O TESTE DOS POSTOS ORDENADOS DE GALTON: UMA ABORDAGEM GEOMÉTRICA

Paulo César de Resende ANDRADE[1]
Lucas Monteiro CHAVES[2]
Devanil Jaques de SOUZA[2]

▪ RESUMO: Este trabalho apresenta a teoria do teste de Galton do ponto de vista geométrico. Tal abordagem permite tratar a combinatória subjacente ao teste como um problema em passeios aleatórios unidimensionais. Com isto, a contagem de caminhos em um reticulado torna-se ferramenta suficiente para a obtenção da distribuição da estatística do teste.

▪ PALAVRAS-CHAVE: Princípio da reflexão; contagem de caminhos; distribuição uniforme.

1 Introdução

Considere duas amostras independentes $X = (x_1, x_2, ..., x_n)$ e $Y = (y_1, y_2, ..., y_n)$ de duas populações, e suas respectivas ordenações decrescentes $a_1 > a_2 > ... > a_n$ e $b_1 > b_2 > ... > b_n$, em que a_i representa o *i-ésimo* maior elemento da amostra X e b_j representa o *j-ésimo* maior elemento da amostra Y. A estatística do teste de Galton consiste da contagem do número de pares (a_k, b_k) tais que $a_k > b_k$. Denominaremos tais pares de ***pares efetivos***. Suponha-se, por exemplo, que uma amostra X seja obtida de uma população, após um tratamento, e Y outra amostra da mesma população, antes do tratamento. Supondo, sem perda de generalidade, que o tratamento aumente a média populacional de forma considerável, espera-se que o número pares efetivos seja próximo de *n*. No caso do tratamento ser inócuo, esse número deve ser próximo de $n/2$.

Essa ideia foi usada por Galton, em 1876, em dados obtidos por Charles Darwin, segundo Hodges (1955). O mesmo Hodges (1955) cita ainda que o teste de Galton foi extensivamente revisto por Fisher (1945).

Apesar da simplicidade desse teste, a combinatória a ele subjacente é complexa e com resultados surpreendentes. Neste artigo será desenvolvida toda a teoria combinatória

[1] Universidade Federal dos Vales do Jequitinhonha e Mucuri – UFVJM, Instituto de Ciência e Tecnologia – Campus II, Rodovia MGT 367, km 583, nº 5000, Alto da Jacuba, CEP: 39100-000, Diamantina, MG, Brasil. E-mail: *paulo.andrade@ict.ufvjm.edu.br*
[2] Universidade Federal de Lavras – UFLA, Departamento de Ciências Exatas, Caixa Postal 37, CEP: 37200-000, Lavras, MG, Brasil. E-mail: *lucas@dex.ufla.br / DevanilJaques@dex.ufla.br*

desse teste, de um ponto de vista totalmente geométrico. Uma abordagem geométrica para o estudo de problemas relacionados a passeios aleatórios simplificou de tal forma a combinatória subjacente que constituiu a principal motivação para a segunda edição do clássico livro de Feller, "An Introduction to Probability Theory and Its Applications", segundo o próprio Feller (1968), pág. 676, nota de rodapé. No entanto, como visa abordar de maneira geral os passeios aleatórios, Feller (1968) não explora totalmente as possibilidades de abordar geometricamente o teste de Galton. Neste trabalho, o método geométrico é desenvolvido em detalhes ausentes no livro e, como a leitura do texto de Feller (1968) é difícil, os autores acreditam que este pode ser uma referência didática e completa da teoria subjacente ao teste de Galton. A abordagem geométrica de problemas estatísticos, em uma perspectiva geral, pode ser encontrada em Saville (1991).

2 Teste de Galton

Seja $W = (w_1, w_2, ..., w_{2n})$ a amostra conjunta, ordenada de maneira decrescente, obtida pela união das amostras X e Y. A sequência $(w_1, w_2, ..., w_{2n})$ é transformada em uma sequência de $(+1)'s$ e $(-1)'s$ da seguinte forma: se $w_i \in X$ é substituído por $+1$ e, se $w_i \in Y$, por -1. O resultado é uma dentre todas as sequências possíveis, de tamanho $2n$, compostas com $n(+1)'s$ e $n(-1)'s$. Como na maioria das abordagens não paramétricas, os valores numéricos da amostra são irrelevantes, sendo consideradas apenas as ordenações desses valores. Nesse sentido, qualquer sequência de tamanho $2n$, composta com n elementos iguais a $+1$ e n elementos iguais a -1, corresponde a uma amostra ordenada W. Para calcular o número dessas sequências basta supor uma fila de $2n$ caixas vazias e escolher n dessas caixas para colocar os $(+1)'s$, ficando assim determinadas as posições dos $(-1)'s$. É trivial que o número de sequências possíveis é dado por $\binom{2n}{n} = \frac{(2n)!}{n!n!}$.

