Conteúdo

Parte I

Revisão dos conceitos básicos de Bioestatística

Capítulo 1

Variáveis aleatórias e sua classificação

1.1 Introdução

O conceito de variável aleatória é fundamental para o desenvolvimento de toda a análise. A sua compreensão torna-se decisiva para a correcta interpretação dos fenómenos por si "modelizados" entendendo por esta afirmação a expressão, em termos analíticos, dos fenómenos naturais que elas permitem caracterizar.

Nas secções seguintes iremos proceder á sua definição e apresentaremos um modelo de classificação que nos pareceparticularmente adaptado à utilização epidemiológica que

iremos fazer do conceito.

De seguida apresentaremos os seus atributos essenciais.

Definição de variável aleatória

Analisemos a seguinte experiência.

Procedemos a uma tiragem em quatro mãos, de seis cartas cada, tendo obtido os resultados apresentados na Tabela 1.1

Tabela 1.1 Extracção de cartas

1ª Mão	2ª Mão	3ª Mão	4ª Mão
Ás de copas	2 de ouros	Dama de paus	Rei de ouros
10 de Espadas	Rei de Paus	2 de ouros	2 de paus
2 de ouros	Ás de copas	9 de paus	10 de copas
5 de paus	7 de paus	Valete de espadas	Rei de paus
Valete de ouros	10 de copas	5 copas	Dama de copas
Rei de espadas	9 de copas	Ás de paus	7 de espadas

Decidimos agora que á saída de qualquer uma das cartas do naipe de espadas atribuíamos o valor 1; igualmente para o naipe de copas o valor 2, para o de ouros o valor 3 e para o de paus o valor 4.

A experiência fica assim sistematizada de acordo com a Tabela 1.2

Tabela 1.2 Extracção de cartas organizada

1ª Mão	2ª Mão	3ª Mão	4ª Mão
2	3	1	3
1	4	3	4
3	2	4	2
4	4	1	4
3	2	2	2
1	2	4	1

Vemos que através deste pequeno "truque" passámos a fazer corresponder a cada uma das cartas extraídas (espadas, copas, ouros ou paus) um número real (1,2,3 e 4).

A regra que permite associar a cada valor do nosso espaço de resultados um e só um valor no conjunto dos números reais é justamente uma variável aleatória.

Podemos assim dizer de uma forma elementar que:

Variável aleatória é uma variável cujo valor é determinado por uma experiência aleatória

Desta forma poderemos formalizar o problema anterior dizendo que a saída de uma carta de naipe de um baralho é representada por uma variável aleatória X que assumirá os seguintes valores da Tabela 1.3

Tabela 1.3 Variável Aleatória

x=1 se a carta for de espadas x=2 se a carta for de ouros

x=3 se a carta for de copas x=4 se a carta for de paus

A generalização deste exemplo leva-nos a compreender que a maioria dos fenómenos da natureza pode assim ser representados por variáveis aleatórias, sejam elas a distribuição do peso ou da altura dos médicos do Centro de Saúde, a distribuição da glicemia nos utentes da consulta, ou de qualquer outra população, a chegada de doentes á urgência, ou ao SAP , ou o aparecimento de neoplasia brônquica nos trabalhadores da indústria cerâmica.

Evidente se torna que, se é verdade que todos estes fenómenos são representados por variáveis aleatórias, também é verdade que o simples enunciado das situações acima referidas demonstra que as variáveis em questão devem ser diferentes umas das outras.

Iremos de seguida proceder, primeiro à sua classificação, e de seguida, apresentar os parâmetros que as definem.

Classificação das variáveis aleatórias

A classificação das variáveis aleatórias pode ser encaradas de diversas formas.

Seguiremos, por nos parecer bastante adequado ao nosso objectivo, o critério sugerido por Kleinbaum (1998).

Assim as variáveis aleatórias poderão ser

classificadas de acordo com o intervalo, a orientação descritiva e os níveis de mensuração.

Classificação segundo o intervalo

Conforme se verifiquem, ou não, intervalos entre os valores observados de uma variável, esta será classificada como discreta ou contínua.

De uma forma mais precisa diremos que uma variável é discreta se entre dois valores observados existir um valor que não é possível observar.

Uma variável é contínua se entre dois valores observados existir sempre um valor potencialmente observável.

Variáveis aleatórias discretas são as que representam, no exemplo acima, a chegada de doentes á urgência ou ao SAP, ou o aparecimento de neoplasia brônquica nos trabalhadores da indústria cerâmica.

Variáveis contínuas são as que representam a distribuição do peso, altura ou da glicémia dos doentes da consultaou de qualquer outra população.

A Figura 1.1 apresenta um gráfico representativo de uma variável discreta, assim como a

Figura 1.1 Variável discreta

A Figura 1.1 apresenta um gráfico representativo de uma variável discreta, assim como a Figura 1.2 um histograma que representa uma variável contínua.

Voltaremos á representação gráfica das variáveis com mais pormenor, mais á frente quando tratarmos dos métodos gráficos em análise de dados.

Refira-se aqui que é frequente em análise de dados agrupar em blocos contíguos valores de variáveis discretastratando-as assim como se de variáveis contínuas.se tratassem.

É contudo bem mais frequente, sobretudo em epidemiologia, dividir as variáveis contínuas em classes e posteriormente encará-las como se de variáveis discretas se tratassem.

O exemplo mais flagrante é o da recodificação da idade em grupos etários, de todos bem conhecida; outro exemplo, bastante comum, é o de classificar em graus de risco variáveis contínuas como o índice de massa corporal, transformando assim em "Muito elevado", "Elevado", "Normal".

Esta questão da chamada "optimização das variáveis" que engloba entre outras a construção de escalas de classificação, para organização de inquéritos, etc. é bastante importante e encontra-se particularmente bem estudada em várias obras como a de Abrhamsom (1979).

Merece contudo uma referência especial, dedicada principalmente a todos aqueles que tenham necessidade prática de fazer estudos que envolvam questionários, inquéritos, etc., a obra "Inquéritos de Saúde na Comunidade" editada pela Oficina Pan-Americana da Saúde (1986).

Embora já com alguns anos, não deixa de ser um ponto de paragem quase obrigatório para esta área, até pela sua clareza e sentido

Figura 1.2 Variavel continua

prático.

Classificação segundo a orientação descritiva

Uma outra forma de descrever as variáveis é baseada no facto de elas descreverem ou poderem ser descritas por outras variáveis. Esta classificação tem mais a ver com os objectivos do estudo, do que com a subjacente estrutura matemática da variável, em si (KIeinbaum, op. citada).

Se a variável em estudo se apresentar de forma que, sozinha ou em conjunto com, outras, explique o comportamento do outras variáveis dizemos que se trata de uma variável independente ou de um predictor.

Se por outro, lado a variável for ela mesma explicada por uma ou mais variáveis independentes ou predictores será uma variável dependente.

Nos estudos epidemiológicos assumem grande importância as variáveis, que embora não pertencendo intrinsecamente ao estudo, isto é, não sendo predictores, interferem nos resultados da variável dependente. Estas variáveis são habitualmente chamadas do covariatos ou variáveis de confundimento.

A sua importância será analisada mais adiante.

Um bom exemplo destas variáveis é apresentado na Tabela 1.4

Tabela 1.4 Variação da pressão arterial face a vários factores

Var P. A (em mm Hg)	Sexo	Idade	Peso	Cons. Tabaco (Maços)
1.0	M	23	75	0
-2.8	F	54	86	0
1.2	M	76	90	1
-0.5	F	46	65	2

A variação da pressão arterial é apresentada como variável dependente do factores como o sexo, a idade, o peso, e o consumo do tabaco, que são as variáveis independentes. Como variáveis de confundimento poderão ser consideradas a ingestão do contraceptivos, a existência de patologias concomitantes, etc.

Classificação segundo o nível do mensuração

Este esquema de classificação lida com a precisão do medição da variável.

Consideraremos três níveis: nominal, ordinal e de intervalo. Uma variável é nominal quando assume valores que simplesmente apontam para categorias diferentes.

No caso, da variável sexo esta assumirá os valores 1 e 0 conforme se trate do sexo masculino ou feminino. A este nível não se considera, portanto nenhuma indicação que reflicta qualquer grau do dependência, de ordenação ou de hierarquia.

0 exemplo da tiragem das cartas, referido no inicio do capitulo, é um exemplo de uma variável ordinal, uma vez que não existe nenhuma hierarquia entre OS vários naipes do baralho.

Para uma variável ordinal já se pressupõe alguma ordenação entre as categorias.

Uma variável que represente uma resposta a um estimulo será deste tipo se forem estabelecidas categorias do tipo, "fraco" "médio" "forte". No caso do, para a16m do sentido ordinal, ser possível graduar a distância entre as categorias falamos do uma variável do intervalo.

Estas variáveis devem ser expressas em termos do uma certa unidade de medida física, ou globalmente aceite. Um bom exemplo do uma variável do intervalo será a que represente a distribuição do peso, da altura. da pressão arterial, etc.

Sobreposição do critérios

Sobressai do tudo o que acima se disse que os vários critérios de classificação se entrecruzam entre si e, que portanto, somente em condições excepcionais, ou de mero interesse explicativo ou académico as coisas se podem encarar do uma forma estanque.

Assim é óbvio que se uma variável tem do ser discreta para ser nominal, também é verdade que uma variável discreta pode ser nominal, ordinal ou do intervalo.

Igualmente uma variável continua pode ser ordinal ou do intervalo, sendo que existem variáveis ordinais e nominais que não são continuas.

Veja-se, o exemplo da idade.

Habitualmente é considerada como uma variável do intervalo e continua, contudo se for recodificada em grupos etários (grupo 1, grupo 2, grupo 3, etc. ..) será tratada como uma variável discreta e nominal. Unia ideia conviria ficar assente ao fechar esta parte.

Em Ciências Sociais c Humanas, entre as quais a Medicina se engloba, a variabilidade e a interacção entre os vários fenómenos são uma constante.

Só por razões metodológicas, ou pedagógicas, se recorrem a simplificações que na prática não correspondem a nenhum facto real.

Por essa mesma razão o uso da Estatística Multivariada, do que estudaremos alguns aspectos ao longo deste manual, assume uma importância fundamental no nosso dia a dia.

As classificações e as sistematizações, têm aqui e noutros ramos como na própria Medicina Clínica, um sentido clarificador, ajudando a educar o espírito ea preparar verdadeiro racciocinio epidemiológico indispensável a quem queira singrar na senda da

investigação em saúde.

Não obrigam a Natureza a comportar-se conforme pretendemos. Teremos do "continuar a aprender com ela e a utilizar os nossos meios do investigação, cada vez mais poderosos, para irmos mais longe na descoberta das suas ainda longínquas profundezas".

Variáveis Aleatórias e sua caracterização

Distribuição Probabilistica e Parâmetros.

0 termo variável aleatória é utilizado para indicar uma variável cujos valores observados se podem considerar como resultado de uma experiência estocástica ou aleatória. (Kleinbaum, op cit).

Uma variável aleatória assume assim valores numa população os quais só podem em regra ser materializados através da colheita de informação realizada sobre amostras æ entendidas como subgrupos do grupo mais dilatado que é a população æ.

Não é assim possível conhecer, na pratica, os valores que a variável assume sem que se procedam. a experiências a partir das quais se podem inferir os mesmos.

Contudo, recordemos que as variáveis aleatórias têm um comportamento probabilístico que convém à partida estudar.

Fundamentalmente podemos afirmar que, sem entrar em maiores considerações de índole matemática, uma variável aleatória fica caracterizada desde que se conheçam a sua distribuição probabilistica (função de probabilidade e função de distribuição) e os respectivos parâmetros (valor esperado e variância).

Consideremos separadamente os casos discreto e continuo.

Seja então X uma variável aleatória que assume valores probabilísticos para valores (discretos) de x

A distribuição probabilística de uma variável aleatória discreta lista todos os valores possíveis que a variável pode assumir e as correspondentes probabilidades

A Figura 1.3 apresenta um gráfico da distribuição de uma variável aleatória discreta.

Como se pode observar a cada valor lido no eixo dos xx (abcissas) corresponde um valor de probabilidade çido no eixo dos yy (ordenadas)

A distribuição probabilistica de uma variável aleatória possui duas características que convém absolutamente reter

Figura 1.3 Distribuição probabilística de uma variável discreta

· A soma das probabilidades atribuídas a todos os valores possíveis de X é igual a 1, isto é

· A probabilidade atribuida a cada valor da variável aleatória X está contido entre 0 e 1, isto é $0<P(x)<1$ para cada valor de x.

Estas duas propriedades são essenciais e derivam do próprio conceito de probabilidade. De facto não pode ser esquecido que a probabilidade de um evento acontecer é 0 ou 1(pele branca ou negra, masculino ou feminino, morto ou vivo, etc.) e que não existe probabilidade negativa nem probabilidade superior á unidade.

Se por vezes em linguagem comum se, fala do 200 ou 300% de probabilidade do um fenómeno acontecer trata-se de uma figura puramente retórica, a qual não tem qualquer sentido estatístico.

0 exemplo seguinte, extraído de Mann 1994), ilustra o conceito de distribuição probabilistica.

A Tabela 1.5 mostra a frequência e a frequência relativa do número de veículos possuídos por 200 famílias.

Tabela 1.5 Nº de viaturas por família

Nº de viaturas	Frequência	Freq. Relativa	Probabilidade P(X)
0	30	0.15	0.15
1	47	0.235	0.235
2	85	0.425	.425
3	29	0.145	.145
4	9	0.05	0.05
	N=200		P(X)=1.00

Como podemos observar a frequência relativa (n.º de acontecimentos com o atributo/total de acontecimentos) obtidas de uma experiência aleatória pode ser utilizado como probabilidade aproximada.

Se as frequências relativas forem conhecidas para a população em estudo então elas dão a probabilidade real da distribuição; da podemos agora extrair alguma informação útil.

Imaginemos que pretendemos saber qual a probabilidade do uma família ter 2 carros, ou 3 ou mais carros.

De acordo com as propriedades atrás apresentadas para a probabilidade torna-se fácil verificar que por leitura directa do quadro,

Se agora pretendermos conhecer a probabilidade de uma família ter mais de 3 carros, temos:

Note-se que, o que pode ser confirmado visualmente através da Figura 1.3, que existem descontinuidades entre os valores (por ser uma variável discreta) que justificam "os saltos" entre os valores da probabilidade.

No caso de inclui-se o valor da probabilidade e toda a que está á sua direita .

No caso da já só se inclui a probabilidade á direita de 3, uma vez que a descontinuidade leva a que "salte" de um valor para o

seguinte.

Média, Esperança Matemática e Valor Esperado de uma variável aleatória

A média m de uma variável aleatória discreta, a que passaremos a denominar de ora em diante, por razões óbvias, de comodidade, variável discreta "tout-court", é o valor médio que esperamos encontrar se repetirmos uma experiência aleatória um grande número de vezes.

Por esse facto é igualmente costume designar-se (designação fundamental em estatistica matemática) por valor esperado ou esperança matemática.

A ilustração deste conceito pode ser feita com o clássico exemplo da lotaria.

Exemplo: Uma lotaria contem 1515 bilhetes com a seguinte distribuição de prémios:

Tabela 1.6 Valor esperado de uma variável

Prémio em $	0	100	1000	10000
Nº de bilhetes	1000	500	10	5
Freq. Rel	0,66	0,33	0,01	0,003

Pergunta: Qual o preço médio porque se devem vender os bilhetes?

A solução para o problema passa por calcular o valor esperado da variável X que modeliza a

saída dos prémios

Logo cada bilhete deveria ser vendido por 73$00, não considerando claro os aspectos de lucro da operação.

Daqui deriva o seguinte:

A Média de uma variável aleatória discreta X é o valor médio que é esperado obter por repetição de uma experiência efectuada um conjunto alargado de vezes

ou

Este parâmetro m é habitualmente apresentado como uma medida de tendência central uma vez que define o valor "à’ volta do qual" se irão distribuir todos os outros valores.

Este parâmetro m é habitualmente apresentado como uma medida de tendência central uma vez que define o valor "à’ volta do qual" se irão distribuir todos os outros valores.

Variância e Desvio Padrão de uma Variável Discreta

A variância de uma variável discreta, denomina-se s2

Mede a "dispersão, à volta da média".

Pode ser calculada da forma seguinte:

A raiz quadrada de s2 denomina-se desvio padrão.

A interpretação do desvio padrão, mais utilizado na pratica do que a variância, advém das propriedades decorrentes da aplicação do teorema de Tchebyshev que pode ser enunciado do uma forma simplista como:

"Pelo menos [1-(1/k*2)*100% da área total sob uma curva esta’ contida entre K desvios padrões da média. com k qualquer numero maior que 1".

Então se k=2, [1-1/2*2.)]*100%=75%.

Logo 75% da probabilidade estará contida entre 2 desvios padrões à roda da média da forma: e

0 teorema do Tchebyshev será recordado mais à frente e reforçado o seu uso pratico.

Exemplifiquemos do seguida o cálculo do desvio padrão.

Na Tabela 1.7 vemos os valores da variação do massa corporal com determinado medicamento estimulante do apetite;

Na 2ªcoluna apresentamos a respectiva probabilidade.

Tabela 1.7 Cálculo do desvio padrão

x	P(x)
4,5	.32
1,2	.51
-2,3	.17

Expandindo a Tabela 1.7 temos:

Tabela 1.8 Cálculo do desvio padrão expandido

x	P(x)	xP(x)	x2	x2P(x)
4,5	.32	1,440	20,25	6,480
1,2	.51	.612	1,44	.7344
-2,3	.17	.391	5,29	.8993

m=SxP(x)=1,661 Sx2P(x)=8,1137

Logo a média é de 1,661 quilogramas e o desvio padrão:

Assim espera-se que o peso tenha um, ganho médio do 1,6661 quilogramas com um desvio padrão do 2,314.

Convirá neste momento sossegar o leitor que já’ se vê mergulhado em fórmulas e algoritmos de cálculo que tal não irá acontecer.

Hoje em dia as calculadoras estatísticas e os computadores dão estes e outros resultados ao toque do um. dedo.

No final deste capítulo apresentaremos o "output" do STATA 6.0 em vários exemplos que deixarão a questão clara.

A apresentação destes cálculos tem um carácter informativo e serve para que se compreendam os meios que levam aos fins.

Uma vez entendido cada um decidirá do interesse, ou não, em ter este tipo do informação.

Distribuição Probabilistica, média e desvio padrão do uma variável continua.

0 raciocínio desenvolvido na secção anterior aplicasse às variáveis aleatórias do tipo continuo, daqui para a frente denominada pelas mesmas razões do comodidade simplesmente variáveis continuas.

Contudo convém fazer alguns reparos que atestam. as suas diferenças

principais.

Sendo variáveis continuas, por definição podem assumir todos os valores num dado intervalo. Por isso os "saltos" observados nas variáveis discretas não se verificam, aqui

A Figura 1.4 mostra um histograma ao qual se sobrepôs uma curva que demonstra o "aspecto" de uma distribuição probabilística continua, alias já referida anteriormente.

Por esta razão denomina-se a função de probabilidade de uma variável continua, função de densidade de probabilidade.

Por outro lado referisse que o cálculo das probabilidades numa área sob uma curva continua não pode ser efectuado através do somatório utilizado atrás, mas sim com o

Figura 1.4 Histograma de uma variável com distribuição normal

recurso a uma ferramenta matemática mais complexa que é o integral.

Assim embora se alterem, e compliquem, as regras de cálculo, todos os conceitos anteriormente definidos de distribuição probabilística, de média e de variância mantêm-se válidos.

Por ser mais difícil a formulação no campo teórico destes conceitos, na área das variáveis aleatórias continuas, guardaremos para quando estudarmos à frente, algumas variáveis continuas de grande interesse em estatística e epidemiologia a apresentação de exemplos concretos.

Para terminar esta secção apresentamos na Caixa 1 a formulação mais rigorosa dos conceitos aqui desenvolvidos.

Como se disse na introdução a leitura destas, e de outras caixas é absolutamente dispensável para o leitor que não sinta qualquer atracção pela estatística matemática.

A sua inclusão deve-se, tão sómente, à necessidade de proporcionar a outros leitores, se os houver, uma visão mais rigorosa da matéria.

Contudo a compreensão geral do corpo do texto não passa minimamente pela sua leitura.

O leitor decidirá.

Quem estiver interessado em aprofundar os conhecimentos nestes aspectos mais formais, poderá comsultar as obras de Tannis e Hoggs (1997) ou a de Fisher e van Belle (1993)

A primeira é um excelente livro de estatística matemática que nos servirá de referência em todos os aspectos com ela relacionados.

A segunda é um óptimo livro de Bioestatística que cobre as diferentes áreas com alguma profundidade e que servirá optimamente a quem pretenda ter uma obra de referência na matéria.

