Universidade de São Paulo
Faculdade de Saúde Pública
Departamento de Epidemiologia

V Programa de Verão - 2002

Stata - *Básico*

Denise Pimentel Bergamaschi
José Maria Pacheco de Souza
Gizelton Pereira Alencar
Milena Baptista Bueno

Índice

1 - Iniciando o trabalho no *Stata*

Stata [Estata ou Esteita] - *Stata Corporation*

- *Intercooled Stata*
- *Versão resumida - Short Stata*
- Versão simplificada *StataQuest*

Existem versões do programa para 3 sistemas: *Windows*, *Unix* e *Macintosh*. Atualmente está na versão 7.

Este curso: Intercooled Stata versão 6 para sistema *Windows*.

.

O *Stata* é descrito em um manual com 5 volumes e em *Hamilton* (1998).

Cada comando está associado a um arquivo-**help** que pode ser acessado durante a utilização do programa.

Informações sobre o *Stata*, bem como atualizações, realização de cursos via *Internet* e lista das dúvidas mais freqüentes podem ser obtidas no *site*: http://www.stata.com.

O *Stata* possui suporte técnico e lista de discussão sobre dúvidas; endereço: http://www.hsph.harvard.edu/statalist.

Estas informações podem ser, também, obtidas pelo **help** disponível no menu principal.

O programa diferencia entre letra maiúscula e minúscula.

1.1 - Iniciando o Stata

Abrir o programa

- diretamente pelo ícone na tela de abertura do *Windows*, ou
- seguindo o caminho **Iniciar**, **Programas**, **Stata, Intercooled Stata 6**

Telas:

Título	Finalidade
Review	Armazenamento dos comandos já utilizados
Variables	Apresentação das variáveis contidas no banco de dados
Stata Results (fundo preto)	Apresentação dos resultados obtidos com a execução dos comandos
Stata Command	Digitação dos comandos a serem executados

janela Stata Command: digitar o comando quando o *prompt* estiver ativo. Executar pressionando a tecla **Enter.** O comando será armazenado na janela Review.

janela Review: o comando pode ser reutilizado e corrigido utilizando-se o mouse ou as teclas **PgUp** (*page up*) e **PgDn** (*page down*)

janela Stata Results: apresenta os resultados da execução do comando.

No *Stata* somente um arquivo de dados pode ser aberto e utilizado de cada vez .

janela Variables: apresenta as variáveis que compõem o banco de dados, matriz retangular onde as colunas representam as variáveis e as linhas as observações, para cada registro.

O *Stata* é basicamente um programa de comandos.

- Forma bem simples de um comando:

comando lista_de_variáveis (***command varlist***)

Ex: usando um banco de dados contendo as variáveis **x** e **y**

o comando para listá-las é : **list x y**

pode ser definida uma condição**: list x y if x>y**

A utilização do **Help** é fortemente recomendada;
clicando-se em **Help** no menu principal, pode-se pesquisar qualquer comando utilizando-se a opção **Contents**, **Search** ou **Stata command**.

O *Stata* trabalha com 4 tipos de arquivos:

tipo de arquivo	Extensão
arquivo que contém os dados	.dta
arquivo que guarda os comandos e resultados obtidos durante a sessão de trabalho	.log
arquivo que contém comandos	.do
arquivo que contém sub-rotinas	.ado

Logo que for iniciado o trabalho no *Stata*, é aconselhável abrir um arquivo **log,** que armazenará todos os comandos e seus resultados (com exceção de gráficos).
Para abrir um arquivo log: clicar sobre o quarto ícone (pergaminho).

O arquivo **log** é um arquivo de tipo somente texto e não permite alteração. Caso seja de interesse, pode-se transformá-lo em documento do *Word* com extensão .doc, para ser manipulado segundo a necessidade.

1.2 – Leitura e salvamento de banco de dados

Via linha de comando

O *Stata* possui seu próprio formato de banco de dados com extensão **.dta**. Para abrir e salvar um banco de dados já existente de nome **banco.dta:**

- **use banco**
- **use banco,clear**
- **use c:\cursosta\banco**

Para salvar um banco de dados de nome **banco.dta:**

- **save banco**

Para salvar o banco com o mesmo nome:

- **save banco,replace**

ou

- **save,replace**

Se os dados não estiverem no formato *Stata*: utilizar o *Stat/Transfer* ou outro pacote que realize conversão de bancos de dados.

Via caixa de diálogo (menu principal)

Pressionar o *mouse* sobre **File** seguido de **Open**. Seleciona-se o sub-diretório que contém o arquivo ***.dta***, marca-se o arquivo e seleciona-se **Open**.

Salvamento do banco: **Save** ou **Save As** na opção **File.**

Arquivos de dados em formato não dta

- **insheet using C:\cursosta\banco2.dat**
- **insheet using C:\cursosta\banco2.raw**

- **infile id nome datadiag tratamen pesoinic sexo using C:\cursosta\banco2.txt**

- **infile id nome _skip(2) pesoinic sexo using C:\cursosta\banco2.txt**

OBS: para a utilização do **infile**, deve-se eliminar a linha contendo o nome das variáveis no banco **.txt**.

1.3 – Criando banco de dados

Entrada de dados diretamente no Stata, pelo teclado

- **input [varlist]**.

Criar um banco de dados com nome **banco1** que contenha as variáveis id, nome, tratamen, pesoinic e sexo; para 5 pacientes, com dados apresentados a seguir.

id	nome	tratamen	pesoinic	sexo
1	A Silva	0	98.4	1
2	G Soares	1	75.5	2
3	V Gomes	1	93.6	2
4	M Costa	0	80.2	1
5	A Cardim	0	70.0	2

- **input id str10 nome tratamen pesoinic sexo**

```
        id      nome          tratamen          pesoinic          sexo
1. 1  "A Silva"   0   98.4  1
2. end
```

A Silva precisou usar aspas porque é um nome com duas palavras. Se fosse Asilva, não precisaria.

Abrir modo de edição clicando sobre o ícone **Data editor** e digitar os dados dos demais registros. Usar **Tab** para entrada horizontal e **Enter** para entrada vertical. Quando terminar, pressionar **Preserve** seguido de **Close** no menu do *Stata* editor.
O arquivo deve ser salvo utilizando a caixa de diálogo, na seqüência: **File**, **Save As**, Sub-diretório - **Cursosta**, nome do arquivo: **banco2.**

O arquivo pode ser salvo como arquivo ASCII com o comando **outfile**:

- **outfile using c:\cursosta\banco2.txt**

Criando arquivo ASCII, externamente ao *Stata*

O arquivo de dados pode ser construído utilizando um editor de texto (*Word for Windows, Wordpad, Notepad* . Os valores devem ser separados por tab, ou vírgula. Na primeira linha do banco pode-se digitar o nome das variáveis. Os valores faltantes devem ser substituídos por valores numéricos (-9, p.ex.). A extensão deve ser .txt, ou .dat. ou .raw. A leitura é com o comando **insheet**.

Outra forma no editor de texto, é entrar com os dados em colunas, separadas por espaço, sem tab ou vírgula. A leitura é com o comando **infile.**

Utilizando qualquer editor de texto, gerar o banco de dados **banco2.raw** (ou .dat ou .txt) onde a primeira linha contém o nome das variáveis e os dados são separados por Tab ou vírgula, para usar **insheet**.
Se o comando a ser utilizado for **infile**, o arquivo texto <u>não</u> deve conter o nome das variáveis.

banco2.raw

id→nome→tratamen→pesoinic→sexo

1→"A Silva"→0→98.4→1

2→"G Soares"→1→75.5→2

3→"V Gomes"→ 1→93.6→2

4→"M Costa"→ 0→80.2→1

5→"A Cardim"→0→70.0→2

→ simboliza o uso de **Tab**

banco2.dat

```
id, nome, tratamen, pesoinic, nome
1 ,   "A Silva" , 0 ,98.4, 1
2 , "G Soares" , 1, 75.5 ,2
3  ,  "V Gomes" , 1 ,93.6, 2
4   , "M Costa" , 0, 80.2, 1
5  , "A Cardim" , 0 ,70.0 ,2
```

banco2.txt

```
1      "A Silva"     0  98.4  1
2    "G Soares"      1  75.5  2
3    "V Gomes"     1  93.6  2
4     "M Costa"     0  80.2  1
5    "A Cardim"     0  70.0  2
```

- **insheet using C:\cursosta\banco2.raw, clear**
- **insheet using C:\cursosta\banco2.dat, clear**
- **infile id str10 nome tratamen pesoinic sexo using C:\cursosta\banco2.txt**

1.4 – Variáveis

Há dois tipos de variáveis no *Stata*: *string* (cadeia de caracteres, palavra) e numérica.

Estas variáveis são armazenadas de formas diferentes que requerem tamanhos diferentes nos registros de memória: *byte*, *int*, *long* e *float* para variáveis numéricas e *str1* até *str80* para variáveis *string* de tamanhos diferentes. Além disto, cada variável pode ter um nome associado a ela (rótulo, *label*) e tem um formato de apresentação

O nome da variável *x* pode ser mudado para *y* usando o comando

- **rename x y**
- **rename datainic datadiag**

O rótulo da variável pode ser definido com o comando

- **label variable x "custo em reais"**
- **label var pesoinic "peso inicial"**

O formato de uma variável numérica pode ser configurado para numérica geral (g) ou formato fixo (f) (com duas casas decimais, por ex.) utilizando

- **format x %7.2g**
- **format x %7.2f**

Variáveis numéricas

Valores faltantes (*missing*) são representados por pontos e são interpretados como valores muito grandes.

