Estadística descriptiva

Ejemplo práctico (datos numéricos)

MNPyR 2023-2

Índice

1. Estadisticas descriptivas básicas

1.1. Media muestral

Definición 1 *(Media muestral)*

Sea $x_1, x_2, ..., x_n$ un conjunto de n datos, entonces la media muestral se define como:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

La media muestral $\overline{X}$ de un conjunto de n datos se puede interpretar como el valor promedio de las n observaciones. Es un valor numérico que nos da una idea de dónde se están centrando los datos.

En R se puede calcular la la media muestral por medio de la función `mean()`.

1.2. Medianda muestral

Definición 2 *(Mediana muestral)*

Sea $x_1, x_2, ..., x_n$ un conjunto de n observaciones, entonces la mediana muestral se define como:

$$\text{Med} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{si } n \text{ es impar} \\ \\ \frac{1}{2}\left[x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n+2}{2}\right)}\right] & \text{si } n \text{ es par} \end{cases}$$

La mediana muestral Med de un conjunto de n datos se puede interpretar como el valor para el cual 50 % de los datos son menores al valor Med y el 50 % de los datos restantes son mayores a dicho valor.

En R se puede calcular la la mediana muestral por medio de la función `median()`.

1.3. Cuantiles y Percentiles

Definición 3 *Sea $x_1, x_2, ..., x_n$ un conjunto de n datos y sea $p \in (0, 1]$, entonces el cuantil al $100 * p\,\%$ es un número q que cumple las siguientes condiciones:*

1. *Una vez que se han ordenado los datos de menor a mayor, al menos una proporción del $100 * p\,\%$ de los datos son menores al valor q.*
2. *La proporción de datos restantes son mayores o iguales al valor q.*

Por ejemplo, supongamos que $p = 0.2$, entonces el cuantil al 20 % es el número q tal que 20 % de los datos caen a la izquierda de q y el 80 % de los datos restantes caen a la derecha de q.

Notas:

1. La mediana corresponde al cuantil al 50 %, es decir con $p = 0.5$
2. El cuantil no es necesariamente único y pueden existir diversas formas de calcularlo.
3. El cálculo de cuantiles es útil cuando el tamaño de la muestra es razonablemente grande.

4. Hay cuantiles que reciben nombres más específicos, por ejemplo:
    - *Percentiles*: Cuantiles al 1 %, 2 %, 3 %,...
    - *Deciles*: Cuantiles al 10 %, 20 %, 30 %, ...
    - *Cuartiles*: Cuantiles al 25 %, 50 % y 75 %

En R se pueden calcular los cuantiles por medio de la función `quantile()`.

2. Gráficas en R

Una forma de entender con mayor claridad la información que se este analizando, es la representación de la información mediante gráficas y tablas. Representar la información de esta manera es muy útil debido a que dan un recurso visual que facilitara el análisis de la información al mostrar una distribución de las variables que estemos estudiando además de patrones.

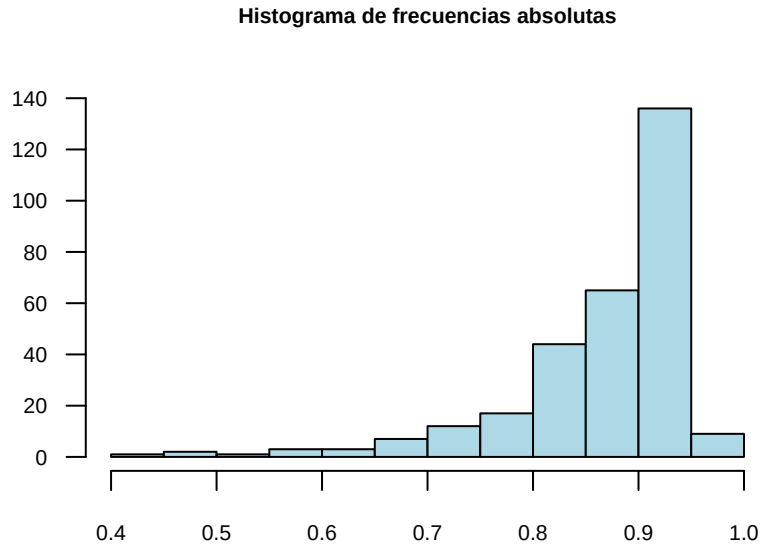
Las gráficas para datos continuos más comunes son:

- Histograma de frecuencias absolutas:
- Histograma de densidades.
- Función de distribución empírica.
- Box Plot (diagrama de caja).

2.1. Histograma de frecuencias absolutas

El histograma de frecuencias absolutas consiste en dividir el rango de los datos en varios intervalos de igual longitud, contar el número de observaciones en cada intervalo y graficar los conteos como longitudes de barra en un histograma.

