Estatística não paramétrica

Aula 5

Manoel Santos-Neto
Atualização: 29 de agosto de 2023

O que você irá aprender nesta aula?

1. Teste χ^2 de aderência.
2. Função de distribuição empírica.

Teste qui-quadrado de Aderência

É comum também denotar o teste χ^2 de aderência por teste χ^2 de "qualidade/bondade" de ajuste. A grande maioria dos testes de hipóteses são referentes a parâmetros de funções de distribuições desconhecidas como por exemplo, média, variância, quantis, etc. Por outro lado, os testes de aderência (qualidade de ajuste) têm por objetivo comparar a amostra obtida com alguma função de distribuição hipotética para verificar se esta distribuição hipotética "ajusta bem" os dados amostrados.

O teste de aderência mais antigo é o teste χ^2, proposto por Pearson (1900, Philosophical Magazine e 1922, Biometrika).

Algumas Aplicações:

1. Retira-se uma amostra de k elementos e o interesse é verificar se as proporções são regidas pelas Leis de Mandel.
2. Considere que N estudantes devam optar cada um por 1 dentre n disciplinas oferecidas. A coordenação do curso tem interesse em verificar se não há preferência por qualquer das disciplinas oferecidas.

Teste qui-quadrado de Aderência

Amostra:

A amostra consiste de N observações independentes de uma v.a. X. Estas N observações são agrupadas em c classes, e o número de observações em cada classe é apresentada na forma de uma tabela de contigência $1 \times c$, isto é

Suposições:

1. A amostra é uma amostra aleatória.
2. A escola de medida é ao menos nominal.

Hipótese de Interesse:

$$\mathcal{H}_0 : \Pr(X \in \text{Classe}_j) = p_j, j = 1, \ldots, c, \quad \text{vs} \quad \mathcal{H}_1 : \Pr(X \in \text{Classe}_j) \neq p_j \text{ para ao menos uma classe.}$$

Teste qui-quadrado de Aderência

Observação:

Considere o vetor $\mathbf{Y} = (Y_1, \ldots, Y_n)$ em que $\forall i = 1, \ldots, c$ temos

$$Y_i := \sum_{k=1}^{N} \mathbb{I}(X_k \in c_i).$$

Note que

$$\mathbf{Y} \sim \mathrm{Multinomial}(N, p_1, \ldots, p_c),$$

em que $p_i = \Pr(X_i \in c_i), i = 1, \ldots, c.$]

Teste qui-quadrado de Aderência

Estatística de teste:

Sob $\mathcal{H}_0$, temos que o número esperado de observações na classe j é dada por

$$E_j = E[Y_j] = Np_j, j = 1, \ldots, c.$$

A estatística de teste é dada por

$$T := \sum_{j=1}^{c} \frac{(O_j - E_j)^2}{E_j}.$$

Sob $\mathcal{H}_0$ a estatística de teste $T \sim X^2_{(c-1)}$. Portanto, rejeitamos $\mathcal{H}_0$, ao nível de significância assintótico α, se $T \geq \chi^2_{(c-1);(1-\alpha)}$. Já o valor-p assintótico é obtido por $\Pr(X^2_{(c-1)} \geq t_{obs})$.

Introdução

Considere $X_1, \ldots, X_n \sim X$ e $x \in \mathbb{R}$ e que o interesse consiste em estimar $F_X(x)\ \forall x \in \mathbb{R}$. Um estimador bastante "simples" é dado pela função de distribuição empírica de X, definida abaixo.

Definição:

Considere $X_1, \ldots, X_n \sim X$. A função de distribuição empírica de X no ponto $x \in \mathbb{R}$ é definida por

$$\widehat{F}_n(x) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(X_i \leq x),$$

que é a proporção de elementos na amostra menores ou iguais a x.

Analogamente, um estimador "simples" de $S_X(x), \quad x \in \mathbb{R}$, é dado pela função de sobrevivência empírica de X:

$$\widehat{S}_n(x) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(X_i > x).$$

