Noes on the correlation coefficient

November 15, 2010

1 Introduction

There are two types of correlation coefficients: the sample correlation coefficient, and the random variable analogue. Here, we will analyze and prove the properties of the random variable version; the properties for the sample version will be nearly identical, and follow from similar arguments.

Given a sample $(X_1, Y_1), ..., (X_k, Y_k)$, the *sample correlation coefficient* is defined to be

$$r \ := \ \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}},$$

where for a sample $(U_1, V_1), ..., (U_k, V_k)$ we use the notation

$$S_{UV} \ = \ \sum_{i=1}^{k}(U_i - \overline{U})(V_i - \overline{V}).$$

The random variable analogue is given by

$$\rho \ := \ \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y},$$

where σ_Z^2 denotes the variance $V(Z)$ of a random variable Z, and where $\mathrm{Cov}(X, Y)$ denotes the covariance, defined to be

$$\mathrm{Cov}(X, Y) \ := \ \mathbb{E}((X - \mu_X)(Y - \mu_Y)) \ = \ \mathbb{E}(XY) - \mu_X \mu_Y.$$

Note: In both cases, if the denominator in the definition of the correlation coefficient is 0, we will just say that the correlation coefficient is *undefined*.

We have that ρ satisfies the following properties

1. $-1 \leq \rho \leq 1$. r also satisfies this proerty.

2. If X and Y are independent, then $\rho = 0$; though, the converse is not true – that is, there exist *dependent* random variables X and Y for which $\rho = 0$.

3. If X and Y are linearly related, in the sense that $Y = \lambda_1 X + \lambda_2$, where $\lambda_1 \neq 0$, then $\rho = \pm 1$, where the sign here matches the sign of λ_1. This also holds for r.

4. Conversely, if $\rho = \pm 1$, then with probability 1 we will have that X and Y are linearly related; that is, there exists $\lambda_1 \neq 0$ and λ_2 for which $\mathbb{P}(Y = \lambda_1 X + \lambda_2) = 1$. Also, if $r = \pm 1$ then $Y_i = \lambda_1 X_i + \lambda_2$ for all i.

5. In these examples above we have intentionally omitted the case $\lambda_1 = 0$, the reason being that if $Y = \lambda_2$ or $X = \lambda_2'$, making X or Y constant random variables, then the correlation coefficient isn't even defined, because $\sigma_X = 0$ or $\sigma_Y = 0$ in those cases. The same goes for r.

2 Proofs of some of the properties of ρ

2.1 Proof that $-1 \leq \rho \leq 1$

We could prove this using a form of the Cauchy-Schwarz inequality for expectation, but that would be cheating, because, in some sense, C-S is *equivalent* to this property about ρ. What we will in fact do is to use the same proof technique for establishing C-S to also establish this property about ρ.

To this end, suppose that t is some real number that we will choose later, and consider the obvious inequality

$$\mathbb{E}((V + tW)^2) \ \geq \ 0, \text{ where } V = X - \mu_X \text{ and } W = Y - \mu_Y.$$

Expanding out the left-hand-side, and using the linearity of expectation, we find that

$$\mathbb{E}(V^2) + 2t\mathbb{E}(VW) + t^2\mathbb{E}(W^2) \ \geq \ 0.$$

Note that the left-hand-side is just a quadratic polynomial in t.

Now, clearly we have that

$$\mathbb{E}(V^2) = \sigma_X^2, \ \mathbb{E}(W^2) = \sigma_Y^2, \text{ and } \mathbb{E}(VW) = \mathrm{Cov}(X, Y);$$

and so, our polynomial inequality becomes

$$\sigma_Y^2 t^2 + 2\mathrm{Cov}(X,Y)t + \sigma_X^2 \ \geq \ 0.$$

From this inequality we find that the only way the left-hand-side could be 0 is if the polynomial has a double-root (i.e. it touches the x-axis in a single point), which could only occur if the discriminant is 0. So, the discriminant must always be negative or 0, which means that

$$4\mathrm{Cov}(X,Y)^2 - 4\sigma_X^2\sigma_Y^2 \ \leq \ 0.$$

In other words,

$$\frac{\mathrm{Cov}(X,Y)^2}{\sigma_X^2\sigma_Y^2} \ \leq \ 1;$$

provided, of course, that the denominator does not vanish.

2.2 Proof that $\rho = \pm 1$ implies X and Y are linearly related

From the proof in the previous subsection, we observe that the only way $\rho = \pm 1$ is if the discriminant of that quadratic polynomial is 0, which would mean that the quadratic polynomial vanishes for some value t_0 for the variable t. This would mean, however, that

$$\mathbb{E}((Y - \mu_Y + t_0 X - t_0\mu_X)^2) \ = \ \mathbb{E}((V + t_0 W)^2) \ = \ 0.$$

The only way this could occur is if $Y - \mu_Y + t_0 X - t_0\mu_X = 0$ with probability 1, which shows that X and Y are linearly related with probability 1.

2.3 Proof that if X and Y are linearly related, then $\rho = \pm 1$

Now suppose that

$$Y \ = \ \lambda_1 X + \lambda_2.$$

Then, we have that $\mu_Y = \lambda_1\mu_X + \lambda_2$; and so,

$$\mathrm{Cov}(X,Y) \ = \ \mathbb{E}((X - \mu_X)(\lambda_1 X - \lambda_1\mu_X)) \ = \ \lambda_1\mathbb{E}((X - \mu_X)^2) \ = \ \lambda_1\sigma_X^2.$$

Also, by properties of variance,

$$\sigma_Y^2 \;=\; V(\lambda_1 X + \lambda_2) \;=\; V(\lambda_1 X) \;=\; \lambda_1^2 \sigma_X^2.$$

From this it follows that

$$\rho \;=\; \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} \;=\; \frac{\lambda_1}{|\lambda_1|},$$

which is ± 1, with the sign determined by the sign of λ_1.

2.4 Independence implies 0 correlation coefficient, but not the converse

If X and Y are independent, then

$$\mathrm{Cov}(X,Y) \;=\; \mathbb{E}((X-\mu_X)(Y-\mu_Y)) \;=\; \mathbb{E}(X-\mu_X)\mathbb{E}(Y-\mu_Y) \;=\; 0 \cdot 0 \;=\; 0;$$

so of course the correlation coefficient is also 0.

The converse, however, is not true. To see this, we begin by defining independent random variables A and B that take on the values ± 1 with equal probability (i.e. probability $1/2$). Then, we define

$$X \;:=\; A + B, \text{ and } Y \;:=\; A - B.$$

We have that

$$\mathrm{Cov}(X,Y) \;=\; \mathbb{E}(XY) - \mu_X \mu_Y \;=\; \mathbb{E}(A^2 - B^2) - 0 \;=\; 0,$$

since $A^2 = B^2 = 1$. Yet, X and Y are *dependent*, since, for example, if $X = 2$, then Y is forced to equal 0.