Revista do

I D P C

Ano 2 Nº 7

Instituto Dante Pazzanese de Cardiologia

Entrevista: o Prof. Moisés Mishel Levy discorre sobre a importância do marketing médico

Questões comuns sobre epidemiologia, estatística e informática

Angiogênese: revascularização transmiocárdica

Um coração movido a pesquisa, aprendizado e ensino na seção *Resgatando nossa história*

ESPAÇO ABERTO

Questões comuns sobre Epidemiologia, Estatística e Informática

Referência: PEREIRA, J. C. R.; PAES, A. T.; OKANO, V. **Espaço aberto: Questões comuns sobre epidemiologia, estatística e informática** Revista do IDPC, São Paulo, v. 7, p. 12-17, 2000

Júlio Cesar Rodrigues Pereira*

Ângela Tavares Paes*

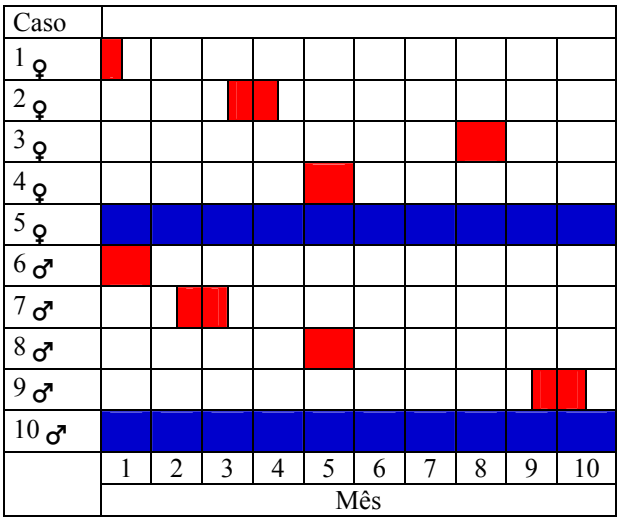
Valdir Okano*

* Laboratório de Epidemiologia e Estatística (LEE) – IDPC

1. QUAL A DIFERENÇA ENTRE PREVALÊNCIA E INCIDÊNCIA DE DOENÇA?

Ambas são medidas de freqüência de ocorrência de doença. Prevalência mede quantas pessoas **estão doentes**, incidência mede quantas pessoas **tornaram-se doentes**. Ambos os conceitos envolvem espaço e tempo – quem **está** ou **ficou** doente *num determinado lugar numa dada época*.

O quadro abaixo representa um espaço e tempo imaginários, com o período de estado de uma doença sendo representado por barras com um comprimento igual a 1 mês. De 10 casos observados, um (caso 1) iniciou o período de observação já doente (barra pela metade), 7 tornaram-se doentes neste período, 2 não registraram doença.



A prevalência da doença seria 80% (8 doentes entre 10 casos), e a incidência seria de 70% (7 tornaram-se doentes entre 10 casos). Neste universo de espaço e tempo, poder-se-ia ter cortes específicos para medidas específicas. Num corte de espaço, sendo os 5 primeiros casos mulheres e os 5 últimos homens, poder-se-ia dizer, por exemplo, que a prevalência é igual em ambos os sexos (4 doentes entre 5 casos cada) embora a incidência seja maior entre homens (4 homens tornaram-se doentes entre 5 casos e 3 mulheres tornaram-se doentes entre 5 casos). Num corte de tempo, poder-se-ia, por exemplo, examinar o mês 3 e concluir que a prevalência de doença é 20% (casos 2 e 7 estão doentes entre os 10 casos) e que a incidência é 10% (caso 2 torna-se doente no mês 3).

Numa abstração teórica de regularidade de ocorrência de doença (sempre a mesma incidência numa dada unidade de tempo e sempre mesma duração), prevalência e incidência obedecem uma relação regulada apenas pelo tempo de duração da doença:

PREVALÊNCIA = INCIDÊNCIA x DURAÇÃO DA DOENÇA

No exemplo, como a premissa de regularidade de ocorrência é rompida (a incidência não é a mesma a cada mês, a unidade de tempo) esta relação não se aplica (a prevalência de 80% não é igual a incidência (70%) vezes a duração da doença – um mês). No entanto, numa amostra grande espera-se que as variações de entrada e saída de observações, bem como de variações de duração de doença, tendam a anular-se fazendo valer a relação entre as duas medidas de ocorrência de doença.

