Planejamento e Análise de Experimentos

10 - Blocagem

Michel Bessani
https://orcslab.github.io/
Departamento de Engenharia Elétrica - DEE

Programa de Pós-Graduação em Engenharia Elétrica - PPGEE

UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

Belo Horizonte

Introdução

No experimento completamente aleatorizado (CRD, *completely randomized design*), abordado no teste ANOVA, as observações são coletadas de forma totalmente aleatorizada entre os níveis do tratamento (único fator de interesse).

Níveis do Tratamento	**Observações**			
1	y_{11}	y_{12}	$\dots$	y_{1n}
2	y_{21}	y_{22}	$\dots$	y_{2n}
$\vdots$	$\vdots$	$\vdots$	$\dots$	$\vdots$
a	y_{a1}	y_{a2}	$\dots$	y_{an}

Níveis do Tratamento	Observações			
1	y_{11}	y_{12}	$\dots$	y_{1n}
2	y_{21}	y_{22}	$\dots$	y_{2n}
$\vdots$	$\vdots$	$\vdots$	$\dots$	$\vdots$
a	y_{a1}	y_{a2}	$\dots$	y_{an}

Nesse design experimental, estamos assumindo, implicitamente, que as condições experimentais são homogêneas, i.e., não existe nenhum outro efeito importante além do fator investigado.

$$y_{ij} = \mu + \tau_i + \epsilon_{ij} \begin{cases} i = 1, \dots, a \\ j = 1, \dots, n \end{cases} , \quad \epsilon_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma^2\right)$$

Entretanto, existem situações em que o experimento terá outros fatores que não são de interesse (a princípio), mas que podem influenciar a variável de resposta. Estes fatores são conhecidos como fatores indesejáveis (*nuisance factors*).

A presença e uma estratégia para lidar com estes fatores indesejáveis foi discutida quando **testes pareados** foram apresentados.

Os fatores indesejados podem ser classificados em três tipos:

1. Desconhecidos (e não controláveis);
2. Conhecidos, mas não controláveis;
3. Conhecidos e controláveis.

A generalização de um experimento pareado para um número arbitrário de níveis de fatores é conhecido como **blocagem**.

É uma forma elegante de organizarmos a coleta de dados para que os efeitos dos fatores indesejáveis sejam removidos na modelagem estatística.

O uso da blocagem é benéfico mesmo na ausência de um fator indesejado conhecido, uma vez que aumenta o poder estatístico do teste ao remover a variabilidade entre repetições do termo residual.

George E.P. Box (1919-2013)

"*Block what you can, randomize what you cannot.*"

Introdução:
Blocagem ou Aleatorização?

A aleatorização (CRD, *Completely Randomized Design*) tenta prevenir que os fatores desconhecidos enviesem as observações.

A blocagem é feita quando sabemos, desde o princípio do planejamento do experimento, que existem fatores que influenciarão a variável de resposta, mas por algum motivo não estamos interessados no efeito desses fatores, e.g.:

- O efeito de diferentes lotes de matéria prima quando comparamos o desempenho de reações químicas;
- O efeito de diferentes problemas de benchmark quando comparamos o desempenho de algoritmos.

Se estivermos interessados no efeito desses fatores eles não serão mais indesejáveis e sim fatores experimentais adicionais.

Introdução:
Exemplo

Um jovem pesquisador decide **comparar um algoritmo padrão** de otimização **com seis versões modificadas** para resolver uma família do Problema de Roteamento de Veículos (PRV). A intenção é verificar se algumas das versões modificadas é **sistematicamente melhor** do que o algoritmo padrão.

Os algoritmos são aplicados em 180 instâncias do problema dividas em 36 grupos de instâncias homogêneas. Os algoritmos são executados 30 vezes em cada instância e o custo (e.g., tempo total das rotas) minimizado é armazenado.

O pesquisador quer um experimento com $\alpha = 0.05$ e poder $\pi = 1 - \beta = 0.8$ para um tamanho de efeito minimamente relevante de $\delta^* = 50$.

Além do efeito do fator experimental (algoritmos), existe um possível fator indesejável devido a variabilidade entre as unidades experimentais (grupos de instâncias).

Se aplicarmos o CRD, o termo residual irá conter toda a variabilidade entre as instâncias, o que resultará em um mascaramento do efeito de interesse (diferença entre os algoritmos).

Como podemos controlar a alocação das execuções dos algoritmos em cada instância, uma possibilidade é fazê-la de forma sistemática para que possamos modelar, além do efeito do fator experimental, os efeitos das instâncias.

Isso pode ser feito executando cada algoritmo para cada um dos 36 grupos de instâncias homogêneas e cada grupo será um **bloco**.

