Primeira avaliação

Estatística não paramétrica

Paulo Ricardo Seganfredo Campana

18 de dezembro de 2023

Questão 1. Justifique o uso da estatística não-paramétrica, citando vantagens, desvantagens e comparando com a estatística paramétrica.

É importante para o uso da inferência estatística quando não temos informação sobre a distribuição dos dados, são técnicas para testes de hipótese mais gerais que requerem menos suposições. Em comparação, a estatística não paramétrica:

- Requer menos suposições sobre a distribuição dos dados, pois em geral muitos testes paramétricos precisam de uma distribuição normal ou assintóticamente normal.
- Funcionam também para dados nominais ou ordinais enquanto que testes paramétricos requerem números contínuos ou discretos.
- Não necessitam de grandes tamanhos de amostras como para testes paramétricos assintóticos.
- Não são tão poderodos quanto testes paramétricos equivalentes quando suas suposições são satisfeitas.
- Os p-valores e regiões críticas de alguns testes são mais dificeis de serem calculados.

Questão 2. Estima-se que cerca de 40% das mulheres que são submetidas à cirurgia do câncer de mama têm algum tipo de efeito após a cirúrgica. Um novo método de cirurgia foi realizado em 18 pacientes e 3 tiveram algum efeito. Ao nível de 5% de significância, podemos afirmar que o novo método é eficiente na redução dos efeitos? Calcule o p-valor.

Usarei um teste binomial acerca da probabilidade de efeito pós cirurgia.

$$\begin{cases} H_0 : p \geqslant 0.4, & \text{(novo método não tem efeito)} \\ H_1 : p < 0.4, & \text{(novo método tem efeito)} \end{cases}$$

O p-valor é a probabilidade de rejeitarmos H_0 quando ela e verdadeira, ou seja quando X tem distribuição $\text{Bin}(18, 0.4)$,

$$\text{p-valor} = P(X \leqslant 3|H_0) = \sum_{x=0}^{3} \binom{18}{x} 0.4^x 0.6^{18-x} = \text{`pbinom(3, 18, 0.4)`} = 0.0327813$$

Como o p-valor do de teste é menor que o nível de 5% de significância, rejeitamos a hipótese de que o novo método não tem efeito, portanto o mesmo é efeciente para a redução dos efeitos pós cirurgia.

Questão 3. Fãs de corrida de cavalos frequentemente sustentam que um corrida em torno de uma pista circular proporciona significativamente vantagem inicial para os cavalos colocados em certas posições dos postos. Em uma corrida de 8 cavalos, a posição 1 é a mais próxima da raia do lado interno da pista e a posição 8 está do lado externo, mais distante da raia. Observou-se uma amostra de 144 vencedores durante um mês de corridas em uma pista circular. O que podemos dizer sobre a suspeita dos fãs, ao nível de significância de 1%?

Usarei um teste Qui-quadrado para aderência do número de vitórias por raias a uma distribuição uniforme discreta.

$$\begin{cases} H_0 : p_1 = p_2 = \cdots = p_8 \\ H_1 : \text{pelo menos algum } p_i \neq p_j \end{cases}$$

Sob H_0, os cavalos de todas as raias teriam as mesmas chances de ganhar, então os valores esperados E_i devem ser iguais: $144/8 = 18$.

Raia	1	2	3	4	5	6	7	8	Total
O_i	29	19	18	25	17	10	15	11	144
E_i	18	18	18	18	18	18	18	18	144
$(O_i - E_i)^2/E_i$	121/18	1/18	0/18	49/18	1/18	64/18	9/18	49/18	294/18

Temos a estatística observada $T = 294/18 = 16.333$ enquanto que o quantil da distribuição de T sob H_0 para 1% de significância é $\chi^2_7(0.99) = 18.475$. Dessa forma não rejeitamos a hipótese nula de que os cavalos de todas as raias tem as mesmas chances de ganhar a corrida.

Questão 4. Verificar, ao nível $\alpha = 0.05$, se os dados abaixo se distribuem segundo uma distribuição normal.

