Inferencia Estadística

Segundo Parcial

15 de mayo de 2023

CUESTIONES

(1 punto) **Cuestión 1**

Enuncia y demuestra el lema de Neyman-Pearson para contrastes no aleatorizados. Sea X una variable aleatoria con distribución Multinomial$(m, \mathbf{p})$, siendo el parámetro m conocido, y $\mathbf{p} = (p_1, p_2, p_3)$ con $p_1+p_2+p_3 = 1$ y $p_i > 0$ para $i = 1, 2, 3$, es decir, X toma valores enteros (m_1, m_2, m_3) con $0 \leqslant m_i \leqslant m$ para $i = 1, 2, 3$ y

$$\Pr[X = (m_1, m_2, m_3)] \quad = \quad \frac{m!}{m_1!\, m_2!\, m_3!}\, p_1^{m_1}\, p_2^{m_2}\, p_3^{m_3}$$

De X se obtiene una muestra aleatoria simple de tamaño n para contrastar $H_0\colon \mathbf{p} = \mathbf{p}_0$, frente a $H_1\colon \mathbf{p} = \mathbf{p}_1$. Aplica el método de Neyman-Pearson para calcular la forma de la región crítica.

(1 punto) **Cuestión 2**

Define la función potencia y explica su significado bajo H_1.

(1 punto) **Cuestión 3**

Explica el test de los rangos con signo de Wilcoxon. Sea una variable continua X simétrica respecto a m_0. Si se define $D = X - m_0$, demuestra que $|D|$ y signo(D) son independientes.

(1 punto) **Cuestión 4**

Calcula la región crítica para contrastar la independencia entre dos variables finitas, usando el test de la razón de verosimilitudes, y comenta cuál es la distribución asintótica bajo H_0.

PROBLEMAS

Problema 1

Se está ensayando un tratamiento nuevo contra la trombocitopenia. Dicho tratamiento consta de dos fases. El tratamiento se aplica a veinte pacientes y se obtienen los siguientes resultados en los recuentos plaquetarios ($\times 10^9$/litro o bien $\times 10^3$/mm^3):

Paciente	1	2	3	4	5	6	7	8	9	10
Recuento basal	88,4	120,4	99,2	96,4	110,1	110,2	86,2	94,7	108,9	111,5
Tras fase 1	111,4	100,8	106,4	104,7	111,8	139,6	130,0	145,5	128,0	137,8
Tras fase 2	104,1	129,3	98,7	118,5	111,0	103,4	92,4	95,2	105,7	130,0

Paciente	11	12	13	14	15	16	17	18	19	20
Recuento basal	105,6	105,3	106,7	107,9	123,4	121,5	85,6	115,2	124,9	94,0
Tras fase 1	119,0	118,7	108,5	122,7	112,9	116,8	144,5	110,1	129,0	110,4
Tras fase 2	101,5	100,6	133,9	139,7	132,2	135,5	113,6	101,2	96,4	104,0

Versión para copiar:

```
c(88.4, 120.4, 99.2, 96.4, 110.1, 110.2, 86.2, 94.7, 108.9, 111.5, 105.6, 105.3, 106.7, 107.9, 123.4, 121.5, 85.6, 115.2, 124.9, 94.0)
c(111.4, 100.8, 106.4, 104.7, 111.8, 139.6, 130, 145.5, 128, 137.8, 119, 118.7, 108.5, 122.7, 112.9, 116.8, 144.5, 110.1, 129.0, 110.4)
c(104.1, 129.3, 98.7, 118.5, 111, 103.4, 92.4, 95.2, 105.7, 130, 101.5, 100.6, 133.9, 139.7, 132.2, 135.5, 113.6, 101.2, 96.4, 104.0)
```

Para las siguientes preguntas, considera un nivel de significación $\alpha = 0{,}05$.

(1 punto) a) ¿Podría considerarse que los recuentos basales siguen una distribución exponencial desplazada, con densidad $f(x) = \lambda e^{-\lambda(x-\theta)} \cdot I(x \geqslant \theta)$?

