2. BIOSTATISTICS

This chapter contains a brief description of some statistical methods in common use in tropical fishery biology and introduces the statistical notation adopted in the manual. It serves as a refresher and reference point, but is not intended as a textbook in its own right.

The amount of literature on statistical methods is staggering, so there is no problem if you want to do further studies in biostatistics. Only two references are given here. The book "Biometry" by Sokal and Rohlf (1981) deals with the theory in a rather accessible way, while "Sampling techniques" by Cochran (1977) is perhaps a bit more complicated, but still recommended as an introduction. However, there are many other textbooks which may be equally useful.

2.1 MEAN VALUE AND VARIANCE

Let us consider a sample of n fish all of one species caught in one trawl haul and let x(i) be the length of fish no. i, i = 1,2,...,n. The "*mean length*" (in general the "*mean value*") of the sample is defined:

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

The two first columns of Table 2.1.1 show an example for n = 27.

The variance, which is a measure of the variability about the mean value is defined as follows:

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

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

Thus, the variance, s^2, is the sum of the squares of the deviations from the mean divided by the number, n, minus one. The third and fourth column of Table 2.1.1 illustrate the calculation of the variance. Note that if all fish in the sample had the same length this would equal the mean length and the variance would be zero. The sum of the deviations (not squared) is always zero. The larger the deviations from the mean value, the larger the variance will be. The two largest values of the square of the deviations from the mean in Table 2.1.1 occurred for the smallest and the largest observations.

The square root of the variance, s, is called the "*standard deviation*". Often one is interested in the variance relative to the size of the mean length, and for that purpose s is the relevant quantity as it has the same unit as the mean. This leads to the relative standard deviation, $s/\bar{x}$, also called the "*coefficient of variation*".

When doing the calculations by hand it is easier to work with a rearranged form of Eq. 2.1.2, which is equivalent to

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

However, as most scientific pocket calculators contain an option for auto-matic calculation of mean and variance the calculations here are illustrated by Eq. 2.1.2, which is conceptually easier to understand.

For many purposes, e.g. for graphical representation, it is convenient to arrange the sample in the form of a "*frequency table*" by dividing the length range into a number of length intervals. The length range for the sample in Table 2.1.1 goes from 11.2 to 19.0 cm. With length groups of 1 cm we need nine length groups to cover the range. Using 10.5 as the lower limit of the first length interval, the intervals and the frequencies of lengths become those shown in the first four columns of Table 2.1.2, which is a so-called length-frequency table.

Table 2.1.1 Mean value, variance and standard deviation of a length-frequency sample

fish no. i	length (cm) x(i)		deviation from mean $x(i)-\bar{x}$	square of deviation from mean $(x(i)-\bar{x})^2$
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 $= \Sigma\, x(i)$		0.00 $= \Sigma(x(i)-\bar{x})$	121.48 $= \Sigma(x(i)-\bar{x})^2$

mean length, $\bar{x}$	:	406.8/27 = 15.07
variance, s^2	:	121.48/(27-1) = 4.67
standard deviation, s	:	$\sqrt{4.67}$ = 2.16
relative standard deviation, $s/\bar{x}$	:	2.16/15.07 = 0.14
standard error, $s/\sqrt{n}$	:	2.16/$\sqrt{27}$ = 0.41

(The concept of standard error is introduced in Section 2.3)

Table 2.1.2 **Mean and variance from a length-frequency sample. (The sample is derived from Table 2.1.1 with a class interval, dL of 1 cm)**

index j	interval (cm) L(j)-L(j)+dL	midpoint (cm) $\bar{L}(j)$	fre-quency 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

mean length, $\bar{x}$	:	407/27	= 15.074, say 15.07
variance, s^2	:	125.86/26	= 4.84
standard deviation, s	:	√4.84	= 2.20
relative standard deviation, $s/\bar{x}$	:	2.20/15.07	= 0.15

Let j be the index of a length group, and let the lower and upper class limit of length group no. j be denoted by respectively:

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

$$\text{or } L(j+1) = L(j) + dL$$

where dL is the "*interval size*". A fish of length x(j) then belongs to length group j when

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

Let F(j) be the frequency of length group j, that is the number of fish observed in length group j. Let $\bar{L}(j) = L(j) + dL/2$ be the midpoint of length group no. j. The calculation of mean value and variance from a frequency table is then performed in the usual way using midpoints to represent the intervals:

$n = \sum_{j=i}^{m} F(j)$ is the total number of observations, where m is the number of length groups,

$\bar{x} = \frac{1}{n} * \sum_{j=1}^{m} F(j)*\bar{L}(j)$ is the mean value and

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

The calculation procedure is shown in Table 2.1.2. The class midpoint $\bar{L}(j)$, and the square of the deviations from the mean are weighted by the number of fish in each class, i.e. the frequency, F(j). The results of Table 2.1.2 deviate slightly from those of Table 2.1.1 because a representation in cm groups produces less precise results than a representation in mm groups.

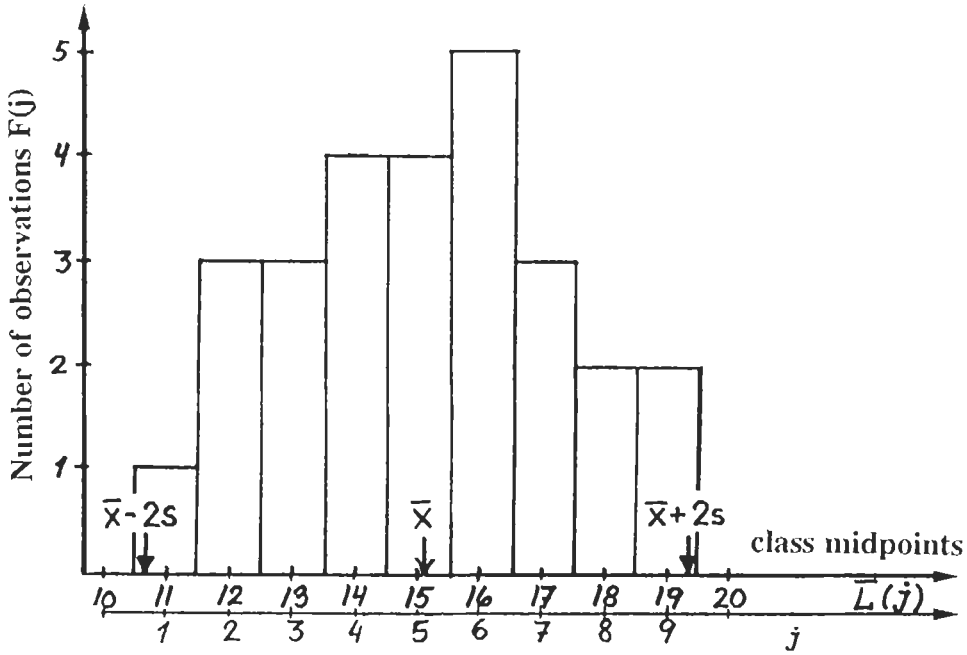


Fig. 2.1.1 **Length-frequency diagram. Graphical representation of the length-frequency sample from Table 2.1.2**

Fig. 2.1.1 shows a graphical representation of the frequency sample. Note that all observations lie in the interval from

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

For the so-called normal distribution (discussed in the next section) we expect about 95% of the observations to be contained in that interval.

(See **Exercise(s)** in Part 2.)

