

# **GUIA PRÁTICO SOBRE PRINCÍPIOS BÁSICOS PARA ESTATÍSTICA APLICADA EM PESQUISAS SOCIAIS E DA SAÚDE**

Por:

Alcides Domingos Cambundo

$$\bar{x}_p = \sum_{i=1}^n \frac{f_i * x_i}{n}$$

## **Índice**

DEDICATÓRIA .....	3
1.PREFÁCIO.....	4
2. OBJECTIVOS DO GUIA PRÁTICO.....	5
3. CONCESSÃO E DEFINIÇÃO DA PESQUISA .....	6
3.1. Definições e conceitos fundamentais .....	6
4. TABULAÇÃO.....	16
5. QUESTIONÁRIO .....	19
6. RECOLHA DE DADOS.....	22
7. CONSTRUÇÃO DE BASE DE DADOS.....	22
8. TRATAMENTO DE DADOS.....	29
9. ANÁLISE DE DADOS .....	40
10. ELABORAÇÃO DE RELATÓRIO ANALÍTICO .....	46
REFERÊNCIAS BIBLIOGRÁFICAS .....	49

## DEDICATÓRIA

Dedico este manual a minha querida esposa Laurinda Cambundo e aos meus preciosos filhos, Adriel e Ivanildo Cambundo, por servirem-me como um baluarte e me terem compreendido ao cederem-me tempo para dedicação na elaboração deste importante guia prático; e ao colega João Lino e irmão Alex Chipuia por incentivarem-me a elaborar o presente material.

## **1.PREFÁCIO**

A estatística continua a ser cada vez mais uma ciência muito indispensável na actualidade, sobretudo no que diz respeito em fornecer informações previamente tratadas e analisadas para servir nas tomadas de decisões, isto em qualquer organização ou instituição, além disso, na elaboração de pesquisas sociais, demográficas, económicas, estudo de mercados, pesquisas em ciências da saúde e em outras áreas, é imprescindível o uso da estatística.

Portanto, fruto da experiência docente e em função dos cinco anos de trabalho como analista de dados no INE, pensei em elaborar este manual com vista a servir a todo leitor que tenha interesse pelos métodos estatísticos para condução de uma pesquisa científica, com ênfase nas áreas de ciências sociais e da saúde. Encorajo em particular, aqueles que têm ou tiveram dificuldades no aprendizado da estatística e na sua aplicabilidade desde os aspectos preliminares e básicos.

Recomenda-se ao leitor que quiser aprender e compreender esta fascinante disciplina, a leitura deste manual que foi escrito para auxiliar e ajudar nos mais diversos acervos de estatística básica que existem, no intuito de facilitar o aprendizado da mesma e tirar o maior proveito em como se proceder para a realização de uma pesquisa de carácter prático e científico.

Estudar os métodos estatísticos hoje, é também conhecer a própria natureza das ferramentas matemáticas disponíveis, para que o leitor possa ter uma visão ampla de suas aplicações, e de como o conhecimento da Matemática nos auxilia a entender melhor o nosso mundo, bem como a compreender o desenvolvimento de avanços tecnológicos, cujas ramificações vão desde o seu emprego na engenharia civil até na medicina. Por fim, espero que este manual atenda os seus objectivos e o conduza de forma agradável pelos fascinantes caminhos da aplicação dos métodos estatísticos.

*O autor*

***Alcides Cambundo***

*(Licenciado em ciências Matemáticas & Analista de Dados)*

## **2. OBJECTIVOS DO GUIA PRÁTICO**

O presente manual ou guia prático tem como objetivos os seguintes:

- Descrever de forma básica, resumida e objectiva o processo de realização de uma pesquisa científica, bem como as ferramentas a serem utilizadas;
- Levar ao aprendizado dos pressupostos necessários para principiar uma pesquisa, bem como os métodos utilizados que permitirão obter noções para análise estatística;
- Desenhar planos de tabulação, questionários para recolha de dados, pequenas bases de dados, noções em tratamento de Big Data, tabelas e gráficos estatísticos;
- Discutir e criticar dados recolhidos do campo para posteriormente servir de análise e objecto de estudo que se transforme em informação estatística;
- Estabelecer as balizas sobre como redigir um relatório analítico que servirá como um “out put” da pesquisa realizada onde os resultados são vistos de forma qualitativa e quantitativa.

### 3. CONCESSÃO E DEFINIÇÃO DA PESQUISA

#### 3.1. Definições e conceitos fundamentais

##### O que é a Estatística?

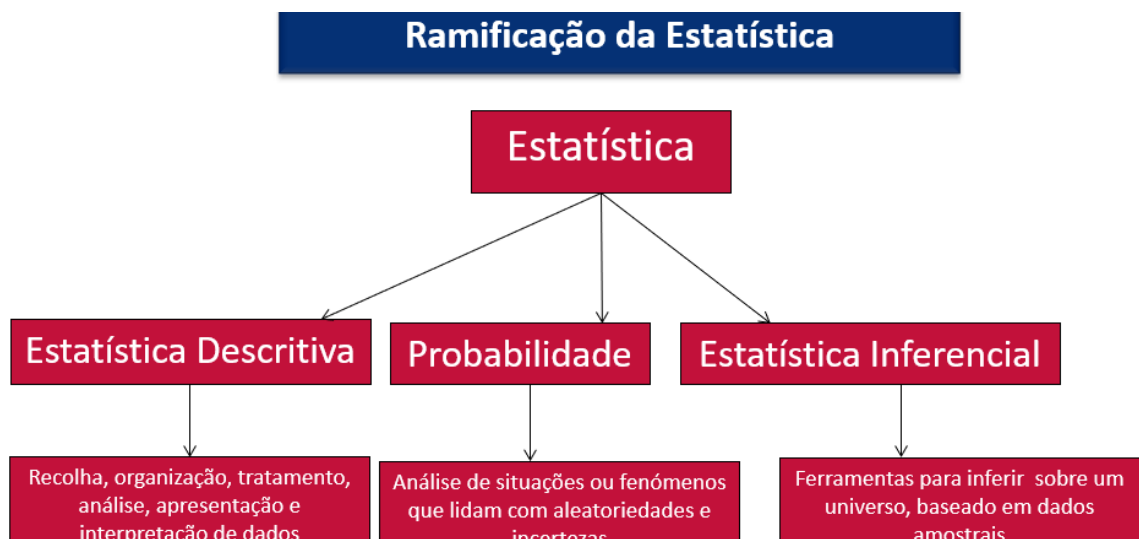
- Etimologia da palavra: do latim status => estado.
- Origem: recolha e apresentação de dados de interesse do Estado, no caso, informações sobre populações e riquezas, fins militares e tributários. Basicamente, a estatística é o conjunto de métodos especialmente apropriado ao tratamento de dados numéricos, afectados por uma multiplicidade de causas. Estes métodos fazem uso da Matemática, e especialmente do cálculo de probabilidades.

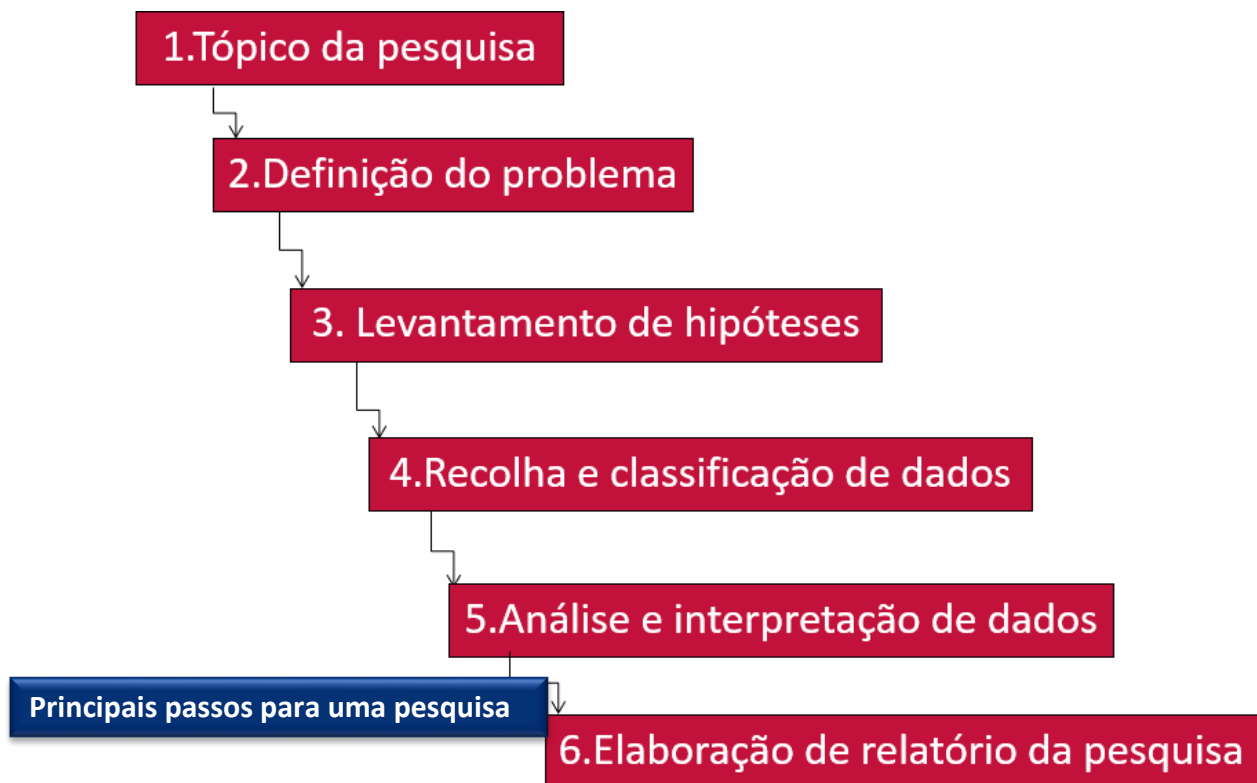
##### Tipos de operações ou pesquisas estatística

**Censo:** É uma operação estatística exaustiva que visa estudar toda população ou universo da pesquisa.

**Inquérito por amostragem:** É uma operação estatística que é feita apenas em uma parte representativa da população, denominada amostra.

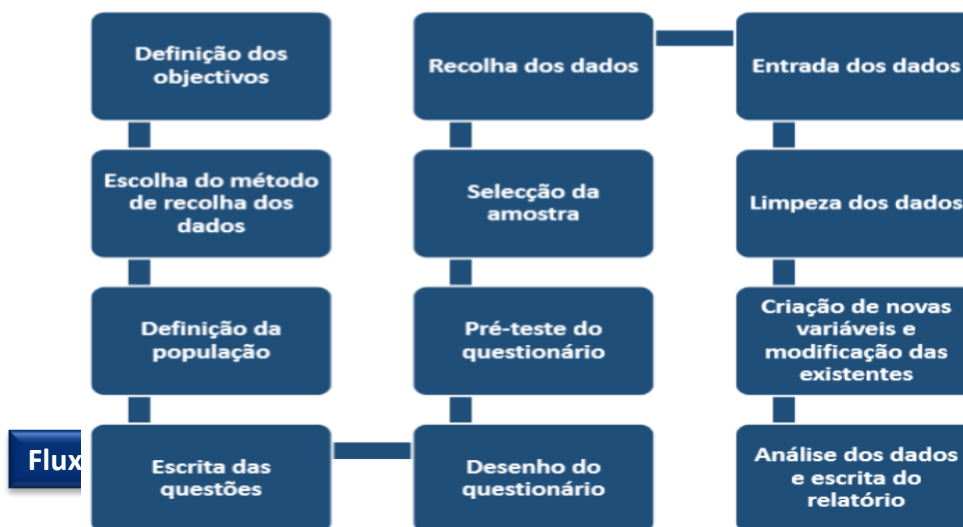
**Sondagem:** É uma operação que se refere geralmente ao acompanhamento e pesquisa feita a partir de amostras para determinar a tendência da maioria das pessoas sobre determinado assunto ou fenómeno.

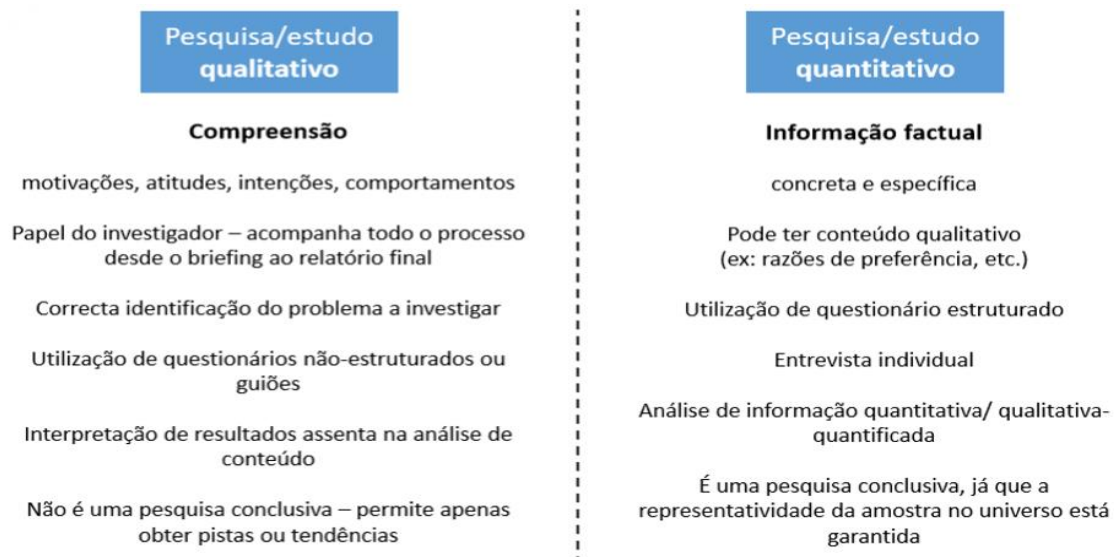
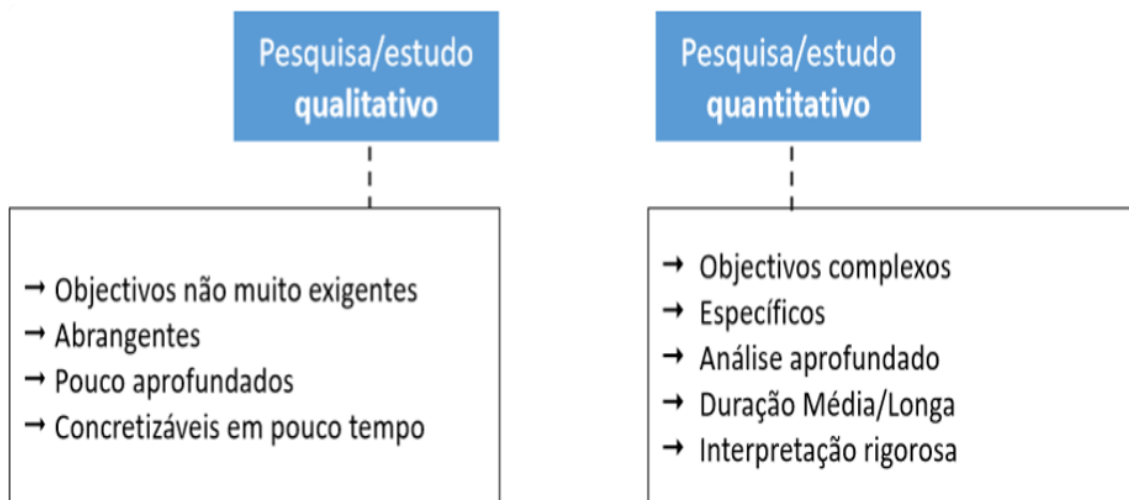




De forma geral, a pesquisa científica é normalizada pelos passos vistos anteriormente.

