Estadística descriptiva con R

Mg. Jesús Gamboa

Noviembre 2022

Introducción

Este documento presenta y desarrolla, de manera teórico - práctica, los puntos concernientes a medidas estadísticas, tablas y gráficos, los cuales son importantes en la fase de análisis exploratorio.

Para la aplicación práctica, se utilizará el conjunto de datos (anonimizado) correspondiente a los postulantes a la UNALM en el ciclo 2018-I. Note que este conjunto de datos es una población pues contiene datos acerca de todos los postulantes en dicho examen de admisión (al menos los que sí rindieron el examen, pues algunos llegaron tarde o no se presentaron). Las variables contenidas en este conjunto de datos son:

- SEXO: Variable de naturaleza cualitativa nominal, cuyas categorías son Femenino y Masculino.
- EDAD: Variable de naturaleza cuantitativa discreta, entendida como la cantidad de años cumplidos al momento de postular a la UNALM.
- TIPO DOCUMENTO: Variable de naturaleza cualitativa nominal, cuyas categorías son DNI (Documento Nacional de Identidad), CE (Carnet de Extranjería) y PAS (Pasaporte)
- PAIS NAC.: Variable de naturaleza cualitativa nominal, que se refiere al país donde nació el / la postulante.
- DIST. NAC.: Variable de naturaleza cualitativa nominal, que se refiere al distrito donde nació el / la postulante.
- PROV. NAC.: Variable de naturaleza cualitativa nominal, que se refiere a la provincia donde nació el / la postulante.
- DEP. NAC.: Variable de naturaleza cualitativa nominal, que se refiere al departamento donde nació el / la postulante.
- TIPO INSTITUCIÓN: Variable de naturaleza cualitativa nominal, que se refiere a la institución de la cual egresó el/la postulante, pudiendo ser Colegio o Universidad.
- GESTIÓN: Variable de naturaleza cualitativa nominal, que se refiere al tipo de Gestión de la institución de egreso. Compuesto por 4 categorías: Privado (aunque en los datos también aparece Privada, se refiere alo mismo), Pública de gestión directa, Pública de gestión privada, Público
- GESTIÓN2: Variable de naturaleza cualitativa nominal, que se refiere al tipo de Gestión de la institución de egreso. En función a la variable anterior, solo considera 2 categorías: Privada o Pública.
- DIST. DOM.: Variable de naturaleza cualitativa nominal, que se refiere al distrito de domicilio del / de la postulante
- PROV. DOM.: Variable de naturaleza cualitativa nominal, que se refiere a la provincia de domicilio del / de la postulante
- DEP. DOM.: Variable de naturaleza cualitativa nominal, que se refiere al departamento de domicilio del / de la postulante

- MODALIDAD: Variable de naturaleza cualitativa nominal, que se refiere a la modalidad bajo la cual postula. Compuesta por más de 10 categorías, siendo la más común Concurso Ordinario.
- OPCIÓN 1: Variable de naturaleza cualitativa nominal, que se refiere a la carrera que el/la postulante elige como primera opción. Compuesta por 12 categorías (las 12 carreras de la UNALM)
- OPCIÓN 2: Variable de naturaleza cualitativa nominal, que se refiere a la carrera que el/la postulante elige como segunda opción, no siendo ésta obligatoria. Compuesta por 12 categorías (las 12 carreras de la UNALM)
- OPCIÓN 3: Variable de naturaleza cualitativa nominal, que se refiere a la carrera que el/la postulante elige como tercera opción, no siendo ésta obligatoria. Compuesta por 12 categorías (las 12 carreras de la UNALM)
- OPCIONES: Variable de naturaleza cuantitativa discreta, que se refiere a la cantidad de carreras elegidas, pudiendo ser 1, 2 o 3.
- INGRESO: Variable de naturaleza cualitativa nominal, que indica si el postulante SÍ ingresó o NO ingresó a la UNALM.
- CARRERA INGRESO: Variable de naturaleza cualitativa nominal, que indica el nombre de la carrera a la cual ingresó el/la postulante, caso contrario coloca "NO INGRESÓ", por lo que acaba estando compuesta por 13 categorías.
- ORDEN MERITO: Variable de naturaleza cualitativa nominal, que se refiere a la posición ocupada por cada postulante una vez que el puntaje final se ha ordenado de mayor a menos.
- PUNTAJE FINAL: Variable de naturaleza cuantitativa continua, que se refiere a la nota final obtenida en el examen de admisión 2018-I.
- PUNTAJE RM: Variable de naturaleza cuantitativa continua, que se refiere a la nota final obtenida en el curso de Razonamiento Matemático en el examen de admisión 2018-I.
- PUNTAJE RV: Variable de naturaleza cuantitativa continua, que se refiere a la nota final obtenida en el curso de Razonamiento Verbal en el examen de admisión 2018-I.
- PUNTAJE MATEMÁTICAS: Variable de naturaleza cuantitativa continua, que se refiere a la nota final obtenida en el curso de Matemática en el examen de admisión 2018-I.
- PUNTAJE FÍSICA: Variable de naturaleza cuantitativa continua, que se refiere a la nota final obtenida en el curso de Fisica en el examen de admisión 2018-I.
- PUNTAJE QUÍMICA: Variable de naturaleza cuantitativa continua, que se refiere a la nota final obtenida en el curso de Química en el examen de admisión 2018-I.
- PUNTAJE BIOLOGÍA: Variable de naturaleza cuantitativa continua, que se refiere a la nota final obtenida en el curso de Biología en el examen de admisión 2018-I.
- AÑO EGRESO COLEGIO: Variable de naturaleza cuantitativa discreta, que indica el año en el que el / la postulante egresó del colegio, si es que aplica (no aplica para las modalidades Egresado de cuarto año o cursando quinto año de Educación Secundaria, Traslados Externos Universidades, Graduados o Titulados, y Participante Libre).

A continuación, se hará lectura de los datos:

```
library(readxl)
datos = read_xlsx("Postulantes.xlsx")
```

Medidas de tendencia central

Son indicadores que permiten resumir el conjunto de datos en un único valor representativo intermedio o central. Estas medidas son: media, mediana y moda.

Media

- También conocido como media aritmética o promedio.
- Se obtiene mediante la suma de todos los valores de una variable de naturaleza cuantitativa entre la cantidad de valores sumados.
- Se expresa en las mismas unidades que la variable. Por ejemplo: si se mide los tiempos en segundos, el promedio estará también en segundos. Si se cuenta el número de frutos por planta, el promedio estará medido también en frutos por planta.

Media aritmética en R

Resolvamos algunas interrogantes a partir del caso establecido:

1. ¿Cuál es la edad promedio de todos los postulantes?

Dos maneras de poder resolverlo:

```
datos$EDAD |> mean()
```

```
## [1] 18.36069
```

```
library(dplyr)
datos |> pull(EDAD) |> mean()
```

```
## [1] 18.36069
```

Respuesta: La edad promedio de los postulantes a la UNALM en el proceso 2018-I fue de 18.36 años.

