Segunda avaliação (A2)

Disciplina: Inferência Estatística
Instrutor: Luiz Max Carvalho
Monitores: Jairon Nóia & Tiago Silva

26 de Novembro de 2022

- O tempo para realização da prova é de 3 horas;
- Leia a prova toda com calma antes de começar a responder;
- Responda todas as questões sucintamente;
- Marque a resposta final claramente com um quadrado, círculo ou figura geométrica de sua preferência;
- A prova vale 80 pontos. A pontuação restante é contada como bônus;
- Apenas tente resolver a questão bônus quando tiver resolvido todo o resto;
- Você tem direito a trazer **uma folha de "cola"** tamanho A4 frente e verso, que deverá ser entregue junto com as respostas da prova.

1. O estatístico e o poeta.

Eu te vejo sumir por aí
Te avisei que a cidade era um vão
Dá tua mão, olha pra mim
Não faz assim, não vai lá, não
Os letreiros a te colorir
Embaraçam a minha visão
Eu te vi suspirar de aflição
E sair da sessão frouxa de rir
Já te vejo brincando gostando de ser
Tua sombra a se multiplicar
Nos teus olhos também posso ver
As vitrines te vendo passar
Na galeria, cada clarão
É como um dia depois de outro dia
Abrindo um salão
Passas em exposição
Passas sem ver teu vigia
Catando a poesia
Que entornas no chão

As Vitrines (Almanaque, 1981) de Chico Buarque (1944-).

O eu-lírico da canção, que vamos chamar aqui de Ivo, pensa em seu amado, Adão. Adão é poeta, e tem a estranha mania de deixar cair seus poemas ao passear pelo shopping. Ivo, muito solícito e perdidamente apaixonado, corre atrás do companheiro catando os papéis que o desastrado deixa cair. Sendo estatístico, Ivo sabe que pode modelar o tempo entre a queda dos poemas como uma variável aleatória exponencial com taxa θ. Ivo quer saber se será capaz de acompanhar Adão na sua jornada sem perder nenhum poema. Para isso, julga que se $\theta \leq \theta_0$, ele será capaz de catar toda a poesia deixada por Adão antes de ser carregada pelo vento.

Suponha que Ivo observa o processo de queda de n poemas e anota o tempo entre cada queda, formando a amostra $Y_1, Y_2, \ldots, Y_n$. Ivo considera a estatística de teste $S = \sum_{i=1}^{n} Y_i$ e constrói o teste δ_c de modo que, se $S \geq c$, ele rejeita a hipótese $H_0 : \theta \leq \theta_0$.

a) (10 pontos) Encontre a função poder do teste de Ivo.

b) (10 pontos) Mostre que a função poder do item anterior é **não-decrescente** em θ;

   Dica: Se X tem distribuição Gama com parâmetros $k \in \mathbb{N}$ e θ, então

$$P_\theta\left(X \leq x\right) = e^{-x/\theta} \sum_{j=k}^{\infty} \frac{1}{j!} \left(\frac{x}{\theta}\right)^j.$$

c) (10 pontos) Encontre uma expressão para o tamanho α_0 do teste δ_c;

d) (10 pontos) O teste em questão é não-viesado? Justifique;

Conceitos trabalhados: função poder; tamanho. **Nível de dificuldade**: fácil.

Resolução: Para responder a), vamos lembrar que a função poder $\pi(\theta \mid \delta_c) = P_\theta\,(\text{Rejeitar } H_0)$. Sendo assim, temos

$$\begin{aligned}\pi(\theta \mid \delta_c) &= P_\theta\,(S \geq c)\,,\\ &= 1 - P_\theta(S < c),\\ &= 1 - F_S\,(c; n, \theta)\,,\end{aligned}$$

onde $F_S\,(x; a, b)$ é a f.d.a. de uma distribuição Gama com forma a e taxa b avaliada em $x \in \mathbb{R}$. Agora precisamos mostrar que $\pi(\theta \mid \delta_c)$ é não descrescente em θ de modo a responder b). Usando a dica, sabemos que

