Universidade de São Paulo
Escola de Artes, Ciências e Humanidades

ACH2053 – Introdução à Estatística – 1º sem. 2023
Professor: José Ricardo G. Mendonça

Gabarito — 2ª Prova — Data: 10 julho 2023 — 19h00 às 20h45

Deve-se preferir o impossível provável ao possível improvável.
Aristóteles (384–322 a. C.), *Poética*, XXIV

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. Uma máquina empacotadeira produz pacotes com massas ("pesos") distribuídas normalmente com média μ e desvio padrão 10 g.

   (a) Quanto deve valer μ para que no máximo 10% dos pacotes tenham menos que 500 g?
   Queremos descobrir o valor de μ tal que $P(X < 500) \leqslant 10\%$. Isso significa que

   $$P\Big(\frac{X-\mu}{\sigma} < \frac{500-\mu}{\sigma}\Big) \leqslant 0{,}1 \Rightarrow P\Big(Z < \frac{500-\mu}{10}\Big) \leqslant 0{,}1,$$

   onde $Z \sim N(0,1)$ é uma variável aleatória normal padrão. Assim, precisamos encontrar o valor de $z = (500-\mu)/10$ tal que $P(Z \leqslant z) = 0{,}1$. Claramente (faça um desenho), $z < 0$ e é igual ao negativo do valor de z' para o qual $P(Z \leqslant z') = 0{,}9$. Consultando uma tabela encontramos $z' \simeq 1{,}28$, de forma que $z = -1{,}28$ e finalmente obtemos $\mu = 512{,}8$ g.

   (b) Para o valor de μ encontrado no item (a), qual é a probabilidade de que a massa total de 10 pacotes escolhidos ao acaso seja inferior a 5,0 kg?
   Para que a massa total $X_1 + X_2 + \cdots + X_{10}$ de 10 pacotes escolhidos ao acaso seja inferior a 5,0 kg, a média amostral $\overline{X} = (X_1 + X_2 + \cdots + X_{10})/10$ deve ser inferior a 500 g. A distribuição da média amostral $\overline{X}$ é uma distribuição normal de média μ e desvio padrão $\sigma/\sqrt{10} = \sqrt{10}$ g, de maneira que estamos procurando a probabilidade

   $$P(\overline{X} < 500) = P\Big(\frac{\overline{X}-\mu}{\sigma/\sqrt{10}} < \frac{500-\mu}{\sigma/\sqrt{10}}\Big) = P(Z < -12{,}8/\sqrt{10}) \simeq P(Z < -4{,}05).$$

   Esse valor não consta da tabela fornecida porque é muito pequeno, de maneira que podemos considerar essa probabilidade igual a zero. Um cálculo usando o comando

`pnorm(-12.8/sqrt(10))` do pacote estatístico R fornece $P(Z < -12{,}8/\sqrt{10}) = 2{,}586 \times 10^{-5}$, aproximadamente 1 vez em cada 39 mil amostras.

Outra maneira completamente equivalente de resolver o problema seria reparar que se cada $X_i \sim N(\mu, \sigma^2)$, então $S_{10} = X_1 + X_2 + \cdots + X_{10} \sim N(10\mu, 10\sigma^2)$ e daí calcular $P(S_{10} < 5000)$ da maneira usual.

A título de controle de qualidade, de hora em hora é retirada da produção uma amostra de 20 pacotes. Se a média da massa na amostra for inferior a 497 g ou superior a 505 g a produção é parada para reajustar a empacotadeira.

(c) Qual é a probabilidade de se efetuar uma parada desnecessária da produção?

Uma parada desnecessária ocorre quando a média amostral $\overline{X} < 497$ g ou $\overline{X} > 505$ g sem que a máquina tenha se descalibrado, isto é, continue mantendo a média $\mu = 512{,}8$ g para a massa dos pacotes individuais. Essa probabilidade vale

$$P(\text{parada}) = P(\overline{X} < 497) + P(\overline{X} > 505),$$

e usando $\mu = 512{,}8$ g para a média da média amostral e $\sigma/\sqrt{20} = \sqrt{5}$ g para o desvio padrão da média amostral encontramos

$$P(\text{parada}) = P(Z < -7{,}07) + P(Z > -3{,}49) \simeq 0 + 0{,}9998 \simeq 1,$$

isto é, com certeza haverá paradas desnecessárias a todo momento, pois com $\mu = 512{,}8$ g, uma amostra de 20 pacotes dificilmente (com probabilidade $< 0{,}15\%$) possuirá média menor que $\mu - 3\sigma = 506{,}1$ g. Isso mostra uma incompatibilidade entre o critério $P(X < 500) \leqslant 10\%$ do item (a) com o critério de parada deste item.

