UNIVERSIDADE DE SÃO PAULO
Gabarito Lista 2: MAE0261 – Introdução à Análise de dados

Alex Monito Nhancololo

Critérios usados na correção

A correção da lista baseou-se nos seguintes aspectos: resolução correta dos exercícios e detalhada, interpretação prática e não técnica dos resultados; organização, limpeza e boa estrutura; justificativas apresentadas sempre que necessárias; símbolos matemáticos e estatísticos escritos adequadamente; uso de régua em trabalhos feitos manualmente, em vez de simples traços à mão livre; presença obrigatória de legenda em todas as tabelas e gráficos; além de clareza e coerência na escrita, com linguagem objetiva e sem ambiguidades.

1) Distribuição Normal: $\mu = 80$ kg, $\sigma = 15$ kg

Para calcular as probabilidades, padronizamos a variável X para uma variável Z com distribuição Normal padrão ($Z \sim N(0, 1)$) usando a fórmula: $z = \frac{X-\mu}{\sigma}$.

(a) Probabilidade de ter menos que 50 kg

Queremos $P(X < 50)$.

$$\begin{aligned}
P(X < 50) &= P(\underbrace{\tfrac{X-\mu}{\sigma}}_{Z} < \tfrac{50-\mu}{\sigma}), \text{ subtraí } \mu \text{ e dividí por } \sigma \text{ ambos os lados} \\
&= P(Z < \tfrac{50-\mu}{\sigma}) \\
&= P(Z < \tfrac{50-80}{15}) \\
&= P(Z < \tfrac{-30}{15}) \\
&= \underbrace{P(Z < -2)}_{\text{este é um valor tabelado, } \textbf{veja aqui}} \\
&= 0,0228 \approx 2.28\%
\end{aligned}$$

Resposta: A probabilidade de uma pessoa adulta selecionada aleatoriamente ter menos que 50 kg é de 2,28%.

(b) Probabilidade de peso entre 50 e 100 kg

$$\begin{aligned}
P(50 < X < 100) &= P\left(\frac{50-\mu}{\sigma} < \underbrace{\tfrac{X-\mu}{\sigma}}_{Z} < \frac{100-\mu}{\sigma}\right) \\
&= P\left(\frac{50-80}{15} < Z < \frac{100-80}{15}\right) \\
&= P\left(\frac{-30}{15} < Z < \frac{20}{15}\right) \\
&= P(-2 < Z < 1.33), \quad \text{Note que: } P(a < X < b) = F_Z(b) - F_Z(a),\ F_Z(b) = \underbrace{\Phi(b)}_{Z\sim N(0,1)} = P(Z < b)
\end{aligned}$$

Aqui, $a = -2$ e $b = 1.33$ e, vamos usar $\Phi(\cdot)$, além de $F_Z(\cdot)$ que é o mais usado para normal padrão.

$$\begin{aligned}
&= \Phi(1.33) - \Phi(-2), \quad \text{Note que tanto } \Phi(\cdot), \text{ quanto } F_Z(\cdot) \text{ representa distribuições acumuladas} \\
&= P(Z < 1.33) - P(Z < -2), \quad \text{Note que estes são valores tabelados, } \textbf{veja aqui} \\
&= 0,9082 - 0,0228 \\
&= 0.8854 \approx 88,54\%
\end{aligned}$$

Resposta: A probabilidade de uma pessoa adulta selecionada aleatoriamente ter entre 50 kg e 100 kg é de 88,54%.

(c) Probabilidade de ter mais que 120 kg

$$\begin{aligned}
P(X > 120) &= P\left(\underbrace{\frac{X-\mu}{\sigma}}_{Z} > \frac{120-\mu}{\sigma}\right) \\
&= P(Z > \frac{120-80}{15}) \\
&= P(Z > \frac{40}{15}) \\
&= P(Z > 2.67), \quad \text{note que } P(Z > a) + P(Z \leq a) = 1, \text{ e, várias tabelas dão valores de } P(Z \leq a) \\
&= 1 - P(Z \leq 2.67) \\
&= 1 - 0,9962 \\
&= 0,0038 \approx 0,38\%
\end{aligned}$$

Resposta: A probabilidade de uma pessoa adulta selecionada aleatoriamente ter mais que 120 kg é de 0,38%.

