2 BIOESTADISTICA

Este capítulo contiene una breve descripción de algunos métodos estadísticos de uso común en la biología pesquera tropical e introduce el sistema de notación estadística adoptado en el manual. Puede servir para repasar nociones y para consulta, pero no pretende ser en sí mismo un texto de estudio.

La cantidad de literatura que existe sobre métodos estadísticos es asombrosa; por lo tanto, quienes quieran aprender más sobre bioestadística no tendrán ningún problema. Aquí señalamos sólo dos referencias: el libro "Biometry" de Sokal y Rohlf (1981), que explica la teoría de un modo más accesible, y "Sampling Techniques", de Cochran (1977), que quizá sea un poco más complicado, pero que también se recomienda como introducción. Sin embargo, existen otros textos igualmente útiles.

2.1 MEDIA Y VARIANZA

Considérese una muestra de n peces, todos de la misma especie, capturados en un lance de arrastre y sea x(i) la talla del i-ésimo pez, i = 1, 2, ... , n. La "*talla media*" (en general la "*media*"), de la muestra se define como:

$$\bar{x} = [x(1) + x(2) + ... + x(n)]/n = \frac{1}{n} * \sum_{i=1}^{n} x(i) \qquad (2.1.1)$$

Las dos primeras columnas de la Tabla 2.1.1 muestran un ejemplo con n = 27.

La varianza, que es una medida de la variabilidad en torno a la media, se define de la siguiente forma:

$$s^2 = \frac{1}{n-1} * [(x(1)-\bar{x})^2 + (x(2)-\bar{x})^2 + ... + (x(n)-\bar{x})^2] =$$

$$= \frac{1}{n-1} * \sum_{i=1}^{n} [x(i)-\bar{x}]^2 \qquad (2.1.2)$$

Así, la varianza, s^2, es la suma de los cuadrados de las desviaciones respecto de la media, dividida por el número n-1. La tercera y cuarta columna de la Tabla 2.1.1 ilustran el cálculo de la varianza. Obsérvese que si todos los peces de la muestra tuviesen la misma talla igualarían la talla media y la varianza sería cero. La suma de las desviaciones (no al cuadrado) es siempre cero. Mientras mayor sea la desviación respecto de la media, mayor será la varianza. En la Tabla 2.1.1, los dos valores más grandes de los cuadrados de las desviaciones respecto de la media se registran en las observaciones más pequeña y más grande.

La raíz cuadrada de la varianza, s, se denomina "*desviación estándar*". A menudo interesa determinar la varianza relativa al tamaño de la talla media; para ello, s es la cantidad apropiada, ya que tiene la misma unidad que la media. Esto conduce a la desviación estándar relativa $s/\bar{x}$, también llamada "*coeficiente de variación*".

Cuando los cálculos se hacen manualmente, es más fácil trabajar con una forma reordenada de la Ecuación 2.1.2, que es equivalente a:

$$s^2 = \frac{1}{n-1} * \left[\sum_{i=1}^{n} x(i)^2 - \frac{1}{n} * \left[\sum_{i=1}^{n} x(i) \right]^2 \right] \qquad (2.1.3)$$

TABLA 2.1.1
Media, varianza y desviación estándar de una muestra de frecuencias de tallas

pez (nº) i	talla (cm) x(i)	desviación respecto a la media $x(i)-\bar{x}$	cuadrado de la desviación respecto a la media $(x(i)-\bar{x})^2$
1	14.2	-0.87	0.75
2	16.3	1.23	1.52
3	14.8	-0.27	0.07
4	13.2	-1.87	3.48
5	16.9	1.83	3.36
6	12.4	-2.67	7.11
7	14.3	-0.77	0.59
8	15.7	0.63	0.40
9	15.3	0.23	0.05
10	11.2 (mín.)	-3.87	14.95
11	12.9	-2.17	4.69
12	13.5	-1.57	2.45
13	18.2	3.13	9.82
14	11.6	-3.47	12.02
15	18.5	3.43	11.79
16	16.3	1.23	1.52
17	15.5	0.43	0.19
18	15.8	0.73	0.54
19	13.2	-1.87	3.48
20	19.0 (máx.)	3.93	15.47
21	12.0	-3.07	9.40
22	17.1	2.03	4.13
23	15.4	0.33	0.11
24	14.6	-0.47	0.22
25	14.0	-1.07	1.14
26	18.1	3.03	9.20
27 = n	16.8	1.73	3.00
Total	406.8 $= \Sigma x(i)$	0.00	121.48 $= \Sigma(x(i)-\bar{x})^2$

talla media, $\bar{x}$	: 406.8/27 = 15.07
varianza, s^2	: 121.48/(27–1) = 4.67
desviación estándar, s	: $\sqrt{4.67}$ = 2.16
desviación estándar relativa, $s/\bar{x}$	: 2.16/15.07 = 0.14
error estándar, $s/\sqrt{n}$	: 2.16/$\sqrt{27}$ = 0.41

(El concepto de error estándar se introduce en la Sección 2.3)

Sin embargo, como la mayoría de las calculadoras científicas de bolsillo tienen la posibilidad de calcular automáticamente la media y la varianza, los cálculos se ilustran aquí con la Ec. 2.1.2 que conceptualmente es más fácil de entender.

Para muchos propósitos, por ejemplo para representaciones gráficas, es conveniente disponer

la muestra en forma de una "*tabla de frecuencias*", dividiendo el recorrido de las tallas en varios intervalos de longitud. En la muestra de la Tabla 2.1.1 el recorrido de las tallas va desde 11.2 a 19.0 cm. Con grupos de tallas de 1 cm se necesitan nueve grupos para cubrir el recorrido. Tomando 10.5 como límite inferior del primer intervalo, los intervalos y las frecuencias de tallas serían los que aparecen en las primeras cuatro columnas de la Tabla 2.1.2, que es la llamada tabla de frecuencias de tallas.

TABLA 2.1.2
Media y varianza de una muestra de frecuencias de tallas. (La muestra se obtuvo de la Tabla 2.1.1, con un intervalo de talla, dL, de 1 cm)

índice	intervalo (cm)	punto medio (cm)	frecuencia			
j	L(j)–L(j)+dL	$\bar{L}(j)$	F(j)	F(j)*$\bar{L}$(j)	($\bar{L}$(j)–$\bar{x}$)	F(j)*($\bar{L}$(j)–$\bar{x}$)^2
1	10.5 - 11.5	11	1	11	-4.074	16.60
2	11.5 - 12.5	12	3	36	-3.074	28.35
3	12.5 - 13.5	13	3	39	-2.074	12.91
4	13.5 - 14.5	14	4	56	-1.074	4.61
5	14.5 - 15.5	15	4	60	-0.074	0.02
6	15.5 - 16.5	16	5	80	0.926	4.29
7	16.5 - 17.5	17	3	51	1.926	11.13
8	17.5 - 18.5	18	2	36	2.926	17.12
9	18.5 - 19.5	19	2	38	3.926	30.83
total			27	407		125.86

talla media, $\bar{x}$: 407/27 = 15.074, es decir 15.07
varianza, s^2 : 125.86/26 = 4.84
desviación estándar, s : $\sqrt{4.84}$ = 2.20
desviación estándar relativa, s/ $\bar{x}$: 2.20/15.07 = 0.15

Sea j el índice de un grupo de tallas y denótense los límites inferior y superior del grupo de tallas j respectivamente por:

$$L(j) = L(1) + (j-1)*dL \text{ y } L(j+1) = L(1) + j*dL,$$

o

$$L(j+1) = L(j) + dL$$

donde dL es la "*amplitud del intervalo*". Entonces, un pez de talla x(j) pertenecerá al grupo de tallas j cuando

$$L(j) <= x(j) < L(j) + dL$$

Sea F(j) la frecuencia del grupo de tallas j, es decir, el número de peces que se observan en ese grupo. Sea $\bar{L}$ (j) = L(j) + dL/2 el punto medio del grupo de tallas j, que es llamado la "*marca de clase*". El cálculo de la media y de la varianza a partir de una tabla de frecuencias se realiza del modo habitual utilizando los puntos medios para representar los intervalos:

$$n = \sum_{j=1}^{m} F(j)$$ es el número total de observaciones, donde m es el número de grupos de longitud.

$$\bar{x} = \frac{1}{n} * \sum_{j=1}^{m} F(j)*\bar{L}(j)$$ es la media, y

$$s^2 = \frac{1}{n-1} * \sum_{j=1}^{m} F(j)*[\bar{L}(j)-\bar{x}]^2$$ es la varianza.

El procedimiento de cálculo se presenta en la Tabla 2.1.2. El punto medio de la clase $\bar{L}(j)$ y el cuadrado de las desviaciones respecto de la media están ponderados por el número de peces de cada clase, es decir, la frecuencia, F(j). Los resultados de la Tabla 2.1.2 se desvían levemente respecto de los de la Tabla 2.1.1 porque una representación en grupos de centímetros produce resultados menos precisos que una representación en grupos de milímetros.

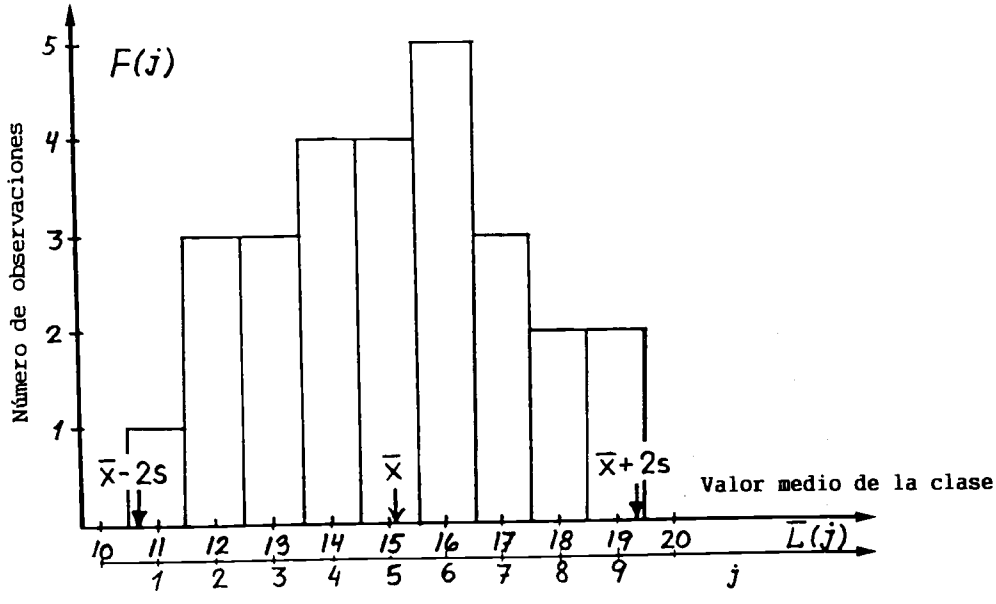


Fig. 2.1.1 Gráfico de frecuencias de tallas. Representación gráfica de la muestra de frecuencias de tallas de la Tabla 2.1.2.