O código de valores faltantes pode ser convertido em valores:

- **mvdecode x,mv(-99)**

substitui todos os valores de *x* iguais a -99, para pontos (.)

- **mvencode x,mv(-99)**

substitui todos os valores de *x* iguais a ponto (.), para -99

Ex:

- **mvdecode pesoinic,mv(-99)**

Definição de rótulos para categorias de variáveis:

- **label define m 1 casado 2 divorciado 3 viuvo 4 solteiro**
- **label values marital m**

Ex:

- **label define s 1 "masculino" 2"feminino"**
- **label values sexo s**

Recodificação de variáveis:

- **recode marital 2 3 =2 4=3** ou
- **recode marital 2/3=2 4=3**

Ex:

- **recode sexo 1=0 2=1**

Variáveis *string*

Variáveis *string* são utilizadas para variáveis com categorias não numéricas, sob a forma de palavras, ou, genericamente, um conjunto de caracteres, com ou sem sentido de palavra.

Uma variável *string*, cujas categorias sejam representadas por caracteres numéricos, pode ser convertida em numérica com o comando:

- **gen varnovanumérica=real(varantigastring)**

Variáveis data

O *Stata* lê variáveis data como tempo decorrido (*elapsed dates)* ou **%d**, que é o número de dias contados a partir de 01 de janeiro de 1960. Assim,

0 corresponde a 01jan1960
1 corresponde a 02jan1960
.
.
.
15000 corresponde a 25jan2001

O *Stata* possui funções para converter datas em **%d**, para imprimir **%d** em formatos compreensíveis e para manipular variáveis **%d**.

Variáveis datas devem ser definidas como variáveis *string* e depois convertidas para **%d**.

No *Word for Windows*, digitar:

```
Id   nome datanasc
1    "A M"     "12/04/1947"
2    "J P"     "5/03/1955"
3    "M G"     "4/08/1957"
```

e salvar como texto: **nasc.txt**

No *Stata*,

- **insheet using c:\cursosta\nasc.txt**
- **list**

```
id       nome    datanasc
  1.        1        A M  12/04/1947
  2.        2        J P   5/03/1955
  3.        3        M G   4/08/1957
```

- **gen dianiver=date(datanasc,"dmy")**
- **list**

```
           id       nome    datanasc   dianiver
  1.        1        A M  12/04/1947      -4647
  2.        2        J P   5/03/1955      -1763
  3.        3        M G   4/08/1957       -880
```

- **desc**

```
Contains data
  obs:             3
 vars:             4
 size:            66 (100.0% of memory free)
-------------------------------------------------
   1. id        byte   %8.0g
   2. nome      str3   %9s
   3. datanasc  str10  %10s
   4. dianiver  float  %9.0g
-------------------------------------------------
Sorted by:
     Note:  dataset has changed since last saved
```

- **format dianiver %d**
- **list**

```
          id       nome    datanasc   dianiver
  1.       1        A M  12/04/1947  12apr1947
  2.       2        J P   5/03/1955  05mar1955
  3.       3        M G   4/08/1957  04aug1957
```

- **gen age2000=(mdy(1,1,2000)-dianiver)/365.25**
- **list**

```
          id       nome    datanasc   dianiver    age2000
  1.       1        A M  12/04/1947  12apr1947   52.72279
  2.       2        J P   5/03/1955  05mar1955   44.82683
  3.       3        M G  4/08/1957   04aug1957   42.40931
```

Uma variável *string* representando data pode ser mostrada como numérica usando a função **date("string1","string2")** onde string1 representa uma data e string2 uma permutação de "dmy" para especificar a ordem de dia, mês e ano na string1. Por exemplo:

- **display date("30/1/1930","dmy")**
- **display date("jan 30 1930"," mdy")**

Ambos retornam o valor –10958 que é o número de dias antes de 1/1/1960.

1.5 - Sintaxe

Os comandos seguem a forma

[by varlist:] command [varlist] [weight] [if exp] [in range] [using filename] [,options]

onde

[by varlist:] instrui *Stata* para repetir o comando para cada combinação de valores nas variáveis listadas em *varlist*;

command é o nome do comando, ex: **list**

[varlist] é a lista de variáveis para as quais o comando é executado

[weight] permite que pesos sejam associados às observações

[if exp] restringe o comando a um subconjunto de observações que satisfazem a expressão lógica definida em *exp*

[in range] restringe o comando àquelas observações cujos índices pertencem a um determinado subconjunto

[using filename] especifica o arquivo que deve ser utilizado

[,options] são específicas de cada comando.

2 - Manipulação de dados

2.1 - Expressões

Existem expressões lógicas, *string* e algébricas, no *Stata*.

Expressões lógicas atribuem 1 (verdadeiro) ou 0 (falso) e utiliza os operadores:

Operador		Significado
<		menor que
<=		menor ou igual a
>		maior que
>=		maior ou igual a
= =		igual a
~=	!=	diferente de
~		não
&		e
\| | | ou |

Ex: **if (y~=2 & z>x) | x= =1**

Significa: se (**y** for diferente de 2 e **z** maior do que **x**) ou **x** for igual a 1

Expressões algébricas utilizam os operadores:

Operador	Significado
+ -	soma, subtração
* /	multiplicação, divisão
^	elevado à potência
sqrt( )	função raiz quadrada
exp( )	função exponencial
log( )	função logarítmica (base 10)
ln( )	função logarítmica (base *e*) - logaritmo natural

2.2 - Observações índice e conjunto de valores

Observações índice

Cada observação está associada a um índice. Por exemplo, o terceiro valor da variável *x* pode ser especificado como **x[3]**. O macro **_n** assume um valor para cada observação e **_N** é igual ao número total de observações. Pode-se referir à penúltima observação da variável *x* escrevendo-se **x[_n-1]**.

Uma variável indexada deve ficar do lado direito de uma asserção. Por exemplo, para substituir a terceira observação da variável *x* pelo valor 2 escreve-se:

- **replace x=2 if _n= =3**

Conjunto de valores

Um conjunto de valores pode ser especificado utilizando-se **if** e **_n** ou utilizando **in range** que possui a sintaxe **f/l (f para *first* e l {letra ele} para *last*)**. Por exemplo, para listar as últimas 10 observações, utiliza-se o comando:

- **list x in –10/l**

Para repetir comandos para variáveis ou categorias de variáveis, utilizar **by varlist;** os dados precisam estar ordenados antes disto, o que é feito utilizando o comando **sort**.

- **sort tratinic**
- **by tratinic: list nome**

2.3 - Gerando variáveis

O comando **generate** iguala uma nova variável a uma expressão que é construída para cada observação

- **generate percent=100*(old-new)/old if old >0**

gera uma nova variável **percent** que pode assumir valor faltante se **old** for um valor faltante ou será igual ao percentual de diminuição de **old** para **new** para cada observação onde **old** é positiva.

O comando **replace** funciona como o comando **generate,** com a diferença que permite que uma variável já existente seja alterada.

- **replace percent =0 if old<=0**

muda os valores faltantes em **percent** para zeros.

Os dois comandos anteriores podem ser escritos em um somente

- **generate percent=cond(old>0, 100*(old-new)/old,0)**

cond faz com que o segundo argumento seja calculado se o primeiro argumento for verdade. Caso contrário executa o terceiro argumento.

Gerando variáveis indicadoras (*dummy*):

Supor a variável glicemia categorizada em (<150mg/l= 0; 150 - 200 mg/l = 1 e >=200mg/l= 2).

- **tab glicemia,gen(gliced)**

gera 3 variáveis *dummy*: gliced1, gliced2 e gliced3

Comando **egen:**

O comando **egen** pode ser função de muitas variáveis simultaneamente.

- **egen average=rmean(m1-m100)**

calcula a média, para cada linha de registro, das 100 variáveis **m1** até **m100**, sendo que os valores faltantes são ignorados. **rmean** trabalha nas linhas**.**

- **egen famsal=mean(salario),by(familia)**

calcula a média da variável **salario** para o conjunto de valores iguais de familia**.** **mean** trabalha na coluna da variável.

Uma variável existente pode ser retirada do banco de dados com o comando **drop.**

- **drop famsal**

Pode-se utilizar, também, o comando **keep varlist**, onde **varlist** é a lista de variáveis que devem permanecer no banco de dados.

2.4 - Mudando a forma de apresentação dos dados

Supor a situação na qual, para um mesmo indivíduo, são obtidas duas ou mais informações, apresentadas no banco de dados c:\curso\repetew.dta.
Os dados estão apresentados como segue, em formato **wide**.

- **use C:\curso\repetew.dta**
- **list**

```
        indiv          x1          x2
 1.         1           2           3
 2.         2           4           5
```

A forma de apresentação dos dados pode ser mudada para o formato **long**, utilizando o comando

- **reshape long x ,i(indiv) j(occ)**
- **list**

```
        indiv         occ           x
 1.         1           1           2
 2.         1           2           3
 3.         2           1           4
 4.         2           2           5
```

e pode ser revertido para a forma anterior (**wide**)

- **reshape wide x ,i(indiv) j(occ)**
- **list**

```
        indiv          x1          x2
 1.         1           2           3
 2.         2           4           5
```

Para os dados em formato **long** pode ser necessário calcular, para cada indivíduo, a média das medidas repetidas (*meanx*), o desvio padrão (*sdx*) e o número de observações repetidas diferentes de *missing* (*num*).

