

Normal vs. t : posterior, preditiva e fator de Bayes

Advanced Bayesian Econometrics

Hedibert Freitas Lopes

Inspere

Problema

Dois modelos para $x_1, \dots, x_n$:

$$M_0 : x_i \mid \theta \stackrel{iid}{\sim} N(\theta, 1)$$

$$M_1 : x_i \mid \theta \stackrel{iid}{\sim} t_\nu(\theta, 1), \quad \nu = 4$$

Em ambos, a priori é $\theta \sim N(0, 1)$.

Objetivos

- $E(\theta \mid x)$, $V(\theta \mid x)$, $DP(\theta \mid x)$ e $\Pr(\theta > 0 \mid x)$
- Preditivas $p(x \mid M_0)$ e $p(x \mid M_1)$, e o fator de Bayes BF_{10}
- Comparação com $p(x \mid \theta = 0)$, ou seja, $H_0 : \theta = 0$

Dados

- Dados 1: 0.18 2.21 1.58 1.82 -0.65 1.29 0.12 2.08 -0.19 -1.84
- Dados 2: -0.32 0.13 0.33 -0.16 2.64 -0.39 -0.03 0.58 1.82 2.68

Dados 1 e 2 têm $n = 10$; dados 3 e 4 (próximo slide) têm $n = 30$.

Dados 3 e 4 ($n = 30$)

Dados 3

1.33	-0.31	1.30	-0.31	1.98	1.13	0.75	-0.20	0.18	1.29
0.32	1.13	-0.45	-0.11	0.45	1.35	0.16	-0.35	0.17	-0.37
1.94	1.84	0.83	0.89	0.36	-0.15	1.05	0.85	-0.11	1.73

Dados 4

0.87	-0.42	0.89	2.12	1.86	-0.02	-0.07	0.53	0.84	-0.36
1.32	3.27	2.28	0.32	0.06	0.70	-0.89	0.28	1.71	0.98
-1.05	3.78	-0.41	0.59	2.69	0.29	1.24	0.92	-0.23	0.76

Posterior conjugada

$$\theta \mid x \sim N\left(\frac{n\bar{x}}{n+1}, \frac{1}{n+1}\right), \quad \Pr(\theta > 0 \mid x) = \Phi\left(\frac{n\bar{x}}{\sqrt{n+1}}\right).$$

Preditiva: como $x = \theta 1 + \varepsilon$, temos $x \sim N_n(0, I_n + 11')$, de modo que

$$p(x \mid M_0) = (2\pi)^{-n/2} (n+1)^{-1/2} \exp\left\{-\frac{1}{2} \left(\sum_i x_i^2 - \frac{(\sum_i x_i)^2}{n+1}\right)\right\}.$$

Sob H_0 : $p(x \mid \theta = 0) = \prod_i \phi(x_i)$.

M_1 : integração de Riemann

Não há forma fechada. Usamos uma grade $\theta_1, \dots, \theta_N$ de pontos médios em $[-2, 2]$, com $N = 10.000$ e $h = 4/N$.

$$p(x \mid M_1) \approx h \sum_{j=1}^N \underbrace{\prod_{i=1}^n t_4(x_i - \theta_j)}_{g(\theta_j)} \phi(\theta_j)$$

$$\pi(\theta_j \mid x) \approx g(\theta_j) / p(x \mid M_1)$$

$$E(\theta \mid x) \approx h \sum_j \theta_j \pi(\theta_j \mid x), \quad \Pr(\theta > 0 \mid x) \approx h \sum_{\theta_j > 0} \pi(\theta_j \mid x)$$

Sob H_0 : $p(x \mid \theta = 0) = \prod_i t_4(x_i)$.

A grade omite cerca de 4.6% da massa a priori. A verossimilhança é desprezível fora de $[-2, 2]$, então o efeito é irrelevante.

