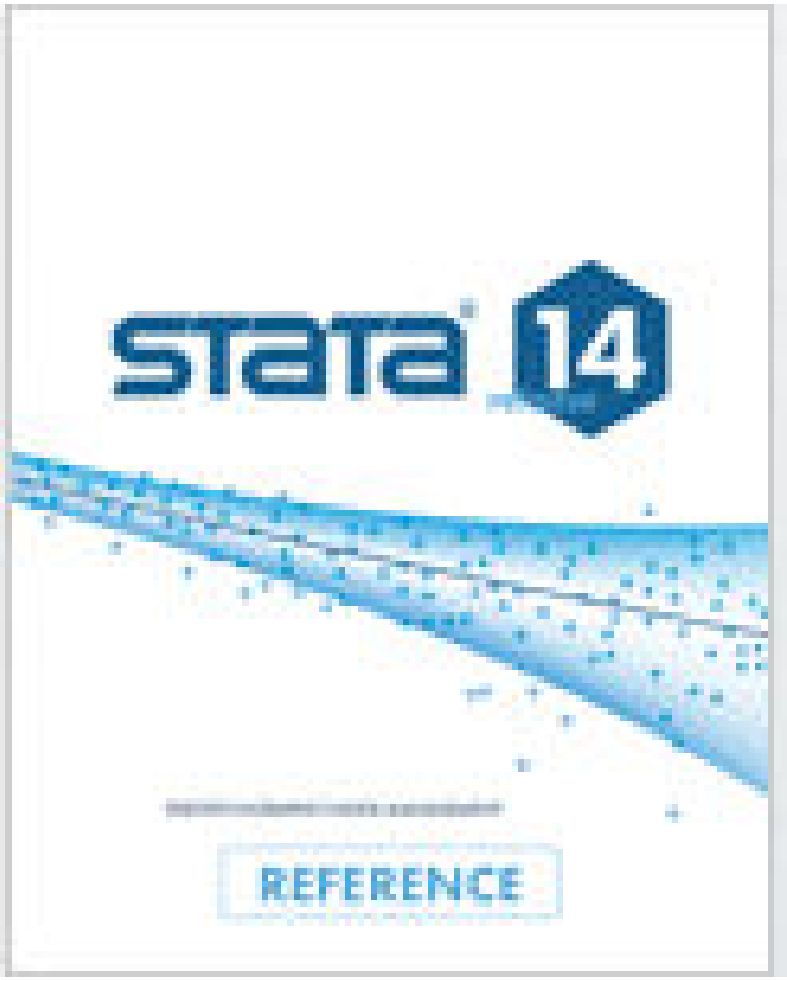
Stata, versão 14



Curso Básico

Ivaldo Olimpio da Silva
Departamento de Medicina Preventiva - FMUSP

2017

Í N D I C E

Objetivo geral do curso

Esta apostila apresenta comandos básicos para manipulação de bases de dados com a utilização do aplicativo STATA, versão 14 e introduz alguns conceitos básicos de estatística referentes aos comandos utilizados.
O leitor interessado em conhecer mais sobre este programa ou aprender teoria estatística mais detalhada deve procurar referências especializadas.

Introdução ao STATA

O STATA possui amplo potencial de utilização e trabalha com bases de dados que ficam armazenadas inteiramente na memória RAM do microcomputador. Por esta razão elabora processamentos de maneira muito rápida.

Em geral, os comandos do STATA tem a forma:
comando nomevar(s) **if....in...., options**

O STATA diferencia letras maiúsculas das minúsculas. Use sempre letras minúsculas quando digitar comandos, e recomendamos que você também use letras minúsculas para os nomes de suas variáveis. O STATA aceita abreviações para comandos e nomes de variáveis, desde que estas abreviações não sejam ambíguas.

Iniciando o STATA

O programa STATA, é iniciado clicando duas vezes no ícone localizado no desktop do Windows.

Janelas do STATA

Cinco janelas são apresentadas quando o STATA é iniciado. São elas:

Review: janela onde são armazenados os comandos.

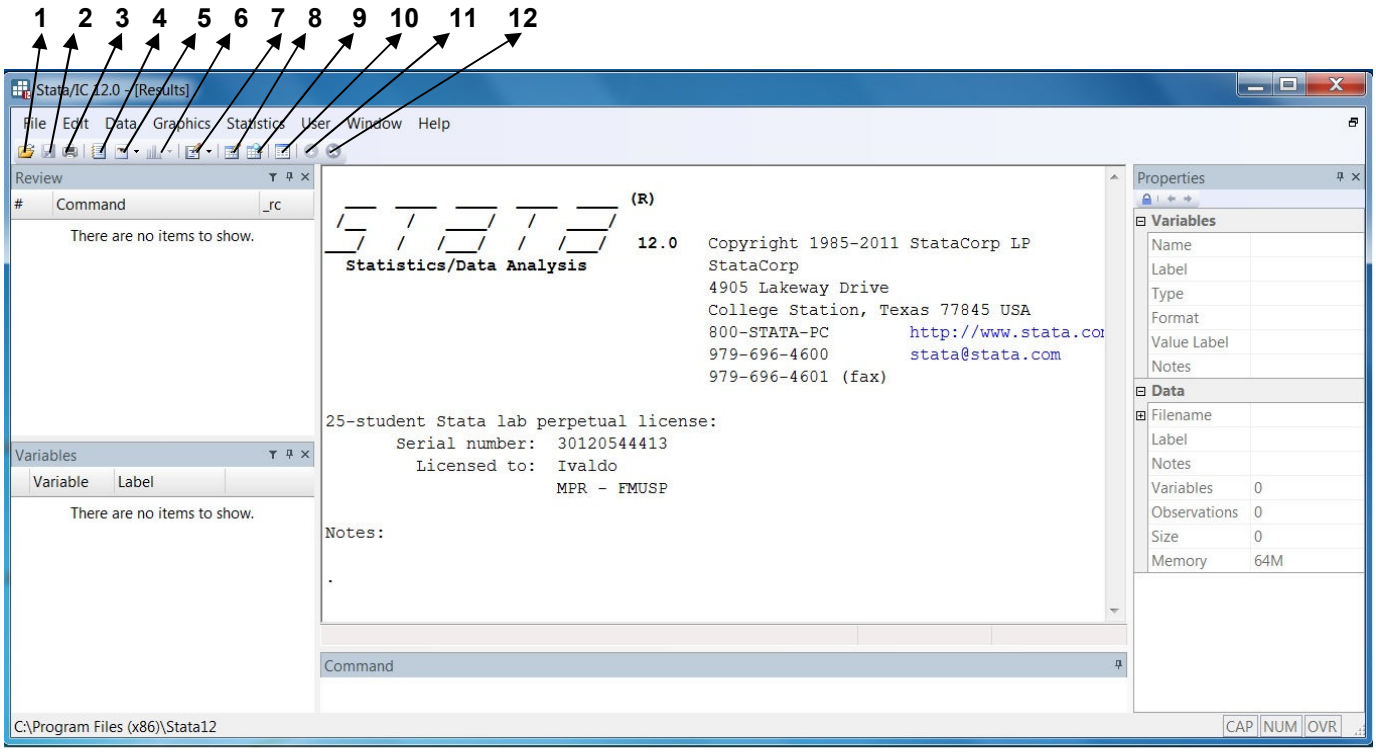
Variables: janela com a lista das variáveis do banco de dados ativo.

Stata Results: janela com os resultados.

Stata Command: janela para digitar os comandos do STATA.

Properties: exibe as características das variáveis e do banco de dados ativo.

Clicando com o botão direito do "*mouse*" , na janela **Review**, ativa-se a opção para salvar os comandos.



O *menu* está disponível na primeira linha e possui os recursos:

File **E**dit **P**refs Data Graphics Statistics **W**indow e **H**elp

Por exemplo, o *menu* "**H**ELP ⇒ SEARCH" é utilizado para procurar ajuda sobre comandos do STATA.

Na segunda linha encontra-se a *Barra de Ferramentas* com os ícones:

(1) *Open (use)*: Abre um banco de dados no formato do STATA.

(2) *Save*: Salva um arquivo no formato do STATA (dta).

(3) *Print Results*: Imprime a janela de resultados.

(4) *Log Begin*: Cria um arquivo do tipo "log" .

(5) New *Viewer*: Exibe a tela de ajuda (Help) em primeiro plano.

(6) *Bring Graph Window to Front*: Exibe a tela de gráficos.

(7) New *Do-file Editor*: Editar arquivo de comandos.

(8) *Data Editor*: Editar arquivo de dados.

(9) *Data Editor (Browse)*: Visualizar/Editar arquivo de dados.

(10) Variables Manager: Gerenciador de variáveis.

(11) Clear More: Limpar/prosseguir a execução do comando.

(12) *Break*: Interrompe a execução de uma tarefa ou comando.

Tipos de arquivos do STATA

.ado	arquivos do programa "ado-files"
.do	do-file
.dta	arquivos de dados formato do STATA
.gph	arquivos gráficos
.log	arquivos textos com resultados(tabelas)

Onde estão os arquivos utilizados no curso ?

Em cada microcomputador foi criado o diretório **C:\Cursos\Stata básico** com todos os bancos de dados que serão utilizados neste curso. É aconselhável que você salve os arquivos neste diretório.

Preparando o STATA para análise

Em primeiro lugar, você deve escolher o banco de dados que irá trabalhar e abrir/carrega-lo no STATA. Note que o STATA só abre bancos de dados no formato **".dta"**. Por isso, caso seu banco de dados não esteja neste formato, antes de iniciar o STATA você deve convertê-lo utilizando o programa *STAT/TRANSFER* que é um programa muito útil e fácil de ser usado.

Para ilustrar, vamos trabalhar com o arquivo **motocobr.dta** que refere-se a um estudo de prevalência de transtornos mentais comuns (depressão e/ou ansiedade) em motoristas e cobradores de ônibus da cidade de São Paulo (Souza, 1996).

Abra o banco de dados clicando no ícone **(1)Open** e, então, selecione o caminho (pasta) onde está o arquivo **motocobr.dta**. Note o que mudou nas janelas do STATA!!

Agora, vamos abrir também um arquivo do tipo **"log"** onde ficarão armazenados todos os resultados gerados a partir da janela de comandos. Isto pode ser feito clicando o ícone **(4)Log Begin** e, então, selecionando o tipo de arquivo=**log**, o caminho e o nome para o arquivo.

Comandos básicos do STATA :

describe	descreve o arquivo de dados em uso
display	calculadora de mão
drop	elimina variáveis ou observações
edit	edita e lista dados
generate	cria ou muda conteúdos de variáveis
graph	cria gráficos
list	lista os valores das variáveis por registro
memory	mudar a quantidade da memória a ser utilizada
recode	recodificar, agrupar códigos(valores)
sort	ordenar os dados
summarize	calcula medidas de tendência central
tabulate	produz/elabora tabelas simples e cruzadas

Utilize o *help* do STATA para obter mais informações sobre estes e outros comandos.

Salvando os comandos

Todos os comandos digitados na janela **Command** são enviados para a janela **Review.** Estes comandos podem ser guardados em um arquivo especial (arquivos tipo **"do"**) para, posteriormente, serem editados e utilizados em uma nova análise.

Para criar um arquivo do tipo **"do"** utilize o botão direito do “mouse” na janela **Review .**

Análise descritiva

Após a coleta de dados e a digitação dos mesmos em um banco de dados apropriado, o próximo passo é a análise descritiva. Esta etapa é fundamental, pois uma análise descritiva detalhada fornece ao pesquisador toda a informação contida no conjunto de dados. Neste enfoque, procura-se obter a maior quantidade possível de informação, buscando responder às questões que estão sendo pesquisadas.

As variáveis podem ser classificadas em contínuas ou categóricas. Por **variável contínua** (ou quantitativa) entende-se as variáveis que podem assumir todos os valores possíveis dentro de um limite especificado. **Variável categórica** (ou qualitativa) é aquela que pode ser classificada em categorias separadas e que não assumem valores intermediários, como por exemplo, sexo e estado civil.

Em geral, uma análise descritiva dos dados é feita com base em medidas de posição e variabilidade. Para variáveis contínuas, as medidas comumente utilizadas são as medidas de tendência central, enquanto as variáveis categóricas são sumarizadas por meio de medidas de freqüência.

Medidas de tendência central:

média aritmética: é a soma de todas as observações dividida pelo número de observações.

mediana: valor central de uma distribuição. Para se obter a mediana, ordena-se as observações em ordem crescente. Se o número de observações for par, a mediana será a média aritmética dos dois valores centrais (n/2 e [(n/2)+1], onde n é o número de observações total da amostra. Se o número de observações for ímpar, a mediana será o valor na posição (n + 1)/2.

frequência: é o número de vezes em que um valor ocorre.

A seguir são apresentados alguns comandos básicos para elaborar uma análise descritiva dos dados:

Aplicação prática-1 - Digitando os comandos na janela **Command**

Digite **describe** ou **desc** e pressione ENTER, aparecerá na janela **Results** o seguinte resultado:

```
Contains data from C:\Motocobr.dta
  obs:           800
 vars:            18                          22 Aug 2000 15:44
 size:        35,200 (96.3% of memory free)
-------------------------------------------------------------------------------
   1. id        long   %12.0g                 id
   2. idade     byte   %8.0g                  idade
   3. pausas    byte   %8.0g                  numero de pausas dia
   4. escola    long   %19.0g      escola     escola
   5. nasc      byte   %8.0g       nasc       procedencia
   6. tsp       int    %11.0g      tsp        tempo em SP
   7. emp       int    %8.0g       emp        tipo de empresa
   8. fun       int    %9.0g       fun        funcao
   9. esc       int    %13.0g      esc        escala da linha do onibus
  10. fol       int    %8.0g       fol        escala de folga
  11. jorn      int    %11.0g      jorn       jornada de trabalho diaria
  12. temp      int    %9.0g       temp       tempo de trabalho na empresa
  13. trans     long   %12.0g      trans      transito diario
  14. banco     long   %12.0g      banco      possibilidade ajuste do assento
  15. fal       int    %8.0g       fal        falta ao trabalho no ultimo mes
  16. sono      int    %10.0g      sono       sono diario
  17. tmc       int    %8.0g       srq        transtorno mental comum
  18. sal       byte   %8.0g       sal        salario mensal
-------------------------------------------------------------------------------
Sorted by:
```

Digite **list in 1** e pressione ENTER

```
Observation 1

          id           27       idade           35      pausas            2
      escola primario com        nasc     nordeste         tsp   11-20 anos
         emp      privada         fun    motorista         esc linha altern
         fol         muda        jorn          > 9        temp     < 4 anos
       trans      intenso       banco          sim         fal          nao
        sono   >= 6 horas         tmc          nao         sal       > 6 sm
```

Para mudar o nome de uma variável, como por exemplo, id para identif, digite

rename id identif

e pressione ENTER

Para observar a mudança. Digite **desc**

Os comandos **tabulate , tab** ou **tab1** produzem tabelas simples ou cruzadas.

tab escola

escola	Freq.	Percent	Cum.
ginasio completo	84	10.50	10.50
primario completo	554	69.25	79.75
primario incompleto	162	20.25	100.00
Total	800	100.00	

tab escola, nolabel

escola	Freq.	Percent	Cum.
0	84	10.50	10.50
1	554	69.25	79.75
2	162	20.25	100.00
Total	800	100.00	

Agora digite: **tab1 escola fun emp**

Aparecerá na tela os seguintes resultados:

-> tabulation of escola

escola	Freq.	Percent	Cum.
ginasio completo	84	10.50	10.50
primario completo	554	69.25	79.75
primario incompleto	162	20.25	100.00
Total	800	100.00	

-> tabulation of fun

funcao	Freq.	Percent	Cum.
motorista	423	52.88	52.88
cobrador	377	47.12	100.00
Total	800	100.00	

-> tabulation of emp

tipo de empresa	Freq.	Percent	Cum.
publica	286	35.75	35.75
privada	514	64.25	100.00
Total	800	100.00	

Criar uma nova variável(**nasc2**) com duas categorias de procedência, SP(código 0) e Outras(código 1).

