PROPERTIES OF THE CORRELATION COEFFICIENT

You may have seen the *covariance* of two random variables in M362K:

$$\mathrm{Cov}(X,Y) = E((X - E(X))(Y - E(Y)).$$

It is related to the variance by

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y).$$

(This is straightforward to establish from the definitions.) You might even have seen the *correlation coefficient for random variables*:

$$\rho = \mathrm{Cov}(X,Y)/(\mathrm{Var}(X)\mathrm{Var}(Y))^{1/2}$$

We can define the *sample covariance* of a collection $x_1, \ldots x_n$, $y_1, \ldots , y_n$ of two variable data as

$$\mathrm{cov}(x,y) = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).$$

Notice that the *sample correlation coefficient* defined in the textbook can be expressed as

$$r = \mathrm{cov}(x,y)/s_x s_y .$$

Also notice that cov(ax,by) = ab(cov(x,y)). (Details left to the student!)

If we use var(x) to denote the sample variance, then we have

$$\mathrm{var}(x + y) = \frac{1}{n-1}\sum_{i=1}^{n} (x_i + y_i - \overline{x+y})^2$$

$$= \frac{1}{n-1}\sum_{i=1}^{n} [(x_i - \bar{x}) + (y_i - \bar{y})]^2 \text{ (since } \overline{x+y} = \bar{x} + \bar{y})$$

$$= \frac{1}{n-1}\sum_{i=1}^{n} [(x_i - \bar{x})^2 + 2(x_i - \bar{x})(y_i - \bar{y}) + (y_i - \bar{y})^2]$$

$$= \mathrm{var}(x) + \mathrm{var}(y) + 2\mathrm{cov}(x,y).$$

Now apply this to $x/s_x + y/s_y$ instead of $x + y$:

$$\mathrm{var}(x/s_x + y/s_y) = \mathrm{var}(x/s_x) + \mathrm{var}(y/s_y) + 2\mathrm{cov}(x/s_x, y/s_y)$$

$$= \mathrm{var}(x)/s_x^2 + \mathrm{var}(y)/s_y^2 + 2\mathrm{cov}(x,y)/s_x s_y$$

$$= 2 + 2r. \qquad \text{(Why?)}$$

Since the sample variance is always ≥ 0 (Why?), this implies that $r \geq -1$.

Now if we had a case with $r = -1$, then we would have

$$\mathrm{var}(x/s_x + y/s_y) = 0,$$

which implies (Why? [Hint: Look at the definition of variance.]) that for each i,

$x_i/s_x + y_i/s_y = \overline{x/s_x + y/s_y}$, which we will call c (since it is a constant). Thus

$$x_i/s_x + y_i/s_y = c, \quad i = 1, 2, \ldots, n, \text{ so}$$

$$y_i = s_y (c - x_i/s_x), \; i = 1, 2, \ldots, n.$$

In other words, the points $(x_1,y_1), \ldots, (x_n, y_n)$ all lie on the straight line with slope $-s_y/s_x$, which is negative.

Similarly, by considering $\mathrm{var}(x/s_x + y/s_y)$, we can show that $r \leq 1$ and that if $r = 1$, then the points $(x_1,y_1), \ldots, (x_n, y_n)$ all lie on the straight line with slope s_y/s_x -- which is positive. (Fill in details!)