En la Figura 2.1.1 aparece una representación gráfica de la muestra de frecuencias. Nótese que todas las observaciones se hallan en el intervalo de

$$\bar{x} - 2*s \quad \text{a} \quad \bar{x} + 2*s$$

A efectos de la llamada distribución normal (que se estudia en la próxima sección), se supone que cerca del 95% de las observaciones están contenidas en ese intervalo.

(Véanse los **Ejercicios** en la Parte 2).

2.2 DISTRIBUCION NORMAL

La Tabla 2.1.2 y la Fig. 2.1.1 muestran como ejemplo un pequeño conjunto de datos de frecuencias

de tallas que se ajustan aproximadamente a la llamada "*distribución normal*". La expresión matemática de una distribución normal es:

$$Fc(x) = \frac{n*dL}{s*\sqrt{2\pi}} * \exp[-(x-\bar{x})^2/(2s^2)] \qquad (2.2.1)$$

donde Fc = "*frecuencia calculada*" o "*frecuencia teórica*", n = número de observaciones, dL = tamaño del intervalo, s = desviación estándar, $\bar{x}$ = talla media y π = 3.14159.

Utilizando los valores n = 27, dL = 1 cm, s = 2.20 y $\bar{x}$ = 15.07 cm, de la Tabla 2.1.2, se tiene:

$$Fc(x) = \frac{27*1}{2.20*\sqrt{2*3.14159}} * \exp[-(x-15.07)^2/(2*4.84)] =$$

$$= 4.896*\exp[-(x-15.07)^2/9.68]$$

Los valores de Fc para una serie de valores de x se listan en la Tabla 2.2.1. Obsérvese que la notación se ha modificado algo ya que ahora se usa el punto medio del intervalo, x, como el argumento en Fc, en lugar del índice del intervalo, j, que se usó como argumento en F en la Tabla 2.1.2.

La Fig. 2.2.1 muestra las frecuencias teóricas junto con el gráfico de barras para F(j) de la Fig. 2.1.1. Como puede verse, Fc(x) da un ajuste aceptable con respecto a las frecuencias de tallas observadas. Este cuadro se observa a menudo cuando se registran frecuencias de tallas de peces que provienen de una cohorte, es decir, peces de aproximadamente la misma edad.

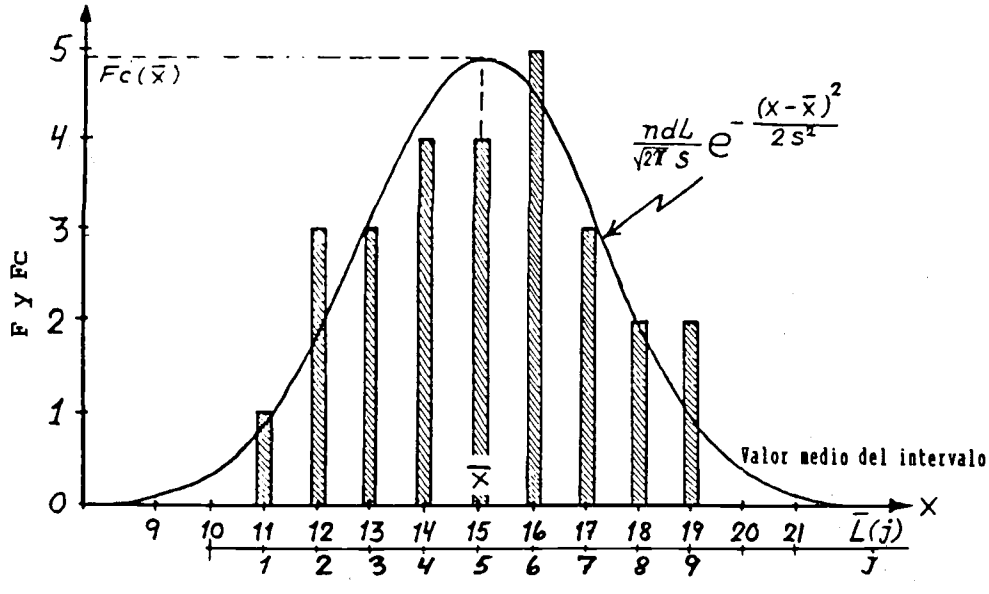


Fig. 2.2.1 Frecuencia teórica, Fc, (la curva de distribución normal) y las frecuencias observadas, F (barras).

TABLA 2.2.1
Frecuencias teóricas correspondientes a la Tabla 2.1.2, en donde x representa la marca de clase (punto medio del intervalo)

x	11	12	13	14	15	16	17	18	19
Fc(x)	0.88	1.85	3.14	4.35	4.89	4.48	3.33	2.02	0.99

La distribución normal se observa en una gran variedad de casos diferentes y de ahí su nombre. Hay otros tipos de distribuciones de probabilidad que se observan en la ciencia pesquera. Entre ellas están la "distribución log-normal", la "distribución binomial negativa" y la "distribución delta". Una diferencia notoria entre éstas y la distribución normal es que son asimétricas, en tanto que la distribución normal es simétrica. La distribución delta, por ejemplo, se utiliza para describir la distribución de probabilidad de la captura por hora de arrastre. Se compone de una distribución log-normal, que describe la distribución de las capturas al arrastre distintas de cero, y una probabilidad especial para la captura cero (véase en la Sección 13.7 la Fig. 13.7.2).

Quizás la característica más importante de la distribución normal sea la que tiene que ver con las medias. Si se toman, por ejemplo, 50 muestras aleatorias de una determinada población, cada una de ellas de 25 observaciones, los cincuenta valores medios se distribuirán (aproximadamente) de forma normal. La media (de cualquier conjunto de observaciones), tiene una distribución (aproximadamente) normal. Este resultado es también válido para las medias de distribuciones log-normales, distribuciones delta o cualquier otro tipo de distribución. Esto significa que las medias de todas las distribuciones observadas en biología pesquera se distribuyen aproximadamente de forma normal.

Si ambos lados de la Ec. 2.2.1 se dividen por n (= tamaño de la muestra), se tiene:

$$Fc(\bar{L}(j))/n = \frac{dL}{s*\sqrt{2\pi}} * \exp\left[-\frac{(\bar{L}(j)-\bar{x})^2}{2s^2}\right] \qquad (2.2.2)$$

con $j = 1, 2 \ldots, 9$ y $dL = 1$.

Los nuevos valores encontrados, Fc(x)/n, sumarán casi 1.0. Cada valor indica la probabilidad de que un pez extraído aleatoriamente pertenezca al intervalo de tallas correspondiente. Es decir, se puede interpretar como la probabilidad de que un pez extraído al azar pertenezca al intervalo de tallas de x–dL/2 a x+dL/2.

Para los nueve intervalos de tallas de la Tabla 2.2.1 se obtiene:

j	intervalo	probabilidad
1	10.5 - 11.5	0.033
2	11.5 - 12.5	0.069
3	12.5 - 13.5	0.116
4	13.5 - 14.5	0.161
5	14.5 - 15.5	0.181
6	15.5 - 16.5	0.166
7	16.5 - 17.5	0.123
8	17.5 - 18.5	0.075
9	18.5 - 19.5	0.037
	Total:	0.961

Así, por ejemplo, hay 181 posibilidades sobre 1000 de que un pez extraído al azar tenga una talla entre 14.5 y 15.5 cm. Si se hubiesen incluido todos los intervalos de tallas (y no sólo los nueve para los que se disponía de observaciones), las probabilidades habrían sumado 1.000.

La distribución normal se usará en los próximos capítulos en los análisis de frecuencias de tallas, porque la distribución por tallas de una cohorte de peces se puede describir mediante una distribución normal. A modo de introducción, estudiaremos algunos de sus aspectos.

Los procedimientos para calcular la media y la desviación estándar (Tabla 2.1.2) pueden aplicarse a cualquier conjunto de datos de frecuencias de tallas. Sin embargo, si por alguna razón el gráfico de las frecuencias observadas no representa la distribución total, los valores obtenidos (de las Ecs. 2.1.1 y 2.1.2) para la media y varianza muestrales estarán sesgados, es decir, pueden no guardar relación con la media y varianza poblacionales. El concepto de "*sesgo*" se tratará con más detalle en la Sección 7.1. Si, por ejemplo, sólo se dispone de las frecuencias en el intervalo de tallas de 10 a 15 cm (o sea, sólo los datos del lado izquierdo), se está en una situación en que la Ec. 2.1.1 (media) y la

Ec. 2.1.2 (varianza) no representan a la población. Como se verá en el Capítulo 3, esto ocurre a menudo cuando se analizan las frecuencias de tallas. Sin embargo, hay varios métodos para resolver el problema.

(Véanse los **Ejercicios** en la Parte 2).

2.3 LIMITES DE CONFIANZA

En esta sección también utilizaremos el ejemplo de una muestra de composición por tallas de los peces de una cohorte. La talla media de la cohorte, $\bar{x}$, se ha estimado a partir de la muestra. Esta estimación suele ser diferente de la media verdadera de la población, que es la que se obtendría si se midiesen todos los peces de esa cohorte en el mar. Generalmente la talla media verdadera se desconoce. Si se tratara de una población de peces cultivados en un estanque, se podría medir la talla media verdadera de esa población, pero en el caso de los peces en libertad es imposible determinar el valor real de cualquier parámetro. En la práctica esto es aplicable también a la población de peces capturados en una pesquería, puesto que no se pueden medir todos los peces capturados. Nos ocuparemos, pues, del grado de precisión de la estimación de la talla media, en otras palabras, de la probable magnitud de la desviación entre la estimación y la media verdadera. Esta incertidumbre acerca de la media verdadera se expresa por medio de los "*límites de confianza*". En el caso de una distribución normal, tales límites de confianza están dados por:

$$\bar{x} - t_{n-1}*s/\sqrt{n} \quad \text{y} \quad \bar{x} + t_{n-1}*s/\sqrt{n} \tag{2.3.1}$$

donde n es el tamaño de la muestra, s la desviación estándar y t(n–1) son los llamados percentiles en la "*distribución t de Student*" (Tabla 2.3.1). El argumento "f" en la distribución t (Tabla 2.3.1) se denomina "*número de grados de libertad*". En general el número de grados de libertad es el número de observaciones menos el número de parámetros. En este caso $\bar{x}$ es el único parámetro, por lo que $f = n-1$ y $t_f = t_{n-1}$ (véase la Tabla 2.3.1).

Los límites de confianza pueden ser calculados con diferentes niveles de precisión, usualmente 90%, 95% y 99%, como se indica en la Tabla 2.3.1. Mientras más alto sea el nivel (porcentaje), serán mayores los cuantiles y por lo tanto los intervalos serán más anchos entre los límites superiores e inferiores.

