MDR 573- - Lista de questões - testes não paramétricos

AUTHOR
Bruno F. Guedes

PUBLISHED
April 30, 2024

```r
#load packages
suppressPackageStartupMessages(library(ggplot2))
suppressPackageStartupMessages(library(dplyr))
suppressPackageStartupMessages(library(tidyr))
suppressPackageStartupMessages(library(purrr))
```

Questão 1:

Questão 1 / 10

```
    A mediana NÃO é equivalente a:
A   2-quantil
B   2º quartil
C   50º percentil
D   5º decil
E   50º postil
```

Resposta

A mediana é um quantil. Tanto faz em quantas partes dividimos o todo, o importante é ser um quantil que separa a metade dos pontos para cada lado, todos os quantis que equivalem a um corte de 50% estão certos. “Postil” é uma palavra que nem consegui localizar na internet, e, se corresponder a um posto específico, certamente não tem nenhuma garantia de um postil de 50 corresponder à mediana. Por exemplo, se os nossos dados são `1:50` , o postil 50 equivale a um percentil de 100%, bem diferente da mediana. Um posto de 50 so corresponderia à mediana se tivessemos 100 pontos.

alternativa E

Questão 2:

Um pesquisador da área médica deseja estudar o efeito de um novo tratamento em pacientes com determinada doença em relação a dois tratamentos convencionais. Em cada um dos três tratamentos foram alocados aleatoriamente dez pacientes distintos. A variável de desfecho com distribuição normal é o número de dias sem fortes desconfortos no período de um ano de tratamento. O teste estatístico mais adequado para o estudo é:

```
A   ANOVA de Welch
B   U de Mann-Whitney
C   H de Kruskal-Wallis
D   Q de Friedman
E   W de Wilcoxon
```

Resposta

Comparação entre-grupos com 3 ou mais grupos com variavel desfecho normal - o teste a ser utilizado é o ANOVA de Welch. O H de Kruskal-Wallis também poderia ser usado aqui, mas não é necessario usar teste não-paramétrico aqui, uma vez que os dados são normais.

Alternativa A

Questão 3:

Um pesquisador da área médica deseja estudar o efeito de um novo tratamento em pacientes com determinada doença em relação ao tratamento convencional. Dez pacientes foram alocados aleatoriamente ao tratamento novo e outros dez pacientes foram alocados ao tratamento convencional. A variável de desfecho com distribuição normal é o número de dias sem fortes desconfortos no período de um ano de tratamento. O teste estatístico mais adequado para o estudo é:

```
A   z
B   t
C   U
D   H
E   Q
F   W
```

Resposta

Aqui temos dois grupos independentes, nos quais comparamos variavel de distribuição normal. Teste t é o teste de escolha.

Alternativa B

Questão 4:

Os testes estatísticos de associação mais adequados entre duas variáveis nominais, duas ordinais e duas intervalares são, respectivamente:

```
A   Qui-quadrado de Pearson, qui-quadrado de Pearson e qui-quadrado de Pearson
B   Qui-quadrado de Pearson, qui-quadrado de Pearson e correlação de Pearson
C   Correlação de Pearson, correlação de Spearman e qui-quadrado de Pearson
D   Correlação de Spearman, qui-quadrado de Pearson e correlação de Pearson
E   Qui-quadrado de Pearson, correlação de Spearman e correlação de Pearson
```

Resposta

Aqui acho que é ligeiramente pegadinha. Porque a palavra usada é “associação”. Para testar associação entre duas variaveis nominais, usamos Qui quadrado. Para duas variaveis ordinais, embora existam testes de correlação, o teste ideal para testar “associação” segue sendo o Qui quadrado. Pra avaliar variaveis intervalares, usando correlaçã de Pearson. Então fico em duvida entre o pedantismo de dizer que "associação é associação", para isso somente qui-quadrado, ou a interpretação mais relaxada do que é associação, que levaria à resposta mais natural com a altenrativa E. Uma coisa que ajuda é que, se a gente for muito criterioso com a necessidade de "associação" no sentido de dependencia como nas variaveis categóricas, não teria nenhum teste razoavel para as variaveis intervalares, então vamos de:

Alternativa E

Questão 5:

Um pesquisador da área médica deseja estudar o efeito de um novo tratamento em pacientes com determinada doença em relação ao tratamento convencional. Cada paciente - de dez pacientes - foi submetido aos dois tratamentos. A variável de desfecho do estudo com distribuição normal é o número de dias sem fortes desconfortos no período de um ano de tratamento. O teste estatístico mais adequado para o estudo é:

```
A   t relacionado
B   Qui-quadrado
C   W
D   U
```