2.2 THE NORMAL DISTRIBUTION

Table 2.1.2 and Fig. 2.1.1 show an example of a small set of length-frequency data that approximately follows the so-called "*normal distribution*". The mathematical expression for a normal distribution is:

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

where Fc = "*calculated frequency*" or "*theoretical frequency*", n = number of observations, dL = interval size, s = standard deviation, $\bar{x}$ = mean length and π = 3.14159.

Using the values n = 27, dL = 1 cm, s = 2.20, $\bar{x}$ = 15.07 cm from Table 2.1.2 we get:

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

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

The values of Fc for a number of different x-values are given in Table 2.2.1. Note that the notation is slightly modified as we now use the interval midpoint, x, as the argument in Fc instead of the interval index, j, as used for argument in F in Table 2.1.2.

Table 2.2.1 Theoretical frequencies corresponding to Table 2.1.2, where x is the class midpoint

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

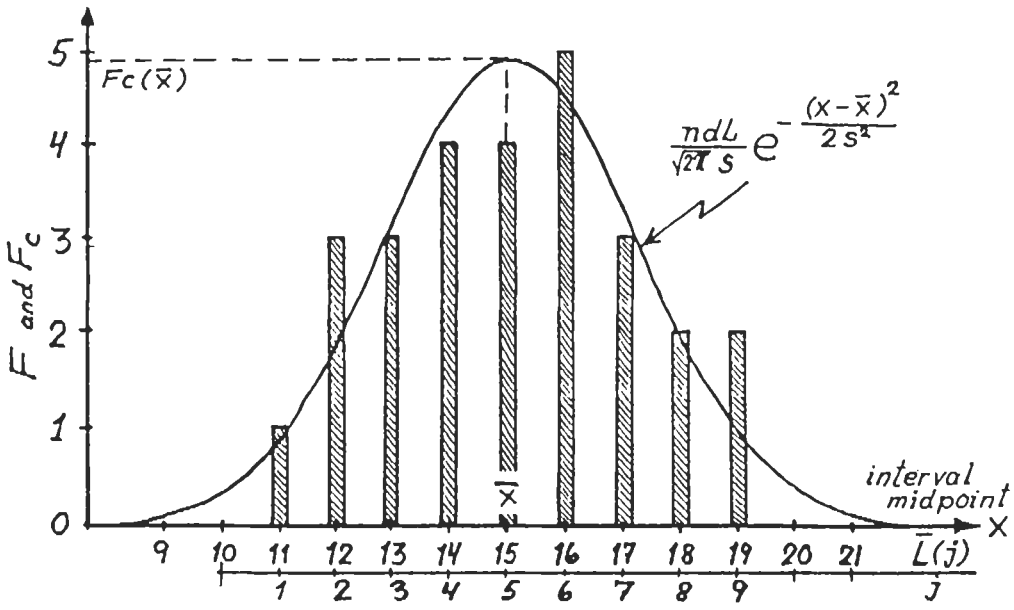


Fig. 2.2.1 The theoretical frequency, Fc, (the normal distribution curve) and the observed frequencies, F, (bars)

Fig. 2.2.1 shows the theoretical frequencies together with the bar diagram for F(j) from Fig. 2.1.1. As can be seen, Fc(x) gives a fair fit to the observed length-frequencies. This picture is often observed when recording length-frequencies of fish originating from one cohort, i.e. fish of the same age.

The normal distribution is observed in a great variety of different cases - hence the name. There are other types of probability distributions observed in fishery science. Examples are the "log-normal distribution", the "negative binomial distribution" and the "delta distribution". A conspicuous difference between these and the normal distribution is that they are skewed, whereas the normal distribution is symmetric. The delta distribution for example, is used to describe the probability distribution for the catch per hour by a trawl. It is composed of a log-normal distribution, which describes the distribution of the non-zero trawl catches and a special probability for zero catch (see Section 13.7, Fig. 13.7.2).

Perhaps the most important feature about normal distributions has to do with mean values. If you take, say, 50 random samples out of a certain population each of, say, 25 single observations, the 50 mean values will be (approximately) normally distributed. Thus, a mean value has a probability distribution. A mean value of any set of observations, is (approximately) normally distributed. This result is also valid for the mean values of log-normal distributions, delta distributions or any other type of distribution. This means that the mean values of all distributions observed in fishery biology are approximately normally distributed.

If we divide both sides of Eq. 2.2.1 by n (= sample size) we get:

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

the new found values, Fc(x)/n, will add up to nearly 1.0. Each value indicates the probability that a randomly drawn fish will belong to the corresponding length interval. That is, they can be interpreted as the probability of a randomly drawn fish to belong to the length interval from x-dL/2 to x+dL/2.

For the nine length intervals of Table 2.2.1 we find:

j	interval	probability
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

Thus for example, the odds are 181 to 1000 that a randomly drawn fish will be of a length between 14.5 and 15.5 cm. If we had included all length intervals and not only the nine for which we had observations, the probabilities would have added up to 1.000.

The normal distribution will be used in length-frequency analyses in the following chapters, because the length distribution of a single cohort of fish can be described by a normal distribution. As an introduction we shall study some of its aspects.

The procedures to calculate the mean and the standard deviation (Table 2.1.2) can be performed on any length-frequency data set. However, if for some reason, the observed frequency diagram does not represent the entire distribution, then the obtained values (from Eqs. 2.1.1 and 2.1.2) for sample mean and variance will be biased, i.e. the sample mean and variance may have no relation to the population mean and variance. The concept of "*bias*" will be further discussed in Section 7.1. If, for example, only the frequencies in the length interval from 10 to 15 cm are available (i.e. only the data for the left hand side) we are in a situation where Eq. 2.1.1 (mean value) and Eq. 2.1.2 (variance) do not represent the population. As will appear in Chapter 3 this is often the case when analyzing length-frequencies. However, there are a number of methods to overcome the problem.

(See **Exercises(s)** in Part 2.)

2.3 CONFIDENCE LIMITS

In this section we shall also use the example of a length composition sample of fish from one cohort. We have estimated the mean length of the cohort, $\bar{x}$, from the sample. Such an estimate is usually different from the true population mean, the mean we would have obtained if all fish of that cohort in the sea had been measured. Usually the true mean length is unknown. If we were dealing with a population of cultured fish in a pond we might be able to measure the true mean length of that population, but for a wild fish stock it is impossible to measure the true value of any parameter. In practice this also applies to the population of fish caught in a fishery, since we will not be in a position to measure all fish caught. We shall deal with the precision of the estimate of the mean length, in other words how great the deviation between the estimate and the true mean is likely to be. This uncertainty about the

true mean is expressed by the "*confidence limits*". In the case of a normal distribution, the lower and upper confidence limits are given by respectively:

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

where n is the sample size, s the standard deviation and t_{n-1} the so-called fractiles in the "*t-distribution*" or "*Student's distribution*" (Table 2.3.1). The argument "f" in the t-distribution (Table 2.3.1) is called the "*number of degrees of freedom*". In general the number of degrees of freedom is the number of observations minus the number of parameters. In this case $\bar{x}$ is the only parameter, so f = n-1 and $t_f = t_{n-1}$ (see Table 2.3.1).

The confidence limits can be calculated at different levels of precision, usually 90%, 95% and 99%, as indicated in Table 2.3.1. The higher the level (percentage), the higher the fractiles and therefore the wider the interval between the lower and upper limits.

