Acta Zool. Fennica 191:57–72
Helsinki 30 June 1992

ISBN 951-9481-36-2 ISSN 0001-7299

Fluctuating asymmetry as a measure of developmental stability: Implications of non-normal distributions and power of statistical tests

A. Richard Palmer & Curtis Strobeck

Palmer, A. R., Department of Zoology, University of Alberta, Edmonton, Alberta T6G 2E9 and Bamfield Marine Station, Bamfield, British Columbia V0R lB0 Canada
Strobeck, C., Department of Zoology, University of Alberta, Edmonton, Alberta T6G 2E9 Canada

Received 15 October 1989

Fluctuating asymmetry (FA) is a widely used measure of developmental stability. Nearly all FA indexes estimate the variance of the frequency distribution of right-minus-left (*R-L*) for a given bilateral character. Differences in these indexes among samples are usually interpreted as reflecting differences in the level of developmental stability. If developmental stability is the ability to correct for small, random developmental perturbations of exclusively environmental origin, then a distribution of *R-L*, which may include both genetically and environmentally caused asymmetries, may not be a good measure of developmental stability. *R-L* distributions that depart significantly from the statistical criteria for ideal FA (mean of zero, normal distribution) are unsuitable as descriptors of developmental stability because a fraction of the asymmetry variation may have a genetic basis. In addition, broad-peaked or bimodal (platykurtic) distributions of *R-L*, which reveal the presence of antisymmetry, also imply genetically based asymmetries and thus seem inappropriate as descriptors of developmental stability. Finally, both skewed distributions and narrow-peaked, long-tailed (leptokurtic) distributions may arise in mixed samples composed of two or more groups of individuals where each exhibits a different form of asymmetry and where one or more forms of asymmetry may have a genetic basis. Statistical techniques for detecting departures from normality and for detecting heterogeneity of variances among samples are briefly reviewed.

1. Introduction

1.1. Background

Fluctuating asymmetry (small, random departures from perfect symmetry) is commonly used as a measure of the developmental stability of bilaterally symmetrical morphological traits (Palmer & Strobeck 1986, Zakharov 1989). Typically, one or more indexes are calculated that express fluctuating asymmetry (FA) as a variance, or an average absolute value, of the difference between the right and left elements of a bilateral pair (*R-L*) for a sample of individuals (Palmer & Strobeck 1986). The larger the FA the lower the developmental stability.

Many studies have examined variation in FA in response to both intrinsic (genetic) and extrinsic (environmental) factors that might influence developmental stability (reviewed in: (Allendorf & Leary 1986, Palmer & Strobeck 1986, Zakharov 1989)). Several patterns have emerged more than once from these descriptive and experimental studies. Developmental stability appears to be reduced:

1) by increased homozygosity,
2) in hybrids between nominal species,
3) by extreme physical conditions, and
4) by pollution or declines in habitat quality.

Significantly, although some studies have found little or no association, none have found patterns opposite to these. Because many of these factors reducing developmental stability correlate with reduced fitness (Allendorf & Leary 1986), Zakharov (1989) has suggested that FA, perhaps together with other measures of developmental stability, may provide a convenient summary measure of overall population "condition" or fitness.

In this paper, we wish to draw attention to some potential difficulties associated with the *interpretation* of differences in empirical estimates of FA (see Palmer & Strobeck (1986) for a discussion of methodological issues). In particular, we wish to emphasize that certain shapes of frequency distributions of *R-L* may weaken or seriously compromise the use of asymmetry variation as a measure of developmental stability for a particular trait. These difficulties arise where the differences in *R-L* among individuals may not be due entirely to developmental noise, which by definition is exclusively non-genetic in origin. If differences in *R-L* among individuals arise in part from genes directly causing greater or lesser departures from symmetry, then differences in asymmetry variation among samples, *no matter how they are calculated*, can no longer be interpreted with confidence as differences in developmental stability.

1.2. Developmental noise, developmental stability, and forms of asymmetry: processes vs. patterns

To try to avoid ambiguity, we use the terms developmental noise and developmental stability as follows. Developmental noise represents the cumulative effects of small, random developmental perturbations or accidents that are exclusively environmental in origin (Waddington 1957, Lewontin 1983). Developmental stability, on the other hand, refers to the ability to correct for these perturbations (Mather 1953, Waddington 1957, Zakharov 1989). These terms are not the converse of each other because one (developmental noise) refers to an exclusively non-genetic phenomenon while the other (developmental stability) refers to a phenomenon that *does* appear to have a genetic basis. As we use them, both terms refer to specific developmental processes each of which may independently influence the frequency distribution of right-minus-left.

The three forms of bilateral asymmetry to which we draw attention — FA, directional asymmetry, and antisymmetry (VanValen 1962) — are merely descriptions of identifiable patterns of asymmetry variation. They refer to frequency distributions of right-minus-left variation with well defined statistical attributes (defined below). Although they may arise as a product of one or more developmental processes, we use these terms exclusively to refer to the shapes of distributions and not to the processes responsible for generating these shapes. This is an important distinction which sometimes is not clearly made in studies of asymmetry variation.

1.3. Fluctuating asymmetry as a measure of developmental stability

FA is used to estimate the level of developmental stability in bilaterally symmetrical organisms based upon three intuitively appealing propositions. First, in the absence of any intrinsic (thermal noise, random developmental accidents, genetic predispositions towards asymmetry) or extrinsic (environmental) perturbations, all individuals in a sample should be perfectly symmetrical. In other words, for bilaterally symmetrical traits the ideal state is clearly defined *a priori*. Note that for traits exhibiting directional asymmetry (DA) or antisymmetry, the ideal state is not clearly defined *a priori*.

Second, biological systems cannot consistently achieve perfect bilateral symmetry even under ideal environmental conditions. Minor inconsistencies during development will deflect developmental trajectories away from perfect symmetry (Waddington 1957). Hence, by chance, one limb will become slightly larger or smaller than its bilateral counterpart. If these deflections are random, independent, and cumulative for each member of a bilateral pair of characters, a frequency distribution of the difference between these right and left members for a sample of individuals should approximate a normal distribution with a mean of zero (ideal FA).

Third, biological systems have the capacity to correct for "accidents" during development. Hence the greater this capacity to correct back to the ideal developmental trajectory for a particular set of environmental conditions, the lower the variance of the difference between right and left members of a bilateral pair (FA).

2. Difficulties in interpretation

2.1. A key assumption

The use of departures from bilateral symmetry to estimate developmental stability depends critically upon one very important assumption: these departures, of a particular character on a particular individual, must not be due to any genetic predisposition towards asymmetry. In other words, regardless of what statistical measure is used to describe them, if departures from bilateral symmetry have arisen through the action of genes or developmental processes specifically directing one member of a bilateral pair to become larger than the other, then it seems to us that such departures may not be *interpreted* as evidence for reduced developmental stability.

