Estatística não paramétrica

Aula 4

Manoel Santos-Neto
Atualização: 24 de agosto de 2023

O que você irá aprender nesta aula?

1. Teste Binomial.

Introdução

Distribuição Binomial:

- n ensaios de Bernoulli independentes;
- Probabilidade de "sucesso" é constante.

Denotando $X : \#\,\text{de sucessos}$, tem-se que

$$\Pr(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, \ldots, n,$$

em que p é a probabilidade de sucesso.

Introdução

Aproximação Normal:

Seja $X \sim Bin(n, p)$, então

$$F_X(x) = \sum_{x=0}^{[x]} \binom{n}{x} p^x (1-p)^{n-x}, \quad \forall x \in \{0, 1, 2, \ldots, n\}.$$

Se n for muito grande, é inviável se utilizar a expressão acima, por exemplo

$$\Pr(X \leq 65) = 1 - \Pr(X \geq 66) = 1 - \sum_{x=66}^{100} \binom{100}{x} p^x (1-p)^{100-x}.$$

Todavia, se $n \to \infty$, $np \to \infty$ e $n(1-p) \to \infty$, tem que

$$\Pr(X \leq x) \to \Phi\left(\frac{x - np}{\sqrt{np(1-p)}}\right),$$

ou em outras palavras,

$$\frac{x - np}{\sqrt{np(1-p)}} \xrightarrow{D} N(0,1), \quad \text{(Teorema do Limite Central de Abraham de Moivre/Pierre-Simon Laplace (1733))}$$

isto é $X \cong N(np, np(1-p))$.

Introdução

A aproximação é melhor quando utilizamos a correção de continuidade, isto é,

$$\Pr(X \leq x) \cong \Phi\left(\frac{x - np + 1/2}{\sqrt{np(1-p)}}\right) \quad \text{e} \quad \Pr(X \geq x) \cong 1 - \Phi\left(\frac{x - np - 1/2}{\sqrt{np(1-p)}}\right).$$

Idéia

Introdução

A aproximação tende a ser melhor quando:

- $n \to \infty$;
- $p \cong 1/2$;
- $np \to \infty$ e $n(1-p) \to \infty$.

Quando $p \to 0$ ou $p \to 1$ de forma que $np \to \lambda_1$ ou $n(1-p) \to \lambda_2$ (eventos raros), então

$$X \xrightarrow{D} \text{Poisson}(\lambda_1).$$

Teste Binomial

Suposições:

- n ensaios independentes
- Variáveis dicotômica
- Probabilidade de "sucesso" é constante.

Sejam X : $\#$ de sucessos nos n ensaios e $p = \Pr(X_i = 1), i = 1, \ldots, n(\text{probabilidade de sucesso})$.

- **<u>Teste de Interesse</u>**

$$\mathcal{H}_0 : p = p_0, \quad p_0 \in (0, 1)\ (\text{especificado}).$$

i) Teste Unilateral

$$\mathcal{H}_1 : p > p_0\ (p < p_0).$$

Rejeitamos $\mathcal{H}_0$, ao nível de significância α, se

Teste Binomial

$$X_{obs} \geq X_{1-\alpha}; \quad \Pr(\text{Bin}(n, p_0) \geq X_{1-\alpha}) = \alpha \; (X_{obs} \leq X_{1-\alpha}; \quad \Pr(\text{Bin}(n, p_0) \leq X_{1-\alpha}) = \alpha).$$

Como a distribuição binomial é discreta, em geral, não é possível determinar testes não aleatorizados com nível α desejado, como por exemplo $\alpha = 0.05$, de forma que o nível de significância α é determinado diretamente da tabela da distribuição binomial.

Pelo motivo acima, uma melhor alternativa é tomar a conclusão com base no $\text{valor-}p$.

$$\text{valor-}p = \Pr(\text{Bin}(n, p_0) \geq X_{obs}) \; (\text{valor-}p = \Pr(\text{Bin}(n, p_0) \leq X_{obs})).$$

ii) Teste Bilateral

$$\mathcal{H}_1 : p \neq p_0.$$

Rejeitamos $\mathcal{H}_0$, ao nível de significância α, se

$$X_{obs} \geq X_{1-\alpha_1} \quad \text{ou} \quad X_{obs} \leq X_{\alpha_2}; \quad \text{com} \quad \alpha_1 + \alpha_2 = \alpha,$$

de forma que

Teste Binomial

$$\Pr(\text{Bin}(n, p_0) \geq X_{1-\alpha_1}) = \alpha_1 \quad \text{e} \quad \Pr(\text{Bin}(n, p_0) \leq X_{\alpha_2}) = \alpha_2.$$

Em geral toma-se por conveniência, o teste simétrico, obtido quando $\alpha_1 = \alpha_2 = \alpha/2$.

