Fluctuating Asymmetry Analyses: A Primer

A. Richard Palmer   Phone: 403-492-3633
Department of BioSciences   FAX: 403-492-9234
University of Alberta   E-mail: rich.palmer@ualberta.ca
Edmonton, AB T6G 2E9
CANADA

and

Bamfield Marine Station
Bamfield, BC V0R 1B0
CANADA

1.0 Fluctuating asymmetry and developmental stability: A cursory overview

The developmental stability of an organism is reflected in its ability to produce an 'ideal' form under a particular set of conditions (Zakharov, 1992). The lower its stability, the greater the likelihood it will depart from this 'ideal' form. Ideal forms are rarely known a priori. However, bilateral structures in bilaterally symmetrical organisms offer a precise ideal, perfect symmetry, against which departures may be compared (Palmer and Strobeck, 1986). Thus they provide a very convenient method for assessing deviations from the norm, and studying the factors that might influence such deviations.

Subtle departures from symmetry are most commonly described by frequency distributions of right - left (R - L). Such frequency distributions usually exhibit one of three patterns (Fig. 1; (VanValen, 1962). Other patterns, however, are theoretically possible (skewed asymmetry,(Palmer and Strobeck, 1992); normal covariant asymmetry (Palmer, et al., 1993); see Sect. 2.3), so no assumptions should be made about distribution shapes or patterns of covariation between R and L.
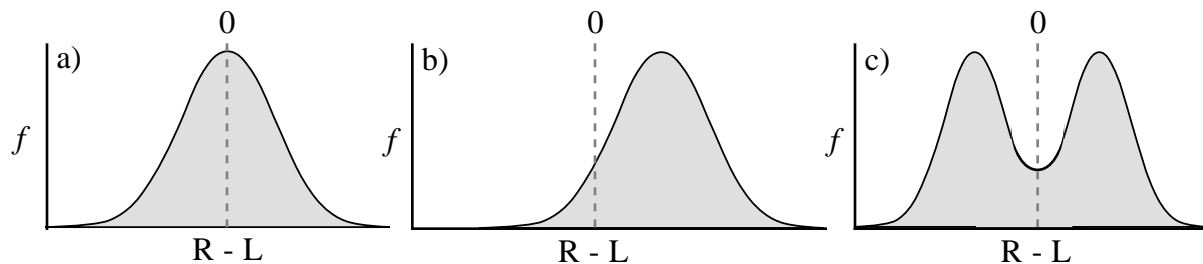
Fig. 1. Three common distributions of R - L in bilateral organisms: a) fluctuating asymmetry (FA), b) directional asymmetry (DA), c) antisymmetry (platykurtic or bimodal).

Fluctuating asymmetry (=FA, Fig. 1a) is widely used as a measure of developmental stability. It either shows no change or it increases with increasing extrinsic (environmental) or intrinsic (predominantly genetic) 'stress' (see reviews in (Palmer and Strobeck, 1986; Leary and Allendorf, 1989; Parsons, 1990; Zakharov, 1992). Deviations from symmetry also appear to correlate with fitness differences, particularly where traits directly affect performance (Thornhill, 1991; Møller, 1994).

A critical assumption underlying studies of FA as a measure of developmental stability is that departures from symmetry in an individual should not have a heritable basis (Palmer and Strobeck, 1992). If departures from symmetry are heritable, then differences in the extent of asymmetry among individuals will have both a genetic and a non-genetic basis. Unless the genetic component can be partitioned out, the variation in R - L can no longer be assumed to describe developmental stability. Debate remains over the reliability of DA and antisymmetry as measures of developmental stability (Palmer and Strobeck, 1992; Graham, et al., 1993; McKenzie and O'Farrell, 1993), so caution is advised when encountering these forms of subtle asymmetries in studies of FA.

Most studies of FA attempt to distinguish differences in the between-sides variation among two or more samples. Since differences between sides are often very small (generally < 5% and often <1% of trait size), great care must be taken during both measurement and analysis to maximize the likelihood of detecting FA differences among samples.

Regrettably, data from published studies of FA are nearly impossible to compare quantitatively for a variety of reasons outlined below. Where discrepancies exist in the literature, I

suspect that these are most commonly due to limited analytical power. I hope that this primer will help illustrate: a) how to conduct FA analyses with the greatest power and precision, b) how to avoid common pitfalls in FA analyses, and c) how to present data in a fashion that allows it to be compared <u>quantitatively</u> among studies. The checklist at the end (Sect. 16.0) should serve as a useful guide.

2.0 Terminology

2.1 Asymmetry in an individual vs pattern of asymmetry variation in a sample

The terms FA, DA and antisymmetry all refer to patterns of variation in a particular trait exhibited by a <u>sample</u> of individuals. One cannot determine to which pattern a departure from symmetry in a single individual belongs without reference to other individuals in the sample. More seriously, one can not conclude that because a particular trait does not depart statistically from FA that DA or antisymmetry are absent for this trait. One can only conclude that, given the sample size used and given the particular precision of measurement, no departure from FA could be detected. For example, small sample sizes or moderate measurement error may obscure statistical departures from FA even though they may exist.

For this reason, some other term would seem advisable to apply to departures from symmetry in an individual. '<u>Subtle asymmetry</u>' might be useful as a general term to refer to slight departures from symmetry in an individual since it makes no presumptions of the distribution to which that departure in that individual might belong. This distinction may seem like splitting hairs, but considerable confusion arises when referring to correlations among traits or correlations among samples, and when trying to discuss the heritability of subtle asymmetries (Sect. 14.1).

2.2 Patterns vs. processes

Care should be taken to use descriptors of patterns of asymmetry variation solely to describe patterns. FA is a <u>pattern</u> of between-sides variation in a sample of individuals. It reflects some developmental compromise between two presumably independent but opposing sets of <u>processes</u> (Fig. 2) grouped loosely under the headings of:

developmental noise = developmental instability- a suite of processes that tend to disrupt precise development, such as a) small, random differences in rates of cell division, cell growth and cell shape change, b) effects of thermal noise on enzymatic processes, c) small, random differences in rates of physiological processes among cells.

developmental stability- a suite of processes that tend to resist or buffer the disruption of precise development, such as a) negative feedback systems that regulate enzyme activity (both concentration and catalytic rate) within and between cells, b) central nervous regulation of non-contiguous structures, or c) hormonal regulation of non-contiguous structures.
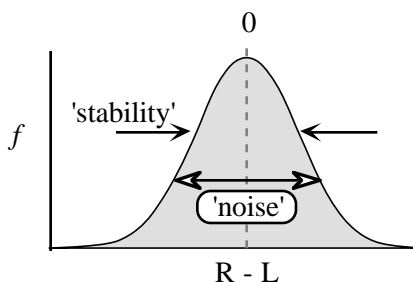

Fig. 2. The magnitude of FA reflects a compromise between two opposing processes: developmental noise and developmental stability.

FA may increase either because developmental noise increases, or because developmental stability decreases (Mather, 1953; VanValen, 1962; Palmer and Strobeck, 1992). Thus differences in FA among samples may arise due to differences in one or the other or both of these phenomena; the differences alone can not delineate their relative contribution. The outcome of these processes influences the extent of homeorhesis, or stabilized flow of development, of (Waddington, 1957) and (Zakharov, 1992).

2.3 Glossary

absolute asymmetry- the absolute value of the difference between the R and L sides of a trait in an individual, |R - L|.

antisymmetry- a pattern of bilateral variation in a sample of individuals, where a statistically significant difference exists between sides, but where the side that is larger varies at random among individuals (Fig. 1c); detected by statistical tests for departures of frequency distributions of (R - L) from normality in the direction of platykurtosis (broad peakedness, or negative values for kurtosis; Sect. 8.0); mean R - L normally zero (VanValen, 1962). May render traits unusable for studies of developmental stability (Sect. 8.0).

bilateral variation- a general term referring to all patterns of between-sides variation in bilaterally symmetrical organisms (FA, DA, antisymmetry, NCA, etc.).

canalization- the ability of a structure to develop along an ideal developmental trajectory under a variety of different environmental conditions (Waddington, 1940). A phenomenon distinct from developmental stability, which refers to the ability of a structure to develop along an ideal developmental trajectory under a particular set of environmental conditions (Zakharov, 1992). A trait that is highly canalized will exhibit little phenotypic plasticity in response to different growth environments, even though it may still exhibit normal levels of developmental noise.

covariant asymmetry- a general term referring to patterns of bilateral variation that involve negative covariation between sides within a sample (i.e. antisymmetry and normal covariant asymmetry) (Palmer and Strobeck, 1992).

developmental instability- see Sect. 2.2.

developmental noise- see Sect. 2.2.

developmental stability- see Sect. 2.2.

directional asymmetry (DA)- a pattern of bilateral variation in a sample of individuals, where a statistically significant difference exists between sides, but the side that is larger is generally the same (Fig. 1b); detected by statistical tests for departures of mean R - L from zero (Sect. 7.0; (VanValen, 1962). May render traits unusable for studies of developmental stability (Sect. 7.0).

fluctuating asymmetry (FA)- a pattern of bilateral variation in a sample of individuals where the mean of R - L is zero and variation is normally distributed about that mean (Fig. 1a); a pattern of bilateral variation that may arise via many processes (Sect. 2.2).

homeorhesis- 'stabilized flow' of a developmental trajectory; a phrase coined by (Waddington, 1957) and encouraged by (Zakharov, 1992); refers to the capacity for a structure to develop along an ideal developmental trajectory under a particular set of environmental conditions. A phenomenon distinct from canalization.

non-directional asymmetry- a general term useful to refer to all forms of subtle asymmetries except directional asymmetry (i.e. FA, antisymmetry, normal covariant asymmetry).

normal covariant asymmetry (NCA)- a hypothetical pattern of bilateral variation in a sample where bilateral traits exhibit continuous negative covariation; on a scatterplot of R vs L for a sample, it would appear as an elliptical distribution of points whose major axis was

perpendicular to the line of perfect symmetry (R=L); it may or may not pose a problem for studies of developmental stability depending on how widespread it is (see Fig. 1h of (Palmer, et al., 1993).

phenodeviant- refers to a trait whose value in an individual lies well outside the 'normal' range for a trait. For each trait described in this manner, one must take care to define precisely, and to defend, what is considered to be 'normal'.

phenotypic plasticity- the capacity for a traits produced by identical genotypes to exhibit different phenotypes under different environmental conditions, the greater the plasticity the lower the canalization.

Population Asymmetry Parameter (PAP)- a phrase coined by (Soulé, 1967) to describe the correlation of asymmetry indices for different traits among populations (i.e. populations with greater asymmetry variation for one trait also exhibit greater asymmetry variation for other traits). Soulé believed such correlations arose because differences in developmental stability were reflected in all traits (Sect. 14.3).

signed asymmetry- the difference between the R and L sides of a trait in an individual with the sign included, (R - L); retains information about the direction of the asymmetry.

skewed asymmetry- a pattern of bilateral variation in a sample of individuals where the frequency distribution of R - L departs significantly from normality in the direction of skew.

subtle asymmetry- a convenient term to refer to departures from symmetry of a trait in a single individual, as opposed to a sample of individuals (Sect. 2.1). Avoids much of the confusion that may arise where terms for the pattern of variation exhibited by a sample are are applied to deviations from symmetry in an individual (Sect. 14.0).

3.0 Indices for describing the level of FA in a sample

3.1 FA indices

Many different indices have been used to describe the level of FA in a sample (Fig. 3, Table 1). The two most frequently used ones are **FA1** & **FA4**. A common modification is to (try to!) correct for size-dependence by dividing by the mean trait size either at the level of individuals (**FA2**, **FA6**, **FA8**) or at the level of samples (**FA3**, **FA7**; see Sect. 10.0 for cautions regarding size correction). Another index (**FA10**) can be obtained from the results of an ANOVA procedure that simultaneously tests for measurement error, size differences among individuals and directional asymmetry (Sect. 6.6).
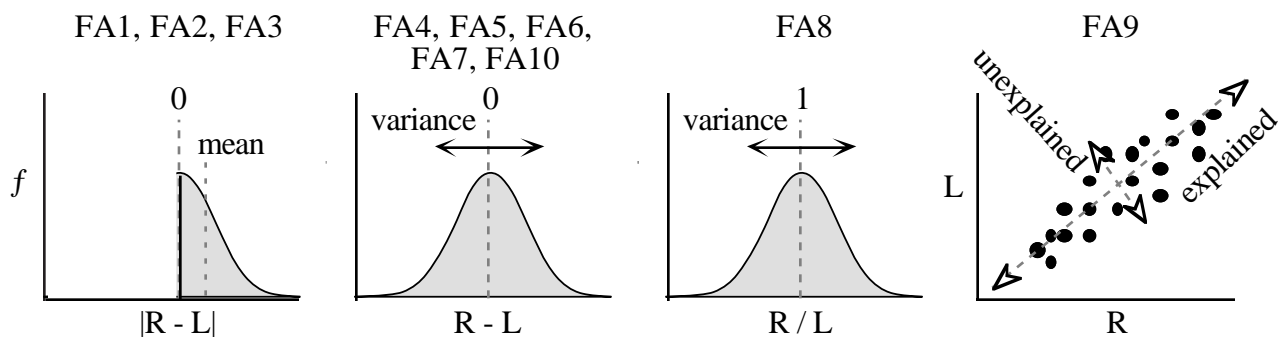

Fig. 3. Graphical illustration of what is being estimated by different FA indices; see Table 1 for actual indices.