Resposta

Para duas amostras de VD com distribuição normal, pareadas (analise intra-grupos), o teste de escolha é o teste t pareado ("relacionado no enunciado")

Alternativa A

Questão 6

Um pesquisador da área médica deseja estudar o efeito de um novo tratamento em pacientes com determinada doença em relação a dois tratamentos convencionais. Em cada um dos três tratamentos foram alocados aleatoriamente seis pacientes distintos. A variável de desfecho com distribuição desconhecida e simétrica em cada tratamento é o número de dias sem fortes desconfortos no período de um ano de tratamento. O teste estatístico mais adequado para o estudo é:

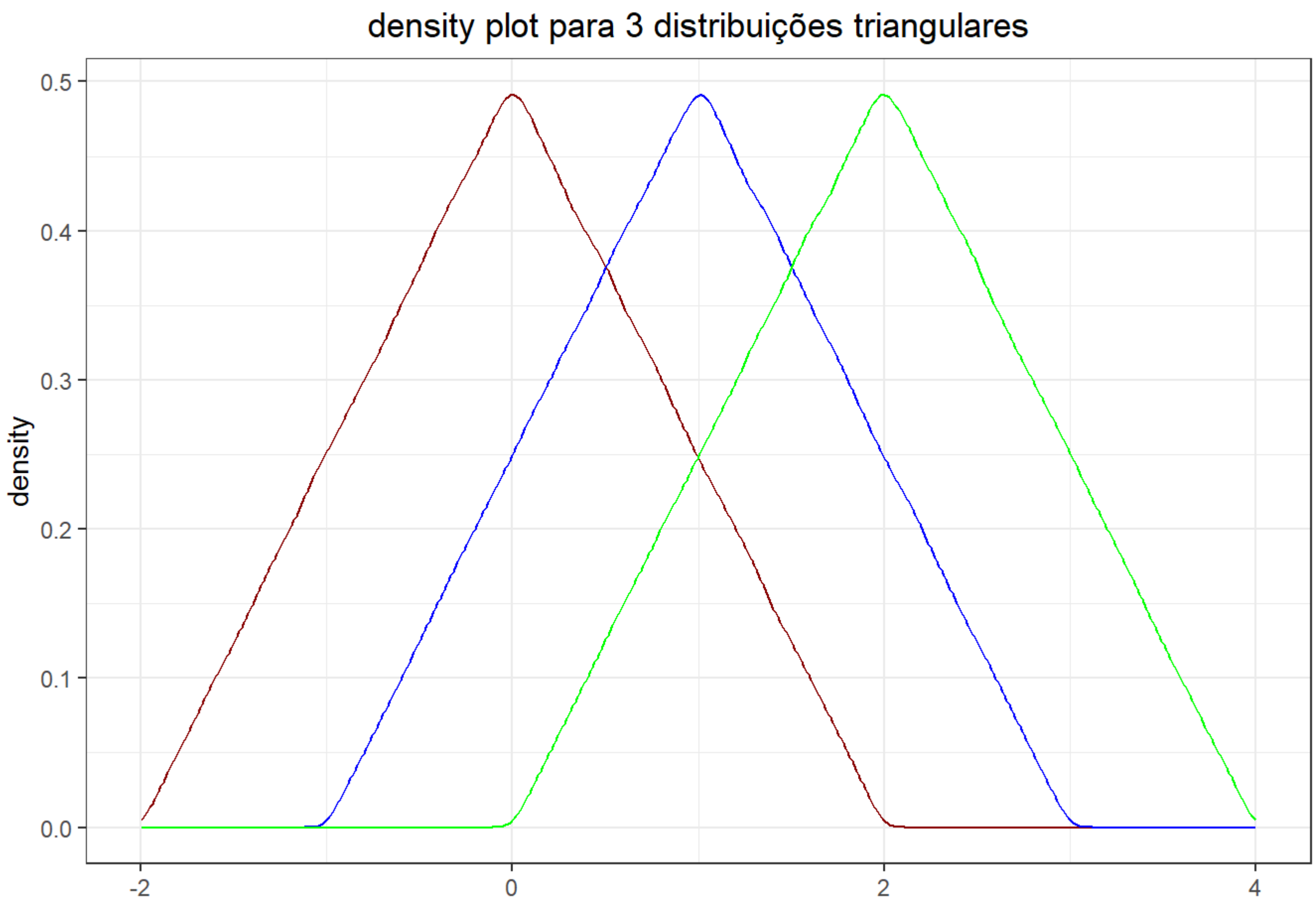
```
A   t independente
B   Qui-quadrado
C   H
D   ANOVA independente
```

