Bioinformática

Metodologias Estatísticas em Bioinformática

Manuela Neves

ISA/ULisboa

15 e 16 de Março 2021

Revisões de Estatística com recurso ao  .
Testes de hipóteses: caso geral
Testes múltiplos. *P-values*

Plano das aulas

1. Referências
2. A comparação de sequências
   - Alinhamento de pares de sequências
3. Breve revisão dos conceitos estatísticos
   - Testes de hipóteses
   - Funções no  para modelos de v.a.'s
   - *P-Values*
4. Tabelas de contingência e testes do qui-quadrado
   - Testes de Independência
   - Testes de Homogeneidade

Referências

- **W. Ewens and G. Grant. (2001)**. *Statistical Methods in Bioinformatics*. An introduction. Statistics for Biology and Health. Springer
- **W. P. Krijnen (2009)**. *Applied Statistics for Bioinformatics using* R. Disponível online
- **M. Manuela Neves (2017)**. *Introdução à Estatística e à Probabilidade com utilização do* R. ISAPress.
- **D. D. Pestana e S. F. Velosa (2008)**. *Introdução à Probabilidade e à Estatística*. Fundação Calouste Gulbenkian.
- **K. Seefeld (2007)**. *Statistics using* R *with Biological Examples*. University of New Hampshire Department of Mathematics & Statistics.

A comparação de sequências

Objectivo: Identificar semelhanças/diferenças entre sequências de DNA, RNA e de Proteínas.

Na comparação de sequências pretende-se

1. Analisar duas ou mais sequências;
2. Identificar diferenças

Para isso efectua-se o Alinhamento de Sequências para

1. Medir a similaridade entre duas ou mais sequências
2. Inferir relações evolucionárias
3. Observar padrões de conservação e variabilidade para predições estruturais e funcionais.

Alinhamento de pares de sequências

- Comparar duas sequências biológicas é como comparar duas "sequências de caracteres".
- Viram que existem diversos métodos para comparar sequências de caracteres
- Do ponto de vista biológico, é possível que a similaridade ocorra devida ao acaso.
- Com o alinhamento pretende-se identificar sequências homólogas numa lista de sequências similares.
- O que vamos tratar nestas duas aulas é referir procedimentos estatísticos que vos permitem interpretar e compreender o que se passa.

A comparação de sequências– testes de hipóteses

Os procedimentos estatísticos para comparar duas sequências iniciam-se pela formulação de uma **hipótese nula, H_0**, como por exemplo:

H_0: dado um par de aminoácidos alinhados, os dois aminoácidos foram gerados por mecanismos independentes

i.e., um pouco mais formalmente, se o aminoácido j ocorre numa qq posição, na 1ª sequência, com probabilidade p_j e o aminoácido k ocorre numa qq posição, na 2ª sequência, com probabilidade p_k, a probabilidade de ocorrer o par (j, k) num dado alinhamento é $p_j p_k$.

Breve revisão dos conceitos estatísticos

Mas a teoria dos Testes de Hipóteses, exige a formulação de uma hipótese alternativa. No caso em estudo é costume considerar

$\boldsymbol{H_1}$: a probabilidade do par $(j,k) = q(j,k)$

Testes de hipóteses

Formular uma **hipótese nula** *versus* uma **hipótese alternativa.**

Na realização de um teste de hipóteses podemos dizer que é necessário executar **cinco etapas** (**quatro** antes da recolha, ou da utilização, dos dados)

Testes de hipóteses

Vamos ilustrar usando duas sequências de DNA, talvez provenientes de duas espécies diferentes. Os | indicam que os pares de nucleótidos são os mesmos em ambas as sequências

```
g g a g a c t g t a g a c a g c t a a t g c t a t a
|   |     |     | | |   | |                 | | |
g a a c g c c c t a g c c a c g a g c c c t t a t c
```

Temos um exemplo de *sequence-matching*

Será que as duas sequências apresentam uma semelhança significativamente superior à que seria de esperar em duas sequências arbitrárias de DNA, daquela espécie, para se poder dizer que há evidência de que provêm de um ascendente comum?

Testes de hipóteses

Se as sequências fossem geradas ao acaso, as 4 letras **a, g, c,** e **t** apresentavam igual probabilidade de ocorrer em qualquer posição, logo as sequências tenderiam a ter a mesma letra em cerca de 1/4 das posições
Mas ... nas 26 posições há 11 comuns

Quão **improvável** seria isto acontecer, se fossem geradas ao acaso?

