

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

INTRODUÇÃO À ANÁLISE ESTATÍSTICA
UTILIZANDO O SPSS 13.0

Elsa Mundstock
Jandyra Maria Guimarães Fachel
Suzi Alves Camey
Marilyn Agranonik

Série B, Número XX
Porto Alegre - maio de 2006

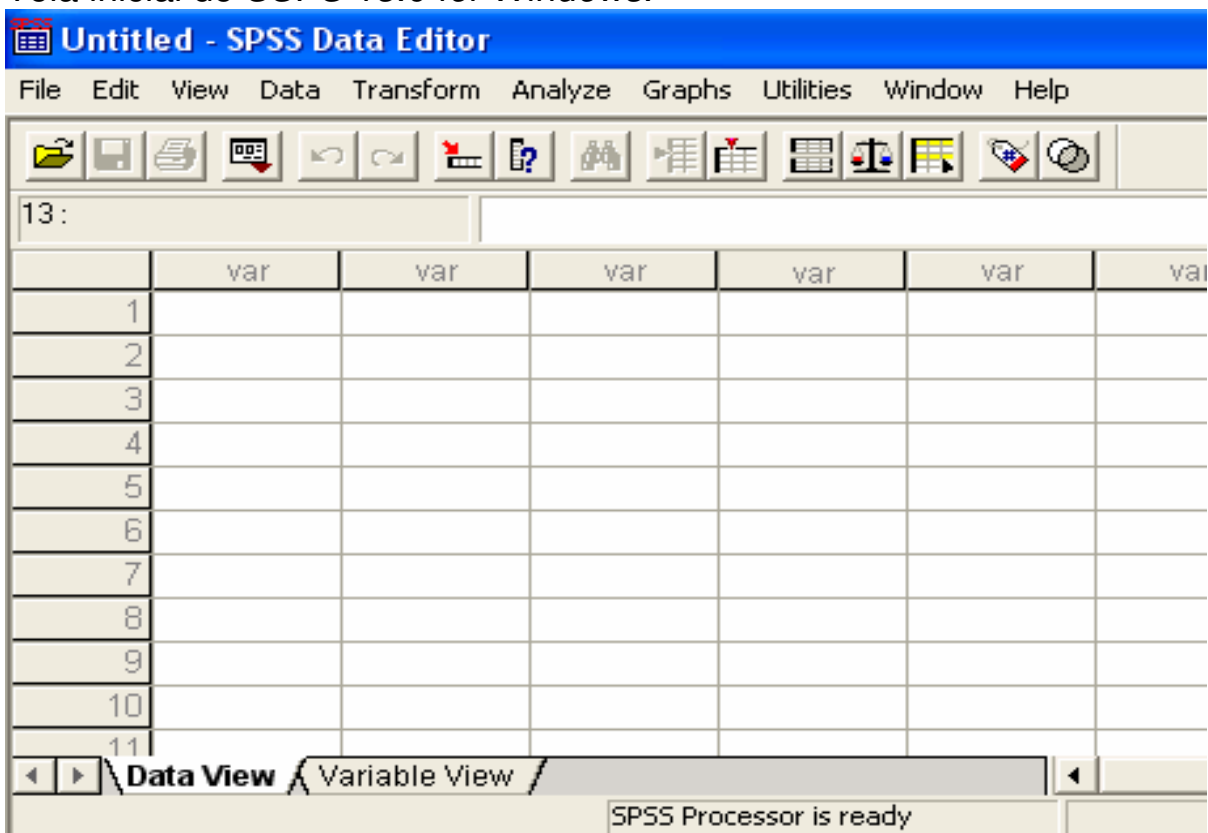
ÍNDICE

1. INTRODUÇÃO AO SPSS	3
1.1 - BANCO DE DADOS: Definição	4
1.2 - COMO CRIAR UM BANCO DE DADOS	4
1.3 - COMO DAR NOME AOS NÍVEIS DE UMA VARIÁVEL	5
1.4 - COMO ACESSAR UM BANCO DE DADOS JÁ EXISTENTE	6
2. INTRODUÇÃO AOS MÉTODOS ESTATÍSTICOS	7
2.1 - TIPOS DE VARIÁVEIS: Definição	7
2.2 - DESCRIÇÃO E EXPLORAÇÃO DE DADOS	7
2.3 - CATEGORIZAÇÃO DE VARIÁVEIS	8
2.3.1 - COMO CATEGORIZAR UMA VARIÁVEL QUANTITATIVA	8
2.3.2 - COMO DAR NOME AOS NÍVEIS DE UMA VARIÁVEL	10
2.4 - COMO CRIAR UMA VARIÁVEL A PARTIR DE UMA DATA	10
2.5 - COMO CRIAR UMA VARIÁVEL ATRAVÉS DA COMBINAÇÃO DE OUTRAS DUAS	10
3. ANÁLISE UNIVARIADA	12
3.1 - VARIÁVEIS QUANTITATIVAS	12
3.1.1 - COMO OBTER AS ESTATÍSTICAS DESCRITIVAS	12
3.1.2 - COMO OBTER UM HISTOGRAMA	12
3.2 - VARIÁVEIS CATEGÓRICAS (QUALITATIVAS)	13
3.2.1 - COMO OBTER A DISTRIBUIÇÃO DE FREQUÊNCIAS	13
3.2.2 - COMO OBTER GRÁFICOS	14
4. ANÁLISE BIVARIADA	16
4.1 - VARIÁVEIS QUANTITATIVAS X QUANTITATIVAS	16
4.1.1 - COMO CALCULAR A CORRELAÇÃO ENTRE DUAS VARIÁVEIS QUANTITATIVAS	16
4.1.2 - COMO OBTER GRÁFICO DE PONTOS (SCATTERPLOT)	16
4.1.3 - COMO OBTER O COEFICIENTE DE CORRELAÇÃO DE PEARSON	18
4.1.4 - COMO FAZER REGRESSÃO LINEAR SIMPLES	19
4.2 - VARIÁVEIS CATEGÓRICAS X CATEGÓRICAS	22
4.2.1 - COMO VERIFICAR A EXISTÊNCIA DE ASSOCIAÇÃO ENTRE VARIÁVEIS CATEGÓRICAS: Teste Qui - Quadrado	23
4.2.2 - COMO CALCULAR OS RESÍDUOS AJUSTADOS	26
4.3 - VARIÁVEIS QUANTITATIVAS X CATEGÓRICAS	27
4.3.1 - COMO FAZER O BOX-PLOT	28
5. COMPARAÇÃO DE MÉDIAS	31
5.1 - COMO COMPARAR MÉDIAS ENTRE DOIS GRUPOS: Teste “t” para Amostras Independentes.	31
5.2 - COMO COMPARAR AS MÉDIAS DE TRES OU MAIS GRUPOS: Análise de Variância – “ANOVA” para um fator”	33
7. MANIPULAÇÃO DE DADOS	36
7.1 SORT CASES	36
7.2 SELECT CASES	37
7.3 SPLIT FILE	39
7.4 MANIPULAÇÃO DE ARQUIVOS	41
7.5 COMO APAGAR ANÁLISES NÃO DESEJADAS NO ARQUIVO DE RESULTADOS “.spo”	41
7.6 COMO REALIZAR ALTERAÇÕES DE DADOS PARA CONSTRUÇÃO DE GRÁFICO	41
8. ESTATÍSTICA NÃO PARAMÉTRICA	43
8.1 TESTE DE KOLMOGOROV-SMIRNOV	43
9. AMOSTRAGEM	45

