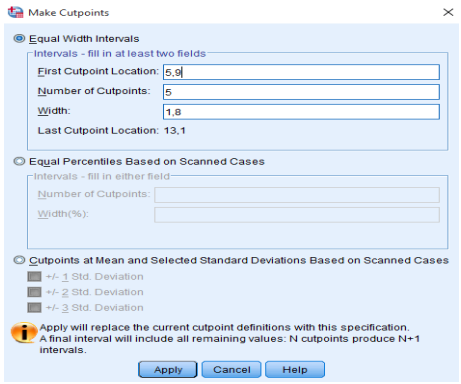
  ⇒ First Cutpoint Location : 5,9

  ⇒ Number of Cutpoints : 5

  ⇒ Width : 1,8

- Apply

  Nota: O número 1,8 é a amplitude de cada uma das classes e 5,9 é o limite superior da primeira classe. Portanto, o número 5 vai aparecer, automaticamente, depois de inserir os números 1,8 e 5,9. O resultado aparece na seguinte janela:

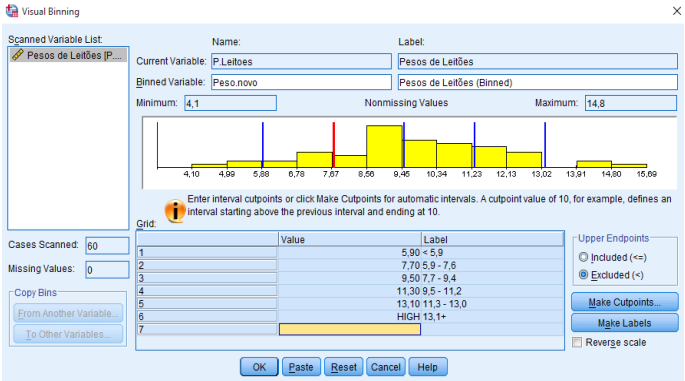


Depois de clicar **Apply**, vai aparecer a janela de **Visual Binning**. Nessa janela pode-se executar as seguintes fases:

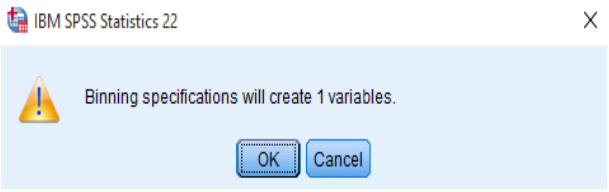
- Make label (para criar novo label automaticamente)

  Nota: Há duas maneiras de construir os intervalos de classe: de forma automática (clicando directamente em Make Labels) ou com o cursor ou seja, manualmente (modifica de imediato as barras que estão sobre o histograma). Neste caso que estamos a estudar os limites das classes são feitos automaticamente.

- OK

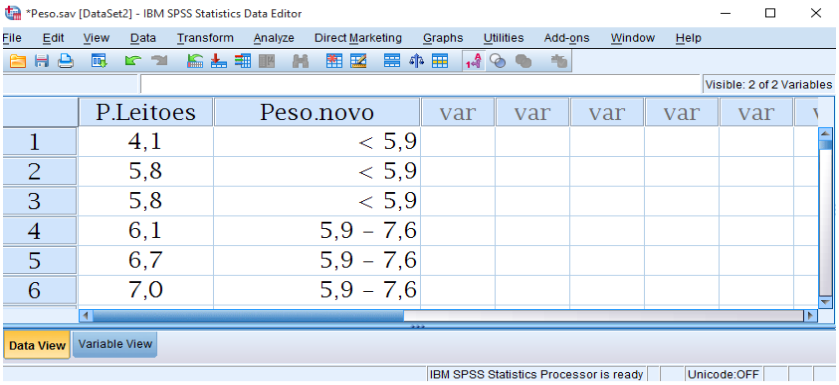


Depois de clicar **OK**, aparece a seguinte janela **Binning specifications will create 1 variables**:

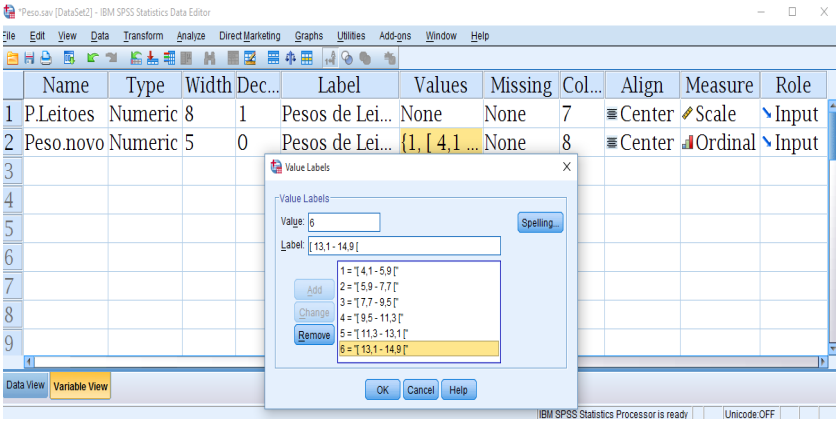


Pode-se clicar **OK**, se se quiser criar e guardar Sinxtaxe, caso contrário, clicar **Cancel**. Neste caso foi escolhido **OK**.

Depois de clicar **OK** vai aparecer o novo nome da variável **Peso.novo** (ver a janela a seguir).



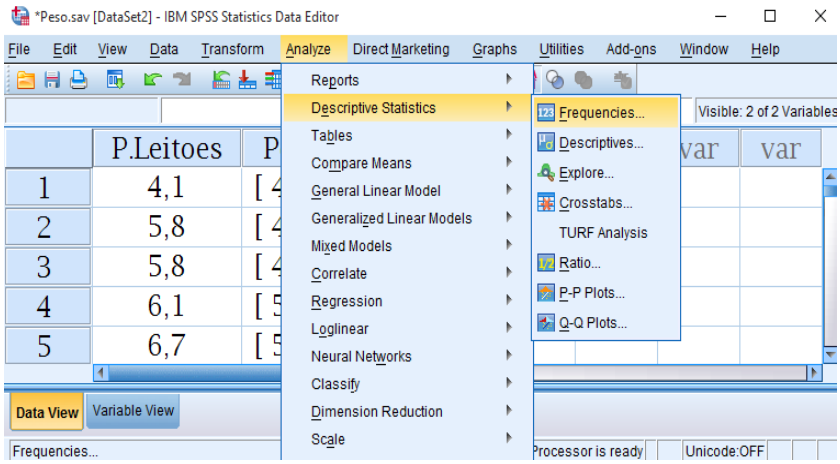
As etiquetas aos valores ou limites das classes da variável *Peso.novo* é atribuída na coluna **Values**, para terminar basta seleccionar **OK**. Como mostra a seguinte janela:



O processo de **Visual Binning**, usado até aqui, será utilizado como modelo para outros exercícios apresentados ao longo deste trabalho.

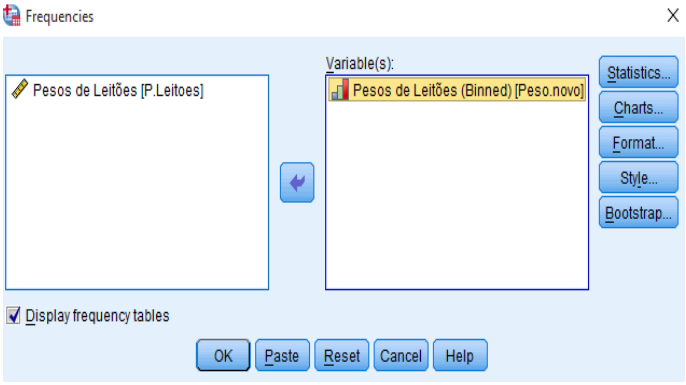
Para ver o resultado da distribuição de frequências, pode ser feito o seguinte procedimento:

- Analize/Descriptive Statistics/Frequencies. . .



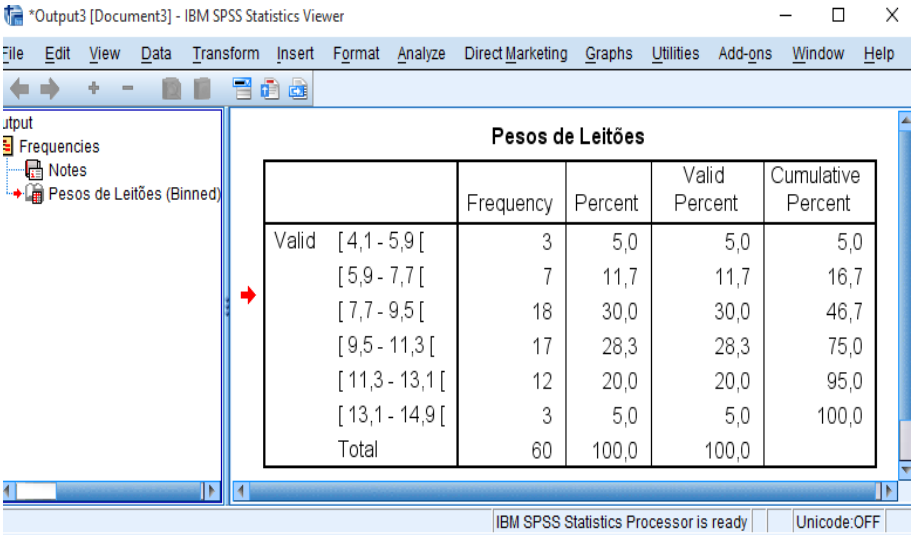
- Variable(s): Pesos de leitões (Binned . . .

- ☑ Display frequency tables



- OK.

A tabela seguinte é o resultado da distribuição de frequências feita pelo SPSS:



Pesos de Leitões

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	[4,1 - 5,9 [	3	5,0	5,0	5,0
	[5,9 - 7,7 [	7	11,7	11,7	16,7
	[7,7 - 9,5 [	18	30,0	30,0	46,7
	[9,5 - 11,3 [	17	28,3	28,3	75,0
	[11,3 - 13,1 [	12	20,0	20,0	95,0
	[13,1 - 14,9 [	3	5,0	5,0	100,0
	Total	60	100,0	100,0	

Resolução com R

A linguagem **R** também pode ser utilizada, como ferramenta, para construir a tabela de distribuição de frequências. Como os dados do *exercício 2* foram introduzidos sob a forma decimal (virgulas), para facilitar que o **R** os possa analisar deve colocar-se os pontos para substituir as virgulas. Para construir a tabela de distribuição de frequência é necessário instalar primeiro o pacote **fdth** (Frequency Distribution Tables, Histograms and Poligons) disponível para o **R**.

Os passos seguintes são as fases de utilização dos comandos desse programa:

```
> peso = c(4.1, 5.8, 5.8, 6.1, 6.7, 7, 7, 7.5, 7.5, 7.5, 7.7, 8.2, 8.8, 9, 9, 9.1, 9.1, 9.1,
+ 9.2, 9.2, 9.2, 9.2, 9.4, 9.4, 9.7, 9.8, 10, 10, 10.2, 10.2, 10.3, 10.6, 10.6, 10.8, 10.9,
+10.9, 11.6, 11.7, 11.8, 11.8, 11.8, 11.8,12, 12.2, 12.3, 12.5, 12.6, 12.7, 8.3, 9.4, 11,
+14, 8.5, 9.5, 11.1,14.2, 8.7, 9.5, 11.1, 14.8)

> peso

 [1] 4.1 5.8 5.8 6.1 6.7 7.0 7.0 7.5 7.5 7.5 7.7 8.2 8.8 9.0 9.0 9.1 9.1 9.1 9.2 9.2 9.2 9.2
[23] 9.4 9.4 9.7 9.8 10.0 10.0 10.2 10.2 10.3 10.6 10.6 10.8 10.9 10.9 11.6 11.7 11.8
[40] 11.8 11.8 11.8 12.0 12.2 12.3 12.5 12.6 12.7 8.3 9.4 11.0 14.0 8.5 9.5 11.1 14.2
[57] 8.7 9.5 11.1 14.8

> tabela= fdt(peso, start=4.1, end=14.9, h=1.8)

> tabela
```

Class limits	f	rf	rf(%)	cf	cf(%)
$[4.1, 5.9)$	3	0.05	5.00	3	5.00
$[5.9, 7.7)$	7	0.12	11.67	10	16.67
$[7.7, 9.5)$	18	0.30	30.00	28	46.67
$[9.5, 11.3)$	17	0.28	28.33	45	75.00
$[11.3, 13.1)$	12	0.20	20.00	57	95.00
$[13.1, 14.9)$	3	0.05	5.00	60	100.00

De um modo geral, os resultados apresentados pelos dois programas são iguais. O número de classes é 6; o total de leitões igual a 60; os limites das classes inferiores são 4.1, 5.9, 7.7, 9.5, 11.3, 13.1; os limites das classes superiores são 5.9, 7.7, 9.5, 11.3, 13.1, 14.9; a amplitude de classe é 5.9 - 4.1 = 1.8.

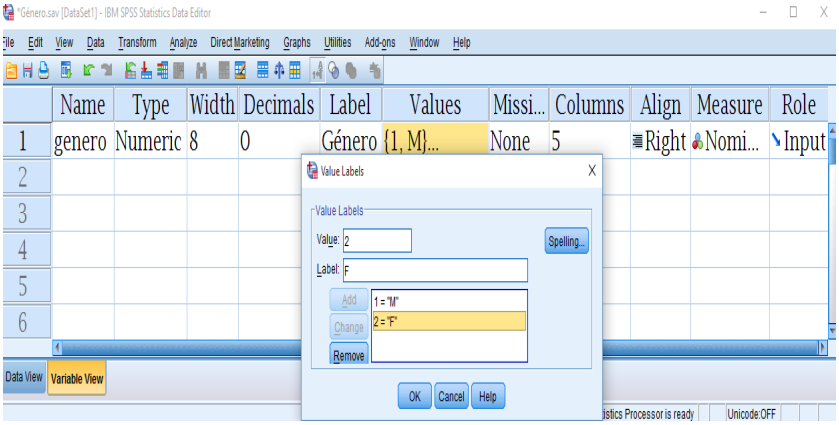
Exercício 3: O género de vinte pessoas escolhidas ao acaso foi:

M	F	F	M	M	M	F	F	M	M
F	M	F	F	M	M	M	M	F	M

Elabora a tabela de frequências absolutas associada (ME-TL. 2014, p.122. Tarefa 40). M: Masculino, F: Feminino.

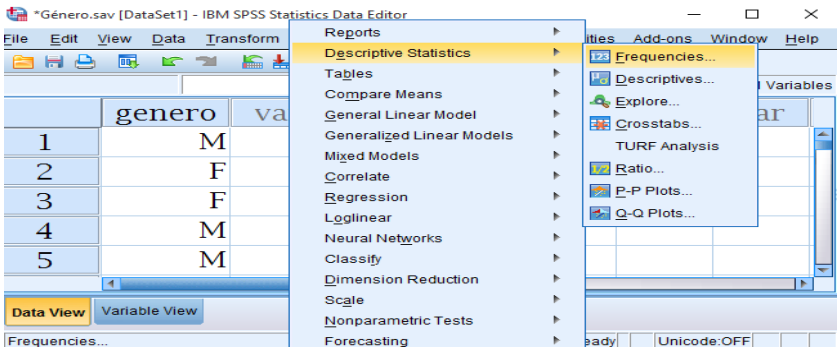
Resolução com SPSS

Estas informações podem ser representadas por números e isso deve ser feito ou transformado na caixa de **Values** da janela do **Variable View**. Esta transformação vai facilitar a análise de SPSS. Essa codificação pode ser feita do seguinte modo: 1 corresponde a Masculino, 2 corresponde a Feminino (ver janela sobre posto o Data Editor). Seguidamente clicar no **OK**:

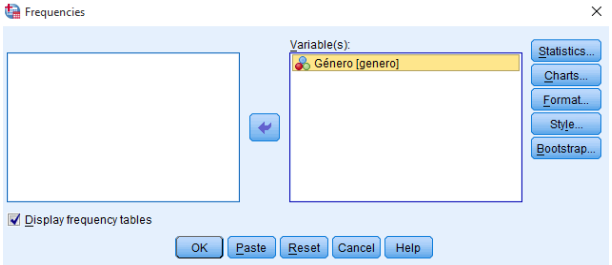


Depois da introdução dos dados no SPSS podem ser executados os seguintes comandos:

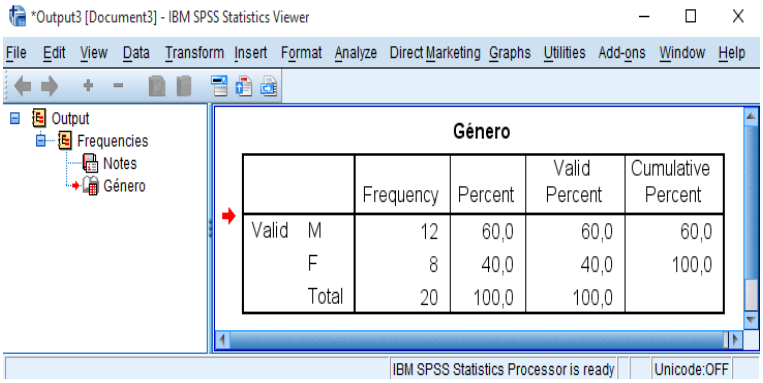
- Analyze/Descriptive Statistics/Frequencies ...



- Variable(s): Género
- ☑ Display frequency tables
- OK



Resultado:



Resolução Com R

> G=c("M", "F", "F", "M", "M", "M", "F", "F", "M", "M", "F", "M", "F", "F", + "M", "M", "M", "M", "F", "M")

> G

[1] "M" "F" "F" "M" "M" "M" "F" "F" "M" "M" "F" "M" "F" "F" "M"

[16]"M" "M" "M" "F" "M"

> f.a = table(G)

> f.r = f.a/sum(f.a)

> f.p = 100 * f.r

> f.cm = cumsum(f.a)

> genero = cbind(f.a, f.r, f.p, f.cm)

> genero

	f.a	f.r	f.p	f.cm
F	8	0.4	40	8
M	12	0.6	60	20

No total dos 20 alunos, 8, ou seja, 40% são do sexo feminino e 12, ou seja, 60% são do sexo masculino.

Capítulo 3

Construções gráficas

O gráfico é uma imagem que mostra visualmente os dados sob a forma de números. Normalmente estes dados vêm de tabela. Em geral, vão sendo utilizados alguns dos gráficos para representar conjuntos de dados:

3.1 Gráfico circular

O gráfico circular é constituído por um círculo dividido em tantas fatias quantas as categorias de variável (Afonso e Nunes, 2011, p.17). O tamanho das fatias é determinado pelo número ou percentagem/proporção de observações nas categorias, i.e., pelas frequências absolutas, n_i, ou pelas relativas, f_i. Este gráfico é utilizado para dados qualitativos. Na figura 3.1 apresenta-se um exemplo genérico de um gráfico circular.

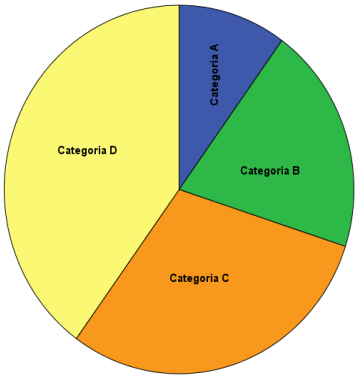
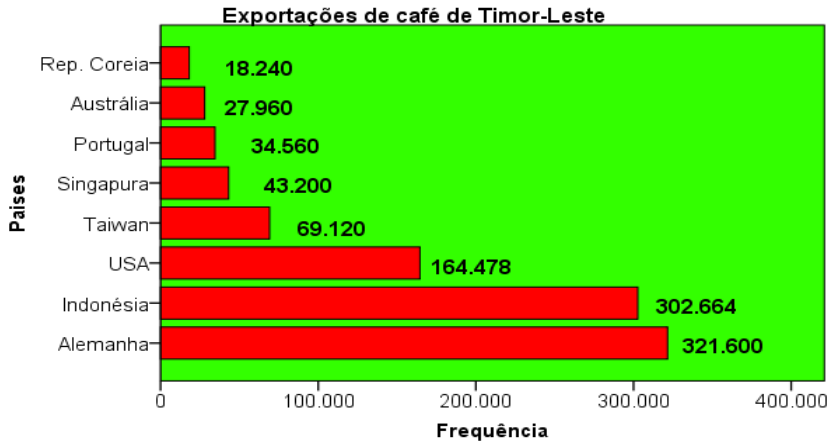


Figura 3.1: Gráfico circular

Exercício 4

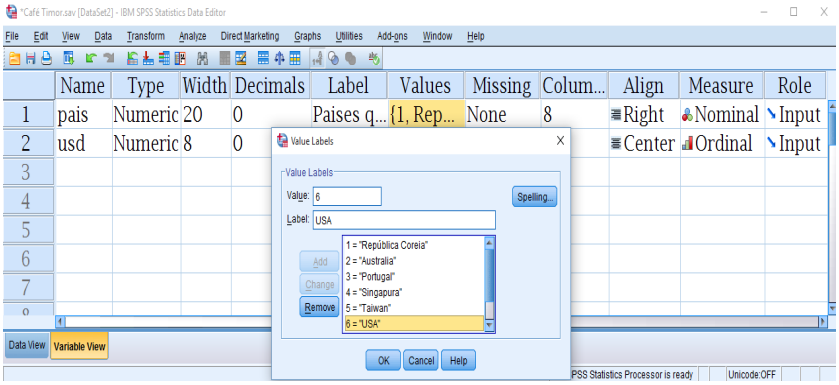
Considera o seguinte gráfico de barras referente às exportações de café de Timor-Leste, no 2º trimestre de 2010, em USD. (Fonte: DNE-Indicadores Estatísticos Trimestrais 2º Trimestre 2010).

Constrói o diagrama circular que represente a informação dada neste gráfico de barra. (ME-TL. 2014, p.164-165. Enunciado número 9 parte 9.4)



Resolução com SPSS

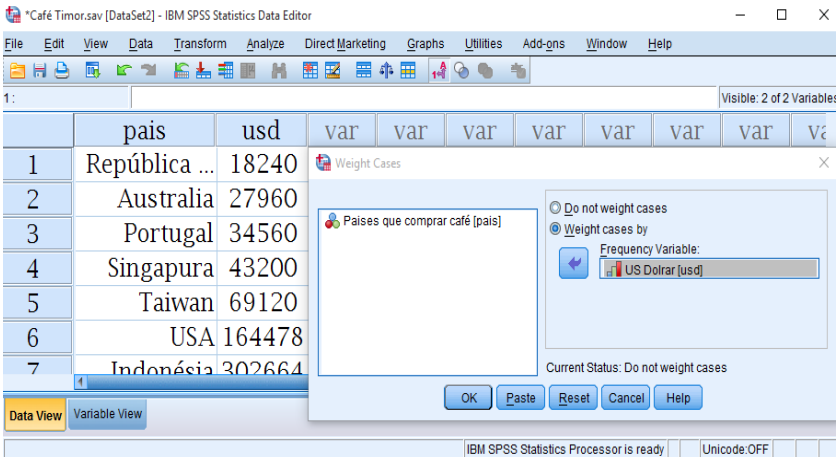
Os dados do gráfico do **exercício 4** são formados por dois tipos de informação: países de destino (tipo qualitativo) e montante de dinheiro por país (tipo quantitativo). No SPSS a cada país deve corresponder o seu montante de dinheiro e os países devem ser colocados por categorias. Para acabar, clicar **OK**. Observe a figura seguinte:



Após executar **OK**, deve fazer-se o processo de **Weight Cases** para estabelecer a correspondência entre cada país e o seu montante de dinheiro, através da execução dos seguintes comandos:

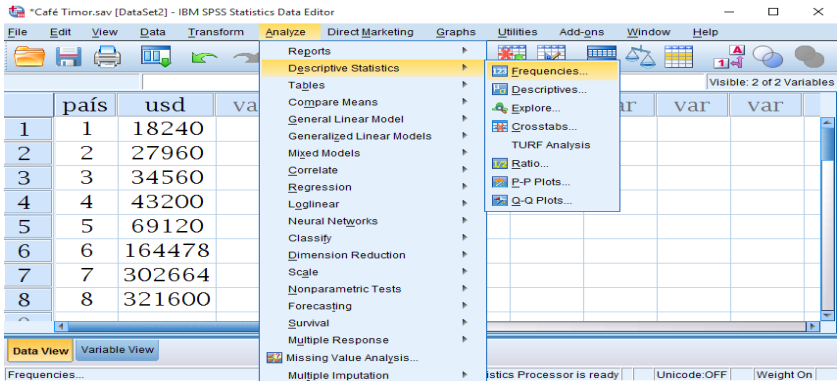
- Data/Weight Cases
- ⊙ Weight Cases By: USD Dollar (usd)
- OK

O **Weight Cases** aparece na janela seguinte (Janela sobre posta do Data Editor):

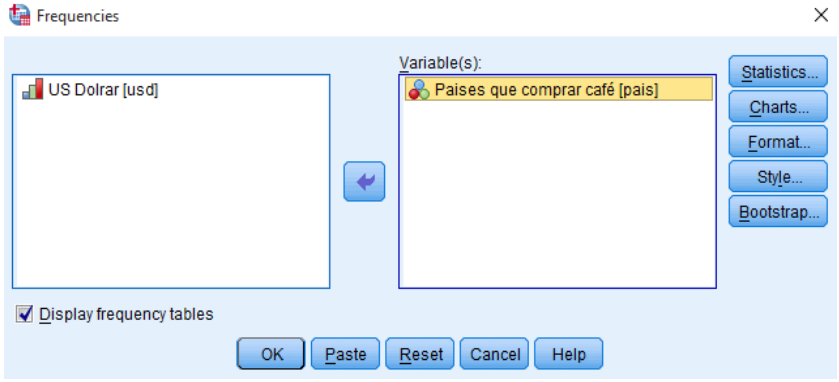


Depois de executar **OK**, já se pode efectuar as fases de construção do gráfico circular como se segue:

- Analyze/Descriptive Estatistics/Frequencies ...