Note-se que estas medidas estão envolvendo uma razão entre **EVENTOS / EXPOSTOS NUM DADO PERÍODO** (um dado mês, todos os meses juntos), sendo evento **ESTAR DOENTE** ou **TORNAR-SE DOENTE**. Esta razão pode ser modificada para incorporação de variações de tempo assumindo a forma **EVENTOS / EXPOSTOS-TEMPO**, vg. doentes/ pessoas-ano. Nesta situação, não se fixa uma data de início e fim para a observação, mas contam-se os eventos registrados durante a observação de um dado número de pessoas por períodos variáveis de tempo. Esta forma de medir ocorrência de doença é chamada **DENSIDADE DE INCIDÊNCIA** ou **FORÇA DE**

MORBIDADE e é muito útil em estudos que acompanham grupos de pessoas durante um período em que sua composição sofre alterações naturais de entrada e saída de indivíduos.

A partir dos conceitos básicos de prevalência e incidência, o pesquisador pode ter várias alternativas para definição de numerador e denominador da razão que mede a ocorrência de doença. Deve, no entanto, estar alerta que para inferir a freqüência de ocorrência de uma doença na população a partir de uma amostra estudada, ele deve considerar correções desta medida para a sensibilidade e especificidade do seu instrumento de medida (questionário, resultado laboratorial, etc.). Para obter esta correção, ele deve aplicar a seguinte fórmula:

$$\text{Freqüência na população} = \frac{Especificidade + (\,frequência\ \ da\ amostra\ -\ 1)}{Especificidade + (Sensibilidade - 1)}$$

2. QUAL A DIFERENÇA ENTRE RISCO RELATIVO E ODDS RATIO?

Risco relativo e odds ratio são medidas de associação entre variáveis que comparam ocorrências de eventos. O **risco relativo** mede quantas vezes a **freqüência relativa** de um evento é maior numa e noutra situação. O **odds ratio** mede quantas vezes o **odds** de um evento é maior numa e noutra situação.

Portanto, para entender risco relativo e odds ratio há que se entender primeiro as medidas de ocorrência de eventos que cada um considera. **Freqüência relativa** é a razão entre o número de eventos observados e o total de observações realizadas, vg. se da observação de 50 pessoas encontra-se 10 eventos, sua freqüência relativa é 10 em 50, ou 1 em 5, ou 20% (20 em 100). **Odds** é a razão entre número de eventos observados e o número de eventos não observados (os 'sim' contra os 'não'), vg. no exemplo o odds seria 10 contra 40, ou 1 contra 4 – não se aplica a redução à base 100 para designação da medida como porcentagem. Odds é uma palavra inglesa para designar chances a favor versus contra e é usada na forma original por falta de um equivalente aceitável em português. Ainda em inglês, ambas as medidas são chamadas de **rate**, uma razão que descreve ritmo de ocorrência: x eventos a cada n observações, x eventos positivos a cada y eventos negativos. A

relação de quantas vezes uma medida é maior que outra é chamada **ratio** – em português, tanto **rate** quanto **ratio** são traduzidas por razão com algum sacrifício de conteúdo semântico.

Ambas as medidas de risco (risco relativo e odds ratio) são razões (**ratios**) de razões (**rates**) para expressar quantas vezes uma medida de ocorrência (**rate**) é maior numa situação, vg. A, quando comparada com outra, vg. B. O pesquisador usa uma ou outra medida na **dependência de como realiza sua observações**. Se sua observação é feita com uma população definida ou com uma amostra representativa desta população, ele tem denominadores que permitem cálculo de freqüências relativas: percentual de eventos na situação A e percentual na situação B – basta que ele conte na sua população ou amostra o número de situações A e B.

Se, no entanto, o pesquisador faz sua observação examinado a ocorrência de eventos num determinado número de indivíduos que dispõe para estudo, alguns na situação A e outros na situação B, ele não tem denominadores para o percentual de eventos em A ou B. Suponha-se que ele observe x eventos entre os indivíduos da situação A e x' entre os indivíduos da situação B: ele não consegue calcular o percentual de x ou x' porque ele não tem os totais para a situação A ou B, ele só tem alguns indivíduos A e alguns indivíduos de B que lhe estavam disponíveis. Nesta circunstância ele só pode calcular x contra não x e x' contra não x'. Portanto, ao invés de comparar as freqüências relativas ele vai comparar os odds, em outras palavras ao invés de calcular o risco relativo ele vai calcular o odds ratio.

Para uma população em *steady-state* (número de pessoas que entram compensa o número de pessoas que saem) a medida de odds ratio eqüivale à medida de risco relativo. Ambas as medidas são usadas para inferências de causa e efeito e servem a propósitos de investigação de etiologia ou história natural de doença. Para medir o eventual impacto que o controle de uma causa de doença pode ter sobre sua ocorrência, o pesquisador dever recorrer a outra medida, o risco atribuível.

3. QUAL A DIFERENÇA ENTRE ASSOCIAÇÃO E CORRELAÇÃO?

Dois eventos se dizem associados quando suas ocorrências variam concomitantemente, sugerindo dependência entre eles. Por exemplo, o trovão e o relâmpago são eventos associados, sempre que um está presente espera-se pela presença do outro.

Para estabelecer-se uma associação verifica-se se as variações de ocorrência dos eventos estudados não pode ser atribuída ao simples acaso. Como nem sempre trovão e relâmpago aparecem juntos, resta uma pergunta se o número de vezes em que se *associam* não pode ser aleatória. Para responder a esta questão busca-se comparar o que se espera (probabilidade) de ocorrências combinadas por força do acaso contra o que se tem de evidências de ocorrências combinadas (o que se observa). Para analisar a associação entre trovão e relâmpago, poder-se-ia registrar suas ocorrências numa tabela de contingência como a seguinte:

		Relâmpago		
		Presente	Ausente	Total
Trovão	Presente	A	b	a + b
	Ausente	C	d	c + d
	Total	a + c	b + d	a+b+c+d

A probabilidade de ocorrência de trovão e relâmpago simultaneamente é o produto da probabilidade de trovão e da probabilidade de relâmpago, e para cada combinação de categorias presente / ausente pode-se fazer cálculo semelhante e identificar-se o número de ocorrências que se espera por simples acaso. Este número de ocorrências esperado pode então ser contrastado com o número de ocorrências observado de forma a ajuizar-se se as coisas estão acontecendo por acaso ou se há dependência entre os eventos. O teste do Qui-quadrado é um teste que compara as ocorrências esperadas e observadas examinando a hipótese de aleatoriedade entre os eventos – se seu resultado é significante (examina-se o valor de *p*) deve-se rejeitar a hipótese de aleatoriedade e concluir que há associação. Uma tabela de contingência pode considerar quaisquer outras formas de categorização dos eventos e a associação entre os eventos pode continuar sendo examinada pelo teste do Qui-quadrado.

A correlação é uma medida específica de associação que permite refinar a informação sobre a associação, quantificando-a. Este refinamento só será possível se houver refinamento das medidas relativas aos eventos: para além de categorias nominais há que haver alguma relação de intensidade ou ordem nas medidas. Em outras palavras os eventos devem poder ser medidos por variáveis quantitativas, ou pelo menos qualitativas ordinais, que permitam examinar a relação entre elas por uma reta de regressão.

A correlação é medida por um coeficiente. O coeficiente de correlação (de Pearson), normalmente designado *r*, é calculado de forma semelhante ao Qui-quadrado, comparando-se valores esperados a observados. A diferença é que o valor esperado considerado para o cálculo do coeficiente de correlação é derivado não de probabilidades de ocorrência, mas de projeções de uma reta de regressão. As razão entre as variações conjuntas dos dois eventos pela suas variações individuais (*product-moment*) resulta no valor 1 se as covariações são perfeitamente idênticas às variações individuais, estabelecendo-se então uma correlação perfeita: qualquer que seja a medida de um evento posso predizer a medida do outro. Quando as variações de medidas entre os dois eventos ocorrerem em sentido contrário (quando um aumenta, o outro diminui), sendo as variações conjuntas equivalentes às variações individuais, o coeficiente ainda terá valor 1, porém com sinal negativo. Daí, o coeficiente de correlação pode variar entre valores de $^{-}1$ a 1, os dois polos de correlação perfeita negativa e positiva, passando pelo valor zero, que significa nenhuma correlação entre os eventos.