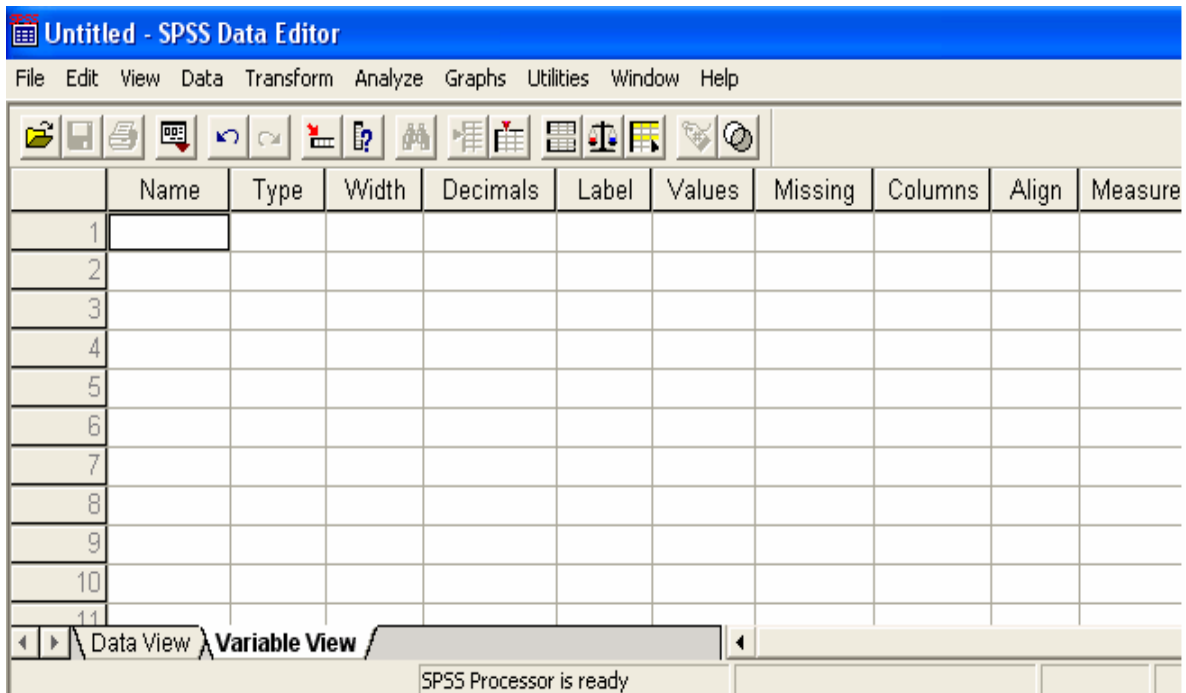
1. INTRODUÇÃO AO SPSS

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

Tela inicial do *SSPS 13.0 for Windows*.



(a)



(b)

Figura 1: Tela inicial do *SPSS 13.0 for Windows*. 1(a): Planilha “Data View”; 1(b): Planilha “Variable View”.

1.1 - BANCO DE DADOS: Definição

Banco de dados é um conjunto de dados registrados em uma planilha, em forma de 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) Clicar em “**File**”; “**New**”; “**Data**”. Aparece a planilha “**data view**”. Na primeira linha estão indicadas as posições das variáveis

- (VAR001, VAR002, etc.), e uma margem vertical numerada a partir de 1 (como mostrado na Figura1a).
- b) Na primeira coluna, correspondendo à VAR001, vamos criar uma variável, por exemplo, “NumCaso” com o número do questionário ou do caso em estudo.
 - c) Para registrar as características da variável, clicar **duas vezes** sobre o nome da coluna. Aparece a planilha “**variable view**” na qual cada variável está definida em uma linha.
 - d) Na primeira coluna (“**Name**”), digitar o nome da variável (NumCaso). Para o nome das variáveis utilize 8 dígitos no máximo, não utilize espaço em branco nem os símbolos “-“, “.” e “/”.
 - e) Clicar na coluna “**Type**” para definir o tipo de variável, aparece a janela “**Variable Type**” onde se deve deixar a opção “**Numeric**”. Se a variável for alfa-numérica (texto) escolha a opção “**String**”. Preferencialmente use sempre a modalidade “Numeric” para variáveis categóricas, como por exemplo, sexo, estado civil, município, etc. criando-se um código para as categorias.
 - f) No caso de não-resposta ou respostas que não se desejam considerar para o tratamento estatístico, como por exemplo, respostas não corretas, etc..., clicar na coluna “**Missing**”, abre-se a janela (“**Missing Values**”), registrar, na opção “**Discrete Missing Values**”, o código de não-resposta, preferencialmente **9, 99, 999, etc.** Clicar em “**OK**”. A melhor opção para não resposta é deixar o espaço em branco no banco de dados.
 - g) Retornar à planilha “**data view**” e passar a digitar, em cada linha da coluna identificada, o valor da variável.
 - h) À medida que o BANCO DE DADOS vai sendo registrado é importante salvar as informações digitadas, para tanto se procede da seguinte forma: Clicar em “**File**”, “**Save as**”... (abre-se a janela do caminho desejado) e criar um nome para o Banco de dados, que terá automaticamente a terminação “**.sav**”.

1.3 - COMO DAR NOME AOS NÍVEIS DE UMA VARIÁVEL

É conveniente registrar no banco de dados os nomes das categorias de variáveis categóricas. Por exemplo, para a variável **sexo**, os códigos poderiam ser: “0” = “masculino” e “1” = “feminino”. Para registrar estes nomes, clicar **2 vezes** sobre a variável **sexo**, abrindo a planilha “**Variable View**” e proceder da seguinte forma:

- a) Clicar em “**Values**”. Abre-se a janela “**Value Labels**”:
- b) Em “**Value**”, digitar **0**;
- c) Em “**Value Label**”, digitar masculino;
- d) Clicar em “**ADD**”;
- e) Procede-se da mesma forma para os demais níveis de categorização: digitar “**1**” para “**Value**” e “feminino” para “**Value Label**”, seguindo-se por “**ADD**”
- f) Clicar 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, número do caso, ou ainda código do registro, pois facilita a localização de informações no caso de serem identificados 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) Clicar em “**File**”, “**Open**”, “**Data**”, abrir o arquivo que se deseja. Usaremos como exemplo o arquivo chamado “**World95.sav**” que se encontra disponível junto com o programa SPSS.

2. INTRODUÇÃO AOS MÉTODOS ESTATÍSTICOS

2.1 - TIPOS DE VARIÁVEIS: Definição

Devemos distinguir entre dois tipos básicos de variáveis:

Variáveis quantitativas. Aquelas que podem ser mensuradas através de escalas quantitativas, isto é, escalas que têm 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. Variáveis medidas originalmente em categorias. Ex: **Sexo, Profissão, Religião, Município, Região...**

2.2 - DESCRIÇÃO E EXPLORAÇÃO DE DADOS

O objetivo básico deste procedimento é introduzir técnicas que permitam organizar, resumir e apresentar os dados, 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 (ex: média, desvio padrão, valor mínimo, valor máximo) ou gráficos (ex: histograma). Se as variáveis são qualitativas usaremos tabelas de frequência ou gráficos (ex: de setores, também conhecido como pie, 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 exemplificar, usaremos uma variável categorizada utilizando 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 um quarto dos casos, ou seja, 25% do tamanho total da amostra.

Os passos necessários para categorizar uma variável utilizando os “quartis” são os seguintes:

1. Calcular os quartis da variável em questão, neste caso, População (**populatn**):
 - a) Clicar em “**Analyze**”, “**Descriptive Statistics**”, “**Frequencies**”;
 - b) Selecionar a variável que se deseja categorizar na janela esquerda e clicar →;
 - c) Retirar a opção de “**Display Frequency Tables**”, a fim de que não venha listada a totalidade de casos da variável (no estudo em pauta o número é de 109 casos);
 - d) Clicar em “**Statistics**” e assinalar “**Quartiles**”;
 - e) Clicar em “**Continue**”; “**OK**”.

