Modelagem paramétrica Bayesiana

Prof. Vinícius D. Mayrink
EST088 - Inferência Bayesiana

Sala: 4073 Email: vdm@est.ufmg.br

1^o semestre de 2025

2.1 - População e amostra aleatória

O progresso da ciência é frequentemente atribuído à experimentação.

O pesquisador realiza um experimento e obtém **dados**.

As conclusões, em geral, vão além dos materiais e operações envolvidos no experimento. Podemos generalizar, partindo de um particular experimento para a classe de todos os experimentos similares.

Este tipo de extensão para o caso mais geral é chamado de inferência.

O papel da Estatística é fornecer técnicas para realizar a inferência e para medir o grau de incerteza em tal procedimento.

Incerteza é medida em termos de probabilidade.

Exemplo: Suponha que temos um recipiente contendo 10 milhões de sementes de flores. Sabemos que cada semente irá produzir flor branca ou vermelha.

Qual seria a porcentagem de sementes que geram flores brancas neste recipiente?

A única maneira de responder esta questão corretamente é plantar cada semente e observar o número de flores brancas obtidas.

Entretanto, esta tarefa não é razoável de ser executada, pois exige muito esforço, tempo e não permitiria a venda das sementes.

Solução: podemos plantar algumas sementes e, baseado nas cores observadas, faremos uma afirmação sobre a porcentagem desejada.

População: é um conjunto contendo todos os elementos do problema em discussão. Deseja-se obter informação sobre estes elementos.

Alguns exemplos de conjuntos população:

- Peças produzidas por uma fábrida em determinado ano;
- Preços do pão em certo dia em Belo Horizonte;
- As sequências hipotéticas de caras e coroas obtidas com o lançamento de uma moeda infinitas vezes;
- Produção de leite por animal em uma fazenda;
- Um conjunto hipotético de infinitas medidas de velocidades de veículos em determinado trecho de uma rodovia.

A população alvo precisa ser bem definida. Ela pode ser real ou hipotética.

Podemos fazer afirmações probabilísticas sobre uma população se a amostra é selecionada seguindo alguns cuidados.

Notação (esclarecimento): Na 1^a parte da disciplina adotou-se que: $p_Y(\bullet)$ representa uma f.m.p. e $f_Y(\bullet)$ representa uma f.d.p. Isso foi importante para ressaltar o tipo de variável (discreta ou contínua) sendo avaliada. A partir de agora iremos adotar $f_Y(\bullet)$ como uma notação mais geral indicando tanto f.m.p. quanto f.d.p.

Notação (esclarecimento): Seja $\theta = (\theta_1, \theta_2, \cdots, \theta_k)^\top$ um vetor com k parâmetros definidos na formulação da f.d.p. ou f.m.p. de uma variável aleatória Y. Iremos escrever $f_{Y|\theta}(y)$ para expressar a presença de θ na distribuição de Y.

Dizemos que a distribuição de Y é **idexada** por θ.

Alguns exemplos para esclarecer a notação genérica θ são:

- Se $Y \sim$ Poisson(λ), então $k = 1$ e $\theta = \{\lambda\}$;
- Se $Y \sim$ Gama(α, β), então $k = 2$ e $\theta = \{\alpha, \beta\}$;
- Se $Y \sim$ N(μ, σ^2), então $k = 2$ e $\theta = \{\mu, \sigma^2\}$.

Sejam $Y_1, Y_2, \cdots, Y_n$ variáveis aleatórias cuja distribuição conjunta pode ser fatorada, por independência condicional dado θ, como segue:

$$f_{Y_1,\cdots,Y_n|\theta}(y_1,\cdots,y_n) = f_{Y_1|\theta}(y_1)\, f_{Y_2|\theta}(y_2) \cdots f_{Y_n|\theta}(y_n) = \prod_{i=1}^{n} f_{Y_i|\theta}(y_i),$$

sendo $f_{Y_i|\theta}(y_i)$ a f.d.p. ou f.m.p. de Y_i indexada por θ.

Dizemos aqui que $Y_1, Y_2, \cdots, Y_n$ é uma **amostra aleatória** de tamanho n obtida de uma população com f.d.p. ou f.m.p. conjunta $f_{Y_1,\cdots,Y_n|\theta}(y_1,\cdots,y_n)$.

Este é o caso mais simples, em que o mesmo parâmetro θ indexa a distribuição de probabilidade de todos os Y_i's.

Em um caso mais geral, assuma $\theta = \{\theta_1, \cdots, \theta_n\}$ e $\theta_i = (\theta_{i1}, \theta_{i2}, \cdots, \theta_{ik})^\top$. Podemos escrever:

$$f_{Y_1,\cdots,Y_n|\theta}(y_1,\cdots,y_n) = f_{Y_1|\theta_1}(y_1)\, f_{Y_2|\theta_2}(y_2) \cdots f_{Y_n|\theta_n}(y_n) = \prod_{i=1}^{n} f_{Y_i|\theta_i}(y_i),$$

sendo $f_{Y_i|\theta_i}(y_i)$ a f.d.p. ou f.m.p. de Y_i indexada por um parâmetro específico θ_i.

Notação (esclarecimento): Para simplificar a escrita envolvendo amostras aleatórias, adote que

- $\boldsymbol{Y} = \{Y_1, Y_2, \cdots, Y_n\}$ é um conjunto de variáveis aleatórias.
- $\boldsymbol{y} = \{y_1, y_2, \cdots, y_n\}$ é um conjunto com observações de uma realização de $\boldsymbol{Y}$.

Exemplo: Reconsidere as 10 milhões de sementes em um recipiente e a produção de flores brancas e vermelhas.

- População: Sementes dentro do recipiente.
- Elemento populacional: uma semente.
- Resposta: flor branca ou vermelha.

Não temos um valor numérico associado a cada elemento, mas podemos definir este tipo de resposta. Considere a variável aleatória Y_i tal que:

- $Y_i = 1$, se a i-ésima semente gerar flor branca.
- $Y_i = 0$, se a i-ésima semente gerar flor vermelha.

Seja $Y_1, Y_2, \cdots, Y_n$ um conjunto de n variáveis aleatórias binárias que formam nossa amostra aleatória.

Naturalmente, iremos supor que $Y_j \sim \text{Bernoulli}(p)$, ou seja, adota-se que a probabilidade de uma semente qualquer originar flor branca é igual a p. Podemos escrever a seguinte f.m.p. (condicionada em p):

$$f_{Y_j|p}(y_j) \;=\; p^{y_j}\,(1-p)^{1-y_j} \quad \text{com} \quad y_j = 0 \text{ ou } 1.$$

Para $n = 2$ sementes, temos a distribuição conjunta:

$$f_{Y_1,Y_2|p}(y_1, y_2) \;=\; f_{Y_1|p}(y_1)\;\; f_{Y_2|p}(y_2)$$
$$=\; p^{y_1}\,(1-p)^{1-y_1}\;\; p^{y_2}\,(1-p)^{1-y_2}$$
$$=\; p^{y_1+y_2}\,(1-p)^{2-y_1-y_2}$$

para $\;y_1 \in \{0,1\}\;$ e $\;y_2 \in \{0,1\}$.

Esta distribuição conjunta (bivariada) não é a mesma obtida ao trabalhar com uma variável aleatória W que conta o n^o de sucessos na realização de dois ensaios Bernoulli(θ).

A f.m.p. de W seria:

$$f_W(w|p) = \binom{2}{w} p^w (1-p)^{2-w} \text{ para } w = 0, 1 \text{ ou } 2.$$

Note que $W = Y_1 + Y_2 \sim \text{Binomial}(2, p)$.

A conjunta no slide anterior $f_{Y_1,Y_2|p}(y_1, y_2)$ representa a distribuição amostral levando em conta a ordem de amostragem.

A f.m.p. $f_{W|p}(w)$ ignora este aspecto.

2.2 - Verossimilhança, distribuição *a priori* e *a posteriori*

Nosso ponto de partida será um estudo empírico (pode ser experimental ou observacional) que irá fornecer certo conjunto de dados (amostra) que denotaremos por $\boldsymbol{y}$.

No caso mais simples, teremos $\boldsymbol{y} = \{y_1, y_2, \cdots, y_n\}$.

As observações em $\boldsymbol{y}$ podem ser vistas como uma possibilidade de resultado a ser assumido pelo conjunto de variáveis aleatórias $\boldsymbol{Y} = \{Y_1, Y_2, \cdots, Y_n\}$.

Objetivo: usar $\boldsymbol{y}$ para concluir sobre a distribuição desconhecida $f_{\boldsymbol{Y}}(\bullet)$ de $\boldsymbol{Y}$.

As conclusões sobre $f_{\boldsymbol{Y}}(\bullet)$ estão sujeitas a incertezas diante da aleatoriedade governando $\boldsymbol{Y}$. Devemos certificar que:

- O nível de incerteza é o menor possível, considerando a aleatoriedade de $\boldsymbol{Y}$;
- É possível avaliar o nível de incerteza nas conclusões.

A natureza física do fenômeno que gera $\boldsymbol{y}$, o esquema de amostragem, e outras informações irão colocar limites no conjunto de possíveis escolhas para $f_{\boldsymbol{Y}}(\bullet)$.

Este conjunto (denotado por $\mathcal{F}$) é chamado de **modelo estatístico**.

É intuitivo pensar que as inferências serão mais precisas se formos capazes de selecionar o menor conjunto $\mathcal{F}$ possível, sob o requerimento de que $f_{\boldsymbol{Y}}(\bullet) \in \mathcal{F}$.

Em algumas situações, $\boldsymbol{Y}$ é encarado como uma amostra aleatória com componentes independentes e identicamente distribuídos.

A princípio, $\mathcal{F}$ pode ser qualquer conjunto de f.m.p.'s ou f.d.p.'s, mas existe uma categoria de tais conjuntos que possui importante papel, tanto do ponto de vista teórico quanto aplicado.

Este caso ocorre quando todos os elementos de $\mathcal{F}$ são funções com a mesma formulação matemática, identificadas apenas pelas diferentes especificações de θ, que variam em $\Theta \in \mathbb{R}^k$

$$\mathcal{F} = \left\{ f_{\boldsymbol{Y}|\theta}(\bullet) \ : \ \theta \in \Theta \subseteq \mathbb{R}^k \right\}.$$

Notação: $f_{\boldsymbol{Y}|\theta}(\bullet)$ é uma f.m.p. ou f.d.p. cujo suporte é um subconjunto de $\mathbb{R}^m$, sendo k e m inteiros positivos.

- θ é chamado de parâmetro (escalar, vetor ou matriz);
- Θ é chamado de espaço paramétrico;
- $\mathcal{F}$ é chamado de classe paramétrica ou modelo paramétrico.

Os elementos de $\mathcal{F}$ estão associados aos elementos de Θ.

Em particular, existe um valor $\theta_* \in \Theta$, associado a $f_{\boldsymbol{Y}}(\bullet)$, que é chamado de "valor real" do parâmetro.

