Universidade de São Paulo
Instituto de Ciências Matemáticas e de Computação

Estatística Computacional

Professor: Mário de Castro

Anotações de aula.

1º Semestre - 2017

Universidade de São Paulo
Instituto de Ciências Matemáticas e de Computação

Estatística Computacional

Professor: Mário de Castro

Digitado e revisado por Matheus Hisatugu, quaisquer dúvidas, sugestões, correções ou comentários: matheus.hisatugu@usp.br

1º Semestre - 2017
Ultima atualização em 7 de janeiro de 2020

Sumário

1 Simulação estocástica e números pseudoaleatórios

Simulação estocástica é a *arte* de gerar amostras de variáveis aleatórias em um ambiente computacional e usar essas amostras para a obtenção de um certo resultado.

O bom uso da simulação estocástica fornece resultados *aproximados*.

Geração de variáveis aleatórias:

- Variáveis aleatórias *iid* Uniforme(0,1);
- Procedimentos para geração de números **pseudoaleatórios**. As sequências geradas, embora sejam determinísticas, devem ter a "aparência" de aleatoriedade;
- Testes de geradores de números aleatórios;
- Período da sequência, precisão, repetibilidade, portabilidade.

Alguns geradores:

Gerador 1: Gerador de von Newmann (1949)

1. X_0 : número de quatro algarismos decimais (semente). Faça $i = 0$;
2. Calcular x_i^2. Se necessário acrescentar zeros à esquerda para que x_i^2 seja escrito como $d_7d_6d_5d_4d_3d_2d_1d_0$, em que $d_i \in \{0, 1, \ldots, 9\}$ para $i = 0, 1, \ldots, 7$;
3. Fazer $x_{i+1} = d_5d_4d_3d_2$ (meio do quadrado);
4. Faça $i = i + 1$ e retorne ao passo 2;
5. Divida os elementos por 10 000.

Exemplos:
a) {2 100, 4 100, 8 100, 6 100, 2 100,...}.
b) {3 792, 3 792,...}.

Gerador 2: Geradores congruenciais lineares (Lemuer, 1951)

Gerar uma sequência de inteiros $x_1, x_2, \ldots, x_n$ em $\{0, 1, \ldots, M-1\}$, M "grande". Fazer $u_i = x_i/M, i = 1, \ldots, n$,

$$x_i = (ax_{i-1} + c) mod M, \qquad i = 1, \ldots, n,$$

sendo que $a, c, x_0 \in \{0, 1, \ldots, M-1\}$, x_0 é chamado **semente** e *mod* representa o resto da divisão inteira.

Exemplos:
a) Gerador do IMSL: $a = 16\ 807$; $c = 0, M = 2^{31} - 1$, período: $2^{31} - 2 = 2\ 147\ 483\ 646$.
b) IBMRANDU: $a = 2^{16} + 3, c = 0, M = 2^{31}$.

Exemplo em R:

- Considere **todas** as matrizes 2×2 $\begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix}$, em que $a_1, a_2, a_3, a_4 \in \{0, 1, \ldots, 9\}$. Quantas são?
- Problema: Calcule o determinante de todas estas matrizes, conte o número de vezes que cada diferente valor ocorre (frequência) e apresente o resultado em um gráfico.

```
#Venables & Smith (1992, notes on S-Plus, p. 19)
prod2 <- outer(0:9, 0:9)
frdet <- table(outer(prod2, prod2, FUN="-"))
plot(frdet, xlab = "Determinante", ylab = "Frequencia", col = "blue")
```

- N: número de alunos, N é par, existem $\frac{N(N-1)}{2} = \binom{N}{2}$ pares.

Obs.: Função *sample*: selecionar amostras, sempre diferentes, exemplo: $sample(N, 2)$.

```
N <- 22
g1 <- sample(N, N/2, replace = F)
cbind(g1, (1:N)[-g1])
```

Recomenda-se usar semente fixa, devido a garantia de verificação dos dados.
Obs.: $set.seed(n)$ fixa o número n como semente.

2 Geração de amostras aleatórias

A geração de amostras aleatórias (**a.a.**) de uma distribuição Uniforme(0,1) está implementada em diversas linguagens e pacotes. Em R, utiliza-se a função $runif(a, b)$ (r de *random*).

2.1 Método de inversão (ou método da transformação)

Definição: A função de distribuição acumulada (f.d.a) de uma v.a. X é definida como:

$$F(x) = P(X \leq x), \text{ para } x \in \mathbb{R}$$

Note que $F(x) \in [0, 1]$

Definição: A função inversa generalizada de $F(x)$ é definida como:

$$F^-(x) = \min\{t \in \mathbb{R} : F(t) \geq x\}, \text{ para } x \in [0, 1]$$

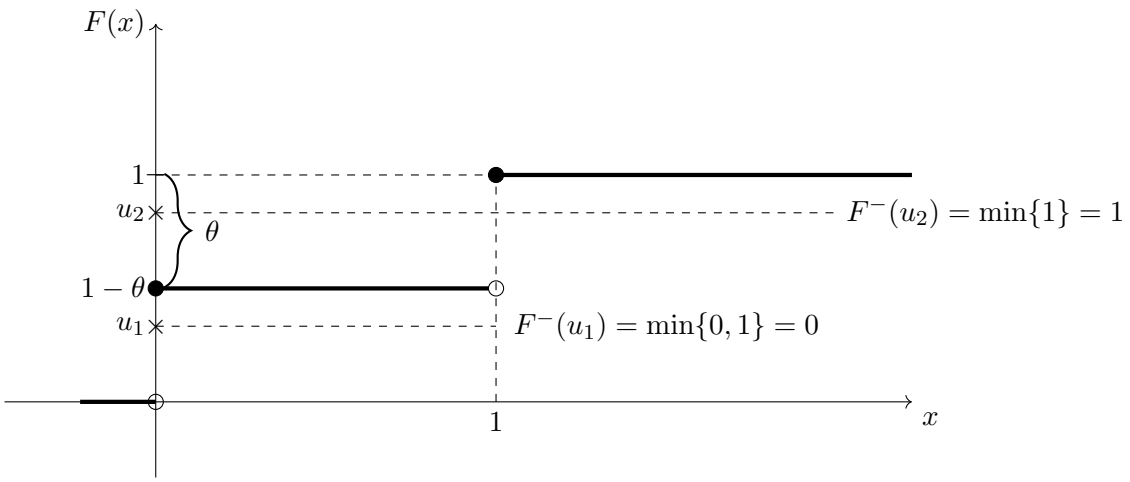
Obs.: De fato, temos inf no lugar de min.

Exemplo: $X \sim Bernoulli(\theta), \theta \in (0, 1)$.

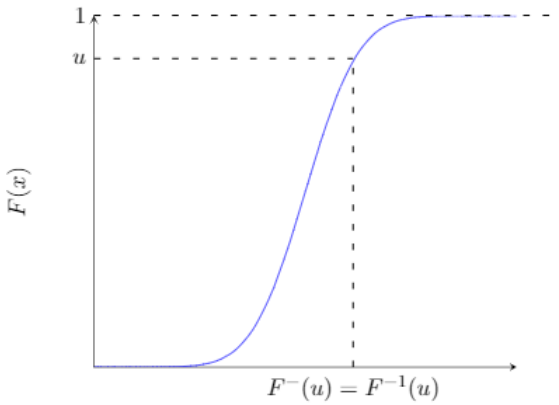
Obs.: $P(X = 0) = 1 - \theta, P(X = 1) = \theta$.

Neste caso,

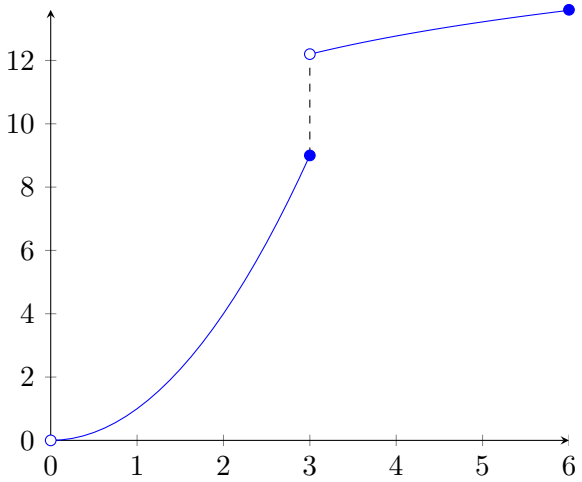
$$F(x) = \begin{cases} 0, & \text{se } x < 0, \\ 1 - \theta, & \text{se } 0 \leq x < 1, \\ 1, & \text{se } x \geq 1. \end{cases}$$



Exemplo: $F(x)$ é contínua e estritamente crescente. Neste caso, $F^{-}(x) = F^{-1}(x)$. **Obs.:** Função descontínua não tem inversa.



Exemplo: Função com ponto descontínuo.



Resultado: Se $U_1, \ldots, U_n$ é uma a.a. da distribuição $U(0,1)$, então $X_1 = F^{-}(u_1), \ldots, X_n = F^{-}(u_n)$ é uma a.a. da v.a. X cuja f.d.a é $F(x)$.

Exemplo: $X \sim Bernoulli(\theta)$

$$F^{-}(u) = \begin{cases} 0, & \text{se } 0 \leq u < 1-\theta, \\ 1, & \text{se } 1-\theta < u \leq 1. \end{cases}$$

Tomar $\theta = 0,6$ e $n = 4$, portanto:

$$F^{-}(u) = \begin{cases} 0, & \text{se } 0 \leq u < 0,4, \\ 1, & \text{se } 0,4 < u \leq 1. \end{cases}$$

u	0,7422	0,6596	0,2052	0,4689
$F(u)$	1	1	0	1

Exemplo: $X \sim Weibull(\alpha, \beta), \alpha > 0, \beta > 0$. A f.d.a de X é dada por:

$$F(x) = \begin{cases} 1 - \exp\left\{ -\left(\frac{x}{\alpha}\right)^{\beta} \right\}, & \text{se } x > 0, \\ 0, & \text{se } x \leq 0. \end{cases}$$

Se $\beta = 1$, então $X \sim Exponencial(\alpha)$. Precisamos obter $F^{-}(u)$, para $u \in (0,1)$. Temos então que para $u \in (0,1)$

$$F^{-}(u) = \alpha\{-\log(1-u)\}^{\frac{1}{\beta}},$$

Se $U \sim U(0,1)$, então $1 - U \sim U(0,1)$

Obs.: ter a mesma distribuição não implica em ter os mesmos valores, apenas as mesmas probabilidades.

$$X_1 \stackrel{d}{=} X_2 \qquad P(X_1 \leq a) = P(X_2 \leq a), \forall a$$

Aplicando o método da inversão temos que: $X_1 = \alpha\{-\log(u_1)\}^{\frac{1}{\beta}}, \ldots, X_n = \alpha\{-\log(u_n)\}^{\frac{1}{\beta}}$ é uma a.a. da distribuição $Weibull(\alpha, \beta)$

Dificuldade do método: achar $F^{-}(u)$.

Tomar $\alpha = \beta = 1$, ou seja, $X \sim Exponencial(1)$. Tomar $n = 4$.

u	0,9513	0,5216	0,2025	0,4031
$F(u)$	0,0498	0,6508	1,597	0,9083

Em R: $rexp(4, rate = 1)$

Obs.: O método da inversão é um método geral, mas requer a expressão de $F^{-}(u)$, que nem sempre é possível de ser obtida. Por exemplo, se $X \sim N(\mu, \sigma^2)$. Em R: função $rnorm(n, \mu, \sigma)$ e $rweibull(n, \alpha, \beta)$, em que n é o número de valores que queremos gerar.

Exemplo:

x	0	1	2	3
$P(X = x)$	$\frac{1}{10}$	$\frac{4}{10}$	$\frac{2}{10}$	$\frac{3}{10}$

Escreva o código em R para gerar uma amostra de tamanho n desta distribuição.

2.2 Método da rejeição

Chamado de método de aceitação-rejeição. Proposto por von Newmann (1951).

Problema: Gerar uma a.a. de uma v.a. X contínua com função densidade $f(x)$ em que a função $F^{-}(x)$ não é conhecida ou não é simples de ser obtida.

Y é uma v.a. contínua com função densidade $g(y)$ e temos um método para gerar uma a.a. de Y.

Além disso, suponhamos que:

$$\frac{f(x)}{g(x)} \leq M, \forall x \text{ no domínio de } f \text{ e } g, \text{ em que } 1 \leq M < \infty,$$

Obs.: Também pode ser usado com v.a. discretas.

Sendo assim, $f(x) \leq Mg(x)$. Dizemos que $Mg(x)$ envelopa $f(x)$

Algoritmo da rejeição:

1. Gerar y (uma observação da v.a. Y);
2. Gerar $u \sim U(0,1)$;
3. Se $u \leq \dfrac{f(y)}{Mg(y)}$, faça $x = y$. Caso contrário, retorne ao passo 1;

4. Repetir os passos 1 a 3 até obter n observações.

Recomenda-se tomar $M = \max_x \frac{f(x)}{g(x)}$. Qualquer outro valor $M^* > M$ poderia ser usado.

Propriedade: O número de tentativas até obter um valor da distribuição de X tem distribuição geométrica com média igual a M. A probabilidade de aceitação é $\frac{1}{M}$.

Obs.: Para x no domínio de X pode ser provado que:

$$P\left(Y \leq y \middle| u \leq \frac{f(y)}{Mg(y)}\right) = F_X(x).$$

Exemplo: Gerar uma a.a. de $X \sim N(0,1)$ utilizando a distribuição de Laplace (exponencial dupla) padrão com função densidade:

$$g(y) = \frac{1}{2}\exp(-|y|), y \in \mathbb{R}.$$

Lembrar que $f(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\}, x \in \mathbb{R}$.

Notar que $g(y)$ é simétrica em relação a $y = 0$, ou seja, $g(y) = g(-y), \forall y \in \mathbb{R}$.