Volviendo al ejemplo de la Sección 2.1 (Tabla 2.1.2), si se quieren calcular, por ejemplo, los límites de confianza del 95% para la talla media de los peces de la población de la que se extrajo la muestra, se utiliza el percentil 95% de la distribución t (Tabla 2.3.1), con n–1 = 26 grados de libertad, y se inserta en la Ec. 2.3.1.:

$$t_{n-1}*s/\sqrt{n} = 2.06*2.20/\sqrt{27} = 0.87, \text{ mientras } \bar{x} = 15.07$$

los límites de confianza al 95% serán:

límite inferior: $\bar{x} - 0.87 = 15.07 - 0.87 = 14.20$
límite superior: $\bar{x} + 0.87 = 15.07 + 0.87 = 15.94$

Así, se tiene un "95% de confianza" de que la verdadera talla media se sitúa en algún lugar entre 14.20 y 15.94; en otras palabras, si el muestreo se repitiese 100 veces bajo las mismas condiciones, cabe prever que 95 de las medias se situarían entre 14.20 y 15.94. El intervalo entre el límite inferior y el límite superior se llama "*intervalo de confianza*".

Para el ejemplo utilizado anteriormente, los intervalos de confianza en los niveles de 90% y 99% son respectivamente [14.35, 15.79] y [13.89, 16.25], de los cuales el primero es más angosto y el segundo más ancho que el intervalo del 95%.

TABLA 2.3.1
Valores de los cuantiles de la distribución t (Distribución t de Student)*

grados de libertad f	cuantiles 90% t_f	95% t_f	99% t_f	grados de libertad f	cuantiles 90% t_f	95% t_f	99% t_f
1	6.31	12.71	63.66	15	1.75	2.13	2.95
2	2.92	4.30	9.93	16	1.75	2.12	2.92
3	2.35	3.18	5.84	17	1.74	2.11	2.90
4	2.13	2.78	4.60	18	1.73	2.10	2.88
5	2.02	2.57	4.03	19	1.73	2.09	2.86
6	1.94	2.45	3.71	20	1.73	2.09	2.85
7	1.90	2.37	3.50	25	1.71	2.06	2.79
8	1.86	2.31	3.36	30	1.70	2.04	2.75
9	1.83	2.26	3.25	40	1.68	2.02	2.70
10	1.81	2.23	3.17	50	1.67	2.01	2.68
11	1.80	2.20	3.11	60	1.67	2.00	2.66
12	1.78	2.18	3.06	80	1.67	1.99	2.64
13	1.77	2.16	3.01	100	1.66	1.98	2.63
14	1.76	2.15	2.98	∞	1.65	1.96	2.58

*El uso de la letra t en este contexto es universal. En este manual t también se utiliza para representar la edad de un pez. Esta tabla ha sido repetida en la última página de este volumen para facilitar su empleo.

La cantidad $s/\sqrt{n}$ es la desviación estándar de la estimación de la talla media (también llamado el "*error estándar*"), de modo que $\bar{x}$ tiene la varianza (compare con la Tabla 2.1.1):

$$VAR(\bar{x}) = s^2/n \qquad (2.3.2)$$

Así, cuánto más grande sea la muestra, más precisa será la estimación de $\bar{x}$ (este tema se tratará más detenidamente en la Sección 7.2).

La Ec. 2.3.2 deriva de dos reglas generales para variables aleatorias que se aplican repetidamente en este manual, a saber:

$$VAR\ (Cx) = C^2*VAR(x) \qquad (2.3.3)$$

$$VAR\ \left(\sum_{i=1}^{n} x\right) = n*VAR(x) \qquad (2.3.4)$$

donde C es una constante. Por ejemplo, si la varianza de x es s^2, entonces la varianza de 3x será $9s^2$; o bien, si las observaciones originales se suman de tres en tres, la varianza de $x_1 + x_2 + x_3$ será $3*s^2$.

Las afirmaciones anteriores sobre los límites de confianza se aplican sólo a estimaciones "insesgadas" de la media. Cuando las muestras están sesgadas, independientemente de cuántos peces se muestreen y se midan, siempre se obtendrán estimaciones de la media que serán diferentes de la media verdadera.

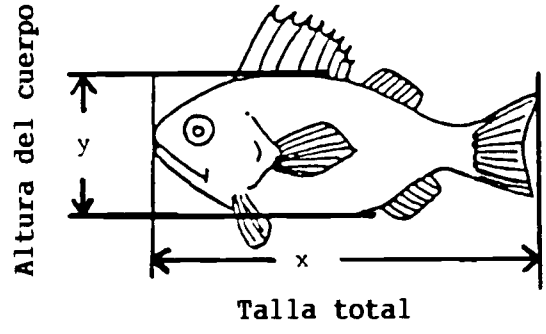
Supóngase que se desea estimar la talla media de cierta especie de peces capturados en una pesquería comercial (téngase presente que los peces capturados son los que se desembarcan más los que se descartan en el mar). Si se muestrean sólo ejemplares de los desembarques, y no de los peces que se descartan en el mar, normalmente inferiores a cierta talla, se obtiene una estimación sesgada de la talla media de los peces capturados. La talla media de la captura estará sobreestimada, sea cual

sea el tamaño de las muestras que se tomen en el lugar de desembarque. Sólo se obtendrá una estimación insesgada de la talla media de los pescados desembarcados.

(Véanse los **Ejercicios** en la Parte 2).

2.4 ANALISIS DE REGRESION LINEAL SIMPLE

Este método se utiliza para describir la variación de una cantidad, por ejemplo, la altura corporal de un pez, como función lineal de otra cantidad, por ejemplo, la talla. La teoría exige que la cantidad que aparece en el eje horizontal (la variable independiente), se mida con absoluta precisión. Sin embargo, el método se emplea a menudo sin cumplir este requisito. El efecto de la inexactitud de los valores de la variable independiente es que la pendiente de la línea se hace más plana (más cercana a cero).



Supóngase que se ha medido la talla total y la altura corporal de una muestra de 7 pescados.

La Tabla 2.4.1 muestra las tallas totales, x(i), y las alturas corporales, y(i), i= 1, 2, ..., 7.

Como es de suponer, la altura del cuerpo tiende a aumentar cuando la talla aumenta. Si las proporciones corporales de un pez permaneciesen constantes para todos los tamaños, la altura sería proporcional a la talla, y podría describirse por medio del modelo:

$$y(i) = b*x(i) \qquad (2.4.1)$$

donde b es una constante, también llamada "*parámetro*". El trazo en este modelo siempre pasa por el origen, el punto donde el eje horizontal x, y el eje vertical y, se encuentran. Se puede incluir una posible desviación de la proporcionalidad entre x e y introduciendo un segundo parámetro, a, y utilizar, en lugar de la Ec. 2.4.1, el siguiente modelo:

$$y(i) = a + b*x(i) \qquad (2.4.2)$$

TABLA 2.4.1
Ejemplo de mediciones de las tallas totales, x, y las respectivas alturas del cuerpo, y

i	1	2	3	4	5	6	7
x(i)	11.2	12.4	13.5	15.7	17.1	18.5	19.0
y(i)	3.0	3.2	4.0	4.8	4.8	4.9	5.6

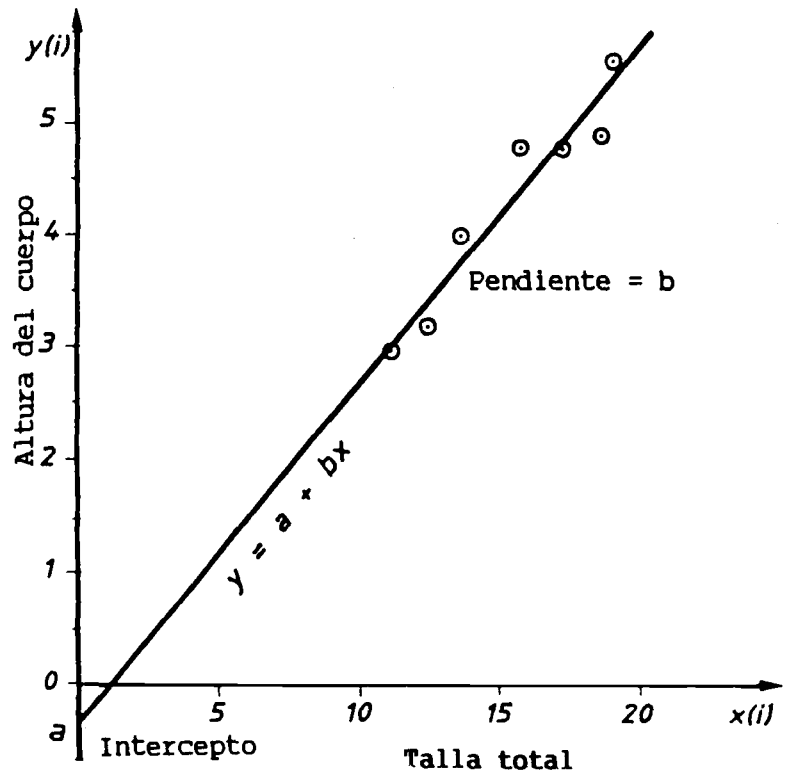


Fig. 2.4.1 Diagrama de dispersión de la altura del cuerpo (y) contra la talla total (x), también denominado "gráfico de y sobre x".

donde a indica el intercepto con el eje y de la línea que corresponde a los puntos. La Fig. 2.4.1. muestra el "*gráfico*" (o "*diagrama de dispersión*") de y(i) respecto de x(i).

La Ec. 2.4.2 implica que un pez de talla cero tiene altura "a", lo cual no tiene sentido excepto si "a" es igual a cero. Sin embargo, si se consideran sólo las tallas de un cierto rango (por ejemplo, las superiores a 5 cm), el modelo de dos parámetros puede dar un mejor ajuste a las observaciones que el modelo de un parámetro, porque el supuesto de proporcionalidad entre talla y altura no se cumple estrictamente.

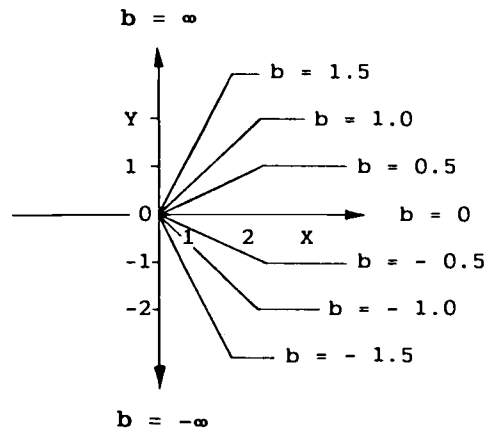
El modelo matemático de la Ec. 2.4.2 se llama "*modelo lineal*" porque los pares (x,y) que lo forman están en línea recta. Con a = -0.32 y b = 0.30 se obtiene la línea recta que se muestra en la Fig. 2.4.1. Con estos valores de a y b, la línea de la Fig. 2.4.1 representa un buen ajuste para los pares observados (x,y).

