

# The secrets of Coding

MXOLX VFDHV DUXVH GWKLV FLSKH U



For a set of n observations of a variable X:  $x_1, x_2, x_3, x_4, \dots, x_n$ .

The mean  $\bar{x} = \frac{\sum x}{n}$  and the variance  $s^2 = \frac{\sum x^2 - n\bar{x}^2}{n}$

**Properties of  $\Sigma$ :**  $\Sigma x = n\bar{x}$ ,  $\Sigma a = na$ ,  $\Sigma ax = a\Sigma x$  where a is a constant.

Now, suppose the original data needed some sort of moderation and we don't want to go through the lengthy calculation of the new mean and variance. We are going to find ways of simply writing down the **new** mean and variance based on the **old** mean and variance.

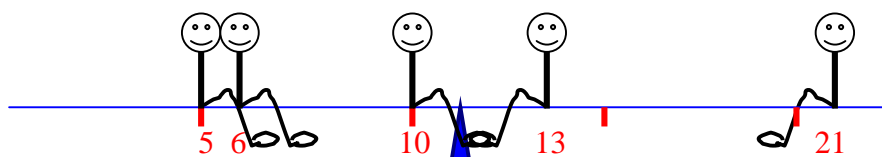
**Code:  $y = x + a$  (all the observations have a constant added to them).**

$$\bar{y} = \frac{\sum y}{n} = \frac{\sum (x+a)}{n} = \frac{\sum x + \sum a}{n} = \frac{\sum x}{n} + \frac{na}{n} = \bar{x} + a$$

The **new** mean is simply the **old** mean with the constant a added.

$$\begin{aligned} \text{The new variance } s_y^2 &= \frac{\sum y^2 - n\bar{y}^2}{n} = \frac{\sum (x+a)^2 - n(\bar{x}+a)^2}{n} \text{ since } \bar{y} = \bar{x} + a \\ &= \frac{\sum (x^2 + 2ax + a^2) - n(\bar{x}^2 + 2a\bar{x} + a^2)}{n} = \frac{\sum x^2 + 2a\sum x + \sum a^2 - n\bar{x}^2 - 2an\bar{x} - na^2}{n} \\ &= \frac{\sum x^2 + 2an\bar{x} + na^2 - n\bar{x}^2 - 2an\bar{x} - na^2}{n} = \frac{\sum x^2 - n\bar{x}^2}{n} = s_x^2 \text{ the variance of X.} \end{aligned}$$

**The new variance is exactly the same as the old variance. The spread has not changed!**



**If everyone moves along the seesaw by the same amount, the mean will change but the variance won't.**

**Code:  $y = mx + c$  (a linear transformation of all the observations).**

$$\bar{y} = \frac{\sum y}{n} = \frac{\sum (mx+c)}{n} = \frac{\sum mx + \sum c}{n} = \frac{m\sum x}{n} + \frac{nc}{n} = m\bar{x} + c$$

The **new** mean undergoes exactly the same transformation.

$$\begin{aligned} \text{The new variance } s_y^2 &= \frac{\sum y^2 - n\bar{y}^2}{n} = \frac{\sum (mx+c)^2 - n(m\bar{x}+c)^2}{n} \\ &= \frac{\sum (m^2x^2 + 2mcx + c^2) - n(m^2\bar{x}^2 + 2mc\bar{x} + c^2)}{n} = \frac{m^2\sum x^2 + 2mc\sum x + \sum c^2 - nm^2\bar{x}^2 - 2mcn\bar{x} - nc^2}{n} \\ &= \frac{m^2\sum x^2 + 2mcn\bar{x} + nc^2 - nm^2\bar{x}^2 - 2mcn\bar{x} - nc^2}{n} = m^2 \left[ \frac{\sum x^2 - n\bar{x}^2}{n} \right] = m^2 s_x^2 \end{aligned}$$

The new **variance** is the variance of X multiplied by  $m^2$ . The c again, has no effect on the spread.

The new **standard deviation** can be found by multiplying the old SD by m.  $S_y = m S_x$ .

**Code:  $y = \frac{x-a}{b}$  (a linear transformation if it is written  $y = \frac{1}{b}x - \frac{a}{b}$  of all the observations).**

It can be proved that the **new** mean is obtained from the old mean using the same transformation.  $\bar{y} = \frac{\bar{x}-a}{b}$  and the

standard deviation is only affected by b.  $s_y = \frac{1}{b} s_x$ . The variances are related by  $s_y^2 = \frac{1}{b^2} s_x^2$ .

