António Paula Brito de Pina

Uma introdução ao EpiInfo

2005

Gabinete de Investigação e Estatística
Delegação Regional do Algarve do Instituto da Droga e Toxicodependência

Índice

I - Apresentação do programa

O EpiInfo é um software que tem as suas raízes em 1985, quando o Center of Disease Control dos EUA o criou, então baseado em MS-DOS, para distribuição gratuita em todo o Mundo.

Este software teve como objectivo apoiar investigadores e gestores de bases de dados na área da saúde e tem evoluído continuamente, recebe o contributo não remunerado de muitos programadores e estatísticos e, finalmente, permite a tradução para muitas línguas, nomeadamente o português.

O actual EpiInfo em ambiente Windows continua a ser periodicamente actualizado (várias vezes ao longo do ano) e está disponível na Internet em www.cdc.gov/epiinfo. Através deste URL poder-se-á fazer as contínuas actualizações que vão aparecendo e participar em fóruns, tanto em inglês como em português (em www.lampada.uerj.br/). É possível ainda fazer perguntas sobre o programa ao próprio *staff* e obter apoio muito rápido.

Os únicos requisitos essenciais a ter no nosso computador é o Windows 98 ou superior instalado e cerca de 260 megabytes livres no disco duro.

Em poucas palavras, quais as funcionalidades deste programa?

1º- Permite construir uma base de dados simples ou relacional, ou ainda ler uma base de dados do Access ou Excel.

2º- Permite calcular a dimensão de uma amostra a seleccionar aleatoriamente.

3º - Permite caracterizar os nossos dados de forma descritiva, ou seja, percentagens e intervalos de confiança, Médias e Desvios-padrões, Medianas e Quartis.

4º - Permite efectuar vários testes de homogeneidade tanto de uma forma paramétrica (t de Student, ANOVA) como não paramétrica (U de Mann-Whitney, Kruskal-Wallis, Qui-quadrado).

5º- Finalmente, permite ainda, para os mais experientes, fazer regressão linear simples e múltipla, regressão logística e análise de sobrevivência.

6º - Permite ainda, tal como o Access, pré-configurar relatórios, esteticamente mais agradáveis, onde os dados são automaticamente actualizados.

Existem ainda outras funcionalidades como o cálculo de risco relativo e Odds Ratio, funções gráficas, etc.

Finalmente, a partir do Menu Inicial é possível ainda encriptar os dados, tendo acesso aos mesmos apenas mediante uma *password* (em Utilities e depois, em Epi Lock), é possível solicitar ajuda através do Help, ou ainda preparar uma aplicação personalizada, com cores, menus e botões específicos.

Nesta introdução ao EpiInfo fala-se também de Estatística, atendendo este programa tratar-se de um software sobretudo estatístico e de epidemiologia.

No entanto, as informações em Estatística e Epidemiologia estão sobretudo desenvolvidas noutro trabalho nomeado *Investigação e Estatística com o EpiInfo* (a consultar em http://www.saudepublica.web.pt/03-Investigacao/031-EpiInfoInvestiga/introdução_estatística.htm), onde também se aprofunda o tema da metodologia de investigação, questão fundamental para enquadrar correctamente a utilização da Estatística.

II - Criar uma base de dados

Uma base de dados no EpiInfo é um "Project" que pode ter várias apresentações ou "Views", por exemplo, um "Project" pode ter uma "View" para médicos e outra para administrativos. Assim, para criar uma base de dados é necessário criar um "Project" e, pelo menos, uma "View" da seguinte forma:

Clicar em MAKEVIEW e depois:

1º - No Menu FILE, seleccione "New" para criar uma nova *View* (claro que, caso quisesse abrir uma já existente, seria "Open")

2º - Dê um nome ao novo "Project" - ex.: Experiência

3º - Dê um nome à "View" - ex.: Terapeuta

4º - Agora vai criar as variáveis ou campos da "View", colocando o cursor no local onde quer ver aparecer a variável e clicando sempre no botão direito do rato. Aparece-lhe uma janela onde escreve o nome da variável e a caracteriza como sendo de texto (até 255 caracteres), numérica, data, multilinha (de texto mas sem limite, ideal para comentários), etc.

Nesta caixa de diálogo é possível também criar variáveis que já tem valores predeterminados (legal values) e dispô-las num botão que as desenrole (pull down button), ou numa caixa de botões de verificação com cruz, e ainda criar um botão para uma outra base de dados relacionada.

Para criar uma segunda variável é necessário, ou repetir o mesmo processo para reabrir mais uma caixa de diálogo, ou simplesmente fazer ENTER (automaticamente a variável seguinte será colocada no espaço vazio seguinte).

Para reeditar e modificar uma variável já existente é só clicar duas vezes por cima do seu nome.

Para mover é clicar e arrastar com o rato.

Para alterar as dimensões é só clicar e arrastar as pequenas setas que aparecem nos limites do campo.