Nossas inferências serão direcionadas para determinar θ_*.

Espaço amostral: conjunto $\mathcal{Y}$ de todas os possíveis resultados amostrais $\boldsymbol{y}$ compatíveis com o modelo paramétrico em consideração.

Notação (esclarecimento): na 1^a parte do curso havíamos adotado o símbolo **U** para indicar conjunto Universo ou espaço amostral de eventos. No atual contexto de amostras aleatórias, iremos adaptar essa notação para $\mathcal{Y}$.

Seja $\mathcal{Y}_\theta$ o suporte da f.m.p. ou f.d.p. $f_{\boldsymbol{Y}|\theta}(\bullet)$, o espaço amostral é dado por:

$$\mathcal{Y} = \cup_{\theta \in \Theta} \mathcal{Y}_\theta.$$

Frequentemente, $\mathcal{Y}$ é o mesmo para todas as possíveis escolhas de θ.
Nesta situação, $\mathcal{Y}_\theta$ coincide com $\mathcal{Y}$.

Exemplo: Assuma que $Y \sim \text{Binomial}(n, \theta)$.

Se $\theta \in (0, 1)$, temos:

- $\mathcal{Y}_\theta$ será o mesmo para todo θ.
- $\mathcal{Y}_\theta$ coincidirá com o espaço amostral $\mathcal{Y} = \{0, 1, 2, \cdots, n\}$

Se $\theta \in [0, 1]$, temos:

- $\mathcal{Y}_{\theta=0} = \{0\}, \quad \mathcal{Y}_{\theta=1} = \{n\} \quad \text{e} \quad \mathcal{Y}_{\theta\in(0,1)} = \{0, 1, 2, \cdots, n\}$.
- Logo $\mathcal{Y} = \cup_{\theta\in[0,1]} \mathcal{Y}_\theta = \{0, 1, \cdots, n\}$.

Exemplo: Se dois valores são amostrados independentemente da $N(\theta, 1)$, então $\boldsymbol{y} = (y_1, y_2)^\top$, sendo $y_i \in \mathbb{R}$,

$\mathcal{Y} = \mathbb{R}^2, \quad \boldsymbol{Y} \sim N[\ (\theta, \theta)^\top,\ I_2\ ],$

$f_{\boldsymbol{Y}|\theta}(\boldsymbol{y}) = (2\pi)^{-1/2} \exp\{-\frac{1}{2}(y_1 - \theta)^2\}\ (2\pi)^{-1/2} \exp\{-\frac{1}{2}(y_2 - \theta)^2\},$

- se não houver qualquer restrição para θ, temos $\Theta = \mathbb{R}$.
- se existir restrição (ex.: saber que $\theta > 0$) então $\Theta = \mathbb{R}^+$.

Considere um dado modelo estatístico do tipo

$$\mathcal{F} = \{f_{\boldsymbol{Y}|\theta}(\bullet) : \theta \in \Theta \subseteq \mathbb{R}^k\}$$

Quando uma amostra $\boldsymbol{y}$ é observada, o valor da f.m.p. ou f.d.p. $f_{\boldsymbol{Y}|\theta}(\bullet)$ dependerá apenas de θ.

Esta função fornece a f.m.p. ou f.d.p. para observar aquilo que de fato observamos, ou seja, $\boldsymbol{y}$.

Se precisarmos estabelecer um ranking envolvendo dois elementos de Θ (considere θ' e θ''), então uma quantidade relevante e útil para esta tarefa será a razão $f_{\boldsymbol{Y}|\theta'}(\boldsymbol{y})/f_{\boldsymbol{Y}|\theta''}(\boldsymbol{y})$, desde que o denominador não seja zero.

Como esta razão não muda caso ambos os termos sejam multiplicados por uma constante positiva "c", independente de θ, então para comparar os elementos de Θ a quantidade relevante será proporcional a $f_{\boldsymbol{Y}|\theta}(\boldsymbol{y})$.

Definição: Para o modelo estatístico $\mathcal{F} = \{f_{\boldsymbol{Y}|\theta}(\bullet) : \theta \in \Theta \subseteq \mathbb{R}^k\}$ a partir do qual uma amostra $\boldsymbol{y} \in \mathcal{Y}$ foi observada, usamos o termo *função de verossimilhança*, ou simplesmente *verossimilhança*, para a função escrita por $f_{\boldsymbol{Y}|\theta}(\boldsymbol{y})$.

Esta função depende de $\boldsymbol{y}$. Para uma amostra $\boldsymbol{y}'$ teremos um valor real positivo de $f_{\boldsymbol{Y}|\theta}(\boldsymbol{y}')$ diferente de $f_{\boldsymbol{Y}|\theta}(\boldsymbol{y}'')$.

Existem duas formas distintas de enxergar a verossimilhança:

- Se olharmos $f_{\boldsymbol{Y}|\theta}(\boldsymbol{y})$ como função de $\boldsymbol{y}$, teremos a distribuição conjunta de $Y_1, \cdots, Y_n$.
- Se olharmos $f_{\boldsymbol{Y}|\theta}(\boldsymbol{y})$ como função de θ, não teremos uma distribuição de probabilidade.

Exemplo: Admita uma amostra $\boldsymbol{Y} = \{Y_1, Y_2, \cdots, Y_n\}$ de n indivíduos que foram entrevistados sobre o hábito de fumar. Adote $Y_i = 1$, se a i-ésima pessoa fuma (0 caso contrário).

Condicional na proporção θ de fumantes na população, a amostra $\boldsymbol{Y}$ é i.i.d. com distribuição Bernoulli(θ).

$f_{\boldsymbol{Y}|\theta}(y_1, \cdots, y_n) = \prod_{i=1}^{n} f_{Y_i|\theta}(y_i)$.

$= \prod_{i=1}^{n} \theta^{y_i} (1-\theta)^{1-y_i} = \theta^{\sum_{i=1}^{n} y_i} (1-\theta)^{n-\sum_{i=1}^{n} y_i}$.

Espaço paramétrico: $\Theta = [0, 1]$.

Espaço amostral: $\mathcal{Y} = \{(y_1, \cdots, y_n) : y_i \in \{0, 1\} \text{ para } i = 1, \cdots, n\}$.

Exemplo (continuação):

Se $n = 2$, teremos: $\mathcal{Y} = \{(0,0),(0,1),(1,0),(1,1)\}$.

Visão Bayesiana: considere $n = 2$ e fixe θ.

Verossimilhança com $\theta = 1/2$:

$$\theta^{\sum_{i=1}^{n} y_i} (1-\theta)^{n-\sum_{i=1}^{n} y_i} = (1/2)^{\sum_{i=1}^{2} y_i} (1/2)^{2-\sum_{i=1}^{2} y_i}.$$

Análise dos possíveis resultados

Amostra	$f_{\boldsymbol{Y}\mid\theta}(y_1, y_2)$
(0,0)	$(1/2)^0 (1/2)^2 = 1/4$
(1,0)	$(1/2)^1 (1/2)^1 = 1/4$
(0,1)	$(1/2)^1 (1/2)^1 = 1/4$
(1,1)	$(1/2)^2 (1/2)^0 = 1/4$

Exemplo (continuação):

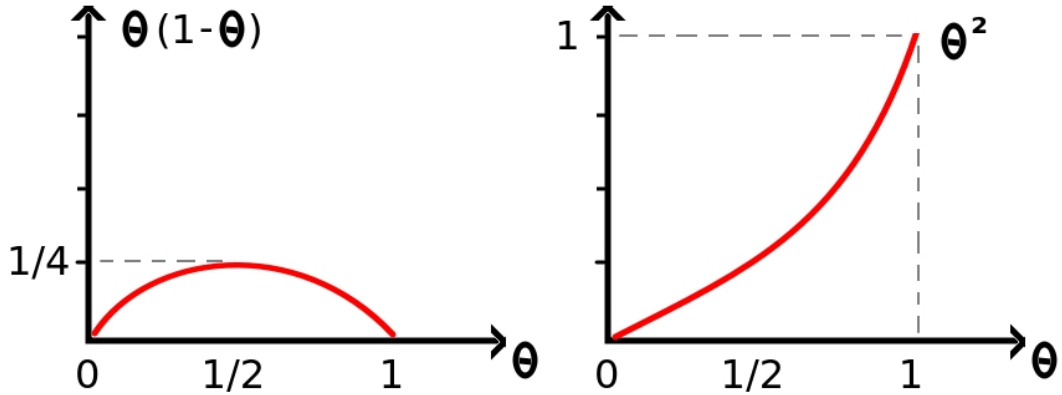
Visão Frequentista: considere $n = 2$ e fixe $\boldsymbol{y}$.

Verossimilhança com $\boldsymbol{y} = (1, 0)$:

$$\theta^{\sum_{i=1}^{n} y_i} (1-\theta)^{n-\sum_{i=1}^{n} y_i} = (\theta)^1 (1-\theta)^{2-1} = \theta(1-\theta).$$

Verossimilhança com $\boldsymbol{y} = (1, 1)$:

$$\theta^{\sum_{i=1}^{n} y_i} (1-\theta)^{n-\sum_{i=1}^{n} y_i} = (\theta)^2 (1-\theta)^{2-2} = \theta^2.$$



$f_{\boldsymbol{Y}|\theta}(\boldsymbol{y})$ é uma quantidade não negativa, e na maioria dos casos é positiva para todo Θ. Sendo assim, podemos definir a função de *log-verossimilhança* como:

$$\ell_{\boldsymbol{Y}|\theta}(\boldsymbol{y}) \;=\; \ln[f_{\boldsymbol{Y}|\theta}(\boldsymbol{y})],$$

com a convenção de que $\ell_{\boldsymbol{Y}|\theta}(\boldsymbol{y}) = -\infty$ se $f_{\boldsymbol{Y}|\theta}(\boldsymbol{y}) = 0$.

A log-verossimilhança é muito útil na visão clássica da inferência Estatística. A análise frequentista está bastante focada em maximizar a função de verossimilhança. Isto é um problema de otimização que requer derivação. Derivar a versão log é mais fácil.

A log-verossimilhança será menos usada sob a visão Bayesiana. Ela irá aparecer em situações para garantir estabilidade numérica em algorítmos computacionais a serem estudados mais adiante.

Exemplo: Considere uma amostra aleatória $\boldsymbol{y} = \{y_1, y_2, \cdots, y_n\}$ proveniente da distribuição $N(\mu, \sigma^2)$, sendo $\theta = (\mu, \sigma^2)$ variando no espaço $\mathbb{R} \times \mathbb{R}^+$.

