Atividade Avaliativa

Análise Estatística Para Várias Populações

Ana Maria Alves da Silva

2024-10-20

```
df <- read_excel(
"/Users/anamaria/especializacao/modulo_6/atividade/Nascidos_vivos_no_municipio_de_Sao_Paulo.xls")

# Tamanho do conunto de dados
print(dim(df))
```

Carregamento e tratativa inicial dos dados

```
## [1] 65535    37
```

- Verificando se é um dataframe:

```
is_data_frame <- is.data.frame(df)
print(is_data_frame)
```

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

- Verificando um resumo dos dados:

```
summary_df <- summary(df)
print(summary_df)
```

```
##    LOCNASC            IDADEMAE          ESTCIVMAE            ESCMAE
##  Length:65535       Length:65535       Length:65535       Length:65535
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##   QTDFILVIVO         QTDFILMORT          GESTACAO           GRAVIDEZ
##  Length:65535       Length:65535       Length:65535       Length:65535
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
```

```
##
##     PARTO             CONSULTAS           MESNASC             ANONASC
##  Length:65535       Length:65535       Length:65535       Length:65535
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##      SEXO              APGAR1             APGAR5              PESO
##  Length:65535       Length:65535       Length:65535       Length:65535
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##    IDANOMAL          RETROALIM           NATURALMAE       CODUFNATU
##  Length:65535       Length:65535       Min.   :  0.000   Length:65535
##  Class :character   Class :character   1st Qu.:  0.000   Class :character
##  Mode  :character   Mode  :character   Median :  0.000   Mode  :character
##                                        Mean   :  2.836
##                                        3rd Qu.:  0.000
##                                        Max.   :800.000
##   ESCMAE2010         SERIESCMAE         RACACORMAE         QTDGESTANT
##  Length:65535       Length:65535       Length:65535       Length:65535
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##   QTDPARTNOR         QTDPARTCES          IDADEPAI          SEMAGESTAC
##  Length:65535       Length:65535       Length:65535       Length:65535
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##   TPMETESTIM         CONSPRENAT         MESPRENAT          TPAPRESENT
##  Length:65535       Length:65535       Length:65535       Length:65535
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##   STTRABPART         STCESPARTO         TPNASCASSI         ESCMAEAGR1
##  Length:65535       Length:65535       Length:65535       Length:65535
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##    TPROBSON
##  Length:65535
##  Class :character
##  Mode  :character
```

##
##
##

Observe que todas as variáveis estão do tipo texto. Conforme necessário realizaremos a alteração para o tipo numérico.

Resoluções:

Questão 1: A pontuação Apgar é uma avaliação rápida da condição do recém-nascido, realizada nos primeiro e quinto minuto de vida. As variáveis analisadas serão APGAR1 e APGAR5, que representam, respectivamente, as pontuações de Apgar nesses momentos. Verifique se houve uma mudança significativa na condição dos recém-nascidos entre o 1º e o 5º minuto após o nascimento. Em outras palavras, teste as hipóteses:

- H_0 : Não há diferença significativa entre as pontuações de Apgar no 1º e no 5º minuto, ou seja:

  $H_0 : \mu_1 = \mu_2$,

  onde

  - μ_1 é a média das pontuações de Apgar no 1º minuto após o nascimento;
  - μ_2 é a média das pontuações de Apgar no 5º minuto após o nascimento.

- H_1 : Há uma diferença significativa entre as pontuações de Apgar no 1º e no 5º minuto, ou seja,

  $H_1 : \mu_1 \neq \mu_2$.

Como estamos comparando as condições de um mesmo recém-nascido em dois momentos, os dados são considerados pareados.

Se for observada uma diferença significativa entre as pontuações APGAR1 e APGAR5, realize um teste t unilateral adequado para verificar se a pontuação de Apgar no 1º minuto é significativamente menor (ou maior) do que no 5º minuto. Interprete todos os resultados com um nível de significância de $\alpha = 1\%$.

