PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO DE JANEIRO
Departamento de Engenharia Industrial
Rua Marquês de São Vicente, 225
22453-900 – Rio de Janeiro
Brasil

ENG1536 – Inferência Estatística

Laboratório – Guia de Estudo 8

A principal utilidade dos fatores que aprendemos no estudo anterior é investigar se o valor esperado de uma variável aleatória de interesse muda de um subgrupo a outro. Uma metodologia para essa investigação é a Análise de Variância (ANOVA). Nela, a variabilidade da variável de interesse ao redor da sua média é decomposta em termos associados aos fatores, e testamos formalmente a hipótese nula de que a variabilidade total não seja significativamente maior (no sentido estatístico) do que a variabilidade que resta depois de levado em conta o efeito de cada fator.

Para entender essa metodologia e como ela é implementada no R, usaremos dados simulados num estudo ANOVA de 1 fator. Suponha que uma pesquisa pedagógica tenha sido realizada com 500 alunos para testar o efeito de três diferentes metodologias de ensino sobre o aprendizado num curso. O aprendizado é medido pela nota final atingida pelos alunos no curso. Nos termos que você já aprendeu, trata-se de um estudo com um fator (a metodologia) que tem três níveis (metodologia A, metodologia B e metodologia C). Nos dados que vamos simular, a metodologia influenciará a nota do aluno:

```
metodologia <- factor(sample(c("mA","mB","mC"),500,replace=TRUE))
notas <- 0; for(i in 1:500)
            { if(metodologia[i] == "mA")
                      {notas[i]=runif(1,4,8)}
                      else { if(metodologia[i] == "mB")
                                          {notas[i]=runif(1,5,9)}
                                          else {notas[i]=runif(1,6,10)} } }
```

O primeiro comando acima sorteia a metodologia de ensino de cada um dos 500 alunos, já no formato de fator. Depois, um laço de repetição sorteia as notas dos alunos assim: a nota de um aluno que tenha estudado pela metodologia A é sorteada de uma distribuição uniforme entre 4 e 8; a nota de um aluno que tenha estudado pela metodologia B é sorteada de uma distribuição uniforme entre 5 e 9; e a nota de um aluno que tenha estudado pela metodologia C é sorteada de uma distribuição uniforme entre 6 e 10. O laço de repetição foi formatado para facilitar a sua compreensão, mas consiste num único comando. (Se você tiver dificuldade para entendê-lo, revise o estudo sobre estruturas de programação.)

Através da função `tapply`, que você já aprendeu, vê-se que as notas nessa amostra de alunos parecem ser influenciadas pela metodologia: `tapply(notas,metodologia,mean)`. Podemos prever que esse comando revelará que a média amostral dos alunos da metodologia A foi próxima de 6; a dos alunos da metodologia B, próxima de 7; e a dos alunos da metodologia C, próxima de 8. (Justifique essa previsão com o seu conhecimento de probabilidade.) Porém, a questão é: essa diferença na amostra é suficientemente grande para podermos rejeitar a hipótese de que as metodologias de ensino não têm impacto sobre o desempenho dos alunos? Em outras palavras, a diferença é estatisticamente significativa? Neste exercício, somos os criadores dos dados e sabemos que, sim, o verdadeiro valor esperado das notas dos alunos depende da metodologia de ensino. Mas um pesquisador no mundo real, diante apenas das 500 notas observadas, só poderia tirar conclusões de forma responsável e científica sobre o impacto das metodologias se fizesse um teste formal de hipótese.

Uma maneira de testar é pela análise de variância. Em R, usaremos duas novas funções: `lm` e `anova`. A função `lm` serve para estimar modelos lineares (*linear models*) e ela será o tema principal do próximo guia de estudo, dedicado às regressões lineares. A função `anova` é aquela que nos interessa mais agora. Digite:

```
anova(lm(notas ~ metodologia))
```

