Bioinformática – 2014/2015

Módulo:
Modelos de Probabilidade em Bioinformática
Estimação e Inferência (algumas notas)

Tópicos do Programa

- **Modelos de Probabilidade em Bioinformática**
  - Introdução. Probabilidade Condicional. Teorema da probabilidade total e Teorema de Bayes (revisões). Exemplos de aplicação.
  - Medidas da qualidade de testes de diagnóstico: sensibilidade e especificidade.
  - Modelos de Probabilidade mais usados em Bioinformática. Propriedades e aplicações.
- **Estimação e Inferência (algumas notas)**
  - Métodos de Estimação (breve abordagem)
  - Testes a contagens: o teste do qui-quadrado de ajustamento. Testes em tabelas de contingência.
  - Aplicação de métodos de reamostragem em Bioinformática - o *bootstrap*.

Manuela Neves
manela@isa.ulisboa.pt

Referências Bibliográficas

- **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 (2009)**. *Introdução à Estatística e à Probabilidade. Apontamentos de Apoio à U.C. Estatística* (disponíveis no portal Fenix da U.C.)
- **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.

Índice

Capítulo 1

Modelos de Probabilidade em Bioinformática

1.1 Introdução

A Bioinformática refere-se à utilização e criação de métodos algorítmicos, computacionais e estatísticos, para resolver problemas teóricos e práticos de dados biológicos.

A estrutura do DNA

Uma sequência de DNA ou sequência genética é uma cadeia de **4 letras (bases)** – $\boldsymbol{A, C, G, T}$ representando a estrutura primária de uma molécula, formada pela junção de um grande número de nucleotídeos, e que contém a informação genética codificada.

Vários problemas de interesse surgem:

- dada uma sequência, que metodologia estatística se pode usar para a descrever?
- é possível determinar a que tipo de organismo uma sequência pertencerá, pela análise do conteúdo da sequência?
- dadas duas sequências que apresentam semelhanças, serão elas significativas para assegurar que têm o mesmo ascendente?

Necessário:

- Modelos Probabilísticos adequados
- Métodos Estatísticos de análise de palavras

Consideremos o seguinte segmento de DNA

$$5' \ldots GGATCGAAGCTAAG \ldots 3'$$

Questões que por exemplo se colocam:

1. Que padrão de bases aparece com frequência "anormal" numa dada sequência?
2. Como avaliar aquela "surpresa"?

Temos então necessidade de utilizar regras probabilísticas.

1.2 Teoria da Probabilidade–Revisões

Probabilidade condicional

Sejam A e B dois acontecimentos definidos em Ω

Definição Chama-se **probabilidade condicional de A dado B** ou **probabilidade de A se B** e representa-se por $\boldsymbol{P(A|B)}$, com $P(B) > 0$ a

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(AB)}{P(B)}. \tag{1.1}$$

Teorema das probabilidades compostas

Se $P(A) > 0,\ P(B) > 0,$

$$\boldsymbol{P(AB) = P(A)\ P(B|A) = P(B)\ P(A|B)}. \tag{1.2}$$

Generalização a três acontecimentos

Sejam $A,\ B,\ C$ tais que $P(A) > 0$, $P(B) > 0$ e $P(C) > 0$, tem-se,

$$P(ABC) = P(A)P(B|A)P(C|AB) = P(B)P(C|B)P(A|BC) =$$
$$= P(C)P(A|C)P(B|AC).$$

Definição Dois acontecimentos A e B dizem-se mutuamente independentes se e só se

$$P(A \cap B) = P(A)\ P(B). \tag{1.3}$$

Da definição conclui-se que se A e B são independentes então

$P(A|B) = P(A)$ se $P(B) > 0$ e $P(B|A) = P(B)$ se $P(A) > 0$.

Teorema Se A e B são independentes A e $\overline{B}$, $\overline{A}$ e B e $\overline{A}$ e $\overline{B}$, também são independentes.

Definição – Independência de três acontecimentos

Os acontecimentos A, B e C dizem-se **mutuamente independentes** também se diz apenas independentes se e só se

$$P(ABC) = P(A)\ P(B)\ P(C); \qquad P(AB) = P(A)P(B);$$
$$P(AC) = P(A)P(C) \qquad P(BC) = P(B)P(C).$$

Teorema da probabilidade total

Sejam $A_1, A_2, ..., A_n$ acontecimentos definindo uma partição sobre Ω, i.e.,
$A_1 \cup A_2 \cup .... \cup A_n = \Omega$ e $A_i \cap A_j = \emptyset$, $\forall i, j (i \neq j)$.
Se $P(A_i) > 0$, então para qualquer acontecimento $B \in \Omega$ tem-se

$$\boldsymbol{P(B) = \sum_{i=1}^{n} P(B|A_i)\ P(A_i).} \tag{1.4}$$

Teorema de Bayes

Sejam $A_1, A_2, ..., A_n$ acontecimentos formando uma partição de Ω , onde $P(A_i) > 0$. Seja B um outro acontecimento de Ω, tal que $P(B) > 0$. Então para $k = 1, ..., n$ tem-se

$$\boldsymbol{P(A_k|B) = \frac{P(A_k).P(B|A_k)}{\sum_{i=1}^{n} P(A_i).P(B|A_i)}}. \tag{1.5}$$

Resolução de um exemplo de Aplicação – Cálculo de probabilidades marginais e conjuntas com recurso à condicional

Exemplo

Considere uma sequência de DNA e uma distribuição conjunta (hipotética) de nucleótidos em duas posições adjacentes, apresentada no seguinte quadro:

	Nucl.pos. 2			
Nucl. pos. 1	A	C	G	T
A	0.2	0	0	0
C	0.1	0.1	0.1	0.1
G	0	0.1	0.1	0
T	0.1	0.1	0	0

Que informação se pode retirar desta tabela?

1.3 Aplicação - Testes de Diagnóstico

Análises clínicas são meios auxiliares de diagnóstico, i.e., são meios de rastreio para o diagnóstico de uma dada doença. Um teste de diagnóstico permite identificar numa população de indivíduos "saudáveis" os que têm uma probabilidade elevada de possuir a doença.

Num teste de diagnóstico há dois tipos de erro possíveis:

- o teste é aplicado a um indivíduo **doente** e dá um **resultado negativo** - negativo falso - NF;
- o teste é aplicado a um indivíduo **são** e dá um **resultado positivo** - positivo falso - PF.

A situação pode resumir-se no seguinte quadro

Resultado	Doente	São $\equiv$ Não Doente	Total
Positivo	PV	PF	$P\,(+)$
Negativo	NF	NV	$N\,(-)$
Total	D	$\overline{D}$	n

Estudar a validade do teste significa saber a sua **sensibilidade (S)** e a sua **especificidade (E)**. Como se definem?

Definição – Chama-se **sensibilidade (S)** de um teste à proporção de positivos entre os doentes, isto é, exprime-se por

$$S = P(\text{Positivo}|\text{Doente}) = P(+|D) = \frac{PV}{PV + NF} \tag{1.6}$$

Definição – Chama-se **especificidade (E)** de um teste à proporção de negativos entre os sãos, isto é, exprime-se por

$$E = P(\text{Negativo}|\text{São}) = P(-|\overline{D}) = \frac{NV}{PF + NV} \tag{1.7}$$

Definição – Chama-se **prevalência da doença** ao quociente D/n, i.e., é a proporção de doentes na população.

Então o que gostaríamos de ter era:
- só positivos verdadeiros, i.e, sensibilidade 100% e
- não ter positivos falsos, portanto especificidade também 100%.

Mas ... não é possível ter as duas situações optimizadas... portanto há que decidir qual o risco que se pretende controlar.

Definição – Chama-se **valor preditivo** do teste à probabilidade de decisão correcta, face aos resultados do teste.

O valor preditivo positivo é $P(D|+) = \frac{PV}{P} = \frac{PV}{PV+PF}$

O valor preditivo negativo é $P(\overline{D}|-) = \frac{NV}{N} = \frac{NV}{NF+NV}$

Exercício 1.1 *(Galvão de Mello - vol I))*

Um teste para a detecção de diabetes tem para S e E os valores 52.9% e 99.4%, respectivamente. Admite-se que a prevalência de casos é, para todas as idades cerca de 8 em 1000.
i) Qual a probabilidade de um indivíduo cujo teste deu positivo ser doente?
ii) Qual a probabilidade de um indivíduo cujo teste deu negativo não ser doente?

Vamos resolver?

Resposta: i) 0.4118 ii) 0.9962

1.4 Exercícios

1. A composição base de um certo genoma decorre de acordo com as seguintes probabilidades $p_G = p_C = 0.3$ e $p_A = p_T = 0.2$. Estamos interessados em sequências de palavras duplas em que a ocorrência de cada letra se admite independente. Há portanto 16 palavras diferentes. Bases "purinas" definem-se por $R = \{A, G\}$ e bases "pirimidinas" definem-se por $Y = \{C, T\}$. Seja E o acontecimento "a primeira letra é uma purina" e F o acontecimento "a segunda letra é" A ou C ou T.

   (a) Determine $P[E]$, $P[F]$ e $P[F|E]$.

   (b) * Analisadas 50 palavras duplas qual a probabilidade de se realizar 5 vezes o acontecimento E definido na alínea anterior.

   (c) * Se nas 50 palavras consideradas na alínea anterior se verificar 20 ocorrências do acontecimento E o que poderá dizer sobre este acontecimento? Justifique convenientemente.

   *–para fazer mais adiante

2. Numa dada população 55% dos indivíduos têm excesso de peso, 20% têm tensão arterial elevada e 60% têm excesso de peso ou tensão arterial elevada.

   (a) Poderemos dizer que a tensão arterial é independente do excesso de peso? Justifique.

   (b) Escolhido um indivíduo ao acaso determine a probabilidade de ter excesso de peso e não ter tensão arterial elevada.

