2 BIOESTATÍSTICA

Este Capítulo trata de uma breve descrição de alguns métodos estatísticos geralmente usados em biologia pesqueira tropical e introduz a notação estatística adoptada neste manual. Fazendo uma rápida revisão do assunto, serve como ponto de referência, não devendo, assim, ser tomado como um manual de estatística propriamente dito.

Existe muita bibliografia, sobre métodos estatísticos, disponível para quem quizer aprofundar o tema da biostatística. Entre muitas outras obras relevantes sobre a matéria, indica-se, a título de referência, os livros "Biometry" de Sokal e Rohlf (1981) que trata a teoria estatística de forma acessível e "Sampling techniques" de Cochran (1977) que, embora pareça um pouco mais complicado, pode também ser recomendado como um texto básico. No entanto, existem muitos outros manuais que podem ser igualmente úteis.

2.1 VALOR MÉDIO E VARIÂNCIA

Vamos considerar uma amostra de n peixes de uma mesma espécie capturados em um lance de arrasto. Seja x(i) o comprimento do peixe nº i, onde i = 1,2,..,n. O "*comprimento médio*" (em geral o "*valor médio*") da amostra é definido por:

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

As duas primeiras colunas da Tabela 2.1.1 mostram um exemplo para n = 27.

A variância, que é uma medida de variabilidade sobre o valor médio, é definida por:

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

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

Assim, a variância, s^2, é a soma dos quadrados dos desvios em relação ao valor médio dividido pelo número n menos um. A terceira e a quarta coluna da Tabela 2.1.1 ilustram o cálculo da variância. Note que, se todos os peixes da amostra tivessem o mesmo comprimento, o valor seria igual ao comprimento médio e a variância seria zero. A soma dos desvios (não elevados ao quadrado) será sempre zero. Por outro lado, quanto mais elevados fôrem os desvios em relação ao valor médio maior será a variância. Os dois maiores valores dos quadrados dos desvios sobre a média ocorrem para a maior e menor observação conforme pode ser visto na Tabela 2.1.1.

A raiz quadrada da variância, s, é chamada "*desvio padrão*". Usualmente estamos interessados na variância relativa do valor médio e por isso s é a quantidade relevante pois possui a mesma unidade da média. Isto conduz à definição do desvio padrão relativo $s/\bar{x}$, também chamado "*coeficiente de variação*".

Para facilitar os cálculos manuais a Eq. 2.1.2 pode ser rearranjada em uma equivalente como se segue:

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

No entanto, como a maioria das calculadoras de bolso, contém uma opção para o cálculo automático da média e variância, ilustraremos os cálculos com a Eq. 2.1.2 que, conceitualmente, é de mais fácil entendimento.

Por várias razões, como por exemplo a representação gráfica, é conveniente ordenar a amostra na forma de uma "*tabela de frequências*", dividindo a gama de comprimentos da amostra por um certo número de intervalos de classe de comprimento. No caso da Tabela 2.1.1 a gama de comprimentos varia de 11.2 cm a 19.0 cm. Se tomarmos classes de 1 cm necessitaremos nove classes de comprimento para cobrir toda a gama observada. Usando 10.5 como limite inferior do primeiro intervalo de classe, os intervalos e frequências dos comprimentos estão representados nas primeiras quatro colunas da Tabela 2.1.2, que é chamada tabela de frequências de comprimento.

Tabela 2.1.1 Valor médio, variância e desvio padrão de uma amostra de distribuição de frequências

nº do peixe	compri-mento (cm)	desvio da média	quadrado dos desvios da média
i	x(i)	$x(i)-\bar{x}$	$(x(i)-\bar{x})^2$
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 (min.)	-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 (max.)	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	0.00	121.48
	$= \Sigma\, x(i)$	$= \Sigma(x(i)-\bar{x})$	$= \Sigma(x(i)-\bar{x})^2$

comprimento médio, $\bar{x}$	:	406.8/27 = 15.07
variância, s^2	:	121.48/(27-1) = 4.67
desvio padrão, s	:	$\sqrt{4.67}$ = 2.16
desvio padrão relativo, $s/\bar{x}$	:	2.16/15.07 = 0.14
erro padrão, $s/\sqrt{n}$	:	2.16/$\sqrt{27}$ = 0.41

(O conceito de erro padrão é introduzido na Secção 2.3)

Seja j o índice da classe de comprimento. Denotamos o limite inferior e superior da respectiva classe de comprimento da seguinte maneira:

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

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

onde dL é o "*tamanho do intervalo de classe*". Um peixe de comprimento x(j) pertence à classe de comprimento j, onde:

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

Seja F(j) a frequência da classe de comprimento j, isto é o número de peixes observados na classe j. Seja $\bar{L}(j) = L(j) + dL/2$ o ponto médio da classe de comprimento correspondente. O cálculo do valor médio e da variância a partir de uma tabela de distribuição de frequências é efectuado normalmente usando-se o ponto médio para representar o intervalo de classe:

$n_{\cdot} = \sum_{j=i}^{m} F(j)$ é o número de observações totais, onde m é o número de classes de comprimento,

$\bar{x} = \frac{1}{n} * \sum_{j=1}^{m} F(j)*\bar{L}(j)$ é o valor médio e

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

O procedimento dos cálculos é mostrado na Tabela 2.1.2. O ponto médio da classe $\bar{L}(j)$, e o quadrado dos desvios em relação à média estão ponderados com o número de peixes em cada classe, ou seja, pela frequência F(j). Os resultados da Tabela 2.1.2 diferem ligeiramente dos obtidos na Tabela 2.1.1 porque a representação em classes de cm resulta numa precisão menor que a representação em classes de mm.

Tabela 2.1.2 Média e variância de uma amostra de frequências de comprimento. (A amostra é derivada da Tabela 2.1.1 com intervalos de comprimento, dL de 1 cm)

índice	intervalo (cm)	ponto médio (cm)	frequência			
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

comprimento médio, $\bar{x}$	:	407/27	= 15.074, digamos 15.07
variância, s^2	:	125.86/26	= 4.84
desvio padrão, s	:	$\sqrt{4.84}$	= 2.20
desvio padrão relativo, $s/\bar{x}$	:	2.20/15.07	= 0.15

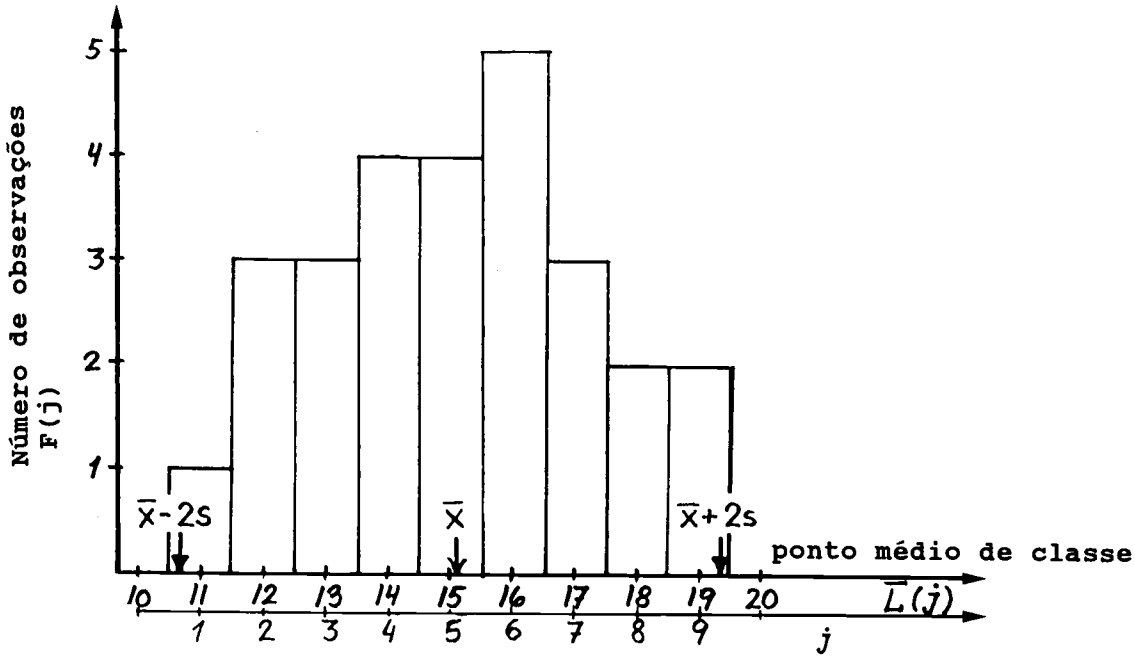


Fig. 2.1.1 Diagrama de frequências de comprimento. Representação gráfica das frequências da amostra da Tabela 2.1.2

A Fig. 2.1.1 mostra a representação gráfica da amostra de frequências. Observe que as observações encontram-se na faixa de intervalo de:

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

No caso de uma distribuição normal (a ser discutida na secção seguinte) espera-se que 95% das observações estejam contidas naquele intervalo.

