PROBABILIDAD Y ESTADISTICA 2001

SEGUNDO PARCIAL
PARTE B

Índice General

1 TEST DE ALEATORIEDAD.

Vamos a verficar si la serie de observaciones (muestra) se puede considerar como aleatoria simple (m.a.s.) , es decir, si se puede aceptar que las observaciones proceden en forma **aleatoria e independiente** de una **misma distribución** (i.i.d.)

1.1 Test de rachas de "subidas y bajadas"

1.1.1 Muestra 1

Se definen las variables

$$U_i = \begin{cases} 1 & \text{si } X_i \leq X_{i+1} \\ 0 & \text{caso contrario} \end{cases} \qquad i = 1, 2, \ldots, n-1$$

y se cuentas las "rachas"

X	U	Rachas
70.1		
70.4	1	
75.8	1	1
67.5	0	2
68.4	1	
73.6	1	
76.9	1	3
75.7	0	
71.4	0	
70.3	0	4
72.1	1	5
69.8	0	6

Estadístico de prueba
$\mathbf{R}$ = número de total rachas = 6

Como

$$\mathbf{R} < \frac{2n-1}{3} = \frac{2\,(12)-1}{3} = 7.6667$$

planteamos

$\mathbf{H}_0$	$\mathbf{H}_1$
$X_1, ..., X_n$ es una m.a.s.	"pocas rachas"

("pocas rachas" = hay una tendencia de los datos a estar agrupados en orden creciente o en orden decreciente).

Usando la Tabla:

n	**R** "pocas rachas"	α^*	**R** "muchas rachas"	α^*
$\downarrow$	$\downarrow$			$\downarrow$
$\downarrow$				
12				
	1	.0000		
	2	.0000		
	3	.0005		
	4	.0082	11	.0113
	5	.0529	10	.0821
	6	**.1918**	9	.2720
$\longrightarrow$	7	.4453	8	.5547

Por lo tanto

$$\boxed{\alpha^* = 0.1918 \geq 0.1}$$

Conclución: No rechazo $\mathbf{H}_0$

1.1.2 Muestra 2

Se definen las variables

$$U_i = \begin{cases} 1 & \text{si } Y_i \leq Y_{i+1} \\ 0 & \text{caso contrario} \end{cases} \qquad i = 1, 2, \ldots, n-1$$

y se cuentas las "rachas"

Y	U	Rachas
74.3		
74.1	0	1
75.4	1	2
67.4	0	3
69.3	1	
70.5	1	4
70.1	0	
69.9	0	
68.7	0	5
70.3	1	
70.7	1	
71.1	1	
74.4	1	6
70.2	0	7

Estadístico de prueba
R = número de total rachas = 7

Como

$$\mathbf{R} < \frac{2n-1}{3} = \frac{2\,(14)-1}{3} = 9$$

planteamos

$\mathbf{H}_0$	$\mathbf{H}_1$
$X_1, ..., X_n$ es una m.a.s.	"pocas rachas"

("pocas rachas" = hay una tendencia de los datos a estar agrupados en orden creciente o en orden decreciente).

Usando la Tabla:

n $\downarrow$ $\downarrow$	$\mathbf{R}$ "pocas rachas" $\downarrow$	α^* $\downarrow$	$\mathbf{R}$ "muchas rachas" $\downarrow$	α^* $\downarrow$
14	1	.0000		
	2	.0000		
	3	.0000		
	4	.0007	13	.0046
	5	.0079	12	.0391
	6	.0441	11	.1536
$\longrightarrow$	7	.1534	10	.3722
	8	.3633	9	.6367

Por lo tanto

$$\boxed{\alpha^* = 0.1534 \geq 0.1}$$

Conclución: No rechazo $\mathbf{H}_0$

1.2 Test de correlación de rangos de Spearman.

1.2.1 Muestra 1

Datos ordenados X_i^*	*Indice (dato original)* i	*Rango* $R(X_i)$
67,5	4	1
68,4	5	2
69,8	12	3
70,1	1	4
70,3	10	5
70,4	2	6
71,4	9	7
72,1	11	8
73,6	6	9
75,7	8	10
75,8	3	11
76,9	7	12

y como $n = 12$ tenemos que:

Estadístico de prueba
$\mathbf{R}_s = 1 - 6\dfrac{\sum\limits_{i=1}^{i=n}(R(X_i) - i)^2}{n(n^2-1)} = 1 - 6\dfrac{264}{12\left((12)^2 - 1\right)} = 0,076923077$

Como es estadístico de correlación de Spearman dió positivo planteamos:

$\mathbf{H}_0$	$\mathbf{H}_1$
$X_1, ..., X_n$ es un m.a.s.	hay tendencia "creciente"

Usando la tabla tenemos que:

n	α^*	**0.1**	**0.05**
$\downarrow$ $\vdots$ $12 \longrightarrow$	$\uparrow$ $\vdots$ $\mathbf{R}_s$ =0,076923077	$\downarrow$ $\vdots$ 0.406	$\downarrow$ $\vdots$ 0.503

por lo tanto

$$\boxed{\alpha^* \geq 0.1 \text{ (p-valor)}}$$

Conclusión: No se rechaza $\mathbf{H}_0$.

1.2.2 Muestra 2

Datos ordenados Y_i^*	*Indice (dato original)* i	*Rango* $R(Y_i)$
67.4	4	1
68.7	9	2
69.3	5	3
69.9	8	4
70.1	7	5
70.2	14	6
70.3	10	7
70.5	6	8
70.7	11	9
71.1	12	10
74.1	2	11
74.3	1	12
74.4	13	13
75.4	3	14

y como $n = 14$ tenemos que:

Estadístico de prueba
$\mathbf{R}_s = 1 - 6\dfrac{\sum\limits_{i=1}^{i=n}(R(Y_i) - i)^2}{n(n^2-1)} = 1 - 6\dfrac{490}{14\left((14)^2 - 1\right)} = -0.076923077$

Como es estadístico de correlación de Spearman dió negativo planteamos:

$\mathbf{H}_0$	$\mathbf{H}_1$
$X_1, ..., X_n$ es un m.a.s.	hay tendencia "decreciente"

Usando la tabla tenemos que:

n	α^*	**0.1**	**0.05**
$\downarrow$	$\uparrow$	$\downarrow$	$\downarrow$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$14 \longrightarrow$	$-\mathbf{R}_s = 0.076923077$	0.367	0.484

por lo tanto

$$\boxed{\alpha^* \geq 0.1 \text{ (p-valor)}}$$

Conclusión: No se rechaza $\mathbf{H}_0$.

2 TEST DE COMPARACIÓN DE DOS MUESTRAS

2.1 Test de Kolmogorov-Smirnov

Tenemos una muestra $X_1, ..., X_{12}$ iid con distribución F (continua) y una muestra $Y_1, ..., Y_{14}$ iid con distribución G (continua) e independientes entre si.

Planteamos

$$\begin{cases} (\mathrm{H}_0)\ F = G \\ (\mathrm{H}_1)\ F \neq G \end{cases}$$

Apliquemos **el test de Kolmogorov - Smirnov**

Estadístico de prueba
$\mathbf{D} = \sup_{t \in R} \left\{ \left| F_n^X(t) - F_m^Y(t) \right| \right\}$ donde
$F_n^X(t) =$ (cantidad de valores muestrales de $X \leq t)n$
$F_m^X(t) =$(cantidad de valores muestrales de$Y \leq t)m$

(con $n = 12$ y $m = 14$) resultando

$$\mathbf{D} = \sup_{t \in R} \left\{ \left| F_m^X(t) - F_n^Y(t) \right| \right\} = 0,25$$

por lo tanto

$$mn\mathbf{D} = (12)(14)(0.25) = 42$$

resultando, de la tabla que

$$\boxed{\alpha^* > 0.2}$$

Conclusión: No se rechaza $\mathbf{H}_0$.

2.2 Tets de Mann -Whitney-Wilcoxon sobre "corrimientos".

Sea $X_1, ..., X_{12}$ una m.a.s. con distribución F (continua y desconocida) e $Y_1, ..., Y_{14}$ una m.a.s. con distribución G (continua y desconocida) e indedendientes entre sí.