Perante a independência condicional das variáveis aleatórias Y_i's dado θ, temos:

$$f_{\boldsymbol{Y}|\theta}(\boldsymbol{y}) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\{-\tfrac{1}{2\sigma^2}(y_i - \mu)^2\}$$

$$= (2\pi\sigma^2)^{-n/2} \exp\{-\tfrac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2\}$$

$$= (2\pi\sigma^2)^{-n/2} \exp\{-\tfrac{1}{2\sigma^2} [\sum_{i=1}^{n} y_i^2 - 2\mu \sum_{i=1}^{n} y_i + n\mu^2]\}.$$

Log-verossimilhança:

$$\ell_{\boldsymbol{Y}|\theta}(\boldsymbol{y}) = -\tfrac{n}{2} \ln(2\pi\sigma^2) - \tfrac{1}{2\sigma^2} [\sum_{i=1}^{n} y_i^2 - 2\mu \sum_{i=1}^{n} y_i + n\mu^2].$$

Exemplo: Considere uma amostra aleatória $\boldsymbol{y} = \{y_1, y_2, \cdots, y_n\}$ proveniente da distribuição Poisson(θ). Neste caso, Y_i é uma variável aleatória tipicamente usada para modelar contagens e θ é a taxa de ocorrência do evento de interesse na unidade de tempo ou espaço sob avaliação.

θ varia no espaço $\Theta = \mathbb{R}^+$.

Perante a independência condicional das Y_i's dado θ, temos:

$$f_{\boldsymbol{Y}|\theta}(\boldsymbol{y}) \;=\; \prod_{i=1}^{n} \frac{\theta^{y_i}}{y_i!} \exp\{-\theta\} \;=\; \frac{\theta^{\sum_{i=1}^{n} y_i}}{\prod_{i=1}^{n} y_i!} \exp\{-n\theta\}.$$

Log-verossimilhança:

$$\ell_{\boldsymbol{Y}|\theta}(\boldsymbol{y}) \;=\; \sum_{i=1}^{n} y_i \;\; \ln(\theta) \;-\; \sum_{i=1}^{n} \ln(y_i!) \;\; - \;\; n\theta.$$

A função de verossimilhança conecta a informação pré-experimental (expressa pela escolha do modelo) com a informação experimental contida em $\boldsymbol{y}$.

Portanto, de certa forma, a verossimilhança contém todas as evidências experimentais relativas ao problema de inferência em questão. Esta afirmação é feita sem levar em conta qualquer informação sobre θ que, por qualquer razão, não foi acomodada no modelo, tais como opiniões pessoais ou resultados de estudos similares.

Princípio da verossimilhança (versão fraca): Para um modelo estatístico $\mathcal{F} = \{f_{\boldsymbol{Y}|\theta}(\bullet) \ : \ \theta \in \Theta\}$, dois pontos $\boldsymbol{y}'$ e $\boldsymbol{y}'' \in \mathcal{Y}$ tal que $f_{\boldsymbol{Y}|\theta}(\boldsymbol{y}') \propto f_{\boldsymbol{Y}|\theta}(\boldsymbol{y}'')$ devem levar às mesmas conclusões inferenciais.

Esta é a versão mais fraca do princípio da verossimilhança. A seguir apresentamos uma versão mais forte que diz que as conclusões coincidem mesmo quando os dois pontos observados referem-se a modelos distintos e espaços amostrais distintos.

Princípio da verossimilhança (versão forte): Dada uma observação $\boldsymbol{y}$ do modelo estatístico $\mathcal{F}_{\boldsymbol{Y}} = \{f_{\boldsymbol{Y}|\theta}(\bullet) \ : \ \theta \in \Theta\}$ e uma observação $\boldsymbol{z}$ do modelo $\mathcal{F}_{\boldsymbol{Z}} = \{f_{\boldsymbol{Z}|\theta}(\bullet) \ : \ \theta \in \Theta\}$ tais que $f_{\boldsymbol{Y}|\theta}(\boldsymbol{y}) \propto f_{\boldsymbol{Z}|\theta}(\boldsymbol{z})$, então elas devem levar às mesmas conclusões inferenciais.

Note que a definição acima assume nos dois modelos e o mesmo espaço paramétrico Θ.

Exemplo: Admita dois experimentos em que observações binárias (0 ou 1) são obtidas em sequência. Cada observação tem probabilidade de sucesso (valor 1) igual a θ, independente dos resultados das demais observações.

Experimento 1:

- o n^o total de observações n é fixado antes da execução.
- o resultado do experimento é dado pelo n^o de sucessos.

Experimento 2:

- o n^o de sucessos requeridos é escolhido antes da execução.
- o resultado é dado pelo n^o de falhas observadas antes de parar a sequência de observações (regra de parada: quando atingir o n^o de sucessos requerido).

As distribuições de probabilidade associadas a estas configurações são:

- Experimento 1: Binomial com f.m.p. $f_{W|\theta}(w) = \binom{n}{w} \theta^w (1-\theta)^z$.

  Para $z = n - w$ e $w \in \{0, 1, 2, \cdots, n\}$ sendo o n^o de sucessos.

- Experimento 2: Binomial Negativa com f.m.p.

$$f_{Z|\theta}(z) = \binom{w+z-1}{z} \theta^w (1-\theta)^z.$$

  Para $z \in \{0, 1, 2, \cdots\}$ sendo o n^o de falhas e assuma w escolhido antes do experimento.

No slide anterior, para qualquer um dos experimentos temos:

$w = n^o$ de sucessos e $z = n^o$ de falhas.

Através destas funções de probabilidade temos que as verossimilhanças são:

- $f_{W|\theta}(w) = c_1\, \theta^w (1-\theta)^z$, com $c_1 = \binom{n}{w}$.
- $f_{Z|\theta}(z) = c_2\, \theta^w (1-\theta)^z$, com $c_2 = \binom{w+z-1}{z}$.

Note que as contantes c_1 e c_2 não dependem de θ.

Se as quantidades w e z coincidirem nos dois experimentos, teremos $f_{W|\theta}(w) \propto f_{Z|\theta}(z)$ e assim as duas modelagens irão determinar a mesma inferência.

Exemplo: O supervisor de uma fábrica realiza um experimento de verificação de defeitos em 6 peças selecionadas ao acaso. Ele resolveu estudar a probabilidade θ de encontrar uma peça defeituosa entre as fabricadas. Para responder este ponto, o supervisor procurou a ajuda de um estatístico.

Considere a seguinte variável aleatória relacionada à i-ésima peça:

$Y_i = 1$ indica defeituosa e $Y_i = 0$ indica não defeituosa.

É natural assumir que:

- a probabilidade θ é a mesma em todas as avaliações de peças;
- temos independência condicional entre as avaliações dado θ;
- $Y_i|\theta \sim$ Bernoulli(θ);
- Modelo: $f_{Y_i|\theta}(y_i) = \theta^{y_i}(1-\theta)^{1-y_i}$.

Exemplo (continuação):

Primeiro pensamento do estatístico (antes de ver resultados): $\theta = 0.5$

Informação adicional dada pelo supervisor: $\boldsymbol{y} = \{1, 1, 1, 1, 1, 0\}$.

Após observar $\boldsymbol{y}$, o estatístico deve continuar pensando que $\theta = 0.5$?

A resposta é não! Ele deve atualizar seu pensamento sobre θ perante a informação $\boldsymbol{y}$. Isso será feito mais adiante usando o Teorema de Bayes.

Suposição: A ordem de 1's e 0's em $\boldsymbol{y}$ não é relevante.

A informação amostral importante é: "observou-se cinco 1's e um 0".

Verossimilhança: $f_{\boldsymbol{Y}|\theta}(\boldsymbol{y}) = \theta^5 (1-\theta)^1$.

Exemplo (continuação):

Na visão frequentista, a estimação de θ pode considerar três raciocínios diferentes:

- Estimação via máxima verossimilhança: $\hat{\theta} = \frac{\sum_{i=1}^{6} y_i}{6} = 5/6 = 0.83$.
- Seja $W = \sum_{i=1}^{6} y_i = n^o$ de peças defeituosas em 6 experimentos.

  Veja que $W \sim$ Binomial(6,θ) e $E[W/6 \mid \theta] = \frac{E[W|\theta]}{6} = \frac{6\ \theta}{6} = \theta$.

  Em outras palavras, $W/6$ é um estimador não viciado para θ.

  Neste caso, $\hat{\theta} = W/6 = 5/6 = 0.83$.
- Seja $Z = n^o$ de experimentos antes de observar a 1^a peça **sem** defeito. Neste caso, $Z \sim$ B.Negativa$(r, 1-\theta)$ com $r = 1$ sucesso (peça **sem** defeito) e probabilidade de sucesso $1-\theta$.

$$f_{Z|\theta}(z) = \binom{r+z-1}{z}(1-\theta)^r\theta^z$$

$E(Z|\theta) = \frac{r[1-(1-\theta)]}{(1-\theta)} = \frac{\theta}{1-\theta}$ e EMV de $\theta = \frac{z}{z+1} = 5/6 = 0.83$.

Exemplo (continuação):

Na inferência clássica, modelos diferentes com verossimilhanças proporcionais podem determinar alguns tipos de conclusões distintas. Considere, por exemplo, os seguintes cálculos de valores-p (sob a hipótese $\theta = 0.5$):

$$P(W \geq 5) = \sum_{y=5}^{6} \binom{6}{y} 0.5^y \, 0.5^{6-y} = 0.109;$$

$$P(Z \geq 5) = \sum_{z=5}^{6} \binom{1+z-1}{z} 0.5^1 \, 0.5^z = 0.031.$$

O mesmo não ocorre na inferência Bayesiana.

Exemplo (continuação):

Análise Bayesiana para o exemplo discutido nos últimos slides.

O Teorema de Bayes estabelece que:

$$f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y}) = \frac{f_{\theta,\boldsymbol{Y}}(\theta,\boldsymbol{y})}{f_{\boldsymbol{Y}}(\boldsymbol{y})} = \frac{f_{\boldsymbol{Y}|\theta}(\boldsymbol{y})\, f_{\theta}(\theta)}{\int_0^1 f_{\theta,\boldsymbol{Y}}(\theta,\boldsymbol{y})\, d\theta}$$

- $f_{\boldsymbol{Y}|\theta}(\boldsymbol{y})$ é a função de verossimilhança;
- $f_{\theta}(\theta)$ é a distribuição *a priori*;
- $f_{\theta,\boldsymbol{Y}}(\theta,\boldsymbol{y})$ é a distribuição conjunta de θ e $\boldsymbol{Y}$;
- $f_{\boldsymbol{Y}}(\boldsymbol{y})$ é conhecida como distribuição preditiva (não depende de θ);
- $f_{\theta|\boldsymbol{Y}}(\theta \mid \boldsymbol{y})$ é a distribuição *a posteriori*.

Exemplo (continuação):

Nos tópicos apresentados até aqui, sobre o uso do Teorema de Bayes, a distribuição *a priori* foi sempre atribuída a uma variável aleatória não observável que denotávamos por X.

O presente exemplo mostra um ponto central da inferência Bayesiana. A variável aleatória não observável pode ser um parâmetro desconhecido que se deseja estimar no modelo estatístico adotado.

A postura Bayesiana subjetivista estabelece que devemos expressar nossa incerteza sobre quantidades desconhecidas do modelo através de uma afirmação probabilística. Em outras palavras, adota-se uma distribuição de probabilidade para θ.

