UTILIZACIÓN DE MÉTRICAS RIEMANNIANAS EN ANALISIS DE DATOS MULTIDIMENSIONALES Y SU APLICACIÓN A LA BIOLOGÍA

JOSE Mª OLLER SALA

BARCELONA, 25 de NOVIEMBRE de 1982.

nocido, como las reacciones leucemoides mielocíticas. Los hiperplasias linfocíticas pueden ser debidas a infecciones víricas, por ejemplo la mononucleosis infecciosa producida por el virus de Epstein-Barr, pueden ser causadas por un descenso relativo de las demás lineas celulares, o pueden tener un origen desconocido, como las reacciones leucemoides linfocíticas. Las hiperplasias eritrocíticas estan con frecuencia asociadas a diversos tipos de anemias, hemorragias periódicas, etc. Las anemias megaloblásticas, como la anemia perniciosa, cursan con la aparición de elementos de la serie megaloblástica, ausente en una médula osea normal.

Hay enfermedades de naturaleza maligna que implican cambios en las proporciones celulares de la médula ósea. Entre éstas, las leucemias. Se distinguen dos grandes grupos de leucemias, las crónicas y las agudas. Las leucemias crónicas se subdividen a su vez en tres grupos, la leucemia linfática crónica (LLC), la leucemia mielocítica crónica (LMC) y la leucemia mielomonocitica crónica (LMMC).

La LLC se caracteriza por una leucocitosis elevada en sangre periférica (del orden de 10^5 glóbulos blancos/mm^3 es común) siendo la mayoría de los leucocitos, linfocitos maduros. A nivel de médula ósea se aprecia una infiltracción linfocítica que puede sobrepasar el 90% de los elementos celulares.

La LMC se caracteriza también por una elevada leucocitosis en sangre periférica tan marcada o más que en el caso anterior,

pero caracterizada esta vez por un aumento relativo de la serie mieloide, con la aparición de elementos inmaduros, como metamielocitos, mielocitos, etc. y hasta algún mieloblasto. A nivel de médula ósea hay una notable hiperplasia granulocítica con un aumento relativo de formas inmmaduras. Suele haber un apreciable aumento de células basófilas, hecho que las distingue de algunas reacciones leucemoides. Existen variantes eosinofílicas de la enfermedad (leucemia a eosinófilos) que clínicamente no se distinguen prácticamente del tipo común. En la etapa final de la enfermedad suele haber un aumento de células blásticas (mieloblastos), transformándose en un cuadro parecido o igual a una leucemia aguda.

La leucemia LMMC, se diferencia de la anterior por la aparición de células de características intermedias entre los monoblastos y los mieloblastos. Así, se habla de blastos mielomonociticos o de mielomonocitos. Otra diferencia es que la cifra de leucocitos en sangre periférica no suele estar demasiado alejada de la normalidad.

En cuanto a las leucemias agudas se subdividen en nueve grupos, todos ellos caracterizados a nivel medular por un gran aumento del número de blastos (mieloides, linfoides, etc.).

La LAM1, es la leucemia mielocítica aguda, sin maduración celular. La mayor parte de las células de la serie granulocítica son mieloblastos.

La LAM2, es la leucemia mielocítica aguda, con maduración celular. Más de la mitad de los elementos de la serie granulocítica son mielocitos u otras formas más maduras, aunque continua habiendo una notable cantidad de mieloblastos.

La LAM3, es la leucemia aguda progranulocítica. Los elementos presentes en mayor número (a veces casi la totalidad) son progranulocitos.

La LAM5 es la leucemia aguda mielomonocítica. Hay un gran aumento de blastos mielomonocíticos.

La LAM6 es la también llamada eritroleucemia. Es la versión aguda de la mielosis eritrémica crónica o síndrome de Di Guglielmo. Hay un aumento de la serie granulocítica (mieloblastos) y de la megalocítica (megaloblastos, promegaloblastos) a expensas de la serie eritrocítica.

La LAL1, es la leucemia linfática aguda, tipo 1, caracterizada por células predominantes, de la serie linfocítica, de pequeño tamaño. Hay linfocitos (linfoblastos) de tipo B y T.

La LAL2, es también una leucemia linfática aguda, pero esta se caracteriza por células predominantes de mayor tamaño y morfología más irregular. Hay células tipos B y T.

La LAL3 es la leucemia linfática aguda, tipo 3. Las células predominantes son parecidas a las que se observan en el linfoma de Burkitt. Todos tienen marcadores B.

Existen otros tipos posibles de leucemias a considerar, como las tricoleucemias (reticuloendoteliosis leucémica), las leucemias de células plasmáticas, etc., pero éstas son más infrecuentes y a veces la fase terminal de algún proceso neoplástico. Así por ejemplo en la fase final de un mieloma puede desarrollarse una leucemia de células plasmáticas.

Otra de las enfermedades que ocasionan variaciones en los porcentajes celulares es la mielomatosis, proceso neoplásico consistente en la aparición de mielomas, tumores de la médula ósea con abundantes plasmocitos, proplasmocitos y plasmoblastos. En el mielograma aparece la línea plasmocítica muy aumentada.

Mayor información sobre las enfermedades que afectan a la médula ósea puede encontrarse en Miale (1982).

8.2.3. Una distancia en el conjunto de los resultados

El conjunto de los resultados de un mielograma puede identificarse con la variedad:

$$E = \{(p^1,\ldots p^k) \in R^k \;/\; p^i \geq 0 \quad i=1,..,k \; \sum_{i=1}^{k} p^i \leq 1 \} \qquad (12)$$

que aparece en el capítulo IV asociada a la distribución multinomial.

La distancia entre dos puntos $(p^1,\ldots,p^k)$ y $(q^1,\ldots,q^k)$ de E, vendrá dada por:

$$d = \sqrt{2} \arccos \left(\sum_{i=1}^{k+1} \sqrt{p^i q^i} \right) \qquad (13)$$

con $p^{k+1} = 1-p^1-\ldots-p^k$, $q^{k+1} = 1-q^1-\ldots-q^k$.

En base a (13) se podrá definir una partición del conjunto de resultados, teniendo en cuenta (4).

8.3. ANALISIS CLINICOS Y RESULTADOS

En un principio se pensó en elegir como población base, a todos los enfermos a los que había de efectuarseles por primera vez un mielograma, ingresados en un Servicio de Hematología de una Ciudad Sanitaria de la Seguridad Social. Se hubiese escogido entonces una muestra controlada de los mismos, para iniciar el algoritmo de clasificación. Sin embargo, sólo se tuvo acceso a resultados de análisis con la impresión diagnóstica correspondiente, no a las historias clínicas completas con los resultados del primer mielograma efectuado. Entonces se escogió como muestra controlada 56 casos clínicos descritos en Miale (1982), conscientes de que las probabilidades "a priori" estimadas a partir de dicha muestra, diferiran notablemente de las probabilidades "a priori" de la población del servicio antes citado.

Sin embargo, como el algoritmo tiene prevista una realimentación, introduciendo como un nuevo elemento de la muestra controlada a cada nuevo caso diagnosticado (posee una cierta "capacidad de aprendizaje") a medida que se diagnostiquen nuevos casos, se mejorará la estimación de las probabilidades a priori.

Las clases patológicas consideradas, un tanto condicionadas por la muestra controlada (casos clínicos "de libro", Miale (1982)), han sido las siguientes:

Clase I. Normales. Caso clínico 1.

Clase II. Hiperplasias granulocíticas. Casos clínicos 2,3,11,12,14, 15,16.

Clase III. Hiperplasias eritrocítica. Casos clínicos 4,5,6,7,18,47, 53,57.

Clase IV. Hiperplasia linfocítica. Casos clínicos 9,50,56.

Clase V. Anemias megaloblástica. Casos clínicos 10,19,20,21,22.

Clase VI. LLC. Casos clínicos 40,41

Clase VII. LMMC. Caso clínico 27.

Clase VIII LMC. Caso clínico 18,29

Clase IX. LAM1. Casos clínicos 30,33.

Clase X. LAM2. Casos clínicos 31,32,36.

Clase XI. LAM3. Caso clínico 34.

Clase XII. LAM4. Casos clínicos 24,25,26,28.

Clase XIII LAM5. Caso clínico 23.

Clase XIV. LAM6. Caso clínico 37.

Clase XV. LAL (LAL1,LAL2 y LAL3). Casos clínicos 35,38,39,48.

Clase XVI. Mielomatosis. Casos clínicos 42,43,44,45.

Clase XVII. Varios. Casos clínicos 8,17,46,49,51,52,55.

Las variables estudiadas han sido los porcentajes de los siguientes tipos celulares:

Variable	Células	Variable	Células
1	Mieloblastos	12	Linfoblastos
2	Progranulocitos	13	Prolinfocitos+linfocitos
3	Mielocitos	14	Monoblastos
4	Metamielocitos	15	Promonocitos+Monocitos
5	Segmentados+bandas,Neutrófilos	16	Serie plasmocítica
6	Segmentados+bandas,Eosinófilos	17	Serie megalocítica
7	Segmentados+bandas,Basófilos	18	Células reticulares
8	Pronormoblastos	19	Serie megacariocítica
9	Normoblastos basófilos	20	Blastos mielomonociticos
10	Normoblastos policromatófilos	21	Monocitos+Promonocitos (mielomonocitos)
11	Normoblastos ortocrómicos	22	Otras células

Se han escogido y agrupado estas 22 variables para que los resultados de los mielogramas de los casos clínicos obtenido "de libro" fueran compatibles con los resultados de mielogramas que efectúa la Seguridad Social.

Se han escrito programas de ordenador para ejecutar el algoritmo propuesto en el apartado primero. Para cada análisis se obtienen las probabilidades a priori y a posteriori de las distintas clases patológicas, la clase patológica más probable, la probabilidad de

error y el cambio de información. Se ejecutó dicho programa con 15 resultados de mielogramas de la Seguridad Social con su correspondiente impresión diagnóstica. A continuación, vamos a describir los resultados.

Análisis 1

Resultados mielograma:

Variables	porcentaje	variables	porcentaje	variables	porcentaje
1	3.0	8	0.0	15	0.0
2	13.0	9	1.0	16	1.0
3	21.0	10	1.0	17	0.0
4	22.0	11	7.0	18	0.0
5	21.0	12	0.0	19	0.0
6	3.0	13	0.0	20	0.0
7	7.0	14	0.0	21	0.0
				22	0.0

fue clasificado, por el algoritmo, como un caso de LMC (Clase VIII), coincidiendo con la impresión diagnóstica.

Análisis 2

Variable	porcentaje	variable	porcentaje	variable	porcentaje
1	0.0	8	0.0	15	0.0
2	10.1	9	2.0	16	3.0
3	17.2	10	11.1	17	0.0
4	18.2	11	14.1	18	0.0
5	19.2	12	0.0	19	0.0
6	5.1	13	0.0	20	0.0
7	0.0	14	0.0	21	0.0
				22	0.0

fue clasificado como perteneciente al grupo XVII (varios), mientras que la impresión diagnóstica era "médula ósea dentro de los límites de la normalidad". Cabe destacar, sin embargo, la ausencia de linfocitos y unos porcentajes algo elevados en la serie granulocítica, en particular los progranulocitos, (por esto fue clasificado por el algoritmo como "varios" en vez de "normal").

Análisis 3

Variable	porcentaje	Variable	porcentaje	Variable	porcentaje
1	2.0	8	0.0	15	0.0
2	5.0	9	1.0	16	2.0
3	10.0	10	10.0	17	0.0
4	14.0	11	20.0	18	2.0
5	24.0	12	0.0	19	0.0
6	5.0	13	5.0	20	0.0
7	0.0	14	0.0	21	0.0
				22	0.0

fue clasificado como perteneciente al grupo XVII (varios), mientras que la impresión diagnóstica era, al igual que el caso anterior, "médula ósea dentro de los límites de la normalidad". Sin embargo, se aprecia un aumento (con respecto la normalidad) de los normoblastos ortocrómicos, aunque no se puede hablar de hiperplasia eritrocítica. Probablemente por esta razón fuese clasificado en el grupo de "varios" en vez de catalogarlo en el grupo "normal".

Análisis 4

Variable	porcentaje	Variable	porcentaje	Variable	porcentaje
1	2.2	8	3.3	15	0.0
2	6.7	9	8.9	16	2.2
3	16.7	10	11.1	17	0.0
4	17.8	11	23.3	18	0.0
5	4.4	12	0.0	19	0.0
6	0.0	13	3.3	20	0.0
7	0.0	14	0.0	21	0.0
				22	0.0

fue clasificado como un caso de hiperplasia eritrocítica, coincidiendo con la impresión diagnóstica.

Análisis 5

Variable	porcentaje	Variable	porcentaje	Variable	porcentaje
1	0.0	8	4.0	15	0.0
2	6.1	9	4.0	16	42.4
3	8.1	10	5.0	17	0.0
4	5.1	11	12.1	18	0.0
5	6.1	12	0.0	19	0.0
6	0.0	13	2.0	20	0.0
7	0.0	14	0.0	21	0.0
				22	0.0

fue clasificado como un caso de mielomatosis, coincidiendo con la impresión diagnóstica.

Análisis 6

Variable	porcentaje	Variable	porcentaje	Variable	porcentaje
1	0.97	8	0.97	15	0.00
2	7.77	9	2.9	16	0.97
3	8.74	10	11.65	17	0.0
4	17.47	11	18.44	18	0.0
5	19.41	12	0.0	19	0.0
6	3.89	13	3.89	20	0.0
7	0.0	14	0.0	21	0.0
				22	0.0

fue clasificado como perteneciente al grupo XVII (varios), mientras que la impresión diagnóstica fue de "médula ósea dentro de los límites de la normalidad". Sin embargo, se aprecian algunas desviaciones en los porcentajes de la variable 2 (progranulocitos) que se halla aumentada (corrimiento hacia la izquierda de la serie granulocítica) y un aumento de normoblastos ortocrómicos, variable 11. Por esto probablemente no fue clasificado como normal.

Análisis 7

Variable	porcentaje	Variable	porcentaje	Variable	porcentaje
1	0.0	8	0.0	15	0.0
2	1.0	9	0.0	16	1.0
3	2.0	10	2.0	17	0.0
4	1.0	11	4.0	18	1.0
5	29.0	12	57.0	19	0.0
6	1.0	13	1.0	20	0.0
7	0.0	14	0.0	21	0.0
				22	0.0

fue clasificado como un caso de LLA (clase XV) coincidiendo con la impresión diagnóstica.

Análisis 8

Variable	porcentaje	Variable	porcentaje	Variable	porcentaje
1	29.0	8	2.0	15	0.0
2	6.0	9	3.0	16	2.0
3	5.0	10	16.0	17	0.0
4	7.0	11	16.0	18	1.0
5	8.0	12	0.0	19	0.0
6	0.0	13	5.0	20	0.0
7	0.0	14	0.0	21	0.0
				22	0.0

fue clasificado como un caso de LAM1, mientras que la impresión diagnóstica fue un caso de LAM2. Aunque, por lo general, los casos de LAM1 suelen presentar un porcentaje más elevado de mieloblastos, los casos LAM2 presentan corrientemente un mayor porcentaje de formas maduras de la serie granulocítica, por esta razón fue probablemente clasificado como LAM1.

Análisis 9

Variable	porcentaje	Variable	porcentaje	Variable	porcentaje
1	0.0	8	7.0	15	0.0
2	3.0	9	10.0	16	0.0
3	8.0	10	19.0	17	0.0
4	11.0	11	22.0	18	0.0
5	12.0	12	0.0	19	0.0
6	4.0	13	4.0	20	0.0
7	0.0	14	0.0	21	0.0
				22	0.0

fue clasificado como un caso de hiperplasia eritrocítica (clase III), coincidiendo con la impresión diagnóstica.

Análisis 10

Variable	porcentaje	Variable	porcentaje	Variable	porcentaje
1	75.0	8	1.0	15	0.0
2	9.0	9	0.0	16	0.0
3	7.0	10	0.0	17	0.0
4	1.0	11	3.0	18	0.0
5	4.0	12	0.0	19	0.0
6	0.0	13	0.0	20	0.0
7	0.0	14	0.0	21	0.0
				22	0.0

fue clasificado como un caso de LAM1, coincidiendo con la impresión diagnóstica.

Análisis 11

Variable	porcentaje	Variable	porcentaje	Variable	porcentaje
1	0.0	8	0.0	15	0.0
2	0.0	9	1.0	16	0.0
3	0.0	10	2.0	17	0.0
4	0.0	11	1.0	18	1.0
5	0.0	12	93.0	19	0.0
6	0.0	13	2.0	20	0.0
7	0.0	14	0.0	21	0.0
				22	0.0

fue clasificado como un caso de LAL, coincidiendo con la impresión diagnóstica.

Análisis 12

Variable	porcentaje	Variable	porcentaje	Variable	porcentaje
1	0.0	8	0.0	15	0.0
2	1.0	9	2.0	16	1.0
3	2.0	10	7.0	17	0.0
4	2.0	11	3.0	18	1.0
5	3.0	12	0.0	19	0.0
6	1.0	13	77.0	20	0.0
7	0.0	14	0.0	21	0.0
				22	0.0

fue clasificado como un caso de LLC, coincidiendo con la impresión diagnóstica.

Análisis 13

Variable	porcentaje	Variable	porcentaje	Variable	porcentaje
1	0.0	8	0.0	15	0.0
2	0.0	9	0.0	16	82.0
3	1.0	10	2.0	17	0.0
4	1.0	11	4.0	18	0.0
5	3.0	12	0.0	19	0.0
6	3.0	13	4.0	20	0.0
7	0.0	14	0.0	21	0.0
				22	0.0

fue clasificado como un caso de mielomatosis, coincidiendo con la impresión diagnóstica.

Análisis 14

Variable	porcentaje	Variable	porcentaje	Variable	porcentaje
1	4.76	8	2.86	15	0.0
2	11.42	9	2.86	16	0.0
3	36.19	10	0.95	17	0.0
4	19.05	11	1.90	18	0.0
5	7.62	12	0.0	19	0.0
6	6.67	13	2.86	20	0.0
7	2.86	14	0.0	21	0.0
				22	0.0

fue clasificado como un caso de LMC, coincidiendo con la impresión diagnóstica.

Análisis 15

Variable	porcentaje	Variable	porcentaje	Variable	porcentaje
1	0.0	8	1.0	15	0.0
2	0.0	9	3.0	16	0.0
3	0.0	10	6.0	17	0.0
4	3.0	11	8.0	18	0.0
5	3.0	12	0.0	19	0.0
6	3.0	13	73.0	20	0.0
7	0.0	14	0.0	21	0.0
				22	0.0

fue clasificado como un caso de LLC, coincidiendo con la impresión diagnóstica.

8.4. DISCUSION

De los quince análisis considerados el algoritmo ha coincidido con la impresión diagnóstica en once casos. El resultado debe considerarse como bastante satisfactorio, dado que la muestra controlada era pequeña (sólo había un individuo en el grupo "normal") y además las discrepancias entre el resultado de aplicar el algoritmo y la impresión diagnóstica no han sido extremas, e incluso tal vez hayan sido razonables.

El siguiente paso sería introducir estos quince casos en la muestra controlada, para ser utilizados en el diagnóstico de nuevos casos. De disponer de un mayor número de casos, bien diagnosticados, sería interesante trabajar con más clases patológicas, incluyendo por ejemplo la mielosis eritrémica, la tricoleucemia (retículo endoteliosis maligna), etc.

También sería de interés que se aumentara el número de células muestradas por enfermo (hasta 500 por ejemplo) en vez del habitual contaje sobre 100. Igualmente sería de interés distinguir más tipos de células, aumentando las posibilidades del algoritmo.

Los resultados anteriores deben considerarse como un ejemplo de como puede automatizarse el diagnóstico de enfermedades, siempre que procuremos expresar en términos cuantitativos los resultados de los análisis.

9. UNA ALTERNATIVA AL TEST T DE STUDENT PARA MUESTRAS INDEPENDIENTES.

Resumen:

En el presente capítulo se plantea un contraste sobre las medias y las varianzas de dos poblaciones normales univariantes independientes, utilizando una función distancia, desarrollada en el capítulo 5, y considerando la distribución de los estadísticos que intervienen en el mismo. Dichos resultados se aplican al estudio de la influencia de la alcohol deshidrogenasa y el tamaño del cuerpo, en *Drosophila melanogaster*.

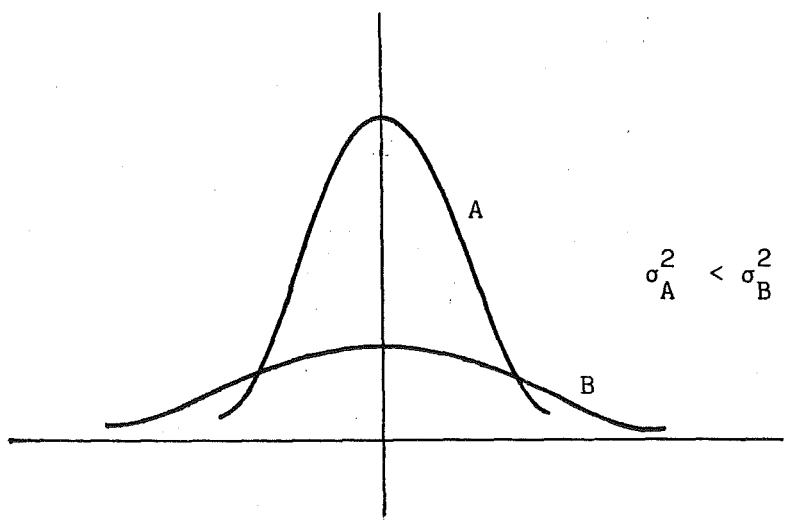
Sumario:

9.1. PLANTEO GENERAL

9.1.1. Test estadístico, necesidad y descripción.

El test t de Student de muestras independientes se utiliza, como es bien sabido, para hacer inferencia estadística acerca de las medias de dos poblaciones normales univariantes independientes y de varianza común y desconocida. Con cierta frecuencia las varianzas poblacionales no coinciden en ambas poblaciones, problema estudiado por Behrens y Fisher, teniéndose que aplicar otro test de hipótesis, basado también en una t, Lindgren (1976).

Sin embargo, otras veces, puede tener interés contrastar si dos poblaciones normales univariantes independientes son iguales. Para ello no basta que las esperanzas de ambas poblaciones coincidan, hecho que queda patente en el dibujo, donde se representan dos funciones de densidad normales univariantes, A y B, con la misma esperanza:



sino que es preciso que coincidan también sus varianzas.

Ello puede resultar interesante, por ejemplo, al estudiar variables fisiológicas, que deben mantenerse, debido a la homeostasis, entre unos límites estrechos. No sólo tiene interés el valor medio que toma la variable, sino que importan también las fluctuaciones de la variable respecto a éste, medidas por la varianza. Cuanto menor ésta, tanto mejor es la regulación.

El contraste puede plantearse, más formalmente, como:

$$\begin{aligned} &H_o: \mu_A = \mu_B \quad \sigma_A = \sigma_B \\ &H_1: \mu_A \neq \mu_B \text{ ó } \sigma_A \neq \sigma_B \end{aligned} \tag{1}$$

y puede ser resuelto a través del test sugerido al final del capítulo III, aplicado al caso de una distribución normal univariante.

Es claro que si $\mu_A = \mu_B$ y $\sigma_A = \sigma_B$ entonces la distancia entre ambas poblaciones es $\Delta = d(A,B) = 0$, y podemos replantear el contraste (1) en términos de distancias:

$$\begin{aligned} &H_o: \quad \Delta = 0 \\ &H_1: \quad \Delta > 0 \end{aligned} \tag{2}$$

Si disponemos, para cada población, de una muestra de tamaño N_A y N_B respectivamente, la región crítica, de nivel de significación ε, que tomaremos, vendrá dada por:

$$w_\varepsilon = \{x \in \mathbb{R}^{N_A+N_B} / U(x) \geq A_\varepsilon\} \tag{3}$$

siendo

$$U(x) = \frac{N_A\, N_B}{N_A+N_B}\, D^2 \qquad (4)$$

y D^2 la distancia, al cuadrado, entre A y B, calculada a través de las estimaciones máximo-verosímiles de los parámetros μ_A,σ_A, μ_B y σ_B, que son respectivamente: $\bar{x}_A$, s_A, $\bar{x}_B$ y s_B. Para el cálculo explícito de D tendremos que considerar los siguientes casos:

$$\text{a) } \bar{x}_A = \bar{x}_B \Rightarrow D = \sqrt{2}\, \left| \ln \frac{s_B}{s_A} \right|$$

$$\left.\begin{array}{l} \text{b) } \bar{x}_A \neq \bar{x}_B \\ (\bar{x}_B-\bar{x}_A)^2 \leq 2|s_A^2-s_B^2| \end{array}\right\} \Rightarrow D = \sqrt{2}\, |\operatorname{argsech}(|c_1|\, s_A)-\operatorname{argsech}(|c_1|\, s_B)| \qquad (5)$$

$$\left.\begin{array}{l} \text{c) } \bar{x}_A \neq \bar{x}_B \\ (\bar{x}_B-\bar{x}_A)^2 > 2|s_A^2-s_B^2| \end{array}\right\} \Rightarrow D = \sqrt{2}\, (\operatorname{argsech}(|c_1|\, s_A)+\operatorname{argsech}(|c_1|\, s_B))$$

y la constante c_1 viene dada por:

$$c_1 = \frac{\sqrt{2}\, (\bar{x}_B-\bar{x}_A)}{\sqrt{\frac{(\bar{x}_B-\bar{x}_A)^4}{4} + (s_A^2-s_B^2)^2 + (s_A^2+s_B^2)(\bar{x}_B-\bar{x}_A)^2}} \qquad (6)$$

9.1.2. Distribución asintótica bajo H_o.

Para el cálculo efectivo de A_ε, de la expresión (3), debe tenerse en cuenta que la distribución asintótica del estadístico U, bajo la hipótesis nula H_o, es una ji-cuadrado con 2 grados de libertad.

Por tanto habrá que hallar A_ε de forma que si χ^2 es una variable aleatoria que sigue una distribución ji-cuadrado con 2 grados de libertad, se cumple:

$$\text{Prob}\,[\chi^2 > A_\varepsilon] = \varepsilon \qquad (7)$$

9.1.3. Distribución asintótica bajo H_1.

La expresión (112) del capítulo III permite hallar la distribución asintótica de D (estimación máximo-verosimil de la distancia entre dos poblaciones) dado Δ (distancia real entre las dos poblaciones). Para nuestro caso particular, podremos escribir, siendo $i=\sqrt{-1}$, lo siguiente:

$$f(D,\Delta) = \frac{N_A\,N_B}{N_A+N_B}\,D\,e^{-\frac{1}{2}((N_AN_B)/(N_A+N_B))(D^2+\Delta^2)}\,J_o\left(i\,\frac{N_A\,N_B}{N_A+N_B}\,D\,\Delta\right) \qquad (8)$$

equivalente a:

$$f(D,\Delta) = \frac{N_A\,N_B}{N_A+N_B}\,D\,e^{-\frac{1}{2}((N_AN_B)/(N_A+N_B))(D^2+\Delta^2)}\left(\sum_{n=o}^{\infty}\left(\frac{N_AN_B}{2(N_A+N_B)}\,D\Delta\right)^{2n}\frac{1}{n!\Gamma(n+1)}\right)$$

(9)

Expresión que puede ser utilizada para el cálculo de la función de potencia, del contraste de hipótesis, para muestras grandes:

$$\beta(\Delta) = \int_{A_\varepsilon}^{\infty} f(D,\Delta)\,dD \qquad (10)$$

También puede ser usada las expresiones (8) ó (9) para resolver

contrastes del tipo:

$$H_o: \Delta = \Delta_o$$
$$H_1: \Delta \neq \Delta_o \text{ (o bien } \Delta > \Delta_o \text{ ó } \Delta < \Delta_o) \qquad (11)$$

puesto que nos permitiría conocer la distribución del estadístico D bajo la hipótesis nula, $\Delta = \Delta_o$.

9.1.4. Distribución para muestras pequeñas.

En orden a evaluar cuales deben ser los tamaños muestrales mínimos para poder utilizar la distribución asintótica como buena aproximación, se han generado, en distintas situaciones (diferentes parámetros, diferentes tamaños muestrales) y mediante el empleo de ordenador, muestras de poblaciones normales univariantes, calculándose la distancia estimada entre dos muestras y hallando la distribución empírica del estadístico U ó D que aparecen en (4), que por el teorema de Glivenko-Cantelli, Rios (1977), converge casi seguramente y uniformemente a la distribución real.

Para obtener números aleatorios que sigan una distribución normal, se han generado primero números aleatorios uniformemente distribuidos en el intervalo (0,1), siguiendo el método del generador lineal congruencial, Yakowitz (1977), y a partir de dos números aleatorios independientes y uniformemente distribuidos en el intervalo (0,1), a través de una variante del método de Box-Müller, obtener un número aleatorio que se distribuya según una distribución normal de esperanza μ y desviación típica σ, Yakowitz (1977).

Veamos a continuación varias tablas de resultados, que han sido obtenidas, cada una de ellas, a partir de dos series de 10000 muestras aleatorias simples de tamaño n, obtenidas ambas de una población normal univariante de media μ y desviación típica σ. Para cada uno de éstos 10000 pares de muestras, se procede a estimar, por máxima verosimilitud, μ y σ. Cada par de estimaciones de μ y σ permite el cálculo del estadístico $U = \frac{N_A N_B}{N_A+N_B} D^2$, con $N_A = N_B = n$.

TABLA I

$n=15$, $\mu=0$, $\sigma=1$

U	Función de distribución empírica de U	Función de distribución ji-cuadrado, 2 gra. lib.	Diferencia
0.0100	0.0059	0.0050	0.0009
0.0201	0.0113	0.0100	0.0013
0.0506	0.0259	0.0250	0.0009
0.1030	0.0468	0.0500	-0.0032
0.2110	0.0969	0.1000	-0.0031
0.5750	0.2380	0.2500	-0.0120
1.3900	0.4611	0.5000	-0.0389
2.7700	0.7081	0.7500	-0.0419
4.6100	0.8672	0.9000	-0.0328
5.9900	0.9239	0.9500	-0.0261
7.3800	0.9548	0.9750	-0.0202
9.2100	0.9770	0.9900	-0.0130
10.600	0.9861	0.9950	-0.0089

diferencia máxima en valor absoluto: -0.0419 (aprox. 4%)

TABLA II

$n=30$, $\mu=0$, $\sigma=1$

U	Función de distribución empírica de U	Función de distribución ji-cuadrado, 2 gra. lib.	Diferencia
0.0100	0.0045	0.0050	-0.0005
0.0201	0.0105	0.0100	0.0005
0.0506	0.0261	0.0250	0.0011
0.1030	0.0486	0.0500	-0.0014
0.2110	0.0995	0.1000	-0.0005
0.5750	0.2482	0.2500	-0.0018
1.3900	0.4789	0.5000	-0.0211
2.7700	0.7235	0.7500	-0.0265
4.6100	0.8817	0.9000	-0.0183
5.9900	0.9373	0.9500	-0.0127
7.3800	0.9650	0.9750	-0.0100
9.2100	0.9851	0.9900	-0.0049
10.600	0.9926	0.9950	-0.0024

diferencia máxima en valor absoluto: -0.0265 (aprox. 2,5%)

TABLA III

n=15, μ=0, σ=0.1

U	Función de distribución empírica de U	Función de distribución ji-cuadrado,2 gra. lib.	Diferencia
0.0100	0.0043	0.0050	-0.0007
0.0201	0.0097	0.0100	-0.0003
0.0506	0.0234	0.0250	-0.0016
0.1030	0.0454	0.0500	-0.0046
0.2110	0.0928	0.1000	-0.0072
0.5750	0.2268	0.2500	-0.0232
1.3900	0.4622	0.5000	-0.0378
2.7700	0.7009	0.7500	-0.0491
4.6100	0.8623	0.9000	-0.0377
5.9900	0.9229	0.9500	-0.0271
7.3800	0.9542	0.9750	-0.0208
9.2100	0.9780	0.9900	-0.0120
10.600	0.9867	0.9950	-0.0083

diferencia máxima en valor absoluto: -0.0491 (aprox. 5%)

TABLA IV

n=30, μ=0, σ=0.1

U	Función de distribución empírica de U	Función de distribución ji-cuadrado,2 gra. lib.	Diferencia
0.0100	0.0045	0.0050	-0.0005
0.0201	0.0090	0.0100	-0.0010
0.0506	0.0235	0.0250	-0.0015
0.1030	0.0470	0.0500	-0.0030
0.2110	0.0954	0.1000	-0.0046
0.5750	0.2339	0.2500	-0.0161
1.3900	0.4815	0.5000	-0.0185
2.7700	0.7268	0.7500	-0.0232
4.6100	0.8808	0.9000	-0.0192
5.9900	0.9360	0.9500	-0.0140
7.3800	0.9630	0.9750	-0.0120
9.2100	0.9826	0.9900	-0.0074
10.600	0.9903	0.9950	-0.0047

diferencia máxima en valor absoluto: -0.0232 (aprox. 2.5%)

TABLA V

n=15, μ=0, σ=10

U	Función de distribución empírica de U	Función de distribución ji-cuadrado,2 gr. lib.	Diferencia
0.0100	0.0044	0.0050	-0.0006
0.0201	0.0099	0.0100	-0.0001
0.0506	0.0250	0.0250	0.0000
0.1030	0.0491	0.0500	-0.0009
0.2110	0.0955	0.1000	-0.0005
0.5750	0.2317	0.2500	-0.0183
1.3900	0.4655	0.5000	-0.0345
2.7700	0.7090	0.7500	-0.0410
4.6100	0.8630	0.9000	-0.0370
5.9900	0.9226	0.9500	-0.0274
7.3800	0.9545	0.9750	-0.0205
9.2100	0.9776	0.9900	-0.0124
10.600	0.9861	0.9950	-0.0089

diferencia máxima en valor absoluto: -0.041 (aprox. 4%)

TABLA VI

n=30, μ=0, σ=10

U	Función de distribución empírica de U	Función de distribución ji-cuadrado, 2 gr. lib.	Diferencia
0.0100	0.0043	0.0050	-0.0007
0.0201	0.0086	0.0100	-0.0014
0.0506	0.0244	0.0250	-0.0006
0.1030	0.0492	0.0500	-0.0008
0.2110	0.0952	0.1000	-0.0048
0.5750	0.2423	0.2500	-0.0077
1.3900	0.4860	0.5000	-0.0140
2.7700	0.7347	0.7500	-0.0153
4.6100	0.8919	0.9000	-0.0081
5.9900	0.9418	0.9500	-0.0082
7.3800	0.9691	0.9750	-0.0059
9.2100	0.9846	0.9900	-0.0054
10.600	0.9916	0.9950	-0.0034

diferencia máxima en valor absoluto: -0.0153 (aprox. 1.5%)

TABLA VII

$n=15$, $\mu=1$, $\sigma=1$

U	Función de distribución empírica de U	Función de distribución ji-cuadrado, 2 gr. lib.	Diferencia
0.0100	0.0046	0.0050	-0.0004
0.0201	0.0094	0.0100	-0.0006
0.0506	0.0215	0.0250	-0.0035
0.1030	0.0439	0.0500	-0.0061
0.2110	0.0887	0.1000	-0.0113
0.5750	0.2297	0.2500	-0.0203
1.3900	0.4722	0.5000	-0.0278
2.7700	0.7107	0.7500	-0.0393
4.6100	0.8685	0.9000	-0.0315
5.9900	0.9247	0.9500	-0.0253
7.3800	0.9560	0.9750	-0.0190
9.2100	0.9767	0.9900	-0.0133
10.600	0.9853	0.9950	-0.0097

diferencia máxima en valor absoluto: -0.0393 (aprox. 4%)

TABLA VIII

$n=30$, $\mu=1$, $\sigma=1$

U	Función de distribución empírica de U	Función de distribución ji-cuadrado, 2 gr. lib.	Diferencia
0.0100	0.0053	0.0050	0.0003
0.0201	0.0095	0.0100	-0.0005
0.0506	0.0239	0.0250	-0.0011
0.1030	0.0474	0.0500	-0.0026
0.2110	0.0966	0.1000	-0.0034
0.5750	0.2407	0.2500	-0.0093
1.3900	0.4793	0.5000	-0.0207
2.7700	0.7254	0.7500	-0.0246
4.6100	0.8810	0.9000	-0.0190
5.9900	0.9358	0.9500	-0.0142
7.3800	0.9655	0.9750	-0.0095
9.2100	0.9850	0.9900	-0.0050
10.600	0.9909	0.9950	-0.0041

diferencia máxima en valor absoluto: -0.0246 (aprox. 2.5%)

En vista de los resultados de las simulaciones, expuestos en las tablas anteriores, parece razonable tomar como valor crítico, para muestras de tamaño próximo a 15 y nivel de significación del 5%, 7.12 y para muestras de tamaño próximo a 30, con el mismo nivel de significación, 6.48. Para tamaños muestrales intermedios es aconsejable interpolar linealmente y para tamaños muestrales más elevados puede servir la aproximación asintótica. Si consideramos el nivel de significación del 1%, para muestras de tamaño 15 tomaremos como valor crítico 11.54 y para muestras de tamaño 30, 10.28. Para tamaños muestrales más elevados utilizaremos la aproximación asintótica. En breve se va a efectuar un estudio más sistemático sobre los valores críticos al utilizar muestras pequeñas.

A continuación vamos a evaluar la potencia del test ante una situación concreta. Vamos a suponer que tenemos dos poblaciones normales univariantes independientes, A y B, con $\mu_A=0$, $\sigma_A=1$, $\mu_B=1$ y $\sigma_B=1$. La distancia real entre éstas es $d(A,B) = 0.9803$. Se han generado 10000 muestras de tamaño n de la población A y 10000 muestras de tamaño n de la población B. Para cada muestra de A y de B se evaluan $\bar{x}_A$, s_A, $\bar{x}_B$ y s_B. Tendremos pues 10000 pares de estimaciones de los parámetros de las poblaciones A y B. Para cada uno de éstos pares de estimaciones se calcula la distancia entre A y B, hallándose la distribución empírica del estadístico D. Los resultados se resumen en las tablas IX y X, para n=15 y n=30 respectivamente.

TABLA IX

n=15

D	Función de distribución empírica de D	D	Función de distribución empírica de D
0.000	0.00	1.109	0.50
0.516	0.05	1.158	0.55
0.641	0.10	1.206	0.60
0.730	0.15	1.256	0.65
0.799	0.20	1.311	0.70
0.857	0.25	1.371	0.75
0.912	0.30	1.441	0.80
0.965	0.35	1.517	0.85
1.012	0.40	1.613	0.90
1.062	0.45	1.773	0.95

TABLA X

n=30

D	Función de distribución empírica de D	D	Función de distribución empírica de D
0.00	0.00	1.038	0.50
0.624	0.05	1.072	0.55
0.706	0.10	1.106	0.60
0.767	0.15	1.142	0.65
0.813	0.20	1.182	0.70
0.859	0.25	1.224	0.75
0.902	0.30	1.270	0.80
0.940	0.35	1.322	0.85
0.973	0.40	1.387	0.90
1.007	0.45	1.490	0.95

Consideremos ahora el caso n=15 y nivel de significación $\varepsilon=0.05$. El valor crítico para el estadístico U hemos quedado en tomarlo igual a 7.12. Ello implica que el valor crítico del estadístico D es 0.975. De la tabla IX deducimos, por interpolación, que la potencia del test es del 64%.

Si consideramos n=30, con el nivel de significación del 5%, debido a que el valor crítico del estadístico U es 6.48, el valor crítico del estadístico D será 0.657. De la tabla X, por interpolación, deducimos que la potencia del test es del 93%.

9.2. APLICACION AL ESTUDIO DE LA RELACION DE LA ALCOHOL DESHIDROGENASA Y LA SELECCION POR EL TAMAÑO EN *DROSOPHILA MELANOGASTER*

Se ha estudiado, en *D. melanogaster* el locus que determina la sintesis de la alcohol deshidrogenasa (Adh), distinguiéndose dos alelos, que designaremos como + y -, que determinan la síntesis de dos alcohol deshidrogenasas separables electroforéticamente, intentando establecer alguna relación entre el genotipo, para este locus, y el tamaño del cuerpo, reflejado por el tamaño del ala.

9.2.1. Descripción de experimentos y resultados

Se han utlizado parte de los resultados obtenidos en la tesis doctoral de Serra (1977), y los detalles experimentales pueden encontrarse en la misma. Resumiendo, consideraremos tres líneas LH1, LL1

y RWL1. En la primera se ha efectuado selección por el tamaño del ala, escogiendo a los más grandes de cada generación para formar la siguiente, en la segunda se ha efectuado selección por el tamaño del ala, escogiendo a los más pequeños y en la tercera no se ha efectuado selección artificial alguna. Se ha tormado una muestra de cada línea en las generaciones 0, 2 y 5 (teniendo en cuenta que la generación 0 consta de una única línea: aún no se ha efectuado selección por tamaño). Para cada individuo de la muestra se le ha medido la longitud del ala y determinado el sexo (♂,♀) y el genotipo (++, +- ó --). Así podemos considerar varias poblaciones estadísticas según la línea generación, genotipo y sexo. Por ejemplo, la línea LH1, genotipo ++, generación 5, sexo ♂, la línea RWL1, generación 2, genotipo +-, sexo ♀, etc. En cada una de éstas se ha estimado la media y la desviación típica de la longitud del ala, obteniendo los siguientes resultados experimentales, resumidos en la tabla XI.