Returning to the example given in Section 2.1 (Table 2.1.2) we want to calculate, for example, the 95% confidence limits for the mean length of fish in the population from which the sample was drawn. We use the 95% fractile of the t-distribution (Table 2.3.1) with n-1 = 26 degrees of freedom and insert into Eq. 2.3.1:

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

the 95% confidence limits are:

$$\text{lower limit: } \bar{x} - 0.87 = 15.07 - 0.87 = 14.20$$
$$\text{upper limit: } \bar{x} + 0.87 = 15.07 + 0.87 = 15.94$$

Thus, we are "95% confident" that the true mean length lies somewhere between 14.20 and 15.94, or in other words, if sampling was repeated 100 times under the same conditions we would expect the means to lie 95 times between 14.20 and 15.94. The interval between the lower limit and the upper limit is called the "*confidence interval*".

Table 2.3.1 Fractiles of the t-distribution (Student's distribution) *)

degrees of freedom f	fractiles 90% t_f	95% t_f	99% t_f	degrees of freedom f	fractiles 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

*) The use of the letter t in this context is universal. In this manual t is also used to represent the age of a fish. This table has been repeated on the last page of this volume for easy reference

For the example used above the confidence intervals at the 90% and 99% levels are respectively [14.35,15.79] and [13.89,16.25], of which the first is narrower and the second wider than the 95% interval.

The quantity $s/\sqrt{n}$ is the standard deviation of the estimate of the mean length (also called the "*standard error*") so that $\bar{x}$ has the variance (see Table 2.1.1):

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

Thus, the larger the sample, the more precise is the estimate of $\bar{x}$ (this subject will be discussed further in Section 2.8).

Eq. 2.3.2 follows from two general rules for random variables which are applied repeatedly in this manual. They are:

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

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

where C is a constant. For instance, when the variance of x is s^2 then the variance of 3x is $9s^2$; or, when the original observations are summed three by three, then the variance of $x_1+x_2+x_3$ is $3*s^2$.

The above statements about confidence limits apply only to "unbiased" estimates of the mean value. In cases when samples are biased, no matter how many fish we sample and measure we shall always get estimates of the mean value which are different from the true mean value.

Suppose we want to estimate the mean length of a certain fish species actually caught in a commercial fishery (note: fish caught are the fish landed plus the fish discarded at sea). Thus, if we sample only from the landings, and not the fish, usually below a certain size, which are discarded at sea, we get a biased estimate of the mean length of the fish caught. The mean length of the catch will be over-estimated, no matter how many fish we sample at the landing site. We can only get an unbiased estimate of the mean length of the fish that has been landed.

(See **Exercise(s)** in Part 2.)

2.4 ORDINARY LINEAR REGRESSION ANALYSIS

This method is used when we want to describe the variation of one quantity, e.g., the body depth of a fish, as a linear function of another quantity, e.g., the total length. The theory requires that the quantity on the horizontal axis (the independent variable) is measured with absolute precision. The method is often applied, however, when this requirement is violated. The effect of the inaccuracy of the values of the independent variable is that the slope of the line becomes flatter (closer to zero).

Suppose we have measured both the total length and the body depth of a sample of 7 fish.

Table 2.4.1 shows the total lengths, x(i), and the corresponding body depths, y(i), i = 1,2,...,7.

Table 2.4.1 **Sample of total lengths, x, and corresponding body depths, y**

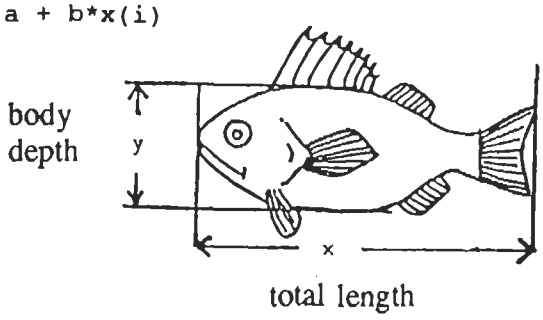
i	1	2	3	4	5	6	7
x(i)	11.2	12.4	13.5	15.7	17.1	18.5	19.0
y(i)	3.0	3.2	4.0	4.8	4.8	4.9	5.6

As can be expected, the body depth tends to increase when the total length increases. If the body proportions of a fish would remain constant for all sizes, its body depth would be directly proportional to its length, and this could be described by the model:

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

where b is a constant, also called a "*parameter*". The plot of this model always passes through the origin, the point where the x-axis and y-axis meet. We may allow for a deviation from proportionality between x and y by introducing a second parameter, a, and use instead of Eq. 2.4.1 the model:

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



where a indicates the intercept with the y-axis of the line that fits to the points. Fig. 2.4.1 shows the "*plot*" (or the "*scatter diagram*") of y(i) against x(i).

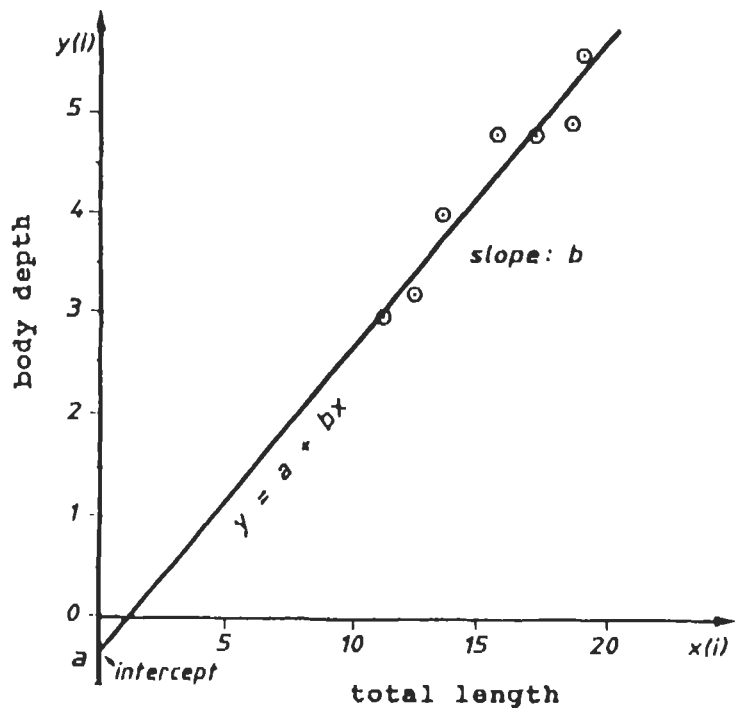


Fig. 2.4.1 **Scatter diagram of body depth (y) against total length (x), also called the "plot of y on x"**

An implication of Eq. 2.4.2 is that a fish of zero length has depth a, which makes no sense except when a is zero. However, if only lengths in a certain range are considered (e.g. only lengths above 5 cm), the two-parameter model may give a better fit to the observations than the one-parameter model, because the assumption of proportionality between length and depth is not strictly fulfilled.

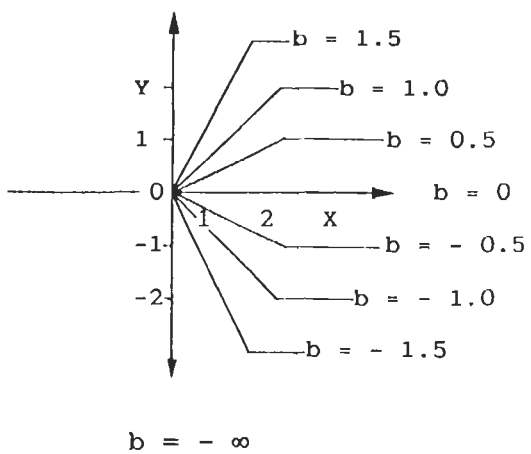
The mathematical model of Eq. 2.4.2 is called a "*linear model*" because pairs (x,y) which conform to the model, lie on a straight line. With a = -0.32 and b = 0.30 we get the straight line shown in Fig. 2.4.1. With these values of a and b the line in Fig. 2.4.1 fits well to the observed pairs of (x,y).