$$\pi(\theta \mid \delta_c) = 1 - e^{-c/\theta} \sum_{j=k}^{\infty} \frac{1}{j!} \left(\frac{c}{\theta}\right)^j,$$

de modo que $\frac{\partial}{\partial \theta}\pi(\theta \mid \delta_c) \geq 0$. Outro bom argumento é esboçar o gráfico da função poder e mostrar que ela não pode decrescer. O tamanho de δ_c é dado por

$$\alpha_0 := \sup_{\theta \in \Theta_0} \pi(\theta \mid \delta_c).$$

Como a função poder é não descrescente, temos que $\alpha_0 = \pi(\theta_0 \mid \delta_c)$, respondendo c). Em d), temos que o teste de fato é não-viesado, pois a função poder é não descrescente em θ, de modo que para todo par $\theta \in \Theta \setminus \Theta_0$ e $\theta' \in \Theta_0$ temos que $\pi(\theta' \mid \theta) \leq \pi(\theta \mid \theta)$. ■

Comentário: Esta é uma questão parecida com a Q1 da A2 de 2020, mas neste caso Ivo mede os tempos entre as quedas dos poemas. Uma questão simples e conceitual para esquentar os músculos.

2. PO-KÉ-MON!

Suponha que a Liga Internacional de Pokemon (LIP) tenha um sistema de *pokescores* que podem assumir qualquer valor real. Quanto maior o *pokescore* de uma jogadora, mais alto no ranking mundial ela está. A liga se organiza em times de n jogadores.

Para entrar na liga, um time precisa ter um *pokescore* médio superior a θ_0, isto é, a média dos pokescores de seus jogadores precisa ser maior que θ_0. Suponha que os *pokescores* dentro de um time são distribuídos de acordo com uma distribuição Normal com média θ e variância σ^2, conhecida. Queremos desenvolver um método para incluir times num torneio automaticamente, baseado nos *pokescores* dos seus integrantes.

a) (5 pontos) Encontre uma quantidade pivotal para θ;

b) (5 pontos) Utilizando a quantidade do item anterior, construa um intervalo de confiança de 95% para θ;

c) (10 pontos) A partir do intervalo encontrado, é possível testar $H_0 : \theta \leq \theta_0$? Como?

d) (10 pontos) Se σ^2 fosse desconhecida, como você modificaria o teste do item anterior?

e) (5 pontos) Se aplicarmos os testes em (c) e (d) para selecionar times automaticamente, seremos injustos com alguns times, isto é, vamos deixar de incluir times que de fato se encaixam na condição de seleção. Com que probabilidade isso acontece?

f) (5 pontos) Se quisermos diminuir a probabilidade do item anterior, o que podemos fazer? Que consequências isso tem?

Conceitos trabalhados: quantidade pivotal; intervalo de confiança; equivalência entre ICs e testes. **Nível de dificuldade**: fácil.
Resolução: Existem várias respostas possíveis para a), algumas mais úteis (para os itens subsequentes) que outras. Por exemplo,

$$W_n := \bar{X}_n - \theta$$

é pivotal, com distribuição Normal com média 0 e variância σ^2/n. Uma escolha um pouco mais sábia é

$$Z_n := \sqrt{n}\frac{(\bar{X}_n - \theta)}{\sigma},$$

que tem distribuição normal-padrão. Para responder b), temos, mais uma vez, algumas opções: podemos construir intervalos unilaterais ou bilaterais. A partir de Z_n, podemos construir um intervalo de confiança conseguimos construir intervalos usando a normal-padrão. Para um intervalo unilateral, podemos escolher $c_U = \Phi^{-1}(0.05)$ e fazer

$$I_1(\boldsymbol{X}_n) = \left(-\infty, \bar{X}_n + |c_U|\frac{\sigma}{\sqrt{n}}\right),$$

ou

$$I_2(\boldsymbol{X}_n) = \left(\bar{X}_n - |c_U|\frac{\sigma}{\sqrt{n}}, \infty\right).$$