Para exercício proponho que sejam criadas as seguintes variáveis:

Variável	Tipo
Nome	texto
Data de Nascimento	data em formato DD-MM-YYYY
Data Actual	data em formato DD-MM-YYYY
Idade	numérica com 3 algarismos (###)
Peso	numérica com 3 algarismos (###)
Sexo	texto - legal com dois valores (Homem; Mulher)*
Comentários	multilinha

* Seleccione o tipo de variável em "Type" como texto (o que aliás está já sempre seleccionado por defeito). Seguidamente clique no botão "Legal Values". Aparece-lhe uma nova caixa de diálogo onde lhe aparece duas opções: "Use Existing Table" a clicar se já tivéssemos criado os dois valores numa tabela, o que não é o caso actual, e "Create New" que é a opção a escolher neste caso. Depois de criar os dois valores - "Homem", "Mulher" - sair fazendo OK.

Se o espaço da primeira página não chega para colocar todas as variáveis pretendidas, poderemos adicionar mais uma página, clicando em "ADD PAGE".

Neste caso particular, foi criada uma variável *Idade* que poderá ser o resultado de uma operação matemática a partir dos valores de duas variáveis: *Data Actual* e *Data de Nascimento*.

Se quisermos que esta operação seja feita automaticamente quando introduzirmos os dados, temos que previamente criar um pequeno Programa e para isso clicamos, dentro do ecrã MAKEVIEW, no botão PROGRAM, situado à esquerda.

Aqui abrir-se-á uma janela onde em “Choose field where actions will occur” vamos seleccionar *Idade* e depois clicamos em ASSIGN. Esta variável seleccionada para *ASSIGN* é aquela que receberá o resultado dos cálculos. Neste caso particular, poderemos atribuir-lhe uma função já existente no EpiInfo, que produz o número de anos resultante da diferença entre duas datas. Esta função, assim como todas as outras que já existem no EpiInfo, poderão ser consultadas no Manual do EpiInfo ou no próprio Help do software. Neste caso, a função a digitar será Years(DataDeNascimento,DataActual) precedida de um sinal de igual ”=”. No final o editor deverá ter escrito: ASSIGN Idade=Years(DataDeNascimento,DataActual).

Agora é só salvar e sair do PROGRAM.

Este pequeno exemplo serve apenas para familiarizar o leitor para as possibilidades de automatizar previamente muitas funções e operações no EpiInfo.

III - Introduzir dados

1º- Sair do MAKEVIEW.

2º- Clicar em ENTER DATA.

3º- Abrir o "Project" e a "View" já existente e introduzir, simplesmente, os dados.

Para continuarmos o nosso exercício proponho que se preencham os dados para cinco fichas, da seguinte forma:

Nome	Data de Nascimento	Data Actual	Peso	Sexo	Comentários
João	13-12-1973	15-12-2000	60	homem	Não quis colaborar
Maria	10-07-1961	15-12-2000	55	mulher	Muito faladora
Manuel	20-11-1980	15-12-2000	80	homem	Vê demasiada televisão ...
Luís	11-05-1950	15-12-2000	75	homem	
Ana	15-08-1985	15-12-2000	50	mulher	

Repare-se como automaticamente, depois de preenchermos os valores para as duas datas, aparece a *Idade* calculada pelo EpiInfo!

Depois de preencher muitas fichas é natural que se queira ocasionalmente encontrar uma delas. Para nos movermos pela base de dados bastará clicar nos botões situados no canto inferior esquerdo. Para encontrar uma ficha ou um conjunto de fichas com características particulares bastará clicar em FIND.

Para modificar ou corrigir valores entrados, poder-se-á voltar a entrar os novos valores depois de encontrar a ficha pretendida. No entanto, por vezes é conveniente ver a base de dados em matriz (ou seja, listando todos os valores numa tabela) de forma a detectar incorrecções, porque é mais cómodo termos uma visão de conjunto neste caso, do que ver e corrigir apenas uma ficha de cada vez. Para termos esta visão do conjunto é útil clicar em LIST na secção ANALYSIS e seleccionar "update" para fazer as correcções (ATENÇÃO: para analisar os dados nesta secção ANALYSIS é necessário sempre começar pelo comando READ, ou seja, é necessário sempre fazer o computador ler previamente a nossa base de dados).

IV - Criar, modificar ou recodificar variáveis numa base dados já existente

A. Deletar variáveis: através do MAKEVIEW entramos no "Project" e "View" pretendida. CLICAR com o botão direito do rato na variável que pretendemos deletar e seleccionar "Delete".

   Neste caso, para exercício, vamos deletar a variável *Idade*. Ao deletar esta variável vamos também apagar o pequeno programa que permitia o seu cálculo. Se a voltarmos a criar dentro do ecrã MAKEVIEW, os dados anteriormente inseridos não serão objecto de qualquer cálculo. Para voltar a recriar esta variável, de forma que os dados anteriormente inseridos sejam recalculados *à posteriori* , teremos que a voltar a criar de uma outra forma, através do ecrã ANALYSYS, o que será explicado adiante.

B. Deletar fichas: através do ENTER entramos no "Project" e "View" pretendida e após abrirmos as fichas que pretendemos deletar, marcamo-las como deletadas (botão à esquerda).

C. Criar variáveis novas para preenchimento de dados: através do MAKEVIEW entramos no "Project" e "View" pretendida. Clicar com o botão direito do rato e criamos a nova variável. Posteriormente, através do comando ENTER teremos que preencher os dados.

