Lista 3 - Planejamento de Experimento

Gustavo Almeida Silva

Exercício 1

Para demonstrar que:

$$E(QMTrat) = \sigma^2 + \frac{J}{I-1}\sum_i ti^2$$

Temos que:

$$QMTrat = SQTrat/I-1 = \frac{\frac{1}{j}\sum_i y_i^2 - \frac{1}{IJ}(\sum_{ij} y_{ij}^2)}{I-1}$$

Assim

$$E(QMTrat) = \frac{1}{J(I-1)}E[\sum_i(\sum_j y_{ij})^2 - \frac{1}{I}(\sum_{ij} y_{ij})^2]$$

$$= \frac{1}{J(I-1)}E[\sum_i(\sum_j m + t_i + e_{ij})^2 - \frac{1}{I}(\sum_{ij} m + t_i + e_{ij})^2]$$

$$= \frac{1}{J(I-1)}E[\sum_i(Jm + Jt_i + \sum_{ij} e_{ij})^2 - \frac{1}{I}(IJm + \sum_{ij} e_{ij})^2]$$

$$= \frac{1}{J(I-1)}E[\sum_i(J^2m^2+J^2t_i^2+(\sum_j e_{ij})^2+2J^2mt_i+2Jm\sum_j e_{ij}+wt_i\sum_j e_{ij})$$

$$-\frac{1}{I}(I^2J^2m^2 + 2IJm\sum_{ij} e_{ij} + (\sum_{ij} e_{ij})^2)]$$

$$= \frac{1}{J(I-1)}[J^2 \sum_i ti^2 + E[\sum_i (\sum_j e_{ij})^2] - E(\frac{1}{I}(\sum_{ij} e_{ij})^2)]$$

$$= \frac{1}{J(I-1)}[J^2 \sum_i t_i^2] + \sum_i E[e_{i1}^2 + e_{i2} + ... + e_{ij}^2) - \frac{1}{I}E(e_{i1}^2 + e_{i2}^2 + ... + e_{ij^2})]$$

$$= \frac{1}{J(I-1)}[J^2 \sum_i t_i^2 + \sum_i J\sigma^2 - \frac{1}{J} IJ\sigma^2]$$

$$\frac{1}{J(I-1)}[J^2 \sum_i t_i^2 + \sigma^2 J(I-1)]$$

$$= \sigma^2 + \frac{J}{I-1} \sum_i ti^2$$

Exercício 2

Dado as seguinte tabela de dados de 4 métodos distintos de estimar a frequência do fluxo de inundação de uma mesma bacia hidrográfica, deseja-se testar se os 4 métodos resultam estimativas equivalentes para essa bacia.

```
df_hidro_wider |>
kbl(
booktabs = TRUE,
escape = FALSE,
caption = "Valores Coletados"
) |>
kable_classic() |>
kable_styling(
font_size = 12,
latex_options = "HOLD_position"
)
```

Table 1: Valores Coletados

1	2	3	4
0.34	0.91	6.31	17.15
0.12	2.94	8.37	11.82
1.23	2.14	9.75	10.95
0.70	2.36	6.09	17.20
1.75	2.86	9.82	14.35
0.12	4.55	7.24	16.82

Considerando que se trata de um experimento inteiramente casualizado, utilizou-se o pacote *ExpDes* para construção do modelo e visualização e teste das suposições dos resíduos

```
model_1 = ExpDes::crd(df_hidro$método, df_hidro$valores)
```

Para um DIC, o modelo é dado por:

$$y_{ij} = m + t_{ij} + e_{ij}$$

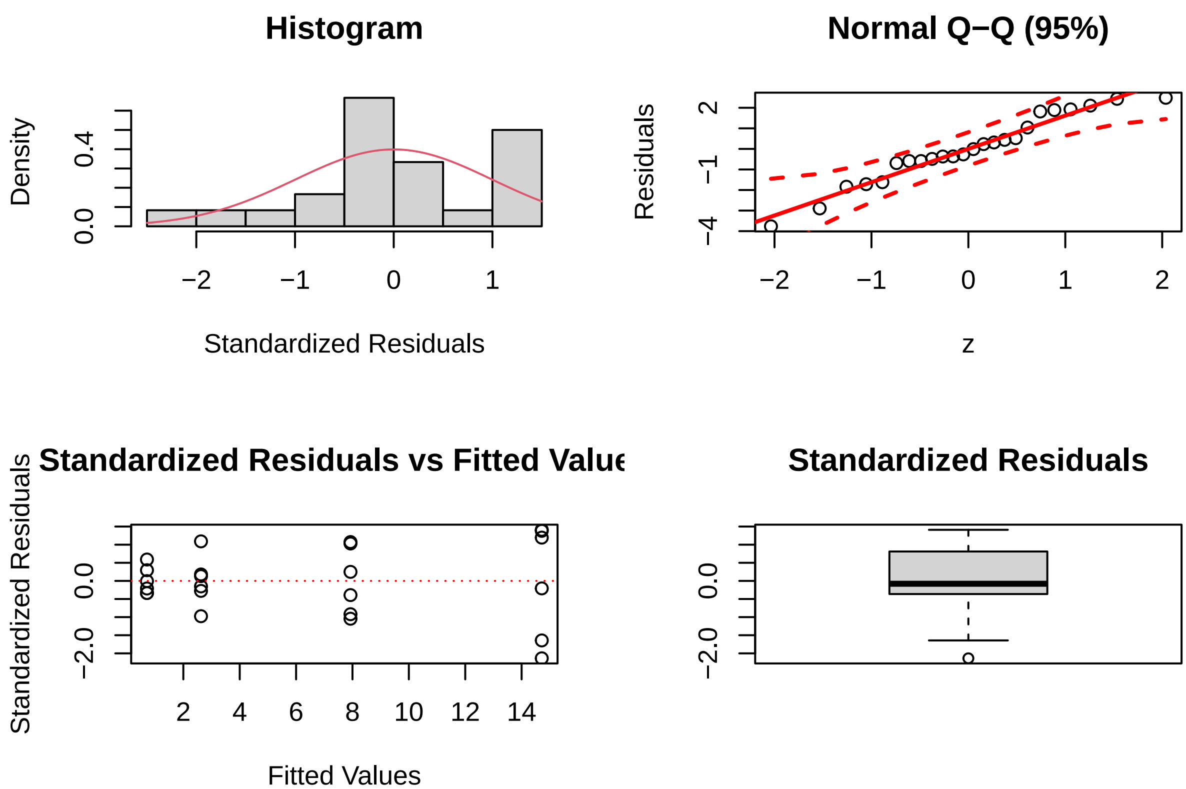
Análise de Resíduos

O modelo assume as seguintes pressuposições a respeito dos resíduos:

- Seguem distribuição Normal de média 0
- São Independentes
- São Homocedasticos

Primeiramente, verificou-se as suposições analisando os seguintes gráficos

```
ExpDes::plotres(model_1)
```



Analisando os gráficos, desenvolveu-se as hipóteses de:

- Pelo histograma e qqplot, os resíduos não fogem da hipótese de normalidade, testes não paramétricos foram aplicados para uma melhor conclusão
- Pelo gráfico de valores preditos x resíduos padronizados é possivel observar a forma de um funil indicando violação da hipótese de homocedasticidade

Para uma melhor conclusão sobre as suposições, utilizou-se testes não parametricos

Teste de Normalidade Para testar a normalidade dos resíduos, utilizou-se os seguintes testes:

Teste de Lilliefors, Teste de Cramer-Von Misses, Teste de Anderson-Darling

Onde os resultados foram anotados na seguinte tabela:

```
resi = model_1$residuals |>
  as.data.frame() |>
  dplyr::rename('residuals'='model_1$residuals')

test_resid = c(shapiro.test(resi$residuals) |> {\(x) x$p.value}(),
```

```r
                nortest::cvm.test(resi$residuals) |> {\(x) x$p.value}(),
                nortest::lillie.test(resi$residuals) |> {\(x) x$p.value}())
resid_values = data.frame('Teste' = c('Shapiro', 'Cramer', 'Lilliefors'),
                           'P-valor'= test_resid)

resid_values|>
kbl(
booktabs = TRUE,
escape = FALSE,
caption = "Teste para Normalidade dos resíduos"
) |>
kable_classic() |>
kable_styling(
font_size = 12,
latex_options = "HOLD_position"
)
```

Table 2: Teste para Normalidade dos resíduos

Teste	P.valor
Shapiro	0.3814231
Cramer	0.4432093
Lilliefors	0.3808247

Dado que os 3 testes não rejeitaram a hipótese de normalidade dos resíduos com 95% de confiança, é razoável considerar sua normalidade

Teste de Homecedasticidade Para testar a homecedasticidade dos resíduos, utilizou-se os seguintes testes:

Teste de Bartlett, Teste de Levene, Teste de Hartley

```r
test_resid = c(bartlett.test(g = df_hidro$método, x = model_1$residuals) |> {\(x) x$p.va
                car::leveneTest(y = model_1$residuals, group = df_hidro$método) |> {\(x)
                PMCMRplus::hartleyTest(x = model_1$residuals, g = df_hidro$método |> as.f
resid_values = data.frame('Teste' = c('Bartlett', 'Levene', 'Hartley'), 'P-valor'= test_

resid_values|>
kbl(
booktabs = TRUE,
escape = FALSE,
caption = "Teste para Homecedasticidade dos resíduos"
```

```
) |>
kable_classic() |>
kable_styling(
font_size = 12,
latex_options = "HOLD_position"
)
```

Table 3: Teste para Homecedasticidade dos resíduos

Teste	P.valor
Bartlett	0.0293467
Levene	0.0135706
Hartley	0.0276092

Os 3 testes rejeitaram a hipótese de homocedasticidade dos resíduos com 95% de confiança, e portanto, é razoável concluir que eles são heterocedasticos. Assim, uma transformação foi aplicada nos dados na tentativa de alterar esse cenário

Transformação Box Cox

Buscando resolver o problema de heterocedasticidade dos dados, aplicou-se a técnica de transformação de dados Box Cox

O valor λ é dado por:

$$\lambda = 1 - \frac{\beta_1}{2}$$

Onde, β_1 é dado por:

$$log(var(T_i)) = \beta_0 + \beta_1 log(m(T_i)) + c$$

```
list_hidro = list('1' = c(0.34, 0.12, 1.23, 0.7, 1.75, 0.12),
     '2' = c(0.91, 2.94, 2.14, 2.36, 2.86, 4.55),
     '3' = c(6.31, 8.37, 9.75, 6.09, 9.82, 7.24),
     '4' = c(17.15, 11.82, 10.95, 17.2, 14.35, 16.82))

box_cox_lambda = lm(list_hidro |> lapply(function(i) var(i) |> log()) |> unlist() ~
     list_hidro |> lapply(function(i) mean(i) |> log()) |> unlist() )
1- (box_cox_lambda$coefficients[2]/2) # aplicar raiz nos dados
```

```
## unlist(lapply(list_hidro, function(i) log(mean(i))))
##                                                0.5535286
```

Calculado $\beta_1 = 0.8929$, tem-se que $\lambda = 0.5535$, que diz que uma transformação raiz quadrada é indicada nos dados

