Planejamento e Análise de Experimentos

8 - Análise de Variância (ANOVA)

Michel Bessani
Departamento de Engenharia Elétrica - DEE

Programa de Pós-Graduação em Engenharia Elétrica - PPGEE

UFMG
UNIVERSIDADE FEDERAL DE MINAS GERAIS

Belo Horizonte

Introdução

Vimos os principais conceitos e estratégias para compararmos a média de uma amostra com uma referência ou entre duas amostras.

Entretanto, existem casos em que podemos estar interessados em inferir sobre a diferença na média de mais de duas amostras.

Vamos desenvolver os conceitos principais relacionados com esse tipo de teste através de um exemplo relacionado com a operação de manufatura de papel.

Análise de Variância (ANOVA)

Exemplo: A resistência à tração é uma característica importante para certas aplicações industriais de papel.

Uma conjectura razoável é que essa característica é influenciada pelo tipo de fibra de madeira utilizada durante a manufatura.

Um engenheiro de processos decide investigar, utilizando uma planta piloto como sua unidade experimental, quais de quatro tipos de fibra de madeira resultam em diferenças relevantes da resistência à tração.

O orçamento limitado para o experimento permite apenas seis ciclos de produção para cada tipo de fibra de madeira.

Com essas informações, sabemos que o experimento possui um **único fator** (fibra de madeira) com $a = 4$ **níveis** (Tipos A, B, C e D de fibra) e $n = 6$ **repetições** para cada nível.

A **variável de resposta** é a resistência a tração do papel (medida, e.g., em kPa) e o engenheiro está interessado em descobrir qual dos tipos de fibras resulta na maior resistência do papel.

A diferença minimamente relevante foi definida como 5 kPa e é razoável uma estimativa máxima para o desvio padrão desse processo em $\hat{\sigma} = 6$ kPa.

As probabilidades dos erros foram definidas como $\alpha = 0.1$ e $\beta = 0.2$.

O experimento possui um **único fator** (fibra de madeira) com $a = 4$ **níveis** (Tipos A, B, C e D de fibra) e $n = 6$ **repetições** para cada nível.

A **variável de resposta** (y) é a resistência a tração do papel (medida, e.g., em kPa).

A Tab. 1 apresenta como é o resultado do experimento.

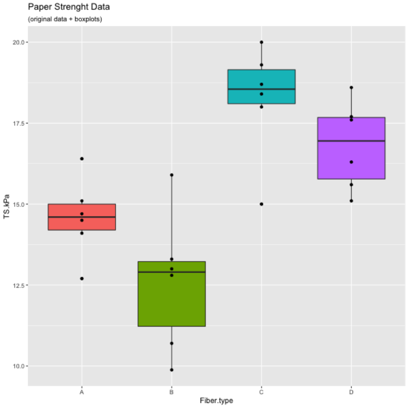
Tabela: Resultado do experimento

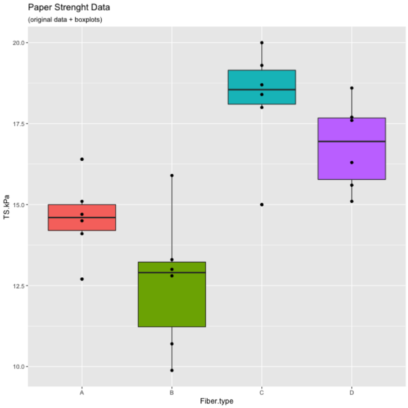
Tratamento $a = 4$ níveis	**Observações** y_{ij}	**Médias** $\bar{y}_{i.}$
1 (A)	$y_{11},\ y_{12},\ y_{13},\ y_{14},\ y_{15},\ y_{16}$	$\bar{y}_{1.}$
2 (B)	$y_{21},\ y_{22},\ y_{23},\ y_{24},\ y_{25},\ y_{26}$	$\bar{y}_{2.}$
3 (C)	$y_{31},\ y_{32},\ y_{33},\ y_{34},\ y_{35},\ y_{36}$	$\bar{y}_{3.}$
4 (D)	$y_{41},\ y_{42},\ y_{43},\ y_{44},\ y_{45},\ y_{46}$	$\bar{y}_{4.}$