Obs.: $|y| = \begin{cases} -y, & \text{se } y \leq 0, \\ y, & \text{se } x > 0. \end{cases}$

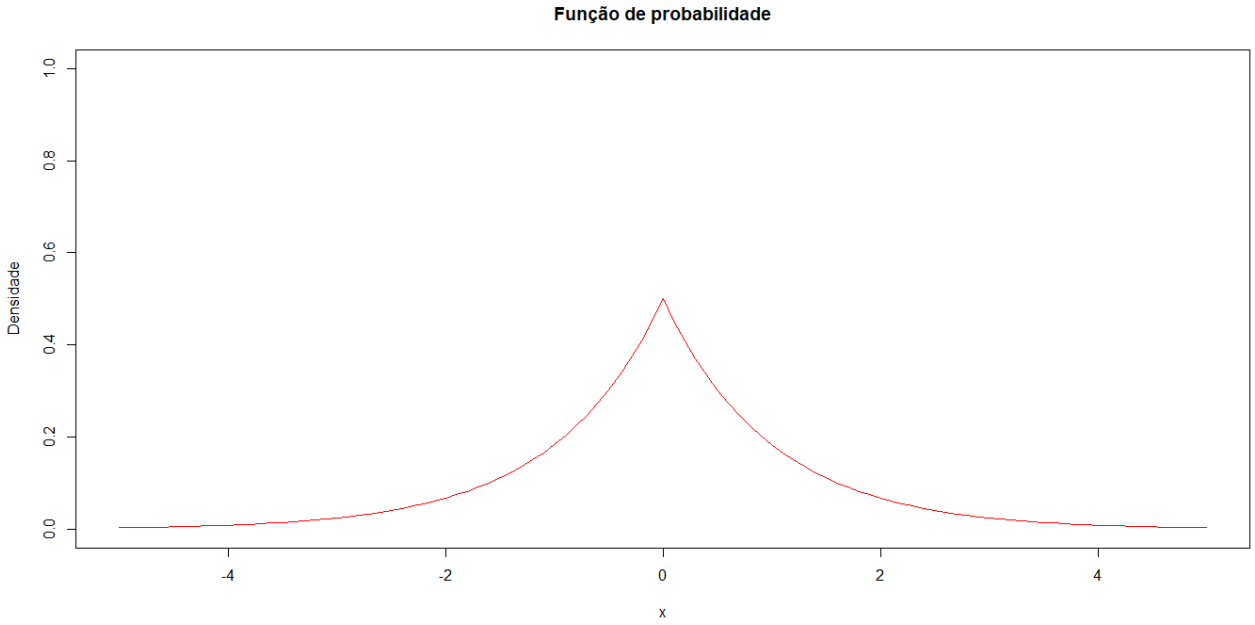


Figura 2.1: Função densidade de probabilidade da distribuição de Laplace(0,1).

A f.d.a de Y é dada por:

$$F_Y(y) = \begin{cases} \frac{1}{2}\exp\{y\}, & \text{se } y \leq 0, \\ 1 - \frac{\exp\{-y\}}{2}, & \text{se } y \geq 0. \end{cases}$$

$F_Y(y)$ é uma função contínua, monótona e crescente. Isso significa que existe $F_Y^{-1}(y)$, dada por:

$$F_Y^{-1}(y) = \begin{cases} \log(2u), & \text{se } 0 < u \leq \frac{1}{2}, \\ -\log(2(1-u)), & \text{se } \frac{1}{2} < u < 1. \end{cases}$$

sendo que:

$$\lim_{u\to 0} F_Y^{-1}(u) = -\infty \qquad \text{e} \qquad \lim_{u\to 1} F_Y^{-1}(u) = \infty.$$

Calculamos $h(x) = \dfrac{f(x)}{g(x)} = \dfrac{\frac{1}{\sqrt{2\pi}}\exp\{-\frac{x^2}{2}\}}{\frac{1}{2}\exp\{-|x|\}} = \sqrt{\dfrac{2}{\pi}}\exp\left\{\dfrac{-x^2}{2} + |x|\right\}$ para $x \in \mathbb{R}$, cujo alor máximo é atingido quando $x \pm 1$.

Portanto, $M = h(-1) = h(1) = \sqrt{\dfrac{2e}{\pi}} = 1,3154$.

Obs.: No método de rejeição, a distribuição da variável auxiliar Y condicionada no evento aceitação (do valor gerado de Y) coincide com a distribuição de X. Em R existem diversas funções para a geração de a.a., todas com a primeira letra "r" (random).

Exemplo: $rbinom(n, x, p), rexp(n, \lambda), rnorm(n, \mu, \sigma), rwilcox(nn, m, n)$. **Obs.:** Verificar a parametrização utilizada.

3 Solução de problemas usando simulações

Em Estatística, simulações de Monte Carlo são bastante usadas.

3.1 Cálculo de esperança de v.a.

X é uma v.a. com função densidade $f(x)$, sendo que;

$$f(x) = P(X = x), \text{ se } X \text{ for discreta.}$$

X assume valores no conjunto A, que pode ser o conjunto $\mathbb{R}$. A função g está definida no conjunto A. Supomos ainda que $\mathbb{E}[g(X)] \in \mathbb{R}$.

Problema: Calcular $\theta = \mathbb{E}[g(X)]$.

Temos então que se X é uma v.a. contínua,

$$\mathbb{E}[g(X)] = \int_A g(x)f(x)dx,$$

se X é uma v.a. discreta,

$$\mathbb{E}[g(X)] = \sum_{x\in A} g(x)f(x) = \sum_{x\in A} g(x)P(X = x),$$

$X_1, \ldots, X_R$ é uma a.a. de X. Logo,

$$\hat{\theta} = \frac{1}{R}\sum_{j=1}^{R} g(X_j) \to \theta = \mathbb{E}[g(X)], \text{ quando } R \to \infty.$$

O resultado se justifica pela lei forte dos grandes números. Como $\hat{\theta}$ é uma média amostral, temos que:

$$Var(\hat{\theta}) = \frac{Var(g(X))}{R} = \frac{\sigma^2}{R},$$

Supondo que $0 < \sigma^2 < \infty$, um estimador de $Var(\hat{\theta})$ é dado por:

$$\widehat{Var}(\hat{\theta}) = \frac{1}{R}\frac{1}{R-1}\sum_{j=1}^{R}\{g(X_j) - \hat{\theta}\}^2.$$

O erro padrão do estimador $\hat{\theta}$ é estimado por:

$$ep(\hat{\theta}) = \sqrt{\widehat{Var}(\hat{\theta})} \qquad \text{(erro padrão de Monte Carlo).}$$

Pelo teorema central do limite:

$$\sqrt{R}(\hat{\theta} - \theta) \xrightarrow{D} N(0, \sigma^2), \text{ quando } R \to \infty.$$

Um intervalo de confiança de 95% aproximado para θ é dado por:

$$\hat{\theta} \pm 1,96 ep(\hat{\theta}).$$

Obs.: Se $X_1, \ldots, X_n$ é uma a.a. de X, então $g(X_1), \ldots, g(X_n)$ é uma a.a. de $g(X)$.

Um caso particular de interesse ocorre quando $g(X) = \mathbb{I}(X \in B)$, em que $B \subset A$ é um conjunto e

$$\mathbb{I}(X \in B) = \begin{cases} 0, & \text{se } X \notin B, \\ 1, & \text{se } X \in B. \end{cases} \qquad (g(X) \sim Bernoulli)$$

Neste caso,

$$\theta = \mathbb{E}[g(X)] = 0P(X \notin B) + 1P(X \in B) = P(X \in B).$$

A distribuição de $g(X)$ é $Bernoulli(\theta)$ com variância $\sigma^2 = \theta(1-\theta)$, de modo que:

$$ep = \frac{\hat{\theta}(1-\hat{\theta})}{R}.$$

Notar que $\hat{\theta}$ é a proporção amostral de "sucessos" $(X \in B)$.

Obs.: X pode ser um vetor.

Exemplo: Eleição com dois candidatos C_1 e C_2.

Em uma urna temos n_1 votos para C_1 e n_2 votos para C_2, com $n_1 > n_2$ e sem votos nulos ou em branco. A apuração é feita sorteando cédula na urna, uma a uma. Para cada cédula retirada, são atualizados os totais de votos de C_1 e C_2.

Calcule a probabilidade de ocorrer empate em pelo menos uma das atualizações.

Exemplos: $n_1 = 3$; $\quad n_2 = 2$

Sequência de apuração: $(C_1, C_2, C_1, C_2, C_1)$

	Sorteio				
	1	2	3	4	5
C_1	1	1	2	2	3
C_2	0	1	1	2	2

(Em R, função any: qualquer um que satisfaça a condição)

Sequência: $(C_1, C_1, C_1, C_2, C_2) \to$ não ocorrem empates.

Resposta: $\dfrac{2n_2}{n_1 + n_2}$.

Uma solução por enumeração requer analisar todas as sequências de tamanho $n_1 + n_2$ com n_1 símbolos c_1 e n_2 símbolos c_2. O número de sequências é $\binom{n_1+n_2}{n_2} = \binom{n_1+n_2}{n_1}$.

Se $n_1 = 30$ e $n_2 = 20$, o número de sequências é da ordem de 10^{13}. Em R, $choose(50, 20) = choose(50, 30)$. Sugestão: representar $C_1 = 1$ e $C_2 = -1$ (resultado 0: houve empate).

Conjunto de N observações

Distribuição exata da variância amostral em amostras de $n < N$ observações, com reposição. Número de amostras: N^n.

$$\mathbb{E}[g(X)] = \sum_{x \in A} g(X)P(X = x) = \frac{1}{k}\sum_{x \in A} g(X) \text{ (indicador de sucesso, } g(X) = \mathbb{I}(X \in B)).$$

B = conjunto de todas as sequências em que ocorre pelo menos um empate na apuração.

No exemplo, obtivemos uma aproximação de Monte Carlo para $\mathbb{E}[g(X)]$ (ex.: apuração dos votos dos candidatos).

3.2 Cálculo de integrais

Calcular $\theta = \int_A g(x)dx$, em que $g(\cdot)$ é uma função definida no conjunto A com valores reais.

Exemplo: Calcular $\int_{-\infty}^{\infty} \arcsin(x^2)dx$.

Considere X uma v.a. com vvalores no conjunto A e com função densidade $f(x) > 0, \forall x \in A$. Escrevemos:

$$\theta = \int_A \frac{g(x)}{f(x)} f(x)dx = \int_A g^*(x)f(x)dx = \mathbb{E}[g^*(x)],$$

supondo que $\mathbb{E}[g^*(x)] \in \mathbb{R}$. Os resultados da Seção 3.1 podem ser apllicados.

O Método de Monte Carlo pode ser vantajoso no cálculo aproximado de integrais múltiplas. Calcular:

$$\theta = \int_0^1 \int_0^1 \dots \int_0^1 g(x_1, x_2, \dots, x_n)d_{x_1}d_{x_2}\dots d_{x_n}.$$

O número de simulações é R.

Note que $g^* = g$, pois $f(x_1, \dots, x_k) = 1(uniforme)$.

Geramos R conjuntos de k observações de variáveis $U(0, 1)$, todas independentes e iid:

$$\begin{cases} u_{11} & u_{12} & \dots & u_{1k} & \longrightarrow g(u_{11}, \dots, u_{1k}) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ u_{R1} & u_{R2} & \dots & u_{Rk} & \longrightarrow g(u_{R1}, \dots, u_{Rk}) \end{cases}$$

A estimativa é dada por:

$$\hat{\theta} = \frac{1}{R}\sum_{j=1}^{R} g(u_{j1}, \dots, u_{jk}).$$

O erro padrão de Monte Carlo é calculado da mesma forma que na Seção 3.1.

3.3 Avaliação de propriedades de estimadores e estatísticas de teste

Geralmente, métodos de estimação permitem obter estimadores com propriedades assintóticas ($n \to \infty$). Por exemplo, método dos momentos e método de máxima verossimilhança.

Exemplo: Problema com um parâmetro θ. X é uma v.a. cuja distribuição depende de θ. $X_1, \dots, X_n$ é uma a.a. da distribuição de X. Em condições bastante gerais, o estimador de máxima verossimilhança de θ (ou seja,$\hat{\theta}$) é consistente e assintóticamente eficiente com distribuição normal, ou seja,

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{D} N(0, \frac{1}{I_F(\theta}), \text{ quando } n \to \infty,$$

em que $I_F(\theta)$ é a informação de Fisher de uma observação.

De forma **aproximada**, $\hat{\theta} \sim N\left(\theta, \frac{1}{nI_F(\theta)}\right)$, em que θ é o verdadeiro valor do parâmetro.

Como o estimador é consistente,

$$\lim_{n\to\infty} P(|\hat{\theta} - \theta| > \varepsilon) = 0, \forall \varepsilon > 0.$$

Exemplo: $X \sim Poisson(\theta)$ sendo que $\mathbb{E}(X) = \theta = Var(X)$.
O estimador de máxima verossimilhança (EMV) de θ é $\hat{\theta} = \overline{X}_n$.
Neste caso, $\mathbb{E}(\hat{\theta}) = \mathbb{E}(\overline{X}_n) = \theta, \forall n$, e $Var(\hat{\theta}) = \frac{Var(X)}{n} = \frac{\theta}{n}, \forall n$.
Pelo teorema central do limite,

$$\sqrt{n}(\overline{X}_n - \theta) \xrightarrow{D} N(0, \theta), \text{ quando } n \to \infty.$$

De forma aproximada,

$$\overline{X}_n \sim N\left(\theta, \frac{\theta}{n}\right).$$

Obs.: A distribuição **exata** de $\overline{X}_n$ é discreta.
De outra forma,

$$\sqrt{n}\frac{(\overline{X}_n - \theta)}{\sqrt{\theta}} \xrightarrow{D} N(0, 1).$$

Esquema de um estudo de simulação

Algumas perguntas (envolvendo estimadores cuja distribuição assintótica for obtida)

1. Como o viés do estimador varia em função do tamanho da amostra (n)?
2. Como a variância do estimador varia em função de n?
3. Como se comporta a distribuição do estimador em função de n?