3. Dados 3 acontecimentos A, B e C de um espaço de resultados Ω, com $P(C) > 0$, mostre que:
$$P(A \cap B|C) = P(A|B \cap C).P(B|C)$$

4. Os sintomas febre, cansaço e dores no corpo estão associados em 60% dos casos às gripes e em 40% às constipações. A auto-medicação é muito frequente face as estes sintomas. Verifica-se que 40% das vezes os medicamentos ingeridos para o tratamento da gripe são os aconselhados para as constipações e em 70% das situações os medicamentos utilizados para tratamento das constipações são os indicados para a gripe.

   (a) Qual a probabilidade de o medicamento ingerido ser realmente o indicado?

   (b) Sabendo que o medicamento era o apropriado para a doença, qual a probabilidade de o doente ter tido gripe?

5. Um teste para detecção de um dado tipo de vírus foi aplicado a 900 portadores e a 2400 não portadores do vírus, tendo-se obtido os seguintes resultados:

Resultado	Portador	Não Portador
Positivo	832	183
Negativo	68	2217

(a) Calcule a probabilidade de um indivíduo escolhido ao acaso, de entre os submetidos ao teste:

i. Ter um resultado positivo no teste.

ii. Ter resultado positivo no teste e ser portador do vírus.

iii. Não ser portador do vírus e ter resultado negativo.

iv. Não ser portador da doença sabendo que o teste deu negativo.

(b) Determine a sensibilidade e a especificidade do teste.

6. Um teste de diagnóstico permite observar 1000 indivíduos semanalmente. Sabendo que $P(D) = 0.02$, $S = 95\%$ e $E = 90\%$, onde D, S e E designam "um indivíduo ter doença", "sensibilidade do teste" e "especificidade do teste", respectivamente, determine:

(a) O número de testes positivos esperados por semana.

(b) O número esperado de falsos positivos.

7. Num grupo de um milhão de pessoas que fizeram um teste específico A, para detecção do vírus HIV, obteve-se os seguintes resultados:

	Estado de saúde da pessoa		
Resultado do teste	Portador de HIV	Não Portador de HIV	Total
Teste Positivo	4 885	73 630	78 515
Teste Negativo	115	921 370	921 485
Total	5 000	995 000	1 000 000

Se uma pessoa for seleccionada aleatoriamente daquele grupo, determine a probabilidade de:

(a) Obter um resultado positivo no teste e não ser portadora de HIV.

(b) Não sendo portadora de HIV obter um resultado positivo no teste.

(c) Tendo um teste negativo não ser portadora de HIV.

(d) Calcule a sensibilidade e a especificidade deste teste.

8. **Realização de 2 testes**
Sejam T_1 e T_2 dois testes para a doença D. Suponha que $P(D) = 0.008$, $S_1 = 52.9\%$, $E_1 = 99.4\%$, $S_2 = 51\%$ e $E_2 = 98\%$. Sabendo que ao aplicar os testes T_1 e T_2 ambos deram resultados positivos, qual o valor preditivo conjunto?

9. Um teste T para uma dada doença tem sensibilidade S e especificidade E. São feitas duas aplicações (independentes, obviamente) deste teste a um indivíduo. Pretende calcular-se a probabilidade de que o individuo tenha efectivamente a doença se foram positivos os resultados das duas aplicações.

   (a) Defina "sensibilidade" e "especificidade" do teste.

   (b) Sendo S=0.99 e E=0.95 exprima o valor preditivo positivo em função da prevalência, ao aplicar uma vez o teste.

   (c) Obtenha a expressão da sensibilidade e da especificidade do resultado da aplicação dos dois testes.

   (d) Obtenha a expressão do valor preditivo positivo, no caso das duas aplicações do teste.

10. Numa população admite-se que uma dada doença tem uma prevalência de 0.01. Sempre que uma pessoa está doente, a doença é detectada (resultado positivo de um teste) em 80% dos casos.

    (a) De entre as pessoas que se sabe serem portadores da doença seleccionaram-se ao acaso 15 pacientes a que se aplicou o teste. Determine a probabilidade de:

        i. Em pelo menos dois doentes o teste não detectar a doença.

        ii. Em quantos doentes se espera que o teste dê positivo?

    (b) Admita que o teste dá resultado positivo em 10% dos casos em que é aplicado. Indique:

        i. A sensibilidade e a especificidade do teste

        ii. O valor preditivo positivo do teste.

Soluções de alguns Exercícios

2. **(a)** Não **(b)** 0.40

4. **(a)** 0.48 **(b)** 0.75

5. **(a)** **i.** 0.3076
ii. 0.2521
iii. 0.6718
iv. 0.9702
(b) S=832/900; E=2217/2400

6. $P(D) = 0.02$, $S = P(+|D) = 0.95$ e $E = P(-|ND) = 0.90$

(a) Como $P(+) = 0.117$ logo o número de testes positivos esperados por semana é 117.

(b) $P(+ \cap \overline{D}) = 0.098$ o número esperado de falsos positivos é 98.

8. $P(D) = 0.008$, $S_1 = 52.9\%$, $E_1 = 99.4\%$, $S_2 = 51\%$ e $E_2 = 98\%$.

Sensibilidade dos dois testes, $S^* = S_1 \times S_2 = 0.2698$

A – ambos os testes darem resultado positivo

Especificidade, $E^* = 1 - (1 - E_1)(1 - E_2)$.

O valor preditivo conjunto – probabilidade de um individuo estar doente se ambos deram positivos

$$P(D|A) = \frac{P(D \cap A)}{P(A)} \qquad P(A) = P(A \cap D) + P(A \cap \overline{D}) = 0.0023$$

Então o valor preditivo conjunto positivo é 0.9477, isto é 94.77%

Verifique que o valor preditivo posiyivo da aplicação só do 1º teste é 41.24%

9. (b) $S = 0.99$ e $E = 0.95$

$P(D) = PV + NF$, com PV a probabilidade de um positivo verdadeiro e NF a probabilidade de um negativo falso.

$PV = 0.99P(D)$ e $NV = 0.95(1 - P(D))$. Então o valor preditivo positivo em função da prevalência é:

$$P(D|+) = \frac{0.99P(D)}{0.94P(D) + 0.05}$$

(c) A sensibilidade e a especificidade do resultado da aplicação dos dois testes, são respectivamene dadas por $S^* = S^2$ e $E^* = 1 - (1 - E)^2$.

(d) $P(D|A)$, onde A representa "os dois testes dão resultado positivo" é dado por

$$P(D|A) = \frac{S^2\ P(D)}{S^2\ P(D) + (1 - E)^2 P(\overline{D})}$$

e agora é só substituir.

1.5 Variável aleatória - Revisões

Definição Chama-se **variável aleatória (v.a.)** e costuma representar-se por $\boldsymbol{X}$, uma função cujo valor é determinado pelo resultado de uma experiência aleatória.

Tipos de variáveis aleatórias

- **Discretas** as que assumem um conjunto finito ou infinito numerável de valores.

  Exemplo: – número de vezes que é contado o nucleótido A numa sequência de ADN com um dado comprimento

- **Contínuas** as que são susceptíveis de tomar qualquer valor real num dado intervalo, que pode ser a recta real (definição grosseira)

  Exemplo: – tempo que decorre até se verificar a 1ª ocorrência de um dado fenómeno

1.6 Modelos de Probabilidade - Revisões

Iremos considerar os modelos de probabilidade mais utilizados em Bioinformática:

- **Modelos Discretos** - uniforme, binomial, geométrico, Poisson,...
- **Modelos Contínuos** - uniforme, normal ou de Gauss, exponencial, gama, ...

Exemplo: Qual a distribuição de probabilidade do número de vezes que um dado padrão pode ocorrer numa sequência aleatória de DNA?

Seja X o número de vezes que, por exemplo, A ocorre numa sequência aleatória de comprimento n.

Se for possível admitir a **independência** de ocorrência de uma qq letra em cada posição da sequência ...

... ter-se-á $\mathrm{X} \frown \boldsymbol{Bin(n, p_A)}$, onde p_A designa a probabilidade de ocorrência de A numa qq posição.

Exercício 1.2 *Admita-se uma molécula de DNA em que cada base tem a mesma probabilidade de ocorrer. Numa sequência de comprimento $n = 1000$, observa-se 280 ocorrências de A. Qual a probabilidade de se observar aquela ocorrência (ou um número mais extremo) sob a hipótese formulada?*

Cálculo directo - uso da distribuição binomial e recurso ao R.
Pretende-se $P[X \geq 280] = 1 - P[X \leq 279]$

```
>1-pbinom(279,1000,0.25)
[1] 0.01643666
```

Uma nota: o R, por omissão, calcula $P[X \leq 280]$, se lhe indicarmos a cauda para a direita (i.e. lower.tail=F) ele considera que o procedimento resultou de $P[X \leq 280] = 1 - P[X > 280]$, isto é, o 280 não é contado.

