

Inferência Estatística

Luiz Max de Carvalho[lmax.fgv@gmail.com]

Disciplina da graduação em Matemática Aplicada
Escola de Matemática Aplicada (EMAp/FGV), Rio de Janeiro.

27 de Outubro de 2021

Bem-vindas (os)!

Este é um curso de 60 (sessenta) horas sobre Inferência Estatística.

Princípios:

- △ Em uma palavra: Liberdade;
- △ Construção conjunta do conhecimento;
- △ Pontualidade na entrega das tarefas;
- △ Participação em aula;

Burocracia:

- ☐ Horário de atendimento: Segundas e quartas de 13:30h a 14:00h.
 - ◇ Por favor, mandar e-mail com antecedência de 24h para marcar;
- ☐ Podem escrever por e-mail (ou carta) quando quiserem;
- ☐ Teremos duas avaliações ($A1$ e $A2$) e 4 (quatro) trabalhos (T_i , $i = 1, 2, 3, 4$).
Trabalhos valerão 20% do grau final.
- ☐ Sejam $N1 = A1 + T1 + T2$ e $N2 = A2 + T3 + T4$.
A **nota final** será $NF := (N1 + N2)/2$.

- Desigualdade de Markov;
- Desigualdade de Chebychev;
- Convergência;
- Lei(s) dos grandes números;
- Teorema(s) Central(is) do Limite;

Teorema 1 (Desigualdade de Markov)

Seja X uma variável aleatória não-negativa e $t > 0$. Então

$$\Pr(X \geq t) \leq \frac{E[X^n]}{t^n}. \quad (1)$$

Prova: Assumindo que X é absolutamente contínua, escrever $E[X]$ explicitamente e usar linearidade e monotonicidade da integral. Para $n = 1$ e X discreta, ver DeGroot, página 349, Teorema 6.2.1 \square

¹Em homenagem a Andrey Andreyevich Markov (1856–1922).

Teorema 2 (Desigualdade de Chebychev)

Seja Y uma variável aleatória com média $E[Y] =: \mu$ e variância $\text{Var}(Y) =: \sigma^2$, ambas finitas. Mais uma vez, $t > 0$. Então

$$\Pr(|Y - \mu| \geq t) \leq \frac{\text{Var}(Y)}{t^2}. \quad (2)$$

Prova: Notar que $E[(Y - \mu)^2] = \sigma^2$ e aplicar Markov. Ver DeGroot, página 349, Teorema 6.2.2 \square

²Em homenagem a Pafnuty Lvovich Chebyshev (1821–1894).

A média amostral

Considere uma **amostra aleatória** X_1, X_2, \dots, X_n , $n \in \mathbb{N}$ de variáveis aleatórias de uma mesma distribuição com média $E[X_i] = \mu$ e variância $\text{Var}(X_i) = \sigma^2$.

Definição 1 (Média amostral)

A *média amostral* de X_1, X_2, \dots, X_n é

$$\bar{X}_n := \frac{1}{n} \sum_{i=1}^n X_i. \quad (3)$$

Teorema 3 (Média e variância em uma amostra i.i.d.)

Sejam X_1, X_2, \dots, X_n variáveis aleatórias independentes e identicamente distribuídas, com média μ e variância σ^2 . Temos que (i) $E[\bar{X}_n] = \mu$ e (ii) $\text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$.

Prova: Para (i), usar a linearidade da esperança e o fato de as variáveis serem identicamente distribuídas – note a falta de menção à independência. Para (ii), usar a soma das variâncias de variáveis independentes, além do fato de serem identicamente distribuídas. Ver DeGroot, página 350, Teorema 6.2.3.

Exemplo: determinando o tamanho de amostra (1)

Vamos estudar o exemplo 6.2.3 de DeGroot. Suponha que uma moeda justa é lançada n vezes. Seja X_i a variável aleatória que é 1 se o i -ésimo lançamento dá cara e 0 caso contrário. Considere $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$.

Pergunta 1 (Quantos lançamentos?)

Quantos lançamentos devemos fazer para que

$$\Pr(0.4 \leq \bar{X}_n \leq 0.6) \geq 0.7 ?$$

Resolução: Primeiro, faça $S_n = \sum_{i=1}^n X_i$ e deduza que $E[S_n] = np = n/2$ e $\text{Var}(S_n) = np(1-p) = n/4$. Agora:

$$\Pr(0.4 \leq \bar{X}_n \leq 0.6) = \Pr\left(\frac{4n}{10} \leq S_n \leq \frac{6n}{10}\right).$$

Subtraia $E[S_n]$ dos dois lados da desigualdade para obter

$$\Pr\left(\frac{4n}{10} \leq S_n \leq \frac{6n}{10}\right) = \Pr\left(\left|S_n - \frac{n}{2}\right| \leq \frac{n}{10}\right).$$

Exemplo: determinando o tamanho de amostra (2)

Note que, usando a desigualdade de Chebychev, temos uma cota superior para

$$\Pr\left(\left|S_n - \frac{n}{2}\right| \geq \frac{n}{10}\right) = 1 - \Pr\left(\left|S_n - \frac{n}{2}\right| \leq \frac{n}{10}\right).$$

Portanto

$$\Pr\left(\left|S_n - \frac{n}{2}\right| \geq \frac{n}{10}\right) \leq \frac{100}{4n}$$

e então

$$\Pr(0.4 \leq \bar{X}_n \leq 0.6) = \Pr\left(\left|S_n - \frac{n}{2}\right| \leq \frac{n}{10}\right) \geq 1 - \frac{25}{n}.$$

Resolvendo $1 - 25/n = 0.7$ obtemos $n \geq 84$.

Exemplo: determinando o tamanho de amostra (3)

Agora, vamos usar o que sabemos **especificamente** sobre este problema. Usando uma tabela de probabilidades binomiais ou rodando um programa como:

```
calcula_p <- function(n){
  prob <- pbinom(q = round(0.6*n), size = n, p = .5)
  - pbinom(q = round(0.4*n)-1, size = n, p = .5)
  return(prob)
}

n <- 10
p <- calcula_p(n)
alvo <- 0.7
erro <- (p-alvo)^2
while(erro > .001){
  n <- n + 1
  p <- calcula_p(n)
  erro <- (p-alvo)^2
  if(n > 10000) break
}
```

Exemplo: determinando o tamanho de amostra (4)

obtemos $n = 15$ e $p = 0.6982422$.

Conclusão: a desigualdade de Chebychev é frouxa, isto é, ela dá uma cota superior para a probabilidade de interesse, mas essa cota pode ser muito maior do que o valor exato. Por outro lado, a desigualdade é válida para qualquer variável aleatória cuja variância exista e seja finita.

Ideia 1 (Sem almoço grátis)

Se uma técnica ou resultado é muito geral, isto é, se aplica a muitas situações, há grandes chances de não fornecer uma resposta muito precisa. O contrário também é verdadeiro: se desenvolvemos uma técnica elaborada para uma classe restrita de problemas, geralmente vamos obter respostas precisas, mas nossa técnica não será aplicável a muitos tipos de problemas. Em Estatística não existe almoço grátis.

Definição 2 (Convergência em probabilidade)

Dizemos que uma sequência de variáveis aleatórias converge em probabilidade para b se, para todo $\epsilon > 0$, temos

$$\lim_{n \rightarrow \infty} \Pr(|Z_n - b| < \epsilon) = 1.$$

Neste caso, escrevemos $Z_n \xrightarrow{P} b$.

Em algumas situações, chamamos a convergência em probabilidade de convergência fraca.

Lei(s) dos Grandes Números (LGN)

A lei fraca dos grandes números é um resultado fundamental da Teoria de Probabilidade, extremamente útil em Estatística.

Teorema 4 (Lei Fraca dos Grandes Números)

Sejam X_1, X_2, \dots, X_n variáveis aleatórias independentes e identicamente distribuídas, com média μ e variância σ^2 . Então

$$\bar{X}_n \xrightarrow{P} \mu.$$

Prova: Usando o teorema ?? e a desigualdade de Chebychev, temos

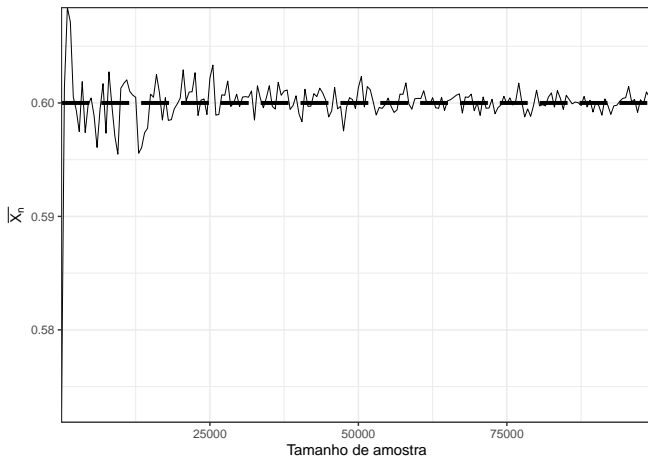
$$\Pr(|\bar{X}_n - \mu| < \epsilon) \geq 1 - \frac{\sigma^2}{n\epsilon^2},$$

e, portanto,

$$\lim_{n \rightarrow \infty} \Pr(|\bar{X}_n - \mu| < \epsilon) = 1. \quad \square$$

LGN: exemplo

$X_1, \dots, X_n \sim \text{Beta}(3, 2)$. $E[X] = \alpha/(\alpha + \beta) = 3/5$.



Definição 3 (Convergência quase certa)

Dizemos que uma sequência de variáveis aleatórias $(Z_n)_{n \geq 1}$ converge quase certamente para b se

$$\Pr \left(\lim_{n \rightarrow \infty} Z_n = b \right) = 1.$$

Esse modo de convergência é por vezes chamado de convergência forte.

Observação: convergência quase certa implica convergência em probabilidade.

Teorema 5 (Lei forte dos grandes números)

Sejam X_1, X_2, \dots variáveis aleatórias independentes e identicamente distribuídas, com média μ . Então

$$\Pr \left(\lim_{n \rightarrow \infty} \bar{X}_n = \mu \right) = 1.$$

Teorema(s) Central(is) do Limite

O Teorema Central do Limite um dos resultados mais importantes da Estatística.

Teorema 6 (Teorema Central do Limite (Lindeberg e Lévy)³)

Sejam X_1, X_2, \dots, X_n variáveis aleatórias independentes e identicamente distribuídas, com média μ e variância σ^2 . Então, para cada x , temos

$$\lim_{n \rightarrow \infty} \Pr \left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq x \right) = \Phi(x),$$

onde

$$\Phi(x) := \frac{1}{\sqrt{2\pi}} \int_0^x \exp \left(-\frac{t^2}{2} \right) dt,$$

é a função de distribuição (cumulativa) normal padrão.

Prova: Ver Casella & Berger (2002), página 237, teorema 5.5.14.

³Jarl Waldemar Lindeberg (1876–1932) e Paul Pierre Lévy (1886–1971).

Teorema Central do Limite: interpretação

- Sabemos que a variável aleatória padronizada $Y_n := (\bar{X}_n - \mu) / \sigma$ tem média 0 e variância 1, por construção;
- O teorema ?? nos diz que se tomamos uma amostra grande de uma distribuição com média μ e variância σ^2 , a variável aleatória $\sqrt{n}Y_n$ terá, aproximadamente, distribuição **normal** com média 0 e desvio padrão 1, chamada *distribuição normal padrão*;
- Isto equivale a dizer que $\bar{X}_n \sim \text{normal}(\mu, \sigma^2/n)$;
- Note que o teorema vale para qualquer variável aleatória cujos dois primeiros momentos existam, seja ela discreta ou contínua!

Teorema Central do Limite: aplicação

Pergunta 2 (Quanto vale p ?)

Suponha que X_1, \dots, X_{12} são variáveis aleatórias independentes com distribuição uniforme entre 0 e 1. Defina

$$p := \Pr \left(\left| \bar{X}_n - \frac{1}{2} \right| \leq 0.1 \right).$$

Quanto vale p ?

Resolução: Lembremos que a variável padronizada $Z = \sqrt{n}(\bar{X}_n - E[X])/\sqrt{\text{Var}(X)}$ terá distribuição aproximadamente normal padrão. Se $X \sim \text{uniforme}(0, 1)$, sabemos que $E[X] = 1/2$ e $\text{Var}(X) = 1/12$. Nos aproveitando do fato de que \sqrt{n} e σ coincidem nesse exemplo, escrevemos

$$\Pr \left(\left| \bar{X}_n - \frac{1}{2} \right| \leq 0.1 \right) = \Pr \left(12 \left| \bar{X}_n - \frac{1}{2} \right| \leq 0.1 \times 12 \right) = \Pr(|Z| < 1.2),$$

de modo que $p \approx \Phi(1.2) - \Phi(-1.2) = 0.7698607$. O valor exato, que não discutiremos como obter, é $p = 0.7667213$.

O que aprendemos?

- 💡 Desigualdades de Markov e Chebychev: extremamente gerais (mas não muito precisas!);
- 💡 Convergência fraca (convergência em probabilidade ou medida), $Z \xrightarrow{p} b$;
- 💡 Lei (fraca) dos grandes números: a média amostral converge para a média populacional à medida que a amostra aumenta, $\bar{X}_n \xrightarrow{p} \mu$;
- 💡 Teorema Central do Limite: para amostras grandes o suficiente,

$$\bar{X}_n \sim \text{normal}(\mu, \sigma^2/n).$$

Leitura recomendada

 DeGroot seções 6.2 e 6.3;

 * Casella & Berger, seções 5.2 e 5.5;

 * Nosso repositório
(https://github.com/maxbiostat/Statistical_Inference_BSc).

▶▶ Próxima aula: DeGroot, seção 7.1;

O que é e para que serve Inferência Estatística?

- ? Esta moeda é justa?
- ? Esta droga “funciona”?
- ? Quantos casos de Dengue teremos mês que vem?
- ? Renda básica universal aumenta o PIB?

Todas essas perguntas podem ser abordadas com as ferramentas que a Estatística nos fornece.

Ideia 2 (A gramática da Ciência)

A Estatística é a gramática da Ciência⁴. O mundo é incerto; medições são imperfeitas. A Estatística é a linguagem que nos permite expressar e quantificar as incertezas associadas às afirmações científicas através da teoria de probabilidades⁵.

⁴Título do livro de Karl Pearson (1857–1936) (“[The Grammar of Science](#)”), publicado em 1982.

⁵Chamada por E.T. Jaynes (1922–1998) de lógica da Ciência (“[Probability Theory: The Logic of Science](#)”).

Definição 4 (Modelo estatístico: informal)

DeGroot, def 7.1.1, pág. 377 Um modelo estatístico consiste na identificação de variáveis aleatórias de interesse (observáveis e potencialmente observáveis), na especificação de uma distribuição conjunta para as variáveis aleatórias observáveis e na identificação dos parâmetros (θ) desta distribuição conjunta. Às vezes é conveniente assumir que os parâmetros são variáveis aleatórias também, mas para isso é preciso especificar uma distribuição conjunta para θ .

Modelo estatístico: definição formal

Definição 5 (Modelo estatístico: formal)

McCullagh, 2002. Seja \mathcal{X} um espaço amostral qualquer, Θ um conjunto não-vazio arbitrário e $\mathcal{P}(\mathcal{X})$ o conjunto de todas as distribuições de probabilidade em \mathcal{X} . Um modelo estatístico paramétrico é uma função $P : \Theta \rightarrow \mathcal{P}(\mathcal{X})$, que associa a cada $\theta \in \Theta$ uma distribuição de probabilidade P_θ em \mathcal{X} .

Exemplos:

- Faça $\mathcal{X} = \mathbb{R}$ e $\Theta = (-\infty, \infty) \times (0, \infty)$. Dizemos que P é um modelo⁶ estatístico normal se para cada $\theta = \{\mu, \sigma^2\} \in \Theta$,

$$P_\theta(x) \equiv \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}.$$

- Faça $\mathcal{X} = \mathbb{N} \cup \{0\}$ e $\Theta = (0, \infty)$. P é um modelo estatístico Poisson se para $\lambda \in \Theta$,

$$P_\lambda(k) \equiv \frac{e^{-\lambda} \lambda^k}{k!}, \quad k = 0, 1, \dots$$

⁶Note o abuso de notação: estritamente falando, P_θ é uma **medida** de probabilidade e não uma *densidade* como apresentamos aqui.

Exemplo: como sempre, moedas.

Pergunta 3 (Esta moeda é justa?)

Suponha que uma moeda tenha sido lançada dez vezes, obtendo o seguinte resultado:

KKKCKCCCKC

- a) Esta moeda é justa?*
- b) Quanto eu espero ganhar se apostar R\$ 100,00 que é justa?*

Podemos formalizar o problema ao, por exemplo, assumir que cada lançamento é uma variável aleatória Bernoulli com probabilidade de cara (K), p . Desta forma $X_i = 1$ se o lançamento deu cara e $X_i = 0$ caso contrário. E queremos saber se $p = 1/2$. Por ora, não temos as ferramentas necessárias para responder a essa pergunta, mas voltaremos a ela no futuro.

Definição 6 (Afirmação probabilística)

Dizemos que uma afirmação é probabilística quando ela utiliza conceitos da teoria de probabilidade para falar de um objeto. Exemplos:

- $\Pr(\bar{Y}_n \in (0, 1)) \leq 2^{-n}$;
- $E[X \mid Y = y] = 2y + 3$;
- $\text{Var}(X) = 4p^2$.
- $\Pr(\text{Var}(X) \leq 4p^2) \leq p^2$

Definição 7 (Inferência Estatística)

Uma inferência estatística é uma afirmação probabilística sobre uma ou mais partes de um modelo estatístico. Considerando o exemplo ??, queremos saber:

- *Quantos lançamentos até termos 80% de certeza de que a moeda é justa?*
- *Quanto vale $E[\bar{X}_n]$;*
- $\Pr(X_n = 1 \mid X_{n-1} = 1)$.

Definição 8 (Estatística)

Suponha que temos uma coleção de variáveis aleatórias $X_1, X_2, \dots, X_n \in \mathbf{X} \subseteq \mathbb{R}^n$ e uma função $r : \mathbf{X} \rightarrow \mathbb{R}^m$. Dizemos que a variável aleatória $T = r(X_1, X_2, \dots, X_n)$ é uma **estatística**.

São exemplos de estatísticas:

- A média amostral, \bar{X}_n ;
- A soma, $\sum_{i=1}^n X_i$;
- O mínimo, $\min(X_1, X_2, \dots, X_n)$;
- $r(X_1, X_2, \dots, X_n) = a, \forall X_1, X_2, \dots, X_n, a \in \mathbb{R}$.

Tipos de Inferência Estatística

- **Predição:** prever o valor de uma variável aleatória (ainda) não observada; No exemplo ??, qual será o valor do próximo lançamento, X_{n+1} ;
- **Decisão Estatística:** Acoplamos o modelo estatístico a uma decisão a ser tomada. Devo emprestar esta moeda ao Duas-Caras? Aqui, temos a *noção* de **risco**.;
- **Desenho experimental:** Quantas vezes é preciso lançar esta moeda para ter 95% de certeza de que ela é (ou não) justa? Quantas pessoas precisam tomar uma droga para sabermos se ela funciona? Onde devemos cavar para procurar ouro/petróleo?;

O que aprendemos?

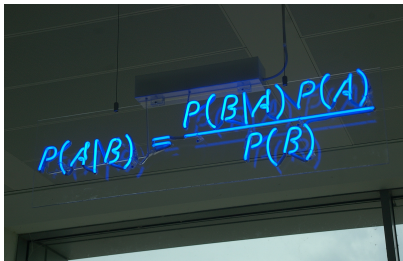
- 💡 Modelo estatístico;
- 💡 Inferência Estatística;
- 💡 Estatística (amostral);
- 💡 Tipos de inferências:
 - ◇ Predição;
 - ◇ Decisão;
 - ◇ Desenho experimental.

 DeGroot seção 7.1;

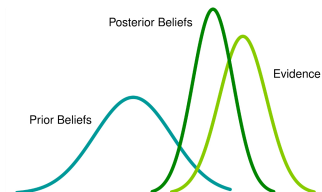
 * [McCullagh, 2002](#).

▶▶ Próxima aula: DeGroot, seção 7.2;

- Os paradigmas bayesiano e frequentista;
- Distribuição *a priori* e *a posteriori*;
- Função de verossimilhança;



A photograph of a chalkboard with the Bayes' theorem formula written in blue chalk. The formula is
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$



Definição 9 (Permutabilidade)

Permutabilidade. Uma coleção finita de variáveis aleatórias X_1, X_2, \dots, X_n com densidade conjunta f é dita **permutável** se

$$f(x_1, x_2, \dots, x_n) = f(x_{\pi(1)}, x_{\pi(2)}, \dots, x_{\pi(n)}),$$

para qualquer permutação $\pi = \{\pi(1), \pi(2), \dots, \pi(n)\}$ dos seus elementos. Uma coleção infinita é permutável se qualquer subconjunto finito é permutável.