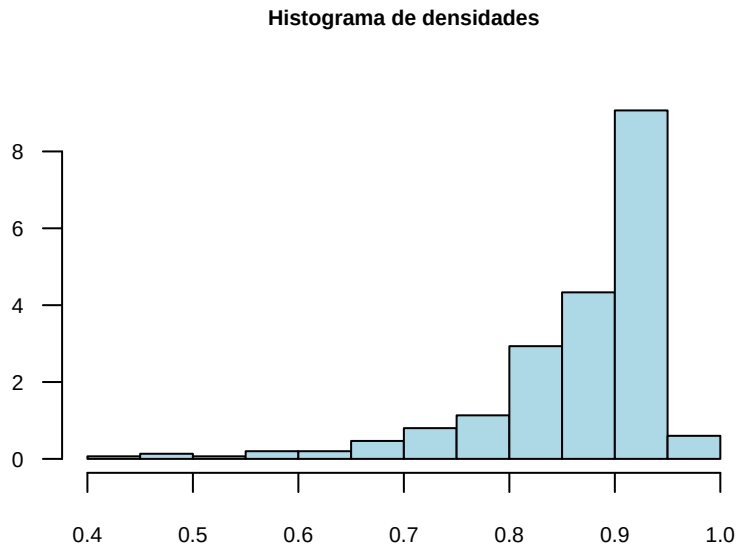
Se utiliza la función `hist()` con el argumento `freq=TRUE`, es decir, `hist(freq = TRUE)` para graficar el histograma de frecuencias absolutas.



2.2. Histograma de densidades

El histograma de densidades consiste en dividir el rango de los datos en varios intervalos de igual longitud, contar el número de observaciones en cada intervalo y graficar las frecuencias relativas como el área de las barras en un histograma.

Se utiliza la función `hist()` con el argumento `freq=FALSE`, es decir, `hist(freq = FALSE)` para graficar el histograma de densidades.



2.3. Función de distribución empírica

Definición 4 *(Función de distribución empírica)*

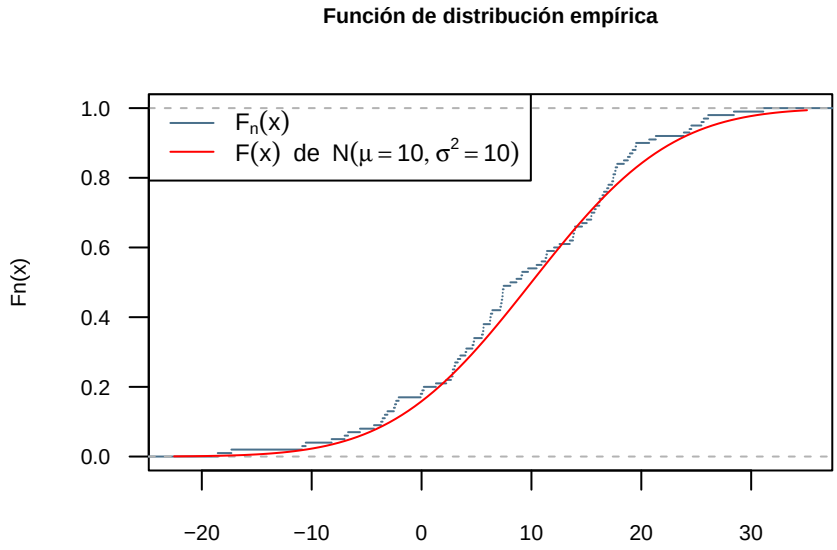
Sea $x_1, x_2, ..., x_n$ un conjunto de n datos, entonces la función de distribución empírica se define como:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{x_i \leq x\}}$$

o de forma equivalente:

$$F_n(x) = \begin{cases} 0 & si\ x < x_{(1)} \\ \frac{i}{n} & si\ x_{(i)} \leq x < x_{(i+1)},\ i \in \{1, ..., n-1\} \\ 1 & si\ x \geq x_{(n)} \end{cases}$$

La función de distribución empírica también es conocida como función de distribución muestral. En R se utiliza la función `ecdf()` (empirical cumulative distribution function) para calcular $F_n(x)$ y las funciones `plot(ecdf())` para graficar dicha función.



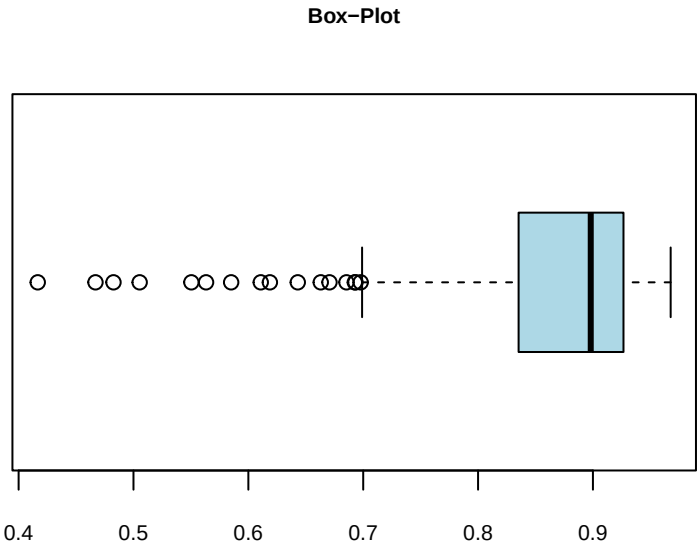
2.4. Box Plot

También conocida como diagrama de caja. En esta gráfica los cuartiles Q_3 (3-er Cuartil) y Q_1 (1-er Cuartil) de los datos están representados por la parte superior e inferior de la "caja" respectivamente, y la mediana Q_2 está representada por un segmento de línea horizontal dentro de la "caja".