(Ver **Exercício(s)** na Parte 2).

2.2 A DISTRIBUIÇÃO NORMAL

Na Tabela 2.1.2 e na Fig. 2.1.1 é apresentado um pequeno conjunto de dados de frequências de comprimento que seguem aproximadamente uma "*distribuição normal*". A expressão matemática que define uma distribuição normal é:

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

onde Fc = "*frequência calculada*" ou "*frequência teórica*", n = número de observações, dL = amplitude do intervalo de classe, s = desvio padrão, $\bar{x}$ = valor médio (comprimento médio) e π = 3.14159.

Usando os valores do exemplo da Tabela 2.1.2: n = 27, dL = 1 cm, s = 2.20, $\bar{x}$ = 15.07 cm:

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

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

Tabela 2.2.1 Frequências teóricas correspondentes à Tabela 2.1.2

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

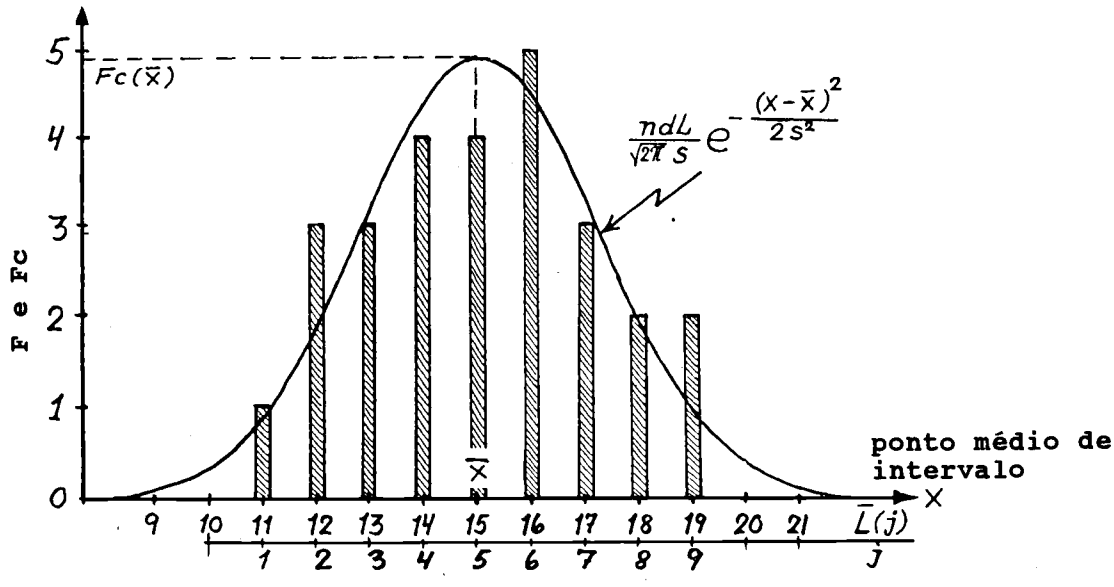


Fig. 2.2.1 Frequências teóricas, Fc, (curva de distribuição normal) e frequências observadas, F, (barras)

Os valores de Fc para diferentes valores de x são dados na Tabela 2.2.1. Observe que a notação foi ligeiramente modificada pois passamos a utilizar o ponto médio da classe x como argumento para Fc, em vez do índice j usado como argumento para F na Tabela 2.1.2.

A Fig. 2.2.1 mostra o gráfico das frequências teóricas (calculadas) em conjunto com o diagrama de barras das frequências observadas F(j) da Fig. 2.1.1. Como pode ser visto, Fc(x) produz um bom ajuste aos dados de frequências de comprimento observadas. Esta figura é geralmente observada quando tomamos frequências de comprimento de peixes originados de uma coorte, ou seja, peixes com aproximadamente a mesma idade.

A distribuição normal é geralmente observada em um grande número de situações, daí o nome "normal". Porém, outros tipos de distribuições probabilísticas, podem ser observadas no campo das ciências pesqueiras. Como exemplos podemos citar a "distribuição log-normal", a "distribuição binomial negativa", e a "distribuição delta". A diferença evidente entre estas e a distribuição normal é que são assimétricas, enquanto que a distribuição normal é simétrica. A distribuição delta por exemplo é usada para descrever a distribuição probabilística das capturas por hora de arrasto e é composta de uma distribuição log-normal, que descreve a distribuição das capturas por lances diferentes de zero e uma probabilidade especial para as capturas igual a zero (ver Secção 13.7, Fig. 13.7.2).

Talvez o mais importante aspecto da distribuição normal seja relacionado com os valores médios. Se tomarmos um número de 50 amostras de uma população, com 25 observações cada, os 50 valores médios serão distribuidos (aproximadamente) como uma normal. Assim, o valor médio tem uma distribuição probabilística. O valor médio de qualquer grupo de observações é distribuido (aproximadamente) como uma normal. Este resultado também é válido para

valores médios de distribuições log-normal, distribuições delta ou qualquer outro tipo de distribuição. Isto significa que os valores médios de todas as distribuições observadas em biologia pesqueira possuem uma distribuição aproximadamente normal.

Se dividirmos ambos os lados da Eq. 2.2.1 por n (= tamanho da amostra), obteremos:

$$Fc(x)/n = \frac{dL}{\sqrt{2\pi}} * \exp\left[-\frac{(x-\bar{x})^2}{2s^2}\right], \quad x = 1,2,3,... \qquad (2.2.2)$$

Os novos valores encontrados, Fc(x)/n, quando acumulados aproximam-se de 1.0. Cada valor indica a probabilidade de um peixe obtido aleatoriamente pertencer ao correspondente intervalo de comprimento. Isto é, os valores podem ser interpretados como a probabilidade de um peixe amostrado aleatoriamente pertencer ao intervalo de comprimento de x-dL/2 a x+dL/2.

Para os nove intervalos de classe da Tabela 2.2.1 encontramos:

j	intervalo	probabilidade
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

Isto significa, por exemplo, que a probabilidade de um peixe tirado aleatoriamente pertencer ao intervalo 14.5 a 15.5 será de 181 em 1000. Caso tivéssemos incluído na nossa amostra todos os intervalos de comprimento, (não somente os nove observados) os valores acumulados das probabilidades chegariam a 1.000.

Uma vez que a distribuição de comprimentos de uma única coorte de peixes pode ser descrita como uma distribuição normal, esta será usada nas análises de frequências de comprimento nos capítulos seguintes. Como introdução abordaremos a seguir alguns dos seus aspectos.

O procedimento para calcular a média e o desvio padrão (Tabela 2.1.2) pode ser efectuado para qualquer conjunto de dados de frequências de comprimento. No entanto, se por alguma razão, o diagrama de frequências observadas não representar a inteira distribuição, então os valores obtidos (das Eqs. 2.1.1 e 2.1.2) para a média e variância da amostra serão enviesados, ou seja, podem não ter qualquer relação com a média e variância da população. O conceito de "*erro*" será melhor discutido na Secção 7.1. Se, por exemplo, somente as frequências no intervalo de comprimento de 10 e 15 cm estiverem disponíveis (isto é, somente dados para o lado esquerdo), estamos numa situação onde a Eq. 2.1.1 (valor médio) e a Eq. 2.1.2 (variância) não representam a população. Como será visto no Capítulo 3, isto ocorre sempre quando analisamos frequências de comprimento, no entanto, existem várias formas de contornar o problema.

(Ver **Exercício(s)** na Parte 2).

