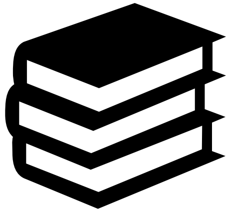
MÉTODOS EXATOS®
CURSOS AVANÇADOS PARA ANÁLISE DE DADOS

Análise Bayesiana de Dados

Análise bayesiana

- Resume-se na construção de uma distribuição de **probabilidade posterior** via o Teorema de Bayes. Esta distribuição resulta da combinação de **informações prévias**, sumarizadas em uma **distribuição denominada priori**, com dados estatísticos descritos por algum modelo probabilístico e resumidos na **função de verossimilhança**.

Distribuições prioris



$$p\left(\theta = \frac{1}{4}\right) = 0.2$$

$$p\left(\theta = \frac{1}{6}\right) = 0.6$$

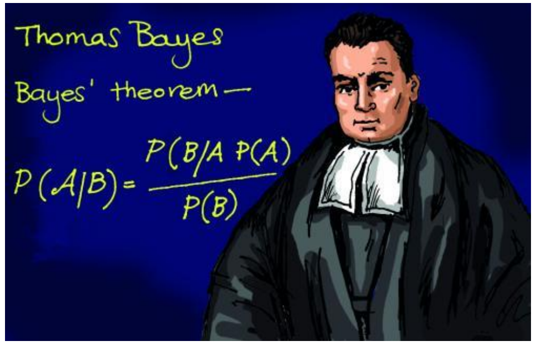
$$p\left(\theta = \frac{3}{4}\right) = 0.2$$

$$\boldsymbol{\theta \sim U(0, 1)}$$

Verossimilhança

$$P(x|\theta) \sim Bin(n, \theta)$$

Distribuições Posteriores



$$P(\theta|X)$$

Análise bayesiana

Exemplo 4.1:

- Um estudo investigou se uma substância abortiva RU 486 poderia ser um contraceptivo para o “dia seguinte”. As 800 participantes do estudo foram divididas em 2 grupos. Enquanto as mulheres de um grupo recebiam o tratamento padrão as outras mulheres recebiam RU 486.
- Das 402 mulheres que receberam RU 486, nenhuma ficou grávida enquanto 4 das 398 mulheres que receberam o tratamento padrão engravidaram.
- Os médicos envolvidos com o estudo queriam saber se o novo tratamento é um contraceptivo mais efetivo que o tratamento padrão.

Solução:

- Qual é a probabilidade de que ao haver uma gravidez esta tenha ocorrido no grupo RU?

$$Bin(n = 4, \theta)$$

$$\theta = \frac{402}{800} \cong 0.50$$

$$x = 0 \ e \ n = 4$$

Distribuições Prioris:

$$P(\theta = 1/4) = 0.2$$

$$P(\theta = 1/2) = 0.6$$

$$P(\theta = 3/4) = 0.2$$

Verossimilhança:

$$p(X|\theta) \sim Bin(n, \theta)$$

$$dbinom(x = 0, size = 4, prob = 1/4)$$

$$dbinom(x = 0, size = 4, prob = 1/2)$$

$$dbinom(x = 0, size = 4, prob = 3/4)$$

Inferência bayesiana

Exemplo 4.1:

- Um estudo investigou se uma substância abortiva RU 486 poderia ser um contraceptivo para o "dia seguinte". As 800 participantes do estudo foram divididas em 2 grupos. Enquanto as mulheres de um grupo recebiam o tratamento padrão as outras mulheres recebiam RU 486.
- Das 402 mulheres que receberam RU 486, nenhuma ficou grávida enquanto 4 das 398 mulheres que receberam o tratamento padrão engravidaram.
- Os médicos envolvidos com o estudo queriam saber se o novo tratamento é um contraceptivo mais efetivo que o tratamento padrão.

Solução:

- Qual é a probabilidade de que ao haver uma gravidez esta tenha ocorrido o grupo RU?

θ	$p(\theta)$	$p(X\|\theta)$	$p(X\|\theta) * p(\theta)$	$p(\theta\|X)$
$1/4$	0.2	0.317	0.0633	0.623
$1/2$	0.6	0.063	0.0375	0.369
$3/4$	0.2	0.004	0.0008	0.008
Soma	1.0		0.1016	1.000

Probabilidades posteriores para famílias conjugadas de distribuições

Definição:

- Seja F uma família de distribuições para a verossimilhança $p(x|\theta)$ e P uma família de distribuição para a priori $p(\theta)$. Dizemos que F e P são **famílias conjugadas** de distribuições se a distribuição posterior $p(\theta|X)$ também for um membro de P.