Las líneas punteadas se extienden desde los extremos de la caja hasta los valores adyacentes que se definen como sigue.

- El valor adyacente inferior (bigote inferior) se define como la observación más pequeña que es mayor o igual a $Q_1 - 1.5 * IQR$, con $IQR = Q_3 - Q_1$, el rango intercuartil.
- El valor adyacente superior (bigote superior) se define como la observación más grande que es menor o igual a $Q_3 + 1.5 * IQR$, con $IQR = Q_3 - Q_1$, el rango intercuartil.

Si cualquier observación cae fuera del rango de los dos valores adyacentes ("bigotes"), se denomina valor "outlier", "atípico" o "extremo" y se representa como un punto individual. Se utiliza la función `boxplot()` para graficar el Box Plot.



3. Ejemplo

3.1. Datos

Trabajaremos con los datos `Estadísticas censales a escalas geoelectorales` del `Censo de Población y Vivienda 2020` disponibles en el siguiente enlace.

El propósito del Censo 2020 es producir información sobre el volumen, la estructura y la distribución espacial de la población, así como de sus principales características demográficas, socioeconómicas y culturales. El Censo de Población y Vivienda 2020 (Censo 2020) se realizó del **2 al 27 de marzo de 2020**.

Nuestro objetivo será calcular estadísticas descriptivas y visualizar los datos sobre **porcentajes** de:

- Viviendas particulares habitadas que disponen de teléfono celular.
- Población sin afiliación a servicios de salud.
- Hogares censales con persona de referencia mujer.

Debemos tener en cuenta las siguientes consideraciones:

- Trabajaremos todo a nivel de los distritos electorales, será necesario entonces trabajar con la base de datos `INE_DISTRITO_2020.CSV`.
- Para identificar las variables (columnas) con las que vamos a trabajar necesitaremos el descriptor de indicadores a nivel de los distritos electorales, será necesario entonces trabajar con el descriptor `Descriptor_indicadores_ECEG_Distrito_2020.csv.csv`

3.2. Lectura de los datos y descriptor

Para importar los datos en R se utiliza la función `read.csv()`. El argumento `encoding = "latin1"` sirve para que se lean de forma correcta los acentos.

Nota: Si `encoding = "latin1"` no sirve, entonces probar con `encoding = "UTF-8"`

```
# 1.1 Lectura de datos y descriptor ---------------------------------------

## Guardamos la base en el DataFrame "Datos"
Datos <- read.csv("INE_DISTRITO_2020.CSV",
                  encoding = "latin1")

## Guardamos el descriptor de variables en el DataFrame "Descriptor"
Descriptor <- read.csv("Descriptor_indicadores_ECEG_Distrito_2020.csv.csv",
                       encoding = "latin1")
```

3.3. Variables de interés

Con ayuda del descriptor de variables `Descriptor_indicadores_ECEG_Distrito_2020.csv.csv` identificaremos las variables que utilizaremos (Cuadro 1) para calcular los porcentajes y realizar el análisis descriptivo.

Variable	Descripción
NOM_ENT	Entidad Federativa
VPH_CEL	**Viviendas particulares habitadas** que disponen de teléfono celular
PSINDER	**Población** sin afiliación a servicios de salud
HOGJEF_F	**Hogares censales** con persona de referencia mujer
POBTOT	Población total
TVIVPARHAB	Total de viviendas particulares habitadas
TOTHOG	Total de hogares censales

Cuadro 1: Variables de interés.

Una vez que identificamos el nombre de las variables que vamos a utilizar, seleccionamos dichas variables de nuestro DataFrame `Datos`.

```
# 1.2 Variables de interés ------------------------------------------------

## Selección de variables
Datos <- Datos[,c("ENTIDAD", "DISTRITO", "NOM_ENT", "VPH_CEL", "PSINDER", "HOGJEF_F",
                  "POBTOT", "TVIVPARHAB", "TOTHOG")]
```

Para continuar con nuestro análisis, es necesario verificar la estructura de nuestros datos y corregirla si es necesario.

```
# Estructura de los datos
str(Datos)

# Cambianos la variable "NOM_ENT" de tipo caracter a tipo factor
Datos$NOM_ENT <- factor(Datos$NOM_ENT)

# Estructura de los datos
str(Datos)
```

3.4. Porcentajes (definir nuevas variables)

Es importante obtener los porcentajes respecto a la población de la que se esta haciendo referencia, para el primer punto por ejemplo, la población a la que se hace referencia son las **viviendas particulares habitadas**, de este modo para calcular el porcentaje correspondiente debemos dividir por el **total de viviendas particulares habitadas** (`TVIVPARHAB`).

El segundo punto se refiere a la población en general, será entonces **total de población** (`POBTOT`) la variable que usaremos para calcular el porcentaje de población sin afiliación a servicios de salud (`PSINDER`).

Para el tercer punto la población a la que se hace referencia son los **hogares censales**, entonces ocuparemos la variable referente al **total de hogares censales** (`TOTHOG`) para calcular el porcentaje correspondiente.