Note-se que de uma ou de outra forma, os processos são quase o mesmo, havendo apenas a necessidade de ligeiras adaptações em função dos objectivos que a pesquisa se propõe alcançar, porém, quando nos cingimos particularmente de uma pesquisa estatística, o fluxograma padrão é o que vamos apresentar de seguida.





## População ou universo vs amostra

**População ou universo:** é o conjunto de seres animados ou inanimados que apresentam pelo menos uma característica em comum. Sendo N o número total de elementos da população, o mesmo pode ser representado pela letra

maiúscula X, tal que  $X_N = X_1; X_2; \dots; X_N$ . A delimitação da população consiste em explicitar que pessoas ou coisas, fenómenos etc. serão pesquisados, enumerando suas características comuns, como, por exemplo, sexo, faixa etária, organização a que pertencem, comunidade onde vivem etc.



**Amostra:** É uma porção ou parte, convenientemente seleccionada do universo (população); é um subconjunto do universo. Sendo  $n$  o número de elementos da amostra, esta pode ser representada pela letra minúscula  $x$ , tal que  $x_n = x_1; x_2; \dots x_n$ . Onde  $x_n < X_N$  e  $n \leq N$ .

**Amostragem:** É o processo metódico pelo qual é feito a determinação de uma amostra para determinada pesquisa. Ela pode ser probabilística e não-probabilística.

**Amostragem não probabilística:** É uma amostragem por conveniência, não fazendo uso de uma forma probabilística por serem imediatamente disponíveis. Não pode ser objecto de certos tipos de tratamento estatístico, o que diminui a possibilidade de inferir para todo os resultados obtidos para a amostra. Por isso é que ela é pouco utilizada.

**Amostragem probabilística:** Baseia-se na escolha aleatória dos pesquisados, significando o aleatório que a selecção se faz de forma que cada elemento da população tenha a mesma probabilidade de ser escolhido ( $1/N$ ).

A amostragem probabilística permite a utilização de tratamento estatístico, que possibilita compensar erros amostrais e outros aspectos relevantes para a representatividade e significância da amostra. Divide-se em: aleatória simples, sistemática, aleatória de múltiplo estágio, por área, por conglomerados ou grupos, de vários degraus ou estágios múltiplos, de fases múltiplas (multifásica ou em várias etapas), estratificada e amostra-tipo (amostra principal, amostra a priori ou amostra padrão).

**Erro amostral ( $\varepsilon$ ):** É a máxima diferença que o pesquisador admite suportar entre a verdadeira média populacional ( $\mu$ ) que ele não conhece e a média amostral ( $\bar{x}$ ) a ser calculada da amostra, no caso em uma população infinita, ou seja,  $|\mu - \bar{x}| < \varepsilon$ .

Ou ainda, o erro amostral será a máxima diferença que o pesquisador admitir suportar entre a verdadeira proporção  $p$  que ele não conhece e  $\hat{p}$  será a proporção (frequência relativa a ser calculada da amostra), ou seja,  $|p - \hat{p}| < \varepsilon$ . Em ambos casos, ele é expresso em decimais.

**Erro não amostral:** Ocorre quando os dados amostrais são recolhidos incorrectamente,

devido a uma amostra tendenciosa, instrumento de medida ou recolha defeituoso, anotações erradas, má percepção do conceito do que se pretende saber ou responder por parte do anotador ou inquirido, etc.

### Dimensionamento da amostra

Tomemos como elementos fundamentais para determinar uma amostra simples:

$\epsilon$ –*Erro amostral*;

$N$ –*Tamanho da população*;

$n$ –*Tamanho da amostra*;

$\alpha$ –*nível de significância*;

$p$ – *Probabilidade de selecção*;

$W_p$ – *Ponderador(peso) amostral*

Bem, em geral para determinação da amostra de uma determinada pesquisa, há que se ter em conta vários factores na amostragem, além dos elementos notáveis que enumeramos acima, podemos ainda ter em conta a composição da população, além de saber se ela é finita ou infinita, os recursos necessários para implementação da pesquisa, etc... mas vamos ainda assumir que pretendemos simplesmente determinar uma amostra simples para uma pesquisa bem delimitada de forma a “doc”. Levando em conta que a probabilidade de seleccionar-se um elemento da população é calculado pela fórmula  $p = \frac{n}{N}$ ; o ponderador amostral  $W_p$ , é o inverso da probabilidade  $p$ , ou seja,  $W_p = \frac{1}{p} \Rightarrow W_p = \frac{1}{\frac{n}{N}} \Rightarrow W_p = \frac{N}{n}$ , de sorte que :  $p \times W_p = 1$  ou  $p \times W_p = 100\%$ , quando estas forem expressas em percentagem.

Geralmente por convecção, para pesquisas bem delimitadas e não muito complexas, a literatura considera valores fixos para o nível de significância " $\alpha$ " de 95%, alguns casos ainda para 99% sobretudo em pesquisas da área de saúde, porém o normal é mesmo 95%. Isto quer dizer que, em um experimento repetido cem vezes numa dada mostra, chegaremos as mesmas conclusões 95 vezes, ou seja, em 100 vezes de estudo, visaremos o mesmo

comportamento 95 vezes no elemento em questão, com uma margem de erro de 5%. O que garante uma boa fiabilidade no estudo em causa.

Inicialmente tomemos,  $n = \frac{N \times n_0}{N + n_0}$ , onde  $n_0$  é a primeira aproximação amostral, assim sendo,  $n_0 = \frac{1}{\varepsilon^2}$  por convenção. Introduzindo  $n_0$  na fórmula acima obtemos:  $n = \frac{N}{1 + \varepsilon^2 N}$  (\*).

(\*) é denominada como uma das fórmulas para determinar uma amostra simples.

Sendo que cada elemento  $x_n$  da amostra “n” representa um certo “ $X_N$ ” da população “N” com o ponderador  $W_p$  e a probabilidade de selecção p.

**Exemplo 1:** Um empresário pretende investir no ramo de geladaria no bairro do Kikolo em Luanda, sabendo que o número de máquinas de gelados para o município de Cacuaco com vista a responder a demanda populacional é de cerca de 1500 máquinas. Assumindo o nível de significância a 95% e um erro amostral de 0,05. Quantas máquinas o empresário deve adquirir para o bairro do Kikolo?

**Resolução:**

Pelo problema apresentado temos:  $N = 1500$ ;  $\varepsilon = 0,05$ ;  $n = ?$

Usando a fórmula (\*), tem-se que:  $n = \frac{N}{1 + \varepsilon^2 N} \Rightarrow n = \frac{1500}{1 + 0,05^2 \times 1500} = 315,79 \cong 316 \Rightarrow n = 316$ .

logo: Para o seu investimento, o empresário deve adquirir 316 máquinas de produção de gelado, em função do nº de habitantes residentes no bairro do Kikolo (segundo o Censo populacional 2014, estimado em cerca de 601 599 pessoas).

### Dimensionamento e composição da amostra

No processo de amostragem geralmente há que se considerar dois aspectos que são:

**Dimensionamento:** É a parte do processo da amostragem que diz respeito a determinação do tamanho da amostra, ou seja, da quantidade de elementos que a pesquisa deverá suportar.

**Composição:** É a parte do processo da amostragem que diz respeito a composição da amostra, ou seja, em função do tamanho selecciona-se os elementos que constituirão a amostra previamente dimensionada.

**Procedimento no dimensionamento da amostra**

1. Analise o questionário, ou roteiro da entrevista e escolha uma variável que julgue mais importante para o estudo. Se possível escolha mais do que uma;
2. Verifique o nível de mensuração da variável: se nominal, ordinal, intervalar ou quantitativa;
3. Considere o tamanho da população: infinita ou finita;
4. Se a variável escolhida for intervalar ou quantitativa e a população considerada **infinita**, deve-se determinar o tamanho da amostra pela fórmula:  $n = \left(\frac{Z \cdot \sigma}{\varepsilon}\right)^2$

Onde: Z- abscissa da curva normal padrão, fixado um nível de confiança.

$$z = \begin{cases} 2, & \text{se } \alpha = 95,5\%; \\ 1,96, & \text{se } \alpha = 95\%; \\ 2,57, & \text{se } \alpha = 99\% \end{cases} \quad \text{Geralmente, utiliza-se } Z=1,96 \text{ ou } 2$$

$\sigma$  –Desvio padrão da população, expresso na unidade variável. Pode-se determinar de pelo menos três maneiras:

- Especificações técnicas;
- Resgatar o valor de estudos semelhantes;
- Fazer conjecturas sobre possíveis valores.

$\varepsilon$ - Erro amostral, expresso na unidade da variável

5. Se a variável escolhida for intervalar ou quantitativa e a população **finita**, tem-se:

$$n = \frac{Z^2 * \sigma^2 * N}{\varepsilon^2(N - 1) + Z^2 * \sigma^2}$$

6. Se a variável escolhida for nominal ou ordinal, e a população considerada **infinita**, pode-se determinar o tamanho da amostra pela fórmula:  $n = \frac{Z^2 * \hat{p} * \hat{q}}{\varepsilon^2}$

Onde:  $\hat{p}$ -Estimativa da verdadeira proporção de um dos níveis da variável escolhida. Por exemplo, se a variável escolhida for parte da empresa,  $\hat{p}$  poderá ser a estimativa da verdadeira proporção de grandes empresas do sector que está sendo estudado. Será expresso em decimais. Assim, se  $\hat{p}=30\%$ , teremos  $\hat{p}=0,30$  e  $\hat{q}=1-\hat{p}=1-0,30 \Rightarrow \hat{q} = 0,70$ .

7. Se a variável escolhida for nominal ou ordinal, e a população considerada **finita**, pode-se

$$\text{determinar o tamanho da amostra pela fórmula: } n = \frac{Z^2 * \hat{p} * \hat{q} * N}{\varepsilon^2(N-1) + Z^2 * \hat{p} * \hat{q}}$$

**Nota:** Se o pesquisador/investigador escolhe mais de uma variável, deve optar pelo maior n obtido. Sabendo que quanto maior for o tamanho da amostra menor será a margem de erro.

### Exemplo 2

Suponha que a variável escolhida num estudo seja o peso de certa peça e que a população é infinita. Pelas especificações do produto o desvio padrão (dispersão em torno da média) é de 10 kg, admitindo-se um nível de confiança de 95,5% e um erro amostral de 1,5 kg, determine o tamanho da amostra do referido estudo.

**Resolução:** Mediante as informações dadas, uma vez que a população é infinita, vamos usar a fórmula:  $n = \left(\frac{Z \cdot \sigma}{\varepsilon}\right)^2$ , Já que peso é uma V. contínua escalar ou intervalar. Logo,  $n = \left(\frac{2 \cdot 10}{1,5}\right)^2 \Rightarrow n = 177,78 \Rightarrow n \cong 178$ .

### Exemplo 3

Sob as mesmas condições do exemplo anterior, determinar o tamanho da amostra do referido estudo considerando que a população seja finita de 600 peças.

**Resolução:** Como a população é finita, usaremos a fórmula:  $n = \frac{Z^2 \cdot \sigma^2 \cdot N}{\varepsilon^2 (N-1) + Z^2 \cdot \sigma^2}$ , Daqui,  $n = \frac{2^2 \cdot 10^2 \cdot 600}{1,5^2 (600-1) + 2^2 \cdot 10^2} \Rightarrow n = 137,319 \Rightarrow n \cong 138$ .

### Exemplo 4

Suponha que a variável escolhida num estudo seja a proporção de eleitores favoráveis ao principal candidato a comissão de moradores do bairro Ngangula (Namibe) e que o pesquisador tenha matéria para assumir que essa percentagem seja de 30%. Admitindo uma população infinita e que se deseja um nível de confiança de 99% e um erro amostral de 2% (ou seja: que a diferença entre a verdadeira proporção de eleitores do candidato x e a estimativa a ser calculada na amostra seja no máximo de 2%). Determine n nestas condições.

**Resolução:** Como a população é infinita, usaremos a fórmula:  $n = \frac{Z^2 \cdot \hat{p} \cdot \hat{q}}{\varepsilon^2}$ , Assim:  $Z = 2,57$ ,  $\hat{p} = 30\% = 0,30$ ,  $\hat{q} = 1 - 0,30 \Rightarrow \hat{q} = 0,70$ ,  $\varepsilon = 2\% = 0,02$ .

Logo,  $n = \frac{2,57^2 \cdot 0,30 \cdot 0,70}{0,02^2} = 3467,57 \Rightarrow n \cong 3468$ .

### Exemplo 5

Admitindo os mesmos dados na questão anterior, e que a população de eleitores seja finita de 20 000 eleitores. Determine o tamanho da amostra nestas condições.

**Resolução:** Como a população é finita, usaremos a fórmula  $n = \frac{Z^2 \cdot \hat{p} \cdot \hat{q} \cdot N}{\varepsilon^2(N-1) + Z^2 \cdot \hat{p} \cdot \hat{q}}$ , Assim:

$$n = \frac{2,57^2 \cdot 0,3 \cdot 0,7 \cdot 20\,000}{0,02^2(20\,000-1) + 20\,000^2 \cdot 0,3 \cdot 0,7} = 2955,329 \Rightarrow n \cong 2\,956.$$

**Nota:** Em geral, quando não tiver condições ou conjecturas de prever o valor de  $\hat{p}$ , admite-se que  $\hat{p} = 0,50$ , desta forma obtém-se o maior tamanho possível da amostra, admitindo-se constantes os demais elementos.

### Composição da amostra

Basicamente, existem dois métodos para composição da amostra: probabilístico e não probabilístico ou intencional.

Como já definimos anteriormente o método probabilístico, exige que cada elemento da população possua determinada probabilidade de ser selecionado. Normalmente possuem a mesma probabilidade. Assim, se N for o tamanho da população, probabilidade de cada elemento será  $\frac{1}{N}$ . Trata-se do método que garante cientificamente a aplicação das técnicas estatísticas de inferências. Somente com base em amostragem probabilísticas é que se podem realizar inferências ou induções sobre a população a partir do conhecimento da amostra.

**Amostragem aleatória simples:** É o processo mais elementar e frequentemente utilizado. Atribui-se a cada elemento da população um nº distinto. Se a população for numerada utilizam-se “rótulos”. Efectuam-se sucessivos sorteios até completar-se o tamanho da amostra n. Para realizar os sorteios, utilizam-se “tábuas de nºs aleatórios” que consistem em tabelas que apresentam sequências dos dígitos de 0-9 distribuídos aleatoriamente. Se por exemplo, a população tem 500 elementos (N=500), pode-se numerá-las de 000-499. Escolhendo uma posição de qualquer linha da tabela de nºs aleatórios, faz-se o sorteio, ou seja, retiram-se conjuntos de três dígitos para se escolherem os elementos que irão compor a amostra. Assim imagine que a sequência de nºs aleatórios seja, 244, 345, 459, 129, 423,... logo, os elementos de números 244-345-459-129-423... serão os componentes/elementos da amostra. Se o número sorteado exceder o maior nº dos elementos rotulados, abandona-se o nº sorteado, prosseguindo-se o processo.

**Amostragem sistemática:** Trata-se de uma variação da amostragem aleatória simples, conveniente quando a população está ordenada segundo algum critério, como fichas em uma pasta de arquivos, listas telefônicas...