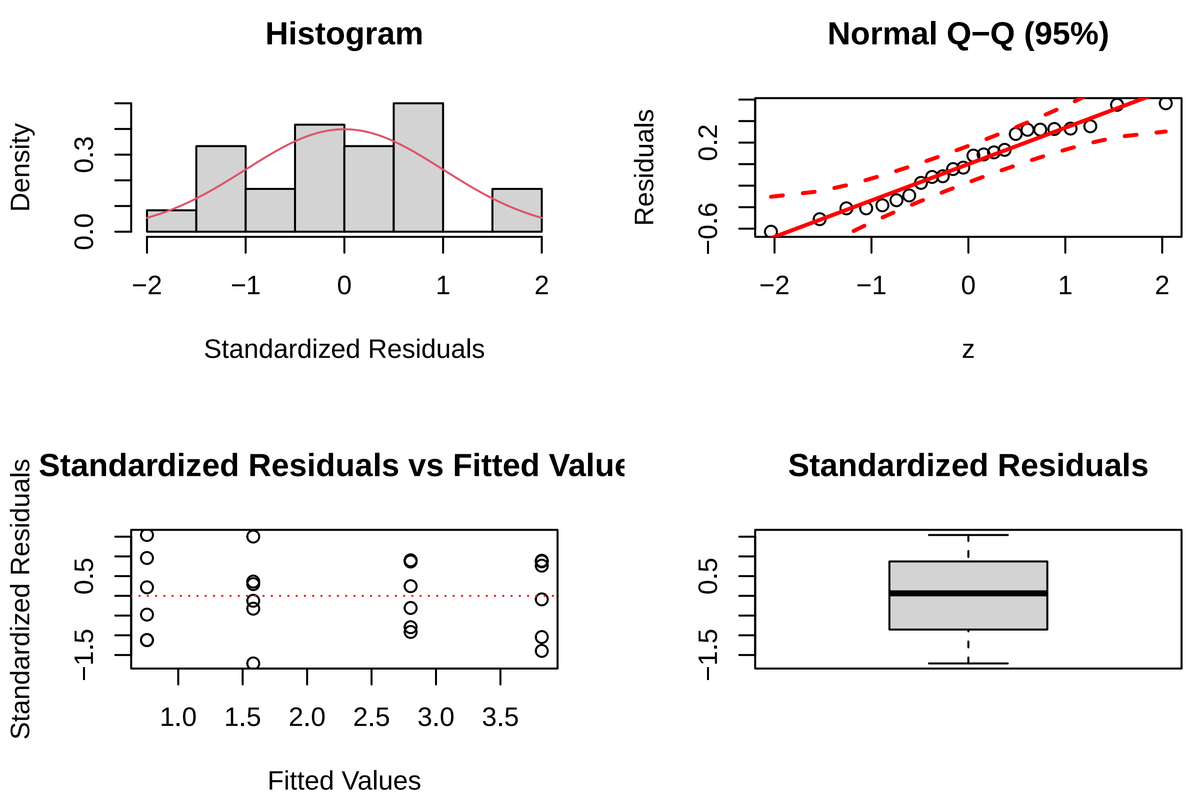
Análise dos Dados Transformados

Como indicado pelo valor λ da transformação Box-Cox, foi aplicada a função raiz quadrada nos dados buscando resolver a quebra a suposição de homocedasticidade dos resíduos.

```
df_hidro = df_hidro |>
  dplyr::mutate(valores = sqrt(valores))
model_2 = ExpDes::crd(df_hidro$método, df_hidro$valores)
```

Os resíduos do modelo com os dados transformados tiveram o seguinte comportamento

```
ExpDes::plotres(model_2)
```



Observando os gráficos dos resíduos, notou-se que a aparência de funil no gráfico de valores preditos x resíduos foi substituida por um retângulo, indicando a homocedasticidade dos resíduos.

Aplicando novamente os teste de normalidade e homocedasticidade, obteu-se os resulatdos:

Teste de Normalidade Para testar a normalidade dos resíduos, utilizou-se os seguintes testes:

Teste de Lilliefors, Teste de Cramer-Von Misses, Teste de Anderson-Darling

Onde os resultados foram anotados na seguinte tabela:

```
resi = model_2$residuals |>
  as.data.frame() |>
  dplyr::rename('residuals'='model_2$residuals')

test_resid = c(shapiro.test(resi$residuals) |> {\(x) x$p.value}(),
               nortest::cvm.test(resi$residuals) |> {\(x) x$p.value}(),
               nortest::lillie.test(resi$residuals) |> {\(x) x$p.value}())
resid_values = data.frame('Teste' = c('Shapiro', 'Cramer', 'Lilliefors'),
                          'P-valor'= test_resid)

resid_values|>
kbl(
booktabs = TRUE,
escape = FALSE,
caption = "Teste para Normalidade dos resíduos"
) |>
kable_classic() |>
kable_styling(
font_size = 12,
latex_options = "HOLD_position"
)
```

Table 4: Teste para Normalidade dos resíduos

Teste	P.valor
Shapiro	0.4141262
Cramer	0.4474381
Lilliefors	0.4013792

Dado que os 3 testes não rejeitaram a hipótese de normalidade dos resíduos com 95% de confiança, é razoável considerar sua normalidade

Teste de Homecedasticidade Para testar a homecedasticidade dos resíduos, utilizou-se os seguintes testes:

Teste de Bartlett, Teste de Levene, Teste de Hartley

```
test_resid = c(bartlett.test(g = df_hidro$método, x = model_2$residuals) |> {\(x) x$p.va
               car::leveneTest(y = model_2$residuals, group = df_hidro$método) |> {\(x)
               PMCMRplus::hartleyTest(x = model_2$residuals, g = df_hidro$método |> as.f
resid_values = data.frame('Teste' = c('Bartlett', 'Levene', 'Hartley'), 'P-valor'= test_

resid_values|>
kbl(
booktabs = TRUE,
escape = FALSE,
caption = "Teste para Homecedasticidade dos resíduos"
) |>
kable_classic() |>
kable_styling(
font_size = 12,
latex_options = "HOLD_position"
)
```

Table 5: Teste para Homecedasticidade dos resíduos

Teste	P.valor
Bartlett	0.9134261
Levene	0.8679782
Hartley	0.9017147

Dado que os 3 testes não rejeitaram a hipótese de homocedasticidade dos resíduos com 95% de confiança, é razoável considerar que são homocedasticos e portanto a transformação raiz quadrada nos dados resolvou o problema de heterocedasticidade