## Coding with a bivariate distribution

For n observations of two variables X:  $x_1, x_2, x_3, x_4, \dots, x_n$  and Y:  $y_1, y_2, y_3, y_4, \dots, y_n$

We can work out the correlation coefficient and the regression coefficient  $b = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2}$

Often with two variables, the values of X are in arithmetic progression and sometimes the values of y.

x	200	210	220	230	240
y	80	92	104	110	120

If we code these values using  $X = \frac{x-p}{q}$  and  $Y = \frac{y-u}{v}$  and use  $p=200, q=10, u=100$  and  $v=2$ , we can simplify the values in the table and rewrite them in terms of the coded values:

X (subtract 200 and divide by 10)	0	1	2	3	4
Y (subtract 100 and divide by 2)	-10	-4	2	5	10

There are other reasons for coding but this time we can see that the figures become more manageable.

We can calculate the coded regression coefficient and then decode it to get the original coefficient.

If the original regression line is  $y = a + bx$  and the coded line is  $Y = A + BX$

$$B = \frac{\sum XY - n\bar{X}\bar{Y}}{\sum X^2 - n\bar{X}^2} = \frac{\sum \frac{x-p}{q} \frac{y-u}{v} - n \frac{\bar{x}-p}{q} \frac{\bar{y}-u}{v}}{\sum \left(\frac{x-p}{q}\right)^2 - n \left(\frac{\bar{x}-p}{q}\right)^2} = \frac{\frac{1}{qv} \sum \frac{xy-ux-py+pu}{1} - n \frac{\bar{x}\bar{y}-u\bar{x}-p\bar{y}+pu}{1}}{\frac{1}{q^2} \sum \frac{x^2-2px+p^2}{1} - n \frac{\bar{x}^2-2p\bar{x}+p^2}{1}}$$

$$= \frac{q}{v} \frac{\sum xy - nu\bar{x} - np\bar{y} + npu - n\bar{x}\bar{y} + nu\bar{x} + np\bar{y} - npu}{\sum x^2 - 2np\bar{x} + np^2 - n\bar{x}^2 + 2np\bar{x} - np^2} = \frac{q}{v} \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2} = \frac{q}{v} b$$

$B = \frac{q}{v} b$  and so, to decode we use  $b = \frac{v}{q} B = \frac{\text{ystep}}{\text{xstep}} B$  and note that it is independent of p and u.

The coded correlation coefficient is  $r = \frac{\sum XY - n\bar{X}\bar{Y}}{\sqrt{(\sum X^2 - n\bar{X}^2)(\sum Y^2 - n\bar{Y}^2)}} =$

$$\frac{\sum \frac{x-p}{q} \frac{y-u}{v} - n \frac{\bar{x}-p}{q} \frac{\bar{y}-u}{v}}{\sqrt{(\sum \left(\frac{x-p}{q}\right)^2 - n \left(\frac{\bar{x}-p}{q}\right)^2)(\sum \left(\frac{y-u}{v}\right)^2 - n \left(\frac{\bar{y}-u}{v}\right)^2)}}$$

$$= \frac{\frac{1}{qv} \sum \frac{xy-ux-py+pu}{1} - n \frac{\bar{x}\bar{y}-u\bar{x}-p\bar{y}+pu}{1}}{\frac{1}{qv} \sqrt{(\sum \frac{x^2-2px+p^2}{1} - n \frac{\bar{x}^2-2p\bar{x}+p^2}{1})(\sum \frac{y^2-2uy+u^2}{1} - n \frac{\bar{y}^2-2u\bar{y}+u^2}{1})}}$$

$$= \frac{\sum xy - nu\bar{x} - np\bar{y} + npu - n\bar{x}\bar{y} + nu\bar{x} + np\bar{y} - npu}{\sqrt{(\sum x^2 - 2np\bar{x} + np^2 - n\bar{x}^2 + 2np\bar{x} - np^2)(\sum y^2 - 2nu\bar{y} + nu^2 - n\bar{y}^2 + 2nu\bar{y} - nu^2)}}$$

$$= \frac{\sum xy - n\bar{x}\bar{y}}{\sqrt{(\sum x^2 - n\bar{x}^2)(\sum y^2 - n\bar{y}^2)}} \text{ which is the formula for the correlation coefficient of the original values.}$$

**Any coding of the form  $X = \frac{x-p}{q}$  and  $Y = \frac{y-u}{v}$  or indeed a linear coding  $X = mx + c$  and  $Y = ny + k$**

**produces no effect on the correlation coefficient.**