Table 1. FA indices (modified from Table 1 of Palmer & Strobeck, 1986; see original paper for references).

Form of size corr.	Unsigned (=absolute) asymmetry |R-L|	Signed asymmetry (R-L)	Ratio between sides (R/L)
none	**FA1:** mean |R-L|	**FA4:** var(R-L) **FA5:** $\sum(R\text{-}L)^2/N$	
by individual	**FA2:** mean [|R-L|/((R+L)/2)]	**FA6:** var[(R-L)/((R+L)/2)]	**FA8:** var(log(R/L))
by sample	**FA3:** mean |R-L|/mean((R+L)/2)	**FA7:** var(R-L)/mean((R+L)/2)	

Other indices for single traits:

    FA9: $1 - r^2$ of correlation between R & L (i.e. % bilateral variation not due to positive covariation).

    FA10: $s^2_i = (\underline{MS_{sj}} - \underline{MS_m})/\underline{M}$ from a sides x individuals ANOVA (from Table 3 of Palmer & Strobeck, 1986); describes the magnitude of non-directional asymmetry after partitioning out measurement error (see Sect. 6.6 for a full explanation).

Indices based on multiple traits per individual:

    FA11: asymmetry in an individual $(A_i) = \sum|R\text{-}L|$ for all traits; the index for a sample is $\sum A_i /N$ where N= number of individuals in the sample (Leary, et al., 1985).

    FA12: asymmetry in an individual (A_i)= total number of asymmetrical traits in an individual, independent of how large the deviation is between sides; the index for a sample is $\sum A_i /N$ where N= number of individuals in the sample (Leary, et al., 1985).

    FA13: Generalized index of overall fluctuating asymmetry (GFA); a multivariate measure of average deviation from symmetry for multiple metrical traits (Zhivotovsky, 1992).

3.2 Pros & cons of different FA indices

  FA1-   Pros: easily computed; yields a number that is intuitively easy to understand; unbiased estimator of the sample standard deviation; readily used in the recommended ANOVA procedure for testing for differences among 3 or more samples (Palmer and Strobeck, 1992); not as sensitive to outliers as FA4; may allow more sophisticated multi-trait tests for differences among samples.

              Cons: will be very biased if either DA or antisymmetry are present; only 87.6% as efficient as the sample standard deviation (Kendall and Stuart, 1951), hence some statistical power is lost that might be important where sample sizes or differences between samples are small; sensitive to size-dependence of \|R - L\|.

              Recommendation: probably the most generally useful index for moderate to large sample sizes (30+). May be used with smaller sample sizes, but at the sacrifice of some statistical power. Use only where DA and antisymmetry are absent (Sects. 7.0, 8.0), and where \|R - L\| does not change with overall size (Sect. 10.0).

  FA2-   Pros: as for FA1.

              Cons: as for FA1 except not biased by size-dependence of \|R - L\|.

              Recommendation: Use only where clear evidence exists of a size dependence of \|R - L\| among individuals within a sample (Sect. 10.2).

  FA3-   Pros: as for FA1.

              Cons: as for FA1 except not biased by size-dependence of \|R - L\|.

              Recommendation: Use only where clear evidence exists of a size dependence of \|R - L\|

among samples (Sect. 10.3).

FA4- Pros: easily computed; lends itself to the most powerful test for differences between two samples (F test; (Lehmann, 1959); more efficient than FA1 for estimating the between-sides variation; not biased by DA.

Cons: more sensitive to outliers than FA1; will be very biased if antisymmetry is present; tests for differences among 3 or more samples are more sensitive to departures from normality than FA1 (Palmer and Strobeck, 1992); yields units that are not as intuitively easy to understand as FA1; sensitive to size-dependence of |R - L|.

Recommendation: A useful descriptor of FA particularly where only two samples are to be compared (Sect. 13.1). Use only where antisymmetry is absent (Sect. 8.0), and where |R - L| does not change with overall size (Sect. 10.0).

FA5- Pros: yields an estimate of the between-sides variance with one additional degree of freedom than FA4 because the mean is assumed to be zero (i.e. one degree of freedom is not used to estimate the mean); statistical power is gained that might be important where sample sizes, or differences between samples, are small.

Cons: harder to compute than FA1 or FA4; will be very biased if either DA or antisymmetry are present; more sensitive to outliers than FA1; sensitive to size-dependence of |R - L|.

Recommendation: The main value of this index is the one additional degree of freedom it yields over FA4. Where sample sizes are small (on the order of 20 or less,) or where differences in FA between samples are small, this index may yield additional needed statistical power. Use only where DA and antisymmetry are absent (Sects. 7.0, 8.0), and where |R - L| does not change with overall size (Sect. 10.0).

FA6- Pros: as for FA4.

Cons: as for FA4 except not biased by size-dependence of |R - L|.

Recommendation: Use only where clear evidence exists of a size dependence of |R - L| among individuals within a sample (Sect. 10.2).

FA7- Pros: as for FA4.

Cons: as for FA4 except not biased by size-dependence of |R - L|.

Recommendation: Use only where clear evidence exists of a size dependence of |R - L| among samples (Sect. 10.3).

FA8- Pros: easily computed; not biased by size-dependence of |R - L|.

Cons: may yield artificial differences among samples if |R - L| does not depend on trait size but trait size varies widely among individuals or samples since the value of this index depends on both the difference between-sides and mean trait size; restricted number of tests for among-sample differences; not widely used in studies of FA; units not as intuitively easy to understand as FA1 or FA4.

Recommendation: Not recommended. Basically equivalent to FA6, but does not yield numbers that are as well defined statistically.

FA9- Pros: easily computed; expresses between-sides variance as a proportion of the total variation between sides and among individuals; not biased by DA.

Cons: is very dependent on overall trait size variation in the sample since, for a given level of developmental noise, the index decreases as the range of body sizes increases (Angus, 1982); will be very biased if antisymmetry is present.

Recommendation: Not recommended. Do not use except in conjunction with other indices.

FA10- Pros: this is the only index that allows the measurement error variance to be partitioned out of the total between-sides variance; not biased by DA; lends itself to the most powerful test for differences between two samples (F test; (Lehmann, 1959).

Cons: more difficult to compute than FA1 or FA4; degrees of freedom approximate and depend on the difference between MS_{inter} and MS_{err}; will be very biased if

antisymmetry is present; yields units that are not as intuitively easy to understand as FA1; sensitive to size-dependence of |R - L|; tests for differences among 3 or more samples are more sensitive to departures from normality than FA1 (Palmer and Strobeck, 1992).

Recommendation: The principle advantage to this index is that the measurement error variance may be partitioned out of the non-directional asymmetry variance (although taking the average of repeated measurements reduces the impact of measurement error, residual variation due to measurement error still remains, see Sect. 6.6). This yields a more accurate estimate of the true non-directional asymmetry variance, but the degrees of freedom for the estimate depend upon the relative sizes of the non-directional asymmetry and measurement error variances (Palmer & Strobeck, 1986, pg. 408); in other words, the greater the measurement error, the lower the confidence in the estimate of the non-directional asymmetry variance (Sect. 6.1). Probably most useful when accompanied by other indices (e.g. FA1 or FA4).

FA11- Pros: easily computed; combines information from multiple traits and may thus yield a more accurate estimate of the overall average asymmetry of an individual where the average deviation from symmetry of all traits is comparable; combines information from all traits into a single index for a sample.

Cons: will be very biased if either DA or antisymmetry are present in any trait; obscures differences in variability among characters; yields values that may not be compared quantitatively with other studies; may be biased by one or a few traits where some traits are more variable than others.

Recommendation: Use only a) in conjunction with other indices for individual traits, b) where tests are needed for correlations between the inferred level of developmental stability in an individual (e.g. its 'average' asymmetry over several traits) and some other factor such as fitness (Sect. 15.0), performance, heterozygosity, etc., or c) where one wishes to pool information from all traits to increase the likelihood of detecting differences among samples. However, see Sect. 14.2 to determine whether differences in FA among individuals are significant statistically or not and Sect. 13.3 for a test that pools information from multiple traits using conventional FA indices.

FA12- Pros: easily computed; combines information from multiple traits and may thus yield a more accurate estimate of the overall average asymmetry of an individual; combines information from all traits into a single index for a sample; less likely to be biased than FA11 by one or two extreme values or by one or two extremely variable traits.

Cons: may only validly be applied to meristic traits or traits where one may identify 'phenodeviants' (Sect. 2.3); will be very biased if either DA or antisymmetry are present in any trait; obscures differences in variability among characters; yields values that may not be compared quantitatively with other studies; less able to reveal differences in overall average asymmetry among individuals than FA11 because no account is taken of the magnitude of the difference between sides in a given trait.

Recommendation: Use only in parallel with FA11 where concerns exist about differences in the relative variability of different traits.

FA13- Pros: combines information from all traits into a single index for a sample even where the average deviation from symmetry of all traits is not comparable.

Cons: difficult to compute; may not be valid for meristic traits where differences between sides are small; will be very biased if either DA or antisymmetry are present in any trait; obscures differences in developmental stability among characters; yields values that may not be compared quantitatively with other studies.

Recommendation: Use only in conjunction with other indices for individual traits, and where one wishes to pool information from all traits to increase the likelihood of detecting differences among samples. See also Sect. 13.3 for a test that pools information from multiple traits using conventional FA indices.

3.3 Relationships among FA indices

The indices of Table 1 differ in the degree to which they may be biased by overall trait size variation, by the presence of DA and antisymmetry, and by departures from normality of the frequency distribution of R - L. However, for a truly normal distribution where: i) mean (R - L)= 0, ii) |R - L| does not depend upon overall trait size, and iii) measurement error is absent, the relations among expected values of the indices are:

FA1= 0.798 * $\sqrt{\textbf{FA4}}$ = $\sqrt{\textbf{FA4}}$ / ($\sqrt{(\pi/2)}$) (Kendall and Stuart, 1951)

FA5= **FA4** (Palmer and Strobeck, 1986)

FA10= **FA4** / 2 (Palmer and Strobeck, 1986)

FA2, **FA3**, **FA6**, **FA7**, **FA8, FA9-** relations to each other and other indices not are defined because values depend either on mean size (all but FA9) or size range (FA9) of traits.

3.4 General recommendations regarding FA indices

i) Every index of Table 1 is sensitive to antisymmetry, hence tests for antisymmetry should be conducted for all traits (Sect. 8.0). Traits exhibiting antisymmetry should be used only with caution if at all (Sect. 1.0).

ii) Many indices of Table 1 are sensitive to DA, hence tests for DA should also be conducted for all traits (Sect. 7.0). Traits exhibiting DA should be used only with caution if at all (Sect. 1.0).

iii) Never assume |R - L| <u>depends on</u> overall trait size. Indices that incorporate size scaling (FA2, FA3, FA6, FA7) should only be used where clear evidence exists that |R - L| depends on overall trait size (Sect. 10.0). If used indiscriminately, size-scaled indices can generate artificial differences in FA where samples differ in overall size but not in the average absolute difference between sides.

iv) Never assume |R - L| is <u>independent</u> of overall trait size. Where |R - L| depends on trait size, artificial differences in FA among samples may arise where samples differ in overall size.

v) Because indices differ in their sensitivity to other factors (outliers, DA, trait size variation, other departures from normality), tabulations of FA indices should include values for at least two FA indices (e.g. FA1, FA4, or FA10) to illustrate the dependence of differences in FA on the choice of index (Sect. 12.0).

vi) Where two or more indices are presented, one should be FA10, because this is the only index that effectively removes the impact of measurement error from the between-sides variance (Sect. 6.0).

4.0 Choice of traits

4.1 Pros & cons of meristic vs metrical traits

meristic <u>Pros:</u> may often (but not always!, see (Jagoe and Haines, 1985)) be measured without error thus avoiding the need for replicate measurements; may be fixed early in development and thus not sensitive body size variation or age (Taning, 1952; Leary, et al., 1985) but see (Strawn, 1961; Angus and Schultz, 1983) for contrary evidence).
<u>Cons:</u> differences between sides are often small (e.g. 1 or 2); quantizing of variation may obscure subtle departures from FA (particularly antisymmetry); where most asymmetrical individuals differ by only 1 between sides, all FA indices depend on how close mean trait size is to an integer value (Sect. 4.2); the degree to which indexes of FA describe the level of developmental noise depends on how individual meristic elements are added to or lost from a group (Sect. 4.2).
<u>Recommendation:</u> Great care must be taken with traits that rarely differ by more than

1 between sides in a given sample (Sect. 4.2). However, when traits are easy to count with little or no error, exhibit moderate variation between sides, and counts are independent (or nearly so) of body size and age, meristic traits can be ideal for studies of FA.

metrical Pros: variation is continuous so the ability to detect differences between sides, or departures from FA, is limited only by measurement precision and accuracy; virtually any trait may be used.

Cons: measurement error generates the same pattern of between-sides variation as FA and hence some estimate of measurement error is necessary to confirm that between-sides differences are greater than measurement error; between-sides variation often increases with overall trait size so a correction for size-dependence of |R - L| may be needed.

Recommendation: Any use of metrical traits in studies of FA variation should be accompanied by clear estimate of the size of measurement error relative to FA (Sect. 6.0). Unless measurement error is less than roughly 25% of the between-sides variation, all traits should be measured at least twice. Because measurement error can obscure differences in FA among samples, particular care must be taken to document the contribution of measurement error in studies reporting negative results.

Care must also be taken to test for, and correct if necessary, a dependence of |R - L| on overall size for each trait (Sect. 10.0).

