Estatística não paramétrica

Aula 6

Manoel Santos-Neto
Atualização: 31 de agosto de 2023

O que você irá aprender nesta aula?

1. Teste de Kolmogorov.

Introdução

O teste de Kolmogorov é um teste de "qualidade de ajuste" proposto por Kolmogorov (1933) e consiste em testar se uma determinada amostra provêm de uma distribuição conhecida, X, com função de distribuição $F_0(x)$.

Teste de Kolmogorov versus teste χ^2

- Pode ser utilizado para dados ordinais em tabelas de contigência, ao contrário do teste χ^2;
- Trata os valores individualmente, sem a necessidade de agrupamento;
- É possível criar "faixas de confiança";
- Em geral, é mais poderoso, principalmente em pequenas amostras (Slakter, 1965);
- Teste pode ser aplicado, sem restrição, para pequenas amostras, ao contrário do teste χ^2.

Suposição:

$X_1, \ldots, X_n \overset{\text{iid}}{\sim} F_X(x)$.

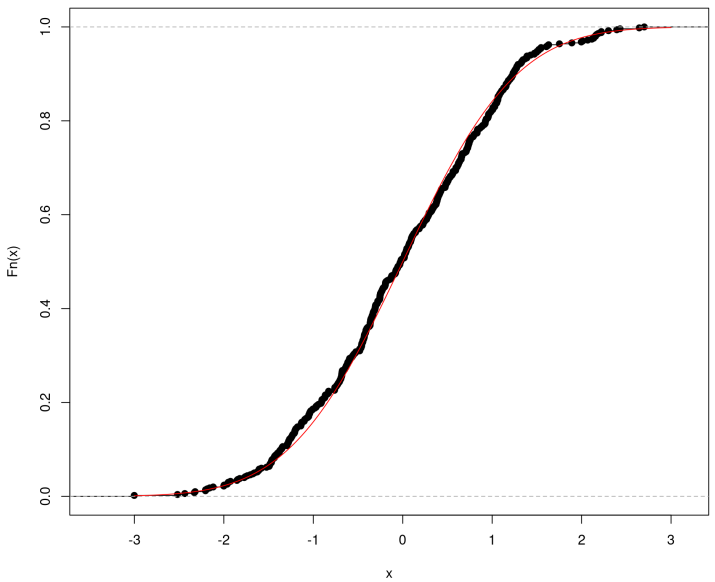
Introdução

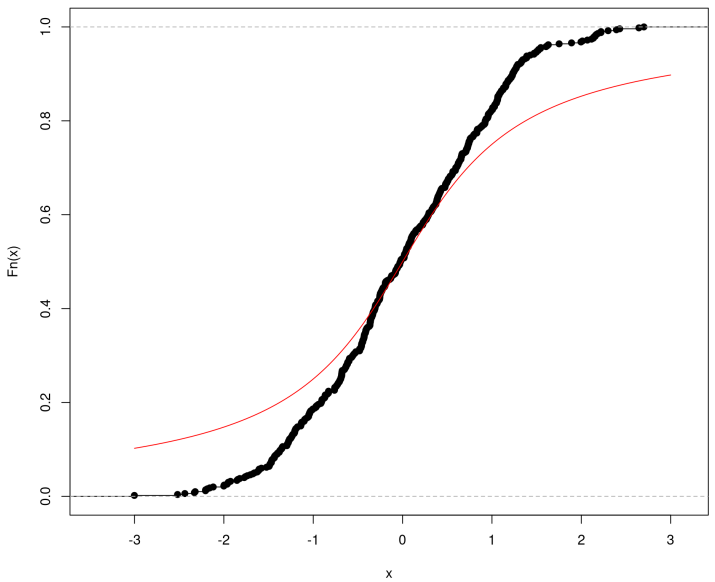
Interesse:

Testar se $F_X(x) = F_0(x)\, \forall x \in \mathbb{R}$, $F_0(x)$ completamente especificada.

Idéia:

Comparar $\widehat{F}_n(x)$(estimador consistente de $F_X(x)$) com $F_0(x)$.





Introdução

Podemos comparar $\widehat{F}_n(x)$ e $F_0(x)$ através de alguma distância $d(\widehat{F}_n(x), F_0(x))$.

Teste bilateral:

$\mathcal{H}_0 : F_x(x) = F_0(x); \forall x \in \mathbb{R}$

$\mathcal{H}_1 : \exists x \in \mathbb{R}\ F_x(x) = F_0(x)$

A estatística de teste utilizada, denotada por T_n é definida como sendo a maior distância vertical entre $\widehat{F}_n(x)$ e $F_0(x)$, isto é,

$$T_n := \sup_x \left| \widehat{F}_n(x) - F_0(x) \right|.$$

Observação: Se $\widehat{F}_n(x) \equiv F_0(x)$, então pelo Teorema de Glivenko-Cantelli, temos que $T_n \overset{\text{q.c}}{\to} 0$.

