

Um teste bayesiano para o coeficiente de correlação intraclass

Julio M Singer

em colaboração com
González-López, V.A., Tanaka, N.I. e Lima, A.C.P.

Departamento de Estatística
Universidade de São Paulo
www.ime.usp.br/~jmsinger

Avaliar o efeito de medidas realizadas em m -plicata na precisão de estimadores

- Introdução
- Apresentação de um exemplo envolvendo estimação de uma média
- Análise por intermédio de intervalos de confiança
- Especificação do problema estatístico
- O teste bayesiano
- Conclusão

Medidas em **triplicata** (m -plicata) comuns em muitos estudos de diversas áreas.

Fagan, T.C., Conrad, K.A., Mayshar, P.V., Mackie, M.J. and Hagaman, R.M. (1988). Single versus triplicate measurements of blood pressure and heart rate. **Hypertension**, **11**: 282-284.

Nutter, F.W. Jr, Gleason, M.L., Jenco, J.H. and Christians, N.C. (1993). Assessing the accuracy, intra-rater repeatability, and inter-rater reliability of disease assessment systems. **Phytopathology**, **83**: 806-812.

Paquet, J., Lacroix, C., Audet, P. and Thibault, J. (2000). Electrical conductivity as a tool for analysing fermentation processes for production of cheese starters. **International Dairy Journal**, **10**: 391-399.

Razões apresentadas: melhorar precisão, tradição

Estimação da quantidade de óleo em suco de limão

- 60 **amostras** de suco, cada uma dividida em três **alíquotas**
- Quantidade de óleo (kg óleo/ton suco) medida em cada alíquota (A, B e C) por meio de análise química
- Objetivo é estimar a quantidade média de óleo em suco de limão
- Precisão da estimativa importante para determinar possíveis usos (consumo, fabricação de cosméticos etc.)
- Dúvida: é preciso trabalhar com triplicatas?
- **Singer et al. (2007, Chemometrics and Intelligent Laboratory Systems)**

Quantidade de óleo no suco (kg óleo/ton suco)

Amostra	A	B	C	Amostra	A	B	C	Amostra	A	B	C
1	5.29	5.10	5.13	21	5.66	5.64	5.46	41	4.90	4.75	4.84
2	5.34	5.34	5.27	22	5.62	5.49	5.73	42	4.88	4.57	4.54
3	5.20	5.07	5.08	23	5.36	5.33	5.46	43	4.80	4.82	4.94
4	5.43	5.38	5.36	24	4.91	5.01	4.86	44	5.29	5.29	5.10
5	5.18	5.03	5.02	25	5.28	5.35	5.14	45	4.53	4.66	4.63
6	5.33	5.07	5.07	26	5.02	4.80	4.64	46	4.39	4.49	4.39
7	5.16	5.40	5.23	27	5.57	5.54	5.29	47	4.50	4.51	4.52
8	4.91	5.10	4.84	28	5.09	5.22	4.95	48	4.82	4.80	4.66
9	5.07	5.01	4.87	29	5.58	5.45	5.32	49	5.06	4.96	4.94
10	4.85	4.76	4.54	30	5.04	4.90	4.94	50	5.20	4.97	5.11
11	5.31	5.42	5.52	31	5.79	5.65	5.58	51	5.63	5.75	5.63
12	5.12	5.40	5.27	32	5.46	5.38	5.36	52	5.38	5.51	5.14
13	5.29	5.47	5.13	33	5.21	5.20	5.07	53	5.37	5.06	5.13
14	5.04	5.09	4.98	34	4.84	4.98	4.91	54	5.06	5.20	5.07
15	5.11	5.11	5.11	35	5.27	5.11	5.25	55	5.15	5.32	4.99
16	4.96	5.07	4.94	36	5.06	5.08	4.89	56	4.74	4.74	4.64
17	5.36	5.06	5.10	37	5.10	5.24	5.05	57	4.48	4.4	4.37
18	5.36	5.40	5.33	38	5.32	5.51	5.22	58	4.26	4.12	4.37
19	5.39	5.13	5.34	39	4.80	4.70	4.58	59	4.46	4.37	4.62
20	5.49	5.60	5.28	40	5.18	4.83	4.80	60	5.20	4.93	5.07

