

Ângelo José Gonçalves Bós

# Epi Info<sup>®</sup>

## Sem Mistérios

### Um Manual Prático



EPI INFO® SEM MISTÉRIOS:  
UM MANUAL PRÁTICO



Pontifícia Universidade Católica  
do Rio Grande do Sul

**Chanceler**

Dom Dadeus Grings

**Reitor**

Joaquim Clotet

**Vice-Reitor**

Evilázio Teixeira

**Conselho Editorial**

Ana Maria Mello

Armando Luiz Bortolini

Augusto Buchweitz

Beatriz Regina Dorfman

Bettina Steren dos Santos

Carlos Graeff Teixeira

Clarice Beatriz de C. Sohngen

Elaine Turk Faria

Érico João Hammes

Gilberto Keller de Andrade

Helenita Rosa Franco

Jane Rita Caetano da Silveira

Lauro Kopper Filho

Luciano Klöckner

Nédio Antonio Seminotti

Nuncia Maria S. de Constantino

**EDIPUCRS**

Jerônimo Carlos S. Braga – Diretor

Jorge Campos da Costa – Editor-Chefe

**Ângelo José Gonçalves Bós**  
Instituto de Geriatria e Gerontologia  
Programa de pós-graduação em Gerontologia Biomédica  
Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS)

# EPI INFO® SEM MISTÉRIOS: UM MANUAL PRÁTICO



2012

© EDIPUCRS, 2012  
CAPA Luana Santos e Thatiana Pandolfo  
REVISÃO DE TEXTO Caren Capaverde  
EDITORAÇÃO ELETRÔNICA Andressa Rodrigues



EDIPUCRS – Editora Universitária da PUCRS  
Av. Ipiranga, 6681 – Prédio 33  
Caixa Postal 1429 – CEP 90619-900  
Porto Alegre – RS – Brasil  
Fone/fax: (51) 3320 3711  
e-mail: edipucrs@pucrs.br - www.pucrs.br/edipucrs.

#### Dados Internacionais de Catalogação na Publicação (CIP)

---

B741e Bós, Ângelo José Gonçalves  
Epi Info® sem mistérios : um manual prático [recurso eletrônico] / Ângelo José Gonçalves Bós. – Dados eletrônicos – Porto Alegre : EDIPUCRS, 2012.  
211 p.  
Sistema requerido: Adobe Acrobat Reader  
Modo de acesso: <http://www.pucrs.br/edipucrs/>  
ISBN: 978-85-397-0157-5

1. Informática Médica. 2. Epi Info® (Programa de Computador). I. Título.

CDD 610.28

---

Ficha Catalográfica elaborada pelo Setor de Tratamento da Informação da BC-PUCRS.

TODOS OS DIREITOS RESERVADOS. Proibida a reprodução total ou parcial, por qualquer meio ou processo, especialmente por sistemas gráficos, microfílmicos, fotográficos, reprográficos, fonográficos, videográficos. Vedada a memorização e/ou a recuperação total ou parcial, bem como a inclusão de qualquer parte desta obra em qualquer sistema de processamento de dados. Essas proibições aplicam-se também às características gráficas da obra e à sua editoração. A violação dos direitos autorais é punível como crime (art. 184 e parágrafos, do Código Penal), com pena de prisão e multa, conjuntamente com busca e apreensão e indenizações diversas (arts. 101 a 110 da Lei 9.610, de 19.02.1998, Lei dos Direitos Autorais).

## Agradecimentos

Desde a publicação da primeira edição deste livro tenho recebido apoio de muitos pesquisadores, muitos desses verdadeiros amigos virtuais. Não posso deixar de mencionar e agradecer o incansável Jean Ezequiel Limongi, do Centro de Controle de Zoonoses de Uberlândia que, desde novembro de 2008, tem me mandado e-mails com relativa frequência sempre com suas dúvidas intrigantes. A querida Sumire Sakabe, outra pessoa sensacional que muito me estimulou a manter este livro atualizado. Agradeço a todos os alunos tanto dos cursos de Epi Info® no CEARGS quanto das disciplinas de Métodos Quantitativos de Pesquisa e de Demografia e Epidemiologia Gerontológica do Programa de Gerontologia Biomédica da PUCRS, pois foram eles que me incentivaram a reeditar este livro. Particularmente, tenho que agradecer a Mayara Sartori, que revisou grande parte das figuras deste livro e a Denielli da Silva Gonçalves Bós (sobrinha e parceira de muitas empreitadas) por terem me ajudado a atualizar algumas partes do livro.

Ainda sou eternamente grato à Pró-Reitoria de Extensão (PROEX) da PUCRS, ao Centro de Estudos da AIDS do RGS (CEARGS) e à faculdade de Saúde Pública das Ilhas Fiji que acreditaram e proporcionaram condições para a realização das aulas que desenvolvi e que culminaram no desenvolvimento da primeira edição deste livro. Em especial, agradeço o apoio excepcional do professor e amigo Emílio A. Jeckel-Neto e do Monsenhor Urbano Zilles.

Pessoas especiais merecem agradecimentos especiais principalmente por terem tido papel importante na formação deste autor. Meus pais Amaury Lobato Bós e Aurea Gonçalves Bós, ambos falecidos, mas de quem recebi muito apoio para seguir sempre em frente na conquista do meu espaço, ganhando, de cada um, habilidades importantes que culminaram na criação deste projeto que hora se cristaliza com a segunda edição. Ao professor Yukio Moriguchi, meu pai da Geriatria e da pesquisa, a quem agradeço por ter me estimulado a seguir a carreira acadêmica e científica. Ao Dr. Newton Luiz Terra ou simplesmente Tio Terra, uma pessoa que sempre valorizou minhas qualidades. Pessoas também especiais que infelizmente ficaram de fora da primeira edição, mas que estão sempre no meu coração: Antônio Miguel Bós, Paulo Consoni, Ayako Kamijo, Evaldo Nascimento, Nair Mônica e os saudosos Antonio Carlos e Valdemarina.

Não posso deixar de agradecer ao CDC por desenvolver esse programa e mantê-lo atualizado corrigindo constantemente os problemas detectados.

Quero finalmente agradecer ao Programa de Gerontologia Biomédica e à Pró-reitoria de pesquisa e Pós-graduação. Também à EDIPUCRS, de quem recebi o apoio à publicação das edições deste livro.

À professora Helen Abbey (*in memoriam*) e ao professor Larry Brant, cujos ensinamentos e métodos de ensino me ajudaram a desenvolver meu interesse pela docência, dedico novamente este livro.

## Sumário

Prefácio da primeira edição .....	9
Geração e acesso ao conhecimento .....	9
Prefácio da segunda edição .....	11
Introdução .....	14
<b>Capítulo 1</b>	
HISTÓRICO E PRINCIPAIS COMPONENTES DO EPI INFO® .....	18
<b>Capítulo 2</b>	
USANDO O EPI MAP .....	30
<b>Capítulo 3</b>	
CRIANDO UM BANCO DE DADOS.....	46
<b>Capítulo 4</b>	
GERENCIANDO OS DADOS.....	79
<b>Capítulo 5</b>	
IMPORTANDO DADOS E MANIPULANDO VARIÁVEIS .....	96
<b>Capítulo 6</b>	
ADICIONANDO DADOS EPIDEMIOLÓGICOS AO EPI MAP.....	106
<b>Capítulo 7</b>	
CRIANDO GRÁFICOS COM O COMANDO GRAPH .....	122
<b>Capítulo 8</b>	
ANALISANDO DADOS, UMA INTRODUÇÃO .....	144
<b>Capítulo 9</b>	
CONSTRUINDO E INTERPRETANDO TABELAS .....	154
<b>Capítulo 10</b>	
CALCULANDO E COMPARANDO MÉDIAS .....	163



## **Capítulo 11**

ESTUDANDO VARIÁVEIS NUMÉRICAS COM A REGRESSÃO LINEAR.....	174
--	-----

## **Capítulo 12**

ENTENDENDO A REGRESSÃO LOGÍSTICA E AS MEDIDAS DE RISCO .....	187
---	-----

## **Capítulo 13**

USANDO CURVAS DE KAPLAN-MEIER E REGRESSÃO DE DANO DE COX NA ANÁLISE DE SOBREVIDA .....	202
---	-----

## **Prefácio da primeira edição**

### **Geração e acesso ao conhecimento**

A Ciência é muito mais que o simples acúmulo de conhecimento, é permitir que o próprio conhecimento seja repensado e reconstruído constantemente. A atual perspectiva de que todo conhecimento é passível de ser aprimorado, questionado e até mesmo substituído é fundamental para quem quer se aventurar na área científica. Nunca na história da humanidade a Ciência avançou de forma tão rápida e abrangente quanto nos dias de hoje. Conceitos como risco, acaso, aleatoriedade e incerteza, por exemplo, foram sendo incorporados no linguajar das pessoas. A Estatística é uma das importantes ferramentas necessárias para entender esses avanços, especialmente no processo de tomada de decisão frente à incerteza.

O constante avanço científico também remete a uma outra importante questão que é a da necessidade de preservar o rigor metodológico. Durante muito tempo houve uma confusão conceitual entre rigidez e rigor metodológico. Neste período, a Ciência reconhecia apenas uma abordagem como tendo rigor metodológico suficiente para ser aceito como válido. Ao longo do século XX, essa perspectiva foi contestada e alterada. Múltiplas abordagens metodológicas foram também aceitas como válidas, desde que mantido o rigor na sua aplicação. Ao contrário de reduzir a importância da Estatística, essa flexibilização fez com que a necessidade de familiarização com análises estatísticas fosse ampliada a outros segmentos científicos e sociais.

Um dos limitantes da difusão da Estatística era a dificuldade de obtenção de resultados através de análises muito trabalhosas e de difícil compreensão. A associação da Estatística com a Informática foi extremamente benéfica nesse sentido. Os inúmeros programas estatísticos vieram a facilitar a análise de dados. Cálculos que antes demoravam até mesmo dias para serem realizados estavam disponíveis de maneira mais rápida e confiável. Algumas dificuldades, contudo, ainda permaneciam: os ambientes de informática nem sempre eram muito amigáveis ao usuário, a montagem dos bancos de dados eram difíceis e poucos tinham os recursos financeiros e de informática necessários para utilizá-los.

O programa Epi Info® é um exemplo de como se pode progressivamente suplantando essas dificuldades. O programa foi concebido para ser utilizado por pessoas de diferentes nacionalidades e níveis educacionais. A sua característica de maior impacto, sem dúvida, foi a de ser um programa de livre distribuição. Isso

possibilitou que muitos usuários tivessem acesso a formas de análise estatística antes apenas disponíveis mediante grandes investimentos. A democratização de meios para permitir a geração e a compreensão de conhecimentos é fundamental para a construção de uma sociedade humana mais justa.

Os programas de informática disponibilizam inúmeros recursos que muitas vezes não são utilizados por que os usuários desconhecem até mesmo a sua existência, ou não sabem como acessá-los adequadamente. Muitas vezes são utilizados menos de 5% dos recursos oferecidos pelos sistemas.

As reações frente a conteúdos ainda não conhecidos variam desde a atração até a aversão. Os manuais disponibilizados pelos fabricantes ou autores deveriam ser elementos facilitadores para o adequado acesso ao sistema informatizado, porém não são lidos por muitos dos seus usuários. Sem o uso dos manuais, o aprendizado acaba por ser feito através de descoberta casual, tentativa, erro e informações prestadas por outros usuários, que também passaram por este mesmo processo. Os motivos alegados para não ler um manual são vários: tamanho dos manuais, ansiedade em utilizar o sistema, falta de hábito de estudo, texto de difícil compreensão e estrutura do material montada para quem já o conhece e não para quem o está explorando. Em muitos casos, todas essas características estão presentes em um mesmo manual, tornando-o acessível apenas para os usuários que já dominam a sua utilização, o que é um contrassenso. Um manual deveria ser um documento altamente amigável, isto é, deveria ser escrito de forma a possibilitar que o leitor se sentisse simultaneamente atraído e à vontade em explorar o sistema. Deve permitir que um iniciante se sinta tranquilo para avançar, para testar, para se apropriar do uso adequado do sistema que está buscando se familiarizar. Por outro lado, deve permitir que um usuário já mais qualificado também se sinta contemplado com suas explicações e orientações. O manual, como o seu nome mesmo diz, é algo para se ter à mão, é um volume de consulta e orientação, é um recurso para aprender e para sanar dúvidas.

No presente volume, o Professor Ângelo Bós conseguiu consolidar a sua experiência de usuário do Epi Info<sup>®</sup>, como pesquisador e professor, em um volume acessível. Este é um manual elaborado justamente com essas características, permitindo que os surpreendentes recursos disponíveis no programa Epi Info<sup>®</sup> sejam adequadamente compreendidos e utilizados.

Porto Alegre, abril de 2004.  
Professor José Roberto Goldim, PhD.

## Prefácio da segunda edição

“Qual é o  $n$  da amostra?” Esta é uma das primeiras perguntas que fazem tanto os que entendem de estatística quanto aqueles que querem que os outros pensem que eles entendem de estatística.

As outras perguntas a serem feitas, somente as fazem aqueles que, de fato, entendem de estatística. São aquelas relacionadas à coleta dos dados, à constituição do banco de dados e, fundamental, à escolha dos testes estatísticos adequados à pesquisa proposta e à interpretação dos resultados obtidos por meio dos testes utilizados.

Atualmente, existe um grande número de softwares que facilitam enormemente o trabalho de montagem de bancos de dados e a aplicação das – muitas vezes – extensas e complexas fórmulas matemáticas utilizadas nos testes estatísticos. Se, por um lado, essas ferramentas tornam o trabalho mais fácil para o pesquisador, também trazem embutido um risco muito grande, que é o de se obter com rapidez um resultado completamente errado ou inadequado. A facilidade em usar também traz consigo a facilidade de errar. Por isso, é essencial que um pesquisador tenha uma sólida formação em estatística se pretende usar essa poderosa ferramenta nas suas investigações científicas.

Na área das Ciências da Saúde, pode-se dizer que o pioneirismo na utilização de dados quantitativos e a sua análise para fundamentar uma argumentação se deve a Florence Nightingale, a enfermeira britânica que, ao publicar o seu livro *Notes on Nursing*, em 1860, revolucionou a maneira de cuidar de pacientes hospitalizados. Durante a Guerra da Criméia, iniciada em 1854, com a declaração de guerra entre, de um lado, a Rússia e, de outro, Inglaterra, França e Turquia, o exército britânico foi duramente criticado por uma série de erros militares e pelas suas péssimas instalações hospitalares. Florence foi enviada para lá para introduzir e supervisionar o serviço de enfermagem nos hospitais de campanha.

Além do fato de ser mulher num ambiente militar em plena época Vitoriana, Florence se deparou com resistências e críticas de toda ordem ao tentar reformular o sistema hospitalar vigente. Os soldados eram, muitas vezes, deitados diretamente no chão sujo e rodeados de insetos e ratos. As cirurgias eram realizadas em condições sem um mínimo de higiene e o pós-operatório era um martírio que desrespeitava a condição humana. Além das mortes nos campos de batalha, muitos soldados morriam porque eram levados aos hospitais. Porém, não morriam por causa dos ferimentos da luta, mas de cólera e tifo.

Ao deparar-se com essa realidade, Florence iniciou sua árdua tarefa de tentar mudar os hábitos e a cultura arraigados num sistema que se caracterizava pelo descuido no tratamento das pessoas e, principalmente, pela mortalidade de muitos que tinham problemas de saúde simples e que perdiam sua vida por causa do próprio sistema hospitalar. A maneira que Florence escolheu para convencer os responsáveis pelo sistema hospitalar e seus superiores hierárquicos foi a demonstração lógica a partir de dados, números, tabelas e gráficos. O que para nós, hoje, parece óbvio, naquele momento provocou tal repercussão que convenceu a todos da necessidade de mudanças drásticas. Da mesma maneira, foi possível demonstrar o resultado que tais mudanças promoveram posteriormente nas condições de saúde das tropas e dos hospitais.

Ao utilizar os dados coletados criteriosamente nos hospitais de campanha desde o momento da sua chegada à zona de combate, organizá-los em bancos de dados apropriados e realizar cálculos de índices de mortalidade, Florence demonstrou que suas ações eram eficazes. O simples fato de estabelecer um suprimento de água fresca, preparar refeições que incluíam frutas e vegetais e substituir alguns equipamentos hospitalares fez com que as taxas de mortalidade se alterassem significativamente. Em alguns meses, a taxa que era de 60% quando Florence e sua equipe chegaram ao front baixou para 42,7% no verão de 1855. Na primavera do mesmo ano, a taxa de mortalidade nos hospitais de campanha britânicos estava em 2,2%!

Essa maneira de apresentar e defender sua argumentação e os resultados obtidos não foi um processo intuitivo. Nightingale tinha uma sólida formação em Matemática, era extremamente cuidadosa com seus registros e rigorosa no tratamento dos dados e nos cálculos que fazia.

Ao retornar para Londres, continuou aplicando e refinando as suas metodologias de coleta e análise de dados em todos os hospitais militares. Assim, demonstrou que a taxa de mortalidade de soldados na faixa etária entre 20 e 35 anos de idade em tempo de paz era o dobro da dos civis por causa das condições sanitárias dos hospitais militares. A divulgação do caso provocou uma investigação formal, fazendo com que a Rainha Vitória e o primeiro-ministro Lord Palmerston estabelecessem a Comissão Real sobre a Saúde nas Forças Armadas.

As contribuições de Florence Nightingale para a estatística hospitalar e as repercussões que tiveram nos sistemas de saúde fizeram com ela se tornasse a primeira mulher eleita como membro da Real Sociedade de Estatística do Reino Unido.

A história de Florence Nightingale nos leva a imaginar: e se ela tivesse à sua disposição um computador com o software Epi Info instalado? Inicialmente, podemos pensar que ela ficaria extremamente feliz e entusiasmada com o poder e eficiência da ferramenta que estaria à sua disposição. Mas, inteligente como era, se daria conta de imediato que necessitaria aprender como utilizar o programa para tirar dele resultados confiáveis e informações consistentes. Ela iria precisar de um manual como este “Epi Info sem mistérios: um manual prático”, escrito brilhantemente pelo professor Ângelo Bós e que hoje temos em mãos.

Escrito para quem está pretendendo usar o Epi Info®, para quem já usa e para quem esqueceu como se faz “aquele” teste especificamente, esta obra – não por acaso – está em sua segunda edição. Com sua estrutura passo a passo apresentada de maneira objetiva e simplificada, tem por trás a longa experiência e extenso conhecimento do professor Bós no tratamento de dados quantitativos e do seu ensino na área das Ciências da Saúde.

O usuário deste manual terá uma excelente ajuda para obter o máximo possível dos recursos que o Epi Info® coloca à sua disposição. Mas, como o exemplo de Florence Nightingale nos mostra, é preciso ter conhecimentos sólidos, perspicácia e preocupação com as pessoas – e não somente com os números – para que o trabalho da investigação científica tenha relevância para a sociedade.

Porto Alegre, abril de 2010.

Emilio A. Jeckel Neto

Professor titular da Faculdade de Biociências – PUCRS.

## Introdução

Epi Info<sup>®</sup> é um pacote que contém uma série de programas desenvolvidos para o *Microsoft Windows*. Os programas foram criados pelo Centro para o Controle e Prevenção de Doenças (*CDC*), do governo americano, e são destinados ao uso por profissionais de saúde que conduzem investigações de epidemias, administração de bancos de dados para vigilância de saúde pública e outras tarefas, além de ser um banco de dados para uso geral e aplicações estatísticas. Com Epi Info<sup>®</sup>, alunos e profissionais de saúde, bem como outros trabalhadores podem desenvolver um questionário rapidamente ou podem personalizar o processo de entrada de dados e gerenciar a entrada e a análise de um banco de dados.

São produzidas análises estatísticas, gráficos e tabelas com comandos simples como *READ*, *FREQ*, *LIST*, *TABLES* e *GRAPH*. Um componente chamado Epi Map exibe mapas geográficos com dados do Epi Info<sup>®</sup> ou de outro banco de dados.

Por ser desenvolvido pelo Governo Americano, o Epi Info<sup>®</sup> é um programa de domínio público, ou seja, pode ser baixado a partir da Internet gratuitamente. Cópias de CD-ROM e manuais impressos, assim como a tradução do programa, podem ser disponibilizados por vendedores privados.

As principais vantagens que o Epi Info<sup>®</sup> oferece frente a outros programas é que ele é um programa, além de gratuito, completo, pois proporciona a criação de um formulário de entrada de dados, um programa de introdução de dados e um programa que não só analisa, mas, também, cria gráficos e relatórios. Muitos pesquisadores já estão acostumados a criar o banco de dados em um programa, criando uma planilha de dados, depois utilizar outro programa para a análise de dados e, finalmente, um programa gráfico para gerar gráficos e tabelas. O leitor terá a oportunidade, neste manual, de observar e se familiarizar com as inúmeras aplicações que o Epi Info<sup>®</sup> proporciona.

O presente livro nasceu a partir de aulas desenvolvidas pelo autor nos cursos: “Introdução ao uso do Epi Info<sup>®</sup> 2002”, “Análise de dados com o Epi Info<sup>®</sup> 2002” e “Curso internacional de introdução e análise de dados com o Epi Info<sup>®</sup> 2002” promovidos pelo Programa de Gerontologia Biomédica e Pró-reitoria de Extensão da Pontifícia Universidade Católica do Rio Grande do Sul, também nos cursos lecionados pelo autor e promovidos pelo Centro de Estudos sobre AIDS do Rio Grande do Sul (CEARGS) e pela School of Public Health das ilhas Fiji, Pacífico Sul. O livro, assim, assume não somente

as características de um manual, mas também de um curso teórico-prático com exercícios a serem desenvolvidos durante o decorrer dos capítulos. Recomenda-se ao leitor, na sua primeira utilização, seguir a leitura desde o primeiro capítulo, executando os exercícios de todos os capítulos. Os arquivos utilizados durante a execução dos exercícios propostos podem ser obtidos nas pastas *Dados e Exercícios* da página [www.pucrs.br/igge/epiinfo](http://www.pucrs.br/igge/epiinfo). Aconselha-se a cópia dos mesmos no computador do leitor, em uma pasta (folder) própria (ex.: EpiAula). Entretanto, o livro pode ser utilizado, também, como um manual, pois destaca os comandos e suas opções facilitando ao leitor localizá-los. O índice remissivo ajudará o leitor a encontrar as informações que necessita para determinado comando. No decorrer dos capítulos, o leitor terá a oportunidade de se familiarizar com outros sites como o do IBGE (Instituto Brasileiro de Geografia e Estatística) e do DATASUS (Banco de Dados do Sistema Único de Saúde do Ministério da Saúde Brasileiro), onde conseguirá algumas informações e arquivos necessários à execução de exercícios propostos e quem sabe até utilizar esses dados em sua própria pesquisa.

Alguns pesquisadores, assim como alguns alunos do autor, podem ter tido a oportunidade de utilizar a versão até então mais popular do Epi Info®, a versão 6.0 para DOS. No decorrer dos capítulos, eles vão observar diversas semelhanças da versão atual com a antiga e terão certa facilidade em utilizar alguns dos comandos desta versão. O autor preocupou-se em utilizar algumas das características em comum a ambas versões para facilitar a adaptação dos usuários do sistema anterior.

O livro segue, em parte, o roteiro das aulas desenvolvidas. Em alguns capítulos é necessária a conexão com a Internet. Inicialmente, expõe-se um pouco da história do programa familiarizando o leitor a algumas expressões e peculiaridades do Epi Info®. Imediatamente, o mesmo passa a familiarizar-se com o site do Epi Info®, onde poderá, não somente baixar o programa de instalação, mas também observar a existência de novas versões. Ainda neste capítulo, observam-se os passos necessários à instalação do programa, bem como onde procurar sua tradução para o português, além de apresentar o menu principal do Epi Info® e seus diversos componentes. Como o desenvolvimento da versão em português muitas vezes não acompanha a atualização do Programa Epi Info®, no livro é utilizada a versão em inglês. Na experiência do autor, quando o usuário está familiarizado com a versão em inglês poderá utilizar a versão em português sem problema. Entretanto, caso o mesmo, inicialmente, familiarizar-se com a versão em português e passar a não poder



mais utilizá-la, porque o programa foi atualizado, por exemplo, esse terá muitas dificuldades em utilizar a versão original em inglês. Nos capítulos, cada comando é descrito em sua versão original em inglês com a versão (caso existente) em português entre parênteses. Os comandos e funções originais estão sempre digitados em itálico para facilitar o entendimento.

No segundo capítulo é introduzido o uso do Epi Map, um programa desenvolvido para gerar mapas de distribuição epidemiológica. Um exercício prático sobre como criar um mapa de distribuição da população brasileira será executado. Após esse exercício, visita-se o site do IBGE, onde se conseguem mapas mais detalhados do que o disponível no site do Epi Info<sup>®</sup> sobre as Unidades da Federação do Brasil.

O terceiro capítulo exemplifica como se cria um banco de dados, seus passos e as principais opções que o Epi Info<sup>®</sup> disponibiliza para facilitar o trabalho de digitação dos dados, que é assunto também abordado neste capítulo. Descrevem-se detalhadamente os diferentes tipos de variáveis utilizados pelo Epi Info<sup>®</sup>, bem como as características de cada um deles. Também são abordados alguns instrumentos de controle de introdução de dados, agrupamento de variáveis, criação automática de variáveis e criação de banco de dados acessório. Cuidados na digitação dos dados e características do programa de introdução dos dados são particularizados durante a execução do exercício proposto.

No quarto capítulo, o leitor tem o primeiro contato com o programa mais importante do Epi Info<sup>®</sup>: o programa análise de dados. Este capítulo aborda noções de gerenciamento de dados: abrir um banco de dados, observar e editar valores, juntar dois bancos de dados diferentes, criar novas variáveis e, finalmente, salvar o banco de dados editado. É também utilizado o programa editor do Epi Info<sup>®</sup>, que facilita a execução de comandos semelhantes, além de ser muito semelhante ao encontrado na versão 6.0 do Epi Info<sup>®</sup>.

O quinto capítulo elucida como se importa um banco de dados criado por outro programa ou pela versão *DOS* do Epi Info<sup>®</sup>. Observa-se a utilidade do comando *IF (Se)*, como se salva e executa um programa de comandos, como se cria um questionário (ficha) a partir de um banco de dados e como se observa a frequência de determinadas variáveis. Finalmente, o leitor observará como se compacta o arquivo do Epi Info<sup>®</sup>.

No capítulo sexto, observa-se como obter dados epidemiológicos para a criação de mapas. São importados dados de mortalidade e da população brasileira visitando o site do DATASUS. Para tanto são abordadas noções

de utilização do programa *Excel* do Microsoft, configurando os bancos de dados importados do site brasileiro para a posterior importação e utilização pelo Epi Info®. O capítulo também aborda como juntar dois arquivos que têm conteúdos (variáveis) diferentes. Finalizando o capítulo, observa-se como se pode calcular a taxa de mortalidade com o *Epi Info*® e como esses dados são transferidos para um mapa.

O capítulo sétimo versa sobre o comando de criação de gráficos. Os diferentes tipos de gráficos disponibilizados pelo Epi Info® serão minuciosamente descritos, abordando, principalmente, a sua utilidade com exemplos práticos.

Noções de análise de dados são introduzidas no capítulo 8. O capítulo revisa conceitos básicos, como tipos de variáveis, importância da análise descritiva, hipótese nula, hipótese alternativa e os tipos de erros relacionados às hipóteses. Neste contexto, o programa *StatCalc*, do Epi Info®, é utilizado para calcular o tamanho amostral de uma pesquisa.

No capítulo nono, é abordada a noção de análise de dados categóricos, com teste do Qui-quadrado, utilizando inicialmente o programa *Tables* do *StatCalc*. Na continuação, a construção de tabelas, a partir de um banco de dados, é também explicada, utilizando o programa *Analysis* do Epi Info®, junto com noções de interpretação da distribuição dos valores nas tabelas, bem como a interpretação da significância do Qui-quadrado na rejeição da sua hipótese nula.

O cálculo de médias usando o Epi Info®, com sua interpretação e análise, é o assunto do capítulo décimo. Tanto o teste t de Student quanto a Análise de Variância são frutos de revisão teórica e suas aplicabilidades são discutidas.

Regressão linear, suas bases teóricas e sua utilização, bem como sua expressão gráfica, é o assunto do capítulo décimo primeiro do livro que é seguido da Regressão Logística no capítulo décimo segundo, onde são revistas noções de análise de medidas de risco utilizando também o programa *Tables* do *StatCalc*. O último capítulo do livro trata das bases teóricas da análise de sobrevivência e a utilização de dois instrumentos de análise do Epi Info®: Curvas de Kaplan – Meier e da Regressão de Dano de Cox.

# Capítulo 1

## HISTÓRICO E PRINCIPAIS COMPONENTES DO EPI INFO®

### Breve Histórico do Epi Info®

Até metade dos anos 70, os programas de computação eram inflexíveis: tanto os comandos e cálculos estatísticos quanto os dados a serem analisados eram colocados no computador manualmente. Esse processo exigia a participação de um programador de computador que, muitas vezes, não estava presente ou disponível. A ideia de se tornar os programas de computação, usados em saúde pública, mais flexíveis, possibilitando que epidemiologistas utilizassem os programas sem a necessidade de um programador, surgiu em 1976. Nesse ano, um auditório do CDC, convertido em quartel-general para a crise da Gripe Suína de 1976, estava cheio de epidemiologistas e um Equipamento *Digital PDP 11*, um minicomputador do tamanho de um refrigerador. Naquele momento, um programa chamado *SOCRATES*, escrito em linguagem *FORTRAN* pelo programador Rick Curtis, possibilitou aos epidemiologistas definirem questões, entrada de dados e produzir resultados em formato tabular (tabelas) sem a presença de um programador.

A primeira versão do Epi Info® foi lançada somente em setembro de 1985. Era chamado de “Sistema de Análise Epidemiológica”, conforme a figura 1.1.

```
Epidemiologic Analysis System
Centers for Disease Control
Epidemiology Program Office
Jeffery A. Dean - Author

September 1, 1985

A.  Date Entry and Editing   <EnterInfo>
B.  Data Conversion Routines <Convert>
C.  Data Output and Analysis <Crosstab>
D.  Return to Master Menu

Enter option letter desired:
C:\EAS>
```

Figura 1.1. Tela principal da primeira versão do Epi Info®.

Fonte: site do Epi Info®: [www.cdc.gov/epiInfo/epiinfo1.htm](http://www.cdc.gov/epiInfo/epiinfo1.htm), último acesso em 20/02/2010.

O programa era basicamente composto de três comandos digitados diretamente no monitor e foi criado em linguagem *DOS*. A primeira versão possibilitava somente a introdução e edição dos dados (programa *EnterInfo*), criação de regras de conversão de dados (programa *Convert*) e construção de tabelas de convergência (*Crosstab*). A primeira versão do Epi Info® produzia tabelas simples de contingência sem nenhuma estatística, conforme a figura 1.2. Cinco meses após o lançamento do programa, um Manual de 27 páginas foi criado e o nome do programa *EnterInfo* foi chamado de Epi Info®.

```
EAS> TABLES VANILLA ILL
```

VANILLA		ILL		Total
N	Y	N	Y	
N	18	3		21
Y	11	43		54
Total		29	46	75

```
EAS>
```

Figura 1.2. Exemplo de tabela produzida pela primeira versão do Epi Info®.  
Fonte: site do Epi Info®: [www.cdc.gov/epiinfo/epiinfo1.htm](http://www.cdc.gov/epiinfo/epiinfo1.htm), último acesso em 20/02/2010.

Em 1987, foi lançado o Epi Info®, versão 2, escrito em linguagem *Turbo Pascal*, que tinha um manual de 112 páginas, um menu de navegação com o cursor, o *EPED*, processador de textos para a criação de questionários; *ENTER*, *ANALYSIS*, um calculador epidemiológico chamado *STATCALC*, *EPICURVE* para a criação de histogramas, um programa para juntar bancos de dados e programas *IMPORT* e *CONVERT* para a conversão de arquivos. O comando *TABLES* começa a oferecer resultados estatísticos.

```
EPI> read c:\epi5\data\palufin
```

```
EPI> tables sexe palu
```

SEXE	PALU		Total
	0	1	
1	354	62	416
2	458	94	552
9	17	4	21
Total	829	160	989

```
An expected cell is < 5. Chi square is not valid.
Chi square = -17190.22
2 degrees of freedom.
p value = 1.00000000
```

```
EPI>
```

Figura 1.3. Exemplo de tabela produzida pelo Epi Info®, versão 2.  
Fonte: site do Epi Info®: [www.cdc.gov/epiinfo/epiinfo23.htm](http://www.cdc.gov/epiinfo/epiinfo23.htm), último acesso em 20/02/2010.

A partir da versão 4 do Epi Info®, a Organização Mundial da Saúde (OMS) iniciou a colaborar com a CDC na elaboração e distribuição do programa, tornando-o internacionalmente conhecido. A versão mais popular foi o Epi Info® 6.0 (figura 1.4), lançado em 1992 e que continua sendo utilizado até hoje. Até 1997, haviam sido realizadas 145 mil cópias das versões *DOS* do Epi Info® em 117 países. O manual e programa da versão *DOS* foram traduzidos para 13 idiomas.



Figura 1.4. Tela principal do Epi Info®, versão 6.

Fonte: site do Epi Info®: [www.cdc.gov/epiinfo/epiinfo6.htm](http://www.cdc.gov/epiinfo/epiinfo6.htm), último acesso em 20/02/2010.

Em 1994, foi criada uma página na Internet, possibilitando o *download* do programa gratuitamente e o acesso a versões mais atualizadas do Epi Info®. Uma revisão interna das visitas ao servidor do Epi Info® site observou que, em 2001, foram documentados mais de 250 mil *downloads*. A versão 2000, lançada em 2001, transporta o Epi Info® para a tecnologia *Windows*, utilizando o formato do programa de gerenciamento de dados *Microsoft Access*. No final de 2002, uma nova versão do Epi Info® para *Windows* foi lançada. A popularidade do Epi Info® tem crescido muito com essas novas versões, o site menciona mais de 1 milhão de cópias baixadas. Lembramos que, como os programas de instalação podem ser copiados livremente, muito mais computadores devem ter o Epi Info® instalado. Novas versões têm sido lançadas com frequência. Para buscar atualizações, é recomendável ao usuário do Epi Info® visitar o site do programa. Atualizações deste livro também são disponíveis no site [www.pucrs.br/igge/epiinfo](http://www.pucrs.br/igge/epiinfo).

## Sobre o Epi Info®

O Epi Info® é escrito em programação *Visual Basic* e usa o formato de arquivo *Microsoft Access* (*.mdb*) como uma forma de padronizar o banco de dados tornando-o acessível para outros programas. Embora os dados do Epi Info® sejam armazenados em formato *.mdb*, outros formatos de arquivos podem ser lidos, analisados, importados ou exportados pelo programa.

O Epi Info® inclui um programa que usa o Sistema de Informação Geográfica (*GIS*), chamado Epi Map, executado por meio de um programa atualmente denominado *Criar Mapas (Create Maps)*. Esse sistema foi construído a partir do programa *MapObjects* da empresa Environmental Systems Research (ESRI), produtores também do *ArcView*. Epi Map é compatível com dados de *GIS* de numerosos locais de Internet nos populares formatos *ESRI*. Veremos no capítulo seguinte como conseguir mapas brasileiros no site do *IBGE*.

A versão para *Windows* do Epi Info® mantém muitas características do conhecido Epi Info® para *DOS*, enquanto oferece as vantagens do ambiente *Windows*, como a facilidade de uso, gráficos, fontes e impressão “indolor”. Os programas, documentação e materiais pedagógicos estão no domínio público (embora “Epi Info” seja marca registrada de *CDC*), e pode ser copiado livremente, distribuído ou traduzido.

## Como conseguir o Epi Info®

Existem basicamente duas maneiras de se conseguir a versão mais atual do programa do Epi Info®.

### 1. Download a partir do Website do CDC:

O CDC mantém sempre a versão mais atual do programa. Para baixar o programa e instalá-lo devemos seguir os seguintes passos:

- a. Ao visitar o site [www.cdc.gov/epiinfo](http://www.cdc.gov/epiinfo) vão surgir duas opções: “*Web Install*”, que teoricamente instala o programa diretamente do website sem salvar os arquivos de instalação no computador (esta opção salva espaço), e “*Download Setup.exe*”, que baixa o programa completo de instalação. Na prática, ambos os comandos salvam o programa no computador, necessitando pelo menos 70 megabites de espaço no computador.
- b. Recomendamos selecionar a opção “*Download Setup.exe*”, pois assim é possível salvar o programa em um local conhecido e reinstalar o programa se necessário ou até mesmo salvar em um CD-ROM para ser ofertado a outras pessoas. O tempo de instalação é praticamente o mesmo, visto que o *Web Install* literalmente baixa o programa no computador do usuário em uma pasta temporária.

- c. Uma nova janela vai aparecer, então escolha “salvar em disco” e clica *OK*;

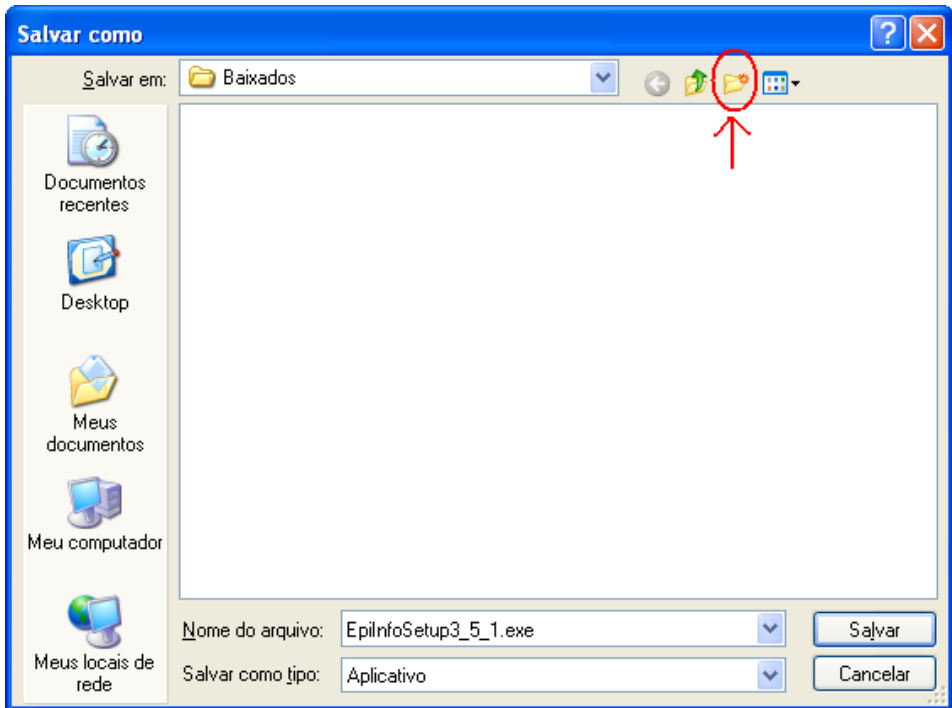


Figura 1.5. Janela padrão de salvamento de arquivo em ambiente *Windows* destacando a opção de criação de nova pasta. Fonte: Windows XP.

- d. A nova janela vai aparecer (figura 1.5), oferecendo opções de salvamento. É aconselhável ter um diretório (folder ou pasta) para armazenar os arquivos que são baixados da Internet. Para criar um novo *Folder*, clicamos no ícone que está circulado na figura. Isso cria uma nova pasta, que automaticamente se chama “*Nova pasta*” ou “*New folder*”. Podemos mudar o nome do novo *Folder* para, por exemplo, “*Baixados*”, sem precisarmos clicar em lugar nenhum. Após trocarmos de nome, apertamos o *ENTER* (*↵*) duas vezes. Depois, basta clicar em “*salvar*” ou “*save*”. Na opção *WebInstall*, um arquivo provisório é também salvo no computador. Nesse caso, as orientações são muito parecidas.


- e. Após salvar o arquivo, deve-se procurar novamente o novo folder no *Meu Computador*. O arquivo baixado é na realidade um arquivo que junta vários arquivos comprimidos. Ao clicar duas vezes no arquivo fará com que o arquivo se descomprima temporariamente. Aguarde um pouco até aparecer uma tela escrita “Epi Info®” e uma janela com o escrito “Next”. Clicar “Next” quatro vezes. E a instalação ficará completa. Observe que o programa é instalado num novo diretório chamado *Epi\_Info*.

**2.** Através do site [www.pucrs.br/igg/epiinfo](http://www.pucrs.br/igg/epiinfo), no diretório *Programas*: Basta clicar no arquivo *EpiInfosetup.exe* quatro vezes.

## Como instalar a versão (tradução) em português

Para instalar a versão em português, temos duas maneiras:

**1.** Conseguindo a versão em português para o Epi Info®, através do site: <http://www.pucrs.br/igg/epiinfo> pasta *Programas*.

Clicando à esquerda do mouse no link “*Tradutor para Português*”, selecione “Save link as” (Salvar destino como...). Assim, irá abrir uma janela com as opções de download. Na janela, aconselhamos escolher a opção salvar o arquivo em disco. Devemos salvar no mesmo lugar onde salvamos o arquivo de instalação. Devemos abrir o arquivo com o winzip. Se o seu computador não tiver o winzip e não conseguir abrir o arquivo, você poderá baixar o programa no site ([www.winzip.com](http://www.winzip.com)). Neste caso, a janela do *Winzip* mostrará somente um arquivo: *Português.exe*. Devemos clicar no ícone *Extract* . Na nova janela (figura 1.6), na esquerda em cima, existe uma opção “*Extract to*” e uma janela pequena. Nessa janela devemos colocar o folder *TRANSEXE* do *Epi\_Info* (diretório onde o Epi Info® foi instalado) conforme a figura a seguir:



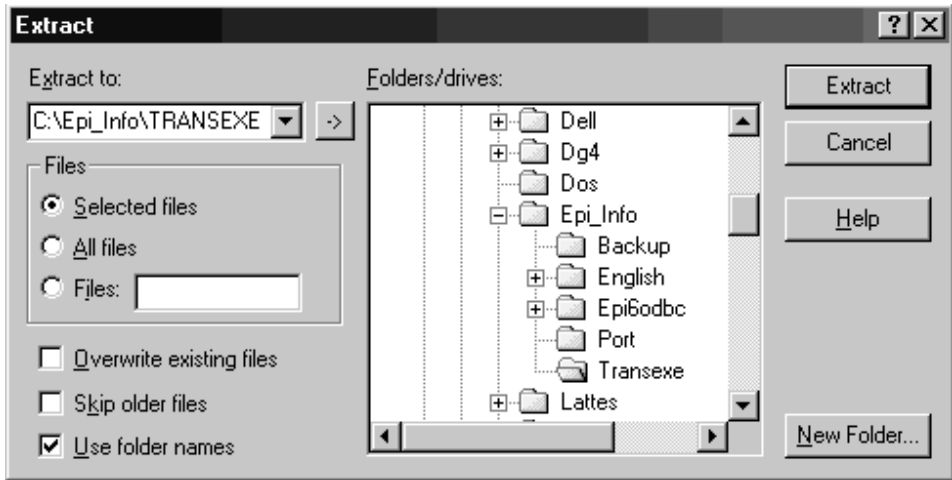


Figura 1.6. Tela de extração de arquivos compactados.  
Fonte: Winzip.

Feito isso, podemos clicar “*Extract*” que está na janela à direita. A tradução ainda não está instalada.

**2.** Através de um disquete (ou CD) de instalação: copiar o arquivo *Português.exe* para o diretório *C:\Epi\_Info\TRANSEXE*.

Devemos, agora, iniciar o *Epi Info*<sup>®</sup>, fechando as janelas e clicando no Ícone do *Epi Info*<sup>®</sup> (EPI) na tela principal do computador. Com o *Epi Info*<sup>®</sup> aberto, vemos que há um menu superior. Nesse menu, escolhemos a opção “*Settings*” e depois “*Install Language*”. Na janela escolhemos “*Port*” e depois clicamos em “*Install*”. Surgirá uma mensagem dizendo que uma nova janela aparecerá e que essa janela poderá não fechar automaticamente. Clicando “*OK*”, observamos uma janela preta do *Prompt DOS* extraindo dois arquivos. Se na barra superior aparecer a palavra “*finished*” e a janela continuar aberta, fechar essa janela clicando no “*X*” do canto superior direito dela. Finalmente pode-se sair do programa de instalação de linguagem e escolher a linguagem Português, no mesmo menu “*Settings*”, agora “*Choose Language*” escolher “*Port*” e “*OK*”. Observa-se que, na tela, à esquerda, em cima, logo abaixo do menu *Programs* aparece os dizeres “*Language=Português*”.

## Problemas com o uso da tradução para o Português

Algumas vezes, o Epi Info® “des-seleciona” a linguagem automaticamente. Se isso acontecer, basta escolher a linguagem novamente. A linguagem não necessita ser instalada outra vez. Infelizmente, nem todos os comandos foram traduzidos e, às vezes, encontramos dificuldades em usar a linguagem em português, pois ela interfere em alguns comandos, principalmente o comando *Criar Ficha (Make View)*. Em algumas versões do *Windows*, como o *Windows 2000*, a tradução não funciona. Além disso, como o Epi Info® está sempre sendo atualizado, novas versões podem não ser compatíveis com a linguagem. Por essa razão, recomendamos a não utilização dessa linguagem, ou desistir dela quando começar a dar problema. Em função desses problemas, neste manual é utilizada a versão em inglês.

## Componentes de Epi Info®

O Epi Info® tem vários componentes; podemos dividi-los em componentes de tela e componentes de menu, de acordo com a sua disposição (figura 1.7).



Figura 1.7. Tela principal do Epi Info®, versão 3.5.1.  
Fonte: Epi Info®.

**1.** *Componentes de tela:* são componentes que aparecem imediatamente quando se executa o Epi Info®, eles aparecem também no Menu Programas.

Versão em inglês	Versão em português
Make View	Criar Ficha
Enter Data	Entrar Dados
Analyze Data	Analisar Dados
Create Reports	Criação de Relatório
Create Maps	Criar Mapas (Epi Map)

**2.** *Componentes de menu:* são componentes que estão localizados na barra de menu da janela do Epi Info®. Esses componentes possuem funções de configuração (troca ou instalação de linguagem e escolha da versão do Banco de Dados), programas, já vistos anteriormente, edição da imagem do Epi Info®, e utilitários onde temos o tradicional *Statcalc*, cuja utilidade será observada nos capítulos oitavo, nono e décimo primeiro. No menu *Ajuda (Help)* também há dois programas de treinamento: o programa Tutorial e exercícios.

## Menu Edit (Editar)

O menu *Edit* possui basicamente comandos que permitem personalizar a tela do Epi Info®. São quatro comandos:

**1.** *Picture (tela):* altera o fundo da tela do Epi Info®. Clicando nesse comando, o programa nos direciona a escolher outro arquivo tipo figura ou gráfico para substituir. Podemos escolher outras figuras existentes no próprio diretório do Epi Info® ou em outro diretório qualquer como, por exemplo, no folder “*imagens*” dos *meus documentos*.

**2.** *Edit this menu:* abre uma tela de edição do arquivo *EpiInfo.MNU* que, em linguagem Visual Basic, nos permite alterar todos os comandos e arquivos gerenciados pelo Epi Info®. É recomendável não alterar esse arquivo, principalmente

se não entendermos como ele é escrito. É interessante ter uma cópia em algum lugar seguro. Para tanto basta clicar no menu Arquivo e selecionar *Save as (Salvar como)*.

**3.** *Bottons on or off*: somente esconde ou mostra os botões dos programas de tela.

**4.** *Move/Resize button <Shift F2>*: Permite que alteremos o tamanho e posição dos botões dos programas de tela. Isso pode ser útil quando queremos destacar um programa a ser utilizado mais constantemente. É útil principalmente para pessoas que têm “medo” de computador e não conseguem “enxergar” o programa que queremos que ela use. Podemos, com esse comando, também apagar algum dos programas, o que é muito perigoso, pois é difícil recuperarmos um botão apagado. Esse comando, na realidade, faz uma edição do arquivo *EpiInfo.MNU*.

## Menu Settings (Configurações)

Além de instalar e modificar a linguagem, o menu *Settings* proporciona outras configurações que podem ser úteis:

**1.** *Choose Epi Info® Database Version*: possibilita trabalharmos com uma versão de arquivo *MS ACCESS 97* ou *2000*. Isso é importante se quisermos editar o arquivo do banco de dados do Epi Info® usando esse programa, o que, na experiência do autor, é desaconselhável, mas algumas vezes necessária. Por definição, o Epi Info usa o formato *MS ACCESS 2000*.

**2.** *Set Working Directory*: automaticamente o Epi Info® determina que o diretório de trabalho seja o *Epi\_Info*. É interessante modificá-lo, mas existem outras maneiras de trabalhar com outros diretórios, como veremos mais adiante neste livro.

**3.** *Set INIFile Directory*: modifica o local do *EpiInfo.INI*, programa que estabelece algumas instruções iniciais ao *Windows*. Modificar esse arquivo só é aconselhável para pessoas que tenham experiência com essa linguagem.

## Menu Utilities (Utilidades)

**1.** *StatCalc*: Programa escrito em 1993 que faz alguns cálculos estatísticos interativamente; veremos a maioria deles nos capítulos seguintes.

**2.** *Datacompare*: compara dois bancos de dados e descreve as diferenças. Isso vale tanto em questão de dados diferentes quanto na inexistência de alguns dados. Esse programa é particularmente importante quando usamos o artifício da dupla digitação dos dados para minimizar o erro na introdução dos dados.

**3.** *Table-to-View*: cria uma ficha (*View*) para uma tabela (banco de dados) já existente; veremos mais adiante sua utilidade.

**4.** *Visualize Data*: Compacta e corrige o arquivo do banco de dados, propicia apagar “subarquivos” temporários. A compactação não interfere na utilização do arquivo tornando-o sim mais “leve”. Às vezes temos que utilizar essa ferramenta quando criamos muitas variáveis novas. Veremos, mais adiante, no capítulo quinto, sua utilidade e execução.

**5.** *EpiLock*: Programa que serve para codificar ou comprimir um arquivo. Codificação é importante para salvarmos o arquivo de espíões, e a compressão serve para diminuir o tamanho do arquivo para transporte, distribuição ou armazenamento.

**6.** *Compact*: Compacta o banco de dados, o mesmo programa que está contido no Visualize Data.

**7.** *Word Processor*: Chama-se o processador de texto (na maioria das vezes, o WordPad).

## Exercício:

1. Salve o arquivo *EpiInfo.MNU* no diretório *Epi\_Info/Backup*.
2. Modifique a tela da apresentação do Epi Info®.
3. Altere as posições dos botões da tela.
4. Modifique a linguagem para inglês.
5. Restaure o Menu do Epi Info®, copiando novamente o arquivo *EpiInfo.MNU* salvo no diretório Backup para o diretório principal do Epi Info® (opcional).

## Capítulo 2

# USANDO O EPI MAP

O Epi Map é um programa que permite construir uma variedade de mapas com dados de distribuição geográfica e que oferece um grau alto de compatibilidade com o programa *ArcView* do Instituto de Pesquisa de Sistemas Ambiental, Inc. (*ESRI*).

### **Baixando mapas para o Epi Map no site do Epi Info®**

Vários mapas são distribuídos gratuitamente no site do Epi Info® ([www.cdc.gov/EpiInfo](http://www.cdc.gov/EpiInfo)), junto com link para várias informações epidemiológicas tanto dos Estados Unidos quanto de outros países. Assim, há no site do Epi Info® um *link* chamado *Maps* à direita da tela na seção de *download*. Esse *link* nos leva para o que o site chama de *Resources for Creating Public Health Maps*. Nessa página encontram-se, então, os *links* que mencionamos. O *link* que nos interessa agora é o primeiro: *Download free shapefiles for Epi Info®*. Como o nome diz, é gratuito e nos permite baixar os arquivos de formato dos mapas. Clicando na primeira palavra, o site nos direciona a uma página onde estão todos os continentes. Clicando no continente América do Sul teremos o nosso Brasil. Observamos que o programa solicita um local para salvar. Podemos salvar o arquivo em um diretório novo no *c:Epi\_Info*, chamado, por exemplo, *Mapas*. O arquivo a ser salvo chama-se de *BR.exe*. Esse é um arquivo tipo Zip que contém cinco arquivos compactados, que podem ser extraídos mesmo sem o programa *Winzip*.

Uma cópia desse arquivo, junto com todos os outros mapas do site do Epi Info®, pode ser encontrada também no site [www.pucrs.br/igg/epiinfo](http://www.pucrs.br/igg/epiinfo) no diretório *Programas/Maps/South America*.

### **Descompactando o BR.EXE**

Para descompactar os arquivos do mapa do Brasil, localizaremos o arquivo *BR.exe* no diretório *c:/Epi\_Info/Mapas*, usando ou o *Meu computador* ou o *Windows Explorer*. Clicando duas vezes no arquivo aparecerá uma mensagem avisando da descompressão dos arquivos. Ao clicar *OK*, aparece uma nova janela, conforme a figura 2.1, informando o local onde serão descompactados os arquivos. Automaticamente, o Epi Info® posiciona um

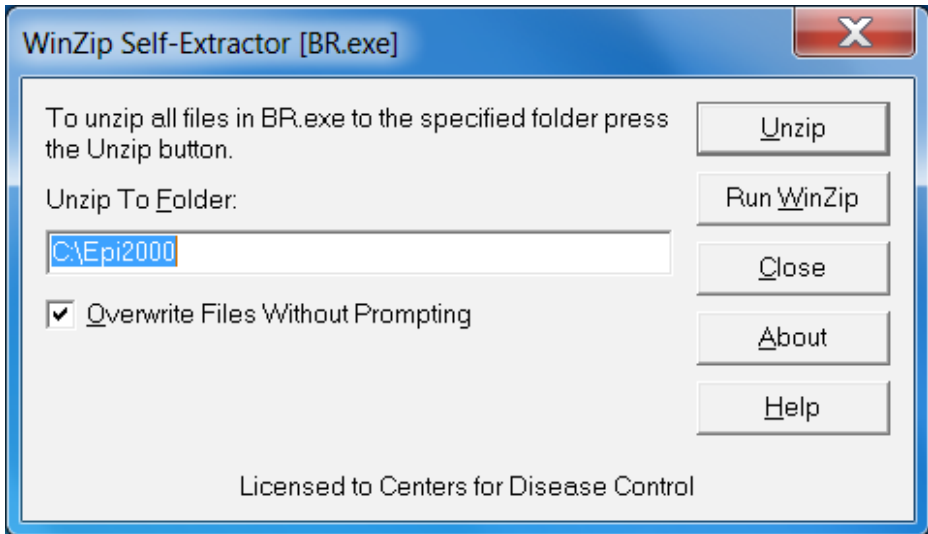


Figura 2.1. Tela de autoextração de arquivos.

Fonte: Winzip self-extractor.

Não nos interessa criar um novo diretório com esse nome, pois irá nos atrapalhar no futuro. É aconselhável mudar o destino para *c:/Epi\_Info/Mapas*, pois queremos deixar os mapas organizados. Uma nova mensagem informa a descompressão bem-sucedida dos arquivos. Veremos, a seguir, o conteúdo do arquivo descomprimido.

### Utilizando o programa Create Map (Criar mapa)

Como vimos, o Epi Map é um programa que permite construir uma variedade de mapas com dados de distribuição geográfica. Vimos também que o Epi Map oferece um alto grau de compatibilidade com o programa *ArcView* do Instituto de Pesquisa de Sistemas Ambiental, Inc. (*ESRI*). Na tela principal do Epi Info®, clicamos no programa *Create Map*. A tela que aparece (figura 2.2) é muito estimulante, pois aparece uma foto aérea de algum local. Mas os mapas não são assim tão reais.



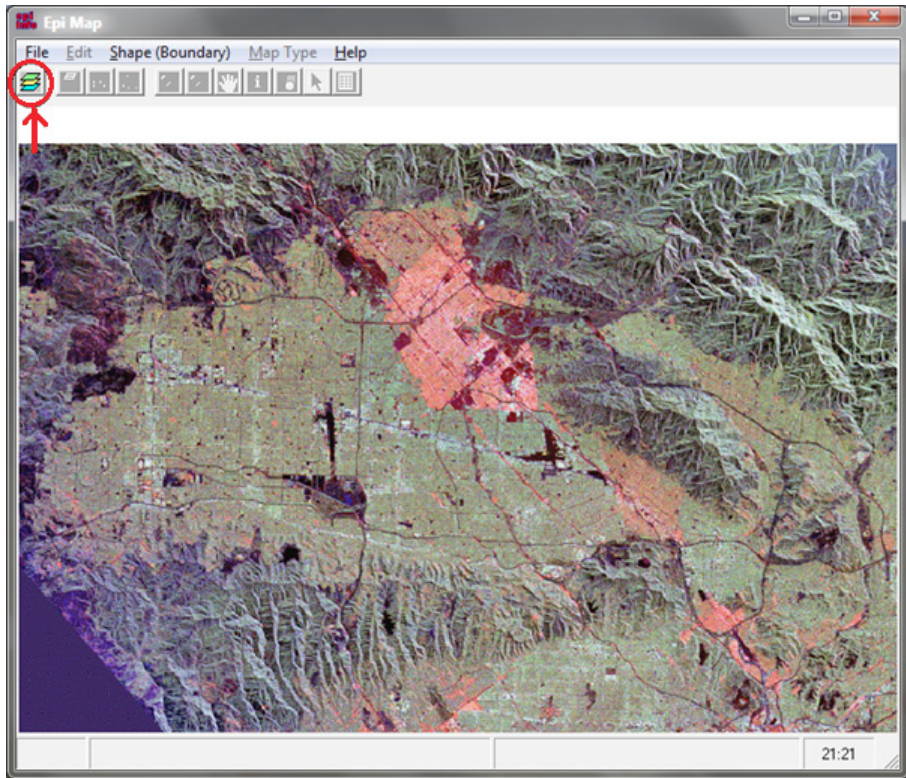


Figura 2.2. Tela inicial do Epi Map, destacando o ícone do *Map Manager*.

Fonte: Epi Map.

Os arquivos dos mapas distribuídos pelo site do Epi Info® contêm, na realidade, cinco diferentes arquivos compactados, um desses arquivos tem a extensão *.shp*, que significa “Shape”, ou formato em inglês. Esse arquivo *Br.shp* contém os formatos dos estados brasileiros e conecta com os outros arquivos. Existe também outro arquivo chamado *Br.dbf*, que é escrito em *Dbase IV*, ele é um arquivo de banco de dados que contém algumas informações sobre o Brasil e os estados. Para iniciar, devemos abrir um formato (Shape) através do *Map Manager* (*Gerenciador de Mapas*).

### Utilizando o Map Manager (Gerenciador de Mapas)

Na primeira janela do programa *Create Map*, aparece somente um ícone, o primeiro à esquerda, que é o ícone do administrador ou gerenciador de mapas (*Map Manager* – circulado na figura 2.2).

