Tarea 3

Code

La Belleza de los Datos

Consideramos un extracto de las respuestas de la Encuesta CASEN 2017, adjuntas en la siguiente tabla:

Tabla 1. Variables a utilizar de la encuesta CASEN 2017

Variables	Descripción
region	Número de la región
sexo	1 (hombre), 2 (mujer)
edad	Edad en años
e_civil	Estado civil
o1	Trabajó al menos una hora
o2	Actividad de al menos una hora
o3	Ausencia temporal al empleo
o6	Búsqueda de trabajo
o25a_hr	Horas que tarda en viaje al trabajo
o25a_min	Minutos que tarda en viaje al trabajo
o25b	Veces a la semana en que realiza el viaje

Preparación de entorno

Iniciación de librerías

Hide

```
library(readr)
library(dplyr)
library(ggplot2)
library(tidyr)
library(knitr)
```

Abrir la base de datos

Hide

```
casen <- read_csv("Tarea_3_casen.csv")
```

```
Rows: 216439 Columns: 12
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
dbl (12): region, expc, sexo, edad, ecivil, o1, o2, o3, o6, o25a_hr, o25a_min, o25b

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.
```

Hide

```
View(Tarea_3_casen)
```

Pregunta 1

Presentamos en una tabla de calidad los estadísticos descriptivos del tiempo de viaje al trabajo.

Primero creamos una base de datos para trabajar en el problema.

Hide

```
df <- casen
```

a) Chile

Luego, creamos una variable llamada `totalmin` , que corresponde al total de minutos de trabajo (horas más minutos), y `total_time` , que corresponde al tiempo en formato `ddttmm` (para trabajar en `dplyr`).

Hide

```
df <- df %>%
  mutate(
    totalmin = ifelse(is.na(o25a_hr),
                      0,
                      o25a_hr * 60) + ifelse(is.na(o25a_min),
                                             0,
                                             o25a_min)
  )
minutes_to_hhmm <- function(minutes) {
  hours <- floor(minutes / 60)
  mins <- minutes %% 60
  sprintf("%02d:%02d", hours, mins)
}
df <- df %>%
  mutate(total_time = sapply(totalmin, minutes_to_hhmm))
```

Luego, filtramos la base de datos df para las personas que trabajaron al menos una hora. Esto es, o1=1. También filtramos para las instancias donde el tiempo de viaje (`o25a_min` y `o25a_hr`) son 99, porque de la tarea anterior también sabemos que son `NAs` .

Hide

```
df <- df %>%
  mutate(o1 = as.factor(o1)) %>%
  filter(o1 == 1) %>%
  filter(o25a_min != 99) %>%
  filter(o25a_hr != 99)
```

Y realizamos una tabla de calidad:

Hide

```
# calculamos los estadísticos descriptivos
stats <- df %>%
  summarise(
    Media = mean(totalmin, na.rm = TRUE),
    Mediana = median(totalmin, na.rm = TRUE),
    Desviacion_estandar = sd(totalmin, na.rm = TRUE),
    Minimo = min(totalmin, na.rm = TRUE),
    Primer_cuartil = quantile(totalmin, 0.25, na.rm = TRUE),
    Tercer_cuartil = quantile(totalmin, 0.75, na.rm = TRUE),
    Maximo = max(totalmin, na.rm = TRUE)
  )

# Crear una tabla con los estadísticos descriptivos
tabla_calidad <- as.data.frame(t(stats))
colnames(tabla_calidad) <- "Valor"
tabla_calidad$Estadistico <- rownames(tabla_calidad)
rownames(tabla_calidad) <- NULL

# Mostrar la tabla usando knitr::kable
kable(tabla_calidad, col.names = c( "Valor", "Estadístico"), caption = "Tabla de Estadísticos
Descriptivos de la Variable totalmin. (Fuente: CASEN 2017)")
```

Tabla de Estadísticos Descriptivos de la Variable totalmin. (Fuente: CASEN 2017)

Valor	Estadístico
30.12542	Media
20.00000	Mediana
31.25440	Desviacion_estandar
0.00000	Minimo
10.00000	Primer_cuartil
40.00000	Tercer_cuartil
239.00000	Maximo

Hide

```
NA
```

a) Región Metropolitana

Utilizamos la misma base de datos `df` , sin embargo, debemos filtrar para la RM, esto es `region=13`

Hide

```
dfRM <- df %>%
  mutate(region = as.factor(region)) %>%
  filter(region == 13)
```

Luego realizamos la misma Tabla de calidad, pero esta vez con el filtro ya hecho.