Se $\hat{\theta}$ é o EMV de θ, então

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{D} N\left(0, \frac{1}{I_F(\theta)}\right), \text{ quando } n \to \infty.$$

Alguns passos:
Digamos que a distribuição de X tem um parâmetro θ:

1. Selecionamos θ_0, que é o verdadeiro valor de θ. Selecionamos um tamanho de amostra(n) e o número de repetições da simulaçãoR. R é tipicamente de ordem 10^3 ou 10^4;
2. Gerar uma a.a. $X_1, \ldots, X_n$ da distribuição de X e calcular $\hat{\theta}$;
3. Repetir o passo 2 R vezes obtendo $\hat{\theta}_1, \ldots, \hat{\theta}_R$;
4. O viés simulado é dado por $\theta_0 - \mathbb{E}(\hat{\theta})$

$$\theta_0 - \frac{1}{R}\sum_{j=1}^{R} \hat{\theta}_j = \theta_0 - \overline{\hat{\theta}}.$$

O viés simulado relativo é dado por:

$$1 - \frac{\overline{\hat{\theta}}}{\theta_0}, \text{ que pode ser expresso em \%.}$$

5. O desvio-padrão simulado do estimador $\hat{\theta}$ é dado por:

$$\left\{ \frac{1}{R-1} \sum_{j=1}^{R} (\hat{\theta}_j - \bar{\hat{\theta}})^2 \right\}^{\frac{1}{2}} = dps.$$

Dizemos que $Var(\hat{\theta}) = \sigma^2(\theta)$. Por exemplo, se $\hat{\theta}$ é o EMV de θ, então $\sigma^2(\theta) = \dfrac{1}{nI_F(\theta)}$, aproximadamente.

O desvio padrão do estimador é $\sigma(\theta)$. Calculamos:

$$\frac{1}{R} \sum_{j=1}^{R} \sigma(\hat{\theta}_j) \text{ e comparamos com } dps.$$

6. A raiz quadrada do erro quadrático médio (REQM) simulado é dado por:

$$REQM = \left\{ \frac{1}{R} \sum_{j=1}^{R} (\hat{\theta}_j - \theta_0)^2 \right\}^{\frac{1}{2}}$$

7. Para cada repetição j, um intervalo de confiança aproximado de $(1-\alpha)\%$ para θ é dado por:

$$\hat{\theta}_j \mp z_{1-\frac{\alpha}{2}} \sigma(\hat{\theta}_j),$$

Se este intervalo contém θ_0, dizemos que é um intervalo correto, $j = 1, \ldots, R$.

Se m é o número de intervalos corretos, a probabilidade de cobertura do intervalo é dada por $\frac{m}{R}$, que esperamos ser próxima de $1-\alpha$.

Obs.: Recomenda-se realizar o estudo com diferentes valores de n e θ_0.

8. A distribuição assintótica do estimador pode ser avaliada usando o resultado

$$\frac{\sqrt{n}(\hat{\theta} - \theta_0)}{\sigma(\theta_0)} \xrightarrow{D} N(0,1), \text{ quando } n \to \infty,$$

$\dfrac{\sqrt{n}(\hat{\theta} - \theta_0)}{\sigma(\theta_0)}, j = 1, \ldots, R$ podem ser representados por um histograma ou pela função distribuição empírica e comparados com a distribuição $N(0,1)$.

Em alguns estudos devem ser informados:

1. Equipamento computacional (processador, memória RAM, etc.);
2. Sistema operacional e versão;
3. Linguagem de programação e/ou pacotes estatísticos;
4. Gerador de números aleatórios e a semente.

4 Métodos de reamostragem

Reamostragem, ou em inglês resampling methods: a partir de uma amostra, pode-se gerar outras amostras a partir dessa. Sem necessitar fazer outra amostragem.

4.1 Método do *bootstrap*

Método computacional para resolver problemas de inferência como, por exemplo, estimação do erro padrão de um estimador, estimação do viés de um estimador, construção de intervalos de confiança e testes de hipóteses.

4.1.1 Função distribuição empírica

Se X é uma v.a. com f.d.a $F(x)$ denotamos a função distribuição empírica por $\hat{F}(x)$ ou $\hat{F}_n(x)$. Em uma a.a. de tamanho n, $\hat{F}(x)$ é dada por:

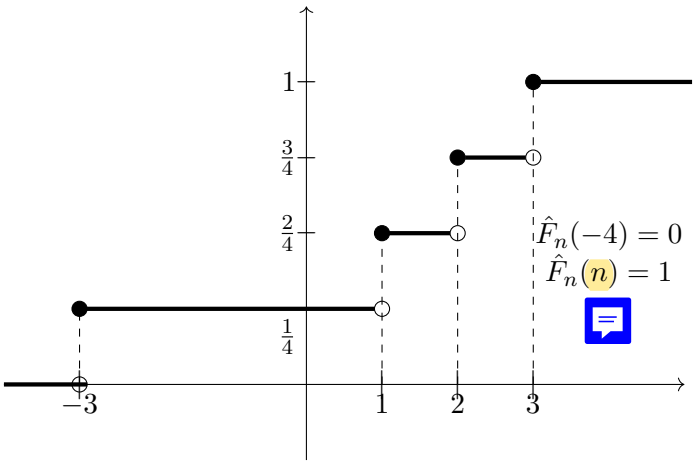
$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(X_i \leq x), \text{ para } x \in \mathbb{R}, \text{ sendo que:}$$

$$\mathbb{I}(X_i \leq x) = \begin{cases} 0, & \text{se } X_i > x, \\ 1, & \text{se } X_i \leq x. \end{cases}$$

$\hat{F}_n(x)$ tem todas as propriedades de uma f.d.a.

Na amostra os valores observados são $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$. Supondo que $x_1 \neq x_2 \neq \ldots \neq x_n$, $\hat{F}_n(x)$ tem saltos de altura igual a $\frac{1}{n}$ em cada diferente valor.

Exemplo: $n = 4; x_1 = 1, x_2 = -3, x_3 = 2, x_4 = 3$



$\hat{F}_n(x)$ corresponde a uma v.a. discreta com no máximo n valores diferentes.

Para x fixado, calculamos

$$\mathbb{E}[\hat{F}_n(x)] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(X_i \leq x)\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\mathbb{I}(X_i \leq x)],$$

Notando que

$$\begin{aligned}\mathbb{I}(X_i \leq x) &\sim Bernoulli(P(X_i \leq x) \\ &\sim Bernoulli(F(x)).\end{aligned}$$

Portanto,

$$\mathbb{E}[\hat{F}_n(x)] = \frac{1}{n}nF(x) = F(x),$$

ou seja, $\hat{F}_n(x)$ é estimador não viesado de $F(x), \forall x \in \mathbb{R}$.

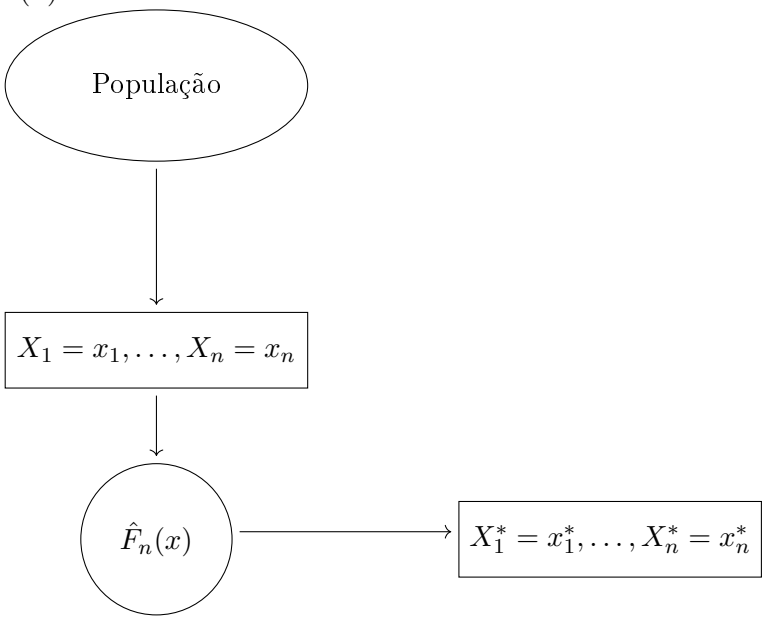
Calculamos:

$$Var(\hat{F}_n(x)) = Var\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(X_i < x)\right) = \frac{1}{n^2}nF(x)\{1-F(x)\}$$
$$= \frac{1}{n}F(x)\{1-F(x)\}$$

Portanto, $\lim_{n\to\infty} Var(\hat{F}_n(x)) = 0$.

Sendo assim, $\hat{F}_n(x)$ é um estimador consistente para $F(x)$, ou seja, $\hat{F}_n(x) \longrightarrow F(x)$ pela lei forte dos grandes números.

As propriedades de $\hat{F}_n(x)$ indicam que a função distribuição empírica pode ser usada no lugar da f.d.a $F(x)$.



Uma amostra *bootstrap* é uma a.a obtida da distribuição com f.d.a $\hat{F}_n(x)$, que é denotada por $X_1^*, \ldots, X_n^*$ com valores $x_1^*, \ldots, x_n^*$. Obter uma amostra *bootstrap* equivale a obter uma amostra elatória de tamanho n com reposição dos valores $(x_1, \ldots, x_n)$. O número total de amostras *bootstrap* é n^n. Por exemplo, se $n = 100$, o número total de amostras *bootstrap* é 10^{200}.

Obs.: Em R, uma amostra *bootstrap* de dados (vetor $n \times 1$) é obtida com o comando *sample(dados,n,replace=T)*

Exemplo: $x_1 = 1, x_2 = -3, x_3 = 2, x_4 = 3$.

Uma amostra*bootstrap*: $x_1^* = 2, x_2^* = 1, x_3^* = 3, x_4^* = 1$.

Obs.: Em R, $\hat{F}_n(x)$ é obtida com a função *ecdf*.

```
Fn <- ecdf(dados)
plot(Fn)
```

Em um problema de estimação devemos considerar uma função $\theta(F)$ ou simplesmente θ, que depende de F.

Por exemplo, pode ser $\mathbb{E}(X)$, mediana(X), ou $P(X > x_0)$. Um estimador para θ é denotado por:

$$\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n).$$

Como estimar o erro padrão de $\hat{\theta}$ usando amostras *bootstrap*?

Geralmente não é viável utilizar o número total de amostras *bootstrap*. Utilizamos B amostras *bootstrap* em que B é "grande", mas muitor menor que $n^n (B \ll n^n)$. B é escolhido como 10 000 em muitas aplicações. O método foi proposto por Bradley Efron (1979).

Quando $B < n^n$, obtemos soluções aproximadas por métodos de Monte Carlo. Quanto maior o valor de B, melhor a aproximação.

Passos:

1. Calcular $\hat{F}_n(x)$;
2. Obter uma amostra *bootstrap* $X^*_{b1}, \ldots, X^*_{bn} \overset{iid}{\sim} \hat{F}_n(x)$ com valores $x^*_{b1}, \ldots, x^*_{bn}$;
3. Calcular $\hat{\theta}$ para a amostra *bootstrap* b, obtendo $\hat{\theta}^*_b$;
4. Repita os passos 2 e 3 obtendo $\hat{\theta}^*_1, \ldots, \hat{\theta}^*_B$.

O erro padrão *bootstrap* é dado por:

$$ep^*(\hat{\theta}) = \left\{ \frac{1}{B-1} \sum_{b=1}^{B} (\hat{\theta}^*_b - \overline{\hat{\theta}}^*)^2 \right\}^{\frac{1}{2}}, \text{ em que } \overline{\hat{\theta}}^* = \frac{1}{b} \sum_{b=1}^{B} \hat{\theta}^*_b.$$

Método do *bootstrap*

Utilizado para obter estimativa do erro padrão de um estimador. O método envolve dois níveis de aproximação:

1. $\hat{F}_n$ é uma aproximação de F e,
2. Em geral, o número total de amostras da distribuição *bootstrap* (n^n) é muito grande. Portanto, é impraticável obter a distribuição *bootstrap* exata.

Os resultados são obtidos utilizando simulações de Monte Carlo com B amostras. Recomenda-se $B = 5\,000$ ou $10\,000$.

Obs.: $\hat{\theta}$ é um estimador de θ com esperança $\mathbb{E}_F(\hat{\theta})$, em que "$F$" indica a distribuição da população. $\mathbb{E}_F(\hat{\theta})$ indica a esperança calculada com a distribuição empírica $(\hat{F}_n)$.

4.1.2 ~~Erro padrão~~

Escrevemos $\hat{\theta}$ como sendo $\hat{\theta} = g(x_1, \ldots, x_n)$. Uma amostra da distribuição empírica $\hat{F}$ tem valores $x_{i_1}, x_{i_2}, \ldots, x_{i_n}$, sendo que $i, j \in \{1, 2, \ldots, n\}$.

Temos n^n possíveis amostras, cada uma com probabilidade $\frac{1}{n^n}$. Calculamos:

$$\mathbb{E}_F(\theta) = \frac{1}{n^n} \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \cdots \sum_{i_n=1}^{n} g(x_{i_1}, x_{i_2}, \ldots, x_{i_n}).$$

Uma aproximação baseada em simulações de Monte Carlo com B amostras *bootstrap* é dada por:

$$\frac{1}{B} \sum_{b=1}^{B} g(x^*_{b,1}, x^*_{b,2}, \ldots, x^*_{b,n}).$$

Obs.: Na situação de duas amostras, devemos observar se são independentes ou pareadas. Com amostras independentes $X_1, X_2, \ldots, X_n$ e $Y_1, Y_2, \ldots, Y_m$ tomamos amostras *bootstrap* das distribuições empíricas de X e Y e calculamos o estimador de θ. Por exemplo, se $\theta = \mathbb{E}(X) - \mathbb{E}(Y)$. Para amostras *bootstrap* $x^*_1, x^*_2, \ldots, x^*_n$ e $y^*_1, y^*_2, \ldots, y^*_m$.

Calculamos $\hat{\theta}^* = \overline{x}^* - \overline{y}^*$.

