Statistics - Miscellaneous (3 pages; 21/8/16)

(1) Linear Coding

To establish the effect on $\bar{x}, s$ & r of the transformation

$y = ax + b$

$\bar{y} = \frac{\sum y_i}{n} = \frac{\sum(ax_i+b)}{n} = \frac{(a\sum x_i)+nb}{n} = a\bar{x} + b$

${s_y}^2 = \frac{S_{yy}}{n}$, where $S_{yy} = (\sum {y_i}^2) - n\bar{y}^2$

$= (\sum(ax_i + b)^2) - n(a\bar{x} + b)^2$

$= (a^2 \sum {x_i}^2) + (2ab \sum x_i) + nb^2 - na^2\bar{x}^2 - 2nab\bar{x} - nb^2$

$= a^2\{(\sum {x_i}^2) - n\bar{x}^2\}$, since $\sum x_i = n\bar{x}$

$= a^2 S_{xx}$

Hence ${s_y}^2 = \frac{a^2 S_{xx}}{n} = a^2 {s_x}^2$, and so $s_y = as_x$

Suppose that we are considering the correlation between x_i and z_i.

Then $r_{xz} = \frac{S_{xz}}{\sqrt{S_{xx}S_{zz}}}$ and $r_{yz} = \frac{S_{yz}}{\sqrt{S_{yy}S_{zz}}}$

Now $S_{yz} = (\sum y_i z_i) - n\bar{y}\bar{z}$

$= (\sum(ax_i + b)z_i) - n(a\bar{x} + b)\bar{z}$

$= (a \sum x_i z_i) + (b \sum z_i) - na\bar{x}\bar{z} - nb\bar{z}$

$= a\{(\sum x_i z_i) - n\bar{x}\bar{z}\}$, since $\sum z_i = n\bar{z}$

$= aS_{xz}$

Hence $r_{yz} = \frac{aS_{xz}}{\sqrt{a^2 S_{xx} S_{zz}}} = \frac{S_{xz}}{\sqrt{S_{xx} S_{zz}}} = r_{xz}$

(2) Formula for the Sample Variance

$s^2 = \frac{1}{n}\left\{(\sum {x_i}^2) - n\overline{x}^2\right\}$

[Note: The unbiased estimate of the population variance is

$\frac{1}{n-1}\left\{(\sum {x_i}^2) - n\overline{x}^2\right\}$, which allows for the fact that the data are being used to estimate the population mean.]

Starting from $s^2 = \frac{1}{n}\sum(x_i - \overline{x})^2$ (the average squared deviation from the mean), we have:

$s^2 = \frac{1}{n}\left\{\sum\left({x_i}^2 - 2x_i\overline{x} + \overline{x}^2\right)\right\}$

$= \frac{1}{n}\left\{(\sum {x_i}^2) - 2\overline{x}(\sum x_i) + n\overline{x}^2\right\}$

$= \frac{1}{n}\left\{(\sum {x_i}^2) - 2\overline{x}(n\overline{x}) + n\overline{x}^2\right\}$

$= \frac{1}{n}\left\{(\sum {x_i}^2) - n\overline{x}^2\right\}$

A useful check is as follows:

If all the n data items are the same, then each $x_i = \overline{x}$

and $\sum {x_i}^2 = n\overline{x}^2$, so that $s^2 = 0$; as expected, since there is no variance amongst the x_i.

Notes

(i) $(\sum {x_i}^2) - n\overline{x}^2$ is often denoted S_{xx}

(ii) $(\sum {x_i}^2) - n\overline{x}^2$ can also be written as $(\sum {x_i}^2) - \frac{(\sum x_i)^2}{n}$

(iii) It is tempting to write $\frac{1}{n}\{(\sum {x_i}^2) - n\overline{x}^2\}$ as $\frac{1}{n}(\sum {x_i}^2) - \overline{x}^2$,

but the $\frac{1}{n}$ in the first expression has the advantage of being associated with taking an average (of the squared deviations from the mean).