Correlation and Causation in Biological Systems with Applications to Asymmetry

Robin E. Owen
Department of Chemical & Biological Sciences
Mount Royal University
4825 Mount Royal Gate SW
Calgary, Alberta
Canada, T3E 6K6
(403) 440-6167
rowen@mtroyal.ca

Abstract:

Significant correlations between variables in biological systems often occur. However correlation does not imply direct causation as most associations arise through multiple interactions. Students will be introduced to correlation and be able to calculate simple and partial correlation coefficients. They will understand how to distinguish between conspicuous, directional and fluctuating asymmetry. They will gather data on (1) Fiddler Crabs where they will make hypotheses about the correlation between body size and large and small cheliped size in males, and (2) human total fingerprint ridge counts and use an index based on the correlation coefficient to test for fluctuating asymmetry.

Introduction

Associations between variables are common in biological systems. A "natural" measurement of this association is the correlation coefficient, and significant (i.e. non-zero) correlations occur in morphology, ecology and genetics. However correlation does not imply direct causation as most associations between variable arise through multiple causes and interactions. In this laboratory students will gain some appreciation of the conceptual basis of correlation and how to use and interpret the linear correlation coefficient correctly. They will be introduced to the basic theory of correlation and be able to calculate the Pearson Product Moment Correlation Coefficient, compare two correlation coefficients and to be able to calculate and use partial correlation coefficients. They will also be introduced to asymmetry in biological systems, and understand how to distinguish between conspicuous asymmetry, directional asymmetry and fluctuating asymmetry. They will then apply this theory in two laboratory exercises where they will gather data on (1) Fiddler Crabs where they will make hypotheses about the correlation between body size and large and small cheliped size in males. They will test these hypotheses by making measurements on preserved specimens. (2) They will then gather and analyze class data on human total fingerprint ridge counts and use an index based on the correlation coefficient to test for fluctuating asymmetry.

Student Outline

Correlations between variables are common in all areas of biology, for instance morphological measurements on the same organism often show significant (i.e. non-zero) correlations. Similarly the activity or behaviour of animals can often be correlated to environmental or ecological factors. However it is easy to fall into the trap of thinking that correlation proves causation. Although correlation is a very useful tool for investigation in biology, it must be interpreted correctly and with caution. This laboratory exercise will help you to do this.

Objectives

1. To understand the conceptual basis of correlations in biological systems.
2. To understand the use and the limitations of correlation analysis in biology.
3. To learn how to calculate the Pearson Product Moment Correlation Coefficient and to test for its significance.
4. To learn how to compare two correlation coefficients.
5. To understand how partial correlations arise and to be able to calculate partial correlation coefficients.
6. To be able to distinguish between conspicuous asymmetry, directional asymmetry and fluctuating asymmetry.
7. To be able to calculate two indices of fluctuating asymmetry and to understand the role of the correlation coefficient in analyzing asymmetry.

The conceptual basis of correlation

If we have measurements of two variables we are often interested to discover if there is an association between them. An obvious first step is to calculate the correlation coefficient, and this is usually the Pearson Product Moment Correlation Coefficient, although in some circumstances Kendall's Rank Correlation Coefficient may be more appropriate. If we find a significant correlation then this may lead us to hypothesize about a causal relationship between the variables (Sokal and Rohlf, 1981). However the finding of a significant correlation does not imply that one variable is the direct cause of the other; for instance they may both result from a common cause. Variables may be related to one another in a variety of ways, some quite complex. Figure 1, adapted from Sokal and Rohlf (1981), shows some possible ways in which the observed correlation between two variables may arise.

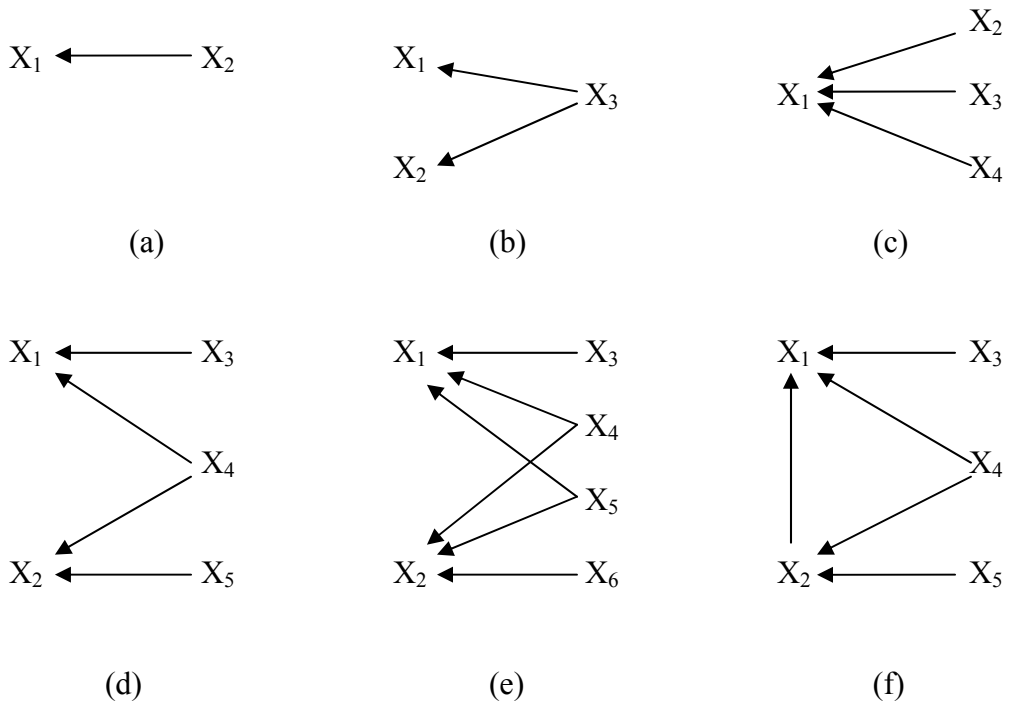

Figure 1. Different explanations for the observed correlation r_{12} between X_1 and X_2 (modified from Sokal and Rohlf, 1981). It is assumed that there are only linear relationships between variables.

The situation shown in Fig. 1 (a), where one variable is entirely determined by one other, is likely to be rare. Even situation (b), where both X_1 and X_2 are completely determined by a common cause is unlikely. In both (a) and (b) $\boldsymbol{r_{12}}$ must equal 1. In reality situation (c) is more likely to apply. An example of this would be the correlation between the age (X_2) and weight (X_1) of an animal (Sokal and Rohlf, 1981). Other factors (X_3, X_4) will inevitably also affect the animal so) $\boldsymbol{r_{12}}$ will be less than 1. Most correlations in nature probably have an even more complex basis as illustrated in (d), (e) and (f).

2. Correlation versus Regression

The correlation coefficient only measures the degree of association between two variables and implies no simple causal relationship between the two (see above). However in many instances, there may be a known or hypothesized functional dependence of one variable on the other. What this means is that the magnitude of one of the variables – the **dependent variable (Y)** – is determined by (is a function of) the magnitude of the second variable – the **independent variable (X)**. In experimental situations the independent variable is controlled completely by the investigator and so the exact effect of incrementally changing the independent variable can be quantified. This is represented by the regression equation: $\hat{Y}_i = a + bX_i$, where $\hat{Y}_i$ is the predicted value of Y, *a* is the Y-intercept and *b* the slope of the regression line (note *a* and *b* are sample estimates of the parameters α and β). We are said to have regressed Y on X. The example shown in Fig. 2 (below) is from Zar (1984). As temperature increases the O_2 consumption decreases in a linear fashion. There obviously must be a direct cause-effect relationship in this case.