Pasemos ahora al problema de determinar la línea, es decir, de cómo estimar los parámetros a y b. Tal como se hizo para la media (vea la Sección 2.3), se mostrará también la forma de calcular los límites de confianza de las estimaciones a y b. Este procedimiento se denomina "*análisis de regresión lineal simple*". Es probablemente la técnica estadística que más se usa en biología pesquera. Los parámetros tienen nombres especiales: a se denomina "*intercepto*" y b "*pendiente*". El intercepto es la distancia desde el punto (0,0) en el gráfico (x,y) hasta el punto donde la "*línea de regresión*"

$$y = a + b*x$$

intercepta al eje y (véase la Fig. 2.4.1).

La pendiente "b" indica el grado de inclinación de la línea. Si b=0, la línea es paralela al eje x. Si b es positivo, la pendiente es ascendente. Si b es negativo, la pendiente es descendente.



La variable del eje horizontal, x, se denomina "*variable independiente*", y la del eje vertical, y, "*variable dependiente*". La línea de regresión se determina como la línea que reduce al mínimo la suma de los cuadrados de las desviaciones entre la línea y = a + b*x respecto a los pares de observaciones, (x(i), y(i)). Se dice que a y b se estiman por el "*método de los mínimos cuadrados*", es decir, se buscan aquellos valores de a y b que reduzcan al mínimo:

$$\sum_{i=1}^{n} [y(i) - a - b*x(i)]^2 \qquad (2.4.3)$$

donde n es el número de pares de observaciones (n = 7 en el ejemplo). Las desviaciones entre la línea y las observaciones se ilustran en la Fig. 2.4.2. El supuesto en el que se basa el análisis de regresión es que cada y(i) se distribuye normalmente con media a + b*x(i) y con varianza constante, es decir, una varianza que no depende del valor de x(i). La fórmula para la estimación de esta varianza común difiere sólo ligeramente de la que se presentó en la Sección 2.1. La llamada "*varianza respecto de la línea de regresión*" es:

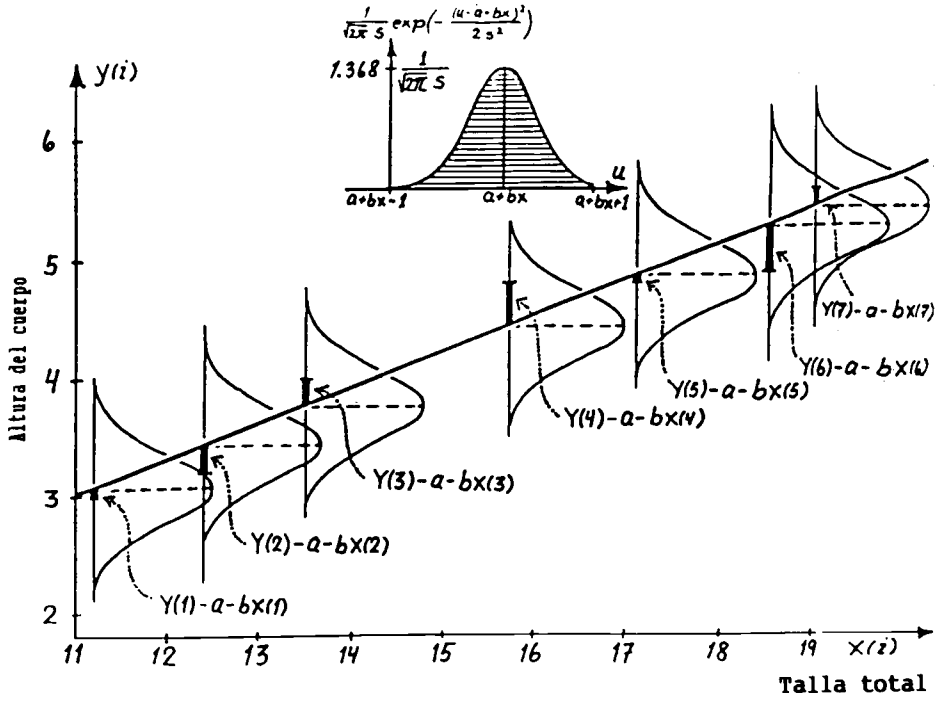


Fig. 2.4.2 Ilustración de los supuestos trás el analisis de regresión lineal simple. Cada y(i), para un x(i) dado, se distribuye normalmente con varianza común.

$$s^2 = \frac{1}{n-2} * \sum_{i=1}^{n} [y(i) - a - b*x(i)]^2 \qquad (2.4.4)$$

Hay n–2 grados de libertad (el número por el que se divide la suma) porque se tienen dos parámetros, *a y b.*

Las estimaciones de los parámetros a (intercepto) y b (pendiente) se obtienen de la siguiente manera:

$$b = \frac{\sum_{i=1}^{n} x(i)*y(i) - \frac{1}{n} * \sum_{i=1}^{n} x(i) * \sum_{i=1}^{n} y(i)}{\sum_{i=1}^{n} x(i)^2 - \frac{1}{n} * \left[\sum_{i=1}^{n} x(i)\right]^2} \qquad (2.4.5)$$

TABLA 2.4.2

Procedimiento de cálculo para el análisis de regresión lineal simple. Los resultados marcados con #) no se utilizan en el cálculo de a y b, pero se señalan aquí para su uso posterior

	talla total		altura del cuerpo		
i	x(i)	$x(i)^2$	y(i)	$y(i)^2$	x(i)*y(i)
1	11.2	125.44	3.0	9.00	33.60
2	12.4	153.76	3.2	10.24	39.68
3	13.5	182.25	4.0	16.00	54.00
4	15.7	246.49	4.8	23.04	75.36
5	17.1	292.41	4.8	23.04	82.08
6	18.5	342.25	4.9	24.01	90.65
7=n	19.0	361.00	5.6	31.36	106.40
Σ	107.4	1703.60	30.3	136.69	481.77
	Σx(i)	$\Sigma x(i)^2$	Σy(i)	$\Sigma y(i)^2$	Σx(i)*y(i)

$\bar{x} = 15.343$

$\frac{1}{n}*(\Sigma x(i))^2 = 1647.82$

$\Sigma x(i)^2 - \frac{1}{n}*(\Sigma x(i))^2 = 55.78$

$sx^2 = 9.296$ #)

$sx = 3.049$ #)

$\bar{y} = 4.329$

$\frac{1}{n}*(\Sigma y(i))^2 = 131.16$ #)

$\Sigma y(i)^2 - \frac{1}{n}*(\Sigma y(i))^2 = 5.534$ #)

$sy^2 = 0.922$ #)

$sy = 0.960$ #)

$\frac{1}{n}*\Sigma x(i)*\Sigma y(i) = 464.89$

$\Sigma x(i)*y(i) - \frac{1}{n}*\Sigma x(i)*\Sigma y(i) = 16.88$ sxy = 2.814 #)

$$b = \frac{\Sigma x(i)*y(i) - \frac{1}{n}*\Sigma x(i)*\Sigma y(i)}{\Sigma x(i)^2 - \frac{1}{n}*(\Sigma x(i))^2} = \frac{16.88}{55.78} = 0.303$$

$a = \bar{y} - \bar{x}*b = 4.329 - 15.343*0.303 = -0.315$

$$a = \bar{y} - \bar{x}*b \tag{2.4.6}$$

donde $\bar{y}$ y $\bar{x}$ son las medias de y y x, definidas por la Ec. 2.1.1.

En la Tabla 2.4.2 los procedimientos de cálculo para estimar a y b se demuestran utilizando los datos de la Tabla 2.4.1. Así, la línea de regresión estimada pasa a ser:

$$y = -0.315 + 0.303*x \tag{2.4.7}$$

Para calcular los límites de confianza de a y b se requiere la suma de los cuadrados de las desviaciones de x e y. Las varianzas de x e y están definidas por la Ec. 2.1.3:

$$sx^2 = \frac{1}{n-1} *[\Sigma x(i)^2 - \frac{1}{n} *\{\Sigma x(i)\}^2] \tag{2.4.8}$$

Para su uso en la siguiente sección se introduce una expresión similar para sy^2. Para ser utilizada en la siguiente sección se introduce aquí la "covarianza" que esta definida por la ecuación:

$$sxy = \frac{1}{n-1} *[\Sigma x(i)*y(i) - \frac{1}{n}*\Sigma x(i)*\Sigma y(i)] \tag{2.4.9}$$

El procedimiento para el cálculo de la varianza respecto de la línea de regresión que conduce a la Ec. 2.4.4 se demuestra en la Tabla 2.4.3. Sin embargo, la varianza respecto de la línea se puede obtener más fácilmente a partir de sy y sx:

$$s^2 = \frac{n-1}{n-2} * [sy^2 - b^2*sx^2] \tag{2.4.10}$$

Con los resultados de la Tabla 2.4.2, la Ec. 2.4.10 pasa a ser:

$$s^2 = \frac{6}{5} *(0.922 - 0.303^2*9.297) = 0.085$$

Las varianzas de las estimaciones de b y a son:

$$sb^2 = \frac{1}{n-2} *[(sy/sx)^2 - b^2] \tag{2.4.11}$$

y

$$sa^2 = sb^2*[\frac{n-1}{n}*sx^2 + \bar{x}^2] \tag{2.4.12}$$

Con los resultados de la Tabla 2.4.2 se tiene:

$$sb^2 = \frac{1}{7-2}*[\frac{0.922}{9.297} - 0.303^2] = 0.00147, \qquad sb = 0.038$$

$$sa^2 = 0.00147*(\frac{7-1}{7}*9.297 + 15.343^2) = 0.3578, \qquad sa = 0.598$$

TABLA 2.4.3
Cálculo de la varianza respecto de la línea, a partir de la Ec. 2.4.4

i	x(i)	y(i)	a+b*x(i)	$[y(i)-(a+b*x(i))]^2$
1	11.2	3.0	3.079	0.0062
2	12.4	3.2	3.442	0.0587
3	13.5	4.0	3.776	0.0504
4	15.7	4.8	4.442	0.1281
5	17.1	4.8	4.866	0.0044
6	18.5	4.9	5.291	0.1525
7	19.0	5.6	5.442	0.0250
$s^2 = 0.4252/(7-2) = 0.085$				suma: 0.4252

Los límites de confianza para el intercepto a y la pendiente b son:

$$a: [a - sa*t_{n-2},\ a + sa*t_{n-2}] \tag{2.4.13}$$

$$b: [b - sb*t_{n-2},\ b + sb*t_{n-2}] \tag{2.4.14}$$

Los límites de confianza del 95% de a y b para el ejemplo con n = 7 peces y $t_{(7-2)} = 2.57$ (Tabla 2.3.1) son:

$$a: [-0.315 - 0.598*2.57,\ -0.315 + 0.598*2.57] = [-1.85,\ 1.22]$$

$$b: [0.303 - 0.038*2.57,\ 0.303 + 0.38*2.57] = [\ 0.21,\ 0.40]$$

Obsérvese que el intervalo de confianza para el intercepto comprende el cero. Esto significa que la hipótesis de que la altura corporal es directamente proporcional a la talla (por lo tanto que "a = 0") no puede ser rechazada por los límites de confianza del 95%. Se dice entonces que "a" no es significativamente diferente de cero al nivel del 95%.