Dados 1 ($\bar{x} = 0.66$)

	$E(\theta x)$	$V(\theta x)$	DP	$\Pr(\theta > 0 x)$
M_0	0.6000	0.0909	0.3015	0.9767
M_1	0.6563	0.1518	0.3896	0.9529

	$p(x M)$	$p(x \theta = 0)$	$p(x \theta = 0)/p(x M)$
Normal	7.597×10^{-9}	3.479×10^{-9}	0.458
t_4	8.510×10^{-9}	5.266×10^{-9}	0.619

$$BF_{10} = \frac{p(x | M_1)}{p(x | M_0)} = 1.120 \quad (BF_{01} = 0.893)$$

Dados 2 ($\bar{x} = 0.728$)

	$E(\theta x)$	$V(\theta x)$	DP	$\Pr(\theta > 0 x)$
M_0	0.6618	0.0909	0.3015	0.9859
M_1	0.4479	0.1244	0.3527	0.9012

	$p(x M)$	$p(x \theta = 0)$	$p(x \theta = 0)/p(x M)$
Normal	3.810×10^{-8}	1.136×10^{-8}	0.298
t_4	3.707×10^{-8}	4.756×10^{-8}	1.283

$$BF_{10} = \frac{p(x | M_1)}{p(x | M_0)} = 0.973 \quad (BF_{01} = 1.028)$$

Dados 3 ($n = 30$, $\bar{x} = 0.622$)

	$E(\theta x)$	$V(\theta x)$	DP	$\Pr(\theta > 0 x)$
M_0	0.6023	0.0323	0.1796	0.9996
M_1	0.5823	0.0371	0.1927	0.9987

	$p(x M)$	$p(x \theta = 0)$	$p(x \theta = 0)/p(x M)$
Normal	3.110×10^{-17}	6.263×10^{-19}	0.020
t_4	1.937×10^{-18}	1.063×10^{-19}	0.055

$$BF_{10} = \frac{p(x | M_1)}{p(x | M_0)} = 0.062 \quad (BF_{01} = 16.06)$$

Dados 4 ($n = 30$, $\bar{x} = 0.828$)

	$E(\theta x)$	$V(\theta x)$	DP	$\Pr(\theta > 0 x)$
M_0	0.8016	0.0323	0.1796	0.999996
M_1	0.6724	0.0429	0.2071	0.9994

	$p(x M)$	$p(x \theta = 0)$	$p(x \theta = 0)/p(x M)$
Normal	3.736×10^{-22}	9.830×10^{-26}	0.0003
t_4	6.245×10^{-22}	1.525×10^{-23}	0.024

$$BF_{10} = \frac{p(x | M_1)}{p(x | M_0)} = 1.672 \quad (BF_{01} = 0.598)$$

Como $H_0 : \theta = 0$ está aninhada em cada modelo,

$$\frac{p(x \mid \theta = 0)}{p(x \mid M)} = \frac{\pi(\theta = 0 \mid x)}{\pi(\theta = 0)}.$$

Verificação (M_0 , dados 1):

$$\frac{\phi(0; 0.6, 0.0909)}{\phi(0; 0, 1)} = \frac{0.1826}{0.3989} = 0.458.$$

O valor coincide com a tabela.

- M_0 só enxerga $\bar{x}$. A variância a posteriori é sempre $1/(n+1)$ ($1/11$ ou $1/31$), e a média é $\bar{x}$ contraída em direção a zero.
- M_1 enxerga a configuração dos dados.
 - Dados 2: os valores 2.64 e 2.68 caem nas caudas. A média vai a 0.45 e $\Pr(\theta > 0)$ cai para 0.90.
 - Dados 1: a dispersão é simétrica (inclui -1.84). A posterior alarga (DP 0.39), mas a média não é puxada para baixo.
- $BF_{10} \approx 1$ nos dois casos. Com $n = 10$, os dados não discriminam caudas normais de caudas t_4 .
- O teste de H_0 muda com a verossimilhança.
 - Dados 2 sob a normal: cerca de 3:1 contra $\theta = 0$.
 - Dados 2 sob a t_4 : 1.28 a favor de $\theta = 0$.

A evidência contra a nula vinha dos dois valores extremos, que a t descarta.

- **Com $n = 30$ os dados já discriminam as caudas.**
 - Dados 3 não têm valores extremos: $BF_{01} = 16$, evidência forte a favor da normal.
 - Dados 4 têm 3.27 e 3.78: $BF_{10} = 1.67$, leve preferência pela t_4 .
- **Robustez da média a posteriori.**
 - Dados 3: M_0 e M_1 quase coincidem (0.60 vs. 0.58).
 - Dados 4: a t desconta os extremos e a média cai de 0.80 para 0.67.
- **$H_0 : \theta = 0$ é rejeitada nos dois modelos.**
 - A magnitude da evidência, porém, depende da verossimilhança.
 - Dados 4: cerca de 3800:1 contra H_0 sob a normal, e cerca de 41:1 sob a t_4 .
 - A normal exagera a evidência ao tratar os extremos como informativos sobre θ .