Vamos visualizar os dados coletados:

```
> paper <- read.table(file   = "../data files/paper_strength.csv",
+                     header = TRUE,
+                     sep    = ",")
```

```
> library(ggplot2)
> ggplot(paper,
+     aes(x    = Fiber.type,
+         y    = TS.kPa,
+         fill = Fiber.type)) +
+  geom_boxplot() +
+  geom_point()
```





O boxplot indica:

- Diferença entre os tratamentos;
- Variabilidade diferente para cada nível;
- Assimetria no nível B;
- Possível *outlier* no nível C.

Essas características terão que ser ponderadas durante a análise.

Análise de Variância (ANOVA): Modelo Estatístico

Podemos escrever um modelo estatístico aditivo (linear) para os dados:

$$y_{ij} = \underbrace{\mu_i + \epsilon_{ij}}_{\text{Modelo das Médias}} \qquad \begin{cases} i = 1, \ldots, a \\ j = 1, \ldots, n \end{cases}$$

onde μ_i é a média de cada nível e ϵ_{ij} é o resíduo (erro aleatório ou variabilidade não modelada).

Podemos ainda alterar o modelo para termos um termo referente ao **tamanho do efeito** de cada nível:

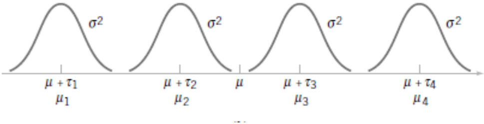
$$y_{ij} = \mu_i + \epsilon_{ij} = \underbrace{\mu + \tau_i + \epsilon_{ij}}_{\text{Modelo dos Efeitos}} \quad \begin{cases} i = 1, \dots, a \\ j = 1, \dots, n \end{cases}$$

onde μ é a média global, τ_i é o efeito do i-ésimo nível e ϵ_{ij} é o resíduo.

Na criação do teste estatístico para a existência das diferenças nas médias dos grupos, será utilizado o modelo dos efeitos com a consideração de algumas premissas sobre os resíduos:

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

Se as premissas forem verdadeiras, então é esperado que as populações sejam distribuídas da seguinte forma:



Uma vez que estamos interessados em testar se existe diferença no valor médio devido aos níveis, podemos testar se existe diferença no tamanho dos efeitos τ_i.

Nesse caso o teste de hipóteses é da seguinte forma:

$$\begin{cases} H_0 : \tau_i = 0, \ \ \forall i \in \{1, 2, \ldots, a\} \\ H_1 : \exists \ \tau_i \neq 0 \end{cases}$$

Se o processo de coleta de observações for em ordem aleatória e com condições experimentais constantes, teremos um **experimento completamente randomizado** (CRD, *Completely Randomized Design*).

Tal estratégia de modelar o efeito médio de cada nível do fator é conhecida como **modelo de efeitos fixos**.

É uma abordagem apropriada para testar hipóteses em situações em que os níveis dos fatores são definidos de forma arbitrária pelo experimentador.

Como a inferência é feita sobre os valores médios de cada nível, os resultados não podem ser extrapolados para níveis similares que não foram testados.

Existem outros modelos, como modelos de efeitos aleatórios ou de mistura de efeitos (fixos e aleatórios), os quais não serão explorados aqui.

Voltando ao nosso modelo de efeitos fixos:

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

Temos que os efeitos dos tratamentos τ_i são um deslocamento em relação a média global μ.

Por construção:

$$\sum_{i=1}^{a} \tau_i = 0$$

A variabilidade total das observações pode ser expressa como a *soma dos quadrados* (SS), a qual representa a soma do quadrado dos desvios de cada observação y_{ij} para a média amostral global $\bar{y}_{\bullet\bullet}$:

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{\bullet\bullet})^2$$

Onde $\bullet$ indica o somatório em um índice e $^{-}$ indica uma operação de média.

