Taller 2 en grupo 1.5 puntos nota final

Estadística

02 enero, 2023

Índice

Nombre del grupo: Grupo NOMBRE

Autores

1. Florit Ensenyat, Jordi
2. Girón Rodríguez, Pau
3. Fornés Reynés, Josep Gabriel
4. Ferrer Fernández, Marc

INSTRUCCIONES

Comentarios:

- Para hacer los cálculos solicitados en los apartados anterior se deben eliminar los valores no disponibles (`NA`) de las variables.
- Siempre que sea posible se deben utilizar las funciones de R explicadas en clase para resolver los ejercicios.

- Debe redactar un documento utilizando Rmarkdown con las respuestas a estas preguntas y que incluya el código R utilizado. También debe generar (Knit) una versión HTML del documento.

El documento, en formato .Rmd y .html o .pdf , se debe **entregar a Aula Digital antes del 22 de diciembre**.

Parte 1. Descripción de datos tidyverse

Considera el conjunto de datos `examenes.csv` que contiene las siguientes variables:

- `gender`: sexo del estudiante masculino ("male") o femenino ("female").
- `race/ethnicity`: raza del estudiante (grupos desde el A hasta el E).
- `parental level of education`: nivel educativo de los padres desde algo de estudios secundarios ("some high school") hasta master ("master degree").
- `lunch`: tipo de precio que paga el estudiante por la comida que recibe en el centro educativo: normal ("standard") o con descuento ("free/reduced").
- `test preparation course`: si el estudiante ha tomado un curso de preparación para el examen de acceso a la Universidad, dos posibles valores: lo completó ("completed"), no lo tomó ("none").
- `math score`: nota que obtuvo el estudiante en la parte de matemáticas del examen de acceso a la Universidad. Valores del 0 al 100, donde el 100 es la máxima puntuación.
- `reading score`: nota que obtuvo el estudiante en la parte de lectura del examen de acceso a la Universidad. Valores del 0 al 100, donde el 100 es la máxima puntuación.
- `writing score`: nota que obtuvo el estudiante en la parte de redacción del examen de acceso a la Universidad. Valores del 0 al 100, donde el 100 es la máxima puntuación.

A continuación te presento la estructura del conjunto de datos:

```
library(readr)
datos <- read_csv("data/examenes.csv")
library(tidyverse)
glimpse(datos)
```

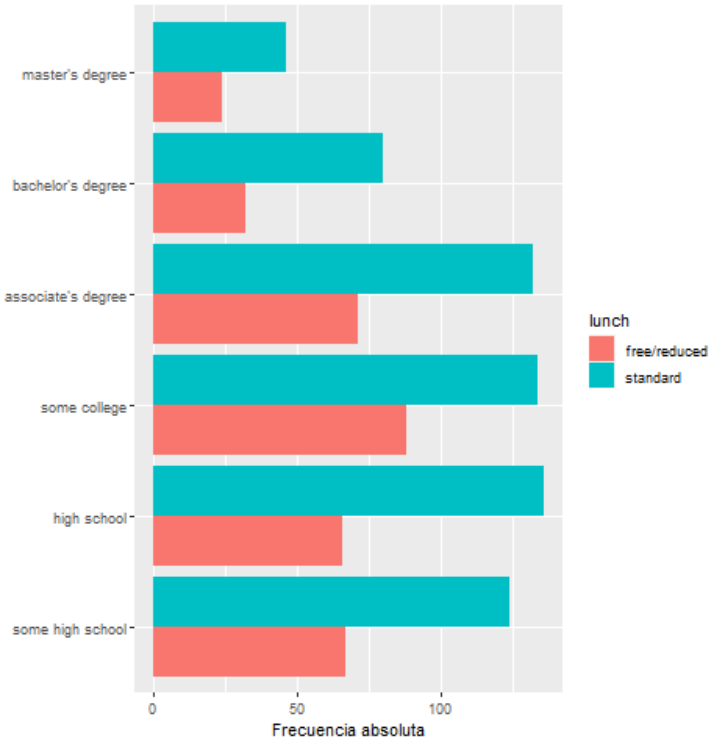
```
## Rows: 1,000
## Columns: 8
## $ gender                         <chr> "male", "female", "male", "male", "male"~
## $ `race/ethnicity`               <chr> "group A", "group D", "group E", "group ~
## $ `parental level of education` <chr> "high school", "some high school", "some~
## $ lunch                          <chr> "standard", "free/reduced", "free/reduce~
## $ `test preparation course`      <chr> "completed", "none", "none", "none", "co~
## $ `math score`                   <dbl> 67, 40, 59, 77, 78, 63, 62, 93, 63, 47, ~
## $ `reading score`                <dbl> 67, 59, 60, 78, 73, 77, 59, 88, 56, 42, ~
## $ `writing score`                <dbl> 63, 55, 50, 68, 68, 76, 63, 84, 65, 45, ~
```