Para construir um intervalo bilateral, fazemos $c_B = \Phi^{-1}(0.025)$ e então

$$I_3(\boldsymbol{X}_n) = \left(\bar{X}_n - |c_B|\frac{\sigma}{\sqrt{n}}, \bar{X}_n + |c_B|\frac{\sigma}{\sqrt{n}}\right),$$

é um intervalo com a cobertura desejada. A resposta de c) é sim: podemos, por exemplo, usar $I_2(\boldsymbol{X}_n)$ e desenhar um teste da forma

$$\delta_2 = \begin{cases} \text{Rejeitar } H_0, \text{ se } \theta_0 \in I_2(\boldsymbol{X}_n), \\ \text{Falhar em rejeitar } H_0 \text{ caso contrário.} \end{cases}$$

Este teste tem tamanho α e é não-viesado. Se não soubéssemos o valor de σ^2, poderíamos construir a quantidade pivotal

$$Q_n = \sqrt{n}\frac{\bar{X}_n - \theta_0}{\sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{n-1}}},$$

que tem distribuição t de Student com $n-1$ graus de liberdade. Isso nos leva a um novo intervalo da forma

$$I_4(\boldsymbol{X}_n) = \left(\bar{X}_n - |t_U| \frac{\sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}{n-1}}}{\sqrt{n}}, \infty \right),$$

onde t_U é o quantil α de uma t de Student com $n-1$ graus liberdade. Com I_4 em mãos, desenhamos um teste como anteriormente:

$$\delta_4 = \begin{cases} \text{Rejeitar } H_0, \text{ se } \theta_0 \in I_4(\boldsymbol{X}_n), \\ \text{Falhar em rejeitar } H_0 \text{ caso contrário.} \end{cases}$$

A resposta de e) tem a ver com aceitar H_0 quando ela é falsa, isto é, quando $\theta > \theta_0$. Este é um erro do tipo II e acontece com probabilidade $1 - \pi(\theta \mid \delta_4) = 0.975$. No mesmo ímpeto, poderiámos responder f) dizendo que é possível construir testes onde o erro do tipo II fica controlado. A consequência é, em geral, que a taxa de erro do tipo I (falsos positivos) tende a aumentar. ■

Comentário: Esta questão é bem conceitual e procura testar os conhecimentos sobre testes no caso normal. Havia várias maneiras de responder corretamente às questões.

3. Run, Joey, run![1]

O modelo linear (de regressão) é um dos cavalos de batalha da Estatística, sendo aplicado em problemas de Finanças, Medicina e Engenharia. Vamos agora estudar como utilizar as propriedades deste modelo para desenhar experimentos com garantias matemáticas de desempenho e obter estimadores de quantidades de interesse.

a) (10 pontos) Uma prática comum em regressão é a de **centrar** a variável independente (covariável), isto é subtrair a média; isto facilita a interpretação do intercepto e também simplifica alguns cálculos importantes. Mostre que no caso com a covariável centrada, $\hat{\beta}_0$ e $\hat{\beta}_1$ são independentes;

b) (10 pontos) Mais uma vez considerando o caso centrado, mostre como obter o número de observações n que faz com que a variância do estimador de máxima verossimilhança do intercepto seja menor que $v > 0$;

c) (10 pontos) Mostre como obter um estimador não-viesado da quantidade $\theta = a\beta_0 + b\beta_1 + c$, com $a, b, c \neq 0$, e encontre o seu erro quadrático médio.

d) (10 pontos) Quando $x_{\text{pred}} = \bar{x}$, mostre como obter o número de observações n necessário para que o intervalo de predição de $100(1-\alpha_0)\%$ para a variável-resposta (Y) tenha largura menor ou igual a $l > 0$ com probabilidade pelo menos γ.

   Dicas:(i) A expressão dependerá *também* da variância dos resíduos, σ^2 e (ii) Você não precisa calcular n, apenas mostrar o procedimento para obtê-lo.

[1] Linear regression is a war horse of Statistics. The horse in 'War Horse' (2011) is named Joey.