Então, caso se pretenda $P[X \geq 280]$, se queremos a cauda direita, deve considerar-se $P[X > 279]$, para começar a contar em 280

```
>pbinom(279,1000,0.25,lower.tail=F)
[1] 0.01643666
```

Em vez de fazermos o cálculo directo, pode considerar-se **uma aproximação** que resulta da utilização do **Teorema Limite Central**– no caso da distribuição binomial tem-se:

Se $X \frown Bin(n,p)$ e $np > 5$ e $nq > 5$ então

$$\frac{X - np}{\sqrt{npq}} \sim \mathcal{N}(0,1)$$

Usando novamente o R temos, por exemplo:

```
>1-pnorm(280,250,sqrt(1000*.25*.75))
[1] 0.01422987
```

Na verdade dever-se-ia ter em conta a correcção por continuidade, ... **o que conduz a um resultado melhor**

```
> 1-pnorm(279.5,250,sqrt(1000*.25*.75))
[1] 0.01560537
```

1.7 Mais Modelos de Probabilidade Discretos

Outras distribuições de probabilidade discretas surgem em Bioinformática para contar a ocorrência de características de interesse. Aconselha-se uma revisão dos seguintes modelos:

- **A distribuição uniforme discreta**

- **A distribuição geométrica** que conta o número de provas até ao 1º sucesso – $\boldsymbol{X \frown Geom(p)}$

  $\boldsymbol{P[X = x] = (1-p)^{x-1}p \qquad x = 1, 2, 3, \ldots}$

  ou o número de sucessos até ao 1º insucesso – designe-se por $\boldsymbol{Y \frown Geom(p)}$

  $\boldsymbol{P[Y = y] = (1-p)\, p^y \qquad y = 0, 1, 2, 3, \ldots}$

  ou ainda o número de insucessos até ao 1º sucesso – designe-se por $\boldsymbol{V \frown Geom(p)}$

  $\boldsymbol{P[V = v] = (1-p)^v\, p \qquad v = 0, 1, 2, 3, \ldots}$

  Exercício 1.3 *Considere esta última definição da v.a. com distribuição geométrica.*

  1. *Mostre que* $P[V \geq m] = q^m, \qquad m = 0, 1, 2, 3, \ldots.$
  2. *Prove a propriedade da falta de memória de* V*, i.e,*

  $$P[V \geq m + n | V \geq m] = P[V \geq n].$$

  Para a **distribuição geométrica**, e consoante a definição adoptada, temos

  $$E[X] = \frac{1}{p}; \qquad Var[X] = \frac{1-p}{p^2};$$

  $$E[Y] = \frac{p}{1-p}; \qquad Var[Y] = \frac{p}{(1-p)^2};$$

  $$E[V] = \frac{1-p}{p}; \qquad Var[V] = \frac{1-p}{p^2}.$$

- **A distribuição de Poisson** – $\boldsymbol{X \frown Pois(\lambda)}$, ($\lambda > 0$ designa o número médio de sucessos no domínio em estudo.)

  Relembre-se que se $X \frown Pois(\lambda)$, tem-se

  $$P[X = x] = \frac{e^{-\lambda}\lambda^x}{x!} \qquad \text{para} \qquad x = 0, 1, 2, \cdots$$

  e para a distribuição de Poisson tem-se

  $$E[X] = \lambda; \qquad Var[X] = \lambda.$$

1.8 Modelos de Probabilidade Contínuos

Os modelos de probabilidade contínuos mais usuais em Bionformática são:

- **A distribuição uniforme** — $\boldsymbol{X \frown Unif(a,b)}$

$$\boldsymbol{f(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & x \leq a \vee x \geq b \end{cases}}$$

- **A distribuição exponencial** — $\boldsymbol{X \frown Exp(\beta)}$

$$\boldsymbol{f(x) = \begin{cases} \frac{1}{\beta} e^{-x/\beta} & x > 0, \quad \beta > 0 \\ 0 & x \leq 0 \end{cases}}$$

Aplicações A distribuição exponencial é uma distribuição de grande importância em Biologia molecular - como exemplo, modela o tempo de vida (aleatório) de moléculas de RNA.

Molecules that do not undergo any kind of aging process while still active, and which are not actively degraded by other means, would be as likely to degrade at any time, irrespective of their age–Ewens and Grant, pag.43

Nota: Esta distribuição, tal como a geométrica (que é uma discreta já referida atrás) gozam da propriedade da falta de memória.

Exercício 1.4

1. *Verifique a propriedade da falta de memória na distribuição exponencial.*
2. *Vamos mostrar que: "Dado h pequeno, a probabilidade de uma variável com distribuição exponencial tomar valores no intervalo $(x, x+h)$ dado que o seu valor excede x é aproximadamente proporcional ao comprimento do intervalo.*

- **A distribuição gama** – esta distribuição tem como caso particular a distribuição exponencial. É particularmente importante porque pode ter várias formas. $\boldsymbol{X \frown Gama(\alpha, \beta)}$

$$\boldsymbol{f(x) = \begin{cases} \frac{1}{\beta^{\alpha}\, \Gamma(\alpha)} x^{\alpha-1}\, e^{-x/\beta} & x > 0 \qquad \alpha > 0, \quad \beta > 0 \\ 0 & x \leq 0 \end{cases}}$$

onde $\Gamma(\alpha) = \int_0^{+\infty} x^{\alpha-1}\, e^{-x}\, dx$, é a **função gama**.

Alguns gráficos da função densidade, para vários valores de α e β.

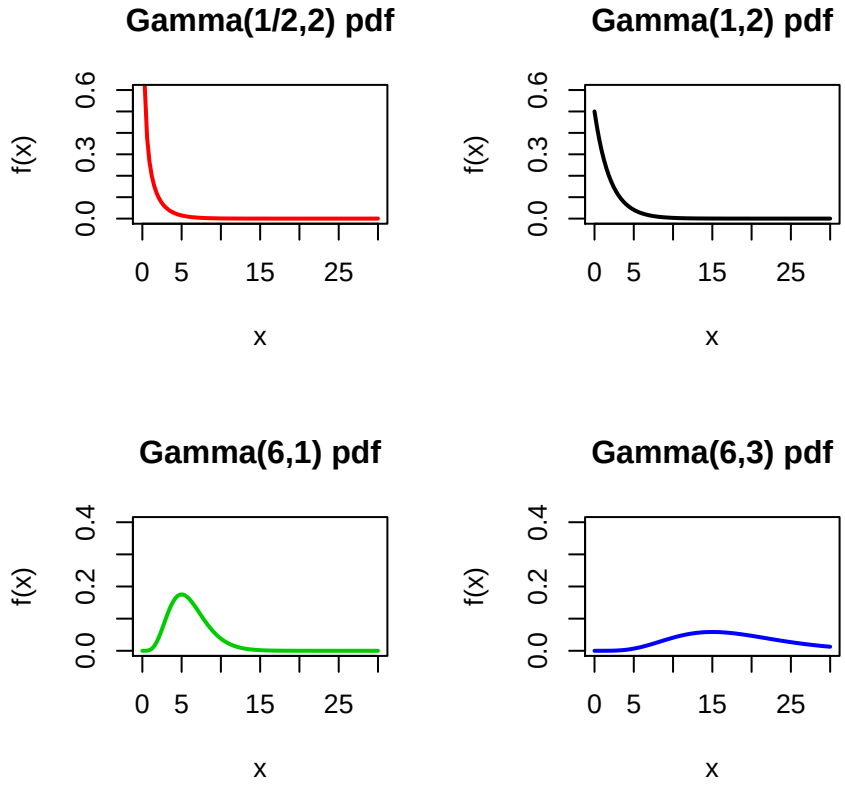


Figura 1.1: Gráficos da função densidade de uma v.a. com distribuição $G(1/2, 2)$, $G(1, 2)$, $G(6, 1)$ e $G(6, 3)$, da esquerda para a direita, cima para baixo, respectivamente.

1.9 Várias variáveis aleatórias

Em quase todas as aplicações lida-se com **várias variáveis aleatórias**, $Y_1, Y_2, \ldots, Y_n$.

1º caso. Admitamos que $Y_1, Y_2, \ldots, Y_n$ são variáveis aleatórias discretas. O vector $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$ tem uma lei de probabilidade a que se chama **distribuição de probabilidade conjunta**, $\boldsymbol{P_Y(y)} = P[Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n]$.

No caso de as n variáveis serem independentes

$$\boldsymbol{P_Y(y)} = \Pi_{i=1}^{n} P[Y_i = y_i] \tag{1.8}$$

Exercício 1.5 *Determine a lei de probabilidade conjunta do vector* $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$, *onde*

- $Y_i \frown Poisson(\lambda_i)$, $i = 1, \ldots, n$, *independentes.*

- *Y_i $(i = 1, \ldots, n)$, são os resultados de n provas de Bernoulli, independentes, cada uma com probabilidade de sucesso p.*

Consideremos agora o caso de haver **dependência** entre as as n variáveis.

Comecemos com um exemplo muito importante – de uma **distribuição de probabilidade (discreta) conjunta** quando as variáveis aleatórias individuais são dependentes – é o caso da **distribuição multinomial** – é a generalização da binomial ao caso de termos uma sequência de $\boldsymbol{n}$ provas, mas em que em cada prova pode haver mais de dois resultados possíveis.

Suponhamos então que cada prova tem k resultados possíveis:

$E_1, \quad E_2, \quad ...E_k \qquad$ com probabilidades
$p_1, \quad p_2, \qquad ...p_k \qquad p_i \geq 0 \qquad$ e $\qquad \sum_{i=1}^{k} p_i = 1.$

Pretendemos determinar a probabilidade de em n provas independentes observar
x_1 vezes o acontecimento E_1
x_2 vezes o acontecimento E_2
...
x_k vezes o acontecimento $E_k, \qquad$ com $x_1 + x_2 + ... + x_k = n.$

Sejam $X_1, X_2, ..., X_k$ as variáveis aleatórias que designam o número de vezes que sai cada um dos acontecimentos nas n provas.
A probabilidade associada ao vector $\quad \boldsymbol{X} = (X_1, X_2, ..., X_k) \quad$ é

$$\boldsymbol{P}_X(\boldsymbol{x}) = \frac{n!}{x_1!\ x_2!...x_k!}\ p_1^{x_1}\ p_2^{x_2}...p_k^{x_k} = \frac{n!}{\Pi_{i=1}^{k}(x_i!)}\Pi_{i=1}^{k}p_i^{x_i}. \tag{1.9}$$

Exercício 1.6

1. *De acordo com a teoria genética, um certo cruzamento de porcos da índia resultará em descendência vermelha, preta e branca na proporção de 8:4:4.*

   Determine a probabilidade de que, entre 8 descendentes, 5 sejam vermelhos, 2 pretos e 1 branco.