D. Criar variáveis a partir das existentes. Por exemplo, suponhamos que a partir das variáveis "Data de Nascimento" e "Data Actual", já existentes e com muitos dados inseridos, queremos novamente recriar a variável *Idade*:

   1º- Entrar na secção ANALYSIS.

   2º- Ler a base de dados (READ), atendendo que nesta secção é sempre necessário começar pelo comando READ, ou seja, é necessário sempre fazer o computador ler previamente a nossa base de dados.

   3º- Definir e nomear uma nova variável através do comando "DEFINE". Nomeá-la-emos *Idade*. Fazer OK.

   4º- Clicar no comando ASSIGN (que significa “atribuir” um valor) e seleccionar a recém-criada variável *Idade*.

5º- Ver no Manual do EpiInfo a lista de "Functions and Operators" (ou no próprio Help do software). Tal como já foi explicado, nesta lista existe uma função Years que nos dá o número de anos entre uma primeira data e uma segunda data. A sintaxe do comando a escrever é:

Years(DataDeNascimento,DataActual)

Atenção: deve-se escrever tudo sem espaços e utilizando os nomes das variáveis que o próprio EpiInfo fornece no item "Available Variables".

Fazer OK e a nova variável está criada.

Para o confirmar basta clicar à esquerda em LIST, escolher as variáveis a listar (obviamente uma delas será *Idade* ...) e visualizar se tudo funcionou bem.

Finalmente, vamos também criar uma outra variável que não é consequência de uma operação lógica, mas sim uma simples recodificação personalizada. Suponhamos que queremos dividir a nossa amostra em dois grupos - "Jovens", i.e., com menos de 18 anos, "Adultos", i.e., os que têm idades superiores. Suponhamos que a nova variável se chamará *Grupos*. Após definir da mesma forma a nova variável, vamos clicar no comando "RECODE" - abre-se uma janela de diálogo: no local denominado "From" seleccionamos a variável original (neste caso *Idade*) e no local "To" seleccionamos a nova variável definida (neste caso será *Grupos*).

Depois preenchemos os intervalos de valores que queremos recodificar, por ex., de 0 a 17 anos, ou de LOVALUE (valor mais baixo) a 17 anos, e no local do novo valor codificado escrevemos "Jovens". Fazemos "Enter" e continuamos para o grupo seguinte: de 18 a 80 anos ou de 18 a HIVALUE (valor mais alto) e no local do novo valor codificado escrevemos "Adultos".

Uma informação importante: quando um intervalo se sobrepõe a outro, são sempre os limites do primeiro que contam.

No entanto, é preciso clarificar que este processo cria provisoriamente as novas variáveis. Se sairmos da secção ANALYSIS ou, simplesmente, voltarmos a ler a base de dados (através do READ) todo este trabalho se perde.

Se queremos que as novas variáveis se tornem permanentes, teremos que salvar esta base de dados com o mesmo nome ou com outro nome através do comando WRITE (ao clicar em WRITE teremos que seleccionar um nome para a FILENAME e um nome para a DATATABLE).

Na janela de diálogo aberta há sempre duas opções a escolher: APPEND e REPLACE. Se optarmos pelo APPEND, as fichas da primeira base de dados são adicionadas às fichas já existentes da nova base de dados, caso existam (ou seja, a informação já existente é conservada). Se optarmos pelo REPLACE a informação da nova base de dados é totalmente substituída pela primeira.

No entanto, geralmente não é necessário criar novas bases de dados porque será sempre possível, com facilidade, recriar as variáveis provisórias, caso se salve simplesmente o programa, ou seja, a sequência de comandos que executámos para as criar, tal como vamos explicar seguidamente.

V - Criar programas para analisar os dados

Ainda em ANALYSIS, a área inferior direita corresponde ao "Program Editor", onde se poderá criar programas. Estes programas não são mais que sequências de comandos, como por exemplo, os que acabámos de construir para a criação de novas variáveis. Se salvarmos um tal programa, poderemos em qualquer momento voltar a corrê-lo automaticamente, ou seja, poderemos recriar as mesmas variáveis provisórias.

Na área do "Program Editor" poderemos ver os últimos comandos que executámos. Neste caso, há comandos que não nos interessa salvar: por ex., o comando LIST que executámos anteriormente apenas para verificar que tudo estava a funcionar bem. Assim, o melhor será seleccioná-lo e deletá-lo. Posteriormente, salva-se o programa dando-se um nome - ex.: "Criação de novas variáveis" dentro do Project "Experiência". Agora, para verificar que tudo funciona, recomendo sair de ANALYSIS e voltar a entrar, ler a base de dados, abrir o programa e clicar em RUN. Finalmente, com o comando LIST verifique se as variáveis foram mesmo criadas.

Criar programas pode ser muito útil se queremos fazer a mesma análise estatística muitas vezes, como é a recriação frequente de novas variáveis provisórias ou quando queremos aplicar os mesmos testes em subgrupos diferentes da amostra, nomeadamente para cada sexo ou grupo etário, social, etc.

VI - Efectuar análises estatísticas

A maioria dos comandos que permitem fazer análises estatísticas estão na secção ANALYSIS.

