STATIST 1.4.1
Manual do Usuário

Jakson Alves de Aquino

jalvesaq@gmail.com

5 de setembro de 2006

Sumário

1 Introdução **1**

2 Avisos aos usuários do Windows **1**

3 Instalação **2**

4 Abrindo o programa **2**

5 Menu **3**

6 Statist e Gnuplot **3**
6.1 Diagrama em caixa 4
6.2 UTF-8 4

7 Dados **5**
7.1 O formato do arquivo 5
7.2 Nomes de colunas e rótulos de variáveis 6
7.3 Valores omissos 7
7.4 Lendo e salvando arquivos 8

8 Manipulação de bases de dados **8**
8.1 Extraindo colunas de um banco de dados com largura fixa 8
8.2 Extraindo uma amostra de uma base de dados 9
8.3 Recodificando uma base dados 9
8.4 Selecionando casos e computando novas variáveis 11
8.5 Ordenando a base de dados 12
8.6 Mesclando arquivos de dados 12

9 Modo não interativo **13**

10 Dicas úteis **14**

1 Introdução

`Statist` é um programa leve e fácil de usar. Todas as funções estão em um menu interativo: você tem apenas que escolher o que estiver precisando. `Statist` é um Software Livre sob a licença GNU GPL e não é acompanhado de absolutamente nenhuma garantia.

Este documento é uma tradução incompleta e não literal do original escrito por Dirk Melcher, mas novo material foi adicionado ao texto. Sou grato a Bernhard Reiter pela revisão do texto e sugestões de melhorias.

2 Avisos aos usuários do Windows

Usuários do GNU/Linux estão muito mais acostumados com programas em modo texto. Uma característica útil nesse sistema é a *completação* de linhas de comando: após digitar as primeiras letras de um arquivo ou diretório de nomes longos, podemos pressionar a tecla "Tab" para que o sistema complete o resto para nós. Os emuladores de terminal onde digitamos os comandos podem salvar e rolar de volta muitas linhas já impressas na tela. E o mais importante, GNU/Linux é um software livre que permite a qualquer pessoa inspecionar o que o computador faz. E muitos podem consertar problemas, o que o torna mais seguro. Logo que puder, experimente o `statist` em um sistema operacional livre, como o Linux ou o FreeBSD.

Para criar gráficos com o `statist` você precisará de uma versão do `gnuplot` que venha com o `pgnuplot`. No Windows, você não pode enviar comandos para o `gnuplot` através do `statist`. É necessário digitar os comandos na janela do `gnuplot`. Mas atenção, não feche a janela do `gnuplot`. Você pode fechar somente o gráfico. Se fechar a janela do `gnuplot` será necessário reiniciar o `statist` para ser possível criar gráficos novamente.

Alguns dos softwares utilizados para manipular arquivos de dados não fazem parte do `statist`, mas eles também estão disponíveis para o Windows. Por favor, procure na Internet por gnucoreutils, que é um dos pacotes do GnuWin32.

Na pasta `C:\Arquivos de Programas\statist` pode ser encontrada a documentação do `statist`, bem como um exemplo de arquivo de configuração. Você pode renomear este arquivo para `statistrc.txt` e editá-lo conforme suas preferências.

Infelizmente, `statist` não pode produzir *output* colorido no DOS.

3 Instalação

1. Abra um terminal.

2. Descompacte o código fonte, compile o programa e, como superusuário, instale-o. Isto é, digite:

```
tar -xvzf statist-1.4.1.tar.gz
cd statist-1.4.1
make
# opcional, se você tem "check" instalado
make check

# instalação como root para todos os usuários
su -
cd diretório-para/statist-1.4.1
make install
exit
```

Essa é a instalação padrão, que deve funcionar na maioria das distribuições GNU/Linux. Se as instruções acima não são suficientes para o seu caso, por favor veja o arquivo README.pt. Ele traz detalhes sobre como instalar o `statist` a partir do código fonte.

4 Abrindo o programa

Você pode simplesmente digitar:

```
statist arquivo_de_dados
```

Entretanto, existem algumas opções que podem ser úteis, e, então, para abrir o `statist`, será preciso digitar:

```
statist [ options ] data_file [ options ]
```

A única opção que você precisa memorisar é `--help`, ou simplesmente `-h`, que exibirá a lista das opções disponíveis.

