Actividad Formativa 04

Análisis de datos de cuestionario (Bivariado)

Dany Lopez (dxlopez@ul.cl) - Ximena Catalán (xrcatala@uc.cl)

Contenido

1 Objetivo

Este material tiene como propósito proporcionar una guía para desarrollar análisis bivariado de ítems de cuestionarios utilizados en Educación Superior. Extenderemos los análisis realizados en la actividad práctica anterior. Para ello, se trabajará nuevamente con el cuestionario que utilizó para transcribir las respuestas de estudiantes ver aqui. Para realizar análisis bivariado, usaremos los ocho ítems tipo Likert del módulo *Aprendizaje* (AP01–AP08) y trataremos de relacionarlos con los ítems de caracterización.

Esta guía está pensada para que la pueda desarrollar en clases y también como material de estudio complementario fuera del horario oficial de clases.

> **!** Importante
>
> Esta guía está preparada para abordar los análisis utilizando EXCEL en conjunto con BlueSky (opción 1), y también utilizando el lenguaje R (opción 2). Sugerimos que aquellas personas que están comenzando en esta área que vean los ejemplos de la opción 1, y aquellas personas que ya cuentan con experiencia en análisis en R, que opten por la opción 2.

2 Introducción

2.1 Análisis cuantítativo de ítems

El análisis cuantitativo de ítems con distinta tipología requiere determinar procedimientos acordes al nivel de medición y al objetivo de la indagación. En un primer nivel, tenemos el análisis univariado, que describe cada ítem por separado de manera tal de caracterizar su distribución. En ítems dicotómicos y nominales se reportan frecuencias y porcentajes. En ítems ordinales tipo Likert se examinan además de la distribución por categorías, promedios, mediana y/o rangos (aunque para la estimación de promedios y mediana se requiere una justificación acorde). En items numéricos se resumen tendencia central y dispersión (promedio, desviación estándar, rango, curtosis, entre otros).

Por su parte, el análisis bivariado estudia relaciones entre dos variables, y los métodos estadísticos para realizarlo varía según la tipología de las variables. Cuando el interés es relacional entre múltiples variables, se consideran modelos que integran varios predictores y controlan covariables. Según el caso, pueden usarse modelos lineales o generalizados (lineal tradicional para respuestas continuas, logísticos para respuestas dicotómicas, multinomiales para Likert). También, se emplean análisis factoriales para comprender la estructura subyacente de un un set de ítems, y enfoques específicos para mediciones de atributos latentes (por ejemplo, modelos de respuesta al ítem y/o teória clásica de medición).

A modo de resumen, a continuación se muestra una tabla que resume los tres tipos de análisis de datos descrito anteriormente En la práctica, usualmente se implementan estos tres tipos de análisis o combinaciones de algunas de ellas.