Não podemos deixar de referir, de entre outras, duas obras em língua portuguesa, as de Galvão de Melo (1971), mas de que existe uma edição recentee de Bento Murteira (1990).

São duas excelentes obras do dois grandes mestres da estatística portuguesa.

Previne-se contudo o leitor para o facto da sua consulta necessitar de alguma preparação previa, nomeadamente na área da matemática, além de algum esforço no seu estudo, devido ao nível que globalmente atingem.

Principais Variáveis Aleatórias de importância em Epidemiologia e Bioestatistica

Iremos de seguida apresentar. sumariamente. as características das principais variáveis de interesse e uso em Epidemiologia e Bioestatistica.

Começaremos pelas variáveis discretas onde abordaremos a Binomial e a Poisson, passando depois às variáveis continuas onde visitaremos as variáveis

Normal, T de Student, Qui Quadrado e F.

No final do capítulo apresentaremos na Caixa 2 as expressões analíticas, bem como os seus valores esperados.

Dispensamo-nos de dar grande ênfase à descrição das tabelas das várias distribuições, que aliás não incluiremos, não só porque este aspecto vem explicado em todos os livros de referência, mas também porque o uso das calculadoras e dos computadores evita essas consultas.

Preferimos mais enfatizar o sentido da sua interpretação do que insistir na técnica da consulta.

Contudo os Ieitores interessados poderão recorrer às obras indicadas, especialmente o livro do Man (op.cit.), que descreve minuciosamente essa operação.

Variáveis Aleatórias Discretas: Distribuição Binomial

Conforme, fizemos em casos anteriores, vamos por comodidade, denominar esta variável simplesmente por binomial.

Analisemos o exemplo seguinte:

Exemplo: Um clínico geral sabe, pelo estudo do seu ficheiro, que a probabilidade do encontrar um doente hipertenso é de 20%.

Feita uma amostra aleatória do seu ficheiro gostaria de saber qual a probabilidade de encontrar 3 doentes hipertensos em 5 fichas retiradas.

0 caso acima é o exemplo de uma experiência binomial. em que nos interessa conhecer o aparecimento æ sucesso æ ou não æ insucesso æ de um dado fenómeno.

Para que uma experiência se possa considerar binomial deverá preencher os seguintes requisitos:

· Existirem n tentativas idênticas;

· Cada tentativa só poder ter dois resultados;

· As possibilidades de ocorrência dos resultados mantenham-se constante ao longo da experiência;

· As tentativas serem independentes

Portanto, voltando ao caso do clínico geral, o que ele pretende é, numa recolha de 5 fichas do seu ficheiro æ tentativas æ encontrar 3 doentes hipertensos æ sucesso æ, sendo que qualquer outro valor será interpretado como insucesso.

É preciso frisar que o Clínico Geral deve, após retirar uma ficha repô-la de novo no ficheiro, a fim de manter a probabilidade de sucesso constante, e deve igualmente assegurar que cada recolha seja independente da anterior (i. e. não seja influenciada pelo resultado prévio).

Nota: Se a amostragem for feita sem reposição, isto é se após retirada a ficha esta não regressar ao ficheiro, a distribuição a utilizar não será a binomial mas sim a distribuição hipergeométrica, de que falaremos mais à frente.

Estamos assim perante uma experiência binomial, a qual pode ser descrita por uma variável aleatória binomial.

A variável aleatória X que representa o número de sucessos p em n tentativas é denominada variável aleatória com distribuição binomial, ou simplesmente binomial

Assim a variável binomial é utilizada para explicitar a possibilidade de x sucessos em n tentativas de uma experiência binomial.

Para calcular a probabilidade de x sucessos em n tentativas só necessitamos de conhecer n e p, ou seja o número de

tentativas e a probabilidade "teórica" do sucesso.

Estes são os parâmetros da distribuição binomial.

A expressão analítica da distribuição binomial tem a seguinte forma:

com

Como se pode observar o primeiro elemento desta expressão é uma combinação de n elementos x a x, a qual é calculada pela expressão da direita. Esta invoca um factorial (x!,) que corresponde ao produto do um numero por todos os que estão antes do si, excepto o 0.

Exemplo: 3!=3x2xl, 5!=5x4x3x2xl, etc.

Referimos o leitor para a Caixa 2 onde encontra pormenores sobre a expressão analítica bom como a média e a variância da distribuição Binomial.

Voltando ao nosso exemplo temos então:

- n=5 d de tentativas
- p=0,20 probabilidade, de encontrar num doente hipertenso no ficheiro
- x=3 n.º de sucessos pretendido encontrar em 5 tentativas

Logo pela fórmula

P(x=3)=0,0512

Logo o clínico Geral terá cerca de 5% de probabilidade de encontrar 3 doentes hipertensos em 5 fichas. retiradas do um ficheiro em que a probabilidade "teórica" do encontrar

um doente hipertenso é de 20%.

Outra forma do calcular a probabilidade é recorrendo á tabela da Binomial que tem a função de distribuição tabelada para o número do tentativas (n) a probabilidade (p) e o número de sucessos pretendido (x).

Assim introduzindo, os valores, o cruzamento da linha e coluna respectiva lê o valor da probabilidade.

Nota: Quando n é muito grande torna-se difícil - ou mesmo impossível - o cálculo da. probabilidade através de tabelas, Nesse caso utilizamos a chamada aproximação quer à normal quer à Poisson, que ajudarão a resolver o problema. Voltaremos a este assunto mais á frente.

Chamamos a atenção para a seguinte nota pratica.

Nota: As tabelas e calculadoras lêem a probabilidade quer a direita, quer à esquerda do valor que procuramos; ou seja: Algumas vezes obtemos o valor na aba. direita da curva representativa da. distribuição, noutras na aba esquerda. Nessa medida há que ter cautela quando se fazem as leituras, verificando previamente qual a forma como tal é encarado na tabela, ou na calculadora. Senão ao pedirmos o valor de P(X>x) podemos obter o valor inverso, ou seja P(X<x) caso a tabela leia à esquerda. Para evitar este problema dever-se-á consultar a imagem que acompanha as tabelas, na qual vem marcada a zona sombreada, ou seja a aba que ela lê. Nas calculadoras a consulta do manual esclarecerá as duvidas. Contudo lembramos que como a soma de toda a probabilidade é igual à unidade se o resultado obtido for o inverso do lido bastará subtrai-lo da unidade para obter o valor exacto.

Variáveis Aleatórias Discretas: Distribuição do Poisson

A Distribuição de Poisson assim denominada em homenagem ao matemático francês Simeon D. Poisson é outra variável de grande interesse em Bioestatistica e Epidemiologia.

A sua utilização centra-se na descrição de fenómenos que se sucedem em intervalos do tempo definidos, sendo por isso muito útil para descrever situações de sequência temporal.

Os exemplos clássicos referem o número de doentes que acorrem a um serviço de urgência num intervalo de tempo definido (uma hora, por exemplo), a passagem de viaturas num ponto definido, ou a chegada de chamadas telefónicas a uma central num dado período.

Fique claro que a noção de "sucessão de acontecimentos num determinado espaço de tempo" é o ponto crucial da utilização desta variável.

Tem um parâmetro único l que representa o número médio de ocorrências no período de tempo em apreço.

A expressão algébrica da distribuição de Poisson é a seguinte:

Esta expressão além do nosso já

conhecido factorial, contém ainda um novo elemento que é a constante e.

Trata-se da chamada constante de Euler, numero irracional de valor aproximado 2,71828.

Exemplo: De acordo com as estatísticas do um Serviço de Urgência o número médio de doentes. no grupo etário dos 15-24, que acedem ao serviço por hora é de 12. Qual a probabilidade de numa hora acederem ao serviço 9 doentes?

Assim temos:

- l=12 n.^{o} médio de acessos no grupo etário 15-24 numa hora
- x=9 n^{o} de acessos cuja probabilidade pretendemos calcular

Resolvendo na fórmula vamos encontrar:

P(x=9)=0.0874

Logo a probabilidade da chegada de 9 doentes numa hora pertencendo ao grupo etário referido é de cerca do 9%.

A distribuição de Poisson está tabelada para vários valores de l e de x. Por isso não será necessário o leitor fazer contas e utilizar a fórmula de cálculo.

Tudo o que necessita saber é o valor de l e de x.

No final deste capitulo apresentaremos a forma prática de obter estes valores utilizando o STATA.

Variáveis Aleatórias Continuas: Distribuição Normal

Entramos agora no campo das variáveis continuas e destas sobressai pela sua importância em toda a estatística a variável aleatória com distribuição normal, ou simplesmente normal

A sua importância deriva do facto desta distribuição ser a que melhor se adapta aos vários fenómenos da natureza, incluindo os da área da medicina.

Assim a distribuição do colesterol, da glicemia, do ácido úrico na população assume uma distribuição normal, tal como a distribuição do peso ou da altura, só para citar alguns exemplos comezinhos.

A sua representação gráfica é a bem conhecida curva em forma de sino, que recordamos na Figura 1.5.

As principais características da distribuição normal são as seguintes:

A área total sob a curva é igual a 1;

A curva é simétrica em relação á média

As duas abas da curva extendem-se indefinidamente

Os parâmetros da distribuição normal são a média m e a variância s2, que como já vimos é usualmente substituída pelo desvio padrão s.

Como se torna evidente é portanto possível imaginar como que uma família de distribuições normais, variando umas das outras pelo valor da média e do desvio padrão.

A Figura 1.6 mostra várias normais com a mesma média e desvio-padrão diverso, assim como um conjunto com médias doferentes e um mesmo desvio-padrão

Distribuição Normal Estandardizada

Como se pode antever da observação da a distribuição normal seria impossível do tabelar se não pudéssemos encontrar uma forma de a "regularizar".

Facilmente se compreende que não era possível estabelecer tabelas para distribuições com as mais variadas médias e desvio-padrão

Facilmente se compreende que não era possível estabelecer tabelas para distribuições com as mais variadas médias e desvio-padrão

Encontrou-se então a solução construindo uma variável normal estandardizada com média 0 e variância 1 como mostra a na qual todas as outras "normais" se podem converter.

Desta forma podemos então estabelecer uma referência com a qual podemos proceder ao cálculo, da área sob a curva, que é o mesmo que dizer, probabilidade.

Figura 1.5 Distribuição normal

Desta forma podemos então estabelecer uma referência com a qual podemos proceder ao cálculo, da área sob a curva, que é o mesmo que dizer, probabilidade.

Trataremos para já do cálculo das probabilidades, indo, do seguida à técnica da conversão da normal à sua forma estandardizada da Figura 1.7

Terminaremos esta secção com a apresentação de algumas aplicações práticas envolvendo, a normal.

Figura 1.6 Normais com médias e desvio-padrão diferentes

Como referimos a normal estandardizada tem média 0 e variância 1. Assim podemos calcular (sabendo, que a soma da probabilidade, sob a curva é a unidade), a probabilidade contida entre qualquer ponto do eixo dos xx.

Note-se que no caso da distribuição estandardadizada os valores dos quantis da mesma, não costumam ser referidos como x, mas como z, como se pode ver naFigura 1.8

Imaginemos agora que pretendíamos saber qual a probabilidade contida entre os pontos

Figura 1.7 Normal estandardizada

sombreados. verifiquemos que a estes pontos corresponde o valor de -1.960 e 1.960, uma vez que eles estão simetricamente colocados.

Consultando uma tabela vemos então que essa probabilidade corresponde a:

P(-1.96<z<1.960)=0.95

Ou seja 95% da probabilidade está contida entre estes dois pontos.

Como se pode observar a distribuição é simétrica pelo que basta ler os valores de um lado para conhecer os resultados do outro lado e o valor total.

Neste momento antevê-se o leitor, algo intrigado, perguntando aos seus botões:

Figura 1.8 Áreas sob a curva normal estandardizada

Mas afinal para que serve isto?

Dentro do algum tempo irá verificar que a interpretação dos testes de hipóteses, a determinação de intervalos do confiança, etc. vão depender da leitura de probabilidades sob a curva normal estandardizada.

Não desespere portanto. Um pequeno esforço agora será recompensado com mais á vontade no futuro.

Estandardização de uma Distribuição Normal

Explicada a forma do ler as probabilidades numa distribuição normal estandardizada. resta-nos o problema do saber como converter uma distribuição normal à sua forma estandardizada.

Por outras palavras, converter unia distribuição normal com. média m e desvio padrão s, numa distribuição normal com média 0 e desvio padrão 1.

Ou seja, converter os quantis da distribuição normal x, nos quantis z da distribuição normal estandardizada.

Tal conversão é operada através da fórmula:

com m e s a média e o desvio padrão da, distribuição normal de X.

Portanto para calcularmos o valor z a partir de um valor x calculamos a diferença entre este valor e a média m e dividimos o resultado pelo desvio padrão s.

Um exemplo clarificará as coisas.

Exemplo: Seja X uma variável aleatória normalmente distribuída com uma média de 25 e um desvio padrão de 4. Calcule a área:

a) entre x=25 e x=32 b) entre x=18 e x=34

Solução: A referida distribuição tem m=25 e s=4

a) Em primeiro lugar temos de estandardizar a variável convertendo os valores de x para z. Assim:

Empregando a formula para x=25

Para x=32 vem:

Portanto o problema do cálculo da área entre x=25 e x=32, converteu-se no problema do cálculo da área da distribuição normal estandardizada entre 0 e 1.75, que qualquer tabela nos fornece e que é igual a:

P(25<x<32)=P(0<x<1.75)=.4599

b.) A solução da segunda parte do problema segue os mesmos passos.

Primeiro estandardizar os valores x=18 e x=34:

z=18-15/4=-1,75 e z=34-25/4=2,25

Logo o nosso problema de calcular a P(18<x<34) converte-se no problema de calcular a P(-1.75<z<2.25) valor que podemos facilmente ler da tabela da normal e que é igual a:

P(-1.75<z<2.25)=.4599+4878=.9477

Para terminar esta secção iremos apresentar uma aplicação pratica da distribuição normal, a qual ajudara' a tornar mais visível a sua utilidade.

Todos os livros do introdução à estatística como o já citado Mann (op. cit.) contem abundantes exercícios de estandardização, bem como exemplos práticos que o leitor mais interessado poderá consultar onde poderá' praticar,

Exemplo: Determinado laboratório produz vários tipos de soros. A maquina de enchimento dos respectivos balões está ajustada para colocar 500 mI, de soro em cada balão de 500 ml. Contudo a quantidade real do soro que cada balão contém não é exactamente 500 ml; varia do balão para balão.

Sabe-se que a quantidade liquida de soro neste tipo de balão tem uma média de 500 mI com. um desvio padrão de 15 ml.

a) Qual a probabilidade que um balão de soro seleccionado ao acaso contenha entre 490 e 520 ml?

b) Qual a percentagem de balões que contêm entre 505 e 507 ml?

SOLUÇÃO: Seja X o peso liquido do cada balão de soro. Então X tem média 500 e desvio padrão 15

a) A probabilidade do um balão conter entre 490 e 520 mI é dada pela área sob a curva normal entre 490 e 520. Para a determinar vamos estandardizar; Assim:

z=(490-500)/15=.66 e z=(520-500)/15=1.33

Então:

Da leitura da tabela da normal vemos que:

Logo temos cerca de 65% de probabilidade de encontrar um balão com uma quantidade do soro entre 490 e 520 ml.

b) Mesmos passos:

z=(505-500)/15=.33 e z=(507-500)/15=.46

Logo pretendemos:

i. e.

Da leitura da tabela vemos que:

Ou seja cerca de 4,8% dos balões de soro, conterão entre 505 e 507 ml...

Variáveis Aleatórias Continuas com Distribuição t de Student Qui Quadrado e F

Estas distribuições assumem igualmente grande importância em Bioestatistica, em aplicações bem definidas.

Não é possível neste momento avançar com mais detalhes acerca da sua utilização.

Nos capítulos seguintes irão sendo introduzidas à medida que a compreensão do texto tal o permita.

Vamo-nos limitar aqui a apresentar o seu "aspecto" e, tecer algumas considerações brevíssimas sobre elas.

Na Caixa 2 apresentamos as expressões analíticas, bem como a média e a variância de algumas distribuições

Em relação a Qui-Quadrado e t de Student apresentamos só a média e variância, a fim do não ter do invocar funções complexas sem qualquer interesse prático para o leitor: a distribuição F fica para depois.

A Figura 1.9 mostra o gráfico da distribui ção t de Student assim denominada em homenagem ao estatístico inglês William

Gosset,que publicava os seus artigos sob o pseudónimo de Student

Em tudo faz lembrar a distribuição normal; repare-se contudo, que as suas abas são mais "Pesadas", isto é

"demoram" mais tempo a aproximar-se do eixo horizontal.

É uma distribuição que iremos empregar abundantemente ao longo dos próximos capítulos.

Em tudo faz lembrar a distribuição normal; repare-se contudo, que as suas abas são mais "Pesadas", isto é "demoram" mais tempo a aproximar-se do eixo horizontal.

Figura 1.9 Distribuição t de Student com 5 graus de liberdade

É uma distribuição que iremos empregar abundantemente ao longo dos próximos capítulos.

A Figura 1.10 apresenta a distribuição de Qui-Quadrado e a Figura 1.11a F

respectivamente.

São bastante semelhantes entre si. Note-se que não são simétricas.

A sua aplicação será desenvolvida tal como no caso da t de Student nos capítulos próprios, sendo que desem-

Figura 1.10 Distribuição de Qui-Quadrado com 5 graus de liberdade

penham papel de relevo tanto no campo da Estatística. como no da Epidemiologia.

Para o estudo destas distribuições é necessário introduzir o conceito de graus de liberdade (r).

0 número de graus de liberdade é definido, "grosso modo" como "o númerode observações que podem ser escolhidas livremente".

Figura 1.11 Distribuição F com 10 e 15 graus de liberdade

Atente-se no seguinte exemplo:

Exemplo: Imaginemos que conhecemos 4 valores, cuja média é 20.

Consequentemente a soma desses 4 valores é 20*4=80.

Então quantos valores, de entre 4, podemos escolher livremente de modo a que a sua soma perfaça 80?.

A resposta é que podemos escolher livremente 4-1=3 elementos.

Suponhamos que escolhemos 27, 8 e 19. Com estes valores e a informação que a sua média iguala 80,então o quarto valor será 80-(27+8+19)=26.

Logo tendo escolhido três elementos, o quarto fica automaticamente determinado.

É evidente que a nossa liberdade. de escolha dos três elementos é total. escolhidos estes o quarto fica automaticamente determinado. Logo dizemos que neste caso, temos r=4- 1=3.

Convirá reter para já a ideia, que irá ser mais desenvolvida adiante, que quando se faz inferência estatística baseada em parâmetros que não conhecemos, e que por isso temos de estimar perdemos graus do liberdade, ou seja ficamos mais condicionados, por cada valor que desconhecemos e que logo temos que "adivinhar".

Voltaremos a esta matéria em capítulos seguintes.

Usando o STATA 6.0

Introdução

A utilização de "software" estatístico será uma constante ao longo deste manual. Ao longo de cada capítulo apresentaremos os comandos necessários p ara a solução das matérias apresentadas.

Iremos aqui introduzir o "software" que escolhemos para nos acompanhar o ST.AIA 6.0.

Este pacote de "software" estatístico é escrito pela Stata Corporation, empresa sediada no Texas.

É um pacote genérico, isto é. cobrindo todas as áreas da estatística, mas que se encontra particularmente preparado para lidar com os problemas da epidemiologia, uma vez que nos seus módulos se encontram vários comandos específicos para matérias do foro epidemiológico, conforme veremos a seu tempo.

Outro aspecto particularmente interessante é que, ao contrário do outros concorrentes que vendem módulos isolados para actualização de áreas especificas, o STATA é trimestralmente actualizado.

Na verdade a Stata Corporation publica quatro vezes por ano o Stata Technical Bulletin, revista que contem as novidades sobre o programa e que se faz acompanhar do um disco com todos os "add-ons" descritos nesse número.

Assim o programa fica vivo, nunca se desactualizando; para além do mais quem não quiser comprar o STB pode descarregá-lo da Internet, inclusive de "sites" portugueses onde ele se encontra.

0 programa é igualmente multiplataforma, escrito para DOS, Windows, Macintosh, Unix, etc. e apresenta-se em dois formatos principais. 0 STATA propriamente dito, e o STATAQUEST.

Este ultimo é uma versão mais ligeira, que se dedica fundamentalmente à aprendizagem.

Possui menus pendentes que facilitam a sua utilização. É mais limitado no numero máximo de observações que pode manipular. Tem contudo um preço irrisório e, acompanha-se de um excelente manual.

Contudo o STATA pode ser usado com menus pendentes, através do um "add-on" que se pode obter na Net.

É esta ultima versão que vamos usar sempre

que possível.

