Modelación Estadística

Actividada 2: Estadística Descriptiva

Dr. Edlin Guerra

Análisis exploratorios y estadística descrptiva

Hagamos un análisis exploratorio a los datos de los pinguinos de la Antártida del género *Pygoscelis*. Lo primero que debe hacer es instalar el paquete de datos `palmerpenguins` y cárguelo en su sesión. Luego haga lo siguiente:

1. Busque en la pestaña *Help* para averiguar qué es el paquete *palmerpenguin*.
2. Identifique la base de datos `penguins` y cárguela en su ambiente global con `data(penguins)`.
3. Identifique cuántas variables hay, cuál es la naturaleza de cada una de ellas (tipo de variable, escala), así como cuáles pueden ser consideradas causales y cuáles respuesta.
4. Efectué un gráfico de dispersión entre las variables `body_mass_g` y `flipper_lemgth_mm`. Mida el grado de asociación ¿cómo lo haría? Una forma de hacer estas cosas es graficando la asociación, y estimando la correlación (método que se desarrollará en otro laboratorio, pero acá es pide con fines demostrativos)

```
plot(penguins$body_mass_g, penguins$flipper_length_mm)

#¿qué hace cor()?
cor(penguins$body_mass_g, penguins$flipper_length_mm)

#Si no obtuvo resultado, trate de resolverlo con el argumento "use"
```

Usemos el paquete `ggplot2` para mejorar el gráfico:

```
pp <- ggplot(data = penguins, aes(x = flipper_length_mm,
                                  y = body_mass_g,
                                  colour = species))+
      geom_point()
```

```
pp

#Mejoremos con capas
pp +  theme_bw()+
      xlab("Largo de aleta (mm)")+
      ylab("Masa corporal (g)")+
      scale_y_continuous(breaks = seq(2600,6400,400))+
      scale_x_continuous(breaks = seq(170,240,5))

#¿cuál gráfico le gustó más?
```

Llegados a este punto, vamos hacer algo de estadística descriptiva. Esta es la parte que deberán entregar como tarea. Calculen promedio, varianza, desviación estándar, valor mínimo y máximo, cuartiles, simetría y curtosis a la variable *masa corporal*. Usen para ello las funciones recomendadas y responda las siguientes preguntas:

```
# Para facilitar cálculos, vamos a remover los datos sin registro
# (identificados con NA), usando el siguiente código:

penguins2 <- penguins %>%
  na.omit()
```

PREGUNTAS

1. Copia el comando que sigue. ¿Qué se calculó?

```
xx<-penguins2$body_mass_g
sum(xx, na.rm = T)/length(xx)
```

2. Busca y aplica una función que ejecute la linea de comando anterior. PISTA: escribe `?mean` en la cónsola y presiona *enter* en tu teclado para ampliar tu búsqueda.
3. Calcula la mediana de la masa corporal usando la función correspondiente.
4. Calcula la varianza y desviación estándar de la masa corporal usando el comando `var()`
5. ¿Cuál es la diferencia entre estas dos fórmulas? ¿Representan lo mismo?

```
sum((xx-mean(xx))^2)/length(xx)
sum((xx-mean(xx))^2)/(length(xx)-1)
```

6. Con base en el valor de la varianza y usando operadores aritméticos, calcula la desviación estándar de la masa corporal. Confirma tu resultado usando la función `sd()`.

7. Explore el rango de la masa corporal identificando mínimos y máximos con la función `min()` y `max()`, respectivamente.
8. Ahora estime los cuartiles de la masa corporal con la función `quantile()`
9. Describa la forma de la distribución de la masa corporal usando la simetría y curtosis con las funciones `skewness()` y `kurtosis()`.
10. Todas estas estimaciones ignoran las posibles diferencias en la masa corporal entre las especies. ¿Qué le dice este gráfico?

```
boxplot(body_mass_g~species, data = penguins2)
```

11. Calcule estos estimadores para cada especie usando el paquete **dplyr** y sus funciones `group_by()` y `summarize()`. Estas líneas de comando lo ayudarán (interprete los resultados):

```
library(dplyr)

penguins %>%
  group_by(species) %>%
  summarise(media = mean(body_mass_g, na.rm =T),
            desviacion = sd(body_mass_g, na.rm =T),
            simetria = skewness(body_mass_g, na.rm =T),
            curtosis = kurtosis(body_mass_g, na.rm =T))
```

12. Usando como guía el libro digital R Graphics Cookbook, genere: (i) una distribución de frecuencias con histograma, (ii) una distribución de frecuencias basada en densidad, (iii) un diagrama de cajas que incluya promedio. En los tres casos la masa corporal debe distinguirse por especie. Recuerde cargar las respuestas a estas preguntas en el formato de **Google Doc** generado en el Google Classroom.