2.2. Ideal fluctuating asymmetry

The conventional interpretation of FA as a measure of developmental stability arises as follows. Consider the wing lengths of a sample of adult fruit flies. For both the left and the right wing, a frequency distribution of lengths will exist (Fig. 1a, b). Some unknown fraction of this variation for each wing will be due to genetic differences among individuals (solid region) and the remainder will be due to the effects of environment on wing length (stippled region). On a given individual, of course, both the right wing and the left wing are a product of the same genome. Hence, if the wings are normally symmetrical, calculating the difference between the right and left wing cancels out any genetic effects on wing length (Fig. 1c). In addition, any effects on average wing length of the specific microenvironmental conditions experienced by an individual fly [the creode for those conditions (Waddington 1957)] will also cancel out. This frequency distribution of R-L is thus commonly interpreted as a product of non-genetic variation in symmetry (Fig. 1c). It has a parametric mean of zero and the variation is normally distributed about this mean. These two statistical criteria *define* ideal FA (VanValen 1962).

An important point we wish to emphasize here is that if either of these two statistical criteria are not met for a given distribution of R-L, one may no longer *interpret* the variation in R-L as a product of pure developmental noise. As a consequence, although one may compute one or more of the nine indexes commonly used to describe FA (Palmer & Strobeck 1986), differences in these computed values among samples for a given character may no longer be attributed exclusively to differences in developmental stability or developmental noise. Some unknown fraction of the variation may have a genetic basis. As with

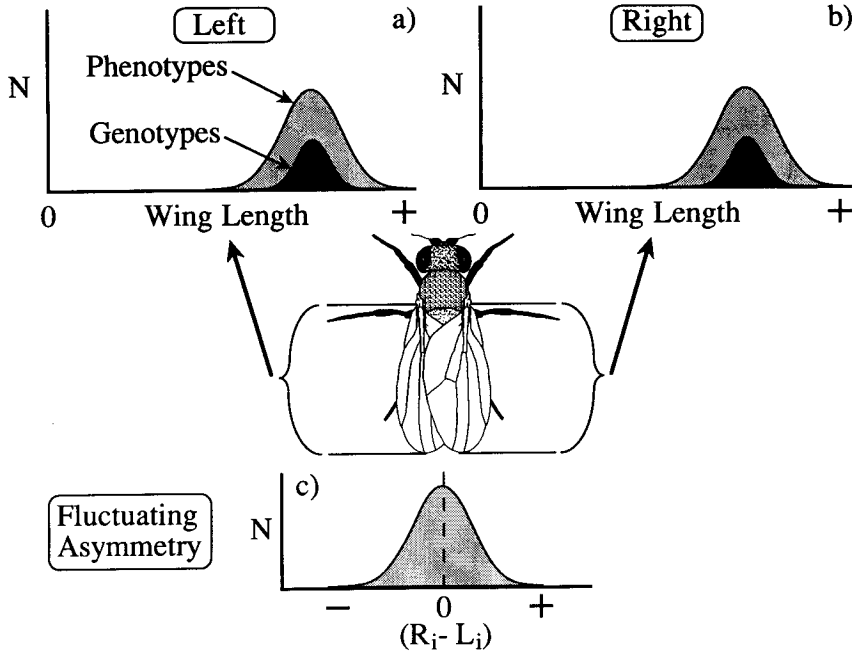

Fig. 1. A graphical illustration of variation in fruit fly wing lengths (a,b) and ideal fluctuating asymmetry (c), illustrating the conventional interpretation of fluctuating asymmetry. R_i = size of a trait on the right side, L_i = size of the bilaterally paired trait on the left side. Here and in later figures we use a convention to represent genetic and environmental components of bilateral variation that requires some explanation. Solid frequency distributions represent genetically induced variation, and stippled frequency distributions represent total phenotypic variation (including developmental noise). The exclusively stippled region under the curve is intended to represent environmentally induced variation only. Two aspects of these curves, however, are not technically correct. First, the area under each should be the same (the area under any frequency distribution is one). Second, if environmental variation is added to underlying genetic variation, the combined distribution should be broader with a lower peak. We use this heuristic convention a) because we wish to emphasize that some subset of the total phenotypic variation has a genetic basis, and b) because it avoids introducing potentially distracting elements to the figures.

any other morphological character exhibiting variation, one can no longer separate genetic from environmental effects. We attempt to buttress this argument in the next three sections.

2.3. Other "pure" forms of asymmetry

Frequency distributions of *R-L* will not always be normally distributed about a mean of zero (ideal FA, Fig. 2a). Two other types of frequency distributions of *R-L* (Fig. 2b, c) represent fundamentally different types of asymmetry (VanValen 1962): directional asymmetry (DA) and antisymmetry. For traits exhibiting DA (Fig. 2b), the variation of *R-L* is normally distributed about the mean, but the mean of the distribution departs significantly from zero.

Antisymmetry (Fig. 2c) presumably results from a genetic predisposition of individuals towards asymmetry but, within a given sample, some individuals develop a left bias while others develop a right bias. Extreme forms of antisymmetry are clearly bimodal, as observed in the signalling claws of fiddler crabs (*Uca*) where one claw is many times larger than the other, but where right-handed and left-handed individuals occur with roughly equal frequency in nearly all species (Yamaguchi 1977, Davis 1978). Subtler degrees of antisymmetry (see Fig. 3a below) will be apparent only as broad-

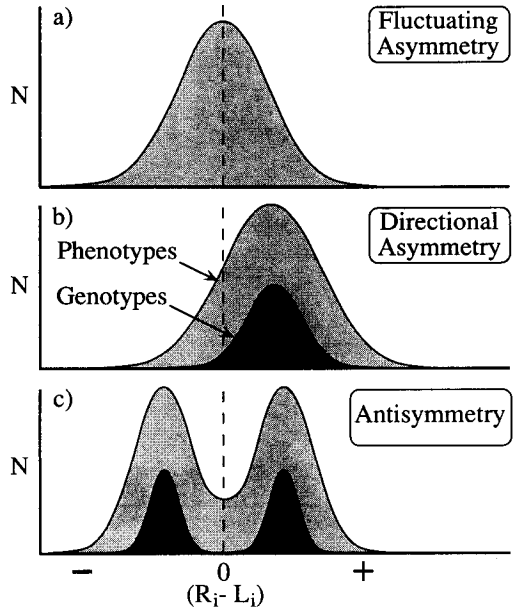

Fig. 2. Three "pure" forms of bilateral asymmetry: a) fluctuating asymmetry, b) directional asymmetry, and c) antisymmetry. See Fig. 1 for an explanation of the convention for representing genetic and environmental variation.