2. *Verifique, considerando por exemplo o par (X_1, X_2), que a factorização apresentada em (1.8) já não se verifica.*

No caso de X e Y serem duas variáveis aleatórias quaisquer, podemos definir agora um conceito análogo ao de acontecimentos condicionais.

Define-se **distribuição condicional** como:

$$P[X = x_i|Y = y_j] = \frac{P[X = x_i, Y = y_j]}{P[Y = y_j]} \qquad \text{com} \quad P[Y = y_j] > 0$$

Generalizando a definição de distribuição condicional a n variáveis

$$P[Y_{i+1} = y_{i+1}, \ldots, Y_k = y_k|Y_1 = y_1, \ldots, Y_i = y_i] = \frac{P[Y_1 = y_1, \ldots, Y_k = y_k]}{P[Y_1 = y_1, \ldots, Y_i = y_i]}$$

com $\quad P[Y_1 = y_1, \ldots, Y_i = y_i] > 0$

Se houver independência, tem-se,

$$P[Y_{i+1} = y_{i+1}, \ldots, Y_k = y_k|Y_1 = y_1, \ldots, Y_i = y_i] = \Pi_{j=i+1}^{k} P[Y_j = y_j]$$

2º caso. Se as variáveis $Y_1, Y_2, \ldots, Y_n$ são variáveis aleatórias contínuas o vector $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$ tem uma **densidade de probabilidade conjunta**, $\boldsymbol{f_Y}(y_1, y_2, \ldots, y_n)$, que no caso de as n variáveis sejam independentes se pode exprimir como:

$$\boldsymbol{f_Y(y)} = \Pi_{i=1}^{n} f_{Y_i}(y_i). \tag{1.10}$$

Aplicação

Consideremos que temos n variáveis aleatórias, $X_1, X_2, \ldots, X_n$ i.i.d., cada uma com a mesma distribuição de X.

Como exemplo considere que as n variáveis modelam o tempo de vida de um tipo de proteínas, i.e., vamos admitir que seguem uma lei exponencial, $X_i \frown Exp(\beta)$.

Se há n dessas moléculas numa célula, mostre que o tempo até à degradação da 1ª molécula segue também uma lei exponencial.

Resposta: O que se pretende é determinar a distribuição de $\boldsymbol{X_{min} \equiv X_{1:n}} := \min(X_1, \cdots X_n)$.

- Comece por mostrar que $F_{X_{1:n}}(x) = 1 - e^{-nx/\beta}$, onde $\quad F_{X_{1:n}}(x) = P[X_{1:n} \leq x]$, portanto $\quad \boldsymbol{X_{min} \frown Exp(\beta/n)}$.
- Mostre ainda que

$$f_{X_{1:n}}(x) = (n/\beta)e^{-nx/\beta}$$

Exercício 1.7 *Determine a função de distribuição cumulativa e a função densidade do mínimo no caso de uma uniforme em (0,L)*

Resumo de distribuições no R

As principais distribuições de interesse estão construídas no R. Antes de passarmos à estimação e inferência, vejamos o resumo no seguinte quadro:

Nome da distribuição no R	Função	Parâmetros	Exemplo (f.dist.cumul.)
Binomial	binom	n, p	pbinom(x, n, p)
Chisquare	chisq	$df = m$	pchisq(x, m)
Exponential	exp	$r = 1/\beta$	pexp(x, r) $r - rate$
FDist	f	$df_1 = m, df_2 = n$	pf(x, m, n)
GammaDist	gamma	α, β (shape, scale)	pgamma$(x, \alpha, scale = 1/rate)$
Geometric	geom	p	pgeom(x, p)
Normal	norm	μ, σ	pnorm(x, μ, σ)
Poisson	pois	λ	ppois(x, λ)
TDist	t	$df = m$	pt(x, m)
Uniform	unif	min,max	punif(x, min, max)

1.10 Exercícios

1. Considere duas variáveis aleatórias X e Y cuja distribuição conjunta de probabilidade é dada no seguinte quadro:

X \ Y	1	3	6	9
2	0.11	0.05	0.2	0.08
3	0.2	0.02	0	0.1
7	0	0.05	0.1	0.09

   (a) Determine as distribuições marginais de X e de Y.

   (b) $Z = XY$. Determine a distribuição de probabilidade de Z.

   (c) Determine $E[X + Y]$.

2. Considere $X_1, X_2, \ldots, X_n$ variáveis aleatórias independentes que se admite seguirem uma distribuição uniforme $U(0, L)$.

   (a) Determine a função densidade de $X_{n:n}$, que designa $\max(X_1, X_2, \ldots, X_n)$.

   (b) Calcule o valor médio de $X_{n:n}$.

3. Suponha que as pessoas possuidoras de um dado gene X contraem uma dada doença em 50% dos casos. Suponha ainda que a prevalência do gene X é 1/1000 e a prevalência da doença acima referida é 1%.

(a) Determine a probabilidade de que uma pessoa escolhida ao acaso possua o gene X se tem a doença.

(b) * Observa-se um grupo de 300 pessoas seleccionadas ao acaso, verificando-se que 5 delas possuem a doença. Considera que o valor de 1% é admíssivel para a prevalência daquela doença? Justifique convenientemente a resposta.
*–para fazer mais adiante

4. Seja Y uma variável aleatória que designa o tempo de vida de uma molécula e se admite ter distribuição exponencial com valor médio α . Obtenha a função distribuição cumulativa do Mínimo de uma amostra de n variáveis i.i.d. a Y. Obtenha o valor médio desse Mínimo.

Soluções

1. (a)

x_i	2	3	7
p_i	0.44	0.32	0.24

y_j	1	3	6	9
p_j	0.31	0.12	0.3	0.27

(b)

z_i	2	3	6	9	...	63
p_i	0.11	0.2	0.05	0.02	...	0.09

(c) $E[X+Y] = E[X] + E[Y]$.

2. Se $X_1, X_2, \ldots, X_n$ seguem uma distribuição uniforme $U(0, L)$, tem-se $f(x) = 1/L, \quad 0 < x < L$ e $f(x) = 0$ para outros valores de x.

(a) É preciso calcular primeiro $F(x)$
$F(x) = \int_{-\infty}^{x} f(t)dt = x/L \quad 0 \leq x < L$, $F(x) = 0$ se $x < 0$ e $F(x) = 1$ se $x \geq L$
$F_{X_{n:n}}(x) = (x/L)^n$ se , $0 \leq x < L$ então
$f_{X_{n:n}}(x) = (n/L)(x/L)^{n-1}$, $0 < x < L$ e nula nos outros valores

(b) $E[X] = (nL)/(n+1)$

3. $P(D|X) = 0.5 \quad P(X) = 0.001 \quad P(D) = 0.01$

(a) $P(X|D) = 0.05$

(b) * esta alínea só pode ser feita depois de revistos os testes (capítulo 2.)
Temos que realizar um teste, das hipóteses $H_0 : p_D = 0.01 \quad vs \quad H_1 : p_D \neq 0.01$
Uma possibilidade é realizar uum teste do qui-quadrado cujo valor calculado é $X^2_{calc} = 1.346$. Como o nível crítico é $\chi^2_{0.05,(1)} = 3.84146$ somos levamos a concluir que não se rejeita H_0 portanto não temos razões para duvidar que o valor 1% possa ser a prevalência da doença.

4. $F_{Y_{min}}(y) = 1 - exp(-ny/\alpha)$, para $y > 0$. $\quad E[Y_{min}] = \alpha/n$.

Capítulo 2

Estimação e Inferência (algumas notas)

Os modelos que foram exemplificados até aqui necessitam da estimação dos parâmetros.

A **Estatística** tem como objectivo analisar dados e ajudar a interpretar a informação neles contida (análise de dados) e tirar conclusões ou fazer previsões a partir da análise desses dados (inferência estatística.) A Inferência Estatística integra duas grandes áreas: estimação e testes de hipóteses. Ambas têm grande utilização em Bioinformática

A Teoria da Estimação foi já introduzida na disciplina **Estatística**, onde conceitos como **estimador, estimativa, precisão de um estimador** foram dados. Os métodos então utilizados designam-se habitualmente por Métodos de estimação clássicos. Iremos apresentar dois métodos de estimação clássicos.

2.1 Métodos de Estimação

O Método dos Momentos

Introduzido por Karl Pearson no início do século XX, foi o primeiro método de estimação que foi estabelecido e que tem uma filosofia muito simples.

A ideia base deste método consiste em "utilizar os momentos da amostra para estimar os correspondentes momentos da população". Consiste em tomar como estimadores dos parâmetros desconhecidos as soluções das equações que se obtêm igualando os momentos teóricos aos momentos empíricos.

É um método de aplicação geral, tendo como única condição que a distribuição tenha um número suficiente de momentos(teóricos).

Dada a amostra aleatória $(X_1, X_2, \cdots, X_n)$ chama-se **momento de ordem** k e representa-se por μ'_k a $\mu'_k = E[X^k]$

O **momento empírico** correspondente é $\quad M'_k = (1/n)\sum_{i=1}^n X_i^k$.

Sejam $\theta_1, ..., \theta_k$ parâmetros desconhecidos de uma v.a. X.

O método dos momentos consiste em igualar momentos teóricos e momentos empíricos, i.e.,

$$\begin{array}{l} E[X] = m'_1 = \frac{1}{n}\sum_{i=1}^n x_i \\ E[X^2] = m'_2 = \frac{1}{n}\sum_{i=1}^n x_i^2 \\ \vdots \\ E[X^k] = m'_k = \frac{1}{n}\sum_{i=1}^n x_i^k \end{array}$$

com $m'_k = \frac{1}{n}\sum_{i=1}^n x_i^k$ calculado à custa da amostra $(x_1, ..., x_n)$.