We shall now look into the problem of determining the line, i.e. how to estimate the parameters a and b. Just as we did for the mean value (cf. Section 2.3) we shall also show how the confidence limits of the estimates of a and b are calculated. This procedure is called "*ordinary linear regression analysis*". This method is probably the most commonly used statistical technique in fishery biology. There are special names for the parameters: a is called the "*intercept*" and b is called the "*slope*". The intercept is the distance from the point (0,0) in the (x,y) diagram to the point where the "*regression line*"

```
y = a + b*x
```

intersects with the y-axis (see Fig. 2.4.1).

The slope, b, indicates how steep the line is. If b = 0 the line is parallel to the x-axis. If b is positive the slope is ascending. If b is negative the slope is descending.



The variable on the horizontal axis, x, is called the "*independent variable*" and the variable on the vertical axis, y, is called the "*dependent variable*". The regression line is determined as the line which minimizes the sum of squares of deviations between the line y = a + b*x and the pairs of observations, (x(i),y(i)). We say that a and b are estimated by the "*least squares method*", i.e. we search for the values of a and b which minimize:

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

where n is the number of pairs of observations (n = 7 in the example). The deviations between the line and the observations are illustrated in Fig. 2.4.2. The assumption behind the regression analysis is that each y(i) is normally distributed with mean value a + b*x(i), and with a constant variance, i.e., a variance which is not dependent on the value of x(i). The following formula to estimate this common variance differs only slightly from the one introduced in Section 2.1. The so-called "*variance about the regression line*" is:

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

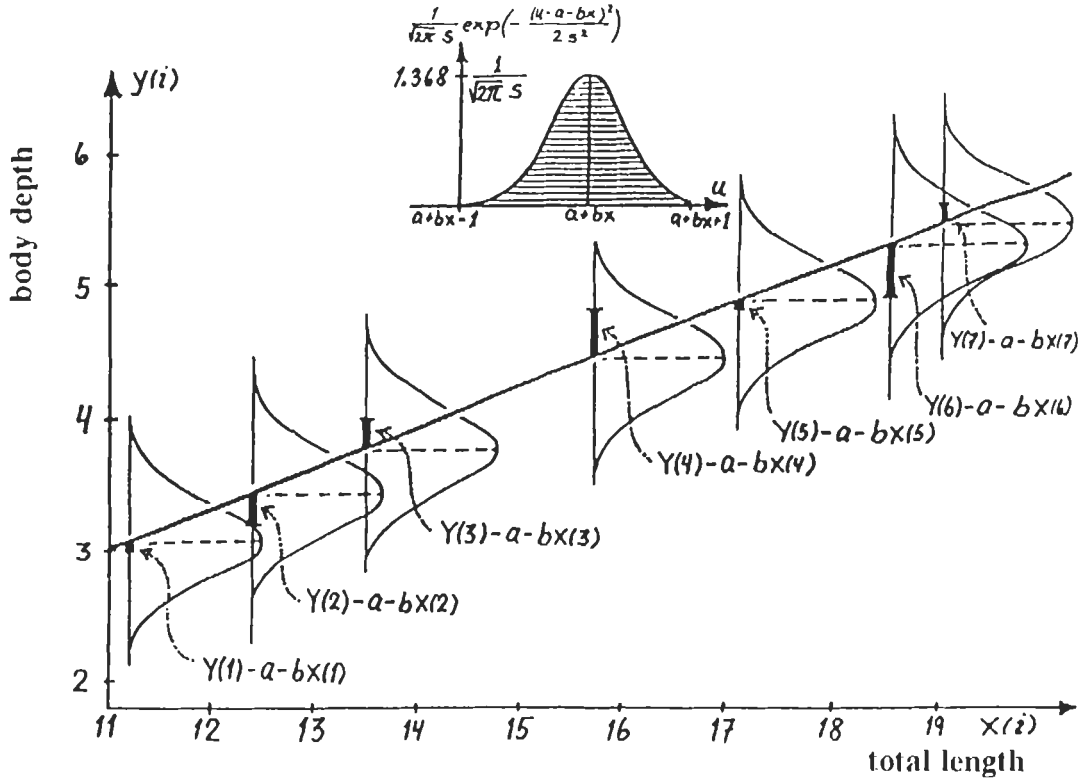


Fig. 2.4.2 **Illustration of the assumptions behind ordinary linear regression analysis. Each y(i) for a given x(i) is normally distributed with a common variance**

There are n-2 degrees of freedom (the number by which the sum is divided) because we have two parameters, a and b.

Estimates of the parameters a (intercept) and b (slope) are obtained by:

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

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

where $\bar{y}$ and $\bar{x}$ are the mean values of y and x as defined by Eq. 2.1.1.

In Table 2.4.2 the calculation procedures to estimate a and b are demonstra-ted using the data from Table 2.4.1. Thus, the estimated regression line becomes:

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

To calculate the confidence limits of a and b we need the sum of squares of deviations of x and y. The variances of x and y are defined by Eq. 2.1.3 as follows:

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

and a similar expression for s_y^2. For use in the next section we introduce the "*covariance*":

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

Table 2.4.2 The calculation procedure for ordinary linear regression analysis. Results marked by #) are not used in the calculation of a and b, but are derived here for subsequent use

i	total length x(i)	$x(i)^2$	body depth 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)

$\bar{x} = 15.343$

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

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

$s_x^2 = 9.296$ #)

$s_x = 3.049$ #)

$\bar{y} = 4.329$

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

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

$s_y^2 = 0.922$ #)

$s_y = 0.960$ #)

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

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

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

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

Table 2.4.3 Calculation of variance about the line from 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
$s^2 = 0.4252/(7-2) = 0.085$				sum: 0.4252

The procedure for the calculation of variance about the regression line leading to Eq. 2.4.4 is demonstrated in Table 2.4.3. However, the variance about the line can be obtained more easily from sy and sx:

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

Given the results from Table 2.4.2, Eq. 2.4.10 becomes:

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

The variances of the estimates of b and a are:

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

and

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

Given the results from Table 2.4.2 we get:

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

$$sa^2 = 0.00147*\left[\frac{7-1}{7}*9.297+15.343^2\right] = 0.3578, \quad sa = 0.598$$

The confidence limits for the intercept a and the slope b are respectively:

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

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

The 95% confidence limits of a and b for the example with n = 7 fish and $t_{7-2} = 2.57$ (Table 2.3.1) become:

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

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

Note that the confidence interval for the intercept a contains zero. This means that the hypothesis that body depth is directly proportional to length, (thus that "a = 0") cannot be rejected by the 95% confidence limits. We say that a is not significantly different from 0 at the 95% level.

If we have a good reason to assume that a = 0 then the estimated value should be replaced by 0 if the estimate is not significantly different from 0. Then, however, b must be recalculated as follows:

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

Our present estimate is based on only seven fish. If we had measured 200 fish the estimate of the standard deviation, sa, would be smaller (cf. Eqs. 2.4.11 and 2.4.12). Let us assume for example, that $\bar{x}$, $\bar{y}$, sx, sy, a and b were the same for a sample size of n = 200 as those estimated for a sample size of n = 7 (which might well happen). Although the estimates of a and b turn out to have the same value, their standard deviations, sa and sb, will be different.

