MATD49-Estatística não paramétrica

6 – Testes de Wald-Wolfowitz

Kim Samejima

IME-UFBA

Teste de Wald–Wolfowitz para tendência I

- Considere uma amostra de tamanho N de uma v.a. X, $\{x_1, ..., x_N\}$ tal que as unidades amostrais possam ser categorizadas em dois grupos diferentes: A e B;
- Se X é contínua, empates nos valores de X, em tese, não ocorrem, no sentido de que $P(X_{n+k} = x_n) = 0$ para todo n e $k \neq 0$;
- Desta forma, teríamos uma ordenação natural $x_{(1)}, ..., x_{(N)}$. Ainda teríamos também uma ordenação natural dentro de cada grupo A e B: $x_{A,(1)}, \ldots, x_{A,(n_A)}$ e $x_{B,(1)}, \ldots, x_{B,(n_B)}$;
- Logo, se não há diferença entre as distribuições de $X|A$ e $X|B$, a ordem dos elementos dos grupos A e B na amostra deve ser aleatória;
- Por exemplo, se $n_A = 5$ e $n_B = 6$, poderíamos ter
  - AABABABBBAB, ou seja, o primeiro e segundo elementos da amostra são do grupo A, o terceiro é de B etc.;
  - Por outro lado também poderíamos ter BBBBBBAAAAA;

Teste de Wald–Wolfowitz para tendência II

- No primeiro exemplo teríamos uma mistura aleatória dos grupos A e B, sugerindo naquela amostra que não há diferença entre as distribuições de X nos grupos A B, ao passo que no segundo temos que a X no grupo B assume valores menores do que no grupo A.

- Sob a hipótese de que a distribuição de X não muda nos grupos A e B:

$$H_0 : F_{X|A}(x) = F_{X|B}(x),$$

desejamos avaliar se a ordem que estas observações aparecem possui tendência ou é aleatória. A hipótese alternativa será:

$$H_a : F_{X|A}(x) \neq F_{X|B}(x).$$

- Defina a variável R como o número total de grupos repetidos ($RUNS$) que temos na categorização da amostra em grupos A e B. Por exemplo:
  - Em AABABABBBAB, teríamos $R = 8$;
  - Em BBBBBBAAAAA, teríamos $R = 2$;
- Rejeitamos H_0 para valores pequenos de R.

Teste de Wald–Wolfowitz para tendência III

- Sob H_0, a distribuição de R é dada por:

$$f_R(r) = \begin{cases} 2\dfrac{\dbinom{n_A - 1}{r/2 - 1}\dbinom{n_B - 1}{r/2 - 1}}{\dbinom{n_A + n_B}{n_A}} & \text{, se } r \text{ é par} \\ \dfrac{\dbinom{n_A - 1}{(r-1)/2}\dbinom{n_B - 1}{(r-3)/2} + \dbinom{n_A - 1}{(r-3)/2}\dbinom{n_B - 1}{(r-1)/2}}{\dbinom{n_A + n_B}{n_A}} & \text{, se } r \text{ é impar} \end{cases} \tag{1}$$

Demonstração.

Em linhas gerais, é dada considerando o princípio fundamental da contagem. Veja [Gibbons and Chakraborti, 2011], Teorema 3.2.2. Valores tabelados para esta distribuição podem ser encontrados na mesma referência. □

Teste de Wald–Wolfowitz para tendência IV

- É possível mostrar que, para valores de n_A e n_B grandes (>20), podemos aproximar a distribuição de R por uma distribuição Normal:

$$R \overset{a}{\sim} N(\mu, \sigma^2),$$

em que:

$$\mu = \frac{2n_A n_B}{N} + 1,$$

$$\sigma = \sqrt{\frac{2n_A n_B(2n_A n_B - N)}{N^2(N-1)}}.$$

Demonstração.

TCL. $\square$

Exemplo [Gibbons and Chakraborti, 2011] I

Verificar que uma distribuição de Qui-quadrado pode ser aproximada por uma distribuição Normal para um número de g.l. grande, ao nível 5%.

- Vamos então avaliar a aderência de seus quantis utilizando os quantis da normal;
- Duas ditribuições de tamanho 8 foram geradas aleatoriamente para as variáveis χ^2_{18} e normal padrão;
- O resultado foi:

Normal	-1.91	-1.22	-0.96	-0.72	0.14	0.82	1.45	1.86
χ^2_{18}	4.90	7.25	8.04	14.10	18.30	21.21	23.10	28.12

- Antes de ordenar as observações, é preciso padronizar os dados da distr. de qui-quadrado para ter média zero e variância 1, assim como a normal. Vamos então, subtrair a média e dividir pelo seu desvio-padrão:

Normal	-1.91	-1.22	-0.96	-0.72	0.14	0.82	1.45	1.86
$(\chi^2_{18} - \nu)/\sqrt{2\nu}$	-2.18	-1.79	-1.66	-0.65	-0.05	0.54	0.85	1.69

- Os dados ordenados ficam:

$$-2.18, -1.91, -1.79, -1.66, -1.22, -0.96, -0.72, -0.65, -0.05, 0.14, 0.54, 0.82, 0.85, 1.45, 1.69, 1.86$$

Exemplo [Gibbons and Chakraborti, 2011] II

- Vamos agora calcular o número de *runs*:

(i)	1	2	3	4	5	6	7	8
$X_{(i)}$	-2.18	-1.91	-1.79	-1.66	-1.22	-0.96	-0.72	-0.65
R	Q	N	Q	Q	N	N	N	Q
(i)	9	10	11	12	13	14	15	16
$X_{(i)}$	-0.05	0.14	0.54	0.82	0.85	1.45	1.69	1.86
R	Q	N	Q	N	Q	N	Q	N

- Logo o número de *runs* neste exemplo é $R = 12$.
- Consultando os valores tabelados da Tabela D de [Gibbons and Chakraborti, 2011], temos que o p-valor para $n_A = n_B = 8$ e $R = 12$ é $p = 0.9$. Logo, não rejeitamos H_0 ao nível 5%.

Aspecto Computacional I

No R:

```
### DescTools::RunsTest: ###
RunsTest(x, y = NULL, alternative = c("two.sided", "less", "greater"),
                 exact = NULL, correct = TRUE, na.rm = FALSE, ...)
        x: Vetor com as categorias dos grupos ou vetor numerico.
        y: Vetor numerico para ser comparado com x.
           Se x eh numerico e y=NULL, o R testa x contra sua mediana.
        alternative: Hipotese alternativa.
        exact: Aprox. pela Normal? (TRUE/FALSE).
        na.rm: Remover NAs? (TRUE/FALSE).

### randtests::runs.test ###
        runs.test(x, alternative, threshold, pvalue, plot).
        x: Vetor numerico.
        alternative: Hipotese alternativa
                     ("two.sided", "left.sided", "right.sided").
        threshold: Ponto de corte para construir os grupos de x.
        pvalue: Aproximado ou exato? ("normal", "exact").
        plot: plotar o grafico? (TRUE/FALSE).
```

Referências I

Gibbons, J. D. and Chakraborti, S. (2011).
Nonparametric statistical inference.
Crc Press, Cop.