Para criar a variável **nasc2**, recodificar e inserir um rótulo (*label*), utilize os comandos:

tab nasc (tabela de freqüência)

```
tab nasc

procedencia |      Freq.     Percent        Cum.
------------+-----------------------------------
         SP |        281       35.12       35.12
   RJ/MG/ES |        135       16.88       52.00
     outros |         48        6.00       58.00
   nordeste |        336       42.00      100.00
------------+-----------------------------------
      Total |        800      100.00
```

gen nasc2=nasc (criar variável **nasc2**)
recode nasc2 0=0 1=0 2=1 3=1 (recodificar variável **nasc2**)

label var nasc2 "Grupos de Procedência" (insere *var label*)

tab nasc2

```
  Grupos de |
Procedência |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        416       52.00       52.00
          1 |        384       48.00      100.00
------------+-----------------------------------
      Total |        800      100.00
```

label define cproc 0 "sudeste" 1 "outros" (insere *value label*)

label val nasc2 cproc

tab nasc2

```
  Grupos de |
Procedência |      Freq.     Percent        Cum.
------------+-----------------------------------
    sudeste |        416       52.00       52.00
     outros |        384       48.00      100.00
------------+-----------------------------------
      Total |        800      100.00
```

O comando **summarize** ou **sum** é utilizado para calcular média, desvio padrão, mínimo, máximo, etc.

summarize idade

```
Variable |     Obs        Mean   Std. Dev.       Min        Max
---------+--------------------------------------------------------
   idade |     800       37.69   10.52532         17         67
```

sum idade, detail

```
                          idade
-------------------------------------------------------------
      Percentiles      Smallest
 1%           21             17
 5%           22             19
10%           24             19       Obs                 800
25%           30             19       Sum of Wgt.         800

50%           37                      Mean              37.69
                        Largest       Std. Dev.       10.52532
75%           45             65
90%           53             66       Variance        110.7824
95%           58             66       Skewness         .440607
99%           63             67       Kurtosis        2.555018
```

Aplicação prática-2 - Utilizando os menus: **Data** e **Statistics**

2.1 - Para descrever o arquivo e suas variáveis, clique no menu **Data**, opção: describe data e describe data in memory.

2.2 – Para editar o banco de dados, clique no menu **Data**, opção: Data editor .

2.3 – Para produzir tabelas simples, clique no menu **Statistics**, opção: Summaries, tables & tests → Tables → Multiple one-way tables.

2.4 Criar a variável *nasc3* a partir da variável *nasc*. Clique no menu **Data**, opção: Create or change variables → Create new variable

Acrescentar *label* para a variável *nasc3* e *label* para os valores de *n*asc3.

Label var : menu **Data**, opção: Data utilitis → Label utilitis → Label variables.

Label val : menu **Data**, opção: Data utilitis → Label utilitis → Assign value label.

2.5 Calcular a média, mediana, desvio padrão, ... Clique no menu **Data**, opção Describe data → Summary statistics ou menu **Stastistics**, opção: Summaries, tables & tests → Summary statistics.

Aplicação prática-3

3.1 – Ler/abrir o arquivo : ***motocobr.dta*** Clique no menu **File**, opção *open.*
Abrir arquivo(**log**) para armazenar os resultados: Clique no botão: *Begin log; esolha (*.log) em tipo de arquivo;* digite ***motcob*** *em nome do arquivo ;* e clique no botão *SALVAR.*

3.2 – Produzir tabela de freqüência simples para as variáveis IDADE e FUN.

tab1 idade fun

tab1 idade fun , nolabel

3.3 – Criar nova var´ IDADER, idade recodificada(agrupada) nas faixas:
ate 30 ; 31 a 40 ; 41 a 50 ; 51 e mais

generate idader= idade

recode idader 17/30=1 31/40=2 41/50=3 51/67=4

3.4 - Inserir *labels* para a variável IDADER

label var idader "idade agrupada"

label define c_idade 1 " ate 30" 2 "31 - 40" 3 "41 - 50" 4 "51 e mais"

label val idader c_idade

3.5 - Produzir tabelas:

tab idader

tab idader fun

tab idader fun , row col cel chi

3.6 - Salvando os arquivos:

(1) arquivo de dados(**dta**) : menu **File**, opção Save as ... e digite **motcob** em nome do arquivo.

(2) arquivo de resultados(**log**) : clique no botão : *Log Begin* e escolha a opção *close log file* e clique no botão *OK* para salvar.

Minimize a tela do STATA e Acesse o aplicativo Word.

Abra o arquivo **motcob.log** no Word e observe os resultados. As tabelas e os resultados poderão ser salvos como um arquivo-doc. Fechar o Word.

(3) arquivo de comandos(**do**) : clique na da janela ***Review*** *com o botão* direito do mouse , e *escolha a opção Save Review Contents ... e digite* **motcob** em nome do arquivo.

Visualizando o arquivo do : clique no botão *Do-file Editor* para abrir o arquivo **motcob.do**.

Gráficos

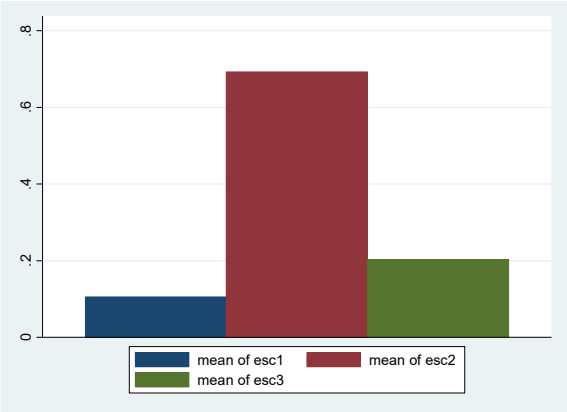
O comando *graph* do STATA possui várias opções. Em geral, gráficos de barra(*bar*) e de setores(*pie*) são usados para mostrar a distribuição de variáveis categóricas, enquanto histogramas e *box-plots* são usados para mostrar a distribuição das variáveis quantitativas.

Para obter um gráfico para a variável escola , utilize os comandos:

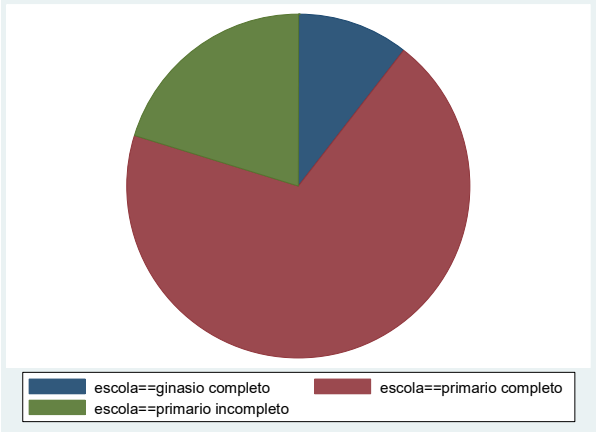
tab escola, gen(escola) (gerar três var´s para as categorias de escola)

escola	Freq.	Percent	Cum.
ginasio completo	84	10.50	10.50
primario completo	554	69.25	79.75
primario incompleto	162	20.25	100.00
Total	800	100.00	

gr bar escola1 escola2 escola3



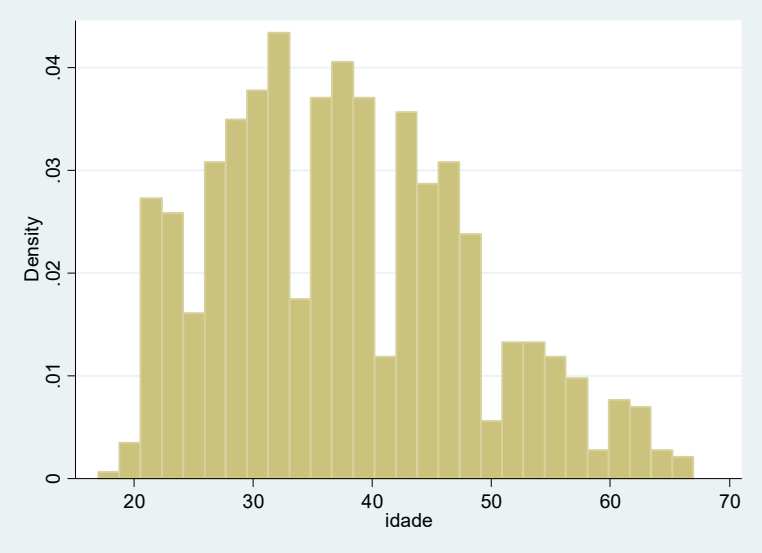
gr pie escola1 escola2 escola3



A cada novo gráfico que o Stata gerar, o anterior será "perdido", por isso, às vezes é desejável salvar um gráfico antes de gerar outro. O gráfico pode ser salvo de duas maneiras diferentes: a primeira é copiar cada gráfico e colar em um outro arquivo "fora" do STATA, por exemplo, um arquivo do Word. A outra maneira é salvar a janela com o gráfico como uma figura, utilizando o *menu*: **File → Save graph** .

Para obter um histograma da variável: idade, digite:

histogram idade



Para melhorar a apresentação visual do histograma, utilize o opção `xlabel` e `ylabel`. O número de retângulos do histograma pode ser modificado pela opção `bin(x)`. Para sobrepor ao seu histograma uma curva normal com média e desvio padrão, adicione a opção *normal*.

histogram idade, percent normal ytitle(percentual) ylabel(#5) xtitle(Idade) bin(10)

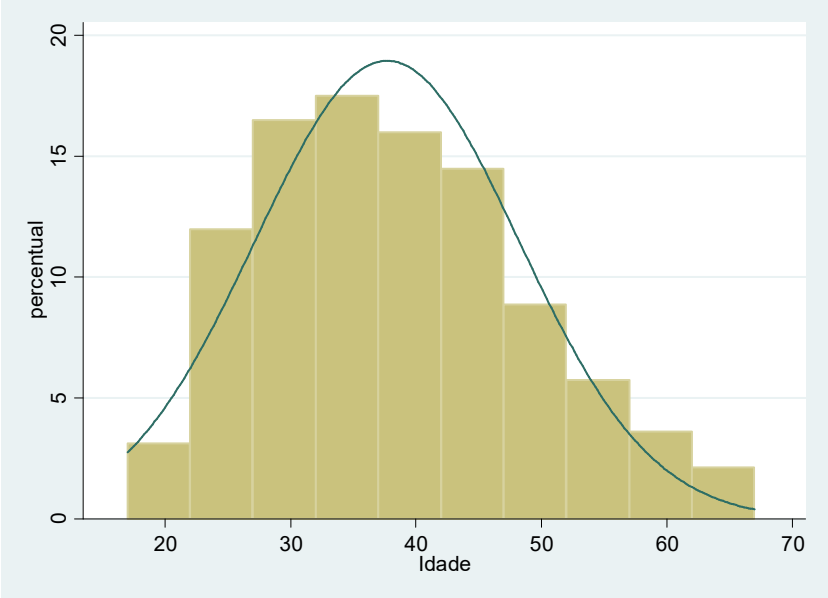
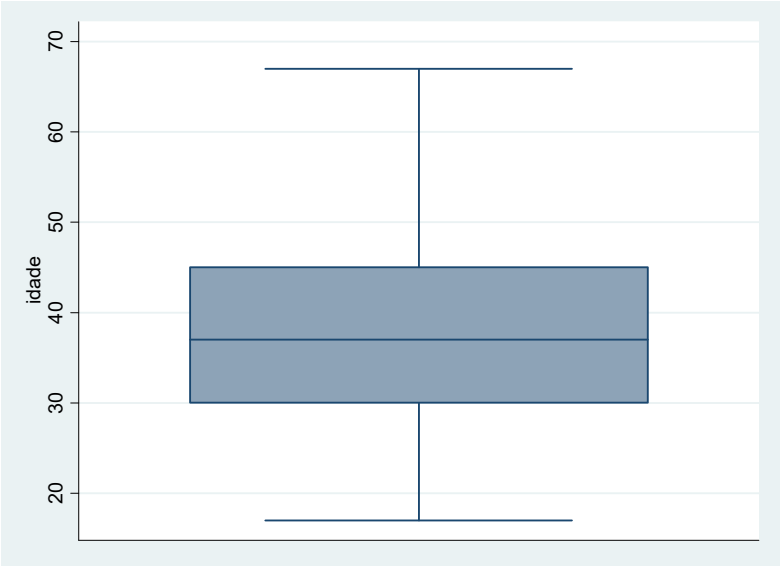


Gráfico do tipo b*ox-plot* para a variável: idade, pode ser obtido com o comando

gr box idade



gr box idade, by(fun)

