1 Mean and SD, Shifting and Scaling, Regression

1.1 Mean and standard deviation

Given a numeric list $\vec{x} = (x_1, x_2, \ldots, x_n)$ we write

$$\bar{x} \;=\frac{1}{n}\sum x \;=\; \frac{1}{n}\sum_{j=1}^{n} x_j \;=\; \frac{x_1 + x_2 + \cdots + x_n}{n}$$

for its **mean** or **average** and

$$s_x \;=\sqrt{\frac{1}{n-1}\sum (x-\bar{x})^2} \;=\; \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(x_j-\bar{x})^2}$$

(the "root–mean–square" of the differences $x_j - \bar{x}$) for its **standard deviation**.
Given a numeric list $\vec{x} = (x_1, x_2, \ldots, x_n)$, we write

$$\vec{\tilde{x}} \;= (\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n) \quad \text{where } \tilde{x}_j \;=\; \frac{x_j - \bar{x}}{s_x} \tag{1.1}$$

for the corresponding list of **z–scores**.

Change of scale:

	average	**standard deviation**
x_j	$\bar{x}$	SD_x
$y_j = x_j \pm c$	$\bar{y} = \bar{x} \pm c$	$\mathrm{SD}_y \;=\; \mathrm{SD}_x$
$u_j = x_j \cdot c \; (c \geqq 0)$	$\bar{u} = \bar{x} \cdot c$	$\mathrm{SD}_u \;=\; \mathrm{SD}_x \cdot c$
$v_j = x_j \cdot c \; (c < 0)$	$\bar{u} = \bar{x} \cdot c$	$\mathrm{SD}_v \;=\; \mathrm{SD}_x \cdot (-c)$
$w_j = x_j \cdot c + b$	$\bar{w} = \bar{x} \cdot c + b$	$\mathrm{SD}_w \;=\; \mathrm{SD}_x \cdot \lvert c \rvert$
$\tilde{x}_j = (x_j - \bar{x}) / \mathrm{SD}_x$	$\bar{\tilde{x}} = 0$	$s_{\tilde{x}} \;=\; 1$

If you can relate x_j and y_j by a change of scale $y_j = a \cdot (x_j + b)$ then both lists have the same standard unit: $\tilde{y}_j \;=\; \tilde{x}_j$ for all j

1.2 Notation for regression

When doing regression for a scatter diagram, note that each observation is not a single number x_j but rather a pair of numbers (x_j, y_j). There are two ways to place the arrows when writing such a scatter diagram as a list:

$$\overrightarrow{(x,y)} := (\vec{x}, \vec{y}) \;:=\; \big((x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\big)$$

The **five point summary**. for such a scatter diagram is $\boxed{\bar{x}, s_x, \bar{y}, s_y, r}$ where we compute the **correlation coefficient** r as follows:

a. Build associated lists $\vec{\tilde{x}}, \vec{\tilde{y}}$ of standard units:

$$\tilde{x}_j \;:=\; \frac{x_j - \bar{x}}{s_x}, \qquad \tilde{y}_j \;:=\; \frac{y_j - \bar{y}}{s_y}.$$

b. Create a new list $\vec{p}$ from the products $p_j := \tilde{x}_j \cdot \tilde{y}_j$.

c. Take the adjusted average of that list:

$$r := r_{x,y} := \frac{n}{n-1}\bar{p} \;=\; \frac{p_1 + p_2 + \cdots + p_n}{n-1} \;=\; \frac{1}{n-1}\sum_{j=1}^{n}\left(\frac{x_j - \bar{x}}{s_x} \cdot \frac{y_j - \bar{y}}{s_y}\right) \;=\; \frac{1}{n-1}\sum_{j=1}^{n} z_{x_j} z_{y_j}.$$

where $z_{x_j} = \frac{x_j - \bar{x}}{s_x}$ is the z–score of x_j, etc.

$$r \;=\; \frac{\sum\limits_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y})}{\sqrt{\sum\limits_{j=1}^{n}(x_j - \bar{x})^2 \;\sum\limits_{j=1}^{n}(y_j - \bar{y})^2}} \;=\; \frac{\sum\limits_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y})}{(n-1)s_x s_y}.$$

Note that there is an **ERROR** in the formula on p.157 of the SDM text. It reads, incorrectly,

$$r \;=\; \frac{\sum\limits_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y})}{\sqrt{\sum\limits_{j=1}^{n}(x_j - \bar{x})^2(y_j - \bar{y})^2}} \qquad \text{instead of the correct} \qquad r \;=\; \frac{\sum\limits_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y})}{\sqrt{\sum\limits_{j=1}^{n}(x_j - \bar{x})^2 \;\sum\limits_{j=1}^{n}(y_j - \bar{y})^2}}\,.$$

SD line and regression line for a scatter diagram

Both lines go through the **point of averages** with coordinates $(\bar{x}, \bar{y})$.

(1.2) The SD line has slope $m \;=\; \dfrac{s_y}{s_x}$ if $r > 0$,

(1.3) $m \;=\; -\dfrac{s_y}{s_x}$ if $r < 0$,

(1.4) The regression line has slope $m \;= r \cdot \dfrac{s_y}{s_x}$ always.