Você também pode criar e editar o arquivo `~/.statistrc`, para configurar algumas opções. Se você tem o privilégio de se poder se tornar superusuário, poderá criar e editar o arquivo `/etc/statistrc`. Opções passadas pela linha de comando têm prioridade sobre as lidas do arquivo `statistrc`. Você pode encontrar um exemplo de `statistrc` no diretório onde está a documentação do programa (usualmente `/usr/share/doc/statist`). Finalmente, você poderá modificar algumas opções durante a execução do `statist` escolhendo no menu principal o item *Preferências*.

5 Menu

O programa possui um menu interativo simples que o torna muito fácil de usar. Não há necessidade de lembrar comandos. Digitando o número '0' você é levado ao menu de nível superior, ou termina o programa se já estiver no *Menu principal*. Uma dica é importante: se escolher um item do menu por engano, você pode cancelar o processo pressionando a tecla <Enter> antes de digitar qualquer valor ou responder a qualquer questão. Fazendo isso, o último menu é impresso novamente.

Se você escolher um procedimento estatístico do menu, lhe serão pedidos os nomes das variáveis. Freqüentemente, não é necessário digitar o nome da coluna completo quando se está entrando os nomes das variáveis para análises. Por exemplo, se você tem uma coluna cujo nome é:

esse_eh_um_nome_realmente_muito_grande

e se não há nenhuma outra coluna começando com a letra 'e', você pode digitar simplesmente 'e' para o nome da coluna. Finalmente, se você quiser selecionar todas as colunas, basta digitar "todas" como nome da primeira coluna.

Na verdade, o processo todo é auto-explicativo, e você seria capaz de usar o programa mesmo sem ler essa breves explicações.

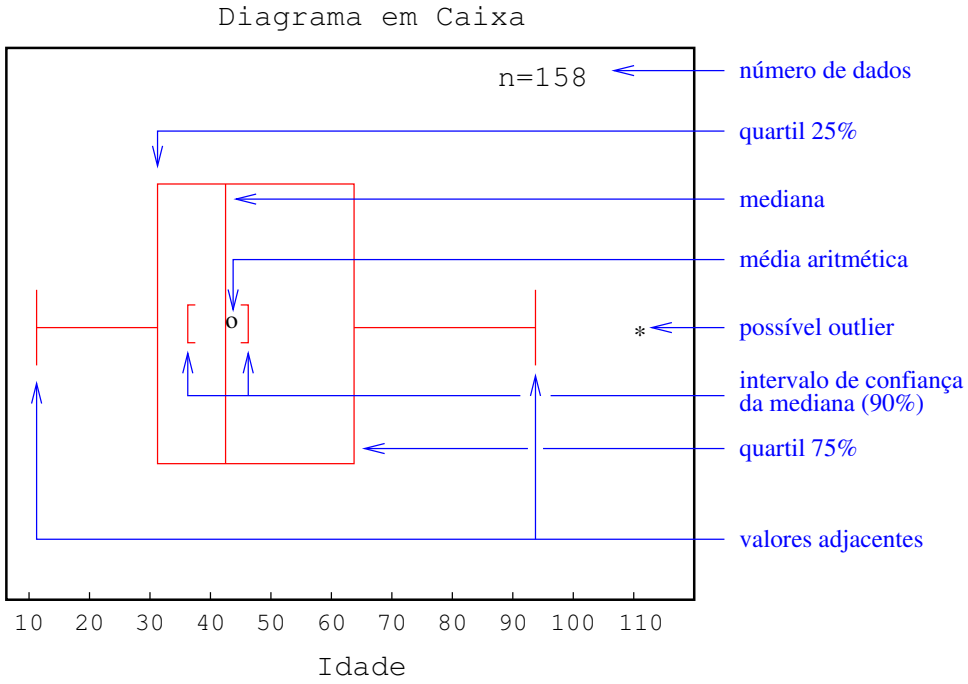
6 Statist e Gnuplot

Gnuplot é um programa interativo que faz apresentações gráfica de dados e funções, e o `statist` cria gráficos do `gnuplot` para algumas funções. Normalmente, você não tem que abrir o `gnuplot` manualmente. O pré-requisito para produzir os gráficos é simplesmente que o programa esteja instalado em um dos diretórios em que o sistema operacional costuma procurar pelos arquivos binários (*PATH*).