- **use C:\curso\repetew.dta**
- **reshape long x ,i(indiv) j(occ)**
- **preserve**
- **collapse (mean) meanx=x (sd) sd=x (count) num=x, by(indiv)**
- **list meanx sdx num**

```
        meanx        sdx        num
 1.       2.5   .7071068          2
 2.       4.5   .7071068          2
```

- **restore**
- **reshape wide x ,i(indiv) j(occ)**
- **list**

```
        indiv         x1         x2
 1.         1          2          3
 2.         2          4          5
```

Mas, também, direto, com mais comandos**:**

- **egen meanx=rmean(x1 x2)**
- **egen sdx=rsd(x1 x2)**

3. Descrição de dados

3.1 - Gráficos

A sintaxe básica para a elaboração de gráficos é:

- **graph varlist, options**

Em **options** deve-se especificar o tipo de gráfico desejado.

Os gráficos não aparecem no arquivo log. Deve-se abrir um arquivo .doc previamente; obtido o gráfico, clicar em **copy graph** na barra do Stata e depois **colar** no doc.

Boxplot

- **graph x, box**

produz um boxplot da variável x

- **graph x y, box**

cria dois boxplots, um para x e outro para y, em um conjunto de eixos ortogonais.

- **by group: graph x,box**

fornece um boxplot para cada categoria de group, em dois conjuntos de eixos ortogonais independentes.

- **graph x,by(group) box**

cria boxplots, um para cada categoria de group, em um mesmo par de eixos

Diagrama de dispersão

- **graph x y**

fornece um diagrama de dispersão de x e y

- **graph x y z,twoway**

fornece um diagrama de dispersão de x e y contra z em um par de eixos

- **graph x y z,twoway s(io) c(l.) xlabel ylabel t1("diagrama de dispersão")**

s(io) faz com que os pontos em x fiquem invisíveis e os pontos em y fiquem representados por pequenos círculos (o). Neste caso está sendo utilizada a opção **symbol()**.

c(l.) faz com que os pontos em x sejam conectados por linhas retas e os de y não sejam conectados. Aqui está sendo utilizada a opção **connect()**.

As opções **xlabel** e **ylabel** fazem com que os eixos *X* e *Y* sejam rotulados utilizando valores redondos (sem estas opções serão apresentados somente os valores mínimo e máximo).

A opção **t1("diagrama de dispersão")** faz com que seja apresentado um título principal no topo do gráfico. **b1()**, **l1()** e **r1()** produzem títulos principais na base, na esquerda e direita. **t2()**, **b2()**, **l2()** e **r2()** produzem títulos secundários em cada um dos lados.

O gráfico deve ser produzido em um único comando; assim, se diferentes símbolos forem utilizados para diferentes grupos, em um diagrama de dispersão, cada grupo deve ser representado por variáveis separadas (sem valores faltantes e somente para observações pertencentes àquele grupo).

- **gen y1=y if group==1**
- **gen y2=y if group==2**
- **graph y1 y2 x,s(dp)**

Produz um diagrama de dispersão onde **y** é representado por losangos **(d)** no grupo 1 e sinal de mais (plus) **(p)** no grupo2.

A variável que representa o eixo X deve ser ordenada antes:

- **sort _varX**

Histograma

Para desenhar um histograma utilizar o comando: **graph x, options**

- **graph x**

desenha um histograma da variável x.

- **graph x, bin(10)**

desenha um histograma da variável **x** em 10 intervalos de classe. O número de intervalos pode variar, de acordo com os dados.

- **graph x, bin(10) norm**

desenha um histograma da variável x com 10 intervalos de classe e sobrepõe uma curva normal com a média e o desvio padrão observados.

- **graph x, bin(10) norm(média desviopadrão)**

desenha um histograma da variável x com 10 intervalos de classe e sobrepõe uma curva normal com média e desvio padrão definidos.

- **graph x, bin(10) xlabel ylabel t1("distribuição da variável x")**

desenha um histograma da variável x com 10 intervalos de classe, apresenta os rótulos dos eixos e o título, no topo do gráfico.

- **graph x, bin(10) xlabel ylabel by(y)**

desenha um histograma da variável x, com 10 intervalos de classe, para cada categoria da variável y.

3.2 – Tabelas e resumo dos dados

Os dados que serão utilizados nesta sessão constituem uma amostra de 118 pacientes psiquiátricos, do sexo feminino e estão disponíveis em D.J. Hand et al. ***A Handbook of Small Data Sets***. Chapman &Hall, London,1994. As variáveis estudadas foram:

- **age**: idade em anos
- **iq**: escore de inteligência (-99 = ignorado)
- **anxiety**: ansiedade (1= nenhuma, 2= leve, 3= moderada, 4=severa, -99=ignorado)
- **depress**: depressão (1=nenhuma, 2= leve, 3= moderada, 4=severa, -99=ignorado)
- **sleep**: você pode dormir normalmente? (1=sim, 2=não, -99=ignorado)
- **sex**: você perdeu interesse em sexo? (1=não, 2=sim)
- **life**: você tem pensado recentemente em acabar com sua vida? (1=não, 2=sim)
- **weight**: mudança no peso durante os últimos 6 meses (em libras)

Objetivo	Comandos
abrir o banco de dados	insheet using c:\cursosta\fem.dat,clear
verificar quais são as variáveis que compõem o banco de dados	describe ou desc
construir uma tabela de freqüências simples de cada variável	tab1 _all ou, tab age tab life

Objetivo	Comandos
remover os valores faltantes	mvdecode _all,mv(-99) ou removendo os valores faltantes de cada variável: recode sleep –99=.
recodificar a variável **sleep**, para ficar consistente com o restante dos códigos (1=não e 2=sim)	recode sleep 1=2 2=1
fornecer rótulos (*labels*) para as variáveis	label define sn 1 nao 2 sim label values sex sn label val sleep sn label val life sn ou, em um único comando: for var sex life sleep: label values X sn
Fornecer um resumo da variável **iq**	summ iq ou summ iq,d
Fornecer um resumo da variável **iq** segundo life	sort life by life: summ iq,d
comparar as médias e desvios padrão de **iq** segundo life	table life,contents(mean iq sd iq)
fornecer um rótulo para a variável **weight**	label variable weight “mudanca de peso nos ultimos 6 meses”
fornecer rótulo para a variável **life**	label variable life “voce pesnsou em terminar sua life recentemente?”
fazer o gráfico boxplot da variável **weight** segundo **life**	graph weight,box by(life) b1(“voce pensou recentemente em terminar sua vida?”)

Objetivo	Comandos
fazer o gráfico qq-plot para verificar normalidade da distribuição da variável **weight**	qnorm weight, gap(5) xlab ylab t1("qq plot para normalidade") onde, gap(5) é usado para diminuir o espaço entre o eixo vertical e o título do eixo
Desenhar um histograma da variável **weight** em 6 intervalos de classe.	graph weight, bin(6) xlabel(-5, -2.5, 0, 2.5, 5, 7.5) ylabel t1("distribuição de perda de peso nos ultimos 6 meses")
Desenhar um histograma da variável **weigh**t em 6 intervalos de classe, segundo a variável **life**	graph weight, bin(6) xlabel(-5, -2.5, 0, 2.5, 5, 7.5) ylabel by(life)
Criar uma variável **ageg** contendo a variável age em intervalos de classes de 5 anos	gen ageg=age recode ageg 25/29=1 30/34=2 35/39=3 40/44=4 45/49=5 label define id 1 "25-29" 2 "30-34" 3 "35-39" 4 "40-44" 5 "45-49" label val ageg id tab ageg

4. Análise de dados epidemiológicos

Banco de dados: c:\cursosta\fem2 .dta

Comparação de médias:

Para comparar as variáveis quantitativas pode-se utilizar o teste *t de "Student"* que assume que as observações nos dois grupos são independentes; as amostras foram retiradas de populações com distribuição normal, com mesma variância. Um teste alternativo, não paramétrico, que não necessita destas pressuposições, é o teste *U* de *Man-Whitney*.

Coeficiente de correlação:

Também é possível calcular correlações entre variáveis contínuas. Se se quiser testar se o coeficiente de correlação de *Pearson* é estatisticamente diferente de zero, o Stata apresenta um teste que pressupõe que as variáveis são normais bivariadas. Se esta pressuposição não for feita, pode-se utilizar a correlação de postos de *Spearman*. Se as variáveis forem categóricas é possível utilizar a estatística de *Kendall* como medida de associação.

Associação entre variáveis:

Para as variáveis qualitativas nominais pode-se utilizar o teste qui-quadrado, de *Pearson*.

4.1 – Teste de hipóteses para uma e duas médias e intervalos de confiança

Objetivo	Comandos
Testar a diferença entre as variâncias da variável **weight** segundo **life**	sdtest weight,by(life)
Testar se existe diferença entre a mudança média de peso nos dois grupos da variável **life**	ttest weight,by(life)
Apresentar o intervalo de confiança para as médias de **weight** segundo **life**	sort life ci weight,by(life)
Construir o intervalo de confiança de 95% para uma amostra de 100 pessoas, média observada igual a 2 e desvio padrão populacional igual a 2,5	cii 100 2 2,5
Testar a hipótese de que a média observada da variável **weight** ($\bar{x}_{obs} = 1{,}585$) é igual à média populacional ($\mu = 2$)	ttest weight=2

4.2 – Teste de hipóteses para associação e intervalo de confiança para proporção

objetivo	Comandos
Construir um intervalo de confiança (exato) para a proporção de pacientes que pensaram em terminar sua vida	tab life cii 117 65
Testar a hipótese de que a proporção de pacientes que pensaram em terminar suas vidas é igual a 0,5	recode life 2=1 1=0 bitest life=0.5 Ou bitesti 117 65 0.5
Verificar a existência de associação entre as variáveis **depres** e **life**	tab life depres,col chi2 OBS: teste exato de Fisher tab life depress,exact
Verificar a existência de associação entre as variáveis **sex** e **life**, apresentando o teste χ^2 e o teste exato de Fisher	tab life sex,row chi2 exact

4.3 – Teste de hipóteses para correlação

objetivo	comandos
Calcular a correlação entre as variáveis **weight**, **iq** e **age**	corr weight iq age Se o número de pares de observações for diferente para cada conjunto de duas variáveis, utilizar pwcorr weight iq age,obs sig
Calcular a associação entre as variáveis **depres** e **anxiety**	ktau depress anxiety

Exercício suplementar no capítulo de exercícios, como exercício 2.