2. ¿Cuál es la edad promedio de los postulantes que sí ingresaron?

```
datos |> filter(INGRESO=="SÍ") |> pull(EDAD) |> mean()
```

```
## [1] 18.80337
```

Respuesta: La edad promedio de los postulantes que sí ingresaron a la UNALM en el proceso 2018-I fue de 18.80 años.

3. ¿Aquellos que ingresaron tienen mayor edad promedio que quienes no ingresaron?

```
datos |> group_by(INGRESO) |> summarize(mean(EDAD))
```

```
## # A tibble: 2 x 2
##   INGRESO `mean(EDAD)`
##   <chr>          <dbl>
## 1 NO              18.3
## 2 SÍ              18.8
```

Respuesta: Sí. La edad promedio de los postulantes que sí ingresaron a la UNALM en el proceso 2018-I supera en medio año a la edad promedio de los que no ingresaron.

4. ¿Qué carrera elegida como opción 1 presenta los postulantes con mayor edad promedio? ¿y en cuál encontramos a los postulantes más jóvenes?

```
datos |> group_by(`OPCIÓN 1`) |> summarize(mean(EDAD))
```

```
## # A tibble: 12 x 2
##    `OPCIÓN 1`    `mean(EDAD)`
##    <chr>                <dbl>
##  1 Agronomía             19.0
##  2 Agrícola              18.9
##  3 Ambiental             18.0
##  4 Biología              18.1
##  5 Economía              18.5
##  6 Estadística           19.1
##  7 Forestal              18.4
##  8 Gestión               18.3
##  9 Industrias            18.5
## 10 Meteorología          18.7
## 11 Pesquería             19.5
## 12 Zootecnia             18.4
```

Respuesta: La carrera de Ingeniería Ambiental tiene los postulantes con menor edad promedio (18.0), mientras que Pesquería a los mayores (19.5 años en promedio).

Mediana

- Indica el valor central para una variable ordenada de menor a mayor.
- Se obtiene mediante el ordenamiento de los datos de una variable de menor a mayor y buscando el valor central. En caso haya dos valores centrales (si el tamaño del conjunto de datos es par) entonces se toma la semisuma de dichos valores.
- Se expresa en las mismas unidades que la variable. Por ejemplo: si se mide los tiempos en segundos, la mediana estará también en segundos. Si se cuenta el número de frutos por planta, la mediana estará medido también en frutos por planta.

Mediana en R

1. ¿Cuál es la mediana del puntaje final? ¿Cómo se interpreta?

```
datos$`PUNTAJE FINAL` |> median()
```

```
## [1] 8
```

```
datos |> pull(`PUNTAJE FINAL`) |> median()
```

```
## [1] 8
```

Respuesta: La mediana es 8 puntos. Esto significa que al menos la mitad (50% de los postulantes obtuvo 8 o menos de nota final)

2. ¿Cuál es la mediana del puntaje obtenido por los postulantes nacidos en Perú? ¿Cómo se interpreta?

```
datos |> filter(`PAIS NAC.` == "Perú") |> pull(`PUNTAJE FINAL`) |> median()
```

```
## [1] 8
```

Respuesta: La mediana es 8 puntos. Esto significa que al menos la mitad (50% de los postulantes nacidos en Perú obtuvo 8 o menos de nota final)

3. ¿Los postulantes que provienen de colegio tienen menor puntaje promedio que aquellos que vienen de universidades?

```
datos |> group_by(`TIPO INSTITUCIÓN`) |> summarize(median(`PUNTAJE FINAL`))
```

```
## # A tibble: 2 x 2
##   `TIPO INSTITUCIÓN` `median(\`PUNTAJE FINAL\`)`
##   <chr>                                   <dbl>
## 1 Colegio                                  7.96
## 2 Universidad                             10.4
```

Respuesta: Sí, mientras que al menos el 50% de los postulantes que solo acabaron el colegio tiene 7.96 o menos de nota, mientras que este valor es de 10.4 puntos para quienes vienen de universidad.

Moda

- Indica el valor que más se repite en una variable. Se puede obtener para variables de tipo cualitativo o cuantitativo.
- Se obtiene contando la frecuencia de aparición de cada valor o categoría y seleccionando aquel más frecuente. Podría haber empate de modo que exista más de una moda, o podría no haber un valor que se repite más veces que los demás y en ese caso no habría moda. De ser alguno de esos los casos, reportarlo junto a la interpretación.
- Se expresa en las mismas unidades que la variable. Por ejemplo: si se mide los tiempos en segundos, la moda estará también en segundos. Si se cuenta el número de frutos por planta, la moda estará medido también en frutos por planta.

Moda en R

1. ¿Cuál es la carrera que más postulantes han elegido como primera opción? ¿Y como segunda? ¿Y como tercera

```
library(modeest)
datos |> pull(`OPCIÓN 1`) |> mfv()
```

```
## [1] "Ambiental"
```

```
datos |> pull(`OPCIÓN 2`) |> mfv(na_rm = TRUE)
```

```
## [1] "Agronomía"
```

```
datos |> pull(`OPCIÓN 3`) |> mfv(na_rm = TRUE)
```

```
## [1] "Agronomía"
```

```
# na_rm = TRUE evita los "valores perdidos" cuando los postulantes no marcaron 2da y/o 3ra opción.
```

Respuesta: La carrera más elegida en 1ra opción es Ingeniería Ambiental. Tanto 2da como 3ra opción se elige Agronomía.

2. ¿La carrera más elegida en primera opción es la misma para hombres y mujeres?

```
datos |> group_by(SEXO) |> summarise(mfv(`OPCIÓN 1`, na_rm=TRUE))
```

```
## # A tibble: 2 x 2
##   SEXO      `mfv(\`OPCIÓN 1\`, na_rm = TRUE)`
##   <chr>     <chr>
## 1 Femenino  Ambiental
## 2 Masculino Ambiental
```

Medidas de posición

- Permite indicar la ubicación relativa de un valor, una vez que los datos han sido ordenados de menor a mayor. Por ejemplo el percentil 90 corresponde al valor tal que existe un 90% menor o igual a él.
- Se puede obtener para variables de naturaleza cuantitativa, y se expresa en las mismas unidades de la variable. Por ejemplo, si se recolecta una variable medida en kilómetros, los percentiles estarán en kilómetros. Se tienen los casos particulares:
  - Percentiles: Dividen los datos en 100 partes, de modo que se tiene un percentil 1, 2, 3, . . . , 99.
  - Cuartiles: Dividen los datos en 4 partes. Se tiene el cuartil 1 (o percentil 25), cuartil 2 (o percentil 50 o mediana) y cuartil 3 (o percentil 75).
  - Deciles: Dividen los datos en 10 partes, de modo que se tienen 9 deciles, desde el decil 1 (o percentil 10), decil 2 (o percentil 20), . . . , decil 9 (o percentil 90)

Medidas de posición en R

1. ¿Cuál es el percentil 32 de la Nota de Razonamiento Matemático? ¿Cómo se interpreta?

```
datos |> pull(`PUNTAJE RM`) |> quantile(probs = c(0.32))
```

```
##      32%
## 10.09524
```

Respuesta: El percentil 32 es igual a 10.10, esto significa que al menos el 32$ de los postulantes obtuvo una nota menor o igual a 10.10 en Razonamiento Matemático.