Você poderá refinar ou personalizar seus gráficos se conhecer a sintaxe do `gnuplot`. Para isso, escolha a opção do menu *Diversos* | *Digitar comandos do gnuplot*. Você pode mudar muitas coisas, como cores e tipos de linhas, rótulos dos eixos, etc... Mesmo para os que não sabem a sintaxe do `gnuplot`, é fácil mudar pelo menos os títulos dos gráficos e dos eixos, porque uma lista dos últimos comandos enviados para o `gnuplot` são impressos na tela. As mudanças serão aplicadas ao gráfico que estiver sendo exibido com o comando do `gnuplot` "replot".

Os gráficos do `gnuplot` podem ser desabilitados se o `statist` for chamado com a opção `--noplot`. Isso pode ser útil se você for, por exemplo, trabalhar no modo não interativo ou se sua base de dados for muito grande, o que tornaria muito lenta a geração de gráficos pelo `gnuplot`.

6.1 Diagrama em caixa



Você provavelmente não terá problemas para interpretar os gráficos do `statist`. O único que pode merecer alguma explicação é o *Diagrama em caixa*. A figura acima mostra o significado de cada parte desse gráfico.

6.2 UTF-8

Os gráficos feitos pelo `statist` através do gnuplot podem apresentar caracteres estranhos se o seu ambiente está configurado para UTF-8 e a sua língua possui caracteres que não pertencem ao alfabeto inglês. O problema é que o gnuplot normalmente irá interpretar títulos e rótulos como se estivessem codificados com caracteres de um único byte, como o ISO-8859-1 (Latin 1, usado pelo português), mesmo que o mapa de caracteres do seu emulador de terminal esteja configurado para UTF-8. É possível misturar letras de diferentes conjuntos de caracteres em um único gráfico (grego e Latin 1, por exemplo), mas para isso é preciso configurar o gnuplot corretamente. Por favor, acesse o endereço abaixo para saber maiores detalhes:

http://statist.wald.intevation.org/utf8.html

7 Dados

7.1 O formato do arquivo

O `statist` lê dados de arquivos de texto simples (ASCII). Se o programa não for chamado com um o nome de um arquivo ASCII como parâmetro, ele irá imediatamente perguntar pelo

nome de um arquivo de dados. Sem um arquivo de dados, não há nada a fazer, a não ser que você declare a opção `--nofile` ao chamar o programa, com o objetivo de digitar manualmente os dados (escolha o item do menu: *Administração dos dados* | *Ler coluna do terminal*). Entretanto, raramente será razoável fazer isso. Seria mais confortável usar um editor de textos ou uma planilha eletrônica, como o *OpenOffice Calc* ou o *Gnumeric*. Nesse caso, salve o arquivo como .csv.

Mas, atenção, o `statist` sempre usa um ponto como separador de decimais quando está trabalhando com arquivos de dados. Como no português o separador de decimais é a vírgula, antes de digitar seus dados, você pode tentar abrir a planilha eletrônica em um terminal com a região configurada para "C", como abaixo:

```
export LC_ALL=C
oocalc &
```

Se você realmente precisar usar um arquivo em que os separadores de decimais são vírgulas, o `statist` converterá em pontos as vírgulas usadas como separadores de decimais em números que estejam entre aspas duplas. Se não for este o caso, será necessário definir manualmente o delimitador de decimais. O `statist` talvez lhe faça perguntas sobre o formato do arquivo. Caso contrário, escolha o item do menu *Administração dos dados* | *Configurar formato dos arquivos*. Alternativamente, inicie o `statist` como no exemplo:

```
statist datafile.csv --dec ","
```

Um arquivo de dados do `statist` consiste de uma ou mais colunas de dados. As colunas devem ser separadas umas das outras por espaços em branco, caracteres de tabulação, vírgulas ou ponto-e-vírgulas. Estes caracteres são ignorados e, portanto, é possível ter qualquer quantidade deles entre dois campos. Por exemplo, o `statist` irá ler os mesmos dados dos dois arquivos abaixo:

```
#Example data-file for statist
  1  3  5  6
  7  8  9 10
 11 12 13 14
```

```
#Example data-file for statist
 1,3,"5",6
 ,7 8 ;, 9 10
 11;12;13;14;;
```

Como você pode perceber pelo exemplo acima, comentários começam com o símbolo '#' e são ignorados pelo programa. Linhas em branco também são ignoradas.