With n = 200 Eq. 2.4.11 gives sb = 0.006098, while Eq. 2.4.12 gives sa = 0.0091 and t_{198} = 1.97 (Table 2.3.1). Thus, sa and sb become smaller, and consequently the confidence interval of a becomes smaller:

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

The estimate of a would now be significantly different from 0. In that case we can conclude that the odds are less than 5% that the true value of a is larger than -0.30 or smaller than -0.33.

(See **Exercise(s)** in Part 2.)

2.5 THE CORRELATION COEFFICIENT AND FUNCTIONAL REGRESSION

The "*correlation coefficient*", r, is a measure of the linear association between two quantities, both of which are subject to random variation. The total length and body depth sample from Section 2.4 is an example of two such quantities. In this case seven fish were drawn at random. By accident we could have drawn seven fish all of (nearly) the same length. In that case the sample would not be suitable for estimation of the length/depth relationship because the confidence limits of a and b would become very wide.

The correlation coefficient can be used only when both measurements are allowed to vary randomly. If we had selected seven fish with predetermined lengths rather than random lengths (e.g. had selected the lengths 4, 6, 8, 10, 12, 14 and 16 cm for the length/depth sample) the calculation of a correlation coefficient for this sample would be incorrect.

The correlation coefficient is defined as:

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

where sxy is defined by Eq. 2.4.9 and sx and sy are defined by Eq. 2.4.8.

Inserting the slope ($b = sxy/sx^2$), Eq. 2.5.1 becomes:

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

The range of r is: $-1.0 \leq r \leq 1.0$. r is negative if y tends to decrease when x increases and r is positive if y tends to increase when x increases. This statement also holds for the slope b and it follows from Eq. 2.5.2: As sx/sy is always positive (cf. the definition Eq. 2.4.8) r has the same sign as the slope b. The extreme cases, r = 1 or r = -1 occur when all pairs (x,y) lie exactly on a straight line. The closer r approaches zero the less pronounced is the linear association between y and x. When r = 0, x and y are independent of each other.

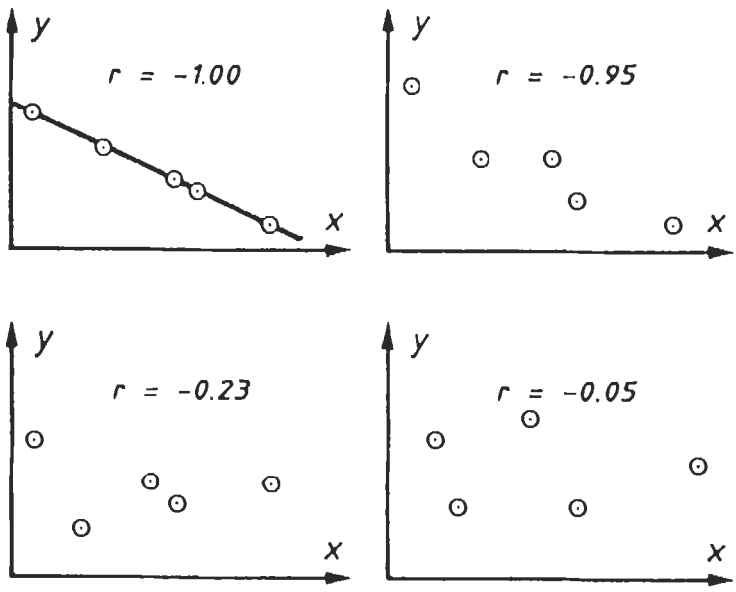


Fig. 2.5.1 Examples of correlation coefficients

Fig. 2.5.1 shows four examples of scatter diagrams with different values of r. For the example of Table 2.4.2 we get:

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

Let us call r1 (lower) and r2 (upper) the 95% confidence limits for r. They can be calculated from the expressions:

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

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

where "tanh" is the "hyperbolic tangent" which is standard on many scientific pocket calculators.

With r from the example (r = 0.961, n = 7) the 95% confidence limits become: [r1,r2] = [0.75,0.99]. The 99% confidence limits can be obtained by replacing the number 1.96 by 2.58 in Eq. 2.5.3.

Often we are interested in knowing whether zero lies in the confidence interval, *viz.* what the odds are that the linear association is due to chance. In this example the odds are less than 5% that the linear association is due to chance, because zero is not in the confidence interval.

In the example of regressing body depth on total length, the length was chosen as the independent variable and the body depth as the dependent variable. However, there is no special reason for this choice. Our sample consists of seven randomly chosen fish. We did not control what their lengths and body depths would be, thus, we could as well have made the opposite choice for dependent and independent variables.

One of the assumptions behind linear regression analysis is that the independent variable cannot be a random variable. The independent variable must be something of which we can determine the values beforehand. For example, if the independent variable is the time the sample is taken, it can be determined beforehand. We could decide to collect a sample on the first day of each month. If we measure time in units of years and start with time zero on the first of January, the independent variable would take the values: 0, 1/12, 2/12, 3/12 ... etc. These values are clearly not random variables.

In the case of the seven fish chosen at random in the example above the situation is that because they were chosen at random out of a normal distribution of fish lengths a correlation analysis can be performed on them. On the other hand, we are able to decide on the lengths beforehand. We could choose the four smallest fish and the three biggest. We could also decide, as we did, to take them as they come. Only in the latter case is it permissible to do both kinds of analysis. In the former case only regression analysis will do. On the other hand, it would probably be a more effective way of doing the regression analysis because of the great distance between the observations on the horizontal axis. This would cause a small variance of the slope. Choosing the fish at random, most of them are likely to be medium-sized and contribute little to the determination of the slope which might show a large variance.

Another question is whether we would have obtained a different result using the body depth as the independent variable, thus plotting fish length as a function of body depth. First it must be considered whether depth is as precisely measured as length. If it is not, the slope would be biased (flattened) as already mentioned. However, there are problems even if the two variables are measured with the same accuracy.

Taking now the body depth as the independent variable we get what is called an "*inverse regression*". Only in the exceptional case that all observations lie on the regression line (i.e. if r = 1 or r = -1) the same result would be obtained for the inverse regression as for the ordinary regression. The equation y = a + b*x (Eq. 2.4.2) is mathematically equivalent to:

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

or $x = A + B*y$ where $A = -a/b$ and $B = 1/b$ (2.5.4)

Carrying out the inverse regression (Eq. 2.5.4) we find that

$$A = 2.139 \text{ and } B = 3.05$$

The equation: x = 2.139 + 3.05*y can be converted into:

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

which can be compared with the result found for the original regression (Eq. 2.4.7: y = -0.315 + 0.303*x). Thus, the inverse regression gives a result that differs from that of the original regression analysis.