Intervalos de confiança (95%) para a quantidade média de óleo no suco de limão (kg/ton)

Dados utilizados	Limite inferior	Limite superior	Amplitude
Primeira alíquota	5.04	5.21	0.17
Segunda alíquota	5.00	5.18	0.18
Terceira alíquota	4.94	5.11	0.17
Média das 3 alíquotas	5.00	5.16	0.16

Definição operacional dos objetivos

- **Problema 1:** Amostra de tamanho n de distribuição Normal
 - Em que situação a utilização de m -plicas ($m \geq 2$) melhora a precisão na estimação da média
- **Problema 2:** Estimação da média de distribuição Normal com precisão fixada.
 - Custo de obtenção de uma unidade amostral = A
 - Custo de realização de cada medida = R
 - Qual dos seguintes planos experimentais tem menor custo
 - Obter n_s unidades amostrais independentes e realizar uma única medida em cada uma
 - Obter $n_c < n_s$ unidades amostrais independentes e realizar ($m \geq 2$) medidas em cada uma

Problema 1

- Modelo de efeitos aleatórios para dados gaussianos em m -plicata

$$y_{ij} = \mu + a_i + e_{ij},$$

com $a_i \sim N(0, \sigma_a^2)$ e $e_{ij} \sim N(0, \sigma_e^2)$, indep, $i = 1, \dots, n$ e $j = 1, \dots, m$.

- Dependência entre observações realizadas na mesma unidade amostral quantificada pelo **coeficiente de correlação intraclasse**

$$\rho = \sigma_a^2 / (\sigma_a^2 + \sigma_e^2)$$

- IC(95%) baseado nos dados de **uma observação** (ex: a primeira)

$$\bar{y}_{+1} \pm 1.96 \hat{\sigma} / \sqrt{n}$$

- $\bar{y}_{+1} = n^{-1} \sum_{i=1}^n y_{i1}$
- $\hat{\sigma}^2 = (n-1)^{-1} \sum_{i=1}^n (y_{i1} - \bar{y}_{+1})^2$ estimativa de $\sigma^2 = \sigma_a^2 + \sigma_e^2$

Problema 1

- IC(95%) baseado na **média** das m observações realizadas na mesma unidade amostral

$$\bar{y}_{++} \pm 1.96 \sqrt{\frac{\hat{\sigma}_a^2 + \hat{\sigma}_e^2/m}{n}},$$

- $\bar{y}_{++} = n^{-1} \sum_{i=1}^n \bar{y}_{i+}$
 - $\bar{y}_{i+} = m^{-1} \sum_{j=1}^m y_{ij}$,
 - $\hat{\sigma}_e^2 = [n(m-1)]^{-1} \sum_{i=1}^n \sum_{j=1}^m (y_{ij} - \bar{y}_{i+})^2$
 - $\hat{\sigma}_a^2 + \hat{\sigma}_e^2/m = (n-1)^{-1} \sum_{i=1}^n (\bar{y}_{i+} - \bar{y}_{++})^2$.
- Lembrando que $\rho = \sigma_a^2 / (\sigma_a^2 + \sigma_e^2)$ e $\sigma^2 = \sigma_a^2 + \sigma_e^2$, o comprimento do IC é

$$2 \times \frac{1.96}{\sqrt{n}} \sqrt{\hat{\sigma}_a^2 + \hat{\sigma}_e^2/m} = 2 \times 1.96 \frac{\hat{\sigma}}{\sqrt{n}} \sqrt{\hat{\rho} + (1 - \hat{\rho})/m}.$$