(Recordar que X es la muestra de **menor cantidad de datos**)

Estadístico de prueba
$\mathbf{T}_X =$ "suma de rangos de X en la muestra conjunta ordenada"

- El estadístico de prueba $\mathbf{T}_X$ se determina de la siguiente manera:
  - Se juntan **todos** los datos y se obtiene una muestra

$$Z_1, Z_2, ..., Z_N \qquad \text{donde } N = 12 + 14 = 26$$

  - Se ordenan de menor a mayor (los valores que se repitan se escriben repetidos)

$$Z_1^* \leq Z_2^* \leq ... \leq Z_N^*$$

- Se asignan rangos $R(Z_i)$ = "lugar que ocupa Z_i en la muestra ordenada"
- **En este caso hay rangos "empatados" (datos iguales). Se reasignan los rangos, atribuyendole a cada empate el promedio de los rangos asignados inicialmente.**
- Se calcula $\mathbf{T}_X$ = suma de rangos de los datos $X_1, ..., X_{12}$ (siendo $\mathbf{T}_Y$ = suma de rangos de los datos $Y_1, ..., Y_{14}$, se debe verificar que $\mathbf{T}_X + \mathbf{T}_Y = \dfrac{N(N+1)}{2}$)

Luego se obtiene

$$\mathbf{T}_X = 173$$

Como

$$\mathbf{T}_X \geq \frac{n(n+m+1)}{2} = \frac{12(12+14+1)}{2} = 162$$

Planteamos

$\mathbf{H}_0$	$\mathbf{H}_1$
$F = G$	$F(t) = G(t-\theta) \qquad (\forall t \in R,\ \theta > 0)$

(es decir que

$\mathbf{H}_0$	$\mathbf{H}_1$
$X \sim Y$	$X \sim Y + \theta$ con $\theta > 0$ ("hay un corrimiento")

)

Luego, usando la aproximación "normal" (incluida la "corección de continuidad") determinamos el p-valor:

$$z_R(\alpha^*) = \frac{\mathbf{T}_X - \frac{1}{2} - \frac{n(n+m+1)}{2}}{\sqrt{\frac{nm(n+m+1)}{12}}} = 0,848668425 \Leftrightarrow \boxed{\alpha^* = 1 - \Phi(z_R(\alpha^*)) = 0,198032859}$$

Conclusión: No se rechaza $\mathbf{H}_0$

Coclusión primaria: No hay evidencia de diferencia significativa entre los niveles de contaminación de las ciudades

NOTA: no se realizan tets no-parámetricos sobre "dispersión o escala", pues las muestras son pequeñas; y en el curso vimos la versión "asintótica" de dichos tests.

3 TEST DE NORMALIDAD (ajuste).

Vamos a aplicar el test de D'Agostino y el test de Shapiro - Wilks para determinar si los datos son normales (aunque los parámetros μ y σ sean desconocidos)

Sea una m.a.s. $X_1, ..., X_n$ con distribución F (desconocida), planteamos

$\mathbf{H}_0$	$\mathbf{H}_1$
$F =$ **Normal**	$F \neq$ **Normal**

3.1 El Test de Normalidad de D'Agostino.

Estadístico de prueba
$\mathbf{E} = \sum_{k=1}^{k=n} \frac{\left(k - \left(\frac{n+1}{2}\right)\right) X_{(k)}}{\sqrt{n^3 (n-1) s_n^2}}$ donde $X_{(1)} \leq ... \leq X_{(n)}$ es la muestra ordenada

3.1.1 Muestra 1

$$\text{Siendo } n = 12 \qquad \text{y} \qquad s_n^2 = \frac{\sum_{i=1}^{i=n} \left(X_i - \overline{X}_n\right)^2}{n-1} = 9,258787879$$

se tiene que:

$$\mathbf{E} = 0,280325137$$

usando la tabla

n	$\boldsymbol{\alpha} = \mathbf{0.20}$		$\boldsymbol{\alpha} = \mathbf{0.10}$	
$\downarrow$	$\downarrow$		$\downarrow$	
12	$\mathbf{K}_{1,\alpha}(n)$	$\mathbf{K}_{2.\alpha}(n)$	$\mathbf{K}_{1,\alpha}(n)$	$\mathbf{K}_{2.\alpha}(n)$
	0.2653	0.2841	0.2598	0.2849

luego

$$\boxed{\alpha^* > 0.2}$$

Conclusión: No se rechaza $\mathbf{H}_0$.