Solução questão 1: Primeiramente vamos filtrar apenas os dados que precisamos para realizar o que se pede nessa questão, isso é os dados das colunas APGAR1 e APGAR5 não podem ser nulos e devem ser numéricos.

```
df_questao1 <- df %>%
  mutate(APGAR1 = as.numeric(APGAR1),
         APGAR5 = as.numeric(APGAR5)) %>%
  filter(!is.na(APGAR1) & !is.na(APGAR5))

print(df_questao1)
```

```
## # A tibble: 65,386 x 37
##    LOCNASC IDADEMAE ESTCIVMAE ESCMAE QTDFILVIVO QTDFILMORT GESTACAO GRAVIDEZ
##    <chr>   <chr>    <chr>     <chr>  <chr>      <chr>      <chr>    <chr>
##  1 3       25       2         5      01         00         5        1
##  2 1       31       5         5      00         00         4        1
```

```
##  3 1       34       2         5      01         00         5         1
##  4 1       32       1         4      03         01         5         1
##  5 1       33       4         4      00         00         5         2
##  6 1       17       5         4      00         00         5         1
##  7 1       30       2         5      02         01         5         1
##  8 1       16       1         4      00         00         5         1
##  9 1       23       2         4      00         00         5         1
## 10 1       19       1         4      00         00         5         1
## # i 65,376 more rows
## # i 29 more variables: PARTO <chr>, CONSULTAS <chr>, MESNASC <chr>,
## #   ANONASC <chr>, SEXO <chr>, APGAR1 <dbl>, APGAR5 <dbl>, PESO <chr>,
## #   IDANOMAL <chr>, RETROALIM <chr>, NATURALMAE <dbl>, CODUFNATU <chr>,
## #   ESCMAE2010 <chr>, SERIESCMAE <chr>, RACACORMAE <chr>, QTDGESTANT <chr>,
## #   QTDPARTNOR <chr>, QTDPARTCES <chr>, IDADEPAI <chr>, SEMAGESTAC <chr>,
## #   TPMETESTIM <chr>, CONSPRENAT <chr>, MESPRENAT <chr>, TPAPRESENT <chr>, ...
```

Vamos análisar um resumo dos dados considerando apenas as colunas que temos interesse para que possamos entender um pouco sobre nossos dados.

```
resumo_1 = summary(df_questao1[, c("APGAR1", "APGAR5")])
print(resumo_1)
```

```
##      APGAR1           APGAR5
##  Min.   : 0.000   Min.   : 0.00
##  1st Qu.: 8.000   1st Qu.: 9.00
##  Median : 9.000   Median :10.00
##  Mean   : 8.425   Mean   : 9.47
##  3rd Qu.: 9.000   3rd Qu.:10.00
##  Max.   :99.000   Max.   :99.00
```

Como os dados são pareados, isso é, cada observação de APGAR1 possui uma observação correspondente em APGAR5, é apropriado realizarmos um t-teste pareado pois o t-teste pareado é uma ferramenta utilizada para determinar se as médias de duas amostras dependentes são estatisticamente diferentes. Podemos observar também, que no resumo dos dados que apresentamos acima obtemos que a média do APGAR1 é 8.425 enquanto a média do APGAR5 é 9.47.

Usaremos a opção *paired = TRUE* especifica que é um teste pareado enquanto a *alternative = "two.sided"* permite identificarmos se existe qualquer diferença estatisticamente significativa entre as pontuações APGAR1 e APGAR5, independentemente da direção da diferença. Já o parâmetro *conf.level = 0.99* ajusta o nível de confiança para 99%, que corresponde a um nível de significância de 1%, que é solicitado na questão.

Observamos que caso a se a hipótese alternativa fosse $H_1 : \mu_1 < \mu_2$, implicando que estamos apenas interessados em testar se a pontuação no 1º minuto é significativamente menor do que no 5º minuto usariamos *alternative = "less"*

```
teste_t <- t.test(df_questao1$APGAR1,
                  df_questao1$APGAR5,
                  paired = TRUE,
                  alternative = "two.sided",
                  conf.level = 0.99)

print(teste_t)
```

```
##
##  Paired t-test
##
## data:  df_questao1$APGAR1 and df_questao1$APGAR5
## t = -289.08, df = 65385, p-value < 2.2e-16
## alternative hypothesis: true mean difference is not equal to 0
## 99 percent confidence interval:
##  -1.053519 -1.034910
## sample estimates:
## mean difference
##       -1.044214
```

```r
if (teste_t$p.value < 0.01) {
  cat("Existe uma diferença significativa entre as pontuações APGAR1 e
      APGAR5.\n")
  if (mean(df_questao1$APGAR1) > mean(df_questao1$APGAR5)) {
    cat("A pontuação no 1º minuto é significativamente maior que no 5º
        minuto.\n")
  } else {
    cat("A pontuação no 5º minuto é significativamente maior que no 1º
        minuto.\n")
  }
} else {
  cat("Não existe uma diferença significativa entre as pontuações APGAR1 e
      APGAR5.\n")
}
```

```
## Existe uma diferença significativa entre as pontuações APGAR1 e
##       APGAR5.
## A pontuação no 5º minuto é significativamente maior que no 1º
##         minuto.
```