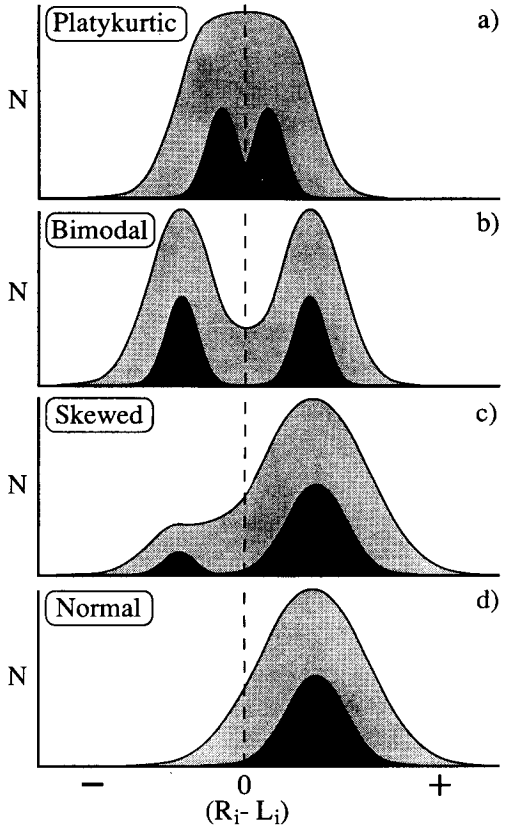

Fig. 3. Different degrees of antisymmetry, and a possible explanation for skewed distributions of *R-L* variation. a) weak antisymmetry, b) strong antisymmetry, c) a skewed distribution resulting from a mixture of antisymmetric and directionally symmetric variation (e.g. 3b and 3d) in a single sample, and d) directional asymmetry. See Fig. 1 for an explanation of the convention for representing genetic and environmental variation.

peaked, effectively unimodal (platykurtic) distributions of *R-L*. These subtler forms of antisymmetry pose perhaps the greatest difficulty to studies of developmental stability.

An important difference between these latter two forms of asymmetry and FA (Fig. 2a) is that, for both DA and antisymmetry, some unknown fraction of the variation in *R-L* may have a genetic basis (solid curves in Figs. 2b, c). In other words, in these latter two forms of asymmetry, individuals are genetically or developmentally directed to become asymmetrical. Consequently, for these two types of bilateral asymmetry, the variation in *R-L* may no longer be a product of pure developmental noise.

2.4. Composite distributions of more than one form of asymmetry: skew

Up to this point, we have considered each form of bilateral asymmetry, FA, DA, and antisymmetry, in isolation. Nothing, of course, precludes the joint occurrence of two or more forms of asymmetry for a given character in a single population. This possibility permits distributions of *R-L* having other shapes.

A skewed distribution of *R-L* variation (Fig. 3c) may result from a mixed sample where some individuals are predisposed towards antisymmetry (Fig. 3b) and the remainder exhibit di-

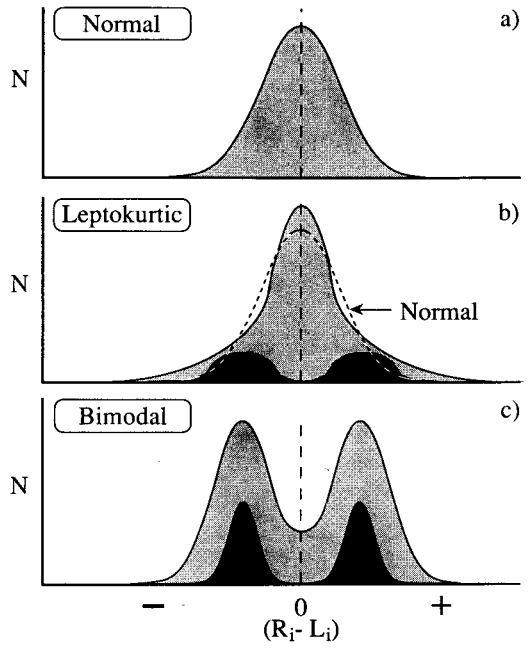

Fig. 4. A leptokurtic distribution of *R-L* variation (b) arising from a mixture of fluctuating asymmetry (a) and antisymmetry (c) in a single sample (Leptokurtic I). See Fig. 1 for an explanation of the convention for representing genetic and environmental variation.

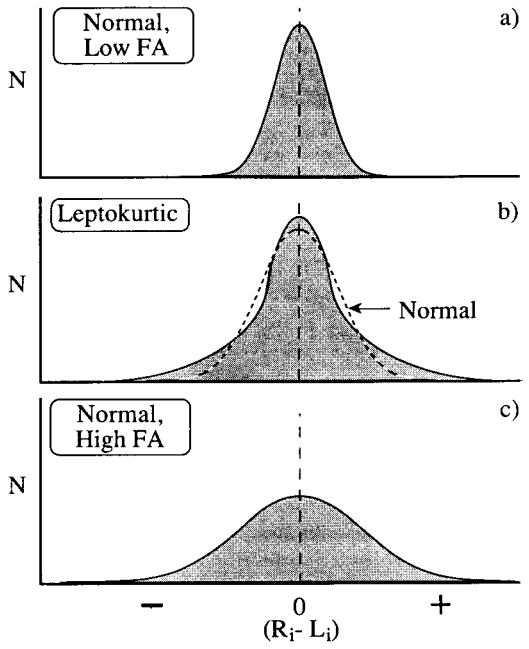

Fig. 5. A leptokurtic distribution of *R-L* variation (b) arising from a mixture of low fluctuating asymmetry (a) and high fluctuating asymmetry (c) in a single sample (Leptokurtic II). See Fig. 1 for an explanation of the convention for representing genetic and environmental variation.

rectional asymmetry (Fig. 3d). The extent and direction of skew will of course depend upon

a) the magnitude of both antisymmetry and DA,
b) the direction of DA, and
c) the relative frequency of the two forms in the mixed sample.

If both antisymmetry and DA have a genetic basis, as seems likely, some unknown fraction of the variation in *R-L* in the mixed sample will also have a genetic basis (solid curve, Fig. 3c). Once again, summary indexes of the variation of *R-L* can no longer be interpreted as reflecting pure developmental noise.

2.5. Composite distributions of more than one form of asymmetry: leptokurtosis

One type of leptokurtic distribution of *R-L* variation (narrow-peaked, long-tailed; Fig. 4b) may arise when a sample includes a mixture of two groups of individuals one of which exhibits FA (Fig. 4a) and the other of which exhibits antisymmetry (Fig. 4c). For convenience, we will call this Type I leptokurtosis. Such a distribution would arise only if individuals prone to antisymmetry made up a minority of the sample. If individuals prone to antisymmetry formed a majority, the composite distribution of *R-L* would probably be difficult to distinguish from pure antisymmetry unless it was clearly trimodal. As for the skewed distribution discussed above, variation in *R-L* can no longer be interpreted as pure developmental noise for this form of leptokurtosis.