Esta distribuição de probabilidade $f_\theta(\theta)$ que expressa nosso conhecimento inicial sobre θ (antes de observar os dados) é denominada **distribuição a priori**.

Exemplo (continuação):

Como escolher a distribuição *a priori* neste caso?

Deve-se levar em consideração os seguintes aspectos:

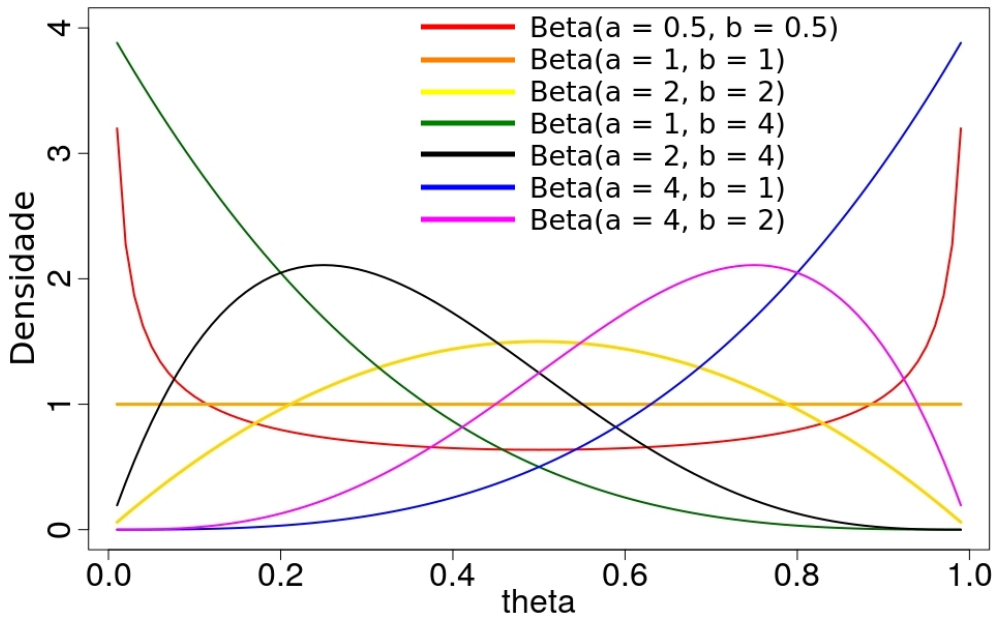
- Espaço paramétrico: $0 \leq \theta \leq 1$;
- Antes de observar $\boldsymbol{y}$ ou conversar com o supervisor, o estatístico não possui muita experiência profissional na área para entender sobre a frequência de peças defeituosas. Desta forma, ele deve ser prudente e expressar grande incerteza sobre o valor de $\theta \in [0, 1]$.
- Expressar pequena incerteza, significa assumir uma distribuição de probabilidade com variância baixa. Por outro lado, expressar grande incerteza implica em adotar uma distribuição de probabilidade com variância alta.
- A distribuição Beta é uma opção interessante aqui, pois ela admite diferentes formatos e está definida exatamente para o intervalo $[0, 1]$.

Exemplo (continuação):

Distribuição *a priori*: $\theta \sim \text{Beta}(a, b)$.

$$f_\theta(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\,\theta^{a-1}(1-\theta)^{b-1} \quad \text{para} \quad 0 \leq \theta \leq 1 \quad \text{e} \quad a, b > 0.$$

Consequentemente, $E(\theta) = \frac{a}{a+b}$ e $Var(\theta) = \frac{ab}{(a+b)^2(a+b+1)}$.



Exemplo (continuação):

Para expressar grande incerteza *a priori* sobre θ, o estatístico pode especificar:
$a = b = 1 \quad \Rightarrow \quad \theta \sim U(0, 1)$.

Dizemos que esta opção é uma especificação **vaga** ou **pouco informativa** com
$E(\theta) = \frac{1}{1+1} = 0.5 \quad \text{e} \quad Var(\theta) = \frac{1}{12} = 0.083$.

Se o estatístico quiser valorizar um pouco mais a ideia de que $\theta = 0.5$ é um valor mais razoável do que os extremos $\theta = 0$ ou $\theta = 1$, ele pode optar por especificar:
$\theta \sim \text{Beta}(2, 2)$.

Esta segunda opção possui
$E(\theta) = \frac{2}{2+2} = 0.5 \quad \text{e} \quad Var(\theta) = \frac{4}{80} = 0.050$.

Perceba que a Beta$(1, 1)$ é **menos informativa** do que a Beta$(2, 2)$.
Esta afirmação é baseada na diferença de magnitude das variâncias
$(0.083 > 0.050)$.

Exemplo (continuação):

O próximo passo da análise é aplicar a regra de Bayes para determinar a distribuição *a posteriori* $f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})$.

Esta distribuição é uma atualização da incerteza inicial sobre θ expressa na distribuição *a priori* $f_{\theta}(\theta)$.

Acrescentar a informação dos dados $\boldsymbol{y}$ provoca alterações na f.d.p. estabelecida *a priori*. Nosso objetivo será avaliar a distribuição *a posteriori* que é formada por consequência dessas alterações.

Na avaliação da distribuição *a posteriori*, iremos explorar suas características como: esperança, moda, mediana, desvio padrão, variância, simetria, etc.

Exemplo (continuação):

Obtendo a distribuição preditiva:

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \int_0^1 f_{\theta,\boldsymbol{Y}}(\theta,\boldsymbol{y})\, d\theta = \int_0^1 f_{\boldsymbol{Y}|\theta}(\boldsymbol{y})\, f_\theta(\theta)\, d\theta,$$

$$= \int_0^1 \theta^{\sum_{i=1}^n y_i}(1-\theta)^{n-\sum_{i=1}^n y_i} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1}\, d\theta,$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 \theta^{\sum_{i=1}^n y_i + a - 1}(1-\theta)^{n-\sum_{i=1}^n y_i + b - 1}\, d\theta.$$

Denote $a^* = \sum_{i=1}^n y_i + a$ e $b^* = n - \sum_{i=1}^n y_i + b$.

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 \theta^{a^*-1}(1-\theta)^{b^*-1}\, d\theta.$$

Dentro da integral é possível identificar o núcleo da f.d.p. de uma Beta(a^*,b^*).

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a^*)\Gamma(b^*)}{\Gamma(a^*+b^*)} \int_0^1 \frac{\Gamma(a^*+b^*)}{\Gamma(a^*)\Gamma(b^*)} \theta^{a^*-1}(1-\theta)^{b^*-1}\, d\theta.$$

A integral resulta em 1, então

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a^*)\Gamma(b^*)}{\Gamma(a^*+b^*)}, \quad \text{tal que } a^* \text{ e } b^* \text{ estão escritos acima.}$$

Exemplo (continuação):

Regra de Bayes estabelece: $f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y}) = \dfrac{f_{\boldsymbol{Y}|\theta}(\boldsymbol{y})\, f_\theta(\theta)}{f_{\boldsymbol{Y}}(\boldsymbol{y})} \propto f_{\boldsymbol{Y}|\theta}(\boldsymbol{y})\, f_\theta(\theta).$

$$f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y}) \propto \theta^{\sum_{i=1}^n y_i}(1-\theta)^{n-\sum_{i=1}^n y_i} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1}.$$

Podemos aproveitar as contas no slide anterior.

$$\propto \theta^{a^*-1}(1-\theta)^{b^*-1}, \quad \text{tal que} \quad a^* = \sum_{i=1}^n y_i + a \quad \text{e} \quad b^* = n - \sum_{i=1}^n y_i + b.$$

Note que os termos que não estavam atrelados a θ foram descartados como parte da constante normalizadora. A expressão acima é o núcleo da Beta(a^*, b^*).

Conclusão: $\theta|\boldsymbol{Y} \sim \text{Beta}(a^*, b^*)$ é a distribuição *a posteriori* desejada.

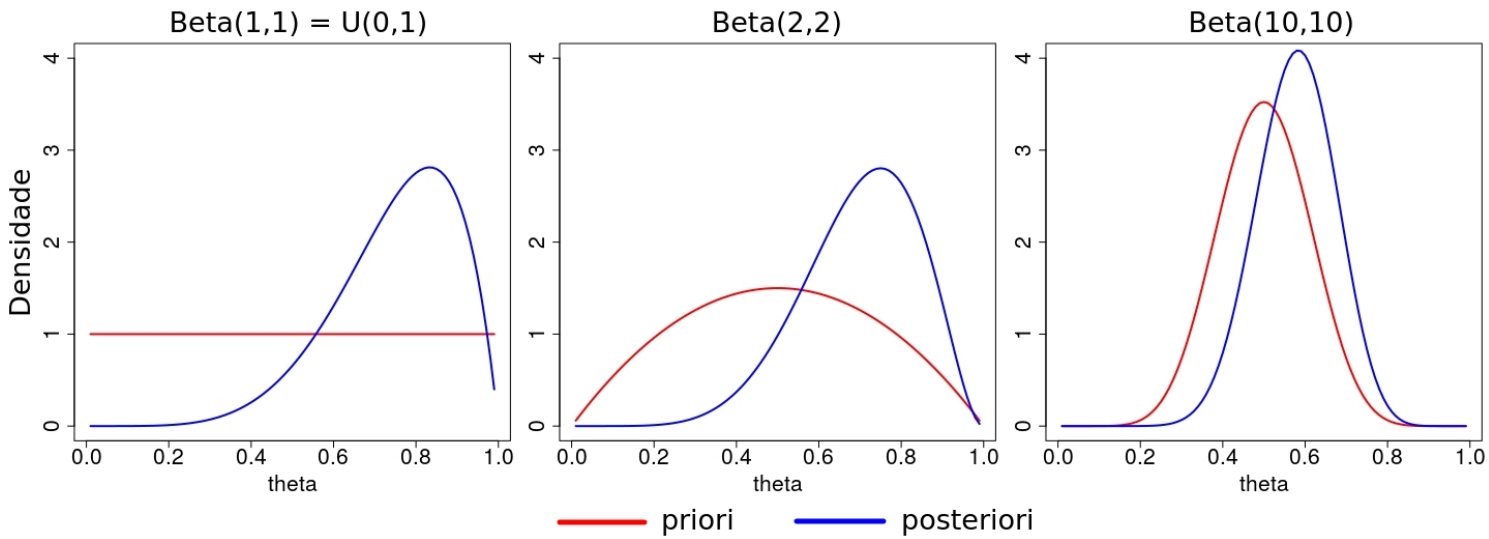
Veja que usar o modelo Binomial Negativo ou o modelo Binomial determinará a mesma conclusão acima, pois sua verossimilhanças são proporcionais e qualquer termo livre de θ será incorporado na constante de normalização.

Exemplo (continuação):

Lembrete: $\boldsymbol{y} = \{1, 1, 1, 1, 1, 0\}$ é a amostra observada neste problema.

Então: $\theta | \boldsymbol{Y} \sim \text{Beta}(a^*, b^*)$ com $a^* = 5 + a$ e $b^* = 6 - 5 + b$.