Realizando uma manipulação algébrica, a SS_T pode ser divida em duas partes, a variabilidade intra-grupo e a inter-grupo:

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{\bullet\bullet})^2 = \underbrace{n\sum_{i=1}^{a}(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2}_{SS_{Níveis}} + \underbrace{\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i\bullet})^2}_{SS_E}$$

Se dividirmos as somas dos quadrados por seus respectivos graus de liberdade obtemos as médias quadráticas.

$$MS_E = \frac{SS_E}{a\left(n-1\right)} \qquad MS_{Níveis} = \frac{SS_{Níveis}}{a-1}$$

Utilizaremos as médias quadráticas dos níveis e do resíduo para realizar o nosso teste. Os valores esperados dessas quantidades são:

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

Alguns comentários sobre esses valores esperados:

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

O MS_E é um estimador não enviesado da variância dos resíduos.

Enquanto que $MS_{Níveis}$ é enviesado por um termo que é proporcional a soma dos quadrados do efeito dos níveis τ_i.

Relembrando as Hipóteses a serem testadas:

$$\begin{cases} H_0 : \tau_i = 0, \ \ \forall i \in \{1, 2, \ldots, a\} \\ H_1 : \exists \ \tau_i \neq 0 \end{cases}$$

Logo, sob a nossa $H_0 : \tau_i = 0, \ \ \forall i \in \{1, 2, \ldots, a\}$, teremos:

$$E\left[MS_E\right] = E\left[MS_{Níveis}\right] = \sigma^2$$

Pode ser mostrado que, sob H_0, a estatística:

$$F_0 = \frac{MS_{Níveis}}{MS_E}$$

é distribuída de acordo com uma distribuição F com $a-1$ graus de liberdade no numerador e $a(n-1)$ graus de liberdade no denominador, $F_{(a-1),a(n-1)}$.

Se H_0 for falso, o valor observado de $MS_{Níveis}$ será maior do que MS_E, o que resultará em valores elevados de F_0.

Região de rejeição

Rejeitar H_0 com α de nível de significância se

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

Análise de Variância (ANOVA): Análise Computacional

```
> my.model <- aov(TS.kPa ~ Fiber.type,
+               data = paper)
>
> summary.aov(my.model)

Df Sum Sq Mean Sq F value   Pr(>F)
Fiber.type   3 110.77   36.92   13.62 4.56e-05 ***
Residuals   20  54.24    2.71
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A tabela do teste ANOVA fornece, além das informações das fontes de variabilidade (níveis e erro aleatório) com seus respectivos graus de liberdade (***D**egrees of **f**reedom*), soma dos quadrados e quadrados médios, o valor da estatística de teste F_0 e o seu p-valor correspondente.

Nesse exemplo, o p-valor$= 4.56 \times 10^{-5}$ sugere que H_0 pode ser rejeitada em favor da H_1. Qual a interpretação dessa rejeição?

Vamos relembrar como definimos o teste de hipótese para o ANOVA:

$$\begin{cases} H_0 : \tau_i = 0, \;\; \forall i \in \{1,\ 2, \ldots,\ a\} \\ H_1 : \exists\ \tau_i \neq 0 \end{cases}$$

A rejeição da H_0 indica que existe *ao menos um nível* com um efeito significativamente diferente de zero. Mas qual?

Para finalizarmos a análise, precisamos responder duas perguntas:

- As premissas do teste foram atendidas?
- Quais efeitos médios são diferentes, e qual o tamanho dessa diferença?

Voltando ao nosso modelo aditivo de efeitos fixos:

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

Precisamos verificar três premissas:

1. Independência;
2. Homoscedasticidade, i.e., igualdade da variância entre os grupos;
3. Normalidade.

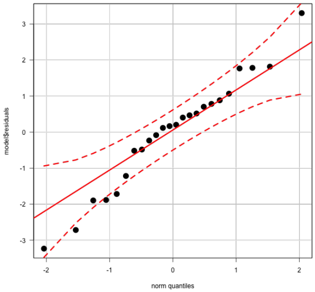
Os resíduos do modelo podem ser obtidos da seguinte forma:

$$\epsilon_{ij} = y_{ij} - \hat{y}_i = y_{ij} - (\hat{\mu} + \hat{\tau}_i) = y_{ij} - \bar{y}_{i\bullet}$$

ou obtidos diretamente do modelo ajustado no R: `my.model$residuals`

A premissa de normalidade dos resíduos pode ser avaliada pelo teste de Shapiro-Wilk ou por análise gráfica:

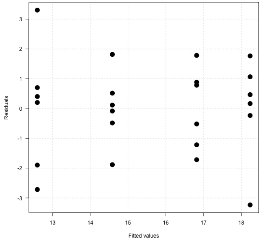
```
> shapiro.test(model$residuals)
Shapiro-Wilk normality test
data:  my.model$residuals
W = 0.9722, p-value = 0.7225