Resposta

Aqui acho que é o caso que o senso comum leva a responder *H de Kruskal-Wallis*, porque é o teste não paramétrico de escolha para comparar 3 distribuições independentes, e a literatura diz ser especialmente valido usar esse teste quando os pressupostos de testes paramétricos (normalidade para anova) não são atingidos. Testes baseados em postos como o de Mann-Whitney e K-W ainda tem alguns pressupostos, como a simetria. Mas nesse caso esses pressupostos estão atingidos. Então parece um bom caso para KW. Fico só na duvida porque na aula claramente demonizaram todos os testes não-paramétricos, então não sei se não estão testando a gente aqui e querem que a gente coloque ANOVA mesmo assim. Fui de KW, pq esse caso tem todos os pressupostos teoricos para usar o teste, seja ele o melhor na teoria dos profs ou não. Tentei fazer uns testes com distribuição triangular, que é simétrica mas não é normal:

```
# criar função que simula distribuição triangular

generate_triangle <-
    \(center, n, one_tail_range) {
        triang_data <- \() {
            runif(n,
                  min = center/2 - one_tail_range,
                  max = center/2 + one_tail_range)
        }
        output <- triang_data() + triang_data()
        output
    }

# criar distribuições triangulares

set.seed(1)

triang_1 <-
    generate_triangle(center = 0,
                      n = 1e6,
                      one_tail_range = 1)
triang_2 <-
    generate_triangle(center = 1,
                      n = 1e6,
                      one_tail_range = 1)
triang_3 <-
    generate_triangle(center = 2,
                      n = 1e6,
                      one_tail_range = 1)

#plotar as distribuições triangulares
ggplot2::ggplot() +
  geom_density(data = data.frame(triang_1 = triang_1),
               aes(x = triang_1),
               color = 'dark red') +

  geom_density(data = data.frame(triang_2 = triang_2),
               aes(x = triang_2),
               color = 'blue') +

  geom_density(data = data.frame(triang_3 = triang_3),
               aes(x = triang_3),
               color = 'green') +
  theme_bw() +
  xlab("") +
```

```
  ggtitle('density plot para 3 distribuições triangulares') +
  theme(plot.title = element_text(hjust = 0.5))
```




```
# criar data.frames com amostras desses triangulos

# função que tira subamostra com 6 elementos de uma distribuição
six_elements <-
    \(x) {
        sample(x,
               size = 6,
               replace = FALSE)
    }

# função que cria data.frame a partir das 3 distribuições

set.seed(1)

create_data.frame <-
    \(list_of_distributions){
        lapply(list_of_distributions,
               \(x) six_elements(x)) |>
            as.data.frame() |>
            setNames(c("grupo_1", "grupo_2", "grupo_3")) |>
            tidyr::pivot_longer(cols = everything(),
                                names_to = "grupo",
                                values_to = "valor")
    }
```

```
# criar 1000 data.frames
list_of_data.frames <-
    lapply(1:1000,
           \(x) create_data.frame(list(triang_1,
                                       triang_2,
                                       triang_3)))

## criar as funções que rodam os testes
## anova
teste_aov <-
    \(x) {
        aov(data = x,
            formula = valor ~ grupo)
    }

teste_kw <-
    \(x) {
        kruskal.test(formula = valor ~ grupo,
                     data = x)
    }

# aplicar os testes nas 1000 data.frames

list_of_tests <-
    lapply(
        list(
            teste_aov,
            teste_kw
            ),
       \(x) lapply(list_of_data.frames,
                   \(y) x(y)
                   )
       )
# obter os valores p dos testes

list_of_summaries <-
    list(aov = lapply(list_of_tests[[1]],
                      \(x) summary(x)[[1]]["grupo", "Pr(>F)"]),
         kw = lapply(list_of_tests[[2]],
                     \(x) x[["p.value"]])
         )

data.frame_of_summaries <-
    list_of_summaries |>
    lapply(unlist) |>
    as.data.frame()
```

O bloco acima simulou 1000 data.frames criadas a partir de subamostras das distribuições triangulares. Abaixo vamos verificar como os dois testes se comportam:

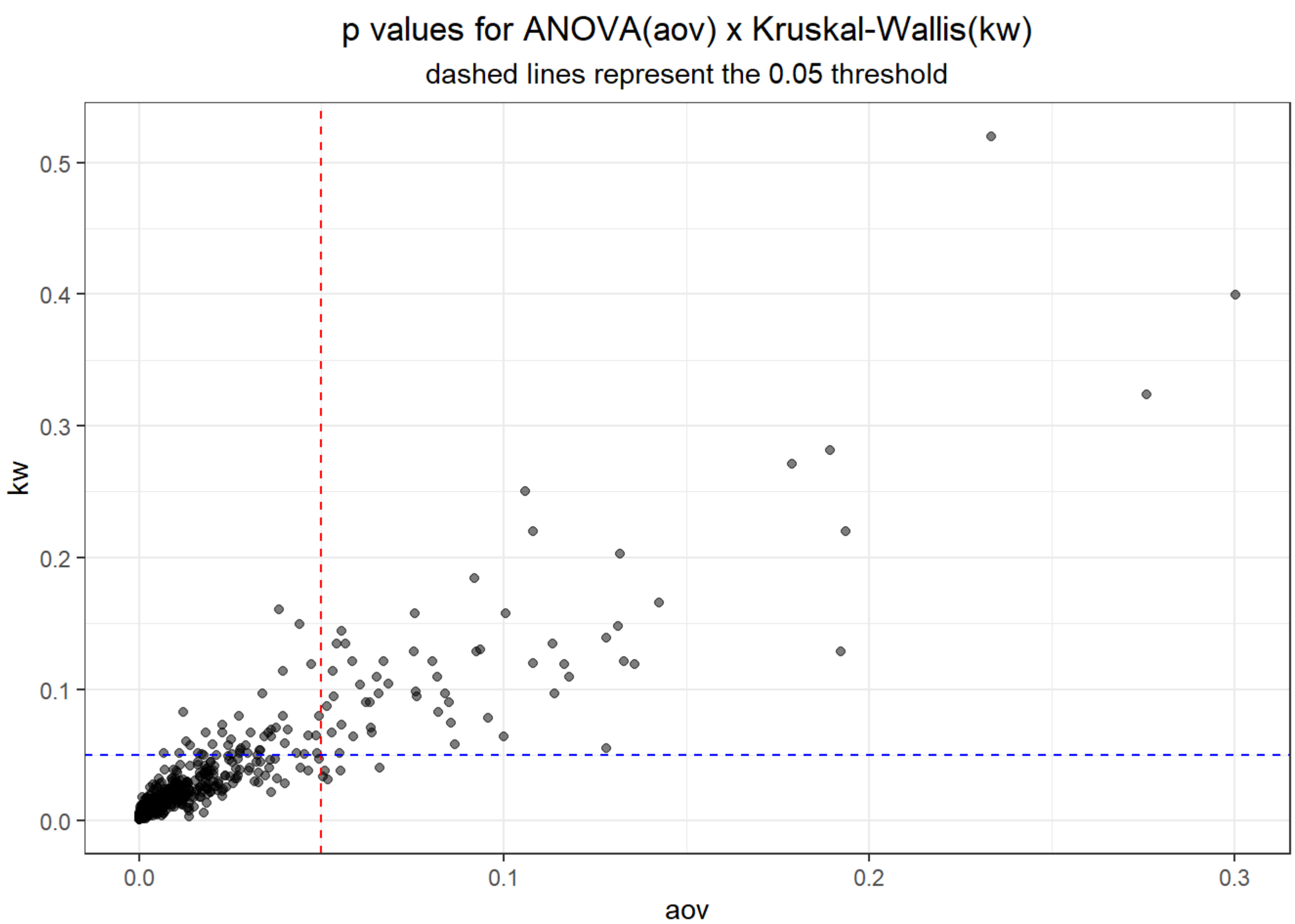
```
# explorar os dados
## proporção de valores de p < 0.05
data.frame_of_summaries |>
```

```
    dplyr::summarise(across(everything(),
                     \(x) mean(x < 0.05)))
```

```
    aov   kw
1 0.937 0.9
```

Os dois rejeitam H0 de forma consistente, o aov um pouco melhor.

```
# plotar valores de p em cada um dos eixos
library(ggplot2)

data.frame_of_summaries |>
    ggplot(aes(x = aov,
               y = kw)) +
    geom_point(alpha = 0.5) +
    geom_vline(xintercept = 0.05,
               color = "red",
               linetype = "dashed") +
    geom_hline(yintercept = 0.05,
               color = "blue",
               linetype = "dashed") +
    ggtitle("p values for ANOVA(aov) x Kruskal-Wallis(kw)", subtitle = "dashed lines represent
    theme_bw() +
    theme(plot.title = element_text(hjust = 0.5),
          plot.subtitle = element_text(hjust = 0.5))
```



Da pra ver, pelo grafico acima, que é mais comum apenas kw falhar em rejeitar a H0 (quadrante superior esquerdo) do que apenas o aov (inferior direito), embora em varios casos ambos tenham falhado (superior direito). Então mudei minha resposta para D - ANOVA.

