TAMANHO DA AMOSTRA PARA TESTES DE HIPÓTESES NÃO PARAMÉTRICOS

Belmer Garcia Negrillo

RELATÓRIO TÉCNICO Nº 30/89

Resumo: É analizado o problema do tamanho da amostra total $N = \sum n_i$, quando os n_i são iguais e as probabilidades de erro α e β são controladas. No caso de n_i diferentes é determinado o incremento necessário no tamanho total da amostra, para que o teste tenha o mesmo poder, que no caso de n_i iguais. Por último são dados alguns exemplos.

Abstract: The total sample size $N = \sum n_i$ is determined in the case of equal n_i for controlled error probabilities α and β. Tests of independence and the nonparametric tests for the one–two–and k–sample problems are discussed. For unequal n_i the increment of N necessary to attain the same power as in the case of equal n_i is determined.

Universidade Estadual de Campinas
Instituto de Matemática, Estatística e Ciência da Computação
IMECC – UNICAMP
Caixa Postal 6065
13.081 – Campinas – SP
BRASIL

Setembro – 1989

$$\text{onde} \quad \rho = \frac{\sigma(T)}{\sigma_0(T)}$$

o poder do teste é igual a $1-\beta$ se

$$\frac{u_0(T)-u(T)}{\rho\sigma_0(T)} + \frac{Z_\alpha}{\rho} = -Z_\beta$$

do que resulta

$$U(T) = \left(\frac{u(T)-u_0(T)}{\sigma_0(T)}\right)^2 = (Z_\alpha + \rho Z_\beta)^2$$

$U(T)$ é conhecido como o fator de não centralidade para o teste T.

Em geral o valor de ρ é desconhecido, assumamos $\sigma(T)$ fechado a $\sigma_0(T)$, isto é, $\rho = 1$, então para determinar o tamanho da amostra, devemos resolver a equação

$$U(T) = (Z_\alpha + Z_\beta)^2$$

No caso paramétrico para uma amostra, temos que $T = \overline{x}$ e

$$n = (Z_\alpha + Z_\beta)^2/[(u-u_0)/\sigma]^2$$

No caso não-paramétrico, como exemplo, vamos considerar alguns testes.

TESTE DO SINAL (TESTE DE LOCAÇÃO PARA UMA AMOSTRA)

A estatística do teste é dada por

$$S = \#(\text{observações positivas})$$

temos que $u(S) = Np$ e $\sigma^2(S) = Np(1-p)$ onde $p = P_r(x > 0)$, assim temos que sob H_0, $p = \frac{1}{2}$ do que resulta

$$u_0(S) = \frac{N}{2}\,,\ \sigma_0^2(S) = \frac{N}{4} \ \text{e} \ U(S) = 4N(p-\frac{1}{2})^2$$

Para determinar o tamanho da amostra, temos que resolver a equação

$$U(S) = (Z_\alpha + Z_\beta)^2 \quad \text{se} \ \rho = 1$$

e o requerido tamanho de amostra é dado por

$$N(S) = \frac{(Z_\alpha + Z_\beta)^2}{4(p - \frac{1}{2})^2}$$

Se $\rho \neq 1$, temos que

$$N(\rho) = \frac{(Z_\alpha + \rho Z_\beta)^2}{4(p - \frac{1}{2})^2}$$

neste caso do teste do sinal podemos usar

$$\rho = 2\sqrt{p(1-p)}$$

Na bibliografia especializada, também é sugerida

$$N(a) = [(Z_\alpha + Z_\beta)/\text{arc sen}(2p - 1)]^2$$

ou o uso da distribuição Binomial ($N(b)$).

Por exemplo para $\alpha = \beta = 0.10$ e vários valores de p temos de NOETHER (1987)

p	$N(S)$	$N(\rho)$	$N(a)$	$N(b)$
0.60	164.4	161.0	162.1	162
0.6666	59.2	55.8	56.9	57
0.75	26.3	22.9	24.0	24

TESTE DE WILCOXON (TESTE DE SIMETRIA EM ZERO)

A estatística do teste é dada por

$$W = \#[\text{positivos } (x_i + x_j)] \quad 1 \leq i \leq j \leq N$$

e temos que

$$u(W) = Np + \frac{1}{2}N(N-1)p'$$

onde

$$p = P_r(x > 0) \quad \text{e} \quad p' = P_r(x + x' > 0)$$

x, x' são duas observações independentes.