Leptokurtic distributions of *R-L* may arise in two other ways. John Graham has pointed out to us that if a sample includes a mixture of two groups of individuals one of which exhibits low FA (Fig. 5a) and one of which exhibits high FA (Fig. 5c), the joint distribution (Fig. 5b) will be

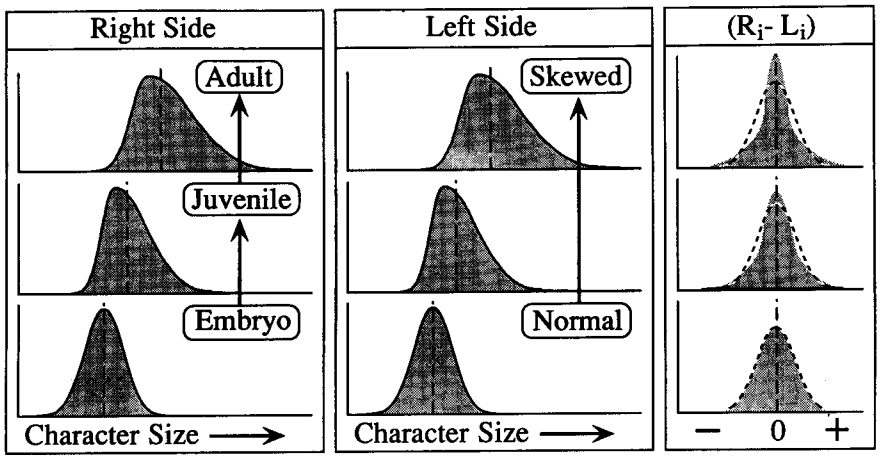

Fig. 6. A leptokurtic distribution of *R-L* variation (right-most three graphs) arising as a product of allometry during development (Leptokurtic III; see text for a more complete explanation). See Fig. 1 for an explanation of the convention for representing genetic and environmental variation.

leptokurtic (Type II leptokurtosis). Unlike Type I leptokurtosis (Fig. 4b), all the variation in *R-L* in Type II leptokurtosis would be due to developmental noise. Quantitative differences among samples in the variance of *R-L* would thus reflect *average* differences in developmental stability or developmental noise. However, this would obscure the fact that some individuals had low developmental stability while it was higher in others. Furthermore, variance differences among samples could arise due to changes in the proportion of the sample made up of individuals from the high- versus the low-FA group.

Finally, a leptokurtic distribution of *R-L* may also arise in a sample where the allometric growth of a trait on one side of the body depended on the degree to which it departed from bilateral symmetry early in ontogeny (Fig. 6). Consider a trait at an early stage in development, where both the right and left sides exhibited independent variation due exclusively to developmental noise. The distribution of *R-L* at this stage would meet the criteria for ideal FA (bottom row of panels, Fig. 6). If, as individuals grew larger, the allometric growth of the trait depended on whether it was initially over- or under-developed relative to the mean, then the frequency distribution of the trait for each side would become progressively more skewed with increasing body size (middle and upper panels for "right side" and "left side", Fig. 6). If the initial over- or under-development on each side was independent of that on the other, then the distribution of *R-L* differences would become progressively more leptokurtic (Type III leptokurtosis; right-most three panels Fig. 6). In view of the complexity of this mechanism, we are unsure whether a Type III leptokurtic distribution of *R-L* would include a genetic component or not.

2.6. A simple algebraic representation of asymmetry

A simple algebraic model will illustrate more explicitly the differences between *R* and *L* for a given bilateral trait. Following Palmer & Strobeck (1986):

$$R_i = \mu + s_i + (D/2)A_i + \alpha_i/2 + r_i, \text{ and} \quad (1)$$

$$L_i = \mu + s_i - (D/2)A_i - \alpha_i/2 + l_i. \quad (2)$$

where R_i and L_i represent the right and left members for individual i, and the remaining variables are defined and explained in Table 1. They include, among others, two to represent the mean and variance of directional asymmetry (D and α_i, respectively), one to represent the contribution of antisymmetry (A_i) and two to represent the

independent contribution of developmental noise to each side (r_i and l_i representing the developmental noise experienced by the right and left elements respectively). This formulation thus makes explicit what components of bilateral variation will still be present in the difference between R and L:

$$R_i - L_i = DA_i + \alpha_i + r_i - l_i \quad (3)$$

Clearly, only if the mean directional asymmetry (D) and the directional asymmetry variation (α_i) are both zero does the difference between R and L reflect pure developmental noise (r_i and l_i). Note that, according to this formulation, a component of directional asymmetry variation (α_i) may still remain even if no average directional asymmetry (D) exists. This would result from a negative covariance between sides that would not be apparent in a distribution of $R_i - L_i$ — this distribution would still be normal with a mean of zero. To the extent that this covariance has a genetic basis, even normal distributions of R-L centered on zero (supposedly "pure" FA) may have a small genetic component. Because we are concerned primarily with the implications of distributions that *depart* from normality in this paper, we will explore the full implications of this covariance elsewhere (Palmer et al. (submitted)).

2.7. Can samples departing from ideal FA be "corrected"?

The preceding argument has emphasized that, for many distributions departing from ideal FA, some fraction of the variation in R-L may have a genetic

Table 1. Components of variation of bilateral characters and their interpretation.

Population mean (μ)	Mean: Average character size: $(\Sigma R_i + \Sigma L_i)/2N$, where R = right, L = left and N = number of individuals. Variation: Not applicable.
Character size variation (s_i)	Mean: 0. Variation: Normal, genetic or environmental. Either genetic or environmental effects may increase or decrease R and L equally in an individual, relative to the population mean.
Population directional asymmetry (D)	Mean: Average directional asymmetry of a population: $\Sigma(R_i - L_i)/N$. Variation: Not applicable.
Antisymmetry (A_i)	Mean: A_i will take the value +1 or −1 with probability p and $q = 1-p$ respectively. p may vary from 0 to 1. For example, when antisymmetry is absent, $p = 1$. When antisymmetry is present, and right- and left- biased individuals occur with equal frequency, then $p = q = 0.5$. When antisymmetry is present, and right- and left-biased individuals occur with unequal frequency, p will be greater than or less than 0.5 depending upon whether right-biased or left-biased individuals are more common. Hence, the mean of A_i is $2p-1$ with a variance of $4pq$. Variation: Binomial, genetic or environmental. As noted by Palmer & Strobeck (1986), antisymmetry may arise via two different mechanisms, either a) via a mixture of right- and left-biased genotypes (Type I Antisymmetry), or b) via chance or an environmental stimulus that produces right- or left- biased phenotypes from a single genotype (Type II Antisymmetry).
Directional asymmetry variation (α_i)	Mean: 0. Variation: Normal, genetic.
Developmental noise of right and left sides respectively (r_i, l_i)	Mean: 0. Variation: Normal, environmental.

basis (e.g. see equation 3 and the summary of distribution shapes in Table 2). Hence, the variation in *R-L* is not pure developmental noise and a statistical correction cannot eliminate this problem. A simple graphical example should make this clear.

