Memo: updating statistics

Michel Bierlaire*

July 8, 2016

Consider a series of values x_i, $i = 1, \ldots, n$. We are interested in the following quantities:

- the mean:
$$\bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i, \tag{1}$$

- the sum of squares:
$$S_n = \sum_{i=1}^{n} (x_i - \bar{x}_n)^2, \tag{2}$$

- the variance
$$V_n = \frac{S_n}{n-1}, \tag{3}$$

- the standard deviation
$$s_n = \frac{V_n}{\sqrt{n}}. \tag{4}$$

We collect a set of m additional pieces of data x_i, $i = n+1, \ldots, n+m$, and we would like to update the above quantities.

*Ecole Polytechnique Fédérale de Lausanne, Transport and Mobility Laboratory, CH-1015 Lausanne, Switzerland. Email: michel.bierlaire@epfl.ch

1 Mean

$$\begin{aligned}
\bar{x}_{n+m} &= \frac{1}{n+m}\sum_{i=1}^{n+m} x_i && (5)\\
&= \frac{1}{n+m}\left(\sum_{i=1}^{n} x_i + \sum_{i=n+1}^{n+m} x_i\right) && (6)\\
&= \frac{1}{n+m}\left(n\bar{x}_n + \sum_{i=n+1}^{n+m} x_i\right) && (7)\\
& && (8)
\end{aligned}$$

2 Sum of squares

Define

$$e = \bar{x}_n - \bar{x}_{n+m} \tag{9}$$

$$\begin{aligned}
\sum_{i=1}^{n}(x_i - \bar{x}_{n+m})^2 &= \sum_{i=1}^{n}(x_i - \bar{x}_n + \bar{x}_n - \bar{x}_{n+m})^2 && (10)\\
&= \sum_{i=1}^{n}(x_i - \bar{x}_n + e)^2 && (11)\\
&= \sum_{i=1}^{n}\left((x_i - \bar{x}_n)^2 + e^2 + 2e(x_i - \bar{x}_n)\right) && (12)\\
&= S_n + ne^2 + 2e\sum_{i=1}^{n}(x_i - \bar{x}_n) && (13)\\
&= S_n + ne^2 && (14)
\end{aligned}$$

Therefore,

$$S_{n+m} = \sum_{i=1}^{n+m} (x_i - \bar{x}_{n+m})^2 \tag{15}$$

$$= \sum_{i=1}^{n} (x_i - \bar{x}_{n+m})^2 + \sum_{i=n+1}^{m} (x_i - \bar{x}_{n+m})^2 \tag{16}$$

$$= S_n + ne^2 + \sum_{i=n+1}^{m} (x_i - \bar{x}_{n+m})^2 \tag{17}$$

$$\tag{18}$$

3 Variance

$$V_{n+m} = \frac{S_{n+m}}{n+m-1} \tag{19}$$

4 Standard deviation

$$s_{n+m} = \frac{V_{n+m}}{\sqrt{n+m}}. \tag{20}$$