Estatística não paramétrica
Problema de posição

Fernando Lucambio

Departamento de Estatística
Universidade Federal do Paraná

25 de março de 2024

Seja $X_1, X_2, \cdots, X_n$ uma amostra aleatória de tamanho n de alguma distribuição desconhecida F. Seja p um número real positivo satisfazendo $0 < p < 1$ e $\kappa_p(F)$ denotando o quantil de ordem p da distribuição F, ou seja, $\kappa_p(F)$ é tal que

$$P\Big(X \leq \kappa_p(F)\Big) = p,$$

isto devido a que no seguinte estudo vamos considerar F absolutamente contínua.

O problema de posição consiste em verificar se $H_0 : \kappa_p(F) = \kappa_0$, sendo κ_0 um valor dado, contra alguma das alternativas $\kappa_p(F) > \kappa_0$, $\kappa_p(F) < \kappa_0$ ou $\kappa_p(F) \neq \kappa_0$. O problema de posição e simetria consiste em verificar se $H_0 : \kappa_{0.5}(F) = \kappa_0$ e F é simétrica contra $H_0 : \kappa_{0.5}(F) \neq \kappa_0$ ou F não é simétrica.

Seja $X_1, X_2, \cdots, X_n$ uma amostra aleatória com função de densidade comum f. O problema aqui é verificarmos

$$H_0 : \kappa_p(f) = \kappa_0 \qquad H_1 : \kappa_p(f) > \kappa_0,$$

onde $\kappa_p(f)$ é o quantil de ordem p para f, $0 < p < 1$. Vejamos como construir a estatística de teste.

Seja $i(X_1, X_2, \cdots, X_n)$ o número de elementos positivos em

$$X_1 - \kappa_0, X_2 - \kappa_0, \cdots, X_n - \kappa_0.$$

Observemos que $P(X_i = \kappa_0) = 0$, desde que X_i seja uma variável aleatória do tipo contínua.

Pode ser demonstrado que o teste uniformemente mais poderoso para verificar H_0 versus H_1 é dado por

$$\varphi(x_1, x_2, \cdots, x_n) = \begin{cases} 1, & i(x_1, x_2, \cdots, x_n) > c, \\ \gamma, & i(x_1, x_2, \cdots, x_n) = c, \\ 0, & i(x_1, x_2, \cdots, x_n) < c, \end{cases}$$

onde c e γ são escolhidos de forma que

$$\sum_{k=c}^{n} \binom{n}{i} q^i p^{n-i} + \gamma \binom{n}{c} q^c p^{n-c} = \alpha,$$

com $q = 1 - p$. Isto é devido a que, sob H_0 verdadeira, $\kappa_p(f) = \kappa_0$, de maneira que $P(X \leq \kappa_0) = p$ e $P(X \geq \kappa_0) = q$ e $i(X) \sim Binomial(n, q)$.

A mesma estatística de teste pode ser utilizada em qualquer outra situação, ou seja, utiliza-se também caso $H_0 : \kappa_p(f) \leq \kappa_0$ versus $H_1 : \kappa_p(f) \geq \kappa_0$, $H_0 : \kappa_p(f) \geq \kappa_0$ versus $H_1 : \kappa_p(f) \leq \kappa_0$ e $H_0 : \kappa_p(f) = \kappa_0$ contra $H_1 : \kappa_p(f) \neq \kappa_0$.

As expressões do p-valor para este teste podem ser obtidas de maneira geral quando $p = 0.5$. Por exemplo, se a alternativa é de cauda superior $H_1 : \kappa_p(f) \geq \kappa_0$, o p-valor para o teste de sinal é dado pela probabilidade binomial na cauda superior

$$\sum_{i=i(x)}^{n} \binom{n}{i} 0.5^n,$$

sendo $i(x) = i(x_1, x_2, \cdots, x_n)$ o valor observado da estatística de teste. Poderíamos gerar tabelas e aplicar o teste de sinal exato para qualquer tamanho de amostra. Sabemos que a aproximação normal à binomial é especialmente boa quando $p = 0.5$.

Portanto, para valores moderados de n digamos, pelo menos 12, a aproximação normal pode ser usada para determinar as regiões de rejeição.

Como esta é uma aproximação contínua a uma distribuição discreta, uma correção de continuidade de 0.5 pode ser incorporada aos cálculos. Por exemplo, para a alternativa, $H_1 : \kappa_{0.5}(f) \geq \kappa_0$, H_0 é rejeitada para $i(x_1, x_2, \cdots, x_n) \geq \kappa_\alpha$, sendo κ_α satisfazendo

$$\kappa_\alpha = 0.5n + 0.5 + 0.5\sqrt{n}z_\alpha.$$

Similarmente, o p-valor aproximado é

$$1 - \Phi\left(\frac{\kappa_0 - 0.5 - 0.5n}{\sqrt{0.25n}}\right).$$

O teste do sinal para as diferentes hipóteses está programado no pacote **DescTools**, função **SignTest**.