- Note que uma amostra permutável não precisa ser independente;
- Note também que IID \implies permutável;
- A intuição é simples: simetria.

Exemplo 1

Ensaio Clínico (DeGroot, exemplo 7.1.3). Suponha que estamos interessados na taxa de recrudescência (“recaída”) de uma determinada doença entre pacientes tratados com uma droga. Seja X_i a variável aleatória que indica se o i -ésimo paciente recrudesceu ($X_i = 1$) ou não ($X_i = 0$). Seja P a proporção de indivíduos que recrudescem num grupo grande de pacientes. Se P é desconhecida, podemos modelar X_1, X_2, \dots como variáveis aleatórias Bernoulli IID com parâmetro p **condicional** a $P = p$. Em notação estatística:

$$X_1, X_2, \dots \mid P = p \sim \text{Bernoulli}(p).$$

Assuma que X_1, X_2, \dots é uma sequência permutável infinita. Agora chamemos de P_n a proporção de pacientes que recrudescem nos n primeiros pacientes. Podemos mostrar que o limite $\lim_{n \rightarrow \infty} P_n = \lim_{n \rightarrow \infty} \sum_{i=1}^n X_i / n$ existe com probabilidade 1 e que pode ser visto como a proporção P .

No Exemplo ?? podemos encarar o problema de duas maneiras:

- A) P é uma variável aleatória e X_1, X_2, \dots são Bernoulli(p) **condicional** ao evento $P = p$, $p \in (0, 1)$.
- B) Para uma constante fixa (e inobservável) p , X_1, X_2, \dots tem distribuição Bernoulli com parâmetro p – isto é, indexada por $p \in (0, 1)$.

Uma diferença *sutil*, não é? A tradição estatística que entende parâmetros como variáveis aleatórias como em A) é chamada de **Estatística bayesiana**⁷. Já os que aderem à abordagem B) são chamados **frequentistas** – ou ortodoxos, como Jaynes gosta de chamá-los. Neste curso veremos conceitos e exemplos destas duas escolas de pensamento.

⁷Em homenagem ao reverendo inglês Thomas Bayes (1701 – 1761).

Exemplo 2 (Duração de componentes)

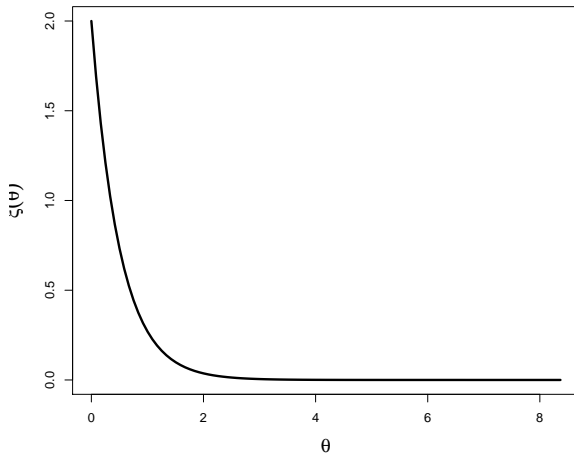
Duração de componentes eletrônicos (DeGroot, exemplo 7.2.1). Suponha que uma empresa esteja interessada em saber o quanto duram os produtos que ela produz. Se representamos os tempos de duração de n objetos como n variáveis aleatórias X_1, X_2, \dots, X_n IID com distribuição exponencial com parâmetro θ de modo que

$$f(x_i | \theta) = \theta \exp(-\theta x_i), x_i > 0.$$

Observação: $n / \sum_{i=1}^n X_i \xrightarrow{P} \theta$.

Aqui, θ é a taxa de falha dos componentes, e é um parâmetro de interesse.

Suponha que uma pessoa experiente na empresa diga que a taxa de falha é mais ou menos 0.5/ano. Como representamos esta informação?



A distribuição *a priori*

Definição 10 (Distribuição *a priori*)

*Se tratamos o parâmetro θ como uma variável aleatória, então a distribuição a priori, que também chamaremos simplesmente de priori, é a distribuição que damos a θ **antes** de observarmos as outras variáveis aleatórias de interesse. Em geral, vamos denotar a função de densidade/massa de probabilidade da priori por $\xi(\theta)$.*

Exemplos:

- Podemos dizer que a probabilidade de uma moeda cair cara, p , tem distribuição uniforme entre 0 e 1;
 - ◊ Ou que tem distribuição $\text{Beta}(2, 2)$;
- A altura média dos jogadores de basquete do CR Flamengo tem distribuição normal com média $\mu_0 = 200\text{cm}$ e variância $\sigma_0^2 = 25\text{cm}^2$;
- A posição de Júpiter em relação ao Sol hoje tem coordenadas X, Y, Z , de modo que $X \sim \text{Normal}(\mu_x, 1)$, $Y \sim \text{Normal}(\mu_y, 1)$, $Z \sim \text{Normal}(\mu_z, 1)$.

Distribuição *a posteriori*

Definição 11 (Distribuição *a posteriori*)

Considere o problema estatístico com parâmetro θ e variáveis aleatórias observáveis X_1, X_2, \dots, X_n . A distribuição condicional de θ dados os valores observados das variáveis aleatórias, $\mathbf{x} := \{x_1, x_2, \dots, x_n\}$ é a **distribuição *a posteriori*** de θ .

Denotamos por $\xi(\theta | \mathbf{x})$ a f.d.p./f.m.p. condicional a $X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$.

Teorema 7 (Distribuição *a posteriori*: derivação)

Considere a amostra aleatória X_1, X_2, \dots, X_n de uma distribuição com f.d.p./f.m.p. $f(\mathbf{x} | \theta)$. Se a distribuição *a priori* é $\xi(\theta)$, temos

$$\xi(\theta | \mathbf{x}) = \frac{\xi(\theta) \prod_{i=1}^n f(x_i | \theta)}{g_n(\mathbf{x})}, \quad \theta \in \Omega. \quad (4)$$

Chamamos $g_n(\mathbf{x})$ de distribuição marginal de X_1, X_2, \dots, X_n .

Prova: Usar a premissa de amostra aleatória para escrever $f(x_1, x_2, \dots, x_n | \theta)$, escrever a distribuição conjunta de θ e \mathbf{x} e computar $g_n(\mathbf{x})$ usando a lei da probabilidade total.

Distribuição *a posteriori*: exemplo

Continuando com o Exemplo ??, fica claro que

$$f(\mathbf{x} \mid \theta) = \prod_{i=1}^n f(x_i \mid \theta) = \theta^n \exp(-S\theta),$$

onde $S = \sum_{i=1}^n x_i$. Desta forma, temos

$$f(\mathbf{x} \mid \theta)\xi(\theta) = \theta^n \exp(-(S+2)\theta).$$

Para obter $g_n(\mathbf{x})$, computamos

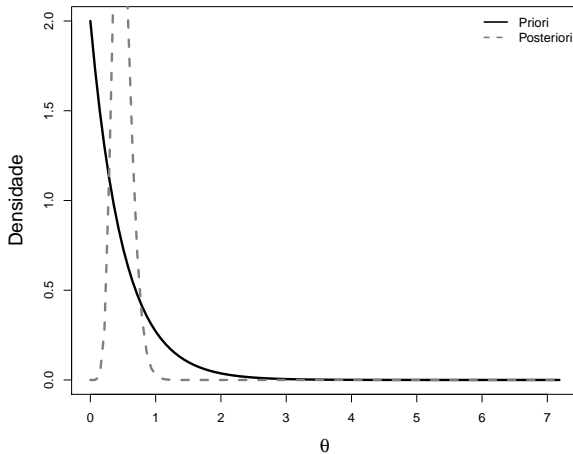
$$g_n(\mathbf{x}) = \int_0^\infty t^n \exp(-(S+2)t) dt = \frac{\Gamma(n+1)}{(S+2)^{n+1}}.$$

Concluimos que

$$\xi(\theta \mid \mathbf{x}) = \frac{(S+2)^{n+1}}{\Gamma(n+2)} \theta^{n+1} \exp(-(S+2)\theta),$$

ou seja, $\theta \mid \mathbf{x} \sim \text{Gama}(n+1, \sum_{i=1}^n x_i + 2)$.

Distribuição *a posteriori*: exemplo (cont.)



A função de verossimilhança

Note que o denominador em (??) não depende do parâmetro, θ . Deste modo, podemos escrever

$$\xi(\theta | \mathbf{x}) \propto f(\mathbf{x} | \theta)\xi(\theta),$$

querendo dizer que os dois lados de \propto são iguais a não ser talvez por uma constante que independe de θ . Por vezes podemos escrever também $\xi(\theta | \mathbf{x}) \propto_{\theta} f(\mathbf{x} | \theta)\xi(\theta)$.

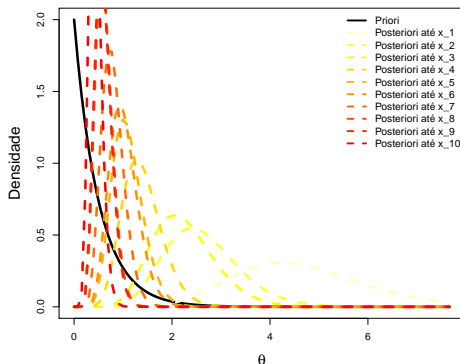
Definição 12 (Função de verossimilhança)

*Quando encaramos a f.d.p./f.m.p. $f(x_1, x_2, \dots, x_n | \theta)$ como uma função do parâmetro θ , chamamos esta função de **função de verossimilhança**, e podemos denotá-la como $L(\theta; \mathbf{x})$ ou, quando a notação não criar ambiguidade, simplesmente $L(\theta)$.*

Aprendizado bayesiano sequencial

Ainda sobre o Exemplo ??, considere a primeira observação x_1 e a distribuição *a posteriori* baseada apenas nesta observação: $\xi_1(\theta | x_1) \propto f(x_1 | \theta)\xi(\theta)$. Se assumirmos que X_1, X_2, \dots, X_n são condicionalmente independentes dado θ , podemos escrever

$$\xi(\theta | x_1, x_2) \propto f(x_1, x_2 | \theta)\xi(\theta) = f(x_1 | \theta)f(x_2 | \theta)\xi(\theta) = f(x_2 | \theta)\xi_1(\theta | x_1).$$



Dentro do paradigma bayesiano, a predição de novos valores da(s) variável(is) aleatória(s) é feita a partir da distribuição *a posteriori*,

$$p(X_{n+1} = x_{n+1} \mid x_1, x_2, \dots, x_n) = \int_{\Omega} f(x_{n+1} \mid \theta) \xi(\theta \mid x_1, x_2, \dots, x_n) d\theta. \quad (5)$$

Chamamos a distribuição condicional em (??) de **distribuição preditiva *a posteriori***. Em contraste, temos a **distribuição preditiva *a priori***:


$$p(x_{n+1}) = \int_{\Omega} f(x_{n+1} \mid \theta) \xi(\theta) d\theta, \quad (6)$$

que é útil na aplicação de modelos bayesianos na prática, mas não será explorada aqui.

O que aprendemos?

- 💡 Bayesianismo X frequentismo;
“Parâmetros como variáveis aleatórias ou constantes fixas e não-observáveis.”
- ⌚ Distribuição *a priori*, $\xi(\theta)$;
“Nosso grau de crença antes de observamos dados.”
- 📄 Função de verossimilhança, $L(\theta) \propto f(\mathbf{x} \mid \theta)$;
“Codifica (toda) a informação sobre o modelo contida nos dados.”
- ⌚ Distribuição *a posteriori*, $\xi(\theta \mid \mathbf{x}) \propto L(\theta)\xi(\theta)$;
“Nossa crença atualizada a partir da informação contida em $L(\theta)$.”

 DeGroot seção 7.2;

 * Capítulo 1 de Schervish, M. J. (2012). Theory of statistics. Springer Science & Business Media.

▶▶ Próxima aula: DeGroot, seção 7.3;

- **Exercícios recomendados**

- DeGroot, seção 7.2: exercícios 2, 3 e 10.

Prioris conjugadas

- Prioris conjugadas
 - ◇ Bernoulli;
 - ◇ Poisson;
 - ◇ Normal;
- Interpretação dos hiperparâmetros.

Teorema 8 (Posteriori Bernoulli)

Sejam X_1, X_2, \dots, X_n uma amostra aleatórias de variáveis aleatórias Bernoulli com parâmetro p , $0 < p < 1$, desconhecido. Suponha que a distribuição a priori de p é uma distribuição Beta com parâmetros $\alpha > 0$ e $\beta > 0$. Seja $y = \sum_{i=1}^n X_i$. Então

$$\xi(p \mid X_1, X_2, \dots, X_n) = \frac{1}{B(\alpha + y, \beta + n - y)} p^{\alpha+y-1} (1 - p)^{\beta+(n-y)-1}.$$

Prova: Escrever a conjunta condicional como produto das marginais condicionais e notar que se obtêm o núcleo de uma distribuição Beta.

Prioris conjugadas

Definição 13 (Hiperparâmetros)

Seja $\xi(\theta \mid \phi)$ a distribuição a priori para o parâmetro θ , indexada por $\phi \in \Phi$. Dizemos que ϕ é (são) o(s) **hiperparâmetro(s)** da priori de θ .

Definição 14 (Priori conjugada)

Suponha que X_1, X_2, \dots sejam condicionalmente independentes dado θ , com f.d.p./f.m.p. $f(x \mid \theta)$. Defina

$$\Psi = \left\{ f : \Omega \rightarrow (0, \infty), \int_{\Omega} f \, dx = 1 \right\},$$

onde Ω é o espaço de parâmetros. Dizemos que Ψ é uma **família de distribuições conjugadas** para $f(x \mid \theta)$ se para toda $f \in \Psi$ e toda realização x de $\mathbf{X} = X_1, X_2, \dots, X_n$,

$$\frac{f(x \mid \theta)f(\theta)}{\int_{\Omega} f(x \mid \theta)f(\theta) \, d\theta} \in \Psi.$$

Isto é, uma família de prioris é conjugada para uma determinada verossimilhança se a posteriori está na mesma família.

Se $X \sim \text{Beta}(a, b)$, $\text{Var}(X) = \frac{ab}{(a+b)^2(a+b+1)}$. Na situação do Teorema ??, temos

$$V_n := \text{Var}(p \mid \mathbf{x}) = \frac{(\alpha + y)(\beta + n - y)}{(\alpha + \beta + n)^2(\alpha + \beta + n + 1)}. \quad (7)$$

Podemos usar a expressão em (??) para desenhar um experimento. Por exemplo, podemos coletar dados até que $V_n \leq 0.01$ (ver exercício 2, seção 7.3 de DeGroot).

Teorema 9 (Posteriori para taxa da Poisson)

Suponha que X_1, X_2, \dots, X_n formam uma amostra aleatória com distribuição Poisson com taxa $\theta > 0$, desconhecida. Suponha que a distribuição a priori para θ é uma distribuição Gama com parâmetros $\alpha > 0$ e $\beta > 0$. Então

$$\xi(\theta \mid \mathbf{x}) = \frac{(\beta + n)^{\alpha+S}}{\Gamma(\alpha + S)} \theta^{\alpha+S-1} e^{-(\beta+n)\theta}, \quad (8)$$

onde $S = \sum_{i=1}^n x_i$.

Prova: Análoga ao exemplo Bernoulli.

Teorema 10 (Distribuição *a posteriori* da média de uma normal)

Suponha que X_1, X_2, \dots, X_n formam uma amostra aleatória com distribuição normal com média desconhecida θ e variância $\sigma^2 > 0$, conhecida e fixa. Suponha que $\theta \sim \text{Normal}(\mu_0, v_0^2)$ a priori. Então

$$\xi(\theta | \mathbf{x}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\theta - \mu_1)^2}{2v_1^2}\right), \quad (9)$$

onde

$$\mu_1 := \frac{\sigma^2\mu_0 + nv_0^2\bar{x}_n}{\sigma^2 + nv_0^2} \quad \text{e} \quad v_1^2 := \frac{\sigma^2v_0^2}{\sigma^2 + nv_0^2} \quad (10)$$

Prova: Escrever as densidades relevantes sem as constantes de proporcionalidade, completar o quadrado (duas vezes) e notar que se obtém o núcleo de uma normal (Gaussiana).

Interpretando a média *a posteriori*

Podemos reescrever μ_1 como


$$\mu_1 = \frac{\sigma^2}{\sigma^2 + nv_0^2} \mu_0 + \frac{nv_0^2}{\sigma^2 + nv_0^2} \bar{x}_n. \quad (11)$$

Observação 1 (Média *a posteriori* como média ponderada)

No caso normal, a média *a posteriori* pode ser vista como uma **média ponderada** entre a média *a priori* e a média amostral, sendo os pesos dados pela variância (conhecida) da distribuição dos dados e a variância da priori, v_0^2 .

O que aprendemos?

- 💡 Prioris conjugadas;
- 💡 Análise conjugada de
 - ◇ Bernoulli;
 - ◇ Poisson;
 - ◇ Normal.

 DeGroot seção 7.3;

▶▶ Próxima aula: DeGroot, seção 7.4;

- **Exercícios recomendados**

- DeGroot, seção 7.3: exercícios 2, 17, 19, 21.

- Estimador e estimativa;
- Função de perda;
- Estimador de Bayes;
- Consistência do estimador de Bayes;
- Estimador de Bayes para grandes amostras;
- Limitações.

Definição 15 (Priori imprópria)

Seja $\xi : \Lambda \rightarrow (0, \infty)$, $\Omega \subseteq \Lambda$, uma função tal que $\int_{\Omega} \xi(\theta) d\theta = \infty$. Se utilizamos ξ como uma p.d.f. para θ , dizemos que ξ é uma **priori imprópria** para θ .

Exemplo 3 (Priori imprópria para a taxa de uma Poisson)

Suponha que X_1, X_2, \dots, X_n formam uma amostra aleatória com distribuição Poisson com taxa $\theta > 0$, desconhecida. Desta vez, fazemos a escolha de hiperparâmetros $\alpha = \beta = 0$, o que leva a

$$\xi(\theta) = \frac{1}{\theta}.$$

A posteriori passa a ser

$$\xi(\theta | \mathbf{x}) = \frac{n^S}{\Gamma(S)} \theta^{n-1} e^{-S\theta},$$

onde $S = \sum_{i=1}^n x_i$.

Definição 16 (Estimador)

Sejam X_1, X_2, \dots, X_n variáveis aleatórias com distribuição conjunta indexada por θ . Um **estimador** de θ é qualquer função real $\delta : X_1, X_2, \dots, X_n \rightarrow \mathbb{R}^d$, $d \geq 1$.

Definição 17 (Estimativa)

Dizemos que o valor de δ avaliado nas realizações de X_1, X_2, \dots, X_n , $\mathbf{x} = \{x_1, x_2, \dots, x_n\}$, $\delta(\mathbf{x})$ é uma **estimativa** de θ .

Definição 18 (Função de perda)

Uma função de perda é uma função real em duas variáveis

$$L : \Omega \times \mathbb{R}^d \rightarrow \mathbb{R},$$

em que dizemos que o estatístico perde $L(\theta, a)$ se o parâmetro vale θ e a estimativa dada vale a .

Exemplos de funções de perda são $L(\theta, a) = (\theta - a)^2$ e $L(\theta, a) = |\theta - a|$.

Observação 2 (Perda esperada *a priori*)

Se escolhermos uma *priori* $\xi(\theta)$, nossa perda esperada, **antes** de observar os dados é

$$E_{\xi}[L(\theta, a)] = \int_{\Omega} L(\theta, a) \xi(\theta) d\theta.$$

Vemos então que a escolha da distribuição *a priori* está inextrincavelmente ligada à função de perda.

Definição 19 (Estimador de Bayes)

Considere a perda esperada a posteriori:

$$E_{\theta|x} [L(\theta, a)] = E[L(\theta, a) | \mathbf{x}] = \int_{\Omega} L(\theta, a) \xi(\theta | \mathbf{x}) d\theta.$$

Dizemos que δ^* é um **estimador de Bayes** se, para toda realização $\mathbf{X} = \mathbf{x}$,

$$E[L(\theta, \delta^*(\mathbf{x})) | \mathbf{x}] = \min_{a \in \mathcal{A}} E[L(\theta, a) | \mathbf{x}].$$

- Em outras palavras, um estimador de Bayes é uma função real dos dados que minimiza a perda esperada com respeito à posteriori dos parâmetros.

Estimador de Bayes sob perda quadrática

Suponha que a função de perda seja

$$L(\theta, \delta^*) = (\theta - \delta^*)^2.$$

Dizemos que a função de perda é **quadrática**. Temos o seguinte resultado:

Teorema 11 (δ^* sob perda quadrática)

(DeGroot, Corolário 7.4.1)

Seja θ um parâmetro tomando valores reais. Sob perda quadrática,

$$\delta^*(x) = E[\theta | \mathbf{X} = x] = \int_{\Omega} \theta \xi(\theta | x) d\theta.$$

Prova: Escrever a perda esperada *a posteriori* explicitamente, usar a lei de esperanças e minimizar a expressão resultante com respeito ao estimador (ex. diferenciar e igualar a derivada a zero).

