Técnicas de Inferencia Estadística II

Primera prueba parcial (A)

Alumno: ____

Grupo: _________ Fecha: ______________

Ejercicio 1. El fichero de datos `pardals.csv` corresponde a un estudio de variables biométricas realizado por Cuadras y colaboradores. El fichero contiene medidas de 49 aves recogidas casi moribundas después de un temporal. Las variables son:

Nombre	Descripción
TT	tamaño total
LA	longitud del ala
LPC	longitud del pico a cabeza
LH	longitud del húmero
LQE	longitud de la quilla del esternón
SB	(= 1) si sobrevive (= 0) si no sobrevive

La siguiente salida de R muestra la lectura y las cinco primeras observaciones:

```
> Datos <- read.table("pardals.csv",header=TRUE,dec=",",sep=";")
> Datos
    TT  LA  LPC   LH  LQE SB
1  156 245 31.6 18.5 20.5  1
2  154 240 30.4 17.9 19.6  1
3  153 240 31.0 18.4 20.6  1
4  153 236 30.9 17.7 20.2  1
5  155 243 31.5 18.6 20.3  1
```

a) Realice un contraste de igualdad de medias del tamaño total, `TT`, entre las aves que sobreviven y las que no. Plantee las hipótesis nula y alternativa, así como los supuestos necesarios sobre las variables para la validez del contraste. Utilice α=0.05. Calcule e interprete el p-valor. (2 puntos)

El contraste pedido es $\begin{matrix} H_0: \mu_{NS} = \mu_S \\ H_1: \mu_{NS} \neq \mu_S \end{matrix}$ donde μ_{NS} es la media del tamaño total en aves que no sobreviven y μ_S en las que sobreviven.

Los supuestos necesarios para realizar este contraste son: (i) independencia de las muestras y observaciones; (ii) los tamaños totales de las aves se distribuyen normales. Notar que el número de aves total es 49, pero en cada grupo hay menos de 30 aves.

El contraste en R sería:

```
>
t.test(Datos$TT[Datos$SB==0],Datos$TT[Datos$SB==1],alternative="two.sided"
)
```

```
        Welch Two Sample t-test

data:  Datos$TT[Datos$SB == 0] and Datos$TT[Datos$SB == 1]
t = 1.0155, df = 46.108, p-value = 0.3152
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.028801  3.124040
sample estimates:
mean of x mean of y 
 158.4286  157.3810 
```

Por tanto, el p-valor = 0.3152 > α=0.05 y no rechazaríamos la hipótesis nula, esto es, no hay evidencia suficiente en contra de la igualdad de medias de los tamaños en los dos grupos de aves.

b) Con estas medidas se puede obtener el siguiente indicador de sobrevivencia:

```
beta <- c(3.6414,-0.0374,-0.0064,-0.0224,0.2487,0.0167)

IS <- beta[1] + beta[2]*Datos$TT + beta[3]*Datos$LA +
      beta[4]*Datos$LPC + beta[5]*Datos$LH + beta[6]*Datos$LQE
```

Realice un contraste de que la media del indicador de sobrevivencia es menor en las aves que no sobreviven respecto a las que sobreviven. Plantee las hipótesis nula y alternativa, así como los supuestos necesarios sobre las variables para la validez del contraste. Utilice α=0.05. Calcule e interprete el p-valor. (2 puntos)

El contraste pedido es $\begin{matrix} H_0: \mu_{NS} = \mu_S \\ H_1: \mu_{NS} < \mu_S \end{matrix}$ donde μ_{NS} es la media del indicador de sobrevivencia en aves que no sobreviven y μ_S en las que sobreviven.

Los supuestos necesarios para realizar este contraste, nuevamente son: (i) independencia de las muestras y observaciones; (ii) los tamaños totales de las aves se distribuyen normales. Notar que el número de aves total es 49, pero en cada grupo hay menos de 30 aves.

El contraste en R sería:

```
> t.test(IS[Datos$SB==0],IS[Datos$SB==1],alternative="less")

        Welch Two Sample t-test

data:  IS[Datos$SB == 0] and IS[Datos$SB == 1]
t = -1.7124, df = 45.761, p-value = 0.0468
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
         -Inf -0.001110567
sample estimates:
mean of x mean of y 
0.4006371 0.4573148 
```

Por tanto, el p-valor = 0.0448 < α=0.05 y rechazaríamos la hipótesis nula, esto es, aceptaríamos que la media del indicador de sobrevivencia es menor en aves que no sobreviven.