Define-se então uma variável aleatória discreta G, que conta, em cada sequência, o número pares efetivos. Sendo assim, G assume valores no conjunto $\{0, 1, ..., n\}$ e tem função de probabilidade, supondo que todas as sequências possíveis são equiprováveis, dada por:

$$P[G = k] = \frac{\textit{número de seqüências com } k \textit{ pares efetivos}}{\binom{2n}{n}} \quad (k = 0, ..., n). \tag{1}$$

A questão que demanda uma teoria combinatória complicada é o cálculo do número de todas as sequências com um determinado número de pares efetivos. Por exemplo, seja a sequência de tamanho 10, $(-1, +1, +1, +1, -1, +1, -1, -1, +1, -1)$. Observe que: o primeiro -1 precede o primeiro $+1$, significando que $b_1 > a_1$; para todo $i \in \{2, 3, 4, 5\}$, o *i-ésimo* $+1$ precede o *i-ésimo* -1, o que significa que $a_i > b_i$, e, portanto, $G = 4$ como pode ser visto na Figura 1, caso se recomponha a sequência original:

$$\begin{array}{cccccccccc} -1 & +1 & +1 & +1 & -1 & +1 & -1 & -1 & +1 & -1 \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ b_1 & a_1 & a_2 & a_3 & b_2 & a_4 & b_3 & b_4 & a_5 & b_5 \end{array}$$

Figura 1 - Ordenação original para a sequência (-1,+1,+1,+1,-1,+1,-1,-1,+1,-1).

Como caracterizar todas as outras sequências para as quais $G = 4$? A teoria combinatória que responde a essa pergunta faz uso da seguinte construção:

Denotando por $u = (u_1, u_2, ..., u_{2n})$ uma sequência de $(+1)'s$ e $(-1)'s$, define-se a soma parcial $S_j = \sum_{i=1}^{j} u_i \quad (j = 1, ..., 2n)$. Observe que:

(i) Se $S_{2k-1} > 0$, então, dos $2k-1$ primeiros elementos de u, pelo menos k são iguais a $+1$ e no máximo $k-1$ são iguais a -1. Significa que, nos $2k-1$ primeiros elementos de u, está presente o $+1$ correspondente a a_k e não está o -1 correspondente ao b_k, ou seja, $a_k > b_k$. Como vale a inversa, se $a_k > b_k$ então $S_{2k-1} > 0$, segue que $a_k > b_k$ se, e somente se, $S_{2k-1} > 0$.

(ii) É trivial que se $S_{2k-1} > 0$ então $S_{2k} \geq 0$.

(iii) A desigualdade $a_k > b_k$ ocorre i vezes se, e somente se, existem i pares $(2k-1, 2k)$ tais que $S_{2k-1} > 0$ e, portanto, $S_{2k} \geq 0$. No exemplo anterior tem-se $a_k > b_k$ nos pares (3, 4), (5, 6), (7, 8) e (9, 10).

Esse procedimento para contar o número de sequências com determinada quantidade de pares efetivos é técnico e trabalhoso. Feller (1968) propôs uma abordagem geométrica para o problema, que é, ao mesmo tempo, didática e possibilita um tratamento unificado a vários problemas combinatórios semelhantes.

3 Abordagem geométrica

A ideia é: Considerem, no plano cartesiano, os pontos de coordenadas inteiras e abscissas não negativas. Partindo do ponto $P_0 = (0,0)$, identifica-se uma sequência $u = (u_1, u_2, ..., u_{2n})$ com um caminho no plano, dado pelos pontos $P_0 \to P_1 \to ... \to P_{2n}$, em que, $P_1 = (1,1)$ caso $u_1 = 1$, $P_1 = (1,-1)$ caso $u_1 = -1$ e, $P_j = P_{j-1} + (1, u_j)$, de modo que, o conjunto de todos os caminhos possíveis, compõe um reticulado quadrado, com uma das diagonais no eixo das abscissas com vértices em $(0,0)$, (n,n), $(n,-n)$ e $(2n,0)$. A Figura 2 mostra o caminho correspondente à sequência $(-1,+1,+1,+1,-1,+1,-1,-1,+1,-1)$.