Em cada um dos blocos, a ordem da execução deve ser aleatorizada para prevenir o efeito de outros fatores desconhecidos na variável de resposta.

Para o caso do experimento computacional, a aleatorização intra-bloco é desnecessária, mas para casos de sistemas menos controláveis é importante.

Esse planejamento experimental é conhecido como Experimento Completamente Aleatorizado com Blocos (RCBD, *randomized complete block design*) com um fator de blocagem.

Experimento Completamente Aleatorizado com Blocos (RCBD)

Um Experimento Completamente Aleatorizado com Blocos assume:

- Uma observação por bloco[1];
- Blocos independentes;
- Independência entre a aleatorização intra-bloco.

A não consideração de dependência (i.e., usar blocos que não são independentes) pode resultar em pseudoreplicação, o que leva a consideração de graus de liberdade incorretos na estatística de teste e, consequentemente, um aumento na taxa de Erro Tipo I.

[1] Para múltiplas observações por bloco deve-se utilizar um experimento aleatorizado generalizado com blocos (GRBD, *Generalized randomized block design)*.

Experimento Completamente Aleatorizado com Blocos (RCBD): Aplicando no Exemplo

Para atender a esses requisitos, vamos considerar nossa variável de resposta como o desempenho médio de um algoritmo em um grupo particular de instâncias.

O desempenho de cada algoritmo será a média entre o custo observado em cada instância (30 execuções), e depois a média intra-grupo (média das 5 instâncias homogêneas). Assim, teremos 36 observações independentes para cada algoritmo.

Observações[2]			
Bloco 1	Bloco 2	...	Bloco 36
$\bar{y}_{original}$	$\bar{y}_{original}$	...	$\bar{y}_{original}$
$\bar{y}_{mod1}$	$\bar{y}_{mod1}$	...	$\bar{y}_{mod1}$
$\vdots$	$\vdots$		$\vdots$
$\bar{y}_{mod6}$	$\bar{y}_{mod6}$	...	$\bar{y}_{mod6}$

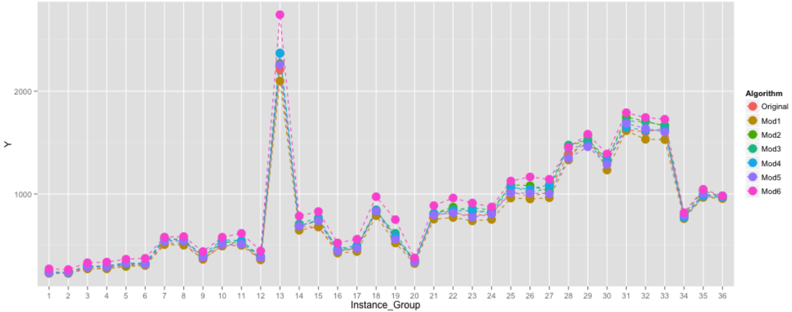
[2] médias das 30 execuções nas 5 instâncias homogêneas em cada bloco.

```
# Load data
> data <- read.table(algo.csv", header = TRUE)
# Aggregate data (algorithm means by instance group)
> aggdata <- with(data, aggregate(x    = Result,
                                  by   = list(Algorithm, Group),
                                  FUN = mean))
# Rename columns and coerce factor variables
> names(aggdata) <- c("Algorithm", "Instance_Group", "Y")
> for (i in 1:2) aggdata[, i] <- as.factor(aggdata[, i])

> levels(aggdata$Algorithm) <- c("Original",
                                 unlist(lapply(X    = "Mod",
                                               FUN = paste0, 1:6)))
> summary(aggdata)
    Algorithm   Instance_Group       Y
 Original:36   1       :  7    Min.   : 227.3
 Mod1    :36   2       :  7    1st Qu.: 440.0
 Mod2    :36   3       :  7    Median : 756.8
 Mod3    :36   4       :  7    Mean   : 821.0
 Mod4    :36   5       :  7    3rd Qu.:1020.5
 Mod5    :36   6       :  7    Max.   :2743.6
 Mod6    :36   (Other):210$
```

```
library(ggplot2)
> p <- ggplot(aggdata, aes(x = Instance_Group, y = Y,
                           group = Algorithm, colour = Algorithm))

> p + geom_line(linetype=2) + geom_point(size=5)
```



Experimento Completamente Aleatorizado com Blocos (RCBD): Modelagem Estatística

No caso geral, temos a níveis do fator experimental e b níveis do fator bloqueado. O modelo estatístico é dado por:

$$y_{ij} = \mu + \tau_i + \beta_j + \epsilon_{ij} \begin{cases} i = 1, \ldots, a \\ j = 1, \ldots, b \end{cases} \quad , \quad \epsilon_{ij} \sim \mathcal{N}\left(0, \sigma^2\right)$$