Nota: La función `attach()` nos permite acceder fácilmente a las "columnas" de un DataFrame. De modo que, en vez de escribir `DataFrame$columna`, podemos usar simplemente el nombre de la `columna`. Es importante usar `detach()` cuando se termina de usar `attach()`.

```
# 1.3 Porcentajes (nuevas variables) ------------------------------------

attach(Datos)
## Porcentaje de viviendas particulares habitadas que disponen de teléfono celular.
Datos$Porcentaje_VPH_CEL <- VPH_CEL/TVIVPARHAB

## Porcentaje de población sin afiliación a servicios de salud.
Datos$Porcentaje_PSINDER <- PSINDER/POBTOT

## Porcentaje de hogares censales con persona de referencia mujer.
Datos$Porcentaje_HOGJEF_F <- HOGJEF_F/TOTHOG
detach(Datos)
```

Una vez definidos los porcentajes podemos continuar con el análisis descriptivo.

4. Resultados

4.1. Porcentaje de viviendas particulares habitadas que disponen de teléfono celular

```
# 2. Estadísticas descriptivas ------------------------------------------

# Media muestral ------------------------------------------------------------
mean(Datos$Porcentaje_VPH_CEL)
```

```
## [1] 0.8672127
```

```
# Mediana muestral ----------------------------------------------------------
median(Datos$Porcentaje_VPH_CEL)
```

```
## [1] 0.898097
```

```
# Cuantiles y Percentiles ---------------------------------------------------
quantile(Datos$Porcentaje_VPH_CEL,
         probs = c(.25, .5, .75))
```

```
##       25%       50%       75%
## 0.8358418 0.8980970 0.9264946
```

Cuadro 2: Porcentaje de viviendas particulares habitadas que disponen de teléfono celular

Media	25 %	Mediana	75 %
86.72 %	83.58 %	89.81 %	92.65 %

Podemos interpretar las estadísticas del Cuadro 2 de la siguiente manera:

Para el año 2020, en los distritos electorales de todo el país:

- En promedio el porcentaje de viviendas particulares habitadas que disponen de teléfono celular es de 86.72 %.
- 25 % de los distritos electorales tienen un porcentaje de viviendas particulares habitadas que disponen de teléfono celular menor o igual a 83.58 %.
- 50 % de los distritos electorales tienen un porcentaje de viviendas particulares habitadas que disponen de teléfono celular menor o igual a 89.81 %.
- 75 % de los distritos electorales tienen un número de viviendas particulares habitadas que disponen de teléfono celular menor o igual a 92.65 %.

Para complementar el análisis se presentan las gráficas de la Figura 1

```
# 3. Visualización de datos ------------------------------------------------

# 3.1 Histograma de frec. abs ----------------------------------------------
hist(Datos$Porcentaje_VPH_CEL,
     breaks = "Sturges",
     freq = TRUE,
     col = "lightblue",
     main = "Histograma de frecuencias absolutas",
     cex.main = 0.7,
     cex.sub = 0.7,
     xlab = "% de viviendas particulares habitadas que disponen de teléfono celular",
     ylab = "",
     cex.lab = 0.7,
     cex.axis = 0.7,
     las = 1)

# 3.5 Box-Plot -------------------------------------------------------------
boxplot(Datos$Porcentaje_VPH_CEL,
        col = "lightblue",
        main = "Box-Plot",
        cex.main = 0.7,
        xlab = "% de viviendas particulares habitadas que disponen de teléfono celular",
        cex.lab = 0.7,
        cex.axis = 0.7,
        las = 1,
        horizontal = TRUE)
```

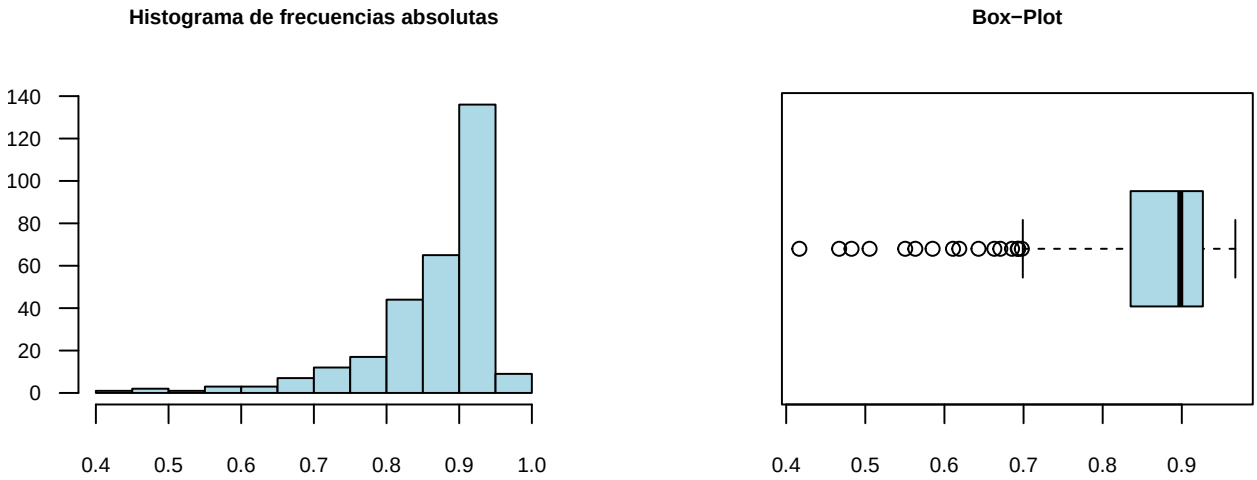


Figura 1: Porcentaje de viviendas particulares habitadas que disponen de teléfono celular