4.2 Two idiosyncrasies of meristic traits

Idiosyncrasy I- Although meristic traits may be advantageous in some circumstances for studies of FA, particular care must be taken where most asymmetrical individuals in a sample differ between sides by only 1 (see (Swain, 1987) for a general discussion of this problem). Swain notes that in such cases the number of asymmetrical individuals in the sample depends heavily on how close the mean count is to a whole integer value. Consider the following hypothetical situation. If bristles are uniformly spaced along a limb segment, and the spacing between bristles remains constant on average, then even where the uncertainty in bristle position (developmental noise, shaded frequency distributions at segment bases in Fig. 4) remains the same, all measures of FA will depend on the average length of the limb segment and hence the average bristle number. If a limb segment had seven bristles on average, and the length of the segment was such that the average position of the terminal bristle was well in from end of the segment (Pop. A, Fig. 4), then even if the position of this bristle varied, most individuals would still be symmetrical, and only a few would have one fewer or one extra on a given side. If, however, the segment was just a little shorter, and the average position of the terminal bristle was precisely at the end of the segment (Pop. B, Fig. 4), the bristle is much more likely to be missing from one side or the other. In this hypothetical example, the terminal bristle would be missing half of the time and the average bristle count would be 6.5. Hence, even though the underlying developmental instability in terms of placement of the bristle is the same in both cases, the observed asymmetry variation is larger for Pop. B, simply because the segment on which it occurs is shorter.

This impact of a non-integer mean on the variance of the between-sides difference is particularly apparent when considering the absolute value of R - L (right-most histograms, Fig. 4). Where the mean count is a whole integer value (Pop. A), a large peak occurs at zero with a small peak off to one side, but if the mean count is a half-integer value (Pop. B), equal frequencies exist in both the zero and +1 class. Clearly, the computed FA will be larger for Pop. B than Pop. A, even though the level of developmental stability is the same for both populations (shaded frequency distributions at segment bases). This arises simply because it is not possible to have 6.5 elements on one side of an individual.
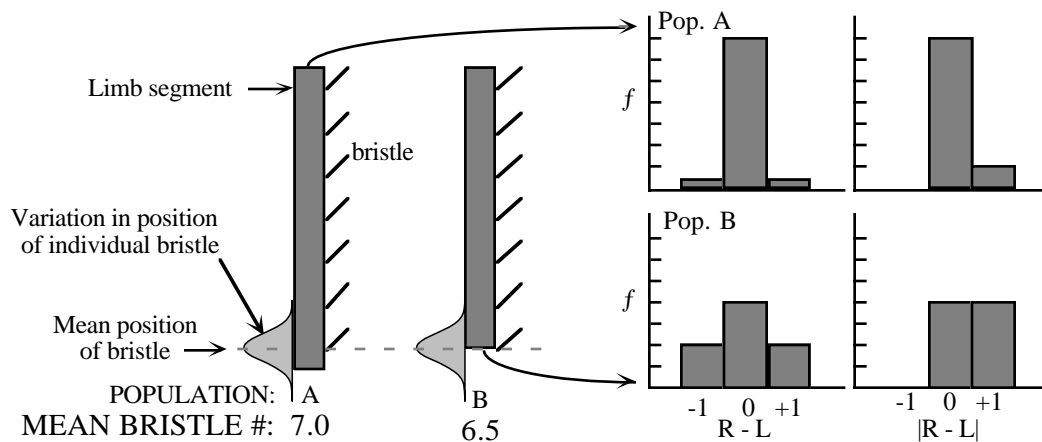
Fig. 4. Effect of non-integer mean on FA variation when sides differ by only 1. In population B, the limb segment is slightly shorter than that for population A. Shaded frequency distributions indicate the hypothetical frequency of occurrence of bristle #7 at various positions near the end of the limb segment for both populations. Cross-hatched frequency distributions indicate the frequency of symmetrical and asymmetrical individuals for both signed (left) and unsigned (right) asymmetry for each population.
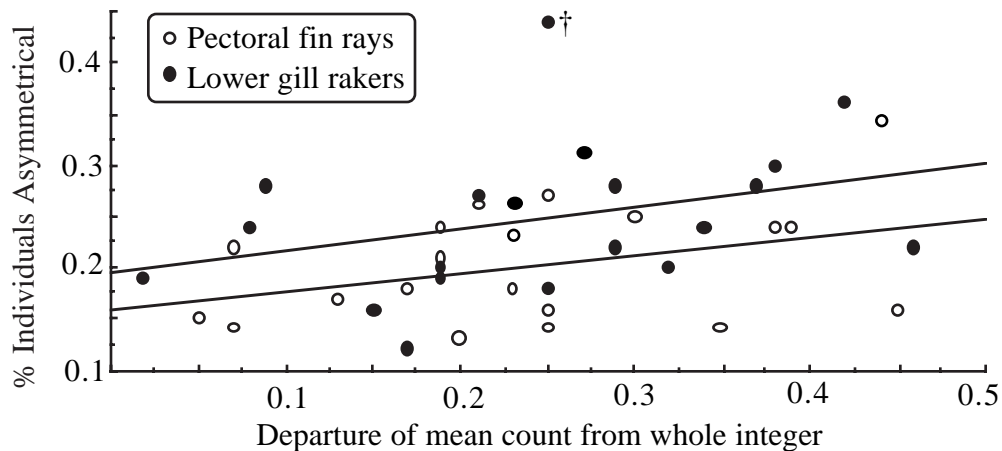
Fig. 5. Effect of deviation of sample mean count from a whole integer value on asymmetry variation among 20 populations of rainbow trout, Onchorynchus mykiss (common slope P= 0.0171 ANCOVA, †- outlier excluded from analysis; data from Tables 2 & 5 of (Leary, et al., 1992); three other meristic traits did not exhibit such an association).

Such concerns are not entirely hypothetical, as illustrated by data for pectoral fin rays and lower gill rakers in rainbow trout (Fig. 5). Should such an association be detected, one solution might be to convert asymmetry estimates to residuals from the regression of asymmetry against departure of mean count for a trait from a whole integer. Swain (1987) recommends that care be taken when $\leq$20% of the sample depart from symmetry by no more than 1 and the mean count is a whole integer, or where $\leq$1% of the sample depart from symmetry by no more than 1 and the mean count is halfway between integer values.

<u>Idiosyncrasy II-</u> Variation in the number of meristic elements in a series can arise via two fundamentally different mechanisms (Scharloo, 1991). In one, elements are added or lost sequentially from terminal positions in a series (e.g. fin rays in fishes). In the second, elements may be present or absent independently at each of several fixed positions in a series (e.g. certain bristles in <u>Drosophila</u>). The sensitivity of a particular meristic trait to developmental noise may very well depend on the mechanism by which meristic elements are added or lost from the series, though this has yet to be explored theoretically.

4.3 Single vs multiple traits

As a rule, multiple traits are preferred when testing for differences in developmental stability <u>among samples</u> (Leary and Allendorf, 1989; Palmer and Strobeck, 1992). If several traits yield a concordant pattern of variation (e.g. see (Zakharov, et al., 1991)), even if very subtle, one has greater confidence that the pattern is somehow 'real', and not an artifact of something peculiar to a particular trait (Palmer and Strobeck, 1992). Thus, wherever possible, examine patterns of FA variation in several traits (Sect. 12.0).

Furthermore, when using Levene's test for differences in FA among samples, additional power may be gained by using multiple traits to conduct a two-way factorial ANOVA (traits x samples) on |R - L|, and testing for an overall effect of 'samples' (Sect. 13.3).

4.4 Choose traits that are developmentally independent

Although subtle asymmetries in different traits seem rarely to correlate among individuals <u>within a sample</u> (Palmer and Strobeck, 1986), such correlations are occasionally found (Waddington, 1955; Chang, et al., 1960; Bader, 1965; Brown and Wolpert, 1990a). If the goal of a study is to be able to infer differences in developmental stability <u>among samples</u>, the less developmentally correlated the traits the more robust the conclusion. If, however, the goal of a study is to be able to make inferences about the developmental basis of subtle deviations from symmetry <u>in individuals</u>, then a comparison of closely connected versus distantly connected traits could be quite informative (Leamy, 1993).

4.5 Choose traits that exhibit 'ideal' FA

In studies of developmental stability, traits that exhibit 'ideal' FA (mean (R - L)= 0, variation= normal) should be selected over those that exhibit DA or antisymmetry, where possible. For traits that exhibit DA or antisymmetry, some of the between-sides difference in an individual may be due to heritable variation for asymmetry (Palmer and Strobeck, 1992). If so, then the variation in such traits in a sample of individuals will reflect a mix of developmental noise, heritable differences for asymmetry, and effects of genotype-environment interactions for asymmetry. As with any trait whose development is affected by both non-genetic and genetic factors, the relative contribution of these sources can only be assessed via some form of heritability analysis. Debate remains over the reliability of traits that exhibit DA or antisymmetry as indicators of developmental stability (Palmer and Strobeck, 1992; Graham, et al., 1993; McKenzie and O'Farrell, 1993), so caution is advised when encountering these forms of subtle asymmetries.

Laboratory studies, for example of the effect of stress on developmental stability, would seem less vulnerable to concerns about DA and antisymmetry since the genetic makeup of experimental groups may be controlled by the investigator. So long as DA and antisymmetry are small, and experimental groups are reasonably homogeneous genetically, differences in developmental stability (however measured) among experimental groups should not be too seriously confounded. However, where DA and antisymmetry are moderate or large, the possibility arises that some of the differences among experimental groups might arise due to genotype-environment interaction (i.e. the effect of genes contributing to DA or antisymmetry may depend on the environment in which the individual develops), in which case differences among groups would not be due to differences in developmental stability. In addition, the between sides variance of traits exhibiting antisymmetry cannot be compared quantitatively with that for traits

exhibiting ideal FA.

Concerns about DA and antisymmetry are more serious for studies of developmental stability in natural populations. If deviations from symmetry are heritable in traits that exhibit DA or antisymmetry, then apparent differences in developmental stability among populations could arise due to differences in the the amount of heritable variation for asymmetry.

Ideally, a preliminary survey should test for departures from FA of R - L in a range of traits. As a precaution, traits exhibiting significant DA (Sect. 7.0), or antisymmetry or skew (Sect. 8.0) should be excluded from further study. If this is not practical, see Sect. 9.0 for some guidelines for avoiding spurious associations.

5.0 Sample sizes

Tests for differences in FA are effectively tests for differences in variances among samples. Even using an F-test to test for differences in FA between two samples (the most powerful test in this case), the ability to detect, for example, a two-fold difference in the variance depends very much on the sample size: a sample size of 10 will only reveal a significant difference ($\alpha = 0.05$) 25% of the time, a sample size of 25 will only reveal a significant difference 50% of the time, and a sample size of 40 will only reveal a significant difference 75% of the time (Fig. 6). For a detailed discussion of this problem, see (Smith, et al., 1982).
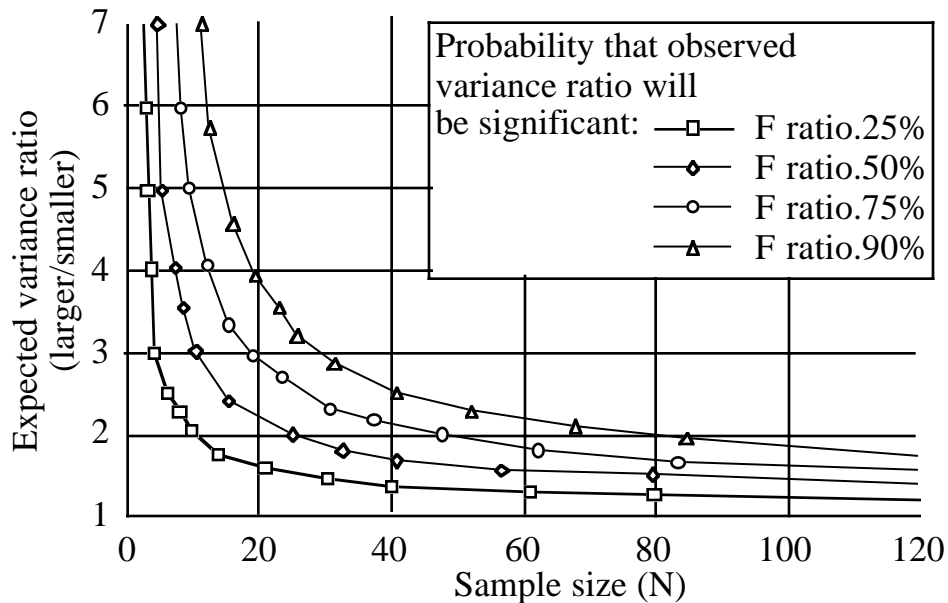
Fig. 6. Effect of sample size on ability to detect a difference between two variances using an F-test (modified from Smith et al., 1982).   The smaller the difference between the variances, the larger the sample size must be to be able to detect the difference at a given probability. For example, to detect a true difference between variances of 2 even 50% of the time requires a sample size of at least 20. To detect such a difference 75% of the time, requires a sample size of 40.

As in any statistical analysis, the recommended sample size depends on the magnitude of the differences in the 'signal' (i.e. FA) among samples. As a rough rule of thumb, N= 30 should be considered a minimum sample size in studies of FA. Where concerns exist about possible departures from normality, N= 40 or 50 would be more desirable since the size of the sample will greatly influence the ability to detect departures from normality (Sect. 8.0).

