Vantagens e Desvantagens dos Métodos não Paramétricos

Vantagens

- Não é necessário fazer suposições sobre a distribuição de probabilidades da variável de interesse na população.

- Podem ser aplicados a dados que não são exatos do ponto de vista numérico, mas que se disponham em uma escala ordinal. Existem técnicas para dados em escala nominal.

- Simplicidade nos cálculos.

- Aplicabilidade a pequenas amostras. (Muitas vezes, as amostras são pequenas pela própria natureza da variável).

Desvantagens

- Se aplicarmos um teste não paramétrico na situação em que é possível um teste paramétrico, teremos uma perda de eficiência (o correspondente teste paramétrico é mais poderoso)

- Não existem testes não paramétricos para alguns problemas para os quais existe um teste paramétrico.

Escala de Medidas

Nominal - quando as medidas servem apenas para classificar os elementos em classes ou categorias, não existindo uma ordenação entre essas categorias.

Ex: Sexo, Cor de Olhos.

Ordinal (ou escala por postos) - quando existir uma ordenação natural entre os elementos de diferentes categorias. O valor numérico da medida (caso exista) serve apenas como meio de ordenar as categorias.

Ex: Classe Social

Classificação do IBGE

A B C D E

A ↓ + 20 salários mínimos mensais

E ↓ 2 ou menos salários mínimos mensais

Intervalar - além da medida representar uma ordenação, como na escala ordinal, consegue-se dimensionar a magnitude da diferença entre dois elementos.

Exige um ponto zero (origem) e uma unidade de medida, que são no entanto arbitrários.

Ex: Temperatura

Escala: graus centígrados

Unidade °C

Origem °0 ou

Escala: Graus Fahrenheit

Unidade: °F

Origem: 32°F

Relação: $F = \frac{9}{5} C + 32$

Razão - quando a escala possuir todas as características da intervalar, com a exceção de que admite um verdadeiro ponto zero. (o zero corresponde à inexistência da quantidade a ser medida).

Ex: Peso : em Kg
Unidade Kg
Origem: 0

em gramas
Unidade gr
Origem : 0

Peso em gramas = 1000. Peso em Kg

Escala de Temperatura Kelvin é do tipo razão porque o zero corresponde à inexistência de energia. Também é denominada escala de temperaturas absoluta.

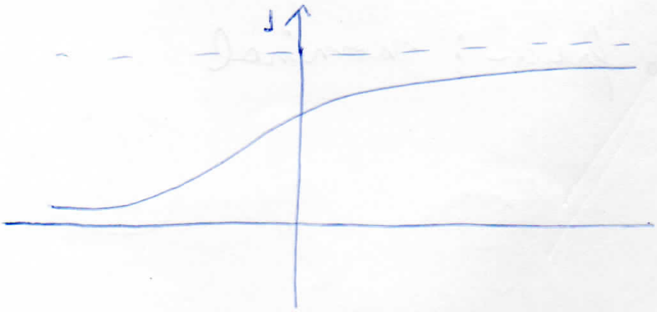
As estatísticas a serem calculadas e os testes a serem utilizados dependem do tipo de medida que foi tomada.

Escala	Estatísticas	Testes
nominal	Moda	não paramétricos
ordinal	Mediana Percentis r_s (Spearman) τ (Kendall)	
intervalar	Média Desvio Padrão Correlação de Pearson	Paramétricos e não paramétricos
razão	Média Geométrica Coeficiente de Variação $\left(\frac{DP}{\text{média}}\right)$	

Maioria dos métodos paramétricos - escala pelo menos intervalar

Função Distribuição Empírica

X v.a. , $F(x) = P(X \leq x) \rightarrow$ função distribuição acumulada



Idéia

amostra : $X_1, X_2 \ldots X_n$

$F_n(x)$ = proporção dos X_i menores ou iguais a x , $-\infty < x < \infty$

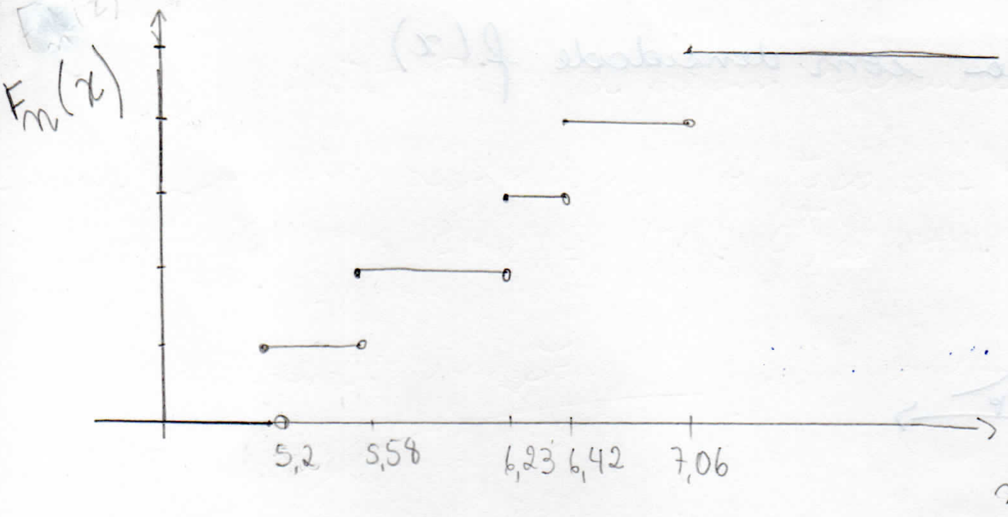
$F_n(x) \rightarrow$ f. d. empírica ou amostral

$F_n(x)$ é um estimador de $F(x)$.