2. ¿Cuáles son los cuantiles de la nota de Razonamiento Verbal? ¿Cómo se interpretan?

```
datos |> pull(`PUNTAJE RV`) |> quantile()
```

```
##       0%      25%      50%      75%     100%
## -0.85714  6.47619  8.47619 10.57143 17.90476
```

Respuesta: El percentil 25 o cuartil 1 es 6.48, esto significa que al menos el 25% de los postulantes obtuvo 6.48 o menos en su nota de Razonamiento Verbal.

El percentil 50 o cuartil 2 es 8.48, esto significa que al menos el 50% de los postulantes obtuvo 8.48 o menos en su nota de Razonamiento Verbal.

El percentil 75 o cuartil 3 es 10.57, esto significa que al menos el 75% de los postulantes obtuvo 10.57 o menos en su nota de Razonamiento Verbal.

3. ¿En qué caso se obtuvo un mayor decil 4 en nota de Química: cuando provienen de institución pública o privada? ¿y decil 9?

```
datos |> filter(!is.na(GESTIÓN2)) |> group_by(GESTIÓN2) |>
  summarise(quantile(`PUNTAJE RV`, probs = c(0.4, 0.9)))
```

```
## Warning: Returning more (or less) than 1 row per `summarise()` group was deprecated in
## dplyr 1.1.0.
## i Please use `reframe()` instead.
## i When switching from `summarise()` to `reframe()`, remember that `reframe()`
##   always returns an ungrouped data frame and adjust accordingly.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## # A tibble: 4 x 2
## # Groups:   GESTIÓN2 [2]
##   GESTIÓN2 `quantile(\`PUNTAJE RV\`, probs = c(0.4, 0.9))`
```

```
##   <chr>                                                <dbl>
## 1 Privada                                               7.90
## 2 Privada                                              12.7
## 3 Pública                                               7.43
## 4 Pública                                              11.8
```

Respuesta: Tanto el decil 4 como el 9 son meores para quienes provienen de una institución pública. 7.43 versus 7.90 para el decil 4, y 11.8 versus 12.7 para el decil 9.

Medidas de dispersión

- Permiten medir la variabilidad o heterogeneidad de un conjunto de datos a través de diversas medidas, tales como el rango, el rango intercuartil, la varianza, la desviación estándar y el coeficiente de variación.

Rango

- El rango permite medir la amplitud de una variable de naturaleza cuantitativa.
- Se obtiene como la diferencia entre el valor máximo y mínimo, por lo que su valor puede estar afectado por valores extremos u *outliers*.
- Se expresa en las mismas unidades que la variable. Por ejemplo: si se mide los tiempos en segundos, el rango estará también en segundos. Si se cuenta el número de frutos por planta, el rango estará medido también en frutos por planta.

Rango en R

1. ¿Cuál es el rango de la edad y cómo se interpreta?

```
library(ggQC)
datos |> pull(EDAD) |> QCrange()
```

```
## [1] 38
```

Respuesta: La amplitud de la variable Edad es de 38 años, es decir esa es la diferencia entre el postulante de mayor y menor edad.

2. ¿En las mujeres el rango es menor que en los hombres?

```
datos |> group_by(SEXO) |> summarize(QCrange(EDAD))
```

```
## # A tibble: 2 x 2
##   SEXO      `QCrange(EDAD)`
##   <chr>               <dbl>
## 1 Femenino               27
## 2 Masculino              37
```

Respuesta: Sí, la amplitud de la edad en mujeres es 10 años menor que en los hombres.

Rango intercuartil

- El rango permite medir la amplitud de una variable de naturaleza cuantitativa en su 50% central.
- Se obtiene como la diferencia entre el cuartil 3 y cuartil 1, por lo que su valor no se ve afectado por valores atípicos, extremos u *outliers*.
- Se expresa en las mismas unidades que la variable. Por ejemplo: si se mide los tiempos en segundos, el rango intercuartil estará también en segundos. Si se cuenta el número de frutos por planta, el rango intercuartil estará medido también en frutos por planta.

Rango intercuartil en R

1. ¿Cuál es el rango intercuartil de la edad y cómo se interpreta?

```
datos |> pull(EDAD) |> IQR()
```

```
## [1] 2
```

Respuesta: La amplitud de la variable Edad en su 50% central es de 2 años. Note que, a diferencia del rango (38 años), esta es una cantidad bastante menor, lo que indica que en los datos centrales las edades son muy parecidas y solo difieren en dos años. Cuando el rango y rango intercuartil difieren mucho, es posible que haya algún valor atípico.

2. ¿Qué tipo de institución presenta la menor variación en el 50% central de las notas de Razonamiento verbal, considerando solo a los ingresantes?

```
datos |> filter(INGRESO == "SÍ") |> group_by(`TIPO INSTITUCIÓN`) |>
  summarize(IQR(`PUNTAJE RV`, na.rm=TRUE)) |> rename(RIC=2) |> arrange(RIC)
```

```
## # A tibble: 2 x 2
##   `TIPO INSTITUCIÓN`   RIC
##   <chr>              <dbl>
## 1 Universidad         2.48
## 2 Colegio             3.62
```

Note que se renombró la segunda columna dándole el nombre RIC y luego se ordenó de menor a mayor con la función arrange, la cual también pertenece a dplyr

Respuesta: Los ingresantes que provienen de Universidad presentaron la menor variación en el 50% central de las notas alcanzadas en razonamiento verbal.

Varianza

- La varianza permite medir qué tanto los datos se desvían de su promedio.
- Se obtiene mediante la siguiente fórmula, para el caso poblacional y muestral, respectivamente:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} \qquad s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

- Se expresa en unidades al cuadrado, lo que complica su interpretación. Por ejemplo: si se mide los tiempos en segundos, la varianza estará en segundos al cuadrado. Si se cuenta el número de frutos por planta, la varianza estará medido en (frutos por planta)^2.

Varianza en R

1. ¿Cuál es la varianza de la nota de Química?

```
datos |> pull(`PUNTAJE QUÍMICA`) |> var(na.rm=TRUE)
```

```
## [1] 30.94235
```

Respuesta: La varianza de la nota de Química es de 30.94 puntos^2

2. ¿Quién presenta mayor varianza de edad: ¿hombres o mujeres?

```
datos |> group_by(SEXO) |> summarize(var(EDAD, na.rm=TRUE))
```

```
## # A tibble: 2 x 2
##   SEXO      `var(EDAD, na.rm = TRUE)`
##   <chr>                         <dbl>
```

```
## 1 Femenino                      3.91
## 2 Masculino                     7.16
```

Respuesta: La varianza en los hombres (7.16 puntos^2) es mayor que en las mujeres (3.91 puntos^2). Sin embargo, para comparar varianzas, se sugiere verificar que las medias sean iguales o muy similares, o en su defecto, utilizar el coeficiente de variación, que se abordará en una subsiguiente sección.