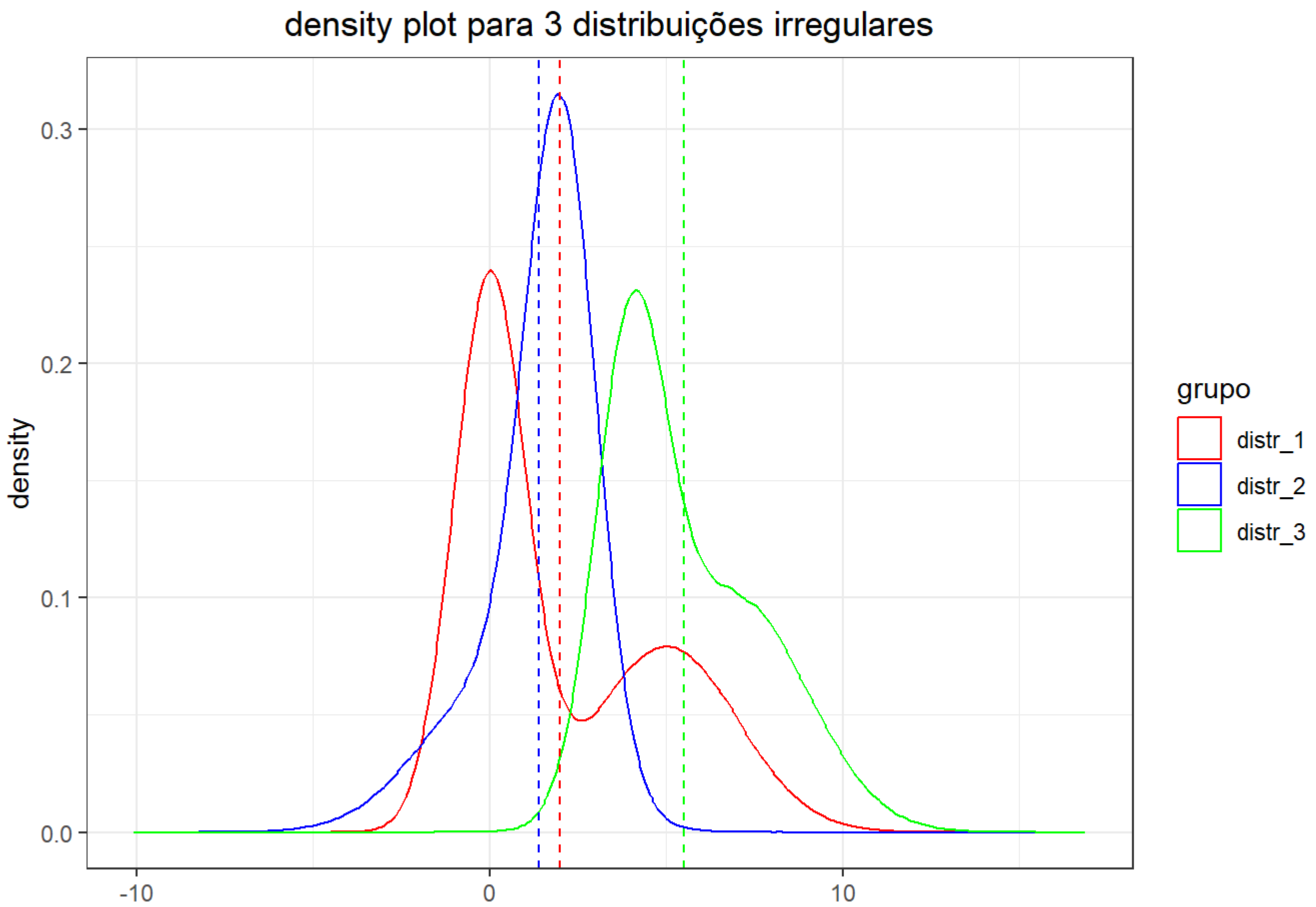
Alternativa D

Questão 6 - simulação bonus

Agora vamos tentar uma simulação semelhante, mas ao inves das distribuições triangulares, vamos usar distribuições irregulares:

```
# criando as distribuições irregulares e sua data.frame

set.seed(1)

two_normal_distrs_1 <-
  list(rnorm(n = 1e6, mean = 0, sd = 1),
       rnorm(n= 1e6, mean = 5, sd = 2))

two_normal_distrs_2 <-
  list(rnorm(n = 1e6, mean = 2, sd = 1),
       rnorm(n= 1e6, mean = 0, sd = 2))

two_normal_distrs_3 <-
  list(rnorm(n = 1e6, mean = 4, sd = 1),
       rnorm(n= 1e6, mean = 7, sd = 2))

distr_1 <- c(sample(two_normal_distrs_1[[1]], 1e5*6),
             sample(two_normal_distrs_1[[2]], 1e5*4))

distr_2 <- c(sample(two_normal_distrs_2[[1]], 1e5*7),
             sample(two_normal_distrs_2[[2]], 1e5*3))

distr_3 <- c(sample(two_normal_distrs_3[[1]], 1e5*5),
             sample(two_normal_distrs_3[[2]], 1e5*5))



data <-
  data.frame(distr_1, distr_2, distr_3) |>
  tidyr::pivot_longer(everything(),
                      names_to = "grupo",
                      values_to = 'valores')

ggplot2::ggplot() +
  geom_density(data = data,
               aes(x= valores,
                   color = grupo)) +
  scale_color_manual(values = c('distr_1' = 'red',
                                'distr_2' = 'blue',
                                'distr_3' = 'green'))+
  geom_vline(xintercept =
               mean(filter(data, grupo == 'distr_1') |> pull(valores)),
```

```
             linetype = 'dashed',
             color = 'red') +
  geom_vline(xintercept =
               mean(filter(data, grupo == 'distr_2') |> pull(valores)),
             linetype = 'dashed',
             color = 'blue') +
  geom_vline(xintercept =
               mean(filter(data, grupo == 'distr_3') |> pull(valores)),
             linetype = 'dashed',
             color = 'green') +
  theme_bw() +
  xlab("") +
  ggtitle('density plot para 3 distribuições irregulares') +
  theme(plot.title = element_text(hjust = 0.5))
```