Entra aqui conhecimento e propriedades das sequências aleatórias.
A observação das 11 identidades (*matches*) em 26, $11/26 = 0.42$, dá alguma informação de que algo mais que o acaso ocorreu.

Estamos a "estimar", com base nos dados, um valor hipotético desconhecido, parâmetro, e pretendemos tomar decisões sobre "quanto acreditamos naquele valor".

Passos num Teste de hipóteses

Passo 1

Formular as hipótes nula, H_0, e alternativa, H_1.

Breves Notas:

- A escolha das hipóteses nula e alternativa deve ser feita antes da recolha dos dados.
- O objectivo do teste é rejeitar ou não rejeitar H_0 com recurso a procedimentos estatísticos adequados e usando os dados.
- O que significa dizer "a hipótese nula é aceite"? – significa dizer– não há evidência estatística para a rejeitar a favor da hipótese alternativa.

Testes de hipóteses

Mas a hipótese nula pode ser aceite porque a alternativa pode não explicar suficientemente melhor os dados.

Por isso, melhor que dizer aceitar ... deve dizer-se **não rejeitar** H_0

Voltemos ao exemplo

Vamos escolher H_0 e H_1.

$H_0 : p = 0.25$, significando que cada um dos quatro nucleótidos aparece numa posição qq com probabilidade 0.25, independentemente dos outros nucleótidos, portanto as duas sequências foram geradas ao acaso;

e podemos especificar que a hipótese alternativa é $p > 0.25$, ou também por exemplo, $p = 0.35$, como podia ocorrer se fossem relacionadas.

Portanto no nosso exemplo, é natural considerar a hipótese alternativa $p > 0.25$

Passo 2

Escolha do erro de decisão

Notas:

A decisão de não rejeitar ou rejeitar H_0 baseada nos dados, pode ser incorrecta.

	não rej. H_0	rej. H_0
H_0 verd.	decisão correcta	erro tipo I
H_0 falsa	erro tipo II	decisão correcta

Testes de hipóteses

Os erros da decisão de rejeitar ou não rejeitar H_0 são designados, respectivamente por **erro de 1^a espécie** ou **erro de tipo I** e **erro de 2^a espécie** ou **erro de tipo II,** sendo as probabilidades associadas a cada um dos erros habitualmente designadas por

$\alpha = P$ (erro de tipo I) = P (rejeitar $H_0 | H_0$ verdadeiro)
$\beta = P$ (erro de tipo II) = P(não rejeitar $H_0 | H_0$ falso).

A α é costume chamar nível de significância do teste e a $1 - \beta$ = P (rejeitar $H_0 | H_0$ falso) potência do teste.

Testes de hipóteses

Situação ideal – ter as probabilidades arbitrariamente pequenas de ter um erro Tipo I e um erro Tipo II, o que não é possível assegurar, a menos que o número de observações fosse tão grande quanto quiséssemos.

O dilema é resolvido, vendo que há assimetria nas implicações dos dois erros.

Por exemplo, no exemplo *sequence-matching*,

- pode haver mais preocupação em fazer uma afirmação falsa positiva – de que as 2 sequências são semelhantes, se não há semelhança
- e menos preocupação numa conclusão falsa negativa – dizer não há semelhança, quando há semelhança

Testes de hipóteses

Habitualmente

Fixa-se um valor para a probabilidade do erro de Tipo I, α (muito baixo 1% ou 5%).
A teoria dos testes foi desenvolvida assegurando que fixado α, o erro de Tipo II tem a menor probabilidade

Neste Passo 2 – fixa-se o valor de α

Testes de hipóteses

Passo 3

Determinação da **estatística de teste** – é a variável que, calculada a partir dos dados leva à tomada de decisão — conduz à aceitação ou rejeição da hipótese nula.

No exemplo *sequence-matching* uma **estatística de teste possível** é Y – v.a. que conta o número total de *matches*.

Algumas vezes a escolha da estatística de teste pode não ser simples!!

Passo 4

Neste passo determina-se o **valor da estatística de teste** com base nos valores observados.

Exemplo com o nosso problema

Seja então, Y, número total de *matches* a estatística de teste.
Quer a hipótese alternative fosse $p = 0.35$ ou $p > 0.25$ a hipótese nula $p = 0.25$ era rejeitada a favor da alternativa quando o valor observado y de Y é suficientemente grande, i.e., é maior que algum valor de significância K.

