CARACTERES DISCRETOS Y CONTINUOS

Conocimientos previos requeridos

álgebra elemental

Conceptos previos requeridos

constante
variable
variable continua
variable discreta
muestra
muestra aleatoria
población
requerimientos de mantenimiento
consumo
sistema de producción
homocigosis, heterocigosis
gen, alelo, genotipo, fenotipo

Campo de interés de las secciones

AC – teórico - académico
EP – estaciones de prueba, pruebas diseñadas, pruebas con datos de campo
GE – cría y mejoramiento de poblaciones generales
PL – cría y mejoramiento de núcleos o planteles

Objetivos de aprendizaje

Identificación del tipo de variabilidad expresada en diferentes caracteres **[EP, GE, PL]**
Identificación de caracteres de importancia económica **[EP, GE, PL]**

CARACTERES DISCRETOS Y CONTINUOS

Lo diversidad de caracteres individuales considerados relevantes para la producción animal son numerosos y variados. Por su relación con el tipo de herencia que manifiestan, resulta conveniente distinguir entre caracteres que exhiben variación *discreta* de aquéllos que varían de manera *continua*. Ejemplos de los primeros son la tipificación en grupos sanguíneos, la presencia o ausencia de cuernos o astas, los colores de piel o fibras, y una extensa serie de anormalidades de origen genético (e.g. enanismo) identificables por signos o síndromes particulares. En general, *cualquier carácter cuya variación pueda clasificarse en términos de presencia o ausencia, o pertenencia a un limitado número de tipos o clases, es un carácter discreto.*

Los ejemplos de caracteres continuos son más abundantes. *Cualquier carácter medible en alguna escala y que no califique como discreto (i.e. para el que no sea posible identificar clases sin introducir algún tipo de arbitrariedad) es un carácter continuo*. Ejemplos de caracteres continuos son el peso al destete de un ternero, la producción de leche de una vaca, el tiempo que tarda un caballo en correr una milla, la proporción de suarda de un vellón, el espesor de grasa dorsal de un cerdo, etc.

Los extremos de esta simple clasificación se identifican sin dificultad. Sin embargo, existen caracteres que varían a lo largo de dimensiones difíciles de evaluar y que no pueden adjudicarse a clases sin introducir criterios arbitrarios. Ejemplos de esta índole son la resistencia o predisposición a enfermedades, las dificultades de parto, la habilidad materna, etc. Cuando no resulta posible identificar alguna variable continua subyacente estrechamente vinculada al atributo de interés (e.g. nivel de inmunoglobulinas en la resistencia a enfermedades, apertura pelviana en las dificultades de parto) la opción habitual es definir clases en forma arbitraria y tratar la variabilidad del carácter como discreta y posiblemente *ordinal* (e.g. no resistente, poco resistente, muy resistente), más que *nominal* (e.g. resistente, no resistente). Por otra parte, algunas variables intrínsecamente discretas, tales como el número de huevos que pone una gallina durante un ciclo de producción, se tratan habitualmente como variables continuas.

Para algunos caracteres discretos se conoce su modo de herencia. En muchos casos, la variabilidad observada se debe a la acción de uno o pocos pares de genes.

En contraste, la variabilidad genética responsable de las variaciones continuas se supone debida a la acción de muchos genes actuando en forma interactiva. Los métodos de estudio de ambos tipos de herencia son, en consecuencia, diferentes.

Caracteres discretos

La herencia de caracteres discretos puede estudiarse, al menos, a tres escalas de interés: individuos, familias, y poblaciones.

A *escala individual*, interesa generalmente caracterizar el *genotipo* de un individuo a partir de información usualmente incompleta (e.g. se sabe que el individuo no sufre determinada anormalidad de herencia simple, pero no se sabe si es portador de la misma, i.e. si es heterocigota para la característica).

A *escala familiar* suele interesar la identificación del *modo o mecanismo de herencia* de un carácter a partir de la información de un conjunto de parientes.

Cuando el interés se centra en *poblaciones*, el objetivo generalmente consiste en determinar tendencias futuras de las *frecuencias* (i.e. proporciones relativas) de alelos que controlan caracteres de interés económico, tanto con el propósito de reducir la frecuencia de alelos indeseables como de introducir alelos deseables, prevenir la pérdida de variabilidad genética, etc.

