UNIVERSITAT DE BARCELONA

Universitat Oberta de Catalunya

Inferencia multivariante

Francesc Carmona

19 de febrero de 2019

1. En un estudio sobre las plagas de diferentes artrópodos en las encinas de las dehesas, la siguiente tabla presenta la distribución conjunta de las variables aleatorias discretas: x_1: diferentes tipos de artrópodos, que toma seis valores posibles P_1 = Orugas, P_2 = Hemipteros, P_3 = Hormigas, P_4 = Arañas, P_5 = Coleópteros y P_6 = Otros y x_2: tipo de dehesa, que toma los valores A = dehesa con pastos, B = dehesa con matorral y C = dehesa cultivada.

	A	B	C
P_1	0.130	0.110	0.090
P_2	0.105	0.065	0.080
P_3	0.070	0.035	0.045
P_4	0.080	0.040	0.080
P_5	0.010	0.020	0.020
P_6	0.005	0.010	0.005

   a) Calcular las distribuciones marginales[1].

   b) Calcular la distribución condicionada de las diferentes clases de artrópodos en la dehesa cultivada[2].

   c) Calcular la distribución condicionada de las dehesas para las arañas[3].

2. (∗∗) Dada la función de densidad conjunta $f(x, y) = 6x$ definida en la región $R = \{(x, y) \ / \ 0 < x < 1, 0 < y < 1 - x\}$,

   a) Comprobar que las densidades marginales de ambas variables son $f(x) = 6x(1 - x)$, para $0 < x < 1$ y $f(y) = 3(1 - y)^2$, para $0 < y < 1$.

   b) Comprobar que las densidades condicionadas son $f(y|x) = \frac{1}{1-x}$, para $0 < y < 1-x$ y $f(x|y) = \frac{2x}{(1-y)^2}$, para $0 < x < 1 - y$.

3. Consideremos un vector aleatorio $\mathbf{x} = (x_1, x_2)$ con distribución uniforme en el rectángulo $[0, 2] \times [3, 4]$.

   a) Especificar la función de densidad conjunta de $\mathbf{x}$ y calcular $E(\mathbf{x})$ y $\text{var}(\mathbf{x})$.

   b) Dada una muestra aleatoria simple $\mathbf{x}_1, \ldots, \mathbf{x}_{30}$ de $\mathbf{x}$, calcular $E(\bar{\mathbf{x}})$ y $\text{var}(\bar{\mathbf{x}})$.

   c) Generar una muestra aleatoria de $\mathbf{x}$ de tamaño 30, dibujarla en un gráfico de dispersión y marcar los puntos $\bar{\mathbf{x}}$ y $E(\bar{\mathbf{x}})$. Calcular también la matriz de covarianzas muestrales $\mathbf{S}$.

   d) Generar 40 muestras de tamaño 5, calcular sus medias muestrales y dibujarlas en un gráfico de dispersión, donde también se representa $E(\bar{\mathbf{x}})$. Repertir este proceso en gráficos distintos para 40 muestras de tamaño 20 y otras 40 de tamaño 50. ¿Qué se observa?

[1] En el caso discreto, las integrales son sumas.
[2] La distribución condicionada de $\mathbf{x}_1$, para un valor concreto de $\mathbf{x}_2 = \mathbf{x}_2^0$ es $f(\mathbf{x}_1|\mathbf{x}_2^0) = f(\mathbf{x}_1, \mathbf{x}_2^0)/f(x_2^0)$.
[3] La distribución condicionada de $\mathbf{x}_2$, para un valor concreto de $\mathbf{x}_1 = \mathbf{x}_1^0$ es $f(\mathbf{x}_2|\mathbf{x}_1^0) = f(\mathbf{x}_1^0, \mathbf{x}_2)/f(x_1^0)$.

4. (∗) **Esperanza y varianza condicionadas**

   Para calcular la esperanza y la matriz de covarianzas del vector $\mathbf{x}_1$ dados los valores del vector $\mathbf{x}_2$, particionamos el vector aleatorio en dos partes $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)'$ con vector de medias $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)'$ y la matriz de covarianzas del vector $\mathbf{x}$ en bloques asociados a estos dos vectores, como:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$$