2.3 LIMITES DE CONFIANÇA

Nesta Secção usaremos também como exemplo uma amostra de composição por comprimentos de uma coorte cujo comprimento médio, $\bar{x}$, foi estimado. Tal estimação difere, naturalmente da média real da população, uma vez que é impossível medir todos os peixes, daquela coorte, que se encontram no mar. Geralmente o comprimento médio real é desconhecido. No caso em que se lida com peixes criados em cativeiro é possível medir o comprimento médio real, mas no caso de um manancial no seu ambiente natural é impossível medir o valor real de qualquer parâmetro. Na prática, como é impossível medir todos os peixes capturados, isto aplica-se também, a populações de peixes provenientes da pesca. Passaremos então a tratar da precisão das estimações do comprimento médio, ou seja, qual o tamanho do desvio entre a estimação e a média real. O grau de incerteza em relação à média real é expresso pelos "*limites de confiança*" e no caso de uma distribuição normal, os limites de confiança inferior e superior são dados, respectivamente, por:

$$\bar{x} - t_{n-1}*s/\sqrt{n} \text{ e } \bar{x} + t_{n-1}*s/\sqrt{n} \qquad (2.3.1)$$

onde n é o tamanho da amostra, s o desvio padrão e t_{n-1} é o chamado quantil na "*distribuição-t*" ou "*distribuição de Student*" (Tabela 2.3.1). O argumento "f" na distribuição-t (Tabela 2.3.1) é chamado o "*número de graus de liberdade*". Em geral o número de graus de liberdade é o número de observações menos o número de parâmetros. No caso em questão $\bar{x}$ é o único parâmetro, assim f = n-1 e $t_f = t_{n-1}$ (ver Tabela 2.3.1).

Os limites de confiança podem ser calculados a diferentes níveis de precisão, usualmente 90%, 95% e 99%, como indicado na Tabela 2.3.1. Quanto maior o nível (percentagem), maior o quantil e portanto maior é o intervalo entre o limite superior e o inferior.

Retornando ao exemplo dado na Secção 2.1 (Tabela 2.1.2), pretende-se calcular, os limites de confiança a 95%, do comprimento médio do peixe na população da qual a amostra foi tirada. Usamos o quantil de 95% da distribuição-t (Tabela 2.3.1) com n-1 = 26 graus de liberdade e substituimos na Eq. 2.3.1:

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

Os limites de 95% de confiança são:

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

Assim, estamos "95% confiantes" que o verdadeiro comprimento médio é um valor entre 14.20 e 15.94. Por outras palavras, se a amostra for repetida 100 vezes, sob as mesmas condições, esperamos que as médias obtidas caiam 95 vezes entre 14.20 e 15.94. O intervalo entre o limite inferior e o limite superior é chamado "*intervalo de confiança*".

No exemplo usado acima os intervalos de confiança ao nível de 90% e 99% são respectivamente [14.35, 15.79] e [13.89, 16.25], dos quais o primeiro é mais pequeno e o segundo é maior do que o intervalo a 95%.

A quantidade $s/\sqrt{n}$ é o desvio padrão da estimação do comprimento médio (também chamado de "*erro padrão*") de modo que a variância de $\bar{x}$ é dada por (ver a Tabela 2.1.1):

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

Assim, quanto maior a amostra, mais precisa é a estimação de $\bar{x}$ (este assunto será melhor discutido na Secção 7.2).

Tabela 2.3.1 Quantils da distribuição-t (Distribuição de Student)*

graus de liberdade f	quantil 90% t(f)	95% t(f)	99% t(f)	graus de liberdade f	quantil 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

*) O uso da letra t neste contexto é universal. Neste manual t também é usado para representar a idade de um peixe. Esta tabela está repetida na última página deste volume para facilitar consultas

A Eq. 2.3.2 deriva-se de duas regras gerais para variáveis aleatórias que são repetidamente aplicadas neste manual e são:

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

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

onde C é uma constante. Por exemplo, quando a variância de x é s^2 então a variância de 3x é $9s^2$; ou, quando as observações originais são somadas três vezes, então a variância de $x_1+x_2+x_3$ é $3*s^2$.

As afirmações acima sobre limites de confiança aplicam-se somente para estimações "não viciadas" do valor médio. Em casos onde as amostras são viciadas, não importando quantos peixes foram tirados e medidos, obteremos sempre estimações do valor médio diferentes do valor real.

Supondo que queremos estimar o comprimento médio de uma certa espécie de peixes capturados, efectivamente, numa pescaria comercial (note que os peixes capturados correspondem a peixes desembarcados mais aqueles devolvidos ao mar). Assim, se amostramos somente os desembarques e não os peixes, geralmente abaixo de certo tamanho, que são devolvidos ao mar, obteremos uma estimação viciada do comprimento médio dos peixes capturados. O comprimento médio da captura será sobrestimado, não importando quantos peixes sejam amostrados nos desembarques. Sendo assim, podemos obter apenas uma estimação não viciada do comprimento médio dos peixes desembarcados.

(Ver **Exercício(s)** na Parte 2).

2.4 ANÁLISE DE REGRESSÃO LINEAR SIMPLES

Este método é utilizado quando queremos descrever a variação de uma quantidade, por exemplo, a altura do corpo do peixe, como uma função linear de outra quantidade, por exemplo, o comprimento total do corpo. A teoria requer que a quantidade no eixo horizontal (a variável independente) seja medida com precisão absoluta. Contudo, por vezes, o método é aplicado com a violação deste requerimento e o efeito da imprecisão dos valores da variável independente é que o declive da recta fique mais "achatado" (próximo de zero).

Suponha que medimos o comprimento total e a altura do corpo de uma amostra de 7 peixes.

A Tabela 2.4.1 mostra o comprimento total, x(i), e a correspondente altura do corpo y(i), i = 1,2,...,7.

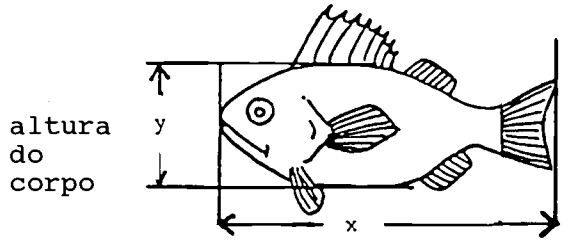


Tabela 2.4.1 Amostra de comprimentos totais, x, e alturas do corpo correspondentes, y

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

Como seria de esperar, a altura do corpo tende a crescer com o aumento do comprimento total. Se as proporções de um peixe permanecerem constantes para todos os tamanhos do peixe, a altura será proporcional ao tamanho, e poderá ser descrita pelo modelo:

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

onde b é uma constante, também chamado "*parâmetro*". O gráfico deste modelo passa sempre pela origem, o ponto de intersecção do eixo dos x com o eixo dos y. Podemos ainda permitir um desvio na proporcionalidade entre x e y introduzindo um segundo parâmetro, a, e usar, em vez da Eq. 2.4.1, o seguinte modelo:

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

onde a indica o ponto de intersecção da recta que se ajusta aos pontos com o eixo dos y. A Fig. 2.4.1 mostra o gráfico (ou "*diagrama de dispersão*") de y(i) contra x(i).

Uma implicação da Eq. 2.4.2 é que um peixe de comprimento zero tem uma altura a, o que não faz sentido, a não ser que a seja zero. No entanto, se somente uma certa gama de comprimentos for considerada (por exemplo comprimentos acima de 5 cm), o modelo com dois parâmetros dará um melhor ajuste às observações que o modelo de um parâmetro, pois o pressuposto da proporcionalidade entre comprimento e altura não é estritamente verificado.

O modelo matemático da Eq. 2.4.2 é chamado "*modelo linear*" porque os pares (x,y) que se ajustam ao modelo seguem uma linha recta. Com a = -0.32 e b = 0.30 obtemos a linha recta mostrada na Fig. 2.4.1. Com estes valores de a e b a recta na Fig. 2.4.1 ajusta-se bem aos pares observados (x,y).