Con individuos y familias el interés se centra en la presencia o ausencia de atributos. A nivel de poblaciones, las variables de interés son las frecuencias de alelos, genotipos y fenotipos.

El tratamiento de caracteres continuos requiere de un enfoque estadístico diferente cuyos rudimentos se tratan a continuación.

Caracteres continuos

Una limitante intrínseca de la información continua es que requiere un contexto de referencia para adquirir sentido. Si se afirma :'ese toro tiene cuernos', no hace falta referencia alguna para valorar la presencia de cuernos. En contraste, si se afirma: 'ese toro pesa 400 kg', la información no es suficiente para valorar el peso del toro; para ello debe disponerse de un contexto de referencia tal como el peso de un grupo comparable de toros, o la variabilidad 'habitual' del peso de esos toros. La información cuantitativa *per se*, sin mención a una población de referencia carece, en general, de sentido. Por ese motivo, la información de tipo continuo se maneja habitualmente como *muestras*, i.e. series o conjuntos de datos tomados (usualmente al azar) de un conjunto mayor de referencia, la *población*, que idealmente contiene todos los casos.

Sea la serie de datos: Y_i = 11, 12, 14, 9, 12, 11, 8

¿Qué puede decirse acerca de ella?

Comenzando por lo más elemental, puede afirmarse que está constituida por 7 *elementos* o *casos*, lo que implica definir el *tamaño de la población* (N) o de la muestra (n). Posiblemente lo siguiente sea indicar que los casos *varían entre* 8 y 14 lo que determina el *rango* o diferencia entre valores extremos.

Si la serie constituyera una muestra aleatoria y a partir de esta información hubiera que responder a la pregunta:

'¿cuál sería el valor más probable de un nuevo dato tomado al azar de la misma población?'

una de las respuestas posibles que intuitivamente parecen razonables implicaría calcular el valor promedio de los datos ya disponibles. El concepto estadístico equivalente es el de operador *esperanza* (E):

$$\boxed{E(Y) = \frac{\sum_{i=1}^{N} Y_i}{N}}$$

La esperanza de una variable es su valor promedio en la población.

Si la población de referencia está constituida únicamente por esos N datos, entonces la esperanza de la variable es igual a la media poblacional (μ):

$$\boxed{E(Y) = \frac{\sum_{i=1}^{N} Y_i}{N} = \mu}$$

Por el contrario, si los datos representan una muestra de tamaño n tomada al azar de la población, entonces la esperanza de la variable es igual a un *estimador* (indicado por el símbolo ^) del verdadero valor del parámetro. A ese estimador en particular se lo conoce como *media muestral* (simbolizada también por una barra horizontal sobre el nombre de la variable):

$$E(Y) = \frac{\sum_{i=1}^{n} Y_i}{n} = \hat{\mu} = \overline{Y}$$

El Apéndice I detalla reglas de aplicación del operador esperanza.

Si consideramos a la serie de datos indicada más arriba como una muestra de tamaño n = 7, entonces:

$$E(Y) = \overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{11+12+14+9+12+11+8}{7} = 11$$

Si se agregara a la serie un nuevo dato (i.e. n_8) tomado al azar de la población original, esperaríamos que su valor resultara cercano a 11.

Considérese otra serie de datos para la misma variable tomada al azar de la misma población:

Y_i = 12, 10, 12, 10, 12, 11, 13

Si se calcula su media:

$$E(Y) = \overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{12+10+12+10+12+11+13}{7} = 11{,}43$$

resulta similar a la anterior.

Sin embargo, la *distribución* de los datos alrededor del valor medio parece mucho más compacta (i.e. más agrupada alrededor del valor medio) en esta segunda muestra. Si se requiriera predecir un dato futuro tomado al azar de la misma población, la confianza depositada en la predicción sería probablemente mayor si se contara con la segunda muestra en lugar de la primera. La razón es intuitiva: si los valores se dispersan menos alrededor del valor medio en la segunda muestra que en la primera, es probable que, si esa tendencia se mantiene (y no hay razón para pensar que no se mantenga...), evaluaremos la probabilidad que el nuevo dato sea cercano a la media como mayor en el segundo caso que en el primero.

