Estatística não paramétrica
Testes de aleatoriedade

Fernando Lucambio

Departamento de Estatística
Universidade Federal do Paraná

03 de abril de 2024

Imagine uma fila de dez pessoas esperando para comprar um ingresso em uma sala de cinema em uma tarde de sábado, suponha que observemos o arranjo de cinco homens e cinco mulheres na fila como sendo

H, M, H, M, H, M, H, M, H, M.

Isso seria considerado um arranjo aleatório por gênero? Intuitivamente, a resposta é não, já que a alternância dos dois tipos de símbolos sugere mistura intencional por pares.

Esse arranjo é um caso extremo, assim como a configuração

H, H, H, H, H, M, M, M, M, M,

como um agrupamento intencional. Nas situações menos extremas, a aleatoriedade de um arranjo pode ser testada estatisticamente usando a teoria das corridas.

Dada uma sequência ordenada de um ou mais tipos de símbolos, uma corrida é definida como sendo uma sucessão de um ou mais tipos de símbolos que são seguidos e processados por um símbolo diferente ou nenhum símbolo.

As pistas para a falta de aleatoriedade são fornecidas por qualquer tendência dos símbolos de exibir um padrão definido na sequência.

Tanto o número de realizações quanto os comprimentos das realizações, que obviamente estão inter-relacionados, devem refletir a existência de algum tipo de padrão. Os testes de aleatoriedade podem, portanto, ser baseados em qualquer critério ou em alguma combinação deles.

Poucas corridas, muitas corridas, uma corrida de comprimento excessivo, muitas corridas de excesso de comprimento excessivo, etc. podem ser usadas como critérios estatísticos para rejeição da hipótese nula de aleatoriedade, uma vez que essas situações devem ocorrer raramente em uma sequência verdadeiramente aleatória.

A alternativa à aleatoriedade é muitas vezes simplesmente falta de aleatoriedade. Em um teste baseado no número total de realizações, tanto poucas quanto muitas corridas sugerem falta de aleatoriedade. Uma hipótese nula de aleatoriedade seria consequentemente rejeitada se o número total de realizações fosse muito grande ou muito pequeno. No entanto, as duas situações podem indicar diferentes tipos de falta de aleatoriedade.

No exemplo do cinema, uma sequência com muitas realizações, tendendo a alternar os sexos, pode sugerir que o filme é popular entre adolescentes e jovens adultos, enquanto o outro arranjo extremo pode resultar se o filme é mais popular entre as crianças mais novas.

O teste de aleatoriedade é uma adição importante à teoria estatística, porque as bases teóricas para quase todas as técnicas clássicas, bem como os procedimentos livres de distribuição, começam com a suposição de uma amostra aleatória. Se essa suposição for válida, toda ordem sequencial é irrelevante.

No entanto, se a aleatoriedade das observações for suspeita, as informações sobre ordem, que quase sempre estão disponíveis, podem ser usadas para testar uma hipótese de aleatoriedade.

Os símbolos estudados para o padrão podem surgir naturalmente, como no exemplo do cinema ou podem ser impostos artificialmente. Assim, os testes de aleatoriedade são aplicáveis a dados qualitativos e quantitativos.

No último caso, a dicotomia é geralmente efetuada comparando-se a magnitude de cada número com um ponto focal, comumente a mediana ou a média da amostra e nada se cada observação excede ou é excedida por este valor. Ambas as técnicas usam a informação sobre magnitudes relativas de números adjacentes na sequência ordenada pelo tempo. Essas técnicas, chamadas de teste de subidas e descidas e o teste de von Neumann, usam mais as informações disponíveis e são especialmente efetivas quando a alternativa para a aleatoriedade é uma tendência ou autocorrelação.

Suponhamos uma sequência ordenada de n elementos de dois tipos, n_1 do primeiro tipo e n_2 do segundo tipo, onde $n_1 + n_2 = n$. Se r_1 for o número de realizações de elementos do tipo 1 e se r_2 for o número de realizações do tipo 2, o número total de realizações na sequência será $r = r_1 + r_2$. Para derivar um teste de aleatoriedade baseado na variável R, o número total de realizações, precisamos da distribuição de probabilidade de R quando a hipótese nula de aleatoriedade for verdadeira.