Si hay buenas razones para suponer que a = 0, entonces el valor estimado debería sustituirse por 0 si la estimación no es significativamente diferente de 0. Sin embargo, después hay que volver a calcular b como sigue:

$$b = \frac{\Sigma x(i)*y(i)}{\Sigma x(i)^2} \tag{2.4.15}$$

La estimación actual se basa sólo en 7 peces. Si se hubiesen medido 200, la estimación de la desviación estándar, sa, sería menor (véanse las Ecs. 2.4.11 y 2.4.12). Supóngase, por ejemplo, que $\bar{x}$, $\bar{y}$, sx, sy, a y b sean los mismos para una muestra de tamaño n = 200 que los estimados para una muestra de tamaño n = 7 (lo que es perfectamente posible). Aun cuando las estimaciones de a y b resulten tener el mismo valor, sus desviaciones estándar, sa y sb, serán diferentes.

Con n = 200, la Ec. 2.4.11 da sb = 0.006098, mientras que la Ec. 2.4.12 da sa = 0.0091 y $t_{198} = 1.97$ (Tabla 2.3.1). Así pues, sa y sb se vuelven más pequeñas y, en consecuencia, el intervalo de confianza de a es menor:

$$a: [-0.315 - 0.0091*1.97\ ,\ -0.315 + 0.0091*1.97] = [-0.33,-0.30]$$

La estimación de a será ahora significativamente diferent de 0. En este caso se puede concluir que las posibilidades de que el valor verdadero de a sea mayor que -0.30 o menor que -0.33 son inferiores a un 5%.

(Véanse los **Ejercicios** en la Parte 2).

2.5 EL COEFICIENTE DE CORRELACION Y LA REGRESION FUNCIONAL

El "*coeficiente de correlación*", r, es una medida de la asociación lineal entre dos cantidades, ambas sujetas a una variación aleatoria. La muestra de tallas estándar y alturas corporales de la Sección 2.4 es un ejemplo de tales cantidades. En ese caso se extrajeron siete pescados al azar. Por casualidad, estos podrían haber sido todos aproximadamente de la misma talla. En tal caso, la muestra no sería adecuada para estimar la relación talla/altura, ya que los límites de confianza a y b se volverían muy amplios.

El coeficiente de correlación sólo se puede usar cuando se permite que ambas medidas varíen aleatoriamente. Si se hubiese elegido siete pescados con tallas predeterminadas en vez de extraerlos al azar (por ejemplo, si se hubiesen seleccionado las tallas de 4, 6, 8, 10, 12, 14, y 16 cm para la muestra de talla/altura), el cálculo de un coeficiente de correlación para esta muestra sería incorrecto.

El coeficiente de correlación se define como:

$$r = \frac{sxy}{sx*sy} \tag{2.5.1}$$

donde sxy está definida por la Ec. 2.4.9 y sx y sy por la Ec. 2.4.8. Si se introduce la pendiente ($b = sxy/sx^2$), la Ec. 2.5.1 se transforma en:

$$r = b*(sx/sy) \tag{2.5.2}$$

El recorrido de r es: $-1.0 \leq r \leq 1.0$. Por lo tanto, r es negativo si y tiende a disminuir cuando x aumenta, y r es positivo si y tiende a aumentar cuando x aumenta. Esta afirmación también vale para la pendiente b y se desprende de la Ec. 2.5.2: como sx/sy es siempre positiva (véase la definición de la Ec. 2.4.8), r tiene el mismo signo que la pendiente b. Los casos extremos, r=1 ó r=-1, ocurren cuando todos los pares (x,y) están situados exactamente en una línea recta. Cuánto más se aproxime r a cero, tanto menos pronunciada será la asociación lineal entre y y x. Cuando r = 0, x e y son independientes una de otra.

La Fig. 2.5.1 muestra cuatro ejemplos de diagramas de dispersión con diferentes valores de r. Para el ejemplo de la Tabla 2.4.2 se tiene:

$$r = \frac{2.814}{3.049*0.960} = 0.961$$

Sean r1 (inferior) y r2 (superior) los límites de confianza del 95% de r. Estos pueden calcularse a partir de las expresiones:

$$r1 = \tanh\left[0.5*\ln\left(\frac{1+r}{1-r}\right) - 1.96/\sqrt{n-3}\ \right]$$

$$r2 = \tanh\left[0.5*\ln\left(\frac{1+r}{1-r}\right) + 1.96/\sqrt{n-3}\ \right] \tag{2.5.3}$$

"tanh" es la "tangente hiperbólica", que está incorporada en muchas calculadoras científicas de bolsillo.

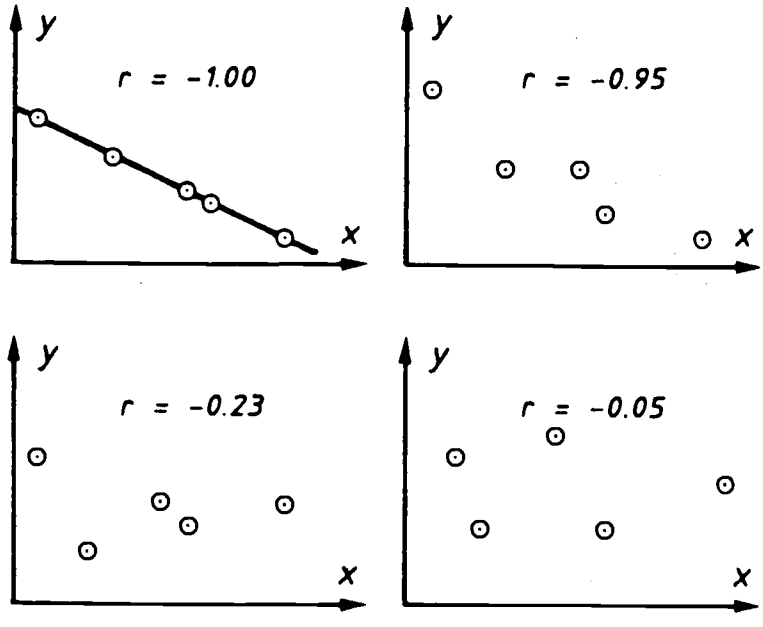


Fig. 2.5.1 Ejemplos de coeficientes de correlación.

Tomando r del ejemplo (r = 0.961, n = 7), los límites del 95% son: [r1 , r2] = [0.75 , 0.99]. Los límites de confianza del 99% se pueden obtener sustituyendo el número 1.96 por 2.58 en la Ec. 2.5.3.

A menudo interesa saber si el cero está incluido en el intervalo de confianza, es decir, qué posibilidades hay de que la asociación lineal se deba al azar. En este ejemplo las posibilidades son inferiores al 5%, ya que el cero no está en el intervalo de confianza.

En el ejemplo de la regresión de la altura corporal sobre la talla total, se eligió la talla como variable independiente y la altura del cuerpo como variable dependiente. Sin embargo, no hay una razón especial para esta elección. La muestra consiste en siete pescados escogidos al azar. No se controló cuáles serían sus longitudes y alturas corporales, por lo que podría haberse elegido igualmente la opción contraria para las variables dependiente e independiente.

Uno de los supuestos en que descansa el análisis de regresión lineal es que la variable independiente no puede ser una variable aleatoria. Tiene que ser algo cuyos valores se puedan determinar de antemano. Por ejemplo, si la variable independiente es el momento en que se toma la muestra, puede ser determinada de antemano. Se podría decidir tomar una muestra el primer día de cada mes. Si el tiempo se mide en unidades de años y se considera como fecha cero el primero de enero, la variable independiente tomará los valores: 0, 1/12, 2/12, 3/12 ... etc. Esta no es, evidentemente; una variable aleatoria.

Volviendo a los siete peces del ejemplo anterior, como se escogieron al azar de una distribución normal de tallas, se les puede aplicar un análisis de correlación. Por otro lado, las tallas se pueden decidir previamente. Se podrían escoger los cuatro peces más pequeños y los tres más grandes, o, como se hizo efectivamente, tomarlos como vengan. Sólo en este último caso se pueden realizar ambos tipos de análisis. En el primer caso, sólo sería posible el análisis de regresión. Por otra parte, probablemente sería una forma más efectiva de hacer el análisis de regresión, dada la gran distancia entre las observaciones en el eje horizontal. Esto causaría una pequeña varianza de la pendiente. Al elegir los peces al azar, la mayoría de ellos serán probablemente medianos y contribuirán muy poco a la determinación de la pendiente, que podría mostrar una varianza grande.

Otra cuestión es si se hubiera obtenido un resultado diferente utilizando la altura del cuerpo como variable independiente, representando entonces la talla como función de la altura corporal. En primer lugar, hay que ver si la altura se puede medir con la misma precisión que la talla. Si no es así, la pendiente estará sesgada (aplanada), como ya se ha mencionado. Pero de todas formas habrá problemas, incluso si las dos variables se miden con la misma exactitud.

Tomando la altura del cuerpo como variable independiente se obtiene lo que se llama una "*regresión inversa*". Sólo en el caso excepcional de que todas las observaciones estuvieran en la línea de regresión (es decir, si r = 1 ó r = -1) se obtendría con la regresión inversa el mismo resultado que con la regresión simple. La ecuación y = a + b*x (Ec. 2.4.2) es matemáticamente equivalente a:

$$x = -a/b + y/b$$

ó

$$x = A + B*y \text{ donde } A = -a/b \text{ y } B = 1/b \qquad (2.5.4)$$

Llevando a cabo la regresión inversa (Ec. 2.5.4) encontramos que A = 2.139 y B = 3.05.

La ecuación: x = 2.139 + 3.050*y se puede convertir en:

$$y = -0.701 + 0.328*x$$

que puede compararse con el resultado obtenido para la regresión original (Ec. 2.4.7: y = - 0.315 + 0.303*x). Como se ve, el resultado de la regresión inversa difiere del análisis de regresión original.