(d) Peso Q_{95} tal que $P(X < Q_{95}) = 0,95$

Primeiro precisamos encontrar o valor z na distribuição Normal Padrão tal que $P(Z < z) = 0,95$. Este valor pode ser obtido consultando a tabela.

$$P(Z < z) = \Phi(z) = 0,95 \Longrightarrow z_{95\%} = 1,645$$

Para encontrar o peso Q95, recorremos à mesma fórmula de padronização acima usada, mas agora levando em consideração que temos o z, σ mas não temos o peso $(X = Q95)$

$$\begin{aligned}
z_{0,95} &= 1,645,\ \sigma = 15 \\
Z &= \frac{X-\mu}{\sigma} \\
Z \times \sigma &= X - \mu \\
Z \times \sigma + \mu &= X = Q_{95} \\
X = Q_{95} &= \mu + z \cdot \sigma \\
&= 80 + 1,645 \cdot 15 \\
&= 104,675\text{kg}
\end{aligned}$$

Resposta: O peso Q95 é 104,675 kg. Isso significa que 95% das pessoas adultas dessa população têm peso inferior a 104,675 kg.

(e) Intervalo simétrico em torno de 80 kg com probabilidade 95%

Como o intervalo é simétrico, queremos um intervalo $[80 - c; 80 + c]$ tal que $P(80 - c < X < 80 + c) = 0,95$. Imagine uma distribuição normal (a curva/sino), se a área central é 95%, então as duas caudas da distribuição somam $1 - 0,95 = 0,05$, onde 1 é 100% da área abaixo da curva da distribuição. Como o intervalo é simétrico, cada cauda tem uma área de $0,05/2 = 0,025 = 2,5\%$. Assim, temos que:
O valor de Z inferior (z_{inf}) deixa $2,5\%$ à sua esquerda, então

$$P(Z < z_{\text{inf}}) = \Phi(z_{\text{inf}}) = 0,025 \Longrightarrow z_{\text{inf}} = -1,96.$$

O valor de Z superior (z_{sup}) deixa $97,5\%$ à sua esquerda $(1 - 0,025)$, então

$$P(Z < z_{\text{sup}}) = \Phi(z_{\text{sup}}) = 0,975 \Longrightarrow z_{\text{sup}} = 1,96$$

$$\begin{aligned}
P(80 - c < X < 80 + c) &= 0,95 \\
P\left(\frac{(80-c)-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{(80+c)-\mu}{\sigma}\right) &= 0,95 \\
P\left(\underbrace{\frac{(80-c)-\mu}{\sigma}}_{z_{\text{inf}}} < Z < \underbrace{\frac{(80+c)-\mu}{\sigma}}_{z_{\text{sup}}}\right) &= 0,95 \\
\Phi\left(\frac{(80+c)-\mu}{\sigma}\right) - \Phi\left(\frac{(80-c)-\mu}{\sigma}\right) &= 0,95, \text{ veja exercício 1b) } (*)
\end{aligned}$$

$$\begin{cases}\Phi\left(\dfrac{(80+c)-\mu}{\sigma}\right)=\Phi(z_{\text{sup}}),\\ \Phi\left(\dfrac{(80-c)-\mu}{\sigma}\right)=\Phi(z_{\text{inf}}),\end{cases} \implies \begin{cases}\Phi\left(\dfrac{(80+c)-80}{\sigma}\right)=\Phi(z_{\text{sup}}),\\ \Phi\left(\dfrac{(80-c)-80}{\sigma}\right)=\Phi(z_{\text{inf}}),\end{cases} \implies \begin{cases}\Phi\left(\dfrac{c}{\sigma}\right)=\Phi(z_{\text{sup}}),\\ \Phi\left(\dfrac{-c}{\sigma}\right)=\Phi(z_{\text{inf}}),\end{cases} \implies \begin{cases}\dfrac{c}{\sigma}=z_{\text{sup}},\\ \dfrac{-c}{\sigma}=z_{\text{inf}}.\end{cases}$$

$\implies c = z_{\text{sup}} \cdot \sigma = 1,96 \times 15 = 29.4$

Portanto, o intervalo é $[80-c; 80+c] = [\,80-29.4;\ 80+29.4\,] = [50,6 \text{ kg}; 109,4 \text{ kg}]$

$\implies$ Este exercício poderia ter sido resolvido facilmente usando noção de intervalo de confiança:

$$IC_{95\%}: \left[\mu - \underbrace{\frac{z\times\sigma}{\sqrt{n}}}_{\text{Margem de erro}};\mu + \underbrace{\frac{z\times\sigma}{\sqrt{n}}}_{\text{Margem de erro}}\right], \text{ como o sorteio é de uma pessoa, } n=1, Z_{95\%}=1.96$$

$$IC_{95\%}: [\mu - z_{95\%}\times\sigma; \mu + z_{95\%}\times\sigma]$$
$$IC_{95\%}: [80-1.96\times 15; 80+1.96\times 15]$$
$$IC_{95\%}: [50,6 \text{ kg}; 109,4 \text{ kg}]$$