Exemplo:

```
> library(DescTools)
> x = c(203,168,187,235,197,163,214,233,179,185,197,216)
> SignTest(x, mu = 195, alternative = "greater")

One-sample Sign-Test

data:  x
S = 7, number of differences = 12, p-value = 0.3872
alternative hypothesis: true median is greater than 195
98.1 percent confidence interval:
 179 Inf
sample estimates:
median of the differences
                      197
```

O teste do sinal de amostra única, descrito aqui, pode ser modificado para se aplicar à amostragem de uma população bivariada.

O teste de sinal perde informação, pois ignora a magnitude da diferença entre as observações e o quantil hipotético. O teste de postos sinalizados de Wilcoxon fornece um teste alternativo de posição e simetria que também leva em conta a magnitude dessas diferenças.

Seja $X_1, X_2, \cdots, X_n$ uma amostra aleatória com função de distribuição F absolutamente contínua que é simétrica a respeito da mediana m. O problema é testar

$$H_0 : m = m_0$$

contra as alternativas usuais unilaterais ou bilaterais.

Sem perda de generalidade, assumimos que $m_0 = 0$. Então $F(-x) = 1 - F(x)$ para todo $x \in \mathbb{R}$.

Por exemplo, para testar

$$H_0 : F(0) = \frac{1}{2} \quad \text{ou} \quad m = 0,$$

primeiro organizamos $|X_1|, |X_2|, \cdots, |X_n|$ em ordem crescente de magnitude e atribuímos postos $1, 2, \cdots, n$ mantendo-se a par dos sinais originais de X_i.

Por exemplo, se $n = 4$ e $|X_2| < |X_4| < |X_1| < |X_3|$ o posto de $|X_1|$ é 3, de $|X_2|$ é 1, de $|X_3|$ é 4 e de $|X_4|$ é 2.

Definimos

$$\begin{cases} T^+ &= \text{soma dos postos dos } X \text{ positivos}, \\ T^- &= \text{soma dos postos dos } X \text{ negativos} \end{cases}.$$

Então, considerando H_0 verdadeira, esperamos que T^+ e T^- sejam o mesmo.

Observe que

$$T^+ + T^- = \sum_{i=1}^{n} i = \frac{n(n+1)}{2},$$

de maneira que T^+ e T^- são linearmente relacionados e oferecem critérios equivalentes. Definimos agora

$$Z_i = \begin{cases} 1, & \text{caso } X_i > 0 \\ 0, & \text{caso } X_i < 0 \end{cases}, \qquad i = 1, 2, \cdots, n,$$

e escrevemos $r(|X_i|) = r_i$, para o posto de $|X_i|$.

Então $T^+ = \sum_{i=1}^{n} r_i Z_i$ e $T^- = \sum_{i=1}^{n} r_i (1 - Z_i)$. Também,

$$T^+ - T^- = -\sum_{i=1}^{n} r_i + 2\sum_{i=1}^{n} r_i Z_i = 2\sum_{i=1}^{n} r_i Z_i - \frac{n(n+1)}{2}.$$

As estatísticas T^+ e T^- são conhecidas como estatísticas de Wilcoxon. Um grande valor de T ou, equivalentemente, um pequeno valor de T significa que a maioria dos grandes desvios de 0 são positivos e, portanto, rejeitamos H_0 em favor da alternativa, $H_1 : m > 0$.

H_0	H_1	Rejeitamos H_0 se
$m = 0$	$m > 0$	$T^+ > c_1$
$m = 0$	$m < 0$	$T^+ < c_2$
$m = 0$	$m \neq 0$	$T < c_3$ ou $T > c_4$

Vamos encontrar agora a distribuição de T^+.

Seja

$$Z_{(i)} = \begin{cases} 1, & \text{se o } |X_i| \text{ que tem posto } i \text{ é } > 0, \\ 0, & \text{caso contrário} \end{cases}$$

É claro que $T^+ = \sum_{i=1}^{n} iZ_{(i)}$. As variáveis aleatórias

$$Z_{(1)}, Z_{(2)}, \cdots, Z_{(n)},$$

tem por distribuição Bernoulli, são não correlacionadas mas não necessariamente igualmente distribuídas.

