UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA
Caderno de Matemática e Estatística
Série B: Trabalho de Apoio Didático

INTRODUÇÃO AOS MÉTODOS ESTATÍSTICOS

utilizando o software SPSS Versão 8.0

Patrícia Klaser Biasoli
Jandyra Maria Guimarães Fachel
Suzi Alves Camey

Série B, Número 57
Porto Alegre - junho de 2001

1. INTRODUÇÃO AO SOFTWARE SPSS

Tela inicial do *SSPS 8.0 for Windows.*

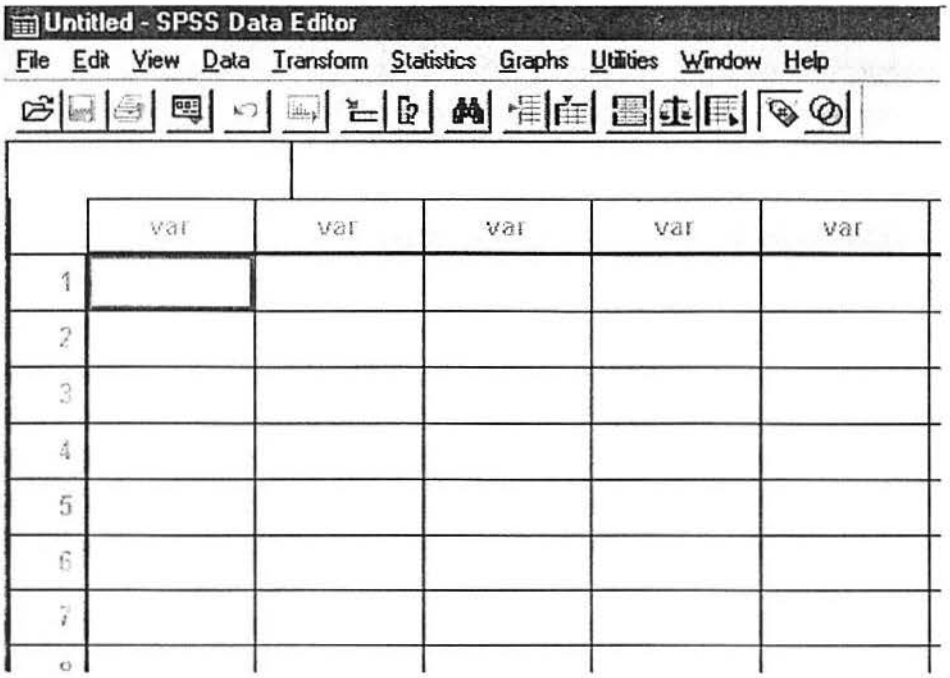


Figura 1: Tela inicial do *SSPS 8.0 for Windows.*

O pacote estatístico **SPSS** (Statistical Package for Social Science – for Windows) é uma das ferramentas disponíveis para análise de dados utilizando **técnicas estatísticas** básicas e avançadas. É um software estatístico de fácil manuseio e é internacionalmente utilizado há muitas décadas, desde suas versões para computadores de grande porte.

1.1- BANCO DE DADOS: Definição

Banco de dados é um conjunto de dados registrados em uma planilha, matriz, com "n" linhas, correspondentes aos casos em estudo e "p" colunas, correspondentes às variáveis em estudo ou itens de um questionário.

O número de casos (número de linhas da matriz) deve ser, em geral, **maior** do que o número de variáveis em estudo (número de colunas).

1.2- COMO CRIAR UM BANCO DE DADOS

Para se criar um BANCO DE DADOS novo procede-se da seguinte forma:

a) Clica-se em **"File"**; **"New"**; **"Data"**; aparece uma planilha onde na primeira linha estão indicadas as posições das variáveis, e uma margem numerada de **1** a **n** (como mostrado na Figura1);
b) Na primeira coluna, correspondendo à VAR001, cria-se uma variável, por exemplo "NumCaso" com o número do questionário ou do caso em estudo;
c) Para serem registradas as variáveis em cada coluna, clica-se **duas vezes** sobre a coluna **"Var"** . Aparece uma janela **"Define Variable"** onde em **"Variable Name"**, digita-se o nome da variável desejada (8 dígitos no máximo). Para o nome das variáveis não utilize espaço em branco, nem os símbolos "-", "." e "/";
d) No caso de não-resposta ou respostas que não se deseja considerar para o tratamento estatístico, como por exemplo, respostas não corretas, etc... a melhor opção é deixar o espaço em branco no banco de dados. Também pode-se utilizar códigos usuais de não-resposta, como p.ex. 9, 99, 999. Neste caso, clica-se na opção **"Missing Values"**, abre-se uma janela e registra-se, na opção **"Discrete Missing Values"** o código de não-resposta, preferencialmente **9, 99, 999, etc**; Clica-se em **"Continue"** e clica-se em **"OK";**
e) Para definir se a variável é ou não numérica, clica-se no botão **"Type"**, aparece uma janela **"Define Variable Type"** onde deve-se deixar a opção **"Numeric"** se a variável for numérica; ou **"String"** se a variável for alfa-numérica, isto é quando a resposta é aberta e vamos digitar uma frase ou parágrafo. Preferencialmente use sempre a modalidade "Numeric" para variáveis categóricas, como por exemplo, sexo, estado civil, profissão e cria-se um código para as categorias;
f) Passa-se a digitar, em cada linha da coluna identificada, o valor correspondente da variável fornecido pelo respondente no questionário ou o valor gravado para cada registro (caso a ser estudado);

g) A medida que o BANCO DE DADOS vai sendo registrado é importante salvar as informações digitadas, para tanto procede-se da seguinte forma: Clica-se em **"File"** , **"Save as"**... (abre-se a janela do caminho desejado) e cria-se um nome para o Banco de Dados, que terá automaticamente a terminação **.sav**.

1.3- COMO DAR NOME AOS NÍVEIS DE UMA VARIÁVEL EM UM BANCO DE DADOS QUE ESTÁ SENDO CRIADO

É conveniente registrar no banco de dados os nomes (labels) das categorias das variáveis categóricas. Por exemplo, para a variável **sexo**, os códigos poderiam ser: **"1" = "masculino"** e **"2" = "feminino"**. Para registrar estes nomes, clica-se **2 vezes** sobre a variável **sexo**, obtendo-se a janela **"Define Variable"**.

No exemplo, para dar o nome aos níveis (1,2) da variável **"sexo"** procede-se da seguinte forma:

a) Clica-se em **"Labels"**. Abre-se uma nova janela - **"Define Labels"**:
b) No espaço **"Value"**, digita-se "**1**";
c) No espaço **"Value Label"**, digita-se a denominação desejada, no caso, masculino;
d) Clica-se em **"ADD"**;
e) Procede-se da mesma forma para os demais níveis de categorização: digita-se "**2**" para **"Value"** e "feminino" para **"Value Label"**, seguindo-se por **"ADD"**
f) Clica-se em **"Continue"**;
g) Clica-se em **"OK"**.

OBSERVAÇÃO:

A manipulação do BANCO DE DADOS nos permite:

- Criar e recodificar variáveis;
- realizar análise de dados através de estatísticas descritivas, gráficos, etc;
- selecionar casos para análise, repetir a análise para grupos de casos diferentes.

É importante dar-se ao arquivo o nome mais claro possível para facilitar sua localização e acesso. Os arquivos de dados são do tipo " ***. sav** "

RECOMENDAÇÃO: A primeira coluna da matriz deve corresponder ao número do questionário, ou número do caso, ou ainda código do registro, o que é necessário para facilitar a localização de informações após serem identificados possíveis equívocos de digitação.

1.4- COMO ACESSAR UM BANCO DE DADOS JÁ EXISTENTE

Para acessar um banco de dados já existente, procede-se da seguinte maneira:

a) Iniciar o programa **SPSS** (clicar 2 vezes sobre o ícone);
b) Ir na opção **"File", "Open", "Data"** , abrir o arquivo que se deseja. Neste manual usaremos como exemplo o arquivo chamado **"Employee data.sav"** que se encontra disponível junto com o programa SPSS.

2. INTRODUÇÃO AOS MÉTODOS ESTATÍSTICOS

2.1- TIPOS DE VARIÁVEIS: Definição

Do ponto de vista estatístico, para decidirmos qual a análise estatística apropriada devemos distinguir entre dois tipos básicos de variáveis:

Variáveis quantitativas que são aquelas que podem ser mensuradas através de escalas quantitativas, isto é, escalas que tem unidades de medida. **Ex.: Renda Familiar** (medida em R$ ou em salários mínimos); **Idade** (medida em anos, ou meses); **Faturamento de uma Empresa** (R$, US$); **Nº de Empregados** (Nº), **Peso** (em Kg), **Altura** (em cm)...

Variáveis qualitativas ou categóricas que são as variáveis medidas originalmente em categorias. Ex: **Sexo, Profissão, Religião, Município, Região, ...**

2.2- DESCRIÇÃO E EXPLORAÇÃO DE DADOS

O objetivo básico deste procedimento é de introduzir técnicas que permitam organizar, resumir e apresentar resultados, de tal forma que possam ser interpretados de acordo com os objetivos da pesquisa e o tipo de variável.

Um primeiro passo para analisar qualquer banco de dados é analisar uma por uma das variáveis (o que será denominado de **análise univariada**). Se as variáveis são quantitativas usamos estatísticas descritivas (média, desvio padrão, valor mínimo, valor máximo). Se as variáveis são qualitativas usamos tabelas de freqüência, gráficos, por exemplo de pizza ou barra.