No comando acima, a função `anova` organiza o resultado da função `lm` numa tabela de análise de variância. Mostramos a seguir, a título de ilustração, o output do comando acima para uma simulação específica de notas. Os valores que você obterá no seu computador serão diferentes, mas as conclusões gerais devem ser as mesmas:

```
Analysis of Variance Table

Response: notas
             Df Sum Sq Mean Sq F value    Pr(>F)
metodologia   2 318.64 159.320  121.68 < 2.2e-16 ***
Residuals   497 650.75   1.309
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Na tabela acima, a coluna mais importante para você entender a essência da análise de variância é aquela intitulada `Sum Sq`, que marcamos em negrito. Ela contém a "soma de quadrados" (*sum of squares*) no fator e nos resíduos. Veja como se chegou ao número **318,64** na linha `metodologia` da tabela:

(i) Calculamos a variância amostral das notas dentro de cada subgrupo de alunos, segundo as metodologias a que pertencem. Com a amostra que gerou a tabela acima, as variâncias de notas em cada subgrupo: 1,2178 (A), 1,4064 (B) e 1,2971 (C). Cada uma dessas variâncias das notas de alunos que estudaram pela mesma metodologia nós não temos como explicar.

(ii) Multiplicamos cada variância pelo número de alunos no subgrupo menos 1, obtendo a "soma de quadrados" em cada metodologia. Na amostra da ilustração, essas somas de quadrados são 185,1078, 233,4644 e 232,1780. (Quantos alunos há em cada subgrupo?)

(iii) Somamos as três somas de quadrados, obtendo a parcela da variabilidade das notas que não é explicada pelo fator metodologia. Na amostra, a soma é **650,7502**. (Veja na linha `Residuals` da tabela acima.)

(iv) Multiplicamos a variância total das notas por 500-1, obtendo a "soma total de quadrados". Desta, subtraímos a soma de quadrados calculada no passo anterior. O que calculamos assim é a parcela da variabilidade das notas que é explicada pelo fator metodologia. A soma total de quadrados na amostra é 969,3895. Por fim: 969,3895 – **650,7502** = **318,6393**.

Como interpretar esse último número? De início, olhando só para as notas, temos 969,39 de "variabilidade" para explicar. É uma medida de o quanto as notas dos 500 alunos (não separados em subgrupos) varia ao redor da média de todos eles. Daquela variabilidade total, cerca de 1/3 é explicado pelas metodologias de ensino (318,64 ÷ 969,39 ≅ 0,33). Os outros 2/3

da variabilidade das notas não conseguimos explicar pela metodologia e são, portanto, um "resíduo" da análise. Isso não significa que outros fatores não possam explicar. A metodologia ANOVA admite dois ou mais fatores, mas não cobriremos isso neste estudo.

Ainda falta realizar o teste formal de hipótese. Ele se baseia na estatística F, que você estudou nas aulas teóricas. A estatística é calculada dividindo-se a soma de quadrados média do fator pela soma de quadrados média residual. Na ilustração, a estatística F para o fator metodologia é 159,32÷1,309 = 121,68. Sob a hipótese nula de que o fator não explica a variável de interesse, essa estatística é uma variável aleatória F. A última coluna da tabela mostra o p-valor associado a essa hipótese nula. Na ilustração, vemos que o p-valor associado ao fator metodologia é extremamente pequeno, indicando que a metodologia é muito significativa para explicar a variação das notas (como já sabíamos, pela simulação). Na sua tabela, os valores serão diferentes, mas você também deverá ver um p-valor muito pequeno para o fator metodologia, coerente com a construção dos dados simulados.

Lista de exercícios 8

Obs: Não utilize laços de repetição em nenhum dos exercícios.

(1) Depois de simular um vetor de 500 notas como descrito no guia, escreva um comando em R para calcular variância amostral das notas dentro de cada subgrupo de alunos segundo as metodologias a que pertencem.

(2) Escreva um comando para obter o número de alunos em cada metodologia.

(3) Calcule com um só comando a soma de quadrados explicada para o fator metodologia. Confirme que é o mesmo número obtido com a função `anova`.

(4) Como é obtida a primeira coluna da tabela ANOVA, intitulada `Df`, que contém o número de graus de liberdade associado ao fator e aos resíduos?

(5) Calcule manualmente (no R) a estatística de teste F para o fator e obtenha, com a função adequada em R, o p-valor associado a ela. Confirme que o mesmo é dado na última coluna da sua tabela ANOVA.

* * *