Como em geral não é possível determinar um teste não aleatorizado com nível α especificado, como por exemplo $\alpha = 0.05$, um critério bastante utilizado é através do valor-p.

Neste caso,

$$\text{valor-}p = 2\min\{\Pr(\text{Bin}(n, p_0) \geq X_{obs}); \Pr(\text{Bin}(n, p_0) \leq X_{obs})\}.$$

Quando $n \to \infty$, podemos utilizar o teste obtido através do TLC. A estatística de teste é dada por:

$$Z_{teste} = \frac{X_{obs} - np_0}{\sqrt{np_0(1 - p_0)}}.$$

Teste Binomial

$\mathcal{H}_1$	**Rejeita** $\mathcal{H}_0$	valor-p
$p > p_0$	$Z_{teste} \geq z_{1-\alpha}$	$\Pr(Z \geq Z_{teste})$
$p < p_0$	$Z_{teste} \leq z_{\alpha}$	$\Pr(Z \leq Z_{teste})$
$p \neq p_0$	$\lvert Z_{teste}\rvert \geq z_{1-\alpha/2}$	$2\min\{\Phi(Z_{teste}); 1-\Phi(Z_{teste})\}$

Intervalo de Confiança

i) Usando o TLC

$$IC_{1-\alpha}(p) = \left[\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right],$$

em que $\hat{p} = X_{obs}/n$.

Teste Binomial

ii) Exato

Se n for pequeno ou $p \approx 1/0$, o IC baseado no TLC pode não ser apropriado (no que tange a taxa de cobertura).

Idéia (Intervalo de Clopper-Pearson): Determinar para quais valores de p, temos:

$$\Pr(\text{Bin}(n,p) \geq X_{obs}) = \alpha/2 \rightarrow \sum_{x=x_{obs}}^{n} \binom{n}{x} p^x (1-p)^{n-x} = \alpha/2 \rightarrow \hat{p}_I,$$

e

$$\Pr(\text{Bin}(n,p) \leq X_{obs}) = \alpha/2 \rightarrow \sum_{x=0}^{x_{obs}} \binom{n}{x} p^x (1-p)^{n-x} = \alpha/2 \rightarrow \hat{p}_S.$$

Logo, um IC exato para p ao nível de confiança de $(1-\alpha)\%$ (note que este intervalo é simétrico) é

$$IC_{1-\alpha}(p) = [\hat{p}_I; \hat{p}_S].$$

Teste Binomial

No R

Exemplo: Em uma amostra de tamanho 20 foram observados 5 sucessos. Testar se $p > 0.2$ com nível de significância de 5%.

```
binom.test(x = 5, #numero de sucessos observados
           n = 20, #tamanho da amostra
           p = 0.2, #hipotese nula
           alternative = "greater", #hipotese alternativa (> ~ greater, < ~less e != ~two.sided)
           conf.level = 0.95 #nivel confiança
           )
```

Teste Binomial

```r
binom.test(x = 5, #numero de sucessos observados
           n = 20, #tamanho da amostra
           p = 0.2, #hipotese nula
           alternative = "greater", #hipotese alternativa (> ~ greater, < ~less e != ~two.sided)
           conf.level = 0.95 #nivel confiança
           )
```

```
##
##      Exact binomial test
##
## data:  5 and 20
## number of successes = 5, number of trials = 20, p-value = 0.3704
## alternative hypothesis: true probability of success is greater than 0.2
## 95 percent confidence interval:
##  0.1040808 1.0000000
## sample estimates:
## probability of success
##                   0.25
```

O resultado apresenta em sequência o número de sucessos, o número de tentativas e o valor-p. Além disso, é apresentada a hipótese nula, seguida pelo intervalo de confiança calculado. Por fim, temos a estimativa da probabilidade de sucesso calculada a partir da amostra.