Una forma de soslayar el problema de elegir la variable independiente cuando ambas variables son aleatorias, es utilizar el llamado "análisis de regresión funcional" (véase Ricker, 1973). Este

método estima una pendiente (que se denominará b' para distinguirla de la pendiente b de la regresión simple) por medio de las expresiones:

$$b' = sy/sx \quad \text{si } r > 0$$
$$b' = -sx/sy \quad \text{si } r < 0 \qquad (2.5.5)$$

y el intercepto:

$$a' = \bar{y} - b'*\bar{x} \qquad (2.5.6)$$

Este tipo de análisis da un resultado que puede considerarse un compromiso entre la regresión simple y su contraparte inversa.

Con los resultados de la Tabla 2.4.2 se tiene:

$$b' = 0.960/3.049 = 0.315 \text{ y } a' = 4.329 - 0.315*15.343 = -0.504$$

e

$$y = -0.504 + 0.315*x$$

El análisis de regresión funcional se ha mencionado aquí para que la exposición resulte más completa. Sin embargo, su aplicabilidad tiene ciertas limitaciones bastante complicadas, sobre las que no podemos detenernos ahora.

Hasta ahora hemos estimado las tres regresiones siguientes:

1. Análisis de regresión simple original: y = -0.315 + 0.303*x
2. Análisis de regresión funcional : y = -0.504 + 0.315*x
3. Análisis de regresión simple inversa : y = -0.701 + 0.328*x

La Fig. 2.5.2 muestra las tres líneas de regresión. Obsérvese que todas atraviesan el punto ($\bar{x},\bar{y}$), y que un aumento de la pendiente está compensado parcialmente por una disminución del intercepto.

(Véanse los **Ejercicios** en la Parte 2).

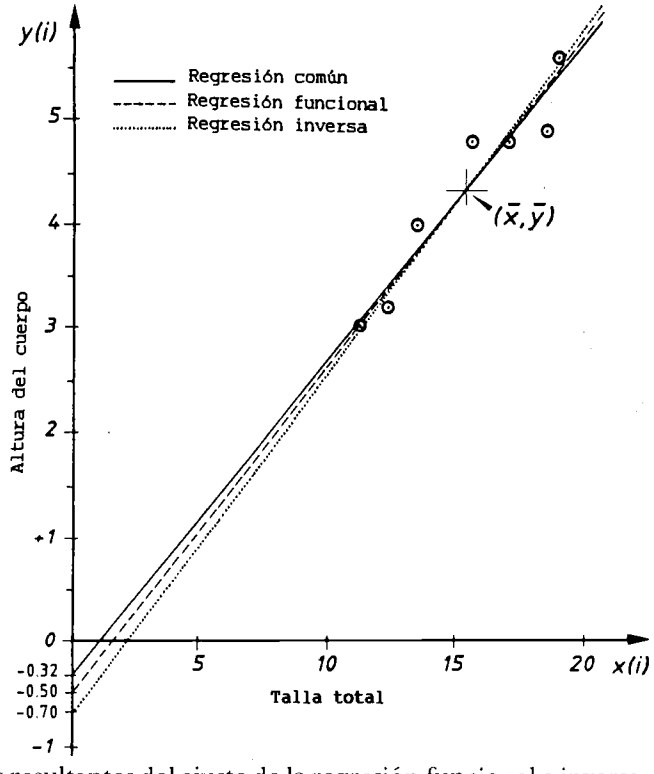


Fig. 2.5.2 Líneas resultantes del ajuste de la regresión funcional e inversa, comparada con la línea de regresión común.

2.6 TRANSFORMACIONES DE LINEALIZACION

Las funciones lineales son matemáticamente fáciles y además tienen la ventaja de que pueden interpretarse gráficamente sin ningún problema. Sin embargo, muchas de las relaciones funcionales que se observan en biología pesquera no son lineales. Afortunadamente, estas funciones no lineales pueden transformarse a menudo en funciones lineales, lo que significa que pueden tratarse de la forma descrita en las secciones precedentes. A continuación se ofrecen varios ejemplos de la aplicación de las transformaciones de funciones no lineales en funciones lineales en biología pesquera.

Ejemplo 1: Relación talla-peso

Este es un ejemplo famoso: la relación funcional entre la talla y el peso corporal del pez. La Fig. 2.6.1 muestra un gráfico del peso respecto de la talla de la baga (*Nemipterus marginatus*). Claramente, ésta no es una relación lineal. La curva de la Fig. 2.6.1 es de la función:

$$W(i) = q*L(i)^b \qquad (2.6.1)$$

donde W(i) es el peso corporal del iésimo pez, L(i) es la talla, y q y b son los parámetros. La Ec. 2.6.1 se denomina generalmente "*relación talla-peso*" y puede transformarse en una ecuación lineal tomando logaritmos a ambos lados de la ecuación:

$$\ln W(i) = \ln q + b*\ln L(i) \qquad (2.6.2)$$

ó

$$y(i) = a + b*x(i) \qquad (2.6.2a)$$

donde $y(i) = \ln W(i)$, $x(i) = \ln L(i)$ y $a = \ln q$.

Con la Ec. 2.6.2a se puede ahora hacer la estimación de a y b por análisis de regresión lineal. Los datos de entrada aparecen en la Tabla 2.6.1 y el correspondiente diagrama de dispersión en la Fig. 2.6.2. Los resultados son:

$a = -4.538$, $b = 3.057$, $sx = 0.3311$, $sy = 1.0161$, $n = 16$,

$\bar{x} = 2.727$ e $\bar{y} = 3.799$

Ya que $a = \ln q$, podemos obtener q de la relación talla-peso original (Ec. 2.6.1), al obtener el antilogaritmo de a:

$$q = \exp a = \exp(-4.538) = 0.0107$$

De esta forma, la relación estimada entre W (en gramos) y L (en centímetros) resulta:

$$W = 0.0107*L^{3.057}$$

(La transformación de los logaritmos a valores normales introduce un sesgo que no es examinado aquí).

Ahora se pueden calcular los límites de confianza al 95% para b, usando los valores de sx, sy, n y t_{14} (véase la Tabla 2.3.1) en la Ec. 2.4.11:

$$sb^2 = \frac{1}{16-2} * \left[\left\{\frac{1.0161}{0.3311}\right\}^2 - 3.057^2\right] = 0.0052$$

$$sb = 0.072 \quad y \quad sb*t_{n-2} = 0.072*2.15 = 0.155$$

El intervalo de confianza de b al 95% es [(3.057 – 0.155), (3.057 + 0.155)] o sea [2.90, 3.21]. Estos límites de confianza muestran que sólo el primer decimal de la estimación es significativo (véase la Sección 2.3), por lo que el valor verdadero de b podría perfectamente ser 3.0.

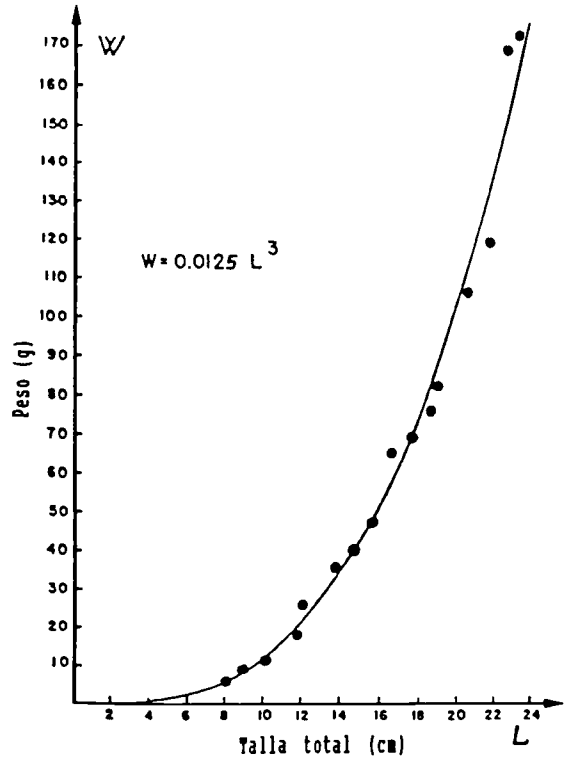


Fig. 2.6.1 Relación talla-peso de *Nemipterus marginatus* en el Mar Meridional de China. (Basada en datos de la Tabla 2.6.1).

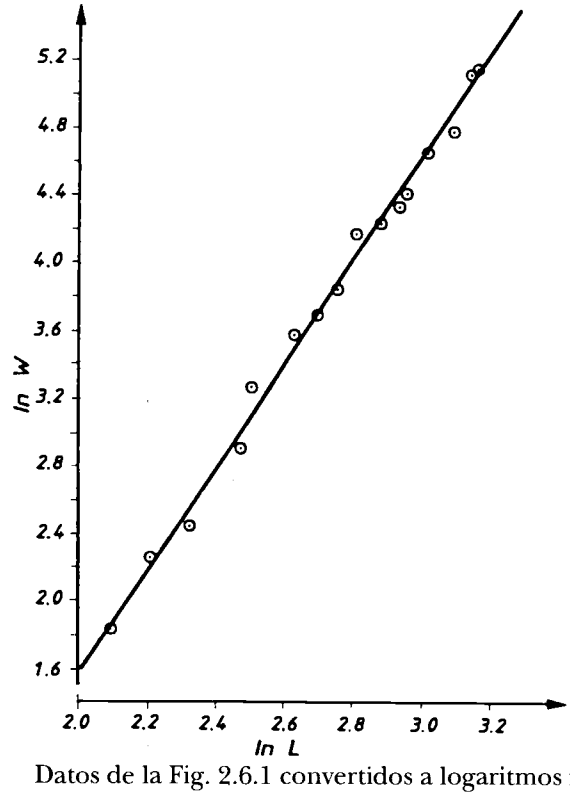


Fig. 2.6.2 Datos de la Fig. 2.6.1 convertidos a logaritmos naturales.

TABLA 2.6.1
Datos para la estimación de la relación talla-peso de la "baga" (*Nemipterus marginatus*) del Mar de China Meridional (Pauly, 1983)

i	L(i)	W(i)	ln L(i) x(i)	ln W(i) y(i)
1	8.1	6.3	2.092	1.841
2	9.1	9.6	2.208	2.262
3	10.2	11.6	2.322	2.451
4	11.9	18.5	2.477	2.918
5	12.2	26.2	2.501	3.266
6	13.8	36.1	2.625	3.586
7	14.8	40.1	2.695	3.691
8	15.7	47.3	2.754	3.857
9	16.6	65.6	2.809	4.184
10	17.7	69.4	2.874	4.240
11	18.7	76.4	2.929	4.336
12	19.0	82.5	2.944	4.413
13	20.6	106.6	3.025	4.669
14	21.9	119.8	3.086	4.786
15	22.9	169.2	3.131	5.131
16	23.5	173.3	3.157	5.155
		suma	43.629	60.786
		media	2.7268	3.7991
		sx	0.3311	
		sy		1.0161

Como el peso de un pez (en gramos) es aproximadamente igual a su volumen (en centímetros cúbicos) y su volumen suele ser proporcional a su largo al cubo, L^3, es posible esperar que el valor de b en las Ecs. 2.6.1 y 2.6.2 se acerque a 3.0.