Através do resultado do t-teste pareado identificamos que há uma diferença significativa entre as pontuações APGAR no 1º e no 5º minuto e além disso determinamoss que a pontuação no 5º minuto é significativamente maior, realizaremos um t-teste unilateral para confirmar se a pontuação do 5º minuto é significativamente maior do que a do 1º minuto, com um nível de significância de 1%, para isso usaremos a opção *alternative = "greater"*

```r
teste_t_unilateral <- t.test(df_questao1$APGAR5,
                             df_questao1$APGAR1,
                             paired = TRUE,
                             alternative = "greater",
                             conf.level = 0.99)

print(teste_t_unilateral)
```

```
##
##  Paired t-test
##
## data:  df_questao1$APGAR5 and df_questao1$APGAR1
## t = 289.08, df = 65385, p-value < 2.2e-16
## alternative hypothesis: true mean difference is greater than 0
## 99 percent confidence interval:
```

```
##  1.035811      Inf
## sample estimates:
## mean difference
##        1.044214
```

```r
if (teste_t_unilateral$p.value < 0.01) {
  cat("A pontuação no 5º minuto é significativamente maior que no 1º minuto
      com um nível de significância de 1%.\n")
  cat("P-valor:", teste_t_unilateral$p.value, "\n")
} else {
  cat("Não há evidência suficiente para afirmar que a pontuação no 5º minuto
      é significativamente maior que no 1º minuto com um nível de significância
      de 1%.\n")
}
```

```
## A pontuação no 5º minuto é significativamente maior que no 1º minuto
##       com um nível de significância de 1%.
## P-valor: 0
```

O t-teste unilateral pareado nos informaque a pontuação de Apgar no 5º minuto é significativamente maior do que no 1º minuto, com um nível de confiança de 99%. A diferença média de 1.044214, com um intervalo de confiança que não inclui o zero e se estende ao infinito, reforça que esta diferença é positiva e significativa.

Com base nos testes que realizamos podemos concluir que a diferença entre as pontuações Apgar no primeiro e quinto minuto é clinicamente relevante sugerindo que há uma melhoria no estado de saúde dos recém-nascidos nos primeiros cinco minutos após o nascimento.

Questão 2: Verifique se há uma diferença significativa no peso médio ao nascer entre recém-nascidos do sexo masculino e feminino. Os dados disponíveis incluem informações sobre o peso ao nascer (em gramas) e o sexo dos recém-nascidos. As variáveis consideradas para a análise são:

- PESO: peso ao nascer (em gramas).
- SEXO: sexo dos recém-nascidos (masculino ou feminino)

O objetivo é testar a hipótese de que não há diferença no peso médio ao nascer entre os dois grupos.

Hipóteses de interesse:

- H_0: Não há diferença significativa entre os pesos médios ao nascer de recém-nascidos do sexo masculino e feminino, ou seja, as médias são iguais.
- H_1: Existe uma diferença significativa entre os pesos médios ao nascer de recém-nascidos do sexo masculino e feminino.

Ou seja,

$$H_0 : \mu_M = \mu_F \quad \text{contra} \quad H_1 : \mu_M \neq \mu_F,$$

onde μ_M é o peso médio dos recém-nascidos do sexo masculino e μ_F é o peso médio dos recém-nascidos do sexo feminino.

Caso seja observada uma diferença significativa entre os pesos médios, realize um teste t unilateral adequado para verificar se o peso médio entre meninos é significativamente maior ou menor do que entre as meninas. Todos os resultados deverão ser interpretados ao nível de significância de $\alpha = 1\%$.