Cuestión 1. 1 punto

a. Describe lo que se calcula con el siguiente código

```
datos <- drop_na(datos)
datos %>% group_by(gender)%>%
  summarise(frecuencia=length(gender))%>%
  mutate(porcentaje=frecuencia/sum(frecuencia)*100)
df<- datos %>% group_by(`race/ethnicity`) %>%
 summarise(frecuencia=length(`race/ethnicity`)) %>%
 arrange(desc(frecuencia))
df
```

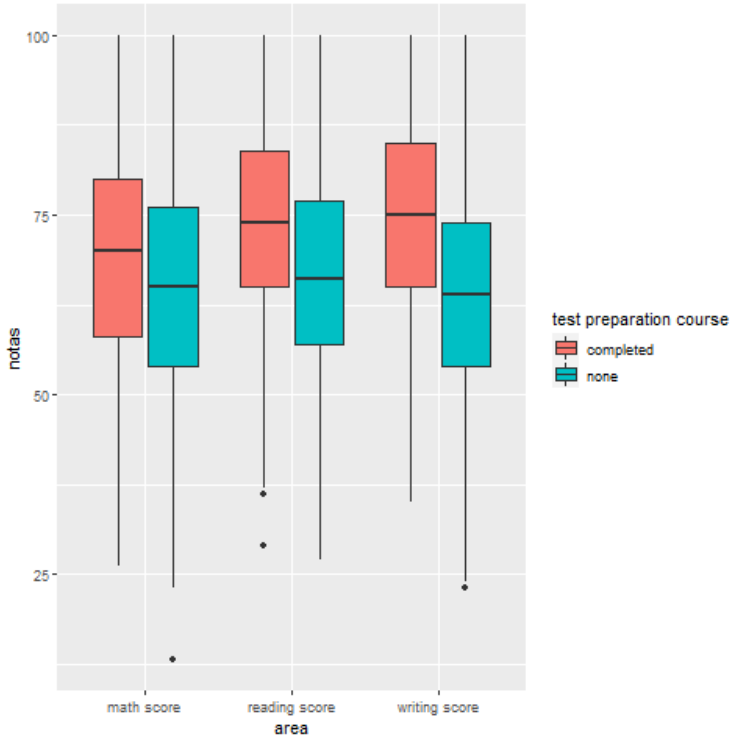
b. Da el código de ggplot2 que genera este gráfico. Comenta los resultados