> library(car)
> qqPlot(my.model$residuals,
pch = 16,
lwd = 3,
cex = 2,
las = 1)
```



O teste ANOVA é relativamente robusto a violações moderadas da normalidade, desde que as outras premissas sejam validadas (ou se o tamanho amostral for grande o suficiente).

A premissa de homoscedasticidade pode ser verificada pelo teste de Fligner-Killen ou pelo gráfico dos resíduos pelos valores ajustados.

```
> fligner.test(TS_kPa ~ Fiber.type, data = paper)
Fligner-Killeen test of homogeneity of
variances
data:  data:  TS.kPa by Fiber.type
Fligner-Killeen:
med chi-squared = 1.0622, df = 3,
p-value = 0.7862
>
> plot(x = my.model$fitted.values,
+       y = my.model$residuals)
```



ANOVA é relativamente robusto a violações modestas da homoscedasticidade desde que as observações sejam balanceadas.

A premissa de independência deve ser garantida (de acordo com o conhecimento do experimentador) na etapa de planejamento, coleta e análise.

Isso inclui a não realização de pseudoreplicações e ordenação dos efeitos, entre outros.

Pode-se utilizar o teste de Durbin-Watson da autocorrelação desde que as observações sejam ordenadas por um fator incontrolável e com possível influencia nos dados, e.g., a ordem de coleta dos dados.

O teste ANOVA, assim como os outros testes, é sensível a violação da premissa de independência. A atenção a fatores com possibilidade de influência na independência pode ajudar a evitar violações dessa premissa.

Se decidirmos pela rejeição da H_0 e as premissas do teste ANOVA forem validadas, i.e., temos uma evidência sólida de que pode-se confiar nos resultados do teste, precisaremos determinar quais níveis do fator são significativamente diferentes.

$$\begin{cases} H_0 : \tau_i = 0, \;\; \forall i \in \{1,\ 2, \ldots,\ a\} \\ H_1 : \exists\ \tau_i \neq 0 \end{cases}$$

Sempre que possível, iremos planejar antecipadamente as comparações posteriores ao teste.

A definição de hipóteses Post-hoc, i.e., após realizarmos um primeiro teste de hipóteses, é um ponto comum de viés na análises e deve ser feito com o máximo de cuidado.

Comparações Múltiplas

O planejamento de comparações múltiplas deve ser guiado pela **questão técnica** que motivou a realização do experimento.

Sempre que possível, o pesquisador deve optar por fazer o menor número de comparações necessárias para responder a questão motivadora.

Um menor número de comparações resultará no maior poder para o teste, ou em um menor número de observações.

Questão usuais:

- Como um nível se compara aos outros (teste *todos contra um*)?
- Como os níveis se comparam entre si (teste *todos contra todos*)?

Comparações por pares de todos contra todos são realizadas quanto estamos interessados em verificar quais níveis são significativamente diferentes e sem utilizar nenhuma informação inicial ou interesse específico em algum nível ou ranqueamento.

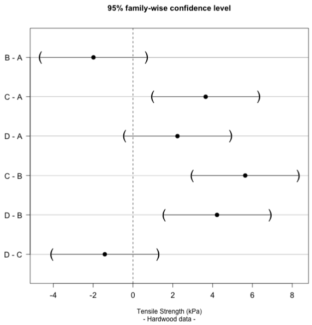
Nesse caso, o número de comparações será $K = a(a-1)/2$, onde a é o número de níveis.

Para realizar a comparação *todos contra todos*, é comum utilizar a Abordagem de Diferença Significativa Honesta (HSD, *Honest Significant Difference*) de **Tukey**.

Esse método resulta em um poder maior do que a realização de múltiplos testes t com valores ajustados de α.

Ilustrando com o exemplo da fibra de madeira:

```
> library(multcomp)
> mc1    <- glht(my.model, linfct = mcp(Fiber.type = "Tukey"))
> mc1_CI <- confint(mc1, level = 0.95)

Simultaneous Confidence Intervals
Multiple Comparisons of Means: Tukey Contrasts
95% family-wise confidence level


Linear Hypotheses:
           Estimate lwr     upr
B - A == 0 -1.9867  -4.6478  0.6745
C - A == 0  3.6500   0.9889  6.3111
D - A == 0  2.2333  -0.4278  4.8945
C - B == 0  5.6367   2.9755  8.2978
D - B == 0  4.2200   1.5589  6.8811
D - C == 0 -1.4167  -4.0778  1.2445

> plot(mc1_CI)
```



Comparações aos pares de *todos contra um* normalmente surgem no contexto de experimentos com um grupo de controle:

- Comparação de um método proposto com um bem aceito;
- Comparação de diferentes abordagens contra uma padrão (ou um grupo placebo).

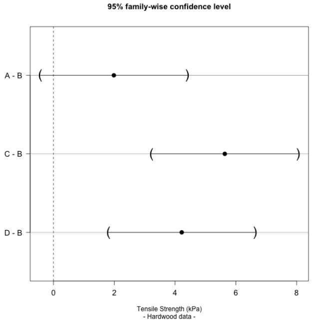
Nesses casos, o número de comparações é $K = a - 1$, onde a é o número de níveis.

O teste de **Dunnett** é usualmente utilizado para esse tipo de comparação devido a sua sensibilidade.

Assumindo, em nosso exemplo, que o nível *B* é o padrão e queremos comparar todos os outros com ele:

```
> paper$Fiber.type<-relevel(paper$Fiber.type, ref = "B")
> model2          <-aov(TS.kPa ~ Fiber.type, data = paper)
> mc2             <-glht(model2, linfct = mcp(Fiber.type = "Dunnett"))
> mc2_CI          <-confint(mc2, level = 0.95)
```

```
Simultaneous Confidence Intervals
Multiple Comparisons of Means: Dunnett
Contrasts 95% family-wise confidence level
Linear Hypotheses:
Estimate lwr     upr
A - B == 0  1.9867  -0.4275  4.4008
C - B == 0  5.6367   3.2225  8.0508
D - B == 0  4.2200   1.8058  6.6342