7.2 Nomes de colunas e rótulos de variáveis

Quando o `statist` lê um arquivo de dados, cada coluna é considerada uma variável. A primeira coluna receberá o nome 'a', a segunda será 'b', etc. Entretanto, se seu banco de dados tiver muitas variáveis você poderá preferir que as variáveis tenham nomes mais significativos. A primeira linha do banco de dados que não seja um comentário pode conter os nomes das colunas. O `statist` tentará detectar os nomes usando um algoritmo muito simples. Ele verifica se todos

campos da primeira linha que não seja comentário se iniciam com uma letra do alfabeto inglês. Se qualquer um dos campos se iniciar com um caractere que não esteja entre 'a' e 'z' ou entre 'A' e 'Z', o `statist` considerará que o arquivo de dados não contém um cabeçalho. Se ele errar na detecção do cabeçalho você poderá dizer ao `statist` como proceder, escolhendo o item do menu *Administração dos dado* | *Configurar formato dos arquivos*. Outra solução é o uso da opções de linha de comando: `--header` ou `--noheader`.

Você pode, ainda, explicitamente incluir no arquivo de dados a informação de que o cabeçalho está presente, escrevendo no início da linha com os nomes das colunas os caracteres "#%". Nesta última alternativa, como nas linhas de comentário, o primeiro caractere deve ser '#', mas esse símbolo deve ser seguido por '%'. Com sua configuração padrão, o `statist` irá ler os dois exemplos abaixo de arquivos de dados com o simples comando "`statist arquivo`":

```
#%kow kaw ec50
0.34 4.56 0.23
1.23 5.45 6.76
6.78 1.34 9.60
```

```
kow kaw ec50
0.34 4.56 0.23
1.23 5.45 6.76
6.78 1.34 9.60
```

O número de nomes para as variáveis deve ser exatamente o mesmo do número de colunas. Somente letras, números e '_' podem ser usados nos nomes, e letras acentuadas podem causar problemas. Se você usar a opção `--labels arq_rótulos`, o `statist` irá usar os títulos das colunas e os rótulos das variáveis presentes no `arq_rótulos`. Quando executando alguns gráficos e análises, os nomes das colunas e os valores das variáveis serão substituídos por seus rótulos. Um `arq_rótulos` é simplesmente uma lista com os nomes das colunas mais seus rótulos seguida da lista dos seus valores com os rótulos. As informações para diferentes colunas são separadas por uma linha em branco, como no exemplo:

```
stat Você gosta de estatística?
0 Não
1 Sim
2 Sem resposta

cor Qual sua cor favorita?
0 Vermelho
1 Verde
2 Azul
3 Outra
```

No exemplo acima, o arquivo de dados tem uma coluna nomeada "stat" e outra, "cor". Os valores da variável "stat" são sempre "0", "1", ou "2". Você pode usar usar o mesmo arquivo com rótulos para diferentes arquivos de dados. Não há problema se algumas colunas permanecerem sem rótulos, ou se alguns rótulos não acharem sua coluna na base de dados. Portanto, se você tiver uma base de dados com centenas de colunas e quiser trabalhar com vários subconjuntos dos dados que compartilham algumas colunas, você poderá preparar um único arquivo de rótulos.

Se você escolher no menu a opção *Ler outro arquivo*, os rótulos até então não usados serão automaticamente aplicados às colunas adicionadas. Nota: rótulos muito extensos demandarão muito espaço, e a tabela *Comparação de médias* poderá não ficar bem ajustada à tela. Se você tiver rótulos grandes, poderá executar uma *Comparação de médias* com apenas um pequeno número de colunas simultaneamente.