Testes de hipóteses

Se o erro de Tipo I for escolhido igual a 5%, K é tal que

$$\text{Prob(hipótese nula ser rejeitada| verdadeira)}=$$
$$Prob(Y > K|p = 0.25) = 0.05$$

Se estamos a trabalhar com variáveis discretas, pode não ser possível encontrar um valor K que dê exactamente aquele valor do erro Tipo I.

Para o cálculo de K e outras quantidades de interesse em modelos de variáveis aleatórias, vamos recordar as facilidades do 

Funções no R para modelos de v.a.'s

- **d**função $(x, ...)$ - permite obter a função massa de probabilidade (modelo discreto) ou a função densidade (modelo contínuo) em x;
- **p**função$(q, ...)$ - permite obter a função de distribuição cumulativa, i.e., devolve a probabilidade de a variável ser menor ou igual a q;
- **q**função $(p, ...)$ - permite calcular o quantil associado à probabilidade p;
- **r**função $(n, ...)$ - permite gerar uma amostra de n números pseudo-aleatórios do modelo especificado.

Significado:

density, probability, quantile, random

Testes de hipóteses

No caso do nosso exemplo, verifique-se que

$Prob(Y > 10|p = 0.25) = 0.0400845$ e
$Prob(Y > 9|p = 0.25) = 0.09085561$
Então a escolha de K é feita de modo conservativo, i.e., deve considerar-se $\boldsymbol{K = 10}$

Verifique que se, por exemplo, tivesse $n = 100$, $\alpha = 0.05$ e $p = 0.25$
$Prob(Y > 31|p = 0.25) = 0.069$ e
$Prob(Y > 32|p = 0.25) = .044$
Usamos o valor conservativo 32 para K.

Nota

Verifique que o uso do comando `qbinom(0.95,n,0.25)` lhe permite obter o valor do K – quantil de probabilidade 0.95

Testes de hipóteses

Aquela dificuldade ocorre quando a estatística de teste é uma v.a. discreta.

Em sequências muito longas, pode usar-se a aproximação da binomial pela distribuição normal.

Exemplo: $n = 1000000$ e $\alpha = 0.05$.
K pode determinar-se considerando

$$Prob[X \geq K + 1/2] = 0.05$$

sendo $X \sim \mathcal{N}(\mu, \sigma)$, com $\mu = 1000000 \times 0.25 = 250000$ e $\sigma^2 = 1000000 \times 0.25 \times 0.75 = 187500$, considerando correcção de continuidade.

Obtém-se $K = 250711.74$ na prática pode usar-se o valor conservativo $K = 250712$

Testes de hipóteses

Passo 5

Finalmente nesta fase vamos usar os dados!!!
Agora determina-se o valor da estatística de teste e verifica-se se é igual ou mais extremo que o "ponto de significância" calculado.

Rejeita-se a Hipótese nula se o valor calculado for superior a K.
Caso contrário (aceita-se) não se rejeita H_0.

P-Values

Um procedimento de teste equivalente ao que foi descrito baseia-se no cálculo do chamado *P-value* do valor encontrado.

Já não se calcula o Passo 4, em vez dele ... a partir dos dados, calcula-se a probabilidade de se obter um valor igual ou mais extremo do observado para a estatística do teste, sob H_0

É esta probabilidade que se chama *P-value*.
Se *P-value* $\leq$ probabilidade do erro Tipo I — a hipótese nula é rejeitada; caso contrário não se rejeita

Exemplo

Dada a hipótese nula $H_0 : p = 0.25$

Qual a probabilidade de se observarem 11 ou mais *matches* numa sucessão de comprimento 26, (exemplo em estudo)?

Sendo $Y \frown Binomial(26, p)$ tem-se $P[Y \geq 11 | p = 0.25] \approx 0.04$

Este é o *P-value* associado ao valor observado 11.

Por exemplo se $n = 1000$ e se encontraram 278 *matches*, o P-value pode ser determinado usando a aproximação da binomial pela normal como

$$Prob(X \geq 277.5)$$

Cálculo do *P-value* no caso de um teste de alternativa bilateral

Exemplo Queremos testar se uma moeda é equilibrada. Realizámos o lançamento 100 vezes e verificámos que, por exemplo, a face “moeda” saíu 58 vezes.
O *P-value* é a probabilidade de obter 58 ou mais ou 42 ou menos dado que para uma alternativa bilateral temos que considerar mais extremos os valores para ambas as caudas.

Exercício: Calcule o *P-value* associado a esta experiência

Cálculo do *P-value* no caso de um teste de alternativa bilateral

O exemplo acabado de tratar é um caso particular de cálculo do *P-value*, quando a distribuição da estatística de teste, neste caso, é simétrica.