4.2. Porcentaje de población sin afiliación a servicios de salud

```
# 2. Estadísticas descriptivas ---------------------------------------------

# Media muestral -----------------------------------------------------------
mean(Datos$Porcentaje_PSINDER)
```

```
## [1] 0.2627213
```

```
# Mediana muestral ---------------------------------------------------------
median(Datos$Porcentaje_PSINDER)
```

```
## [1] 0.2661254
```

```
# Cuantiles y Percentiles --------------------------------------------------
quantile(Datos$Porcentaje_PSINDER,
         probs = c(.25, .5, .75))
```

```
##       25%       50%       75%
## 0.2007350 0.2661254 0.3103405
```

Cuadro 3: Porcentaje de población sin afiliación a servicios de salud

Media	1-Cuartil	2-Cuartil	3-Cuartil
26.27 %	20.07 %	26.61 %	31.03 %

Podemos interpretar las estadísticas del Cuadro 3 de la siguiente manera:

Para el año 2020, en los distritos electorales de todo el país:

- En promedio el porcentaje de población sin afiliación a servicios de salud es de 26.27 %.

- 25 % de los distritos electorales tienen un porcentaje de población sin afiliación a servicios de salud menor o igual a 20.07 %.
- 50 % de los distritos electorales tienen un porcentaje de población sin afiliación a servicios de salud menor o igual a 26.61 %.
- 75 % de los distritos electorales tienen un porcentaje de población sin afiliación a servicios de salud menor o igual a 31.03 %.

Para complementar el análisis se presentan las gráficas de la Figura 2

```
# 3. Visualización de datos ------------------------------------------------

# 3.1 Histograma de frec. abs ----------------------------------------------
hist(Datos$Porcentaje_PSINDER,
     breaks = "Sturges",
     freq = TRUE,
     col = "lightblue",
     main = "Histograma de frecuencias absolutas",
     cex.main = 0.7,
     cex.sub = 0.7,
     xlab = "% de viviendas particulares habitadas que disponen de teléfono celular",
     ylab = "",
     cex.lab = 0.7,
     cex.axis = 0.7,
     las = 1)

# 3.5 Box-Plot -------------------------------------------------------------
boxplot(Datos$Porcentaje_PSINDER,
        col = "lightblue",
        main = "Box-Plot",
        cex.main = 0.7,
        xlab = "% de viviendas particulares habitadas que disponen de teléfono celular",
        cex.lab = 0.7,
        cex.axis = 0.7,
        las = 1,
        horizontal = TRUE)
```