Nesta secção, novamente realço que é sempre necessário começar pelo comando READ, ou seja, é necessário sempre fazer o computador ler previamente a nossa base de dados. Seguidamente poder-se-á analisar os dados utilizando comandos diferentes de acordo com os nossos objectivos, nomeadamente:

1º- FREQUENCIES: Percentagens e Intervalos de confiança (adequado a variáveis qualitativas ou categóricas) - experimente FREQUENCIES de "Sexo".

2º- TABLES: devolve-nos tabelas de contingência entre duas variáveis (também adequado a variáveis qualitativas ou categóricas); no caso de variáveis binomiais calcula-nos automaticamente o valor do Qui-quadrado, Prova de Fisher, Odds Ratio, Risco relativo e respectivos intervalos de confiança - experimente TABLES de "Sexo" (na "Exposure variable") e "Grupos" (na "Outcome variable").

3º- MEANS: Média e Desvio-padrão, Mediana e Quartis, Moda, valor máximo e mínimo (adequado a variáveis quantitativas) - experimente fazer Means de "Idade"; é possível ainda aplicar a prova t de Student ou ANOVA, e em alternativa, de U de Mann-Whitney ou Kruskal-Wallis - experimente fazer Means de "Idade" cruzada com "Sexo".

Outros comandos, como os que são utilizados na regressão linear, regressão logística e na análise de sobrevivência não são especificamente referidos porque devem ser utilizados apenas por quem tem uma formação sólida em estatística.

Finalmente, na secção STATCALC é possível efectuar diversos cálculos com dados entrados directamente do teclado, nomeadamente:

- calcular a dimensão de uma amostra seleccionada aleatoriamente;
- aplicar o Qui-quadrado (prova de homogeneidade ou de independência) a tabelas;
- calcular o Odds Ratio, Risco Relativo e respectivos intervalos de confiança;
- executar a prova da tendência linear do Qui-quadrado.

VII - Compreender os resultados estatísticos no EpiInfo

A- Para que servem os testes estatísticos?

Os testes estatísticos , nomeadamente os seus resultados em termos de probabilidades "p" de significância estatística, são sempre apenas um desenvolvimento da teoria das probabilidades.

Um exemplo muito simples pode dar-nos uma ideia do que estou a dizer.

Suponhamos que atiramos 10 vezes uma moeda ao ar. É evidente que se a moeda não estiver viciada, em princípio deveremos obter cerca de 5 "caras" e 5 "coroas". No entanto, é muito provável que o resultado não seja tão claro assim. É muito provável que tenhamos 6 "coroas" e 4 "caras", por exemplo... Evidentemente, a probabilidade de obtermos 7 "coroas" e 3 "caras" há-de ser menor, embora ainda assim, possa suceder. Finalmente, a probabilidade de obtermos um resultado muito extremo como 9 "coroas" e 1 só "cara" é ainda menor, embora também possa suceder excepcionalmente. Claro que quanto mais excepcional for o resultado, mais acreditaremos que a moeda está viciada pois a probabilidade de tal acontecer numa moeda não viciada é cada vez mais pequena.

A estatística mede as probabilidades associadas a estes acontecimentos e assim, pode ajudar-nos a tirar as nossas conclusões sobre os factos (segundo os estatísticos existe uma pequena *nuance*: a medição é feita ao contrário da teoria das probabilidades, ou seja, em vez de se partir de uma população para a medição teórica da probabilidade do evento , parte-se de um evento concreto para a estimativa da sua probabilidade).

Aplicando o mesmo princípio, existem testes que comparam duas amostras e nos dizem qual a probabilidade de estas serem diferentes. Por exemplo, através do comando MEANS no EpiInfo, podemos ver se as mulheres são diferentes dos homens quanto à variável idade. Tal como no caso da moeda, há a possibilidade de haver diferenças entre a idade dos dois grupos, mas isto poderá ser devido apenas ao acaso e não a *verdadeiras* diferenças. Os testes estatísticas medem sempre a probabilidade de as diferenças encontradas serem devidas ao acaso, partindo do pressuposto que na *verdade* não existem diferenças. Se a probabilidade encontrada for pequena, teremos mais confiança em afirmar que as mulheres e os homens têm idades diferentes.

Geralmente em ciências da saúde, quando estas probabilidades são inferiores a 5%, ou seja, há menos de 5 possibilidade em 100 de suceder um determinado resultado, nós consideramos que são estatisticamente significativas.

É importante também referir que existem 3 formas de aplicar os testes estatísticos:

1º - Provas de conformidade, ou seja, para verificar se há diferenças entre uma amostra e uma população (ex.: na minha amostra tenho 20 mulheres e 80 homens e eu sei que na população a proporção de mulheres é de 30%. Será que existe diferença entre a minha amostra e a população? Por outras palavras, até que ponto a minha amostra é representativa da população?).

2º - Provas de homogeneidade, ou seja, para verificar se há diferenças entre dois grupos (um exemplo já referido será verificar se há diferença entre homens e mulheres quanto à idade).

3º- Provas de independência, ou seja, verificar se duas variáveis simétricas são independentes (ex.: se a cor dos olhos é independente ou está associada `a raça).

O EpiInfo permite facilmente fazer as provas de homogeneidade e de independência (que, em termos práticos, são semelhantes) mas infelizmente, não permite executar as provas de conformidade.