4.4 - Estimação

Todos os comandos de estimação, por exemplo, **regress**, **logistic**, **poisson**, seguem a mesma estrutura em sua sintaxe:

[xi:] command depvar [model] [weights],options

que pode ser combinado com **by varlist:**, **if exp** e **in range**. A variável resposta é especificada por **depvar** e as variáveis explanatórias, pelo **model**o.

- **regress resp x**

ajusta um modelo de regressão de resp (variável contínua) em x

- **tab y,gen(z)**
- **regress resp z2 z3**

constrói variáveis *dummy* para a variável y, representada em 3 categorias; ajusta um modelo de regressão de resp em z2 e z3**,** tendo z1 como basal (variáveis dummy).

Alternativamente, pode-se optar por utilizar o comando **xi:** no começo do comando, que faz com que variáveis *dummy* sejam criadas e adicionadas ao modelo

- **xi: regress resp i.z**

ajusta um modelo de regressão de resp em z2 e z3.

4.5 – Análise de medidas de efeito

Nesta sessão será utilizado o banco de dados originário de um ensaio clínico onde pacientes com câncer de pulmão foram alocados aleatoriamente para receber dois tipos diferentes de quimioterapia (terapia seqüencial e alternada). A variável resposta foi classificada em 4 categorias: doença progressiva, sem mudança, remissão parcial e remissão completa. Os dados foram publicados por Holtbrugge e Schumacher (1991). A análise principal será avaliar as duas terapias.

Distribuição de pacientes com câncer de pulmão segundo sexo , tipo de terapia e resultado do tratamento

Terapia	Sexo	Doença Progressiva	Sem Mudança	Remissão Parcial	Remissão completa
seqüencial	Masculino	28	45	29	26
	Feminino	4	12	5	2
alternada	Masculino	41	44	20	20
	Feminino	12	7	3	1

.

- **infile fr1 fr2 fr3 fr4 using C:\cursosta\bancos\tumour.dat**

Gerando as variáveis indicadoras terapia e sexo:

- **gen terapia=int((_n-1)/2)**
- **sort terapia**
- **by terapia:gen sex=_n**
- **label define t 0 “seq” 1 “alt”**
- **label values terapia t**
- **label define s 1 “masculino” 2 “feminino”**
- **label values sexo s**

Transformando o banco no formato long:

- **reshape long fr,i(terapia sexo) j(outc)**

Verificando se deu certo:

- **table sexo out [freq=fr], by(terapia) row col** ou
- **table sexo out terapia [freq=fr], row col scol**

terapia and sexo	outc 1	2	3	4
seq				
masculino	28	45	29	26
feminino	4	12	5	2
alt				
masculino	41	44	20	20
feminino	12	7	3	1

	terapia and outc													
	seq					alt					Total			
sex	1	2	3	4	Total	1	2	3	4	Total	1	2	3	4	Total
masculino	28	45	29	26	128	41	44	20	20	125	69	89	49	46	253
feminino	4	12	5	2	23	12	7	3	1	23	16	19	8	3	46
Total	32	57	34	28	151	53	51	23	21	148	85	108	57	49	299

Expandindo o banco, com cada um dos 299 indivíduos tendo seu próprio registro com suas respectivas variáveis:

- **expand fr**

Repetindo os comandos anteriores de tabela para verificar que os resultados são os mesmos:

- **table sexo out [freq=fr], by(terapia) row col** ou
- **table sexo out terapia [freq=fr], row col scol**

Transformando a variável resposta em uma variável dicotômica:

- **gen melhora=outc**
- **recode melhora 1/2 = 0 3/4 = 1**

Calculando os odds de melhora segundo terapia:

- **tabodds melhora terapia**

```
------------+------------------------------------------------------------
   terapia  |      cases     controls       odds      [95% Conf. Interval]
------------+------------------------------------------------------------
       seq  |         62           89    0.69663        0.50372   0.96341
       alt  |         44          104    0.42308        0.29740   0.60187
------------+------------------------------------------------------------
Test of homogeneity (equal odds): chi2(1)  =     4.18
                                  Pr>chi2  =   0.0409

Score test for trend of odds:     chi2(1)  =     4.18
                                  Pr>chi2  =   0.0409
```

Cuidado! O programa considera caso o valor 1 e controle o valor 0, portanto melhora=1 = caso e piora= 0= controle.

Calculando a odds ratio:

- **mhodds melhora terapia**

```
Maximum likelihood estimate of the odds ratio
Comparing terapia==1 vs terapia==0
------------------------------------------------------------------
Odds ratio    chi2(1)        P>chi2        [95% Conf. Interval]
------------------------------------------------------------------
.607320         4.18        0.0409         0.374628   0.984544
------------------------------------------------------------------
```

Lembrando: terapia 1= seqüencial e terapia 0= alternada.

Análise estratificada:

- **cc melhora terapia**

```
                | terapia                |             Proportion
                |   Exposed   Unexposed  |     Total     Exposed
----------------+------------------------+----------------------
          Cases |        44          62  |       106      0.4151
       Controls |       104          89  |       193      0.5389
----------------+------------------------+----------------------
          Total |       148         151  |       299      0.4950
                |                        |
                |      Point estimate    |  [95% Conf. Interval]
                |------------------------+----------------------
     Odds ratio |         .6073201       |  .3767754    .9790055 (Cornfield)
 Prev. frac. ex. |         .3926799       |  .0209945    .6232246 (Cornfield)
 Prev. frac. pop |         .2115995       |
                +-------------------------------------------
```

Reforçando a lembrança: caso=melhora, controle= piora, exposto=alternado, não exposto= seqüencial**.**

Regressão logística , comandos logit |logistic

- **logit melhora terapia**

```
Iteration 0:   log likelihood = -194.40888
Iteration 1:   log likelihood = -192.30753
Iteration 2:   log likelihood = -192.30471

Logit estimates                                   Number of obs   =        299
                                                  LR chi2(1)      =       4.21
                                                  Prob > chi2     =     0.0402
Log likelihood = -192.30471                       Pseudo R2       =     0.0108

------------------------------------------------------------------------------
 melhora |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
 terapia |  -.4986993   .2443508     -2.041   0.041       -.977618   -.0197805
   _cons |   -.361502   .1654236     -2.185   0.029      -.6857263   -.0372777
------------------------------------------------------------------------------
```

O algoritmo precisa de 3 iterações para convergir. O coeficiente de terapia representa a diferença no *log odds* (de uma melhora) entre as terapias alternada e seqüencial. O valor negativo indica que a terapia seqüencial é superior à terapia alternada. O valor de *p* associado à estatística *z* do teste de *Wald* é 0,041. A estatística *z* é igual ao coeficiente dividido pelo erro padrão. Este valor de *p* é assintoticamente igual ao valor de *p* derivado do teste da razão de verossimilhança entre o modelo incluindo somente a constante e o modelo incluindo a variável terapia (chi2(1)=4,21). -2 vezes o logaritmo da razão de verossimilhança é igual a 4,21 com distribuição aproximada qui quadrado, com 1 grau de liberdade, com valor *p*= 0,040.

- **logit melhora terapia,or**

```
Logit estimates                           Number of obs   =        299
                                          LR chi2(1)      =       4.21
                                          Prob > chi2     =     0.0402
Log likelihood = -192.30471               Pseudo R2       =     0.0108

------------------------------------------------------------------------------
 melhora | Odds Ratio   Std. Err.  z    P>|z|   [95% Conf. Interval]
---------+--------------------------------------------------------------------
 terapia | .6073201   .1483991  -2.041  0.041   .3762061    .9804138
------------------------------------------------------------------------------
```

- **logistic melhora terapia**

```
Logit estimates                                   Number of obs   =        299
                                                  LR chi2(1)      =       4.21
                                                  Prob > chi2     =     0.0402
Log likelihood = -192.30471                       Pseudo R2       =     0.0108

------------------------------------------------------------------------------
 melhora | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
 terapia |   .6073201   .1483991    -2.041   0.041       .3762061    .9804138
------------------------------------------------------------------------------
```

- **logistic melhora terapia sex**

```
Logit estimates                                   Number of obs   =        299
                                                  LR chi2(2)      =       7.55
                                                  Prob > chi2     =     0.0229
Log likelihood = -190.63171                       Pseudo R2       =     0.0194

------------------------------------------------------------------------------
 melhora | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
 terapia |   .6051969   .1486907    -2.044   0.041       .3739084    .9795537
     sex |   .5197993   .1930918    -1.761   0.078       .2509785    1.076551
------------------------------------------------------------------------------
```

- **lrtest,saving(2)**

- **logistic melhora terapia**

```
Logit estimates                                   Number of obs   =        299
                                                  LR chi2(1)      =       4.21
                                                  Prob > chi2     =     0.0402
Log likelihood = -192.30471                       Pseudo R2       =     0.0108

------------------------------------------------------------------------------
 melhora | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
 terapia |   .6073201   .1483991    -2.041   0.041       .3762061    .9804138
------------------------------------------------------------------------------
```

