Análise Multivariada I

Débora dos Santos Farias

06 junho, 2021

Teste MANOVA em R:

Na análise multivariada de variância (MANOVA), mais de uma variável dependente é avaliada simultaneamente.

Pressupostos da MANOVA:

- Normalidade multivariada;
- Homogeneidade das matrizes covariâncias-variâncias entre os grupos(testar M de Box);
- Se há outliers multivariados;
- Linearidade entre os pares de observações;

Vantagens:

- MANOVA permite que olhemos para os dados de forma multivariada;
- Pode ser uma ferramenta com maior poder para detectar alguns efeitos;
- Além disso, como fazemos um único teste, isso diminui a chance de cometer erros do tipo I (rejeita-se $H0$ quando na realidade é verdadeiro)

Metodologia:

As análises foram feitas com algumas bibliotecas (pacotes) disponíveis no *software* R, que serão esmiuçado mais à frente. O banco de dados contém informações oriundas do exemplo de MANOVA usando dados de Tstsuoka, 1988, p. 274. Onde os indíviduos foram submetidos a testes em que primeiro comparou a existência de feitos dos dois métodos de ensino de taquigrafia (método A e método B) sobre a combinação da precisão e da velocidade (variáveis dependentes) e depois foram medidas as três condições de prática (C1 = 2 horas de instrução por dia durante 6 semanas; C2 = 3 horas de instrução por dia durante 4 semanas; C3 = 4 horas de instrução por dia durante 3 semanas) sobre a análise simultaneamente das variáveis dependentes.

Pacotes necessários:

```
library(tidyverse)
library(dbplyr) # manipulação dos dados
library(rstatix) # para análise fáceis
library(emmeans) # comparação entre pares
library(MVN) # normalidade multivariada
library(GGally) # analisar a linearidade, fazer uma matriz
```

Tabela inicial:

```
##    metodo praticar velocidade precisao
## 1       A       C1         36       26
## 2       A       C1         34       22
## 3       A       C1         28       21
## 4       A       C1         34       23
## 5       A       C1         34       21
## 6       A       C1         29       19
## 7       A       C1         48       25
## 8       A       C1         28       20
## 9       A       C1         34       21
## 10      A       C1         38       20
## 11      A       C2         46       17
## 12      A       C2         34       21
## 13      A       C2         31       17
## 14      A       C2         31       18
## 15      A       C2         36       23
## 16      A       C2         26       19
## 17      A       C2         35       16
## 18      A       C2         33       19
## 19      A       C2         23       15
## 20      A       C2         30       14
## 21      A       C3         26       14
## 22      A       C3         31       14
## 23      A       C3         30       16
## 24      A       C3         34       16
## 25      A       C3         30       13
## 26      A       C3         27       13
## 27      A       C3         21       12
## 28      A       C3         31       15
## 29      A       C3         37       14
## 30      A       C3         29       14
## 31      B       C1         42       25
## 32      B       C1         47       24
## 33      B       C1         51       29
## 34      B       C1         35       25
## 35      B       C1         37       26
## 36      B       C1         44       28
## 37      B       C1         44       25
## 38      B       C1         49       24
## 39      B       C1         43       24
## 40      B       C1         36       26
```

```
## 41      B       C2          32       18
## 42      B       C2          39       19
## 43      B       C2          37       17
## 44      B       C2          31       17
## 45      B       C2          36       19
## 46      B       C2          32       19
## 47      B       C2          31       17
## 48      B       C2          41       21
## 49      B       C2          36       18
## 50      B       C2          40       20
## 51      B       C3          28       11
## 52      B       C3          28       10
## 53      B       C3          25       10
## 54      B       C3          22       12
## 55      B       C3          27       11
## 56      B       C3          25       12
## 57      B       C3          33       14
## 58      B       C3          31       13
## 59      B       C3          28       12
## 60      B       C3          23       11
```

Verificação da normalidade multivariada

Teste de Henze-Zirkler - pacote MVN: verifica normalidade uni e multivariada

```
mvn(data = dados[,c(1,3,4)], subset = "metodo", mvnTest = "hz")
```

```
## $multivariateNormality
## $multivariateNormality$A
##            Test        HZ   p value MVN
## 1 Henze-Zirkler 0.6590596 0.1276548 YES
##
## $multivariateNormality$B
##            Test        HZ   p value MVN
## 1 Henze-Zirkler 0.6268565 0.1564967 YES
##
##
## $univariateNormality
## $univariateNormality$A
##           Test   Variable Statistic   p value Normality
## 1 Shapiro-Wilk velocidade    0.9351    0.0671    YES
## 2 Shapiro-Wilk  precisao     0.9518    0.1892    YES
##
## $univariateNormality$B
##           Test   Variable Statistic   p value Normality
## 1 Shapiro-Wilk velocidade    0.9754    0.6944    YES
## 2 Shapiro-Wilk  precisao     0.9269    0.0407    NO
##
##
## $Descriptives
## $Descriptives$A
##              n     Mean Std.Dev Median Min Max  25th 75th      Skew  Kurtosis
## velocidade 30 32.13333  5.7038   31.0  21  48 29.00   34 0.7509630  1.074564
```

```
## precisao   30 17.93333  3.8231   17.5  12  26 14.25   21 0.3070827 -1.045536
##
## $Descriptives$B
##             n     Mean  Std.Dev Median Min Max  25th  75th       Skew
## velocidade 30 35.10000 7.774184   35.5  22  51 28.75 40.75 0.21702839
## precisao   30 18.56667 5.980908   18.5  10  29 12.25 24.00 0.06357061
##              Kurtosis
## velocidade -0.9309418
## precisao   -1.4040533
```