One way to circumvent the problem of choosing the independent variable when both variables are random variables is to use the so-called "*functional regression analysis*" (see Ricker, 1973). This method estimates a slope (which we call b' to distinguish it from slope b of the ordinary regression analysis) by the expressions:

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

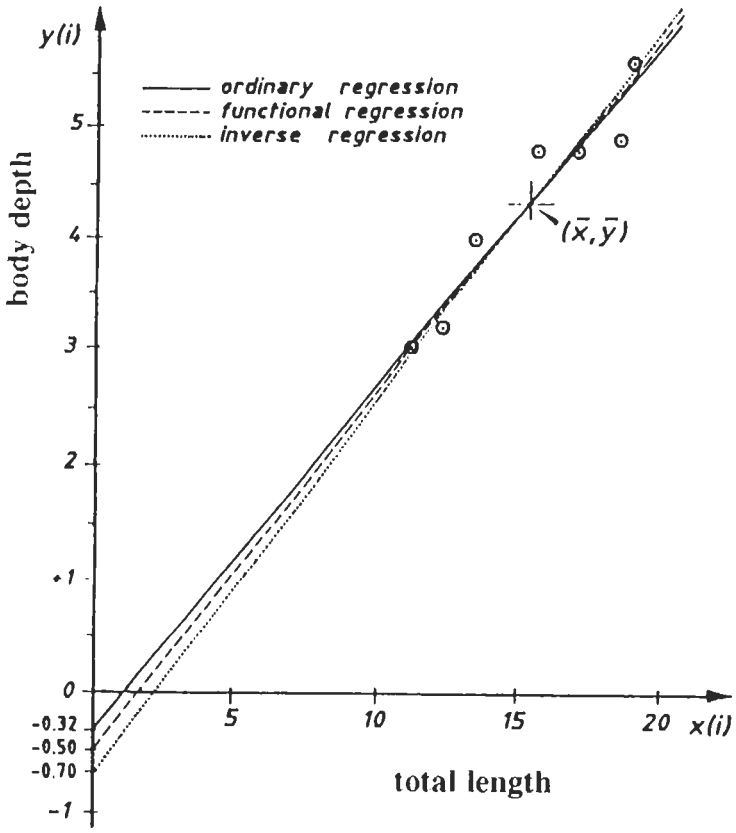


Fig. 2.5.2 **Functional and inverse regression lines compared to the original regression line**

and the intercept:

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

This type of analysis gives a result that may be considered a compromise between the original ordinary regression and its inverse counterpart.

With the results from Table 2.4.2 we get:

$b' = 0.960/3.049 = 0.315$ and $a' = 4.329 - 0.315*15.343 = -0.504$

and $y = -0.504 + 0.315*x$

Functional regression analysis is mentioned here for the sake of completeness. There are some rather intricate limitations to its applicability which we cannot go into here.

The following three regression lines have now been estimated:

1. Original ordinary regression analysis: $y = -0.315 + 0.303*x$
2. Functional regression analysis: $y = -0.504 + 0.315*x$
3. Inverse ordinary regression analysis: $y = -0.701 + 0.328*x$

Fig. 2.5.2 shows the three regression lines. Note that all three lines pass through the point $(\bar{x},\bar{y})$ and that an increase in slope is partly balanced by a decrease of the intercept.

(See **Exercise(s)** in Part 2.)

2.6 LINEAR TRANSFORMATIONS

Linear functions are mathematically easy and also have the advantage that they can be graphically interpreted without any problem. However, many functional relationships observed in fishery biology are not linear. Fortunately, such non-linear functions can often be transformed into linear functions, which means that after transformation they can be dealt with in the way described in the foregoing sections. Several examples are given below of the application of transformations from non-linear functions to linear functions in fishery biology.

Example 1: Length-weight relationship

Here we consider a famous example, namely the functional relationship between total length and body weight of fish. Fig. 2.6.1 shows a plot of weight on length of the threadfin bream, *Nemipterus marginatus*. Clearly, this is not a linear relationship. The curve in Fig. 2.6.1 is of the function:

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

where W(i) is the body weight of fish no. i, L(i) is the total length and q and b are parameters. Eq. 2.6.1 is usually called the "*length-weight relationship*". It can be transformed into a linear equation by taking logarithms on both sides:

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

or

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

where

$$y(i) = \ln W(i),\ x(i) = \ln L(i) \quad \text{and} \quad a = \ln q.$$

With Eq. 2.6.2a we are now in a position to carry out the estimation of a and b by linear regression analysis. Input data are shown in Table 2.6.1 and the corresponding scatter diagram in Fig. 2.6.2. The results are:

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

$$\bar{x} = 2.727 \quad \text{and} \quad \bar{y} = 3.799$$

Since a = ln q we can obtain q of the original length-weight relationship (Eq. 2.6.1) by taking the antilog of a:

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

Thus, the estimated relationship between W (in g) and L (in cm) becomes:

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

(The back-transformation from logarithms introduces a bias which we will not go into here.)

Table 2.6.1 **Data for estimation of a length-weight relationship for the threadfin bream (*Nemipterus marginatus*) from the South China Sea (from 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
		sum	43.629	60.786
		mean	2.7268	3.7991
		sx and sy	0.3311	1.0161

We can also calculate the 95% confidence limits of b, using the values of sx, sy, n and t_{14} (see Table 2.3.1) in Eq. 2.4.11:

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

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

The 95% confidence interval for b is [(3.057-0.155),(3.057+0.155)] or [2.90, 3.21]. These confidence limits tell us that only the first decimal in the estimate of b is significant (compare Section 2.3), thus the true value of b could just as well be 3.0.

Since the weight of a fish (in grammes) is approximately equal to its volume (in cubic cm), and since its volume is often proportional to the cube of its length, L^3, we would expect that the value of b in Eqs. 2.6.1 and 2.6.2 is close to 3.0.

Since the confidence interval calculated above supports this hypothesis we can simplify the length-weight relationship by replacing the estimate b = 3.057 by b = 3.0. This implies that a new estimate of the intercept a has to be obtained. Since the new straight line with b = 3.0 also passes through the point $(\bar{x},\bar{y})$ we can calculate the new intercept a using Eq. 2.6.2a:

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

From a we obtain the corresponding new value for q

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

Thus, the new relationship becomes:

$$W = 0.0125*L^3$$

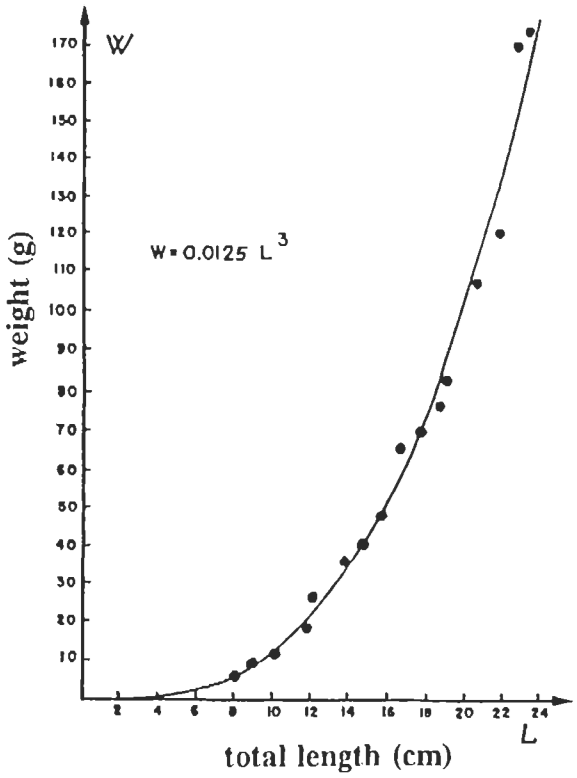


Fig. 2.6.1 **Length-weight relationship of *Nemipterus marginatus* in the South China Sea. (Based on data from Table 2.6.1)**