6.0 Measurement error

6.1 Why is measurement error a particular concern in studies of FA?

Measurement error is a particularly troubling concern in studies of FA because the 'data' to be compared among samples are measures of the variation between-sides. Not only can measurement error obscure differences among samples, as in any analysis, but it can also give the illusion that the between-sides variation is large and does not differ among samples. Consider, for example, a bilaterally symmetrical mechanical part machined to within 0.0001 mm on each side. If 50 such parts were given to a student and s(he) was asked to measure the asymmetry of these parts with calipers to a precision of 0.1 mm, the frequency distribution of R - L would, of course, be normally distributed about a mean of zero (e.g. see Fig. 9c). Virtually all of the apparent 'between-sides' variation, however, would simply be due to measurement error, not to variation among individual parts.

More seriously, measurement error not only affects confidence in estimates of FA, as it does for any trait, but it also biases estimates of the between-sides variance (Fig. 7). For a conventional trait where one is attempting to estimate the mean, increasing measurement error simply increases the variance about the mean, but it has no impact on the mean itself (Fig. 7a). However, for FA, where one is attempting to estimate a variance, increasing the measurement error actually increases the variance (Fig. 7b). The greater the measurement error, the greater the impact on the estimate of the between-sides variance (compare middle vs. lower panel of Fig. 7b). Fig. 7 thus illustrates why a quantitative estimate of measurement error is essential in studies of FA. Without it, the 'observed FA' contains some unknown mix of 'true FA' and measurement error. Fig. 7 also illustrates why an FA index such as FA10 (Sect. 3.1, 6.6) is particularly valuable. This index formally attempts to separate the fraction of the between-sides variance due to FA from that due to measurement error. All other indices simply lump these two components of the between-sides variance together, and thus yield estimates of FA that are not comparable quantitatively among traits or among studies.
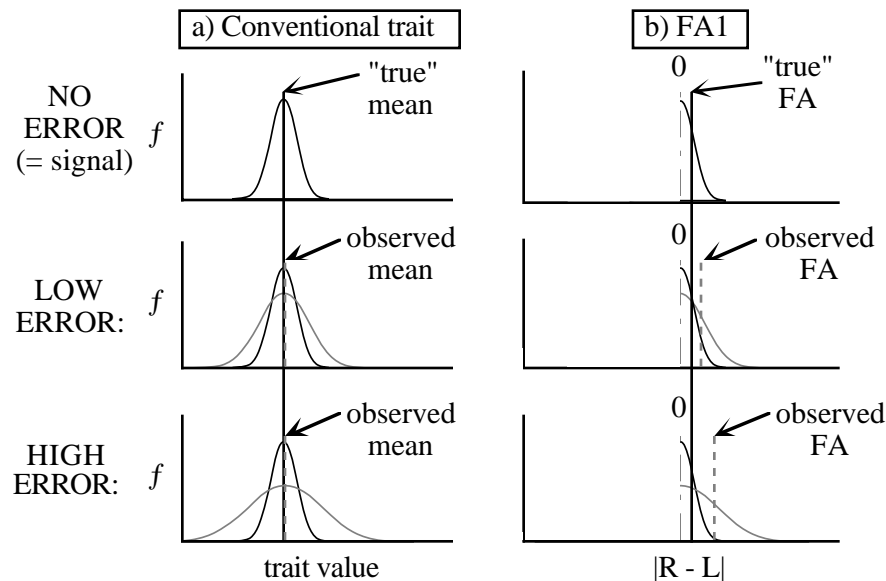
Fig. 7. Impact of measurement error on conventional traits (a) and estimates of FA1 (b). Although only one FA index is illustrated, the impact of measurement error is the same for all indices except FA10. Solid curves and solid vertical lines indicate the 'true' values for a sample. Dashed curves and uniformly-dashed vertical lines indicate the values actually observed when measurement error is added to the 'true' values. Short/long dashed line indicates the expected value for perfect symmetry.

The statistical impact measurement error has on estimates of FA may be more apparent if one examines the effect it has on the effective size of the sample in which one is attempting to estimate FA (Fig. 8). As measurement error increases relative to the underlying non-directional asymmetry, the effective size of the sample with which one estimates non-directional asymmetry decreases. For example, if the measurement error variance is only 25% of the non-directional asymmetry variance (for two replicate measurements per side), the effective sample size is 80% of that actually measured. However, if the measurement error variance is 100% of the non-directional asymmetry variance, the effective sample size for estimating the non-directional asymmetry variance is only 42%.
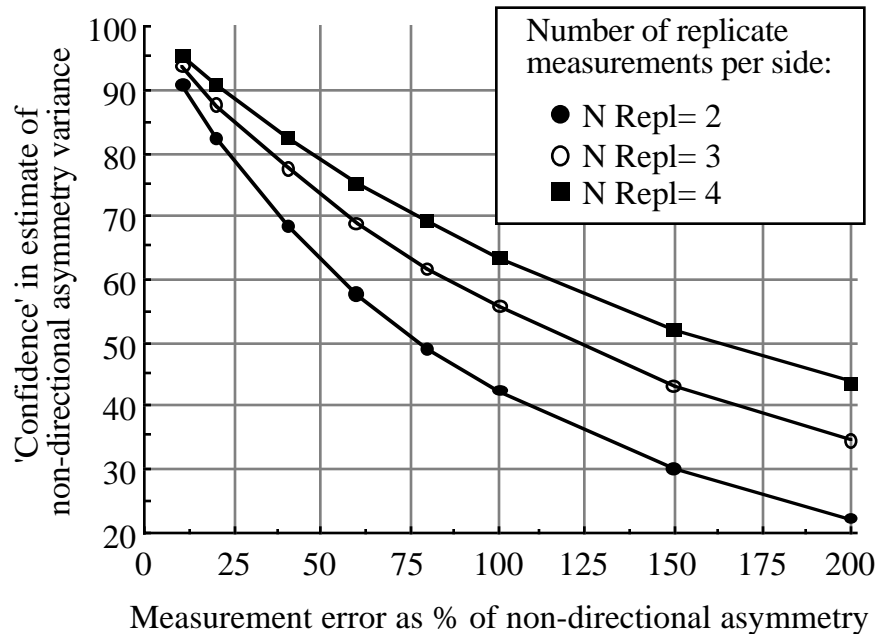
Fig. 8. Impact of measurement error on the 'confidence' with which the magnitude of non-directional asymmetry (FA, antisymmetry, etc.) may be estimated. X axis= 100 * (σ_m^2/σ_i^2) from Table 2b, Y axis= 100 * [(approximate df for σ_i^2) / (df for number of genotypes, J - 1)] from Table 2d.

Because of the bias it introduces into studies of FA, accurate estimates of measurement error are essential.

6.2 Error in meristic traits

In principle, meristic traits can be counted without error. In practice, error can arise from several sources. First, unless counting is conducted by an image analysis system, human observers invariably make some mistakes. These may be sufficiently infrequent that they can be ignored. However, when beginning a study, one sample should be counted at least twice to assess how frequent such errors are.

Commonly, establishing the last element in a series of meristic traits that decrease in size may require a subjective assessment (Fig. 9a, b), as may establishing whether a meristic element falls inside or outside an arbitrarily defined area. This subjectivity can lead to an error that is no different from the measurement error of metrical traits. If subjectivity is involved when counting meristic traits, an assessment of the magnitude of 'measurement' error is essential (Sect. 6.7).

Measurement error for meristic traits:

a) Bristles, setae or fin rays: b) Scales:

Measurement error for metrical traits:

c) Measurement error ONLY (◆ , □):   d) 'True' FA variation between sides (↔ , ■):

Fig. 9. Examples of measurement error for meristic and metrical traits. (a,b) errors due to subjective decisions about meristic traits, (c) 'artificial' FA created solely by measurement error,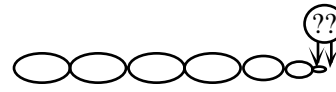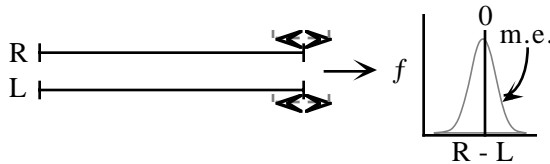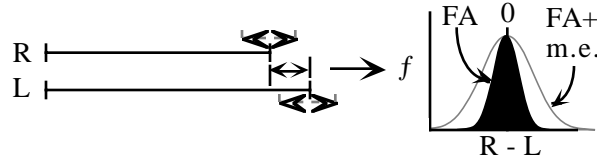 (d) a combination of 'true' FA and measurement error. m.e.- measurement error, R- trait on right side, L- trait on left side. Note that measurement error should be expressed relative to FA rather than to trait size [(R + L)/2].

6.3 Error in metrical traits

With metrical traits, measurement error is unavoidable and can seriously bias estimates of FA (Sect. 6.1). Measurement error can create the illusion of FA, even where it is negligible (Fig. 9c), or it will artificially inflate estimates of FA (Fig. 9d). Furthermore, Fig. 9d illustrates how meaningless it is to report measurement error as a percent of overall trait size [e.g. as a % of (R+L)/2]. Clearly what matters in studies of FA is measurement error as a percent of between-sides variation (compare solid vs. dashed lines with arrows, Fig. 9d).

Ideally, all metrical traits should be measured at least twice. However, where sample sizes are large, or where measurements are difficult to obtain, repeated measurements of all traits may not be feasible. In these cases, tests for the impact of measurement error (Sect. 6.5) should be conducted for each trait on a subsample of at least 30 individuals.

6.4 Quantizing error in image analysis systems

Video image analysis systems (e.g. Morphosys, Jandel) offer a seductively simple, and seemingly more reliable method for quantifying the size and shape of structures. For conventional morphometric studies they can be quite useful. For studies of FA, however, great care must be taken to determine the size of the pixellation error relative to the average difference between sides.

Pixellation error is a problem for studies of FA because most imaging systems have rather low resolution (on the order of 600 x 600 pixels per image). When a structure spans 200 pixels, the pixellation error would seem to be quite small (0.5%). However, since the average difference between sides in FA studies is on the order of 1% of trait size, pixellation error can make up a substantial fraction of the average between-sides difference.

6.5 Recommended procedure for conducting repeated measurements

To obtain a totally unbiased estimate of measurement error, repeat measurements should be conducted 'blind' (i.e. without reference to earlier measurements). Furthermore, since subjective 'preferences' of the observer may vary with time, repeat measurements are more reliably conducted one or more days apart. Given the bias that measurement error introduces (Sect. 6.1), the more reliable the estimate of measurement error, the more reliable will be the estimate of FA. Thus,

where replicate measurements are to be taken, the most reliable estimate of measurement error will be obtained if one complete set of measurements is finished, and several days allowed to elapse, before the repeat set of measurements is started. Ideally, the time between replicate measurements should be roughly equal to the total elapsed time spanned by one complete set of measurements of all treatment groups.

6.6 Tests for the significance of FA relative to measurement error in metrical traits

Palmer & Strobeck (1986) describe a test for determining whether the between-sides variation is significantly larger than the measurement error. The test is a simple two-way ANOVA (sides x individuals, Table 2b), and should be conducted routinely as a part of any study of FA. If the interaction variance is not significant ($\underline{MS}_{sj}$ of Tables 2b,c), then tests for FA differences among samples are not justified.

A crucial point to keep in mind is that this ANOVA procedure tests for the significance of all between-sides variation relative to measurement error, including antisymmetry and FA (Table 2c). In other words, this test asks: Does the difference between sides vary more AMONG GENOTYPES (=individuals) than would be expected, given the size of the measurement error? Hence the ANOVA procedure can not establish whether the between-sides variation arises due to antisymmetry or to FA. Some other procedure must be conducted to confirm that the observed between-sides variation is statistically indistinguishable from FA (Sect. 8.0).

If the interaction variance ($\underline{MS}_{sj}$) is significantly greater than the measurement error ($\underline{MS}_m$), then measurement error may be partitioned out to yield a more accurate estimate of the underlying between-sides variance (σ_i^2, Table 2d). Of all the different indices for FA, only FA10 (σ_i^2 of Table 2d) provides an estimate of the between-sides variance after removing the effects of measurement error.

Table 2. ANOVA procedure for testing the significance of FA relative to measurement error. This test also simultaneously tests for the presence of DA and for trait-size differences among individuals.

Table 2a) Data layout

Individual #	Grouping Var. Side	Repl.	Trait 1	Trait 2	etc.
A1	R	1	M1	M1	
A1	R	2	M2	M2	
A1	L	1	M1	M1	
A1	L	2	M2	M2	
A2	R	1	M1	M1	
A2	R	2	M2	M2	
A2	L	1	M1	M1	
A2	L	2	M2	M2	
A3	R	1	M1	M1	
A3	R	2	M2	M2	
A3	L	1	M1	M1	
A3	L	2	M2	M2	
etc.					

Layout of analysis for 1 trait

Individual (random)	Side (fixed) Right	Left
A1	M1	M1
	M2	M2
A2	M1	M1
	M2	M2
A3	M1	M1
	M2	M2
etc.		