Pelo valor de $p > 0.05$ há evidências de normalidade multivariada para todos os grupos, isso ao nível de 5% de significância.

```
mvn(data = dados[,2:4], subset = "praticar", mvnTest = "hz")
```

```
## $multivariateNormality
## $multivariateNormality$C1
##            Test        HZ   p value MVN
## 1 Henze-Zirkler 0.5192989 0.2220955 YES
##
## $multivariateNormality$C2
##            Test        HZ   p value MVN
## 1 Henze-Zirkler 0.2029432 0.9412597 YES
##
## $multivariateNormality$C3
##            Test        HZ  p value MVN
## 1 Henze-Zirkler 0.2071513 0.935054 YES
##
##
## $univariateNormality
## $univariateNormality$C1
##           Test   Variable Statistic   p value Normality
## 1 Shapiro-Wilk velocidade    0.9401    0.2413    YES
## 2 Shapiro-Wilk  precisao     0.9574    0.4939    YES
##
## $univariateNormality$C2
##           Test   Variable Statistic   p value Normality
## 1 Shapiro-Wilk velocidade    0.9741    0.8387    YES
## 2 Shapiro-Wilk  precisao     0.9690    0.7341    YES
##
## $univariateNormality$C3
##           Test   Variable Statistic   p value Normality
## 1 Shapiro-Wilk velocidade    0.9852    0.9827    YES
## 2 Shapiro-Wilk  precisao     0.9487    0.3477    YES
##
##
## $Descriptives
## $Descriptives$C1
##             n  Mean  Std.Dev Median Min Max 25th  75th       Skew  Kurtosis
## velocidade 20 38.55 7.037307   36.5  28  51   34 44.00 0.18551728 -1.238998
## precisao   20 23.70 2.754900   24.0  19  29   21 25.25 0.01033085 -1.040669
##
## $Descriptives$C2
```

```
##             n Mean  Std.Dev Median Min Max 25th  75th      Skew    Kurtosis
## velocidade 20 34.0 5.241535   33.5  23  46   31 36.25 0.1499956 -0.04969503
## precisao   20 18.2 2.117595   18.0  14  23   17 19.00 0.1933498 -0.22464690
##
## $Descriptives$C3
##             n  Mean  Std.Dev Median Min Max  25th 75th      Skew   Kurtosis
## velocidade 20 28.30 4.040584     28  21  37 25.75   31 0.1063245 -0.5756776
## precisao   20 12.85 1.785173     13  10  16 11.75   14 0.1082340 -1.0493990
```

Portanto há indícios de normalidade multivariada para todos os grupos, ao nível de 5% de siginificância.

Verificação de outliers Multivariados:

Pela distância de Mahalanobis (outliers = $p < 0.001$) - pacote rstatix

```
dados  %>% select(c(1,3,4)) %>% group_by(metodo) %>%
  doo(~mahalanobis_distance(.)) %>%
  filter(is.outlier == TRUE)
```

```
## # A tibble: 0 x 5
## # ... with 5 variables: metodo <fct>, velocidade <int>, precisao <int>,
## #   mahal.dist <dbl>, is.outlier <lgl>
```

Pode-se observar que não há outliers multivariados, pelo distância de Mahalanobis.

```
dados %>% select(2:4) %>% group_by(praticar) %>%
  doo(~mahalanobis_distance(.)) %>%
  filter(is.outlier == TRUE)
```

```
## # A tibble: 0 x 5
## # ... with 5 variables: praticar <fct>, velocidade <int>, precisao <int>,
## #   mahal.dist <dbl>, is.outlier <lgl>
```

Percebe-se que não indícios de outliers multivariados, avaliados pela distância de Mahalanobis.