Cuestión 2. 1 punto

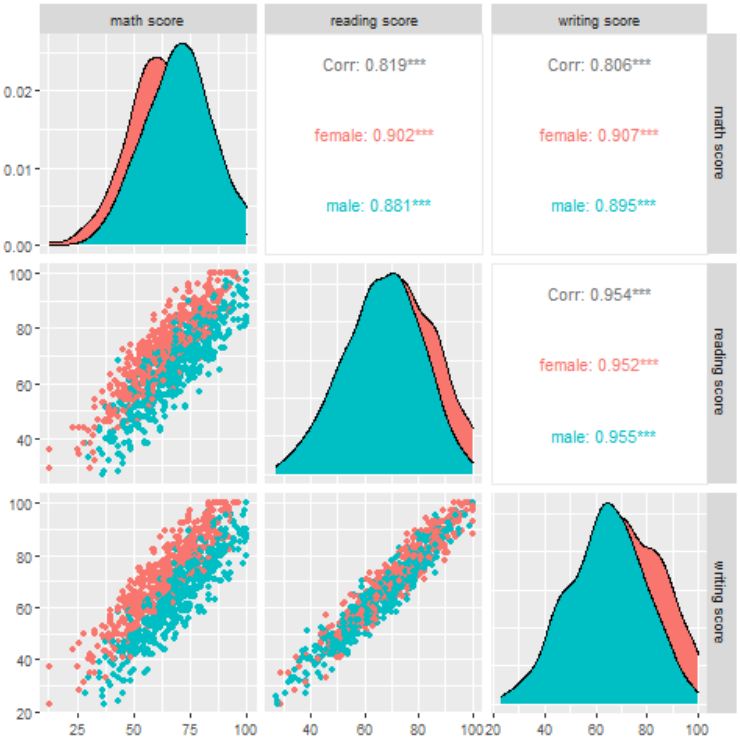
Explica lo que se obtiene en la tibble `df2` y dibuja y analiza el gráfico que se genera en el contexto del problema.

```
df2<- datos %>%
  tidyr::pivot_longer(
    cols=contains("score"),
    names_to="area", values_to="notas") %>%
  select("area", `test preparation course`, "notas")
```



Cuestión 3. 1 punto

Genera el gráfico (con `ggplot2`) e INTERPRETA el siguiente gráfico



Parte 2. Estadística Inferencial

Nos piden analizar los datos de la web de Airbnb para Mallorca de septiembre de 2022 y junio de 2022 (se adjuntan) los ficheros .

Cargad en un `dataframe` los datos del fichero `listings.csv` (descomprimido a partir de `listings.csv.gz`).

Vamos a cargar los datos y seleccionar algunas variables `price, review_scores_rating` y `neighbourhood_cleansed`.

```
#library(tidyverse)
data_june=readr::read_csv("data/listings_mallorca_june_2022.csv")#
```

```
## Rows: 18298 Columns: 74
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (24): listing_url, name, description, neighborhood_overview, picture_ur...
## dbl  (37): id, scrape_id, host_id, host_listings_count, host_total_listings_...
## lgl   (8): host_is_superhost, host_has_profile_pic, host_identity_verified, ...
## date  (5): last_scraped, host_since, calendar_last_scraped, first_review, la...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
print(object.size(data_june),units="MB",stardard="SI")
```

```
## 47.1 Mb
```

```
#glimpse(data_june)
gsub(pattern="[\\$]|[,]",replacement="",data_june$price[1:10])
```

```
##  [1] "118.00" "173.00" "120.00" "300.00" "372.00" "195.00" "237.00" "286.00"
##  [9] "135.00" "86.00"
```

```
as.numeric(gsub(pattern="[\\$]|[,]",replacement="",data_june$price[1:10]))
```

```
##  [1] 118 173 120 300 372 195 237 286 135  86
```

```
data_june$price=as.numeric(gsub(pattern="[\\$]|[,]",replacement="",data_june$price))
head(data_june$price)
```

```
## [1] 118 173 120 300 372 195
```

```
class(data_june$price)
```

```
## [1] "numeric"
```

```
data_june= data_june %>% select(price,review_scores_rating,neighbourhood_cleansed)
glimpse(data_june)
```

```
## Rows: 18,298
## Columns: 3
## $ price                   <dbl> 118, 173, 120, 300, 372, 195, 237, 286, 135, 86~
## $ review_scores_rating    <dbl> 4.73, 4.17, NA, 5.00, 5.00, 4.82, 5.00, 4.90, N~
## $ neighbourhood_cleansed <chr> "Sóller", "Pollença", "Sóller", "Alcúdia", "Mur~
```

Pregunta 1. 1 punto

a. Calcular una estimación puntual de la media para la variable `price` y el error estándar del estimador.

b. Calcular un intervalo de confianza, al nivel de confianza del 95%, para la variable `price`.