Resposta: O intervalo simétrico em torno de 80 kg que contém 95% dos pesos da população é $[50,6 \text{ kg}; 109,4 \text{ kg}]$

2) Seja X a população dos pesos, pelo exercício sabemos que $X \sim N(\mu,\sigma^2)$, $\mu = 80$, $\sigma = 15$. Sorteiou-se uma amostra de 9 pessoas ($n = 9$), e pelo exercício, sabemos que $\bar{X} \sim N(\mu_{\bar{X}}, \sigma^2_{\bar{X}})$, $\mu_{\bar{X}} = 80$ e $\sigma_{\bar{X}} = \dfrac{15}{\sqrt{9}} = 5$. Assim, $\bar{X} \sim N(80, 5^2)$

A probabilidade do peso médio estar entre 70 kg e 90 kg é dada por:

$$\begin{aligned}
P(70 < \bar{X} < 90) &= P\left(\frac{70-\mu_{\bar{X}}}{\sigma_{\bar{X}}} < \frac{\bar{X}-\mu_{\bar{X}}}{\sigma_{\bar{X}}} < \frac{90-\mu_{\bar{X}}}{\sigma_{\bar{X}}}\right)\\
&= P\left(\frac{70-\mu_{\bar{X}}}{\sigma_{\bar{X}}} < Z < \frac{90-\mu_{\bar{X}}}{\sigma_{\bar{X}}}\right)\\
&= P\left(\frac{70-80}{5} < Z < \frac{90-80}{5}\right)\\
&= P\left(\frac{-10}{5} < Z < \frac{10}{5}\right)\\
&= P(-2 < Z < 2)\\
&= \Phi(2) - \Phi(-2), \quad \text{veja exercício 1b)}\\
&= 0,9772 - 0,0228\\
&= 0,9544 \approx 95,44\%.
\end{aligned}$$

Resposta: A probabilidade do peso médio de uma amostra de 9 pessoas estar entre 70 kg e 90 kg é de 95,44%.

3) Pacotes de café (X): $\mu = 500$ g, $\sigma = 20$ g

(a) Probabilidade de que o peso médio de um pacote seja menor que 480 é $P(X < 480)$

$$P(X<480) = P(\underbrace{\tfrac{X-\mu}{\sigma}}_{Z} < \tfrac{480-\mu}{\sigma})$$
$$= P(Z < \frac{480-500}{20}) \quad = P(Z < \frac{-20}{20}) = P(Z < -1) \quad = \Phi(-1) = 0,1587 \approx 15,87\%$$

Resposta: A probabilidade de um pacote de café selecionado aleatoriamente pesar menos que 480 gramas é de 15,87%.

(b) Probabilidade de que o peso médio de 4 pacotes de café seja menor que 480 g é $P(\bar{X} < 480)$

$$\mu_{\bar{X}} = \mu = 500 \text{ g}, \sigma_{\bar{X}} = \underbrace{\frac{\sigma}{\sqrt{n}}}_{\text{Erro padrão}} = \frac{20}{\sqrt{4}} = \frac{20}{2} = 10 \text{ g}$$

$$\begin{aligned} P(\bar{X} < 480) &= P\left(\frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} < \frac{480 - \mu_{\bar{X}}}{\sigma_{\bar{X}}}\right) \\ &= P(Z < \frac{480 - 500}{10}) \\ &= P(Z < -2) \\ &= \Phi(-2) \\ &= 0,0228 \approx 2,28\% \end{aligned}$$

Resposta: A probabilidade de que o peso médio de 4 pacotes de café seja menor que 480 gramas é de 2,28%.

(c) Intervalo de confiança ($n = 10$**,** $\bar{x} = 490$**,** $s = 25$**), graus de liberdade (g.l)** $= n - 1$ **e nível de confiança de 95%**

A distribuição t de Student é simétrica.

$$IC_{95\%} = \left[\ \bar{x} - \frac{t_{(\frac{\alpha}{2}, n-1)} \cdot s}{\sqrt{n}}\ ;\ \bar{x} + \frac{t_{(\frac{\alpha}{2}, n-1)} \cdot s}{\sqrt{n}}\ \right]$$

$$IC_{95\%} = \left[\ 490 - \frac{t_{(0{,}025,\, 9)} \cdot 25}{\sqrt{10}}\ ;\ 490 + \frac{t_{(0{,}025,\, 9)} \cdot 25}{\sqrt{10}}\ \right]$$

com $\alpha = 1 - 0,95 = 0,05$, $\quad t_{0{,}025,9} \approx 2,262$, veja os valores **aqui**

$$IC_{95\%} \approx \left[490 - \frac{2,262 \cdot 25}{\sqrt{10}}\ ;\ 490 + \frac{2,262 \cdot 25}{\sqrt{10}}\right]$$

$$IC_{95\%} \approx \left[490 - \frac{56.55}{\sqrt{10}}\ ;\ 490 + \frac{56.55}{\sqrt{10}}\right]$$

$$IC_{95\%} \approx [490 - 17.88\ ;\ 490 + 17.88]$$

$$IC_{95\%} \approx [472,12\ ;\ 507.88]$$

Resposta: O intervalo de 95% de confiança para o peso médio de um pacote de café é $[472,12 \text{ g}, 507,88 \text{ g}]$.