Verificação da presença de outliers univariados - pacote rstatix - por grupo:

```
dados %>% group_by(metodo) %>%
  identify_outliers(velocidade)
```

```
## # A tibble: 3 x 6
##   metodo praticar velocidade precisao is.outlier is.extreme
##   <fct>  <fct>         <int>    <int> <lgl>      <lgl>
## 1 A      C1               48       25 TRUE       FALSE
## 2 A      C2               46       17 TRUE       FALSE
## 3 A      C3               21       12 TRUE       FALSE
```

Pode-se observar que não há outliers extremos univariados na variável velocidade.

```
dados %>% group_by(metodo) %>%
  identify_outliers(precisao)
```

```
## [1] metodo     praticar   velocidade precisao   is.outlier is.extreme
## <0 rows> (or 0-length row.names)
```

Não houve outliers univariados na variável precisão.

```
dados %>% group_by(praticar) %>%
  identify_outliers(velocidade)
```

```
## # A tibble: 2 x 6
##   praticar metodo velocidade precisao is.outlier is.extreme
##   <fct>    <fct>       <int>    <int> <lgl>      <lgl>
## 1 C2       A              46       17 TRUE       FALSE
## 2 C2       A              23       15 TRUE       FALSE
```

```
dados %>% group_by(praticar) %>%
  identify_outliers(precisao)
```

```
## # A tibble: 1 x 6
##   praticar metodo velocidade precisao is.outlier is.extreme
##   <fct>    <fct>       <int>    <int> <lgl>      <lgl>
## 1 C2       A              36       23 TRUE       FALSE
```

Percebe-se que não há indicativos de outliers extremos univariados nas variáveis velocidade e precisão.

Verificação da homogeneidade da suposição de covariâncias (pacote rstatix):

O Teste M de Box pode ser usado para verificar a igualdade de covariância entre os grupos. Isso é o equivalente a uma homogeneidade de variância multivariada. Este teste é considerado altamente sensível. Portanto, a significância para este teste é determinada em $\alpha = 0.001$.

$H0$: as matrizes de variâncias-covariâncias são homogêneas

$H1$: as matrizes de variâncias-covariâncias não são homogêneas

```
box_m(dados[ ,3:4], dados$metodo)
```

```
## # A tibble: 1 x 4
##   statistic p.value parameter method
##       <dbl>   <dbl>     <dbl> <chr>
## 1      9.05  0.0286         3 Box's M-test for Homogeneity of Covariance Matric~
```

Como $p - valor$ é maior que 0.001, nesse caso há indícios que essas variâncias são homogêneas.

```
box_m(dados[ ,3:4], dados$praticar)
```

```
## # A tibble: 1 x 4
##   statistic p.value parameter method
##       <dbl>   <dbl>     <dbl> <chr>
## 1      9.45   0.150         6 Box's M-test for Homogeneity of Covariance Matric~
```

Pelo o valor de p maior que 0.001, pode-se considerar que existe homogeneidade das matrizes de variâncias e covariâncias.

Verificação da homogeneidade das variâncias - teste de Levene (pacote rstatix):

$H0$: As variâncias são homogêneas

$H1$: As variâncias não são homogêneas

Procedimento:

1. Reúna as variáveis de resultado em pares de valores-chave;
2. Agrupar por variável;
3. Calcule o teste de Levene

```
dados %>%
  gather(key = "variable", value = "value", velocidade,precisao) %>%
  group_by(variable) %>%
  levene_test(value ~ metodo)
```

```
## # A tibble: 2 x 5
##   variable     df1   df2 statistic      p
##   <chr>      <int> <int>     <dbl>  <dbl>
## 1 precisao       1    58      7.02 0.0104
## 2 velocidade     1    58      4.29 0.0427
```

Como $p < 0.05$, portanto não há indícios de homogeneidade das variâncias.

```
dados %>%
  gather(key = "variable", value = "value", velocidade,precisao) %>%
  group_by(variable) %>%
  levene_test(value ~ praticar)
```

```
## # A tibble: 2 x 5
##   variable     df1   df2 statistic      p
##   <chr>      <int> <int>     <dbl>  <dbl>
## 1 precisao       2    57      1.74 0.184
## 2 velocidade     2    57      3.00 0.0578
```

Percebe-se que o p foi maior que 0.05, então pode-se considerar as variâncias iguais.

Verificação da presença de multicolinearidade ($r > 0.9$) - pacote rstatix:

Idealmente, a correlação entre as variáveis de resultado deve ser moderada, não muito alta. Uma correlação acima de 0.9 é uma indicação de multicolinearidade, o que é problemático para MANOVA.

```
dados %>% cor_test(velocidade, precisao)
```

```
## # A tibble: 1 x 8
##   var1       var2       cor statistic        p conf.low conf.high method
##   <chr>      <chr>    <dbl>     <dbl>    <dbl>    <dbl>     <dbl> <chr>
## 1 velocidade precisao  0.75      8.61 5.89e-12    0.611     0.843 Pearson
```

Portanto não houve multicolinearidade, avaliada pela correlação de Pearson ($r = 0,75$).