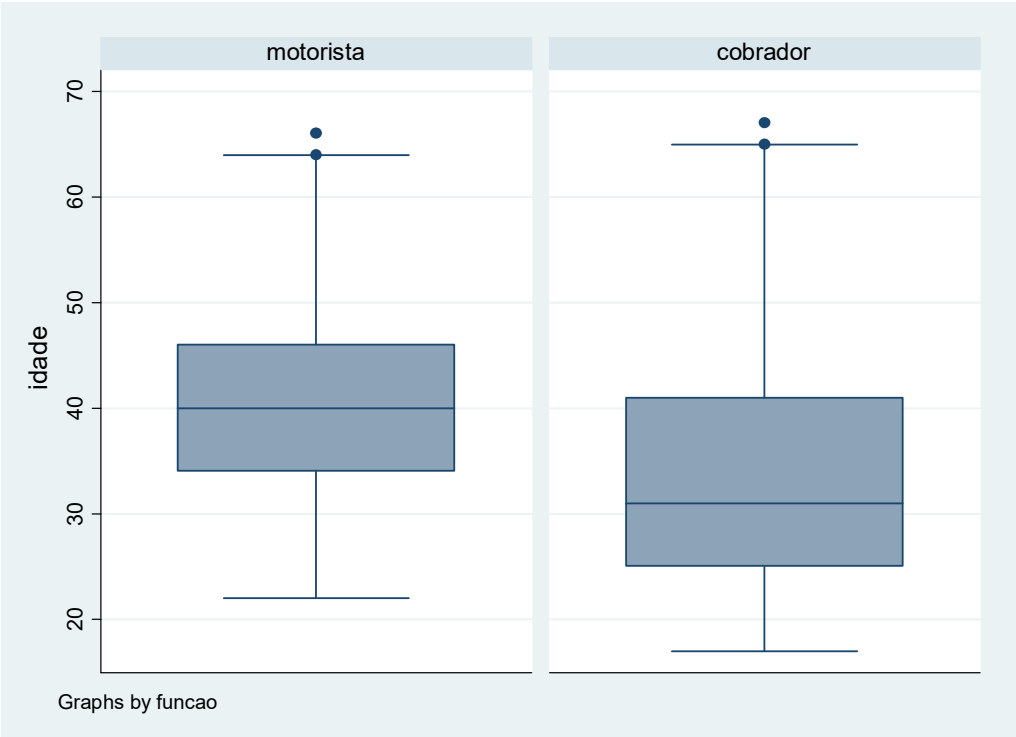
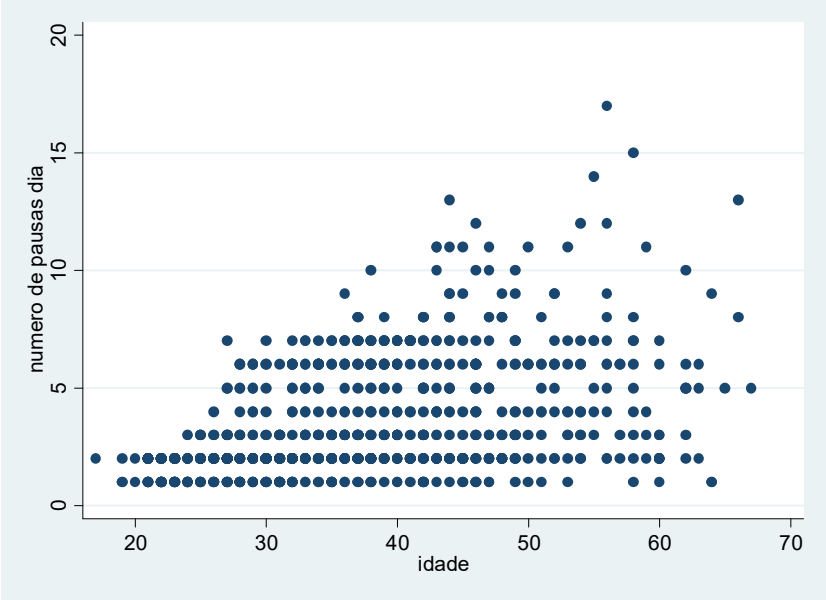
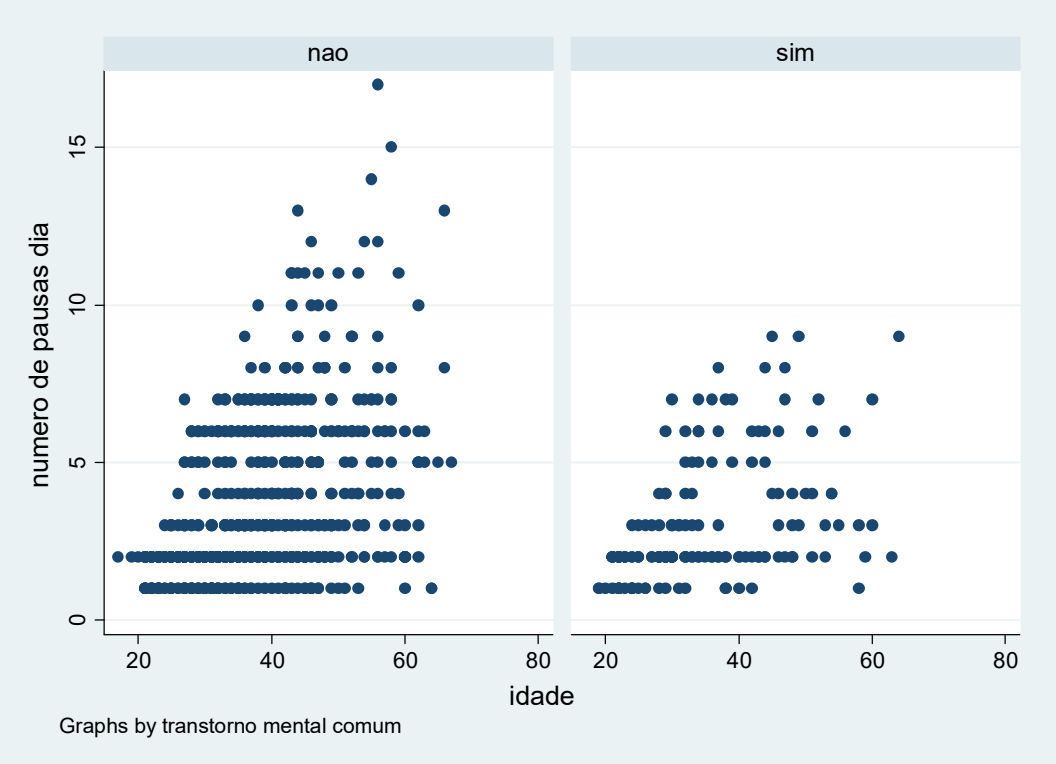


Gráfico de dispersão(*scatter-plot*) para as variáveis: idade e pausas:

```
Scatter  pausas idade
```




```
Scatter  pausas idade,  by(tmc)
```



Testes de hipóteses

Testes de hipóteses consistem em testar a significância estatística e quantificar o grau em que a variabilidade da amostra pode ser responsável pelos resultados observados no estudo. Para isto, define-se uma hipótese nula (H_0) e uma hipótese alternativa (H_a), que podem representar, por exemplo:

H_0 : não existe diferença entre exposição e doença

H_a: existe diferença entre exposição e doença.

Relacionando duas variáveis categóricas

A seguir são ilustradas algumas maneiras de relacionar duas ou mais variáveis categóricas.

Suponha que você queira investigar se os trabalhadores que têm mais transtorno mental comum (TMC) faltam mais ao trabalho, ou seja, se existe uma associação entre TMC e a falta ao trabalho. Para isto, você pode construir uma tabela 2X2 usando o comando *tabulate* ou, de forma abreviada, *tab*

```
tab  tmc  fal
```

```
transtorno | falta ao trabalho no
    mental |      último mês
     comum |       não        sim |     Total
-----------+----------------------+----------
       não |       485        160 |       645
       sim |       100         55 |       155
-----------+----------------------+----------
     Total |       585        215 |       800
```

A tabela acima não mostra com clareza se as duas variáveis analisadas estão associadas.

Uma opção simples é analisar as porcentagens destas variáveis em relação aos totais observados. Os subcomandos **row**, **col** e **cel** fornecem, respectivamente, as porcentagens das linhas, colunas e do total:

tab tmc fal, row col cel

```
transtorno | falta ao trabalho no
    mental |      último mês
     comum |       não        sim |     Total
-----------+----------------------+----------
       não |       485        160 |       645
           |     75.19      24.81 |    100.00
           |     82.91      74.42 |     80.63
           |     60.62      20.00 |     80.63
-----------+----------------------+----------
       sim |       100         55 |       155
           |     64.52      35.48 |    100.00
           |     17.09      25.58 |     19.38
           |     12.50       6.88 |     19.38
-----------+----------------------+----------
     Total |       585        215 |       800
           |     73.13      26.88 |    100.00
           |    100.00     100.00 |    100.00
           |     73.13      26.88 |    100.00
```

Avaliando a associação de duas variáveis com o teste Qui-quadrado de Pearson

Ainda com o objetivo de estudar a associação entre *função do empregado* e *presença de falta no último mês*, vamos usar o teste Qui-quadrado de Pearson para testar a significância da associação. Para isto, utilize a opção **chi**.

tab tmc fal, row chi

```
transtorno | falta ao trabalho no
    mental |      último mes
     comum |       não        sim |     Total
-----------+----------------------+----------
       não |       485        160 |       645
           |     75.19      24.81 |    100.00
-----------+----------------------+----------
       sim |       100         55 |       155
           |     64.52      35.48 |    100.00
-----------+----------------------+----------
     Total |       585        215 |       800
           |     73.13      26.88 |    100.00

          Pearson chi2(1) =   7.2500   Pr = 0.007
```

Considerações a respeito da validade do teste Qui-quadrado de Pearson

O teste Qui-quadrado de Pearson segue, aproximadamente, um distribuição chamada Qui-quadrado (χ^2). Para amostras grandes esta suposição é razoável. No entanto, as seguintes regras podem ser usadas para garantir a validade do uso do teste:

- para tabelas 2 x 2, o teste χ^2 pode ser usado :
  - se o tamanho total da amostra (N) é maior do que 40,
  - se N está entre 20 e 40 e o menor valor esperado é maior ou igual a 5
- para tabelas de dimensões maiores :
  - o teste χ^2 é válido se não mais do que 20% dos valores esperados forem menores do que 5 e nenhum for menor do que 1.

Caso o teste χ^2 não seja adequado, uma opção é utilizar o teste exato de Fisher obtido com o subcomando **`exact`**.

```
tab  tmc  fal,  row  exact
```

```
transtorno | falta ao trabalho no
    mental |      último mes
     comum |       não        sim |     Total
-----------+----------------------+----------
       não |       485        160 |       645
           |     75.19      24.81 |    100.00
-----------+----------------------+----------
       sim |       100         55 |       155
           |     64.52      35.48 |    100.00
-----------+----------------------+----------
     Total |       585        215 |       800
           |     73.13      26.88 |    100.00

           Fisher's exact =                 0.009
   1-sided Fisher's exact =                 0.005
```

Relacionando três variáveis categóricas

Utilize o comando **tabulate**, com a opçãp **if**(se), como mostrado a seguir:

tab tmc fal if fun==1, row chi

```
transtorno | falta ao trabalho no
    mental |      último mes
     comum |       não        sim |     Total
-----------+----------------------+----------
       não |       203         76 |       279
           |     72.76      27.24 |    100.00
-----------+----------------------+----------
       sim |        62         36 |        98
           |     63.27      36.73 |    100.00
-----------+----------------------+----------
     Total |       265        112 |       377
           |     70.29      29.71 |    100.00

          Pearson chi2(1) =   3.1308   Pr = 0.077
```

tab tmc fal if fun==0, row chi

```
transtorno | falta ao trabalho no
    mental |      último mes
     comum |       não        sim |     Total
-----------+----------------------+----------
       não |       282         84 |       366
           |     77.05      22.95 |    100.00
-----------+----------------------+----------
       sim |        38         19 |        57
           |     66.67      33.33 |    100.00
-----------+----------------------+----------
     Total |       320        103 |       423
           |     75.65      24.35 |    100.00

          Pearson chi2(1) =   2.8861   Pr = 0.089
```

Manipulação de variáveis contínuas

Construção de intervalos de confiança para a média

A média é uma medida pontual e não fornece nenhuma informação a respeito da variabilidade dos dados. Este procedimento não permite julgar qual a possível magnitude do erro que estamos cometendo. Daí surge a idéia de construir o intervalo de confiança, que é definido como o intervalo dentro do qual se encontra a verdadeira magnitude do efeito com um certo grau de certeza.

O comando abaixo ilustra a construção do intervalo de confiança (IC) para a média da variável **idade**.

```
ci  idade
```

```
    Variable |     Obs         Mean    Std. Err.       [95% Conf. Interval]
-------------+--------------------------------------------------------------
       idade |     800        37.69    .3721263        36.95954    38.42046
```

Com base na amostra deste estudo, podemos dizer, com 95% de confiança, que o verdadeiro valor para a idade média dos motoristas e cobradores está entre 37,0 e 38,4 anos.

Note que, quando não especificamos um determinado nível de confiança, o programa assume $\gamma = 95\%$ para o cálculo do intervalo. No entanto, é possível mudar este valor usando a opção **level**.

No exemplo abaixo, o IC foi construído com confiança de 90%.

```
ci  idade,  level(90)
```

```
    Variable |     Obs         Mean    Std. Err.       [90% Conf. Interval]
-------------+--------------------------------------------------------------
       idade |     800        37.69    .3721263         37.0772     38.3028
```

O IC também pode ser utilizado para testar se a média de interesse é estatisticamente igual, com um certo coeficiente de confiança, a um determinado valor de interesse.

De maneira análoga, podemos fazer um teste de hipótese para avaliar a mesma questão: “Será que a idade média dos motoristas e cobradores é estatisticamente diferente de 35 anos?”