Hide

```
# calculamos los estadísticos descriptivos
stats <- dfRM %>%
  summarise(
    Media = mean(totalmin, na.rm = TRUE),
    Mediana = median(totalmin, na.rm = TRUE),
    Desviacion_estandar = sd(totalmin, na.rm = TRUE),
    Minimo = min(totalmin, na.rm = TRUE),
    Primer_cuartil = quantile(totalmin, 0.25, na.rm = TRUE),
    Tercer_cuartil = quantile(totalmin, 0.75, na.rm = TRUE),
    Maximo = max(totalmin, na.rm = TRUE)
  )

# Creamos la tabla para presentarla
RM <- as.data.frame(t(stats))
colnames(RM) <- "Valor"
RM$Estadistico <- rownames(RM)
rownames(RM) <- NULL

# Mostramos la tabla usando knitr::kable
kable(RM, col.names = c( "Valor", "Estadístico"), caption = "Tabla de Estadísticos Descriptiv
os de la Variable totalmin. (Fuente: CASEN 2017)")
```

Tabla de Estadísticos Descriptivos de la Variable totalmin. (Fuente: CASEN 2017)

Valor	Estadístico
40.60433	Media
30.00000	Mediana
34.14251	Desviacion_estandar
0.00000	Minimo
15.00000	Primer_cuartil
60.00000	Tercer_cuartil
239.00000	Maximo

Hide

```
NA
```

Respecto de las medias, los tiempos de viaje en la Región Metropolitana son significativamente mayores que a nivel nacional (10 minutos más aproximadamente). Por tanto, los tiempos de viaje al trabajo en esta región son más largos. Esto puede deberse a que existe mayor tráfico en la capital y probablemente también se

saturen más los modos de transporte público en horarios de trabajo. Esta tendencia se ve también en los cuartiles (primer cuartil con 5 minutos más y tercer cuartil con 20 minutos más) donde siempre fue superior en la RM.

b Probabilidades

Se pregunta por la probabilidad de que, siendo mujer y realizando el viaje al menos 5 veces a la semana, la persona encuestada se demore una hora o más en llegar al trabajo.

Primero, mutamos la variable sexo para que sea factor. Luego, filtramos para quienes hacen este viaje 5 veces o más en su semana.

Hide

```
df <- df %>%
  mutate(sexo = as.factor(sexo)) %>% # factor
  filter(o25b > 4) # 5 viajes o más
```

Calculamos el total de mujeres.

Hide

```
total_mujeres
```

```
[1] 26941
```

Luego calculamos el número de mujeres que se demoran 60 min o más en su viaje.

Hide

```
mujeres_60_mas <- df %>%
  filter(sexo == 2 & totalmin >= 60) %>%
  nrow()
mujeres_60_mas
```

```
[1] 4422
```

Y obtenemos la proporción

Hide

```
probabilidad <- mujeres_60_mas / total_mujeres
probabilidad
```

```
[1] 0.1641364
```

Pregunta 2

Reiniciamos el estudio con casen

Hide

```
df <- casen
```

a) Ocupación

Presentamos una tabla de calidad de frecuencia relativa de las personas ocupadas y no ocupadas de la muestra (y también las inactivas para ser consencuente con el formato CASEN).

Para ello, debemos considerar quiénes están ocupados y quiénes no. Debemos considerar que el nivel de ocupación es sólo para mayores de 15 años. Personas ocupadas las definiremos como (condicional `OR`):

- Quien trabajó.
- Quien realizó una actividad no remunerada (como labores de cuidado).
- Quien está ausente temporalmente de su empleo.

Personas desocupadas como (condicional `AND`):

- Quien no trabajó.
- Quien no realizó actividades.
- Quien no estaba ausente temporalmente en su ocupación.
- Quien está buscando trabajo remunerado o realizando gestión para iniciar actividad

Personas inactivas como (condicional `AND`),

- Personas que no están ausentes de sus ocupaciones
- Personas que no buscan trabajo remunerado o iniciar actividad.

Ocupadas	No ocupadas	Inactivas
o1=1 o o2=1 o o3=1	o1=2 y o2=2 y o3=2 y o6=1	o3=2 y o6=2

Generamos la variable de ocupación.