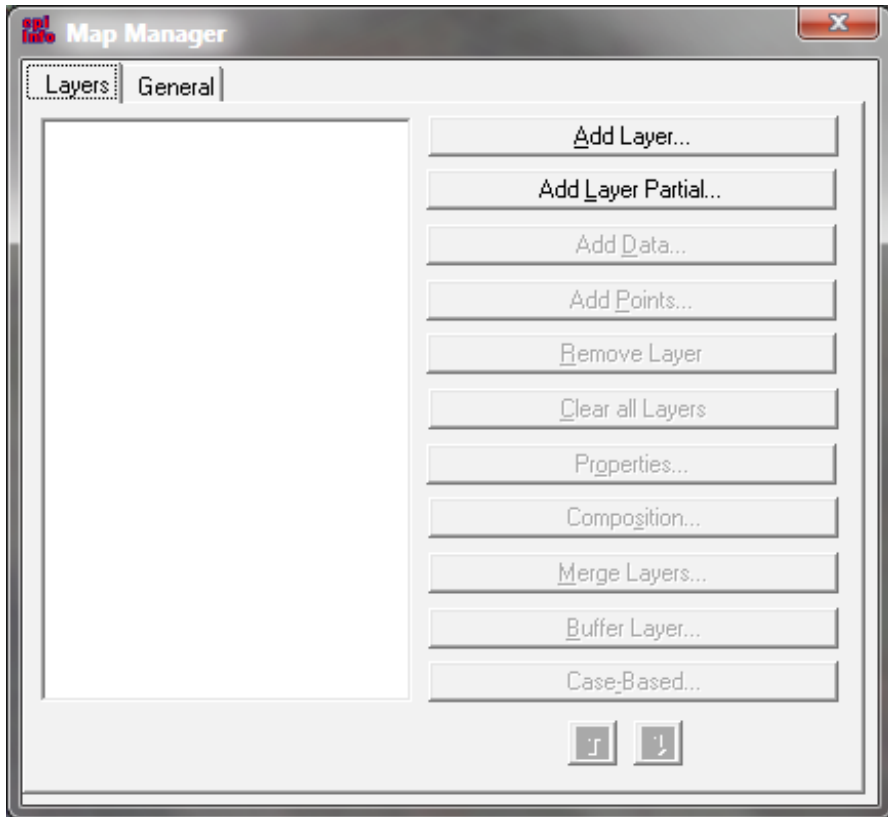


Figura 2.3. Janela inicial do *Map Manager*.

Fonte: Epi Map.

O Gerenciador de Mapas (*Map Manager*) é a ferramenta mais importante do programa Epi Map. Na janela inicial, figura 2.3, aparecem somente duas possíveis opções: *Add Layer* e *Add Layer Partial*. *Layer* (folha, ou camada) será o nosso mapa, ou seja, o arquivo *.shp*. Automaticamente, o programa procura por um arquivo *.shp* no diretório do Epi Info®, assim devemos abrir o folder *Mapas*, pois lá salvamos o *Br.shp*. Selecionamos o arquivo e clicamos *Open* (abrir). Surgirá, assim, um mapa do Brasil todo em branco e, na janela do *Map Manager*, outras opções tornar-se-ão possíveis (figura 2.4):

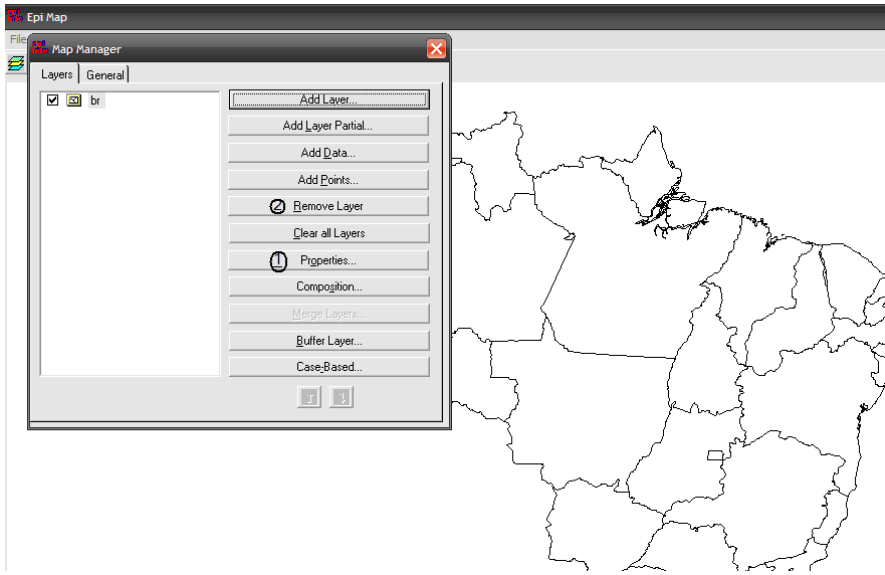


Figura 2.4. Tela do *Map Manager*, destacando as opções *Properties* e *Remove Layer* na janela do *Map Manager*.  
Fonte: Epi Map.

A opção mais importante é a das propriedades (*Properties... (I)*). Imediatamente observaremos algumas funções dessa opção, clicando na barra correspondente. Aparecerão seis opções de apresentação gráfica (figura 2.5).

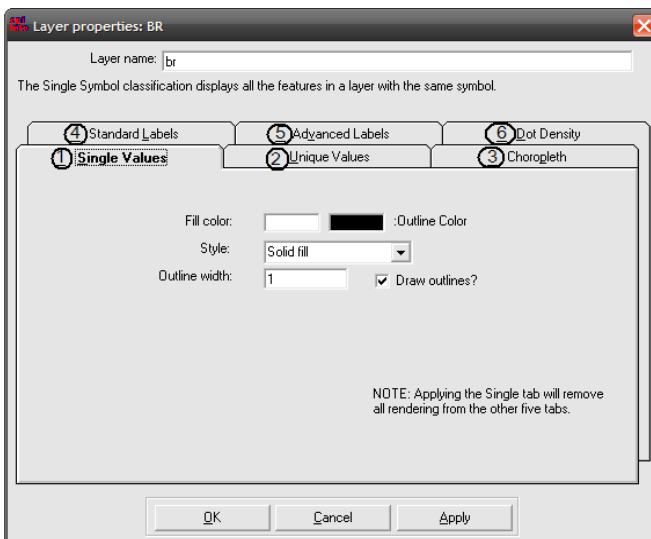


Figura 2.5. Janela de opções do *Properties ...* do *Map Manager*.  
Fonte: Epi Map.

Cada opção de gráfico terá uma explicação sucintamente descrita logo abaixo da janela do nome da folha (*Layer name*:), que, por sinal, pode ser mudado, mas não afetará o nome no *Gerenciador de Mapas (Map Manager)*.

**1.** *Single*, estabelece um único padrão para todas as características da folha ou mapa.

**2.** *Unique*, coloca todas as diferentes características ou informações contidas no banco de dados do mapa em cores distintas.

**3.** *Choropleth*, estabelece intervalos dos valores dos dados que desejamos mostrar. Vamos ver como essa opção funciona. O banco de dados original do mapa oferece a população de cada Estado (*POP\_ADMIN*), a área do Estado em quilômetros quadrados (*SQKM\_ADMIN*) ou em milhas quadradas (*SQMI\_ADMIN*). Essa opção é escolhida no *Numeric Field (1)*.

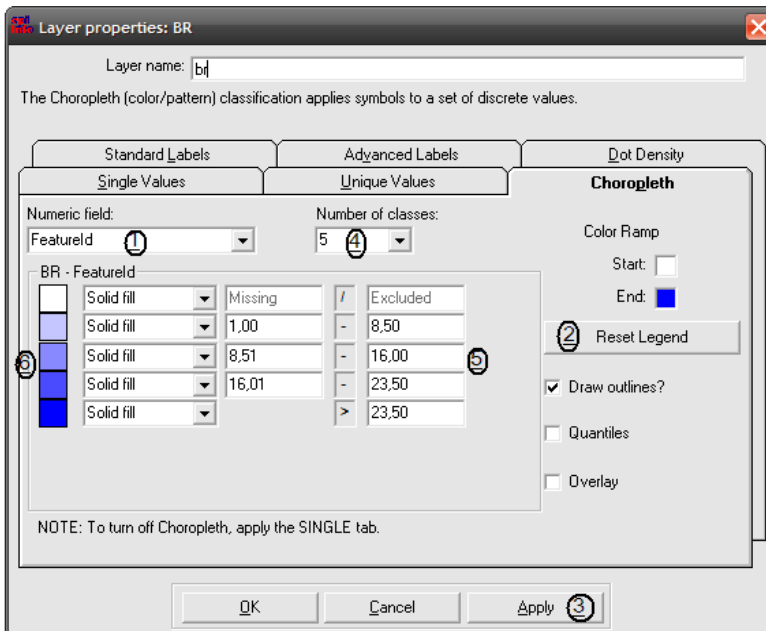


Figura 2.6. Janela da opção *Choropleth*, propriedades de tela do *Map Manager*.  
Fonte: Epi Map.

Toda vez que mudarmos alguma coisa na opção, ela só será efetiva se clicarmos em *Reset Legend* (2) e poderemos visualizar no mapa clicando em *Apply* (3) que só fica acessível quando clicamos no *Reset Legend* primeiro. Podemos modificar o número de classes na opção (4). Observamos que os intervalos que automaticamente aparecem não são inteiros. Podemos modificá-los colocando o número inteiro na coluna da esquerda. Observamos que, os primeiros valores são fixos para valores ausentes ou excluídos. Podemos modificar somente os valores a partir da segunda linha. Primeiro colocamos zero na primeira casa acessível (a segunda de cima para baixo à esquerda). Às vezes necessitamos colocar esse valor e clicar no *Reset Legend* duas vezes, por isso não se preocupe se o valor voltar ao original na primeira vez. Após isso, poderemos arredondar os outros intervalos. Mantendo o mesmo número de classes, 5, podemos ver que os intervalos poderiam ser arredondados para valores a cada 8 000 000. Assim, colocaríamos esse valor na terceira casa da esquerda e clicamos no *Reset Legend*. Observamos que a segunda casa da direita torna-se 7 999 999.99. A quarta casa da esquerda poderia ser, então, 16 000 000. Não temos mais casa na esquerda, então colocamos o próximo valor na quinta casa da direita: 23 999 999.99. Novamente, para observarmos o resultado, clicamos no *Reset Legend* e depois no *Apply*. Podemos modificar as cores de cada uma das classes conforme nosso “bel-prazer”. Para tanto basta clicarmos duas vezes no quadrado colorido ao lado de cada classe (6). Aparecerá uma janela com várias cores e tonalidades para escolhermos. Nesse caso específico não necessitamos clicar no *Reset Legend* antes de clicar no *Apply*, basta clicar *OK* na janela das cores. Para fechar a janela basta clicar *OK*, mas continuaremos a observar as outras opções de mapas.

**4.** Outras opções do *Gerenciador de Mapas* são as *Std Labels* (Etiquetas padrão) que coloca os valores da característica que queremos, entre elas podemos colocar o nome dos Estados. Observe que junto a essa opção aparecem algumas zonas que o banco de dados chama de zonas de litígio (entre o Ceará e o Piauí e o Amazonas e o Pará). Nessa opção poderemos também colocar os valores numéricos de cada Estado, no caso a população total ou a área. A próxima opção (*Adv Labels*) é muito parecida, mas só permite colocar as variáveis não numéricas.

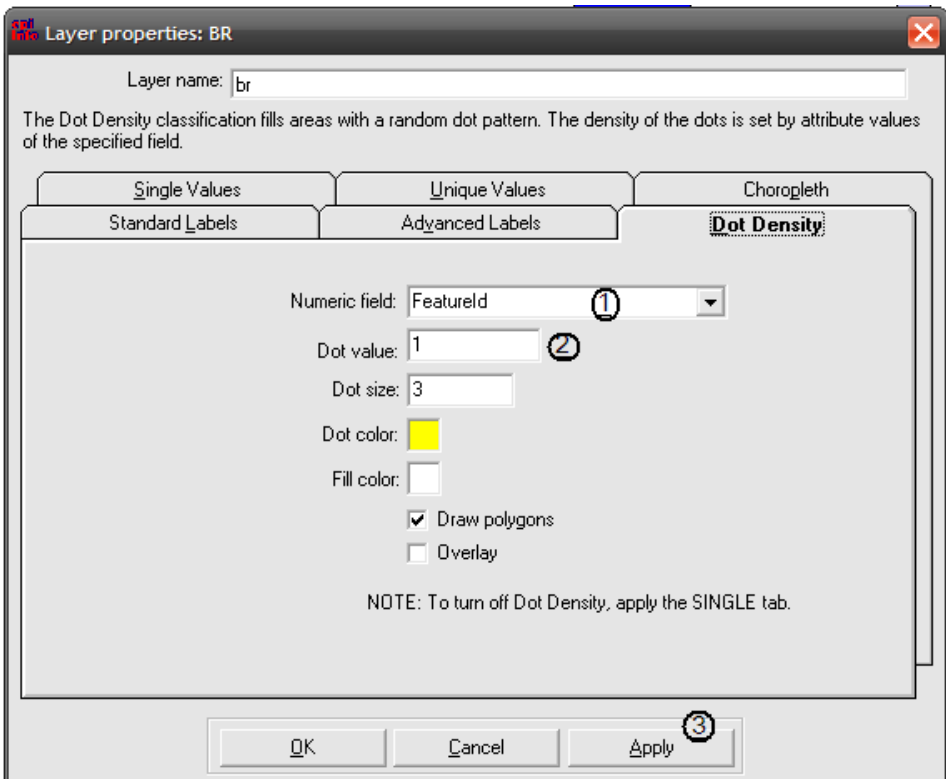


Figura 2.7. Janela da opção *Dot Density*, propriedades de tela do *Map Manager*.

Fonte: Epi Map.

A última opção (figura 2.7), *Dot Density*, é muito perigosa, pois se não modificarmos nada e, sem querer, clicamos em *Apply*, o programa coloca um ponto para cada valor da variável, isso demora “séculos”, principalmente se selecionamos a característica População (*Pop\_Admin*). Como ilustração dessa opção, podemos selecionar essa mesma característica anterior: (1) população. Depois modificamos o valor de cada ponto (*Dot Value* (2)) para 1 000 000, conforme a figura 2.7. Podemos, assim, clicar em *Apply* (3) e observarmos o resultado. Aparecem pontos correspondentes a cada milhão de habitantes; alguns Estados possuem menos de um milhão de habitantes, por isso não recebem nenhum ponto.

Novamente selecionamos o *Choropleth*, aplicamos, para então fecharmos a janela das propriedades do gráfico e fechamos o *Gerenciador de Mapas*. Assim, veremos mais algumas opções interessantes sobre esse programa. A legenda das classes da população aparece no canto esquerdo, mas podem ser arrastadas com nosso mouse para mais perto do gráfico.

## Opções do Epi Map

Clicando no botão da direita do mouse, em qualquer lugar no mapa, ou no menu *view*, obtemos uma janela com algumas opções do Epi Map, conforme a figura 2.8.

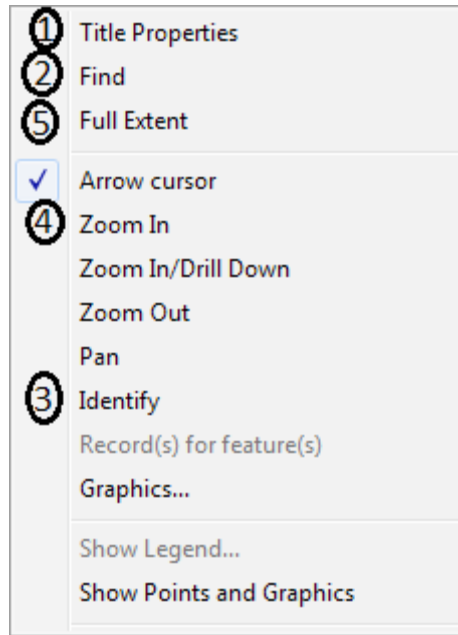



Figura 2.8. Janela de opções do Epi Map.  
Fonte: Epi Map.

Podemos definir as propriedades do título para o gráfico assim como criá-lo (*Title Properties (1)*), mas essa opção cria um título na parte superior ou inferior do gráfico e é difícil movê-lo. Como veremos mais adiante, no capítulo 7, é muito inconveniente criarmos um título em um gráfico, pois, após salvar o mesmo, não poderemos editá-lo ajustando para as exigências do local onde queremos publicá-lo ou mostrá-lo. Podemos localizar determinado local no mapa ou mostrar os dados de um local utilizando respectivamente os comandos *Find (2)* e *Identify (3)*. É possível ainda esconder a legenda das classes (*Legend*) e ampliar uma parte do mapa através do *Zoom in*. Poderemos, por exemplo, ampliar algumas ilhas no mapa à direita. Para voltar ao tamanho máximo, clicamos em *Full Extent*. Outra opção interessante é o *Identity*, que nos mostra informações sobre cada Estado que clicarmos. Para adicionarmos um texto e também um título

no lugar que quisermos, clicamos na opção *Graphics* ou no ícone que tem um triângulo, um quadrado e um círculo (📐) no menu do programa. Esse comando gera outros ícones: , o primeiro adiciona um texto onde clicarmos. Podemos, com essa opção, adicionar nomes ou valores ao gráfico. Finalmente podemos salvar nosso mapa como um arquivo tipo mapa, para editá-lo mais tarde, ou como um arquivo bitmap (.bmp) (menu *Arquivo, Salvar como BMP*) que pode ser adicionado a um texto. Infelizmente, como esse arquivo é muito grande, podemos transformá-lo em um formato mais compactado como .jpg, usando um editor gráfico ou o próprio *Paintbrush* do *Windows*, dependendo da configuração do computador. Usaremos o Epi Map de novo, mais adiante, mostrando como adicionar novos dados ao mapa.

## Como conseguir mapas mais detalhados do Brasil

O *IBGE* (Instituto Brasileiro de Geografia e Estatística) mantém um site que disponibiliza vários tipos de gráficos brasileiros. Nesse site poderemos baixar mapas, por exemplo, de qualquer Unidade da Federação por município. O endereço do site é <http://mapas.ibge.gov.br/divisao/viewer.htm>. Note que não temos o “www” no endereço. Caso o link não funcione, procure o site [www.pucrs.br/igg/epiinfo](http://www.pucrs.br/igg/epiinfo), folder atualizações para obter a versão atualizada do link e novas instruções a respeito do site.

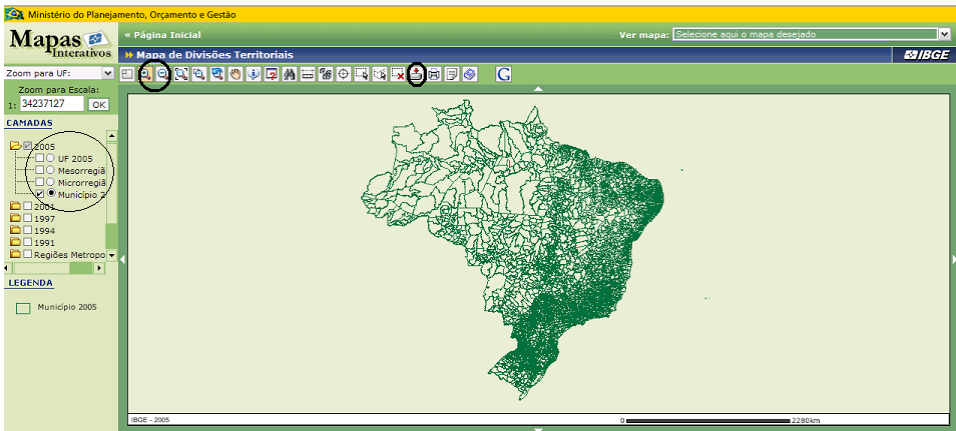


Figura 2.9. Tela da página de geração de mapas do IBGE, destacando as camadas dos mapas, ícones de ampliação, redução e de importação de mapas.

Fonte: <http://mapas.ibge.gov.br/divisao/viewer.htm>, último acesso em 4 de abril de 2011.

Automaticamente, depois de algum tempo, abre-se uma tela com o mapa do Brasil com as regiões delimitadas, conforme figura 2.9. Na



janela, à esquerda, temos as camadas onde podemos modificar o que queremos visualizar (no círculo) ou selecionar (no quadrado) no mapa. Vemos que existem as opções de visualização por Unidade da Federação, Mesorregião, Microrregião e finalmente, Município. Mais do que uma opção de visualização é possível, até mesmo todas. Somente uma camada pode ser ativa. Uma camada ativa (segunda coluna, círculo) significa que podemos procurar no mapa a localização de uma cidade, por exemplo. Veremos mais adiante a sua utilidade.

Na parte superior do mapa vemos alguns ícones. Veremos a função deles. O primeiro da esquerda faz aparecer ou desaparecer uma versão compacta do mapa no canto superior esquerdo. Os dois próximos ícones da esquerda para a direita aumentam ou diminuem o mapa. Selecionando o ícone da lupa com o sinal de mais, aumentamos o tamanho do local que circularmos com o mouse. Como exercício, selecionaremos a região Sul. Depois do ícone da lupa com sinal menos, vemos, na sequência, os ícones que retornam à extensão total, extensão prévia e movimentação do mapa com o mouse (ícone da mão). Os três próximos itens referem-se ao conteúdo do mapa. O ícone do “i” dá informações ao local no mapa que clicarmos. As informações são basicamente a localização, quanto a região, unidade da federação, mesorregião, microrregião, município e área, dependendo do nível que estamos visualizando. Assim, se clicarmos na Região Sul, só teremos como informação a região e a área. Podemos utilizar o ícone do binóculo para localizarmos um município. Mas, para localizar uma cidade, por exemplo, a cidade de Catuípe, devemos ter a camada Município ativa e visível. Feito isso, clicamos no ícone do binóculo que abre uma caixa, onde colocamos o nome da cidade: Catuípe (com acento e primeira letra maiúscula). Ao clicarmos o botão *Localizar*, temos a cidade em amarelo, no meio do mapa. Se quisermos informações sobre a cidade, devemos clicar no próximo ícone, que parece uma planilha com um ponto de interrogação. A janela é um pouco mais complicada, por isso ela é apresentada na figura 2.10.

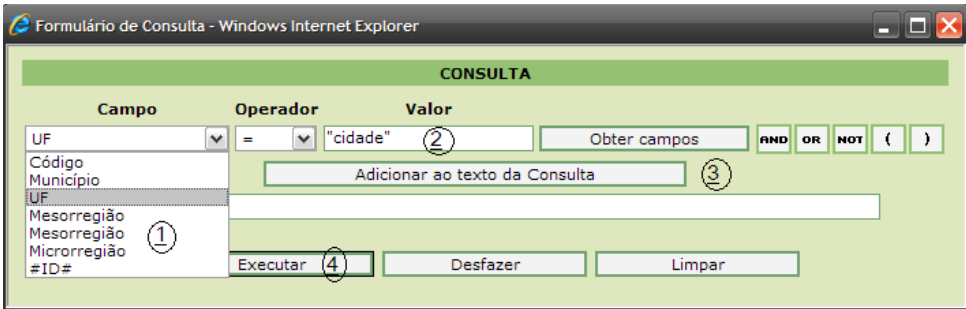


Figura 2.10. Janela de consulta de dados do programa de geração de mapas do IBGE.

Fonte: <http://map.ibge.gov.br/divisao/viewer.htm>, último acesso em 04 de abril de 2011.

No primeiro compartimento (*Campo (1)*) devemos escolher o que desejamos, no caso: Município. O *Operador* serve para escolhermos locais com valores acima ou abaixo de determinado nível que determinamos no compartimento *Valor (2)*. No caso, colocaremos de novo a palavra Catuípe. Temos que adicionar esse valor na janela de busca clicando na barra correspondente (3). Finalmente podemos clicar no botão *Executar (4)*, para observarmos as informações em uma nova janela. Deverão aparecer informações sobre o município, localização quanto a região, unidade da federação, mesorregião, microrregião e, finalmente, a área do município. Podemos selecionar uma região pela localização, clicando no ícone do quadrado tracejado, que é o próximo ícone (*Seleção por retângulo*). O ícone à direita da *Seleção por retângulo* é o chamado *Buffer*, que seleciona uma zona ao redor da localização previamente selecionada, a uma atribuída distância, no caso, poderíamos selecionar as cidades que ficam 10 quilômetros ao redor de Catuípe. Para tanto, selecionamos a opção município na janela *Destacar as Feições de (1)*, colocamos o valor na janela correspondente (2) e, finalmente, clicamos em *Criar Buffer (3)*, conforme a figura 2.11.

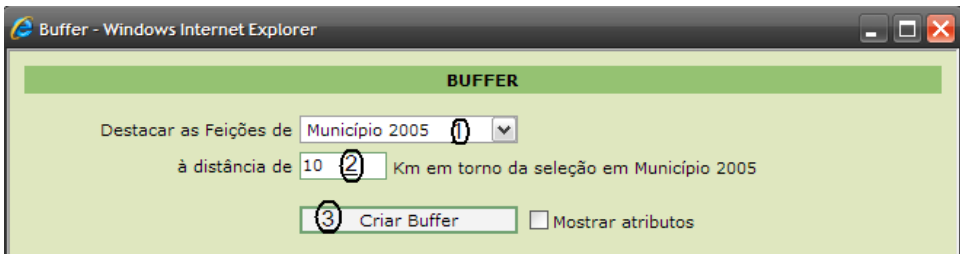


Figura 2.11. Janela de criação de *Buffer* do programa de geração de mapas do IBGE.

Fonte: <http://map.ibge.gov.br/divisao/viewer.htm>, último acesso em 4 de abril de 2011.

O próximo ícone tem um formato que deveria ser o de uma borracha justamente para apagarmos a seleção. Vamos clicar nesse ícone para limparmos as seleções realizadas até aqui. Como temos duas seleções, município e entorno, temos que clicar duas vezes no ícone. Os próximos ícones são para medir distâncias (régua) e para “alternar lista de camadas/legendas”. Passando pelo ícone de imprimir, finalmente chegamos ao ícone que salva o mapa (📄, *Download*). Como exercício, vamos visualizar somente as mesorregiões com o mapa mostrando todo o Rio Grande do Sul. Evidentemente que acabaremos visualizando mais do que o Estado, mas iremos corrigir isso, posteriormente. Clicando no ícone do *Download*, abre uma janela informando a área selecionada (longitudes e latitudes superiores e inferiores) e um botão com o comando *Extrair*. Nova janela se abre informando que o arquivo foi criado e indicando o *Link* para baixar o arquivo (*Download*). Ao clicarmos nesse *Link*, surge a janela de salvar ou abrir o arquivo; clicamos em abrir. O comando abre o arquivo com o *WinZip* e observamos que existem três arquivos: *Meso.shp*, *Meso.dbf* e *Meso.shx*. Vamos extrair todos os arquivos no diretório: *C:/Epi\_Info/Mapas*, criando um novo subdiretório *RS*, pois os nomes dos arquivos extraídos são muito genéricos.

Quando importamos arquivos desse site, algumas vezes temos que corrigir o nome de alguma localidade. Por exemplo, nomes com acento podem ser importados errados. Por isso devemos observar se os nomes das localidades estão corretos. Esses nomes estão no arquivo com extensão *.dbf*.

### **Editando arquivo .dbf**

Os arquivos com a extensão *.dbf* são arquivos em formato *dBase*. O site do *IBGE* cria os arquivos na versão *IV* do programa. Podemos observar e editar esse tipo de arquivo usando o *Excel*. Com o *Excel* aberto, clicamos em *Abrir (Open)*. Devemos modificar o tipo de arquivo para arquivo do *dBase*, no compartimento inferior da janela de abertura, conforme a figura 2.12.

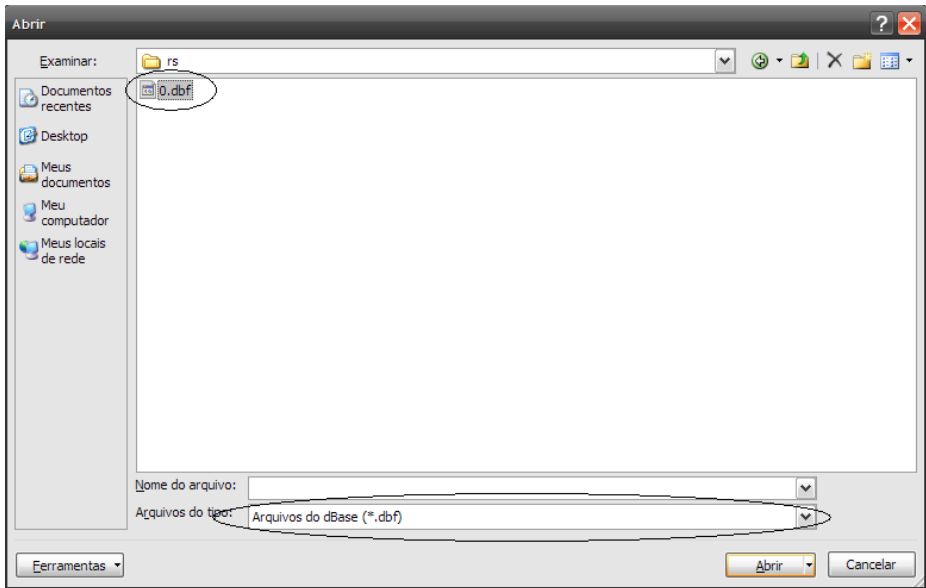


Figura 2.12. Janela de seleção de arquivo a abrir pelo *Excel*, destacando o nome do arquivo e o tipo de arquivo. Fonte: Microsoft Excel.

Selecionamos o arquivo *meso.dbf* localizado no diretório que o havíamos salvo. Clicando no botão *Abrir (open)*, visualizamos os dados. Observamos que foram salvas informações a mais do que queríamos, pois temos mesorregiões inclusive de outros Estados. Não podemos apagar esses dados extras, pois eles estão relacionados a outros arquivos. Podemos modificar e corrigir alguns nomes que estejam mal escritos. Se isso ocorrer, devemos salvar o arquivo com o formato e nomes originais (*dBase IV* e *meso*), mesmo que o programa recomende que não. Depois de salvo o arquivo em formato *dBase*, quando desejamos sair do *Excel*, uma mensagem indicando que o arquivo não foi salvo aparece. Devemos ignorar essa mensagem, pois está informando que o arquivo não foi salvo em formato *Excel*. O Epi Map não trabalha com arquivos em formato *Excel*, por isso devemos manter o formato original. Como os outros arquivos são chamados *meso*, devemos manter o mesmo nome, ou mudar todos os nomes.

### Selecionando regiões em um arquivo .shp (Add Layer Partial)

Como vimos, o arquivo *meso.shp* que importamos do IBGE possui regiões que não pertencem ao Rio Grande do Sul. Como desejamos ter um mapa só com essas regiões, podemos utilizar o comando *Adicionar camada parcial*

(*Add layer partial*) do Epi Map. Após abrir o programa, escolhemos o arquivo *meso.shp*. Modificando um pouco a largura das colunas obteremos a janela conforme a figura 2.13.

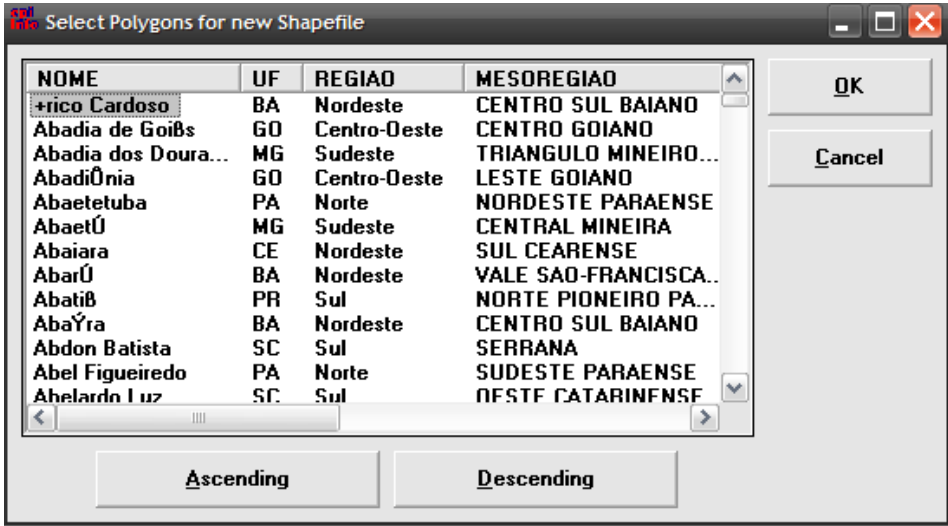


Figura 2.13. Janela de seleção de componentes de um mapa pelo Epi Map.  
Fonte: Epi Map.

Se clicarmos na barra do nome da Unidade da Federação (*NOMEUF*), a mesma passa a ser a primeira da lista e passa a ficar em ordem alfabética. Prestando um pouco de atenção, notamos que as primeiras localidades do Rio Grande do Sul são: Lagoa Mirim e Lagoa dos Patos. Vamos selecionar todas as localidades do Rio Grande do Sul, exceto as duas lagoas. Após selecionar as sete mesorregiões do Estado, clicamos *OK*. Aparecerá uma janela para salvarmos os arquivos com um nome que seja fácil lembrar, como *Mesorregião do RS*. Finalmente podemos salvar o mapa como está, utilizando o mesmo nome do arquivo *.shp*.

## Exercício:

1. Baixe o mapa contendo as Regiões do Brasil, no site do *IBGE*, salvando como *Regiões do BR* no diretório *c:/Epi\_Info/Mapas/BR* (extraia os arquivos).
2. Faça um mapa *Choropleth* das áreas das regiões, arredonde as faixas e salve como *Regiões BR.map*.
3. Edite o arquivo *Regiões BR.dbf* usando o *Excel*, observe como os nomes das regiões estão escritos (principalmente observe se estão com os nomes completos, com ou sem acento etc.). Não importa se os nomes estejam em maiúsculas ou minúsculas. Como não necessita nenhuma alteração, simplesmente feche o arquivo.

## Capítulo 3

# CRIANDO UM BANCO DE DADOS

### Estrutura do arquivo Epi Info®

Os arquivos do Epi Info® são chamados de *Projetos (Projects)* e têm o formato do *MS-ACCESS*. O arquivo do *MS-ACCESS* é um banco de dados, por isso tem a extensão *.mdb* (Microsoft DataBase). Esse tipo de arquivo pode conter muitos “subarquivos”, assim como os arquivos *EXCEL*, que podem ter muitas planilhas. O Epi Info® salva dois diferentes principais “subarquivos”: *Tabela (Table)* e *Ficha (View)*. Também há outros arquivos diferentes, onde são salvos códigos e “links”, mas aqueles dois quase sempre estão presentes. *Tabela (Table)* é o real banco de dados e contém todos os dados. Esse arquivo é essencial. O arquivo *ficha* ou em inglês *View* corresponde ao “arquivo formulário” do *MS-ACCESS*, mas tem uma estrutura diferente. O subarquivo *ficha* salva os códigos e a tela de introdução de dados, bem como descrição das variáveis.

### Criando um projeto e uma ficha (Make View) no Epi Info®

A primeira coisa que temos que fazer para introduzir dados a partir do nada para o Epi Info® é construir uma *View (Ficha)*. O programa para isso é o *Make View (Criar Ficha)*. Antes, porém, devemos verificar se a linguagem selecionada do programa é o inglês, pois o português nos traz alguns problemas. Para tanto, basta olhar no canto superior esquerdo da tela principal do Epi Info® e observaremos a linguagem. Vimos, no primeiro capítulo, como modificar a linguagem do Epi Info®.

Na tela do Epi Info® clicamos em *Make View (Criar Ficha)*, para iniciar a criação da nossa ficha.

O *Make View* tem duas janelas: a janela da direita que é a janela do registro e a janela da esquerda a que guia as diferentes páginas do questionário.

Primeiro temos que abrir ou criar um projeto (arquivo de banco de dados) que o Epi Info® chama de *Project*. No nosso caso, criaremos um arquivo novo: Menu *File (Arquivo)*, clicar em *New (novo)* de acordo com a figura 3.1.

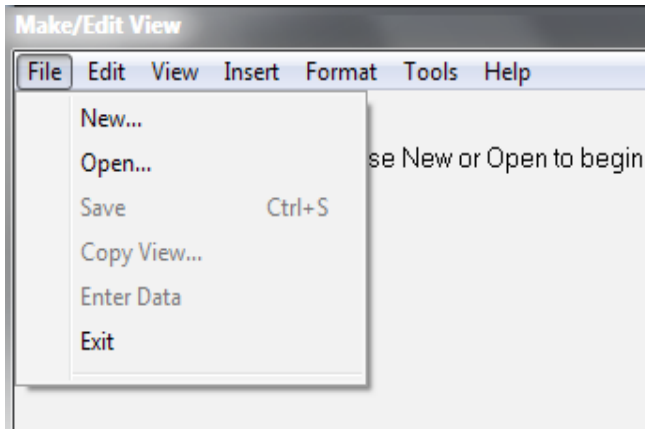


Figura 3.1. Janela de opções do menu *File* do *Make View*.  
Fonte: Epi Info®.

Observe que o último arquivo usado fica registrado na parte baixa do comando. Uma janela nova, muito comum do *Windows*, aparecerá: a janela de abrir ou salvar. Vamos salvar o arquivo como: “*Seunome*”. Isso cria um arquivo em branco. Antes de salvar, vamos criar um novo diretório, para armazenar nossos arquivos, dentro do *C:\Epi\_info*, chamaremos de *EpiAula*. Como vimos, o banco de dados de Epi Info® tem os *Views (Fichas)*. Epi Info® pede um nome agora para a nova ficha. Podemos chamar “*Dados*” e clicar *OK*. A tela muda e uma grade aparece. Automaticamente, o programa estabelece que os campos serão alinhados à grade. Podemos mudar essa regra no Menu *Format (Formato)*, *Settings (configurações)*. Também podemos mudar algumas outras configurações, como aumentar ou diminuir o espaço entre as linhas da grade.

### **Criando campos (variáveis)**

Para criar um campo novo (*Field*) devemos clicar no botão da direita do mouse no lugar onde queremos que ele fique. Isso nos leva a outra janela (*Field Definition*) cujos componentes serão descritos a seguir e observados na figura 3.2.



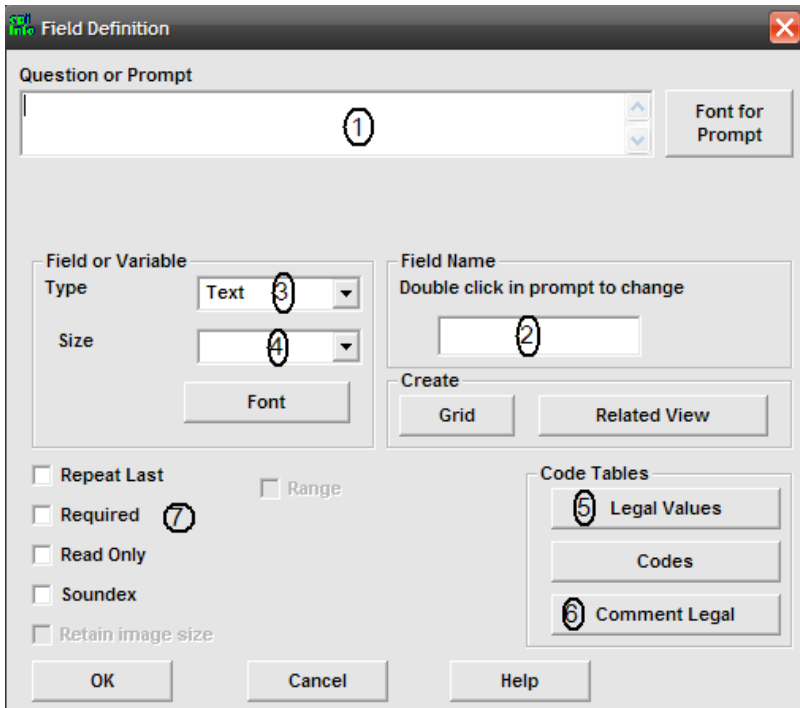


Figura 3.2. Janela *Field Definition* do *Make View*.  
Fonte: Epi Info®.

**1.** *Question or Prompt (Pergunta)*: é o primeiro componente da janela e está localizado na parte superior da mesma. É suposto que descrevamos a pergunta relacionada à variável que estamos criando neste componente. Isso pode ser bastante longo, com vários caracteres. Essa pergunta aparecerá em alguns resultados quando analisarmos os dados.

**2.** *Field Name (Nome do Campo)*: automaticamente o programa sugestiona um nome do campo. O nome do novo campo será criado conforme o *Prompt*: utilizando as primeiras palavras da pergunta. Podemos mudar o nome do campo usando um mais reconhecível ou mais conveniente para nós. Esse nome se tornará o nome da variável, por isso é importante observarmos como ele é denominado. O nome da variável pode ser longo (mais de 40 caracteres), aceita acentos, mas não aceita caracteres especiais nem espaços, quando o definimos usando o *Make View*. O nome da

variável no Epi Info® não pode iniciar com números, podendo ter tantos números quanto quisermos, a partir do segundo caractere. É aconselhável que o campo não tenha mais que 8 caracteres, pois alguns programas (como o SPSS) não aceitam mais do que esse número de caracteres. Construindo nossas variáveis somente com 8 caracteres não teremos dificuldades ao importar o banco de dados em outro programa. Também não é aconselhável usarmos acentos nos nomes das variáveis justamente pelo mesmo motivo.

**3.** *Field or Variable Type*: existem diferentes tipos de variáveis em Epi Info®. Observaremos esses tipos logo após terminarmos de observar mais algumas opções da tela de definição de campo (*Field Definition*).

**4.** *Field or Variable Size*: define o tamanho do campo ou variável, veremos sua utilidade para cada um dos tipos de campo.

**5.** *Legal Values (Valores Legais)*: codifica valores válidos como resposta às variáveis do tipo texto.

**6.** *Comment Legal (Comentários Legais)*: é parecido com os valores legais, mas propicia a colocação de um comentário ao lado do valor, somente o valor é armazenado na tabela de dados.

**7.** Outras opções: *Repeat Last (Repetir o último)* repete o último valor colocado; é interessante quando temos uma variável constante para grupos de amostras (por exemplo, pessoas avaliadas em um mesmo dia). *Required (Necessário)* não deixa seguir a introdução de dados se nenhum valor for colocado, muitas vezes teremos problemas ao colocar esse tipo de opção. *Read Only (Só leitura)*, não permite que anotemos o valor da variável, é importante quando criamos automaticamente um valor para ela e *Range (Faixa)* estabelece uma faixa de valores válidos para variáveis numéricas.

Como um exemplo, construiremos um questionário que pergunta algumas informações sobre nós mesmos, conforme as orientações da tabela 3.1.

Como queremos juntar nossos dados posteriormente, é importante que o nome dos campos (*Field Name*) e os seus tamanhos sejam exatamente os

mesmos para todos. Assim, é importante seguir as instruções da tabela 3.1, principalmente a da sugestão do nome do campo.

Tabela 3.1. Lista de variáveis do questionário a ser utilizado como exemplo durante a explicação dos diversos tipos de variáveis do *Make View*.

Grupo	Pergunta (Prompt)	Nome do campo (Variável)	Tipo de variável	Valor Legal/ Formato/Faixa
	Questionário	QUESTIONARIO	Text/ Label	
Identificação	Registro	REGISTRO	Número	###
	Data de Nascimento	DATANASC	Data	DD/MM/YYYY
	Data do Exame	DATAEXAME	Data	DD/MM/YYYY
	Idade (anos)	IDADE	Número	### (0-130)
	Gênero	GENERO	Texto	Masculino Feminino
	Cor (raça)	COR	Texto	Branca Negra Mulata
	Estado civil	ESTADOCIV	Texto	Solteiro Casado Viúvo Separado
Dados clínicos	Altura (cm)	ALTURA	Número	### (70-250)
	Peso atual (kg)	PESOATUAL	Número	###.# (20-200)
	Peso aos 20 a (kg)	PESO20A	Número	###.# (20-200)
	Tabagismo	TABAGISMO	Opção	Atual – fuma atualmente Exfumante – fumou por mais de um ano Nunca – todos os outros
	Atividade Física Regular	ATIVIDADE	Check Box	Yes / No

Fonte: criação do autor.

## Field or Variable Type (tipos de campos ou variáveis)

Seguindo a sequência do exemplo, temos os seguintes tipos de campo:

**1.** *Label/Title*: de fato não é uma variável, porque só insere um texto ou comentário. Todavia, pode ser utilizado para gerar informações importantes para o responsável pelo preenchimento. No caso usaremos para gerar um título que será o título do nosso questionário: colocaremos como *Question or Prompt (1)* a palavra “Questionário”, no nome do Campo (3) “QUEST”, e como Tipo: *Label/Title*. Podemos modificar a fonte do Prompt clicando no botão à direita da janela (1).

**2.** *Number* define uma variável numérica. O Epi Info® trabalha as variáveis numéricas com símbolos ###. ## (sustenidos com marcador decimal ponto, mesmo como no Epi Info® 6.0). O Epi Info® oferece um número restrito de opções de formatos no compartimento *Size (4)*. Mas podemos, e devemos, definir o tamanho da variável utilizando o símbolo “#” tantas vezes quantas necessárias para configurarmos o número de dígitos que desejamos para a variável.

A segunda variável do nosso questionário é numérica: registro, no caso, chamada de “REGISTRO”. É importante termos uma variável que identifique cada um dos nossos registros para podemos revisá-los posteriormente ou usar como chave de ligação com outro banco de dados. O registro deve ser único, infelizmente o Epi Info® não evita que coloquemos dois registros iguais, por isso temos que tomar cuidado.

**3.** *Date*: insere datas; esse tipo de variável oferece diferentes padrões de opções que temos que escolher de acordo com a necessidade. Podemos também digitar o formato das datas manualmente ou o escolher no compartimento *Size (4)*. No nosso caso, usaremos esses tipos para definir as duas variáveis seguintes: Data de nascimento, chamada “DATANASC”, e Data do Exame, chamada “DATAEXAME” e será para nós a data de hoje com o formato DD/MM/YYYY.

**4.** Variável IDADE: Sempre que construimos um banco de dados precisamos pensar no tamanho máximo de cada valor

das variáveis. Esse é o caso, em nosso exemplo, da variável IDADE. A maioria das pessoas tem menos de 100 anos, mas, até mesmo quando pensamos que ninguém ultrapassará aquele nível, deveremos definir o tamanho da variável com três dígitos, exceto, evidentemente, se tivermos trabalhando somente com crianças. Podemos definir uma faixa de idade para evitar que alguém, menos avisado, entre com dados fora de um limite razoável, assim, clicamos em *Range (Faixa)* e definimos como mínimo 0 (zero) e máximo 130 (desconfie da veracidade de alguém com mais de 130 anos!). Isso vai depender da nossa amostra. Se estivermos trabalhando com uma população pediátrica, deveremos adequar a faixa a essa clientela. Observe que no *Prompt* podemos colocar a unidade de sua medida, no caso a idade estará sendo medida em anos (poderia ser em dias, se a pesquisa fosse com recém-nascidos).

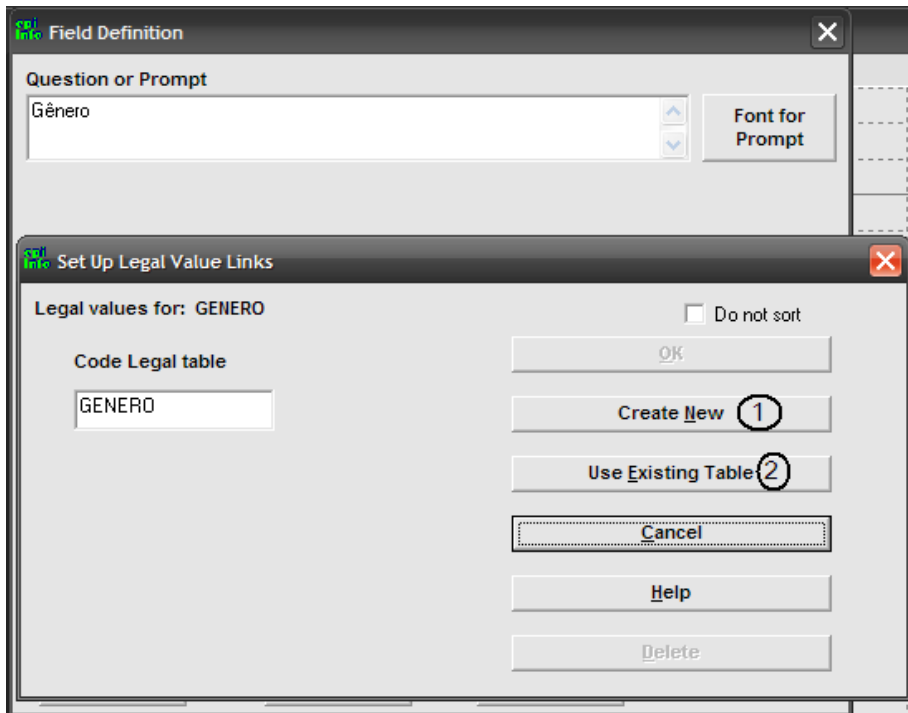


Figura 3.3. Janela de criação de *Legal Values* do *Field Definition*.  
Fonte: Epi Info®.

**5.** *Text* (texto) com *Legal Values* (*Valores Legais*): nossas três próximas variáveis serão variáveis tipo texto. *Text*: é o tipo que automaticamente aparece toda vez que desejamos criar um campo, cria variáveis que aceitam textos e vários tipos de caracteres. Não precisamos decidir o número máximo de caracteres especificando seu tamanho (4), veremos o porquê. O tipo de variável *TEXT* aceita *Legal Values* (5) ou, observada na primeira figura desta sessão. “*Legal Values*” são valores preestabelecidos que a variável pode ter. Essa característica previne a introdução de dados com textos mal escritos, incompletos ou semelhantes, mas diferentes do que queremos. Também nos ajuda a introduzir os dados, como veremos mais adiante. No caso do exercício, a variável *GENERO* possuirá dois valores legais: Masculino ou Feminino. Para criar os valores legais devemos que clicar em *Legal Values* (5) da janela do *Field Definition*, que nos levará a uma nova janela (figura 3.2) que sugere o nome da variável como um código novo, mas oferece a possibilidade de usarmos um código de uma tabela (*Table*) existente (2). Ao clicarmos em *CreateNew* (1), uma nova janela (figura 3.4) aparecerá para colocamos, então, os dois valores: Masculino e Feminino, pulando de uma linha para a outra com o *Enter*. Automaticamente, os valores serão dispostos em ordem alfabética, podemos evitar isso se checarmos a opção logo acima do *OK* (*Do not sort*), conforme a figura. (Esse comando não afetará a forma como serão mostrados os resultados da análise.) Faremos o mesmo para as outras duas variáveis: cor (*COR*) e estado civil (*ESTADOCIV*). É salutar preservar as maiúsculas.

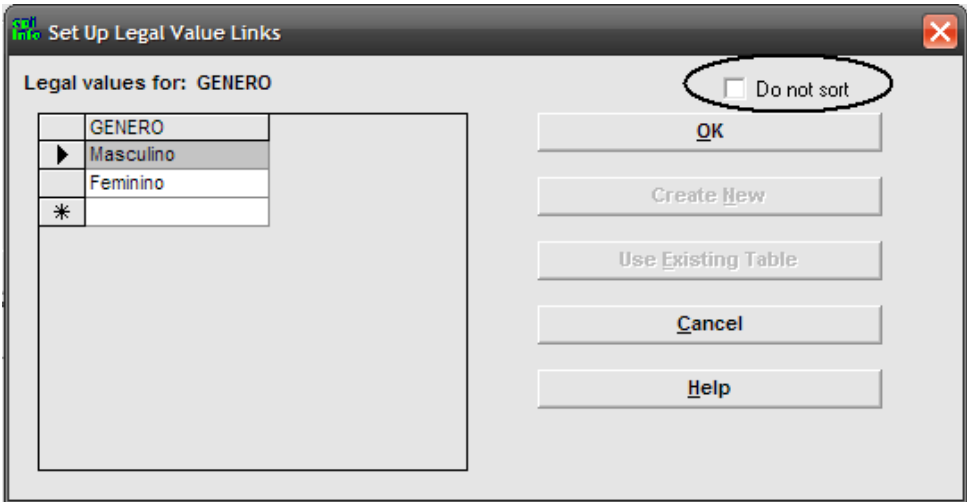


Figura 3.4. Janela de definição dos valores no *Legal Values* do *Field Definition*, destacando a opção não ordenar (*Do not sort*).

Fonte: Epi Info®.

**6.** Variáveis ALTURA, PESOATUAL, PESO20: serão definidas como variáveis numéricas (*Number*) e definir as faixas conforme orientado.

**7.** *Text com Comment Legal (Comentário legal)*: os valores das variáveis tipo texto ou categóricas devem ser sucintos, com poucas letras e de preferência com palavras únicas. Entretanto, às vezes necessitamos fornecer uma descrição maior da definição do valor. Esse pode ser o caso da variável TABAGISMO, a pessoa que introduzir os dados pode ficar confusa principalmente entre ex-fumante e nunca fumante. Assim usamos o artifício (ou artimanha) do *Comment Legal* que aceita colocarmos uma descrição junto ao valor da variável. Praticamente é o mesmo procedimento do *Legal Value*, a única diferença é que, depois do valor, devemos colocar um hífen ( – ) e então a descrição que desejamos. Tanto o valor quanto o comentário aparecerão quando entrarmos com os dados, mas somente o valor será colocado na tabela de dados.

**8. Check Box:** insere uma caixa de checagem pequena, trabalha como uma variável de Yes/No; usaremos esse tipo para definir a variável ATIVIDADE. No caso, os valores omissos serão sempre considerados como resposta negativa, o que nem sempre é interessante.

**9. Option:** cria uma caixa com as opções desejadas. Tem a mesma função do tipo texto com Legal Values, mas os mesmos aparecem mais claramente. Seguindo os passos da figura 3.5, selecionamos o tipo *Option* na lista de tipos de variáveis (1) e clicamos OK (2). Vai aparecer outra janela onde devemos colocar o número de opções que desejamos (3). Podemos aumentar ou diminuir o número de opções posteriormente. No caso estamos criando uma variável com 3 opções. Após colocarmos o número de opções devemos apertar a tecla *enter*, surgindo os espaços para colocarmos as opções desejadas (4). Finalizando isso podemos clicar OK (5), para observarmos o resultado que aparece em destaque ainda na figura 3.5.

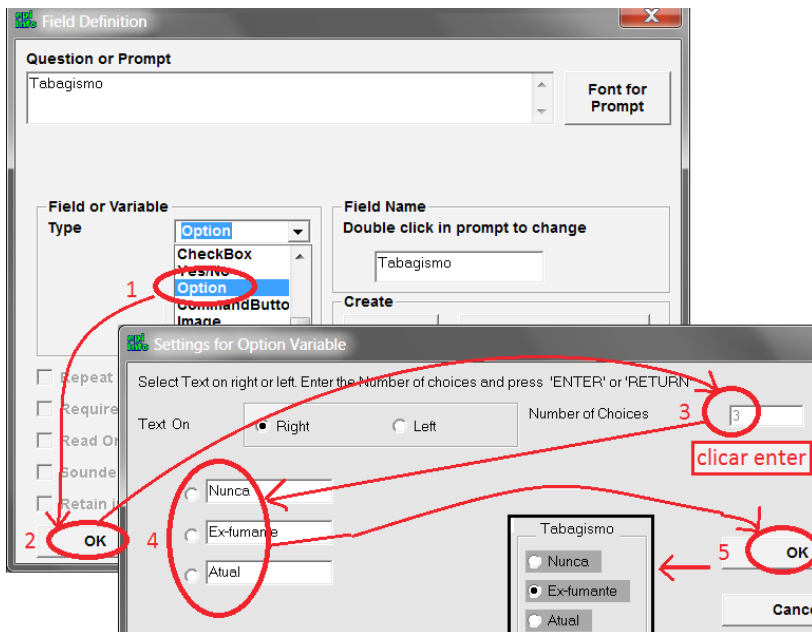


Figura 3.5 Passos para a criação de uma variável *Option* (Opção), no exemplo a variável Tabagismo, destacando no quadrado o resultado dos comandos.


Fonte: Epi Info®.



## 10. Outros tipos de variáveis:

- a. *Yes/No*: insere uma variável que aceita só dois tipos de resposta. É praticamente igual à variável Check Box, mas permite que tenhamos valores omissos, quando as pessoas não respondem.
- b. *Multiline*: cria uma variável onde podemos colocar um texto com várias linhas.

### Modificando posição das variáveis e o tamanho da janela de resposta na ficha

Podemos facilmente modificar a posição da variável clicando em cima dela e arrastando com o mouse até o local que desejamos. Para aumentarmos ou diminuirmos o tamanho da janela de resposta devemos clicar em cima da janela uma vez. Nas variáveis com *Legal Value* devemos clicar o mouse sobre a janela e pressionar a tecla *Alt*. Isso faz aparecer números no fundo da janela e pequenos quadrados ou pontos quadriláteros nos quatro cantos e no meio da mesma. Devemos posicionar o mouse exatamente em cima do ponto do meio no canto direito da janela, pois isso fará com que a ponta do mouse modifique a sua forma, passando a ter duas setas horizontais, conforme a figura: . Com o mouse assim posicionado podemos arrastar para a direita ou esquerda até chegarmos ao tamanho desejado. Essa opção é particularmente importante para a variável TABAGISMO, onde temos os comentários dos valores legais. Só podemos alterar o tamanho da janela de entrada de dados quando temos uma ficha nova sem dados na tabela correspondente.

### Agrupando variáveis em uma Ficha (comando Group)

As variáveis podem ser agrupadas na ficha, segundo suas características ou simplesmente pelo desejo do seu criador. Essa característica poderá ser usada quando estivermos analisando essas variáveis. Esse grupo de variáveis poderá ser movido facilmente na página e poderá ser usado para gerar uma análise separada. Para englobarmos as variáveis REG, DATANASC, DATAEXAME, IDADE, GENERO, COR, ESTADOCIV, devemos clicar algo acima da primeira variável e arrastando de tal forma a criar um quadrado ao redor das variáveis. Com o quadro ativo clicamos no menu *Insert, Group*. Surge uma janela (conforme a figura 3.6) onde colocamos uma denominação ou descrição do grupo, que deverá ser diferente das variáveis.

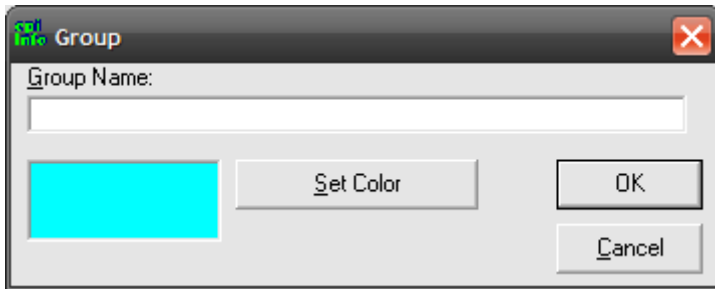


Figura 3.6. Janela de denominação de grupo do *Make View*.  
Fonte: Epi Info®.

Essa descrição, sem espaço, aparecerá como variável em algumas análises no Epi Info®. Podemos também selecionar a cor do fundo do grupo. Como outro exercício, vamos agrupar as variáveis que estão identificadas como “Dados Clínicos” na tabela 3.1, margem da esquerda. Devemos escrever a palavra “Clinicos”, sem acento, pois ele será transformado em outra letra ou caractere.

Podemos incluir uma nova variável dentro do grupo simplesmente arrastando a mesma para dentro da caixa do grupo. A caixa também pode ser aumentada ou diminuída de tamanho. É bom movimentarmos um pouco a caixa do grupo para saber se todas as variáveis estão realmente no grupo. Às vezes, uma variável fica nas margens do grupo e não é incluída, se ela não estiver incluída ela não se movimentará junto com o resto do grupo.