Sob H_0 temos que $p = p' = \frac{1}{2}$ do que resulta

$$u_0(W) = N(N+1)/4$$

e

$$\sigma_0^2(W) = N(N+1)(2N+1)/24$$

então

$$U(W) = \frac{[N(p - \frac{1}{2}) + \frac{1}{2}N(N-1)(p' - \frac{1}{2})]^2}{N(N+1)(2N+1)/24}$$
$$\doteq 3N(p' - \frac{1}{2})^2$$

e o requerido tamanho da amostra é

$$N(W) = \frac{(Z_\alpha + Z_\beta)^2}{3(p' - \frac{1}{2})^2}$$

Se $x = U + \eta$ $(\eta > 0)$ onde U é uma variável aleatória simétrica em zero, alguns valores de p e p' são conhecidos para certas distribuições.

se $U \sim$ uniforme $(-\frac{1}{2}, \frac{1}{2})$

$$p = \frac{1}{2} + \eta \quad p' = \frac{1}{2} + 2\eta(1-\eta) \quad \eta < \frac{1}{2}$$

se $U \sim n(0.1)$

$$p = \Phi(\eta) \quad p' = \Phi(\eta\sqrt{2})$$

se $U \sim$ Laplace $(0.1))$

$$p = 1 - \frac{1}{2}e^{-\eta} \quad p' = 1 - \frac{1}{2}(1+\eta)e^{-2\eta}$$

se $U \sim$ Cauchy (0.1)

$$p = p' = \frac{1}{2} + \frac{1}{\pi} tang^{-1} \eta$$

TESTE DE WILCOXON (TESTE DE LOCAÇÃO PARA DUAS AMOSTRAS)

Dado duas amostras aleatórias independentes $X_1 \ldots X_m$ e $Y_1 \ldots Y_n$ com $N = m+n$ e desejamos testar a hipótese nula de que as duas amostras vem da mesma população, contra a hipótese alternativa de que as observações X tendem a ser maiores que as observações Y.

A estatística do teste de Wilcoxon é dada por

$$W = \sum_{i=1}^{m} R_i$$

onde R_i é o posto de X_i na amostra combinada e a estatística de Mann Whitney é dada por

$$M = W - min\ W = W - \frac{1}{2}m(m+1)$$

sabemos que $u(M) = mnp''$ onde $p'' = P_r(x > y)$ e sob H_0 $p'' = \frac{1}{2}$ do que resulta

$$u_0(M) = \frac{1}{2}mn \qquad \sigma_0^2(M) = \frac{mn(N+1)}{12}$$

fazendo $m = n = \frac{1}{2}N$ encontramos que

$$U(M) = 3N^2(p'' - \frac{1}{2})^2/(N+1)$$

e

$$N = (Z_\alpha + Z_\beta)^2 / 3(p'' - \frac{1}{2})^2$$

TESTE DA MEDIANA PARA DUAS AMOSTRAS

Dadas duas amostras aleatórias independentes $X_1, \ldots, X_m$; $Y_1, \ldots, Y_n$ com $N =$

$m+n$. Se R_i é o posto X_i na amostra combinada temos que os escores da mediana são

$$a(R_i) = 1 \qquad R_i > \frac{1}{2}(N+1)$$

$$0 \qquad R_i \leq \frac{1}{2}(N+1)$$

e a estatística do teste para N par é

$$S = \sum_{i=1}^{m} a(R_i)$$

com $u(S) = mp''$ e $\sigma^2(S) = p''\frac{1}{2} \cdot \frac{m\,n}{N-1}$

Sob H_0 $p'' = 1/2$ do que resulta

$$u_0(S) = \frac{1}{2}m \quad \text{e} \quad \sigma_0^2(S) = \frac{1}{4}\,\frac{m\,n}{N-1}$$

assim teremos

$$U(S) = \frac{m^2(p'' - \frac{1}{2})^2}{\frac{1}{4}\,\frac{m\,n}{N-1}} \text{ e fazendo } m = n = \frac{1}{2}N$$

temos

$$U(S) = 4(p'' - \frac{1}{2})^2(N-1)$$

Substituindo $N-1$ por N e resolvendo a equação

$$U(S) = (Z_\alpha + Z_\beta)^2$$

temos que

$$N(S) = (Z_\alpha + Z_\beta)^2 \;/\; 4(p' - \frac{1}{2})^2$$

TESTE DE KRUS KALL-WALLIS (para $c \geq 2$ amostras independentes)