Estimador de Bayes sob perda absoluta

Teorema 12 (δ^* sob perda absoluta)

(DeGroot, Corolário 7.4.2)

Suponha que a função de perda é dada por

$$L(\theta, \delta^*) = |\theta - \delta^*|.$$

Dizemos que a função de perda é **absoluta**.

Seja θ um parâmetro tomando valores na reta. Sob perda absoluta, $\delta^*(\mathbf{x})$ é a **mediana** a posteriori, isto é,

$$\int_{-\infty}^{\delta^*(\mathbf{x})} \xi(\theta | \mathbf{x}) d\theta = \frac{1}{2}.$$

Prova: Decompor a perda esperada em duas integrais de funções não-negativas utilizando as propriedades da função valor absoluto e aplicar a regra de Leibnitz duas vezes para encontrar o ponto de mínimo.

O estimador de Bayes em grandes amostras

- Sob condições brandas de regularidade, à medida que o tamanho de amostra cresce, a influência da priori diminui.

Exemplo 4 (Proporção de itens defeituosos)

Suponha que estamos interessados na proporção θ de itens defeituosos em uma linha de produção. Suponha ainda que

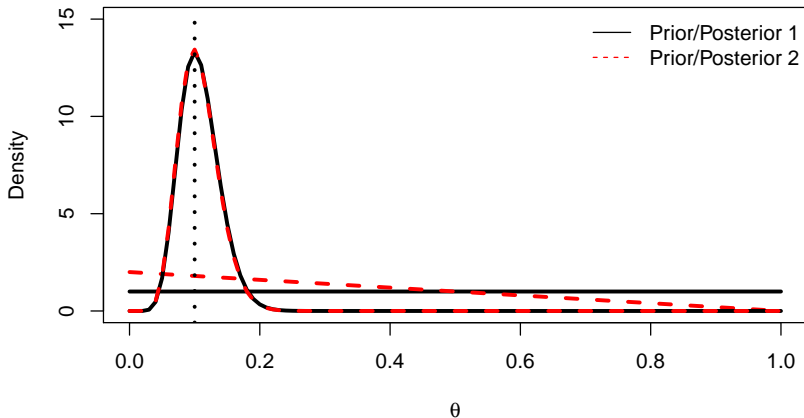
- Priori 1: $\xi_1(\theta) = 1, 0 < \theta < 1$;*
- Priori 2: $\xi_2(\theta) = 2(1 - \theta), 0 < \theta < 1$;*
- Dados: de $n = 100$ itens observados, $y = 10$ apresentaram defeito.*

Perguntas:

- $\bar{x}_n = ?$
- $E_1[\theta | \mathbf{x}] = \int_0^1 \theta \xi_1(\theta | \mathbf{x}) d\theta = ?$

Proporção de itens defeituosos: prioris e posteriores

Ver também exemplo 7.3.3 de DeGroot.



Consistência do estimador de Bayes

Definição 20 (Estimador consistente)

Seja $\delta_1, \delta_2, \dots, \delta_n$ uma sequência de estimadores de θ . Se quando $n \rightarrow \infty$ a sequência converge para θ , dizemos que esta é uma sequência consistente de estimadores.

Observação 3 (A média amostral é consistente para o caso Bernoulli)

Se X_1, X_2, \dots, X_n são i.i.d. Bernoulli com parâmetro θ condicional a θ , temos pela LGN: $\bar{X}_n \xrightarrow{P} \theta$.

Observação 4 (O estimador de Bayes é consistente para o caso Bernoulli)

Para $\alpha > 0$ e $\beta > 0$ fixos, a média a posteriori vale

$$\delta^*(\mathbf{x}) = E[\theta \mid \mathbf{x}] = \frac{\alpha + y}{\alpha + \beta + n},$$

onde $y = \sum_{i=1}^n x_i$. É fácil ver que $\delta^*(\mathbf{x}) \xrightarrow{P} \bar{x}_n \xrightarrow{P} \theta$.

O que aprendemos?

- 💡 Estimador;
“Um estimador é qualquer função real dos dados”
- 💡 Função de perda;
“Uma função real que quantifica a perda incorrida por uma estimativa incorreta”
- 💡 Estimador de Bayes;
“Um estimador que minimiza a perda esperada *a posteriori*”
- 💡 Propriedades e limitações do estimador de Bayes;
“À medida que o tamanho da amostra cresce, o estimador se aproxima do valor verdadeiro, a influência da priori diminui, mas precisamos sempre de uma função de perda bem especificada”

Leitura recomendada

 DeGroot seção 7.4;

 * Casella & Berger, seção 7.2.3.

▶▶ Próxima aula: DeGroot, seções 7.5 e 7.6;

- **Exercícios recomendados**

- DeGroot, seção 7.4: exercícios 2, 4, 7, 11 e 14.

Tópicos da aula

- Estimador de máxima verossimilhança (EMV);
 - ◇ Existência e unicidade;
 - ◇ Invariância do EMV;
 - ◇ Consistência do EMV;
- Limitações;

Definição 21 (Estimador de máxima verossimilhança)

Para cada possível vetor (de observações) \mathbf{x} , seja $\delta(\mathbf{x}) \in \Omega$ um valor de $\theta \in \Omega$ de modo que a função de verossimilhança, $L(\theta) \propto f(\mathbf{x} \mid \theta)$, atinge o máximo. Dizemos que $\hat{\theta} = \delta(\mathbf{X})$ é o **estimador de máxima verossimilhança** de θ (Fisher, 1922)⁸. Quando observamos $\mathbf{X} = \mathbf{x}$, dizemos que $\delta(\mathbf{x})$ é uma estimativa de θ . Dito de outra forma,

$$\max_{\theta \in \Omega} f(\mathbf{X} \mid \theta) = f(\mathbf{X} \mid \hat{\theta}).$$

⁸Ronald Aylmer Fisher (1890-1962), biólogo e estatístico inglês. Para a história do desenvolvimento do EMV, ver [Aldrich \(1997\)](#).

Mudando de paradigma

Na Definição ??, vemos θ com um número real que indexa a distribuição de probabilidade conjunta dos dados.

- Poderíamos trocar⁹ $f(x | \theta)$ por $f(x; \theta)$;
- Com o EMV, procuramos um valor de θ de modo que a probabilidade de observarmos $\mathbf{X} = \mathbf{x}$ seja máxima;
- Isso não nos diz nada sobre o quão provável $\hat{\theta}$ é;
- θ não é uma quantidade aleatória, portanto não admite afirmações probabilísticas.

⁹Mas não vamos, pois a notação fica clara em quase todos os contextos.

Exemplos

- Exponencial;
- Bernoulli;
- Normal;
 - ◊ μ desconhecida, σ^2 conhecida;
 - ◊ μ conhecida, σ^2 desconhecida;
 - ◊ μ e σ^2 ambas desconhecidas.

- Exponencial: $\hat{\theta} = 1/\bar{X}_n$;
- Bernoulli $\hat{\theta} = \bar{X}_n$;
- Normal;
 - ◊ $\hat{\mu} = \bar{X}_n$;
 - ◊ $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2$;
 - ◊ $\hat{\theta} = \left\{ \hat{\mu} = \bar{X}_n, \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \right\}$.

Exemplo 5 (EMV para uniforme)

Suponha que X_1, X_2, \dots, X_n perfazem uma amostra aleatória de uma distribuição uniforme no intervalo $[0, \theta]$, $\theta \in \mathbb{R}, \theta > 0$. Considere a f.d.p.

$$f(x | \theta) = \begin{cases} \frac{1}{\theta}, & 0 \leq x \leq \theta, \\ 0, & \text{caso contrário.} \end{cases} \quad (12)$$

A f.d.p. conjunta é

$$f_n(\mathbf{x} | \theta) = \begin{cases} \theta^{-n}, & 0 \leq x_i \leq \theta \ (i = 1, 2, \dots, n), \\ 0, & \text{caso contrário,} \end{cases} \quad (13)$$

e o EMV é $\hat{\theta} = \max(x_1, x_2, \dots, x_n)$.

Observação 5 (Existência do EMV)

A existência do EMV pode depender de detalhes irrelevantes acerca do espaço de parâmetros, Ω .

Exemplo 6 (Não existência do EMV)

Considere o Exemplo ??, mas agora com uma f.d.p. um pouco diferente:

$$f(x | \theta) = \begin{cases} \frac{1}{\theta}, & 0 < x < \theta, \\ 0, & \text{caso contrário.} \end{cases} \quad (14)$$

É fácil mostrar que, nesse caso, o EMV não existe.

Observação 6 (Unicidade do EMV)

Mesmo quando existe, o EMV nem sempre é único.

Exemplo 7 (EMV para uma uniforme num intervalo de tamanho 1)

Suponha que X_1, X_2, \dots, X_n perfazem amostra aleatória de uma distribuição uniforme no intervalo $[\theta, \theta + 1]$. A densidade conjunta é

$$f_n(\mathbf{x} \mid \theta) = \begin{cases} 1, \theta \leq x_i \leq \theta + 1, (i = 1, 2, \dots, n), \\ 0, \text{ caso contrário.} \end{cases} \quad (15)$$

Defina $m := \min(x_1, x_2, \dots, x_n)$ e $M := \max(x_1, x_2, \dots, x_n)$. Podemos reescrever (??) como

$$f_n(\mathbf{x} \mid \theta) = \begin{cases} 1, M - 1 \leq \theta \leq m, (i = 1, 2, \dots, n), \\ 0, \text{ caso contrário.} \end{cases} \quad (16)$$

Conclusão: $\hat{\theta}$ é qualquer valor no intervalo $[M - 1, m]$.

Invariância do EMV

Suponha que estamos interessados em uma transformação do parâmetro θ , $\phi(\theta)$. Por exemplo, se X_1, X_2, \dots, X_n são Bernoulli com parâmetro p , podemos estar interessados na *chance* $\omega = \phi(p) = p/(1 - p)$.

Teorema 13 (Invariância do EMV)

Considere uma função $\phi : \Omega \rightarrow \mathbb{R}$. Se $\hat{\theta}$ é um EMV para θ , então $\phi(\hat{\theta})$ é um EMV para $\omega = \phi(\theta)$.

Prova: Defina a *verossimilhança induzida*:

$$L^*(\omega) := \sup_{\{\theta: \phi(\theta)=\omega\}} L(\theta),$$

e note que o supremo desta função sobre Ω é precisamente o EMV. Ver Casella & Berger, Teorema 7.2.10 (pág. 320) ou DeGroot, Teorema 7.6.2 (pág. 427).

Exemplo: O EMV para o quadrado da média de uma normal, μ^2 , é \bar{X}_n^2 .

Consistência do EMV

Sob condições de regularidade, o EMV é consistente, isto é $\hat{\theta}_{EMV} \rightarrow \theta$.

Teorema 14 (Consistência do EMV)

Defina $l(\theta) := \log f_n(\mathbf{x} \mid \theta)$ e assuma que $X_1, X_2, \dots, X_n \sim f(\theta_0)$, isto é, que θ_0 é o valor verdadeiro do parâmetro. Denote $E_{\theta_0}[g] := \int_{\mathcal{X}} g(\mathbf{x}, \theta_0) f(\mathbf{x} \mid \theta_0) d\mathbf{x}$. Suponha que

- $f(\mathbf{x}_i \mid \theta)$ tem o mesmo suporte;
- θ_0 é ponto interior de Ω ;
- $l(\theta)$ é diferenciável;
- $\hat{\theta}_{EMV}$ é a única solução de $l'(\theta) = 0$.

Então,

$$\hat{\theta}_{EMV} \rightarrow \theta.$$

Prova: (rascunho) mostrar que, para todo $\theta \in \Omega$,

$$\frac{1}{n} \sum_{i=1}^n \log f(X_i \mid \theta) \rightarrow E_{\theta_0} [\log f(\mathbf{X} \mid \theta)],$$

e aplicar a desigualdade de Jensen.


O que aprendemos?

- 💡 Estimador de máxima verossimilhança (EMV);
“Encontrar o valor do parâmetro que maximiza a probabilidade observar os dados obtidos”
- 💡 Invariância ;
“O EMV é invariante a transformações dos parâmetros; se $\hat{\theta}$ é o EMV para θ , $\phi(\hat{\theta})$ é o EMV para $\phi(\theta)$ ”
- 💡 Consistência;
“Sob condições brandas de regularidade, o EMV converge para valor verdadeiro à medida que $n \rightarrow \infty$ ”
- 💡 Limitações;
“O EMV não existe necessariamente, e, mesmo quando existe, não precisa ser único”

Leitura recomendada

 DeGroot seções 7.5 e 7.6;

 * Casella & Berger, seção 7.2.2.

 * Schervish (1995), seção 5.1.3.

▶▶ Próxima aula: DeGroot, seção 7.6 (pág. 432 em diante);

• Exercícios recomendados

■ DeGroot,

Seção 7.5: exercícios 1, 4, 9 e 10;

Seção 7.6: exercícios 3, 5 e 11.

- Estimadores de Bayes vs EMV;
- Método dos momentos.

Argumentos assintóticos:

$$L(\theta) \approx \exp \left[-\frac{(\theta - \hat{\theta})^2}{2V_n(\theta)/n} \right].$$

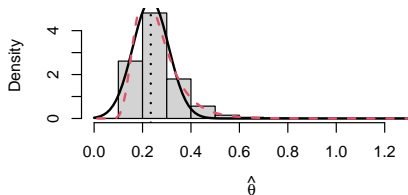
Exemplo 8 (Exemplo 7.6.11 em DeGroot)

$X_1, X_2, \dots, X_n \sim \text{Exponencial}(\theta),$

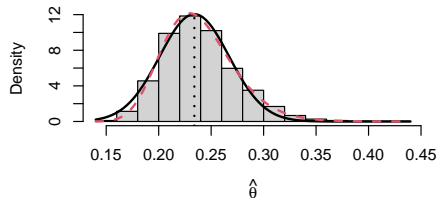
- $\hat{\theta}_{EMV} = (\bar{X}_n)^{-1} \implies E[\hat{\theta}_{EMV}] = \theta$ e $\text{Var}[\hat{\theta}_{EMV}] = \theta^2$. Pelo método Delta, temos $\hat{\theta}_{EMV} \approx \text{Normal}(\theta, \theta^2/n)$;
- Se escolhemos uma priori gama para θ com hiperparâmetros $\alpha > 0$ e $\beta > 0$, temos $E_{\theta|x}[\theta] = (\alpha + n)/(\beta + S_n)$ e $\text{Var}_{\theta|x}(\theta) = (\alpha + n)/(\beta + S_n)^2$. Fazendo $\alpha, \beta \ll n$, temos um argumento análogo para o estimador de Bayes.

EMV para taxa de uma exponencial (DG ex. 7.6.11)

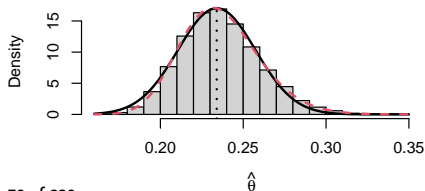
Tamanho de amostra = 10



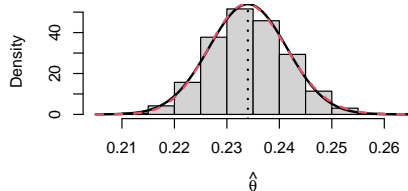
Tamanho de amostra = 50



Tamanho de amostra = 100



Tamanho de amostra = 1000



Inferência para uma Uniforme em $(0, \theta)$

Exemplo 9 (Exemplo 7.6.14 em DeGroot)

$X_1, X_2, \dots, X_n \sim \text{Uniforme}(0, \theta)$. Definindo $Y = \max(X_1, X_2, \dots, X_n)$ temos

$$g_n(y | \theta) = n \frac{y^{n-1}}{\theta^n}.$$

Como já discutido, temos $\hat{\theta}_{EMV} = \max(x_1, x_2, \dots, x_n) = y_n$ e portanto

- $E[\hat{\theta}_{EMV}] = \frac{n}{n+1}\theta$;
- $\text{Var}(\hat{\theta}_{EMV}) = \frac{n}{(n+1)^2(n+2)}\theta^2$.

Do lado bayesiano, vamos obter a posteriori (com uma priori imprópria):

$$\xi(\theta | \mathbf{x}) = \begin{cases} \frac{(n-1)y_n^{n-1}}{\theta^n}, & y_n < \theta, \\ 0, & \text{caso contrário.} \end{cases} \quad (17)$$

Isto nos leva a

- $E[\hat{\theta}_{Bayes}] = \frac{n-1}{n-2}y_n$;
- $\text{Var}(\hat{\theta}_{Bayes}) = \frac{n-1}{(n-2)^2(n-3)}y_n^2$.

Método dos momentos (MM)

Algumas vezes, obter o EMV ou o estimador de Bayes envolve dificuldades numéricas (ex. estimar os parâmetros de uma distribuição Gama). Nestas situações, podemos encontrar um estimador para os parâmetros que relacione os momentos empíricos com os teóricos.

Definição 22 (Método dos momentos)

Suponha que X_1, X_2, \dots, X_n formam uma amostra aleatória com distribuição conjunta $f_n(X_1, X_2, \dots, X_n | \theta)$, $\theta \in \Omega \subseteq \mathbb{R}^k$ e que o k -ésimo momento existe. Defina $\mu_j(\theta) = E[X_1^j | \theta]$ e suponha que $\mu : \Omega \rightarrow \mathbb{R}^k$ é biúnivoca, de modo que sua inversa é

$$\theta = M(\mu_1(\theta), \dots, \mu_k(\theta)).$$

*Dados os momentos amostrais $m_j := \frac{1}{n} \sum_{i=1}^n X_i^j$, $j = 1, \dots, k$, o **estimador de momentos (EMM)** de θ é*

$$\hat{\theta}_{EMM} = M(m_1, \dots, m_k).$$

Exemplo

Exemplo 10

$X_1, X_2, \dots, X_n \sim \text{Gama}(\alpha, \beta)$, com $\alpha > 0$ e $\beta > 0$ desconhecidos. Para começar,

- $\mu_1(\theta) = \alpha/\beta$;
- $\mu_2(\theta) = (\alpha + 1)\alpha/\beta^2$.

Agora equacionamos com os momentos amostrais ("empíricos"): $\mu_1(\theta) = \bar{x}_n$ e $\mu_2(\theta) = \frac{1}{n} \sum_{i=1}^n x_i^2$ para obter

- $\hat{\alpha} = \frac{(\bar{x}_n)^2}{\frac{1}{n} \sum_{i=1}^n x_i^2 - (\bar{x}_n)^2} = \frac{(\bar{x}_n)^2}{\bar{s}^2}$;
- $\hat{\beta} = \frac{\bar{x}_n}{\frac{1}{n} \sum_{i=1}^n x_i^2 - (\bar{x}_n)^2} = \frac{\bar{x}_n}{\bar{s}^2}$.

Observação 7

O método dos momentos também pode ser usado para obter chutes iniciais para procedimentos numéricos nos métodos mais avançados (EMV, Bayes).

Teorema 15 (Consistência do EMM)

Suponha que X_1, X_2, \dots, X_n formam uma amostra aleatória com distribuição conjunta $f_n(X_1, X_2, \dots, X_n | \theta)$, $\theta \in \Omega \subseteq \mathbb{R}^k$ e que o k -ésimo momento existe. Mais uma vez, suponha que a inversa M existe e é contínua. Então o EMM é consistente para θ .


Prova: Pela LGN, $m_i \xrightarrow{P} \mu_i(\theta)$. Assumindo que M é contínua, temos que $M(m_1, \dots, m_k) \xrightarrow{P} M(\mu_1(\theta), \dots, \mu_k(\theta)) = \theta$ (DeGroot, Teorema 6.2.5).

O que aprendemos?

- 💡 EMV vs Bayes;
“Em várias situações, à medida que $n \rightarrow \infty$, os estimadores 'convergem' ”
- 💡 Nem sempre EMV \approx Bayes;
“Verossimilhanças descontínuas e/ou pequenos tamanhos de amostra”
- 💡 Método dos momentos (MM);
“Quando os momentos são funções inversíveis dos parâmetros, podemos obter estimadores em função dos momentos amostrais”
- 💡 Consistência do MM;
“Sob condições brandas de regularidade, o EMM converge para valor verdadeiro à medida que $n \rightarrow \infty$ ”
- 💡 Limitações do MM;
“Raras as situações em que tudo se alinha de modo que o EMM exista em forma fechada”

Leitura recomendada

 DeGroot seção 7.6;

 * Schervish (1995), capítulo 7.

▶▶ Próxima aula: DeGroot, seções 7.7 e 7.8;

- **Exercícios recomendados**

- DeGroot, seção 7.6: exercícios 20, 22 e 23.

- Estatística suficiente;
- Teorema da fatorização;
- Suficiência conjunta;
- Suficiência mínima;

Um exemplo motivador

Suponha que os tempos de falha de um modelo de lâmpada podem ser modelados como $X_1, X_2, \dots, X_n \sim \text{expo}(\theta)$.