### **Criando variáveis com preenchimento automático (Check Code)**

Podemos automaticamente preencher algumas variáveis. Temos, na ficha que estamos criando, três variáveis: data do exame (DATAEXAME), data de nascimento (DATANASC) e idade (IDADE). A data do exame será para nós, no exercício, a data do dia do preenchimento, e a idade pode ser calculada como o número de anos entre a data de nascimento e a data de hoje. A data do dia do preenchimento pode ser inserida automaticamente após, por exemplo, da variável registro (REG). Igualmente, a idade pode ser calculada tão logo tenhamos colocado a data de nascimento. Assim entendendo, passaremos ao comando *Check Code* (*Código Check*), única opção do menu *View*. Abre uma janela, como na figura 3.7. Na primeira opção, superiormente colocada (1), escolhemos a variável que desejamos que o comando ocorra, antes ou depois (2) da introdução do dado nela, no nosso caso depois da variável REG. É muito comum, no Epi Info®, termos um número bastante grande de opções que nos deixam bastante

atrapalhados. Temos que sempre lembrar o que queremos, no caso trabalhar com uma variável, especificamente atribuir um valor a uma variável. Selecionamos, então, o comando *Variables (Variáveis)* (3), onde temos o comando *Assign (Atribuir)* (4), esse comando é muito útil e aparecerá mais vezes em várias partes deste livro.

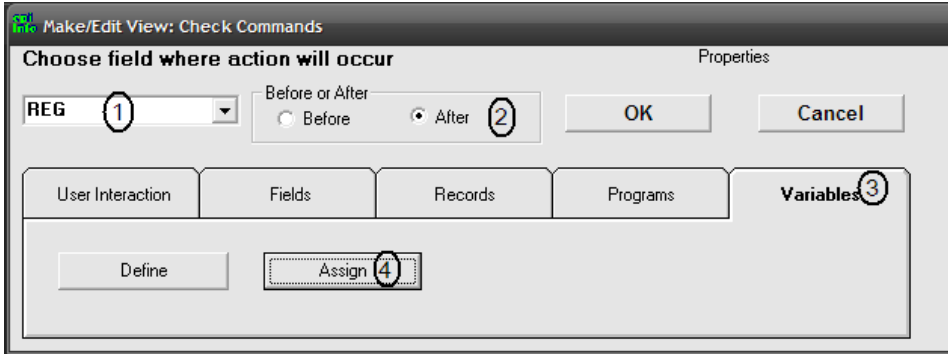


Figura 3.7. Janela de criação de *Check Code* do *Make View*.  
Fonte: Epi Info®.

### Atribuindo a data do dia (Assign Systemdate)

Clicando no botão *Assign* (4) da janela do *Check Code* abre uma janela (figura 3.8) onde podemos atribuir um valor a uma variável. A variável à qual queremos atribuir o valor deve ser escolhida ou escrita no primeiro compartimento da janela do *Assign* (1). Escrevendo o comando *SYSTEMDATE* no compartimento = *Expression* (2), introduziremos a data (dia, mês e ano) do sistema do computador que estamos usando. É esperado que essa data seja a data do dia em que estamos usando o computador, mas é bom checar o relógio do mesmo. Como vemos na janela, podemos usar letras minúsculas ou maiúsculas nesse comando.

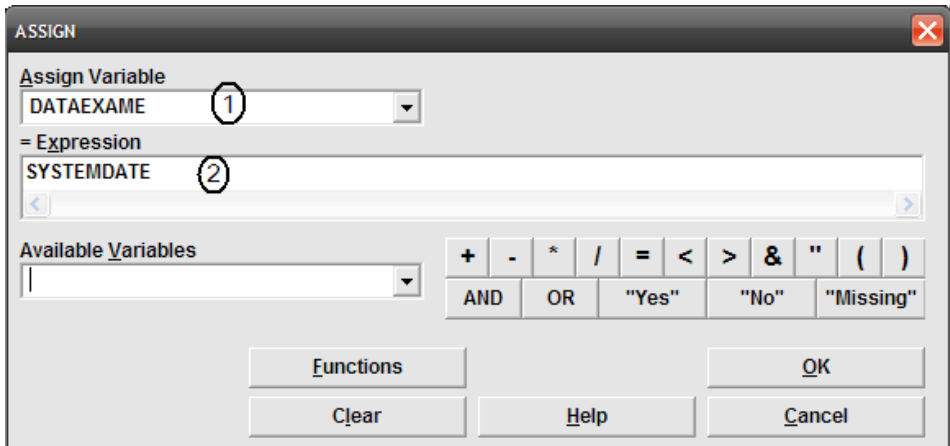


Figura 3.8 Janela do *Assign* do Epi Info® demonstrando a atribuição da data do sistema a uma variável.  
Fonte: Epi Info®.

Ao clicarmos *OK*, na janela do *Assign*, observamos que o comando será adicionado no compartimento inferior da janela do *Check Code*, esse comando, por sinal, não é muito complicado e poderia ter sido escrito diretamente no compartimento se soubéssemos como. Clicamos em *Save* no compartimento dos comandos para salvar o que fizemos.

### Calculando os anos entre duas datas (Years)

Podemos calcular os anos entre duas datas; no nosso caso, para calcular a idade entre a data de nascimento e a data de hoje. No nosso exercício, desejaremos calcular a idade após introduzirmos a data de nascimento, pois a data de hoje será colocada tão logo digitemos o registro. Escolheremos, inicialmente, a variável *DATANASC* na janela das variáveis do comando *Check Code* (1) e, no compartimento de comandos das variáveis, poderemos simplesmente digitar: *Assign idade=years(datanasc,dataexame)*. Atenção: cuidado com a posição das variáveis!

Clicamos *Save* novamente e finalmente *OK* para fecharmos a janela. Esses comandos somente serão executados quando terminarmos de entrar os valores das variáveis correspondentes. Podemos editar as variáveis *DATAEXAME* e *IDADE* clicando duas vezes em cima delas e selecionar a opção *Read Only*, isso evitará que coloquemos outros valores que não os determinados pelo nosso programa.

## Funções do Epi Info®

Além do *Systemdate* e do *Years*, existem outras funções do Epi Info® que podem ser observadas clicando na barra *Functions (Funções)* do comando *Assign (Atribuir)*. Elas são dispostas na figura 3.9 e são descritas junto a alguns operadores como operações aritméticas. As funções são divididas, então, em funções de data (*Date Functions*), sistema (*System*), tempo (*Time*) e de texto (*Text*). Descreveremos a seguir algumas delas:

### Functions and Operators

OPERATORS	<a href="#">ROUND</a>	TIME FUNCTIONS
<a href="#">ARITHMETIC</a>	<a href="#">NUMTODATE</a>	<a href="#">HOURS</a>
<a href="#">COMPARISON</a>	<a href="#">NUMTOTIME</a>	MINUTES
BOOLEANS	DATE FUNCTIONS	SECONDS
<a href="#">AND</a>	<a href="#">YEARS</a>	<a href="#">HOUR</a>
<a href="#">OR</a>	<a href="#">MONTHS</a>	<a href="#">MINUTE</a>
<a href="#">XOR</a>	<a href="#">DAYS</a>	<a href="#">SECOND</a>
<a href="#">NOT</a>	<a href="#">YEAR</a>	TEXT FUNCTIONS
NUMERIC	<a href="#">MONTH</a>	<a href="#">TXTTONUM</a>
<a href="#">EXP</a>	<a href="#">DAY</a>	<a href="#">TXTTODATE</a>
<a href="#">SIN, COS, TAN</a>	SYSTEM FUNCTIONS	<a href="#">SUBSTRING</a>
<a href="#">LOG</a>	<a href="#">SYSTEMDATE</a>	<a href="#">UPPERCASE</a>
<a href="#">LN</a>	<a href="#">SYSTEMTIME</a>	<a href="#">FINDTEXT</a>
<a href="#">ABS</a>	<a href="#">ENVIRON</a>	<a href="#">FORMAT</a>
<a href="#">RND</a>	<a href="#">EXISTS</a>	
<a href="#">TRUNC</a>	<a href="#">FILEDATE</a>	

Figura 3.9. Lista de funções e operadores do Epi Info®.

Fonte: Epi Info®.

**1. Funções de Data (Date Functions):** Já vimos um exemplo de função de data, a função *Years*, que conta o número de anos entre duas datas. As funções *Months* e *Days* são muito parecidas, pois contam, respectivamente, o número de meses e dias entre duas datas. As outras três funções extraem uma parte de uma data, respectivamente o ano, o mês e o dia, quando usamos as funções *Year*, *Month* e *Day*. Note que a diferença está somente no fato de uma função ser plural ou singular, por isso devemos tomar cuidado. O uso das últimas três funções é muito parecido. Por exemplo, se quisermos obter somente o ano de uma variável data como a *DATANASC*, criando uma variável chamada *ANONASC*, colocaríamos: *Assign Anonasc=Year(Datanasc)*.

**2.** *Funções de Sistema (System Functions)*: Dentre as funções de sistema, as mais interessantes são justamente as que se referem à data e ao tempo do sistema: *Systemdate* e *Systemtime*, cujas utilizações são semelhantes e já foram abordadas anteriormente.

**3.** *Funções de Tempo (Time Functions)*: As funções de tempo usam a mesma estrutura e têm a mesma utilidade que as funções de data.

**4.** *Funções de Texto (Text Functions)*: As funções de texto, *Txttonum* e *Txttdate*, servem para manipular variáveis do tipo texto, transformando-as em variável numérica ou variável tipo data. Isso é útil quando, erroneamente, essas variáveis foram definidas, no início, como variáveis tipo texto. A função *Format* tem a mesma função que as anteriores, mas possibilita a transformação em vários formatos. É aconselhável a leitura do “ajuda” para obter maiores detalhes sobre o seu uso.

**5.** Funções *Substring* e *Findtext* ajudam a procurar informações dentro de variáveis do tipo texto. *Substring* extrai o conteúdo de uma faixa determinada da variável. Por exemplo, podemos extrair quatro letras da esquerda para a direita da variável a partir da segunda letra: *Assign variavelnova=substring(variavelvelha, 2,4)*.

**6.** Já a função *Findtext* atribui a posição em que se encontra o parâmetro desejado.

**7.** Finalmente, a função *Uppercase* transforma todos os valores da variável texto em maiúscula.

## Adicionando páginas (Add/Insert Page)

Quando temos um questionário muito grande, podemos acrescentar uma ou mais páginas à nossa ficha. O comando *Add page* acrescenta uma página antes e o comando *Insert page* acrescenta uma página depois da página atual.

## Limite de tamanho da Ficha

Cada ficha do Epi Info<sup>®</sup> pode abrigar até 250 variáveis, neste número não estão incluídos os *Lable/Title*, mas inclui as variáveis que definem os grupos. Podemos criar, entretanto, várias fichas dentro de um mesmo projeto.

## Criando botão de ligação entre fichas (Related View)

Podemos criar uma nova ficha a ser preenchida logo após a introdução dos dados ou quando temos algum dado especial que desejamos preencher quando nos deparamos com algum caso especial. Por exemplo, poderemos criar uma ficha para colocar dados específicos para mulheres como “data da primeira menstruação”, “número de gravidezes” etc. Esse botão é criado como se fosse uma variável, colocamos o *Prompt* e o nome da variável exatamente da mesma forma como se estivéssemos colocando uma variável *Text*. No caso, poderemos colocar no *Prompt* “Dados sobre as Mulheres” e como nome do campo “MULHERES”. Temos, entretanto, que clicar na opção *Related View*. Isso fará aparecer uma nova janela (figura 3.9) que possui duas opções importantes: O botão será sempre ativo (*Any time*) ou ativo quando certas condições forem verdadeiras (*Only when ...*) (1). No caso, se quisermos que o botão seja ativo somente quando a variável GÊNERO for Feminino devemos, além de checar o botão correspondente, colocar o seguinte comando: *GÊNERO= "Feminino"*. Atenção: não esquecer as aspas, colocar exatamente como aparece na figura 3.10 no compartimento (3).

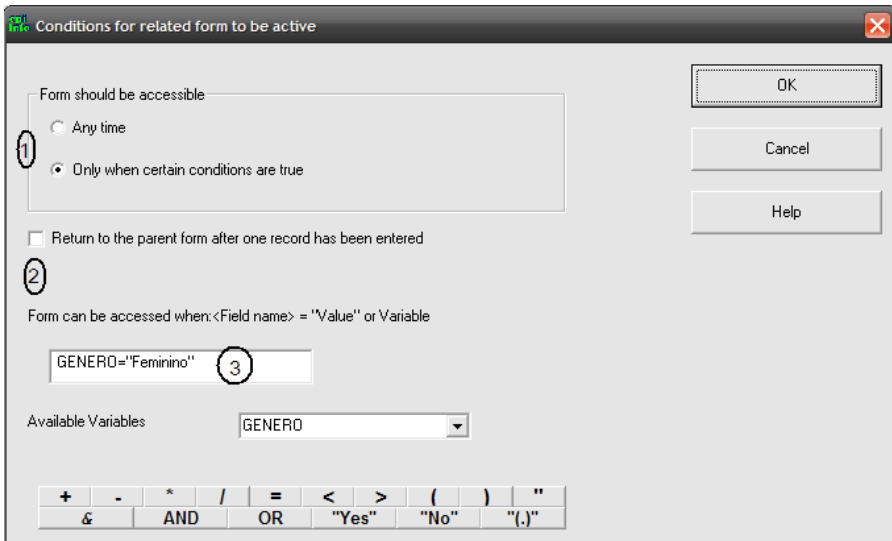


Figura 3.10. Janela de criação de *Related View* no *Make View*.  
Fonte: Epi Info®.

Outra opção é podermos ou não voltar à janela inicial, após completarmos o preenchimento da ficha (2), vamos checar a opção. É importante ressaltar que a nova ficha criará uma nova tabela de dados, independente da tabela criada pela ficha inicial. Assim, devemos incluir algumas variáveis, principalmente a variável REG (Registro) se quisermos relacionar as duas tabelas posteriormente. Ao clicarmos *OK* surge a opção de relacionarmos a uma ficha existente ou criarmos outra. Vamos criar outra ficha, por isso clicamos nessa opção. O botão surge na tela da ficha que estamos criando.

### Editando o botão de ligação entre fichas

Quando colocamos o mouse em cima do botão de ligação entre fichas, sem clicar no mesmo, aparece uma caixa orientando para algumas opções: *Alt Click to resize*, com a tecla *Alt* pressionada clicamos com o mouse em cima do botão, aparecerão os pontos quadriculados nos cantos e no meio das bordas do botão, usando o mouse, podemos alterar o tamanho. *Shift left click to move*, com a tecla *Shift* pressionada, que aciona as maiúsculas, podemos arrastar o botão com o mouse até o local que desejamos que ele fique posicionado. *Shift right click to edit*, com a tecla *Shift* pressionada ou não, clicando com o botão da direita do mouse, voltamos à janela de edição onde podemos inclusive deletar o botão. A última dica é a mais importante: *Control click*



to go to related file, pressionando então a tecla *Ctrl* e clicando somente uma vez em cima do botão, movemos para uma nova ficha, onde colocaremos as seguintes variáveis: REG, MENARCA. Voltamos à janela clicando no botão *Home* ou *Back* que estão à esquerda da janela, logo abaixo dos comandos de gerenciamento de páginas. Antes surgirá uma janela perguntando o nome da tabela a ser criada. No caso, o programa sugere o mesmo nome do campo: “MULHERES”. Aceitando, voltamos à janela da primeira ficha.

### Modificando a ordem de entrada dos dados (Order of FieldEntry [Taborder])

Quando criamos o questionário, o Epi Info® ordena os dados conforme sua posição na página: horizontalmente da esquerda para a direita e de cima para baixo. Se mudarmos a posição da variável, sua ordem também será alterada. Mas podemos modificar essa ordem, se quisermos, talvez porque sentimos que seria melhor de outro jeito.

Para modificar a ordem de entrada de campo devemos usar o programa *Criar Ficha*. Com a ficha aberta deveremos abrir o menu *Edit, Order of FieldEntry (Taborder)*. Uma janela aparecerá (figura 3.11). Nessa janela, *Tab Index* significa o número da ordem de entrada do campo. Podemos mudar essa ordem selecionando a variável que queremos e clicando nos botões *Up (1)* ou *Down (2)* até chegar ao lugar ou ordem desejada. Repetimos esse procedimento com as outras variáveis até que todas estejam na ordem que quisermos.

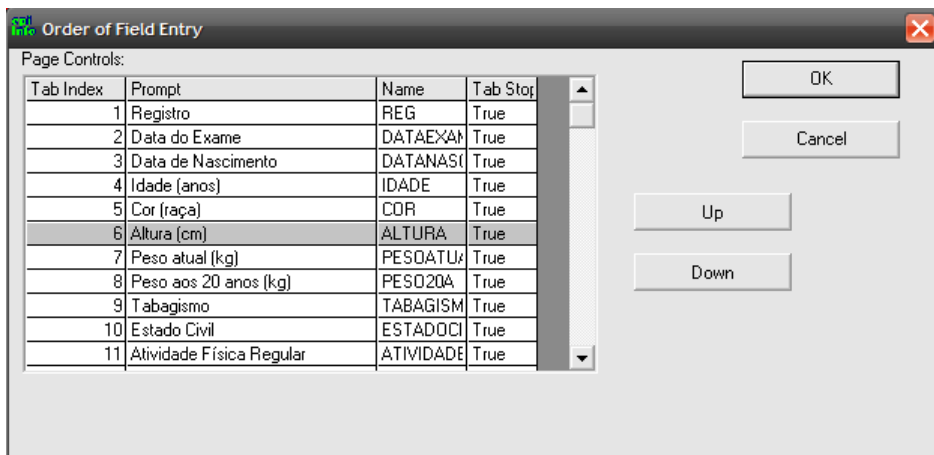


Figura 3.11. Janela do *Order of FieldEntry* do *Make View* destacando os botões de movimentação das variáveis. Fonte: Epi Info®.

Nossa tela ficaria mais ou menos conforme a figura 3.12.

Figura 3.12. Visão final da ficha criada pelo exemplo.

Fonte: Epi Info®.

Depois de ter concluído isso, podemos passar para a próxima fase, que é a de introdução de dados. Para tanto, precisamos fechar o programa *Criar Ficha (Make View)*.

## Criando a tabela de dados do View

Quando terminamos de criar a ficha e clicamos *exit* no menu *File*, o programa nos pede um nome para a tabela nova, criando um arquivo de dados em branco, podemos pôr o mesmo nome que usamos para a ficha. Quando colocamos os dados em uma tabela, o programa cria uma variável chave única que identifica cada um dos registros, essa chave é chamada *UniqueKey*. O programa nos pergunta se queremos iniciar essa variável pelo número 1 ou outro valor. Ela não nos ajuda muito, por isso tanto faz iniciarmos essa contagem a partir do número 1 ou não, assim deixamos como está e simplesmente clicamos *OK* para criarmos a tabela.

## Como criar duas fichas com as mesmas informações

Para que os valores de uma ficha sejam transportados para outra necessitamos copiá-los para variáveis permanentes. As variáveis permanentes mantêm o valor que posteriormente deverá ser copiado para as variáveis da outra ficha.

Os passos para a criação dessas variáveis são (figura 3.13):

1. No programa *Criar Ficha* (MakeView) abrir o *CheckCode* no menu *View*.
2. Na janela que pede para escolher o campo onde a ação vai ocorrer (Choose Field where action will occur), escolher a opção *DEFINEVARIABLES*.
3. Definir cada variável como permanente (sugestão: para transportar a variável “nome” criar a variável *nome\_transp*).

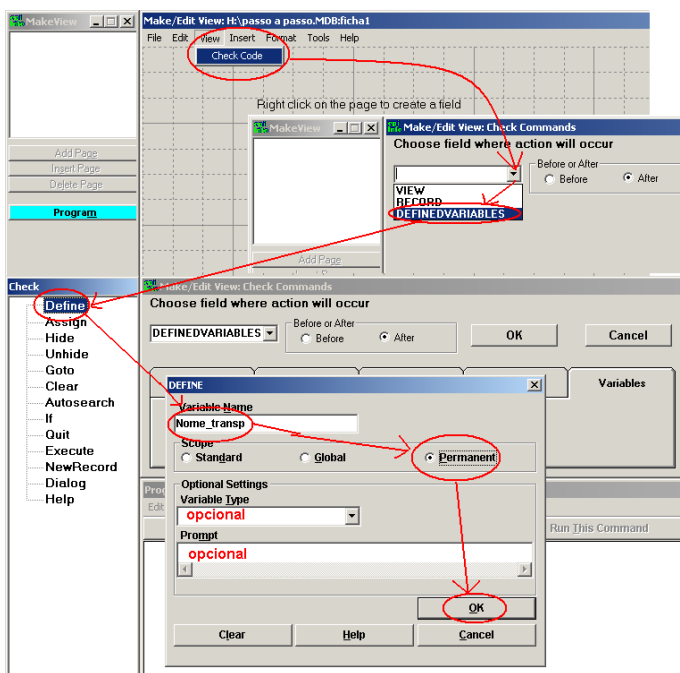


Figura 3.13 Passos para a criação de variáveis permanentes.  
Fonte: Epi Info®.

Depois de criadas as variáveis, necessitamos criar uma variável que executará a cópia dos valores para as variáveis de transporte. Pode ser uma variável *Check-code*. No exemplo, estou criando a variável “copia” e colocando no prompt para selecionar a opção para enviar os dados para a ficha 2.

Assumindo que as variáveis que queremos copiar da ficha inicial para a outra ficha já estejam definidas, devemos criar os comandos de cópia:

**1.** Menu View => Checkcode, escolher a variável *copia*.  
Vamos criar um comando para copiar os valores somente quando a variável for escolhida (valor=sim), usando um comando *if copia=(+)* como na figura 3.14.

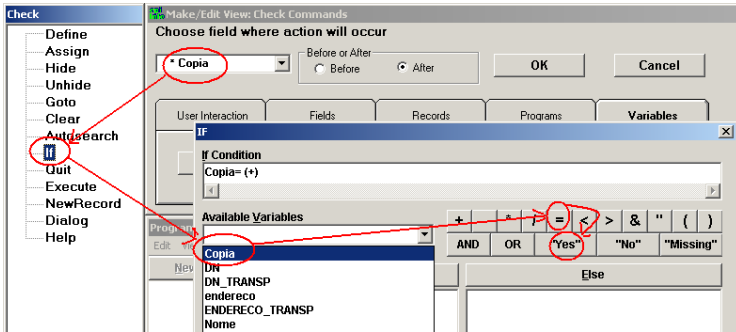


Figura 3.14. Passos para a criação do comando que copia os valores entre as fichas somente quando desejamos.

Fonte: Epi Info®.

**2.** Atribuímos então para as variáveis transporte os valores das variáveis da ficha 1. No exemplo teríamos conforme o ilustrado na figura 3.15.

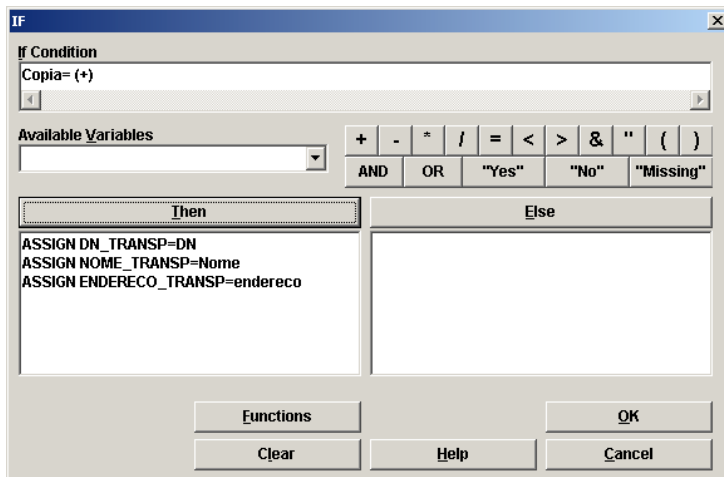


Figura 3.15. Ilustração dos comandos que copiam os valores desejados.

Fonte: Epi Info®.

**3.** Clicamos em *ok* para confirmar.

**4.** Agora necessitamos criar a ficha 2 criando um botão de ligação. Antes disso necessitamos salvar a ficha 1 e fechá-la. Dessa maneira a tornamos “mãe” da próxima ficha.

**5.** Abrimos novamente a ficha 1 e criamos o botão de ligação. O botão de ligação nada mais é que uma nova variável, no caso chamada ficha2, onde escolhemos a opção “Related View” na janela de definição da variável. Após, abre uma janela que nos permitirá definir se a ficha estará acessível sempre ou quando alguma condição for verdadeira. Como queremos utilizar a variável *copia* como comando para copiar as variáveis, vamos determinar que essa variável precisa ser selecionada para que a ficha seja acessível. Os comandos para isso são simples, conforme vemos na figura 3.16.

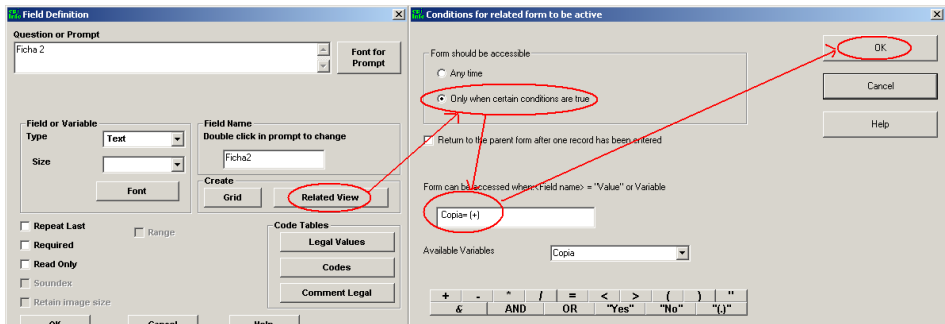


Figura 3.16. Comandos para tornar a ficha 2 visível.

Fonte: Epi Info®.

**6.** Agora necessitamos criar a nova ficha, ao clicarmos OK, o programa vai perguntar se queremos criar uma ficha nova ou associar o comando a uma ficha existente.

**7.** Com o botão *Ctrl* apertado clicamos em *Ficha2*. Isso nos levará a uma ficha nova onde devemos criar as variáveis que desejamos copiar. Os nomes das variáveis a serem copiadas devem ser exatamente iguais aos da ficha 1. Esse procedimento é na realidade somente uma sugestão, mas vai facilitar muito. Podemos também criar as variáveis com a opção de somente leitura, conforme vimos anteriormente na figura 3.2. É importante que o tipo e formato da variável seja idêntico à variável da ficha 1.

8. Deveremos criar uma variável que nos dê o comando para copiar os valores das variáveis *transporte* para a ficha 2. No exemplo a seguir, a variável é *confirma*. É importante que ela seja a primeira variável da ficha, ficando na posição mais superior e à esquerda da mesma para que seja a primeira variável a receber o cursor.

9. No comando *CheckCode* criamos, finalmente, os comandos que vão copiar os valores após sairmos da variável *confirma*, conforme a figura 3.17.

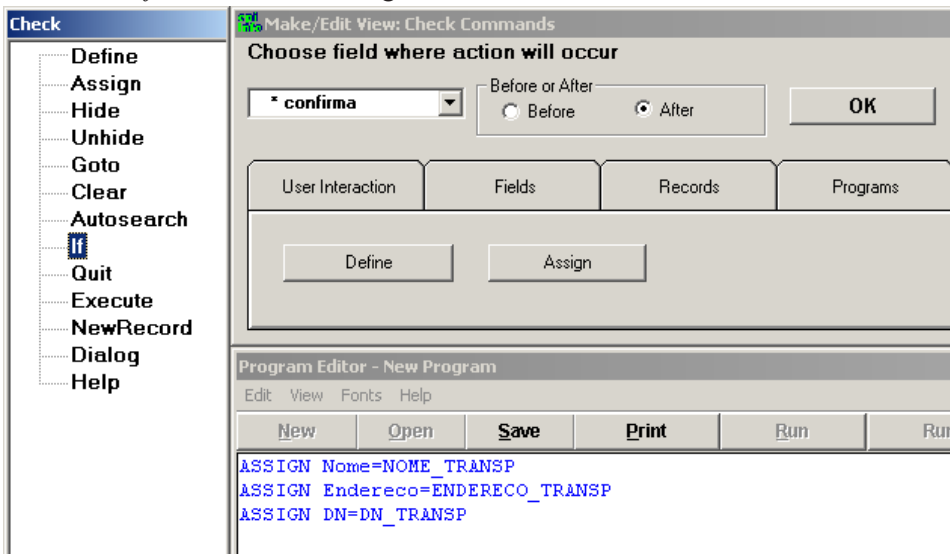


Figura 3.17. Comandos necessários para copiar os valores das variáveis que vão ter os mesmos resultados em ambas as fichas.

Fonte: Epi Info®.

## Introduzindo dados (Entrar Dados – Enter Data)

É uma etapa que não requer muita tecnologia, mas descreveremos algumas dicas para facilitar nosso trabalho e evitar maiores transtornos. Para tanto, vamos abrir o arquivo que queremos usando o programa *Entrar Dados (Enter Data)*. Clicar no menu *File (Arquivo)*, então *Open (abrir)* e selecionamos o arquivo, no caso escolheremos o arquivo *Seunome.mdb*. Depois de clicar *Open*, uma janela abrirá perguntando que ficha (*View*) queremos abrir. Temos duas fichas: *Dados* e *Mulheres*, como queremos a ficha dados, selecioná-la-emos.

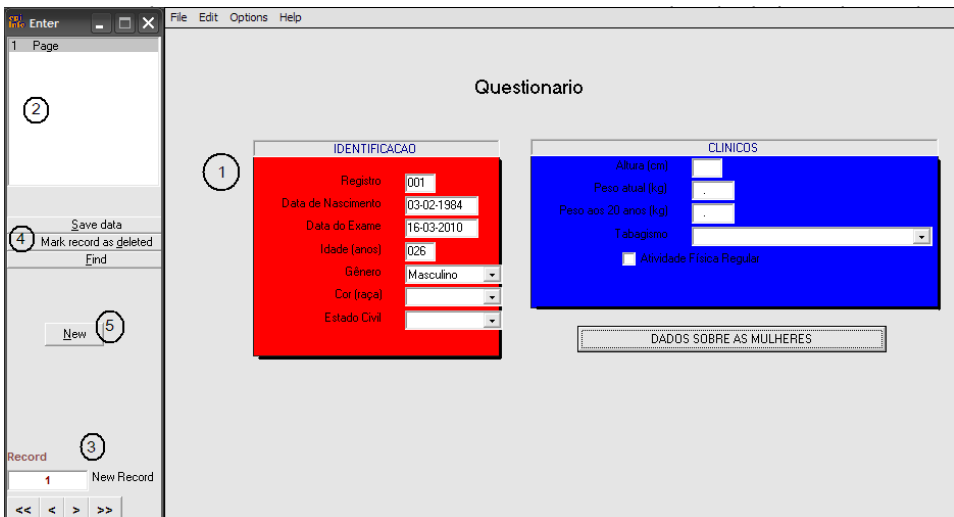


Figura 3.18. Tela do *Data Enter* do Epi Info® destacando seus componentes.  
Fonte: Epi Info®.

A primeira janela, a maior à direita, é a de introdução dos dados (marcada como (1) na ilustração). Os dados são colocados ou editados nesse componente. Para mover entre uma variável e outra usamos a tecla *TAB* ou *Enter*. Se preenchermos totalmente o espaço da variável, automaticamente o cursor passará para a próxima variável, conforme definida pela ordem de entrada. À direita em cima na tela, temos a janela das páginas (2), nela podemos observar qual página está aberta, no caso estamos trabalhando com a página 1, que é única. As páginas movem automaticamente depois que a última variável de cada página é introduzida, até mesmo se a página estiver em branco, exceto se existir uma variável cujo valor foi definido como necessário (*Required*). Igualmente, quando terminamos de colocar os dados na última página, um registro novo é aberto. Se, por alguma razão, quisermos colocar um registro novo antes que terminemos de colocar todos os dados do registro atual, podemos clicar no botão *New* (5). Na parte mais baixa da janela, à esquerda (3), o programa mostra o número do registro que estamos trabalhando e se aquele registro é novo (*New Record – Registro novo*) ou não. Se quisermos, podemos mover para o registro prévio clicando no sinal <. Isso fará com que o cursor mova para o registro prévio (na mesma página). Se o registro que estamos trabalhando é novo e não tem nenhum dado, o Epi Info® nos dará a opção para salvar o registro em branco. Fazendo isso, podemos nos mover ao registro prévio sem salvar ou podemos cancelar o comando e retornar ao

registro novo. Se quisermos pular para o primeiro registro, clicamos em <<. Igualmente, se quisermos mover para o próximo registro clicamos > e para o último registro em >>. Esse comando só moverá para um próximo registro existente, não começa um registro novo.

### Apagando registros (Mark Record as Deleted)

Se quisermos apagar um registro, podemos fazer isso clicando em *Mark Record as Deleted* (*Marcar registro como apagado*) (4). Como o comando diz, na realidade o programa não irá apagar o registro, somente o “esconde”. Poderemos recuperá-lo posteriormente clicando na opção *Recuperar*.

O Epi Info® salvará o arquivo automaticamente, após cada adição de ficha, mas sempre é um hábito saudável salvar o arquivo de vez em quando. Fazendo isso, nosso trabalho não sofre surpresas desagradáveis. Para esse propósito temos um botão chamado *Save data* (*Salvar dados*) (4). Como o Epi Info® salva os dados automaticamente enquanto introduzimos os dados, também é um bom hábito manter pelo menos uma cópia de segurança do projeto que estamos trabalhando.

### Características de introdução de algumas variáveis

**1.** *Variáveis Numéricas com ponto decimal:* quando temos que colocar uma variável com valores decimais, devemos colocar, evidentemente, os números inteiros primeiro e, depois, o caráter decimal, ponto, mesmo se não tivermos preenchido todo o espaço inteiro. Por exemplo, para colocar o peso de uma mesma pessoa magra: 22.5. Deveremos digitar exatamente assim: “ 2 2 . 5 “. Observe que terá um espaço entre o segundo 2 e o “. 5 “. Não precisamos nos preocupar com isso, pois o Epi Info® introduzirá o correto valor antes de salvar os dados.

**2.** *Variáveis Check Box:* não precisamos clicar com o mouse no local da variável, é bastante apertar a barra de espaço do teclado e a caixa ficará selecionada. Para desselecionar, basta apertarmos a barra de espaço novamente.

**3.** *Data contendo ano:* podemos digitar somente os últimos dois dígitos do ano, o Epi Info® transformará automaticamente em quatro dígitos. Porém, deveremos ter cuidado porque valores iguais ou superiores a 50 o programa converterá para 19NN.



Porém, se tivermos um valor inferior, por exemplo, 49, o programa transformará em 2049! Por isso, é um bom hábito sempre introduzir todos os quatro dígitos do ano.

**4.** Variáveis tipo *Texto* com *valores legais*: não precisamos escrever a resposta inteira, nem usar o mouse para selecionar a resposta, na caixa de respostas. Basta digitar as primeiras letras da resposta. Precisamos introduzir mais de uma letra se tivermos duas ou mais respostas que começam com as mesmas letras. Por exemplo, se temos duas variáveis que começam com a letra “s” teremos que colocar as primeiras duas letras. Se tivermos duas ou mais respostas que começam com “se” deveremos colocar três, e assim por diante, até que coloquemos um número suficiente de letras de forma que a resposta fique única.

**5.** *Variáveis Yes/No*: elas funcionam como uma variável tipo *texto* com *valores legais*: precisamos colocar só a primeira letra: “y” ou “n”.

## Exercício:

1. Crie uma ficha para o questionário fornecido abaixo, denomine o arquivo ApoE4 e tabela Grupo1. Chame os nomes dos campos como:
  - ID (identificação ###),
  - GENERO (texto Masc/Fem),
  - VISITA (data da primeira visita (MM-DD-YYYY)),
  - IDADE (número ###.#),
  - COLESTEROL (###),
  - PAS (primeiro valor, pressão sistólica, ###),
  - PAD (segundo valor, pressão diastólica, ###),
  - IMC (##.##),
  - CIGARRO (Nunca, Atual, Exfumante),
  - TEMPO (##.#),
  - EVENTO (yes/no),
  - APOE4 (check box).

Inclua informação sobre o modo como os dados deveriam ser introduzidos. (Observação: a pressão arterial não necessita aparecer como no formulário, pois deve ser definida com duas variáveis separadas.)

2. Introduza os dados conforme os questionários abaixo:

*Questionário para Alelo ApoE4 e Evento Coronário.*

ID: 1

Gênero: Masc

Data da primeira visita: 5/15/80

Idade na primeira visita: 17.2

Estado na primeira visita:

Colesterol total: 167 (mg/dl)

Pressão arterial: 107/74(mmHg)

Índice de Massa Corporal: 20.81 (kg/m<sup>2</sup>)

Fumou alguma vez: Nunca

Sobre o evento:

Teve evento coronariano? Não

Tempo de acompanhamento: 18.6

Alelo ApoE4: Não

---

*Questionário para Alelo ApoE4 e Evento Coronário.*

ID: 2

Gênero: Fem

Data da primeira visita: 6/8/88

Idade na primeira visita: 19.5

Estado na primeira visita:

Colesterol total: 106 (mg/dl)

Pressão arterial: 99 / 54(mmHg)

Índice de Massa Corporal: 20.80 (kg/m<sup>2</sup>)

Fumou alguma vez: Atual

Sobre o evento:

Teve evento coronariano? Não

Tempo de acompanhamento: 10.5

Alelo ApoE4: Sim

*Questionário para Alelo ApoE4 e Evento Coronário.*

ID: 3

Gênero: Masc

Data da primeira visita: 3/27/90

Idade na primeira visita: 19.8

Estado na primeira visita:

Colesterol total: 175 (mg/dl)

Pressão arterial: 130/90(mmHg)

Índice de Massa Corporal: 21.53 (kg/m<sup>2</sup>)

Fumou alguma vez: Nunca

Sobre o evento:

Teve evento coronariano? Não

Tempo de acompanhamento: 8.7

Alelo ApoE4: Não

---

*Questionário para Alelo ApoE4 e Evento Coronário.*

ID: 4

Gênero: Masc

Data da primeira visita: 7/20/71

Idade na primeira visita: 19.9

Estado na primeira visita:

Colesterol total: 166 (mg/dl)

Pressão arterial: 110/80(mmHg)

Índice de Massa Corporal: 20.34 (kg/m<sup>2</sup>)

Fumou alguma vez: Nunca

Sobre o evento:

Teve evento coronariano? Não

Tempo de acompanhamento: 27.4

Alelo ApoE4: Não

---

*Questionário para Alelo ApoE4 e Evento Coronário.*

ID: 5

Gênero: Masc

Data da primeira visita: 7/20/69

Idade na primeira visita: 21.3

Estado na primeira visita:

Colesterol total: 206 (mg/dl)

Pressão arterial: 123/88(mmHg)

Índice de Massa Corporal: 34.44 (kg/m<sup>2</sup>)

Fumou alguma vez: Nunca

Sobre o evento:

Teve evento coronariano? Não

Tempo de acompanhamento: 29.4

Alelo ApoE4: Sim

*Questionário para Alelo ApoE4 e Evento Coronário.*

ID: 6

Gênero: Masc

Data da primeira visita: 7/16/87

Idade na primeira visita: 21.3

Estado na primeira visita:

Colesterol total: 158 (mg/dl)

Pressão arterial: 115/65(mmHg)

Índice de Massa Corporal: 21.7 (kg/m<sup>2</sup>)

Fumou alguma vez: Exfumante

Sobre o evento:

Teve evento coronariano? Não

Tempo de acompanhamento: 11.4

Alelo ApoE4: Sim

---

*Questionário para Alelo ApoE4 e Evento Coronário.*

ID: 7

Gênero: Masc

Data da primeira visita: 12/8/90

Idade na primeira visita: 21.4

Estado na primeira visita:

Colesterol total: 124 (mg/dl)

Pressão arterial: 140/90(mmHg)

Índice de Massa Corporal: 25.07 (kg/m<sup>2</sup>)

Fumou alguma vez: Exfumante

Sobre o evento:

Teve evento coronariano? Não

Tempo de acompanhamento: 8

Alelo ApoE4: Não

---

*Questionário para Alelo ApoE4 e Evento Coronário.*

ID: 8

Gênero: Fem

Data da primeira visita: 6/8/91

Idade na primeira visita: 21.5

Estado na primeira visita:

Colesterol total: 146 (mg/dl)

Pressão arterial: 109/70(mmHg)

Índice de Massa Corporal: 22.61 (kg/m<sup>2</sup>)

Fumou alguma vez: Exfumante

Sobre o evento:

Teve evento coronariano? Não

Tempo de acompanhamento: 7.5

Alelo ApoE4: Sim

## Capítulo 4

# GERENCIANDO OS DADOS

### Estrutura do Programa Analysis Data (Análise de Dados):

Quando abrimos o programa *Analysis Data* (Análise de Dados) aparece uma tela (figura 4.1) com três janelas: *Analysis* à esquerda, *Analysis Output* em cima e *Program Editor* abaixo:

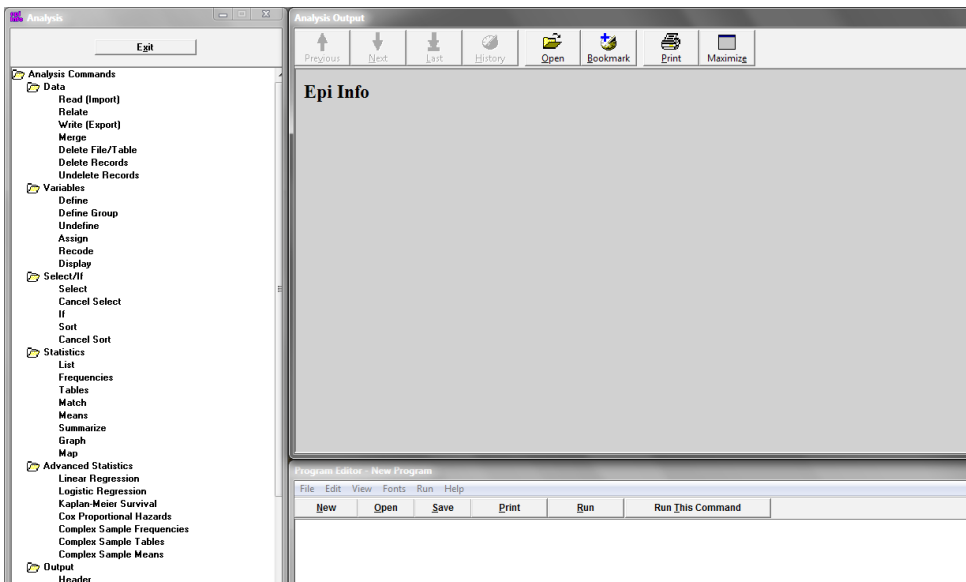


Figura 4.1. Tela do *Analysis* do Epi Info®.

Fonte: Epi Info®.

Na janela *Analysis* estão os comandos de análise, cujos resultados serão vistos na janela *Analysis Output*. Na janela *Program Editor* aparecerá o comando em sua linguagem de programação.

### Abrindo banco de dados no Epi Info®

Para abrimos um banco de dados para análise ou visualização, utilizamos o comando *Read (Import)* [*Abrir (Importar)*]. Automaticamente, quando clicamos nesse comando, abrirá uma janela posicionando para abrir o



último projeto (arquivo) utilizado. Observamos na janela *READ* (figura 4.2) que o primeiro compartimento chama-se *Current Project* (Projeto atual). O compartimento informa o projeto que está ativo, mas não podemos alterar o projeto nessa janela. Se o projeto ativo não for o projeto (arquivo *.mbd*) que queremos utilizar, devemos clicar no botão *Change Project* (Mudar Projeto). Se abrirmos um outro banco de dados, sem alterarmos o projeto ativo, criará-se um vínculo desnecessário com o projeto antigo.

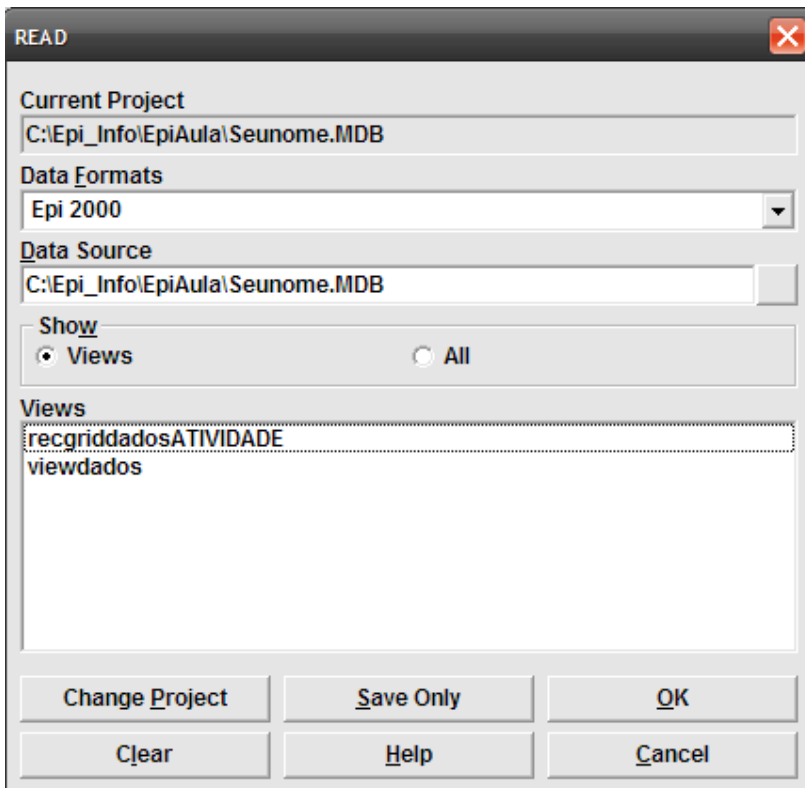


Figura 4.2. Janela do comando *Read* do *Analysis*.  
Fonte: Epi Info®.

## Mudando o projeto ativo (Change Project)

Ao clicarmos em *Change Project* (*Mudar Projeto*), surgirá então a janela para localizarmos o arquivo que desejamos abrir ou criar. Para criar um novo arquivo, basta colocar um nome novo. Devemos prestar atenção onde ele será salvo. Criar um arquivo novo pode ser importante quando estamos importando

um banco de dados em outro formato, por exemplo, *Epi 6* ou *Excel*.

Vamos selecionar o arquivo criado no capítulo anterior: *Seunome.mdb*. (o leitor poderá utilizar o arquivo *Exemplo.mdb*, que se encontra no folder *Dados* do Site [www.pucrs.br/igg/epiinfo](http://www.pucrs.br/igg/epiinfo)).

Após abrirmos um projeto (banco de dados), aparecerá, na janela chamada *Views*, o nome do questionário ou tabela de dados existente precedido da palavra *view* (no exemplo: *ViewDados*). Devemos selecioná-la clicando com o mouse uma vez e, em seguida, clicamos *OK* para abrir a tabela de dados. Veremos que, na janela *Analysis Output* aparece o nome do arquivo e a respectiva tabela que estamos trabalhando (1). Na linha seguinte aparece o número de fichas válidas com detalhes a respeito da data da última atualização. Na janela *Program Editor* aparece o comando em sua linguagem de programação (2).

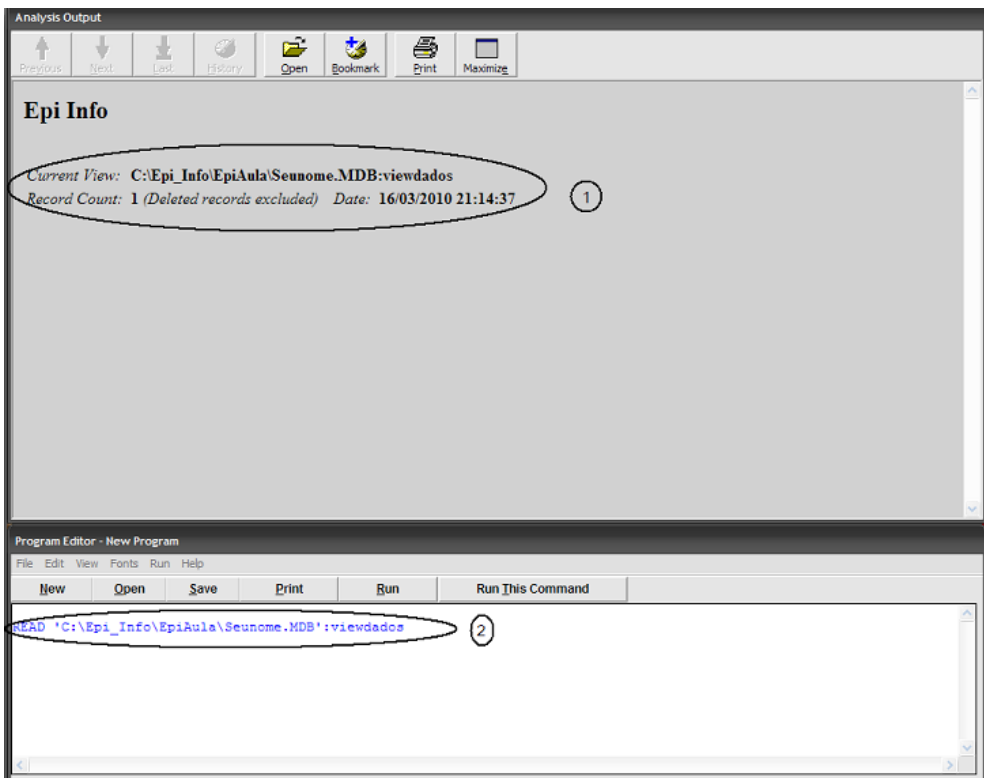


Figura 4.3. Janela de resultado (*Analysis Output*) e programa editor (*Program Editor*) criados pelo comando *Read*.

Fonte: Epi Info®.

## Observando os dados (Comando List)

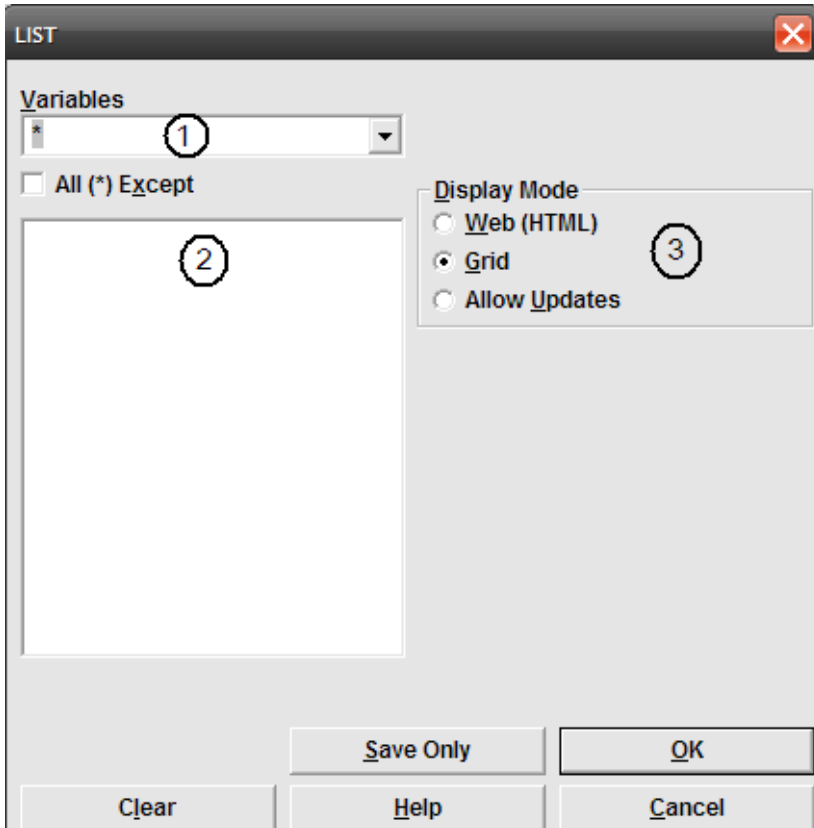


Figura 4.4. Janela do comando *List* do *Analysis*.

Fonte: Epi Info®.

O comando *List* é o primeiro comando do grupo *Statistics* da janela *Analysis Command*. Ele é utilizado para visualizar ou editar os dados. Clicando uma vez no comando, aparecerá a janela do comando (figura 4.4): No item *Variables* (1), podemos escolher as variáveis que desejamos. O Epi Info® dispõe as variáveis em ordem alfabética, não respeitando a ordem em que as colocamos na tabela. Devemos selecionar uma variável de cada vez ou todas, aceitando o comando asterisco (\*), que seleciona todas. No caso, desejamos mostrar todas as variáveis. Poderíamos também selecionar só as variáveis que não desejássemos, assim deveríamos marcar a caixa *All (\*) Except* e selecionar as variáveis não desejadas. Na lista obteremos as variáveis selecionadas também em ordem alfabética não importando a ordem de seleção que escolhermos.

## Tipo de saída dos dados no List

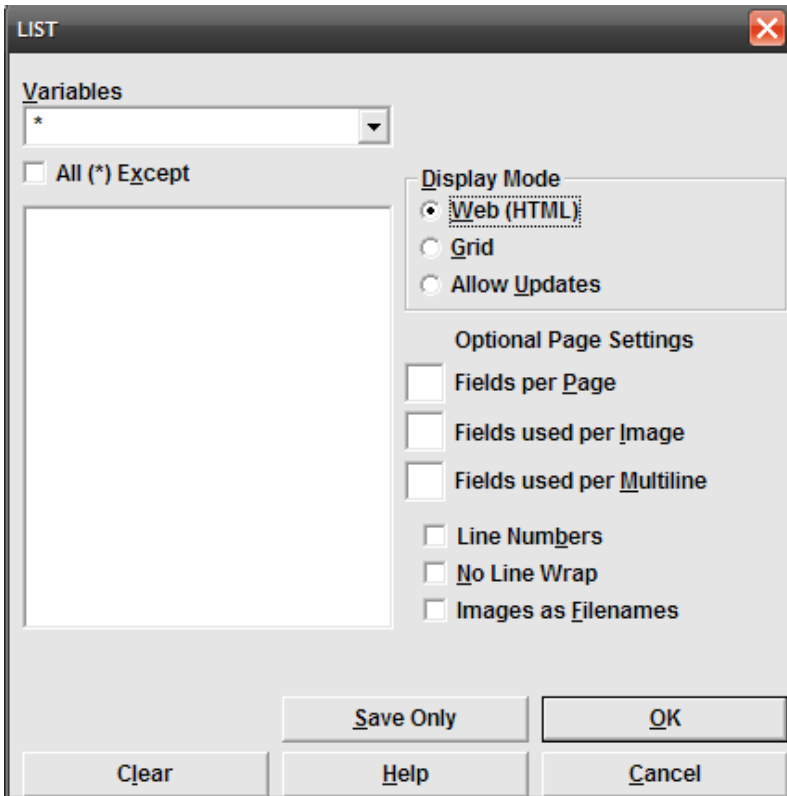


Figura 4.5. Janela do comando *List* do *Analysis* usando a opção *Web (HTML)*.  
Fonte: Epi Info®.

Automaticamente o Epi Info® posiciona a opção do tipo de saída (*Display Mode (3)*) para o formato de *Grade (Grid)*. Essa opção dispõe os dados em colunas muito parecidas com planilhas tipo *Excel*. Podemos alterar a largura das colunas colocando o cursor do mouse entre os nomes das colunas, essa opção não permite atualização. Se quisermos ter a oportunidade de editar os dados, devemos selecionar a opção *Allow Updates (Permitir Atualização)*. Ela é útil para corrigir valores incorretos que descobrimos durante a análise de dados. Iremos ver alguns exemplos de seu uso mais adiante. Se quisermos imprimir as variáveis, devemos utilizar outra opção. A opção *Web (HTML)* salva as variáveis em formato de página da Web, permitindo que se imprima ou que se edite posteriormente. Algumas opções adicionais são possíveis conforme

a figura 4.5. Podemos selecionar quantas colunas desejamos por página na opção *Fields per Page*, no caso devemos colocar o número de variáveis. As observações são normalmente identificadas pelo número da linha em que elas estão posicionadas. A opção *Line Numbers* não necessita ser selecionada.

## Juntando dois bancos de dados com o comando Merge

O comando *Merge* é útil para juntar dois bancos de dados, que devem ter pelo menos algumas variáveis idênticas, com dados a serem adicionados ou atualizados. Pode ser o caso de dois pesquisadores colocarem dados separadamente, ou quando utilizamos mais do que um computador para colocar os dados e desejamos transferir esses dados para um único computador. O comando é o último comando da secção *Data* da janela *Analysis Command*.

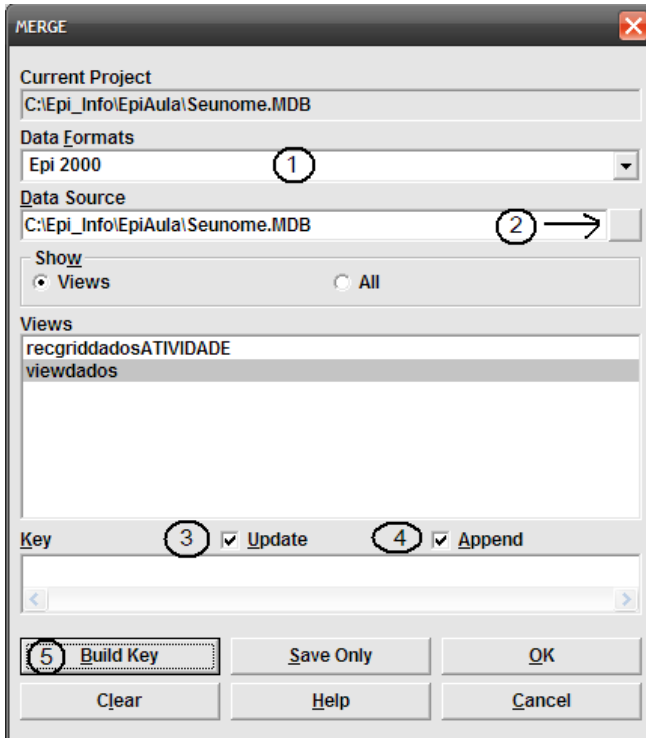


Figura 4.6. Janela do comando *Merge* do *Analysis* destacando seus diversos componentes.  
Fonte: Epi Info®.

A janela do comando *Merge* é muito parecida com a janela do *Read*. Se o arquivo que queremos juntar não está no formato Epi Info® 2000,

devemos selecionar o formato na janela *Data Formats (1)*. Selecionamos o local do banco de dados que queremos juntar na janela *Data Source*, clicando no ícone da direita que tem três pontos (2). No exercício vamos abrir o arquivo *Avaliacao.mdb* que está no folder *Exercícios* do site [www.pucrs.br/igg/epiinfo](http://www.pucrs.br/igg/epiinfo). Selecionamos o *ViewPalma*. Note que devemos estar com o arquivo *Seunome.mdb* ou *Exemplo.mdb* aberto.

O comando *Merge* possui duas opções importantes: o *Update* (Atualiza) e o *Append* (Adiciona) (3). O *Update* vai juntar os arquivos procurando somente dados em comum que tenham a mesma ficha, identificada por uma chave que veremos a seguir. Quando encontra uma ficha semelhante, atualiza os dados pelo arquivo novo, alterando os dados do arquivo inicialmente aberto. Quando a variável é a mesma (exatamente a mesma) os valores são atualizados. Quando a variável é nova, ou com nome diferente, não acrescenta essa variável ao banco de dados antigo (que estava aberto antes de iniciar o comando *Merge*). É importante destacar que o Epi Info® trata variáveis com nomes parecidos como variáveis diferentes. Por exemplo, as variáveis *EST\_CIV* e *ESTCIV* são consideradas variáveis distintas. A opção *Append* adiciona fichas novas ao banco de dados antigo. Igualmente, nessa opção, variáveis com nomes não idênticos serão consideradas variáveis diferentes.