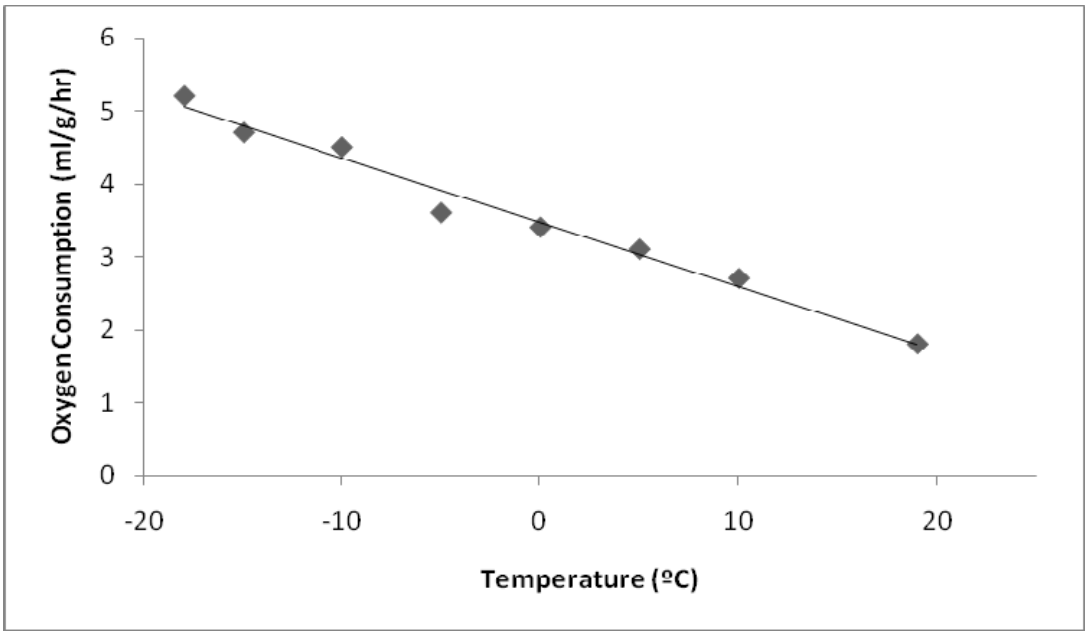

Figure 2. Regression of oxygen consumption of birds (Y) on temperature (X). Data from Zar (1984).

Regression is an extremely useful and widespread statistical technique in biology and in the sciences generally. Nevertheless there are many situations where there is no clear-cut cause-effect relationship between variables and so correlation must be used instead.

3. Measuring correlation

(a) Bivariate Data

If we have paired or **bivariate** data, i.e. a measurement of two variables (X, Y) on an individual, then there may be an association or a **correlation** between them.

The **Linear Correlation Coefficient** (or Pearson Product Moment Correlation Coefficient) *r* measures the strength of the *linear* relationship between the paired X and Y values in a sample. The sample statistic *r* estimates the parameter ρ (*rho*). It is assumed that the variates follow the **bivarate normal distribution**.

(b) The Linear Correlation Coefficient

The correlation coefficient is given by the formula:

$$\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j} \qquad (1)$$

where ρ_{ij} = the (parametric) covariance of X_i and Y_j
ρ_i = the standard deviation of X_i
ρ_j = the standard deviation of Y_j

The corresponding sample statistic is:

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}} \qquad (2)$$

where $\sum xy = \sum (X_i - \overline{X})(Y_j - \overline{Y})$, which is the **covariance** of X and Y. Note that this is how covariance is defined, but you should NOT actually calculate the covariance this way, instead use the computational formula:

$$\sum xy = \sum X_i Y_j - \frac{\left(\sum X_i\right)\left(\sum Y_j\right)}{n} \qquad (3)$$

We also have the sum of squares:

$$\sum x^2 = \sum X_i^2 - \frac{\left(\sum X_i\right)^2}{n} \qquad (4)$$

$$\sum y^2 = \sum Y_j^2 - \frac{\left(\sum Y_j\right)^2}{n} \qquad (5)$$

The null hypothesis tested is that there is no linear correlation, i.e. H_O: $\rho = 0$, and the H_A: $\rho \neq 0$. The test is usually two-tailed, but a one-tailed test can be done. The formulas use summation notation. See Appendix A for a summary.

(c) Properties of *r*

1. The values of the correlation coefficient can only vary from minus one to plus one;
$$-1 \leq r \leq 1$$

2. The value of *r* does not change if all values of either variable are converted to a different scale.

3. The linear correlation coefficient only measures the strength of a *linear* relationship.

4. It does not matter which variable you designate as "X" and which as "Y"

4. Correlations in biological systems

(a) Morphological Correlations.

Measurements made on two different structures on the same individual will often be correlated, but as illustrated in Fig. 1, it is unlikely that one variable will be the direct cause of the other. For example consider two size measurements of bumble bee (*Bombus* spp.) queens. One is the length of the radial cell (*mm*) in the wing (Fig. 3) the other is the wet weight (*mg*). Figure 4 shows a scatter plot of these variables for 25 queens of the species *Bombus bifarius*.

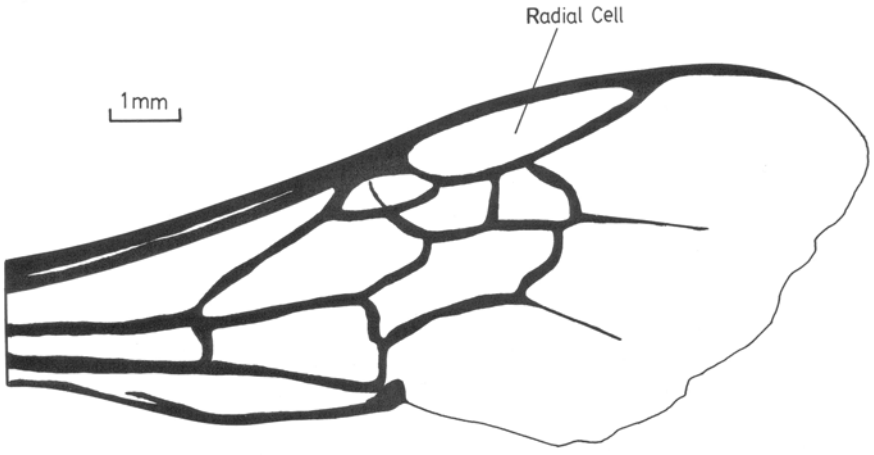

Figure 3. Drawing of a forewing of a queen bumble bee (*Bombus rufocinctus*) showing the radial cell.

(i) Examples

Radial cell length X (mm), vs. body mass Y (mg) of 25 queens of the bumble bee *Bombus bifarius*.

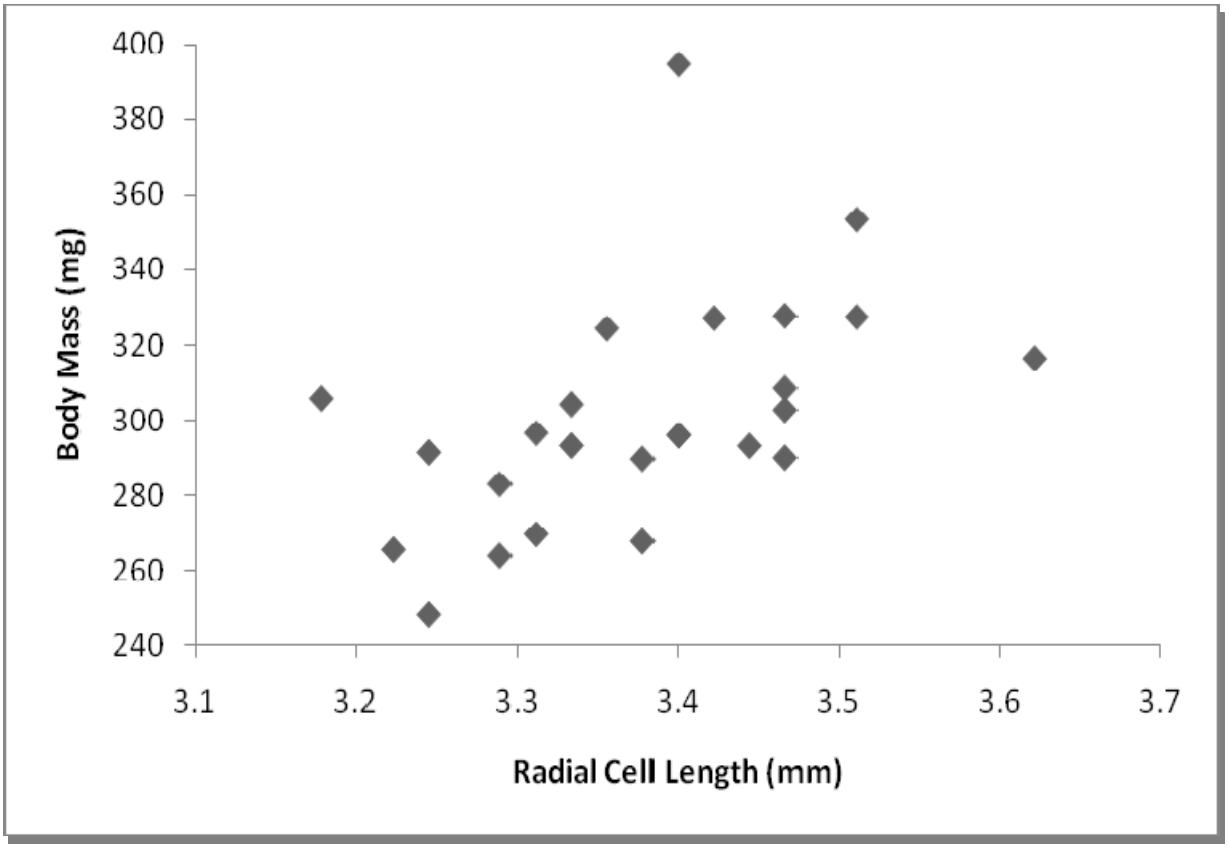

Figure 4. A scatter plot of radial cell length and body mass measured on 25 queens of the bumble bee *Bombus bifarius*.