Ya que el intervalo de confianza calculado anteriormente apoya esta hipótesis, se puede simplificar la relación talla-peso, reemplazando la estimación b = 3.057 por b = 3.0. Esto implica que se debe obtener una nueva estimación para el intercepto a. Como la nueva línea recta con b = 3.0 también pasa por el punto $(\bar{x},\bar{y})$, se puede calcular el nuevo intercepto a usando la Ec. 2.6.2a:

$$a = \bar{y} - b*\bar{x} = 3.799 - 3.0*2.727 = -4.382$$

Con el valor de a se obtiene el nuevo valor correspondiente a q:

$$q = \exp(-4.382) = 0.0125$$

Así, la nueva relación que se obtiene queda definida por la ecuación:

$$W = 0.0125*L^3$$

Ejemplo 2: Linealización de una distribución normal

En la Sección 2.2 (Ec. 2.2.1), la expresión matemática de una distribución normal está dada como:

$$Fc(x) = \frac{n*dL}{s*\sqrt{2\pi}} * \exp[-(x-\bar{x})^2/(2s^2)]$$

Esta ecuación puede transformarse en una regresión lineal con los dos pasos siguientes:

Paso 1: *Conversión de la distribución normal en una parábola*

Tomando los logaritmos a ambos lados de la Ec. 2.2.1 se obtiene:

$$\ln Fc(x) = \ln\left[\frac{n*dL}{s*\sqrt{2\pi}}\right] - (x-x^2)/(2s^2) \qquad (2.6.3)$$

Considerando ln Fc(x) como la variable dependiente, y, y x como la variable independiente, se obtiene una relación funcional entre y y x, que puede representarse gráficamente por una parábola con la fórmula general:

$$y = a + bx + cx^2$$

Introduciendo los valores utilizados en el ejemplo de la Tabla 2.1.2 se tiene:

$$y = \ln[27*1/2.2\sqrt{2\pi}\,] - (x-15.07)^2/(2*2.2^2) = 1.59 - (x-15.07)^2/9.68$$

cuyo gráfico se muestra en la Fig. 2.6.3.

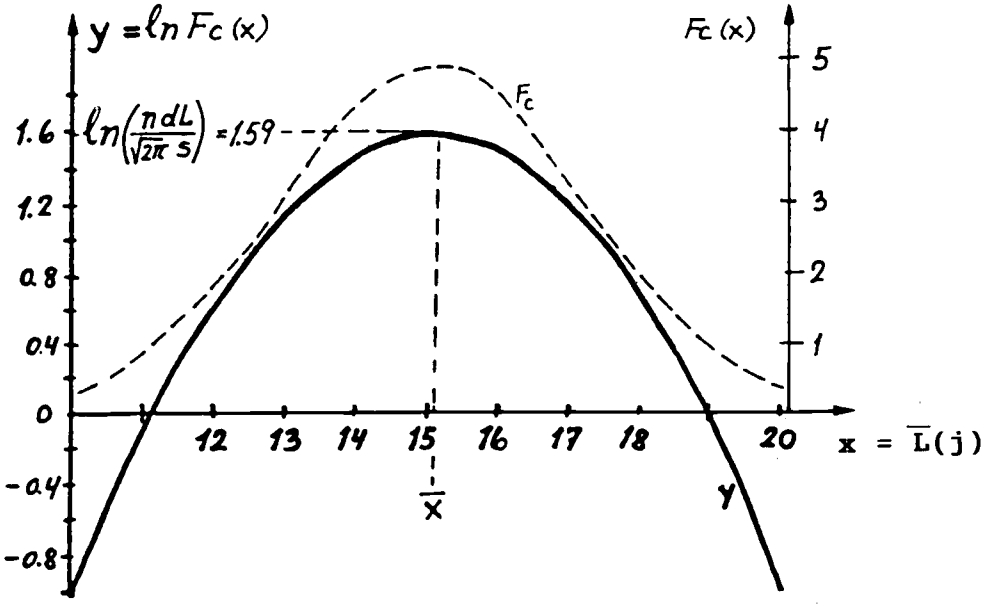


Fig. 2.6.3 Distribución normal transformada al aplicar logaritmos naturales (y) junto con la distribución original (Fc).

Paso 2: *Conversión de la parábola en una línea recta*

Cuando se calculan y grafican las diferencias entre puntos establecidos sobre una parábola, utilizando como referencia valores regularmente espaciados sobre el eje x, siempre se logra una línea recta. Para dos valores de x consecutivos, se le resta a la función de ln Fc(x) del valor de x mayor la correspondiente al valor menor de x. Se procede de esta manera en forma consecutivamente con los restantes valores (x,x-1). Así, se obtiene una serie de diferencias que son positivas en el lado izquierdo, cero para el punto máximo de la parábola y negativos en el lado derecho. El resultado de estos cálculos se ilustra en las Fig. 2.6.4.

Para explicar matemáticamente este proceso se introduce una nueva variable independiente, y', que representa la diferencia determinada entre el logaritmo de la frecuencia a una cierta clase de longitud (x) y el logaritmo de la frecuencia correspondiente a la clase anterior.

$$y' = \ln Fc(x + dL) - \ln Fc(x) \tag{2.6.4}$$

Esto también se puede expresar como:

$$y' = \Delta \ln Fc(x + dL/2)$$

donde Δ (delta) designa una "pequeña" diferencia entre dos valores de la función. De allí que y' se debe graficar respecto a una nueva variable independiente, z, que equivale a la talla x más la mitad del intervalo de talla utilizado:

$$z = x + dL/2$$

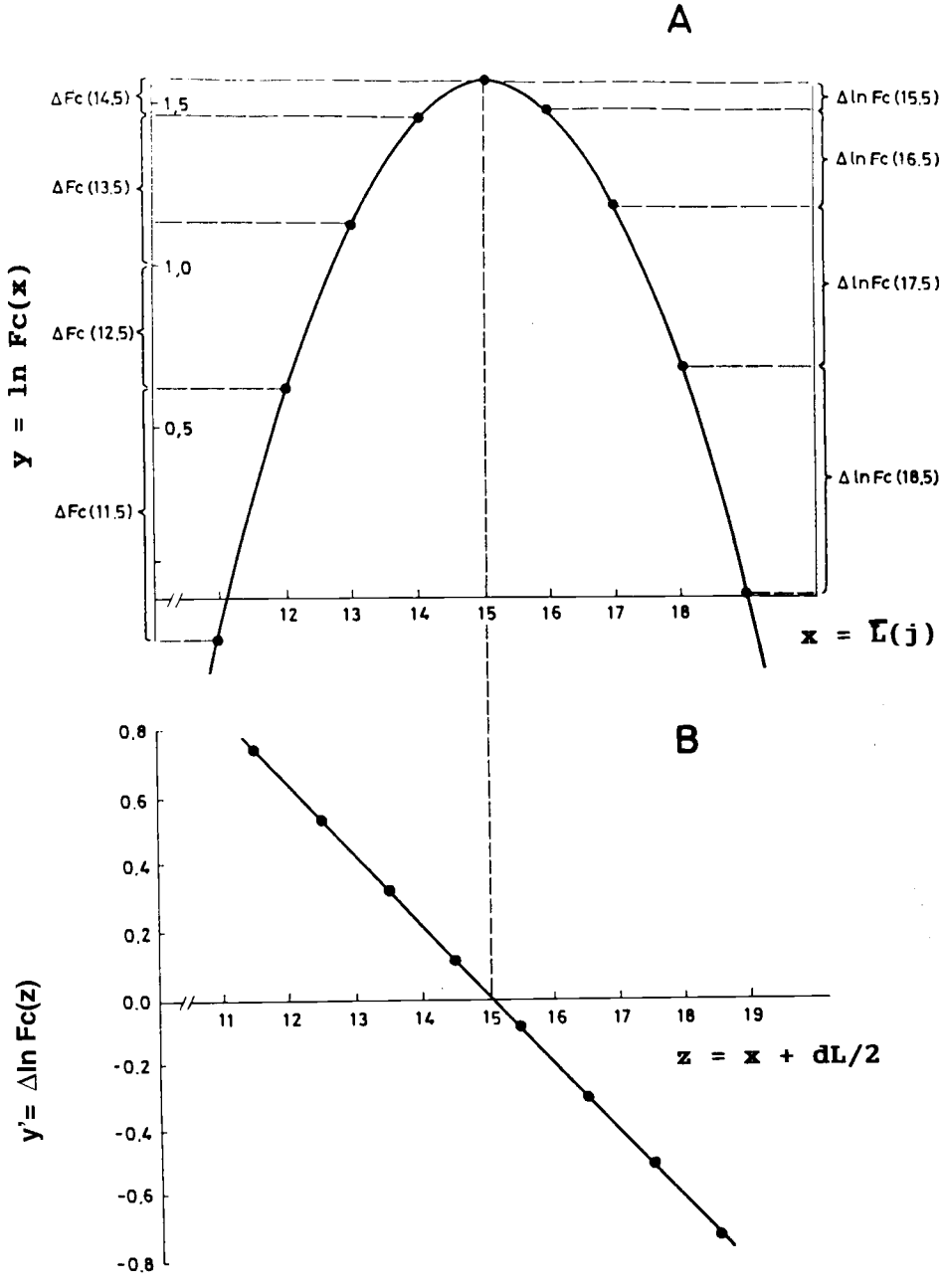


Fig. 2.6.4a Estimación de la media y la varianza mediante el método de Bhattacharya. A: Parábola y diferencias entre puntos equidistantes sobre el eje x. B: Gráfico de Bhattacharya con las diferencias entre marcas de clases consecutivas. Datos de la Tabla 2.6.2.

Ahora hay que insertar la Ec. 2.6.3 en la Ec. 2.6.4 como sigue:

$$y' = \Delta \ln Fc(x+dL/2) = \Delta \ln Fc(z) =$$

$$= \left[\ln\left(\frac{n*dL}{s*\sqrt{2\pi}}\right) - \frac{(x+dL-\bar{x})^2}{2s^2}\right] - \left[\ln\left(\frac{n*dL}{s*\sqrt{2\pi}}\right) - \frac{(x-\bar{x})^2}{2s^2}\right] =$$

$$= \left[\frac{-(x+dL-\bar{x})^2+(x-\bar{x})^2}{2s^2}\right]$$

Una vez resueltos los cuadrados y las sumas, esta ecuación se convierte en una relativamente más sencilla:

$$y' = \frac{dL*\bar{x}}{s^2} - \frac{dL}{s^2} * (x+dL/2) \qquad (2.6.5)$$

o

$$y' = a + b*z \quad \text{con} \quad a = dL*\bar{x}/s^2,\ b = -dL/s^2 \quad y \quad z = x+dL/2$$

A partir de la pendiente, b, y del intercepto, a, se obtiene la varianza:

$$s^2 = -dL/b \qquad (2.6.6)$$

y la media

$$\bar{x} = -a/b \qquad (2.6.7)$$

Esa regresión es uno de los elementos principales del método descrito por Bhattacharya (1967) para separar dos o más distribuciones normales (Sección 3.4.1). Se le llama el *"gráfico de Bhattacharya"*. La Tabla 2.6.2 y la Fig. 2.6.4a muestran un ejemplo sobre este particular. En este caso, los valores teóricos de la Tabla 2.2.1 se han utilizado como "observaciones" y corresponden a frecuencias de tallas, Fc, y a las marcas de clase, x. Estas se ajustan exactamente al modelo. En este caso, la media y varianza estimada a través del gráfico de Bhattacharya deberían ser iguales a las obtenidas por el método tradicional (Tabla 2.1.2). No obstante, pequeñas diferencias son posibles debido al ajuste de los valores a través de la regresión lineal. En la parte B de la Fig. 2.6.4a se encuentran graficadas las diferencias entre los logaritmos de dos frecuencias consecutivas respecto a la marca de clase de los valores de x.