RESULTADOS:

Frequencies

Statistics		
Population in thousands		
N	Valid	109
	Missing	0
Percentiles	25	5000,00
	50	10400,00
	75	37100,00

2. Criar uma variável com 4 categorias, definidas pelos quartis, da seguinte maneira:

Categoria	Intervalo de valores
1	Mínimo até 5000,00
2	5001,00 até 10400,00
3	10401,00 até 37100,00
4	37101,00 até o Máximo no Banco de Dados

Para categorizar a variável “**populatn**”, usando os limites dados pelos quartis procede-se da seguinte forma:

- a) Clicar em “**Transform**”, “**Recode**”, “**Into Different Variables**”;
- b) Localizar, na janela à esquerda, a variável a ser categorizada (**populatn**) e clicar na →;
- c) Digitar um novo nome para a variável de saída (**Output Variable**), por exemplo POPREC e clicar em “**Change**”;
- d) Clicar em “**Old and New Values**”;
- e) Clicar em “**Range (lowest through)**” e digitar o valor obtido para o primeiro quartil, no caso 5000,00;
- f) Em “**New Value**”, digitar **1**;
- g) Clicar em “**ADD**”;
- h) Assinalar “**Range**”, colocando: 5001,00 até (Through) 10400,00 (segundo quartil);
- i) Na opção “**New Value**”, digitar **2**;
- j) Clicar em “**ADD**”;
- k) Assinalar “**Range**”, 10401,00 até (Through) o terceiro quartil 37100,00;
- l) Na opção “**New Value**”, digitar **3**;
- m) Clicar em “**ADD**”;
- n) Clicar em “**Range**” (**Through Highest**) e digitar o valor imediatamente superior ao 3º quartil, no caso 37101,00;
- o) Na opção “**New Value**”, digitar **4**;
- p) Clicar em “**ADD**”; “**Continue**” “**OK**”.

A nova variável “**POPREC**” corresponde à variável “**populatn**” categorizada, sendo esta automaticamente incluída no banco de dados que estamos utilizando (**World95.sav / Arquivo Data**).

2.3.2 - COMO DAR NOME AOS NÍVEIS DE UMA VARIÁVEL

No banco de dados, clicar **2 vezes** sobre a nova variável “**Poprec**”, obtendo-se a planilha “**Variable View**”.

Para dar o nome aos níveis (1, 2, 3 e 4) da nova variável “**Poprec**” procede-se da seguinte forma:

- a) Clicar em “**Values**”. Abre-se uma nova janela - “**Value Labels**”;
- b) Em “**Value**”, digitar **1**;
- c) Em “**Value Label**”, digitar pop. inferior a 5000,00;
- d) Clicar em “**ADD**”;
- e) Procede-se da mesma forma para os demais níveis de categorização: 2, 3 e 4;
- f) Clicar em “**Continue**”; “**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, 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 clicar na ↑;
- e) Localizar na janela abaixo de “**Target Variable**” a variável **bdate** e clicar na → (a variável selecionada deve ficar entre os parênteses);
- f) Clicar em “**OK**”.

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

Nesta seção, será utilizado o banco “**GSS93.sav**”, que também se encontra disponível junto com o programa SPSS.

Para criar uma variável a partir da combinação de outras duas, como por exemplo, combinar a variável sexo (**sex**) e a variável raça (**race**) utilizaremos o seguinte procedimento para criar a variável SEXRACE.

Sabendo que a variável SEX é categorizada da seguinte forma:

1-Male e 2-Female

e a variável RACE é categorizada da seguinte forma:

1- White, 2-Black e 3-Other

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

1- White Male,

2- White Female,

3- Black Male

4- Black Female

5- Other Male

6- Other Female

Então se procede da seguinte forma:

- a) Selecionar “**Transform**”, “**Compute**”;
- b) Em “**Target Variable**” digite o nome da nova variável, por exemplo SEXRACE;
- c) Na janela “**Numeric Expression**” digite 1;
- d) Clicar 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, **sex**), clicar na →;
- h) Digitar =1 & na janela ao lado da variável **sex**;
- g) Selecionar na janela ao lado a variável **race** e clicar na →;
- h) Na janela ao lado da variável **race** digitar =1;
- i) Após esse procedimento a expressão na janela deve ser a seguinte: **sex=1 & race=1**;
- j) Clicar em “**Continue**” e “**OK**”, (a variável **SEXRACE** aparecerá no final do banco de dados,
- k) Para criar as demais categorias da variável **SEXRACE** procede-se de maneira análoga, alterando o código na janela “**Numeric Expression**” para 2, 3, 4, 5 e 6 e a expressão da janela “**Include if case satisfies condition**”.

3. ANÁLISE UNIVARIADA

3.1 - VARIÁVEIS QUANTITATIVAS

3.1.1 - COMO OBTER AS ESTATÍSTICAS DESCRITIVAS

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

- Clicar em “**Analyze**”, “**Descriptive Statistics**”, “**Descriptives**”;
- Localizar na janela à esquerda a variável de interesse (por exemplo, mortalidade infantil) e clicar na →;
- Clicar em “**Options**”, e assinalar as opções desejadas;
- Clicar em “**Continue**”; “**OK**”;
- Os resultados da análise estatística aparecem na janela de resultados (**OUTPUT**), que poderá ser salva, dando origem a um arquivo do tipo “.spo” (**SPSS output**).

EXEMPLO:

Descriptives

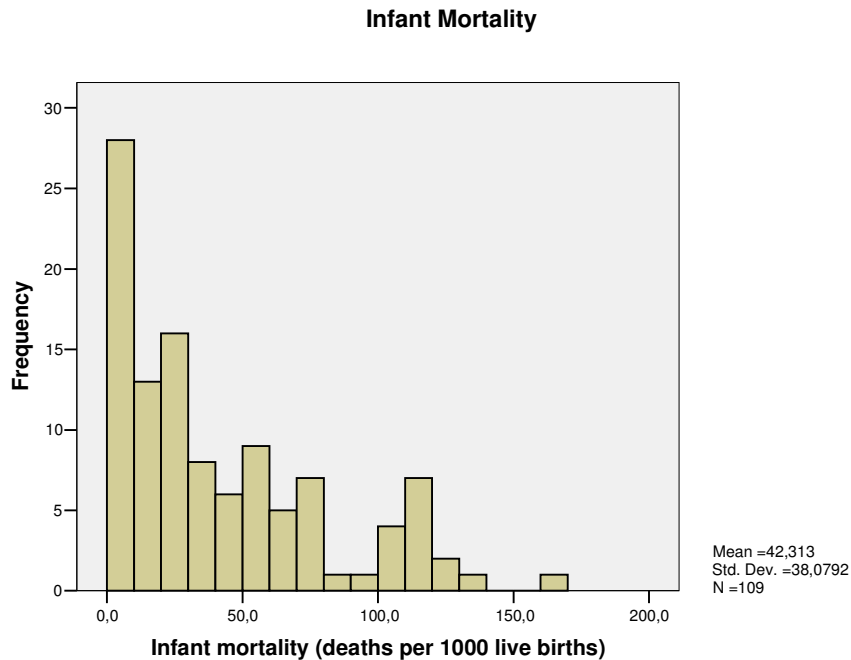
Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Infant mortality (deaths per 1000 live births)	109	4,0	168,0	42,313	38,0792
Valid N (listwise)	109				

3.1.2 - COMO OBTER UM HISTOGRAMA

- Clicar em “**Graphs**”, “**Histogram**”
- Localizar na janela a variável desejada,
- Após ter selecionado a variável (neste caso, **babymort**), clicar na →;
- Pode-se clicar na opção “**Titles**” para dar um título ao histograma.
- Clicar em “**OK**”

EXEMPLO: Histograma da variável “**Infant Mortality**”



3.2 - VARIÁVEIS CATEGÓRICAS (QUALITATIVAS)

3.2.1 - COMO OBTER A DISTRIBUIÇÃO DE FREQUÊNCIAS

Para calcular as frequências procede-se da seguinte forma:

- Clicar em “**Analyze**”, “**Descriptive Statistics**”, “**Frequencies**”;
- Selecionar a variável desejada (neste caso, **region**), clicar na →;
- Selecionar “**Display frequency tables**”;
- Clicar em “**OK**”.