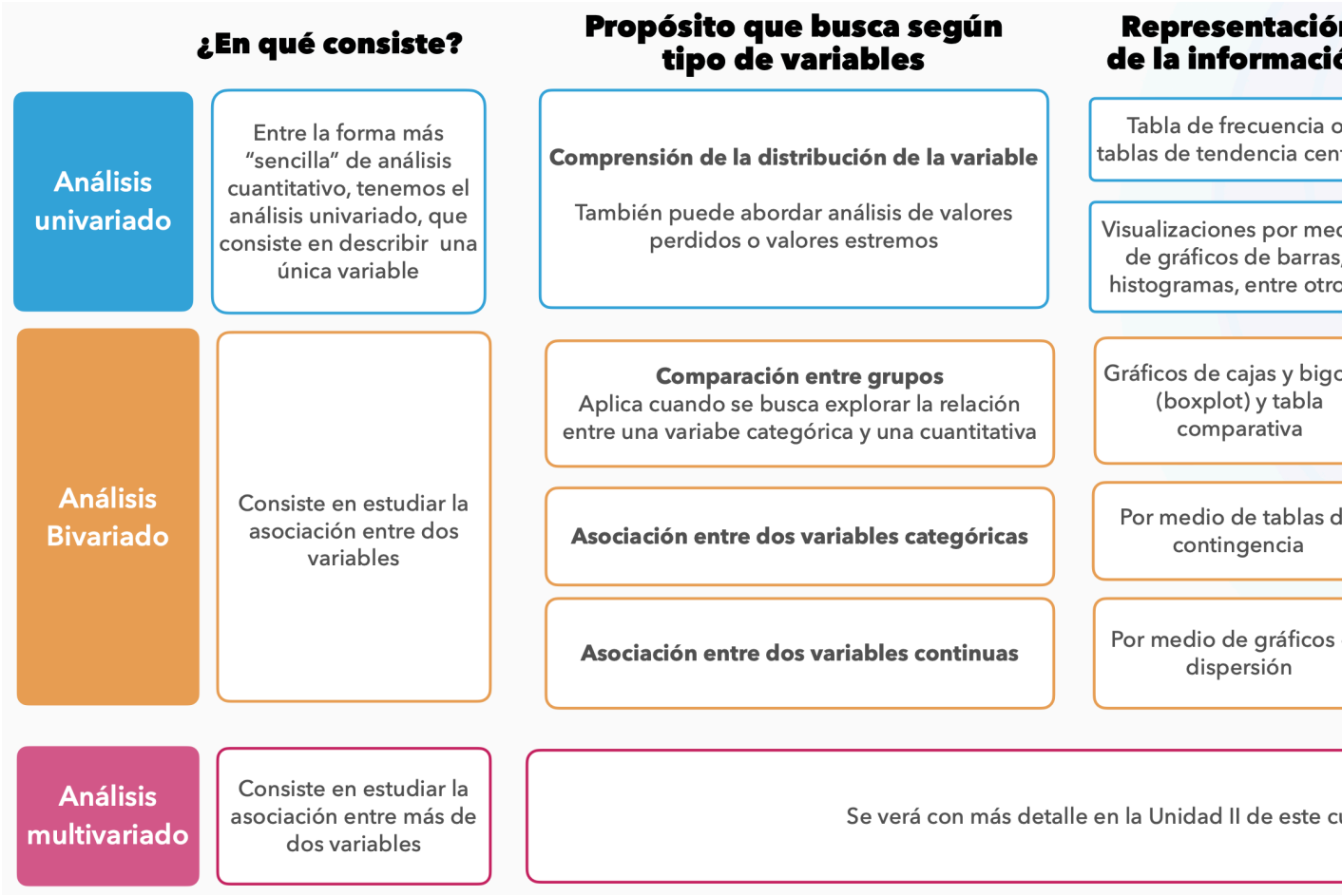

Figure 1: Creación propia.

2.2 Análisis bivariado de ítems

Con el análisis bivariado ya estamos en condiciones de explorar asociació entre variables e incluso, responder preguntar de investigación que aborden como tema central la relación entre dos variables.

2.2.1 Introducción general

El análisis bivariado estudia la relación entre dos variables medidas en una muestra. En cuestionarios, cada variable suele ser un ítem con categorías como "sí/no", escalas tipo Likert, respuestas nominales o continuas. Este análisis permite responder si existe relación entre dos ítems, cuál es su dirección y magnitud, y si la evidencia es consistente con una asociación no debida al azar. En Educación Superior se usa para comprender diversos constructos, por ejemplo: vínculos entre indicadores de docencia, experiencia estudiantil y bienestar, o entre carga de trabajo y estrés académico.

2.2.2 Tipos de preguntas de investigación que aborda

Las preguntas típicas que se responden con análisis bivariado se formulan en términos de asociación y contraste. Una primera clase plantea preguntas so bre si existe relación entre dos ítems y con qué fuerza, por ejemplo entre satisfacción con la enseñanza y percepción de aprendizaje. Una segunda clase contrasta perfiles de respuesta entre dos categorías, por ejemplo si quienes declaran haber asistido a tutorías muestran respuestas de forma distinta respecto de quienes no asistieron. Una tercera clase examina la dirección y el patrón de cambio entre categorías ordenadas, por ejemplo si a medida que aumenta el acuerdo con presentar una mayor carga académica también muestran un aumenta en el nivel de estrés experimentado. En todos los casos, el foco está en una relación entre dos variables y no en múltiples variables a la vez.

2.2.3 Procedimiento recomendado para efectuar análisis bivariado

Paso 1 (análisis univariado): Primero se identifica el tipo y la escala de medición de cada ítem y, en caso de aplicar, se explicita el tratamiento de valores perdidos. Luego se construye la tabla de frecuencias, tablas de tendencia central u otro tipo de representación que permita commprender la distribución de cada ítem por separado.