Calcula-se o intervalo da amostragem “k” ( $k=N/n$ ) devendo escolher-se apenas a parte inteira da divisão. Utilizando-se a tábua dos nºs aleatórios, sorteia-se um nº x entre 1 e a, formando-se a amostra dos elementos correspondentes aos nºs x; x+k; x+2k; x+3k;...

Por exemplo, suponhamos que para uma pesquisa cuja população seja finita igual a 66 324 e o tamanho da amostra ficou em 3 635. Logo:  $k = \frac{66\,324}{3\,635} = 18,25 \Rightarrow k = 18$ .

Imagine que recorremos a uma tabela aleatória simples (*ver anexo no final*) cujo dia da amostragem (coluna) foi considerado 29, assim sendo  $2+9=11$  (coluna), devemos fazer o cruzamento na tabela entre coluna 11 vs linha 18, neste caso o primeiro valor sorteado é 6. Portanto, seguindo a fórmula, os elementos da população numerados por 6; 24; 42;...irão compor a amostra.

## 4. TABULAÇÃO

### Definição

**Tabulação:** É a projecção da disposição dos dados em tabelas/quadros, neste caso em linhas e colunas possibilitando maior facilidade na verificação das inter-relações entre eles. É uma parte do processo técnico de análise estatística, que permitirá sintetizar os dados de observação, conseguidos pelas diferentes categorias e representá-los graficamente.

Geralmente, a tabulação é feita antes da elaboração do questionário e recolha de dados, como antevisão dos indicadores e objectivos que se pretende alcançar na pesquisa.

Para elaborar uma tabulação, é necessário ter em conta os objectivos da pesquisa bem como as variáveis fundamentais que constituem as variáveis chaves e derivados.

Eis algumas variáveis a ter-se em conta: Localização geográfica da pesquisa, período, género, idade, nível de escolaridade, situação profissional, renda, despesas, etc... como dissemos antes, as variáveis a constar na tabulação dependem ou dependerão da definição e dos objectivos que a pesquisa se propõe alcançar.

### Procedimentos a observar durante a elaboração da tabulação

Para elaboração da tabulação, há necessidade de se obedecer determinados critérios, desde as variáveis para associação em cruzamento bem como para elaboração dos títulos das tabelas. Assim sendo destacamos o seguinte:

- **Itens principais (variáveis de classificação principal)**

Geralmente as variáveis de classificação principal ou itens principais, são aquelas que durante a elaboração da tabela cruzada, são predispostas na linha, normalmente são unidades administrativas, género, idade, nível de escolaridade, ou outras variáveis. É importante salientar que as variáveis de classificação principal podem variar de acordo a pesquisa e a natureza da disposição dos dados, mas vale lembrar que no cruzamento são comumente usados na linha.



- **Itens essenciais (variáveis de classificação essencial)**

Geralmente as variáveis de classificação essencial ou itens essenciais, são aquelas que durante a elaboração da tabela cruzada, são predispostas na coluna, normalmente são variáveis determinantes na informação que pretendemos previsualizar ou observar. Porém, elas podem em algum momento serem alternadas de classificação tabular, tal como para as variáveis principais de acordo a pesquisa e a natureza da disposição dos dados, usualmente elas aparecem na coluna (vertical).

- **Elaboração dos títulos**

Para elaboração do título de uma tabela, há que se ter em conta as variáveis principais e essenciais, bem como a unidade numérica (**Número, percentagem ou proporção**) a ser expressa na informação que se pretende apresentar na tabela. Deste modo, obedecendo os princípios, primeiro devemos expressar a unidade numérica, seguido das variáveis de classificação principal associado ao conectivo de cruzamento “**por**” e depois o conectivo “**segundo**” posteriormente as variáveis ou itens de cruzamento essencial.

Ao elaborarmos as tabelas/quadros, há que se ter em conta que cada uma delas deve possuir sua numeração de acordo a quantidade de tabelas, para o efeito, normalmente a tabulação é feita em Excel dado a sua agilidade e especificidade na elaboração de tabelas. A seguir vejamos alguns exemplos de tabulação.

Bairros	Rendimento médio
Total	
Bairro 1	
Bairro 2	
Bairro 3	
Bairro 4	
Bairro 5	

<b>Ítem principal</b>	<b>Ítem essencial</b>
<b>Tabela 2- Número de pessoas entrevistadas por Nível de escolaridade, segundo o valor médio da despesa na cesta básica durante o mês</b>	
Nível de escolaridade	<b>Despesa média mensal</b>
<b>Total</b>	
Ensino primário	
I Ciclo	
II Ciclo	
Ensino superior	
Sem nenhum nível	

### Observação:

Por vezes o número de variáveis principais numa tabela é mais que uma e que, se tivermos que descrevê-las no título, o mesmo fica muito extenso, para estes casos podemos simplesmente fazer menção dessas variáveis como um todo. Por exemplo, suponhamos que para uma tabela as variáveis de classificação principal são, **Província, estado civil, faixas etárias e nível de escolaridade**...ao descrever no título pormenorizadamente seria: Número de pessoas entrevistadas por província, estado civil, faixa etária e nível de escolaridade, segundo...

Porém, desta forma o título fica muito extenso, podemos simplificar para seguinte maneira: Número de pessoas entrevistadas por itens seleccionados, segundo...

## 5. QUESTIONÁRIO

### Conceitos e definições

**Questionário:** É um instrumento de recolha de dados, constituído por uma série ordenada de perguntas, que devem ser respondidas por escrito ou verbal, com ou sem a presença do entrevistador. Normalmente ele deve ser feito depois do plano de tabulação.

Quanto ao formato, o questionário pode ser em **papel** (físico) ou **digital** indexada em um aparelho electrónico que geralmente tem sido: computadores, tablets, telemóveis, etc.

Quanto a classificação da natureza das perguntas, o questionário pode ser classificado em:

**Fechadas ou dicotómicas** - Também denominadas limitadas ou de alternativas fixas, são aquelas em que geralmente o informante escolhe sua resposta entre duas opções: sim e não ou escolhe sua resposta em um conjunto de opções de respostas já pré-definidas.

#### Exemplo:

1) Os Cristãos devem ou não formar um partido político?

1. Sim ( )

2. Não ( )

2) Você concorda ou não ao celibato dos padres?

1. Sim, concordo ( )

2. Não concordo ( )

3) Qual a sua opinião relativamente a implementação de políticas habitacionais no país?

1. Péssimas políticas ( )

2. Políticas não inclusivas ( )

3. Boas políticas, porém exclusivas ( )

Estes tipos de perguntas, embora restrinjam a liberdade das respostas, facilitam o trabalho do pesquisador e também o tratamento e análise de dados: as respostas são mais objectivas. É o tipo de questionário usualmente empregue e recomendado nas operações estatísticas.

**Abertas** – Também chamadas livres ou não limitadas, são as que permitem ao informante responder livremente, usando linguagem própria, e emitir opiniões.

**Exemplo:**

1. Qual é sua opinião sobre a legalização do aborto?

---

---

---

2. Em sua opinião, quais são as principais causas da delinquência em Luanda?

---

---

---

Estes tipos de perguntas possibilitam investigações mais profundas, porém dificulta a resposta ao próprio informante, o tratamento dos dados e a análise é difícil, complexa, cansativa e demorada. Portanto, não é um tipo de questionário usual para operações estatísticas. Apesar de que nos questionários do tipo fechados, pode conter alguma pergunta com opção de resposta em aberto, geralmente são intituladas como “outro”.

A elaboração de um questionário requer a observância de normas precisas, a fim de aumentar sua eficácia e validade. Em sua organização, devem-se levar em conta os tipos, a ordem, os grupos de perguntas bem como a formulação das mesmas e também tudo aquilo que se sabe sobre percepção" (Augras, 1974, p.143)".

O questionário deve ser limitado em extensão e em finalidade. Se for muito longo, causa fadiga; se curto demais, corre o risco de não oferecer suficientes dados. Deve conter de 20 a 30 perguntas e demorar cerca de 30 minutos para ser respondido. É claro que este número não é fixo. Outro aspecto importante do questionário é a indicação da entidade ou organização responsável da pesquisa. Por exemplo: AS-K LIMITADA.

Deve estar acompanhado por instruções definidas e notas explicativas, para que o informante tome ciência do que se deseja dele.

O aspecto material e a estética também devem ser observados: tamanho, facilidade de manipulação, espaço suficiente para as respostas, a disposição dos itens, de forma a facilitar o tratamento dos dados.

## Exemplo de um questionário de perguntas fechadas

### SECÇÃO 1: CARACTERIZAÇÃO DOS INDIVÍDUOS/RESPONDENTES

PARA TODAS AS PESSOAS DOS 15-35 ANOS DE IDADE																		
SEXO	IDADE	BAIRRO DE RESIDÊNCIA		ESTADO CIVIL	ESCOLARIDADE							SITUAÇÃO DE EMPREGO						
P1_1	P1_2	P1_3		P1_4	P1_5		P1_6		P1_7		P1_8		P1_9	P1_10				
QUAL É O SEXO DO(A) NOME?	QUAL É A IDADE DO(A) NOME?	EM QUE BAIRRO O(A) NOME VIVE?	QUAL É O ESTADO CIVIL DO(A) NOME?	(O) NOME ALGUMA VEZ FREQUENTOU A ESCOLA/UNIVERSIDADE?		QUAL É O NÍVEL QUE (O) NOME FREQUENTA/FREQUENTOU?		(O) NOME CONCLUÍU ESTE NÍVEL?		POR QUE (O) NOME NUNCA FREQUENTOU A ESCOLA?		POR QUE (O) NOME NÃO CONCLUÍU ESTE NÍVEL?		(O) NOME TRABALHA POR CONTA PRÓPRIA OU DE OUTREM?				
Masculino	1	Bairro 1	1	Solteiro(a)/nunca viveu em união de facto	1	Sim, ainda frequenta	1	Primário	1	Sim	1=>1_10	Não tenho documentos	1	Por falta de recursos	1	Sim, trabalha por conta própria	1	
Feminino	2	Bairro 2	2	Casado (a)	2	Sim, mas já não frequenta	2	1º Ciclo do Ensino Secundário	2	Não	2=>1_9	Por falta de dinheiro	2	Reprovei/desisti	2	Sim, trabalha por conta de outrem	2	
	Registe a idade em anos completos.	Bairro 3	3	União de facto	3	Não, nunca frequentou	3=>P1_8	1º Ciclo do Ensino Secundário	3			Por falta de pais/direção	3	Por doença	3	Não trabalha	3	
		Bairro 4	4	Divorciado (a)	4			Ensino superior	4			Não tem escola na zona onde vivo	4	Desleixo/desinteresse	4			
		Bairro 5	5	Separado (a)	5				Desleixo/desinteresse		5	Ainda estou a estudar	5					
		Bairro 6	6	Viúvo (a)	6				Não gosto de estudar		6	Outra razão	6					
		ANOS												Outra razão	7			
		CÓDIGO		CÓDIGO		CÓDIGO		CÓDIGO			CÓDIGO		CÓDIGO		CÓDIGO		CÓDIGO	

## 6. RECOLHA DE DADOS

### Recolha de dados

É uma etapa da pesquisa em que se inicia a aplicação dos instrumentos elaborados e das técnicas seleccionadas, a fim de se efectuar a recolha dos dados previstos.

Exige da pessoa que a faz, paciência, perseverança e esforço pessoal, além do cuidadoso registo dos dados e de um bom preparo anterior. São vários os procedimentos para a realização da recolha de dados, que variam de acordo com as circunstâncias ou com o tipo de pesquisa.

Em linhas gerais, destacamos os tipos de recolha mais usuais: recolha documental, observação, entrevista/questionário, formulário, testes, análise de conteúdo e histórico de vida.

Outro aspecto importante é o perfeito enquadramento das tarefas organizacionais e administrativas com as científicas, obedecendo aos prazos estipulados, aos orçamentos previstos, ao preparo do pessoal. Quanto mais planeamento for feito previamente, menos desperdício de tempo haverá no trabalho de recolha.

O rigoroso controlo na aplicação dos instrumentos de pesquisa é factor fundamental para evitar erros e defeitos resultantes de entrevistadores inexperientes ou de informantes tendenciosos, que podem afectar negativamente os dados e a pesquisa em geral.

Com a evolução tecnológica, a recolha de dados pode ser feita mediante um aparelho electrónico onde é indexado o questionário, isso ajuda na redução de erros durante a recolha.

## 7. CONSTRUÇÃO DE BASE DE DADOS

### Base de dados resultante da recolha

**Base de dados:** é um conjunto de arquivos alfanuméricos, devidamente codificados e sistematizados que geralmente resultam da recolha de dados.

É um arquivo onde é armazenado todos os dados da pesquisa, a partir do qual é produzida a informação estatística mediante análise de dados.

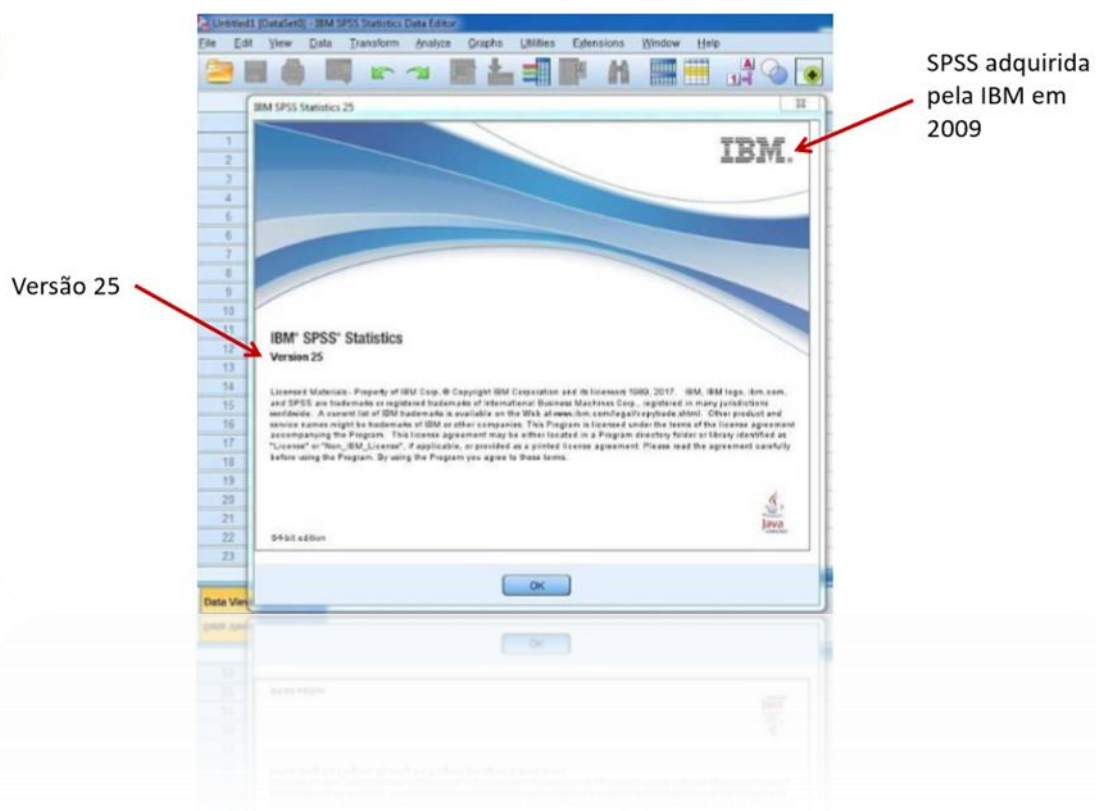
A elaboração de base de dados requer uma sensibilidade a partir da inserção de dados, este processo é feito normalmente de forma computacional, mediante aplicativos apropriados. Existem vários aplicativos/programas para elaboração de bases de dados, cujos mesmos servem para tratamento de dados, como já fizemos menção anteriormente, como é o caso de

Access, Excel, cspro, spss, etc... uns mais acessíveis e outros mais complexos, de acordo a natureza da pesquisa e sua abrangência. Para o nosso caso, vamos abordar um pouco sobre o SPSS STATISTICS quanto a sua utilidade na construção de bases de base de dados.