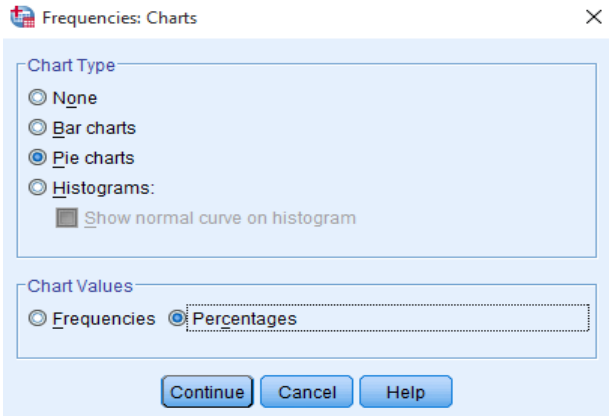


- Variable(s): pais
- ☑ Display frequency tables



- Chart:
  - Chart Type: ⊙ Pie Chart

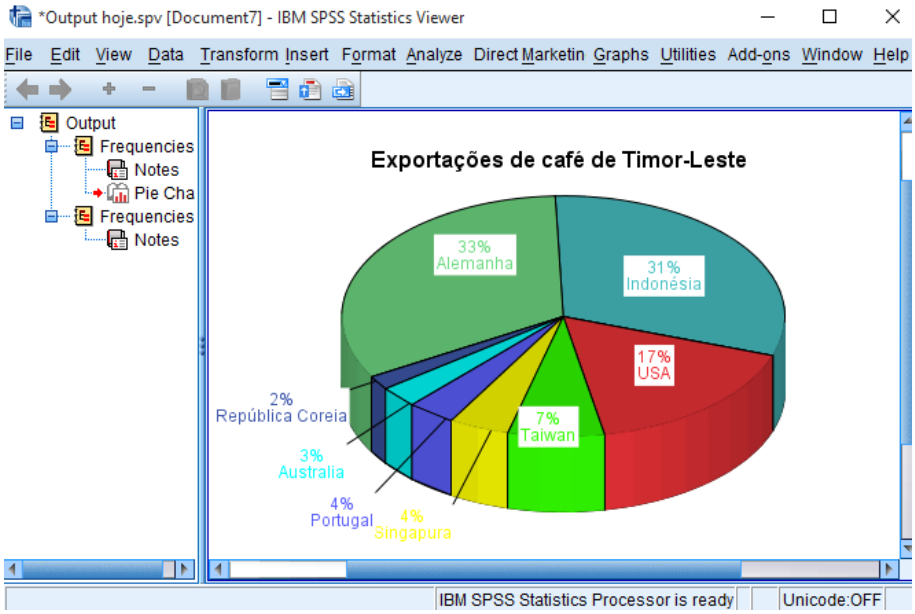
- Chart Values: ⊙ Frequencies



► Continue

► OK

Após executar **OK**, vai aparecer uma janela que contém o gráfico. Nessa janela podem ser acrescentados vários elementos que farão parte do gráfico que aparece na janela **Chart Editor**. O resultado final é o seguinte:



Resolução com R

Para construir o gráfico de **Pie 3D** é necessário instalar primeiro o pacote **plotrix** disponível para o **R**. Por último, pode inserir os dados em **R**, como se mostra a seguir:

```
> caf =c(18240, 27960, 34560, 43200, 69120, 164478,302664, 321600)
> caf
[1] 18240  27960  34560  43200  69120  164478  302664  321600
```

> percentagem=round(caf/sum(caf) ∗ 100, digits=0)

> percentagem

[1] 2 3 4 4 7 17 31 33

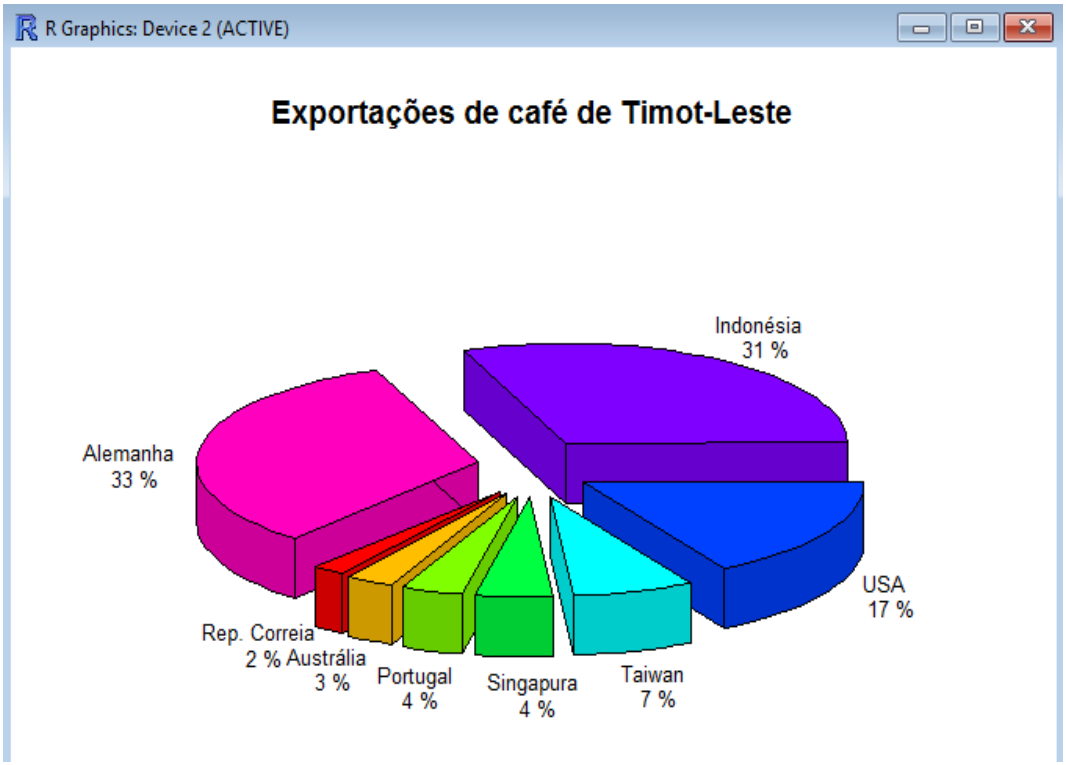
> destino=c("Rep.Correia", "Austrália", "Portugal", "Singapura", "Taiwan", "USA",
+ "Indonésia", "Alemanha")

> destino

[1] "Rep.Correia" "Austrália" "Portugal" "Singapura" "Taiwan" "USA" "Indonésia"
[8] "Alemanha"

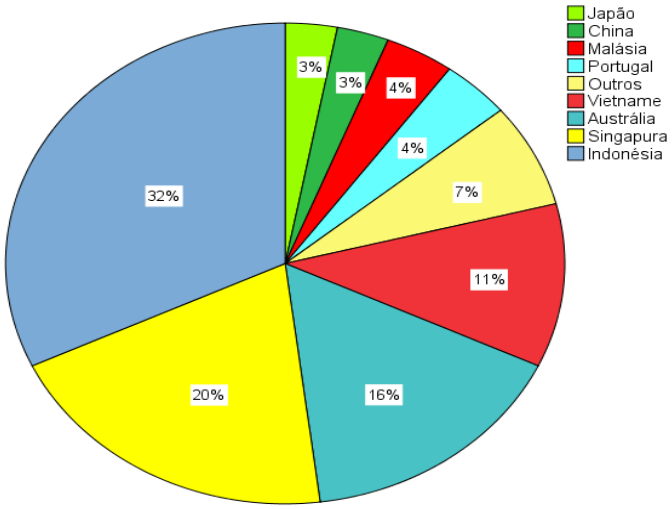
> labels=paste(destino,"\n",percentagem,"%", sep=)

pie3D(destino, main = "Exportações de café de Timot-Leste", labels = labels, explode = 0.2, labelcex = 0.8, start = 4)



3.2 Gráfico de barras

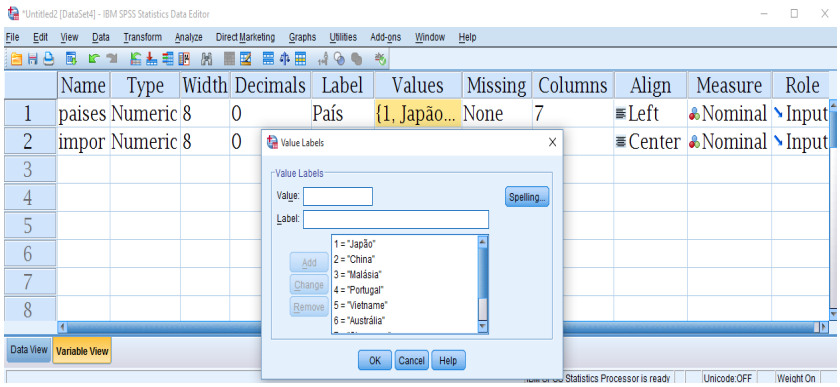
Exercício 5: Considere o seguinte gráfico circular correspondente à estrutura das importações em Timor-Leste nos primeiros 8 meses de 2009 (Fonte: DNE a partir de dados das alfandegas de TL). Constrói o gráfico de barras de frequências relativas com base nos dados de gráfico circular (ME-TL, 2014. p.164. Enunciado número 8 parte 8.2)



Resolução com SPSS

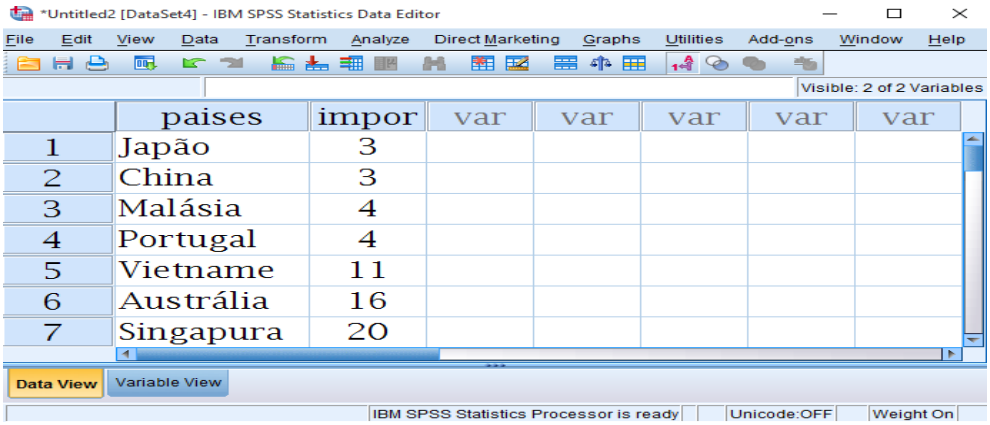
Observando o gráfico circular do exercício 5, verifica-se que existe: uma variável representa o países e a percentagem das importações. Os pontos seguintes explicam como se introduz estes dados neste programa:

- A variável país deve ser codificada como mostra a figura seguinte:



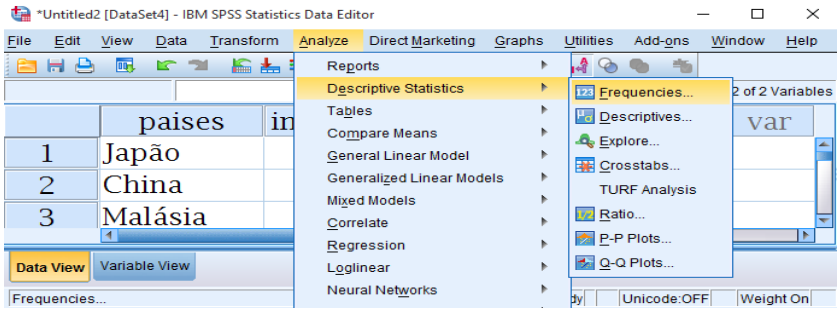
- OK

Os nomes dos países e a percentagem das importações são apresentados abaixo:

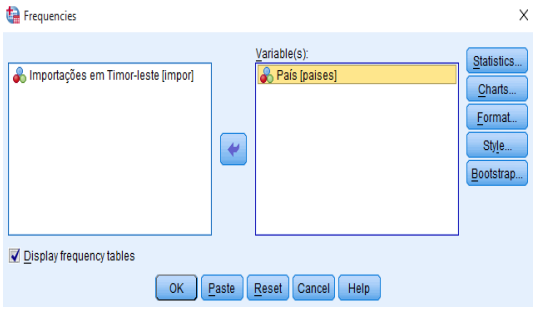


Depois de codificar a variável e introduzir os dados no programa, a fase seguinte deve ser ponderar ou seja **Weigth Cases** os dados dessas duas variáveis. Para fazer isto pode seguir a instrução do exercício 1 fase 2. Neste caso, a variável impor é que foi ponderada. Por último, pode ser feita a construção do gráfico de barras do seguinte modo:

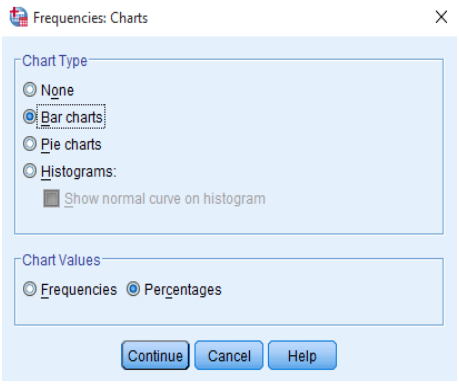
- Analize/Descriptive/Frequencies...



- Variable(s): paises
- ☑ Display frequency tables



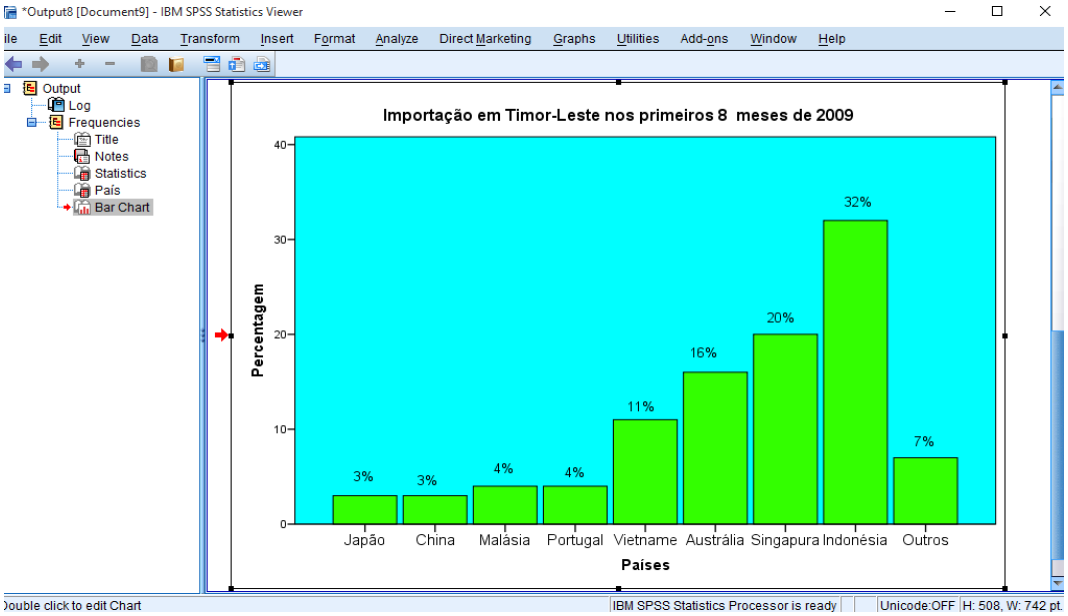
- Chart:
  - Chart Type: ⊙ Bar Chart
  - Chart Values: ⊙ Percentages



- Continue

► OK

Para modificar o gráfico, clicar duas vezes na área do gráfico e na janela de **Properties**. Depois, pode efectuar as modificações desejadas. A janela seguinte mostra o resultado obtido por este programa:



Resolução com R

Para fazer o gráfico de barras em **R**, basta executar os seguintes comandos:

> impor= c(0.03, 0.03, 0.04, 0.04, 0.11, 0.16, 0.2, 0.32, 0.07)

> impor

[1] 0.03 0.03 0.04 0.04 0.11 0.16 0.20 0.32 0.07

> paises=c("Jap", "Chin", "Malá", "Port", "Viet", "Aus", "Sing", "Indo", "Out")

> paises

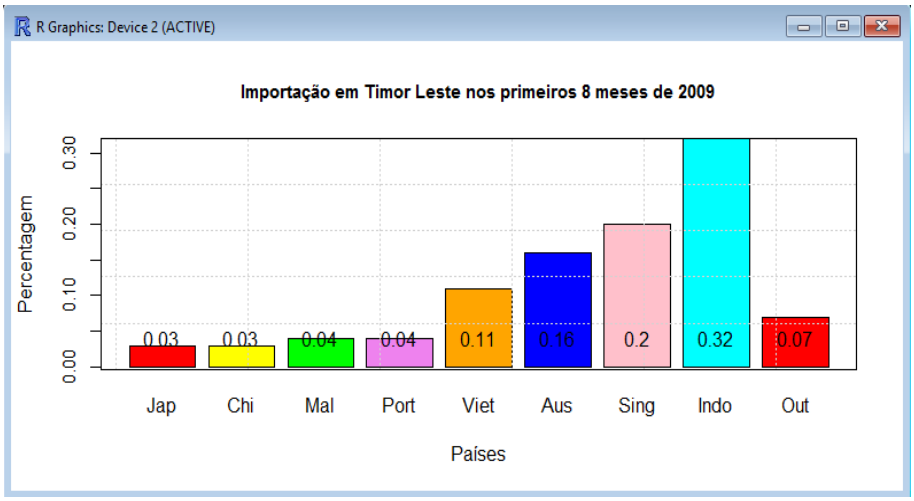
[1] "Jap" "Chin" "Malá" "Port" "Viet" "Aus" "Sing" "Indo" "Out"

> colors =c("red", "yellow", "green", "violet", "orange", "blue", "pink", "cyan")

> colors

[1] "red" "yellow" "green" "violet" "orange" "blue" "pink" "cyan"

> barplot (impor, names.arg = paises, cex.main = 0.9, cex.axis = 0.9, ylab = "País", xlab = "Importações ", main = "Importação em Timor-Leste nos primeiros 8 meses de 2009 ", col = colors)



3.3 Gráficos de frequências acumuladas

Exercício 6: O dono de um restaurante contou o número de almoço servidos durante 24 dias, os resultados foram o seguinte:

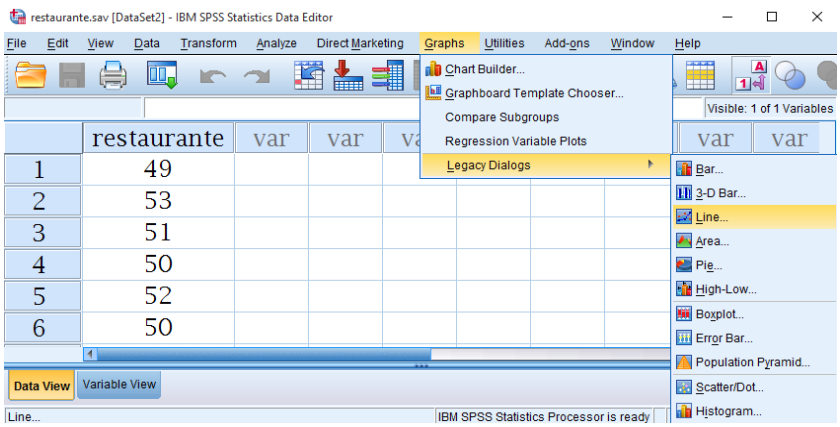
49	53	51	50	52	50	52	50	50	51	49	53
50	49	51	48	51	50	50	51	52	50	51	49

Constrói a função cumulativa correspondente, usando frequências absoluta acumuladas. (ME-TL. 2014, p.126. Tarefa 44)

Resolução com SPSS

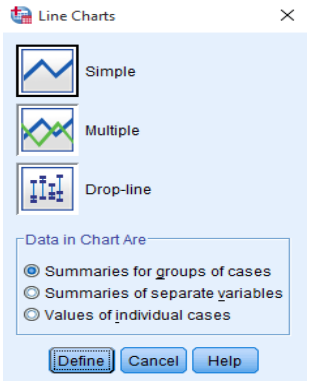
Depois de inserir e representar o conjunto dos dados por uma variável (restaurante), podem ser executados os seguintes comandos no SPSS para construir o gráfico de frequências acumuladas.

- Graphs/Legacy dialog/Line ...

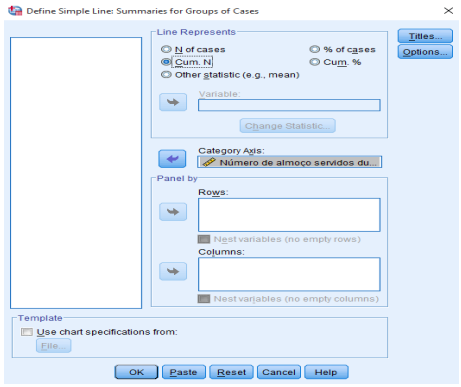


- Simple:

- Data in Chart Are: ⊙ summaries for groups of cases
- Define



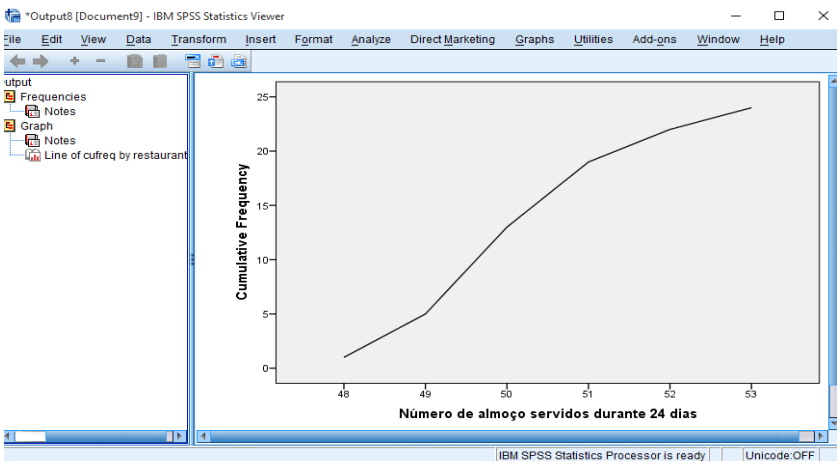
- Line represents: ⊙ Cum.N
- Categori Axis: Número de almoço servidos...



- OK

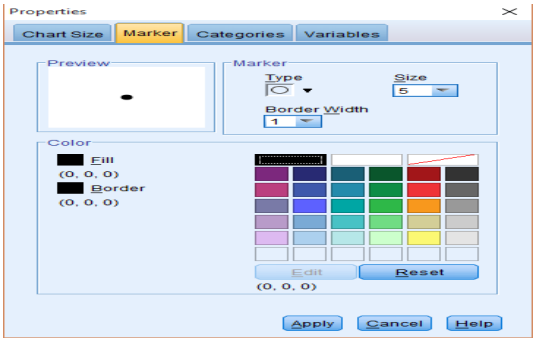
Depois de clicar **OK**, aparece uma janela, que apresenta um gráfico de linha. Nela pode ser feito o gráfico de Frequência Acumulada, através da execução dos seguintes passos:

1. Clicar duas vezes na área do gráfico. Aparece logo a janela **Chart Editor**.



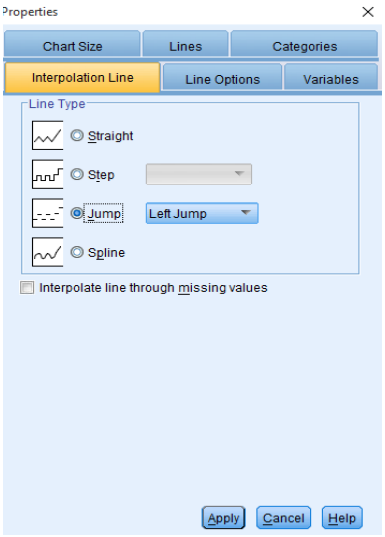
2. Nesta janela, basta fazer um clique no botão direito do rato, em qualquer área do gráfico e, seguidamente, escolher **Add Markers**. Depois de executar esta opção vai aparecer logo a janela **Properties**. Nela pode escolher:

- Marker para dar ou modificar a cor dos pontos segundo as necessidades
- Apply
- Close

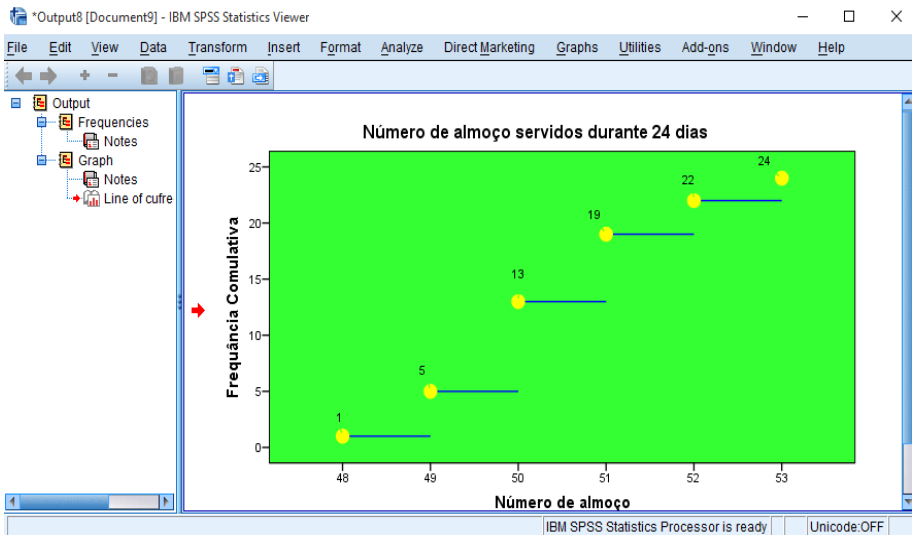


Depois de clicar em **Close**, vai aparecer logo um gráfico no qual ainda é preciso acrescentar os degraus, por isso são necessárias mais formatações. Para editar este gráfico, basta fazer duplo clique sobre a linha recta. Em seguida vai aparecer a janela **Properties** e nela pode executar os seguintes comandos:

- Interpolation Line
- Line type: Jump/Left Jump
- Apply/ Close



A figura seguinte é o Gráfico de frequência acumulada do exercício 4.



Solução com R

O programa de **R** pode ser utilizado para construir vários tipos de gráficos de função cumulativa. Neste caso seria interessante de construir este gráfico utilizando a frequência relativa acumulativa como mostra nos seguintes:

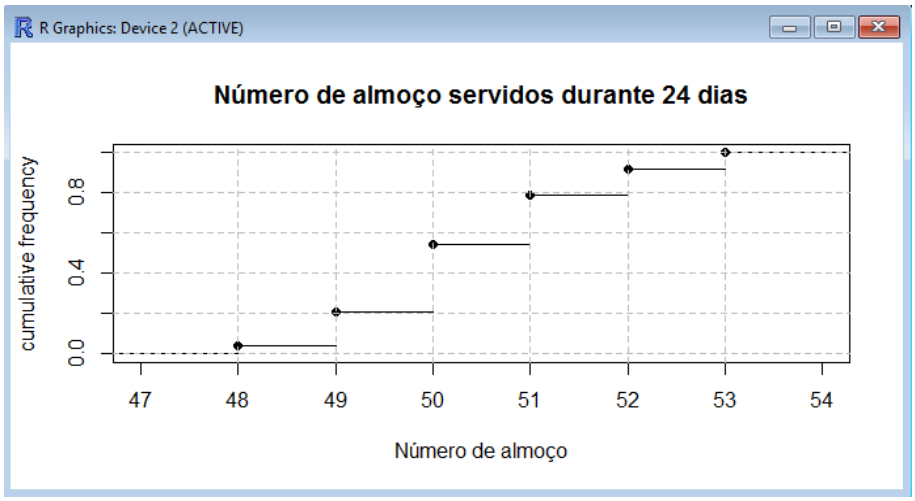
> restaurante =c(49, 53, 51, 50, 52, 50,52, 50, 50, 51, 49, 53,50, 49, 51, 48, 51, 50, + 50, 51, 52, 50, 51, 49)

> restaurante

> [1] 49 53 51 50 52 50 52 50 50 51 49 53 50 49 51 48 51 [18] 50 50 51 52 50 51 49

> plot(ecdf(restaurante), xlab = "Número de almoço ", ylab = "Cumulative frequency ", main = "Número de almoço servidos durante 24 dias ")

> abline(h = seq(0, 1, 0.2), v = restaurante, col = "gray", lty = 2)



A partir das figuras finais apresentadas pelo SPSS e pelo **R** verifica-se que ambos apresentam um gráfico de função cumulativa de seis degraus. Nestes dados existem valores repetidos. Por isso, as diferenças de altura entre cada ponto e o seu anterior no eixo Y não são iguais. O gráfico final do SPSS apresenta a frequência absoluta acumulada enquanto o do **R** mostra a frequência relativa acumulada.

3.4 Histograma

Os dados que foram agrupados em classes de frequências podem ser apresentados sob a forma de um histograma. Estes são gráficos de barras onde a largura de cada barra representa a amplitude da classe e a altura corresponde à frequência absoluta ou quantidade dos elementos que pertencem a esta classe.

3.5 Polígono de frequências

Polígono é um gráfico semelhante ao histograma. A diferença é que o histograma é um gráfico formado por várias barras, enquanto o polígono é formado pela linha recta que une os pontos coordenados. Cada ponto tem como coordenadas o ponto médio do intervalo da classe e a frequência da classe.