Com dados pareados $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$, tomamos amostras *bootstrap* $i_1, i_2, \ldots, i_n$ do conjunto $\{1, 2, \ldots, n\}$. Uma amostra *bootstrap* é dada por: $(x^*_{i_1}, x^*_{i_1}), (x^*_{i_2}, y^*_{i_2}), \ldots, (x^*_{i_n}, y^*_{i_n})$

Exemplo:

```
sample(n,n,replace=T) #amostra de indices

library(datasets)
data(iris)
x <- iris[1:50, 2]
theta <- median(x)
n <- length(x)
B <- 5000
thetas <- c()
for(b in 1:B) {
        thetas[b] <- median(sample(x, n, replace = T))
}
hist(thetas, freq=F)
sd(thetas)
```

4.1.3 Estimativa *bootstrap* do viés

Se $\hat{\theta}$ é um estimador de θ, o viés de $\hat{\theta}$ é definido como: $\mathbb{E}(\hat{\theta}) - \theta$, ou de forma mais completa:

$$E_F(\hat{\theta}) - \theta(F).$$

O viés *bootstrap* é definido como:

$$E_F(\hat{\theta}^*) - \hat{\theta}.$$

O viés *bootstrap* pode ser estimado usando simulações de Monte Carlo com B amostras *bootstrap*. A expressão é:

$$\frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^* - \hat{\theta} = \overline{\hat{\theta}}^* - \hat{\theta}.$$

Um estimador de θ corrigido pelo víes é dado por:

$$\hat{\theta} - (\overline{\hat{\theta}}^* - \hat{\theta}) = 2\hat{\theta} - \overline{\hat{\theta}}^*.$$

A variância do estimador corrigida tende a ser maior do que a variância de $\hat{\theta}$, mas pode haver ganho em termos de EQM (*bias-variance tradeoff*).

4.1.4 Intervalo de confiança *bootstrap*

a) Intervalo percentil:

Usando B amostras *bootstrap*, obtemos: $\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_B^*$.

Tomando $\alpha \in (0,1)$, um IC de $100(1-\alpha)\%$ para θ tem limites dados pelos percentis $100(\alpha/2)$ e $100(1-\alpha/2)$ dos valores $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$.

Para um IC de 95% de confiança, os percentis são 2,5% e 97,5%.

Obs.: Em R, se "thetab" é o vetor de estimativas *bootstrap* de θ, o comando: *quantile(thetab, probs=c(0.0025,0.975), type = 6)*

b) Intervalo *t bootstrap*:

Quando for viável sua implementação, é recomendado. Definimos:

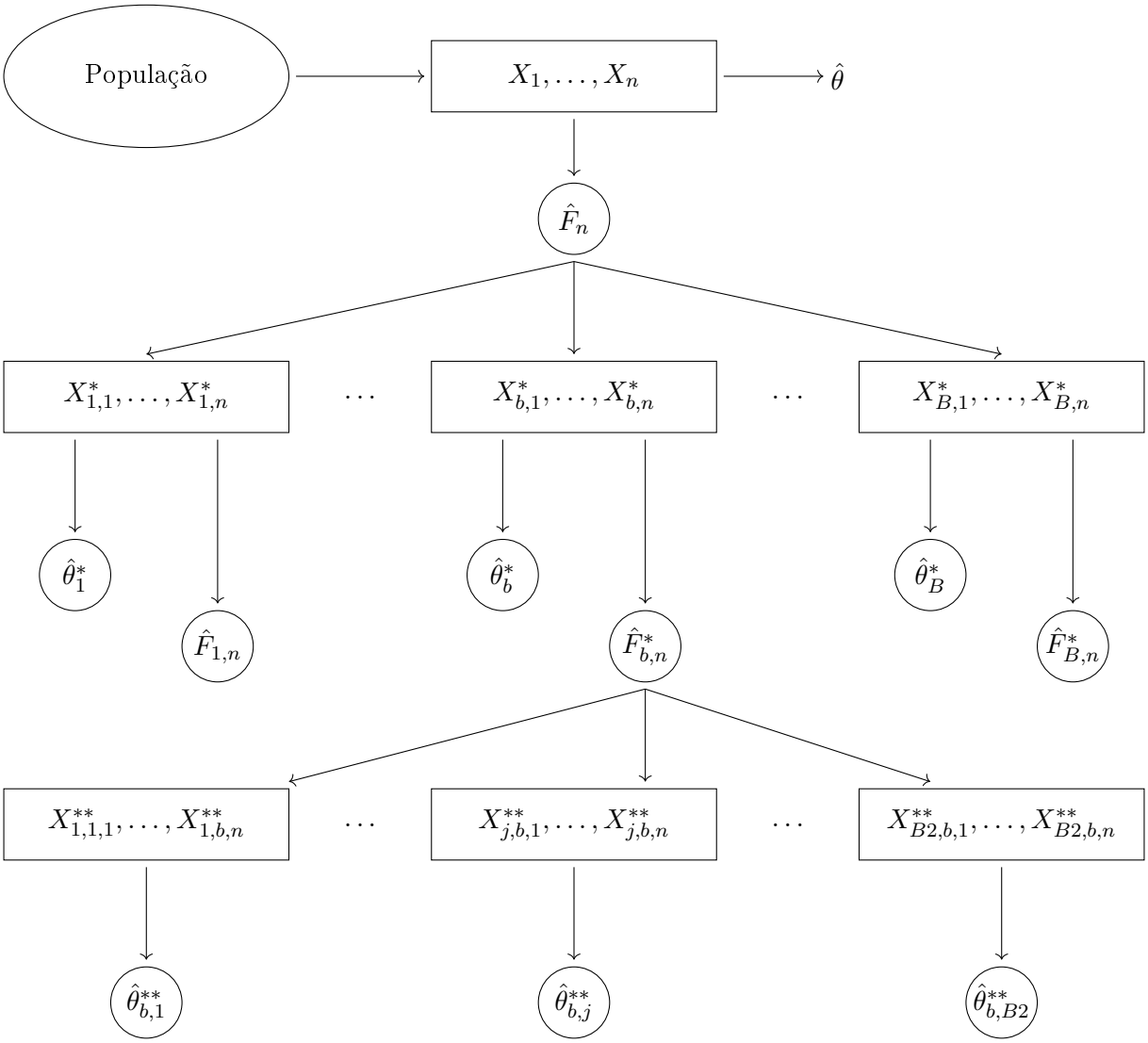
$$t = \frac{\hat{\theta} - \theta}{ep(\hat{\theta})}. \tag{1}$$

A versão *bootstrap* de t é:

$$t_b^* = \frac{\hat{\theta}_b^* - \hat{\theta}}{ep_b^*(\hat{\theta})}, b = 1, \ldots, N. \tag{2}$$

Se a distribuição de t na expressão (1) é t-student com k graus de liberdade podemos obter um IC para θ.

A obtenção de $ep_b^*(\theta)$ é baseada no Método *Bootstrap* iterado (*iterated bootstrap*), para $b = 1, \ldots, B$.



Na $b-$ésima amostra *bootstrap* obtemos $\hat{\theta}_{b,1}^{**}, \ldots, \hat{\theta}_{b,B2}^{**}$, calculamos:

$$\overline{\hat{\theta}}_b^{**} = \frac{1}{B_2} \sum_{m=1}^{B_2} \hat{\theta}_{b,m}^{**}, \text{ para } b = 1, \ldots, B \qquad (B2 < B).$$

Em seguida, calculamos:

$$ep_b^*(\hat{\theta}) = \left\{ \frac{1}{B_2 - 1} \sum_{m=1}^{B_2} (\hat{\theta}_{b,m}^{**} - \overline{\hat{\theta}}_b^{**})^2 \right\}.$$

Na expressão (2) calculamos $t_b^*, b = 1, \ldots, B$.

***Bootstrap* iterado**, temos: $t_1^*, t_2^*, \ldots, t_B^*$. Além disso, $q_{\frac{\alpha}{2}}$ e $q_{1-\frac{\alpha}{2}}$ representam os quantis amostrais de $(t_1^*, \ldots, t_B^*)$. Um intervalo de confiança para θ é dado por:

$$(\hat{\theta} - q_{1-\frac{\alpha}{2}} ep^*(\hat{\theta}), \hat{\theta} - q_{\frac{\alpha}{2}} ep^*(\hat{\theta})),$$

em que,

$$ep^*(\hat{\theta}) = \left\{ \frac{1}{B-1} \sum_{b=1}^{B} (\hat{\theta}_b^* - \overline{\hat{\theta}}^*) \right\}^{1/2}.$$

4.1.5 Tópicos adicionais

1) *Bootstrap* paramétrico

Modelo para $X : X \sim$ *distribuição*(γ). Com base em uma a.a. obtemos um estimador $\hat{\gamma}$ para γ. Amostras *bootstrap* $X_1^*, \ldots, X_n^*$ são geradas de *distribuição*$(\hat{\gamma})$.

2) Testes *bootstrap*

Para obter o valor-p de um teste, amostras *bootstrap* devem ser geradas com a hipótese nula H_0, que pode representar dificuldades. Testes de permutação são mais usados.

3) *Bootstrap* em modelos de regressão

$$Y_i = g(\boldsymbol{x}_i, \boldsymbol{\beta}) + \varepsilon_i.$$

em que $\boldsymbol{x}_i$ e $\boldsymbol{\beta}$ podem ser vetores e os erros $\boldsymbol{\varepsilon}_i$ são independentes com $\mathbb{E}(\varepsilon_i) = 0$ e $Var(\varepsilon_i) = \sigma^2$, para $i = 1, \ldots, n$.

No modelo linear, $g(\boldsymbol{x}, \boldsymbol{\beta}) = \boldsymbol{x}^T\boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$.

Obtemos estimativas $\hat{\beta}$ para β e calculamos os resíduos $e_i = Y_i - g(x_i, \hat{\beta}), i = 1, \ldots, n$

Uma amostra *bootstrap* dos resíduos é dada por: $e_1^*, e_2^*, \ldots, e_n^*$.

Calculamos: $Y_i^* = g(x_i, \hat{\beta}) + e_i^*, i1, \ldots, n$.

Daí, obtemos: $\hat{\beta}^* = \begin{bmatrix} Y_1^*, x_1 \\ Y_2^*, x_2 \\ \vdots \\ Y_n^*, x_n \end{bmatrix}$

Pacote boot e resample em R.

4.2 Método *jackknife*

Proposto por Quenouille (1949). Seja $X_1, \ldots, X_n \overset{iid}{\sim} F$ com observações $x_1, \ldots, x_n$. Problema: estimar $\theta = E_F(X)$.

$\hat{\theta} = \overline{X}$ é um estimador não viesado e consistente. O erro padrão de $\hat{\theta} = \overline{X}$ é dado por:

$$\hat{\sigma}(\overline{X}) = \left\{ \frac{1}{n(n-1)} \sum_{i=1}^{n} (X_i - \overline{X})^2 \right\}^{\frac{1}{2}}.$$

Considere:

$$\overline{X}_{[i]} = \frac{1}{n-1} \sum_{\substack{j=1 \\ j \neq i}} X_j,$$

que representa a média amostral sem a $i-$ésima observação, para $i = 1, \ldots, n$.

Considere:

$$\overline{X}_{[\cdot]} = \frac{1}{n}\sum_{i=1}^{n}\overline{X}_{[i]}.$$

De fato $\overline{X}_{[\cdot]} = \overline{X}$.

Considere:

$$\hat{\sigma}_j = \left\{\frac{n-1}{n}\sum_{i=1}^{n}(\overline{X}_{[i]} - \overline{X}_{[\cdot]})^2\right\}^{\frac{1}{2}}.$$

De fato $\hat{\sigma}_j = \hat{\sigma}(\overline{X})$.

Considere que θ é uma característica qualquer da população (por exemplo, a mediana) com estimador $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$. Neste caso, $\hat{\theta}_{[i]} = \hat{\theta}(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$, para $i = 1, \ldots, n$ e

$$\overline{\hat{\theta}}_{[\cdot]} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{[i]}.$$

O erro padrão *jackknife* é dado por:

$$\hat{\sigma}_j = \left\{\frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}_{[i]} - \overline{\hat{\theta}}_{[\cdot]})^2\right\}^{\frac{1}{2}}.$$

(Vantagem em relação a *bootstrap*: mais leve).

Obs.: Se x é o vetor com os dados, $mean(x[-i])$ calcula $\overline{X}_{[i]}$.

Exemplo: θ é a mediana de X. Observações estão no vetor $x(n \times 1)$.

O estimador $\hat{\theta}$ é a mediana amostral. Em R, função *median*().

```
thetaci <- c()
n <- length(x)
for(i in 1:n) {
        thetaci[i] <- median(x[-i])
}
epJ <- (n-1)*sd(x)/sqrt(n)
```

Obs.: O método *jackknife* não é tão intensivo computacionalmente quanto o método *bootstrap*.

Obs.: Se $\hat{\theta} = \overline{X}$, então $\overline{\hat{\theta}}_{[\cdot]} = \overline{X}_{[\cdot]} = \overline{X}$.

De fato,

$$\begin{aligned}\overline{X}_{[\cdot]} &= \frac{1}{n}\sum_{i=1}^{n}\overline{X}_{[i]} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{n-1}\sum_{\substack{j=1\\ j\neq i}}^{n}X_j = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\sum_{\substack{j=1\\ j\neq i}}^{n}X_j + X_i - X_i\right)\\ &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\sum_{j=1}^{n}X_j - X_i\right) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(n\overline{X} - X_i)\\ &= \frac{1}{n(n-1)}(n^2\overline{X} - n\overline{X}) = \frac{1}{n(n-1)}n\overline{X}(n-1) = \overline{X}.\end{aligned}$$

4.2.1 Estimativa *jackknife* para o viés

Definimos $E_n = E_F[\hat{\theta}(X_1, \ldots, X_n)]$

Diversos estimadores têm esperança que pode ser escrita como:

$$E_n = \theta + \underbrace{\frac{a_1(F)}{n} + \frac{a_n(F)}{n^2} + \ldots}_{\text{viés}}$$

em que $a_1(F), a_2(F), \ldots,$ são funções que não dependem de θ.

Em geral, $E_n \neq \theta$. O viés é da ordem $\frac{1}{n}$.

Calculamos:

$$E_F(\bar{\hat{\theta}}_{[\cdot]}) = \frac{1}{n}\sum_{i=1}^{n} E_F(\hat{\theta}_{[i]}) = \frac{1}{n}\sum_{i=1}^{n} E_F[\hat{\theta}(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)]$$
$$= \frac{1}{n}\sum_{i=1}^{n} E_{n-1} \cancel{E_{n-1}},$$

sendo que,

$$E_{n-1} = \theta + \frac{a_1(F)}{n-1} + \frac{a_2(F)}{(n-1)^2} + \ldots$$

Calculamos agora $nE_n - (n-1)E_{n-1}$ e obtemos:

$$n\theta + \cancel{a_1(F)} + \frac{a_2(F)}{n} + \frac{a_3(F)}{n^2} + \ldots$$
$$- (n-1)\theta - \cancel{a_1(F)} - \frac{a_2(F)}{n-1} - \ldots$$
$$= \theta \underbrace{- \frac{a_2(F)}{n(n-1)} + a_3(F)\left\{\frac{1}{n^2} + \frac{1}{(n-1)^2}\right\} + \ldots}_{\text{viés}}$$

A expressão anterior diferente de θ por um termo de ordem $\frac{1}{n^2}$ (que é um ganho em relação a $\frac{1}{n}$).