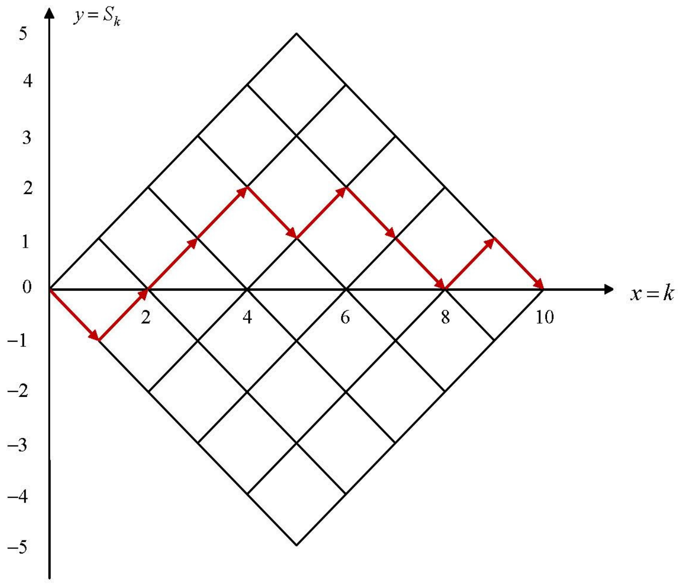

Figura 2 - Caminho dado pela sequência $(-1,+1,+1,+1,-1,+1,-1,-1,+1,-1)$.

Note que:

(i) As coordenadas de um ponto sobre o caminho são (k, S_k) e um caminho só pode tocar o eixo Ox nas abscissas pares, uma vez que $S_j = 0$ implica em j ser par.

(ii) Se $S_{2k-1} > 0$ então o segmento de reta entre os pontos coordenados $(2k-1, S_{2k-1})$ e $(2k, S_{2k})$ está, necessariamente, acima do eixo Ox.

(iii) Se um caminho toca o eixo Ox em k_1 e k_2 e fica estritamente acima de Ox no intervalo (k_1, k_2) então k_1 e k_2 são números pares, o número de segmentos no intervalo é a quantidade par $k_2 - k_1$ e o número de pares $(2t-1, 2t)$, com $\frac{k_1+2}{2} \le t \le \frac{k_2}{2}$, tais que $S_{2t-1} > 0$, ou seja, o número de pares efetivos, é $\frac{k_2 - k_1}{2}$.

Conclusão: Uma sequência tem j pares efetivos se, e somente se, seu caminho correspondente tem $2j$ segmentos acima do eixo Ox.

Na contagem dos caminhos tem-se o seguinte fato fundamental:

Teorema: O Princípio da Reflexão (Feller, 1968).

O número de caminhos de A até B que tocam ou cruzam o eixo Ox é igual ao número de caminhos de A' até B, em que A' é a imagem refletida de A em relação ao eixo Ox.

Demonstração:

Seja α um caminho de A até B e seja k_0 o primeiro ponto do eixo Ox pertencente a α. Refletindo a parte do caminho α de A até k_0 em relação ao eixo Ox obtém-se um caminho α' de A' até B. Após k_0, α e α' são coincidentes. Da correspondência biunívoca $\alpha \leftrightarrow \alpha'$ o resultado segue.

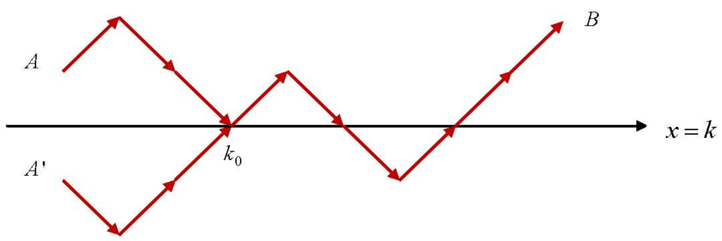

Figura 3 - Princípio da reflexão.

