Universidade de São Paulo
Escola de Artes, Ciências e Humanidades

ACH2053 – Introdução à Estatística – 1º sem. 2025
Professor: José Ricardo G. Mendonça

Gabarito — 2ª Prova — Data: 02 jul. 2025 — 19h00 às 20h45

Statistics: the mathematical theory of ignorance.
Morris Kline (1908–1992)

Na resolução dos problemas, explique seu raciocínio e o que você está fazendo de forma que eu possa acompanhá-lo(a). Soluções "mágicas" ou "geniais" não serão aceitas sem explicações.

Problemas

1. Um procedimento de controle de qualidade foi planejado para garantir um máximo de 10% de peças defeituosas na produção. Periodicamente sorteia-se uma amostra de 20 peças e, havendo mais de 3 peças defeituosas, a produção é parada para verificação.

   (a) Qual é a probabilidade de realizar uma parada desnecessária da produção?

   Mesmo que a produção esteja sob controle, com $p = 0{,}1$, existe uma probabilidade de que numa amostra de 20 peças haja mais de 3 peças defeituosas. Com X denominando o número de peças defeituosas, essa probabilidade vale

   $$\mathbb{P}(X > 3) = 1 - \mathbb{P}(X \leqslant 3) = 1 - \sum_{k=0}^{3} \mathbb{P}(X = k),$$

   com $\mathbb{P}(X = k) = \binom{20}{k} p^k (1-p)^{20-k}$, uma probabilidade binomial. Fazendo a conta encontramos[(*)]

   $$\begin{aligned}\mathbb{P}(X > 3) &= 1 - \sum_{k=0}^{3} \binom{20}{k} p^k (1-p)^{20-k} \\ &\simeq 1 - 0{,}121577 - 0{,}270170 - 0{,}285180 - 0{,}190120 = 0{,}132953,\end{aligned}$$

   isto é, a probabilidade de realizar uma parada desnecessária da produção é de aproximadamente 13,3%.

   Suponha que a produção esteja sob controle, com a proporção de peças defeituosas $p \leqslant 10\%$, e que as peças sejam vendidas em caixas de 100 unidades.

[(*)] Os valores de $\mathbb{P}(X = k)$ foram calculados com o comando `dbinom(0:3,20,0.1)` em R.

(b) Qual é a probabilidade de que uma caixa possua mais que 10% de peças defeituosas?

Precisamos calcular $\mathbb{P}(X > 10)$ com $X \sim \text{Bin}(100; 0{,}1)$,

$$\mathbb{P}(X > 10) = 1 - \mathbb{P}(X \leqslant 10) = 1 - \sum_{k=0}^{10} \binom{100}{k} (0{,}1)^k (0{,}9)^{100-k}.$$

Podemos calcular esse número usando a distribuição binomial, como no item (a),[(†)] ou, reparando que $\min(np, n(1-p)) > 5$, usar a aproximação normal $N(\mu, \sigma^2)$ para a distribuição binomial $\text{Bin}(n, p)$ com $\mu = np$, $\sigma^2 = np(1-p)$ e a correção de continuidade. Lançando mão desta aproximação temos, com $\mu = np = 10$ e $\sigma = \sqrt{np(1-p)} = 3$, que

$$\mathbb{P}_B(X \geqslant 11) \simeq \mathbb{P}_N(X > 10{,}5) = \mathbb{P}_N\left(\frac{X-\mu}{\sigma} > \frac{10{,}5-\mu}{\sigma}\right) \simeq \mathbb{P}_N(Z > 0{,}167).$$

Agora, $\mathbb{P}_N(Z > 0{,}167) = 1 - \mathbb{P}_N(Z \leqslant 0{,}167) = 1 - \Phi(0{,}167)$, e consultando uma tabela da distribuição normal padrão encontramos

$$\mathbb{P}_N(Z > 0{,}167) \simeq 1 - 0{,}566 = 0{,}434.$$

Esse valor é muito próximo ao valor exato dado por $0{,}416844\ldots$.