OBS: Na situação em que você tem multicolinearidade, pode considerar a remoção de uma das variáveis de resultado que está altamente correlacionada.

Verificação de linearidade (pacote GGally):

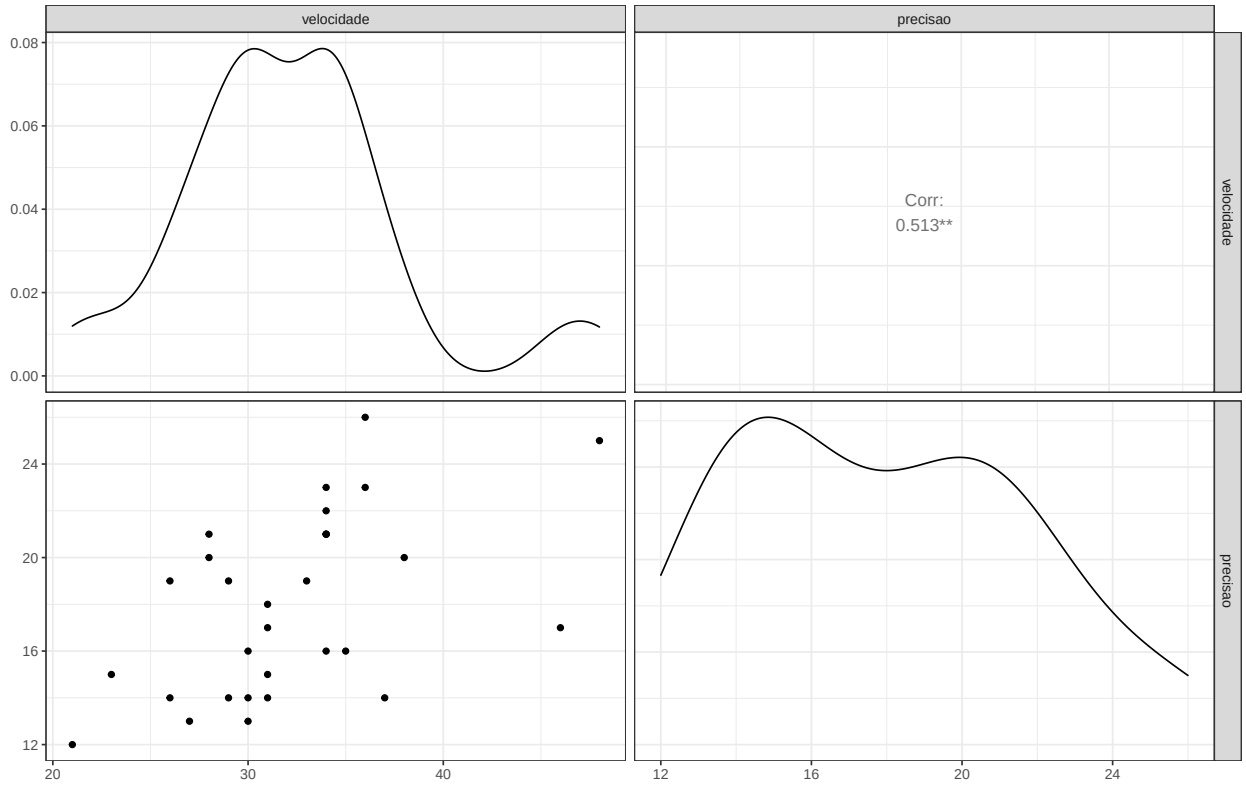
A relação das variáveis dependentes deve ser linear para cada grupo.

```
# Uma matriz de gráfico de dispersão por grupo
results <- dados %>%
  select(velocidade, precisao, metodo) %>%
  group_by(metodo) %>%
  doo(~ggpairs(.) + theme_bw(), result = "plots")
results
```

