LGN5830 - Biometria de Marcadores Genéticos

Tópico 2: Verossimilhança

Antonio Augusto Franco Garcia
http://about.me/augusto.garcia
augusto.garcia@usp.br

Departamento de Genética
ESALQ/USP
2015

STATISTICAL GENETICS LAB · ESALQ / USP

UNIVERSIDADE DE SÃO PAULO
DEPARTAMENTO DE GENÉTICA

Conteúdo

Definições

Regras

- Adição
$$P(A \text{ ou } B) = P(A) + P(B) - P(A \text{ e } B)$$
- Adição (eventos mutuamente exclusivos)
$$P(A \text{ ou } B) = P(A) + P(B)$$
- Subtração
$$P(A) = 1 - P(\text{não } A)$$
- Multiplicação
$$P(A \text{ e } B) = P(A) \times P(B|A)$$
- Multiplicação (A e B independentes)
$$P(A \text{ e } B) = P(A) \times P(B)$$
- Notação: $P(A \text{ e } B) = P(A \cap B) = P(A, B)$

Probabilidade Condicional

Dois dados com cores diferentes

- Se eu jogar os dois dados simultaneamente, qual é a probabilidade de obter soma 3?
  - # resultados possíveis: $6 \times 6 = 36$
  - # resultados com soma 3: 2 ($\{1,2\}$, $\{2,1\}$)
  - Resp: $P(\text{soma } 3) = 2/36$

Probabilidade Condicional

Dois dados com cores diferentes

- Suponha agora que um dos dois dados foi jogado antes, e o resultado foi 1
- Qual a probabilidade de obter soma 3?
  - # resultados possíveis: 6
  - # resultados com soma 3: 1 ($\{1,2\}$)
  - Resp: $P(\text{soma } 3 | \text{valor } 1 \text{ em um dos dados}) = 1/6$

Probabilidade Condicional

$P(A|B)$

$$P(A|B) = \frac{P(A,B)}{P(B)}$$

Atenção

- Note a relação entre probab. condicional e a regra da multiplicação
- O que significam $P(A|B) = 1$ e $P(A|B) = 0$?
- Eventos independentes: $P(A,B) = P(A) \times P(B)$

Exemplo anterior

$$P(A|B) = \frac{P(A,B)}{P(B)} = \frac{\frac{1}{36}}{\frac{1}{6}} = \frac{1}{6}$$

Eventos independentes

Moeda "honesta"

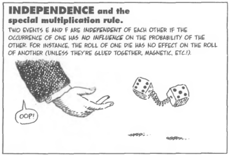
- Qual a probabilidade de obter uma sequência de 4 caras?
- Resp: $\left(\frac{1}{2}\right)^4$

Eventos independentes

Moeda "honesta"

Qual a probabilidade de obter uma sequência de 4 caras?

Eventos independentes



Um caso simples

Doença, Genótipo

	mm	Mm	MM	
R	0.10	0.21	0.47	0.78
S	0.05	0.09	0.08	0.22
	0.15	0.30	0.55	1

- $P(D = R) = 0.78$
- $P(G = Mm) = 0.30$
- $P(D = R|G = MM) = \frac{P(D=R,G=MM)}{P(G=MM)} = \frac{0.47}{0.55} = 0.85$
- $P(D = R, G = MM) = P(D = R)\,P(G = MM|D = R) = 0.78 \times \frac{0.47}{0.78} = 0.47$
- Note que $P(D = R).P(G = MM) = 0.78 \times 0.55 = 0.429$

Teorema de Bayes

Thomas Bayes, 1701–1761

$$P(A|B) = \frac{P(A)\,P(B|A)}{P(B)}$$

- $P(A)$: "priori"
- $P(A|B)$: "posteriori"
- $P(B|A)/P(B)$: suporte que B fornece para A

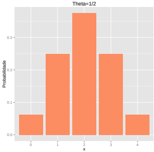
Variável Discreta

Exemplo - Distribuição Binomial

- Seja θ a proporção de indivíduos Aa numa população grande e homogênea, proveniente de um retrocruzamento.
- Neste caso, temos teoricamente 50% dos indivíduos com este genótipo ($\theta = 1/2$)
- Qual a probabilidade de observarmos x genótipos Aa numa amostra de 4 indivíduos ($n = 4$)?
  - $P(x) = \left(\begin{array}{c} n \\ x \end{array} \right) \theta^x (1-\theta)^{(n-x)}$
- Note que estamos assumindo que os eventos são independentes!

Distribuição Binomial

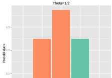
Exemplo



Distribuição Binomial

Exemplo

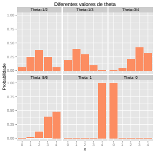
- Qual a probabilidade de observarmos 3 genótipos Aa ($x = 3$) numa amostra de 4 indivíduos ($n = 4$)?
  - $P(3) = \left(\begin{array}{c} 4 \\ 3 \end{array} \right) (1/2)^3 \, [1 - (1/2)]^{(4-3)} = 1/4$



Distribuição Binomial

Outras distribuições

- E se θ tiver outros valores?