Esta distinção é importante ter presente principalmente quando formos aplicar o Qui-quadrado no EPiInfo. O Qui-quadrado poderá ser aplicado como Prova de conformidade ou como Prova de homogeneidade/independência mas, repete-se, o Qui-quadrado do EpiInfo só serve para estas últimas provas e nunca para a de conformidade.

Será importante também referir previamente uma limitação genérica da estatística: nem sempre aquilo que é estatisticamente significativo é importante! Repare-se: suponhamos que queremos saber se as mulheres são diferentes dos homens quanto à idade. Se aplicarmos um teste estatístico os seus resultados são tanto mais estatisticamente significativos quando maiores forem as diferenças entre os dois grupos, e também, quanto maior for a dimensão da amostra estudada. Isto significa que por vezes, pequeníssimas diferenças entre os dois grupos podem ser estatisticamente significativas se a amostra tiver grandes dimensões. Mas a questão que aqui se põe é: são estas pequeníssimas diferenças importantes do ponto de vista clínico, social, etc.?

Neste caso particular, o aprendiz de estatística tenderá a valorizar um qualquer resultado estatisticamente significativo, mesmo que este não tenha interesse absolutamente nenhum, atendendo ter a ver com pequeníssimas diferenças...

Posto isto vamos agora partir para a compreensão dos testes estatísticos no EpiInfo.

B- Comando FREQUENCIES: os intervalos de confiança das proporções

Quando solicitamos o comando "FREQUENCIES" para uma variável como o sexo, por ex., o EpiInfo dá-nos os valores das percentagens de cada sexo e o Intervalo de Confiança de 95% para as mesmas percentagens.

Este intervalo de confiança só tem interesse se as percentagens em causa são de uma amostra seleccionada aleatoriamente[1] de uma população mais vasta.

Por exemplo, suponhamos que na nossa base de dados "Experiência" colhemos aleatoriamente a amostra de 5 elementos, de todos os utentes dum serviço de saúde, no qual verificamos que 3 são homens e 2 são mulheres. Neste caso, poderemos dizer que nesta amostra existem 60% (3/5=0,6) de homens e 40% (2/5=0,4) de mulheres. Mas será que as percentagens de cada sexo, em toda a população de utentes, são também estas? Nunca o saberemos ao certo com estes dados. No entanto, poderemos dizer que, aplicando a prova que o EpiInfo aplica, poderemos *acreditar* com uma confiança de

[1] A selecção aleatória de uma amostra implica que cada elemento seleccionado tenha exactamente a mesma probabilidade de o ser.
Existem vários métodos de selecção aleatória:

1- Selecção aleatória simples: é necessário ter uma listagem de toda a população, por exemplo, de todos utentes dum serviço de saúde, aos quais se atribui um número de 1 a x. Posteriormente, ao acaso, seleccionam-se alguns elementos de toda a lista, geralmente com a ajuda de uma tabela de números aleatórios.
2- Selecção aleatória sistemática: é necessário também ter uma listagem de toda a população. Depois seleccionam-se elementos de x em x intervalos, ou seja, selecciona-se 1 elemento de 10 em 10 da listagem.
3- Selecção aleatória por conglomerados: é necessário uma listagem dos conglomerados, por exemplo, uma listagem de escolas da região. Após se seleccionar aleatoriamente um pequeno número de escolas desta listagem, estudam-se todos os alunos de cada escola seleccionada.
4- Selecção aleatória estratificada: é necessária a definição dos vários estratos da população e ter uma listagem de todos os elementos de cada estrato para fazer uma selecção aleatória de uma amostra por cada estrato.

Os métodos de selecção não aleatória poderão ser utilizados em casos especiais, mas têm o defeito de nunca garantirem minimamente a representatividade da amostra.

95% que a percentagem de homens na população estará algures entre 14,7% e 94,7% e a percentagem de mulheres entre 5,3% e 85,3%[2].

Repare-se que neste caso os intervalos de confiança são muitíssimos dilatados, atendendo que a amostra em causa conta com apenas 5 elementos, pelo que o erro de amostragem é enorme.

É evidente que quanto maior for a nossa amostra, mais pequeno será o intervalo de confiança e por isso, mais provável será obtermos extrapolações precisas das *verdadeiras* percentagens da população.

Mas atenção: mesmo este intervalo não é uma certeza pois tem uma confiança de 95%, ou seja, há sempre uma probabilidade de 5% de a verdadeira percentagem estar fora destes limites...

É claro que, se as percentagens em causa forem calculadas tendo por base não uma amostra mas toda a população, os intervalos de confiança que o EpiInfo automaticamente *vomita* não têm significado absolutamente nenhum, pelo que devem ser ignorados. Um exemplo é quando um médico introduz os dados de todo o seu ficheiro clínico e depois quer saber a percentagem de cada sexo para o seu ficheiro. Se o resultado for 35% de mulheres, é mesmo 35% sem qualquer dúvida ou intervalo de confiança, pois ele quis saber a percentagem de mulheres do seu ficheiro que, neste caso, está totalmente informatizado.

Finalmente, tenha-se em atenção que se a amostra não é aleatória, também não será legítimo falar-se em intervalos de confiança para a população, porque aqui a amostra não será representativa de nenhuma população conhecida.