3.6 Polígono de frequências acumuladas

Um polígono de frequência acumulada é um gráfico de linhas onde são representadas frequências absolutas, N_i, ou relativas, F_i, acumuladas. A frequência acumulada para valores inferior ao limite inferior da primeira classe é nula. A frequência acumulada para valores superiores ao limite superior da ultima classe é n, se forem representada as frequências N_i, ou 1, se forem representadas as frequências F_i. Afonso e Nunes (2011, p.20)

Exercício 7: Pediu-se aos alunos de uma turma 10º ano que cronometrassem o tempo gasto no percurso de casa a escola, num determinado dia. Os dados recolhidos, em minutos foram os seguintes:

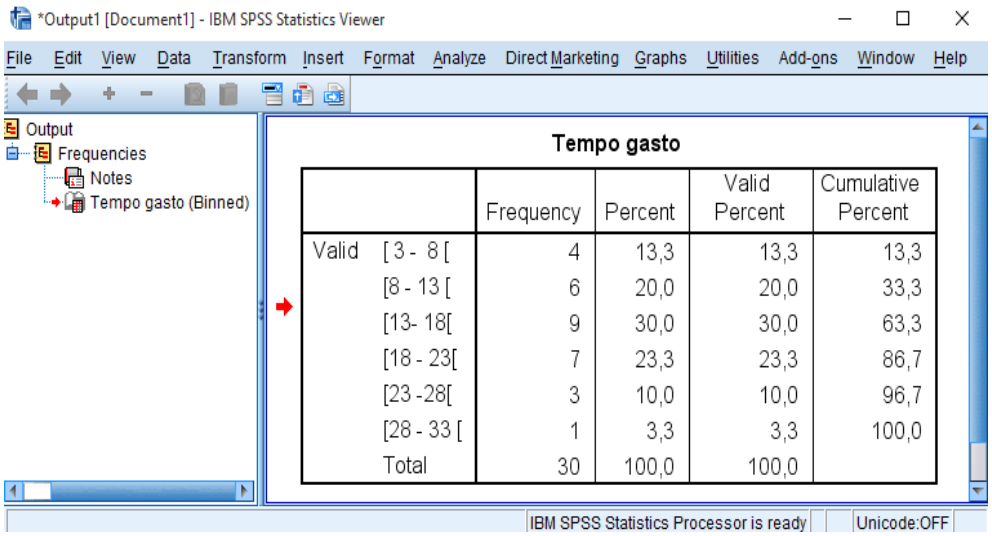
3	5	15	4	11	12	17	10	24	6
18	8	21	30	14	13	16	7	23	18
20	19	27	12	10	22	14	22	15	14

a) Agrupa os dados em classes e elabora a tabela de frequências simples.

b) Constrói o polígono de frequências e o respectivo polígono de frequências. (ME-TL. 2014, p.162. Exercícios e Problemas nº 2 parte 2.2).

Resolução com SPSS

Com o SPSS, os dados referidos no exercício 7 serão transformados em tabela de distribuição de frequências. Na sua elaboração pode seguir-se o processo usado no exercício 2, usada como modelo de resolução dos exercícios. A janela seguinte apresenta a tabela de distribuição de frequências:

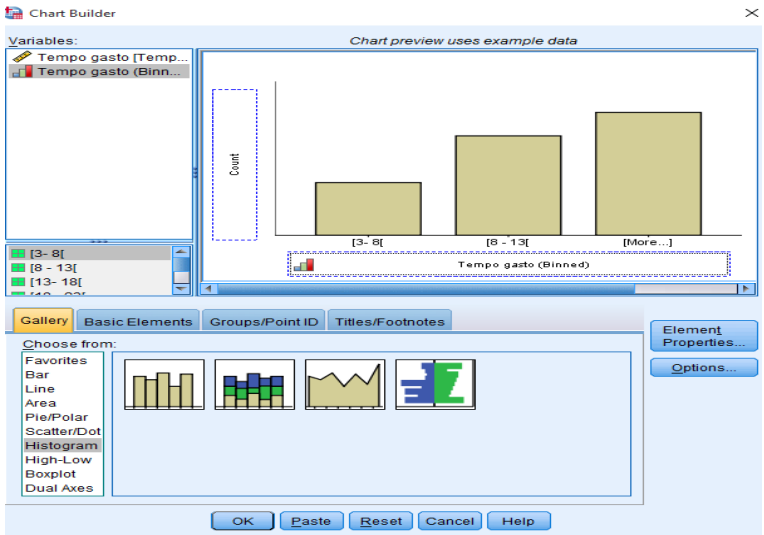


Tempo gasto

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	[3 - 8 [	4	13,3	13,3	13,3
	[8 - 13 [	6	20,0	20,0	33,3
	[13- 18[	9	30,0	30,0	63,3
	[18 - 23[	7	23,3	23,3	86,7
	[23 -28[	3	10,0	10,0	96,7
	[28 - 33 [	1	3,3	3,3	100,0
	Total	30	100,0	100,0	

Para fazer o histograma, pode executar-se os seguintes comandos:

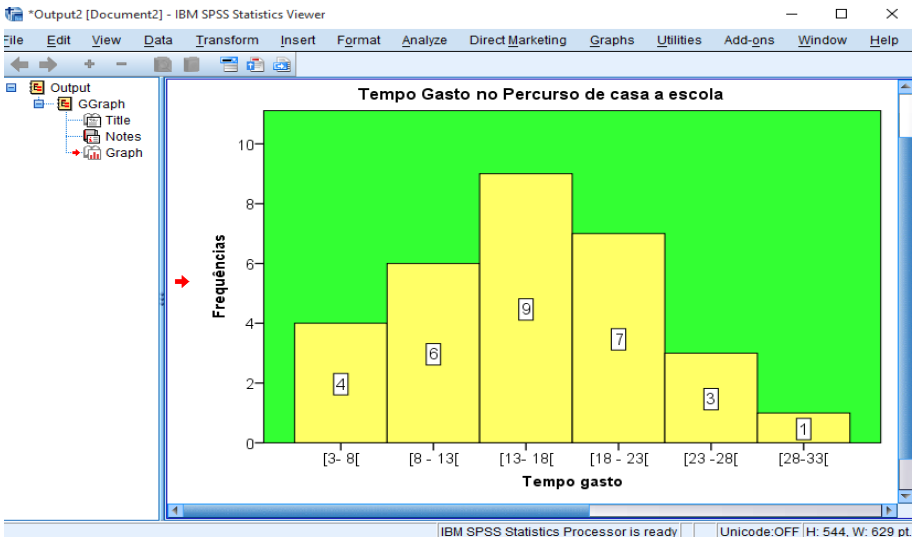
- Graphs/Chart Builder...
- Choose from: seleccionar **Histogram**. Em seguida arrastar o gráfico **Simple Histogram** e colocá-lo na parte superior **Chart preview uses example data**
- Arrastar a variável Tempo gasto[binned] e colocá-la no eixo X.



O Tempo gasto[Binned] é a nova variável obtida por recodificação (Transform/Visual Binnig). Arrastar esta variável e colocá-la no eixo X do gráfico histograma. Automaticamente, cada classe dos intervalos vai ter correspondência à sua frequência de classe no eixo Y.

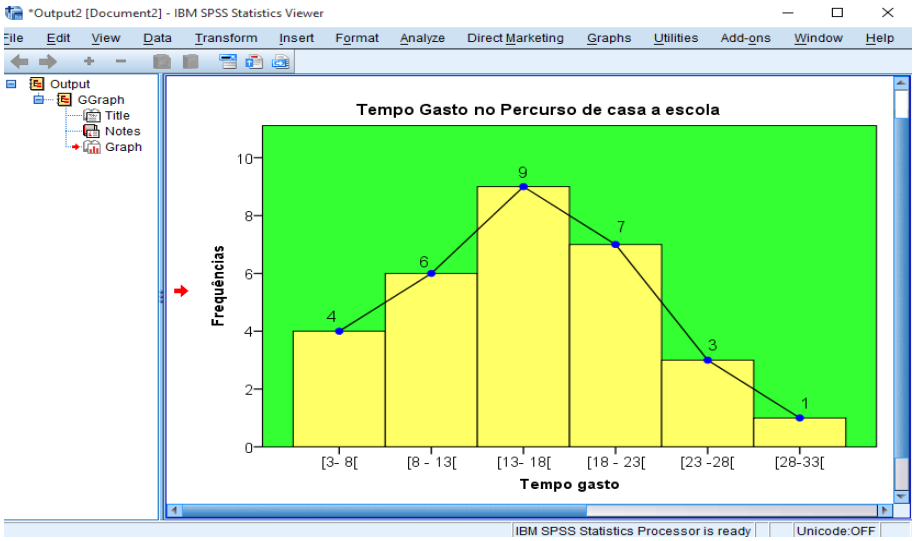
- OK

Para modificar o histograma basta clicar duas vezes na área do gráfico. Aparecerá a janela **Properties**. Nela podem ser feitas as transformações dos gráficos como se desejar.



Pode seguir-se o mesmo procedimento para construir os gráficos de: polígono de frequências e polígono de frequências acumuladas. Nas janelas seguintes são apresentados os resultados da construção:

Polígono de Frequências dos tempos gastos

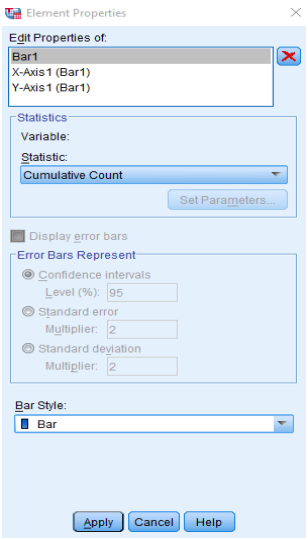


Polígono de Frequências acumulada

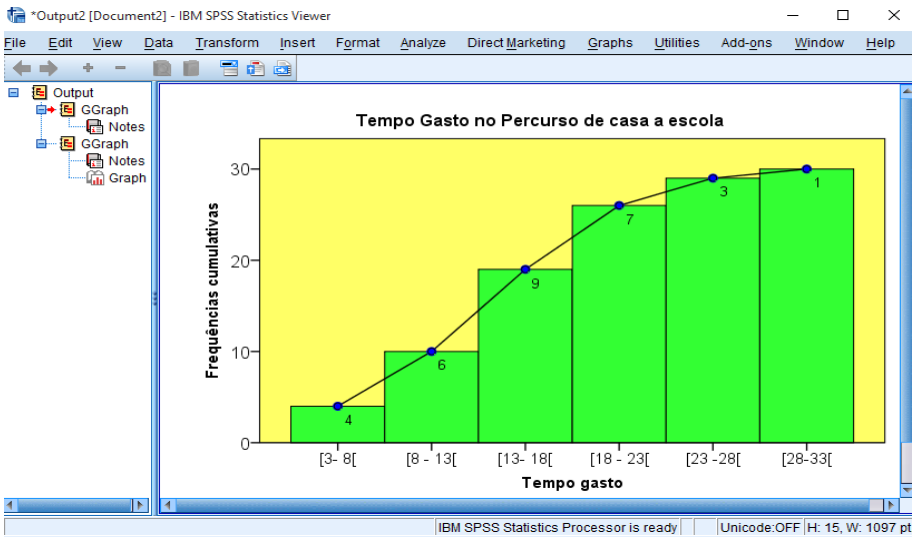
Para construir o polígono de frequência acumulada, pode executar-se os seguintes comandos:

- Graphs/Chart Builder...
- Choose from: seleccionar **Histogram**. Em seguida arrastar o gráfico **Simple Histogram** e colocá-lo na parte superior **Chart preview uses example data**
- Arrastar a variável Tempo gasto[binned] e colocá-la no eixo X.

  Na janela **Element Properties** executar:
- Statistic: Cumulative Count
- Bar Style: Bar
- Apply



A figura seguinte é o gráfico de polígono de frequências acumulada



Resolução pela linguagem R

Antes de fazer os histogramas no **R**, os dados deste exercício devem ser transformados em tabela de distribuição de frequências. Neste caso, será utilizada a função Frequency Distribution Tables (**fdt**) que faz parte do **fdth-package** (Frequency distribution tables, histograms and polygons). O **fdth** pode ser instalado com os seguintes passos:

1. Clicar no package/instal package
2. Seleccionar país, neste caso foi seleccionado Portugal (Lisbon)
3. OK

   Depois de clicar em **OK**, aparece logo uma janela de package e nela pode escolher:

   (a) fdth

   (b) OK e aguardar o processo de instalação.

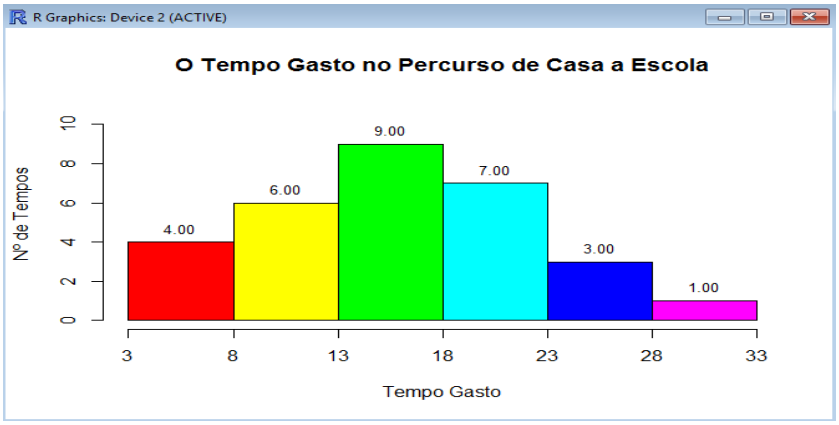
Depois de instalar esta função pode inserir os dados, a função fdt para construir a tabela de distribuição de frequências, histograma e polígono de frequência no **R**. Pode ser visto o seguinte:

```
> percurso=c(3, 15, 15, 4, 11, 12, 17, 10, 24, 6, 18, 8, 21, 30, 14, 13, 16, 7, 23, 18, +
  20, 19, 27, 12, 10, 22, 14, 22, 15, 14)

> percurso
 [1] 3 15 15 4 11 12 17 10 24 6 18 8 21 30 14 13 16 7 23 18 20 19 27 12 10 22 14 22
[29] 15 14

> tabela =fdt(percurso, start = 3, end = 33, h=5 )

> tabela
```

Class limits	f	rf	rf(%)	cf	cf(%)
$[3, 8)$	4	0.13	13.33	4	13.33
$[8, 13)$	6	0.20	20.00	10	33.33
$[13, 18)$	9	0.30	30.00	19	63.33
$[18, 23)$	7	0.23	23.33	26	86.67
$[23, 28)$	3	0.10	10.00	29	96.67
$[28, 33)$	1	0.03	3.33	30	100.00

Histograma simples

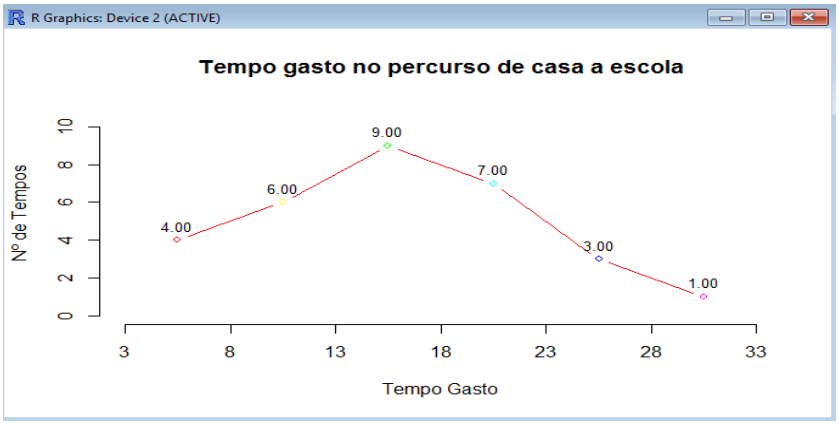
> plot(tabela, main = "Tempo gasto no percurso de casa a escola", xlab = "Tempo Gasto", ylab = "Nº de Tempos", col = rainbow(6), v = TRUE, cex = .8)



Observando os gráficos de barras produzidos pelos programas de estatística, percebe-se que existem seis barras que correspondem aos números de classes, as barras estão justapostas. Entre 3 e 33 minutos para fazer o percurso de casa à escola, o mais provável é fazê-lo de 13 a 18 minutos com um número de frequência de 10 alunos. É pouco provável que cheguem alunos à escola para além de 23 minutos.

Polígono de frequência

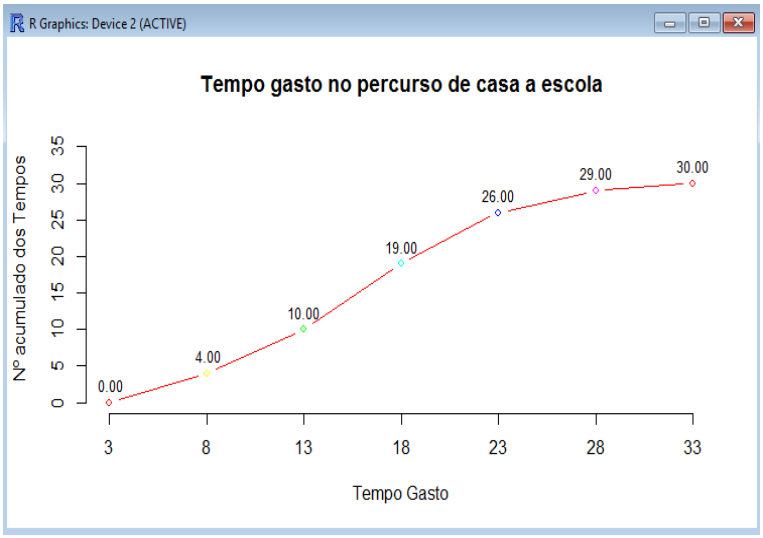
> plot(tabela,type ='fp', main = "Tempo gasto no percurso de casa a escola", col = rainbow(6), v = TRUE, cex = .8,xlab = "Tempo Gasto", ylab = "Nº de Tempos")



Polígono de frequência acumulada

> plot(tab,type = 'cfp', main = "O Tempo Gasto no Percurso de Casa a Escola", col = rainbow (6), v = TRUE, cex = .8, xlab = "Tempo Gasto ", ylab = "Nº acumulado dos Tempos")

> grid(ny = 7, col = "black", box())



Exercício 8: Num teste de 79 perguntas aplicado a 620 pessoas, o número de respostas certas está representado na tabela seguinte:

Nº de respostas corretas	Nº de pessoas
$[0, 10[$	40
$[10, 20[$	60
$[20, 30[$	75
$[30, 40[$	90
$[40, 50[$	105
$[50, 60[$	85
$[60, 70[$	80
$[70, 80[$	85

a) Constrói um histograma e um polígono de frequências absolutas da distribuição.

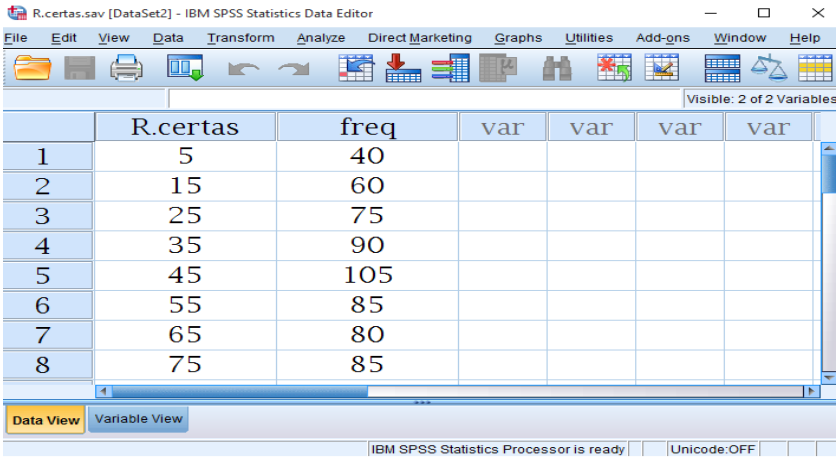
b) Constrói um histograma de frequências relativas acumuladas e o respetivo polígono.

Resolução com SPSS

Para construir os gráficos deste exercício, é necessário determinar primeiro o ponto médio de cada classe. Os pontos médios são os seguintes:

$$x'_1 = \frac{0+10}{2} = 5 \qquad x'_5 = \frac{40+50}{2} = 45$$
$$x'_2 = \frac{10+20}{2} = 15 \qquad x'_6 = \frac{50+60}{2} = 55$$
$$x'_3 = \frac{20+30}{2} = 25 \qquad x'_7 = \frac{60+70}{2} = 65$$
$$x'_4 = \frac{30+40}{2} = 35 \qquad x'_8 = \frac{70+80}{2} = 75$$

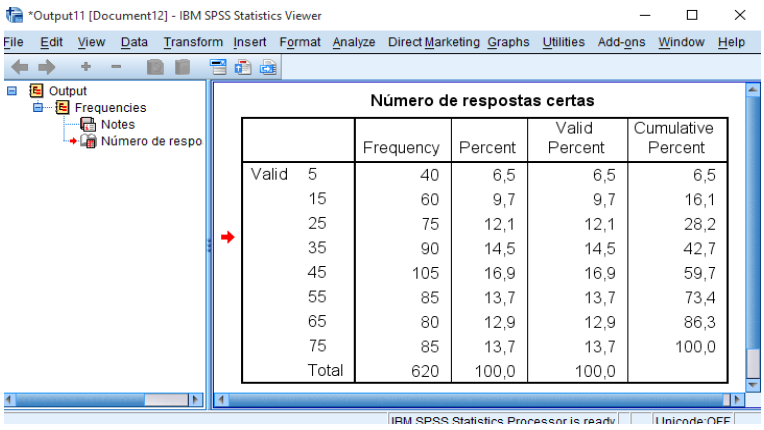
Depois de calcular manualmente os pontos médios, abrir uma nova janela de SPSS para introduzir os pontos médios e as frequências, como mostra a janela seguinte:



	R.certas	freq	var	var	var	var
1	5	40				
2	15	60				
3	25	75				
4	35	90				
5	45	105				
6	55	85				
7	65	80				
8	75	85				

Histograma e um polígono de frequências absolutas

Depois de inserir os pontos médios e as frequências, deve fazer-se o processo de **Weight Cases** que pode ser feito do mesmo modo como se fez no exercício 10. A janela seguinte mostra o processo de ponderação.

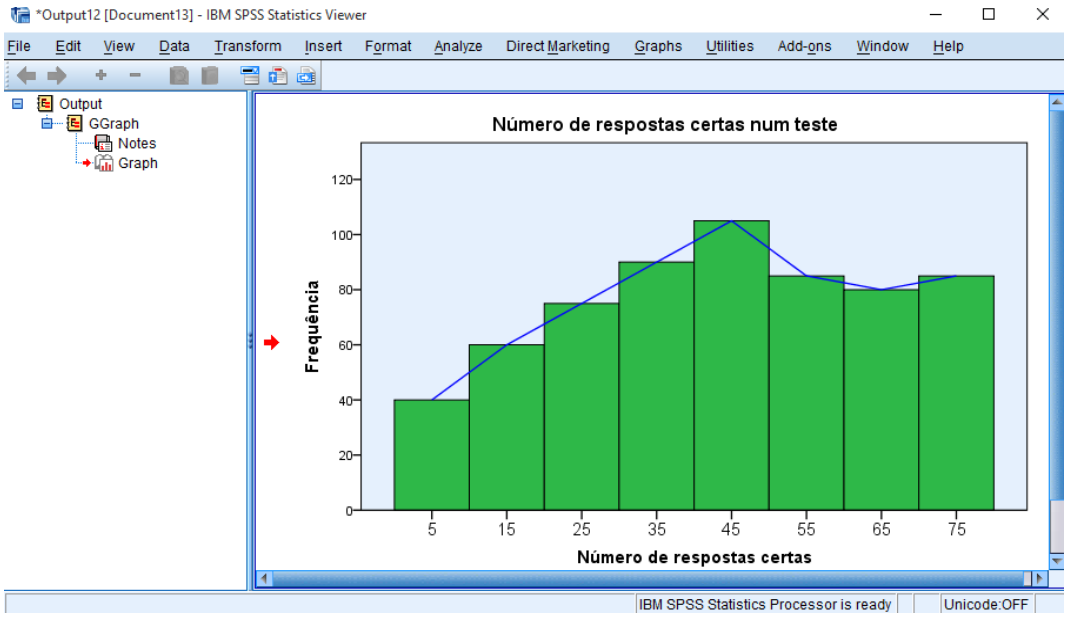


		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	5	40	6,5	6,5	6,5
	15	60	9,7	9,7	16,1
	25	75	12,1	12,1	28,2
	35	90	14,5	14,5	42,7
	45	105	16,9	16,9	59,7
	55	85	13,7	13,7	73,4
	65	80	12,9	12,9	86,3
	75	85	13,7	13,7	100,0
	Total	620	100,0	100,0	

Para construir o histograma e polígono de frequência absoluta basta executar os seguintes comandos:

- Graphs/Chart Builder ...
- Choose from: seleccionar **Histogram**. Em seguida arrastar o gráfico **Simple Histogram** e colocá-lo na parte superior **Chart preview uses example data**
- Arrastar a Frequência e colocá-la no eixo X e o Número de respostas certas no eixo Y.
- OK

Depois de clicar em **OK**, vai aparecer uma janela que apresenta uma figura simples de histograma. Para adicionar ou alterar o gráfico através de **Chart Editor**, basta fazer o duplo clique na área do gráfico. O resultado da construção é o seguinte:

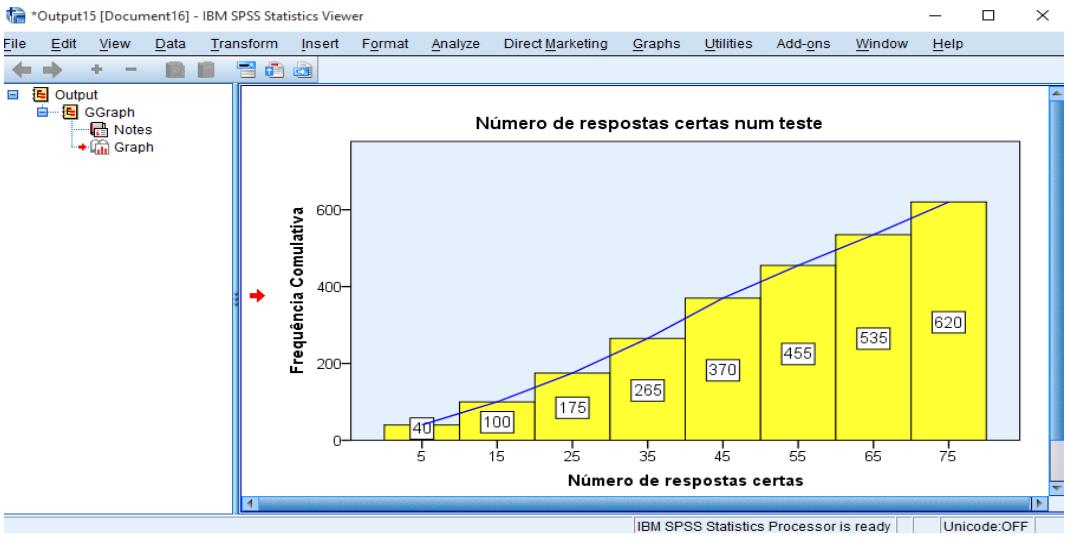


Histograma de frequências relativas acumuladas e o respetivo polígono

Para construir o histograma de frequências relativas acumuladas e o respectivo polígono basta executar os seguintes comandos:

- Graphs/Chart Builder ...
- Choose From: seleccionar **Histogram**. Em seguida arrastar o gráfico **Simple Histogram** e colocá-lo na parte superior **Chart preview uses example data**
- Arrastar a Frequência e colocá-la no eixo Y e o Número de respostas certas no eixo Y. Na janela **Element Properties** executar:
- Statistic: Cumulative Sum
- Bar Style: Bar
- Apply
- OK

Depois de clicar em **OK**, vai aparecer uma janela que apresenta uma figura simples de histograma. Para adicionar ou alterar o gráfico através de **Chart Editor**, basta fazer o duplo clique na área do gráfico. A figura seguinte é Histograma de frequências relativas acumuladas e o respetivo polígono



Resolução com R

Histograma e um polígono de frequências absolutas

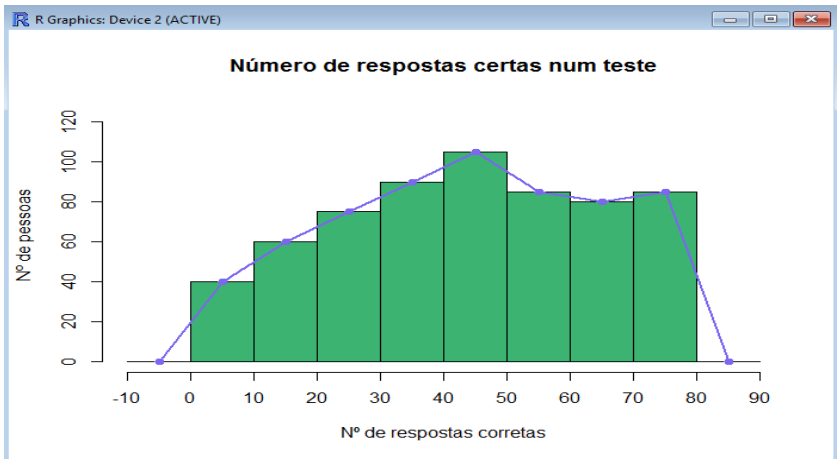
> tb.correta= make.fdt(f = c(0, 40, 60, 75, 90, 105, 85, 80, 85, 0), start = -10, end = 90)

> tb.correta

Class limits	f	rf	rf(%)	cf	cf(%)
$[-10, 0)$	0	0.00	0.00	0	0.00
$[0, 10)$	40	0.06	6.45	40	6.45
$[10, 20)$	60	0.10	9.68	100	16.13
$[20, 30)$	75	0.12	12.10	175	28.23
$[30, 40)$	90	0.15	14.52	265	42.74
$[40, 50)$	105	0.17	16.94	370	59.68
$[50, 60)$	85	0.14	13.71	455	73.39
$[60, 70)$	80	0.13	12.90	535	86.29
$[70, 80)$	85	0.14	13.71	620	100.00
$[80, 90)$	0	0.00	0.00	620	100.00

> plot(tb.correta, main = "Número de respostas certas num teste", ylab = "Nº de pessoas", xlab = "Nº de respostas corretas", col = 'mediumseagreen')

> lines(-5+10 * (0:9), tb.correta $table$f, type = "o", col = 'mediumslateblue', lwd = 2, pch = 19)



Histograma de frequências relativas acumuladas e o respetivo polígono

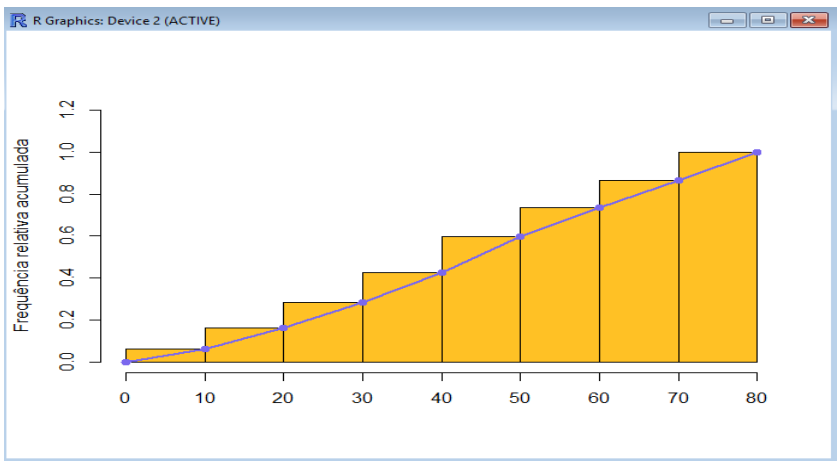
> tb.acumulada= make.fdt(f = c(40, 60, 75, 90, 105, 85, 80, 85), start = 0, end = 80)

> tb.acumulada

Class limits	f	rf	rf(%)	cf	cf(%)
[0, 10)	40	0.06	6.45	40	6.45
[10, 20)	60	0.10	9.68	100	16.13
[20, 30)	75	0.12	12.10	175	28.23
[30, 40)	90	0.15	14.52	265	42.74
[40, 50)	105	0.17	16.94	370	59.68
[50, 60)	85	0.14	13.71	455	73.39
[60, 70)	80	0.13	12.90	535	86.29
[70, 80)	85	0.14	13.71	620	100.00

> plot(tb.acumulada, type = 'cdh', col = 'goldenrod1', ylab = "Frequência relativa acumulada", xlab =)

> lines(10 ∗ (0:8), c(0,tb.acumulada $table [,6]/100),type="o", col = 'mediumslateblue', lwd = 2, pch = 19)



3.7 Diagrama de caixa e bigodes

O diagrama de caixa e bigodes é um diagrama representado em forma de caixa rectangular onde, em cada lado, tanto direito como esquerdo, existe um intervalo ou bigode. O diagrama possui valor mínimo (extremo inferior), primeiro quartil, segundo quartil (mediana), terceiro quartil, valor máximo (extremo superior).