(c) Qual é a probabilidade de que uma caixa não possua peças defeituosas?

A probabilidade vale $\mathbb{P}(X = 0) = \binom{100}{0} p^0 (1-p)^{100} = 0{,}9^{100} \simeq 0{,}000027$, isto é, somente 1 em 37000 caixas (um valor mais preciso é 1 em 37648 caixas).

2. Uma pesquisa entre estagiários do Bacharelado em Ciências Anfibológicas, da UniCorner, revelou o seguinte conjunto de dados para os valores de suas bolsas-auxílio (em múltiplos de salário mínimo):

1,2	2,6	1,8	1,4	3,3	2,7	1,1	4,1
1,8	1,9	1,8	2,4	3,9	2,3	1,6	1,7
1,8	1,3	2,8	1,7	1,0	1,9	1,3	1,2

Supondo que as bolsas-auxílio são uniformemente distribuídas, use o método dos momentos para estimar os parâmetros da distribuição.

No método dos momentos comparamos os momentos da distribuição com os momentos empíricos obtidos a partir dos dados. A distribuição uniforme $U(a, b)$ possui dois parâmetros, a e b, e seus momentos são

$$\mu_1 = \mathbb{E}(X) = \frac{1}{2}(a+b) \quad \text{e} \quad \mu_2 = \mathbb{E}(X^2) = \frac{1}{b-a}\int_a^b x^2 \,\mathrm{d}x = \frac{1}{3}\cdot\frac{b^3-a^3}{b-a} = \frac{1}{3}(a^2+ab+b^2).$$

[(†)] O comando `1-pbinom(10,100,0.1)` em R retorna o valor `0.4168445` para essa probabilidade.

O primeiro momento poderia ter sido consultado no formulário, enquanto o segundo momento, reescrito como $\mu_2 = \mathbb{E}(X^2) - \mathbb{E}(X)^2 + \mathbb{E}(X)^2 = \mathrm{Var}(X) + \mu_1^2$, também poderia ser consultado no formulário (verifique esta última fórmula).

Igualando os momentos teóricos e empíricos obtemos o sistema de duas equações (não lineares) a duas incógnitas:

$$\mu_1 = \frac{1}{2}(a+b) = m_1,$$
$$\mu_2 = \frac{1}{3}(a^2+ab+b^2) = m_2.$$

Da primeira equação obtemos $b = 2m_1 - a$, que substituído na segunda equação fornece

$$a^2 - 2m_1 a + 4m_1^2 = 3m_2,$$

uma equação do segundo grau para a cujas raízes são dadas por $a = m_1 \pm \sqrt{3m_2 - 3m_1^2}$. Inserindo os valores numéricos para m_1 e m_2, a saber,

$$m_1 = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{24}(1{,}2 + 2{,}6 + \cdots + 1{,}2) = \frac{48{,}6}{24} = 2{,}025,$$
$$m_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 = \frac{1}{24}(1{,}2^2 + 2{,}6^2 + \cdots + 1{,}2^2) = \frac{114{,}8}{24} \simeq 4{,}783,$$

obtemos as soluções $a \simeq 3{,}456$ e $b = 2m_1 - a \simeq 0{,}594$ ou vice-versa, $a \simeq 0{,}594$ e $b = 3{,}456$. Como deve ser $b > a$, ficamos com a segunda solução, que fornece as estimativas de momentos para os parâmetros a e b da distribuição uniforme $U(a,b)$ a partir dos dados fornecidos. De fato, se gerarmos 1000 números aleatórios a partir da distribuição uniforme com esses parâmetros com o comando `X <- runif(1000,a,b)` em R obtemos, em uma das simulações, `mean(X)` $= 2{,}023002$ e `mean(X^2)` $= 4{,}771686$.

3. Seja $X_1, \ldots, X_n$ uma amostra aleatória simples de uma variável aleatória

$$X \sim f(x;\theta) = \frac{\theta^4}{6} x^3 \mathrm{e}^{-\theta x}$$

de parâmetro $0 < \theta < \infty$ no intervalo $0 < x < \infty$. Encontre o estimador de máxima verossimilhança $\hat{\theta}$ para θ.