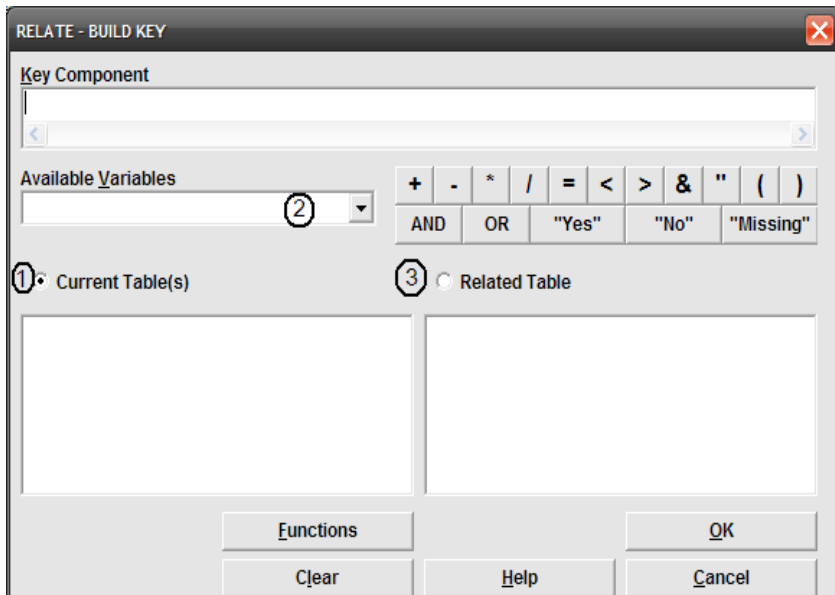


Figura 4.7. Janela do comando *Relate* do *Merge* destacando os compartimentos *Current table(s)*, *Available Variables* e *Related Table*.

Fonte: Epi Info®.

Mais um passo é necessário para juntarmos dois bancos de dados: criar uma chave de ligação. Essa chave (*Key*) ensinará o Epi Info® quais fichas são iguais e serão atualizadas e quais são diferentes e serão adicionadas. Na janela do *Merge* (*juntar*) clicamos no botão *Build Key* (5) situado em baixo à esquerda. Se a opção não está disponível é por que nenhum *View* foi selecionado. Surgirá a janela *Relate – Build Key* (figura 4.7), com três componentes principais: (1) *Current Table(s)* que se refere à tabela inicialmente aberta (antiga). *Available Variables* (2) onde selecionaremos a(s) variável(is) que vão compor a chave de relacionamento e o botão *Related Table* que se refere à nova tabela a ser juntada. Na maioria das vezes, temos uma única variável que identifica as fichas. No caso do nosso exercício, a variável REG é a que será utilizada. Com o *Current Table(s)* selecionado, procuramos na janela *Available Variables* a variável REG. As variáveis estarão, como sempre, em ordem alfabética, mas podemos digitar algumas letras da variável e ela aparecerá mais facilmente. Depois de selecionada a variável da tabela atual (antiga), passaremos para definição da variável da tabela nova (tabela relacionada). Para tanto devemos selecionar o botão *Related Table*. Observamos que a variável REG da tabela antiga está na janela logo abaixo do *Current Table(s)*. Novamente, na lista de variáveis, procuramos a variável da tabela nova, no caso tem o mesmo nome da tabela antiga: REG. Feito isso, clicamos *OK* duas vezes para retornar à janela *Merge*. Com a chave pronta podemos finalizar o processo clicando *OK*. O programa solicita um nome para realizar a ligação dos dois bancos de dados. Podemos colocar qualquer palavra como, por exemplo, “junta”.

Aparecerá, na janela *Analysis Output*, o resultado do processo de agrupamento, com mais fichas que anteriormente. Podemos verificar o processo pelo comando *List*.

### **Criando novas variáveis (comando Define)**

Podemos criar novas variáveis, por exemplo, a partir de variáveis já existentes. Esse processo, no Epi Info®, obedece algumas etapas. Primeiro, devemos definir um nome para essa variável, criando assim uma variável em branco, ou virtual. No exercício, devemos criar uma variável para Índice de Massa Corporal, ou simplesmente IMC. O IMC é um parâmetro muito utilizado que relaciona o peso e a altura de uma pessoa e proporciona critérios para definir se uma pessoa é obesa ou magra. Assim, depois de criado, utilizaremos o IMC para criar uma variável chamada *OBESIDADE*. Enquanto não salvamos o arquivo, essas variáveis serão sempre temporárias.

Primeiro vamos definir as variáveis: na secção *Variables (Variáveis)*, clicaremos no primeiro comando, *Define (Definir)*. Na janela, colocamos o nome das variáveis, uma de cada vez. Manteremos selecionada a opção *Standard* que mantém a variável somente durante a execução do Epi Info®. Variáveis podem ser definidas como *Permanent(e)* quando desejamos utilizá-las em todas as vezes que utilizamos o Epi Info®, enquanto variável *Global* manter-se-á na memória do programa, enquanto estivermos usando o mesmo projeto. No exercício, nossas variáveis serão padrões. Colocamos então o nome da primeira variável (IMC) e clicamos *OK*. Repetimos o processo para criar a variável *OBESIDADE*.

### Usando comando Assign (Atribuir)

Usamos o comando *Assign (Atribuir)* para transformar variáveis, principalmente as numéricas. No caso do nosso exercício, vamos usar as variáveis *ALTURA* e *PESOATUAL* para calcular o IMC. O comando é o terceiro da secção *Variables (Variáveis)*. Clicando nele obteremos a janela observada na figura 4.8.

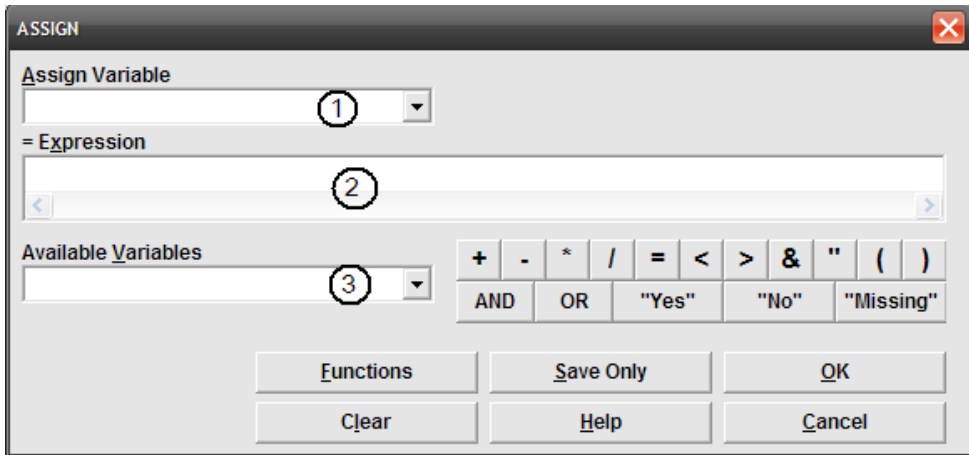
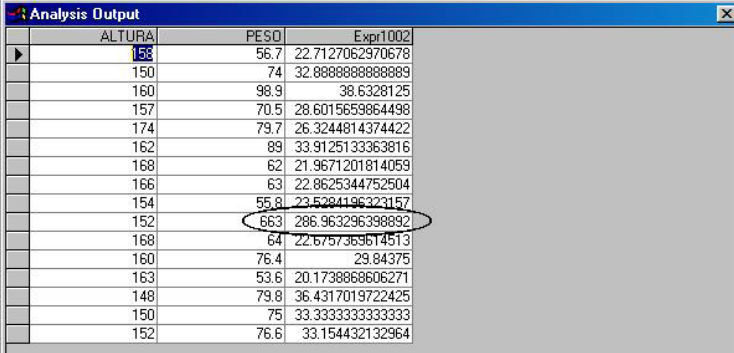


Figura 4.8. Janela do comando *Assign* destacando seus três compartimentos.  
Fonte: Epi Info®.

O primeiro compartimento, *Assign Variable (1)*, serve para definir para qual variável será atribuído o valor. No segundo, *= Expression (2)*, será colocada a fórmula que definirá o valor e as variáveis da fórmula serão selecionadas no já conhecido *Available Variables (3)*. No exercício, para calcular o IMC, devemos então colocar essa variável no primeiro compartimento (*Assign Variable*). O IMC é definido como o peso, em quilos, dividido pela altura, em metros, ao quadrado.



Nossa fórmula será então:  $PESOATUAL/(ALTURA/100)^2$ , onde devemos dividir a altura por 100, porque ela está em centímetros, a expressão  $^2$  eleva ao quadrado nosso valor. Ao clicar *OK*, o programa não nos informa se nossa fórmula está correta ou não. Por isso, observaremos o resultado usando o *List* para *ALTURA*, *PESOATUAL* e *IMC*, mas com a opção *Update* checada, que permite atualização.



ALTURA	PESO	Expr1002
153	56.7	22.7127062970678
150	74	32.8888888888889
160	98.9	30.6328125
157	70.5	28.6015659864498
174	79.7	26.3244814374422
162	89	33.9125133363816
168	62	21.9671201814059
166	63	22.8625344752504
154	55.8	23.5284196323157
152	663	286.963296398892
168	64	22.6757363614513
160	76.4	29.84375
163	53.6	20.1738888606271
148	79.8	36.4317019722425
150	75	33.3333333333333
152	76.6	33.154432132964

Figura 4.9. Janela de resultado do comando *List* após a criação da variável *IMC*.

Fonte: Epi Info®.

Observamos na figura 4.9 que um dos *IMC* (chamado aqui de *Expr1002*) não é “compatível com a vida”, pois tem um valor muito maior que os outros. Observamos também que o valor de peso desse registro foi introduzido de forma incorreta: 663. Revisando o questionário da mesma pessoa observamos que o valor correto é 66,3. Podemos corrigir na própria janela, pois estamos usando a opção que permite atualização. Em algumas versões de *Windows*, o valor de *IMC* será corrigido automaticamente. Caso isto não acontecer, poderemos também corrigir o *IMC* na janela, mas existe uma forma mais segura de recalculá-lo mesmo rodando o comando *Assign* novamente.

## Utilizando o Program Editor

Quando desejamos repetir um comando, basta localizá-lo na janela de programação: *Program Editor*, que fica situada na parte inferior à direita na tela. Colocamos o cursor em cima do programa que fez o *Assign* do *IMC*, dentro da janela de programação, clicando somente uma vez. Não necessitamos selecionar o programa, basta que o cursor fique piscando em qualquer parte dele. Feito isso, clicamos na última barra da direita da janela *Program Editor*, chamada *Run This Command*. Novamente podemos fazer o *List* para verificar o resultado do novo *Assign*. Para tanto, podemos novamente

utilizar o programa que está no *Program Editor*, dessa vez apagando a palavra “update”, e clicar novamente na barra *Run This Command*.

## Utilizando o Recode

Utilizamos o *Recode*, quarto comando da secção *Variables*, para recodificar variáveis categóricas (texto). O comando *Recode* utiliza somente uma única variável para recodificar outra ou ela mesma. No exercício, utilizaremos o comando para definir níveis de obesidade, baseados no IMC. Utilizaremos o seguinte critério:

Se IMC menor que 20 kg/m<sup>2</sup> então OBESIDADE = magro

Se IMC maior ou igual a 20 e menor que 25 então OBESIDADE = normal

Se IMC maior ou igual a 25 e menor que 30 então OBESIDADE = sobrepeso

Se IMC maior ou igual a 30 então OBESIDADE = obeso

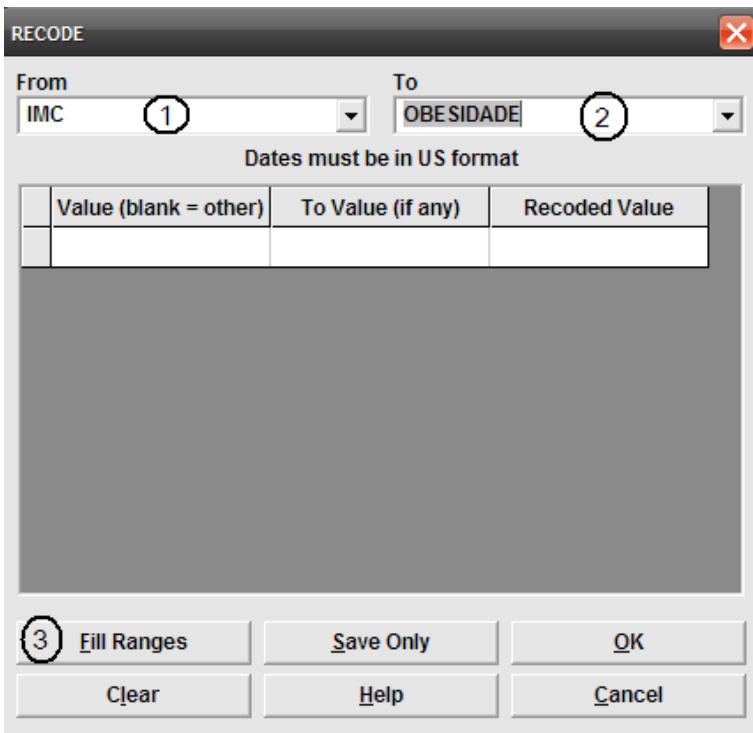


Figura 4.10. Janela do comando *Recode* destacando os compartimentos *From*, *To* e *Fill Ranges*.  
Fonte: Epi Info®.

Observamos que o critério mais baixo é 20 e que se modifica a cada 5 kg/m<sup>2</sup> até 30Kg/m<sup>2</sup>. Existem duas maneiras de colocar esses códigos. A maneira mais fácil nos ensinará como colocar os códigos da outra maneira.

Vamos abrir a janela do *Recode*. No primeiro compartimento superior à esquerda (*From (1)*) da janela do comando (figura 4.10) colocamos a variável que vai definir os critérios de recodificação. A variável a ser recodificada será colocada ao lado, no compartimento *To (2)*. No exercício temos duas variáveis, mas poderia ser a mesma variável. Antes de iniciarmos a definir os intervalos, devemos observar se o valor que usamos como limite pertence à margem superior ou inferior. No caso do IMC, os valores 20, 25 e 30 pertencem à margem inferior, pois esta está classificada como “maior ou igual”, as margens superiores não incluem o “igual”. Observamos que os intervalos têm tamanho fixo de 5 nos grupos. A maneira mais fácil de definir os intervalos é usando a opção *Fill Ranges*, no botão em baixo à esquerda marcado com (3).

## Usando a opção Fill Ranges

The image shows a dialog box titled "RECODE" with a close button in the top right corner. It contains two dropdown menus: "From" (set to "IMC") and "To" (set to "OBESIDADE"). Below these is the text "Dates must be in US format". There are three input fields: "Start" (20), "End" (30), and "By" (5). A "Reverse" checkbox is checked. Red circles with numbers 1, 2, 3, and 4 highlight the "From" field, "End" field, "By" field, and the "Reverse" checkbox, respectively. At the bottom, there are buttons for "Clear", "Help", "OK", and "Cancel".

Figura 4.11. Janela *Fill Ranges* do comando *Recode* destacando seus compartimentos.  
Fonte: Epi Info®.

A opção *Fill Ranges* do *Recode* produz uma nova janela que nos facilita “montar” os códigos. Como vemos na figura 4.11, existem três compartimentos importantes: o *Start*, que define o valor do intervalo menor, o *End*, que define o valor do intervalo maior, e, finalmente, o *By* que define o intervalo entre os grupos. No caso do exercício, o valor do intervalo menor é 20, que é o limite máximo para definição de magro, o valor do intervalo maior é 30, pois é o menor valor para a definição de obeso, e o intervalo entre os grupos é 5. Assim nossa janela ficará conforme a figura 4.11. Como os valores de referência pertencem aos limites inferiores dos intervalos temos que selecionar a opção *Reverse*, por motivos que veremos a seguir.

RECODE

From: IMC To: OBESIDADE

Dates must be in US format

Value (blank = other)	To Value (if any)	Recoded Value
30	HIVALUE	>=30
25	30	25 - <30
20	25	20 - <25
LOVALUE	20	<20

Buttons: Fill Ranges, Save Only, OK, Clear, Help, Cancel

Figura 4.12. Janela do comando *Recode* mostrando os códigos para a variável OBESIDADE.  
Fonte: Epi Info®.

Ao clicarmos *OK*, o programa retorna para a janela anterior (figura 4.12) com os intervalos já construídos: o código *HIVALUE* define que o

intervalo incluirá o valor maior que existir na nossa amostra, independente do valor (corresponderia ao infinito positivo). O código *LOVALUE* corresponderia ao menos infinito (ou infinito negativo). Observamos que nas colunas de definição dos valores de referência são colocados somente os números ou os dois códigos de limite máximo e mínimo.

Observamos também que a primeira definição inclui o valor ( $\geq 30$ ). Se não tivéssemos colocado a opção *Reverse*, a primeira opção seria  $\leq 20$ , o que não corresponderia a nossa definição de Magro ( $< 20$ ). Para finalizar podemos substituir os códigos na coluna *Recoded Value* pelos códigos que tínhamos: obeso, sobrepeso, normal e magro, conforme a sequência do exercício e clicamos OK.

The image shows a dialog box titled 'Fill Ranges' for the variable 'DISLIPIDEMIA'. It contains four input fields: 'Start' with the value '200', 'End' with '200', and 'By' with '200'. To the right of these fields is a checkbox labeled 'Reverse', which is currently unchecked.

Figura 4.13. Detalhe da janela *Fill Ranges* para a variável *DISLIPIDEMIA*.  
Fonte: Epi Info®.

Em outro exemplo, vamos definir a variável *DISLIPIDEMIA* como sendo valores maiores que 200 como *DISLIPIDEMIA sim*, e 200 ou menores como sendo *DISLIPIDEMIA não*. Nesse caso, o valor de referência pertence ao limite superior do intervalo. Podemos também utilizar o *Fill Range* novamente. O limite máximo e mínimo será igual: 200; como temos que colocar um *By*, colocamos 200 novamente, como na figura 4.13. Observamos que no caso não usamos a opção *Reverse*, porque o limite mínimo no valor de referência pertence ao limite superior e não ao inferior.

### Salvando o arquivo (comando *Write*)

Para salvar as variáveis que criamos e torná-las, assim, permanentes, devemos salvar o nosso arquivo. Com esse objetivo usamos o comando *Write (Export)* da secção *Data*. Na figura 4.14 observamos que o programa automaticamente seleciona a opção *Append* enquanto, na maioria das vezes, vamos querer utilizar a opção *Replace (1)*. Imediatamente devemos

marcar o botão correspondente para que não venhamos a nos arrepender quando for tarde demais.

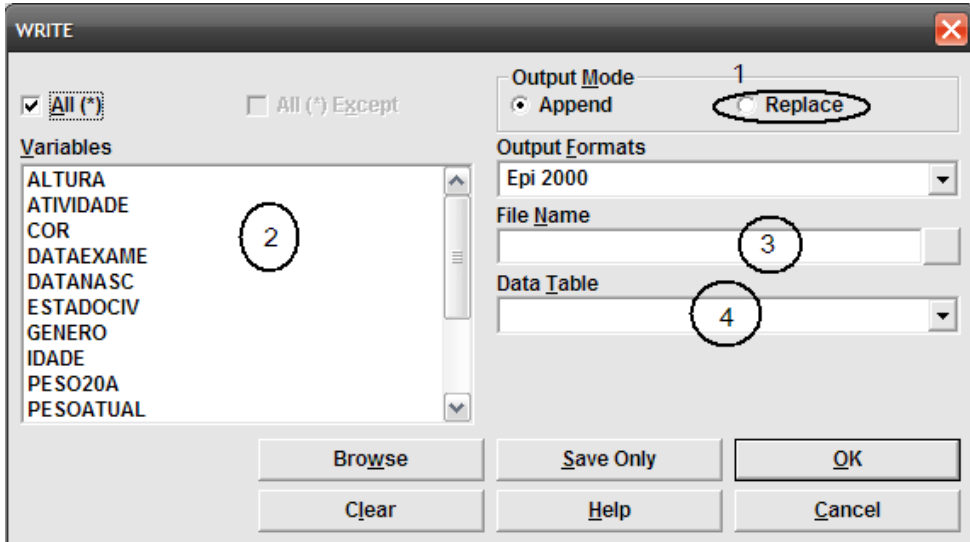


Figura 4.14. Janela do comando *Write* destacando seus componentes.  
Fonte: Epi Info®.

No compartimento *Variables*, podemos escolher quais variáveis queremos salvar ou, ao contrário, não salvar. No compartimento *File Name* escolhemos o nome do arquivo que queremos salvar. No exercício, salvaremos no mesmo nome que tínhamos aberto: *Seunome.mdb*. Necessitamos ainda definir em qual tabela serão salvos os dados, na mesma tabela ou em outra no compartimento *Data Table*. No exercício, usaremos a mesma tabela: *Dados*. Podemos salvar o arquivo em outro formato que não o de Epi Info®, mas isso fica como lembrança aqui.

## Salvando o programa Obesidade

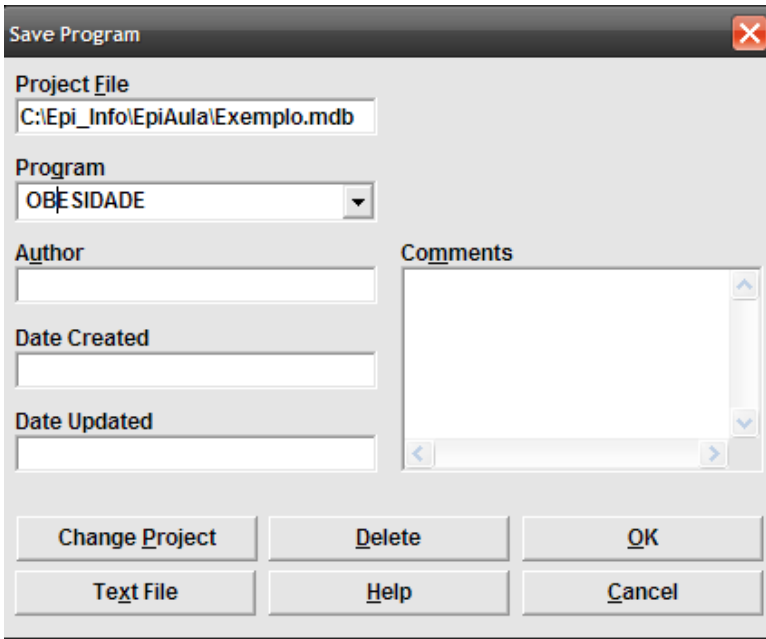


Figura 4.15. Janela *Save Program* do *Program Editor*.  
Fonte: Epi Info®.

Muitas vezes criamos alguns comandos que nos serão muito úteis no futuro. Por exemplo, os comandos que definem e recodificam a variável *OBESIDADE*. Podemos salvar esses comandos para um uso futuro. Assim, na janela *Program Editor*, apagamos todos os comandos que não pertencem aos códigos da definição e recodificação ficando só os comandos:

```
DEFINE obesidade
RECODE imc TO obesidade
30 - HIVALUE = "obeso"
25 - 30 = "sobrepeso"
20 - 25 = "normal"
LOVALUE - 20 = "magro"
END
```

No *Program Editor*, escolher *Save*. Aparecerá a janela, onde colocaremos o nome do programa: *Obesidade*, conforme a figura 4.15.

## Exercício:

1. Abra o arquivo *ApoeA.mdb* tabela Grupo1, criado no capítulo passado (ou contido no folder *Dados* do site [www.pucrs.br/igg/epiinfo](http://www.pucrs.br/igg/epiinfo)), modificando o projeto ativo. Junte com o arquivo *ApoeB.mdb* tabela Grupo2, no folder *Exercícios* do site [www.pucrs.br/igg/epiinfo](http://www.pucrs.br/igg/epiinfo), usando como *Build Key* a variável “ID”.
2. Defina e crie a variável *OBESIDADE* utilizando o programa criado durante a aula. (Detalhe: clique em *OPEN* no *Program Editor*, para achar o programa é necessário mudar o projeto na janela de abertura.) Execute o programa e faça o *List* do IMC e obesidade.
3. Defina as variáveis *IDADEEVENTO*, *DISLIPIDEMIA* e *HIPERTENSÃO*.
4. Recodifique *DISLIPIDEMIA* a partir do nível de *COLESTEROL* sendo >200 sim, 200 ou menos não. (Cuidado com a posição das variáveis!!=> na primeira janela vai o *COLESTEROL*.)
5. Recodifique *HIPERTENSÃO* a partir do nível de PAS sendo >=140 sim, menos não.
6. Atribua *IDADEEVENTO=IDADE+TEMPO*.
7. Salve com o mesmo nome original: *ApoeA*, tabela Grupo1 (cuidado com a opção *Adicionar*, o nome da tabela não pode ter espaço nem acento).



## Capítulo 5

# IMPORTANDO DADOS E MANIPULANDO VARIÁVEIS

### Abrindo banco de dados em formato Epi Info® 6.0

Como vimos, o Epi Info® trabalha com um banco de dados em formato *MS-ACCESS*. Esse formato nos permite trabalhar com várias tabelas; no Epi Info® essas tabelas (*tables*) são relacionadas a uma ficha (*View*). Se quisermos usar um arquivo em formato Epi Info® versão 6.0 poderemos transformar esse em um com formato Epi Info®, pelo programa *Make View*.

Com o programa *MakeView (CriarFicha)* sem qualquer arquivo aberto, devemos clicar no menu *Tools (Ferramentas)*, a primeira opção *Import Epi 6 REC file (Importar arquivo .Rec do Epi 6)*. Selecionando o arquivo que queremos abrir, o programa pede o nome do novo arquivo que queremos criar (arquivo novo) ou o nome do arquivo onde queremos colocar a *ficha* e a *tabela* (se o arquivo já existe). Depois de fazer isso, clicamos em *Abrir (Open)* e esperamos uns segundos. O programa então pede o nome da tabela de dados que queremos criar. Automaticamente o programa oferece o mesmo nome do arquivo do *Epi 6*, podemos aceitar aquela sugestão ou usar a vantagem de usar um nome mais longo. No exercício, ao final do capítulo, o leitor terá a oportunidade de executar esse comando.

### Importando dados em formato Excel

Vamos ver como se importa um arquivo do *Excel* no Epi Info®. Quando desejamos importar informações de um banco de dados com outra formatação, devemos tomar alguns cuidados. Antes de abrirmos o arquivo devemos ver se ele está estruturado apropriadamente para podermos importá-lo sem dificuldade. Em primeiro lugar, os dados devem estar todos em uma mesma planilha ou que tenham, em todas as planilhas, uma variável que identifique cada um dos registros (no caso do *Excel*, linhas). Essa variável pode ser, por exemplo: *REGISTRO*, *NOME* ou outro indicador. As variáveis devem estar indicadas na primeira linha, sem espaço, sem acento e com nomes distintos como na figura 5.1.

REG	DATANASC	DATAEXAME	IDADE	GENERO	COR	ESTADOCIV	IMC	ALTURA	PESOATUAL	PESO20A	TABAGISMO	ATIVIDADE	COLESTEROL	PAS	PAD
45	11/05/1926	24/04/1995	69	Masculino	Branca	Casado	28,52	168	80,5	55	Nunca	No	167	107	74
46	14/11/1947	02/11/1995	48	Masculino	Branca	Casado	31,30	145	65,8	45	Nunca	No	106	99	54
47	14/11/1947	02/11/1995	48	Masculino	Branca	Casado	31,30	145	65,8	45	Nunca	No	175	130	90
53	30/12/1949	19/12/1995	46	Feminino	Branca	Casado	25,69	168	72,5	60	Nunca	Yes	166	110	80
60	02/11/1940	20/10/1995	55	Masculino	Branca	Casado	28,70	152	66,3	62	Atual	No	206	123	88
61	02/11/1940	20/10/1995	55	Masculino	Branca	Casado	26,93	161	69,8	62	Atual	No	158	115	65
62	30/03/1947	18/03/1995	48	Masculino	Preta	Casado	33,13	168	93,5	55	Exfumante	No	124	140	90
63	29/04/1948	18/04/1995	47	Feminino	Branca	Casado	20,31	160	52	50	Nunca	Yes	146	109	70
64	15/01/1958	06/01/1996	38	Feminino	Branca	Casado	25,55	160	65,4	60	Atual	No	170	120	75
65	13/02/1939	30/01/1995	56	Feminino	Preta	Casado	24,33	159	61,5	50	Nunca	No	214	110	71
66	07/09/1942	25/08/1995	53	Masculino	Branca	Casado	25,85	170	74,7	65	Exfumante	No	149	127	86
67	07/09/1942	25/08/1995	53	Masculino	Branca	Casado	25,85	170	74,7	65	Exfumante	No	167	145	97,5
68	10/06/1952	31/05/1995	43	Masculino	Branca	Casado	20,18	175	61,8	61	Atual	Yes	200	135	90
76	10/03/1944	26/02/1995	51	Masculino	Branca	Casado	27,16	170	78,5	66	Nunca	No	125	117	66
77	30/08/1944	18/08/1995	51	Feminino	Branca	Casado	33,85	158	84	52	Nunca	No	156	104	70
78	29/05/1961	21/05/1995	34	Masculino	Preta	Casado	26,29	175	80,5	82	Atual	No	109	98	70
79	11/01/1956	01/01/1996	40	Masculino	Branca	Casado	24,94	174	75,5	55	Atual	Yes	220	108	60
80	26/01/1938	12/01/1996	58	Masculino	Mulata	Casado	32,45	167	90,5	56	Nunca	No	148	100	65
84	17/05/1952	07/05/1995	43	Feminino	Branca	Casado	18,76	154	44,5	29	Nunca	No	150	117,5	70
85	14/12/1937	30/11/1995	58	Masculino	Mulata	Casado	20,19	155	48,5	55	Exfumante	No	146	101	75
86	21/11/1961	13/11/1995	34	Feminino	Branca	Casado	28,71	160	73,5	50	Nunca	No	177	110	78
87	27/11/1954	17/11/1995	41	Feminino	Branca	Casado	34,13	155	82,5	50	Nunca	No	183	122	88
88	12/10/1925	25/09/1995	70	Masculino	Branca	Casado	28,89	169	82,5	65	Exfumante	No	146	116	80
89	21/05/1959	12/05/1995	36	Masculino	Branca	Casado	21,99	162	57,7	52	Atual	No	196	106	70
90	25/07/1941	12/07/1995	54	Masculino	Branca	Casado	29,76	169	85	68	Atual	Yes	153	152	100
91	27/06/1955	17/06/1995	40	Masculino	Mulata	Casado	30,10	170	87	55	Nunca	Yes	150	117	76
94	01/06/1956	23/05/1995	39	Feminino	Branca	Casado	24,96	158	62,3	50	Nunca	No	164,5	99	63,5
95	17/03/1949	06/03/1995	46	Feminino	Branca	Casado	29,20	165	79,5	65	Exfumante	Yes	182	129	66
96	15/01/1950	04/01/1996	46	Feminino	Branca	Casado	32,89	150	74	77	Nunca	Yes	172	97	73
97	02/09/1944	21/08/1995	51	Feminino	Preta	Casado	31,33	147	67,7	55	Nunca	No	221	124	78
101	28/08/1937	11/08/1995	58	Masculino	Branca	Casado	29,02	165	79	60	Exfumante	No	150	155	100

Figura 5.1. Detalhe de um arquivo com formato *Excel* pronto para ser importado.

Fonte: Microsoft Excel, arquivo *Avaliacao.xls*.

Além disso, nas planilhas que queremos importar não devem conter figuras, outras tabelas ou valores e informações além dos dados que desejamos. Para evitar possíveis incompatibilidades é recomendável não utilizar acentos ou espaços na primeira linha.

Partindo dessas premissas, podemos importar os dados utilizando o programa *Analysis Data*, comando *Read (Import)*. No compartimento *Data Formats*, escolhemos a versão mais apropriada do *Excel*; dependendo da versão que foi utilizada. Costumo, na dúvida, escolher a versão mais atual: *Excel 8.0*, que é utilizada pelo *Office 2000*. No exercício vamos importar o banco de dados *Avaliacao.xls* que está no folder *Exercicio* do site [www.pucrs.br/igg/epiinfo](http://www.pucrs.br/igg/epiinfo). Observamos que o programa automaticamente mostra as planilhas (*Worksheets*) do arquivo, por isso ele demora um pouco, pois necessita ler o arquivo primeiro. Vamos escolher a planilha “*avaliacao*”. Observe que não alteramos o projeto anteriormente em uso. Após clicar *OK* surgirá uma janela confirmando se a primeira linha possui o nome das variáveis. Como sempre, após abrir um banco de dados, o programa informa o número de observações

e o dia da última atualização do arquivo lido. Sempre é importante, após importarmos ou lermos um banco de dados, que não nos é familiar; observar quais os tipos de variáveis dispomos. Assim, após importarmos os dados, deveremos fazer um *List* de todas as variáveis.

### Utilizando o *Recode* (revisão)

Observamos, no banco de dados recém-importado, que existe a variável *IMC*, mas não existem as variáveis *OBESIDADE*, *DISLIPIDEMIA* e *HIPERTENSÃO*. Utilizaremos o *Recode*, de novo, para recodificar essas variáveis. Antes disso devemos defini-las, usando o comando *Define*.

Utilizaremos o seguinte critério para OBESIDADE:

Se *IMC* maior ou igual a 30 então *OBESIDADE*= obeso.

Se *IMC* maior ou igual a 25 e menor que 30 então *OBESIDADE*= sobrepeso.

Se *IMC* maior ou igual a 20 e menor que 25 então *OBESIDADE* = normal.

Se *IMC* menor que 20 kg/m<sup>2</sup> então *OBESIDADE*= magro.

Nosso comando deve ficar assim:

```
RECODE imc TO obesidade
```

```
30 - HIVALUE = "obeso"
```

```
25 - 30 = "sobrepeso"
```

```
20 - 25 = "normal"
```

```
LOVALUE - 20 = "magro"
```

```
END
```

Finalmente, utilizaremos o seguinte critério para Dislipidemia:

Se *COLESTEROL*  $\geq 240$  então *DISLIPIDEMIA* = Elevada.

Se *COLESTEROL*  $\geq 200$  e  $< 240$  então *DISLIPIDEMIA* = Moderada.

Se *COLESTEROL*  $< 200$  então *DISLIPIDEMIA* = Normal.

Nosso comando deve ficar assim:

RECODE COLESTEROL TO DISLIPIDEMIA

240 – HIVALUE = “Elevada”

200 – 240 = “Moderada”

LOVALUE – 200 = “Normal”

END

## Utilizando o comando IF

Podemos utilizar o comando *IF* para recodificar uma variável baseada em duas outras variáveis. Por exemplo, podemos chamar uma pessoa de hipertensa se ela tiver a pressão sistólica maior que 140 ou diastólica maior que 90. Para tanto usamos o comando *IF*.

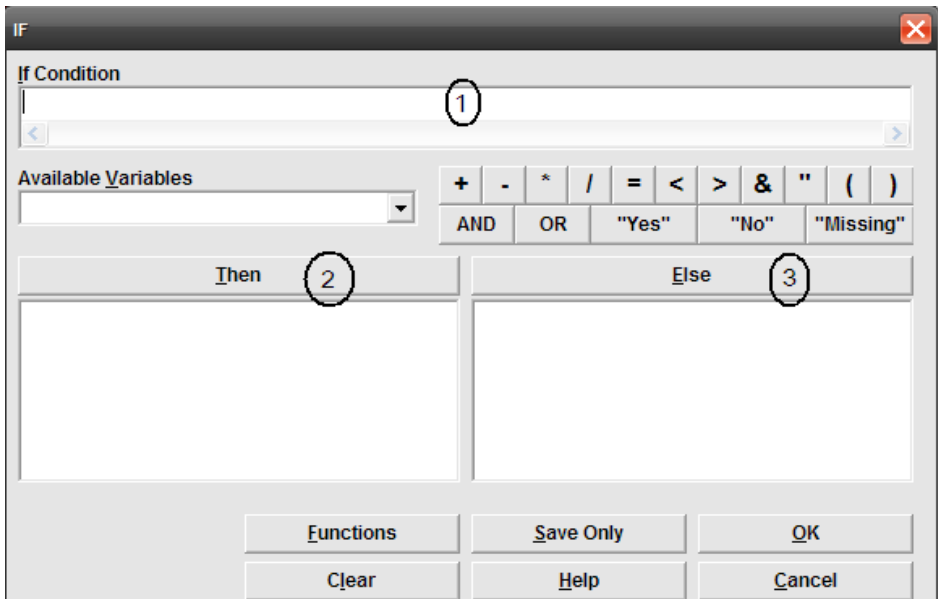


Figura 5.2. Janela do comando *IF* destacando os compartimentos *If Condition*, *Then* e *Else*.

Fonte: Epi Info®.

No compartimento (1) da janela do comando *IF* (figura 5.2), colocamos nossa condição:  $PAS > 140$  or  $PAD > 90$ . Na janela (2) colocamos o que queremos que seja feito nessa condição: `assign HIPERTENSAO= "Sim"`

ou o símbolo de SIM do Epi Info® (+). No *else* (3) colocamos um novo assign: HIPERTENSAO = (-). A janela vai ficar assim:

```
IF PAS>140 or PAD >90 THEN
ASSIGN HIPERTENSAO= "yes"
ELSE
ASSIGN HIPERTENSAO="no"
END
```

Vamos salvar esse programa para utilizá-lo novamente no exercício de hoje. Antes disso, devemos salvar o nosso projeto como *Avaliacao.mdb* e tabela *Avaliacao*. Após vamos apagar todos os comandos que não nos interessam como o *Read*, o *List* e o *Write*, salvaremos o programa como *classificacao*. Fechamos o *Analysis Data*.

### Criando uma ficha a partir da tabela (*Make View from Data Table*)

Quando criamos um programa em outro formato e salvamos os dados no formato Epi Info®, será conveniente criarmos também uma *ficha* (*View*). Para tanto, utilizamos o programa *Make View*; no menu *Tools* encontramos o *Make View from Data Table* ou no próprio menu *Utilities* do Epi Info® escolhemos a opção *Table-to-View*. Clicando em qualquer um dos comandos, a janela de seleção de arquivo será aberta. Vamos selecionar o arquivo que recém criamos: *Avaliacao.mdb*. Após isso o programa pedirá a tabela que será a única do arquivo: *Avaliacao*. Surgirá uma nova janela ilustrada na figura 5.3.

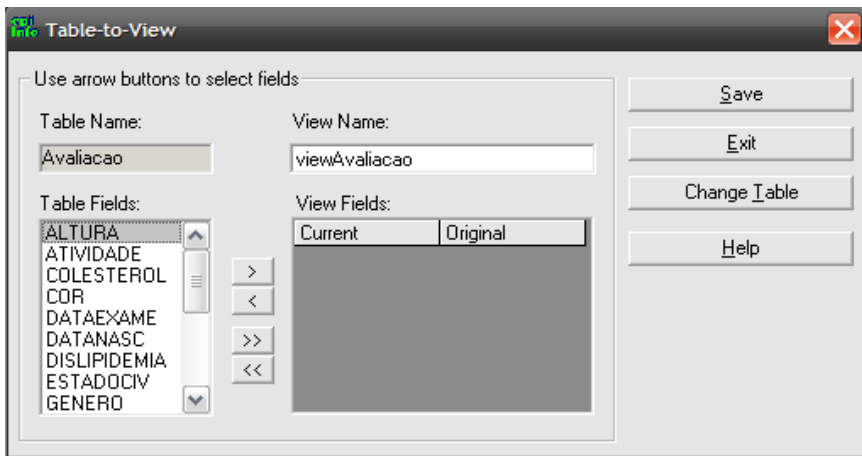


Figura 5.3. Janela de criação de *ficha* a partir de uma *tabela*.

Fonte: Epi Info®.

Para incluir todas as variáveis na nossa ficha devemos clicar em >>. Finalmente clicamos em *Save*. Às vezes, por algum motivo, esse comando não funciona, isso tem ocorrido principalmente em computadores que estão interligados por rede (sistema de Network). Quando isso acontece não podemos criar a ficha por esse método. Esse problema não tem ocorrido com a versão atual do programa. Mas, na impossibilidade de criar uma, podemos trabalhar com um arquivo mesmo sem ficha.

## Observando frequências de variáveis (comando *Frequencies*)

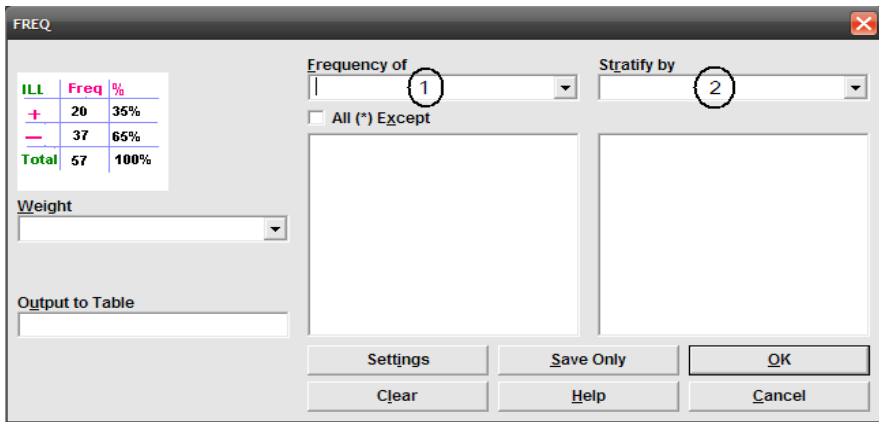


Figura 5.4. Janela do comando *Frequencies* destacando os compartimentos *Frequency of* e *Stratify by*.  
Fonte: Epi Info®.

Na secção *Statistics* do *Analysis*, com o comando *Frequencies*, que é o segundo comando, podemos observar a frequência de distribuição das variáveis de tipo texto. Na realidade poderíamos ver a frequência das respostas de todas as variáveis, mas as variáveis numéricas contínuas são pouco apropriadas para isso. Como elas são contínuas, a maioria dos valores vai ser única. Valores como 12 e 12,01 serão caracterizados como valores distintos e frequências computadas distintamente. Assim, frequências são instrumentos de análise descritiva para as variáveis de tipo texto (também chamadas categóricas). No exercício, vamos calcular as frequências das variáveis: *ESTADOCIV* e *COR*. Na janela do *Frequencies* (figura 5.4), observamos dois compartimentos importantes. No compartimento do *Frequency of* (1) procuramos as variáveis cujas frequências queremos calcular e no compartimento *Stratify by* (2) podemos selecionar a(s) variáveis que desejamos estratificar. Inicialmente, vamos calcular somente as frequências das variáveis, não estratificando. E, depois, calcularemos as frequências de maneira estratificada

para o gênero. Estratificar significa calcular as frequências, da variável a ser observada, para cada valor da variável a ser estratificada. Assim nosso resultado trará distribuição da frequência para cada valor de *GENERO*.

## Estrutura da Janela de Resultado (*OUTPUT*) do Epi Info®

Observamos na figura 5.5 a janela *Output* com os resultados do comando *Frequencies*.

C:\Epi\_Info\EpiAula\OUT1.htm

Previous Next Last History Open Bookmark Print

### Epi Info

[Results Library](#)  
 Current View: C:\Epi\_Info\EpiAula\Avaliacao.xls:AvaliacaoS  
 Record Count: 66 Date: 17/03/2010 16:20:29

**FREQ COR ESTADOCIV STRATAVAR = GENERO**

[Next Procedure](#)  
[COR, GENERO=Feminino](#)  
[COR, GENERO=Masculino](#)  
[ESTADOCIV, GENERO=Feminino](#)  
[ESTADOCIV, GENERO=Masculino](#)

**COR, GENERO=Feminino**

[Forward](#)

COR	Frequency	Percent	Cum Percent
Branca	29	82,9%	82,9%
Mulata	4	11,4%	94,3%
Preta	2	5,7%	100,0%
<b>Total</b>	<b>35</b>	<b>100,0%</b>	<b>100,0%</b>

**95% Conf Limits**

Branca 66,4% 93,4%  
 Mulata 3,2% 26,7%  
 Preta 0,7% 19,2%

Figura 5.5. Janela de resultados do *Analysis* do Epi Info® destacando o caminho e nome do arquivo e a tabela criada pelo comando *Frequencies*.

Fonte: Epi Info®, arquivo *Avaliacao.mdb*.

O Epi Info® salva os resultados em um arquivo. Observamos que, na barra mais superiormente colocada (normalmente de coloração azul) existe o caminho (ou endereço) e nome do arquivo, que será, por definição, *OUTNN.htm*, onde NN serão números sequenciais conforme o número de arquivos

existentes no nosso diretório de trabalho. A extensão *.htm* denomina o formato do arquivo gerado. No caso é um arquivo com formato de página de Internet. A janela *Output* informa o comando solicitado e depois a tabela. A tabela do comando *Frequencies* dispõe os valores da variável na primeira coluna, em ordem alfabética. A segunda coluna dispõe as frequências e as próximas mostram a percentagem de cada categoria e a percentagem acumulativa. Na última coluna, o programa cria um tipo de gráfico de barra que expressa a distribuição das frequências. Finalmente calcula um intervalo de confiança.

## Comprimindo o Arquivo de Dados

É importante destacar que, assim como no *MS-ACCESS*, o banco de dados do Epi Info® pode ser comprimido para salvar espaço. Isso não prejudica em nada o arquivo e nos ajuda, principalmente, quando estamos trabalhando em um disquete ou em um disco rígido com falta de espaço ou ainda desejamos transportar o arquivo. Algumas vezes podemos corrigir problemas no nosso banco de dados através desse programa. As versões anteriores do Epi Info® (versão 1 e 2) possuíam dois programas para compactar o banco de dados: um executado a partir do menu *File* do *Enter data (Entrar dados)*, onde tínhamos a opção *Compress database (Compactar banco de dados)*; e a opção *Compact & Repair MDB Utility... (Compactar e Reparar MDB)* do programa *Visualize data (Visualizar dados)* que fica no menu *Utilities (Ferramentas)* da tela principal do Epi Info®. O primeiro programa, também chamado simplesmente de *Compact*, às vezes não funcionava, o que fazia desaparecer o arquivo compactado, nesse caso era preferível utilizar o segundo programa (*Database Compact Utility*). A versão 3.2 possui somente um programa de compactação, justamente o *Compact* que parece funcionar adequadamente. Esse programa pode ser executado tanto pelo *Enter data (Entrar dados)* quanto pelo *Visualize data*, ou ainda no menu *Utilities (Ferramentas)* que possui a nova opção *Compact*. Vamos ver como funciona e também como evitar maiores problemas ao usar esse programa. A partir do programa *Visualize data (Visualizar dados)*, da tela principal do Epi Info®. No menu *File* do programa, escolhe-se a opção *Compact & Repair MDB Utility... (Compactar e Reparar MDB)*. Onde aparecerá a janela exposta na figura 5.6.



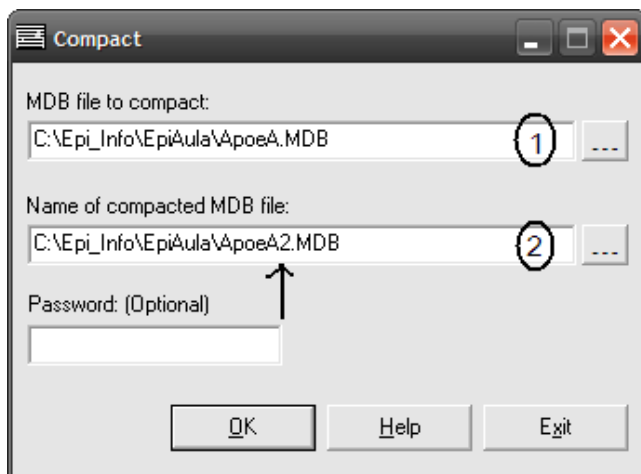


Figura 5.6. Janela do comando *Compact* destacando os compartimentos de denominação dos arquivos a compactar e compactado.

Fonte: Epi Info®.

Ao lado do primeiro compartimento temos o ícone (1) que clicamos para escolher o arquivo que desejamos. No compartimento abaixo (2) escolhemos primeiro o mesmo arquivo e clicamos *Abrir*, depois, na tela do *Compact*, alteramos o nome, por exemplo, adicionando uma letra, conforme a seta ilustrada na figura 5.6. É importante trocarmos o nome, pois se o programa não funcionar direito, poderemos perder os nossos dados para sempre!

Após isso podemos clicar em *OK*. O programa, então, mostra uma caixa informando que o processo foi executado. Antes de sair do programa, devemos confirmar se o novo arquivo foi realmente criado, clicando em um dos ícones de seleção, para ver se o arquivo aparece na tela. Após verificar isso, podemos sair do programa clicando em *Exit*. Assim podemos “deletar” o arquivo antigo, após termos certeza de que o novo pode ser utilizado sem problemas.

## Exercício:

1. Importe o banco de dados *Apoe4.REC* em versão *EPI 6*, que se encontra no folder *Exercício* do site [www.pucrs.br/igg/epiinfo](http://www.pucrs.br/igg/epiinfo), utilizando o programa *Analyze Data*. Crie e defina as variáveis: *OBESIDADE*, *DISLIPIDEMIA*, *HIPERTENSAO*, usando o programa “*classificacao*” que está no arquivo *Avaliacao.mdb*.
2. Salve o arquivo como *Apoe4.MDB* tabela *Grupo1* (sem espaço).
3. Crie uma ficha para o projeto, utilizando o *Makeview*.
4. Calcule as frequências de *HIPERTENSAO*, *DISLIPIDEMIA* e *OBESIDADE*.
5. Comprima o arquivo como *Apoe.mdb*.

Obs.: Como vamos utilizar esse arquivo nos próximos capítulos, preste atenção onde o mesmo será salvo.

## Capítulo 6

# ADICIONANDO DADOS EPIDEMIOLÓGICOS AO EPI MAP

Vimos, anteriormente, como criar um mapa usando o Epi Map. Naquele momento criamos um mapa baseado nos dados originalmente fornecidos. Vamos ver um exemplo de como adicionar dados novos ou que nos interessem.

### **Como obter dados sobre saúde do Brasil ([www.datasus.gov.br](http://www.datasus.gov.br))**

O Ministério da Saúde do Brasil mantém um site onde estão acessíveis vários dados sobre o país. Esse site é chamado *DATASUS*. Nele podemos encontrar vários dados, não só demográficos, mas socioeconômicos, e, evidentemente, de saúde, como mortalidade e morbidade. Mortalidade se refere às mortes que ocorreram em determinado período, enquanto morbidade refere-se ao número de pessoas com determinada doença em um determinado momento (prevalência) ou o número de casos novos que ocorreram em um período de tempo (incidência). Normalmente, mortalidade e morbidade são expressas em taxas, onde temos o número de pessoas com a característica, dividido pelo número total de pessoas. No *DATASUS* podemos obter o número de mortes por determinada doença tanto no nível de país, como um todo, por Unidades da Federação ou por cidades. Faixas etárias e gênero são outras variáveis disponíveis para cruzamento. Também está acessível a população estimada ou recenseada com todas as variáveis anteriores disponíveis, exceto causa morte, desde o ano de 1980. Assim, para calcularmos a taxa de mortalidade específica para determinada doença devemos baixar tanto o número de mortes quanto a população do ano ou anos desejados. Algumas taxas de morbidade e mortalidade são acessíveis diretamente no site do *DATASUS*, dispensando então o cálculo mencionado. Entretanto, nem sempre teremos disponível a taxa da doença e cruzamento desejado. Por isso veremos como obter tanto o número de mortes quanto a população, importando para o Epi Info® onde calcularemos as taxas de mortalidade e, finalmente, transportaremos os dados para o Epi Map onde criaremos um mapa desses valores.

## Modificando a linguagem do sistema do computador para importarmos dados do DATASUS (Control Panel)

Antes de abrirmos o navegador (*browser*), devemos modificar ou confirmar se a linguagem do sistema do computador é o português. Para tanto, devemos clicar no botão *Iniciar* (*Start*), que é o mais inferiormente colocado no nosso computador, à esquerda. Temos a opção *Painel de Controle* (*Control Panel*), conforme a figura 6.1.

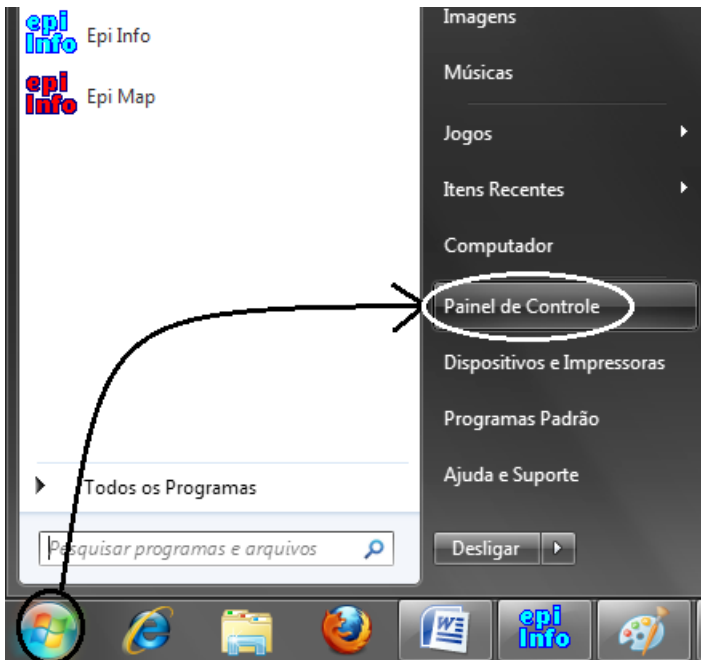


Figura 6.1. Detalhe da tela principal do ambiente *Windows* demonstrando o caminho para a localização do comando *Painel de Controle* – *Control Panel*.  
Fonte: Windows 7.

Ao abrirmos o *Painel de Controle*, devemos procurar a opção *Relógio Idioma e Região* que levará à opção *Região e Idioma*, que é um globo com um relógio na frente (figura 6.2).

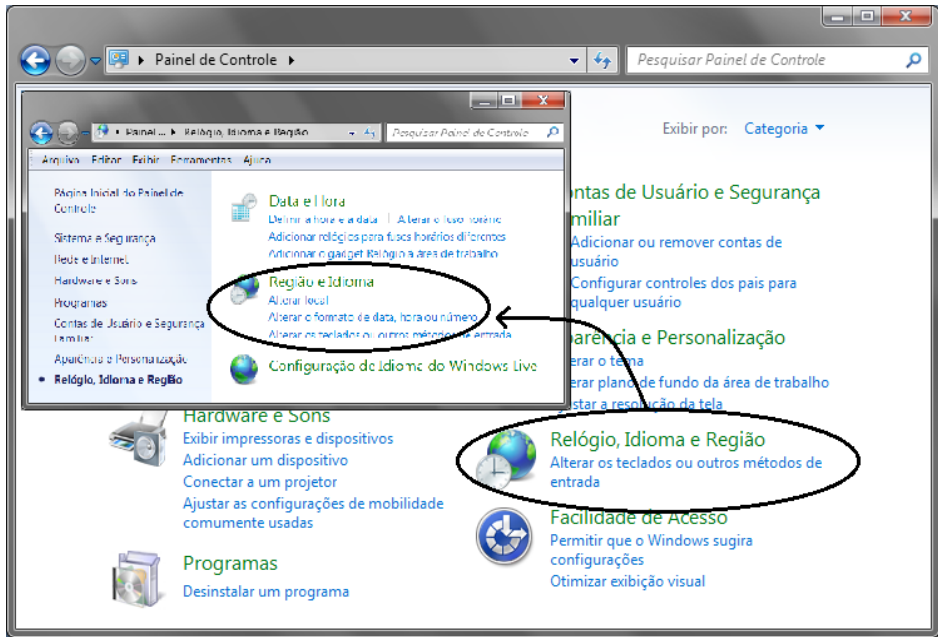


Figura 6.2 Painel de Controle mostrando o ícone da opção *Região e Idioma*.

Fonte: Windows 7.

Ao clicarmos duas vezes no ícone, aparecerá uma janela onde escolheremos o formato de uma das línguas portuguesas, que pode ser tanto a padrão quanto a brasileira (figura 6.3). O computador talvez pergunte se queremos reiniciá-lo. Se isso acontecer, devemos fechar os programas abertos e salvar os documentos que estamos usando, se não quisermos perder o trabalho até então realizado. Note que a linguagem portuguesa pode atrapalhar cálculos de variáveis com casas decimais, realizados pelo Epi Info<sup>®</sup>. Por isso, devemos modificar a linguagem para o inglês quando quisermos, por exemplo, calcular médias de índice de massa corporal (ICM), que normalmente é definido com duas casas decimais.

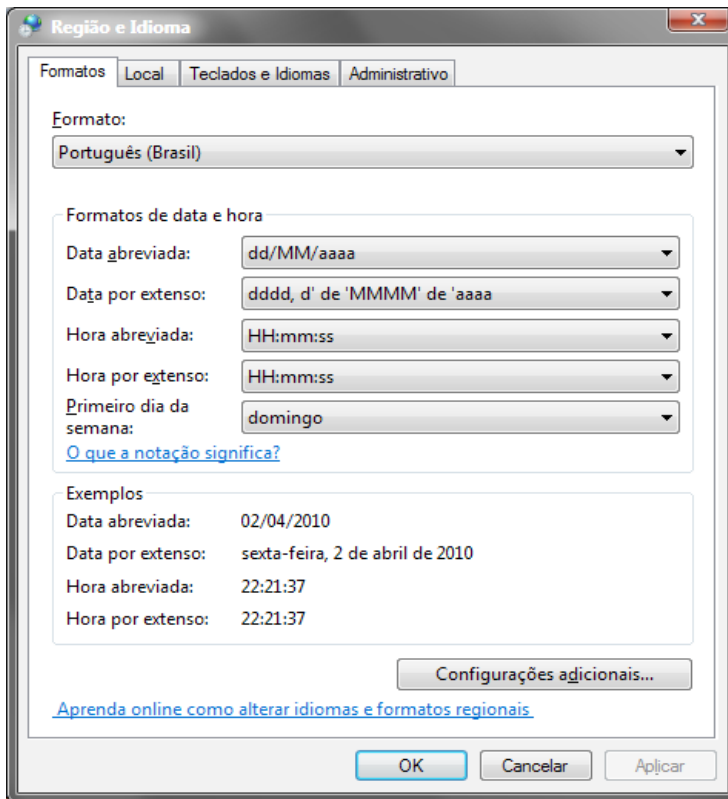
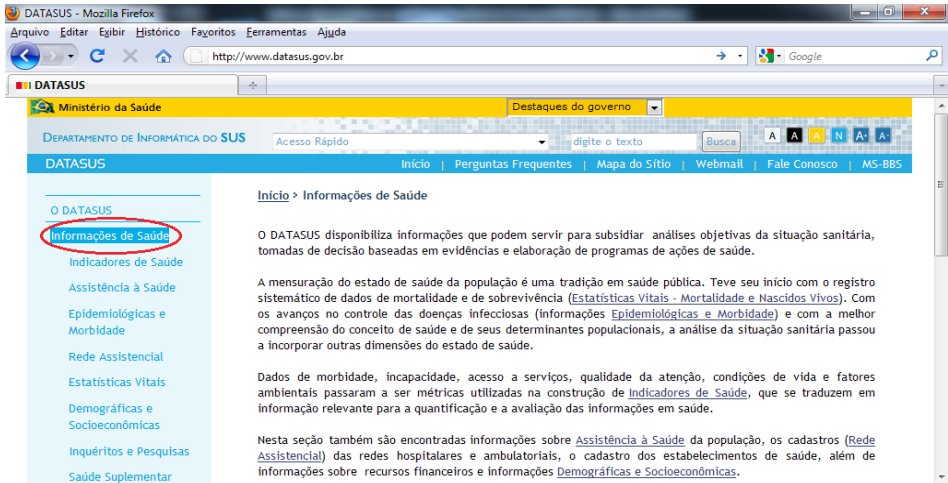


Figura 6.3 Janela da *Região e Idioma* com o formato para Português (Brasil) selecionado.  
Fonte: Windows 7.