7.3 Valores omissos

O `statist` pode lidar com arquivos de dados contendo valores omissos (*missing values*), e há dois modos de indicar que um valor está ausente. O primeiro é usar uma *string* específica no lugar do valor ausente. Por padrão, o `statist` interpreta a *string* "M" como indicadora de valor omisso, mas você pode escolher uma *string* diferente no `statistrc`, com o argumento de linha de comando `--na-string <string>`, ou no item do menu *Administração dos dado* | *Configurar formato dos arquivos*.

Devido ao fato do `statist` interpretar qualquer quantidade de caracteres ignoráveis (" `",;\t`") como um único delimitador de campo, dois separadores de campo adjacentes não serão interpretados pelo `statist` como valor omisso. Pelo contrário, ele reportará que a linha possui menos colunas do que deveria. Esse é o comportamento padrão, mas ele pode ser modificado no `statistrc`, com o argumento de linha de comando `--sep <caractere>` ou, mais uma vez, escolhendo o item do menu *Administração dos dado* | *Configurar formato dos arquivos*. Com essa opção, somente um caractere específico será interpretado como separador de campos. Logo, os seguintes arquivos de dados serão considerados iguais, mas o segundo precisa da opção `--sep ","`:[1]

```
  1   3   5   6                    1,3,5,6
  7   M   9 10                    7,,9,10
 11  12   M 14                    11,12,,14
```

Cada coluna da base de dados é salva em um arquivo temporário binário, onde todos os valores são guardados como números reais (tipo *double*, na linguagem C). Esses arquivos são apagados quando você sai do `statist`. Os valores omissos são guardados como o menor número *double* possível, ou seja, $-1,7976931348623157 \times 10^{308}$. Você precisa ter certeza de que este número não está presente em seu banco de dados como um número válido, pois ele seria interpretado como valor omisso.

Antes de cada análise, o `statist` lê a coluna selecionada do arquivo temporário e armazena os valores na memória (ram). Dependendo da análise a ser feita, o programa poderá apagar somente os valores omissos ou apagar uma linha inteira quando uma das colunas tem um valor omisso naquela posição. Os dados, entretanto, somente são apagado na cópia dos arquivos temporários que estão na memória. Os arquivos temporários permanecem intactos até você fechar o programa. Por exemplo, se você escolher *Regressões e correlações* | *Correlação linear múltipla* serão apagadas todas as colunas que têm valores omissos em qualquer uma das colunas. Esta

[1] Mesmo com a opção `--sep`, o algoritmo padrão é utilizado para interpretar a linha com os nomes das colunas. Logo, não é permitido haver nomes vazios para colunas.

análise pode ser usada se cada linha do seu banco de dados representa um único caso, o que é muito comum em ciências sociais. A opção *Testes* | *Teste-t para comparação das médias de duas amostras* irá apagar todos os valores omissos, mas a existência de um valor omisso em uma coluna não irá fazer a linha inteira ser apagada. Você poderia fazer essa análise se, por exemplo, as colunas de seu banco de dados representam séries diferentes de experimentos semelhantes e você gostaria de comparar dois conjuntos de resultados.

7.4 Lendo e salvando arquivos

Se você quiser trabalhar com somente um sub-conjunto do seu banco de dados, você pode salvar colunas em um arquivo de texto, escolhendo a opção *Administração dos dados* | *Exportar colunas para arquivo de texto*. Você pode também ler os dados de vários arquivos simultaneamente (*Administração dos dados* | *Ler outro arquivo*). Quando você *Ler outro arquivo*, novas colunas são adicionadas à base de dados e, se o nome de uma coluna no novo arquivo já estiver em uso no banco de dados atual, o símbolo "_" será afixado ao novo rótulo.

Outra possibilidade é juntar colunas (*Manipulação dos dados* | *Juntar colunas*). Nesse caso, as colunas selecionadas serão concatenadas em uma maior.