As can be seen from the figure there is an obvious association between the variables, but we need to test this statistically to determine if the association is significant and to calculate a quantitative measure of the intensity of the association. We do this by calculating the sample correlation coefficient, *r,* and using it to test the hypothesis that the parametric value of the correlation is significantly different from zero. Formally we state: the **null hypothesis** H_O: $\rho = 0$, versus **the alternative hypothesis** H_A: $\rho \neq 0$.

The calculations are as follows:

X (mm)	X^2	Y (mg)	Y^2	XY
3.44	11.8641	293.0	85849.00	1009.222
3.37	11.4093	289.6	83868.16	978.2044
3.37	11.4093	267.6	71609.76	903.8933
3.46	12.0177	308.3	95048.89	1068.773
3.22	10.3827	265.4	70437.16	855.1778
3.46	12.0177	327.6	107321.8	1135.6800
3.40	11.5600	295.9	87556.81	1006.0600
3.31	10.9634	296.4	87852.96	981.4133
3.17	10.0982	305.6	93391.36	971.1289
3.31	10.9634	269.5	72630.25	892.3444
3.28	10.8167	263.8	69590.44	867.6089
3.46	12.0177	302.6	91566.76	1049.0130
3.51	12.3279	327.3	107125.30	1149.1870
3.42	11.7116	327	106929.00	1119.0670
3.33	11.1111	304.1	92476.81	1013.6670
3.62	13.1204	316.3	100045.70	1145.7090
3.46	12.0177	289.9	84042.01	1004.9870
3.51	12.3279	353.2	124750.2	1240.1240
3.28	10.8167	282.8	79975.84	930.0978
3.24	10.5264	248.2	61603.24	805.2711
3.40	11.5600	295.8	87497.64	1005.7200
3.35	11.2597	324.2	105105.60	1087.871
3.40	11.5600	394.5	155630.30	1341.3000
3.24	10.5264	291.3	84855.69	945.1067
3.33	11.1111	293.1	85907.61	977.0000
Σ = 84.44	285.4982716	7533	2292668	25483.63

$n = 25$

$\Sigma X = 84.44$ $\Sigma Y = 7533$ $\Sigma XY = 25483.63$

$\Sigma X^2 = 285.4982716$ $\Sigma Y^2 = 2292668$

$\Sigma x^2 = 22825.14$ $\Sigma y^2 = 0.25815$ $\Sigma xy = 38.83133$

Therefore,

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}} = \frac{38.83133}{\sqrt{(22825.14)(0.25815)}} = 0.506$$

This is the calculated r. To test H_O: $\rho = 0$ vs. H_A: $\rho \neq 0$ we compare the calculated value with the critical value from Table B1 (Appendix B); Critical value: $r_{0.05,\ 23} = 0.396$

Therefore we reject H_O and we can say that there is a significant linear correlation at the $\alpha = 0.05$ level.
Consider another species, *B. nevadensis* ($n = 20$);

RC length (mm)	**Mass (mg)**
4.77	767.0
4.64	632.3
4.64	813.9
4.46	630.7
4.75	756.0
4.73	796.3
4.68	780.6
4.68	771.0
4.66	841.6
4.64	786.2
4.53	731.0
4.59	776.0
4.95	740.0
4.51	728.6
4.75	824.2
4.87	775.2
4.82	823.6
4.82	825.3
4.57	822.9
4.75	1137.0

$\Sigma X = 93.88$ $\Sigma Y = 15759.4$ $\Sigma XY = 74050.97$

$\Sigma X^2 = 441.004636$ $\Sigma Y^2 = 12608722.14$

$\Sigma x^2 =$ _____ $\Sigma y^2 =$ _____ $\Sigma xy =$ _____

Therefore,

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}} = \frac{\_\_\_\_\_}{\sqrt{(\_\_\_\_\_)(\_\_\_\_\_)}} = \_\_\_\_\_$$

Critical value: $r_{0.05,\ 18} = 0.444$

What do you conclude about the significance of the correlation?
Figure 5 shows data from both species plotted together.

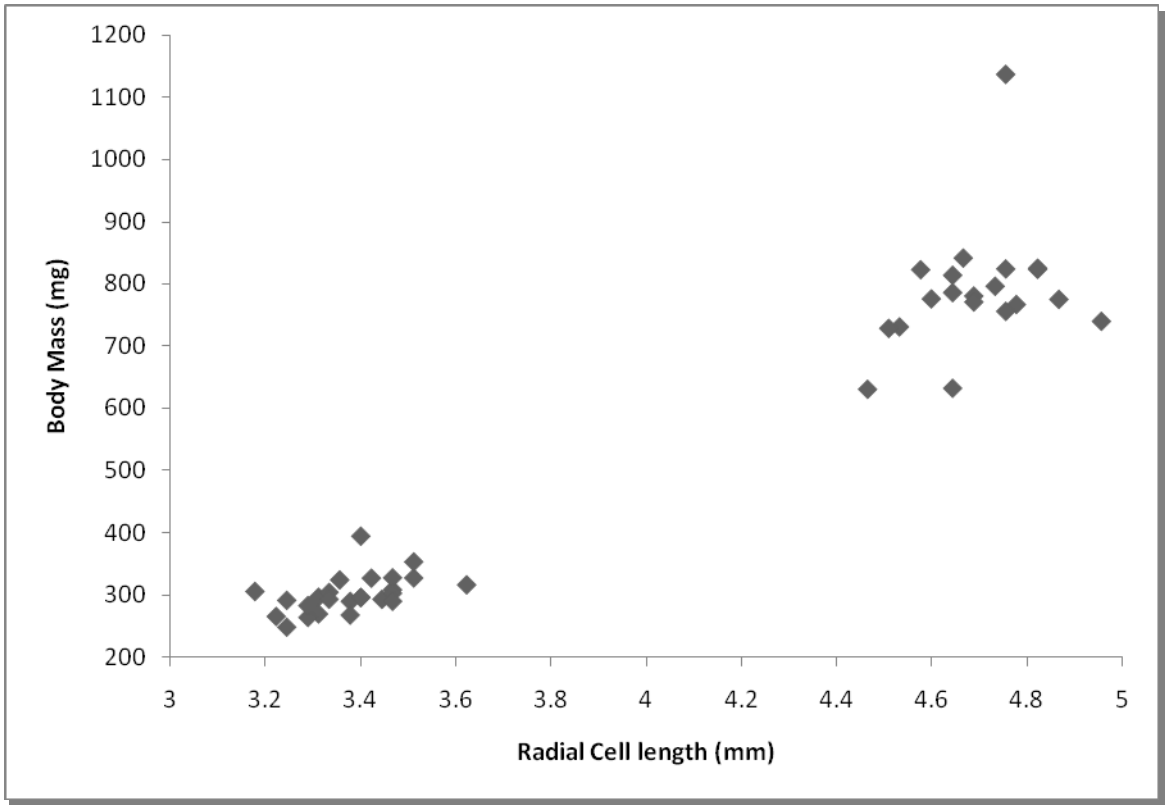

Figure 5. A scatter plot of radial cell length and body mass measured on 25 queens of the bumble bee *Bombus bifarius* and 20 queens of the larger species *B. nevadensis*.