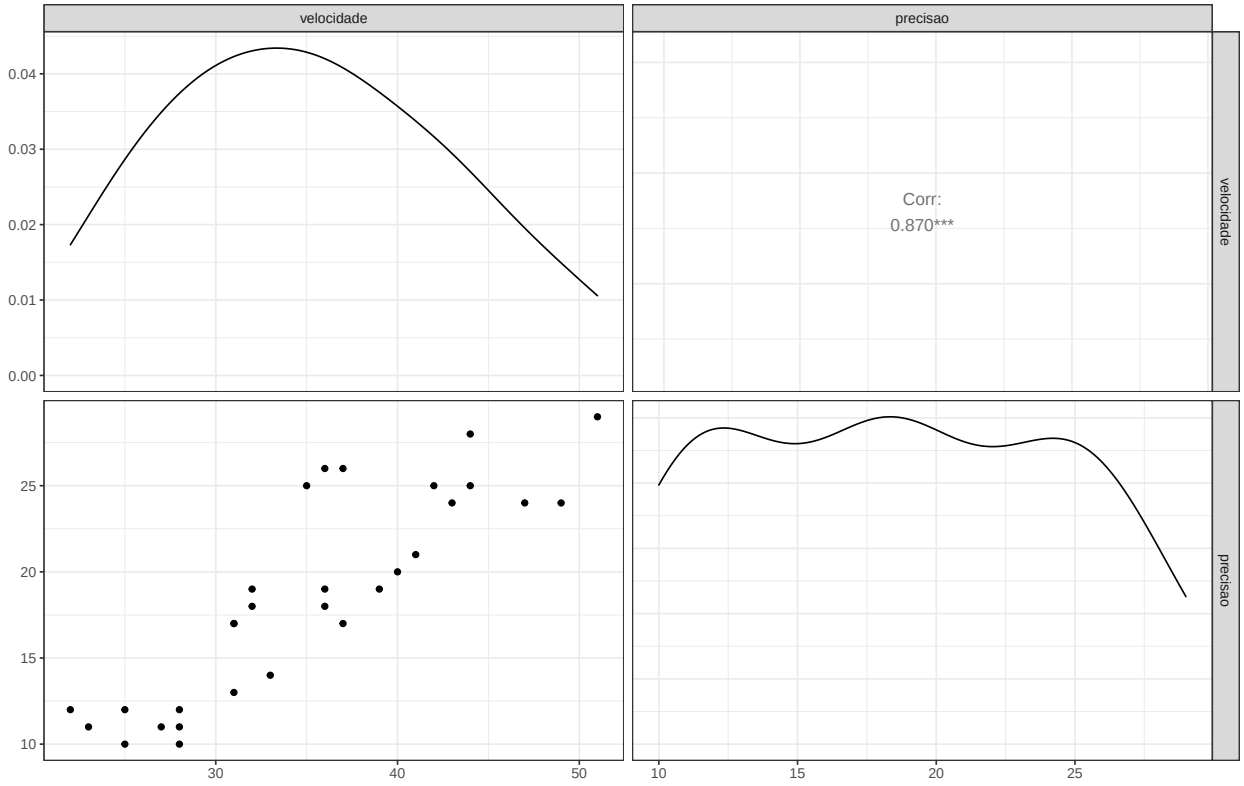
```
## # A tibble: 2 x 2
##   metodo plots
##   <fct>  <list>
## 1 A      <gg>
## 2 B      <gg>
```

```
results$plots
```

```
## [[1]]
```




```
##
## [[2]]
```



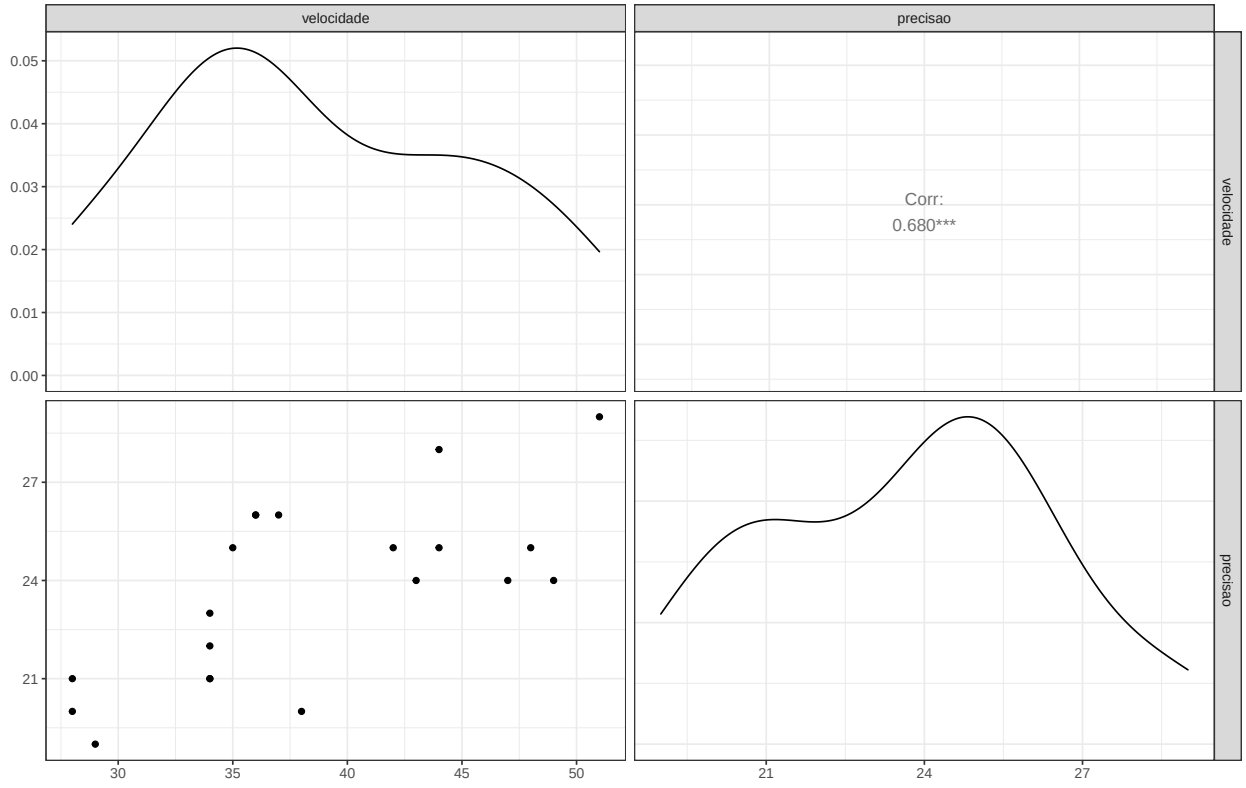
Há indícios de relação linear entre velocidade e precisão em cada grupo de método, conforme avaliado pelo gráfico de dispersão.

```
# Uma matriz de gráfico de dispersão por grupo
results1 <- dados %>%
  select(velocidade, precisao, praticar) %>%
  group_by(praticar) %>%
  doo(~ggpairs(.) + theme_bw(), result = "plots")
results
```