Dadas $c \geq 2$, amostras aleatórias independentes $X_{i1}, \ldots, X_i n_i$ para $i = 1, \ldots, c$ com $N = \sum_{i=1}^{c} n_i$ e se desejamos testar a hipótese nula de que as c amostras vem da mesma população temos que a estatística linear de postos para i-ésima amostra é

$$S_i = \sum_{j=1}^{n_i} R_{ij} - \frac{1}{2}n_i(n_i+1) \quad i = 1, \ldots, c$$

onde R_{ij} é o posto de X_{ij} na amostra conjunta.

Sabemos que

$$u(S_i) = n_i(N - n_i)p^*$$

onde

$$p^* = P_r(X_i > X_j) \quad \text{para algum} \quad i \neq j = 1, \ldots, (c-1)$$

e sob H_0 $p^* = \frac{1}{2}$, assim temos que

$$u_0(S_i) = \frac{1}{2}n_i(N - n_i)$$

$$\sigma_0^2(S_i) = \frac{1}{12}n_i(N - n_i)(N + 1)$$

do que resulta

$$U(S_i) = 12n_i(N - n_i)^2(p^* - \frac{1}{2})^2/(N + 1)$$

fazendo

$$n_i = \frac{1}{c}N$$

temos

$$U(S_i) = 12N(c-1)(p^* - \frac{1}{2})^2/c^2$$

resolvendo a equação

$$U(S_i) = (Z_\alpha + Z_\beta)^2$$

temos que

$$N = (Z_\alpha + Z_\beta)^2 c^2 \ / \ 12(c-1)(p^* - \frac{1}{2})^2$$

observemos que para $c = 2$, o valor de N coincide com o teste de Wilcoxon para duas amostras.

TESTE DE KENDALL (TESTE DE INDEPENDÊNCIA)

Dada a amostra aleatória bivariada $(x_1, y_1) \ldots (x_N, y_N)$, desejamos testar a hipótese que as duas variáveis aleatórias são independentes. A estatística do teste é

$$C = \#(\text{pares concordantes}) \quad \text{isto é } (x - x')(y - y') > 0$$

Sabemos que

$$u(C) = \frac{1}{2}N(N-1)P_c$$

onde P_c é a probabilidade de concordância, sob H_0 temos que $P_c = \frac{1}{2}$ do que resulta

$$u_0(C) = \frac{1}{4}N(N-1) \quad \text{e} \quad \sigma_0^2(C) = N(N-1)(2N+5)/72$$

e

$$N(C) = (Z_\alpha + Z_\beta)^2/9(P_c - \frac{1}{2})^2$$

A relação entre P_c e o coeficiente de correlação de Kendall (τ) é dada por

$$P_c = \frac{1}{2}(1+\tau)$$

TAMANHO TOTAL DA AMOSTRA, QUANDO AS AMOSTRAS SÃO DE TAMANHOS DIFERENTES

Para o mesmo tamanho total de amostras $N = \sum n_i$, os testes com amostras de tamanhos iguais são mais poderosos que aqueles em que os n_i são diferentes. Se $k \geq 1$ é a razão entre o maior e o menor n_i, de Hsieh (1987). Temos que o tamanho total da amostra quando os n_i são diferentes é

$$N_k = N(k+1)^2/4k$$

onde N é o tamanho total da amostra quando os n_i são iguais.

BIBLIOGRAFIA

1) HSIEH, F.Y. (1987) " A Simple Method of Sample Size Calculation for unequal-sample-size designs that use the Logrank or t-test." Statistics Medicine, vol. 6 pag 577-581.

2) NEGRILLO, B.G. (1988) "Metodos não paramétricos univariados e multivariados" - IMECC - UNICAMP.

3) NOETHER, G.E. (1987) "Sample Size Determination for some Common nonparametric tests", American Statistical Association vol. 82 nº 398 pag 645-647.