Com base neste resultado, obtemos o estimador

$$\tilde{\theta} = n\hat{\theta} - (n-1)\hat{\theta}_{[\cdot]} = \hat{\theta} - \underbrace{(n-1)(\hat{\theta}_{[\cdot]} - \hat{\theta})}_{\text{viés}},$$

em que $\hat{\theta}$ é a esperança do estimador.

Outra proposta de estimador *jackknife* para o viés é: $n(\hat{\theta}_{[\cdot]} - \hat{\theta})$.

Exemplo: Parâmetro θ : mediana. Estimador $\hat{\theta}$: mediana amostral. x vetor $n \times 1$ com os dados.

```
n <- length(x)
thetac <- median(x) #theta chapeu
thetacp <- c() #theta [.]
for(i in 1:n) {
        thetacp[i] <- median(x[-i])
}
thetacpb <- mean(thetacp) #theta chapeu barra [.]
vies <- n*(thetacpb - thetac)
thetatil <- thetac - vies
```

4.3 Validação cruzada *(cross validation)*

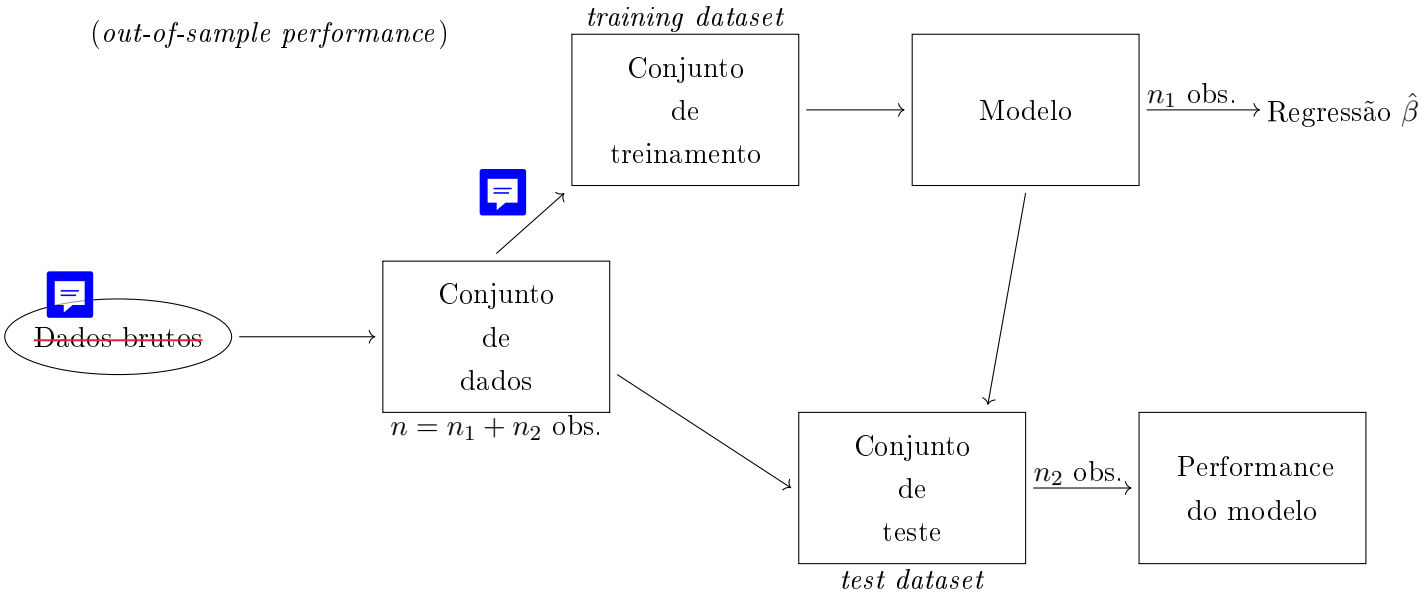
$$\cancel{\sqrt{n}(\hat{\alpha} - \alpha) \xrightarrow{D} N(0, \alpha^2(1 + \alpha\psi_1(1+\alpha)))}.$$

$$z = \frac{\sqrt{n}(\hat{\alpha} - \alpha)}{\alpha\sqrt{1 + \alpha\psi_1(1+\alpha)/\sqrt{n}}} \xrightarrow{D} N(0,1).$$

($\sqrt{n}$ erro padrão assintótico).

Em diversas aplicações de modelos estatísticos (por exemplo, modelos de regressão) uma etapa importante é a verificação da capacidade preditiva do modelo.

Tipicamente, a capacidade preditiva é avaliada utilizando dados que **NÃO** foram usados para ajustar o modelo.



O procedimento é chamado de método de extensão (*holdout method*).

Em um modelo de regressão com vvariável resposta Y, a performane do modelo pode ser avaliada com base no erro de predição. Uma medida usada é o erro relativo:

$$\frac{1}{n_2}\sum_{i \in teste}\left|\frac{Y_i - \hat{Y}_i}{Y_i}\right| 100\%,$$

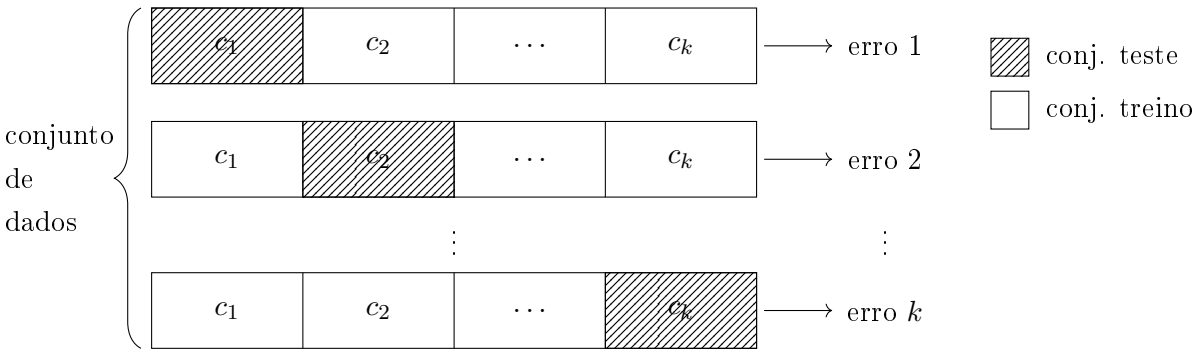
sendo que $\hat{Y}_i$ é calculado com base no vetor de coeficientes $\hat{\boldsymbol{\beta}}$ obtido como ajuste do modelo ao conjunto de dados de treinamento.

Um refinamento do modelo consiste em dividir o conjunto de dados três partes:

1. Conjunto de treinamento (ex.: 50%);
2. Conjunto de validação (*validation dataset, ex.: 25%*);
3. Conjunto de teste (ex.: 25%).

Com aplicações sucessivas do método de extensão obtemos o método de validação cruzada $k-$dobras (*k-fold cross validation*).

O conjunto de dados é dividido (por sorteio) em k conjuntos disjuntos com aproximadamente o mesmo número de observações.



A capacidade do modelo é avaliada usando a média dos k erros.

A situação extrema ocorre quando o número de observações de cada conjunto é igual a 1, de modo que $k = n$. Neste caso, obtemos o método de eliminação individual (*leave-one-out crossvalidation*).

O método *10-fold cross validation* é bastante usado em aplicações de Aprendizagem de Máquina (*Machine Learning*).

Obs.: Problemas em que a variável resposta Y tem valores conhecidos em pelo menos uma parte dos dados: aprendizagem supervisionada (*supervised learning*). Problemas em que no conjunto de dados a vvariável resposta não tem valores conhecidos: aprendizagem não supervisionada (*unsupervised learning*).

5 Testes de permutação (modo alternativo de testes de hipóteses)

Testes de permutação (em inglês, *permutation tests*) são testes baseados em amostras **sem** reposição. As amostras envolvem todas as permutações possíveis dos índices das observações.

5.1 Amostras independentes

Os dados são observações de duas amostras aleatórias $X_1, \ldots, X_{n_1}$ e $Y_1, \ldots, Y_{n_2}$, independentes entre si e $n = n_1 + n_2$. As funções distribuição de X e Y são F e G, respectivamente. Devemos testar a hipótese:

$$H_0 : F = G \qquad vs \qquad H_1 : F \neq G$$

Para realizar o teste, utilizamos uma estatística de teste D.

Suponha que H_0 é rejeitada quando o valor de D é "grande". Para os valores observados $x_1, \ldots, x_{n_1}$ e $y_1, \ldots, y_{n_2}$, o valor da estatística de teste é d_{obs}. O valor-p do teste é dado por:

$$p = P_0(D \geq d_{obs}) \qquad (P_0 : H_0 \text{ é verdadeira})$$

Se H_0 for verdadeira, as n observações podem ser combinadas em uma única amostra.

Exemplo:

$x_1 = 0,9$	$y_1 = 1,8$
$x_2 = 1,7$	$y_2 = 1,1$
$x_3 = 1,3$	
$n_1 = 3$	$n_2 = 2$

Se H_0 for verdadeira, obtemos uma amostra de $n = 5$ observações:

$$\begin{array}{ll} x'_1 = 1,8 & y'_1 = 0,9 \\ x'_2 = 1,3 & y'_2 = 1,1 \quad \text{(uma amostra de permutação)} \\ x'_3 = 1,7 & \end{array}$$

Se $D = \overline{X} - \overline{Y}$, obtemos $d_{obs} = 1,3 - 1,45 = -0,15$. Na amostra de permutação obtemos: $D' = 1,6 - 1 = 0,6$. O número total de permutações é $\binom{n}{n_1} = \binom{n}{n_2}$, pois $n = n_1 + n_2$.

A distribuição exata condicional (nos dados observados) é baseada nas $\binom{n}{n_1}$ permutações das observações. A distribuição exata condicional de D é discreta.

Para cada permutação nos n_1 rótulos "x" e n_2 rótulos "y" no conjunto $1, 2, \ldots, n_1, n_1 + 1, \ldots, n_1 + n_2$, calculamos o valor D, denotado por d'. O número de permutações em que $d' \geq d_{obs}$ é denotado por k. Cada permutação tem probabilidade $\dfrac{1}{\binom{n}{n_1}}$.

O valor-p exato condicional é dado por:

$$p = \frac{k}{\binom{n}{n_1}}$$

O número de permutações $\binom{n}{n_1}$ pode ser muito grande, significando que não é viável percorrer todas as permutações. Neste caso, o valor-p do teste pode ser aproximado pelo Método de Monte Carlo baseado em R permutações. Uma aproximação para o valor-p é dada por:

$$p \approx \frac{k' + 1}{R + 1} \qquad (*)$$

em que k' é o número de permutações em que $d' \geq d_{obs}$. Na expressão $(*)$ está incluída a amostra observada. Desta forma evitamos $p = 0$.

Obs.: As observações estão nos vetores x e y.

```
dados <- c(x, y)
n1 <- length(x)
n2 <- length(y)
n <- n1+n2
dadosl <- sample(dados)
xl <- dadosl[1:n1]
yl <- dados[(n1+1):n]
```

Obs.: Para aplicar o teste de permutação às hipóteses $H_0 : \mu_x - \mu_y = 0$ e $H_1 : \mu_x - \mu_y \neq 0$, devemos verificar se as variâncias das duas variáveis são iguais.

A estatística pode ser $\overline{X} - \overline{Y}$.

Quando a hipótese alternativa for bilateral, calculamos dois valores-p, p_1 e p_2, correspondendo às duas hipóteses alternativas unilaterais. O valor-p dado por:

$$2 \cdot \min(P_1, P_2).$$

A estatística de teste D com d_{obs}.

P_1 é calculado analisando a desigualdade $D \geq |d_{obs}|$. P_2 é calculado analisando a desigualdade $D \leq -|d_{obs}|$.

5.2 Testes de associação

Os dados são uma amostra aleatória dos pares $(X_1, Y_1), \ldots, (X_n, Y_n)$.

A medida de associação pode ser o coeficiente de correlação de Pearson, de Kendall, de Spearman ou o coeficiente de correlação de concordância. O objetivo é testar a hipótese nula em que o valor da medida de associação é igual a 0.

Por exemplo, tomando o coeficiente de correlação linear de Pearson, testamos:

$$H_0 : \rho = 0 \qquad vs \qquad H_1 : \rho \neq 0, \text{ em que}$$

$$\rho = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}.$$

Se H_0 for verdadeira, a associação é nula e as observações de X e Y podem ser permutadas fixando uma das duas variáveis.

O número total de amostras de permutação é $n!$, que pode ser muito grande (por exemplo, se $n = 25$, este número é da ordem de 10^{25}).

Um estimador para ρ é dado por:

$$\hat{\rho} = \frac{S_{XY}}{\sqrt{S_X^2 S_Y^2}} \text{(S representa variâncias e a covariância)}$$

No cálculo exato do valor-p, para cada amostra de permutação verificamos se:

$$\hat{p}' \geq |\underbrace{\hat{p}_{obs}}_{amostras}| \qquad \text{e} \qquad \hat{p}' \leq -|\hat{p}_{obs}|,$$

com totais k_1 e k_2.

O valor-p exato (condicional) é dado por:

$$p = \frac{2 \cdot \min(k_1, k_2)}{n!}.$$

Uma aproximação de Monte Carlo é dada por:

$$p \approx 2 \cdot \min(P_1', P_2'),$$

em que,

$$P_1' = \frac{K_1' + 1}{R + 1} \qquad e \qquad P_2' = \frac{K_2' + 1}{R + 1},$$

e R é o número de simulações de Monte Carlo.

5.3 Teste global em regressão múltipla

Modelo de regressão múltipla com variável resposta Y e k variáveis explicativas. O vetor de coeficientes é denotado por: $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^T$ em que β_0 é o intercepto.

O objetivo é testar:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0 \qquad vs \qquad H_0 : \beta_j \neq 0$$

para pelo menos um $j \in \{1, \ldots, k\}$.