Um outro resultado útil na contagem de caminhos é:

Lema 1: O número de caminhos de um ponto (c,d) até outro ponto (e,f), com $e > c$, é dado por:

$$\binom{e-c}{\frac{e-c+f-d}{2}}.$$

Demonstração: Como S_j é a soma dos j primeiros elementos do caminho, segue que $S_c = d$ e $S_e = f$.

Sejam: p_c, o número de $(+1)'s$ do caminho, até o *c-ésimo* termo;

q_c, o número de $(-1)'s$ do caminho, até o *c-ésimo* termo;

p_e, o número de $(+1)'s$ do caminho, até o *e-ésimo* termo;

q_e, o número de $(-1)'s$ do caminho, até o *e-ésimo* termo.

Segue que $p_e - p_c$ é o número de $(+1)'s$ no caminho entre o $(c+1)-$*ésimo* termo e o *e-ésimo* termo. $q_e - q_c$ é o número de $(-1)'s$ no mesmo intervalo.

Um caminho de (c,d) até (e,f) fica determinado quando se distribuem $p_e - p_c$ $(+1)'s$ em $e-c$ posições, o que pode ser feito de $\binom{e-c}{p_e - p_c}$ maneiras.

Tem-se que:

$$\begin{aligned} p_c + q_c &= c \\ p_e + q_e &= e \\ p_c - q_c &= d \\ p_e - q_e &= f \end{aligned} \quad \Rightarrow \quad \begin{aligned} p_c &= \frac{c+d}{2} \\ p_e &= \frac{e+f}{2} \\ q_c &= \frac{c-d}{2} \\ q_e &= \frac{e-f}{2} \end{aligned}$$

Logo,

$$\binom{e-c}{p_e - p_c} = \binom{e-c}{\frac{e+f}{2} - \frac{c+d}{2}} = \binom{e-c}{\frac{e-c+f-d}{2}}.$$

Observe que, o número de caminhos de $(0,0)$ até (k,s) é $\binom{k}{\frac{k+s}{2}}$ e, de $(0,0)$ a $(2n,0)$ é $\binom{2n}{n}$.

Teorema: Feller (1968), pág. 94:

O número de caminhos de $(0,0)$ até $(2n,0)$ com exatamente $2k$ segmentos acima do eixo Ox é independente de k e é igual a

$$\frac{1}{n+1}\binom{2n}{n}.$$

Demonstração:

(i) Caso $k = n$, basta contar os caminhos de $(0,0)$ até $(2n,0)$ que estejam acima do eixo Ox, isto é, caminhos que não tocam a reta $y = -1$. A Figura 4 mostra um desses caminhos.

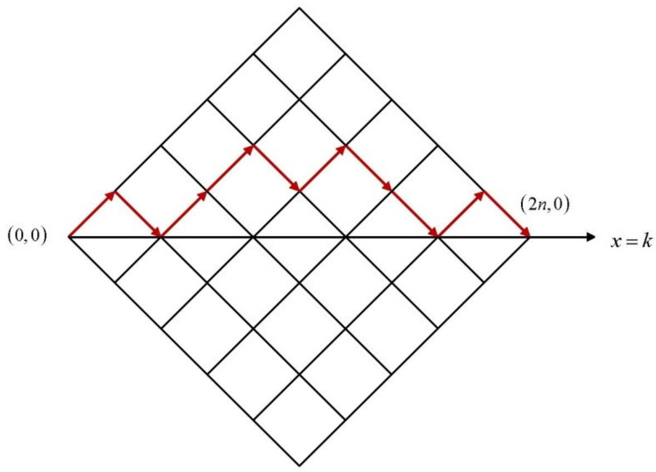

Figura 4 - Caminho acima do eixo $Ox.$

Para contar esses caminhos, faz-se uso do princípio da reflexão. Prolonga-se para a esquerda e para baixo o reticulado de forma que o ponto $(0,0)$, da Figura 4, torne-se o ponto $(1,1)$, conforme Figura 5, e que a reta original $y=-1$ torne-se a reta $y=0.$