Para isto, podemos usar o teste t de Student :

`ttest  idade = 35`

```
One-sample t test

------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
   idade |     800       37.69    .3721263    10.52532    36.95954    38.42046
------------------------------------------------------------------------------
Degrees of freedom: 799

                           Ho: mean(idade) = 35

     Ha: mean < 35               Ha: mean ~= 35               Ha: mean > 35
       t =   7.2287                t =   7.2287                 t =   7.2287
   P < t =   1.0000            P > |t| =   0.0000           P > t =   0.0000
```

Comparação entre médias de duas amostras independentes

Suponha agora que você queira avaliar se a idade média difere segundo a função do trabalhador. Neste caso, utiliza a opção **by**:

`ttest  idade,  by(fun)`

```
Two-sample t test with equal variances

------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
motorist |     423    40.74468    .4227253    8.694175    39.91377    41.57559
cobrador |     377     34.2626    .5833693    11.32698    33.11552    35.40967
---------+--------------------------------------------------------------------
combined |     800       37.69    .3721263    10.52532    36.95954    38.42046
---------+--------------------------------------------------------------------
    diff |            6.482081    .7097834                5.088818    7.875344
------------------------------------------------------------------------------
Degrees of freedom: 798

               Ho: mean(motorist) - mean(cobrador) = diff = 0

     Ha: diff < 0                 Ha: diff ~= 0                 Ha: diff > 0
       t =   9.1325                 t =   9.1325                 t =   9.1325
   P < t =   1.0000             P > |t| =   0.0000           P > t =   0.0000
```

Considerações a respeito da validade do teste t de Student

O teste t assume que a distribuição da variável resposta é aproximadamente normal e o desvio padrão é o mesmo em cada grupo a ser comparado.

Então, no caso acima, estamos assumindo que o desvio padrão da variável IDADE (variável resposta) é o mesmo para motoristas e cobradores. Esta suposição precisa ser verificada, o que pode ser feito com o comando:

```
sdtest idade, by(fun)
```

```
Variance ratio test

------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
motorist |     423    40.74468    .4227253    8.694175    39.91377    41.57559
cobrador |     377     34.2626    .5833693    11.32698    33.11552    35.40967
---------+--------------------------------------------------------------------
combined |     800       37.69    .3721263    10.52532    36.95954    38.42046
------------------------------------------------------------------------------

                    Ho: sd(motorist) = sd(cobrador)

            F(422,376) observed   = F_obs           =    0.589
            F(422,376) lower tail = F_L   = F_obs   =    0.589
            F(422,376) upper tail = F_U   = 1/F_obs =    1.697

   Ha: sd(1) < sd(2)         Ha: sd(1) ~= sd(2)          Ha: sd(1) > sd(2)
  P < F_obs = 0.0000     P < F_L + P > F_U = 0.0000     P > F_obs = 1.0000
```

Quando o teste acima (teste de homocedasticidade) indicar que as variâncias não são iguais nos dois grupos, devemos usar um teste que considere esta desigualdade. Isto pode ser feito com o uso da opção **unequal**:

```
ttest idade, by(fun) unequal
```

```
Two-sample t test with unequal variances

------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
motorist |     423    40.74468    .4227253    8.694175    39.91377    41.57559
cobrador |     377     34.2626    .5833693    11.32698    33.11552    35.40967
---------+--------------------------------------------------------------------
combined |     800       37.69    .3721263    10.52532    36.95954    38.42046
---------+--------------------------------------------------------------------
    diff |            6.482081    .7204279                5.06763    7.896533
------------------------------------------------------------------------------
Satterthwaite's degrees of freedom:  702.063

                Ho: mean(motorist) - mean(cobrador) = diff = 0

     Ha: diff < 0                 Ha: diff ~= 0                 Ha: diff > 0
       t =   8.9975                 t =   8.9975                 t =   8.9975
   P < t =   1.0000             P > |t| =   0.0000             P > t =   0.0000
```

Comparação entre médias de duas amostras dependentes

Quando as amostras não são independentes dizemos que as observações são correlacionadas e neste caso, o teste *t-pareado* é mais indicado pois leva em conta a correlação existente entre as observações.

Um exemplo de amostras dependentes é o estudo onde dois observadores diferentes fizeram medições da prega cutânea de 15 indivíduos distintos. As medidas são observadas no mesmo indivíduo, portanto, dizemos que as amostras dos 2 observadores são dependentes.

O banco de dados do estudo descrito anteriormente chama-se ***Prega.dta***. Neste arquivo, os valores foram cadastrados de modo que cada indivíduo tem seus dados representados em uma coluna diferente. As variáveis são descritas a seguir:

id = identificação do indivíduo
observA = medida da prega cutânea segundo o observador A
observB = medida da prega cutânea segundo o observador B

Para realizar o teste t-pareado basta digitar

```
ttest  observa=observb
```

```
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
 observa |      15    23.84667    2.041145    7.905321    19.46885    28.22449
 observb |      15    21.56667    1.842221    7.134891     17.6155    25.51784
---------+--------------------------------------------------------------------
    diff |      15        2.28    .5819672    2.253949    1.031805    3.528196
------------------------------------------------------------------------------

               Ho: mean(observa - observb) = mean(diff) = 0

  Ha: mean(diff) < 0         Ha: mean(diff) ~= 0         Ha: mean(diff) > 0
       t =   3.9177               t =   3.9177               t =   3.9177
   P < t =   0.9992           P > |t| =   0.0015           P > t =   0.0008
```

Observando o resultado acima, o que você conclui?

Teste de Normalidade

A distribuição normal é descrita por dois parâmetros: a média e o desvio padrão da população e apresenta as características: (a) a média, mediana e a moda coincidem;

(b) a curva é simétrica ao redor da média; e (c) as extremidades da curva, em ambos os lados da média, se estendem próximas da linha do eixo-x, sem nunca tocá-la.

Para exemplificar, utilizaremos o banco de dados **lbw.dta**, estudo com crianças de baixo peso ao nascer. Inicialmente, elaboramos as medidas de tendência central para as variáveis: *lwt*(peso da mãe) e *bwt*(peso do bebe) com o comando *summarize*.

sum lwt bwt, det

lwt

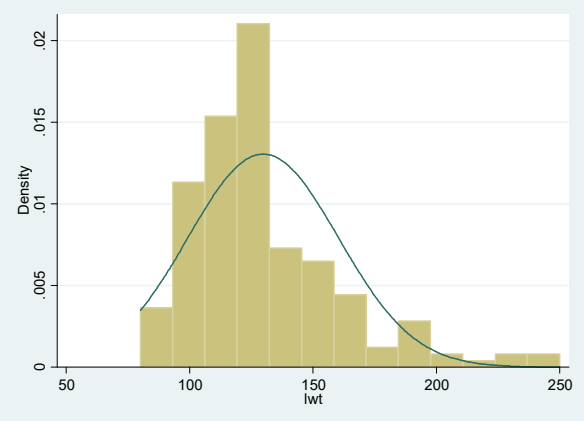
Obs	189	Mean	129.8148
Std. Dev.	30.57938	Variance	935.0985
Skewness	1.390855	Kurtosis	5.309181
Percentiles			
25%	110		
50%	121		
75%	140		

bwt

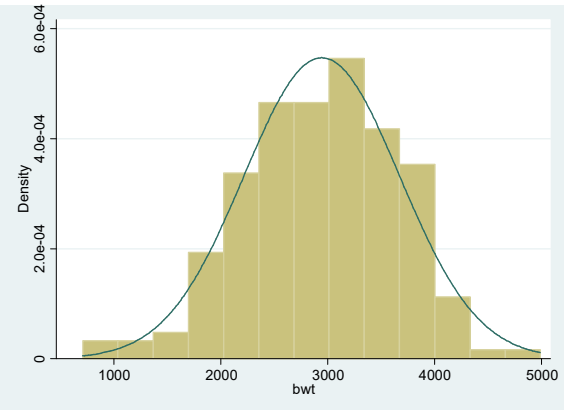
Obs	189	Mean	2944.656
Std. Dev.	729.0224	Variance	531473.7
Skewness	-.2084993	Kurtosis	2.889143
Percentiles			
25%	2414		
50%	2977		
75%	3475		

Elaborar gráficos para as variáveis peso da mãe e peso do bebe:

histogram lwt, normal



histogram bwt, normal



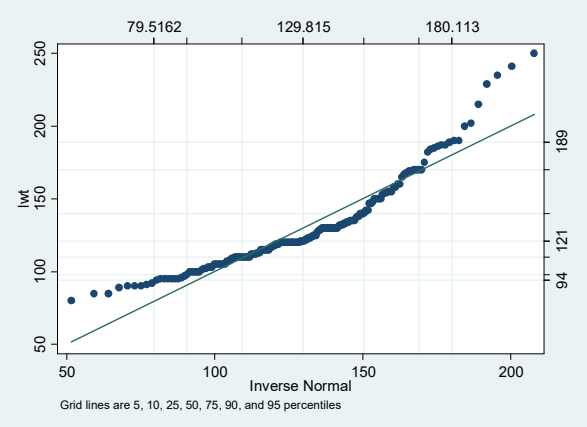
Teste de *Shapiro-Wilk*, testa a hipótese de que os dados da amostra estão normalmente distribuídos.

swilk lwt bwt

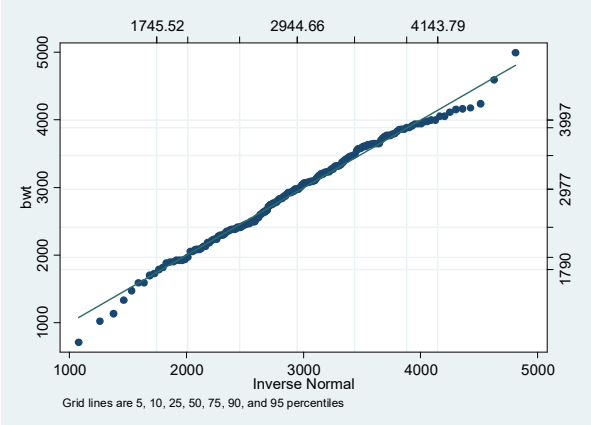
```
                    Shapiro-Wilk W test for normal data

    Variable        |     Obs       W           V         z       Prob>z
--------------------+-------------------------------------------------------
(peso da mãe) lwt   |     189    0.89396     15.060     6.222     0.00000
(peso do bebe) bwt  |     189    0.99258      1.053     0.119     0.45247
```

Observando os resultados, o valor de p < 0.05, indica que os dados(peso da mãe) se afastam da curva normal e o valor de p = 0,45247 para o peso dos bebes, indica que os dados não se afastam da curva normal. Os gráficos abaixo ilustram os resultados:

qnorm lwt , grid



qnorm bwt , grid



Análise de Variância - ANOVA

O teste paramétrico ANOVA(análise de variância), testa a diferença entre as médias de três ou mais grupos independentes. *One-way* ANOVA, com um fator compara o egeito de uma variável preditora(independente) sobre uma variável contínua(desfecho). A análise de variância, calcula a diferença do valor observado de cada indivíduo em

relação à média de seu grupo e a diferença do valor observado de cada indivíduo em relação à média total. O teste **F** para duas variâncias comporta uma razão, cujo numerador representa a variância entre as médias comparadas(variância entre os grupos) e o denominador, a variância entre os indivíduos dentro de cada grupo. Se a variabilidade entre os grupos for significativamente maior do que a variabilidade dentro das amostras, há indícios de que pelo menos duas médias diferem entre si. Se as variâncias dentro dos grupos é igual à variância entre os grupos, o valor **F** será igual a 1, indicando que não há diferença significativa entre os grupos.

Para exemplificar o teste ANOVA, observe a média de peso ao nascer(*bwt*) nos três grupos da variável raça(*race*):

tab race, sum(bwt)

```
           |           Summary of bwt
      race |        Mean   Std. Dev.       Freq.
-----------+------------------------------------
         1 |   3103.7396   727.72424          96
         2 |   2719.6923   638.68388          26
         3 |   2804.0149   721.30115          67
-----------+------------------------------------
     Total |   2944.6561   729.02242         189
```

O comando abaixo testa a diferença entre as médias de peso do bebe*(bwt*) entre as três categorias da variável raça(*race*):

oneway bwt race

```
                       Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      5070607.63      2   2535303.82      4.97     0.0079
 Within groups        94846445    186   509927.124
------------------------------------------------------------------------
    Total           99917052.6    188   531473.684

Bartlett's test for equal variances:  chi2(2) =   0.6545  Prob>chi2 = 0.721
```

Os resultados exibem significância estatística: F(2,186)=4,97 e p=0,0079. Indicando que há, pelo menos, uma diferença entre as raças testadas, rejeitando-se a hipótese nula(a variabilidade entre os grupos é grande em relação a variabilidade dentro das raças). Significando que ao nascer o peso médio dos bebes não é igual nas três raças.

Para identificar onde está a diferença, são necessários testes de múltiplas comparações de médias. Métodos mais conhecidos: *Scheffe* e *Bonferroni*.

oneway bwt race, bonferroni

```
                  Analysis of Variance  (omitida)

                          Comparison of bwt by race
                                (Bonferroni)
Row Mean-|
Col Mean |     white      black
---------+----------------------
  black  |   -384.047
         |      0.048
         |
  others |   -299.725    84.3226
         |      0.027      1.000
```

Diferenças significativas entre as raças: white com black e white com others.

oneway bwt race, scheffe

```
                 Analysis of Variance   (omitida)

                          Comparison of bwt by race
                                  (Scheffe)
Row Mean-|
Col Mean |    white      black
---------+----------------------
  black  |   -384.047
         |      0.054
         |
  others |   -299.725    84.3226
         |      0.033      0.878
```

Diferenças significativas entre as raças: . white com others.

Conversão de banco de dados - programa Stat/Transfer

Como foi comentado anteriormente, o STATA trabalha apenas com bancos de dados no formato "**dta**". O banco de dados que iremos utilizar agora **(*Plasma.xls*)** está no formato EXCEL e, portanto, deve ser convertido para o formato de um banco de dados do STATA. A conversão deve ser feita por meio do STAT/TRANSFER. Antes de inicializar o STATA, utilize-o para converter o arquivo *Plasma.xls* em *Plasma.dta*.

Clique duas vezes no ícone **Stat Transfer** na área de trabalho

Na opção *transfer*, há as seguintes alternativas:

Input file type: das várias opções, escolha *Excel*

File specification: clique em *Browse* para localizar o arquivo **Plasma.xls**.

Output file type: das várias opções, escolha *STATA.*

File specification: exibi a pasta e o nome do arquivo convertido.

Clique em **Transfer.** Quando o programa terminar clique em EXIT.

Relação entre duas variáveis contínuas

Correlação linear de Pearson

Em muitas situações, é de interesse quantificar a força da relação linear entre duas variáveis contínuas, sem designar uma como resposta e outra como explicativa.