No modelo normal, o teste é realizado com uma estatística de teste W cuja distribuição é ~~$F_{n-(k+1),k}$~~ quando H_0 é verdadeira.

$$W = \frac{SQE}{(n-k)},$$

o teste é unilateral.

Se H_0 for verdadeira, as k variáveis explicativas não têm efeito sobre a variável resposta.

$$\begin{bmatrix} \boldsymbol{x}_1^T & Y_1 \\ \vdots & \vdots \\ \boldsymbol{x}_i & Y_i \\ \vdots & \vdots \\ \boldsymbol{x}_n^T & Y_n \end{bmatrix}, \text{ em que } \boldsymbol{x}_i = (1, x_{i_1}, \ldots, x_{i_n})^T.$$

Amostras de permutação são obtidas permutando os elementos do vetor $(Y_1, \ldots, Y_n)^T$.

O teste de H_0 pode ser realizado usando a estatística W, mas o valor-p é aproximado por simulações de Monte Carlo. Se k' é o número de amostras de permutação em que $W' \geq W_{obs}$, então $p \approx \dfrac{k'+1}{R+1}$.

Obs.: Fisher (1936) já reconhecia a importância dos testes de permutação, impraticável nos anos 30 do século XX.

6 Métodos de Monte Carlo com Cadeias de Markov

Em inglês, *Markov Chain Monte Carlo Methods*, ou também conhecido como métodos MCMC.

$\underset{\sim}{X}$ é um vetor $p \times 1 (p > 1)$ com função densidade $f(\underset{\sim}{x})$ que não depende de parâmetro desconhecido.

O problema consiste em gerar uma amostra aleatória $\underset{\sim}{X}_1, \ldots, \underset{\sim}{X}_n$ da distribuição de $\underset{\sim}{X}$.

Os métodos MCMC são convenientes quando **não** é possível resolver o problema de forma exata.

Os métodos MCMC permitiram uma (r)evolução em muitas áreas da Estatística (por exemplo, em inferência bayesiana).

Contribuições importantes: Metropolis & outros (1953) e Geman & Geman (1984).

6.1 Amostrador de Gibbs

Em inglês, *Gibbs sampler*, foi proposto em problemas de computação gráfica utilizando a distribuição de Gibbs.

Gelfand e Smith (1990) contribuíram para difundir o uso de métodos como o amostrador de Gibbs em problemas de Estatística.

As amostras são geradas utilizando os componentes $X_1, X_2, \ldots, X_p$ do vetor $\underset{\sim}{X}$ ou então blocos dos componentes de $\underset{\sim}{X}$.

Exemplo: $\underset{\sim}{X} = \begin{bmatrix} X_1 \\ X_2 \\ \hline X_3 \\ X_4 \\ \hline X_5 \end{bmatrix}$, em que $X_1, X_2; X_3, X_4; X_5$ formam diferentes blocos.

A notação $f_i(x_i)$ indica $f(x_i | \underset{\sim[i]}{x})$ em que $\underset{\sim[i]}{x} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_p)^T$.

$f_i(x_i)$ é chamada de distribuição condicional completa de x_i (*full conditional distribution*), $i = 1, \ldots, p$.

Para aplicar o amostrador, devemos ter como gerar amostras das distribuições condicionais completas.

Passos do amostrador

1. Contador de iterações: $j = 1$. Escolher valores iniciais $x_1^{(0)}, x_2^{(0)}, \ldots, x_p^{(0)}$;

2. Obtenha um novo vetor:

$$\underset{\sim}{x}^{(j)} = (x_1^{(j)}, x_2^{(j)}, \ldots, x_p^{(j)})^T \text{ a partir de } \underset{\sim}{x}^{(j-1)}$$

utilizando as distribuições condicionais completas, gerando:

$$\begin{aligned}
x_1^{(j)} &\sim f(x_1|x_2^{(j-1)}, \ldots, x_p^{(j-1)}) \\
x_2^{(j)} &\sim f(x_2|x_1^{(j)}, x_3^{(j-1)}, \ldots, x_p^{(j-1)}) \\
x_3^{(j)} &\sim f(x_3|x_1^{(j)}, x_2^{(j)}, \ldots, x_p^{(j-1)}) \\
&\vdots \\
x_p^{(j)} &\sim f(x_p|x_1, x_2, \ldots, x_{p-1})
\end{aligned}$$

3. Atualize o contador de iterações de j para $j+1$ e retorne ao passo 2 até que o processo convirja.

Exemplo:

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N_2\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \right),$$

sendo que $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$ e σ_{12} são conhecidos.

Para implementar o amostrador, necessitamos das distribuições condicionais:

$$X_1|X_2 = x_2 \qquad \text{e} \qquad X_2|X_1 = x_1$$

Pode ser provado que:

$$X_1|X_2 = x_2 \sim N\left(\mu_1 + \frac{\sigma_{12}}{\sigma_2}(x_2 - \mu_2); \sigma_1^2 - \frac{\sigma_{12}^2}{\sigma_2^2}\right)$$

e

$$X_1|X_2 = x_2 \sim N\left(\mu_2 + \frac{\sigma_{12}}{\sigma_1}(x_1 - \mu_1); \sigma_2^2 - \frac{\sigma_{12}^2}{\sigma_1^2}\right)$$

Tomamos $\mu_1 = 2, \mu_2 = 1, \sigma_1^2 = 3, \sigma_2^2 = 1, \sigma_{12} = \frac{1}{2}$

```
R <- 10000
mi1 <- 2
mi2 <- 1
sig1 <- sqrt(3)
sig2 <- 1
sig12 <- 0.5
x1 <- x2 <- c()
x1[1] <- 37
x2[1] <- 35
sigx1 <- sqrt(sig1^2-sig12^2/sig2^2)
sigx2 <- sqrt(sig2^2-sig12^2/sig1^2)
```

```
for(j in 1:R) {
  x1[j+1] <- rnorm(1, mi1+sig12*(x2[j]-mi2)/sig2), sigx1)
  x2[j+1] <- rnorm(1, mi2+sig12*(x1[j+1]-mi1)/sig1, sigx2)
}

plot(x1[-1], type="l", xlab = "iteracao")
plot(x2[-1], type="l", xlab = "iteracao")
plot(x1[-1], x2[-1], pch = 20)
#retirar as primeiras 100 observacoes => x1[-(1:100]
hist(x1[-(1:100)], freq = F, main = "")
lines(density(x1[-(1:100)]), col = "red")
acf(x1[-(1:100)])
```

(este não é o método mais eficiente para gerar amostras da distribuição normal).

O amostrador de Gibbs gera uma cadeia de Markov que, em condições bastante gerais, converge para a distribuição de $\underset{\sim}{X}$.

O amostrador é válido quando as distribuições condicionais completas $f_i(x_i), i = 1, \ldots, p$ determinam a distribuição conjunta $f(\underset{\sim}{x})$. Pode ser provado que esta condição é em geral (mas nem sempre) satisfeita.

Calculamos:

$$\begin{aligned} f_i(x_i) &= f(x_i|\underset{\sim[i]}{x}) \\ &= \frac{f(x_i, \underset{\sim[i]}{x})}{f(\underset{\sim[i]}{x})} \\ &= \frac{f(\underset{\sim}{x})}{f(\underset{\sim[i]}{x})} \propto f(\underset{\sim}{x}) \end{aligned}$$

Para obter $f_i(x_i)$ analisamos a ~~expansão~~ de $f(\underset{\sim}{x})$ e separamos a parte que envolve x_i, que é chamada de núcleo (*kernel*) de $f_i(x_i)$.

Tentamos associar o núcleo de $f_i(x_i)$ com uma distribuição conhecida.

Exemplo: $X \sim Poisson(\theta)$, temos que: $f(x) = P(X = x) = \dfrac{e^{-\theta}\theta^x}{x!}$, e o núcleo de $f(x)$ é $\dfrac{\theta^x}{x!}$.

~~**Obs.:** No teste de permutação para o teste global com regressão múltipla (Seção 5.3) a estatística de teste é:~~

$$~~F = \frac{SQReg/(p-1)}{SQRes/(n-p)}~~$$

Nos métodos MCMC recomenda-se um certo número de iterações iniciais (*burn-in*) antes que se considere que as cadeias convergiram. As iterações seguintes são usadas para obter os resultados.

Para minimizar a correlação entre as observações pode ser necessário tomar um espaçamento (*thinning*) entre as iterações.

Devemos gerar uma amostra de tamanho M. As primeiras M_0 iterações são descartadas e tomamos um espaçamento igual a $k, k \geq 1$. O número de iterações é igual a $M_0 + kM$

Existem estatísticas para testar convergência. Utilizamos também o gráfico das observações geradas.

Recomenda-se simular mais de uma cadeia com valores iniciais afastados entre si.

Exemplo: $Y_1, \ldots, Y_n$ são valores observados. A distribuição de $(X_1, X_2)^T$ tem função densidade, tal que:

$$f(x_1, x_2) \propto x_2^{\frac{n}{2}+1} \exp\left\{ -\frac{x_2}{2} \sum_{i=1}^{n} (y_i - x_i)^2 \right\} \exp\left\{ -\frac{x_1^2}{2} \right\} \exp\left\{ -\frac{x}{2} \right\},$$

em que $x_1 \in \mathbb{R}$ e $x_2 \in (0, \infty)$.

Devemos gerar uma amostra da distribuição de $(X_1, X_2)^T$

Iniciamos obtendo as distribuições condicionais completas. Temos que:

$$\begin{aligned} f_2(x_2) &= f(x_2|x_1) \\ &\propto x_2^{\frac{n}{2}+1} \exp\left\{ -x_2 \left[1 + \frac{\sum_{i=1}^{n} (Y_i - x_1)^2}{2} \right] \right\} \\ &= x_2^{(\frac{n}{2}+2)-1} \exp\left\{ -x_2 \left[1 + \frac{\sum_{i=1}^{n} (Y_i - x_1)^2}{2} \right] \right\} \end{aligned}$$

Portanto,

$$x_2|x_1 \sim Gama\left(forma = \frac{n}{2} + 2, taxa = 1 + \frac{\sum_{i=1}^{n} (Y_i - x_1)^2}{2} \right)$$

Temos que:

$$f_1(x_1) \propto \exp\left\{ -\frac{x_1^2}{2} - \frac{x_2}{2} \sum_{i=1}^{n} (Y_i - x_1)^2 \right\}$$

Pode ser provado que

$$X_1|X_2 = x_2 \sim N\left(\frac{x_2 \sum_{i=1}^{n} Y_i}{1 + nx_2}; \frac{1}{1 + nx_2} \right)$$

Médias ergódicas

$$\overline{X}_m^j = \frac{1}{j} \sum_{a=1}^{j} X_{a,m}, j = 1, \ldots, R.$$

6.2 Algoritmo de Metropolis-Hastings

O algoritmo de Hastings (1970) generaliza o algoritmo de Metropolis e outros (1953).

$\underset{\sim}{X}$ é um vetor aleatório $p \times 1, p \geq 1$. O algoritmo gera uma cadeia de Markov que converge, em condições bastante gerais, para a distribuição de $\underset{\sim}{X}$. O valor de $\underset{\sim}{X}$ na iteração j é $\underset{\sim}{x}^{(j)}$.

O valor seguinte, $\underset{\sim}{x}^{(j+1)}$, é gerado a partir de um valor candidato (*candidate*) $\underset{\sim}{y}$ obtido de uma distribuição proposta (*proposal*) $q(\underset{\sim}{y}, \underset{\sim}{x}^{(j)})$, que deve ser escolhida de forma cuidadosa. O candidato $\underset{\sim}{y}$ é aceito com probabilidade igual a:

$$\alpha = \alpha(\underset{\sim}{x}^{(j)}, \underset{\sim}{y}) = \min\left(1, \frac{f(\underset{\sim}{y}) q(\underset{\sim}{x}^{(j)}, \underset{\sim}{y})}{f(\underset{\sim}{x}^{(j)}) q(\underset{\sim}{y}, \underset{\sim}{x}^{(j)})} \right)$$

Se o candidato é aceito, tomamos:

$$\underset{\sim}{x}^{(j+1)} = \underset{\sim}{y}; \text{ caso contrário, } \underset{\sim}{x}^{(j+1)} = \underset{\sim}{x}^{(j)}$$

O algoritmo pode ser aplicado sem que se conheça $f(\underset{\sim}{x})$. Basta conhecermos que $f(\underset{\sim}{x}) \propto h(\underset{\sim}{x})$, de modo que:

$$\alpha = \min\left(1, \frac{h(y)q(\underset{\sim}{x}^{(j)}, \underset{\sim}{y})}{h(\underset{\sim}{x})q(\underset{\sim}{y}, \underset{\sim}{x}^{(j)})}\right)$$

Em muitas situações não é simples obter uma distribuição proposta para o vetor $\underset{\sim}{X}$, que é $p \times 1$. O algoritmo Metropolis-Hastings por componentes é bastante usado. O número máximo de componentes é p.

O candidato Y_i é gerado de uma distribuição proposta $q_i(Y_i, x_i^{(j)}, \underset{\sim[i]}{x}^{(j-1)})$. A probabilidade de aceitação do candidato Y_i é:

$$\begin{aligned}\alpha_i &= \alpha_i(\underset{\sim[i]}{x}^{(j)}, x_i^{(j)}, y_i) \\ &= \min\left(1, \frac{f(y_i|x_{[i]}^{(j)})q_i(x_i^{(j)}, y_i, x_{[i]}^{(j)})}{f(x_i^{(j)}|x_{[i]}^{(j)})q_i(y_i, x_i^{(j)}, x_{[i]}^{(j)})}\right)\end{aligned}$$

Notar que $f(x_i^{(j)}|x_{[i]}^{(j)})$ representa a distribuição condicional completa de $x_i^{(j)}$.

Quando as distribuições propostas são as distribuições condicionais completas, obtemos:

$$\alpha_i = \min(1, 1) = 1$$

e o algoritmo de Metropolis-Hastings se reduz ao amostrador de Gibbs.

Em vários problemas com $p > 1$ componentes de $\underset{\sim}{X}$, é possível gerar amostras usando as condicionais completas para alguns componentes de $\underset{\sim}{X}$ e para os demais componentes, aplicamos o algoritmo de Metropolis-Hastings. Este algoritmo é conhecido como amostrador de Gibbs com passos de Metropolis-Hastings (*Metropolis within Gibbs*).