3.7.1 Construir o gráfico de extremos e quartis para n par

Exercício 9: O número de mensagens SMS recebidas em 18 dias consecutivos foram as indicadas a seguir:

9	10	13	14	15	16
19	19	20	21	25	25
32	32	34	36	37	58

Constrói um diagrama de extremos e quartis ou caixa de bigodes[1](ME-TL. 2014, p.142. Tarefa 65)

Para construir o diagrama de extremos e quartis com o programa SPSS e linguagem **R** é necessário calcular, em primeiro lugar, os valores máximo e mínimo, mediana, quartis e atípicos dos dados. Os dados do exercício 8.1 estão por ordem crescente. O número de observações é $n = 18$ (par). O valor mais pequeno é 9 e o maior é 58. A seguir, apresenta-se o processo para calcular os quartis:

$$\begin{aligned} 1^{o} \text{ quartil: } k &= \frac{n+2}{4} = \frac{18+2}{4} = 5 \\ 2^{o} \text{ quartil: } k &= \frac{2n+2}{4} = \frac{n+1}{2} = \frac{18+1}{2} = 9.5 \\ 3^{o} \text{ quartil: } k &= \frac{3n+2}{4} = \frac{3 \cdot 18+2}{4} = 14 \end{aligned}$$

O primeiro quartil está na quinta posição, equivale a 15, a mediana dos dados é $\frac{20+21}{2} = 20.5$ e o último quartil, ou seja, terceiro quartil, situa-se na décima quarta posição e é igual a 32. Para calcular a *Amplitude interquartil (AIQ)* destes dados basta determinar: $AIQ = Q_3 - Q_1 = 32 - 15 = 17$.

[1]Pergunta interpretada pelo pesquisador

Portanto:

$$\begin{aligned} 1.5 \cdot AIQ &= 1.5 \cdot 17 = 25.5 \\ 3 \cdot AIQ &= 3 \cdot 17 = 51 \\ Q_1 &= 15 \\ Q_3 &= 32 \end{aligned}$$

O passo seguinte é identificar o valor Atípico ou Outliers.

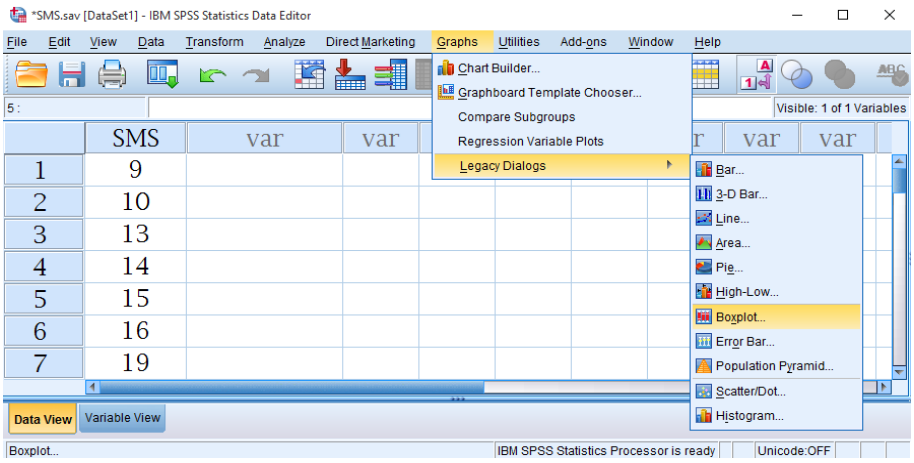
$$\begin{aligned} AI &= Q_1 – 1.5 \cdot AIQ \\ AI &= 15 - 25.5 = -10.5 \\ AS &= Q_3 + 1.5 \cdot AIQ \\ AS &= 32 + 25.5 = 57.5 \end{aligned}$$

Portanto, qualquer valor menor que -10.5 ou maior que 57.5 vai ser considerado como valor atípico. Logo, o único valor atípico é 58.

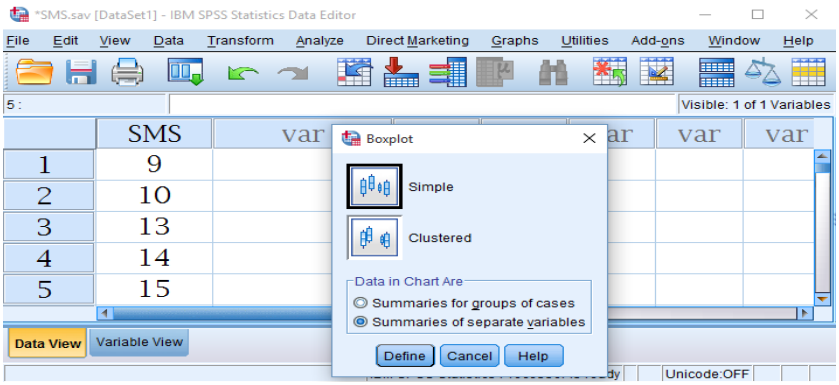
Resolução com SPSS

Para construir os quartis dos dados do exercício 8, pode executar os seguintes comandos:

- Graphs/Legacy Dialog/Boxplot...

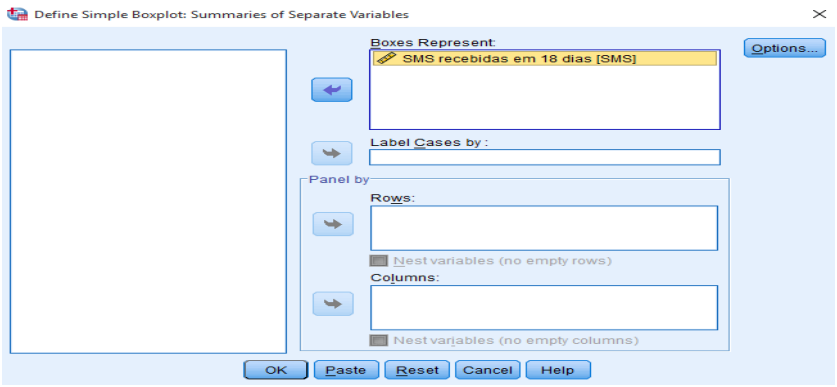


- Simple
- Summaries of seperate variables
- Define

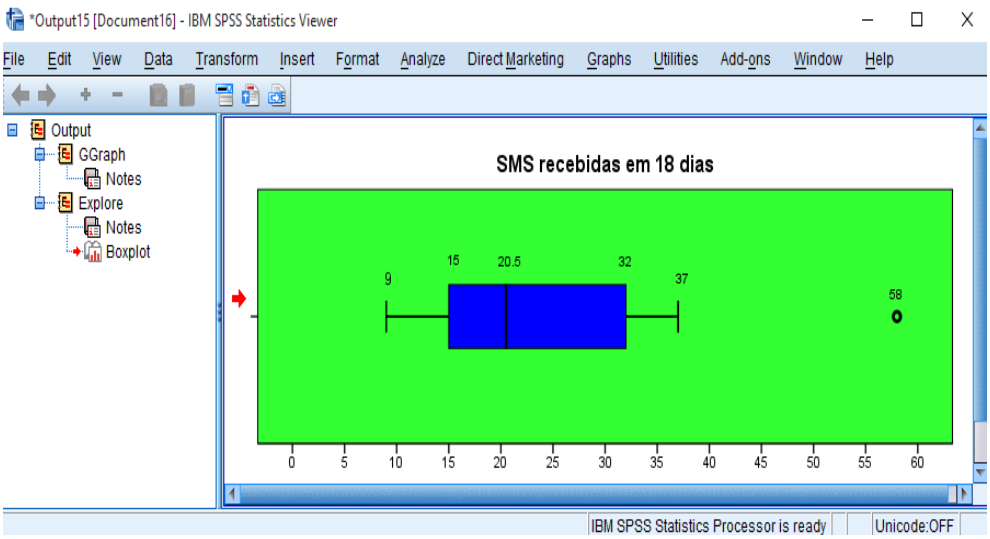


Depois de executar no **Define** vai aparecer logo uma janela de **Define Simple Boxplot** e nela pode ser executado o que se segue:

- Boxes Represent: SMS recebidas em 18 dias [SMS]
- OK



Depois de clicar em **OK**, vai aparecer uma janela que apresenta uma figura simples de Boxplot. Para adicionar ou alterar o gráfico através de **Chart Editor**, basta fazer o duplo clique no gráfico. O resultado da construção é o seguinte:

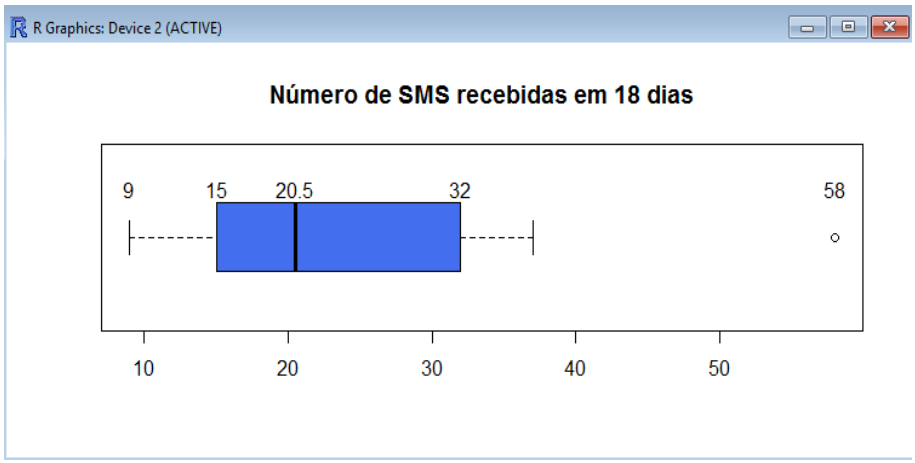


Resolução com R

O diagrama de extremos e quartis, para n par, pode ser obtido através da execução da seguinte sequência de comandos:

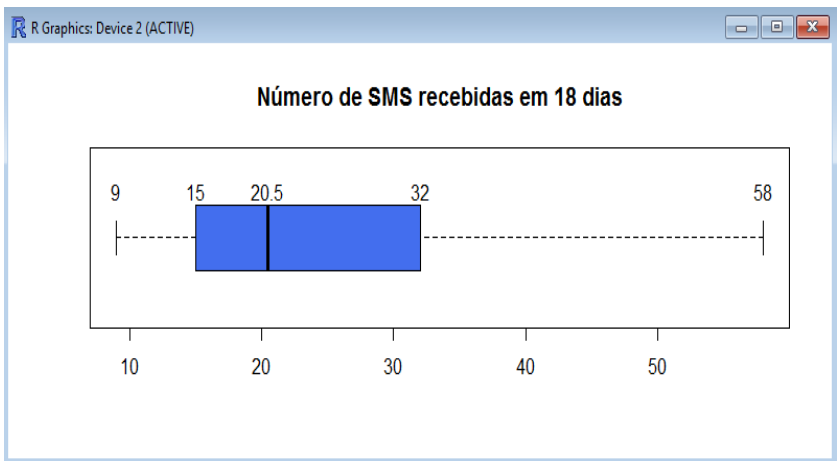
```
> SMS=c(9, 10, 13, 14, 15, 16, 19, 19, 20, 21, 25, 25, 32, 32, 34, 36, 37, 58)
> SMS
 [1] 9 10 13 14 15 16 19 19 20 21 25 25 32 32 34 36 37 58
> boxplot(SMS, col = "royalblue2", main = "Número de SMS recebidas em 18 dias",
  horizontal = TRUE)
> text(x = fivenum(SMS), labels = fivenum(SMS), y = 1.28)
```

O gráfico de extremos e quartis é o seguinte:



Para excluir o valor atípico, basta acrescentar *range = 0* na função de *boxplot()* :

```
> boxplot(SMS, range=0, col = "royalblue2", main = "Número de SMS recebidas em
  18 dias", horizontal = TRUE)
> text(x = fivenum(SMS), labels = fivenum(SMS), y = 1.28)
```



No SPSS é impossível excluir os valores atípicos, enquanto que no **R** é possível excluí-los. Quer nos programas de estatística quer no manual do aluno, a mediana é definida da mesma forma e é um elemento que separa a metade dos dados de baixo e de cima. Os dois outros quartis (Q_1 e Q_2) são obtidos como mediana das metades de baixo e de cima, não incluindo o Q_2. Dos gráficos apresentados pelo programa SPSS e **R**, conclui-se que o valor mínimo dos dados é 9 e o máximo é 58. O primeiro quartil é 15, o segundo quartil ou mediana corresponde a 20.5 e, por último, o terceiro quartil é 32; a maior concentração está entre 15 e 32.

3.7.2 Construir o gráfico de extremos e quartis para n ímpar

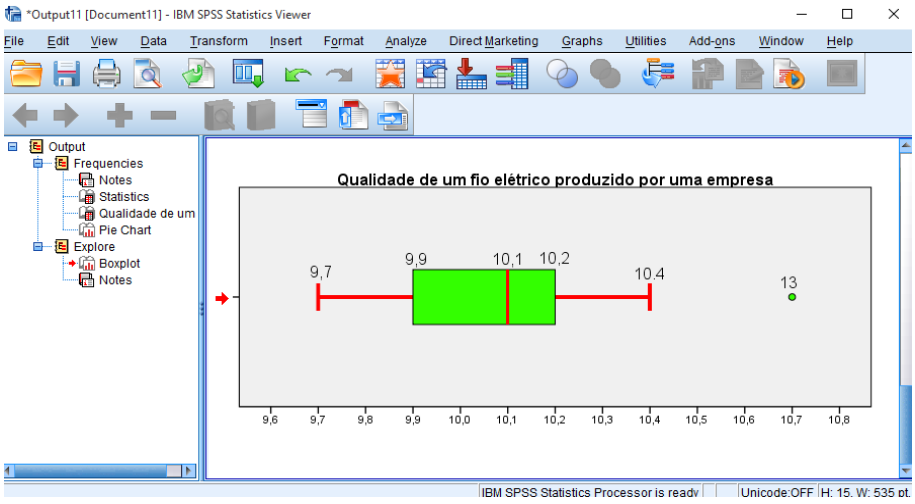
Exemplo 2 Num controlo de qualidade a um fio elétrico produzido por uma empresa, realizou-se a medição da longitude do fio e registaram-se os resultados:

10.4	10.3	9.8	10.2	10	10.2	10.7
10.1	9.8	9.9	10	10.2	9.7	

Desenhe o diagrama de extremos e quartis[2] (ME-TL. 2014, p.145. Tarefa 68)

Resolução com SPSS

A construção do diagrama de extremos e quartis, para n ímpar, no SPSS poder-se-ia tentar realizar da mesma forma que na resolução¸ do *Exercício 9*. No entanto existe um problema adicional que é a forma diferente de obtenção dos quartis relativamente ao métodos seguido no manual. O SPSS segue o método inclusivo, incluindo a mediana Q_2 nas metades de cima e de baixo, ao passo que o manual usa o método exclusivo, que exclui a mediana de ambas as metades. Na janela seguinte é apresentado o resultado da construção.



[2]Pergunta interpretada pelo pesquisador

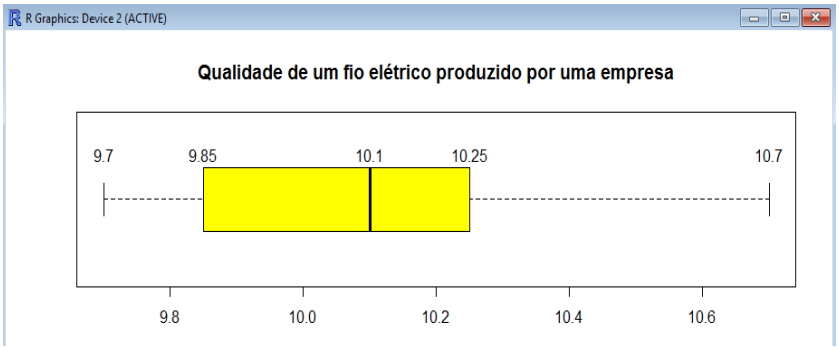
Nota: Note se que um Boxplot não é um diagrama de extremos quartis, pois, apresenta em separados as observações atípicas. É o caso da observação 10.7 no exemplo anterior.

Resolução com R

O diagrama de extremos e quartis tal como é feito no manual, para n ímpar, não pode ser obtido a partir da instrução *boxplot()*, pois os quartis são calculados usando o método inclusivo. No entanto pode ser obtido através da execução da seguinte sequência de comandos:

```
> qualidade=c(10.4,10.3,9.8, 10.2, 10, 10.2, 10.7, 10.1, 9.8, 9.9, 10, 10.2, 9.7)

> qualidade
 [1] 10.4 10.3 9.8 10.2 10.0 10.2 10.7 10.1 9.8 9.9 10.0 10.2 9.7

> qboxplot(qualidade, type = 6, range = 0, col = "yellow", horizontal = T, main
="Qualidade de um fio elétrico produzido por uma empresa")

> text(x = quantile(qualidade,type=6), labels = quantile(qualidade, type=6), y =
1.28)
```

Nota: Para construir o diagrama de extremos e quartis utilizado o comando *qboxplot()* é necessário instalar primeiro o pacote **KMmisc** disponível para o **R**.



Nota: O diagrama de extremos e quartis, produzido pelo **R**, não contém o valor atípico porque a função *qboxplot()* também possui a opção *range = 0*.

Calcular manualmente, as cinco medidas e os valores atípicos

Depois de construir o diagrama de extremos e quartis com o programa estatístico e **R**, o passo seguinte é fazer um cálculo manual das cinco medidas (máximo, mínimo, mediana, primeiro quartil e terceiro quartil) de n ímpar e valores atípicos. Finalmente, comparar a resolução feita no livro *Matemática 12o Ano de Escolaridade: Manual do Aluno* com a de SPSS e **R**.

Serão utilizados dois métodos de calcular quartis: método inclusivo e método exclusivo. Segundo Fernandes e Pinto (2015, p.36) o método é inclusivo, "quando o conjunto de dados tem um número ímpar de elementos e o elemento correspondente a Q_2 é incluído em ambas as metades do conjunto de dados para cálculo dos Q_1 e Q_3"; O método é exclusivo, "quando o conjunto de dados tem um número ímpar de elementos e o elemento correspondente ao Q_2 não é incluído em nenhuma das metades do conjunto de dados para cálculo dos Q_1 e Q_3".

1. Método inclusivo

O método inclusivo é utilizado pelo SPSS para construir o Boxplot. No **R** é também usado pela instrução boxplot().

$$\begin{aligned} 1^{\text{o}}\ \text{Quartil}(Q_1): \quad k &= \frac{n+3}{4} = \frac{13+3}{4} = 4 \\ 2^{\text{o}}\ \text{Quartil}(Q_2): \quad k &= \frac{n+1}{2} = \frac{13+1}{2} = 7 \\ 3^{\text{o}}\ \text{Quartil}(Q_3): \quad k &= \frac{3n+1}{4} = \frac{3\cdot 13+1}{4} = 10 \end{aligned}$$

$$\begin{array}{cccccccccccccc} 9.7 & 9.8 & 9.8 & \underset{\substack{\uparrow \\ \mathbf{Q_1}}}{9.9} & 10 & 10 & \underset{\substack{\uparrow \\ \mathbf{Q_2}}}{10.1} & 10.2 & 10.2 & \underset{\substack{\uparrow \\ \mathbf{Q_3}}}{10.2} & 10.3 & 10.4 & 10.7 \end{array}$$

Logo: $Q_2 = 10.1$, $Q_1 = 9.9$ e $Q_3 = 10.2$

2. Método exclusivo

Segue-se o processo para calcular os quartis pelo método exclusivo que é um método utilizado no livro *Matemática 12º Ano de Escolaridade: Manual do Aluno*:

$$\begin{aligned} 1^{\text{o}}\ \text{Quartil}(Q_1): \quad k &= \frac{n+1}{4} = \frac{13+1}{4} = 3.5 \\ 2^{\text{o}}\ \text{Quartil}(Q_2): \quad k &= \frac{n+1}{2} = \frac{13+1}{2} = 7 \\ 3^{\text{o}}\ \text{Quartil}(Q_3): \quad k &= \frac{3n+3}{4} = \frac{3\cdot 13+3}{4} = 10.5 \end{aligned}$$

$$\begin{array}{cccccccccccccc} 9.7 & 9.8 & \underset{\substack{\uparrow \\ \mathbf{Q_1}}}{9.8} & 9.9 & 10 & 10 & \underset{\substack{\uparrow \\ \mathbf{Q_2}}}{10.1} & 10.2 & 10.2 & 10.2 & \underset{\substack{\uparrow \\ \mathbf{Q_3}}}{10.3} & 10.4 & 10.7 \end{array}$$

Portanto: $(Q_2 = 10.1)$, $Q_1 = \dfrac{9.8+9.9}{2} = 9.85$ e $Q_3 = \dfrac{10.2+10.3}{2} = 10.25$

3. Calcular valores atípicos

- Determinar a amplitude interquartil (AIQ):

$$AIQ = Q_3 – Q_1 = 10.25 – 10.1 = 0.15$$

- Determinar os valores $AI = Q_1 – 1.5 \cdot AIQ$ e $AS = Q_3 + 1.5 \cdot AIQ$:

$$\begin{aligned} AI &= Q_1 – 1.5 \cdot AIQ \\ AI &= 9.85 - 1.5 \cdot 0.15 = 8.35 \\ AI &= 9.85 - 0.225 = 9.625 \\ AS &= Q_3 + 1.5 \cdot AIQ \\ AS &= 10.25 + 1.5 \cdot 0.15 = 11.75 \\ AS &= 10.25 + 0.225 = 10.475 \end{aligned}$$

Portanto, qualquer valor menor que 9.625 ou maior que 10.475 vai ser considerado como valor atípico. Logo, o único valor atípico é 10.7.

Conclusão:

Observando os dois gráficos produzidos pelos SPSS e **R** e a resolução manual dos quartis concluímos que, sendo n ímpar, a mediana é igual a $\frac{n+1}{2}$.

O **R** possui a opção *range = 0* para excluir o valor atípico, enquanto que o SPSS não possui comando para tal efeito.

O **R** permite calcular quartis de 10 maneiras diferentes: o método inclusivo usado no *boxplot()* e *fivenum()*, e 9 tipos diferentes de quantis usados no *qboxplot()* e *quantile()* através da opção *type=...* , nenhuma delas coincidindo com o método exclusivo. No entanto, fazendo a separação em número de observações par e ímpar, o método exclusivo coincide com o de *boxplot()*, *qboxplot(...,type=2)* e *qboxplot(...,type=5)* no caso par, e com o de *qboxplot(...,type=6)* no caso ímpar.

O SPSS calcula quantis usando o *type=6* do **R** quando faz descrição estatística dos dados. Mas no boxplot o SPSS utiliza as "dobradiças"de Tukey (método inclusivo) usadas também pelo boxplot do **R**.

No manual, os quartis são obtidos através do método exclusivo e assim só com o **R** se consegue sempre representar o diagrama de extremos e quartis tal como é apresentado aos alunos. O *SPSS* só consegue fazer uma representação coincidente no caso de haver um número de observações par e na ausência de valores atípicos.

3.8 Medidas descritivas

Exercício 11: Num controlo de qualidade a um fio elétrico produzido por uma empresa, realizou-se a medida da longitude do fio e registaram-se os resultados:

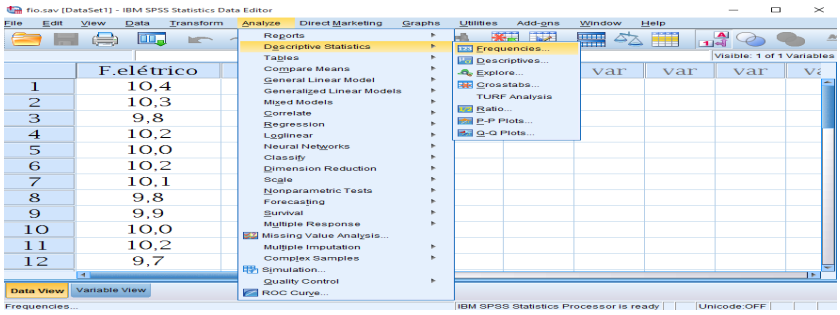
10.4	10.3	9.8	10.2	10	10.2
10.1	9.8	9.9	10	10.2	9.7

1. Determine a média, a mediana e a amplitude das medições.
2. Determina a variância e desvio padrão da distribuição. (ME-TL. 2014, p.145. Tarefa 68).