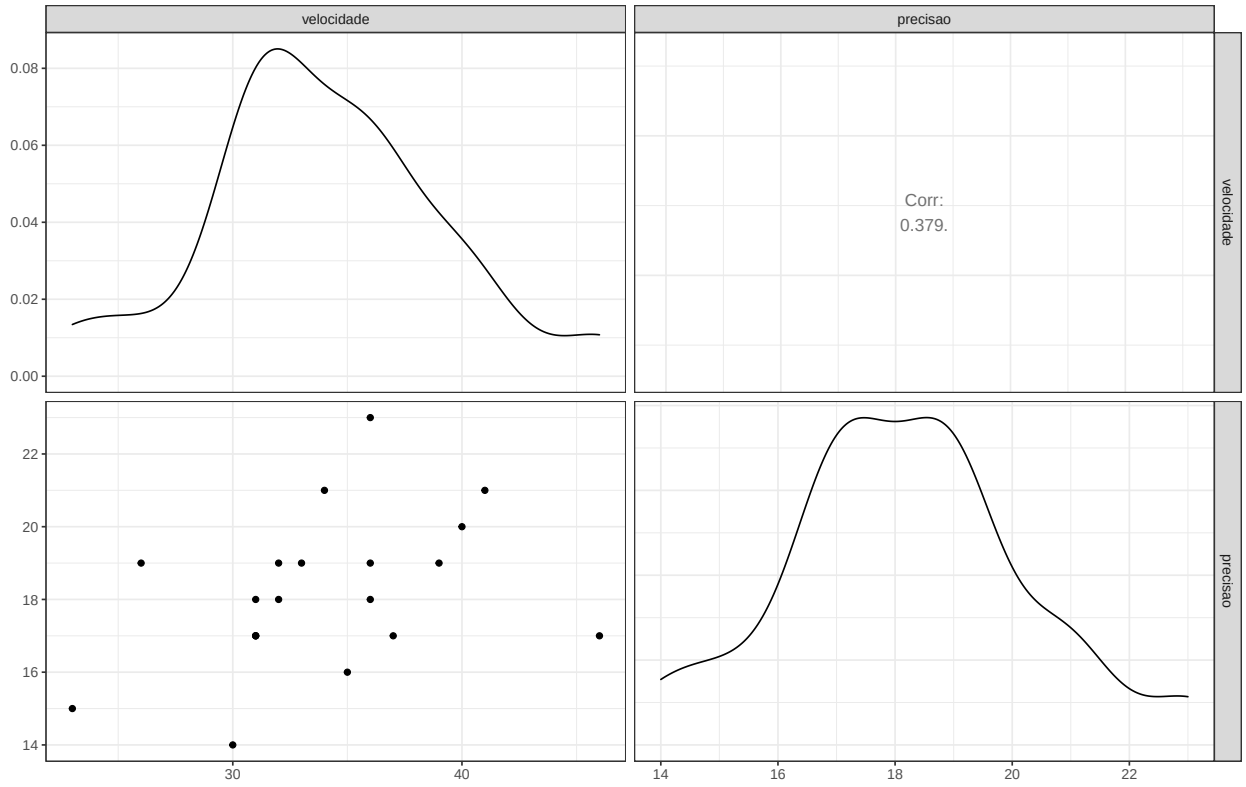
```
## # A tibble: 2 x 2
##   metodo plots
##   <fct>  <list>
## 1 A      <gg>
## 2 B      <gg>
```

```
results1$plots
```