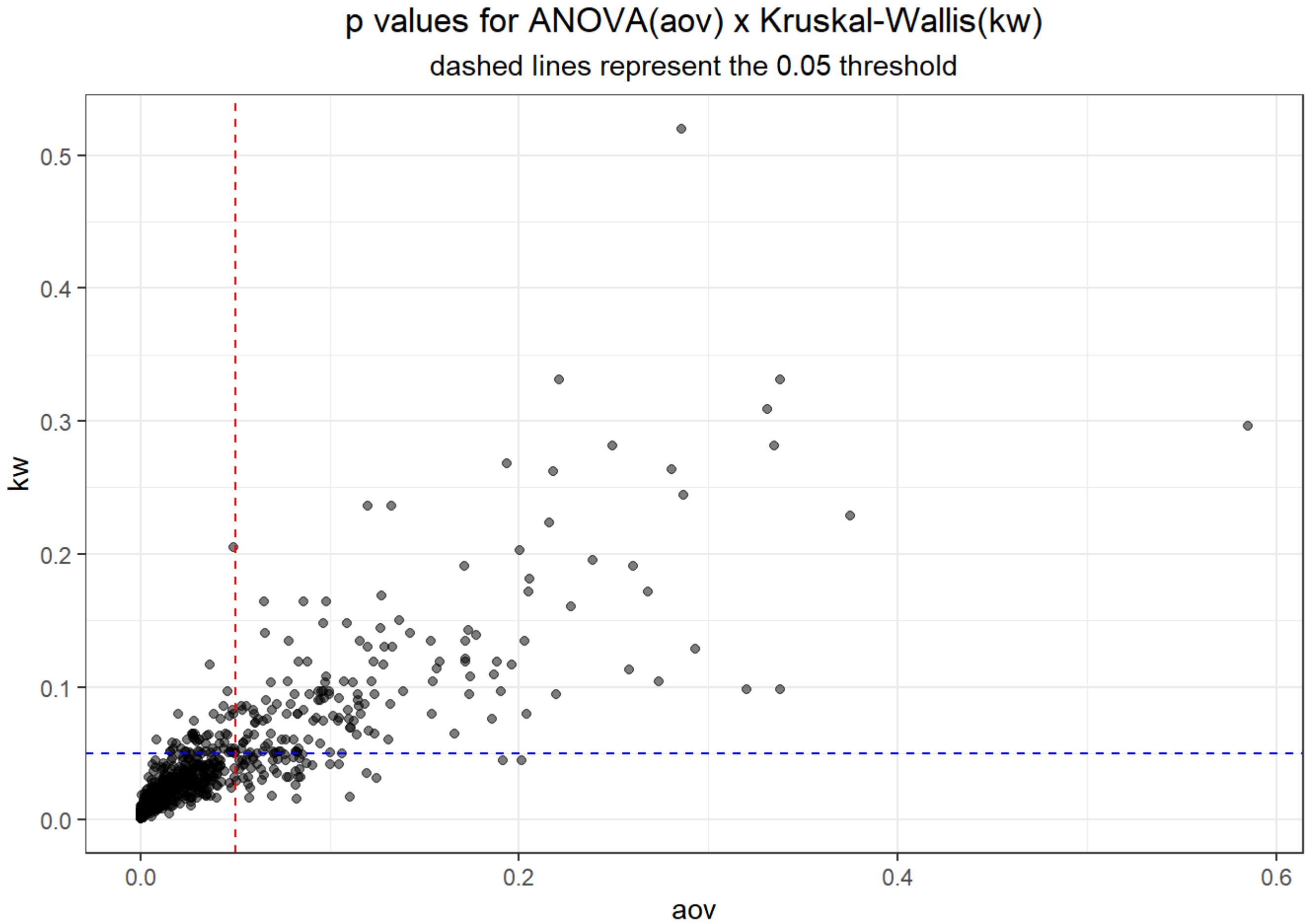

```
# criar 1000 data.frames
list_of_data.frames <-
  lapply(1:1000,
         \(x) create_data.frame(list(distr_1,
                                     distr_2,
                                     distr_3)))


# aplicar os testes nas 1000 data.frames

list_of_tests <-
  lapply(
```

```
    list(
      teste_aov,
      teste_kw
    ),
    \(x) lapply(list_of_data.frames,
                \(y) x(y)
    )
  )
# obter os valores p dos testes

list_of_summaries <-
  list(aov = lapply(list_of_tests[[1]],
                    \(x) summary(x)[[1]]["grupo", "Pr(>F)"]),
       kw = lapply(list_of_tests[[2]],
                   \(x) x[["p.value"]])
  )

data.frame_of_summaries <-
  list_of_summaries |>
  lapply(unlist) |>
  as.data.frame()

# explorar os dados
## proporção de valores de p < 0.05
data.frame_of_summaries |>
  dplyr::summarise(across(everything(),
                          \(x) mean(x < 0.05)))
```

```
    aov    kw
1 0.797 0.796
```

O desempenho aqui foi parecido

```
# plotar valores de p em cada um dos eixos
library(ggplot2)

data.frame_of_summaries |>
  ggplot(aes(x = aov,
             y = kw)) +
  geom_point(alpha = 0.5) +
  geom_vline(xintercept = 0.05,
             color = "red",
             linetype = "dashed") +
  geom_hline(yintercept = 0.05,
             color = "blue",
             linetype = "dashed") +
  ggtitle("p values for ANOVA(aov) x Kruskal-Wallis(kw)", subtitle = "dashed lines represent t
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5))
```



Questão 7