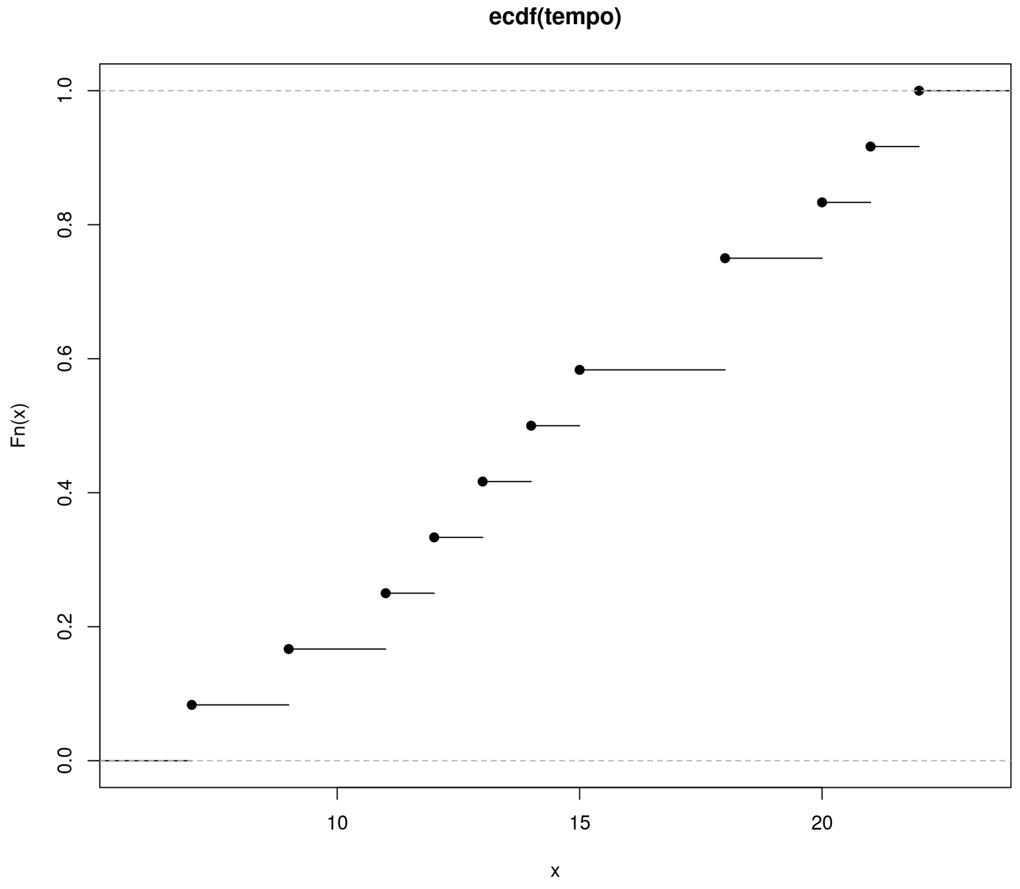
Exemplo (Controle de Qualidade)

Uma empresa de lâmpadas deseja invertigar o tempo (100h) de funcionamento de um novo tipo de lâmpada. Para isso, testou-se 12 lâmpadas e foi observado os tempos e em que ela pararam de funcionar, obtendo os seguintes valores: 20, 18, 7, 9, 12, 13, 22, 21, 14, 15, 18, 11.

- Desenhe o gráfico da função de distribuição empírica do tempo (100h) de funcionamento das lâmpadas. Qual a estimativa da probabilidade de uma lâmpada durar mais que 20000 horas?

Exemplo (Controle de Qualidade)

```
tempo <- c(20, 18, 7, 9, 12, 13, 22, 21, 14, 15, 18, 11)
plot(ecdf(tempo))
```



Exercício

Desenhe o gráfico da função de sobrevivência empírica de X definida no exemplo anterior.

Propriedades

Teorema:

Considere $X_1, \ldots, X_n \sim X$. Então $\forall x \in \mathbb{R}, \widehat{F}_n(x) \equiv (1/n)Y; \quad Y \sim \text{Bin}(n, F_X(x))$.

De maneira similar temos que, $\forall x \in \mathbb{R}$

$$\widehat{S}_n(x) \equiv (1/n)Y^\star; \quad Y^\star \sim \text{Bin}(n, S_X(x)).$$

Do teorema segue de forma imediata que

$$\mathrm{E}\left[\widehat{F}_n(x)\right] = F_X(x) \quad \forall x \in \mathbb{R}$$

e

$$\mathrm{Var}\left[\widehat{F}_n(x)\right] = \frac{F_X(x)(1 - F_X(x))}{n}. \quad \forall x \in \mathbb{R}$$

Exercício

Encontre $\mathrm{E}\left[\widehat{S}_n(x)\right]$ e $\mathrm{Var}\left[\widehat{S}_n(x)\right]$.y

Teorema (Glivenko-Cantelli)

Teorema:

Seja $X_1, \ldots, X_n \sim X$ então

$$\sup_{x \in \mathbb{R}} \left| \widehat{F}_n(x) - F_X(x) \right| \to 0.$$

Observação 1: A distância $\sup_{x \in \mathbb{R}} \left| \widehat{F}_n(x) - F_X(x) \right|$ é conhecida como distância de Kolmogorov, e serve como base para alguns testes não-paramétricos como veremos no decorrer do curso.

Observação 2: O teorema de Glivenko-Cantelli implica diretamente também que

$$\sup_{x \in \mathbb{R}} \left| \widehat{S}_n(x) - S_X(x) \right| \to 0.$$

Observação 3: Perceba que se tivermos interesse em testar hipóteses a respeito de $F_X(x)$, i.e, $\mathcal{H}_0 : F_X(x) = p_0$ é equivalente ao teste de quantis $\mathcal{H}_0 : x = x_{p_0}$.

Intervalo de confiança para $F_X(x)$

Para um dado valor $x \in \mathbb{R}$, podemos construir um intervalo de confiança (IC) para $F_X(x)$ usando a distribuição de $\widehat{F}_n(x)$, obtendo um IC exato com base na distribuição binomial, como já distutido na aula anterior. Uma outra opção é utilizar a aproximação normal.

$$\frac{\widehat{F}_n(x) - F_X(x)}{\sqrt{\frac{\widehat{F}_n(x)(1-\widehat{F}_n(x)}{n}}} \rightarrow N(0,1),$$

de forma que um IC de nível $(1-\alpha)$ para $F_X(x)$ é dada por

$$IC(1-\alpha, F_X(x)) = \left[\widehat{F}_n(x) \pm z_{(1-\alpha/2)}\sqrt{\frac{\widehat{F}_n(x)(1-\widehat{F}_n(x)}{n}}\right].$$

Exercício: Baseado na aproximação obtenha $\forall x \in \mathbb{R}$ um IC de nível $(1-\alpha)$ para $S_X(x)$.