OBSERVAÇÃO:
Não podemos calcular média, variância ou desvio-padrão de variáveis qualitativas ou variáveis categóricas

2.3- CATEGORIZAÇÃO DE VARIÁVEIS

2.3.1- COMO CATEGORIZAR UMA VARIÁVEL QUANTITATIVA

Para categorizar variáveis quantitativas procede-se da seguinte forma:

a) Clicar na opção **"Transform"**, **"Recode"**, **"Into Different Variables"**;
b) Localizar na janela à esquerda a variável a ser categorizada;
c) Após selecionar a variável, clicar na →;
d) Digitar um novo nome para a nova variável **"Output Variable"** e clica-se em **"Change"**;
e) Clicar em **"Old and New Values"**, aparece uma janela denominada **"Recode Into Different Variables: Old and New Values";**
f) Clicar em **"Range" (Lowest Through...)** e digitar o primeiro ponto de corte da variável quantitativa, isto é o limite superior da primeira classe;
g) Na opção **"New Value"**, digita-se **1**;
h) Clica-se em **"ADD"**;
i) Assinala-se novamente **"Range"** (o intervalo entre mínimo e máximo), colocando-se os valores correspondentes: um dígito superior ao limite superior do intervalo anterior até **(Through)** o próximo limite de intervalo;
j) Na opção **"New Value"**, digita-se **2**;
k) Clica-se em **"ADD"**;
l) Assinala-se **"Range"**, colocando-se os seguintes valores: um dígito superior ao limite superior do intervalo anterior até **(Through)** o próximo intervalo;
m) Na opção **"New Value"**, digita-se **3**, e assim sucessivamente até esgotar a categorização desejada;
n) Clica-se em **"ADD"**;
o) Para o último intervalo, clicar em **"Range" (Through Highest)** e digitar o valor mínimo obtido para a última categoria;
p) Na opção **"New Value"**, digita-se o código do último intervalo;
q) Clica-se em **"ADD"**; **"Continue"** , **"OK".**

2.3.2- COMO CATEGORIZAR A VARIÁVEL PELO USO DOS QUARTIS

Os quartis são pontos de corte na escala da variável de tal forma que cada grupo formado, a partir destes pontos de corte, terá 25% do tamanho total da amostra.

Os passos necessários para categorizar uma variável utilizando-se os **"quartis"** são os seguintes: Inicialmente calcula-se os quartis da variável em questão, neste caso Salário (**Salary**):

a) Clica-se em **"Statistic"**, **"Summarize"**, **"Frequencies"** ;
b) Localiza-se a variável que se deseja categorizar na janela esquerda;
c) Após ter selecionado a variável, clica-se na →;
d) Retira-se a opção de **"Display Frequency Tables"**, a fim de que não venham listados a totalidade de casos da variável (no estudo em pauta o número é de 474 casos);
e) Clica-se no botão **"Statistics"**;
f) Assinala-se apenas **"Quartis"**;
g) Clica-se em **"Continue"**; **"OK"**.

RESULTADOS:

Frequencies

Statistics

Current Salary

N	Valid	474
	Missing	0
Percentiles	25	$24,000.00
	50	$28,875.00
	75	$37,162.50

CONCLUSÃO:

O primeiro intervalo obtido irá do valor mínimo apresentado no Banco de Dados (no caso em estudo 15.750) até o valor do 1º quartil = 24.000,00.

O segundo intervalo irá do valor imediatamente superior ao 1º quartil (24.001,00) até o valor do 2º quartil = 28.875,00 .

O terceiro intervalo irá do ponto imediatamente superior ao 2° quartil (28.876,00) até o valor do 3° quartil = 37.162,50 .

O quarto intervalo vai do ponto imediatamente superior ao 3º quartil (37.163,00) até o valor máximo existente (highest) no Banco de Dados.

Pode-se, então, dar continuidade ao procedimento relativo à categorização da variável **"Salary"**, usando os limites dados pelos respectivos quartis acima identificados. Sendo assim, procede-se da seguinte forma:

a) Clicar na opção **"Transform"**, **"Recode"**, **"Into Different Variables"**;
b) Localizar, na relação à esquerda, a variável a ser categorizada (**salary**);
c) Após ter selecionado a variável clicar na →;
d) Digitar um novo nome para a variável de saída (**Output Variable**) - por exemplo SALREC - e clicar em **"Change"**;
e) Clicar em **"Old and New Values"**;
f) Clicar em **"Range (lowest through)"** e digitar o valor obtido para o primeiro quartil, no caso 24000,00;
g) Na opção **"New Value"**, digita-se **1**;
h) Clica-se em **"ADD"**;
i) Assinala-se **"Range"**, colocando-se os seguintes valores: 24.001,00 até (Through) o segundo quartil 28.875,00;
j) Na opção **"New Value"**, digita-se **2**;
k) Clica-se em **"ADD"**;
l) Assinala-se **"Range"**, colocando-se os seguintes valores 28.876,00 até (Through) o terceiro quartil 37.162,50;
m) Na opção **"New Value"**, digita-se **3**;
n) Clica-se em **"ADD"**;
o) Clica-se em **"Range" (Through Highest)** e digitar o valor obtido para o quarto quartil, no caso 37.163,00;
p) Na opção **"New Value"**, digita-se **4**;
q) Clica-se em **"ADD"**;
r) Clica-se em **"Continue"**;
s) Clica-se em **"OK"**.

O resultado deste conjunto de procedimentos é a obtenção da nova variável **"SALREC"** que é a variável **"Salary"** categorizada, sendo esta automaticamente incluída no banco de dados que estamos utilizando (**Employee data. sav / Arquivo Data**).

2.3.3- COMO DAR NOME AOS NÍVEIS DE UMA VARIÁVEL

No banco de dados, clica-se **2 vezes** sobre a nova variável **"Salrec"**, obtendo-se uma nova janela **"Define Variable"**.

Para dar o nome aos níveis (1, 2, 3 e 4) em que foi categorizada a nova variável **"Salrec"** procede-se da seguinte forma:

a) Clica-se em **"Label"**. Abre-se uma nova janela - **"Define Labels"**: (nome da variável) **"Salrec"**;
b) No espaço **"Value"**, digita-se **1**;
c) No espaço **"Value Label"**, digita-se a denominação desejada, por exemplo, salário baixo;
d) Clica-se em **"ADD"**;
e) Procede-se da mesma forma para os demais níveis de categorização: 2, 3 e 4;
f) Clica-se em **"Continue"**;
g) Clica-se em **"OK"**.

2.4- COMO CRIAR UMA VARIÁVEL A PARTIR DE UMA DATA

Para criar uma variável, p.ex. Idade, a partir do ano de nascimento, ou outra variável data, utilizamos a função XDATE.YEAR (datevalue) a partir da variável data de nascimento, que no exemplo é **BDATE**:

a) Selecionar **"Transform", "Compute"**;
b) Em **"Target Variable"** digite o nome da nova variável, por exemplo AGE;
c) Na janela **"Numeric Expression"** digite 2001-;
d) Na janela **"Functions"** selecionar a opção XDATE.YEAR(datevalue);
e) Clica-se na ↑;
f) Localizar na janela abaixo de **"Target Variable"** a variável desejada,
g) Após ter selecionado a variável (neste caso, **bdate**), clica-se na →, (a variável selecionada deve ficar entre os parênteses);
h) Clica-se em **"OK"**.

2.5- COMO CRIAR UMA VARIÁVEL ATRAVÉS DA COMBINAÇÃO DE OUTRAS DUAS

Para criar uma variável a partir da combinação de outras duas, como por exemplo, combinar a variável sexo (gender) e a variável raça (minority) utilizaremos o seguinte procedimento para criar a variável GENRACE.

Sabendo que a variável GENDER é categorizada da seguinte forma:

0-Male e 1-Female

e a variável MINORITY é categorizada da seguinte forma:

0-No (White) e 1-Yes(Minority)

pode-se criar a variável GENRACE com as seguintes categorias:

1-Minority Male
2-Minority Female
3-White Male
4-White Female

Então procede-se da seguinte forma:

a) Seleciona-se **"Transform", "Compute", "Into Different Variables"**;
b) Em **"Target Variable"** digita-se o nome da nova variável, por exemplo GENRACE;
c) Na janela **"Numeric Expression"** digita-se **1**;
d) Clica-se em **"if"**;
e) Selecione a opção **"Include if case satisfies condition"**;
f) Localizar na janela abaixo de **"Include if case satisfies condition"** a variável desejada,
g) Após ter selecionado a variável (neste caso, **gender**), clica-se na →;
h) Digita-se **=1** & na janela ao lado da variável **gender**;
i) Selecionar na janela ao lado a variável **minority** e clica-se →;
j) Na janela ao lado da variável **minority** digitar **=0**;
k) Após esse procedimento a expressão na janela deve ser a seguinte: **gender=1&minority=0;**
l) Clica-se em **"Continue"** e **"OK"** (a variável GENRACE aparecerá no final do banco de dados),
m) Para criar as demais categorias da variável GENRACE procede-se de maneira análoga, alterando o código na janela **"Numeric Expression"** para 2, 3 e 4 e a expressão da janela **"Include if case satisfies condition".**

Resultado das janelas:

"Numeric Expression"	**"Include if case satisfies condition"**
1	Gender=1&minority=0
2	Gender=1&minority=1
3	Gender=0&minority=0
4	Gender=0&minority=1

3. ANÁLISE DE DADOS: ANÁLISE UNIVARIADA

3.1- VARIÁVEIS QUANTITATIVAS

3.1.1- COMO OBTER AS ESTATÍSTICAS DESCRITIVAS

Para calcular as principais estatísticas descritivas procede-se da seguinte forma:

a) Clica-se em **"Statistics"**, **"Summarize"** , **"Frequencies"**;
b) Localizar na janela a esquerda a variável de interesse;
c) Após ter selecionado a variável desejada, clica-se na →;
d) Clica-se no botão **"Statistics"**, e assinalam-se as opções desejadas;
e) Clica-se em **"Continue"**;
f) Clica-se em **"OK"**;
g) Obtêm-se os resultados da análise estatística solicitada que serão adicionados em um arquivo de resultados (**OUTPUT**), que, ao salvá-lo, dá origem a um arquivo do tipo **"*.spo"** (**SPSS output**).