C- Comando TABLES e a secção STATCALC: o Qui-quadrado

- Tabelas 2x2

Se executar o comando TABLES na base de dados que já criámos ("Experiência") em que pomos Sexo como variável de exposição e Grupos (etários) como variável "output" depararemos com a seguinte tabela:

[2] Novamente, segundo os Estatísticos, a interpretação deverá ser feita com uma pequena *nuance*: um Intervalo de confiança de 95% significará que se seleccionarmos um número infinito de amostras sempre da mesma forma aleatória, em 95% delas, estarão incluidos os valores "verdadeiros" no intervalo de confiança resultante.

		Variável "output" (Grupos)		
		Adultos	Jovens	Total
Variável de exposição (Sexo)	Homem	3	0	3
	Mulher	1	1	2
	Total	4	1	5

Também na secção STATCALC é possível construir tabelas mas, neste caso, directamente do teclado, e não a partir de uma base de dados já existente.

Seja como for, face aos dados da tabela exemplificada em cima, pode-se perguntar se o grupo de "Adultos" é diferente do grupo de "Jovens" quanto ao sexo.

Vamos supor que na verdade não existe uma diferença entre os dois grupos quanto ao sexo. Novamente, mesmo que seja esta a verdade, é também possível que quando colhemos uma amostra o resultado possa apresentar uma diferença por questões ligadas ao acaso.

É claro que quanto maior for a dimensão da amostra, mais fácil será identificar diferenças verdadeiras entre os dois grupos. Também, quanto maior for a diferença no género sexual entre os dois grupos, mais provável será a existência desta diferença *verdadeira* entre os dois grupos.

O Qui-quadrado mede a probabilidade de os dois grupos da nossa amostra serem semelhantes, partindo do pressuposto que, na verdade, são mesmo semelhantes na população donde provêm. Se a probabilidade for alta poderemos concluir que não há diferenças estatisticamente significativas. Se a probabilidade for baixa (particularmente menor que 5%) poderemos concluir que o grupo de "Adultos" é diferente do grupo de "Jovens" quanto ao sexo, e de forma estatisticamente significativa.

No entanto, o Qui-quadrado tem limitações, nomeadamente, deverá ser substituído pela Prova exacta de Fisher quando os valores esperados nas células da tabela são inferiores a 5.

Assim, eu recomendo que se verifique sempre se somos avisados - "Warning: the expected value of a cell is < 5. Fisher Exact Test should be used". Nestes casos, evidentemente utilizaremos o "p" unilateral de Fisher ("1-tailed P-value"). Quando não recebemos este aviso poderemos utilizar o valor "p" do Qui-quadrado não corrigido.

No caso do nosso exemplo o valor "p" do Qui-quadrado seria 0,17 mas o valor a utilizar deveria ser o de Fisher, ou seja 0,40 (o que significaria que as eventuais diferenças não seriam estatisticamente significativas).

- Outras tabelas (com mais de 2 filas e/ou colunas)

Neste caso a Prova de Fisher não é aplicável (só o é para tabelas de 2x2) pelo que o seu resultado nunca aparece.

O comando TABLES produz a tabela e calcula o Qui-quadrado mas, nestes casos, não nos avisa quando os valores esperados nas células da tabela são inferiores a 5.

Só a secção STATCALC nos dá tais avisos pelo que eu recomendo sempre reproduzir estas tabelas, com mais de duas filas ou colunas, nesta secção. Deverá ser aceite o valor de "p" proposto excepto quando somos avisados que o valor esperado de uma célula é inferior a 5. Nestes casos, como já não podemos utilizar a Prova de Fisher, resta-nos agregar a tabela de forma a conter menos colunas ou filas, e voltar a aplicar o Qui-quadrado.

Também na secção STATCALC existe a possibilidade muito interessante de se fazer a prova da tendência linear do Qui-quadrado.

Suponhamos que temos uma tabela do género:

		Doenças cardíacas (Variável resultado)	
		Sim (casos)	Não (controlos)
Variável de exposição (Consumo de tabaco)	1 (não fuma)	5	85
	2 (1-15 cigarros/dia)	6	54
	3 (>15 cigarros/dia)	9	41

Fonte:Massons, J.M.D. - Métodos estadísticos en ciencias de la salud, UD 10 - Barcelona, 11ª Ed, ISBN: 84-8049-189-2, 1999.

Neste caso, temos uma variável de exposição ordenada e uma variável resultado dicotómica. Se, fizermos o Qui-quadrado obteremos o seguinte resultado: p=0,0629 .

Este resultado diz-nos que não há diferenças estatisticamente significativas (para um nível de significância convencionado de 0,05) entre os doentes e não doentes quanto

ao seu consumo de tabaco, mas não tem em conta o efeito crescente da variável exposição.

Se entrarmos em conta com este efeito, não só tornamos mais potente o teste como poderemos verificar existir uma relação linear entre as duas variáveis. É o que faz a Prova da tendência linear do Qui-quadrado cujo p=0,0206. Ou seja, há uma relação linear estatisticamente significativa entre o nível de consumo de tabaco e a existência de doença cardíaca.

Esta Prova da tendência linear só poderá ser aplicada quando a variável resposta seja dicotómica e a variável exposição seja quantitativa ou ordinal (variável de categorias ordenadas em três ou mais níveis).