Suponha que dois técnicos, Afonso e Bruna, medem cada um três lâmpadas, obtendo:

- $x_A = \{1.64, 1.37, 0.13\}$ meses;
- $x_B = \{0.48, 0.87, 1.79\}$ meses;

O chefe dos dois, Astolfo, suspeita que o tempo de falha seja, em média, 2 meses com desvio padrão de mais ou menos 1 mês. Para cada uma das amostras

- (i) Compute o estimador de Bayes θ sob perda quadrática;
- (ii) Estime θ por máxima verossimilhança.

Definição 23 (Estatística suficiente)

Seja X_1, X_2, \dots, X_n uma amostra aleatória de uma distribuição indexada pelo parâmetro θ . Seja $T = r(X_1, X_2, \dots, X_n)$ uma estatística. Dizemos que T é uma **estatística suficiente** para θ se e somente se

$$f(X_1, X_2, \dots, X_n \mid T = t, \theta) = f(X_1, X_2, \dots, X_n \mid T = t, \theta'), \quad \forall \theta, \theta' \in \Omega,$$

isto é, se a distribuição condicional da amostra dado o valor da estatística não depende de θ .

No exemplo anterior, tanto $\hat{\theta}_{\text{Bayes}}$ quanto $\hat{\theta}_{\text{EMV}}$ dependem de X_1, X_2, \dots, X_n apenas através de $r(X_1, X_2, \dots, X_n) = T = \sum_{i=1}^n X_i$.

Uma observação importante

Definição 24 (Aleatorização auxiliar)

Suponha que T é suficiente para θ . O processo de simular $X'_1, \dots, X'_n \mid T = r(X_1, X_2, \dots, X_n)$ de modo que

$$f(X_1, X_2, \dots, X_n \mid \theta) = f(X'_1, \dots, X'_n \mid \theta), \forall \theta \in \Omega,$$

é chamado de **aleatorização auxiliar** (em inglês, *auxiliary randomisation*).

Observação 8 (A busca por bons estimadores)

Na busca por bons estimadores, estamos justificados em restringir a busca a funções de estatísticas suficientes.

Justificativa: Suponha que o estatístico A tem à sua disposição X_1, X_2, \dots, X_n , enquanto B tem acesso somente a $T = r(X_1, X_2, \dots, X_n)$. Se T é suficiente, B pode sempre fazer uma aleatorização auxiliar e gerar X'_1, \dots, X'_n com exatamente a mesma distribuição conjunta condicional a θ .

Teorema da fatorização (TF)

Teorema 16 (Teorema da fatorização)

Suponha que X_1, X_2, \dots, X_n perfazem uma amostra aleatória com f.d.p./f.m.p $f(x | \theta)$, $\theta \in \Omega$. Uma estatística $T = r(X_1, X_2, \dots, X_n)$ é suficiente para θ se, e somente se, para todo $x \in \mathcal{X}$ e $\theta \in \Omega$ existem u e v não negativas tal que

$$f_n(x | \theta) = u(x)v[r(x), \theta].$$

Prova: (Para v.a.s discretas). Para a “ida” notar que T é uma função determinística de X , ou seja, $\Pr(T = t | X = x, \theta) = 1$ e que só precisamos considerar $x \in \{y : r(y) = t\}$. Para a “volta”, mostrar que T suficiente implica que $\Pr(X = x | T = t, \theta)$ é função apenas de x . Ver DeGroot, Teorema 7.7.1 e Casella & Berger, Teorema 6.2.6.

- Poisson;
- $f(x | \theta) = \theta x^{\theta-1}$, $x \in (0, 1)$ e $\theta > 0$;
- Normal;

O que acontece, por exemplo, no caso Normal com μ e σ^2 desconhecidos?

Definição 25 (Suficiência conjunta)

*Dizemos que um conjunto de estatísticas $\mathbf{T} = \{T_1, \dots, T_k\}$ é **suficiente** (conjuntamente) se que a distribuição condicional conjunta de X_1, X_2, \dots, X_n dado $T_1 = t_1, \dots, T_k = t_k$ não depende de θ .*

Observação 9 (TF para estatísticas suficientes conjuntas)

Para o caso de estatísticas suficientes conjuntas, vale um Teorema da fatorização:

$$f_n(\mathbf{x} \mid \theta) = u(\mathbf{x})v[r_1(\mathbf{x}), \dots, r_k(\mathbf{x}), \theta].$$

Suficiência conjunta – exemplos

- Normal;
- Uniforme;

Observação 10 (Transformações biunívocas de estatísticas suficientes)

Se $\mathbf{T} = \{T_1, \dots, T_k\}$ são estatísticas suficientes conjuntas, e $h : \mathcal{T} \rightarrow \mathbb{R}$ é um mapa inversível, então $\mathbf{T}' = h(\mathbf{T})$ também são suficientes conjuntas.

Primeiro um exemplo motivador:

Definição 26 (Estatísticas de ordem)

Seja $\mathbf{X} = X_1, X_2, \dots, X_n$ uma amostra aleatória. Dizemos que Y_1, Y_2, \dots, Y_n são **estatísticas de ordem** se Y_1 é o menor valor de \mathbf{X} , Y_5 é o quinto menor valor e assim por diante.

Teorema 17 (Estatísticas de ordem são suficientes conjuntas)

Seja X_1, X_2, \dots, X_n uma amostra aleatória com f.d.p/f.m.p. $f(x | \theta)$. As estatísticas de ordem Y_1, Y_2, \dots, Y_n são suficientes conjuntas para θ .

Prova: Usar o fato de que a conjunta é o produto das marginais e a comutatividade da multiplicação em \mathbb{R} .

Definição 27 (Suficiência mínima)

Uma estatística T é dita **mínima suficiente** se T é suficiente e é função de qualquer outra estatística suficiente. Um vetor $\mathbf{T} = \{T_1, \dots, T_k\}$ é dito **minimamente suficiente conjunto** se é função de qualquer outro vetor de estatísticas suficientes conjuntas.

Observação 11 (Estatísticas de ordem são minimamente suficiente conjuntas no caso Cauchy)

$$f_n(\mathbf{x} \mid \theta) = \frac{1}{\pi^n \prod_{i=1}^n [1 + (x_i - \theta)^2]} \quad (18)$$

Teorema 18 (EMV e Bayes são suficientes)

Se a função de verossimilhança admite fatorização como no Teorema ??, os estimadores de Bayes e de máxima verossimilhança são estatísticas minimamente suficientes.

Prova:



- EMV: notar que $f_x(\mathbf{x} \mid \theta) \propto v[r(\mathbf{x}), \theta]$;
- Bayes: escrever a perda esperada a posteriori explicitamente usando a verossimilhança na forma do TF.

Ver Teoremas 7.8.3 e 7.8.4 de DeGroot.

O que aprendemos?

- 💡 Estatística suficiente;
 “Uma estatística T é suficiente para θ se $\Pr(\mathbf{X} = \mathbf{x} \mid T = t, \theta)$ não depende de θ .”
- 💡 Teorema da fatorização;
 “Se T é suficiente para θ , podemos escrever a verossimilhança como o produto entre uma função que não depende de θ e uma função que só depende de \mathbf{X} através de T .”
- 💡 Os estimadores de Bayes e de máxima verossimilhança são minimamente suficientes.

Leitura recomendada

-  DeGroot seções 7.7 e 7.8;
-  * Casella & Berger (2002), seção 6.2.
- ▶▶ Próxima aula: DeGroot, seção 7.9;

- **Exercícios recomendados**

- 📖 DeGroot.

- Seção 7.7: exercícios 4, 7, 13, 16;

- Seção 7.8: exercícios 3, 8, 12, 16.

Como avaliar um estimador?

Definição 28 (Notação conveniente)

Para as próximas computações, é conveniente definir Para $g : \mathcal{X}^n \rightarrow \mathbb{R}$, escrevemos

$$E_{\theta}[g] = \int_{\mathcal{X}} \cdots \int_{\mathcal{X}} g(\mathbf{x}) f_n(\mathbf{x} | \theta) dx_1 \cdots dx_n = \int_{\mathcal{X}^n} g(\mathbf{x}) f_n(\mathbf{x} | \theta) d\mathbf{x}.$$

Agora podemos definir o **erro quadrático médio** (EQM) de um estimador $\delta(\mathbf{X})$:

Definição 29 (Erro quadrático médio)

$$R(\theta, \delta) := E_{\theta} [\{\delta(\mathbf{X}) - \theta\}^2].$$

Seja \mathbf{T} uma estatística suficiente. Podemos definir o seguinte estimador

Definição 30 (Estimador condicionado)

$$\delta_0(\mathbf{T}) := E_{\theta} [\delta(\mathbf{X}) \mid \mathbf{T}].$$

Como \mathbf{T} é suficiente, podemos escrever, simplesmente,

$$\delta_0(\mathbf{T}) = E [\delta(\mathbf{X}) \mid \mathbf{T}].$$

Com essas definições em mãos, estamos preparados para enunciar um dos teoremas mais importantes da Estatística:

Teorema 19 (Teorema de Rao-Blackwell¹⁰)

Seja $\delta(\mathbf{X})$ um estimador, \mathbf{T} uma estatística suficiente para θ e seja $\delta_0(\mathbf{T})$ como na definição ???. Então vale que

$$R(\theta, \delta_0) \leq R(\theta, \delta).$$

¹⁰O estatístico indo-estadunidense Calyampudi Radhakrishna Rao (1920-) e o estatístico estadunidense David Harold Blackwell (1919-2010) provaram o resultado independentemente no final dos anos 1940.

Prova do TRB

Primeiro, notemos que, para qualquer função g e variáveis aleatórias X e Y , valem os seguintes fatos:

- $(E[g(X) | Y])^2 \leq E[\{g(X)\}^2 | Y]$;
Desigualdade de Cauchy-Schwarz¹¹, também obtida, nesse caso, rearranjando a expressão da variância.
- $E\{E[X | Y]\} = E[X]$ (lei da esperança total).

Fazendo $g(X) = (\delta(\mathbf{X}) - \theta)^2$, obtemos

$$(E[\delta(\mathbf{X}) | \mathbf{T}] - \theta)^2 \leq E[(\delta(\mathbf{X}) - \theta)^2 | \mathbf{T}] \quad (19)$$

Note que $(E[\delta(\mathbf{X}) | \mathbf{T}] - \theta)^2 = [\delta_0(\mathbf{T}) - \theta]^2$. Agora, tomamos esperanças nos dois lados de (19) para obter:

$$\begin{aligned} R(\theta, \delta_0) &= E[(\delta_0(\mathbf{T}) - \theta)^2] \leq E\{E[(\delta(\mathbf{X}) - \theta)^2 | \mathbf{T}]\} \\ &= E[(\delta(\mathbf{X}) - \theta)^2] = R(\theta, \delta). \quad \square \end{aligned}$$

¹¹Em homenagem ao matemático francês Augustin-Louis Cauchy (1789-1857) e ao matemático alemão Karl Hermann Amandus Schwarz (1843-1921).

O conceito de admissibilidade diz respeito à relação entre estimadores.

Definição 31 (Admissibilidade)

Um estimador δ é dito **inadmissível** se existe outro estimador δ_0 tal que $R(\theta, \delta_0) \leq R(\theta, \delta)$ para todo $\theta \in \Omega$ e existe $\theta' \in \Omega$ tal que $R(\theta', \delta_0) < R(\theta', \delta)$. Nesse caso, dizemos que δ_0 domina δ . O estimador δ_0 é **admissível** se (e somente se) não há nenhum estimador que o domine.

Observação 12 (Estimadores admissíveis e o Teorema de Rao-Blackwell)

O Teorema de Rao-Blackwell diz que todo estimador condicionado em uma estatística suficiente é admissível.


Exemplo 11 (Estimadores no caso normal)

- Estimando μ através da mediana amostral;
- Estimando $\sqrt{\sigma^2}$.


O que aprendemos?

- 💡 Teorema de Rao-Blackwell;
 “Quando T é uma estatística suficiente, todo estimador condicionado em T tem menor EQM”
- 💡 Estimador admissível;
 “Um estimador é admissível quando domina todos os outros estimadores ”
- 💡 Caso normal;
 “No caso normal, qualquer estimador de μ que não seja função de \bar{X}_n é inadmissível. O mesmo vale para qualquer estimador de $\sqrt{\sigma^2}$ que não seja função de $\sum_{i=1}^n X_i$ e $\sum_{i=1}^n X_i^2$.”

Leitura recomendada

 DeGroot, seção 7.9;

 * Casella & Berger (2002), seção 7.3.

 * Schervish (1995), Teorema 3.20.

▶▶ Próxima aula: DeGroot, seções 8.7 e 8.8;

- **Exercícios recomendados**

- DeGroot, Seção 7.9: exercícios 2, 3, 6 e 10.

Em Estatística, a palavra viés tem um significado preciso e tem a ver com a esperança da distribuição de um estimador.

Definição 32 (Estimador não-viesado)

Um estimador $\delta(\mathbf{X})$ de uma função $g(\theta)$ é dito **não-viesado** se $E_{\theta}[\delta(\mathbf{X})] = g(\theta)$ para todo $\theta \in \Omega$. Um estimador que não atende a essa condição é dito viesado. O **viés** de δ é definido como $B_{\delta}(\theta) := E_{\theta}[\delta(\mathbf{X})] - g(\theta)$.

Exemplo 12 (Tempos de falha de lâmpadas)

Lembremos do exemplo das lâmpadas da fábrica de Astolfo. Neste caso, não é difícil mostrar que $E[\hat{\theta}_{EMV}] = \frac{n}{n-1}\theta = 3\theta/2$. Desta forma, o viés do EMV é $B_{\hat{\theta}_{EMV}}(\theta) = 3\theta/2 - \theta = \theta/2$. É possível encontrar $\delta(\mathbf{X})$ não-viesado? Esse estimador é bom?

Estimadores não-viesados sempre?

Quando avaliamos estimadores, o erro quadrático médio e o viés são alguns *aspectos* a serem considerados, mas há um compromisso (*trade-off*) entre eles, de certa forma.

Observação 13 (Erro quadrático, variância e viés)

$$R(\theta, \delta) = \text{Var}_{\theta}(\delta) + [B_{\delta}(\theta)]^2.$$

No exemplo das lâmpadas, é possível mostrar que $\delta_2(\mathbf{X}) = 1/S$ tem o menor EQM, mas tem viés $B_{\delta_2}(\theta) = \frac{n-2}{n-1}\theta = \theta/2$, assim como o EMV.

Estimador não-viesado da variância

A variância amostral como a temos definido até aqui é viesada. Uma pequena modificação leva a um estimador não viesado da variância.

Teorema 20 (Estimador não-viesado da variância)

Seja $\mathbf{X} = \{X_1, X_2, \dots, X_n\}$ uma amostra aleatória, com $E[X_1] = m$ e $\text{Var}(X_1) = v < \infty$. Então

$$\delta_1(\mathbf{X}) = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

é um estimador não-viesado de v .

Prova: usar a igualdade

$$\sum_{i=1}^n (X_i - m)^2 = \sum_{i=1}^n (X_i - \bar{X}_n)^2 + n(\bar{X}_n - m)^2$$

e usar a linearidade da esperança e o fato de que temos uma amostra aleatória.

Não-viesamento é uma característica desejável, mas nem sempre um estimador não-viesado (i) existe ou (ii) é um bom estimador.

- Não existência. Exemplo: $X_1, X_2, \dots, X_n \sim \text{Bernoulli}(p)$, estimador para \sqrt{p} ?
- Estimador não-viesado ruim: $X \sim \text{Geometrica}(p)$. Quais as propriedades do estimador não viesado, $\delta(X)$?

Exemplo 13 (Estudando chegada de clientes)

Exemplo 8.8.1 em DeGroot. Suponha que Palmirinha esteja interessada em estudar quantos clientes chegam à sua loja de pamonha num determinado intervalo. Para isso, ela vai modelar o fenômeno como um processo de Poisson:

$$Y(\Delta_t) \sim \text{Poisson}(\theta\Delta_t),$$

isto é, o número Y de clientes num intervalo de tempo Δ_t tem distribuição Poisson com média $\theta\Delta_t$. Palmirinha pode

- *Fixar um número n de clientes a serem observados e marcar o tempo, X que leva para chegarem n clientes ou;*
- *Fixar um determinado intervalo de tempo, t , e contar o número Y de clientes que chegam neste intervalo.*

Pergunta: *qual desenho é melhor para estimar θ ?*

Como medir a quantidade de informação (sobre um parâmetro θ) contida em uma amostra aleatória? A (matriz de) informação de Fisher oferece a resposta.

Definição 33 (Informação de Fisher)

Seja X uma variável aleatória com f.d.p/f.m.p. $f(x | \theta)$, $\theta \in \Omega \subseteq \mathbb{R}$. Suponha que $f(x | \theta)$ é duas vezes diferenciável com respeito a θ . Defina $\lambda(x | \theta) = \log f(x | \theta)$ e

$$\lambda'(x | \theta) = \frac{\partial \lambda(x | \theta)}{\partial \theta} \quad \text{e} \quad \lambda''(x | \theta) = \frac{\partial^2 \lambda(x | \theta)}{\partial \theta^2}. \quad (20)$$

Definimos a **informação de Fisher** como

$$I(\theta) = E_{\theta} [\{\lambda'(x | \theta)\}^2] \stackrel{(1)}{=} -E_{\theta} [\lambda''(x | \theta)] = \text{Var}_{\theta} (\lambda'(x | \theta)). \quad (21)$$

Prova de ⁽¹⁾: diferenciar sob o sinal da integral e usar a regra da cadeia.

- Bernoulli;
- Normal;

Informação de Fisher: exemplos

- Bernoulli;

$$I(p) = \frac{1}{p(1-p)}.$$

- Normal;

$$I(\mu) = \frac{1}{\sigma^2}.$$

Teorema 21 (Informação de Fisher em uma amostra aleatória)

Seja $\mathbf{X} = \{X_1, X_2, \dots, X_n\}$ uma amostra aleatória e seja $I_n(\theta) = E_\theta [-\lambda_n''(\mathbf{X} | \theta)]$ a informação de Fisher da amostra. Então

$$I_n(\theta) = nI(\theta).$$

Prova: Usar as propriedades do log, da derivada e a lei de esperanças. Ver DeGroot, Teorema 8.8.2.

Voltando ao dilema de Palmirinha

Podemos usar a informação de Fisher para analisar os desenhos propostos por Palmirinha. Não é difícil derivar

$$I_X(\theta) = \frac{n}{\theta^2} \quad \text{e} \quad I_Y(\theta) = \frac{t}{\theta}.$$

Portanto, os desenhos são equivalentes se $n = t\theta$, o que não ajuda muito, já que θ é desconhecido. Por outro lado, vemos que neste caso não é possível decidir entre os desenhos baseado apenas na informação de Fisher.

Extra: faça uma análise Bayesiana deste problema, derivando a esperança *a priori* da informação de Fisher sob os dois desenhos.

O Teorema de Cramér-Rao

Outro uso importante da informação de Fisher é encontrar uma cota inferior para a variância de um estimador. Para isso, empregamos um dos resultados mais importantes da Estatística:

Teorema 22 (Teorema de Cramér-Rao¹²)

Seja $\mathbf{X} = \{X_1, X_2, \dots, X_n\}$ uma amostra aleatória com f.d.p./f.m.p $f(\mathbf{x} | \theta)$, com as mesmas premissas da definição ???. Suponha que $T = r(\mathbf{X})$ é uma estatística com variância finita. Seja $m(\theta) = E_\theta(T)$ uma função diferenciável de θ . Então,

$$\text{Var}_\theta(T) \geq \frac{[m'(\theta)]^2}{nI(\theta)}, \quad (22)$$

com igualdade apenas se existem u e v tal que

$$T = u(\theta)\lambda'_n(\mathbf{X} | \theta) + v(\theta).$$

Prova: Usar Cauchy-Schwarz e diferenciar sob o sinal da integral.

¹²Em homenagem ao estatístico indo-estadunidense Callyampudi Radhakrishna Rao (1920-) e ao matemático sueco Harald Cramér (1893–1985).

Se T é um estimador não-viesado, temos uma expressão útil para a cota de Cramér-Rao.

Observação 14 (Variância de um estimador não-viesado)

Se T é um estimador não-viesado de θ , temos

$$\text{Var}_{\theta}(T) \geq \frac{1}{nI(\theta)}$$

Prova: T é não viesado $\implies m(\theta) = \theta \implies m'(\theta) = 1 \forall \theta \in \Omega$ \square

Com esse Teorema de Cramér-Rao em mãos, estamos em posição de definir um critério de otimalidade para estimadores.

Definição 34 (Estimador eficiente)

Um estimador $\delta(\mathbf{X})$ é dito **eficiente** de (sua esperança) $m(\theta)$ se

$$\text{Var}_{\theta}(\delta) = \frac{[m'(\theta)]^2}{nI(\theta)}.$$

Exemplo 14

Seja X_1, X_2, \dots, X_n uma amostra aleatória de uma distribuição Poisson com parâmetro θ . Podemos mostrar que \bar{X}_n é um estimador eficiente de θ .