Al ser la $\mathsf{Exp}(\lambda;\theta)$ una distribución continua y H_0 compuesta, intentaremos primero usar Kolmogórov+Smirnov+Lílifors (si no, habría que recurrir a una prueba χ^2). Para ello, habrá que ver si el estadígrafo KSL es de libre distribución, lo que es esperable ya que la familia exponencial desplazada depende sólo de un parámetro de localización y de otro de escala. Los estimadores MV de los parámetros son: $\hat\theta = \min X_i$, $\hat\lambda = n/\sum(X_i - \hat\theta)$. Sean

- $Y_i := \lambda(X_i - \theta) \hookrightarrow \mathsf{Exp}(1;0) = \mathsf{Exp}(1)$
- $\hat\theta_Y = \min Y_i = \lambda(\min X_i - \theta) = \lambda(\hat\theta - \theta)$.
- $\hat\lambda_Y = n/\sum(Y_i - \hat\theta_Y) = n/\sum[\lambda(X_i - \theta) - \lambda(\hat\theta - \theta)] = n/\lambda\sum(X_i - \hat\theta) = \hat\lambda/\lambda$.

Entonces:

$$\begin{aligned}
D_n &= \sup_{x\in\mathbb{R}} \left|\hat F(x) - F_{\hat\lambda,\hat\theta}(x)\right| \\
&= \sup_{x\geqslant\hat\theta} \left|\frac{\sum[X_i \leqslant x]}{n} - (1 - e^{\hat\lambda(x-\hat\theta)})\right| \\
&= \sup_{x\geqslant\hat\theta} \left|\frac{\sum[\lambda(X_i-\theta) \leqslant \lambda(x-\theta)]}{n} - (1 - e^{\hat\lambda(x-\hat\theta)})\right| \\
&= \sup_{\substack{x\geqslant\hat\theta \\ y:=\lambda(x-\theta)}} \left|\frac{\sum[Y_i \leqslant y]}{n} - (1 - e^{\frac{\hat\lambda}{\lambda}[\lambda(x-\theta)-\lambda(\hat\theta-\theta)]})\right| \\
&= \sup_{y\geqslant\hat\theta_Y} \left|\frac{\sum[Y_i \leqslant y]}{n} - (1 - e^{\hat\lambda_Y[y-\hat\theta_Y]})\right| \qquad \text{independiente de } (\lambda,\theta) \\
&= \sup_{y\in\mathbb{R}} \left|\hat F_Y(x) - F_{\hat\lambda_Y,\hat\theta_Y}(y)\right|
\end{aligned}$$

luego D_n tiene distribución libre que puede generarse a partir de muestras $\mathsf{Exp}(1)$. Una implementación en R básico podría ser:

```
x <- c(88.4, 120.4, 99.2, 96.4, 110.1, 110.2, 86.2, 94.7, 108.9, 111.5,
      105.6, 105.3, 106.7, 107.9, 123.4, 121.5, 85.6, 115.2, 124.9, 94)
n <- length(x)
ksl <- function (x) ks.test((x-min(x))/mean(x-min(x)),pexp)$statistic
t0 <- ksl(x)
d <- replicate (1000000, ksl(rexp(n)))
mean (d >= t0) # 0.0133
```

(1 punto) b) ¿Se produce alguna variación significativa entre las distribuciones de los recuentos?

En primer lugar hacemos una prueba ómnibus (global) que tenga en cuenta que se trata de muestras dependientes. La única que hemos visto es la de Friedman:

```
y <- c(111.4, 100.8, 106.4, 104.7, 111.8, 139.6, 130, 145.5, 128,
  137.8, 119, 118.7, 108.5, 122.7, 112.9, 116.8, 144.5, 110.1,
  129, 110.4)
z <- c(104.1, 129.3, 98.7, 118.5, 111, 103.4, 92.4, 95.2, 105.7, 130,
  101.5, 100.6, 133.9, 139.7, 132.2, 135.5, 113.6, 101.2, 96.4, 104)
friedman.test (cbind (x, y, z)) # p-value = 0.01057
```

Como hay diferencias significativas, realizamos un análisis dos a dos mediante Friedman y Wilcoxon pareado:

```
friedman.test (cbind (x, y))    $ p.value # 0.007290358 *
wilcox.test (x, y, paired=TRUE) $ p.value # 0.006066037 *
friedman.test (cbind (x, z))    $ p.value # 0.1797125
wilcox.test (x, z, paired=TRUE) $ p.value # 0.07014859
friedman.test (cbind (y, z))    $ p.value # 0.07363827
wilcox.test (y, z, paired=TRUE) $ p.value # 0.164957
```