No caso geral, para **testes bilaterais**, adopta-se:

– Sendo T a estatística de teste e t_{obs} o valor da estatística, sob a hipótese H_0, para os dados observados, o *p-value* do teste é assim calculado:

- $2P[T < t_{obs}|H_0]$ se t_{obs} for reduzido;
- $2P[T > t_{obs}|H_0]$ se t_{obs} for elevado.

(t_{obs} é reduzido (elevado) se a estimativa que se obtém para o parâmetro a testar é inferior (superior) ao valor especificado em H_0)

Os testes do qui-quadrado

Uma vez que estamos a falar de testes de hipóteses, vamos referir uns testes muito importantes nas vossas aplicações

Os testes do qui-quadrado em tabelas de contingência

Tabelas de contingência

Suponhamos que os indivíduos de uma amostra são classificados de acordo com dois critérios (factores) A e B (qualitativos ou quantitativos).
Consideremos ***r*** níveis do critério A e ***c*** níveis do critério B. Portanto os **n** valores observados são classificados de acordo com 2 diferentes factores (critérios).
É costume apresentar as frequências observadas o_{ij} na célula (i, j) de uma tabela a que se chama **tabela de contingência**

	B_1	$\cdots$	B_j	$\cdots$	B_c	
A_1	o_{11}	$\cdots$	o_{1j}	$\cdots$	o_{1c}	$o_{1.}$
A_2	o_{21}	$\cdots$	o_{2j}	$\cdots$	o_{2c}	$o_{2.}$
.	.	.	.	.	.	.
A_r	o_{r1}	$\cdots$	o_{rj}	$\cdots$	o_{rc}	$o_{r.}$
	$o_{.1}$	$\cdots$	$o_{.j}$	$\cdots$	$o_{.c}$	

$\sum_{i=1}^{r} \sum_{j=1}^{c} o_{ij} = n$ e o_{ij} representa o número de elementos da amostra classificados nas categorias A_i e B_j.

Testes de independência

Se a tabela de contingência resultou da classificação dos **n** indivíduos da amostra segundo os níveis de cada um dos critérios, regra geral pretende-se com este estudo inferir da eventual existência de alguma relação ou associação entre os dois critérios de classificação. As hipóteses a testar são:

$\mathbf{H}_{0:}$ *A* e *B* são independentes *vs* $\mathbf{H}_{1:}$ *A* e *B* não são independentes

A estatística do teste é

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - e_{ij})^2}{e_{ij}},$$

onde e_{ij} representa a estimativa da frequência esperada, se a hipótese H_0 fosse verdadeira, i.e. $e_{ij} = \frac{o_{i.}o_{.j}}{n}$

Se H_0 verdadeira, $\boldsymbol{X^2} \sim \chi^2_{(r-1)(c-1)}$.

Rejeita-se a hipótese H_0 se $X^2_{cal} > \chi^2_{\alpha,(r-1)(c-1)}$

Exercício

Realizou-se uma experiência para verificar a eficácia de uma nova vacina contra a gripe, a qual foi administrada numa pequena comunidade. A vacina era gratuita e tinha de ser administrada em duas doses, separadas por um período de duas semanas. Nem todos apareceram à vacinação e alguns que tomaram a 1ª dose, não apareceram para receber a 2ª dose. Na primavera seguinte, recolheu-se a seguinte informação sobre 1000 dos habitantes da dita comunidade:

	Não vacinado	Uma dose	Duas doses
Gripe	24	9	13
Não gripe	289	100	565

Com base nos resultados, verifique se existe evidência suficiente que indique existência de associação entre a administração da vacina e a ocorrência ou não de gripe.

Resolução do Exercício no R

Pretendemos testar a hipótese nula, de que não há relação entre a ocorrência de gripe a administração, i.e, pretendemos testar a hipótese de que são independentes.