De forma similar ao CRD, temos por construção:

$$\sum_{i=1}^{a} \tau_i = 0 \qquad \sum_{j=1}^{b} \beta_j = 0$$

No RCBD, estamos interessados apenas no efeito do fator experimental. Consequentemente, o teste de hipóteses se refere apenas a esses coeficientes.

$$\begin{cases} H_0 : \tau_i = 0, \;\; \forall i = 1, \cdots, a \\ H_1 : \exists\, \tau_i \neq 0 \end{cases}$$

A variabilidade total das observações pode ser particionada da seguinte maneira:

$$
\begin{aligned}
SS_T &= \sum_{i=1}^{a}\sum_{j=1}^{b}\left(y_{ij} - \bar{y}_{..}\right)^2 \\
&= \underbrace{b\sum_{i=1}^{a}\left(\bar{y}_{i.} - \bar{y}_{..}\right)^2}_{SS_{\text{levels}}} + \underbrace{a\sum_{j=1}^{b}\left(\bar{y}_{.j} - \bar{y}_{..}\right)^2}_{SS_{\text{blocks}}} + \underbrace{\sum_{i=1}^{a}\sum_{j=1}^{b}\left(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}\right)^2}_{SS_E}
\end{aligned}
$$

Cada soma de quadrados dividida pelo seu grau de liberdade é um quadrado médio:

$$MS_{\text{levels}} = \frac{SS_{\text{levels}}}{a-1} \qquad MS_{\text{blocks}} = \frac{SS_{\text{blocks}}}{b-1} \qquad MS_E = \frac{SS_E}{(a-1)(b-1)}$$

O valor esperado dos quadrados médios com os tratamentos e blocos fixos pode ser escrito como:

$$E\left[MS_{\text{levels}}\right] = \sigma^2 + \frac{b\sum_{i=1}^{a}\tau_i^2}{a-1} \qquad E\left[MS_{\text{blocks}}\right] = \sigma^2 + \frac{a\sum_{j=1}^{b}\beta_j^2}{b-1}$$

$$E\left[MS_E\right] = \sigma^2$$

$$\begin{cases} H_0 : \tau_i = 0, \ \ \forall i = 1, \cdots, a \\ H_1 : \exists\, \tau_i \neq 0 \end{cases}$$

$$E\left[MS_{\text{levels}}\right] = \sigma^2 + \frac{b \sum_{i=1}^{a} \tau_i^2}{a-1} \qquad E\left[MS_E\right] = \sigma^2 \qquad F_0 = \frac{MS_{\text{levels}}}{MS_E}$$

Sob a H_0, temos que a estatística de teste F_0 é distribuída de acordo com a distribuição F com $(a-1)$ graus de liberdade no numerados e $(a-1)(b-1)$ no denominador.

$$MS_{\text{levels}} = \frac{SS_{\text{levels}}}{a-1} \qquad MS_E = \frac{SS_E}{(a-1)(b-1)}$$

A região crítica para esse teste é dada por:

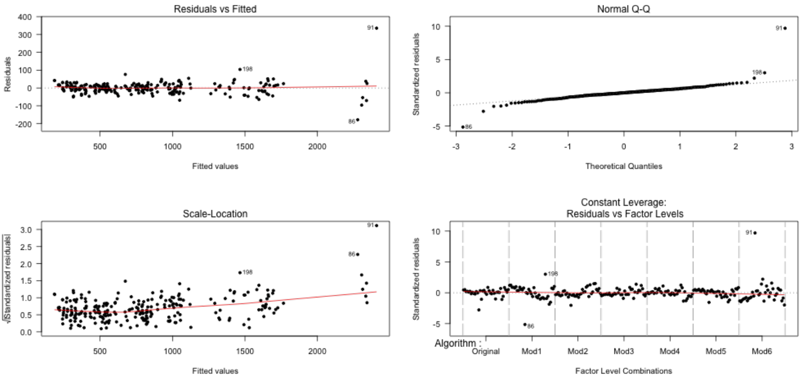
$$f_0 > F_{\alpha;(a-1),(a-1)(b-1)}$$

```
# First model
> model <- aov(Y~Algorithm+Instance_Group, data=aggdata)

> summary(model)
                Df   Sum Sq Mean Sq F value Pr(>F)
Algorithm        6   359949   59991   41.75 <2e-16 ***
Instance_Group  35 60438639 1726818 1201.86 <2e-16 ***
Residuals      210   301725    1437
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> summary.lm(model)$r.squared
[1] 0.9950618
```

```
> par(mfrow = c(2, 2))
> plot(model, pch = 20, las = 1)
```