Distribuição assintótica de um estimador eficiente

Podemos usar o TCL para estudar a distribuição assintótica de um estimador eficiente.

Teorema 23 (Distribuição assintótica de um estimador eficiente)

Assumindo as condições de regularidade usuais, considere δ um estimador eficiente de $m(\theta)$. Assuma também que $m'(\theta) \neq 0 \forall \theta \in \Omega$. Então a distribuição assintótica de

$$\frac{\sqrt{nl(\theta)}}{m'(\theta)} [\delta - m(\theta)]$$

é normal padrão.

Prova: Ver DeGroot, Teorema 8.8.4. Escrever $E_{\theta}[\delta]$ e $\text{Var}_{\theta}(\delta)$ explicitamente, usar a condição $\delta = u(\theta)\lambda'_n(\mathbf{X} | \theta) + v(\theta)$, e aplicar as leis de esperanças e variâncias.

Observação 15 (Normalidade Assintótica do EMV)

Supondo que o EMV possa ser derivado ao resolver a equação $\lambda'_n(\mathbf{X} | \theta) = 0$ e que $\lambda''_n(\mathbf{X} | \theta)$ e $\lambda'''_n(\mathbf{X} | \theta)$ satisfazem certas condições técnicas, $\sqrt{nl(\theta)} (\hat{\theta}_{EMV} - \theta)^2$ tem distribuição aproximadamente normal padrão.


O que aprendemos?

- 💡 Viés;
“Um estimador viesado é aquele cuja esperança não coincide com a função estimada”
- 💡 Informação de Fisher;
“A informação de Fisher é uma quantidade derivada de uma distribuição que mede a quantidade de informação contida em uma amostra aleatória advinda desta distribuição”
- 💡 Cramér-Rao;
“A desigualdade de Cramér-Rao dá uma cota inferior para a variância de um estimador”
- 💡 Distribuição assintótica de estimadores eficientes (e EMV);
“Sob condições de regularidade, vale um TCL para estimadores eficientes e para o EMV”

Leitura recomendada

 DeGroot seções 8.7 e 8.8;

 * Casella & Berger (2002), seção 7.3.

 * Schervish (1995), Teorema 5.13.

►► Próxima aula: DeGroot, seções 8.1 e 8.2;

• Exercícios recomendados

■ DeGroot.

Seção 8.7: exercícios 4, 6, 11 e 13;

Seção 8.8: exercícios 5, 7 e 10.

- Distribuição amostral de uma estatística;
- A família qui-quadrado de distribuições Gamma;
- Exemplos.

Se $\mathbf{X} = \{X_1, X_2, \dots, X_n\}$ é uma amostra aleatória, $T = r(X_1, X_2, \dots, X_n)$ é uma variável aleatória, e portanto, faz sentido falar da distribuição de T .

Exemplo 15 (Distribuição amostral de uma proporção)

(Exemplo 8.1.1 em DeGroot)

Suponha que estamos interessados na proporção de pacientes que se recrudescem após tratamento com uma determinada droga. Para uma amostra de n pacientes, podemos modelar os desfechos como variáveis aleatórias i.i.d. Bernoulli com parâmetro θ e computar $T = n^{-1} \sum_{i=1}^n X_i$ como estimativa de θ . Deste modo, temos

$$\Pr(T = t) = \begin{cases} \binom{n}{nt} \theta^{nt} (1 - \theta)^{n(1-t)}, & t = \frac{0}{n}, \frac{1}{n}, \dots, \frac{n-1}{n}, \frac{n}{n}, \\ 0, & \text{caso contrário.} \end{cases} \quad (23)$$

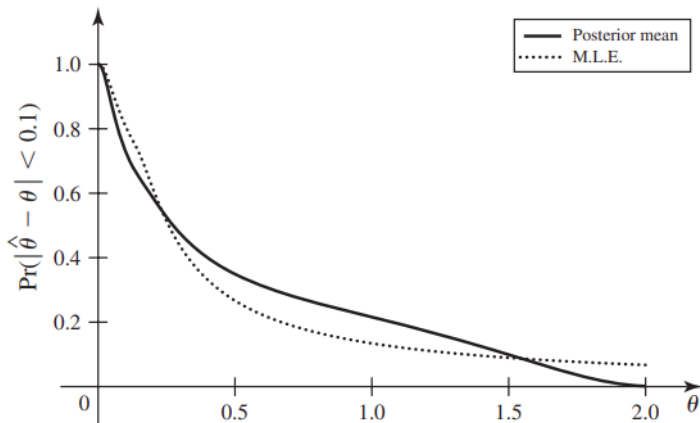
Chamamos (??) de **distribuição amostral** de T .

Relembre o exemplo das lâmpadas de Astolfo:

$$\hat{\theta}_{\text{Bayes}} = \frac{\alpha + n}{\beta + S}; \hat{\theta}_{\text{EMV}} = \frac{n}{S}.$$

Podemos perguntar,

$$\Pr \left(|\hat{\theta} - \theta| < a \right) = ?$$



Definição 35 (Distribuição qui-quadrado)

Dizemos que uma variável aleatória Y tem distribuição **qui-quadrado** com m graus de liberdade quando

$$f_Y(y) = \frac{1}{2^{m/2}\Gamma(m/2)} y^{m/2-1} e^{-y/2}, \quad y > 0. \quad (24)$$

Vemos que Y tem função geradora de momentos

$$\psi(t) = \left(\frac{1}{1-2t} \right)^{m/2}, \quad t < 1/2.$$

$E[Y] = ?$, $\text{Var}(Y) = ?$

Alguns resultados úteis

Teorema 24 (Soma de variáveis aleatórias qui-quadrado)

Se X_1, X_2, \dots, X_n são variáveis aleatórias independentes com graus de liberdade m_i , então $W = \sum_{i=1}^n X_i$ tem distribuição qui-quadrado com graus de liberdade $m = \sum_{i=1}^n m_i$.

Prova: Segue da soma de variáveis aleatórias Gama.

Teorema 25 (Distribuição do quadrado de uma variável aleatória Normal padrão)

Se $X \sim \text{Normal}(0, 1)$, $Y = X^2$ tem distribuição qui-quadrado com $m = 1$.

Prova: Escrever a acumulada de Y , diferenciar e usar a regra da cadeia.

Observação 16 (Distribuição da soma de quadrados de normais padrão)

Se X_1, X_2, \dots, X_n são variáveis aleatórias Normal padrão, então $Z = \sum_{i=1}^n X_i^2$ tem distribuição qui-quadrado com n graus de liberdade.

Prova: Imediato dos dois últimos teoremas.

Distribuição da variância amostral

Vamos a um exemplo motivador. No caso Normal, quando μ é conhecida, temos o estimador de máxima verossimilhança para a variância:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2.$$

Isso nos leva às duas próximas observações

Observação 17 (Uma transformação linear do EMV)

$$\frac{n\hat{\sigma}^2}{\sigma^2} \sim \text{qui-quadrado}(n).$$

Prova: Notar que $Z_i = (X_i - \mu)/\sigma$ são Normal padrão e aplicar a observação ??.

Observação 18 (Distribuição do EMV da variância)

$$\hat{\sigma}^2 \sim \text{Gama} \left(\frac{n}{2}, \frac{n}{2\sigma^2} \right).$$

Prova: Exercício 13 da seção 8.2 de DeGroot.

Exemplo 16 (Concentração de ácido no queijo)

Suponha que estamos interessados em medir a concentração de um certo ácido em pedaços de queijo produzidos por uma fábrica. Ao longo dos anos, grande acúmulo de dados permitiu afirmar que a distribuição populacional da concentração é Normal com parâmetros μ e σ^2 . Suponha que amostramos n pedaços e medimos as concentrações X_1, X_2, \dots, X_n . Então

$$Y = \frac{1}{n} \sum_{i=1}^n |X_i - \mu|^2$$

é uma medida de quanto estas amostras desviam da concentração típica μ . Suponha que uma diferença de concentração u é o suficiente para dar gosto diferente ao queijo. Podemos calcular $\Pr(Y \leq u^2)$ para quantificar o risco de isso acontecer.

O que aprendemos?

- 💡 Distribuição amostral;
“Estatísticas e estimadores são variáveis aleatórias e têm distribuições amostrais”
- 💡 A distribuição qui-quadrado;
“A soma de quadrados de variáveis aleatórias gaussianas é um tipo especial de distribuição Gama”
- 💡 Avaliação probabilística de estimadores;
“Podemos utilizar a distribuição amostral para fazer afirmações sobre quantidades como $|\hat{\theta} - \theta|$ ”

Leitura recomendada

 DeGroot seções 8.1 e 8.2;

▶▶ Próxima aula: DeGroot, seções 8.3 e 8.4;

- **Exercícios recomendados**

- DeGroot.

- Seção 8.1: exercícios 1, 2, 3 e 9;

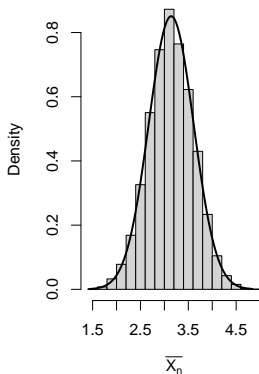
- Seção 8.2: exercícios 4, 7, 10 e 13.

- Distribuição conjunta de \bar{X}_n e \bar{S}_n^2 ;
- No caso Normal, $\bar{X}_n \perp\!\!\!\perp \bar{S}_n^2$ são independentes!
- Distribuição t de Student.

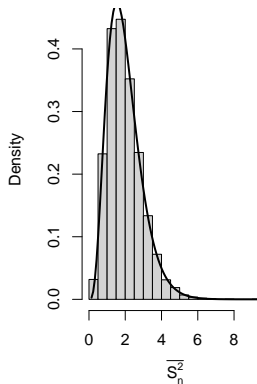
Distribuição de \bar{X}_n e \bar{S}_n^2

- $\bar{X}_n \sim \text{Normal}\left(\mu, \frac{\sigma^2}{n}\right)$;
- $\bar{S}_n^2 \sim \text{Gama}\left(\frac{n-1}{2}, \frac{n}{2\sigma^2}\right)$

Média amostral



Variância amostral



Um Teorema importante

Aqui vamos ver um caso especial do Teorema de Basu¹³, que fala que os dois primeiros momentos amostrais da distribuição Normal são independentes.

Teorema 26 (Independência da média e variância amostrais na Normal)

Seja X_1, X_2, \dots, X_n uma amostra aleatória de uma distribuição Normal com parâmetros μ e σ^2 . Então a média amostral, \bar{X}_n e a variância amostral, \bar{S}_n^2 , são independentes. Ademais, $\bar{X}_n \sim \text{Normal}\left(\mu, \frac{\sigma^2}{n}\right)$ e $\bar{S}_n^2 \sim \text{Gama}\left(\frac{n-1}{2}, \frac{n}{2\sigma^2}\right)$.

Prova: Troca de variáveis em duas dimensões; propriedades de matrizes ortogonais. Ver Teorema 8.3.1 em DeGroot (prova na pág. 476).

¹³Debabrata Basu (1924–2001) foi um importante estatístico indiano.

Exemplo

Suponha que queremos determinar o tamanho de amostra, n , de modo que os EMVs da média μ e do desvio padrão σ estejam “perto” dos seus valores verdadeiros. Formalmente, queremos encontrar n tal que

$$\Pr \left(|\hat{\mu} - \mu| \leq \frac{1}{5}\sigma \text{ e } |\hat{\sigma} - \sigma| \leq \frac{1}{5}\sigma \right) \geq \frac{1}{2},$$

seja satisfeito.

A distribuição t de Student

Qual a distribuição de $\frac{\sqrt{n}(\bar{X}_n - \mu)}{\hat{\sigma}}$? A resposta é a distribuição t de “Student”¹⁴

Definição 36 (A distribuição t)

Considere duas variáveis aleatórias, $Y \sim \text{Qui-quadrado}(m)$ e $Z \sim \text{Normal}(0, 1)$ e defina a variável aleatória

$$X = \frac{Z}{\sqrt{\frac{Y}{m}}}.$$

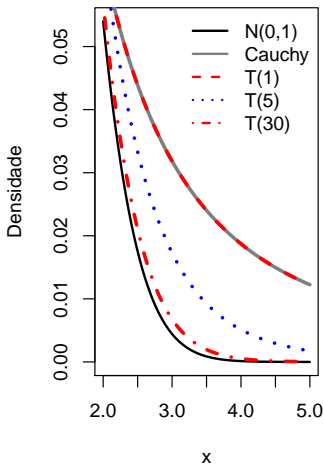
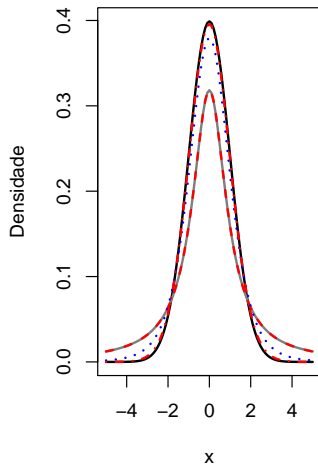
Dizemos que X tem distribuição **t de Student com m graus de liberdade**. Sabemos ainda que

$$f_X(x) = \frac{\Gamma(\frac{m+1}{2})}{\sqrt{m\pi}\Gamma(\frac{m}{2})} \left(1 + \frac{x^2}{m}\right)^{-\frac{m+1}{2}}, \quad x \in (-\infty, \infty).$$

Para $m > 2$, $E[X] = 0$ (porquê?) e $\text{Var}(X) = m/(m - 2)$.

¹⁴William Sealy Gosset (1876–1937) foi um estatístico inglês que, em 1908, publicou o resultado acima sob o pseudônimo “Student”, ou estudante/aluno.

Comparando a t com outras distribuições



Teorema 27 (Distribuição amostral do estimador não-viesado da variância)

Considere o estimador

$$\hat{\sigma}' = \sqrt{\frac{\Delta^2}{n-1}},$$

onde $\Delta^2 = \sum_{i=1}^n (X_i - \bar{X}_n)^2$. Então

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\hat{\sigma}'} \sim \text{Student}(n-1).$$

Prova: Ver Teorema 8.4.2 em DeGroot. Defina $Z = \sqrt{n}(\bar{X}_n - \mu)/\sigma$ e $Y = \Delta^2/\sigma^2$. Então $Z \sim \text{Normal}(0, 1)$ e $Y \sim \text{Qui-quadrado}(n-1)$. Faça

$$U = \frac{Z}{\sqrt{\frac{Y}{n-1}}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sqrt{\frac{\Delta^2}{n-1}}}, \quad (25)$$

e note que $U \sim T(n-1)$ \square

O que aprendemos?

- 💡 Independência dos momentos amostrais da Normal;
 “Numa amostra aleatória Normal, \bar{X}_n e \bar{S}_n^2 são independentes e $\bar{X}_n \sim \text{Normal}\left(\mu, \frac{\sigma^2}{n}\right)$ e $\bar{S}_n^2 \sim \text{Gama}\left(\frac{n-1}{2}, \frac{n}{2\sigma^2}\right)$.”
- 💡 A distribuição t de Student;
 “A diferença padronizada entre a média amostral e a média populacional (μ) tem distribuição t de Student, que não depende de σ^2 ”

Leitura recomendada

 DeGroot seções 8.3 e 8.4;

▶▶ Próxima aula: DeGroot, seção 8.5;

- **Exercícios recomendados**

- DeGroot.

- Seção 8.3: exercício 8;

- Seção 8.4: derivar a densidade da Distribuição t de Student.

Intervalos de confiança

- Intervalos de confiança;
- Caso normal: média;
- Intervalos de confiança unilaterais;
- Estatística pivotal.

Lembremos que

$$U = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sqrt{\frac{\Delta^2}{n-1}}} \sim T(n-1). \quad (26)$$

Para $c > 0$, podemos computar $\Pr(-c < U < c) = \gamma$:

$$\begin{aligned} \Pr\left(-c < \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sqrt{\frac{\Delta^2}{n-1}}} < c\right) &= \gamma, \\ \Pr\left(\bar{X}_n - \frac{c\hat{\sigma}'}{\sqrt{n}} < \mu < \bar{X}_n + \frac{c\hat{\sigma}'}{\sqrt{n}}\right) &= \gamma, \\ T_{n-1}(c) - T_{n-1}(-c) &= 2T_{n-1}(c) - 1 = \gamma. \end{aligned}$$

Concluimos que $c = F_T^{-1}\left(\frac{1+\gamma}{2}; n-1\right)$.

Definição de intervalo de confiança

O conceito de **intervalo de confiança** é fundamental em Estatística e nas aplicações em Ciência.

Definição 37 (Intervalo de confiança)

Seja $\mathbf{X} = \{X_1, X_2, \dots, X_n\}$ uma amostra aleatória, cada variável aleatória com p.d.f. $f(x | \theta)$, e considere uma função real $g(\theta)$. Sejam $A(\mathbf{X})$ e $B(\mathbf{X})$ duas estatísticas de modo que valha

$$\Pr \{A(\mathbf{X}) < g(\theta) < B(\mathbf{X})\} \geq \gamma. \quad (27)$$

Dizemos que $I(\mathbf{X}) = (A(\mathbf{X}), B(\mathbf{X}))$ é um **intervalo de confiança** de $100\gamma\%$ para $g(\theta)$. Se a desigualdade for uma igualdade para todo $\theta \in \Omega$, dizemos que o intervalo é **exato**.

No caso do intervalo de confiança para o parâmetro de média, temos

$$\Pr \{A(\mathbf{X}) < g(\mu) < B(\mathbf{X})\} \geq \gamma,$$

com $g(\mu) = \mu$ e

$$A(\mathbf{X}) = \bar{X}_n - \frac{c\hat{\sigma}'}{\sqrt{n}} = \bar{X}_n - \frac{c\sqrt{\sum_{i=1}^n (X_i - \bar{X}_n)^2}}{\sqrt{n(n-1)}},$$
$$B(\mathbf{X}) = \bar{X}_n + \frac{c\hat{\sigma}'}{\sqrt{n}} = \bar{X}_n + \frac{c\sqrt{\sum_{i=1}^n (X_i - \bar{X}_n)^2}}{\sqrt{n(n-1)}}.$$

Interpretação de um intervalo de confiança

ATENÇÃO: a interpretação de um intervalo é crucial. Muita gente confunde o que um intervalo de confiança significa!

Observação 19 (Um intervalo de confiança não é uma afirmação sobre o(s) parâmetro(s)!)

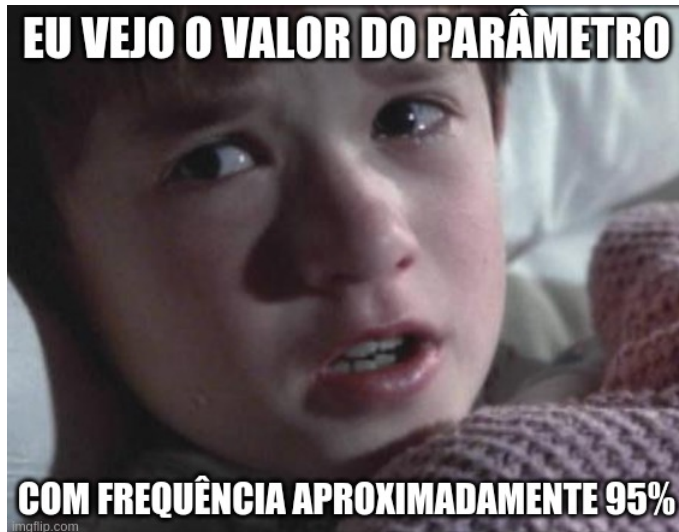
A afirmação probabilística da forma $\Pr \{A(\mathbf{X}) < g(\theta) < B(\mathbf{X})\} = \gamma$ diz respeito à distribuição conjunta das variáveis aleatórias $A(\mathbf{X})$ e $B(\mathbf{X})$ para um valor fixo de θ – e, portanto, de $g(\theta)$.

Ideia 3 (Intervalos de confiança são procedimentos)

Como de costume na teoria ortodoxa (frequentista), o foco da construção de um intervalo confiança está em dar garantias probabilísticas **com relação à distribuição dos dados**. Dizer que $\Pr \{A(\mathbf{X}) < g(\theta) < B(\mathbf{X})\} = \gamma$ é dizer que, se eu gerasse M grande amostras aleatórias $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \dots, \mathbf{X}^{(M)}$ de tamanho n e construísse M intervalos $I(\mathbf{X}^{(1)}), I(\mathbf{X}^{(2)}), \dots, I(\mathbf{X}^{(M)})$, eu esperaria encontrar:

$$\frac{1}{M} \sum_{i=1}^M \mathbb{I} \left(g(\theta) \in I(\mathbf{X}^{(i)}) \right) \approx \gamma.$$

I see $g(\theta)$...



Intervalos unilaterais

Em várias situações, estamos interessados em uma cota superior ou inferior para $g(\theta)$.