D - Comando MEANS

D.1. - Interpretar as Medianas, Quartis e Modas

Aqui será importante fazer um pequeno parêntesis para classificarmos as variáveis quanto à sua escala de valor. Compreender esta classificação é de importância crucial para utilizarmos de forma adequada a estatística. Sumariamente, poderemos classificar as variáveis da seguinte forma:

1- Variáveis qualitativas nominais: são variáveis cujos valores não tem uma relação de ordem entre eles, por ex., o Sexo e Raça.

Para este tipo de variáveis poder-se-á fazer o estudo das proporções em FREQUENCIES e aplicar-se o Qui-quadrado em TABLES ou na secção STATCALC.

2- Variáveis qualitativas ordinais, cujos valores não são métricos mas incluem relações de ordem. É o caso da variável "Peso" medida em 3 níveis (pouco pesados, pesados, muito pesados).

Para este tipo de variáveis poder-se-á fazer tudo quanto é possível fazer-se para as variáveis nominais, mas também adicionalmente é possível estudar as medianas, quartis, modas, e aplicar o Kruskal-Wallis no comando MEANS.

3- Variáveis quantitativas, cujos valores são medidos numa escala métrica, como por ex., a "Idade", ou o "Peso" medido em gramas.

Para este tipo de variáveis poder-se-á fazer tudo quanto é possível fazer-se para as variáveis nominais e ordinais, mas também adicionalmente é possível estudar as médias, desvios-padrão, e aplicar o ANOVA dentro do comando MEANS.

Execute agora o comando MEANS para "Idade", no Project "Experiência" que já construímos.

Enquanto os comandos FREQUENCIES e TABLES são apropriados para variáveis qualitativas (nominais ou ordinais), o comando MEANS é apropriado especialmente para variáveis quantitativas. Isto porque não é possível pedir-se a Média da variável Sexo, mesmo quando por razões de codificação atribuímos números códigos aos seus valores (ex.: sexo masculino=1; sexo feminino=2). Neste caso, se fosse pedido o comando MEANS, os resultados não teriam evidentemente significado.

Apesar de tudo, o comando MEANS pode ter interesse para um tipo particular de variáveis qualitativas - as variáveis ordinais (como é um exemplo já referido dos três níveis de "Peso" ou dos três níveis de fumadores, onde é claro que existe uma relação de ordem). Nestes casos, as Médias continuam a não ter significado, mas as Medianas já poderão tê-lo.

A Média aritmética assim como o Desvio-padrão que lhe está associado, são conceitos que geralmente oferecem poucas dúvidas pelo que aqui não os abordaremos directamente.

O conceito de Mediana, no entanto, gera muitas confusões: a Mediana é simplesmente o valor que se situa a meio da fila ordenada de valores, desde o mais baixo ao mais alto. Assim, tem que haver uma relação de ordem nos valores, pelo que a Mediana pode ser calculada tanto para as variáveis ordinais como para as quantitativas puras. Um exemplo: no caso da base de dados "Experiência", já construída, temos cinco elementos cujas idades são ordenadas através do comando MEANS:

15
20
27
39
50

O número 27 representa o valor que está a meio, ou seja, é a Mediana.

O número 20 representa o valor que está a meio da primeira metade, ou seja, é o primeiro Quartil ou Percentil 25.

O número 39 representa o valor que está a meio da segunda metade, ou seja, é o terceiro Quartil ou Percentil 75.

Claro que a mediana é também o segundo Quartil e o Percentil 50.

A Moda é o valor mais frequente (ou seja, o que "está na moda"...). Neste caso, como não existe nenhum valor mais frequente, o EpiInfo dá-nos o menor valor, o que não tem significado absolutamente nenhum, podendo mesmo induzir-nos em erro. O que se passa é que quando existem várias Modas, o EpiInfo apresenta sempre a menor: ou seja, se numa amostra existem 10 pessoas com 20 anos e 10 pessoas com 30 anos, sendo todas as outras idades menos frequentes, sucede que existem duas Modas, mas o EpiInfo vai referir apenas a que apresenta o menor valor ou seja, dirá que 20 anos é o valor mais frequente. Por isto, se nos interessa referir a Moda, convém verificar se não há outro valor tão frequente na nossa amostra. Para isto basta executar o comando FREQUENCIES, que nos dá as frequências de todos os valores.

Qual a diferença de interpretação entre a Mediana e a Média?

Em primeiro lugar a Mediana pode ser utilizada tanto em variáveis quantitativas como em variáveis qualitativas ordinais, enquanto a Média só pode ser utilizada em variáveis quantitativas.

Em segundo lugar, no caso das variáveis quantitativas, embora a Média seja um valor mais fácil de entender, tem o defeito de nos induzir em erro se a nossa amostra tiver valores muito extremos. Por exemplo, na distribuição de idades da nossa amostra a Média é de 30,2 e a Mediana de 27. Imagine que o indivíduo mais velho tinha não 50 anos de idade mas sim 100 anos. Isto faria com que a Média *saltasse* para 40,2, ou seja, seria superior a quase todos os valores individuais, mas a Mediana continuaria a ser 27. Se olharmos para todos os 5 valores individuais da nossa amostra, verificamos que o n.º 27 é melhor representante da distribuição global da idade na nossa amostra que o erróneo n.º 40,2.