Conceitos trabalhados: Regressão linear; desenho experimental; quantidades derivadas.
Nível de dificuldade: médio.
Resolução: Para resolver a) vamos perceber que quando substituímos a covariável original X por $X' = X - \bar{x}$ temos $\bar{x}' = 0$ e portanto $\operatorname{Cov}\left(\hat{\beta}_0, \hat{\beta}_1\right) = -\frac{\bar{x}'\sigma^2}{s_x^2} = 0$. Para afirmarmos que $\hat{\beta}_0$ e $\hat{\beta}_1$ são independentes é preciso lembrar que estes estimadores têm distribuição conjunta Normal bivariada; quando a covariância é zero, sabemos que são independentes. A resposta de b) pode ser deduzida ao lembrar que no caso centrado, a variância de $\hat{\beta}_0$ é σ^2/n. Desta forma, precisamos apenas encontrar n tal que $\sigma^2/n < v$, isto é $n > \sigma^2/v$. Como sabemos que os estimadores dos coeficientes são não-viesados, podemos encontrar $\hat{\theta} = a\hat{\beta}_0 + b\hat{\beta}_1 + c$ como nosso estimador não-viesado de θ. O EQM de tal estimador é a sua variância:

$$
\begin{aligned}
E[(\hat{\theta} - \theta)^2] = \operatorname{Var}(\hat{\theta}) &= a^2 \operatorname{Var}(\hat{\beta}_0) + b^2 \operatorname{Var}(\hat{\beta}_1) - ab \operatorname{Cov}(\hat{\beta}_0, \hat{\beta}_1), \\
&= a^2\sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{s_x^2}\right) + b^2\frac{\sigma^2}{s_x^2} + ab\frac{\bar{x}\sigma^2}{s_x^2}, \\
&= \sigma^2\left(\frac{a^2}{n} + \frac{a^2\bar{x}^2}{s_x^2} + \frac{b^2}{s_x^2} + \frac{ab\bar{x}}{s_x^2}\right).
\end{aligned}
$$

Por fim, vamos responder d). Note que a expressão necessária aqui é a do intervalo de predição:

$$
\hat{Y} \pm c(n, \alpha_0) \cdot \hat{\sigma}_r' \cdot \sqrt{\left[1 + \frac{1}{n} + \frac{(x_{\text{pred}} - \bar{x})^2}{s_x^2}\right]},
$$

onde

$$
c(n, \alpha_0) := T^{-1}\left(1 - \frac{\alpha_0}{2}; n - 2\right),
$$

e

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

Quando $x_{\text{pred}} = \bar{x}$ a expressão se reduz um pouco e podemos deduzir que a largura do intervalo é

$$
\hat{l} = 2 \cdot c(n, \alpha_0) \cdot \hat{\sigma}_r' \sqrt{\left[1 + \frac{1}{n}\right]}.
$$

Desejamos, portanto, encontrar n tal que

$$
\begin{aligned}
\Pr\left(\hat{l} < l\right) &\geq \gamma, \\
\Pr\left(\hat{\sigma}_r' < \frac{l}{2 \cdot c(n, \alpha_0) \cdot \sqrt{\left[1 + \frac{1}{n}\right]}}\right) &\geq \gamma,
\end{aligned}
$$

isto é conseguimos reduzir nossa afirmação probabilística a uma afirmação com respeito à f.d.a. (ou CDF) de $\hat{\sigma}_r'$. Para completar nossos cálculos só precisamos nos lembrar que $n\hat{\sigma}_r'/\sigma^2$ tem distribuição qui-quadrado com $n-2$ graus de liberdade (De Groot, Teorema 11.3.2) e, portanto,

$$\Pr\left(\hat{\sigma}_r' \leq a\right) = F_\chi\left(\frac{\sigma^2}{n}a; n-2\right).$$

■

Comentário: Nesta questão, retirada *ipsis litteris* da A2 2021, trabalhamos os efeitos de centrar a variável independente na distribuição dos estimadores dos coefficientes. Além disso, trabalhamos ideias de desenho experimental, determinando o tamanho amostral necessário para que a banda de predição na média da variável independente tenha uma certa largura com alta probabilidade.