Para um teste de nível α, rejeitamos $\mathcal{H}_0$ se $T_n \geq t$, em que

$$\Pr_{\mathcal{H}_0}(T_n \geq t) = \alpha.$$

Teste bilateral

Como a função de distribuição empírica $\widehat{F}_n(x)$ é descontínua e a função de distribuição hipotética pode ser contínua, uma alternativa é considerar:

$$T_n^1 = \sup_x \left|\widehat{F}_n(x_{(i)}) - F_0(x_{(i)})\right| \quad \text{e} \quad T_n^2 = \sup_x \left|\widehat{F}_n(x_{(i-1)}) - F_0(x_{(i)})\right|.$$

Essas estatísticas medem as distâncias verticais entre os gráficos das duas funções, teóricas e empírica, nos pontos $x_{(i-1)}$ e $x_{(i)}$. Com isso, podemos utilizar como estatística de teste:

$$T_n = \max\{T_n^1, T_n^2\}.$$

No R

```
ks.test(x, y, ...,
        alternative = c("two.sided", "less", "greater"),
        exact = NULL, simulate.p.value = FALSE, B = 2000)
```

Teste Unilateral

Teste unilateral:

$\mathcal{H}_0 : F_x(x) \leq F_0(x); \forall x \in \mathbb{R} \quad \text{vs} \quad \mathcal{H}_1 : \exists x \in \mathbb{R}\ F_x(x) > F_0(x)$

A estatística de teste é dada por

$$T_n^+ := \sup_x \left\{ \widehat{F}_n(x) - F_0(x) \right\}.$$

Rejeitamos $\mathcal{H}_0$ ao nível α se $T_n^+ \geq t_1$, em que

$$\Pr_{\mathcal{H}_0}(T_n^+ > t_1) = \alpha.$$

Teste Unilateral

Teste unilateral:

$\mathcal{H}_0 : F_x(x) \geq F_0(x); \forall x \in \mathbb{R} \quad \text{vs} \quad \mathcal{H}_1 : \exists x \in \mathbb{R}\ F_x(x) < F_0(x)$

A estatística de teste é dada por

$$T_n^- := \sup_x \left\{ F_0(x) - \widehat{F}_n(x) \right\}.$$

Rejeitamos $\mathcal{H}_0$ ao nível α se $T_n^- \geq t_2$, em que

$$\Pr_{\mathcal{H}_0}(T_n^- > t_2) = \alpha.$$

Distribuição da estatística de teste

Para calcular o $\text{valor-}p$, devemos saber a distribuição EXATA de T_n, T_n^+ e T_n^-.

- Quando $F_0(x)$ é contínua, então sob $\mathcal{H}_0$, a função de distribuição EXATA de T_n^+ e T_n^- é dada por

$$G_{T_n^+} := \Pr(T_n^+ \leq x) = 1 - x \sum_{j=0}^{[nx(1-x)]} \binom{n}{j} \left(1 - x - \frac{j}{n}\right)^{n-j} \left(x + \frac{j}{n}\right)^{j},$$

em que $[b]$ é o maior número inteiro menor ou igual a b.

Quando $n \to \infty$, mostra-se que $\sqrt{n}T_n^+$ e $\sqrt{n}T_n^-$ tem função de distribuição aproximadamente igual a

$$\Pr\left(\frac{T_n^+}{\sqrt{n}} \leq x\right) \cong \lim_{n\to\infty} G_{T_n^+}\left(\frac{x}{\sqrt{n}}\right) = 1 - \exp\{-2x^2\}.$$

Já T_n tem função de distribuição aproximadamente igual (sob $\mathcal{H}_0$) a

$$\Pr(T_n \leq x) = G(x)^2,$$

dado que $\{\omega \in \Omega : T_n \leq x\} \leftrightarrow \{\omega \in \Omega : T_n^+ \leq x \quad \text{e} \quad T_n^- \leq x\}$.

Exemplo

Verifique se os dados abaixo podem ser ajustados por uma distribuição de Poisson com média igual a 1.2. Considere $\alpha = 0.05$.

Exemplo

X_i	f_i	f_0	$F_0(x)$	$F_n(x)$	$\lvert F_n(x) - F_0(x) \rvert$	$\lvert F_n(x_{(i-1)}) - F_0(x_{(i)}) \rvert$
0	15	18	0.25	0.30		
1	25					
2	10					
3	5					
4	4					
5	1					

Exercício

Considere uma amostra de 5 elementos com os seguinte valores ordenados: 0.28, 0.47, 0.54, 0.63, 0.68. Teste ao nível de significâcnia de 5% se esta amostra é oriunda de uma distribuição uniforme no intervalor $(0, 1)$.