Consider a trait exhibiting statistically significant DA (Fig. 7a). As outlined above, the variation in *R-L* will have both a genetic component (some individuals will be genetically predisposed towards greater asymmetry than others, solid region) and a developmental noise component (stippled region). The mean of the distribution may be "corrected" to be zero via the appropriate arithmetic (Fig. 7b). This, however, achieves only a *statistical* correction. Although the resulting distribution now meets the statistical criteria for ideal FA, the genetic component of the variation in *R-L* is still present. Hence this variation may still not be interpreted as developmental noise.

Regrettably, statistical corrections for DA have been applied in a number of studies (e.g. Mather 1953, VanValen 1962, Soulé 1967, Livshits et al. 1988, among others) thereby rendering their conclusions about differences in developmental stability among samples open to doubt. If a trait exhibits statistically significant DA, we suggest it should be eliminated from consideration as a measure of developmental stability.

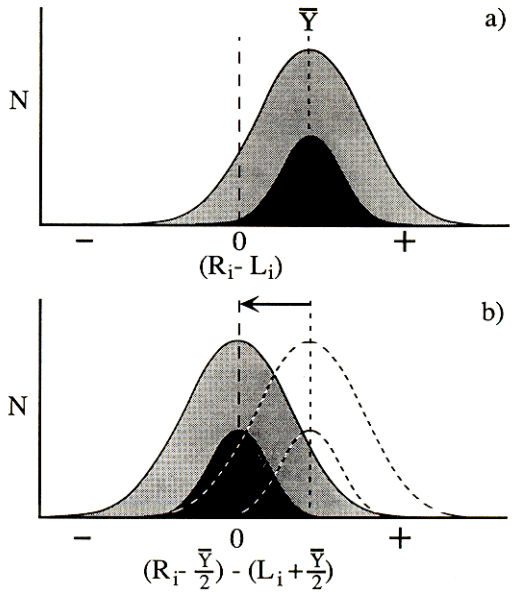

Fig. 7. A graphical illustration of the consequences of "correcting" for directional asymmetry: a) a trait exhibiting directional asymmetry before correction, b) the same trait after correction. Note that correction does not eliminate the genetic fraction of the variation. See Fig. 1 for an explanation of the convention for representing genetic and environmental variation.

2.8. Can antisymmetry be interpreted as decreased developmental stability?

Some have considered that antisymmetry and FA are merely different manifestations of developmental noise (e.g. McKenzie & Clarke 1988). Along with VanValen (1962, p. 126), we feel this is unlikely. Consider the extreme case of fiddler crabs mentioned above (see 'Other pure forms of asymmetry'). Male fiddler crabs exhibit an extreme form of antisymmetry in their signal-

Table 2. Shapes of *R-L* distributions and their possible significance. Origin of variation: E = environmental including developmental noise, G = genetic.

	Shape	Mean	Form of asymmetry	Origin of variation	See Fig.
a	Normal	0	Ideal fluctuating asymmetry (FA)	E[a]	2a
b	Normal	≠0	Directional asymmetry (DA)	E & G	2b
c	Platykurtic	0	Weak antisymmetry	E & G	3a
d	Bimodal	0	Strong antisymmetry	E & G	3b
e	Skewed	≠0	Mixed DA & antisymmetry	E & G	3c
f	Leptokurtic I	0	Mixed FA & antisymmetry	E & G	4b
g	Leptokurtic II	0	Mixed high FA & low FA	E[a]	5b
h	Leptokurtic III	0	Allometrically amplified FA	?	6

[a]This assumes no covariance between *R* and *L* (see conclusion to Section 2.6.)

ling claws but the claws of female fiddler crabs are not conspicuously asymmetrical (Yamaguchi 1977). Do male fiddler crabs somehow experience greater developmental noise (are they less developmentally stable) than females? Clearly not. Similarly, American lobsters *Homarus americanus* also exhibit extreme antisymmetry in their master claws — one claw develops into a crusher claw and the other remains a cutter claw. In very young juveniles, both claws begin as cutter claws. Which claw becomes the crusher claw is determined by which receives more use handling hard objects (Govind & Pearce 1986). Does this mean lobsters experience more developmental noise or are less developmentally stable than female fiddler crabs? Again, clearly not.

A reasonable question to ask is whether a bimodal distribution centered on zero (e.g. Fig. 2c) can even arise as a product of developmental noise. This is equivalent to asking whether developmental noise can give rise to a frequency distribution of *R-L* whose maximum is not at zero, as in Fig. 2c. We prove below that, given two simple assumptions, polymodal distributions of *R-L* will always have a maximum at zero. Hence a bimodal distribution of *R-L*, as found in pure antisymmetry, can arise only if one or both assumptions are violated.

Theorem: If 1) r_i and l_i are independently distributed, and 2) they have the same frequency distribution, $f(x)$, then the frequency distribution $g(a)$, where $a = r_i - l_i$, has a global maximum at $a = 0$.

Proof: Because the frequency distributions of r_i and l_i are both $f(x)$, and because they are independently distributed, the joint frequency distribution of r_i and l_i is $f(r_i)f(l_i)$. The frequency of distribution of $r_i - l_i$ is thus

$$g(r_i - l_i) = g(a) = \int_{-\infty}^{+\infty} f(x) f(x - a)\, dx. \quad (4)$$

To prove that only a single global maximum exists at $r_i - l_i = 0$ it is both necessary and sufficient to show that

$$g(0) > g(a) \text{ for all } a \neq 0 \quad (5)$$

which requires that

$$\int_{-\infty}^{+\infty} f^2(x)\, dx > \int_{-\infty}^{+\infty} f(x) f(x - a)\, dx. \quad (6)$$

This may be rewritten as

$$\int_{-\infty}^{+\infty} f^2(x) - f(x) f(x - a)\, dx > 0 \quad (7)$$

Subtracting a constant from x has no effect on the integral, hence $x–a$ may be substituted for x and

$$\int_{-\infty}^{+\infty} f^2(x)\, dx = \int_{-\infty}^{+\infty} f^2(x - a)\, dx. \quad (8)$$

Therefore, the sum of one-half of each of these integrals is still

$$\int_{-\infty}^{+\infty} f^2(x)\, dx = \int_{-\infty}^{+\infty} \frac{f^2(x)}{2}\, dx + \int_{-\infty}^{+\infty} \frac{f^2(x - a)}{2}\, dx. \quad (9)$$

Substituting from equation 9 into equation 7 means we must now show that

$$\int_{-\infty}^{+\infty} \frac{f^2(x)}{2} - f(x) f(x - a) + \frac{f^2(x - a)}{2}\, dx > 0. \quad (10)$$

But this reduces to

$$\int_{-\infty}^{+\infty} \frac{[f(x) - f(x - a)]^2}{2}\, dx > 0. \quad (11)$$

Equation 11 is always true because $f(x)$ and $f(x–a)$ must differ over some interval. Only if $f(x)$ was a constant for all values of x would $f(x)$ always equal $f(x–a)$ and hence would inequality 11 not be true. However, the integral of such a function would either be infinite or zero, and we know that for a frequency distribution

$$\int_{-\infty}^{+\infty} f(x) = 1. \quad (12)$$

A simple numerical example may illustrate this point more intuitively. Consider a population in which the *R* and *L* limbs of a bilateral pair can only occur in one of two discrete sizes (e.g. 1 unit or 3

units in length). If a) a limb on either side has a probability p of being size 1 and $q = 1 - p$ of being size 3, and b) the size of the limb on one side is completely independent of that on the other, then the frequency distribution of *R-L* will be trimodal, with a maximum peak of height $p^2 + q^2$ at zero, and two equivalent peaks of height pq at +2 units and –2 units. Hence even with only two discrete states for each side, so long as they occur independently on each side, the frequency distribution of *R-L* will be trimodal with a maximum at zero. Increasing the number of possible states will only increase the number of modes, which will always be odd in number.

We feel that different levels of antisymmetry should not be interpreted as representing different levels of developmental stability (see also VanValen 1962). A peculiar mechanism of development might exist, where antisymmetry arises early in ontogeny due to some form of developmental constraint and then converges on bilateral symmetry to varying extents depending on how developmentally stable an organism is. However, even if such a mechanism was shown to exist, it is not clear how this type of developmental stability could be compared to the developmental stability inferred from differences in FA.

3. Statistical considerations

3.1. Testing for departures from normality

Because the use of differences in FA to infer differences in developmental stability among samples depends heavily on whether the frequency distributions of *R-L* are normal, we review briefly the relative power of tests for departures from normality. Shapiro (1968) conducted an extensive computer simulation using eight tests for normality, five sample sizes, and 44 distributions. They concluded that the *W* test (Shapiro & Wilk 1965) was overall more robust than the other tests, which included tests for skewness and kurtosis, and the Kolmogorov-Smirnov test (*K-S* test). A combination of the tests for skewness and kurtosis, however, performed almost as well as, and sometimes surpassed, the *W* test. Significantly, a closer inspection of their results revealed that the only distributions where tests for kurtosis and skewness failed completely were discrete distributions that were approximately normal, i.e., the binomial and Poisson distributions. The *K-S* test, on the other hand, exhibited surprisingly poor power over most non-normal distributions but worked well if the distributions were discrete (e.g. binomial or Poisson) and approximately normal.

These simulations suggest to us that tests for skewness and kurtosis when taken together are probably the most useful way to detect departures from normality for metrical traits. Although the *W* test is the most robust overall, the joint use of skew and kurtosis statistics has two advantages. First, unlike the *W* test, these two descriptive statistics are widely available on microcomputer and mainframe statistical packages, and their significance levels are easily calculated (Sokal & Rohlf 1981). Second, these two statistics provide a more complete description of how a distribution departs from normality.

Even where meristic characters are used to describe bilateral asymmetry, tests for skew and kurtosis still appear to be the best way to detect departures from normality. Although the *K-S* test was more effective at detecting departures from normality for certain discrete distributions (Shapiro et al. 1968), this test rejects such distributions as non-normal because of the discrete nature of the data, and not because the underlying processes are fundamentally non-normal. In other words, the *K-S* test may reject as non-normal distributions of *R-L* based on meristic characters, even though they might have arisen exclusively from developmental noise.

3.2. Power of tests for differences among three or more samples

As we have argued elsewhere (Palmer & Strobeck 1986), the most useful descriptor of FA is the variance. Therefore testing for differences in FA among samples amounts to testing whether the variances are equal for two or more samples. For two samples the *F* test is the most powerful test for equality of variances if the two samples are normally distributed (Lehmann 1959). For more than two samples at least 20 different tests are available (Conover et al. 1981). Four of the more

common tests, Bartlett's, F_{MAX}, Scheffé's, and Levene's are summarized in Table 3. To determine which of these tests would be most powerful and robust for detecting differences of FA among populations, we conducted a series of Monte Carlo simulations. Although these simulations repeat to a large extent those reported previously in other studies, we provide here the power curves for these four tests. These curves make more apparent the manner in which the tests lose statistical power for distributions whose shapes depart from normality.

In our simulations, the performance of each test was examined for three symmetric distributions: the normal distribution, a platykurtic distribution (the uniform distribution), and a leptokurtic distribution. These distributions were generated as follows. Normally and uniformly distributed random deviates were obtained from IMSL subroutines GGNQF and GGUBS respectively. The leptokurtic distribution was generated as the difference between two lognormal deviates $[\exp(y\sqrt{\log 2}) - \exp(z\sqrt{\log 2})]/2$ where y and z are normal deviates N[0,1]. We examined two cases where the variances differed among three populations either as a) 1.0, 1.0, and x, or b) $1/x$, 1.0, and x, respectively. To determine the effect of sample size, population sizes of 30 and 100 were used. For a sample size of 30, the values of x were $2^{i/2}$ (where $i = -4, -3, -2, -1, 0, 1, 2, 3, 4$) and when the sample size was 100, the values of x were $2^{i/4}$ (where $i = -4, -3, -2, -1, 0, 1, 2, 3, 4$). For each model, population size, and value of x, samples for the three populations were drawn independently from the appropriate IMSL subroutine. A P-value of 0.05 was used as the criterion for rejection in all simulations.

The power curves generated from our simulations revealed that some tests were more sensitive to departures from normality than others (Figs. 8 and 9). These curves describe the probability of rejection given that the null hypothesis (variances for all three samples equal) is false. Hence, for a given variance, the higher the power the greater the ability of a test to detect true differences in the variances among samples. At the same time, to conform to the conventions for statistical significance, the probability of rejecting the null hypothesis when true (i.e. when all variances are in fact equal), should be $P = 0.05$. Thus, the preferred test should have a) a higher power to detect true differences in the variances, and b) the proper rejection rate ($P = 0.05$) when the variances are equal.

Three conclusions can be drawn from these power curves. First, Scheffé's test was consistently less powerful than the others when the distributions were normal, regardless of whether

Table 3. Four common tests for heterogeneity of variances (heteroscedasticity) among three or more samples. s_i^2 = variance of R-L for sample i, n_i = number of individuals per sample i, N = number of samples.