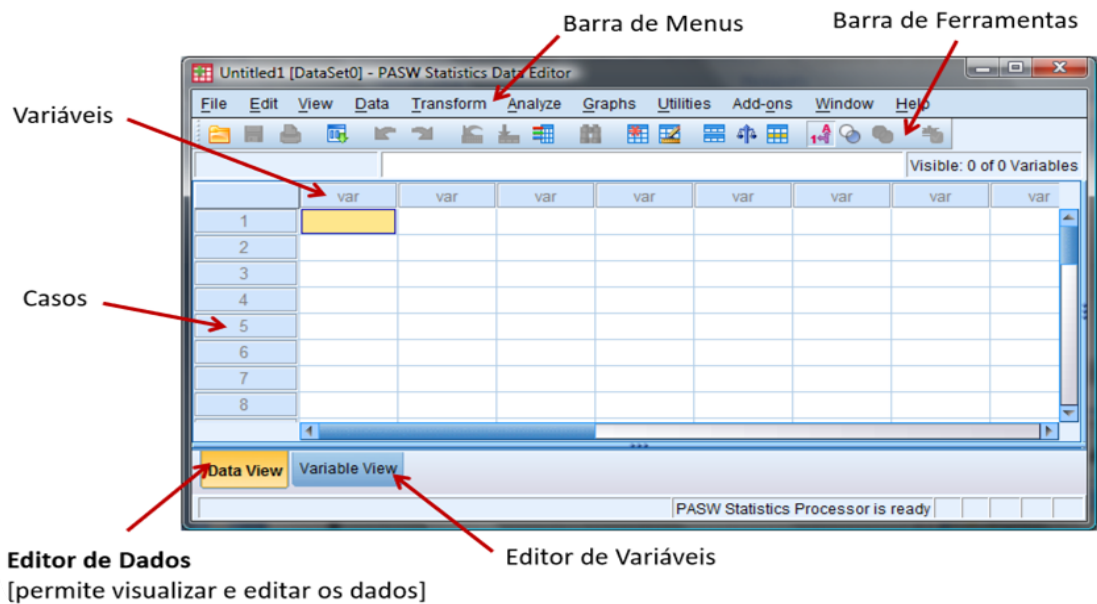
O SPSS é um software/aplicativo (programa de computador) do tipo científico. Originalmente o nome é acrónimo de Statistical Package for the Social Sciences - pacote estatístico para as ciências sociais, é um produto da empresa de tecnologia “IBM”. Este software tem sido muito útil para aplicação analítica, mineração de dados e textos que no final traduz-se em informação para apoio a tomada de decisão, sobretudo em pesquisas de mercado, “business intelligence”, pesquisas de saúde pública, estudos de viabilidade económica, estudo analíticos de trabalhos de fim de curso, pesquisas de opiniões etc.

### Visão geral do IBM SPSS STATISTICS

#### Editor de Dados



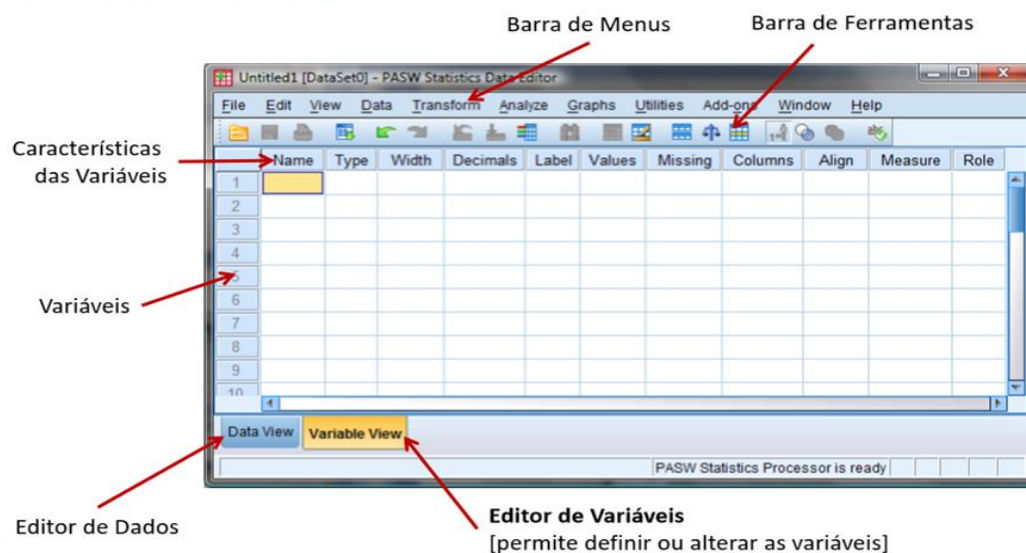
## Editor de Dados: Data View



[permite visualizar e editar os dados]  
Editor de Dados

Editor de Variáveis

## Editor de Dados: Variable View

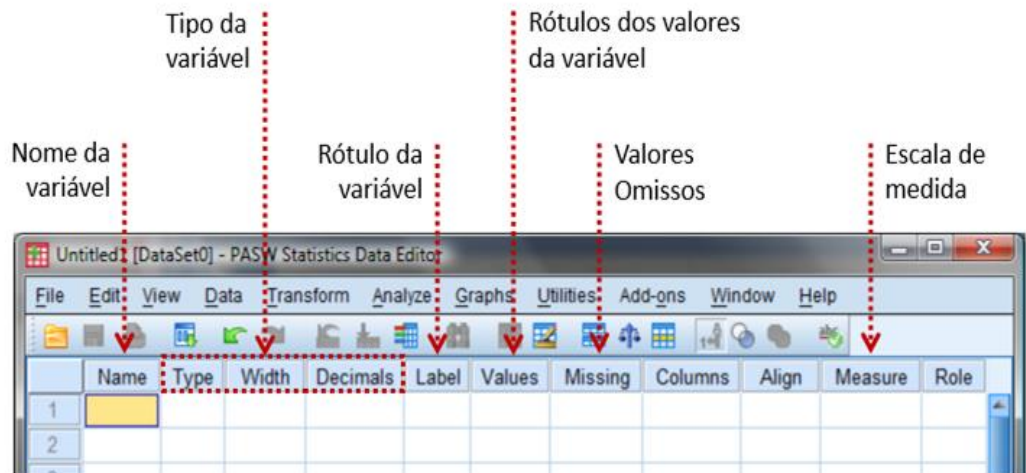


Editor de Dados

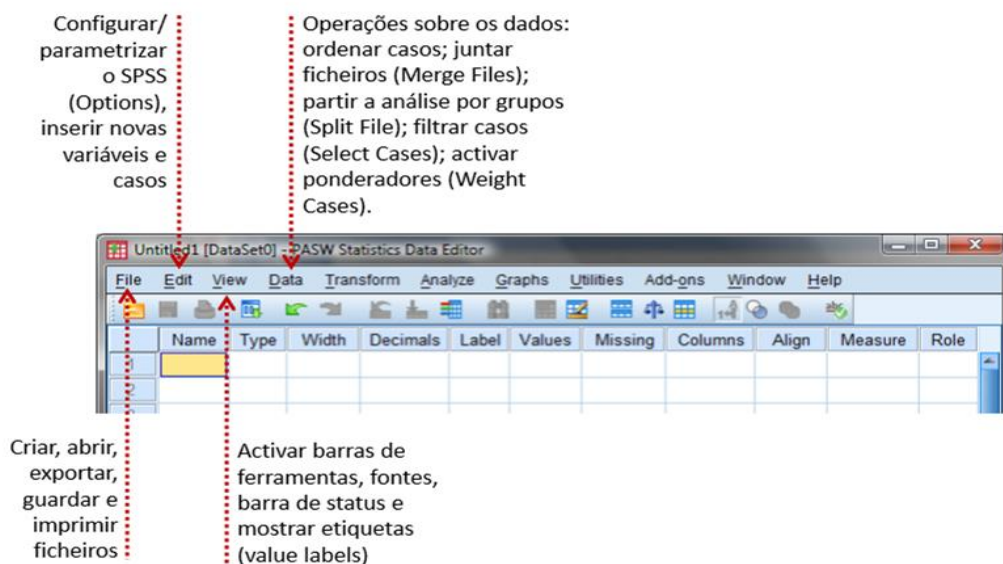
[permite definir ou alterar as variáveis]  
Editor de Variáveis



## Editor de Dados: Variable View



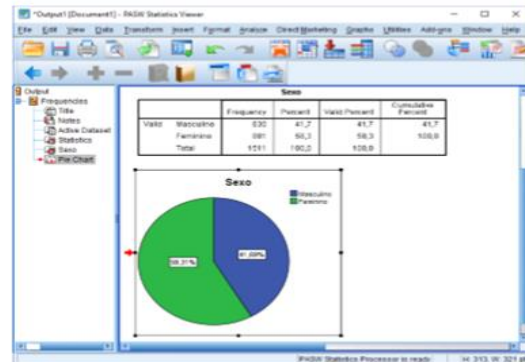
## Editor de Dados: Variable View



## Apresentação de Resultados: Output Viewer

Na janela **Output Viewer**, são apresentados todos os resultados estatísticos, tabelas e gráficos. O **Output** indexa os resultados do output num único menu fixo, disponível na parte esquerda do écran.

No lado direito do écran surgem os resultados da análise.



Existem ainda outras janelas, que nos permitem modificar tabelas, gráficos, etc., nomeadamente: **Pivot table editor** (edição de tabelas); **Chart editor** (edição de gráficos); **Syntax editor** (comandos SPSS).

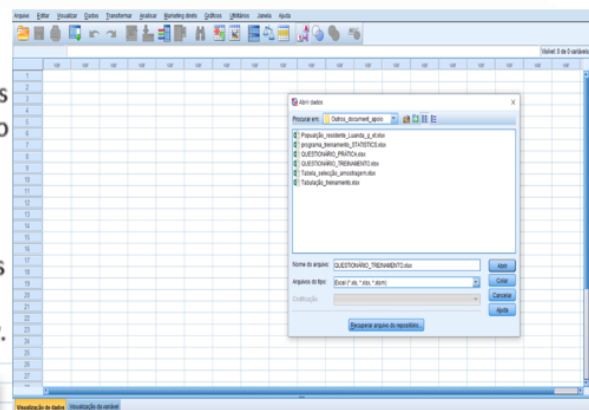
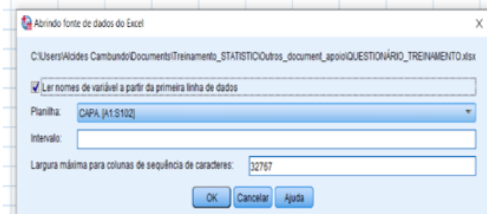
Os dados são gravados num ficheiro **SPSS Data Editor(\*.sav)** e os resultados num ficheiro **SPSS Viewer(\*.spo)**

## Importação de Dados

O **SPSS** permite a utilização de ficheiros de vários tipos, tais como Excel (\*.xls), dBase (\*.dbf), Texto (\*.txt), entre outros:

**File > Open > Data**

o **SPSS** reconhece o nome das variáveis e os seus valores, sendo necessário definir as restantes características dessas variáveis no **Variable View**.



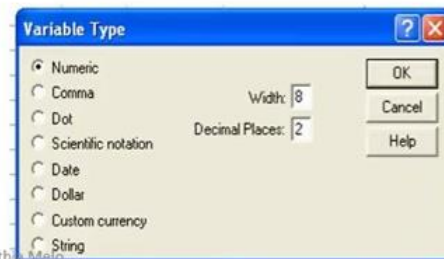
Portanto, usando o SPSS a base de dados é criada mediante a um questionário de recolha (Inquérito, Censo...), neste caso deve-se introduzir primeiramente o questionário da pesquisa no dicionário do SPSS "MENU VARIABLE VIER OU VISUALIZAÇÃO DE VARIÁVEIS" seguindo as instruções seguintes:

- No momento da introdução do questionário no SPSS, devemos ir ao menu visualização da variável (variable view) fazendo assim o preenchimento das principais colunas: **Nome**, **Rótulo**, **Valores** (caso haja) e **Medida** sendo que as outras colunas são pré-definidas;

- Para inserção dos dados no SPSS, resultante da recolha temos que ir ao menu Visualização de dados (Data view) e fazer a inserção dos dados de acordo a codificação prévia do questionário.

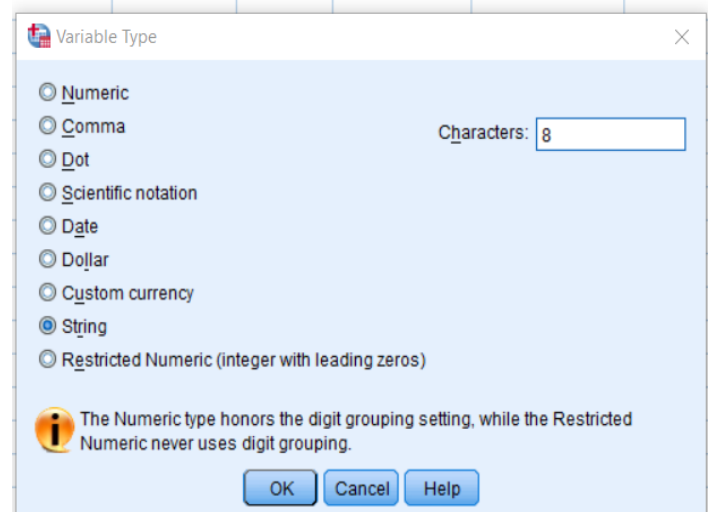
Na criação da base de dados, há que se ter em conta as 11 colunas na janela de visualização da variável (variable view):

- 1 • **Name:** ou nome em Português, corresponde a um código que representa o nome da variável que será trabalhada.
- 2 • **Type:** ou tipo, representa o tipo de formatação da variável. O SPSS permite que as variáveis sejam codificadas de diferentes formas.



## Tipos de variáveis no SPSS

Tipo	Descrição
<i>Numeric</i>	Configura a variável como um número, permitindo ajustar a largura da variável em número de caracteres ( <i>Width</i> , no caso igual a 8) e o número de casas decimais ( <i>Decimal Places</i> , no caso igual a 2).
<i>Comma</i>	Ajusta o separador de casas decimais de vírgula para ponto. Também possibilita configurar a largura e as casas decimais.
<i>Dot</i>	Ajusta o separador de casas decimais de ponto para vírgula. Também possibilita configurar a largura e as casas decimais.
<i>Scientific notation</i>	Apresenta o número em notação científica, multiplicado por um expoente de 10. Também possibilita configurar a largura e as casas decimais.
<i>Date</i>	Ajusta a configuração de datas no SPSS. Diferentes alternativas encontram-se disponíveis.
<i>Dollar</i>	Ajusta a apresentação de moedas. Também possibilita configurar a largura e as casas decimais.
<i>Custom currency</i>	Também ajusta a configuração da apresentação de moedas, possibilitando diferentes configurações customizáveis.
<i>String</i>	Configura a variável como um <i>string</i> , uma sequência não necessariamente numérica de caracteres, podendo apresentar números, textos ou símbolos. Permite ajustar o número de caracteres da variável.