Dados 4: incerteza sobre ν

Seja $\nu \in \{1, \dots, 7\}$ com $\Pr(\nu) = 1/7$. Para cada ν , calculamos $p(x | \nu)$ por Riemann, como antes:

$$p(x | \nu) \approx h \sum_j \prod_i t_\nu(x_i - \theta_j) \phi(\theta_j), \quad \Pr(\nu | x) = \frac{p(x | \nu)}{\sum_{k=1}^7 p(x | k)}.$$

ν	$p(x \nu)$	$E(\theta x, \nu)$	DP	$\Pr(\theta > 0 x, \nu)$	$\frac{p(x \theta = 0, \nu)}{p(x \nu)}$	$\Pr(\nu x)$
1	3.268×10^{-24}	0.6007	0.2152	0.9972	0.0966	0.0009
2	1.383×10^{-22}	0.6319	0.2135	0.9985	0.0576	0.0388
3	4.096×10^{-22}	0.6550	0.2101	0.9991	0.0362	0.1149
4	6.245×10^{-22}	0.6724	0.2071	0.9994	0.0244	0.1752
5	7.487×10^{-22}	0.6860	0.2045	0.9996	0.0175	0.2101
6	8.087×10^{-22}	0.6969	0.2024	0.9997	0.0131	0.2269
7	8.309×10^{-22}	0.7058	0.2005	0.9998	0.0102	0.2331

Dados 4: média sobre modelos (BMA) e leitura

Posterior de θ integrando ν : $\pi(\theta | x) = \sum_{\nu} \Pr(\nu | x) \pi(\theta | x, \nu)$

$$E(\theta | x) = 0.6850, \quad V = 0.0422, \quad DP = 0.2055, \quad \Pr(\theta > 0 | x) = 0.9995, \quad E(\nu | x) = 5.17.$$

- $\Pr(\nu | x)$ **cresce com ν e se achata a partir de $\nu \approx 5$** . Cauchy ($\nu = 1$) é descartada e $\nu = 2$ é pouco provável. Os dados não distinguem $\nu = 5, 6, 7$.
- **Comparação com a normal** ($p(x | M_0) = 3.736 \times 10^{-22}$): $BF_{\nu, N}$ vai de 0.009 ($\nu = 1$) a 2.22 ($\nu = 7$). Com $\nu \geq 3$, a t é preferida, mas só moderadamente.
- $E(\theta | x, \nu)$ **cresce de 0.60 a 0.71 com ν** . Caudas mais pesadas descontam mais os valores 3.27 e 3.78. A BMA (0.685) fica entre a t_4 (0.672) e a normal (0.802).
- $H_0 : \theta = 0$ é **rejeitada em todos os ν** . A evidência cresce com ν , de 10:1 ($\nu = 1$) a 98:1 ($\nu = 7$). Integrando ν , $p(x | \theta = 0)/p(x) = 0.020$, ou seja, cerca de 50:1.

Dados 5 ($n = 30$): dados

-1.90	-1.80	-0.94	-0.78	-0.57	-0.49	-0.34	-0.30	-0.19	0.34
0.44	0.47	0.47	0.50	0.59	0.60	0.82	0.83	0.84	1.05
1.51	1.55	1.69	1.75	1.76	1.99	2.40	4.44	6.84	23.19

- $\bar{x} = 1.559$ e mediana = 0.595.
- Sob a normal, $E(\theta | x) = 1.508$ e $p(x | M_0) = 1.78 \times 10^{-137}$.
- Com $\nu \leq 3$, $p(x | \nu)$ é da ordem de 10^{-28} : o fator de Bayes contra a normal é da ordem de 10^{108} .
- Para evitar *underflow*, trabalhe em log com log-sum-exp.

Dados 5: incerteza sobre ν

Com $\Pr(\nu) = 1/7$ e Riemann em $[-2, 2]$ (10.000 bins):

ν	$p(x \nu)$	$E(\theta x, \nu)$	DP	$\Pr(\theta > 0 x, \nu)$	$\frac{p(x \theta = 0, \nu)}{p(x \nu)}$	$\Pr(\nu x)$
1	6.545×10^{-29}	0.5644	0.2373	0.9902	0.2572	0.3129
2	1.213×10^{-28}	0.5680	0.2326	0.9922	0.2216	0.5797
3	2.011×10^{-29}	0.5740	0.2266	0.9941	0.1811	0.0961
4	2.134×10^{-30}	0.5805	0.2219	0.9954	0.1489	0.0102
5	2.135×10^{-31}	0.5871	0.2182	0.9963	0.1240	0.0010
6	2.241×10^{-32}	0.5938	0.2152	0.9970	0.1043	0.0001
7	2.547×10^{-33}	0.6003	0.2128	0.9975	0.0886	0.00001

BMA: $E(\theta | x) = 0.568$, $DP = 0.233$, $\Pr(\theta > 0 | x) = 0.992$, $E(\nu | x) = 1.81$ e $p(x | \theta = 0)/p(x) = 0.228$.