Paso 2: Se selecciona el formato de presentación y las medidas de asociación según el nivel de medición. En caso de aplicar, también se seleccionan las pruebas estadísticas necesarias. Después se visualizan los resultados para una mayor interpretación de la relación (esto se verá en el módulo de Visualización de ítems más adelante).

Paso 3: Se reporta e interpreta la relación observada, la mgnitud de la asociación (si aplica) y las pruebas estadísticas (si aplica).

3 Actividades prácticas

A continuación se propone una actividad de análisis bivariado en dos entornos complementarios, Excel y R, orientada a describir la distribución entre una variables categóricas con una variable ordinal. El objetivo es producir una tabla clara junto con un breve texto interpretativo como ilustración de lo que tendrá que realizar en su investigación en caso de enfrentarse al mismo desafio. Para realizar estas actividades, usaremos el cuestionario con el que se transcribieron las respuestas de 30 estudiantes. Comenzaremos describiendo algunos de los ítems del módulo de caracterización para luego análizar los ítems del módulo de Aprendizaje. A continuación se describe la base de datos con la que se realizarán los análisis univariados.

3.1 Descripción de los datos

Los datos que usaremos provienen del módulo de Aprendizaje del cuestionario VOCES que se aplicó a 30 estudiantes de primer año de diversas carreras de una universidad privada.

A continuación se muestran los primeros 10 registros de la base de datos que utilizaremos para realizar los análisis univariados.

```
db <- readxl::read_excel('./02_data/01_database.xlsx',
                         sheet='Base_datos',col_names =T, na=c("","  ")
                        )

codebook <- readxl::read_excel("./02_data/01_database.xlsx",
                               sheet='Diccionario',col_names =T, na=c("","  ")
                              )
```

folio	CA01	CA02	CA03	CA04	AP01	AP02	AP03	AP04	AP05	AP06	AP07	AP08
XXX01	F	21	2022	No	5	5	1	5	2	2	2	1
XXX02	F	21	2022	No	NA	5	1	5	2	2	2	4
XXX03	F	21	2022	Si	5	1	5	4	5	3	1	2
XXX04	M	20	2021	Si	2	4	4	1	2	3	2	2
XXX05	No indica	20	2022	No	1	3	1	1	4	4	4	2
XXX06	No indica	20	2022	Si	5	2	2	3	2	4	4	1
XXX07	No indica	19	2022	No	3	2	1	2	1	1	1	3
XXX08	No indica	19	2023	No	2	1	3	2	1	2	5	2
XXX09	No indica	20	2022	No	5	3	2	4	5	4	2	2
XXX010	No indica	21	2021	Si	4	5	2	5	4	1	1	2

También se adjunta el diccionario para comprender el significado de las columnas y las escalas de medición de cada ítem.

Caracterizacion	Folio		FOLIO	identificador		
Caracterizacion	Sexo	Sexo	CA01	categorica		F = 'Femenino' M = 'Masculino' No indica = 'Prefiero no indicar'
Caracterizacion	Edad	Edad	CA02	continua		Edad medida en años
Caracterizacion	Período ingreso carrera	Año ingreso a la carrera	CA03	Entero		Variable medida en años
Caracterizacion	Primera persona estudios universitarios	¿Eres la primera persona en tu familia en estudiar en la Universidad?	CA04	binaria		Si No
Aprendizaje	Aprendizaje Superficial	He tenido problemas para encontrar el sentido a las cosas que tengo que estudiar.	AP01	Acuerdo-Desacuerdo	5	1 = Totalmente en desacuerdo 2 = En desacuerdo 3 = Ni de acuerdo ni en desacuerdo 4 = De acuerdo 5 = Totalmente de acuerdo
Aprendizaje	Aprendizaje Superficial	Muchas de las cosas que he aprendido permanecen en mi mente como ideas sin relación.	AP02	Acuerdo-Desacuerdo	5	1 = Totalmente en desacuerdo 2 = En desacuerdo 3 = Ni de acuerdo ni en desacuerdo 4 = De acuerdo 5 = Totalmente de acuerdo
Aprendizaje	Aprendizaje Superficial	Los temas que estudiamos son presentados de una manera tan complicada que a menudo no puedo entender qué significan.	AP03	Acuerdo-Desacuerdo	5	1 = Totalmente en desacuerdo 2 = En desacuerdo 3 = Ni de acuerdo ni en desacuerdo 4 = De acuerdo 5 = Totalmente de acuerdo
Aprendizaje	Aprendizaje Superficial	Tengo que estudiar una y otra vez cosas que realmente no me hacen mucho sentido.	AP04	Acuerdo-Desacuerdo	5	1 = Totalmente en desacuerdo 2 = En desacuerdo 3 = Ni de acuerdo ni en desacuerdo 4 = De acuerdo 5 = Totalmente de acuerdo
Aprendizaje	Aprendizaje profundo	Las ideas que he encontrado	AP05	Acuerdo-Desacuerdo	5	1 = Totalmente en