Table 2b) Variance components in a mixed model, two-way ANOVA*

Source of variation	Mean squares label	df	Expected mean squares	Mean squares used to test for
Sides	MS_s	$(S\text{-}1)$	$\sigma_m^2 + M(\sigma_i^2 + (\frac{J}{S-1})\sum\alpha^2)$	Directional asymmetry
Genotypes**	MS_j	$(J\text{-}1)$	$\sigma_m^2 + M(\sigma_i^2 + S\sigma_j^2)$	Size/shape variation§
Side x Geno.	MS_{sj}	$(S\text{-}1)(J\text{-}1)$	$\sigma_m^2 + M\sigma_i^2$	Nondirectional asymmetry†
Meas. error	MS_m	$SJ(M\text{-}1)$	σ_m^2	

* From Table 3b of Palmer & Strobeck (1986). $S=$ number of sides, $J=$ number of 'genotypes' (='individuals'), $M=$ number of replicate measurements, $\sum\alpha^2=$ added fixed variance component due to 'sides' (i.e. directional asymmetry), $\sigma_j^2=$ added random variance component due to 'genotypes' (='individuals'; reflects variation in overall trait size or shape among individuals), $\sigma_i^2=$ added random variance component due to non-directional asymmetry (reflects variation in the between-sides difference among individuals such as FA and antisymmetry), $\sigma_m^2=$ random variance component due to measurement error.

** 'Genotypes' refers to genetically unique individuals and may be considered as 'individuals' in most cases. Where clonemates are available, a more detailed analysis is possible that can distinguish between two forms of antisymmetry (see Table 3a of Palmer & Strobeck, 1986).

§ Indicates variation in overall trait size or shape among individuals; this may be useful for testing the effectiveness of size scaling.

† Includes <u>all</u> forms of non-directional asymmetry including FA, antisymmetry and NCA.

Table 2c) Significance tests of results from ANOVA (abbrev. as in Table 2b)

Significance test	F ratio	degrees of freedom numerator	denominator
Nondirectional asymmetry (i.e. FA if antisymmetry is absent) relative to measurement error	$\underline{MS}_{sj} / \underline{MS}_m$	$(\underline{S}\text{-}1)(\underline{J}\text{-}1)$	$\underline{SJ}(\underline{M}\text{-}1)$
Directional asymmetry	$\underline{MS}_s / \underline{MS}_{sj}$*	$(\underline{S}\text{-}1)$	$(\underline{S}\text{-}1)(\underline{J}\text{-}1)$
Variation due to overall trait size or shape among genotypes (=individuals)	$\underline{MS}_j / \underline{MS}_{sj}$*	$(\underline{J}\text{-}1)$	$(\underline{S}\text{-}1)(\underline{J}\text{-}1)$

* Note that the 'error' term from this ANOVA yields no information about the variation between sides, it simply reflects measurement error. Only the sides x genotypes (=individuals) interaction contains information about between-sides variation. Hence the test for directional asymmetry asks: is the mean difference between sides greater than expected given the amount of non-directional asymmetry, as reflected in $\underline{MS}_{sj}$? The same logic underlies the test for variation due to trait size or shape.

Table 2d) <u>Partitioning measurement error variance from between-sides variance</u> (abbrev. as in Table 2b)

The total between-sides variance ($\underline{MS}_{sj}$) includes a contribution from both measurement error and from all forms of non-directional asymmetry. A better estimate of the true between-sides variance (σ_i^2) may be obtained by partitioning measurement error out of the sides x genotypes (=individuals) mean squares of the ANOVA results (Palmer and Strobeck, 1986):

non-directional asymmetry =   (between-sides MS - meas. error MS) / # replicate meas.

$$\sigma_i^2 = (\underline{MS}_{sj} - \underline{MS}_m) / \underline{M}$$

approx. degrees of freedom for $\sigma_i^2 = \dfrac{\left(MS_{sj} - MS_m\right)^2}{\left(\dfrac{\left(MS_{sj}\right)^2}{(S-1)(J-1)} + \dfrac{\left(MS_m\right)^2}{S\,J\,(M-1)}\right)}$

Note that as the magnitudes of measurement error and between-sides variance become more similar, the approximate degrees of freedom for σ_i^2 decrease. In other words, the larger the measurement error, the lower the certainty in the estimate of the between-sides variance (see Fig. 8).

<u>6.7 Tests for the significance of FA relative to counting error in meristic traits</u>

Unfortunately, for meristic traits, it is less clear how to test for the significance of non-directional asymmetry relative to counting error. Where the differences between sides are large, for example an average of four or five, the ANOVA technique (Sect. 6.6) should probably work adequately. However, where the true difference between sides is only one or two, and a low level of scoring error may exist, no formal tests have been suggested. A likelihood ratio or g-test (Sokal and Rohlf, 1981) comparing the counts in each asymmetry class (e.g. -1, 0, +1) vs. the counts in each class of deviations between replicates (e.g. -1, 0, +1) might be used to yield an estimate of the statistical significance of the between-sides variation relative to counting error.

7.0 Directional Asymmetry

<u>7.1 Why test for DA in studies of FA?</u>

Tests for the presence of DA should be conducted in studies of FA for two reasons: 1) the presence of DA artificially inflates the values of certain FA indices (FA1, FA2, FA3, FA5, Table 1; Sect. 3.2), and 2) if a trait exhibits DA, some portion of the between-sides variation may have a genetic basis, hence the between-sides variance may not purely be a product of developmental noise (Palmer and Strobeck, 1992). Although the biasing effect on FA indices may be removed statistically by 'correcting' for DA, the possibility of heritable variation for DA remains and can not be eliminated without a formal heritability analysis. Hence, if other traits are available that do not exhibit DA, these will be preferred because they raise fewer concerns about inferences based on patterns of bilateral variation among samples

<u>7.2 Tests for DA</u>

Many tests are possible and appropriate for DA (Table 3), since the question is simple: <u>Is one side significantly larger than the other on average?</u> One advantage to the factorial ANOVA procedure (Table 3d; Sect. 6.6) is that the significance of DA may be tested at the same time as that of FA relative to measurement error, hence this procedure is the recommended one. Care should be taken to apply the Sequential Bonferroni test (Sect. 11.0) when conducting multiple tests for DA.

Table 3. Tests for significance of directional asymmetry.

Test	Description of test
a) one-sample t-test	tests for a departure of the mean of (R - L) from an expected mean of zero
b) paired t-test	tests the consistency of the direction and magnitude of (R - L) among individuals
c) nested ANOVA (individuals within sides)	tests for a significant difference between mean R and mean L in a sample of individuals
d) factorial ANOVA (Table 2) (sides x individuals)	tests for a significant difference between mean R and mean L in a sample of individuals relative to the the between-sides variation <u>after accounting for measurement error</u>; also tests significance of non-directional asymmetry and overall trait size variation among individuals.

8.0 Departures from normality (e.g. antisymmetry and skew)

8.1 Why test for departures from normality in studies of FA?

Frequency distributions of R - L may depart from normality in several ways, including skew (a long tail on one side), leptokurtosis (narrow peaked & long-tailed) or platykurtosis (broad peaked & short-tailed, or bimodal). Traits exhibiting significant departures from normality should be viewed with skepticism as valid measures of developmental stability because they do not conform to the pattern of bilateral variation expected due to developmental noise. Because developmental noise is most precisely viewed as the cumulative consequence of subtle, random variation in perhaps several developmental processes that affect each side independently (Sect. 2.2), it should generate perfectly normal distributions of R - L. Although processes considered to be developmental noise may potentially give rise to non-normal distributions of R - L (Palmer and Strobeck, 1992), processes <u>not</u> considered to be developmental noise could also give rise to such distributions. Hence, unless alternative explanations can be ruled out, it seems safest to exclude traits exhibiting significant departures from normality from studies of developmental stability.

Tests for the presence of antisymmetry should be conducted in studies of FA for the same two reasons as tests for DA (Sect. 7.1): 1) antisymmetry artificially inflates the values of <u>all</u> FA indices (FA1 - FA10, Table 1; Sect. 3.2), and 2) if a trait exhibits antisymmetry, some portion of the between-sides variation may have a genetic basis, hence the between-sides variance may not purely be a product of developmental noise (Palmer and Strobeck, 1992). Unlike DA, however, no statistical 'corrections' for antisymmetry have been suggested. In addition, estimates of the between sides variance based on traits that exhibit antisymmetry can not be compared quantitatively with those that exhibit ideal FA (Sect. 4.5). As for DA, if other traits are available that do not exhibit antisymmetry, these will be preferred because they raise fewer concerns about inferences based on patterns of bilateral variation among samples.

The relevance of skewed or leptokurtic (long-tailed) distributions of R - L to studies of FA is somewhat less clear, though they may also signal the presence of a genetic component to subtle asymmetries (Palmer and Strobeck, 1992). Until more is known about their developmental significance, traits exhibiting significant skew or leptokurtosis should be avoided.

Note, however, that a common source of skew or leptokurtosis in studies of FA is statistical outliers. For example, the R - L difference in one or two individuals in a large and

otherwise normally distributed sample may deviate unusually far from zero. Such deviations could arise due to extreme measurement errors, or they might reflect truly unusual deviations from symmetry due to prior injury or trauma. If measurement error is suspected, outlying individuals could simply be remeasured. Alternatively, if injury or trauma are suspected, such individuals can legitimately be excluded from samples used to estimate FA, because injury or trauma are not normally considered 'developmental noise' (Sect. 2.2).

8.2 Tests for departures from normality: general comments

Numerous tests for departures from normality are available but their statistical power varies depending on the form of departure (Shapiro, et al., 1968). After a review of prior studies, Palmer & Strobeck (1992) suggested that conventional skew and kurtosis statistics provide the most useful descriptions of and tests for departures from normality because a) they are readily computed by most commercial statistical packages, b) their standard errors are known (Table 4), and c) these two statistics taken together provide additional useful descriptors of the form of between-sides variation. A visual inspection of frequency distributions of R - L may also be particularly helpful when attempting to interpret apparent departures from normality.

Table 4. Descriptors of departures from normality, their standard errors and interpretation (Sokal and Rohlf, 1981). See Sokal & Rohlf (1981) pg. 114 for computation of the skew and kurtosis statistics themselves.

Descriptor	Standard error (s)	Large sample approximation (n> 150)	Interpretation of departure from normality
skew (g_1)	$\sqrt{\dfrac{6n(n-1)}{(n-2)(n+1)(n+3)}}$	$\sqrt{\dfrac{6}{n}}$	negative= long tail to left (skewed to L) positive= long tail to right (skewed to R)
kurtosis (g_2)	$\sqrt{\dfrac{24n(n-1)^2}{(n-3)(n-2)(n+3)(n+5)}}$	$\sqrt{\dfrac{24}{n}}$	negative= platykurtic, broad or bimodal peak and short tails positive= leptokurtic, narrow peak and long tails

NOTE ADDED 12/6/96: The SE for kurtosis (as published in most textbooks) assumes a symmetrical distribution of g_2. For small sample sizes (<100), g_2 is strongly skewed to the right (it can never be less than -2.0). Therefore the above SE yields a greatly exaggerated type II error when testing for platykurtosis of small samples.

As with all such testing, care should be taken to apply the Sequential Bonferroni test (Sect. 11.0) when conducting multiple tests.

8.3 Tests for departures from normality in either meristic or metrical traits

Since for a truly normal distribution the expected values for both skew (g_1) and kurtosis (g_2) are zero and the degrees of freedom infinite (Sokal and Rohlf, 1981), the significance test is a straightforward one-sample t-test (see Table 4 for standard errors and interpretations of departures from normality):

skew $t_s = g_1 / s_{g1}$ (compare versus values of t for ∞ df)
kurtosis $t_s = g_2 / s_{g2}$ (compare versus values of t for ∞ df)

Unfortunately, the power to detect departures from normality with these statistics drops off radically as sample size decreases, hence other tests may be more useful for small samples (<30; Sect. 8.4).

8.4 Tests for departures from normality in small samples of metrical traits

The Kolmogorov-Smirnov Test offers a useful non-parametric test for departures from normality in small samples (Sokal and Rohlf, 1981). Unfortunately, because of the 'steppiness' of the frequency distribution of R - L for most meristic traits, this test will yield many 'false' positive results if the number of asymmetry categories is small (< 10). Thus it is not very useful when applied to meristic traits. For metrical traits, however, the variation in R - L is continuous and hence the test may be applied with more confidence. See Sokal & Rohlf (1981, pp. 716 - 721) for worked examples of this test.

9.0 What to do when traits depart from ideal FA

For the reasons outlined above (Sects. 7.1, 8.1), if traits are found to exhibit DA, or departures from normality, some care should be taken to show that inferred differences in developmental stability (however measured) among samples are not confounded by those traits that depart from 'ideal' FA (Sect. 4.5). The principle concern is that both DA and antisymmetry can bias certain FA indices, and thus render them unreliable as measures of developmental stability (Sects. 7.1, 8.1). If necessary, the presence of DA may be removed statistically by subtracting (mean(R-L))/2 from the side with the larger mean and adding it to the smaller side of all individuals in a sample. However, concerns remain about the validity of inferring levels of developmental stability using such 'corrected' variation in R - L (Palmer and Strobeck, 1992). Regrettably, no methods have yet been suggested for statistically 'correcting' for the presence of antisymmetry or skew in distributions of R - L.

Should traits be found to depart from 'ideal' FA (Sect. 4.5), and should it not be possible to complete a study without incorporating such traits in the final analyses, several questions should be asked to determine if such traits might be biasing any conclusions:

Do traits exhibiting DA, antisymmetry or skew appear to reveal greater differences in developmental stability among samples than those exhibiting ideal FA (Sect. 4.5)?

Do traits exhibiting greater DA, antisymmetry or skew appear to reveal a greater range in developmental stability among samples than those exhibiting lesser DA, antisymmetry or skew?

