Estatística não paramétrica
Medidas de associação em classificações múltiplas

Fernando Lucambio

Departamento de Estatística
Universidade Federal do Paraná

Abril de 2025

Suponha temos um conjunto de dados completo com I linhas e J colunas, com uma entrada em cada uma das $I \times J$ células. Sob a hipótese nula de populações idênticas, os dados podem ser considerados como uma única amostra aleatória de tamanho IJ da população comum.

O paralelo a este problema na estatística clássica é a Análise Variância. Vamos estudar alguns procedimentos análogos não paramétricos à Análise de Variância tudo paralelo, no sentido de que os dados são apresentados na mesma forma.

Vamos primeiro revisar as técnicas da Análise de Variância, abordagem para testar a hipótese nula de que os efeitos são todos o mesmo. O modelo é geralmente escrito

$$X_{ij} = \mu + \beta_i + \theta_j + \epsilon_{ij}, \quad \text{para } i = 1, 2, \cdots, I \text{ e } j = 1, 2, \cdots, J.$$

Os termos β_i e θ_j são os efeitos por fila e columa, respectivamente. No modelo teórico, os erros ϵ_{ij} são variáveis aleatórias independentes, normalmente distribuídas com média zero e variância σ^2_ϵ. A estatística de teste para a hipótese nula de efeitos de coluna iguais ou, equivalentemente,

$$H_0 : \theta_1 = \theta_2 = \cdots = \theta_J,$$

é a relação

$$\frac{(I-1)\sum_{j=1}^{J}(\overline{X}_j - \overline{X})^2}{\sum_{i=1}^{I}\sum_{j=1}^{J}(X_{ij} - \overline{X}_I - \overline{X}_j + \overline{X})^2},$$

onde

$$\overline{X}_i = \frac{1}{J}\sum_{j=1}^{J} X_{ij}, \qquad \overline{X}_j = \frac{1}{I}\sum_{i=1}^{I} X_{ij} \qquad \text{e} \qquad \overline{X} = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} X_{ij}.$$

Se todas as suposições do modelo forem atendidas, esta estatística de teste tem distribuição F-Fisher com graus de liberdade $J-1$ e $(I-1)(J-1)$.

Os dois primeiros paralelos deste desenho que consideraremos são os problemas de k-amostras relacionadas e o de k-amostras combinadas. A correspondência pode surgir de duas maneiras diferentes, mas ambas são de certa forma análogas aos modelos de blocos randomizados.

O esquema de agrupamento em blocos é importante, uma vez que o propósito de tal projeto é minimizar as diferenças entre as unidades no mesmo bloco. Dividindo o campo em I blocos, os gráficos dentro de cada bloco podem ser mantidos próximos. Quaisquer diferenças entre parcelas dentro do mesmo bloco podem ser atribuídas a diferenças entre os tratamentos e o efeito de bloco pode ser eliminado da estimativa do erro experimental.

Sob a hipótese de populações idênticas, temos uma única amostra aleatória de tamanho $\sum_{i=1}^{k} n_i = N$ da população comum. A mediana geral δ das amostras agrupadas é uma estimativa da mediana dessa população comum. Portanto, uma observação de qualquer uma das *k* amostras é tão provável que seja acima de δ como abaixo dela.

O conjunto de *N* observações apoiará a hipótese nula se, para cada uma das *k* amostras, cerca de metade das observações nessa amostra forem inferiores à mediana da grande amostra. Um teste baseado neste critério é atribuído a Mood (1950) e Brown Mood (1948, 1951).

Como no caso de duas amostras, a mediana geral δ será definida como a observação de classificação $(N+1)/2$ na amostra ordenada agrupada se N for ímpar e qualquer número entre as duas observações com postos $N/2$ e $(N+2)/2$ se N for par.

Então, para cada amostra separadamente, as observações são dicotomizadas de acordo como são menores que δ ou não. Defina a variável aleatória U_i como o número de observações na amostra i que são menores que δ e seja t o número total de observações que são menores que δ.

Então, pela definição de δ, temos

$$t = \sum_{i=1}^{k} u_i = \begin{cases} N/2 & \text{caso } N \text{ seja par} \\ (N-1)/2 & \text{caso } N \text{ seja ímpar} \end{cases}.$$

	Amostra 1	**Amostra 2**	$\cdots$	**Amostra k**	**Total**
$< \delta$	u_1	u_2	$\cdots$	u_k	t
$\geq \delta$	$n_1 - u_1$	$n_2 - u_2$	$\cdots$	$n_k - u_k$	$N - t$
Total	n_1	n_2	$\cdots$	n_k	N

Sob a hipótese nula, cada um dos $\binom{N}{t}$ possíveis conjuntos de t observações tem a mesma probabilidade de estar na categoria menor que δ e o número de dicotomizações com este resultado amostral particular é $\prod_{i=1}^{k} \binom{n_i}{u_i}$. Portanto, a distribuição de probabilidade nula das variáveis aleatórias é a extensão multivariada da distribuição hipergeométrica ou

$$f(u_1, u_2, \cdots, u_k \,|\, t) = \binom{n_1}{u_1}\binom{n_2}{u_2}\cdots\binom{n_k}{u_k} \Big/ \binom{N}{t}.$$

Se algum ou todos os U_i diferirem muito de seu valor esperado $n_i\theta$, onde θ denota a probabilidade de que uma observação da população comum seja menor que δ, a hipótese nula poderia ser rejeitada.