Os gráficos abaixo comparam três escolhas de distribuições *a priori* feitas pelo estatístico. A distribuição *a posteriori* correspondente está em azul em cada painel.



Exemplo (continuação):

Note que a 3ª especificação *a priori* é a **mais informativa** (possui variância menor). Ela tem grande influência sobre a distribuição *a posteriori* trazendo a f.d.p. azul para uma região mais próxima de 0.5.

Quando uma distribuição **vaga** é adotada *a priori*, a influência dos dados através da verossimilhança é marcante (a f.d.p. azul concentra massa de probabilidade na região acima de 0.6).

A distribuição *a posteriori* concentrando massa de probabilidade na região do espaço paramétrico acima de 0.6 é uma forte indicação de alta probabilidade de peça defeituosa.

O processo de estimação na inferência Bayesiana é focado na avaliação sobre como a f.d.p. *a posteriori* distribui sua massa de probabilidade no espaço paramétrico de θ.

Algumas medidas que ajudam nessa avaliação *a posteriori* são sugeridas no próximo slide.

Exemplo (continuação):

Medidas relacionadas à distribuição *a posteriori* que auxiliam na inferência:

- Média $\hat{\theta}_{me} = E(\theta|\boldsymbol{Y}) = \int_0^1 \theta\, f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})\, d\theta = \frac{a^*}{a^*+b^*}$.
- Mediana $\hat{\theta}_{md}$ é tal que $\int_0^{\hat{\theta}_{md}} f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})\, d\theta = 0.5$.
- Moda $\hat{\theta}_{mo}$ é o valor de θ que maximiza $f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})$.
- Variância $Var(\theta|\boldsymbol{Y}) = \int_0^1 (\theta - \hat{\theta}_{me})^2\, f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})\, d\theta = \frac{a^* b^*}{(a^*+b^*)^2(a^*+b^*+1)}$.
- Desvio padrão $\sqrt{Var(\theta|\boldsymbol{Y})}$

As três primeiras medidas são úteis para estimação pontual.
Seus valores serão nossas estimativas sobre o verdadeiro valor de θ.

As duas últimas medidas indicam o **grau de incerteza** *a posteriori*. Variabilidade *a posteriori* grande ocorre quando a f.d.p. espalha muito a massa de probabilidade no espaço paramétrico. Teríamos alta incerteza sobre o verdadeiro valor de θ (ele pode estar em uma região muito ampla).

Exemplo (continuação):

Um aspecto que afeta o grau de incerteza *a posteriori* é a distribuição *a priori*. Reveja gráficos comparativos alguns slides atrás.

Especificações *a priori* muito informativas (indicando variância e incerteza inicial baixas) ajudam a determinar distribuições *a posteriori* com menos variabilidade. Entretanto, esse tipo de escolha é perigoso, pois se a informação inicial estiver equivocada, a estimação *a posteriori* do parâmetro alvo terá vício.

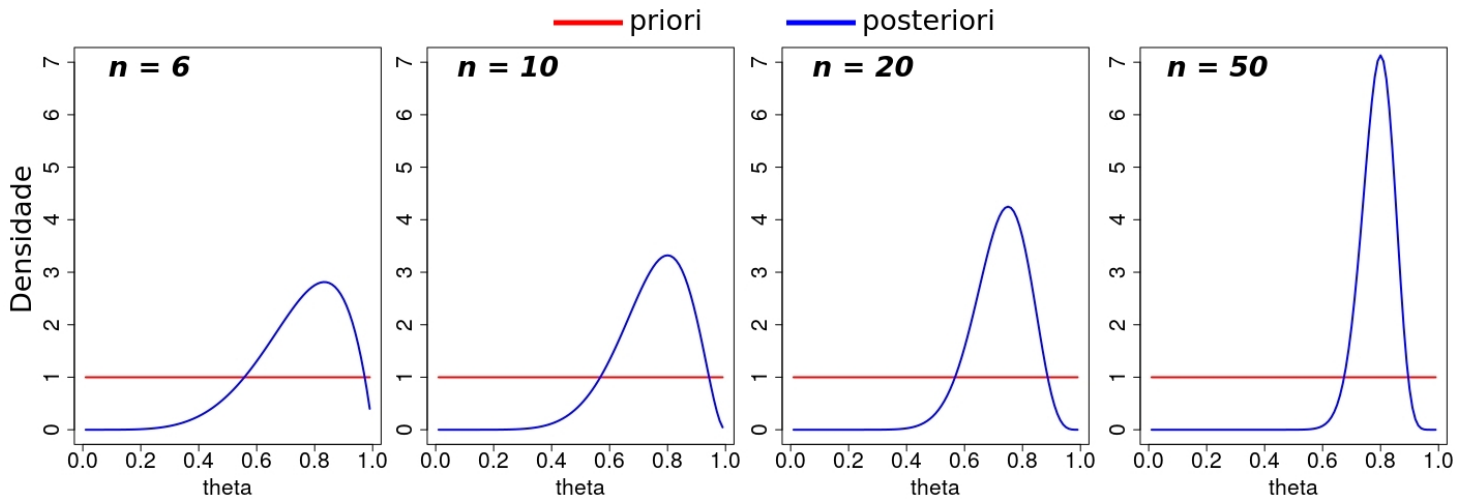
Um 2^o ponto que também impacta o grau de incerteza *a posteriori* é o tamanho amostral n. Quanto maior for n, menor será a variância *a posteriori*.

Suponha que o supervisor da fábrica, coletou informações de peças adicionais:

- Cenário 1 (amostra original, $n = 6$): $\sum_{i=1}^{6} y_i = 5$;
- Cenário 2 (original + 4 peças, $n = 10$): $\sum_{i=1}^{10} y_i = 8$;
- Cenário 3 (original + 14 peças, $n = 20$): $\sum_{i=1}^{20} y_i = 15$;
- Cenário 4 (original + 44 peças, $n = 50$): $\sum_{i=1}^{50} y_i = 40$;

Exemplo (continuação):

Assumindo a especificação *a priori* $\theta \sim \text{Beta}(1, 1)$, teremos as seguintes f.d.p's *a posteriori* em cada cenário.



Veja que, ao aumentar n, a f.d.p. azul concentra cada vez mais a massa de probabilidade na região próxima de 0.8.

A incerteza (a variabilidade) *a posteriori* reduz conforme o tamanho amostral aumenta.

Exemplo (continuação):

$$E(\theta|\boldsymbol{Y}) = \frac{a^*}{a^*+b^*} = \frac{a+\sum_{i=1}^n y_i}{(a+\sum_{i=1}^n y_i)+(b+n-\sum_{i=1}^n y_i)} = \frac{a+\sum_{i=1}^n y_i}{a+b+n} = \frac{a}{a+b+n} + \frac{\sum_{i=1}^n y_i}{a+b+n}$$

$$= \frac{a+b}{a+b}\frac{a}{a+b+n} + \frac{n}{n}\frac{\sum_{i=1}^n y_i}{a+b+n} = \frac{a+b}{a+b+n}\frac{a}{a+b} + \frac{n}{a+b+n}\frac{\sum_{i=1}^n y_i}{n} = \frac{a+b}{a+b+n}E(\theta) + \frac{n}{a+b+n}\bar{y}.$$

Este resultado indica que, neste modelo, a média *a posteriori* pode ser escrita como uma mistura envolvendo a média *a priori* $E(\theta) = a/(a+b)$ e a média amostral $\bar{y}$. O peso da componente $E(\theta)$ depende de n no denominador. O peso da componente $\bar{y}$ depende de n no numerador e denominador.

- Se $n \to \infty \Rightarrow E(\theta|\boldsymbol{Y}) \to \bar{y}$, ou seja,
  a informação inicial terá pouca influência na distribuição *a posteriori*.
- Se $a \to \infty$ e/ou $b \to \infty$, para n fixo, temos $E(\theta|\boldsymbol{Y}) \to E(\theta)$, ou seja,
  a distribuição *a priori* é muito informativa.

Revisitamos agora uma modelagem, explorada algumas aulas atrás, para um problema envolvendo a localização de uma partícula em um teste nuclear.

Suponha que $\boldsymbol{Y} = \{Y_1, Y_2, \cdots, Y_n\}$ é uma amostra aleatória tal que $Y_i \sim \text{Exp}(\theta)$ com média $1/\theta > 0$. Deseja-se estimar o parâmetro desconhecido θ que indexa a distribuição dos Y_i's.

Por independência condicional dos Y_i's dado θ, escrevemos a verossimilhança:

$$f_{\boldsymbol{Y}|\theta}(\boldsymbol{y}) \;=\; \prod_{i=1}^{n} \theta \;\exp\{-\theta y_i\} \;=\; \theta^n \;\exp\{-\theta \;\sum_{i=1}^{n} y_i\}$$

Espaço paramétrico: $\quad \theta \in \mathbb{R}^+$.

Diante deste espaço paramétrico, uma distribuição *a priori* interessante para descrever a incerteza inicial sobre θ é a Ga(a,b) com $a > 0$ e $b > 0$.

$$f_\theta(\theta) \;=\; \frac{b^a}{\Gamma(a)}\; \theta^{a-1} \;\exp\{-b\theta\} \quad \text{para} \quad \theta > 0.$$

Obtendo a distribuição preditiva:

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \int_0^\infty f_{\theta,\boldsymbol{Y}}(\theta,\boldsymbol{y})\, d\theta = \int_0^\infty f_{\boldsymbol{Y}|\theta}(\boldsymbol{y})\, f_\theta(\theta)\, d\theta$$

$$= \int_0^\infty \theta^n \exp\{-\theta \textstyle\sum_{i=1}^n y_i\} \, \frac{b^a}{\Gamma(a)} \theta^{a-1} \exp\{-b\theta\}\, d\theta$$

$$= \frac{b^a}{\Gamma(a)} \int_0^\infty \theta^{a+n-1} \exp\{-(b + \textstyle\sum_{i=1}^n y_i)\theta\}\, d\theta$$

Núcleo da $\text{Ga}(a+n, b+\sum_{i=1}^n y_i)$ dentro da integral.

$$= \frac{b^a}{\Gamma(a)} \, \frac{\Gamma(a+n)}{(b+\sum_{i=1}^n y_i)^{a+n}} \int_0^\infty \frac{(b+\sum_{i=1}^n y_i)^{a+n}}{\Gamma(a+n)} \, \theta^{a+n-1} \exp\{-(b + \textstyle\sum_{i=1}^n y_i)\theta\}\, d\theta$$

A integral resulta em 1, então teremos

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{b^a}{\Gamma(a)} \, \frac{\Gamma(a+n)}{(b+\sum_{i=1}^n y_i)^{a+n}} = \frac{\Gamma(a+n)}{\Gamma(a)} \, b^a \, (b + \textstyle\sum_{i=1}^n y_i)^{-(a+n)}$$

para $y_i > 0, \quad i = 1, 2, \cdots, n.$

Obtendo a distribuição *a posteriori* de θ:

$$f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y}) = \frac{f_{\boldsymbol{Y}|\theta}(\boldsymbol{y}|\theta)\ f_\theta(\theta)}{f_{\boldsymbol{Y}}(\boldsymbol{y})} \propto f_{\boldsymbol{Y}|\theta}(\boldsymbol{y}|\theta)\ f_\theta(\theta) \quad \text{para} \quad \theta > 0.$$

$$\propto \theta^n \exp\{-\theta \textstyle\sum_{i=1}^n y_i\}\ \frac{b^a}{\Gamma(a)}\theta^{a-1} \exp\{-b\theta\}.$$

$$\propto \theta^{a+n-1}\ \exp\{-(b + \textstyle\sum_{i=1}^n y_i)\theta\}.$$

É possível identificar aqui o núcleo de uma distribuição Ga(a^*,b^*) com:

$$a^* = a + n \quad \text{e} \quad b^* = b + \textstyle\sum_{i=1}^n y_i.$$

Esta é a distribuição *a posteriori* de θ.