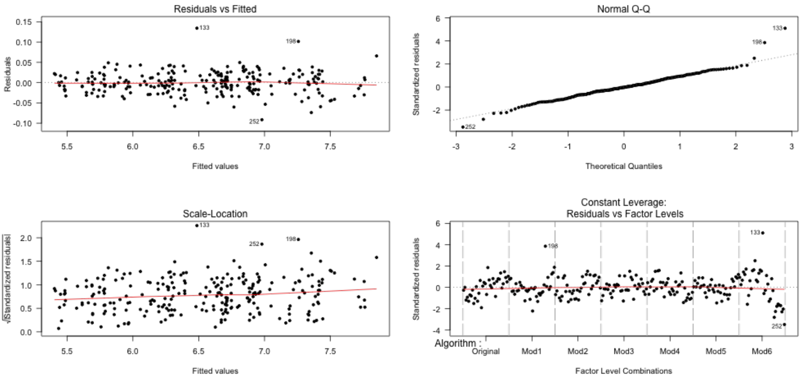


```
# Trying with log-transformed response variable
> model2 <- aov(log(Y)~Algorithm+Instance_Group, data=aggdata)

> summary(model2)
                 Df Sum Sq Mean Sq F value Pr(>F)
Algorithm         6   0.60  0.1002   120.4 <2e-16 ***
Instance_Group   35  89.92  2.5690  3089.0 <2e-16 ***
Residuals       210   0.17  0.0008
---

> summary.lm(model2)$r.squared
[1] 0.9980742
```

```
> par(mfrow = c(2, 2))
> plot(model2, pch = 20, las = 1)
```



Se existirem dois ou mais fatores que devem ser bloqueados, é possível adicionar mais termos ao modelo estatístico.

A soma dos quadrados para esse novo termo é similar aos já apresentados para o caso com um único bloco.

Existe uma abordagem de planejamento experimental específica que é conhecida como Cubo Latino[3] que pode ser utilizada nesse caso.

[3] https://online.stat.psu.edu/stat503/lesson/4/4.3

Experimento Completamente Aleatorizado com Blocos (RCBD): Eficiência da Blocagem

Vamos analisar a estatística de teste F_0, a qual é a razão entre o quadrado médio do fator experimental e o quadrado médio do erro:

$$F_0 = \frac{MS_{\text{levels}}}{MS_E}$$

Uma vez que bloqueamos o fator indesejado, estamos removendo seu efeito do erro do modelo linear.

$$y_{ij} = \mu + \tau_i + \beta_j + \epsilon_{ij} \begin{cases} i = 1, \ldots, a \\ j = 1, \ldots, b \end{cases} \quad , \; \epsilon_{ij} \sim \mathcal{N}\left(0, \sigma^2\right)$$

Se esse experimento fosse realizado como um CRD, i.e., considerando apenas o fator experimental, o modelo linear seria:

$$y_{ij} = \mu + \tau_i + \epsilon_{ij} \begin{cases} i = 1, \ldots, a \\ j = 1, \ldots, n \end{cases} , \ \epsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$$

Consequentemente, a variabilidade devido ao efeito bloqueado, β_j no slide anterior, estaria presente no erro e aumentaria o valor do MS_E.

Dessa forma, seria necessário observar um MS_{levels} maior para se chegar a mesma conclusão que a obtida no experimento com blocos.

Após a realização de um RCBD e de ajustarmos o modelo linear aos dados, é possível calcular a *eficiência relativa da blocagem (E)*, a qual quantifica quão maior deveria ser o número de observações para obtermos em um CRD o mesmo poder estatístico que no RCBD.

Isso pode ser calculado utilizando os quadrados médios retornados na tabela do modelo linear ajustado:

$$E = \frac{(b-1)MS_{\text{Blocks}} + b(a-1)MS_{\text{E}}}{(ab-1)MS_{\text{E}}}$$

Por exemplo, um valor de 1.3 indica que um CRD precisaria de 30% mais observações para ter o mesmo poder. Para o exemplo, esse valor poder ser calculado e é $E = 431.6$.

Experimento Completamente Aleatorizado com Blocos (RCBD): Tamanho Amostral

A determinação do tamanho amostral necessário para um RCBD é similar ao procedimento para um CRD.

O número de observações calculadas com as fórmulas para o CRD será o número de blocos necessários e teremos uma observação por bloco.