8 Manipulação de bases de dados

8.1 Extraindo colunas de um banco de dados com largura fixa

Para extrair colunas de um banco de dados de largura fixa e salvá-las em um arquivo de dados do `statist`, digite:

```
statist --xcols arq_config basedados_orig nova_base
```

O conteúdo do `arq_config` é simplesmente uma lista de nomes de variáveis e suas respectivas posições no banco de dados de largura fixa, como no exemplo abaixo:

```
nasc 1-4
sexo 8
renda 11-15
```

Com o arq_config acima, o `statist` iria ler a seguinte base de dados:

```
1971 522   2365
19609991  32658
19455632
19674131  32684
```

E produzir:

```
#%nasc sexo    renda
1971    2       2365
1960    1       32658
1945    2       M
1967    1       32684
```

O `statist` não adicionará os caracteres "`#%`" à primeira linha do arquivo se ele tiver sido iniciado com a opção de linha de comando `--header` ou se o aquivo `statistrc` estiver com a opção "`autodetect_header = yes`". A *string* usada para valores omissos também pode ser modificada pela linha de comando e pelo *statistrc*. As colunas são separadas por um espaço em branco, a não ser que você tenha escolhido algo diferente com a opção `--sep`. Valores não numéricos são extraídos e incluídos na `nova_base`, embora o `statist` não seja capaz de lê-los. Você teria que substituí-los por códigos numéricos.

8.2 Extraindo uma amostra de uma base de dados

Se você for trabalhar com uma base de dados muito grande que ainda não conhece muito bem, poderá ser conveniente iniciar a exploração dos dados usando uma amostra da base de dados. As análises, é claro, serão mais rápidas do que se fossem feitas na base inteira. Após descobrir quais análises são mais relevantes para sua pesquisa você poderá refazer essas análises na base de dados original.

Para extrair uma porcentagem das linhas de uma base de dados, abra o `statist` com os seguintes parâmetros:

```
statist --xsample porcentagem basedados arq_destino
```

onde "porcentagem" deve ser um número inteiro entre 1 e 99. A nova base de dados, `arq_destino` será criada com *aproximadamente* a porcentagem solicitada de linhas extraídas de `basedados`.

8.3 Recodificando uma base dados

Para algumas tarefas de manipulação de dados, precisaremos utilizar alguns programas que não fazem parte do pacote do `statist`, mas que estão disponíveis no GNU/Linux (e podem ser instalados no DOS/Windows). Para pequenos arquivos de dados, com poucas variáveis, você pode usar seu editor de texto ou planilha de cálculos preferidos. Entretanto, se seu arquivo for muito grande, ou tiver muitas variáveis, poderá ser mais conveniente usar as ferramentas descritas aqui e nas seções seguintes.

Algumas vezes, precisamos recodificar alguns valores em uma base de dados. Suponha, por exemplo, que num certo arquivo de dados o valor "999" significa valor omisso para a variável idade, e que em algumas análises nós queremos usar "faixa etária" e, em outras, "idade". Como ainda precisaremos da variável "idade", precisamos recodificar "idade" em uma variável diferente. Para criar nossa nova base de dados teríamos que digitar:

```
awk '{if(/idade/) {print $0 "\t" "faixa"}
        else {
          if(NF == 0) {print $0}
            else {
              if ($2 <= 20){id1 = 1} else
                if ($2 > 20 && $2 <= 50){id1 = 2} else
                if ($2 > 51 && $2 < 999){id1 = 3} else
                {id1 = "M"}
              {print $0 "\t" id1}
            }
       }
     }' datafile.csv > newfile.csv
```

A expressão entre aspas simples são comandos do `awk`. Com esse comando, o `awk` leria o seguinte arquivo:

```
sexo idade
2      23
1      88
2      10
2      36
3      999
1      55
```

E produziria:

```
sexo  idade   faixa
0       23      2
1       88      3
0       10      1
0       36      2
M       999     M
1       55      3
```

À primeira vista, o comando do `awk` pode parecer complexo, mas posso explicá-lo:

`$`: O símbolo '`$`' significa "campo", isto é, uma coluna de um banco de dados do `statist`.

`$0`: tem um significado especial: a *linha inteira*.

`if(/#/) {print $0 "\t" "faixa"}`: Se a linha tiver o símbolo '#', imprima a linha inteira mais um caractere de tabulação mais a expressão "faixa". Essa linha contém os nomes das nossas colunas (a não ser que tenham sido inseridos comentários no arquivo de dados).

`if(NF == 0) {print $0}`: Se o número de campos for igual a zero, imprima a linha inteira.