Una medida razonable de dispersión surge de promediar los valores absolutos de los desvíos (i.e. Y_i – E(Y)):

$$D(Y) = \frac{\sum_{i=1}^{n} \left| Y_i - E(Y) \right|}{n}$$

Otra medida posible consiste en promediar los cuadrados de los desvíos:

$$V(Y) = \frac{\sum_{i=1}^{n} \left[Y_i - E(Y) \right]^2}{n}$$

A esta última expresión se la conoce como varianza (V) y tiene ciertas propiedades deseables que han popularizado su uso. En el Apéndice I se detallan reglas de aplicación del operador varianza. De la definición de esperanza resulta claro que:

$$V(Y) = \frac{\sum_{i=1}^{n} \left[Y_i - E(Y) \right]^2}{n} = E\left[Y - E(Y) \right]^2$$

Así como se distinguen media poblacional y media muestral, se distinguen también varianza poblacional (usualmente indicada con σ^2):

$$\sigma^2 = \frac{\sum_{i=1}^{N} (Y_i - \mu)^2}{N}$$

y varianza muestral (usualmente indicada con s^2):

$$s^2 = \frac{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}{n - 1}$$

La razón de dividir por n – 1 en lugar de hacerlo por n tiene que ver con la cantidad de información independiente utilizada. Para calcular cada desvío, previamente debe calcularse el valor de la media muestral. Definido ese valor, el número de desvíos independientes posibles es uno menos que el número total. A cada dato le corresponde un desvío, pero una vez calculados n – 1 desvíos, el valor del último desvío corresponde a la suma de los ya calculados restada de cero.

En la práctica, se utilizan fórmulas de trabajo equivalentes pero más sencillas de operar (derivación en Apéndice II):

$$\sigma^2 = \frac{\sum_{i=1}^{N} Y_i^2 - \frac{\left(\sum_{i=1}^{N} Y_i\right)^2}{N}}{N} \qquad s^2 = \frac{\sum_{i=1}^{n} Y_i^2 - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}}{n-1}$$

Para la primera muestra:

$$s_1^2 = \frac{11^2 + \cdots + 8^2 - \frac{(11 + \cdots + 8)^2}{7}}{7-1} = 4,0$$

Para la segunda:

$$s_2^2 = \frac{12^2 + \cdots + 13^2 - \frac{(12 + \cdots + 13)^2}{7}}{7-1} = 1,29$$

Otro atributo de interés para variables continuas es la *covarianza*, que cuantifica el grado de asociación entre variables. Se calcula como el promedio de los productos cruzados de los desvíos. Si X e Y son dos variables aleatorias:

$$\mathrm{cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

La fórmula de trabajo correspondiente es:

$$\mathrm{cov}(X, Y) = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n}}{n-1}$$

El grado de asociación entre las dos muestras consideradas previamente puede calcularse como:

$$\mathrm{cov}(Y_1, Y_2) = \frac{12 \times 11 + \cdots + 13 \times 8 - \frac{[(12 + \cdots + 13)(11 + \cdots + 8)]}{7}}{7-1} = 0,33$$

El *signo* de una covarianza indica el tipo de asociación entre variables. Una covarianza positiva se origina cuando valores altos de una variable se corresponden con valores altos de la otra y lo mismo ocurre para los valores bajos. Por el contrario, cuando valores altos de una variable se corresponden con valores bajos de la otra, la covarianza será negativa. Si no existe un patrón de asociación entre valores correspondientes, la covarianza se aproximará a cero.

El *valor absoluto* de una covarianza carece de sentido por si mismo. Para valorar su magnitud debe comparársela con la variabilidad de las variables cuyo grado de asociación se desea estimar. Una forma de hacerlo es a través del *coeficiente de correlación* (r) que relaciona la covarianza a la *media geométrica* (Apéndice III) de las variables:

$$\boxed{r_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sqrt{V(X)V(Y)}}}$$

El coeficiente de correlación puede variar entre –1 (correlación perfecta y negativa) y +1 (correlación perfecta y positiva). Valores cercanos a cero indican ausencia de asociación. Para el ejemplo considerado:

$$\boxed{r_{Y_1,Y_2} = \frac{\mathrm{cov}(X,Y)}{\sqrt{V(X)V(Y)}} = \frac{0{,}33}{\sqrt{4 \times 1.29}} = 0{,}147}$$

Lo que indica que las muestras no parecen estar asociadas. Ello es esperable tratándose de muestras aleatorias de una misma variable. La utilidad de la covarianza, el coeficiente de correlación, y otras medidas de asociación consiste en permitir evaluar el grado de relación entre *diferentes* variables medidas sobre los mismos sujetos o entre mediciones de la misma variable en diferentes sujetos, más que entre diferentes muestras de una misma variable.

Caracteres de importancia económica

Cualquier carácter, tanto discreto como continuo, puede ser de importancia económica si su variabilidad influye en el resultado económico de la empresa. Las fibras pigmentadas en vellones blancos constituyen un ejemplo. Con independencia de otros factores, a mayor proporción de ellas, menores serán los ingresos que se perciban debido a una reducción de calidad de la lana comercializada. En la medida en que el carácter presente variabilidad heredable, será un carácter de importancia económica que podría mejorarse genéticamente.

Para muchos caracteres de importancia económica, particularmente los que exhiben variabilidad continua, la situación no es tan clara. Es seguramente conveniente desde un punto de vista económico lograr terneros más pesados al destete o vacas con mayor producción de leche por cabeza. Sin embargo, la mayor producción debe estar asociada a mayor eficiencia de producción (e.g. mejor uso del forraje o suplementos, menores pérdidas productivas por stress de frío o calor, etc.) ya que de lo contrario puede verificarse el absurdo de obtener peores resultados económicos con animales más productivos. Ello es de ocurrencia común cuando las mayores producciones se logran en base a mayor consumo de recursos en lugar de incrementos de eficiencia.

Como no existen animales que puedan ser los mejores en todas y cada una de las características de importancia económica, el número de éstas a tomar en cuenta en un plan de mejoramiento debe limitarse únicamente a las más importantes. Por ese motivo, el mejoramiento animal debe apoyarse en modelos de producción animal que representen, de la manera más realista posible, las acciones e interacciones que determinan la eficiencia de producción a nivel de empresa. De esa realidad debe partirse para identificar caracteres medibles en individuos y cuya variabilidad contribuya en forma significativa al resultado de eficiencia productiva evaluado a nivel de población. Mejorar genéticamente el peso al destete de los terneros tiene sentido si ello no incrementa los requerimientos de mantenimiento de madres de potencialmente mayor tamaño, o el nivel de dificultades de parto.

Resulta evidente entonces que el tipo de caracteres a mejorar y su importancia relativa

- no pueden definirse en abstracto sino vinculados a un sistema de producción y
- pueden resultar diferentes en diferentes sistemas.

Los ejemplos de mejoramiento genético más exitosos son los de caracteres de importancia económica fácilmente evaluables en ambos sexos a edad temprana, que guardan estrecha relación con la eficiencia de producción del sistema y cuyo mejoramiento se debe a un uso más eficiente de recursos y no simplemente a mayor uso de éstos.

En cualquier sistema de producción animal, la mayor parte del consumo de energía se verifica simplemente para satisfacer los requerimientos de mantenimiento de los animales. Los incrementos de producción por cabeza son deseables siempre que sean al menos proporcionales a los incrementos de consumo que seguramente los acompañan. De esa forma, los requerimientos de mantenimiento se 'diluyen' en una mayor producción aumentando, en consecuencia, la eficiencia y mejorando el resultado económico.

Bibliografía

Loether, H.J. y McTavish, D.G. 1988. Descriptive and inferential statistics. 3rd. Edition. Alvin and Bacon. Boston, USA.

McClave, J.T. y Dietrich, F.H. II. 1985. Statistics.3rd. Edition. Dellen-McMillan. San Francisco, Cal. USA.