Aquelas igualdades dão-nos estimativas que são a concretização dos estimadores correspondentes.

Os cálculos não são complicados, mas no cálculo dos momentos empíricos aparecem potências de expoente elevado quando há muitos parâmetros, conduzindo a estimativas instáveis.

Por isso, como regra prática deve evitar-se recorrer ao método dos momentos para mais de quatro parâmetros.

O método dos momentos–exemplo

Consideremos $X \frown \mathcal{N}(\mu, \sigma)$. Quais os estimadores de momentos de μ e σ?

Tem-se :

$$\begin{array}{lll} E[X] = \mu & \text{e} & M'_1 = \frac{1}{n}\sum_{i=1}^n X_i = \overline{X} \\ E[X^2] = \sigma^2 + \mu^2 & \text{e} & M'_2 = \frac{1}{n}\sum_{i=1}^n X_i^2 \end{array}$$

donde

$$\begin{array}{l} \mu^* = \overline{X} \\ \sigma^{2*} = \frac{1}{n}\sum_{i=1}^n X_i^2 - \mu^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2 \end{array}$$

Da resolução da equação que estabelece a igualdade dos momentos obtemos os estimadores dos momentos.

O Método da Máxima Verosimilhança

Este método é mais complexo mas regra geral produz melhores estimadores (podendo nalguns casos ser coincidente com os obtidos pelo método dos momentos)

Seja X uma v.a. cuja distribuição depende de um parâmetro θ.

Definição Seja $(X_1, X_2, ..., X_n)$ uma amostra aleatória. Chama-se **verosimilhança da amostra** e representa-se por $L(\theta|x_1, x_2, ..., x_n)$ a

$$L(\theta|x_1, \cdots, x_n) = \begin{array}{ll} f(x_1, ..., x_n|\theta) & \text{caso contínuo} \\ P(X_1 = x_1, ..., X_n = x_n|\theta) & \text{caso discreto} \end{array}$$

é a "probabilidade de se observar uma dada amostra aleatória".

$$L(\theta|x_1, ..., x_n) = \begin{array}{ll} \prod_{i=1}^{n} f(x_i|\theta) & \text{caso contínuo} \\ \prod_{i=1}^{n} P(X_i = x_i|\theta) & \text{caso discreto} \end{array}$$

O método da máxima verosimilhança, proposto por Fisher no início do século, consiste em escolher como estimativa de θ o valor que maximiza a verosimilhança $L(\theta|x_1, ..., x_n)$.

Muitas vezes (e sob certas condições) o máximo é obtido mais facilmente por derivação de $logL$, i.e., a estimativa de máxima verosimilhança é o valor de θ tal que

$$\frac{\partial \log L}{\partial \theta} = 0 \quad .$$

A solução desta equação, $\hat{\theta}(x_1, ..., x_n)$ é a estimativa de máxima verosimilhança, que é uma realização da v.a. $\hat{\Theta} = \hat{\Theta}(X_1, ..., X_n)$, a que se chama **estimador de máxima verosimilhança**.

2.2 Testes de Hipóteses do Qui-quadrado

Vamos aqui considerar testes de hipóteses de grande importância em Bioinformática aplicados a dados de **contagens - os testes do qui-quadrado**.

Será bom recordarmos que a realização de um **Teste Estatístico** envolve 5 etapas:

1. Formulação das hipóteses:
Hipótese nula – $\mathbf{H}_0$ e **Hipótese alternativa** – $\mathbf{H}_1$

2. Escolher uma variável aleatória – **Estatística do teste** que sob a hipótese $\mathbf{H}_0$ terá distribuição conhecida (pelo menos aproximadamente).

3. Definir a **região de rejeição** ou **região crítica** – **RC** (conjunto de valores da estatística que são pouco "plausíveis" caso $\mathbf{H}_0$ seja verdadeira, portanto levam a rejeitar $\mathbf{H}_0$).

Note-se que até aqui ainda não precisámos de dados, só agora vamos usar a amostra observada.

4. Calcular **o valor** da estatística do teste, para a amostra observada.

5. **Se o valor calculado $\in$ RC $\longrightarrow$ rejeita-se H_0**

Se o valor calculado $\notin$ RC $\longrightarrow$ não se rejeita H_0

As decisões de rejeitar a hipótese nula ou não rejeitar a hipótese nula envolvem erros.

Chama-se **erro de primeira espécie** ou **erro de tipo I** à rejeição de H_0 quando ela é verdadeira. À probabilidade de rejeitar H_0 se H_0 for verdadeira chama-se **nível de significância do teste**, habitualmente é denotada por $\boldsymbol{\alpha}$ e os valores usuais para α são $\boldsymbol{\alpha} = 0.05$ ou $\boldsymbol{\alpha} = 0.01$

Os **testes do qui-quadrado** que iremos considerar usam a contagem do número de valores que se observam em classes ou categorias, de uma variável de interesse. Podem ser aplicados ao estudo de uma só variável qualitativa (categórica) ou quantitativa mas categorizada ou então ao estudo da relação de duas ou mais variáveis classificadas em categorias.

Iremos considerar **testes do qui-quadrado de ajustamento** e **testes do qui-quadrado em tabelas de contingência** (testes de independência e de homogeneidade).

2.2.1 Teste do qui-quadrado de ajustamento

O objectivo deste teste é comparar as contagens observadas com as contagens que seriam de esperar se uma dada hipótes de modelo ou distribuição da população fosse verificada.

O teste sugerido por Karl Pearson (para o caso de as observações se encontrarem classificadas em k classes) consiste em considerar a estatística de teste:

$$\boldsymbol{X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}}$$

$k-$ nº de classes; O_i- frequência observada, ou seja a frequência absoluta na classe i; E_i- frequência esperada, i.e., é o número esperado de observações na classe i, dada então por $E_i = n\ p_i$, com p_i a probabilidade associada à classe i, se a hipótese $\mathbf{H_0}$ for verdadeira; n é o número (total) de observações independentes.

Interpretação: – o valor de X^2 em cada classe é uma medida da proximidade entre os dados e a Hipótese nula. Quanto menor for o valor de X^2 mais plausível é a hipótese $\mathbf{H_0}$.

Então neste teste, as hipóteses a formular são:

$\mathbf{H}_0$: a distribuição está de acordo com um dado modelo
$\mathbf{H}_1$: a distribuição não está de acordo com aquele modelo

sendo, como se viu, a estatística do teste $\boldsymbol{X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}}$

Pearson mostrou que, sob a validade da hipótese nula (e ainda verificadas algumas condições que veremos adiante), X^2 tem assintoticamentre distribuição $\chi^2_{(k-1)}$, $\boldsymbol{X^2 \sim \chi^2_{(k-1)}}$, onde k é o número de classes em que as observações foram agrupadas.

A região crítica, para um nível de significância α, é $\quad$ **RC:** $\boldsymbol{X^2_{cal} > \chi^2_{\alpha,(k-1)}}$

Portanto **a regra de decisão é:**

- Se $\boldsymbol{X^2_{calc} > \chi^2_{\alpha(k-1)}}$ - Rejeita-se H_o a um nível de significância α;
  Caso contrário - não se rejeita H_o

Exemplo

A descendência originada pelo cruzamento de dois dados tipos de plantas pode ser qualquer um dos três genótipos que representaremos por CC, cC e cc. Um modelo teórico de sucessão genética indica que os genótipos CC, cC e cc devem aparecer na razão $1:2:1$. Efectuou-se o cruzamento de 90 plantas daqueles dois tipos, tendo sido os descendentes classificados geneticamente de acordo com o registado na tabela:

Genótipos	CC	cC	cc
	18	44	28

Estarão estes dados de acordo com o modelo genético?

Vamos começar por formular as seguintes hipóteses:

$\mathbf{H}_0$: a distribuição está de acordo com um dado modelo
$\mathbf{H}_1$: a distribuição não está de acordo com aquele modelo
ou seja
$\mathbf{H}_0$ **:** $p_1 = 0.25, \quad p_2 = 0.5, \quad p_3 = 0.25$
$\mathbf{H}_{1:}$ pelo menos duas das probabilidades são diferentes do formulado.

Resolução:

	CC	cC	cc
O_i	**18**	**44**	**28**
p_i	**0.25**	**0.5**	**0.25**
$E_i = np_i$	**22.5**	**45**	**22.5**

$X^2_{cal} = \frac{(18-22.5)^2}{22.5} + \frac{(44-45)^2}{45} + \frac{(28-22.5)^2}{22.5} = 2.27$ $\quad$ e $\quad$ $\chi^2_{0.05,(2)} = 5.99.$

Conclusão: Como $X^2_{cal} < \chi^2_{0.05,(2)}$, não se rejeita $\mathbf{H}_0$ a um nível de significância de 5%, portanto não há razões para duvidar que os dados estejam de acordo com o modelo genético.

Testes do qui-quadrado - Algumas notas

- A variável X^2, tem distribuição assintótica. Qual a dimensão que a amostra deverá ter para que a aproximação à distribuição $\chi^2_{(k-1)}$ seja válida?
  - Alguns autores consideram que a aproximação só é válida se a frequência esperada, E_i, verifica $E_i \geq 5$;
  - Cochran (1954) sugeriu o seguinte critério, que é o que vamos seguir:
    - todas as classes deverão ter $E_i \geq 1$;
    - 80% das classes devem apresentar $E_i \geq 5$.