- **lrtest,using(1)**

- **lrtest, using(1) model(2)**

```
Logistic:  likelihood-ratio test                  chi2(1)     =       3.35
                                                  Prob > chi2 =     0.0674
```

Modelo linear generalizado (glm)

- **glm melhora terapia, family(binomial) link(logit) eform**

```
Iteration 1 : deviance =  385.2854
Iteration 2 : deviance =  384.6098
Iteration 3 : deviance =  384.6094
Iteration 4 : deviance =  384.6094

Residual df =       297                     No. of obs =         299
Pearson X2  =  298.9998                     Deviance   =  384.6094
Dispersion  =  1.006733                     Dispersion =  1.294981

Bernoulli distribution, logit link
------------------------------------------------------------------------------
     melhora | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     terapia |   .6073201   .1483995    -2.04   0.041     .3762057     .980415
------------------------------------------------------------------------------
```

- **glm melhora terapia, family(binomial) link(log) eform**

```
Iteration 1 : deviance =  534.9433
Iteration 2 : deviance =  391.1093
Iteration 3 : deviance =  384.6570
Iteration 4 : deviance =  384.6094
Iteration 5 : deviance =  384.6094
Iteration 6 : deviance =  384.6094

Residual df =       297                     No. of obs =         299
Pearson X2  =       299                     Deviance   =  384.6094
Dispersion  =  1.006734                     Dispersion =  1.294981

Bernoulli distribution, log link
------------------------------------------------------------------------------
     melhora |      e^coef   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     terapia |   .7240628   .1155715    -2.02   0.043     .5295555    .9900132
------------------------------------------------------------------------------
```

5- Análise de sobrevida

Pacientes com dependência a heroína, internados em uma clínica de tratamento com metadona. O evento de interesse é abandono do tratamento. Os pacientes ainda internados no término do estudo estão registrados na variável **status** (1 se o paciente abandonou o tratamento, 0 caso contrário). As variáveis explanatórias para a saída do tratamento são dose máxima de metadona, detenção prisional e clínica onde foi internado. Estes dados foram coletados e analisados por Caplehorn e Bell (1991). Variáveis estudadas:

id: identificação do paciente

clinic: clínica de internação (1, 2)

status: variável de censura (1 - abandono, 0 - em tratamento)

time: tempo de tratamento

prison: tem registro de encarceramento (1) ou não (0)

dose: dose máxima de metadona

Os dados estão disponíveis no banco C:\cursosta\heroina

5.1 - Apresentação dos dados

Declarando os dados como sendo na foram "st" (survival time)

- **stset time, failure(status)**

```
     failure event:  status ~= 0 & status ~= .
obs. time interval:  (0, time]
 exit on or before:  failure

--------------------------------------------------------------------------
      238  total obs.
        0  exclusions
--------------------------------------------------------------------------
      238  obs. remaining, representing
      150  failures in single record/single failure data
    95812  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =      1076
```

Resumindo os dados

stsum

```
failure _d:  status
   analysis time _t:  time

         |               incidence       no. of    |------ Survival time -----|
         | time at risk     rate        subjects        25%       50%       75%
---------+-------------------------------------------------------------------
   total |        95812   .0015656           238        212       504       821
```

São 238 pacientes, com tempo mediano de "sobrevida" de 504 dias. Se a taxa de incidência (hazard ratio) for constante, é estimada como 0,0016 abandonos por dia, que corresponde a 150 abandonos/95812 dias.

Pode-se realizar a análise para cada clínica:

- **strate clinic**

```
          failure _d:  status
   analysis time _t:  time

Estimated rates and lower/upper bounds of 95% confidence intervals
(238 records included in the analysis)

   clinic     _D         _Y      _Rate     _Lower     _Upper
        1    122      59558  0.0020484  0.0017154  0.0024462
        2     28      36254  0.0007723  0.0005333  0.0011186
```

Calculando o hazard ratio:

```
. display 0.0020484/0.0007723
2.6523372
```

ou

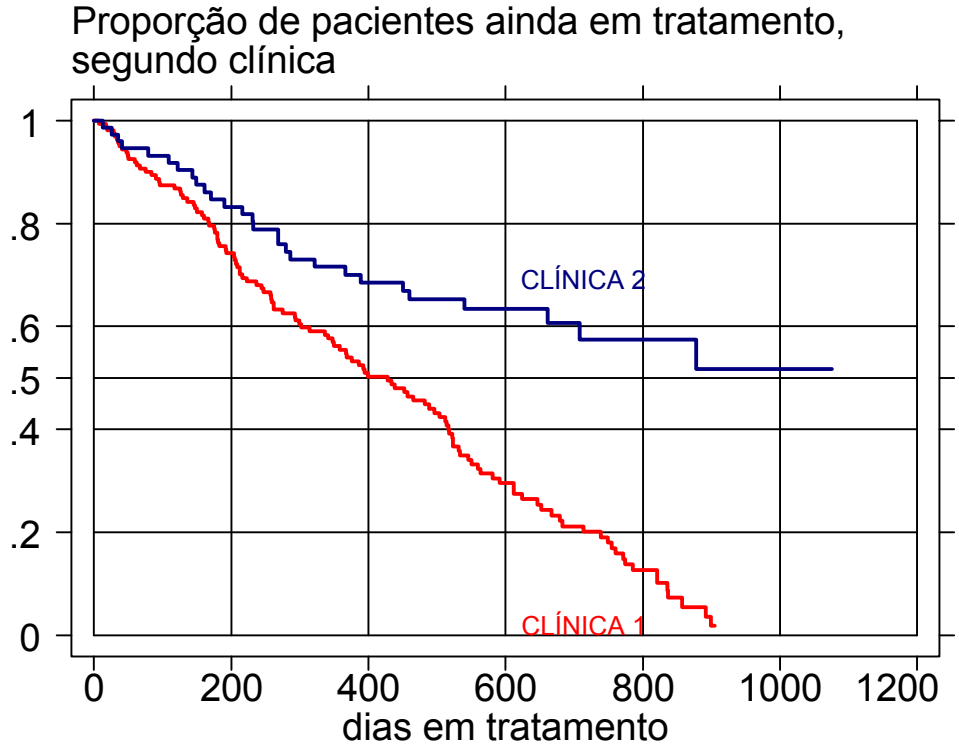
- **stsum,by(clinic)**

```
         failure _d:  status
   analysis time _t:  time

         |               incidence       no. of    |------ Survival time -----|
clinic   | time at risk     rate        subjects        25%       50%       75%
---------+-------------------------------------------------------------------
       1 |        59558   .0020484           163        192       428       652
       2 |        36254   .0007723            75        280         .         .
---------+-------------------------------------------------------------------
   total |        95812   .0015656           238        212       504       821
```

5.2- Curvas Kaplan-Meier

Construindo gráficos das curvas Kaplan-Meier

- **set textsize 150**
- **sts graph, by(clinic) xlabel(0 200 400 600 800 1000 1200) xline(0 200 400 600 800 1000 1200) ylabel(0 .2 .4 .5 .6 .8 1) yline(0 .2 .4 .5 .6 .8 1) b2(dias em tratamento)t1(Proporção de pacientes ainda em tratamento,) t2(segundo clínica)**



Realizando o teste para igualdade das funções de sobrevida:

- **sts test clinic**

```
         failure _d:  status
   analysis time _t:  time


Log-rank test for equality of survivor functions (teste Mantel-Cox)
-------------------------------------------------

       |  Events
clinic |  observed       expected
-------+-------------------------
1      |       122          90.91
2      |        28          59.09
-------+-------------------------
Total  |       150         150.00

             chi2(1) =      27.89
             Pr>chi2 =     0.000
```

- **stcox clinic**

```
failure _d:  status
  analysis time _t:  time

Iteration 0:   log likelihood =  -705.6619
Iteration 1:   log likelihood = -690.57156
Iteration 2:   log likelihood = -690.20742
Iteration 3:   log likelihood = -690.20658
Refining estimates:
Iteration 0:   log likelihood = -690.20658

Cox regression -- Breslow method for ties

No. of subjects =        238                Number of obs   =      238
No. of failures =        150
Time at risk    =      95812
                                  LR chi2(1)      =    30.91
Log likelihood  =   -690.20658                Prob > chi2     =   0.0000

------------------------------------------------------------------------------
          _t |
          _d | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      clinic |   .3416238   .0726424    -5.05   0.000     .2251904    .5182585
------------------------------------------------------------------------------
```

5.3 - Modelo de Cox (utilizando clinicas como estrato e as outras variáveis como explanatórias)