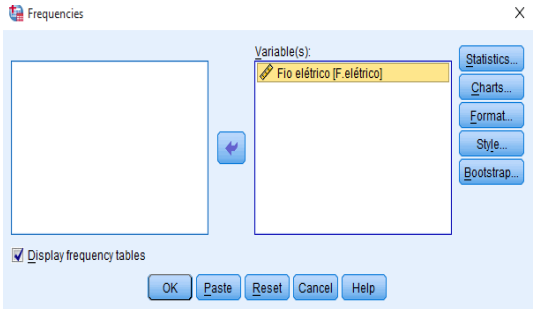
Resolução com SPSS

Depois de introduzir os dados no SPSS, pode executar os seguintes comandos:

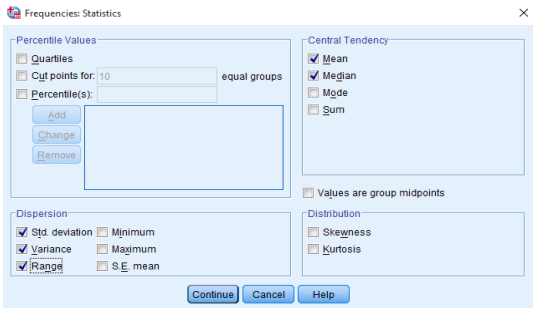
- Analyze/Descritive Statistics/Frequencies...



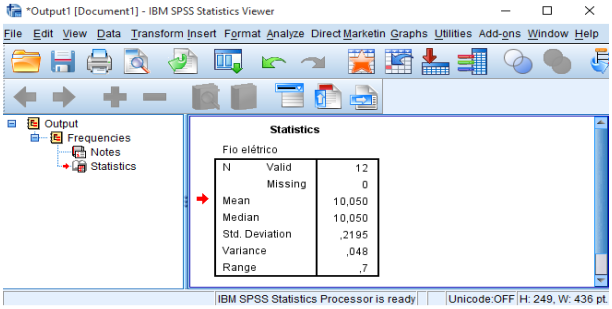
- Variable(s): Fio elétrico[F.elétrico]
- ☑ Display frequency tables



- Statistics: ☑ Mean, ☑ Median, ☑ Std. Deviation, ☑ Variance ☑ Range
- Continue
- OK



O resultado obtido é apresentado na tabela seguinte:



Resolução com R

> fio=c(10.4, 10.3, 9.8, 10.2, 10, 10.2, 10.1, 9.8, 9.9, 10, 10.2, 9.7)

> fio

[1] 10.4 10.3 9.8 10.2 10.0 10.2 10.1 9.8 9.9 10.0 10.2 9.7

> média=mean(fio)

> média

[1] 10.05

> mediana=median(fio)

> mediana

[1] 10.05

> desviopadrão =sd(fio)

> desviopadrão

[1] 0.2195036

> variância= var(fio)

> variância

[1] 0.04818182

> range =max(fio)-min(fio)

> range

[1] 0.7

Da análise dos dados que estão guardados no objecto de *fio*, pode concluir-se que o programa **R** deu a mesma solução que SPSS. Além disso, os valores que aparecem nas janelas de saídas dos programa de estatística são média, mediana, variância, desvio padrão e amplitude dos dados. respetivamente, 10.05, 10.05, 0.04818182, 0.2195036 e 0.7

Resolução do manual

1. Variância populacional de dados univariados

   A formula da variância populacional para dados univariados no livro do aluno(METL. 2014, p.144) é dada por:

$$\sigma^2 = \sum_{i=1}^{n} \frac{f_i(x_i - \bar{x})^2}{n},$$

   onde x_i Esta fórmula só faz sentido no caso de se assumir uma população finita de tamanho n. De um modo mais geral, σ^2 obtém-se a partir de uma variável aleatória X e considerando o valor esperado

$$E((X - \mu)^2),$$

   onde μ é a média de X.

2. Variância amostral para dados univariados

   A fórmula geral da variância amostral para dados univariados é dada por:

$$S^2 = \sum_{i=1}^{n} \frac{f_i(x_i - \bar{x})^2}{n-1}$$

Para calcular a variância e o desvio padrão é necessário determinar, em primeiro lugar, o valor da média dos dados:

$$\begin{aligned} \bar{x} &= \frac{9.7 + 9.8 \cdot 2 + 9.9 + 10 \cdot 2 + 10.1 + 10.2 \cdot 3 + 10.3 + 10.4}{12} \\ &= \frac{120.6}{12} = 10.05 \end{aligned}$$

x_i	f_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$f_i(x_i - \bar{x})^2$
9,7	1	-0.35	0,1225	0,1225
9,8	2	-0.25	0,0625	0,125
9,9	1	-0.15	0,0225	0,0225
10	2	-0,05	0,0025	0,005
10,1	1	0.05	0,0025	0,0025
10,2	3	0,15	0,0225	0,0675
10,3	1	0,25	0,0625	0,0625
10,4	1	0,35	0,1225	0,1225
Total				0,53

A Variância populacional será:

$$\begin{aligned} \sigma^2 &= \sum_{i=1}^{n} \frac{f_i(x_i - \bar{x})^2}{n} \\ \sigma^2 &= \frac{0.53}{12} \\ \sigma^2 &= 0,04 \end{aligned}$$

O Desvio padrão populacional é:

$$\begin{aligned} \sigma^2 &= 0,04 \\ \sigma &= \sqrt{0,04} \\ \sigma &= 0.2 \end{aligned}$$

A Variância amostral será:

$$\begin{aligned} S^2 &= \sum_{i=1}^{n} \frac{f_i(x_i - \bar{x})^2}{n-1} \\ S^2 &= \frac{0.53}{11} \\ S^2 &= 0,05 \end{aligned}$$

O Desvio padrão amostral é:

$$\begin{aligned} S^2 &= 0,05 \\ S &= \sqrt{0,05} \\ S &= 0.2 \end{aligned}$$

Observando os dois resultados apresentados por SPSS e **R**, conclui-se que ambos estão a calcular a variância e o desvio padrão amostral. Ao calcular estas duas medidas, o SPSS

executou segundo os comandos de *var()* e *sd()* do **R**. O resultado é diferente daquele que se encontra no manual do aluno. Para calcular o valor da variância populacional, exactamente igual à resolução do manual, é necessário executar os seguintes comandos do **R**.

> varp=function(fio){sum((fio-mean(fio))^2)/(length(fio))}

> varp(fio)

[1] 0.04416667

> sdp=function(fio)sqrt(sum((fio-mean(fio))^2)/(length(fio)))

> sdp(fio) [1] 0.2101587

Da informação apresentada pelo programa de estatística e **R**, conclui-se que a variância e o desvio padrão amostral são, respectivamente, 0.048 e 0.2195. Por outro lado, os comandos:

varp=function(fio){sum((fio-mean(fio))^2)/(length(fio))} e

sdp=function(fio)sqrt(sum((fio-mean(fio))^2)/(length(fio)))

do **R** calculam a variância e o desvio padrão populacional com valores iguais aos do manual do aluno.

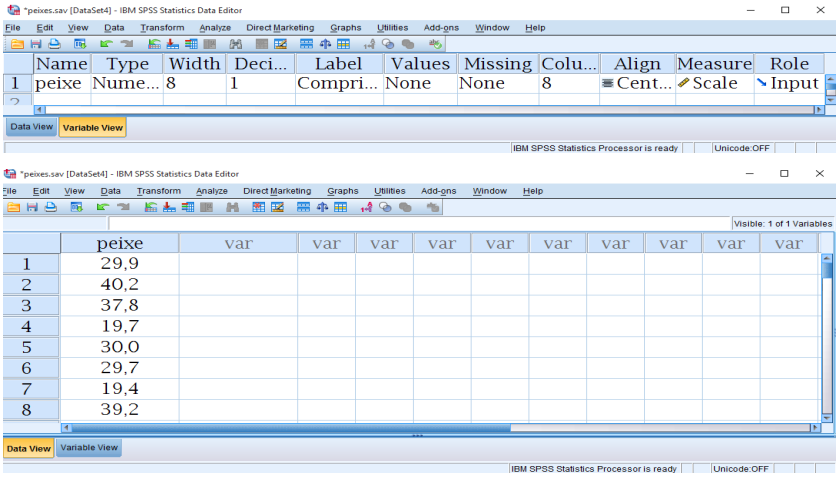
Exercício 12: Os dados que se seguem correspondem ao comprimento, em cm, de uma amostra de peixes:

29,9	40,2	37,8	19,7	30	29,7	19,4	39,2	24,7	20,4
19,1	34,7	33,5	18,3	19,4	27,3	38,2	16,2	36,8	33,1
41,4	13,6	32,2	24,3	19,1	37,4	23,6	33,3	31,6	20,1
17,2	13,3	37,7	12,6	39,6	24,6	18,6	18	33,7	38,2

1. Agrupe os dados em classes e constrói um histograma de frequências absoluta.
2. Determine a variância e desvio padrão dos dados agrupados (ME-TL. 2014, p.146. Tarefa 69).

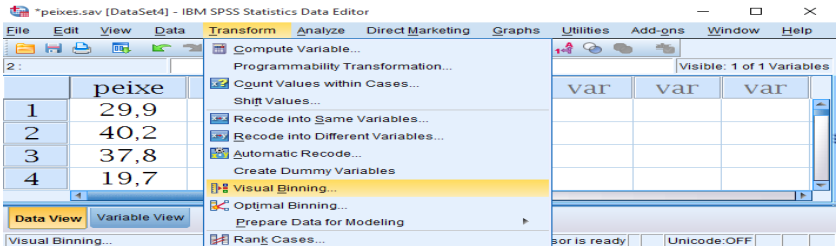
Resolução com o SPSS

Para iniciar, deve abrir-se a janela **Variable View**, Em seguida, cria-se a variável com o nome adequado, por exemplo, *Peixe*. Entretanto, guarda-se o ficheiro com o nome escolhido (*peixes.sav)*. Depois de definir a variável, abre-se a janela **Data View**, para introduzir os dados como se mostra nas duas janelas seguintes:

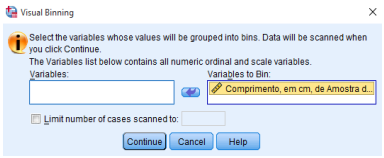


Para agrupar os dados em classes, basta executar os seguintes comandos:

- Transform /Visual Binnig ...



- Variables to Bin: Comprimento, em cm, de Amostra...

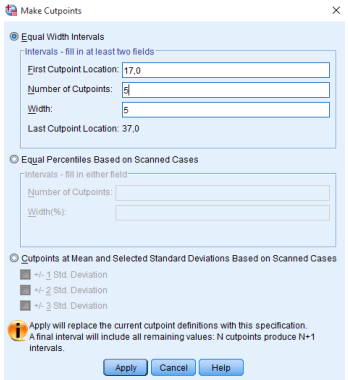


- Continue

  Depois de executar **Continue** aparece a janela **Visual Binnig**. Na mesma janela executar:

- Make Cutpoints...

- Apply



Novamente, na janela de **Visual Binning**. pode-se executar as seguintes fases:

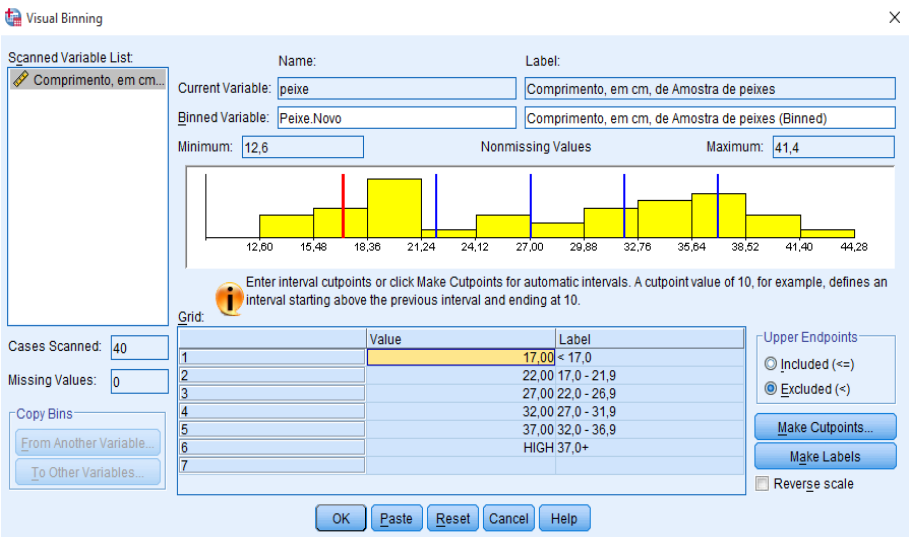
- Binned Variable: Peixe.Novo

  Deve-se escrever o nome da nova variável que é diferente do nome da primeira *Peixe*

- ⊙ Excluded (<)

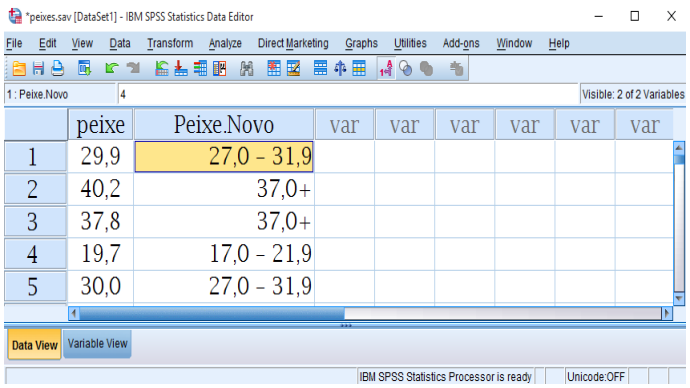
- Make label (para criar novo label automaticamente)

- OK

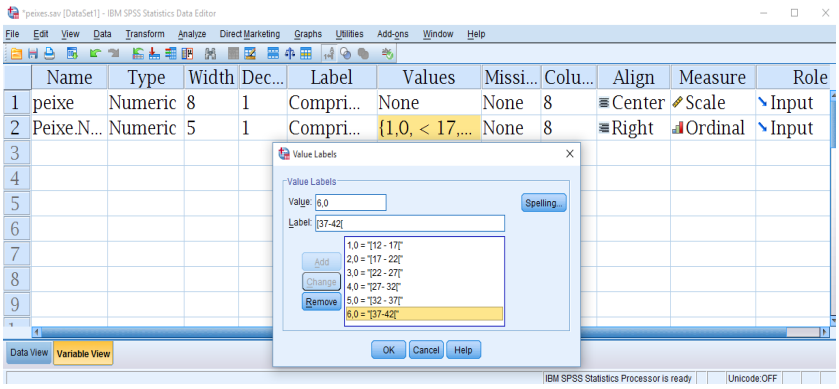


Depois de clicar **OK**, aparece a seguinte janela **Binning specifications will create 1 variables**. Nela, clicar **OK**.

Depois de clicar **OK** vai aparecer o novo nome da variável **Peixe.novo** (ver a janela a seguir).

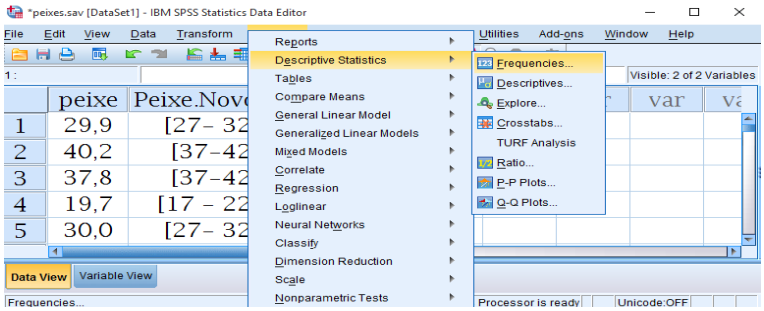


As etiquetas aos valores ou limites das classes da variável *Peixe.Novo* é atribuída na coluna **Values**, para terminar basta seleccionar **OK**. Como mostra a seguinte janela:



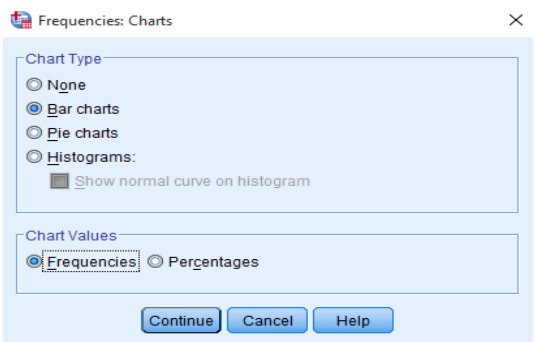
Para ver o resultado da distribuição de frequências, pode ser feito o seguinte procedimento:

- Analize/Descriptive Statistics/Frequencies...



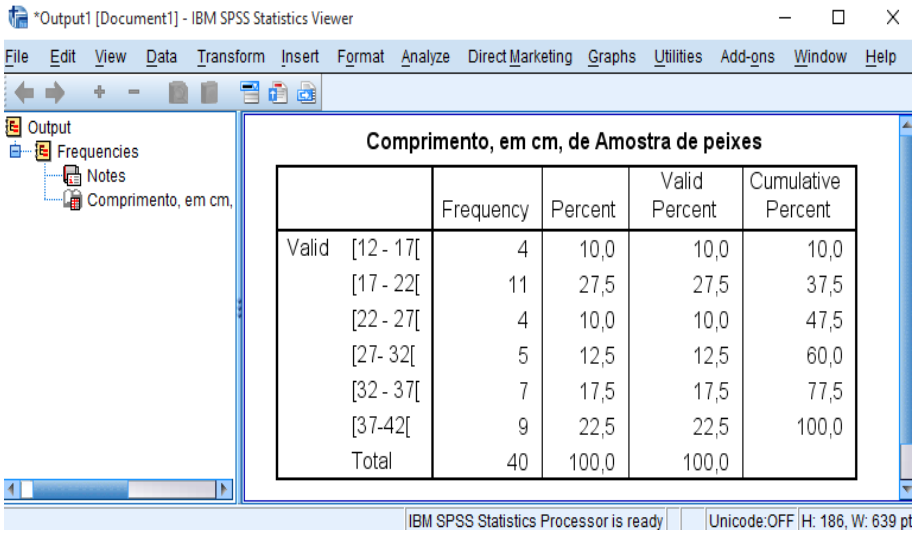
- Variable(s): Comprimento, em...
- ☑ Display frequency tables

- Chart
- Chart Type: ⊙ Bar charts
- Chart Values: ⊙ Frequencies
- Continue



► OK.

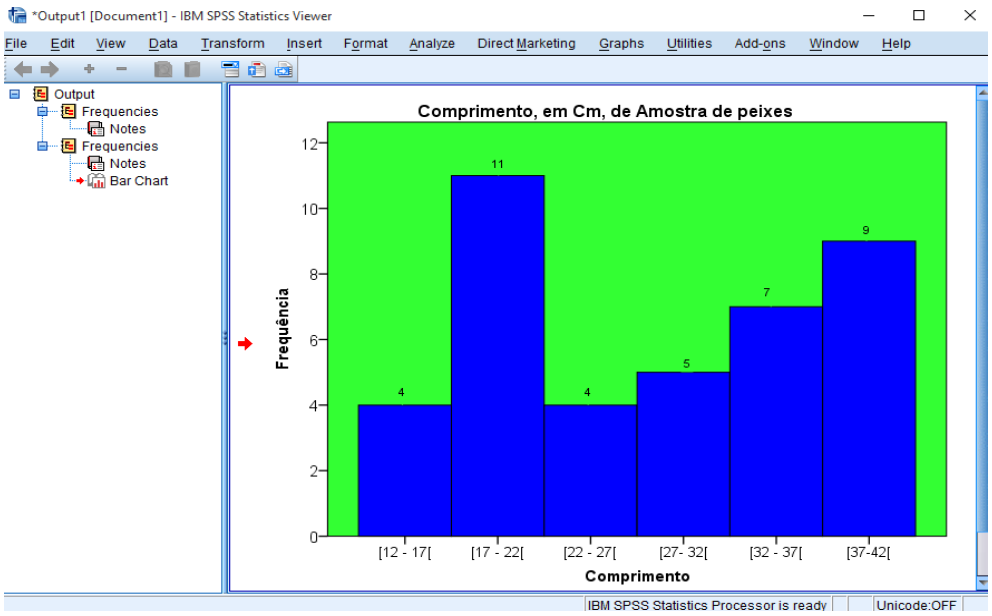
A tabela seguinte é o resultado da distribuição de frequências:



Comprimento, em cm, de Amostra de peixes

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	[12 - 17[	4	10,0	10,0	10,0
	[17 - 22[	11	27,5	27,5	37,5
	[22 - 27[	4	10,0	10,0	47,5
	[27- 32[	5	12,5	12,5	60,0
	[32 - 37[	7	17,5	17,5	77,5
	[37-42[	9	22,5	22,5	100,0
	Total	40	100,0	100,0	

Para adicionar ou alterar o gráfico através de *Chart Editor*, basta fazer o duplo clique no gráfico. O resultado da construção é o seguinte:



Resolução com R

Depois de introduzir os dados na janela de **R**, pode executar os seguintes comandos:

```
> peixe=c(29.9, 40.2, 37.8, 19.7, 30, 29.7, 19.4, 39.2, 24.7, 20.4, 19.1, 34.7, 33.5, 18.3,
+ 19.4, 27.3, 38.2, 16.2, 36.8, 33.1, 41.4, 13.6, 32.2, 24.3, 19.1, 37.4, 23.6, 33.3, 31.6,
+ 20.1, 17.2, 13.3, 37.7, 12.6, 39.6, 24.6, 18.6, 18, 33.7, 38.2)
> [1] 29.9  40.2  37.8  19.7  30.0  29.7  19.4  39.2  24.7  20.4  19.1  34.7 33.5  18.3  19.4
```

[16] 27.3 38.2 16.2 36.8 33.1 41.4 13.6 32.2 24.3 19.1 37.4 23.6 33.3 31.6 20.1

[31] 17.2 13.3 37.7 12.6 39.6 24.6 18.6 18.0 33.7 38.2

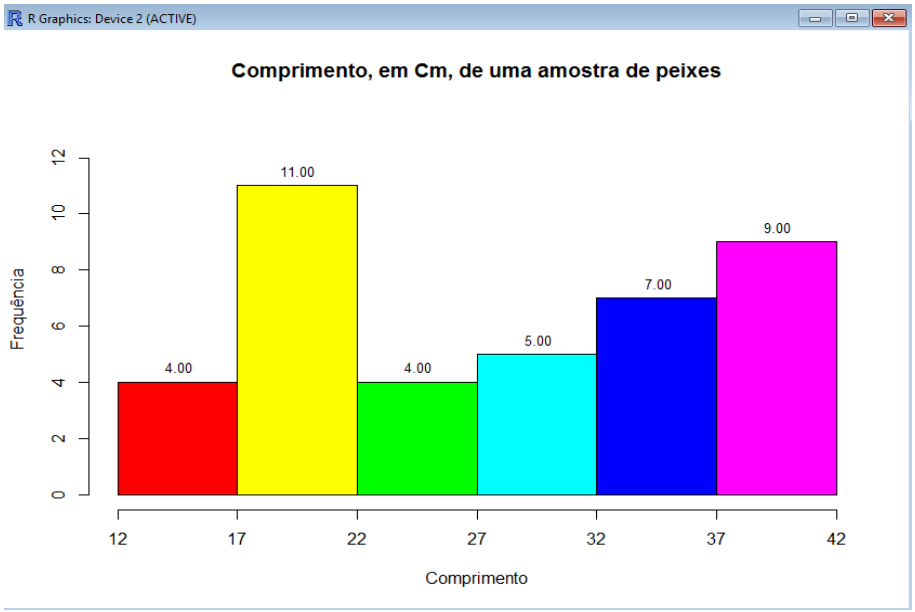
Para construir a tabela de distribuição de frequência é necessário instalar primeiro o pacote **fdth** (Frequency Distribution Tables, Histograms and Poligons) disponível para o **R**.

> tabela= fdt(peixe, start=12, end=42, h=5)

> tabela

Class limits	f	rf	rf(%)	cf	cf(%)
[12, 17)	4	0.10	10.0	4	10.0
[17, 22)	11	0.28	27.5	15	37.5
[22, 27)	4	0.10	10.0	19	47.5
[27, 32)	5	0.12	12.5	24	60.0
[32, 37)	7	0.18	17.5	31	77.5
[37, 42)	9	0.22	22.5	40	100.0

> plot(tabela, main = "Comprimento, em Cm, de uma amostra de peixes", xlab = "Comprimento", ylab = "Frequência", col = rainbow(6), v = TRUE, cex = .8) O resultado da construção é o seguinte:

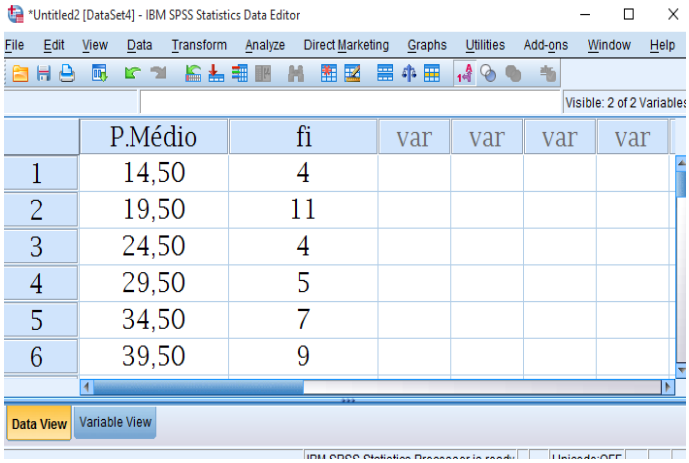


Calcular a variância e desvio padrão dos dados que estão distribuídos numa tabela de frequência **comprimento em cm [Binned]**, só pode ser possível quando utilizar os

pontos médios (calculados manualmente) e as frequências.

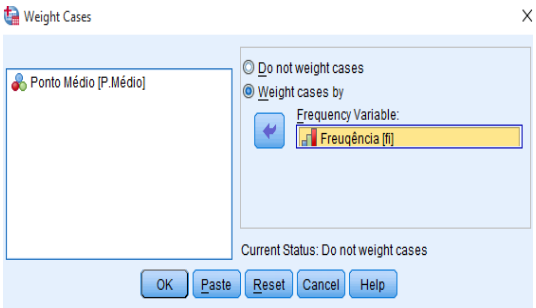
$$x_1' = \frac{12+17}{2} = 14,5 \qquad x_4' = \frac{27+32}{2} = 29,5$$
$$x_2' = \frac{17+22}{2} = 19,5 \qquad x_5' = \frac{32+37}{2} = 34,5$$
$$x_3' = \frac{22+27}{2} = 24.5 \qquad x_6' = \frac{37+42}{2} = 39,5$$

Depois de calcular manualmente os pontos médios, abrir uma nova janela de SPSS para introduzir os os pontos médios e e as frequências. Como mostra a janela seguinte:



Depois de inserir os pontos médios e as frequências, deve fazer-se o processo de **Weight Cases** para estabelecer a correspondência entre cada classe de intervalo e a sua frequência, através da execução dos seguintes comandos:

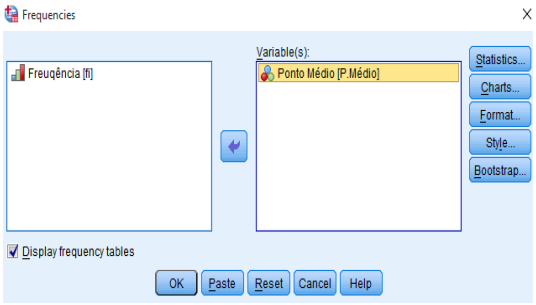
- Data/Weight Cases
- ⊙ Weight Cases By: Frequências



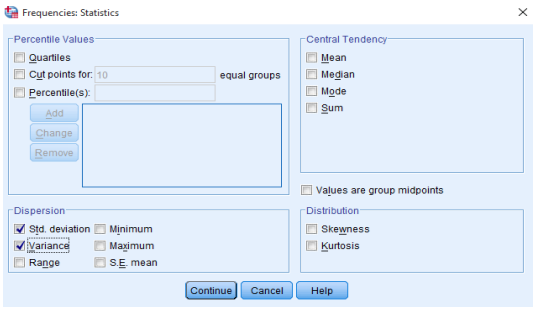
- OK

Para calcular a variância e desvio padrão pode executar os seguintes comandos:

- Analyze/Descriptive Statistics/Frequencies...
- Variable(s): Comprimento de peixe em cm
- ☑ Display frequency tables

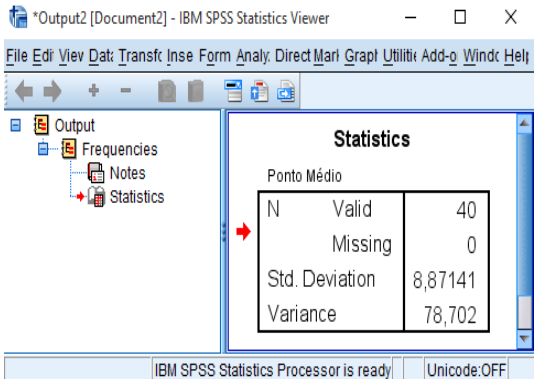


- Statistics: ☑ Std. Deviation, ☑ Variance



- Continue
- OK

O resultado obtido é apresentado na seguinte saída:



Resolução com R

Os dados de *peixe* já estão distribuídos em classes. Essa distribuição é gravada na função de **R** com o nome *tabela*. Em seguida, pode efectuar os seguintes comandos para calcular a variância e o desvio padrão:

```
> var(tabela)
  [1] 78.70192
> sd(tabela)
  [1] 8.87141
```

Dos resultados das análises apresentados pelo programa de estatística e **R** (resolução do exercício 12), conclui-se que existem diferenças principalmente na apresentação da variância e do desvio padrão dos dados. SPSS e **R** mostraram que a variância corresponde ao valor **78.7** e o desvio padrão é igual a **8.9**, enquanto que utilizar a *Variância Populacional* no manual do aluno (p. 144) apresentou o resultado da variância igual a **76.7** e o do desvio padrão igual a **8.8**. Estas diferenças podem ser comprovadas através de uma resolução manual, utilizando a *fórmula 4.2* e a *formula da variância* utilizada no livro do aluno (p.144).