RESULTADO:

Frequencies

Statistics

Region or economic group

N	Valid	109
	Missing	0

Region or economic group

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	OECD	21	19,3	19,3	19,3
	East Europe	14	12,8	12,8	32,1
	Pacific/Asia	17	15,6	15,6	47,7
	Africa	19	17,4	17,4	65,1
	Middle East	17	15,6	15,6	80,7
	Latn America	21	19,3	19,3	100,0
	Total	109	100,0	100,0	

3.2.2 - COMO OBTER GRÁFICOS

Para se obterem os diferentes tipos de gráficos disponíveis no programa procede-se da seguinte forma:

- Clicar em “**Graphs**”, selecionar o gráfico desejado, que ao salvá-lo, dá origem a um arquivo do tipo “**.cht (Chart)**” (arquivo de gráficos).

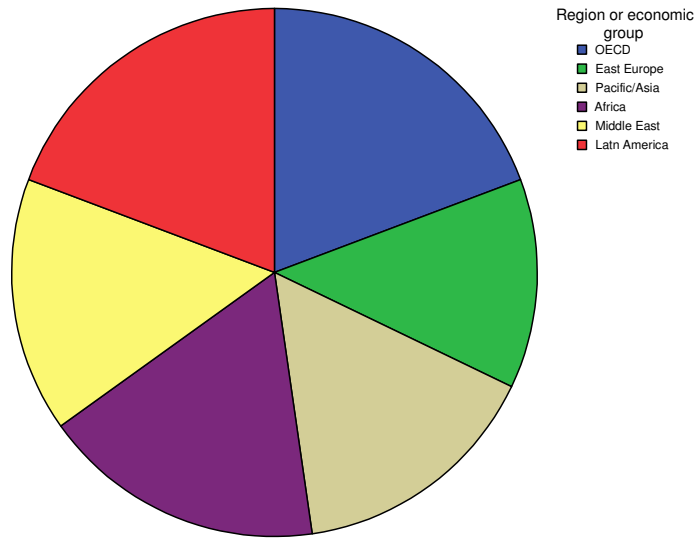
OBSERVAÇÃO:

Com variáveis categóricas, o adequado é fazer gráfico de setores (Pie), de Colunas...

EXEMPLO: Gráfico de Setores (Pie) para a variável “**region**”

- Clicar em “**Graphs**”, selecionar “**Pie**”;
- Selecionar a opção “**Summaries for groups of cases**” e clicar em “**Define**”;
- Na opção “**Define Slices by**” selecionar a variável “**region**”.

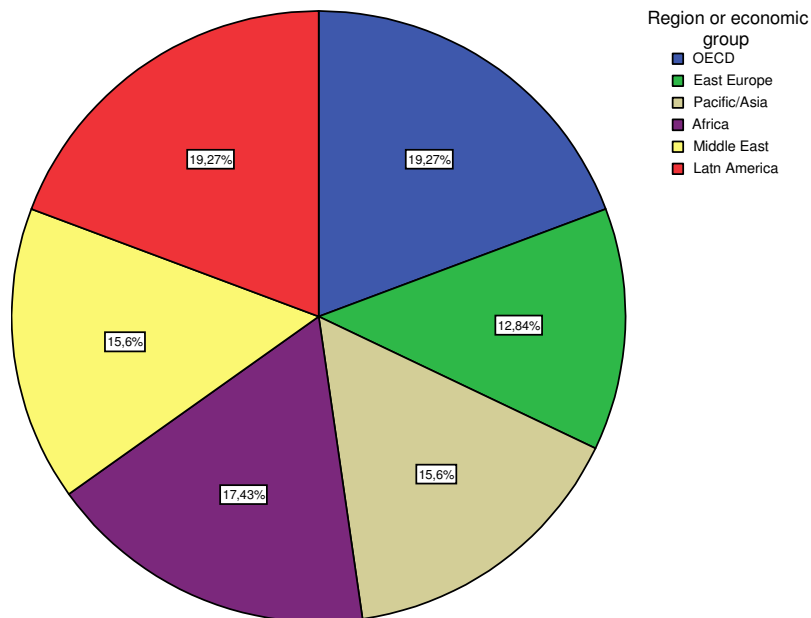
RESULTADO:



Para colocar o valor percentual de cada categoria no gráfico:

- Clicar duas vezes no gráfico;
- Abre o **SPSS Chart Editor**; clicar em **elements/show data labels**;
- Na janela “**properties**”, em “data value labels”, selecionar **percents**;
- Clicar “**OK**”.

RESULTADO:



4. ANÁLISE BIVARIADA

Para realizar uma análise **bivariada**, ou seja, análise da relação entre duas variáveis, utilizam-se 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)**
 - 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.

4.1.2 - COMO OBTER GRÁFICO DE PONTOS (SCATTERPLOT)

O gráfico de pontos (**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).

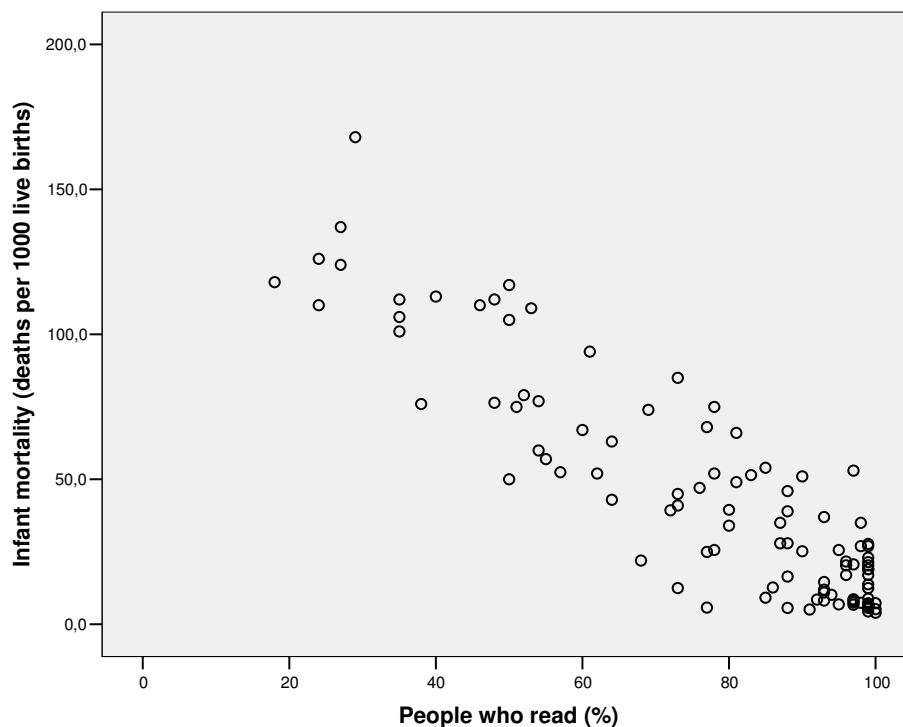
Através deste gráfico podemos visualizar empiricamente a relação entre as variáveis.