Assim, no caso das variáveis quantitativas, quando o valor da Mediana é muito diferente da Média, é aconselhável considerar sempre a Mediana como valor de referência mais importante.

D.2. - Interpretar as provas de homogeneidade (ANOVA e Kruskal-Wallis)

Agora execute o comando MEANS da variável Idade segundo o Sexo ("crosstabulated by value of ...").

Além das Médias, Desvios-padrões, Medianas, Quartis, etc. das idades para os dois sexos, aqui temos ainda os resultados dos típicos testes de homogeneidade nos quais a pergunta é "haverá diferenças entre os dois grupos (masculino/feminino) quanto à idade?".

O EpiInfo *vomita* os resultados do teste ANOVA (correspondem ao teste t de Student quando é aplicado apenas para duas subamostras) e do teste de Kruskal-Wallis (que correspondem ao teste U de Mann-Whitney/Wilcoxon quando é também aplicado apenas para duas subamostras).

O teste ANOVA exige muitos pressupostos pelo que é perigoso ser utilizado por principiantes, especialmente em amostras de pequena dimensão. Em alternativa, recomendo utilizar sempre os resultados do teste de Kruskal-Wallis porque é robusto, muito conservador e não exige nenhum pressuposto. O Kruskal-Wallis pode ser utilizado para variáveis quantitativas e qualitativas ordinais tal como a Mediana.

No entanto, caso haja interesse em utilizar o ANOVA, atendendo que quando se cumprem os seus pressupostos, este teste é de facto um pouco mais potente que o Kruskal-Wallis, recomenda-se fazê-lo só nestas circunstâncias:

1º- A variável a testar terá de ser quantitativa.

2º- Quando os grupos têm dimensões diferentes, deverá existir homogeneidade nas variâncias, ou seja, o "p" do teste de Bartlett, efectuado automaticamente pelo EpiInfo deve ser superior a 0,05. No caso do nosso exemplo é 0,9385 pelo que se conclui estar cumprido este pressuposto.

3º- Quando pelo menos um dos grupos tem menos de 30 elementos, deverão os diversos grupos ter uma distribuição Normal. Infelizmente, o EpiInfo não executa qualquer teste para confirmar esta Normalidade, pelo que se aconselha a nunca aplicar o ANOVA nestes casos.

No caso do nosso exemplo, o grupo de mulheres tem apenas 2 elementos e o grupo de homens apenas 3 elementos, pelo que nunca se deveria utilizar o ANOVA. Apenas o teste de Kruskal-Wallis poderia ser aplicado, não sendo as diferenças estatisticamente significativas (p=0,5637).

E - Calcular a dimensão de uma amostra através do STATCALC

Esta possibilidade de calcular a dimensão de uma amostra é muito útil se queremos partir para o estudo com alguma confiança sobre a possibilidade de, no futuro, podermos extrapolar os nossos resultados para a população. Por outras palavras, a dimensão da amostra tem tudo a ver com a precisão dos intervalos de confiança que queremos vir a ter quando fizermos os nossos cálculos. No entanto, é necessário ter em

conta que esta amostra terá que ser obrigatoriamente seleccionada pelo método aleatório simples ou sistemático.

Para isso, na secção STATCALC poderemos verificar qual a dimensão correcta da nossa amostra, escolhendo "Sample size & power" e depois "Population survey". Teremos que responder seguidamente às perguntas colocadas, nomeadamente:

1º- Qual a dimensão da população total? Experimente pôr 5000.

2º - Qual a frequência que julgamos ser *verdadeira* na população total? É evidente que não estamos certos desta frequência, no entanto, tendo em conta outros estudos ou informações poderemos estimar esta frequência... Quando não fazemos a mínima ideia desta frequência real, poderemos escolher o valor mais conservador que é 50%. Experimente então pôr 50%.

3º - Qual o valor mais errado que admitiríamos obter da nossa amostra? Suponhamos que admitiríamos ter um intervalo de confiança de 50% $\pm$ 10%, ou seja seria obter ou 60% ou 40% como limites. Terá que responder a esta pergunta colocando ou 60 ou 40%.

Veja agora o resultado: terá de ter uma amostra de 94 elementos se quiser obter intervalos de confiança de 95% ("confidence level of 95%"), cujo limites não ultrapassem 60 ou 40%, partindo do pressuposto que a verdadeira proporção é de 50%, e que a amostra será seleccionada pelo método aleatório simples ou sistemático.

Complicado? Talvez, mas melhor que isto só se perguntar directamente a Deus ...

VIII
Bibliografia

- Abramson, J.H. - Survey methods in community medicine - an introduccion to epidemiological and evaluative studies - New York, Churchill Livingstone, 2º ed, 1979.
- Dean, A.G.; et al - Epi Info 2000, a database and statistics program for public health professionals for use on Windows 95, 98, NT, and 2000 computers - Centers for Disease Control and Prevention, Atlanta, Georgia, USA, 2000.
- Doménech Massons, José M. - Métodos Estadísticos en Ciencias de la Salud, Universitat Autònoma de Barcelona, Espanha, 1999.