## Visitando o site do DATASUS

O endereço do site do *DATASUS* é: [www.datasus.gov.br](http://www.datasus.gov.br).



No menu da esquerda temos várias opções, a que escolhermos será *Informações de Saúde*. Para o nosso exercício, vamos utilizar, inicialmente, os dados de mortalidade. Para tanto, devemos clicar em *Estatísticas vitais*. Isso abre várias opções, entre elas a opção *Mortalidade Geral*. Ao clicarmos em cima dessas palavras, aparecerá um mapa do Brasil como na figura.



Os *Links* superiores se referem a tabulações em âmbito nacional. Podemos selecionar dados de um estado clicando no mesmo, tanto no mapa quanto no nome na aba que se abre ao clicarmos na caixa de seleção. Vamos selecionar *Brasil por Região e Unidade da Federação*. Isso nos levará à página de construção da nossa tabela.

## Construindo uma tabela de mortalidade no DATASUS

Figura 6.3. Detalhe da página de tabulação dos dados de mortalidade do site do *DATASUS*.  
 Fonte: <http://tabnet.datasus.gov.br/cgi/deftohtm.exe?sim/cnv/obtuf.def>.

A página mostra três janelas superiores horizontalmente dispostas e 16 outras janelas verticalmente dispostas. Nas primeiras (figura 6.3) determinamos quais valores queremos nas linhas e na coluna da tabela que desejamos. Nesse caso, escolheremos *Ano do óbito* para linha e *Sexo* para Coluna. Na janela chamada *Conteúdo*, temos duas opções: óbitos por residência e óbitos por ocorrência. Às vezes uma pessoa está viajando ou é hospitalizada em um Estado da Federação, que não é o que ela reside, mas acaba falecendo nesse Estado; nesse caso, o óbito por residência será diferente do óbito por ocorrência. Na maioria das vezes, o óbito por residência será utilizado. O óbito por ocorrência será importante quando temos uma causa aguda de morte por atropelamento. Como em idosos a maioria dos óbitos é por doenças crônicas, não importa muito onde a pessoa morreu porque, provavelmente, ela desenvolveu a doença no Estado de residência. Assim, manteremos a opção *Óbitos p/Residência* nessa janela. Na próxima janela, *Períodos disponíveis*, vamos escolher todos os anos disponíveis. Podemos selecionar um ou mais regiões ou Estados nas primeiras janelas e as causas de óbito. Desde 2010, o site disponibiliza somente a partir de 1996, desde quando os óbitos são classificados utilizando o CID 10.



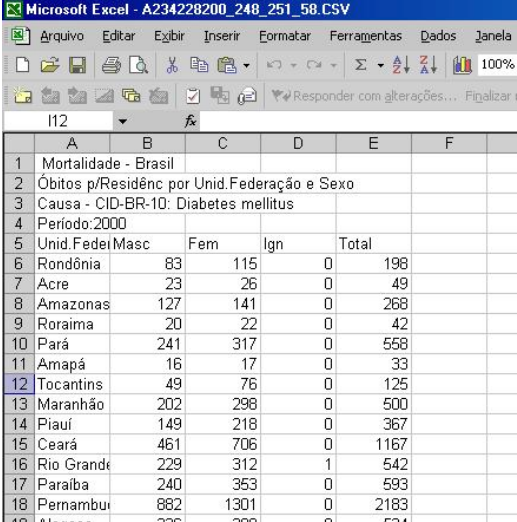
Região	Todas as categorias Região Norte Região Nordeste Região Sudeste
Unid.Federação	Todas as categorias Acre Alagoas Amapá
Capítulo CID-10	Todas as categorias I. Algumas doenças infecciosas e parasitárias II. Neoplasias (tumores) III. Doenças sangue órgãos hemat e transt imunitár
Grupo CID-10	Todas as categorias Doenças infecciosas intestinais Tuberculose Algumas doenças bacterianas zoonóticas
Categoria CID-10	Todas as categorias A00 Colera A01 Febres tifoide e paratifoide A02 Outr infecc p/Salmonella
Causa - CID-BR-10	Todas as categorias 001-031 ALGUMAS DOENÇAS INFECCIOSAS E PARASITÁRIAS . 001 Doenças infecciosas intestinais ... 002 Cólera

Figura 6.4. Detalhe da página de tabulação dos dados de mortalidade do site do *DATASUS*, continuação.  
Fonte: <http://tabnet.datasus.gov.br/cgi/deftohtm.exe?sim/cnv/obtuf.def>, última visita em abril, 2011.

Nas próximas janelas, observadas na figura 6.4, selecionamos alguns valores, como local, causa morte, faixa etária e sexo. Existem quatro janelas para selecionarmos uma causa morte. Na janela *Capítulo CID-10* podemos selecionar os capítulos da Classificação Internacional das Doenças (CID). Podemos escolher um tipo de óbito específico usando uma das duas janelas seguintes, obedecendo ao mesmo critério, conforme o ano de óbito desejado. Como desejamos o número de diabetes em 2000, devemos escolher essa doença na janela *Causa – CID-BR-10*. Observe que as outras janelas correspondentes à causa de morte devem estar na primeira opção: *Todas as categorias*. Nem todas as causas de morte estão disponíveis. Podemos ainda selecionar somente uma faixa etária ou sexo, se desejássemos. Após selecionar *Unid. Federação* na linha, *Sexo* na coluna e *Diabetes mellitus* na *Causa – CID-BR-10*, podemos clicar no botão *Mostra* no final da página. Após algum tempo, dependendo do congestionamento da Internet, obteremos a tabela com as informações desejadas. A tabela é precedida de uma descrição de todas as seleções realizadas. Devemos confirmar se todos os nossos critérios, e somente eles, estão mencionados. Confirmando que estamos com a tabela desejada, podemos clicar no *Link: Cópia como CSV*, que significa “Comma Separated Variables”. Na realidade, o programa separa as variáveis com um ponto e vírgula (;) e não com vírgula como a abreviação sugere. Ao clicarmos no *Link*,

provavelmente abrirá uma janela perguntando se desejamos salvar ou abrir o arquivo. Para facilitar nossa vida, vamos abrir o arquivo. Às vezes, o programa abre automaticamente o arquivo sem nos perguntar. Alguns *navegadores* (*browsers*) necessitam que escolhamos o programa a ser utilizado, esse será o *Excel*. Ao abrir o arquivo, automaticamente inicia o programa *Excel*, mostrando as colunas da tabela dispostas em colunas separadas na planilha, conforme figura. Com o arquivo aberto no *Excel*, devemos transformar o arquivo para poder ser importado pelo Epi Info®, conforme vimos no capítulo 5. Se as colunas não estão dispostas conforme a figura é porque a linguagem do sistema do computador não foi alterada para o português, conforme anteriormente mencionada. Nesse caso devemos salvar como arquivo tipo texto e extensão *.txt*, para depois abrirmos novamente com o *Excel* formato tipo texto, opção campo delimitado, clicar em próximo, delimitador “ponto e vírgula” (;), concluir.

## Preparando o Arquivo do Excel para ser Importado pelo Epi Info



	A	B	C	D	E	F
1	Mortalidade - Brasil					
2	Óbitos p/Residênc por Unid.Federação e Sexo					
3	Causa - CID-BR-10: Diabetes mellitus					
4	Período:2000					
5	Unid.Fed	Masc	Fem	Ign	Total	
6	Rondônia	83	115	0	198	
7	Acre	23	26	0	49	
8	Amazonas	127	141	0	268	
9	Roraima	20	22	0	42	
10	Pará	241	317	0	558	
11	Amapá	16	17	0	33	
12	Tocantins	49	76	0	125	
13	Maranhão	202	298	0	500	
14	Piauí	149	218	0	367	
15	Ceará	461	706	0	1167	
16	Rio Grande	229	312	1	542	
17	Paraíba	240	353	0	593	
18	Pernambuco	882	1301	0	2183	

Figura 6.5. Detalhe da tela do *Excel* mostrando a disposição original dos dados criados pela página Mortalidade do *DATASUS*.

Fonte: Microsoft Excel, arquivo importado do *DATASUS*.

Para importar o arquivo usando o Epi Info®, a primeira linha da planilha deve conter somente o nome das variáveis. Devemos, então, excluir as primeiras linhas da planilha que possuem as descrições da tabela. Para fazer isso podemos selecionar as linhas clicando com o mouse no número que denomina a linha. Com as linhas que desejamos excluir selecionadas,

podemos pressionar as teclas *Ctrl* – (menos) ou clicarmos no menu *Editar* (*Edit*), opção *Excluir* (*Delete*). Aparecerá uma janela pequena perguntando o que se deseja excluir, escolhemos *linha inteira* e, depois, *OK*. No caso do comando *Ctrl* –, não necessitaremos confirmar, se selecionarmos a linha corretamente. Nossa planilha terá, então, o título (rótulo) das colunas na primeira linha e os demais valores nas outras linhas. Temos, muito frequentemente, uma coluna com o sexo ignorado. Essa informação não nos é importante e podemos ou não excluí-la. Para tanto repetimos, mais ou menos, o mesmo procedimento anterior: selecionamos a coluna clicando na letra que a denomina e pressionamos *Ctrl* –. No mapa fornecido pelo site do Epi Info®, as Unidades da Federação estão escritas sem acento. Se quisermos utilizar esse mapa, devemos tirar todos os acentos.



Figura 6.6. Detalhe da tela do *Excel* mostrando o nome da planilha já corrigido.  
Fonte: Microsoft Excel.

Após retirar todos os acentos das Unidades da Federação, temos mais algumas etapas importantes. Notamos, na parte inferior do *Excel*, que o nome da planilha está com o nome do arquivo que contém um conjunto de letras e números sem sentido. Esse nome foi denominado pelo *DATASUS*, mas não nos ajuda a identificar o conteúdo, por isso devemos modificá-lo. Clicando duas vezes, selecionamos o nome, podemos modificá-lo para “obitos” (sem acento) conforme a figura 6.6. As últimas alterações que devemos fazer são modificar os rótulos das colunas, entre outros motivos, para podermos, depois, nos lembrar do que elas significam. Como estamos trabalhando com óbitos, podemos acrescentar *Ob* aos rótulos *Masc*, *Fem* e *Total*, ficando: *ObMasc*, *ObFem*, *ObTotal*, por exemplo. Note que é importante não termos espaços nos nomes das colunas, pois elas tornar-se-ão nomes de variáveis no Epi Info®. Devemos, também, modificar o nome da primeira coluna que denomina as Unidades de Federação tirando o ponto que está no meio, passando, assim a se chamar: *UnidFederação*. Note que não necessitamos transformar os nomes das variáveis para maiúsculas, pois o Epi Info® não faz essa distinção.

## Salvando arquivo em formato Excel

Finalmente, podemos salvar o arquivo, no menu *Arquivo (file)*, clicar em *Salvar como (Save as)*. Na janela colocaremos como nome, por exemplo: *Óbitos Diabete BR 2000*. Devemos modificar o formato do arquivo para *Pasta de trabalho do Excel*, conforme figura 6.7.

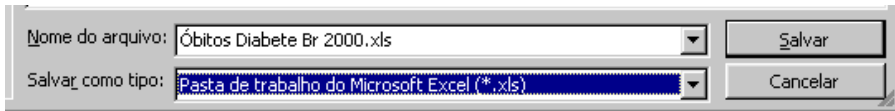


Figura 6.7. Detalhe dos compartimentos de denominação e alteração do tipo de arquivo da janela de salvamento do *Excel*.

Fonte: Microsoft Excel.

Não se esqueça de salvar o arquivo no diretório de trabalho. Observe que o nome do arquivo pode ter espaço e acento, mas o nome da planilha não.

## Obtendo dados da população

Como vimos, necessitamos do número da população brasileira para cada Estado, de ambos os sexos e ano em que temos os óbitos. Para tanto devemos retornar à janela *Informações de Saúde*, retornando três páginas. As informações sobre a população estão na sessão *Informações demográficas e socioeconômicas*. Existem dois *Links* para a população, escolhemos o primeiro, clicando no mesmo, *População residente*, que tem informações desde 1980. A janela que se abre ao clicarmos no *Link* é muito parecida com a janela do óbito, com o mesmo mapa e opções. Clicando em *Região e Unidade da Federação*, abrimos uma janela, também semelhante à janela em que construímos a tabela dos óbitos, tendo no seu conteúdo apenas a população residente. Devemos selecionar os mesmos valores para linha (*Unidade da Federação*) e coluna (*Sexo*); observe que devemos selecionar também o mesmo ano (*2000*) nos *Períodos disponíveis*. Não precisamos fazer mais nenhuma seleção, pois estamos trabalhando com toda a população. Clicando, então, no botão *Mostra*, criamos a tabela que deve ter mais ou menos o mesmo formato da tabela anterior, só que com valores maiores. Clicando no *Link copia como CSV*, selecionamos a opção de abrir o arquivo, que nos retorna ao Excel com as colunas bem definidas. Novamente devemos excluir as linhas de descrição da tabela, deixando somente os rótulos das colunas. Devemos também retirar os acentos das Unidades da Federação. Sugere-se modificar também o nome dos rótulos

para *PopMasc*, *PopFem* e *PopTotal*, para não confundirmos com os dados do óbito. O rótulo *Unid.Federação* deve ser mudado para *UnidFederação*, sem o ponto no meio. Mudamos também o nome da planilha para *População*. E, finalmente, salvamos com o nome *População BR 2000*, não esquecendo de modificar o formato do arquivo para *Pasta de trabalho do Excel*. É recomendável salvar no mesmo diretório que o arquivo do óbito. Feito isso, fechamos o *Excel* e abrimos o Epi Info®.

### **Juntando dois arquivos com variáveis diferentes (comando *Relate* – Relacionar)**

Vimos que, quando temos dois arquivos com as mesmas variáveis, podemos juntá-los com o comando *Merge* (*Juntar*). No comando *Merge* (*Juntar*) somente os dados das variáveis comuns são juntados. Quando temos dois arquivos com variáveis diferentes e queremos juntar essas variáveis em um mesmo arquivo, usamos o comando *Relate* (*Relacionar*).

O comando *Relate* (*Relacionar*) é muito semelhante ao do *Merge* (*Juntar*), primeiro devemos abrir o arquivo que receberá as variáveis novas. No caso vamos abrir o arquivo *Óbitos Diabete BR 2000.xls*, usando o comando *Read* (*Abrir* ou *Ler*), sem esquecer que o arquivo está em formato *Excel* e que temos que selecionar a planilha. Confirmando que a primeira linha possui o nome das variáveis, clicamos *OK* e observamos que o nosso arquivo possui 28 observações: 26 Estados, Distrito Federal e Total. Imediatamente clicamos em *Relate* (*Relacionar*). Observamos que a janela é também muito semelhante à do *Merge* (*Juntar*). Trocando o *Data Format* (*Formato do banco de dados*) para *Excel 8.0*, selecionamos o arquivo *População BR 2000.xls*. Selecionando a planilha *Populacao*, devemos construir a chave de ligação dos dois arquivos através do *Build Key* (*Construir chave*). Escolhemos em ambos os arquivos a variável *UnidFederação*. Nosso comando deve ficar mais ou menos conforme a figura 6.7.

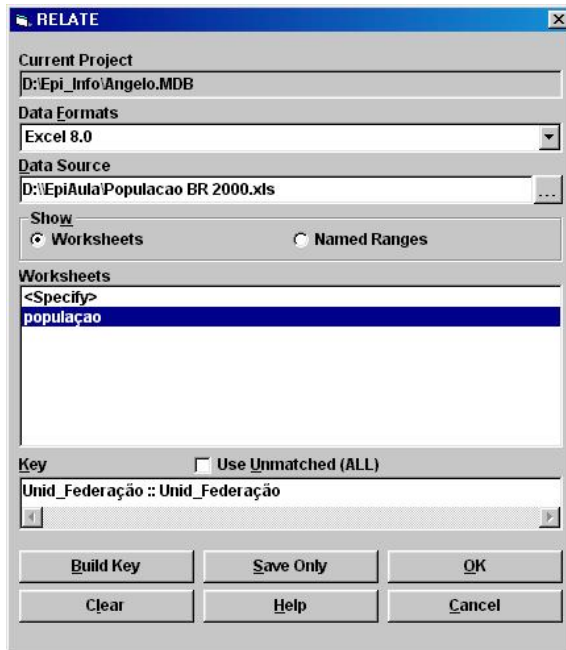


Figura 6.7. Janela do comando *Relate* do *Analysis*.  
Fonte: Epi Info®.

Ao clicarmos *OK*, o programa nos pergunta sobre um *Link* permanente dos arquivos. Como iremos salvar o arquivo em seguida, simplesmente clicamos *OK*, sem acrescentar nenhum nome. Completada essa etapa, poderemos salvar o nosso projeto como *Óbitos BR 2000*, tabela *Diabetes*, usando o comando *Write (Salvar)*, não esquecendo de selecionar a opção *Replace (Substituir)*.

Devemos agora definir as variáveis que conterão as taxas: *TxMasc*, *TxFem*, *TxTotal*, por exemplo. Devemos usar o comando *Define (Definir)*, evidentemente.

## Calculando taxas de mortalidade

As taxas de mortalidade são normalmente calculadas para cada mil, 10 mil ou até 100 mil habitantes. Esses valores variam conforme a raridade da doença: quando temos uma doença com mortalidade baixa, menos de uma morte para 10 mil habitantes, usaremos o valor mortes por 100 mil para denominar a taxa. O nosso caso é uma doença bastante comum, temos mais de 35 mil mortes no ano, para uma população brasileira de 170 milhões, poderemos usar a denominação de mortes por 10 mil habitantes. Usando o *Assign*, teremos nosso comando, demonstrado na figura 6.8.

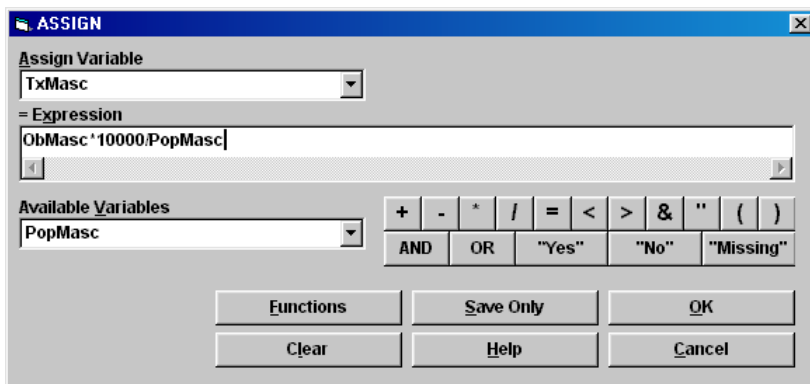


Figura 6.8. Janela do comando *Assign* exemplificando a fórmula para o cálculo da taxa de mortalidade. Fonte: Epi Info®.

Devemos fazer o mesmo para *TxFem* e *TxTotal*.

Salvamos novamente o nosso projeto com o mesmo nome, não esquecendo de selecionar *Replace (Substituir)*.

## Adicionando dados ao Epi Map

Vimos como importar informações de saúde do *Datasus* para o Epi Info®. Apesar de existir um comando *Map* no programa *Analysis* do Epi Info®, esse comando muitas vezes não funciona. Por isso, é recomendável abrir o mapa através do Epi Map. Abrindo o programa e com o *Layer (Camada)* já contendo o mapa do Brasil (*BR.shp*), clicamos no botão *Add data (Adicionar dados)*. Abrimos, então, o projeto *Óbitos BR 2000.mdb*, que contém os dados que desejamos. Após pensar um pouco, o programa abre uma janela onde escolheremos as variáveis que têm os valores que vão juntar os arquivos, no caso o nome das Unidades da Federação (primeira janela em inglês: *ADMIN\_NAME* e na segunda janela em português: *UnidFederação*), assim como o nome das variáveis que serão adicionadas, conforme a figura 6.9.

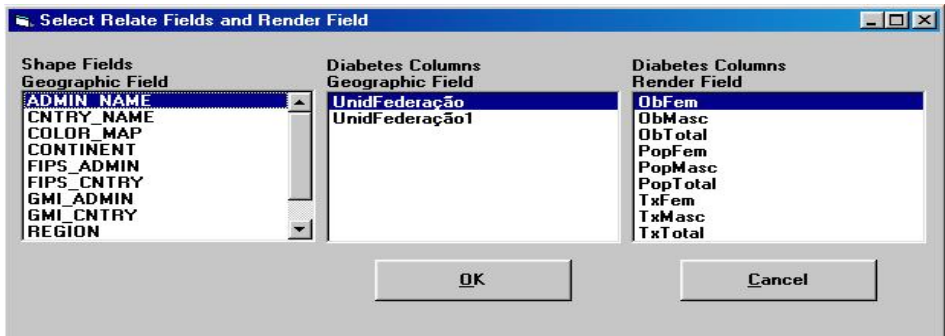


Figura 6.9. Janela de seleção dos campos de relacionamento e variáveis do comando *Add data* do Epi Map.  
Fonte: Epi Info®.

O programa automaticamente seleciona as primeiras variáveis de tipo texto, em ordem alfabética. Casualmente são justamente as variáveis que desejamos juntar. Caso as variáveis que contêm os nomes dos Estados não estejam selecionadas, devemos corrigir isso, selecionando o nome correto.

A seguir o programa informa valores das duas variáveis que não encontraram nomes semelhantes (figura 6.10). No caso, temos quatro zonas de litígio que existem entre os Estados da Paraíba e Piauí, e entre o Estado do Amazonas e do Pará. No arquivo que criamos temos o total das taxas de mortalidade que também não encontra semelhante no mapa.

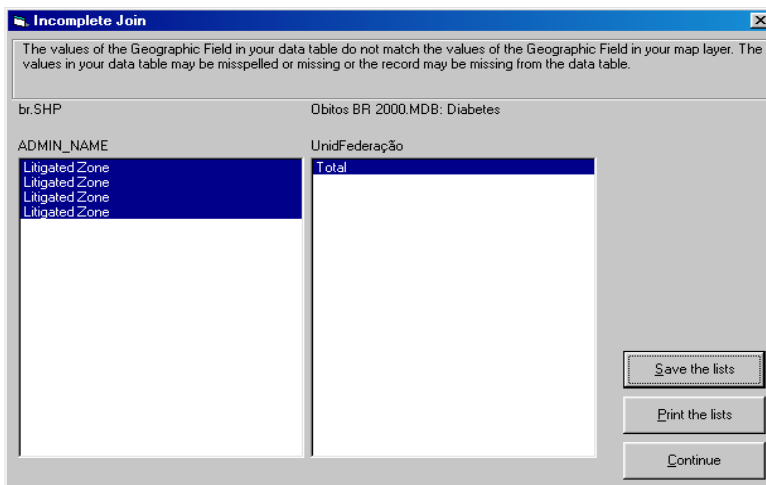


Figura 6.10. Janela de aviso dos campos que não encontraram correspondentes nos dois arquivos, mapa e banco de dados novo.  
Fonte: Epi Info®.



Como esses valores não atrapalham nosso resultado, simplesmente os ignoramos. É importante observar, entretanto, se existem Estados que não encontraram correspondentes nos dois arquivos. Isso pode ocorrer se, por exemplo, um Estado possui acento, como o da Paraíba, cujo acento nem sempre é bem visível. Se ocorrer uma incompatibilidade de nomenclatura entre Estados, eles aparecerão em ambas as janelas. Nesse caso deveremos encerrar o programa e alterar o nome do arquivo criado por nós, adequando o nome do(s) Estado(s) com problema. Independente disso, clicamos em *Continue* para fechar a janela. Se não tivermos nenhum problema, observamos o mapa com a primeira variável do arquivo (ObFem), já com a distribuição por faixas (*Choropleth*) criada. Podemos salvar todo mapa ou cada um dos mapas.

## Exercício:

1. Obtenha dados sobre mortalidade geral e população das Regiões do Brasil para o ano de 2000. Coloque as Regiões na linha e sexo na coluna.
2. Prepare os arquivos para serem importados para o Epi Info® e retire a palavra “Região” antes dos nomes das regiões.
3. Troque o nome das colunas para *obitomasc*, *obitofem* e *obitotot*, no arquivo dos óbitos e para *popmasc*, *popfem* e *poptotal* no arquivo da população.
4. Modifique os nomes das planilhas para *obitos* e *populacao* (sem acento) e salve como pasta do Excel como *Óbito Região BR 2000* e *População Região BR2000* respectivamente
5. Importe o arquivo dos óbitos para o Epi Info® e depois relacione (juntar pelo *Relate*) com a população.
6. Calcule a taxa de mortalidade para cada sexo e total.
7. Salve o projeto como *Óbitos BR 2000* e tabela *RegiaoGeral*.
8. Abra o mapa *Regiões.map*, criado no primeiro capítulo.
9. Adicione dados a partir do projeto *Óbitos Br* tabela *RegiaoGeral*.
10. Faça o mapa Choropleth da Mortalidade Total.

## Capítulo 7

# CRIANDO GRÁFICOS COM O COMANDO GRAPH

Finalmente chegamos ao comando tão esperado de criação de gráficos pelo Epi Info®. Existem muitos tipos de gráficos, por isso vamos ver como criamos alguns deles. Primeiro vamos ter uma visão geral sobre a janela do comando observando a figura 7.1.

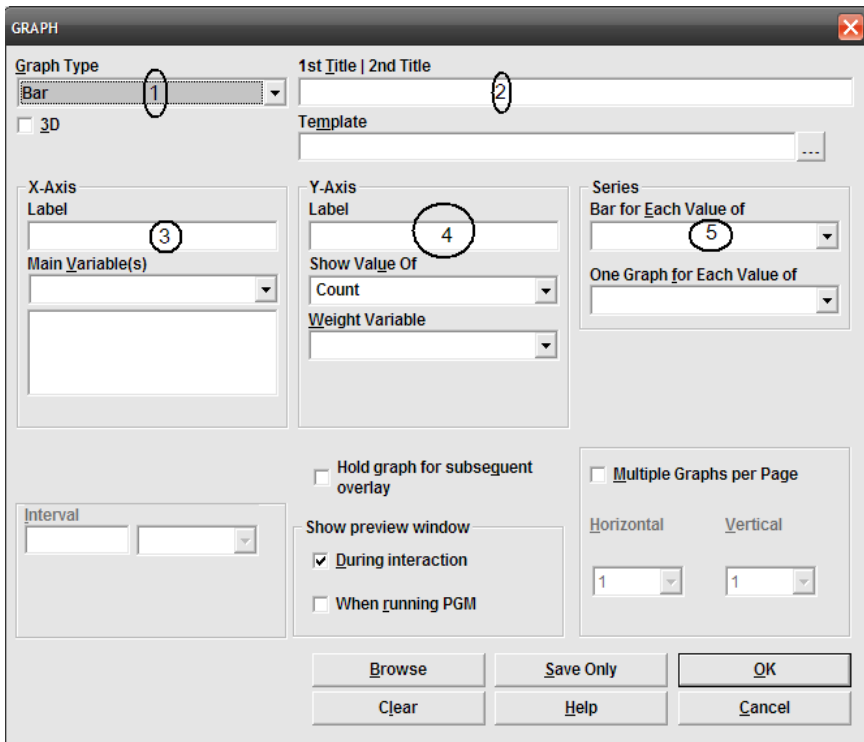


Figura 7.1. Janela do comando *Graph* do *Analysis* destacando os compartimentos principais.

Fonte: Epi Info®.

No compartimento *Graph Type* (1) escolhemos o tipo de gráfico que desejamos, que automaticamente nos dá a opção de gráfico tipo barra; colocamos depois um título que é opcional (2). Logo abaixo do compartimento do tipo de gráfico, existe a opção de criarmos gráficos

tridimensionais. Mais importante é colocarmos a variável que desejamos no eixo do X (3) e, depois, escolhemos que valor vamos colocar no eixo do Y (4). Podemos determinar algumas funções gráficas incluindo uma outra variável no item (5), mas esse comando modificar-se-á conforme o tipo de gráfico que desejamos. Existem várias características de edição dos gráficos, algumas comuns outras específicas de determinado tipo de gráfico. Assim, no primeiro tipo de gráfico, descreveremos as características tanto comuns quanto próprias desse tipo, ressaltando quando a mesma é própria do tipo em questão. Na sequência, descreveremos somente as características próprias de cada gráfico ou as nuances particulares de cada tipo de gráfico.

Gráfico tipo Barra (-): O gráfico tipo barra é o primeiro que aparece quando iniciamos o comando *Graph*. Ele é muito útil para descrever variáveis categóricas, como, por exemplo, a nossa já clássica variável *OBESIDADE* do arquivo *Apoe*. Primeiro deveremos abrir o arquivo utilizado no capítulo 4: *Apoe* e verificar se realmente temos a variável que desejamos. Não se esqueça de modificar o projeto se o arquivo não estiver ativo. Caso não exista essa variável, devemos abrir o arquivo *Apoe.mdb* que está no folder *Dados* do site [www.pucrs.br/igglepiinfo](http://www.pucrs.br/igglepiinfo). Após lermos o projeto e já com o comando *Graphs* aberto, selecionamos a variável *OBESIDADE* como *Main Variable (x)*. Note que o tipo de gráfico que aparece na janela *Graph type* é justamente o *Bar*. Automaticamente temos no eixo do Y a opção *Count* que nos dará o número total de indivíduos em cada grupo de obesidade. Sem selecionar outra opção, clicando em *OK* temos o gráfico, como na figura 7.2, praticamente pronto.

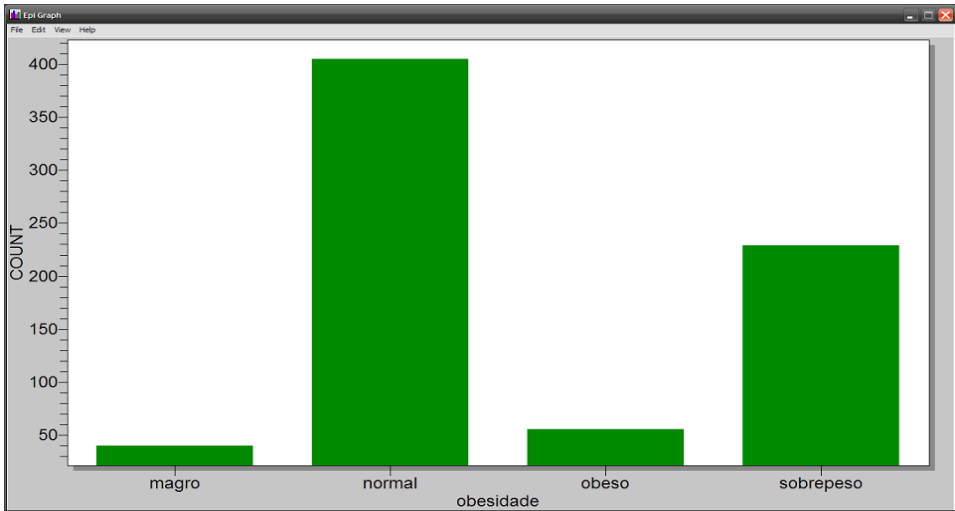


Figura 7.2. Janela de edição do comando *Graph* do *Analysis*.

Fonte: Epi Info®, arquivo *Apo.e.mdb*.

Inicialmente, o Epi Info® coloca o gráfico em uma janela que preenche toda a tela do computador (figura 7.2). Nessa tela, e somente nela, poderemos alterar o formato do gráfico, por isso é tão importante.

### Alterações comuns nos gráficos do Epi Info®

Ainda com o gráfico na janela de edição podemos clicar no botão à direita do *Mouse* e obteremos as seguintes opções (figura 7.3):

1. *Viewing Style*: alterna para monocromo e monocromo com símbolo (no caso do *Bar* não tem símbolo).
2. *Font Size*: três níveis para todo o gráfico: *Large*, *Medium* & *Small*.
3. *Numeric Precision*: define o número de casas decimais que vai aparecer nos eixos.

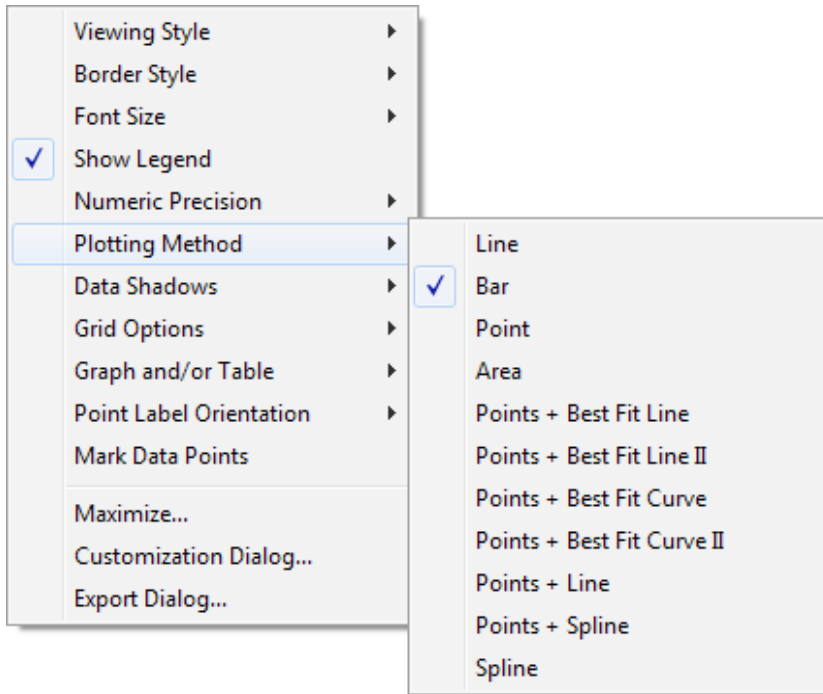


Figura 7.3. Janela das opções de edição do comando *Graph* do *Analysis* com as opções de plotagem abertas.  
Fonte: Epi Info®.

**4. *Plotting Method*:** podemos alterar o tipo de gráfico se nos arrependermos ou quisermos testar outras formas de visualização (figura 7.3.), os gráficos de linha (*Line*), ponto (*Point*), área (*Area*) e *Spline* podem ser criados diretamente selecionando a opção tipo de gráfico no comando *Graph*; veremos alguma descrição de cada um mais adiante.

**5. *Data Shadows*:** definimos se queremos colocar uma sombra ou dar um formato tridimensional às barras.

**6. *Grid Lines*:** dá a opção de colocarmos linhas horizontais, verticais ou ambas no gráfico.

**7. *Grid in Front*:** quando optamos por ter linhas horizontais ou verticais no gráfico; podemos mostrar as linhas na frente do gráfico ao selecionarmos essa opção.

**8.** *Graph and/or Table*: oferece a opção de incluirmos uma tabela junto com o gráfico ou só a tabela, no caso aparecerá o número de indivíduos em cada nível de obesidade com a precisão decimal definida pelo *Numeric Precision*.

**9.** *What to Table*: define se mostra todos os resultados ou somente os que aparecem no gráfico (posteriormente veremos que podemos selecionar quais grupos mostrar).

**10.** *Point Label Orientation*: opção para colocarmos os rótulos do eixo do X na posição horizontal, vertical ou inclinada (*Slanted*).

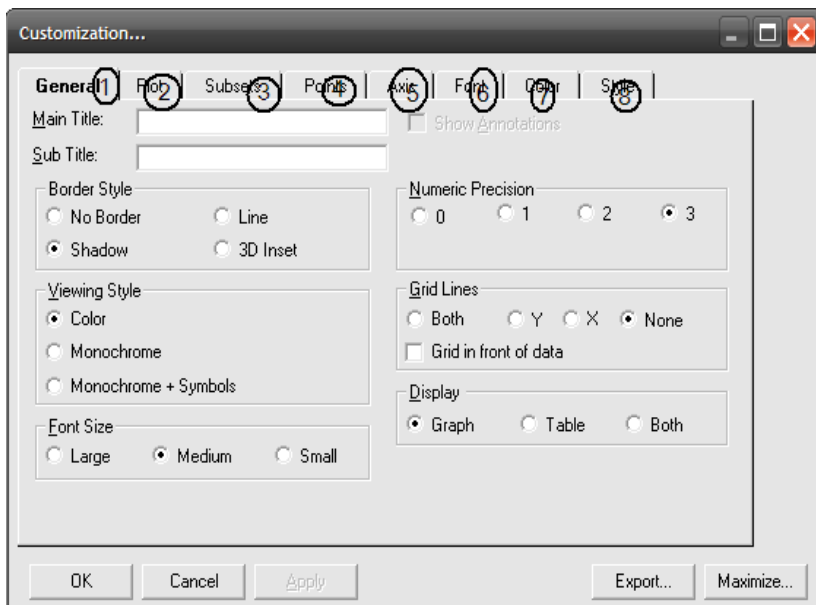


Figura 7.4. Janela do *Customization dialog* do *Graph*.

Fonte: Epi Info®.

**11.** *Customization dialog*: é a opção mais importante e pode ser aberta também se clicarmos com o botão da esquerda do *mouse* duas vezes sobre o gráfico. Na janela, visualizada na figura 7.4, temos quase todas as opções anteriores e outras novas. No item *General (1)*, podemos definir um título e subtítulo, além das opções anteriormente vistas (*Viewing Style*, *Font Size*, *Numeric Precision*, *Grid Lines*, *Display Graph and/or Table* e *Subsets to*

Table). O item *Plot* (2) é semelhante ao *Plotting Method*. *Subsets* (3) é utilizado para selecionar uma variável, quando fazemos um gráfico com mais do que uma variável no X. *Points* (4) é muito importante nesse tipo de gráfico, pois poderemos selecionar quais grupos mostrar ou não. *Axis* (5) é pouco útil no gráfico tipo barra. *Font* (6), nesse podemos escolher fontes separadas para título, subtítulo, rótulos e tabela.

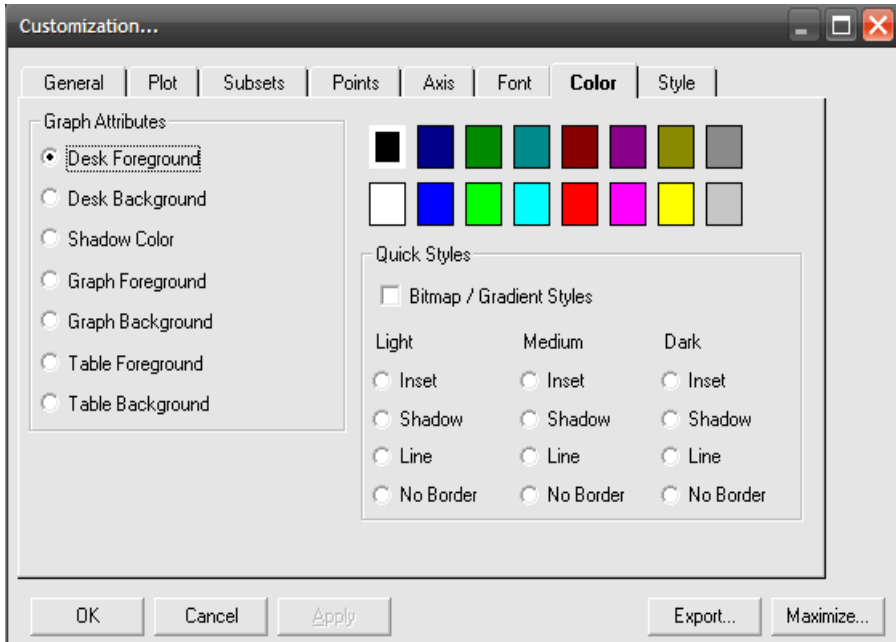


Figura 7.5. Janela da opção *Color* do *Customization dialog*.  
Fonte: Epi Info®.

*Color* (7) (figura 7.5) modifica somente as cores dos textos (*Desk Foreground*), fundo atrás do gráfico (*Desk Background*), cor da sombra do gráfico (*Shadow Color*, que não inclui a sombra das barras), cor das linhas dentro do gráfico (*Graph Foreground*), cor do fundo do gráfico (*Graph Background*) e as cores das tabelas (*Table Foreground* e *Table Background*) quando as mesmas existirem. Finalmente, no item *Style* (8), podemos modificar a cor das barras, quando no caso estamos trabalhando com o tipo *Bar*, e algumas opções que nos serão úteis para outros gráficos como tipo de pontos e tipo de linhas. Essa opção é particularmente útil quando temos duas variáveis categóricas num mesmo gráfico. Vamos ver como funciona



colocando *GENERO* junto com a variável *OBESIDADE* um pouco mais adiante, quando estivermos criando um outro gráfico.

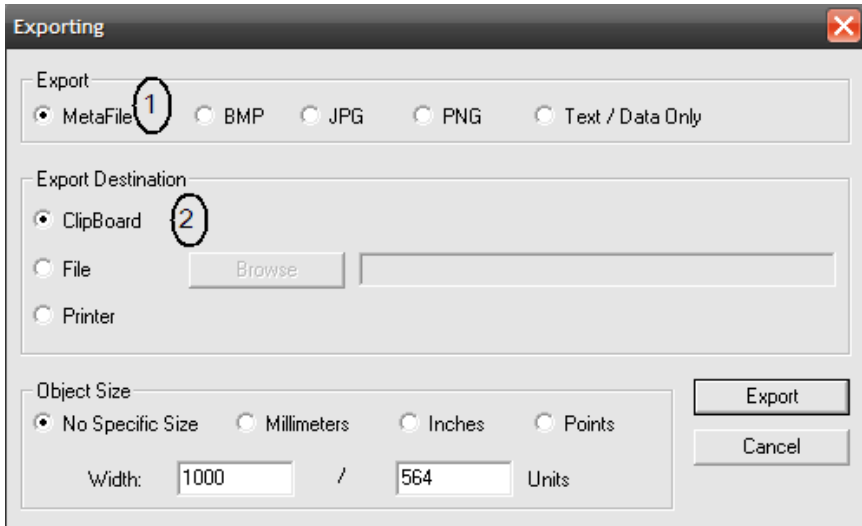


Figura 7.6. Janela do *Export Dialog* do *Graph* destacando o formato e local de exportação do gráfico em edição.

Fonte: Epi Info®.

**12.** *Export Dialog*: opção para exportar o gráfico em diversos formatos (1) e no nome que desejarmos (2). Primeiro devemos escolher a opção formato de exportação. No nosso caso, se exportarmos como *Metafile*, o arquivo ocupará somente 4Kb, se exportarmos como *BMP* ocupará 2.602 Kb, enquanto exportando como *JPG* ocupa 64 Kb. *Metafile* nem sempre é facilmente visualizável por programas. Por exemplo, o programa “Paint” não consegue abrir. Já o *JPG* é o que, depois do *BMP*, pode ser visualizado e editado por um número razoável de programas. Depois de escolhermos o formato, devemos decidir se queremos colocar o gráfico na memória imediata do computador (*Clipboard*), salvar em arquivo (*File*) ou ainda imprimir diretamente mandando para a impressora (*Printer*). Caso quisermos salvar em arquivo devemos, após escolher essa opção, clicar em *Browse* para definirmos o nome do arquivo e onde desejamos salvar. Finalmente clicamos em *Export*.

## Modificando os rótulos

Os rótulos ou textos que aparecem no gráfico podem ser modificados simplesmente clicando em cima deles. Isso fará aparecer uma caixa de edição. No caso podemos clicar no rótulo do eixo “Y” que se chama “*Count*” e modificá-lo para “Número de pessoas” (podemos usar acento espaço e qualquer outro símbolo especial).

Como exemplo, vamos editar o gráfico da *OBESIDADE* selecionando somente os normais, sobrepesos e obesos na opção *Plot*, colocando depois o eixo do “X” iniciando no Zero (escala), na opção *Axies*, mudamos a cor do fundo e da sombra para branco e finalmente a cor das barras para verde. Após modificarmos o rótulo do eixo “Y” para “Número de pessoas”, salvaremos o gráfico como tipo *Metafile* e nome *obesidade*.

Devemos fechar o gráfico para podermos criar outro, pois o programa não permite duas janelas de edição de gráfico abertas. Ao fecharmos o gráfico, observamos que o mesmo passa a aparecer na janela do *Output (Resultado)*. Notamos também que, mesmo clicando várias vezes em cima do gráfico, a janela de edição não aparecerá. É importante mais uma vez ressaltar que, depois de fechada a janela de edição não poderemos mais alterar o gráfico. Podemos, sim, clicando com o botão da direita do mouse, salvar o gráfico, mas com menos opções de formato.

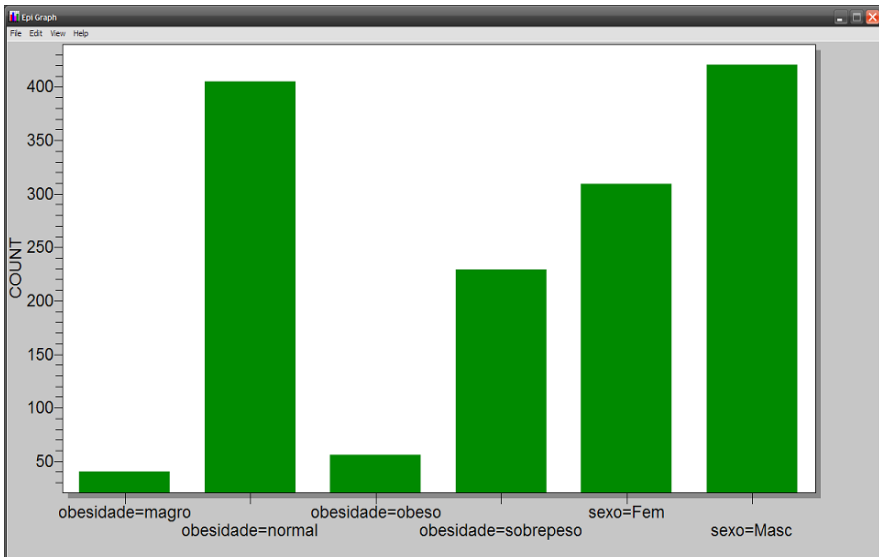


Figura 7.7. Exemplo de gráfico tipo barra com duas variáveis selecionadas.  
Fonte: Epi Info®, arquivo *Apo.e.mdb*.

1. Gráfico tipo *Bar* com duas variáveis categóricas (tipo texto): podemos colocar mais do que uma variável categórica, tipo texto, como *Main Variable* (eixo “X”), mas isso fará com que se crie uma barra para cada diferente valor das variáveis. Assim, se colocarmos as variáveis *OBESIDADE* e *GENERO* juntas, aparecerá um gráfico mais ou menos como na figura 7.7. A linha de comando é: `GRAPH OBESIDADE GENERO GRAPHTYPE = “Bar”` Observamos que todas as colunas têm a mesma cor, tendo tanto a variável *GENERO* quanto a variável *OBESIDADE* a mesma cor. Se quisermos, podemos modificar a cor das barras, mas todas se manterão iguais. Gráficos separados de cada variável talvez ficaria melhor.

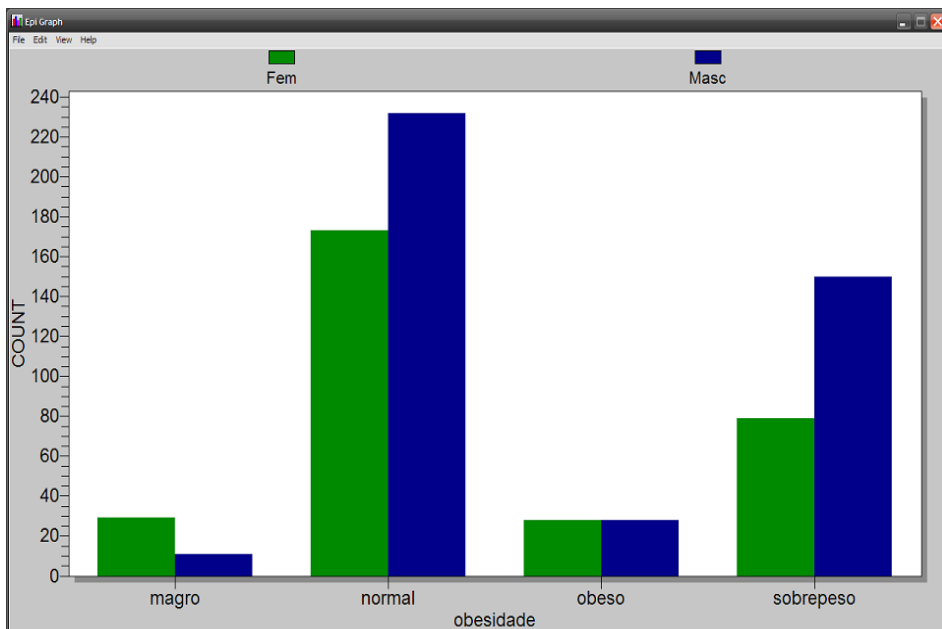


Figura 7.8. Exemplo de gráfico tipo barra utilizando a opção *Bar for each value of*.

Fonte: Epi Info®, arquivo *Apo.e.mdb*.

Outra maneira, talvez mais interessante, de colocarmos duas variáveis categóricas num mesmo gráfico é utilizando a opção *Bar for each value of* que é a opção número (5) da janela de criação de gráficos. A linha de comando é muito semelhante à anterior, mas aparece um asterisco entre *OBESIDADE* e *GENERO*: `GRAPH OBESIDADE * GENERO GRAPHTYPE = “Bar”`. O gráfico aparece com quatro sequências de duas barras com cores diferentes, uma para cada valor

de gênero, conforme o rótulo observado na parte de cima do gráfico na figura 7.8. Esse gráfico nos proporciona observar a distribuição do gênero em cada nível de obesidade. Podemos observar que, no grupo estudado, existem mais mulheres magras e mais homens normais e com sobrepeso e que o número de obesos é igual em ambos os gêneros no nosso banco de dados. A interpretação, no entanto, é um pouco complicada, pois temos mais homens na amostra que mulheres. Melhor seria se tivéssemos escolhido a opção *Show value of (4)* na janela principal do *Graph*, *Count %*, mas fica como lembrança aqui.

**2.** Gráfico tipo *Bar* para descrever a média de variáveis numéricas entre grupos: podemos utilizar o gráfico tipo barra para observar o comportamento de uma variável numérica em cada nível de uma variável categórica. No caso, podemos observar o comportamento da idade para cada nível de obesidade. Para tanto, colocamos *OBESIDADE* como *Main variable (1)*, na janela *Weight Variable (2)* colocamos *IDADE* e finalmente devemos escolher que valor de *IDADE* queremos colocar no gráfico, na opção *Show value of (3)*. Podemos escolher entre colocar a soma de todas as idades (*Sum*), a média (*Average*), o valor máximo (*Maximum*) ou o valor mínimo (*Minimum*) das idades, para cada um dos valores de *OBESIDADE*. No caso, faz mais sentido colocarmos a média. O recorte da nossa janela ficaria com o formato exemplificado na figura 7.9.

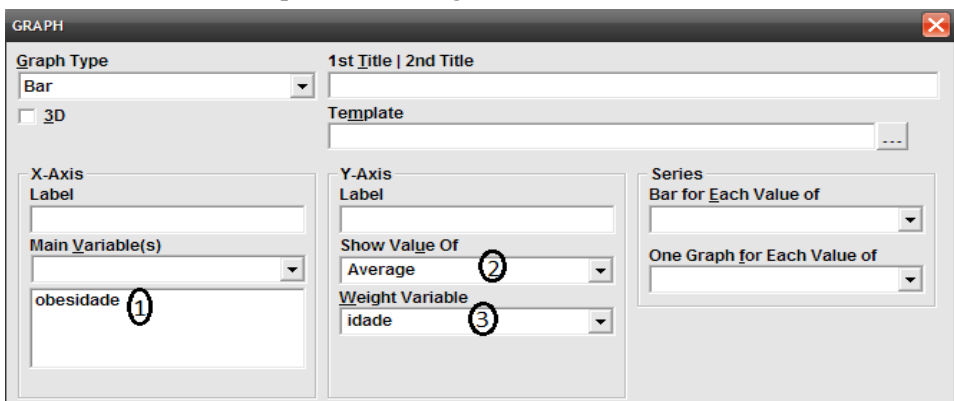


Figura 7.9. Detalhe da janela do *Graph* com comandos para a criação de barras com média de *IDADE* para cada valor de *OBESIDADE*.

Fonte: Epi Info®.

O gráfico teria quatro barras, uma para cada valor de *OBESIDADE*, mas com tamanhos (eixo “Y”) relacionados à média das idades em cada um dos grupos de obesidade. Mais uma vez o Epi Info® inicia a escala do valor da variável “Y” próximo ao melhor valor da amostra. Isto faz com que observemos uma diferença bastante grande nos valores de *IDADE*, estando, no caso, os magros com uma barra muito menor que os outros níveis de *OBESIDADE*. Se prestarmos atenção, o valor médio de idade dos magros está em torno de 47 anos enquanto o valor dos com sobrepeso não chega a 54 anos. Assim, a diferença entre o grupo mais velho e o mais jovem é menor que sete anos. Ao corrigirmos a escala iniciando no “zero”, observamos, conforme a figura 7.10, que as barras não são tão diferentes assim.



Figura 7.10. Gráfico das médias de *IDADE* para cada valor de *OBESIDADE* após a correção da escala do eixo vertical.

Fonte: Epi Info®, arquivo *Apo.e.mdb*.

**3.** Gráfico *Rotate Bar*: é o mesmo gráfico anterior só que com barras horizontais.

**4.** Gráfico tipo Histograma (*Histogram*): é um gráfico de barra usado para as variáveis numéricas como *IDADE*, *IMC* etc. Cada barra é construída com intervalos fixos, que devemos determinar usando a opção *Interval* a partir de um determinado valor (*Ist Value*). Se não colocarmos nenhum valor na opção de

intervalo, o programa criará um gráfico usando cada valor unitário da variável, ou seja, um histograma com inúmeras barras. Como exercício em classe usaremos a variável *idade*. Podemos usar, na realidade, qualquer valor como intervalo, mas faz mais sentido colocarmos intervalos de dez, para termos uma barra para cada década de idade. É também recomendável estabelecermos qual será o primeiro valor do histograma (*1st Value*), do contrário o gráfico iniciará a partir do primeiro valor que pode ser não inteiro. No caso, o primeiro valor da variável *idade* é 17,2 anos. Se não colocarmos nenhum valor inicial isto afetará todos os intervalos, criando, assim, os intervalos 17,2, 27,2, 37,2 e assim por diante. Para evitar isto, colocaremos um valor redondo: 10. O recorte da janela de comando ficará, então, como na figura 7.11.

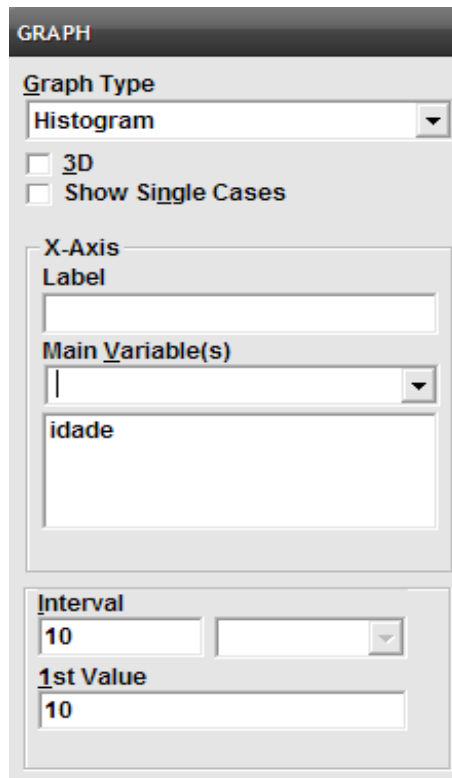


Figura 7.11. Detalhe da janela do *Graph* mostrando as opções do tipo *Histogram*.

Fonte: Epi Info®.

O histograma aparece como um gráfico de barras onde temos no eixo “X” os valores dos intervalos e no eixo “Y” o número de pessoas correspondentes a cada intervalo. Quando temos uma variável numérica, espera-se que tenhamos um formato semelhante ao de um sino, com um número reduzido de pessoas nos intervalos extremos (mínimos e máximos) e um número mais concentrado de pessoas ao redor da média. O gráfico do histograma da idade desta nossa amostra apresenta mais ou menos essa característica (figura 7.12).

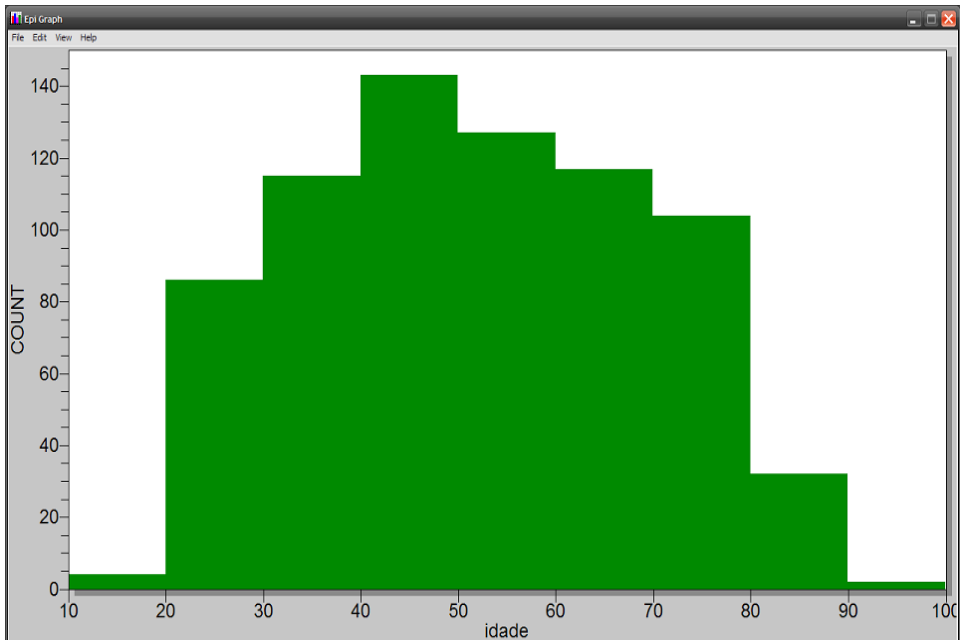


Figura 7.12. Exemplo de gráfico tipo histograma.

Fonte: Epi Info®, arquivo *Apo.e.mdb*.

**5.** Gráficos tipo *Spline* e *Points*: são gráficos semelhantes ao gráfico de barra que produzem respectivamente gráfico de linhas arredondadas conectando os valores ou somente pontos. Ao contrário do tipo “barra”, o comando só aceita variáveis numéricas. No caso só funcionariam se tivéssemos valores numéricos para as variáveis categóricas. Como os gráficos podem ser criados a partir do tipo “barra”, recomenda-se criar primeiro esse tipo de gráfico e editá-lo para *Spline* ou *points*, conforme o desejado.