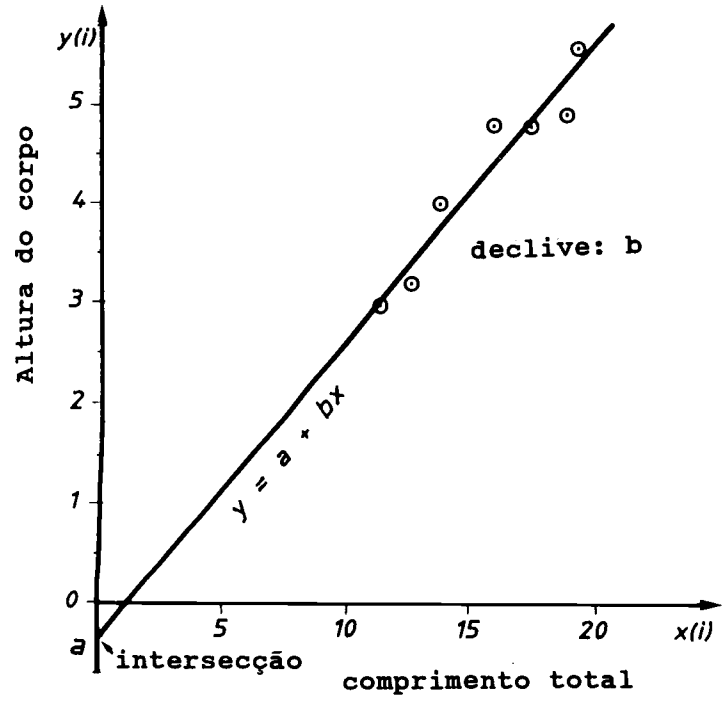


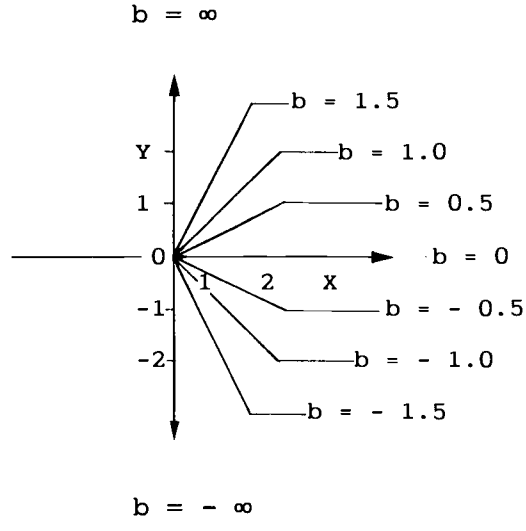
Fig. 2.4.1 Diagrama de dispersão da altura do corpo (y) contra o comprimento total (x), também chamado "gráfico de y contra x"

Vejamos agora a questão da determinação da recta, isto é, como estimar os parâmetros a e b. Exactamente como foi feito para o valor médio (cf. Secção 2.3) demostraremos também como os limites de confiança de a e b são calculados. Este procedimento é chamado "*análise de regressão linear simples*". Esta é, provavelmente, a técnica estatística mais usada em biologia pesqueira. Existem nomes específicos para os parâmetros: a é chamado "*intersecção ou ordenada na origem*" e b o "*declive*". A intersecção é a distância do ponto (0,0) no diagrama (x,y) ao ponto onde a "*recta de regressão*"

$$y = a + b*x$$

intersecta o eixo dos y (ver Fig. 2.4.1).

O declive, b, indica o grau de inclinação da recta. Se b = 0 a recta é paralela ao eixo dos x. Se b é positivo a inclinação é ascendente. Se b é negativo a inclinação é descendente.



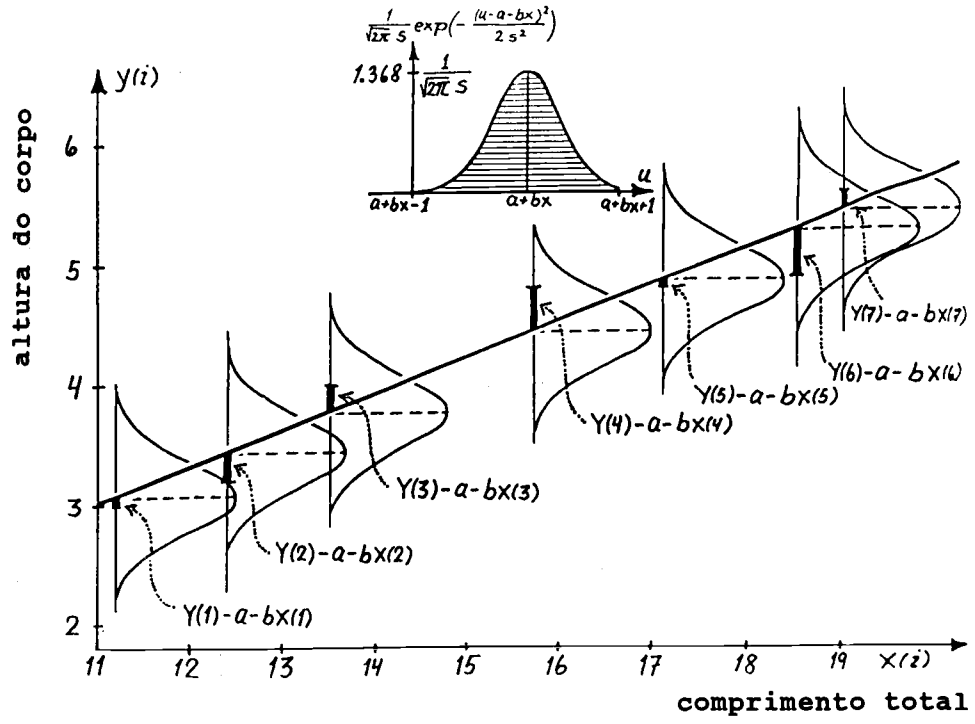


Fig. 2.4.2 Ilustração do pressuposto básico da análise de regressão linear simples. Cada y(i) para um dado x(i) é normalmente distribuido com uma variância comum

A variável no eixo horizontal, x, é chamada "*variável independente*" e a variável no eixo vertical, y, é a "*variável dependente*". A recta de regressão é determinada como sendo aquela que minimiza a soma dos quadrados dos desvios entre a recta y = a + b*x e os pares de observações, (x(i),y(i)). Dizemos que a e b são estimados pelo "*método dos mínimos quadrados*", ou seja, procuramos valores de a e b que minimizem a expressão:

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

onde n é o número de pares de observações (n = 7 no exemplo). Os desvios entre a recta e as observações são ilustrados na Fig. 2.4.2. O pressuposto básico da análise de regressão é que cada y(i) é normalmente distribuído com valor médio a + b*x(i) e com variância constante. Isto é, uma variância que não depende do valor de x(i). A seguinte fórmula para estimar a variância comum difere apenas ligeiramente daquela introduzida na Secção 2.1. A chamada "*variância sobre a recta de regressão*" é:

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

Temos n-2 graus de liberdade (número pelo qual a soma é dividida) porque são dois parâmetros, a e b.

Tabela 2.4.2 **Cálculos para a análise de regressão linear simples. Os resultados marcados com #) não são usados no cálculo de a e b, mas foram calculados para posterior utilização**

i	comprimento total x(i)	$x(i)^2$	altura do corpo 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)

$\bar{x} = 15.343$ | $\bar{y} = 4.329$

$\frac{1}{n}*(\Sigma x(i))^2 = 1647.82$ | $\frac{1}{n}*(\Sigma y(i))^2 = 131.16$ #)

$\Sigma x(i)^2 - \frac{1}{n}*(\Sigma x(i))^2 = 55.78$ | $\Sigma y(i)^2 - \frac{1}{n}*(\Sigma y(i))^2 = 5.534$ #)

$sx^2 = 9.296$ #) | $sy^2 = 0.922$ #)

$sx = 3.049$ #) | $sy = 0.960$ #)

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

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

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

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

Tabela 2.4.3 **Cálculo da variância da recta da Eq. 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		soma:	0.4252

As estimações dos parâmetros a (intersecção) e b (declive) são obtidas por:

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

$$a = \bar{y} - \bar{x}*b \qquad (2.4.6)$$

onde $\bar{y}$ e $\bar{x}$ são os valores médios de y e x conforme definido pela Eq. 2.1.1.

Na Tabela 2.4.2 os cálculos para estimar a e b são demonstrados usando-se os dados da Tabela 2.4.1. Assim, a recta de regressão estimada é:

$$y = -0.315 + 0.303*x \qquad (2.4.7)$$

Para calcular os limites de confiança de a e b necessitamos da soma dos quadrados dos desvios de x e y. As variâncias de x e y são definidas pela Eq. 2.1.3 como se segue:

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

e uma expressão similar para sy^2. Para uso na próxima secção, introduz-se a expressão "*covariância*":

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

O procedimento para o cálculo da variância da recta de regressão é dado pela Eq. 2.4.4 e demonstrado na Tabela 2.4.3. No entanto, a variância da recta de regressão pode ser obtida mais facilmente a partir de sy e sx:

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

Dado os resultados da Tabela 2.4.2, a Eq. 2.4.10 fica:

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

As variâncias das estimações de b e a são:

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

e

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

Dado os resultados da Tabela 2.4.2 obtemos:

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

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

Os limites de confiança para a intersecção a e o declive b são respectivamente:

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

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

Os limites de confiança ao nível de 95% para o exemplo com n = 7 peixes e $t_{7-2} = 2.57$ (Tabela 2.3.1) será:

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

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

Note que o intervalo de confiança para a intersecção, a, contém zero, o que significa que a hipótese de a altura do corpo ser directamente proporcional ao comprimento (ou seja "a = 0") não pode ser rejeitada com 95% de confiança. Diz-se, então, que, a, não é significativamente diferente de 0 ao nível de 95%.