A distribuição de R será encontrada determinando primeiro a distribuição conjunta de R_1 e R_2 e depois a distribuição de sua soma. Uma vez que sob a hipótese nula cada arranjo dos n_1+n_2 objetos é equiprovável, a probabilidade de que $R_1 = r_1$ e $R_2 = r_2$ é o número de arranjos distinguíveis de $n_1 + n_2$ objetos com r_1 corridas tipo 1 e r_2 corridas de objetos do tipo 2 dividido pelo número total de arranjos distinguíveis, que é $n!/n_1!n_2!$.

Teorema

Denotemos por R_1 e R_2, respectivamente, o número de corridas de n_1 objetos do tipo 1 e o número de corridas de n_2 objetos do tipo 2 numa amostra aleatória de tamanho $n = n_1 + n_2$. A distribuição conjunta de R_1 e R_2 é

$$f_{R_1,R_2}(r_1, r_2) = \frac{c\binom{n_1-1}{r_1-1}\binom{n_2-1-1}{r_2-1}}{\binom{n_1+n_2}{n}},$$

quando $r_1 = 1, 2, \cdots, n_1$, $r_2 = 1, 2, \cdots, n_2$, $r_1 = r_2$ ou $r_1 = r_2 \pm 1$ sendo que $c = 2$ se $r_1 = r_2$ e $c = 1$ caso $r_1 = r_2 \pm 1$.

Demonstração. Veja material de consulta.

Corolário

A distribuição marginal de R_1 é

$$f_{R_1}(r_1) = \frac{\binom{n_1-1}{r_1-1}\binom{n_2-1+1}{r_1}}{\binom{n_1+n_2}{n}},$$

quando $r_1 = 1, 2, \cdots, n_1$. Similarmente para R_2 com n_1 e n_2 permutados.

Demonstração. Veja material de consulta.

Teorema

A distribuição de R, o número total de corridas de $n = n_1 + n_2$ objetos, n_1 do tipo 1 e n_2 do tipo 2, tem função de densidade dada por

$$f_R(r) = \begin{cases} \dfrac{2\binom{n_1-1}{\frac{r}{2}-1}\binom{n_2-1}{\frac{r}{2}-1}}{\binom{n_1+n_2}{n_1}}, & r \text{ par} \\ \dfrac{\binom{n_1-1}{\frac{r-1}{2}}\binom{n_2-1}{\frac{r-3}{2}} + \binom{n_1-1}{\frac{r-3}{2}}\binom{n_2-1}{\frac{r-1}{2}}}{\binom{n_1+n_2}{n_1}}, & r \text{ ímpar} \end{cases}$$

para $r = 2, 3, \cdots, n$.

Distribuição nula de R assintótica

Embora a função de densidade de R possa ser utilizada para encontrar sua distribuição exata quaisquer sejam os valores de n_1 e n_2, os cálculos são trabalhosos, a menos que n_1 e n_2 sejam pequenos. Para amostras grandes, uma aproximação para a distribuição nula pode ser usada o que dá resultados razoavelmente bons, desde que n_1 e n_2 sejam maiores que 10.

A fim de encontrar a distribuição assintótica, assumimos que o tamanho total da amostra n tende ao infinito de tal maneira que $n_1/n \to \lambda$ e $n_2/n \to 1-\lambda$, λ fixo e $0 < \lambda < 1$. Para grandes amostras, a média e a variância de R são

$$\lim_{n\to\infty} \mathrm{E}\left(\frac{R}{n}\right) = 2\lambda(1-\lambda) \quad \text{e} \quad \lim_{n\to\infty} \mathrm{Var}\left(\frac{R}{\sqrt{n}}\right) = 4\lambda^2(1-\lambda)^2.$$

Padronizando a variável aleatória obtemos

$$Z = \frac{R - 2n\lambda(1-\lambda)}{2\sqrt{n}\lambda(1-\lambda)}$$

e substituindo R em termos de Z, obtemos a distribuição de R ou $f_Z(z)$. Se os fatores na expressão resultante são avaliados pela fórmula de Stirling, o limite é

$$\lim_{n\to\infty} \ln\left(f_Z(z)\right) = -\ln\left(\sqrt{2\pi}\right) - \frac{1}{2}z^2,$$

o que mostra que a função de probabilidade limite de Z é a densidade normal padrão. Este resultado foi obtido por Wald and Wolfowitz (1940).