Calcular manualmente, a variância e desvio padrão

Construí-se uma tabela com os valores envolvidos no calculo da variância e desvio padrão

Classe	x_i	f_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$f_i (x_i - \bar{x})^2$	$f_i x_i$
$[12 - 17[$	14,5	4	-13,4	178,6	715,6	58
$[17 - 22[$	19,5	11	-8,38	70,14	771,5	214,5
$[22 - 27[$	24,5	4	-3,38	11,39	45,56	98
$[27 - 32[$	29,5	5	1,641	2,641	13,2	147,5
$[32 - 37[$	34,5	7	6,625	43,89	307,2	241,5
$[37 - 42[$	39,5	9	11,63	135,1	1216	355,5
Total		40			3069	1115

Média

$$\bar{x} = \sum_{i=1}^{6} \frac{f_i x_i}{n} = \frac{1115}{40} = 27.9$$

O valor da média obtido é $\bar{x} = 27.9$.

1. Variância e desvio padrão populacional

Para calcular manualmente a Variância populacional e desvio padrão populacional pode ser feito o seguinte cálculo:

$$\begin{aligned} \sigma^2 &= \sum_{i=1}^{n} \frac{f_i(x_i - \bar{x})^2}{n} \\ &= \frac{3069}{40} \\ &= 76,725 \\ \sigma &= \sqrt{\sigma^2} = \sqrt{76,75} = 8.76 \end{aligned}$$

2. Variância e desvio padrão amostral

Para calcular manualmente a Variância amostral e desvio padrão amostral pode ser feito o seguinte cálculo:

$$\begin{aligned} S^2 &= \sum_{i=1}^{n} \frac{f_i(x_i - \bar{x})^2}{n} \\ &= \frac{3069}{39} \\ &= 78,6923 \\ S &= \sqrt{S^2} = \sqrt{78,692} = 8.87 \end{aligned}$$

Conclusão: Observando os dois resultados apresentados por SPSS e **R**, conclui-se que ambos estão a calcular a variância e o desvio padrão amostral, os quais são, respectivamente,78,7 e 8.87. Os resultados são diferentes daqueles que se encontram no manual do aluno, pois aí são utilizados a variância populacional e o desvio padrão populacional.

3.9 Distribuições bidimensionais

Uma das técnicas de estatística que é utilizada para analisar a relação causa e efeito entre duas variáveis quantitativas, é conhecida como análise bivariada.
Exemplos:

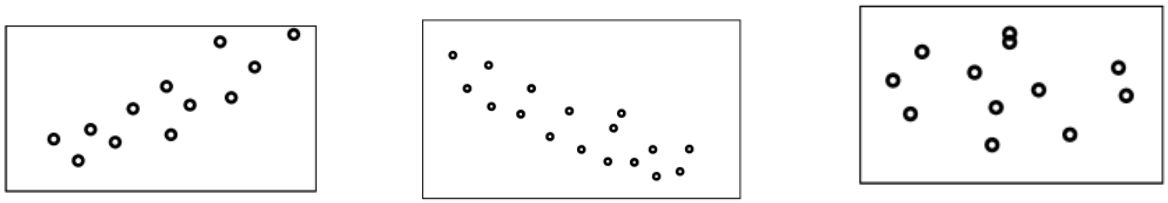
1. A duração do tempo de estudo e o aproveitamento de estudo.

2. A quantidade produzida e a quantidade consumida.

   ”Existe correlação linear positiva entre duas variáveis se a nuvem de pontos se ajustar a uma recta com declive positivo, ou seja, as variáveis evoluem no mesmo sentido. (Se uma cresce a outra também cresce, se uma decresce a outra decresce.) *Figura 1*

Existe correlação linear negativa entre duas variáveis se a nuvem de pontos se ajustar a uma recta com declive negativo, ou seja, as variáveis evoluem em sentido contrário. (Se uma cresce a outra decresce e vice-versa.)*Figura 2*

Existe correlação nula se não há qualquer influência de uma variável na outra. *Figura 3*



(a) Figura 1 (b) Figura 2 (c) Figura 3

Figura 3.2: Diagrama de Disperção

Numa distribuição bidimensional, ao ponto $(\bar{x}, \bar{y})$ chama-se ponto médio da nuvem de pontos ou centro de gravidade da distribuição "(ME-TL,2014,p.148,149)

Recta de regressão

A recta que passa no centro de gravidade da distribuição e melhor se ajusta à nuvem de pontos chama-se recta de regressão.

3.9.1 Coeficiente de correlação

O coeficiente de correlação de pearson, r, mede o grau de associação linear entre x ey (amostra bivariada quantitativa). Este coeficiente assume a Normalidade dos dados e é dado por:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Se $r < 0$

A correlação é negativa, A variação de variáveis é feita em sentidos opostos, isto é uma aumenta quando a outra diminui.

Se $r > 0$

A correlação é positiva, A variação das variáveis é feita no mesmo sentido, isto é, uma aumenta quando a outra também aumenta.

Se $r = 0$

A correlação é nula. (ME-TL. 2014, p.149-150)

Exercício 13: Considere a seguinte distribuição das idades dos elementos dos casais que se encontram numa festa.

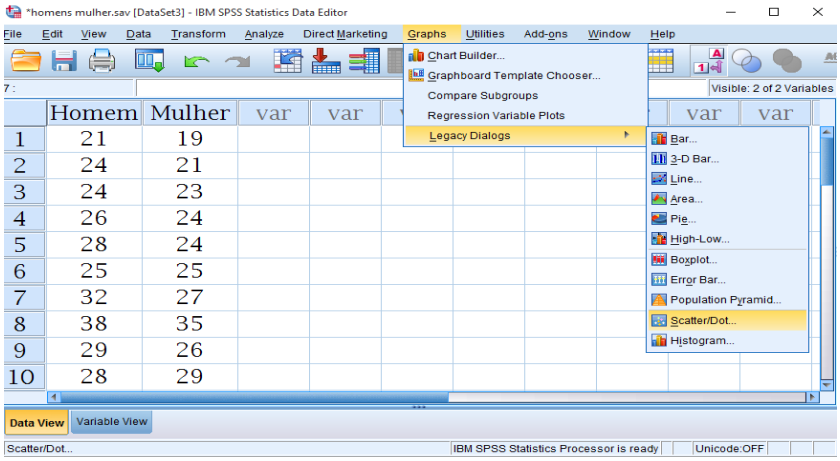
Homem	21	2 4	24	26	28	25	32	38	29	28
Mulher	19	21	23	24	24	25	27	35	26	29

Construa o diagrama de dispersão dos dados fornecidos. (ME-TL. 2014, p.148. Tarefa 71)

Resolução com SPSS

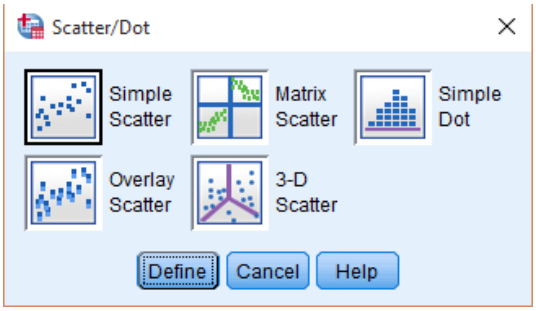
Depois de inserir os dados no SPSS, pode executar o seguinte comando, para construir o diagrama de dispersão:

- Graphs/Legacy Dialog/Scatter/Dot ...



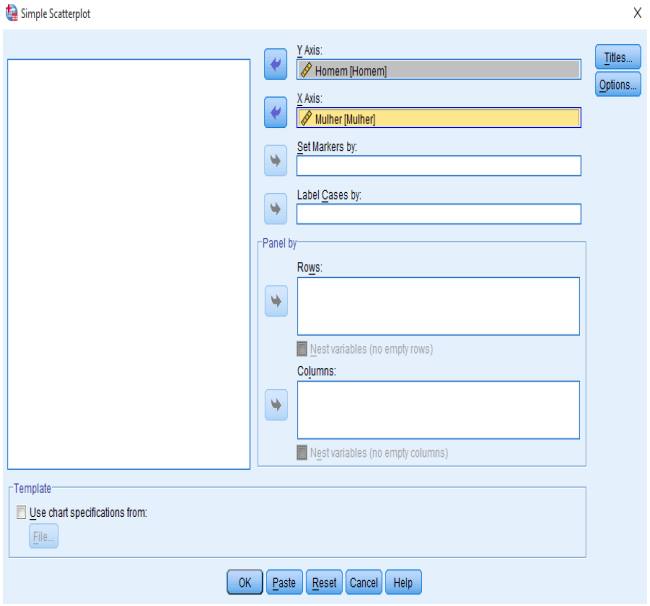
Aparece a janela scatter/Dot e nela pode executar:

- Simple scater
- Define

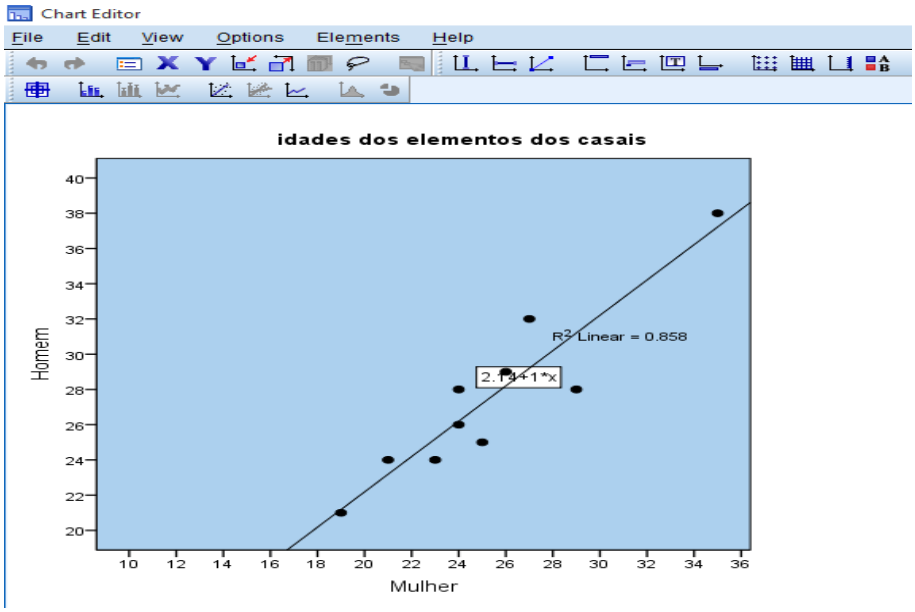


- Y Axis: Homem
- X Axis: Mulher

► OK



Para saber a relação linear e correlação linear das variáveis, clicar duas vezes na área do diagrama. Em seguida, vai aparecer **Chart Editor**. Seleccionar **Add Fit line At Total**. Aparece a janela Propriedade, em **Fit Method** escolher **Linear**, e por último, carregar no **Apply**. O resultado da análise está na seguinte saída:



Resolução com R

Depois de introduzir os dados na janela de **R**, pode executar os seguintes comandos:

> homem = c(21, 24, 24, 26, 28, 25, 32, 38, 29, 28)

> homem

[1] 21 24 24 26 28 25 32 38 29 28

> mulher=c(19, 21, 23, 24, 24, 25, 27, 35, 26, 29)

> mulher

[1] 19 21 23 24 24 25 27 35 26 29

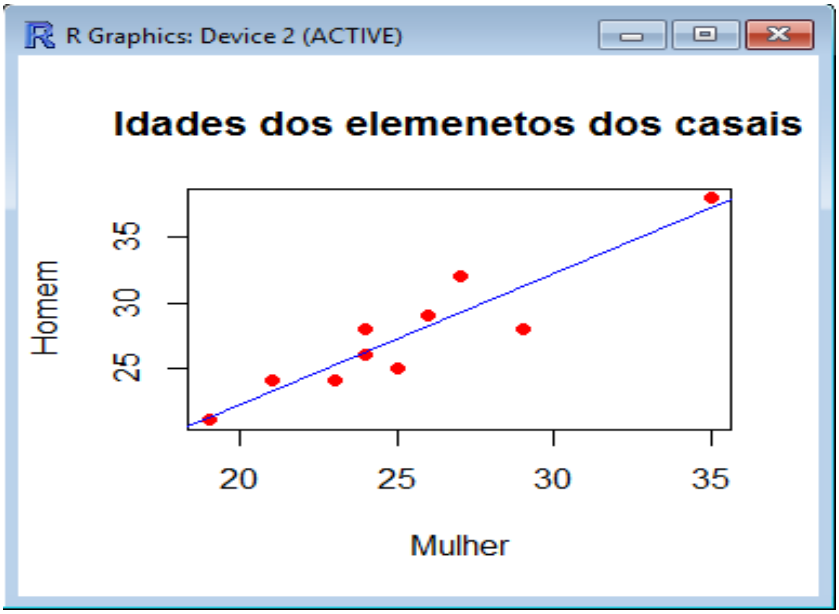
> plot(mulher, homem, xlab = "Mulher", ylab = "Homem", main = "Idades dos Elemenetos dos Casais", pch = 16, col = "red")

> abline(lm(homem mulher), col = "blue")

> cor(mulher, homem)

[1] 0.9263033

O diagrama que pode ser visto é o seguinte:



Os diagramas apresentadas pelo SPSS e pelo **R** são do mesmo tipo, isto é, os dois são idênticos. Pode concluir-se pelo valor de $r > 0$ ou $r = 0.93$, que existe uma correlação linear positiva ou seja, quando a idade do homem aumenta, também aumenta linearmente a idade da mulher.

Capítulo 4

Distribuições de probabilidade

4.1 Variáveis aleatórias discretas

4.1.1 Distribuição binomial

Exercício 14: Num saco há quatro bolas vermelhas e duas azuis, indistinguíveis ao tacto. Considere a seguinte experiência: "tirar uma bola do saco, tomar nota da cor e repor a bola ". Determine a probabilidade de, repetindo a experiência 10 vezes, tirar duas, e só duas, bolas azuis. (ME-TL. 2014, p.155. Tarefa 80)

Resolução com SPSS e R

Antes de resolver este exercício com SPSS e **R**, deve definir primeiro a variável de interesse e o acontecimento p ocorrido em n experiências. A variável de interesse, neste caso, é o X, que corresponde à variável aleatória (número de vezes que se tira a bola azul em 10 extrações). A variável de interesse é o número de sucesso em 10 tentativas, logo, a variável X segue uma distribuição binomial em que n corresponde a 10 tentativas e p é à probabilidade de tirar uma bola azul, ou seja, p=prob(sair uma bola azul na extração) $\frac{2}{6} = \frac{1}{3} \approx 0.35$ (o valor mais próximo que aparece nas tabelas do manual).

A v.a. X segue uma distribuição binomial pode ser representada como, $X \sim B(10; 0.35)$.

Resolução em R

Os comandos a serem executados devem ser os seguintes:

```
> n = 10

> n

  [1] 10
```

> p = 0.35

> p

[1] 0.35

> dbinom (2, size = n, prob = p)

[1] 0.175653

O **R** permite obter este valor como parte da tabela do manual através da execução dos seguintes comandos:

> dbinom (0:10, size = n, prob = p)

[1] 1.346274e-02 7.249169e-02 **1.756530e-01** 2.522196e-01 2.376685e-01

[6] 1.535704e-01 6.890980e-02 2.120302e-02 4.281378e-03 5.123017e-04

[11] 2.758547e-05

O **R** permite representar num gráfico a função de probabilidade da distribuição binomial (Figura a seguir).

> n = 10

> n

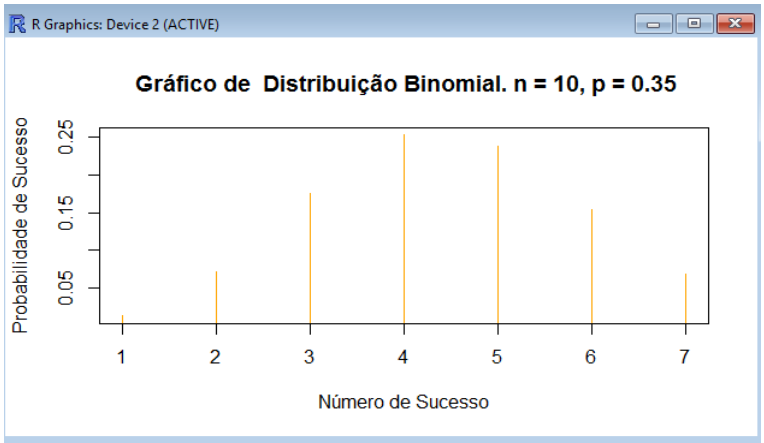
[1] 10

> p = 0.35

> p

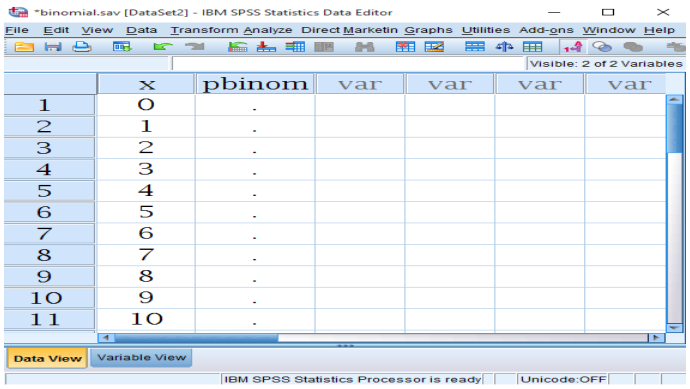
[1]0.35

> plot(dbinom (seq(0,6, by = 1), size = n, prob = p), type = ”h”, col = ”orange”, xlab = ”Número de Sucesso ”, ylab = ”Probabilidade de Sucesso”, main = ”Gráfico de Distribuição Binomial. n = 10, p = 0.35”)



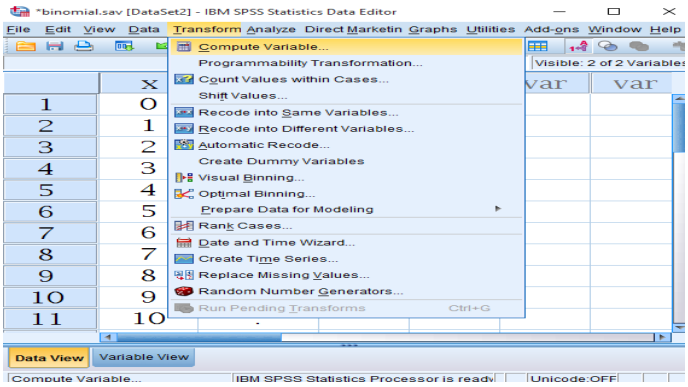
Resolução com SPSS

Insira os parâmetros da distribuição binomial no SPSS. Neste caso, x é considerado como variável que representa o número de bolas azuis extraídas, em 10 tentativas, logo, os valores que o x pode assumir são: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10. A variável *pbinom* corresponde à função de probabilidade da variável aleatória x. Como se mostra na janela seguinte:

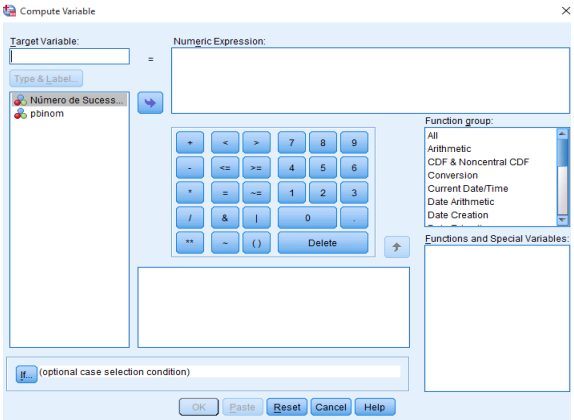


Para calcular os valores de função da probabilidade que o *pbinom* vai ter, pode executar os seguintes comandos:

- Transform/Compute Variable...

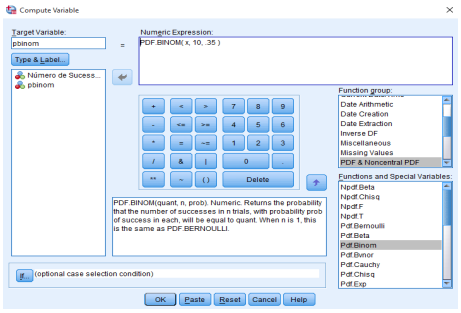


- Function Group: PDF and non central PDF
- Function and Special Variable: Pdf. Binom (Dupla clique)

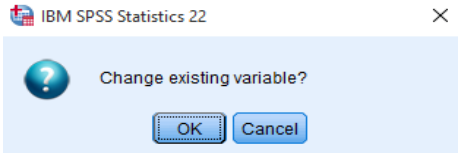


- Preencher os campos da função: Pdf. Binom(x, $n = 10$, $p = .35$)
- Target Variable: pbinom
- OK

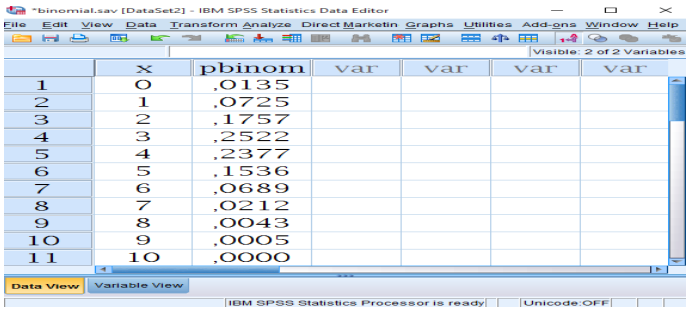
A janela seguinte mostra o processo de aceder e preencher a função Pdf.Binom $(x, 10, .35)$



Depois de clicar **OK** vai aparecer uma janela **Change Existing Variable?**, nela deve executar **OK**.

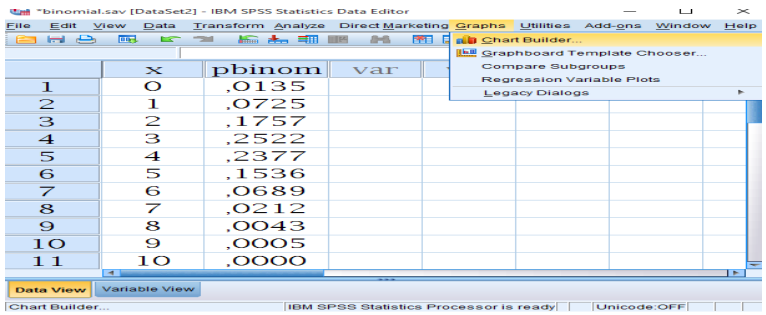


O resultado da análise está apresentado na seguinte janela:



Para construir o gráfico da função de probabilidade de distribuição Binomial para valores de p = 0.35 e com n = 10 no SPSS, pode executar os seguintes comandos:

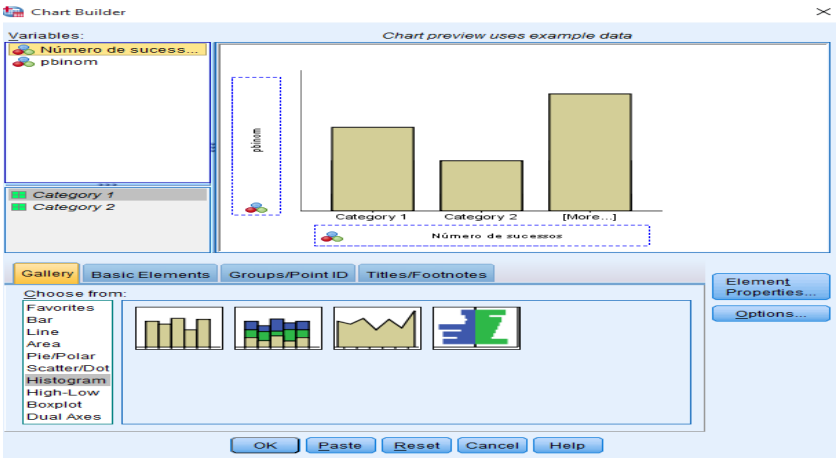
- Graphs/Chart Builder...



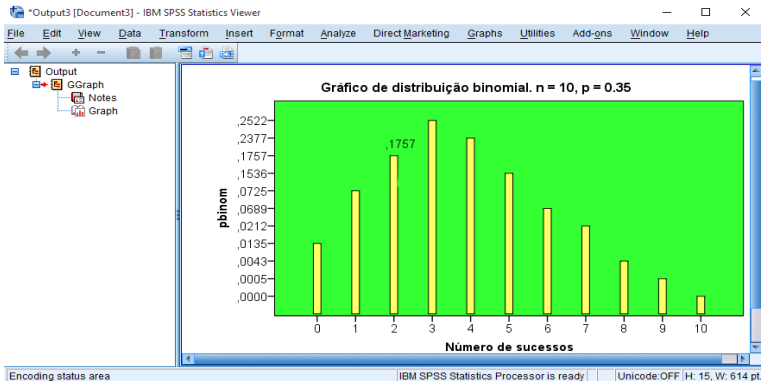
- Choose from: Histogram

  Escolher o Histograma, dupla clic no **Simple Histogram** e em seguida arrastar o Número de sucessos (x) e colocá-lo em X-Axis, Feito isto, arrastar a variável probabilidade (pbinom) e colocá-la em Y-Axis

- OK



Depois de clicar **OK**, vai aparecer uma janela com o gráfico básico. Para o editar pode clicar duas vezes na área do gráfico para aparecer **Chart Editor**. A janela do gráfico está apresentada a seguir:



A respeito dos resultados analíticos dados pelos SPSS e **R**, pode ser concluído que: a probabilidade de saírem exatamente duas bolas azuis em dez tentativas é aproximadamente **0.1757** ou **17.57%**.

4.1.2 Distribuição de Poisson

Exercício 15: O número de reclamações que uma lavandaria recebe por dia é uma variável aleatória seguindo uma distribuição de Poisson com parâmetro 2.5. Determine a probabilidade da lavandaria não receber reclamações num dia. (ME-TL. 2014, p.157.

Tarefa 82)

Resolução com SPSS e R

Seja X a v.a. que representa o número de reclamações que a empresa recebe por dia. Temos que calcular $P(X = 0)$ no exercício. A v.a. X segue uma distribuição de Poisson com média $\lambda = 2.5$.