Solução questão 2:

Primeiramente vamos filtrar apenas os dados que precisamos para realizar o que se pede nessa questão, isso é os dados das colunas Sexo e Peso. Além disso, precisaremos converter os dados da coluna Peso para númericos e verificar se há valores ausentes separadamente em ambas as colunas.

```
df_questao2 <- df[!is.na(df$PESO) & !is.na(df$SEXO) & df$SEXO %in% c("M", "F"),]
df_questao2$PESO <- as.numeric(df_questao2$PESO)
print(df_questao2)
```

```
## # A tibble: 65,523 x 37
##    LOCNASC IDADEMAE ESTCIVMAE ESCMAE QTDFILVIVO QTDFILMORT GESTACAO GRAVIDEZ
##    <chr>   <chr>    <chr>     <chr>  <chr>      <chr>      <chr>    <chr>
##  1 1       25       1         4      02         00         5        1
##  2 3       25       2         5      01         00         5        1
##  3 1       31       5         5      00         00         4        1
##  4 1       34       2         5      01         00         5        1
##  5 1       32       1         4      03         01         5        1
##  6 1       33       4         4      00         00         5        2
##  7 1       17       5         4      00         00         5        1
##  8 1       30       2         5      02         01         5        1
##  9 1       16       1         4      00         00         5        1
## 10 1       23       2         4      00         00         5        1
## # i 65,513 more rows
## # i 29 more variables: PARTO <chr>, CONSULTAS <chr>, MESNASC <chr>,
## #   ANONASC <chr>, SEXO <chr>, APGAR1 <chr>, APGAR5 <chr>, PESO <dbl>,
## #   IDANOMAL <chr>, RETROALIM <chr>, NATURALMAE <dbl>, CODUFNATU <chr>,
## #   ESCMAE2010 <chr>, SERIESCMAE <chr>, RACACORMAE <chr>, QTDGESTANT <chr>,
## #   QTDPARTNOR <chr>, QTDPARTCES <chr>, IDADEPAI <chr>, SEMAGESTAC <chr>,
## #   TPMETESTIM <chr>, CONSPRENAT <chr>, MESPRENAT <chr>, TPAPRESENT <chr>, ...
```

```
summary(df_questao2$PESO)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     235    2860    3180    3128    3480    5640
```

Neste caso, realizaremos um t teste para Amostras Independentes. Este teste é ideal pois os grupos que queremos comparar são constituídos de indivíduos diferentes,recém-nascidos do sexo masculino versus feminino. A hipótese nula é de que não há diferença no peso médio ao nascer entre os dois grupos.

Para isso, filtramos os recém-nascidos do sexo masculino e as recém-nascidas do sexo feminino em nosso conjunto de dados. Assim como na questão 1, usaremos o parâmtro *alternative = "two.sided"*. Usaremos também o parâmetro *var.equal = FALSE* para não assumirmos que as variâncias são iguais entre os grupos, isso é, entre os sexos.

```
peso_masculino <- df_questao2$PESO[df_questao2$SEXO == "M"]
peso_feminino <- df_questao2$PESO[df_questao2$SEXO == "F"]

teste_t_independente <- t.test(peso_masculino,
                               peso_feminino,
                               alternative = "two.sided",
                               var.equal = FALSE,
                               conf.level = 0.99)
print(teste_t_independente)
```

```
##
##  Welch Two Sample t-test
##
## data:  peso_masculino and peso_feminino
## t = 24.165, df = 65518, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##   94.19796 116.67651
## sample estimates:
## mean of x mean of y
##  3179.607  3074.170
```

```r
if(teste_t_independente$p.value < 0.01) {
  cat("A hipótese nula foi rejeitada: existe uma diferença estatisticamente
      significativa no peso ao nascer entre os sexos.\n")
} else {
  cat("A hipótese nula não foi rejeitada: não existe uma diferença
      estatisticamente significativa no peso ao nascer entre os sexos.\n")
}
```

```
## A hipótese nula foi rejeitada: existe uma diferença estatisticamente
##       significativa no peso ao nascer entre os sexos.
```

Conforme os resultados do teste que realizamos, nós rejeitamos a hipótese nula, isso é, existe uma diferença estatisticamente significativa no peso ao nascer entre os sexos. Além disso, a diferença média de aproximadamente 105.437 gramas (calculada usando a diferença entre a média de peso entre os sexos, 3179.607 - 3074.170) entre os sexos é clinicamente significativa, sugerindo que recém-nascidos do sexo masculino tendem a ser mais pesados ao nascer do que os do sexo feminino.