3.1.2 Muestra 2

$$\text{Siendo } n = 12 \qquad \text{y} \qquad s_n^2 = \frac{\sum_{i=1}^{i=n} \left(X_i - \overline{X}_n\right)^2}{n-1} = 5,819120879$$

se tiene que:

$$\mathbf{E} = 0,271943628$$

usando la tabla

n	$\boldsymbol{\alpha} = \mathbf{0.20}$		$\boldsymbol{\alpha} = \mathbf{0.10}$	
$\downarrow$	$\downarrow$		$\downarrow$	
14	$\mathbf{K}_{1,\alpha}(n)$	$\mathbf{K}_{2.\alpha}(n)$	$\mathbf{K}_{1,\alpha}(n)$	$\mathbf{K}_{2.\alpha}(n)$
	0.2669	0.2846	0.2518	0.2853

luego

$$\boxed{\alpha^* > 0.2}$$

Conclusión: No se rechaza $\mathbf{H}_0$.

3.2 El Test de Normalidad de Shapiro - Wilks.

Estadístico de prueba
$\mathbf{W} = \dfrac{(b_n)^2}{(n-1)\, s_n^2}$
donde $b_n = a_n \left(X_{(n)} - X_{(1)}\right) + a_{n-1}\left(X_{(n-1)} - X_{(2)}\right) + \cdots + a_k\left(X_{(n-k+1)} - X_{(k)}\right)$ con $k = \left[\dfrac{n}{2}\right]$ y $X_{(1)} \leq ... \leq X_{(n)}$ es la muestra ordenada (los coeficientes a_i se obtienen de la tabla del test)

3.2.1 Muestra 1

		coeficientes a_i		
$X_{(12)} = 76,9$	$X_{(1)} = 67,5$	$0,5475$		
$X_{(11)} = 75,8$	$X_{(2)} = 68,4$	$0,3325$		
$X_{(10)} = 75,7$	$X_{(3)} = 69,8$	$0,2347$	$\Rightarrow$	$b_n = 9,74309$
$X_{(9)} = 73,6$	$X_{(4)} = 70,1$	$0,1586$		
$X_{(8)} = 72,1$	$X_{(5)} = 70,3$	$0,0922$		
$X_{(7)} = 71,4$	$X_{(6)} = 70,4$	$0,0303$		

por lo tanto

$$\mathbf{W} = 0,932065878$$

usando la tabla

n	$\boldsymbol{\alpha = 0.1}$	$\boldsymbol{\alpha = 0.5}$
$\downarrow$	$\downarrow$	$\downarrow$
12	0.883	0.943

luego

$$\boxed{0.1 < \alpha^* < 0.5}$$

Conclusión: No se rechaza $\mathbf{H}_0$.

3.2.2 Muestra 2

		coeficientes a_i		
$X_{(14)} = 75,4$	$X_{(1)} = 67,4$	$0,5251$		
$X_{(13)} = 74,4$	$X_{(2)} = 68,7$	$0,3318$		
$X_{(12)} = 74,3$	$X_{(3)} = 69,3$	$0,246$	$\Rightarrow$	$b_n = 8,24405$
$X_{(11)} = 74,1$	$X_{(4)} = 69,9$	$0,1802$		
$X_{(10)} = 71,1$	$X_{(5)} = 70,1$	$0,124$		
$X_{(9)} = 70,7$	$X_{(6)} = 70,2$	$0,0727$		
$X_{(8)} = 70,5$	$X_{(7)} = 70,3$	$0,024$		

por lo tanto

$$\mathbf{W} = 0,898422259$$

usando la tabla

n	$\boldsymbol{\alpha = 0.1}$	$\boldsymbol{\alpha = 0.5}$
$\downarrow$	$\downarrow$	$\downarrow$
14	0.895	0.947

luego

$$\boxed{0.1 < \alpha^* < 0.5}$$

Conclusión: No se rechaza $\mathbf{H}_0$.