(ii) Comparing Two Correlation Coefficients

Sometimes it is interesting or important to compare two correlation coefficients to see if the strength of an association is the same or different in two different populations. To do this first involves transforming each r to a z value:

$$z_i = 0.5 \ln\left(\frac{1+r_i}{1-r_i}\right) \qquad (6)$$

This can be calculated directly, or obtained from Table B2 (Appendix B).

Then,

$$z = \frac{z_1 - z_2}{\sigma_{z_1 - z_2}} \qquad (7) \qquad \text{where, } \sigma_{z_1 - z_2} = \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}$$

We test the hypothesis

H_O: $\rho_1 = \rho_2$
H_A: $\rho_1 \neq \rho_2$

Example:
We have two species of bumble bees. The correlation between radial cell length and body mass has been calculated for each sample. We want to determine if the two correlation coefficients are significantly different or not.

B. bifarius	*B. huntii*
$r_1 = 0.51$	$r_2 = 0.59$
$z_1 = 0.563$	$z_2 = 0.678$
$n_1 = 25$	$n_2 = 19$

therefore

$$\sigma_{z_1 - z_2} = \sqrt{\frac{1}{25-3} + \frac{1}{19-3}} = 0.32856$$

and so the calculated

$$z = \frac{0.563 - 0.678}{0.32856} = \frac{-0.115}{0.32856} = -0.350012$$

Since this is a *z*-test (Snedecor and Cochran, 1980) the critical value: $z_{0.05(2)} = \pm 1.96$ we fail to reject the null hypothesis, and so we conclude that the two correlation coefficients are not significantly different.

This means that we can calculate a common correlation coefficient (this is just the weighted average of the two):

$$z_w = \frac{(n_1 - 3)z_1 + (n_2 - 3)z_2}{(n_1 - 3) + (n_2 - 3)}$$

Therefore $$z_w = \frac{(22 \times 0.563) + (19 \times 0.678)}{22 + 19} = 0.61629$$

Thus $r_w = 0.55$

Exercise:
Compare the correlation coefficients calculated for *B. bifarius* and *B. nevadensis*.

Bumble bees

There are about 250 species of bumble bees worldwide (Williams, 2008) and about 50 of these occur in North America. They are all placed in the single genus *Bombus* Latreille. They are large, colourful and conspicuous insects. They are *primitively eusocial* as opposed to the *advanced eusocial* honeybees and ants. A typical colony is shown below.

A *Bombus huntii* colony reared in the laboratory. The queen is the large bee in the centre and the others are her daughters – the workers. In this, as in many species, there is a considerable size dimorphism between the queen and the workers. (Photograph by the author).

Colonies are annual, being founded by overwintered queens emerging from hibernation in the spring. They will have mated the previous fall. They find a nest site (usually an abandoned rodent nest) and rear the first brood of workers by themselves. After the first workers emerge, the colony enters the social stage for 2 – 3 months. Towards the end of summer the colony switches to young queen and male production; the foundress queen and workers eventually die as do the males after leaving the colony and mating with young queens.

Classification:

Kingdom	Animalia
Phylum	Arthropoda
Class	Insecta
Order	Hymenoptera
Family	Apidae
Subfamily	Apinae
Tribe	Bombini
Genus	*Bombus*

(b) Ecological Correlations.

(i) Foraging activity in bumble bees

Correlations abound in ecological situations where a host of factors, many unknown, will influence a measured variable either directly or indirectly. Young and Owen (1989) studied the foraging activity of bumble bee (*Bombus* spp.) workers in a subalpine meadow at an elevation of about 1980 m in the Kananaskis Mountain range of the Southern Canadian Rockies in western Alberta. The number of bumble bees observed foraging in three 10m x 10m study plots was recorded at 30 minutes intervals over five days. The only forage plant available was Yellow Hedysarum (*Hedysarum sulphurescens*) and environmental conditions as well as nectar measurements were recorded. The data were pooled over the study days to give a total (*n*) of 68 observations. Workers of seven bumble bee species occurred in the meadow during the study period: *B. occidentalis*, *B. bifarius*, *B. flavifrons*, *B. frigidus*, *B. mixtus*, *B. melanopygus*, *B. sylvicola.* The latter three accounted for over 80% of the total, and the data for the last two were combined because of their morphological, ecological and taxonomic similarity.

Table 1 (below) shows that foraging activity is strongly positively correlated with both temperature and nectar concentration, and to a lesser extent with the amount of sugar, but this is still highly significant.

Table 1. Correlations coefficients (*r*) between foraging activity of *B. melanopygus* and *B. sylvicola* (numbers of bees of both species combined), ambient temperature, and nectar production in *Hedysarum sulphurescens* for the data pooled over all study days (*n* = 68 observations).

Variable	Foraging Activity
Temperature (*°C*)	0.433****
Concentration (*%*)	0.345***
Sugar (*mg*)	0.292**

*P < 0.05, **P < 0.02, ***P < 0.01, ****P < 0.001, ns - not significant.

Although temperature clearly has the strongest influence, the correlations suggest that sugar concentration, on its own, also has a major effect. However, this may not be this simple, because we also have to consider the effect of temperature on sugar concentration. Table 2 (below) gives the correlations between the environmental variable (temperature and humidity) and the nectar variables.

Table 2. Correlations between environmental variables and nectar production in *Hedysarum sulphurescens* for the data pooled over all study days (n = 68 observations). Product-moment correlation coefficients (r) are given.

	Humidity	**Sugar Conc. (%)**	**Amount of Sugar (*mg*)**	**Nectar Volume (*μl*)**
Temperature (*°C*)	-0.802****	0.320****	0.262*	---------
Humidity		-0.534****	-0.308****	0.106 ns

*P < 0.05, **P < 0.02, ***P < 0.01, ****P < 0.001, ns - not significant.

We can postulate a causal pathway connecting temperature, humidity and sugar concentration; humidity is inversely related to temperature, and in turn sugar concentration is inversely related to humidity, thus leading to a positive correlation between temperature and concentration.

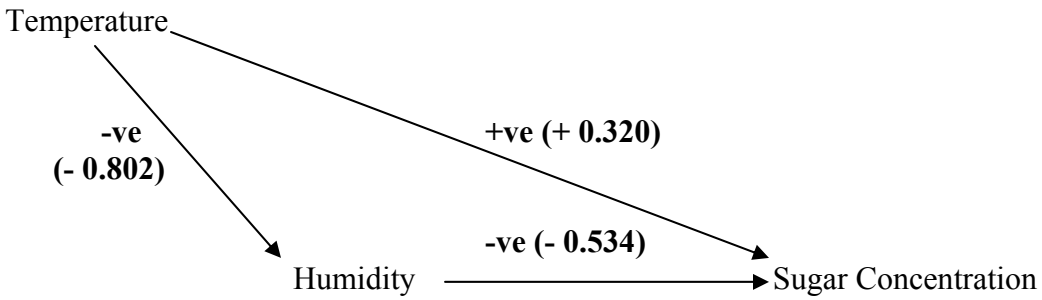

Figure 6. Causal pathways and correlations between environmental variables and sugar concentration in *H. sulphurescens*.

In order to discover the *direct* effect of sugar concentration on the foraging activity of the bees, we need to calculate the *partial correlation coefficient* between foraging activity and sugar concentration.

(ii) Partial correlations

If there are three variables, then there are three *simple correlations* between them; ρ_{12}, ρ_{13}, ρ_{23} (Snedecor and Cochran, 1980). Simple correlations are the type we have dealt with so far. However, the *partial correlation coefficient* $\rho_{12.3}$ is the correlation between variable 1 and 2 in individuals all having the same value of variable 3; i.e. the third variable is held constant so that only 1 and 2 are involved in the correlation (Snedecor and Cochran, 1980).

The sample estimate $r_{12.3}$ of $\rho_{12.3}$ is:

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}} \qquad (8)$$

This can be tested for significance using the Table B1 (Appendix B) of the Critical Values for the Pearson Correlation Coefficient, but using (n - 3) degrees of freedom rather than (n - 2) as for a simple correlation coefficient (Snedecor and Cochran, 1980)
We can now calculate the estimate ($r_{AC.T}$) of the partial correlation coefficient between foraging activity (A) and sugar concentration (C) with the temperature (T) held constant:

$$r_{AC.T} = \frac{0.345 - (0.320)(0.433)}{\sqrt{(1 - 0.433^2)(1 - 0.320^2)}} = 0.242$$

The values of the simple correlations from Tables 1 and 2 are substituted into equation (8). The degrees of freedom, df = 68 – 3 = 65, and therefore from Table B1 (Appendix B) we can conclude that the partial correlation is significant (P < 0.05), but just so. The effect of sugar concentration on its own is clearly much less that the original simple correlation coefficient from Table 1, would indicate.