Um pesquisador da área médica deseja estudar o efeito de um novo tratamento em pacientes com determinada doença em relação a dois tratamentos convencionais. Cada um dos dez pacientes foi submetido aos três tratamentos. A variável de desfecho do estudo com distribuição normal é o número de dias sem fortes desconfortos no período de um ano de tratamento. O teste estatístico mais adequado para o estudo é:

```
A   t pareado
B   Qui-quadrado
C   Q de Friedman
D   ANOVA relacionada
```

Resposta

O teste para variavel continua de distribuição normal, pareada (intra-participantes), com mais de 2 grupos, é o de ANOVA relacionada

Alternativa D

Questão 8

Um pesquisador da área médica deseja estudar o efeito de um novo tratamento em pacientes com determinada doença em relação a dois tratamentos convencionais. Cada um dos dez pacientes foi submetido aos três tratamentos. A variável de desfecho do estudo é ordinal com os níveis é A (0% a 25% de dias sem fortes desconfortos no período de um ano de tratamento), B (>25% a 50%), C (>50% a 75%) e D (>75% a 100%). O teste estatístico mais adequado para o estudo é:

```
A   t relacionado
B   Qui-quadrado de Pearson
C   Q de Friedman
D   ANOVA relacionada
E   W de Wilcoxon
F   H de Kruskal-Wallis
```

Resposta

Aqui, mais uma vez, da muita vontade de responder *Q de Friedman* , porque é o teste de escolha para medidas repetidas (intra-participantes), com n pequeno e variavel ordinal com pouquissimos niveis. Nunca sei se não é mais uma pegadinha pra testar a nossa fé nos testes paramétricos. Se tivermos fé, colocamos ANOVA relacionada. Eu acho que não tenho. E dessa vez fiquei com preguiça de fazer as simulações.