    Sempre que as frequências esperadas de algumas classes forem inferiores a 1 essas classes devem agrupar-se com as adjacentes por forma a atingir a frequência mínima desejada.
- Sempre que para determinar a probabilidade p_i for necessário estimar parâmetros, a distribuição χ^2 virá

$$\chi^2_{(k-1-\text{nº de parâmetros estimados})}$$

Voltemos ao exercício 1.2 e ver outro modo de resolver, utilizando ainda o R.

```
>prob<-c(.25,.75)
>padrao<-c("A"=280,"nA"=720)
>padrao
A  nA
280 720
> chisq.test(padrao, y=NULL,correct=F,p=prob)
        Chi-squared test for given probabilities
data:  padrao
X-squared = 4.8, df = 1, p-value = 0.02846

> chisq.test(padrao,p=prob) #se desse este commando teria
                            #o mesmo resultado
```

Observação — vamos calcular o p-value

```
 >pchisq(4.8,1,lower.tail = F)
```

Exercício 2.1 *Considere uma sequência de DNA de comprimento n=2000, na qual se observaram 520 ocorrências de A, 460 de C, 560 de G e 460 de T.*

Sob a hipótese da independência serão estes valores compatíveis com a hipótese de igual probabilidade de ocorrência de qualquer das bases?

2.2.2 Tabelas de contingência

Atrás foi considerado o estudo de uma amostra de valores observados de uma variável de interesse cujos de dados tinham sido classificados e o teste era aplicado às frequências absolutas.

Se os indivíduos de uma amostra retirada de uma dada população são classificados de acordo com dois critérios A e B (qualitativos ou quantitativos) é costume apresentar as frequências observadas numa tabela a que se chama **tabela de contingência**.

Consideremos $\boldsymbol{r}$ níveis do critério A e $\boldsymbol{c}$ níveis do critério B. O aspecto formal de uma tabela de contingência é:

	B_1	$\cdots$	B_j	$\cdots$	B_c	
A_1	O_{11}	$\cdots$	O_{1j}	$\cdots$	O_{1c}	$n_{1.}$
A_2	O_{21}	$\cdots$	O_{2j}	$\cdots$	O_{2c}	$n_{2.}$
.	.	.	.	.	.	.
A_i	O_{i1}	$\cdots$	O_{ij}	$\cdots$	O_{ic}	$n_{i.}$
.	.	.	.	.	.	.
A_r	O_{r1}	$\cdots$	O_{rj}	$\cdots$	O_{rc}	$n_{r.}$
	$n_{.1}$	$\cdots$	$n_{.j}$	$\cdots$	$n_{.c}$	n

Testes de independência

Na tabela anterior tem-se uma amostra de dimensão n e os seus elementos são classificados de acordo com os dois critérios, verificando-se:

$$\sum_{i=1}^{r}\sum_{j=1}^{c} O_{ij} = n; \qquad n_{i.} = \sum_{j=1}^{c} O_{ij}; \qquad n_{.j} = \sum_{i=1}^{r} O_{ij}$$

onde O_{ij} representa o número de elementos da amostra classificados nas categorias A_i e B_j.

O objectivo do estudo de uma tabela de contingência, como a apresentada, é tentar inferir sobre a existência ou não de alguma associação entre os dois atributos A e B, ou seja, pretende-se testar

$\mathbf{H}_{0:}$ na população em estudo há independância entre os critérios A e B;
$\mathbf{H}_{1:}$ na população em estudo não há independância entre os critérios A e B

Como vimos na UC Estatística, dizemos que há independência entre duas variáveis aleatórias se as probabilidades conjuntas são iguais ao produto das probabilidades marginais. Então as hipóteses doi teste podem apresentar-se formalmente como:

$\mathbf{H}_{0:}$ $p_{ij} = p_{i.} \times p_{.j}, \qquad \forall i, j$
$\mathbf{H}_{1:}$ $p_{ij} \neq p_{i.} \times p_{.j}, \quad$ pelo menos em 2 pares.

A estatística de Pearson do teste de independência é

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

onde $\boldsymbol{E_{ij}}$ representa a frequência esperada, isto é,

$$\boldsymbol{E_{ij} = np_{ij}}$$

Sob a validade da hipótese H_0 tem-se $\boldsymbol{E_{ij}}$ estimada por

$$\boldsymbol{e_{ij} = n\frac{n_{i.}}{n} \times \frac{n_{.j}}{n} = \frac{n_{i.}n_{.j}}{n}}$$

Se H_0 verdadeira, a estatística X^2 tem distribuição assintótica Qui-quadrado com $\boldsymbol{(r-1)(c-1)}$ graus de liberdade.

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

Algumas notas sobre testes de independência.

Também aqui há **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 (tendo em conta se tal junção tem 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.

Exemplo *(retirado de Dagnelie, vol. II)*

Submeteram-se ramos florais da macieira "Golden Delicious", em números sensivelmente iguais, a quatro tratamentos e contou-se o número de frutos produzidos em cada caso, a fim de verificar se existe ou não uma relação entre os diferentes tratamentos e a frutificação.

Vejamos os resultados no seguinte quadro:

Tratamentos	N. de frutos			Totais
	0	1	2ou 3	
A	203	150	6	359
B	266	112	1	379
C	258	126	2	386
D	196	168	17	381

Pretendemos testar a hipótese nula, de que não há relação entre os tratamentos e a frutificação, ou seja, que há independência entre a tipo de tratamento e a frutificação.

Resolução no R

Vamos introduzir a matriz dos dados

```
> trat.frutos<-matrix(c(203,150,6,266,112,1,258,126,2,196,168,17),
+nc=3,byrow=T)
> trat.frutos
     [,1] [,2] [,3]
[1,]  203  150    6
[2,]  266  112    1
[3,]  258  126    2
[4,]  196  168   17
```

Mas para as linhas e colunas ficarem identificadas pode fazer-se

```
> trat.frutos<-matrix(c(203,150,6,266,112,1,258,126,2,196,168,17),
+nc=3,byrow=T,
+ dimnames = list(c("A", "B","C","D"),c("0", "1", " 2/3")))
> trat.frutos
```

```
     0   1   2/3
A  203 150   6
B  266 112   1
C  258 126   2
D  196 168  17
```

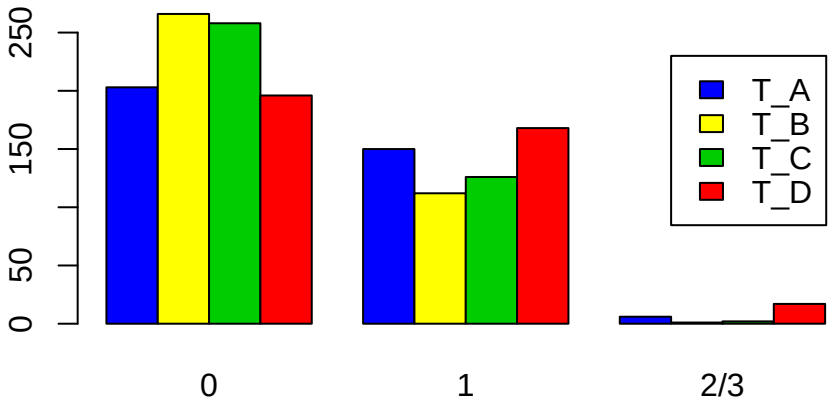
```
> chisq.test(trat.frutos)

        Pearson's Chi-squared test

data:  trat.frutos
X-squared = 53.7199, df = 6, p-value = 8.402e-10
```

Já agora vejamos o gráfico com as contagens

```
> barplot(trat.frutos,names=c("0", "1", "2/3"),
+col=c(4,7,3,2),cex.names=1,beside=T)
> legend("topright",c("A", "B","C","D"),fill=c(4,7,3,2))
```



Como se rejeita a hipótese nula, vamos tentar compreender um pouco mais do que se passará. Vamo ver os valores esperados sob a hipótese nula e ainda as parcelas do cálculo da estatística.

```
> chisq.test(trat.frutos)$expected
          0       1        2/3
A    220.1708 132.6272 6.201993
B    232.4365 140.0159 6.547508
C    236.7296 142.6020 6.668439
D    233.6631 140.7548 6.582060

> chisq.test(trat.frutos)$residuals^2
          0        1         2/3
A    1.339120 2.275646  0.006578742
B    4.846508 5.605742  4.700238162
C    1.911173 1.932835  3.268279112
D    6.070752 5.273709 16.489287743
```

Nota: Verifique-se que a linha correspondente ao tratamento D apresenta valores esperados bem diferentes dos observados, assim como os maiores quadrados dos resíduos obviamente, i.e., este tratamento é o que mais contribui para o valor elevado do X^2.

Interpretação???

Testes de Homogeneidade

No **caso referido atrás** – **teste do qui-quadrado de independência** – considerava-se que se retirava a amostra que depois 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. Então neste caso, **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.

Contudo, o total das linhas ou das colunas da tabela de contingência pode estar sob o controle do investigador, i.e., **uma das margens da tabela pode ser fixa**, no caso de o número de observações em cada nível de um dos critérios ser fixado à partida. Suponhamos (para facilitar a explicação) que os totais em cada linha (critério A) são fixos. Agora estamos interessados em testar se a distribuição dos níveis do critério B é a mesma para todos os níveis do critério A. O teste a realizar chama-se **teste do qui-quadrado de homogeneidade** e há necessidade de estimar as probabilidades dos níveis do critério B (continuando com o que foi suposto).

Exemplo - Recolhe-se uma amostra de condutores, estratificados por classe etária, para averiguar se tiveram algum acidente no ano anterior e, em caso afirmativo se foi de maior ou menor gravidade. O resultados encontram-se na seguinte tabela, registados por grupos de idade.

Idade	Tipo de acidente		
	Nenhum	menor	maior
Inferior a 18	67	10	5
18-25	42	6	5
26-40	75	8	4
40-65	56	4	6
over 65	57	15	1

Realize um teste de homogeneidade do qui-quadrado para verificar se existe diferença na distribuição do tipo de acidente em cada classe etária.