   Entonces, la distribución condicionada $\mathbf{x}_1|\mathbf{x}_2$ tiene esperanza

$$E[\mathbf{x}_1|\mathbf{x}_2] = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)$$

   y matriz de varianzas y covarianzas

$$\text{var}[\mathbf{x}_1|\mathbf{x}_2] = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$$

   a) Obtener las distribuciones condicionadas en la normal bivariante[4] de media cero y matriz de covarianzas

$$\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

   b) Con la matriz de covarianzas

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho & \rho^2 \\ & 1 & 0 \\ & & 1 \end{pmatrix}$$

   probar que la distribución condicionada de (x_1, x_2) dada x_3, tiene como vector de medias $(\mu_1 + \rho^2(x_3 - \mu_3), \mu_2)'$ y matriz de covarianzas

$$\begin{pmatrix} 1-\rho^4 & \rho \\ \rho & 1 \end{pmatrix}$$

5. *a*) Generar una muestra de tamaño n de una normal bivariante $(X_1, X_2) \sim N_2(\mathbf{0}, \mathbf{I})$.[5]

   b) Comprobar que las variables transformadas

$$\begin{aligned} Y_1 &= \mu_1 + \sigma_1 X_1 \\ Y_2 &= \mu_2 + \sigma_2(\rho X_1 + \sqrt{1-\rho^2} X_2) \end{aligned}$$

   siguen una distribución $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ con

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ & \sigma_2^2 \end{pmatrix}$$

   c) Con unos valores concretos $\boldsymbol{\mu} = (2, 1)'$, $\sigma_1 = 1$, $\sigma_2 = 1.5$ y $\rho = 0.6$, hallar con la transformación explicada en el apartado (b) una muestra aleatoria de (Y_1, Y_2) a partir de la muestra del primer apartado,

6. Generar muestras de una distribución normal bivariante $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ por el método siguiente: (1) generar un valor al azar de la distribución marginal de la primera variable con media μ_1 y varianza σ_1^2; (2) generar un valor al azar de la distribución univariante de la segunda variable dada la primera.

   Para generar el segundo valor dado un valor x_1, hay que tener en cuenta que

$$\begin{aligned} E(X_2|x_1) &= \mu_2 + \sigma_{21}\sigma_{11}^{-1}(x_1 - \mu_1) = \mu_2 + \sigma_1^{-1}\sigma_2\rho(x_1 - \mu_1) \\ \text{var}(X_2|x_1) &= \sigma_{22} - \sigma_{21}\sigma_{11}^{-1}\sigma_{12} = \sigma_2^2(1-\rho^2) \end{aligned}$$

[4] La distribución condicionada de una normal es normal.

[5] Con la orden `rnorm()` de **R** se genera una muestra de una distribución normal estandar univariante.

a) Aplicar este método para generar valores al azar de un vector aleatorio con vector de medias $\boldsymbol{\mu} = (0, 5)'$, desviaciones típicas $(\sigma_1, \sigma_2) = (2, 3)$ y correlación $\rho = 0.5$.

b) Utilizar los datos para estimar una densidad bivariante con la función `bivden()` de Everitt(2005)[6] y representarla con un gráfico en perspectiva con la función `persp()`.

c) Dibujar con los mismos datos un diagrama de dispersión y añadir un gráfico de niveles con la función `contour()`.

7. Para una normal bivariante de media $(1, 2)'$ y matriz de covarianzas

$$\begin{pmatrix} 4 & 3 \\ 3 & 9 \end{pmatrix}$$

a) Representarla con un gráfico en tres dimensiones[7].

b) Calcular la probabilidad del rectángulo $P(1 < X_1 \leq 1.5, 2 < X_2 \leq 2.75)$.
Se puede utilizar la función de distribución `pmvnorm()` del paquete mvtnorm de **R**.

8. (∗∗) Dibujar el gráfico de la sección 4.6 del curso online STAT 505

   `https://newonlinecourses.science.psu.edu/stat505/node/36/`

9. (∗∗) **Geometría de la Distribución Normal Multivariante**

   Calcular los semiejes l_j del elipsoide al 95 % para las variables numéricas de los datos `crabs` del paquete `MASS` para los individuos machos de la especie Blue tal como se explica en la página:

   `https://newonlinecourses.science.psu.edu/stat505/node/36/`

   Calcular también el volumen del mismo elipsoide.

10. En la Tabla 1 se muestran los datos de 28 alcornoques que miden los depósitos de corcho (en centigramos) en cada uno de los puntos cardinales: N, E, S, W. El vector de medias es

$$\bar{\mathbf{x}} = (50.536, 46.179, 49.679, 45.179)'$$

y la matriz de covarianzas

$$\mathbf{S} = \begin{pmatrix} 280 & 216 & 278 & 218 \\ & 212 & 221 & 165 \\ & & 337 & 250 \\ & & & 218 \end{pmatrix}$$

Las siguientes variables compuestas explican diferentes aspectos de la variabilidad de los datos:

Contraste eje $N - S$ con eje $E - W$:	$Y_1 = N + S - E - W$
Contraste $N - S$:	$Y_2 = N - S$
Contraste $E - W$:	$Y_3 = E - W$

a) Realizar una matriz de dispersiones de las variables originales dos a dos. Añadir las rectas de regresión y los histogramas en la diagonal.

b) Calcular el vector de medias y la matriz de covarianzas de las variables compuestas.

c) Lo mismo para las variables transformadas y normalizadas (la suma de cuadrados de los coeficientes es 1).