`if ($2 > 20 && $2 <= 50){id1 = 2}`: Se o segundo campo tiver um valor maior do que 20 e menor ou igual a 50, o valor da variável "id1" é 2.

`print $0 "\t" id1`: Imprima a linha inteira mais um caractere de tabulação mais o valor da variável `id1`.

Nós também usamos `awk` para selecionar casos e para computar novas variáveis. Por isso, por favor consulte seu manual ou páginas de info para maiores detalhes sobre como usá-lo (em um terminal, digite `info awk`). Freqüentemente, os comandos do `awk` que usamos começam testando se a linha contém os nomes das colunas e se a linha está vazia.

8.4 Selecionando casos e computando novas variáveis

Podemos usar `awk` para realizar duas outras tarefas: (1) criar uma nova base de dados, selecionando somente alguns casos de um arquivo de dados existente, e (2) computar uma nova variável, usando os valores de variáveis existentes. Aqui mostraremos apenas dois exemplos de como usar o `awk`.

Suponha que a segunda coluna de um arquivo de dados contém a variável "sexo", codificada como '0' para homens e '1' para mulheres, e que nós queremos incluir somente as mulheres em algumas análises Digitando o seguinte comando em um terminal o novo arquivo de dados que precisamos seria criado:

```
awk '{if(/sexo/ || /#/ || $2 > 0) {print $0}
}' arq_dados.csv > novo_arq_dados.csv
```

Estamos dizendo para o `awk` que se ele encontrar a palavra "sexo" ou o símbolo '#' em uma linha (por que ela certamente contém os rótulos das nossas colunas ou um comentário) ou se o segundo campo da linha tiver um número maior do que 0 ele deverá produzir como saída a linha inteira (todos os campos = $0, e "||" significa "ou"). Finalmente, dizemos ao programa que controla a linha de comando que queremos que as linhas impressas sejam redirecionadas da tela para o arquivo novo_arq_dados.csv.

Agora, suponha que você queira calcular um índice usando três variáveis de sua base de dados, e que o índice será a soma das colunas 1 e 2 dividida pelo valor da terceira coluna:

```
awk '{if(/#/ || /var1/) {print $0 "\tind"} else
{{ind = ($1 + $2) / $3}
{print $0 "\t" ind}}}' arqdados.csv > novoarq.csv
```

A seqüência de caracteres "\t" será convertida em um caractere de tabulação horizontal.

Atenção: O `statist` sempre usa ponto como separador de decimais quando está trabalhando com arquivos de dados. Entretanto, o separador de decimais em português é a vírgula, que será usada pelo `awk` para produzir seus resultados. Para evitar esta incompatibilidade, digite o seguinte comando no terminal antes de usar o `awk`:

```
export LC_ALL=C
```

Com o comando acima, a linguagem, os números, etc serão configurados para inglês. Note que programas iniciados no terminal também funcionarão em inglês. Para que o terminal volte a funcionar em português, você tem que digitar "export LC_ALL=pt_BR" ou "export LC_ALL= pt_PT", dependendo do seu país (ou fechar o terminal e abrir outro).

8.5 Ordenando a base de dados

Você pode usar alguns programas disponíveis no Linux (mas que também podem ser instalados no Windows) se quiser ordenar as linhas de uma base de dados inteira, tomando uma ou mais colunas como chaves. Suponha, por exemplo que você queira ordenar a base de dados usando a 12^a coluna como chave. Os seguintes comandos fariam o serviço:

```
head -n 1 arqdados.csv > nomesdascolunas
sort -g -k 12,12 arqdados.csv > ordenado
cat nomesdascolunas ordenado > arqdados_ordenado.csv
```

Com os comandos acima nós ordenamos o banco de dados em três etapas: (1) Criamos o arquivo `nomesdascolunas` contendo a primeira linha de `arqdados.csv`. (2) Criamos o arquivo `ordenado`, uma versão ordenada do banco de dados. Entretanto, o 12^o nome de coluna foi tratado como se fosse um número e a sua linha foi ordenada. Se ela não for mais a primeira linha do arquivo, as colunas do nosso banco de dados estarão sem rótulos. Se for este o caso, deveremos usar o terceiro comando. (3) Concatenamos os arquivos `nomesdascolunas` e `ordenado`, criando o arquivo `arqdados_ordenado.csv`. Por favor, leia os manuais de `head`, `sort` e `cat` para maiores detalhes sobre como usá-los.