Definição 38 (Intervalo de confiança unilateral)

Seja $\mathbf{X} = \{X_1, X_2, \dots, X_n\}$ uma amostra aleatória, cada variável aleatória com p.d.f. $f(x | \theta)$, e considere uma função real $g(\theta)$. Seja $A(\mathbf{X})$ uma estatística que, para todo $\theta \in \Omega$, valha

$$\Pr \{A(\mathbf{X}) < g(\theta)\} \geq \gamma,$$

dizemos que o intervalo aleatório $(A(\mathbf{X}), \infty)$ é chamado um intervalo de confiança **unilateral** de $100\gamma\%$ para $g(\theta)$, ou, ainda, um intervalo de confiança **inferior** de $100\gamma\%$ para $g(\theta)$. O intervalo $(-\infty, B(\mathbf{X}))$, com

$$\Pr \{g(\theta) < B(\mathbf{X})\} \geq \gamma,$$

é definido de forma análoga, e é chamado de intervalo de confiança **superior** de $100\gamma\%$ para $g(\theta)$. Se a desigualdade é uma igualdade para todo $\theta \in \Omega$, os intervalos são chamados exatos.

O conceito de quantidade pivotal é útil na construção de intervalos de confiança.

Definição 39 (Quantidade pivotal)

Seja $\mathbf{X} = \{X_1, X_2, \dots, X_n\}$ uma amostra aleatória com p.d.f. $f(x | \theta)$. Seja $V(\mathbf{X}, \theta)$ uma variável aleatória cuja distribuição é **a mesma** para todo $\theta \in \Omega$. Dizemos que $V(\mathbf{X}, \theta)$ é uma **quantidade pivotal**.

Podemos utilizar quantidades pivotaís para construir intervalos de confiança. Considere uma função $r(v, x)$ tal que

$$r(V(\mathbf{X}, \theta), \mathbf{X}) = g(\theta).$$

Construindo ICs a partir de quantidades pivotaís

Vamos ver como usar $r(v, \mathbf{x})$ para construir um intervalo de confiança.

Teorema 28 (Intervalos de confiança a partir de uma quantidade pivotal)

Seja $\mathbf{X} = \{X_1, X_2, \dots, X_n\}$ uma amostra aleatória com p.d.f. $f(x | \theta)$. Suponha que existe uma quantidade pivotal V , com c.d.f. contínua G . Assuma que existe $r(v, \mathbf{x})$, estritamente crescente em v para todo \mathbf{x} . Finalmente, tome $0 < \gamma < 1$ e $\gamma_1 < \gamma_2$ de modo que $\gamma_2 - \gamma_1 = \gamma$. Então as estatísticas

$$A(\mathbf{X}) = r(G^{-1}(\gamma_1), \mathbf{X}),$$

$$B(\mathbf{X}) = r(G^{-1}(\gamma_2), \mathbf{X}),$$

são os limites de um intervalo de confiança de $100\gamma\%$ para $g(\theta)$.

Prova: Usar a monotonicidade de r e de G e notar que

$$\Pr(A(\mathbf{X}) = g(\theta)) = \Pr(V(\mathbf{X}, \theta) = G^{-1}(\gamma_1)) = 0,$$

e que o mesmo vale para $B(\mathbf{X})$. Ver Teorema 8.5.3 em DeGroot.

Exemplos

- Exponencial;
- Normal, μ e σ^2 desconhecidas;
- Normal, σ^2 conhecida;

Exemplos

- Exponencial: $\theta S \sim \text{Gama}(n, 1)$;
- Normal, μ e σ^2 desconhecidas:
 - ◊ $\frac{\bar{X}_n - \mu}{\hat{\sigma}} \sim \text{Student}(n - 1)$;
 - ◊ $\frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{\sigma^2} \sim \text{Qui-quadrado}(n - 1)$;
- Normal, σ^2 conhecida: $\frac{\bar{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim \text{Normal}(0, 1)$;

Limitações

Intervalos de confiança estão entre as ferramentas mais importantes da Estatística. Isso não quer dizer que não tenham limitações importantes.

- Interpretação; Uma vez observados $a(x)$ e $b(x)$, não é correto dizer que $g(\theta)$ mora em (a, b) com probabilidade γ . “Antes de observarmos o valor tomado por \mathbf{X} , há probabilidade γ de que o intervalo $I(\mathbf{X})$, construído a partir da amostra \mathbf{X} , inclui $g(\theta)$.”

Em geral falamos de **confiança** γ do intervalo $I(\mathbf{X})$.

- Uso da informação; Uma vez que observamos $I(x) = (a(x), b(x))$, pode haver informação extra sobre se $I(x)$ cobre $g(\theta)$ ou não, mas não existe maneira canônica de ajustar o nível de confiança γ à luz desta nova informação. Ver exemplo 8.5.11 em DeGroot.

O que aprendemos?

- 💡 Intervalos de confiança;
 “Um intervalo $(A(\mathbf{X}), B(\mathbf{X}))$ de confiança de $100\gamma\%$ para $g(\theta)$ é tal que $\Pr[A(\mathbf{X}) < g(\theta) < B(\mathbf{X})] \geq \gamma$ ”;
- 💡 Um intervalo de confiança é uma afirmação probabilística sobre **as estatísticas** $A(\mathbf{X})$ e $B(\mathbf{X})$ a partir da **distribuição conjunta dos dados**;
- 💡 Quantidade pivotal “Uma quantidade pivotal é uma função $V(\mathbf{X}, \theta)$ cuja distribuição não depende de θ ”
- 💡 Intervalos de confiança podem ser construídos a partir de quantidades pivotaís;

Leitura recomendada

 DeGroot seção 8.5;

 * Casella & Berger (2002), seção 9.2.

▶▶ Próxima aula: DeGroot, seção 9.1;

- **Exercícios recomendados**

- DeGroot.

Seção 8.5: 1, 4, 5 e 6.

Testes de hipóteses

- Hipótese nula e alternativa;
- Hipóteses simples e compostas;
- Região crítica e estatística teste;
- Função poder;
- Tipos de erro (I e II);
- P-valor;

Hipótese nula e alternativa

No teste de hipóteses estatísticas, identificamos partições do espaço de parâmetros que codificam as hipóteses de interesse.

Definição 40 (Hipótese nula e hipótese alternativa)

Considere o espaço de parâmetros Ω e defina $\Omega_0, \Omega_1 \subset \Omega$ de modo que $\Omega_0 \cup \Omega_1 = \Omega$ e $\Omega_0 \cap \Omega_1 = \emptyset$. Definimos

$$H_0 := \theta \in \Omega_0,$$

$$H_1 := \theta \in \Omega_1.$$

Dizemos que H_0 é a **hipótese nula** e H_1 é a **hipótese alternativa**.

Se $\theta \in \Omega_1$, dizemos que rejeitamos a hipótese nula. Por outro lado, se $\theta \in \Omega_0$ dizemos que não rejeitamos ou falhamos em rejeitar H_0 .

Suponha que Palmirinha recebeu uma carta da Associação Nacional da Pamonha Gourmet (ANPG), dizendo que a pamonha deve ter, no mínimo, 7 mg/L de concentração de amido. Supondo que a concentração de amido tenha distribuição Normal com parâmetros μ (desconhecido) e σ^2 (conhecido), Palmirinha rabisca num papel:

$$H_0 : \mu \in [7, \infty),$$

$$H_1 : \mu \in (0, 7).$$

Hipóteses simples e compostas

Dependendo do tipo de partição do espaço de parâmetros, as hipóteses recebem classificações diferentes.

Definição 41 (Hipótese simples e hipótese compostas)

*Dizemos que uma hipótese H_i , é **simples**, se $\Omega_i = \{\theta_i\}$, isto é, se a partição correspondente é um ponto. Uma hipótese é dita **composta** se não é simples.*

Exemplo 17 (Hipótese simples sobre a média)

Suponha que estamos estudando o efeito de uma droga na redução da pressão arterial. Modelamos esta redução como uma variável aleatória X com esperança $E[X] =: \theta$. É costumaz testar a hipótese $H_0 : \theta = 0$, que chamamos, especificamente nesse caso, de “hipótese de efeito nulo”.

Hipótese unilateral e hipótese bilateral

Em analogia com os intervalos de confiança, também podemos entender as hipóteses como sendo unilaterais ou bilaterais.

Definição 42 (Hipótese unilateral e hipótese bilateral)

Uma hipótese da forma $H_0 : \theta \leq \theta_0$ ou $H_0 : \theta \geq \theta_0$ é dita unilateral (“one-sided”), enquanto hipóteses da forma $H_0 : \theta \neq \theta_0$ são ditas bilaterais (“two-sided”).

Observação 20 (Hipóteses bilaterais como consequência de H_0 simples)

Se H_0 é simples, a hipótese alternativa H_1 será, em geral, bilateral.

Região crítica: exemplo motivador

Exemplo 18 (Teste para a média de uma Normal com variância conhecida)

Suponha que $\mathbf{X} = \{X_1, X_2, \dots, X_n\}$ é uma amostra aleatória de uma Normal com média μ e variância σ^2 conhecida. Queremos testar a hipótese

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu \neq \mu_0.$$

Intuitivamente, queremos rejeitar H_0 se \bar{X}_n está longe de μ_0 . Para isso definimos

$$S_0 := \{x : -c \leq \bar{X}_n - \mu_0 \leq c\},$$

de modo que $S_1 = S_0^C$. Então, seguimos o procedimento:

$$\mathbf{X} \in S_1 \implies \text{rejeitar } H_0,$$

$$\mathbf{X} \in S_0 \implies \text{não rejeitar } H_0.$$

Região crítica e região de rejeição

Uma maneira mais simples de expressar o procedimento acima é definir $T := |\bar{X}_n - \mu_0|$ e rejeitar H_0 se $T \geq c$.

Definição 43 (Região crítica)

O conjunto

$$S_1 := \{x : |\bar{X}_n - \mu_0| \geq c\},$$

é chamado de **região crítica** do teste.

Analogamente, considere a estatística $T = r(\mathbf{X})$ e tome $R \subseteq \mathbb{R}$. Então podemos definir

Definição 44 (Região de rejeição)

Se $R \subseteq \mathbb{R}$ é tal que dizemos que “rejeitamos H_0 se $T \in R$ ”, então R é chamada uma **região de rejeição** para a estatística T e o teste associado.

Começamos com uma observação:

Observação 21 (Correspondência entre região crítica e região de rejeição)

Podemos relacionar os conceitos de região crítica e região de rejeição notando queremos

$$S_1 := \{x : r(x) \in R\}.$$

Ideia 4 (Dividindo o espaço amostral e o espaço de parâmetros)

*Suponha que temos um modelo estatístico dado pela distribuição $f(x | \theta)$, com $x \in \mathcal{X}$ e $\theta \in \Omega$. Desta forma, uma amostra aleatória $\mathbf{X} = \{X_1, X_2, \dots, X_n\}$ mora em \mathcal{X}^n . Para formular uma hipótese estatística, estabelecemos uma partição do espaço de parâmetros Ω em Ω_0 e Ω_1 disjuntos. Isto, por sua vez, induz uma partição $S_0, S_1 \in \mathcal{X}^n$. Estes objetos, embora, relacionados, **não são a mesma coisa**. Por exemplo, nós observamos se $\mathbf{X} \in S_0$ ou $\mathbf{X} \in S_1$, mas raramente “observamos” se $\theta \in \Omega_0$ ou $\theta \in \Omega_1$.*

Nossa capacidade de rejeitar H_0 depende do valor de $\theta \in \Omega$. Esta dependência é capturada pela função poder.

Definição 45 (Função poder)

*Seja δ um procedimento de aceitação/rejeição como visto anteriormente. A **função poder** é definida como*

$$\pi(\theta \mid \delta) := \Pr(\mathbf{X} \in S_1 \mid \theta) = \Pr(T \in R \mid \theta), \theta \in \Omega. \quad (28)$$

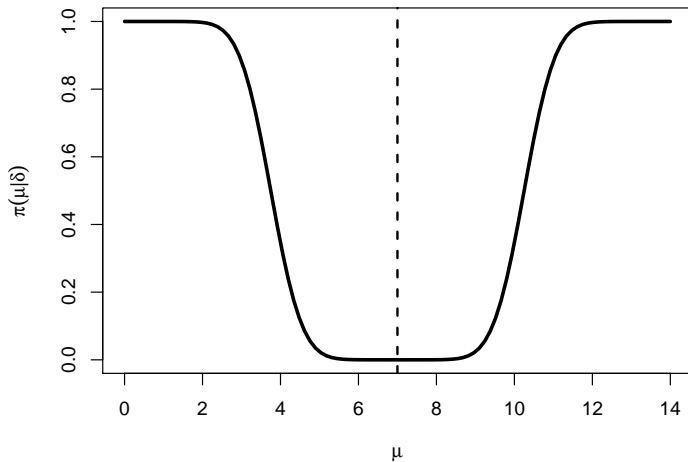
Idealmente, queremos $\pi(\theta \mid \delta) = 1$ para $\theta \in \Omega_1$ (por quê?).

Considere a situação em que X_1, X_2, \dots, X_n vêm de uma Normal com média μ , desconhecida, e variância σ^2 , conhecida.

Exemplo 19 (Função poder no teste para média da Normal (σ^2 conhecida))

Lembrando que $T = |\bar{X}_n - \mu_0|$, e tomando δ como o procedimento descrito acima, escrevemos

$$\begin{aligned}\pi(\mu \mid \delta) &= \Pr(T \in R \mid \mu), \\ &= \Pr(\bar{X}_n \geq \mu_0 + c \mid \mu) + \Pr(\bar{X}_n \leq \mu_0 - c \mid \mu), \\ &= \left\{ 1 - \Phi\left(\sqrt{n} \frac{\mu_0 + c - \mu}{\sigma}\right) \right\} + \Phi\left(\sqrt{n} \frac{\mu_0 - c - \mu}{\sigma}\right).\end{aligned}$$



Tipos de Erro

Quando testamos uma hipótese, nunca estamos livres de cometer um erro. É conveniente classificar os possíveis erros em duas categorias.

Definição 46 (Tipos de erros)

Nome	Erro cometido
<i>Erro tipo I</i>	<i>Rejeitar H_0 quando ela é verdadeira.</i>
<i>Erro tipo II</i>	<i>Falhar em rejeitar H_0 quando ela é falsa.</i>

Isto nos leva a concluir que

Situação	Quantidade	Interpretação
$\theta \in \Omega_0$	$\pi(\theta \delta)$	Pr(Erro tipo I)
$\theta \in \Omega_1$	$1 - \pi(\theta \delta)$	Pr(Erro tipo II)

Idealmente, gostaríamos de um teste δ para o qual as probabilidades de erro fossem as menores possíveis. Infelizmente, em geral, diminuir o erro tipo I implica aumentar o erro tipo II.

Em geral, precisamos encontrar um equilíbrio entre os tipos de erros.

Ideia 5 (Encontrando um balanço entre erro tipo I e tipo II)

Tome $0 < \alpha_0 < 1$. Nós construímos o procedimento δ^ de modo que*

$$\pi(\theta \mid \delta^*) \leq \alpha_0, \forall \theta \in \Omega. \quad (29)$$

Então, entre todos os testes que satisfazem (??), buscamos o teste que tenha $\pi(\theta \mid \delta^)$ máxima em $\theta \in \Omega_1$.*

Tamanho de um teste

Definição 47 (Tamanho/nível de um teste)

Dizemos que um teste, δ , tem **tamanho** ou **nível de significância** $\alpha(\delta)$, com

$$\alpha(\delta) := \sup_{\theta \in \Omega_0} \pi(\theta \mid \delta).$$

Um teste que atende à condição anterior (??) tem que tamanho?

Observação 22 (Tamanho de um teste com H_0 simples)

Se H_0 é simples, então $\alpha(\delta) = \pi(\theta_0 \mid \delta)$.

Exemplo 20 (Teste para o parâmetro de uma uniforme)

Suponha que X_1, X_2, \dots, X_n tem distribuição Uniforme em $[0, \theta]$, com θ desconhecido, e que aventamos as seguintes hipóteses:

$$H_0 : 3 \leq \theta \leq 4,$$

$$H_1 : \theta < 3 \text{ ou } \theta > 4$$

Lembre que $\hat{\theta}_{EMV} = \max\{X_1, X_2, \dots, X_n\}$ e suponha que temos um teste δ da forma

Condição	Ação
$\hat{\theta}_{EMV} \notin (2.9, 4)$	Rejeitar H_0
$\hat{\theta}_{EMV} \in (2.9, 4)$	Falhar em rejeitar H_0 .

- Qual a região de rejeição para δ ?
- Como escrever $\pi(\theta \mid \delta)$?
- Qual o tamanho de δ ?

Em geral, sempre conseguimos construir um teste que tenha o tamanho desejado.

Observação 23 (Construindo um teste de tamanho α_0)

Se $T = r(\mathbf{X})$ é uma estatística, podemos quase sempre encontrar c tal que valha

$$\sup_{\theta \in \Omega_0} \Pr(T \geq c \mid \theta) \leq \alpha_0, \quad (30)$$

ou seja, encontrar c tal que δ tenha tamanho (ou nível de significância) α_0 .

O p-valor

Começamos com uma observação:

Observação 24 (Testes são decisões binárias)

Um teste de hipótese reduz a informação contida nos dados a uma decisão binária: rejeitar ou não H_0 . Se observamos $T = c + 10^{-10}$ ou $T = c + 10^{10}$, tomamos a mesma decisão de rejeitar H_0 ao nível α_0 .

Ao invés disso, podemos reportar o maior nível de significância que ainda levaria à rejeição de H_0 .

Definição 48 (O p-valor)

*Para cada t , seja δ_t o teste que rejeita H_0 se $T \geq t$. Então, quando $T = t$, o **p-valor** vale*

$$p(t) := \sup_{\theta \in \Omega_0} \pi(\theta \mid \delta_t) = \sup_{\theta \in \Omega_0} \Pr(T \geq t \mid \theta), \quad (31)$$

ou seja, o p-valor é o tamanho do teste δ_t .

Exemplo: tá me enganando, parceiro?

Vamos voltar a uma pergunta não respondida lá no início do curso. Suponha que você encontre um “artista” de rua, que joga uma moeda e pede para as pessoas apostarem se vai dar cara ou coroa. Conhecendo estatística e probabilidade, você decide (i) observar o jogo à distância para coletar alguns dados (ii) fazer algumas contas para ver se vale a pena apostar.

Pergunta 4 (Esta moeda é justa? cont. I)

Suponha que uma moeda tenha sido lançada dez vezes, obtendo o seguinte resultado:

KKKCKCCCKC

- a) Esta moeda é justa?*
- b) Quanto eu espero ganhar se apostar R\$ 100,00 que é justa?*


Hoje vamos dar uma resposta parcial à pergunta a).

O que aprendemos?

- 💡 Hipóteses nula e alternativa, simples e composta;
- 💡 Região crítica e região de rejeição;
- 💡 Função poder;
 “O poder de um teste é a probabilidade de rejeitarmos H_0 caso ela seja falsa”
- 💡 Erro tipo I e tipo II;
 - ◊ Tipo I: Rejeitar erroneamente H_0 ;
 - ◊ Tipo II: Falhar em rejeitar H_0 quando ela é falsa.
- 💡 P-valor;
 “O p-valor pode ser interpretado como a probabilidade, sob H_0 , de observarmos uma estatística tão ou mais extrema do que aquela que foi observada”

Leitura recomendada

 DeGroot seção 9.1;

 * DeGroot seções 9.2 e 9.3.

► Próxima aula: DeGroot, seção 9.1 (razão de verossimilhanças);

- **Exercícios recomendados**

- DeGroot.

- Seção 9.1: 3, 8 e 13.

- * Seção 9.1: 19 e 21.

Razões de verossimilhanças

- Intervalos de confiança e testes;
- Razões de verossimilhanças

Intervalos de confiança \equiv testes

De posse de um intervalo de confiança, podemos testar hipóteses sobre uma função dos parâmetros, $g(\theta)$, como mostra o seguinte teorema:

Teorema 29 (Intervalos de confiança e testes são equivalentes)

Suponha que dispomos de dados $\mathbf{X} = \{X_1, X_2, \dots, X_n\}$ com f.d.p. comum $f(x | \theta)$, e estamos interessados em testar as hipóteses:

$$H_0 : g(\theta) = g_0,$$

$$H_1 : g(\theta) \neq g_0,$$

de modo que existe um teste δ_{g_0} com nível α_0 destas hipóteses. Para cada $\mathbf{X} = \mathbf{x}$, defina

$$w(\mathbf{x}) = \{g_0 : \delta_{g_0} \text{ não rejeita } H_0 \text{ dado que } \mathbf{X} = \mathbf{x}\}.$$

Fazendo o nível de confiança do intervalo $\gamma = 1 - \alpha_0$, temos

$$\Pr(g(\theta_0) \in w(\mathbf{X}) | \theta = \theta_0) \geq \gamma, \forall \theta_0 \in \Omega.$$

Prova: Notar que $\Pr(\delta_{g_0} \text{ não rejeita } H_0 | \theta = \theta_0) \geq \alpha_0 = 1 - \gamma$ e concluir que $w(\mathbf{X})$ é uma região de crítica para δ_{g_0} . Ver Teorema 9.1.1 de DeGroot.