RELATÓRIOS TÉCNICOS — 1989

01/89 — **Uniform Approximation of Continuous Functions With Values in [0, 1]** — *João B. Prolla.*

02/89 — **On Some Nonlinear Iterative Relaxation Methods in Remote Sensing** — *A. R. De Pierro.*

03/89 — **A Parallel Iterative Method for Convex Programming with Quadratic Objective** — *Alfredo N. Iusem* and *Alvaro R. De Pierro.*

04/89 — **Fifth Force, Sixth Force, and all that: a Theoretical (Classical) Comment** — *Erasmo Recami* and *Vilson Tonin-Zanchin.*

05/89 — **An Application of Singer's Theorem to Homogeneous Polynomials** — *Raymundo Alencar.*

06/89 — **Summhammer's Experimental Test of the Non-Ergodic Interpretation of Quantum Mechanics** — *Vincent Buonomano.*

07/89 — **Privileged Reference Frames in General Relativity** — *Waldyr A. Rodrigues Jr.* and *Mirian E. F. Scanavini.*

08/89 — **On the Numerical Solution of Bound Constrained Optimization Problems** — *Ana Friedlander* and *José Mario Martínez.*

09/89 — **Dual Extremun Principles for the Heat Equation Solved by Finite Element Methods I** — *Vera Lucia da Rocha Lopes* and *José Vitório Zago.*

10/89 — **Local Convergence Theory of Inexact Newton Methods Based on Structured Least Chance Updates** — *José Mario Martínez*

11/89 — **Real Spin-Clifford Bundle and the Spinor Structure of Space-Time** — *Waldyr A. Rodrigues Jr.* and *Vera L. Figueiredo.*

12/89 — **A Multiplier Theorem on Weighted Orlicz Spaces** — *B. Bordin* and *J. B. Garcia.*

13/89 — **Dual Extremum Principles For The Heat Equation Solved By Finite Element Methods II** — *Vera Lucia da Rocha Lopes* and *José Vitório Zago.*

14/89 — **Dirac and Maxwell Equations in the Clifford and Spin-Clifford Bundels** — *W. A. Rodrigues Jr.* and *E. Capelas de Oliveira.*

15/89 — **Formal Structures, The Concepts of Covariance Invariance, Equivalent Reference Frames, and the Principle of Relativity** — *W. A. Rodrigues Jr., M. E. F. Scanavini* and *L. P. de Alcantara.*

16/89 — **Local Minimizers of a Quadratic Function With a Spherical Constraint** — *José Mario Martínez.*

17/89 — **On Pseudo-Convex Polycircular Domains In Banach Spaces** — *Mário C. Matos.*

18/89 — **On Circular and Special Units of an Abelian Number Field** — *Trajano Nóbrega.*

19/89 — Implementing Algorithms for Solving Sparse Nonlinear Systems of Equations — *Márcia A. Gomes-Ruggiero, José Mario Martínez* and *Antonio Carlos Moretti.*

20/89 — An Algorithm for Solving Nonlinear Least-Squares Problems with a New Curvilinear Search — *José Mario Martínez* and *Rita Filomena Santos.*

21/89 — Covariant Spinors, Algebraic Spinors, Operator Spinors and their Relationship — *V. L. Figueiredo, E. Capelas de Oliveira* and *W. A. Rodrigues Jr.*

22/89 — The Column–Updating Method for Solving Nonlinear Equations in Hilbert Space — *Márcia A. Gomes-Ruggiero* and *José Mario Martínez.*

23/89 — Automorphisms Control Systems and Observability — *Victor Ayala Bravo*

24/89 — Adjoint Functors Arise Everywhere — *T. M. Viswanathan.*

25/89 — New Sum Rules of Special Functions — *E. Capelas de Oliveira.*

26/89 — Minimal Realizations Under Controllability — *Luiz A. B. San Martin* and *Victor Ayala Bravo.*

27/89 — Optimization of Burg's Entropy Over Linear Constraints - *Yair Censor, Alvaro R. De Pierro* and *Alfredo N. Iusem.*

28/89 — Magnetic Monopoles without String in the Kähler–Clifford Algebra Bundle: A Geometrical Interpretation — *Adolfo Maia Jr., Erasmo Recami, Waldyr A. Rodrigues Jr.* and *Marcio A. F. Rosa.*

29/89 — On the Asymptotic Behavior of Some Alternate Smoothing Series Expansion Iterative Methods — *Alvaro R. De Pierro* and *Alfredo N. Iusem.*