A correlação é, portanto, uma medida de associação que expressa sua intensidade, de nula (0) a máxima ($^{-}1$,1). Não se deve confundir o valor de *r* com o valor de *p*: uma vez obtido um coeficiente de correlação, o pesquisador pode se perguntar se aquele valor é confiável, se não seria outro se ele refizesse suas observações. Para responder a esta inquietação, o pesquisador pode fazer um teste de hipótese para verificar se o valor que obteve não poderia ser zero, ou seja, relativo a associação nula. É este teste de hipótese que gera um valor de *p* que normalmente aparece associado ao valor

de *r*. Este valor de *p*, se significante, apenas informa que o valor calculado para *r* é confiável, se não poderia ser igual a zero: é errado à vista de, vg. um $r = 0.20$ com um *p* de 0,01, concluir que os eventos sejam correlacionados – o valor de *r* informa que a correlação é muito pobre (0,2 está muito longe de ⁻1 ou 1) e o valor de *p*, por ser significante, informa que este resultado é confiável.

4. O QUE SIGNIFICA O PODER DE UM TESTE?

Antes de falarmos em "poder de teste" é necessário introduzir alguns conceitos sobre **testes de hipóteses**. Sob o ponto de vista científico, as hipóteses podem ser entendidas como questões levantadas relacionadas ao problema em estudo e que, se respondidas, podem ajudar a solucioná-lo. Uma vez formuladas as hipóteses, estas devem ser comprovadas ou não com o auxílio de testes estatísticos. Os testes estatísticos têm como objetivo fornecer ferramentas que nos permitam validar ou refutar uma hipótese através dos resultados de uma amostra.

Em um teste estatístico duas hipóteses devem ser especificadas: a **hipótese nula (H_0)** e a **hipótese alternativa (Ha)**. A ***hipótese nula*** é aquela que é colocada a prova, ou seja, é aquela que desejamos contestar. Por exemplo, em um estudo com o objetivo de investigar a associação entre diabetes e doença cardiovascular, a hipótese nula poderia ser: "não existe associação entre diabetes e doença cardiovascular" ou, em outras palavras, "a prevalência de doença cardiovascular é semelhante entre diabéticos e não diabéticos". A ***hipótese alternativa*** é aquela que será considerada como aceitável, caso a hipótese nula seja rejeitada. No exemplo dado, a hipótese alternativa poderia ser: "a prevalência de doença cardiovascular é maior entre diabéticos".

Formuladas as hipóteses, o pesquisador cria uma **regra de decisão** baseada no nível descritivo obtido (valor de p). Qualquer que seja a decisão tomada, sempre existem erros. A decisão de rejeitar a hipótese nula quando ela é verdadeira é chamada de **erro do tipo I**. A probabilidade de cometer esse erro é chamada de **nível de significância** e é usualmente representada pela letra grega

alfa (α). Já a decisão de aceitar a hipótese nula quando ela é falsa é denominada **erro do tipo II** e é representado por β.

A capacidade de um teste identificar diferenças ou associações que realmente existem, ou seja, de rejeitar H_0 quando ela é realmente falsa, é denominada **poder do teste** e é definida como 1 - β.

Tabela: Erros possíveis associados ao teste de hipótese

DECISÃO	SITUAÇÃO REAL	
DO TESTE	H_0 verdadeira	H_0 falsa
Não rejeitar H_0 (aceitar Ha)	Decisão correta	**Erro do tipo II**
Rejeitar H_0	**Erro do tipo I**	Decisão correta **(poder)**

No exemplo da diabetes, o erro do tipo I seria afirmar que a prevalência de doença é maior entre os diabéticos quando na realidade ela é igual à dos não diabéticos. O erro do tipo II seria afirmar que as prevalências são iguais quando na verdade a dos diabéticos é maior. Já o poder do teste seria concluir ***corretamente*** que a prevalência entre os diabéticos é maior.

5. NÍVEL DE SIGNIFICÂNCIA E "VALOR DE P" SÃO A MESMA COISA?

NÃO. Esta confusão existe uma vez que é muito comum o uso da palavra "significância" para expressar o resultado de um teste baseado no valor de p. Por exemplo, quando um valor de p é significativo, o pesquisador tende a dizer que o resultado tem "significância estatística" e isto faz com que ele chame erroneamente o valor de p de nível de significância.