Ejercicio 2. Se desea contrastar si la media de los tiempos de ejecución de una tarea disminuye con la experiencia del trabajador/a. Para ello se toma una muestra al azar de n=20 trabajadores/as recién contratados/as se miden sus tiempos y tras un año de contrato se vuelven a medir los tiempos. El contraste a realizar es:

$$\begin{matrix} H_0: \mu_N = \mu_E \\ H_1: \mu_N > \mu_E' \end{matrix}$$

donde μ_N es la media de los tiempos en trabajadores/as noveles y μ_E en trabajadores/as experimentado/as. Se supone que los tiempos de ejecución se distribuyen como una normal y que las varianzas son iguales en ambos momentos de la medición.

a) Calcule el tamaño de la región de rechazo definida por $R = \left\{ \frac{\bar{X}-\bar{Y}}{\sqrt{\frac{S_D^2}{n}}} > 0 \right\}$, donde $\bar{X}$ es la media en la muestra de trabajadores/as al inicio del contrato, $\bar{Y}$ en trabajadores/as tras un año de experiencia y S_D^2 es la cuasi-varianza de la diferencia entre los tiempos al principio y tras un año de experiencia, es decir, $D = X - Y$. (2 puntos)

Debemos darnos cuenta que la muestra es apareada porque al mismo trabajador se le mide en dos momentos distintos. Por tanto, utilizaremos $D = X - Y$, y tenemos que $\bar{D} = \bar{X} - \bar{Y}$. Así que la región $R = \left\{ \frac{\bar{D}}{\sqrt{\frac{S_D^2}{n}}} > 0 \right\}$ y sabemos que $\frac{\bar{D}}{\sqrt{\frac{S_D^2}{n}}}$ bajo la hipótesis nula se distribuye como una $t_{n-1} = t_{19}$, por tanto,

$Pr\left\{ \frac{\bar{D}}{\sqrt{\frac{S_D^2}{n}}} > 0 \middle| H_0 \right\} = Pr\{t_{n-1} > 0\} = 0.5$ puesto que la distribución t es simétrica respecto del cero.

b) Proponga una región de rechazo para este contraste que tenga un tamaño igual a α. (1 punto)

Utilizamos como punto de partida lo que conocemos del apartado anterior, es decir que $\frac{\bar{D}}{\sqrt{\frac{S_D^2}{n}}}$ bajo la hipótesis nula se distribuye como una t_{n-1}. Por tanto, una región de tamaño α sería $R_b = \left\{ \frac{\bar{D}}{\sqrt{\frac{S_D^2}{n}}} > t_{n-1,\alpha} \right\}$.

c) ¿Cuál de las dos regiones tiene una mayor potencia? Justifique su respuesta. (1 punto)

La región R tiene mayor potencia que R_b porque ambas rechazan en la misma dirección, es decir, el estadístico de contraste mayor que una cierta cantidad y si en el apartado b) usamos valores "razonables" de α tendríamos que el tamaño de la región R es mayor que el tamaño de la región R_b. Sabemos que a mayor tamaño o probabilidad de error de tipo I tenemos menor β o probabilidad de error de tipo II y, por tanto, mayor potencia, 1- β.

Ejercicio 3. En el archivo <Portfolios_Formed_on_ME.csv> tenemos los rendimientos mensuales de cuatro portafolios construidos a fines de cada mes de junio utilizando el capital de mercado de junio y los puntos de ruptura de NYSE. Las carteras para julio del año t a junio de t + 1 incluyen todas las acciones de NYSE, AMEX y NASDAQ para las cuales se tienen datos de capital de mercado para junio de t. Los datos están tomados de:

http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

La siguiente salida de R muestra la lectura y las cinco primeras observaciones:

```
> Portafolios <- read.table("Portfolios_Formed_on_ME.csv",sep=";")
> Portafolios
```

```
      V1     V2   V3   V4  V5
1  200001 4213 3664 590 501
2  200002 4161 3618 585 497
3  200003 4105 3566 581 492
4  200004 4061 3532 575 489
5  200005 4023 3496 570 487
```

donde `V1` corresponde a la fecha en formato YYYYMM y `V2` a `V5` son los rendimientos de cuatro portafolios seleccionados.

a) Contraste si la varianza del portafolio `V4` es inferior a la varianza del portafolio `V5`. Plantee las hipótesis nula y alternativa, así como los supuestos necesarios sobre las variables para la validez del contraste. Utilice α=0.05. Calcule e interprete el p-valor. (2 puntos)

El contraste pedido es $\begin{matrix} H_0: \sigma_4^2 = \sigma_5^2 \\ H_1: \sigma_4^2 < \sigma_5^2 \end{matrix}$ donde σ_4^2 es la varianza del portafolio V4 y σ_5^2 es la varianza del portafolio V5.

Los supuestos necesarios para realizar este contraste: (i) independencia de las muestras y observaciones; (ii) los rendimientos deben distribuirse normales. Notar que en este ejercicio estamos contrastando varianzas y por tanto el supuesto de normalidad no puede eludirse aunque los tamaños muestrales sean grandes.

El contraste en R sería:

```
> var.test(Portafolios$V4,Portafolios$V5,alternative="less")

        F test to compare two variances

data:  Portafolios$V4 and Portafolios$V5
F = 2.5801, num df = 227, denom df = 227, p-value = 1
alternative hypothesis: true ratio of variances is less than 1
95 percent confidence interval:
 0.000000 3.211173
sample estimates:
ratio of variances
          2.580112
```

Por tanto, el p-valor = 1 > α=0.05 y no rechazaríamos la hipótesis nula, esto es, no hay evidencia suficiente en contra de la igualdad de las varianzas.