Nos aspectos mais sofisticados teremos do recorrer à linha de comandos a fim, de aproveitar todo o poder do programa.

Digamos ainda que o STATA é programável, sem excessiva dificuldade, pelo que o leitor, se tiver veia informática, poderá escrever os seus próprios módulos.

Acompanha-se de vastíssima documentação e existe uma newslist, de inscrição gratuita, onde o leitor pode obter esclarecimentos sobre as suas dúvidas.

Por outro lado o distribuidor em Portugal, que é a Timberlake Consultants, tem promovido cursos via Internet de aprendizagem do STATA na área da estatística médica, abertos aos interessados.

Mais informações sobre o produto poderão ser obtidas junto do referido distribuidor em Portugal ou através da Net em http://www.stata.com.

Vamos então passar à apresentação do STATA 6.0.

Descrição do STATA 6.0

Na Figura 1.12 podemos observar o aspecto geral do programa.

Como se vê tem 4 janelas, sendo a central a que apresenta os resultados das operações efectuadas.

Por baixo encontra-se a janela em que se escrevem os comandos. Na esquerda em cima vemos a janela que nos apresenta os comandos emitidos, e por baixo a janela com as variáveis que os nossos dados contêm.

Note-se contudo que a Figura 1.12 representa a versão STATAQUEST com menus pendentes. Por isso o écran que o leitor possa obter poderá ser algo diferente.

Contudo basicamente as coisas diferem pouco.

Existe uma barra do menus para os comandos do gestão de ficheiros (File.); para as operações habituais de Copy e Paste (Edit.); para ajustar as preferências do menus, janelas, etc. (Pref.), para gráficos (Graph.), para o manuseamento do dados (Data), para sumarização dos mesmos (Summaries.), para a realização de testes estatísticos (Statistics) e para a utilização do STATA como se do uma calculadora estatística se tratasse (Calculator).

Por baixo existe uma barra com botões com várias funções entre as quais se destacam a de invocar o editor do STATA que é uma verdadeira folha do cálculo.

Figura 1.12 Écran principal do STATA 6.0

A propósito diga-se que o STATA importa dados de varias folhas do cálculo, bem como os formatos "comma-delmited" e "tab-delimited", assim, como o formato ASCII.

Diga-se igualmente que existe um programa 'a parte o Stata Transfer, que se encarrega de importar formatos de quase todos os programas de estatística, folhas do cálculo, bases do dados, etc.

Não vamos, óbvia mente, descrever o STATA em pormenor, limitando-nos, afim do

não tornar o texto muito pesado, a apresentar somente as imagens que consideramos essenciais Para a compreensão do mesmo.

No restante apresentaremos os comandos pela sua ordem sequencial, bem como o "output" do STATA.

Admitimos também que o leitor tenha um conhecimento mínimo do ambiente Windows e das suas aplicações.

Feitas as apresentações passemos então à pratica.

Caso 1 A sumarização dos parâmetros de unta variável.

Tendo aberto o ficheiro que contém os valores da variável que queremos analisar, ou tendo preenchido esses valores no editor do STATA, invocamos o menu Summaries-Means and SDs.

A seguir surge-nos uma caixa de diálogo a pedir o nome da variável que pretendemos analisar. Indicando o sou nome, temos o "output" seguinte:

```
. summ age
Variable | Obs   Mean  Std. Dev.  Min  Max
---------+--------------------------------
     age | 48    55.875 5.659205  47   67
```

Sublinhámos o comando que nos produziu o "output" que vemos. Para uma variável imaginária "age", com 49 observações obtivemos a média de 55.875 e o Dp. do 5.659. 0 STATA diz-nos ainda que o valor mínimo das nossas observações é de 47 e o máximo de 67 anos

Caso 2 Problemas envolvendo a binomial e a Poisson.

Recordemos o caso do clínico geral que pretendia saber qual a probabilidade de encontrar 3 doentes hipertensos em 5 fichas retiradas de um. ficheiro em que a probabilidade de encontrar um. doente hipertenso é de 20%.

Neste caso recorremos ao menu Calculator-Statistical-Tables-Binomial no qual nos apresenta a caixa de diálogo seguinte:

Como se pode observar o STATA pede-nos os parâmetros da distribuição: N^{o} de tentativas e Probabilidade de sucesso, assim como o no do sucessos pretendido. Preenchida a caixa o resultado é como segue:

B(5,0.20) = 3

Pr(k 3) = 0.0512

Pr(k 3) = 0.0579

Pr(k <= 3) = 0.9933

Note-se que o programa nos dá as três leituras da probabilidade; igual a, à direita de, e à esquerda do ponto pretendido.

No nosso caso o valor encontrado é coincidente, com o valor encontrado na
pagina 1-29 P(x=3)=0,0512- como era
previsível.

Para o exemplo da Poisson procedemos da mesma forma.Vamos ao menu Calculator-Statistical-Tables-Poisson; na caixa de dialogo que nos surge,

e que solicita o parâmetro l e o valor que pretendemos, preenchemos os nossos valores.

0 resultado é:

Poisson(12) =9
Pr(k>=9)=0.0874
Pr(k >= 9)= 0.8450
Pr(k<= 9)=0.2424
Coincidente mais uma vez com o resultado da pagina 1-32 P(x=9)=0.0874
Caso 3 Áreas sob a curva normal reduzida

No exemplo, atras referido pretendíamos saber qual a área entre x=25 e x=32 numa normal com média=25 e desvio padrão=4

Aqui existe um pequeno problema. Como sabemos as tabelas dão-nos as leituras numa normal estandardizada, ou como também se diz, reduzida. Ora então vamos ter de estandardizar à mão, e depois pedir as leituras ao STATA.

No nosso caso iá vimos que os valores pretendidos sã 0 e 1.75.

Então vamos saber qual a probabilidade dw1.75

Seguimos por Calculator-Statistical Tables-Normal e vemos a caixa de diálogo

Preenchemos o valor pretendido (1.75) e obtemos o resultado:

Tablesq Z (1.75)
Pr(Z<=1.75)=0.9599
Pr(Z >=1.75)=0.0401
Pr(|Z|>=1.75)=0.0801

Recordemo-nos que a probabilidade á direita do ponto 0, que é a média é de 0,50.

Por outro lado vemos que a probabilidade á direita de 1.75 é do 0.0401.

Logo, por uma simples operação de subtracção 0.50-0.0401=0.4599, resultado que corresponde à probabilidade contida entre 0 e 1.75, conforme tínhamos já visto na pagina 1-40

P(25<x<32)=P(0<x<1.75)=.4599

Vimos assim, neste primeiro contacto, como o STATA 6.0 torna simples as operações mais elementares.

Veremos mais adiante como simplifica igualmente as mais complexas.

Para finalizar lembramos que a estatística, e a informática, se aprendem praticando, pelo que o leitor terá toda a utilidade se puder acompanhar a leitura deste livro com uma cópia do STATA, ou mesmo do STATAQUEST.

Capítulo 2

Métodos gráficos em análise de dados

2.1 Introdução

Todos temos a noção que, de acordo com a famosa máxima, uma imagem vale mais que mil palavras.

E de facto é hoje usual e, diríamos quase inevitável que quem trabalha em estudos mais ou menos científicos, utilize profusamente a imagem æ habitualmente sob a forma de gráficos æ para ajudar a visualizar a informação que se pretende veicular.

Ganharam assim grande popularidade os chamados "gráficos de apresentação" tais como os gráficos de barras, gráficos de tarte ou queijo, (pie charts) e as inevitáveis pirâmides etárias bem conhecidas sobretudo dos médicos de Clínica Geral que as utilizam profusa e quase inevitavelmente nos vários relatórios que ao longo da sua vida profissional vão produzindo.

Contudo devemos recordar a existência de um outro grupo de gráficos a que chamamos por extensão "gráficos analíticos" que se revelam de uma superioridade inequívoca e ocupam um lugar indiscutível nos métodos de análise exploratória de dados.

Reportando-nos aos tempos em que eram feitos á mão, com papel milimétrico e lápis, observamos a sua espantosa evolução, devido á introdução dos computadores e do respectivo "software" estatístico que hoje permitem a sua realização de uma forma prática e expedita.

Mais do que servirem para uma simples apresentação de resultados este conjunto de ferramentas permite sobretudo:

· Analisar a distribuição das variáveis, per si (normalidade, desvio, simetria e "outliers");

· Comparação de duas ou mais distribuições;

· Comparação da distribuição observada de uma dada variável com uma distribuição teórica;

· Post- análise de modelos como o de regressão.

Se recordarmos que a aplicação de certos métodos estatísticos depende de algumas condições, como a normalidade da distribuição das variáveis, entre outras, verificamos a importância de poder avaliar, de início, com a máxima precisão possível as características das variáveis em estudo.

Por outro lado a detecção de "outliers" (valores extremos que se afastam significativamente do centro da distribuição), pelo efeito de enviesamento que introduzem a estatísticas pouco robustas, como a média, é tarefa fundamental na fase de análise exploratória de dados, afim de permitir orientar de forma correcta toda a investigação ulterior.

Por esse facto, e também porque entendemos que este tópico se encontra dispersamente explicado em vários textos, decidimos incluir um capítulo em que iremos descrever sucintamente os vários tipos de "gráficos analíticos" assim como a sua utilização na prática.

Contudo, e porque esta é uma matéria vasta, e do nosso ponto de vista muito importante, recomendamos ao leitor a

consulta de algumas obras de referência entre as quais destacamos:

O livro de Tukey (1977) é de referência absoluta nesta matéria.

Igualmente recomendável é a obra de Chambers e col. (1998) que apresenta uma análise muito cuidada dos vários métodos.

Num plano mais acessível merece referência o livro de Hamilton (1992) o qual apresenta estes métodos de uma forma clara, cobrindo ainda outros aspectos dos modelos de regressão.

Iremos apresentar nas secções seguintes gráficos para análise univariada e para análise bivariada, nomeadamente entre 2 ou mais variáveis e entre estas e uma distribuição teórica.

No capítulo respectivo mostraremos a utilização dos métodos gráficos em análise post-regressão e em ANOVA.

Análise univariada

Quando olhamos para um conjunto de dados, que representam os valores que dada variável, chamemos-lhe X assume na nossa experiência, uma primeira questão se pode colocar ao nosso espírito.

Como interpretar a diferença de valores de uma observação para outra?

Tomemos um exemplo que extraímos de Hamilton(1992).

Durante a década de 70 a cidade de Concord, NW experimentou um aumento de consumo de água, apesar de manter uma população estável.

No final de 1979 este aumento associado a um ano particularmente seco levou a uma redução dos recursos hídricos.

No final de 1980 a Companhia das Águas de Concord desencadeou uma campanha para sensibilizar a população para consumir menos água.

No ano seguinte o consumo de água caiu cerca de 15%, o que foi naturalmente interpretado como resultado positivo da campanha.

Obviamente 15% de declínio não significa que toda a gente tenha reduzido o consumo em 15%. Alguns terão poupado mais água, outros menos.

Em 1981 o Concord Water Study examinou essas variações na redução do consumo de água.

Foram enviados questionários a uma amostra aleatória de residências de Concord inquirindo sobre características demográficas, opiniões e hábitos de poupança (de água).

Estes questionários foram então confrontados com os registos da Companhia das Águas sobre o consumo de água das residências no verão de 1980 e de 1981, isto é antes e depois da campanha de sensibilização.

Foram inquiridas 496 casas.

A primeira questão que se coloca é tão simples como:

Como varia o consumo de água de casa para casa?

A resposta a esta pergunta passa pelo simples cálculo da média e do desvio padrão das amostras de 496 casas respectivamente em 1980 e 1981, conforme mostra a Tabela 1.1

Tabela 1.1 Consumo de água

	1980	1981
Média	2731	2298
Desvio-Padrão	1764	1486

A leitura destes valores leva-nos a perceber uma diminuição do consumo de 2731 para 2298, ou seja de 433 unidades de consumo (no caso do estudo o galão, por ser em unidades anglo-saxónicas).

Um histograma da distribuição do consumo de água em 1981 com uma curva da distribuição normal sobreposta é apresentado na Figura 1.1

Figura 1.1 Histograma do consumo de água em 1981

Este gráfico permite-nos observar um desvio para a direita da distribuição, isto é a aba direita da curva é mais "comprida" do que a aba esquerda, demonstrando este facto um maior "peso" nesta parte da distribuição.

Se ela fosse normal seria simétrica, logo as duas abas da curva estariam identicamente distribuídas æ para usar um linguagem simples, embora pouco rigorosa æ.

Contudo não é isso que se verifica.

O alongamento da aba direita leva-nos a verificar que existe um número excessivo de casas com consumos elevados, pelo menos um número superior ao que seria de esperar se a distribuição fosse normal.

Consequências práticas deste facto, a que os anglo-saxónicos chamam "skewness", de difícil tradução mas que podemos chamar de "desvio á direita", ou "à esquerda".

Se existe desvio o centro da distribuição torna-se "ambíguo" o que dificulta a sumarização da variável e ainda mais, a sua comparação com outras variáveis. Resultado:

Só a média e o desvio padrão são insuficientes para descrever com rigor esta distribuição.

Então que fazer?

É aqui que iremos recorrer a outras estatísticas e outros métodos de análise.

Mediana e Intervalo interquartil

Recordemos que teoricamente a mediana divide uma lista ordenada ao meio, sendo portanto que 50% dos valores se encontram abaixo da mediana e 50% acima.

Se n valores se encontram ordenados por ordem crescente a mediana m é o valor na posição:

A mediana, tal como a média é portanto uma medida de tendência central; contudo têm algumas características que a diferenciam desta última, sendo a que mais nos interessa neste caso, a da sua robustez, isto é, o facto de ser menos influenciado pelos valores extremos.

Assim 25% dos valores estão abaixo do 1º quartil, 50% abaixo do 2º quartil e 75% abaixo do 3º quartil.

Note-se que o 2º quartil equivale assim à mediana.

O percentil divide por seu turno a lista em 100 partes; 1% dos valores estão abaixo do 1º percentil, 10% abaixo do 10º percentil, etc.

O intervalo interquartil (IQR) equivale á distância entre o 1º e o 3º quartil, cobrindo assim o valor da mediana.

Esta representa a tendência central, como atrás se disse, aquela a dispersão.

Devido á sua robustez o seu uso é preferível quando as distribuições são enviesadas ou contém valores extremos ("outliers").

Referimos aqui igualmente o conceito de quantil que nos servirá adiante.

O quantil é similar ao percentil mas expresso em fracções ou proporções em vez de percentagens.

Assim o quantil .25 equivale ao 20º percentil. Igualmente .25 dos valores são-lhe inferiores.

Quando raciocinamos em termos de dados ordenados em ordem crescente podemos dizer que o iésimo valor yi transforma-se no quantil (i-.5)/n.

Assim (i-.5)/n é a fracção de casos abaixo de yi contendo metade dos casos em si.

Por exemplo no estudo de Concord a 400ª casa utiliza 3200 unidades de água. 3200 é portanto o (400-0.5)/496=.81 quantil.

Veremos a sua utilização mais á frente, remetendo o leitor para um estudo mais detalhado dos quantis para o livro de Chambers (1998).

Analisemos agora algumas ferramentas gráficas para a análise de dados, a saber: caixas de bigodes ou gráficos de dispersão interquartil, gráficos de quantil e gráficos de simetria.

Caixa de bigodes, ou gráfico de dispersão interquartil

Trata-se de uma ferramenta poderosa, uma vez que nos dá informação acerca da centragem, dispersão e simetria da distribuição da variável.

A Figura 1.2 apresenta um gráfico desta natureza, na qual podemos apreciar a sua "anatomia".

Como vemos o gráfico tem um corpo central que contem as observações entre o 1º e o 3º quartil. A linha central corresponde á mediana (2º quartil).

Figura 1.2 Gráfico de dispersão interquartil

Os limites superior e inferior correspondem aos valores de -1,5*IQR e +1,5*IQR sendo op IQR (interquartil-range) justamente o valor correspondente à diferença entre o 3º e o 1º quartil. Todos os valores que caem for a destes limites (bigodes) são os valores extremos ou "outliers".

Note-se que ao olhar para um gráfico como o da Figura 1.2 se fica logo com a ideia da dispersão da distribuição (se o IQR for grande) da posição da mediana (mais à direita ou à esquerda, i. e. para cima ou para baixo) e se existem valores extremos.

Apreciemos agora a Figura 1.3

Figura 1.3 Gráfico comparativo de três caixas de bigodes

Aqui o que pretendemos é avaliar o numero de "estalidos" dados com os dedos em função da quantidade de cafeína ingerida.

Vemos que á medida que o número de cafés aumenta o número de estalidos aumenta igualmente; note-se que a mediana da distribuição aumenta igualmente.

Estes gráficos são particularmente úteis quando se pretende desagregar uma observação global (p.ex. uma observação numa população) nos seus vários componentes (estados, províncias, concelhos, etc.)

Iremos utilizar largamente estes gráficos mais adiante.

Gráficos de quantil e de simetria.

Neste caso mais do que observar um padrão de distribuição, podemos observar os vários pontos da distribuição.

Estes gráficos são mais incisivos na apreciação da simetria da distribuição e, no caso dos gráficos de quantil servem para comparar a distribuição em estudo com os quantis de uma distribuição normal ou, em análise bivariada que veremos á frente comparar os quantis de duas distribuições em ordem a comparar a sua similitude.

Voltando ao gráfico da Figura 1.2 que nos mostra o número de divórcios por mil nas idades entre os 15 e os 59 anos em 146 cidades americanas vamos agora apreciar um gráfico de simetria da mesma situação.

Antes porém vejamos na Figura 1.4 um histograma da variável, onde se nota um marcado desvio á direita, que já tínhamos apreciado na caixa de bigodes

Figura 1.4 Histograma dos divórcios por 1000 em 146 cidades dos EUA

Comparando agora este gráfico com o gráfico de quantil-normal da Figura 1.7 vemos igualmente que a distribuição apesar de muito aproximadamente normal apresenta os tais "outliers" que corresponderão a cidades com maior índice de divórcios.

Como se vê todos os gráficos nos vêm ajudar no sentido de notar as particularidades da distribuição.

De seguida apresentamos na Figura 1.5 o gráfico de quantil da mesma distribuição e na Figura 1.6 o gráfico de

Figura 1.5 Gráfico de quantil

simetria, os quais corroboram o que atrás tínhamos observado.

Convém lembrar o conceito de quantil definido na secção anterior.

Refira-se igualmente que a abordagem aqui efectuada é extremamente simplista e pretende somente desvendar o valor diagnóstico destas ferramentas.

Remetemos o leitor para a leitura de obras já citadas como as de Tukey (1979), Chambers (1998) Hamilton (1992), e para

Figura 1.6 Gráfico de simetria

o manual de utilização do Stata 6.0 bem como o outro livro de Hamilton ( )

Antes de passarmos á análise bivariada gostaríamos de enfatizar um ponto.

É de toda a utilidade a utilização exaustiva de todos os métodos de análise, mesmo que representam alguma repetição, a fim de testar todas as hipóteses de normalidade, simetria, etc. das variáveis, condição indispensável para muita da análise ulterior.

Figura 1.7 Gráfico de quantil-normal

Não se fique o leitor pelas primeiras impressões.

Quantas surpresas elas podem encerrar.

Análise bivariada

A análise bivariada estuda a relação entre duas variáveis, sendo que as técnicas que vamos apresentar poderão ser extensas á análise multivariada.

Note-se contudo que em termos gráficos se torna difícil visualizar, digamos a mais de três dimensões, pelo que nesses casos se recorrem a técnicas analíticas próprias, algumas das quais iremos referir.

Fixemo-nos para já na análise bivariada.

Iremos apresentar os gráficos de dispersão por pontos (scatterplots), e gráficos de quantil-quantil

Gráfico de dispersão por pontos

São talvez a ferramenta mais fácil de comparar a relação entre duas variáveis.

A Figura 1.8 mostra um gráfico de dispersão em que se analisa a relação entre a idade e o aumento da pressão sistólica.

Como é de prever verifica-se um aumento da pressão sistólica com a idade.

É um gráfico muito utilizado, devendo contudo chamar-se a atenção para o facto de ser um gráfico, digamos mais visual, e de forma alguma suficiente para tirar conclusões ou grandes inferências.

Figura 1.8 Gráfico de dispersão por pontos

Gráfico de quantil-quantil

Tem uma filosofia semelhante à da Figura 1.7.

Aqui temos a comparação de quantis de uma variável contra os quantis de uma normal com a mesma média e variância.

Neste caso trata-se de comparar os quantis de duas distribuições empíricas. Se as duas fossem idênticas veríamos os pontos ao longo de uma linha diagonal.

Figura 1.9 Gráfico de quantil-quantil

No caso da Figura 1.9 estudamos a comparação entre o numero de divórcios e a percentagem de população negra.