> plot(mc2_CI)
```



Comparações Múltiplas: Tamanho Amostral

As comparações que são realizadas após o teste ANOVA devem ser planejadas anteriormente ao teste, uma vez que elas influenciarão a coleta de dados e o tamanho amostral.

Existem fórmulas para o tamanho amostral apenas para o ANOVA, mas esses são limitados por dependerem de parâmetros contraintuitivos, e.g., a razão entre as variâncias intra-grupo e inter-grupo.

A **abordagem geral** é utilizar ajustes para corrigir o α a fim de reduzir o aumento do Erro Tipo-I e depois calcular o n utilizando as fórmulas para o teste-t de duas amostras.

O aumento do Erro Tipo-I é devido à utilização do mesmo conjunto de dados para realizar múltiplos testes de hipóteses.

Podemos pensar nas comparações múltiplas após o ANOVA como uma sequência de testes-t para a diferença na média de duas populações.

Se as premissas do ANOVA são validadas, já temos algumas informações sobre os dados: *i)*homoscedasticidade e *ii)* a variância comum dos grupos é estimada como o MS_E com $a(n-1)$ graus de liberdade.

Também sabemos que a probabilidade do Erro Tipo-I de cada comparação é α. Se quisermos manter a taxa de erro controlada em um determinado nível enquanto realizamos as comparações múltiplas, será necessário corrigir o valor do α utilizado para cada teste.

Existem diferentes formas de ajustar o α para as comparações dos pares de tal forma a manter a taxa de erro global em um nível controlado.

Um dos métodos mais comuns (e conservador) é o de Bonferroni.

Assumindo que serão realizadas K comparações, o α ajustado será:

$$\alpha_{aj} = \frac{\alpha}{K}$$

Esse valor de α_{aj} deve ser utilizado para calcular o tamanho amostral:

$$n \approxeq 2 \left[\left(t_{\alpha_{aj}/2}^{a(n-1)} + t_{\beta}^{a(n-1)} \right) d^* \right]^2$$

onde d^* é o tamanho do efeito minimamente relevante normalizado.

Caso o interesse seja em calcular o tamanho amostral diretamente para o ANOVA (sem preocupações com as comparações múltiplas post-hoc), as fórmulas são simples como no caso do teste-t.

Essencialmente, o poder (ou tamanho amostral) pode ser calculado para o ANOVA através da equação:

$$F_{(1-\alpha)} = F_{\beta;\phi}$$

Com ambas distribuições F tendo $(a-1)$ graus de liberdade no numerador e $a(n-1)$ graus de liberdade no denominador. O parâmetro de não centralidade ϕ é dado por:

$$\phi = \frac{n \sum_{i=1}^{a} \tau_i^2}{\hat{\sigma}^2}$$

Vamos ilustrar esse cálculo de tamanho amostral. Considere um planejamento experimental com $a = 4$, $\alpha = 0.05$ e uma estimativa do máximo desvio padrão intra-grupos de $\sigma = 7$.

Suponha que o experimentador quer detectar se quaisquer duas médias apresentam diferenças de $\delta^* = 12$ com um poder $(1 - \beta) = 0.8$.

Nessas condições, pode-se assumir um cenário hipotético (conservador) onde temos dois níveis com desvios simétricos em relação à média global e os outros dois sem desvios:

$$\tau = \left\{-\frac{\delta^*}{2}, \frac{\delta^*}{2}, 0, 0\right\}$$

Para esse caso, teremos o parâmetro de não centralidade:

$$\phi = \frac{n\sum_{i=1}^{a}\tau_i^2}{\hat{\sigma}^2} = \frac{4\left((-6)^2 + 6^2 + 0 + 0\right)}{7^2} = 5.88$$

O que permite calcular o tamanho amostral iterando em n até:

$$F_{(1-\alpha)} \leq F_{\beta;\phi}$$

Utilizando `power.anova.test()`:

```
> vartau <- var(tau)
> power.anova.test(groups = 4, between.var = vartau,
+                  within.var = sigma^2, sig.level = alpha,
+                  power = 1 - beta)$n
[1] 8.463358
```

Referências

Principais:

- D.C. Montgomery, G.C. Runger, *Applied Statistics and Probability for Engineers*, Ch. 13. 5th ed., Wiley, 2010.
- N.L. Kerr, *HARKing: Hypothesizing After the Results are Known*, Personality and Social Psychology Reviews 2(3): 196–217, 1998.

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.
- T. Hothorn, F. Bretz, P. Westfall, *Simultaneous Inference in General Parametric Models*. Biometrical Journal 50(3), 346–363, 2008.
- 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.