Problema 1

- Como $\rho < 1$, para $m \geq 2$, o comprimento do IC (média 3 obs) é sempre menor que o comprimento do IC (uma obs) pelo fator

$$\sqrt{\hat{\rho} + (1 - \hat{\rho})/m}$$

- Dada a possibilidade de fazer observações em m -plicata, o problema é saber para que valores de ρ o comprimento do IC (média 3 obs) fica reduzido a $100(1 - r)\%$, ($0 < r < 1$) ou menos do comprimento do IC (uma obs).
- Basta fazer $\sqrt{\rho + (1 - \rho)/m} = (1 - r)$, e observar que $\sqrt{\rho + (1 - \rho)/m}$ é uma função crescente de ρ para se concluir que o resultado desejado vale quando

$$\rho < [m(1 - r)^2 - 1]/(m - 1).$$

Problema 1

- A máxima redução do IC (uma obs) corresponde a $r = 1 - \sqrt{1/m}$ e ocorre quando $\rho = 0$
- Para o exemplo ($m = 3$), a máxima redução é $r = 1 - \sqrt{1/3} = 0,42$

Redução do IC(uma obs)

ρ	0,05	0,10	0,25	0,50	0,75	0,90
Redução	39%	37%	29%	18%	8%	3%

- Como $\hat{\sigma}_a^2 = 0,0992$, $\hat{\sigma}_e^2 = 0,0131$ e $\hat{\rho} = 0,8830$ o que implica que a redução no comprimento do IC é de $100r\% = 4\%$

Problema 2

- Tamanho amostral necessário para comprimento do IC(95%) baseado em uma obs/unit amost seja igual a d é

$$n_s = \left[2 \times 1.96 \sqrt{\hat{\sigma}^2/d} \right]^2$$

- Custo correspondente é $C_s = n_s (A + R)$
- Como desejamos um IC com mesmo comprimento sob ambos os planos experimentais

$$2 \times 1.96 \sqrt{\frac{\hat{\sigma}_a^2 + \hat{\sigma}_e^2/m}{n_c}} = d.$$

- Então

$$\sqrt{\frac{\hat{\sigma}^2}{n_s}} = \sqrt{\frac{\hat{\sigma}_a^2 + \hat{\sigma}_e^2/m}{n_c}} \implies \sqrt{n_c/n_s} = \sqrt{\hat{\rho} + (1 - \hat{\rho})/m}$$

Problema 2

- Como $n_c \leq n_s$
 - Obter o valor de m para $n_c = 2, 3, \dots, n_s - 1$ sob a restrição $m \leq n_s(1 - \rho)/(n_c - \rho n_s)$
 - Calcular o custo $C_c = n_c(A + mR)$
- A solução é o menor valor entre C_c e C_s
- Planilha Excel para cálculos ([www.ime.usp.br/ jmsinger](http://www.ime.usp.br/jmsinger))

- Para ambos os problemas é preciso saber se $\rho \leq c$, c constante
- Teste exato sob normalidade (**Scheffé, 1959, The Analysis of Variance**)
- Conclusões baseadas em amostras piloto, geralmente pequenas, distribuição normal, poder baixo
- Utilização de testes bayesianos: incorporação de informação sobre componentes de variância

O teste bayesiano

- Modelo de efeitos aleatórios para dados gaussianos em m -plicata

$$y_{ij} = \mu + a_i + e_{ij},$$

$a_i \sim N(0, \sigma_a^2)$ e $e_{ij} \sim N(0, \sigma_e^2)$, indep, $i = 1, \dots, n$, $j = 1, \dots, m$.