Estudo simulado usando o modelo Exponencial apresentado nos últimos slides:

Passos para gerar dados artificiais (comandos R):

```
thetareal = 2            # defina o verdadeiro valor de theta;
n = 30                   # defina o tamanho amostral;
y = rexp(n, thetareal)   # gere observações da Exponencial.
```

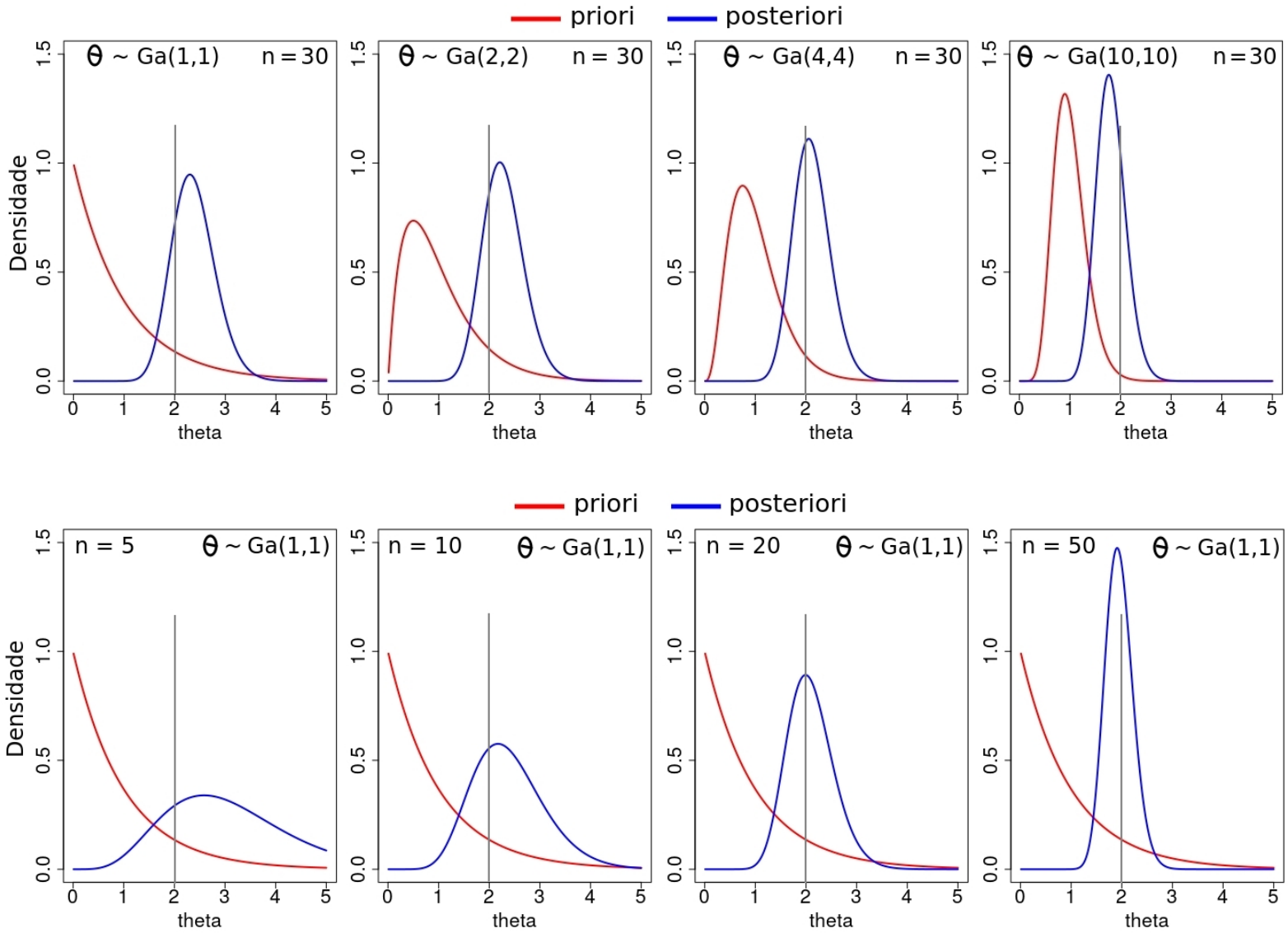
Lembre-se que na prática, não conhecemos o valor real de θ.
Ele é usado acima apenas para gerar a amostra $\boldsymbol{y}$.

A vantagem de um estudo simulado é que poderemos verificar se a distribuição *a posteriori* concentra massa de probabilidade na região próxima a $\theta_{\text{real}} = 2$.

Os gráficos a seguir mostram a distribuição *a posteriori* perante diferentes especificações *a priori* e diferentes valores de n.

Comandos para cálculo da f.d.p. *a priori* e *a posteriori*:

```
theta = seq(0.01, 5, 0.01)
prior = dgamma(theta, a, rate = b)
post = dgamma(theta, a + n, rate = b + sum(y))
```


priori
posteriori
θ ~ Ga(1,1) n = 30
θ ~ Ga(2,2) n = 30
θ ~ Ga(4,4) n = 30
θ ~ Ga(10,10) n = 30
n = 5 θ ~ Ga(1,1)
n = 10 θ ~ Ga(1,1)
n = 20 θ ~ Ga(1,1)
n = 50 θ ~ Ga(1,1)
Densidade
theta

Comentários sobre os gráficos no slide anterior:

- Todas as especificações *a priori* consideradas determinam $E(\theta) = a/b = 1$.
- Conforme $a = b$ aumenta, temos $Var(\theta) = a/b^2$ diminuindo.
- Assumir *a priori* $\theta \sim \text{Ga}(10, 10)$ é ruim, pois $Var(\theta)$ é pequena e $E(\theta) = 1 \neq 2 = \theta_{\text{real}}$. Note que a f.d.p. *a posteriori* é influenciada e desloca um pouco para a esquerda.
- Em geral, para $n = 30$, a verossimilhança domina a informação *a priori*. Veja que a f.d.p. *a posteriori* (azul) coloca massa de probabilidade perto de $\theta_{\text{real}} = 2$ para esse tamanho amostral.
- Conforme n aumenta, é evidente a mudança de formato da f.d.p. *a posteriori*.
- Para $n = 5$ temos $Var(\theta|\boldsymbol{Y})$ grande e alto grau de incerteza sobre θ, entretanto, é notável que as 5 observações ajudaram a melhorar um pouco a informação inicial sobre θ.
- Para $n = 50$ temos $Var(\theta|\boldsymbol{Y})$ pequena e baixo grau de incerteza sobre θ (a f.d.p. *a posteriori* está muito concentrada ao redor de $\theta_{\text{real}} = 2$).

Admita agora que $\boldsymbol{Y} = \{Y_1, Y_2, \cdots, Y_n\}$ é uma amostra aleatória com $Y_i \sim \text{Poisson}(\theta)$, sendo $\theta > 0$ a média esperada de eventos em certo período de tempo/espaço.

Por independência condicional dos Y_i's dado θ, temos a verossimilhança:

$$f_{\boldsymbol{Y}|\theta}(\boldsymbol{y}) = \prod_{i=1}^{n} \frac{\theta^{y_i}}{y_i!} \exp\{-\theta\} = \frac{\theta^{\sum_{i=1}^{n} y_i}}{\prod_{i=1}^{n} y_i!} \exp\{-n\theta\}$$

Espaço paramétrico: $\theta \in \mathbb{R}^+$.

Considere também aqui a distribuição *a priori* Ga(a,b), com $a > 0$ e $b > 0$, para descrever a incerteza inicial sobre θ.

$$f_\theta(\theta) = \frac{b^a}{\Gamma(a)} \theta^{a-1} \exp\{-b\theta\} \quad \text{para} \quad \theta > 0.$$

Distribuição preditiva de $\boldsymbol{Y}$:

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \int_0^\infty f_{\theta,\boldsymbol{Y}}(\theta,\boldsymbol{y})\, d\theta = \int_0^\infty f_{\boldsymbol{Y}|\theta}(\boldsymbol{y})\, f_\theta(\theta)\, d\theta$$

$$= \int_0^\infty \frac{\theta^{\sum_{i=1}^n y_i}}{\prod_{i=1}^n y_i!} \exp\{-n\theta\}\, \frac{b^a}{\Gamma(a)}\theta^{a-1}\exp\{-b\theta\}\, d\theta$$

$$= \frac{1}{\prod_{i=1}^n y_i}\, \frac{b^a}{\Gamma(a)}\, \int_0^\infty \theta^{a+\sum_{i=1}^n y_i-1}\exp\{-(b+n)\theta\}\, d\theta$$

Núcleo da f.d.p. $\text{Ga}(a+\sum_{i=1}^n y_i, b+n)$ dentro da integral.

$$= \frac{1}{\prod_{i=1}^n y_i!}\, \frac{b^a}{\Gamma(a)}\, \frac{\Gamma(a+\sum_{i=1}^n y_i)}{(b+n)^{a+\sum_{i=1}^n y_i}}\, \int_0^\infty \frac{(b+n)^{a+\sum_{i=1}^n y_i}}{\Gamma(a+\sum_{i=1}^n y_i)}\, \theta^{a+\sum_{i=1}^n y_i-1}\exp\{-(b+n)\theta\}\, d\theta$$

A integral resulta em 1.

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{1}{\prod_{i=1}^n y_i!}\, \frac{b^a}{\Gamma(a)}\, \frac{\Gamma(a+\sum_{i=1}^n y_i)}{(b+n)^{a+\sum_{i=1}^n y_i}} \quad \text{para} \quad y_i = 0,1,2,\cdots$$

Distribuição *a posteriori* de θ:

$$f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y}) = \frac{f_{\boldsymbol{Y}|\theta}(\boldsymbol{y}|\theta)\ f_\theta(\theta)}{f_{\boldsymbol{Y}}(\boldsymbol{y})} \propto f_{\boldsymbol{Y}|\theta}(\boldsymbol{y}|\theta)\ f_\theta(\theta), \quad \text{para} \quad \theta > 0.$$

$$\propto \frac{\theta^{\sum_{i=1}^n y_i}}{\prod_{i=1}^n y_i!}\ \exp\{-n\theta\}\ \frac{b^a}{\Gamma(a)}\theta^{a-1}\exp\{-b\theta\}.$$

$$\propto \theta^{a+\sum_{i=1}^n y_i - 1}\exp\{-(b+n)\theta\}.$$

Identifica-se aqui o núcleo da distribuição Ga(a^*,b^*) com:

$$a^* = a + \sum_{i=1}^n y_i \quad \text{e} \quad b^* = b + n.$$

Esta é a distribuição *a posteriori* de θ.