**6.** Gráfico tipo *Pizza* ou *Torta (Pie)*: propicia uma visão da distribuição proporcional dos valores de uma variável categórica ou numérica não contínua. Como exercício, utilizaremos agora a variável *CIGARRO*, simplesmente selecionando o tipo de gráfico *Pie* e colocando a variável *CIGARRO* como *Main Variable(s)*. Ao clicarmos *OK* vemos o gráfico. A janela de configurações deste tipo de gráfico apresenta algumas particularidades observadas nas figuras 7.13 e 7.14.

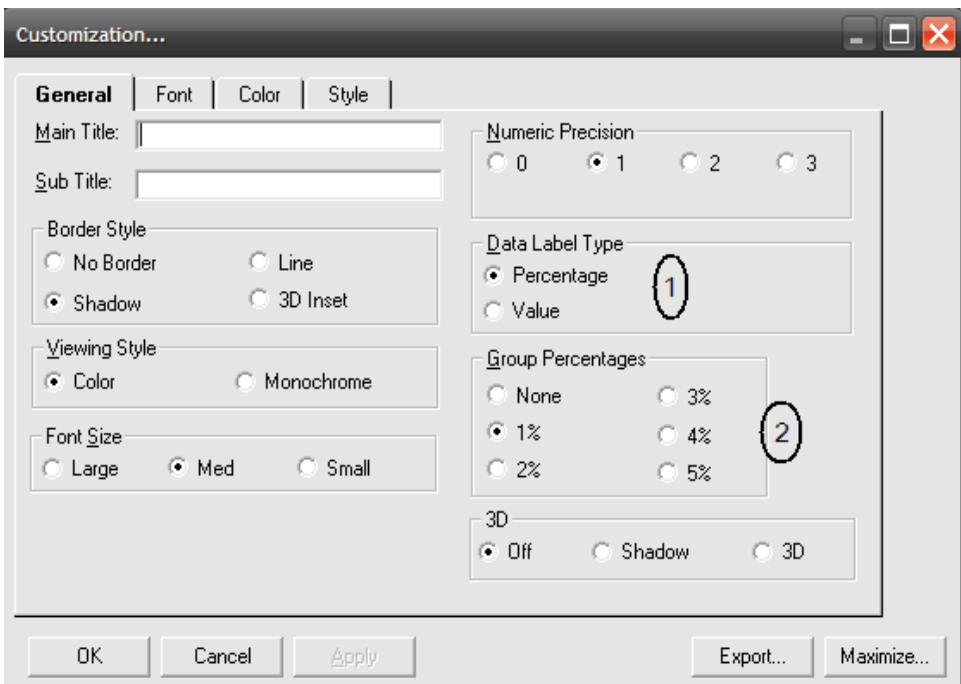


Figura 7.13. Janela de configurações do gráfico tipo *Pizza – Pie*.  
Fonte: Epi Info®.

Na opção *Data Label Type* (1), destacada na figura 7.13, podemos mostrar valores percentuais ou os valores absolutos. No caso dos valores percentuais podemos modificar o número de casas decimais através do comando *Numeric Precision*. *Group Percentages* (2). Podemos agrupar valores pequenos ( $\leq 5\%$ ). Podemos destacar uma fatia. Para tanto devemos clicar duas vezes no rótulo próximo da mesma. Por exemplo, para destacar a fatia atual, clicamos duas vezes na palavra atual ou no número 17%.



Na opção *Style* podemos modificar a cor de cada uma das fatias. Nosso gráfico poderia ficar como na figura 7.14.

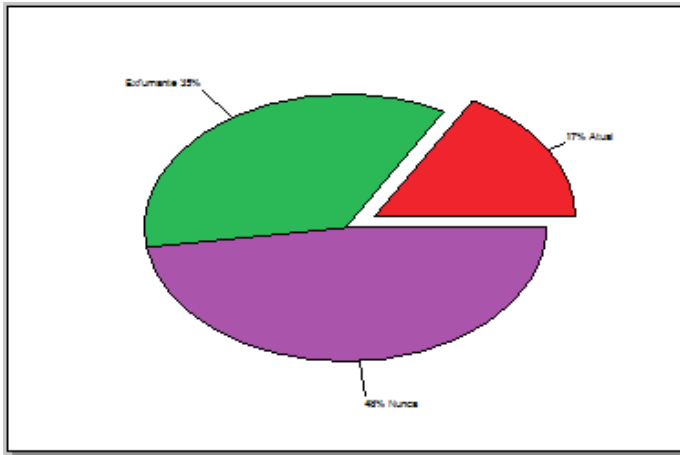


Figura 7.14. Exemplo de gráfico tipo pizza, destacando a fatia atual.  
 Fonte: Epi Info®, arquivo *Apo.e.mdb*.

**7.** Gráfico tipo *área* (*Area*): é semelhante ao gráfico de barras produzindo uma área abaixo de uma linha que une os pontos. Este gráfico pode ser convertido em gráfico de barra, linha ou ponto.

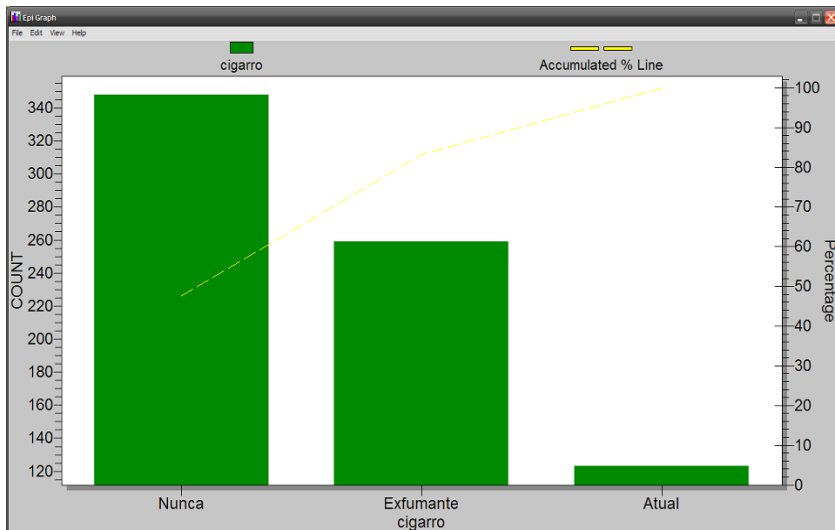


Figura 7.15. Exemplo de gráfico tipo *Pareto*.  
 Fonte: Epi Info®, arquivo *Apo.e.mdb*.

**8.** Gráfico tipo *Pareto*: é um gráfico de barras com a percentagem cumulativa dos valores em uma linha. As barras são dispostas em ordem decrescente de valor. Por exemplo, o gráfico Pareto da variável *CIGARRO* ficaria como na figura 7.16. Observe que esse gráfico tem, na margem vertical direita, a escala percentual da linha.

**9.** Gráfico XY (*Scatter XY*): constrói gráficos de dispersão XY e desenha uma linha de regressão. Necessitamos, assim, de duas variáveis numéricas, por exemplo, *COLESTEROL* e *IDADE*. A primeira variável a ser colocada é a que vai aparecer no eixo horizontal (X), a seguinte aparecerá no eixo vertical (Y). Pode-se colocar mais uma variável no eixo Z que seria um terceiro eixo entre os outros. Abordaremos algumas características da interpretação deste tipo de gráfico na sessão de Análise de Dados, mais adiante neste livro. O gráfico da dispersão do *COLESTEROL* e da *IDADE* ficaria como na figura 7.16.

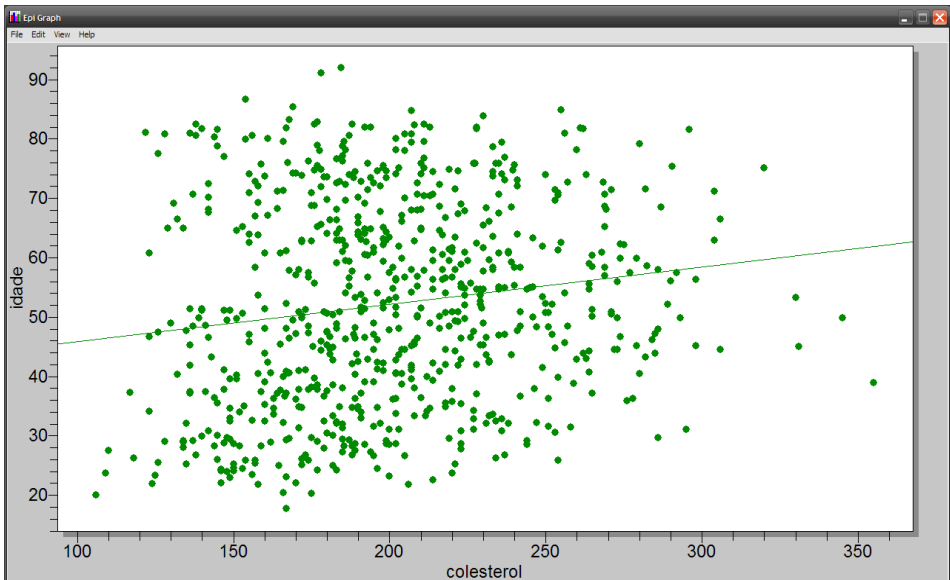


Figura 7.16. Exemplo de gráfico tipo *Scatter XY*.  
Fonte: Epi Info®, arquivo *Apo.e.mdb*.

**10.** Gráfico de barra com mais de um nível (*Stacked bar*):  
Vimos que podemos criar gráficos de barra com mais de uma

variável. Usando a opção *Bar for each value of*, teremos mais do que uma barra para cada valor variável principal. O tamanho de cada barra corresponde ao número de indivíduos em cada nível das duas variáveis. Podemos, entretanto, dividir cada barra da variável principal conforme o número de pessoas pertencentes ao nível da segunda variável. Assim teremos barras com mais de uma cor, cada cor estaria relacionada com o nível da segunda variável. Seria como tivéssemos barras, uma em cima da outra, conforme o número de níveis que a segunda variável possuía. O comando do gráfico tipo *Stacked bar* é semelhante ao comando barra. Uma principal vantagem deste tipo de barra é que, na opção *Plot* do *Customization dialog* podemos selecionar barras percentuais tornando todas as barras iguais e podendo visualizar as diferenças de distribuição percentual com mais facilidade do que com os valores absolutos. As duas versões, com valores absolutos e percentuais das barras de obesidade para valores de gênero ficariam como nas figuras 7.17 e 7.18 respectivamente.

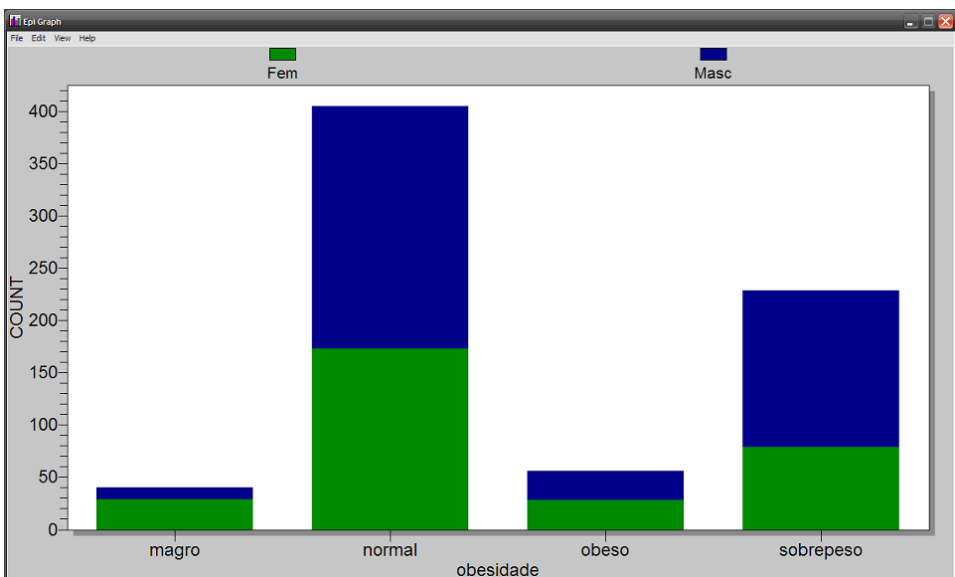


Figura 7.17. Exemplo de gráfico tipo *Stacked bar* para valores absolutos.

Fonte: Epi Info®, arquivo *Apo.e.mdb*.

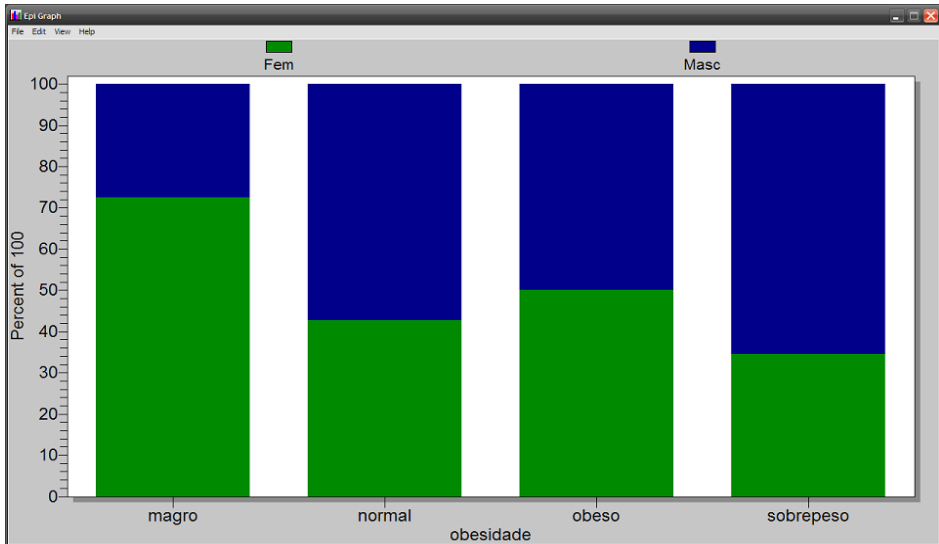


Figura 7.18. Exemplo de gráfico tipo *Stacked bar* para valores percentuais.

Fonte: Epi Info®, arquivo *Apo.e.mdb*.

**11.** Gráfico em formato de vela (*Box-Whisker*): quando queremos mostrar o comportamento de uma variável numérica em diferentes grupos, nem sempre só a média é o suficiente. Alguma informação sobre o grau de dispersão da amostra ao redor dessa média é importante. Para isto existe o gráfico tipo vela (*Box-Whisker*) onde podemos ilustrar esse grau de dispersão. Quando escolhermos esse tipo de gráfico no comando *Graph* do Epi Info®, a opção *Main variable(s)* se torna *Analyze Variables*. Ao contrário do gráfico tipo barra, devemos colocar nossa variável numérica na primeira opção (1), neste caso escolheremos a variável *COLESTEROL* como variável a ser analisada e a variável *GENERO* para grupo (2). Uma terceira opção é necessária, pois devemos informar ao comando que tipo de gráfico(3) queremos: média com 1 e 2 desvios padrão (*Mean-1SD-2SD*), mediana com 25 e 10% (*Median-25%-10%*) ou mediana com 25 e 5% (*Median-25%-5%*). Como queremos a média e desvios padrão, escolheremos a primeira opção. Nosso comando ficaria como na figura 7.19.

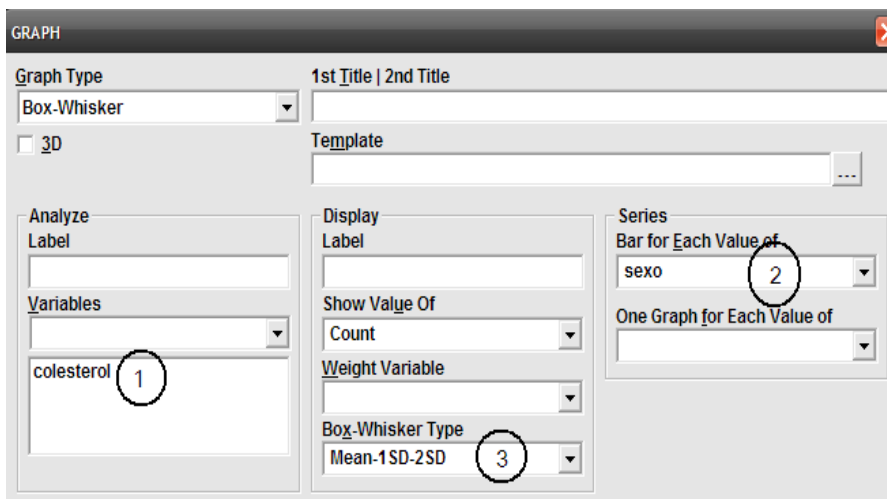


Figura 7.19. Detalhe da janela do *Graph* com os comandos para a criação de gráfico tipo *Box-Whisker*.  
Fonte: Epi Info®.

O gráfico, por sua vez, aparece como uma barra flutuante com dois traços verticais e um ponto central. O ponto central representa justamente a média, a barra flutuante, representa um desvio padrão superior e inferior e os traços dois desvios. Observamos, no gráfico da figura 7.20, que os homens apresentam média superior, mas um desvio padrão maior que as mulheres.

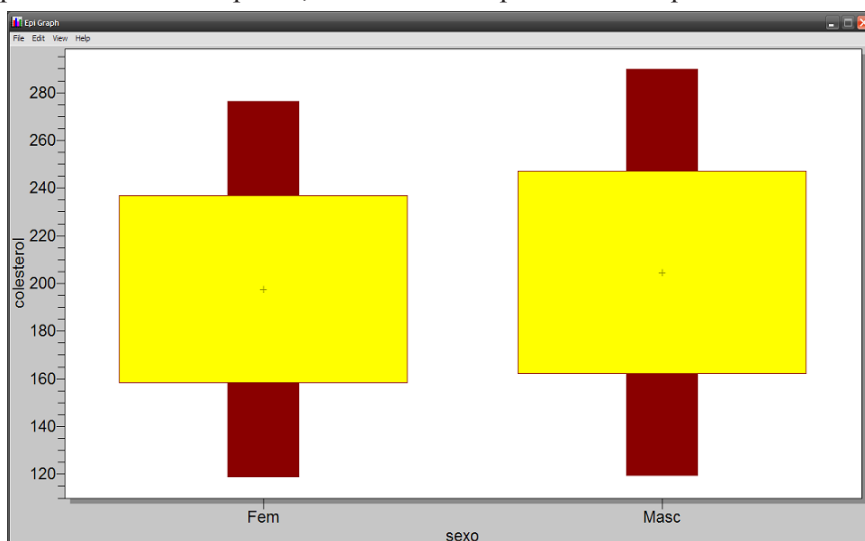


Figura 7.20. Exemplo de gráfico tipo *Box-Whisker*.  
Fonte: Epi Info®, arquivo *Apo.e.mdb*.

**12.** Gráfico de máximos e mínimos (*Hi-Low Graph*): serve para ilustrar o comportamento de duas variáveis, contendo uma os valores máximos e a outra os mínimos da variável grupo. Poderia ser usado, por exemplo, para ilustrar o comportamento da pressão sistólica e diastólica nos dois gêneros, conforme a figura 7.21.

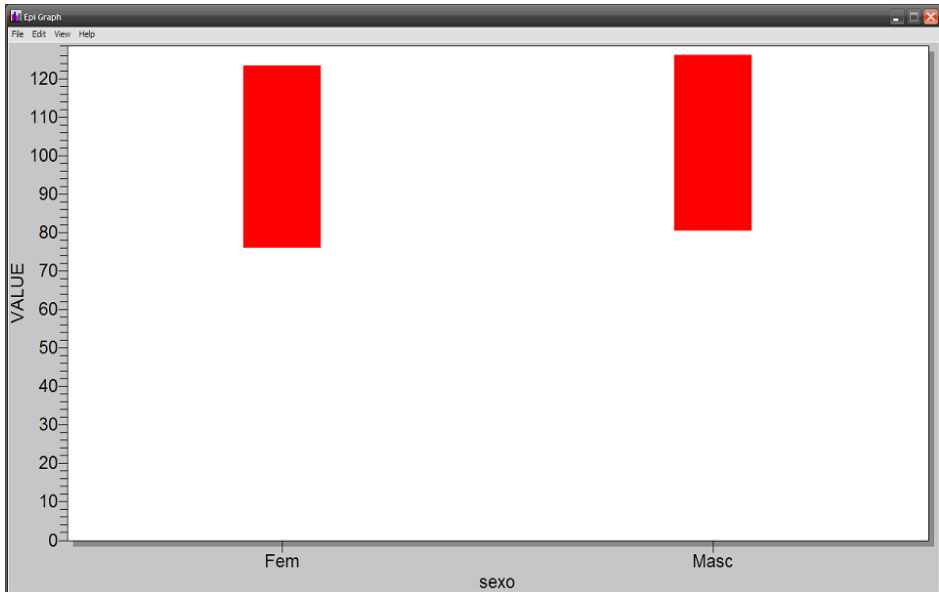


Figura 7.21. Exemplo de gráfico tipo *Hi-Low*.  
Fonte: Epi Info®, arquivo *Apo.e.mdb*.

**13.** Histograma com mais de um nível (*Stacked Histogram*): assim como temos um gráfico de barras com mais de um nível e o histograma que cria barras para variáveis numéricas, podemos também criar histogramas com mais de um nível utilizando o gráfico tipo *Stacked Histogram*. O comando é similar aos dois. Depois de selecionar a variável numérica, o intervalo e o valor inicial, devemos selecionar a variável grupo. Por exemplo, se colocarmos a idade, para cada 20 anos, tendo o valor inicial de 20, agrupando por gênero, teremos o gráfico da figura 7.22.

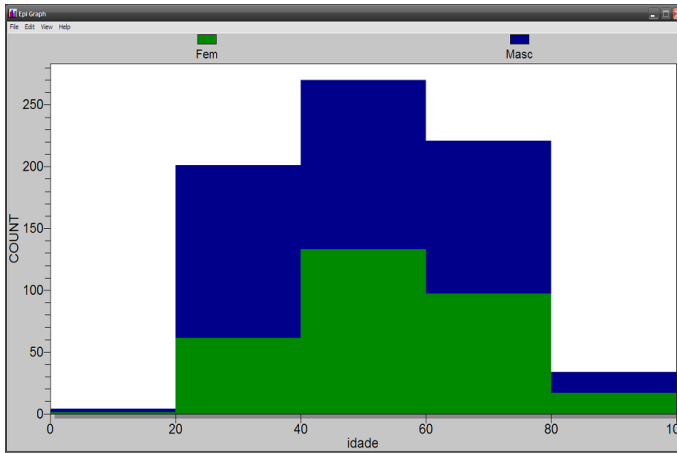


Figura 7.22. Exemplo de gráfico tipo *Stacked Histogram*.  
 Fonte: Epi Info®, arquivo *Apo.e.mdb*.

**14.** Criando pirâmides (*Pyramid Graph*): o gráfico tipo pirâmide é na realidade um gráfico tipo barra horizontal que tem duas séries, uma a direita outra a esquerda. Este gráfico é utilizado para criar as famosas pirâmides populacionais. O comando exige que a variável principal (*Main variable*) seja numérica, mas não dá opções de se criar intervalos. Assim, se criarmos uma variável chamada *DÉCADA* onde teremos as idades divididas em intervalos de 10, poderemos criar um gráfico tipo pirâmide para cada valor de *GENERO* como a figura 7.23.



Figura 7.23. Exemplo de gráfico tipo *Pyramid*.  
 Fonte: Epi Info®, arquivo *Apo.e.mdb*.

## Exercício:

1. Abra (ler) o arquivo *Apoe\**, tabela Grupo1, na Análise (não esqueça de mudar o projeto) e faça os seguintes gráficos:
  - a. Barra para cada *GENERO* para médias de *IDADE*.
  - b. Pizza (torta) da variável *OBESIDADE*, destacando o grupo obeso.
  - c. Gráfico Barra descrevendo a distribuição percentual do nível de *OBESIDADE* para cada valor de *GENERO*.

\* Poderá usar o banco de dados homônimo que se encontra no folder *Dados* do site [www.pucrs.br/igg/epiinfo](http://www.pucrs.br/igg/epiinfo), caso não tenha salvo o projeto em capítulo anterior.



## Capítulo 8

# ANALISANDO DADOS, UMA INTRODUÇÃO

Valemo-nos dos cálculos e testes estatísticos para descrever e interpretar os dados provenientes de um levantamento ou pesquisa. Média, desvio padrão, variância, mediana, valores máximos e mínimos, entre outros, são cálculos e valores muito utilizados na descrição de dados (estatística descritiva). Os testes estatísticos, como a regressão linear, análise de variância, são instrumentos utilizados para testar ou comprovar hipóteses fundamentadas no objetivo da pesquisa e observadas durante a análise descritiva dos dados. Tanto a estatística descritiva quanto os testes estatísticos têm, como fim, a interpretação dos dados colhidos durante a pesquisa, que objetiva chegarmos às conclusões inicialmente pretendidas.

### Importância da análise descritiva

Toda a análise de dados deve começar pela análise descritiva de todas as variáveis que pretendemos usar. Observando o resultado dessas análises podemos identificar possíveis problemas com os dados como números que o autor chama de “incompatíveis com a vida” ou categorias não definidas. Por mais que se tenha cuidado na introdução dos dados sempre devemos esperar que alguma inconsistência possa ter passado despercebida. A análise descritiva também pode nos orientar na análise dos dados apontando diferenças entre grupos em variáveis que possam comprometer nossos resultados. No exemplo que descreveremos a seguir teremos dois grupos de tratamento que devem ser uniformes em vários aspectos, por exemplo, gênero e idade.

A análise descritiva geralmente faz parte do que chamamos de “tabela 1”, onde descrevemos todas as variáveis que iremos usar na nossa análise. Frequência das variáveis categóricas e médias das variáveis numéricas entre diferentes grupos como, por exemplo, gênero ou raça nos ajuda a antecipar os resultados ou a necessidade de uma análise específica como veremos no decorrer do livro.

### O que é hipótese nula

A hipótese que testamos estatisticamente é chamada de hipótese nula. Exemplificando, vamos testar a eficiência de uma medicação. Dividiremos os pacientes em dois grupos, um usando medicação, grupo Tratamento, outro

grupo placebo (grupo Controle). Nossa variável resposta (ou dependente) será mortalidade. Esperamos que exista algum efeito benéfico da medicação, ou seja, que o grupo Tratamento tenha uma mortalidade menor que o grupo Controle, ou seja, mortalidade do grupo Controle ( $m_{controle}$ ) maior que a mortalidade do grupo Tratamento ( $m_{tratamento}$ ). Em outras palavras,  $m_{controle} < m_{tratamento}$  seja maior que zero ( $m_{controle} - m_{tratamento} > 0$ ). (Revise com calma!)

Se a medicação não tiver nenhum efeito, a mortalidade do grupo controle será igual à mortalidade do grupo tratamento, ou seja,  $m_{controle} = m_{tratamento}$  será igual a zero. Chamamos essa hipótese de “hipótese nula”, porque tentaremos provar que ela é falsa, no jargão estatístico, tentaremos rejeitar a hipótese nula. Se conseguirmos rejeitar a hipótese nula, conseguiremos provar que a medicação tem efeito sobre a mortalidade, pois  $m_{tratamento}$  é diferente da  $m_{controle}$ , chamamos esta segunda hipótese de “hipótese alternativa”.

### Por que testamos a hipótese nula?

Se nós acreditarmos que a medicação realmente tem efeito sobre a mortalidade, por que não testar que a diferença é maior que zero? A justificativa para isso é que existem infinitos valores maiores que zero. Se testássemos que a diferença da mortalidade fosse 4 e não conseguíssemos provar isso, teríamos que testar vários outros valores. Testando a hipótese nula, conseguimos concluir se há ou não diferença e, se usarmos o teste adequado, conseguimos provar que a mortalidade do grupo sem medicação é maior que a do grupo com medicação. Por outro lado, é impossível provar uma hipótese, conseguimos rejeitar uma hipótese. Comparamos grosseiramente com o teste de paternidade, antes da era do teste do DNA. Conseguimos comprovar quem não é o pai (rejeitar a paternidade). Quem não conseguir provar que não é o pai assume a paternidade (não conseguiu rejeitar a paternidade). Este exemplo só é válido se não utilizarmos os testes de DNA, pois o pai terá metade dos genes do filho, sempre.

A maioria dos testes estatísticos tem o objetivo de rejeitar a hipótese nula. Entretanto, se não conseguirmos rejeitar a hipótese nula não significa que realmente não existe diferença entre os dois grupos estudados. O não rejeitar a hipótese nula significa tão somente que os resultados são consistentes com a hipótese nula. Podemos comparar novamente com o caso da paternidade, quem não conseguiu provar que não é pai pode, na realidade, não ser o pai da criança.

## Tipos de erros relacionados com as hipóteses

É importante lembrar que sempre trabalharemos com uma amostragem de uma população, tentando, ao observar essa pequena amostragem, concluir dados sobre toda uma população. Por isto, nunca poderemos estar totalmente certos da verdade, pois assumimos sempre algumas incertezas sobre a nossa amostragem.

Infelizmente, a verdade é uma incógnita que tentamos desvendar. Ao testarmos uma hipótese, poderemos assumir dois tipos de erros: rejeitar a hipótese nula quando na realidade ela é verdadeira e concluir que a medicação é eficaz quando na realidade não tem nenhum efeito. Esse erro é chamado de Erro Tipo I (ou alfa). Podemos, inversamente, não rejeitar a hipótese e concluir que a medicação não tem efeito quando na realidade ela é eficaz. (É importante lembrar que a realidade é nossa principal incógnita!) Esse erro é chamado de Erro Tipo II (ou beta).

Não podemos eliminar o risco de cometermos algum desses erros, mas, sim, diminuir a probabilidade de que eles ocorram. A probabilidade de cometermos o Erro Tipo alfa, ou seja, rejeitarmos a hipótese nula quando na realidade ela é falsa, é conhecido como “o nível de ‘significância’ de um teste estatístico”. Quando lermos que um resultado de um teste foi *significativo a um nível de 0,05* significa que a probabilidade de estarmos errados ao rejeitarmos a hipótese nula (e concluir que a hipótese alternativa é erroneamente verdadeira) é de 0,05, ou seja 5%. Podemos também dizer que alfa é a possibilidade de determinado resultado ter sido somente determinado pelo acaso. Alguns autores consideram um alfa entre 0,1 e 0,05 como indicativo de uma diferença estatística. Um alfa menor que 0,01 é normalmente considerado e descrito em artigos como uma diferença muito significativa. Existe uma tendência contemporânea de considerar o erro alfa de 0,05 de significância baixa, mas isso não é uma unanimidade.

A maioria dos artigos dá muita importância ao erro tipo I e muitas vezes se esquece de abordar o erro tipo II, que é igualmente importante. Como vimos, o nível de “significância” usualmente aceito para um teste estatístico é de 0,05. O erro tipo II passa a ser um problema quando, evidentemente, não conseguimos rejeitar a hipótese nula dentro de um nível aceitável. Quanto maior for a diferença entre a mortalidade dos dois grupos, menor será a probabilidade de cometermos um erro tipo II (também chamado de erro beta), o nível de poder normalmente aceito para um teste estatístico é de 0,20. O erro tipo beta está diretamente relacionado com a variância da nossa amostragem (variância = (desvio padrão)<sup>2</sup>). Quanto maior a variabilidade dos dois grupos,

menor será a probabilidade de comprovarmos uma diferença entre dois grupos (a diferença entre os grupos deverá ser maior para podermos rejeitar a hipótese nula). Para calcular a variância utilizamos a seguinte fórmula:

$$S(x) = \frac{\sum (x - \bar{x})^2}{n - 1}$$

Em linguagem humana: variância é igual ao somatório do quadrado das diferenças entre cada valor de  $x$  e a média (quanto mais próximos da média os valores de cada pessoa, menor será essa diferença, conseqüentemente menor será a variância), dividido pelo número da amostragem, menos um. Quanto maior o número de pessoas na nossa amostragem, menor será nossa variância. Assim podemos diminuir a probabilidade de termos um erro tipo beta aumentando o número de participantes de nossa pesquisa. A capacidade de uma amostragem rejeitar a hipótese nula é chamada de “Poder Estatístico”. Assim o erro tipo beta é muito utilizado para calcular o número (N) amostral, ou seja, o número antecipado de pessoas ou participantes em uma pesquisa necessário para se observar significância estatística em uma diferença entre grupos.

## Usando o Programa *StatCalc* para calcular o N Amostral

O programa *StatCalc*, criado em 1993, é usado para cálculos estatísticos rápidos. Basicamente possui três componentes: Tabelas 2xN, cálculo no N amostral e o Qui-quadrado para tendência (figura 8.1).

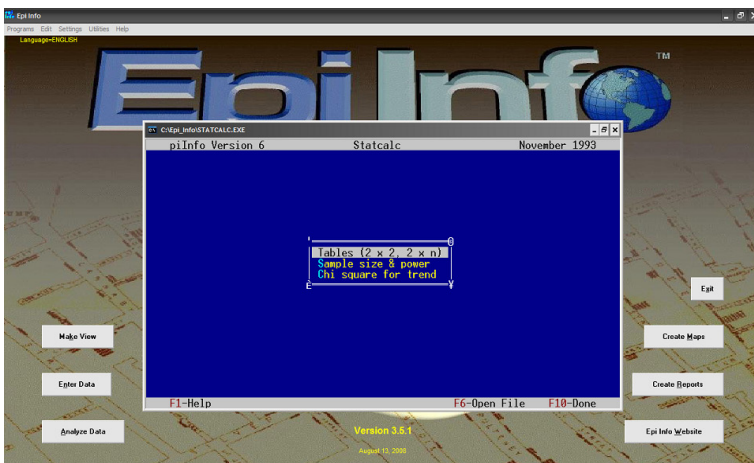


Figura 8.1. Janela inicial do *StatCalc*.  
Fonte: Epi Info®.

*Sample Size & Power*: calcula o número amostral necessário para se conduzir uma pesquisa, baseado em proporções.

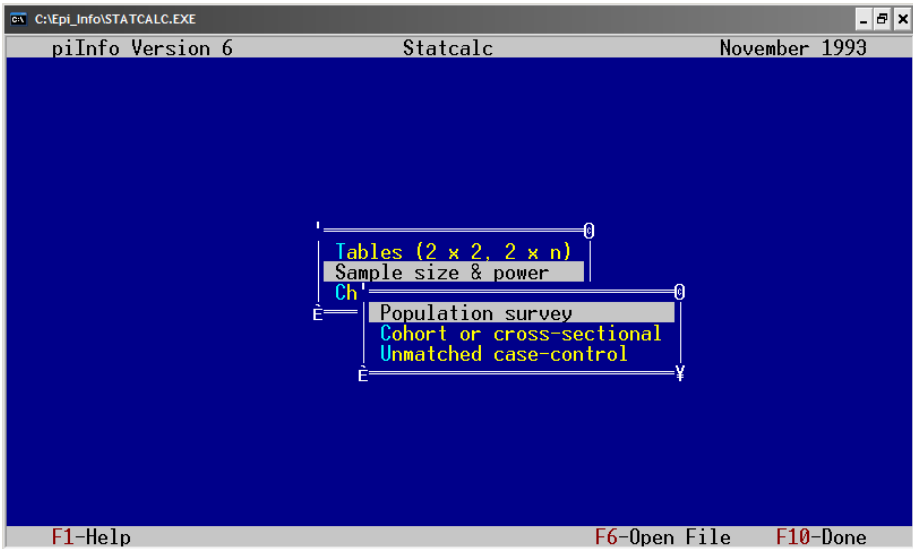


Figura 8.2. Janela do *StatCalc* mostrando as opções do *Sample size & power*.

Fonte: Epi Info®.

O delineamento das pesquisas determinará o tipo de cálculo a ser realizado. Quando desejamos obter uma amostra representativa de uma população para saber a prevalência de determinada característica usamos o *Population survey*.

*Exemplo*: vamos supor que queremos estudar o número de fumantes entre estudantes de uma universidade. Sabendo que o número de fumantes em uma população é normalmente de 30%, e encontramos em um estudo piloto que 45% dos estudantes fumam, queremos saber quantas pessoas devemos estudar para confirmarmos essa hipótese. O número total de estudantes é 10 mil. Assim colocaremos os dados como mostra a figura 8.3.

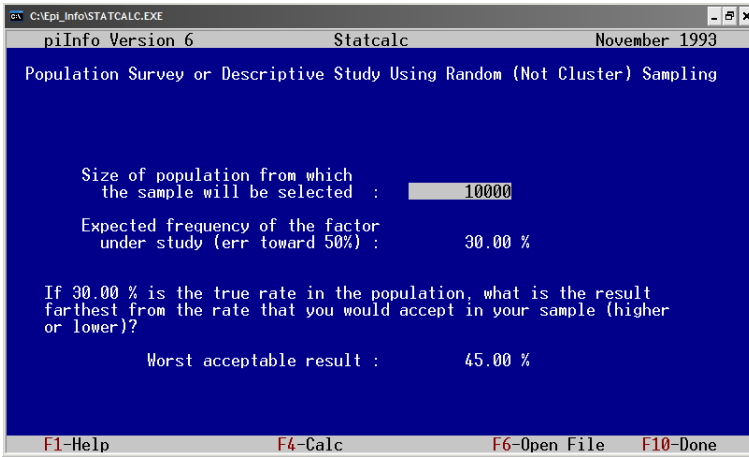


Figura 8.3. Janela de definição dos parâmetros para o cálculo amostra de uma pesquisa populacional pelo *StatCalc*.

Fonte: Epi Info®, dados fictícios.

Pressionando o F4 obteremos o resultado, ilustrado na figura 8.4.

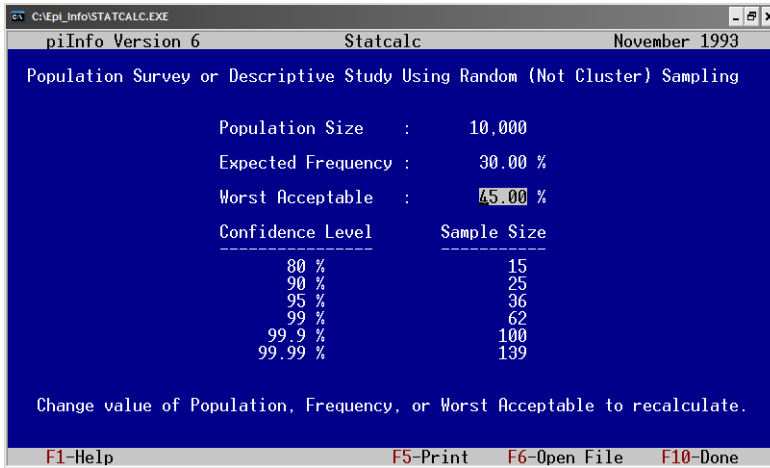


Figura 8.4. Janela com os resultados do cálculo amostral de uma pesquisa populacional pelo *StatCalc*.

Fonte: Epi Info®, dados fictícios.

Neste caso, o N amostral é calculado sobre o intervalo de confiança (*Confidence level*). O intervalo de confiança é calculado como  $(1 - \text{erro } \alpha)$  e, assim, devemos usar o intervalo de 95%. O programa oferece, assim, vários níveis de Intervalo de confiança, para o nível de 95% necessitamos entrevistar 36 pessoas aleatoriamente.

## Exercício :

Calcule o tamanho de amostra para uma pesquisa de população de 700,000 pessoas para uma doença em que é esperada uma prevalência que seja 10%, mas o pior resultado que se espera poderia ser tanto quanto 20%. Dê os valores por um nível de confiança de 95%.

### *N amostral para Estudo de Coorte ou Estudo Transversal*

Estudo de Coorte é um estudo onde acompanhamos um grupo de pessoas com características especiais. Estudo Transversal é um estudo onde temos dados coletados somente uma vez e comparamos a prevalência de determinada doença ou característica em pessoas com determinado fator.

Para calcular o N amostral de um Estudo de Coorte ou de um Estudo Transversal, necessitamos ter uma ideia da proporção de pessoas no grupo exposto ao fator e no grupo não exposto. Significa, proporção de pessoas que têm determinado fator como, por exemplo, o número de pessoas que fumam ou têm desnutrição, a proporção esperada de doentes no grupo não exposto e uma das três alternativas: o risco relativo, a razão de chance ou a proporção esperada de doentes no grupo exposto. Como exemplo vamos calcular o tamanho de amostra de estudo Transversal (*cross-sectional*) para uma pesquisa sobre bronquite crônica e tabagismo, onde sabemos, através da revisão da literatura, que a proporção de tabagismo na população é 30%, entre os não fumantes a frequência da bronquite é 5% e entre os fumantes é 15%. Explicando, temos que a proporção de expostos ao tabagismo é 30%, ou seja, 70 não fumantes para 30 fumantes. Antes devemos confirmar o intervalo de confiança e o poder estatístico que queremos (95% e 80% respectivamente). Finalmente, o programa pede a razão de chance (*odds ratio*), risco relativo ou a proporção de doentes entre os expostos.

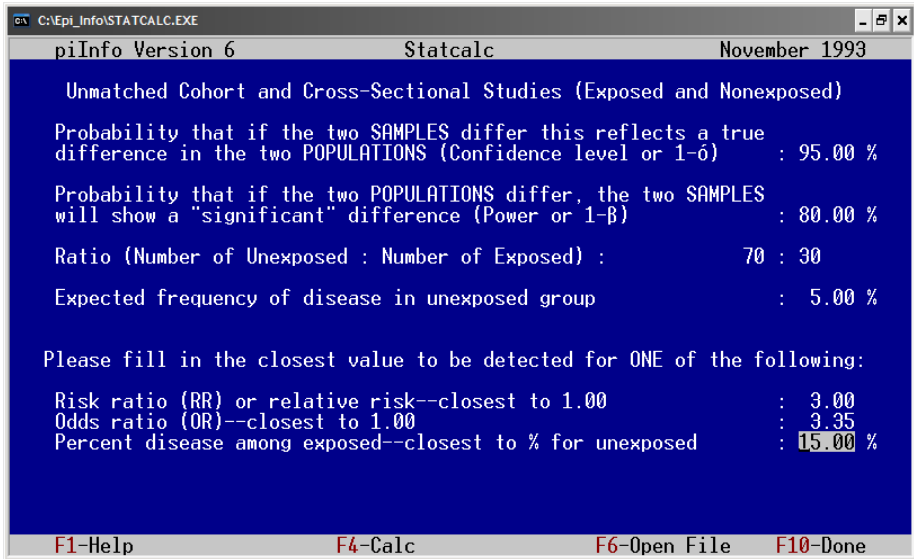


Figura 8.5. Janela de definição dos parâmetros para o cálculo amostral de uma pesquisa tipo Estudo de Coorte ou Transversal pelo *StatCalc*.

Fonte: Epi Info®, dados fictícios.

Como temos somente esse último valor, entramos 15% na última linha, conforme a figura 8.5. Clicando em F4 temos os resultados:

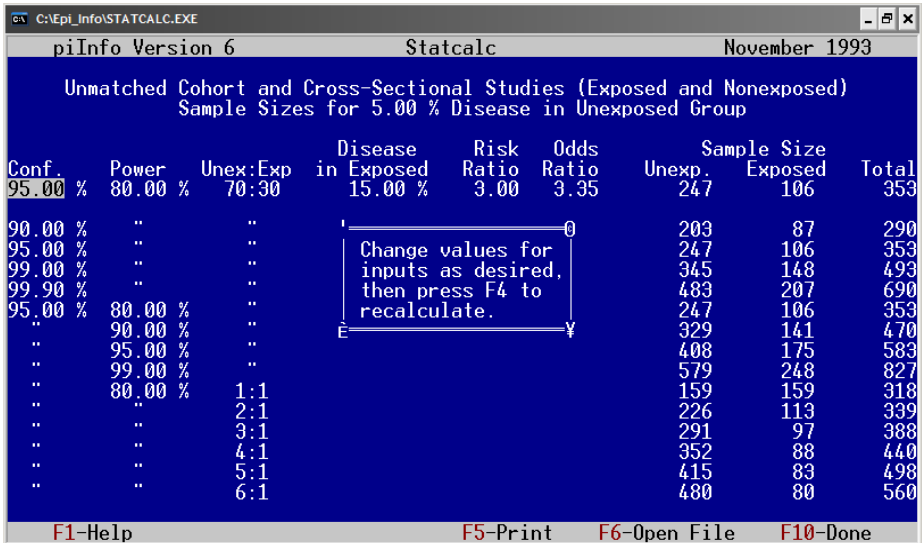


Figura 8.6. Janela de resultados do cálculo amostral de um Estudo de Coorte ou Transversal pelo *StatCalc*.

Fonte: Epi Info®, dados fictícios.



O resultado parece um pouco complicado, mas calcula valores para outros níveis de IC e Poder. O que nos interessa está na primeira linha de números. Temos o número de não expostos que seria 247 e de expostos 106, com o total de 353. Igualmente, podemos modificar os valores tanto da proporção de não expostos/expostos, e a proporção de doentes nos expostos. E devemos recalculá-los clicando F4.

*N amostral para estudo de caso controle*

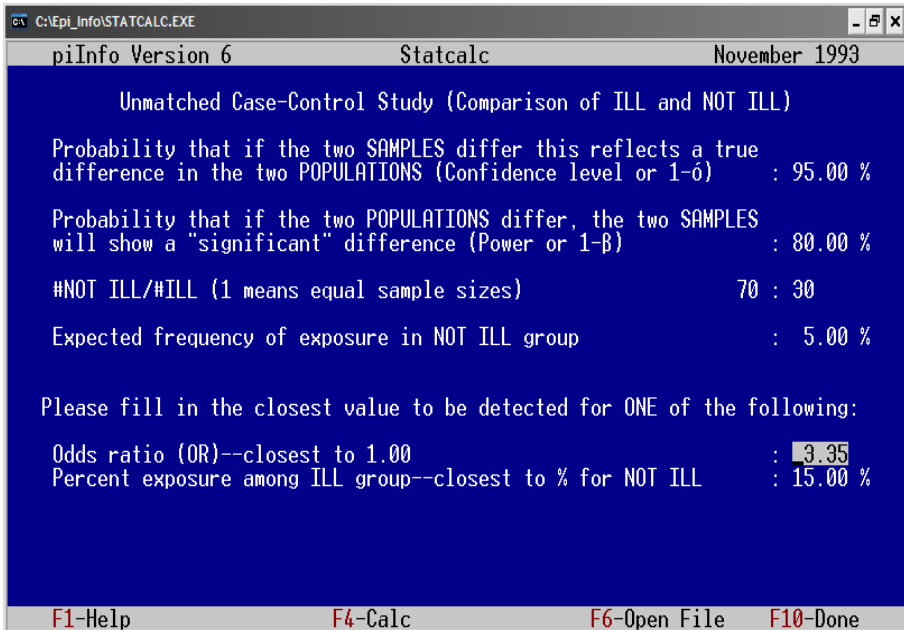


Figura 8.7. Janela de definição dos parâmetros para o cálculo amostral de uma pesquisa tipo Caso Controle pelo *StatCalc*.

Fonte: Epi Info®, dados fictícios.

Para cálculo do N amostral de um estudo de caso controle, as informações necessárias são quase as mesmas do Estudo de Coorte ou Transversal. A única diferença é que não temos o risco relativo, somente a razão de chance e a percentagem de expostos entre os doentes. O resultado também será parecido.

C:\Epi\_Info\STATCALC.EXE

piInfo Version 6 Statcalc November 1993

Unmatched Case-Control Study (Comparison of ILL and NOT ILL)  
Sample Sizes for 5.00 % Exposure in NOT ILL Group

Conf.	Power	NOT ILL :ILL	Exposure in ILL Ratio	Odds Ratio	Sample Size		Total
					NOT ILL	ILL	
95.00 %	80.00 %	70:30	15.00 %	3.35	247	106	353
90.00 %	"	"			203	87	290
95.00 %	"	"			247	106	353
99.00 %	"	"			345	148	493
99.90 %	"	"			483	207	690
95.00 %	80.00 %	"			247	106	353
"	90.00 %	"			329	141	470
"	95.00 %	"			408	175	583
"	99.00 %	"			579	248	827
"	80.00 %	1:1			159	159	318
"	"	2:1			226	113	339
"	"	3:1			291	97	388
"	"	4:1			352	88	440
"	"	5:1			415	83	498
"	"	6:1			480	80	560

Change values for inputs as desired, then press F4 to recalculate.

F1-Help F5-Print F6-Open File F10-Done

Figura 8.8. Janela de resultados do cálculo amostral de uma pesquisa Caso Controle pelo StatCalc.  
Fonte: Epi Info®, dados fictícios.

## Capítulo 9

# CONSTRUINDO E INTERPRETANDO TABELAS

Diferentes testes estatísticos são usados dependendo do tipo de variável que dispomos. Como vimos, as variáveis podem ser numéricas ou categóricas. As variáveis numéricas podem ser contínuas ou não. Exemplos de variáveis contínuas seriam a pressão arterial, o colesterol e o peso. Um exemplo de variável não contínua seria o número de mortes de um hospital.

De maneira geral podemos dizer que a análise dos dados vai ser baseada nas características da variável dependente e das variáveis independentes. A variável dependente é assim chamada, pois é a variável que desejamos observar a relação ou efeito das outras sobre elas. Dizemos que a variável dependente vai “depende” das outras, que são chamadas independentes. Por exemplo, dizemos que o número de mortes estará na dependência do tratamento. No caso a variável dependente vai ser o número de mortes e a variável independente o tratamento.

Assim definindo podemos sintetizar a abordagem de análise de dados conforme a tabela 9.1.

Tabela 9.1 Sistematização da análise dos dados conforme as características das variáveis dependente e independente.

Variável dependente	Variável independente	Teste	Interpretação
Categórica	Categórica	Qui-quadrado	Relação
Numérica	Categórica	t Student ou ANOVA	Comparação
Numérica	Numérica	Regressão linear ou correlação	Correlação
Categórica dicotômica	Numérica	Regressão logística	Associação

### Calculando o $\chi^2$ (qui-quadrado) para variáveis categóricas

O qui-quadrado (identificado pela letra grega qui -  $\chi$  ao quadrado), é o teste estatístico utilizado para determinar se a distribuição de uma tabela teria ocorrido por chance (hipótese nula) ou não (hipótese alternativa). O  $\chi^2$  é

utilizado para estudar a relação entre duas variáveis categóricas ou variáveis numéricas transformadas em categóricas, por exemplo, faixas etárias. Por isto é o teste estatístico calculado pelo comando *Tables* do Epi Info®.

Temos como exemplo hipotético a frequência do uso de óculos entre homens e mulheres. Estudamos 100 pessoas, 50 homens e 50 mulheres, deste grupo teríamos igualmente 50 pessoas usando e 50 pessoas não usando óculos conforme a tabela 9.2.

Tabela 9.2. Distribuição do número de pessoas que usa ou não óculos entre homens e mulheres.

	Usa óculos	Não usa	Total
Homens	30 (60%)	20 (40%)	50 (50%)
Mulheres	20 (40%)	30 (60%)	50 (50%)
Total	50 (50%)	50(50%)	100

Fonte: dados fictícios.

Observamos que o número de homens usando óculos foi de  $30/50 = 60\%$ , enquanto essa frequência nas mulheres foi de  $20/50 = 40\%$ . Notamos assim que os homens usam proporcionalmente mais óculos que as mulheres. Se o gênero não tivesse relação nenhuma com o uso de óculos, 50% dos homens e 50% das mulheres usariam óculos. Assim, na interpretação de uma tabela procuramos sempre comparar o percentual dos totais (fora da tabela) com os percentuais observados. Os percentuais fora da tabela (totais) são os percentuais ESPERADOS, e os percentuais dentro da tabela são os percentuais OBSERVADOS. Se os percentuais observados forem diferentes que os esperados, mais do que a chance dizemos que existe uma relação significativa entre as variáveis, no caso gênero e uso de óculos. O teste do  $\chi^2$  é nada mais que a o somatório dos quadrados das diferenças entre os valores observados e esperados, divididos pelos valores esperados como veremos a diante. Felizmente, na maioria das vezes os programas fazem todo esse cálculo para nós, mas veremos como é feito só para matar nossa curiosidade. Para calcular os valores esperados multiplicamos a percentagem dos valores totais das linhas. No caso teríamos 50% de homens vezes 50 usando óculos, assim o esperado seria termos 25 homens usando óculos. Se não tivermos os valores percentuais podemos usar as fórmulas a partir da tabela:

- Classicamente denominamos a tabela 2x2 com quatro letras a, b, c, e d:

	Usa óculos	Não usa	Total
Homens	a	b	a+b=50
Mulheres	c	d	c+d=50
Total	a+c=50	b+d=50	a+b+c+d=100

Os valores esperados de “a” seria o total da linha (a+b) vezes o total da coluna (a+c) dividido pelo total geral (a+b+c+d), ou seja,  $50 \times 50 / 100 = 25$ .

- Usando o mesmo princípio, todas as fórmulas seriam:

$$a \text{ esperado} = \frac{(a+b) \times (a+c)}{(a+b+c+d)}; \quad b \text{ esperado} = \frac{(a+b) \times (b+d)}{(a+b+c+d)},$$

$$c \text{ esperado} = \frac{(c+d) \times (a+c)}{(a+b+c+d)}; \quad d \text{ esperado} = \frac{(b+d) \times (d+c)}{(a+b+c+d)}$$

- Assim, a tabela esperada se a distribuição ocorresse ao acaso seria:

	Usa óculos	Não usa	Total
Homens	Ea=25	Eb=25	50
Mulheres	Ec=25	Ed=25	50
Total	50	50	100

Notamos que na segunda tabela (tabela esperada)  $25/50=50\%$  dos homens usam óculos, o mesmo esperado para as mulheres. A principal característica dessa tabela é o fato de ser “homogênea”, ou seja, que a distribuição nas casas obedece à distribuição dos subtotais. Finalmente, poderemos calcular o qui-quadrado, que é a soma dos resultados obtidos pela divisão do quadrado das diferenças entre o observado e o esperado de cada célula pelo esperado de cada célula:

$$\chi^2 = \frac{(O_a - E_a)^2}{E_a} + \frac{(O_b - E_b)^2}{E_b} + \frac{(O_c - E_c)^2}{E_c} + \frac{(O_d - E_d)^2}{E_d}$$

$$\chi^2 = \frac{(30-25)^2}{25} + \frac{(20-25)^2}{25} + \frac{(20-25)^2}{25} + \frac{(30-25)^2}{25} = 4,0$$

O valor crítico ( $p < 0,05$ ) do qui-quadrado para uma tabela 2x2 (1 grau de liberdade<sup>1</sup>) é 3,84. Poderemos buscar em uma tabela de distribuição do qui-quadrado e observaremos que, para 1 grau de liberdade, qui-quadrado = 4,0 temos  $p < 0,0455$ . Concluimos que, com a chance de 4,6% de estarmos enganados, existe diferença estatística no uso de óculos entre homens e mulheres. Importante dizer que o qui-quadrado observa a distribuição na tabela como um todo. Não podemos, com o resultado do qui-quadrado, dizer que os homens usam significativamente mais óculos que as mulheres. Mas podemos dizer que a proporção de homens usando óculos foi maior que a de mulheres e que o qui-quadrado foi significativo. Poderemos calcular o qui-quadrado para tabelas nxn (3x2, 4x2, 4x4 etc.), utilizando o mesmo cálculo realizado.

### Utilizando o *StatCalc* para calcular o Qui-Quadrado

O programa *Tables 2xN* calcula o qui-quadrado de tabelas de distribuição sendo a variável da coluna com dois níveis. Como o Epi Info® foi originalmente criado pensando em uma análise de dados epidemiológicos, essa variável da coluna é chamada de *Disease* (doença) possuindo os níveis positivo (com doença) e negativo (sem doença), como pode ser observado na figura 9.1. A outra variável é chamada de exposição que aceita dois ou mais níveis.

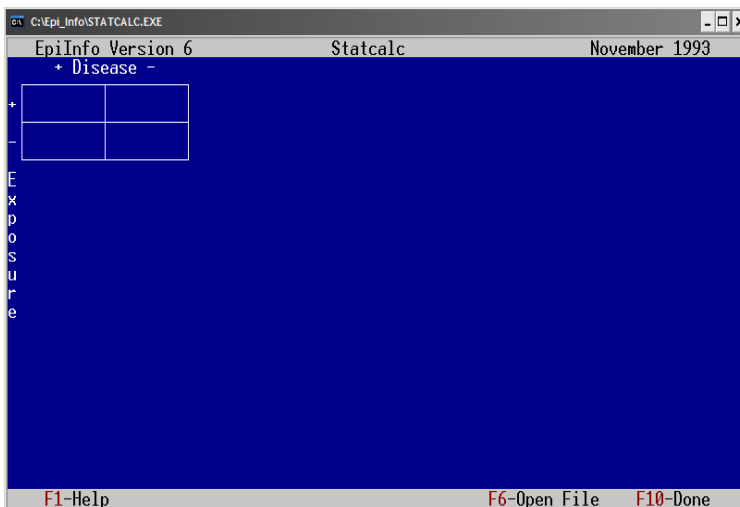


Figura 9.1 Janela de definição dos valores da tabela a ser avaliada pelo *StatCalc*.

Fonte: Epi Info®.

<sup>1</sup> A tabela 2x2 tem um grau de liberdade, pois, sabendo os totais, necessitamos saber somente o valor de uma casa, para podermos calcular as outras. Grau de liberdade é um parâmetro estatístico utilizado para calcular o valor crítico do teste que utilizamos, nesse caso o qui-quadrado.

O *StatCalc* também não aceita o comando do *Mouse*, para navegar nele devemos apertar o *Enter* ( $\zeta$ ) e os comandos de função que ficam na parte de cima do teclado. A sequência de entrada dos números também não muda. Começa no canto esquerdo depois na célula da direita para retornar a esquerda já na célula em baixo e finalmente na célula da direita.

- Vamos digitar a seguinte tabela:

	Falecidos	Sobreviventes	Total
Tratamento	40	232	272
Controle	35	93	128
Total	75	325	400

Devemos colocar primeiro o 40, que corresponde ao número de falecidos (com doença) e em tratamento (exposto). Apertamos o *Enter* e depois 232, sobreviventes em tratamento e *Enter* novamente. Logo a seguir colocamos 35 e *Enter*, e finalmente digitamos 325. Quando terminamos a tabela devemos apertar o *Enter* duas vezes ou *F4*, se seguirmos digitando números logo após cada *Enter*, uma nova linha aparecerá. Após apertarmos dois *Enter* seguidos os resultados da análise da tabela aparecerão, segundo a figura 9.2. No caso estamos trabalhando com uma tabela 2x2, assim os resultados da Razão de Chance (Odds ratio) e do Risco Relativo com seus respectivos intervalos de confiança aparecerão. Discutiremos esses resultados quando estivermos falando sobre a avaliação do risco, no décimo primeiro capítulo deste livro.

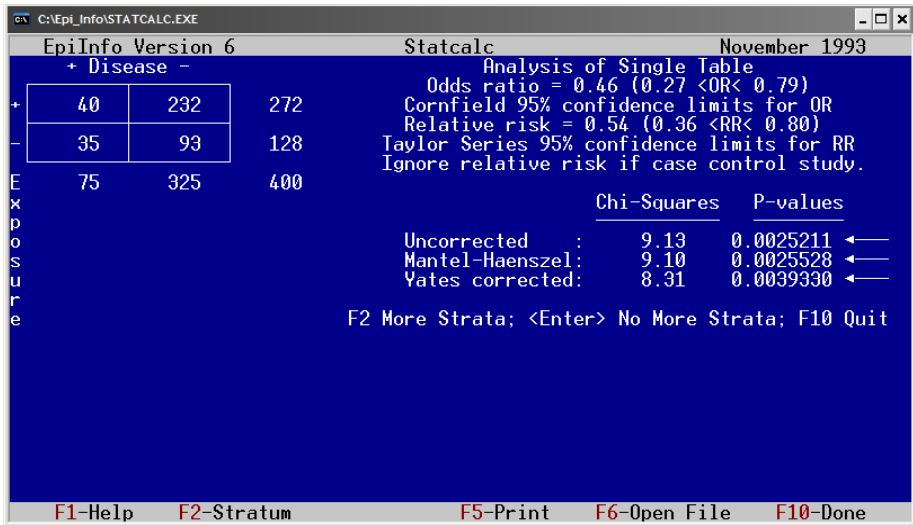


Figura 9.2. Janela com resultados da análise da tabela pelo *StatCalc*.