Como podemos observar o número de divórcios é bem menor nesta população do que seria de esperar (linha diagonal) e mantêm-se bastante constante mesmo com o aumento da população negra.

Deixamos aqui em linhas gerais algumas ferramentas gráficas básicas para análise de dados.

Mais uma vez remetemos os leitores interessados para as obras referidas anteriormente.

Usando o STATA 6.0

Um dos pontos fortes do STATA 6.0 é a sua capacidade de gerar gráficos.

Diga-se contudo que a versão STATAQUEST não possui nos menus todos os comandos necessários para a produção de todo o tipo de gráficos, sobretudo os mais complexos.

O menu Graph tem comandos para todos os gráficos mais usuais e inclusive para gráficos de que não falámos, como os de sucessões cronológicas (time-series) e para gráficos de controlo de qualidade (quality control).

Contudo para gráficos mais complexos teremos de recorrer á linha de comandos. Deixamos na Tabela 1.2 indicação desses comandos.

Algumas notas importantes:

· Os comandos encontram-se em itálico

· Os termos var, var1 e var2 referem-se aos nomes das variáveis

Tabela 1.2 Comandos gráficos do STATA

Gráfico de simetria symplot (var), xlabel ylabel

Gráfico de quantil 1 drop if var==.

2 sort var

3 gen quantile =(_n-.5)/_N

quantile var, ylabel

Gráfico de quantil-normal qnorm var, ylabel xlabel

Gráfico de quantil-quantil qqplot var1 var2, ylabel xlabel

· Note-se que para o gráfico de quantil-normal foram necessários alguns passos iniciais: 1:Eliminam-se os valores em falta (drop if var==.); 2: Ordenam-se os valores da variável (sort var); 3: Geram-se os quantis com a formula explicitada (exactamente como está escrita); 4: Finalmente emite-se o comando quantile com o nome da variável e o programa produz o gráfico.

Capítulo 3

Conceito geral de população e amostra

3.1 Introdução

Imagine-se que alguém pretendia conhecer o valor médio da glicemia dos portugueses.

Uma forma de resolver o problema seria proceder à medição da glicemia de todos os cidadãos portugueses numa determinada data.

Uma outra seria escolher um lote de indivíduos, de entre o conjunto de todos os portugueses, medir a sua glicemia determinar o valor médio, e a partir dai concluir que esse valor representava o real valor médio da glicemia de todos os portugueses.

A apresentação deste exemplo simples leva-nos para a definição de alguns conceitos essenciais, a saber: População, amostra e inferência.

Vejamos: 0 conjunto de todos os cidadãos constitui, no nosso exemplo, uma população, entendida como o conjunto de indivíduos que possuem dado atributo.

Note-se que o conceito de população estatística não implica que ela seja constituída por seres vivos, como no nosso, caso; nesse sentido a população é todo o conjunto de indivíduos (estatisticamente entendido) que detêm o atributo em estudo.

No nosso exemplo coincide com urna população, no sentido demográfico. Mas nem sempre tal acontece.

Convém igualmente notar aqui, que as populações não são obrigatoriamente infinitas. Podem existir populações finitas e até, em certas circunstâncias, o tal conjunto de indivíduos pode funcionar como uma população finita, ou como uma mostra de uma outra população.

Veja-se um exemplo:

Os doentes que frequentam o Centro de Saúde pode ser entendidos como:

· A população dos utentes do Centro de Saúde do meu concelho;

· Uma amostra dos doentes que frequentam os Centros de Saúde (do Distrito ou até do Pais);

Veremos mais à frente situações deste tipo.

Se de uma determinada população se extrair um lote de indivíduos com o atributo em estudo, dizemos que estamos em presença de uma amostra. Entendido assim, uma amostra é um subconjunto da população.

Esta amostra serve-nos para obter informação acerca do atributo em estudo; a partir desta deduziremos o seu valor na população.

Este é o processo de inferência.

Note-se que o atributo em estudo na população recebe o nome de parâmetro.

0 seu correspondente na amostra denomina-se estatística.

Fique claro que de aqui em diante utilizaremos o termo parâmetro quando nos referimos à população, e o termo estatística quando nos referirmos à amostra.

Não existem assim estatísticas populacionais, nem, parâmetros amostrais. A inversa é que é verdadeira.

Neste capitulo iremos abordar a amostra no seu sentido, genérico, bem como as distribuições amostrais de médias e proporções.

Iremos referir igualmente os aspectos elementares da teoria estatística da estimação, abordando a estimação por pontos e por intervalo de confiança.

Não aprofundaremos as questões da amostragem em si, nomeadamente no que concerne ao cálculo da sua dimensão, etc. Guardaremos essa abordagem para a 2ª Parte deste manual, onde para cada tipo de estudo definiremos o tipo, e a forma, de recolha de amostras.

Terminaremos com alguns exemplos práticos, usando o STATA

Distribuição amostral de

Voltemos ao nosso exemplo da glic6mia.

Imaginemos que escolhemos uma amostra de 50 pessoas e medimos a glic6mia de cada uma delas.

De seguida calculamos a média dessas 50 observações.

Repetimos a operação varias vezes e de cada uma delas calculamos a média observada.

Podemos fmalmente calcular a média dessas médias obtendo assim a média amostral da glicemia, a que vamos chamar, uma vez que se trata de uma estatistica, segundo o conceito que definimos na secção anterior.

Diga-se para já que as estatisticas se designam por letras romanas pequenas, enquanto os parâmetros se representam por letras gregas.

Note-se que se tivessemos a possibilidade de repetir esta operação incontáveis vezes e, se soubessemos o valor da média da glicemia na população veriamos que ao fim de essas incontáveis vezes, o valor encontrado na amostra se ia aproximando do valor conhecido na população.

Isto porque é um estimador centrado de m (termo que definiremos à frente), e por isso este é o seu valor esperado.

Assim sendo vemos que deve assim ser entendida como uma variável aleatória, cuja média, variância edistribuição analizaremos de seguida.

Média e variância de

Corno já dissemos atrás a média da distribulgAo amostral de coincide com a média m da população, isto devido ao facto de ser um estimador centrado de m

Assim:

A variância de não é igual a variância s 2.

Esta é igual a s2 divido pelo número de elementos da amostra, ou de urna forma mais usual, falando de desvio padrão, este é igual a s divido pela raiz quadrada de n.

0 quadro seguinte sumariza esta informação

A fórnula do desvio-padrão de só se revela exacta quando a amostra éfeita com reposição, a partir de uma população finita, ou sem reposição quando se trata de uma população infinita.

Duas notas importantes conviria reter:

Nota: A dispersdo da distribuiqdo de é menor que a dispersdã da correspondente populaqao 0 desvio-padrão da distribuiqdo amostral de diminui àmedida que a dimensão da amostra aumenta.

Nota: O desvio-padrão da distribuição amostral de diminui á medida que a dimensão da amostra aumenta

Esta segunda propriedade, que iremos demonstrar com um exemplo prático, e que decorre da leitura óbvia da fórmula de cálculo (note-se que no denominador se encontra a raiz quadrada de n, a qual aumenta àmedida que a dimensão aumenta, logo reduzindo o valor do cociente), denomina-se consistência do estimador, e falremos dela mais adiante.

Veiarnos o exemplo seguinte:

Exemplo: A média do preço/hora dos 5000 trabalhadores de. uma empresa é de 2430$00 com um desvio-padrão de 122$00. Seja a média do preço/hora de uma amostra aleatória de trabalhadores desta companhia.

Calcule-se a média e. o desvio padrão de para urna amostra de dimensão:

a) 30 b)75 c)200

Resposta: Da informação recolhida dos dados sabemos que N=5000 m=2430 s2=522

Para uma mostra de 30 a relação n/N (30/5000) é igual a 0.006.

Ora desde que este cociente seja inferior a 0.05 podemos aplicar a teoria atrás desenvolvida. Então

Estendendo os cálculos para as outras duas amostras vem

Como podemos observar à medida que a dimensão da mostra aumenta a varifincia do estimador diminui. No limite, se tivessemos uma amostra infmita o estimador teria variância zero e portanto o valor estimado scria igual ao valor do parâmetro.

É nisto que consiste a propriedade da consistencia atrás refenida.

Vá o leitor retendo, como norma prática, a influencia da dimensão da amostra no comportamento das estatisticas.

Iremos referir várias vezes, mais adiante, este aspecto, pelo que é convemente que vá já inteniorizando da vantagem de trabalhar corn amostras grandes.

Forma da Distribuição amostral de

A forma da distribuição amostral de está relacionada com as duas situações seguintes:

1 A população a partir da qual a amostra é retirada tem uma distribuição normal.

2 A população a partir da qual a amostra é retirada não tem uma distribuição normal.

Estas situações são particularmente importantes de considerar, sobretudo no que concerne á segunda.

Como veremos nesse caso vai entrar em jogo um dos mais importantes teoremas da estatística - O Teorema do Limite Central- que nos servirá de suporte para um conjunto alçargado de considerações.

Analisemos então o primeiro caso.

Amostras retiradas de populações normalmente distribuidas

Neste caso, se a população da qual a amostra é retirada é normalmente distribuida, com média e desvio-padrão , então:

A média de é igual á média da população

O desvio-padrão de é igual a desde
que

A forma da distribuição amostral de é normal para qualquer valor de n

Note-se a importância desta ultima conclusão que nos servirá de muito ao longo deste manual

Se pudermos garantir a normalidade da população da qual é retirada a amostra, então por mais pequena que seja a dimensão desta, a distribuição amostral de é sempre normal.

E o que se passa então se não for possivel assegurar a normalidade (como tantas vezes sucede) da população da qual é retirada a amostra?

Estamos assim no segundo caso.

Amostras retiradas de populações que não são normalmente distribuída

Num grande número de situações aplica-se a teoria que este caso encerra.

Na realidade quantas vezes nada sabemos acerca da distribuição populacional de uma variável cujos parâmetros pretendemos estimar através das respectivas estatísticas amostrais.

Nessas condições a forma da distribuição amostral de deduz-se a partir do importantíssimo Teorema do Limite Central que podemos sumariamente formular do modo adiante explicado:

Considera-se urna amostra suficientemente grande se n>=30.

.

Teroema do Limite Central

Para uma amostra suficientemente grande a distribuição amostral de é aproximadamente normal independentemente da forma da distribuição populacional.

A média e o desvio-padrão da distribuição amostral de são e

Formulado de outra maneira podemos dizer que qualquer valor de pode ser sempre. transformado (qualquer que seja a sua distribuição) numa variável da forma:

com z com distribuição normal de média 0 e variância 1, ou seja sendo z uma variável estandardizada, desde que n>=30.

Note-se que quando uma. população não tem distribuição normal - ou dizendo de uma forma mais precisa, quando a variável em estudo não tem uma distribuição populacional normal - a forma da distribuição amostral não é exactamente normal, mas sim, aproximadamente (assintóticamente) normal, sendo essa aproximação tanto melhor quanto maior for a dimensão da amostra em causa.

Repita-se, mais urna vez, que o Teorema do Limite Central só deve ser aplicado em amostras maiores que 30 unidades.

Deve-se realçar a importância do Teorema do Limite Central, ao criar um suporte para toda a inferência que pode ser desenvolvida a partir daqui, independentemente da forma da distribuição populacional.

Sem ele tal não seria possível. e note-se, igualmente e mais urna vez, a importância crucial de se trabalhar com amostras grandes.

Só elas podem garantir segurança na inferência, através precisamente do chamado efeito do limite central.

No final deste capitulo apresentaremos alguns exemplos utilizando o STATA que aclararão estas matérias, nomeadamente realçando a importância da dimensão da amostra.

População e Proporções Amostrais

0 conceito de proporção é sobreponivel ao conceito de frequência relativa.

A frequência relativa de urna categoria ou classe dá-nos a proporção na população, ou na amostra que pertence a essa categoria ou classe.

0 trabalho com proporções é bastante frequente em estudos epidemiológicos e clínicos, uma vez que uma boa parte das variáveis que habitualmente se estudam são do tipo discreto; o seu conhecimento alargado é portanto fundamental para quem queira "dominar" a matéria.

Recomenda-se a esses leitores eventualmente interessados a consulta, entre outros, do clássico c excelente livro de Fleiss ( )que cobre de forma exaustiva este assunto.

A proporção populacional, doravante denominada de obtém-se tomando a razão entre o número de elementos numa população com uma dada característica e o número total de elementos da população.

A proporção amostral designada de dá-nos uma razão similar para a amostra.

Assim:

As proporções populacional e amostral denominadas respectivamente de p e calculam-se da seguinte forma:

e =

- N= no total de elementos da população
- n= n.º total de elementos da amostra
- x= n.º de elementos na população ou na amostra que possuem a característica em estudo.

Média Desvio Padrão e forma da distribuição de

De urna forma similar ao que se disse na secção anterior a média de é igual à proporção populacional p, e o desvio-padrão padrão de é igualmente dado por

Se n/N for maiorque.05 então o desvio-padrão é calculado da seguinte forma:

em que é denominado factor de correcção de população finita.

A forma da distribuição de infere-se do teorema do limite central.

A distribuição amostral de é aproximadamente normal para uma amostra suficientemente grande. Suficientemente grande é neste caso quando np>5 e nq>5.

Teoria Estatística da Estimação

Conforme temos vindo a analisar o objectivo primordial da estatística é fazer previsões, ou inferências, a partir de observações de amostras aleatórias.

De uma forma mais precisa podemos dizer: seja X uma variável aleatória definida por um determinado parâmetro, a que chamaremos genericamente l realizem-se n experiências aleatórias e observemos o equivalente na amostra do parâmetro l e que vamos denominar Afinal será a partir deste último, que podemos observar, que vamos inferir o primeiro que logicamente não podemos conhecer.

Então dizemos que é um estimador de l.

Algumas questões se nos colocam, a saber:

- Como é possível determinar um estimador para l ?
- Que garantias temos que este valor "é de confiança"?
- Poderemos quantificar essa confiança?

Não vamos nesta secção responder a todas estas questões; elas irão sendo respondidas passo a passo. Por agora interessa-nos a resposta à primeira.

Problema: Métodos para encontrar um estimador?

A resposta a esta pergunta passa pela clássica distinção entre estimação por pontos e estimação por intervalos de confiança.

Alerte-se o leitor que vamos referir esta matéria de uma forma muito superficial, fazendo referência somente aos aspectos de ordem prática; até porque a teoria da estimação, sobretudo no que se refere à estimação por pontos, é bastante complexa e ultrapassa. nitidamente o objectivo de um texto básico como este.

Contudo o leitor mais interessado poderá consultar qualquer uma das obras citadas, nomeadamente as de Galvão de Melo, Bento Murteira ou Tannis and Hogg, onde encontrará explicações detalhadas sobre métodos de estimação por pontos, como o dos momentos ou o de máxima verosimilhança. este tão utilizado actualmente devido à qualidade dos estimadores que produz.

Para além de uma ou outra referência que iremos fazer aos estimadores de máxima verosimilhança, mais á frente, iremos manter, como dissemos o nosso texto a um nível coloquial, evitando o mais possível o recurso à estatística matemática.

Na Caixa 3 iremos contudo desenvolver superficialmente o método da máxima verosimilhança pela importância que esses estimadores assumem em epidemiologia como atrás referimos e aplicaremos à frente.

Vejamos então:

Para encontrarmos um estimador para. um determinado parâmetro de uma distribuição populacional existem dois métodos usados habitualmente em paralelo.

0 método de estimação pontual, e o método de estimação por intervalos de confiança.

Vejamos em que consiste a sua diferença usando para tal uma analogia bélica.

Imagine-se que pretendíamos acertar num determinado alvo com uma arma de fogo. Imaginemos que a arma era excelente e o atirador possuía pontaria afinada. Nestas condições muito provavelmente o atirador acertaria na "mouche" com todos os tiros que disparasse.

Mas imagine-se agora que o alvo estava a uma distância maior e que o atirador em vez de usar uma espingarda ou uma pistola utilizava uma arma de tiro curvo, como um morteiro.

Então o que iria acontecer é que o atirador não iria acertar em cheio na "mouche" mas sim cobrir urna área que numa boa parte de casos incluiria o alvo pretendido.

Esta comparação um pouco grosseira permite-nos contudo aproximar do que pretendemos.

Assim diremos que no primeiro caso optamos por um instrumento (estimador) que nos permite "palpitar"um valor (estimativa) para o parâmetro a conhecer.

No segundo caso optamos por usar outra técnica, que em vez de uma estimativa nos fornece uma região (intervalo de confiança) que supomos "cobrir" o valor que pretendemos encontrar.

Note-se que o valor do parâmetro que pretendemos estimar é um valor fixo, o que varia são os resultados que os nossos métodos de estimação produzem. Por isso não é correcto, como muitas vezes se ouve dizer, que a média por exemplo de uma distribuição está contida no intervalo tal.

Ora a média não está contida em intervalo nenhum; o fruto da nossa observação é que produziu um determinado intervalo que supomos, com uma determinada probabilidade, englobar a média da dita distribuição.

Já que falamos em média diga-se que é o exemplo mais usual de um estimador pontual (de m) para o qual se calcula um intervalo de confiança, da forma que mostraremos mais adiante.

Quais são as propriedades que um estimador pontual deve ter para que nos
seja útil.

Esta pergunta requer uma complexa resposta que vamos contudo sumarizar assim.

Para que um estimador seja de qualidade ele dever ser:

· consistente,
· não enviesado,
· suficiente
· eficiente

Em duas palavras e sem matemática o que é que isto quer dizer:

Um estimador é consistente se a sua variância tender para zero quando a amostra aumenta indefinidamente; ou seja à medida que a dimensão da amostra aumenta a variância reduz-se, logo o valor aproxima-se do parâmetro.

Um estimador é não enviesado se o seu valor esperado for o parâmetro. Quanto maior for o enviesamento, pior é a "pontaria", para usar a linguagem bélica de atras.

Um estimador é suficiente se nos puder fornecer por si só a máxima informação possível sobre o parâmetro

Finalmente um estimador é eficiente se tiver variância mínima, sendo esta (a variância mínima) definida por uma expressão designada por desigualdade de Frechnet-Cramier-Rao.

Para mais explicações remete-se o leitor para as obras de referência.

0 que interessa reter é que um estimador para ser útil tem de ter propriedades relevantes, sem as quais não se poderá garantir inferência de qualidade.

Por outro lado refira-se que o cálculo do intervalo de confiança serve muitas vezes como uma contraprova no sentido de testar ou confirmar o resultado obtido através da estimação pontual.

Note-se que possui as características atrás referidas, isto é um estimador centrado (não enviesado) e tem variância mínima.

Nem sempre as coisas se passam assim:

Para a variância o estimador s2 é calculado como: não é estimador cventrado de

Para encontrar um estimador centrado de teremos de o calcular da forma:

ou seja, demonstra-se que o estimador centrado da variância se obtém com n-1 (numero de observações menos 1) no denominador, em vez de n como no caso do estimador da média.

Atente o leitor na Figura 1.1

Figura 1.1 Estimação por intervalos de confiança

Pretende-se visualizar o que se disse antes.

Como observamos a estimação do parâmetro é feita através da obtenção de repetidos intervalos confiança (em sucessivas experiências aleatórias).

Note-se que existem pelo menos dois caos em que esses intervalos não englobaram o parâmetro a estimar, pelo que a precisão desses intervalos de confiança terá de ser definida de uma forma probabilistica, do tipo "temos 95% (ou outra) dce probabilidade que o intervalo -x +x englobe o parâmetro a analisar".

Só assim é que ele ficará correctamente definido.

Essa é a matéria que ocupará o próximo capitulo.

Por agora vamos regressar ao convívio do STATA para consolidar as ideias descritas neste capitulo.

Usando o STATA 6.0

Nesta secção mais do que trabalhar corno o STATA na resolução de problemas, eles virão mais tarde, iremos utilizar urna ferramenta disponibilizada no livro de (Stata Concepts) que nos vai permitir visualizar os aspectos discutidos no capitulo.

Fundamentalmente o que podemos reter do que atrás se disse pode-se resumir da seguinte forma:

· Se a amostra for retirada de uma população em que as variáveis em estudo tenham uma distribuição (garantidamente) normal, então qualquer que seja a dimensão da amostra, a distribuição amostra de é igualmente normal.

· No caso de não ser possível garantir essa normalidade, entre em acção o chamado efeito do limite central, e desde que a amostra seja suficientemente grande, a distribuição será aproximadamente (assintoticamente) normal.

· A qualidade da aproximação depende da dimensão da amostra n

È isto que iremos visualizar.