Se existir uma boa razão para assumir que a = 0, então o valor estimado deve ser substituido por 0, se a estimação não fôr significativamente diferente de zero. Assim sendo, b deve ser recalculado como se segue:

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

A estimação é baseada somente em sete peixes. Se tivéssemos medido 200 peixes, a estimação do desvio padrão, sa, seria menor (cf. Eq. 2.4.11 e 2.4.12). Assumindo, por exemplo, que $\bar{x}$, $\bar{y}$, sx, sy, a e b obtidos para a amostra de n = 7, são os mesmos para uma amostra de n = 200 (que poderia muito bem ter ocorrido), embora as estimações de a e b venham a ter o mesmo valor, os seus desvios padrão, sa e sb, serão diferentes.

Com n = 200 a Eq. 2.4.11 dá sb = 0.006098, enquanto que a Eq. 2.4.12 dá-nos sa = 0.0091 e $t_{198} = 1.97$ (Tabela 2.3.1). Assim sa e sb tornam-se bem menores e consequentemente o intervalo de confiança também é menor.

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

A estimação de a seria agora significativamente diferente de zero. Neste caso podemos concluir que para o verdadeiro valor de a ser maior que -0.30 ou menor que -0.33 a probabilidade será menor que 5%.

(Ver **Exercício(s)** na Parte 2).

2.5 O COEFICIENTE DE CORRELAÇÃO E A REGRESSÃO FUNCIONAL

O "*coeficiente de correlação*", r, é uma medida da associação linear entre duas quantidades, ambas sujeitas à variação aleatória. A amostra de comprimento total e altura do corpo da Secção 2.4 é um exemplo das duas quantidades associadas. Neste caso sete peixes foram retirados aleatoriamente. Por acidente poderíamos ter retirado sete peixes com o mesmo comprimento, o que não seria adequado para a estimação da relação comprimento/altura porque os limites de confiança de a e b seriam muito amplos.

O coeficiente de correlação pode ser usado apenas quando ambas as medidas podem variar aleatoriamente. Se tivéssemos relacionado sete peixes com comprimentos predeterminados em vez de comprimentos aleatórios (por exemplo, se tivéssemos relacionado os comprimentos 4, 6, 8, 10, 12, 14 e 16 cm para a amostra comprimento/altura) o cálculo do coeficiente de correlação para esta amostra seria incorreto.

O coeficiente de correlação é definido por:

$$r = \frac{sxy}{sx*sy} \qquad (2.5.1)$$

onde sxy é definido pela Eq. 2.4.9 e sx e sy pela Eq. 2.4.8.

Sendo o declive ($b = sxy/sx^2$), a Eq. 2.5.1 fica:

$$r = b*sx/sy \qquad (2.5.2)$$

A amplitude de variação de r é: $-1.0 \leq r \leq 1.0$. O r é negativo se y tende a decrescer com o aumento de x e é positivo se y tende a aumentar com o aumento de x. Esta afirmação também é válida para o declive b e é derivada da Eq. 2.5.2. Como sx/sy é sempre positivo (cf. definido pela 2.4.8) r tem o mesmo sinal do declive b. Os casos extremos, r = 1 ou r = -1, ocorrem quando todos os pares (x,y) caiem exactamente sobre a recta. Quanto mais próximo de zero for r, menor o grau da associação linear entre y e x. Quando r = 0, x e y são independentes entre si.

A Fig. 2.5.1 mostra quatro exemplos de diagramas de dispersão com diferentes valores de r. Para o exemplo da Tabela 2.4.2 obtemos:

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

Vamos chamar r1 (inferior) e r2 (superior) os limites de 95% de confiança de r, que podem ser calculados pela expressão:

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

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

onde "tanh" é a "tangente hiperbólica", que é uma função encontrada em muitas calculadoras científicas de bolso.

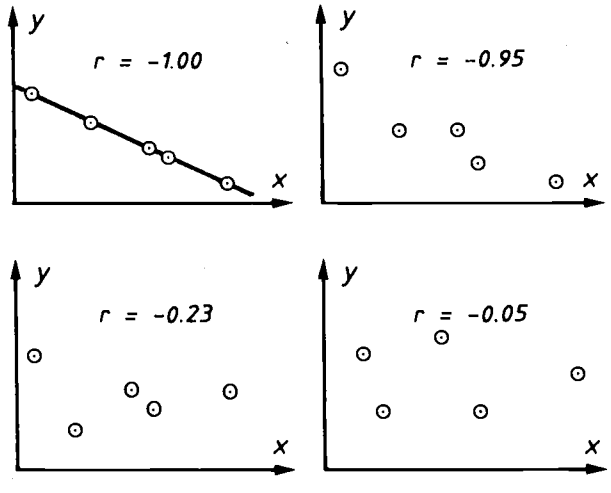


Fig. 2.5.1 Exemplos de coeficientes de correlação

Com r do exemplo (r = 0.961, n = 7) os limites de confiança a 95% seriam: (r1 e r2] = [0.75,0.99]. Os limites de confiança a 99% podem ser facilmente obtidos substituindo-se 1.96 por 2.58 na Eq. 2.5.3.

Frequentemente estamos interessados em saber se zero se encontra dentro dos limites de confiança, para sabermos se há possibilidade de que a associação linear seja apenas um acaso. Neste exemplo a probilidade que a relação linear seja um acaso é menor que 5%, uma vez que zero não se encontra dentro do intervalo de confiança.

No exemplo da regressão da altura do corpo contra o comprimento total, o comprimento foi escolhido como variável independente e a altura como variável dependente. Porém não há qualquer razão especial para esta escolha. A nossa amostra de comprimentos consiste em sete peixes escolhidos aleatoriamente, não se controlando que comprimentos e alturas seriam, podendo assim, ter-se feito a escolha contrária para a variável dependente e independente.

Uma das suposições básicas que fundamenta a análise de regressão linear é que a variável independente não possa ser aleatória. A variável independente deve ser qualquer, da qual podemos determinar valores prévios. Por exemplo, se a variável independente é a data da amostra, ela pode ser previamente determinada. Poderíamos decidir tomar amostras no primeiro dia de cada mês. Se medirmos o tempo em unidade de anos e começarmos com o tempo zero em primeiro de janeiro, a variável independente assumiria valores: 0, 1/12, 2/12, 3/12 ... etc. Estes são, claramente valores, não aleatórios.

No caso dos sete peixes escolhidos aleatoriamente no exemplo acima, a análise de correlação pode ser aplicada, pois a amostra foi retirada aleatóriamente duma distribuição normal de comprimentos. Contudo, estaríamos aptos para decidir os comprimentos préviamente, uma vez que poderíamos escolher os quatro peixes menores e os três maiores. Poderíamos também decidir, como fizemos, retirá-los aleatóriamente. Somente nesta última situação é permitido aplicar as duas análises. No caso dos comprimentos préviamente definidos, só se pode aplicar a análise de regressão. Por outro lado, seria a maneira, provavelmente, mais efectiva de aplicar a análise de regressão devido à distância grande entre as observações no eixo horizontal. Isto resultaria numa pequena variância do declive. Escolhendo os peixes aleatoriamente, seria provável que a maioria deles pertencesse ao intervalo médio de comprimentos, contribuindo pouco para a determinação do declive, cuja variância seria grande.

Uma outra questão é se poderiamos obter um resultado diferente usando a altura do corpo como variável independente. Primeiro, temos que considerar se a medida da altura é tão exacta como a do comprimento. Caso não seja, o declive poderá ser enviesada (achatado). No entanto, continuam a existir certos problemas, mesmo que seja possível medir as duas variáveis com o mesmo grau de precisão.

Usando agora a altura do corpo como variável independente obtem-se a "*regressão inversa*". Somente no caso excepcional de todas as observações caírem na recta de regressão (isto é, r = 1 ou r = -1), os mesmos resultados da regressão simples seriam obtidos para a regressão inversa. A equação y = a + b*x (Eq. 2.4.2) é matematicamente equivalente a:

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

ou $x = A + B*y$ onde $A = -a/b$ e $B = 1/b$ (2.5.4)

Desenvolvendo a regressão inversa (Eq. 2.5.4) obtemos:

$$A = 2.139 \quad e \quad B = 3.05$$

A equação: $x = 2.139 + 3.05*y$ pode ser convertida em:

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

que pode ser comparada com os resultados encontrados para a regressão original (Eq. 2.4.7: $y = -0.315 + 0.303*x$). Assim, os resultados da regressão inversa diferem dos obtidos para a análise de regressão original.

Uma maneira de contornar o problema da escolha da variável independente, quando ambas as variáveis são aleatórias, é usar a "*análise de regressão funcional*" (ver Ricker, 1973). Este método estima o declive (que chamaremos de b' para distinguir do declive b da regressão simples) pela expressão:

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

e a intersecção:

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

Este tipo de análise dá um resultado que pode ser considerado um compromisso entre a regressão simples original e a sua contraparte inversa.

Com os resultados da Tabela 2.4.2 obtemos:

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

e $y = -0.504 + 0.315*x$

A análise de regressão funcional é mencionada aqui apenas a título de complementação. Existem outras limitações relacionadas à sua aplicabilidade que não trataremos aqui.

As três seguintes rectas de regressão foram estimadas:

1. Análise de regressão simples original: $y = -0.315 + 0.303*x$
2. Análise de regressão funcional : $y = -0.504 + 0.315*x$
3. Análise de regressão simples inversa : $y = -0.701 + 0.328*x$

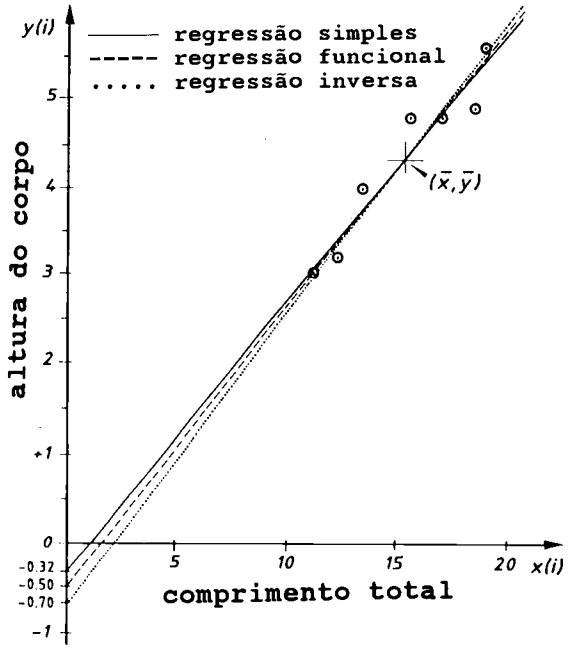


Fig. 2.5.2 Rectas de regressão funcional e inversa comparadas com a recta de regressão original

A Fig. 2.5.2 mostra as três rectas de regressão. Note que todas as três rectas passam pelo ponto $(\bar{x},\bar{y})$ e que um aumento do declive é parcialmente compensado por um decréscimo na intersecção.

*(Ver **Exercício(s)** na Parte 2).*

2.6 TRANSFORMAÇÕES LINEARES

As funções lineares são matematicamente fáceis e possuem ainda a vantagem de serem interpretadas graficamente sem quaisquer problemas. No entanto, muitas relações funcionais observadas em biologia pesqueira não são lineares. Felizmente, tais funções não lineares podem ser sempre convertidas em funções lineares, o que significa que após a conversão podemos tratá-las da maneira descrita nas secções anteriores. Seguem-se vários exemplos de aplicações das transformações de funções não lineares em funções lineares.

Exemplo 1: Relação peso-comprimento

Vamos considerar o famoso exemplo da relação funcional entre o peso e o comprimento do corpo de um peixe. A Fig. 2.6.1 mostra um gráfico do peso contra o comprimento do falso-besugo, *Nemipterus marginatus*. Observa-se, claramente, que esta relação não é linear. A curva na Fig. 2.6.1 é uma função do tipo:

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

onde W(i) é o peso do corpo do peixe nº i, L(i) é o comprimento do corpo e q e b são os parâmetros. A Eq. 2.6.1 é geralmente chamada "*relação peso-comprimento*" e pode ser convertida em uma equação linear aplicando-se logarítmos em ambos os lados da equação:

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

ou

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

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

Com a Eq. 2.6.2a podemos agora proceder à estimação de a e b pela análise de regressão linear. Os dados de entrada são mostrados na Tabela 2.6.1 e o diagrama de dispersão correspondente na Fig. 2.6.2. Os resultados são:

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

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

Como $a = \ln q$ podemos obter q da relação peso-comprimento original (Eq. 2.6.1) através do antilog de a:

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

Assim, a relação estimada entre W (em g) e L (em cm) será:

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

(A transformação inversa dos logarítmos introduz um erro que não será tratado aqui).

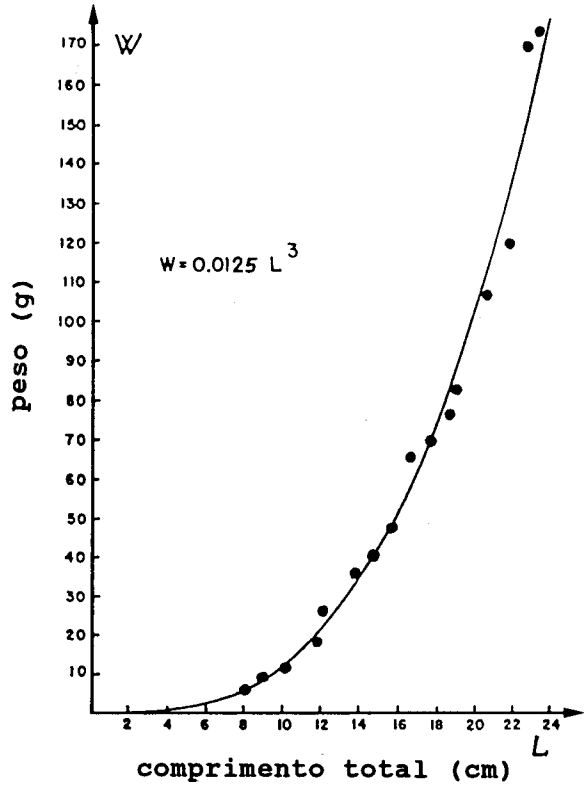


Fig. 2.6.1 Relação peso/comprimento do *Nemipterus marginatus* no Mar do Sul da China (baseado nos dados da Tabela 2.6.1)

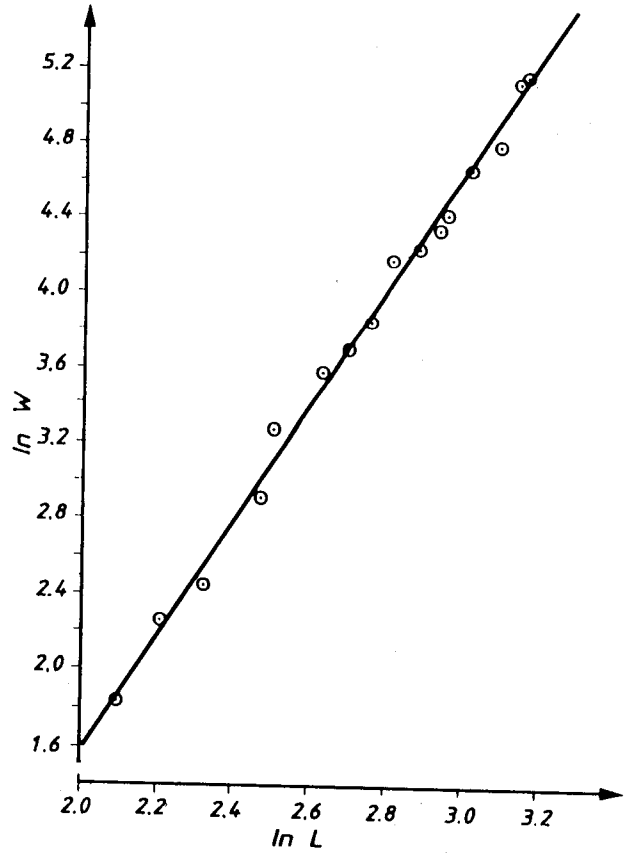


Fig. 2.6.2 Dados da Fig. 2.6.1 convertidos em logarítmos naturais

Tabela 2.6.1 **Dados para a estimação da relação peso-comprimento do falso besugo (*Nemipterus marginatus*) do Mar do Sul da China (de 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
	soma		43.629	60.786
	média		2.7268	3.7991
	sx e sy		0.3311	1.0161

Podemos também calcular os limites de 95% de confiança de b, usando os valores de sx, sy, n e t_{14} (ver Tabela 2.3.1) na Eq. 2.4.11:

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

$$sb = 0.072 \text{ e } sb*t_{n-2} = 0.072*2.15 = 0.155$$

O intervalo de 95% de confiança de b é [(3.057-0.155),(3.057+0.155)] ou [2.90,3.21]. Estes limites de confiança significam que somente a primeira décima na estimação de b é significativa (cf. Secção 2.3), e o verdadeiro valor de b poderia perfeitamente ser 3.0.