Hide

```
df <- df %>%
  mutate(ocup = case_when(
    edad>=15 & o1 == 2 & o2 == 2 & o3 == 2 & o6 == 1 ~ 2,
    edad>=15 & (o1 == 1 | o2 == 1 | o3) == 1 ~ 1,
    edad>=15 & o3 == 2 & o6 == 2 ~ 3,
    TRUE ~ NA_real_
  )) %>%
  mutate(ocup = factor(ocup, levels = c(1, 2, 3))) %>%
  mutate(ocup = recode(ocup,
                       `1` = 'Ocupado',
                       `2` = 'No ocupado',
                       `3` = 'Inactivo')) %>%
  mutate(ocup = factor(ocup, levels = c('Ocupado', 'No ocupado', 'Inactivo')))
```

Y generamos la tabla como sigue:

Hide

```
frecuencias <- df %>%
  filter(!is.na(ocup)) %>%
  count(ocup) %>%
  mutate(frecuencia = n / sum(n))

tabla <- frecuencias %>%
  select(ocup, frecuencia) %>%
  rename(Ocupación = ocup, 'Frecuencia relativa' = frecuencia)

tabla %>%
  kable(caption = 'Tabla de Calidad de Frecuencias Relativas de Ocupación (Fuente: CASEN 201
7)')
```

Tabla de Calidad de Frecuencias Relativas de Ocupación (Fuente: CASEN 2017)

Ocupación	Frecuencia relativa
Ocupado	0.9554308
No ocupado	0.0445692

b) Ocupación por edad

Primero, filtramos por la edad para que corresponda al caso

Hide

```
df <- df %>%
  filter(!is.na(ocup))

head(df)
```

region <dbl>	expc <dbl>	sexo <dbl>	edad <dbl>	ecivil <dbl>	o1 <dbl>	o2 <dbl>	o3 <dbl>	o6 <dbl>	o25a_hr <dbl>
1	33	2	56	8	1	*NA*	*NA*	*NA*	0
1	33	2	21	8	1	*NA*	*NA*	*NA*	0
1	33	1	24	2	1	*NA*	*NA*	*NA*	0
1	33	1	28	2	1	*NA*	*NA*	*NA*	0
1	33	1	26	1	2	1	*NA*	*NA*	0
1	33	2	26	1	2	2	2	2	*NA*

6 rows | 1-10 of 13 columns

Y hacemos la tabla

Hide

```
tabla_calidad <- df %>%
  group_by(ocup) %>%
  summarise(
    Media = mean(edad, na.rm = TRUE),
    Mediana = median(edad, na.rm = TRUE),
    Desviacion_estandar = sd(edad, na.rm = TRUE),
    Minimo = min(edad, na.rm = TRUE),
    Primer_cuartil = quantile(edad, 0.25, na.rm = TRUE),
    Tercer_cuartil = quantile(edad, 0.75, na.rm = TRUE),
    Maximo = max(edad, na.rm = TRUE)
  )

tabla_calidad <- tabla_calidad %>% #trasponemos porque quedaba muy desordenada
  pivot_longer(cols = -ocup, names_to = "Estadístico", values_to = "Valor") %>%
  pivot_wider(names_from = ocup, values_from = Valor)

colnames(tabla_calidad) <- c('Estadístico',
                             'Ocupado',
                             'No ocupado')

tabla_calidad %>%
  kable(caption = 'Estadísticos Descriptivos de la Edad por Ocupación')
```

Estadísticos Descriptivos de la Edad por Ocupación

Estadístico	Ocupado	No ocupado
Media	45.49515	34.64206
Mediana	45.00000	30.00000
Desviacion_estandar	19.40707	13.79090
Minimo	15.00000	15.00000
Primer_cuartil	28.00000	23.00000
Tercer_cuartil	60.00000	45.00000
Maximo	117.00000	99.00000