Now we can construct a more complete picture of the interactions between the various factors that appear to influence bumble bee foraging activity in this subalpine meadow:

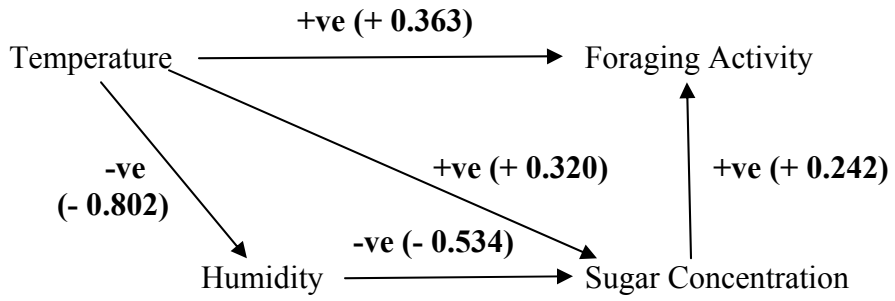

Figure 7. Causal pathways and correlations between environmental variables, sugar concentration and foraging activity**.**

Exercise:

a. Calculate the estimate ($r_{AS.T}$) of the partial correlation coefficient between foraging activity (A) and amount of sugar (S) with the temperature (T) held constant:

b.

$$r_{SC.T} = \frac{\_\_\_\_\_ - (\_\_\_\_)(0.433)}{\sqrt{(1 - 0.433^2)(1 - \_\_\_\_\_^2)}} = \_\_\_\_\_$$

What do you conclude?

c. In Figure 7 (above) the correlation coefficient between temperature and foraging activity is different from that given in Table 1. What is this correlation coefficient? Try to calculate it.

$$r\_\_ = \frac{\_\_\_\_\_ - (\_\_\_\_)(\_\_\_\_)}{\sqrt{(1 - \_\_\_\_^{2})(1 - \_\_\_\_\_^{2})}} = \_\_\_\_\_$$

(c) Genetic Correlations

Another example of correlation in nature is the correlation between relatives in groups or populations. Correlation is one way of measuring the **relatedness** between individuals. There are other methods, one being regression. The following simple derivation of the correlation between parent and child is modified from Li (1976). If we consider a single autosomal gene (i.e. one that is *not* sex-linked) with two alleles (*A* and *a*) in a random mating population, then there nine different mating types in total, see Table 3 below:

Table 3. Parent–Offspring chart for an autosomal gene in a population assumed to be in Hardy-Weinberg equilibrium.

Mating			**Offspring**		
Mother	**Father**	**Frequency**	***AA***	***Aa***	***aa***
AA	*AA*	p^4	p^4		
AA	*Aa*	$2p^3q$	p^3q	p^3q	
AA	*aa*	p^2q^2		p^2q^2	
Aa	*AA*	$2p^3q$	p^3q	p^3q	
Aa	*Aa*	$4\,p^2q^2$	p^2q^2	$2p^2q^2$	p^2q^2
Aa	*aa*	$2pq^3$		pq^3	pq^3
aa	*AA*	p^2q^2		p^2q^2	
aa	*Aa*	$2pq^3$		pq^3	pq^3
aa	*aa*	q^4			q^4

This table can be condensed to give the joint distribution of values for a single parent (mother or father) and child (Table 4).

Table 4. Parent-child correlation in a random mating population.

Parent		**Child**			
		AA	**Aa**	**Aa**	**Total Frequency**
	X	**2**	**1**	**0**	
AA	**2**	p^3	p^2q	*0*	p^2
Aa	**1**	p^2q	pq	pq^2	$2pq$
aa	**0**	*0*	pq^2	q^3	q^2
Total Frequency		p^2	$2pq$	q^2	

Each genotype can be assigned an X value (X in the parent, X' in the child), 2, 1 or 0. We calculate the correlation $r_{XX'}$: we have $\overline{X} = 2p$, $VAR(X) = 2pq$ (the Binomial Variance) which are the same in the parent and the child. Covariance is defined as $COV = \sum XX' - (\overline{X})(\overline{X}')$

Therefore,

$$\begin{aligned} COV(X,X') &= 4p^3 + 4p^2q + pq - (2p)^2 = 4p^3 + 4p^2q + pq - 4p^2 \\ &= 4p^2(p + q - 1) + pq \\ &= pq \end{aligned}$$

thus,

$$r_{XX'} = \frac{COV(X,X')}{\sigma(X)\sigma(X')} = \frac{COV(X,X')}{Var(X)} \frac{pq}{2pq} = \frac{1}{2}$$

as σ(X) = σ(X') in an equilibrium population. Note that the correlation is independent of the gene frequency in the population.

The correlations between some other relatives are given in Table 5 below:

Table 5. Correlations between relatives based on autosomal genes.

Relationship	**Correlation**
Parent-Child	r = ½
Identical Twins	r = 1
Full Sibs	r = ½
Half-Sibs	r = ¼
Grandparent-Grandchild	r = ¼
Uncle-Nephew	r = ¼

5. Correlation and asymmetry

(a) Asymmetry

Dr. A. Richard Palmer of the University of Alberta is an expert on the study of asymmetry in animals, and this brief introduction to asymmetry is largely based on his work, but see Palmer (1994, 2008) for a detailed discussion. The majority of animals are bilaterally symmetrical, with paired internal organs and paired appendages. However the symmetry often is not exact or "perfect". There are two general classes of asymmetry; **conspicuous** and **subtle** (Palmer, 2008). Conspicuous asymmetries are very obvious; an example is the extreme difference in size between right and left claws in some crabs, as shown by the Fiddler Crab (*Uca* spp.). You will observe this in the first laboratory exercise. However many animals exhibit a less obvious, but still definite, type of asymmetry that can only be quantified in a *sample* of individuals, and thus statistical methods must be used to analyze it (Palmer, 1994). Measurements are made on a structure on the right (R) and left (L) sides of each individual in the sample and an index of asymmetry is then calculated. Palmer (1994) distinguishes between different types of asymmetry. In this laboratory we will only be concerned with *fluctuating asymmetry*, but is important to understand the difference between this and *directional asymmetry*. I have taken the definitions of each directly from Palmer (1994):

"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; detected by statistical tests for departures of mean R - L from zero." Palmer (1994).

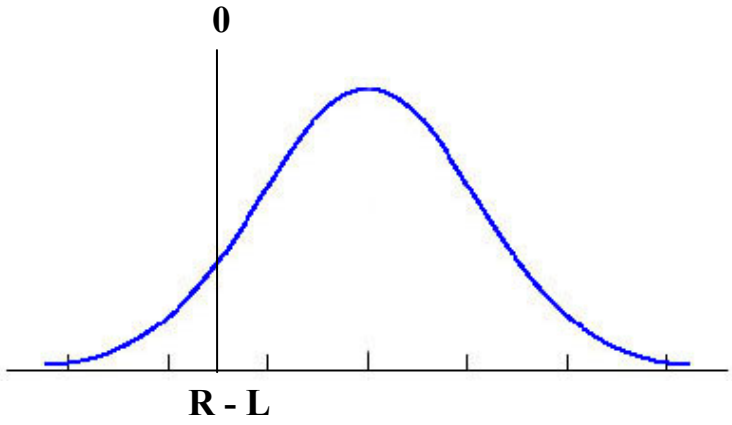

Figure 8. Directional asymmetry. Modified from Palmer Figure 1 (1994).

"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; a pattern of bilateral variation that may arise via many processes." Palmer (1994).

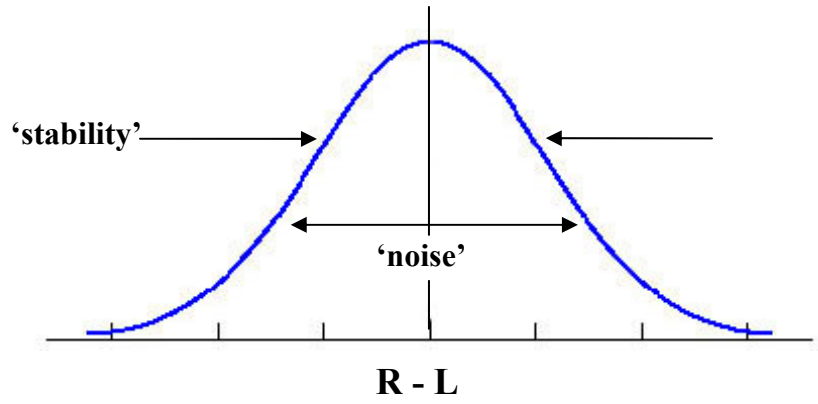

Figure 9. Fluctuating asymmetry. Modified from Palmer Figures 1 and 2 (1994).

Fluctuating asymmetry is often used as an indicator of developmental stability (Palmer, 1994) – see Fig. 9. Greater developmental stability will result in greater symmetry, i.e. less asymmetry.

(b) FA Indices

Palmer and Strobeck (1986) list 13 indices of fluctuating asymmetry. The first 10 (FA1 – FA10) apply only to single traits, the other three are based on multiple traits per individual. In the second laboratory exercise you will calculate FA1 and FA9.

$$\mathbf{FA1 = mean \left| R - L \right|} \qquad (10)$$

This is just the average of the absolute differences between the right and left sides. It is intuitively obvious and easy to calculate, although it will be very biased if there is directional asymmetry present, and it is sensitive size-dependence of $\left| R - L \right|$ (Palmer, 1994). Palmer (1994) notes that it is probably the best index for moderate to large (30+) sample sizes.

$$\mathbf{FA9 = (1 - r^2)} \qquad (11)$$

This is based on the correlation between right and left sides, and represents the percentage bilateral variation not due to positive correlation (Palmer, 1994), Figure 10 (below). It is easily computed and it is not biased by DA, however it is very dependent on overall trait size variation in the sample (Palmer, 1994). Palmer (1994) recommends that it only be used in conjunction with other indices.

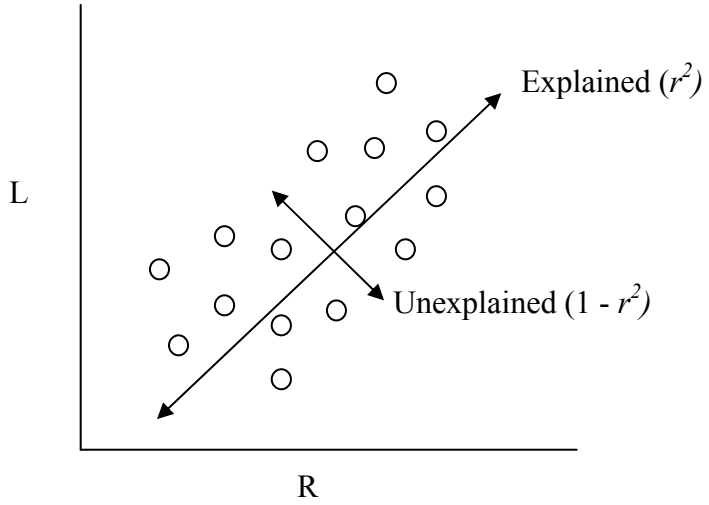

Figure 10. (modified from Figure 3 of Palmer, 1994). Hypothetical correlation between measurements on a structure on the right and left sides of an organism. FA9 is a measure of the amount of variation that is not explained by the linear correlation.

The correlation coefficient squared (r^2) is also called the **Coefficient of Determination** and is usually reported in regression studies as it gives a quick indication of the amount of variation explained by the regression as opposed to that unexplained. FA9 obviously has a non-linear relationship to r (Fig. 11).

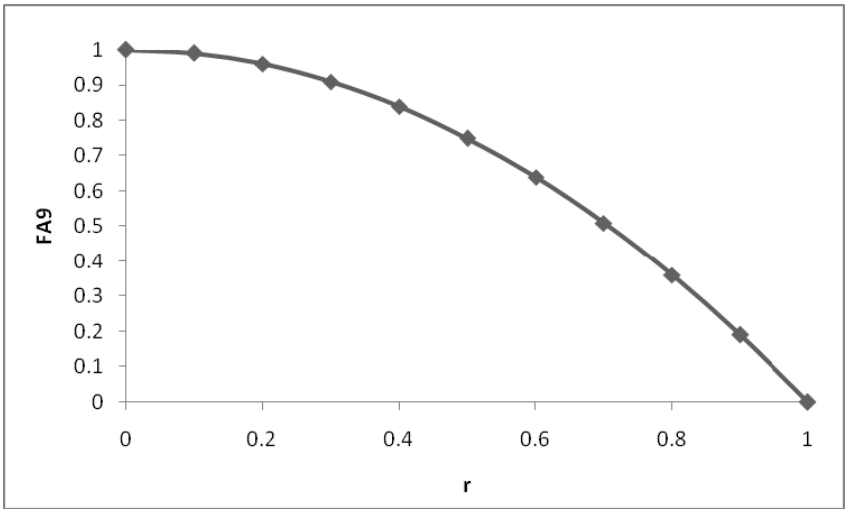

Figure 11. FA9 ($1 - r^2$) as a function of the correlation coefficient, r.

6. Laboratory Exercises

(a) Conspicuous asymmetry and correlation: Fiddler crabs

Fiddler crabs (genus *Uca*, see Textbox) are very interesting because of the conspicuous asymmetry exhibited by the males (Fig. 12).

Figure 12. *Uca pugnax*.
From: De Kay (1844) in Rosenberg (2007).

True crabs have five pairs of limbs and four pair are used for walking, these are the *ambulatory* legs. The two clawed limbs are called the *chelipeds*. On male fiddler crabs, the larger cheliped is called the *major cheliped*, and the smaller is called the *minor cheliped*. Female fiddler crabs have two small chelipeds that are similar to the minor cheliped in the males (Rosenberg, 2007). Both chelipeds grow allometrically with respect to carapace size (Huxley, 1932), but each has a different growth trajectory (Rosenberg 1997). The major cheliped is used in antagonistic (male-male combat) and courtship behaviours (Takeda and Murai, 1993). In some species there is additional asymmetry, as the side bearing the major cheliped has larger ambulatory legs and this is related to the type of courtship waving display (Takeda and Murai, 1993).

Fiddler Crabs:
This information is from Rosenberg (2007) www.fiddlercrab.info

Classification:

Kingdom	Animalia
Phylum	Arthropoda
Class	Crustacea
Sub-class	Malocostraca
Order	Decapoda
Infraorder	Brachyura
Superfamily	Ocypodoidea
Family	Ocypodidae
Subfamily	Ocypodinae
Genus	*Uca*

There are 97 recognized species/subspecies in the genus *Uca*.
The common English name “Fiddler Crab” comes from the feeding of the males, where the movement of the small claw from the ground to its mouth resembles the motion of a person moving a bow across a fiddle (the large claw).

Laboratory exercise:

You will measure the carapace width and the lengths of the major and minor chelipeds in the sample of male Fiddler Crabs provided. You will then calculate the product moment correlation coefficient (*r*) between carapace width and the claw lengths.

Question: Which correlation do you predict to be greater and why?
State this as a hypothesis.

Procedure:

a. Measure large claw and small claw lengths (Fig. 13) and carapace width (Fig. 14) as explained by the laboratory instructor.

b. Plot a scatter diagram for carapace width vs. major cheliped length and for carapace width vs. minor cheliped length.

c. Calculate correlation coefficients and test for significance.

d. State your conclusions.

Figure 13. The major and minor chelipeds of a typical male *Uca* crab.

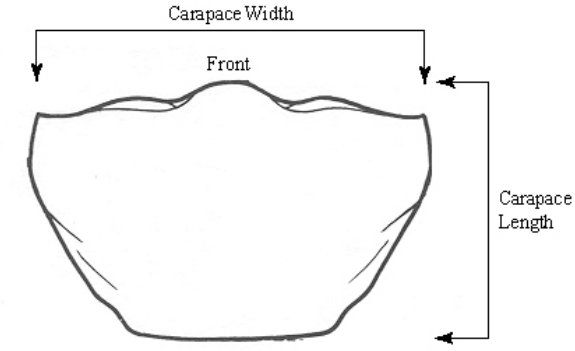

Figure 14. The carapace of a typical Fiddler Crab.
(Modified from Rosenberg 2007).

(b) Fluctuating asymmetry: Human fingerprint ridge numbers.