Así también, el grafico de Bhattacharya entrega una respuesta acerca del número de observaciones que contiene una distribución normal para la cual sólo se dispone de información para ciertas clases de tallas. Reescribiendo la Ec. 2.2.1 se tiene que

$$F(\bar{L}(j)) = n * \frac{dL}{s*\sqrt{2\pi}} * \exp\left[\frac{-[\bar{L}(j)-\bar{x}]^2}{2s^2}\right] \qquad (2.6.8)$$

De esta manera, n se puede estimar inclusive para una simple clase de tallas j, una vez que se han estimado $\bar{x}$ y s^2. Sin embargo, problemas en las muestras causan inexactitud debido a que influencian el número de peces en cada intervalo de longitud, como se puede ver en la Fig. 2.2.1. Cuando se conoce el número de individuos en varias clases de longitud, las frecuencias se pueden sumar para suavizar las desviaciones de cada una de ellas respecto a las correspondientes frecuencias esperadas. Al sumar los valores calculados para las clases i en ambos lados del signo igual y rearreglando los términos se llega a la siguiente expresión:

$$n = \frac{\sum_{j=1}^{i} F[\bar{L}(j)]}{\frac{dL}{s*\sqrt{2\pi}} * \sum_{j=1}^{i} \exp\left[-\frac{[\bar{L}(j)-\bar{x}]^2}{2s^2}\right]} \qquad (2.6.9)$$

TABLA 2.6.2

Estimación del valor medio y la varianza de una distribución normal determinado a través del gráfico de Bhattacharya, utilizando las frecuencias teóricas, Fc(x), de la Tabla 2.1.2, presentados en la Tabla 2.2.1. Los valores se encuentran ilustrados en las respectivas Figs. 2.6.3 y 2.6.4

indice j	$\bar{L}$ (j) (x)	intervalo x– dL/2,x+dL/2	Fc(x)	ln Fc(x) (y)	Δln Fc(z) (y')	x+dL/2 (z)
1	11	10.5-11.5	0.88	-0.128		
					0.743	11.5
2	12	11.5-12.5	1.85	0.615		
					0.529	12.5
3	13	12.5-13.5	3.14	1.144		
					0.326	13.5
4	14	13.5-14.5	4.35	1.470		
					0.117	14.5
5	15	14.5-15.5	4.89	1.587		
					-0.088	15.5
6	16	15.5-16.5	4.48	1.500		
					-0.297	16.5
7	17	16.5-17.5	3.33	1.203		
					-0.500	17.5
8	18	17.5-18.5	2.02	0.703		
					-0.713	18.5
9	19	18.5-19.5	0.99	0.010		
					a = 3.1237 b = -0.2073	(dL = 1)

$\bar{x} = -a/b = 15.07$

$s^2 = -dL/b = 4.82$

$s = 2.20$

TABLA 2.6.2a

Estimación del número total de observaciones mediante el método de Bhattacharya

j	$\bar{L}(j)$	$F[\bar{L}(j)]$	$\exp\left[-\frac{[\bar{L}(j)-\bar{x}]^2}{2s^2}\right]$
1	11	0.88	0.1802
2	12	1.85	0.3778
3	13	3.14	0.6433
4	14	4.35	0.8898
5	15	4.89	0.9996
suma	-	15.11	3.0907

$$n = \frac{15.11}{\frac{1}{2.193 * \sqrt{2\pi}} * 3.0907} = 26.88$$

TABLA 2.6.3
Gráfico de Bhattacharya con las distribuciones de frecuencias de tallas de las muestras presentadas en la Tabla 2.1.2

indice	x (x)	x–dL/2,x+dL/2	F(x)	ln F(x) (y)	Δln F(z) (y')	x+dL/2 (z)
1-2	11.5	10.5-12.5	4	1.386		
					0.560	12.5
3-4	13.5	12.5-14.5	7	1.946		
					0.251	14.5
5-6	15.5	14.5-16.5	9	2.197		
					-0.588	16.5
7-8	17.5	16.5-18.5	5	1.609		
					-0.916	18.5
9	19.5	18.5-20.5	2	0.693		
					a = 3.909	(dL = 2)
					b = -0.263	

$\bar{x}$ = -a/b = 14.8
s^2 = -dL/b = 7.605
s = 2.76

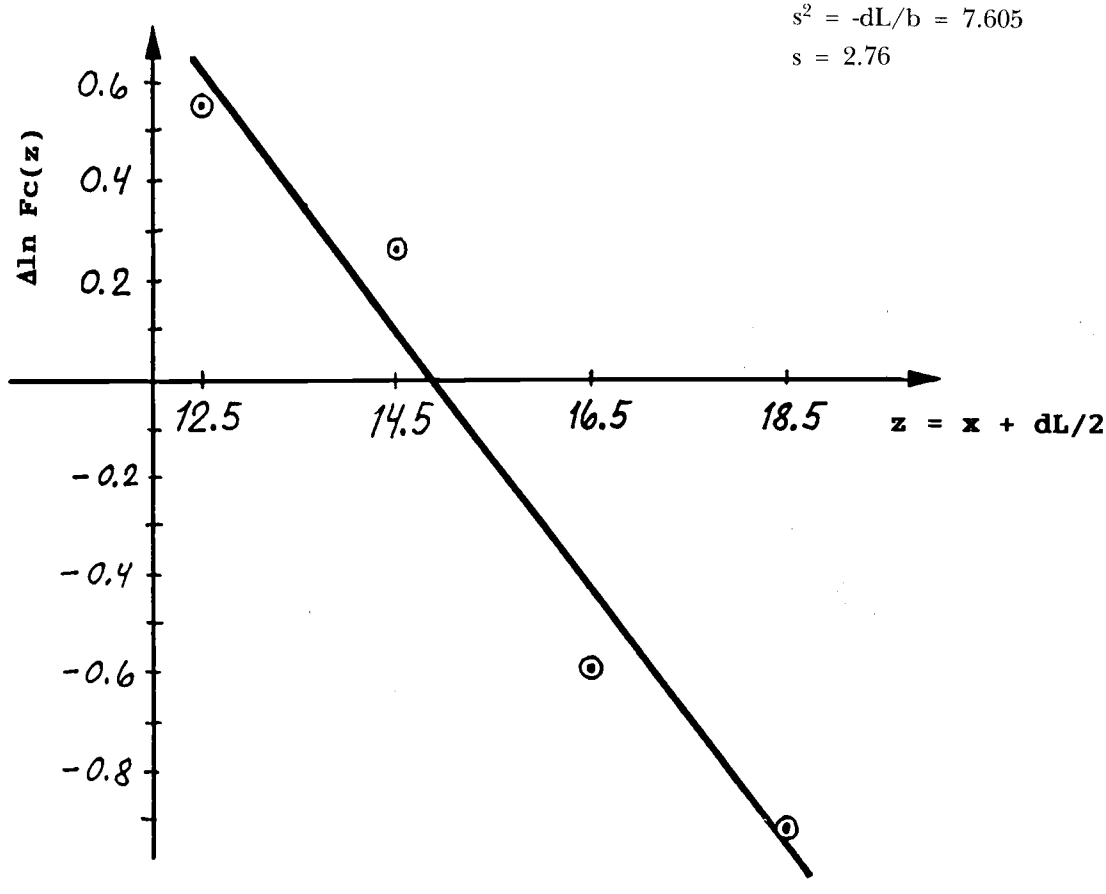


Fig. 2.6.5 Gráfico de Bhattacharya correspondiente a la Tabla 2.6.3.

Las observaciones para los peces con longitudes mayores de $\bar{x}$ en la Fig. 2.6.4a puede que no sean del todo reales, dado que en esas tallas existe sobreposición con los peces más pequeños de la edad superior. Este es el caso que se muestra en la Fig. 1.4.1, por ejemplo entre las edades 1 y 2. De allí que se debería utilizar únicamente el lado izquierdo de las observaciones de la Fig. 2.6.4a (x = 11, 12, 13, 14, 15 cm) para efectuar el gráfico de Bhattacharya, ya que éstos permiten utilizar cuatro puntos para la línea recta que conduce a la estimación de $\bar{x}$ y s^2. Así, con los datos de la Tabla 2.6.2 se obtiene que

$$a = 3.134; \quad b = -0.2081; \quad \bar{x} = 15.06; \quad s^2 = 4.805; \quad s = 2.193$$

El resultado es prácticamente el mismo que el obtenido para la distribución normal completa, debido a que la línea recta que ajusta los valores es casi perfecta (véase la Fig. 2.6.2). Una aplicación práctica de la Ec. 2.6.9 se incluye en la Tabla 2.6.2a. En esta se obtiene como resultado que n = 26.88 cuando el verdadero valor (determinado en la Tabla 2.1.2) es 27.

Una vez que se ha determinado n, se pueden estimar las frecuencias de cada clase de tallas mediante la Ec. 2.6.8. Estos cálculos no se encuentran en la Tabla 2.6.2a, debido a que en este ejercicio las "observaciones" son ahora las frecuencias teóricas.

En la Tabla 2.6.3 se puede ver la estimación de la media y la varianza establecida del gráfico de Bhattacharya, pero ahora utilizando los valores incluidos en la Tabla 2.1.2. En este caso, debido al pequeño tamaño de la muestra, las frecuencias fueron agrupadas en intervalos de dos centímetros. La graficación de estos valores se puede ver en la Fig. 2.6.5. Ahora, tanto la media como la varianza que se indican en la Tabla 2.6.3 son diferentes a los calculados a través del método tradicional (Tabla 2.1.2), debido a lo pequeño de la muestra, al error causado por los intervalos de tallas grandes y debido a diferencias relativas al método estadístico utilizado (análisis de regresión lineal).

(Véanse los **Ejercicios** en la Parte 2).