8.6 Mesclando arquivos de dados

Para mesclar arquivos de dados usando uma variável como chave, usamos um outro programa externo: `join`. Suponha que você tem um arquivo de dados contendo informações sobre pessoas e que algumas das pessoas, de fato, são casadas umas com as outras. Você quer saber qual a média das diferenças de idade entre os maridos e suas esposas. Não é possível fazer análises comparando pessoas em linhas diferentes, somente variáveis em colunas diferentes. Entretanto, seu banco de dados tem uma variável que pode ser usada como chave: *domicílio*. Pessoas que têm o mesmo valor para a variável "domicílio", e que são casadas, são casadas uma com a outra. Você poderia seguir dois passos para atingir seu objetivo: (1) Usar `awk` para criar dois arquivos de dados diferentes, um com os homens casados e outro com as mulheres casadas. (2) Usar `join` para mesclar os dois arquivos em um novo. Se a variável domicílio estiver na primeira coluna dos bancos de dados, bastará digitar:

```
join -e "" mulheres.csv homens.csv > casais.csv
```

O comando acima iria ler os dois arquivos seguintes:

```
domic renda  idade              domic renda  idade
123     4215   23               123     3256   27
124     3251   35               125     4126   25
126     0      20               126     4261   22
127     1241   45               128     3426   60
```

E produzir:

```
domic renda idade renda idade
123 4215 23 3256 27
126 0 20 4261 22
```

Não há problema em haver uma dupla ocorrência de “renda” e “idade” porque, ao abrir o banco de dados, o `statist` irá acrescentar ‘_’ à segunda ocorrência. Se você tiver que mesclar dois arquivos usando mais de uma coluna como chave, você pode usar o `awk` para criar uma coluna chave única que concatene os caracteres de todas as colunas chave. Por exemplo, se as suas variáveis chaves estão nas colunas 2 e 3:

```
awk '{if(/renda/) {print "key" "\t" $0} else {
       if(NF == 0) {print $0} else {
          {print $2$3 "\t" $0}
        }
       }
      }' pessoas.csv > pessoas_com_chave.csv
```

9 Modo não interativo

Se você tem que repetir muitas vezes a mesma análise, você certamente não achará divertido ter que iniciar o `statist` e, pela milésima vez, escolher a mesma seqüência de comandos do menu. Se for este o seu caso, será possível usar o `statist` no modo não interativo. Abra o `statist` com a opção `--silent`, e forneça a ele como parâmetro o arquivo contendo o que você teria que digitar se o `statist` estivesse sendo executado no modo normal. A única diferença é que no modo `silent` o `statist` não imprime a mensagem “Pressione <ENTER> para continuar”, e, portanto, você não deve incluir esse <ENTER> no script. Por exemplo, se você quiser fazer uma correlação entre as variáveis “a” e “b” em um arquivo de dados chamado, digamos, `dia365.csv`, você poderia criar um arquivo com nome de `automatico` com o seguinte conteúdo:

```
2
1
a
b
0
0
```

O próximo passo seria invocar o `statist` com o seguinte comando:

```
statist --silent --noplot dia365.csv < automatico
```

O resultado seria impresso na tela. Entretanto, se você preferir que os resultados sejam salvos em um arquivo chamado relato365 digite:

```
statist --silent --noplot dia365.csv < automatico > relato365
```

10 Dicas úteis

- Por favor, comunique qualquer problema que encontrar neste documento ou no `statist` para: statist-list@itevation.de (em inglês) ou para mim (jalvesaq@gmail.com). Escreva também para fazer sugestões de mudanças no programa ou dizendo quais características você gostaria de ver adicionadas ao `statist`.
- Quando você encontrar uma questão do tipo "Fazer tal coisa? (s/N)", a letra "N" maiúscula significa que se você digitar qualquer letra diferente de "s", ou mesmo se simplesmente pressionar <Enter>, será considerado que sua resposta é "Não".
- A última versão do `statist` está disponível na Internet:

http://statist.wald.intevation.org/