- Dados: $\mathbf{y} = (y_{11}, \dots, y_{1m}, \dots, y_{n1}, \dots, y_{nm})$
- Parâmetros do modelo: $\Psi = (\mu, \sigma_a^2, \sigma_e^2)$
- Parâmetro de interesse: $\rho = \sigma_a^2 / (\sigma_a^2 + \sigma_e^2)$
- Procedimento baseado em σ_a^2, σ_e^2
 - Parâmetros do modelo
 - Simplificação computacional
- $\Theta_0 = \left\{ (\sigma_a^2, \sigma_e^2) \in \mathbb{R}_+^2 : \sigma_a^2 / (\sigma_a^2 + \sigma_e^2) \leq c \right\}$
- $I_{\Theta_0}(\sigma_a^2, \sigma_e^2) = \begin{cases} 1 & \text{se } (\sigma_a^2, \sigma_e^2) \in \Theta_0, \\ 0 & \text{se } (\sigma_a^2, \sigma_e^2) \notin \Theta_0 \end{cases}$
- $E\{I_{\Theta_0} \mid \mathbf{y}\}$: **probabilidade a posteriori de $\rho \leq c$.**

- Verossimilhança expressa em termos dos estimadores MV:

$$L(\Psi | \mathbf{y}) = \frac{\exp\left(-\frac{1}{2}\left[\frac{n(m-1)\hat{\sigma}_e^2}{\sigma_e^2} + \frac{(n-1)(\hat{\sigma}_e^2 + m\hat{\sigma}_a^2)}{\sigma_e^2 + m\sigma_a^2} + \frac{nm(\hat{\mu} - \mu)^2}{\sigma_e^2 + m\sigma_a^2}\right]\right)}{(\sigma_e^2)^{n(m-1)/2}(\sigma_e^2 + m\sigma_a^2)^{n/2}}$$

$$\hat{\mu} = \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m y_{ij}$$

$$\hat{\sigma}_e^2 = \frac{1}{n(m-1)} \sum_{i=1}^n \sum_{j=1}^m (y_{ij} - \bar{y}_i)^2$$

$$\hat{\sigma}_a^2 = \frac{1}{n-1} \sum_{i=1}^n (\bar{y}_i - \hat{\mu})^2 - \frac{\hat{\sigma}_e^2}{m},$$

- Distribuição a posteriori conjunta de $\Psi = (\mu, \sigma_a^2, \sigma_e^2)$:

$$g(\Psi | \mathbf{y}) = L(\Psi | \mathbf{y}) f_{\mu}(\mu) f_{\sigma_a^2}(\sigma_a^2) f_{\sigma_e^2}(\sigma_e^2).$$

- Como interesse recai em ρ , distribuição a posteriori marginal de (σ_a^2, σ_e^2) proporcional a

$$h(\sigma_a^2, \sigma_e^2 | \mathbf{y}) = \int g(\mu, \sigma_a^2, \sigma_e^2 | \mathbf{y}) d\mu$$

- Probabilidade a posteriori de $H_0: \rho \leq c$

$$\begin{aligned} E\{I_{\Theta_0} | \mathbf{y}\} &= C^{-1} \int_0^\infty \int_0^\infty h(\sigma_a^2, \sigma_e^2 | \mathbf{y}) I_{\Theta_0}(\sigma_a^2, \sigma_e^2) d\sigma_a^2 d\sigma_e^2 \\ &= C^{-1} \int_0^\infty \int_0^\infty \int_{-\infty}^\infty g(\Psi | \mathbf{y}) I_{\Theta_0}(\sigma_a^2, \sigma_e^2) d\mu d\sigma_a^2 d\sigma_e^2 \end{aligned}$$

- $C = \int_0^\infty \int_0^\infty \int_{-\infty}^\infty g(\Psi | \mathbf{y}) d\mu d\sigma_a^2 d\sigma_e^2$