Na Figura 1.2 apresentamos a forma de quatro distribuições, a saber a Uniforme, a Exponencial, a Binomial resumidamente chamada de 0- 1. e a Normal

Repare-se que estas distribuições (à excepção da normal logicamente) são bastante "anormais" pelo que foram escolhidas de propósito para experiências de amostragem

Na Figura 1.3 mostramos a distribuição das médias amostrais de 500 amostras retiradas de uma população normal; note-se que apesar da amostra ser de somente 5 unidades a forma da distribuição amostral das médias é normal, como se pode observar pela curva sobreposta

Repare-se agora na Figura 1.4e na Figura 1.5

Aqui temos amostras igualmente de 5 unidades, de populações com distribuição 0-1 e uniforme.

Note-se o enviesamento da distribuição amostral das médias, revelando que nestas condições as coisas se comportam de forma muito pouco regular.

Reparemos agora na Figura 1.6e na Figura 1.7

Figura 1.2 Formas de distribuições populacionais

Voltamos a retirar um conjunto de 50 amostras das mesmas populações: 0-1 e uniforme.

Só que desta vez temos amostras com 50 elementos cada, ou seja aumentamos dez vezes a dimensão de cada amostra.

O resultado é visível: aproximamo-nos claramente de uma distribuição com um aspecto bastante mais normal, embora não possamos em verdade dizer que se trata de uma distribuição absolutamente normal.

Figura 1.3 Amostras de população Normal

É perfeitamente visivel: a melhoria substancial da aproximação é evidente; e como dissémos á medida que a dimensão da amostra for crescendo, a aproximação vai melhorando igualmente.

Figura 1.4 Amostras de população 0-1

Figura 1.5 Amostras de população uniforme

Figura 1.6 Amostras de população 0-1 (50 unidades)

Figura 1.7 Amostras de população uniforme (50 unidades)

.

Caixa 3: Estimação por máxima verosimilhança:

O método de estimação por máxima verosimilhança baseia-se em conceitos queilustraremos por facilidade com um exemplo, retirado de Rothman (1998).

Imaginemos que observamos uma amostra de 1000 individuos para a pesquiza de HIV. O que será mais provável? Encontrar 10, 4 100?

Como se pode ver a questão qeu aqui se pôe é encontrar o estimador a partir dos dados observados.

1º Passo: Construção da função de verosimilhança (likelihood)

Seja X a variável aleatória definida com o parâmetro m, tendo sido retiradas várias amostras independentes e idênticamente distribuidas X1, X2, X3,....,Xk

Nessas mesmas amostras foram observados os valores x1, x2, x3,....,xk para a variável aleatória. O que nos interessa agora é determinar a probabilidade conjunta de:

P(X1=x1),P(X2=x2),P(X3=x3),.....,P(Xk=xk)

A teoria das probabilidades ensina-nos que a probabilidade conjunta é dada pelo somatório do produto das probabilidades individuais, o que se designa em notação por piatório P

Então definimos a função de verosimilhança como:

L(X|t)=Pf(xi|t)

2º Passo: Logaritmo da função de verosimilhança

Como a maioria das funções estatísticas envolvem exponenciais trona-se mais prático trabalhar sem a sua presença, pelo que tomando logaritmos libertamos a expressão das exponenciais; por outro lado como a logaritmização é uma operação monótona não altera o sentido do resultado final.

3º Passo: Encontrar o valor que maximiza a função log-verosimilhança.

Da álgebra retiramos que a função atinge o seu máximo (oun o seu mínimo) no ponto em que a derivada é zero. Logo calculamos a derivada da função e igualando-a a zero obtemos o valor desejado.

Voltando ao nosso exemplo:

Considerando a fórmula da distribuição binomial com amostra de dimensão N, nimero de casos observados y e o parâmetro p

Logo

Seguindo os passos definidos calculamos o valor que maximiza L(p). Feitos os cálculos encontramos o valor de p=0,001

Calculado este valor podemos verificar que a função de verosimilhança tem o valor máximo de:

Este é assim o valor máximo da função.

Imaginemos agora que queriamos testar a hipótese de p=0.005

Define-se então a razão das verosimilhanças, entre o valor hipotético e o valor estimado pelo método da máxima mverosimilhança

Neste caso daria: 0,00334/0,3681=0,0908

Só cerca de 9% dos dados seriam explicados por este valor de p

Ou seja, este valor suporta mal os dados.

Nem sempre os métodos algébricos podem ser práticamente utilizados para este cálculo.

Assim os com putadores utilizam métodos iterativos, iusto é, estimam vários valores para o parâmetro, introduzindo-os

no modelo e vendo qual o que melhor suporta os dados. Quando o encontra termina o processo.

Os métodos e algoritmos de computação estão completamente fora do âmbito deste tipo de texto introdutório

Capítulo 4

Testes de Hipóteses

4.1 Introdução

Os testes de hipóteses realizam-se regularmente ao longo de toda a investigação; testamos se o valor da média de uma variável é igual a um valor teóricamente definido, ou testamos se os resultados de dois tratamentos se podem considerar idênticos.

A discussão teórica dos testes de hipóteses levar-nos-ia longe, mas vamos, como é hábito, abordar a sua formulação geral de uma forma muito elementar, somente para fornecer ao leitor bases para a compreensão da sua interpretação prática.

O Teste de Hipóteses

Reparemos no seguinte problema prático:

A um grupos de doentes aleatóriamente escolhidos foi realizado um teste de diagnóstico para despiste de lesão cardíaca. Anotaram-se os resultados.

Contudo posteriormente foi possível confirmar se os diagnósticos tinham sido correctamente efctuados.

Podemos afirmar que o teste é eficaz?

A tabela sumariza o problema

Tabela 1.1 Testes de Hipóteses:

Teste	Confirmação	
	Positiva	Negativa
Positivo	a	b
Negativo.	c	d

Como observamos existem três situações diferentes.

Nos casos a e d a concordância é perfeita; o teste diagnosticou correctamente a doença.

No caso c o teste deu negativo, quando de facto se estava em presença da doença, e no b foi o inverso: o teste disse positivo mas a doença estava ausente.

Imaginemos agora que em vez de tratamento tínhamos previamente formulado uma hipótese de trabalho (ex. não haver diferenças entre dois tratamentos) e estávamos a testar os resultados a fim de verificar se podíamos aceitar a nossa hipótese nula, ou antes optar por uma hipótese alternativa (haver diferenças de tratamentos).

Nessa situação corríamos dois riscos: o primeiro representado pela situação c, ou seja o de não aceitarmos a hipótese nula quando ela é verdadeira (erro de tipo I ou erro a) e o segundo representado pelo caso b (erro de tipo II ou erro b).

Estes dois erros acompanham-se sendo impossivel elimina-los completamente. Quando diminuimos o erro I aumentamos o II e vice-versa.

Ora quando se realizam testes de hipóteses temos de estabelecer qual o valor máximo aceitavel dos erros I e II.

Usualmente define-se o valor para o erro I (1%, 5%), denominando este valor de nivel de significância.

O erro de tipo II é mais complexo, e o seu manuseamento ultrapassa este texto introdutório.

Assim quando realizamos um teste de hipóteses temos de observar os seguintes passos:.

· Formulação da hipótese nula
· Escolha do teste adequado
· Fixação do nivel de significância
· Interpretação.

O primeiro e terceiro ponto estão abordados. O segundo ponto será visto á frente, conforme os testes a usar.

Vejamos agora o problema da interpretação.

De acordo com aquilo que dissemos atrás podemos deduzir o seguinte:

Uma vez realizado o teste escolhido, este produzirá um dado resultado.

O que pretendemos saber é se este resultado nos permite assumir a hipótese nula, ou não.

Ora bem: ao estabelecermos o nivel de significância (p. ex. a 5%) estamos a dizer que não queremos correr um risco de mais de 5% de nos enganarmos; ou seja dito de outro modo: não queremos correr um risco maior que 5% de rejeitar a hipótese nula quando ela é verdadeira.

Ora o resultado do teste realizado será um valor numérico que corresponderá ao percentil da distribuição que o teste seguir.

Daqui em diante só temos que verificar se esse valor encontrado está acima ou abaixo do limite por nós estabelecido.

Se estiver abaixo dizemos que está na região de aceitação e temos de aceitar a hipótese nula.

Se estiver acima rejeitamos a hipótese nula.

A figura seguinte ajuda a visualizar estas ideias.

Se entre A e B está 95% da probabilidade (o inverso dos 5%, também chamado nível de confiança) então o resultado C "cai" na área de aceitação e o resultado D claramente na área de rejeição.

p-value

A forma anterior de apresentar os resultados torna-se um pouco complexa, uma vez que para cada teste de hipótese teriamos de após saber o resultado consultar a tabela da distribuição e verificar qual a probabilidade que correspondia ao respectivo percentil.

Por isso os computadores passaram a fornecer com cada resultado a respectiva probabilidade associada, que se denomina p-value.

Para efeitos práticos seguimos a seguinte regra:

Sempre que o p-value fôr superior a 0,05, ou 0,01 -conforme o nivel de significância (95% ou 99%)-, aceitamos a hipótese nula.

De outra forma rejeitamos.

Estas simples bases vão-nos permitir compreender e interpretar os testes de hipóteses que iremos analisar ao longo deste texto.

Como habitualmente quem estiver interessado em aprofundar os aspectos teóricos dos testes de hipóteses, nomeadamente determinação de região crítica, etc. poderá consultar as obras de referência, muito em especial a obra de Tannis and Hogg () ou o livro de Bento Murteira ().

Usando o STATA

Para terminar este capítulo vamos observar o STATA resolvendo testes de hipóteses, de duas formas diferentes.

Primeiro vamos utilizar uma folha de dados, e depois vamos utilizar os comandos interactivos.

1º Caso: Comparação de médias em dois grupos independentes a partir de uma folha de dados

Trata-se de uma folha de dados, que acompanha o STATA em que se registam os resultados de um estudo de sobrevivência (48 observações) de doentes portadores de neoplasia, que tomaram dado medicamento.

Vamos atarvés de um teste de hipóteses tentar verificar se existe diferença "estatísticamente significativa" entre a média de idades dos doentes que faleceram e os que se mantiveram vivos no fim do ensaio.

Para tal seguiremos o seguinte caminho nos menus do STATA: Statistics-Parametric Tests-Two-sample test; a caixa seguinte pode suscitar um pouco mais de confusão, uma vez que o programa pergunta se se trata de: 1data variable, 1group variable, ou 2 independent data variables.

Esta questão tem sómente a ver com a forma como os dados estão organizados. Reparese na tabela seguinte:

Age	Death	Death	Live
54	0	54	58
58	1	74	67
67	1		
74	0		

Na tabela da esquerda temos uma variável (age) com todos os casos e uma outra variável (death) que assume dois valores 1 e 0, conforme o doente sobrevive ou não; na tabela da direita temos duas variáveis (death) e (live), que contêm os casos que sobreviveram e os que faleceram. é isso que o STATA nos pergunta. Como estão os dados organizados?

- Se fôr como na tabela do lado esquerdo escolhemos a 1ª opção da pergunta;
- Se fôr como no lado direito escolhemos a 2ª opção.

Note o leitor que esta questão aparecerá em várias caixas de diálogo, sempre que o programa necessite conhecer a organização dos dados para prosseguir os testes.

Atente portanto bem neste aspecto que lhe poderá ser útil mais á frente.

De seguida surge uma caixa para escolher quais as variáveis de dados e quais as variáveis de grupo que mostramos na Figura 1.1

Figura 1.1 Caixa de diálogo do STATA para o teste de hipóteses

Note-se o pormenor de nos ser pedido quer o nível de confiança a que pretendemos o teste e a possibilidade de indicar se a variância é desigual nas duas populações (a analizar préviamente) como atrás já foi aflorado e adiante se voltará a observar.

Os resultados são os do quadro seguinte:

Two-sample t test with equal variances

Group Obs Mean Std. Err. Std. Dev 95% CI

0 17 54.17647 1.317962 5.434097 51.38252-56.97043

1 31 56.80654 1.014327 5.647533 54.73492-58.87798

combined 48 55.875 .8168358 5.659205 54.23174-57.51826

diff -2.629981 1.682285 6.016245 -6.016245-.7562825

Degrees of freedom 46

Ho: mean(0) - mean(1) = diff = 0

Ha: diff < 0 Ha: diff ~= 0 Ha: diff > 0

t = -1.5633 t = -1.5633 t = -1.5633

P < t = 0.0624 P > |t| = 0.1248 P > t = 0.9376

Asua interpretação que poderá parecer um pouco confusa é contudo bem simples.

Em primeiro lugar diz-nos que se trata de um teste t de duas amostras, com variância igual, o que deveria ter sido testado por nós préviamente com qualquer dos métodos atrás estudaddos. A análise do quadro confirma aliás esse pressuposto, uma vez que o desvio-padrão das duas amostras é práticamente igual.

Claro que aqui temos igualmente de admitir a independência ds duas amostras, ou seja que os doentes de uma eram diferentes dos doentes da outra e que o fenómeno em estudo "óbito" era independente de se pertencer a uma ou outra das amostras.

As coisas poderão não ser completamente assim, mas para fins académicos a situaçao presente serve.

A dsitribuição t é neste caso óbvia, uma vez que a amostra é pequena e nada sabemos sobre a variância populacional, a não ser através do seu estimador, o desvio-padrão amostral.

Na parte final do quadro, apresentanos a hipótese nula da diferença das duas médias ser igual a 0 e em seguida apresenta-nos o valor do teste t unilateral e bilateral (valor do meio).

Como nos interessa saber se existe diferença entre as duas médias, (seja ela par amias ou para menos) vanos ler o valor do centro do quadro. Por baixo vem o correspondente p-value de 0.1248, que por ser muito maior que 0.05 nos leva a ter que aceitar a hipótese nula, de não haver diferença na média das idades dos doentes que faleceram e, dos que ficaram vivos.

2º Caso: Comparação de médias em dois grupos independentes a partir da linha de comandos

Esta situação pode ser muito comum e a aptidão do STATA para trabalhar em linha de comandos particularmente útil.

Imagine-se que o leitor está a analizar um artigo científico e pretende confirmar algum dos aspectos do mesmo. Lógicamente que nesse artigo só excepcionalmente virá a folha de dados que serviu de base ao mesmo. Quando muito o autor apresenterá os resultados de uma forma resumida, o que tornará impossível, ao leitor qualquer análise pelos métodos clássicos. Vejamos um exemplo extraido de Fisher e van Belle (1993).

Trata-se de um estudo de Holtzman et all (1975) em que os autores analisaram a relação entre uma dieta pobre em fenilalanina em crianças de 4 anos afectadas de fenilcetonúria. as crianças foram divididas em dois grupos; um terminou a sua dieta e outro prosseguiu; aos 6 anos os níveis de fenilalanina foram determinados com os seguintes resultados, constantes daTabela 1.2 :

Tabela 1.2 Doseamento da fenilalanina

Dieta Terminada Dieta Continuada

N^o de crianças 5 4

Mean da fenilalanina(mg/dl) 26.9 16.7

Desvio-Padrão 4.1 7.3

Com estes dados e, se pretendermos saber se existe diferença estatísticamente significativa entre os dois grupos teremos de utilizar o comando interactivo do STATA com a sintaxe:

ttesti obs1 média1 sd1 obs2 media2 sd2, unequal level()

O uso da opção unequal prende-se com o facto de se desconhecer a variância populacional e também para colocar o teste nas suas condições de execução mais gravosas, a fim de reforçar o valor da inferência; no nosso caso seria:

ttesti 5 26.9 4.1 4 16.7 7.3, unequal level(95)

E o resultado:

Two-sample t test with unequal variances

Obs Mean Std. Err. Std. Dev 95% CI

x 5 26.9 1.833576 4.1 [21.80918 31.99082]

y 4 16.7 4 3.65 7.3 [5.084071 28.31593]

combined 9 22.36667 2.522978 7.568933 [16.54867 28.18466]

diff 10.2 4.084666 [-.668460721.06846]

Satterthwaite's degrees of freedom: 4.49069

Ho: mean(x) - mean(y) = diff = 0

Ha: diff < 0 Ha: diff ~= 0 Ha: diff > 0

t = 2.4971 t = 2.4971 t = 2.4971

P < t = 0.9699 P > |t| = 0.0602 P > t = 0.0301

A leitura é em tudo semelhante ao quadro anterior e a interpretação de igual modo; com um p-value de 0.0602 não podemos rejeitar a hipótese nula de não haver diferença entre as duas dietas.

Contudo este estudo tem um problema relacionado com a sua ausência de potência. De facto com amostras tão pequenas a potência deste teste é fraca (70.44%) a que corresponde um erro tipo II de 0.2956, que é elevado, resultando assim numa inferência fraca, de que se pode suspeitar por um p-value de 0.06, tão próximo do valor limite.

Neste caso seria recomendável tentar uma amostra de maiores dimensões.

Os cáluclos da potência do teste e do correspondente erro b foram igualmente realizados com o STATA, mas estão fora do âmbito deste texto introdutório, embora voltemos a fazer referência a eles mais á frente.

Capítulo 5

Intervalos de confiança

5.1 Introdução

Em capítulos anteriores falámos de estimadores e de estimação e apresentámos, de uma forma muito ligeira, um dos métodos de estimação pontual, o da máxima verosimilhança.

Este método está voltado para os dados e permite encontrar qual o estimador, e a respectiva estimativa, que melhor suporta o conjunto das nossas observações.

Outra forma muito usual de estimar é pelo método dos intervalos de confiança. Toda a gente conhece a nomenclatura usual em todos os trabalhos científicos em que se apresenta uma dado estimador (p. ex. uma média amostral) á qual se acrescenta o respectivo intervalo de confiança, como quando se diz que a prostaglandinémia basal em grávidas submetidas a indução do trabalho de parto com PGE2 intracervical, expressa em pg/ml é de (Cruz A.S et. coll1995).

Note-se que aqui, já possuimos um estimador pontual, a média, mas que pretendemos determinar um "espaço" dentro do qual se espera que por multiplas repetições da experiência aleatória os sucessivos valores encontrados venham a estar incluidos.

Note-se que existe uma proximidade entre a matéria deste en do próximo capítulo, pelo que a compreensão global destes tópicos necessitará da leitura de ambos.

Vamos tecer algumas considerações sobre a interpretação dos intervalos de confiança, apresentar o seu modelo de cálculo nos casos mais usuais utilizando o STATA para proceder à respectiva computação.

No final do capítulo arbodaremos brevemente o cálculo da dimensão da amostra tanto para médias, como para nproporções

Interpretação de um intervalo de confiança

Como vamos ver a seguir no cálculo de um intervalo de confiança intervêm os seguintes elementos:

- A variância do estimador;
- O nível confiança a que pretendemos efectuar o teste (90, 95%, etc.)
- A distribuição amostral do estimador

· A dimensão da amostra

Os dois ultimos aspectos, a distribuição amostral do estimador e a dimensão da amostra estão intimimamente ligados entre si, uma vez que o uso de uma (distribuição) depende do valor da outra (dimensão).

A interpretação do intervalo de confiança deve assim ser feita de uma forma prudente e equilibrada, lembrando que a sua dimensão vari em função inversa do respectivo nivel de confiança préviamente definido.

Assim se pretendermos um intervalo de confiança a, por exemplo, 99% iremos obter um valor muito mais alargado do que se o fizermos a 90%; ou seja em linguagem mais simples.

Quanto maior fôr o risco de errar que estivermos dispostos a correr, mais estreito será o intervalo de confiança.

No limite se quizermos correr um risco mínimo, poderemos obter um intervalo de confiança de tal forma largo que pouca utilidade tenha para nós.

A discussão do interesse dos intervalos de confiança, nomeadamente em estudos epidemiológicos pode ser seguida em Rothman (1998) que apresenta uma curiosa discussão sobre este tema.

Intervalo de confiança da média

Por agora vamos analisar o caso típico revendo de seguida o formulário para os restantes casos:

Caso 1: Intervalo de confiança da média de uma amostra de uma população normal em que a variância é conhecida.

Esta é de longe a situação mais fácil, mas nem sempre a mais usual. temos tudo o que precisamos para calcular o intervalo de confiança para a a média; e uma vez que asseguramos a distribuição normal da variável (a qual pode ter sido analizada com os métodos gráficos atrás descritos) então só temos de estabelecer o nível de confiança com que vamos trabalhar (90%, 95%, 99%, etc.)

Então a fórmula geral será:

A notação refere-se á distribuição normal e ao valor do nivel de significância a (5%) complementar do nível de confiança (95%), assunto a esclarecer melhor no próximo capítulo.

O exemplo seguinte dá-nos o "output" do STATA para o cáculo do intervalo de confiança de uma folha de dados com 58 observações da variação da pressão sistólica.

Variable | Obs Mean Std. Err. [95% Conf.Interval]

———+—————————————-

systolic | 58 18.87931 1.680837 15.51349 22.4513

Como vemos se repetissemos a experiência n vezes iriamos encontrar valores no intervalo 15,51349 e 22,4513

Caso 2: Intervalo de confiança da média de uma amostra de uma população normal em que a variância é conhecida.