Para se obter o gráfico **Scatterplot** (gráfico de pontos) procede-se da seguinte maneira:

- Clicar em **“Graphs”**; **“Scatter”**, abre a janela **“Scatterplot”**, onde se seleciona o tipo de gráfico, neste caso **“Simple”**;
- Clicar em **“Define”**. São apresentadas as variáveis do Banco de Dados, escolhem-se as variáveis, no caso, **“Literacy”** e **“Babymort”**;
- Define-se a variável Y no caso **“Babymort”**, clicar na flecha pertinente e a variável X, no caso **“Literacy”**, clicando-se na flecha correspondente;
- Clicar em **“OK”**. O gráfico é gerado na janela **“Chart”**. “Esta janela pode ser salva em arquivo com a extensão **“.cht”** (arquivo de gráfico).

RESULTADO:

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) Clicar em “**Analyze**”, “**Correlate**”, “**Bivariate**”, abre-se a janela “**Bivariate Correlations**”;
- b) Selecionar as variáveis (no caso “**Literacy**” e “**Babymort**”), clicar na →;
- c) Selecionar a estatística desejada, no caso, Pearson;
- d) Clicar em “**OK**”;

OBSERVAÇÃO:

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

O coeficiente fornece informação do tipo de associação das variáveis 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 ou aproximadamente nulo, significa que não existe correlação linear.

RESULTADO:

Nos resultados aparece uma tabela com 3 linhas em cada célula: o coeficiente de correlação, o resultado do teste de significância desse coeficiente e o número de observações utilizadas no cálculo do coeficiente.

Correlations

Correlations

		People who read (%)	Infant mortality (deaths per 1000 live births)
People who read (%)	Pearson Correlation	1	-,900**
	Sig. (2-tailed)	,	,000
	N	107	107
Infant mortality (deaths per 1000 live births)	Pearson Correlation	-,900**	1
	Sig. (2-tailed)	,000	,
	N	107	109

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

As hipóteses do teste 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 significativa)**

CONCLUSÃO:

Ao analisarmos os dados obtidos, rejeita-se H_0 (hipótese nula) de que **não há correlação** entre “**Literacy**” e “**Babymort**”, uma vez que o valor de **p** (“Sig. 2-tailed”) é menor que 0,001 (muito pequena, neste caso) e conclui-se em favor da hipótese alternativa de que há correlação entre as variáveis em estudo.

Este resultado confirma a configuração do gráfico **Scatterplot**, mostrando que à medida que a taxa de pessoas alfabetizadas aumenta, a mortalidade infantil tende a diminuir.

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 estimar o valor de uma variável em função de outra (Y em função de X). Para aplicar o modelo de regressão

devemos definir *a priori* a variável explicativa ou independente (X) e 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, “Literacy” e “Babymort”, procede-se da seguinte forma:

- Clicar “**Analyze**”, “**Regression**”, “**Linear**”;
- Definir a variável independente “**Literacy**”, e a variável dependente “**Babymort**”;
- Selecionar “**Method Enter**”;
- Na opção “**Statistics**”, selecionar “**Casewise Diagnostics**” para mostrar a tabela com os valores residuais atípicos;
- Na opção “**Save**”, selecionar “**Predicted Values**” / “**Unstandardized**”, para salvar no banco de dados os valores estimados pela reta ajustada;
- Clicar “**OK**”.

RESULTADO:

Regression

Variables Entered/Removed^b

Model	Variables Entered	Variables Removed	Method
1	People who read (%) ^a		Enter

a. All requested variables entered.

b. Dependent Variable: Infant mortality (deaths per 1000 live births)

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,900 ^a	,811	,809	16,7334

a. Predictors: (Constant), People who read (%)

b. Dependent Variable: Infant mortality (deaths per 1000 live births)

INTERPRETAÇÃO: O coeficiente de determinação (R square) é igual a 0,811, este valor indica que 81,1% da variação da variável mortalidade

infantil (Babymort) é explicada pela variável taxa de pessoas alfabetizadas (Literacy) através do modelo de regressão linear simples.

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	126066,8	1	126066,834	450,226	,000 ^a
	Residual	29400,822	105	280,008		
	Total	155467,7	106			

a. Predictors: (Constant), People who read (%)

b. Dependent Variable: Infant mortality (deaths per 1000 live births)

INTERPRETAÇÃO: A tabela acima (ANOVA) analisa o modelo de regressão. A coluna F dessa tabela corresponde ao teste conjunto dos coeficientes de regressão.

Neste exemplo, fizemos regressão simples, logo os 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)	160,732	5,794		27,740	,000
	People who read (%)	-1,507	,071	-,900	-21,219	,000

a. Dependent Variable: Infant mortality (deaths per 1000 live births)

INTERPRETAÇÃO: A equação de regressão é $Y = a + bX$, onde o coeficiente linear da reta é $a = 160,732$ e o coeficiente angular é $b = -1,507$. Como o “sig” de b é menor que 0,001, rejeitamos a hipótese nula de que $\beta = 0$. A partir desta equação podemos estimar (predizer) os valores da variável dependente (**babymort**).

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	Infant mortality (deaths per 1000 live births)	Predicted Value	Residual
1	3,046	168,0	117,027	50,973

a. Dependent Variable: Infant mortality (deaths per 1000 live births)

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	10,026	133,605	42,674	34,4864	107
Residual	-38,888	50,973	,000	16,6543	107
Std. Predicted Value	-,947	2,637	,000	1,000	107
Std. Residual	-2,324	3,046	,000	,995	107

a. Dependent Variable: Infant mortality (deaths per 1000 live births)

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

OBSERVAÇÃO: Os valores de Y estimados por essa equação aparecem na última coluna do banco de dados, pois selecionamos a opção “**Save**” / “**Predicted Values**” / “**Unstandardized**”. Essa coluna tem o nome de **pre-1 (Unstandardized Predicted Value)**. Os resíduos que forem calculados para outras variáveis terão os nomes pre-2, pre-3, etc, esses nomes podem ser alterados pelo usuário.

4.2 - VARIÁVEIS CATEGÓRICAS X CATEGÓRICAS

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

O banco **GSS93.sav**, será utilizado para obter a tabela de contingência e estudar a associação entre “**Sexrace**” e “**Income4**” (salário em categorias). Procede-se da seguinte forma:

- a) Clicar em “**Analyze**”, “**Descriptive Statistics**”, “**Crosstabs**”;
- b) Definir a variável da linha “**Row - Sexrace**”;
- c) Definir a variável da coluna “**Column – Income4**”;
- d) Clicar em “**Statistics**”;
- e) Escolher o tratamento estatístico desejado, no caso, “**Chi-Square**”;
- f) Clicar em “**Continue**”;
- g) Clicar em “**Cell**”, veremos a janela “**Crosstabs : Cell Display**”;
- h) Assinalar as opções “**Observed**”; etc, de acordo com o desejado;
- i) Clicar em “**Continue**”; “**OK**”.

O valor esperado de cada casela na tabela pode ser obtido na janela “**Crosstabs : Cell Display**” assinalando-se também a opção “**Expected**”.