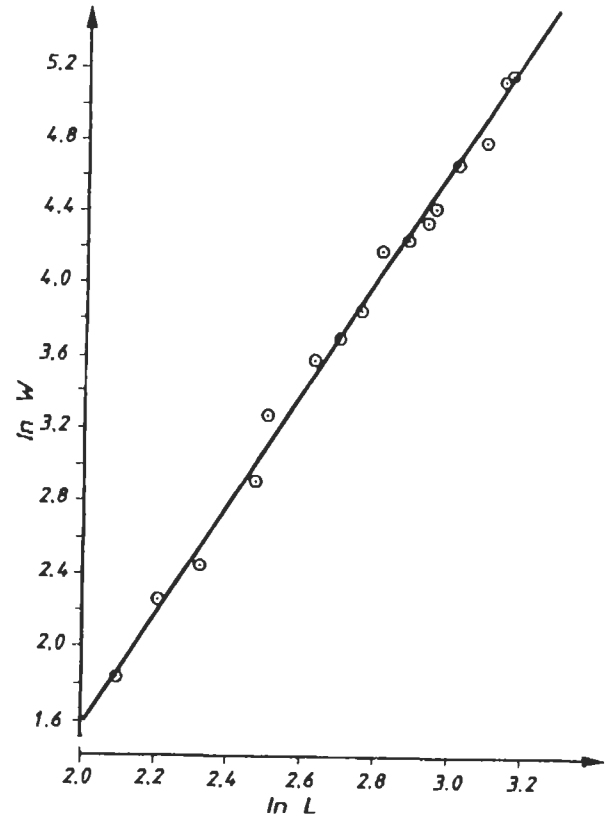


Fig. 2.6.2 **The data from Fig. 2.6.1 converted to natural logarithms**

Example 2: Linearization of a normal distribution

In Section 2.2 (Eq. 2.2.1) the mathematical expression for a normal distribution is given as:

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

This equation can be transformed into a linear regression in the following two stages:

Stage 1: ***Converting a normal distribution into a parabola***

Taking the logarithms on both sides of Eq. 2.2.1 gives:

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

Considering ln Fc(x) as the dependent variable, y, and x as the independent variable, we have hereby obtained a functional relationship between y and x, which can graphically be represented by a parabola which has the general formula:

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

Inserting the values used in the example of Table 2.1.2 we obtain:

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

the graph of which is shown in Fig. 2.6.3.

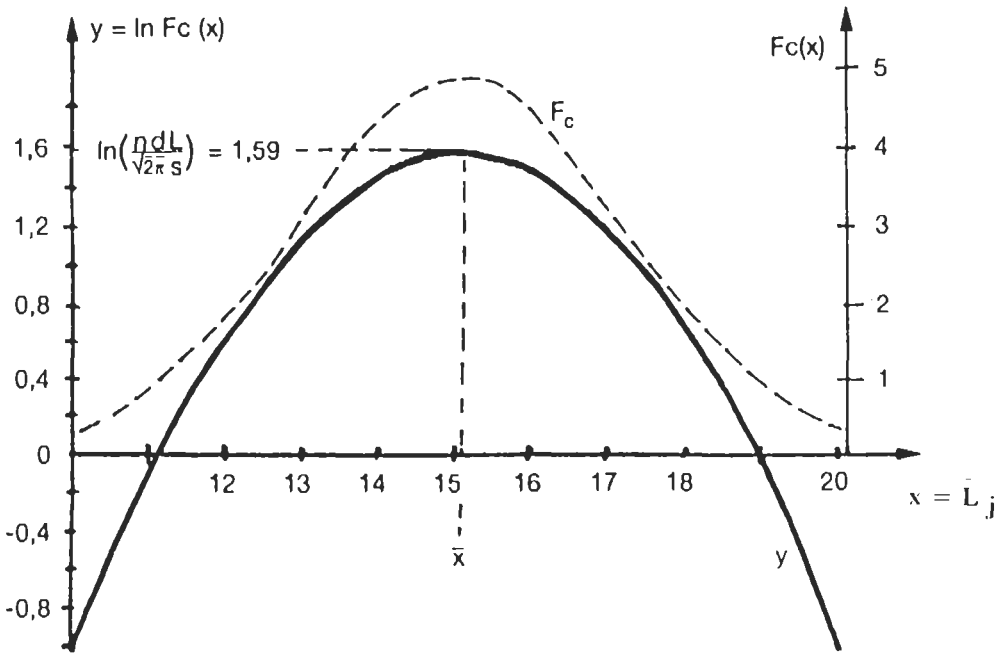


Fig. 2.6.3 **The ln-transformed normal distribution (y) together with the original distribution (Fc)**

Stage 2: *Converting a parabola into a straight line*

When dealing with a parabola differences between points evenly spaced on the x-axis are always found to be linear. Subtracting the functional value (in our case: ln Fc(x)) for the higher value of x from that of the lower value of x gives a series of differences which are positive for the left half of the parabola and negative for the right half. The process and the results of calculating differences are illustrated in Figs. 2.6.4aA and 2.6.4aB respectively.

To explain this process mathematically we introduce a new dependent variable, y', which is the difference between the logarithm of the number in a certain length class and the logarithm of the number in the preceding class.