Exemplo:

Sexualmente maduras

Peixe-galo

- θ: proporção de peixes sexualmente maduras:

  $Beta(\alpha, \beta) \rightarrow \; p(\theta)$ (priori)

- Em uma amostra o objetivo é identificar ou não fêmeas com essas características:

  $Bin(n, \theta) \rightarrow \; p(X|\theta)$ (verossimilhança)

- A distribuição posterior é proporcional ao núcleo de uma nova distribuição beta:

  $\alpha^* = \alpha + x \; e \; \beta^* = \beta + n - x \;\; \rightarrow \; p(\theta|X) \sim Beta(\alpha^*, \beta^*)$

Solução: Se $p(\theta)$ tem distribuição $Beta(\alpha, \beta)$ e $p(x|\theta)$ tem distribuição $Bin(n, \theta)$, então $p(\theta|X)$ tem distribuição $Beta(\alpha^* = \alpha + x, \beta^* = \beta + n - x)$.

Probabilidades posteriores para famílias conjugadas de distribuições

Exemplo 4.1 (continuação):

- Um estudo investigou se uma substância abortiva RU 486 poderia ser um contraceptivo para o "dia seguinte". As 800 participantes do estudo foram divididas em 2 grupos. Enquanto as mulheres de um grupo recebiam o tratamento padrão as outras mulheres recebiam RU 486.
- Das 402 mulheres que receberam RU 486, nenhuma ficou grávida enquanto 4 das 398 mulheres que receberam o tratamento padrão engravidaram.

$\boldsymbol{\theta}$	$\boldsymbol{p(\theta)}$	$\boldsymbol{p(X\|\theta)}$	$\boldsymbol{p(X\|\theta) * p(\theta)}$	$\boldsymbol{p(\theta\|X)}$
$1/4$	0.2	0.317	0.0633	0.623
$1/2$	0.6	0.063	0.0375	0.369
$3/4$	0.2	0.004	0.0008	0.008
Soma	1.0		0.1016	1.000

Solução:

$x = 0 \; e \; n = 4$

$U(0,1) \equiv Beta(1,1)$ (priori)

$Bin(n = 4, \theta)$ (verossimilhança)

$Beta(\alpha^*, \beta^*)$ (posterior)

$Beta(\alpha^* = \alpha + x, \beta^* = \beta + n - x)$

$\alpha^* = \alpha + x \quad = 1 + 0 \quad = 1$

$\beta^* = \beta + n - x = 1 + 4 - 0 = 5$

Aplicação:

- Aferir a opinião do médico sobre a efetividade do RU 486 em relação ao método tradicional.

1. Opinião do médico sobre $\boldsymbol{\theta}$ (antes) ? $r = 0.3 = \alpha/(\alpha + \beta)$ (esperança)

2. Opinião do médico sobre $\boldsymbol{\theta}$ (supondo 1 caso de gravidez p/ grupo RU) ?

$n = 1; \; x = 1 \qquad r^+ = 0.4$

$$\left.\begin{array}{l} Beta(\alpha^* = \alpha + x, \beta^* = \beta + n - x) \\ r^+ = \dfrac{\alpha + x}{[(\alpha + x) + (\beta + n - x)]} = 0.4 \end{array}\right\} \quad \begin{array}{l} \alpha = \dfrac{r(1 - r^+)}{r^+ - r} = 1.8 \\ \beta = \dfrac{(1 - r)(1 - r^+)}{r^+ - r} = 4.2 \end{array}$$

Probabilidades posteriores para famílias conjugadas de distribuições

Atividade:

- **O número médio de defeitos por 30m de fita**, simbolizado por θ, é o parâmetro de interesse que é incerto. Após consulta com o engenheiro de produção responsável pelo controle de qualidade, conclui-se que esta incerteza pode ser convenientemente descrita por uma distribuição **priori gama** com parâmetros $\boldsymbol{\alpha = 2}$ **e** $\boldsymbol{\beta = 10}$. Após terem sido encontrados **4 defeitos** durante a inspeção de **360 m de fita**, qual deverá ser a distribuição de probabilidade para θ?