$H_0 : p_{ij} = p_{i.}p_{.j}, \ \forall(i,j)$
v.s.
$H_1 : p_{ij} \neq p_{i.}p_{.j}$, para pelo menos 2 pares (i,j)

```
gripe<-matrix(c(24,9,13,289,100,565),nc=3,byrow=T,
  dimnames=list(c("Gripe", "Nao.Gripe"),
  c("Nao.Vac.", "1Dose","2Doses")))
gripe
margin.table(gripe,1)
margin.table(gripe,2)
chisq.test(gripe)
chisq.test(gripe)$expected
chisq.test(gripe)$residuals^2
```

Notas

Pressupostos a verificar:

- as frequências esperadas em cada classe não devem ser inferiores a 5, quando o número total de observações é ≤ 20;
- se $n > 20$ não deverá existir mais do que 20% das células com frequências esperadas inferiores a 5, nem deverá existir nenhuma com frequência esperada inferior a 1.
- se nos casos anteriores as condições não se verificarem deve-se juntar linhas ou colunas (desde que tal junção tenha significado).
- a realização de um teste de independência não deve terminar com a rejeição da hipótese nula. Deve analisar-se a contribuição de cada célula para o valor de X^2.

Tabelas de contingência–Testes de Homogeneidade

Nas tabelas de contingência referidas atrás, considerava-se que a amostra de dimensão n era classificada de acordo com cada um dos critérios, i.e., o número de observações que era contado em cada célula era determinado depois de obtida a amostra. Sendo assim, o total das linhas e colunas não está sob o controle do investigador. Diz-se que a tabela de contingência tem margens livres, pois os totais das margens resultam do processo de classificação. O teste realizado chama-se teste do qui-quadrado de independência.

Contudo, o total das linhas ou das colunas de uma tabela de contingência pode estar sob o controle do investigador, i.e., uma das margens da tabela ser fixa. Nesta situação o teste a realizar diz-se ser um teste do qui-quadrado de homogeneidade.

Testes de Homogeneidade

Exemplo - Pretende fazer-se um estudo para averiguar se o comportamento dos condutores face a acidentes de automóvel é diferente consoante a faixa estária.
Em vários grupos de idade, recolheu-se uma amostra de condutores e foi-lhes perguntado se tinham tido algum acidente no ano anterior e, em caso afirmativo se tinha sido de maior ou menor gravidade. O resultados encontram-se na seguinte tabela:

Idade	Tipo de acidente			
	Nenhum	menor	maior	Total
Inferior a 18	67	10	5	82
18-25	42	6	5	53
26-40	75	8	4	87
40-65	56	4	6	66
mais de 65	57	15	1	73

Testes de Homogeneidade

Existirá diferença na distribuição das respostas em cada classe etária?

Portanto agora estamos preocupados em responder à questão: a percentagem (a proporção) de acidentes de cada tipo é a mesma entre as diferentes classes etárias, i.e., as classes etárias apresentam o mesmo comportamento face ao tipo de acidente?

Neste exemplo as linhas representam as subpopulações das quais se retiraram as amostras. Cada elemento da amostra foi depois classificado em cada um dos três critérios: Nenhum acidente, Acidente menor e Acidente maior.

Testes de Homogeneidade

A estatística de teste é a mesma que num teste de independência.

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - e_{ij})^2}{e_{ij}},$$

onde e_{ij} é uma estimativa da frequência esperada.
Se as populações são homogéneas, i.e., se o comportamento face a cada tipo de acidente é o mesmo em cada uma das classes etárias, então sendo a proporção de elementos em cada modalidade da categoria é "a mesma" de subpopulação para subpopulação, por exemplo, se relativamente à modalidade "Nenhum acidente" se tem $67/82 \approx 42/53 \approx ... \approx 297/361$, etc.

Testes de Homogeneidade

Então esperamos encontrar $o_{.1}o_{1.}/n$ observações na 1ª célula, depois ...
As frequências esperadas são então:

$$e_{ij} = \frac{o_{i.}o_{.j}}{n}$$

Se H_0 é verdadeiro a estatística de teste, X^2, tem assintoticamente distribuição Qui-quadrado com $(r-1)(c-1)$ graus de liberdade.

Rejeitamos a hipótese H_0 se o valor calculado, $X^2_{cal} > \chi^2_{\alpha,(r-1)(c-1)}$

Note-se que o teste de realiza da mesma forma que o teste de independência

Resolver o exercício no 

Testes de Independência e de Homogeneidade

Notas conclusivas:

O teste qui-quadrado de Pearson é um teste estatístico aplicado a dados categóricos ou dados classificados.

Um teste de qualidade do ajustamento, que veremos mais tarde, estabelece se uma distribuição de frequências observadas difere de uma distribuição teórica.

Um teste de independência avalia se observações de duas variáveis, expressas numa tabela de contingência, são independentes entre si. Recolhe-se uma amostra de dimensão n e são contados os indivíduos que pertencem a uma categoria (classe) de uma variável e a uma categoria (classe) da outra variável.