Distribuições envolvidas

- $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)$, $\theta_i = \mu + a_i$, $i = 1, \dots, n$,
- Distribuições condicionais
 - $[\theta_i | \mu, \sigma_a^2] = \mathcal{N}(\mu, \sigma_a^2)$
 - $[e_{ij} | \sigma_e^2] = \mathcal{N}(0, \sigma_e^2)$
 - $[y_{ij} | \theta_i, \sigma_e^2] = \mathcal{N}(\theta_i, \sigma_e^2)$
- Distribuições a priori
 - Não informativa para μ : $\mu \sim \mathcal{N}(0, 1e + 06)$
 - Gaussianas inversas para σ_a^2 e σ_e^2 : $\sigma_a^2 \sim \text{IG}(\beta, 1)$, $\sigma_e^2 \sim \text{IG}(\alpha, 1)$
 - Consequentemente, $\rho \sim \text{Beta}(\alpha, \beta)$
 - $\alpha = \beta = 1$ implica distribuição uniforme para ρ
- Distribuição conjunta de $(\mathbf{y}, \mu, \sigma_a^2, \sigma_e^2)$ não tem expressão analítica
- Distribuições a priori de $\mu, \sigma_a^2, \sigma_e^2$ condicionalmente conjugadas
- Amostrador de Gibbs

- Consideremos variáveis aleatórias X, Y
- Gerar amostra da distribuição marginal $f(x)$
- Se $f(x, y)$ tem expressão analítica, utilizar $f(x) = f(x, y)/f(y|x)$
- Em caso, contrário, utilizar $f(x|y)$ e $f(y|x)$ quando disponíveis
 - Iniciar com $Y_0 = y_0$
 - Gerar $X_0 = x_0$ a partir de $f(x|Y_0 = y_0)$ e $Y_1 = y_1$ a partir de $f(y|X_0 = x_0)$
 - Repetir o procedimento para gerar valores de X_k e Y_k , $k = 1, 2, \dots$
 - Sob condições de regularidade, densidade de X_k converge para $f(x)$ com $k \rightarrow \infty$.
 - Na prática, para garantir que amostra foi gerada após convergência: gerar e.g. 15000 valores e desprezar os primeiros e.g. 5000

Implementação do amostrador de Gibbs

- Distribuições condicionais necessárias para o amostrador de Gibbs

$$[\boldsymbol{\theta} | \mathbf{y}, \mu, \sigma_a^2, \sigma_e^2] = \mathcal{N}_n \left(\frac{m\sigma_a^2}{m\sigma_a^2 + \sigma_e^2} \bar{\mathbf{y}} + \frac{\sigma_e^2}{m\sigma_a^2 + \sigma_e^2} \mu \mathbf{1}_n, \frac{\sigma_a^2 \sigma_e^2}{m\sigma_a^2 + \sigma_e^2} \mathbf{I}_n \right)$$

$$[\mu | \mathbf{y}, \boldsymbol{\theta}, \sigma_a^2, \sigma_e^2] = [\mu | \sigma_a^2, \boldsymbol{\theta}] = \mathcal{N} \left(\frac{n(1e + 06)}{n(1e + 06) + \sigma_a^2} \bar{\theta}, \frac{n(1e + 06)}{n(1e + 06) + \sigma_a^2} \frac{\sigma_a^2}{n} \right)$$

$$[\sigma_e^2 | \mathbf{y}, \mu, \boldsymbol{\theta}, \sigma_a^2] = [\sigma_e^2 | \mathbf{y}, \boldsymbol{\theta}] = \text{IG} \left(\alpha + \frac{nm}{2}, 1 + \frac{\sum_{j=1}^m \sum_{i=1}^n (y_{ij} - \theta_i)^2}{2} \right)$$

$$[\sigma_a^2 | \mathbf{y}, \mu, \boldsymbol{\theta}, \sigma_e^2] = [\sigma_a^2 | \mu, \boldsymbol{\theta}] = \text{IG} \left(\beta + \frac{n}{2}, 1 + \frac{\sum_{i=1}^n (\theta_i - \mu)^2}{2} \right)$$

Implementação do amostrador de Gibbs

- *Input*

- Parâmetros α e β definidos a partir de informações sobre σ_a^2 e σ_e^2 ou ρ
- Número de unidades amostrais, n , de réplicas, m , e constante c
- Opcionalmente, hiperparâmetros da distribuição a priori de μ