EXEMPLO:

Frequencies

Statistics

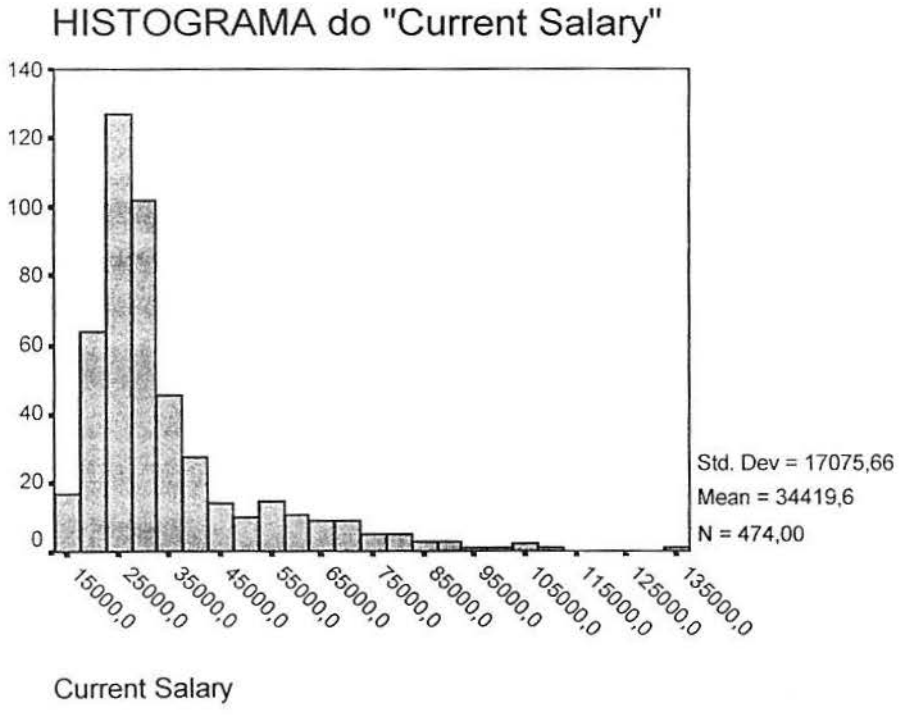
Current Salary

N	Valid	474
	Missing	0
Mean		$34,419.57
Std. Deviation		$17,075.66
Variance		$291578214
Minimum		$15,750
Maximum		$135,000

3.1.2 COMO OBTER UM HISTOGRAMA

a) Clica-se em **"Graphs"**, **"Histogram"**;
b) Localizar na janela a variável desejada;
c) Após ter selecionado a variável (neste caso, **salary**), clica-se na $\rightarrow$;
d) Pode-se clicar na opção **"Titles"** para dar um título ao histograma;
e) Clica-se em **"OK"**

EXEMPLO: Histograma da variável "**Current Salary**"



3.2- VARIÁVEIS QUALITATIVAS OU CATEGÓRICAS

3.2.1- COMO OBTER A DISTRIBUIÇÃO DE FREQÜÊNCIAS

Para calcular as freqüências procede-se da seguinte forma:

a) Clica-se em **"Statistics"**, **"Summarize"** , **"Frequencies"**;

b) Seleciona-se a variável desejada, clica-se na →;

c) Seleciona-se a opção: **"Display frequency tables";**

d) Clica-se em **"OK".**

3.2.2 COMO OBTER GRÁFICOS

Para se obter os vários tipos de gráficos estatísticos disponíveis no programa procede-se da seguinte forma:

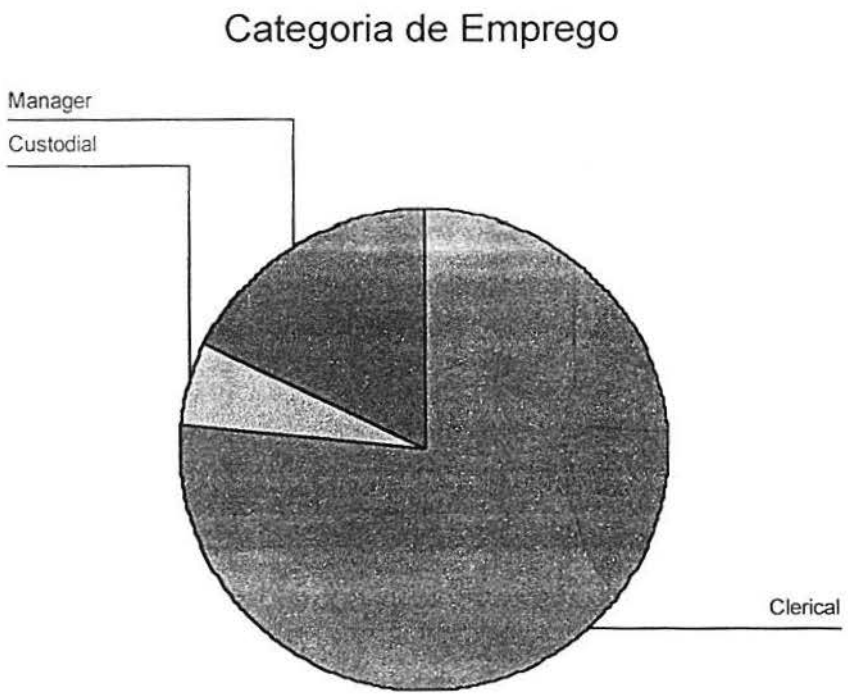
a) Clica-se em **"Graphs"**, seleciona-se o gráfico desejado, que ao salvá-lo, dá origem a um arquivo do tipo **" *.cht (Chart)"** (arquivo de gráficos).

OBSERVAÇÃO:
Como estamos trabalhando com variáveis categóricas, o adequado seria fazer gráfico de Pizza, de Colunas...

EXEMPLO: Gráfico de Pizza para a variável **"Jobcat"**

a) Clica-se em **"Graphs"**, seleciona-se o gráfico **"Pie"**;
b) Seleciona-se a opção **"Summaries for groups of cases"** e clica-se em **"Define"**;
c) Na opção **"Define Slices by"** selecione a variável desejada, nesse exemplo foi selecionada a variável **"Jobcat"**.

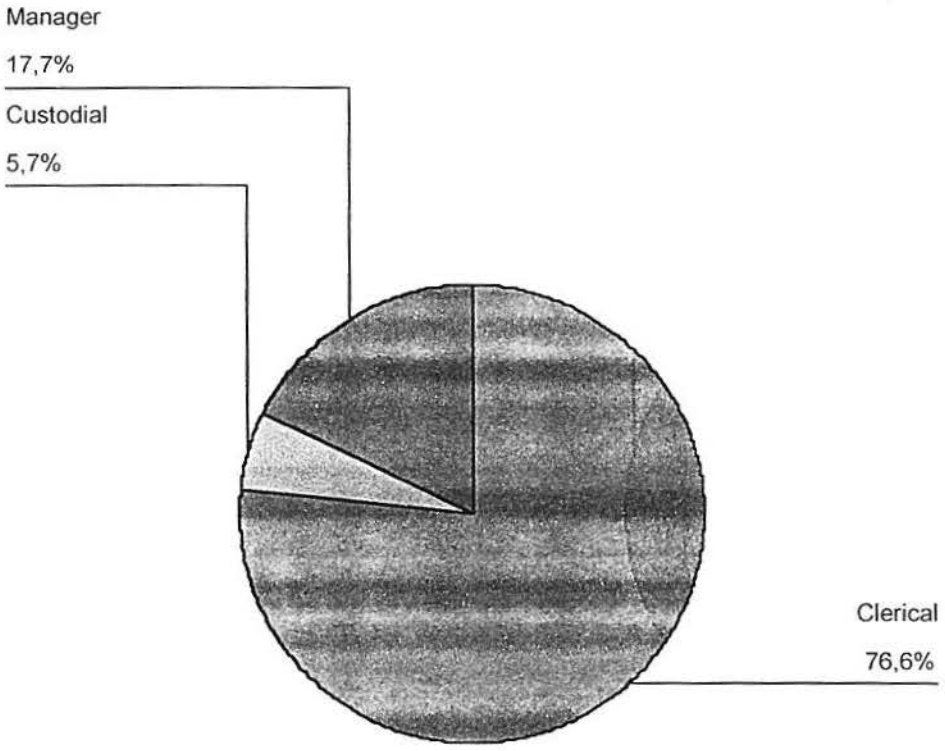
RESULTADO:



Para colocar no gráfico o valor percentual de cada categoria:

a) Clica-se duas vezes no gráfico;
b) Abrirá a janela de edição de gráficos (**SPSS Chart Editor**);
c) Nesta janela, clicar em **chart / options**;
d) Selecionar a opção **percents**;
e) Clica-se em **"OK".**

RESULTADO:



4. ANÁLISE DE DADOS: ANÁLISE BIVARIADA

Para realizar uma análise estatística **bivariada**, ou seja, fazer uma análise da relação entre duas variáveis é preciso utilizar testes estatísticos e/ou gráficos adequados:

a) **Para duas variáveis quantitativas:**
- Gráfico - **"Scatterplot"** de X e Y
- Coeficiente de Correlação de Pearson
- Análise de Regressão Simples

b) **Para duas variáveis categóricas (qualitativas)**
- Medir a associação pelo teste Qui-Quadrado e a Análise dos Resíduos
- Análise de Correspondência
- Gráfico de colunas por estratos da segunda variável

c) **Para uma variável quantitativa e uma qualitativa**
- Categoriza-se a variável quantitativa e procede-se como no item anterior
- Gráfico **"Box-Plot"**, para cada estrato ou categoria da variável qualitativa

4.1 VARIÁVEIS QUANTITATIVAS X QUANTITATIVAS

4.1.1 COMO CALCULAR A CORRELAÇÃO ENTRE DUAS VARIÁVEIS QUANTITATIVAS

Para medir o grau de correlação entre duas variáveis quantitativas, estão disponíveis no programa alguns coeficientes de correlação, entre os quais, o Coeficiente de Correlação de Pearson, que é o mais utilizado para variáveis quantitativas. Os cálculos dos coeficientes são realizados conjuntamente com a aplicação de um teste estatístico que testa se as variáveis são ou não significantemente correlacionadas.

4.1.2 - COMO OBTER GRÁFICO SCATTERPLOT

Para testar se existe correlação significativa entre duas variáveis quantitativas, o gráfico **Scatterplot** deve ser uma etapa preliminar ao cálculo do Coeficiente de Correlação. Neste gráfico, cada ponto representa um par observado de valores das duas variáveis (X,Y). Através deste gráfico podemos visualizar empiricamente a relação entre as variáveis.

Para se obter o gráfico **Scatterplot** (gráfico tipo XY) para a visualização correspondente ao comportamento conjunto das duas variáveis em estudo, procede-se da seguinte maneira:

a) Clica-se em **"Graphs"**;

b) Clica-se em **"Scatter"**, abre uma janela **"Scatterplot"**, onde se seleciona o item **"Simple"**;

c) Clica-se em **"Define"**. São apresentadas as variáveis do Banco de Dados, escolhem-se as variáveis pertinentes ao estudo, no caso, por exemplo, **"Educ"** e **"Salary"**;

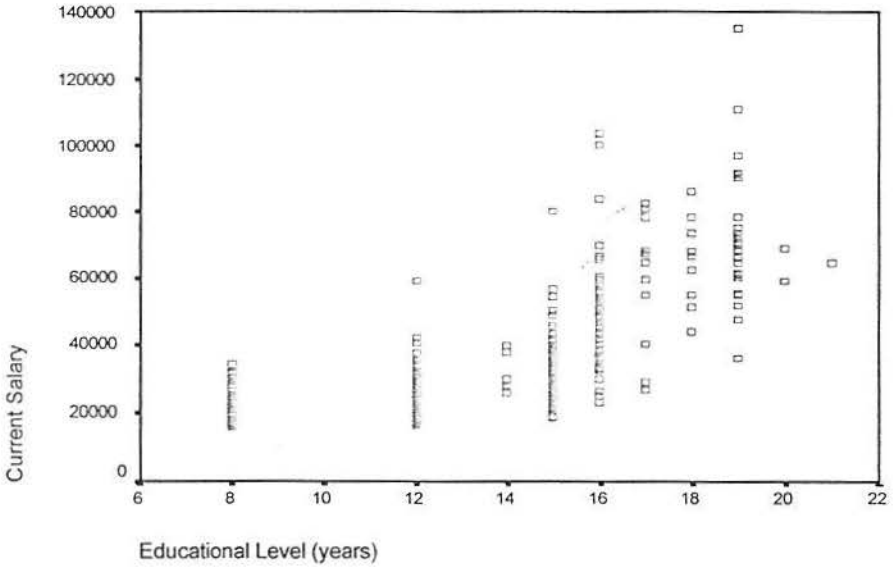
d) Define-se qual a variável dependente (Y) no caso **"Salary"**, clica-se na flecha pertinente e a variável independente (X), no caso **"Educ"**, clicando-se na flecha correspondente;

e) Clica-se em **"OK"**, obtendo-se um gráfico gerado na saída **"Chart"**, extensão **"* . cht"**, (arquivo de gráfico).

RESULTADO:

Gráfico **"Chart Carroussel 1, Scatter of Salary Educ"**

Graph



4.1.3 COMO OBTER O COEFICIENTE DE CORRELAÇÃO DE PEARSON

Para calcular o coeficiente de Correlação de Pearson procede-se da seguinte maneira:

a) Clica-se em **"Statistics"**, **"Correlate"**, **"Bivariate"**, abre-se uma janela **"Bivariate Correlations";**

b) Selecionam-se as variáveis (no caso **"Educ"** e **"Salary"**), clica-se na →;

c) Seleciona-se a estatística desejada, no caso, Pearson;

d) Clica-se em **"OK"**;

OBSERVAÇÃO:

O coeficiente de Correlação Linear de Pearson (r) é uma medida que varia de –1 a +1.

O coeficiente fornece informação através do sinal:

- Se r for positivo, existe uma relação direta entre as variáveis (valores altos de uma variável correspondem a valores altos de outra variável);
- Se r for negativo, existe uma relação inversa entre as variáveis (valores altos de uma variável correspondem a valores baixos de outra variável);
- Se r for nulo, significa que não existe correlação linear.

RESULTADO:

Correlations

Correlations

		Educational Level (years)	Current Salary
Educational Level (years)	Pearson Correlation	1,000	,661**
	Sig. (2-tailed)	.	,000
	N	474	474
Current Salary	Pearson Correlation	,661**	1,000
	Sig. (2-tailed)	,000	.
	N	474	474

**. Correlation is significant at the 0.01 level (2-tailed).

CONCLUSÃO:

Analisando-se os dados obtidos, rejeita-se H_0 (hipótese nula) de que **não há correlação** entre **"Educ"** e **"Salary"**, uma vez que o valor de **p** ("sig") é menor do que 0,05 ($p<0,001$ no caso em estudo) e aceita-se a hipótese alternativa de que há correlação entre as variáveis em estudo.

Este resultado confirma a configuração do gráfico Scatterplot, mostrando que a medida que o nível de escolaridade aumenta, o salário também tende a aumentar.

As hipóteses do Coeficiente de Correlação de Pearson são:

- **Hipótese Nula (H_0): $\rho = 0$ (não existe correlação entre as variáveis)**
- **Hipótese Alternativa (H_1): $\rho \neq 0$ (existe correlação significante)**

4.1.4 - COMO FAZER REGRESSÃO LINEAR SIMPLES

O modelo de regressão linear utiliza-se quando queremos ajustar uma equação linear entre duas variáveis quantitativas com a finalidade, por exemplo, de predizer o valor de uma variável em função de outra (Y em função de X). Para aplicar o modelo de regressão devemos definir a *priori* qual variável é a variável explicativa, ou independente (X) e qual é a variável explicada ou dependente (Y). A relação entre as variáveis deve ser explicada teoricamente dentro da área de estudo.

Para obter a reta de regressão entre duas variáveis, por exemplo **"Salary"** e **"Salbeqin"**, procede-se da seguinte forma:

a) Clica-se **"Statistics", Regression", "Linear"**;
b) Define-se a variável independente **"Salbeqin"**, e a variável dependente **"Salary"**;
c) Seleciona-se **"Method Enter"**;
d) Na opção **"Statistics"**, selecione **"Casewise Diagnostics"** para mostrar a tabela com os valores residuais atípicos;
e) Na opção **"Save"**, selecione **"Predicted Values"** / **"Unstandart"**, para mostrar e salvar no banco de dados os valores preditos pela reta ajustada;
f) Clica-se **"OK"**.

RESULTADO:

Regression

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	Beginning Salary[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: Current Salary

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,880[a]	,775	,774	$8,115.36

a. Predictors: (Constant), Beginning Salary

b. Dependent Variable: Current Salary

INTERPRETAÇÃO: Como o coeficiente de determinação, R^2 (R square) é igual a 0,775, pode-se afirmar que 77,5% da variação da variável salário atual (Salary) é explicada pela variável salário inicial (salbegin).

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1,07E+11	1	1,07E+11	1622,118	,000[a]
	Residual	3,11E+10	472	65858997		
	Total	1,38E+11	473			

a. Predictors: (Constant), Beginning Salary

b. Dependent Variable: Current Salary

INTERPRETAÇÃO: A tabela acima (ANOVA) só tem significado se a regressão é múltipla (mais de uma variável independente) e serve para verificar se pelo menos uma das variáveis explicativas é significante. Neste exemplo, fizemos regressão simples, logo as resultados da tabela ANOVA e da tabela COEFFICIENTS são os mesmos.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1928,206	888,680		2,170	,031
	Beginning Salary	1,909	,047	,880	40,276	,000

a. Dependent Variable: Current Salary

INTERPRETAÇÃO: A equação de regressão linear é **Y = a X + b**, onde o coeficiente linear da reta é **a** = 1928,206 e o coeficiente angular é **b** = 1,909. Como o "sig" de b é menor que 0,05 (nível de significância), rejeitamos a hipótese nula de que β **= 0**, logo a variável é significante, isto é, explica a variável dependente. A partir desta equação podemos estimar (predizer) os valores da variável dependente **(salary)**.