También, si lo deseas puedes mirar el video (Figure 3) donde se explica la base de datos y su diccionario.

https://youtu.be/BTeLsLwsM9g

Figure 2: Explicación base de datos y diccionario.

3.2 Análisis Bivariado

3.2.1 Análisis ítem AP01 según Sexo (CA01)

Caracterice la proporción de sujetos que manifiestan distintos niveles de acuerdo con la afirmación "He tenido problemas para encontrarle sentido a las cosas que tengo que estudiar" (ítem AP01), diferenciando los resultados según género (CA01). Presente los resultados en una tabla que resuma la distribución y luego describa las principales tendencias observadas.

4 En Excel

Notará que puede ser complejo realizar explicaciones que permitan replicar los análisis en Excel. Vea el video y la explicación para replicar los resultados.

https://youtu.be/YpmNabPQbmE

Figure 3: Análisis univariado en Excel.

5 En R

Ahora interesa caracterizar la proporción de sujetos según el sexo (ítem CA01) que manifiestan distintos niveles de acuerdo con la siguiente afirmación: *He tenido problemas para encontrarle sentido a las cosas que tengo que estudiar* (ítem AP01).

En este caso, Sexo (CA01) actúa como variable independiente y el ítem AP01 como dependiente. Es decir, queremos inspeccionar si existe alguna diferencia en los niveles de acuerdo reportados para el ítem AP01 en función del sexo de los individuos.

5.1 Forma 1 (usando funcion table)

Este método funciona muy bien solo para tablas bivariadas, es decir, para realizar tablas de contigencia entre dos ítems. Este es un método rápido pero es poco flexible si quisieramos hacer una tabla cruzada para todos los ítems de la dimensión aprendizaje según sexo. El código se muestra a continuación:

```
library(dplyr)

tabla_1 <- table(db$AP01, db$CA01)
tabla_1 <- prop.table(tabla_1,2)
tabla_1 <- round(tabla_1,2)*100
tabla_1 <- as.data.frame.matrix(tabla_1)

tabla_1$AP01 <- as.character(rownames(tabla_1))
```

Ejecutando el código anterior, se observa la siguiente tabla:

	F	M	No indica	AP01
1	15	14	20	1
2	23	29	20	2
3	15	43	10	3
4	8	0	20	4
5	31	14	30	5
NA	8	0	0	NA

Podemos optimizar un poco más este código. Sería ideal que en la columna AP01 aparecieran los niveles de la escala Likert. Es decir nos gustaría hacer la siguiente conversión:

- 1 = Totalmente en Deacuerdo
- 2 = En desacuerdo
- 3 = Ni de acuerdo ni en Desacuerdo
- 4 = De acuerdo
- 5 = Completamente de acuerdo
- NA = Respuesta perdida

Podemos crear un dataframe con esta información para luego utilizarla en la tabla que construimos anteriormente.

```
library(dplyr)

conversion_likert <- data.frame( likert_numero = c('1',
                                    '2',
```

```
                        '3',
                        '4',
                        '5',
                        'NA'),
              likert_texto = c('Totalmente en Desacuerdo',
                        'En desacuerdo',
                        'Ni de acuerdo ni en Desacuerdo',
                        'De acuerdo',
                        'Completamente de acuerdo',
                        'Respuesta perdida')
            )
```