ANOVA

Buscando verificar sa houve diferença entre os métodos de estimação da frequência do fluxo de inundação de uma bacia hidrográfica, realizou-se uma ANOVA nos dados

Os cálculos foram feitos através das seguintes funções construidas manualmente:

- Função para cálculo dos Graus de Liberdade:

É dado por:

$$
\begin{aligned}
Trat &= L - 1 \\
Res &= L(J - 1)
\end{aligned}
$$

$$
Total = IL - 1
$$

```
gl_anova = function(data_wider){
  trat = ncol(data_wider) - 1
  res =  ncol(data_wider) * (nrow(data_wider) - 1)

  return(list('trat' = trat,
              'res' = res))
}
```

- Função para cálculo das Somas de Quadrados

É dado por:

$$
C = \frac{(\sum_{ij} y_{ij})^2}{IJ}
$$

$$
SQTotal = \sum_{i,j} y_{ij}^2 - C
$$

$$
SQTrat = \frac{1}{J} \sum_{i} T_i^2 - C
$$

$$
SQRes = SQTotal - SQTrat
$$

```
sq_anova = function(data_wider){

  sum_sq = sum(data_wider**2)
  correction = (sum(data_wider)**2) / (ncol(data_wider) * nrow(data_wider))
  sq_total = sum_sq - correction


  sum_ti = data_wider |>
    apply(2, sum)
```

```
  sq_trat = (sum(sum_ti**2)/nrow(data_wider)) - correction

sq_res = sq_total - sq_trat

return(list('sq_total' = sq_total,
            'sq_trat' = sq_trat,
            'sq_res' = sq_res))
}
```

- Função para cálculo dos Quadrados Médios e Teste F

É dado por:

$$QMTrat = \frac{SQTrat}{gl(Trat)}$$

$$QMRes = \frac{SQRes}{gl(Res)}$$

Com os dois valores calculados é possível realizar um teste F, que indica diferença entre os tratamentos testados. Sua estatística de teste é dado por:

$$\frac{QMTrat}{QMRes} \sim F_{(I-1),(J-1)}$$

```
qm_anova = function(sq_trat, sq_res, gl_trat, gl_res){
  qm_trat = sq_trat/gl_trat
  qm_res = sq_res/gl_res

  f_statistic = qm_trat/qm_res
  p_value = 1 - pf(f_statistic, gl_trat, gl_res)

  return(list('qm_trat' = qm_trat,
              'qm_res' = qm_res,
              'f_statistic' = f_statistic,
              'p_value' = p_value))
}
```

Com a funções definidas, foi possível construir a tabela da ANOVA

```
df_hidro_wider = df_hidro |>
  dplyr::group_by(método) |>
  dplyr::mutate(row = dplyr::row_number()) |>
  tidyr::pivot_wider(names_from = método, values_from = valores) |>
  dplyr::select(-row)

gls = gl_anova(df_hidro_wider)
sqs = sq_anova(df_hidro_wider)
qm_f = qm_anova(sqs$sq_trat, sqs$sq_res, gls$trat, gls$res)
```

Construindo a tabela apartir dos valores calculados pelas funções

```
data.frame('G.L' = c(gls[[1]], gls[[2]]),
           'S.Q' = c(sqs[[2]],sqs[[3]]),
           'Q.M' = c(qm_f[[1]], qm_f[[2]]),
           'Valor F' = c(qm_f[[3]],0),
           'P-Valor' = c(qm_f[[4]],0)
) |>
kbl(
booktabs = TRUE,
escape = FALSE,
caption = "ANOVA"
) |>
kable_classic() |>
kable_styling(
font_size = 12,
latex_options = "HOLD_position"
)
```

Table 6: ANOVA

G.L	S.Q	Q.M	Valor.F	P.Valor
3	32.684213	10.8947376	81.04899	0
20	2.688433	0.1344216	0.00000	0

Sabendo que o teste F utilizado possui as seguintes hipóteses:

$$H_0 : \text{Tratamentos não possuem diferenças}$$
$$H_1 : \text{Pelo menos 1 tratamento é diferente de outro tratamento}$$

Analisando o p-valor,onde foi calculando um valor aproximadamente 0, rejeita-se a hipótese de igualdade de tratamento através do teste, com significância de 5%, ou seja, podemos afirmar que há pelo menos um método de estimação de frequência do fluxo de inundação que é diferente de outro método com confiança de 95%