Geralmente, seria impreciso configurar regiões de rejeição de junção para as estatísticas de teste $U_1, \cdots, U_k$, devido à grande variedade de combinações dos tamanhos de amostra $n_1, \cdots, n_k$ e ao fato de que a hipótese alternativa é geralmente bilateral para $k > 2$, como no caso do teste F.

Felizmente, podemos usar outro critério de teste que, embora seja uma aproximação, é razoavelmente preciso, mesmo para N tão pequeno quanto 25, se cada amostra consistir em pelo menos cinco observações. Esta estatística de teste pode ser derivada apelando para a análise do teste de bondade de ajuste.

Cada um dos N elementos da amostra agrupada é classificado de acordo com dois critérios, o número da amostra e sua magnitude em relação a δ. Sejam estas $2k$ categorias denotas por (i, j), onde $i = 1, \cdots, k$ de acordo com número da amostra e $j = 1$ se a observação é menor do que δ e $j = 2$ caso contrário. Vamos denotar as frequências esperadas para a (i, j) categoria por f_{ij} e ϵ_{ij}, respectivamente.

Então

$$f_{i1} = u_i, \qquad f_{i2} = n_i - u_i, \qquad \text{para } i = 1, 2, \cdots, k.$$

e as frequências esperadas quando H_0 é verdadeira estimadas dos dados como

$$e_{i1} = \frac{n_i t}{N}, \qquad e_{i2} = \frac{n_i (N - t)}{N}, \qquad \text{para } i = 1, 2, \cdots, k.$$

O teste da bondade de ajuste para essas $2k$ categorias é então

$$
\begin{aligned}
Q &= \sum_{i=1}^{k}\sum_{j=1}^{2}\frac{(f_{ij}-e_{ij})^2}{e_{ij}} \\
&= \sum_{i=1}^{k}\frac{(u_i-n_it/N)^2}{n_it/N} + \sum_{i=1}^{k}\frac{\left(n_i-u_i-n_i(N-t)/N\right)^2}{n_i(N-t)/N} \\
&= N\sum_{i=1}^{k}\frac{(u_i-n_it/N)^2}{n_it} + N\sum_{i=1}^{k}\frac{\left(n_i-u_i-n_i(N-t)/N\right)^2}{n_i(N-t)} \\
&= N\sum_{i=1}^{k}\frac{(u_i-n_it/N)^2}{n_i}\left(\frac{1}{t}+\frac{1}{N-t}\right) \\
&= \frac{N^2}{t(N-t)}\sum_{i=1}^{k}\frac{(u_i-n_it/N)^2}{n_i}
\end{aligned}
$$

Q tem aproximadamente distribuição do qui-quadrado sob H_0. Os parâmetros estimados a partir dos dados são as $2k$ probabilidades de que uma observação seja menor que δ para cada uma das k amostras e que não sejam menores que δ.

Mas, para cada amostra, essas probabilidades somam 1 e, portanto, há apenas k parâmetros independentes estimados. O número de graus de liberdade para Q é então $2k - 1 - k$ ou $k - 1$. A aproximação qui-quadrado para a distribuição de Q é um pouco melhorada pela multiplicação de Q pelo fator $(N-1)/N$. Então a região de rejeição é

$$Q \in \mathbb{R} \qquad \text{para} \qquad \frac{(N-1)Q}{N} \geq \chi^2_{k-1,\alpha}.$$

Assim como no teste de mediana de duas amostras, as observações empatadas não apresentam um problema. Sugere-se a abordagem conservadora, segundo a qual a decisão é baseada nessa resolução de vínculos que leva ao menor valor de Q.

Exemplo

Um estudo mostrou que 45 por cento dos que dormem normalmente roncam ocasionalmente, enquanto 25 por cento roncam quase o tempo todo.

Mais de 300 patentes foram registradas no Escritório de Patentes dos EUA para dispositivos que pretendiam parar de roncar. Três desses dispositivos são um squeaker costurado na parte de trás da roupa de dormir, um empate para segurar os pulsos para os lados da cama e uma cinta de queixo para manter a boca fechada. Um experimento foi realizado para determinar qual dispositivo é o mais eficaz em parar o ronco ou, pelo menos, reduzi-lo. Quinze homens que são roncadores habituais foram divididos aleatoriamente em três grupos para testar os dispositivos.