Obs. 1: A distribuição proposta deve ser escolhida de modo que seja possível gerar amostras por um método conhecido.

Obs. 2: A distribuição proposta deve ser escolhida de modo que a taxa de aceitação de candidatos pertença ao intervalo $[0,15; 0,5]$ (recomendação).

Obs. 3: É comum que a distribuição proposta tenha média igual a $\underset{\sim}{x}^{(j)}$ ou $\underset{\sim[i]}{x}^{(j)}$ e um parâmetro adicional para controlar a variância (e portanto, controla a taxa de aceitação dos candidatos gerados).

Obs. 4: Tomamos ~~$p = 1$~~ e ~~uma~~ distribuição proposta tal que $q(\underset{\sim}{y}, \underset{\sim}{x}^{(j)}) = q(\underset{\sim}{x}^{(j)}, \underset{\sim}{y})$.

Temos $f(\underset{\sim}{x}) = kh(\underset{\sim}{x})$, k desconhecido.

Neste caso,

$$\alpha = \min\left(1, \frac{f(y)}{f(x^{(j)})}\right) = \min\left(1, \frac{h(y)}{h(x^{(j)})}\right)$$

Se o candidato gerado é y, ~~então~~ $\frac{h(y)}{h(x^{(j)})} > 1, \alpha = 1$ e $x^{(j+1)}$ recebe y.

Se o candidato gerado é y^*, ~~então~~ $\frac{h(y^*)}{h(x^{(j)})} < 1, \alpha = \frac{h(y^*)}{h(x^{(j)})}$.

Nesse caso a probabilidade de aceitar o candidato y^* é igual a $\alpha(< 1)$. Se o candidato for aceito, $x^{(j+1)}$ recebe y^*, caso contrário, $x^{(j+1)}$ recebe $x^{(j)}$.

7 Estimação de parâmetros usando métodos numéricos

Métodos de estimação (por exemplo, mínimos quadrados, máxima verossimilhança) muitas vezes necessitam de métodos iterativos.

7.1 Método de Newton

Encontrar o estimador de máxima verossimilhança (EMV) de $\underset{\sim}{\theta}$ $(p \times 1)$ com base em observações de uma amostra $\underset{\sim}{X}_1, \ldots, \underset{\sim}{X}_n$.

A função verossimilhança é denotada por $\mathcal{L}(\underset{\sim}{\theta}; \underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n)$ ou $\mathcal{L}(\underset{\sim}{\theta})$, com logaritmo $\ell(\underset{\sim}{\theta})$.

O vetor de derivadas primeiras de $\ell(\underset{\sim}{\theta})$ em relação a $\underset{\sim}{\theta}$ é denotado por $U(\underset{\sim}{\theta})$ (escore), com elementos:

$$\frac{\partial \ell(\underset{\sim}{\theta})}{\partial \theta_1}, \ldots, \frac{\partial \ell(\underset{\sim}{\theta})}{\partial \theta_p}.$$

A matriz de derivadas segundas (hessiana) é denotada por $H(\underset{\sim}{\theta})$, que é uma matriz simétrica eom elemento geral:

$$\frac{\partial^2 \ell(\underset{\sim}{\theta})}{\partial \theta_j \partial \theta_m}$$

com dimensão $p \times p$.

A aproximação em série de Taylor de segunda ordem da função $\ell(\underset{\sim}{\theta})$ em torno de $\underset{\sim}{\hat{\theta}}^{(k)}$ é dada por:

$$\ell(\underset{\sim}{\theta}) \approx \ell(\underset{\sim}{\hat{\theta}}^{(k)}) + \underset{\sim}{U}(\underset{\sim}{\hat{\theta}}^{(k)})^T(\underset{\sim}{\theta} - \underset{\sim}{\hat{\theta}}^{(k)}) + \frac{1}{2}(\underset{\sim}{\theta} - \underset{\sim}{\hat{\theta}}^{(k)})^T \underset{\sim}{H}(\underset{\sim}{\hat{\theta}}^{(k)})(\underset{\sim}{\theta} - \underset{\sim}{\hat{\theta}}^{(k)}),$$

devemos obter $\underset{\sim}{\theta}$ que ~~minimiza~~ a aproximação de $\ell(\underset{\sim}{\theta})$ (lado direito da expressão anterior)

Primeiro, obtemos um ponto crítico igualando a $\underset{\sim}{0}$ a derivada do lado direito em relação a $\underset{\sim}{\theta}$, ou seja,

$$\underset{\sim}{U}(\underset{\sim}{\hat{\theta}}^{(k)}) + \underset{\sim}{H}(\underset{\sim}{\hat{\theta}}^{(k)})(\underset{\sim}{\theta} - \underset{\sim}{\hat{\theta}}^{(k)}) = \underset{\sim}{0}.$$

Resolvendo, obtemos:

$$\begin{aligned}
\underset{\sim}{H}(\underset{\sim}{\hat{\theta}}^{(k)})(\underset{\sim}{\theta} - \underset{\sim}{\hat{\theta}}^{(k)}) &= -\underset{\sim}{U}(\underset{\sim}{\hat{\theta}}^{(k)} \\
\underset{\sim}{\theta} - \underset{\sim}{\hat{\theta}}^{(k)} &= [-\underset{\sim}{H}^{-1}(\underset{\sim}{\hat{\theta}}^{(k)})]\underset{\sim}{U}(\underset{\sim}{\hat{\theta}}^{(k)}) \\
\underset{\sim}{\hat{\theta}}^{(k+1)} &= \underset{\sim}{\hat{\theta}}^{(k)} + [-\underset{\sim}{H}^{-1}(\underset{\sim}{\hat{\theta}}^{(k)})]\underset{\sim}{U}(\underset{\sim}{\hat{\theta}}^{(k)}).
\end{aligned}$$

Nesse processo, $\underset{\sim}{\hat{\theta}}^{(0)}, \underset{\sim}{\hat{\theta}}^{(1)}, \ldots,$ são sucessivas atualizações, sendo que $\underset{\sim}{\hat{\theta}}^{(0)}$ é o ponto inicial.

Para $\underset{\sim}{\theta} = \underset{\sim}{\hat{\theta}}$ (EMV de θ), temos que $\underset{\sim}{H}(\underset{\sim}{\hat{\theta}})$ é uma matriz definida negativa. Para $\underset{\sim}{\theta} = \underset{\sim}{\hat{\theta}}^{(k)}$, não podemos garantir que $\underset{\sim}{H}(\underset{\sim}{\hat{\theta}}^{(k)})$ é definida negativa.

O método de Newton não tem a propriedade ascendente, ou seja, não é possível garantir que $\ell(\underset{\sim}{\hat{\theta}}^{(k+1)}) > \ell(\underset{\sim}{\hat{\theta}}^{(k)})$.

Obs.: Supomos que $\underset{\sim}{H}(\underset{\sim}{\theta})$ é inversível.

Em condições bastante gerais, temos que $\mathbb{E}[\underset{\sim}{U}(\underset{\sim}{\theta})] = \underset{\sim}{0}$ e $Cov(\underset{\sim}{U}(\underset{\sim}{\theta})) = \mathbb{E}[-\underset{\sim}{H}(\underset{\sim}{\theta})] = \underset{\sim}{I}_F(\underset{\sim}{\theta})$, que é a matriz de Fisher.

$\underset{\sim}{I}_F(\theta)$ é uma matriz definida positiva, logo $-\underset{\sim}{I}_F(\theta)$ é definida negativa, para todo θ no espaço paramétrico.

O método do escore de Fisher tem atualização dada por:

$$\underset{\sim}{\hat{\theta}}^{(k+1)} = \underset{\sim}{\hat{\theta}}^{(k)} + [\underset{\sim}{I}_F(\underset{\sim}{\hat{\theta}}^{(k)})]^{-1}\underset{\sim}{U}(\underset{\sim}{\hat{\theta}}^{(k)})$$

O passo $[\underset{\sim}{I}_F(\underset{\sim}{\hat{\theta}}^{(k)})]^{-1}\underset{\sim}{U}(\underset{\sim}{\hat{\theta}}^{(k)})$ representa um movimento em uma direção em que $\ell(\underset{\sim}{\theta})$ aumenta de valor (propriedade ascendente). O tamanho do passo deve ser controlado.

Na iteração k tomamos $\lambda^{(k)} = 1$

(∗) Se,

$$\ell(\hat{\underset{\sim}{\theta}}^{(k)}) + \lambda^{(k)}[\underset{\sim}{I}_F(\hat{\underset{\sim}{\theta}}^{(k)})]^{-1}\underset{\sim}{U}(\hat{\underset{\sim}{\theta}}^{(k)}) > \ell(\hat{\underset{\sim}{\theta}}^{(k)})$$

então,

$$\hat{\underset{\sim}{\theta}}^{(k+1)} = \hat{\underset{\sim}{\theta}}^{(k)} + \lambda^{(k)}[\underset{\sim}{I}_F(\hat{\underset{\sim}{\theta}}^{(k)})]^{-1}\underset{\sim}{U}(\hat{\underset{\sim}{\theta}}^{(k)}) > \ell(\hat{\underset{\sim}{\theta}}^{(k)}),$$

caso contrário, fazemos $\lambda^{(k)} = \dfrac{\lambda^{(k)}}{2}$ e retornamos ao ponto (∗).

Obs.: $\underset{\sim}{U}(\hat{\underset{\sim}{\theta}}) = \underset{\sim}{\theta}$.

7.2 Métodos ~~de~~ quase Newton

No método de Newton, a ~~informação observada~~ $\underset{\sim}{H}(\underset{\sim}{\theta})$ é calculada no ponto $\underset{\sim}{\theta} = \hat{\underset{\sim}{\theta}}^{(k)}$. O método requer cálculo de derivadas segundas de $\ell(\underset{\sim}{\theta})$ em relação a $\underset{\sim}{\theta}$. Este cálculo pode ser trabalhoso.

Nos métodos quase Newton, a ~~informação observada~~ é substituída por uma matriz $\underset{\sim}{A}$ mais simples de ser calculada.

Em um destes métodos, a matriz $\underset{\sim}{A}$ é dada por:

$$\underset{\sim}{A}^{(k+1)} = \underset{\sim}{A}^{(k)} - C_k\underset{\sim}{U}_k\underset{\sim}{U}_k^T, \text{ em que } C_k\underset{\sim}{U}_k\underset{\sim}{U}_k^T \text{ é uma matriz de posto 1,}$$

em que C_k é um escalar e $\underset{\sim}{U}_k$ é um vetor $p \times 1$. C_k e $\underset{\sim}{U}_k$ dependem apenas das derivadas primeiras de $\ell(\underset{\sim}{\theta})$. $A^{(0)}$ pode ser a matriz identidade de ordem p. Pode ser provado que $\underset{\sim}{A}^{(k)}$ converge para $\underset{\sim}{H}(\hat{\underset{\sim}{\theta}})$ quando $k \to \infty$.

Existem métodos quase Newton com atualização da matriz usando uma matriz de posto igual a 2. Um dos mais utilizados é o método *BFGS*.

O método *BFGS* está implementado em R na função *optim*. Um dos argumentos da função é o vetor de derivadas primeiras ~~de~~ $\underset{\sim}{U}(\underset{\sim}{\theta})$ (gradiente). Se o gradiente não é informado, o vetor gradiente é calculado numericamente por um método de diferenças finitas.

Obs. 1: A função *optim* é usada para obter o ponto de mínimo de uma função.

Obs. 2: $h : \mathbb{R}^2 \to \mathbb{R}$ é uma função diferenciável. Calcular $\dfrac{\partial h(x_1, x_2)}{\partial x_1}$ no ponto (x_{10}, x_2), temos:

$$\left.\frac{\partial h(x_1, x_2))}{\partial x_1}\right|_{x_1 = x_{10}} = \lim_{\Delta x_1 \to 0} \frac{h(x_{10} + \Delta x_1, x_2) - h(x_{10}, x_2)}{\Delta x_1}.$$

Tomando ~~δ~~x_1 com valor "pequeno" obtemos a aproximação:

$$\frac{h(x_{10} - \Delta x_1, x_2) - h(x_{10}, x_2)}{\Delta x_1} \text{ ou } \frac{h(x_{10} + \Delta x_1, x_2) - h(x_{10} - \Delta x_1, x_2)}{2\Delta x_1},$$

o pacote *numDeriv* em R pormite calcular aproximações numéricas para derivadas.

Existem aproximações baseadas em diferentes valores de Δx_1. Por exemplo, valores de Δx_1 tais que $a_1 > a_2 > \ldots > a_r$.

Obs. 1: A função *fitdistr* usa o método *BFGS* para obter as estimativas de máxima verossimilhança (*MASS*).

Obs. 2: O método *L-BFGS-B* permite restringir o espaço paramétrico. Por exemplo, se um parâmetro for uma variância o intervalo é $(0, \infty)$.

Para a distribuição normal, log-normal, geométrica, exponencial e Poisson, o EMV é calculado de forma exata, assim como o erro padrão.

Exemplo: $X \sim Poisson(\lambda)$, $\quad EMV = \hat{\lambda} = \overline{X}$.

Temos: $Var(\hat{\lambda}) = Var(\overline{X}) = \dfrac{Var(X)}{n} = \dfrac{\lambda}{n}$, de modo que o erro padrão da estimativa é calculado por:

$$\sqrt{\frac{\hat{\lambda}}{n}} = \sqrt{\frac{\overline{X}}{n}}.$$

Para todas as outras distribuições as estimativas de MV dos parâmetros são calculadas de forma iterativa com a função *optim*.

Os erros padrão das estimativas são baseadas na matriz de informações observada que é calculada por aproximação numérica.

Para a maioria das distribuições, é necessário fornecer valores iniciais para as estimativas dos parâmetros:

$$start = list(p1 =, p2 =, etc.).$$

O primeiro argumento da função *fitdistr* é o vetor com os dados. O segundo argumento especifica a distribuição, pode ser um nome ("beta", "lognormal", etc.) ou o nome de uma função densidade (dbeta, dmydensity).

Exemplo:

```
library(MASS)
dnova <- function(t1, t2) {
...
}
a1 <- fitdistr(dados, "gamma")
a2 <- fitdistr(dados, dnova, start = list(t1 = 0.5, t2 = 1))
```

limites superiores e inferiores para os parâmetros podem ser fixados.

$$lower = c(0, 0), \qquad upper = c(1, inf).$$

Neste caso, o método *L-BFGS-B* é usado.