O estimador de máxima verossimilhança é aquele que maximiza a probabilidade de que uma amostra aleatória simples $\boldsymbol{x} = (x_1, \ldots, x_n)$ tenha sido obtida a partir da distribuição,

$$\hat{\theta} = \operatorname*{argmax}_{\theta} L(\boldsymbol{x}\,|\,\theta) = \operatorname*{argmax}_{\theta} \prod_{i=1}^{n} f(x_i;\theta) = \operatorname*{argmax}_{\theta} \prod_{i=1}^{n} \frac{\theta^4}{6} x_i^3 \mathrm{e}^{-\theta x_i}.$$

Podemos manipular o produtório acima para obter a expressão equivalente

$$L(\boldsymbol{x}\,|\,\theta) = \prod_{i=1}^{n} \frac{\theta^4}{6} x_i^3 \mathrm{e}^{-\theta x_i} = \left(\frac{\theta^4}{6}\right)^n \left(\prod_{i=1}^{n} x_i\right)^3 \mathrm{e}^{-\theta \sum_{i=1}^{n} x_i}$$

Para encontrar o máximo de $L(\boldsymbol{x}|\theta)$ em relação a θ, podemos maximizar o logaritmo $\ell(\boldsymbol{x}|\theta)$ de $L(\boldsymbol{x}|\theta)$ em relação a θ,

$$\ell(\boldsymbol{x}|\theta) = \log L(\boldsymbol{x}|\theta) = 4n\log\theta - n\log 6 + 3\log\left(\prod_{i=1}^{n} x_i\right) - \theta\sum_{i=1}^{n} x_i.$$

Derivando $\ell(\boldsymbol{x}|\theta)$ em relação a θ e igualando a zero encontramos

$$\frac{\mathrm{d}\ell(\boldsymbol{x}|\theta)}{\mathrm{d}\theta} = \frac{4n}{\theta} - \sum_{i=1}^{n} x_i = 0,$$

que fornece a solução

$$\hat{\theta} = \frac{4n}{\sum_{i=1}^{n} x_i} = \frac{4}{\bar{x}},$$

onde $\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$ é a média aritmética dos valores amostrais.

4. Para ajudar a decidir o melhor estimador de determinado parâmetro populacional θ entre dois estimadores $\hat{\theta}'$ e $\hat{\theta}''$, simulou-se uma situação em que 1000 amostras de 10 valores cada foram retiradas de uma população para a qual $\theta = 100$, resultando em 1000 valores $t'_1, \ldots, t'_{1000}$ a partir da fórmula para $\hat{\theta}'$ e, igualmente, 1000 valores $t''_1, \ldots, t''_{1000}$ a partir da fórmula para $\hat{\theta}''$. O resumo das estatísticas obtidas a partir desses dados está a seguir:

Estimador $\hat{\theta}_1$	Estimador $\hat{\theta}_2$
$\bar{t}_1 = 102$	$\bar{t}_2 = 100$
$\mathrm{Var}\,(t_1) = 5$	$\mathrm{Var}\,(t_2) = 10$
$\mathrm{med}\,(t_1) = 100$	$\mathrm{med}\,(t_2) = 100$
$\mathrm{mod}\,(t_1) = 98$	$\mathrm{mod}\,(t_1) = 100$

Qual desses estimadores você considera melhor para estimar θ e por quê?

As principais características que tornam um estimador útil são que ele não seja enviesado, $\mathbb{E}(\hat{\theta}) = \theta$, e que ele seja consistente, $\lim_{n\to\infty} \mathbb{E}(\hat{\theta}_n) = \theta$ e $\lim_{n\to\infty} \mathrm{Var}(\hat{\theta}_n) = 0$. De outra forma, dados dois estimadores não enviesados e consistentes, é preferível usar aquele mais eficiente, isto é, aquele que possui a menor variância.