Bonus

a) Histograma

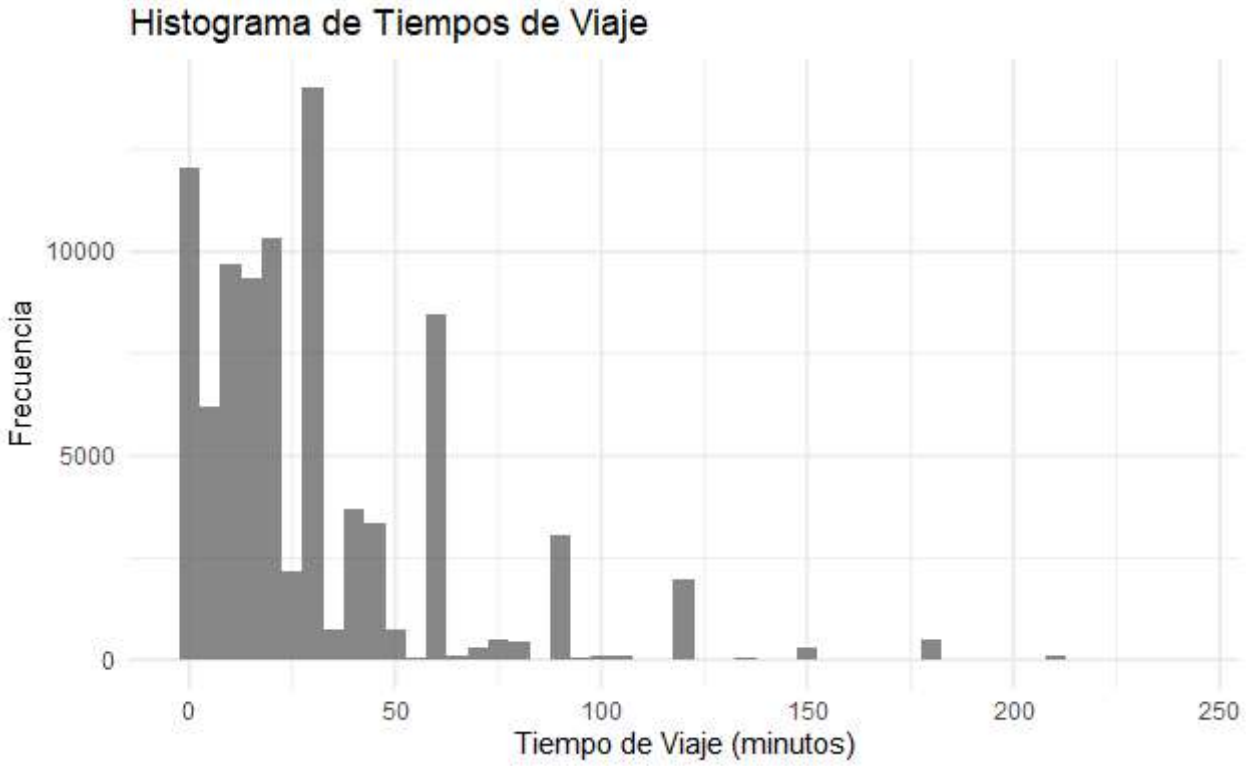
Presentamos en un histograma la distribución de tiempos de viaje de las personas ocupadas de la muestra como sigue.

Hide

```
df <- casen # reiniciamos el análisis
df <- df %>%
  mutate(
    totalmin = ifelse(is.na(o25a_hr),
                      0,
                      o25a_hr * 60) + ifelse(is.na(o25a_min),
                                             0,
                                             o25a_min)
  )
minutes_to_hhmm <- function(minutes) {
  hours <- floor(minutes / 60)
  mins <- minutes %% 60
  sprintf("%02d:%02d", hours, mins)
}
df <- df %>%
  mutate(o1 = as.factor(o1)) %>%
  filter(o1 == 1) %>%
  filter(o25a_min != 99) %>%
  filter(o25a_hr != 99) %>%
  mutate(total_time = sapply(totalmin, minutes_to_hhmm)) %>%
  mutate(ocup = case_when(
    edad>=15 & o1 == 2 & o2 == 2 & o3 == 2 & o6 == 1 ~ 2,
    edad>=15 & (o1 == 1 | o2 == 1 | o3) == 1 ~ 1,
    edad>=15 & o3 == 2 & o6 == 2 ~ 3,
    TRUE ~ NA_real_
  )) %>%
  mutate(ocup = factor(ocup,
                       levels = c(1, 2, 3))) %>%
  mutate(ocup = recode(ocup,
                       `1` = 'Ocupado',
                       `2` = 'No ocupado',
                       `3` = 'Inactivo')) %>%
  mutate(ocup = factor(ocup,
                       levels = c('Ocupado',
                                  'No ocupado',
                                  'Inactivo'))) %>%
  filter(ocup == 'Ocupado')
```

Hide

```
ggplot(df, aes(x=totalmin,)) +
  geom_histogram(binwidth = 5,
                 position = "identity",
                 alpha = 0.7) +
  labs(title = "Histograma de Tiempos de Viaje",
       x = "Tiempo de Viaje (minutos)",
       y = "Frecuencia") +
  theme_minimal()
```



b) No alcanzo

Disponible en GitHub (https://github.com/zoendeloi/MPP-ADI/tree/main/Tarea3)