

The Two Formulae for Correlation

The correlation coefficient is generally defined by one of two formula. Either $r = \frac{z_x z_y}{n-1}$

or $r = \frac{\sqrt{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$. The former is used when describing the correlation in absence of a

regression equation while the latter is used with reference to the regression line (hence the use of \hat{y}).

What is the connection between these two computations? How do you show that they are, in fact, the same measure?

Begin with $r = \frac{\sqrt{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$. In this equation, both y and \hat{y} are compared to the mean

value \bar{y} . To simplify the notation, define $\hat{Y}_i = \hat{y}_i - \bar{y}$, $Y_i = y_i - \bar{y}$, and $X_i = x_i - \bar{x}$. So

$$r = \frac{\sqrt{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{\sqrt{\sum_{i=1}^n \hat{Y}_i^2}}{\sqrt{\sum_{i=1}^n Y_i^2}}. \text{ To further simplify the work, consider } r^2 = \frac{\sum_{i=1}^n \hat{Y}_i^2}{\sum_{i=1}^n Y_i^2}.$$

We know that $y - \bar{y} = b(x - \bar{x})$ is one form of the least squares equation, so $\hat{Y} = bX$ is

our one parameter equation. Using calculus, we see that b is estimated by $b = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}$ (see

details at the end of this note).

So, we have $r^2 = \frac{\sum_{i=1}^n \hat{Y}_i^2}{\sum_{i=1}^n Y_i^2}$ with $\hat{Y} = bX$ and $b = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}$. Then

$$r^2 = \frac{\sum_{i=1}^n (bX_i)^2}{\sum_{i=1}^n Y_i^2} = \frac{b^2 \sum_{i=1}^n X_i^2}{\sum_{i=1}^n Y_i^2} = \frac{\left(\sum_{i=1}^n (X_i Y_i) \right)^2 \sum_{i=1}^n (X_i)^2}{\left(\sum_{i=1}^n X_i^2 \right)^2 \sum_{i=1}^n Y_i^2} = \frac{\left(\sum_{i=1}^n (X_i Y_i) \right)^2}{\left(\sum_{i=1}^n X_i^2 \right) \left(\sum_{i=1}^n Y_i^2 \right)}.$$

Now, rewrite back in terms of x and y .

$$r^2 = \frac{\left(\sum_{i=1}^n (X_i Y_i) \right)^2}{\left(\sum_{i=1}^n X_i^2 \right) \left(\sum_{i=1}^n Y_i^2 \right)} = \frac{\left(\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \right)^2}{\left(\sum_{i=1}^n (x_i - \bar{x})^2 \right) \left(\sum_{i=1}^n (y_i - \bar{y})^2 \right)}$$

Notice that the denominator is composed of the numerators of the variance of x and y . So, divide numerator and denominator by $(n-1)^2$.

$$r^2 = \frac{\left(\sum_{i=1}^n \frac{(x_i - \bar{x})(y_i - \bar{y})}{(n-1)} \right)^2}{\left(\sum_{i=1}^n \frac{(x_i - \bar{x})^2}{(n-1)} \right) \left(\sum_{i=1}^n \frac{(y_i - \bar{y})^2}{(n-1)} \right)} \text{ or } r^2 = \frac{\left(\sum_{i=1}^n \frac{(x_i - \bar{x})(y_i - \bar{y})}{(n-1)} \right)^2}{(s_x^2)(s_y^2)} = \frac{\left(\sum_{i=1}^n \frac{(x_i - \bar{x})(y_i - \bar{y})}{(s_x)(s_y)} \right)^2}{(n-1)^2}.$$

So, rewriting in terms of z -scores we have $r^2 = \frac{\left(\sum_{i=1}^n z_x z_y \right)^2}{(n-1)^2}$ and $r^2 = \frac{\left(\sum_{i=1}^n z_x z_y \right)}{(n-1)}$ as desired.

To show that the least squares estimate of b when $\hat{Y} = bX$ is $b = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}$, we need to consider

the minimization problem, “minimize $S = \sum_{i=1}^n (Y_i - bX_i)^2$ ”. We differentiate S with respect to b .

So, $\frac{dS}{db} = \sum_{i=1}^n 2(Y_i - bX_i)(-X_i) = 0$. Remember the values of X and Y are constants. So,

$\sum_{i=1}^n (-X_i Y_i + bX_i^2) = 0$ and $b \sum_{i=1}^n (X_i^2) = \sum_{i=1}^n (X_i Y_i)$. This means that the least squares estimate of b

is $b = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}$, when the data is centered at (\bar{x}, \bar{y}) .