O sono de cada homem foi monitorado por uma noite por uma máquina que mede a quantidade de ronco em uma escala de 100 pontos enquanto usa um dispositivo. Analise os resultados mostrados abaixo para determinar se os três dispositivos são igualmente eficazes ou não.

Squeaker	Gravata de pulso	Cinta de queixo
73	96	12
79	92	26
86	89	33
91	95	8
35	76	78

A mediana geral da amostra é 78. Como $N = 15$ é ímpar, temos $t = 7$ e os dados são

Grupo	1	2	3
< 78	2	1	4
≥ 78	3	4	1

Calculamos $Q = 3.75$ e $(N1)Q/N = 3.50$.

```
> tabela = as.table( rbind( c(2, 1, 4), c(3, 4, 1)) )
> dimnames(tabela) = list( Mediana = c("Menor", "Maior"),
                           Patentes = c("Squeaker", "Gravata", "Cinta"))
> tabela
        Patentes
Mediana Squeaker Gravata Cinta
  Menor        2       1     4
  Maior        3       4     1
> (Xsq <- chisq.test(tabela))  # Mostra somente o sumário do teste

Pearson's Chi-squared test

data:  tabela
X-squared = 3.75, df = 2, p-value = 0.1534
```

```
> Xsq$observed   # contagens observadas
       Patentes
Mediana Squeaker Gravata Cinta
  Menor        2       1     4
  Maior        3       4     1
> Xsq$expected   # contagens esperados sob H0
       Patentes
Mediana Squeaker  Gravata    Cinta
  Menor 2.333333 2.333333 2.333333
  Maior 2.666667 2.666667 2.666667
> Xsq$residuals  # resíduos de Pearson
       Patentes
Mediana   Squeaker    Gravata      Cinta
  Menor -0.2182179 -0.8728716  1.0910895
  Maior  0.2041241  0.8164966 -1.0206207
> Xsq$stdres     # resíduos padronizados
       Patentes
Mediana   Squeaker    Gravata      Cinta
  Menor -0.3659625 -1.4638501  1.8298126
  Maior  0.3659625  1.4638501 -1.8298126
```

```
> chisq.test(tabela, simulate.p.value = TRUE, B = 10000)

Pearson's Chi-squared test with simulated p-value
                                    (based on 10000 replicates)

data:  tabela
X-squared = 3.75, df = NA, p-value = 0.2949
```

Não há evidências de que as medianas sejam diferentes.

Neste modelo a amostra será apresentada sob a forma de uma tabela com k linhas e n colunas. As linhas indicam números de bloco, assunto ou amostra e as colunas são efeitos do tratamento. As observações em diferentes linhas são independentes, mas as colunas não são por causa de alguma unidade de associação.

Para evitar fazer as suposições necessárias para o teste usual de análise de variância de que os n tratamentos são os mesmos, Friedman (1937, 1940) sugeriu substituir cada observação de tratamento dentro do i-bloco por um número do conjunto

$$\{1, 2, \cdots, n\}$$

que representa a magnitude do tratamento em relação às outras observações no mesmo bloco.

Denotamos as observações ranqueadas por R_{ij}, $i = 1, 2, \cdots, k$, $j = 1, 2, \cdots, n$ de modo que R_{ij} é o posto do j-ésimo efeito do tratamento no i-ésimo bloco. Então $R_{i1}, R_{i2}, \cdots, R_{in}$ é uma permutação dos primeiros n inteiros e $R_{1j}, R_{2j}, \cdots, R_{kj}$ o conjunto de postos dados ao j-ésimo efeito do tratamento em todos os blocos.

Representamos os dados em forma de tabela da seguinte forma:

	1	2	$\cdots$	n	Totais por linha
1	R_{11}	R_{12}	$\cdots$	R_{1n}	$n(n+1)/2$
2	R_{21}	R_{22}	$\cdots$	R_{2n}	$n(n+1)/2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	R_{k1}	R_{k2}	$\cdots$	R_{kn}	$n(n+1)/2$
Totais por coluna	R_1	R_2	$\cdots$	R_n	$kn(n+1)/2$

Os totais das linhas são constantes, mas os totais das colunas são afetados pelas diferenças entre os efeitos do tratamento. Se os efeitos do tratamento forem todos iguais, cada coluna esperada é igual e igual a média da coluna dos totais $k(n+1)/2$.