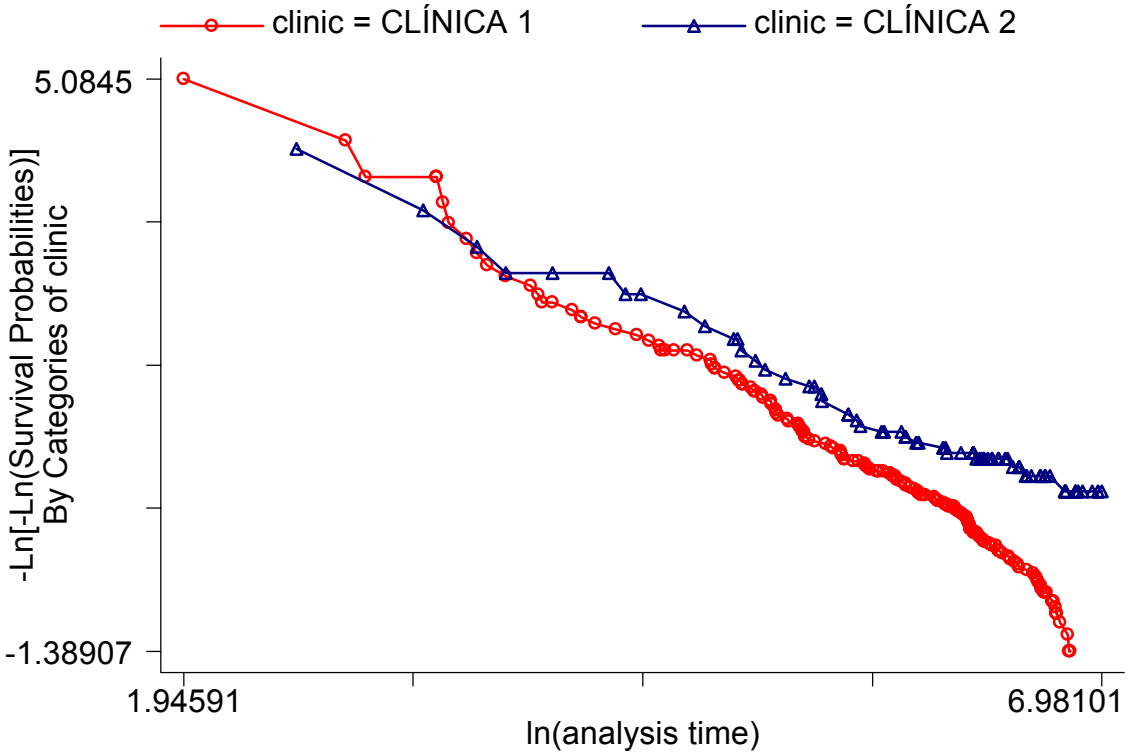
- **stcox dose prison,strata(clinic)**

```
. stcox dose prison,strata(clinic)

         failure _d:  status
   analysis time _t:  time

Iteration 0:   log likelihood = -614.68365
Iteration 1:   log likelihood = -597.73516
Iteration 2:   log likelihood =   -597.714
Refining estimates:
Iteration 0:   log likelihood =   -597.714

Stratified Cox regr. -- Breslow method for ties

No. of subjects =          238                     Number of obs   =       238
No. of failures =          150
Time at risk    =        95812
                                                   LR chi2(2)      =     33.94
Log likelihood  =     -597.714                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
      _t |
      _d | Haz. Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
    dose |   .9654655   .0062418     -5.436   0.000        .953309     .977777
  prison |   1.475192   .2491827      2.302   0.021       1.059418    2.054138
------------------------------------------------------------------------------
                                                         Stratified by clinic
```

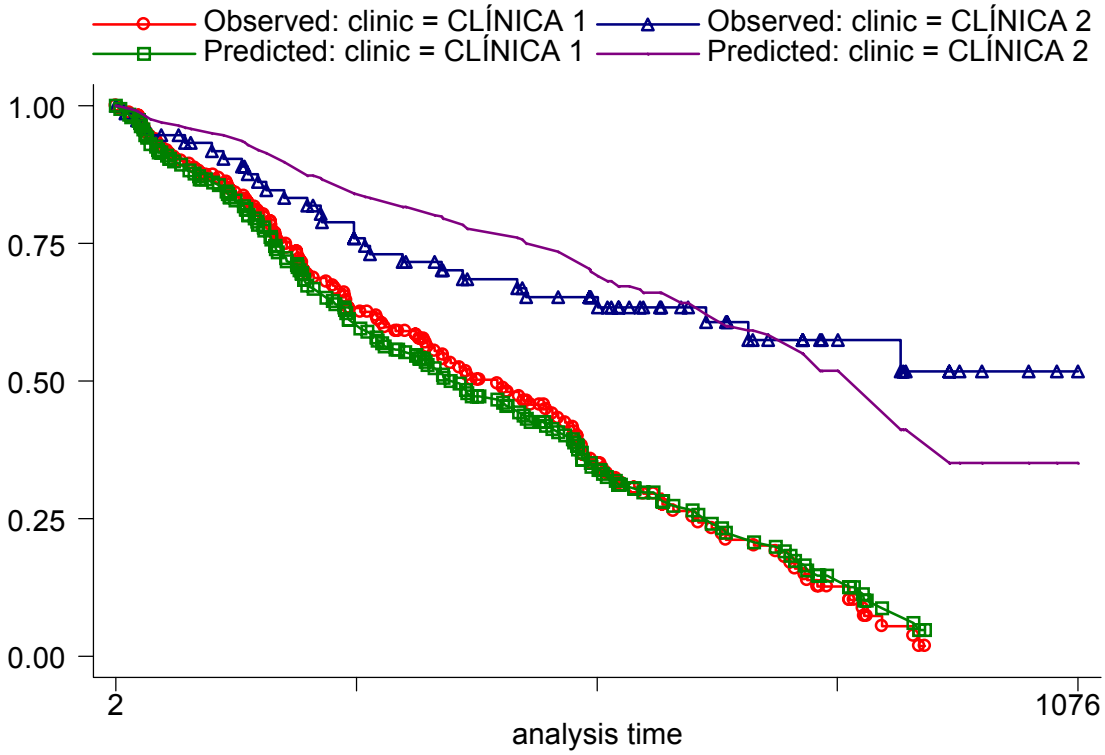
Pacientes com história de prisão tendem a abandonar o tratamento mais rapidamente do que aqueles sem história de prisão. Para cada aumento de uma unidade (1 mg) na dose de metadona, o *hazard* é multiplicado por 0,965, ou seja, maior dose de metadona implica maior tempo no tratamento. Pacientes da clínica ficam mais tempo em tratamento.

Uma questão importante é se o modelo de *hazards* proporcionais de Cox não é infligido quando da comparação entre as clínicas ou da comparação entre prisioneiros e não prisioneiros. A *hazards* ratio deve ser constante no tempo.

- stphplot, by(clinic)



- stcoxkm, by(clinic)



A análise visual indica que a proporcionalidade não se mantém no tempo.

6- Comandos gerais

6.1 – Stata como calculadora

- **display exp**
- **display sqrt(5*(11-3^2))**

```
3.1622777
```

6.2- Cálculo de tamanho de amostra

Para identificar diferença entre duas médias de amostras independentes utilizando o teste t de "Student" bicaudal, com poder de 80% para detectar uma diferença de 1 com nível de significância de 1%, com desvios padrão iguais a 1.

- **sampsi 1 2,sd(1) power (.8) alpha(0.01)**

```
Estimated sample size for two-sample comparison of means

Test Ho: m1 = m2, where m1 is the mean in population 1
                  and m2 is the mean in population 2
Assumptions:

         alpha =   0.0100  (two-sided)
         power =   0.8000
            m1 =        1
            m2 =        2
           sd1 =        1
           sd2 =        1
         n2/n1 =     1.00

Estimated required sample sizes:

            n1 =       24
            n2 =       24
```

6.3 –Guardando resultados em um macro

- **local a=exp**

que pode ser utilizado novamente com o nome da macro entre aspas.

- **local a=5**
- **display sqrt(`a')**

```
2.236068
```

6.4 – Breve introdução a arquivos *.do

Às vezes é necessário realizar uma análise igual para conjuntos de dados diferentes. Isto é possível, armazenando-se os comandos em um arquivo com extensão **.do**, por exemplo, analise.do, que pode ser executado com o comando:

- **do analise**

Uma forma de criar um arquivo *.do é salvando os comandos utilizados durante a sessão de trabalho. Isto pode ser feito selecionando "**save review contents**" do menu da janela "**Review**". Qualquer processador de texto pode ser utilizado para a correção dos comandos, lembrando que o arquivo *.do é texto, em ASCII. A seguir é apresentada uma estrutura básica de um arquivo *.do:

```
/*comentário descrevendo o que o arquivo faz*/
version 6.0
capture log close
log using filename,replace
set more off
command 1
command 2
.
.
log close
exit
```

Onde cada linha significa:

1. as barras e asterisco fazem com que seja ignorado o que está entre eles; são usados para comentários. Também pode ser utilizado simplesmente o asterisco.
2. O comando especificando a versão é útil porque o *Stata* produz versões mais atualizadas e em futura utilização do programa pode ser útil saber para qual versão o programa foi escrito.

3. O comando **capture** faz com que o *Stata* continue rodando mesmo que ocorra um erro na execução de um comando. O comando **capture log close** fecha o arquivo **log** em uso se for aberto outro ou envia mensagem de erro. Outro comando útil é o **quietly** que suprime toda saída exceto as mensagens de erro.
4. O comando **log using filename,replace** abre um arquivo **log** substituindo o já existente.
5. O comando **set more off** faz com que a saída seja apresentada na tela automaticamente, sem ter que manualmente instruir o *Stata* para mostrar o que está faltando.
6. Depois que a análise é feita, o arquivo **.log** é fechado com o comando **log close**.
7. A última linha do programa contendo o comando **exit** não é necessária. Ela é útil para fazer o programa parar de ser rodado.

```
*arquivo exemplo.do*
version 6.0
pause on
stphplot, by(clinic) gap(2) l2(" ")
pause
stcoxkm, by(clinic) gap(1) l2(" ")l1(" ")
sampsi 1 2 , sd(1) p(.8) a(.01)
sampsi 1 3 , sd(1) p(.8) a(.01)
exit
```

6.5 – Macros que contêm resultados de comandos

O *Stata* armazena os resultados de comandos em macros que podem ser acessados com a forma geral **_result(#)**, após o comando.