As hipóteses do Coeficiente Angular β são:

- **Hipótese Nula (H_0): $\beta = 0$**
- **Hipótese Alternativa (H_1): $\beta \neq 0$**

Casewise Diagnostics[a]

Case Number	Std. Residual	Current Salary	Predicted Value	Residual
18	6,074	$103,750	$54,457.17	$49,292.83
103	3,478	$97,000	$68,778.04	$28,221.96
106	4,068	$91,250	$58,237.88	$33,012.12
160	-3,279	$66,000	$92,607.97	-$26,607.97
205	-4,365	$66,750	$102,174.32	-$35,424.32
218	5,914	$80,000	$32,002.04	$47,997.96
274	4,965	$83,750	$43,458.74	$40,291.26
449	3,270	$70,000	$43,458.74	$26,541.26
454	3,577	$90,625	$61,598.51	$29,026.49

a. Dependent Variable: Current Salary

INTERPRETAÇÃO: A tabela **"Casewise Diagnostics"** apresenta os casos em que os valores residuais são atípicos, isto é valores dos resíduos padronizados maiores do que 3 em valor absoluto, mostrando que a diferença entre o valor observado e o valor predito é relativamente grande e isto pode ser um sintoma de que o modelo não está bem ajustado.

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	$19,113.25	$154646.0	$34,419.57	$15,028.59	474
Residual	-$35,424.32	$49,292.83	$.00	$8,106.77	474
Std. Predicted Value	-1,018	8,000	,000	1,000	474
Std. Residual	-4,365	6,074	,000	,999	474

a. Dependent Variable: Current Salary

INTERPRETAÇÃO: Esta tabela mostra um resumo das estatísticas descritivas dos principais resultados da Análise de Regressão.

OBSERVAÇÃO: Os valores de Y determinados por essa equação aparecem na última coluna do banco de dados, pois selecionamos a opção **"Save" / "Predicted Values" / "Unstandart".** Essa coluna tem o nome de **pre-1 (Unstandardized Predicted Value).** Sugere-se mudar do nome da variável (label) para que fiquem identificados as colunas das várias análises que porventura sejam realizadas.

4.2 VARIÁVEIS CATEGÓRICAS X CATEGÓRICAS

4.2.1 - COMO VERIFICAR A EXISTÊNCIA DE ASSOCIAÇÃO ENTRE VARIÁVEIS CATEGÓRICAS: Teste Qui - Quadrado

Para obter a tabela de contingência e verificar se existe associação entre **"Genrace"** e **"Salrec"** (salário em categorias), procede-se da seguinte forma:

a) Clica-se em **"Statistics"**, **"Summarize"**, **"Crosstabs";**
b) Define-se a variável da linha **"Row - Genrace"**;
c) Define-se a variável da coluna **"Column - Salrec"**;
d) Clica-se em **"Statistics";**
e) Escolhe-se o tratamento estatístico desejado, no caso, **"Chi-Square"**;
f) Clica-se em **"Continue"**;
g) Clica-se em **"Cell"**, veremos a janela **"Crosstabs : Cell Display"**;
h) Assinala-se as opções **"Observed"**; etc, de acordo com o desejado;
i) Clica-se em **"Continue"**;
j) Clica-se em **"OK".**

Se desejarmos obter o valor esperado de cada casela na tabela, abre-se a janela **"Crosstabs : Cell Display"** e assinala-se também a opção **"Expected".**

RESULTADOS:

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
GENRACE * salrec	474	100,0%	0	,0%	474	100,0%

GENRACE * salrec Crosstabulation

			salrec				
			salário baixo	sálário médio baixo	salário médio alto	salário alto	Total
GENRACE	Minority Male	Count	8	23	25	8	64
		Expected Count	16,2	15,8	16,1	15,9	64,0
	Minority Female	Count	22	16	2	0	40
		Expected Count	10,1	9,9	10,0	10,0	40,0
	White Male	Count	8	34	57	95	194
		Expected Count	49,1	47,9	48,7	48,3	194,0
	White Female	Count	82	44	35	15	176
		Expected Count	44,6	43,4	44,2	43,8	176,0
Total		Count	120	117	119	118	474
		Expected Count	120,0	117,0	119,0	118,0	474,0

A leitura das caselas na 1ª linha (count) informa a freqüência bruta e a 2ª linha (expect) corresponde ao valor esperado, que é o número de pessoas que seria esperado caso não houvesse nenhuma associação entre as variáveis em estudo, isto é, se as variáveis fossem independentes.

OBSERVAÇÃO: Valor Esperado sob hipótese de independência para o Teste Qui-Quadrado, para cada casela ij é obtido com a fórmula a seguir:

$$\frac{(TL_i \times TC_j)}{TG}$$

TL - total da linha i
TC - total da coluna j
TG - total geral

Quando se deseja obter o percentual correspondente a linha (**Row**) procede-se como anteriormente; só que em **"Cell"**, abre-se a janela **"Crosstabs": "Cell Display"** e assinala-se a opção **"Row"** em **"Percentages"**, obtendo-se a seguinte tabela:

RESULTADOS:

GENRACE * salrec Crosstabulation

			salrec				
			salário baixo	sálário médio baixo	salário médio alto	salário alto	Total
GENRACE	Minority Male	Count	8	23	25	8	64
		Expected Count	16,2	15,8	16,1	15,9	64,0
		% within GENRACE	12,5%	35,9%	39,1%	12,5%	100,0%
	Minority Female	Count	22	16	2	0	40
		Expected Count	10,1	9,9	10,0	10,0	40,0
		% within GENRACE	55,0%	40,0%	5,0%	,0%	100,0%
	White Male	Count	8	34	57	95	194
		Expected Count	49,1	47,9	48,7	48,3	194,0
		% within GENRACE	4,1%	17,5%	29,4%	49,0%	100,0%
	White Female	Count	82	44	35	15	176
		Expected Count	44,6	43,4	44,2	43,8	176,0
		% within GENRACE	46,6%	25,0%	19,9%	8,5%	100,0%
Total		Count	120	117	119	118	474
		Expected Count	120,0	117,0	119,0	118,0	474,0
		% within GENRACE	25,3%	24,7%	25,1%	24,9%	100,0%

Desejando-se os percentuais relativos a coluna (**Columm**) e ao total (**Total**) procede-se da mesma forma que para o cálculo relativo à percentagem da linha. Neste caso, cada casela poderia ter até 5 valores descritos a seguir:

1º linha: valor observado;
2º linha: valor esperado;
3º linha: percentual da linha;
4º linha: percentual da coluna;
5º linha: percentual total.

OBSERVAÇÃO:

Sugere-se que num relatório final de pesquisa, selecione-se apenas o valor observado e um destes percentuais.

RESULTADO:

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	187,827[a]	9	,000
Likelihood Ratio	207,033	9	,000
Linear-by-Linear Association	7,562	1	,006
N of Valid Cases	474		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 9,87.

INTERPRETAÇÃO: Ao serem analisados os dados do exemplo acima e considerando que $p < 0,05$ (sig), rejeita-se a hipótese nula (H_0) de independência entre as variáveis. Sendo assim, conclui-se que há evidências de associação entre **"Genrace"** e **"Salary"**.

As hipótese do teste "Qui-Quadrado" (Chi-Square) são:

- **Hipótese Nula (H_0): As variáveis são independentes.**
- **Hipótese Alternativa (H_1): As variáveis são dependentes.**

4.2.2 COMO CALCULAR OS RESÍDUOS AJUSTADOS

Verificada a associação global entre as variáveis pode-se verificar se há associação local entre categorias, calculando-se os resíduos ajustados. O resíduo ajustado tem distribuição normal com média zero e desvio padrão igual a 1. Desta forma, caso o resíduo ajustado seja **maior que 1,96**, em valor absoluto, pode-se dizer que há evidências de associação significante entre as duas categorias (p. ex. homem branco e salário alto) naquela casela. Quanto maior for o resíduo ajustado, maior a associação entre as categorias.

Para obter os resíduos ajustados procede-se da seguinte maneira:

a) Seleciona-se **"Statistics"** , **"Summarize"**, **"Crosstabs"**;
b) Clica-se em **"Cells"**, abre-se a janela **"Crosstabs": "Cell Display"**;
c) Assinala-se a opção **"Observed"** e **"Adj. Residuals"**;
d) Clica-se em **"Continue"**;
e) Clica-se em **"OK"**.

RESULTADOS:

GENRACE * salrec Crosstabulation

			salrec				
			salário baixo	sálário médio baixo	salário médio alto	salário alto	Total
GENRACE	Minority Male	Count	8	23	25	8	64
		Adjusted Residual	-2,5	2,2	2,8	-2,5	
	Minority Female	Count	22	16	2	0	40
		Adjusted Residual	4,5	2,3	-3,1	-3,8	
	White Male	Count	8	34	57	95	194
		Adjusted Residual	-8,8	-3,0	1,8	10,1	
	White Female	Count	82	44	35	15	176
		Adjusted Residual	8,2	,1	-2,0	-6,3	
Total		Count	120	117	119	118	474

INTERPRETAÇÃO: A associação entre sexo (**gender)** e salário **salrec** (salário em categorias) já foi testada como sendo significativa. Agora a pergunta é: Quais categorias estão associadas localmente? Olhando os resíduos ajustados vemos que os maiores valores (positivos) indicam forte associação entre homem-branco e salário alto, bem como há forte associação entre mulher-branca e salário baixo. Há outras associações locais interessantes na tabela, identifique.