O grau desta associação pode ser medido com o uso do coeficiente de correlação linear de Pearson (***r***), que leva este nome pois foi descrito por Pearson. A correlação entre duas variáveis é positiva se valores mais altos de uma variável estão associados a valores mais altos da outra, e é negativa se os valores de uma variável crescem enquanto os da outra diminuem. O coeficiente de correlação próximo do zero significa que não existe uma relação linear entre as duas variáveis.

O coeficiente de correlação varia de –1 a +1, sendo:

+1: associação positiva perfeita

0: ausência de associação

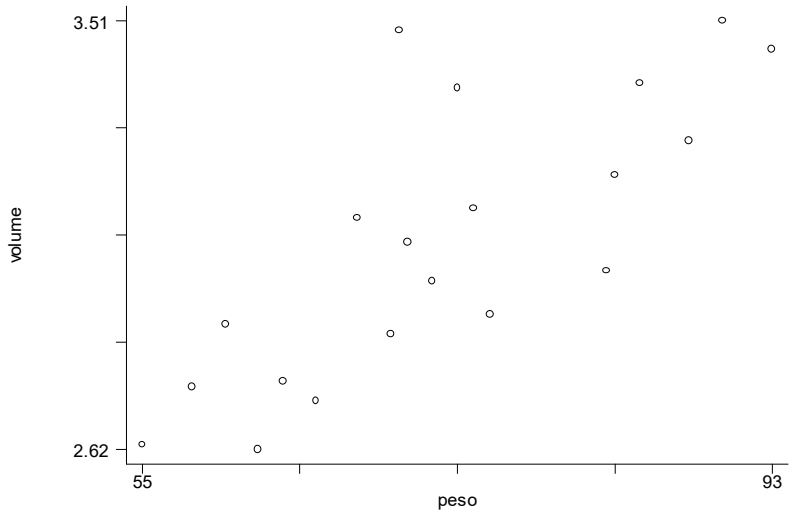
-1: associação negativa perfeita

Aplicação prática:

Utilizando o banco de dados **plasma.dta** vamos verificar se existe uma relação linear entre as variáveis volume plasmático e peso.

A melhor forma de iniciar o estudo da possível relação entre estas duas variáveis contínuas é construir um gráfico de dispersão, utilizando os comandos:

```
scatter volume  peso, ytitle(volume(litros)) xtitle(peso(kg))
```



Observando o gráfico acima, você acha que existe uma correlação linear entre o volume plasmático e o peso dos vinte homens incluídos neste banco de dados?

Para obter o valor do coeficiente de correlação de Pearson podemos utilizar o comando **correlate** (que pode ser abreviado como **corr**):

```
corr  peso  volume
```

```
 (obs=20)

             |     peso   volume
-------------+------------------
        peso |   1.0000
      volume |   0.7803   1.0000
```

A saída apresenta o número de sujeitos utilizados para o cálculo (obs = 20) e o coeficiente de correlação linear entre as variáveis peso e volume, isto é, r = 0,78.

É possível obter os coeficientes de correlação linear entre muitas variáveis contínuas do mesmo banco. Para isto, basta digitar os nomes das variáveis após o comando **corr** (por exemplo, **corr var1 var2 var3 . . .**).

Pode ser usado também o comando **pwcorr** (*pairwise correlation*), que produz o mesmo resultado e permite o uso da opção **sig** que apresenta o nível de significância do coeficiente de correlação apresentado.

```
pwcorr volume peso, sig
```

```
             |   volume     peso
-------------+------------------
      volume |   1.0000
             |
        peso |   0.7803   1.0000
             |   0.0000
             |
```

A saída acima apresenta, abaixo do coeficiente de correlação (***r*** = 0,78), o nível de significância (p = 0,0000).

Regressão linear

A regressão linear apresenta a equação da reta que melhor descreve como a variável *y* aumenta (ou diminui) com um aumento na variável *x*. A escolha de qual será a variável a ser chamada de *y* é importante porque, diferentemente da correlação, as duas alternativas não fornecem o mesmo resultado. A variável *y* é comumente denominada variável **dependente**, e *x* é a variável **independente** ou **explicativa**. A técnica de regressão linear permite:

- estudar a forma da relação entre *x* e *y*, e

- obter o valor esperado de *y* quando conhecemos apenas o valor de *x*.

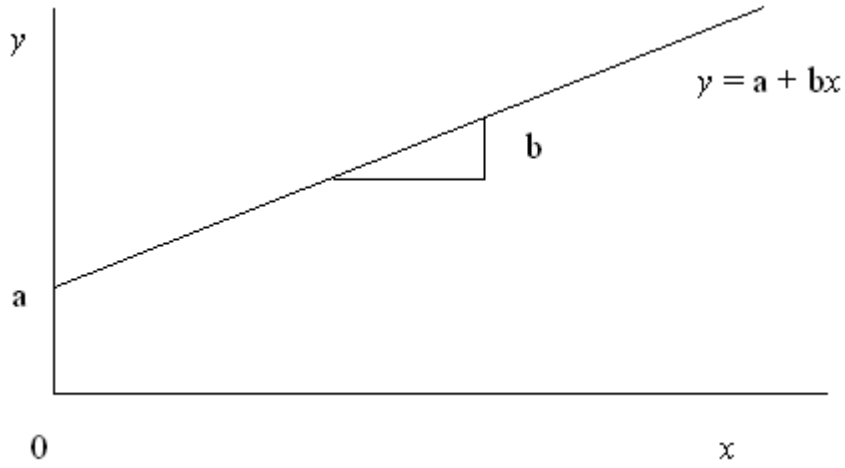
A equação da reta de regressão é:

$$y = a + bx$$

onde **a** é o intercepto e **b** é a inclinação da reta.

a (intercepto): é o ponto onde a reta cruza o eixo *y* e mostra o valor de *y* para *x*=0.

b (inclinação): mostra o aumento em *y* correspondente ao incremento de uma unidade em *x*.



Aplicação prática:

Utilizando os dados de nosso arquivo ***plasma.dta*** vamos utilizar a técnica de regressão linear para obter a reta que melhor exprime a relação linear entre o peso e o volume plasmático dos indivíduos incluídos no banco de dados. Nossa variável independente (*x*) será o peso e a variável dependente (ou resposta) será o volume plasmático (*y*).

Para fazer a regressão linear no STATA utilizaremos o comando **regress**. Para executarmos este comando, a variável dependente aparece em primeiro lugar, seguida da variável explicativa:

```
regress volume peso
```

```
      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  1,    18) =   28.03
       Model |  .967837779     1  .967837779           Prob > F      =  0.0000
    Residual |  .621562203    18  .034531234           R-squared     =  0.6089
-------------+------------------------------           Adj R-squared =  0.5872
       Total |  1.58939998    19  .083652631           Root MSE      =  .18583

------------------------------------------------------------------------------
      volume |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        peso |   .0204617    .003865     5.29   0.000     .0123417    .0285817
       _cons |   1.552716   .2858553     5.43   0.000     .9521564    2.153276
------------------------------------------------------------------------------
```

O resultado deste comando consiste em duas partes. Na primeira há uma tabela que fornece a quantidade de variação da variável **volume** explicada pelo modelo de regressão linear.

A segunda parte do resultado mostra os valores estimados para os parâmetros. O valor estimado para o parâmetro correspondente ao intercepto **a** é chamado **_cons** (constante). O valor estimado do parâmetro **b** é o coeficiente para o **peso**. Na maioria das vezes este é o parâmetro de maior interesse e pode ser chamado de coeficiente de regressão do volume plasmático com o peso.

Na saída apresentada acima, o valor estimado de a (**_cons**) é 1,55 e o valor estimado de b (peso) é 0,02.

A partir da equação geral y = **a** + **b***x*, podemos escrever a equação de regressão utilizando as estimativas obtidas:

$$\text{volume} = 1{,}55 + 0{,}02*(\text{peso})$$

Próximo às estimativas dos parâmetros estão os erros padrão (EP) e os correspondentes testes t e valores de p, que nos ajudam a decidir se cada parâmetro é significantemente diferente de zero. O teste para o coeficiente de regressão é o teste da hipótese nula, ou seja, de não existir relação linear. Finalmente, temos os intervalos de confiança (IC95%) dos valores dos parâmetros estimados.

Observando a saída acima, quais são os EP dos parâmetros estimados e quão forte é a evidência de que existe uma associação linear entre estas duas variáveis?

Depois de ajustar a reta de regressão, é possível calcular o volume plasmático previsto pelo modelo, dado o peso de cada indivíduo, utilizando o seguinte comando:

predict Y

O comando acima gera uma nova variável (de nome Y) onde ficam guardados os valores previstos dos volumes plasmáticos para cada peso observado. Para obter uma lista das 10 primeiras observações digite:

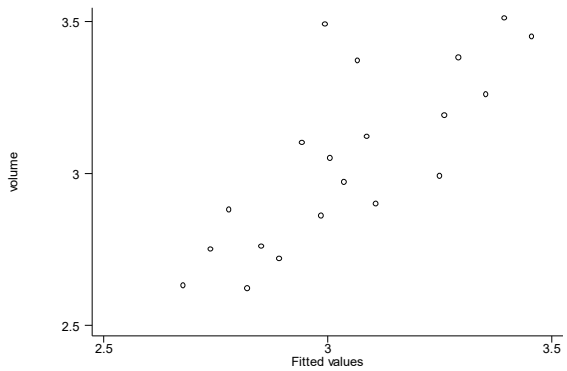
list Y peso in 1/10

```
            Y       peso
 1.  2.739494         58
 2.  2.985034         70
 3.  2.892956       65.5
 4.  3.066881         74
 5.  2.852033       63.5
 6.  2.821341         62
 7.  2.995265       70.5
 8.  3.005496         71
 9.  2.944111         68
10.   3.29196         85
```

Uma maneira descritiva de estudar a adequação do modelo adotado é exibir o diagrama de dispersão dos valores previstos versus os valores observados:

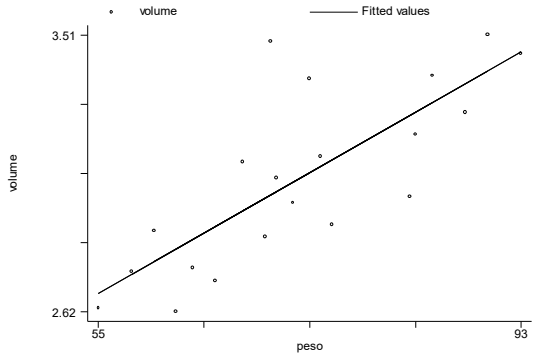
scatter volume Y

O gráfico obtido foi:



Finalmente, para construirmos o gráfico de dispersão mostrando os dados e a reta de regressão ajustada ao modelo utilizamos o comando:

```
scatter volume Y peso, c(. l) s(o i)
```

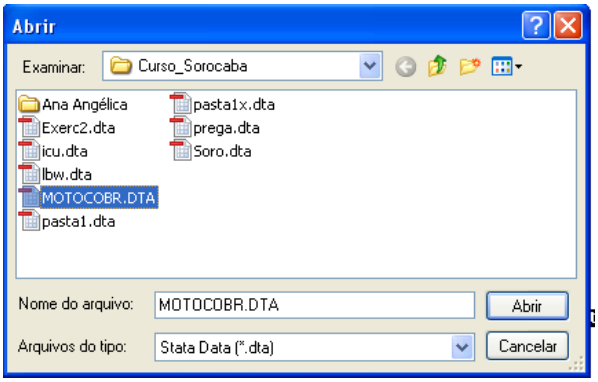


c(. l) significa “não conecte volume” e “conecte Y (valores previstos do volume)”.
s(o i) significa “use pequenos círculos para volume” e “use um símbolo invisível para Y”.

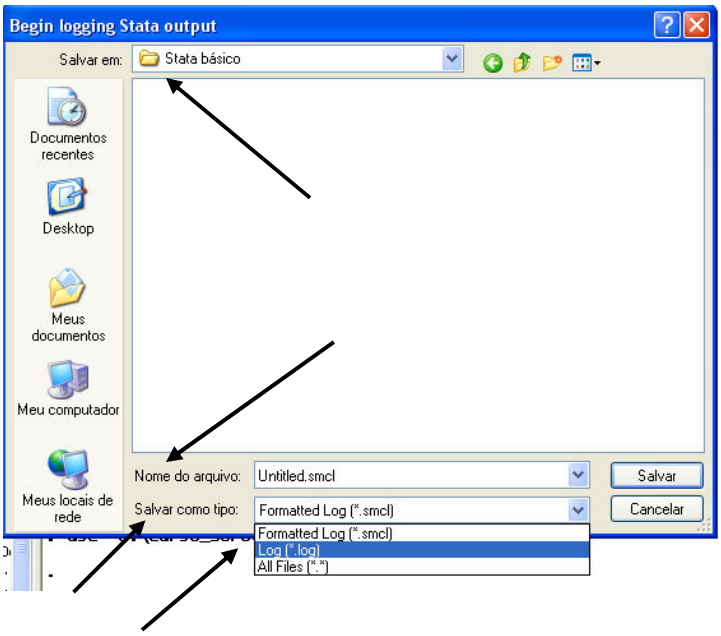
Elaborar tabelas de frequências simples

O primeiro passo para analisar um banco de dados: é produzir tabelas de frequencias simples para as variáveis categóricas e medidas de tendência central para as variáveis contínuas.Para exemplificar esta etapa de análise, vamos inicialmente ler(abrir) o banco de dados **MOTOCOBR.dta** em seguida abrir(criar) um arquivo-log(arquivo para armazenar os resultados).

1) Clique no botão *Open* para abrir o arquivo: **MOTOCOBR.dta**. O Stata lê/abre somente arquivos-dta.



2) Clique no botão *Log Begin* para criar o arquivo-log. escolha a opção *log* em tipo do arquivo e digite um nome para o arquivo-log, por ex.: tabelas.



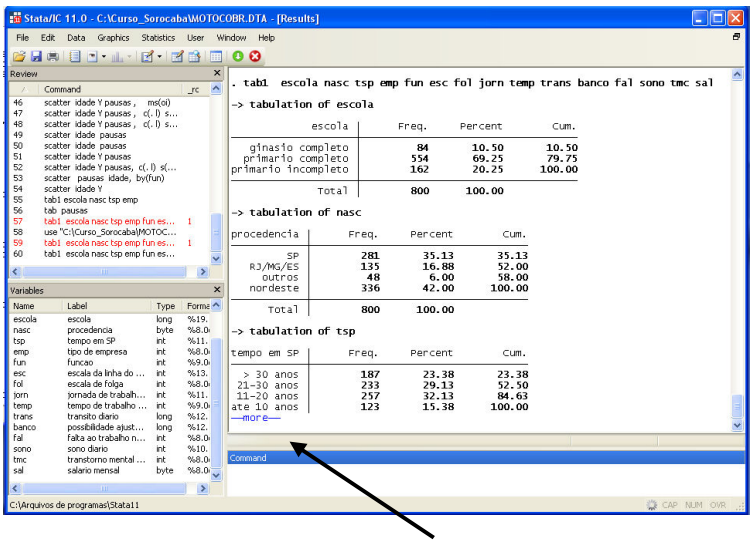
Todas as tabelas serão armazenadas no arquivo **tabelas.log**. Este arquivo poderá ser aberto no aplicativo Word.