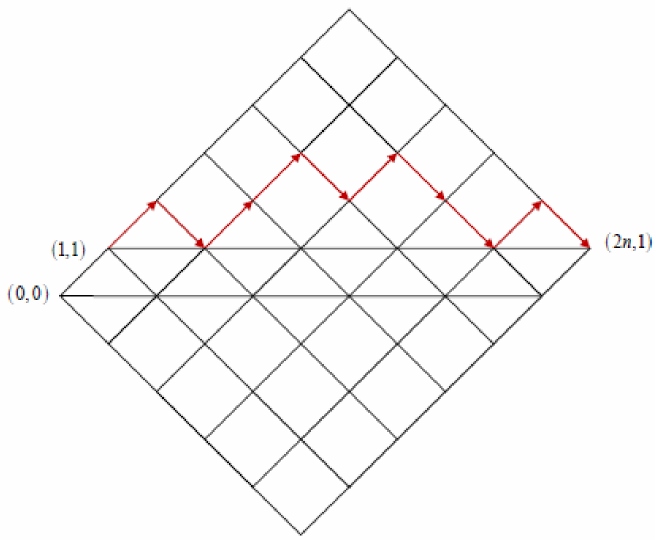

Figura 5 - Prolongamento à esquerda do reticulado.

Com isto, um caminho acima do eixo na Figura 4, ligando $(0,0)$ a $(2n,0)$, torna-se um caminho ligando $(1,1)$ a $(2n+1,1)$ e que passa necessariamente pelo ponto $(2,2)$. Queremos, então, contar os caminhos de $(2,2)$ até $(2n+1,1)$, que estejam, estritamente, acima do novo eixo Ox. Esse número pode ser obtido tomando-se todos os caminhos de $(2,2)$ a $(2n+1,1)$ e subtraindo-se os caminhos de $(2,2)$ a $(2n+1,1)$ que tocam ou cruzam o eixo Ox que, pelo princípio da reflexão, é igual ao número de caminhos de $(2,-2)$ a $(2n+1,1)$. Portanto, pelo Lema 1, tem-se:

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

(ii) O caso $k=0$ corresponde aos caminhos que ficam abaixo do eixo Ox e, por simetria, o número desses caminhos é igual ao caso (i), ou seja, $\frac{1}{n+1}\binom{2n}{n}$.

(iii) O caso $0<k<n$ será provado por indução em n. Para $n=1$ o único caminho possível é $(0,0)\to(1,1)\to(2,0)$, isto é,

$$\frac{1}{n+1}\binom{2n}{n}=\frac{1}{1+1}\binom{2}{1}=1.$$

Suponha o resultado válido para quando se consideram caminhos de comprimento menor que $2n$. Denote por $2r$ a coordenada em que, pela primeira vez, um caminho qualquer, com exatamente $2k$ segmentos acima do eixo Ox, intercepta este eixo. Há duas possibilidades:

$1^a)$ No intervalo $[0,2r]$ o caminho é estritamente positivo, isto é, possui $2r$ segmentos acima do eixo. Neste caso, como o caminho tem $2k$ segmentos positivos, no intervalo $[2r,2n]$, ocorrem os $2k-2r$ segmentos positivos restantes. Pela hipótese de indução, o número de caminhos de $(2r,0)$ até $(2n,0)$, com $2k-2r$ segmentos acima do eixo Ox, é dado por:

$$\frac{1}{n-r+1}\binom{2n-2r}{n-r}\quad(1\le r\le k).$$

É preciso, agora, calcular o número de caminhos estritamente positivos, de $(0,0)$ até $(2r,0)$.

Lema 2: O número de caminhos estritamente positivos, de $(0,0)$ até $(2r,0)$, é dado por:

$$\frac{1}{r}\binom{2r-2}{r-1}.$$

Demonstração: Tais caminhos passam, necessariamente, pelos pontos $(1,1)$ e $(2r-1,1)$. Usando o princípio da reflexão, o número de caminhos estritamente positivos entre $(1,1)$ e $(2r-1,1)$ é dado pelo total de caminhos entre $(1,1)$ e $(2r-1,1)$ menos o total de caminhos entre $(1,-1)$ e $(2r-1,1)$, ou seja,

$$\binom{2r-1-1}{\frac{2r-1-1+1-1}{2}}-\binom{2r-1-1}{\frac{2r-1+1-(-1)}{2}}=\binom{2r-2}{r-1}-\binom{2r-2}{r}$$

$$=\frac{1}{2(2r-1)}\binom{2r}{r}=\frac{1}{r}\binom{2r-2}{r-1}.$$

Segue, então, que o número de caminhos com $2k$ lados acima do eixo Ox de $(0,0)$ até $(2n,0)$, e que são caminhos estritamente positivos de $(0,0)$ até $(2r,0)$, como na Figura 6, é dado pelo produto:

$$\frac{1}{n-r+1}\binom{2n-2r}{n-r}\frac{1}{r}\binom{2r-2}{r-1} \quad (1\leq r\leq k).$$