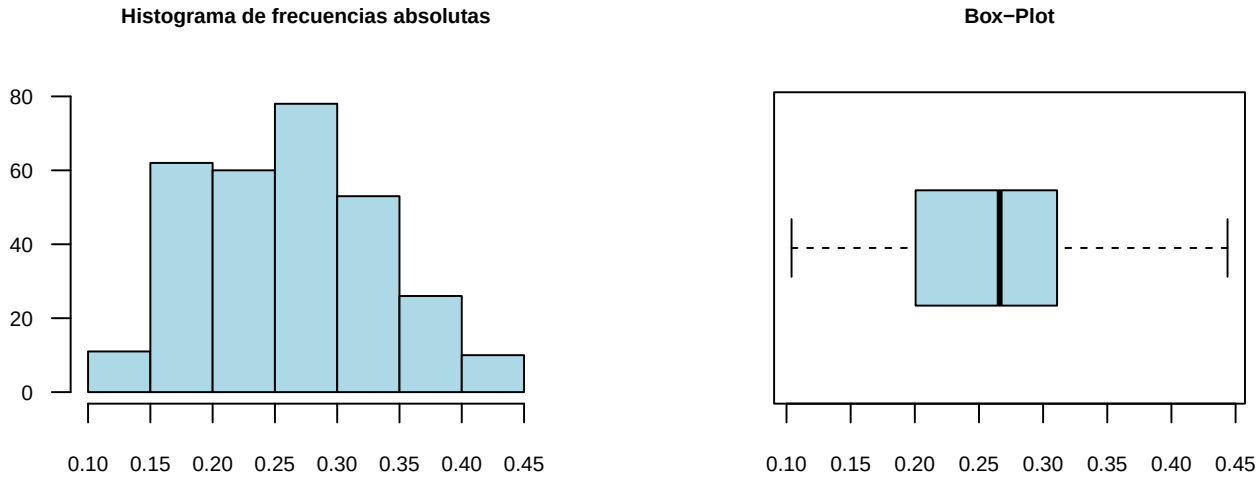


Figura 2: Porcentaje de población sin afiliación a servicios de salud

4.3. Porcentaje de hogares censales con persona de referencia mujer

```
# 2. Estadísticas descriptivas ------------------------------------------

# Media muestral ------------------------------------------------------
mean(Datos$Porcentaje_HOGJEF_F)
```

```
## [1] 0.3254836
```

```
# Mediana muestral ----------------------------------------------------
median(Datos$Porcentaje_HOGJEF_F)
```

```
## [1] 0.322916
```

```
# Cuantiles y Percentiles ---------------------------------------------
quantile(Datos$Porcentaje_HOGJEF_F,
         probs = c(.25, .5, .75))
```

```
##       25%       50%       75%
## 0.2998519 0.3229160 0.3506623
```

Cuadro 4: Porcentaje de hogares censales con persona de referencia mujer

Media	1-Cuartil	2-Cuartil	3-Cuartil
32.55 %	29.99 %	32.29 %	35.07 %

Podemos interpretar las estadísticas del Cuadro 4 de la siguiente manera:

Para el año 2020, en los distritos electorales de todo el país:

- En promedio el porcentaje de hogares censales con persona de referencia mujer es de 32.55 %.
- 25 % de los distritos electorales tienen un porcentaje de hogares censales con persona de referencia mujer menor o igual a 29.99 %.
- 50 % de los distritos electorales tienen un porcentaje de hogares censales con persona de referencia mujer menor o igual a 32.29 %.
- 75 % de los distritos electorales tienen un porcentaje de hogares censales con persona de referencia mujer menor o igual a 35.07 %.

Para complementar el análisis se presentan las gráficas de la Figura 3

```
# 3. Visualización de datos -------------------------------------------

# 3.1 Histograma de frec. abs -----------------------------------------
hist(Datos$Porcentaje_HOGJEF_F,
     breaks = "Sturges",
     freq = TRUE,
     col = "lightblue",
```

```
     main = "Histograma de frecuencias absolutas",
     cex.main = 0.7,
     cex.sub = 0.7,
     xlab = "% de viviendas particulares habitadas que disponen de teléfono celular",
     ylab = "",
     cex.lab = 0.7,
     cex.axis = 0.7,
     las = 1)

# 3.5 Box-Plot ------------------------------------------------------------
boxplot(Datos$Porcentaje_HOGJEF_F,
        col = "lightblue",
        main = "Box-Plot",
        cex.main = 0.7,
        xlab = "% de viviendas particulares habitadas que disponen de teléfono celular",
        cex.lab = 0.7,
        cex.axis = 0.7,
        las = 1,
        horizontal = TRUE)
```