Output

- Gráfico da densidade a priori para ρ
 - Histograma da distribuição a posteriori empírica de ρ
 - Histograma alisado para distribuição a posteriori empírica de ρ
 - Probabilidade de $\rho \leq c$
-
- Comandos disponíveis em <http://www.ime.unicamp.br/~veronica/>

Tabela: Probabilidade a posteriori de $\rho \leq c$ com priori uniforme

Redução no comprimento do IC (95%)	c	$E\{I_{\Theta_0} \mathbf{y}\}$
29%	0.26	<0.001
18%	0.51	0.036
10%	0.72	0.504
7%	0.80	0.790
5%	0.85	0.922
3%	0.91	0.984

Tabela: Probabilidade a posteriori de $\rho \leq c$ com prioris informativas

Redução no comprimento do IC (95%)	c	$E\{I_{\Theta_0} \mathbf{y}\}$		
		Beta(2,10)	Beta(2,2)	Beta(10,2)
29%	0.26	0.006	<0.001	<0.001
18%	0.51	0.572	0.047	0.001
10%	0.72	0.992	0.587	0.200
7%	0.80	1.000	0.800	0.553
5%	0.85	1.000	0.959	0.806
3%	0.91	1.000	0.995	0.962

Figura: Prioris Beta(1,1) e Beta(2,2), posterioris empíricas e alisadas

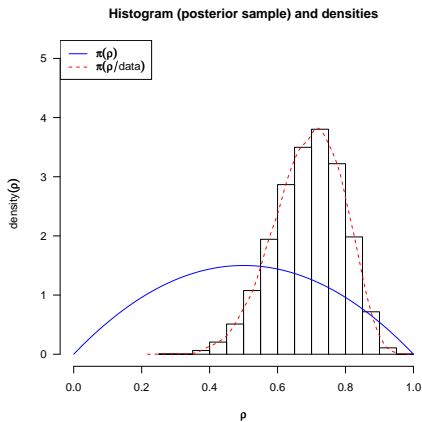
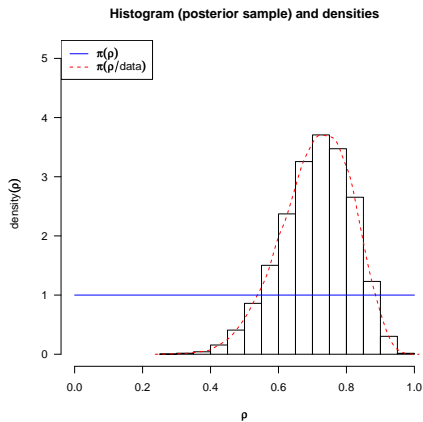
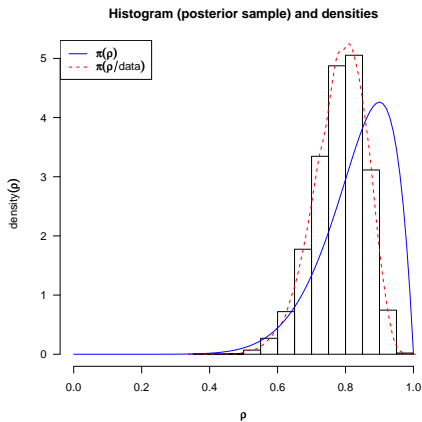
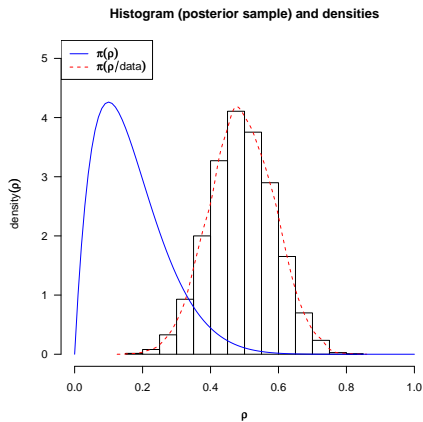


Figura: Prioris Beta(2,10) e Beta(10,2), posterioris empíricas e alisadas



Aprovado pelo chefe