- **A posteriori de ν se inverte em relação aos dados 4.** A massa fica em $\nu = 2$ (0.58) e $\nu = 1$ (0.31). A observação 23.19 só é plausível com caudas muito pesadas.
- $E(\theta \mid x, \nu) \approx 0.56\text{--}0.60$ **para todo ν** , próxima da mediana (0.595) e longe de $\bar{x}$ (1.56). Com ν pequeno, a t praticamente ignora 4.44, 6.84 e 23.19.
- **A normal é descartada.** Ela desloca θ para 1.51 e tem verossimilhança marginal 10^{108} vezes menor.
- $H_0 : \theta = 0$: a evidência contra é só moderada, de 4:1 ($\nu = 1$) a 11:1 ($\nu = 7$). Na BMA, fica em cerca de 4.4:1.
- **A grade $[-2, 2]$ é adequada.** Repetindo em $[-6, 6]$ com 30.000 bins, os resultados são idênticos.

Dados 5 com escala desconhecida: modelos

$$M_0 : x_i \mid \theta, \sigma^2 \stackrel{iid}{\sim} N(\theta, \sigma^2)$$

$$M_1 : x_i \mid \theta, \tau^2, \nu \stackrel{iid}{\sim} t_\nu(\theta, \tau^2), \quad \nu \in \{1, \dots, 7\}, \Pr(\nu) = 1/7$$

Prioris independentes: $\theta \sim N(0, 1)$ e $\sigma^2, \tau^2 \sim G(a, b)$ (forma, taxa).

- Moda $(a - 1)/b = 1$ e $DP = \sqrt{a}/b \approx 0.3$ dão $a = 13.03$; arredondamos para $G(13, 12)$, que tem moda 1, média 1.083 e $DP = 0.300$.
- A priori, $\Pr(\sigma^2 < 2.25) = 0.999$.

Integração de Riemann 2D, por ponto médio:

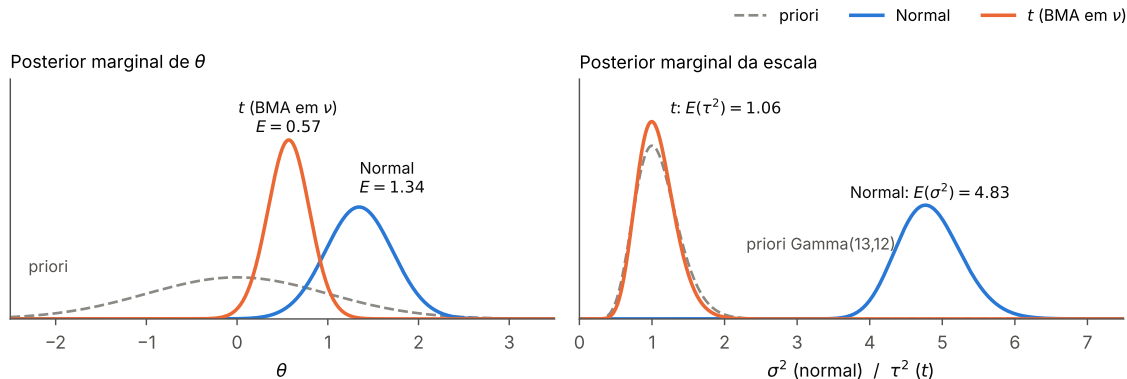
- $\theta \in [-4, 4]$ com 10.000 bins e $\sigma^2, \tau^2 \in (0, 10]$ com 2.000 bins, ou seja, 2×10^7 células por modelo.
- O intervalo $[-2, 2]$ não basta: sob M_0 , 3.9% da massa de θ fica fora dele.
- **Checagem:** sob M_0 , $x \mid \sigma^2 \sim N_n(0, \sigma^2 I_n + 11')$. O resultado semi-analítico (1D em σ^2) coincide com o 2D: $\log p(x \mid M_0) = -138.7426$.
- Grade grossa (2000×1000) e grade fina dão resultados idênticos.