Como o peso de um peixe (em gramas) é aproximadamente igual ao seu volume (em cm cúbicos) e como o seu volume é frequentemente proporcional ao cubo do seu comprimento, L^3, poderíamos esperar que o valor de b nas Eqs. 2.6.1 e 2.6.2 seja próximo de 3.0.

Como o intervalo de confiança calculado acima suporta esta hipótese podemos simplificar a relação peso-comprimento substituindo a estimação de b = 3.057 por b = 3.0. Como a nova recta com b = 3.0 também passa pelo ponto $(\bar{x},\bar{y})$ tambem podemos calcular a nova intersecção a usando a Eq. 2.6.2a:

$$a = \bar{y} - b*\bar{x} = 3.799 - 3.0*1.727 = -4.382$$

De a obtemos o novo valor correspondente de q

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

Assim a nova relação sera:

$$W = 0.0125*L^3$$

Exemplo 2: Linearização da distribuição normal

Na Secção 2.2 (Eq. 2.2.1) a expressão matemática para a distribuição normal foi dada como:

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

Esta equação pode ser convertida em uma função linear através dos seguintes passos:

Passo 1: ***Conversão da distribuição normal em uma parábola***

Aplicando logarítmos em ambos os lados da Eq. 2.2.1 obtemos:

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

Considerando ln Fc(x) como a variável dependente, y, e x como variável independente, obtemos assim uma relação funcional entre y e x, que pode ser representada graficamente por uma parábola cuja fórmula geral é:

$$y = a + b*x + c*x^2$$

Inserindo os valores usados no exemplo da Tabela 2.1.2 obtemos:

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

cujo gráfico é mostrado na Fig. 2.6.3.

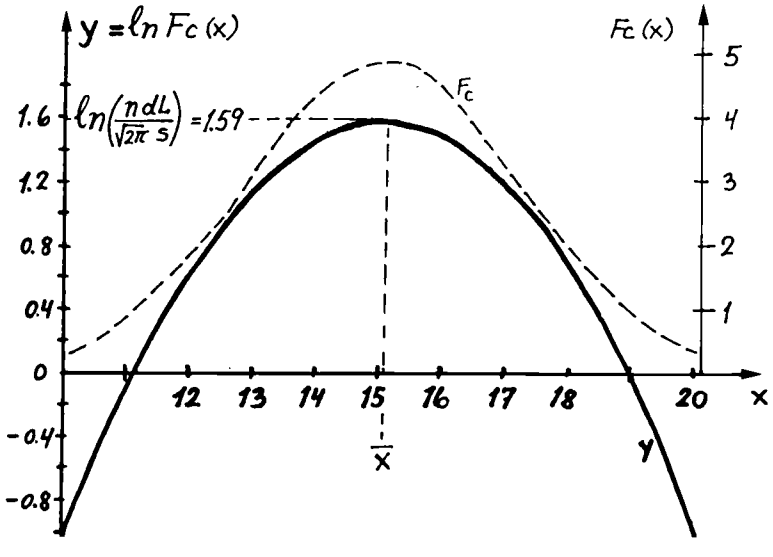


Fig. 2.6.3 Distribuição normal transformada em logarítmo (y) junto com a distribuição original (Fc)

Passo 2: *Conversão duma parábola em uma recta*

Numa parábola a diferença entre os pontos do eixo dos x, mesmo que espaçados, dá sempre uma recta. Subtraindo o valor mais alto de x (neste caso: ln Fc(x)) do valor mais baixo de x resulta uma série de diferenças que são positivas na parte esquerda da parábola e negativas na parte direita da parábola. O procedimento e o resultado das diferenças calculadas estão ilustradas na Figs. 2.6.4A e 2.6.4B,respectivamente.

Para explicar o procedimento matematicamente, introduz-se uma nova variável dependente, y', que é a diferença entre o logaritmo do número numa dada classe e o logaritmo do número na classe seguinte.

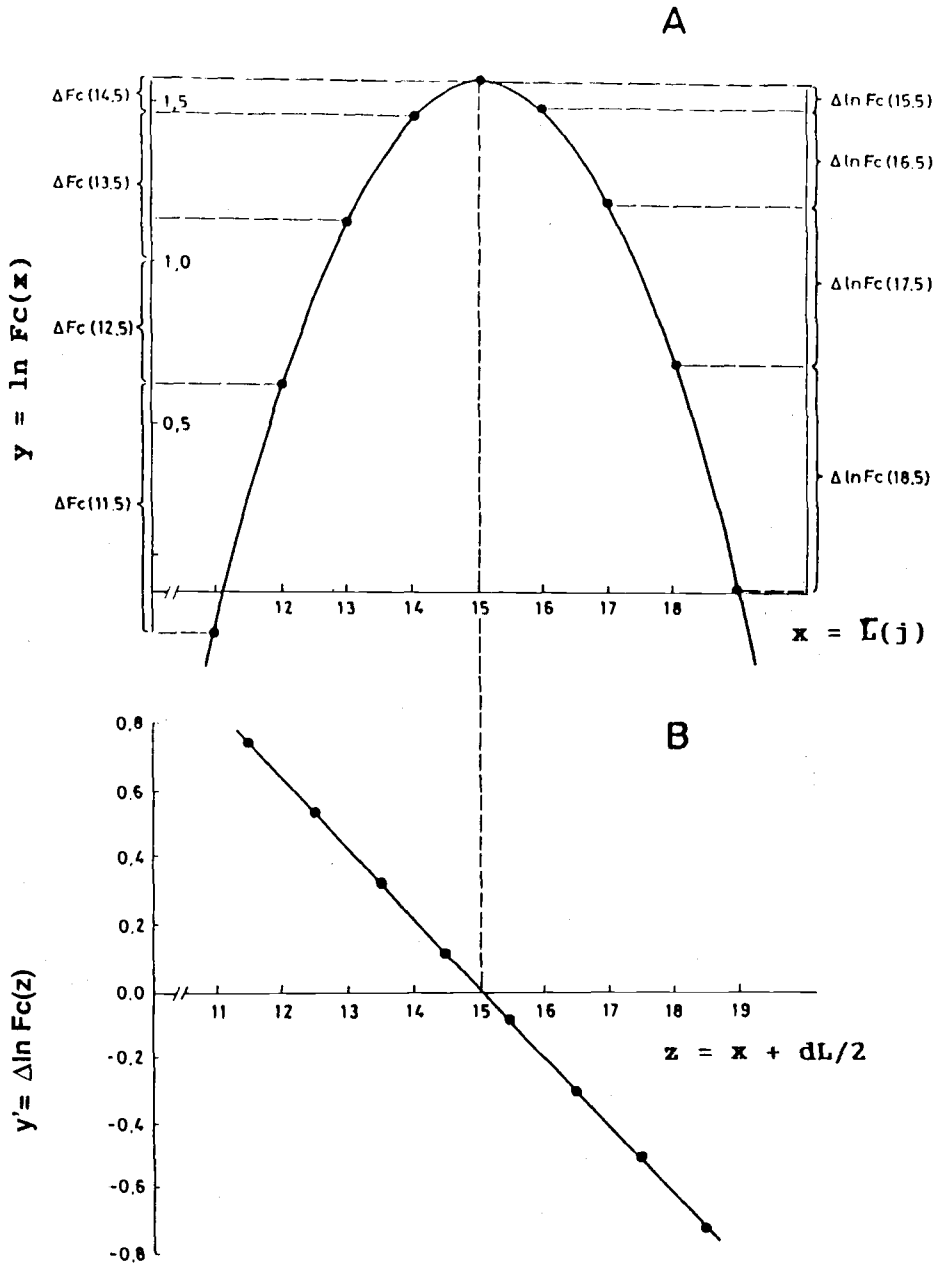


Fig. 2.6.4 **Estimação da média e da variância utilizando o método Bhattacharya. A: Parábola mostrando as diferenças dos pontos equidistantes no eixo dos x. B: Diagrama de Bhattacharya das diferenças contra o ponto médio das classes. Dados na Tabela 2.6.2**

