```
---
title: "Parcial 1"
author: "Laura Valentina Gonzalez Rodriguez"
date: "2023-03-03"
output:
  pdf_document: default
  html_document: default
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
# Se cargan algunas librerias
packages = c("dslabs", "MASS", "scatterplot3d", "car", "dplyr", "kableExtra", "plotrix")
## Se cargan o se instalan y cargan
package.check <- lapply(
  packages,
  FUN = function(x) {
    if (!require(x, character.only = TRUE)) {
      install.packages(x, dependencies = TRUE)
      library(x, character.only = TRUE)
    }
  }
)

load("penguins.RData")
```

Primer Punto

Un investigador cree que la media de las variables medidas es 30 en todas ellas. Realice una prueba de hipótesis para comprobar si esa afirmación es cierta.

Primero se cargaran las variables de interes: longitud del pico (bill length), la profundidad del pico (bill depth) y la logintud de la aleta (fipper length) para realizar el T^2 respecto a H0 = [30,30,30].


```{r}
pengu = penguins2;
data_objetivo = subset(pengu, select = c("bill_length_mm", "bill_depth_mm", "flipper_length_mm" ));data_objetivo

h0 = c(30,30,30);
```
```

A tibble: 333 × 3

bill_length_mm <dbl>	bill_depth_mm <dbl>	flipper_length_mm <int>
39.1	18.7	181
39.5	17.4	186
40.3	18.0	195
36.7	19.3	193
39.3	20.6	190
38.9	17.8	181
39.2	19.6	195
41.1	17.6	182
38.6	21.2	191
34.6	21.1	198

1-10 of 333 rows Previous 1 2 3 4 5 6 … 34 Next

Para eso, se define el codigo en R que realiza la distribución T^2 y la comparativa con la multiplicación de Fisher.

```{r}

T2 = function (X, mu_0) { # Esto es la T^2
  x_bar = colMeans(X);
  S = cov(X);

  n = nrow(X);
  return(n * t(x_bar - mu_0) %*% solve(S) %*% (x_bar - mu_0));
}

F_val = function (X, mu_0, alpha = 0.05) { # Esto es fisher
  p = length(mu_0);
  n = nrow(X)

  df_num = p;
  df_denom = n - p;

  return(qf(1 - alpha, df_num, df_denom));
}

F_val_lone = function (X, alpha = 0.05) { # Esto es fisher
  p = ncol(X);
  n = nrow(X)
```

```r
F_val_lone = function (X, alpha = 0.05) { # Esto es fisher
  p = ncol(X);
  n = nrow(X)

  df_num = p;
  df_denom = n - p;

  return(qf(1 - alpha, df_num, df_denom));
}


T2_F = function(X, mu_0, alpha = 0.05) { # Una vaiana rara que utiliza fisher y maso la T^2
  p = length(mu_0);
  n = nrow(X);

  df_num = (n - 1) * p;
  df_denom = n - p;
  return(df_num / df_denom * F_val(X, mu_0, alpha));
}


T2_test = function(X, mu_0, alpha = 0.05) { # T2 debe ser menor que T2_F para aceptar
  return(T2(X, mu_0) < T2_F(X, mu_0, alpha)); # Si retorna true, se acepta mu_0.
}



```

Se reemplazan los valores en la función anterior obteniendo que:

```r

Valor_t2 = T2(data_objetivo,h0); Valor_t2

Valor_fish = T2_F(data_objetivo,h0); Valor_fish

T2_test(data_objetivo,h0)
```

```
          [,1]
[1,] 68512.99
[1] 7.94378
      [,1]
[1,] FALSE
```

Particularmente para este caso, se tiene que $T^2=68512.99 > 7.94378 = 3.0181F_{3,300}(0.05)$. Por lo cuál, se rechaza la hipotesis de que la media cada una de las variables es 30 ($H_0 : \mu \neq [30,30,30]$) con un nivel de 5% de significancia.

Segundo Punto

En la muestra, ¿cuál es la especie que tiene más alta cada una de las mediciones?

Las mediciones anteriores (longitud del pico (bill length), la profundidad del pico (bill depth) y la logintud de la aleta (fipper length)), se deben agrupar particularmene por la especie, correspondiente a species(1,2,3)

```r
data_especies = subset(pengu, select = c("species","bill_length_mm", "bill_depth_mm", "flipper_length_mm" ));
datos_especies <- split(data_especies, data_especies$species); datos_especies
```

A tibble: 146 × 4

species <dbl>	bill_length_mm <dbl>	bill_depth_mm <dbl>	flipper_length_mm <int>
1	39.1	18.7	181
1	39.5	17.4	186
1	40.3	18.0	195
1	36.7	19.3	193
1	39.3	20.6	190
1	38.9	17.8	181
1	39.2	19.6	195
1	41.1	17.6	182
1	38.6	21.2	191
1	34.6	21.1	198

A tibble: 119 × 4

species <dbl>	bill_length_mm <dbl>	bill_depth_mm <dbl>	flipper_length_mm <int>
2	46.1	13.2	211
2	50.0	16.3	230
2	48.7	14.1	210
2	50.0	15.2	218
2	47.6	14.5	215
2	46.5	13.5	210
2	45.4	14.6	211
2	46.7	15.3	219
2	43.3	13.4	209
2	46.8	15.4	215

1-10 of 119 rows

A tibble: 68 × 4

species <dbl>	bill_length_mm <dbl>	bill_depth_mm <dbl>	flipper_length_mm <int>
3	46.5	17.9	192
3	50.0	19.5	196
3	51.3	19.2	193
3	45.4	18.7	188
3	52.7	19.8	197
3	45.2	17.8	198
3	46.1	18.2	178
3	51.3	18.2	197
3	46.0	18.9	195
3	51.3	19.9	198

1-10 of 68 rows

Previous 1 2 3 4 5 6 7 Ne

Con lo cuál, podemos generar una tabla comparativa para cada subconjunto.

Para la especie 1, Adelie se tiene:

```{r}
grupo_1 <- datos_especies[[1]][,-which(names(datos_especies[[i]]) == "species")];

sapply(grupo_1, function(x) {
  data.frame(
  mean = round(mean(x), digits = 3),
  median = median(x),
  mode = names(which.max(table(x))),
  sd = round(sd(x), digits = 3)
  )
}) %>%  kable() %>% kable_styling(latex_options = "hold_position")
```

	bill_length_mm	bill_depth_mm	flipper_length_mm
mean	38.824	18.347	190.103
median	38.85	18.4	190
mode	41.1	18.5	190
sd	2.663	1.219	6.522

Para la especie 2, Gentoo se tiene:

```{r}
grupo_2 <- datos_especies[[2]][,-which(names(datos_especies[[i]]) == "species")];

sapply(grupo_2, function(x) {
  data.frame(
  mean = round(mean(x), digits = 3),
  median = median(x),
  mode = names(which.max(table(x))),
  sd = round(sd(x), digits = 3)
  )
}) %>%  kable() %>% kable_styling(latex_options = "hold_position")
```

	bill_length_mm	bill_depth_mm	flipper_length_mm
mean	47.568	14.997	217.235
median	47.4	15	216
mode	45.2	15	215
sd	3.106	0.986	6.585

Para la especie 3, Chinstrap se tiene:

```{r}
grupo_3 <- datos_especies[[3]][,-which(names(datos_especies[[i]]) == "species")];

sapply(grupo_3, function(x) {
  data.frame(
  mean = round(mean(x), digits = 3),
  median = median(x),
  mode = names(which.max(table(x))),
  sd = round(sd(x), digits = 3)
  )
}) %>%  kable() %>% kable_styling(latex_options = "hold_position")
```

	bill_length_mm	bill_depth_mm	flipper_length_mm
mean	48.834	18.421	195.824
median	49.55	18.45	196
mode	51.3	17.3	187
sd	3.339	1.135	7.132

Para la medición de bill_length_mm, la media de la especie 3 (Chinstrap) fue la mayor con un valor de 48.834 mm. Para la medición de bill_depth_mm, la media de la especie especie 3 (Chinstrap) fue ligeramente la mayor con un valor de 18.421 mm en compración a la especie 1 (Adelie) con 18.347 mm. Finalmente para flipper_length_mm la media de la especie 2 (Gentoo) fue la mayor con un valor de 217.235 mm.

Tercer Punto

En la muestra, ¿cuál es la isla en la que es más alta cada una de las mediciones?

Las mediciones anteriores (longitud del pico (bill length), la profundidad del pico (bill depth) y la logintud de la aleta (fipper length)), se deben agrupar particularmene por la especie, correspondiente a island(1,2,3)

```{r}
data_isla = subset(pengu, select = c("island","bill_length_mm", "bill_depth_mm", "flipper_length_mm" ));
datos_isla <- split(data_isla, data_isla$island); datos_isla
```

A tibble: 163 × 4

island <dbl>	bill_length_mm <dbl>	bill_depth_mm <dbl>	flipper_length_mm <int>
1	37.8	18.3	174
1	37.7	18.7	180
1	35.9	19.2	189
1	38.2	18.1	185
1	38.8	17.2	180
1	35.3	18.9	187
1	40.6	18.6	183
1	40.5	17.9	187
1	37.9	18.6	172
1	40.5	18.9	180

1-10 of 163 rows

Previous 1 2 3 4 5 6 … 17 N

A tibble: 123 × 4

island <dbl>	bill_length_mm <dbl>	bill_depth_mm <dbl>	flipper_length_mm <int>
2	39.5	16.7	178
2	37.2	18.1	178
2	39.5	17.8	188
2	40.9	18.9	184
2	36.4	17.0	195
2	39.2	21.1	196
2	38.8	20.0	190
2	42.2	18.5	180
2	37.6	19.3	181
2	39.8	19.1	184

1-10 of 123 rows

Previous 1 2 3 4 5 6 … 13 Next

A tibble: 47 × 4

island	bill_length_mm	bill_depth_mm	flipper_length_mm
<dbl>	<dbl>	<dbl>	<int>
3	39.1	18.7	181
3	39.5	17.4	186
3	40.3	18.0	195
3	36.7	19.3	193
3	39.3	20.6	190
3	38.9	17.8	181
3	39.2	19.6	195
3	41.1	17.6	182
3	38.6	21.2	191
3	34.6	21.1	198

1-10 of 47 rows

Previous 1 2 3 4 5 Next

Con lo cuál, podemos generar una tabla comparativa para cada subconjunto.

Para la isla 1, Torgesen, se tiene:

```{r}
grup_1 <- datos_isla[[1]][,-which(names(datos_isla[[i]]) == "island")];

sapply(grup_1, function(x) {
  data.frame(
  mean = round(mean(x), digits = 3),
  median = median(x),
  mode = names(which.max(table(x))),
  sd = round(sd(x), digits = 3)
  )
}) %>%  kable() %>% kable_styling(latex_options = "hold_position")
```

	bill_length_mm	bill_depth_mm	flipper_length_mm
mean	45.248	15.907	209.558
median	45.8	15.6	213
mode	45.2	15	215
sd	4.827	1.828	14.282

Para la isla 2, Biscoe, se tiene:

```{r}
grup_2 <- datos_isla[[2]][,-which(names(datos_isla[[i]]) == "island")];

sapply(grup_2, function(x) {
  data.frame(
  mean = round(mean(x), digits = 3),
  median = median(x),
  mode = names(which.max(table(x))),
  sd = round(sd(x), digits = 3)
  )
}) %>%  kable() %>% kable_styling(latex_options = "hold_position")
```

	bill_length_mm	bill_depth_mm	flipper_length_mm
mean	44.222	18.34	193.187
median	45.2	18.4	193
mode	36	18.5	190
sd	5.947	1.137	7.429

Para la isla 3, Dream, se tiene:

```{r}
grup_3 <- datos_isla[[3]][,-which(names(datos_isla[[i]]) == "island")];

sapply(grup_3, function(x) {
  data.frame(
  mean = round(mean(x), digits = 3),
  median = median(x),
  mode = names(which.max(table(x))),
  sd = round(sd(x), digits = 3)
  )
}) %>%  kable() %>% kable_styling(latex_options = "hold_position")
```

	bill_length_mm	bill_depth_mm	flipper_length_mm
mean	39.038	18.451	191.532
median	39	18.4	191
mode	34.6	17	190
sd	3.028	1.346	6.22

Para la medición de bill_length_mm, la media de la isla 1 (Torgesen) fue la mayor con un valor de 45.248 mm. Para la medición de bill_depth_mm, la media de la isla 3 (Dream) fue ligeramente la mayor con un valor de 18.451 mm. Finalmente para flipper_length_mm la media de la isla 1 (Torgesen) fue la mayor con un valor de 209.558 mm.

Cuarto Punto

Encuentre y grafique la elipse de confianza para la longitud y la profundidad del pico de la especie Adelie en la isla Biscoe. ¿Cuál es el centro de la elipse?

Primero se debe obtener el dataset de dos columnas (bill length y bill depth) filtrado a la especie 1 (Adelie) y la isla 2 (Biscoe).

```{r}
data_elipse = subset(pengu, select = c("species","island", "bill_length_mm", "bill_depth_mm"));
datos_elipse <- split(data_elipse, data_elipse$species);
grupo_especie <- datos_elipse[[1]][,-which(names(datos_especies[[i]]) == "species")];
data_espe <- split(grupo_especie, grupo_especie$island);

isla_especie <- data_espe[[2]][,-which(names(data_espe[[i]]) == "island")];isla_especie

```

A tibble: 55 × 2

bill_length_mm <dbl>	bill_depth_mm <dbl>
39.5	16.7
37.2	18.1
39.5	17.8
40.9	18.9
36.4	17.0
39.2	21.1
38.8	20.0
42.2	18.5
37.6	19.3
39.8	19.1

Con la data seleccionada, se encuentran los ejes de confianza:

```{r}
sqrt_axes <-function(Z,i){
  p = ncol(Z);
  n = nrow(Z);
  einz = eigen(cov(Z));
  return(sqrt(einz$values[i])*sqrt(((p*(n-1))/(n*(n-p)))*F_val(Z,h0)))
}

major_axes <- function(Z){
  p = ncol(Z);
  n = nrow(Z);
  Z_bar = colMeans(Z);
  einz = eigen(cov(Z));
  z1 = Z_bar + sqrt_axes(Z,1) *einz$vectors[1];
  z2 =  Z_bar - sqrt_axes(Z,1)*einz$vectors[1];
  print("Ejes Mayores")
  return(c(z1,z2));
}

minor_axes <- function(Z){
  p = ncol(Z);
  n = nrow(Z);
  Z_bar = colMeans(Z);
  einz = eigen(cov(Z));
  z1 = Z_bar + sqrt_axes(Z,2)*einz$vectors[2];
  z2 =  Z_bar - sqrt_axes(Z,2)*einz$vectors[2];
  print("Ejes Menores")
  return(c(z1,z2));
}

major_axes(isla_especie)
minor_axes(isla_especie)
```

```
[1] "Ejes Mayores"
bill_length_mm  bill_depth_mm bill_length_mm  bill_depth_mm
      37.72647       17.44647       39.31353       19.03353
[1] "Ejes Menores"
bill_length_mm  bill_depth_mm bill_length_mm  bill_depth_mm
      38.43893       18.15893       38.60107       18.32107
```

Y apartir de los ejes, la longitud del mayor y menor

Y apartir de los ejes, la longitud del mayor y menor

```{r}
long_axes <-function(X){
  einz = eigen(cov(X));
  n = nrow(X);
  l1 = sqrt(einz$values[1])*sqrt((1/n)*F_val_lone(X));
  l2 = sqrt(einz$values[2])*sqrt((1/n)*F_val_lone(X));
  return(c(l1,l2))
}
long_axes(isla_especie)
```

```
[1] 0.6157383 0.2275529
```

Adicionalmente sabemos que el promedio de las variables, será el cetro de la elipse.

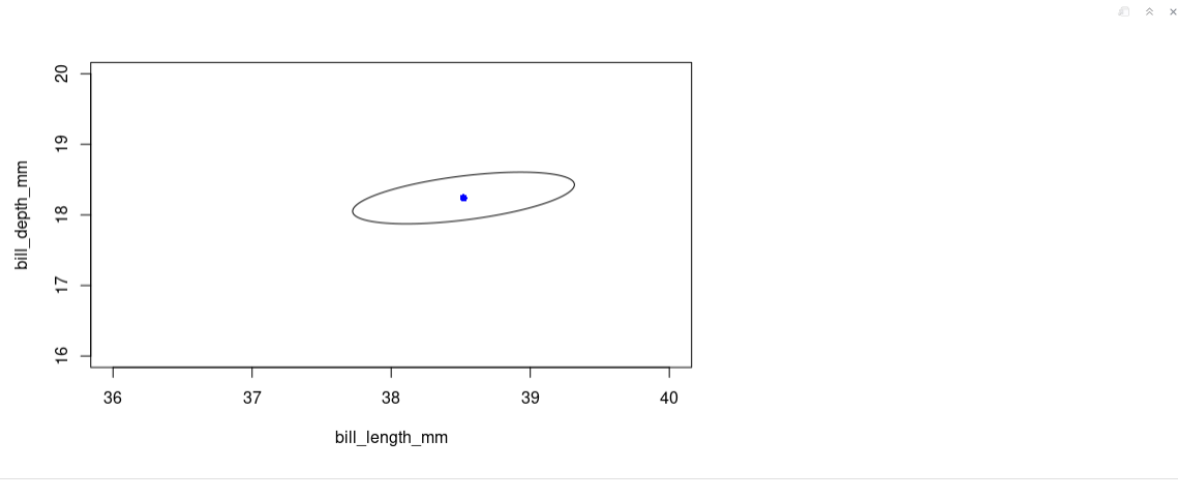
```{r}
colMeans(isla_especie)
```

```
bill_length_mm  bill_depth_mm
         38.52          18.24
```

Finalmente, se realiza la gráfica, de centro en azul:

```{r}
make_hip <- function(Z){
  p = ncol(Z);
  n = nrow(Z);
  angle = atan(eigen(cov(Z))$vectors[2,1]/eigen(cov(Z))$vectors[1,1])
  plot(0,pch='',ylab='bill_depth_mm',xlab='bill_length_mm',xlim=c(36,40),ylim=c(16,20))
  draw.ellipse(x=colMeans(Z)[1],y=colMeans(Z)[2],a=sqrt_axes(Z,1),b=sqrt_axes(Z,2),angle=angle,deg=FALSE)
  points(colMeans(Z)[1], colMeans(Z)[2], col = "blue", pch = 16)
}

make_hip(isla_especie)
```



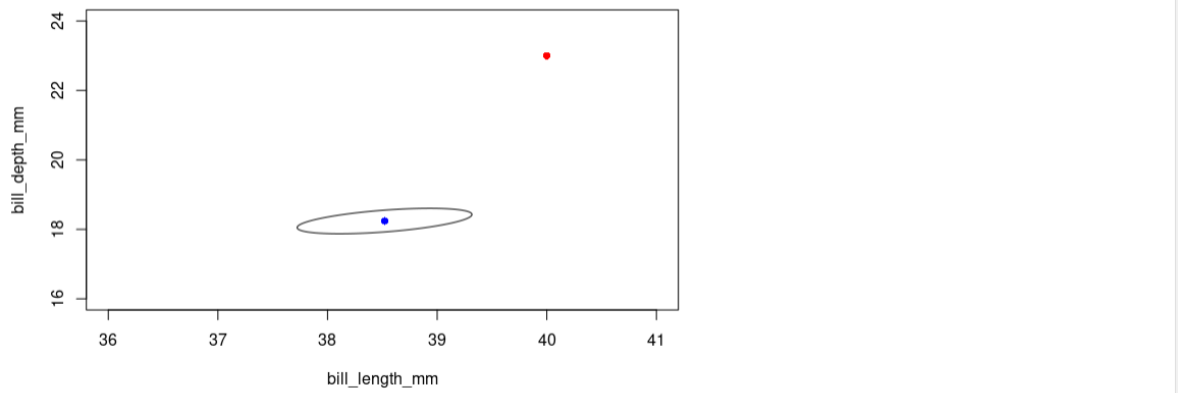
Quinto Punto

Si se afirma que, en la isla Biscoe, la especie Adelie tiene 40cm y 23cm como medias de la longitud y la profundidad del pico respectivamente, ¿ese punto está dentro de la región de confianza encontrada en el punto anterior? ¿Qué puede concluir de esto?

El punto rojo será las medias de la especie Adelie (40,23), generando la gráfica:

```{r}
make_hip_dot <- function(Z,d){
  p = ncol(Z);
  n = nrow(Z);
  angle = atan(eigen(cov(Z))$vectors[2,1]/eigen(cov(Z))$vectors[1,1])
  plot(0,pch='',ylab='bill_depth_mm',xlab='bill_length_mm',xlim=c(36,41),ylim=c(16,24))
  draw.ellipse(x=colMeans(Z)[1],y=colMeans(Z)[2],a=sqrt_axes(Z,1),b=sqrt_axes(Z,2),angle=angle,deg=FALSE)
  points(d[1], d[2], col = "red", pch = 16)
  points(colMeans(Z)[1], colMeans(Z)[2], col = "blue", pch = 16)
}

make_hip_dot(isla_especie, c(40,23))
```



Podemos observar graficamente que para la especie Adelie con medias (40,23) de longitud y profundidad, está afuera de la región de confianza encontrada anteriormente. Esto implica que la probabilidad de que la especie Adelie tenga de medidas 40 de longitud y 23 de profundidad es muy baja, suponiendo que los datos siguen esta distribución. En pocas palabras, es un punto atípico.