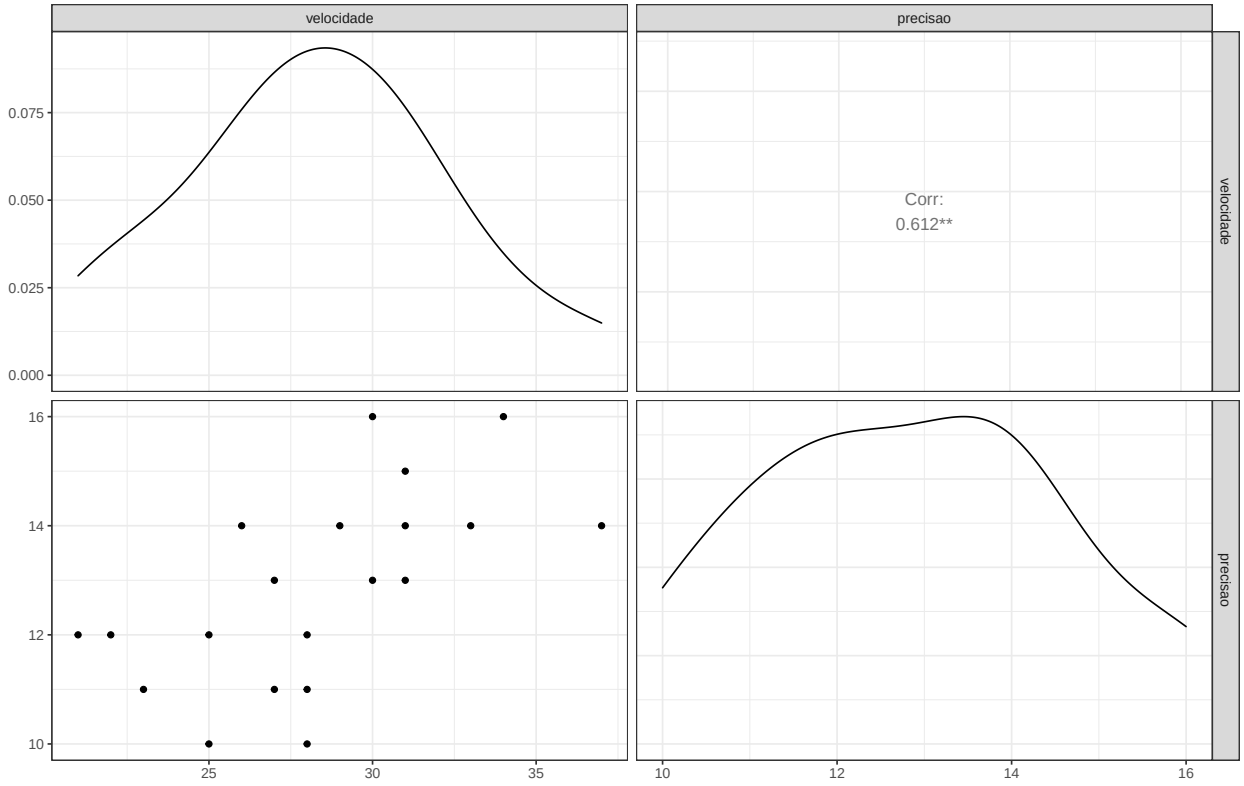
```
## [[1]]
```


velocidade
precisao
Corr:
0.680***
velocidade
precisao
0.05
0.04
0.03
0.02
0.01
0.00
27
24
21
30
35
40
45
50
21
24
27

```
##
## [[2]]
```


velocidade
precisao
Corr:
0.379.
velocidade
precisao
0.08
0.06
0.04
0.02
0.00
22
20
18
16
14
30
40
14
16
18
20
22

```
##
## [[3]]
```



Portanto há relação linear entre velocidade e precisão em cada grupo de prática, segundo avaliado pelo gráfico de dispersão.

Modelo de MANOVA:

Existem quatro tipos diferentes de estatísticas multivariadas que podem ser usadas para calcular MANOVA. São eles: *Pillai*, *Wilks*, *Hotelling-Lawley* ou *Roy*.

```
modelo<-manova(cbind(velocidade,precisao) ~ metodo, data=dados )
summary(modelo, test = "Pillai")
```

```
##           Df  Pillai approx F num Df den Df Pr(>F)
## metodo     1 0.06846   2.0945      2     57 0.1325
## Residuals 58
```