Usarei o teste de Lilliefors para aderência a distribuição normal, primeiro é preciso estimar os parâmetros da distribuição e padronizar os dados com os mesmos.

$$\bar{x} = 9.58, \qquad s = 0.9784$$

X_i	8.50	8.70	8.80	9.10	9.30	9.50	9.60	9.80	11.00	11.50
Z_i	-1.10	-0.90	-0.80	-0.49	-0.29	-0.08	0.02	0.22	1.45	1.96
$\Phi(Z_i)$	0.13	0.18	0.21	0.31	0.39	0.47	0.51	0.59	0.93	0.98
$F_n(Z_i)$	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00
$\lvert\Phi(Z_i) - F_n(Z_i)\rvert$	0.03	0.02	0.09	0.09	0.11	0.13	0.19	0.21	0.03	0.02
$\lvert\Phi(Z_i) - F_n(Z_{i-1})\rvert$	0.13	0.08	0.01	0.01	0.01	0.03	0.09	0.11	0.13	0.08

O maior desvio entre a distribuição empírica e a distribuição normal é de 0.21, o quantil da distribuição de Kolmogorov é $T_{10}(0.95) = 0.409$, como a estatística observada é menor que o quantil da distribuição, não rejeitamos a hipótese de normalidade dos dados.

Questão 5. Usar a amostra seguinte para testar a hipótese nula de aleatoriedade. Use $\alpha = 0.05$.

O teste de iterações para aleatoriedade requer uma variável dicotômica, então irei categorizar os números entre aqueles acima e abaixo da mediana de $\tilde{x} = 22.45$

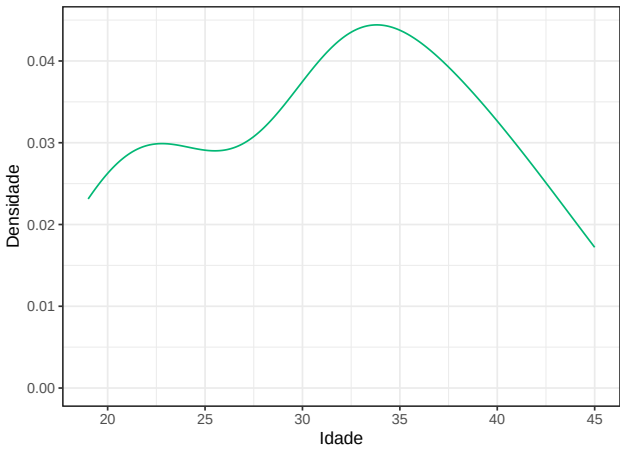
X_i	12.4	31.8	22.2	24.5	17.9	24.6	15.7	27.3	22.7	26.0	14.5	22.8	21.8	21.9	11.5	28.3
Y_i	↓	↑	↓	↑	↓	↑	↓	↑	↑	↑	↓	↑	↓	↓	↓	↑

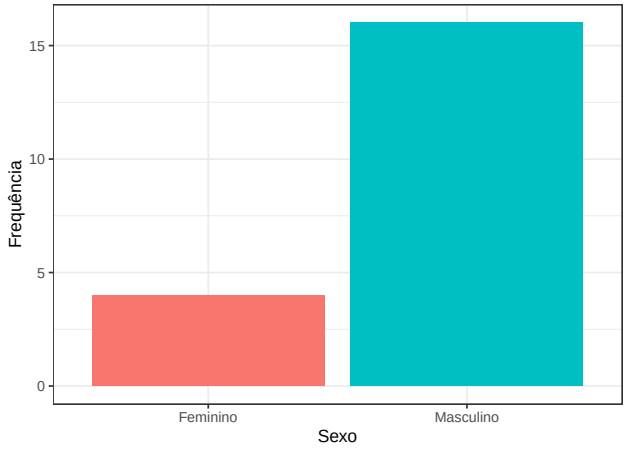
São no total 12 carreiras, em que 8 observações são acima e 8 abaixo da mediana. Assim, a região de aceitação do teste é entre 4 a 14 carreiras segundo a tabela para um nível de significância de 5%, então não rejeitamos a hipótese de aleatóriedade dos dados.