Neste caso, e devido ao facto de se desconhecer a variância populacional é substituida a distribuição normal (Z) pela distribuição t de Student (t) na fórnula anterior. Tudo o mais é idêntico e a fórmula é do tipo:

O número de graus de liberdade da distribuição t é n-1, ou seja o número de observações menos 1.

Caso 3: Intervalo de confiança da média de uma amostra de uma população não normal em que a variância é desconhecida, mas a amostra é de grandes dimensões.

Segue exactamente o formulário do caso anterior, á condição, da amostra ser de grandes dimensões.

Intervalo de confiança de uma proporção

Frequentemente desejamos estimar uma proporção populacional. Poderemos querer saber a proporção de utentes do sexo feminino que frequentam dada consulta, ou a proporção de doentes que melhorou a sua glicemia com o uso de dado hipoglicemiante.

Recordemos que p é a proporção na população e a proporção amostral.

Esta é uma estatística e possui uma distribuição amostral que para amostras grandes:

· É aproximadamente normal

· A média m da distribuição amostral de é igual á proporção populacional p

· O desvio-padrão s da distribuição amostral de é dado por com q=1-p

Nota: Recordemos que no caso das proporções uma amostra é considerada grande se np e nq forem ambos maiores que 5.

Para facilitar as notações vamos denominar a proporção populacional p e a proporção amostral como p.

O intervalo de confiança será portanto da forma

Intervalo de confiança da diferença de duas médias

Iremos considerar aqui as situações mais usuais, em que se admite que as variáveis em estudo têm na população distribuição normal, ou então embora não podendo garantir tal facto, as amostras são de dimensão suficientemente grande.

Vamos igualmente considerar que desconhecemos as variâncias populacionais, uma vez que só excepcionalmente tal facto será do nosso conhecimento.

Consideraremos então em primeiro lugar a situação em que, embora desconhecendo essaqs variâncias populacionais, admitimos que elas sejam iguais e, de seguida o caso em que sejamos levados a suspeitar que elas sejam desiguais.

Não abordaremos aqui casos mais oblíquos, tais como variâncias desconhecidas e desiguais em amostras de pequenas dimensões, uma vez que nessas sistuações teremos de recorrer a métodos mais complexos completamente fora do âmbito deste texto.

1^{o} Caso: Populações normalmente distribuidas ou amostras de grandes dimensões com variâncias supostamente iguais.

O cálculo do intervalo de confiança de uma média segue sempre o método descrito na secção anterior.

O que varia é o desvio-padrão amostral, a que em bom rigor se chama erro-padrão que no caso presente será baseado na média das variâncias das duas amostras com ponderação para a dimensão das mesmas.

O erro-padrão necessário, será assim obtido através de uma fórmula mais complexa, do que para um caso isolado. Calculemos préviamente a variância conjunta, s2, como:

com s1 e s2 os desvio-padrão dos grupos de dimensão n1 e n2

Usando e como notação para as médias das duas amostras e se como o desvio-padrão conjunto (raiz quadrada do valor acima s2) podemos calcular o intervalo de confiança para a diferença das médias da forma:

com o erro-padrão de e a distribuição t com n1+n2-2 graus de liberdade.

2º Caso: Populações normalmente distribuidas ou amostras de grandes dimensões com variâncias supostamente iguais.

Iriamos perder algum tempo e complicar as coisas se tentássemos uma explicação minuciosa da solução deste problema; no fundo ele resume-se ao anterior, mas com algumas nuances importantes.

Assim será util investigar quão diferentes são as variâncias das duas amostras. Lembramos que a variância s2 tem uma distribuição de c2 com n-1 graus de liberdade, e que o quociente de duas c2 é uma F com n1-1 graus de liberdade no numerador e n2-1graus de liberdade no denominador.

Assim podemos fazer um teste de hipóteses para verificar se podemos aceitar, ou não a hipótese nula desse quociente ser igual a 1, ou seja as variâncias iguais.

Isto pode ser feito no STATA recorrendo ao comando sdtest, com a sintaxe sdtest var1=var2, ou recorrendo ao menu Calculator-2 sample test of variance.

Se as variâncias forem claramente desiguais procedemos como da forma anterior, só que teremos que usar formulas alternativas do cálculo do estimador do erro-padrão de ; este será dado pela fórmula:

Quanto aos graus de liberdade, serão igualmente utilizadas formas alternativas como as de Satterwhaite ou a de Welch, cuja complexidade está fora do âmbito deste texto.

Recomenda-se mais uma vez a leitura dos textos já referidos par quem queira aprofundar esta matéria.

Recordemos que a primeira já foi utilizada anteriormente, quando do exemplo de um teste de hipóteses nestas condições, que o leitor poderá consultar na pagina 1-13

Bastará para tanto quando digitar o comando no ttest não esquecer de acrescentar a opção unequal e igualmente a opção Welch, caso pretenda utilizar esta fórmula; se nada a crescentar o STATA usa a fórmula de Satterwhaite. Não parece haver vantagem especial pelo que por uma questão de economia de digitação poderá deixar o programa com a opção por omissão.

Intervalo de confiança da diferença de duas médias em amostras emparelhadas

Os casos que apresentámos atrás referem-se a amostras independentes retiradas de populações em que as variáveis se encontram identicamente distribuidas.

Contudo acontece frequentemente que temos de lidar com casos em que em vez de amostras independentes temos duas colheitas de dados feitas a partir da mesma amostra, por exemplo valores laboratoriais colhidos de doentes antes e depois de administrado dado medicamento.

Neste caso trata-se de duas amostras colhidas de um mesmo grupo em estudo.

Não podemos aqui utilizar a metodologia anterior sem violarmos o principio da independência, pelo que temos de recorrer a outra metodologia.

A melhor forma de apreender a metodologia será recorrendo a um exemplo, este retirado de Altman (1991).

Trata-se de uma folha de dados com o aporte calórico médio diário (em kJ) de 11 mulheres durante os 10 dias pre e post-menstruais (Maznochan et all.). A Tabela 1.1apresenta os dados

Tabela 1.1 Aporte calórico diário pré e post menstrual

Aporte Calórico (kJ)

Caso	Pre-menstrual	Post-menstrual	Diferença
1	5260	3910	1350
2	5470	4220	1250
3	5640	3885	1755
4	6180	5160	1020
5	6390	5645	745
6	6515	4680	1835
7	6805	5265	1540
8	7515	5975	1540
9	7515	6790	725
10	8230	6900	1330
11	8770	7335	1435
Média	6753.6	5433.2	(a) 1320.5
DP	1142.1	1216.8	366.7

Como podemos observar trata-se de duas observações feitas no mesmo grupo de mulheres, só que em tempos diferentes; logo estamos perante um caso de amostras emparelhadas.

O que pretendemos é calcular um intervalo de confiança para a média da diferença entre as duas observações, no nosso caso (a) - ver Tabela 1.1; vamos chmar-lhe .

Então a média da diferença das amostras emparelhadas será dada por:

e o desvio-padrão por:

Na maioria dos casos as amostras são de pequena dimensão (como no caso presente), a que acresce que nada sabemos sobre a variância populacional da diferença das médias.

Assim assumindo que esta tem na população uma distribuição (aproximadamente normal) utilizamos a distribuição t com n-1 graus de liberdade para fazer inferências sobre md. Como nada sabemos sobre sd então o desvio-padrão de é dado por

Voltando ao nosso exemplo temos:

=1320.5

sd=366.7

Logo o intervalo de confiança a 95% para a diferença destas médias será de:

Note-se que o valor 2,28 corresponde ao quantil da t com 10 graus de liberdade, para um teste bilateral.

Mais uma vez podemos recorrer ao STATA para nos fazer os cálculos respectivos. Eis o resultado usando o menu Statistics-Parametric Tests-Paired t Tests

Paired t test.

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
Pre	11	6753.636	344.3631	1142.123	[5986.348	7520.925]
Post	11	5433.182	366.8888	1216.833	[4615.703	6250.661]
diff	11	1320.455	110.5779	366.7455	[1074.072	1566.838]

Ho: mean(Pre - Post) = mean(diff) = 0

Ha: mean(diff) < 0   Ha: mean(diff) ~= 0   Ha: mean(diff)> 0

t = 11.9414   t = 11.9414   t = 11.9414

P < t = 1.0000   P > |t| = 0.0000   P > t = 0.0000

Se atentarmos na linha realçada verificaremos que os resultados são absolutamente coincidentes com os obtidos através do nosso laborioso esforço aritmético.

Mas note-se noutro pormenor; o programa fez igualmente um teste de hipóteses da diferença das médias ser 0. É que na realidade intervalos de confiança e testes de hipóteses encomtram-se particularmente próximos pelo que calculando um fácilmente se chega ao outro.

No final do capítulo apresentaremos uma breve discussão sobre o uso de uns e outros.

Intervalo de confiança para a diferença de duas proporções

Como já fizemos atrás denominemos a proporção populacional de p e a proporção amostral por p.

Para duas amostras independentes e suficientemente grandes a distribuição amostral de p1-p2 é aproximadamente normal com a média

e o desvio-padrão

com q=(1-p)

Daqui fácilmente deriva que o (1-a)100% intervalo de confiança para p1-p2 será da forma:

com

Para além destes intervalos de confiança é possível calcular outros intervalos de confiança, nomeadamente para proporções em amostras emparelhadas.

Contudo a vontade de não adensar este texto leva a deixar de fora aspectos importantes na manipulação de proporções como os problemas da correcção de continuidade da utilização do teste exacto de Fisher, do teste de McNemar, etc.

Essas matérias poderão ser consultadas em qualquer dos livros de referência, nomeadamente Fleiss() Altman(op. cit.) Fisher & van Belle (op. cit) Kleinbaum (op. cit.), etc.

Quem estiver interessado tem no final uma vasta e qualificada bibliografia de fácil escolha.

Cálculo da dimensão da amostra para a estimação da média

Calcular a dimensão de uma amostra é sempre um processo importante, na medida em que trabalhar com amostras de pequenas dimensões enfraquece a inferência, enquanto que trabalhar com amostras de excessivas dimensões encarece, por vezes desnecessáriamente o estudo.

Vamos dar uma ideia muito simples de como tal se pode conseguir.

O que necessitamos saber é o desvio-padrão na população da variável em estudo, ou em sua substituição -com os perigos usuais de enfraquecimento da inferência- o desvio-padrão amostral.

Par além disso necessitamos predeterminar o valor máximo do afastamento do real valor do parâmetro que admitimos aceitável e, que definimos como, o erro máximo da estimativa de m ,E da forma:

recordando que o desvio padrão da média amostral é igual a

Suponhamos agora que predeterminamos o erro máximo e queremos saber a dimensão da amostra que produzirá esse erro máximo. Da expressão acima deriva a fórmula seguinte para cáculo de n z corresponde ao quantil da distribuição normal correspondente ao erro a que determinamos préviamente.

Vejamos um exemplo, imaginário.

Um clínico pretende determinar a dimensão de uma amostra, que a 99%, não se afaste mais de 1% do valor médio da glicémia dos portugueses, sabendo que o desvio-padrão é de 1.25.

Logo E=0.01 e n será:

Como se vê com tão duras condições a amostra mostra-se de dimensões enormes (quais 109000 casos); contudo se diminuissemos as restrições, p. ex. 95% com um E=0,05 já iriamos para um valor de n=2401.

Como se vê tudo depende do risco que pretendemos correr.

Se formos muito exigentes temos amostras grandes.

Se facilitarmos um pouco as amostras diminuem; contudo a força da inferência diminuirá igualmente.

Cálculo da dimensão da amostra para a estimação de uma proporção

Os passos a seguir neste caso são em tudo iguais aos da secção anterior.

Será contudo necessário substituir na fórmula por o qual é como se sabe

Daqui deriva que:

Estas fórmulas de cálculo, simplificadas servem para dar uma ideia da forma como se pode calcular a dimesnão de uma dada amostra, conhecendo os valores referidos.

Recorrendo ao STATA é possivel usar um comando interactivo sampsi que facilita o cálculo da dimensão de amostras e da potência de testes comparando as médias de duas amostras, a média de uma amostra com um valor teórico, proporções de duas amostras e uma proporção com um valor teórico.

Veremos, á frente, a utilização prática prática destes métodos.

"Manual"
2000/12/29
page 58

Capítulo 6

Tabelas de contingência

6.1 Introdução

A maioria dos estudos clínicos ou epiderniológicos apresenta os seus resultados sumarizados sob a bem conhecida forma de tabelas de contingência.

Estas que poderão conter várias linhas e colunas, apresentam normalmente nas colunas os tratamentos ou a exposição e nas linhas os resultados desses tratamentos ou exposições

Como a maioria das variáveis utilizadas em clínica e epidemiologia são habitualmente categoriais, ordinais ou mesmo nominais o seu estudo e análise necessita de técnicas apropriadas as quais podem ser mais fácilmente aplicadas em dados insertos em forma tabular.

Por uma questão de abordagem mais simples iremos em primeiro lugar considerar as tabelas de contingência com duas linhas e duas colunas, ditas 2x2.

Contudo o que dissermos poderá ser extensivo a tabelas, com n linhas e k colunas não perdendo portanto nada do seu sentido.

Iremos por outro lado analisar as tabelas 2x2 inseridas, no contexto dos ensaios clínicos (tratamentos).

Referiremos urna aplicação importante na análise destas tabelas sob certas circunstâncias; o teste ou índice de Kapa.

Abordaremos em seguida as tabelas na perspectiva epidemiológica (exposição/risco).

Terminaremos a secção com a demonstração da aplicação do STATA neste campo.

A Abordagem Clínica

Teste de Independência

Atentemos no exemplo seguinte:

Na Tabela 1.1 observamos a acção de dois tratamentos A e B (placebo) na redução da pressão arterial de um conjunto de doentes, sendo que o critério de "redução da pressão arterial" foi definido no inicio do estudo.

Admita-se igualmente que este conjunto de doentes é constituído por dois subgrupos de doentes em tudo iguais entre si, excepto no que aos tratamentos concerne.

Tabela 1.1 Comparação de tratamentos A e B

A B Total

Não reduziu	506	143	649
Reduziu	624	161	785
Total	1130	304	1434

O investigador encarregue do estudo em causa pretendia averiguar se a redução da pressão arterial observada era independente do tratamento ser A ou B ou, preferencialmente se existia associação entre o tratamento A e a redução da pressão arterial.

Para tal irá proceder a um Teste de Independencia.

0 endereçamento deste problema passa por uma formulação mais geral que demonstraremos com outro exemplo.

Na Tabela 1.2 observamos o número de condutores (100) detidos por condução sob o efeito do álcool, distribuídos por vários grupos etários.

Tabela 1.2 Condutores com excesso de álcool por idades

Idades	16-25	26-35	36-45	46-55	56 e >
Nº de condutores	32	25	19	16	8

Formulemos a seguinte questão:

Ao nível de 1% é possível rejeitar a hipótese nula de que a proporção de pessoas detidas é idêntica em todos os grupos etários?

Para responder a essa questão pomos em marcha o seguinte procedimento.

1º Formulação da hipótese nula

Como existem 5 categorias (grupos etários) a proporção de condutores com excesso de álcool seria a mesma para todos eles se cada um contivesse 1/5 do total dos condutores.

Assim as hipóteses nula e alternativa seriam, formuladas desta forma:

Ho: Proporção de pessoas detidas e igual em todos os grupos etários;

H1: Proporção de pessoas detidas não é igual em todos os grupos etários.

Se a proporção de pessoas detidas fosse igual em todos os grupos etários então a probabilidade de um condutor seleccionado ao acaso pertencer a qualquer um dos cinco grupos seria de 1/5=.20 ou 20%.

2º Selecção do teste a utilizar

A experiência realizada é uma experiência multinomial porque envolve mais de 2 opções possíveis (neste caso os 5 grupos etários). Recorde-se que quando uma experiência só envolve 2 opções chama-se experiência binomial, como atrás se referiu, sendo um caso especial de uma experiência multinomial.

Uma experiência multinomial possui quatro características:

Consistem em n experiências idênticas (tentativas)

Cada tentativa resulta num de k possíveis resultados (k>2)

As tentativas são independentes

As probabilidades dos vários resultados mantem-se constantes ao longo das n tentativas

Assim sendo, e adiantando a condição da hipótese nula (P=.20) esperamos obter em cada uma das células da Tabela 1.2 a frequência de 100x0.20=20 detidos.

Esta frequência denomina-se frequência esperada.

A nossa decisão de aceitar ou rejeitar a hipótese nula baseia-se no afastamento das frequência observadas (0) das frequência esperadas (E).

0 teste será então da forma:

Este teste tem uma distribuição de Qui-quadrado com k-1 (k=nº de resultados possíveis) graus de liberdade.

Este teste é sempre unilateral (aba direita).

Uma condição de aplicabilidade do teste de Qui-quadrado indispensável e a de que as amostras sejam suficientemente grandes de modo a que a frequência esperada em cada categoria seja pelo menos 5.

Se assim não acontecer deve-se ou aumentar a amostra ou combinar duas, ou mais categorias.

3º Região critica

No nosso exemplo trata-se de uma distribuição qui-quadrado com 5-1=4 graus de liberdade.

A leitura da tabela do qui-quadrado diz-nos que para 4 g. 1. e a região critica e igual a 13.270.

4º Calculo do teste

A Tabela 1.3 apresenta os cálculos

Tabela 1.3 Cálculo do Qui-quadrado

Idade	FreqO	p	FreqE	(O-E)	(O-E)2	(O-E)2/E
16-25	32	.20	20	12	144	7.20
26-35	25	.20	20	5	25	1.250
36-45	19	.20	20	-1	1	.0.50
45-55	16	.20	20	-4	16	.800
56 e >	8	.20	20	-12	144	7.200
	n=100					16.500

5º Decisão

Como o valor do Qui-quadrado e de 16.50 ultrapassa o valor critico de 13.27 e cai na região de refeição.

Logo temos de rejeitar Ho e declarar que a nível de 1% a proporção de condutores detidos com excesso de álcool não é a mesma em todos os grupos etários.

Ou dito, de outra forma, existe associação entre o grupo aterro e o numero de condutores detidos por conduzirem com excesso de álcool no sangue.

Assentes estas ideias podemos regressar ao problema inicial, (tratamentos e pressão arterial) que no fundo e uma extensão deste que acabamos de resolver.

Recordemos então a marcha geral da solução.

· Formular Ho e H1;
· Calcular as frequências esperadas para cada célula da tabela;
· Determinar o numero de graus de liberdade;
· Encontrar o valor critico associado ao numero de graus de liberdade e ao nível de significancia do teste;
· Calcular o valor do teste de qui-quadrado;
· Decidir conforme ele caia na região de aceitação ou de refeição.

Estes passos foram anteriormente explicados faltando somente resolver os pontos 2 e 3.

Existe uma regra para calcular as frequências esperadas a saber:

A frequência esperada E para uma célula e calculada como:

E=(Total de linha)x(Total de Coluna)/Dimensao da amostra

Na Tabela 1.4 ficara

Tabela 1.4 Frequências esperadas e observadas

A B

N. Desceu 506(O)

(E)=(1130x649)/1434=511.41 143(O)

(E)=(304x649)/1434=137.58 649

Desceu 624(O)

(E)=(1130x785)/1434=618.58 161(O)

(E)=(304x785)/1434=166.41 785

1130 304 1434

0 numero de graus de liberdade para um teste de independência e calculado:

g.I=P(L-1)x(C-1)

No nosso caso com 2 linhas e 2 colunas e 1.

Note-se que aqui igualmente temos de assegurar que a amostra e suficientemente grande para que as frequências observadas em cada célula sejam superiores a 5.

Ora se considerarmos a=0.05 com 1 g.l o valor critico do Qui-quadrado e de 3.841.

Ora o nosso Qui-quadrado e de 0.4942.

Como se vê bem abaixo do ponto critico pelo que temos de concluir não poder rejeitar a hipótese nula de os dois tratamentos A e B, do qual B e placebo, terem efeitos equivalentes.

Teste de Homogeneidade

Num teste de homogeneidade testamos se 2 ou mais populações são homogéneas (similares) em relação à distribuição de uma dada característica ou atributo.

Podemos, por exemplo, estar interessados em testar a hipótese nula da proporção de doentes de 3 grupos sócio-etários definidos (crianças, activos, idosos) ser equivalente nos ficheiros dos Dr. A e B. do Centro de Saúde.

Vejamos como:

Admitamos que retiramos urna amostra de 250 fichas do ficheiro do Dr. A e 150 fichas do ficheiro do Dr. B.