$$y' = \ln Fc(x+dL) - \ln Fc(x) \qquad (2.6.4)$$

Pode-se expressar também do seguinte modo

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

onde Δ (delta) designa uma "pequena" diferença entre dois valores. y' é graficado contra a nova variável independente z, equivalente a x mais metade do intervalo de comprimento:

$$z = x + dL/2$$

Tabela 2.6.2 **Estimação do valor médio e da variância da distribuição normal do diagrama de Bhattacharya, ilustrado pelas frequências teóricas, Fc(x), da Tabela 2.1.2., apresentadas na Tabela 2.2.1. A Tabela está ilustrada na Figs. 2.6.3 e 2.6.4**

índice 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	(dL = 1)
					b = -0.2073	

$\bar{x} = -a/b = 15.07$ $\quad s^2 = -dL/b = 4.82$ $\quad s = 2.20$

Temos agora que inserir a Eq. 2.6.3 na Eq. 2.6.4 como se segue:

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

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

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

Depois de elevar ao quadrado e somar converte-se numa equação relativamente simples:

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

ou $y' = a + b*z$, onde $z = x + dL/2$

$a = dL*\bar{x}/s^2$ e $b = -dL/s^2$

Do declive, b, e da ordenada na origem, a, obtêm-se a variância e o valor médio, respectivamente:

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

e

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

Esta regressão é um dos elementos principais do método descrito por Bhattacharya (1967) para separar duas ou mais distribuições normais (Secção 3.4.1) e é chamado o "diagrama de Bhattacharya", exemplificado na Tabela 2.6.2 e na Fig. 2.6.4. Neste caso, os valores teóricos, Fc, da Tabela 2.2.1 foram usados como "observações" e seguem exactamente o modelo, a média e a variância estimadas pelo diagrama de Bhattacharya devem ser idênticos aos obtidos pelo método tradicional (como na Tabela 2.1.2). No entanto, se existir alguma pequena diferença, será devida à introdução da regressão linear. A Fig. 2.6.4 mostra o gráfico das diferenças logarítmicas entre 2 frequências consecutivas contra o ponto médio do valor de x.

O diagrama de Bhattacharya também nos dá uma ideia do número de observações numa distribuição normal, no qual somente se conhece as frequências em algumas classes de comprimento. Reescrevendo a Eq. 2.2.1 com as observações actuais, obtêm-se:

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

Assim sendo, n pode ser estimado para uma única classe de comprimento j, uma vez que $\bar{x}$ e s^2 já tenham sido estimados. Erros na amostragem, no entanto, causam inprecisão devido a influenciarem o número de peixes em cada intervalo de classe, cf. Fig. 2.2.1. Quando os valores em várias classes de comprimento forem conhecidos, as frequências podem ser somadas minimizando os desvios de cada frequência esperada. Somando para cada classe i de ambos os lados do sinal de igual e reordenando, obtêm-se

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

As observações para peixes maiores que $\bar{x}$ da Fig. 2.6.4 podem não ser de confiança pois os seus comprimentos sobrepõem-se com os peixes mais pequenos de um grupo de idade mais velho, tal como está ilustrado na Fig. 1.4.1, para o grupo de idades 1 e 2. Neste caso só poderemos utilizar as observações do lado esquerdo da Fig. 2.6.4 (x = 11, 12, 13, 14, 15 cm) do diagrama de Bhattacharya. Neste, 4 pontos formam uma linha recta do qual se estima a $\bar{x}$ e s^2. Calculando dos dados da Tabela 2.6.2:

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

Neste caso o resultado é praticamente o mesmo ao da distribuição normal completa, pois o ajuste da recta aos dados é quase perfeito (ver Fig.

2.6.2). A aplicação da Eq. 2.6.9 está na Tabela 2.6.2a. Determinou-se n = 26.88 sendo o verdadeiro valor 27 (conhecido da Tabela 2.1.2).

Tabela 2.6.2a Estimação do número total de observações através do método 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
soma		15.11	3.0907

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

Uma vez que n seja conhecido, o número em cada classe de comprimento (as fraquências teóricas) podem ser estimadas da Eq.2.6.8. Estes calculos não estão efectuados na Tabela 2.6.2a devido a neste exercício as "observações serem de facto as frequências teóricas.

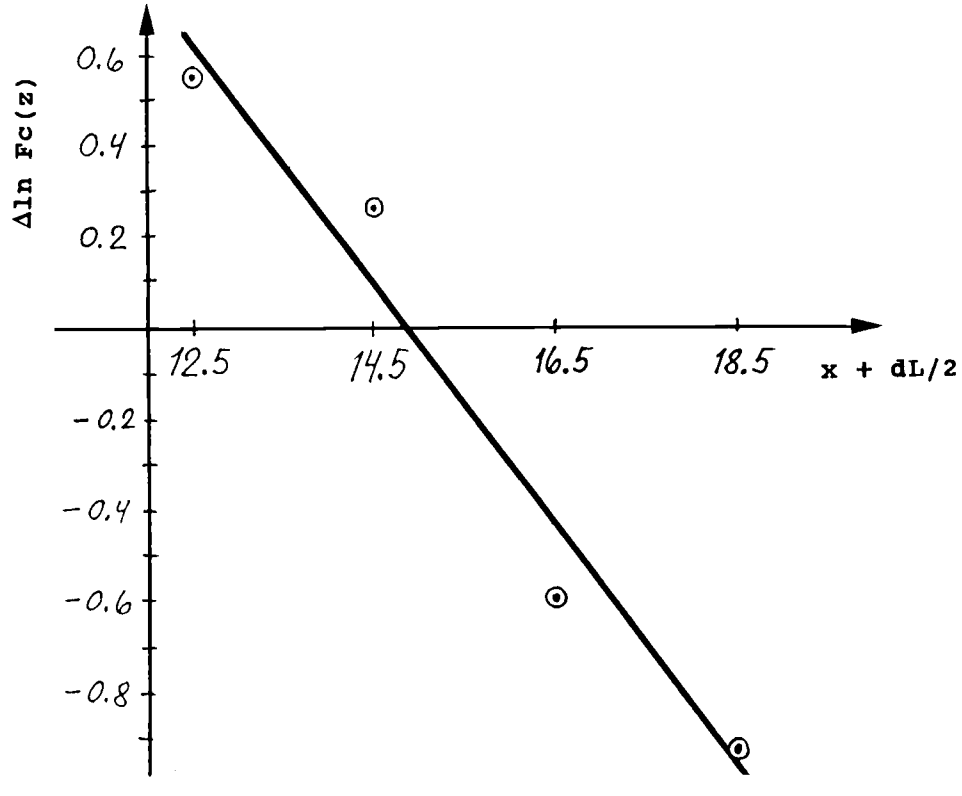


Fig. 2.6.5 Diagrama de Bhattacharya correspondente à Tabela 2.6.3

Tabela 2.6.3 **Diagrama de Bhattacharya correspondente à amostra de distribuição de frequências da Tabela 2.1.2**

índice	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 b = -0.263	(dL = 2)
$\bar{x}$ = -a/b = 14.8				s^2 = -dL/b = 7.605	s = 2.76	

A Tabela 2.6.3 mostra a estimação do valor médio e da variância do diagrama de Bhattacharya, mas agora com as observações actuais dadas na Tabela 2.1.2. Devido ao pequeno tamanho da amostra, as observações são agrupadas em intervalos de 2 cm. A Fig. 2.6.5 mostra o gráfico correspondente. As estimações do valor médio e da variância na Tabela 2.6.3 desviam-se das que foram calculadas pelo método tradicional (Tabela 2.1.2) devido a 1) tamanho pequeno da amostra, 2) erro proveniente de longos intervalos e 3) aplicação de um método estatístico diferente (análise de regressão linear).

(Ver **Exercício(s)** na Parte 2).