8 O algoritmo EM

O método de MV é o método de estimação mais utilizado. O algoritmo EM é um método iterativo para obtenção de estimativas de MV de parâmetros.

O algoritmo tem dois passos: E de esperança e M de maximização.

Foi sistematizado por Dempster, Laird e Rubin (1977). O conceito fundamental para o algoritmo EMé dado faltante (*missing data*).

Para o algoritmo EM, dado faltante tem um significado mais geral. Dado faltante pode significar uma variável auxiliar não observada, mas que facilita a maximização da função verossimilhança.

Os dados observados (ou incompletos) são denotados por Y (pode ser um vetor). A distribuição de Y tem parâmetro (θ) (pode ser um vetor).

Os dados faltantes são denotados por Z. Os "dados" completos são denotados por X, sendo que $X = (Y, Z)$. Os dados são aumentados (*data augmentation*).

O algoritmo EM é mais recomendados quando a função verossimilhança usando X é mais simples do que a função verossimilhança usando Y.

A função densidade ou a função massa de probabilidade de X é denotada por $f_c(x;\theta) = f_c(y, z;\theta)$. Analogamento, definimos $f(y_i;\theta)$.

A relação entre f_c e f é dada por:

$$f(y;\theta) = \int_{s_z} f_c(y,z;\theta)dz, \text{ em que } s_z \text{ denota o espaço amostral de } z.$$

Se $\hat{\theta}^{(k)}$ é a estimativa de θ na iteração k do algoritmo, os passos são os seguintes:

Passo E: Calculamos:

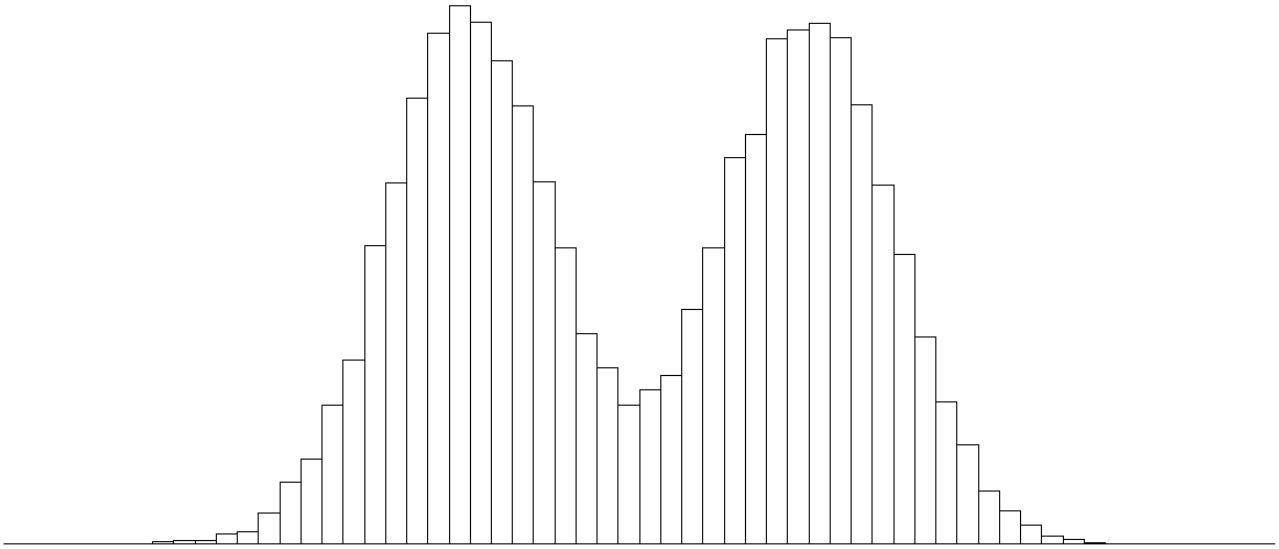
$$Q(\theta;\hat{\theta}^{(k)}) = \mathbb{E}[\log(f_c(y;z;\theta)|Y,\hat{\theta}^{(k)})],$$

sendo que no cálculo da esperança, θ é substituído por $\hat{\theta}^{(k)}$.

Obs.: A função Q é a função de Y com parâmetro θ.

Passo M: Maximizamos $Q(\hat{\theta};\hat{\theta}^{(k)})$ em relação a θ, obtendo $\hat{\theta}^{(k+1)}$. Repetimos o passo E e M até convergir começando com $\theta^{(0)}$.

Exemplo: Dados foram coletados e o seguinte histograma foi observado:



Os dados sugerem uma mistura de duas distribuições normais, ou seja,

$$Y_1 \sim N(\mu_1, \sigma_1^2) \qquad \text{e} \qquad Y_2 \sim N(\mu_2, \sigma_2^2).$$

Definimos $Z \in \{0,1\}$ com $P(Z=1) = \alpha$ e $P(Z=0) = 1-\alpha$, sendo que $z=0$ representa o grupo 1, e $z=1$ representa o grupo 2. A distribuição dos dados Y é dado por:

$$Y = (1-Z)Y_1 + ZY_2,$$

não observamos z (dado faltante).

A função densidade de Y é dada por:

$$f(y) = (1-\alpha)f_1(y;\mu_1,\sigma_1^2) + \alpha f_2(y;\mu_2,\sigma_2^2).$$

Obs.: $\int f(y)dy = 1$ e $f(y) \geq 0$.

O vetor de parâmetros é:

$$\theta = (\alpha, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2).$$

As observações são $y_1, \ldots, y_n$. A função log-verossimilhança é:

$$\ell(\theta; y) = \sum_{i=1}^{n} \log\Big((1-\alpha) f_1(y_i; \mu_1, \sigma_1^2) + \alpha f_2(y_i; \mu_2, \sigma_2^2)\Big).$$

A maximização de $\ell(\theta; y)$ não é simples.

Os dados faltantes são $z_1, \ldots, z_n$. Escrevemos:

$$f_c(y, z; \theta) = \begin{cases} f_1(y_i, \mu_1, \sigma_1^2), & \text{se } z = 0, \\ f_2(y_i, \mu_2, \sigma_2^2), & \text{se } z = 1. \end{cases}$$

Levando em conta a distribuição de Z, escrevemos:

$$L_c(\theta; y, z) = \prod_{i=1}^{n} \Big\{(1-\alpha) f_1(y_i; \mu_1, \sigma_1^2)\Big\}^{1-z_i} \Big\{\alpha f_2(y_i; \mu_2, \sigma_2^2)\Big\}^{z_i},$$

de modo que:

$$\begin{aligned} \ell(\theta; y, z) &= \sum_{i=1}^{n} \Big\{(1-z_i)\log((1-\alpha) f_1(y_i; \mu_1, \sigma_1^2)) + z_i \log(\alpha f_2(y_i; \mu_2, \sigma_2^2))\Big\} \\ &= \sum_{i=1}^{n} (1-z_i)\log(f_1(y_i; \mu_1, \sigma_1^2)) + z_i \log(f_2(y_i; \mu_2, \sigma_2^2)) \\ &\qquad + \sum_{i=1}^{n} (1-z_i)\log(1-\alpha) + z_i \log(\alpha) \end{aligned}$$

Devemos calcular:

$$Q(\theta, \hat{\theta}^{(k)}) = \mathbb{E}[\ell_c(\theta; Y, Z) | Y, \hat{\theta}^{(k)}]$$

Calculamos:

$$\begin{aligned} \gamma_i(\theta) = \mathbb{E}[Z_i | Y, \theta] &= P(Z_i = 1 | Y, \theta) = \\ &= \frac{\alpha f_2(y_i; \mu_2, \sigma_2^2)}{(1-\alpha) f_1(y_i; \mu_1, \sigma_1^2) + \alpha f_2(y_i; \mu_2, \sigma_2^2)} \end{aligned}$$

correspondente ao passo E.

Função log-verossimilhança para os dados completos

$$\ell_c(\theta; y, z) = \sum_{i=1}^{n} \Big\{(1-z_i)\log(f_1(y_i; \mu_1, \sigma_1^2)) + z_i \log(f_2(y_i, \mu_2, \sigma_2^2))\Big\} + \sum_{i=1}^{n} \Big\{(1-z_i)\log(1-\alpha) + z_i \log(\alpha)\Big\}$$

Se Z fosse conhecido, teríamos:

$$\hat{\mu}_1 = \frac{\sum_{i=1}^{n}(1-z_i)Y_i}{\sum_{i=1}^{n}(1-z_i)} : \text{ média ponderada de } Y \text{ com peso } 1 - Z$$

$$\text{e } \sigma_1^2 = \frac{\sum_{i=1}^{n}(1-z_i)(Y_i - \hat{\mu}_i)^2}{\sum_{i=1}^{n}(1-z_i)},$$

z_i é substituído por $\gamma_i(\theta) = \mathbb{E}[Z_i | Y, \theta]$.

Início

Valores inciais $\hat{\alpha}^{(0)}, \hat{\mu}_1^{(0)}, \hat{\sigma}_1^{(0)}, \hat{\mu}_2^{(0)}, \hat{\sigma}_2^{(0)}$

Passo E

$$\hat{\gamma}_i^{(k)} = \hat{\gamma}_i(\hat{\theta}) = \frac{\alpha^{(k)} f_2(y_i; \hat{\mu}_2^{(k)}, \hat{\sigma}_2^{2(k)})}{\{(1-\hat{\alpha}^{(k)}) f_1(y_i; \hat{\mu}_1^{(k)}, \hat{\sigma}_1^{2(k)}) + \hat{\alpha}^{(k)} f_2(y_i; \hat{\mu}_2^{(k)}, \hat{\sigma}_2^{2(k)})\}}$$

Passo M

$$\hat{\mu}_1^{(k+1)} = \frac{\sum_{i=1}^n (1-\hat{\gamma}_i^{(k)}) Y_i}{\sum_{i=1}^n (1-\hat{\gamma}_i^{(k)})}; \qquad \sigma_1^{2(k+1)} = \frac{\sum_{i=1}^n (1-\hat{\gamma}_i^{(k)})(Y_i - \hat{\mu}_1^{(k+1)})^2}{\sum_{i=1}^n (1-\hat{\gamma}_i^{(k)})},$$

$$\hat{\mu}_2^{(k+1)} = \frac{\sum_{i=1}^n \hat{\gamma}_i^{(k)} Y_i}{\sum_{i=1}^n \hat{\gamma}_i^{(k)}}; \qquad \sigma_2^{2(k+1)} = \frac{\sum_{i=1}^n \hat{\gamma}_i^{(k)} (Y_i - \hat{\mu}_2^{(k+1)})^2}{\sum_{i=1}^n (1-\hat{\gamma}_i^{(k)})} \text{ e}$$

$$\hat{\alpha}^{(k+1)} = \frac{\sum_{i=1}^n \hat{\gamma}_i^{(k)}}{n}.$$

Repetir os passos E e M até convergir.

Um critério de convergência pode ser baseado na diferença entre $\hat{\theta}^{(k+1)}$ e $\hat{\theta}^{(k)}$.

Obs. 1: A convergência do algoritmo EM pode necessitar de muitas iterações.

Obs. 2: No passo E deve ser calculada uma esperança condicional. Este cálculo pode ser aproximado por integração de Monte Carlo. Chamamos de algoritmo EM Monte Carlo.

Propriedade ascendente

$$\ell(\hat{\theta}^{(k+1)}) \geq \ell(\hat{\theta}^{(k)}), k = 0, 1, 2, \ldots.$$

Exemplo: Detecção de cliques fraudulentos na internet.

Variáveis observadas:

- C_i : número de cliques do usuário i em um certo intervalo de tempo;
- P_{ij} : número de páginas vistas (PV) no $j-$ésimo clique, $j = 1, \ldots, C_i (P_{ij} \geq 1)$;
- Δ_{ij} : diferença de tempo entre o $j-$ésimo e o $(j+1)-$ésimo clique, para $j = 1, \ldots, C_{i-1}$.

Variável não observada:

$$Z_i = \begin{cases} 0, & \text{se o usuário é regular,} \\ 1, & \text{se o usuário é fraudulento,} \end{cases} \quad i = 1, 2, \ldots, n.$$

Temos $P(Z_i = 1) = \alpha$ e $P(Z_i = 0) = 1-\alpha$, em que $\alpha \in (0, 1)$. Supomos que P_{ij} e Δ_{ij} são v.a's independentes. Supomos que $C_{i-2} \sim Poisson(\lambda_k), P_{ij} \sim Poisson(\mu_k)$ e $\Delta_{ij} \sim Exp(\delta_k), k \in (\{0, 1\}$, dependendo do valor de $Z_i, i = 1, \ldots, n$.

Como consequência,

$$\Delta_i = \sum_{j=1}^{C_{i-1}} \Delta_{ij} \sim Gama(forma = c_{i-1}, escala = \delta_k) \text{ e } P_i | C_i \sim P_{01}(c_i \mu_k),$$

em que $P_i = \sum_{j=1}^{c_i} P_{ij}$.

Os dados observados são denotados por $Y_i = (c_i, P_i, \Delta_i), i = 1, \ldots, n$.

A função densidade condicional de Y_i dado $Z_i = k$ é denotada por f_k. Obtemos:

$$f_k(Y_i|\theta) = f_k(c_i|\theta) f_k(P_i|c_i, \theta) f_k(\Delta_i|c_i, \theta),$$

em que $\theta = (\alpha, \lambda_0, \lambda_1, \mu_0, \mu_1, \delta_0, \delta_1)$.

Aplicando o algoritmo EM, obtemos:

$$\hat{P}(Z_i = 1) = \frac{\hat{\alpha} f_1(D_i|\hat{\theta})}{\hat{\alpha} f_1(D_i|\hat{\theta}) + (1-\hat{\alpha}) f_0(D_i|\hat{\theta})}$$

Comportamentos:

1. Em um certo intervalo de tempo, o número de cliques gerado por um usuário mal intencionado é tipicamente grande.

2. A diferença de tempo entre cliques fraudulentos é tipicamente pequena.

O modelo foi ajustado a um conjunto de dados com $n = 1115$ usuários.
Outra notação:
(a) $P(Z_i = k) = \pi_k, k \in \{0, 1\}, (\pi_0 + \pi_1 = 1)$.
(b) $D = Y$.

Obs.: Problemas de aprendizagem não supervisionado não existe um conjunto de treinamento.