3) **O comando tab1**, produz tabelas de freqüências simples para as variáveis relacionadas. É aconselhável relacionar as variáveis categóricas.

Sintaxe: **tab1 varlist [if] [in] [weight] [, tab1_options]**

Digite o comando abaixo para obter as tabelas.

tab1 escola nasc tsp emp fun esc fol jorn temp trans banco fal sono tmc sal



Clique em "*more*" para prosseguir.

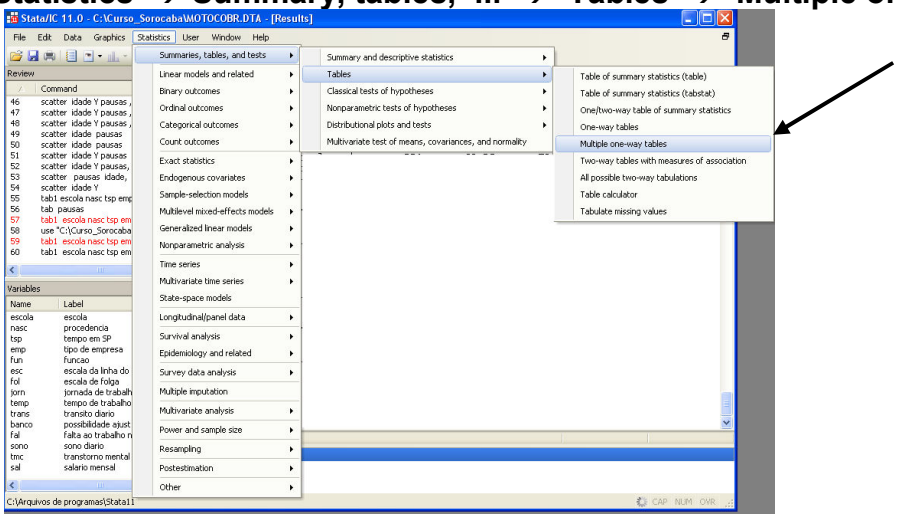
Observe os resultados incluindo a opção: ***nolabel***. Esta opção é utilizada para exibir os códigos(*values*) das variáveis ao invés dos nomes das categorias(*value label*).

tab1 nasc fun , nolabel

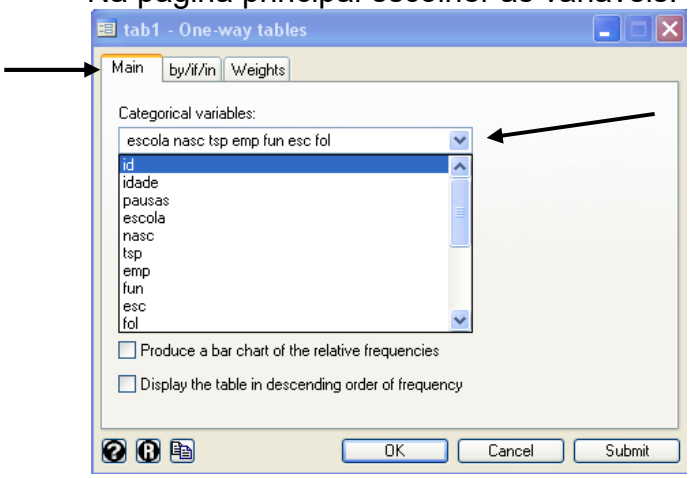
```
procedencia |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         281       35.13       35.13
          1 |         135       16.88       52.00
          2 |          48        6.00       58.00
          3 |         336       42.00      100.00
------------+-----------------------------------
      Total |         800      100.00

-> tabulation of fun

     funcao |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         423       52.88       52.88
          1 |         377       47.13      100.00
------------+-----------------------------------
      Total |         800      100.00
```

Outra maneira de obter as tabelas é utilizar a barra de menus:

Statistics → Summary, tables, ... → Tables → Multiple one-way tables

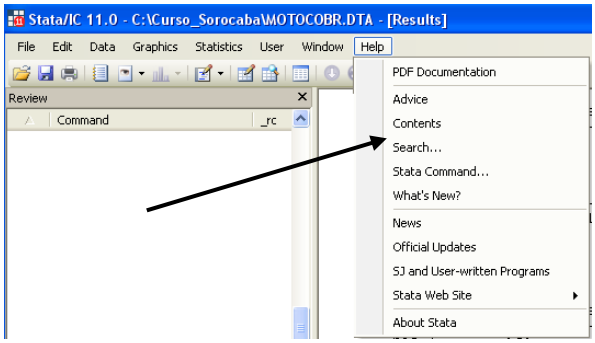


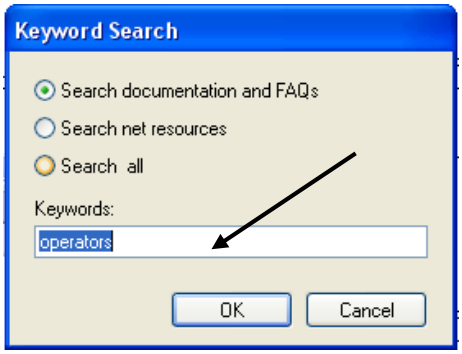
Na página principal escolher as variáveis.



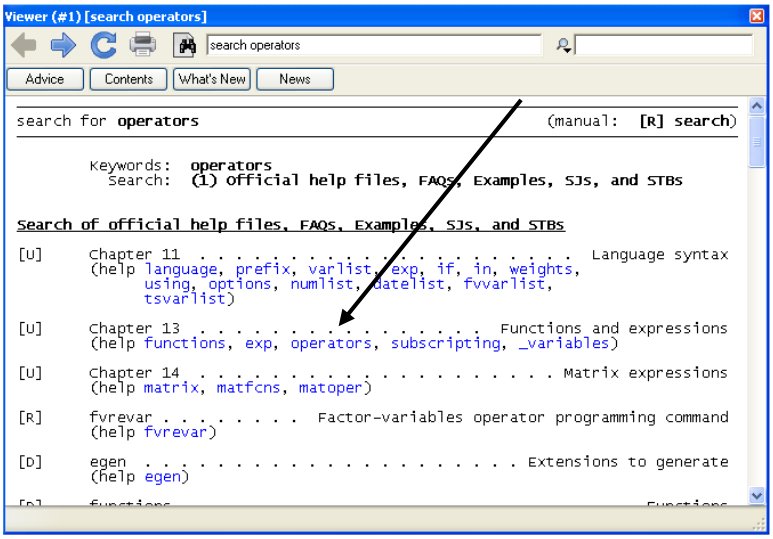
HELP - O Stata possui um sistema de busca muito útil para auxiliar na utilização dos comandos. Por exemplo: Quais são os operadores lógicos utilizados pelo Stata?

Na barra de menus, clique em *Help* e escolha a opção *Search*, digite o conteúdo da busca e clique em *OK* para continuar.





O Stata exibirá a tela:



Clique em ***operators*** para visualizar os operadores lógicos.

```
     Arithmetic               Logical            (numeric and string)
--------------------     ------------------     ---------------------
 +   addition              &   and                >   greater than
 -   subtraction           |   or                 <   less than
 *   multiplication        !   not                >=  > or equal
 /   division              ~   not                <=  < or equal
 ^   power                                        ==  equal
 -   negation                                     !=  not equal
 +   string concatenation                        ~=  not equal

    A double equal sign (==) is used for equality testing.

Examples

. sysuse auto                                . generate weight2 = weight^2
. count if rep78 > 4                         . count if rep78 > 4 & weight < 3000
. list make if rep78 == 5 | mpg > 25
```

Após a observação da utilização do comando *Help*, um procedimento muito utilizado nas análises é quando queremos elaborar tabelas apenas para um sub-grupo da população estudada.

Frequência de ***tmc*** somente para os motoristas(variável ***fun*** igual ao código 0):

tab1 tmc if fun == 0 (observe o if)

```
 transtorno |
     mental |
      comum |      Freq.     Percent        Cum.
------------+-----------------------------------
        nao |        366       86.52       86.52
        sim |         57       13.48      100.00
------------+-----------------------------------
      Total |        423      100.00
```

Frequência **nasc**(procedência) para os motoristas(variável ***fun*** igual a 0) e que apresentaram transtorno mental comum(variáveil ***tmc*** igual a 1):

tab1 nasc if fun == 0 & tmc == 1 (observe o if)

```
-> tabulation of nasc if fun == 0  &  tmc==1

procedencia |      Freq.     Percent        Cum.
------------+-----------------------------------
         SP |         16       28.07       28.07
   RJ/MG/ES |          9       15.79       43.86
     outros |          4        7.02       50.88
   nordeste |         28       49.12      100.00
------------+-----------------------------------
      Total |         57      100.00
```

O comando *summarize* ou simplesmente ***sum*** é utilizado para produzir, elaborar medidas de tendência central para as variáveis contínuas.

<u>Sintaxe:</u> **summarize [varlist] [if] [in] [weight] [, options]**

sum idade pausas

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       idade |       800       37.69    10.52532         17         67
      pausas |       800      3.5325    2.481258          1         17
```

sum idade pausas, detail

```
                          idade
-------------------------------------------------------------
      Percentiles      Smallest
 1%           21             17
 5%           22             19
10%           24             19       Obs                 800
25%           30             19       Sum of Wgt.         800
```

```
50%           37                      Mean              37.69
                        Largest       Std. Dev.      10.52532
75%           45             65
90%           53             66       Variance       110.7824
95%           58             66       Skewness        .440607
99%           63             67       Kurtosis       2.555018

                    numero de pausas dia
-------------------------------------------------------------
      Percentiles      Smallest
 1%             1             1
 5%             1             1
10%             1             1       Obs                 800
25%             2             1       Sum of Wgt.         800

50%             2                     Mean              3.5325
                        Largest       Std. Dev.      2.481258
75%             5             13
90%             7             14       Variance        6.15664
95%             8             15       Skewness       1.435872
99%          11.5             17       Kurtosis       5.380058
```

sum idade pausas if tmc==0

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       idade |       645    38.28372    10.38102         17         67
      pausas |       645    3.643411    2.569525          1         17
```

sum idade pausas if tmc==0 & fun==1

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       idade |       279    34.63799    11.23113         17         67
      pausas |       279    3.250896    2.263598          1         15
```

O comando *recode* é utilizado para agrupar, recodificar valores das variáveis.

Sintaxe: **recode varlist (rule) [(rule) ...] [, generate(newvar)]**

where the most common forms for rule are

rule	Example	Meaning
# = #	3 = 1	3 recoded to 1
# # = #	2 . = 9	2 and . recoded to 9
#/# = #	1/5 = 4	1 through 5 recoded to 4
nonmissing = #	nonmiss = 8	all other nonmissing to 8
missing = #	miss = 9	all other missings to 9

O comando *generate* ou ***gen*** é utilizado para criar, gerar novas variáveis.

Sintaxe: **generate [type] newvar[:lblname] = exp [if] [in]**

O comando *label* é utilizado para atribuir nomes aos códigos e as variáveis.

Sintaxes:

label variable varname ["label"] (nomear variável)

label define lblname # "label" [# "label" ...] (criar variável-label com nomes para os códigos)

label values varlist [lblname|.] [, nofix] (atribuir nomes aos códigos)

label drop {lblname [lblname ...] | _all} (apagar a variável label)

É aconselhável criar novas variáveis a partir das existentes e depois recodificá-las.

Exemplo: criar variável para faixas etárias. Utilizar o valor dos *quartis* para definir os limites para os intervalos.

Criar a variável *faet* a partir da variável idade:

gen faet = idade

Recodificar a variável faet:

recode faet min/30=1 31/37=2 38/45=3 46/max=4

Criar "*labels"* para a variável *faet*:

label var faet "Idade agrupada, segundo os quartis"

label define cfaet 1 "até 30" 2 "31-37" 3 "38-45" 4 "46 e mais"
label val faet cfaet

Elaborar tabela de freqüência para *faet*:

tab faet ou **tab1 faet**

```
      Idade |
  agrupada, |
 segundo os |
    quartis |      Freq.     Percent        Cum.
------------+-----------------------------------
     até 30 |        229       28.63       28.63
      31-37 |        191       23.88       52.50
      38-45 |        193       24.13       76.63
  46 e mais |        187       23.38      100.00
------------+-----------------------------------
      Total |        800      100.00
```

tab faet, nol ou **tab1 faet, nol**

Idade agrupada, segundo os quartis	Freq.	Percent	Cum.
1	229	28.63	28.63
2	191	23.88	52.50
3	193	24.13	76.63
4	187	23.38	100.00
Total	800	100.00	

Elaborar tabelas cruzadas(duas variáveis)

O comando *tabulate* ou ***tab*** produz tabelas cruzadas para duas variáveis.

Sintaxe: Two-way tables:
tabulate varname1 varname2 [if] [in] [weight] [, options]

Two-way tables for all possible combinations - a convenience tool:
tab2 varlist [if] [in] [weight] [, options]

Principais opções:

row	% na linha
col	% na coluna
cel	% do total
chi	calcular x^2 *de Pearson*
exact	calcular x^2 de Fisher

Elaborar tabela cruzada para escolaridade(escola) e transtorno mental(tmc):

tab escola tmc

escola	transtorno mental comum: nao	sim	Total
ginasio completo	70	14	84
primario completo	447	107	554
primario incompleto	128	34	162
Total	645	155	800

tab escola tmc, row col cel chi

```
+-------------------+
| Key               |
|-------------------|
|     frequency     |
|  row percentage   |
| column percentage |
|  cell percentage  |
+-------------------+

                    |   transtorno mental
                    |         comum
             escola |       nao        sim |     Total
--------------------+----------------------+----------
   ginasio completo |        70         14 |        84
                    |     83.33      16.67 |    100.00
                    |     10.85       9.03 |     10.50
                    |      8.75       1.75 |     10.50
--------------------+----------------------+----------
  primario completo |       447        107 |       554
                    |     80.69      19.31 |    100.00
                    |     69.30      69.03 |     69.25
                    |     55.88      13.38 |     69.25
--------------------+----------------------+----------
primario incompleto |       128         34 |       162
                    |     79.01      20.99 |    100.00
                    |     19.84      21.94 |     20.25
                    |     16.00       4.25 |     20.25
--------------------+----------------------+----------
              Total |       645        155 |       800
                    |     80.63      19.38 |    100.00
                    |    100.00     100.00 |    100.00
                    |     80.63      19.38 |    100.00

          Pearson chi2(2) =   0.6655   Pr = 0.717
```

Elaborar tabela cruzada para transtorno mental*(*tmc) e função(fun):

tab fun tmc, row col cel chi exact

```
           |   transtorno mental comum
           |
    funcao |       nao        sim |     Total
-----------+----------------------+----------
 motorista |       366         57 |       423
           |     86.52      13.48 |    100.00
           |     56.74      36.77 |     52.88
           |     45.75       7.12 |     52.88
-----------+----------------------+----------
  cobrador |       279         98 |       377
           |     74.01      25.99 |    100.00
           |     43.26      63.23 |     47.13
           |     34.88      12.25 |     47.13
-----------+----------------------+----------
     Total |       645        155 |       800
           |     80.63      19.38 |    100.00
           |    100.00     100.00 |    100.00
           |     80.63      19.38 |    100.00

          Pearson chi2(1) =  20.0012   Pr = 0.000
            Fisher's exact =                 0.000
```

A tabela aa apresenta a variável **tmc** com as categorias(0=não e 1=sim) .Para uma

interpretação mais adequada dos resultados e calcular: risco relativo e o *odds-ratio* será necessário inverter a ordem das categorias da variável **tmc**.