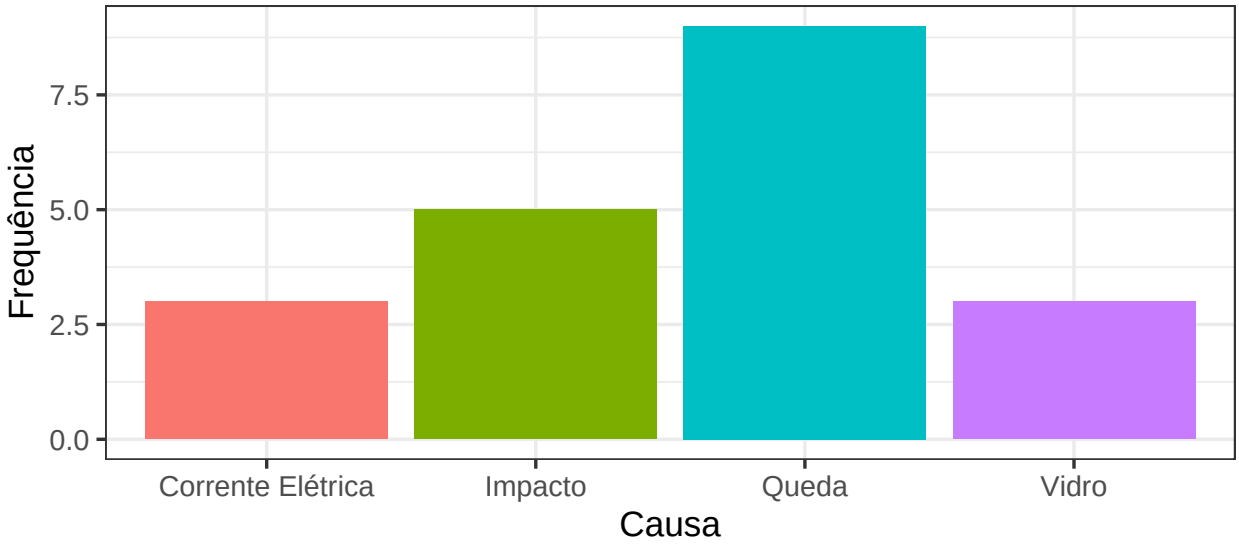
Questão 6. Analise a base de dados `acidentes.txt` para responder as perguntas a seguir.

a) Faça o gráfico que julgar mais adequado para cada uma das variáveis. O que você pode dizer, com base nos gráficos, acerca da distribuição dos dados?

Não podemos atribuir alguma distribuição conhecida a variável idade possívelmente pela pequena quantidade de dados, o sexo e a causa dos acidentes não ocorrem com frequências semelhantes entre seus grupos.







b) Teste a aderência da variável idade à distribuição normal padrão. Conclua;

A um nível de significância de 5%, os diferentes testes não rejeitam a hipótese da idade ter distribuição normal.

```
nortest::ad.test(acidentes$idade)$p.value
## [1] 0.3824371
nortest::cvm.test(acidentes$idade)$p.value
## [1] 0.4082063
nortest::lillie.test(acidentes$idade)$p.value
## [1] 0.4364655
nortest::pearson.test(acidentes$idade)$p.value
## [1] 0.1585976
nortest::sf.test(acidentes$idade)$p.value
## [1] 0.5119692
```

c) Com base no teste anterior, qual medida de posição você julga ser a mais adequada para representar a variável idade? Explique. Teste se essa medida de posição é igual a 30 anos. Conclua;

Como aparentemente a variável pode ter distribuição normal, é possivel testar a média ou a mediana, em ambos os casos concluimos que não rejeitamos a hipótese delas serem igual a 30.

```
t.test(acidentes$idade, mu = 30)$p.value
## [1] 0.336665
wilcox.test(acidentes$idade, mu = 30, exact = FALSE)$p.value
## [1] 0.3538399
```

d) A frequência de acidentes difere entre as causas?

Pelo gráfico parece que sim, mas o teste Qui-quadrado não rejeita a hipótese de que todas as causas de acidentes tenham a mesma probabilidade de ocorrência.

```
acidentes |>
    dplyr::count(causa) |>
    _$n |>
    chisq.test(p = c(0.25, 0.25, 0.25, 0.25)) |>
    _$p.value
## [1] 0.1870417
```

e) A variável sexo é aleatória? E a variável idade?

Dicotomizando a idade pela mediana, o teste de iterações não rejeita a hipótese da idade e do sexo serem aleatórios.

```
acidentes <- acidentes |>
    dplyr::mutate(
        sexo = as.factor(sexo),
        idade2 = as.factor(idade > median(idade))
    )

tseries::runs.test(acidentes$sexo)$p.value
## [1] 0.2992471
tseries::runs.test(acidentes$idade2)$p.value
## [1] 1
```