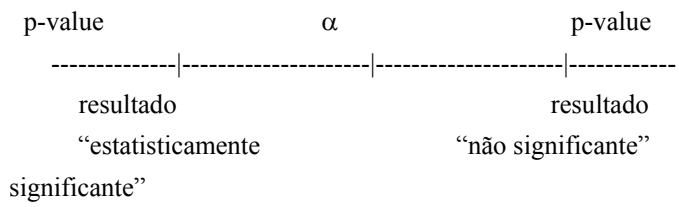
Para melhor entender melhor as diferenças entre nível de significância e valor de p, voltemos às definições. Em um teste estatístico de hipóteses, sempre existe uma hipótese a ser contestada que é a chamada ***hipótese nula***. O **nível de significância** corresponde ao **erro** associado à rejeição hipótese nula, conhecido como **erro do tipo I** (rejeitar a hipótese nula quando esta for verdadeira). Geralmente é expresso pela letra grega α e os valores usualmente adotados são 5%, 1% e 0,1%.

Portanto, o nível de significância é um valor **fixado previamente** pelo pesquisador e sua magnitude vai depender do risco que o pesquisador deseja assumir ao tomar uma decisão incorreta.

Considerando que os pesquisadores ao rejeitarem a hipótese nula costumam dizer que existe "significância estatística" poderíamos definir o **valor de p** como a "probabilidade mínima de erro ao concluir que existe significância estatística" ou seja, "o menor nível de significância (α) que pode ser assumido para se rejeitar H_0". Este valor é **calculado** com base nos dados observados e o pesquisador deve decidir se o valor de p produzido é pequeno o suficiente para afirmar com segurança que o resultado é "estatisticamente significante".

A tomada de decisão consiste em avaliar se o erro calculado está dentro de uma margem de erro tolerável e é sempre baseada na comparação entre os dois valores: o valor de p e o nível de significância. Se o valor do p for menor que o nível de significância (α) deve-se concluir que o resultado é significante pois o erro está dentro do limite fixado. Por outro lado, se o valor de p for superior à α significa que o menor erro que podemos estar cometendo ainda é maior do que o erro máximo permitido, o que nos levaria a concluir que o resultado é não significante pois o risco de uma conclusão errada estaria acima do que se deseja assumir. Segue abaixo um esquema que resume a regra de decisão descrita.

p-value $< \alpha \Rightarrow$ rejeito $H_0 \Rightarrow$ diferença
p-value $> \alpha \Rightarrow$ não rejeito $H_0 \Rightarrow$ igualdade



Em resumo, nível de significância e valor de p embora estejam relacionados são valores distintos. O nível de significância (α) é um valor arbitrado previamente pelo pesquisador, enquanto que o valor de p é calculado de acordo com os dados obtidos e deve ser comparado ao nível de significância fixado para tomada de decisões.

6. QUAL DEVE SER O TAMANHO DA AMOSTRA?

Muitas vezes no delineamento do estudo, o pesquisador se vê obrigado a perguntar "quantos pacientes devo observar?". Esta é uma pergunta muito freqüente em estudos da área médica, porém nem sempre a resposta é única e objetiva. Arbitrar um tamanho adequado de amostra envolve conhecimento da natureza das medidas realizadas, do plano de análise, do nível de erro aceitável para estimativas e etc.

Mesmo com todas essas informações, o tamanho da amostra vai depender também da **viabilidade** de coleta de dados, que envolve principalmente tempo, custos e disponibilidade de casos para serem estudados. Isto não significa que o cálculo de tamanho de amostra seja dispensável. O que desejamos salientar aqui é que ele deve ser utilizado como **planejamento**, isto é, como parte de um estudo bem delineado. Vale ressaltar também que os vícios de seleção ou de informação não serão prevenidos por qualquer definição de tamanho de amostra, mas sim por um plano amostral cuidadoso.

Para o planejamento do tamanho da amostra o investigador precisa estabelecer algumas definições como:

- tipo de estudo que pretende realizar (ex: estudo de prevalência, ensaio clínico, coorte, caso-controle)
- o tipo de medida que deve utilizar (ex: medidas contínuas, categorizadas, prevalência, incidência)
- o tipo de análise (ex:. diferenças entre médias, diferença entre proporções, cálculo de risco)
- a margem de erro que pode assumir para o estudo (ex. o nível de significância e o poder do teste estatístico que pretende aplicar).

Estes conceitos podem ser melhor esclarecidos na *homepage* do Laboratório de Epidemiologia e Estatística (www.lee.dante.br) que apresenta um serviço que calcula tamanhos de

amostra para alguns dos desenhos de pesquisa médica/biológica mais freqüentes, além de oferecer textos de apoio para compreensão de cada item envolvido no cálculo e referências bibliográficas.