```
gen    tmc2 = tmc                                   (criar tmc2)
recode  tmc2   0=1  1=0                             (recodificar tmc2)
label define  ctmc2   0 sim    1 não                (criar variável de value-label)
label  val   tmc2   ctmc2                           (atribuir value-label a tmc2)
label  var   tmc2   "Transtorno Mental-comum"       (atribuir label a tmc2)
tab  fun  tmc2, row col cel chi exact
```

```
           |   Transtorno Mental-comum
    funcao |       sim        não |     Total
-----------+----------------------+----------
 motorista |        57        366 |       423
           |     13.48      86.52 |    100.00
           |     36.77      56.74 |     52.88
           |      7.12      45.75 |     52.88
-----------+----------------------+----------
  cobrador |        98        279 |       377
           |     25.99      74.01 |    100.00
           |     63.23      43.26 |     47.13
           |     12.25      34.88 |     47.13
-----------+----------------------+----------
     Total |       155        645 |       800
           |     19.38      80.63 |    100.00
           |    100.00     100.00 |    100.00
           |     19.38      80.63 |    100.00

          Pearson chi2(1) =  20.0012   Pr = 0.000
           Fisher's exact =                 0.000
   1-sided Fisher's exact =                 0.000
```

Odds ratio(OR) e Risco relativo(RR).

As medidas de força de associação, são utilizadas para medir a associação da variável desfecho com a variável de exposição, isto é, o quanto da probabilidade de ocorrência da variável dependente se deve à sua relação com a variável independente. Duas medidas de associação são utilizadas: *odds ratio*(*OR*) e risco relativo(*RR*). Quando o valor da estimativa for próximo de 1,0, tem-se uma indicação de não-associação. Valores >1 indicam risco maior de ocorrência de desfecho entre os expostos, enquanto valores < 1 indicam proteção para o desfecho entre os expostos.

Odds ratio(*OR*)
(Estudos: caso-controle)

	Casos	Controles
Expostos	a	b
Não-expostos	c	d

OR = ad / bc

Risco relativo (*RR*)
(Estudos: coorte)

	Doença	
	Presença	Ausência
Expostos	a	b
Não-expostos	c	d

RR = a/(a + b) / c/(c + d)

Recodificar a variável **fun** para calcular **OR** e **RR** para os motoristas:

gen fun2=fun
recode fun2 0=1 1=0

Exemplo: Calcular o *odds ratio*(OR) com o comando: **cc**

cc tmc fun2

```
                                                        Proportion
                 |   Exposed   Unexposed  |      Total     Exposed
-----------------+------------------------+------------------------
           Cases |        57          98  |        155      0.3677
        Controls |       366         279  |        645      0.5674
-----------------+------------------------+------------------------
           Total |       423         377  |        800      0.5288
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
      Odds ratio |         .4433757       |    .3029976    .6458508 (exact)
 Prev. frac. ex. |         .5566243       |    .3541492    .6970024 (exact)
 Prev. frac. pop |         .3158519       |
                 +-------------------------------------------------
                             chi2(1) =    20.00  Pr>chi2 = 0.000
```

Exemplo: Calcular o risco relativo(RR) com o comando: **cs**

cs tmc fun2

```
                 | fun3                   |
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+------------
           Cases |        57          98  |        155
        Noncases |       366         279  |        645
-----------------+------------------------+------------
           Total |       423         377  |        800
                 |                        |
            Risk |   .1347518    .2599469  |     .19375
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
 Risk difference |        -.1251952       |   -.1801411   -.0702493
      Risk ratio |         .5183818       |    .3857579    .6966021
 Prev. frac. ex. |         .4816182       |    .3033979    .6142421
 Prev. frac. pop |         .2546556       |
                 +-------------------------------------------------
                             chi2(1) =    20.00  Pr>chi2 = 0.0000
```

Calcular **OR** e **RR** para os cobradores:

cc tmc fun (calcular o **OR**)

cs tmc fun (calcular o **RR**)

Medidas de concordância(coeficiente de Kappa).

A estatística *kappa*, é baseada no número de respostas concordantes, ou seja, no número de casos cujo resultado é o mesmo entre os observadores. Mede o grau de concordância além do que seria esperado tão somente pelo acaso. O *kappa* varia de +1 a -1. O valor +1 representa a total concordância; o valor 0 significa a inexistência de relação entre as classificações e o valor -1 significa que as classificações são opostas. Interpretação da estatística kappa:

< 0,40	pequena concordância
0,41 a 0,60	concordância regular
0,61 a 0,80	boa concordância
> 0,80	excelente concordância

Utilizando o banco de dados **kappa.dta**, estudo de avaliação realizada por dois radiologistas em mamografias(1-normal, 2-doença benigna, 3-suspeita de cancer e 4-cancer). A variável *rada*, indica a classificação do radiologista-A e *radb* do radiologista-B.

tab rada radb

Radiologist A's assessment	Radiologist B's assessment: Normal	benign	suspect	cancer	Total
Normal	21	12	0	0	33
benign	4	17	1	0	22
suspect	3	9	15	2	29
cancer	0	0	0	1	1
Total	28	38	16	3	85

kap rada radb

Agreement	Expected Agreement	**Kappa**	Std. Err.	Z	Prob>Z
63.53%	30.82%	**0.4728**	0.0694	6.81	0.0000

O coeficiente kappa = 0,473, indica que houve 47,3% de acertos(coincidência) entre os dois radiologista, uma concordância moderada(regular).

Testes Não-paramétricos

Os testes não-paramétricos são métodos que compreendem procedimentos que não necessitam do cumprimento de todas as suposições restritivas dos testes paramétricos. São testes que podem ser utilizados com variáveis nominais, ordinais ou quantitativas, são menos robustos que os testes paramétricos. O teste qui-quadrado é um teste não–paramétrico. Utilizar o banco de dados: l**bw.dta**(estudo com crianças de baixo peso).

(1) Teste de Mann-Whitney é o substituto do teste *t* para amostras independentes quando há ruptura dos pressupostos paramétricos. A variável deve ser ordinal ou quantitativa.

Comparar a idade da mãe entre os bebes que apresentaram baixo peso ao nascer(*low*=1) e aqueles que não apresentaram baixo peso(l*ow*=0).

ranksum age, by(low)

```
Two-sample Wilcoxon rank-sum (Mann-Whitney) test

         low |      obs    rank sum    expected
-------------+---------------------------------
           0 |      130       12753       12350
           1 |       59        5202        5605
-------------+---------------------------------
    combined |      189       17955       17955

unadjusted variance   121441.67
adjustment for ties     -483.85
                     ----------
adjusted variance     120957.81

Ho: age(low==0) = age(low==1)
              z =   1.159
    Prob > |z| =   0.2466
```

O teste de *Mann-Whitney*, avalia a semelhança entre as duas séries por meio do grau de intersecção dos *ranks*(postos). Avaliando o teste, pode-se observar que não há diferença entre os grupos p=0,247, pois a mediana dos grupos são semelhantes, isto é mediana para baixo peso é 22 e não baixo peso é 23. Pode-se calcular a mediana com o comandos:

sort low

by low:sum age, det

```
-> low = 0(não)
-------------------------------------------------------------------------------
Obs    130     Mean   23.66154        50%     23

-> low = 1(sim)
-------------------------------------------------------------------------------
Obs     59     Mean   22.30508        50%     22
```

(2) Teste de Kruskal-Wallis conhecido como análise de variância(ANOVA) não-paramétrica em analogia ao teste paramétrico ANOVA de um critério de classificação(*one-way*).

Comparar o peso dos bebes(*bwt*) entre os grupos de etnia(*race*).

kwallis bwt , by(race)

```
Kruskal-Wallis equality-of-populations rank test

  +----------------------+
  | race | Obs | Rank Sum |       chi-squared =     8.590 with 2 d.f.
  |                       |       probability =     0.0136
  |------+-----+----------|
  |    1 |  96 | 10193.00 |
  |    2 |  26 |  2012.00 |      chi-squared with ties =  8.591 with 2 d.f.
  |    3 |  67 |  5750.00 |      probability =     0.0136
  +----------------------+
```

O teste de Kruskal-Wallis, p=0,0136, indica a existência de diferença significativa nos pesos dos bebes entre as etnias.

(3) Teste de Wilcoxon é o substituto do teste *t* para duas amostras dependentes(amostras emparelhadas), baseia-se nos postos(*ranks*) das diferenças dos valores de cada par de observação. A variável deve ser ordinal ou quantitativa.

Digite o comando **clear** para limpar a memória em seguida digite as linhas abaixo para criar o banco de dados de um ensaio clínico placebo para testar droga para induzir o sono.

```
input    paciente    droga    placebo

  1     6.1     5.2
  2       7     7.9
  3     8.2     3.9
  4     7.6     4.7
  5     6.5     5.3
  6     8.4     5.4
  7     6.9     4.2
  8     6.7     6.1
  9     7.4     3.8
 10     5.8     6.3
end

list

        (saída omitida)
```

Comparar o uso de uma droga e placebo para induzir o sono em dez pacientes.

signrank droga = placebo

```
    Wilcoxon signed-rank test

        sign |      obs   sum ranks    expected
-------------+--------------------------------
    positive |        8        50.5        27.5
    negative |        2         4.5        27.5
        zero |        0           0           0
-------------+--------------------------------
         all |       10          55          55

unadjusted variance       96.25
adjustment for ties       -0.13
adjustment for zeros       0.00
                        ----------
adjusted variance         96.13

Ho: droga = placebo
             z =   2.346
    Prob > |z| =   0.0190
```

O teste p=0,019, indica que a droga para induzir o sono é mais efetiva(eficaz) do que o placebo.

(4) Coeficiente de correlação de Spearman utilizado para avaliar a força da associação entre variáveis ordinais e quantitativas, é o substituto mais usado para o coeficiente de correlação de Pearson em caso de ruptura dos pressupostos paramétricos.

Para exemplificar o uso do coeficiente de Spearman, será avaliada a correlação entre peso da mãe e o peso do bebe ao nascer. Arquivo **lbw.dta**.

spearman lwt bwt

```
Number of obs =       189
Spearman's rho =       0.2483

Test of Ho: lwt and bwt are independent
    Prob > |t| =       0.0006
```

O coeficiente é 0.2483. É um valor baixo mas é significante($p<0.05$).

(5) Teste de Mcnemar esse teste é empregado em estudos com variáveis binárias(sim/não) emparelhadas, isto é, deseja-se saber se a proporção de participantes se altera com uma característica em estudo após uma intervenção ou o transcurso do tempo. As medidas coletadas não são independentes.

Para exemplificar o teste de Mcnemar, vamos verificar a relação entre baixo peso ao nascer(*low*) e o hábito de fumar da mãe(*smoke*).

tab1 low smoke

```
        low |      Freq.     Percent        Cum.
------------+-----------------------------------
  (não)   0 |        130       68.78       68.78
  (sim)   1 |         59       31.22      100.00
------------+-----------------------------------
      Total |        189      100.00

      smoke |      Freq.     Percent        Cum.
------------+-----------------------------------
  (não)   0 |        115       60.85       60.85
  (sim)   1 |         74       39.15      100.00
------------+-----------------------------------
      Total |        189      100.00
```

. mcc low smoke

```
                | Controls                |
Cases           |   Exposed   Unexposed   |      Total
----------------+-------------------------+-----------
         Exposed |        30          29   |         59
       Unexposed |        44          86   |        130
-----------------+-------------------------+-----------
           Total |        74         115   |        189