Portanto agora estamos preocupados em responder à questão: **as amostras foram retiradas de populações homogéneas no que se refere a um dos critérios de classificação** - ou seja as classes etárias apresentam o mesmo comportamento face ao tipo de acidente?

Neste exemplo as linhas representam as populações (níveis do critério) 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.

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

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

onde $\boldsymbol{e_{ij}}$ é uma estimativa da Frequência Esperada.

Vejamos, por exempo, se as populações fossem homogéneas, i.e., se o comportamento face aos acidentes for o mesmo em cada uma das classes etárias, então, dado que a proporção da população que pertence, por exemplo, à categoria "Inferior a 18", é $n_{1.}/n = 82/361$, esperamos encontrar $n_{.1}n_{1.}/n$ observações na 1ª célula.

As frequências esperadas são então:

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

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

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

Resolver o exercício no R – note que todo o procedimento é igual ao que se usa num teste de independência, diferindo apenas a interpretação.

2.3 Exercícios

1. 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 a administração da vacina e a ocorrência ou não de gripe.

2. É realizado um teste genético para averiguar qual o alelo que os indivíduos em teste possuem e um teste de diagnóstico para averiguar se o indivíduo possui uma dada doença. Os dados encontram-se no seguinte quadro

	Doença	
	Sim	Não
Alelo l	45	122
Alelo 2	67	38

   Pretende-se averiguar se o tipo de Alelo e a existência ou não de doença são independentes. Responda à questão de forma completa.

3. Foi observada uma sequência de 1000 bp de um genoma de M. genitalium. Admitindo um modelo onde é válida a independência da ocorrência de nucleótidos na sequência, verifique se as frequências observadas para cada dinucleótido, representadas na tabela seguinte, são compatíveis com as seguintes hipóteses probabilísticas associadas a cada uma das bases

$$p_A = 0.45, \quad p_C = 0.09; \quad p_G = 0.09; \quad p_T = 0.37$$

Dinucleótido	AA	AC	AG	AT	CA	CC	CG	CT
Frequência	202	42	39	160	38	8	11	32
Dinucleótido	GA	GC	GG	GT	TA	TC	TG	TT
Frequência	40	10	8	32	175	36	31	135

   Apresente os cálculos necessários e justifique convenientemente a resposta.

   Nota: Tenha em conta que (n-1) é o número de dinucleótidos numa palavra de comprimento n.

4. Para estudar o efeito do tipo de solo no crescimento de uma dada planta, plantaram-se amostras em três tipos de solo e classificou-se o crescimento das plantas em três categorias, tendo-se obtido os resultados seguintes:

	Tipo de Solo		
Crescimento	Barrento	Arenoso	Orgânico
Pobre	16	8	14
Médio	31	16	21
Bom	18	36	25

Com base nestes dados, será que se pode afirmar que o crescimento da planta difere significativamente consoante o tipo de solo ?

5. Suponha que se dispõe de 50 vasos em cada um dos quais foram semeadas 5 sementes, que se deixaram germinar. Ao fim de um certo tempo contou-se o número de sementes germinadas por vaso, tendo-se obtido

$1, 2, 0, 0, 1, 4, 2, 5, 1, 1, 5, 0, 2, 2, 3, 2, 1, 0, 0, 3, 0, 2, 0, 4, 1$

$3, 3, 2, 2, 5, 5, 0, 3, 1, 0, 0, 1, 1, 2, 0, 4, 1, 4, 0, 3, 4, 2, 3, 1, 1$

Poder-se-á admitir que o número de sementes germinadas por vaso segue uma distribuição Binomial (5,p) com p a estimar a partir dos dados?

Soluções de alguns Exercícios

1. Aqui vai o script do R. Tentem correr e terão as respostas que procuram.

```
gripe<-matrix(c(24,9,13,289,100,565),nc=3,byrow=T)
gripe
margin.table(gripe,1)
margin.table(gripe,2)
chisq.test(gripe)
chisq.test(gripe)$expected
chisq.test(gripe)$residuals^2
```

3. O que nos é solicitadoa aqui é a realização de um teste de ajustamento do qui-quadrado. As hipóteses são:

$H_0 : p_A = 0.45;\ p_C = 0.09;\ p_G = 0.09;\ p_T = 0.37 \qquad vs \qquad H_1$: pelo menos 2 diferentes

Obtive o valor para $\chi^2_{cal} = 3.0264$ e a Estatística de Teste é $X^2 \simeq \chi^2_{(15)}$

Apresente os cálculos necessários e justifique convenientemente a resposta.

Nota: Tenha em conta que (n-1) é o número de dinucleótidos numa palavra de comprimento n.

5. Vamos usar e interpretar o seguinte script

```
semente<-c(1,2,0,0,1,4,2,5,1,1,5,0,2,2,3,2,1,0,0,3,
       3,3,2,2,5,5,0,3,1,0,0,1,1,2,0,4,1,4,0,3,4,2,3,1,1,0,2,0,4,1)
semente
length(semente)
mean(semente)
table(semente)

#Note que, como em cada vaso se colocaram 5 sementes, para ajustarmos a
# uma variável binomial terá que ser B(5,p)
#com p - desconhecido, que vamos estimar

p_est<-mean(semente)/5;p_est
p_est

# Vamos agora ao teste do qui-quadrado
x<-0:5
p<-dbinom(x,5,0.372)
p
# deu  0.097678328 0.289302055 0.342740015 0.203024340 0.060131413 0.007123849


par(mfrow=c(1,2))
plot(x,dbinom(x,5,0.372),type="h",ylim=c(0,0.5)) #por exemplo pôr estes limites
plot(table(semente)/length(semente),ylim=c(0,0.5))

val<-c(12,12,10,7,5,4)
pval<-c(0.098, 0.289, 0.343, 0.203,0.06,0.007)
sum(pval) # para confirmar que a soma é 1
chisq.test(val, p = pval)
#vamos guardar para analisarmos as parcelas do qui-quadrado
result<-chisq.test(val, p = pval);result
result$expected
result$residuals^2

#Cálculo do p-value
pchisq(result$statistic,result$parameter-1,lower.tail = F)
```

2.4 Aplicação de métodos de reamostragem em Bioinformática - o *bootstrap*– introdução

.

As ideias iniciais de reamostragem e a orientação para uma linha de trabalho de análise estatística baseada em simulações, surgem nos anos 40 /50 graças ao grande impulso verificado nos computadores.

Os métodos de que falamos, chamados **métodos de reamostragem**, são baseados em procedimentos repetidos sobre muitos conjuntos de réplicas dos próprios dados e

- substituem complicadas e por vezes "grosseiras" aproximações por simulações;
- despertam fortemente a atenção, quer de investigadores teóricos, quer de utilizadores de métodos estatísticos.

O *jackknife* e o *bootstrap* são os métodos de reamostragem mais populares usados na análise estatística que tentam usar a informação existente na amostra para estimar a variabilidade e a distribuição de uma estatística de interesse, principalmente em modelos complexos, para os quais as expressões analíticas são difíceis de tratar.

Primeiros trabalhos devem-se a Quenouille (1949) e a Efron (1979).

Em filogenia este procedimentos começaram a ser utilizados em 1982.

Ao contrário dos procedimentos de estimação já atrás referidos, estes métodos não necessitam da especificação paramétrica do modelo.

A ideia básica da metodologia ***bootstrap não paramétrico***, Efron (1979), consiste em a partir da amostra observada $x_1, x_2, \ldots, x_n$ e tendo um estimador de interesse, calcular versões desse estimador sobre réplicas daquela amostra.

Formalmente tem-se

$\underline{X}_n = (X_1, X_2, \cdots X_n) \overset{i.i.d.}{\sim} F(.|\theta)$ $\qquad$ T_n um estimador de θ

A ideia baseia-se em: a partir da amostra obervada $\underline{x}_n = (x_1, x_2, \ldots, x_n)$ construir a *amostra bootstrap* $\underline{X}^*_n = (X^*_1, X^*_2, \cdots X^*_n)$, constituída por variáveis reamostradas de $\underline{x}_n$ de acordo com a função de distribuição empírica,

$$\widehat{F}_n(x) = \frac{\#\{i : x_i \leq x, 1 \leq i \leq n\}}{n},$$

sendo a **extracção é feita com reposição**, i.e.,

$$P(X^*_i = X_j | \underline{x}_n) = \frac{1}{n} \qquad i, j = 1, \cdots n$$

$$T^*_n := T_n\left(\underline{X}^*_n\right) \quad \textbf{\textit{versão bootstrap}} \textbf{ do estimador } T_n$$

O comportamento da *versão bootstrap* deverá simular o comportamento de T_n.

A distribuição de T_n^*, é usada para aproximar a distribuição de amostragem (desconhecida), de T_n.

Procedimento da metodologia Bootstrap não paramétrico

Recurso à simulação de Monte Carlo.

- dada uma amostra observada $\underline{x}_n = (x_1, x_2, \cdots x_n)$, constrói-se $\widehat{F}_n$ atribuíndo a cada x_i peso $1/n$;
- gera-se uma amostra *bootstrap* $\underline{x}_n^* = (x_1^*, x_2^*, \cdots x_n^*)$, de variáveis $X_i^* \overset{i.i.d.}{\sim} \widehat{F}_n$ e calcula-se $t_n^* = t_n(x_1^*, x_2^*, \cdots, x_n^*)$;
- repete-se, independentemente, B vezes o passo anterior, obtendo assim B réplicas $(t_n^{*,1}, t_n^{*,2}, \cdots t_n^{*,B})$
- calcula-se as estimativas do viés, erro padrão e distribuição de amostragem *bootstrap*

$$\widehat{Vies_B^*}[T_n] = \sum_{i=1}^{B} t_n^{*,i}/B - \theta(\widehat{F}_n) \qquad \widehat{\sigma_B^*}[T_n] = \sqrt{\frac{1}{B-1}\sum_{i=1}^{B}\left(t_n^{*,i} - \overline{t_n^{*,i}}\right)^2}$$

$$\widehat{F_B^*}(t) = \frac{\#\left\{i : t_n^{*,i} \leq t, 1 \leq i \leq B\right\}}{B}, \qquad -\infty < t < +\infty$$

Nota: $\lim_{B\to\infty} \widehat{Vies_B^*} = Vies^* \quad \lim_{B\to\infty} \widehat{\sigma_B^*} = \sigma^* \quad \lim_{B\to\infty} \widehat{F_B^*} = F^*$

Desvio padrão — estimativa *bootstrap* é regra geral bastante boa, com um número pequeno de réplicas, Efron (1993) considera suficiente $B = 200$.

