Práctico ANCOVA

Antes que nada, bajen el modelo de informe con el siguiente comando en la consola:

```
download.file("https://git.io/Informe",
              "Informe-2.Rmd")
```

Ovinos

La cría de ovejas es una actividad típica de Tierra del Fuego. El cordero fueguino tiene muy buena reputación por su gran sabor. Igualmente, las ovejas y carneros también son usados por su carne. En la industria frigorífica se paga menos el kilo animal más graso porque en proporción tiene menos carne.

Un frigorífico quiere saber si las hembras tienen la misma cantidad de grasa que los machos. Para ello han faenado cinco hembras y cinco machos. En cada caso, además de registar el sexo de animal, pesaron la carcasa en kg. y midieron la grasa como la profundida de la misma a la altura de la doceava costilla

```
ovejas <- read.table(url("https://git.io/oveja1.txt"),
           sep = "\t", header = TRUE)
```

Lo primero que vamos a hacer es graficar los datos de grasa por sexo. Para hacerlo, usamos:

```
ggplot(ovejas, aes(x = Kind, y = Fatness)) +
  geom_point()
```

La primer función crea la "base" del gráfico, usando el objeto `ovejas` como datos. Luego, la función `aes` indica que use la columna `Kind` para el eje x y `Fatness` para el y. Luego agregamos que queremos hacer puntos con esos datos con `geom_point`.

Aunque así es medio díficil ver si hay diferencias, podemos agregar un boxplot por encima

```
ggplot(ovejas, aes(Clase, y = Grasa)) +
  geom_boxplot() +
  geom_point(colour = "gray")#color gris para diferenciar de los puntos extremos
```

Este caso es similar al anterior, pero agregamos el gráfico de cajas y barras con `geom_boxplot` y se pone antes que `geom_point` porque las formas se gráfican en el orden que se llaman. De otra forma la caja taparía a los puntos.

Ejercicio 1

```
Solo teniendo en cuenta los datos usadas para los gráficos de arriba.
¿Les parece que pueden llegar a encontar diferencias entre sexos?
¿Qué tipo de test podrían usar para ver si las diferencias son significativas?
```

Probando diferencias

Podemos usar una prueba de **t** para dos muestras independientes. Pero primero tenemos que ver si las varianzas son iguales usando el test de

```
leveneTest(Grasa ~ Clase, data = ovejas)
```

Ejercicio 2

¿Se cumple con la homogeneidad de varianzas?

Ejercicio 3

¿Cuál es la hipótesis nula?¿Y la alternativa?

Hacemos el test usando una formula de la forma: y ~ x. Completen el argumento `alternative` usando `"two.sided"` si su hipótesis nula es dos colas, `"less"` si es a cola izquierda y `"greater"` si es a cola derecha. Además, completen el argumento `var.equal` con `TRUE` si las varianzas son iguales y con `FALSE` si son diferentes

```
t.test(Grasa ~ Clase, data = ovejas, alternative =  , var.equal = T)
```

Analicemos la salida de esta función por un momento. Primero, no indica el nombre del test que se realizo. Luego sigue un bloque donde nos dice que datos se usó; Fatness by Kind (grasa por tipo (sexo)). Y nos da el valor del estadístico t seguido por los grados de libertad (`gl`) y la probabildad de un valor `t` mayor (o menor según la hipótesis que estemos usando). Luego nos indica que cola estamos usando. Seguido por los intervalos de confianza (inferior y superior) Finalmente, nos da los estimadores de las medias para los dos grupos.

Ejercicio 4

¿Encontraron diferencias significativas?

Dado que tenemos los pesos de las carcasas y probablemente los animales más grandes tengan más grasa podemos usar estos datos como covariable. Lo primero que tenemos que hacer es graficar estas dos variables.

```
ggplot(ovejas, aes(Peso_carcasa, Grasa, colour = Clase)) +
  geom_point()
```

Ejercicio 4

¿Qué relación entre las variables logra distinguir en base al gráfico?
¿Qué se pretende lograr al usar el peso como covariable?

Como tenemos dos variables independientes, una cualitiva y otra continua, usamos un análisis de covarianza. La forma de usarlo es similar, una formula para describir la relación entre los datos, inicialmente no sabemos si las pendientes son las mismas para ambos sexos (paralelas). Por lo que ajustamos un modelo con pendientes independientes

```
ovejas.fm1 <- lm(Grasa ~ Clase + Peso_carcasa + Clase:Peso_carcasa, data = ovejas)
```

Analisemos lo que hemos hecho arriba. Primero, llamamos a la función `lm` que ajusta modelos de regresión lineales. La formula que pusimos tiene varios componentes, veamos que significa cada uno de ellos. Primero, la variable de respuesta `Grasa`antes del tilde (`~`), luego de él van las variables explicatorias. Al escribir `Clase` indicamos que la ordenada al origen depende de esa variable ¿Cómo sabe R que es la ordenada al origen y no la pendiente? R pregunta que tipo de variable es; si es categórica es un cambio de ordenada, pero si es numérica (discreta o continua) es una pendiente. Luego, agregamos `Peso_carcasa`, que indica que la variable dependiente cambia linealmente con esa variable ya es numérica. Por lo tanto R, estimará una pendiente. Finalmente, `Clase:Peso_carcasa` indica que tenemos una interacción entre dos variables. Dependiendo del tipo de variables indicará algo distinto la interacción. Para dos variables categóricas, indica un cambio en la media de la combinación de esas variables; para dos variables continuas indica un cambio de pendiente para la multiplicación de ambas; y para una variable categórica y una continua indica un cambio de pendiente según el nivel de la variable categórica. Como la interacción es del último tipo, entonces es un cambio de pendiente de `Peso_carcasa` según la clase. Por último, `data` indica en que objeto están los datos.

Como asignamos el objeto que crea la función `lm` a `ovejas.fm1` no vemos ningún resultado. Pedimos el resumen de nuestro modelo con la función `summary()`.

```
summary(ovejas.fm1)
```

La interpretación de los resultados es similar al caso de una regresión lineal simple. Primero, la llamada que usamos para modelar la regresión. Luego, un resumen de los residuales (mínimo, primer cuartil, mediana, tercer cuartil y máximo). Luego la tabla de coeficientes. Aquí tenemos varias cosas, `(Intercept)` la ordenada al origen, el efecto de `ClaseOveja`, la pendiente de `Peso_carcasa` y el cambio de pendiente en `Peso_carcasa` por `ClaseOveja`.

Detengamonos un poco, y analicemos estos resultados. Para coeficiente estimado, da el error estándar, el estadístico t y la probabilidad de un valor absoluto de t mayor. El estimador de la ordenada al origen es -7.9179 pero no es significativamente diferente de 0 con $\alpha = 0.05$. Pero ¿qué significa esta ordenada al origen? ¿Es de los carneros o las ovejas? Recordemos que, por defecto, R ordena los niveles de los factores alfabeticamente. Por lo tanto, `(Intercept)` corresponde a la ordenada al origen de los carneros. ¿Y que representa `ClaseOveja`? No es la ordenda al origen de las ovejas, sino que es la *diferencia* entre la ordenada al origen base, en este caso los carneros, y la de ovejas: $e_o = \mu_c - \mu_o$. ¿Cómo calculamos la ordenada la origen de ovejas?

$$\mu_o = \mu_c + e_o$$

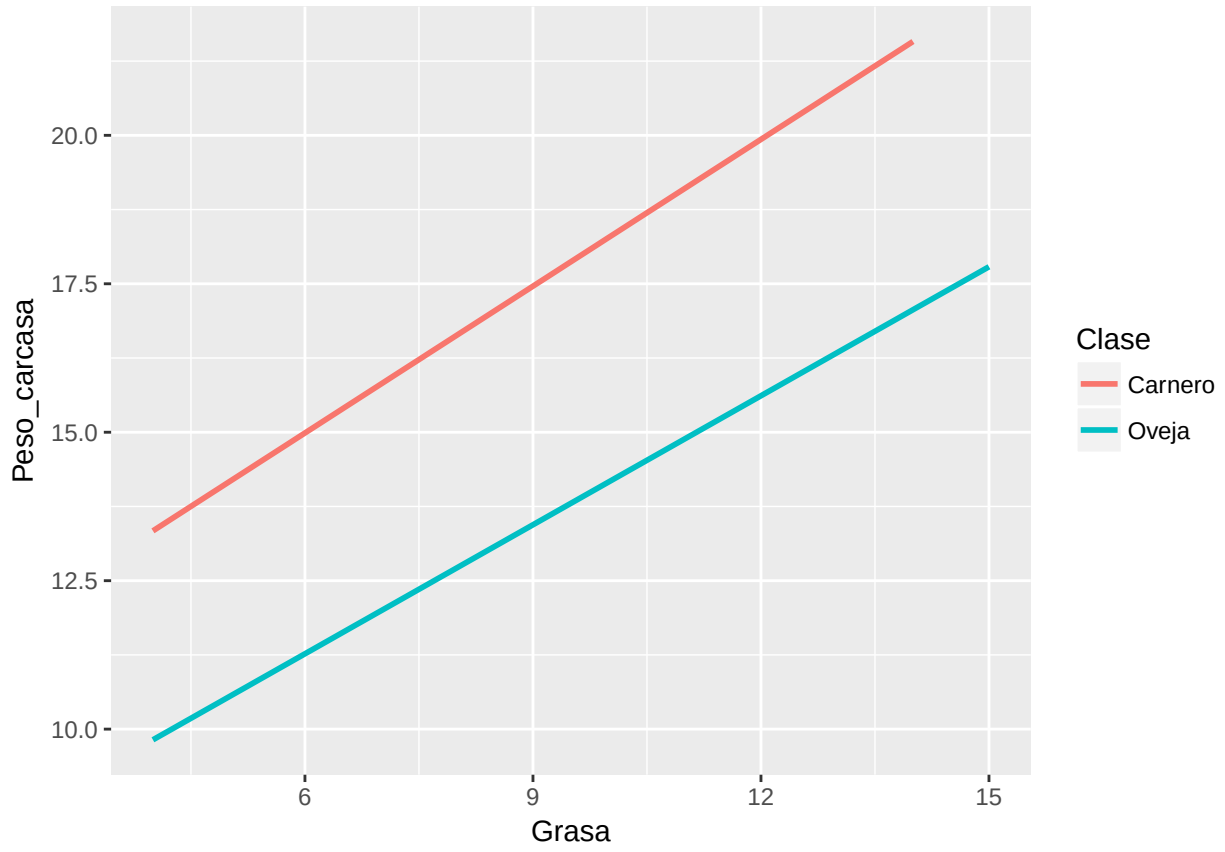
$$\mu_o = -7.9178584 + -0.7303048 = -8.6481633$$

La prueba estadística que se realiza sobre el efecto es para ver si este es significativamente diferente de 0. No rechazar H_0 no implica necesariamente que la ordenada al origen de ese nivel no sea diferente de 0, sino que el efecto (o sea la diferencia) no es significativamente diferente de 0. Por lo tanto, no es significativamente diferente del nivel de base, en este caso los carneros.

De igual forma, el estimador de `Peso_carcasa` es el estimador de la pendiente por cada kg de peso de carcasa para los carneros, que en este caso es significativamente diferente de 0. Como mencionamos arriba, `ClaseOveja:Peso_carcasa` es el *efecto* de oveja sobre la pendiente base. Para obtener la pendiente oveja debemos sumar el efecto a la base: $\beta_o = 0.9593963 + 0.358971 = 1.3183673$ Como en el caso anterior, no rechazar H_0 para el efecto implica que no es significativamente diferente del nivel base, *¡no que es 0!*.

Graficamente, nuestro modelo se ve así:

```
ggplot(ovejas, aes(Grasa, Peso_carcasa, colour = Clase)) +
  stat_smooth(method = "lm", se=FALSE)
```



Podemos pedir la matriz de diseño con la función `model.matrix()` con el objeto del modelo ajustado.

```
model.matrix(ovejas.fm1)
```

Ejercicio 5

Interprete la salida de model.matrix. ¿Qué significa cada columna? ¿Cómo lo usaría con la tabla de coeficientes para obtener el valor esperado para cada **Y**?

Probando el paralelismo

Para realizar un análisis de ANCOVA primero debemos probar que las pendientes sean paralelas. Como son solo dos niveles podemos verlo simplemente mirando la tabla de resumen y ver que el efecto de `ClaseOveja` sobre la pendiente no es significativo. Pero si tenemos más niveles esto no es aconsejable porque estaremos aumentando el error Tipo I. Por eso podemos pedir una tabla de ANOVA para nuestro modelo. La forma de hacerlo es con la función `anova()`:

```
anova(ovejas.fm1)
```

Ejercicio 6

En base a la tabla de ANOVA ¿Es significativa la interacción?

Si las rectas son paralelas, debemos ajustar un nuevo modelo con una sola pendiente en común. Podemos realizarlo actualizando el modelo viejo, con la función `update()`, la sintaxis es objeto del modelo viejo y luego lo que queremos quitar (o agregar).

```
ovejas.fm2 <- update(ovejas.fm1, . ~ . -Clase:Peso_carcasa)
```

Lo que queremos quitar o agregar tiene que ser expresado como una formula y todas las formulas contienen en tilde (`~`). Para quitar una variable se usa el signo menos (`-`) y para agregar el signo más (`+`). Tiene además puntos, estos puntos indican que queremos usar toda la formula anterior.

Volvamos a ver si hay diferencias entre las clases con la función `anova()`

```
anova(ovejas.fm2)
```

Ejercicio 7

¿Hay evidencias de diferencias significativas entre el espesor de la grasa entre ovejas y carneros luego de eliminar el efecto del peso? ¿Cómo se compara con el modelo que no tiene en cuenta el efecto del peso?

Mejorando el modelo

Nuestro modelo indica que no hay diferencias significativas entre ovejas y carneros. Si bien son pocos datos, podemos intentar sacarles un poco más de jugo a nuestros datos. La varianza del estimador aumenta a medida que se aleja de la media de X ya que sigue esta formula:

$$\hat{S}_\alpha = \hat{S}_r \sqrt{\frac{1}{n}(1 + \frac{\bar{x}^2}{n s_x})}$$

Por lo tanto, aumenta a medida que la media de x aumenta. Una forma de eliminar este efecto es centrando x es decir restar la media de x a todos los valores. Probemos esto:

```
ovejas <- transform(ovejas,
                    Peso_centrado = Peso_carcasa - mean(Peso_carcasa))
```

Aquí usamos la función `transform()`, esta función toma una `data.frame` y crea o modifica (si la el nombre de la variable ya está en `data.frame`) según la formula que especifiquemos. En este caso, a cada de valor de peso de carcasa le restamos la media.

Volvamos a ajustar una nueva regresión lineal con la variable que recién creamos:

```
ovejas.fm3 <- lm(Grasa ~ Clase + Peso_centrado, data = ovejas)
```

Y veamos que pasó:

```
summary(ovejas.fm3)
```

Ejercicio 8

En base al resumen del modelo ¿Les parece que hay diferencias significativas en el espesor de la grasa entre las ovejas y carneros? Compruebelo con la función anova()

Por su cuenta

El frigorífico comisionó un nuevo experimento. Ahora quieren ver si los machos castrados (borregos) presentan el mismo contenido de grasa que las ovejas y los carneros. Los animales pastorearon las misma pastura durante un tiempo y luego fueron faenados. Para descargar los datos utilizen este comando

```
read.table(url("https://git.io/ovejas2.txt"), header = TRUE)
```

Analicen los datos y presenten:

1. Un gráfico de dispersión
2. Prueben el paralelismo
3. Prueben si hay diferencias significativas entre las clases de animales.
4. Presenten un gráfico de dispersión más la rectas adecuadas.
5. ¿Qué le recomendarían al frigorífico?