Does the magnitude of developmental stability correlate with the magnitude of DA, antisymmetry or skew among samples?

If the answer to any of these questions is yes, then it may not be possible to say with certainty that differences in the between-sides variance among samples are due to differences in developmental stability.

10.0 Size dependence of FA

10.1 Why is size-dependence a concern in studies of FA?

A dependence of asymmetry on trait size can influence inferences made in studies of developmental stability in three ways: 1) if the magnitude of asymmetry depends on trait size within samples, and some samples have a greater size range than others, spurious differences in FA may arise among samples (i.e. samples with a greater size range will appear to exhibit higher FA), 2) if FA depends on mean trait size among samples, and mean size varies among samples, inferred differences in developmental stability among samples may be an artifact of size variation (i.e. samples with a larger mean size will appear to exhibit higher FA), and 3) if FA depends on mean trait size when comparing different traits, inferred differences in developmental stability among traits may be an artifact of size differences. At the same time, arbitrary 'correction' for

presumed size-dependence can lead to spurious differences among samples (Sect. 3.4). For example, of samples differ in average body size, and the variance of R - L is independent of trait size, then inappropriate correction for size effects will create artificial differences among samples (i.e. those with larger mean size will appear to exhibit artificially lower FA). Furthermore, differences in overall size among samples may partly reflect differences in general 'condition', which itself is often correlated with FA (Zakharov, 1992). Hence, eliminating all size-dependence of FA among samples could partially or completely obscure associations between FA and condition (Palmer and Strobeck, 1986).

Among related traits the variance often increases with the mean, which has lead to considerable discussion about how best to describe relative variation (Lande, 1977; VanValen, 1978). Thus, individuals should ideally be selected in such a manner that mean trait size does not differ among samples. Where this is unavoidable, some form of size-correction may be necessary. The question thus is: Does asymmetry vary with size? If, for the size range examined, an association is weak or absent using unscaled indices (e.g. FA1, FA4, FA5 or FA10), no size correction is needed. If the unscaled index varies significantly with size, then some correction may be required.

Never assume that size-corrected FA indices remove all size effects. Regardless of what kind of size corrections are considered or employed, there is no substitute for visual inspections of scatterplots of asymmetry versus 'size', both before <u>and</u> after conducting any size correction.

10.2 Tests for size dependence of FA within samples

Where size differences exist mainly within samples, analyses of the size-dependence of FA are less clear cut. Note that a regression of (R - L) against some independent measure of body size will yield a zero slope even though the variance of (R - L) clearly depends on size (Fig. 10a). A more appropriate regression is |R - L| against an independent measure of body size (Fig. 10b). If a significant association remains, a scaled asymmetry index may be in order (e.g. FA2, FA6 or FA8 of Table 1).
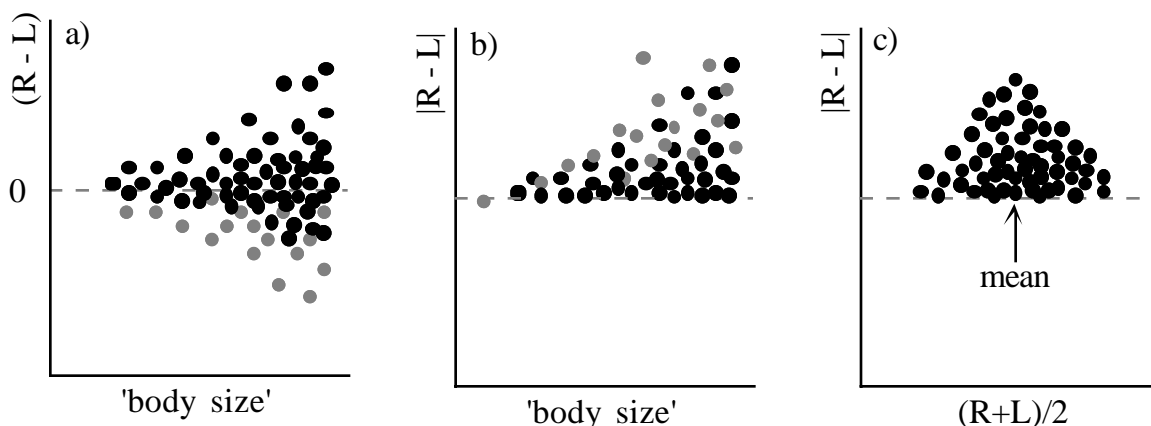
Fig. 10. Examples of size-dependence of asymmetry within a sample. a-b) asymmetry versus some independent measure of 'body size'. c) asymmetry vs. trait size for a single trait exhibiting FA, in the absence of any 'body size' variation (i.e. all apparent variation in trait size is due solely to random variation in each side independently).

Scaling |R - L| by (R+L)/2 for the same trait should be done with care. If body size variation within a sample is small, and asymmetry variation large, much of the apparent variation in (R+L)/2 will arise solely from departures from symmetry (Fig. 10c).

10.3 Tests for size dependence of FA among samples

Where size differences exist mainly among samples, a simple linear regression or visual inspection of plots of log(var(R-L)) vs. mean((R+L)/2) should reveal any size-dependence. Log (var(R-L)) is preferred here because the variance increases as a square of the average difference between sides. If a significant association is detected, some correction may be warranted (e.g. FA3 or FA7 of Table 1). However, if body size is also correlated with the factor under investigation (e.g. stress or heterozygosity), such a correction may mask an association with FA (Palmer and Strobeck, 1986).

11.0 Care when conducting multiple tests

Where a large number of traits or a large number of samples have been analyzed in a single study, care must be taken not to be mislead by 'false' significant results. Recall that a significance level of $P < 0.05$ means that, on average, one out of every 20 tests with completely random data will yield a statistically 'significant' result. The Sequential Bonferroni correction provides a useful guard against such 'false' positive tests. It should be applied to each set of related tests (e.g. all tests for DA or all tests for antisymmetry, etc., in a single study). This procedure is explained and justified in more detail by (Rice, 1989).

The Sequential Bonferroni test ensures the appropriate table-wide probability of type I error (i.e. 'false' positives) and is very simple to conduct (Table 5):

   i)   Rank all the original test results from most significant to least significant.
   ii) Multiply the lowest original P value (P_o) by the total number of tests in the table.
   iii) If the revised P value (P_r) is significant, proceed to the next higher P value in the table.
   iv) Multiply the next lowest original P value (P_o) by the number of tests in the table, minus the number of prior Bonferroni tests up to that point.
   v)   Repeat steps (iii) and (iv) until reaching the first revised P value (P_r) that is <u>not</u> significant. All lower ranked tests beyond this test would are considered not significant.
   vi) Cease computing revised P values. In other words, even though tests #5 and #8 (Table 5) would appear to be significant at the 0.05 level, these would not be considered significant by the Sequential Bonferroni procedure.

Table 5. Illustration of the Sequential Bonferroni procedure to ensure an appropriate table-wide likelihood of 'false' positives (i.e. probability of Type I error at the level of the whole table is at the desired α, normally 0.05); modified slightly from Rice (1989). N= total number of tests in a given table (8 in this example).

Test #	Initial Probability P_o	Bonferroni correction	Revised Probability P_r	Significant at P= 0.01	Significant at P= 0.05
1	0.001	$P_o * N$	0.008	yes	yes
2	0.002	$P_o * (N-1)$	0.014	no	yes
3	0.007	$P_o * (N-2)$	0.042	no	yes
4	0.011	$P_o * (N-3)$	0.055	no	no
5	0.012	$P_o * (N-4)$	0.048	no	no
6	0.031	$P_o * (N-5)$	0.093	no	no
7	0.037	$P_o * (N-6)$	0.074	no	no
8	0.042	$P_o * (N-7)$	0.042	no	no

12.0 Adequate presentation of descriptive data

Anyone who has ever attempted to extract quantitative estimates of FA from the literature for comparison among studies will realize how hopelessly inadequate most presentations of descriptive data are. In some cases, either the number of samples, or the number of traits is too

large to permit a full description. However, in most cases, the authors seem mostly concerned with the qualitative patterns of their own study rather than with a quantitative presentation of their data in a way that makes it usable by others. In view of the extraordinary effort required to collect a respectable set of FA data, and in view of how little we understand about the developmental origins and implications of variation in subtle asymmetries, a detailed presentation of descriptive data when results are published should be a high priority.

A detailed presentation of data should ideally include several descriptors for each trait in each sample (Table 6). The more complete the description of bilateral variation, the more useful and informative the comparisons among published studies will be. The logic behind including these statistics includes:

mean±SE of trait size [(R+L)/2]- to assess dependence of FA indices on trait size among samples and among traits (Sect. 10.0); also indicates the range of trait size variation within a trait in a given sample.

mean±SE of the difference between sides (R - L)- indicates the presence and significance of any directional asymmetry (Sect. 7.0), also the SE may be easily converted to FA4.

skew and kurtosis of frequency distribution of R - L- to assess the magnitude of variation of skew and kurtosis among samples and among traits (Sect. 8.0).

two (ideally three) separate FA indices:
  a) FA1- indicates the average difference between sides, |R-L|, in easy to understand units,
  b) FA4- (the SE of (R - L) may easily be converted to FA4 just by knowing the sample size)
  c) FA10- σ_i^2 provides an estimate of FA after measurement error has been partitioned out (Table 2d, Sect. 6.6).

estimate of measurement error- critical for knowing what fraction of the bilateral variation is due to measurement error and for comparing values of FA quantitatively among studies (Sect. 6.6).

dependence of |R - L| on size within samples- confirms presence/absence of size dependence of subtle asymmetries among individuals within a sample (Sect. 10.0).

Table 6. A recommended 'complete' presentation of descriptive data in studies of FA.

Sample label	Trait	N	(R+L)/2 Mean±SE	Slope±SE¶	(R - L) Mean±SE*	Skew±SE†	Kurtosis±SE†	|R-L|=FA1 Mean±SE	FA10§ MS_m	σ_i^2	df
1	A										
	B		<---- computed AFTER taking average of replicate measurements - - - - ->						(uses replicate		
	C								measurements		
	D								to partition		
2	A								out meas.		
	B								error)		
	C										
	D										
etc.											

¶ Slope ± SE from regression of |R - L| or log|R - L| versus some measure of size (either organism size or trait size, Sect. 10.2).

* for mean (R-L), FA4= N * SE^2.

† SE optional since it may be computed from the sample size (Table 4, Sect. 8.3).

§ MS_m = measurement error mean square, σ_i^2= non-directional asymmetry, df= approx. degrees of freedom for non-directional asymmetry after partitioning out measurement error (Table 2d, Sect. 6.6).

13.0 Significance tests for differences in FA

13.1 Between two samples

Where frequency distribution of R - L have been confirmed to be mean=0, normal (Sects. 7.0, 8.0), the most powerful test for the significance of the difference in FA between two samples is an F-test (Lehmann, 1959) using FA4, FA5, FA6, FA7, FA8 or FA10 (Table 1). All these indices are variances, and an F-test is simply a ratio of the larger over the smaller variance. The significance of this ratio need only be looked up in a statistical table for the appropriate degrees of freedom.

13.2 Among three or more samples

Tests for differences in FA are effectively tests of whether variances are significantly heterogeneous among samples. Many tests for heterogeneity of variances are available, but they vary in their statistical power and their sensitivity to departures from normality (VanValen, 1978; Conover, et al., 1981). After a detailed comparison of four common tests (Bartlett's, F_{MAX}, Levene's and Scheffé's), Palmer & Strobeck (1992) concluded that Levene's test may be the most practical for studies of FA for the following reasons: 1) it was the least sensitive to departures from normality in the direction of leptokurtosis or platykurtosis and hence more likely to yield valid P values where such departures are small, 2) it was only slightly less powerful than Bartlett's and F_{MAX}, both of which were quite sensitive to departures from normality, and 3) it is extremely easy to conduct on commercial statistics packages, even those that do not formally incorporate the test.

Levene's test is disarmingly easy to conduct: it is simply a one-way ANOVA (i.e. a comparison among samples) conducted with the unsigned asymmetry values, |R - L|. Thus in a data file where |R - L| have already been computed to estimate FA1 or FA2, these same data may be used in an ANOVA, with 'samples' as a grouping variable, to test for differences in FA among samples. Even though these data are not normally distributed (they are truncated normal, and hence skewed to the right; Fig. 3a), this test yields very close to the proper Type I error when no differences exist among samples (Palmer and Strobeck, 1992). However, when all the variation in R - L within samples is normal, Levene's test does lose on the order of 10 - 15% statistical power compared to Bartlett's and F_{MAX} tests when the differences among variances are slight.

Another potential advantage to using Levene's test is that a posteriori tests may be conducted among samples if an overall significant difference is detected. Thus, conventional multiple-comparisons tests can identify subsets of samples among which variation is not significantly different. In addition, where comparisons among subsets of data within a larger analysis are desirable and identifiable a-priori, one may conduct a formal analysis of contrasts (Sect. 9.6 of (Sokal and Rohlf, 1981). Levene's test thus provides a robust and versatile method for analyzing FA differences among samples.

13.3 Among samples using multiple traits simultaneously

Since FA variation is often quite subtle, and since the FA of individual traits may not differ much among samples, additional statistical power may be gained by combining information about FA variation from several traits. Here too, a variation on Levene's test (Table 7) may offer an easy and effective way to combine information from several traits: a two-way ANOVA (samples x trait) of variation in |R - L| would yield a single test for overall significance of FA differences among samples. Such an analysis would also yield, via the interaction MS, a test of how consistent traits were at revealing among-sample differences in FA (i.e. a significant interaction term would indicate that some traits were better than others at revealing presumed differences in developmental stability among samples, Table 7b).