Para um teste bilateral de tamanho α usando a aproximação normal, a hipótese nula de aleatoriedade seria rejeitada quando

$$\left| \frac{R - 2n\lambda(1-\lambda)}{2\sqrt{n}\lambda(1-\lambda)} \right| \leq z_{\alpha/2},$$

onde z_γ é um número satisfazendo que $\Phi(z_\gamma) = 1-\gamma$ ou, equivalentemente, z_γ é o $(1-\gamma)$-ésimo quantil da distribuição normal padrão.

A média e variância exatas de R foram obtidas e também podem ser usadas na formação da variável aleatória padronizada, já que a distribuição assintótica permanece inalterada. Estes resultados estão disponíveis na função **runs.test** no pacote **tseries** e na função **runs.test** no pacote **randtests**.

Exemplo: Dados simulados dos quais nos interessa somente o sinal.

```
> x = rnorm(100)
> xx = factor(sign(x))
> xx
  [1] -1 -1 1 -1 1 -1 -1 1 -1 1 -1 -1 1 -1 1 -1 1 -1 -1 -1 -1 1 -1 1 -1
 [30] -1 -1 1 1 1 1 1 1 -1 -1 1 1 -1 -1 -1 1 -1 -1 1 1 1 1 1 -1 1 1 -1
 [59] -1 1 -1 -1 1 1 -1 1 1 1 1 1 1 1 -1 -1 1 -1 1 -1 1 -1 -1 1 1 -1 -1
 [88] 1 -1 -1 -1 -1 -1 1 -1 1 -1 -1 1 1
Levels: -1 1
> tseries::runs.test(xx)

Runs Test

data:  xx
Standard Normal = 0.64157, p-value = 0.5212
alternative hypothesis: two.sided
```

Teste baseado nos postos

Outra maneira de testar a aleatoriedade é comparando a magnitude de cada elemento com a do elemento imediatamente anterior na sequência e computando assim a soma dos quadrados das desvalorizações dos pares de elementos sucessivos.

Se as magnitudes desses elementos forem substituídas por suas respectivas classificações na sequência antes de computar a soma dos quadrados dos desvios sucessivos, podemos obter um teste não paramétrico. Especificamente, seja a sequência de observações ordenada pelo tempo $X_1, X_2, \cdots, X_n$. A estatística de teste

$$NM = \sum_{i=1}^{n-1} \left(\text{posto}(X_i) - \text{posto}(X_{i+1})\right)^2$$

como proposto por Bartels (1982).

Um teste baseado em uma função desta estatística é a versão do teste de razão para aleatoriedade desenvolvido por von Neumann usando a teoria normal e é uma transformação linear do coeficiente de correlação serial por postos introduzido por Wald e Wolfowits (1943).

Pode-se mostrar que a estatística de teste *NM* varia entre $(n-1)$ e $(n-1)(n^2+n-3)/3$ se n é par e entre $(n-1)$ e $[(n-1)(n^2+n-3)/3]-1$ se n for ímpar. A distribuição nula exata do *NM* pode ser encontrada por enumeração e é dada em Bartels (1982) para $4 \leq n \leq 10$ sendo disponibilizada na função **bartels.rank.test** (Bartels Rank Test) no pacote **randtests**.

Exemplo

Exemplo em Bartels (1982). Mudanças nos níveis de estoque de 1968-1969 para 1977-1978 (em milhões) deflacionado pelo índice de preços do Produto Interno Bruto (PIB) australiano (base 1966-1967).

```
> library(randtests)
> x.dados = c(528, 348, 264, -20, -167, 575, 410, -4, 430, -122)
> bartels.rank.test(x, pvalue = "normal")

Bartels Ratio Test

data:  x.dados
statistic = 0.083357, n = 10, p-value = 0.9336
alternative hypothesis: nonrandomness
```