- 7 **Missing** ou ausente, configura o tratamento a ser dado pelo SPSS na existência de valores ausentes na base de dados (missing values). **IDEAL:9999**
- 8 **Columns** ou largura da coluna.
- 9 **Align** ou alinhamento. (LEFT, RIGHT, CENTER)
- 10 **Measure** ou mensuração.



- 11 **Função** comportamento das informações na base de dados (pré-determinada pelo software).

Relativamente a coluna 10 Measure (Medir), o SPSS classifica as variáveis estatísticas Qualitativas e Quantitativas em três formas, nomeadamente:

1. Variáveis qualitativas em estatística:
  - ✓ Nominal;
  - ✓ Ordinal

No SPSS:

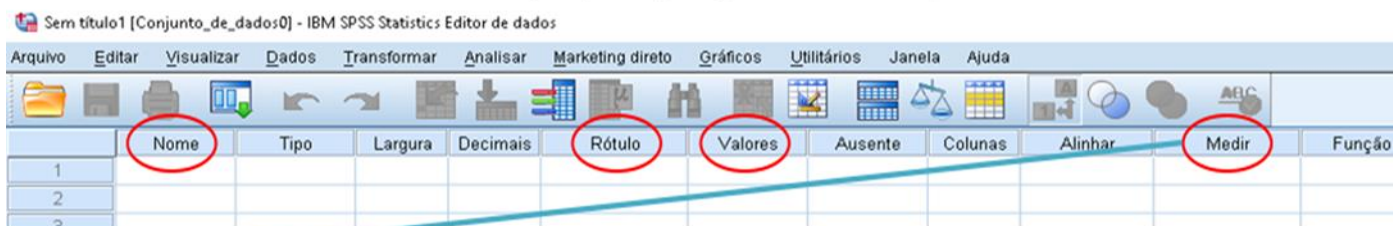
- ✓ Nominal;
- ✓ Ordinal

- 3 **Width:** ou largura, representa o número de caracteres especificado para a variável.
  - 4 **Decimals:** ou decimais, corresponde ao número de casas decimais definido para a variável.
  - 5 **Label:** ou rótulo, corresponde ao rótulo atribuído para a variável.
  - 6 **Values:** ou valores, corresponde aos rótulos atribuídos para os dados contidos na variável.
2. Variáveis quantitativas em estatística:
    - ✓ Discreta;
    - ✓ Contínua



No SPSS as variáveis Quantitativas Discretas e Contínuas são todas incorporadas no nível de medição “escala”

Para a criar a base de dados em SPSS, as principais janelas a serem preenchidas são:



**Medir:** escala de medição das variáveis. Pode ser sob o formato Escalar, Ordinal ou Nominal.

	P1.SEXO	P2.IDADE	P3.NÍVELESCOLARIDADE	P4.ESTADOCIVIL	P5.RESP_FAMIL_DIFICULT						
1	Homens	34.0	2º Ciclo ensino secundário	Casado(a)	Sim						
2	Mulheres	23.0	Ensino superior	Divorciado(a)	Não						
3	Homens	54.0	2º Ciclo ensino secundário	solteiro(a)	Sim						
4	Mulheres	23.0	Ensino superior	separado(a)	Não						
5	Homens	34.0	2º Ciclo ensino secundário	união de facto	Sim						
6	Mulheres	26.0	Ensino superior	viúvo(o)	Não						
7	Homens	32.0	2º Ciclo ensino secundário	separado(a)	Sim						
8	Mulheres	45.0	Ensino superior	união de facto	Não						
9	Homens	32.0	2º Ciclo ensino secundário	separado(a)	Sim						
10	Mulheres	37.0	Ensino superior	Casado(a)	Não						
11											
12											
13											
14											
15											
16											
17											
18											
19											
20											
21											
22											
23											
24											
25											
26											
27											

## 8. TRATAMENTO DE DADOS

É também uma das etapas mais importantes em uma pesquisa, que consiste em dar uma atenção minuciosa aos dados recolhidos, através da (codificação) classificação dos dados, agrupando-os sob determinadas categorias; atribuição de um código, número ou letra, tendo cada um deles um significado. Codificar quer dizer transformar o que é qualitativo em quantitativo, para facilitar não só a análise dos dados, mas também sua comunicação.

Geralmente o processo de tratamento de dados é feito recorrendo a processos automáticos quando a pesquisa é delicada e ampla, além disso é mais confiável e eficiente. Excepto o

tratamento de dados manual, existem vários programas ou aplicativos que têm sido usados para o tratamento de dados, desde Excel, cspro, spss, etc.

**Obs:** O pesquisador deve ter o conhecimento prático de pelo menos um dos aplicativos para o tratamento de dados.

#### ▪ Identificação de outliers

Esta secção visa abordar de forma sucinta e objectiva a forma de identificação e tratamento de valores atípicos observados numa dada variável de uma base de dados. Estes valores denominam-se “outliers”.

A abordagem será realizada na óptica do aplicativo estatístico SPSS da IBM.

Portanto, “**Outliers**” são valores que apresentam uma diferença anormal em relação ao padrão normal verificado em outras observações de uma variável em causa.

Para identificar Outliers no SPSS deve-se criar variáveis que definem os valores limites das observações. Os valores que forem superiores ao limite, requererão uma atenção pormenorizada por serem considerados Outliers, neste caso estamos a falar de limite superior, isto para identificar valores acima, para tal segue a expressão analítica abaixo:

Limite superior =  $Q3 + 2,2 (Q3 - Q1)$

Porém, se tivermos interessados em saber os valores outliers abaixo, a fórmula é: Limite inferior =  $Q3 - 2,2 (Q3 - Q1)$ .

Onde:

Q1- refere-se ao primeiro quartil;

Q3-refere-se ao terceiro quartil;

$(Q3 - Q1)$ - É o intervalo interquartil.

**Nota:** A literatura clássica geralmente utiliza a fórmula seguinte:

Limite =  $Q3 \pm 1,5 (Q3 - Q1)$  ou Limite superior =  $1,5 (Q3 - Q1)$ ,

Mas a literatura moderna viu a necessidade de se reformular o cálculo dos outliers porque a expressão  $Q3 \pm 1,5 (Q3 - Q1)$  faz com que o limite seja muito reduzido devido o factor de multiplicação 1,5, já que Q3 é sempre maior que a diferença interquartil  $(Q3 - Q1)$ , a tendência é

que os valores que devem fazer parte do grupo de valores outliers sejam incluídos também nas observações normais, é daí que surge o factor de correcção  $2,2(Q3-Q1)$ .

Cria-se a variável Outlier, para cada variável a ser estudado, todos os valores superiores ao valor limite superior definido para a variável será considerado como Outliers acima.

Suponhamos como exemplo que queiramos estudar os Outliers da variável S5\_02 (Área cultivada) da base de dados da formação do “RAPP” no âmbito da listagem.

Para tal, primeiro calculemos o valor dos quartis Q3 e a diferença interquartil Q3-Q1.

Portanto, vamos calcular no SPSS, primeiro achemos o valor do terceiro quartil (Q3) seguindo o procedimento conforme a figura abaixo:

The screenshots illustrate the steps in SPSS to calculate the 75th percentile (Q3) for the variable 'Area cultivada' (S5\_02). The first screenshot shows the 'Analisar' menu with 'Estatísticas descritivas' > 'Tabelas personalizadas...'. The second screenshot shows the 'Tabelas personalizadas' dialog box with 'S5\_02' selected and 'Percentil 75' chosen. The third screenshot shows the 'Estatísticas de resumo' dialog box with 'Percentil 75' selected. The fourth screenshot shows the 'Tabelas personalizadas' output window displaying the result: 'Area cultivada (hectares)? Percentil 75 2,000'.

Neste caso estamos a calcular o valor do terceiro quartil (Q3) através do valor do percentil 75, já que  $Q1 = \text{Percentil } 25$ , logo  $Q3 = \text{Percentil } 75$ .

Abaixo segue o Q3 da variável Área cultivada

The screenshot shows the 'Tabelas personalizadas' output window with the following content:

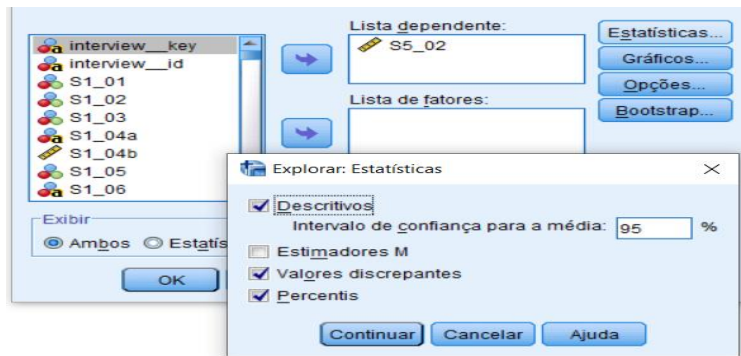
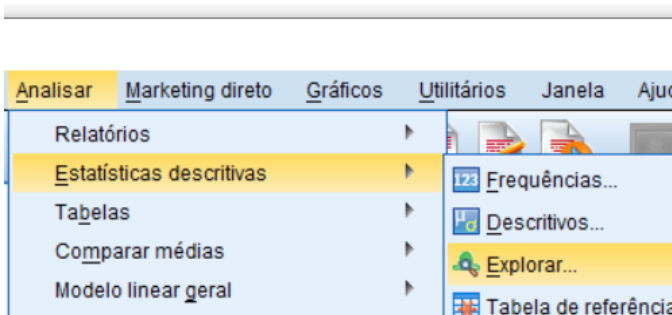
```

* Tabelas personalizadas.
CTABLES
  /VARIABLES=S5_02 DISPLAY=LABI
  /TABLE BY S5_02 (PTILE 75).

Tabelas personalizadas

Area cultivada
(hectares)?
Percentil 75
2,000
  
```

Logo, o valor de **Q3** da variável área cultivada é **2,00**. De seguida calculemos o intervalo interquartil ( $Q3-Q1$ ) também no SPSS.

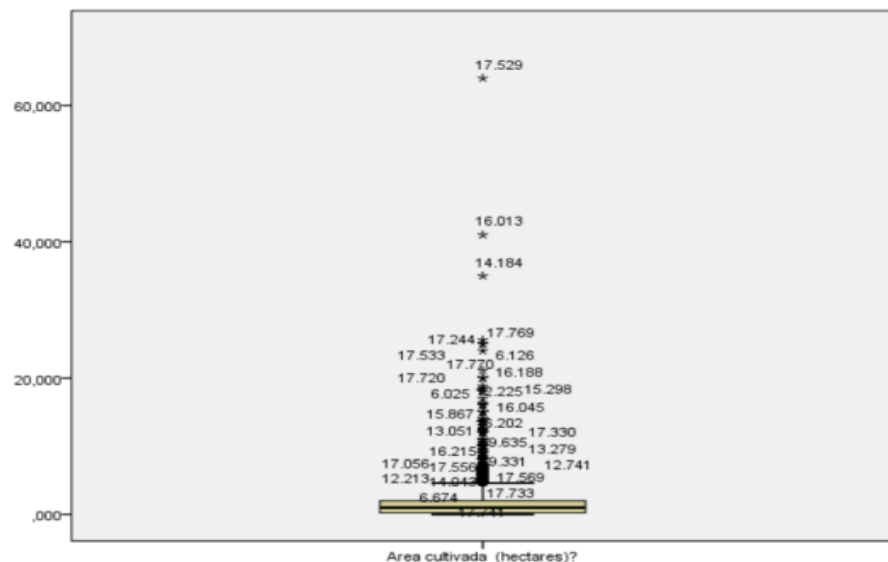


## Explorar

Descritivos			Estadística	Erro Padrão
Área cultivada (hectares)?	Média		1,00000	,018166
	95% Intervalo de Confiança para Média	Limite inferior	1,00000	
		Limite superior	1,00000	
	5% da média aparada		1,00000	
	Mediana		1,00000	
	Variância		4,000	
	Desvio Padrão		2,030072	
	Mínimo		,000	
	Máximo		64,000	
	Intervalo		63,000	
	Intervalo interquartil		1,000	
	Assimetria		6,000	,022
	Curtose		108,000	,044

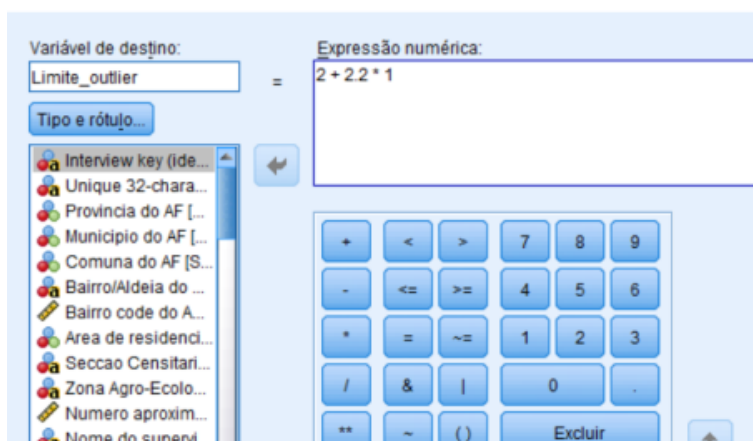
→ Área cultivada (hectares)?

## Área cultivada (hectares)?



Nota-se que o valor do **intervalo interquartil é igual a 1,00**, e o gráfico “Boxplot” ilustra o comportamento dos valores anormais da variável S5\_02 (área cultivada), observa-se por exemplo que os agregados das linhas 14 184, 16 013 e 17 529 apresentam valores muito díspar da média ou dos valores normais, implica dizer que teremos Outliers na variável em estudo em função do limite.

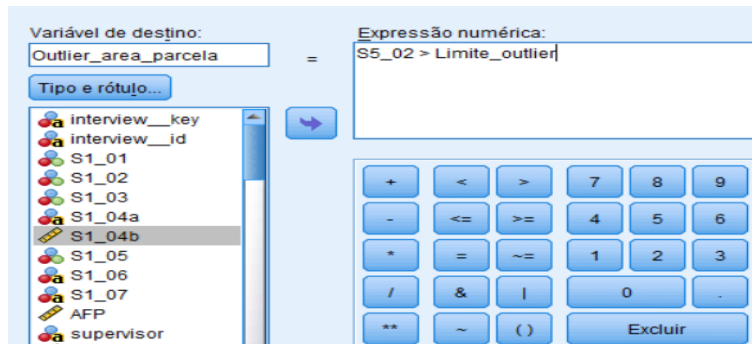
Desta feita, vamos agora no SPSS criar a varável limite:



S5_5	sssys_imd	has_errors	interview...	assignmen...	Limite_outlier
.	,643	,000	ApprovedB...	39,000	4,20
.	,960	,000	ApprovedB...	140,000	4,20
.	,199	,000	ApprovedB...	209,000	4,20
.	,563	,000	ApprovedB...	207,000	4,20
.	,937	,000	ApprovedB...	207,000	4,20
.	,016	,000	ApprovedB...	316,000	4,20
.	,646	,000	ApprovedB...	71,000	4,20
.	,347	,000	ApprovedB...	69,000	4,20
.	,732	,000	ApprovedB...	69,000	4,20
.	,368	,000	ApprovedB...	69,000	4,20
.	,003	,000	ApprovedB...	71,000	4,20
.	,557	,000	ApprovedB...	69,000	4,20
.	,804	,000	ApprovedB...	85,000	4,20
.	,076	,000	ApprovedB...	87,000	4,20



Assim temos o valor “limite\_outlier “ conforme a imagem a direita mostra que ficou em 4,20, logo, todos os valores acima de 4,20 hectares serão considerados como outliers. Criamos agora a variável Outlier da seguinte maneira: Outlier\_area\_parcela= S5\_02> limite\_outlier.