(d) Se o valor de μ da empacotadeira se desregulou para 503 g, qual é a probabilidade de continuar a produção fora dos padrões desejados?

Este problema é o "oposto" do anterior. Se μ se desregula para 503 g, com que probabilidade ainda teremos $P(497 \leqslant \overline{X} \leqslant 505)$? Essa probabilidade vale

$$P(497 \leqslant \overline{X} \leqslant 505) = P\Big(\frac{497-503}{\sqrt{5}} \leqslant \frac{\overline{X}-503}{\sqrt{5}} \leqslant \frac{505-503}{\sqrt{5}}\Big) = P(-2{,}68 \leqslant Z \leqslant 0{,}89),$$

onde, novamente, $Z \sim N(0,1)$. Essa probabilidade vale $P(-2.68 \leqslant Z \leqslant 0.89) = \Phi(0{,}89) - \Phi(-2{,}68) = \Phi(0{,}89) - (1 - \Phi(2{,}68)) = 0{,}813 - (1 - 0{,}996) \simeq 81\%$, que é a probabilidade de a máquina continuar funcionando desregulada.

2. Um pesquisador deseja estimar a média de uma população usando uma amostra que seja grande o suficiente para garantir que a probabilidade de que a média amostral não difira da média da população por mais de 20% de seu desvio padrão possua pelo menos 90% de confiança. Qual deve ser o tamanho da amostra neste caso?

Queremos encontrar o valor de n, o tamanho da amostra, tal que $P(\left|\overline{X}-\mu\right| < 20\%\sigma) \geqslant 90\%$, onde μ é a média da população e σ é o desvio padrão da população. Como a média da média amostral também vale μ e o desvio padrão da média amostral vale $\sigma/\sqrt{n}$, podemos centralizar e normalizar a variável $\overline{X}$ para obter

$$P\left(\left|\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right| < 20\%\sqrt{n}\right) \geqslant 90\% \Rightarrow P(0 < Z < 20\%\sqrt{n}) \geqslant 45\%,$$

onde agora $Z = (\overline{X}-\mu)/(\sigma/\sqrt{n})$ é uma v.a. normal padrão $N(0,1)$. Consultando uma tabela para a distribuição normal padrão encontramos que $P(0 < Z < 20\%\sqrt{n}) \geqslant 45\%$ quando $20\%\sqrt{n} \geqslant 1{,}645$, isto é, quando $n \gtrsim 68$. O pesquisador precisa, portanto, de uma amostra com no mínimo 68 sujeitos para garantir que a média amostral apurada não difira da média da população por mais de 20% do desvio padrão da população.

3. Baseado em ensaios eletroquímicos detalhados, sabe-se que o desvio padrão do tempo de vida de determinado tipo de bateria industrial é de 15 horas. Uma amostra aleatória de 15 baterias exibiu um tempo de vida médio de 120 horas. Supondo que o tempo de vida das baterias é distribuído normalmente, encontre o intervalo de confiança de 96% para o tempo de vida médio real das baterias.

   Queremos encontrar o valor de ε tal que $P(\left|\overline{X}-\mu\right| \leqslant \varepsilon) \geqslant 96\%$, pois daí sabemos que $\overline{X}-\varepsilon \leqslant \mu \leqslant \overline{X}+\varepsilon$ com 96% de probabilidade. Não sabemos o valor de μ da população, mas como o valor esperado da média amostral vale μ e o desvio padrão da média amostral vale $\sigma/\sqrt{n}$, podemos centralizar e normalizar $\overline{X}$ para obter

$$P\left(-\frac{\varepsilon}{\sigma/\sqrt{n}} \leqslant \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \leqslant \frac{\varepsilon}{\sigma/\sqrt{n}}\right) \geqslant 96\% \Rightarrow P(-z_c \leqslant Z \leqslant z_c) \geqslant 96\%,$$

   onde $z_c = \sqrt{n}\varepsilon/\sigma$. Consultando uma tabela (faça também um desenho) encontramos $z_c \simeq 2{,}05$ (em R, o comando `qnorm(0.98)` retorna `2.053749`), e com $n = 15$ e $\sigma = 15\,\text{h}$ encontramos $\varepsilon \simeq 8{,}0\,\text{h}$. Assim, o intervalo de confiança de 96% para o tempo de vida médio real das baterias é dado por

$$\text{IC}(\mu;96\%) = (\overline{X}-\varepsilon, \overline{X}+\varepsilon) = (112, 128)\,\text{h},$$

   isto é, $112\,\text{h} \leqslant \mu \leqslant 128\,\text{h}$ com 96% de confiança.