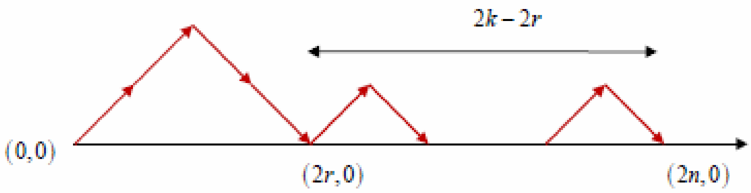

Figura 6 - Caminhos estritamente positivos de $(0,0)$ até $(2r,0)$.

$2^{a})$ No intervalo $[0,2r]$ o caminho é estritamente negativo, isto é, possui $2r$ segmentos abaixo do eixo Ox (Figura 7), isto é, que só tocam o eixo Ox nos extremos.

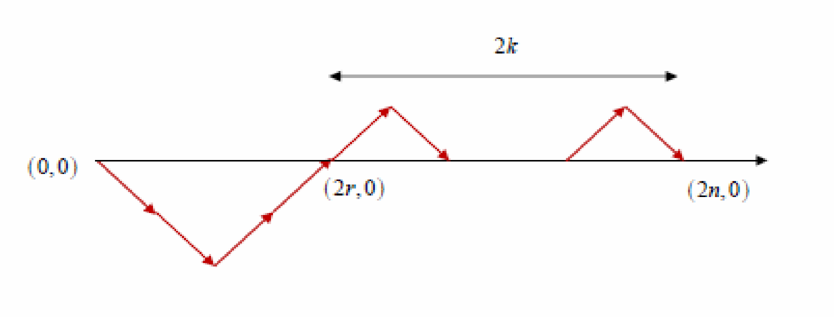

Figura 7 - Caminhos estritamente negativos de $(0,0)$ até $(2r,0)$.

Neste caso, a ocorrência dos $2k$ segmentos acima do eixo se dá no intervalo $[2r,2n]$, o que implica em $n-r \geq k$. Pela hipótese de indução, o número de caminhos de $(2r,0)$ até $(2n,0)$, com $2k$ segmentos acima do eixo é:

$$\frac{1}{n-r+1}\binom{2n-2r}{n-r} \quad (n-k>r).$$

Portanto, o número total desses caminhos é obtido pelo produto:

$$\frac{1}{n-r+1}\binom{2n-2r}{n-r}\frac{1}{r}\binom{2r-2}{r-1}.$$

Observe, então, que as expressões para o primeiro caso e o segundo caso são idênticas. Sendo assim, o número total de caminhos de $(0,0)$ até $(2n,0)$, com $2k$ lados acima do eixo Ox, com $0<k<n$ é obtido pela seguinte soma em r:

$$\sum_{r=1}^{k}\frac{1}{r(n-r+1)}\binom{2n-2r}{n-r}\binom{2r-2}{r-1}+\sum_{r=1}^{n-k}\frac{1}{r(n-r+1)}\binom{2n-2r}{n-r}\binom{2r-2}{r-1}.$$

Fazendo $\rho=n-r+1 \Rightarrow r=n+1-\rho$, $1\leq r\leq n-k \Rightarrow k+1\leq \rho\leq n$, tem-se:

$$\sum_{r=1}^{n-k}\frac{1}{r(n-r+1)}\binom{2n-2r}{n-r}\binom{2r-2}{r-1}=$$

$$=\sum_{\rho=k+1}^{n}\frac{1}{\rho(n-\rho+1)}\binom{2(n+1-\rho)-2}{n+1-\rho-1}\binom{2n-2(n+1-\rho)}{n-(n+1-\rho}$$

$$=\sum_{\rho=k+1}^{n}\frac{1}{\rho(n-\rho+1)}\binom{2n-2\rho}{n-\rho}\binom{2\rho-2}{\rho-1}.$$