Desviación estándar

- La desviación estándar, al igual que la varianza, permite medir qué tanto los datos se desvían de su promedio.
- Se obtiene extrayendo la raíz cuadrada a la varianza, es decir $\sigma = \sqrt{\sigma^2}$ y $s = \sqrt{s^2}$ para parámetro y estimador, respectivamente.
- Se expresa en las mismas unidades que la variable, lo que resuelve la complicación de interpretación de la varianza. Por ejemplo: si se mide los tiempos en segundos, la desviación estándar también estará en segundos. Si se cuenta el número de frutos por planta, la desviación estándar estará medida en frutos por planta.

Desviación estándar en R

1. ¿Cuál es la desviación estándar de la edad de os postulantes?

```
datos |> pull(EDAD) |> sd()
```

```
## [1] 2.333832
```

Respuesta: La desviación estándar es 2.33 años, es decir los datos de edad se desvían de su media en 2.33 años en promedio.

2. ¿Cuál es la desviación estándar de la edad de os postulantes que ingresaron y que no ingresaron?

```
datos |> group_by(INGRESO) |> summarise(sd(EDAD))
```

```
## # A tibble: 2 x 2
##   INGRESO `sd(EDAD)`
##   <chr>        <dbl>
## 1 NO            2.32
## 2 SÍ            2.40
```

Para comparar las desviaciones estándar (así como las varianzas) es conveniente que las medias sean similares o iguales, veamos que sí son similares:

```
datos |> group_by(INGRESO) |> summarise(mean(EDAD))
```

```
## # A tibble: 2 x 2
##   INGRESO `mean(EDAD)`
##   <chr>          <dbl>
## 1 NO              18.3
## 2 SÍ              18.8
```

Respuesta: Entonces, vemos que aquellos que ingresaron solo superan en medio año de edad a quienes no ingresaron. Asimismo, muestran una desviación estándar ligeramente menor (2.40-2.32=0.08 años)

3. ¿Cuál es la desviación estándar de la nota final de los ingresantes a la carrera de Ingeniería Ambiental cuyo orden de mérito se encuentra entre el 1 y el 100?

```
datos |>
  filter(INGRESO == "SÍ" & `CARRERA INGRESO` == "Ambiental" &
```

```
          `ORDEN MERITO` >= 1 & `ORDEN MERITO` <=100) |>
  summarise(sd(`PUNTAJE FINAL`))
```

```
## # A tibble: 1 x 1
##   `sd(\`PUNTAJE FINAL\`)`
##                     <dbl>
## 1                   0.751
```

Coeficiente de variabilidad

- También llamado coeficiente de variación. Es una medida de dispersión relativa, es decir permite realizar comparaciones en cuanto a la variabilidad de un conjunto de datos. Estas comparaciones son útiles cuando se tienen dos variables con unidades distintas o cuando las medias son distintas.
- Por ejemplo si las variables son distintas, para comparar la variabilidad de la altura (en metros) y el peso (en kg), no podemos usar la desviación estándar pues ésta se expresa en las unidades, y no se pueden comparar metros y kg.
- Un segundo ejemplo sería comparar la variabilidad en el salario de un recién egresado y uno que egresó hace 15 años. El hecho de que los promedios sean bastante diferentes, hará naturalmente que la variabilidad en los egresados hace 15 años sea mayor que en los recién egresados, cuyo sueldo debe ser más similar.
- Se obtiene mediante la fórmula:

$$CV = \frac{\sigma}{\mu} \times 100\% \qquad cv = \frac{s}{\overline{x}} \times 100\%$$

- Es adimensional, es decir se expresa en %, sin importar la variable cuantitativa que se esté analizando. Por ejemplo: si se mide los tiempos en segundos o si se cuenta el número de frutos por planta, el cv estará medido en % en ambos casos.

Coeficiente de variabilidad en R

1. ¿Cuál es el coeficiente de variación de la nota de matemáticas?

```
library(sjstats)
datos |> pull(`PUNTAJE MATEMÁTICAS`) |> cv()
```

```
## [1] 0.872405
```

Respuesta: El coeficiente de variación de la nota de matemáticas es 87.24%

2. ¿Cuáles son las dos carreras con mayor variabilidad en las edades de sus ingresantes?

```
datos |> filter(INGRESO=="SÍ") |> group_by(`CARRERA INGRESO`) |>
  summarise(cv(EDAD)) |> rename(cv=2) |> arrange(-cv)
```

```
## # A tibble: 12 x 2
##    `CARRERA INGRESO`     cv
##    <chr>              <dbl>
##  1 Gestión            0.203
##  2 Estadística        0.169
##  3 Ambiental          0.151
##  4 Agronomía          0.141
##  5 Economía           0.131
##  6 Meteorología       0.127
##  7 Pesquería          0.126
##  8 Forestal           0.114
```

```
##  9 Zootecnia         0.0879
## 10 Agrícola          0.0774
## 11 Industrias        0.0744
## 12 Biología          0.0699
```

Respuesta: Las dos carreras con mayor variabilidad son Ingeniería en Gestión Empresarial y Estadística Informática, con 20.3% y 16.9%, respectivamente. Son las carreras cuyos ingresantes tuvieron mayor dispersión de edades.

3. ¿Qué variable tiene mayor dispersión: el número de opciones marcadas, la edad o la nota final?

```
datos |> pull(OPCIONES) |> cv();datos |> pull(EDAD) |> cv(); datos |> pull(`PUNTAJE FINAL`) |> cv()
```

```
## [1] 0.2144668
```

```
## [1] 0.1271102
```

```
## [1] 0.3932047
```

La variable con menor dispersión fue la Edad (cv=12.71%), seguida de la cantidad de opciones (cv=21.31%) y finalmente la nota final (cv=39.32%).

Tablas de frecuencia

Tablas de frecuencia para variables cualitativas

1. ¿Cuántos postulantes ingresaron y qué proporción o porcentaje representan?

```
datos |>
  count(INGRESO) |>
  mutate(fr = n/sum(n))
```

```
## # A tibble: 2 x 3
##   INGRESO     n    fr
##   <chr>   <int> <dbl>
## 1 NO       2824 0.888
## 2 SÍ        356 0.112
```

Respuesta: Ingresaron 356 postulantes por Examen de Admisión, en sus distintas modalidades, lo cual representa el 11.2%.