RESULTADOS:

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
SEXRACE * Total Family Income	1500	100,0%	0	,0%	1500	100,0%

SEXRACE * Total Family Income Crosstabulation

			Total Family Income				Total
			24,999 or less	25,000 to 39,999	40,000 to 59,999	60,000 or more	
SEXRACE 1,00	Count	181	130	104	137	552	
	Expected Count	215,3	110,4	84,6	141,7	552,0	
2,00	Count	285	125	99	196	705	
	Expected Count	275,0	141,0	108,1	181,0	705,0	
3,00	Count	30	10	12	14	66	
	Expected Count	25,7	13,2	10,1	16,9	66,0	
4,00	Count	58	22	5	17	102	
	Expected Count	39,8	20,4	15,6	26,2	102,0	
5,00	Count	13	3	2	5	23	
	Expected Count	9,0	4,6	3,5	5,9	23,0	
6,00	Count	18	10	8	16	52	
	Expected Count	20,3	10,4	8,0	13,3	52,0	
Total	Count	585	300	230	385	1500	
	Expected Count	585,0	300,0	230,0	385,0	1500,0	

A leitura das caselas na 1ª linha (count) informa a freqüência bruta e a 2ª linha (expected count) corresponde ao valor esperado, isto é, o número de pessoas que seria esperado caso não houvesse nenhuma associação entre as variáveis em estudo, ou seja, 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 à 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:

SEXRACE * Total Family Income Crosstabulation

		Total Family Income				Total
		24,999 or less	25,000 to 39,999	40,000 to 59,999	60,000 or more	
SEXRACE 1,00	Count	181	130	104	137	552
	Expected Count	215,3	110,4	84,6	141,7	552,0
	% within SEXRACE	32,8%	23,6%	18,8%	24,8%	100,0%
2,00	Count	285	125	99	196	705
	Expected Count	275,0	141,0	108,1	181,0	705,0
	% within SEXRACE	40,4%	17,7%	14,0%	27,8%	100,0%
3,00	Count	30	10	12	14	66
	Expected Count	25,7	13,2	10,1	16,9	66,0
	% within SEXRACE	45,5%	15,2%	18,2%	21,2%	100,0%
4,00	Count	58	22	5	17	102
	Expected Count	39,8	20,4	15,6	26,2	102,0
	% within SEXRACE	56,9%	21,6%	4,9%	16,7%	100,0%
5,00	Count	13	3	2	5	23
	Expected Count	9,0	4,6	3,5	5,9	23,0
	% within SEXRACE	56,5%	13,0%	8,7%	21,7%	100,0%
6,00	Count	18	10	8	16	52
	Expected Count	20,3	10,4	8,0	13,3	52,0
	% within SEXRACE	34,6%	19,2%	15,4%	30,8%	100,0%
Total	Count	585	300	230	385	1500
	Expected Count	585,0	300,0	230,0	385,0	1500,0
	% within SEXRACE	39,0%	20,0%	15,3%	25,7%	100,0%

Os percentuais relativos à coluna (**Column**) e ao total (**Total**) podem ser obtidos da mesma forma que para o cálculo da percentagem da linha. Cada casela poderia ter até 5 valores, descritos a seguir:

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

OBSERVAÇÃO:

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

RESULTADO:

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	42,956 ^a	15	,000
Likelihood Ratio	44,902	15	,000
Linear-by-Linear Association	6,226	1	,013
N of Valid Cases	1500		

a. 2 cells (8,3%) have expected count less than 5. The minimum expected count is 3,53.

INTERPRETAÇÃO: Considerando que $p < 0,000$ (significance), rejeita-se a hipótese nula (H_0) de independência entre as variáveis. Sendo assim, conclui-se que há evidências de associação entre “**Sexrace**” e “**Income4**”.

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 significativa 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) Selecionar “**Analyze**”, “**Descriptive Statistics**”, “**Crosstabs**”;
- b) Clicar em “**Cells**”, abre-se a janela “**Crosstabs**”: “**Cell Display**”;
- c) Assinalar a opção “**Observed**” e “**Adj. standardized**”;
- d) Clicar em “**Continue**”; “**OK**”.

RESULTADOS:

SEXTRACE * Total Family Income Crosstabulation

			Total Family Income				Total
			24,999 or less	25,000 to 39,999	40,000 to 59,999	60,000 or more	
SEXTRACE	1,00	Count	181	130	104	137	552
		Adjusted Residual	-3,8	2,6	2,9	-,6	
	2,00	Count	285	125	99	196	705
		Adjusted Residual	1,1	-2,1	-1,3	1,8	
	3,00	Count	30	10	12	14	66
		Adjusted Residual	1,1	-1,0	,7	-,8	
	4,00	Count	58	22	5	17	102
		Adjusted Residual	3,8	,4	-3,0	-2,2	
	5,00	Count	13	3	2	5	23
		Adjusted Residual	1,7	-,8	-,9	-,4	
	6,00	Count	18	10	8	16	52
		Adjusted Residual	-,7	-,1	,0	,9	
Total		Count	585	300	230	385	1500

CONCLUSÃO: A associação entre **sex** (sexo) e **income4** (salário em categorias) já foi considerada 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-negra 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) Seleccione “**Simple**” / “**Summaries for groups of cases**”;
- c) Clicar em “**Define**”;
- d) Em **Variable** seleccionar uma variável quantitativa (por exemplo, **Babymort**);
- e) Em **Category Axis**, seleccionar uma variável categórica (por exemplo, **Region**);
- f) Clicar em “**OK**”.

RESULTADO:

Explore

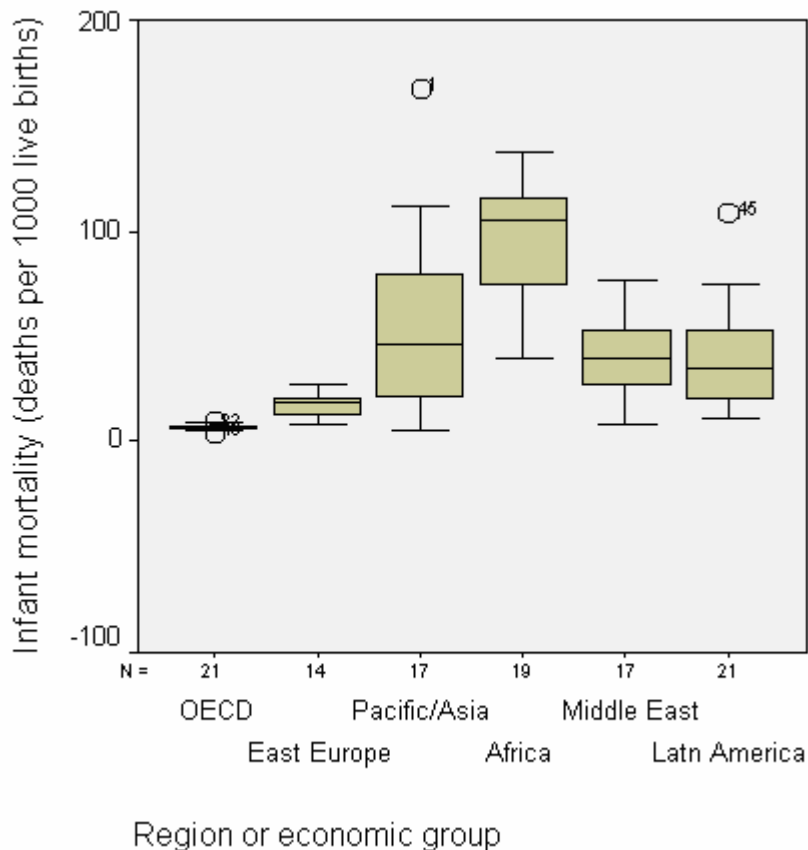
Region or economic group

Case Processing Summary

	Region or economic group	Cases					
		Valid		Missing		Total	
		N	Percent	N	Percent	N	Percent
Infant mortality (deaths per 1000 live births)	OECD	21	100,0%	0	,0%	21	100,0%
	East Europe	14	100,0%	0	,0%	14	100,0%
	Pacific/Asia	17	100,0%	0	,0%	17	100,0%
	Africa	19	100,0%	0	,0%	19	100,0%
	Middle East	17	100,0%	0	,0%	17	100,0%
	Latin America	21	100,0%	0	,0%	21	100,0%

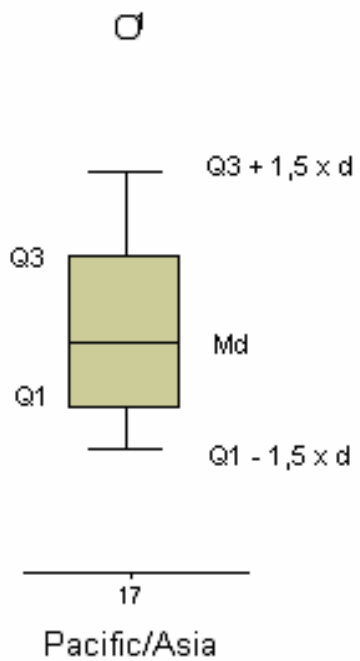
INTERPRETAÇÃO: A tabela acima apresenta o número de casos válidos (**valid**), o número de não respostas (**missing**) e o número total das observações de cada categoria.