- **summarize x**

Fornece o resumo da variável *x* e guarda os resultados em macros numerados:

#	Resultado
1	Número de observações
2	Soma das observações
3	Média
4	Variância
5	Valor mínimo
6	Valor máximo
7	Percentil 5
8	Percentil 10

#	Resultado
9	Percentil 25
10	Percentil 50
11	Percentil 75
12	Percentil 90
13	Percentil 95
14	Assimetria
15	Curtose
16	Percentil 1
17	Percentil 99

Assim, o comando

- **gen xnew=x-_result(3)**

gera uma variável xnew que contém a diferença entre cada observação e a média das observações.

7- Exercício 1

1- iniciar o *STATA*

2- abrir um arquivo **exerc1.log** no sub-diretório C:\cursosta

3- abrir banco de dados existente em **C:\cursosta\bancos\female.dta**

Oito variáveis foram medidas em cada uma das 118 pacientes psiquiátricas do sexo feminino. Os dados apresentados constituem um subconjunto. As variáveis são: idade (age), coeficiente de inteligência (iq), ansiedade (anxiety; 1=no, 2=mild, 3=moderate, 4=severe), depressão (depression, 1=no, 2=mild, 3=moderate, 4=severe), problemas para dormir (sleep; 1=yes, 2=no), perda de interesse por sexo (sex; 1=no, 2=yes), tem pensado em suicídio recentemente? (life; 1=no, 2=yes), ganho de peso, em libras, nos últimos 6 meses (weight).

Conrad, S. *Assignments in Applied Statistics*. Wiley, Chichester.1989 (p.126).

4- estudar as variáveis existentes utilizando o comando **describe**

5- alterar o banco de dados utilizando o Editor

paciente 2	age =43	anxiety =3
paciente 10	sleep=1	life= 1

quando terminar, salve as alterações (utilizando a opção **preserve**) e volte para a janela de comandos.

6- salvar o banco de dados como **C:\cursosta\bancos\female corrigido.dta** (utilizando a opção **Save As** do menu)

7- fechar o arquivo de dados utilizando o comando **clear**

8- verificar se o arquivo **.log** continua aberto, utilizando o quarto ícone (pergaminho) e visualizando-o.

9- fechar (suspender definitivamente) o arquivo **.log**

10- abrir arquivo **.log** como continuação (**append**) do arquivo **.log** anterior

11- abrir arquivo de dados **c:\cursosta\breast.txt**, em formato ASCII (**.txt,** que contém os nomes das variáveis na primeira linha - cabeçalho) utilizando comando **insheet**

Estes dados foram coletados para investigar a associação entre temperatura média anual (^{0}F) e mortalidade por câncer de mama em mulheres de alguns países europeus (Reino Unido, Noruega e Suécia). Lea, AJ (1965) New observations on distribution of neoplasms of female breast in certain European countries. *British Medical Journal*, 1, 488-490.

12- visualizar variáveis do banco utilizando o comando **describe**

13- listar os dados utilizando o comando **list**

14- fechar o arquivo de dados utilizando o comando **clear**

15- abrir arquivo de dados **C:\cursosta\human.dat**, em formato ASCII (que não contém o cabeçalho na primeira linha) utilizando o comando **infile var1 var2 var3.**

Os dados são provenientes de um estudo que investiga um novo método para medir composição corpórea. O estudo fornece a porcentagem de gordura corpórea (%fat), idade (age) e sexo (sex) para 18 adultos normais com idade entre 21 e 61 anos. Mazess RB; Peppler WW & Gibbons M Total body composition by dual-photon (^{153}Gd) absorptiometry. *American Journal of Clinical Nutrition*, 40, 834-839, 1984.

16- visualizar os dados utilizando os comandos **describe** ou **browse**

17- renomear as variáveis: **var1** - age, **var2**- %fat e **var3** - sex

18- salvar o arquivo (sobre o arquivo aberto)

19- fechar o arquivo de dados

20- criar o banco de dados da página 7 da apostila e salvá-lo com o nome **banco1.dta** no sub-diretório **cursosta**. Entre os dados diretamente no *Stata*, utilizando o comando **input**

21- fechar o banco de dados após salvamento

22- criar o mesmo banco de dados em processador de texto, contendo o nome das variáveis no cabeçalho. Salvar como somente texto com o nome **banco2.txt**, no sub-diretório **c:\cursosta.** Não se esqueça de fechar o arquivo no Word, quando terminar.

23- abrir o banco no Stata utilizando o comando **insheet**

24- fechar o banco de dados

25- criar banco de dados utilizando o Editor do Stata, salvando no sub-diretório **c:\cursosta**, com o nome **banco3.dta**

26- fechar arquivo de dados

27- fechar arquivo **.log**

28- abrir arquivo **.log** no Word for Windows

29- salvar o arquivo **.log** como arquivo do Word

30- no Stata: abrir arquivo **.log** com novo nome (**rim.log**)

31- abrir arquivo **c:\cursosta\rim.dat** utilizando o comando

infile var1 var2 var3 var4 var5 var6 var7 using c:\cursosta\rim.dat

32- estude as variáveis do banco

33- utilize o comando **compress** para otimizar o armazenamento dos dados

34- substituir os valores codificados como -99 para valores faltantes (.)

35- renomear as variáveis **var1** para **id**, **var2** para **dias**, **var3** para **censura**, **var4** para **sexo**, **var5** para **tratam** , **var6** para **doador** e **var7** para **idade**.

36- rotular as variáveis: **id "identificacao"**; **dias "tempo ate ocorrer o obito"**; **censura "condicao do paciente no fim do estudo"**; **tratam "tratamento"**; **doador "tipo de doador"**. Dado que o arquivo é um arquivo .dat, os labels não aparecem na janela de variáveis. Portanto, para visualizar os rótulos aplicados é necessário descrever as variáveis.

37- definir rótulos para as categorias das variáveis

variável	codificação	
censura	0 – censura	1 – falha (óbito)
sexo	0 – masculino	1 – feminino
tratam	0 – sem imunossupressor	1– com imunossupressor
doador	0 – vivo	1 – cadáver

38- verificar os rótulos gerados utilizando o comando **tab1** e o nome da variável

39- pedir um resumo das variáveis utilizando o comando **summarize** ou **sum**

40- gerar uma nova variável **idade_30** centrada na média utilizando o comando

gen idade_30 = idade - 30

41- listar as variáveis **idade** e **idade_30**; verificar se a nova variável foi criada corretamente

42- salvar o banco de dados incluindo a nova variável gerada utilizando o comando **save**, **replace**

43- criar banco de dados em formato ASCII que contenha os seguintes dados; incluir o nome das variáveis e salvar como texto somente; salvar com nome **c:\cursosta\data nascimento.txt**

id	datanasc
1	30/03/1954
2	4/07/1928
3	12/02/1961
5	9/07/1987

44- abrir o banco de dados no Stata utilizando o comando **insheet**

45- gerar variável numérica correspondendo à variável data

46- visualizar o que foi feito utilizando o comando **list** ou **browse**

47- visualizar o tipo de variável gerada

48- formatar a variável numérica referente a data, em um formato compreensível e visualizá-la

49- gerar uma variável que corresponda à idade, em anos, da pessoa, em 1º de janeiro de 2001.

50- corrigir a data do paciente 2 para 10/08/1970, pelo Editor do Stata.

51- apagar (jogar fora) a variável data numérica, recriá-la depois da correção e apresentá-la em um formato compreensível.

52- visualizar as modificações

53- fechar arquivo **.log**

54- fechar arquivo de dados

55- salvar comandos utilizados durante a sessão de trabalho

Gabarito – lista de comandos

1- pelo ícone ou **Iniciar**, **Programas**, **Stata**, **Intercooled Stata**

2- clicar no quarto ícone, mudar diretório para **c:\cursosta**, salvar com nome **exerc1.log**, fechar janela do arquivo **.log**

3- **use c:\cursosta\female.dta** ou pelo menu, **File, Open** e seleciona-se o arquivo **female.dta**, no diretório C:\cursosta

4- **describe** ou **desc**

5- utilizar o editor do Stata (10° ícone) para correção. Para salvar, clicar em **preserve**

6- **File, Save As**. Salvar com o nome **female corrigido.dta**

7- clear

8- clicar sobre o 4^0 ícone, escolher a 1^a. opção (**Bring log window to top**); rolar a tela do arquivo **.log**, fechar a janela do arquivo **.log**

9- clicar sobre o 4^0 ícone e selecionar a opção **Close log file** ou utilizar a opção fechar do Windows (X no topo superior direito da janela).

10- clicar sobre 4° ícone, abrir arquivo já existente e escolher opção **append to existing file**

11- **insheet using c:\cursosta\breast.txt**

12- **describe** ou **desc**

13- **list**

14- **clear**

15- **infile var1 var2 var3 using c:\cursosta\human.dat**

16- **describe**

17- **ren var1 age** <E>

ren var2 fat <E>

ren var3 sex <E>

18- **Save As c:\cursosta\human.dta** <E>

19- clear

20- **input id str10 nome tratamem pesoinic sexo**

digitar os dados, quando terminar digite **end**

21- **Save As c:\cursosta\banco1.dta** <E>

clear <E>

22- no Word, digitar nomes das variáveis e valores, separados por TAB. Salvar como somente texto: c:\cursosta\banco2.txt. Confirmar que se quer salvar nesse formato (clicando em Sim).