4.3- VARIÁVEIS QUANTITATIVAS X CATEGÓRICAS

Neste caso os tratamentos estatísticos possíveis são os mesmos utilizados para duas variáveis qualitativas, desde que as variáveis quantitativas sejam categorizadas, logo, procede-se da seguinte forma:

- Categoriza-se a variável quantitativa em classes apropriadas;
- Mede-se a associação aplicando-se o teste Qui-Quadrado e a Análise dos Resíduos;
- Também podemos utilizar gráficos de colunas por estratos da segunda variável e o gráfico **BOX-PLOT** por categorias da segunda variável para apresentação dos dados de forma descritiva, exploratória.

4.3.1 COMO FAZER O BOX-PLOT

a) Clicar em **"Graphs" / "Boxplot"**;
b) Selecione **"Simple" / "Summaries for groups of cases"**;
c) Clicar em **"Define"**;
d) Em **Variable** selecionar uma variável quantitativa (por exemplo, **Current salary**);
e) Em **Category Axis**, selecionar uma variável categórica (por exemplo, **Gender**);
f) Clicar em **"OK"**.

RESULTADO:

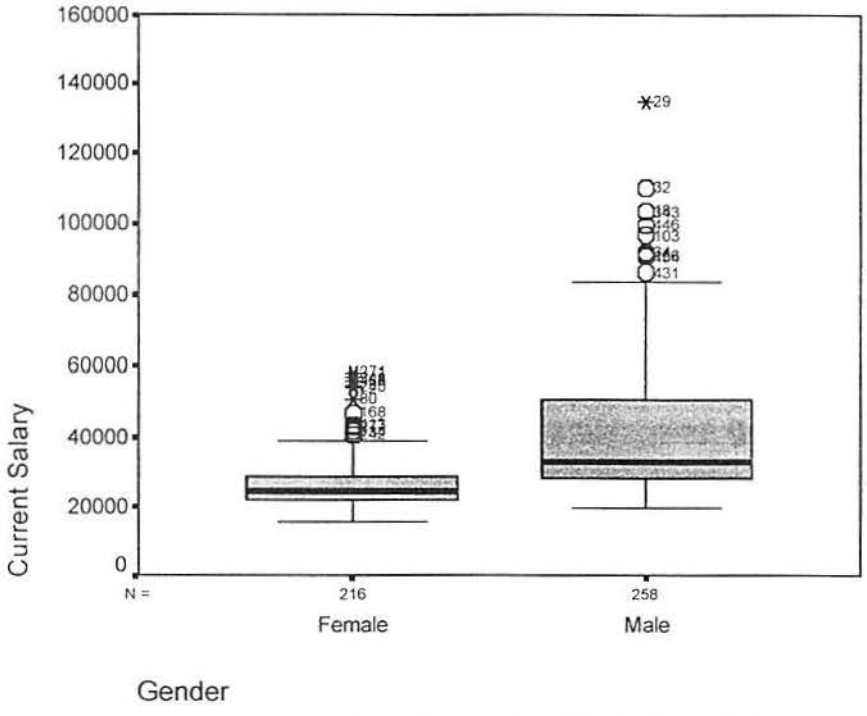
Explore
Gender

Case Processing Summary

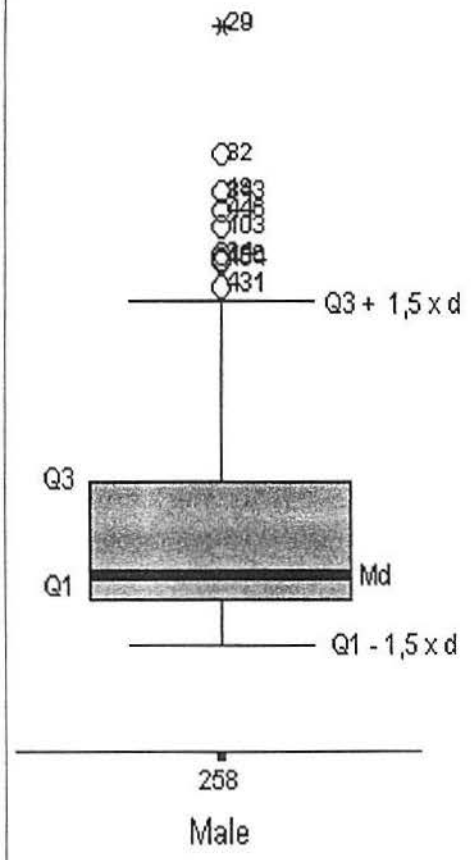
		Cases					
		Valid		Missing		Total	
	Gender	N	Percent	N	Percent	N	Percent
Current Salary	Female	216	100,0%	0	,0%	216	100,0%
	Male	258	100,0%	0	,0%	258	100,0%

INTERPRETAÇÃO: A tabela acima apresenta o número de casos válidos (**valid**), o número de não respostas (**missing**) e o número total das observações de cada categoria.

Current Salary



INTERPRETAÇÃO: Através do Box-plot pode-se ver como as variáveis estão distribuídas em relação à homogeneidade dos dados, valores de tendência central, valores máximos e mínimos e valores atípicos (se existirem). Quando a caixinha (box) é muito "pequena", significa que os dados são muito concentrados em torno da mediana, e se a caixinha for "grande", significa que os dados são mais heterogêneos.



LEGENDA:

Md: Mediana (linha horizontal escura dentro do "box")

Q1: Quartil inferior –1°quartil (limite inferior do "box")

Q3: Quartil superior – 3°quartil (limite superior do "box")

d: diferença interquartílica (d = Q3 – Q1)

o : outlier (valores acima de 1,5 x d)

x : outlier (valores acima de 3,0 x d)

5. ANÁLISE DE DADOS: COMPARAÇÃO DE MÉDIAS

5.1 - COMO COMPARAR MÉDIAS ENTRE DOIS GRUPOS: Teste "t" para Amostras Independentes.

O teste **"t"** é apropriado para comparar as médias de uma variável quantitativa entre dois grupos independentes.

EXEMPLO: Comparar a média de salários entre os sexos masculino e feminino numa empresa.

a) Sexo (masculino, feminino) - Dois grupos (variável que define os grupos).
b) Salário (variável resposta ou de teste).

Para a aplicação do teste **"t"** nesta situação procede-se da seguinte forma:

a) Clica-se em **"Statistics"**, **"Compare Means"**, **Independent Samples t test"**;
b) Clica-se sobre a variável de teste (**Test Variables**): **"Salary"** ou, conforme o caso em estudo, clica-se na variável correspondente;
c) Clica-se sobre a variável de grupo (Grouping Variable) **"Gender"**;
d) Clica-se sobre o botão: **"Define Group"**;
e) Abre-se uma janela, na qual se define qual a categoria correspondente ao **"Group 1"** (no caso masculino) - digitando-se o código da categoria atribuída quando da construção do Banco de Dados, nesse caso **0** e **"Group 2"** (no caso feminino) digitando-se o código **1.**
(*Observação* : No caso de se desejar confirmar os valores atribuídos às variáveis, abre-se a janela **"Utilities"** , **"Variables"**)
f) Clica-se em **"Continue"** e **"OK"** .

RESULTADO:

T-Test

Group Statistics

	Gender	N	Mean	Std. Deviation	Std. Error Mean
Current Salary	Male	258	$41,441.78	$19,499.21	$1,213.97
	Female	216	$26,031.92	$7,558.02	$514.26

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Current Salary	Equal variances assumed	119,67	,000	10,945	472	,000	$15,409.86	$1,407.91	$12,643.32	$18,176.40
	Equal variances not assumed			11,688	344,3	,000	$15,409.86	$1,318.40	$12,816.73	$18,003.00

INTERPRETAÇÃO: Ao serem analisados os dados do exemplo acima e considerando que p é menor que 0,05 (sig), rejeita-se a hipótese nula (H_0) de igualdade das médias dos dois grupos, logo, pode-se concluir que as médias da variável salário são significativamente diferentes entre os dois grupos de sexo.

O teste t de Igualdade de Médias a ser escolhido é o teste para o caso de variâncias desiguais (Unequal), pois o computador realiza previamente o teste de igualdade das variâncias (Teste Levene) e neste exemplo, como o valor de p para o teste Levene é menor do que 0,001, rejeita-se a hipótese de variâncias iguais, logo o teste t a ser utilizado é o que aparece na segunda linha.

As hipóteses do teste Levene de igualdade de variâncias são:

- **Hipótese Nula (H_0): A variância dos dois grupos são iguais.**
- **Hipótese Alternativa (H_1): A variância dos dois grupos são diferentes**

As hipóteses do teste "t" para igualdade de médias entre amostras Independentes são:

- **Hipótese Nula (H_0): A média dos dois grupos são iguais.**
- **Hipótese Alternativa (H_1): A média dos dois grupos são diferentes**

EXERCÍCIO: Verificar se há desigualdade entre a média de idade entre os homens e as mulheres no banco de dados em estudo.