Solução: Se $p(\theta)$ tem distribuição $Gama(\alpha, \beta)$ e se X é uma variável aleatória com distribuição de poisson com parâmetro $n\theta$, onde n denota o número conhecido de intervalos em que X é observado, então a distribuição posterior $p(\theta|X)$ também será gama, com parâmetros $\alpha^* = \alpha + x, e\ \beta^* = \beta + n$.

$$x = 4\ e\ n = \frac{360}{30} = 12$$

$Gama(2, 10)$ (priori)

$Poi(n\theta)$ (verossimilhança)

$Gama(\alpha^*, \beta^*)$ (posterior)

$$Beta(\alpha^* = \alpha + x, \beta^* = \beta + n)$$

$$\alpha^* = \alpha + x \quad = 2 + 4 \quad = 6$$

$$\beta^* = \beta + n \quad = 10 + 12 \quad = 22$$

Probabilidades posteriores para famílias conjugadas de distribuições

Atividade 4.2 (continuação):

- O número médio de defeitos por 30m de fita, simbolizado por θ, é o parâmetro de interesse que é incerto. Após consulta com o engenheiro de produção responsável pelo controle de qualidade, conclui-se que esta incerteza pode ser convenientemente descrita por uma distribuição priori gama com parâmetros $\alpha = 2 \; e \; \beta = 10$. Após terem sido encontrados 4 defeitos durante a inspeção de 360 m de fita, qual deverá ser a distribuição de probabilidade para θ?
- O engenheiro de produção inicialmente avaliou que o procedimento de produção de fitas estava dentro do padrão que, de acordo com sua experiência, corresponderia a uma média de 0.2 falhas por 30 m de fita ($r = 0.2$). Solicitado a sua atribuição se tivesse observado exatamente 10 falhas em 150 m (5 intervalos de 30 m) de fita inspecionadas (média de 2.0 defeitos por intervalo), ele diz que passaria a considerar razoável a média igual a 0.8 ($r^+ = 0.8$). A partir desses valores e das equações a seguir estime os valores a priori de $\alpha \; e \; \beta$.

$x = 4 \; e \; n = \dfrac{360}{30} = 12$

$Gama(2, 10)$ (priori)

$Poi(n\theta)$ (verossimilhança)

$Gama(\alpha^*, \beta^*)$ (posterior)

$Beta(\alpha^* = \alpha + x, \beta^* = \beta + n)$

$$\alpha^* = \alpha + x \quad = 2 + 4 \quad = 6$$

$$\beta^* = \beta + n \quad = 10 + 12 \quad = 22$$

$x = 10; n = 5; r = 0.2; r^+ = 0.8$

$$\alpha = \frac{r(x - nr^+)}{r^+ - r} = \frac{0.2 * (10 - 5 * 0.8)}{0.8 - 0.2} = \mathbf{2}$$

$$\beta = \frac{x - nr^+}{r^+ - r} = \frac{10 - 5 * 0.8}{0.8 - 0.2} = \mathbf{10}$$

Probabilidades posteriores para famílias conjugadas de distribuições

Exemplo 4.3

- 100 leitões de determinada idade foram pesados, passaram a receber uma dieta especial nesta data e, foram pesados novamente 20 dias depois. O interesse está em estimar a média (μ) do ganho de peso em quilos resultante dessa dieta.
- Um veterinário avalia que qualquer valor μ entre 0 kg e 100 kg é igualmente provável e que valores fora desse intervalo são impossíveis. Portanto, para ele a priori p(μ) é não-informativa e tem densidade uniforme em [0, 100].

Solução: Se $p(\mu)$ tem distribuição uniforme e $p(\bar{X}|\mu)$ é normal $N\left(\mu, \frac{S}{\sqrt{n}}\right)$, então $p(\mu|\bar{X})$ tem distribuição normal $N\left(\bar{X}, \frac{S}{\sqrt{n}}\right)$.

Nota: quando $n \leq 30$ substitui-se S por S_c.

- A amostra de 100 leitões resultou num ganho de peso médio de 30 kg com desvio padrão de 10 kg.

Solução: Se $p(\mu)$ tem distribuição $N(\mu_0, \sigma_0)$ e $p(\bar{X}|\mu)$ é $N\left(\mu, \frac{S}{\sqrt{n}}\right)$. Então $p(\mu|\bar{X})$ tem distribuição $N(\mu_1, \sigma_1)$.

Nota: quando $n \leq 30$ substitui-se S por S_c.