[6] Ver la página web `http://biostatistics.iop.kcl.ac.uk/publications/everitt/`.

[7] Por ejemplo, en la página `http://www.stat.cmu.edu/ kass/KEB/RHTML/R/bivariateNormalPerspectives.r.html` tenemos un código que se puede adaptar. No confundir la correlación entre dos variables con su covarianza.

11. (∗∗) Probar que si la matriz de varianzas y covarianzas muestral de unos datos normales es

$$\mathbf{S} = \frac{1}{n}\mathbf{X}'\mathbf{H}\mathbf{X}$$

donde $\mathbf{H}$ es la matriz de centrado, entonces

$$n\mathbf{S} \sim W_p(\mathbf{\Sigma}, n-1)$$

12. Para observar gráficamente que $T^2(p, n-1) \to \chi^2_p$, representar en un mismo gráfico las densidades $T^2(3, 14)$, $T^2(3, 49)$ y χ^2_3

    La aproximación se considera válida si el cociente $n/p > 15$.

13. Con la matriz de datos siguiente

X_1	X_2	X_3
6.0	4.7	3.3
5.0	4.2	2.9
4.5	5.1	3.2
6.2	4.9	4.0
5.4	4.5	3.7
5.7	4.6	2.2
7.2	5.1	4.4
6.5	6.4	5.3
2.8	3.9	2.5
5.6	4.6	3.2
6.1	4.9	4.0
4.5	4.3	3.5

    a) Calcular el vector de medias, la matriz de covarianzas y la matriz de correlaciones.

    b) Calcular la varianza generalizada, la varianza total y el coeficiente de dependencia global $\eta^2 = 1 - |\mathbf{R}|$.

    c) Contrastar las hipótesis univariantes $H_0^{(1)} : \mu_1 = 5$, $H_0^{(2)} : \mu_2 = 4$ y $H_0^{(3)} : \mu_3 = 3$. Utilizar, por ejemplo, la función `t.test` de **R**.

    d) Contrastar la hipótesis $H_0 : (\mu_1, \mu_2, \mu_3)' = (5, 4, 3)'$ con el estadístico T^2 de Hotelling y hallar el estadístico F correspondiente.

14. (∗∗)**Contrastes sobre la matriz de covarianzas de una población normal**

    En una población normal multivariante $N_p(\boldsymbol{\mu}, \mathbf{\Sigma})$ el contraste

$$H_0 : \mathbf{\Sigma} = \mathbf{\Sigma}_0 \qquad H_1 : \mathbf{\Sigma} \neq \mathbf{\Sigma}_0$$

    se realiza con la razón de versosimilitudes.

    Si el logaritmo de la versosimilitud es

$$l = \log L(\boldsymbol{\mu}, \mathbf{\Sigma}|\mathbf{x}) = -\frac{n}{2}\log|2\pi\mathbf{\Sigma}| - \frac{n}{2}\mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{S}) - \frac{n}{2}(\bar{\mathbf{x}} - \boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu})$$

    Bajo la hipótesis nula, los estimadores de los parámetros son $\bar{\mathbf{x}}$ y $\mathbf{\Sigma}_0$, mientras que en general son $\bar{\mathbf{x}}$ y $\mathbf{S}$.

    De modo que

$$l_0 = -\frac{n}{2}\log|2\pi\mathbf{\Sigma}_0| - \frac{n}{2}\mathrm{tr}(\mathbf{\Sigma}_0^{-1}\mathbf{S})$$
$$l_1 = -\frac{n}{2}\log|2\pi\mathbf{S}| - \frac{np}{2}$$

Así que

$$-2\log\lambda = 2(l_1 - l_0) = n\log\frac{|\boldsymbol{\Sigma}_0|}{|\mathbf{S}|} + n\ \mathrm{tr}(\boldsymbol{\Sigma}_0^{-1}\mathbf{S}) - np$$

que asintóticamente tiene distribución $\chi^2_{p(p+1)/2}$.

a) Comprobar los cálculos anteriores.

b) Concretar el test $H_0 : \boldsymbol{\Sigma} = \mathbf{I}$.

c) Dado que para poblaciones normales incorrelación implica independencia, el contraste de independencia es $H_0 : \boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$.
La estimación de los parámetros bajo esa hipótesis es $\bar{\mathbf{x}}$ y $\mathrm{diag}(\mathbf{S})$.
Comprobar que el estadístico para contrastar la independencia es

$$-2\log\lambda = -n\log|\mathbf{R}|$$

cuya distribución asintótica es $\chi^2_{p(p-1)/2}$.

15. Los vectores de medias y las matrices de covarianzas de $n_1 = 24$ tortugas macho y $n_2 = 24$ tortugas hembra, respecto a las variables $X_1 =$ longitud, $X_2 =$ anchura y $X_3 =$ altura del caparazón en mm, son:

$$\bar{\mathbf{x}}_m = \begin{pmatrix} 113.38 \\ 88.29 \\ 40.71 \end{pmatrix} \qquad \mathbf{S}_m = \begin{pmatrix} 132.99 & 75.85 & 35.82 \\ & 47.96 & 20.75 \\ & & 10.79 \end{pmatrix}$$

$$\bar{\mathbf{x}}_h = \begin{pmatrix} 136.00 \\ 102.58 \\ 51.96 \end{pmatrix} \qquad \mathbf{S}_h = \begin{pmatrix} 432.58 & 259.87 & 161.67 \\ & 164.57 & 98.99 \\ & & 63.87 \end{pmatrix}$$

a) Calcular la matriz de covarianzas común $\mathbf{S}$.

b) Descartar la independencia de las variables en los dos grupos con el test $-2\log\lambda = -n\log|\mathbf{R}|$ que sigue asintóticamente[8] una distribución $\chi^2_{p(p-1)/2}$.

c) Comparar de forma univariante las medias de X_1, X_2 y X_3 en las dos poblaciones.

d) Comparar de forma multivariante las medias en las dos poblaciones.

e) Comparar las matrices de covarianzas de las dos poblaciones con el test de la razón de verosimilitudes:

$$-2\log\lambda = n\log|\mathbf{S}| - n_1\log|\mathbf{S}_m| - n_2\log|\mathbf{S}_h|$$

que asintóticamente sigue una distribución $\chi^2_{p(p+1)/2}$.

16. La tabla 2 contiene las medidas de 5 variables biométricas sobre gorriones hembra, recogidos casi moribundos después de una tormenta[9]. Los primeros 21 sobrevivieron mientras que los 28 restantes no lo consiguieron. Las variables observadas son $X_1 =$ longitud total, $X_2 =$ extensión del ala, $X_3 =$ longitud del pico y de la cabeza, $X_4 =$ longitud del húmero y $X_5 =$ longitud del esternón.

Suponiendo normalidad multivariante,

a) Comparar las covarianzas entre el grupo de supervivientes y el de no supervivientes con el test de la razón de verosimilitudes.
Calcular también el test M de Box que se explica en Cuadras(2018)[10] y que se puede aplicar con la función `boxM()` del paquete `heplots`.

[8] Ver Peña (2002) pág. 295 o Cuadras(2018) pág. 52.
[9] También podemos utilizar el *data.frame* `sparrows` del paquete `mAr` e incorporar el factor supervivencia.
[10] Ver la sección 7.5.2.

b) Comparar las medias de las dos poblaciones con el estadístico T^2 de Hotelling.

c) (∗) Comparar las medias de las dos poblaciones con el estadístico Λ de Wilks.

17. Con los datos de los gorriones hembra del ejercicio anterior y teniendo en cuenta que el test M de Box es muy sensible a la no normalidad de los datos:

    a) Comparar la variabilidad de los dos grupos de datos (supervivientes y no supervivientes) con el test de Levene múltiple (no multivariante) que se puede aplicar con la función `leveneTests()` del paquete `heplots`.

    b) (∗∗) Comparar la variabilidad de los dos grupos con una generalización multivariante del test de Levene llamada test de Van Valen. El algoritmo del test es el siguiente:

    1) Estandarizar todas las variables numéricas para que tengan media cero y varianza 1.
    2) Calcular los centroides (vector de medianas) de cada grupo.
    3) Restar a cada observación su centroide y tomar el valor absoluto del resultado.
    4) Calcular las distancias euclídeas de cada observación al origen de coordenadas (es la norma euclídea de las filas de la base de datos).
    5) Comparar estas distancias (variable univariante) con un test t para los dos grupos. También se puede utilizar un test no paramétrico.

    Observemos que los tests propuestos en este ejercicio comparan la varianza multivariante de los grupos y no las matrices de varianzas-covarianzas como los tests del ejercicio anterior.