A tabela corresponde a dados obtidos com uma grande quantidade de amostras, $n = 1000$ amostras de 10 valores cada, de maneira que podemos assumir que os dados já representam os valores no limite de n grande. Nesse limite, a distribuição amostral da média tende a ser normal e simétrica; uma distribuição desse tipo possui média, moda e mediana coincidentes. Assim, embora a partir dos dados apresentados na tabela não seja possível dizer quais estimadores são consistentes, pode-se ver que o estimador $\hat{\theta}_2$ parece ser não enviesado e possuir uma distribuição mais simétrica que a do estimador $\hat{\theta}_1$. Dessa forma, a despeito do estimador $\hat{\theta}_1$ possuir uma variância menor, o que indicaria um estimador mais eficiente, devemos preferir o estimador $\hat{\theta}_2$, pelas suas demais características.

Formulário

Axiomas da teoria das probabilidades

Para quaisquer eventos (subconjuntos) A e B de um espaço amostral Ω valem $P(A) \geqslant 0$, $P(\Omega) = 1$ e $P(A \cup B) = P(A) + P(B)$ se $A \cap B = \varnothing$.

Como consequências dos axiomas da teoria das probabilidade valem $P(\overline{A}) = 1 - P(A)$, $P(\varnothing) = 0$ e $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ para quaisquer eventos $A, B \subseteq \Omega$.

Identidades de Bayes

$$P(A|B)P(B) = P(B|A)P(A), \quad P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i),$$

onde $A_i \cap A_j = \varnothing$ se $i \neq j$ e $A_1 \cup A_2 \cup \cdots \cup A_n = \Omega$.

Algumas distribuições de probabilidade

- Binomial $\mathrm{Bin}(n,p)$: $p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}$, $k = 0, 1, \ldots, n$;
  $\mathbb{E}(X) = np$, $\mathrm{Var}(X) = np(1-p)$.
- Poisson(λ): $p_X(k) = e^{-\lambda} \dfrac{\lambda^k}{k!}$, $k = 0, 1, \ldots$; $\mathbb{E}(X) = \lambda$, $\mathrm{Var}(X) = \lambda$.
- Uniforme $U(a,b)$: $f_X(x) = \dfrac{1}{b-a}$, $a \leqslant x \leqslant b$; $\mathbb{E}(X) = \frac{1}{2}(a+b)$, $\mathrm{Var}(X) = \frac{1}{12}(b-a)^2$.
  Às vezes, por conveniência o domínio da distribuição uniforme é dado como $a < x < b$.
- Exponencial $\mathrm{Exp}(\beta)$: $f_X(x) = \dfrac{1}{\beta} e^{-x/\beta}$, $x \geqslant 0$; $\mathbb{E}(X) = \beta$, $\mathrm{Var}(X) = \beta^2$.
- Normal $N(\mu, \sigma^2)$: $f_X(x) = \dfrac{1}{\sqrt{2\pi}\,\sigma} \exp\Big[-\dfrac{(x-\mu)^2}{2\sigma^2} \Big]$, $-\infty < x < +\infty$;
  $\mathbb{E}(X) = \mu$, $\mathrm{Var}(X) = \sigma^2$.

Transformação entre distribuições normais

Se $F_X(x)$ é a cdf de uma v.a. $X \sim N(\mu, \sigma^2)$ e $\Phi(z)$ é a cdf de uma v.a. padrão $Z \sim N(0,1)$, então

$$F_X(x) = \Phi\Big(\frac{x-\mu}{\sigma} \Big), \quad \Phi(-z) = 1 - \Phi(z).$$

Alguns valores de $\Phi(z) = \mathbb{P}(Z < z) = \dfrac{1}{\sqrt{2\pi}} \displaystyle\int_{-\infty}^{z} \exp\big(-\tfrac{1}{2}u^2\big)\,\mathrm{d}u$ aparecem na tabela abaixo.

z	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998
3.5	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998

Tabela: Valores da distribuição cumulativa normal padrão $\Phi(z) = \mathbb{P}(Z < z)$ para $z \geqslant 0$.