Bartlett's test[a]	$\chi^2 = [\sum (n_i - 1)] \ln s_T^2 - [\sum (n_i - 1) \ln s_i^2]$, where $s_T^2 = [\sum (n_i - 1) \ln s_i^2] / \sum (n_i - 1)$ and n_i = number of observations per sample i. χ^2 significance is assessed for $N - 1$ degrees of freedom.
F_{MAX} test[a]	Given $s_1^2, s_2^2, s_3^2, \ldots s_n^2$, the test statistic is s_{max}^2 / s_{min}^2.
Levene's test[b]	One-way analysis of variance on $\lvert R_i - L_i \rvert$ where each cell corresponds to a single sample.
Scheffé's test[a]	1) randomly divide each sample i of n_i observations into m subsamples (the optimum value of $m = \sqrt{n_i}$.) 2) compute the variance separately for each of the m subsamples 3) conduct a one-way analysis of variance on $\log s_{i,j}^2$ (each sample will thus be represented by m log-transformed variances, the ANOVA tests for differences in average log variance among samples).

[a] described in Sokal & Rohlf (1981); Scheffé's test = Scheffé-Box test in Conover et al. (1981).
[b] described in Snedecor & Cochran (1980).

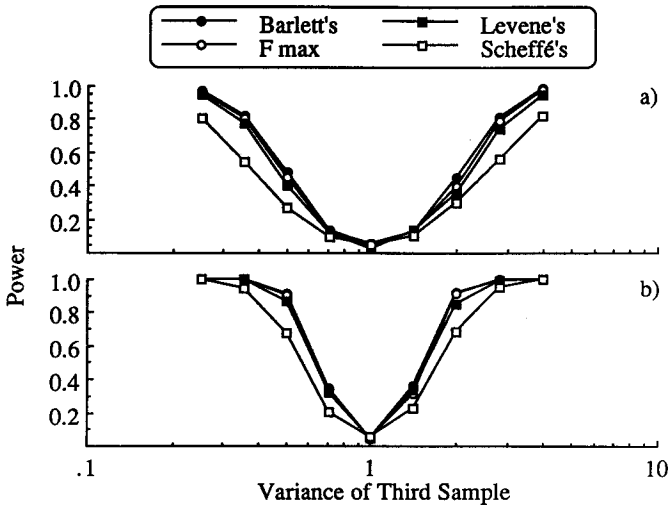

Fig. 8. Power curves for four tests for heterogeneity of variances among three samples when variation is normal: a) only the variance of the third sample varies, b) the variance of both the first and third sample varies. N = 30 for all samples. Scheffé's test was conducted with 5 subsamples of size 6. Power is the percent rejection out of 1000 trials. See Table 3 for descriptions of the tests.

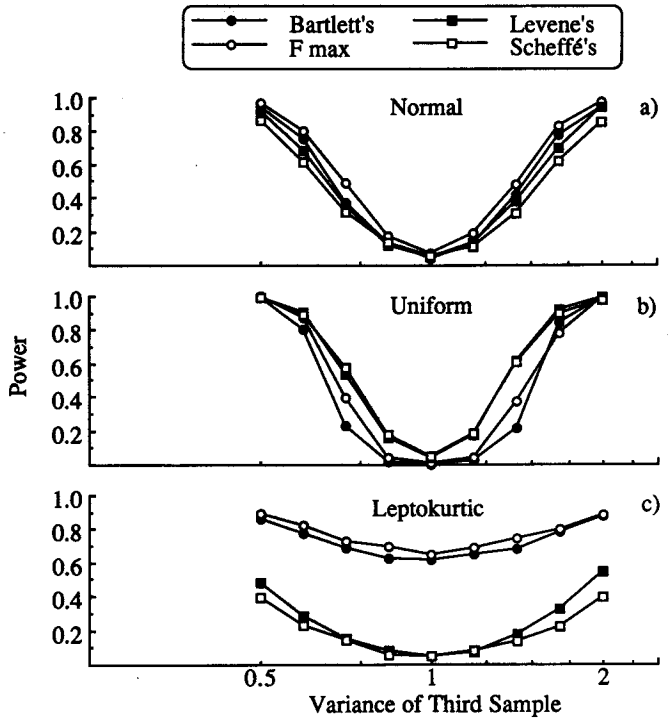

Fig. 9. Power curves for four tests for heterogeneity of variances among three samples when variation is: a) normal, b) uniform (platykurtic), or c) leptokurtic. Only the variance of the third sample varies. N = 100 for all samples. Scheffé's test was conducted with 10 subsamples of size 10. Power-percent rejection out of 1000 trials. See Table 3 for descriptions of the tests.

sample sizes were small or large (Fig. 8a vs 9a), or whether variances differed in one or two of the three simulated populations (Fig. 8a vs 8b). It was either the same as or less powerful than Levene's test when the distributions were not normal (Figs. 9b, c). These simulations confirm the conclusions of Layard (1973) that Scheffé's test is not a powerful test for heterogeneity of variances, hence it seems of dubious value for FA studies. Second, the F_{MAX} test and Bartlett's test were both very sensitive to non-normality (Sokal & Rohlf 1981). Where the distribution was platykurtic (uniform distribution) both tests rejected the null hypothesis when true (i.e. equal variance for all three populations), with a probability of approximately 0.003 instead of 0.05 (Fig. 9b). Where the distribution was leptokurtic this probability was approximately 0.65 instead of 0.05 for both tests (Fig. 9c). Third, Levene's test and Scheffé's test were both robust for all distributions tested. That is, when the null hypothesis was true (equal variances for all three samples) the rejection rate of the null hypothesis was approximately 0.05 as it should have been (Figs. 8 and 9). These results agree with those of Conover et al. (1981), but they did not consider Scheffé's test and did not present results for the uniform distribution.

Therefore we recommend that Levene's test be used for three reasons. First, it appears to be the least sensitive to departures from normality in the direction of leptokurtosis or platykurtosis, and hence more likely to yield correct P values even where these departures are small. Note, however, that Levene's test is rather sensitive to asymmetrical distributions (i.e. skew; Conover et al. 1981). Second, because it involves only an analysis of variance (Table 3), it may be conducted easily with common microcomputer and mainframe statistical packages. Finally, even where distributions were normal, Levene's test was only slightly less powerful than Bartlett's or F_{MAX} (Fig. 8).

4. Conclusions

4.1. Whither developmental stability?

Given that so many departures of *R-L* distributions from ideal FA may signal a genetic component to bilateral asymmetry (Table 2) and may thereby render inferences about levels of developmental stability tenuous, what should one conclude? Does this greatly restrict the utility of developmental stability studies? Unfortunately, we cannot answer this question with much confidence given the information available. Too few data exist to assess how common the above departures from ideal FA are.