Logo, todas as observações da variável S5\_02 superior ao valor do limite\_outlier (4,20) serão todos outliers. Como podemos observar na figura abaixo:

	Limite_outlier	S5_02	Outlier_area_parcela
0	4,20	64,000	1,00
0	4,20	41,000	1,00
0	4,20	35,000	1,00
0	4,20	25,600	1,00
0	4,20	25,000	1,00
0	4,20	25,000	1,00
0	4,20	25,000	1,00
0	4,20	25,000	1,00
0	4,20	24,000	1,00
0	4,20	21,120	1,00
0	4,20	20,000	1,00
0	4,20	20,000	1,00
0	4,20	20,000	1,00
0	4,20	20,000	1,00
0	4,20	20,000	1,00
0	4,20	20,000	1,00
0	4,20	18,824	1,00
0	4,20	18,500	1,00
0	4,20	18,000	1,00
0	4,20	18,000	1,00
0	4,20	17,000	1,00
0	4,20	16,250	1,00
0	4,20	16,250	1,00
0	4,20	16,000	1,00
0	4,20	16,000	1,00
0	4,20	16,000	1,00
0	4,20	16,000	1,00
0	4,20	15,000	1,00
0	4,20	15,000	1,00
0	4,20	15,000	1,00

#### ▪ Imputação de valores numa variável

Muitas das vezes em uma base de dados, verifica-se em observações de determinadas variáveis em que:

- O valor é Outlier identificado (conforme abordado anteriormente);
- Observação não preenchida ou com valores ausentes (missing) devido a factores diversos como resultado da recolha de dados no campo.

Portanto, não é eficiente prosseguir com a análise dos dados quando existem situações descritas acima, sobretudo quando o número de ocorrências é grande estatisticamente falando,

para resolver ou tratar estas diferenças, utiliza-se um método estatístico denominado **imputação de valores**.

A escolha do método de realização da imputação depende da classificação estatística da variável que se pretende imputar, entretanto descreveremos abaixo os métodos a utilizar pelo SPSS:

- Regressão linear múltipla;
- Imputação pelas estatísticas de tendência central (Média, moda, mediana);
- Interpolação linear.

Nota que tanto um quanto outro método, é aplicável desde que as variáveis em estudo cumpram com os pressupostos de classificação, por exemplo, não se deve imputar uma variável que é quantitativa (escalar) utilizando o método da estatística moda e nem se deve fazer imputação de uma variável qualitativa (nominal ou ordinal) por exemplo, com o método da estatística média ou interpolação linear.

De seguida como exemplo, vamos procurar dar o tratamento aos valores outliers que foram verificados na variável S5\_02 (Área cultivada).

Para tal, como a variável é escalar, pode-se utilizar um dos seguintes métodos:

- Regressão linear;
- Imputação através da Média;
- Interpolação linear.

No nosso caso, para fins práticos vamos utilizar os dois primeiros métodos. Verifica-se por exemplo na base de dados em trabalho que, as observações 1 à 27 são outliers.

Base\_listagem\_sem\_dup\_sav\_VT.sav [Conjunto\_de\_dados1] - IBM SPSS Statistics Editor de dados

1: S5_02	64,000							
	sssys_irm	has_errors	interview_st atus	assignment_ id	Limite_outlier	S5_02	Outlier	
1	,678	1,000	ApprovedB...	400,000	4,20	64,000	1,00	
2	,851	2,000	ApprovedB...	400,000	4,20	25,000	1,00	
3	,254	,000	ApprovedB...	52,000	4,20	25,000	1,00	
4	,736	,000	ApprovedB...	158,000	4,20	20,000	1,00	
5	,138	,000	ApprovedB...	52,000	4,20	20,000	1,00	
6	,289	,000	ApprovedB...	3,000	4,20	15,000	1,00	
7	,931	,000	ApprovedB...	83,000	4,20	15,000	1,00	
8	,807	,000	ApprovedB...	246,000	4,20	15,000	1,00	
9	,262	,000	ApprovedB...	246,000	4,20	15,000	1,00	
10	,449	,000	ApprovedB...	32,000	4,20	14,000	1,00	
11	,828	,000	ApprovedB...	400,000	4,20	13,000	1,00	
12	,010	,000	ApprovedB...	154,000	4,20	12,500	1,00	
13	,049	,000	ApprovedB...	141,000	4,20	12,000	1,00	
14	,523	,000	ApprovedB...	18,000	4,20	12,000	1,00	
15	,471	,000	ApprovedB...	369,000	4,20	12,000	1,00	
16	,127	,000	ApprovedB...	4,000	4,20	12,000	1,00	
17	,675	,000	ApprovedB...	153,000	4,20	12,000	1,00	
18	,100	,000	ApprovedB...	18,000	4,20	11,925	1,00	
19	,055	,000	ApprovedB...	183,000	4,20	10,972	1,00	
20	,728	,000	ApprovedB...	300,000	4,20	10,940	1,00	
21	,122	,000	ApprovedB...	256,000	4,20	10,000	1,00	
22	,657	,000	ApprovedB...	400,000	4,20	10,000	1,00	
23	,368	,000	ApprovedB...	3,000	4,20	10,000	1,00	
24	,509	,000	ApprovedB...	358,000	4,20	10,000	1,00	
25	,167	,000	ApprovedB...	397,000	4,20	10,000	1,00	
26	,227	,000	ApprovedB...	9,000	4,20	10,000	1,00	
27	,198	,000	ApprovedB...	210,000	4,20	10,000	1,00	

Além disso, existem um conjunto de observações ausentes (missing) conforme a imagem abaixo.

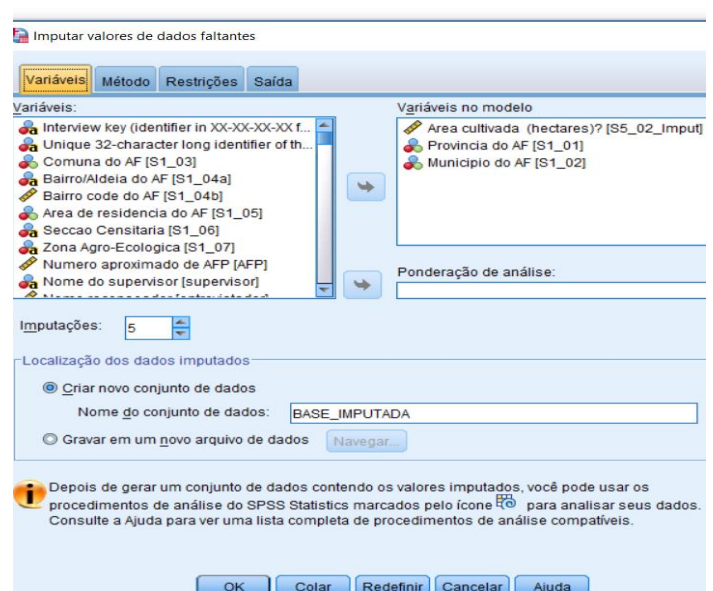
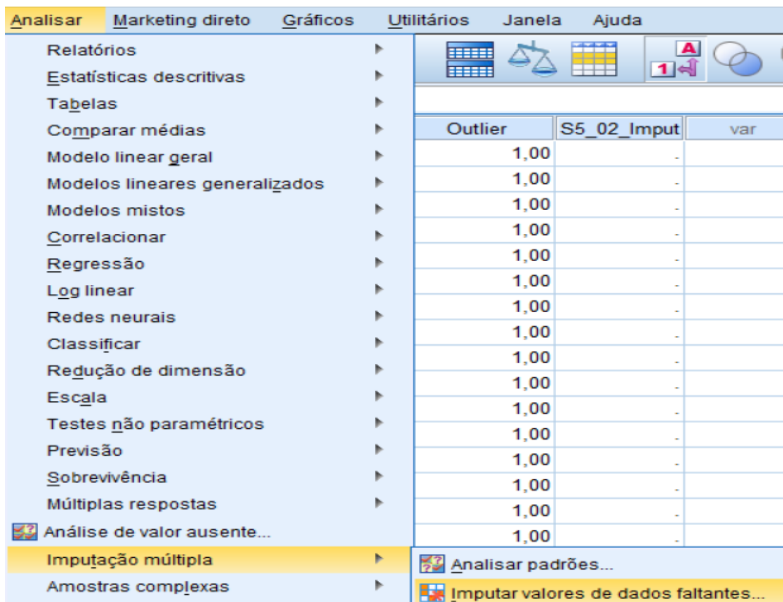
1: S5_02	64.000					
rrors	interview	assignmen...	Limite_outlier	S5_02	Outlier	
12596	.000	ApprovedB...	34.000	4,20		
12597	.000	ApprovedB...	34.000	4,20		
12598	.000	ApprovedB...	34.000	4,20		
12599	.000	ApprovedB...	53.000	4,20		
12600	.000	ApprovedB...	65.000	4,20		
12601	.000	ApprovedB...	65.000	4,20		
12602	.000	ApprovedB...	65.000	4,20		
12603	.000	ApprovedB...	65.000	4,20		
12604	.000	ApprovedB...	65.000	4,20		
12605	.000	ApprovedB...	65.000	4,20		
12606	.000	ApprovedB...	67.000	4,20		
12607	.000	ApprovedB...	119.000	4,20		
12608	.000	ApprovedB...	119.000	4,20		
12609	.000	ApprovedB...	354.000	4,20		
12610	.000	ApprovedB...	100.000	4,20		
12611	.000	ApprovedB...	378.000	4,20		
12612	.000	ApprovedB...	190.000	4,20		
12613	.000	ApprovedB...	190.000	4,20		
12614	.000	ApprovedB...	190.000	4,20		

Deste modo, teremos que efectuar a imputação por duas seguintes razões: Primeiro para substituímos os valores outliers, segundo, para preenchermos os valores “missings”.

Primeiro vamos efectuar a imputação pelo método de **regressão linear múltipla**:

Deve-se inicialmente apagar todas as observações que são outliers na variável em questão, fazendo sempre uma cópia da variável original, assim a variável a considerar como cópia será S5\_02\_Imput.

Sigamos então os passos:



Como podemos observar nas figuras acima, a variável a ser imputada é (área cultivada), neste caso é a variável dependente do modelo e escolhemos como variáveis independentes ou preditivas, a província e município que são estratos geográficos, poderíamos adicionar mais

variáveis produtoras, como as escolhidas serão suficientes para cobrir todos os dados falantes nos estratos geográficos necessários, então ficamos apenas com estas.

Observa-se que durante o processo de imputação é necessário criar uma nova base com conjunto de dados a serem imputados, e no item “imputações” é onde devemos inserir o número de iterações requeridas no modelo, para cada iteração há-de gerar um conjunto de observações na mesma base criada, sendo que a 1ª iteração é com base já a imputação da variável dependente, assim senso, cada iteração tende a mostrar um valor melhorado relativamente ao outro até encontrar um valor óptimo na enésima ou última iteração e por final, o modelo fornece um conjunto de observações como resumo médio de todas as iterações. Para este exemplo, escolhemos aleatoriamente (predefinido pelo SPSS) 5 como o número de imputações.

Imputar valores de dados faltantes

Verificação de dados para resumo de variável

Verificar dados ☐ Número limite de casos verificados Casos: 5000

Resumo da variável:

Variáveis no modelo	Porcentagem ausente	Mín. observado	Máx. observado
S5_02_Imput			
S1_01			
S1_02			

Casos verificados: none

Definir restrições:

Variáveis no modelo	Função	Mín.	Máx.	Arredondam...
S5_02_Imput	Apenas imputar	0		
S1_01	Imputar e usar como preditor			
S1_02	Imputar e usar como preditor			

☒ Excluir variáveis com grandes quantidades de dados faltantes

Porcentagem máxima ausente:

Máximo de coletas de caso: 50

Máximo de coletas de parâmetro: 2

O aumento das coletas de parâmetro máximo pode aumentar significativamente o tempo de análise.

OK Colar Redefinir Cancelar Ajuda

A seguir a janela de introdução das variáveis do modelo, é necessário irmos no menu restrições conforme mostra a figura acima. Neste menu, é onde definimos a **função** de cada variável no modelo, neste caso a variável S5\_02\_Imput é somente variável dependente (a ser imputada), já as demais escolhemos para ambas a função de imputar e usar como produtora. Talvez possa surgir a pergunta porquê usar esta função? A resposta é a seguinte: as variáveis pela qual queremos que elas sejam explicativas ou predito rãs, não devem conter missings, neste caso, devemos prevenir o modelo no sentido de que se eventualmente haja missing na variável preditora, então o modelo primeiramente faz uma pré-imputação interna das variáveis independentes ou produtoras, só assim é que de seguida fará a imputação da variável em causa (dependente). Por outro lado, os itens Mín. e Máx. servem para definir os valores mínimos e máximos que a variável a ser imputado devem assumir, para o nosso caso, precisamos apenas definir o mínimo que é zero (0), já que de ponto de vista geométrico, a área nunca assume valores negativos, para o valor máximo não é necessário pois já temos calculado como o limite

de Outlier. Feitos estes procedimentos, podemos prosseguir clicando em ok. Mais abaixo temos as tabelas resumo do resultado da operação.

Especificações de imputação		
Método de imputação	Automático	5
Número de imputações		
Modelo para variáveis de escala	Regressão linear	
Interações incluídas nos modelos	(nenhum)	
Porcentagem máxima de valores ausentes		100,0%
Número máximo de parâmetros no modelo de imputação		100

Restrições de imputação				
	Função na imputação		Valores imputados	
	Dependente	Preditor	Mínimo	Máximo
Área cultivada (hectares)?	Sim	Não	0	(nenhum)
Provincia do AF	Sim	Sim		
Município do AF	Sim	Sim		

#### Valores imputados

Resultados de imputação		
Método de imputação		Uniforme
Iterações de método de especificação totalmente condicional		n/a
Variáveis dependentes	Imputado	S5_02_Imput
	Não imputado (muitos valores ausentes)	
	Não imputado (nenhum valor ausente)	S1_01,S1_02
Sequência de imputação		S1_01,S1_02,S5_02_Imput

Modelos de imputação				
	Modelo		Valores ausentes	Valores imputados
	Tipo	Efeitos		
Área cultivada (hectares)?	Regressão linear	S1_01,S1_02	6089	30307

Neste caso, temos 30 307 casos imputados na base em geral, quer dizer que para cada uma das 5 imputações, o modelo imputou cerca de 6 089 observações que são efectivamente valores ausentes ou missings.

**Nota:** A base de dados de trabalho tem  $n$  observações (17 835), portanto, a base gerada pelo modelo após a realização da imputação terá  $5 * n + n = 6n$  observações, neste caso, 107 010, isto em função do número de imputações (5) que foi requerida no modelo e a sexta observação é o resumo médio agrupado de todas as imputações. De seguida guardemos a base gerada pelo modelo com o nome criado a quando da imputação.