Procedendo como o caso anterior, colocando em lugar de "**Salary**", a variável **"AGE"** obtêm-se o seguinte resultado:

Group Statistics

	Gender	N	Mean	Std. Deviation	Std. Error Mean
AGE	Male	257	44,1128	9,9310	,6195
	Female	216	45,3380	13,6604	,9295

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
AGE	Equal variances assumed	80,863	,000	-1,127	471	,260	-1,2251	1,0875	-3,3620	,9117
	Equal variances not assumed			-1,097	384,68	,273	-1,2251	1,1170	-3,4213	,9710

Observa-se que, neste caso, não há diferença significativa entre a média de idades entre homens e mulheres. O teste usado também neste caso, é o teste para o caso de variâncias desiguais (Unequal), pois como o valor de p para o teste Levene de igualdade de variâncias é menor do que 0,001, rejeitamos a hipótese de variâncias iguais.

5.2- COMO COMPARAR AS MÉDIAS DE TRÊS OU MAIS GRUPOS: Análise de Variância – "ANOVA" para um fator"

Para comparar a média de três ou mais grupos procede-se da seguinte maneira:

a) Clica-se em **"Statistics", "Compare Means", "One-Way Anova"**;
b) Assinala-se a variável dependente **"Dependent List",** clica-se sobre a seta correspondente (pode-se realizar mais de um teste incluindo outras variáveis na lista, o teste será repetido para cada variável incluída na lista), neste caso utilize **"Salary";**
c) Assinala-se a variável independente **"Factor"**, no caso **"Genrace",** clica-se na flecha correspondente;
d) Clica-se o botão **"Options".**
e) Clica-se na alternativa do quadro Statistics **"Descriptive"** e depois **"Continue";**
f) Clica-se no botão **"Post Hoc".** Aparece uma tela "**One-Way Anova: Post Hoc Multiple Comparisons",** assinala-se a alternativa **"T3 de Dunnet multiple range test"** ou outro teste conforme a escolha;
g) Clica-se em **"Continue",** e clica-se em **"OK".**

RESULTADOS:

Oneway

Descriptives

Current Salary

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Minority Male	64	$32,246.09	$13,059.88	$1,632.49	$28,983.83	$35,508.36	$19,650	$100,000
Minority Female	40	$23,062.50	$3,972.37	$628.09	$21,792.07	$24,332.93	$16,350	$35,100
White Male	194	$44,475.41	$20,330.66	$1,459.66	$41,596.49	$47,354.34	$21,300	$135,000
White Female	176	$26,706.79	$8,011.89	$603.92	$25,514.89	$27,898.69	$15,750	$58,125
Total	474	$34,419.57	$17,075.66	$784.31	$32,878.40	$35,960.73	$15,750	$135,000

ANOVA

Current Salary

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	3,55E+10	3	1,18E+10	54,405	,000
Within Groups	1,02E+11	470	2,18E+08		
Total	1,38E+11	473			

INTERPRETAÇÃO: Ao serem analisados os dados do exemplo acima e considerando que o valor p (**sig**) da ANOVA é menor que 0,05 ($p<0,001$), rejeita-se a hipótese nula (H_0) de igualdade dos quatros grupos, logo, pelo menos uma média de idade difere das demais. Verificada a diferença entre os grupos pode-se identificar qual(is) grupo(s) diferem significativamente, através de um teste de comparações múltiplas.

As hipóteses da Análise de Variância para um fator ("ANOVA – One-Way") são:

- **Hipótese Nula (H_0): A média de todos os grupos são iguais.**
- **Hipótese Alternativa (H_1): Pelo menos uma média difere das demais.**

Post Hoc Tests

Multiple Comparisons

Dependent Variable: Current Salary

Dunnett T3

(I) GENRACE	(J) GENRACE	Mean Difference (I-J)	Std. Error	Sig.
Minority Male	Minority Female	$9,183.59*	2974,607	,000
	White Male	-$12,229.32*	2127,413	,000
	White Female	$5,539.30*	2154,231	,012
Minority Female	Minority Male	-$9,183.59*	2974,607	,000
	White Male	-$21,412.91*	2562,772	,000
	White Female	-$3,644.29*	2585,077	,000
White Male	Minority Male	$12,229.32*	2127,413	,000
	Minority Female	$21,412.91*	2562,772	,000
	White Female	$17,768.62*	1536,302	,000
White Female	Minority Male	-$5,539.30*	2154,231	,012
	Minority Female	$3,644.29*	2585,077	,000
	White Male	-$17,768.62*	1536,302	,000

*. The mean difference is significant at the .05 level.

Como as variâncias da variável salário dos diferentes grupos é muito heterogênea utilizamos um teste de comparações múltiplas que não utiliza a suposição de igualdade de variâncias, este teste é, por exemplo, o teste T3 de Dunnet.

Para facilitar a leitura dos resultados, em vez de utilizar, para identificação dos grupos, a denominação G1, G2, G3, identificam-se os mesmos pelos Labels, o que é melhor para apresentação de relatório. Para tal, procede-se como descrito no capítulo 2.3.3- COMO DAR NOME AOS NÍVEIS DE UMA VARIÁVEL. É útil também que as estatísticas descritivas (Médias, Desvio Padrão, etc) sejam apresentadas, para isto clica-se na alternativa do quadro **"Statistics Descriptive"**, e na opção **"Display Labels"**, além dos comandos anteriores.

6. ANÁLISE DE DADOS: INTRODUÇÃO À ANÁLISE MULTIVARIADA

6.1 ANÁLISE DE COMPONENTES PRINCIPAIS

Análise de Componentes Principais é uma técnica de Estatística Multivariada utilizada para estudar o inter-relacionamento entre um conjunto de variáveis observadas com o objetivo de reduzir a dimensão original do problema a um número menor de variáveis sintéticas ou componentes, que podem ser utilizadas como índices e que preservam a maior parte da informação original contida nas variáveis em estudo. Para esta análise usaremos o banco de dados **"World95.sav"**

a) Clica-se em **"Statistics"**, **"Data Reduction"**, **"Factor"**;

b) Selecionam-se as variáveis desejadas e clica-se na →;

c) Em **"Extration"** selecione em **"Method"** a opção **"Principal Components"**;

d) Clica-se em **"Continue"**;

e) Em **"Rotation"** selecione **"Varimax"**;

f) Clica-se em **"Continue"**;

g) Em **"Scores"** selecione **"Save as variables"**;

h) Clica-se em **"Continue"**;

i) Em **"Option"** selecione **"Sorted by size"**;

j) Clica-se em **"Continue"**;

k) Clica-se em **"OK"**;

RESULTADOS:

Communalities

	Initial	Extraction
People living in cities (%)	1,000	,662
Average female life expectancy	1,000	,978
Average male life expectancy	1,000	,962
People who read (%)	1,000	,842
Population increase (% per year))	1,000	,904
Infant mortality (deaths per 1000 live births)	1,000	,943
Gross domestic product / capita	1,000	,734
Daily calorie intake	1,000	,717
Birth rate per 1000 people	1,000	,941
Death rate per 1000 people	1,000	,928
Fertility: average number of kids	1,000	,870

Extraction Method: Principal Component Analysis.

INTERPRETAÇÃO: Se as comunalidades estimadas forem muito baixas e se as variáveis não forem importantes, devem ser retiradas da análise. Neste caso, as comunalidades são em geral altas.

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	8,132	73,925	73,925	8,132	73,925	73,925	5,133	46,662	46,662
2	1,348	12,251	86,176	1,348	12,251	86,176	4,347	39,514	86,176
3	,663	6,026	92,202						
4	,289	2,631	94,833						
5	,215	1,954	96,787						
6	,182	1,655	98,442						
7	7,566E-02	,688	99,129						
8	4,789E-02	,435	99,565						
9	2,603E-02	,237	99,802						
10	1,505E-02	,137	99,938						
11	6,779E-03	6,163E-02	100,000						

Extraction Method: Principal Component Analysis.

INTERPRETAÇÃO: Para esta Análise de Componentes Principais, a percentagem da variância total das variáveis originais explicada pelos dois fatores ou componentes principais é 86,18%.

Component Matrix[a]

	Component	
	1	2
Average female life expectancy	,966	,212
Infant mortality (deaths per 1000 live births)	-,962	-,132
Average male life expectancy	,945	,261
Birth rate per 1000 people	-,945	,219
Fertility: average number of kids	-,919	,161
People who read (%)	,917	-8,53E-03
Daily calorie intake	,838	-,120
People living in cities (%)	,786	,210
Gross domestic product / capita	,772	-,371
Population increase (% per year))	-,727	,612
Death rate per 1000 people	-,596	-,757

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

INTERPRETAÇÃO: A matriz das componentes é a matriz das cargas fatoriais de cada variável no fator (ou componente) e mostra o peso da variável na composição do fator (ou componente).

Rotated Component Matrix[a]

	Component	
	1	2
Population increase (% per year))	-,950	-,026
Birth rate per 1000 people	-,852	-,465
Gross domestic product / capita	,823	,236
Fertility: average number of kids	-,793	-,491
Daily calorie intake	,706	,468
People who read (%)	,691	,604
Death rate per 1000 people	,058	-,962
Average male life expectancy	,533	,823
Average female life expectancy	,580	,801
Infant mortality (deaths per 1000 live births)	-,631	-,738
People living in cities (%)	,447	,680

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.