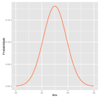


Distribuição Normal

- Grande parte das variáveis (caracteres) estudados na Genética

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

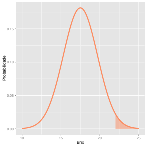
Exemplo: teor de açúcar numa população de cana-de-açúcar



- Concentração em torno da média, dispersão, indivíduos raros, etc

Densidade de Probabilidades

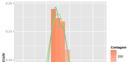
Qual a probab. de selecionar ind. com Brix acima de 22?



Modelo vs dados reais

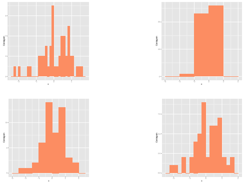
- É óbvio que dados reais não estão "classificados"
- Várias técnicas são empregadas (histogramas, boxplots, ramo-e-folhas, etc

Brix, 200 valores



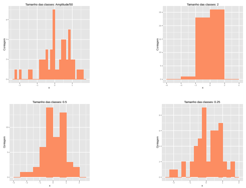
Exercício

Qual(is) conj. foram amostrado(s) em pop. com dist. normal?



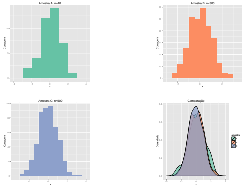
Exercício

Todos eles! ($n = 50$)



Exercício

Tamanho da amostra



História - Tycho Brahe

Movimento de corpos celestiais

Variável Discreta

Retrocruzamento, com interesse em Aa

- Imagine um exp. hipotético não realizado, com 4 indivíduos

$$\underbrace{1, 1, \ldots, 1}_{\text{x}}, \underbrace{0, 0, \ldots, 0}_{\text{n-x}}$$

- Resultados possíveis:

Aa	0	1	2	3	4
Probabilidade	$P(0)$	$P(1)$	$P(2)$	$P(3)$	$P(4)$

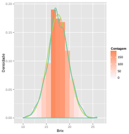
- Média esperada:

$$E(X) = \sum xP(x)$$

- Na distribuição binomial, demonstra-se que $E(X) = np$
- No caso, $E(X) = 4(1/2)$, ou seja, 2 indivíduos com genótipo Aa

Variável Contínua

Brix de 200 indivíduos, cana-de-açúcar



- Qual a média desse experimento, com base no histograma?

Variável Contínua

- Qual a média esperada para uma variável contínua?
- Esperança Matemática:

$$E(X) = \int_{-\infty}^{\infty} x f(x)\, dx$$

- Demonstra-se que, no caso da distribuição normal, $E(X) = \mu$

Alguns conceitos

Experimento

- Conjunto de dados
- Informações sobre como esses dados foram coletados

Inferência estatística

- Desejamos explicitar o modelo que deu origem aos dados
- Usualmente, o modelo envolve um ou mais parâmetros desconhecidos
- Os parâmetros devem ser estimados a partir dos dados

Método da Verossimilhança

- Suponha que um modelo probabilístico tenha sido formulado para um experimento
- Imagine que esse modelo envolva um parâmetro θ
- Desejamos usar os dados para estimar θ
- Formalmente, desejamos determinar quais são os possíveis valores de θ mais plausíveis (prováveis, verossímeis), à luz das observações

Método da Verossimilhança

Exemplo

- Seja θ a proporção de indivíduos Aa numa população grande e homogênea, com 2 alelos para esse loco.
- Desejamos estimar essa proporção.
- Para tanto, selecionamos aleatoriamente n indivíduos e verificamos seu genótipo.
- Após o experimento, notamos que x deles são Aa
- A probabilidade de observarmos esse evento E é $P(E;\theta) =$ probab. de x, de um total de n indivíduos, possuírem o genótipo Aa

$$P(E;\theta) = \binom{n}{x} \theta^x (1-\theta)^{(n-x)}$$

Método da Verossimilhança

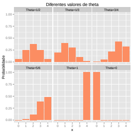
Exemplo

- Suponha que $x = 3$ e $n = 4$
- Note que, nesta situação, θ não é conhecido
- $P(E;\theta) = \binom{4}{3} \theta^3 (1-\theta)^{(4-3)}$
  - Se $\theta = 1/2$, $P(E;\theta) = 0.25$
  - Se $\theta = 1/3$, $P(E;\theta) = 0.10$
  - Se $\theta = 3/4$, $P(E;\theta) = 0.42$
  - Se $\theta = 5/6$, $P(E;\theta) = 0.39$
  - Se $\theta = 1$, $P(E;\theta) = 0$
- Qual valor de θ é mais plausível?

Verossimilhança

Distribuições

- De qual distribuição os dados foram amostrados?