Alternativa C

Questão 9

A VD de teste U de Mann-Whitney NÃO pode ser:

```
A   Nominal
B   Ordinal
C   Intervalar
D   Razão
E   Contagem
```

Resposta

Tudo que pode ser ordenado pode entrar num teste de MU. Então apenas o que não é ordenavel (variavel nominal) é impossivel de ser usado num teste de mann whitney.

Alternativa A

Questão 10

A VD de teste W de Wilcoxon tem que ser:

```
A   Nominal
B   Ordinal
C   Intervalar
```

Resposta

Da aula:

> O teste W de Wilcoxon testa a hipótese nula de igualdade das médias populacionais da VD quantitativa em duas condições dependentes. É, portanto, correspondente ao teste t relacionado.

> Conover (1999) sumariza as suposições: a distribuição das diferenças deve ser simétrica. as diferenças devem ser independentes entre si. todas as diferenças precisam ter a mesma média. a medida das diferenças deve ser pelo menos intervalar.

> A VD não pode ser, consequentemente, ordinal (vide a tabela de Conover, 1999).

Achei muito confuso, porque o senso comum leva a crer que testes de postos sao adequados para variaveis ordinais mas GPT deu uma explicação consistente. Resumindo muito (e talvez mal), o teste W para amostras pareadas precisa primeiro calcular as diferenças entre pares, e depois cria um rank das diferenças. Ou seja, embora seja um teste de postos, os postos surgem em uma etapa intermediaria, e é necessario ter as distancias intervalares para criar os postos, por isso a necessidade de ter variaveis intervalares.

Do GPT-4:

> O requisito de que a variável de resultado seja “pelo menos de intervalo, não ordinal” no contexto do teste W de Wilcoxon para dados emparelhados (também conhecido como teste de sinais e postos de Wilcoxon) pode parecer um pouco confuso, dada sua natureza como um teste não paramétrico baseado em ranqueamento. Aqui está como isso se encaixa na metodologia do teste e suas premissas:

> Natureza dos Testes Baseados em Ranks: Testes baseados em ranks, incluindo o teste de sinais e postos de Wilcoxon, geralmente transformam os dados em ranks, mitigando assim os efeitos de outliers e não-normalidade. Isso os torna adequados para dados não paramétricos que não assumem uma distribuição específica, como a distribuição normal, frequentemente assumida em testes paramétricos.

> Manuseio das Diferenças Emparelhadas: O teste de sinais e postos de Wilcoxon lida especificamente com diferenças emparelhadas. Ele ranqueia essas diferenças, independentemente de seu sinal, e então usa esses ranks para avaliar se a mediana das diferenças é significativamente diferente de zero. O ponto crítico aqui é que o teste calcula diferenças entre pares e depois ranqueia essas diferenças.

> Requisito para Dados de Intervalo: Embora seja verdade que testes baseados em ranks tipicamente não precisem de dados de intervalo e possam operar com dados ordinais, a operação de calcular diferenças (usada no teste de sinais e postos de Wilcoxon) pressupõe que essas diferenças sejam significativas. Isso não é garantido em dados ordinais. Por exemplo, se você tem

> categorias ordinais como “ruim”, “regular”, “bom” e “excelente”, a diferença entre “ruim” e “regular” pode não ser equivalente à diferença entre “bom” e “excelente”. Portanto, para que o teste calcule e interprete essas diferenças com precisão, os dados precisam estar pelo menos em uma escala de intervalo, onde a diferença entre quaisquer dois pontos é consistentemente significativa.

> Aplicabilidade a Dados de Intervalo: Ao exigir que os dados estejam pelo menos em uma escala de intervalo, o teste garante que operações como a subtração, usadas no cálculo das diferenças, sejam válidas e consistentes. Isso permite um ranqueamento significativo dessas diferenças e um teste válido para inferência estatística sobre a tendência central das diferenças.

Alternativa C