Interpretação: Se coletássemos um grande número de amostras de 10 pacotes e construíssemos um intervalo de confiança para cada uma delas, esperaríamos que 95% desses intervalos contivessem o verdadeiro peso médio populacional (μ) dos pacotes de café.

Obs.: Não significa que há 95% de probabilidade de o verdadeiro valor estar neste intervalo específico, mas sim que o método utilizado para construir o intervalo é confiável em 95% das vezes.

4) Letalidade de uma nova doença a partir de uma amostra de 200 infectados, na qual a letalidade foi de 41%.

Temos que o tamanho da amostra (n) é 200 e a proporção amostral $\hat{p} = 41\% = 0,41$

(a) Intervalo de confiança para a proporção populacional de letalidade com 95% de confiança

$$IC_{95\%} = \left[\ \hat{p} - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p}}{n}}\ ;\ \hat{p} + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p}}{n}}\ \right]$$

$$\text{com } \alpha = 1 - 95\% = 0,05,\ \text{ e } z_{\frac{\alpha}{2}} = z_{\frac{0,05}{2}} = z_{0,025} = 1,96$$

$$IC_{95\%} = \left[\ 0,41 - 1,96 \cdot \sqrt{\frac{0,41(1-0,41)}{200}}\ ;\ 0,41 + 1,96 \cdot \sqrt{\frac{0,41(1-0,41)}{200}}\ \right]$$

$$= \left[\ 0,41 - 1,96 \cdot \sqrt{\frac{0,41(0,59)}{200}}\ ;\ 0,41 + 1,96 \cdot \sqrt{\frac{0,41(0,59)}{200}}\ \right]$$

$$= \left[\ 0,41 - 1,96 \cdot \sqrt{\frac{0,2419}{200}}\ ;\ 0,41 + 1,96 \cdot \sqrt{\frac{0,2419}{200}}\ \right]$$

$$= \left[\ 0,41 - 1,96 \cdot \sqrt{0,00012095}\ ;\ 0,41 + 1,96 \cdot \sqrt{0,00012095}\ \right]$$

$$= [\ 0,41 - 1,96 \cdot 0,03478\ ;\ 0,41 + 1,96 \cdot 0,03478\ ]$$

$$= [\ 0,41 - 0,06817\ ;\ 0,41 + 0,06817\ ]$$

$$= [\ 0,34183\ ;\ 0,47817\ ]$$

Resposta: O intervalo de 95% de confiança para a proporção populacional de letalidade é $[0,3418, 0,4782]$, ou seja, entre 34,18% e 47,82%.

(b) Tamanho da amostra para que o intervalo de confiança seja de 3 pontos percentuais para baixo e para cima

Desejamos uma margem de erro E de 3 pontos percentuais, ou seja, $E = 0,03$.

A fórmula para calcular o tamanho da amostra para uma proporção é: $n = \frac{(z_{\alpha/2})^2 \cdot \hat{p} \cdot (1-\hat{p})}{E^2}$, $\hat{p}$ é a proporção estimada e E a margem de erro desejada.

$$\begin{aligned} n &= \frac{(z_{\alpha/2})^2 \cdot \hat{p}\,(1-\hat{p})}{E^2} \\ &\text{com } \alpha = 1 - 95\% = 0,05,\ \text{ e } z_{\frac{\alpha}{2}} = z_{\frac{0,05}{2}} = z_{0,025} = 1,96 \\ &= \frac{(1,96)^2 \cdot 0,41(1-0,41)}{(0,03)^2} \\ &= \frac{3,8416 \cdot 0,41 \cdot 0,59}{0,0009} \\ &= \frac{0,92955504}{0,0009} \\ &\approx 1032,84 \end{aligned}$$

Note que o tamanho da amostra deve ser um número inteiro. Assim, é conveniente arredondar para o próximo número inteiro maior para garantir a margem de erro desejada. Ou seja, $n = 1033$

Resposta: O tamanho da amostra necessário para que o intervalo de confiança tenha uma margem de erro de 3 pontos percentuais é de 1033 infectados.