2. ¿Cuántos postulantes tuvo cada carrera en primera opción y qué porcentaje representa la carrera menos elegida?

```
datos |>
  count(`OPCIÓN 1`) |>
  mutate(fr = n/sum(n)) |>
  arrange(-n)
```

```
## # A tibble: 12 x 3
##    `OPCIÓN 1`       n     fr
##    <chr>        <int>  <dbl>
##  1 Ambiental     1047 0.329
##  2 Agronomía      426 0.134
##  3 Gestión        363 0.114
##  4 Industrias     348 0.109
##  5 Biología       234 0.0736
##  6 Forestal       225 0.0708
```

```
##  7 Zootecnia      164 0.0516
##  8 Economía       146 0.0459
##  9 Agrícola       105 0.0330
## 10 Meteorología    48 0.0151
## 11 Estadística     42 0.0132
## 12 Pesquería       32 0.0101
```

Respuesta: En la tabla se aprecia la cantidad de postulantes en primera opción por cada carrera. Asimismo, la carrera menos demandada fue Pesquería, con solo 32 postulantes, lo que representa 1.01% del total.

3. ¿Cuántos ingresantes hubo por cada carrera? ¿Qué porcentaje representa la carrera con más estudiantes de primer ciclo?

```
datos |>
  filter(INGRESO == "SÍ") |>
  count(`CARRERA INGRESO`) |>
  mutate(fr = n/sum(n)) |>
  arrange(-n)
```

```
## # A tibble: 12 x 3
##    `CARRERA INGRESO`     n     fr
##    <chr>             <int>  <dbl>
##  1 Agronomía            67 0.188
##  2 Zootecnia            42 0.118
##  3 Industrias           30 0.0843
##  4 Agrícola             29 0.0815
##  5 Ambiental            28 0.0787
##  6 Pesquería            28 0.0787
##  7 Biología             27 0.0758
##  8 Forestal             23 0.0646
##  9 Economía             22 0.0618
## 10 Gestión              22 0.0618
## 11 Estadística          19 0.0534
## 12 Meteorología         19 0.0534
```

Respuesta: En la tabla se aprecia la cantidad de ingresantes en cada carrera. Asimismo, la carrera con más ingresantes fue Agronomía con 67 nuevos estudiantes, lo que representó el 18.8% del total de ingresantes.

Tablas de frecuencia para variables cuantitativas discretas

1. ¿Cuántos postulantes marcaron 1, 2 y 3 opciones?

```
datos |>
  count(OPCIONES) |>
  mutate(fr = n/sum(n))
```

```
## # A tibble: 3 x 3
##   OPCIONES     n     fr
##      <dbl> <int>  <dbl>
## 1        1   231 0.0726
## 2        2   398 0.125
## 3        3  2551 0.802
```

Respuesta: 231, 398 y 2551 postulantes eligieron 1, 2 y 3 opciones, respectivamente. El 80%, o 4 de cada 5, marcó 3 opciones.

2. ¿Es factible obtener una tabla de frecuencias para la Edad de los postulantes?

```
datos |>
  count(EDAD) |>
  mutate(fr = n/sum(n)) |>
  print(n = 23)
```

```
## # A tibble: 23 x 3
##     EDAD     n       fr
##    <dbl> <int>    <dbl>
##  1    14     1 0.000314
##  2    15    35 0.0110
##  3    16   359 0.113
##  4    17   856 0.269
##  5    18   845 0.266
##  6    19   489 0.154
##  7    20   232 0.0730
##  8    21   114 0.0358
##  9    22    83 0.0261
## 10    23    68 0.0214
## 11    24    31 0.00975
## 12    25    23 0.00723
## 13    26    12 0.00377
## 14    27    11 0.00346
## 15    28     4 0.00126
## 16    29     3 0.000943
## 17    31     3 0.000943
## 18    32     2 0.000629
## 19    33     5 0.00157
## 20    37     1 0.000314
## 21    41     1 0.000314
## 22    48     1 0.000314
## 23    52     1 0.000314
```

Respuesta: Como se ve en la tabla previa, su tamaño es muy grande, ya que la Edad toma muchos valores, en ese caso se prefiere utilizar una tabla para variable cuantitativa continua.

Tablas de frecuencia para variables cuantitativas continuas

Al construir tablas de frecuencia para variables cuantitativas continuas, los valores deben agruparse en intervalos ya que son teóricamente infinitos. Por ejemplo, existen infinitas posibles notas entre 0 y 20. Así, al dividir en intervalos, lo que se contará (frecuencia absoluta) es la cantidad de observaciones que caen dentro de cada intervalo; a partir de esta frecuencia absoluta (`f`) se calculan la frecuencia relativa (`rf`), la frecuencia porcentual (`rf(%)`), la frecuencia acumulada (`cf`) y la frecuencia porcentual acumulada (`cf(%)`)

Se mostrará a continuación el uso de la función `fdt` del paquete `fdth` considerando que cada intervalo es de una longitud igual o similar (uso de la regla de Sturges). Note que al final siempre debe indicarle que muestre la tabla con la función `print`.

```
library(fdth)
datos |> pull(`PUNTAJE FINAL`) |>
  fdt(breaks = "Sturges") |> print()
```

```
##      Class limits   f   rf rf(%)   cf  cf(%)
##  [0.93555,2.2667)  34 0.01  1.07   34   1.07
##   [2.2667,3.5979) 187 0.06  5.88  221   6.95
##    [3.5979,4.929) 344 0.11 10.82  565  17.77
##    [4.929,6.2602) 453 0.14 14.25 1018  32.01
```

```
##   [6.2602,7.5913) 436 0.14 13.71 1454  45.72
##   [7.5913,8.9225) 422 0.13 13.27 1876  58.99
##   [8.9225,10.254) 408 0.13 12.83 2284  71.82
##   [10.254,11.585) 354 0.11 11.13 2638  82.96
##   [11.585,12.916) 286 0.09  8.99 2924  91.95
##   [12.916,14.247) 162 0.05  5.09 3086  97.04
##   [14.247,15.578)  70 0.02  2.20 3156  99.25
##   [15.578,16.909)  20 0.01  0.63 3176  99.87
##   [16.909,18.241)   4 0.00  0.13 3180 100.00
```

Se puede indicar lo siguiente:

- 34 postulantes alcanzaron una nota mayor o igual a 0.9355 pero menor a 2.2667.
- Una proporción igual a 0.05 de estudiantes obtuvo una nota como mínimo de 12.916 puntos pero inferior a 14.247
- El 8.99% de postulantes registró una nota por lo menos de 11.585 puntos pero menor a 12.916
- 1876 postulantes alcanzaron una nota mayor o igual a 0.9355 pero menor a 8.9225.
- El 58.99% de postulantes obtuvo una nota mayor o igual a 0.9355 pero menor a 8.9225.
- Note que la última frecuencia acumulada es igual a la cantidad de datos, y la última frecuencia acumulada porcentual es igual a 100%