A soma dos desvios dos totais por coluna observados em torno dessa média é zero, mas a soma dos quadrados desses desvios será indicativo das diferenças nos efeitos do tratamento. Portanto, devemos considerar a distribuição amostral da variável aleatória

$$S = \sum_{j=1}^{n} \left(R_j - \frac{k(n+1)}{2} \right)^2 = \sum_{j=1}^{n} \left(\sum_{i=1}^{k} \left(R_{ij} - \frac{n+1}{2} \right) \right)^2$$

sob a hipótese nula de não haver diferença entre os n efeitos do tratamento.

Para este caso nulo, no i-ésimo bloco os postos são atribuídos completamente aleatórios e cada linha na tabela bidirecional constitui uma permutação aleatória dos primeiros n inteiros, se não houver empates.

Há então um total de $(n!)^k$ conjuntos de entradas distinguíveis na tabela $k \times n$ e cada um é igualmente provável. Essas possibilidades podem ser enumeradas e o valor de S calculado para cada um. A função de probabilidade de S é então

$$P(S = s) = \frac{u_s}{(n!)^k},$$

onde u_s é o número dessas atribuições que produzem s como a soma dos quadrados dos desvios totais da coluna.

Os cálculos são consideráveis.

Portanto, fora do intervalo de tabelas existentes, uma aproximação para a distribuição nula é geralmente usada para testes de significância. Seja $\mu = (n+1)/2$, então podemos escrever

$$\begin{aligned} S &= \sum_{j=1}^{n}\sum_{i=1}^{k}(R_{ij}-\mu)^2 + 2\sum_{j=1}^{n}\sum_{1\le i<p\le k}(R_{ij}-\mu)(R_{pj}-\mu) \\ &= k\sum_{j=1}^{n}(j-\mu)^2 + 2U = \frac{kn(n^2-1)}{12} + 2U. \end{aligned}$$

Os momentos de S então são determinados pelos momentos de U, que podem ser encontrados usando as seguintes relações

$$\mathrm{E}(R_{ij}) = \frac{n+1}{2},\ \mathrm{Var}(R_{ij}) = \frac{n^2-1}{12},\ \mathrm{Cov}(R_{ij}, R_{iq}) = -\frac{n+1}{2}.$$

Além disso, pelas suposições de modelo, as observações em diferentes linhas são independentes, de modo que, para todo $i \neq p$, o valor esperado de um produto de funções de R_{ij} e R_{pq} é o produto dos valores esperados e $\mathrm{Cov}(R_{ij}, R_{pq}) = 0$. Então

$$\mathrm{E}(U) = n\binom{k}{2}\mathrm{Cov}(R_{ij}, R_{pj}) = 0,$$

de maneira que $\mathrm{Var}(U) = \mathrm{E}(U^2)$.

Temos que

$$\mathrm{E}(S) =, \frac{kn(n^2-1)}{12}, \qquad \mathrm{Var}(S) = \frac{n^2k(k-1)(n-1)(n+1)^2}{72}.$$

Uma função linear das variáveis aleatórias, definida como

$$Q = \frac{12S}{kn(n+1)} = \frac{12\sum_{j=1}^{n} R_j^2}{kn(n+1)} - 3k(n+1),$$

tem momentos $\mathrm{E}(Q) = n-1$ e $\mathrm{Var}(Q) = 2(n-1)(k-1)/k \approx 2(n-1)$, os quais são os dois primeiros momentos da distribuição qui-quadrado com $n-1$ graus de liberdade. Os momentos mais altos de Q também são intimamente aproximados por momentos superiores correspondentes à distribuição qui-quadrado para k grande. Para todos os efeitos práticos, Q pode ser tratado como uma variável qui-quadrado com $n-1$ graus de liberdade.

Exemplo

Exemplo do livro Hollander and Wolfe (1973), p. 140. Comparação de três métodos: arredondamento (Round Out), ângulo estreito (Narrow Angle) e ângulo amplo (Wide Angle).

Para cada um dos 18 jogadores e os três métodos, o tempo médio de duas corridas a partir de um ponto na primeira linha de base, 35 pés da placa inicial até um ponto a 15 pés da segunda base.

```
> RoundingTimes <-
   matrix(c(5.40, 5.50, 5.55,
            5.85, 5.70, 5.75,
            5.20, 5.60, 5.50,
            ..............
                        5.70, 5.65, 5.55,
            6.30, 6.30, 6.25),
nrow = 22, byrow = TRUE,
dimnames = list(1:22, c("Round Out", "Narrow Angle", "Wide Angle")))
```

```
> friedman.test(RoundingTimes)

Friedman rank sum test

data:  RoundingTimes
Friedman chi-squared = 11.143, df = 2, p-value = 0.003805
```

Com este resultados obtemos forte evidência contra o afirmado na hipótesis nula, quer dizer, os métodos não são equivalentes em relação à velocidade.

III.3.3 Comparações Múltiplas Não Paramétricas e Intervalos de Confiança Simultâneos