INTERPRETAÇÃO: A matriz rotada das componentes serve para ajudar a fazer uma melhor interpretação das componentes. Nem sempre é a melhor solução.

Nesta Análise de Componentes Principais usaremos apenas o primeiro fator (ou componente) que pode ser interpretado como um fator geral da condição sócio-econômica dos países. Este índice pode ser utilizado para análises posteriores.

7. MANIPULAÇÃO DE DADOS

7.1 SORT CASE

Uma necessidade comum na hora da manipulação dos dados é a ordenação dos casos segundo uma ou mais variáveis. Para fazer isto no *SPSS for Windows*, você deve usar o procedimento **Sort Cases** presente no menu **Data**.

Após clicar o menu **Data** opção **Sort Cases**, uma janela é aberta. Movemos para o quadro **Sort by** a variável segundo a qual o arquivo deve ser ordenado. Podemos mover para esse quadro mais do que uma variável. Nesse caso, o arquivo é ordenado, em primeiro lugar, pelos valores da primeira variável presente no quadro e, em segundo lugar, pela segunda variável presente no quadro; a segunda ordenação é feita para os valores comuns da primeira variável.

Podemos escolher também entre ordem crescente ou decrescente de ordenação para cada uma das variáveis. Isso é feito através do quadro **Sort Order** opções **Descending** (decrescente) ou **Ascending** (crescente).

Vamos fazer uma ordenação segundo sexo (ordem crescente) e idade (ordem decrescente). Para isso movemos a variável sexo para ao quadro **Sort Cases** e escolhemos a opção **Ascending** no quadro **Sort** Order. Movemos em seguida a variável idade para o quadro Sort Cases e escolhemos a opção **Descending** no quadro **Sort Order**. Agora, basta clicar **OK** para validar a ordenação.

Note que após a execução deste comando a posição dos indivíduos nas linha ficam completamente alteradas, pois o indivíduo na linha 1 do banco de dados pode não ser o primeiro

caso digitado. Para que esta informação não se perca é essencial que exista uma variável com o número do indivíduo.

7.2 SELECT CASES

Outra ferramenta importante do SPSS é o comando "Select Cases", utilizado quando é necessário fazer a seleção (temporária ou permanente) de parte do arquivo de dados. Digamos que estamos interessados em estudar um segmento específico da amostra. O SPSS possui várias formas de seleção de dados. Falaremos nessa seção de todas elas, mas discutiremos detalhadamente a mais usada de todas. Para maiores detalhes sobre as demais formas de seleção, recomenda-se que o leitor use o manual do *SPSS for Windows*.

Para fazer qualquer tipo de seleção, devemos clicar o menu **Data** opção **Select Cases**.

No quadro central **Select**, estão presentes cinco opções diferentes para seleção. Faremos a seguir uma breve descrição de cada uma delas:

- **All cases** – opção usada por *default*, utiliza todas as observações do banco de dados;
- **If condition is satisfied** – através dessa opção, podemos definir expressões condicionais para seleção de casos; estudaremos essa opção mais detalhadamente mais adiante;
- **Random sample of cases** – podemos selecionar uma porcentagem ou número exato de casos; a seleção é feita aleatoriamente;

- **Based on time or case range** – usamos essa opção quando estamos interessados em selecionar uma faixa específica de valores, por exemplo, os casos do número 100 ao 200; também utilizada para fazer seleções baseadas em datas;
- **User filter variable** – uma variável é escolhida no banco de dados e usada como filtro; todos os casos para os quais a variável filtro assume o valor 0 não serão selecionados.

Você tem duas opções para o tratamento dos casos que não serão selecionados. É através do quadro **Unselected Cases Are** que podemos fazer a escolha:

- **Filtered** – os casos (linhas) que não são selecionados não são incluídos nas análises posteriores, porém, permanecem na janela de dados; caso você mude de idéia e queira usar os casos não selecionados na mesma sessão do SPSS, basta "desligar" o filtro;
- **Deleted** – os casos (linhas) não selecionados são apagados da janela de dados; caso você mude de idéia e queira usar os casos não selecionados, você deverá ler novamente o arquivo de dados original. Neste caso deve-se tomar o cuidado de salvar o banco de dados com outro nome (**File..Save As**).

Suponha que estamos interessados em selecionar as pessoas que trabalham pelo menos 40 horas por semana e que têm até 20 horas de lazer. A função condicional para seleção nesse caso é dada por:

$$\text{trabalho} \geq 40 \ \& \ \text{lazer} \leq 20$$

Portanto, o tipo de seleção de dados que faremos deve possibilitar a criação de sentenças matemáticas lógicas

para seleção dos casos. Para isso, clicamos em **If condition is satisfied** e entramos no retângulo **If..**

Através da janela que é aberta, usamos o retângulo superior para escrever uma função lógica na qual a seleção vai ser baseada. Para a construção da função, podemos usar todas as variáveis que estão no quadro à esquerda e as funções disponíveis no quadro inferior direito.

Uma vez escrita a função que determina a regra de seleção dos casos, clique Continue e você voltará à janela anterior. No quadro inferior (**Unselected cases are**), vamos optar pelo modo **Filtered** (ou seja, os casos não selecionados permanecem na tela de dados, porém, não serão utilizados em análises futuras) e clicar **OK**.

Você pode perceber que, depois de feita a seleção, a janela de dados sofre algumas alterações. As linhas (casos) que não foram selecionadas apresentam uma listra no canto esquerdo da janela de dados. A barra localizada na parte inferior da janela apresenta a mensagem **Filter On**. Além disso, uma coluna de nome filter_$ é adicionada à janela de dados. Essa nova coluna apresenta valor 0 para as linhas que não foram selecionadas e valor 1 para as linhas que foram selecionadas.

Apesar de você conseguir ver os casos que não foram selecionados, qualquer análise efetuada daí para frente não leva em conta esses casos.

Podemos mudar de idéia e querer usar todas as observações para o cálculo das estatísticas. Temos duas maneiras de cancelar a seleção de casos, se a opção **Filtered** foi usada para efetuar a seleção. A primeira delas é ativar a opção **All Cases** da janela de seleção de casos (menu **Select**

Cases) e clicar **OK**. A Segunda maneira é deletar a coluna filter__$ da janela de dados.

7.3 SPLIT FILE

Vamos supor que, após uma série de análises, chegamos à conclusão de que o comportamento dos homens e das mulheres são completamente diferentes com relação às preferências para horas de lazer. Não faz sentido, portanto, apresentar a análise do questionário de opinião sobre lazer com os homens e mulheres juntos. No fundo, o que pretendemos fazer, daqui para frente, são duas análises idênticas, uma para cada sexo.

Para esse tipo de situação, podemos utilizar o procedimento **Split File**, presente no menu **Data**. Por *default* sempre analisamos todos os casos juntos, sem separação por grupos. Por esse motivo, a opção selecionada na janela é **Analyze all cases**. Para poder repetir a análise para as categorias de uma determinada variável, clicamos em **Compare groups** ou **Organize output by groups**, e então o quadro **Groups Based on** fica disponível.

Moveremos para esse quadro a variável (ou variáveis) que definirão os grupos para os quais a análise deve ser repetida. Se mais do que uma variável for selecionada, os grupos serão definidos pela combinação das categorias de todas as variáveis. Podemos ainda escolher se o banco de dados deve ser ordenado pela variável que definirá os grupos (**Sort the file by group variables**) ou se o banco de dados já está ordenado pela variável que definirá os grupos (**File is already sorted**).

No nosso caso, selecionamos a variável sexo e a movemos para o quadro **Groups Based on** e clicamos **OK**. A única mudança que acontece na janela de dados é a mensagem **Split File On** na barra inferior, ou a ordenação dos casos pela variável que definiu os grupos, caso o banco de dados ainda não estivesse ordenado. Porém, qualquer análise ou gráfico feitos de agora em diante vão gerar dois resultados, uma para os homens e outro para as mulheres.

Note que os resultados são apresentados em dois blocos, o primeiro para o sexo masculino e o segundo para o sexo feminino se a opção escolhida foi ou **Organize output by groups**.

Podemos mudar de idéia e querer usar todas as observações para o cálculo das estatísticas. Para cancelar o procedimento **Split File** basta ativar a opção **Analyze all cases** presente na janela de definição da opção **Split File** menu **Data**.

7.4 MANIPULAÇÃO DE ARQUIVOS

Para retornar aos arquivos:

- ***.sav (arquivo de dados)**
- ***.cht (arquivo com cada gráfico realizado)**
- ***.spo (arquivo de resultados)**

procede-se da seguinte maneira:

a) Clica-se na opção de menu **"Window"**;
b) Seleciona-se a janela de saída desejada que consta na lista de arquivos abertos ou disponíveis, clicando **uma vez** sobre sua indicação, obtendo-se a tela desejada.

7.5 COMO APAGAR ANÁLISES NÃO DESEJADAS NO ARQUIVO DE RESULTADOS " *.spo "

Quando inadvertidamente realiza-se um procedimento não desejado, para corrigir o equívoco, procede-se da seguinte forma:

a) Clica-se em **"Edit"**, **"Select"** , **"Output Block"** a partir deste momento será selecionada a última saída executada , o que vai dar origem a uma **"tarja preta"**;
b) Aperta-se o botão **"Delete"**, tornando sem efeito o último procedimento efetuado. Também pode-se apagar outros blocos de resultados, bastando para tal colocar o cursor sobre o bloco que se deseja apagar e repetir a operação explicada acima.