También es posible establecer los intervalos de manera arbitraria, por ejemplo de 0 a 20, de 2 en 2:

```
datos |> pull(`PUNTAJE FINAL`) |>
  fdt(start = 0, end = 20, h = 2) |> print()
```

```
##  Class limits   f   rf rf(%)   cf  cf(%)
##         [0,2)  20 0.01  0.63   20   0.63
##         [2,4) 297 0.09  9.34  317   9.97
##         [4,6) 598 0.19 18.81  915  28.77
##         [6,8) 670 0.21 21.07 1585  49.84
##        [8,10) 632 0.20 19.87 2217  69.72
##       [10,12) 523 0.16 16.45 2740  86.16
##       [12,14) 321 0.10 10.09 3061  96.26
##       [14,16) 106 0.03  3.33 3167  99.59
##       [16,18)  12 0.00  0.38 3179  99.97
##       [18,20)   1 0.00  0.03 3180 100.00
```

Se puede indicar lo siguiente:

- 670 postulantes alcanzaron una nota mayor o igual a 6 pero menor a 8
- Una proporción igual a 0.20 de estudiantes obtuvo una nota como mínimo de 8 puntos pero inferior a 10
- El 10.09% de postulantes registró una nota por lo menos de 12 puntos pero menor a 14
- 2740 postulantes alcanzaron una nota mayor o igual a 0 pero menor a 12
- El 49.84% de postulantes obtuvo una nota mayor o igual a 0 pero menor a 8

Note que por defecto los intervalos siempre son abierto a la derecha y cerrado a la izquierda. Sin embargo, es posible modificar ello indicando el argumento `right=TRUE`:

```
datos |> pull(`PUNTAJE FINAL`) |>
  fdt(start = 0, end = 20, h = 2, right = TRUE) |> print()
```

```
##  Class limits   f   rf rf(%)   cf  cf(%)
##         (0,2]  20 0.01  0.63   20   0.63
```

```
##          (2,4] 300 0.09  9.43  320  10.06
##          (4,6] 602 0.19 18.93  922  28.99
##          (6,8] 669 0.21 21.04 1591  50.03
##         (8,10] 632 0.20 19.87 2223  69.91
##        (10,12] 519 0.16 16.32 2742  86.23
##        (12,14] 321 0.10 10.09 3063  96.32
##        (14,16] 104 0.03  3.27 3167  99.59
##        (16,18]  12 0.00  0.38 3179  99.97
##        (18,20]   1 0.00  0.03 3180 100.00
```

Se debe tener cuidado con las interpretaciones que se desprenden de este caso, por el hecho de haber cambiado el tipo de intervalo:

- 669 postulantes alcanzaron una nota mayor a 6 pero menor o igual a 8
- Una proporción igual a 0.20 de estudiantes obtuvo una nota mayor a 8 puntos pero como máximo de 10
- El 10.09% de postulantes registró una nota mayor a 12 puntos pero menor o igual a 14
- 2742 postulantes alcanzaron una nota mayor a 0 pero menor o igual a 12
- El 50.03% de postulantes obtuvo una nota mayor a 0 pero menor o igual a 8

Gráficos

Los gráficos que veremos son representaciones equivalentes a las tablas de frecuencia. Sin embargo, existen muchas más representaciones gráficas que irán conociendo en los próximos ciclos, y con un acabado mucho más elegante (usando el paquete `ggplot2`).

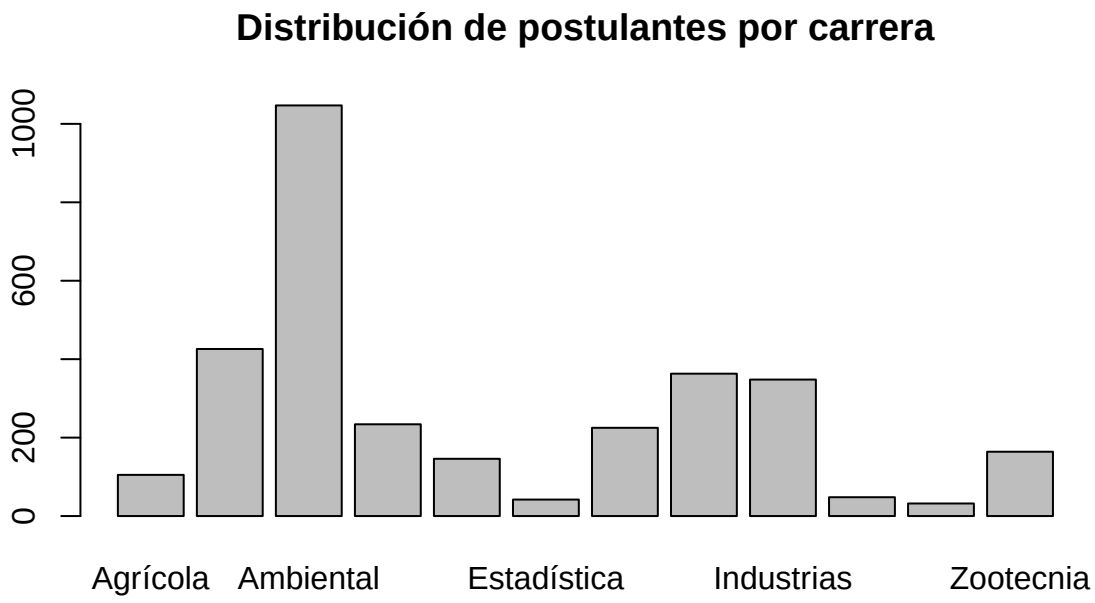
Gráficos para variables cualitativas

Para variables cualitativas se puede utilizar el gráfico de barras, el gráfico circular y el de wafle. No son los únicos, pero sí los más sencillos.

Gráfico de barras

Por ejemplo, una representación gráfica de la cantidad de postulantes, en primera opción, por cada carrera, donde destaca Ingeniería Ambiental:

```
datos |> pull(`OPCIÓN 1`) |> table() |>
  barplot(main = "Distribución de postulantes por carrera")
```



Un segundo ejemplo, para la cantidad de ingresantes por cada carrera:

```
datos |> filter(INGRESO=="SÍ") |>
  pull(`CARRERA INGRESO`) |> table() |>
  barplot(main = "Distribución de ingresantes por carrera")
```