Table 7. Structure of and results from a test that combines information from several traits to test for differences in FA among samples. It is a modified version of Levene's test for heterogeneity of variances (e.g. FA) among samples: a Model I, two-way ANOVA of variation in |R - L|. R_1-L_1= R - L for individual 1, etc., MS= mean squares, P= probability.

Table 7a) Structure of data

Sample (fixed effect)	Trait (fixed effect)				
	Trait 1	Trait 2	Trait 3	Trait 4	etc.
A	$\|R_1 - L_1\|$ $\|R_2 - L_2\|$ $\|R_3 - L_3\|$ … $\|R_n - L_n\|$	$\|R_1 - L_1\|$ $\|R_2 - L_2\|$ $\|R_3 - L_3\|$ … $\|R_n - L_n\|$	$\|R_1 - L_1\|$ $\|R_2 - L_2\|$ $\|R_3 - L_3\|$ … $\|R_n - L_n\|$	$\|R_1 - L_1\|$ $\|R_2 - L_2\|$ $\|R_3 - L_3\|$ … $\|R_n - L_n\|$	
B	$\|R_1 - L_1\|$ $\|R_2 - L_2\|$ $\|R_3 - L_3\|$ … $\|R_n - L_n\|$	$\|R_1 - L_1\|$ $\|R_2 - L_2\|$ $\|R_3 - L_3\|$ … $\|R_n - L_n\|$	$\|R_1 - L_1\|$ $\|R_2 - L_2\|$ $\|R_3 - L_3\|$ … $\|R_n - L_n\|$	$\|R_1 - L_1\|$ $\|R_2 - L_2\|$ $\|R_3 - L_3\|$ … $\|R_n - L_n\|$	
etc.					

Table 7b) ANOVA results

Source of variation	MS	P	Interpretation if significant*
Traits (T)	MS_t	P_t	Some traits are repeatedly less stable developmentally than others.
Samples (S)	MS_s	P_s	When information from all traits is pooled, the level of developmental stability varies among samples.
Interaction (TS)	MS_{ts}	P_{ts}	Extent of FA variation among samples depends on the trait.
Error (w/in sample)	MS_e		

* All MS tested over MS_e. Assumes that all traits have been shown to exhibit 'ideal' FA prior to ANOVA (Sects. 7.0, 8.0).

## 14.0 Correlations of subtle asymmetries among traits

When examining correlations among subtle asymmetries, one must keep in mind that the terms FA, DA and antisymmetry refer to attributes of a sample of individuals (Sect. 2.1). Since all bilateral traits will depart from perfect symmetry at some level of precision, the only way to recognize the pattern of subtle asymmetry to which an individual belongs is with reference to other individuals in a sample. In addition, the general presumption in studies of FA is that as developmental stability decreases the variance of R-L increases. This does not mean that all individuals, or all traits become more asymmetrical, only that the range of variation in R - L increases. Hence, even a sample with very low developmental stability will still exhibit a frequency distribution of R-L whose mean is zero: i.e. many individuals by chance will still not differ significantly from symmetry (Fig. 7a). In other words, for traits exhibiting FA (assuming it arises due to developmental noise), the expected value of R - L is always zero. The expected value of |R - L| depends on the variance, but even here many individuals will still not differ significantly from symmetry (Fig. 7b).

Once one recognizes that the expected value of R - L is zero for all traits that exhibit ideal FA (Sect. 4.5), some initially puzzling patterns begin to make more sense.

14.1 Between parents and offspring (heritability)

How does one assess the heritability of subtle asymmetries? Conventional heritability analyses (parent-offspring regressions) of signed R - L will yield no information about the heritability of developmental stability (Fig. 11a) for the simple reason that regardless of the size of the departure from asymmetry of the parents, the expected value of signed R - L among the offspring is zero. Thus, although the range of variation among offspring should, on average, be greater from parents that were more asymmetrical (curved dashed lines, Fig. 11a), the apparent heritability would be zero (straight dashed line, Fig. 11a). If, on the other hand, subtle deviations from symmetry per se are heritable, we see quite a different pattern: right-biased parents give rise to right-biased offspring, or left-biased parents give rise to left-biased offspring, but the variation among offspring in all the crosses should remain the same (Fig. 11c).

DEPARTURES FROM SYMMETRY DUE TO DEVELOPMENTAL NOISE
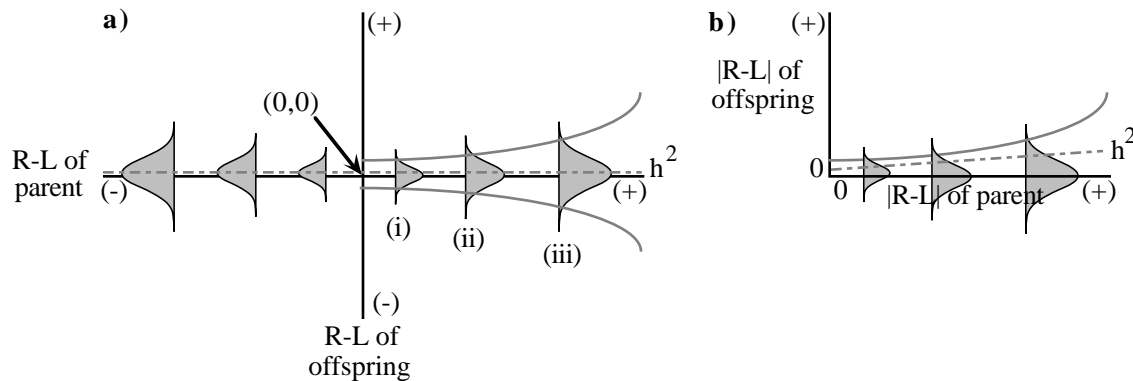
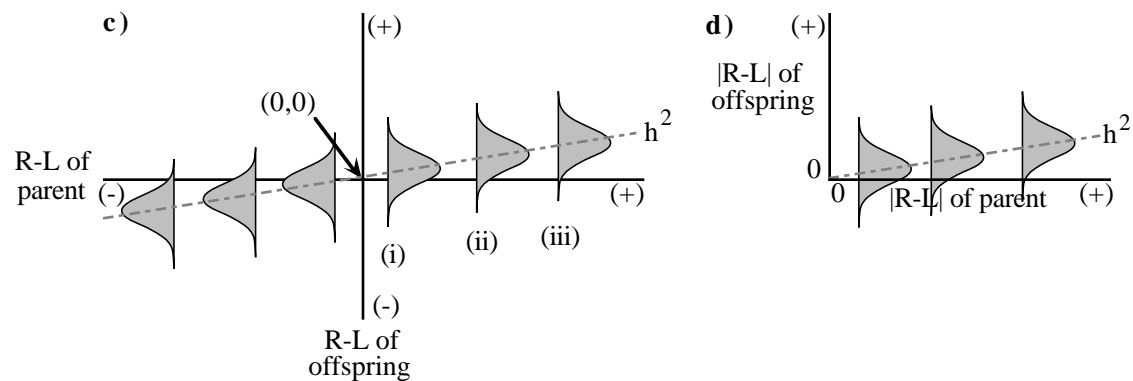DEPARTURES FROM SYMMETRY DUE IN PART TO DIRECT ACTION OF GENES

Fig. 11. Asymmetry correlations between parents and offspring for two cases: deviations from symmetry in an individual are due to developmental noise (a,b) or due in part to the action of genes promoting asymmetry (c,d). For each case, expected associations are illustrated using both signed (R - L) and unsigned |R - L| asymmetry. Frequency distributions represent the expected frequency distribution of asymmetry among offspring from a pair of parents. h^2- estimated heritability (from parent-offspring regression, dashed straight line). Dashed curved lines delineate the range of variation among offspring (a & b only). These figures assume that differences in developmental stability are not due to differences in heterozygosity, but are due to some form of heritable variation for developmental stability (e.g. if extremely homozygous parents were homozygous for different alleles, their offspring would be highly heterozygous).

Since an increase in developmental instability is reflected in an increase in the variance of R - L, the proper form of heritability analysis is a regression of some measure of the <u>variance</u> of R - L among offspring against some measure of the <u>variance</u> of R - L of the parents. Since mean |R - L| is an unbiased estimator of SD(R - L) (Kendall and Stuart, 1951), a regression of offspring |R - L| versus parent |R - L| should yield a reasonable estimate of the heritability of developmental stability (straight dashed line, Fig. 11b). However, a potential problem arises when asymmetries are expressed as absolute values of R - L, since both cases yield a positive slope. In one case, however, this positive slope effectively represents heritable variances (i.e. heritable differences in developmental stability, Fig. 11b), and in the other heritable means (i.e. heritable deviations from symmetry, Fig. 11b).

The different estimates of heritability for (R - L) versus for |R - L| suggest another way in which FA (if due to developmental noise) may be distinguished from DA. For bilateral variation in the form of FA, the estimated heritability of (R - L) should be zero (Fig. 11a) whereas that for |R - L| should be significant and positive (Fig. 11b). On the other hand, for bilateral variation in the form of DA, the estimated heritability should be significant and positive for both (R - L) (Fig. 11c) and |R - L| (Fig. 11d).

More importantly, if a trait exhibiting FA yields a statistically significant heritability of (R - L) (Fig. 11c), this reveals that departures from symmetry in individuals have a heritable basis and hence have <u>not</u> arisen exclusively due to developmental noise. In such a case, FA may <u>not</u> be a valid descriptor of developmental stability. (Palmer, et al., 1993) describe a pattern of bilateral variation where continuous negative covariation may exist between sides (normal covariant asymmetry, NCA). Even though NCA will appear to be ideal FA (Sect. 4.5) when viewed as a frequency distribution of R - L, the negative feedback mechanism implied by such negative covariation between sides may have a heritable basis and thus signal the earliest stages of heritable variation for asymmetry. Such variation may set the stage for the evolution of macroscopic asymmetries (Palmer, et al., 1993).

Finally, given the truncated-normal shape of the distribution of |R - L|, it is not clear how estimates of the heritability of |R - L| should be compared against those for conventional traits, since heritability analyses assume that the distribution of offspring phenotypes more closely resembles those of Fig. 11c.

14.2 Among individuals within samples: Are some individuals more stable developmentally than others?

Why do we so often not see correlations of asymmetries among traits when examining a sample of individuals, when, at the same time, we often do see correlations among traits when examining different samples? The logic used in Sect. 14.1 should make it clear that one should not expect correlations of (R - L) among traits if developmental noise leads to asymmetries in each trait that are independent of those in every other trait. This would be analogous to the pattern in Fig. 11a: even though (R - L) for one trait may depart substantially from zero, the expected value of (R - L) should still be zero for all other traits (e.g. distribution <u>iii</u>).

If signed asymmetries (R - L) do correlate among traits, this may arise via two possible mechanisms. The first would result in some structure-wide or organism-wide DA (analogous to Fig. 11c), where genes or some other factors promoting asymmetry affect several traits in the same individual in the same direction simultaneously. If this mechanism were responsible for the correlation of asymmetries among traits within a sample, such traits would probably exhibit DA when examined independently.

A second mechanism that might account for the correlation of signed asymmetries among traits is developmental noise acting at the level of several traits simultaneously. Unlike the first mechanism, however, all these traits would exhibit ideal FA (Sect. 4.5) when examined independently. Such a pattern would suggest that shared departures from symmetry reflect the

action of developmental noise either a) at a very early stage in ontogeny (e.g. early embryo), before developmental independence is achieved among traits, or b) in substances that regulate the development of several traits simultaneously later in ontogeny. In other words, studies of the correlations of signed asymmetries (R - L) among traits may be very informative about the stage at which suites of traits are influenced by developmental noise.

Since unsigned asymmetry |R - L| <u>does</u> provide an estimate the variance of (R - L) for traits exhibiting ideal FA (Sect. 4.5), one might expect to see correlations of |R - L| among traits within a sample even if all bilateral variation were due to developmental noise. However, one must keep in mind that for a given trait in an individual organism, the unsigned asymmetry |R - L| estimates the variance of (R - L) with only one degree of freedom (i.e. one knows the observed difference between sides, and one assumes that the true mean of R - L is zero). This should be obvious with a little thought: individuals that are very unstable developmentally will nonetheless be symmetrical for some traits some of the time just due to chance (e.g. distribution <u>iii</u> of Fig. 11a). Thus the confidence one has in an estimate of the variance of R - L in an individual (which presumably reflects the level of developmental stability in that individual) is not very high since it is based on a sample with only one degree of freedom. Thus the ability to detect correlations of |R - L| for different traits among individuals within a sample will depend on a) the magnitude of the differences in developmental stability among individuals (e.g. the larger the differences, the greater the likelihood of detecting a correlation) and b) the number of traits measured on each individual (e.g. the larger the number of traits sampled for an individual, the greater the confidence in the estimate of developmental stability of that individual). The widespread lack of such correlations in most studies of FA (Palmer and Strobeck, 1986) may thus be a result of a) limited differences in developmental stability among individuals within a population or b) examining too few traits.

Table 8. Structure of and results from a test that combines information from several traits to test for differences in FA among individuals. It is a modified version of Levene's test for heterogeneity of variances (e.g. FA) among individuals: a mixed Model, two-way ANOVA of variation in |R - L|. R_A-L_A= R - L for individual A, etc., MS= mean squares, P= probability.