Apéndice I

Reglas de aplicación del operador esperanza (E)

1. Si a es una constante:

$$E(a) = a$$

2. Si Y es una variable aleatoria y a es una constante:

$$E(a + Y) = a + E(Y)$$

$$E(aY) = a\,E(Y)$$

3. Si x e y son variables aleatorias estadísticamente independientes:

$$E(XY) = E(X)\,E(Y)$$

Reglas de aplicación del operador varianza (V)

1. Si a es una constante:

$$V(a) = 0$$

2. Si Y es una variable aleatoria y a es una constante:

$$V(aY) = \frac{\sum_{i=1}^{N}\left[aY_i - E(aY_i)\right]^2}{N} = \frac{\sum_{i=1}^{N}\left[aY_i - aE(Y)\right]^2}{N} = \frac{a^2\sum_{i=1}^{N}\left[Y_i - E(Y)\right]^2}{N}$$

$$V(aY) = a^2 V(Y)$$

$$V(a + Y) = V(a) + V(Y) = V(Y)$$

3. Si X e Y son variables aleatorias:

$$V(X+Y)=\frac{\sum_{i=1}^{N}\left[(X_i+Y_i)-E(X_i+Y_i)\right]^2}{N}=\frac{\sum_{i=1}^{N}\left[(X_i+Y_i)-E(X)-E(Y)\right]^2}{N}$$

$$V(X+Y)=\frac{\sum_{i=1}^{N}\left[(X_i-E(X))+(Y_i-E(Y))\right]^2}{N}=$$

$$V(X+Y)=\frac{\sum_{i=1}^{N}\left[(X_i-E(X))^2+(Y_i-E(Y))^2+2(X_i-E(X))(Y_i-E(Y))\right]}{N}$$

$$V(X+Y)=\frac{\sum_{i=1}^{N}(X_i-E(X))^2}{N}+\frac{\sum_{i=1}^{N}(Y_i-E(Y))^2}{N}+\frac{2\sum_{i=1}^{N}(X_i-E(X))(Y_i-E(Y))}{N}$$

El tercer término en esa expresión es dos veces el promedio de los productos cruzados de los desvíos de las variables, una medida de asociación entre variables denominada *covarianza* (cov; ver texto). Entonces:

$$V(X+Y)=V(X)+V(Y)+2\,\mathrm{cov}(X,Y) \qquad \text{y}$$
$$V(X-Y)=V(X)+V(Y)-2\,\mathrm{cov}(X,Y)$$

Reglas de aplicación del operador covarianza (cov)

1. Si X e Y son variables aleatorias y a es una constante:

$$\mathrm{cov}(aX,Y) = a\ \mathrm{cov}(X,Y)$$

2. Si x es una variable alatoria:

$$\mathrm{cov}(X,X) = V(X)$$

Apéndice II

Si Y es una variable aleatoria, entonces:

$$V(Y) = \sigma^2 = \frac{\sum_{i=1}^{N}\left(Y_i - \frac{\sum_{i=1}^{N} Y_i}{N}\right)^2}{N} = \frac{\sum_{i=1}^{N}\left(Y_i^2 - 2Y_i\frac{\sum_{i=1}^{N} Y_i}{N} + \frac{\left(\sum_{i=1}^{N} Y_i\right)^2}{N^2}\right)}{N}$$

$$V(Y) = \frac{\sum_{i=1}^{N} Y_i^2 - 2\sum_{i=1}^{N} Y_i\left(\frac{\sum_{i=1}^{N} Y_i}{N}\right) + \sum_{i=1}^{N}\frac{\left(\sum_{i=1}^{N} Y_i\right)^2}{N^2}}{N} = \frac{\sum_{i=1}^{N} Y_i^2 - 2\frac{\left(\sum_{i=1}^{N} Y_i\right)^2}{N} + N\frac{\left(\sum_{i=1}^{N} Y_i\right)^2}{N^2}}{N}$$

$$V(Y) = \frac{\sum_{i=1}^{N} Y_i^2 - \frac{\left(\sum_{i=1}^{N} Y_i\right)^2}{N}}{N}$$

Apéndice III

La media geométrica (G) de una serie (Y_i) de longitud N se calcula como la raíz enésima del producto de los valores en la serie:

$$\boxed{G(Y_i) = \sqrt[N]{\prod_{i=1}^{N} Y_i}}$$