The individuality of human fingerprints has been recognized at least since the work of Francis Galton, and many studies have been carried out on inheritance of ridge patterns. Since the dermal ridges develop during the embryonic period and remain constant throughout the life of the individual, any environmental influences must be exerted in early embryogenesis.

In this exercise you are going to examine the correlation between ridge counts of individual fingers of right and left hands, and calculate two indices of fluctuating asymmetry.

There are three major groups of dermal ridge patterns: (a) arches, (b) loops, and (c) whorls (see Appendix C).

- The *arch* is the least frequent pattern, subdivided into two groups, plain (the ridges rise slightly over the middle of the finger) and tented (the ridges rise to a point in the center).

- The *loop* pattern is more complicated. It consists of a core and a tri-radius (a point where three ridge groups meet at angles of about 120 degrees). The core is a ridge surrounded by areas of ridges that turn back on themselves. Loops are classified as radial or ulnar depending on the orientation of the core ridge. Radial loops have a triradius that is on the side of the little finger, with the loop opening towards the thumb. The ulnar loop has a tri-radius that is on the side of the thumb, with the loop opening towards the little finger.

- The *whorl* pattern on the other hand, has two tri-radii with the ridges inside the whorl consisting of a number of different patterns.

The Total Ridge Count (TRC) is the sum of ridge counts on all fingers of both hands. In a study by Holt (1968), the average TRC for males was found to be 145 and for females it was found to be 126. Holt (1956) obtained the following correlations ($r \pm$ S.E.) of total finger-ridge counts between parents and offspring:

Father - Child: $r = 0.50 \pm 0.04$
Mother - Child: $r = 0.49 \pm 0.04$

How do these actual correlations compare with the theoretical correlations between relatives calculated in section 4 (c) above? What does this say about the inheritance of total fingerprint ridge count?

(1) Data Collection:

(i) Determine your ridge counts for each finger of both hands. Arches have a ridge count of zero. For loops, count the number of ridges between the tri-radius and the center or core of the pattern. For whorls, a ridge count is made from each tri- radius to the center of the fingerprint and the higher of the two possible counts is recorded. To determine your TRC add together each of your individual ridge counts.

(ii) Enter your data on the class Excel spreadsheet.

(2) Analysis:

(a) Correlation

(i) Graph a scatter plot for each finger.

(iii) Compute the Pearson Product Moment Correlation Coefficient (r) for each finger.

(iv) Construct a matrix of the correlation coefficients:

	T	1	2	3	4
T	----	r_{T1}	r_{T2}	r_{T3}	r_{T4}
1		---	r_{12}	r_{13}	r_{14}
2			---	r_{23}	r_{24}
3				---	r_{34}
4					---

(v) Compare all the correlation coefficients with each other; i.e. do all 10 pairwise comparisons. What do you conclude?

(b) Fluctuating Asymmetry

(i) Calculate FA9 $(1 - r^2)$ for each finger.

(ii) Calculate FA1 the absolute difference, Δ, in ridge counts, $\left| R - L \right|$, for each finger.

(iii) Calculate the mean and SE of FA1.

(iv) What do you conclude?

Materials

No expensive or sophisticated equipment is required. It is useful, although not essential, to have a computer available for the entry of the class data into an Excel spreadsheet.

(a) Fiddler crabs:

You can use either preserved specimens from a biological supply company, or if you can collect them yourself that would be ideal.

1. Preserved Fiddler Crabs Wards Natural Science (http://wardsci.com/)
   Catalog: 68 W 2792 Jar of 10 Can $10.94
   Website: 68 V 2792 Jar of 10 Can $8.75

2. Rulers – measurements can be made to the nearest *mm* with a regular plastic ruler. This is quite good enough for the purposes of this lab.

(b) Fingerprints:

A simple ink pad can be used; alternatively ink strips are more convenient:
Fingerprint ink strips Wards Natural Science (http://wardsci.com/)
Catalog: 15 F 3612 Package of 200 Can $33.50
Website: 15 V 3612 Package of 200 Can $26.50

Notes for the Instructor

This lab can be used as is, or parts of it can be taken and modified as desired depending on the course. It does not matter whether the students have taken statistics or not. If they have, then a review of the theory will not hurt, and they will see some applications. If they have not then the notes and exercises should be a sufficient introduction. My overall objective in this lab is to stimulate students to think and question. The actual results do not matter in themselves, and this lab will always "work". The emphasis should be on constructing and testing hypotheses and if they turn out to be wrong then that is how science proceeds.

Fiddler crabs: The exact points measured (Figs. 13 and 14) can easily be decided upon. A discussion should be initiated by the instructor to bring the class to a consensus about the (biological) hypothesis regarding which claw should show the highest correlation with carapace width. The laboratory instructor should read the paper by Rosenberg (1997). It is important to realise that "The two functions of the major cheliped, display and combat, do not necessarily have the same morphological requirements." (Rosenberg 1997). It can be argued either way. In any case the statistical hypotheses will be H_O: $\rho_1 = \rho_2$ and H_A: $\rho_1 \neq \rho_2$. Some sample data is given in Table 6 (below). This was collected by a zoology class using preserved specimens from Wards Natural Science.

Table 6. Some example Fiddler Crab data. All measurements are in *mm.*

Crab ID	Carapace width	Large claw length	Small claw length
1	1.9	3.4	0.8
2	2	3.2	0.7
3	1.9	3	0.8
4	1.9	3.2	0.8
5	1.8	2.9	0.7
6	1.7	2.9	0.7
7	1.7	2.8	0.6
8	1.7	2.5	0.7
9	1.7	2.6	0.7
10	1.9	2.6	0.9
11	1.6	2.7	0.8
12	2.1	3.9	0.7
13	1.7	2.6	0.5
14	2.1	3.1	0.8
15	1.8	3.3	0.5
16	2.1	3.5	0.8
17	1.7	2.8	0.7
18	2.05	3.6	0.8
19	1.8	3.1	0.8
20	2	3.5	0.8
21	1.7	2.7	0.7

Human fingerprint ridge numbers: Everyone is always interested in themselves, and it is quite striking how much variation there is even in a small class. The exercise gives a lot of practice calculating and comparing correlation coefficients. It should illustrate the point that it is surprising how much analysis can be done with even a relatively small amount of biological data. The results can be discussed in some different ways depending on the interests of the instructor. The meaning of fluctuating asymmetry could be the discussion, and here it is probably a good idea to stress two important point made by Palmer (1994); (1) that for valid conclusions measurements must be very accurate, and (2) one must be very careful in drawing sweeping conclusions from studies such as these. Otherwise the inheritance of human fingerprint ridge number can be discussed in light of the genetic correlations.

Literature Cited

De Kay, J. E. 1844. Zoology of New-York. Part VI. Crustacea. Albany, New York: Carroll and Cook.

Holt, S.B. 1956. Quantitative genetics of dermal ridge-patterns on fingers. Acta Genetica, 6:473-476.

Holt, S. B., 1968.The Genetics of Dermal Ridges. Charles C. Thomas, Springfield, IL.

Huxley, J.S. 1932. Problems of Relative Growth. Methuen & Co., London.

Li, C.C. 1976. First Course in Population Genetics. The Boxwood press, Pacific Grove, California.

Palmer, A. R. 1994. Fluctuating asymmetry analyses: A primer, pp. 335-364. *in* T. A. Markow (ed.), Developmental Instability: Its Origins and Evolutionary Implications. Kluwer, Dordrecht, Netherlands.

Palmer, A.R. 2008. Biological Asymmetry. Retrieved Oct. 23, 2008 from http://www.biology.ualberta.ca/palmer.hp/asym/asymmetry.htm

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

Rosenberg, M.S. 1997. Evolution of shape: differences between the major and minor chelipeds of *Uca pugnax* (Decapoda: Ocypodidae). Journal of Crustacean Biology, 17: 52-59.

Rosenberg, M.S. 2007. Fiddler Crabs (Genus *Uca*). Retrieved Oct. 28, 2008 from http://www.fiddlercrab.info/

Snedecor, G.W. and Cochran, W.G. 1980. Statistical Methods. 7th ed. The Iowa State University Press, Ames, Iowa, U.S.A.

Sokal, R. R. and Rohlf, J. 1981. Biometry. 2nd ed. W.H. Freeman and Company, New York.