Os resultados estão resumidos na Tabela 1.5

Tabela 1.5 Grupos sócio-etários por ficheiro

A B

Crianças 70 34 104

Activos 80 40 120

Idosos 100 76 176

250 150 400

Procedemos então ao teste de homogeneidade a a=0.05

1º Formulação da hipótese nula

Ho=A proporção dos doentes dos três grupos sócio-etários e igual nos dois ficheiros

H1 =A proporção dos doentes dos três grupos sócio-etários não é igual nos dois ficheiros

2º Distribuição a usar

A distribuição a usar e a de Qui-quadrado

3º Determinação da região critica

Com (3-1)*(2-1)=2 g.l. e com a=0.05 a tabela de Qui-quadrado da-nos 7.378.

4º Calculo do valor do teste

0 valor do Qui-quadrado calculado de acordo com o método referido, atrás e de (70-65)2/65+(34-39)2/39+(80-75)2/75+(40-45)2/45+(100-100)2/110+(76-66)2/ 66=4.339

5º Decisão

Como o valor encontrado e inferior ao valor critico, não podemos rejeitar a hipótese nula e afirmar que ambos os médicos tem ficheiros que sob o ponto de vista dos grupos sócio-etários são homogéneos, isto e idênticos.

Estes testes são as bases da utilização das tabelas de contingência na clínica.

Iremos a seu tempo introduzindo outras aplicasses mas que se baseiam nesta estrutura base.

Lembramos ainda que não será necessário fazer os cálculos atrás apontados, uma vez que o STATA procede a analise das tabelas de forma imediata, permitindo varias opções que vão para alem do calculo do teste do qui-quadrado como no final do capitulo veremos.

Não se preocupe portanto, mais uma vez, o leitor.

0 Teste de Kapa

Uma aplicação importante a utilizar em tabelas 2x2 é o teste de Kapa, igualmente por vezes denominado indica de Kapa cuja abordagern iremos fazer na vertente prática deixando as considerações te6ricas para as referências na matéria nomeadamente Hui S. e Walter Sd (1980), Landis e Kock (1975) para alem do já citado Fleiss (—).

Esta técnica pretende abordar um problema muito comurn em estudos clínicos, que vamos exemplificar com um trabalho do autor, Botelho de Sousa,(1990).

Neste trabalho pretendia-se comparar o valor diagnóstico dos testes mais usuais (na altura) no estudo da patologia gástrica.

Note-se que estávamos numa fase emergente da endoscopia, que não era utilizada com a frequência que é hoje, e o que se pretendia conhecer era o grau de concordância diagnóstica dos dois testes, quer no diagnóstico da úlcera, quer da neoplasia gástricas.

Ora o problema que existia, –e que pode existir em muitos estudos comparativos deste tipo–, era o de não possuirmos um instrumento que pudesse produzir um diagn6stico infalível ("gold standard") contra o qual se testaria o resultados dos métodos em estudo.

A forma de ultrapassar essa dificuldade foi a utilização do teste Kapa.

Atente-se na Tabela 1.6

Tabela 1.6 Índice de Kapa

Radiologia

Endoscopia Presente Ausente

Presente a b L1

Ausente c d L2

C1 C2 N

Calculamos primeiro Po (proporção de concordância observada).

Po=a+d/N

Igualmente Pc e a proporção esperada devido ao acaso.

Pc=(R1xC1+R2xC2)/NxN

Daqui calcula-se Kapa da forma:

K=(Po-Pc)/(1-Pc)

Este indica e assim uma medida de concordância com uma correcção para o acaso.

A interpretação de K baseia-se na seguinte escala proposta por Landis e Kock.

O=Nula

0-0.2=Fraca

0.21-0.40=Ligeira

0.41-0.60=Moderada

0.61-0.80=Forte

0.80-1.0=Quasiperfeita

Partindo destas bases observemos como se procedeu.

A aplicação aos dados de que dispúnhamos e que provinham de uma amostra de doentes que frequentavam a valência de GastroEnterologia de um Centro de Saúde da região de Lisboa.

Tabela 1.7 Índice de Kapa para a úlcera gástrica

	Radiologia		
Endoscopia	Presente	Ausente	
Presente	34	11	45
Ausente	9	64	73
	43	75	118

Po=(34+64)/118=0.83

Pc=(45x43+73x75)/118x118=0.53

K=(0.83-0.53)/(1-0.53)=.063

Tabela 1.8 Indica de Kapa para a neoplasia gástrica

	Radiologia		
Endoscopia	Presente	Ausente	
Presente	10	0	10
Ausente	3	105	108
	13	105	118

Po=115/118=0.97

Pc=11470/13924=0.82

K=0.15/0.18=0.83

Portanto de acordo com a escala proposta e uma vez que o grau de concordância era cerca de 63% para a úlcera gástrica e de 83% para a neoplasia gástrica estávamos perante uma associação forte para a úlcera e quase perfeita para a neoplasia.

Note-se que naquele tempo (1990) a literatura apresentava valores de 0.67 e 0.91 respectivamente para a úlcera e para a neoplasia, que não se afastavam muito dos valores por nós encontrados.

Note-se igualmente que sol) o ponto de vista clínico o estudo se apresenta algo ultrapassado, uma vez que na ultima década a melhoria da técnica endoscópica, bem como dos respectivos equipamentos, associada à sua capacidade de utilização para biópsia veio, colocar o problema noutros moldes.

Contudo o que se pretende realçar é o interesse do método, e esse mantem-se actual, a utilizar sempre que se necessite proceder à comparação de duas técnicas diagnósticas sem que seja possível dispor de um "gold standard" com o qual os resultados obtidos possam ser comparados.

A Abordagem epidemiológica

No campo da epidemiologia as tabelas de contingência apresentam-se como uma ferramenta essencial de trabalho sendo sobre elas que uma boa parte da investigação se vai desenvolver.

lremos aqui trazer uma abordagem básica a esta matéria, guardando para posteriores capitulos a análise de casos especiais (estudos longitudinais, caso-controlo, etc.)

Convirá aqui recordar, antes do mais, alguns conceitos básicos de Epidemiologia.

As três medidas de frequ6ncia de doença mais comummente utilizadas são razões, proporções e taxas que se designam nos livros anglosax6nicops por "ratios", "proportions" e "rates".

Uma razão é uma expressão da relação entre duas quantidades. As duas quantidades não estarão relacionadas entre si, mas na prática muitas vezes tal acontece.

Ex: Razão dos sexos = N^{o} Homens/N^{o} Mulheres

Uma proporção é uma razão na qual o numerador está contido no denominador.

Ex: N^{o} de mulheres que desenvolvem cancro das mama/N^{o} total de mulheres portadoras de cancro

Uma taxa reflecte uma noção mais dinâmica da evolução no tempo e "grosso modo" pode dizer-se que relaciona o conjunto de indivíduos que detêm o atributo, no numerador, com o número de indivíduos em risco no denominador.

Ex: N^{o} de mulheres que desenvolvem cancro das mama/N^{o} total de mulheres no período (atendendo a que todas as mulheres estão em risco de contrair cancro da mama).

Estes termos são por vezes indiscriminadamente utilizados até mesmo em linguagem epidemiológica.

Essa controvérsia pode ser acompanhada em textos como o de Elandt-Johnson (1975), Kleibaum(), Rothman(), Lllienfeld() e Jeniceck() entre outros.

Para efeitos práticos interessa-nos recordar que este conjunto de medidas de frequência pode ser categorizado em dois grandes grupos:

A Prevalência e a Incidência.

A Prevalência mede o número de casos existentes enquanto a Incidência mede o número de casos novos num dado período de tempo.

Assim a Prevalência dá-nos uma visão da dimensão da doença, enquanto a incidência permite analisar o ritmo a que esta se desenvolve.

Nesta linha a literatura tem vindo a definir dois componentes na Incidência: A Incidência Cumulativa, ou simplesmente Incidência e a Densidade de Incid49ncia segundo Miettinem ().

A Incidência Cumulativa é definida para a clínica como o número de casos novos numa população definida, num dado período de tempo, dividida pela população em risco no mesmo período de tempo.

Define o risco da doença para a s pessoas numa dada população.

Ex: Num almoço de casamento adoecerarn de gastroenterite 30 de 150 convivas, pelo que o risco (Incidência) de adoecer neste almoço foi de 30/150=0.02

(assumindo que todos os convivas comeram o mesmo menu).

Usualmente esta taxa de incidência é denominada taxa de ataque, termo muito utilizado sobretudo em doenças transmissíveis de inicio brusco e de curta duração.

A Densidade de Incidência calcula-se como o conjunto de novos casos numa dada população, num dado período de tempo, dividido pelosomat6rio do tempo que cada pessoa esteve em risco, ou observação, totalizado para todas as pessoas na população em estudo.

· Ex: Foi estudado um grupo de 500 pessoas durante 5 anos e anotado o número de casos de cancro da tir6ide.

· Observaram-se 10 casos no período de 5 anos.

· As pessoas estiveram no estudo durante o tempo seguinte:

· No 1º ano 500 pessoas

· No 2º ano observamos 420 (80 saíram do estudo por várias razões, abandono, 6bito, etc.)

· No 3º ano observamos 415

· No 4º ano 400

· No 5º ano 395

· Logo o total de pessoas-ano é de:

· (500* 1)+(420*2)+(415*3)+(400*4)+(395*5)=6160

Logo a densidade de incidência é de 10/6160 pessoas ano ou aproximadamente 1,62 casos por 1000 pessoas/ano

Contráriamente à incidência cumulativa a densidade de incid6ncia não indica o risco individual na população; representa a velocidade (taxa) A qual os novos casos acorrem.

A densidade de incid6ncia utiliza informação mais especifica no que concerne aos membros de urna população porque usa o tempo que cada indivíduo foi observado (ou esteve exposto).

Toma-se particularmente útil em estudos longitudinais "follow-up" nos quais vArias pessoas são observadas (ou estilo expostas) por períodos diferentes.

Em relação à Prevalência é usual definir a Prevalência Estacional ("Period Prevalence") como o número de casos de doença existentes numa população durante um dado período dividido pelo população susceptível nesse mesmo período.

Define-se igualmente Prevalência Pontual ("Point Prevalence") do mesmo modo só que para um dado momento (dia do mês, fim do ano, etc.) em vez de para um dado período de tempo.

Convirá igualmente nesta breve introdução alargar um pouco a refer6ncia a medidas de associação, úteis para avaliar se existe relação entre a exposição a um determinado factor de risco e a eclosão de dada situação patológica.

Note-se que a medida de associação apropriada depende do modelo de estudo utilizado.

Desenvolveremos mais em pormenor, nos capitulos a eles destinados, os aspectos que se prendem com esses modelos.

Por agora ficamos no campo dos conceitos básicos.

Nos estudos de coorte a medida de associação mais utilizado é o risco relativo.

Este define-se como o quociente entre as taxas de ataque (atrás definida) no grupo dos expostos e a taxa de ataque no grupo dos não expostos.

RR=Ie/Io

Um risco relativo de 1 indica igualdade nos dois grupos. Acima de 1 indica que os expostos têm um risco acrescido, abaixo de 1 indica que os exposto têm menos risco (efeito protector).

A Tabela 1.9 mostra-nos um exemplo de Gross (1976) citado por Stroup D. F e Teutsch (1998).

Tabela 1.9 Intoxicação alimentar em Oswego

	Doentes	Sãos	Total
Expostos	43	11	54
Não Expostos	3	18	21
	46	29	75

Trata-se de uma intoxicação alimentar num jantar de igreja em Oswego, N.Y.

Dos 75 convivas 46 desenvolveram sintomas gastrointestinais.

A taxa de ataque nos expostos foi de 43/54=0.80 e nos não expostos de 3/21 =0.14

0 risco relativo foi de 0.80/=. 1 4=5.7, o que significa que nos expostos o risco de contracção da afecção foi de 5.7 vezes superior ao grupo dos não expostos.

Mais à frente veremos a forma de.realizar testes estatísticos para determinar se esta diferença é significativa sob o ponto de vista estatístico (i.e. o risco ser o mesmo nos dois grupos), Fleiss (1982).

Referem-se ainda outros indicadores como o risco atribuído.

Este define-se como a diferença entre a taxa de ataque nos expostos e nos não expostos.

RA=Ie-Io

Igualmente se define a fracção etiológica do risco, Lillienfeld (1984), como "a máxima proporção de doença que pode ser atribuída a uma característica ou factor etiológico".

FER=Ie-Io/Ie

É possível calcular a FER quando se está em presença de vários factores de risco simultâneos, Kleinbaum et all (1993).

Note-se igualmente que a definição de FER assume que todas as outras variáveis externas se distribuam igualmente no grupo dos expostos e dos não expostos.

Voltando ao exemplo vemos que o risco atribuído é de

0.80-0.14=0.66

e a FER e de

FER=0.80-0.14/0.80=0.66/0.80=0.825

ou seja, 82,5% dos casos são atribuíveis ao risco, neste caso ingestão de determinado alimento.

Nos estudos de caso-controlo a medida de associação mais usual é a "odds ratio ".

Diga-se de passagem que a tradução do termo "odds ratio" tem sido muito tentada e pouco conseguida.

De facto termos como "razão probabilistica" ou "razão do produto dos contrários" soam um pouco destoados da linguagem comum pelo que na literatura se continua a utilizar o termo "odds ratio". É o que iremos fazer.

Recorde-se que em estudos de caso-controlo não se conhecem as taxas de doença pelo que lidamos com probabilidades ("odds").

Daqui que a OR seja a razão entre as probabilidades de exposição nos casos e nos controlos.

Na Tabela 1.10 mostramos a forma de calcular a OR.

Tabela 1.10 Cálculo da OR

	Caso	Controlo
Expostos	a	b
N. Expostos	c	d

Se a doença é rara a OR aproxima-se do risco relativo, Schlesselman (1982).

A interpretação da OR é similar à do risco relativo.

Uma palavra final para a questão do ajustamento das taxas.

Quando, se comparam as taxas de doenças entre populações com diferentes estruturas etárias podem ocorrer diferenças motivadas não pelo risco acrescido em si, mas pela presença numa população de um número de pessoas mais susceptíveis de contrair doença. devido á idade.

Nestas condições torna-se necessário. ajustar ou estandardizar as taxas a fim de remover o efeito da idade e torna-las portanto comparáveis entre si.

Esta matéria que é bastante importante, sobretudo em taxas de mortalidade, está bem desenvolvida em. textos de Bioestatistica, como os de Fleiss (1981) e Colton entre outros.

Voltaremos a estas matérias mais à frente.

Contudo para os leitores mais interessados no desenvolvimento e aprofundamento destes assuntos recomenda-se a leitura dos textos já referidos, nomedamente Rothman e Greenland (1998) para a discussão dos conceitos de incidência e prevalência.

Para os aspectos relacionados com intervalos de confiança e testes de hipóteses para as taxas referidas recomenda-se a leitura das obras de Kleinbaum (1989) e Fleiss (1982).

Uma abordagem mais simplificada da matéria pode ser encontrada em Lillienfeld (1994).

Vamos terminar este capitulo com a apresentação introdut6ria do STATA nestas matérias, campo em que este "software" se mostra particularmente, útil.

Usando o STATA

Vamos utilizar o STATA para testes de independência e de homogeneidade.

Primeiro vamos utilizar os menus pendentes para uma "dataset" previamente introduzida.

Depois vamos observar um conjunto de comandos particularmente úteis, comandos interactivos que permitem flexibilizar o uso, do programa.

Veremos finalmente exemplos da área da epiderniologia onde se torna indispensável o recurso aos comandos interactivos.

1º Exemplo: Teste de Independência

Trata-se de uma "dataset" que resume a redução da pressão arterial de um grupo de doentes a quem foram administrados quatro tratamentos.

0 que pretendemos saber é se as diferenças na redução da pressão arterial são iguais para todos os grupos.

Para isso vamos emitir os seguintes comandos:

Summaries–Tables–Two-way(cross-tabulation)

Surge uma caixa em que nos é pedido a indicação da variável para a linha (row) e para a coluna (columm).

Indicamos na linha a variável systolic e coluna drug.

A caixa dá ainda a opção de nos dar as percentagens de linha, coluna, célula e todas as estatísticas, associadas à tabela.

Para evitar o excesso de informação nas Mulas vamos somente pedir o total de estatisticas.

Fazemos Enter e o resultado é o seguinte:

Pearson chi2(93) = 109.1125 Pr = 0. 121

likelihood-ratio chi2(93) =

Cramer's V = 0.7919

gamma= -0.3974 ASE = 0.083

Kendall's tau-b = -0.3447 ASE = 0.072

Como podemos observar o "output" envolve para além do teste de Qui-quadrado outros testes que complementam a informação. Por agora detemo-nos no qui-quadrado.

Informa-nos que temos um chi2(93) ou seja com 93 graus de liberdade (32-1) linhasx(4-1) colunas cujo valor é de 109.1125. Diz-nos igualmente que o p-value associado é de 0. 121 o que nos impede de rejeitar a hip6tese nula de não haver diferenças nos vários tratamentos.

Os outros testes confirmam esta decisão.

Voltemos agora ao nosso exemplo inicial (também de tratamentos na hipertensão) e imaginemos que queríamos aplicar o teste não a uma "dataset" construida linha a linha, mas sim a uma tabela de que já conhecemos os valores das várias células.

Para tal recorremos ao comando interactivo e escrevemos na linha de comandos:

tabi 506 143\624 161, chi2

tabi é o comando para a tabela com o sufixo i, de interactivo e aplica-se a todos os comandos em que é possível introduzir dados directamente do teclado, sem que o programa os vá ler à "dataset".

Temos depois os valores das linhas separadas pelo "backslash"(\).Finalmente após a vírgula (indispensável) introduzimos a opção chi2 para que o programa nos deia o resultado do teste de Qui-quadrado.

0 resultado e o seguinte:

. tabi 506 143\624 161, chi2

	col		
row	1	2	Total
1	506	143	649
2	624	161	785

```
————+——————————-+————-
   Total |    1130      304 |     1434

      Pearson chi2(1) =   0.4942   Pr = 0.482
```

Como vemos o programa "fez a tabela" e deu-nos o resultado que já tínhamos calculado no inicio do capítulo.

Os comandos interactivos são muito úteis no STATA e permitem muita flexibilidade no uso do programa. Veiamos outra situação

2º Exemplo. Cálculo da taxa de incidência

Para este fim vamos utilizar o comando interactivo iri.

0 STATA trata as tabelas da seguinte forma:

Tabela 1.11 Incidência

	Expostos	Não expostos
Casos	a	b
Pessoas-Ano	N1	N2

E o comando iri é da forma:

iri (a b N1 N2)

A este comando podernos ainda adicionar opções como p. ex. o nível a que queremos efectua-lo (90, 9.5%, etc.)

Irnaginernos o exemplo com os seguintes valores:

a=41, b=15, N1=28010 e N2=19017 com a=0.05

iri 41 15 28010 19017, level(95)

```
      Exposed Unexposed      Total
Cases   41      15      56
Person-Time     28010   19017   47027
Incidence Rate  .0014638        .0007888        .0011908
Point Estimate          (95% Conf. interval)
Inc. Rate Diff. .000675 .0000749–0012751
Inc. Rate Ratio 1.855759        1.005722–3.60942 (exact)
Attr. Frac. Exp.        .4611368        .005689–.7229472 (exact)
Attr. Frac. Pop.        .337681
(mid) Pr(K>=41)         .0177 (exact)
(mid) 2*Pr(K>=41)               .0355 (exact)
```

como vernos o STATA calcula a taxa de incidência, nos expostos e nos não expostos, calcula o risco relativo (RR), o risco atribuído (RA) e a fracção etiológica do risco (FER) procedendo ainda ao cálculo dos intervalos de confiança para estes indicadores.

3º Exemplo: Estudos de coorte e de caso-controlo-

Neste caso usam-se os comandos csi e cci respectivamente.

As tabelas serão do tipo:

Tabela 1.12 Coorte e Caso-Controlo

	Expostos	Não Expostos
Casos	a	b

Não Casos c d

Para a=7 b= 12 c=9 d=2 a 95% vem:
.csi 7 12 9 2, level(95)

	Exposed	Unexposed	Total
Cases	7	12	19
Non Cases	9	2	11
Total	16	14	30
Risk	.4375	.8571429	.633333

Point Estimate		95% Conf. Interval
Risk diff	-.4196429	-.7240828—.1152029
Risk Ratio	.5104167	.2814332–.9257086
Prev. frac.exp.	.4895833	.0742914–.7185668
Prev. frac.pop.	.2611111	

chi2(1)=5.66 Pr>chi2=0.0173

.cci 7 12 9 2, level(95)

	Exposed	Unexposed	Total	Prop.total Exposed
Cases	7	12	19	0.3684
Controls	9	2	11	0.8182
Total	16	14	30	0.5333

	Point estimate	95% Conf. Interval
Odds Ratio	.1296296	.0246233–.7180882(comfield)
Prev. fracc.ex.	.8703704	.2819118–.9753767(comfield)
Prev. fracc. pop.	.7121212	

chi2(1)=5.66 Pr>chi2=0.0173

Note-se a semelhança dos quadros e note-se igualmente que nos resultados do caso-controlo os valores do risco são substituídos pela OR.