Conjunto de confiança

O conjunto $w(\mathbf{X})$ definido acima pode ser entendido como um conjunto de confiança para $g(\theta)$.

Definição 49 (Conjunto de confiança)

Se um conjunto aleatório $w(\mathbf{X})$ satisfaz

$$\Pr(g(\theta_0) \in w(\mathbf{X}) \mid \theta = \theta_0) \geq \gamma,$$

para todo $\theta_0 \in \Omega$, então chamamos $w(\mathbf{X})$ de um **conjunto de confiança** para $g(\theta)$.

Isso nos leva ao seguinte teorema

Teorema 30 (Testando hipóteses a partir de conjuntos de confiança)

Suponha que dispomos de dados $\mathbf{X} = \{X_1, X_2, \dots, X_n\}$ com f.d.p. comum $f(x \mid \theta)$ e que $w(\mathbf{X})$ é um conjunto de confiança para uma função de interesse $g(\theta)$. Então para todo valor g_0 assumido por $g(\theta)$ existe um teste δ_{g_0} , de nível α_0 que rejeita $H_0 : g(\theta) = g_0$ se e somente se $g(\theta_0) = g_0 \notin w(\mathbf{X})$.

Prova: Trivial. Ver DeGroot, Teorema 9.1.2.

Exemplo

Vamos aplicar os conceitos discutidos ao caso Normal com variância conhecida.

Exemplo 21 (Teste para média da Normal com variância conhecida)

Suponha que $\mathbf{X} = \{X_1, X_2, \dots, X_n\}$ formam uma amostra aleatória de uma distribuição Normal com média μ e variância σ^2 , conhecida. Considere testar a hipótese

$$H_0 : \mu = \mu_0,$$

$$H_1 : \mu \neq \mu_0.$$

Seja $\alpha_0 = 1 - \gamma$. Lembre-se de que o teste de tamanho α_0 , δ_{μ_0} é rejeitar H_0 se $|\bar{X}_n - \mu_0| \geq c$, $c := \Phi^{-1}(1 - \alpha_0/2) \sigma \sqrt{n}$. Esta última desigualdade pode ser manipulada algebricamente para obter o intervalo de confiança exato

$$(A(\mathbf{X}), B(\mathbf{X})) = (\bar{X}_n - c, \bar{X}_n + c),$$

de modo que $\Pr(A(\mathbf{X}) < \mu_0 < B(\mathbf{X}) | \mu = \mu_0) = \gamma$.

Da mesma forma que intervalos de confiança podem ser uni- ou bilaterais. Considere testar a hipótese

$$H_0 : g(\theta) \geq g_0,$$

$$H_1 : g(\theta) < g_0.$$

Podemos testar esta hipótese a partir de um intervalo de confiança da forma $I_l = (A(\mathbf{X}), \infty)$: se $g(\theta) \notin I_l$ então rejeitamos H_0 .

Testes de razão de verossimilhanças

Considere testar

$$H_0 : \theta \in \Omega_0,$$

$$H_1 : \theta \in \Omega_1.$$

Em certas situações, podemos utilizar a função de verossimilhança para quantificar a evidência em favor de H_0 .

Definição 50 (Teste de razão de verossimilhanças)

A estatística

$$\Lambda(\mathbf{x}) = \frac{\sup_{\theta \in \Omega_0} f_n(\mathbf{x} \mid \theta)}{\sup_{\theta \in \Omega} f_n(\mathbf{x} \mid \theta)},$$

é chamada uma **estatística de razão de verossimilhanças**. Um teste de razão de verossimilhanças, δ_k é um teste que rejeita H_0 se $\Lambda(\mathbf{x}) \leq k$ para uma constante k .

Exemplo 22 (Teste de razão de verossimilhanças para uma hipótese simples)

Suponha que X_1, X_2, \dots, X_n são uma amostra aleatória de uma distribuição Bernoulli com parâmetro p . Assim, temos $Y = \sum_{i=1}^n X_i$ e $Y \sim \text{Binomial}(n, p)$. Considere testar a hipótese $H_0 : p = p_0$, $H_1 : p \neq p_0$. Depois de observarmos $Y = y$, a função de verossimilhança é

$$f(x | p) = \Pr(Y = y | p) = \binom{n}{y} p^y (1 - p)^{n-y}.$$

Como neste exemplo $\Omega_0 = \{p_0\}$ e $\Omega_1 = (0, 1) \setminus \{p_0\}$,

$$\Lambda(x) = \frac{p_0^y (1 - p_0)^{n-y}}{\sup_{p \in (0,1)} p^y (1 - p)^{n-y}}.$$

O supremo no denominador é atingido no EMV, $\hat{p} = y/n$, de modo que

$$\Lambda(x) = \left(\frac{np_0}{y} \right)^y \left(\frac{n(1 - p_0)}{n - y} \right)^{n-y}.$$

Para mais detalhes, ver código no repositório do curso.

Um teorema útil

Sob certas condições de regularidade, podemos fazer afirmações sobre a distribuição assintótica de $\log \Lambda(\mathbf{X})$.

Teorema 31 (Teorema de Wilks¹⁵)

Suponha que temos um espaço de parâmetros com k coordenadas, $\theta = (\theta_1, \theta_2, \dots, \theta_k)$ e desejamos testar a hipótese (simples) da forma

$$H_0 : \theta_j = \theta_0^j, j = 1, 2, \dots, k,$$

$$H_1 : \theta_j \neq \theta_0^j, j = 1, 2, \dots, k.$$

Então, sob condições de regularidade, temos que, à medida que $n \rightarrow \infty$,

$$-2 \log \Lambda(\mathbf{x}) \xrightarrow{d} \chi^2(k),$$




Prova: Avançada, não será dada aqui. Ver Teorema 9.1.4 de DeGroot. Para a demonstração, ver Teorema 7.125 de Schervish (1995).

¹⁵Em homenagem a Samuel Wilks (1906-1964), matemático estadunidense.

O que aprendemos?

- 💡 Intervalos de confiança podem ser utilizados para testar hipóteses;
- 💡 Testes podem ser bicaudais $(1 - \alpha_0/2)$ quando unicaudais $((1 + \alpha_0)/2)$;
- 💡 Razões de verossimilhanças
 “A razão entre o supremo da função de verossimilhança tomado no espaço em que H_0 é verdadeira (Ω_0) e o mesmo supremo tomado sobre todo o espaço de parâmetros (Ω)”
- 💡 Teorema de Wilks;
 “À medida que o tamanho de amostra aumenta, menos duas vezes o logaritmo da razão de verossimilhanças tende em distribuição para uma Qui-quadrado com k graus de liberdade”

Leitura recomendada

-  DeGroot seção 9.1;
-  * Schervish (1995), capítulos 4.5.5 e 7.5 .
-  * Casella & Berger (2002), seção 8.2.
- ▶▶ Próxima aula: DeGroot, seção 9.5;

Teste t (de Student)

- Teste não-viesado;
- O teste t (unilateral e bilateral);
- Teste t pareado;
- Teste t para duas amostras;
 - ◊ Variâncias iguais (homogeneidade);
 - ◊ Variâncias proporcionais.
- Propriedades e exemplos;

Teste não-viesado

Em analogia com o conceito de estimador não-viesado, podemos também classificar testes de hipótese em viesados ou não-viesados.

Definição 51 (Teste não viesado)

Suponha que desejamos testar a hipótese

$$H_0 : \theta \in \Omega_0,$$

$$H_1 : \theta \in \Omega_1.$$

*através do teste δ . Dizemos que δ é **não-viesado** se (e somente se) para $\theta \in \Omega_0$ e $\theta' \in \Omega_1$, vale*

$$\pi(\theta \mid \delta) \leq \pi(\theta' \mid \delta),$$

ou seja, se a função poder é pelo menos tão grande no espaço onde H_0 é falsa (Ω_1) quanto no espaço em que H_0 é verdadeira (Ω_0).

Teste t: motivação

Suponha que estamos interessados em testar hipóteses sobre a média (μ) de uma distribuição Normal quando a variância (σ^2) é desconhecida. Sabemos que é possível encontrar uma quantidade pivotal $(\bar{X}_n - \mu)/\hat{\sigma}'$ tal que é possível construir um intervalo de confiança para μ da forma $\bar{X}_n - c\hat{\sigma}'/\sqrt{n}, \bar{X}_n + c\hat{\sigma}'/\sqrt{n}$, onde $c = T^{-1}(\gamma; n - 1)$ é a f.d.a. inversa de uma distribuição t de Student com $n - 1$ graus de liberdade.

Pergunta 5 (Como falar sobre μ quando ambas μ e σ^2 são desconhecidas?)

Suponha que Palmirinha esteja interessada em testar a hipótese de que a média da concentração de amido na sua pamonha seja maior que $\mu_0 = 7\text{mg/L}$, ou seja,

$$H_0 : \mu \geq \mu_0,$$

$$H_1 : \mu < \mu_0,$$

mas ela desconhece a variância do processo, σ^2 . Como testar hipóteses sobre $\theta = (\mu, \sigma^2)$?

Exemplo

Palmirinha pode começar computando a estatística

$$U := \sqrt{n} \frac{\bar{X}_n - \mu_0}{\hat{\sigma}'},$$

onde $\hat{\sigma}' = \sqrt{\sum_{i=1}^n (X_i - \bar{X}_n)^2 / (n-1)}$ como já estudado. A partir daí, pode construir um teste δ_c que rejeita H_0 se $U \leq c$, para uma constante real c definida.

Observação 25 (A distribuição de U sob H_0)

Quando H_0 é verdadeira, em particular, quando $\mu = \mu_0$, U tem distribuição t de Student com $n - 1$ graus de liberdade, não importando o valor de σ^2 . Isto é, U é pivotal.

Definição 52 (Teste t)

Um teste δ_c que rejeita H_0 se $U \geq c$ (equiv. $U \leq c$), com $c = T^{-1}(1 - \alpha_0; n - 1)$ é chamado um teste t (unicaudal) de tamanho α_0 .

Teorema 32 (Propriedades do teste t)

Suponha que δ_c rejeita H_0 se $U \geq c$. Então

- i) $\mu = \mu_0 \implies \pi(\mu, \sigma^2 \mid \delta_c) = \alpha_0$;
- ii) $\mu < \mu_0 \implies \pi(\mu, \sigma^2 \mid \delta_c) < \alpha_0$;
- iii) $\mu > \mu_0 \implies \pi(\mu, \sigma^2 \mid \delta_c) > \alpha_0$;
- iv) $\lim_{\mu \rightarrow -\infty} \pi(\mu, \sigma^2 \mid \delta_c) = 0$;
- v) $\lim_{\mu \rightarrow \infty} \pi(\mu, \sigma^2 \mid \delta_c) = 1$;
- vi) δ_c é não-viesado e tem tamanho α_0 .

Prova: Ver Teorema 9.5.1. de DeGroot.

P-valor para o teste t

Lembre-se de que o p-valor é a probabilidade, **sob** H_0 , de observarmos uma estatística tão ou mais extrema do que a que foi observada.

Teorema 33 (P-valor para um teste t unicaudal)

Suponha que observarmos $U = u$ e seja $T(\cdot; n - 1)$ a f.d.a. de uma distribuição t de Student com $n - 1$ graus de liberdade. Para a hipótese

$$H_0 : \mu \geq \mu_0,$$

$$H_1 : \mu < \mu_0,$$

o p-valor vale $T(u; n - 1)$, enquanto para a hipótese

$$H_0 : \mu \leq \mu_0,$$

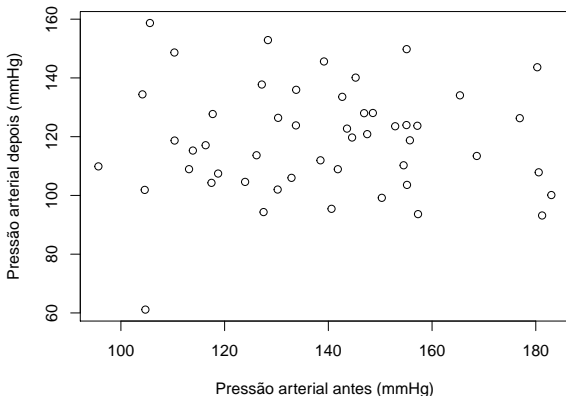
$$H_1 : \mu > \mu_0,$$

o p-valor vale $1 - T(u; n - 1)$.

Prova: Notar que δ_c depende de $c = T^{-1}(1 - \alpha_0; n - 1)$. Ver Teorema 9.5.2 em DeGroot.

Teste t pareado: motivação

Suponha que estamos interessados em medir o efeito de uma droga sobre a pressão arterial sistólica de um grupo de pacientes. Suponha que medimos as pressões arteriais de n pacientes antes (X) e depois (Y) de administrar a droga. Vamos supor que $X_i \sim \text{Normal}(\mu_{\text{antes}}, \sigma^2)$ e $Y_i \sim \text{Normal}(\mu_{\text{depois}}, \sigma^2)$.



Teste t pareado: execução

Estamos, portanto, interessados na hipótese¹⁶

$$H_0 : \mu_{\text{antes}} \leq \mu_{\text{depois}},$$

$$H_1 : \mu_{\text{antes}} > \mu_{\text{depois}}.$$

Podemos modelar a variável $Z_i = X_i - Y_i$ e sabemos que $Z_i \sim \text{Normal}(\mu_Z = \mu_{\text{antes}} - \mu_{\text{depois}}, 2\sigma^2)$. Desta forma, estamos interessados em testar hipóteses sobre μ_Z a partir de \mathbf{Z} . Em particular, a hipótese acima se traduz em

$$H_0 : \mu_Z \leq 0,$$

$$H_1 : \mu_Z > 0,$$

uma hipótese que podemos testar utilizando um teste t unicaudal como já discutido.

¹⁶Note que, neste caso, esta é a hipótese razoável a ser testada, porque estamos interessados apenas em rejeitar a hipótese de que a droga **não aumenta** a pressão arterial dos pacientes. Drogas que não tem efeito ou causam aumento da pressão não costumam ser aprovadas pela ANVISA.

Teste t para duas amostras

Considere agora a situação em que dispomos de dois conjuntos de dados, $\mathbf{X} = \{X_1, X_2, \dots, X_m\}$ e $\mathbf{Y} = \{Y_1, Y_2, \dots, Y_n\}$ e queremos estudar diferenças nas médias. Novamente, vamos modelar os processos como distribuições normais: $X_i \sim \text{Normal}(\mu_1, \sigma_1^2)$, $i = 1, 2, \dots, m$ e $Y_j \sim \text{Normal}(\mu_2, \sigma_2^2)$, $j = 1, 2, \dots, n$. Sob a premissa de homogeneidade $\sigma_1^2 = \sigma_2^2 = \sigma^2$, podemos testar a hipótese

$$H_0 : \mu_1 \leq \mu_2,$$

$$H_1 : \mu_1 > \mu_2,$$

computando a estatística

$$U = \frac{\sqrt{m+n-2}(\bar{X}_m - \bar{Y}_n)}{\sqrt{(\frac{1}{m} + \frac{1}{n})(S_X^2 + S_Y^2)}},$$

onde $\bar{X}_m = (1/m) \sum_{i=1}^m X_i$, $\bar{Y}_n = (1/n) \sum_{j=1}^n Y_j$, $S_X^2 = \sum_{i=1}^m (X_i - \bar{X}_m)^2$ e $S_Y^2 = \sum_{j=1}^n (Y_j - \bar{Y}_n)^2$. O teste procede analogamente ao que já foi discutido.

Relaxando a premissa de homogeneidade

Até aqui assumimos variâncias iguais, tanto no teste pareado quanto no teste para duas amostras. Podemos relaxar a premissa de igualdade das variâncias um pouco se assumirmos que $\sigma_2^2 = k\sigma_1^2$, isto é, que a razão entre as variâncias é conhecida. Neste caso, a estatística teste vale

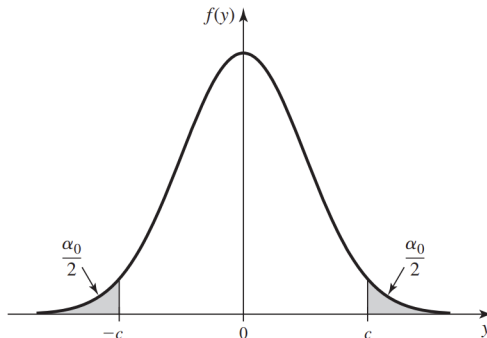
$$U_k = \frac{\sqrt{m+n-2}(\bar{X}_m - \bar{Y}_n)}{\sqrt{\left(\frac{1}{m} + \frac{k}{n}\right) \left(S_X^2 + \frac{S_Y^2}{k}\right)}}.$$

Quando as variâncias são diferentes e desconhecidas e não conhecemos k , temos o problema de Behrens-Fisher¹⁷ que é muito mais difícil de tratar.

¹⁷Em homenagem ao químico alemão Walter-Ulrich Behrens (1902–1962) e ao biólogo e estatístico britânico Ronald Aylmer Fisher (1890-1962).

O teste t bicaudal (bilateral)

No caso do teste t pareado, podemos estar interessados apenas em testar $\mu_{\text{antes}} = \mu_{\text{depois}}$, o que levaria a uma hipótese alternativa composta e um teste bicaudal (bilateral). Situação parecida acontece no caso de duas amostras quando queremos testar $\mu_1 = \mu_2$. Nesses casos, podemos facilmente adaptar os testes discutidos para acomodar a hipótese bilateral. Em ambos os casos, podemos fazer o teste “rejeite H_0 se $|U| \geq T^{-1}(1 - \alpha_0/2; n - 1)$ ”, e este terá tamanho α_0 .



O Teste t como um TRV (LRT)

Podemos também entender o teste t como um teste de razão de verossimilhanças. Em particular, temos para um teste t unicaudal,

$$\begin{aligned}\Lambda(\mathbf{x}) &= \frac{\sup_{(\mu, \sigma^2): \mu > \mu_0} f_n(\mathbf{x} \mid \mu, \sigma^2)}{\sup_{(\mu, \sigma^2)} f_n(\mathbf{x} \mid \mu, \sigma^2)}, \\ &= \begin{cases} \left(\frac{\hat{\sigma}^2}{\sigma_0^2} \right)^{n/2} & \text{se } \bar{x}_n > \mu_0, \\ 1, & \text{caso contrário,} \end{cases}\end{aligned}$$

onde $\hat{\sigma}^2$ é o EMV da variância e $\hat{\sigma}_0^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu_0)^2$. Onde o teste t tradicional rejeita H_0 se $U \geq c$, sua formulação TRV rejeita H_0 se $\Lambda(\mathbf{x}) \leq k$. A relação entre c e k é




$$c = \sqrt{\left[\left(\frac{1}{k^2} \right)^{1/n} - 1 \right] (n-1)},$$

o que estabelece que o teste t é um teste de razão de verossimilhanças.

O que aprendemos?

- 💡 O teste t permite comparar a média de um conjunto de dados com um valor postulado μ_0 ;
- 💡 Permite também comparar as médias de duas amostras, pareadas ou independentes;
- O teste t é não-viesado e pode ser escrito como um teste de razão de verossimilhanças;

Leitura recomendada

-  DeGroot seções 9.5 e 9.6;
-  * Casella & Berger (2002), seção 8.
- ▶▶ Próxima aula: DeGroot, seção 9.7;
- **Exercícios recomendados**
 -  DeGroot Seção 9.5: exercício 8.

Testes para igualdade de variâncias

- A distribuição F;
- Comparação de variâncias de duas normais;
- Propriedades;
- P-valor;

A distribuição F

Sejam $Y \sim \text{Qui-quadrado}(m)$ e $W \sim \text{Qui-quadrado}(n)$. Então

$$X = \frac{Y/m}{W/n},$$

tem distribuição F com m e n graus de liberdade, com f.d.p.

$$f_X(x) = \frac{\Gamma\left(\frac{m+n}{2}\right) m^{m/2} n^{n/2}}{\Gamma\left(\frac{n}{2}\right) \Gamma\left(\frac{m}{2}\right)} \cdot \frac{x^{m/2} - 1}{(mx + n)^{(m+n)/2}}, \quad x > 0.$$

Teorema 34 (Propriedades da distribuição F)

- i) Se $X \sim F(m, n)$, então $\frac{1}{X} \sim F(n, m)$;
- ii) Se $Y \sim \text{Student}(n)$, então $Y^2 \sim F(1, n)$.

Prova: Transformação de v.a.s padrão. Exercício para a leitora.