Estudo simulado usando o modelo Poisson apresentado nos últimos slides:

Gerando dados artificiais (comandos R):

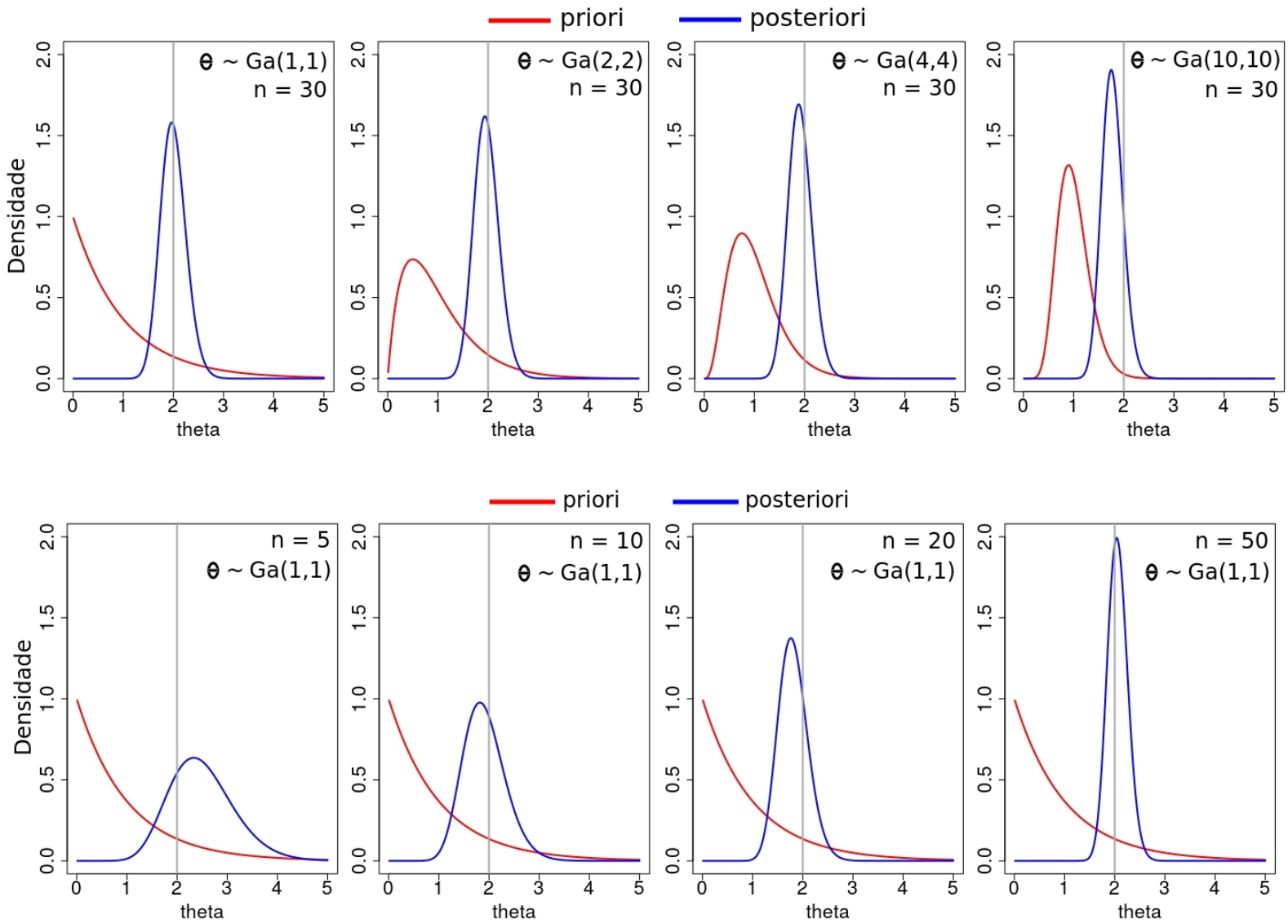
```
thetareal = 2              # defina o verdadeiro valor de theta;
n = 30                     # defina o tamanho amostral;
y = rpois(n, thetareal)    # gere observações da Exponencial.
```

Na prática, θ_{real} é deconhecido. O pesquisador tem acesso apenas à amostra $\boldsymbol{y}$.

Os gráficos a seguir mostram a distribuição *a posteriori* perante diferentes especificações *a priori* e diferentes valores de n.

Comandos para cálculo da f.d.p. *a priori* e *a posteriori*:

```
theta = seq(0.01, 5, 0.01)
prior = dgamma(theta, a, rate = b)
post = dgamma(theta, a + sum(y), rate = b + n)
```


priori
posteriori
Densidade
θ ~ Ga(1,1)
n = 30
θ ~ Ga(2,2)
n = 30
θ ~ Ga(4,4)
n = 30
θ ~ Ga(10,10)
n = 30
n = 5
θ ~ Ga(1,1)
n = 10
θ ~ Ga(1,1)
n = 20
θ ~ Ga(1,1)
n = 50
θ ~ Ga(1,1)
theta

Comentários sobre os gráficos no slide anterior:

- Todas as especificações *a priori* indicam $E(\theta) = a/b = 1$.
- $a = b$ aumentando $\Rightarrow$ $Var(\theta) = a/b^2$ diminuindo.
- $\theta \sim \text{Ga}(10, 10) \Rightarrow Var(\theta)$ pequena e $E(\theta) = 1 \neq 2 = \theta_{\text{real}}$;
  f.d.p. *a posteriori* é influenciada e desloca um pouco para a esquerda.
- Para $n = 30$, a verossimilhança domina a informação *a priori*;
  f.d.p. *a posteriori* (azul) tem massa de probabilidade perto de $\theta_{\text{real}} = 2$.
- n aumentando $\Rightarrow$ f.d.p. *a posteriori* muda formato.
- $n = 5$ determina $Var(\theta|\boldsymbol{Y})$ grande e incerteza alta sobre θ;
  5 observações já ajudam a melhorar um pouco a informação inicial sobre θ.
- $n = 50$ determina $Var(\theta|\boldsymbol{Y})$ pequena e incerteza baixa sobre θ;
  f.d.p. *a posteriori* concentra-se ao redor de $\theta_{\text{real}} = 2$.

Distribuições preditivas:

Note que, nos estudos anteriores, foi possível obter a distribuição preditiva $f_{\boldsymbol{Y}}(\boldsymbol{y})$ que aparece no denominador da regra de Bayes.

Entretanto, o cálculo da distribuição *a posteriori* foi feito usando o símbolo de proporcionalidade "$\propto$", o qual indica que podemos focar apenas no numerador $f_{\boldsymbol{Y}|\theta}(\boldsymbol{y})\, f_{\theta}(\theta)$, i.e. verossimilhança $\times$ priori. Esta é a parte principal do Teorema de Bayes, pois é aqui que aparece θ na formulação.

Os pontos centrais nesta opção de cálculo ignorando a preditiva é:

1. Manter todos os termos atrelados a θ e eliminar (atribuir à constante normalizadora) aqueles que estão livres de θ.
2. Organizar o resultado para tentar reconhecer o núcleo de uma distribuição de probabilidade existente na literatura.

Apesar da preditiva ser usualmente desconsiderada no cálculo da distribuição *a posteriori*, elas determinam algumas conclusões interessantes.

A **distribuição preditiva a priori** fornece a probabilidade de observarmos uma amostra $\boldsymbol{y} = \{y_1, y_2, \cdots, y_n\}$.

Ela descreve o comportamento conjunto das variáveis aleatórias $Y_1, Y_2, \cdots, Y_n$, sem considerar θ (que é eliminado com a integração).

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \int_{\Theta} f_{\boldsymbol{Y},\theta}(\boldsymbol{y}, \theta)\, d\theta = \int_{\Theta} f_{\boldsymbol{Y}|\theta}(\boldsymbol{y})\, f_{\theta}(\theta)\, d\theta.$$

Esta foi a conta feita nos exemplos anteriores.

Um outro tipo de preditiva pode ser obtida levando em conta o contexto de avaliação de dados observados e dados futuros. Ela é denominada **distribuição preditiva a posteriori**.

Seja $\boldsymbol{y} = \{y_1, y_2, \cdots, y_n\}$ o conjunto de valores observados a partir de uma população com distribuição conjunta condicional $f_{\boldsymbol{Y}|\theta}(\boldsymbol{y})$.

Admita a f.d.p. *a priori* $f_{\theta}(\theta)$.

Considere o conjunto $\boldsymbol{Y}^\star = \{Y^\star_{n+1}, Y^\star_{n+2}, \cdots, Y^\star_m\}$ representando variáveis aleatórias medindo observações futuras.

Suposição de independência condicional:

Dado θ, o conjunto $\{\boldsymbol{Y}, \boldsymbol{Y}^\star\} = \{Y_1, \cdots, Y_n, Y^\star_{n+1}, \cdots, Y^\star_m\}$ tem distribuição conjunta que fatora como segue:

$$f_{\boldsymbol{Y},\boldsymbol{Y}^\star|\theta}(\boldsymbol{y}, \boldsymbol{y}^\star) = \prod_{i=1}^{n} f_{Y_i|\theta}(y_i) \prod_{j=n+1}^{m} f_{Y^\star_j|\theta}(y^\star_j).$$

Note que:

$$f_{\boldsymbol{Y}^\star|\boldsymbol{Y}}(\boldsymbol{y}^\star|\boldsymbol{y}) = \frac{f_{\boldsymbol{Y},\boldsymbol{Y}^\star}(\boldsymbol{y}, \boldsymbol{y}^\star)}{f_{\boldsymbol{Y}}(\boldsymbol{y})} = \frac{\int_\Theta f_{\boldsymbol{Y},\boldsymbol{Y}^\star,\theta}(\boldsymbol{y}, \boldsymbol{y}^\star, \theta)\, d\theta}{\int_\Theta f_{\boldsymbol{Y},\theta}(\boldsymbol{y}, \theta)\, d\theta}$$

$$= \frac{\int_\Theta f_{\boldsymbol{Y}^\star|\boldsymbol{Y},\theta}(\boldsymbol{y}^\star|\boldsymbol{y})\, f_{\boldsymbol{Y},\theta}(\boldsymbol{y}, \theta)\, d\theta}{\int_\Theta f_{\boldsymbol{Y},\theta}(\boldsymbol{y}, \theta)\, d\theta} = \frac{\int_\Theta f_{\boldsymbol{Y}^\star|\boldsymbol{Y},\theta}(\boldsymbol{y}^\star|\boldsymbol{y})\, f_{\boldsymbol{Y}|\theta}(\boldsymbol{y})\, f_\theta(\theta)\, d\theta}{\int_\Theta f_{\boldsymbol{Y}|\theta}(\boldsymbol{y})\, f_\theta(\theta)\, d\theta}$$

O termo $f_{\boldsymbol{Y}^\star|\boldsymbol{Y},\theta}(\boldsymbol{y}^\star|\boldsymbol{y})$ é a verossimilhança dos dados futuros.