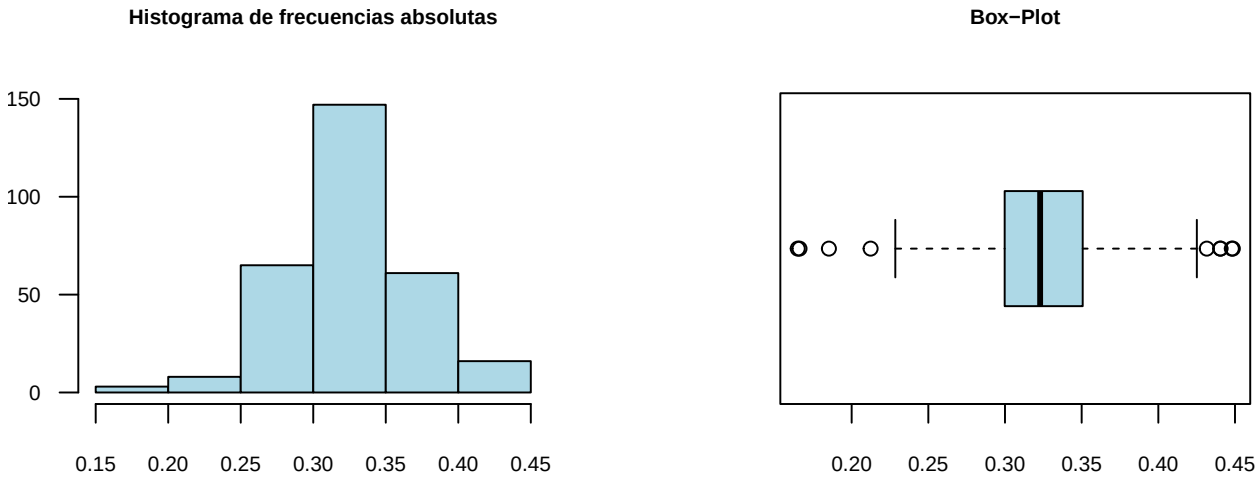


Figura 3: Porcentaje de de hogares censales con persona de referencia mujer

5. Visualización de datos 2 (Librerías: `ggplot2`, `dplyr` y `sf`)

5.1. Box Plot con información categórica

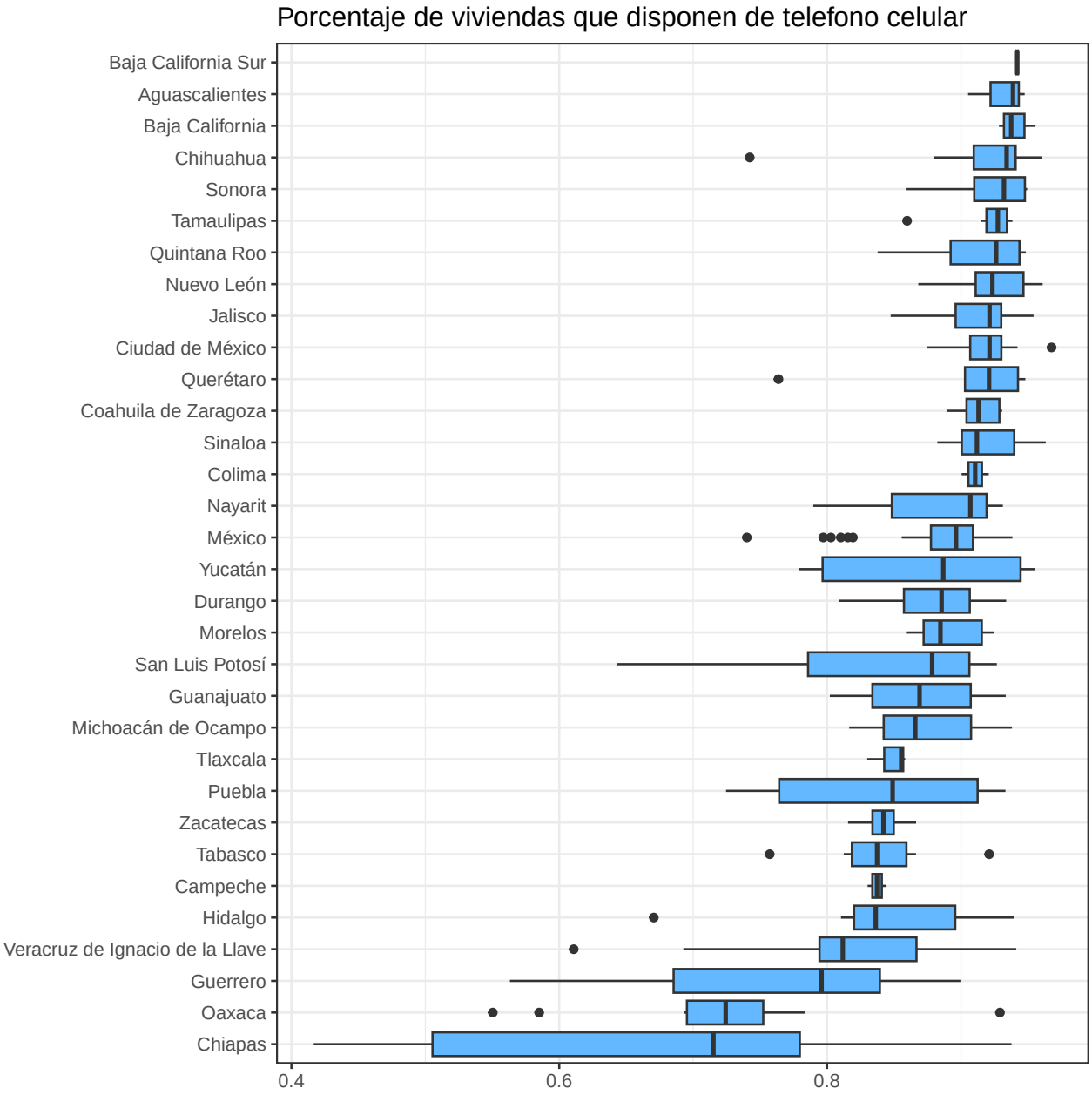
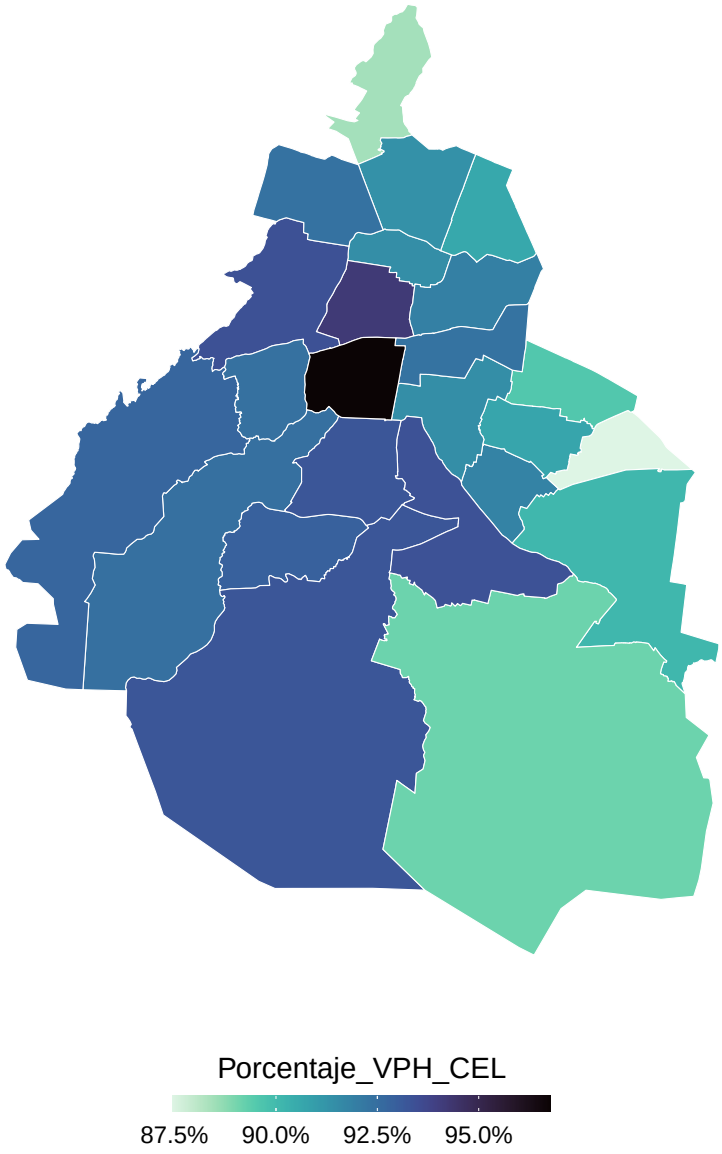