Siguiendo el criterio de Bonferroni, al comparar con $\alpha/3 \approx 0{,}017$ hay diferencias significativas sólo entre el recuento basal y el recuento tras fase 1. Nótese que Friedman aplica corrección por empates, pero `wilcox.test` emite avisos, a los que no daremos importancia ya que todos los p-valores quedan del mismo lado del nivel de significación que los de Friedman. Otra forma:

```
pairwise.wilcox.test (c(x,y,z), rep(1:3,each=20), paired=TRUE)
```

Se aplica automáticamente la corrección de Holm y se muestra que sólo la primera pareja tiene un p-valor (corregido) menor que α.

Otra estrategia válida sería pensar en diferencias:

```
dxy <- x-y ; shapiro.test(dxy) # p-value = 0.5855 => gausiana
dxz <- x-z ; shapiro.test(dxz) # p-value = 0.8793 => gausiana
dyz <- y-z ; shapiro.test(dyz) # p-value = 0.4492 => gausiana
t.test(dxy, mu=0) # p-value = 0.004171
t.test(dxz, mu=0) # p-value = 0.06837
t.test(dyz, mu=0) # p-value = 0.1324
## en vez de t.test(dxy, mu=0) se puede hacer
## t.test(x, y, paired=TRUE) ... etcétera
```

Se llega a la misma conclusión que antes.

(1 punto) c) Para que el tratamiento se considere válido, tiene que conseguir que la mediana poblacional del recuento sea superior a 110. ¿Hay evidencias de que lo consigue tras alguna de las dos fases? (Por población se entiende el conjunto de pacientes del que se ha extraído la muestra.)

La variable es continua. Como no estoy seguro de su simetría (requerida por la prueba de rangos con signo de Wilcoxon), usaré la prueba de los signos, con

$$H_0 : \text{mediana} = 110 \qquad H_1 : \text{mediana} > 110$$

```
prueba.signos <- function (muestra)
   binom.test (sum(muestra>110), length(muestra), # < 110
               alternative="greater") $ p.value   # "less"
prueba.signos (y) # 0.005908966 *
prueba.signos (z) # 0.7482777
```

La única significativa es la muestra tras fase 1, incluso aplicando corrección de Bonferroni.

Otra opción es comprobar si la distribución puede seguir un modelo simétrico. De especial interés es el caso gausiano, que permitiría usar la prueba t, más potente que Wilcoxon:

```
shapiro.test (y) # p-value = 0.148  => gausiana
t.test      (y, mu=110, alternative="greater") # p-value = 0.00140 *
wilcox.test (y, mu=110, alternative="greater") # p-value = 0.00211 *
shapiro.test (z) # p-value = 0.0193 => no gausiana
ksl <- function (x) ks.test(x,punif,min(x),max(x))$statistic
mean (replicate (1e5, ksl (runif(n))) >= ksl(z))
## p-valor = 0.08697 => puede ser unifome => simétrica => wilcox
wilcox.test (z, mu=110, alternative="greater") # p-value = 0.3544
```

Se obtienen los mismos resultados que antes.

Problema 2

El fichero 'Inspeccion.RData' contiene datos registrados sobre la recaudación diaria de un cierto bar en distintos días que se les realizó una inspección sorpresa. Se dispone de las siguientes variables:

Recaudacion: Recaudación diaria (en euros);

DiaSemana: Día de la semana en que se realizó la inspección.

Basándote en los datos recogidos y trabajando a nivel de significación $\alpha = 0{,}05$, responde a las siguientes preguntas:

(1 punto) a) Justifica mediante un estudio inferencial que la distribución de la recaudación para cada día de la semana es aproximadamente una normal con la misma varianza.

(1 punto) b) ¿Hay diferencias en la recaudación en los distintos días de la semana? Realiza un estudio descriptivo previo (tanto numérico como gráfico), basa tus conclusiones en los resultados de un contraste de hipótesis y concluye con un estudio a posteriori.

(1 punto) c) ¿Se distribuyen las inspecciones uniformemente durante los días de la semana? Realiza un estudio descriptivo previo y basa tus conclusiones en los resultados de un contraste de hipótesis.

¡ JUSTIFICA TODAS LAS RESPUESTAS !