Solución

a)

```
media = mean(data_june$price)
media
```

```
## [1] 286.2452
```

```
n = length(data_june$price)
error = sd(data_june$price)/sqrt(n)
error
```

```
## [1] 6.333008
```

b)

```
alpha = 1-0.95
intconf = c(media-qt(1-(alpha/2),df=n-1)*error,media+qt(1-(alpha/2),df=n-1)*error)
intconf
```

```
## [1] 273.8319 298.6585
```

Pregunta 2. 2 puntos

a. Supongamos que un responsable de Airbnb asegura que el porcentaje de los valores de `review_scores_rating` mayor o igual que 4.5 es del 79.5% . Contrastad esta hipótesis con los datos de Mallorca.

b. Calcular un intervalo de confianza del 95% asociado a este contraste por el método exacto, el de Wilson y el Laplace.

Solución

a) Queremos contrastar la siguiente hipótesis: H0: mu = 4.5 H1: mu > 4.5

```
n = length(data_june$review_scores_rating)
##n
media = mean(data_june$review_scores_rating)
##media
t.test(data_june$review_scores_rating, mu=4.5, alternative="greater", conf.level=0.95)
```

```
##
##  One Sample t-test
##
## data:  data_june$review_scores_rating
## t = 26.817, df = 12662, p-value < 2.2e-16
## alternative hypothesis: true mean is greater than 4.5
## 95 percent confidence interval:
##  4.635406      Inf
## sample estimates:
## mean of x
##  4.644254
```

b)

Método exacto:

```
epitools::binom.exact(table(data_june$review_scores_rating>4.5)["TRUE"],length(data_june$review_scores_
```

```
##         x     n proportion     lower     upper conf.level
## TRUE 9565 18298  0.5227347 0.5154671 0.5299951       0.95
```

Método Wilson:

```
epitools::binom.wilson(table(data_june$review_scores_rating>4.5)["TRUE"],length(data_june$review_scores
```

```
##         x     n proportion     lower     upper conf.level
## TRUE 9565 18298  0.5227347 0.5154936 0.5299663       0.95
```

Método Laplace:

```
epitools::binom.approx(table(data_june$review_scores_rating>4.5)["TRUE"],length(data_june$review_scores
```

```
##         x     n proportion     lower     upper conf.level
## TRUE 9565 18298  0.5227347 0.5154976 0.5299719       0.95
```

Pregunta 3. 2 puntos

Considera ahora los datos de `price` para del mes junio de 2022 de las dos zonas de Mallorca con más apartamentos vacacionales

```
sort(table(data_june$neighbourhood_cleansed),decreasing = TRUE)[1:4]
```

```
##
##          Pollença Palma de Mallorca           Alcúdia          Santanyí
##              2625              1923              1897              1015
```

a. Decidid si las varianzas del precio en las dos zonas son iguales o diferentes. Considera que las distribuciones de los valores de precio en las poblaciones son normales.
b. Dad un intervalo de confianza del 95% para comparar las varianzas. Interpretar adecuadamente el resultado.

Solución

a)

```
data_june %>%
  filter(neighbourhood_cleansed=="Palma de Mallorca") %>%
  select(1)
```

```
## # A tibble: 1,923 x 1
##    price
##    <dbl>
##  1    80
##  2   171
##  3    65
##  4   110
##  5    75
##  6    65
##  7    25
##  8    46
##  9   238
## 10   120
## # ... with 1,913 more rows
```

```
##EnvStats:: varTest(palma,conf.level=0.95)$conf.int
```

b)

Pregunta 4. 2 puntos

a. A partir de los resultados del apartado anterior contrastad la hipótesis de que los precios medios en las dos ciudades son iguales contra que son distintos.
b. Calcular un intervalo de confianza del 95% para la diferencia de precios.

Solución