Infant mortality (deaths per 1000 live births)



INTERPRETAÇÃO:

Através do Box-plot pode-se observar 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 - 1º quartil (limite inferior do “box”)

Q3: Quartil superior - 3º quartil (limite superior do “box”)

d: diferença interquartílica ($d = Q3 - Q1$)

o : outlier (valores acima de $1,5 \times d$)

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

5. 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 na empresa.

- a) Sexo (masculino, feminino) - Dois grupos (variável que define os grupos).
- b) Idade no 1º. Casamento (Agewed) - Variável resposta ou de teste.

Para a aplicação do teste “t” nesta situação procede-se da seguinte forma:

- a) Clicar em “**Analyze**”, “**Compare Means**”, “**Independent Samples t test**”;
- b) Clicar sobre a variável de teste (Test Variables): “**Agewed**” ou, conforme o caso em estudo, clicar na variável correspondente;
- c) Clicar sobre a variável de grupo (Grouping Variable) “**Gender**”;
- d) Clicar em: “**Define Group**”;
- e) Abre-se uma janela, na qual se define 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 **1** e “**Group 2**” (no caso feminino) digitando-se o código **2**.
(*Observação:* No caso de se desejar confirmar os valores atribuídos às variáveis, abrir a janela “**Utilities**”, “**Variables**”)
- f) Clicar em “**Continue**” e “**OK**”.

RESULTADO:

T-Test

Group Statistics

Respondent's Sex		N	Mean	Std. Deviation	Std. Error Mean
Age When First Married	Male	492	24,16	4,87	,22
	Female	710	21,84	4,93	,18

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Age When First Married	Equal variances assumed	,342	,559	8,066	1200	,000	2,32	,29	1,76	2,88
	Equal variances not assumed			8,085	1064,66	,000	2,32	,29	1,76	2,88

INTERPRETAÇÃO: Ao serem analisados os dados do exemplo acima vemos o seguinte:

- Observa-se o resultado do teste para variâncias iguais (Teste de Levene). Neste exemplo, o valor de p para o teste Levene é 0,559, não se rejeita a hipótese de variâncias iguais.
- O teste t a ser utilizado é o que aparece na primeira linha (Equal variances assumed), considerando que $p < 0,000$ (Sig 2-tailed), rejeita-se a hipótese nula (H_0) de igualdade das médias dos dois grupos, logo, pode-se concluir que as médias da variável **agewed** são significativamente diferentes entre os dois grupos de sexo.

As hipóteses do teste Levene de igualdade de variâncias são:

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

As hipóteses do teste “t” para igualdade de médias entre Amostras Independentes são:

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

5.2 - COMO COMPARAR AS MÉDIAS DE TRES 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) Clicar em “**Analyze**”, “**Compare Means**”, “**One-Way Anova**”;
- b) Assinalar a variável dependente em “**Dependent List**”, clicar 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 “**Infant mortality**”;
- c) Assinalar a variável independente “**Factor**”, no caso “**Region**”, clicar na flecha correspondente;
- d) Clicar o botão “**Options**”.
- e) Clicar na alternativa do quadro “**Statistics**” “**Descriptive**” e depois “**Continue**”;
- f) Clicar no botão “**Post Hoc**”. Aparece uma tela “**One-Way Anova: Post Hoc Multiple Comparisons**”, assinalar a alternativa “**Tukey**” ou outro teste conforme a escolha;
- g) Clicar em “**Continue**”, “**OK**”.

RESULTADOS:

Oneway

Descriptives

Infant mortality (deaths per 1000 live births)

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
OECD	21	6,91	1,17	,26	6,38	7,44	4,0	9,2
East Europe	14	16,89	5,48	1,47	13,73	20,06	8,7	27,0
Pacific/Asia	17	53,88	46,44	11,26	30,00	77,76	4,4	168,0
Africa	19	94,18	28,65	6,57	80,37	107,99	39,3	137,0
Middle East	17	41,39	19,18	4,65	31,53	51,25	8,6	76,4
Latn America	21	39,11	24,52	5,35	27,95	50,28	10,2	109,0
Total	109	42,31	38,08	3,65	35,08	49,54	4,0	168,0

ANOVA

Infant mortality (deaths per 1000 live births)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	88983,515	5	17796,703	27,108	,000
Within Groups	67619,443	103	656,499		
Total	156602,958	108			

INTERPRETAÇÃO: No exemplo acima o valor p (Sig) da ANOVA é $p < 0,001$, então, rejeita-se a hipótese nula (H_0) de igualdade das médias dos seis grupos, logo, pelo menos duas médias de mortalidade infantil diferem entre si. Um teste de comparações múltiplas (post-hoc) permite identificar qual(is) grupo(s) diferem.

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

- **Hipótese Nula (H_0):** As médias de todos os grupos são iguais.
- **Hipótese Alternativa (H_1):** Pelo menos duas médias diferem entre si.

Post Hoc Tests

Multiple Comparisons

Dependent Variable: Infant mortality (deaths per 1000 live births)

Dunnnett T3

(I) Region or economic group	(J) Region or economic group	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
OECD	East Europe	-9,983*	8,841	,000	-15,109	-4,857
	Pacific/Asia	-46,972*	8,359	,010	-85,018	-8,927
	Africa	-87,269*	8,113	,000	-109,134	-65,404
	Middle East	-34,484*	8,359	,000	-50,215	-18,753
	Latn America	-32,204*	7,907	,000	-49,788	-14,620
East Europe	OECD	9,983*	8,841	,000	4,857	15,109
	Pacific/Asia	-36,989	9,247	,060	-75,025	1,046
	Africa	-77,286*	9,025	,000	-99,390	-55,182
	Middle East	-24,501*	9,247	,001	-40,604	-8,399
	Latn America	-22,221*	8,841	,008	-40,163	-4,280
Pacific/Asia	OECD	46,972*	8,359	,010	8,927	85,018
	East Europe	36,989	9,247	,060	-1,046	75,025
	Africa	-40,297	8,554	,064	-81,986	1,393
	Middle East	12,488	8,788	,991	-27,291	52,267
	Latn America	14,768	8,359	,972	-25,556	55,092
Africa	OECD	87,269*	8,113	,000	65,404	109,134
	East Europe	77,286*	9,025	,000	55,182	99,390
	Pacific/Asia	40,297	8,554	,064	-1,393	81,986
	Middle East	52,785*	8,554	,000	27,463	78,107
	Latn America	55,065*	8,113	,000	28,621	81,508
Middle East	OECD	34,484*	8,359	,000	18,753	50,215
	East Europe	24,501*	9,247	,001	8,399	40,604
	Pacific/Asia	-12,488	8,788	,991	-52,267	27,291
	Africa	-52,785*	8,554	,000	-78,107	-27,463
	Latn America	2,280	8,359	1,000	-19,841	24,400
Latn America	OECD	32,204*	7,907	,000	14,620	49,788
	East Europe	22,221*	8,841	,008	4,280	40,163
	Pacific/Asia	-14,768	8,359	,972	-55,092	25,556
	Africa	-55,065*	8,113	,000	-81,508	-28,621
	Middle East	-2,280	8,359	1,000	-24,400	19,841

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

As variâncias da variável mortalidade infantil dos diferentes grupos são muito heterogêneas, por esta razão utilizamos um teste de comparações múltiplas que leva em conta esta desigualdade de variâncias, por exemplo, o teste T3 de Dunnett.