Fonte: Epi Info®, dados fictícios.

Nota-se uma das características do Epi Info® é de colocar uma seta ressaltando que o teste de significância é positivo. Para nova análise aperta-se a função *F2*, se quisermos imprimir *F5*, se quisermos terminar *F10*.



## Exercício:

1. Calcule o Qui-Quadrado para a tabela seguinte:

	Diabéticos	Não diabéticos
Fumantes	39	59
Ex-Fumantes	54	62
Nunca Fumantes	28	45

## Construindo tabelas usando o Analysis Data do Epi Info®: comando *Tables*

O comando *Tables* é o terceiro comando da secção *Statistics* e fornece alguns resultados estatísticos. Como se trata de tabelas teremos assim duas variáveis categóricas. Como vimos o teste estatístico para testar a relação entre duas variáveis categóricas é o Qui-Quadrado. O Qui-quadrado nada mais faz que determinar se existe um “balanço” das casas na tabela observada ao comparar essa tabela a uma tabela aleatória, chamada de tabela esperada. Se a tabela observada tiver uma distribuição nas casas diferente das esperadas dizemos que uma variável está relacionada com a outra, pois produz uma distribuição não aleatória. Vamos ver um exemplo, usando o banco de dados *Avaliacao.mdb* e as variáveis: *ESTADOCIV* e *GENERO*.

Na janela do comando temos dois compartimentos importantes observados na figura 9.3.

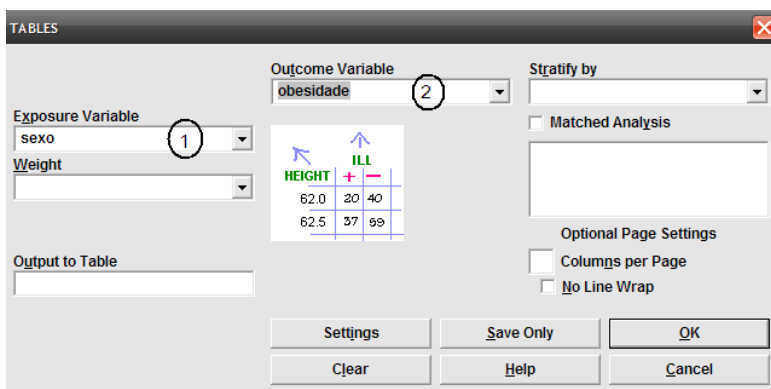


Figura 9.3. Janela do comando *Tables* do *Analysis* mostrando os compartimentos de definição das variáveis da tabela.

Fonte: Epi Info®.

Os valores vão ficar na vertical no compartimento chamado *Exposure Variable (1)* e na horizontal no compartimento *Outcome Variable (2)*. O motivo pelo qual esses compartimentos são chamados dessa forma é devido ao fato do Epi Info® ter sido projetado para o trabalho com levantamentos epidemiológicos. No nosso caso não nos interessaria muito qual a variável colocar como variável expositiva, mas como o sexo é que define qual estado civil, na nossa hipótese este faz mais sentido ser colocado ali. Os testes estatísticos são fornecidos automaticamente sem necessitarmos de qualquer configuração manual. Após selecionadas as variáveis podemos clicar OK.

No *Analysis Output*, observamos a tabela, na figura 9.4, com algumas análises estatísticas que vamos interpretar. Primeiro observamos como o programa dispõe os resultados. Dentro de cada casa da tabela temos a frequência, embaixo temos a percentagem na linha e a percentagem da coluna. No nosso caso observamos que 13 mulheres (37%) são casadas contra 22 (71%) dos homens, já a proporção de mulheres viúvas é bem maior que de viúvos, 40% contra 6,5% respectivamente(1). Esse achado corrobora com a nossa hipótese. Abaixo da tabela temos os resultados da análise estatística. O qui-quadrado foi 11,2 com um  $p = 0,0104$ , sendo significativo.

TABLES sexo obesidade

[Next Procedure](#)

[Forward](#)

**OBESIDADE**

sexo	magro	normal	obeso	sobrepeso	TOTAL
<b>Fem</b>	29	173	28	79	309
Row %	9,4	56,0	9,1	25,6	100,0
Col %	72,5	42,7	50,0	34,5	42,3
<b>Masc</b>	11	232	28	150	421
Row %	2,6	55,1	6,7	35,6	100,0
Col %	27,5	57,3	50,0	65,5	57,7
<b>TOTAL</b>	40	405	56	229	730
Row %	5,5	55,5	7,7	31,4	100,0
Col %	100,0	100,0	100,0	100,0	100,0

**Single Table Analysis**

**Chi-square df Probability**

22,0435    0,0001    2

Figura 9.4. Janela do resultado do comando *Tables* destacando alguns resultados.

Fonte: Epi Info®, arquivo *Avaliacao.mdb*.

2. Usando o banco de dados *ApoE\**, tabela Grupo1, construa e interprete as tabelas:
  - a. SEXO X OBESIDADE.
  - b. SEXO X HIPERTENSAO.
  - c. SEXO X DISLIPIDEMIA.

\*Poderá usar o banco de dados homônimo que se encontra no folder *Dados* do site [www.pucrs.br/igg/epiinfo](http://www.pucrs.br/igg/epiinfo).

## Capítulo 10

# CALCULANDO E COMPARANDO MÉDIAS

### Comparando duas médias pelo teste t de Student

O teste t de Student é utilizado para comparar médias de dois grupos; necessita que a variável a ser medida (dependente) seja contínua e, evidentemente, a variável a ser comparada (independente ou variável resposta) seja dicotômica (categórica com somente dois níveis). O teste é utilizado para resolver a seguinte hipótese nula:  $H_0 : \mu_1 - \mu_2 = 0$  (média do grupo 1 menos a média do grupo 2 igual a zero, ou seja, não há diferença entre as médias dos dois grupos), com duas possíveis hipóteses alternativas:  $H_a : \mu_1 - \mu_2 \neq 0$  teste de duas caudas (as duas médias são diferentes) ou  $H_a : \mu_1 - \mu_2 > 0$  teste de uma cauda (uma média de 1 é maior que a média de 2). Assume dois diferentes grupos, que poderão ser de tamanhos diferentes.

Para calcular o t não pareado utilizamos a seguinte fórmula:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left( \frac{\sum (X_1 - \bar{X}_1)^2 + \sum (X_2 - \bar{X}_2)^2}{n_1 + n_2 - 2} \right) \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

Observamos na fórmula que, quanto maior a diferença entre as médias, maior vai ser o t. No denominador existem dois fatores que podem influenciar: o somatório das diferenças entre os diversos valores de x de um grupo com a média do mesmo, isto determina que se a variância dos grupos for pequena, menor vai ser o denominador, portanto maior o valor do t; por último, devemos comentar que o número de observações influencia inversamente o denominador da equação. Por isto, quanto maior a diferença das médias, menor a variabilidade dos grupos e, principalmente, maior o número de observações, maior serão as chances de encontrarmos um t significativamente grande. Quanto maior o t maior seria a diferença entre os dois grupos e, conseqüentemente, menor o erro alfa. O erro alfa estaria relacionado à área comum existente entre os dois grupos.

## Comparando duas ou mais médias

Supomos que estamos testando o efeito de três tratamentos diferentes para hipertensão. Um grupo recebeu diurético outro betabloqueador e outro bloqueadores do cálcio. Se usarmos o t de Student para comparar as três médias, obteremos três valores diferentes de p (erro tipo I ou alfa). Assumindo um alfa de 0,05, teremos que a probabilidade de não estarmos errados ao rejeitar a hipótese nula para cada um dos testes é de 95%. Entretanto, a probabilidade de não estarmos errados ao rejeitar a hipótese nula, em nenhum dos testes, será o produto das três possibilidades, ou seja:  $0,95 \times 0,95 \times 0,95 = 0,86$ . A probabilidade de não cometermos o erro tipo I em nenhuma das três comparações possíveis é, então, 0,86. Em outras palavras, a probabilidade de ocorrer um erro tipo I nas três comparações será de 0,14. Por isso, a probabilidade de não ocorrer um erro tipo I, rejeitarmos a hipótese nula quando na realidade ela for verdadeira, nessa tríplice comparação será maior que a desejada de 0,05. Quanto maior o número de comparações, maior será a possibilidade de ocorrer um erro tipo I.

## Entendendo a Análise de Variância

A técnica apropriada para analisar variáveis contínuas quando desejamos comparar dois ou mais grupos (variável categórica) é a Análise de Variância, também conhecida como ANOVA.

Os princípios envolvidos na Análise de Variância são os mesmos que o t-teste. Dentro da hipótese nula, deveremos ter a seguinte situação: existe uma grande população e, se tivermos diversas amostras de um tamanho determinado dessa população, teremos várias médias, que devem variar, aleatoriamente, entorno da média global de toda a população. Se as amostras variarem em torno da média total, mais do que o esperado pela chance, talvez tenhamos algum fator, além da chance, influenciando essa alteração. Talvez as amostras colhidas não sejam da mesma população. Poderemos, assim, rejeitando a hipótese nula de que todas as médias são iguais, concluir que as médias diferem entre si mais do que o esperado pela chance.

A ANOVA calcula o F estatístico usando uma fórmula menos complicada, mas mais trabalhosa que a do cálculo do t de Student. Tentaremos explicar um pouco, mas não se preocupe se não conseguir entender o todo. Essencialmente, queremos observar se a variabilidade dos valores entre os grupos (externa) é maior do que a variabilidade dos valores dentro de cada grupo (interna). As variabilidades são calculadas pela soma dos valores dos grupos e as somas

dos quadrados desses valores (no inglês SS – sum of square). Inicialmente, calculamos o que se chama de soma dos quadrados total (SS total) e soma dos quadrados externa (SS externa), usando as seguintes fórmulas:

$$SS \text{ total} = \sum \sum x^2 - \frac{(\sum \sum x)^2}{\sum n} \quad SS \text{ externa} = \sum \left( \frac{(\sum x)^2}{n} \right) - \frac{(\sum \sum x)^2}{\sum n}$$

$$SS \text{ interna} = SS \text{ total} - SS \text{ externa}$$

onde  $\sum \sum x^2$  = soma dos quadrados dos valores de todos os grupos

$\sum \sum x$  = soma dos valores de todos os grupos

$\sum n$  = número de observações em todos os grupos

A SS total é, assim, calculada pela soma dos quadrados dos valores dos grupos menos o que poderemos chamar de média dos quadrados: o quadrado da soma dos valores dos grupos dividido pelo número total de observações. A SS externa, que representa a variabilidade entre os grupos, é calculada pela soma das razões entre os quadrados das somas dos valores de cada grupo e seu correspondente número de observações, menos a média dos quadrados. A soma dos quadrados interna (SS interna), que corresponde à variabilidade dentro dos grupos, será calculada, assim pela diferença entre o SS total e o SS externo. Finalmente, dividimos tanto a SS interna quanto a SS externa pelos seus correspondentes graus de liberdade. Os graus de liberdade interna são calculados pelo número de grupos menos 1, enquanto os graus de liberdade externa são o número total de observações de todos os grupos menos o número total de grupos. Veremos mais adiante um exemplo prático. O valor do F será então o resultado da divisão da média externa pela interna. O F compara assim as duas variabilidades entre os grupos e dentro dos grupos. Quanto maior a média externa em relação à média interna, maior será o valor de F. Valores da distribuição do F aparecem em várias tabelas, e se o valor obtido da nossa análise for maior que o valor crítico tabelado, podemos rejeitar a hipótese nula. O F assume os dois tipos de graus de liberdades vistos anteriormente: um fornecido pela variância entregrupos e outro pela variância interna dos grupos. O primeiro grau de liberdade é calculado pelo número de grupos comparados, chamado de  $k$ , subtraído de 1 ( $k-1$ ). O grau de liberdade correspondente à variância interna é igual a  $N-k$ , onde  $N$  é o número total de resultados (se tivermos três experimentos realizados em 100 pessoas teremos  $300-3=297$  graus de liberdade). Teremos a oportunidade de observar na prática como esses valores são obtidos pela ANOVA e como o F é calculado.

## Configurando o computador para realizar cálculos estatísticos

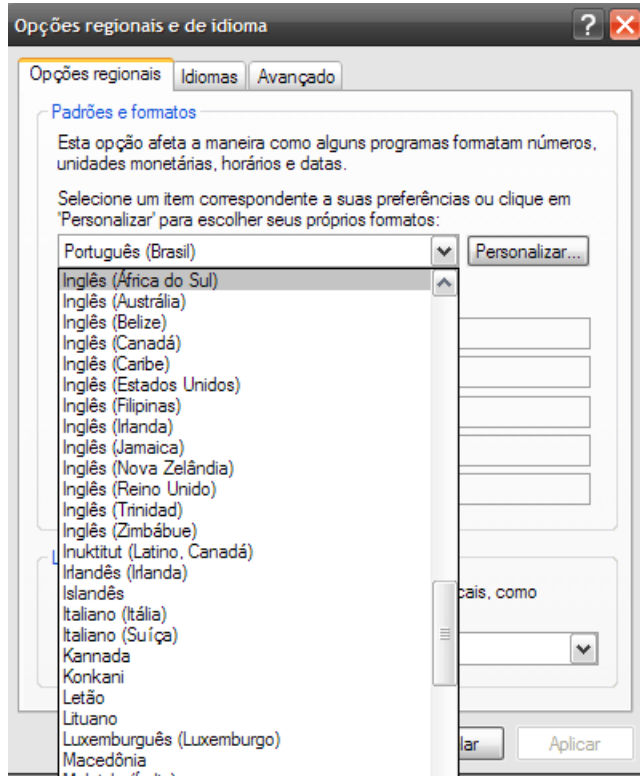


Figura 10.1. Janela de seleção da língua do sistema do computador.  
Fonte: Windows 7.

Na maioria das vezes nossos computadores estão configurados para utilizar a língua portuguesa. A língua portuguesa utiliza vírgula no lugar de ponto decimal. A maioria das versões de *Windows* faz com que o Epi Info<sup>®</sup> se “atrapalhe” ao fazer cálculos usando vírgula como marcador decimal. Isso faz com que variáveis com valores decimais como peso ou altura tenham suas médias alteradas. Esse problema também poderá afetar outros cálculos estatísticos. Por isso, devemos alterar as *Opções Regionais* (*Regional Options*) na janela de *Painel de Controle* do *Windows*. Para tanto devemos abrir o Meu computador e localizar a pasta *Painel de Controle*, abrir *Opções Regionais* e alterá-las escolhendo uma das tantas línguas inglesas, observadas na figura 10.1. Em alguns computadores isso não é necessário, como no *Windows 2000*, por exemplo. Recomenda-se testar o Epi Info<sup>®</sup> antes.

## Calculando e comparando médias usando o Epi Info®: comando *Means*

Podemos calcular médias e comparar as mesmas usando testes estatísticos através do quinto comando do *Statistics* chamado *Means*. Normalmente, testes estatísticos são usados para verificar se as diferenças das médias de dois ou mais grupos poderiam ter ocorrido ao acaso (Hipótese nula). Vimos que o teste t de Student é usado para comparar duas médias e, quando temos mais que duas médias, utilizamos a ANOVA (Análise de Variância). Na realidade ambos os testes se equivalem, porque a ANOVA também serve para testar diferenças entre as médias de dois grupos. Por isso o Epi Info® calcula também a ANOVA. No exercício vamos calcular a média do peso para cada sexo usando o banco de dados *Avaliacao*. Na janela ilustrada na figura 10.2 temos como *Means of* a variável *IDADE*, e como *Cross-tabulate by Value of* a variável *GENERO*.

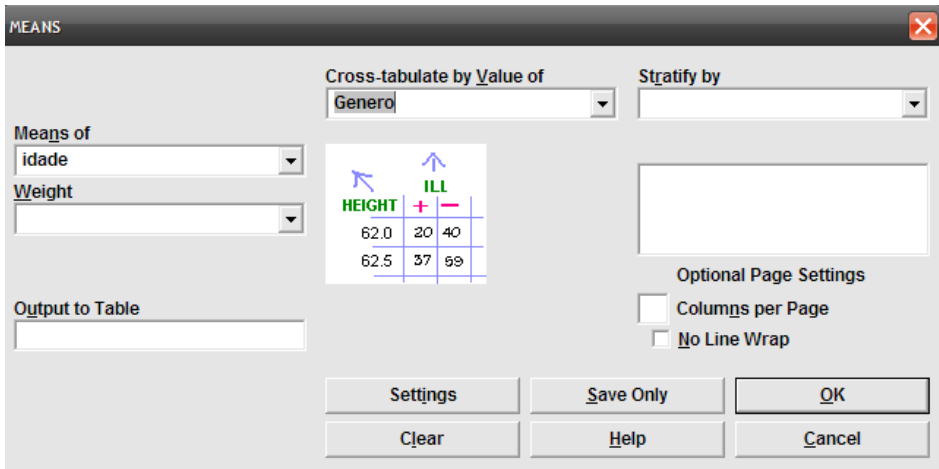


Figura 10.2. Janela do comando *Means* do *Analysis*.

Fonte: Epi Info®.

Temos que alterar um pouco a configuração, por isso vamos clicar em *Settings*.



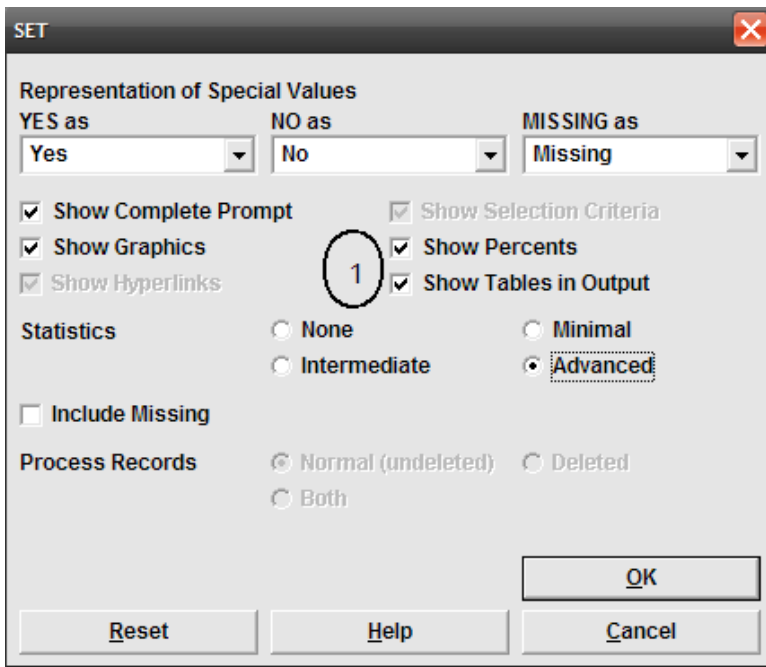


Figura 10.3. Janela da opção *Settings* do *Analysis* destacando as opções *Show Percents* e *Show Tables in Output*.

Fonte: Epi Info®.

A janela do *Settings* do comando *Means* (figura 10.3) é muito semelhante aos outros comandos do Epi Info®. Vemos na figura as configurações automáticas deste comando. O Epi Info®, utilizando uma única configuração, não permite configurações personalizadas para cada tipo de comando. Existem assim algumas configurações que não nos ajudam para o comando *Means*, são os casos das opções *Show Percents* e *Show Tables in Output* (1). Esta última opção cria uma tabela com todos os valores da variável, o que não nos interessa. Então vamos desseleccionar essas opções. Não existe diferença entre as opções de estatística intermediária e avançada. Clicamos *OK* para fechar essa janela e retornar à janela do *Means* e *OK* novamente para executar o comando. Essa configuração não é salva automaticamente e devemos repeti-la toda vez que quisermos fazer uma tabela. Podemos, todavia, alterar essa configuração no último comando do Epi Info: *Options Set*. Devemos resseleccionar essas opções para fazer tabelas.

O resultado do comando *Means* nos fornece vários cálculos que são muito importantes. Na figura 10.4 observamos que todos os resultados não cabem em uma só tela.

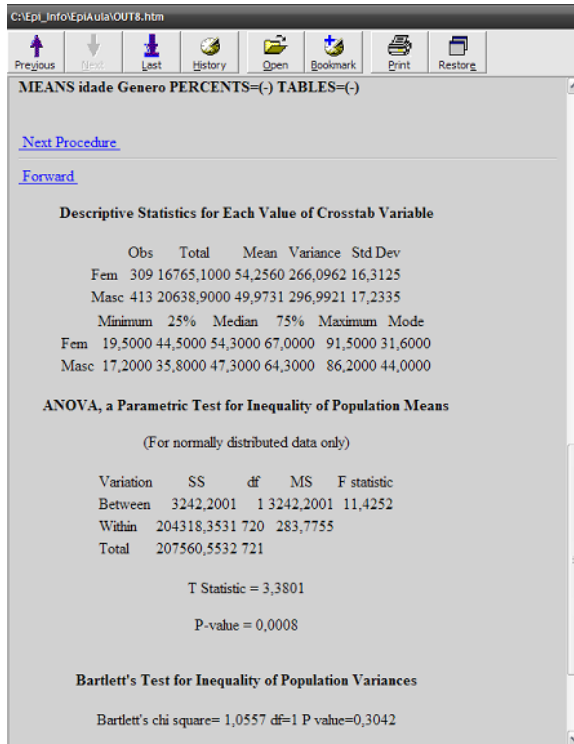


Figura 10.4. Janela com resultados do comando *Means* do *Analysis*.  
Fonte: Epi Info®, arquivo *Avaliacao.mdb*.

Descendendo de cima para baixo os resultados, temos uma análise descritiva dos resultados para cada valor da variável independente (no caso gênero). Temos uma linha para o gênero feminino e outra para o masculino. Inicialmente, o programa fornece o número de observações, depois o somatório das idades, que não nos interessa muito, e, finalmente, a média, variância e o desvio padrão. É importante, inicialmente, determinar se o número total das observações, número de pessoas no gênero masculino mais o número de pessoas no gênero feminino, é igual ao total de observações no banco de dados. Isso é importante para verificar se o programa está calculando adequadamente as médias com os valores não inteiros, quando estamos usando a linguagem portuguesa no sistema, ao contrário do orientado anteriormente. No caso temos o número total de 66 pessoas que corresponde ao número total de pessoas no banco de dados. Entretanto, a variável *IDADE* não se presta para observar se a linguagem afeta os cálculos, por que só temos valores inteiros (anos). Observando as médias, vemos que não existe muita diferença nas idades entre os dois gêneros, sendo as mulheres

um pouco mais idosas. Depois devemos observar as variâncias, no caso, temos uma variância maior para as mulheres. Gostaria de destacar a importância do segundo grupo de análise que é justamente os valores mínimos e máximos. Eles nos ajudam a observar se existem observações com valores colocados errados. Por exemplo, se tivéssemos um valor mínimo de 5 poderíamos pensar que, para uma pessoa adulta, como todas as pessoas dessa amostra, esse valor não é correto. Assim observando, deveremos tentar corrigi-lo e repetir a análise.

Depois dessa análise descritiva, iniciam-se os resultados da ANOVA e do teste t, que, no nosso exercício, não foram significativas. O Epi Info<sup>®</sup> chama a média externa de MS Between e a interna de MS Within. O F estatístico é calculado então dividindo a média das somas entre os grupos (MS Between) e a média das somas dentro dos grupos (MS Within) ( $135,9363/153,3334=0,89$ ). É importante observarmos a similaridade da ANOVA e o t de Student quanto aos resultados. Apesar de terem um cálculo diferente chegam a resultados idênticos, pois se elevarmos o t de Student ao quadrado chegaremos ao mesmo valor do F calculado pela ANOVA ( $0,9416^2=0.8866$ ).

O teste mais importante quando estamos trabalhando com o teste t de Student e a ANOVA é se as variâncias são homogêneas ou não. Assim, o Epi Info<sup>®</sup> fornece o teste de Bartlett, mostrado na figura 10.5, que testa a probabilidade das variâncias serem heterogêneas ou não iguais. No nosso caso, foi significativo, observando existir diferenças nas variâncias dos dois gêneros. Como a variância dos dois gêneros não é homogênea, o resultado da ANOVA e do t de Student não é válido. Devemos usar um teste não paramétrico como o teste Kruskal-Wallis que é fornecido pelo programa. No exercício, esse teste não foi significativo, confirmando que não existem diferenças entre as idades dos homens e mulheres no grupo estudado.

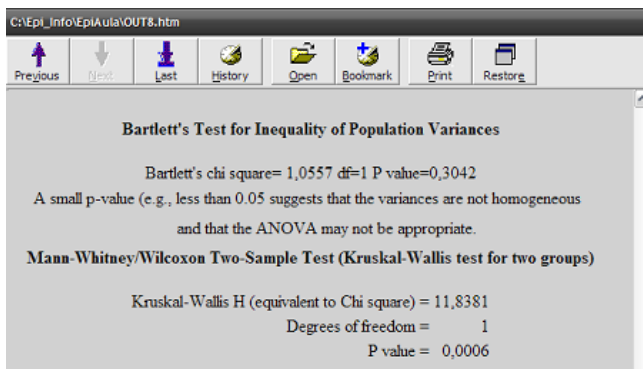


Figura 10.5. Janela com a continuação dos resultados do comando *Means* do *Analysis*.  
 Fonte: Epi Info<sup>®</sup>, arquivo *Avaliacao.mdb*.

## Comparando médias entre mais de três grupos

Num novo exercício, vamos comparar as médias de idade entre os grupos de tabagismo. Ao invés de colar a figura vou selecionar o resultado e colá-lo no texto:

### MEANS IDADE tabagismo PERCENTS=(-) TABLES=(-)

#### Descriptive Statistics for Each Value of Crosstab Variable

	Obs	Total	Mean	Variance	Std Dev	
<b>Atual</b>	19	921.0000	48.4737	68.3743	8.2689	
<b>Exfumante</b>	14	838.0000	59.8571	76.4396	8.7430	
<b>Nunca</b>	33	1763.0000	53.4242	208.7519	14.4482	
	<i>Minimum</i>	<i>25%</i>	<i>Median</i>	<i>75%</i>	<i>Maximum</i>	<i>Mode</i>
<b>Atual</b>	34.0000	43.0000	50.0000	54.0000	65.0000	50.0000
<b>Exfumante</b>	46.0000	53.0000	59.5000	67.0000	70.0000	67.0000
<b>Nunca</b>	34.0000	40.0000	48.0000	68.0000	83.0000	75.0000

#### ANOVA, a Parametric Test for Inequality of Population Means

(For normally distributed data only)

Variation	SS	df	MS	F statistic
<b>Between</b>	1044.7610	2	522.3805	3.6959
<b>Within</b>	8904.5117	63	141.3415	
<b>Total</b>	9949.2727	65		

P-value =0.0304

#### Bartlett's Test for Inequality of Population Variances

Bartlett's chi square=	8.3767	df=2	P value=0.0152
------------------------	--------	------	----------------

A small p-value (e.g., less than 0.05) suggests that the variances are not homogeneous and that the ANOVA may not be appropriate.

**Mann-Whitney/Wilcoxon Two-Sample Test (Kruskal-Wallis test for two groups)**

Kruskal-Wallis H (equivalent to Chi square) =	7.0843
Degrees of freedom =	2
P value =	0.0290

Vemos que os fumantes atuais são mais jovens que os outros dois grupos. Vemos também que o grupo dos nunca fumantes apresenta uma maior variância que os outros. O teste de Batlett foi significativo, orientando para não se usar o resultado da ANOVA, assim usaremos, novamente, o teste de Kruskal-Wallis, que foi significativo.

**Testes de multicomparação (teste de Bonferroni)**

Quando calculamos a ANOVA para mais de dois grupos e chegamos a um teste significativo, nos deparamos com o seguinte dilema: sabemos que existem diferenças entre os grupos, pelo menos um grupo é diferente dos demais, mas será que todos são diferentes entre si? Temos, no nosso exemplo, que a idade dos atuais fumantes é a menor de todas, assim ela, com certeza, é significativamente menor que a idade dos ex-fumantes, que é a maior. Mas como saber se a idade dos atuais fumantes é significativamente menor que os nunca fumantes e se os nunca fumantes são significativamente mais jovens que os ex-fumantes? Para resolver este dilema existem os chamados testes de multicomparação. Existem vários, mas o Epi Info<sup>®</sup> não fornece nenhum. Um teste que podemos realizar é o teste de Bonferroni que ajusta o p crítico ao número de comparações. Assim, usando o mesmo motivo de o porquê da necessidade da ANOVA quando temos mais de dois grupos. Quando temos três grupos precisamos fazer três comparações: A com B; B com C e A com C. Se usarmos o limite de p de 0,05 temos ( $0,95 \times 0,95 \times 0,95 = 0,86$ ), se dividirmos 0,05 por 3 chegamos a um p de 0,01666, calculando o p total nas três comparações temos ( $0,983 \times 0,983 \times 0,983 = 0,95$ ). Assim, conseguimos ajustar o p para múltipla comparação, dividindo o mesmo pelo número de comparações necessárias para comparar todos os grupos. Podemos utilizar essa estratégia quando temos uma ANOVA significativa.

**Exercício:**

Usando o banco de dados *ApoE\**, tabela Grupo1, observar quais as variáveis numéricas (COLESTEROL, PAS, PAD, IMC, IDADE) têm suas médias significativamente diferentes entre APOE4 sim e não.

\*Poderá usar o banco de dados homônimo que se encontra no folder *Dados* do site [www.pucrs.br/igg/epiinfo](http://www.pucrs.br/igg/epiinfo).

## Capítulo 11

# ESTUDANDO VARIÁVEIS NUMÉRICAS COM A REGRESSÃO LINEAR

### Testando a relação entre variáveis numéricas contínuas

Quando desejamos estudar o comportamento de duas variáveis numéricas devemos, inicialmente, construir gráficos de dispersão XY (*Scatter XY*). Para observar algumas das características dos testes utilizados para testar a relação das variáveis numéricas vamos construir dois gráficos ALTURA x IMC e PESOATUAL x IMC.

Para tanto vamos utilizar o arquivo *Avaliacao.mdb*, não esquecendo de trocar o projeto se o mesmo não estiver ativo.

Para fazer o gráfico de dispersão XY, usa-se o comando *Graph* na secção *Statistics*. Escolhendo a opção *Scatter XY*, colocamos primeiro a variável X (ALTURA); depois a variável Y (IMC). Como resultado teremos o gráfico da figura 11.1.

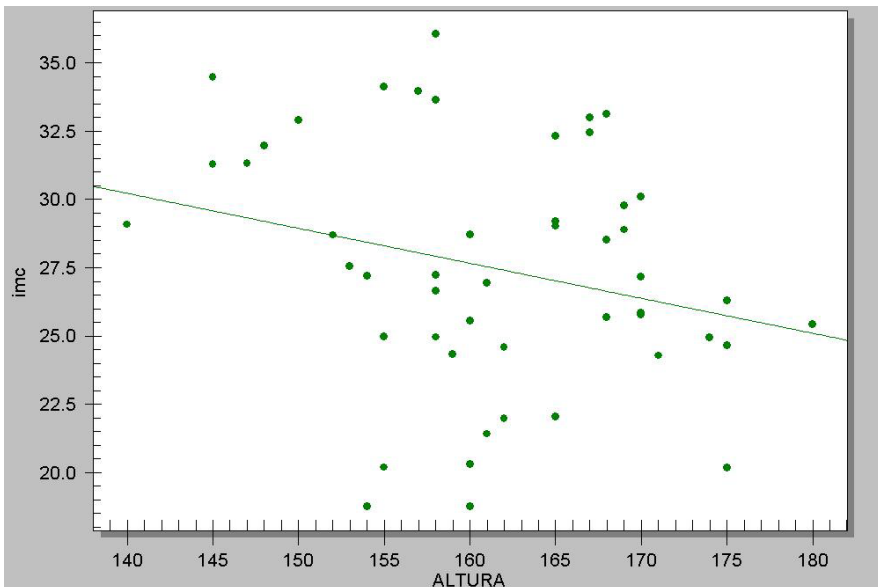


Figura 11.1 Gráfico da distribuição do IMC (Índice de Massa Corporal) em função da ALTURA criado pelo comando *Graph*.

Fonte: Epi Info®.

Observamos que o Epi Info® fornece automaticamente uma linha retilínea e pontos, que correspondem a cada uma das observações do banco de dados. Essa reta é a chamada Linha de Regressão que procura ilustrar, ou descrever, a relação entre as duas variáveis. Vamos ver, em seguida, como ela é calculada. Notamos que essa linha é descendente, mais ou menos metade das observações está acima e a outra, logicamente, está abaixo dessa linha. Vemos também que existe uma certa dispersão ao redor dessa linha, algumas observações estão próximas e outras mais afastadas. Vamos fazer o mesmo gráfico para IMC e PESOATUAL, simplesmente fechando a janela do gráfico e trocando, no *Program Editor* a palavra “altura” por “pesoatual” (não importa se for maiúscula ou minúscula) e clicamos no botão *Run This Command*, que é o último à direita no menu da janela do *Program Editor*.

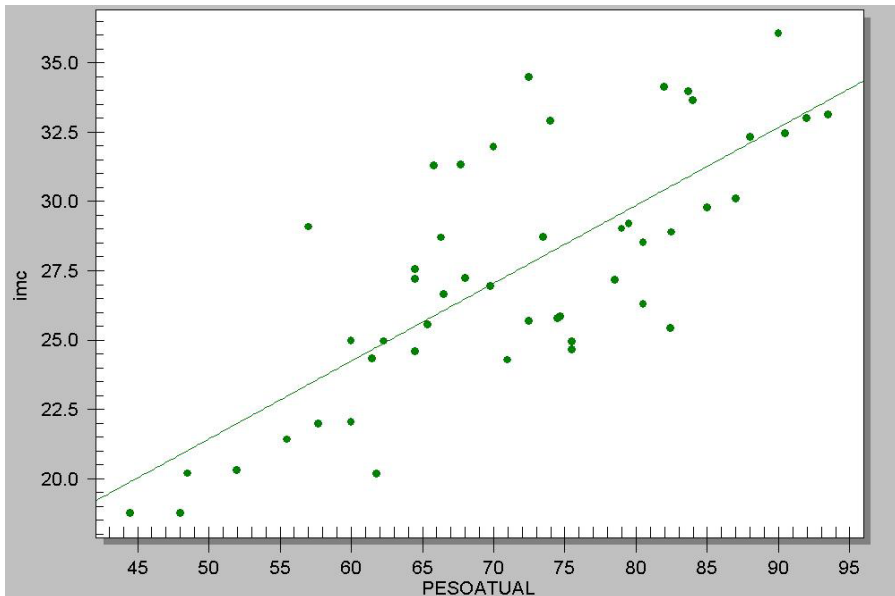


Figura 11.2. Distribuição do IMC (Índice de Massa Corporal) em função do PESOATUAL criado pelo comando *Graph*.

Fonte: Epi Info®.

Observamos o novo gráfico na figura 11.2 que apresenta algumas características semelhantes e outras diferentes. Igualmente ao gráfico anterior este gráfico também apresenta uma reta com mais ou menos metade da amostra acima e outra metade abaixo da reta. Observamos, entretanto, que a reta apresenta-se ascendente e as observações mais próximas da reta (menos dispersão). Existem, assim, dois importantes fatores a observar quando



analisamos a associação entre duas variáveis numéricas: o grau de dispersão das observações e a análise da reta. O grau de dispersão é medido pelo Coeficiente de Correlação e seu valor independe de qual variável é dependente ou independente; ela independe também do tipo de unidade utilizada para medir as variáveis (kg ou libras, metros ou polegadas, por exemplo).

## Entendendo o Coeficiente de Correlação

Dois variáveis contínuas podem estar correlacionadas uma com a outra. Por exemplo, pessoas mais altas tendem a serem mais pesadas. O Coeficiente de Correlação (também chamado de “ $r^2$ ”) é uma medida da associação entre duas variáveis contínuas. O  $r^2$  pode variar de 0 a 1. Se for zero, significa que as duas variáveis não têm nenhuma correlação. Se o  $r^2 = 1$ , significa que existe perfeita correlação, ou seja, uma variável pode prever outra. O coeficiente de correlação está relacionado com o grau de dispersão de duas variáveis. Quanto mais próximas forem as observações em torno de uma reta imaginária (linha de regressão), maior vai ser o  $r^2$ .

A fórmula para o cálculo do  $r$  é a seguinte:

$$r = \frac{n \sum (XY) - (\sum X)(\sum Y)}{\sqrt{n \sum X^2 - (\sum X)^2} \sqrt{n \sum Y^2 - (\sum Y)^2}}$$

Elevando ao quadrado teremos o  $r^2$ .

Observamos que poderíamos muito bem trocar a posição do X e do Y na fórmula sem alterar o seu resultado, corroborando com o que foi explicado.

## Entendendo a Regressão Linear

Para representarmos graficamente a associação entre duas variáveis, utilizamos a Regressão Linear, que poderá ser simples, se utilizarmos apenas uma variável dependente (resposta) e uma variável independente (fator). A Regressão Linear múltipla utiliza também uma variável resposta, que deverá ser contínua, e dois ou mais fatores, que poderão ser contínuos ou não. Iniciaremos com a Regressão simples. Ilustraremos essa relação pelo gráfico abaixo, com cinco diferentes pessoas, A, B, C, D e E com seus pesos e alturas.

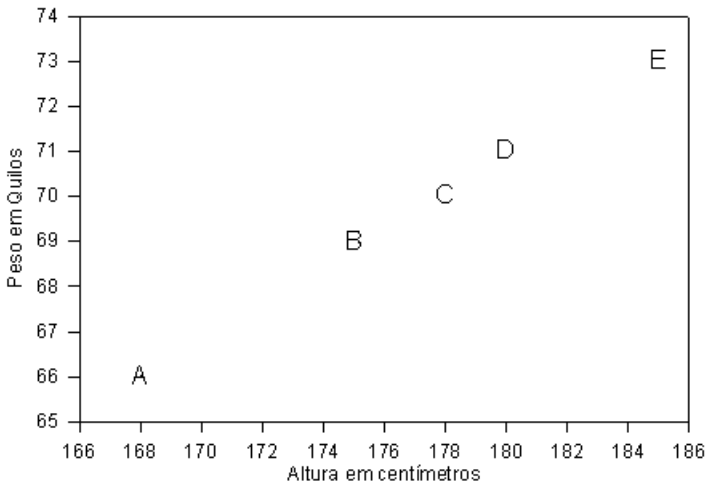


Figura 11.3. Distribuição de cinco pessoas conforme seu peso e altura.  
Fonte: dados fictícios.

Observamos que seria possível traçar uma linha entre os resultados. Anteriormente, estudamos como podemos descrever uma linha no gráfico. Chamávamos de Equação de Primeiro Grau da Reta. A fórmula da equação de primeiro grau era, basicamente,  $y=a+bx$ ; onde  $a$  era o ponto de intersecção da reta no eixo  $y$ , ou seja, quando  $x=0$ ; já o  $b$  era o ângulo da reta. A regressão linear calcula esses parâmetros da equação. Vejamos no gráfico da figura 11.4.

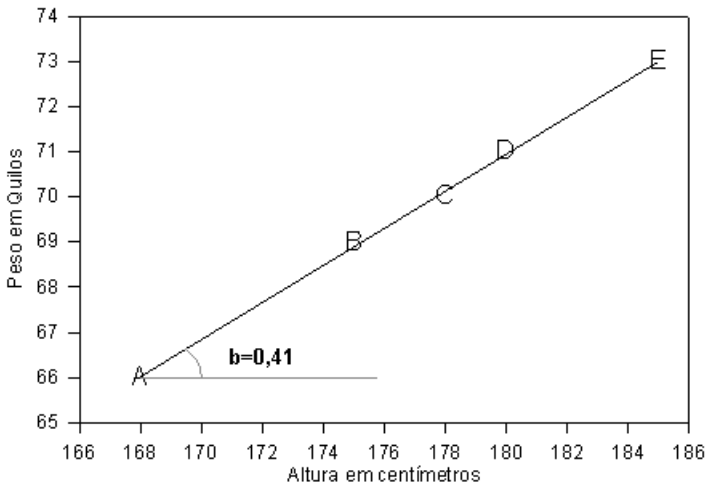


Figura 11.4. Distribuição de cinco pessoas conforme seu peso e altura e a regressão linear dessa distribuição, destacando o coeficiente de regressão.  
Fonte: dados fictícios.

A fórmula da linha seria  $\text{peso} = -2,85 + \text{Altura} * 0,41$ . *Intercept* é o  $a$  (ponto de intersecção) que corresponde ao valor do peso se a altura fosse igual a zero (-2,85), usando os resultados da Regressão. A altura é multiplicada por  $b$  que também é chamado de coeficiente de regressão e corresponde ao valor alterado no peso quando a altura sobe uma unidade de medida, no caso, uma pessoa com um centímetro a mais teria 0,41 quilos a mais que a outra. O valor de  $a$  é um tanto impróprio, pois nos diz pouca coisa a respeito dos dados que temos, pois jamais vamos encontrar alguém com altura zero, enquanto o  $b$  tem um valor expressivo, pois nos indica um valor preditivo de variação. Poderemos, com os valores de  $a$  e  $b$ , prever o quanto pesaria uma pessoa de 170 cm:

$\text{Peso} = a + b \times \text{altura}$ , ou seja  $-2,85 + 0,41 \times 170 = 66,85$  (quase 67) quilos.

Em termos estatísticos temos duas hipóteses nulas: que o Intercepto ( $a$ ) é igual a zero e que o Coeficiente de Regressão ( $b$ ) é igual a zero. Se o Coeficiente de Regressão fosse igual a zero, teríamos uma linha paralela ao eixo do X. Nesse caso, para qualquer valor de altura, o peso seria igual ao Intercepto, provando assim que a altura não afetaria o peso.

Vamos ver, no nosso exemplo, como o Epi Info® calcula esses parâmetros. Primeiro vamos calcular a Regressão para  $Y = \text{IMC}$  e  $X = \text{ALTURA}$ . Na secção *Advanced Statistics* clicamos em *Linear Regression* e obtemos a janela da figura 11.5.

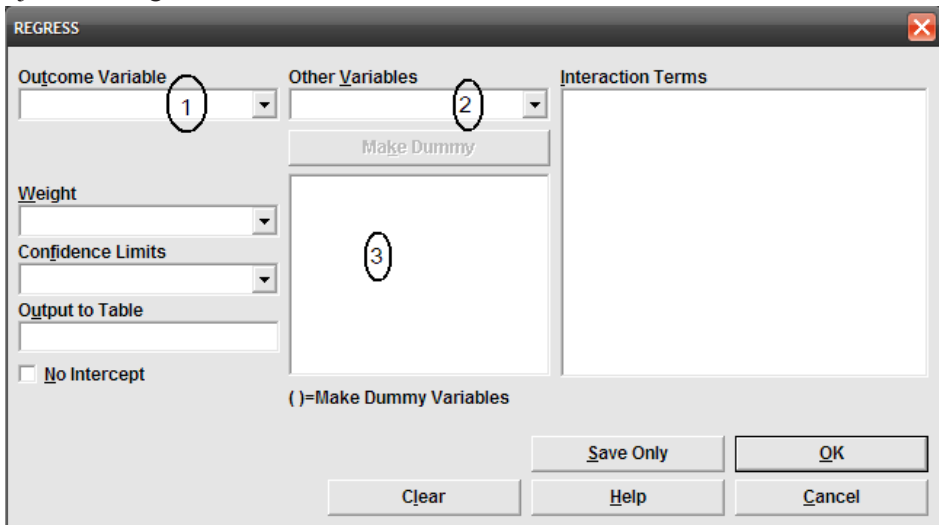


Figura 11.5. Janela do comando *Linear Regression* do *Analysis* destacando os compartimentos mais importantes. Fonte: Epi Info®.

No primeiro compartimento (1) selecionamos o que o Epi Info® chama de *Outcome Variable (Variável de Saída)*, que poderíamos chamar de Variável Resposta, mas usualmente chamamos de Variável Dependente. No caso específico é a variável Y da nossa equação, no nosso exemplo será o IMC. No compartimento *Other Variables (2)* seriam as Variáveis Independentes, como podemos notar pode ser mais do que uma. No nosso caso será ALTURA. Mais tarde usaremos mais do que uma variável independente, mas por enquanto vamos ver como o programa nos mostra os resultados:

### Linear Regression

Variable	Coefficient	Std Error	F-test	P-Value
ALTURA	-0.177	0.055	10.4070	0.001992
CONSTANT	56.365	8.849	40.5705	0.000000

Observamos os coeficientes dos parâmetros da Regressão: ALTURA=-0,177, com seu teste estatístico (F) e significância ( $p=0,001992$ ), e o que o Epi Info® chama de *CONSTANT*, que é o Intercepto, no caso 56,365, igualmente com seu teste estatístico e significância.

Observamos que o coeficiente para ALTURA é negativo, que concorda com o gráfico que mostra uma linha descendente e significativo. Raramente o Intercepto não será significativo, mas, de qualquer forma, ele não é importante para nós, pois não nos interessa saber o IMC de uma pessoa que não tem altura. A interpretação do coeficiente da altura é muito importante: na nossa amostra, pessoas com um centímetro a mais de altura são, em média, 0,177 kg/m<sup>2</sup> IMC menor.

Nossa equação ficaria assim:  $IMC = 56,365 - \text{Altura} * 0,177$

Vimos no gráfico que existia uma dispersão muito grande das observações. Essa dispersão é medida pelo “Coeficiente de Correlação”, a seguir:

Correlation Coefficient: $r^2=$	0.14
---------------------------------	------

Observamos que altura explicaria somente 14% da variabilidade do IMC.

Vamos ver o outro exemplo,  $IMC=PESOATUAL$ , bastando alterar o comando na janela Program Editor e clicar em Run This Command. Teremos as seguintes respostas:

Variable	Coefficient	Std Error	F-test	p-Value
pesoatual	0.277	0.034	65.3074	0.000000
CONSTANT	7.988	2.492	10.2784	0.002115
Correlation Coefficient: r <sup>2</sup> =		0.51		

A equação para o cálculo do IMC seria:  $IMC = 7,988 + \text{Pesoatual} * 0,277$

Observamos que, corroborando com o gráfico que fizemos anteriormente, o coeficiente de regressão para o peso é positivo e com valor absoluto maior que para a altura. Para cada diferença de um quilo há uma diferença de 0,277 kg/m<sup>2</sup> no IMC. Vemos também que o coeficiente de correlação é também maior, pois existe menos dispersão entre as duas variáveis. O peso explicaria mais de 50% da variabilidade do IMC.

O principal motivo para o peso ter uma maior correlação que a altura no IMC está relacionado à fórmula utilizada para o cálculo desse. Na fórmula do IMC, o peso é colocado diretamente, enquanto a altura é elevada ao quadrado. A elevação ao quadrado faz com que a altura tenha assim uma menor correlação.

## Usando variáveis categóricas na Regressão Linear

Vimos que, por definição, a Regressão Linear serve para observar a relação entre duas variáveis numéricas. Entretanto, podemos introduzir no modelo uma variável categórica, por exemplo, a variável GÊNERO. Neste caso, gostaríamos de descobrir se a linha de regressão  $IMC = \text{PESOATUAL}$  é igual ou diferente nos dois gêneros. Acrescentamos, assim, uma nova hipótese nula, de que o coeficiente para o gênero = 0. Se o coeficiente para o gênero for zero, o gênero não altera essa relação. Vamos colocar a variável GÊNERO no nosso comando: *REGRESS IMC = PESOATUAL GÊNERO*. Observamos os resultados:

Variable	Coefficient	Std Error	F-test	P-Value
pesoatual	0.284	0.029	96.5528	0.000000
genero (Masculino/Feminino)	-3.500	0.667	27.5633	0.000002
CONSTANT	9.173	2.107	18.9616	0.000051
Correlation Coefficient: r <sup>2</sup> =	0.66			

A equação para essa análise seria:  $IMC = 9,173 + \text{Pesoatual} * 0,284 + \text{Gênero} * -3,500$

Como a variável gEnero tem dois níveis, o programa já fornece a interpretação: Masculino/Feminino, ou seja, masculino em relação ao feminino. O programa determina, automaticamente, um valor numérico para cada valor de genero, sendo o masculino 1 e feminino zero, ou referencial. No caso, o coeficiente de gEnero fornece a diferença que existe nos homens em relação às mulheres. Se fossemos calcular o IMC das mulheres nós utilizaríamos somente a fórmula:

$$IMC = 9,173 + PESOATUAL * 0,284$$

A fórmula para os homens seria:

$$IMC = (9,173 - 3,5) + PESOATUAL * 0,284$$

Observamos que a única diferença é no Intercepto e essa diferença foi significativa. O coeficiente CONSTANT na nossa regressão se refere então ao intercepto das mulheres, sendo o coeficiente GENERO a diferença entre homens e mulheres no Intercepto (9173 para mulheres e 5,673 para homens). Nesta equação, homens e mulheres alterariam da mesma forma seus IMC, em função do peso. Notamos que o valor do coeficiente Pesoatual não alterou muito com a introdução desse fator, ao contrário do coeficiente intercepto, que alterou de 7,988 para 9,173. No gráfico teríamos: homens e mulheres com linhas paralelas, sendo a dos homens, 3,5 kg/m<sup>2</sup>, inferiormente colocada.

Vamos fazer esse gráfico para entender a equação gerada pela Regressão Linear. Calcularemos, usando a equação como seriam os valores de IMC preditivos para homens e mulheres. Primeiro devemos definir uma variável nova, chamaremos de IMCC. Usaremos o comando If, para calcular diferentemente os valores de IMCC para cada um dos gêneros. No primeiro compartimento do comando If colocaremos genero="feminino" e no compartimento Then atribuiremos a fórmula para o gênero feminino: assign IMCC= 9,173+0,284\*pesoatual, no compartimento else colocamos a fórmula para o masculino assign IMCC= 5.673 + 0,284\*pesoatual. Conforme a figura 11.6.

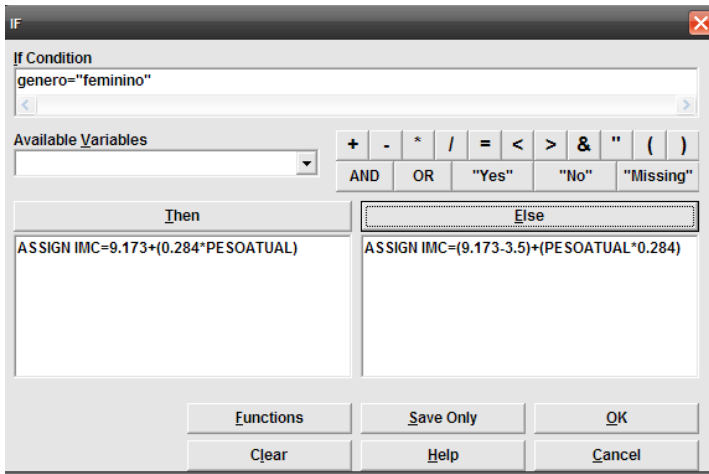


Figura 11.6. Comando *IF* do *Analysis* ilustrando a atribuição fórmulas distintas para cada valor de *GENERO* no cálculo do IMC.  
 Fonte: Epi Info®.

Finalmente podemos construir o gráfico PESOATUAL X IMCC usando a opção *Series* para cada valor de *Genero*, observado na figura 11.7.

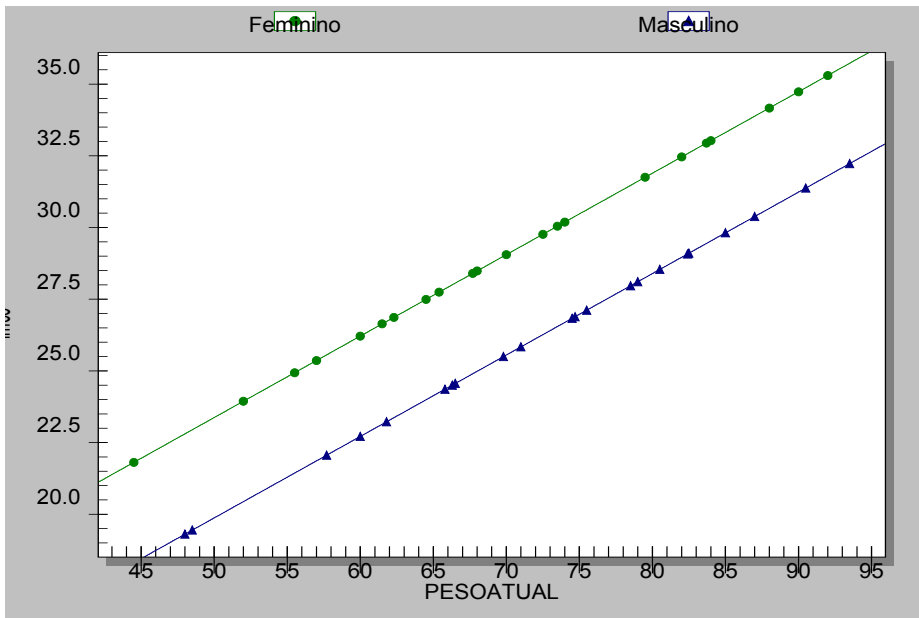


Figura 11.7. Distribuição do IMC calculado para cada valor de *GENERO* em função do *PESOATUAL*.  
 Fonte: Epi Info®, arquivo *Avaliacao.mdb*.

Vemos que, utilizando o resultado da regressão para calcular o IMC, teríamos duas retas paralelas com 3,5 unidades de IMC de diferença. Mas vamos criar o gráfico PESOATUAL X IMC para cada valor de gênero e observar se as duas retas são realmente paralelas. O gráfico apresenta-se na figura 11.8.

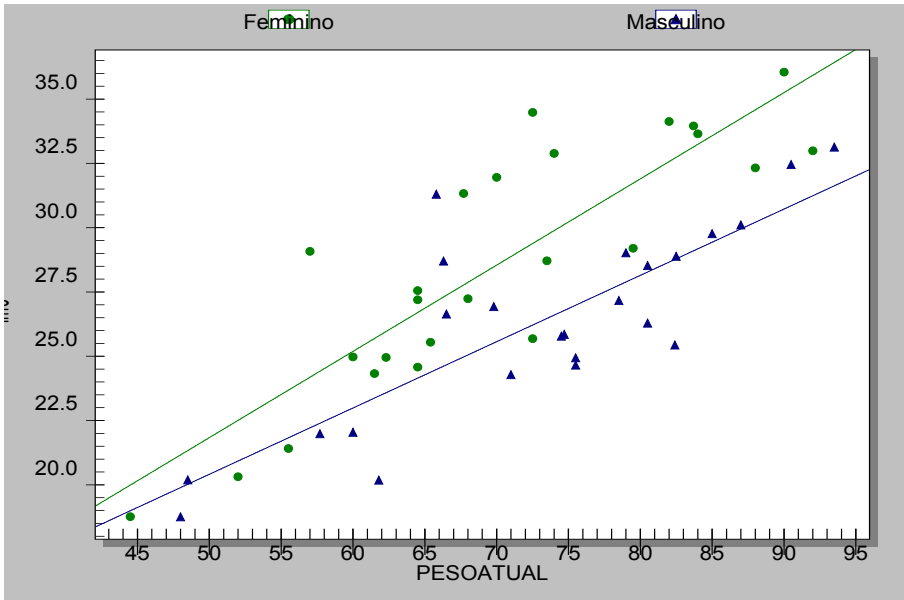


Figura 11.8. Distribuição do IMC medido para cada valor de *GENERO* em função do *PESOATUAL*.  
Fonte: Epi Info®, arquivo *Avaliacao.mdb*.

Observamos que os homens, em triângulo no gráfico, ficam, em sua maioria, inferiormente colocados. Mas as duas curvas não são, aparentemente, paralelas. O gráfico mostra que a linha das mulheres tem um ângulo maior que a dos homens, levando a acreditar que um aumento no peso das mulheres está relacionado a um aumento maior no IMC do que aconteceria nos homens. Como medir ou verificar se isso é significativo ou não? Usamos um artifício chamado de interação. Nossa fórmula ficará:  $IMC = peso\ genero\ peso * genero$ , onde teremos mais um coeficiente para medir a diferença no efeito do peso sobre o IMC nos diferentes gêneros. No comando da Regressão Linear, para criar uma interação, devemos selecionar as variáveis desejadas colocando-as na lista de variáveis do modelo. Depois devemos marcá-las clicando em cada uma delas conforme a figura 11.9 (1). A barra que fica entre a janela de seleção das variáveis e as variáveis selecionadas muda de nome, para *Make Interaction* (2); clicando nessa barra a fórmula da interação passa para a janela de interação, ficando a janela final semelhante à observada na figura 11.10.



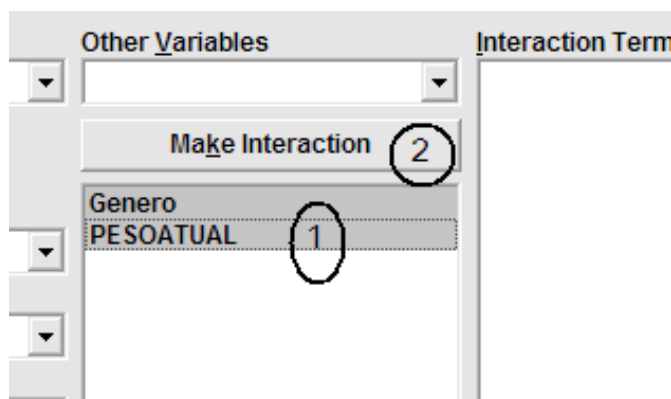


Figura 11.9. Detalhe do comando *Linear Regression* mostrando como criar uma interação entre duas variáveis.  
Fonte: Epi Info®.

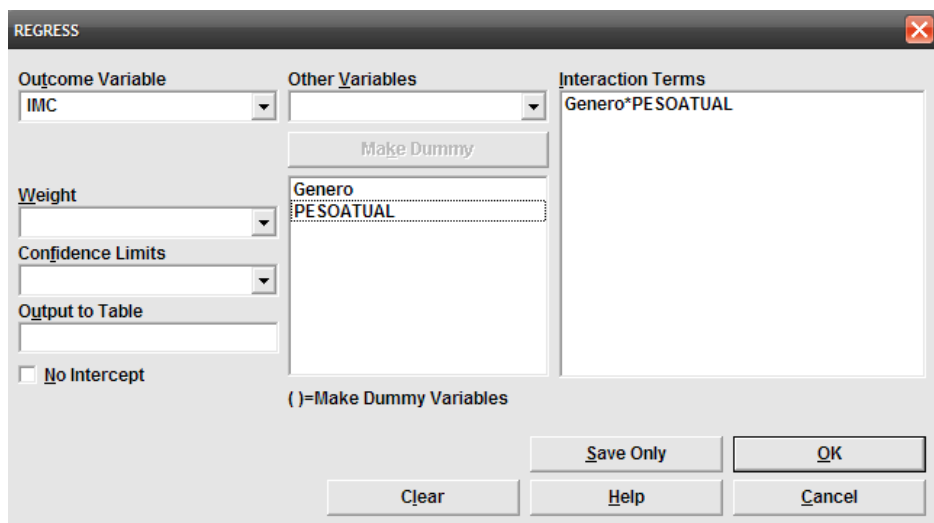


Figura 11.10. Janela do comando *Linear Regression* mostrando os comandos para a criação de uma regressão linear com interação entre GÊNERO e PESOATUAL.  
Fonte: Epi Info®.