\*Base\_imput.sav [IMPUT] - IBM SPSS Statistics Editor de dados

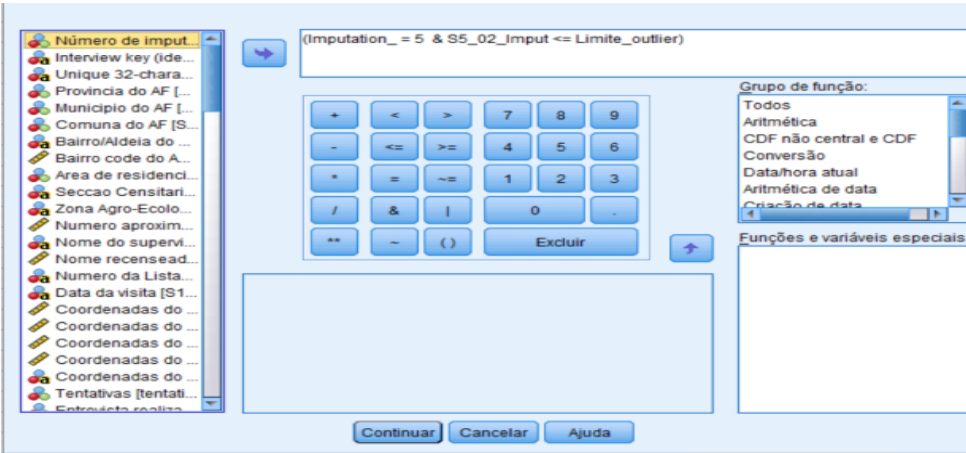
	has_errors	interview_status	assignment_id	Limite_outlier	S5_02	Outlier	S5_02_Imput
107000	,000	ApprovedB...	47,000	4,20	6,750	1,00	.
107001	,000	ApprovedB...	280,000	4,20	.	.	.
107002	,000	ApprovedB...	101,000	4,20	.	.	.
107003	,000	ApprovedB...	38,000	4,20	.	.	.
107004	,000	ApprovedB...	4,000	4,20	.	.	.
107005	,000	ApprovedB...	13,000	4,20	.	.	.
107006	,000	ApprovedB...	287,000	4,20	.	.	.
107007	,000	ApprovedB...	291,000	4,20	.	.	.
107008	,000	ApprovedB...	212,000	4,20	.	.	.
107009	,000	ApprovedB...	226,000	4,20	.	.	.
107010	,000	ApprovedB...	263,000	4,20	.	.	.

Portanto, esta é a base de dados com casos imputados, como podemos observar a ultima linha marca o caso nº 107 010, o que quer dizer que este ficheiro de dados contém este nº de observações. Entretanto, para trabalho, temos que filtrar nesta base apenas um conjunto de dados com apenas um nº de imputação da variável Imputation\_, por exemplo filtraremos a base com a 5ª imputação, neste caso no SPSS será Select if (Imputation\_=5). Geralmente escolhe-se a última imputação, pois alcança a optimalidade até a última observação da base de dados, além de apresentar um desvio padrão nulo relativamente a média.

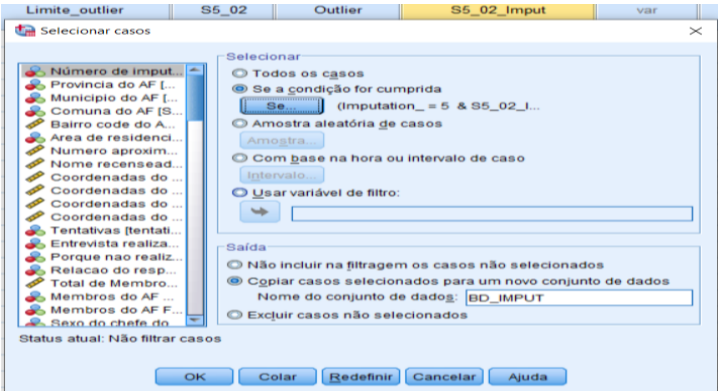
Descritivos

Estatísticas descritivas						
Número de imputação		N	Mínimo	Máximo	Média	Desvio Padrão
Dados originais	Area cultivada (hectares)?	11 746	,000	4,000	1,04396	1,015667
	N válido (de lista)	11 746				
1	Area cultivada (hectares)?	17 797	,000	4,000	1,00000	,000000
	N válido (de lista)	17 797				
2	Area cultivada (hectares)?	17 820	,000	4,000	1,00000	,000000
	N válido (de lista)	17 820				
3	Area cultivada (hectares)?	17 761	,000	4,000	1,00000	,000000
	N válido (de lista)	17 761				
4	Area cultivada (hectares)?	17 824	,000	4,000	1,00000	,000000
	N válido (de lista)	17 824				
5	Area cultivada (hectares)?	17 835	,000	5,031	1,00000	,000000
	N válido (de lista)	17 835				
Agrupado	Area cultivada (hectares)?	17 807,0			1,00000	
	N válido (de lista)	17 807,0				

Aplicando o filtro, especificado acima e tendo em conta o limite\_outlier temos:



Feito isso, devemos renomear a base e salvar

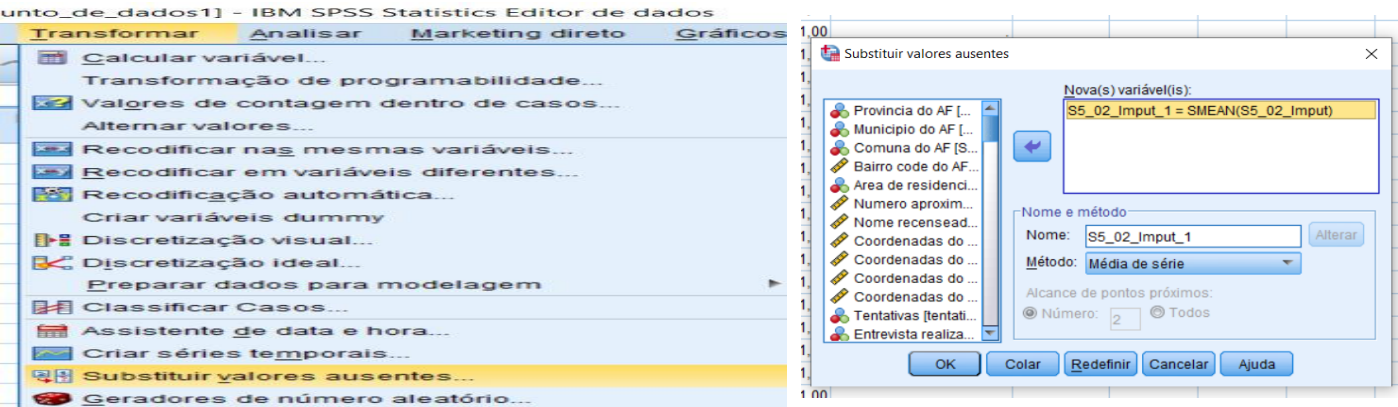


Finalmente temos a base de dados com casos imputados e sem outliers na variável em causa, conforme mostra a figura a seguir:

Limite_outlier	S5_02	S5_02_Imput
4,20	64,000	1,647
4,20	41,000	,234
4,20	35,000	,678
4,20	25,600	1,850
4,20	25,000	1,837
4,20	25,000	,177
4,20	25,000	,160
4,20	24,000	1,130
4,20	21,120	1,531
4,20	20,000	2,161
4,20	20,000	1,224



De seguida, vamos agora fazer a imputação através da Média da série, para tal, vamos no SPSS em substituir valores ausentes como mostra a figura a seguir:



Para os devidos efeitos, a variável original permanecerá intacta, entretanto a nova variável imputada será S5\_02\_Imput\_1 como observa-se na figura acima a direita. Abaixo segue a tabela resumo da observação:

**Substituir valores ausentes**

Variáveis de Resultado					
	Variável de Resultado	N de Valores Ausentes Substituídos	Número de Caso de Valores Não ausentes		Criando Função
			Primeiro	Último	
1	S5_02_Imput_1	6089	1	17835	SMEAN (S5_02_Imput)

Neste caso temos 6 089 observações imputados e o número total de observações na base de dados é de 17 835 casos. Abaixo mostramos e descrevemos os resultados da imputação:

Limite_outlier	S5_02	Outlier	S5_02_Imput	S5_02_Imput_1
4,20	64,000	1,00	-	1,044
4,20	41,000	1,00	-	1,044
4,20	35,000	1,00	-	1,044
4,20	25,600	1,00	-	1,044
4,20	25,000	1,00	-	1,044
4,20	25,000	1,00	-	1,044
4,20	25,000	1,00	-	1,044
4,20	24,000	1,00	-	1,044
4,20	21,120	1,00	-	1,044
4,20	20,000	1,00	-	1,044
4,20	20,000	1,00	-	1,044
4,20	20,000	1,00	-	1,044
4,20	20,000	1,00	-	1,044
4,20	20,000	1,00	-	1,044
4,20	18,824	1,00	-	1,044
4,20	18,500	1,00	-	1,044
4,20	18,000	1,00	-	1,044
4,20	18,000	1,00	-	1,044
4,20	17,000	1,00	-	1,044
4,20	16,250	1,00	-	1,044
4,20	16,250	1,00	-	1,044
4,20	16,000	1,00	-	1,044
4,20	16,000	1,00	-	1,044
4,20	16,000	1,00	-	1,044
4,20	15,000	1,00	-	1,044
4,20	15,000	1,00	-	1,044

A primeira coluna é a do limite definido na secção anterior sobre os outliers, lembrando que o limite assumido para a variável área cultivada é de até 4,20 Há, a segunda coluna é a variável original área cultivada onde inclui as observações outliers, a terceira é a variável Outlier que define efectivamente os casos nesta condição, a quarta coluna é da variável cópia a ser imputada a qual anteriormente designamos por S5\_02\_Imput onde inclui as observações normais e retiradas os casos outliers de modo a tornar missings nestas observações, já última coluna, diz respeito a variável já imputada (S5\_02\_Imput\_1) através do método de imputação da média da série. Como podemos notar por exemplo, as observações da variável S5\_02 agora é totalmente diferente da variável imputada S5\_02\_Imput\_1, é esta que deve ser considerada para análises, visto que já foi corrigido os valores atípicos.

Portanto, a verificação de casos outliers e processo de imputação de valores missings estão estritamente ligados e são muito importantes no decurso do tratamento de dados estatísticos, pois permitem assegurar a análise estatística e por conseguinte a qualidade da informação em particular de uma dada variável e de um modo geral, de uma base de dados.

## 9. ANÁLISE DE DADOS

### Conceitos e definições

**A análise de dados:** é a etapa técnica que é caracterizada fundamentalmente com a manipulação de elementos de análise estatística de modo a relacionar as evidências quantitativas e qualitativas.

Uma análise estatística abrange a fase de recolha e da avaliação dos dados corporativos, bem como dos relatórios de tendências e a análise efectiva da totalidade e também em amostras individuais.

Existem basicamente dois tipos de análise de dados estatísticos: a descritiva e a modelagem.

A primeira é usada para resumir os dados, descrevendo-os, enquanto a outra é utilizada para testar uma hipótese e tirar conclusões, procurando assim uma significância entre os dados.

A análise descritiva é assente na manipulação de medidas estatísticas tais como medidas de localização (média, moda, mediana, quartis, decís), dispersão (Variância e desvio padrão), bem como na elaboração de tabelas cruzadas, gráficos, cartogramas... e fazer a interpretação destes. Para tal, é importante compreender e saber as linguagens analíticas em estatística.



**Média ponderada** ( $\bar{x}_w$ ): Quando temos uma questão de resposta em escala, pode ser interessante chegar a um indicador único de valorização, através do cálculo de uma média ponderada.

Exemplificando, para uma questão de satisfação temos a seguinte escala de respostas, com os respectivos pesos de acordo com a opinião dos inquiridos:

1.Totalmente insatisfeito	10
2.Insatisfeito	15
3.Nem satisfeito, nem insatisfeito	25
4.Satisfeito	30
5.Totalmente satisfeito	20

A média ponderada calcula-se da seguinte forma:  
 $\bar{x}_w = (1*10+2*15+3*25+4*30+5*20)/100=3,35$   
 Em média os inquiridos estão entre Nem satisfeito, nem insatisfeito (3) e Satisfeito (4).

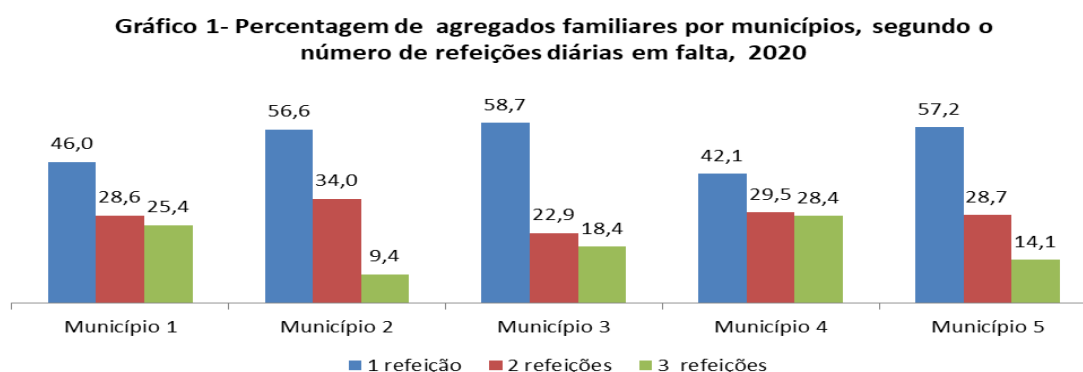
#### ▪ Tipos de gráficos mais usuais na análise descritiva

Para uma boa análise de dados, os gráficos são visualmente atraentes facilitando uma melhor comparação, de padrões e de tendências dos dados. Por exemplo, em vez de analisar diferentes colunas de números de uma folha de cálculo ou tabela, poderemos visualizar rapidamente os aumentos ou diminuições quantitativas das variáveis em estudo através de gráficos.

Desta feita, vamos destacar alguns tipos de gráficos utilizados em análise estatística

#### ▪ Gráfico de colunas

O gráfico de colunas mostra alterações nos dados ao longo de um período de tempo ou ilustra comparações entre itens. As categorias são organizadas na vertical e os valores na horizontal de modo a realçar a variação ao longo do tempo. Exemplo:



**Fonte:** Pesquisa sobre SAM, 2020

- **Gráfico de barras**

**Gráfico 2-Percentagem de agregados familiares, segundo a dependência para conseguir uma refeição, 2020**

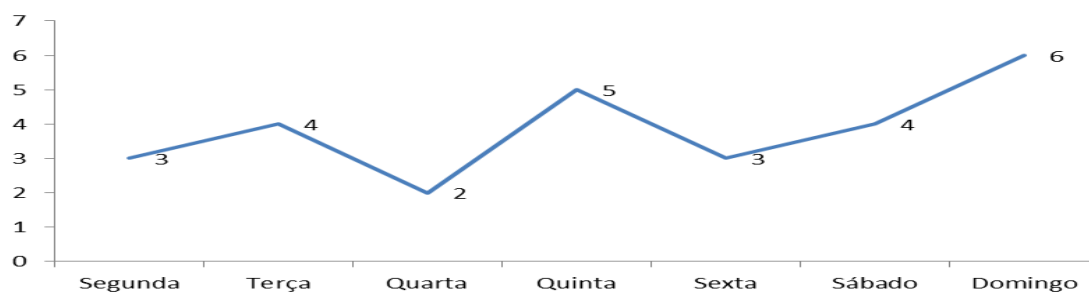


O gráfico de barras, geralmente é utilizada para ilustrar comparações entre itens individuais e quando os títulos dos itens são muito extensos. Exemplo:

- **Gráfico de linhas**

O gráfico de linhas é utilizado geralmente para mostrar tendências em dados em intervalos iguais ou também para comparar séries temporais. Exemplo:

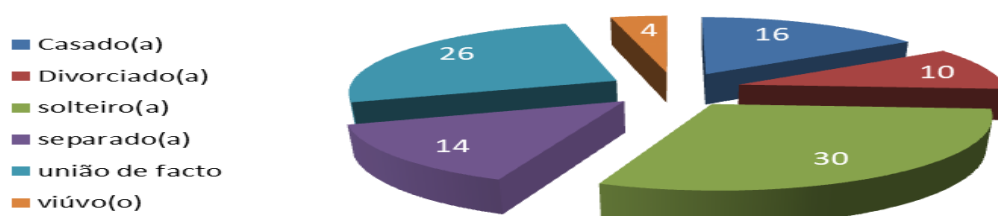
**Gráfico 3-Média de refeições por agregado familiar ao longo dos últimos 7 dias de semana, antes do inquérito, 2020**



- **Gráfico circular**

O gráfico circular ou também chamado de “pitsa” é utilizado geralmente para mostrar o tamanho proporcional dos itens que constituem uma série de dados relativamente à soma dos itens. Exemplo:

**Gráfico 4- Distribuição percentual da população inquirida, segundo o estado civil, 2020**



- **Algumas técnicas em análise de dados**

Em análise de dados, muitas das vezes quando estes estiverem em percentagem as vezes há necessidade de expressá-los em fracções como parte de um todo ou ainda como uma proporção.