McNemar's chi2(1) =      3.08    Prob > chi2 = 0.0792
Exact McNemar significance probability       = 0.1006
```

O teste de McNemar não detectou uma diferença significativa na proporção de fumantes com o baixo peso ao nascer(p=0.1006).

Criando Gráficos

Utilizar o banco de dados: **lbw.dta**(estudo com crianças de baixo peso).

Gerar/criar variáveis indicadoras para *low*(0=não; 1=sim), *race*(1=branca; 2=negra; 3=outras) e *smoke*(0=não, 1=sim) com os comandos:

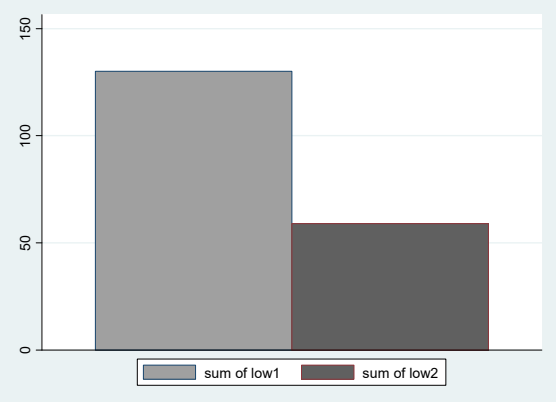
tab low, gen(low) (cria as variáveis: low1 e low2)
tab race, gen(race) (cria as variáveis: race1 , race2 e race3)
tab smoke , gen(smoke) (cria as variáveis: smoke1 e smoke2)

Inserir *labels*:

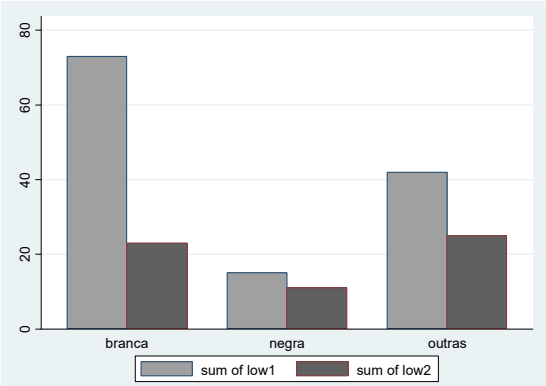
```
label define crace 1 branca 2 negra 3 outras
label val race crace
label define clow 0 "peso normal" 1 "baixo peso"
label val low clow
label define csmoke 0 não 1 sim
label val smoke csmoke
tab1 low race smoke
```

1 Gráfico de barras.

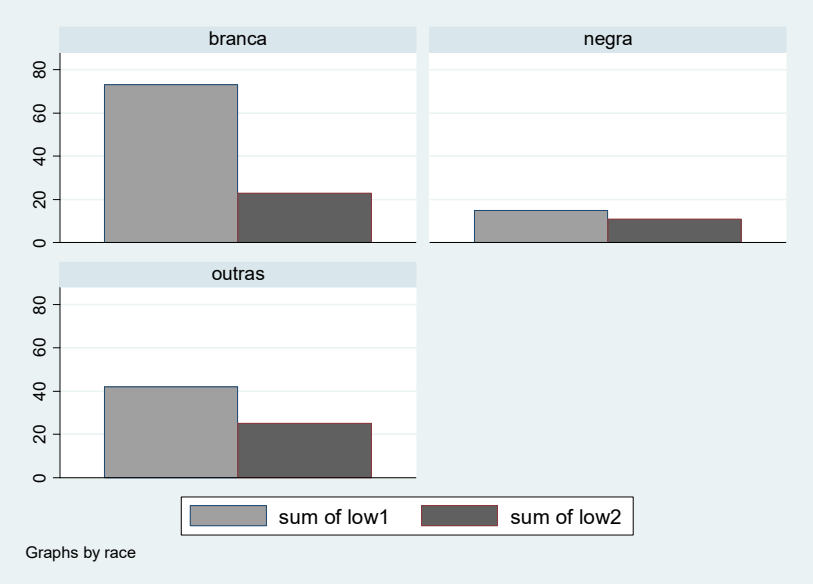
graph bar (sum) low1 low2



graph bar (sum) low1 low2, over(race)

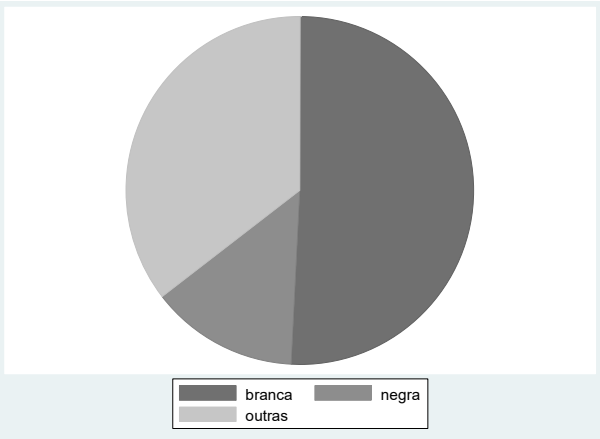


graph bar (sum) low1 low2, by(race)

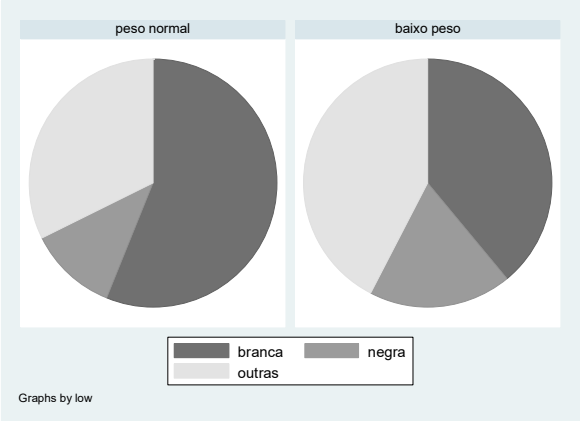


2 Gráfico de setores.

graph pie, over(race)

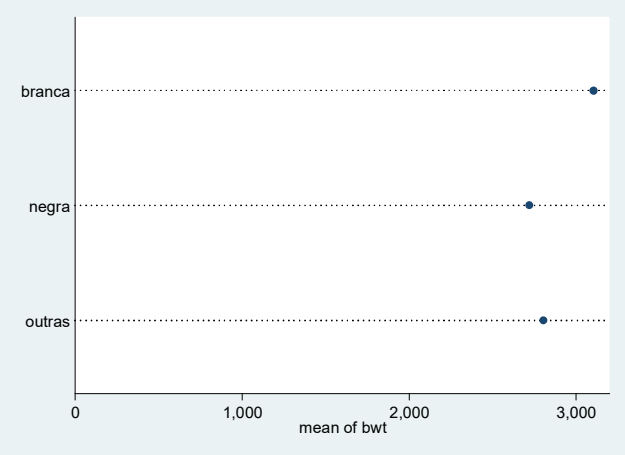


graph pie, over(race) by(low)



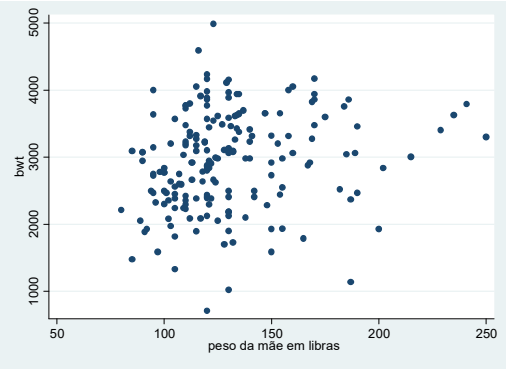
3 Gráfico de pontos.

graph dot (mean) bwt , over(race)

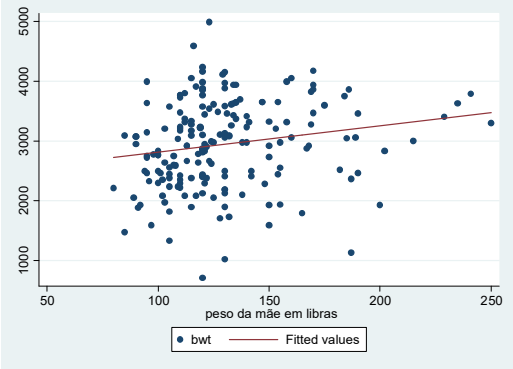


4 Gráfico de dispersão.

twoway (scatter bwt lwt)

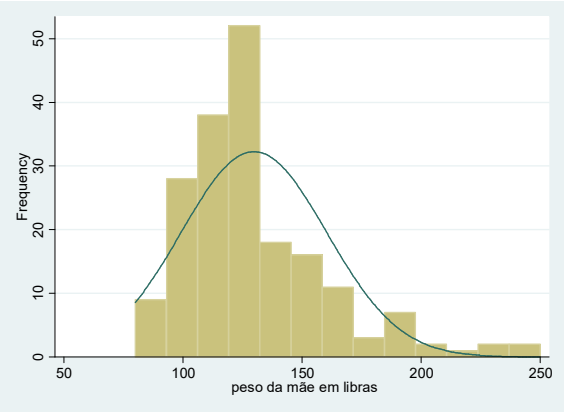


twoway (scatter bwt lwt) (lfit bwt lwt)

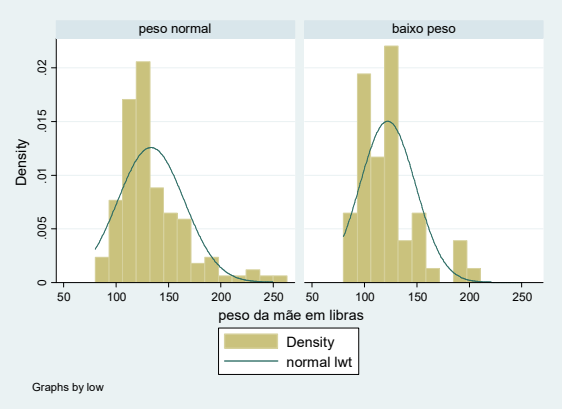


5 Gráfico Histograma.

histogram lwt, normal

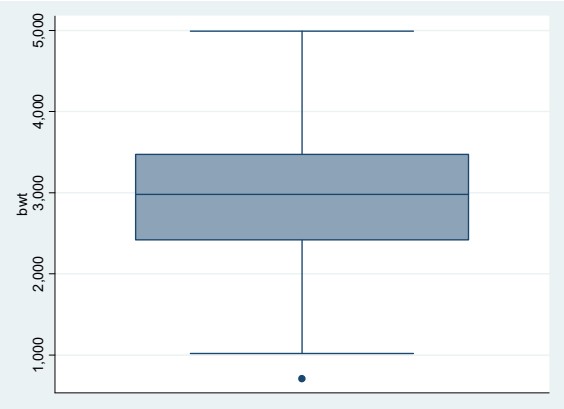


histogram lwt, normal by(low)

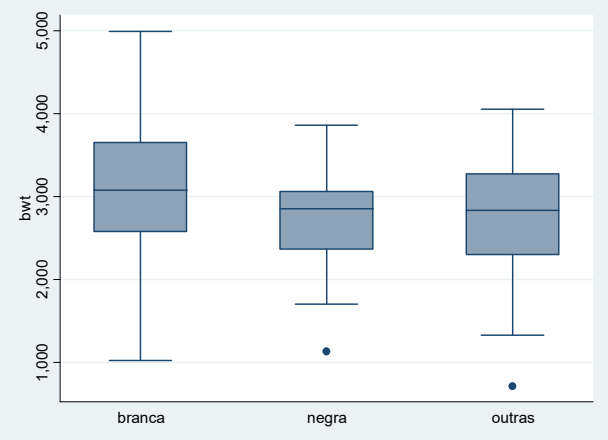


6 Gráfico Box-plot.

graph box bwt



graph box bwt, over(race)

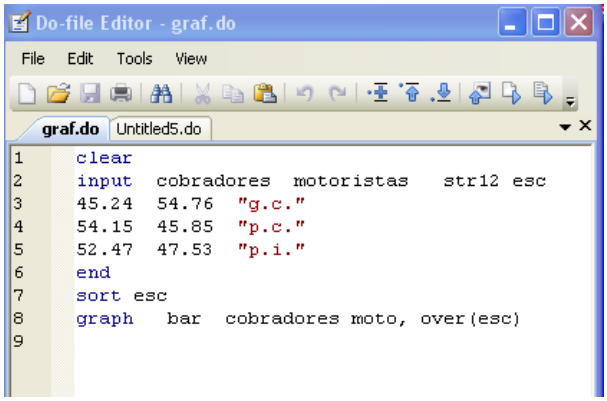


Criando Arquivos-do

Criando um arquivo **"do"** para obter gráfico de barras. Clique no botão: *New Do-file Editor*. Digite as linhas abaixo e salve com o nome **graf.do**.

```
clear
input  cobradores  motoristas   str12 esc
45.24  54.76  "g.c."
54.15  45.85  "p.c."
52.47  47.53  "p.i."
end
sort esc
graph  bar  cobradores moto, over(esc)
```

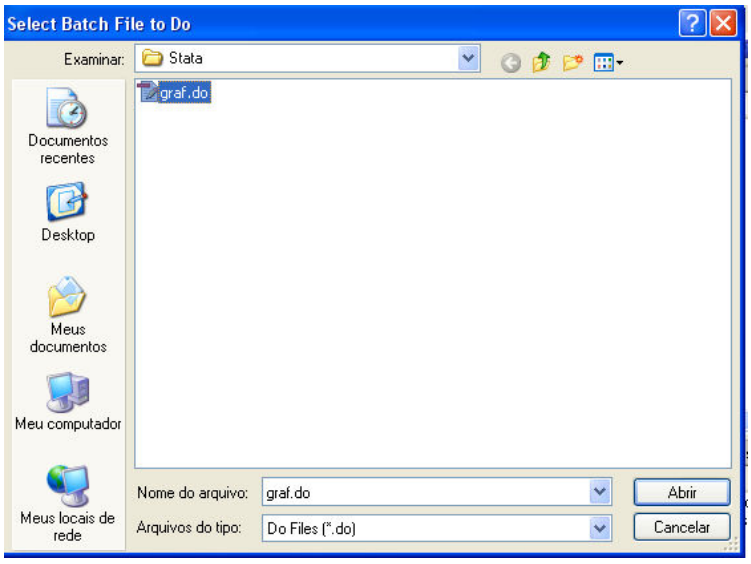
Tela do Do-file Editor



Após salvar o arquivo fechar a janela do editor *Do-file*.

Para correr(*run*) um arquivo-do:

clique no no menu: **File** → ***do*** escolha o arquivo ***graf.do*** e clique no botão ***Abrir.***



Bibliografia

Altman, D. E. (1991). **Practical Statistics for Medical Research**. London, Chapman & Hall.

Berquó, E. S.; Souza, J. M. P.; Gotlieb, S. L. D. (1981). **Bioestatística**. São Paulo, Editora Pedagógica Universitária.

Kirkwood, B. R. (1988). **Essentials of Medical Statistics**. Oxford, Blackwell Science Publications.

Callegari-Jacques, Sidia M. (2003). Bioestatística: princípios e aplicações – Porto Alegre: Artmed.

Souza, M.F.M. (1996). **Um estudo sobre o risco de distúrbios psiquiátricos menores entre motoristas e cobradores do sistema de ônibus urbano na cidade de São Paulo**. Tese de mestrado. Faculdade de Medicina - USP.

Jekel, James F. **Epidemiologia, Bioestatística e medicina preventiva** / James F. Jekel, David L. Katz e Joan G. Elmore; trad. Jair Ferreira – 2.ed. – Porto Alegre : Artmed, 2005.