7. O QUE SÃO VÍRUS DE COMPUTADOR E COMO POSSO ME PROTEGER?

Os vírus de computador são pequenos programas que executam tarefas sem que o usuário tenha solicitado. Por exemplo, um vírus pode ser programado para apagar arquivos do computador sem o conhecimento do usuário. Inicialmente os vírus se anexavam somente aos arquivos de programas (arquivos com extensão .EXE e .COM) ou então no setor de inicialização do disco rígido ou do disco flexível (setor de *boot* do disco). Com o aparecimento de editores de texto e leitores de e-mails que permitem o uso de linguagem de programação (macros), surgiu um novo tipo de vírus que contamina arquivos de documentos, planilhas eletrônicas e e-mails. O caso mais recente é o vírus **ILOVEYOU** que foi amplamente noticiado pela mídia.

Um vírus de computador não aparece sozinho. Sua propagação é semelhante aos vírus biológicos, ou seja, é necessário haver troca de informação entre um computador saudável e outro que esteja contaminado. A troca de informação pode ser feita por meio de mensagens eletrônicas ou disquetes com arquivos. Um vírus somente contaminará o computador se abrirmos estes arquivos. É possível ter arquivos com vírus sem que o mesmo contamine o computador, isto é, podemos ter um arquivo do Word infectado por vírus, mas se não abrirmos o arquivo não existirá contaminação. Não existe antivírus totalmente eficiente, algumas REGRAS BÁSICAS podem ajudar a proteger o computador contra esses programas mal intencionados.

1- Nunca execute programas ou abra arquivos que não foram solicitados, o fato do e-mail vir de um amigo não é garantia de que ele redigiu a mensagem ou então que o arquivo esteja livre de qualquer código estranho. Atualmente a maior forma de disseminação de vírus são os documentos e cartões animados que são anexados aos e-mails. Ser criterioso na leitura dos anexos em mensagens é muito importante para não ter o equipamento comprometido por vírus.

2- Pense antes de abrir qualquer link que veio em um e-mail. Existem sites mal intencionados que querem descobrir informações privadas no seu computador. Recuse os links de mensagens de propagandas não solicitadas.

3- Desabilite o *boot* pelo disquete. Ao ligar seu computador, verifique que não existe disquete no driver de disco flexível. O *boot* acidental por disquete pode contaminar seu computador com vírus.

4- Tenha sempre um antivírus instalado e atualizado no computador. Muitos fabricantes de antivírus lançam definições de vírus a cada 15 dias ou menos. Um antivírus de última versão não serve para nada se definição de vírus estiver desatualizada, para que ele seja eficiente é necessário atualizá-lo constantemente.

5- Mantenha seus programas atualizados. Os fabricantes dos programas freqüentemente publicam atualizações dos seus produtos, por exemplo, o site http://officeupdate.microsoft.com/brasil tem várias informações sobre atualizações necessárias do Microsoft Office. Outro site útil é o http://windowsupdate.microsoft.com com atualizações do Windows e também do Internet Explorer.

6- Sempre desative as macros nos documentos do Word caso não conheça a origem do documento. O Word 97 mostra uma tela alertando o usuário que arquivo contém macro e pergunta se o documento será aberto com as macros ativa ou não. Verifique com o autor do documento se ele programou alguma macro antes de abrir o documento, a presença de macros em documentos é um forte indício da presença de vírus.

7- Tenha várias cópias de segurança dos seus arquivos. Nunca acredite que vírus aparece somente nos computadores dos outros e que seu equipamento não falha. Imprevistos acontecem e a melhor segurança é ter cópias repetidas dos arquivos importantes.

8- Verifique qualquer comportamento estranho no computador: demora na execução de programas, travamentos constantes e avisos de falta de memória que antes não existiam. Alguns vírus tornam os programas extremamente lentos, consumindo uma grande quantidade de memória.

9- Se tiver o computador comprometido por vírus, a melhor coisa a fazer é não entrar em pânico. Um procedimento mal planejado pode comprometer todo seu equipamento, muitas vezes sendo necessário reinstalar todos os programas e o sistema operacional. Tente limpar usando um antivírus atualizado.