4 TEST PARAMETRICOS

Asumiendo la normalidad de los datos, de acuerdo con la parte anterior, realizamos:

4.1 Test paramétrico sobre "dispersión"

Sea $X_1, X_2, \ldots, X_n$ una m.a.s con distribución $N(\mu_X, \sigma_X^2)$ e $Y_1, Y_2, \ldots, Y_m$ una m.a.s con distribución $N(\mu_Y, \sigma_Y^2)$ e independientes entre si,, donde $\sigma_X^2 = \mathbf{Var}(X)$ y $\sigma_Y^2 = \mathbf{Var}(Y)$ son desconocidas

Calculamos las estimaciones de las varianzas

$$s_n^2(X) = 9,258787879 \qquad \text{y} \qquad s_m^2(Y) = 5,819120879$$

Planteamos

$\mathbf{H}_0$	$\mathbf{H}_1$	**Estadístico de prueba**	**Región crítica**
$\sigma_X^2 = \sigma_Y^2$	$\sigma_X^2 > \sigma_Y^2$	$\mathbf{F} = \dfrac{s_n^2(X)}{s_m^2(Y)}$	$\mathbf{F} \geq f_\alpha(n-1, m-1)$

resultando

$$\mathbf{F} = \frac{s_n^2(X)}{s_m^2(Y)} = \frac{9,258787879}{5,819120879} = 1.591097362$$

y usando la Tabla con $\alpha = 0.1$

$$\begin{array}{cccc} n & \longrightarrow & 10 & 12 \\ m & & \downarrow & \downarrow \\ \downarrow & & \downarrow & \downarrow \\ 13 & \rightarrow & 2,14 & 2,1 \end{array}$$

por lo tanto $\alpha^* > 0.1$

Conclusión: No Rechazamos $\mathbf{H}_0$

4.2 Test de comparación de medias.

Sea $X_1, X_2, \ldots, X_n$ una m.a.s con distribución $N\left(\mu_X, \sigma_X^2\right)$ e $Y_1, Y_2, \ldots, Y_m$ una m.a.s con distribución $N\left(\mu_Y, \sigma_Y^2\right)$ e independientes entre si,, donde $\sigma_X^2 = \mathbf{Var}\left(X\right)$ y $\sigma_Y^2 = \mathbf{Var}\left(Y\right)$ son desconocidas pero iguales

$\mathbf{H}_0$	$\mathbf{H}_1$	**Estadístico de prueba**	**Región crítica**
$\mu_X = \mu_Y$	$\mu_X \neq \mu_Y$	$\mathbf{T} = \dfrac{\overline{X}_n - \overline{Y}_m}{s_p\sqrt{\dfrac{1}{n} + \dfrac{1}{m}}}$	$\lvert\mathbf{T}\rvert \geq t_{\frac{\alpha}{2}}\left(n+m-2\right)$

calculando

$$\begin{aligned} \overline{X}_n &= 71,83333333 \qquad \text{y} \qquad \overline{Y}_m = 71,17142857 \\ s_n^2\left(X\right) &= 9,258787879 \qquad \text{y} \qquad s_m^2\left(Y\right) = 5,819120879 \end{aligned}$$

tenemos que

$$s_p^2 = \frac{(n-1)\, s_n^2\left(X\right) + (m-1)\, s_m^2\left(X\right)}{n+m-2} = 7.395634921$$

y por lo tanto

$$\mathbf{T} = \frac{\overline{X}_n - \overline{Y}_m}{s_p\sqrt{\dfrac{1}{n} + \dfrac{1}{m}}} = \frac{71.83333333 - 71.17142857}{\sqrt{(7.395634921)\left(\dfrac{1}{12} + \dfrac{1}{14}\right)}} = 0.61869$$

usando la tabla

$$\begin{array}{cccc} \alpha \longrightarrow & 0.4 & \dfrac{\alpha^*}{2} & 0.25 \\ n \downarrow & & \uparrow & \\ 24 & 0.256 & \mathbf{T} = 0.61869 & 0,685 \end{array}$$

por lo tanto el p-valor α^* cumple que

$$0.25 < \frac{\alpha^*}{2} < 0.4 \Leftrightarrow \boxed{0.5 < \alpha^* < 0.8}$$

Conclusión: No se rechaza $\mathbf{H}_0$.

5 Conclusión final

Ambas muestras son normales con la misma media y la misma varianza.

No hay diferencia significativa entre los niveles de contaminación de ambas ciudades.