4. A distribuição geométrica descreve a probabilidade de observar um sucesso no k-ésimo ensaio de Bernoulli após observar $k-1$ fracassos e é dada por

$$P(X=k) = \underbrace{(1-p)\cdots(1-p)}_{k-1 \text{ fracassos}}\, p = (1-p)^{k-1}p.$$

   Suponha que uma série de experimentos resulta nos dados $x_1, \ldots, x_n$ representando o número de tentativas realizadas antes de obter sucesso em cada experimento. Encontre uma estimativa de máxima verossimilhança para o parâmetro p da distribuição.

A função de máxima verossimilhança em nosso problema vale

$$L(p;x_1,\ldots,x_n)=\prod_{i=1}^{n}P(X_i=x_i)=\prod_{i=1}^{n}(1-p)^{x_i-1}p=(1-p)^{\sum_{i=1}^{n}(x_i-1)}p^n.$$

Essa função atinge seu máximo no valor de $p=\hat{p}$ para o qual $L'(\hat{p};x_1,\ldots,x_n)=0$. Derivando $L(p;x_1,\ldots,x_n)$ em relação a p encontramos que

$$L'(p;x_1,\ldots,x_n)\big|_{p=\hat{p}}=-\Big(\sum_{i=1}^{n}(x_i-1)\Big)(1-\hat{p})^{\sum_{i=1}^{n}(x_i-1)-1}\hat{p}^n+(1-\hat{p})^{\sum_{i=1}^{n}(x_i-1)}n\hat{p}^{n-1}=0$$

quando

$$-\frac{\hat{p}}{(1-\hat{p})}\Big(\sum_{i=1}^{n}(x_i-1)\Big)+n=0\Rightarrow\hat{p}=\frac{n}{\sum_{i=1}^{n}x_i}=\frac{1}{\bar{x}}.$$

Encontramos, portanto, que o estimador de máxima verossimilhança $\hat{p}$ do parâmetro p de uma distribuição geométrica é dado pelo inverso da média amostral dos dados. Esse resultado coincide com aquele que seria obtido a partir do método dos momentos, que fornece $\hat{p}=1/E(X)$ com $X\sim\text{Geom}(p)$.

Poderíamos também ter realizado os cálculos (talvez mais facilmente) usando o logaritmo da função de máxima verossimilhança, $\ell(p;x_1,\ldots,x_n)=\log L(p;x_1,\ldots,x_n)$.

Boas férias!

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 $P(\bar{A})=1-P(A)$, $P(\varnothing)=0$ 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 $N(\mu,\sigma^2)$: $f_X(x) = \dfrac{1}{\sqrt{2\pi}\,\sigma} \exp\Big[-\dfrac{(x-\mu)^2}{2\sigma^2}\Big]$, $-\infty < x < +\infty$;
  $\mathbb{E}(X) = \mu$, $\mathrm{Var}(X) = \sigma^2$.

Determinação do tamanho de uma amostra

Para que $\mathbb{P}(\big|\overline{X} - \mu\big| \leqslant \varepsilon) \geqslant \gamma$, onde $\varepsilon > 0$ é o erro amostral tolerado e $0 < \gamma < 1$ é o coeficiente de confiança, usando o CLT para $\overline{X}$ encontramos que devemos ter $\mathbb{P}(-z_\gamma \leqslant Z \leqslant z_\gamma) \simeq \gamma$, onde $Z \sim N(0,1)$ e $z_\gamma = \sqrt{n}\varepsilon/\sigma$ com σ^2 a variância da população a partir da qual $\overline{X}$ foi obtido.

Estimadores de mínimos quadrados (EMQ) e de máxima verossimilhança (EMV)

- Dado um conjunto de dados (x_1,y_1), …, (x_n,y_n) e um modelo $Y = g(X;\theta) + \varepsilon$ com $\mathbb{E}(\varepsilon) = 0$, o EMQ dos parâmetros θ é aquele que minimiza $S(\theta) = \displaystyle\sum_{i=1}^{n} \big[Y_i - g(X_i;\theta)\big]^2$.
- Dado um conjunto de valores $x_1,\ldots,x_n$ extraídos de uma população com distribuição $f(x;\theta)$, o EMV dos parâmetros θ é aquele que minimiza $L(\theta;x_1,\ldots,x_n) = \displaystyle\prod_{i=1}^{n} f(x_i;\theta)$ ou, equivalentemente, $\ell(\theta;x_1,\ldots,x_n) = \displaystyle\sum_{i=1}^{n} \log f(x_i;\theta)$.

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 $z \geqslant 0$.