Sites úteis sobre vírus e programas antivírus:

- http://www.splitnet.com
- http://www.symantec.com.br
- http://www.nai.com.br

8. COMO ORGANIZAR UM BANCO DE DADOS COM UMA PLANILHA DE CÁLCULO?

Em geral quando falamos em banco de dados pensamos em programas complexos criados por programadores. Atualmente é possível manter um banco de dados usando somente uma planilha de cálculo como por exemplo o Microsoft Excel.

Para inserir os dados em uma planilha do Excel, as variáveis de interesse devem estar dispostas em colunas e os indivíduos em linhas, sendo que cada linha corresponde à um único indivíduo, ou seja, dados de um mesmo indivíduo não devem estar em linhas diferentes.

	A	B	C	D	E	F
1	**nome**	**sexo**	**idade**	**peso**	**Altura**	**fumante**
2	José Pedro	0	26	80	1,68	1
3	...	...	...	...	...	...

Deve-se evitar o uso de palavras no preenchimento das células do banco de dados pois o computador interpreta de forma distinta letras maiúsculas e minúsculas, por exemplo, *Sim* e *sim* são duas palavras distintas para o computador embora tenham o mesmo significado. Para facilitar a análise dos dados e eliminar erros de digitação é preferível que se use códigos para indicar o valor de uma variável. No exemplo acima podemos usar 0 = *masculino* e 1 = *feminino*, 0 = *não* e 1 = *sim*. Como podemos ter um grande número de variáveis e códigos, é interessante manter uma lista com uma descrição das variáveis e seus respectivos códigos usados no banco de dados. Esta lista é também conhecida como ***code book,*** segue abaixo um exemplo:

Variável	**Descrição**	**Tipo de medida**
TABAG	Tabagismo	0 = não fumante 1 = ex fumante 2 = fumante atual
PESO	Peso do paciente em kg	quantitativa

Uma confusão muito freqüente é o uso do Word para se manter uma base de dados. Geralmente os dados são dispostos na forma de colunas dentro de um texto. Os dados desta forma não são facilmente transportados para pacotes estatísticos como o SPSS.

Nos casos em que há a necessidade de checar a consistência dos dados no momento da entrada (digitação), é comum o uso de programas desenvolvidos em Acccess ou alguma linguagem de programação de alto nível como o Visual Basic ou C. Exceto nesses casos especiais, o Excel é bastante satisfatório para se criar uma base de dados.

9. COMO USAR A INTERNET PARA PESQUISA?

A internet tem a filosofia de manter informações distribuídas pela rede. O grande problema é encontrar estas informações nos milhões de sites que existem. Para facilitar a busca, existem sites que foram desenvolvidos com o propósito de serem *listas amarelas* da internet. Existem vários sites de busca onde colocamos uma palavra chave e o site lista as páginas que possam conter alguma informação relacionada à palavra. Alguns sites de busca são:

- http://www.cade.com.br
- http://www.radaruol.com.br
- http://www.tay.com.br
- http://www.altavista.com.br

Se o interesse da busca for por algum artigo científico, o melhor lugar para se procurar é no site da empresa que publica a revista, por exemplo, o JAMA (*Journal of the American Medical Association*) pode ser acessado pelo site http://jama.ama-assn.org . Em alguns sites é necessário que o leitor faça uma assinatura para ter acesso ao texto na íntegra; outros oferecem o texto gratuitamente, o usuário somente precisa preencher um cadastro. Este é o caso do *American Heart Journal* (http://www.medscape.com/mosby/AmHeartJ/public/journal.AmHeartJ.html). Se a biblioteca da instituição de pesquisa possui a assinatura da revista em papel, basta a bibliotecária

criar um cadastro institucional no site da revista para que a biblioteca possa obter o artigo na íntegra.

O Lee (Laboratório de Epidemiologia e Estatística – IDPC: http://www.lee.dante.br) oferece o acesso a vários textos completos publicado pela Elsevier, Academic Press e High Wire Press. Este serviço é oferecido pela FAPESP aos centros de pesquisa do estado de São Paulo. O acesso a esta biblioteca virtual somente é permitida através da rede do Lee ou por uma instituição cadastrada no Probe (Programa Biblioteca Eletrônica).

Em geral os textos completos são oferecidos no formato eletrônico PDF que exige que o usuário tenha o programa Acrobat Reader para a sua leitura. O Acrobat Reader pode ser obtido gratuitamente no site http://www.adobe.com.