0 manual do STATA contem ampla informação sobre as variadas possibilidades e opções destes comandos, que têm obviamente o seu equivalente para aplicar em "datasets", cs (cohort study) e cc (case-control study), sem o respectivo, i de "imediate".

Voltaremos a este assunto nos capitulos dedicados aos estudos de, coorte e caso-controlo.

Capítulo 7

Análise de variância (Anova)

7.1 Introdução

Em capítulos anteriores estudámos a forma de testar hipóteses acerca de médias de duas populações, usando a distribuição nomral, a t de student, etc.

Um problema bem mais frequente é o da comparação das médias de mais de duas populações.

Deste assunto trata o método que iremos descrever neste capítulo, abordando únicamente os seus aspectos mais essenciais e elementares.

Análise de Variância (ANOVA)

Consideremos um caso simples através de um exemplo extraído de Rao (1998).

Exemplo: Num estudo observacional a fim de determinar os efeitos da menopausa nos níveis de colesterol, o observador mediu o colesterol de mulheres divididas em sete grupos, como mostra a

O que se pretende saber é se existe alguma diferença entre a média do colesterol nos 7 grupos.

Para responder a esta questão vamos utilizar um modelo conhecido por ANOVA (Análise de Variância) sobre o qualiremos tecer alguns comentários breves.

O teste de Análise de Variância é feito utilizando a distribuição F, a qual está representada na Figura 1.1.

Figura 1.1 Distribuição F com 10 e 15 graus de liberdade

Possui as seguintes caracteristicas:

É uma distribuição contínua desviada para a direita;

Tem n1 graus de liberdade no numerador e n2 graus de liberdade no denominador

As unidades de uma distribuição F, denominadas F, são sempre não negativas

Posto isto, recordemos que a ANOVA é utilizada para testar a hipótese nula que as médias de três, ou mais populações,são equivalentes, contra a hipótese alternativa de o não serem.

Este teste, que é o mais simples como (one-way) ANOVA, unidireccional uma vez que que se compara uma média em várias populações.

Contudo o processo pode ser mais complexo, bidimensional (two-way) ou multidireccional (N-way) se estiverem em jogo não um mas dois ou mais factores.

Assim, no nosso exemplo, como pretendemos conhecer a diferença das médias do colesterol, levando em consideração vários tipos de tratamento em cada grupo, teriamos de recorrer a uma análise de tipo multidimensional.

Recordemos que o processo se denomina de Análise de Variância, porque é baseado na análise da variação dos dados obtidos das diversas amostras.

Vamos de seguida:

- Fazer uma referência às condições de operacionalidade do teste;
- Tecer alguns comentários a essas condições;
- Apresentar o teste, em si;
- Fazer menção á análise estatística post-teste;
- Mencionar brevemente alguns testes "post-hoc" (post-exame);

Condições Gerais de Operacionalidade do Teste

As populações das quais as amostras são retiradas devem ter distribuição aproximadamente normal

As populações das quais as amostras são retiradas devem ter variância igual

As amostras retiradas de diferentes populações devem ser independentes e as variáveis devem estar identicamente distribuidas

Comentários ás condições anteriores e sua violação

Tal como nos testes paramétricos dos capítulos anteriores, a violação das condições atrás referidas introduzirá distorções nos resultados que linitarão a inferência.

Convém contudo notar aqui que no caso da ANOVA esta é um pouco menos sensível à violação da normalidade da distribuição das variáveis e, mesmo a algumas pequenas alterações da igualdade das variâncias desde que as amostras sejam iguais e não muito pequenas.

Contudo devemos notar que a desigualdade das variâncias, sobretudo em amostras pequenas e com grnades variâncias internas pode ser desastrosa para os resultados.

Voltamos aqui a reiterar alguns conselhos já anteriormente referidos:

- Trabalhar com amostras grandes;
- Fazer uma boa análise univariada, usando os métodos gráficos ou os métodos algébricos que o STATA (e outros programas) contêm: Shapiro-Wilk, Shapiro-Francia, etc.

Vemos que o STATA produz igualmente o teste de Bartlett para testar a igualdade das variâncias. Recordamos contudo que este teste, tal como outros testes para a comparação das variâncias como o de Hartley Fmax, assumem a distribuição normal das populações e podem conduzir a resultados anormais, caso esse pressuposto não se verifique.

Teste Estatístico

O teste estatístico F para um teste de hipóteses usando ANOVA é dado pela razão de duas variâncias entre as amostras (MSB) e a variância dentro das amostras (MSW).

O cálculo destes valores é explicado com o exemplo seguinte:

Exemplo: 15 estudantes doram aleatorizados em três grupos a fim de analizar o efeito de três tratamentos diferentes (medindo-se o valor de dado parâmetro em estudo).

Tabela 1.2 Distribuição de tratamentos

Tratamento 1	Tratamento 2	Tratamento 3
48	55	84
73	85	68
51	70	95
65	69	74
87	90	67

Vamos utilizar a notação seguinte:

x=o valor de cada observação;

k=o número de amostras diferentes (ou tratamentos)

ni= a dimensão da amostra i

Ti=a soma dos valores da amostra i

n=o número total de valores em todas as amostras (n1+n2+n3+.......+ni)

Sa=a soma dos valores em todas as amostras (T1+T2+T3.....+Ti)

Sa2=a soma dos quadrados dos valores em todas as amostras

Desenvolvendo a tabela fica:

Tabela 1.3 Cálculo da ANOVA

Tratamento1	Tratamento 2	Tratamento 3
48	55	84
73	85	68
51	70	95
65	69	74
87	90	67
T1=324	T2=369	T3=388
n1=5	n2=5	n3=5

Capítulo 8

Regressão

8.1 Introdução

Sem dúvida que um dos aspectos mais interessantes de toda a investigação é o de tentar estabelecer relações (de preferência causiais) entre variáveis de forma a que se possa inferir qualquer coisa do tipo "determinado acontecimento é provocado pela presença de uma causa que concluimos ser o fenómeno tal".

Estas relações causa-efeito, são sempre difíceis de determinar, quiça mesmo impossiveis, mas existe sempre a esperança, da nossa parte, de o conseguir. Muitas vezes contudo ficamos um pouco aquém e, em vez de se chegar a um verdadeiro nexo de causalidade contentamo-nos com uma aproximação do tipo "este acontecimento está relacionado com aquele efeito".

Repare-se nesta importante diferença: duas variáveis podem estar relacionadas entre si, sem que daí se possa concluir que uma é causa da outra.

A problemática da discussão dos chamados "critérios de causalidade" sai fora do âmbito deste texto; contudo qualquer bom livro de Epidemiologia já citado como o de Rothman () abordam a questão, havendo até um tratamento mais clássico, mas contudo interessante, sobretudo para quem começa, no livro de Jeniceck().

O que vamos abordar neste capítulo é a introdução a um método de análise de relação entre variáveis, muito usual e quando bem empregue qualificado, conhecido por modelo de regressão.

Claramennte o que pretendemos encontrar é algo que nos permita *descrever* a relação entre as variáveis e, daí poder prever o valor que a variável pode assumir num indivíduo, conhecendo sómente o valor da outra.

O tema é vasto e complexo, sendo particularmente bem apresentado em textos de Econometria, como os de Griffith(), etc.

Contudo existem igualmente excelentes obras na área da Investigação em Saúde, aliás já citadas, como as de Hamilton(), Kleinbaum(), etc. que recomendamos aos leitores que queiram melhor conhecer o assunto.

Aqui iremos sómente falar do modelo de regressão linear simples, que explicaremos adiante; abordaremos apenas o essencial, fugiremos á matemática, que aqui se torna mais pesada, sobretudo em modelos com várias variáveis, iremos interpretar o "output" do computador, e edeixaremos no fim do capítulo alguns apêndices sobre temas mais avançadas, cuja leitura como de costume será absolutamente dispensável para quem não esteja interessado na matéria.

8.2 O modelo de regressão linear simples

Começaremos por um exemplo, retirado de Thuesen *et all.*, citado por Altman(1991) explicando depois os conceditos de *linear* e *simples*.

Trata-se de 24 observações de diabéticos de tipo 1, aos quais foi medida a glucose em jejum (em mmol/l) *fbg* e a média circunferencial da velocidadede de encurtamento do ventriculo esquerdo *vcf*.

O que pretendemos saber é se essa média é influencada pelos valores da glucose em jejum e, mais ainda, saber qual o valor dessa média para um dado caso sabendo qual o valor da glucose em jejum desse indivíduo.

No primeiro caso trata-se de uma inferência sobre a média, no segundo de uma previsão individual.

Vejamos como responder a estas duas questões.

Antes do mais vamos definir alguns termos: À variável *vcf* chamamos de variável dependente, enquanto á variável *fbg* iremos chamar variável dependente, ou predictor.

Digamos em palavras simples que a segunda influencia a primeira e é justamente essa influência que desjamos analisar.

Um modelo de regressão é um modelo matemático que descreve as relações entre duas ou mais variáveis.

O modelo de regressão simples inclui sómente duas variáveis; uma independente e uma dependente.

A variável dependente é a que se pretende explicada, e a variável independente é a que se utiliza para explicar a variação da variável dependente.

A relação entre as duas variáveis, num modelo de regressão é expresso por uma equação matemática denominada *equação de regressão*, ou *modelo*.

Esta equação quando grafada pode produzir vários tipos de curvas, incluindo uma linha recta.

Neste caso diz-se que o modelo assume a forma de *regressão linear*; nos outros casos dir-se-á que se trata de uma *regressão não linear*.[1]

Resumida e sintéticamente podemos dizer que:

A regressão linear é um modelo de regressão simples em que a relação entre as duas variáveispode ser descrita por uma equação do tipo:

$$Y = A + BX$$

que é, como se sabe, a equação de uma recta que intercepta o eicxo dos YY no ponto A e tem um declive B, como na Figura 8.1

[1]O termo regressão foi utilizado pela primeira vez por Sir Francis Galfour (1822-1911) que estudou a relação

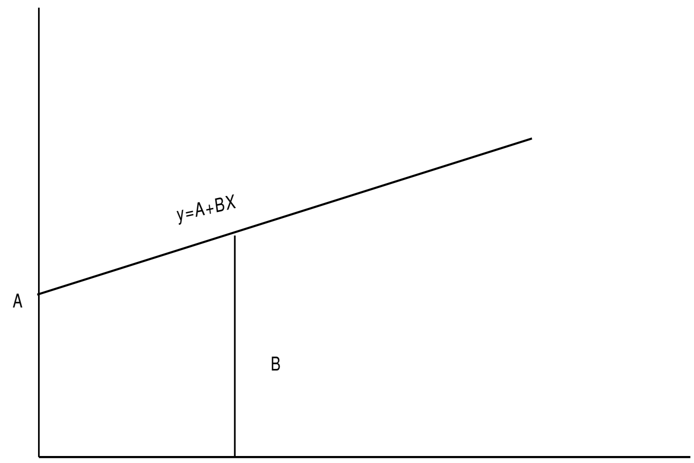

Figura 8.1: Recta de Regressão

Os respectivos coeficientes A e B designam-se por constante (A) e declive (B).

Genéricamente podemos dizer que A representa o valor que a variável dependente (Y) assume quando a variável indenpendente (X) assume o valor 0, representando B o incremento na variação da variável dependente (Y) quando a variável independente (X) aumenta de uma unidade.

8.3 Modelo matemático e modelo estatístico

O que acabamos de expôr é genéricamente o conceito de relação linear simples, materializado por uma expressão que nos conduz à equação de uma recta, bem conhecida da Geometria Analítica.

Se as coisas se passassem assim no mundo real, então sabendo um qualquer valor de X era possivel determinar com toda a precisão o respectivo valor de Y; digamos que não existiria margem para qualquer erro.

Mas na verdade as coisas não são assim; a natureza não se comporta como um modelo matemático rígido e, torna-se necessário espelhar na formulação do problema o sentido probabilístico ou estocástico do modelo, introduzindo um novo elemento que é o erro aleatório ou resíduo, aqui definido por ε.

Assim a nossa equação passará a assumir a forma:

$$Y = A + BX + \varepsilon$$

Note-se a diferença fundamental; passámos de um modelo rígido para um modelo dinâmico que reflecte as variações introduzidas pelo acaso e pela variabilidade dos fenómenos da natureza.

---

entre as alturas das crianças e a altura dos respectivos progenitores.

É este o modelo que nos interessa, e sobre o qual nos vamos debruçar um pouco mais.

8.4 Estimação de A e B

Como já sabemos de capítulos anteriores desconhecemos a forma como as coisas se passam na população (encarada sob um ponto de vista estatístico) em geral.

Os nossos modelos servem para, a partir de amostras -com as características próprias definidas- podermos inferir, com um grau de certeza também préviamente estabelecido, o valor dos parâmetros populacionais.

Aqui passa-se algo de igual: A e B são desconhecidos; temos de encontrar métodos para os estimar a partir da nossa amostra aleatória.

Digamos então que na prática o nosso modelo assume a forma:

$$y = a + bx + e$$

em que a, b, e e, são *estimadores* de A, B e ε.

Quais os métodos utilizados para produzir esta estimação?.

Como de costume vamos deixar de lado os detalhes -os quais no caso presente até podem ser bem complexos- e vamos dar uma ideia muito geral do problema, passando ao "output" do computador para melhor interpretação.

Fundamentalmente dois métodos são utilizados para a estimação referida: o dos *mínimos quadrados*, e o da *máxima verosimilhança*.

O primeiro, conhecido em linguagem anglo-saxónica por OLS de "Ordinary Least Squares", é um método clássico -utilizado no tempo dos cálculos "à mão"-e dificil de implementar quando aunenta o número de variáveis independentes (regressão multipla).

O segundo é um método mais sofisticado, utilizado pelos computadores, que produz estimadores de boa qualidade, mas cuja complexidade ultrapassa o âmbito deste texto; contudo os interessados encontrarão no livro de Hills() um excelente texto sobre a matéria, podendo ainda recorrer ao livro de ...que explica a utilização da técnica no STATA.

No Apêndice 8.1 temos uma descrição do método dos míminos quadrados (OLS), chamando a atenção do leitor aqui e agora para os elementos seguintes, suficientes no nosso ponto de vista para entender o modelo.

Vejamos a figura 8.2

Para o valor k de x, x_k recolhemos por observação o valor y_k . Contudo se introduzirmos na equação da recta o valor de x_k só por sorte iriamos obter o valor y_k; era necessário que o nosso modelo fosse de uma perfeição matemática absoluta para que tal acontecesse.

O que iremos obter, então, será um valor $\hat{y}_k$ que se afastará, mais ou menos do valor y_k encontrado, conforme o nível de ajustamento do modelo.

Podemos então dizer que para o modelo $Y = A + BX$ a média, ou valor esperado $E[Y|X]$ será igual a $A + BX$ estimado por $\overline{y}= a + bx$ e $\hat{y}_k = a + bx_k$; o valor $\hat{y}_k - y_k$ será o valor residual ou e.

Atente-se na figura 8.3

Repare-se que fazendo várias observações os vários valores encontrados se afastam, para cima ou para baixo, do valor médio $a + bx$

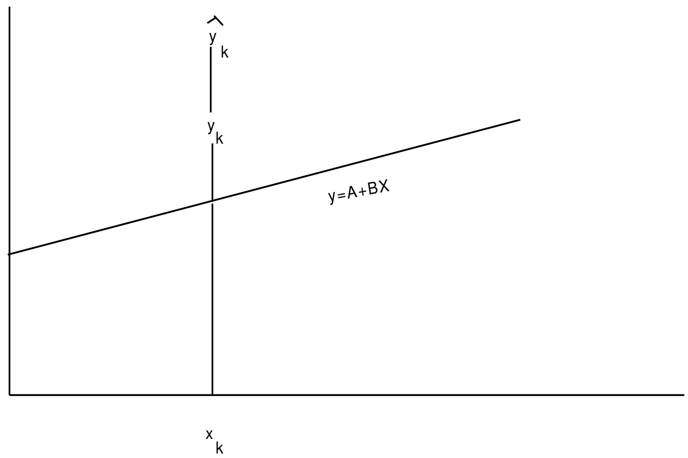

Figura 8.2: Regressão, valores observados e estimados

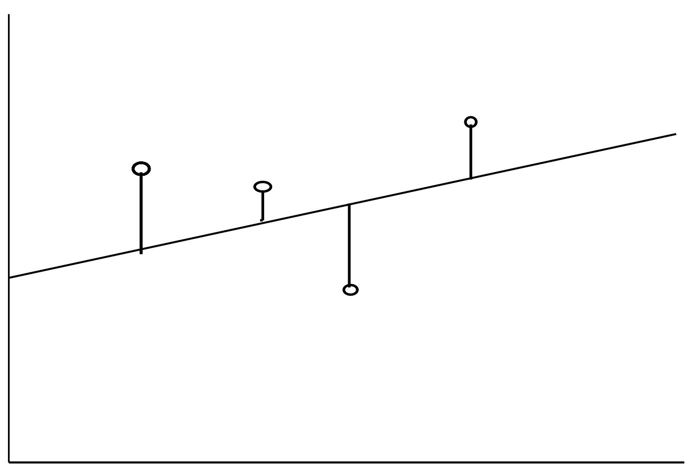

Figura 8.3: Regressão e valores observados

Então demonstra-se (ver apêndice 8.1) que a recta ideal será aquela que minimize o $\sum(y_k\text{-}\hat{y}_k)$.

Contudo para evitar somatórios negativos utilizou-se o expediente de elevar ao quadrado esse somatório.

A recta que melhor se ajusta aos nossos dados será aquela que minimize o somatório dos quadrados dos resíduos $SSR=\sum e^2=\sum(y_k\text{-}\hat{y}_k)^2$.

Recordemos igualmente duas expressões que nos vão ser úteis: $SSE=\sum(\hat{y}_k\text{-}\bar{y})^2$ e SST=SSR+SSE.

Regressemos ao nosso exemplo inicial.

Pedindo ao STATA que proceda á regressão de y sobre x, ou seja *vcf* sobre *fbg* temos:

```
. regress vcf fbg

         |       SS       df       MS                  Number of obs =      23
---------+------------------------------                    F(  1,    21) =    0.00
   Model |  6.3340e-06     1  6.3340e-06                    Prob > F      =  0.9927
Residual |   1.5396547    21  .073316891                    R-squared     =  0.0000
---------+------------------------------                    Adj R-squared = -0.0476
   Total |  1.53966104    22  .069984593                    Root MSE      =  .27077

------------------------------------------------------------------------------
     vcf |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     fbg |   .0000181   .0019437      0.009   0.993       -.004024    .0040601
   _cons |   1.358401   .0647288     20.986   0.000        1.22379    1.493012
------------------------------------------------------------------------------
```

Por agora vamos ficar com a visão desta tabela; a sua explicação será detalhada nas secções seguintes.

8.5 Condições do modelo de regressão

O modelo de regressão é algo exigente e por isso torna-se necessário assegurar-mo-nos de alguns aspectos essenciais a saber:

1. O erro aleatório, ou resíduo (termo mais usual) tem uma média 0 e um desvio-padrão fixo, ao longo de todas as observações de valor σ_ε.

2. Os resíduos associados com diferentes observações devedm ser independentes entre si.

3. Para qualquer valor de *x* a distribuição dos resíduos é normal com média 0 e desvio-padrão definido em 1.

Deve notar-se ainda que o modelo de regressão linear se aplica a *variáveis dependentes contínuas,* independentemente da variável indcependente poder ser contínua ou discreta; as observações da variável X devem ser independentes entre si e, assumir-se cada observação como um valor fixo.

A violação destes princípios acarretará graves imprecisões na estimação dos parâmetros. Veremos no Apêndice 8.2 alguns desses casos e suas respectivas consequências.

8.5.1 Coeficiente de determinação

Um dos elementos a interpretar no "output" da tabela da regressão é o chamado *coeficiente de determinação* R^2.

No fundo pretende responder a questões do tipo:

- Qual a qualidade do modelo de regressão?
- De que forma a variável independente explica a variável dependente?
- Qual, de uma forma geral é a "força" da relação entre as variáveis?

Calcula-se como o quociente entre o somatório dos quadrados do modelo definido como, já trás dissemos:

SSE(Explained summ of squares)=$\sum(\hat{y}_k - \bar{y})^2$ e o total da soma dos quadrados definida como SST(Total summ of squares)=SSR+SSE.

Logo $R^2 = \frac{SSE}{SST}$; como se vê este coeficiente diz-nos qual a parte da variabilidade geral explicada pelo modelo e portanto o quanto ele se adapta aos nossos dados.

O coeficiente de determinação varia entre 0 e 1 e é, como atrás se disse um dos componenetes do "output" do programa.

Inferência sobre b