Testando a igualdade de duas variâncias

Suponha $X_i \sim \text{Normal}(\mu_1, \sigma_1^2), i = 1, 2, \dots, m$ e
 $Y_j \sim \text{Normal}(\mu_2, \sigma_2^2), j = 1, 2, \dots, n$. Estamos interessados em testar

$$H_0 : \sigma_1^2 \leq \sigma_2^2,$$

$$H_1 : \sigma_1^2 > \sigma_2^2.$$

Para isso, vamos computar a estatística de teste

$$V = \frac{S_X^2/(m-1)}{S_Y^2/(n-1)},$$

onde $S_X^2 = \sum_{i=1}^m (X_i - \bar{X}_m)^2$ e $S_Y^2 = \sum_{j=1}^n (Y_j - \bar{Y}_n)^2$.

Definição 53 (O teste F)

O teste F de homogeneidade (igualdade de variâncias) é o teste δ_c que rejeita H_0 se $V \geq c$, para uma constante positiva c .

Em primeiro lugar, podemos fazer afirmações sobre a distribuição de (uma transformação de) V .

Teorema 35 (A distribuição de V)

Seja $V = \frac{S_X^2/(m-1)}{S_Y^2/(n-1)}$, então:

$$\frac{\sigma_2^2}{\sigma_1^2} V \sim F(m-1, n-1).$$

Além disso, se $\sigma_1^2 = \sigma_2^2$, $V \sim F(m-1, n-1)$.

Prova: Notar que S_X^2/σ_1^2 e S_Y^2/σ_2^2 tem distribuição qui-quadrado com $m-1$ e $n-1$ graus de liberdade, respectivamente. Ver Teorema 9.7.3 de DeGroot.

Seja $G(x; m-1, n-1)$ a f.d.a. de uma distribuição F com $m-1$ e $n-1$ graus de liberdade. Da mesma forma, defina $G^{-1}(p; m-1, n-1)$ como a f.d.a. inversa.

Então, se $V = v$:

- Para a hipótese $H_0 : \sigma_1^2 \leq \sigma_2^2$, o p-valor vale $p = 1 - G(v; m-1, n-1)$;
- Para a hipótese $H_0 : \sigma_1^2 \geq \sigma_2^2$, o p-valor vale $p = G(v; m-1, n-1)$;
- Para a hipótese bicaudal $H_0 : \sigma_1^2 \neq \sigma_2^2$, o p-valor vale $p = 2 \min \{1 - G(v; m-1, n-1), G(v; m-1, n-1)\}$;

Mais propriedades do teste F

Analogamente ao teste t, podemos enunciar o seguinte teorema sobre o teste F.

Teorema 36 (Propriedades do teste F)

Suponha que estamos testando $H_0 : \sigma_1^2 \leq \sigma_2^2$. Então

- i) $\pi(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2 \mid \delta_c) = 1 - G\left(\frac{\sigma_2^2}{\sigma_1^2}c; m-1, n-1\right)$;*
- ii) $\sigma_1^2 = \sigma_2^2 \implies \pi(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2 \mid \delta_c) = \alpha_0$;*
- iii) $\sigma_1^2 < \sigma_2^2 \implies \pi(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2 \mid \delta_c) < \alpha_0$*
- iv) $\sigma_1^2 > \sigma_2^2 \implies \pi(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2 \mid \delta_c) > \alpha_0$;*
- v) $\lim_{\sigma_1^2/\sigma_2^2 \rightarrow 0} \pi(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2 \mid \delta_c) = 0$;*
- vi) $\lim_{\sigma_1^2/\sigma_2^2 \rightarrow \infty} \pi(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2 \mid \delta_c) = 1$;*
- vii) δ_c é não-viesado e tem tamanho α_0 .*

Prova: Omitida aqui. Ver Teorema 9.7.4 de DeGroot.

O que aprendemos?


- 💡 A distribuição F aparece quando tomamos a razão de variáveis aleatórias Qui-quadrado;
- 💡 Para comparação das variâncias de duas amostras a estatística teste tem distribuição F com $m - 1$ e $n - 1$ graus de liberdade;
- O teste F , como seu primo o teste t , é não viesado e tem tamanho α_0 .

 DeGroot seção 9.7;

 * Casella & Berger (2002), seção 8.

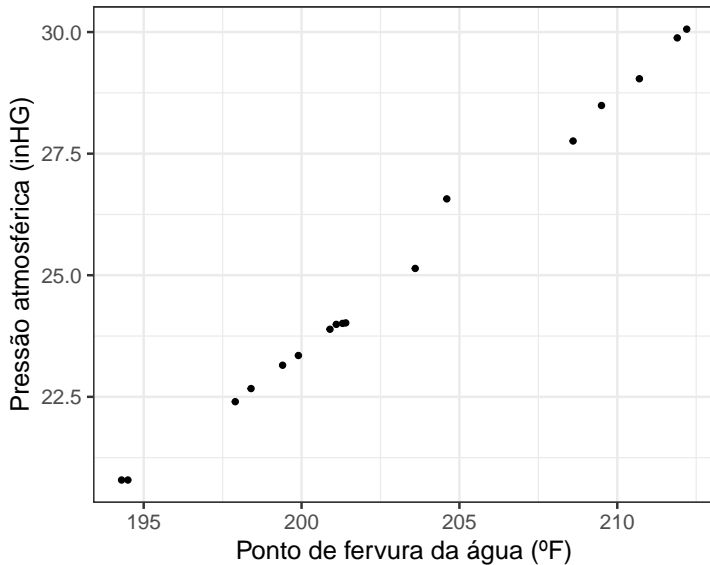
▶▶ Próxima aula: DeGroot, seção 11;

- **Exercícios recomendados**

-  Derivar a função de densidade de probabilidade de uma distribuição F (Teorema 9.7.1 de DeGroot).

-  Derivar o teste F como um teste de razão de verossimilhanças.

- Mínimos quadrados;
- Modelo de regressão simples (univariado);
 - ◇ Formulação;
 - ◇ Premissas.
- Distribuição amostral dos estimadores;
- Intervalos de confiança para os coeficientes;
- Testes para os coeficientes;
- Predição: pontual e intervalar.



Mínimos quadrados

Suponha que estamos interessados na reta

$$y_i = \beta_0 + \beta_1 x_i. \quad (32)$$

- β_0 é chamado o **intercepto** (*intercept*) da reta;
- β_1 é chamado o **coeficiente angular** (*slope*) da reta.

Teorema 37 (A linha de mínimos quadrados)

Sejam $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ uma coleção de n pontos. Os valores dos coeficientes que minimizam a soma de quadrados são

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}, \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \end{aligned}$$

onde $\bar{x} = (1/n) \sum_{i=1}^n x_i$ e $\bar{y} = (1/n) \sum_{i=1}^n y_i$.

Prova: Escrever a equação de estimação, $Q = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2$, diferenciar Q com respeito aos coeficientes e igualar a zero. Ver Teorema 11.1.1 em DeGroot.

O modelo linear

Podemos construir um modelo estatístico explícito para a relação entre as variáveis¹⁸ \mathbf{X} e Y :

$$E[Y \mid \mathbf{X} = x_1, x_2, \dots, x_P] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_P x_P. \quad (33)$$

Terminologia:

- Y é chamada de desfecho, **variável-resposta** ou variável dependente;
- \mathbf{X} são chamados covariáveis, **preditores** ou, ainda, variáveis independentes;
- $\beta = \{\beta_0, \beta_1, \dots, \beta_P\}$ são os **coeficientes de regressão**.

Podemos então idealizar o seguinte modelo

Ideia 6 (Modelo linear)

$$Y_i = \beta_0 + \sum_{j=1}^P \beta_j x_{ij} + \epsilon_i, \quad \epsilon_i \sim \text{Normal}(0, \sigma^2).$$

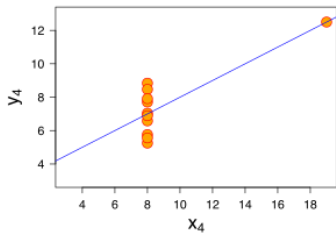
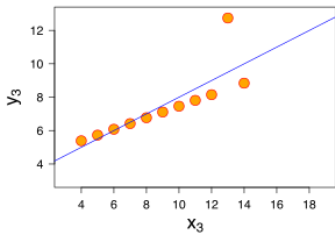
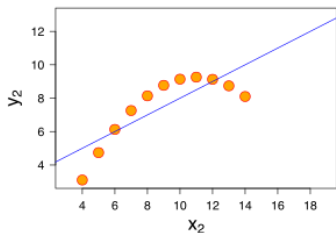
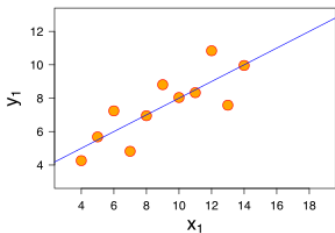
¹⁸Em notação de matrizes, $E[Y] = \mathbf{X}^T \beta$.

Premissas (importante!)

Como todo modelo, a regressão linear se apoia em premissas sobre os dados e o seu processo gerador.

- P1. O(s) preditor(es) é (são) conhecido(s);
- P2. Normalidade: dados os preditores \mathbf{X} , a resposta Y tem distribuição normal;
- P3. Linearidade na média: a esperança condicional de Y é dada por $\beta_0 + \sum_{j=1}^P \beta_j x_{ij}$;
- P4. Variância comum (**homocedasticidade**): a variância condicional de Y_i é σ^2 para todo $i = 1, 2, \dots, n$;
- P5. Independência (condicional): dados os valores de \mathbf{X} , os valores de Y são independentes entre si.

Cuidado! Quarteto de Anscombe¹⁹



¹⁹Em homenagem ao estatístico Britânico Francis Anscombe (1918-2001).

Um teorema interessante

No modelo linear, a solução de mínimos quadrados e a de máxima verossimilhança coincidem!

Teorema 38 (EMV para os coeficientes de uma regressão linear (simples))

Sob as premissas já listadas, os estimadores de máxima verossimilhança para $\theta = (\beta_0, \beta_1, \sigma^2)$ são

$$\begin{aligned}\hat{\beta}_{0EMV} &= \bar{y} - \hat{\beta}_{1EMV}\bar{x}, \\ \hat{\beta}_{1EMV} &= \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \\ \hat{\sigma}_{EMV}^2 &= \frac{1}{n} \sum_{i=1}^n \left(y_i - (\hat{\beta}_{0EMV} + \hat{\beta}_{1EMV}x_i) \right)^2,\end{aligned}$$

ou seja, os estimadores de máxima verossimilhança dos coeficientes minimizam a soma de quadrados da reta estimada.

Prova: Ver Teorema 11.2.1 de DeGroot.

Distribuição amostral dos estimadores

Sob as premissas já discutidas, podemos fazer afirmações sobre a distribuição amostral dos estimadores obtidos:

Teorema 39 (Distribuição amostral dos estimadores dos coeficientes)

$$\hat{\beta}_{0EMV} \sim \text{Normal} \left(\beta_0, \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{s_x^2} \right) \right),$$

$$\hat{\beta}_{1EMV} \sim \text{Normal} \left(\beta_1, \frac{\sigma^2}{s_x^2} \right),$$

$$\text{Cov} \left(\hat{\beta}_{0EMV}, \hat{\beta}_{1EMV} \right) = -\frac{\bar{x}\sigma^2}{s_x^2},$$

onde $s_x = \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}$.

Prova: Usar as leis de esperanças e variâncias. Ver Teorema 11.2.2 de DeGroot.

Intervalos de confiança para os coeficientes

Podemos computar intervalos de confiança para os coeficientes da regressão linear de maneira muito similar ao que já vimos para o caso da média da Normal.

Teorema 40 (Intervalos de confiança para os coeficientes de uma regressão linear)

$$\hat{\beta}_0 \pm \hat{\sigma}' c \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{s_x^2}} \quad \text{e} \quad \hat{\beta}_1 \pm c \frac{\hat{\sigma}'}{s_x},$$

$$\hat{\beta}_0 + \hat{\beta}_1 x_{pred} \pm c \hat{\sigma}' \sqrt{\frac{1}{n} + \frac{(x_{pred} - \bar{x})^2}{s_x^2}},$$

onde $c = T^{-1}(1 - \frac{\alpha}{2}; n - 2)$ e

$$\hat{\sigma}' := \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n - 2}}.$$

Prova: Usar o Teorema 11.3.5 de DeGroot e os valores apropriados de c_0 e c_1 .

Testes de hipóteses para o coeficiente angular

Em geral, estamos interessados em testar a hipótese

$$H_0 : \beta_1 = \beta^*,$$

$$H_1 : \beta_1 \neq \beta^*.$$

Para tanto, podemos computar a estatística

$$U_1 = s_x \frac{\hat{\beta}_1 - \beta^*}{\hat{\sigma}'},$$

e computar o p-valor como

$$\Pr(U_1 \geq |u_1|) + \Pr(U_1 \leq -|u_1|).$$

Notando que U_1 tem distribuição t de Student com $n - 2$ graus de liberdade sob H_0 , podemos computar o p-valor exatamente.

Resultados bem similares valem para testar hipóteses sobre β_0 ou \hat{Y} .

Predição pontual

Suponha que queremos prever o valor de Y para um certo x_{pred} que não foi observado no experimento. Podemos compor nossa predição (pontual) como

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_{\text{pred}}. \quad (34)$$

Teorema 41 (Erro quadrático médio da predição)

A predição como em (??) tem erro quadrático médio (EQM) igual a

$$E \left[\left(\hat{Y} - Y \right)^2 \right] = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_{\text{pred}} - \bar{x})^2}{s_x^2} \right).$$

Prova: Ver Teorema 11.2.3 de DeGroot.

Observação 26 (EQM fora da amostra)

O EQM aumenta quanto mais longe x_{pred} estiver dos valores de X que foram medidos (observados).

Muitas vezes estamos interessados em produzir um *intervalo* para a nossa predição, ao invés de um único valor (predição pontual). Nesta situação, podemos fazer uso do seguinte teorema:

Teorema 42 (Intervalos de **predição** para \hat{Y})

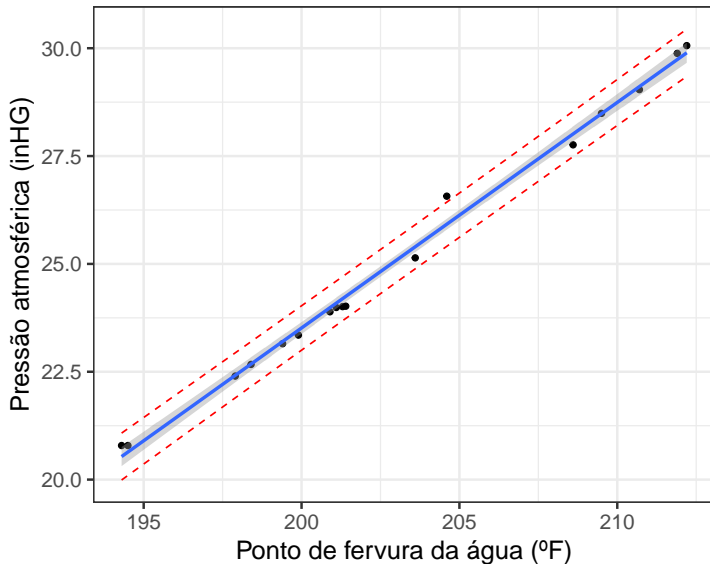
A probabilidade de $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_{pred}$ estar no intervalo

$$\hat{Y} \pm T^{-1} \left(1 - \frac{\alpha_0}{2}; n - 2 \right) \hat{\sigma}' \sqrt{\left[1 + \frac{1}{n} + \frac{(x_{pred} - \bar{x})^2}{s_x^2} \right]},$$

é $1 - \alpha_0$.

Prova: Ver Teorema 11.3.6 de DeGroot.

Intervalos de confiança e de predição: ilustração



O que aprendemos?

- 💡 O modelo linear permite modelar a relação (linear) entre uma (ou mais) variável(is) independente(s) e uma variável dependente;
- 💡 A estimação dos coeficientes pode ser feita por mínimos quadrados;
- 💡 A solução de mínimos quadrados é também a solução de máxima verossimilhança!
- 💡 Podemos aplicar a teoria Normal para testar hipóteses sobre os coeficientes e calcular intervalos de confiança;
- 💡 Podemos produzir previsões sobre a variável dependente para valores não-observados da(s) variável(is) independente(s).


Leitura recomendada


 DeGroot seções 11.1, 11.2 e 11.3;

 * Casella & Berger (2002), seção 11.3.

▶▶ Próxima aula: DeGroot, seção 9.9;

• Exercícios recomendados

 DeGroot, seção 11.1: exercício 3.

 DeGroot, seção 11.2: exercícios 2, 3 e 6.

 * Bônus: DeGroot, seção 11.2: exercício 19 (valendo 0.5 na média).

Testes de hipótese: discussão

- Como construir um teste que **quase sempre** rejeita H_0 ;
- Significância estatística vs significância prática;
- Rapidinhas.

Um teste esquisito

Suponha que temos X_1, X_2, \dots, X_n vindos de uma distribuição Normal com média θ e variância 1 e queremos testar as hipóteses

$$H_0 : \theta = 0,$$

$$H_1 : \theta = 1.$$

Seguindo o exemplo 9.2.5 de DeGroot, podemos escrever

$$\eta(\mathbf{x}) = \frac{f_1(\mathbf{x})}{f_0(\mathbf{x})},$$

e compor um teste que rejeita H_0 quando $\eta(\mathbf{x}) > c$. Isto é equivalente a construir um teste de tamanho α_0 , de modo que valha

$$\Pr(\bar{X}_n \geq c' \mid \theta = 0) = \alpha_0,$$

o que nos leva a concluir que $c' = \frac{1}{2} + \frac{\log(c)}{n}$ e que $c = \Phi^{-1}(1 - \alpha_0)/\sqrt{n}$.

Qual o problema?

Primeiro, vamos lembrar que, para um teste δ ,

$$\alpha(\delta) := \Pr(\text{Rejeitar } H_0 \mid \theta = 0),$$

$$\beta(\delta) := \Pr(\text{Não rejeitar } H_0 \mid \theta = 1).$$

O problema aqui é que para este teste temos

n	$\alpha(\delta)$	$\beta(\delta)$	c
1	0.05	0.74	0.72
25	0.05	3.97×10^{-4}	2.3×10^{-4}
100	0.05	8×10^{-15}	2.7×10^{-15}

Ou seja, quando temos $n = 100$ observações, os dados podem ser trilhões de vezes mais prováveis sob H_0 e ainda assim vamos rejeitar a hipótese nula.

Podemos pensar em duas soluções (complementares) para o problema posto.

Ideia 7 (Ajustando o nível de significância com o tamanho da amostra)

Em várias situações, por exemplo como a mostrada acima, faz sentido ajustar (diminuir) o nível de confiança do teste com o tamanho da amostra de modo a balancear os erros do tipo I e II.

Ideia 8 (Minimizar uma combinação linear das probabilidades de erro)

Poderíamos balancear os erros ao minimizar

$$a\alpha(\delta) + b\beta(\delta).$$

Lehmann (1958)²⁰ propôs a restrição $\beta(\delta) = c\alpha(\delta)$, que tem a vantagem de forçar que ambos os tipos de erro diminuam à medida que obtemos mais dados.

Ver seções 9.2 e 9.8 de DeGroot.

²⁰Lehmann, Erich L. "Significance level and power." The Annals of Mathematical Statistics (1958): 1167-1176.

Suponha que eu estou testando uma nova droga, e o parâmetro θ mede o efeito da droga. Em geral, estamos interessados em testar a hipótese

$$H_0 : \theta \leq 0,$$

$$H_1 : \theta \geq 0.$$

Quando o tamanho de amostra é muito grande, seremos capazes de detectar, com alta probabilidade (poder) se $\theta = 0.000003$ ou $\theta = 0$.

Acontece que uma droga com $\theta = 0.000003$ não oferece nenhuma vantagem prática. Portanto, ao se realizar um teste de hipótese e rejeitar H_0 , não podemos concluir que “a droga funciona”, pelo menos não num sentido médico.

Ideia 9 (Significância estatística não implica relevância prática)

Responda rápido

- a) O que é a função poder de um teste de hipótese e o que esperamos observar em um teste não-enviesado?
- b) Se testarmos uma hipótese um número suficiente de vezes ela eventualmente será rejeitada. Explique esta afirmação e suas consequências.
- c) O que é o p-valor de um teste?
- d) É correto afirmar que uma hipótese nula é falsa se ela for rejeitada? É correto afirmar que uma hipótese alternativa é verdadeira se a nula for rejeitada? Justifique.
- e) Um intervalo de confiança nível de 95% para θ é calculado a partir de n observações. É correto afirmar que o parâmetro verdadeiro θ_0 está dentro deste intervalo com probabilidade 95%? Justifique.
- f) Explique como podemos obter um conjunto de confiança a partir de um teste de hipótese.

O que aprendemos?

- 💡 Rejeição eventual; “Se coletarmos uma quantidade suficiente de dados, podemos rejeitar qualquer hipótese nula”
- 💡 Significância estatística \neq significância prática/científica!

Leitura recomendada

 DeGroot seções 9.2, 9.3 e 9.9;

- **Exercícios recomendados**

- DeGroot, seção 9.9: exercícios 2 e 3.