Dados 5 com escala desconhecida: resultados

Modelo	$\log p(x \cdot)$	$E(\theta x)$	DP	$\Pr(\theta > 0 x)$	$E(\text{escala} x)$	DP	$\frac{p(x \theta = 0)}{p(x \cdot)}$	$\Pr(\nu x)$
Normal	-138.74	1.3427	0.3726	0.9998	4.8302	0.4469	0.0046	—
t_1	-65.01	0.5656	0.2335	0.9907	0.9562	0.2430	0.2405	0.3193
t_2	-64.46	0.5675	0.2375	0.9908	1.0903	0.2512	0.2472	0.5536
t_3	-66.08	0.5738	0.2384	0.9914	1.1938	0.2592	0.2342	0.1090
t_4	-68.03	0.5819	0.2396	0.9920	1.2796	0.2664	0.2200	0.0155
t_5	-69.98	0.5910	0.2412	0.9925	1.3540	0.2729	0.2064	0.0022
t_6	-71.84	0.6005	0.2430	0.9930	1.4204	0.2788	0.1939	0.0003
t_7	-73.59	0.6101	0.2449	0.9934	1.4808	0.2843	0.1826	0.00006
M_1 (BMA)	-65.81	0.5679	0.2364	0.9908	1.0624	0.2632	0.2431	—

$\log BF_{10} = -65.81 - (-138.74) = 72.93$, ou seja, $BF_{10} \approx 4.7 \times 10^{31}$. Além disso,
 $E(\nu | x) = 1.83$.

Dados 5 com escala desconhecida: posteriores marginais



Sob a normal, σ^2 precisa ser grande para acomodar 23.19, e isso alarga e desloca a posterior de θ . Sob a t , a escala fica na região da priori e a observação extrema é absorvida pelas caudas.

Dados 5: escala fixa vs. escala desconhecida

- **A normal melhora muito com σ^2 livre, mas continua sendo descartada.**
 - $\log p(x \mid M_0)$ sobe de -314.9 para -138.7 .
 - BF_{10} cai de $\sim 10^{108}$ para $\sim 10^{31}$.
- **Há conflito entre priori e dados sob M_0 .** $E(\sigma^2 \mid x) = 4.83$, mas a priori quase não tem massa acima de 2.25. A normal só explica 23.19 inflando a variância de todas as observações.
- **Sob M_1 quase nada muda.**
 - $\Pr(\nu \mid x)$: (0.32, 0.55, 0.11, ...) contra (0.31, 0.58, 0.10, ...) com $\tau^2 = 1$ fixo.
 - $E(\theta \mid x) = 0.568$ nos dois casos.
 - $E(\tau^2 \mid x, \nu)$ cresce com ν : caudas mais leves exigem escala maior.
- **$H_0 : \theta = 0$ depende da escala.**
 - Normal com $\sigma^2 = 1$: razão 2.7×10^{-15} , uma evidência absurda contra H_0 .
 - Normal com σ^2 livre: 0.0046, cerca de 217:1.
 - t (BMA): 0.243, cerca de 4:1, como antes.

```
x <- c(0.18,2.21,1.58,1.82,-0.65,1.29,0.12,2.08,-0.19,-1.84)
n <- length(x); nu <- 4
# M0: analitico
m0 <- n*mean(x)/(n+1); v0 <- 1/(n+1); p0 <- 1-pnorm(0,m0,sqrt(v0))
py0 <- (2*pi)^(-n/2)*(n+1)^(-1/2)*exp(-0.5*(sum(x^2)-sum(x)^2/(n+1)))
# M1: Riemann (ponto medio) em [-2,2]
N <- 10000; h <- 4/N; th <- -2 + h*((1:N)-0.5)
g <- sapply(th, function(t) prod(dt(x,nu))) * dnorm(th)
py1 <- sum(g)*h; post <- g/py1
m1 <- sum(th*post)*h; v1 <- sum((th-m1)^2*post)*h
p1 <- sum(post[th>0])*h
# H0: theta = 0
l0N <- prod(dnorm(x)); l0T <- prod(dt(x,nu))
c(BF10=py1/py0, H0vsM0=l0N/py0, H0vsM1=l0T/py1)
```