Exemplo

$x_1 = 6{,}23 \quad x_2 = 5{,}58 \quad x_3 = 7{,}06 \quad x_4 = 6{,}42 \quad x_5 = 5{,}20$

$$F_n(x) = \begin{cases} 0 , & x < 5{,}2 \\ 1/5 , & 5{,}2 \leq x < 5{,}58 \\ 2/5 , & 5{,}58 \leq x < 6{,}23 \\ 3/5 , & 6{,}23 \leq x < 6{,}42 \\ 4/5 , & 6{,}42 \leq x < 7{,}06 \\ 1 , & x \geq 7{,}06 \end{cases}$$



f, escada

c) Quantis

O quantil de ordem p $(0 \leq p \leq 1)$ da v. aleatória X de uma dist. de probab. é um número x_p tal que

p = prob
x_p valor de X

$P(X < x_p) \leq p \qquad P(X > x_p) \leq 1 - p$

↳ área a esquerda de x_p é no máximo p, mas se for muito menor que p, a área à direita de x_p será maior que $1-p$.

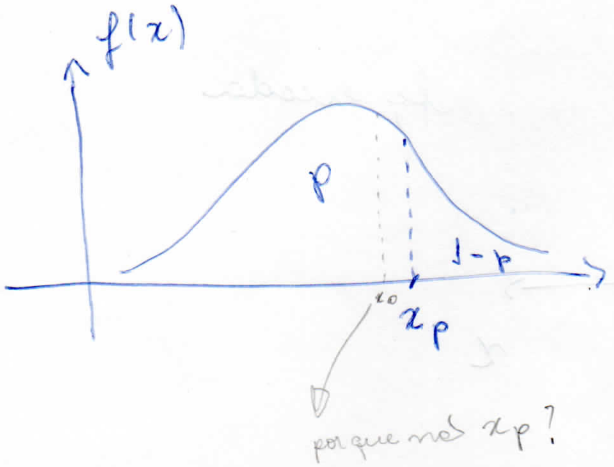
$p = 0,5 \Rightarrow x_p$ é a mediana

$p = 0,75$, $p = 0,25 \Rightarrow$ 3º e 1º quartis

$p = 0,4 \Rightarrow$ (x_p é o percentil de ordem 40%) quadragésimo percentil

Obs: Caso mais do que um valor satisfaça a definição de quantil, x_p = média entre o maior e o menor valor satisf. à def.

Ex: X v.a. contínua com densidade $f(x)$

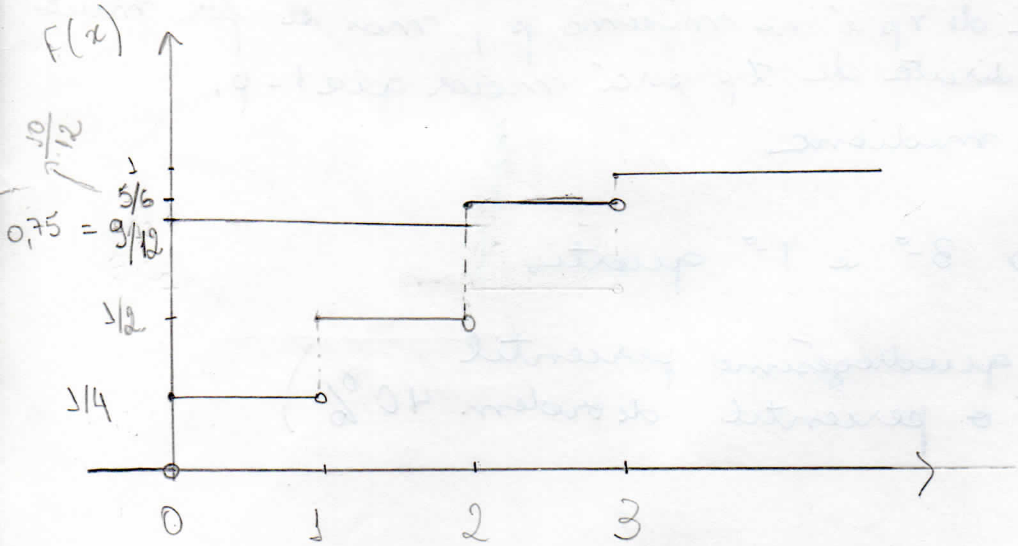


$P(X < x_0) \leq p$

$P(X > x_0) > 1-p$

Ex: X v.a. discreta com dist. de Prob:

$P(X=0) = 1/4 = P(X=1) \quad P(X=2) = \frac{1}{3} \quad P(X=3) = \frac{1}{6}$



$$\boxed{\begin{array}{l} P(X < x_p) \leq p \\ P(X > x_p) \leq 1-p \end{array}}$$

$x_{0,75} = 2 \qquad P(X<2) = 0,5 \qquad 1-p = 1-0,75 = 0,25$

$P(X>2) = 1/6 < 0,25$

Poderia ser 1,9? $P(X < 1,9) = 0,5 < 0,75 \mid P(X>1,9) = \frac{1}{3} + \frac{1}{6} = 0,5 > 0,25$

ser 2,1? $P(X < 2,1) = 0,5 + \frac{1}{3} = \frac{5}{6} > 0,75$

$x_{0,5} = 1,5$ $\quad \begin{cases} P(X<1) = \frac{1}{4} \leq 0,5 \\ P(X>1) = 0,5 \leq 1-0,5 \end{cases}$

$$\begin{cases} P(X<2) = 0,5 \leq 0,5 \\ P(X>2) = 1/6 \leq 0,5 \end{cases}$$

$$P(X<1,3) = 0,5 \leq 0,5$$

$$P(X>1,3) = 0,5 \leq 0,5$$

$x_{0,5} \in [1, 2]$ Verificar que é somente este intervalo.

$\therefore x_{0,5} = 1,5$

d) Quantis amostrais

Sejam $X_1, X_2 \dots X_n$ uma amostra aleatória da v. a. X. O quantil amostral de ordem p é o valor Q_p tal que a proporção dos X_i menores que Q_p é $\leq p$ e a prop. dos X_i maiores que Q_p é $\leq 1-p$.

$x_p \rightarrow$ parâmetro

$Q_p = \hat{x}_p$

Procedimentos não Paramétricos - Definição
(Conover - pag 118)

Um metodo estatístico é não paramétrico se satisfaz no mínimo a um dos seguintes critérios:

1 - O método pode ser usado em dados em uma escala nominal.

2 - O método pode ser usado para dados em uma escala ordinal.

3 - O método pode ser usado em dados em escala intervalar ou razão, quando a distribuição da variável aleatória que produziu os dados não foi especificada, ou foi especificada exceto para um nº infinito de parâmetros desconhecidos.

Leitura: Conover

Pag 114 - Seção 2.5 até Asymptotically Distribution-Free inclusive (pag 117)

Pag 85 - Desde "An approximate Confidence Interval" até computer Assistance (pag 88)

2 - Testes baseados na Distribuição Binomial

2.1 - Inferência para a proporção p.

Dados: Amostra de tamanho n consistindo dos resultados de n repetições independentes de um experimento (ensaios de Bernoulli). Cada resultado é classificado em uma de duas categorias mutuamente exclusivas: categoria 1 ou categoria 2.

T - número de resultados pertencentes à categoria 1

$n-T$ - número de resultados pertencentes à categoria 2

Suposições:

a) Os n ensaios são independentes.

b) Cada um deles tem probabilidade p de resultar categoria 1, p constante nos n ensaios.

A v.a. T tem distribuição binomial com parâmetros n e p.

$$T \sim B(n, p)$$

Se p é desconhecido, inferência sobre p.

Teste de Hipóteses

$H_0: p = p^*$ $\quad 0 \leq p^* \leq 1$

$H_a: p \neq p^*$

Intervalo de Confiança para p

Amostra aleatória de tamanho n

Estimador de p: $\hat{p} = \dfrac{T}{n}$

$$E(\hat{p}) = E\left(\frac{T}{n}\right)$$

$T \sim B(n, p)$ $\quad E(T) = np$ $\quad Var(T) = np(1-p)$

$$\therefore E(\hat{p}) = \frac{np}{n} = p$$

$$Var(\hat{p}) = \frac{1}{n^2} Var(T) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}.$$

1) Para n grande

$$\frac{T - np}{\sqrt{np(1-p)}} \begin{matrix} (\div n) \\ (\div(n)) \end{matrix} = \frac{\frac{T}{n} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \approx N(0,1)$$

$$\xrightarrow{\mathcal{D}} N(0,1)$$

Fixado γ, coeficiente de confiança, obtemos z_γ tal que

$$P\left(-z_\gamma \leq \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \leq z_\gamma\right) = P\left(-z_\gamma \leq Z \leq z_\gamma\right) = \gamma$$

$Z \sim N(0,1)$

Intervalo de Confiança para p com coef. γ é

$$\left[\hat{p} - z_\gamma\sqrt{\frac{p(1-p)}{n}} \; , \; \hat{p} + z_\gamma\sqrt{\frac{p(1-p)}{n}}\right]$$

Como p é desconhecido:

a) Intervalo de Confiança Otimista

$$\left[\hat{p} - z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \; ; \; \hat{p} + z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right]$$

amplitude do intervalo : $L = 2z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

b) Intervalo de Confiança Conservador

Como $p(1-p) \leq \frac{1}{4}$

$$\left[\hat{p} - z\sqrt{\frac{1}{4n}} \; , \; \hat{p} + z\sqrt{\frac{1}{4n}} \right]$$