Como a hipótese nula foi rejeitada e confirmamos que existe uma diferença significativa entre os pesos médios dos sexos, podemos prosseguir realizando um t teste unilateral com a hipótese alternativa de que o peso médio dos recém-nascidos do sexo masculino é maior que o peso médio das recém-nascidas do sexo feminino. Para realizar esse t teste, usaremos o parâmetro *alternative = "greater"*.

```r
teste_t_unilateral <- t.test(peso_masculino, peso_feminino,
                             alternative = "greater",
                             var.equal = FALSE, conf.level = 0.99)


print(teste_t_unilateral)
```

```
##
##  Welch Two Sample t-test
##
## data:  peso_masculino and peso_feminino
## t = 24.165, df = 65518, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 99 percent confidence interval:
##  95.28658      Inf
## sample estimates:
## mean of x mean of y
##  3179.607  3074.170
```

```r
if(teste_t_unilateral$p.value < 0.01) {
  cat("A hipótese alternativa foi aceita: o peso médio ao nascer dos
      recém-nascidos do sexo masculino é significativamente maior do que o das
      recém-nascidas do sexo feminino com um nível de significância de 1%.\n")
} else {
  cat("A hipótese alternativa não foi aceita: não há evidência suficiente para
      afirmar que o peso médio ao nascer dos recém-nascidos do sexo masculino é
      maior do que o das recém-nascidas do sexo feminino com um nível de
      significância de 1%.\n")
}
```

```
## A hipótese alternativa foi aceita: o peso médio ao nascer dos
##       recém-nascidos do sexo masculino é significativamente maior do que o das
##       recém-nascidas do sexo feminino com um nível de significância de 1%.
```

Como a hipótese alternativa foi aceita, temos que os resultados são conclusivos ao indicar que os recém-nascidos do sexo masculino nascem significativamente mais pesados que as recém-nascidas do sexo feminino, com uma diferença que é estatisticamente significativa.

Questão 3: Usando os dados do SINASC, deseja-se verificar se existe uma associação entre a duração da gestação (GESTACAO) e o tipo de gravidez (GRAVIDEZ). A variável GESTACAO indica o período gestacional em semanas, enquanto GRAVIDEZ refere-se ao tipo de gravidez, como única, dupla, ou tripla. Para isso, será realizado um teste de independência entre essas duas variáveis categóricas. Para simplificar a análise, algumas categorias das variáveis deverão ser agrupadas: na variável GRAVIDEZ, as categorias "dupla" e "tripla ou mais" serão unidas e formarão a nova categoria "dupla ou mais"; na variável GESTACAO, as categorias "menos de 22 semanas" e "de 22 a 27 semanas" serão agrupadas e formarão uma nova categoria chamada "menos de 27 semanas".

O objetivo é testar se a duração da gestação está associada ao tipo de gravidez.

Hipóteses de interesse:

- H_0: As variáveis GESTACAO (semanas de gestação) e GRAVIDEZ (tipo de gravidez) são independentes, ou seja, não há associação entre o período gestacional e o tipo de gravidez.
- H_1: As variáveis GESTACAO e GRAVIDEZ não são independentes, ou seja, há uma associação entre o período gestacional e o tipo de gravidez.

Interprete detalhadamente todos os resultados obtidos. Considere um nível de significância de $\alpha = 1\%$.

Solução questão 3:

Primeiramente vamos filtrar apenas os dados que precisamos para realizar o que se pede nessa questão, isso é os dados das colunas GESTACAO e GRAVIDEZ. Realizaremos a limpeza e tratativas dos dados conforme o necessário, realizando os agrupamentos solicitados.

```r
df$GESTACAO <- factor(df$GESTACAO, levels = c("1", "2", "3", "4", "5", "6"))
df$GRAVIDEZ <- factor(df$GRAVIDEZ, levels = c("1", "2", "3"))
```

```r
df$GESTACAO <- ifelse(df$GESTACAO %in% c("1", "2"),
                      "menos de 27 semanas", df$GESTACAO)
df$GESTACAO <- ifelse(df$GESTACAO %in% c("3", "4"), "27 a 36 semanas",
                      df$GESTACAO)
df$GESTACAO <- ifelse(df$GESTACAO %in% c("5", "6"), "37 semanas ou mais",
                      df$GESTACAO)
df$GRAVIDEZ <- ifelse(df$GRAVIDEZ %in% c("2", "3"), "múltipla", "única")
```