Takeda, S. and Murai, M. 1993. Asymmetry in male fiddler crabs is related to the basic pattern of claw-waving display. Biological Bulletin, 184: 203-208.

Young, C.G. and Owen, R.E. 1989 Foraging activity of bumble bee (*Bombus* spp.) workers on *Hedysarum sulphurescens* in a subalpine meadow. Canadian Field Naturalist, 103:341-347.

Williams, P.H. 2008. Bombus bumblebees of the world. Retrieved Oct. 1, 2008 from http://www.nhm.ac.uk/research-curation/projects/bombus/index.html

Zar, J.H. 1984. Biostatistical Analysis, 2nd ed. Prentice-Hall Inc., Englewood Cliffs, N.J.

About the Author

Robin Owen graduated from the University of Toronto with a B.Sc. and then a Ph.D. in Zoology. In 1989 he joined Mount Royal as a Full-time faculty member. His research interests are centered on the evolutionary genetics of the Hymenoptera, and include bumble bee genetics, ecology and taxonomy, theoretical population genetics of X-linked and haplodiploid systems, insect mimicry and bee conservation genetics.

Appendix A: Summation Notation

The symbol Σ is used to mean "the sum of", and the quantities to be added immediately follow the Σ. Thus, Σ*X* represents the sum of the "x-values", i.e. the sum of a set of numbers labeled X. These can be integers:

$$\sum_{x=1}^{5} X = 1+2+3+4+5 = 16$$

A function of x can follow Σ,

$$\sum_{x=1}^{5} X^2 = 1^2+2^2+3^2+4^2+5^2 = 55$$

or $$\sum_{x=0}^{3} (X-2) = (0-2)+(1-2)+(2-2)+(3-2) = -2$$

In these examples, X is called the "index of summation". There is one term of the sum for each value of the index of summation.

Different indices of summation may be used. Given an arbitrary set of numbers, for instance 3, 5, -2, 4, 7, let X_1 represent the first number 3,
X_2 represent the second number 5, etc.
then the sum can be written,

$$\sum_{i=1}^{5} X_i = X_1 + X_2 + X_3 + X_4 + X_5$$

In this case the index of summation is i, which is indicated by putting i, under the Σ.
e.g.

$$\sum_{i=1}^{3} (X_i - 4)^2 = (X_1-4)^2 + (X_2-4)^2 + (X_3-4)^2 = (3-4)^2 + (5-4)^2 + (-2-4)^2 = 38$$

The index of summation normally takes all integral values, from the one under the Σ to the one on top of the Σ,
e.g.

$$\sum_{i=3}^{11} X_i = X_3 + X_4 + X_5 + ... + X_{10} + X_{11}$$

Sometimes, when the range of the index of summation is clear from the context, some information may be omitted: $\sum_i X_i$ (or even $\sum X_i$) indicates the sum over all I for which x_i is defined. The range may also be indicated in other, self-explanatory ways, such as $\sum_{i<5} X_i$.

Rules of Operation:

1. $\sum_{i=1}^{n} c = nc$, c is any constant.

The sum of a constant, c, is the constant times the number of terms in the sum,
e.g.

$$\sum_{i=1}^{4} 3 = 3+3+3+3 = 3\times 4 = 12$$

2. $$\sum_{i=1}^{n}(X_i + c) = \sum_{i=1}^{n} X_i + nc$$

e.g.

$$\sum_{i=1}^{4}(X_i + 3) = (3+3)+(5+3)+(-2+3)+(4+3) = (3+5-2+4)+(3+3+3+3)$$

$$= \sum X_i + 3\times 4 = 10 + 12 = 22$$

NOTE: do not confuse this expression with $\sum_i X_i + 3 = 10 + 3 = 13$

3. $$\sum_{i=1}^{n}(cX_i) = cX_1 + cX_2 + \ldots + cX_n = c(X_1 + X_2 + \ldots + X_n) = c\sum_{i=1}^{n} X_i ,$$

where c is a constant and a common factor.

4. Sometimes two or more numbers, X_i, Y_j, say are associated with each value of the index i.

$$\sum_{i=1}^{n}(X_i + Y_j) = (X_1 + Y_1) + (X_2 + Y_2) + \ldots + (X_n + Y_n) = \sum X_i + \sum Y_j$$

e.g.

i	X_i	Y_j	$(X_i + Y_j)$
1	3	-4	-1
2	2	-1	1
3	5	0	5
4	-4	-1	-5
5	-1	5	4
	5	-1	4

The "mean" of the numbers, $X_1, X_2, \ldots X_n$ is denoted by $\overline{X}$ and is defined by

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ , hence } \sum X_i = n\overline{X} .$$

Appendix B: Statistical Tables

Table B1.
Critical Values of the Pearson Product Moment Correlation Coefficient, *r*

*df**	α = 0.05	α = 0.01
3	0.878	0.959
4	0.811	0.917
5	0.754	0.875
6	0.707	0.834
7	0.666	0.798
8	0.632	0.765
9	0.602	0.735
10	0.576	0.708
11	0.553	0.684
12	0.532	0.661
13	0.514	0.641
14	0.497	0.623
15	0.482	0.606
16	0.468	0.590
17	0.456	0.575
18	0.444	0.561
19	0.433	0.549
20	0.423	0.537
23	0.396	0.505
25	0.381	0.487
30	0.349	0.449
35	0.325	0.418
40	0.304	0.393
45	0.288	0.372
50	0.273	0.354
60	0.250	0.325
70	0.232	0.302
80	0.217	0.283
90	0.205	0.267
100	0.195	0.254

*Degrees of Freedom; for a sample size n, $df = n - 2$.

Note: H_O: $\rho = 0$, versus H_A: $\rho \neq 0$, reject H_O if the absolute value of r is greater than the critical value in the table.

Table B2.
Table of $z = (½)ln[(1+r)/(1-r)]$ to transform the correlation coefficient.

r	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
0.0	0.000	0.010	0.020	0.030	0.040	0.050	0.060	0.070	0.080	0.090
0.1	0.100	0.110	0.121	0.131	0.141	0.151	0.161	0.172	0.182	0.192
0.2	0.203	0.213	0.224	0.234	0.245	0.255	0.266	0.277	0.288	0.299
0.3	0.310	0.321	0.332	0.343	0.354	0.365	0.377	0.388	0.400	0.412
0.4	0.424	0.436	0.448	0.460	0.472	0.485	0.497	0.510	0.523	0.536
0.5	0.549	0.563	0.576	0.590	0.604	0.618	0.633	0.648	0.662	0.678
0.6	0.693	0.709	0.725	0.741	0.758	0.775	0.793	0.811	0.829	0.848
0.7	0.867	0.887	0.908	0.929	0.950	0.973	0.996	1.020	1.045	1.071
0.8	1.099	1.127	1.157	1.188	1.221	1.256	1.293	1.333	1.376	1.422
0.90	1.472	1.478	1.483	1.488	1.494	1.499	1.505	1.510	1.516	1.522
0.91	1.528	1.533	1.539	1.545	1.551	1.557	1.564	1.570	1.576	1.583
0.92	1.589	1.596	1.602	1.609	1.616	1.623	1.630	1.637	1.644	1.651
0.93	1.658	1.666	1.673	1.681	1.689	1.697	1.705	1.713	1.721	1.730
0.94	1.738	1.747	1.756	1.764	1.774	1.783	1.792	1.802	1.812	1.822
0.95	1.832	1.842	1.853	1.863	1.874	1.886	1.897	1.909	1.921	1.933
0.96	1.946	1.959	1.972	1.986	2.000	2.014	2.029	2.044	2.060	2.076
0.97	2.092	2.109	2.127	2.146	2.165	2.185	2.205	2.227	2.249	2.273
0.98	2.298	2.323	2.351	2.380	2.410	2.443	2.477	2.515	2.555	2.599
0.99	2.646	2.700	2.759	2.826	2.903	2.994	3.106	3.250	3.453	3.800

Appendix C: Scoring Fingerprints

Examples of the three basic fingerprint patterns:

(a) Arch. (no triradius) There is no line of count and the score is 0.

(b) Loop with (one triradius). The triradius is on the left at the junction of three ridge systems. The white line joining the point of the triradius to the centre of the loop illustrates the method of ridge counting. The number of ridges cutting the line is 13.

(c) Whorl (two triradii). The white lines are two lines of count, one from each triradius to the centre of the whorl. The ridge count on the left of the pattern is 17, that on the right is 8. The higher count is used.