A variabilidade intra-grupo, necessária para calcular o tamanho amostral necessário, corresponde a variância residual estimada, i.e., a variabilidade não explicada pelo fator experimental e pelo fator bloqueado.

Essa variabilidade pode ser estimada a partir de um estudo piloto ou informações de estudos passados.

Similar ao CRD, pode ser conveniente utilizar as comparações múltiplas post-hoc como base para o cálculo do tamanho amostral, especialmente nos casos em que o custo para coletar os dados não é alto.

Existem também Curvas Operacionais Características[4]. (OCR, *Operating Characteristic Curves*) que podem ser utilizadas. As curvas consideram diferentes valores de:

- Tamanho do efeito (δ);
- Desvio padrão do erro (σ);
- Número de blocos (b);
- Número de fatores experimentais (a).

[4] Veja 4.1.3 de D.C. Montgomery, Design and Analysis of Experiments, Ch.4-5, 7th ed., Wiley, 2005.

Experimento Completamente Aleatorizado com Blocos (RCBD): Comparações Múltiplas

Para as comparações múltiplas post-hoc, adota-se um procedimento similar ao apresentado para o ANOVA.

Os testes-t pareados são utilizados para comparar os pares e o número de graus de liberdade do erro $(a-1)(b-1)$ é utilizado na distribuição de referência.

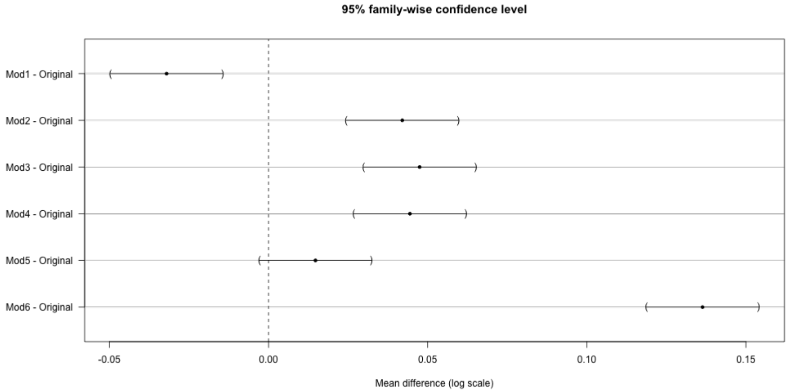
As comparações podem ser *todos contra todos* ou *todos contra um*, a depender da motivação do experimento.

No nosso exemplo, o pesquisados queria verificar se alguma das seis modificações seria sistematicamente melhor do que uma versão padrão de um algoritmo de otimização para uma família de problemas de roteamento de veículos.

Para responder a pergunta motivadora, pode-se realizar uma comparação de *todos contra um* para verificar se alguma das 6 versões modificadas é melhor do que a padrão.

Utiliza-se o teste de Dunnett devido a sua sensibilidade superior ao fazer esse tipo de comparação.

```
> library(multcomp)
> duntest     <- glht(model2, linfct = mcp(Algorithm = "Dunnett"))
> duntestCI   <- confint(duntest)
> par(mar = c(5, 8, 4, 2), las = 1)
> plot(duntestCI, xlab = "Mean difference (log scale)")
```



Um RCBD é uma generalização do pareamento e é uma forma elegante de remover efeitos indesejados da análise.

A consequência de utilizar a blocagem é um aumento da sensibilidade do teste, necessitando de um menor tamanho amostral.

Na prática, a realização de um experimento com blocagem resulta em pouco aumento do custo do experimento se comparado com um CRD, devendo ser o planejamento escolhido sempre que possível.

O planejamento e análise de experimento com blocos incompletos (observações faltantes) também pode aplicado[5].

[5] Veja D.C. Montgomery, Design and Analysis of Experiments, Ch.4-5, 7th ed., Wiley, 2005.

Referências

- D.C. Montgomery, *Design and Analysis of Experiments*, Ch.4-5, 7th ed., Wiley, 2005.
- Campelo, F., Takahashi, F. *Sample size estimation for power and accuracy in the experimental comparison of algorithms*. Journal of Heuristics, 25, 305–338, 2019.

Adicionais:

- P. D. Ellis, *The Essential Guide to Effect Sizes*. 1st ed., Cambridge, 2010.
- J.P. Schaffer, *Multiple Hypothesis Testing*, Annual Reviews on Psychology 46, 561–584, 1995.
- P. Mathews, *Sample Size Calculations: Practical Methods for Engineers and Scientists*. Ch. 8, 1st ed., MMB, 2010.
- Lee, S., & Lee, D. K., *What is the proper way to apply the multiple comparison test?* Korean journal of anesthesiology, 71(5), 353-360, 2018.