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

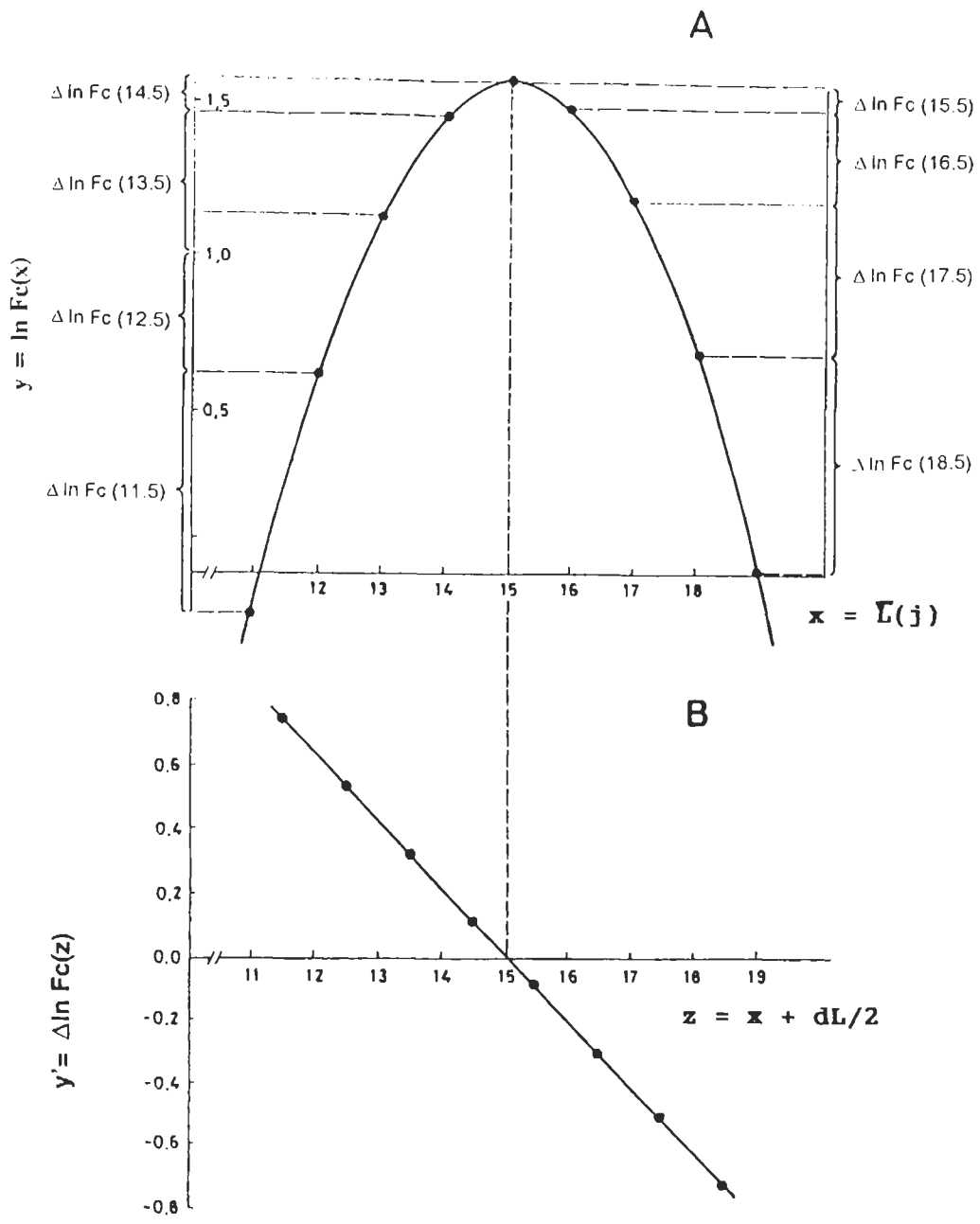


Fig. 2.6.4a **Estimation of mean and variance by Bhattacharya's method. A. The parabola and the differences between equidistant points on the x-axis. B. The Bhattacharya plot of differences against class midpoints. Data in Table 2.6.2**

This can also be expressed as

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

where Δ (delta) designates a "small" difference between two function values. y' is to be plotted against a new independent variable, z, which is equivalent to x plus half the length interval:

$$z = x + dL/2$$

We now have to insert Eq. 2.6.3 into Eq. 2.6.4 as follows:

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

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

After squaring and summing this can be converted into a relatively simple equation:

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

or $y' = a + b*z$, where $z = x + dL/2$

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

From the slope, b, and the intercept, a, we get the variance and mean value, respectively, by:

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

and

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

This regression is one of the main elements of the method described by Bhattacharya (1967) for separating two or more normal distributions (Section 3.4.1). We call it the "*Bhattacharya plot*". Table 2.6.2 and Fig. 2.6.4a show an example. In this case the theoretical frequencies, Fc, and the class midpoints, x, from Table 2.2.1 have been used as "observations". These conform exactly to the model. In this case the mean and variance estimated by the Bhattacharya plot are almost the same as those obtained by the traditional method (as in Table 2.1.2). Any small difference will be due to the introduction of a regression analysis. Part B of Fig. 2.6.4a shows the plot of the differences between the logarithms of two consecutive frequences against the *midpoints* of the x-values.

The Bhattacharya plot also gives a clue to the number of observations in a normal distribution of which only the frequencies in some size classes are known. Rewriting Eq. 2.2.1 with the actual observations we get

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

Thus n can be estimated even for a single size class j once $\bar{x}$ and s^2 have been estimated. Accidents of sampling, however, cause inaccuracy because they influence the number of fish in each class interval, cf. Fig. 2.2.1. When the numbers in several size classes are known the

frequencies can be summed smoothing the deviations from each of the expected frequencies. Summing for i classes on both sides of the equality sign and rearranging gives

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

Table 2.6.2 **Estimation of mean value and variance of a normal distribution from the Bhattacharya plot, illustrated by the theoretical frequencies, Fc(x), of Table 2.1.2, presented in Table 2.2.1. Details of the table are illustrated in Figs. 2.6.3 and 2.6.4a**

index j	$\bar{L}(j)$ (x)	interval 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	

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

Table 2.6.2a **Estimation of the total number of observations with the Bhattacharya method**

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
sums		15.11	3.0907

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

Table 2.6.3 Bhattacharya plot corresponding to the length-frequency sample of Table 2.1.2

index	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

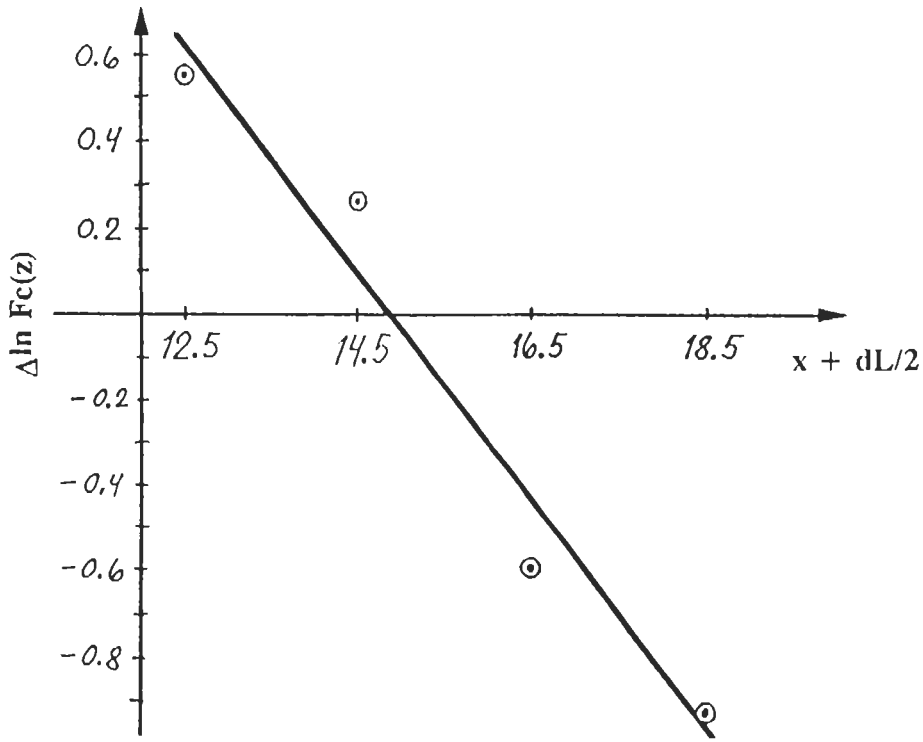


Fig. 2.6.5 Bhattacharya plot corresponding to Table 2.6.3

The observations for fish larger than $\bar{x}$ in Fig. 2.6.4a might not be reliable because their sizes overlap with the smaller fish of an older age group such as illustrated in Fig. 1.4.1, age groups 1 and 2. In that case we could use only the left hand observations of Fig. 2.6.4a (x = 11, 12, 13, 14, 15 cm) for the Bhattacharya plot which provides four points on the straight line from which to estimate $\bar{x}$ and s^2. We find with data from Table 2.6.2:

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

The result is practically the same as for the entire normal distribution because the fit of the straight line to the data is in this case almost perfect (see Fig. 2.6.2). The application of Eq. 2.6.9 is shown in Table 2.6.2a. We find n = 26.88 when the true value (known from Table 2.1.2) is 27.

Once n is known the numbers in each size class (the theoretical frequencies) can be estimated from Eq. 2.6.8. These calculations are not carried out in Table 2.6.2a because in this exercise the "observations" are actually the theoretical frequencies.

Table 2.6.3 shows the estimation of mean and variance by the Bhattacharya plot, but now with the actual observations given in Table 2.1.2. Because of the small sample size the observations have been grouped into 2 cm intervals. Fig. 2.6.5 shows the corresponding plot. The estimates of mean and variance obtained in Table 2.6.3 deviate from those calculated by the traditional method (Table 2.1.2) because of 1) the small sample size, 2) the bias introduced by large length intervals and 3) the use of a different statistical method (linear regression analysis).

(See **Exercise(s)** in Part 2.)