    Nota: Anderson (2006) sugirió utilizar como centroide el punto que minimiza la suma de distancias euclídeas de las observaciones a ese punto en cada grupo. Para comparar la variabilidad de los grupos se puede utilizar la distribución F del ANOVA de las distancias o un test de permutaciones. La función `betadisper()` del paquete `vegan` implementa este procedimiento.

18. **Los datos de Iris de Edgar Anderson**

    El *data.frame* `iris` de **R** contiene los famosos datos de Iris de Fisher(1936) o de Anderson(1935) con las variables longitud del sépalo, anchura del sépalo, longitud del pétalo y anchura del pétalo, medidas en centímetros sobre tres especies de flores del género Iris: Iris setosa, Iris versicolor e Iris virginica.

```
data(iris)
help(iris)
attach(iris)
```

    a) Realizar un tratamiento descriptivo de los datos.

    b) Dibujar un boxplot múltiple con el paquete lattice.

    c) Representar las tres poblaciones conjuntamente en gráficos de dispersión para las variables numéricas dos a dos.

    d) Comparar las matrices de covarianzas de las tres especies con el test de la razón de verosimilitudes y con el test M de Box.

    e) Realizar un MANOVA de un factor, la especie, y contrastar si las medias de las especies se pueden considerar estadísticamente distintas.

    f) (∗∗) Realizar una representación canónica de las tres poblaciones con regiones confidenciales para los individuos medios de cada grupo.

    Utilizar la función `candisc()` del paquete candisc o consultar

    `http://erre-que-erre-paco.blogspot.com/2010/03/analisis-canonico-de-poblaciones.html`

Tabla 1: Depósitos de corcho en centigramos de 28 alcornoques en las cuatro direcciones cardinales.

N	E	S	W	N	E	S	W
72	66	76	77	91	79	100	75
60	53	66	63	56	68	47	50
56	57	64	58	79	65	70	61
41	29	36	38	81	80	68	58
32	32	35	36	78	55	67	60
30	35	34	26	46	38	37	38
39	39	31	27	39	35	34	37
42	43	31	25	32	30	30	32
37	40	31	25	60	50	67	54
33	29	27	36	35	37	48	39
32	30	34	28	39	36	39	31
63	45	74	63	50	34	37	40
54	46	60	52	43	37	39	50
47	51	52	43	48	54	57	43

Tabla 2: Medidas biométricas sobre unos gorriones.

Supervivientes					No supervivientes				
X_1	X_2	X_3	X_4	X_5	X_1	X_2	X_3	X_4	X_5
156	245	31.6	18.5	20.5	155	240	31.4	18.0	20.7
154	240	30.4	17.9	19.6	156	240	31.5	18.2	20.6
153	240	31.0	18.4	20.6	160	242	32.6	18.8	21.7
153	236	30.9	17.7	20.2	152	232	30.3	17.2	19.8
155	243	31.5	18.6	20.3	160	250	31.7	18.8	22.5
163	247	32.0	19.0	20.9	155	237	31.0	18.5	20.0
157	238	30.9	18.4	20.2	157	245	32.2	19.5	21.4
155	239	32.8	18.6	21.2	165	245	33.1	19.8	22.7
164	248	32.7	19.1	21.1	153	231	30.1	17.3	19.8
158	238	31.0	18.8	22.0	162	239	30.3	18.0	23.1
158	240	31.3	18.6	22.0	162	243	31.6	18.8	21.3
160	244	31.1	18.6	20.5	159	245	31.8	18.5	21.7
161	246	32.3	19.3	21.8	159	247	30.9	18.1	19.0
157	245	32.0	19.1	20.0	155	243	30.9	18.5	21.3
157	235	31.5	18.1	19.8	162	252	31.9	19.1	22.2
156	237	30.9	18.0	20.3	152	230	30.4	17.3	18.6
158	244	31.4	18.5	21.6	159	242	30.8	18.2	20.5
153	238	30.5	18.2	20.9	155	238	31.2	17.9	19.3
155	236	30.3	18.5	20.1	163	249	33.4	19.5	22.8
163	246	32.5	18.6	21.9	163	242	31.0	18.1	20.7
159	236	31.5	18.0	21.5	156	237	31.7	18.2	20.3
					159	238	31.5	18.4	20.3
					161	245	32.1	19.1	20.8
					155	235	30.7	17.7	19.6
					162	247	31.9	19.1	20.4
					153	237	30.6	18.6	20.4
					162	245	32.5	18.5	21.1
					164	248	32.3	18.8	20.9