La construcción del dataframe se llama conversion_likert y se muestra a continuación

likert_numero	likert_texto
1	Totalmente en Desacuerdo
2	En desacuerdo
3	Ni de acuerdo ni en Desacuerdo
4	De acuerdo
5	Completamente de acuerdo
NA	Respuesta perdida

Entonces, ahora reemplazamos los valores numéricos de la escala Likert por su grado de acuerdo correspondiente. Para ello, utilizamos la función left_join(). Una vez que hacemos esta unión entre tablas, eliminamos la columna AP01.

```
library(dplyr)


tabla_1 <- tabla_1 %>%
          dplyr::left_join(. ,conversion_likert,
                    by=join_by(AP01 == likert_numero)) %>%
          dplyr::arrange(AP01) %>%
          relocate(., likert_texto, .before = 1) %>%
          dplyr::select(-c(AP01))
```

La tabla se ve de la siguiente forma

likert_texto	F	M	No indica
Totalmente en Desacuerdo	15	14	20
En desacuerdo	23	29	20
Ni de acuerdo ni en Desacuerdo	15	43	10
De acuerdo	8	0	20
Completamente de acuerdo	31	14	30
Respuesta perdida	8	0	0

5.2 Forma 2 (usando libreria **dplyr** y **tidyr**)

Esta opción es más compleja que la versión anterior, pero mucho más flexible. Usamos la librería dplyr para realizar el proceso. Antes se usar esta librería, vamos a escribir un pseudo-código para explicar de manera funcional cómo realizar el procesamiento de los datos. Este pseudocóigo expresa un flujo lógico por medio del cual se realizan acciones a la base de datos para obtener un resultado deseado. El pseudo código para este caso quedará expresado por la siguiente secuencia

5.2.0.1 Pseudo código:

Usa la base de datos de nombre db %LUEGO%
Selecciona las columnas AP01 y CA01 %LUEGO%
Agrupa por AP01 y CA01 %LUEGO%
Calcula la frecuencia total de casos %LUEGO%
desagrupa %LUEGO%
Agrupa por CA01 %LUEGO%
construye la columna de nombre total %LUEGO%
construye otra columna de nombre porcentaje %LUEGO%
selecciona las variables de nombre AP01, CA01 y porcentaje %LUEGO%

El pseudo-código anterior se traduce entonces en un código funcional que luego puede usarse como referencia para la implementación de funciones que provienen de la libreria dplyr. Para ello, se resaltó en color azul todos los procesos que se traducen en funciones usando la libreria dplyr. Por ejemplo, el proceso Selecciona columna se traduce en la función (o código) select() en la libreria dplyr. De forma análogca, el proceso Agrupa porse traduce en la función (o código) group_by() en la libreria dplyr.

Note además que el operador %LUEGO% expresa el flujo lógico en el procesamiento de los datos, y corresponde a la función lógica que se verá expresada por el símbolo pipe %>%. El pseudo-código presentado tiene como función devolver una tabla en formato Long, luego de haber seleccionado, agrupado y calculado. Este flujo se traduce en una tabla que es guardada en una entidad de nombre tabla_AP01_sexo.

Aplicamos el pseudo código anterior para implementar un código en R usando la libreria dplyr.Este código ejecutará los procesos para analizar nuestros datos.

```
library(dplyr)

tabla_AP01_sexo <- db %>%
        dplyr::select(AP01,CA01) %>%
        dplyr::group_by(AP01, CA01) %>%
```

```
        dplyr::summarise(frecuencia = n()) %>%
        dplyr::ungroup() %>%
        dplyr::group_by(CA01) %>%
        dplyr::mutate(total     = sum(frecuencia),
                      porcentaje = round(frecuencia/total*100,2)) %>%
        dplyr::select(AP01,CA01, porcentaje)
```

La tabla se muestra a continuación

AP01	CA01	porcentaje
1	F	15.38
1	M	14.29
1	No indica	20.00
2	F	23.08
2	M	28.57
2	No indica	20.00
3	F	15.38
3	M	42.86
3	No indica	10.00
4	F	7.69
4	No indica	20.00
5	F	30.77
5	M	14.29
5	No indica	30.00
NA	F	7.69

La tabla anterior necesitamos convertirla a un formato wide. Para ello, utilizamos la función pivot_wider()

```
library(dplyr)

tabla_AP01_sexo_w <- tidyr::pivot_wider(tabla_AP01_sexo,
                          names_from = CA01,
                          values_from = porcentaje)
```

El resultado queda

AP01	F	M	No indica
1	15.38	14.29	20
2	23.08	28.57	20
3	15.38	42.86	10
4	7.69		20
5	30.77	14.29	30
NA	7.69		

De manera similar al ejemplo univariado, cambiamos los valores de la escala likert por su respectivo nivel de acuerdo.

```
library(dplyr)
library(tidyr)

tabla_AP01_sexo_w <-tabla_AP01_sexo_w %>%
              dplyr::left_join(. ,conversion_likert,
                               by=join_by(AP01 == likert_numero)) %>%
              dplyr::arrange(AP01) %>%
              relocate(., likert_texto, .before = AP01) %>%
              dplyr::select(-c(AP01)) %>%
              dplyr::mutate(across(everything(), ~replace_na(.x, 0))) #reemplaza NA por 0


names(tabla_AP01_sexo_w)[1] <- codebook %>%
                          dplyr::filter(item_codigo=='AP01') %>%
                          dplyr::select(Pregunta)
```

El resultado final se ve a continuación.

He tenido problemas para encontrar el sentido a las cosas que tengo que estudiar.	F	M	No indica
Totalmente en Desacuerdo	15.38	14.29	20
En desacuerdo	23.08	28.57	20
Ni de acuerdo ni en Desacuerdo	15.38	42.86	10
De acuerdo	7.69	0.00	20
Completamente de acuerdo	30.77	14.29	30
Respuesta perdida	7.69	0.00	0

En un solo código, toda la instrucción anterior quedaría

```
library(dplyr)



 enunciado_P01<-  codebook %>%
          dplyr::filter(item_codigo=='AP01') %>%
          dplyr::select(Pregunta) %>%
          pull(.)


tabla_sexo_P01 <- db %>%
            dplyr::select(AP01,CA01) %>%
            dplyr::group_by(AP01, CA01) %>%
            dplyr::summarise(frecuencia = n()) %>%
            dplyr::ungroup() %>%
```

```
        dplyr::group_by(CA01) %>%
        dplyr::mutate(total     = sum(frecuencia),
                      porcentaje = round(frecuencia/total*100,2)) %>%
        dplyr::select(AP01,CA01, porcentaje)   %>%
# Convertir a formato WIDE
       tidyr::pivot_wider(.,
                      names_from = CA01,
                      values_from = porcentaje) %>%
# Convertir escalas numericas con sus categorias en texto
       dplyr::left_join(. ,conversion_likert,
                      by=join_by(AP01 == likert_numero)) %>%
       dplyr::arrange(AP01) %>%
       relocate(., likert_texto, .before = AP01) %>%
       dplyr::select(-c(AP01)) %>%
       dplyr::mutate(across(everything(), ~replace_na(.x, 0))) %>%
# Incluimos el enunciado y cambiamos nombre de las columnas
       dplyr::rename_with(~c(enunciado_P01,"Femenino","Masculino"), c(likert_texto,F,M))
```

Las respuesta vacias deberían considerarse en un 0%.

He tenido problemas para encontrar el sentido a las cosas que tengo que estudiar.	Femenino	Masculino	No indica
Totalmente en Desacuerdo	15.38	14.29	20
En desacuerdo	23.08	28.57	20
Ni de acuerdo ni en Desacuerdo	15.38	42.86	10
De acuerdo	7.69	0.00	20
Completamente de acuerdo	30.77	14.29	30
Respuesta perdida	7.69	0.00	0

5.2.1 Análisis ítems AP01-AP08 según Sexo (CA01)

Ahora interesa caracterizar la proporción de sujetos que manifiestan distintos niveles de acuerdo cada uno de los ítems de las dimensiones Aprendizaje Superficial (AP01-AP04) y Aprendizaje Profundo (AP05-AP08), diferenciando los resultados según género (CA01). Presente los resultados en una tabla que resuma la distribución y luego describa las principales tendencias observadas.

6 En Excel

Se entregará la solución el día 22 de septiembre.

7 En R

Se entregará la solución el día 22 de septiembre.