7. MANIPULAÇÃO DE DADOS

7.1 SORT CASES

Uma das necessidades na hora da manipulação dos dados no dia-a-dia é a ordenação dos casos segundo uma ou mais variáveis. Para fazer isso no *SPSS for Windows*, usar o procedimento **Sort Cases** presente no menu **Data**.

Após clicar em **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 no quadro e, em segundo lugar, pela segunda variável 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 idade (ordem decrescente) **dentro** dos códigos de sexo (ordem crescente). 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 executar a ordenação.

Note que após a execução deste comando a posição dos indivíduos nas linhas fica completamente alterada, pois o indivíduo na linha 1 do banco de dados após ordenado 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

Uma outra necessidade é 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:

- **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;
- **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 é completamente diferente 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 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.

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 podem-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.

7.6 COMO REALIZAR ALTERAÇÕES DE DADOS PARA CONSTRUÇÃO DE GRÁFICO

Caso haja uma variável selecionada que não é a desejada ou deseja-se substituí-la, para um novo estudo, procede-se da seguinte maneira:

- a) Clica-se sobre a variável a ser substituída;

- b) Clica-se sobre a seta correspondente que deverá estar com o sentido voltado para a esquerda; com este procedimento deixa-se o campo livre para a próxima variável;
- c) Completa-se este campo de acordo com o item "d" do procedimento "COMO OBTER GRÁFICO DE CORRELAÇÃO".

OBSERVAÇÃO:

Se o campo destinado à variável estiver ocupado, não há disponibilidade para substituição automática (a flecha indicativa estará em cinza claro, da mesma forma que ficam quaisquer botões quando não estão em disponibilidade).

Caso se deseje enriquecer o gráfico quanto a sua forma de apresentação (mudar o tamanho dos "**labels**", colocar em negrito, colocar grades, trocar de cor, etc.) clica-se no "**Edit**" da barra de ferramentas. O resultado é o aparecimento de uma nova barra de fontes e de um novo menu de funções, sobre as quais basta **um clicar** para obter-se o resultado desejado.

8. ESTATÍSTICA NÃO PARAMÉTRICA

8.1 TESTE DE KOLMOGOROV-SMIRNOV

Para verificar se uma variável segue determinada distribuição procede-se da seguinte maneira:

- a) Clicar em “**Analyze**”, “**Non-Parametric Tests**”, “**1-Sample KS**”;
- b) Assinalar a variável dependente em “**Dependent List**”, clicar 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 “**Infant mortality**”;
- c) Assinalar a distribuição em relação a qual a variável será testada em **Test Distribution**. Neste caso, distribuição Normal;
- d) Clicar o botão “**Options**”.
- e) Clicar na alternativa do quadro Statistics “**Descriptive**”;
- f) Clicar em “**Continue**”, “**OK**”.

RESULTADO:

NPar Tests

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Infant mortality (deaths per 1000 live births)	109	42,313	38,079	4,0	168,0

One-Sample Kolmogorov-Smirnov Test

		Infant mortality (deaths per 1000 live births)
N		109
Normal Parameters ^{a,b}	Mean	42,313
	Std. Deviation	38,079
Most Extreme Differences	Absolute	,169
	Positive	,169
	Negative	-,157
Kolmogorov-Smirnov Z		1,769
Asymp. Sig. (2-tailed)		,004

a. Test distribution is Normal.

b. Calculated from data.

INTERPRETAÇÃO: Ao analisarmos os dados obtidos, rejeita-se H_0 (hipótese nula) de que **a variável Infant mortality segue distribuição Normal**, uma vez que o valor de **p** (“Asymp. Sig. 2-tailed”) é menor que 0,004 (muito pequena, neste caso) e conclui-se em favor da hipótese alternativa de que a distribuição da mortalidade infantil não deve ser Normal.

As hipótese do Teste de Kolmogorov-Smirnov são:

- **Hipótese Nula (H_0): A variável segue distribuição Normal.**
- **Hipótese Alternativa (H_1): A variável não segue distribuição Normal.**

9. AMOSTRAGEM

Para acessar o módulo que permite selecionar diferentes tipos de amostras no SPSS 13.0, procede-se da seguinte maneira:

a) Analyze; complex samples; select a sample

Abre uma janela “sampling wizard”

Escolhe-se o que se quer fazer:

- Desing a sample
- Edit a sample desing
- Draw a sample

Utilizando “**desing sample**”:

a) Em file, escolher o local onde deseja salvar os resultados. É gravado um arquivo do tipo **.csplan**.

b) Clicar em avançar. Aparece a janela “**desing variables**”, pode-se escolher estratificar ou criar clusters para a amostra. Por exemplo, para criar extratos por região, selecionar a variável **region** na janela da esquerda, clicar em → passando a variável para o quadro “**stratify by**”, clicar em avançar.

c) Uma nova janela aparece. Deve-se escolher o método de seleção da amostra dentro dos estratos definidos anteriormente (region). Em **type**, pode-se escolher entre “**simple random sampling**” (amostra aleatória simples) “**simple sistematic**”(amostra sistemática simples) “**simple sequential**” (amostra seqüencial simples) “**PPS**” “**PPS systematic**” e “**PPS sequential**”

Para a amostra aleatória simples existem as opções de amostragem com ou sem reposição.

Após escolher o método, clicar em **Continue**.

d) Aparece uma janela, na qual deve-se especificar o tamanho da amostra. Colocar o valor escolhido em “value”, clicar em **Continue**.

e) Uma nova janela “**output variables**” aparece. Aqui é possível salvar variáveis como tamanho da população (**population size**), tamanho da amostra (**sample size**), proporção da amostra (**sample proportion**) e peso da amostra (**sample weight**). Clicar em avançar.

f) Na próxima janela (“**plan summary**”) aparece um resumo do que foi pedido até então. Pode-se escolher criar mais um estágio de

amostragem ou terminar por aqui. Se a escolha for o último caso, concluir.

RESULTADOS:

➤ **No banco de dados:** as variáveis selecionadas na janela “**output variables**” aparecem nas últimas colunas do banco:

- InclusionProbability_1_ Inclusion (Selection) Probability for Stage 1
- SampleWeightCumulative_1_: Cumulative Sampling Weight for Stage 1
- PopulationSize_1_: Population Size for Stage 1
- SampleSize_1_: Sample Size for Stage 1
- SampleWeight_Final_: Final Sampling Weight

➤ **No output:**

Complex Samples: Selection

Summary for Stage 1

Region or economic group	Number of Units Sampled		Proportion of Units Sampled	
	Requested	Actual	Requested	Actual
OECD	10	10	47,6%	47,6%
East Europe	10	10	71,4%	71,4%
Pacific/Asia	10	10	58,8%	58,8%
Africa	10	10	52,6%	52,6%
Middle East	10	10	58,8%	58,8%
Latn America	10	10	47,6%	47,6%

Plan File: C:\ESTAP\SPSS13\teste3.csplan