Probabilidades posteriores para famílias conjugadas de distribuições

Exemplo 4.3 (continuação)

- Um outro veterinário, porém especialista em suinocultura, expõe suas incertezas sobre o aumento médio resultante desta dieta. Com base em sua experiência em estudos com outras dietas similares ele se decide por 45 kg.
- Após essa declaração ele é solicitado a fornecer a probabilidade que atribui ao seguinte evento : "μ é maior que 60 kg". Ele atribui a probabilidade de 0.1 a essa afirmação.

```
> qnorm(0.9, 0 , 1)
[1] 1.281552
```

$$\frac{60-45}{\sigma_0} = 1.28 \quad \rightarrow \sigma_0 \cong 11.7$$

$$Precisão\ a\ priori\colon\ c_0 = \frac{1}{\sigma_0^2}$$

$$Precisão\ amostral\colon c = \frac{n}{S^2}$$

$$Precisão\ posterior\colon\ c_1 = c_0 + c$$

- Reescrevendo a priori e a verossimilhança em termos de c_0, c e c_1, pode-se mostrar que seu produto é proporcional a uma nova distribuição normal com parâmetros:

$$\mu_1 = \frac{c_0}{c_1}\mu_0 + \frac{c}{c_1}\bar{X}$$

$$\sigma_1 = \frac{1}{\sqrt{c_1}}$$

```
media0 <- 45
sigma0 <- 11.7
S <- 10
Xbarra <- 30
n <- 100
c0 <- 1/sigma0^2
c <- n/S^2
c1 <- c0 + c
media1 <- (c0/c1)*media0 + (c/c1)*Xbarra
sigma1 <- 1/sqrt(c1)
```

$\cong 30.1$ (media1)

$\cong 0.99$ (sigma1)

Famílias Conjugadas para Amostras da Distribuição Normal

- Considerando que a distribuição normal é bidimensional.
- Para suposições de que sigma não é conhecido ou não é bem representado pelo desvio padrão da amostra.
- Quando a distribuição marginal $p(\mu|x)$ é de interesse.

Etapas:

1. Calcular a posterior conjugada bidimensional: $\boldsymbol{p(\mu, \sigma|x)}$
2. Obtém-se a distribuição marginal: $\boldsymbol{p(\mu|x)}$

Priori Informativa

Precisão: $\tau = \frac{1}{\sigma^2} \quad \rightarrow N^*(\mu, \tau)$

- A densidade de uma variável aleatória X com distribuição normal com média μ e precisão τ, $N^*(\mu,\tau)$, será:

$$p(x|\mu,\tau) = \sqrt{\frac{\tau}{2\pi}} \exp\left[-\frac{\tau(x-\mu)^2}{2}\right]$$

- A distribuição priori conjunta será definida como sendo uma distribuição normal-gama para o vetor aleatório (μ,τ). Isto é:

$$p(\mu,\tau) = p(\mu|\tau) * p(\tau) \qquad p(\mu|\tau) \sim N^*(\mu_0, \lambda_0\tau), \lambda > 0 \qquad p(\tau) \sim Gama(\alpha_0, \beta_0)$$

Solução: Se (μ,τ) tem distribuição priori normal-gama com parâmetros $(\mu_0, \lambda_0, \alpha_0, \beta_0)$ e $p(\bar{X}|\mu,\tau)$ têm distribuição $N^*(\mu, n\tau)$, então $p(\mu,\tau|\bar{X},S^2)$ terá uma distribuição normal-gama:

$$\boldsymbol{p(\mu|\tau,\bar{x},S^2) \sim N^*(\mu_1, \lambda_1\tau) \qquad p(\tau|\bar{x},S^2) \sim Gama(\alpha_1, \beta_1)}$$

com parâmetros:

$$\boldsymbol{\mu_1 = \frac{\lambda_0\mu_0 + n\bar{X}}{\lambda_0 + n} \qquad \alpha_1 = \alpha_0 + \frac{n}{2} \qquad \lambda_1 = \lambda_0 + n \qquad \beta_1 = \beta_0 + \frac{(n-1)S^2}{2} + \frac{n\lambda_0(\bar{X}-\mu_0)^2}{2(\lambda_0+n)}}$$

$$\boldsymbol{Nota}: \quad \boldsymbol{c_0 = \lambda_0\tau \qquad c_1 = \lambda_1\tau \qquad c = n\tau}$$

Priori Informativa

Atividade 4.3 (reavaliando)

- Voltando ao exemplo dos leitões, imagine que deseja-se estimar a média de ganho de peso, mas a precisão também é desconhecida. Assim, uma possível solução é utilizar a família conjugada normal-gama.
- Para isso determinou-se os seguintes parâmetros: $\mu_0 = 45, \lambda_0 = 2, \alpha_0 = 2 \; e \; \beta_0 = 100.$
- Considere agora que uma amostra de $n = 20$ leitões resultou num ganho de peso médio de 30 kg com desvio padrão de 10 kg.

$$\boldsymbol{\mu_1 = \frac{\lambda_0\mu_0 + n\bar{X}}{\lambda_0 + n}} \qquad \boldsymbol{\alpha_1 = \alpha_0 + \frac{n}{2}} \qquad \boldsymbol{\lambda_1 = \lambda_0 + n} \qquad \boldsymbol{\beta_1 = \beta_0 + \frac{(n-1)S^2}{2} + \frac{n\lambda_0(\bar{X} - \mu_0)^2}{2(\lambda_0 + n)}}$$

- Aplicando as expressões dadas, obtenha os parâmetros para a distribuição posterior da família normal-gama.