O termo $f_{\boldsymbol{Y}|\theta}(\boldsymbol{y})$ é a verossimilhança dos dados observados.

Perceba que a distribuição preditiva *a priori* $f_{\boldsymbol{Y}}(\boldsymbol{y})$ aparece no denominador da expressão no slide anterior. Essa distribuição não depende de θ e, por esta razão, pode ser inserida dentro da integral do numerador.

$$f_{\boldsymbol{Y}^\star|\boldsymbol{Y}}(\boldsymbol{y}^\star|\boldsymbol{y}) = \frac{\int_\Theta f_{\boldsymbol{Y}^\star|\boldsymbol{Y},\theta}(\boldsymbol{y}^\star|\boldsymbol{y})\, f_{\boldsymbol{Y}|\theta}(\boldsymbol{y})\, f_\theta(\theta)\, d\theta}{\int_\Theta f_{\boldsymbol{Y}|\theta}(\boldsymbol{y})\, f_\theta(\theta)\, d\theta}$$

$$= \int_\Theta f_{\boldsymbol{Y}^\star|\boldsymbol{Y},\theta}(\boldsymbol{y}^\star|\boldsymbol{y}) \frac{f_{\boldsymbol{Y}|\theta}(\boldsymbol{y})\, f_\theta(\theta)}{f_{\boldsymbol{Y}}(\boldsymbol{y})}\, d\theta = \int_\Theta f_{\boldsymbol{Y}^\star|\boldsymbol{Y},\theta}(\boldsymbol{y}^\star|\boldsymbol{y})\, f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})\, d\theta$$

Compare as preditivas:

- Preditiva *a priori* $f_{\boldsymbol{Y}}(\boldsymbol{y}) = \int_\Theta f_{\boldsymbol{Y}|\theta}(\boldsymbol{y})\, f_\theta(\theta)\, d\theta = E_\theta[f_{\boldsymbol{Y}|\theta}(\boldsymbol{y})]$.
- Preditiva *a posteriori*
  $f_{\boldsymbol{Y}^\star|\boldsymbol{Y}}(\boldsymbol{y}^\star|\boldsymbol{y}) = \int_\Theta f_{\boldsymbol{Y}^\star|\boldsymbol{Y},\theta}(\boldsymbol{y}^\star|\boldsymbol{y})\, f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})\, d\theta = E_{\theta|\boldsymbol{Y}}[f_{\boldsymbol{Y}^\star|\boldsymbol{Y},\theta}(\boldsymbol{y}^\star|\boldsymbol{y})]$.

Exemplo: Voltamos ao problema de peças defeituosas produzidas em uma fábrica. Lembre-se que uma amostra de tamanho $n = 6$ peças foi avaliada e indicou o resultado $\boldsymbol{y} = \{1, 1, 1, 1, 1, 0\}$ ($1 =$ com defeito, $0 =$ sem defeito).

O parâmetro de interesse é θ, representando a probabilidade de classificar uma peça qualquer como defeituosa.

Modelo: $Y_i|\theta \sim \text{Bernoulli}(\theta)$ para $Y_i \in \{0, 1\}$ e $\theta \in [0, 1]$.

Adotando independência condicional entre os Y_i's, a verossimilhança dos dados observados será:

$$f_{\boldsymbol{Y}|\theta}(\boldsymbol{y}) = \prod_{i=1}^{n} f_{\boldsymbol{Y_i}|\theta}(\boldsymbol{y_i}) = \theta^{\sum_{i=1}^{n} y_i} (1-\theta)^{n-\sum_{i=1}^{n} y_i} = \theta^5 (1-\theta)^{6-5}.$$

Distribuição *a priori*: $\theta \sim \text{Beta}(a, b)$.

Distribuição *a posteriori* (rever slides, conta já feita):

$\theta|\boldsymbol{y} \sim \text{Beta}(a^*, b^*)$, com

$a^* = a + \sum_{i=1}^{n} y_i = a + 5$ e $b^* = b + n - \sum_{i=1}^{n} y_i = b + 6 - 5$.

O objetivo aqui é avaliar a probabilidade de uma futura 7^a peça (ainda não observada) ser defeituosa.

A independência condicional estabelece que

$f_{Y_7^\star|\boldsymbol{Y},\theta}(y_7^\star|\boldsymbol{y}) = f_{Y_7^\star|\theta}(y_7^\star) = \theta^{y_7^\star}(1-\theta)^{1-y_7^\star}$; solução simples, mas depende de θ.

Alternativamente, trabalhe com a preditiva *a posteriori*

$$f_{Y_7^\star|\boldsymbol{Y}}(y_7^\star = 1|\boldsymbol{y}) = \int_0^1 f_{Y_7^\star,\theta|\boldsymbol{Y}}(y_7^\star = 1, \theta|\boldsymbol{y})\, d\theta = \int_0^1 f_{Y_7^\star|\boldsymbol{Y},\theta}(y_7^\star = 1|\boldsymbol{y})\, f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})\, d\theta.$$

Supondo *a priori* $\theta \sim \text{Beta}(1,1)$, teremos *a posteriori* $\theta|\boldsymbol{y} \sim \text{Beta}(6,2)$. Logo

$$f_{Y_7^\star|\boldsymbol{Y}}(y_7^\star = 1|\boldsymbol{y}) = \int_0^1 \theta^1(1-\theta)^{1-1}\, \frac{\Gamma(6+2)}{\Gamma(6)\Gamma(2)}\theta^{6-1}(1-\theta)^{2-1}\, d\theta.$$

$$= \frac{\Gamma(6+2)}{\Gamma(6)\Gamma(2)} \int_0^1 \theta^{1+6-1}(1-\theta)^{2-1}\, d\theta.$$ Núcleo da Beta(7,2) dentro da integral.

$$= \frac{\Gamma(6+2)}{\Gamma(6)\Gamma(2)}\, \frac{\Gamma(7)\Gamma(2)}{\Gamma(7+2)} \int_0^1 \frac{\Gamma(7+2)}{\Gamma(7)\Gamma(2)}\, \theta^{7-1}(1-\theta)^{2-1}\, d\theta.$$ Integral resulta em 1.

Conclusão: $f_{Y_7^\star|\boldsymbol{Y}}(y_7^\star = 1|\boldsymbol{y}) = \frac{\Gamma(8)}{\Gamma(6)\Gamma(2)}\, \frac{\Gamma(7)\Gamma(2)}{\Gamma(9)} = 7/9 = 0.778$

Formato mais geral da preditiva *a posteriori* avaliada nos últimos slides:

Assuma $\boldsymbol{Y} = \{Y_1, \cdots, Y_n\}$.

$$f_{Y^\star_{n+1}|\boldsymbol{Y}}(y^\star_{n+1}|\boldsymbol{y}) = \int_0^1 \binom{1}{y^\star_{n+1}} \theta^{y^\star_{n+1}} (1-\theta)^{1-y^\star_{n+1}} \times$$

$$\times \frac{\Gamma(a+b+n)}{\Gamma(a+\sum_{i=1}^n y_i)\Gamma(b+n-\sum_{i=1}^n y_i)} \theta^{a+\sum_{i=1}^n y_i - 1} (1-\theta)^{b+n-\sum_{i=1}^n y_i - 1} \, d\theta.$$

Lembrete: $a^* = a + \sum_{i=1}^n y_i$ e $b^* = b + n - \sum_{i=1}^n y_i$.

$$= \binom{1}{y^\star_{n+1}} \frac{\Gamma(a^*+b^*)}{\Gamma(a^*)\Gamma(b^*)} \int_0^1 \theta^{y^\star_{n+1}+a^*-1} (1-\theta)^{1-y^\star_{n+1}+b^*-1} \, d\theta.$$

Núcleo da Beta$(y^\star_{n+1} + a^*, 1 - y^\star_{n+1} + b^*)$ dentro da integral.

$$= \binom{1}{y^\star_{n+1}} \frac{\Gamma(a^*+b^*)}{\Gamma(a^*)\Gamma(b^*)} \frac{\Gamma(y^\star_{n+1}+a^*)\Gamma(1-y^\star_{n+1}+b^*)}{\Gamma(a^*+b^*+1)} \times$$

$$\times \int_0^1 \frac{\Gamma(a^*+b^*+1)}{\Gamma(y^\star_{n+1}+a^*)\Gamma(1-y^\star_{n+1}+b^*)} \theta^{y^\star_{n+1}+a^*-1} (1-\theta)^{1-y^\star_{n+1}+b^*-1} \, d\theta.$$

Integral resulta em 1.

Conclusão: $f_{Y^\star_{n+1}|\boldsymbol{Y}}(y^\star_{n+1}|\boldsymbol{y}) = \binom{1}{y^\star_{n+1}} \frac{\Gamma(a^*+b^*)}{\Gamma(a^*)\Gamma(b^*)} \frac{\Gamma(a^*+y^\star_{n+1})\Gamma(b^*+1-y^\star_{n+1})}{\Gamma(a^*+b^*+1)}$,

para $y^\star_{n+1} \in \{0,1\}$.

A expressão acima é um caso particular da distribuição Beta-Binomial(κ,α,β) cuja f.m.p. é dada por:

$$f_{W|\kappa,\alpha,\beta}(w) = \binom{\kappa}{w} \frac{\Gamma(\alpha+\beta)\,\Gamma(\alpha+w)\,\Gamma(\beta+\kappa-w)}{\Gamma(\alpha)\,\Gamma(\beta)\,\Gamma(\alpha+\beta+\kappa)},$$

para $\alpha > 0$, $\beta > 0$, $\kappa \in \{1,2,3,\cdots\}$ e $w \in \{0,1,2,\cdots,\kappa\}$.

$E(W) = \kappa\frac{\alpha}{\alpha+\beta}$ e $Var(W) = \frac{\kappa\alpha\beta}{(\alpha+\beta)^2} \frac{\alpha+\beta+\kappa}{\alpha+\beta+1}$.

Temos $Y^\star_{n+1}|\boldsymbol{Y} \sim$ Beta-Binomial$(1, a^*, b^*)$ com $E(Y^\star_{n+1}|\boldsymbol{Y}) = 1\frac{a^*}{a^*+b^*}$.

.....

Observação: Se $Z|\theta \sim$ Binomial(κ,θ) e $\theta \sim$ Beta(α,β), então $Z \sim$ Beta-Binomial(κ,α,β).