23- **insheet using c:\cursosta\banco2.txt**

24- **clear**

25- acessar o Editor do Stata pelo 10º. ícone. Digitar os dados, organizando as variáveis por coluna, sem entrar com o nome da variável. Depois da digitação, pressione na opção **preserve.** Fechar a janela do Editor. Salvar o arquivo utilizando **File, Save As**. com o nome **banco3.dta**

26- **clear**

27- no 4º ícone, escolher a opção **Close log file**

28- no Word, abrir arquivo .log

29- salvar como arquivo do Word (Documento do Word)

30- **log using c:\cursosta\rim.log** ou pelo menu: no 4º ícone, abrir arquivo rim.log

31- **infile var1 var2 var3 var4 var5 var6 var7 using c:\cursosta\rim.dat**

32- **describe** ou **desc** e **browse**

33- **compress**

34- **mvdecode var*, mv(-99)**

35- **ren var1 id**

ren var2 dias

ren var3 censura

ren var4 sexo

ren var5 tratam

ren var6 doador

ren var7 idade

36- **label variable id "identificacao"**

label var dias "tempo ate ocorrer o obito"

label var censura "condicao do paciente no fim do estudo"

label var tratam "tratamento"

label var doador "tipo de doador"

describe ou **desc**

37- **label define cen 0"censura" 1"falha"**

label val censura cen

label define s 0"masculino" 1"feminino"

label val sexo s

label define trat 0"sem imunossupressor" 1"com imunossupressor"

label val tratam trat

label define doa 0"vivo" 1"cadaver"

label val doador doa

38- **tab1 censura sexo tratam doador**

39- **sum** ou **summarize**

40- **gen idade_30=idade-30**

41- **list idade idade_30**

42- **save, replace**

43- no Word, digitar nomes das variáveis e valores, separados por TAB. Salvar como somente texto: **c:\cursosta\data nascimento.txt**. Confirmar que se quer salvar nesse formato (clicando em Sim).

44- **insheet using "c:\cursosta\data nascimento.txt"**

45- **gen data = date(datanasc, "dmy")**

46- **list** ou **browse**

47- **describe** ou **desc**

48- **format data %d**

49- **gen age2001 = (mdy(1,1,2001) - data) / 365.25**

50- utilizar o editor do Stata (10° ícone) para correção. Primeiramente, alterar a variável do tipo string datanasc. Para salvar, clicar em **preserve**

51- **drop data**

gen data = date(datanasc, "dmy")

format data %d

52- **describe** ou **desc** e **list** ou **browse**

53- no 4º ícone, escolher a opção **Close log file**

54- **clear**

55- Na janela Review, clicar sobre a caixa no canto superior esquerdo e escolher **Save Review Contents**. O arquivo terá extensão .do que poderá ser utilizado como a base para de um arquivo de programa.

.

8- Exercício 2

Exercício suplementar, página 32:

1. Faça o resumo da variável **weight** segundo nível de depressão (variável **depres**);

2. Faça a tabela que contém somente o peso médio e o desvio padrão da variável perda de peso (**weight**) para os níveis da variável **depres**;

3. Procure no **Help** a sintaxe do comando para realizar o *teste U de Mann-Witney*;

4. Compare as mudanças de peso segundo a variável **depres**, utilizando o *teste U de Mann-Witney*;

5. Faça um histograma da variável **age** e salve-o em um arquivo **doc**.

6. Faça um **boxplot** da variável **weight** segundo níveis da variável depres.

7. Crie um arquivo do conteúdo estes comandos e execute-o. Use a opção **saving(filename),replace** para salvar o gráfico e investigue o gráfico depois, utilizando o comando **graph using filename**.

Gabarito - exercício suplementar, página 32

56- **use "C:\cursosta\female.dta", clear**

sort depressi

by depressi: sum weight

57- **table depressi, contents(mean weight sd weight)**

58- Help, Contents. Em "Command:", digitar Mann-Whitney. Clicar na opção signrank (o teste de Mann-Whitney é feito pelo comando ranksum).

---

help for signrank, signtest, ranksum (manual: [R] signrank)

Sign and rank tests

---

signrank varname = exp [if exp] [in range]

signtest varname = exp [if exp] [in range]

ranksum varname [if exp] [in range], by(groupvar)

Description

---

signrank tests the equality of matched pairs of observations using the Wilcoxon matched-pairs signed-ranks test. The null hypothesis is that both distributions are the same.

signtest also tests the equality of matched pairs of observations. It does this by calculating the difference between varname and the expression. The null hypothesis is that the median of the differences is zero; no further assumptions about the distributions are made. This, in turn, is equivalent to the hypothesis that the true proportion of positive (negative) signs is one-half.

ranksum tests the hypothesis that two independent samples (i.e., unmatched data) are from populations with the same distribution using the Wilcoxon rank-sum test which is also known as the Mann-Whitney two-sample statistic. Note

that the by() "option" is not optional.

Options
---
by(groupvar) is not optional. It specifies the name of the grouping variable.

Examples

```
. signrank mpg1 = mpg2
. signtest mpg1 = mpg2
. ranksum mpg, by(treatment)
```

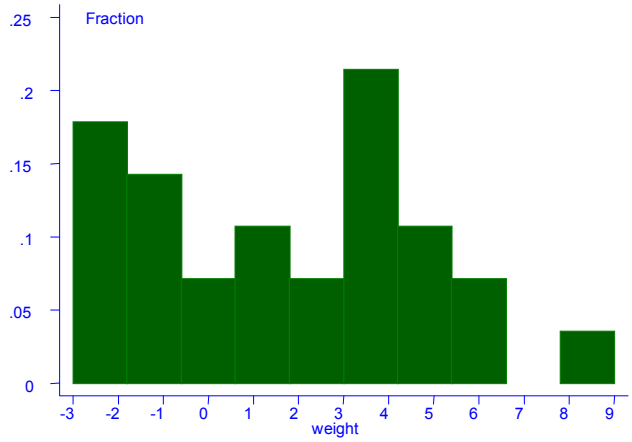
Also see

Manual: [R] signrank
On-line: help for kwallis, nptrend, runtest, ttest

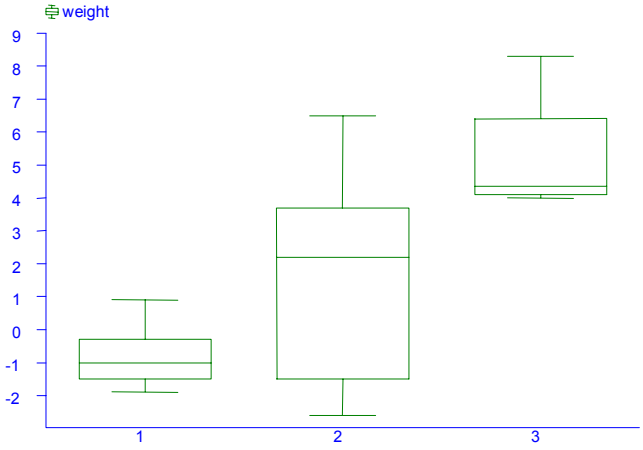
59- **ranksum weight, by(life)**

60- **graph weight, bin(10) xlab(-3,-2,-1,0,1,2,3,4,5,6,7,8,9) ylab(0,0.05,0.10,0.15,0.20,0.25) saving("C:\cursosta\histograma.gph", replace)**

Edit, Copy Graph. Abrir o Word, colar no documento e salvá-lo em um arquivo do Word.



61- **graph weight, by(depressi) box ylab(-2,-1,0,1,2,3,4,5,6,7,8,9) saving("C:\cursosta\boxplot.gph", replace)**



62- Na janela Review, clicar no ícone superior esquerdo e escolher Save Review Contents. Dar um nome para o arquivo, por exemplo, compara.do, no subdiretório C:\cursosta. Abrir o 9^0 ícone (Do-file Editor), e abrir o arquivo - File, Open, compara.do. Clicar em Abrir. Editar o arquivo deixando somente os comandos corretos.

```
use "C:\cursosta\female.dta", clear
recode weight -99=.
sort depressi
by depressi: summarize weight
table depressi, contents(mean weight sd weight)
ranksum weight, by(life)
graph weight, bin(10) xlab(-3,-2,-1,0,1,2,3,4,5,6,7,8,9)
ylab(0,0.05,0.10,0.15,0.20,0.25) saving("C:\cursosta\histograma.gph", replace)
sort depressi
graph weight, by(depressi) box ylab(-2,-1,0,1,2,3,4,5,6,7,8,9) sa-
ving("C:\cursosta\boxplot.gph", replace)
```

graph using "C:\cursosta\histograma.gph"

graph using "C:\cursosta\boxplot.gph"

9- Bibliografia

Caplehorn J e Bell J. Methadone dosage and the retention of patients in maintenance treatment. *The medical Journal of Australia*, 154:195-9, 1991.

Conrad S. *Assignments in Applied Statistics*. Wiley, Chichester, 1989 (p.126).

Hamilton LC. *Statistics with Stata 5*. Duxbury Press, Belmont, CA, 1998.

Hand DJ et al. *A Handbook of Samall Data Sets*. Chapmen e Hall, London, 1994.

Holtbrugge W e Schumacher M. A comparison of regression models for the analysis of ordered categorical data. *Applied Statistics*, 40:249-59, 1991.

Lea AJ New observations on distribution of neoplasms of female breast in certain European countries. *British Medical Journal*, 1, 488-490, 1965.

Mazess RB; Peppler WW & Gibbons M Total body composition by dual-photon (^{153}Gd) absorptiometry. *American Journal of Clinical Nutrition*, 40, 834-839, 1984.

StataCorp Stata Statistical Software: release 6.0. Stata Corporation, 1999.