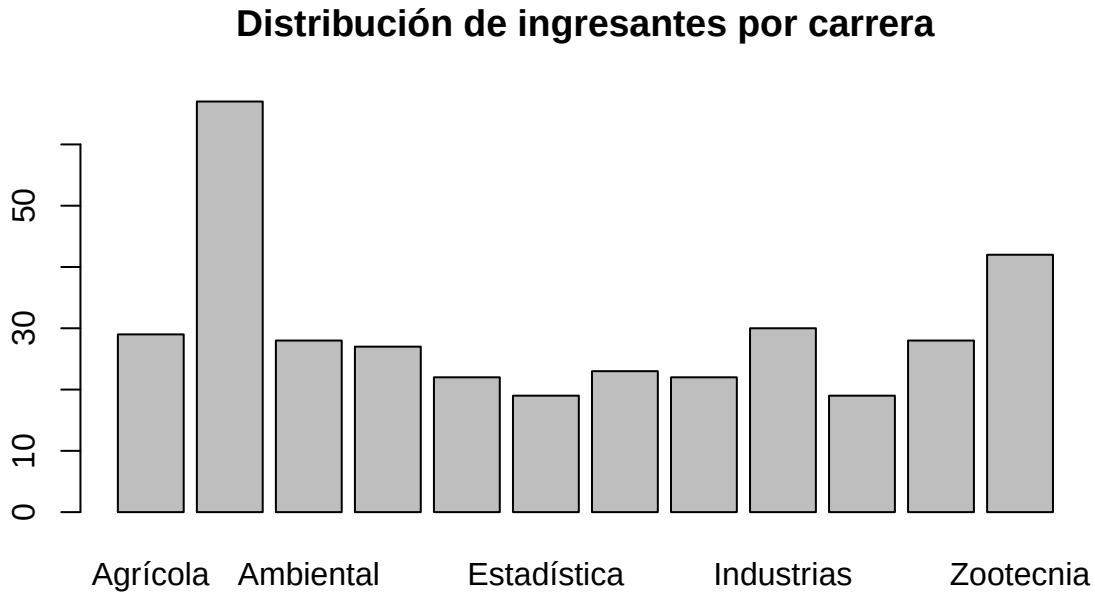
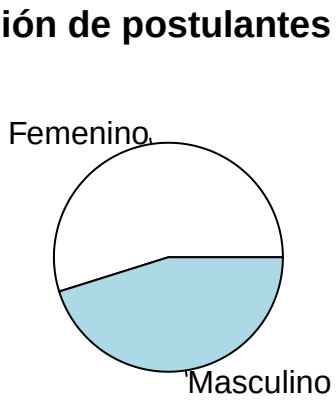


Gráfico circular

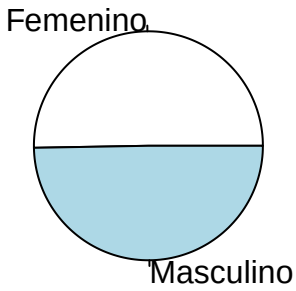
Para el caso del gráfico de pye o de torta:

```
datos |> pull(SEXO) |> table() |>
  pie(main = "Distribución de postulantes por sexo")
```



Se aprecia que existen más postulantes mujeres que varones.

```
datos |> filter(INGRESO=="SÍ") |>  pull(SEXO) |> table() |>
  pie(main = "Distribución de ingresantes por sexo")
```



Sin embargo, cuando se comparan para postulantes, la proporción de hombres y mujeres es casi la misma.

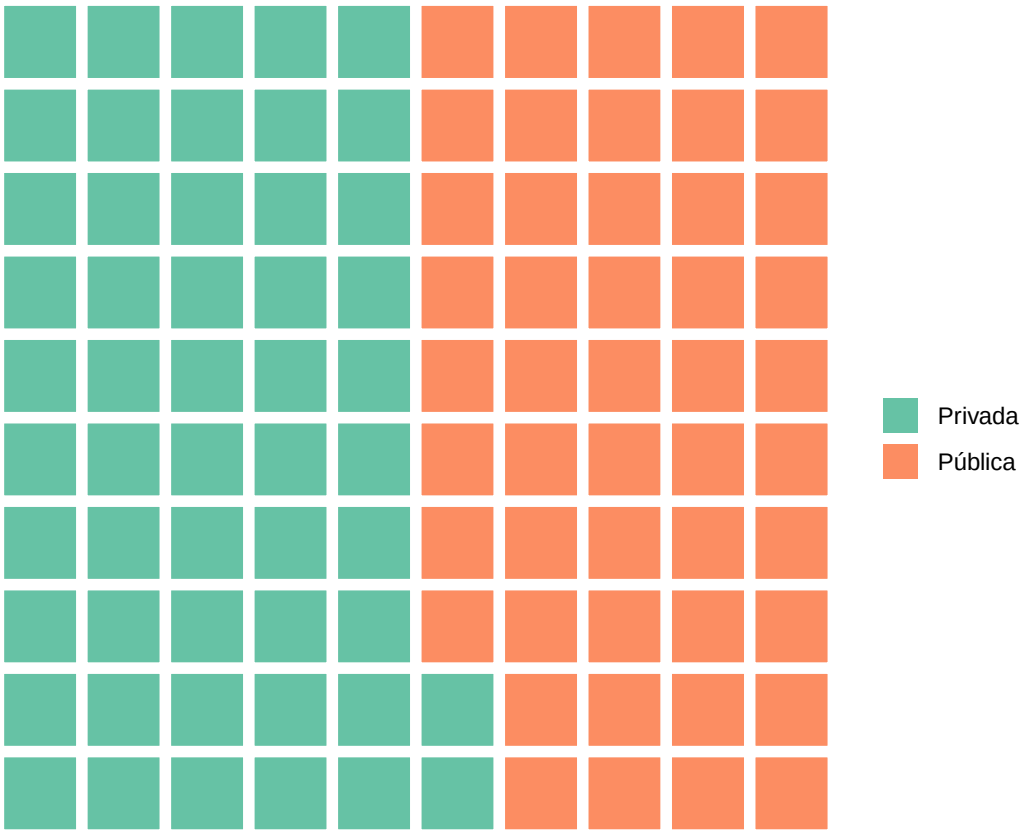
Gráfico de waffle

```
library(waffle)

datos |>
  count(GESTIÓN2) |>
  mutate(PORC = round(100*n/sum(n))) -> tabla

tabla |> pull(PORC,GESTIÓN2) -> partes

waffle(partes, rows = 10)
```

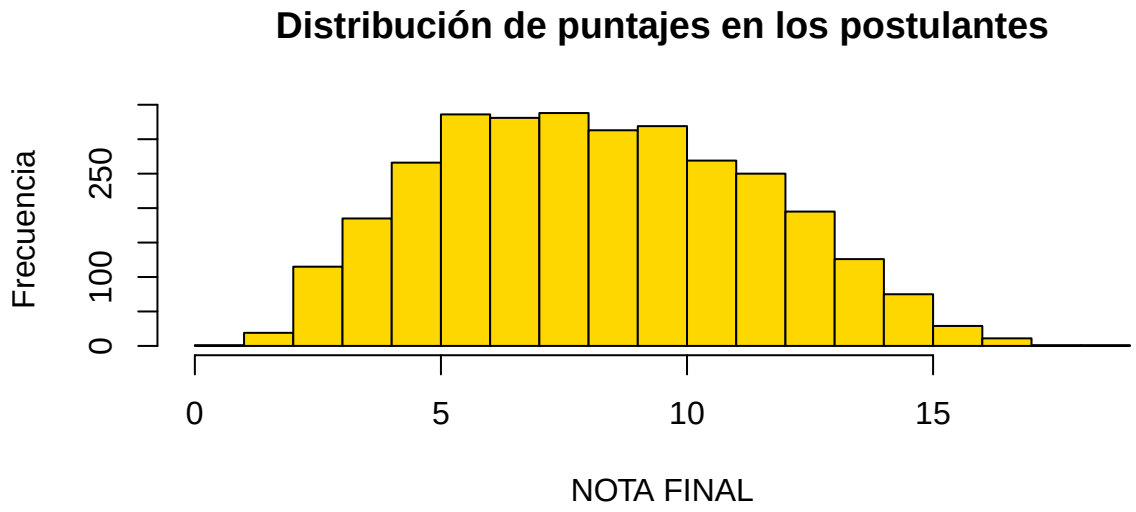


Gráficas para variables cuantitativas

Cuando se tiene una variable cuantitativa, se tiena una diversidad de gráficas.