Distribuição Posterior Marginal

$\boldsymbol{p(\mu|x)}$

- Em que μ é uma função conhecida da variável T, que tem distribuição de Student padronizada com $g = 2\alpha_1$ graus de liberdade:

$$\mu = \sqrt{\frac{\beta_1}{\lambda_1\alpha_1}} * T + \mu_1$$

- Então:

$$E(\mu|x) = \mu_1 se\ \alpha_1 > \frac{1}{2} \qquad V(\mu|x) = \frac{\beta_1}{\lambda_1(\alpha_1 - 1)}\ se\ \alpha_1 > 1$$

- Portanto, quando lidamos com uma normal-gama obtém-se que a solução analítica marginal para μ é a distribuição de probabilidade de Student:

$$p(\mu|x) \sim St(2\alpha_1, \mu_1, \sqrt{\frac{\beta_1}{\lambda_1\alpha_1}})$$

Priori não-informativa

- Uma priori não-informativa pode ser utilizada nas situações em que não há conhecimentos que permitam especificar uma distribuição priori informativa conjugada normal-gama.

$$p(\mu,\tau) = \frac{1}{\tau}\ para\ -\infty < \mu < +\infty\ e\ \tau > 0$$

Solução: Se (μ,τ) tem distribuição priori não-informativa $p(\mu,\tau) = \frac{1}{\tau}$ e $p(\bar{X}|\mu,\tau)$ têm distribuição $N^*(\mu, n\tau)$, então, para $n \geq 2, p(\mu,\tau|\bar{X},S^2)$ terá distribuição normal-gama:

$$\boldsymbol{p(\mu|\tau,\bar{x},S^2) \sim N^*(\mu_1,\lambda_1\tau)} \qquad \boldsymbol{p(\tau|\bar{x},S^2) \sim Gama(\alpha_1,\beta_1)}$$

com parâmetros:

$$\boldsymbol{\mu_1 = \bar{X}} \qquad \boldsymbol{\alpha_1 = \frac{n-1}{2}} \qquad \boldsymbol{\lambda_1 = n} \qquad \boldsymbol{\beta_1 = \frac{(n-1)S^2}{2}}$$

$$n = 20;\ \mu_1 = 30;\ \lambda_1 = 20;\ \alpha_1 = 9.5\ e\ \beta_1 = 950$$

	μ		σ	
$Percentil$	I	NI	I	NI
2.5%	26.8	25.3	7.9	7.6
50.0%	31.3	30.0	10.3	10.1
97.5%	35.8	34.6	14.2	14.6

Hiper-parâmetros da Priori Normal-Gama

Etapas para a construção dos hiper-parâmetros

- Determinar um valor razoável para μ_0 a partir dos conhecimentos existentes: $\boldsymbol{\mu_0 = 45\ kg}$
- Especifique valores mínimo e máximo para a variável aleatória X.

$$V = \left(\frac{y_2 - y_1}{6}\right)^2 \qquad (ex: X = \{15; 75\} \rightarrow V = 100)$$

- Divida em 2 componentes que expressam, respectivamente, a variabilidade de X e a incerteza.

$$V_x = (1-p)V \quad (ex: p = 0.5 \rightarrow V_x = 50) \qquad V_\mu = pV \quad (ex: p = 0.5 \rightarrow V_\mu = 50)$$

- Fixe α_0 em algum valor pequeno para caracterizar uma priori "mente-aberta": $\boldsymbol{\alpha_0 = 2}$
- Como $\mathrm{E}(\tau) = \frac{\alpha_0}{\beta_0} = 1/V_x$, segue que $\beta_0 = \alpha_0 * V_x$: $\boldsymbol{\beta_0 = 2 * 50 = 100}$
- Sabendo que a variância da distribuição marginal da priori para μ é igual a

$$V(\mu) = \frac{\beta_0}{\lambda_0(\alpha_0 - 1)} = V_\mu \qquad têm - se\ que \qquad \lambda_0 = \frac{\beta_0}{(\alpha_0 - 1) * V_\mu} = \frac{100}{(2-1) * 50}$$

$\boldsymbol{\lambda_0 = 2}$