This dearth of data highlights our first and perhaps most important conclusion. Future studies of developmental stability should

a) include more explicit descriptions of the distributions of *R-L* (all four statistical moments: mean, variance, skew, and kurtosis) for the variables examined, and
b) describe the results of statistical tests for departures from ideal FA for these characters.

Alternatively, histograms of *R-L* distributions for selected traits or selected samples crucial to the main conclusions of the study could suffice. Without the benefit of such basic descriptive information, studies of developmental stability will remain very difficult to interpret with much confidence, and meaningful quantitative comparisons among studies will be virtually impossible.

Second, where a single trait is used to characterize developmental stability, the *interpretation* of differences among samples will be much more sensitive to departures from ideal FA. From a practical point of view, nearly all distributions of *R-L* will depart from ideal FA in some way or another if only by chance. Without additional data, one may or may not be able to determine with any confidence if these departures have influenced the conclusions of a study or not. If several developmentally independent traits are examined instead of one, however, and the qualitative differences among experimental groups or among natural populations are for the most part consistent across these traits, then differences among samples are much more likely to reflect real differences in organism-wide developmental stability or developmental noise. Studies based upon multiple traits will thus be more convincing than those based on one or a few traits.

Finally, skeptics may feel that our cautions are largely hypothetical. In particular they may

question whether any evidence exists to support our peculiar notion of composite populations where, for example, both FA and antisymmetry may co-occur for a given trait in a single sample (e.g. Fig. 4b). A recent paper, however, reports precisely the kind of composite distribution about which we have raised concern. McKenzie & Clarke (1988) present striking evidence for a single allele (or block of tightly linked alleles) that induces antisymmetry in the sheep blowfly *Lucilia cuprina* (see their Fig. 6). The sample of individuals lacking this allele exhibited ideal FA. Although they interpret the variation erroneously as FA, their data suggest that this antisymmetry allele increased in frequency shortly after the application of pesticides, and then declined in frequency as *Lucilia* adapted to the pesticide (Clarke & McKenzie 1987). Composite distributions of asymmetry variation are thus clearly possible.

4.2. Limitations of a classical study

Among the many studies of FA, one of the most extensive and thought-provoking is that of Mather (1953) on homeostasis in *Drosophila melanogaster*. He examined the effects of sex differences, environmental stress, inbreeding, crosses between inbred strains, and selection for both increased and decreased FA. Although an admirable study in a great many ways, Mather's conclusions all depend critically on the assumption that the single trait with which he estimated developmental stability (sternopleural chaetae number) exhibited ideal FA. If it exhibited some form of antisymmetry (e.g. Fig. 3a, b), as VanValen (1962, p. 126) suggested it did, the majority of the patterns he reported may have reflected changes in the frequency of genes affecting antisymmetry rather than in developmental stability. Unfortunately, because he did not present any histograms of *R-L* or statistics describing the shape of this distribution, this possibility cannot be ruled out. We hope future studies of this magnitude do not suffer from such oversights.

Acknowledgments. This research was supported by NSERC Operating Grants A7245 (ARP) and A0502 (CS). We thank John Graham and two anonymous reviewers for their thoughtful comments on drafts of the manuscript. We are particularly grateful to Vladimir Zakharov and all of his colleagues and coworkers at the Koltzov Institute of Developmental Biology in Moscow for providing us an opportunity to present and explore these ideas, and for hosting us so graciously.

References

Allendorf, F. W. & Leary, R. F. 1986: Heterozygosity and fitness in natural populations of animals. — In: Soulé, M. E. (ed.), Conservation biology. The science of scarcity and diversity: 57–76. Sinauer Assoc., Inc., Sunderland, MA (USA).

Clarke, G. M. & McKenzie, J. A. 1987: Developmental stability of insecticide resistant phenotypes in blowfly; a result of canalizing natural selection. — Nature (Lond.) 325:345–346.

Conover, W. J., Johnson, M. E. & Johnson, M. M. 1981: A comparative study of tests for homogeneity of variances with applications to the Outer Continental Shelf Bidding data. — Technometrics 23:351–361.

Davis, T. A. 1978: Reversible and irreversible lateralities in some animals. — Behav. Brain Sci. 2:291–293.

Govind, C. K. & Pearce, J. 1986: Differential reflex activity determines claw and closer muscle asymmetry in developing lobsters. — Science 233:354–356.

Layard, M. W. J. 1973: Robust large-sample tests for homogeneity of variances. — J. Amer. Stat. Assoc. 68:195–198.

Lehmann, E. L. 1959: Testing statistical hypotheses. — Wiley, New York. 369 pp.

Lewontin, R. C. 1983: Gene, organism and environment. — In: Bendall, D. S. (ed.), Evolution from molecules to men: 273–285. Cambridge Univ. Press, New York, NY.

Livshits, G., Davidi, L., Kobyliansky, E., Ben-Amital, D., Levi, Y. & Merlob, P. 1988: Decreased developmental stability as assessed by fluctuating asymmetry of morphometric traits in preterm infants. — Amer. J. Med. Gen. 29:793–805.

Mather, K. 1953: Genetical control of stability in development. — Heredity 7:297–336.

McKenzie, J. A. & Clarke, G. M. 1988: Diazinon resistance, fluctuating asymmetry and fitness in the Australian sheep blowfly, Lucilia cuprina. — Genetics 120:213–220.

Palmer, A. R. & Strobeck, C. 1986: Fluctuating asymmetry: measurement, analysis, patterns. — Ann. Rev. Ecol. Syst. 17:391–421.

Palmer, A. R., Strobeck, C. & Chippindale, A. K. (submitted): Bilateral variation and the evolutionary origins of asymmetry. — Amer. Nat.

Shapiro, S. S. & Wilk, M. B. 1965: An analysis of variance test for normality (complete samples). — Biometrika 52:591–611.

Shapiro, S. S., Wilk, M. B. & Chen, H. J. 1968: A comparative study of various tests for normality. — J. Amer. Stat. Assoc. 63:1342–1372.

Sokal, R. R. & Rohlf, J. F. 1981: Biometry. 2nd ed. — Freeman, San Francisco, CA. 859 pp.

Soulé, M. E. 1967: Phenetics of natural populations. II. Asymmetry and evolution in a lizard. — Amer. Nat. 101:141–160.

VanValen, L. 1962: A study of fluctuating asymmetry. — Evolution 16:125–142.

Waddington, C. H. 1957: The strategy of the genes. — George Allen Unwin, London.

Yamaguchi, T. 1977: Studies on the handedness of the fiddler crab, Uca lactea. — Biol. Bull. 152:424–436.

Zakharov, V. M. 1989: Future Prospects for Population Phenogenetics. — In: Turpaev, T. M. & Yablokov, A. V. (eds.), Physiology and general biology reviews. Harwood Academic Publishers GmbH, New York. 79 pp.