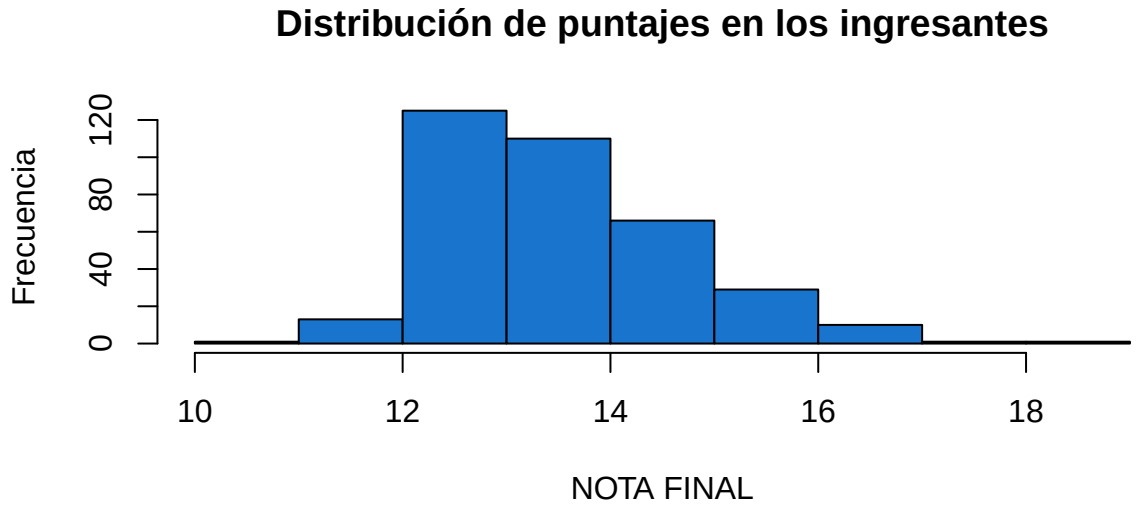
Histograma

```
datos |> pull(`PUNTAJE FINAL`) |>
  hist(col = "gold", xlab = "NOTA FINAL",
       ylab = "Frecuencia",main = "Distribución de puntajes en los postulantes")
```



Esta gráfica permite ver las frecuencias en intervalos, y así la forma de la distribución de los datos. En este caso se aprecia casi una simetría alrededor de los 8 puntos.

```
datos |> filter(INGRESO=="SÍ") |> pull(`PUNTAJE FINAL`) |>
  hist(col = "dodgerblue3", xlab = "NOTA FINAL",
       ylab = "Frecuencia", main = "Distribución de puntajes en los ingresantes")
```



Sin embargo, si se considera a solo los ingresantes, hay una asimetría hacia la derecha bastante marcada, con una gran cantidad de notas concentradas entre 12 y 14.

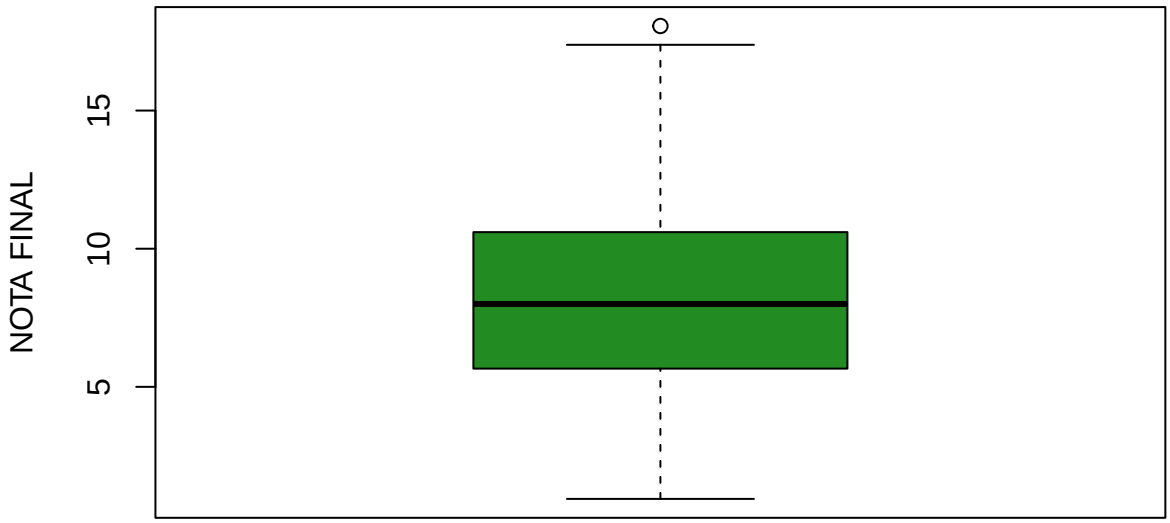
Más colores in R: http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf

Boxplot o diagrama de cajas

El diagrama de cajas permite mostrar el percentil 25 o cuartil 1 en la base de la caja y el cuartil 3 en su "techo", mientras que la línea al medio corresponde al cuartil 2 o mediana.

Muestra además el o los posibles valores atípicos con círculo(s) fuera de los límites establecidos por los bigotes (líneas verticales).

```
datos |> pull(`PUNTAJE FINAL`) |>
  boxplot(col = "forestgreen", ylab = "NOTA FINAL")
```



Se puede verificar viendo lo siguiente:

```
datos |> pull(`PUNTAJE FINAL`) |> quantile()
```

```
##     0%    25%    50%    75%   100%
##  0.945  5.660  8.000 10.600 18.060
```

Ojiva

Permite ver las frecuencias acumuladas:

```
datos |>
  pull(`PUNTAJE FINAL`) |>
  fdt(start=0,end=20,h=2,right=FALSE) |>
  plot(type = "cfp",
       col  = "darkblue",
       pch  = 18,
       xlab = "Nota final",
       ylab = "Frecuencia acumulada",
       main = "Distribución acumulada de las notas finales")
```