Temos

$$\begin{aligned} \mathrm{E}\big(Z_{(i)}\big) &= P\big(Z_{(i)} = 1\big) = P\big([r(|X_i|) = i,\, X_j > 0] \text{ para algum } j\big) \\ &= P\big(i\text{-ésima estatística de ordem em} \\ &\qquad |X_1|, |X_2|, \cdots, |X_n| \text{ corresponde a um } X_j \text{ positivo}\big) \end{aligned}$$

Então

$$\mathrm{E}(Z_{(i)}) = \int_0^\infty n\binom{n-1}{i-1}[F_{|X|}(u)]^{i-1}[1-F_{|X|}(u)]^{n-i}f(u)\mathrm{d}u$$

a qual pode ser escrita como

$$\mathrm{E}(Z_{(i)}) = n\binom{n-1}{i-1}\int_0^\infty [F(u)-F(-u)]^{i-1}[1-F(u)+F(-u)]^{n-i}f(u)\mathrm{d}u,$$

onde f é a função de densidade de X.

Além disso

$$\mathrm{Var}(Z_{(i)}) = \mathrm{E}\Big(Z_{(i)}\big(1-\mathrm{E}(Z_{(i)})\big)\Big)$$

e

$$\mathrm{Cov}(Z_{(i)}, Z_{(j)}) = 0, \qquad i \neq j.$$

Sob H_0, X é simétrica a respeito de 0, de modo que $F(0) = \frac{1}{2}$ e $F(-u) = 1 - F(u)$, para todo $u > 0$.

Então

$$\mathrm{E}\big(Z_{(i)}\big) = n\binom{n-1}{i-1}\int_0^\infty \big(2F(u)-1\big)^{i-1}\big(2-2F(u)\big)^{n-i} f(u)\mathrm{d}u.$$

Escolhendo $\nu = 2F(u) - 1$, temos que

$$\mathrm{E}\big(Z_{(i)}\big) = \frac{n}{2}\binom{n-1}{i-1}\int_0^1 \nu^{i-1}(1-\nu)^{n-i}\mathrm{d}\nu$$

$$= \frac{n}{2}\binom{n-1}{i-1}B(i, n-i+1) = \frac{1}{2}.$$

Os momentos de T^+, em geral, são dados por

$$\mathrm{E}(T^+) = \sum_{i=1}^{n} i\mathrm{E}\left(Z_{(i)}\right)$$

e

$$\mathrm{Var}(T^+) = \sum_{i=1}^{n} i^2\mathrm{E}\left(Z_{(i)}\left(1-\mathrm{E}\left(Z_{(i)}\right)\right)\right).$$

De maneira que, considerando H_0 verdadeira, temos

$$\mathrm{E}_{H_0}(T^+) = \frac{1}{2}\sum_{i=1}^{n} i = \frac{n(n+1)}{4}$$

e

$$\mathrm{Var}_{H_0}(T^+) = \frac{1}{4}\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{24}.$$

Observe que $T^+ = 0$ se todas as diferenças tiverem sinais negativos e $T^+ = n(n+1)/2$ se todas as diferenças tiverem sinais positivos. Aqui, uma diferença significa uma diferença entre as observações e o valor postulado da mediana.

A estatística T^+ é completamente determinada pelas funções indicadoras $Z_{(i)}$, assim o espaço amostral pode ser considerado como um conjunto de $(z_1, z_2, \cdots, z_n)$, onde cada z_i é 0 ou 1, de cardinalidade 2^n.

Sob H_0, $m = m_0$ e cada arranjo é igualmente provável. Portanto

$$P_{H_0}(T^+ = t) = \frac{\left\{ \begin{array}{c} \text{no. de maneiras de atribuir sinais + ou - aos} \\ \text{inteiros 1,2,..., n para que a soma seja t} \end{array} \right\}}{2^n}.$$

Observemos que toda atribuição tem uma atribuição conjugada com sinais de mais e menos trocados, de modo que, este conjugado T^+ é dado por

$$\sum_{i=1}^{n} i(1 - Z_{(i)}) = \frac{n(n+1)}{2} - \sum_{i=1}^{n} iZ_{(i)}.$$

Assim, sob H_0, a distribuição de T^+ é simétrica em relação à média $n(n+1)/4$.

Exemplo:

Para os dados -0.465, 0.120, -0.238, -0.869, -1.016, 0.417, 0.056, 0.561 queremos verificar se $H_0 : m = 1.0$ versus $H_1 : m > 1.0$.

```
> x = c(-0.465,0.120,-0.238,-0.869,-1.016,0.417,0.056,0.561)
> library(ggplot2)
> dados = data.frame(x=c(rep(" ",8)), dados = x)
> qplot( x=x, y=dados, data=dados , geom=c("boxplot","jitter"))
> hist(x, xlab = "x", col = "green", border = "red",
       xlim = c(-1.5,1.5), ylim=c(0,0.8), breaks = 5,
       freq = F, ylab="dados", main = "")
> lines(density(x)); box(); grid()
> wilcox.test(x, mu = -1, alternative = "greater")

Wilcoxon signed rank exact test

data:  x
V = 35, p-value = 0.007813
alternative hypothesis: true location is greater than -1
```

Exemplo: