Inferencia Estadística — 12 de mayo de 2025 — parcial 2

Instrucciones.

- Primera parte: TEORÍA. Ordenador apagado. Tienes hasta una hora para resolverla y entregar los folios con tu respuesta.
- Segunda parte: PROBLEMAS. En cuanto entregues la teoría, podrás encender el ordenador exclusivamente para usar R y acceder al campus virtual. No hay una "tarea" en el campus virtual para subir documentos. Has de entregar todas tus respuestas en papel: para un gráfico, esbózalo a mano dejando claro lo que quieras destacar; para código fuente, escríbelo con la mayor precisión posible. Has de entregar los folios en dos bloques separados: problema 1; problema 2.
- Justifica todas las respuestas.

Teoría (4 puntos).

1. (1 punto) A una I.A. se le planteó el siguiente problema:

   ```
   Para comparar la eficacia de tres medicamentos se consideró una muestra
   de 21 personas, cada una de las cuales fue asignada aleatoriamente a uno
   de los grupos. Posteriormente se estudió el tiempo que tardaba en hacer
   efecto con los siguientes resultados:

   x1 = c(3.88, 6.00, 4.15, 5.48, 5.13, 4.77, 5.08)

   x2 = c(1.54, 6.52, 2.22, 20.51, 0.63, 0.37, 4.47)

   x3 = c(7.60, 5.51, 8.62, 4.34, 6.62, 8.12, 7.64)

   ¿Puedes describir los resultados de cada grupo y analizar si existen
   diferencias significativas en el tiempo que tarda en actuar cada
   medicamento?
   ```

   La respuesta dada por esa I.A. a este problema fue la siguiente:

   ```
   > ## Solución I.A.
   > # Datos
   > x1 <- c(3.88, 6.00, 4.15,  5.48, 5.13, 4.77, 5.08)
   > x2 <- c(1.54, 6.52, 2.22, 20.51, 0.63, 0.37, 4.47)
   > x3 <- c(7.60, 5.51, 8.62,  4.34, 6.62, 8.12, 7.64)
   > # Crear un data frame
   > grupo <- factor(c(rep("Medicamento1", length(x1)),
   +                   rep("Medicamento2", length(x2)),
   +                   rep("Medicamento3", length(x3))))
   > tiempo <- c(x1, x2, x3)
   > datos <- data.frame(Grupo=grupo, Tiempo=tiempo)
   > # Resumen descriptivo usando numSummary
   > numSummary(datos$Tiempo, groups=datos$Grupo)
                    mean        sd   IQR   0%   25%  50%   75%  100% data:n
   Medicamento1 4.927143 0.7355659 0.845 3.88 4.460 5.08 5.305  6.00      7
   Medicamento2 5.180000 7.1081925 4.410 0.37 1.085 2.22 5.495 20.51      7
   Medicamento3 6.921429 1.5299938 1.815 4.34 6.065 7.60 7.880  8.62      7
   > # Análisis de varianza
   > anova_result <- aov(Tiempo ~ Grupo, data=datos)
   > summary(anova_result)
   ```

```
            Df Sum Sq Mean Sq F value Pr(>F)
Grupo        2   16.5   8.253   0.464  0.636
Residuals   18  320.5  17.803
```

Estudia si la solución propuesta es correcta o comenta los inconvenientes que pueda tener.

2. (1 punto) Define la potencia de un test de hipótesis y explica su significado bajo $\boldsymbol{H_0}$ y bajo $\boldsymbol{H_1}$.
3. (1 punto) Explica el test de los signos para contrastar las hipótesis $\boldsymbol{H_0}$: mediana $= \boldsymbol{M_0}$, $\boldsymbol{H_1}$: mediana $> \boldsymbol{M_0}$, a nivel de significación $\boldsymbol{\alpha}$.
4. (1 punto) Desarrolla el test de la razón de verosimilitudes para bondad de ajuste de una variable aleatoria que toma $\boldsymbol{k}$ valores posibles y tiene todos los parámetros completamente especificados.

Problema 1 (3 puntos). El número de ciberataques diarios a cierta central mareomotriz sigue una distribución binomial $\boldsymbol{B(m, 1/2)}$. En una muestra aleatoria simple de 30 días se han obtenido los siguientes valores: **(10, 10, 11, 14, 9, 14, 15, 12, 12, 7, 9, 9, 12, 10, 13, 11, 12, 17, 10, 13, 15, 9, 12, 8, 10, 10, 6, 10, 14, 10)**.

a) (1 punto) Halla una estimación de $\boldsymbol{m}$ por el método de los momentos y por máxima verosimilitud.

b) (1 punto) Contrasta las hipótesis $\boldsymbol{H_0}$: $\boldsymbol{m = 20}$, $\boldsymbol{H_1}$: $\boldsymbol{m = 22}$ con nivel de significación $\boldsymbol{\alpha = 0{,}05}$.

c) (1 punto) Comprueba si puede considerarse que la muestra del enunciado proviene de una población con distribución de Poisson.

Ten en cuenta que, si $\boldsymbol{X}$ sigue una distribución $\boldsymbol{B(m, p)}$, entonces tiene función masa de probabilidad $\boldsymbol{f(x) = \binom{m}{x} p^x (1-p)^{m-x}}$, donde $\boldsymbol{\binom{m}{x} = \frac{m!}{x!(m-x)!}}$; esperanza $\boldsymbol{E(X) = mp}$ y varianza $\boldsymbol{Var(X) = mp(1-p)}$.
Recuerda que, si no encuentras una expresión exacta en algún apartado, puedes emplear métodos numéricos, como el de Montecarlo.

a) momentos: $\boldsymbol{\hat{m} = 2\bar{x} = 22{,}26667}$

```
> x <- c(10, 10, 11, 14, 9, 14, 15, 12, 12, 7, 9, 9, 12,
  10, 13, 11, 12, 17, 10, 13, 15, 9, 12, 8, 10, 10,
  6, 10, 14, 10)
> 2*mean(x)
[1] 22.26667
```

e.m.v.: $\boldsymbol{\hat{m} = 22}$

```
> L <- function(m) prod(dbinom(x,m,1/2))
> v <- 0:100000 ; v[which.max(sapply(v,L))]
[1] 22
```

b) se trata de dos hipótesis simples, por lo que el contraste más potente se consigue con el método de Neyman y Pearson:

$$\begin{aligned} \boldsymbol{L(m)} &= \prod_{i=1}^{n} \left[\frac{\boldsymbol{m!}}{\boldsymbol{x_i!\,(m-x_i)!}} \times \frac{\boldsymbol{1}}{\boldsymbol{2^m}} \right] = \frac{\boldsymbol{m!^n}}{\boldsymbol{2^{mn} \prod_{i=1}^{n} x_i!\,(m-x_i)!}} \\ \boldsymbol{\Lambda} &= \frac{\boldsymbol{L(22)}}{\boldsymbol{L(20)}} = \boldsymbol{(22\times 21)^n} \prod_{i=1}^{n} \frac{\boldsymbol{(20-x_i)!}}{\boldsymbol{(22-x_i)!}} = \frac{\boldsymbol{(22\times 21)^n}}{\boldsymbol{\prod_{i=1}^{n}(22-x_i)(21-x_i)}} \end{aligned}$$

la distribución de $\boldsymbol{\Lambda}$ bajo $\boldsymbol{H_0}$ puede obtenerse por Montecarlo:

```
n <- length(x) # 30
Landa <- function (x) (22*21)^n / prod ((22-x) * (21-x))
dis <- replicate(1000000, Landa(rbinom(n,20,1/2)))
```

se rechaza H_0 para valores grandes de Λ, por lo que el p-valor puede hallarse así:

```
mean (dis >= Landa(x)) # 0.001651
```

y la región crítica puede hallarse así:

```
quantile (dis, .95) # 3.077231e+18
Landa (x)           # 7.566794e+19 en la R.C.
```

si se quiere evitar números tan grandes se pueden usar logaritmos, por ejemplo así:

```
logL <- function(x,m) sum(dbinom(x,m,1/2,log=TRUE))
Landa <- function (x) logL(x,22) - logL(x,20)
```

c) se estima el parámetro λ mediante el estimador máximo-verosímil:

```
landa <- mean(x) # 11.13333
```

frecuencias esperadas:

```
n * c(ppois(6,landa), dpois(7:16,landa), 1-ppois(16,landa))
# 2.198958 1.844684 2.567185 3.175703 3.535616 3.578472
# 3.320027 2.843305 2.261105 1.678242 1.167777 1.828926
```

juntamos para obtenerlas mayores que 5 en al menos[1] el 80 %; por ejemplo,

```
n * (p <- c(ppois(6,landa)+sum(dpois(7:8,landa)),
            sum(dpois(9:10,landa)), sum(dpois(11:12,landa)),
            sum(dpois(13:14,landa)), 1-ppois(14,landa)))
# 6.610827 6.711319 6.898499 5.104410 4.674945
```

agrupamos de igual manera las observadas y aplicamos la prueba χ^2 de Pearson de bondad de ajuste con un parámetro estimado:

```
obs <- table (x)
# valor 6  7  8  9 10 11 12 13 14 15 17
# frec. 1  1  1  4  8  2  5  2  3  2  1
obs <- c(sum(obs[c("6","7","8")]), sum(obs[c("9","10")]),
         sum(obs[c("11","12")]), sum(obs[c("13","14")]),
         sum(obs[c("15","17")])) # 3 12  7  5  3
1 - pchisq (chisq.test (obs, p=p) $ statistic, 5 - 1 - 1)
# p-valor = 0.08053617
```

luego no puede rechazarse que provenga de una población Poisson

d) el método de los momentos da números reales y no enteros, pero fundamentalmente da estimaciones imposibles si $\hat{m} < \max x_i \iff \bar{x} < \max x_i/2$

e) intervalo de confianza mediante *t-bootstrap* paramétrico:

```
> n <- length(x)
> m <- v[which.max(sapply(v,L))] # 22
> dis <- replicate(10000,{xb<-rbinom(n,m,1/2)
          L <- function (m) prod(dbinom(xb,m,1/2))
```

[1] Página 178 de https://doi.org/10.1080/00031305.2017.1286260

```
          ((mMV<-v[which.max(sapply(v,L))]) - m) / sqrt(mMV/4)})
> m - sqrt(m/4)*quantile(dis,c(.975,.025))
   97.5%     2.5%
20.08515 24.09762
```

intervalo de confianza mediante *bootstrap* básico paramétrico:

```
> n <- length(x)
> m <- v[which.max(sapply(v,L))]
> dis <- replicate(10000, {xb<-rbinom(n,m,1/2)
  L<-function(m)prod(dbinom(xb,m,1/2));v[which.max(sapply(v,L))]-m})
> m - quantile(dis,c(.975,.025))
97.5%  2.5%
   20    24
```

asintótico: $\boldsymbol{X \hookrightarrow B(m,1/2) \approx N(m/2,\sqrt{m}/2) \Longrightarrow \hat{m} = 2\bar{X} \overset{\sim}{\hookrightarrow} N\left(m,\sqrt{\hat{m}/n}\right)}$

```
> m + c(-1,1) * qnorm(.975) * sqrt(m/n)
[1] 20.32159 23.67841
```

Problema 2 (3 puntos). La empresa Blackout Consulting quiere estudiar el consumo energético en la ciudad de Villaluz. Para ello, realiza un estudio del consumo energético (medido en kWh) durante los cuatro lunes del mes de abril de 2025, tomando una muestra aleatoria simple de 100 viviendas escogidas al azar cada uno de los días. Los datos obtenidos se encuentran en el fichero 'BlackoutConsulting.RData'. Basándote en los datos recogidos y asumiendo independencia en las distintas mediciones, responde a las siguientes preguntas a nivel de significación $\boldsymbol{\alpha = 0{,}01}$:

a) (1 punto) Estudia mediante un gráfico y un test de hipótesis adecuados si la distribución del consumo energético cada día estudiado sigue una distribución normal.

b) (1 punto) ¿Hay diferencias en el consumo energético los distintos días estudiados? Realiza un estudio descriptivo previo (tanto numérico como gráfico), y basa tus conclusiones en los resultados de un contraste de hipótesis.

c) (1 punto) Considerando el tamaño muestral homogéneo e igual a 100 en cada unos de los cuatro grupos y asumiendo la distribución exponencial en cada grupo, estima mediante simulación por el método de Monte Carlo la potencia del test utilizado en el apartado anterior en los dos casos siguientes: (1) las medias en los cuatro grupos son iguales a 12kWh; (2) las medias en tres de los cuatro grupos son iguales a 12kWh y en el cuarto grupo es igual a 8kWh. Interpreta los resultados obtenidos.