Figura 4: Ejemplo de una gráfica Box Plot por grupos con `ggplot2`

5.2. Ciudad de México

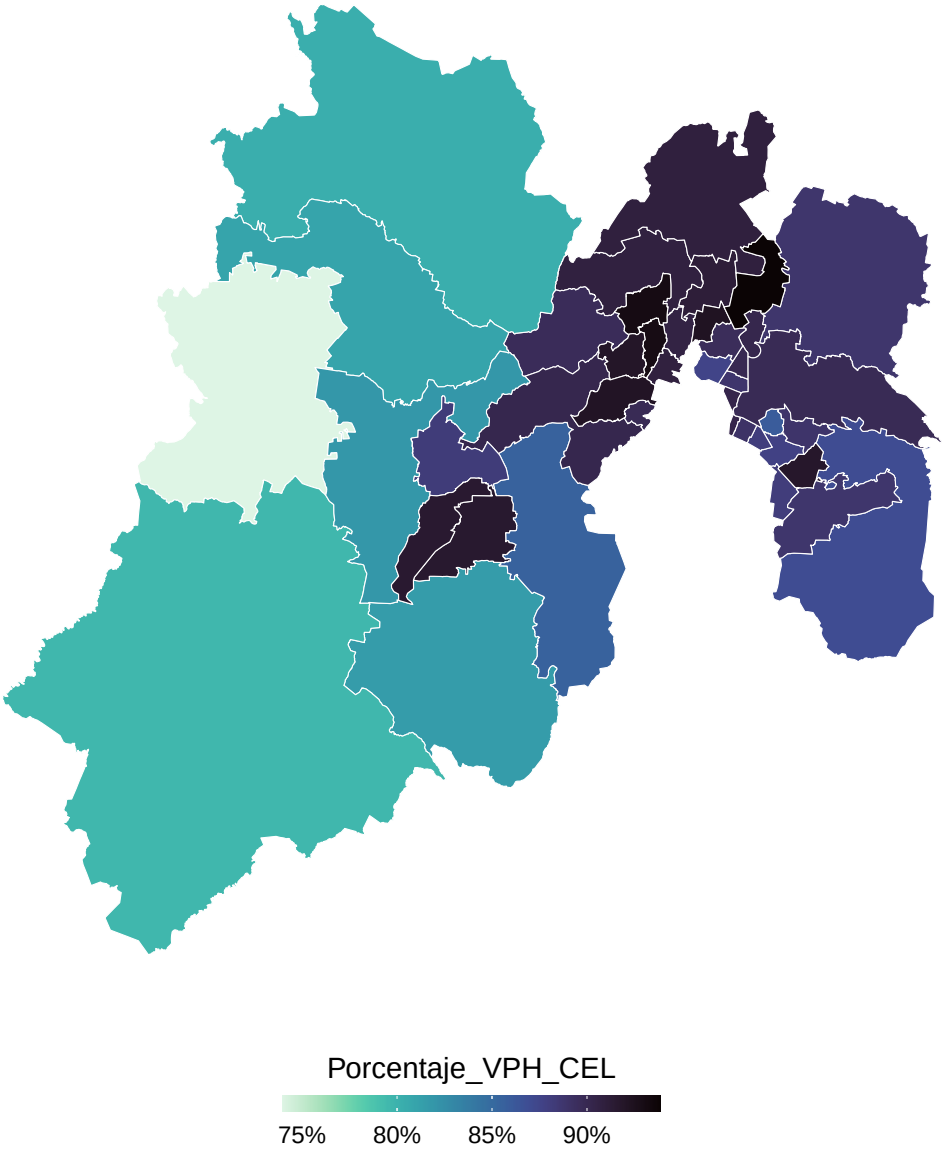


Fuente: Censo 2020 INEGI

5.3. Estado de México

Estado de México

Porcentaje de viviendas particulares habitadas que disponen de teléfono celular



Fuente: Censo 2020 INEGI