Sendo assim,

$$\sum_{r=1}^{k}\frac{1}{r(n-r+1)}\binom{2n-2r}{n-r}\binom{2r-2}{r-1}+\sum_{r=1}^{n-k}\frac{1}{r(n-r+1)}\binom{2n-2r}{n-r}\binom{2r-2}{r-1}=$$

$$=\sum_{r=1}^{k}\frac{1}{r(n-r+1)}\binom{2n-2r}{n-r}\binom{2r-2}{r-1}+\sum_{\rho=k+1}^{n}\frac{1}{\rho(n-\rho+1)}\binom{2n-2\rho}{n-\rho}\binom{2\rho-2}{\rho-1}$$

$$=\sum_{r=1}^{n}\frac{1}{r(n-r+1)}\binom{2n-2r}{n-r}\binom{2r-2}{r-1}.$$

Como o somatório anterior não depende de k, seu valor pode ser calculado da seguinte forma: o número total de caminhos de $(0,0)$ até $(2n,0)$ menos o número de caminhos acima do eixo Ox, com $k=n$ e $k=0$, casos (i) e (ii) dividido por $n-1$, que é o número de valores que k assume, de 1 até $n-1$, isto é:

$$\frac{1}{n-1}\left[\binom{2n}{n}-\frac{1}{n+1}\binom{2n}{n}-\frac{1}{n+1}\binom{2n}{n}\right]=\frac{1}{n-1}\binom{2n}{n}\left[1-\frac{2}{n+1}\right]$$

$$=\frac{1}{n-1}\binom{2n}{n}\left[\frac{n-1}{n+1}\right]=\frac{1}{n+1}\binom{2n}{n}.$$

Sendo assim,

$$P[G=k]=\frac{\textit{número de seqüências com } k \textit{ pares efetivos}}{\binom{2n}{n}}$$

$$=\frac{\frac{1}{n+1}\binom{2n}{n}}{\binom{2n}{n}}=\frac{1}{n+1}$$, o que encerra a prova do teorema.

Tem-se então um resultado surpreendentemente não intuitivo que enganou até a Galton: A distribuição da estatística do teste, que conta o número de pares efetivos da amostra ordenada conjunta, é uniforme.

Considerações finais

Retornando ao caso dos dados de Darwin submetidos à apreciação de Galton, os dados eram constituídos de duas amostras de mesmo tamanho, de duas populações e a pergunta era se essas populações tinham a mesma média. As amostras tinham tamanhos iguais a 15 e número de pares efetivos igual a 13. Galton concluiu que as médias eram diferentes. No entanto, caso as médias fossem iguais, a distribuição do número de pares efetivos seria uniforme e, a probabilidade de se ter 13, 14 ou 15 pares efetivos, puramente devidos ao acaso, seria de $\frac{3}{16} = 18,75\%$, o que torna, no mínimo temerária, a conclusão de Galton.

Um estudo comparativo entre os poderes dos testes de postos com sinal, Mann-Whitney, Galton e Teste t, encontrado em Andrade *et al.* (2003), concluíram que, para pequenas diferenças entre as médias populacionais, o teste de Galton se mostrou superior aos demais.

ANDRADE, P. C. R.; CHAVES, L. M.; SOUZA, D. J. A geometric approach to Galton's ordered ranks test. *Rev. Bras. Biom.*, São Paulo, v.30, n.1, p.124-135, 2012.

- *ABSTRACT: A complete description of the theory of Galton's test is presented. The approach is geometric, in the sense of putting the underlying combinatorics as one-dimensional random walk problem, using counting paths in a grid.*

- *KEYWORDS: Reflection principle; counting paths; uniform distribution.*

Referências

ANDRADE, P. C. R.; CHAVES, L. M.; FERREIRA, D. F., Proposta de um teste não-paramétrico de sinal com postos para dados independentes de duas populações. *Rev. Mat. Estat.*, São Paulo, v.21, n.2, p.7-23, 2003.

HODGES JR., J. L. Galton's rank-order test. *Biometrika*, London, v.42, p.261-262, 1955.

FELLER, W., *An introduction to probability theory and its applications*. New York: John Wiley, 1968. v.1, 460p.

FISHER, R. A. *The design of experiments*. 4.ed. Edinburgh: Oliver and Boyd, 1945.

SAVILLE, D. J.; WOOD, G. R. *Statistical methods*: the geometric approach. New York: Springer, 1991. 560p.

Recebido em 24.02.2012

Aprovado após revisão 24.05.2012