### Vejamos nosso resultado.

Variable	Coefficient	Std Error	F-test	P-Value
Pesoatual	0.319	0.038	69.4905	0.000000
genero (Masculino/Feminino)	2.310	4.195	0.3034	0.583791
pesoatual * genero (Masculino/Feminino)	-0.081	0.058	1.9675	0.165780
CONSTANT	6.632	2.766	5.7477	0.019587
<b>Correlation Coefficient: <math>r^2=0.67</math></b>				

A equação fica assim:  $IMC = 6,632 + pesoatual * 0,319 + genero * 2,310 + pesoatual * genero * -0,081$ , onde *genero* será igual a 1 para os homens e Zero para mulheres. Notamos primeiro que o coeficiente *genero* mudou para positivo, mostrando que a linha dos homens inicia maior, mas não é significativamente maior que zero, pois *p* é maior que 0,05; homens e mulheres iniciam em pontos não significativamente diferentes. O coeficiente da interação *peso\*genero* também não foi significativo. Continuamos tendo duas fórmulas para o cálculo do *IMC*, uma para cada valor de gênero:

- Para as mulheres será simplesmente:  $IMC = 6,632 + pesoatual * 0,319$ .
- Para os homens passa para  $IMC = (6,632 + 2,310) + Pesoatual (0,319 - 0,081)$ .

Os coeficientes *CONSTANT* e *pesoatual* se referem então às mulheres e os coeficientes *genero* e *pesoatual\*genero* às diferenças entre os homens e as mulheres. Entretanto, no nosso exemplo, a interação não foi significativa. Na nossa interpretação final, o gráfico demonstrou linhas não paralelas, mas a diferença dos ângulos entre homens e mulheres não foi significativa. Quando uma interação não é significativa, ela não necessita aparecer nos resultados, então podemos, sim, somente citar que ela não foi significativa.

Notamos também que, na equação com interação, o coeficiente *genero* não foi significativo, pois a interação “roubou” a significância. Por isto, no nosso caso específico, a interação deve ser removida do nosso modelo final. Exceto se quisermos mostrar essa diferença e divulgarmos que não é significativa. Novamente, um resultado não significativo pode ser importante e não significa que não exista diferença no comportamento do peso em relação *IMC* entre homens e mulheres, pois estamos tratando de uma amostragem. Quem sabe numa amostra diferente ou maior o coeficiente *Peso\* genero* seja significativo.

## Exercício:

Utilizando o projeto Apoe\*, tabela Grupo1, calcular a regressão linear de: IMC, COLESTEROL, PAD, PAS e TEMPO como variáveis dependentes e as variáveis IDADE e APOE4 como variáveis independentes; utilizando a interação IDADE\*APOE4, e descrever o resultado.

1.  $IMC = IDADE \text{ APOE4 } IDADE * APOE4.$
2.  $COLESTEROL = IDADE \text{ APOE4 } IDADE * APOE4.$
3.  $PAS = IDADE \text{ APOE4 } IDADE * APOE4.$
4.  $PAD = IDADE \text{ APOE4 } IDADE * APOE4.$
5.  $Tempo = Idade \text{ Apoe4 } Idade * Apoe4.$

\*Poderá usar o banco de dados homônimo que se encontra no folder *Dados* do site [www.pucrs.br/igg/epiinfo](http://www.pucrs.br/igg/epiinfo).

## Capítulo 12

# ENTENDENDO A REGRESSÃO LOGÍSTICA E AS MEDIDAS DE RISCO

### Compreendendo as medidas de risco

As medidas de risco são utilizadas para avaliar a relação existente entre um fator de risco e um evento. Existem três formas de se avaliar esta relação:

- 1.** Escolher duas amostras, uma com (caso) e outra sem o evento (controle), e estudar a prevalência do fator.
- 2.** Escolher uma amostra aleatória de uma população e observar a prevalência do evento e dos fatores.
- 3.** Estudar um grupo de pessoas com ou sem o fator e observar quantos vão desenvolver o evento “desejado”.

O primeiro método, chamado caso-controle, é particularmente útil quando temos um evento (doença) raro, pois teríamos que ter uma amostra da população muito grande para podermos pegar ao acaso um número expressivo de casos. Por exemplo, se a prevalência de determinada doença é um caso para mil habitantes, teremos que observar pelo menos 5 mil pessoas escolhidas aleatoriamente para termos alguma chance de conseguir cinco casos. Observa-se que estudos caso-controle são mais “rentáveis” que os levantamentos populacionais ou transversais, que são o segundo método de avaliação do risco.

O estudo de coorte, o terceiro método, observa a incidência de novos casos durante um período de tempo. O delineamento não somente é baseado na hipótese que as pessoas com fator de risco vão ter maior incidência do evento, mas também avalia a função do tempo. A outra hipótese dos estudos de segmento é que os indivíduos com fator de risco vão ter maior probabilidade de desenvolver determinado evento antes do que o grupo livre do fator de risco. É o caso da mortalidade, pois todas as pessoas vão morrer, mas as que tiverem fatores de risco a probabilidade de morrer mais cedo será maior.

Os instrumentos de avaliação do risco podem concluir que determinado fator não tem relação com o evento. Isto ocorre quando a proporção de eventos é igual nos dos grupos, com e sem fator. Assim, é possível concluir

que possui relação de risco, quando se prova que a presença do fator eleva o número de casos, ou, ainda, que o fator pode ser “protetor”, quando se observa, proporcionalmente, menos eventos no grupo com fator.

Como estamos trabalhando com proporção, seria o primeiro caso se tivermos 50% de casos com fator e 50% de casos sem fator. Essa proporção de casos e não casos seria  $50/50 = 1$ . Se tivermos mais casos com fator, teremos um número maior que 1 e, inversamente, se tivermos menos casos com fator, teremos um número menor que 1. Neste último caso, onde menos gente teria o evento, nos levaria a crer que o fator é protetor. Vimos que o número 1 é o “divisor de águas”: mais que 1 significaria fator de risco, menos que 1 fator protetor.

Toda a amostra nos induz a um erro, pois tentamos, através de um número reduzido, estimar a verdade. Para sabermos se o cálculo do risco é significativo necessitamos calcular o “intervalo de confiança de 95%”. Dentro deste intervalo estaria o risco “real”. Para que o risco seja significativo, ambos os valores, limite superior ou limite inferior do intervalo, devem ser ou menores ou maiores que 1. Em outras palavras, o intervalo não deve conter o valor 1.

Dependendo do desenho da nossa pesquisa poderemos usar dois tipos de medidas de risco: o Risco Relativo e a Razão de Chance.

## Calculando o Risco Relativo

Para sabermos se determinada exposição é fator de risco para a doença, calculamos o Risco Relativo. O cálculo do Risco Relativo é muito simples, pois dividimos a proporção de doentes expostos pela proporção de doentes não expostos.

**Exemplo:** Considerando uma população de 100.000 habitantes e a distribuição da mesma em uma tabela 2x2 sobre o fato de ter ou não uma determinada doença e/ou um determinado fator de risco:

	Doença		Total
	Sim	Não	
Exposto	a= 300	b=9.700	a+b= 10.000
Não exposto	c=700	d=89.300	c+d= 90.000
Total	a+c=1.000	b+d=99.000	a+b+c+d=100.000

$$\begin{aligned} \text{incidência em expostos} &= a/(a+b) \Rightarrow 300/10.000=0,03 \\ \text{incidência em não expostos} &= c/(c+d) \Rightarrow 700/90.000=0,00784 \\ \text{RR} &= 0,03/0,00784 = 3,939 \end{aligned}$$

Significa: o risco de uma pessoa exposta ter a doença é 3,94 vezes o risco da não exposta, ou 2,94 vezes maior. No caso temos um estudo transversal.

O Risco Relativo, por ser um simples cálculo de chances maiores em determinado grupo comparado com outro, pode nos levar a falsas interpretações quando um terceiro fator desconhecido esteja afetando ambos. Algum pesquisador desavisado poderá associar câncer de pele e catarata. No entanto, ambos estão associados à exposição ao sol sem proteção. A retirada de um fator de risco deve afetar o aparecimento do outro, o que não vai ocorrer com a cirurgia de catarata e o câncer de pele. O uso de creme protetor solar para evitar o câncer de pele não vai afetar o aparecimento de catarata.

### Calculando a Razão de Chance

Outro cálculo utilizado é a Razão de Chance (Odds Ratio), que seria o cálculo de o quanto a chance de ter a doença com exposição seria maior que o sem exposição ao fator. Para isso usamos a fórmula: número de expostos doentes x não expostos não doentes, divididos pelos expostos não doentes e não expostos doentes.

**No exemplo:** Razão de Chance =  $axd/bc = 300 \times 89.300 / 700 \times 9.700 = 3.946$

### Características do Risco Relativo (RR) e da Razão de Chance (RC)

1. Quanto mais rara for a doença, mais RR=RC.
2. Se RR for igual a 1, então RC sempre será igual a 1.
3. Se RR for maior que 1, então RC sempre será maior que 1.
4. Igualmente, se RR for menor que 1, então RC sempre será menor que 1.
5. Se trocarmos o fator de risco pelo efeito, a razão de chance não muda, mas o valor do risco relativo altera.

## Calculando o Risco usando o Epi Info®

A maneira mais fácil de calcular o risco quando temos já a tabela com os valores dos casos e eventos é utilizar o programa *tables 2xn* do StatCalc. Como vimos no capítulo 9, o StatCalc fica no menu *utilities*. Vamos calcular o exemplo anterior criando uma tabela 2x2.

Escolhemos assim a opção *tables 2x2* e surge a tabela para colocarmos os dados sempre apertando a tecla *Enter*. Na sequência colocamos 300, 9700, 700, 89300. Clicando *Enter* duas vezes teremos o resultado na figura 12.1.

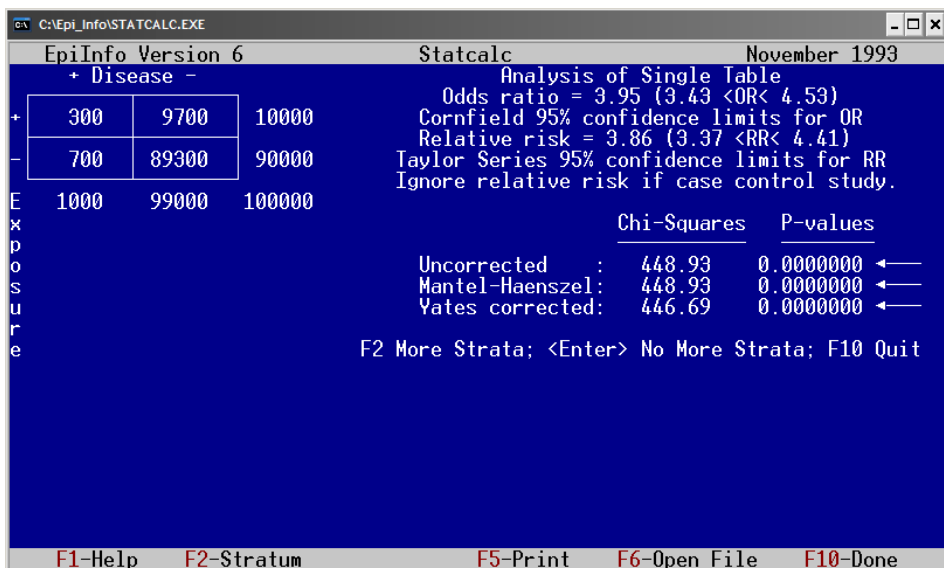


Figura 12.1. Resultado da análise da tabela do exemplo 1 pelo *StatCalc*.

Fonte: Epi Info®, dados fictícios.

Vimos que os valores são muito grandes, pois representam um estudo transversal populacional; existem também mais sadios que doentes, pois a prevalência da doença é de 1/100. Como vimos anteriormente, estudos populacionais são muito difíceis, principalmente quando trabalhamos com doenças raras. Assim nos valemos de estudos de caso-controle, onde sabemos de antemão quem é doente e sadio. Geralmente, a proporção de doentes será maior que o da população total e podemos escolher a mesma quantidade de casos e controles ou um pouco mais controles que casos. No estudo de caso-controle, devemos escolher os casos e controles sem controlarmos, nesse momento, quem é exposto ou não. Assim, se escolhermos uma amostra menor, a proporção de expostos

e não expostos deverá ser representativa, ou seja, parecida com o que acontece na população.

Vamos ver qual das medidas de risco seria mais adequada para o cálculo do risco em estudo de caso-controle. No exemplo acima, teríamos um estudo de caso-controle onde observamos apenas metade dos casos (500) e 1.000 controles (1/99), obedecendo às proporções do fator de risco teremos:

	Doentes	Controles	Total
Exposto	300/2=150	9.700/99=98	248
Não exposto	700/2=350	89.300/99=902	1.252
Total	500	1.000	1.500

Usando o *StatCalc* para calcular os riscos teremos o resultado na figura 12.2.

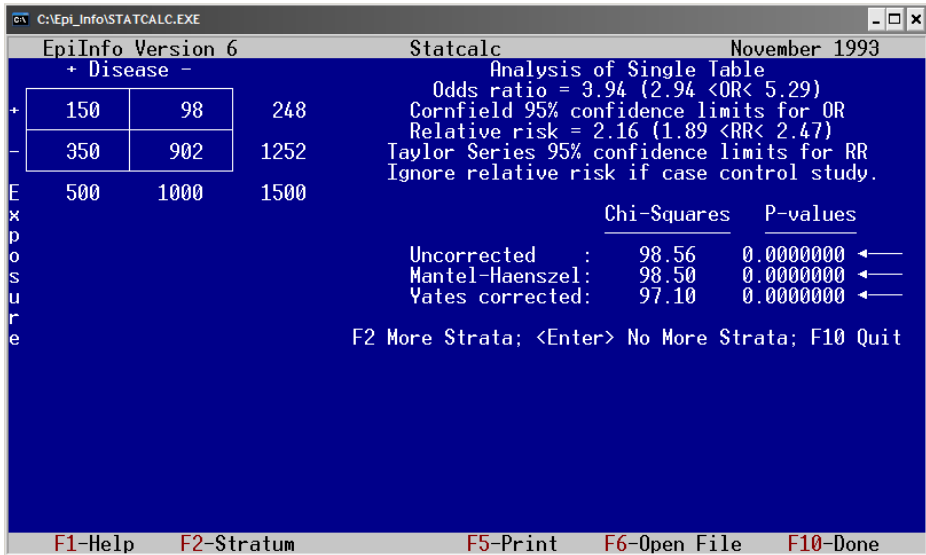


Figura 12.2. Resultado da análise da tabela do exemplo 2 pelo *StatCalc*.

Fonte: Epi Info®, dados fictícios.

Observamos que, como a proporção do fator de risco em ambos os grupos, doentes e controles, é o mesmo do estudo transversal, o resultado do cálculo do risco deveria ser o mesmo. Entretanto, o Risco Relativo calculado no caso-controle é bem diferente do calculado no estudo transversal. Já na Razão de Chance, os resultados de ambos os estudos



são muito semelhantes, comprovando que a Razão de Chance pode ser utilizada nos dois desenhos.

Se trocarmos a posição do fator, pelo evento, teremos o resultado conforme a figura 12.3.

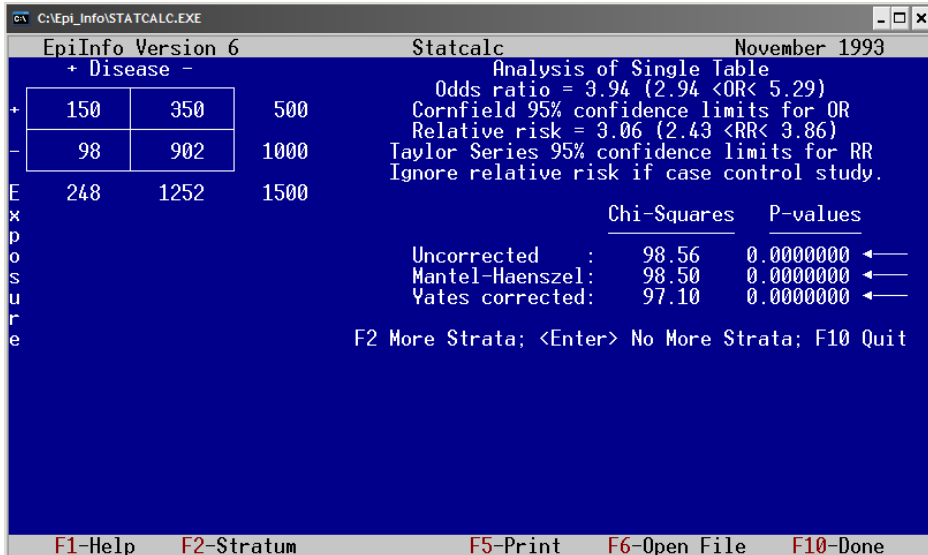


Figura 12.3. Resultado da análise da transposição da tabela do exemplo 2 pelo *StatCalc*.

Fonte: Epi Info®, dados fictícios.

Novamente, o resultado do cálculo da Razão de Chance mantém-se e o resultado do Risco Relativo se altera.

## Usando a Regressão Logística

Utilizamos a regressão logística para calcular a Razão de Chance de uma pesquisa cujo desenho seja caso-controle com mais de um fator de risco. Como vimos, a regressão tenta descrever a relação de duas ou mais variáveis em uma fórmula. No caso do estudo do risco teremos nossa variável dependente categórica com dois únicos valores possíveis: ter ou não ter o evento. Esse evento pode ser um evento adverso ou favorável: morte ou ganhar na loteria, infelizmente a Razão de Chance do primeiro é maior que a do segundo. Quando montamos um estudo sobre o risco de um evento, devemos, inicialmente, caracterizar ou descrever claramente que critérios serão utilizados para caracterizar o evento. A fórmula da regressão logística é:

$$RC = e^{bx}$$

onde  $e$  é o logaritmo natural e  $b$  o coeficiente do fator ou variável  $x$ . Vamos ver como ele é calculado no Epi Info®.

No nosso banco de dados (*Apo*e), temos, como evento, o episódio coronariano, caracterizado como angina, infarto ou alterações no eletrocardiograma, que sugerem alterações isquêmicas das artérias coronarianas. Sabemos que essas alterações estão relacionadas à Aterosclerose das artérias e que possui muitos fatores de risco e muitos deles estão relacionados entre si. Por exemplo, sabemos que a idade é um fator de risco, quanto maior a idade maior a chance de desenvolver um evento coronariano.

Para calcular esse risco podemos utilizar a Regressão Logística. Neste único caso, o Epi Info® EXIGE que a variável resposta ou dependente (*Outcome*) seja *Sim/Não* (*Yes/No*). Para sabermos se a variável *EVENTO*, do banco de dados *Apo*e e que possui as informações sobre o nosso caso, é uma variável com formato *Yes/No*, não basta olharmos no *list*. Fazendo um *list* dessa variável, notamos que ela seus valores são todos 1 ou 0 (figura 12.4).

The screenshot shows a window titled "Analysis Output" with a table. The first column is labeled "evento" and contains the following values: 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0.

evento
0
0
1
0
0
0
0
0
0
0
0
1
0
0
0

Figura 12.4. Detalhe do resultado do comando *List* mostrando os valores originais da variável *EVENTO*.

Fonte: Epi Info®, arquivo *Apo*e.mdb.

Entretanto, quando usamos o comando *Display*, que nos mostra as características das variáveis do projeto em uso, observamos, na figura 12.5, que a variável é, na realidade, uma variável tipo texto, provavelmente porque o arquivo havia sido importado do *Excel*.

The screenshot shows a web browser window titled 'C:\Epi\_Info\EpiAula\OUT9.htm' with a toolbar containing 'Previous', 'Next', 'Last', 'History', 'Open', 'Bookmark', 'Print', and 'Maximize'. Below the toolbar is the heading 'DISPLAY DBVARIABLES'. A table lists database variables with columns: Variable, Table, Field Type, Format/Value, Special Info, and Prompt. The 'evento' row is circled in red.

Variable	Table	Field Type	Format/Value	Special Info	Prompt
ano	Apoe4	NUMBER	#####		ano
apoe4	Apoe4	NUMBER	###		apoe4
cigarro	Apoe4	TEXTBOX			cigarro
colesterol	Apoe4	NUMBER	#####		colesterol
Dislipidem	Apoe4	TEXTBOX			Dislipidem
evento	Apoe4	NUMBER	###		evento
Hipertensa	Apoe4	YES/NO			Hipertensa
idade	Apoe4	NUMBER	#####		idade
id	Apoe4	NUMBER	###		id
IMC	Apoe4	NUMBER	#####		IMC
obesidade	Apoe4	TEXTBOX			obesidade

Figura 12.5. Resultado do comando *Display* destacando as características originais da variável *EVENTO*.  
Fonte: Epi Info®, arquivo *Apoe.mdb*.

Devemos, então, transformar a variável *EVENTO* em variável tipo *Yes/No*. O que parece fácil exige alguns cuidados, pois não basta usar o *Recode* da variável com os valores “Yes” e “No”. Existem as siglas especiais que o Epi Info® utiliza: (-) para não e (+) para sim. Teremos nosso comando conforme a figura 12.6.

The screenshot shows the 'RECODE' dialog box. The 'From' dropdown is set to 'evento' and the 'To' dropdown is also set to 'evento'. Below these are two columns: 'Value (blank = other)' and 'To Value (if any)'. The 'Recoded Value' column shows the result of the recoding.

Value (blank = other)	To Value (if any)	Recoded Value
0		(-)
1		(+)

At the bottom of the dialog are buttons for 'Fill Ranges', 'Save Only', 'OK', 'Clear', 'Help', and 'Cancel'.

Figura 12.6. Janela do comando *Recode* mostrando os códigos para a transformação da variável *EVENTO* em variável tipo *Yes/No*.  
Fonte: Epi Info®.

Agora poderemos usar a Regressão Logística. No menu *Advanced Statistics*, escolhemos *Logistic Regression*. Colocando como *Outcome Variable*, *EVENTO* e como *Other Variables*, *IDADE*, conforme a figura 12.7.

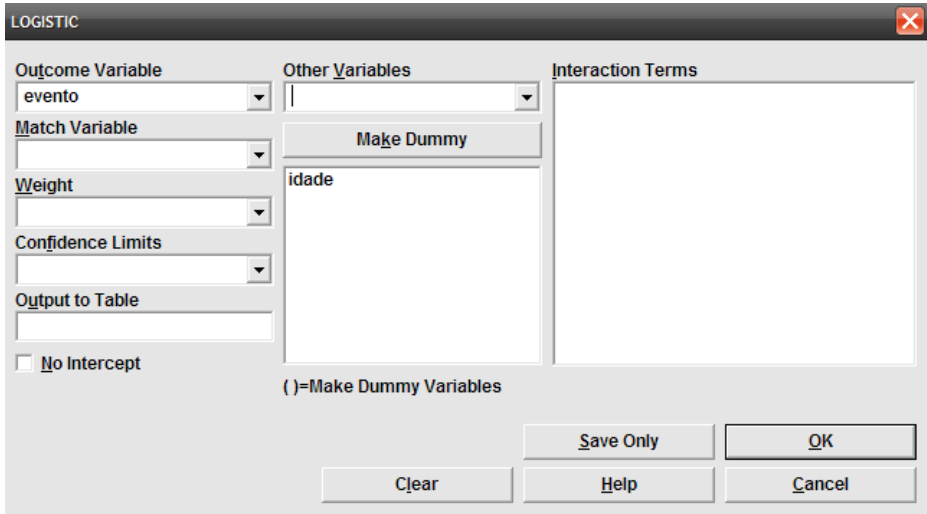


Figura 12.7. Janela do comando *Logistic Regression*.  
Fonte: Epi Info®.

Como resultado, teremos a figura 12.8.

Unconditional Logistic Regression							
Term	Odds Ratio	95% C.I.	Coefficient	S.E.	Z-Statistic	P-Value	
IDADE	1.0396	1.0254 1.0541	0.0389	0.0070	5.5153	0.0000	
CONSTANT	*	*	-3.9567	0.4274	-9.2576	0.0000	
Convergence:		Converged					
Iterations:		5					
Final -2*Log-Likelihood:		545.7608					
Test	Statistic	D.F.	P-Value				
Score	32.5189	1	0.0000				
Likelihood Ratio	33.2749	1	0.0000				

Figura 12.8. Janela do resultado do comando *Logistic Regression*.  
Fonte: Epi Info®, arquivo *Apo.e.mdb*.

O resultado fornece todas as informações de que precisamos. O coeficiente para idade é 0,0389 que, aplicando a fórmula  $RC=e^{0,0389}$ , temos justamente o valor dado para a razão de chance, ou seja, 1,0396. Na interpretação, uma pessoa com um ano a mais tem um risco de 1,0389 vezes o risco de ter o evento, ou 3,9% a mais de ter o evento. O intervalo de confiança de 95% vai de 1,0254 a 1,0541, como não inclui o 1, então a razão de chance é significativa.

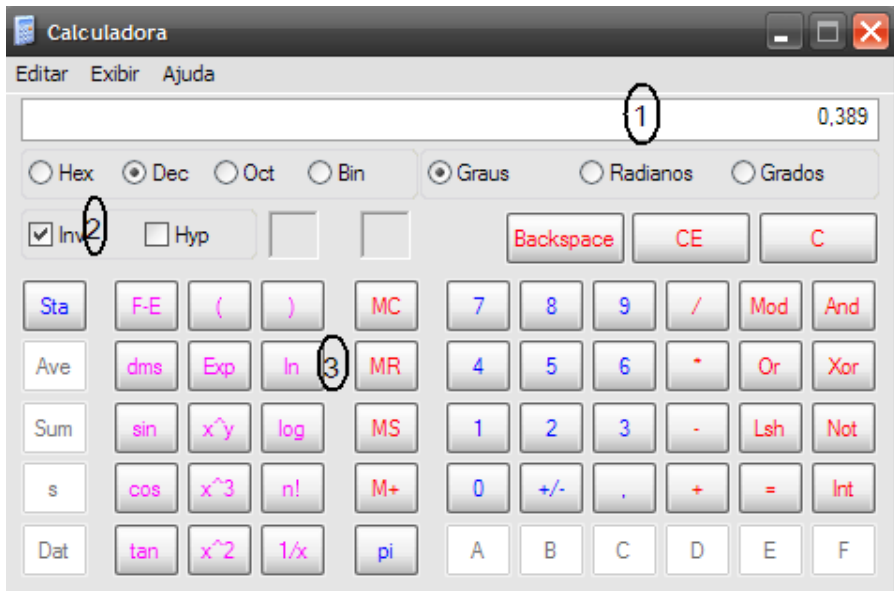


Figura 12.9. Janela da calculadora do *Windows* destacado os comandos para o cálculo do logaritmo natural de um número.

Fonte: Windows 7.

*Importante:* uma pessoa 10 anos mais velha não terá uma razão de chance de 10,396 ( $1,0396 \times 10$ ) mas de  $e^{(10 \times 0,0389)}$ , ou seja 1,4755. Para calcular isso podemos usar a calculadora do *Windows*.

No botão *iniciar* do *Windows* temos a opção *programas, acessórios, calculadora (calculator)*. No menu *Exibir (View)* da calculadora (figura 12.9) selecionamos *Científica*. Primeiro temos que colocar o valor do coeficiente já multiplicado por 10: 0,389 (1). Selecionamos a opção *Inv* (invertido) (2) depois clicamos em *ln* (logaritmo natural) (3). E, assim, temos o resultado (1,4755).

Se quisermos que o Epi Info® descreva automaticamente essa Razão de Chance, com intervalos de 10 anos entre as idades, basta criar uma variável nova atribuindo a idade dividida por 10. Igualmente, outras variáveis podem ser manipuladas dessa forma como colesterol, pressão arterial etc.

Vamos rapidamente criar *idade10*, *colesterol10*, *PAS10* e *PAD10*: antes temos que definir essas variáveis, usando o comando *Define*. Podemos usar o diálogo da esquerda ou, simplesmente, na janela do *Command Program* digitar os comandos:

- Define Idade10;
- Define colesterol10;
- Define PAS10,
- Define PAD10.

A seguir atribuiremos os valores correspondentes às variáveis criadas, usando o comando *Assign*:

- Assign Idade10=idade/10;
- Assign colesterol10=colesterol/10;
- Assign PAS10=PAS/10 e, finalmente,
- Assign PAD10=PAD/10.

## Usando Regressão Logística com múltiplas variáveis

A Regressão Logística, diferentemente da Razão de Chance calculada para uma tabela 2x2 vista acima, pode ser usada para avaliar o impacto de duas ou mais variáveis independentes sobre a variável resposta. Algumas variáveis independentes podem influenciar outras no advento que se está estudando. Assim é sempre interessante a comparação entre os resultados da análise da regressão simples onde cada uma das variáveis independentes é analisada separadamente e a análise múltipla onde todas as variáveis são incluídas. Uma utilização muito feliz da regressão logística multivariada foi usada por Bós e Bós em um artigo recente, publicado na *Revista de Saúde Pública* (vol. 38 nº 1, 2004) no qual se observou o efeito de algumas variáveis independentes sobre local de atendimento médico. Para tanto foi analisado o banco de dados da Pesquisa “Idosos do Rio Grande do Sul, suas condições de vida”, realizada pelo Conselho Estadual do Idoso com a colaboração de várias Universidades do Estado. Esse estudo entrevistou 7.280 idosos, não institucionalizados, em 1995. A análise buscou entender que fatores proporcionariam o idoso a utilizar a rede privada de atendimento e a rede pública. No questionário original foi perguntado ao idoso qual o tipo de atendimento conseguido na última vez em que buscou

um atendimento médico. A variável dependente então foi definida como Yes = rede privada, No = rede pública. O principal objetivo da regressão foi estimar como cada variável dependente (chamada no artigo de variável explicativa) afeta a chance (pelo cálculo da Razão de Chance) do idoso ter utilizado a rede privada de atendimento. As variáveis independentes foram: gênero, idade (em anos), escolaridade (número de anos de estudo), renda individual (em salários mínimos), renda familiar (também em salários mínimos), tamanho da família, participação na renda (percentagem) e autoavaliação da saúde.

O artigo descreve muito apropriadamente as características da análise, por isso reproduziremos aqui alguns trechos. Recomendamos a leitura do texto original.

Três modelos diferentes foram apresentados: o modelo completo (multivariado), que inclui todas as variáveis explicativas; o modelo sem renda, que exclui as variáveis relacionadas à renda: individual, familiar e participação na renda da família; e o modelo com variáveis individuais, onde cada fator explicativo foi analisado individualmente. Esse último modelo é simplesmente uma sequência de análises bivariadas, em que a relação entre cada variável explicativa e a variável dependente é analisada isoladamente. A comparação entre os resultados dos três modelos permite uma análise mais detalhada e rigorosa de como cada variável independente influencia a escolha do local de atendimento. Por exemplo, se o coeficiente para uma determinada variável apresenta valores distintos entre o modelo completo e o sem renda, o impacto desta variável é influenciado pelo nível de renda. Se o valor for maior no modelo sem renda do que no completo, a influência da variável é explicada, parcialmente, pela renda. No caso extremo em que o valor é zero no modelo completo, mas significativo no sem renda, toda a influência dessa variável seria explicada pela renda. A interpretação é invertida, se o valor do coeficiente for menor no modelo sem renda do que no completo. Usando o caso extremo de novo, se o valor é significativo no modelo completo, mas zero no modelo sem renda, toda a influência da renda seria explicada por essa variável.

Na interpretação dos resultados, as razões de chance com valores acima de 1,0 indicam que a presença ou o maior valor do fator explicativo aumentam

as chances do idoso ter utilizado a rede privada de atendimento. Valores abaixo de 1,0 indicam que a presença ou o maior valor do fator explicativo diminuem essas chances. Como a alternativa fundamental à rede privada é a rede pública de atendimento, os resultados podem ter a direção invertida e aplicados à rede pública.

Apresentamos na tabela 12.1, a seguir, o resultado mais importante do artigo.

Tabela 12.1. Razões de Chance, dos modelos de regressão logística completo, sem renda e com variáveis individuais na predição da escolha da rede privada de atendimento. \* $p < 0,05$ ; \*\* $p < 0,01$ .

Variável	Modelo completo	Modelo sem renda	Variáveis Individuais
<b>Gênero: feminino</b>	1,1571*	1,1284*	1,0749
<b>Idade</b>	1,0266**	1,0267**	1,0213**
<b>Escolaridade</b>	1,0429**	1,1360**	1,1369**
<b>Renda individual</b>	1,0707*	-----	1,2132**
<b>Renda familiar</b>	1,2019**	-----	1,2411**
<b>Tamanho da família</b>	0,8520**	0,9720	0,9499**
<b>Participação na renda</b>	0,9968	-----	0,9990
<b>Saúde</b>	1,0266	1,1147**	1,2117**

Fonte: *Revista de Saúde Pública*, vol. 38, n. 1, 2004, reproduzido com permissão.

Estudando os resultados do modelo completo, observa-se que as variáveis seguintes foram significativas no uso da rede privada de atendimento:

- 1.** idosas têm 15% mais chances que os idosos;
- 2.** com o aumento na idade em um ano, as chances aumentam em 2,7%;
- 3.** com o aumento na escolaridade em um ano, as chances aumentam em 4%;
- 4.** com o aumento na renda individual em um salário mínimo, as chances aumentam em 7%;



**5.** com o aumento na renda familiar em um salário mínimo, as chances aumentam em 20%;

**6.** com uma pessoa a mais na família, as chances diminuem em 15%.

Por outro lado, a participação na renda familiar e a avaliação pessoal sobre a saúde não foram significativas.

Comparando os resultados do modelo completo com o modelo sem renda, observa-se que, em quase todas as variáveis, o valor das razões de chance é alterado – em maior ou menor grau – indicando que a renda afeta como estas variáveis influenciam a escolha do local de atendimento. Destaca-se que, no modelo sem renda, a avaliação pessoal da saúde se tornou significativa, enquanto o tamanho da família teve uma redução no seu nível de significância.

Resultados semelhantes são observados na comparação entre os outros modelos e o uso individual de cada variável explicativa.

Os resultados do trabalho mostram que, à medida que a sua renda aumenta, o idoso tende a escolher a rede privada de atendimento. A renda familiar teve um impacto muito mais expressivo de que a renda individual do idoso. Com um aumento na renda familiar em um salário mínimo, as chances de o idoso utilizar a rede privada aumentam em 20%. Com um aumento idêntico na renda individual, as chances aumentam em apenas 7%. Por outro lado, o idoso retém um certo grau de independência sobre o uso de seus recursos financeiros individuais, já que a razão de chance para a renda individual foi significativa e maior que um.

**Exercício:**

Utilizando o projeto Apoe\* tabela Grupo1, calcular a Razão de Chance da presença de evento usando os seguintes modelos:

1. Evento = Idade10.
2. Evento = Colesterol10.
3. Evento = PAS10.
4. Evento = PAD10.
5. Evento = Idade10 + Colesterol10 + PAS10 + PAD10.

Comente sobre a Razão de Chance de cada variável no modelo simples e no modelo complexo.

\*Poderá usar o banco de dados homônimo que se encontra no folder *Dados* do site [www.pucrs.br/igg/epiinfo](http://www.pucrs.br/igg/epiinfo).

## Capítulo 13

# USANDO CURVAS DE KAPLAN-MEIER E REGRESSÃO DE DANO DE COX NA ANÁLISE DE SOBREVIDA

### Introdução à análise de sobrevida

A análise de sobrevida pode ser realizada quando temos um acompanhamento de um grupo. Dentro do desenho estatístico, esse grupo é chamado de grupo coorte, por isso é também chamado de Estudo de Coorte. Quando desejamos estudar um evento usando um Estudo de Coorte, devemos ter um momento inicial onde todos os seus integrantes estejam livres do evento ou que tenham a mesma característica. Podemos estudar o aparecimento de doença aguda coronariana ou o reaparecimento da mesma. No primeiro caso devemos excluir pessoas que tenham já de antemão a doença coronariana. No segundo caso, podemos estudar pessoas que tiveram Infarto Agudo do Miocárdio e estudar quem irá desenvolver um segundo. Faz mais sentido, nesse caso, termos todos os integrantes desse acompanhamento com história confirmada de infarto prévio. De qualquer forma, temos um momento no qual se inicia o acompanhamento. Nesse momento, vários dados podem e devem ser colhidos. Esses dados devem ser fatores que potencialmente vão afetar o aparecimento do evento, sejam fatores de risco ou de proteção. Temos assim o ponto inicial. Não necessariamente todas as pessoas que fazem parte do grupo de coorte devem iniciar no mesmo momento, mas todas devem ter o mesmo tipo de avaliação, seguindo os mesmos critérios de diagnóstico do evento a ser estudado.

### Entendendo as curvas de Kaplan-Meier

São curvas usadas para estimar a probabilidade de sobrevida a um evento por unidade de tempo. São usadas para desenvolver curvas de sobrevida quando nem todos os tempos de sobrevida são conhecidos. Explicando, temos um grupo de coorte que será acompanhado para o evento morte. Nem todas as pessoas morrerão durante o tempo de observação, pois só podemos acompanhar essas pessoas durante certo tempo. Igualmente, podemos perder o seguimento de algumas pessoas

durante o acompanhamento, por desistência, mudança de endereço ou outra forma de contato. Assim teremos três grupos: um grupo que vai morrer durante o acompanhamento, esse grupo saberemos assim o tempo de acompanhamento antes do falecimento; teremos um grupo que vai ainda estar vivo e outro grupo que vamos perder o acompanhamento. Nesses dois últimos grupos não sabemos, com certeza, quando eles falecerão, mas temos uma data mínima de sobrevivência, na qual, até essa data, todos estavam vivos, esses dois casos são chamados de grupos censurados. Na realidade, como, dependendo do desenho do nosso estudo, indivíduos podem iniciar o acompanhamento em diferentes períodos, teremos diferentes tempos de acompanhamento antes do evento. Se uma pessoa não desenvolveu um evento, mesmo que não seja óbito, até determinada idade ou período, não significa que ela não desenvolverá o evento, por isso é lógico chamar este caso de caso censurado. No cálculo da probabilidade de sobrevida, cada vez que ocorre um evento, no caso morte, a probabilidade de sobrevida é ajustada para os participantes ainda ativos naquele período. Entendemos por período o tempo de acompanhamento, independente de quando cada participante iniciou. Assim, se ao iniciarmos o nosso acompanhamento tínhamos 120 pessoas e durante a primeira morte se deu aos 2 meses de acompanhamento, mas nesse momento tínhamos somente 110 ativos, incluindo os mortos, temos uma probabilidade de sobrevida de  $108/110 = 0,98$  ou 98%. Assim a curva de Kaplan-Meier é formada por “degraus” formados cada vez que um evento ocorre. Essas curvas de sobrevida ou sobrevivência podem ser criadas para diferentes grupos, mas usando somente uma variável independente que defina os mesmos. Vamos ver no nosso exemplo usando o mesmo banco de dados da Regressão Logística.

### **Criando curvas de Kaplan-Meier no Epi Info®**

Para abrir a janela dos comandos da Análise de Kaplan-Meier clicamos no comando correspondente na seção *Advanced Analysis*. Usaremos o banco de dados *ApoE*<sup>2</sup>, tabela Grupo1 neste exemplo. Os comandos para fazer as curvas de Kaplan-Meier estão na figura 13.1.

---

<sup>2</sup> Poderá usar o banco de dados homônimo que se encontra no folder *Dados* do CD-ROM que acompanha este livro.

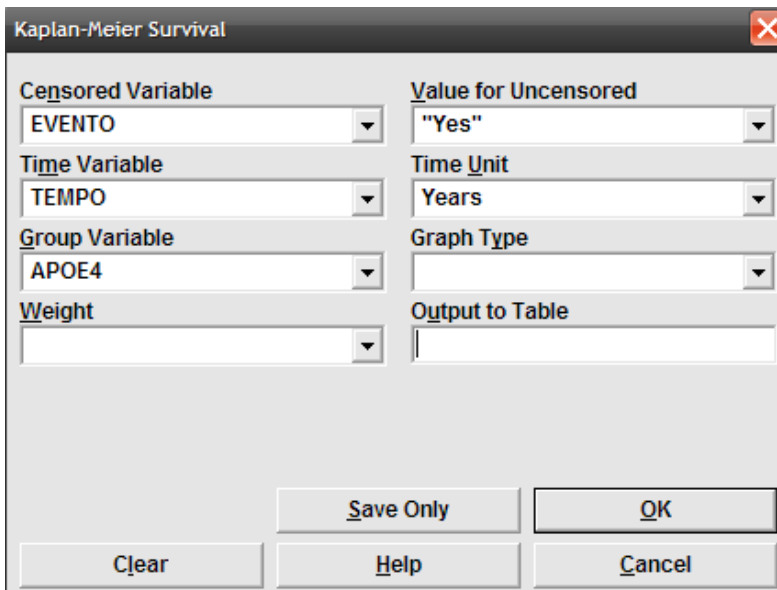


Figura 13.1. Janela do comando *Kaplan-Meier Survival*.

Fonte: Epi Info®.

A *Censored Variable* (1) é justamente a variável que contém informações sobre o evento, no nosso caso chama-se *EVENTO*. O valor para o caso não censurado (*Value for Uncensored* (2)) é o valor atribuído ao evento positivo, no caso “Yes”. A *Time Variable* (3) é a variável que informa o tempo decorrente entre o início do acompanhamento e o evento, para os casos de evento positivo ou do tempo decorrente entre o início do acompanhamento e o último período no qual o participante estava ativo. Essa variável é, no exercício, *TEMPO*. No nosso caso específico, como estamos interessados somente em evento cardíaco, a pessoa pode ter falecido de outra causa, desistido do trabalho, sumido ou ainda estar ativo. Todos esses casos serão enquadrados como censurados e terão a variável *EVENTO* igual a “No”. Para que o resultado da curva seja correto devemos informar a unidade de medida de tempo utilizada por essa variável no item: *Time Units* (4), no nosso caso estamos tratando de anos. Podemos criar uma curva única ou para dois ou mais grupos escolhendo a variável no item *Group Variable* (5), no nosso caso escolheremos a variável *APOE4*. Finalmente podemos escolher o tipo de curva que queremos no item *Graph Type* (6), na quase totalidade escolheremos *Survival Probability*, que automaticamente aparecerá mesmo se não a selecionamos. Clicando *OK* teremos nosso resultado:

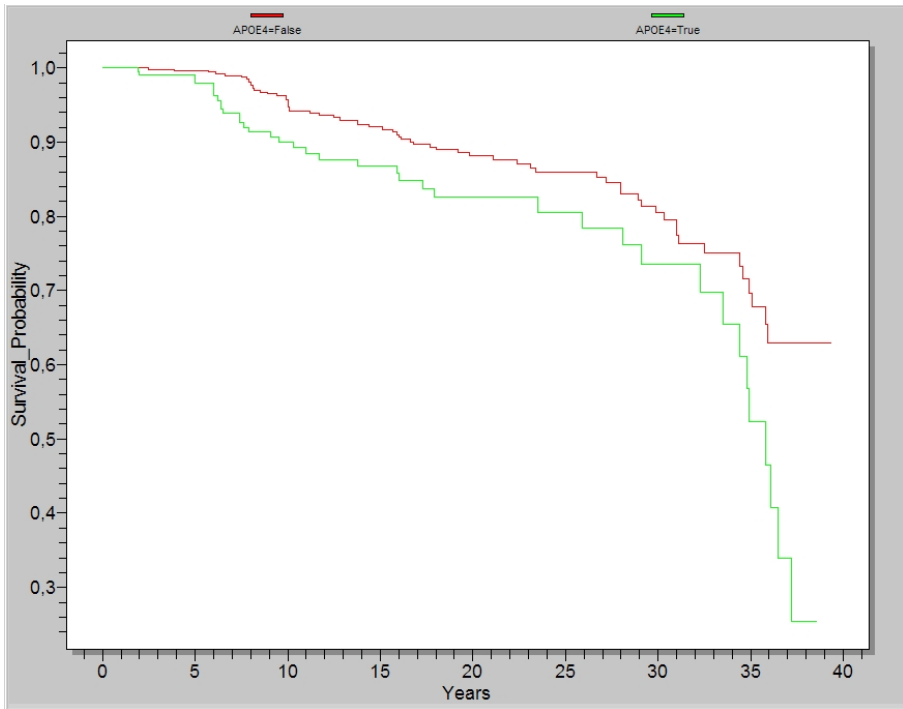


Figura 13.2. Gráfico do resultado do comando *Kaplan-Meier Survival*.

Fonte: Epi Info®, arquivo *Apo.e.mdb*.

Observamos que cada valor de *APOE4* apresenta-se em curva distinta com a curva *APOE4=No* ou grupo sem o alelo 4 da *Apo.e* apresentam uma curva superiormente colocada e que, ao contrário, a curva do grupo com *APOE4* é inferior. Isso significa que o grupo *APOE4* apresenta uma probabilidade de sobrevivência menor que o grupo sem o alelo 4. Podemos analisar essas curvas de duas maneiras: usando o parâmetro probabilidade de sobrevivência ou o tempo de sobrevivência. Normalmente se descreve o tempo de 50% de sobrevivência, no nosso caso não somente o grupo de *APOE4* tem valor para a sobrevivência de 50%, assim podemos usar uma sobrevivência de 70%. Na interpretação teremos uma sobrevivência de 70% do Grupo com *APOE4* de 32 anos e do grupo sem o alelo de 35 anos. Na outra interpretação teremos uma sobrevivência de 84% para os com *APOE4* e de 88% para os sem. O Epi Info® oferece dois tipos de análise estatística para as curvas Kaplan-Meier: o *Log-Rank*, que é um teste que compara os valores observados com os esperados, e o teste *Wilcoxon*, que é um teste não paramétrico e que assume uma mortalidade constante, e é útil para comparar dois tratamentos. No nosso caso ambos os testes foram

significativos, confirmando que as curvas são diferentes estatisticamente. Infelizmente o programa não nos deixa editar o gráfico, mas podemos salvá-lo clicando no botão da direita do *Mouse*.

## Usando a Regressão Proporcional de Dano de Cox

A Regressão Proporcional de Dano de Cox, ou chamada de Regressão de Cox, é usada para calcular uma estimativa do Risco Relativo, chamado de Razão de Dano (*Hazard Ratio*). Podemos calcular o risco de vários fatores e testar a influência dos mesmos no desenvolvimento de um evento. O comando, visto na figura 13.3, é muito semelhante ao das Curvas de *Kaplan-Meier*, mas possui a opção *Other Variables*. Igualmente, usaremos o banco de dados *ApoE*, tabela Grupo1.

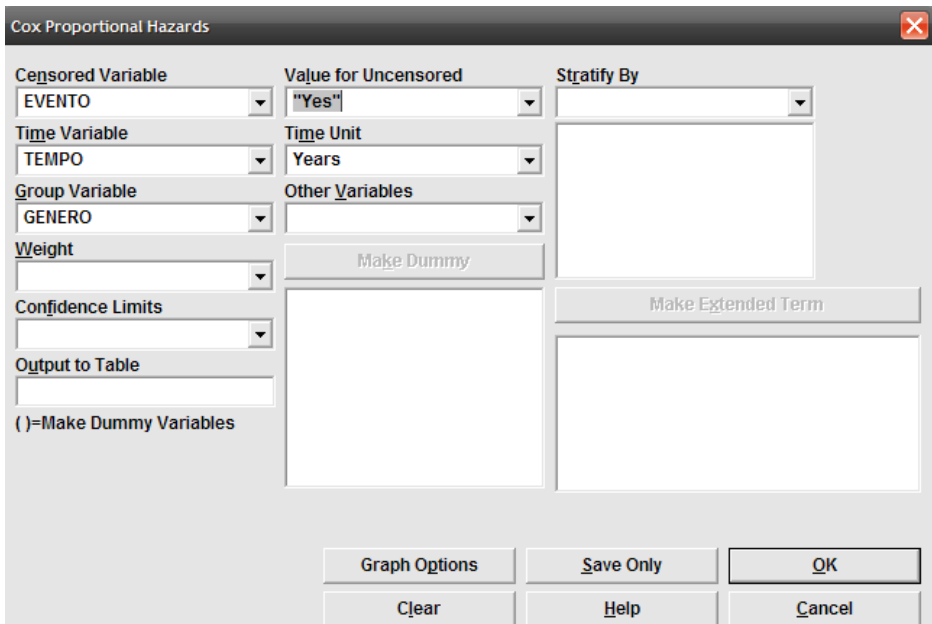


Figura 13.3. Janela do comando *Cox Proportional Hazards*.

Fonte: Epi Info®.

O resultado do modelo de Cox não só fornece a curva de probabilidade de sobrevivência como também a razão de dano que é interpretada da mesma forma que o Risco Relativo e, logicamente, fornece o Intervalo de Confiança para podermos interpretar os resultados. No resultado da regressão, usando somente o sexo como variável, teríamos:

```

COXPH tempo = sexo * evento ( 1 ) TIMEUNIT="Years" GRAPH=sexo

```

[Next Procedure](#)

**Cox Proportional Hazards**

Term	Hazard Ratio	95% C.I.	Coefficient	S. E.	Z-Statistic	P-Value
sexo (Masc/Fem)	0.8445	0.5185 1.3755	-0.169	0.2489	-0.6791	0.497

**Convergence:** Converged  
**Iterations:** 4  
**-2 \* Log-Likelihood:** 1169.6468

Test	Statistic	D.F.	P-Value
Score	0.4623	1	0.4965
Likelihood Ratio	0.4556	1	0.4997

Figura 13.4. Resultado da análise estatística realizada pelo comando *Cox Proportional Hazards*.  
 Fonte: Epi Info®, arquivo *Apo.e.mdb*.

Interpretando os dados observados veríamos que a razão de dano dos homens, em relação às mulheres, seria de 0,8445, não significativa, pois o intervalo de confiança varia de 0,5185 e 1,3755. A regressão de Dano de Cox é uma regressão que tem componentes paramétricos e não paramétricos. A razão de dano é um resultado paramétrico, pois estabelece um parâmetro a partir dos dados. O modelo como um todo é analisado não parametricamente pela Razão de Verossimilhança, em inglês: *Likelihood Ratio*. No caso, ambos critérios não são significativos. Por não ser significativo, o modelo do Epi Info® não forneceu a curva de sobrevivência.

## Entendendo a noção de proporcionalidade

Uma noção muito importante que devemos salientar é a da proporcionalidade. A regressão de Cox é chamada proporcional, pois assume que o risco cresce proporcionalmente conforme a variável risco, corroborando com a noção de ser um teste paramétrico. Assim, uma pessoa com nível de risco mais elevado terá um risco proporcionalmente maior. Muitas vezes há variáveis que não respeitam essa proporcionalidade, uma dessas variáveis é a idade sobre alguns eventos. A mortalidade não cresce proporcionalmente com a idade, mas,



sim, exponencialmente. Assim, para evitarmos problemas de interpretação devemos usar a variável idade categorizada. Essa variável também é um forte preditor de evento coronariano, tão forte que poderá afetar a interpretação.

Neste caso devemos estratificar nossa análise por faixa etária. Quanto maiores os grupos etários (10 ou 20 anos de intervalo), menor será o número de interpretações. Vamos criar uma variável chamada *DUODECADA*, colocando um valor para cada 20 anos conforme os comandos:

```
DEFINE Duodecada
RECODE idade TO Duodecada
60 - HIVALUE = 60
40 - 60 = 40
LOVALUE - 40 = 20
END
```

Adicionando a variável *DUODECADA*, *APOE4* e *SEXO* do modelo e escolhendo a variável *DUODECADA* como variável grupo, teremos os resultados observados na figura 13.5.

Cox Proportional Hazards							
Term	Hazard Ratio	95%	C.I.	Coefficient	S. E.	Z-Statistic	P-Value
<b>Duodecada(40/20)</b>	<u>3.3009</u>	<u>1.5729</u>	<u>6.9271</u>	1.1942	0.3782	3.1575	<u>0.0016</u>
<b>Duodecada(60/20)</b>	<u>49.4035</u>	<u>21.5344</u>	<u>113.3398</u>	3.9	0.4237	9.2054	<u>0.0</u>
<b>apoe4</b>	<u>2.0615</u>	<u>1.3715</u>	<u>3.0987</u>	0.7235	0.2079	3.4793	<u>0.0005</u>
<b>sexo (Masc/Fem)</b>	1.5462	0.9387	2.5469	0.4358	0.2546	1.7115	0.087
<b>Convergence:</b>	Converged						
<b>Iterations:</b>	6						
<b>-2 * Log-Likelihood:</b>	1014.5659						
Test	Statistic	D.F.	P-Value				
<b>Score</b>	177.271	4	0.0				
<b>Likelihood Ratio</b>	155.5365	4	0.0				

Figura 13.5. Resultado da análise estatística realizada pelo comando *Cox Proportional Hazards* utilizando a variável *DUODECADA*.

Fonte: Epi Info®, arquivo *Apo.e.mdb*.

Observamos que, comparando com a *DUODECADA* menor que 40 anos (*DUODECADA* = 20), as demais possuem razões de dano muito maiores.

Note o aumento desproporcional das faixas etárias mais elevadas. Enquanto a *DUODECADA 40* possui uma razão de dano 3,3 x a razão do grupo jovem, para a *DUODECADA 60* ou mais esse valor é 49 x.

Observamos na figura abaixo que as curvas produzidas pela Regressão de Cox produz uma probabilidade de sobrevida da *DUODECADA 40* maior que a *DUODECADA 60* refletindo que as pessoas que iniciam o acompanhamento entre 40 e 59 anos têm uma probabilidade de se manter livres de doença por mais tempo que pessoas com 60 ou mais anos. Infelizmente o grupo jovem (*DUODECADA 20*) apresenta uma curva menor, o que não é verdade. Aparentemente o programa posiciona erradamente os valores da variável grupo e, infelizmente, não nos dá a opção de corrigi-lo.

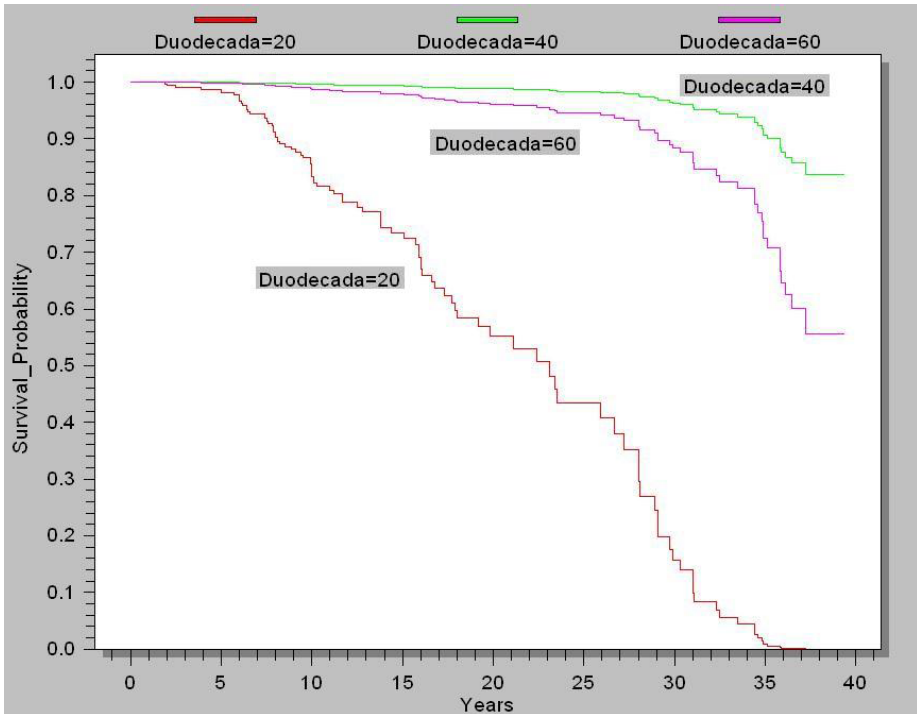


Figura 13.6. Gráfico criado pelo comando *Cox Proportional Hazards* para cada valor de *DUODECADA*.  
Fonte: Epi Info®, arquivo *Apoe.mdb*.

Usando as curvas de Kaplan-Meier, entretanto, podemos observar, na figura 13.7, um comportamento mais coerente com os resultados.

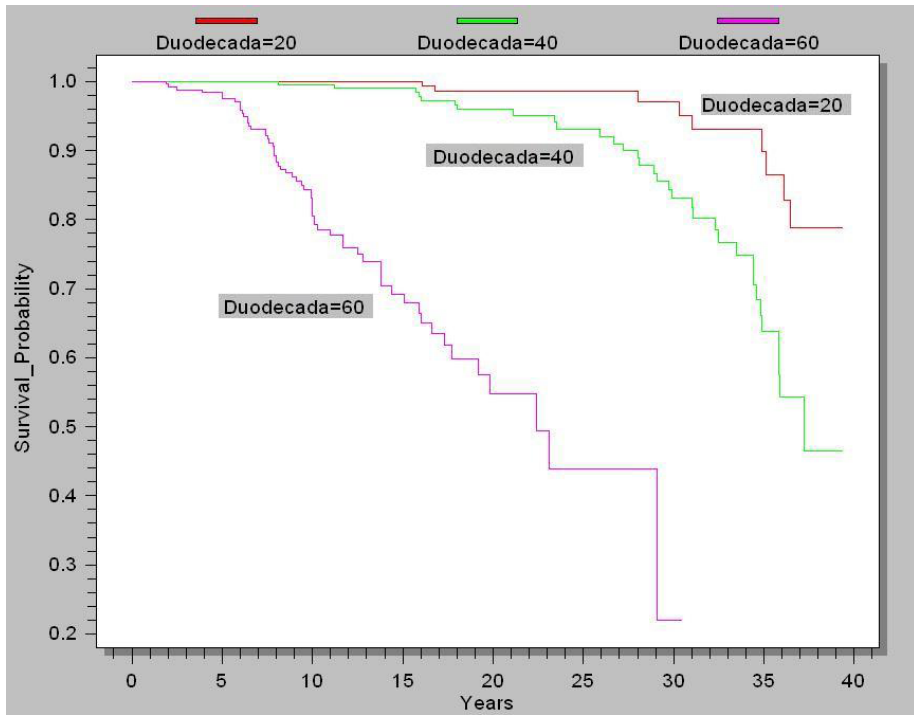


Figura 13.7. Gráfico criado pelo comando *Kaplan-Meier Survival* para cada valor de *DUODECADA*.  
Fonte: Epi Info®, arquivo *Apoe.mdb*.

Confirmando haver algum problema na configuração do gráfico no programa do Cox do Epi Info®, esperamos que esse problema seja corrigido em versões futuras.

**Exercício:**

Usando o mesmo banco de dados (*ApoE\**), calcule as razões de dano dos seguintes modelos:

- a. EVENTO (TEMPO) = APOE4.
- b. EVENTO (TEMPO) = APOE4 COLESTEROL10 (ou a variável DISLIPIDEMIA).
- c. EVENTO (TEMPO) = APOE4 PAS10 (ou a variável HIPERTENSAO).
- d. EVENTO (TEMPO) = APOE4 COLESTEROL10 (ou DISLIPIDEMIA) PAS10 (ou HIPERTENSAO) SEXO DUODECADA.

Comente a respeito do efeito da introdução das outras variáveis sobre os valores da razão de dano da variável APOE4.

\*Poderá usar o banco de dados homônimo que se encontra no folder *Dados* do site [www.pucrs.br/igg/epiinfo](http://www.pucrs.br/igg/epiinfo).