- Note que é mais fácil rejeitar do que aceitar

Método da Verossimilhança

Definição

- A função de verossimilhança de θ é definida como $L(\theta) = c \cdot P(E;\theta)$
- Função de Verossimilhança: função densidade de probabilidade das observações, interpretada como uma função dos parâmetros que determinam a distribuição (Siegmund e Yakir, 2007)
- Edwards (1992): The *likelihood* $L(H|R)$, of the hypothesis H given data R, and a specific model, is proportional to $P(R|H)$, the constant of proportionality being arbitrary.

Método da Verossimilhança

Definição

- Likelihood is the hypothetical probability that an event that has already occurred would yield a specific outcome. The concept differs from that of a probability in that a probability refers to the occurrence of future events, while a likelihood refers to past events with known outcomes. (http://mathworld.wolfram.com/Likelihood.html)
- $L(\theta) \propto P(E;\theta)$
- $L(\theta) \propto \theta^x(1-\theta)^{(n-x)}$ (no caso da dist. binomial)
- A constante c, por não depender dos parâmetros, normalmente é desconsiderada

Verossimilhança

Definição

- Sorensen e Gianola (2002): Sejam **y** os dados observados, resultado de um processo estocástico caracterizado por um modelo com distribuição (densidade) $p(\mathbf{y}|\theta)$
- A distribuição (densidade) das observações é portanto $p(\mathbf{y}|\theta)$
- A verossimilhança $L(\theta)$ ou $L(\theta|\mathbf{y})$ é obtida com base em uma "inversão" deste conceito
- Por definição: $L(\theta|\mathbf{y}) \propto p(\mathbf{y}|\theta)$

Exemplo - genótipos Aa

- Dados: $\mathbf{y_i}$ $(i = 1,\ldots,n; n = 4)$

$$\underbrace{1,1,\ldots,1}_{\text{x}},\underbrace{0,0,\ldots,0}_{\text{n-x}}$$

- $p(\mathbf{y_i}|\theta) = \prod_{i=1}^{n} p(\mathbf{y_i}|\theta)$

$$\underbrace{\theta,\theta,\ldots,\theta}_{\text{x}},\underbrace{(1-\theta),(1-\theta),\ldots,(1-\theta)}_{\text{n-x}}$$

- $p(\mathbf{y_i}|\theta) = \left(\begin{array}{c} n \\ x \end{array}\right)\theta^x(1-\theta)^{n-x}$
- Verossimilhança:

$$L(\theta|\mathbf{y}) \propto \theta^x(1-\theta)^{n-x}$$

Peso de indivíduos amostrados numa pop. F_2

- Um modelo possível: $y_i \sim N(\mu,\sigma^2)$, sendo $\theta = (\mu,\sigma^2)$
- Verossimilhança:

$$L(\theta|\mathbf{y}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(y_i-\mu)^2}{2\sigma^2}}$$

- Qual seria um modelo para estudar a variação do peso dos alunos da sala de aula?

Método da Verossimilhança

- Para simplificar, é usual trabalharmos com o log de $L(\theta)$
- Qual a razão?

Atenção Os pontos de máximo e mínimo não se alteram após o uso do logaritmo (função *monótona*)

- Notação: $l(\theta) = \log_e L(\theta) = \log L(\theta)$
- $\frac{d\,l(\theta)}{d\theta}$ é dita função score
- $I(\theta) = -\frac{d^2 l(\theta)}{d\theta^2}$ é dita função de informação de Fisher

Estimador de Máxima Verossimilhança

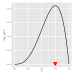
Exercício

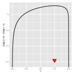
1. Qual a função de verossimilhança do exemplo anterior (binomial)?
2. Qual a função score?
3. Qual é o ponto de máximo de $l(\theta)$, dito $\hat{\theta}$?

Estimador de Máxima Verossimilhança

Exercício

1. $L(\theta) \propto \theta^x (1-\theta)^{(n-x)}$
2. $l(\theta) = x\ \log(\theta) + (n-x)\ \log(1-\theta)$
3. $\hat{\theta} = \frac{x}{n}$





MLE

- $\hat{\theta} = 3/4$ é o MLE de θ

MLE

Distribuição Normal

1. $L(\theta|\mathbf{y}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(y_i-\mu)^2}{2\sigma^2}}$
2. $\hat{\mu} = \frac{\sum_{i=1}^{n} y_i}{n} = \bar{\mathbf{y}}$
3. $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n}$

MLE e Quadrados Mínimos

- Sob normalidade, os *MLE's* também são estimadores de quadrados mínimos

Principais Referências

Gonick, L; Smith, W.
The Cartoon Guide to Statistics
Editora Harper Perennial, 1993

Kalbfleisch, J.G.
Probability and Statistical Inference
Editora Springer-Verlag, 1985 Volume 1

Edwards, A.W.F.
Likelihood (expanded edition)
The John Hopkins University, 1992

Distribuição de Probabilidades Esperança Matemática Verossimilhança Referências

Principais Referências

Sorensen, D.; Gianola, D.
Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics
Editora Springer-Verlag, 2002

Koller, D.; Friedman, N.
Probabilistic Graphical Models: Principles and Techniques
MIT Press, 2009