Table 8a) <u>Structure of data</u>

Individual (random effect)	Trait 1	Trait 2	Trait 3	Trait 4	etc.
A	$|R_A - L_A|$	$|R_A - L_A|$	$|R_A - L_A|$	$|R_A - L_A|$	
B	$|R_B - L_B|$	$|R_B - L_B|$	$|R_B - L_B|$	$|R_B - L_B|$	
C	$|R_C - L_C|$	$|R_C - L_C|$	$|R_C - L_C|$	$|R_C - L_C|$	
etc.					

(Trait (fixed effect) spans Trait 1–Trait 4)

Table 8b) <u>ANOVA results</u>

Source of variation	MS	P	Interpretation if significant*
Traits (T)	MS_t	P_t	Some traits are repeatably less stable developmentally than others.
Individuals (I)	MS_i	P_i	When information from all traits is pooled, the level of developmental stability varies among individuals.
Interaction (TI)	MS_{ti}	P_{ti}	(used as error term for testing MS_t and MS_i)

* MS_t and MS_i both tested over MS_{ti}, since no error MS is available. This analysis assumes no significant interaction between traits and individuals (e.g. the rank order of the underlying instability of traits is the same for all individuals) and assumes that all traits have been shown to exhibit 'ideal' FA prior to ANOVA (Sects. 7.0, 8.0).

Kendall's Coefficient of Concordance (Sokal and Rohlf, 1981) is commonly used to test for correlated differences <u>among individuals</u> in the levels of asymmetry for several traits. However, the ANOVA test suggested in Sect. 13.3 could also be used here (Table 8). If the effect of 'individuals' was significant (Table 8b), this would indicate that some individuals exhibited consistently higher average asymmetry than others when the information from multiple traits was pooled.

14.3 Among samples

Unlike the situation for correlations among traits within a sample (Sect. 14.2), correlations of FA indices for different traits among two or more samples often are significant (Palmer and Strobeck, 1986). Such correlations among <u>samples</u> are the basis of Soulé's (Soulé, 1967) 'Population Asymmetry Parameter'. Correlations of asymmetry indices for different traits among populations are more likely because the average asymmetry in a sample of individuals is a better indicator of average developmental stability of that sample than is the observed asymmetry in a single individual. In other words, with a sample of individuals, a single trait provides a better estimate of the level of average developmental stability than it does for a single individual.

Kendall's Coefficient of Concordance (Sokal and Rohlf, 1981) is also used to test for correlated differences <u>among samples</u> in the levels of asymmetry for several traits. Here too, the ANOVA test suggested in Sect. 13.3 would also be appropriate. If the effect of 'Samples' was significant (Table 7b), this would indicate that multiple traits exhibited consistently higher FA in some samples than in others.

15.0 Correlations between asymmetry and fitness: Can asymmetry predict mate choice?

Groups of individuals that exhibit differences in fitness sometimes exhibit parallel differences in FA: lower fitness is associated with higher FA (Biémont, 1983; Clarke and McKenzie, 1987; Quattro and Vrijenhoek, 1989). This seductive association has prompted others to test for, and find, an association between the magnitude of subtle asymmetry and mate selection: males more asymmetrical for a particular trait were less preferred by females in both barn swallows (Møller, 1992) and scorpion flies (Thornhill, 1991). For the reasons outlined above (Sect. 14.2), the deviation from symmetry of a single character in an individual should be a very poor predictor of the developmental stability of that individual. So if females are selecting males based on differences in developmental stability (as deduced from subtle differences in asymmetry), and deviation from symmetry in a single trait is a very poor predictor of developmental stability, how can these results be explained?

Significantly, in both cases the traits whose asymmetry correlated with mate choice also directly affected performance. In barn swallows, asymmetry in the outer tail feathers affects maneuverability (Møller, 1991), and in scorpion flies, forewing asymmetry affects foraging efficiency (Thornhill, 1992) and hence (presumably) the volume of pheromone output upon which females based their choice. Hence, the traits examined in these studies were not just arbitrary indicators of developmental stability.

Future attempts to use subtle asymmetries to predict mate choice should either a) measure asymmetry in a large number of traits to ensure an accurate estimate of the developmental stability of individuals, or b) focus on traits whose asymmetry directly affects performance, because one or a few randomly selected traits will be a very poor predictor of developmental stability unless differences in developmental stability among individuals are large (see Sect. 14.2).

16.0 Checklist for studies of FA

_____ Has FA variation been examined in several developmentally independent traits? Because the 'signal' of FA variation is often very weak, and can sometimes be biased in unexpected ways, correlated patterns of FA variation among several traits can greatly strengthen inferences about levels of developmental stability (Sect. 4.0).

_____ If meristic traits have been used, have concerns about the dependence of FA on the proximity of mean (R - L) to whole integer values been adequately addressed? (Sect. 4.2)

_____ Are sample sizes large enough to be able to detect differences in FA among samples? (Sect. 5.0)

_____ Has the measurement protocol been described in sufficient detail? What landmarks were used to measure particular features? Were repeat measurements conducted totally blind (Sect. 6.5)? Did more than one 'observer' take measurements, and if so did the estimate of measurement error include differences among observers? Did any dimensions share a common landmark?

_____ Was measurement error estimated accurately? In other words, did it include all possible sources of variation: temporal variation in observer accuracy, among-observer variation if more than one observer took measurements, variation in ease of measurement for live and dead specimens, digitizing or quantizing error (Sect. 6.5)?

_____ Is FA significantly larger than measurement error in all traits? Because measurement error yields the same pattern of between-sides variation as FA, such a test is essential to all studies of FA (Sect. 6.0). This is perhaps the most common (and important!) oversight in studies of FA.

_____ Do all traits exhibit ideal FA of R - L (i.e. mean zero, normal variation)? If not, have those traits departing from ideal FA in due to DA, antisymmetry or skew been eliminated from the analyses (Sects. 7.0, 8.0)? If they must be included, have tests been conducted to verify they they have not biased FA indices or biased the conclusions about differences among samples (Sect. 9.0)?

_____ Does asymmetry correlate with trait size? If so, have the appropriate corrections for such dependence been applied (Sect. 10.0) and has the size correction been confirmed to eliminate any dependence of asymmetry on size? If not, have the proper FA indices been used (i.e. those that do not correct for trait size; Sect. 3.0)?

_____ Where multiple related statistical tests have been conducted, have the P-values been corrected to reflect this fact? (Sect. 11.0)

_____ Have the data describing bilateral variation among traits and among samples been presented in sufficient detail? Has more than one index of FA been included with these descriptive data (Sects. 3.4, 12.0)?

_____ Have the proper statistical tests been used to test for differences in FA among samples? (Sect. 13.0)

_____ If heritability estimates for asymmetries are presented have the analyses distinguished between heritability of developmental stability vs. heritability of subtle asymmetries? (Sect. 14.1)

_____ If deviations from symmetry have been used as an index for fitness, have a sufficiently large number of traits been used to obtain valid differences among individuals? (Sects. 14.2, 15.0)

_____ In the final paper, have terms referring to patterns (which are observed) been kept distinct from those referring to processes (which are inferred)? (Sect. 2.2)

_____ Have the indices used to estimate FA been identified unambiguously in ALL table and figure legends? (Sect. 3.0)

Acknowledgements

References

Angus, R. A. 1982. Quantifying fluctuating asymmetry: not all methods are equivalent. Growth 46: 337-342.

Angus, R. A., and R. H. Schultz. 1983. Meristic variation in homozygous and heterozygous fish. Copeia 1983: 287-299.

Bader, R. S. 1965. Fluctuating asymmetry in the dentition of the house mouse. Growth 29: 291-300.

Biémont, C. 1983. Homeostasis, enzymatic heterozygosity and inbreeding depression in natural populations of Drosophila melanogaster. Genetica 61: 179-189.

Brown, N. A., and L. Wolpert. 1990a. The development of handedness in left-right asymmetry. Development 109: 1-9.

Chang, K. S. F., F. K. Hsu, S. T. Chan, and B. Chan. 1960. Scrotal asymmetry and handedness. J. Anat. 94: 543-548.

Clarke, G. M., and J. A. McKenzie. 1987. Developmental stability of insecticide resistant phenotypes in blowfly; a result of canalizing natural selection. Nature 325: 345-346.

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

Graham, J. H., D. C. Freeman, and J. M. Emlen. 1993. Antisymmetry, directional asymmetry, and chaotic morphogenesis. Genetica (special issue): ??

Jagoe, C. H., and T. A. Haines. 1985. Fluctuating asymmetry in fishes inhabiting acidified and unacidified lakes. Can. J. Zool. 63: 130-138.

Kendall, M. G., and A. Stuart. 1951. The Advanced Theory of Statistics. Hafner, London. pp.

Lande, R. 1977. On comparing coefficients of variation. Syst. Zool. 26: 214-217.

Leamy, L. 1993. Morphological integration of fluctuating asymmetry in the mouse mandible. Genetica (this volume):

Leary, R. F., and F. W. Allendorf. 1989. Fluctuating asymmetry as an indicator of stress in conservation biology. Trends Ecol. Evol. 4: 214-217.

Leary, R. F., F. W. Allendorf, and R. L. Knudson. 1985. Developmental instability and high meristic counts in interspecific hybrids of salmonid fishes. Evolution 39: 1318-1326.

Leary, R. F., F. W. Allendorf, and R. L. Knudson. 1992. Genetic, environmental, and developmental causes of meristic variation in rainbow trout. Acta Zool. Fenn. 191: 79-95.

Leary, R. F., F. W. Allendorf, R. L. Knudson, and G. H. Thorgaard. 1985. Heterozygosity and developmental stability in gynogenetic diploid and triploid rainbow trout. Heredity 54: 219-225.

Lehmann, E. L. 1959. Testing Statistical Hypotheses. Wiley, New York. 369 pp.

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

McKenzie, J. A., and K. O'Farrell. 1993. Modification of developmental stability and fitness: Malathion-resistance in the Australian sheep blowfly, Lucilia cuprina. Genetica (special issue): ??

Møller, A. P. 1991. Sexual ornament size and the cost of fluctuating asymmetry. Proc. Roy. Soc. Lond. B 243: 59-62.

Møller, A. P. 1992. Female swallow preference for symmetrical male sexual ornaments. Nature 357: 238-240.

Møller, A. P. 1994. Sexual selection in the barn swallow (Hirundo rustica). IV. Patterns of fluctuating asymmetry and selection against asymmetry. Evolution 48: (in press).

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

Palmer, A. R., and C. Strobeck. 1992. Fluctuating asymmetry as a measure of developmental stability: Implications of non-normal distributions and power of statistical tests. Acta Zool. Fenn. 191: 57-72.

Palmer, A. R., C. Strobeck, and A. K. Chippindale. 1993. Bilateral variation and the evolutionary origin of

macroscopic asymmetries. Genetica (in review):

Parsons, P. A. 1990. Fluctuating asymmetry: An epigenetic measure of stress. Biol. Rev. 65: 131-145.

Quattro, J. M., and R. C. Vrijenhoek. 1989. Fitness differences among remnant populations of the endangered Sonoran topminnow. Science 245: 976-978.

Rice, W. R. 1989. Analyzing tables of statistical tests. Evolution 43: 223-225.

Scharloo, W. 1991. Canalization: Genetic and developmental aspects. Ann. Rev. Ecol. Syst. 22: 65-93.

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

Smith, B. H., S. M. Garn, and P. E. Cole. 1982. Problems of sampling and inference in the study of fluctuating dental asymmetry. Amer. J. Phys. Anth. 58: 281-289.

Sokal, R. R., and J. F. Rohlf. 1981. Biometry. 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.

Strawn, K. 1961. A comparison of meristic means and variances of wild and laboratory-raised samples of the fishes, Etheostoma grahami and E. lepidum (Percidae). Texas J. Sci. 13: 127-159.

Swain, D. P. 1987. A problem with the use of meristic characters to estimate developmental stability. Amer. Nat. 129: 761-768.

Taning, A. 1952. Experimental study of meristic characters in fishes. Biol. Rev. 27: 169-193.

Thornhill, R. 1991. Female preference for the pheromone of males with low fluctuating asymmetry in the Japanese scorpionfly (Panorpa japonica: Mecoptera). Behav. Ecol. 3: 277-283.

Thornhill, R. 1992. Fluctuating asymmetry, interspecific aggression and male mating tactics in 2 species of Japanese scorpionflies. Behavioral Ecology and Sociobiology 30: 357-363.

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

VanValen, L. 1978. The statistics of variation. Evolutionary Theory 4: 33-43.

Waddington, C. H. 1940. Organizers and genes. Cambridge Univ. Pr., Cambridge. pp.

Waddington, C. H. 1955. On a case of quantitative variation on either side of the wild type. Zeitschr. Indukt. Abstammun Vererbungslehre 87: 208-228.

Waddington, C. H. 1957. The Strategy of the Genes. George Allen Unwin, London. pp.

Zakharov, V. M. 1992. Population phenogenetics: Analysis of developmental stability in natural populations. Acta Zool. Fenn. 191: 7-30.

Zakharov, V. M., E. Pankakoski, B. I. Sheftel, A. Peltonen, and I. Hanski. 1991. Developmental stability and population dynamics in the common shrew, Sorex-Araneus. Amer. Nat. 138: 797-810.

Zhivotovsky, L. A. 1992. A measure of fluctuating asymmetry for a set of characters. Acta Zool. Fenn. 191: 37-77.