Resolução em R

> lambda = 2.5

> lambda

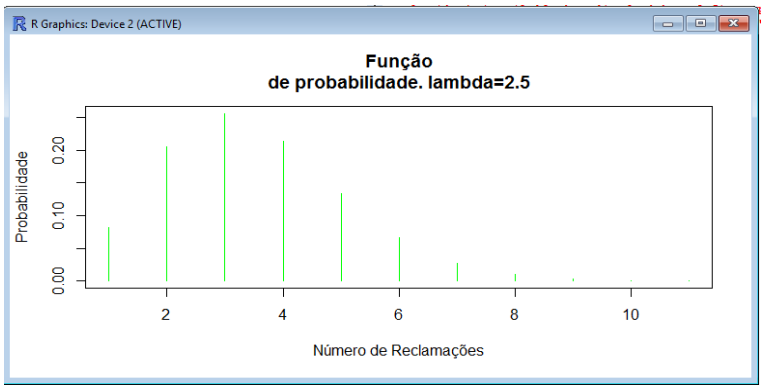
[1] 2.5

> dpois(0, lambda)

[1] 0.082085

Gráfico da função de probabilidade da distribuição de Poisson em R

> plot(dpois(seq(0, 10, by = 1), lambda = 2.5), type = "h", col = "green", xlab = Número de Reclamações", ylab = "Probabilidade", main = "Função de probabilidade para lambda = 2.5")



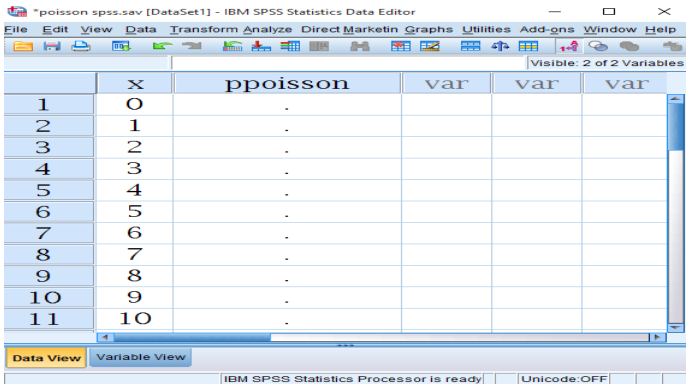
O programa **R** permite outra maneira de conseguir o mesmo valor através da execução dos seguintes comandos:

> dpois(0:5, lambda = 2.5)

[1] **0.08208500** 0.20521250 0.25651562 0.21376302 0.13360189 0.06680094

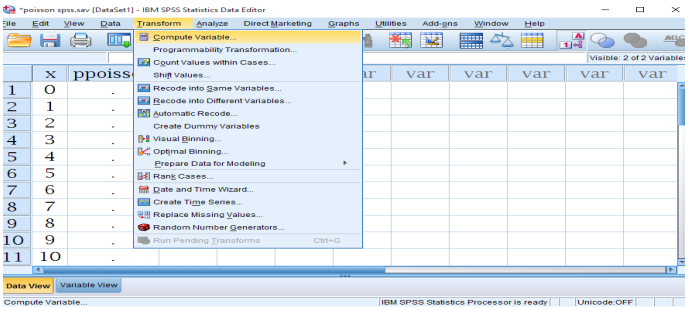
Resolução em SPSS

Insira os parâmetros da distribuição de Poisson no SPSS. Neste caso x é uma variável que representa o número de reclamações e que consideramos tomar os valores: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10. A variável *ppoisson* corresponde à função de probabilidade $P(X = x)$, como se mostra na janela seguinte:

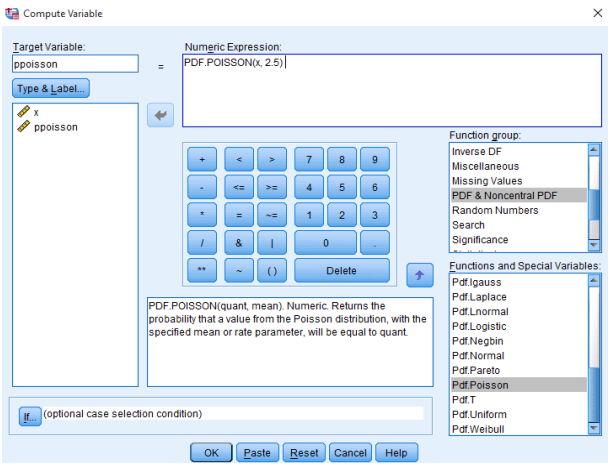


Para calcular os valores da probabilidade $P(X = x)$, pode executar os seguintes comandos:

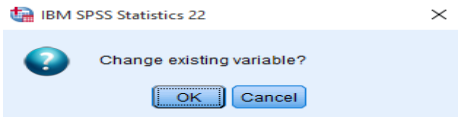
- Transform/Compute Variable



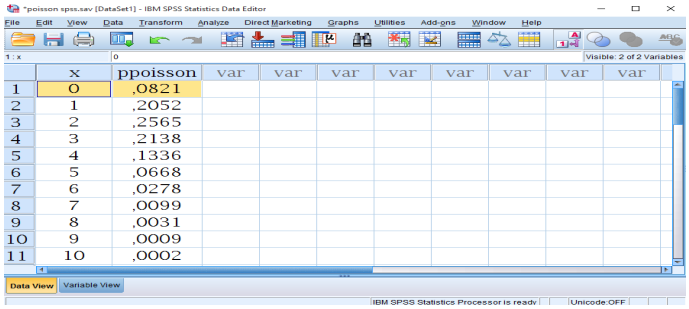
- Function Group: PDF and non central PDF
- Function and Special Variable: Pdf poisson
- Preencher os campos da função: Pdf.poisson (x, 2.5)
- Target Variable: ppoisson
- OK



Depois de clicar **OK**, vai aparecer uma janela **Change Existing Variable?**. A seguir deve executar **OK**.



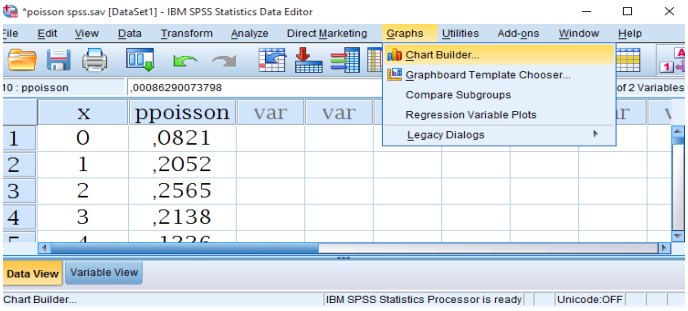
O resultado da análise está apresentado na seguinte janela:



	x	ppoisson
1	0	,0821
2	1	,2052
3	2	,2565
4	3	,2138
5	4	,1336
6	5	,0668
7	6	,0278
8	7	,0099
9	8	,0031
10	9	,0009
11	10	,0002

Para construir o gráfico da função de distribuição de poisson no SPSS pode executar os seguintes comandos:

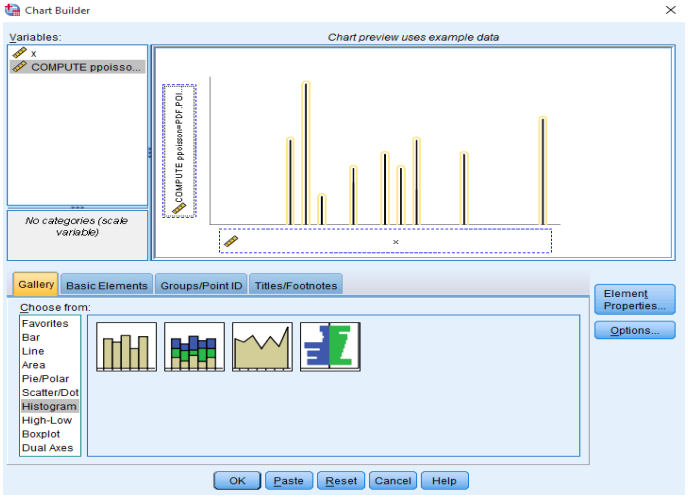
- Graphs/Chart Builder...



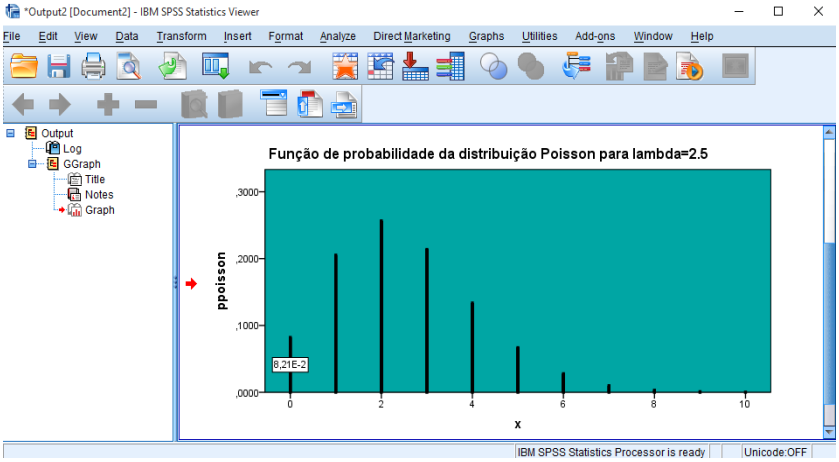
- Choose from: Histogram

  Escolher o Histograma e clicar no Simple Histogram e em seguida arrastar o número de reclamações e colocá-lo em X-Axis. Feito isto, arrastar a variável COMPUTE ppoisson=PDF... e coloca-la em Y-Axis.

- OK



Depois de clicar **OK**, vai aparecer uma janela com o gráfico básico. Para o editar, pode clicar duas vezes na área do gráfico para aparecer **Chart Editor**. A janela do gráfico está apresentada a seguir:



A observação dos resultados analíticos dados pelos SPSS e **R** permite concluir que a probabilidade da empresa não receber reclamações num dia é, aproximadamente, **0.0821** ou **8.21%**.

4.2 Variáveis aleatórias contínuas

4.2.1 Distribuição normal

Exercício 16: A distribuição dos pesos dos alunos de uma escola segue uma distribuição normal com $\mu = 64$ e $\sigma = 10$ em Kg. Determina a percentagem de alunos que pesam entre 54 Kg e 74 Kg. (ME-TL. 2014, p.160. Tarefa 85).

X: variável aleatória que representa os pesos dos alunos da escola. X é uma variável aleatória com distribuição normal de média $\mu = 64$ Kg e desvio padrão é $\sigma = 10$ Kg.

Resolução à mão:

O intervalo entre 54 Kg e 74 Kg pode ser identificado como sendo o intervalo entre $\mu - \sigma$ e $\mu + \sigma$. Logo, a percentagem pedida é 68.27%, tal como consta no manual.

Apresentamos de seguida resoluções usando o SPSS e o **R**, mais como curiosidade para ilustrar as potencialidades destes programas. Os métodos exibidos são facilmente generalizáveis para quaisquer outros intervalos.

Resolução com SPSS:

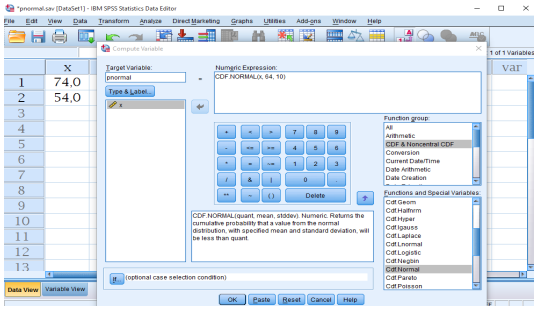
Depois de abrir o SPSS, pode começar por definir ou criar os nomes das variáveis na janela de **Variable View**. Neste caso a primeira variável é X. Depois de definir a variável

X, pode abrir o Data View localizado no fundo, à esquerda do Variable View, para poder aceder à coluna da variável X. Ainda nessa coluna, introduzir o número 74 na primeira linha e primeira coluna e o outro número 54 vai ser colocado na segunda linha e primeira coluna.

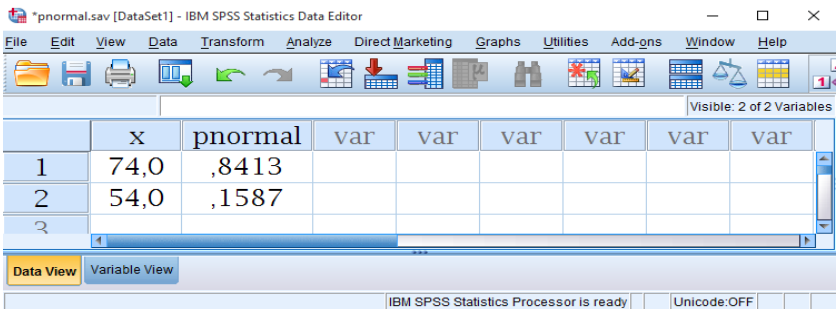
Para obter o valor de probabilidade da distribuição de normalidade dos pesos dos alunos executam-se os seguintes comandos:

- Transform/Compute Variable ...
- Function Group: CDF and non central CDF
- Function and Special Variable: Cdf. Normal
- Preencher os campos da função/Numeric Expression: CDF. NORMAL $(x, 64, 10)$
- Target Variable: pnormal
- OK

A janela seguinte mostra o processo de aceder e preencher a função CDF. NORMAL $(x, 64, 10)$.

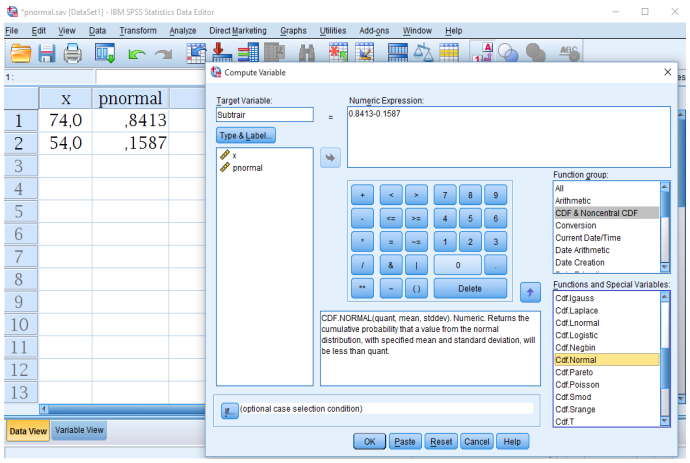


Depois de clicar **OK**, vai aparecer a janela **Change Existing Variable?**, A seguir deve clicar **OK**. O resultado da análise está apresentado na seguinte janela:

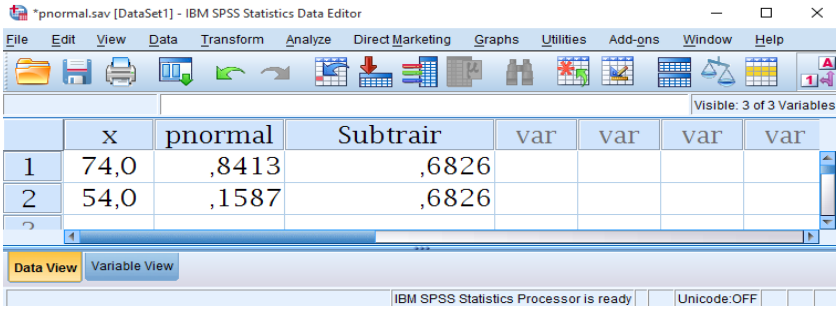


O resultado da análise mostrou que o valor $P(X < 74)$ está a corresponder a 0.8413, enquanto $P(X > 54)$ equivale a 0.1587. Para calcular a percentagem de alunos que pesam entre 54 Kg e 74 Kg, basta subtrair o valor 0.1587 a 0.8413, como mostra no comando abaixo:

- Transform/Compute Variable
- Preencher os campos da função/Numeric Expression:(0.8413 - 0.1587)
- Target Variable: Subtrair
- OK



Novamente se verifica que a probabilidades obtida é aproximadamente 0.6827, mais concretamente 0.6826 neste caso, como mostra a seguinte janela:



Resolução com R

Para calcular a probabilidade dos valores entre 54 Kg e 74Kg, pode-se executar o seguinte comando:

```
> pnorm(74, 64, 10) - pnorm(54, 64, 10)
  [1] 0.6826895
```

Sugestão: Para gerar os valores associados à Distribuição Normal Padrão, pode-se executar o comando abaixo.

```
> pnorm(1:5)
  [1] 0.8413447 0.9772499 0.9986501 0.9999683 0.9999997
```

O **R** contém os comandos para construir o gráfico da função densidade de probabilidade da distribuição normal, como se mostra a seguir:

> x = seq(44, 84, by = .01)

> y = dnorm(x, mean = 64, sd = 10, log = FALSE)

> rx = seq(54, 74, by = .1)

> ry = numeric(2 ∗ length (rx))

> ry [1 : length(rx)]

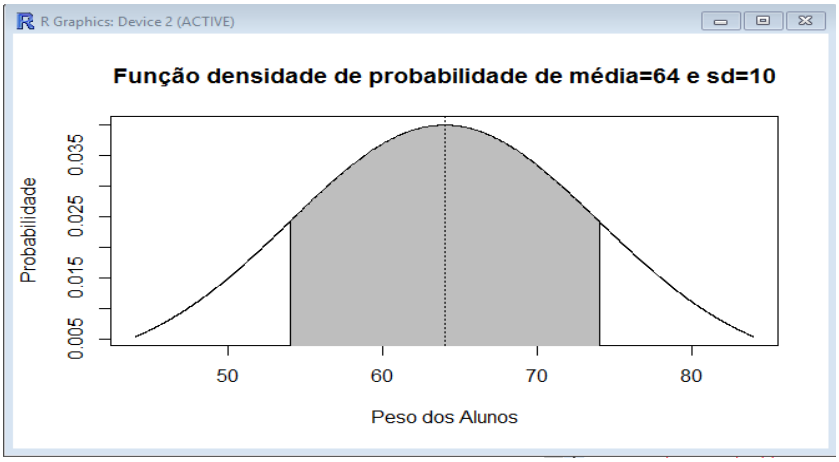
1. <= dnorm(rx, mean = 64, sd = 10, log = FALSE)

> rx = c(rx, rev(rx))

> plot(x, y, "l", xlab = "Peso dos Alunos", ylab = "Probabilidade", main = Função densidade de probabilidade de média = 64 e sd = 10 ")

> polygon(rx, ry, col = "gray")

> abline(v = 64, h = 0, lty = 3)



A percentagem de alunos de uma escola que pesam entre 54 Kg e 74 Kg é **0.6827**, ou seja, **68.27 %**. Este resultado está representado na parte sombreada do gráfico.

Conclusão: Observando estas duas resoluções(SPSS e **R**) percebe-se que se chega sempre à mesma conclusão, isto é, que a probabilidade de que os alunos pesarem entre 54 Kg e 74 Kg é 68.27%. Para além disso, bastaria modificar os valores 54 e 74 nas instruções para se determinar a probabilidade relativa a intervalos diferentes.

Capítulo 5

Conclusões e Sugestões

5.1 Conclusões

Este trabalho seria uma referência inicial importante para desenvolver o conhecimento e aumentar a capacidade do pesquisador na utilização dos programas de estatística. O SPSS e **R** são ferramentas que constituem uma mais-valia e que podem ser utilizadas pelo pesquisador na sua actuação diária como professor de matemática, principalmente na disciplina de estatística. Vão ser úteis também para todas as pessoas, principalmente os alunos, que estão interessadas em trabalhar com a estatística, utilizando estes programas no seu quotidiano, ou na sala de aula.

A aprendizagem da estatística na sala de aula influencia positivamente o desenvolvimento do raciocínio e domínio dos alunos nesta matéria. Para isso, requer-se que o ensino de estatística seja atrativo pelos métodos e instrumentos auxiliares. Alguns desses instrumentos são o programa de estatística SPSS e linguagem **R**. Esses programas de estatística auxiliam na resolução dos problemas, quer aos professores quer aos alunos, tanto na sala de aula como no profissional.

O ensino de estatística acompanhado pelos referidos programas é mais inovador, atrativo, criativo, fácil e motivador. Deste modo, o ensino de estatística não se vai centrar apenas na transmissão de conhecimentos como: contagem de números, operações com algarismos, elaboração de tabelas e gráficos sem significado para os alunos.

Embora estes programas não venham a ser utilizados no exame trimestral ou nacional, eles são ferramentas alternativas para professores compreenderem os conceitos de estatística. Como exemplo, poderei dizer que muitos alunos já calculam, manualmente, a média e a mediana dos dados de uma ou mais variáveis da mesma população ou amostra. Contudo, têm dificuldade em estabelecer a relação, ou fazer a comparação entre os valores da média e da mediana sem os visualizar num gráfico. Com o apoio destes programas, os alunos podem comparar as diferenças das duas medidas através de gráficos apresentados

pelos programas.

Estes programas serão ferramentas poderosas no ensino de estatística na sala de aula. Embora a facilidade de contagens e construções de tabelas e gráficos pelos programas não sejam objectivos principais no ensino e aprendizagem de estatística, estas ferramentas, se devidamente utilizadas, vão funcionar como meio para desenvolver conceitos, raciocínio, facilitar a resolução de problemas e o cálculo de grandes quantidades de dados e permitir aos professores a fácil interpretação dos mesmos.

5.2 Sugestões ao governo

Para facilitar o pôr em funcionamento a utilização destes programas, pelos professores e alunos, nas escolas em Timor-Leste, é necessário que o governo faça um plano de investimento na educação, através da instalação de redes de computadores, e facilite aos professores a aquisição de mais formação, tanto no interior do país como no exterior, principalmente na utilização dos programas no ensino de estatística, sobretudo SPSS e **R**. É também necessário fazer uma revisão imediata dos manuais escolares de matemática, principalmente o manual do aluno do 12º ano de escolaridade de Timor-Leste , para salientar a importância dos programas de matemática, sobretudo dos programas de estatística, no ensino e aprendizagem.

5.3 Sugestões aos futuros pesquisadores

NCTM (1991), defende que a pesquisa académica nunca está completa ou acabada. Ela poderá sempre ser refeita, considerando outros aspectos, ou actualizada com novas informações, tecnologia dos computadores e programas de estatística, tanto qualitativa quanto quantitativamente, de acordo com o momento social e político em que ela é desenvolvida. Esta pesquisa não pretende ser diferente. A sua principal contribuição é servir de fonte de informações, em relação à formação dos alunos existentes em Timor-Leste e para outros pesquisadores que poderão, no futuro, aprofundar e melhorar o que aqui está registado.

Os exercícios, colectados e analisados pelos SPSS e **R** neste trabalho, buscam também ser uma fonte de consulta para os próximos pesquisadores, ao elaborar e desenvolver esta pesquisa. Ela foi um estudo bibliográfico realizado no âmbito do curso de Mestrado em Matemática para Professores da Faculdade de Ciências da Universidade do Porto. Pensando assim, estão apresentadas, abaixo, três sugestões para o próximo pesquisador de Timor-Leste que tenha interesse em continuar ou melhorar esta pesquisa:

1. Seria melhor usar estes programas de estatística para realizar pesquisa nas escolas,

em Timor-Leste, tanto no ensino básico como no ensino superior. Assim, o futuro pesquisador poderia ter mais conhecimentos acerca das dificuldades e capacidade de raciocínio apresentadas pelos professores e alunos, na sala de aula, de forma individual ou em trabalho de grupo.

Bibliografia

[1] Afonso, Anabela e Nunes Carla (2011). *Estatística e Probabilidades: Aplicação e Soluções em SPSS*. Lisboa: Escolar Editoral.

[2] Ferreira, Maria João e Tavares, Isabel. (2009). *Um Mundo Para Conhecer Os Números: Dossiê*. Acedido a 9 de Dezembro de 2015, ás 01:21, de http://www.alea.pt/html /statofic/html/dossier/doc/publicacao__2009__web.pdf.

[3] Fernandes, Susana E Pinto, Mónica, *Afinal, o que são e como se calculam os quartis?*, Universidade do Algarve acedido do http://gazeta.spm.pt/getArtigo?gid=468

[4] Laureano, Raul M. S e Botelho, Maria do Carmo. (2012). *SPSS: O Meu Manual de Consulta Rapidas*, 2ª Edição. Lisboa: Edições sílabo.

[5] Maroco, João. (2003). *Análise Estatística: Com Utilização do SPSS*. Lisboa: Edições Sílabo.

[6] Ministério da Educação de Timor-Leste (ME-TL, 2014). *Matemática 12º Ano de Escolaridade: Manual do Aluno*. 1ª edição.

[7] *National Council of Teachers of Mathematics (1991). Normas para Currículo e a Avaliação em Matemática Escolar*. Lisboa: APM e IIE (tradução portuguesa dos Standards do NCTM, 1989).

[8] Nunes, Cristina. F. (2012). *Probabilidades e Estatística: 275 Problemas Resolvidos (Utilização* **R***)*. Lisboa: Escolar Editora.

[9] Pestana, Dinis Duarte e Velosa Sílvio Filipe. (2010). *Introdução à Probabilidade e à Estatística*.Vol I. 4ª Edição. Lisboa: Fundação Calouste Gulbenkia.

[10] Ponte, Joao Pedro. (1997). *As Novas Tecnologias e A Educação*. Lisboa: Texto Editora.

[11] Torgo, Luis. (2009). *A Linguagem* **R***: Programação para a Análise de Dados*. Lisboa: Escolar Editora.

Anexo A

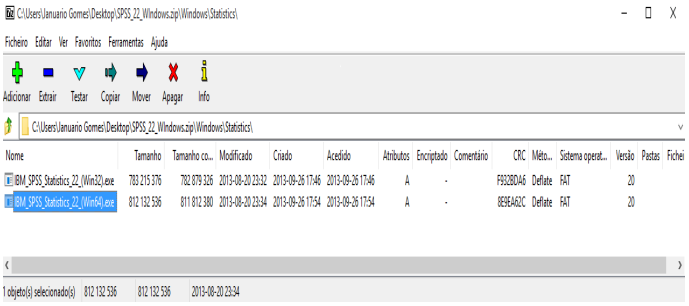
Baixar e instalar o IBM SPSS Statistics 22

Nesta secção vai ser apresentado o processo de aceder, baixar e instalar o programa de estatística no servidor da U. Porto. O IBM SPSS Statistics 22, utilizado neste estudo, é um programa que se encontra disponível para baixar no sítio http://atlas.up.pt.
Nota: Antes de instalar este programa, os estudantes devem saber que:

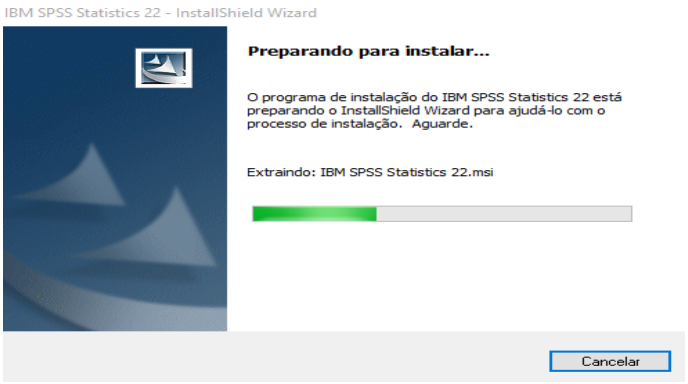
1. É necessário estar inscrito, ou seja, pertencer a esta universidade;
2. A máquina utilizada, sobretudo o computador, tem, obrigatoriamente, de estar configurada pelo Centro de Informática (CI) da FCUP.

Depois de baixar este programa no computador, pode efectuar-se o seguinte processo de instalação:

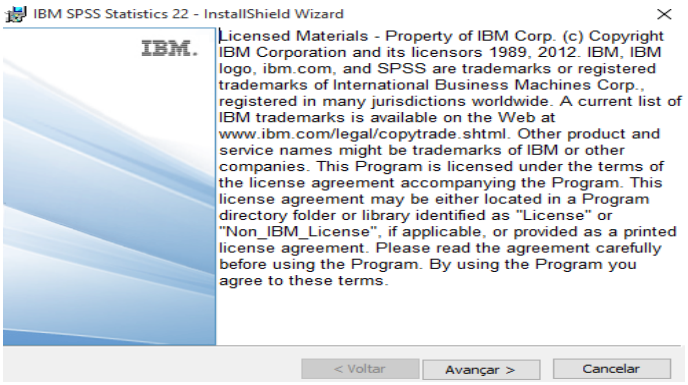
- Abrir o programa IBM SPSS statistics 22 e sobre ele efectuar duplo clic para iniciar a instalação.



Depois de clicar duas vezes sobe o programa, aparece uma janela que mostra que o processo de instalação está a iniciar (ver janela seguinte):



Aguardar que o processo de iniciação decorra, até aparecer a janela *Licensed Materials*, conforme figura abaixo. Pressione o botão **Avançar**.