Para realizarmos o que se pede nessa questão, construiremos uma tabela de contigência com os dados categóricos que foram solicitados e depois realizaremos um teste Chi-quadrado.

```r
tabela_contigencia <- table(df$GRAVIDEZ, df$GESTACAO)
print(tabela_contigencia)
```

```
##
##            27 a 36 semanas 37 semanas ou mais menos de 27 semanas
##   múltipla            1159                595                  80
##   única               5975              57357                 341
```

```r
print(addmargins(tabela_contigencia))
```

```
##
##            27 a 36 semanas 37 semanas ou mais menos de 27 semanas   Sum
##   múltipla            1159                595                  80  1834
##   única               5975              57357                 341 63673
##   Sum                 7134              57952                 421 65507
```

```r
chi_test <- chisq.test(tabela_contigencia)

print(chi_test)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tabela_contigencia
## X-squared = 5815.5, df = 2, p-value < 2.2e-16
```

```r
if (chi_test$p.value < 0.01) {
  cat("Rejeitamos a hipótese nula. Existe uma associação significativa entre a
  duração da gestação e o tipo de gravidez com um nível de significância
  de 1%.\n")
} else {
  cat("Aceitamos a hipótese nula, não existe uma associação significativa entre
  a duração da gestação e o tipo de gravidez com um nível de significância
  de 1%.\n")
}
```

```
## Rejeitamos a hipótese nula. Existe uma associação significativa entre a
##   duração da gestação e o tipo de gravidez com um nível de significância
##   de 1%.
```

A elevada estatística Chi-quadrado reforça que a duração da gestação está significativamente associada ao tipo de gravidez. A duração da gestação varia de maneira previsível com base no tipo de gravidez (única ou múltipla). Rejeitamos a hipótese nula com um alto grau de certeza devido ao valor p ser muito pequeno.

Além disso, esta descoberta é estatisticamente robusta e tem implicações significativas para a prática médica e o planejamento de políticas de saúde. Apontando para a necessidade de abordagens diferenciadas no acompanhamento e intervenção durante a gravidez, dependendo de ser uma gravidez única ou múltipla.

Questão 4: Verifique se há uma diferença estatisticamente significativa entre os pesos médios ao nascer (em gramas) dos recém-nascidos, de acordo com a raça/cor das mães. As variáveis utilizadas são: PESO, representando o peso ao nascer, e RACACORMAE, cujos níveis são: 1 – Branca, 2 – Preta, 3 – Amarela, 4 – Parda e 5 – Indígena.

Para essa análise, aplique a Análise de Variância (ANOVA) de um fator (adequada), verificando se o peso médio dos recém-nascidos difere significativamente entre os grupos de raça/cor das mães.

Caso a ANOVA aponte uma diferença significativa entre os grupos, realize um teste de comparações múltiplas adequado. Utilize o teste de Tukey se as variâncias forem homogêneas, ou o teste de Games-Howell caso as variâncias não sejam homogêneas, para identificar quais grupos raciais apresentam diferenças significativas entre os pesos médios dos recém-nascidos.

Interprete detalhadamente todos os resultados obtidos. Considere um nível de significância de $\alpha = 1\%$.

Solução questão 4:

Primeiramente vamos filtrar apenas os dados que precisamos para realizar o que se pede nessa questão, isso é os dados das colunas PESO e RACACORMAE. Realizaremos a limpeza e tratativas dos dados conforme o necessário, além disso, observe que no conjunto de dados o peso está em gramas.

```
df$PESO <- as.numeric(as.character(df$PESO))
df$RACACORMAE <- factor(df$RACACORMAE, levels = c("1", "2", "3", "4", "5", "9"),
                            labels = c("Branca", "Preta", "Amarela", "Parda",
                                       "Indígena", "Ignorado"))


questao_4 <- df %>%
  filter(!is.na(PESO), !is.na(RACACORMAE), RACACORMAE != "Ignorado")

print(dim(questao_4))
```

```
## [1] 65511    37
```

