ESTIMACION DEL NUMERO DE LOCI QUE AFECTA A UNA CARACTERISTICA CUANTITATIVA

OCTAVIO MARTINEZ DE LA VEGA

T E S I S

PRESENTADA COMO REQUISITO PARCIAL

PARA OPTAR AL GRADO DE

MAESTRO EN CIENCIAS

ESPECIALIDAD DE ESTADISTICA EXPERIMENTAL

Universidad Autonoma Agraria Antonio Narro

PROGRAMA DE GRADUADOS

Buenavista, Saltillo, Coah.

NOVIEMBRE DE 1987.

Tesis elaborada bajo la supervisión del comité particular de asesoría y aprobada como requisito parcial, para optar al grado de

MAESTRO EN CIENCIAS ESPECIALIDAD
DE ESTADISTICA EXPERIMENTAL

COMITE PARTICULAR

Asesor principal: _____
Dr. Rolando Cavazos Cadena

Asesor: _____
MC. José Antonio Díaz García

Asesor: _____
MC. Félix de Jesús Sánchez Pérez

_____
Dr. Eleuterio López Pérez
Subdirector de Asuntos de Postgrado

Buenavista, Saltillo, Coahuila. Noviembre 1987

Deseo expresar mi más profundo agradecimiento al Dr. Rolando - Cavazos Cadena, por el asesoramiento de este trabajo, así como por su - ayuda durante el desarrollo de mis estudios.

A los maestros MC. José Antonio Díaz García y MC. Félix de Jesús Sanchez Pérez por su apoyo y consejo durante el desarrollo de la investigación.

A mi esposa Yolanda, porque gracias a la confianza que depositó en mi y por su comprensión, pude culminar el presente trabajo.

A las instituciones que con su apoyo económico me permitieron desarrollar mis estudios: la Universidad Autónoma de San Luis Potosí, - la Secretaría de Educación Pública y el Consejo Nacional de Ciencia y - Tecnología.

COMPENDIO

Estimación del Número de Loci que Afecta a una Característica Cuantitativa.

POR

OCTAVIO MARTINEZ DE LA VEGA

MAESTRIA

ESTADISTICA EXPERIMENTAL

UNIVESIDAD AUTONOMA AGRARIA ANTONIO NARRO

BUENAVISTA, SALTILLO, COAHUILA. NOVIEMBRE DE 1987.

Dr. Rolando Cavazos Cadena -Asesor-

Palabras clave: Estimación, número de loci, carácter cuantitativo, estimación insesgada.

En este trabajo se estudia la estimación del número de loci -o pares de genes- que afectan la herencia de una característica cuantitativa (o continua). Se formula un modelo estadístico que surge de consideraciones biológicas. Se denota por k al número de loci relevantes y se propone un conjunto de estimadores para k, algunos de los cuales ya se encuentran citados en la literatura genética, sin embargo se conoce muy poco de sus propiedades. Aquí demostramos el hecho de que ninguno de estos estimadores de k tiene esperanza finita, siendo sin embargo -

consistentes. Mas aun, nuestro principal resultado es la inexistencia - de estimadores insesgados de k. Concluimos con algunos comentarios so - bre nuestros supuestos y resultados, así como sobre su significado biológico.

ABSTRACT

Estimation of the Number of Loci Affecting
a Quantitative Trait

BY

OCTAVIO MARTINEZ DE LA VEGA

MASTER OF SCIENCE

EXPERIMENTAL STATISTICS

UNIVERSIDAD AUTONOMA AGRARIA ANTONIO NARRO

BUENAVISTA, SALTILLO, COAHUILA. NOVEMBER 1987.

Dr. Rolando Cavazos Cadena -Advisor-

Key words: Estimation, number of loci, quantitative trait, unbiased estimation.

This work is concerned with the estimation of the number of loci -or gene pairs- that affect a quantitative (or continuos) trait. A - Statistical Model arising from a set of biological considerations is - formulated. The number of relevant loci is denoted by k, and then a set of estimators for k is proposed. Some of these estimators are aviable - in the genetical literature, but too little is known about their properties. Here we prove that each of these estimators does not has finite - expectation, however they are consistent. Moreover, our main result is

the nonexistence of unbiased estimators for k. We conclude with some - brief comments about our basic assumtions and results, as well as on - their biological meaning.

INDICE DE CONTENIDO

PREFACIO

La Genética Cuantitativa estudia la herencia de caracterísi - cas que se miden en una escala continua, como son el peso o la longitud de una planta o un animal. Se ha comprobado experimentalmente que éstas características comunmente están afectadas por un número (fijo pero desconocido) de pares de genes. En éste trabajo se estudia la posibilidad de estimar dicho parámetro.

Para ello, en el capítulo uno se presenta una breve reseña histórica del problema, pasando a formular el modelo genético para el fenómeno en cuestión.

En el capítulo dos se plantea la contraparte estadística del - modelo, incluyendo las variables de interés, sus funciones de densidad así como sus esperanzas y varianzas. Tanto el modelo genético como el - estadístico implican una serie de hipótesis que se hacen explícitas conforme se desarrolla la presentación.

En el capítulo tres se presentan estimadores de momentos para el número de loci, que son función de esperanzas y varianzas de las variables del modelo. Enseguida se demuestra el hecho de que dichos estimadores no tienen esperanza finita, y por ello no son aplicables en este caso los métodos comunes de selección de estimadores.

En el capítulo cuatro se demuestra el resultado más importante de éste trabajo, a saber, que no existen estimadores insesgados para el número de loci. Su importancia radica en que *a priori* sabemos que -

cualquier estimador que se haya propuesto, o se proponga en el futuro, será sesgado o bien no tendra esperanza finita (como es el caso de los estimadores propuestos en el capítulo tres).

El trabajo concluye, en el capítulo cinco, con algunos comentarios sobre los casos en que las hipótesis propuestas para el modelo no se cumplen y algunas consideraciones prácticas sobre la selección del mejor estimador para el número de loci.

CAPITULO 1

INTRODUCCION: EL MODELO GENETICO

El supuesto básico de la Genética Cuantitativa es que la herencia de características continuas puede ser explicado por el efecto de varios pares de genes, es decir de varios loci diferentes (Mather y Jinks, 1977; Mayo, 1980). En este contexto surge naturalmente la pregunta sobre cuantos loci están influyendo en la herencia de una característica. El objetivo de este trabajo es estudiar la posibilidad de obtener estimadores de este parámetro, con propiedades 'deseables' como son el insesgamiento y la consistencia (para una definición de estos conceptos vea el capítulo tres, o bien: Koroliuk, 1981).

En la siguiente sección se presenta un boquejo del desarrollo histórico del problema que nos ocupa, lo cual permitirá una comprensión mas aguda de este y facilitará el planteamiento del modelo genético así como de su contraparte estadística.

1.1 Desarrollo Histórico

En el siglo pasado se publicó el trabajo "Experimentos de Hibridación en Plantas", en el cual se sientan las bases de la Genética, con las leyes de segregación y transmisión independiente de caracteres (Mendel, 1865). Es evidente que el éxito de esta investigación se debe, ademas de al indiscutible genio del autor, a una elección adecuada del material experimental (plantas del genero Pisum) y sobre todo de las características que eligió para observar. Estas últimas fueron caracteres

discretos, bien definidos y clasificables en una sola categoría cada uno (semillas lisas o rugosas, etc.). Al parecer Mendel nunca intentó dilucidar el mecanismo de la herencia de características esencialmente continuas (como el peso o longitud de un organismo). No obstante su importancia el trabajo de Mendel permaneció prácticamente desconocido hasta el año de 1900, en que DeVries, Correns y Tchermark lo redescubrieron, confirmaron y extendieron (Jamenson, 1977).

Con anterioridad a Mendel varios investigadores habian intentado comprender el mecanismo de la herencia, sin embargo elegian características continuas, por lo cual los resultados de sus experimentos fueron demasiado confusos para darles una interpretación coherente.

Mientras que el trabajo de Mendel permanecía en el olvido, varios investigadores trabajaban sobre el problema de la herencia, enfocandose sobre todo a características continuas (ver por ejemplo Baldwin, 1896). De particular importancia resultan los trabajos del gran Estadístico Francis Galton, que a partir de 1876 propone su "Ley de la Herencia Ancestral", la cual pretendía explicar la herencia de las características continuas por medio de la contribución de todos los ancestros. Inclusive, en el año de 1889 Galton publica un artículo donde presenta observaciones del pedigree de perros de caza, que parecian confirmar su teoría, la cual declara universal (Galton, 1889). Aún cuando erroneo, el punto de vista de Galton tiene el gran mérito de ser la primera aproximación cuantitativa al problema de la herencia de caracteres continuos, que además está expresada en lenguaje matemático. Posteriormente su teoría fúe utilizada contra el enfoque mendeliano de la herencia.

El redescubrimiento de las leyes de Mendel en 1900 da lugar a una fuerte polémica: por un lado los "estadísticos", que defienden la -

teoría de Galton de la herencia ancestral y por otro los "mendelianos", que no aceptan lo anterior y han confirmado experimentalmente los trabajos de Mendel. En 1902 Yule (un Matemático competente) publica una revisión de los dos puntos de vista y apunta que estos no son forzosamente contradictorios, y que ambos deben de integrarse en una teoría única de la herencia (Yule, 1906). Sin embargo las dos teorías no son "perfectamente consistentes" como pretendia Yule. Por ejemplo, Castle (1903) publica los análisis de cruzamientos entre lineas de ratones, comparando el ajuste de los datos con lo esperado bajo ambas teorias y concluye - que se ajustan a las leyes de Mendel, pero de ninguna manera a la teo - ría de la herencia ancestral de Galton. El año siguiente se publica "Sobre una Teoría Alternativa de la Herencia con especial referencia a las Leyes de Mendel" (Pearson, 1904). En este trabajo se desarrolla la teoría de la correlación utilizada previamente por Galton y se duda aún de la universalidad de las leyes de Mendel, así como de su aplicación a características continuas y a poblaciones naturales. También Bateson en - la reimpresión de los trabajos de Mendel (ver Mendel, 1865) duda de la posible extensión de las leyes mendelianas a poblaciones naturales y - menciona la posible coexistencia de otras leyes de un orden de complejidad mucho mayor. Con respecto a ello, en el año de 1908 se publican dos artículos casi simultaneos, pero totalmente independientes, donde se demuestra que en una población de cruzamiento libre se puede presentar - cualquier proporción fenotípica, sin contravenir con ello las leyes de Mendel (Hardy, 1908; Weinberg, 1908).

En el año siguiente se publican dos artículos donde se demuestra que, si se excluye el principio de dominancia, hay poca contradic - ccion entre Galton y Mendel (Pearson, 1909a; Pearson 1909b). -

También Weinberg opina que la controversia no tiene fundamento (ver Jamenson, 1977). Un año después, East (1910) muestra que la variación de color en los granos de maíz depende de dos loci, y que por tanto se puede dar una interpretación mendeliana a variación que es aparentemente continua. Posteriormente junto con Nilsson-Ehle establece la hipótesis de factores múltiples en la herencia cuantitativa, esto es, la hipótesis de que la herencia de características continuas puede estar determinada por varios loci que actuan al mismo tiempo. Como ya se mencionó esta hipótesis ha pasado a ser el principio básico de la Genética Cuantitativa.

El descubrimiento y confirmación de este principio permitieron un enorme desarrollo de la Genética Cuantitativa y de Poblaciones, que fue llevado a cabo por Haldene (1926) Wright (1932) Fisher (1958) y otros (ver ademas Robinson, 1986). Debido a la necesidad de poner rápidamente en operación principios prácticos, derivados del desarrollo teórico que se dió entre 1910 y 1930, muchos puntos de importancia han quedado relativamente poco estudiados. Uno de los problemas que al parecer necesitan de un mayor desarrollo es el de la estimación de parámetros genéticos y particularmente la estimación del número de loci involucrados en la herencia de características continuas. Parece evidente que un avance en el conocimiento teórico de las propiedades que posean los estimadores de este parámetro, pudiera redundar en el mejoramiento de las estrategias de selección para características continuas, que suelen ser las de mayor importancia económica.

Autores contemporaneos mencionan la importancia de estimar este parámetro y proponen algunos estimadores que son función de medias y varianzas de las poblaciones estudiadas (Mather y Jinks, 1977). Estos

autores llaman "factores efectivos" a la estimación del número de loci, por considerar que sus estimadores dan el número mínimo de de loci involucrados. Estos estimadores y algunos otros estan mencionados también por Mayo (1980) que recalca la importancia práctica de estimar dicho parámetro, pero menciona que todos los estimadores tienen "error" o "falta de precisión" en la estimación. Esta dificultad es insalvable, pues como veremos en el capítulo cuatro, no exsisten estimadores insesgados del número de loci involucrados en la herencia de una característica cuantitativa. Es importante mencionar el poco trabajo teórico que se ha realizado en cuanto a la estimación de parámetros genéticos, ya que inclusive en textos de nivel avanzado se menciona muy poco al respecto.

En la siguiente sección se presenta el modelo genético que se asume en este trabajo, así como los supuestos, o hipótesis de trabajo, que involucra.

1.2 El Modelo

Para poder establecer una correspondencia entre lo que ocurre en la realidad y lo que se puede inferir de ella, es decir, para modelar un fenómeno natural, es necesario hacer ciertos supuestos que simplifiquen la situación real, lo suficiente para poder manejar la inferencia, pero no tanto que el modelo resulte estéril. Esto es lo que se pretende en esta sección con respecto a la herencia de características continuas: presentar un modelo que esté de acuerdo con la evidencia disponible, que permita hacer suposiciones que simplifiquen esta situación y que pueda resultar fructífero en lo que respecta a la solución del problema planteado.

Asumiremos que se tienen dos poblaciones que difieren en una característica cuantitativa de interés, por ejemplo: dos razas de ganado con una producción media de leche diferente o dos lineas de trigo con una altura media desigual. Llamaremos P_1 y P_2 a estas poblaciones, siendo la media aritmética de la característica en la población P_1 significativamente menor que la correspondiente media aritmética en P_2. En principio, podemos pensar que la herencia de ésta característica se debe al efecto de k loci, donde k puede ser cualquier entero igual o mayor que uno. En adelante denotaremos por k al número de loci relevantes.

Nuestra primera hipótesis es la siguiente: los individuos de ambas poblaciones son genotípicamente homogéneos y homocigotos para los k loci en cuestión. Así pues el genotipo de los individuos de P_1 puede representarse como

$$a_1a_1, a_2a_2, \ldots a_ka_k \qquad (1.2.1)$$

donde a_j representa al gen que se encuentra en el j-esimo loci, para $j = 1, 2, \ldots k$.

Consecuentemente representaremos el genotipo de los individuos de P_2 como

$$A_1A_1, A_2A_2, \ldots A_kA_k \qquad (1.2.2)$$

donde A_j representa, como antes, el gen que se encuentra en el j-esimo loci, para $j = 1, 2, \ldots k$.

Lo anterior implica las primeras dos hipótesis, a saber:

H1: Las poblaciones P_1 y P_2 son genéticamente homogéneas, esto es: todos los individuos de P_1 tienen el mismo genotipo (representado por 1.2.1) y todos los individuos de P_2 tienen el mismo genotipo (representado por 1.2.2).

H2: Todos los individuos de P_1 y P_2 son homocigotos para los k loci.

Las anteriores aseveraciones resultan biológicamente factibles en el caso de que se utilicen poblaciones consanguíneas de animales o "líneas puras" de plantas, ya que en ambos casos se supone que existe muy poca variabilidad genética dentro de estas poblaciones.

Al cruzar entre si individuos de las dos poblaciones (P_1 y P_2) obtendremos la primera generación filial, que denotaremos como F_1; lo anterior se expresa en el siguiente diagrama de cruza.

POBLACION:	P_1	$\times$	P_2
GENOTIPOS:	$a_1a_1, a_2a_2, \ldots a_ka_k$		$A_1A_1, A_2A_2, \ldots A_kA_k$
GAMETOS:	$a_1, a_2, \ldots a_k$		$A_1, A_2, \ldots A_k$
POBLACION RESULTANTE:	F_1		
GENOTIPO:	$A_1a_1, A_2a_2, \ldots A_ka_k$		(1.2.3)

DIAGRAMA 1.2.1

Si la población F_1 se reproduce, ya sea por autofecundación o bien al cruzarse entre si los individuos que la integran, se producirá la segunda generación filial, población que aquí denotaremos por F_2. La proporción en que se presentan los gametos de F_1 y por tanto las proporciones de los genotipos de F_2 dependen de la distribución física de los k loci en los cromosomas de la especie de que se trate. Lo más sencillo, aún cuando no siempre realista, es suponer que los k loci se heredan en forma independiente, esto es que no se encuentran ligados; de manera que la probabilidad de recombinación entre dos loci cualesquiera es de ½. En otras palabras, estamos suponiendo que los loci siguen la ley de la segregación de caracteres de Mendel. Lo anterior se formaliza en la

siguiente hipótesis:

H3: Los k loci implicados en la herencia de la característica en cuestión se heredan independientemente.

Puesto que los k loci en los individuos de la poblacion F_1 estan en estado heterocigótico ($A_j a_j$) el gameto puede contener el gen A_j o bien el gen a_j, y esto para cada j-ésimo locus, j = 1, 2, ... k; entonces existen

$$2 \times 2 \times \ldots \times 2 \ (k \text{ veces}) = 2^k$$

gametos diferentes y por la hipótesis H3 tenemos que cada uno de ellos tiene la misma probabilidad de formarse, a saber $(\frac{1}{2})^k$.

Con respecto al número de genotipos que se pueden producir, tenemos que para cada j-ésimo locus se pueden producir tres genotipos diferentes:

$$A_j A_j \ , \ A_j a_j \ \text{ y } \ a_j a_j \ ,$$

de manera que el número total de genotipos posibles es igual a

$$3 \times 3 \times \ldots \times 3 \ (k \text{ veces}) = 3^k$$

En el diagrama 1.2.2 se presenta la obtención de F_2.

En la práctica de los métodos de selección es común que se realicen retrocruzas, esto es, cruzamientos de los individuos progenitores con individuos de la población F_1. Así pues definimos a la poblacion R_1 como los individuos resultado de la cruza de F_1 con P_1 y a R_2 como los individuos resultado de la cruza de F_1 con P_2. En los diagramas 1.2.3 y 1.2.4 se muestra la obtención de estas poblaciones. Es importante hacer notar que en estos casos P_1 y P_2 producen un solo tipo de gameto.

POBLACION:	F_1	$\times$	F_1
GENOTIPOS:	$A_1a_1, A_2a_2, \ldots A_ka_k$		$A_1a_1, A_2a_2, \ldots A_ka_k$
GAMETOS:	1º $A_1, A_2, \ldots A_k$		1º $A_1, A_2, \ldots A_k$
	2º $A_1, A_2, \ldots a_k$		2º $A_1, A_2, \ldots a_k$
	.		.
	.		.
	.		.
	2^kº $a_1, a_2, \ldots a_k$		2^kº $a_1, a_2, \ldots a_k$

POBLACION RESULTANTE: F_2

GENOTIPOS:

1º $A_1A_1, A_2A_2, \ldots A_kA_k$

2º $A_1A_1, A_2A_2, \ldots A_ka_k$

.
.
.

3^kº $a_1a_1, a_2a_2, \ldots a_ka_k$

DIAGRAMA 1.2.2

POBLACIONES:	P_1	F_1
GENOTIPOS:	$a_1a_1, a_2a_2, \ldots a_ka_k$	$A_1a_1, A_2a_2, \ldots A_ka_k$
GAMETOS:	$a_1, a_2, \ldots a_k$	1º $A_1, A_2, \ldots A_k$
		2º $A_1, A_2, \ldots a_k$
		.
		.
		.
		2^kº $a_1, a_2, \ldots a_k$

POBLACION RESULTANTE: R_1

GENOTIPOS:

1º $A_1a_1, A_2a_2, \ldots A_ka_k$

2º $A_1a_1, A_2a_2, \ldots a_ka_k$

.

.

.

2^kº $a_1a_1, a_2a_2, \ldots a_ka_k$

DIAGRAMA 1.2.3

POBLACIONES:	P_2	F_1
GENOTIPOS:	$A_1A_1, A_2A_2, \ldots A_kA_k$	$A_1a_1, A_2a_2, \ldots A_ka_k$
GAMETOS:	$A_1, A_2, \ldots A_k$	$1º\ A_1, A_2, \ldots A_k$
		$2º\ A_1, A_2, \ldots a_k$
		$\vdots$
		$2^k º\ a_1, a_2, \ldots a_k$

POBLACION RESULTANTE: R_2

GENOTIPOS:

$1º\ A_1A_1, A_2A_2, \ldots A_kA_k$

$2º\ A_1A_1, A_2A_2, \ldots A_ka_k$

$\vdots$

$2^k º\ A_1a_1, A_2a_2, \ldots A_ka_k$

DIAGRAMA 1.2.4

CAPITULO 2

EL MODELO ESTADISTICO

Para el objetivo de éste trabajo, podemos considerar que una variable aleatoria es una función con dominio en un conjunto llamado "espacio muestral", y contradominio en los números reales. Es decir, una variable aleatoria asigna un número real a cada elemento de un conjunto que consta de todos los resultados posibles de un experimento; a éste conjunto se le da el nombre de "espacio muestral" (Sanin, 1984). Un modelo estadístico puede definirse entonces como un conjunto de medidas de probabilidad sobre el espacio muestral. En nuestro caso estamos interesados en variables aleatorias definidas sobre el espacio muestral que forman los conjuntos de individuos (poblaciones) presentados en el capítulo anterior, es decir, estamos interesados en las mediciones de la característica de interés en los diferentes grupos genéticos: P_1, P_2 F_1, F_2, R_1 y R_2. La característica de interés puede ser cualquier medición que se haga en los individuos de la población, como por ejemplo medidas de peso, longitud, etc. Nuestro interés aquí es encontrar, o definir, el conjunto de medidas de probabilidad que se ajuste al modelo genético presentado en el capítulo anterior. Es decir, queremos encontrar las variables aleatorias que nos permitan describir e interpretar ese modelo.

En Genética se considera que el valor fenotípico, esto es el valor observable de la característica de interés, puede ser explicado por una expresión de la forma

$$X = G + \varepsilon, \qquad (2.0.1)$$

donde X denota la medida de la característica observable o valor fenotípico; G representa el valor genotípico y ε la desviación por efectos medioambientales o "error experimental". Esta expresión presupone que el valor genotípico y los efectos medioambientales son independientes (ver por ejemplo Kempthorne, 1969 ó Bulmer, 1985).

2.1 Variables Aleatorias Implicadas

Según el principio fundamental de la Genética Cuantitativa, la herencia de una característica continua puede ser explicada por el efecto de un número k de loci. En nuestro caso esto quiere decir que el genotipo de los individuos, y mas específicamente cada gen en particular, tiene efecto sobre la característica de interés. en adelante denotaremos al gen y a su respectivo efecto genotípico con el mismo símbolo: A_j y a_j para j = 1, 2, ... k. El significado de estos símbolos se aclara por el contexto.

El efecto G es el producto de la interacción de todos los genes, y por tanto debemos especificar la forma funcional de G en nuestro modelo. En principio G puede depender de una manera mas o menos complicada de los efectos génicos. Sin embargo, la evidencia experimental disponible muestra, que en la mayoría de los casos la hipótesis de que G es una forma lineal produce resultados satisfactorios, en el sentido de que el ajuste observado es estadistícamente bueno (Kempthorne, 1969).

Enseguida, veremos cual es la expresión del efecto fenotípico en los casos de interés cuando, en cada caso G es una función lineal de los efectos génicos.

Primero veremos el caso de la variable aleatoria $X_1 : P_1 \to \mathbb{R}$, - la cual representa la medida de la característica de interés en la po - blación P_1. recuerde que el genotipo de los individuos de P_1 es

$$a_1 a_1 , \; a_2 a_2 , \; \cdots \; a_k a_k$$

(vea 1.2.1) y que X_1 está influenciada también por el medio ambiente, - podemos escribir

$$X_1 = 2 \sum_{j=1}^{k} a_j + \varepsilon_1 \qquad (2.1.1)$$

donde a_j representa el efecto del j-ésimo gen, $j = 1, 2, \ldots$; (de esta forma $2a_j$ será el efecto total del j-ésimo locus) y ε_1 denota el efecto medioambiental que altera la expresión del genotipo.

Continuando con la obtención de las variables aleatorias impli_ cadas, denotaremos por $X_2 : P_2 \to \mathbb{R}$ a la medida de la característica de interés en la población P_2. Así pues, recordando que el genotipo de los individuos de P_2 (dado en 1.2.2) es

$$A_1 A_1 , \; A_2 A_2 , \; \cdots \; A_k A_k$$

podemos escribir

$$X_2 = 2 \sum_{j=1}^{k} A_j + \varepsilon_2 \qquad (2.1.2)$$

donde A_j representa el efecto del j-ésimo gen, $j = 1, 2, \ldots$; y como antes ε_2 denota el efecto medioambiental.

Como se mencionó en el planteamiento del modelo genético, se - supone que la medida de la característica de interés en la población P_1 es menor que en la población P_2. Aquí se hace algo más, se supone que - todos los genes que aumentan la característica de interés se encuentran reunidos en el genotipo de los individuos de P_2, esto es, se plantea la siguiente hipótesis

blación P_1. recuerde que el genotipo de los individuos de P_1 es

$$a_1 a_1 , \; a_2 a_2 , \; \ldots \; a_k a_k$$

(vea 1.2.1) y que X_1 está influenciada también por el medio ambiente, - podemos escribir

$$X_1 = 2 \sum_{j=1}^{k} a_j + \varepsilon_1 \qquad (2.1.1)$$

donde a_j representa el efecto del j-ésimo gen, j = 1, 2, ... ; (de esta forma $2a_j$ será el efecto total del j-ésimo locus) y ε_1 denota el efecto medioambiental que altera la expresión del genotipo.

Continuando con la obtención de las variables aleatorias implicadas, denotaremos por $X_2 : P_2 \to \mathbb{R}$ a la medida de la característica de interés en la población P_2. Así pues, recordando que el genotipo de los individuos de P_2 (dado en 1.2.2) es

$$A_1 A_1 , \; A_2 A_2 , \; \ldots \; A_k A_k$$

podemos escribir

$$X_2 = 2 \sum_{j=1}^{k} A_j + \varepsilon_2 \qquad (2.1.2)$$

donde A_j representa el efecto del j-ésimo gen, j = 1, 2, ... ; y como antes ε_2 denota el efecto medioambiental.

Como se mencionó en el planteamiento del modelo genético, se - supone que la medida de la característica de interés en la población P_1 es menor que en la población P_2. Aquí se hace algo más, se supone que - todos los genes que aumentan la característica de interés se encuentran reunidos en el genotipo de los individuos de P_2, esto es, se plantea la siguiente hipótesis

Primero veremos el caso de la variable aleatoria $X_1 : P_1 \to \mathbb{R}$, - la cual representa la medida de la característica de interés en la po - blación P_1. recuerde que el genotipo de los individuos de P_1 es

$$a_1 a_1 , \; a_2 a_2 , \; \ldots \; a_k a_k$$

(vea 1.2.1) y que X_1 está influenciada también por el medio ambiente, - podemos escribir

$$X_1 = 2 \sum_{j=1}^{k} a_j + \varepsilon_1 \qquad (2.1.1)$$

donde a_j representa el efecto del j-ésimo gen, j = 1, 2, ... ; (de esta forma $2a_j$ será el efecto total del j-ésimo locus) y ε_1 denota el efecto medioambiental que altera la expresión del genotipo.

Continuando con la obtención de las variables aleatorias impli cadas, denotaremos por $X_2 : P_2 \to \mathbb{R}$ a la medida de la característica de interés en la población P_2. Así pues, recordando que el genotipo de los individuos de P_2 (dado en 1.2.2) es

$$A_1 A_1 , \; A_2 A_2 , \; \ldots \; A_k A_k$$

podemos escribir

$$X_2 = 2 \sum_{j=1}^{k} A_j + \varepsilon_2 \qquad (2.1.2)$$

donde A_j representa el efecto del j-ésimo gen, j = 1, 2, ... ; y como antes ε_2 denota el efecto medioambiental.

Como se mencionó en el planteamiento del modelo genético, se - supone que la medida de la característica de interés en la población P_1 es menor que en la población P_2. Aquí se hace algo más, se supone que - todos los genes que aumentan la característica de interés se encuentran reunidos en el genotipo de los individuos de P_2, esto es, se plantea la siguiente hipótesis

H4: Para cada $j = 1, 2, \ldots$, se tiene que $A_j > a_j > 0$.

El hecho de poner $a_j > 0$ se debe a que generalmente las características medidas en seres vivos son positivas, además ésta hipótesis se hace por simplicidad, pero los resultados del capítulo cuatro dependen solamente de que $A_j > a_j$.

Observación: es interesante notar que la hipótesis H4 excluye el fenómeno conocido como "variación transgresiva". Este fenómeno, que en ocasiones se manifiesta en la práctica genética, consiste en que individuos de la generación F_2 presenten la característica de interés en un grado mayor o menor que cualquiera de sus progenitores. Generalmente esto se puede explicar por el hecho de que no todos los genes que aumentan la característica (A_j en nuestro caso) estan reunidos en el genotipo de uno de los progenitores, sino que ambos progenitores poseen algunos de estos genes. Claramente la anterior hipótesis (H4) deja fuera del modelo la posibilidad de éste tipo de variación.

Al obtener la generación F_1, producto de la crúza de P_1 y P_2 puede suceder que la media aritmética de la característica de interes en F_1 sea intermedia a las de P_1 y P_2. Esto confirmará que los efectos de los genes son aditivos, lo cual se formaliza en la siguiente hipótesis.

H5: Los efectos de los genes son totalmente aditivos.

De no cumplirse lo anterior se puede suponer que existen efectos de dominancia y/o espistasis. Aquí se estudia el caso de aditividad total (vea el capítulo cínco para algunos comentarios de los casos en que existen efectos de dominancia o epistasis).

Denotaremos por $X_3 : F_1 \to \mathbb{R}$ a la medida de la característica de interés en los individuos de F_1. Recordando que el genotipo de estos es

$$A_1a_1,\ A_2a_2,\ \ldots\ A_ka_k,$$

(vea diagrama 1.2.1), podemos escribir

$$X_3 = \sum_{j=1}^{k} (A_j + a_j) + \varepsilon_3 \qquad (2.1.3)$$

donde como antes ε_3 representa los efectos medioambientales.

En principio los efectos de los k loci pueden ser diferentes, sin embargo el tratamiento posterior del modelo se simplifica mucho bajo la siguiente condición

H6: $A_1 = A_2 = \ldots = A_k = A$; $a_1 = a_2 = \ldots = a_k = a$.

De este modo estamos suponiendo que el efecto de cada uno de los loci es el mismo. Esta misma hipótesis es propuesta por Mather y Jinks (1977), otros autores como simplificación en el problema que nos ocupa (ver Mayo, 1980).

Con ésta hipótesis (H6) las expresiones para X_1, X_2 y X_3 se simplifican quedando como sigue

$$X_1 = 2ka + \varepsilon_1, \qquad (2.1.4)$$

$$X_2 = 2kA + \varepsilon_2, \qquad (2.1.5)$$

$$X_3 = k(A + a) + \varepsilon_3. \qquad (2.1.6)$$

Dada la anterior simplificación, pasaremos a expresar la medida de la característica de interés en la población F_2. Como se vio en el diagrama 1.2.2, los gametos que produce la F_1 pueden tener 0, 1, ... k genes A. El número de genes con efecto A en estos gametos es una variable aleatoria con distribución binomial de parámetros k, $\frac{1}{2}$.

Para ver esto llamemos Q al número de genes A que contiene un gameto,

estamos interesados en $P[Q = m]$ para $m = 0, 1, 2, \ldots k$. Si tomamos en cuenta que en cada uno de los k loci de F_1 se encuentran los genes A y a y dado que el proceso de meiosis (foramación de gametos) escoge al - azar a uno de estos genes para formar parte del gameto, y puesto que - los k loci son independientes (por H3), es claro que el número de for - mas (o arreglos) en que un gameto puede contener m genes A (para $m = 0$, 1, 2, ... k) estará dado por el coeficiente binomial

$$\binom{k}{m} = \frac{k!}{m!\,(k - m)!} ; \qquad (2.1.7)$$

(ver por ejemplo Parzen, 1979). Como se mencionó en la sección 1.2, el número de gametos diferentes que se pueden formar en los individuos de F_1 es 2^k, y por H3 cada uno de ellos tiene la misma oportunidad de formarse, a saber $(\frac{1}{2})^k$. De éste modo concluimos que

$$P[Q = m] = (\tfrac{1}{2})^k \binom{k}{m} \quad ; m = 0, 1, \ldots k. \qquad (2.1.8)$$

que es precisamente la función de probabilidad de la distribución binomial con k ensayos y probabilidad de éxito igual a $\frac{1}{2}$. Como es usual escribiremos

$$Q \sim B(k; \tfrac{1}{2})$$

Los individuos de F_2 son el producto de la unión de dos game - tos de F_1 (vea el diagrama 1.2.2), y cada uno de estos gametos lleva un número $Q = m$ de genes con efecto A, por lo tanto el número de genes A que posea el individuo de F_2 será la suma del número de genes A de cada uno de los gametos. Es decir, si denotamos por Q_1 al número de genes en los individuos de F_2 y como Q, Q' al número de genes A que tienen cada uno de los dos gametos de F_1, tendremos que

$$Q_1 = Q + Q'$$

y puesto que Q y Q' son variables aleatorias independientes tales que

$$Q \sim B(k;\tfrac{1}{2}),$$

$$Q' \sim B(k;\tfrac{1}{2}),$$

tenemos que

$$Q_1 \sim B(2k;\tfrac{1}{2}).$$

Lo anterior debido a que la suma de variables aleatorias independientes con distribución binomial resulta en una variable aleatoria con distribución también binomial con parámetros como se indica (ver - Feller, 1986).

Ahora note que el número de genes a en los individuos de F_2 es

$$2k - Q_1$$

y entonces, si denotamos mediante $X_4 : F_2 \to \mathbb{R}$ a la variable aleatoria - definida como la medida de la característica de interés en la población F_2 , tenemos que

$$X_4 = Q_1A + (2k - Q_1)a + \varepsilon_4$$

$$= 2ka + Q_1(A - a) + \varepsilon_4 \qquad (2.1.10)$$

donde ε_4 es el efecto medioambiental.

Como se vio anteriormente las poblaciones R_1 y R_2 se forman - por cruzas entre $P_1 \times F_1$ y $P_2 \times F_1$ respectivamente. En ambas cruzas la segregación se presenta solamente en los gametos de F_1. Así pues, los - individuos de R_1 tendrán k genes a, procedentes de P_1, mientras que el número de genes A que reciben de F_1 es, como antes, una variable aleatoria con distribución binomial, que aquí denotaremos por Q_2, y tal que

$$Q_2 \sim B(k;\tfrac{1}{2}) \qquad (2.1.11)$$

Por tanto, si denotamos por X_5 a la medida de la característica de -

interés en la población R_1, definida por

$$X_5 = ka + Q_2 A + (k - Q_2)a + \varepsilon_5$$

$$= 2ka + Q_2(A - a) + \varepsilon_5 \qquad (2.1.12)$$

donde ε_5 denota los efectos medioambientales.

Similarmente los individuos de R_2 tendrán k genes con efecto A que recibirán de P_2, mientras que el número de genes A que reciban de F_1 es, como en el caso anterior, una variable aleatoria que denotaremos por Q_3 y que claramente satisface

$$Q_3 \sim B(k;\tfrac{1}{2}) \qquad (2.1.13)$$

Por lo tanto, denotando como $X_6 : R_2 \to \mathbb{R}$ a la medida de la característica de interés en la población R_2 tenemos que

$$X_6 = kA + Q_3 A + (k - Q_3)a + \varepsilon_6$$

$$= k(A + a) + Q_3(A - a) + \varepsilon_6 \qquad (2.1.14)$$

donde como es costumbre ε_6 denota los efectos medioambientales.

En los párrafos anteriores hemos determinado la forma funcional del valor genotípico (G en la ecuación 2.0.1), ahora nos resta determinar la función de densidad de las variables ε_i, i = 1, 2, ... 6; que representan los efectos medioambientales sobre la característica de interés. Estos efectos son el resultado de la interacción de un gran número de factores y generalmente se ha visto que se ajustan adecuadamente a una distribución normal (Elandt-Johnson, 1971). Lo anterior se formaliza en la siguiente hipótesis.

H7: para i = 1, 2, ... 6 se tendrá que $\varepsilon_i \sim N(0;\sigma^2)$ donde $\sigma^2 \in \mathbb{R}^+$.

La anterior hipótesis es de utilización muy general en Genética Estadística, particularmente en la prueba de hipótesis mediante

análisis de varianza (Kempthorne, 1969).

Al final del capítulo uno se mencionaron los parámetros que determinan la distribución de las variables implicadas. Es importante hacer notar que con las hipótesis presentadas en esta sección (H4 a H7), hemos especificado los parámetros que determinan estas distribuciónes. Así, H4 implica que todos los genes que aumentan la característica estan reunidos en P_2, H5 determina la completa aditividad de los efectos génicos y por lo tanto deja fuera del modelo los posibles efectos de dominancia y epistasis. H6 simplifica el tratamiento del modelo y H7 especifica la distribución de los efectos medioambientales, implicando con ello la independencia de estos con el valor genotípico.

En la siguiente sección se presentan las funciones de densidad de las variables implicadas, así como sus esperanzas y varianzas.

2.2 Funciones de densidad.

Como se mencionó al principio de éste capítulo, podemos considerar un modelo estadístico como una colección de medidas de probabilidad P_θ, donde θ es el vector de parámetros que determinan estas medidas, es decir, que determinan las funciones de densidad de las variables. El vector de parámetros en nuestro caso esta dado por

$$\theta = (k, A, a, \sigma^2) \qquad (2.2.1)$$

donde cada uno de los componentes tiene el siguiente significado:

i) k, el número de loci involucrados, que es un entero igual o mayor a uno.

ii) A y a, los efectos génicos, que son números reales sujetos a la condición $A > a > 0$.

iii) σ^2, la varianza de los efectos medioambientales, la cual pertenece al conjunto de los números reales positivos.

Con lo anterior se define el espacio de parámetros θ como

$$\theta = \{ \theta = (k, A, a, \sigma^2) \mid k = 1, 2, \ldots ; A > a > 0 ; \sigma^2 > 0 \} \qquad (2.2.2)$$

Ahora vamos a encontrar las densidades de las variables $X_1, \ldots, X_6$. Para esto usaremos la siguiente notación: para una variable aleatoria W, la correspondiente función de densidad cuando $\theta \in \Theta$ es el verdadero valor del parámetro se denotará por $f_W(\cdot;\theta)$, mientras que el símbolo $\phi(y;\mu;\sigma^2)$ denotará a la función de densidad normal unidimensional con media μ y varianza σ^2, es decir

$$\phi(y;\mu;\sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp(-((y-\mu)^2/2\sigma^2)); \; y \in \mathbb{R}, \; \mu \in \mathbb{R}, \; \sigma^2 \in \mathbb{R}^+$$

(ver por ejemplo Mood et al; 1974).

En el siguiente teorema se presentan las funciones de densidad de cada una de las variables aleatorias definidas en la sección precedente. Algunas de las propiedades de estas funciones serán útiles para decidir sobre la existencia de estimadores insesgados de k, tema que se trata en el capítulo cuatro.

TEOREMA 2.1.1

Considere las variables X_i, i = 1, 2, ... 6 definidas en la sección 2.1 de este capítulo. Para cada $\theta \in \Theta$, las correspondientes densidades estan dadas como sigue

i) $f_{X_1}(y;\theta) = \phi(y; 2ka; \sigma^2)$,

o equivalentemente

$$X_1 \sim N(2ka; \sigma^2).$$

ii) $f_{X_2}(y;\theta) = \phi(y;\ 2kA;\ \sigma^2)$,

o equivalentemente

$$X_2 \sim N(2kA;\ \sigma^2).$$

iii) $f_{X_3}(y;\theta) = \phi(y;\ k(A + a);\ \sigma^2)$.

o equivalentemente

$$X_3 \sim N(k(A + a);\ \sigma^2)$$

iv) $f_{X_4}(y;\theta) = (\frac{1}{2})^{2k} \sum_{s_1=0}^{2k} \binom{2k}{s_1} \phi(y;\ 2ka + s_1(A - a);\ \sigma^2)$.

v) $f_{X_5}(y;\theta) = (\frac{1}{2})^{k} \sum_{s_2=0}^{k} \binom{k}{s_2} \phi(y;\ 2ka + s_2(A - a);\ \sigma^2)$.

vi) $f_{X_6}(y;\theta) = (\frac{1}{2})^{k} \sum_{s_3=0}^{k} \binom{k}{s_3} \phi(y;\ k(A + a) + s_3(A - a);\ \sigma^2)$.

todo lo anterior para $y \in \mathbb{R}$, $\theta \in \Theta$. //

DEMOSTRACION.

i) Es claro que X_1 tiene distribución normal, puesto que ε_1 se distribuye normalmente, y X_1 es la suma de una constante (2ka) y ε_1. Como la media y la varianza de ε_1 son 0 y σ^2 respectivamente, el resultado es inmediato (ver Brunk, 1965). Los incisos (ii) y (iii) se pueden demostrar en forma similar.

Para demostrar (iv) recordemos que X_4 fué definida en 2.1.10 - como

$$X_4 = 2ka + Q_1(A - a) + \varepsilon_4,$$

donde

$$Q_1 \sim B(2k;\tfrac{1}{2}), \quad \varepsilon_4 \sim N(0;\ \sigma^2).$$

Entonces, por el teorema de la probabilidad total (ver Mood et al., - 1974) se tiene

$$P_\theta[X_4 < y] = \sum_{s_1=0}^{2k} P_\theta[X_4 < y \mid Q_1 = s_1]\ P_\theta[Q_1 = s_1] \qquad (2.2.3)$$

Debido a la independencia de Q_1 y ε_4, la distribución condicional de X_4 dado un valor de Q_1, digamos s_1, es normal con media

$$2ka + s_1(A - a)$$

y varianza σ^2. Luego

$$P_\theta[X_4 < y \mid Q_1 = s_1] = \int_{-\infty}^{y} \phi(w;\ 2ka + s_1(A - a);\ \sigma^2)dw;$$

por otro lado sabemos que

$$P_\theta[Q_1 = s_1] = (\tfrac{1}{2})^k \binom{k}{s_1}$$

y entonces substituyendo estos valores en la expresión (2.2.3), y derivando con respecto a y, vemos que la función de densidad de X_4 es la indicada en el enumerado (iv) del teorema.

Finalmente (v) y (vi) se demuestran de forma similar. //

En adelante los símbolos E_θ y V_θ denotarán los operadores esperanza y varianza respectivamente, evaluados bajo la condición de que $\theta \in \Theta$ es el verdadero valor del parámetro.

El siguiente teorema presenta las esperanzas y varianzas de las variables implicadas en el modelo. Estas se utilizarán en el capítulo siguiente para obtener estimadores del número de loci.

TEOREMA 2.2.2

Bajo las hipótesis H_i, i = 1, 2, ... 7, lo siguiente ocurre

i) $$E_\theta[X_1] = 2ka,$$

$$V_\theta[X_1] = \sigma^2.$$

ii) $$E_\theta[X_2] = 2kA,$$

$$V_\theta[X_2] = \sigma^2.$$

iii) $$E_\theta[X_3] = k(A + a),$$

$$V_\theta[X_3] = \sigma^2.$$

iv) $$E_\theta[X_4] = k(A + a),$$

$$V_\theta[X_4] = \tfrac{1}{2}k(A - a)^2 + \sigma^2.$$

v) $$E_\theta[X_5] = k(\tfrac{1}{2}A + \tfrac{3}{2}a),$$

$$V_\theta[X_5] = \tfrac{1}{4}k(A - a)^2 + \sigma^2.$$

vi) $$E_\theta[X_6] = k(\tfrac{3}{2}A + \tfrac{1}{2}a),$$

$$V_\theta[X_6] = \tfrac{1}{4}k(A - a)^2 + \sigma^2. \ //$$

DEMOSTRACION

Las partes (i), (ii) y (iii) son claras, puesto que

$$\varepsilon_i \sim N(0;\sigma^2) \ , \ i = 1, 2, \ldots 6;$$

y el componente genético en los tres casos es una constante, a saber, 2ka, 2kA y k(A + a) respectivamente.

Para ver que (iv) se satisface recordemos que los componentes genético y medioambiental son variables aleatorias independientes. Ahora, debido a que

$$Q_1 \sim B(2k;\tfrac{1}{2}),$$

obtenemos que

$$E_\theta[Q_1] = k \quad y \quad V_\theta[Q_1] = \tfrac{1}{2}k$$

y por lo tanto

$$\begin{aligned} E_\theta[X_4] &= E_\theta[2ka + Q_1(A - a) + \varepsilon_4] \\ &= 2ka + E_\theta[Q_1](A - a) + E_\theta[\varepsilon_4] \\ &= 2ka + k(A - a) + 0 \\ &= k(A + a) \end{aligned}$$

Además

$$V_\theta[X_4] = V_\theta[2ka + Q_1(A - a) + \varepsilon_4]$$

$$= 0 + V_\theta[Q_1](A - a)^2 + V_\theta[\varepsilon_4]$$

$$= \tfrac{1}{2}k(A - a)^2 + \sigma^2.$$

Las restantes partes del teorema se demuestran en forma similar. //

Con la demostración del teorema 2.2.2 concluimos la presentación del modelo estadístico para el fenómeno de interés: la herencia de una característica cuantitativa. En el capítulo siguiente se presenta un conjunto de estimadores para k, y se demuestra que dichos estimadores son consistentes pero no tienen esperanza finita.

CAPITULO 3

ESTIMADORES PARA EL NUMERO DE LOCI

En los capítulos anteriores hemos presentado un modelo para el fenómeno de interés: la herencia de una característica cuantitativa. Sin embargo, lo único que podemos conocer con certeza del modelo son los valores de las observaciones de las variables aleatorias, esto es las mediciones de la característica. Para que el modelo tenga alguna utilidad es necesario calcular los parámetros que lo determinan y para ello tendremos que utilizar las observaciones. Solamente en casos triviales es posible tener certeza acerca del verdadero valor del parámetro después de un número finito de observaciones del fenómeno; en la mayoría de los casos tendremos que aceptar un cierto grado de incertidumbre en nuestro cálculo. Es por esto que al resultado del cálculo se le llama "estimación" y a la función de las observaciones que nos permite llegar al resultado se le conoce como "estimador" (ver Brunk, 1965). En principio existen un gran número de estimadores posibles, sin embargo algunos de ellos resultan deseables porque permiten reducir, o acotar, el grado de incertidumbre que se tiene sobre el verdadero valor del parámetro.

Como se ha mencionado anteriormente, el objetivo de este trabajo es estudiar la estimación del número de loci que afectan la herencia de una característica cuantitativa. A continuación se presentan algunos conceptos útiles a este respecto.

Existen dos formas de estimación conceptualmente diferentes: - la estimación puntual y la estimación por intervalos (ver por ejemplo, Rao, 1965). La primera de ellas tiene como resultado final una estima - ción que es un punto en el espacio de parámetros, en cambio en la esti- mación por intervalos el resultado final de la estimación es una región en el espacio de parámetros, para la que se tiene cierta "confianza" de que contenga al verdadero valor del parámetro. En este trabajo nos limi taremos a la primera opción, esto es a la estimación puntual de k.

Dado el problema de la estimación puntual existen varios méto- dos para encontrar estimadores, como son el método de máxima verosimili tud, mínima ji cuadrada, mínima distancia y el método de los momentos - (ver por ejemplo Mood et al., 1974). En este trabajo nos restringiremos a la obtención de estimadores de k por el método de los momentos.

En este capítulo se proponen algunos estimadores de k obteni - dos por el método de los momentos, se estudia su consistencia y se de - muestra el hecho de que no tienen esperanza finita, discutiendo algunas implicaciones de ello.

3.1 Estimadores de Momentos de k.

El método de los momentos consiste básicamente en substituir - los momentos poblacionales por los correspondientes momentos muestrales, obtenidos a partir de las observaciones. Así, una función parametral - que depende de los momentos poblacionales se estima a partir de la co - rrespondiente función evaluada en los momentos muestrales (ver por ejem plo Wilks, 1962).

Aquí utilizaremos el método de los momentos de la siguiente ma nera: primero expresaremos a k como función de medias y varianzas de -

las variables de interés y entonces se substituirán estos momentos po - blacionales por los correspondientes momentos muestrales (medias aritmé ticas y varianzas muestrales), para obtener estimadores de k.

En adelante supondremos que se cuenta con n observaciones de - cada una de las variables definidas en el capítulo anterior (X_1, X_2, ... X_6) y denotaremos por $\mathbf{X}_i$, $\mathbf{X}_i \in \mathbb{R}^{n \times 1}$, al vector de n observaciones - de la i-ésima variable; i = 1, 2, ... 6. Esto es

$$\mathbf{X}_i = \begin{pmatrix} X_{i_1} \\ X_{i_2} \\ \vdots \\ X_{i_n} \end{pmatrix} \quad ; \; i = 1, 2, \ldots 6,$$

donde X_{ij} representa la j-ésima observación de la variable i.

También definiremos los momentos muestrales $\overline{X}_i$ y S_i (media - aritmética y varianza muestral respectivamente) como

$$\overline{X}_i = \frac{1}{n} \sum_{j=1}^{n} X_{ij}$$

$$S_i = \frac{1}{n-1} \sum_{j=1}^{n} (X_{ij} - \overline{X})^2$$

para i = 1, 2, ... 6. Es un hecho bien conocido que $\overline{X}_i$ y S_i son estima dores insesgados de varianza uniformemente mínima para sus correspon - dientes momentos poblacionales (ver Mendenhall et al., 1986).

En el siguiente ejemplo se presentan funciones de esperanzas y varianzas que producen estimadores de k al aplicar el método de los mo- mentos.

EJEMPLO 3.1.1

i) Recordando que

$$E_\theta[X_1] = 2ka,$$

$$E_\theta[X_2] = 2kA,$$

$$V_\theta[X_4] = \tfrac{1}{2}k(A - a)^2 + \sigma^2,$$

y

$$V_\theta[X_3] = \sigma^2.$$

podemos ver que la siguiente identidad es válida para todo $\theta \in \Theta$.

$$k = \frac{(E_\theta[X_2] - E_\theta[X_1])^2}{8(V_\theta[X_4] - V_\theta[X_3])}$$

En efecto, usando las igualdades anteriores, se verifica que

$$\frac{(E_\theta[X_2] - E_\theta[X_1])^2}{8(V_\theta[X_4] - V_\theta[X_3])} = \frac{(2kA - 2ka)^2}{8(\frac{1}{2}k(A - a)^2 + \sigma^2 - \sigma^2)}$$

$$= \frac{k^2(A - a)^2}{k\,(A - a)^2}$$

$$= k.$$

Entonces, substituyendo en la expresión anterior a los momentos poblacionales (esperanzas y varianzas), por sus correspondientes estimadores muestrales $\bar{X}$ y S obtenemos un estimador de k, dado por

$$\frac{(\bar{X}_2 - \bar{X}_1)^2}{8\,(S_4^2 - S_3^2)}$$

ii) De forma similar, recordando que

$$V_\theta[X_5] = V_\theta[X_6] = \tfrac{1}{4}k(A - a)^2 + \sigma^2,$$

podemos ver que la función

$$\frac{(E_\theta[X_2] - E_\theta[X_1])^2}{\frac{1}{2}(V_\theta[X_5] + V_\theta[X_6] - 2V_\theta[X_3])}$$

es idéntica a k para todo $\theta \in \Theta$; en efecto

$$\frac{(E_\theta[X_2] - E_\theta[X_1])^2}{\frac{1}{2}(V_\theta[X_5] + V_\theta[X_6] - 2V_\theta[X_3])} = \frac{(2kA - 2ka)^2}{\frac{1}{2}(\frac{1}{4}k(A - a)^2 + \sigma^2 + \frac{1}{4}k(A - a)^2 + \sigma^2 - 2\sigma^2)}$$

$$= \frac{k^2(A - a)^2}{k\ (A - a)^2}$$

$$= k,$$

y aplicando el método de momentos obtenemos el estimador de k dado por

$$\frac{(\overline{X}_2 - \overline{X}_1)^2}{\frac{1}{2}(S_5^2 + S_6^2 - 2S_3^2)} \quad . \ //$$

Como lo muestra el ejemplo anterior, en nuestro caso el procedimiento para encontrar estimadores de momentos de k se reduce a buscar funciones de esperanzas y varianzas que resulten en la forma

$$\frac{K^2(A - a)^2}{k\ (A - a)^2} = k \qquad (3.1.1)$$

y entonces substituir las esperanzas y varianzas (poblacionales) por sus respectivos estimadores muestrales.

En este trabajo nos restringiremos a estudiar estimadores que sean función de dos medias aritméticas y dos varianzas muestrales, específicamente los estimadores

$$\hat{K}_h = \frac{(\overline{X}_i - \overline{X}_j)^2}{c_h\ (S_l^2 - S_m^2)} \qquad (3.1.2)$$

donde h = (i, j, l, m) y las parejas (i,j) y (l,m) toman valores en los siguientes conjuntos

(i,j) : (2,1), (3,1), (5,1), (6,1), (2,3), (2,5), (2,6), (3,5), (6,3), (6,5), (4,1), (2,4), (4,5), (6,4);

(l,m) : (4,3), (4,2), (4,1), (5,3), (5,2), (5,1), (6,3), (6,2), (6,1);

la constante $c_h \in \mathbb{R}$ se relaciona de modo que se cumpla la condición

$$\frac{(E_\theta[X_i] - E_\theta[X_j])^2}{c_h(V_\theta[X_l] - V_\theta[X_m])} = k \qquad (3.1.3)$$

3.2 Consistencia de los Estimadores Propuestos.

Hasta el momento hemos considerado al número de observaciones en un estimador como una constante n. Sin embargo, para estudiar algu - nas de sus características, podemos considerar la sucesión de estimado- res

$$T_1, T_2, \ldots T_n, T_{n+1}, \ldots$$

es decir, la sucesión de estimadores T con un número creciente de obser vaciones. Una de las propiedades deseables que puede tener un estimador es que al aumentar el número de observaciones (tamaño de la muestra) - sea más probable que el valor de la estimación esté cercano al verdade- ro valor del parámetro, o función paramétrica que se pretende estimar. En otras palabras, es deseable que el estimador "converga en probabili- dad"a la función que se quiere estimar. A ésta propiedad se le llama - "consistencia", y se define formalmente a continuación.

DEFINICION 3.2.1 (Estimador Consistente).

Sea $\{P_\theta \mid \theta \in \Theta\}$ un modelo con observaciones $Y = (Y_1, Y_2, \ldots Y_n)$ y $\tau(\theta)$ una función que se desea estimar. Una sucesión de estimadores

$$\{ T_n = T(Y, n) \}$$

de $\tau(\theta)$ se dice que es una sucesión consistente de estimadores de $\tau(\theta)$- si para cada número positivo δ se tiene que

$$\lim_{n \to \infty} P_\theta(|T_n - \tau(\theta)| \leq \delta) = 1$$

o en forma equivalente

$$\lim_{n \to \infty} P_\theta (|T_n - \tau(\theta)| > \delta) = 0 \quad .//$$

Intuitivamente esta definición nos dice que si tenemos un esti mador consistente y aumentamos indefinidamente el número de -

Observación.

Todas las combinaciones posibles (i,j), (l,m) (con i, j, l, m como en 3.1.2) y una constante adecuada resultan en un estimador de k, $\hat{k}_h$. Las diferentes combinaciones de (i,j) con (l,m) generan un total de 14 x 9 = 126 estimadores "distintos" de k, que son función exclusivamente de dos medias y dos varianzas muestrales. Esta lista es exhaustiva para éste tipo específico de estimadores, aún cuando existen muchos más que son o combinaciones de los anteriores o funciones de un número mayor de medias y varianzas (ver punto (ii) del ejemplo 3.1.1, o capítulo cinco). No obstante lo anterior, los resultados de las siguientes secciones de éste capítulo y del capítulo cuatro son muy generales. //

Algunos de los estimadores presentados en 4.1.2 se encuentran citados en la literatura consultada. Así por ejemplo, Mather, Jinks y Mayo presentan el estimador

$$\frac{(\bar{X}_2 - \bar{X}_1)^2}{8(S_4^2 - S_3^2)},$$

sin embargo, no hacen mención de sus propiedades estadísticas, salvo que coinciden en señalar que este estimador (y otros de su tipo) son sensibles al incumplimiento de las hipótesis planteadas. En otras palabras, cuando una o varias de las hipótesis no se cumplen, no se puede tener "confianza" en la estimación (Mather y Jinks, 1977; Mayo, 1980). En el capítulo cinco se discute mas ampliamente este punto.

Una faceta muy importante del problema de la estimación es el estudio de las propiedades de los estimadores, puesto que solamente conociendo dichas propiedades se podrá manejar adecuadamente la inferencia. En la siguientes secciones se estudian algunas propiedades de los estimadores propuestos.

obsrevaciones podremos estar "casi seguros" de que nuestra estimación está cercana al verdadero valor del parámetro.

Un hecho bien conocido es que los estimadores $\overline{X}$ y S^2 son consistentes (ver por ejemplo Mendenhall et al. , 1986).

A continuación se presenta un teorema sobre la consistencia de estimadores que son función de estimadores consistentes. Se omite la demostración, que puede encontrarse en varios textos de Estadística Matemática (por ejemplo Mendenhall et al., 1986).

TEOREMA 3.2.1

Sea T_n estimador consistente de $\tau(\theta)$ y T_n' estimador consistente de $\tau'(\theta)$, entonces

i) $T_n + T_n'$ es estimador consistente de $\tau(\theta) + \tau'(\theta)$.

ii) $T_n T_n'$ es estimador consistente de $\tau(\theta)\tau'(\theta)$.

iii) T_n / T_n' es estimador consistente de $\tau(\theta) / \tau'(\theta)$, siempre que $\tau'(\theta) \neq 0$. //

El hecho de que los estimadores $\overline{X}$ y S^2 son consistentes y el teorema anterior nos conducen al siguiente resultado

TEOREMA 3.2.2 (Sobre la consistencia de $\hat{R}_h$).

Los estimadores $\hat{R}_h$ propuestos en 4.1.2 para k, o equivalentemente, la sucesión

$$\hat{R}_{h,1} , \hat{R}_{h,2} , \ldots \hat{R}_{h,n} , \hat{R}_{h,n+1} , \ldots$$

converge en probabilidad a k. //

DEMOSTRACION

Notando que cada $\hat{R}$ es funciòn de estimadores consistentes ($\bar{X}_i$, $\bar{X}_j$, S_l^2, S_m^2), y aplicando reiteradamente el resultado del teorema 3.2.1 se demuestra el teorema. //

3.3 Inexistencia de la Esperanza de los Estimadores Propuestos.

En la sección 3.1 se presentó un conjunto de estimadores $\{\hat{R}_h\}$ del número de loci, y en la sección anterior se vió que éste conjunto posee la propiedad deseable de consistencia. Ahora se debería de continuar con la resolución del problema escogiendo uno o varios de estos estimadores como "mejores estimadores" del número de loci. Sin embargo, los posibles criterios de selección como el insesgamiento (ver capítulo cuatro para una definición) o la varianza uniformemente mínima (ver por ejemplo Mood et al., 1974), involucran la obtención de la esperanza de los estimadores y como demostraremos en esta sección, los estimadores propuestos en 3.1.2 no tienen esperanza finita.

Lo anterior quiere decir que la integral definida por la función

$$E_\theta[\,|\hat{R}_h|\,]$$

en general no converge, o dicho de otra manera, es infinita. Para ver esto escribiremos

$$E_\theta[\,|\hat{R}_h|\,] = E\left\{\frac{(\bar{X}_i - \bar{X}_j)^2}{|c_h(S_l^2 - S_m^2)|}\right\} \qquad (3.3.1)$$

Como el numerador y denominador de la ecuación anterior son independientes (ver por ejemplo Searle, 1971) podemos escribir la expresión anterior como

$$E_\theta[|R_h|] = \frac{1}{c_h} E_\theta[(\bar{X}_i - \bar{X}_j)^2]\, E_\theta[1/(|S_l^2 - S_m^2|)].$$

Veremos que el término

$$E_\theta[1/(|S_l^2 - S_m^2|)] ,$$

en general no es finito.

Recordando que S_m^2 es un estimador insesgado de σ^2 (esto es la varianza de los efectos medioambientales ε_i), mientras que S_l^2 es un estimador insesgado de $c_l k(A - a)^2 + \sigma^2$, donde c_l es una constante que depende solamente de la elección de $l = 4, 5, 6$. Así pues se sabe que $(n - 1)S_m^2$ tiene una distribución ji cuadrada centrada con $n-1$ grados de libertad, esto es

$$(n - 1)S_m^2 \sim \chi^2_{(n-1)}.$$

La distribución de S_l^2 es mas complicada, puesto que las variables X_l , $l = 4, 5, 6$; $i = 1, 2, \ldots n$ tienen una densidad que es una mezcla de variables aleatorias con distribución normal y variables con distribución binomial (ver sección 2.2). En general

$$X_{li} = 2ka + Q_{li}(A - a) + \varepsilon_i ,$$

y entonces el vector de observaciones $X_l \in \mathbb{R}^{n \times 1}$ lo podemos escribir como

$$X_l = 2ka\ \mathbb{1} + (A - a)Q_l + \varepsilon_l ,$$

donde el vector $\mathbb{1}$ es de orden $n \times 1$ y

$$Q_l = \begin{pmatrix} Q_{l1} \\ Q_{l2} \\ \vdots \\ Q_{ln} \end{pmatrix} , \quad \varepsilon_l = \begin{pmatrix} \varepsilon_{l1} \\ \varepsilon_{l2} \\ \\ \varepsilon_{ln} \end{pmatrix} .$$

Note que entonces

$$S_1^2 = X_1' (I - J / n) X_1$$

donde I es la matriz identidad de orden n x n y J es la matriz con to - dos sus componentes uno y de orden n x n. Entonces, si fijamos el valor de Q_1 en un punto particular q, la distribución de X_1 será

$$N(2ka\,1\!\!1 + (A - a)q;\ \sigma^2 I),$$

y entonces S_1^2 tendrá una distribución ji cuadrada no centrada, a saber

$$S_1^2 \mid Q_1 = q \sim \chi^2_{n-1,\lambda q},$$

donde el parámetro de no centralidad λq está dado por

$$\lambda q = \tfrac{1}{2}(2ka\,1\!\!1 + (A - a)q')(I - J/n)(2ka\,1\!\!1 + (A - a)q)$$

(ver Searle, 1971).

Denotemos por U a S_1^2, así pues hemos visto que la densidad de U dado $Q_1 = q$ es $\chi^2_{n-1,\lambda q}$, más explícitamente

$$f_{U|Q=q}(u|Q=q) = \exp(-\lambda q) \sum_{r=0}^{\infty} \frac{(\lambda q)^r\, u^{\frac{1}{2}n+r-2}\ \exp(-\frac{1}{2}u)}{r!\ 2^{\frac{1}{2}n+r+1}\ \Gamma(\frac{1}{2}(n-1)+r)},$$

por lo tanto la densidad de U está dada por

$$f_U(u;\theta) = \sum_q P_\theta[Q_1 = q]\ f_{U|Q=q}(u|Q_1 = q),$$

o sea una mezcla de densidades ji cuadrada no centradas, ponderadas por la probabilidad del evento $Q_1 = q$, para cada uno de los valores posi - bles de q.

Así pues tenemos que

$$E_\theta[1/(|S_1^2 - S_m^2|)] = E_\theta[1/(|U - V|)]$$

$$= \int_{u=0}^{\infty}\int_{v=0}^{\infty} 1/(|u - v|)\ \textstyle\sum_q P_\theta[Q=q]\ f_{U|Q}(u|q)\ f_V(v)\ du\ dw,$$

donde $f_V(v)$ es la densidad de $V = S_m^2$, o sea la densidad ji cuadrada centrada con n - 1 grados de libertad. Haciendo el cambio de variable - $w = u - v$, tenemos que la ecuación anterior se expresa como

$$\int_{v=0}^{\infty} \int_{w>-v}^{\infty} (1/|w|) \sum_q P_\theta[Q=q]\, f_{W+V}(w+v)\, f_V(v)\, dw\, dv =$$

$$\int_{v=0}^{\infty} \int_{w>-v} (1/|w|) \sum_q P_\theta[Q=q] \exp(-\lambda q) \sum_{r=0}^{\infty} \frac{(\lambda q)^r (w+v)^{\frac{1}{2}n+r-2} \exp(-\frac{1}{2}(w+v))}{r!\, 2^{\frac{1}{2}n+r-1}\, \Gamma(\frac{1}{2}(n-1)+r) 2^{\frac{1}{2}(n-1)}}$$

$$\frac{\exp(-\frac{1}{2}v)}{\Gamma(\frac{1}{2}(n-1))}\, dw\, dv =$$

$$\sum_q P_\theta[Q=q] \exp(-\lambda q) \sum_{r=0}^{\infty} (\lambda q)^r / (r! 2^{\frac{1}{2}(n+r-1)} \Gamma(\frac{1}{2}(n-1)+r)\, 2^{\frac{1}{2}(n-1)}\, \Gamma(\frac{1}{2}(n-1)))$$

$$\int_{v=0}^{\infty} \int_{w>-v}^{\infty} (1/|w|)(w+v)^{\frac{1}{2}(n+r-2)} \exp(-\tfrac{1}{2}(w+v)) v^{\frac{1}{2}(n-2)} \exp(-\tfrac{1}{2}v)\, dw\, dv \quad (3.3.2)$$

Es claro que esta serie de integrales diverge con tan solo que alguno - de sus términos diverja. Sin embargo, ahora veremos que cualquiera de - las integrales de la serie anterior es ∞. Para obtener este resultado - defina

$$I(v) = \int_{w>-v} (1/|w|)(w+v)^{\frac{1}{2}(n-2)} \exp(-\tfrac{1}{2}(w+v)) v^{\frac{1}{2}(n-2)} \exp(-\tfrac{1}{2}v)\, dw,\ v>0.$$

Entonces es claro que

$$I(v) \geq \int_{-v/2}^{v/2} (1/|w|)(w+v)^{\frac{1}{2}(n-2)} \exp(-\tfrac{1}{2}w-v)\, v^{\frac{1}{2}(n-2)} dw = I'(v)$$

y claramente

$$I'(v) \geq \int_{-v/2}^{v/2} (1/|w|)(\tfrac{1}{2}v)^{\frac{1}{2}(n-2)} \exp(-\tfrac{7}{4}v)\, v^{\frac{1}{2}(n-2)} dw$$

$$= (\tfrac{1}{2}v)^{\frac{1}{2}(n-2)} \exp(-7/4v)\, v^{\frac{1}{2}(n-2)} \int_{-v/2}^{v/2} (1/|w|)\, dw.$$

De ésta última expresión podemos confirmar la no existencia de la esperanza de $\hat{R}_h$, pues la integral

$$\int_{-v/2}^{v/2} (1/|w|)\, dw \ , \ v>0,$$

no converge. Luego, para todo $v>0$, $I(v)=\infty$ pues

$$I(v) \geq I'(v),$$

y ya que las integrales dobles de 3.3.2 se obtienen integrando $I(v)$ en $[0,\infty)$, vemos que la suma de integrales diverge.

El hecho de que los estimadores propuestos no tengan esperanza impide utilizar criterios ya establecidos para escoger el "mejor" de dichos estimadores. Por lo tanto poco se puede decir a este respecto. Solamente cabe señalar que dada la consistencia de los estimadores, el utilizar tamaños de muestra grandes resulta muy recomendable. Además, es posible usar estimadores que involucren todas las variables implicadas en el modelo. Esto se discute con mayor amplitud en el capítulo cinco.

CAPITULO 4

IMPOSIBILIDAD DE LA ESTIMACION INSESGADA DEL NUMERO DE LOCI

El hecho de que los estimadores propuestos no tengan esperanza así como la forma de las distribuciones de las variables de interes, inducen a pensar que en el caso que nos ocupa no existen estimadores de k cuya esperanza sea igual al valor del parámetro. A esta propiedad se le conoce como "insesgamiento", y es definida y ejemplificada en la siguiente sección.

4.1 Estimadores Insesgados.

Un estimador es una función conocida de las observaciones de una variable aleatoria y por lo tanto es en si mismo una variable aleatoria. La esperanza (o valor esperado) es un concepto muy útil en la solución de problemas que involucran variables aleatorias, pues se puede pensar en ella como un promedio ponderado de los posibles valores de la variable, donde los valores mas probables reciben mayor peso.

DEFINICION 4.1.1 (Estimador Insesgado).

Sea $\{P_\theta \mid \theta \in \Theta\}$ un modelo estadístico y $\tau: \Theta \to \mathbb{R}$ una función. Denote por $\mathbf{Y} = (Y_1, Y_2, \ldots Y_n)$ al vector de observaciones correspondientes al modelo. Una variable aleatoria $T = T(\mathbf{Y})$ se dice un estimador insesgado de $\tau(\theta)$ si y sólo si

$$E_\theta[T] = \tau(\theta) \text{ para todo } \theta \in \Theta. \quad //$$

Los estimadores insesgados resultan intuitivamente atractivos o deseables, pues como se ha dicho, la medida de tendencia central de su distribución coincide con la función paramétrica que se pretende estimar. Ademas, dadas las propiedades lineales del operador esperanza, = muchas veces resulta mas conveniente trabajar con estimadores insesgados que con aquellos que no lo son. Por ejemplo, si T_1 y T_2 son estimadores insesgados de $\tau(\theta)$ entonces cualquier combinación lineal

$$\alpha T_1 + \beta T_2 ,$$

donde α y β son constantes tales que $\alpha + \beta = 1$, será también un estimador insesgado de $\tau(\theta)$ (ver por ejemplo Brunk, 1965).

Es importante hacer notar que el criterio de insesgamiento permite en muchos casos restringir la búsqueda de estimadores a aquellos que tienen tal propiedad, y cuando existen de ellos se puede seleccionar al estimador de varianza uniformemente mínima (Rao, 1965).

No obstante, la existencia de estimadores insesgados no está de ninguna manera garantizada. El siguiente ejemplo presenta un caso, trivial pero ilustrativo, de la no existencia de estimadores insesgados para algunas funciones paramétricas.

EJEMPLO 4.1.1

En el experimento de lanzar una moneda y observar la cara que queda arriba, definamos la variable aleatoria Y como uno si la moneda muestra "águila" y cero de otro modo. De esta forma se tiene una variable aleatoria con distribución Bernoulli, con parametro p, donde p es la probabilidad de obtener águila. O, simbolicamente

$$Y \quad \text{Ber}(p), \; p \in (0,1).$$

Ahora supongamos que se cuenta con una muestra de cinco observaciones de la variable: $Y = (Y_1, Y_2, \ldots Y_5)$ y que se quiere encontrar un estimador insesgado de la función $\tau(p) = p^{11}$ (la probabilidad de once águilas seguidas). Veremos que no existe estimador insesgado de tal función. De hecho, supongamos que existe un estimador $T(Y)$ tal que

$$E_p[T(Y)] = p^{11}$$

pero en general se tiene que

$$E_p[T(Y)] = \sum_{y_1, \ldots y_5 = 0 \text{ó } 1} T(y)\, p^{\Sigma^5 y_i}\, (1-p)^{\Sigma^5 y_i}$$

Puesto que la expresión anterior es un polinomio en $\bar{p}$ de grado cinco, se concluye que no puede ser igual a $\tau(p)$, puesto que esta última es un polinomio de grado mayor. //

En el ejemplo anterior es claro que al aumentar el número de observaciones de la variable aleatoria es posible encontrar, aplicando alguno de los métodos conocidos, estimadores insesgados de $\tau(p)$; por supuesto no siempre es así, como queda demostrado en el siguiente ejemplo.

EJEMPLO 4.1.2

Considere una muestra $Y = (Y_1, Y_2, \ldots Y_n)$, $n \geq 1$ de la distribución **Ber**(p) y $\tau(p) = 1/(1-p)$. Entonces para cualquier $n \geq 1$, no existe ningun estimador insesgado de $\tau(p)$.

Para ver esto sea $T = T(Y)$ algún estadístico y note que

$$E_p[T(Y)] = \sum_{y_i = 0 \text{ ó } 1} T(Y)\, p^{\Sigma y_i}\, (1-p)^{n-\Sigma y_i},$$

y entonces $E_p[T(Y)]$ es un polinomio en p, de grado a lo mas n. Sin embargo $\tau(p) = 1/(1-p)$ no es un polinomio y por esta razón no es posible tener

$$E_p[T(Y)] = \tau(p) \text{ para todo } p \in [0,1),$$

es decir, ningún estadístico T puede ser estimador insesgado de $\tau(p)$.//

En el caso particular que nos ocupa la función paramétrica para la cual nos interesa obtener estimadores es

$$\tau(\theta) = k,$$

en otras palabras, nos interesa estimar el número de loci involucrados en la herencia de una característica de interés.

En principio estamos interesados en encontrar estimadores insesgados de k, sin embargo como demostraremos mas adelante, tales estimadores no existen. Este resultado, que aunque intuitivamente claro, requiere de una demostración no trivial y es nuestro principal resultado en este capítulo. En dicha demostración utilizaremos los conceptos de identificabilidad de una parametrización y estadístico completo. En la siguiente sección se presenta el primero de ellos.

4.2 Identificabilidad de una Parametrización.

DEFINICION 4.2.1 (Parametrización Identificable).

Para un modelo $\{P_\theta \mid \theta \in \Theta\}$ diremos que la parametrización

$$\theta \to P_\theta$$

es identificable si y sólo si para todo par $\theta_0, \theta_1 \in \Theta$, con $\theta_0 \neq \theta_1$ se tiene que

$$P_{\theta_0} \neq P_{\theta_1},$$

o equivalentemente

$$F_X(x; \theta_0) \neq F_X(x; \theta_1) \text{ para al menos un } x,$$

donde para cada $\theta \in \Theta$, $F(\cdot;\theta)$ es la función de distribución de P_θ.//

En otras palabras, esto quiere decir que una parametrización es identificable cuando a parámetros diferentes les corresponden siempre medidas de probabilidad diferentes en el modelo, y esto último es equivalente a decir que las funciones de distribución son distintas.

El siguiente teorema aclara y ejemplifica la importancia del concepto de identificabilidad en el problema de la estimación insesgada de funciones paramétricas.

TEOREMA 4.2.1

Sea $\{P_\theta \mid \theta \in \Theta\}$ un modelo y suponga que

$$\theta_0 \neq \theta_1 \text{ y } P_{\theta_0} = P_{\theta_1} .$$

Es decir, la parametrización no es identificable. entonces, existe una función $\tau : \Theta \to \mathbb{R}$ para la cual no existe estimador insesgado. //

DEMOSTRACION

Simplemente seleccione la función τ de modo que

$$\tau(\theta_0) \neq \tau(\theta_1), \qquad (4.2.1)$$

esto es posible ya que $\theta_0 \neq \theta_1$. Además, note que para cualquier estadístico $T = T(Y)$ con esperanza finita se tiene que

$$E_{\theta_0}[T] = E_{\theta_1}[T], \qquad (4.2.2)$$

lo anterior usando el hecho de que $P_{\theta_0} = P_{\theta_1}$. a partir de (4.2.1) y (4.2.2) vemos que no se puede tener

$$\tau(\theta) = E_\theta[T] \text{ para todo } \theta \in \Theta ,$$

y por lo tanto no existe estimador insesgado de $\tau(\theta)$. //

En el siguiente teorema veremos que no es posible obtener estimadores insesgados de k (el número de loci involucrados), cuando se

utilizan estimadores que dependen solamente de una observación de las - variables X_1 , X_2 , X_3 . En la demostración se utiliza el hecho de que para estas tres variables la parametrización definida en 2.1.13 no es - identificable.

TEOREMA 4.2.2

Para i = 1, 2, 3, no existe estimador $T(X_i)$ tal que

$$E_\theta[T(X_i)] = k \text{ para todo } \theta \in \Theta .$$

Es decir, no existe estimador insesgado de k que dependa solamente de - las primeras tres variables del modelo.

DEMOSTRACION

Denotemos por $P_{X_1,\theta}$ la distribución de X_1 correspondiente al vector de parámetros θ . Entonces, como se vio en el teorema 2.2.1

$$P_{X_1,\theta} = N(2ka; \sigma^2).$$

Ahora, si tomamos dos valores diferentes de θ , digamos

$$\theta_0 = (k_0, A_0, a_0, \sigma^2_0) , \quad \theta_1 = (2k_0, A_0, \tfrac{1}{2}a_0, \sigma^2_0), \qquad (4.2.3)$$

vemos que estos inducen la misma distribución, a saber

$$N((2k_0a_0; \sigma^2_0),$$

es decir, la parametrización no es identificable. Ahora, supongamos que existe $T(X_1)$ tal que

$$E_\theta[T(X_1)] = k \text{ para todo } \theta \in \Theta,$$

entonces tomando θ_0 y θ_1 en (4.2.3), concluimos que

$$E_{\theta_0}[T(X_1)] = k_0 \text{ y } E_{\theta_1}[T(X_1)] = 2k_0.$$

Sin embargo, puesto que

$$P_{X_1,\theta_0} = N(2k_0a_0; \sigma^2_0) = P_{X_1,\theta_1},$$

debemos tener que

$$E_{\theta_0}[T(X_1)] = E_{\theta_1}[T(X_1)],$$

o equivalentemente $k_0 = 2k_0$ lo cual es contradictorio. Luego no existe estimador insesgado para que dependa solamente de X_1. Para X_2 y X_3 la demostración es semejante. //

El teorema anterior puede extenderse facilmente a estimadores que sean función de un número finito de n observaciones de las varia - bles X_1, X_2, X_3; es decir, a estimadores que sean función de los vectores X_i, i = 1, 2, 3.

TEOREMA 4.2.3

Para cada i = 1, 2, 3, no existe estimador $T(X_i)$ tal que

$$E[T(X_i)] = k \text{ para todo } \theta \in \Theta. \text{ //}$$

DEMOSTRACION

Puesto que las observaciones del modelo son variables aleato - rias independientes y dentro de cada uno de los vectores (o muestras) - identicamente distribuidas, es fácil ver que la distribución de estos es

$$X_1 \sim N(2ka\mathbb{1}, \sigma^2 I),$$

$$X_2 \sim N(2kA\mathbb{1}, \sigma^2 I),$$

$$X_3 \sim N(k(A + a)\mathbb{1}, \sigma^2 I),$$

donde $\mathbb{1} \in \mathbb{R}^{n\times 1}$, $I \in \mathbb{R}^{n\times n}$ son como antes el vector unitario de dimensión $n \times 1$ y la matriz identidad de orden $n \times n$ (ver por ejemplo Searle, 1971).

Así pues si tomamos dos valores diferentes de θ digamos

$$\theta_0 = (k_0, A_0, a_0, \sigma^2_0) \quad y \quad \theta_1 = (2k_0, \tfrac{1}{2}A_0, \tfrac{1}{2}a_0, \sigma^2_0)$$

vemos que estos valores inducen las mismas distribuciones, a saber

$$N(2k_0 a_0 \mathbb{1}; \sigma^2_0 I),$$

$$N(2k_0 A_0 \mathbb{1}; \sigma^2_0 I),$$

$$N(k_0(A_0 + a_0)\mathbb{1}; \sigma^2_0 I),$$

y por lo tanto la parametrizaciones de cada X_i no son identificables. -

Ahora si suponemos que existen estimadores insesgados del número de loci tenemos que

$$E_{\theta_0}[T(X_i)] = k_0, \quad E_{\theta_1}[T(X_i)] = 2k_0,$$

sin embargo puesto que θ_0 y θ_1 inducen la misma distribución (para cualquiera de los tres vectores) se tiene que

$$E_{\theta_0}[T(X_i)] = E_{\theta_1}[T(X_i)],$$

lo cual conduce a la contradicción

$$k_0 = 2k_0.$$

Por lo tanto no existe estimador insesgado que sea función de un número finito de observaciones de las variables X_i, i = 1, 2, 3. //

Observación.

El significado biológico de los teoremas 4.2.2 y 4.2.3 es claro: no podemos obtener estimadores (insesgados) del número de loci cuando contamos únicamente con observaciones de poblaciones que son genéticamente homogeneas, es decir que no presentan segregación. Existe posibilidad de detectar el efecto del número de loci solamente cuando exista variabilidad genética entre los individuos de la población, como ocurre en las poblaciones F_2, R_1 y R_2.//

En el siguiente teorema veremos que la parametrización presentada es identificable en el caso de las variables: X_4, X_5 y X_6, que estan definidas como observaciones en las poblaciones F_2, R_1 y R_2 respectivamente. Por supuesto, lo anterior no significa que existan estimadores insesgados que sean función de estas variables, por lo cual para probar que no existen estimadores insesgados de k, que sean función de estas variables tendremos que utilizar argumentos diferentes.

TEOREMA 4.2.4

La parametrización $P_{X_i,\theta}$ es identificable. En otras palabras, si θ_0, $\theta_1 \in \Theta$; $\theta_0 \neq \theta_1$, entonces

$$P_{X_i,\theta_0} \neq P_{X_i,\theta_1}, \quad i = 4, 5, 6. \; //$$

DEMOSTRACION (para X_4)

Tenemos que ver que el hecho de que dos funciones de distribución de X_4 digamos $F_{X_4}(\cdot;\theta_0)$ y $F_{X_4}(\cdot;\theta_1)$ sean iguales implica que $\theta_0=\theta_1$. Puesto que la igualdad de dos funciones de distribución es equivalente a la igualdad de las correspondientes funciones generadoras de momentos (ver Lindgren, 1968) veremos que

$$M_{X_4,\theta_0}(t) = M_{X_4,\theta_1}(t) \text{ implica que } \theta_0 = \theta_1,$$

donde $M_{X_4,\theta}(t)$ denota la función generadora de momentos de X_4.

Recordemos que la función de densidad de cada X_4 correspondiente a θ esta dada por

$$f_{X_4}(y;\theta) = \left(\tfrac{1}{2}\right)^{2k} \sum_{s=0}^{2k} \binom{2k}{s} \phi(y;\, 2ka + s(A-a);\, \sigma^2), \qquad (4.2.4)$$

ahora poniendo

$$b_{ks} = \left(\tfrac{1}{2}\right)^{2k} \binom{2k}{s}, \quad c_{\theta s} = 2ka + s(A-a), \qquad (4.2.5)$$

entonces la densidad en 4.2.4 puede expresarse como

$$f_{X_4}(y;\theta) = \sum_{s=0}^{2k} b_{ks}\ \phi(y;\ c_{\theta s}\ ;\ \sigma^2),$$

y la correspondiente funcion generadora de momentos queda como

$$M_{X_4,\theta}(t) = \exp(\tfrac{1}{2}\sigma^2 t^2) \sum_{s=0}^{2k} b_{ks}\ \exp(c_{\theta s} t)$$

Ahora tomemos dos funciones generadoras de momentos de X_4 idénticas para todo t y correspondientes a

$$\theta_0 = (k_0, A_0, a_0, \sigma^2_0) \text{ y } \theta_1 = (k_1, A_1, a_1, \sigma^2_1);$$

veremos que el hecho de que

$$M_{X_4,\theta_0}(t) = M_{X_4,\theta_1}(t) \qquad (4.2.6)$$

implica que

$$\theta_0 = \theta_1$$

Tomando logaritmos en ambos lados de 4.2.6 obtenemos

$$\tfrac{1}{2}\sigma^2_0 t^2 + \log\Big(\sum_{s_0=0}^{2k_0} b_{k_0 s_0} \exp(c_{\theta_0 s_0} t)\Big) = \tfrac{1}{2}\sigma_1^2 t^2 + \log\Big(\sum_{s_1=0}^{2k_1} b_{k_1 s_1} \exp(c_{\theta_1 s_1} t)\Big)$$

(4.2.7) Note que para $t > 0$, $k = 1, 2, \ldots$ se tiene que

$$\exp(2kat) \leq \sum_{s=0}^{2k} b_{ks}\ \exp(c_{\theta s} t) \leq \exp((2ka + 2k(A-a))t)$$

y entonces el logaritmo en la ecuación anterior esta entre $2kat$ y $-(2ka + 2k(A-a))t$. Así dividiendo ambos términos de 4.2.7 entre t^2 y dejando que t tienda a infinito obtenemos

$$\tfrac{1}{2}\sigma^2_0 = \tfrac{1}{2}\sigma^2_1, \quad \sigma^2_0 = \sigma^2_1,$$

y entonces podemos escribir 4.2.7 como

$$\sum_{i=0}^{2k_0} b_{k_0 i} \exp(c_{\theta_0 i} t) = \sum_{j=0}^{2k_1} b_{k_1 j}\ \exp(c_{\theta_1 j} t) \qquad (4.2.8)$$

multiplique ambos lados de 4.2.8 por $\exp(-t c_{\theta_1 2k_1})$ para obtener

$$\textstyle\sum_i\ b_{k_0 i} \exp(t(c_{\theta_0 i} - c_{\theta_1 2k_1})) = \sum_j b_{k_1 j} \exp(t(c_{\theta_1 j} - c_{\theta_1 2k_1})) \qquad (4.2.9)$$

En la ecuación 4.2.9 tome el límite cuando $t \to \infty$. Debido a que

$$c_{\theta_1 j} < c_{\theta_1 2k_1}$$

para $j < 2k_1$ el límite del lado derecho de 4.2.9 es $b_{k_1 2k_1}$ el cual debe de ser igual al límite correspondiente del lado izquierdo. Como este límite (cuando $t \to \infty$) es finito y ya que $b_{k_0 i} > 0$ para $i=0, 1, \ldots 2k_0$ debemos de tener que

$$c_{\theta_0 i} \leq c_{\theta_1 2k_1} \quad i = 0, 1, 2, \ldots 2k_0$$

por lo tanto

$$c_{\theta_0 i} < c_{\theta_1 2k_1} \quad i = 0, 1, \ldots 2k_0 - 1, \text{ y}$$

$$c_{\theta_0 2k_0} \leq c_{\theta_1 2k_1} \tag{4.2.10}$$

Si la desigualdad estricta ocurre en 4.2.10 el límite cuando $t \to \infty$ del lado izquierdo de 4.2.9 es cero y se tendría

$$0 = b_{k_1 2k_1} > 0$$

lo cual es contradictorio. Por lo anterior se concluye que

$$c_{\theta_0 2k_0} = c_{\theta_1 2k_1} \tag{4.2.11}$$

por lo tanto

$$\lim_{t \to \infty} \Sigma_i b_{k_0 i} \exp(t(c_{\theta_0 i} - c_{\theta_1 2k_1})) = b_{k_0 2k_0}$$

lo que implica que

$$b_{k_0 2k_0} = b_{k_1 2k_1}$$

o equivalentemente

$$(\tfrac{1}{2})^{2k_0} = (\tfrac{1}{2})^{2k_1}$$

y así obtenemos

$$k_0 = k_1. \; //$$

Como mencionamos anteriormente, el hecho de que la
ción sea identificable en el caso de X_4, X_5 y X_6 implica que
rio utilizar otro tipo de argumentos para demostrar la no ex
estimadores insesgados del número de loci.

En la siguiente sección se presenta la nocion de es
completo, un comentario sobre su importancia y un ejemplo de
de estadísticos. Estos conceptos se utilizan para demostrar
tencia de estimadores insesgados del número de loci, primero
en que solamente se utiliza una observación de X_4, X_5 ó X_6 y
mente en el caso general.

4.3 Inexistencia de Estimadores Insesgados de k : Caso Ge

Para iniciar esta sección, daremos las definiciones
portantes conceptos en la teoría de la estimación puntual: la
dístico suficiente y la de familia completa de distribuciones
tadístico. Como se vera mas adelante estas nociones resultan
tes en el problema que nos ocupa.

DEFINICION 4.3.1 (Estadístico Suficiente).

Sea $\{P_\theta \mid \theta \in \Theta\}$ un modelo estadístico con observaci

$$Y = (Y_1, Y_2, \ldots Y_n).$$

Un estadístico $S = S(Y)$ se dice suficiente si y sólo si la fu
distribución condicional de Y dado S,

$$F_{Y|S}(y|S=s),$$

no depende del valor de θ.//

La importancia del concepto anterior reside en que, de cierta manera, el estadístico suficiente "concentra" la información que dan las observaciones; es decir, si se conoce el valor del estadístico suficiente, entonces no es necesario conocer los valores de todas las observaciones, pues estos no diran más acerca de θ, de lo que se puede inferir del valor de S (ver por ejemplo Mood et. al., 1974).

A continuación se presenta la definición de familia completa de distribuciones de un estadístico.

DEFINICION 4.3.2 (Familia Completa).

Sea el modelo $\{P_\theta \mid \theta \in \Theta\}$ con observaciones X, y $T = T(X)$ un estadístico. La familia de distribuciones de T, es decir, el conjunto

$$\{F_T(\cdot;\theta),\ \theta \in \Theta\}$$

se dice "familia completa" siempre que se satisfaga lo siguiente:

Si $g(\cdot)$ es una función de valores reales cuyo dominio incluye todos los valores de T y si

$$E_\theta[g(t)] = 0 \text{ para todo } \theta \in \Theta,$$

entonces se tiene que

$$P_\theta[g(T) = 0\ ] = 1 \text{ para todo } \theta \in \Theta.$$

En este caso también se dice que T es un estadístico completo. //

En otras palabras, la familia de distribuciones de un estadístico T se dice completa si la única variable aleatoria que es función de T y cuya esperanza es idénticamente cero, es la variable aleatoria cero.

En general la importancia de los anteriores conceptos en el problema de la estimación puntual reside en que permite juzgar las

diferentes funciones que se eligen como estimadores, y seleccionar las que presenten características deseables. Por ejemplo, si se tiene un estimador insesgado, que es función de un estadístico completo y suficiente, el teorema de Lehmann-Sheffe nos asegura que este sera de varianza uniformemente mínima, entre los estimadores insesgados y además único (ver Mood et al. 1974). En nuestro caso particular, estos conceptos serán de utilidad en la demostración de la inexistencia de estimadores insesgados para el número de loci.

El siguiente ejemplo de estadísticos suficientes y completos será utilizado más adelante.

EJEMPLO 4.3.1

En este ejemplo I es un intervalo de longitud menor que cero e I' es un intervalo de los reales positivos de longitud mayor que cero.

i) Sean $X_1, \ldots X_n$ variables aleatorias con distribución

$$N(\mu;\ \sigma_0^2),$$

donde σ_0^2 es fija y conocida y μ es fija y desconocida, esto es

$$X_i \sim N(\mu;\ \sigma_0^2) \quad \text{para } i = 1, 2, \ldots n;\ \mu \in I.$$

Entonces, $T = \sum_{i=1}^{n} X_i$ es un estadístico completo y suficiente (para μ).

ii) Sean $X_1, \ldots X_n$ variables aleatorias identicamente distribuidas, tales que

$$X_i \sim N(\mu;\sigma^2) \text{para } i = 1, 2, \ldots n\ ;\ \mu \in I,\ \sigma^2 \in I'.$$

Entonces $T = (\sum_{i=1}^{n} X_i\ ,\ \sum_{i=1}^{n} X_i^2\ )$ es un estadístico completo y suficiente para $\theta = (\mu, \sigma^2)$.

Nota: Por la importancia de la distribución normal en aplicaciones estadísticas el ejemplo anterior es bien conocido. Puede encontrarse una demostración en Mood et. al. (1974) o bien Lindgren (1968)

A continuación ilustraremos la importancia del concepto de completitud para resolver nuestro problema. En el teorema 4.2.2 vimos que no existen estimadores insesgados de k que sean función de X_1, X_2 ó X_3. ahora extenderemos este resultado para X_4, X_5 y X_6.

TEOREMA 4.3.1

Para i = 4, 5 ó 6, no existe un estadístico $T = T(X_4)$ tal que

$$E_\theta[T] = k \text{ para todo } \theta \in \Theta. \; //$$

DEMOSTRACION (Para X_4).

Veremos que no existe $T(X_4)$ insesgado para k. Recordemos que la función de densidad de X_4 se puede escribir como

$$f_{X_4}(y;\theta) = \sum_{s=0}^{2k} b_{ks} \exp(-\tfrac{1}{2}\sigma^2(y - 2ka - sc)^2),$$

donde

$$b_{ks} = (\tfrac{1}{2})^{2k} \binom{2k}{s}, \quad c = A - a.$$

Así vemos que

$$E_\theta[T(X_4)] = k$$

es equivalente a

$$\int \sum_{s=0}^{2k} b_{ks} \exp(-\tfrac{1}{2}\sigma^2(y-2ka-sc)^2)\, T(y)\, dy = k,$$

donde el símbolo "$\int$" denota integración entre $-\infty$ y $+\infty$. Ahora poniendo $x = y - sc$ tenemos

$$\int \sum_{s=0}^{2k} b_{ks} \exp(-\tfrac{1}{2}\sigma^2(x-2ka)^2) T(x+sc)\, dx = k \qquad (4.3.1)$$

Entonces $E_\theta[T(X_4)] = k$ para todo $\theta \in \Theta$ es equivalente a que 4.3.1 vale para todo θ. Por el lema del apéndice podemos suponer que T es una función continua, por lo tanto

$$g(x) = \sum_{s=0}^{2k} b_{ks}\ T(x+sc) - k \qquad (4.3.2)$$

tambien es continua. Para concluir que $g(x) = 0$ para todo $x \in \mathbb{R}$, es suficiente ver que g se anula fuera de un conjunto nulo. fijando k, σ^2 y c tenemos que 4.3.1 es equivalente a

$$E_a[\,g(X)] = 0,\ a > 0,$$

donde $X \sim N(2ka;\sigma^2)$. Por el ejemplo 4.3.1 se tiene que

$$P_a[\,g(X) = 0\,] = 1 \text{ para todo } a > 0,$$

entonces $g(x) = 0$, salvo para x en un conjunto nulo, y como mencionamos antes, esto implica que $g(x) = 0$ para todo $x \in \mathbb{R}$, pues g es continua.

Para finalizar la demostración veremos que T es identicamente k. De hecho, supongamos que $T(x_0) \neq k$, $x_0 \in \mathbb{R}$. Sin perdida de generalidad, supongamos que

$$T(x_0) > k,$$

luego $T(x_0 + h) > k$ si $|h| > \zeta$ para algún $\zeta > 0$, pues T es continua. Así pues, tomando c tal que

$$0 < c < \frac{\zeta}{2k}$$

se tiene que

$$0 < sc < \zeta \text{ para } s = 0, 1, \ldots 2k,$$

Por lo tanto, concluimos que

$$T(x_0 + sc) > k,\ s = 0, 1, \ldots 2k,$$

lo que a su vez implica que

$$\sum_{s=0}^{2k} b_{ks}\, T(x_0 + sc) > k,$$

y entonces $g(x_0) > 0$ lo cual es imposible pues $g(x) = 0$ para todo x. Similarmente, si tomamos $T(x_0) < k$ concluiremos que $g(x_0) < k$, lo cual contradice el hecho de que $g = 0$. Por lo tanto concluimos que

$$T(x) = k \text{ para todo } x \in \mathbb{R}.$$

Sin embargo, k es desconocido, y tomando otro valor k^* diferente de k tendríamos

$$T(x) = k \text{ y } T(x) = k^* \text{ para todo } x \in \mathbb{R},$$

lo cual es claramente imposible. Esto demuestra el teorema, pues la demostración para X_5 y X_6 es similar. //

Hasta el momento hemos visto que no existen estimadores insesgados del número de loci, que sean función de una sola observación de las variables implicadas (teoremas 4.2.2 y 4.3.1) o de un número finito de observaciones de X_1, X_2 ó X_3 (teorema 4.2.3). Cabe preguntarse si no existirá algún estimador insesgado de k que sea función de un número cualquiera de observaciones de todas las variables implicadas. La respuesta a esta pregunta es negativa, como se ve en el siguiente teorema.

NOTACION

Sea $Y_i \in \mathbb{R}^{6\times 1}$, $i = 1, 2, \ldots n$; el vector formado por las i-ésimas observaciones ordenadas del modelo presentado, esto es

$$Y_i' = (X_{1i}, X_{2i}, X_{3i}, X_{4i}, X_{5i}, X_{6i}), \quad i = 1, 2, \ldots n.$$

Ahora bien, el vector $Y \in \mathbb{R}^{6n\times 1}$ estará formado por los vectores Y_1, Y_2, $\ldots$ Y_n, donde Y_i es como arriba; esto es

$$Y' = (Y_1, Y_2, \ldots Y_n).$$

Recordando las definiciones de las variables aleatorias X_1, X_2, ... X_6, presentadas anteriormente (ver sección 2.1), vemos que Y_i se puede escribir como

$$Y_i = 2ka1 + cV_i + \varepsilon_i , \quad i = 1, 2, \ldots n;$$

donde $V_i \in \mathbb{R}^{6x1}$ es un vector aleatorio definido como

$$V_i' = (0, 2k, k, Q_{1i}, Q_{2i}, Q_{3i}+k),$$

donde, como antes

$$Q_{1i} \sim B(2k; \tfrac{1}{2}),$$

$$Q_{2i} \sim B(k; \tfrac{1}{2}),$$

$$Q_{3i} \sim B(k; \tfrac{1}{2}),$$

y $\varepsilon_i \in \mathbb{R}^{6x1}$ es un vector aleatorio, formado por

$$\varepsilon_i = (\varepsilon_{1i}, \varepsilon_{2i}, \ldots \varepsilon_{6i}),$$

donde consecuentemente ε_{ij} es el efecto medioambiental de la j-ésima observación de la variable i, y por lo tanto

$$\varepsilon_{ij} \sim N(0; \sigma^2).$$

Por extensión podemos escribir

$$Y = 2ka1\!\!1 + cV + \varepsilon ,$$

donde ahora $1\!\!1$ es el vector unitario $1\!\!1 \in \mathbb{R}^{6nx1}$, y $V \in \mathbb{R}^{6nx1}$ es el vector aleatorio formado por los V_i, esto es

$$V' = (V_1 , V_2 , \ldots V_n),$$

y $\varepsilon \in \mathbb{R}^{6nx1}$ es el vector aleatorio cuyos componentes son los ε_i definidos antes

$$\varepsilon = (\varepsilon_1 , \varepsilon_2 , \ldots \varepsilon_n).$$

De acuerdo con la anterior notación la función de densidad de Y puede escribirse como

$$f_Y(y;\theta) = \sum_v P_v \exp(-\tfrac{1}{2}\sigma^2(\| y - 2ka\mathbb{1} - cv \|)^2),\ y \in \mathbb{R}^{6n\times 1},\ \theta \in \Theta .$$

donde la sumatoria es sobre todos los valores posibles de V, y las P_v - son las constantes definidas por

$$P_v = P_\theta[V = v] .$$

ademas para $w \in \mathbb{R}^{m\times 1}$, $m = 1, 2, \ldots$

$$\| w \|^2 = \sum_{i=1}^{m} w_i^2 .$$

Podemos conceptualizar la distribución de Y como una mezcla de densidades normales con diferentes medias. Es claro que la familia de - densidades de Y no es completa pues, por ejemplo, si tomamos n = 2 y - $T(Y) = Y_1 - Y_2$ tenemos que

$$E_\theta[T] = 0 \text{ para todo } \theta \in \Theta ,$$

sin embargo, obviamente $T(Y) \neq 0$ para un sinnúmero de valores de Y = y.

TEOREMA 4.3.2

Sea Y como antes. Entonces no existe un estimador T(Y) tal - que

$$E_\theta[T(Y)] = k \text{ para todo } \theta \in \Theta . \ //$$

DEMOSTRACION (Por reducción al absurdo).

Suponga que existe un estimador insesgado de k, es decir un estadístico T(Y) tal que

$$E_\theta[T(Y)] = k \qquad (4.3.3)$$

veremos que esto nos conduce a una contradicción. Note que 4.3.3 es - equivalente a

$$\int \cdots \int \sum_v P_v T(y) \exp(-\tfrac{1}{2}\sigma^2(\| y-2ka\mathbb{1} -cv\|^2)) dy = k$$

y haciendo el cambio de variable $x = y - cv$ se obtiene

$$\int \cdots \int \sum_v P_v T(x + cv) \exp(-\tfrac{1}{2}\sigma^2(\|x - 2ka\mathbb{1}\|^2)) dx = k \qquad (4.3.4)$$

Ahora si Y^* es un vector aleatorio de dimensión 6n tal que

$$Y^* \sim N(2ka\mathbb{1} ; \sigma^2 I),$$

vemos que la ecuación 4.3.4 se puede escribir como

$$E_\theta[\sum_v P_v T(Y^* + cv)] = k \text{ para todo } \theta \in \Theta. \qquad (4.3.5)$$

Por el ejemplo 4.3.1 tenemos que para la distribución de Y^*,

$$S = S(y^*) = (y^{*\prime}\mathbb{1} , \|y^*\|^2),$$

es un estadístico completo y suficiente. Por lo tanto

$$E_\theta[\sum_v P_v T(Y^* + cv) | S=s] = k, \qquad (4.3.6)$$

no depende de θ.

Ahora sean a, b, c, d números reales tales que

$$a < b \text{ y } c < d,$$

y definiendo

$$J_1 = (a,b], \ J_2 = (c,d]$$

tenemos que

$$E_\theta[\sum_v P_v T(Y^* + cv) \ I[S \in J_1 \times J_2] \mid S = s]$$

$$= I[s \in J_1 \times J_2] \ E_\theta[\sum_v P_v T(Y^* + cv) \mid S = s]$$

$$= I[s \in J_1 \times J_2]k,$$

donde $I[\cdot]$ denota intervalo y la primera igualdad se debe al teorema de substitución, mientras que la segunda se obtiene de 4.3.6. Luego, vemos que

$$E_\theta[\sum_v P_v T(y^* + cv)\ I[S \in J_1 \times J_2] \mid S = s] = I[S \in J_1 \times J_2]\, k$$

y tomando esperanza respecto a S, concluimos que

$$E_\theta[\sum_v P_v T(y^*+cv)\ I[S \in J_1 \times J_2]] = kP_\theta[S \in J_1 \times J_2] \qquad (4.3.7)$$

donde hemos usado el teorema de la doble esperanza (Rao, 1965).

Ahora tome $a=(r-\varepsilon)\sqrt{6n}$, $b=r\sqrt{6n}$, $c = 0$, $d=0$ y utilizando la notación del lema A2 del apéndice vemos que

$$I[S \in J_1 \times J_2] = I[(r-\varepsilon)\sqrt{6n} < y^* 1\!\!1 \leq r\sqrt{6n},\ 0 < \|y^*\|^2 \leq r^2]$$

$$= I[y^* \in G_{r\varepsilon}],$$

y concluimos a partir de 4.3.7 que para todo $\theta \in \Theta$

$$E_\theta[\sum_v P_v\ T(y^* + cv)\ I[y^* \in G_{r\varepsilon}] = k\ P_\theta[y^* \in G_{r\varepsilon}] \qquad (4.3.8)$$

Ahora defina $g(y^*,\theta)$ mediante

$$g(y^*, \theta) = \sum_v P_v T(y^* + cv) - k$$

y se obtiene que 4.3.8 es equivalente a

$$E_\theta[g(y^*,\theta)\ I[y^* \in G_{r\varepsilon}]] = 0 \text{ para todo } \theta \in \Theta. \qquad (4.3.9)$$

Usando el lema A1 del apéndice podemos suponer que $g(y^*,\theta)$ es una función continua. Así para cada θ fijo se tiene que dado $\xi > 0$ existe $\xi_1 > 0$ tal que

$$| y^* - r1\!\!1 / \sqrt{6n} | < \xi_1$$

lo cual implica que

$$|g(y^*,\theta) - g(r1\!\!1/\sqrt{6n},\ )| < \xi \qquad (4.3.10)$$

Ahora usando el lema A3 del apéndice vemos que

$$y^* \in G_{r\varepsilon}$$

implica que

$$\| y^* - r\mathbb{1}/\sqrt{6n} \| \leq \sqrt{2r\varepsilon} < \xi_1 \text{ si } \varepsilon < (\xi_1/\sqrt{2r})^2$$

y entonces

$$| g(y^*,\theta) - g(r\mathbb{1}/\sqrt{6n},\theta) | < \xi \quad \text{para} \quad y^* \in G_{r\varepsilon} \tag{4.3.11}$$

siempre que

$$\varepsilon < (\delta_1/\sqrt{2r})^2 .$$

Entonces

$$| E_\theta[g(y^*,\theta)\ I[y^* \in G_{r\varepsilon}] - E_\theta[g(r\mathbb{1}/\sqrt{6n},\theta)\ I[y^* \in G_{r\varepsilon}] |$$

$$= | E_\theta[(g(y^*,\theta) - g(r\mathbb{1}/\sqrt{6n},\theta))\ I[y^* \in G_{r\varepsilon}]|$$

$$\leq \xi E_\theta[I[y^* \in G_{r\varepsilon}]] \tag{4.3.12}$$

donde $\varepsilon < (\xi_1/\sqrt{2r})$ y hemos usado 4.3.11 para obtener la desigualdad. -

Por otro lado, usando 4.3.9 junto con 4.3.12 vemos que

$$| E_\theta[g(r\mathbb{1}/\sqrt{6n},\theta)\ I[Y^* \in G_{r\varepsilon}] | \leq \xi\ E_\theta[I[Y^* \in G_{r\varepsilon}]]$$

y entonces

$$| g(r\mathbb{1}/\sqrt{6n},\theta) |\ P_\theta[y^* \in G_{r\varepsilon}] \leq \xi\ P_\theta[Y^* \in G_{r\varepsilon}]$$

y como la probabilidad que aparece en la desigualdad anterior es positiva (vea el lema A2 del apéndice) vemos que

$$|g(r\mathbb{1}/\sqrt{6n},\theta)| \leq \xi$$

y como $\xi > 0$ y $\theta \in \Theta$ son arbitrarios, vemos que

$$g(r\mathbb{1}/\sqrt{6n},\ ) = 0$$

es decir,

$$\sum_v P_v\ T(r\mathbb{1}/\sqrt{6n} + cv) = k \tag{4.3.13}$$

es valido para todo $\theta \in \Theta$. Tomando el límite cuando $c \to \infty$ en 4.3.13 concluimos que

$$T(r1\!\!1 / \sqrt{6n}) = k.$$

Tomando otro valor de k, digamos $\tilde{K}$ se tendría que

$$T(r1\!\!1 / 6n) = k \quad y \quad T(r / 6n) = \tilde{K},$$

lo cual es imposible. Con lo anterior se concluye que no existen estimadores insesgados del número de loci.

CAPITULO 5

DISCUSION Y OBSERVACIONES FINALES

Hasta este punto el presenta trabajo ha consistido en plantear un modelo genético y estadístico para la herencia de características - continuas. Dicho modelo sirve como marco para el planteamiento del problema de la estimación del número de loci que es el objetivo de ésta tesis. A este respecto, en el capítulo tres se presentó un conjunto de estimadores $\{\hat{K}_h\}$, $h = (i, j, l, m)$ acerca de los cuales se demostró que son consistentes y que no tienen esperanza finita. Esto motivó el estudio de la posibilidad de estimación insesgada de K, concluyendo en el - capítulo cuatro con la demostración de que no existen estimadores insesgados.

Como sucede a menudo, el planteamiento y solución parcial de - un problema conduce hacia otros problemas, más profundos, más sutiles o mas específicos.

En nuestro caso surgen un gran número de interrogantes: ¿son - biologicamente válidas las hipótesis planteadas?, ¿que sucede con el modelo (y por ende con la estimación de K) si estas hipótesis no se cumplen?, ¿como seleccionar el mejor estimador?, ... etc. En este capítulo no se pretende contestar dichas interrogantes, sino solamente plantearlas adecuadamente y discutir muy brevemente sus expectativas de solución.

En la siguiente sección se discute el incumplimiento de las hipótesis que se plantearon al formular el modelo, y en la sección 5.2 -

se comenta sobre los criterios de elección de estimadores en el caso - que nos ocupa, concluyendo con algunos comentarios sobre los problemas que parecen relevantes en este caso.

5.1 Incumplimiento de las Hipótesis

Las hipótesis formuladas al presentar el modelo fueron:

H1: Las poblaciones P_1 y P_2 son genéticamente homogeneas.

H2: Todos los individuos de P_1 y P_2 son homocigotos para los k loci.

H3: Los k loci se heredan independientemente.

H4: Para cada j = 1, 2, ... , se tiene que

$$A_j > a_j > 0$$

H5: Los efectos de los genes son totalmente aditivos.

H6: $A_1 = A_2 = \ldots = A_k = A$, $a_1 = a_2 = \ldots = a_k = a$.

H7: Para i = 1, 2, ... 6 se tiene que

$$\varepsilon_i \sim N(0;\sigma^2).$$

Analizaremos rápidamente cada una de estas hipótesis comentando cuando se puede suponer biologicamente válida, como es el modelo si no se cumple y que pasa con los estimadores $\hat{K}_h$ en ese caso.

Con respecto a H1, podremos suponerla válida cuando se trate - de líneas puras (de plantas) o poblaciones muy consanguíneas de animales. Si este no es el caso se tendrá que las variables X_1 y X_2 no estaran adecuadamente representadas por 2.1.4 y 2.1.5, en el sentido de que existirá una cierta diferencia genética dentro de los individuos P_1 y - P_2, lo cual incidirá sobre la varianza de X_1 y X_2 pues ahora

$$V_\theta[X_1] = \sigma^2_{g1} + \sigma^2,$$

$$V_\theta[X_2] = \sigma^2_{g2} + \sigma^2 ,$$

donde σ^2_{g1} y σ^2_{g2} son varianzas originadas por las diferencias genéticas de los individuos. Si estas varianzas son aproximadamente iguales - ($\sigma^2_{g1} \simeq \sigma^2_{g2}$), la estimación de k no sufrirá muchos cambios puesto que - por la ley del equilibrio de Hardy y Weinberg (ver Elandt-Jhonson, 1971) se tendrá que el termino σ^2_g se agregará a las varianzas de todas las variables y por lo tanto la condición 3.1.3 se seguirá cumpliendo. Si la diferencia entre σ^2_{g1} y σ^2_{g2} es muy grande habrá que buscar un estimador que sea menos sensible a esta diferencia, lo cual se puede hacer utilizando todas las variables para estimar $k(A - a)^2$ (ver siguiente sección).

La hipótesis H2 es un complemento de H1, pues especifica que, a más de ser idénticos, los individuos de las poblaciones progenitoras deben de tener dos genes de idéntico efecto en cada locus (homocigotas). Las condiciones biológicas en que ésto se puede cumplir son las mismas *que en el caso anterior. Si H2 no se cumple, se presentará un componente* de varianza extra en la varianza de X_3, o sea la variable de interés en F_2. Es decir

$$V_\theta[X_3] = \sigma^2_s + \sigma^2$$

donde σ^2_s es la varianza debida a la segregación de los loci que se encuentren en estado heterocigotico. Si este es el caso, debe de evitarse utilizar estimadores que contengan a S^2_3, pues ésta no será un estimador insesgado de σ^2 sino que

$$E_\theta[S^2_3] = \sigma^2_s + \sigma^2$$

Sin embargo, dado que H1 si se cumple seguiremos teniendo que

$$V_\theta[X_1] = V_\theta[X_2] = \sigma^2$$

Las varianzas de X_4, X_5 y X_6 siguen siendo las mismas.

Con respecto a H3, es imposible predecir en que situación biológica puede considerarse realista, puesto que los diferentes loci que afectan una característica parecen distribuirse al azar dentro de los cromosomas. Es fácil ver el efecto que tendrá que los genes esten ligados, si fijamos nuestra atención en el caso límite de que dos loci esten completamente ligados. De ser así, estos dos loci (diferentes) se comportarían como un solo locus para efectos de segregación, con lo cual se reduciría la varianza genética, esto es $c_h k(A - a)^2$, y por lo tanto se subestimaría (con los $\hat{K}_h$ presentados) el número de loci. Este punto es tratado más a fondo por Mather y Jinks (1977).

En lo que concierne a H4 es difícil predecir en que situaciones biológicas pueda ser cierta, pero en caso de violarse, es decir, si se tiene que

$$A_j > a_j \text{ para } j = 1, 2, \ldots l \ (l < k) \text{ y}$$
$$A_j < a_j \text{ para } j = l, l+1, \ldots k.$$

Esto altera las esperanzas de las variables, y por tanto la estimación de k mediante los $\hat{K}_h$ propuestos. Además, H6 tendría que modificarse para ser consistente con lo anterior, quedando:

H6*: $A_1 = A_2 = \ldots A_l = A$, $A_{l+1} = A_{l+2} = \ldots A_k = a$

$a_1 = a_2 = \ldots a_l = a$, $a_{l+1} = a_{l+2} = \ldots a_k = A$

Los efectos del incumplimiento de H4 deben estudiarse para casos específicos de los estimadores planteados.

En el caso de que los efectos de los genes no sean totalmente aditivos, esto es, cuando no se cumpla H5, podemos pensar que existen efectos de dominancia y/o epistasis, dentro y/o entre los diferentes -

loci involucrados. Aquí solamente revisaremos el caso de dominancia - planteando dos hipótesis accesorias H5* y H5**

H5*: Los efectos de dominancia en los k loci son iguales a d.

h5**: No existen efectos de epistasis.

Con lo anterior el modelo quedaría como sigue

$$X_1 = 2ka + \varepsilon_1 \tag{5.1.1}$$

$$X_2 = 2kA + \varepsilon_2 \tag{5.1.2}$$

$$X_3 = k(A + a) + \varepsilon_3 \tag{5.1.3}$$

$$X_4 = 2N_1A + N_2(A + a + d) + 2N_3a + \varepsilon_4 \tag{5.1.4}$$

$$X_5 = 2ka + Q_5(A - a + d) + \varepsilon_5 \tag{5.1.5}$$

$$X_6 = k(A + a + d) + Q_6(A - a - d) + \varepsilon_6 \tag{5.1.6}$$

donde d es el efecto de dominancia del j-ésimo locus, j = 1, 2, ... k, y por H5* es igual para todos los loci, el vector (N_1, N_2, N_3) es variable aleatoria tal que

$$(N_1, N_2, N_3) \sim M(k, \tfrac{1}{4}, \tfrac{1}{2}, \tfrac{1}{4})$$

y como antes

$$Q_5 \sim B(k, \tfrac{1}{2}),$$

$$Q_6 \sim B(k, \tfrac{1}{2}).$$

Es fácil ver que en este caso

$$E_\theta[X_1] = k(A + a + d),\ V_\theta[X_1] = \sigma^2;$$

$$E_\theta[X_2] = k(A + a + \tfrac{1}{2}d),\ V_\theta[X_2] = k(\tfrac{1}{2}(A - a) + \tfrac{1}{4}d) + \sigma^2;$$

$$E_\theta[X_3] = k(\tfrac{1}{2}A + \tfrac{1}{2}a + \tfrac{1}{2}d),\ V_\theta[X_3] = \tfrac{1}{4}k(A - a + d) + \sigma^2;$$

$$E_\theta[X_4] = k(\tfrac{1}{2}A + \tfrac{1}{2}a + \tfrac{1}{2}d),\ V_\theta[X_4] = \tfrac{1}{4}k(A - a + d) + \sigma^2.$$

Nota: Las esperanzas y varianzas de X_1 y X_2 no se alteran, puesto que - los efectos de dominancia solamente se presentan en los loci heterocigotos.

Es claro que en estos casos los estimadores $\hat{K}_h$ no cumplen la - condición 3.1.3, y por tanto pierden la propiedad de ser asintóticamente convergentes en probabilidad a k, es decir, los $\hat{K}_h$ ya no serán consistentes. En este caso es recomendable buscar el $\hat{K}_h$ que minimice s en la condición

$$\frac{(E_\theta[X_i] - E_\theta[X_j])^2}{c_h(V_\theta[X_l] - V_\theta[X_m])} = k + s$$

También es posible encontrar funciones de esperanzas y varianzas que - sean iguales a k, y por lo tanto, los estimadores de momentos resultantes al substituir los momentos poblacionales por los muestrales, serán estimadores consistentes de k. Esto se puede lograr manipulando adecuadamente las expresiones para esperanzas y varianzas presentadas como - resultado de la modificación de H5: H5* y H5**.

La hipótesis H6 de igualdad de los efectos génicos, fué pro - puesta como una forma de simplificar el tratamiento del problema. Si se abandona dicha hipótesis, y únicamente se pone la condición

$$1/k(\sum_{i=1}^{k} A_i) = A, \quad 1/k(\sum_{i=1}^{k} a_i) = a, \qquad (5.1.8)$$

las ecuaciones 2.1.1, 2.1.1 y 2.1.3 representan adecuadamente a X_1, X_2 y X_3. Para obtener una expresión adecuada de X_4, X_5 y X_6 es más sencillo utilizar como variable aleatoria el valor posible de cada locus, en vez de utilizar el número de genes de cada tipo (Q_i, i = 1, 2, 3 en los capítulos dos, tres y cuatro). Es decir, sabemos que cada i-ésimo loci, - i = 1, 2, ... puede estar en tres estados posibles: homocigoto A_i (A_iA_i), heterocigoto (A_ia_i) y homocigoto a_i (a_ia_i), de manera que si definimos

las variables aleatorias independientes Y_i, $i = 1, 2, \ldots k$ con función de probabilidad dada por

$$P_\theta[Y_i = y_i] = \begin{cases} \frac{1}{4} \text{ si } y_i = 2A_i \\ \frac{1}{2} \text{ si } y_i = A_i + a_i \\ \frac{1}{4} \text{ si } y_i = 2a \\ 0 \text{ de otro modo.} \end{cases}$$

entonces tendremos que $X_4 : F_2 \to \mathbb{R}$ queda definida por

$$X_4 = (\sum_{i=1}^{k} Y_i) + \varepsilon_4 . \qquad (5.1.9)$$

La función de probabilidad de Y es debida a que la probabilidad de que un locus cualquiera quede en estado homocigoto (A_iA_i o a_ia_i) es de ¼, mientras que la probabilidad de que quede en estado heterocigoto ($A_i a_i$) es de ½.

Así mismo definiendo Z_i, $i = 1, 2, \ldots k$ por la función de probabilidad

$$P_\theta[Z_i = z_i] = \begin{cases} \frac{1}{2} \text{ si } z_i = A_i + a_i \\ \frac{1}{2} \text{ si } z_i = 2a_i \\ 0 \text{ de otro modo.} \end{cases}$$

tenemos que $X_5 : R_1 \to \mathbb{R}$ se define por

$$X_5 = (\sum_{i=1}^{k} Z_i) + \varepsilon_5 \qquad (5.1.10)$$

y definiendo a W_i, $i = 1, 2, \ldots k$ como la variable aleatoria que cumple

$$P_\theta[W_i = w_i] = \begin{cases} \frac{1}{2} \text{ si } w_i = A_i + a_i \\ \frac{1}{2} \text{ si } w_i = 2A_i \\ 0 \text{ de otro modo.} \end{cases}$$

tenemos que $X_6 : R_2 \to \mathbb{R}$ queda

$$X_6 = (\sum_{i=1}^{k} W_i) + \varepsilon_6 . \qquad (5.1.11)$$

Dado lo anterior, es fácil comprobar que las esperanzas y varianzas de X_4, X_5 y X_6 quedan dadas por las expresiones siguientes

$$E_\theta[X_4] = \sum_{i=1}^{k}(A_i + a_i),\ V_\theta X_4 = \tfrac{1}{2}\sum_{i=1}^{k}(A_i - a_i)^2 + \sigma^2,$$

$$E_\theta[X_5] = \sum_{i=1}^{k}(\tfrac{1}{2}A_i + \tfrac{3}{2}a_i),\ V_\theta X_5 = \tfrac{1}{4}\sum_{i=1}^{k}(A_i - a_i)^2 + \sigma^2,$$

$$E_\theta[X_6] = \sum_{i=1}^{k}(\tfrac{3}{2}A_i + \tfrac{1}{2}a_i),\ V_\theta[X_6] = \tfrac{1}{4}\sum_{i=1}^{k}(A_i - a_i)^2 + \sigma^2 .$$

Dada la condición 5.1.8 es fácil ver que las esperanzas de X_4, X_5 y X_6 no varian con respecto al modelo que conserva H6 (ver capítulo dos), sin embargo, todas las varianzas seran mayores, esto es

$$\tfrac{1}{2}\sum_{i=1}^{k}(A_i - a_i)^2 + \sigma^2 \geq \tfrac{1}{2}k(A - a)^2 + \sigma^2$$

$$= \tfrac{1}{2}k\{1/k(\sum_{i=1}^{k}A_i - \sum_{i=1}^{k}a_i)^2\} + \sigma^2,$$

y por lo tanto los estimadores $\hat{R}_h$ siempre subestimaran el valor de k puesto que

$$\frac{k^2 (A - a)^2}{k(A - a)^2} \geq \frac{k^2 (A - a)^2}{\sum_{i=1}^{k}(A_i - a_i)^2}$$

donde se cumple la condición 5.1.8. Lo anterior se puede comprobar fácilmente por medio de la desigualdad de Jensen (Rao, 1965).

En lo que se refiere a H7, la hipótesis sobre la normalidad de los errores o efectos medioambientales, la literatura reporta que es un supuesto generalmente aceptable (Bulmer, 1985; Kempthorne, 1969; Mayo, 1980) y parece que su incumplimiento no tendría mucha importancia en la estimación de k, siempre que la esperanza de la distribución de los $\hat{R}$ exista y sea cero y la varianza $\sigma^2 > 0$.

En la siguiente sección se discute brevemente sobre los posi - bles criterios de elección de un estimador para k.

5.2 Elección del Estimador de k

Como se ha mencionado, el hecho de que los estimadores $\hat{K}_h$ propuestos no tengan esperanza finita impide utilizar los criterios comu - nes de selección del mejor estimador. De aquí la importancia de que en el futuro se estudie la obtención de estimadores razonables con esperanza finita, o bien se demuestre que no existen. Por el momento nos queda el problema eminentemente práctico de seleccionar alguno de los estimadores $\hat{K}_h$ disponibles.

En muchos casos no se dispondrá de observaciones de todas las variables X_1 a X_6, y en este caso se tendrá que seleccionar un estima - dor que solamente utilice las variables disponibles. No nos ocuparemos aquí de esa situación.

En otros casos se tendrán razones para sospechar el incumpli - miento de alguna(s) de la(s) hipótesis. No discutiremos aquí la sele - cción de un estimador en esos casos (ver sección anterior). De hecho, - esta situación así como la del párrafo precedente amerita un trabajo - aparte.

Supondremos entonces que se cuenta con n observaciones de cada una de las variables y que las hipótesis H1 a H6 se cumplen.

Dada la consistencia de los estimadores propuestos es adecuado pensar que un estimador que utilice todas las observaciones tendrá ma - yor probabilidad de estar cercano al valor del parámetro. Ninguno de - los estimadores $\hat{K}_h$ cumple con ello, pero es fácil ver que existen -

expresiones, función de esperanzas y varianzas de las seis variables, - que son iguales a k, por ejemplo

$$\frac{(E_\theta[X_2] - E_\theta[X_1])^2 + (E_\theta[X_3] - E_\theta[X_5])^2 + (E_\theta[X_6] - E_\theta[X_4])^2}{26/4(V_\theta[X_4] + V_\theta[X_5] + V_\theta[X_6] - V_\theta[X_3] - V_\theta[X_2] - V_\theta[X_1])} = k$$

y aplicando el método de los momentos obtenemos el estimador

$$\hat{K}^* = \frac{(\overline{X}_2 - \overline{X}_1)^2 + (\overline{X}_3 - \overline{X}_5)^2 + (\overline{X}_6 - \overline{X}_4)^2}{26/4(S_4^2 + S_5^2 + S_6^2 - S_1^2 - S_2^2 - S_3^2)}$$

que se puede verificar que es consistente y no tiene esperanza finita.

También es posible obtener estimadores que utilicen todas las observaciones haciendo combinaciones lineales de los $\hat{K}_h$, por ejemplo

$$\hat{K}^{**} = 1/3(\hat{K}_{(2,1,4,3)} + \hat{K}_{(3,5,5,2)} + \hat{K}_{(6,4,6,1)})$$

sin embargo, dentro del conjunto de los estimadores de k que utilizan - todas las observaciones tampoco es posible seleccionar al "mejor" (se - gún el criterio de error cuadrático mínimo), pues es posible demostrar (en una forma similar a la demostración dada en la sección 3.3) que estos estimadores no tienen esperanza finita.

Por lo tanto, y solamente para buscar un posible criterio de - selección, se expandió $\hat{K}_h$ en serie de Taylor, alrededor de los momentos poblacionales. El resultado de dicha expansión hasta el tercer - término fué

$$E_\theta[\hat{K}_h] = k + \frac{\sigma_l^2 + \sigma_j^2}{c_h c_{lm} nk(A-a)^2} + \frac{(n-1)(\mu_l^4+\mu_m^4) - n(n-3)(\sigma_l^4+\sigma_m^4)}{n(n-1)c_{lm}^2 k^2 (A-a)^4} + R_\theta$$

donde μ^r y σ^r representan el r-ésimo momento alrrededor de el orígen y alrededor de μ respectivamente, y por R_θ el residuo. Dado que no existe $E\hat{K}$ finita, es claro que la serie anterior no converge, sin embargo podemos minimizar el valor de

$$\frac{\sigma_1^2 + \sigma_1^2}{c_h c_{1m} nk(A-a)^2} + \frac{(n-1)(\mu_1^4 + \mu_m^4) - n(n-3)(\sigma_1^4 - \sigma_m^4)}{n(n-1)c_{1m}^2 k^2 (A-a)^4}$$

para un θ_0 y n fijos si escogemos los valores máximos de c_h y c_{1m}, para el conjunto R_h. Con este criterio, los estimadores seleccionados resultan

$$R_h^* = \frac{(\bar{X}_2 - \bar{X}_1)^2}{8(S_1^2 - S_m^2)} \qquad , (1,m) = (4,3), (4,2), (4,1).$$

Ahora bien, si en vez de utilizar S_m^2 con m = 1, 2, 3, utilizamos el estimador S^{2*} dada por una ponderación de S_1^2, S_2^2 y S_3^2, es decir

$$S^{2*} = 1/3(S_1^2 + S_2^2 + S_3^2)$$

obtenemos el estimador

$$R^{**} = \frac{(\bar{X}_2 - \bar{X}_1)^2}{8(S_4^2 - S^{2*})}$$

que, de alguna manera poco precisa, puede decirse que es mejor.

Para concluir este trabajo es adecuado resaltar que los problemas de estimación de parámetros genéticos tienen una enorme importancia práctica, pues de su solución puede depender el diseño más adecuado de técnicas de mejoramiento genético de especies cultivadas, y por tanto - un aumento en la producción agropecuaria. ///

LITERATURA CITADA

Baldwin, J. M. 1896. A New Factor in Evolution. En: Jamenson, D. L. - (Ed.) 1977. Evolutionary Genetics. Dowden, Hutch & Ross, - Inc. USA. p. 45-50.

Bulmer, M. G. 1985. The Mathematical Theory of Quantitative Genetics. - Clarendon Press. Oxford, USA. 253 p.

Brunk, H. D. 1965. An Introduction to Mathematical Statistics. 2ª ed. - Blaisdell Publishing Co. USA. 429 p.

Castle, W. E. 1903. The Laws of Heredity of Galton and Mendel, and Some Laws Governing Race Improvement by Selection. En: Jamenson, D. L. (Ed.) 1977. Evolutionary Genetics. Dowden, Hutch & - Ross, Inc. USA. p. 90-109.

East, E. M. 1910. A Mendelian Interpretation of Variation that is Apparently Continous. En: Jamenson, D. L. 1977. Evolutionary - Genetics. Dowden, Hutch & Ross, Inc. USA. p. 137-155.

Elandt-Johnson, R. C. 1971. Probability Models and Statistical Methods in Genetics. John willey & Sons. USA. 591 p.

Fisher, R. A. 1958. The Genetical Theory of Natural Selection. Dover - Publications Inc. USA. 291 p.

Galton, F. 1889. The Average Contribution of Each Several Ancestor to - the Total Heritage of the Offspring. En: Jamenson, D. L. - (Ed.) 1977. Evolutionary Genetics. Dowden, Hutch & Ross, - Inc. USA. p. 32-44.

Haldane, J. B. S. 1926. A Mathematical Theory of Natural and Artificial Selection. Parts I to V. En: Robinson, H. F. 1986. (Ed.) Papers on Quantitative Genetics and Related Topics. North Carolina State Univ. USA. p. 195-250.

Hardy, G. H. 1908. Mendelian Proportions in a Mixed Population. En: - Jamenson, D. L. (Ed.) 1977. Evolutionary Genetics. Dowden, Hutch & Ross, Inc. USA. p. 113-114.

Jamenson, D. L. (Ed.) 1977. Evolutionary Genetics. Dowden, Hutch & Ross, Inc. USA. 332 p.

Kempthorne, O. 1969. An Introduction to Genetic Statistics. The Iowa - State Univ. Press. USA. 545 p.

Koroliuk, V. S. 1981. Manual de la Teoría de Probabilidades y Estadística Matemática. Mir. Moscu. URSS. 579 p.

Lindgren, B. W. 1968. Statistical Theory. 3ª ed. McMillan Publishing Co. USA. 613 p.

Mather, K. and Jinks, J. L. 1977. Introduction to Biometrical Genetics. Cornell Univ. Press. USA. 231 p.

Mayo, O. 1980. The Theory of Plant Breeding. Oxford Univ. Press. UK. 293 p.

Mendel, G. 1865. Experiments in Plant Hybridisation. En: Jamenson (Ed.) 1977. Evolutionary Genetics. Dowden, Hutch & Ross, Inc. USA. p. 51-83.

Mendenhall, W., Scheaffer, R. L. y Wackerly, D. D. 1974. Estadística Matemática con Aplicaciones. Grupo Editorial Iberoamericana. Mex. 751 p.

Mood, A. M., Graybill, F. A. and Boes, D. C. 1974. Introduction to the Theory of Statistics. 3ª ed. McGraw-Hill. USA. 564 p.

Parzen, E. 1979. Teoría Moderna de Probabilidades y sus Aplicaciones. Limusa. Mex. 509 p.

Pearson, K. 1904. On a Generalized theory of Alternative Inheritance, with Spetial Reference to Mendel's Laws. En: Jamenson, D. L. (Ed.) 1977. Evolutionary Genetics. Dowden, Hutch & Ross, Inc USA. p. 85-89.

_____1909a. The Theory of Ancestral Gametic Contributions in Heredity. En: Jamenson, D. L. (Ed.) 1977. Evolutionary Genetics. Dowden, Hutch & Ross, Inc. USA. p. 126-131.

_____1909b. On the Ancestral Gametic Correlation of a Mendelian Population Mating at Random. En: Jamenson, D. L. (Ed.) 1977. Evolutionary genetics. Dowden, Hutch & Ross, Inc. USA. p. 132-136.

Rao, R. C. 1965. Linear Statistical Inference and its Aplications. 2ª ed. John Wiley & Sons. USA. 625 p.

Robinson, H. F. 1986. (Ed.) Papers on Quantitative Genetics and Related Topics. North Carolina State Univ. USA. 354 p.

Sanin, O. I. 1984. Teoría de la Probabilidad. Limusa. Mex. 447 p.

Searle, S. R. 1971. Linear Models. John Wiley & Sons, Inc. USA. 532 p.

Wilks, S. 1962. Mathematical Statistics. John Wiley & Sons, Inc. USA. 644 p.

Wright, S. 1932. Evolution in Mendelian Populations. En: Robinson, H. F. 1986. (Ed.) Papers on Quantitative Genetics and Related Topics. North Carolina State Univ. USA. 467-477.

Weinberg, W. 1908. On the Demostration of Inheritance in Men. En: Jamenson, D. L. (Ed.) 1977. Evolutionary Genetics. Dowden, Hutch & ross, Inc. USA. p. 115-125.

Yule, G. U. 1906. On the Theory of Inheritance of Quantitative Compound Characters on the Basis of Mendel's Laws -A Preliminary Note-. En: Jamenson, D. L. (Ed.) 1977. Evolutionary Genetics. Dowden, Hutch & Ross, Inc. USA. p. 110-112.

APENDICE

LEMA AO

Sea Y un vector n-dimensional tal que

$$Y \sim N(\mu;\ \sigma^2 I)$$

donde $\mu \in W \subset \mathbb{R}^n$ y $\sigma^2 > 0$. Suponga que $T: \mathbb{R}^n \to \mathbb{R}$ es una función tal que

$$E_{\mu\sigma^2}[|T(Y)|] < \infty \text{ para todo } \mu \in W,\ \sigma^2 > 0. \qquad (A1)$$

Entonces, para cada $\sigma^2 > 0$ fija

$$Te(\mu) = E_{\mu\sigma^2}[T(Y)],\ \mu \in W \qquad (A2)$$

es una función continua. //

DEMOSTRACION

Sin pérdida de generalidad, suponga que en (A2), $\sigma^2 = 1$. Así veremos que para cada $\mu_0 \in W$

$$\lim_{\mu \to \mu_0} Te(\mu) = Te(\mu_0),$$

es decir

$$\lim_{\mu \to \mu_0} E_{\mu 1}[T(Y)] = E_{\mu_0 1}[T(Y)] \qquad (A3)$$

Ahora, usando (A1) con $\mu = \mu_0$ y $\sigma^2 = 2$ vemos que

$$\int \exp(-\|y - \mu_0\|^2 / 4)\ |T(y)|\ dy < \infty$$

y entonces podemos encontrar un número positivo $M = M(\varepsilon) > 0$ tal que

$$\int_{\|y-\mu_0\| > M} \exp(-\tfrac{1}{4}\|y - \mu_0\|^2)\ |T(y)|\ dy < \varepsilon, \qquad (A4)$$

donde $\varepsilon > 0$ es un número positivo arbitrario.

Observe que

$$\|y - \mu\|^2 = \|y - \mu_0\|^2 + 2(y - \mu_0)'(\mu - \mu_0) + \|\mu - \mu_0\|^2$$

$$\geq \|y - \mu_0\|^2 - 2\|y - \mu_0\| \, \|\mu - \mu_0\| + \|\mu - \mu_0\|^2, \qquad (A5)$$

donde usamos la desigualdad de Cauchy-Schwaz. Ahora recuerde que

$$ab \leq (a^2 + b^2)/2$$

y poniendo

$$a = \|y - \mu_0\|, \; b = 2\|\mu - \mu_0\|$$

se concluye que

$$2\|y - \mu_0\| \, \|\mu - \mu_0\| \leq (\|y - \mu_0\|^2 + 4\|\mu - \mu_0\|^2)/2$$

Luego, usando la última desigualdad junto con (A5) vemos que

$$\|y - \mu\|^2 \geq \|y - \mu_0\|^2 - \tfrac{1}{2}(\|y - \mu_0\|^2 + 4\|\mu - \mu_0\|^2) + \|\mu - \mu_0\|^2$$

$$= \tfrac{1}{2}\|y - \mu_0\|^2 - \|\mu - \mu_0\|^2.$$

Luego,

$$\exp(-\tfrac{1}{2}\|y - \mu\|^2) \leq \exp(-\tfrac{1}{4}\|y - \mu_0\|^2 + \tfrac{1}{2}\|\mu - \mu_0\|^2,$$

y vemos que

$$\int_{\|y-\mu_0\| > M} \exp(-\tfrac{1}{2}\|y - \mu\|^2 \, |T(y)| \, dy \leq \left(\int_{\|y-\mu_0\| \geq M} \exp(-\tfrac{1}{4}\|y - \mu_0\|^2 \, |T(y)| \, dy \right)$$

$$\exp(\tfrac{1}{2}\|\mu - \mu_0\|^2 \leq \exp(\tfrac{1}{2}\|\mu - \mu_0\|^2) \qquad (A6)$$

Es claro que

$$\exp(-\tfrac{1}{2}\|y - \mu_0\|^2) \leq \exp(-\tfrac{1}{4}\|y - \mu_0\|^2)$$

y entonces

$$\int_{\|y-\mu_0\| \geq M} \exp(-\tfrac{1}{2}\|y - \mu_0\|^2)|T(y)| \, dy \leq \int_{\|y-\mu_0\| \geq M} \exp(-\tfrac{1}{4}\|y - \mu_0\|^2)|T(y)| dy$$

Así

$$\left| \int_{\|y-\mu_0\| \geq M} \exp(-\tfrac{1}{2}\|y-\mu\|^2)T(y)dy - \int_{\|y-\mu_0\| \geq} \exp(-\tfrac{1}{2}\|y-\mu_0\|^2)T(y)dy \right| \leq \varepsilon(1+\exp(\tfrac{1}{2}\|\mu-\mu_0\|^2) \quad (A$$

Ahora note que para $\|y-\mu_0\| \leq M$

$$\|y-\mu\|^2 \geq \|y-\mu_0\|^2 - 2M\|\mu-\mu_0\| + \|\mu-\mu_0\|^2;$$

vea (A5), y similarmente

$$\|y-\mu\|^2 \leq \|y-\mu_0\|^2 + 2M\|\mu-\mu_0\| + \|\mu-\mu_0\|^2 .$$

Esto último puede obtenerse usando la ecuación (A5) junto con

$$(Y-\mu_0)'(\mu-\mu_0) \leq \|y-\mu_0\| \, \|\mu-\mu_0\| \leq M\|\mu-\mu_0\|$$

Entonces, para $\|y-\mu_0\| \leq M$

$$\exp(-\tfrac{1}{2}\|y-\mu_0\|^2)(\exp(-M\|\mu-\mu_0\| + \|\mu-\mu_0\|^2/2)) \leq \exp(-\tfrac{1}{2}\|y-\mu\|^2)$$

$$\leq \exp(-\tfrac{1}{2}\|y-\mu_0\|^2)(\exp(M\|\mu-\mu_0\| - \|\mu-\mu_0\|^2/2))$$

y entonces

$$\left|\exp(-\tfrac{1}{2}\|y-\mu\|^2) - \exp(-\tfrac{1}{2}\|y-\mu_0\|^2)\right| \leq \exp(-\tfrac{1}{2}\|y-\mu_0\|^2)\max(\exp(M\|\mu-\mu_0\| - \|\mu-\mu_0\|^2/2)-1)$$

$$= \exp(-\tfrac{1}{2}\|y-y_0\|^2)\,\psi(\mu,\mu_0)$$

Ahora, a partir de esta desigualdad se obtiene que

$$\left| \int_{\|y-\mu_0\| \leq M} \exp(-\tfrac{1}{2}\|y-\mu\|^2)T(y)dy - \int_{\|y-\mu_0\| \leq M} \exp(-\tfrac{1}{2}\|y-\mu_0\|^2)T(y)dy \right|$$

$$\leq \int_{\|y-\mu_0\|} \left|\exp(-\tfrac{1}{2}\|y-\mu\|^2) - \exp(-\tfrac{1}{2}\|y-\mu_0\|^2)\right| \, |T(y)| \, dy$$

$$\leq \left(\int_{\|y-\mu_0\| \leq M} \exp(-\tfrac{1}{2}\|y-\mu_0\|^2) \, |T(y)| \, dy\right) \psi(\mu,\mu_0)$$

$$\leq E_{\mu_0,1}(|T(Y)|)\psi(\mu,\mu_0) \qquad (A8)$$

Para concluir observe que (A7) y (A8) implican que

$$\left| \lim_{\mu \to \mu_0} \int \exp(-\tfrac{1}{2}\|y-\mu\|^2) \, T(y) \, dy - \int \exp(-\tfrac{1}{2}\|y-\mu_0\|^2)T(y) \, dy \right| \leq 2\varepsilon \qquad (A9)$$

puesto que $\psi(\mu,\mu_0) \to 0$ si $\mu \to \mu_0$. Entonces, ya que $\varepsilon > 0$ en (A9) es arbitrario concluimos que (A3) es válido. //

LEMA A1

Suponga que Y es un vector aleatorio n-dimensional con distribución $N(2ka\mathbb{1}; \sigma^2 I)$ con $a > 0$ y $\sigma^2 > 0$ arbitrarios.

Suponga que

i) $E_{a\sigma^2}[|T(Y)|] < \infty$ para toda a, $\sigma^2 > 0$

ii) $E_{a\sigma^2}[\sum_v P_v\, T(Y + cv)] = k$ para todo a, $\sigma^2 > 0$, $c > 0$.

Entonces, existen un número positivo δ^2 y una función continua Te tales que al reemplazar T por Te en (ii) la condición se siga cumpliendo.

DEMOSTRACION

Usando los argumentos del lema A0 es fácil ver que $|T(Y)|$ tiene esperanza finita si $Y \sim N(\mu, \sigma^2 I)$ para $\mu \in \mathbb{R}^n$ arbitrario y $\sigma^2 > 0$. Ahora defina $Te(\mu)$ como en el lema A0

$$Te(\mu) = (1/\sqrt{2\pi\delta^2})(\exp(-\tfrac{1}{2}\|y-\mu\|^2/\delta^2))T(y)\, dy,$$

donde δ^2 es un número fijo.

Por el lema A0, $Te(\mu)$ es una función continua de $\mu \in \mathbb{R}^n$. Ahora tome a, $c > 0$ y $\sigma^2 > \delta^2$ fijos. Escriba

$$\sigma^2_1 = \sigma^2 + \delta^2,$$

y observe que si

$$y \sim N(2ka; \sigma^2_1 I),$$

entonces Y puede escribirse como

$Y = Y^* + Z$, $Y^* \sim N(2ka\mathbb{1}; \sigma^2 I)$, $Z \quad N(0; \delta^2 I)$

Entonces

$$E_{a\sigma^2_1}[T(Y + cv)] = E[T(Y^* + Z + cv)]$$
$$= E[E[T(Y^* + cv + Z) \mid Y^* = y^*]]$$

$$= E[E[T(Y^* + cv + Z) \mid Y^*]] \qquad (A.1.1)$$

por el teorema de la doble esperanza, sin embargo

$$E[T(Y^* + cv + Z) \mid Y^* = y^*] = E[T(y^* + cv + z)]$$

$$= E_{y^* + cv,\ \delta^2}[T(Y)]$$

$$= Te(y^* + cv) \qquad (A.1.2)$$

donde usamos que $Z \sim N(0,\ \delta^2 I)$, y entonces

$$y^* + cv + Z \sim N(y^* + cv;\ \delta^2 I),$$

donde Te es continua y ademas tenemos que usando (A.1.1) y (A.1.2) se tiene

$$E_{a_1, \sigma_1^2}[T(Y + cv)] = E_{a_1, \sigma_1^2}[Te(Y^* + cv)]$$

y por lo tanto

$$k = E_{a,\ \sigma_1^2}[\sum_v P_v\ T(y + cv)]$$

$$= E_{a,\ \sigma_1^2}[\sum_v P_v Te(y^* + cv)]$$

lo cual concluye la demostración.

LEMA A2

Sea $G_{r\varepsilon} \in \mathbb{R}^{6n \times 1}$ la región para cada $r > 0$, $\varepsilon > 0$ tal que

$$G_{r\varepsilon} = \{y^* \in \mathbb{R}^{6n \times 1} \mid \|y^*\|^2 \leq r^2,\ y\mathbf{1} > (r-\varepsilon)\sqrt{6n}\}$$

$$= \{y^* \mid \|y^*\|^2 \leq r^2,\ r\sqrt{6n} \geq y^*\mathbf{1} > (r-\varepsilon)\sqrt{6n}\}$$

entonces

i) $G_{r\varepsilon}$ contiene un cubo de dimensión $6n$.

ii) $\| y^* - r\mathbf{1} / \sqrt{6n} \|^2 \leq 2r\varepsilon$

DEMOSTRACION

(i) es claro, mientras que (ii) se ve facilmente por la si - guiente desigualdad

$$\| y^* - r\mathbf{1} \| / \sqrt{6n} \|^2 \leq y^{*\prime} y^* + r^2 - 2r(r-\varepsilon)$$

$$\leq r^2 + r^2 - 2r^2 + 2r\varepsilon$$

$$\leq 2r\varepsilon. \; //$$