Vale lembrar antes as seguintes definições de ponto de vista Matemático:

**Proporção:** Dado dois números inteiros  $a$  e  $b$  ( $b \neq 0$ ), chama-se proporção de  $a$  relativamente  $b$ , ao quociente de  $a$  sobre a soma entre  $a$  e  $b$ , ou seja,  $\frac{a}{a+b}$ .

**Razão:** Dado dois números inteiros  $a$  e  $b$  ( $b \neq 0$ ), chama-se razão de  $a$  relativamente  $b$ , ao quociente entre  $a$  e  $b$ , ou seja,  $\frac{a}{b}$ .

**Percentagem:** A percentagem é geralmente uma unidade para expressar uma grandeza estatística diante de um fenómeno resultante de uma proporção, ou razão em cada 100 observações. Neste caso, o valor máximo da proporção vale 1, e em percentagem é expresso como 100%. Portanto, para obtermos a percentagem, basta multiplicar a proporção ou a razão por 100.

**Taxa:** É uma unidade estatística normalmente usada para expressar a prevalência, variação, intensidade, o estado de uma situação ou fenómeno através de um indicador num período de tempo e espaço geográfico, é expresso geralmente em percentagem, por mil, etc...

Frequentemente também tem-se recorrido a transformação de expressões aritméticas. Vejamos por exemplo:

No agregado familiar “a” com 11 membros residentes, verificou-se que 2 deles possuem telemóvel.

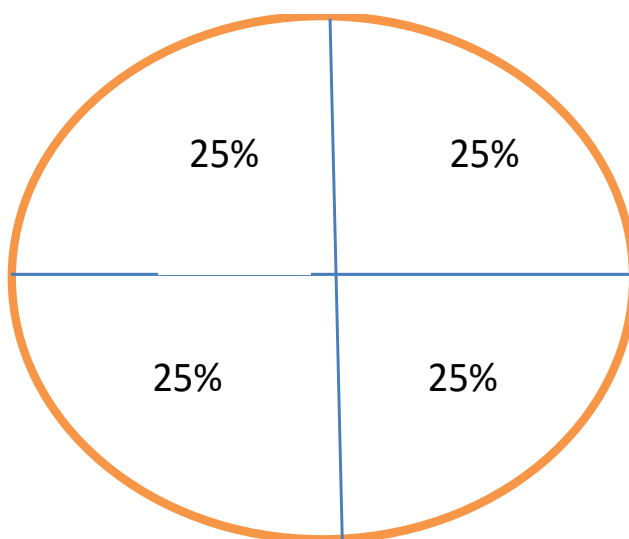
Quantos porcentos equivalem a pessoas que possuem telemóvel?

**Solução:** Sabendo que o total de membros no agregado “a” é 11, neste caso é o nosso universo que em percentagem vale a 100%.

Para respondermos a nossa pergunta inicial basta acharmos a proporção dos que têm telemóvel, ou seja, temos que dividir o número de membros no agregado familiar que têm telemóvel pelo total de membros, neste caso é:  $2/11=0,27$ . O valor obtido é a proporção de membros que usam o telemóvel.

Para expressar em percentagem basta multiplicar por 100, assim temos:  $0,27 \times 100\% = 27\%$ .

Se quisermos transformar este resultado em fracção como parte de um todo, temos que ter em conta que o nosso universo é 100%, assim sendo vamos dividir por exemplo este universo de ponto de vista geométrico em 4 partes iguais (cada uma das partes vale 25%).



Pelo facto do universo estar dividido em 4 partes iguais, cada uma dessas partes chamar-se-á de um quarto, que matematicamente escreve-se  $\frac{1}{4}$ . Já que  $\frac{1}{4}=0,25$  (em percentagem é 25%).

Notemos que no exemplo dado, a percentagem de membros que têm um telemóvel é de 27%, ou seja,  $27\% = 25\% + 2\%$ .

Podemos dizer que a percentagem de membros no agregado que possuem um telemóvel é igual a um pouco mais de um quarto (27%).

▪ **Outros argumentos para relacionar o fenómeno observado vs Universo**

- ✓ Verificar se o valor percentual do indicador observado encontra-se na classe das dezenas;
- ✓ Se a alínea anterior está satisfeita, então devemos arredondar este valor percentual para um número inteiro, caso necessite;
- ✓ Ao arredondar tem que se ter em conta os princípios de arredondamento, se o último algarismo for menor que 5, arredondamos por defeito, caso contrário arredondamos por excesso;
- ✓ Feito isto, expressamos o valor percentual em fracção, em que o denominador deve ser 100, e o numerador o valor arredondado na alínea anterior;
- ✓ Simplificamos a fracção obtida tornando-a irredutível (Se for conveniente);

- ✓ O numerador obtido após simplificação será “x” e o denominador “y”, que no final ler-se-á: x em cada y.

Este procedimento, de ponto de vista estatístico significa que estamos a reduzir o nosso universo ou população (y - denominador) para cada elemento observado na amostra (x- numerador), com objectivo de facilitar a compreensão analítica e qualitativa da informação.

Façamos este exercício com o resultado obtido no exemplo dado (27%).

Vinte e sete por cento, deve ser arredondado por excesso para trinta, já que o último algarismo 7 é > 5, então escrevemos:  $x = 27\% \cong 30\%$  (expressando este valor em fracção temos:  $30\% = 30/100 = 3/10$  (de forma simplificada) a fracção obtida ler-se-á como: cerca de 3 em cada 10.

**Nota:** estas técnicas permitem que a relação textual e quantitativa no relatório, seja mais atraente e menos monótona, de modo a aliciar e situar melhor o leitor na compreensão do fenómeno em análise.

Com base as instruções e exemplo anterior, Faça os exercícios abaixo:

- 1.Considere agora que o agregado “a” tem 9 membros e apenas 6 possuem telemóvel.
- 2.Faça o mesmo exercício para o valor 20,3%.

## 10. ELABORAÇÃO DE RELATÓRIO ANALÍTICO

Quando os dados são irrelevantes, inconclusivos, insuficientes, não se pode nem confirmar nem contestar a hipótese, e tal facto deve ser apontado agora não apenas sob o ângulo da análise estatística, mas também correlacionada. No relatório, é necessário assinalar:

- As discrepâncias entre os factos obtidos e os previstos nas hipóteses;
- A comprovação ou a contestação da hipótese, ou ainda, a impossibilidade de realizá-la;
- Qual é o valor da generalização dos resultados para o universo, no que se refere aos objectivos determinados;
- Como as provas obtidas mantêm a sustentabilidade da teoria, determinam sua limitação ou, até, a sua rejeição.

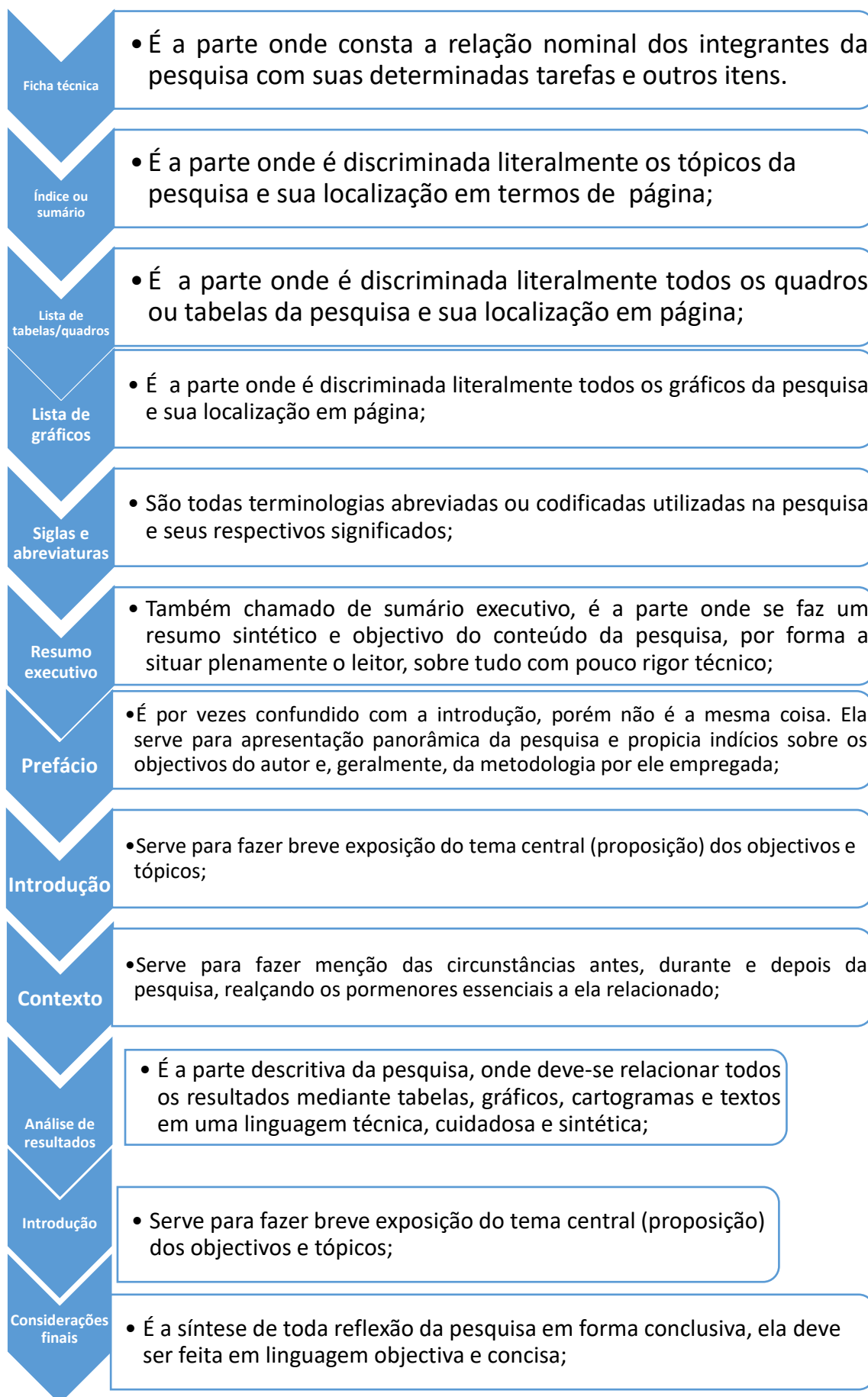
### 10.1. Estrutura de um relatório de pesquisa

Em geral, a elaboração de um relatório de pesquisa científica obedece a seguinte estrutura:

- Apresentação;
- Resumo;
- Índice;
- Introdução;
- Metodologia;
- Fundamentação teórica;
- Análise e apresentação dos resultados;
- Conclusões;
- Recomendações;
- Apêndices;
- Anexos;
- Referências Bibliográficas.

Para o presente caso, tendo em conta a especificidade da nossa temática, vamos nos basear apenas em estruturas de relatórios de pesquisas não puramente académicas, mas sim em publicações estatísticas do tipo relatório analítico.

A estrutura que poderemos adoptar é a seguinte:



## Anexos

- Geralmente constituídos de elementos esclarecedores de outra autoria, devem ser limitados, incluindo apenas o estritamente necessário à compreensão de partes do relatório, mas também pode-se incluir elementos da pesquisa que não cabem no corpo descritivo do relatório;

## Bibliografia

- Inclui todas as obras já apresentadas na pesquisa, acrescidas das que foram sendo sucessivamente utilizadas durante a execução da pesquisa e a redação do relatório;



**Anexo:** Tabela de auxílio para selecção aleatória simples

Tabela de números aleatórios para amostragem simples												
Dia de referência												
		1	2	3	4	5	6	7	8	9	10	11
k=N/n	1	1	1	1	1	1	1	1	1	1	1	1
	2	2	1	2	2	1	2	2	2	2	1	1
	3	2	2	1	3	2	3	1	1	3	3	2
	4	2	2	2	3	1	3	4	3	3	3	1
	5	1	3	1	1	3	1	5	4	3	2	5
	6	3	5	6	2	4	2	2	1	4	6	5
	7	7	2	7	5	2	2	2	1	4	6	3
	8	4	6	3	8	6	5	2	8	3	2	5
	9	8	9	6	7	7	3	4	8	1	7	4
	10	10	8	4	10	1	3	1	5	1	10	2
	11	4	2	10	1	5	7	11	5	7	1	7
	12	5	2	12	7	8	1	10	4	10	2	11
	13	7	4	13	5	1	8	4	11	13	4	10
	14	11	11	3	8	8	3	14	7	11	1	4
	15	9	15	1	3	3	1	5	12	2	4	8
	16	4	15	12	7	9	9	14	14	2	1	15
	17	12	5	10	7	3	11	11	6	3	17	10
	18	4	13	11	16	17	5	18	9	3	4	6
	19	17	8	12	8	8	15	12	3	6	9	1
	20	1	12	19	20	18	1	1	10	9	17	17
	21	13	7	1	7	13	11	16	18	12	3	12
	22	3	9	6	19	20	14	3	16	19	7	18
	23	3	8	7	18	16	14	16	7	7	15	8
	24	6	23	15	4	19	21	15	17	6	24	3
	25	24	3	18	7	4	19	4	15	22	25	8

## REFERÊNCIAS BIBLIOGRÁFICAS

1. Marconi & Lakatos, Fundamentos de Metodologia Científica, 5ª edição, 310p, editora atlas. S.Paulo, Brasil, 2003.
2. L. Helena, Visão geral do processo de pesquisa por questionário, 100p, ISEGI da UNOVA, Lisboa, Portugal, 2018;
3. M. Fernando de Pol, Introdução à Estatística e conceitos de amostragem, 48p, UFPR, 2015;
4. Bussab, WO; Morettin, PA. Estatística básica. São Paulo: Saraiva, 2002. 526 p. [Cap. 1 e 10];
5. F. Jairo & M.Gilberto, curso de estatística, 6ª edição, 318p, editora atlas. S.Paulo, Brasil, 2011;
6. Programa Mundial de Censo Agropecuário 2020, Volume 1, definições e conceitos, 217 p, Roma, Itália, 2015;
7. Google, gestor de buscas