Guía de trabajo variables cuantitativas continuas

Diplomado en Análisis de datos con R para la acuicultura

true

29 abril 2021

Introducción ¿Cuál es la diferencia entre variable y dato?

Las **variables** son las características que se pueden medir en un individuo o en un ambiente y los **datos** son los valores que puede tomar esa variable.

¿Qué es una variable aleatoria?

Es una variable cuyo valor se determina por el azar. Las variables aleatorias se representan por letras mayúsculas (**X**) y sus valores numéricos por letras minúsculas (**xi**).

Objetivos de aprendizaje Los objetivos de aprendizaje de esta guía son:

1. Observar y predecir una variable aleatoria continua con distribución normal.

2. Elaborar un reporte dinámico en formato pdf.

Clasificación de variables cuantitativas

Tipo de variable	Descripción
Variables discretas:	Una variable ***Y*** es **discreta** si puede tomar valores puntales, pueden tener un número finito o infinito de valores.
Variables continuas:	Una variable ***Y*** es **continua** si puede tomar cualquier valor dentro de un intervalo del conjunto de los números reales. La probabilidad de que tome un valor cualquiera es 0 debido a que existe un número infinito de posibilidades en el intervalo.

Clasificación de variables cualitativas o categóricas

Tipo de variable	Descripción
Variables nominales:	Sus valores representan categorías que no obedecen a una clasificación intrínseca.
Variables ordinales:	Sus valores representan categorías con alguna clasificación intrínseca.

Distribución de una variable

Distribución empírica (observación): Los datos de una muestra, obtenidos de forma aleatoria de una población, pueden ser usados para observar su comportamiento o distribución.

Distribución de probabilidad (predicción): Las variables aleatorias tienen diferentes distribuciones de probabilidad subyacentes, lo que nos permite predecir su comportamiento y realizar inferencia estadística.

Distribución normal: Las variables cuantitativas continuas siguen una distribución normal.

¿Cómo puedo observar y predecir el comportamiento de una variable? Funciones clave

1.- Tabla de distribución de frecuencia
table()

2.- Histograma
hist()

3.- Gráfica x-y de puntos "p", lineas "l" o ambas "b".
plot()

4.- Gráfico de cajas y bigotes
boxplot()

5.- Mediante la función de densidad empírica
density()

6.- Mediante la función de distribución acumulada empírica
ecdf()

¿Qué puedo medir de una variable aleatoria continua? Funciones clave

1.- ***Medidas de tendencia central***:

media **mean()** y mediana **median()**.

2.- ***Medidas de dispersión***:

varianza **var()**, desviación estándar **sd()**.

3.- ***Concentración de datos en cuantiles***.

quantile().

Librerías a usar en esta guía {stat} , {graphics}, {readxl}

Ejercicios

Ejercicio 1. Elaborar archivo Rmarkdown

Elabore un archivo o file con extensión **.Rmd** y configurelo para exportar el resultado como un documento dinámico **pdf**. Utilice el siguiente ejemplo para completar la información de **metadatos**: Título: Reporte variables continuas, nombre del autor: Su nombre.

Luego guarde inmediatamente su script como **script__3__nombre__apellido.Rmd**. Al finalizar la actividad deberá exportar y almacenar este ***script*** en su carpeta drive de tareas.

Ejercicio 2. Configuración del reporte

En el primer bloque de códigos o **chunk** configure los comandos de la siguiente manera ***knitr::opts_chunk$set(echo = TRUE)*** y cargue las librerías **readxl**, **stats** y **graphics** usando la función ***library()***.

```
knitr::opts_chunk$set(echo = TRUE)
library(readxl)
library(stats)
library(graphics)
```

Ejercicio 3. Borrar información de la plantilla

Borre los bloques de códigos **R** que se generan automáticamente con cada archivo **.Rmd** y reemplacelos por nuevos bloques de códigos con el botón verde **+C** que se encuentra en la parte superior del panel de códigos.

Ejecute cada uno de los siguientes ejercicios en uno o más bloques de códigos diferentes. Sea ordenado y documente su reporte adecuadamente.

Ejercicio 4. Importar y explorar datos

Cree un objeto llamado **dat** e importe el set de datos **shrimp** usando la función ***read_excel()*** de la librería **readxl**. Explore el set de datos usando las funciones **head()**, **tail**, **summary()** y **str()**.

```
dat <- read_excel("shrimp.xlsx")
head(dat)
```

```
## # A tibble: 6 x 2
##   sample_id Weight
##       <dbl> <chr>
## 1         1 17.2
## 2         2 18.8
## 3         3 27.8
## 4         4 20.4
## 5         5 20.6
## 6         6 28.6
```

```
tail(dat)
```

```
## # A tibble: 6 x 2
##   sample_id Weight
##       <dbl> <chr>
## 1       195 13.4
## 2       196 30.0
## 3       197 23.0
## 4       198 13.7
## 5       199 16.9
## 6       200 14.1
```

```
summary(dat)
```

```
##    sample_id         Weight
##  Min.   :  1.00   Length:200
##  1st Qu.: 50.75   Class :character
##  Median :100.50   Mode  :character
##  Mean   :100.50
##  3rd Qu.:150.25
##  Max.   :200.00
```

```
str(dat)
```

```
## tibble[,2] [200 x 2] (S3: tbl_df/tbl/data.frame)
##  $ sample_id: num [1:200] 1 2 3 4 5 6 7 8 9 10 ...
##  $ Weight   : chr [1:200] "17.2" "18.8" "27.8" "20.4" ...
```

Ejercicio 5. Corrección de variables

Note que en el ejercicio anterior la variable **weight** fue erradamente codificada como caracter o texto (chr) en vez de número. Use la función **as.numeric()** para corregir este error. Vuelva a ejecutar los comandos **summary()** y **str()** para comprobar que las variables están adecuadamente codificadas.

```
dat$Weight <- as.numeric(dat$Weight)
summary(dat)
```

```
##    sample_id          Weight
##  Min.   :  1.00   Min.   : 8.50
##  1st Qu.: 50.75   1st Qu.:16.90
##  Median :100.50   Median :19.70
##  Mean   :100.50   Mean   :19.96
##  3rd Qu.:150.25   3rd Qu.:22.82
##  Max.   :200.00   Max.   :36.20
```

```
str(dat)
```

```
## tibble[,2] [200 x 2] (S3: tbl_df/tbl/data.frame)
##  $ sample_id: num [1:200] 1 2 3 4 5 6 7 8 9 10 ...
##  $ Weight   : num [1:200] 17.2 18.8 27.8 20.4 20.6 28.6 22.3 13.7 16.6 17.8 ...
```

Ejercicio 6. Observar el comportamiento de una variable

A partir del set de datos **shrimp** elabore un histograma y un boxplot de la variable cuantitativa continua **weight**. Use las funciones **hist()**, **boxplot()**.

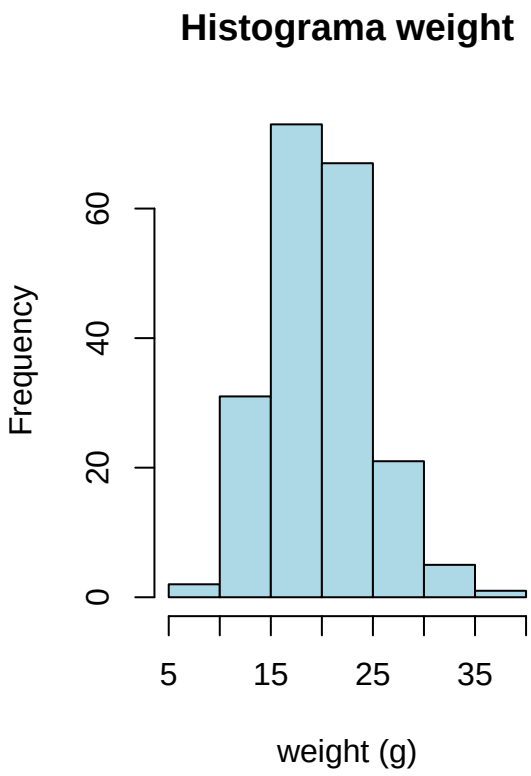
Use este comando **par(mfrow=c(1,2))** para unir las gráficas en un solo panel con dos columnas.

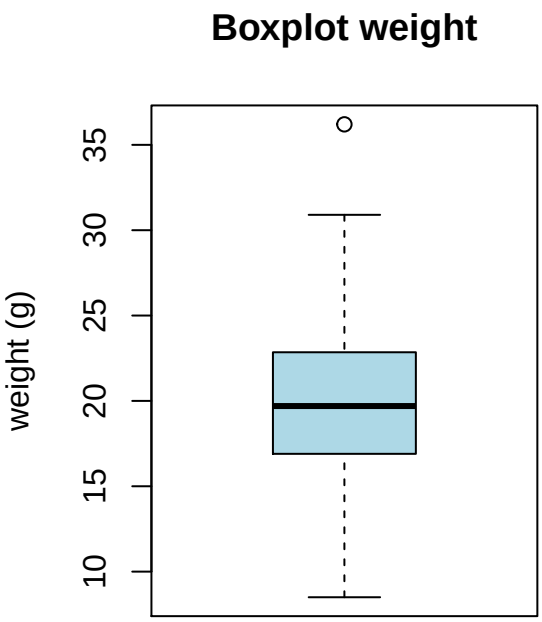
Investigue que representa el circulo que aparece en el Boxplot.

```
par(mfrow=c(1,2))

hist(dat$Weight, col="light blue", main = "Histograma weight", xlab = "weight (g)")

boxplot(dat$Weight, col="light blue", main = "Boxplot weight", ylab = "weight (g)")
```





Ejercicio 6. Métricas del set de datos shrimp

Calcule las siguientes métricas del set de datos: promedio, desviación estándar, rango y cuantiles.

```
mean(dat$Weight)
```

```
## [1] 19.958
```

```
sd(dat$Weight)
```

```
## [1] 4.710923
```

```
range(dat$Weight)
```

```
## [1]  8.5 36.2
```

```
quantile(dat$Weight)
```

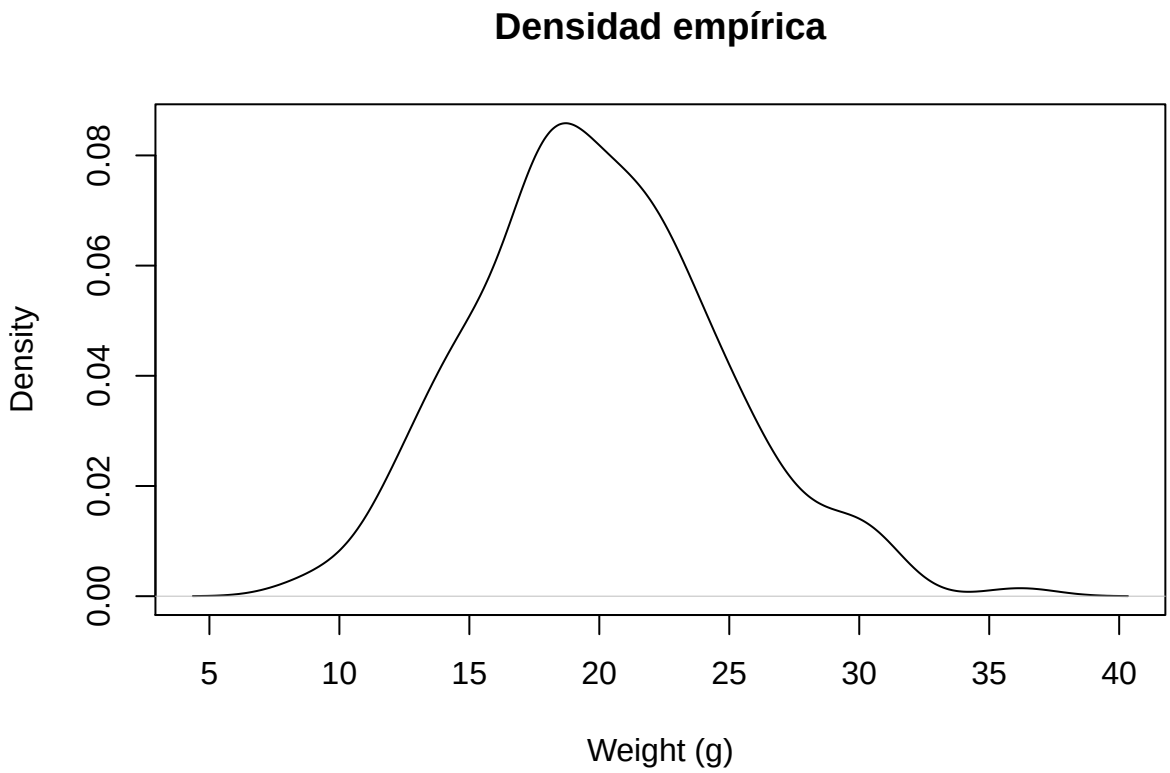
```
##     0%    25%    50%    75%   100%
##  8.500 16.900 19.700 22.825 36.200
```

Ejercicio 7. Función de densidad

Usando la función **plot()** elabore:

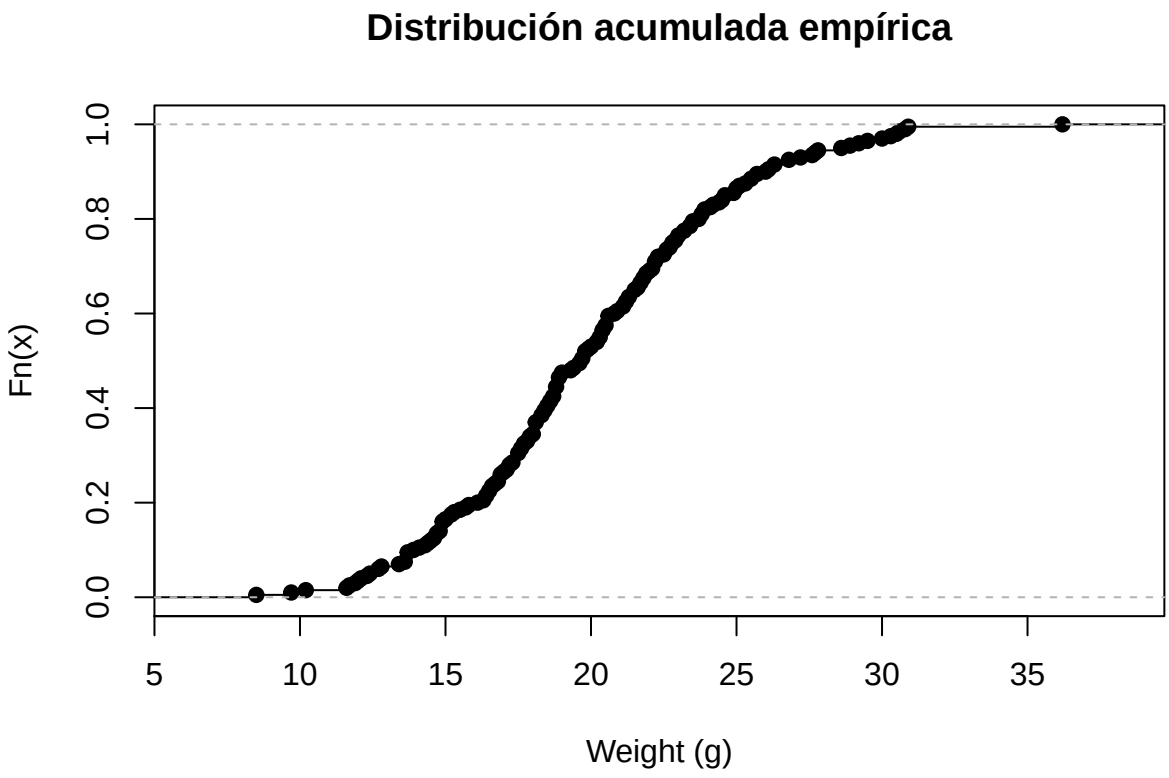
a). Gráfico con la densidad empírica. Debe incluir la función **density()** dentro de plot.

```
# Densidad empírica.
plot(density(dat$Weight), main="Densidad empírica", xlab="Weight (g)")
```



b). Gráfico con la distribución acumulada empírica. Debe incluir la función **ecdf()** dentro de plot.

```
# Distribución acumulada empírica.
plot(ecdf(dat$Weight), main="Distribución acumulada empírica", xlab="Weight (g)")
```



Ejercicio 8. Crear una función y predecir datos observados

Utilice la función de distribución acumulada empírica **ecdf()** para determinar que proporción de camarones es menor a 20 g, y que proporción es mayor de 30 g:

a). Primero cree una función de distribución acumulada empírica para los datos del peso de sus camarones.

```
Fn <- ecdf(dat$Weight)
Fn
```

```
## Empirical CDF
## Call: ecdf(dat$Weight)
##  x[1:125] =    8.5,    9.7,   10.2,  ...,   30.9,   36.2
```

b). Calcule la proporción de camarones menores de 20 g.

```
# Fn(x) returns the percentiles for x
paste0(Fn(20)*100, "%")
```

```
## [1] "53%"
```

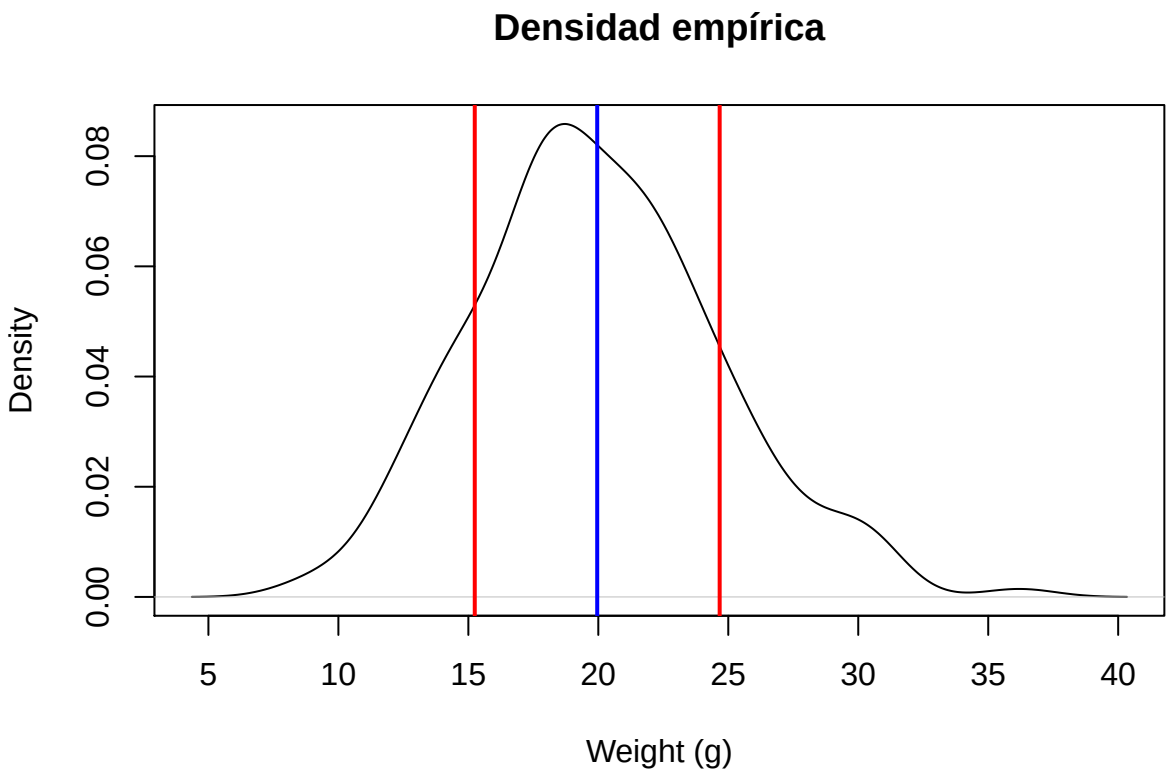
c). Calcule la proporción de camarones mayor a 30 g.

```r
# 1- Fn(x) returns 1 - the percentiles for x
paste0(((1 - Fn(30))*100), "%")
```

```
## [1] "3%"
```

Ejercicio 9. Proporción de datos entorno a la media

En la siguiente figura, la linea roja representa 1 de sobre y bajo la media, la linea azul representa la media.



¿Qué proporción de los datos está contenido entre una desviación estándar hacia arriba y hacia abajo de la media?.

a). Calcule la proporción de datos 1 desviación estándar sobre la media.

```r
p1 <- 1 - Fn(19.958 + 4.710923)
p1
```

```
## [1] 0.15
```

b). Calcule la proporción de datos 1 desviación estándar bajo la media.

```r
p2 <- Fn(19.958 - 4.710923)
p2
```

```
## [1] 0.175
```

c). Calcule la proporción de datos entre 1 desviación estándar arriba y abajo de la media multiplicado por 100.

```
paste0(((1 - (p1 + p2))*100),"%")
```

```
## [1] "67.5%"
```

```
# Este valor es muy cercano al 68% teórico de una variable con distribución normal.
```