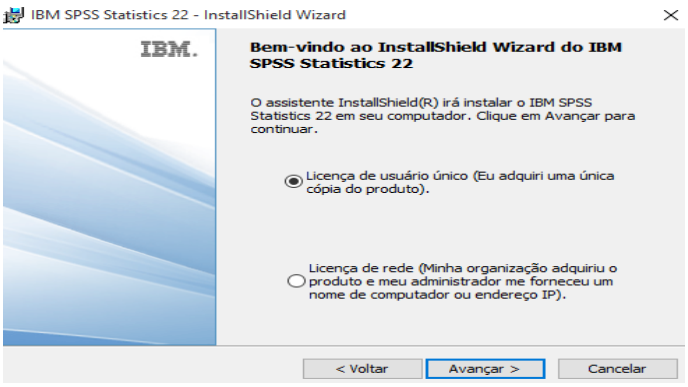


Na janela seguinte, escolher:

- ⊙ Licença de usuário único

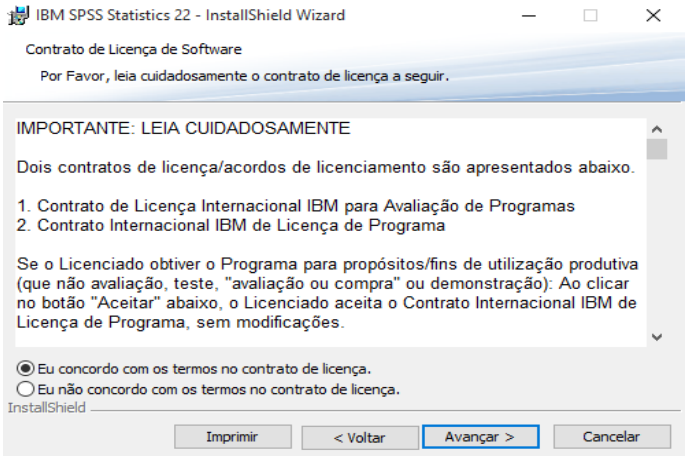
  Nota: Existem duas opções, cada usuário tem a possibilidade de escolher uma delas.

- Avançar



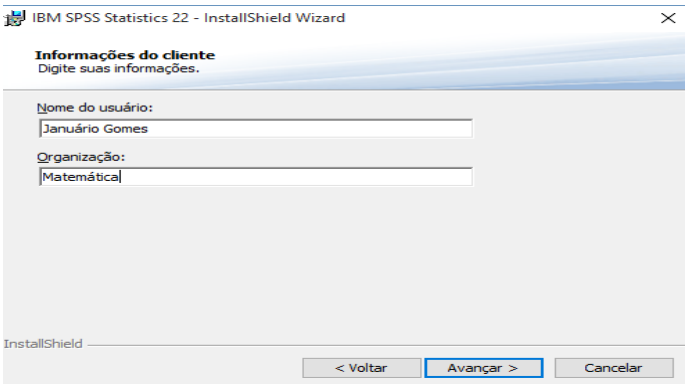
Depois de pressionar o botão **Avançar**, aparece a janela *Informações do cliente*. Podem ser deixadas em branco as duas linhas: *Nome do usuário* e *Organização*. Neste caso, foram preenchidas, respectivamente, por *Januário Gomes* e *Matemática*.

- Seleccionar Avançar

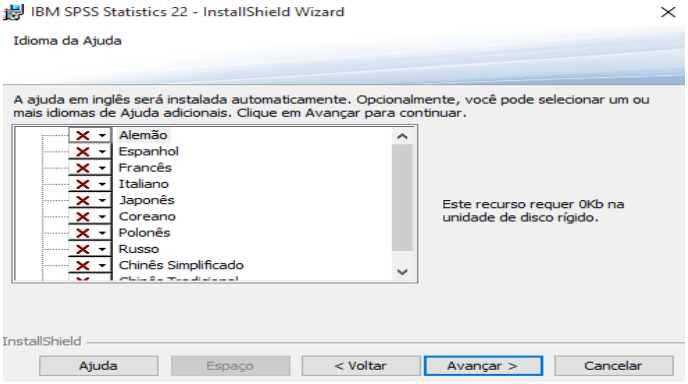


Em seguida, clicar em:

- ⊙ Eu concordo com os termos no contrato de licença
- Avançar

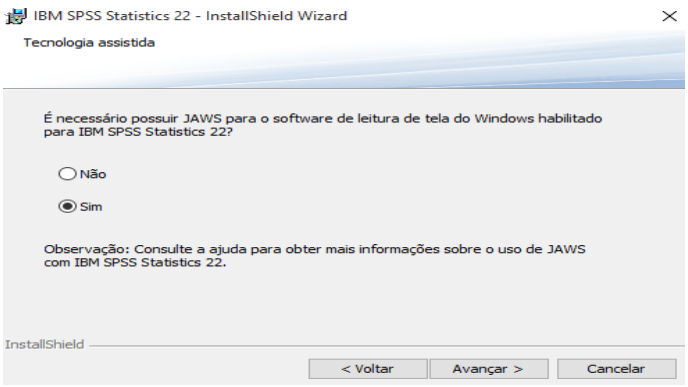


- Seleccionar Avançar



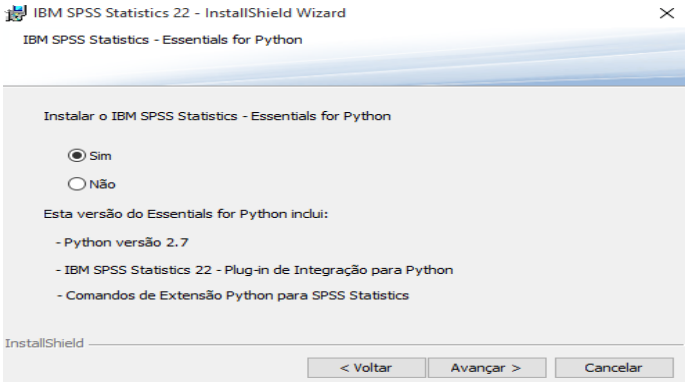
Seguidamente, aparece o menu que pergunta: *É necessário possuir JAWS?* As opções que devem ser assinaladas são:

- ⊙ Sim
- Avançar



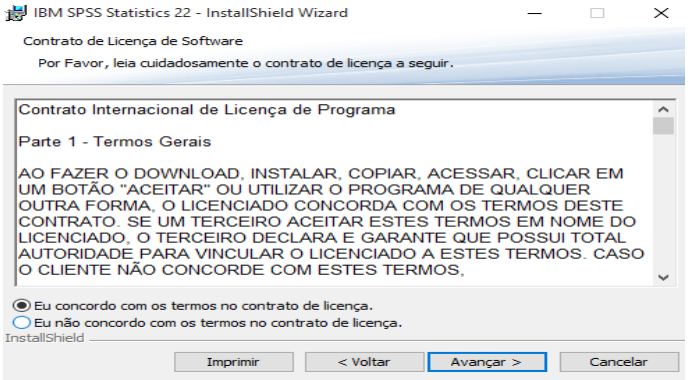
Novamente, na janela seguinte, escolher:

- ⊙ Sim
- Avançar



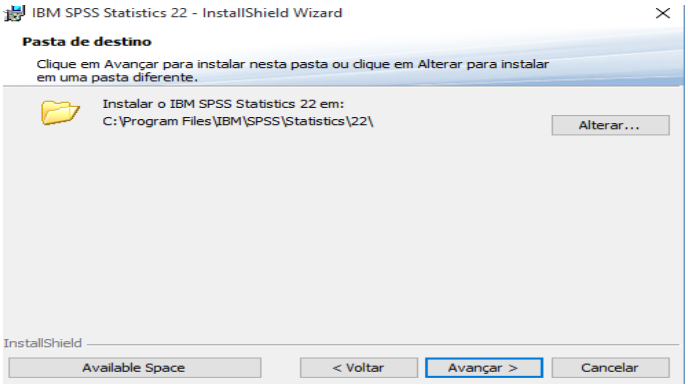
No menu seguinte:

- ⊙ Eu concordo com os termos no contrato de licença
- Avançar

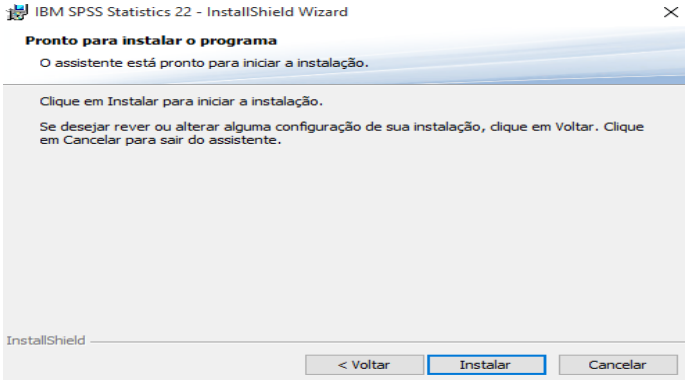


Para iniciar a instalação, escolher a opção:

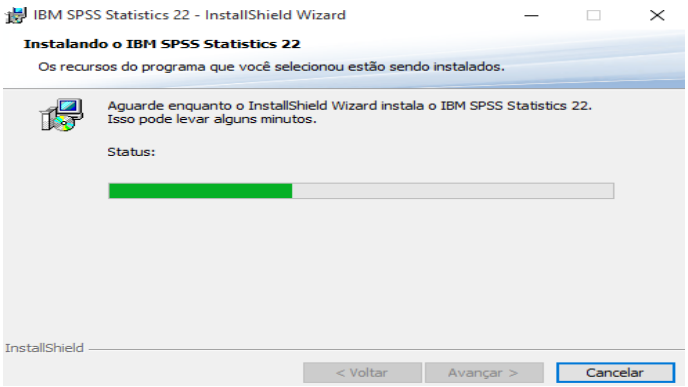
- Avançar



- Instalar



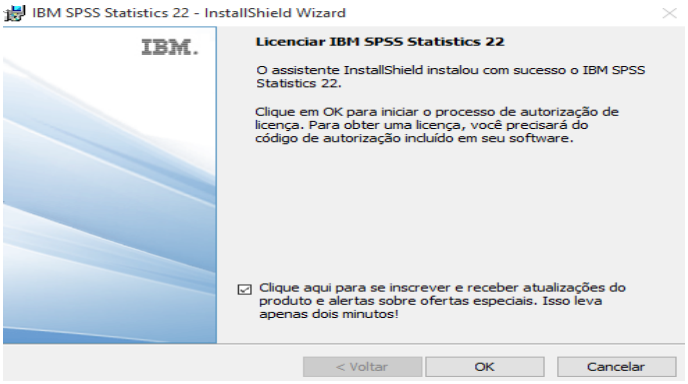
A instalação do programa de SPSS está a começar e deve-se aguardar até o processo finalizar.



Depois de o programa finalizar a instalação, aparece a janela seguinte, onde se deve seleccionar:

- ☑ Clique aqui para se inscrever ...

- OK

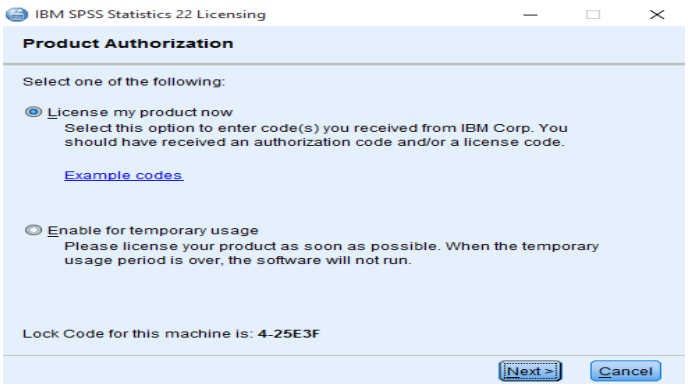


Na janela *Product Autorizatio*, seleccionar:

- ⊙ License my product now

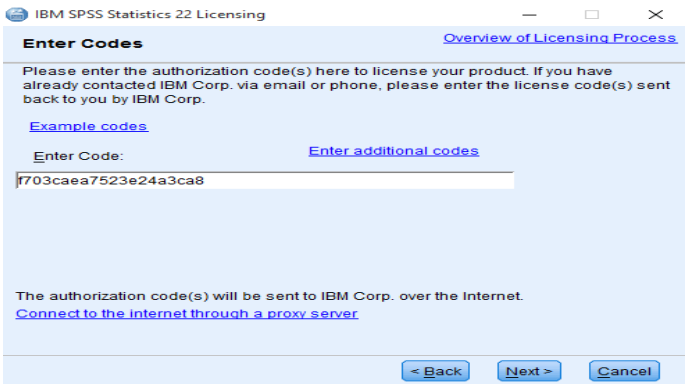
Nota: Foi escolhida a primeira opção porque este programa de SPSS foi baixado no servidor da Universidade do Porto e também se encontra disponível para activar o código de autorização.

- Next

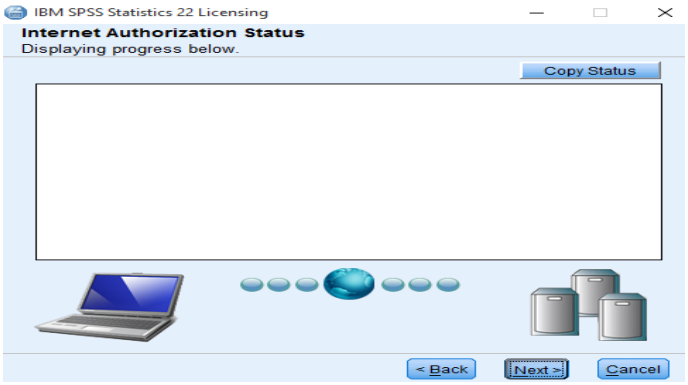


- Inserir a chave de autorização.

- Next

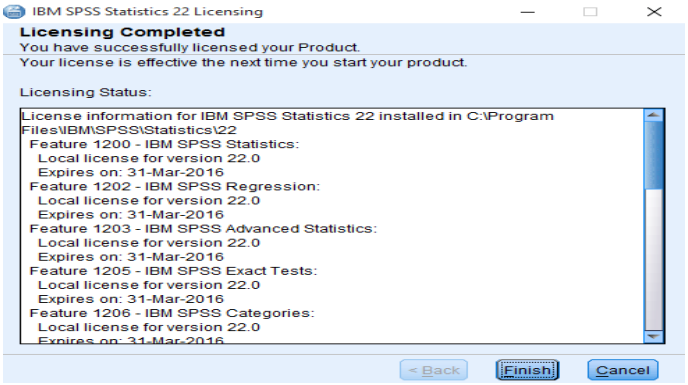


- Next



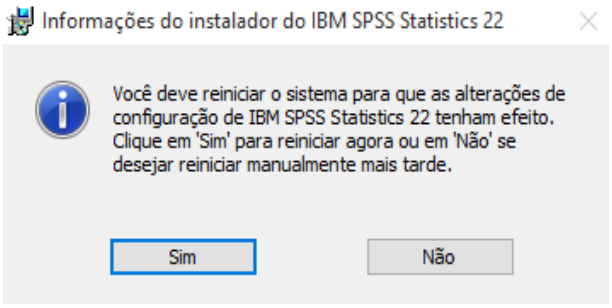
Na janela seguinte aparece *Licensing Completed*. Pressionar:

- Finish



Finalmente, aparece uma informação que sugere reiniciar a máquina ou computador. A melhor opção a escolher é:

- Sim

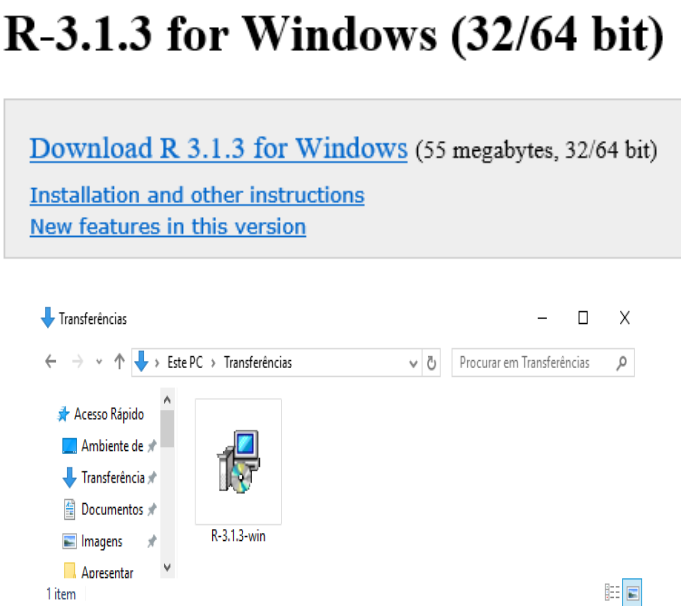


Anexo B

Baixar e instalar o R

Para utilizar o **R**, primeiramente deve instalar este programa na máquina ou computador do usuário como se mostra a seguir:

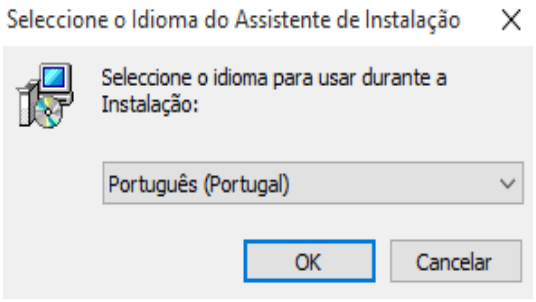
1. Aceder à página oficial do programa, http://www.r-project.org/.
2. Baixar o programa Download R-3.1.3 for Windows (ver as duas janelas seguintes).



Sugestão: Os iniciantes que querem instalar este programa no computador, é melhor tentarem obter a versão actual na página do projecto **R**, na internet, porque este programa, em qualquer momento, aparece disponível em nova versão. Os utilizadores que já o têm no computador, devem conferi-lo, para o actualizar pela nova versão. Depois de baixar o programa no computador, podem executar-se as seguintes instruções, para iniciar a instalação:

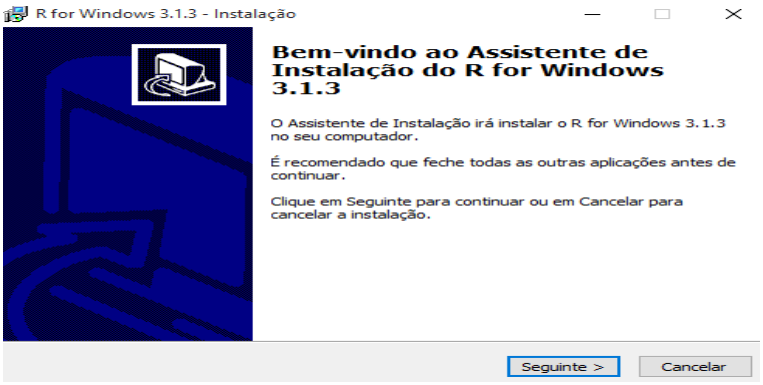
- Duplo clic no programa;

- Seleccionar idioma. Neste caso, foi escolhido português (Portugal).
- OK



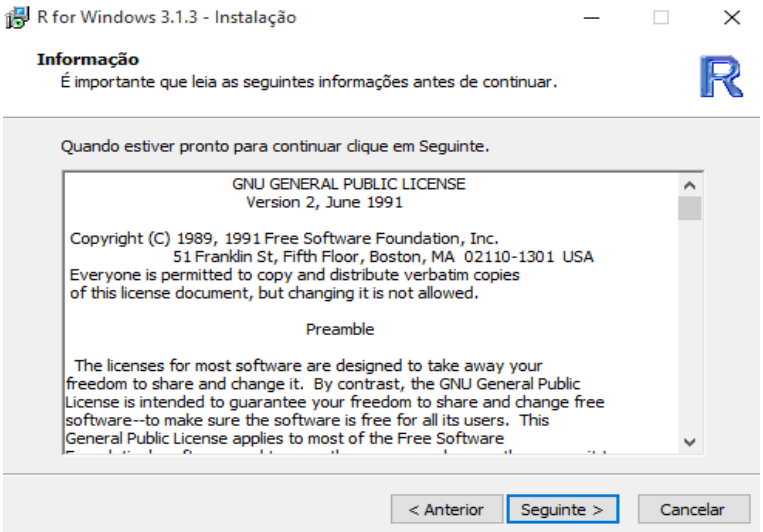
Na janela que se segue:

- Seguinte



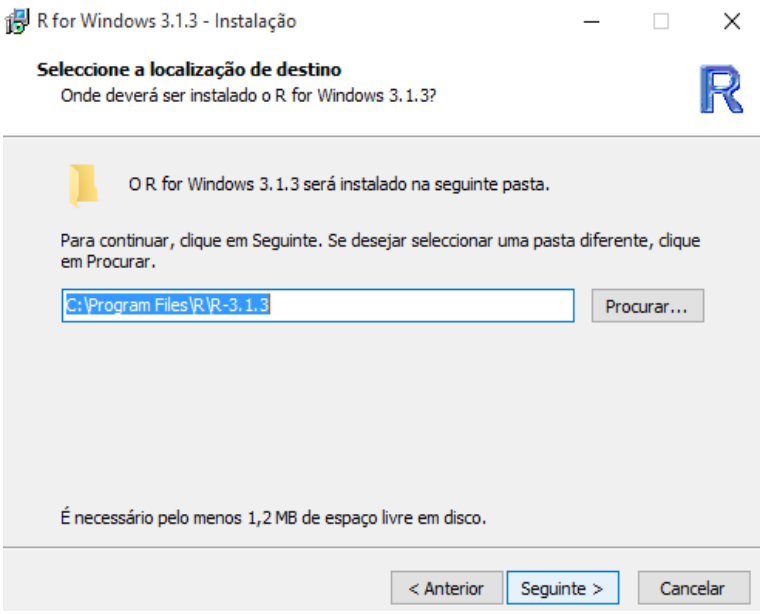
Depois de clicar **Seguinte**, aparece uma janela de informação que sugere aos usuários a leitura das informações, antes de aceitar os termos do programa. Para continuar, basta seleccionar:

- Seguinte



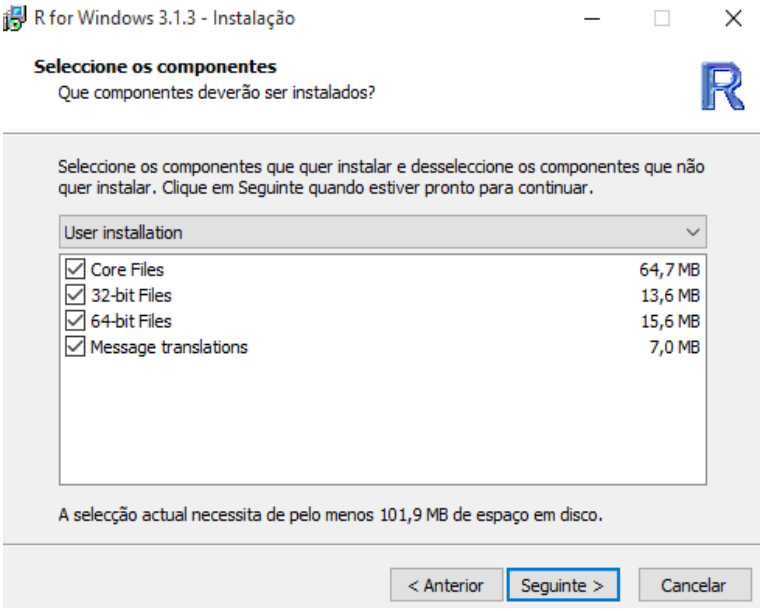
Na janela seguinte seleccione a localização de destino:

- Seguinte



- Seguinte

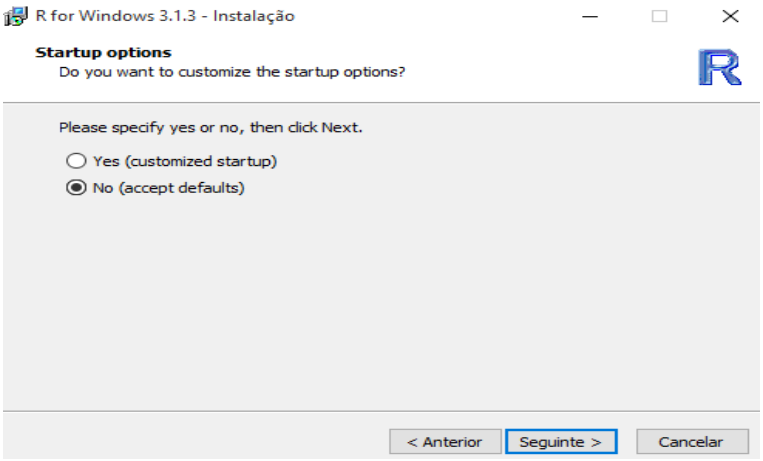
O menu seguinte pergunta que componentes deverão ser instaladas. Sem precisar de escolher nenhuma destas opções, basta executar:



Mais uma vez aparece nova janela que apresenta uma pergunta com duas opções Yes (customized startup) e No (accept defaults). Nela pode ser feito:

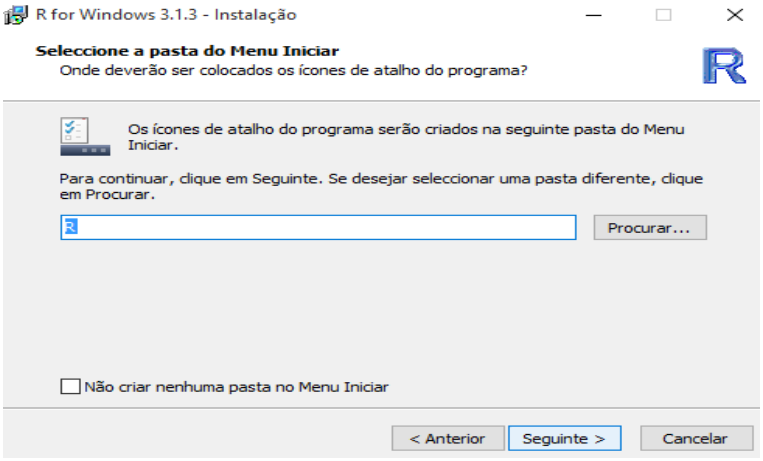
- No (accept defaults) - a escolha desta opção, mostra que o usuário está a manter a opção de aceitar o padrão deste programa.

- Seguinte

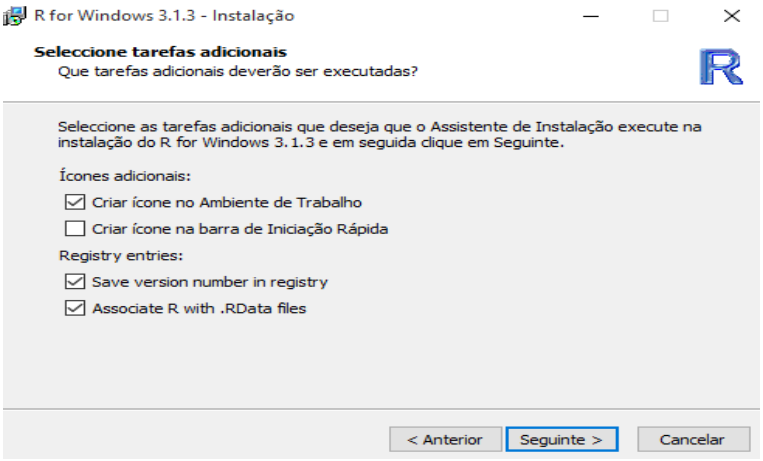


Na janela " Seleccione a pasta do menu iniciar", basta clicar em:

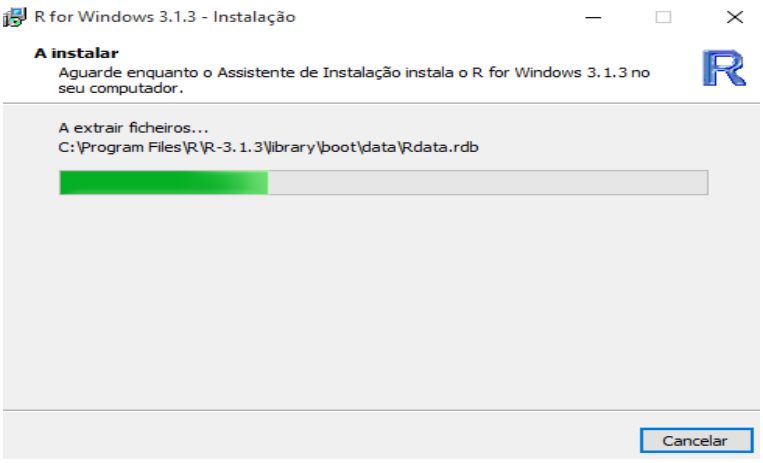
- Seguinte



- Seguinte



Aguardar o processo de instalação.



Para finalizar a instalação, pressionar o botão **Concluir**.