Código (R): incerteza sobre ν

```
# x <- dados 4
N <- 10000; h <- 4/N; th <- -2 + h*((1:N)-0.5)
nus <- 1:7; post <- matrix(0, N, 7); py <- numeric(7)
for (k in nus) {
  lg <- sapply(th, function(t) sum(dt(x-t, k, log=TRUE))) + dnorm(th, log=TRUE)
  g <- exp(lg); py[k] <- sum(g)*h; post[,k] <- g/py[k]
}
pnu <- py/sum(py) # Pr(nu | x), priori uniforme
Eth <- colSums(th*post)*h # E(theta | x, nu)
pbma <- post %*% pnu # posterior de theta integrando nu
c(E=sum(th*pbma)*h, Ppos=sum(pbma[th>0])*h, Enu=sum(nus*pnu))
```

Código (R): Riemann 2D com escala desconhecida

```
# x <- dados 5;  priori: theta ~ N(0,1), s2 ~ Gamma(13, rate=12)
Nt <- 10000; ht <- 8/Nt;  th <- -4 + ht*((1:Nt)-0.5)
Ns <- 2000;  hs <- 10/Ns; s2 <- 0 + hs*((1:Ns)-0.5)
lse <- function(a) { m <- max(a); m + log(sum(exp(a-m))) } # log-sum-exp
logpost2d <- function(loglik) { # loglik(theta, s2) -> matriz Ns x Nt
  J <- outer(dgamma(s2,13,12,log=TRUE), dnorm(th,log=TRUE), "+") + loglik
  lpy <- lse(J) + log(hs*ht); P <- exp(J - lpy)
  list(lpy=lpy, pth=colSums(P)*hs, ps2=rowSums(P)*ht)
}
llN <- t(sapply(s2, function(v) sapply(th, function(t) sum(dnorm(x,t,sqrt(v),log=
  TRUE))))))
llT <- function(nu) t(sapply(s2, function(v) sapply(th, function(t)
  sum(dt((x-t)/sqrt(v), nu, log=TRUE)) - length(x)/2*log(v))))
M0 <- logpost2d(llN)
M1 <- lapply(1:7, function(nu) logpost2d(llT(nu))) # lento em R puro: vetorize!
lp <- sapply(M1, '[[' , "lpy"); pnu <- exp(lp - lse(lp))
```

Por que Monte Carlo? A maldição da dimensionalidade

Riemann com m pontos por dimensão exige m^d avaliações de $p(x | \vartheta) p(\vartheta)$.

Medimos aqui $\approx 3.7 \times 10^{-7}$ s por avaliação ($n = 30$, código vetorizado). Com $m = 100$:

d	avaliações	tempo
1	10^2	0.04 ms
2	10^4	4 ms (nossa grade: 2×10^7 pontos, 7.5 s por modelo)
3	10^6	0.4 s
5	10^{10}	1 hora
7	10^{14}	1.2 ano
10	10^{20}	1.2 milhão de anos
20	10^{40}	10^{26} anos (idade do universo: 1.4×10^{10})

- **Quase toda a grade é desperdiçada.** No nosso caso 2D, 99.9% da massa a posteriori está em 3.4% das células. Se esse padrão se repetir por dimensão, a fração útil é $\approx 0.18^d$: em $d = 10$, uma célula a cada 25 milhões.
- **A grade exige saber onde está a posteriori.** Vimos isso com $[-2, 2]$, que não bastou sob M_0 .

Por que Monte Carlo? O caminho adiante

Ideia: se $\vartheta^{(1)}, \dots, \vartheta^{(M)} \sim \pi(\vartheta | x)$, então

$$E[g(\vartheta) | x] \approx \frac{1}{M} \sum_{m=1}^M g(\vartheta^{(m)}), \quad \text{erro} = O(M^{-1/2}) \text{ qualquer que seja } d.$$

Com N pontos no total, a regra do ponto médio tem erro $O(N^{-2/d})$. Monte Carlo passa a ser melhor já a partir de $d \approx 4$.

- **Os pontos vão para onde está a massa a posteriori**, não para uma grade cega.
- **MCMC** (Metropolis, Hastings, Gibbs) amostra de $\pi(\vartheta | x) \propto p(x | \vartheta)p(\vartheta)$ sem conhecer a constante $p(x)$, nem a forma fechada.
- **Liberta a modelagem:** prioris não conjugadas, caudas pesadas com ν desconhecido, hierarquias e variáveis latentes. Na volatilidade estocástica com $T = 1000$, por exemplo, $d \approx 1000$.
- **A modelagem passa a ser guiada pelo problema, não pela conta.**
 - Sem Monte Carlo: conjugação, aproximações normais ou problemas de dimensão 1 ou 2.
 - Com Monte Carlo: os modelos que os dados pedem.

Próximas aulas: Monte Carlo, amostragem por importância, MCMC e SMC.