Viés — convergência é mais difícil de atingir.

Intervalos de confiança *bootstrap* — há vários procedimentos, mas aqui vamos apenas referir um deles:

O métodos dos percentis — $\left(t_{\alpha/2}^*, t_{1-\alpha/2}^*\right)$, que são os percentis empíricos dos valores bootstrap $t_n^{*,i}$.

Veremos um exemplo simples de aplicação do *Bootstrap* em árvores de filogenia, *(Ewens and Grant).*

2.5 Exercícios Finais e de Revisão

1. Os dados que se encontram na tabela a seguir apresentada correspondem a um estudo sobre o efeito da droga Mesalamina em pacientes com problemas de úlcera moderados. As duas variáveis são o Tratamento utilizado e o Resultado. Esta última variável corresponde ao estado da úlcera seis semana após o início do tratamento. Neste estudo participaram 131 indivíduos.

Resultado	Tratamento		
	Placebo	Dose Baixa	Dose Elevada
Em Remissão	2	6	6
Melhorou	8	13	15
Manteve-se	10	11	16
Piorou	22	14	8

   (a) Qual a proporção de pacientes com úlceras em remissão?

   (b) De entre todos os indivíduos, qual a percentagem que tomou placebo e a úlcera redimiu ou melhorou?

   (c) Considere que lhe foram inicialmente fornecidos apenas os dados relativos ao Resultado piorou. Admite-se que a probabilidade de resposta negativa ao tratamento neste caso é a mesma para ambas as doses mas que quando se aplica placebo a probabilidade de piorar é o triplo da que se verifica para cada dose. Os dados dão compatíveis com esta suposição? Justifique convenientemente.

   (d) Poder-se-á afirmar que existe associação entre o tratamento e o resultado obtido? Consulte o Anexo que contém alguns dados em falta que deve completar(valores A e B) e responda à questão colocada.

Anexo

```
> droga<-matrix(c(2,6,6,8,13,15,10,11,16,22,14,8),nc=3,byrow=T,
+ dimnames = list(c("Remissao", "Melhorou",
 "Manteve","Piorou"),c("Placebo", "Dose baixa","Dose
  elevada")));droga

         Placebo Dose baixa Dose elevada
Remissao       2          6            6
Melhorou       8         13           15
Manteve       10         11           16
Piorou        22         14            8
>
> chisq.test(droga)
```

```
        Pearson's Chi-squared test

data:  droga
X-squared = 12.8631, df = A, p-value = 0.04526

>
> chisq.test(droga)$expected
          Placebo Dose baixa Dose elevada
Remissao  4.48855    4.70229      4.80916
Melhorou 11.54198   12.09160     12.36641
Manteve  11.86260   12.42748     12.70992
Piorou   14.10687   14.77863     15.11450

> (chisq.test(droga)$residuals)^2
           Placebo Dose baixa Dose elevada
Remissao 1.3797061 0.35813423        B
Melhorou 1.0869583 0.06824447    0.5608567
Manteve  0.2924538 0.16396740    0.8516654
Piorou   4.4163940 0.04102265    3.3488473
```

2. Uma variável aleatória diz-se ter distribuição de Pareto se a sua função densidade é da forma

$$f(x) = \alpha\ x^{-\alpha-1}, \quad x > 1 \quad \alpha > 0$$

(α parâmetro desconhecido)

   (a) Com base numa amostra aleatória de dimensão n determine o estimador de máxima verosimilhança de α.

   (b) Tendo observado a seguinte amostra
   2 2.1 1.7 4.2 1.3 1.3 2 4 1.9 2.5 3.6 4.2 3.5
   determine uma estimativa de α.

3. Num estudo sobre a disseminação de sementes motivada pelos dejectos de aves (a acção dos ácidos do tracto digestivo sobre a casca de muitos tipos de sementes é importante para a sua germinação) dividiu-se um terreno em 50 parcelas e contou-se o número de sementes em cada uma delas. Os resultados obtidos encontram-se resumidos na tabela abaixo:

Nº de observações por parcela	0	1	2	3	8	
Nº de parcelas	11	13	14	7	5	(n=50)

   Pretende-se testar a hipótese $H_0 :\ X \frown Poisson(\lambda)$. Resolva o exercício de acordo com as seguintes alíneas:

(a) Determine o estimador de máxima verosimilhança para com base numa amostra aleatória.

(b) Calcule a estimativa de máxima verosimilhança para a amostra observada.

(c) Realize o teste proposto.

4. Numa sondagem telefónica é perguntado aos entrevistados o seu grau de concordância com a afirmação "Fumar deve ser proibido em locais públicos". Os resultados obtidos foram os seguintes:

Sexo	Grau de concordância				
	Concorda forte/	Concorda	Neutro	Discorda	Discorda forte/
Feminino	40	28	16	37	5
Masculino	16	25	11	25	11

Com base nestes dados poder-se-á concluir que os dois sexos diferem no que respeita ao seu grau de concordância sobre a proibição de fumar em lugares públicos? Elabore um estudo completo, tendo em conta o que lhe é dado no output abaixo

```
> respostas<-matrix(c(40,16,38,25,16,11,37,25,5,11),nr=2,
+ dimnames = list(c("Female", "Male"),
+ c("St.Agree", "Agree", "Neutral","Disagree","St.Disagree")))

> respostas
       St.Agree Agree Neutral Disagree St.Disagree
Female       40    38      16       37           5
Male         16    25      11       25          11
> chisq.test(respostas)

        Pearson's Chi-squared test

data:  respostas
X-squared = 8.5748, df = 4, p-value = 0.07265

> (chisq.test(respostas)$residuals)^2
       St.Agree       Agree     Neutral   Disagree St.Disagree
Female 1.058824 0.001633987 0.009414877 0.01097858    2.287815
Male   1.636364 0.002525253 0.014550265 0.01696690    3.535714

> chisq.test(respostas)$expected
       St.Agree  Agree  Neutral  Disagree  St.Disagree
Female       34   38.25  16.39286 37.64286    9.714286
Male         22   24.75  10.60714 24.35714    6.285714
```

Soluções de alguns Exercícios

3. **(a)** Pretende-se o estimador de máxima verosimilhança para λ construído com base numa amostra aleatória.

Se $X \frown Poisson(\lambda) \Rightarrow P[X = x] = \frac{e^{-\lambda}\lambda^x}{x!}$. Então dada a amostra aleatória $X_1, X_2, \cdots, X_n$, a verosimilhança é definida como

$$L(\lambda|x_1, x_2, \cdots, x_n) = \Pi_{i=1}^n \; P[X = x_i|\lambda] = \frac{e^{-n\lambda}\lambda^{\sum_i x_i}}{\Pi_i x_i!}$$

Vamos agora logaritmizar e depois derivar em ordem a λ

$$lnL(\lambda|x_1, x_2, \cdots, x_n) = -n\lambda + \sum_i x_i ln(\lambda) - ln(\Pi_i x_i!)$$

Derivando, igualando a zero e resolvendo a equação em ordem a λ, obtemos a estimativa de máxima verosimilhança para $\lambda \quad \hat{\lambda} = \sum_i x_i/n$, sendo o estimador associado definido como $\hat{\lambda} = \sum_i X_i/n$

(b) λ é então estimado pela média da amostra, portanto $\hat{\lambda} = \overline{x} = 2.04$

(c) O teste a realizar é um teste de ajustamento do qui-quadrado.

Pretende-se testar a hipótese $H_0 : \; X \frown Poisson(\lambda)$.

Para realizarmos este teste é necessário considerar uma estimativa de λ. Vamos então considerar $\hat{\lambda} = 2.0$ e construir o quadro de apoio à realização do teste:

Valores da variável	0	1	2	3	≥ 4	Total
O_i	11	13	14	7	5	50
p_i	0.135	0.271	0.271	0.180	0.143	1
e_i	6.75	13.55	13.55	9	7.15	50

com O_i – frequências observadas; p_i - probabilidades calculadas sob a hipótese nula (ver exemplo abaixo); $e_i = np_i$ – frequências esperadas.

Por exempplo $\quad p_0 = P[X = 0] = 0.135; \quad p_1 = P[X = 1] = 0.271$ e como depois de se ter observado o valor 3 só apareceu 8 e a variável que se está a colocar em hipótese nula toma um nº infinito de valores vamos considerar na última classe $P[X \geq 4] = 0.143$.

Cálculo do valor da estatística de teste

$X^2_{calc} = \sum_{i=1}^5 \frac{(O_i - e_i)^2}{e_i} = \frac{11 - 6.75)^2}{6.75} + \cdots + \frac{5 - 7.15)^2}{7.15} = 3.8040$

A região crítica é definida como $X^2_{calc} > \chi^2_{0.05,(5-1-1)} \Leftrightarrow X^2_{calc} > \chi^2_{0.05,(3)} = 7.81$

Dado que $X^2_{calc} \not> \chi^2_{0.05,(3)}$ não rejeitamos H_0, portanto a um nível de significância de 5% podemos admitir que X segue uma lei de Poisson.