Então não há evidências de efeitos entre os métodos nas variáveis dependentes combinadas (velocidade, precisão).

```
modelo1<-manova(cbind(velocidade,precisao) ~ praticar, data=dados )
summary(modelo1,test = "Pillai")
```

```
##           Df  Pillai approx F num Df den Df    Pr(>F)
## praticar   2 0.81555   19.624      4    114 2.608e-12 ***
## Residuals 57
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Portanto houve efeito entre as práticas nas variáveis dependentes combinadas (velocidade, precisão).

ANOVA univariada:

Uma MANOVA estatisticamente significativa pode ser seguida por ANOVA univariada examinando separadamente, cada variável dependente. O objetivo é analisar as variáveis dependentes específicas que contribuíram para o efeito global significativo.

OBS: `welch_anova_test()`: pode ser usado quando a suposição de homogeneidade da variância é violada, como para esse exemplo.

Procedimento:

1. Reúna as variáveis de resultado em pares de valores-chave;
2. Agrupar por variável;
3. Teste de cálculo ANOVA unilateral

```
mode <- dados %>%
  gather(key = "variable", value = "value", velocidade,precisao) %>%
  group_by(variable)

mode %>% welch_anova_test(value ~ metodo)
```

```
## # A tibble: 2 x 8
##   variable   .y.       n statistic   DFn   DFd     p method
## * <chr>      <chr> <int>     <dbl> <dbl> <dbl> <dbl> <chr>
## 1 precisao   value    60      0.24     1  49.3 0.627 Welch ANOVA
## 2 velocidade value    60      2.84     1  53.2 0.098 Welch ANOVA
```

Portanto também não há evidências de efeitos dos métodos sobre cada uma das variáveis dependentes.

```
mode1 <- dados %>%
  gather(key = "variable", value = "value", velocidade,precisao) %>%
  group_by(variable)

mode1 %>% welch_anova_test(value ~ praticar)
```

```
## # A tibble: 2 x 8
##   variable   .y.       n statistic   DFn   DFd        p method
## * <chr>      <chr> <int>     <dbl> <dbl> <dbl>    <dbl> <chr>
## 1 precisao   value    60     114       2  37.0 1.54e-16 Welch ANOVA
## 2 velocidade value    60      18.1     2  36.3 3.46e- 6 Welch ANOVA
```

Então há indícios de efeitos das práticas sobre cada uma das variáveis dependentes.

Teste de comparações multiplas:

Uma ANOVA univariada estatisticamente significativa pode ser seguida por múltiplas comparações de pares para determinar quais grupos são diferentes.

Pela médias marginais estimadas (Pacote emmeans)

```
dados %>% emmeans_test(velocidade ~ praticar, p.adjust.method = "bonferroni")
```

```
## # A tibble: 3 x 9
##   term    .y.     group1 group2    df statistic         p     p.adj p.adj.signif
## * <chr>   <chr>   <chr>  <chr>  <dbl>     <dbl>     <dbl>     <dbl> <chr>
## 1 pratic~ veloci~ C1     C2        57      2.58   1.25e-2   3.75e-2 *
## 2 pratic~ veloci~ C1     C3        57      5.81   2.92e-7   8.77e-7 ****
## 3 pratic~ veloci~ C2     C3        57      3.23   2.05e-3   6.14e-3 **
```

Portanto há evidências que o grupo $C1$ e $C2$, $C1$ e $C3$, $C2$ e $C3$ diferem estatisticamente entre si, pelo teste de Bonferroni, ao nível de 5% de significância, para a variável dependente velocidade.

```
dados %>% emmeans_test(precisao ~ praticar, p.adjust.method = "bonferroni")
```

```
## # A tibble: 3 x 9
##   term     .y.      group1 group2    df statistic        p    p.adj p.adj.signif
## * <chr>    <chr>    <chr>  <chr>  <dbl>     <dbl>    <dbl>    <dbl> <chr>
## 1 praticar precisao C1     C2        57      7.71 2.07e-10 6.22e-10 ****
## 2 praticar precisao C1     C3        57     15.2  9.99e-22 3.00e-21 ****
## 3 praticar precisao C2     C3        57      7.50 4.65e-10 1.39e- 9 ****
```

Então há indícios que o grupo $C1$ e $C2$, $C1$ e $C3$, $C2$ e $C3$ diferem estatisticamente entre si, pelo teste de Bonferroni, ao nível de 5% de significância, para a variável dependente precisão.

Conclusão:

A partir dos resultados obtidos, pode-se concluir que a MANOVA mostrou que não há efeito dos métodos(A,B) sobre a velocidade e a precisão [Traço de Pillai $= 0.06846; F(2, 57) = 2.0945; p > 0.1325$], já para as práticas (C1,C2,C3) sobre a velocidade e a precisão mostrou efeitos [Traço de Pillai $= 0.81555; F(4, 114) = 19.624$]. ANOVA univariadas subsequentes mostraram que não há de efeitos dos métodos sobre cada uma das variáveis dependentes, e para as práticas sobre cada uma das variáveis dependentes mostrou efeitos significativos. O teste de Bonferroni mostrou que há diferenças entre C1 e os demais tanto para velocidade quanto para precisão.