Agora, deverimaos verificar se os dados seguem uma distribuição normal mas note que nosso conjunto de dados possui mais de 65 mil observações. Caso nosso conjunto de dados fosse um pouco menor, ou seja, até 5 mil observações, poderiamos usar o teste de Shapiro-Wilk para verificar na normalidade dos dados, o que não é o nosso caso. Aqui, nós não realizaremos teste de normalidade, pois aplicaremos o Teorema Central do Limite, isso é, quando o tamanho da amostra é suficientemente grande, a distribuição amostral da média tende a ser aproximadamente normal, independentemente da distribuição original dos dados. Assim, a ANOVA pode ser aplicada, pois a normalidade das amostras individuais não é tão crítica.

Mas, antes de realizar a ANOVA, vale a pena realizar o teste de Levene para verificar a homogeneidade das variâncias dos grupos. Neste caso, temos que a hipótese nula é:

- H_0: Não há diferença no peso médio ao nascer entre os diferentes grupos de raça/cor das mães.

Enquanto a hipótese alternativa é:

- H_1: Existe pelo menos uma diferença no peso médio ao nascer entre os diferentes grupos de raça/cor das mães.

```
levene_resultado <- leveneTest(PESO ~ RACACORMAE, data = questao_4)
print(levene_resultado)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value    Pr(>F)
## group     4  5.9864 8.206e-05 ***
##       65506
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Como o valor p é muito menor que o nível de significância de 1%, rejeitamos a hipótese nula de homogeneidade de variâncias. Portanto, há diferenças estatisticamente significativas nas variâncias entre os grupos, indicando que as variâncias não são iguais. Neste caso, usareamos a ANOVA de Welch.

```
oneway.test(PESO ~ RACACORMAE, data = questao_4, var.equal = FALSE)
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  PESO and RACACORMAE
## F = 6.7232, num df = 4.00, denom df = 553.91, p-value = 2.743e-05
```

O resultado do teste ANOVA de Welch mostra que há diferenças estatisticamente significativas nos pesos médios ao nascer entre diferentes grupos raciais/culturais das mães, nesse caso, rejeitamos a hipótese nula, ou seja, existe pelo menos uma diferença no peso médio ao nascer entre os diferentes grupos de raça/cor das mães. Faremos agora o teste de Games-Howell, este teste é útil para identificar especificamente quais grupos diferem entre si quando a homogeneidade das variâncias não é assumida.

```
library(PMCMRplus)

# Teste de Games-Howell
games_howell_resultado <- summary(gamesHowellTest(PESO ~ RACACORMAE,
                                    data = questao_4, var.equal = FALSE))
```

```
##
##  Pairwise comparisons using Games-Howell test

## data: PESO by RACACORMAE

## alternative hypothesis: two.sided

## P value adjustment method: none
```

```
## H0

##                           q value   Pr(>|q|)
## Preta - Branca == 0      -6.916 1.0138e-05 ***
## Amarela - Branca == 0    -2.127 0.56040870
## Parda - Branca == 0      -1.836 0.69233606
## Indígena - Branca == 0    1.613 0.78435062
## Amarela - Preta == 0      0.327 0.99937040
## Parda - Preta == 0        5.683 0.00056434 ***
## Indígena - Preta == 0     2.535 0.38485963
## Parda - Amarela == 0      1.724 0.74043651
## Indígena - Amarela == 0   2.270 0.49774236
## Indígena - Parda == 0     1.764 0.72366240

## ---

## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
print(games_howell_resultado)
```

```
##
##  Pairwise comparisons using Games-Howell test

## data: PESO by RACACORMAE

##          Branca  Preta   Amarela Parda
## Preta    1e-05   -       -       -
## Amarela  0.56041 0.99937 -       -
## Parda    0.69234 0.00056 0.74044 -
## Indígena 0.78435 0.38486 0.49774 0.72366

##
## P value adjustment method: none

## alternative hypothesis: two.sided
```

O resultado do teste de Games-Howell indica diferenças significativas entre mães pretas e brancas e entre mães pretas e pardas com variações estatísticas relevantes que podem necessitar de investigação adicional para entender as causas subjacentes. As demais comparações não mostraram diferenças estatísticas significativas.