相関係数の求め方(標準偏差と分散の求め方を含む)

x 軸のデータと y 軸のデータに関連性があるかの判定には、相関係数を用いるのが一般的。

相関係数と言えば、一般的には「ピアソンの積率相関係数」ということらしい。
相関係数 r は次の式で表される。

$$相関係数 = \frac{共分散}{\sqrt{x軸の分散 \times y軸の分散}}$$

x 軸データを $x_i(i=1 \to n)$ 、その平均を $\bar{x}$ とすると分散は次の式で表される。

$$分散 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

不偏分散とか言うのが正しいらしいが、n で割るか n-1 で割るかはいろいろある模様。
但し、実は(後述するが)相関係数を求めるのには影響がない。

y 軸についての分散も同様。共分散は、次の式となる。

$$共分散 = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \times (y_i - \bar{y})}{n-1}$$

尚、標準偏差の式は次の通り。

$$標準偏差 = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

見ての通り分散の平方根となっているので、標準偏差を求める過程で分散も求められる。
寧ろ、分散を元に標準偏差を求めていると言うべきか。
標準偏差については、お馴染みだと思うので詳細は省略。

ここまでで、相関係数(と標準偏差)を求める手段は揃った。実装に入る前に、式を変形してみる。
分散を求める式をよく見ると、平均値との差を(2 乗してから)集計するので集計前に平均値が必要になる。
平均値自体は全データにアクセスしないと得られないので、都合 2 回アクセスする必要ができて都合が悪い。
しかし、次のように変形することができる。
その理屈を説明できるほど判っていないのが難だが……
便宜上、ついでに平均を使って書き直してある。

$$分散 = \frac{\sum(x_i \times x_i) - \frac{\sum x_i \times \sum x_i}{n}}{n-1} = \frac{\sum(x_i \times x_i) - \bar{x} \times \sum x_i}{n-1}$$

また、共分散も同じように変形できる。

$$共分散 = \frac{\sum(x_i \times y_i) - \frac{\sum x_i \times \sum y_i}{n}}{n-1} = \frac{\sum(x_i \times y_i) - \bar{x} \times \sum y_i}{n-1}$$

すると、当然標準偏差はこうなる。

$$標準偏差 = \sqrt{\frac{\sum(x_i \times x_i) - \frac{\sum x_i \times \sum x_i}{n}}{n-1}} = \sqrt{\frac{\sum(x_i \times x_i) - \bar{x} \times \sum x_i}{n-1}}$$

これらの式は事前に平均値を求めておく必要がないので、データには一回アクセスすれば済む。
つまり、データを読み捨てるケースでも使えると言う次第。
実はこの方式、関数電卓の標準偏差の機能そのものだったり。
データエントリー時に、総計（ $\sum x_i$ ）と平方和（ $\sum (x_i^2)$ ）と件数(n)を更新するだけで済むので。

ここでやっと本題の相関係数の式に戻って。

$$相関係数 = \frac{共分散}{\sqrt{x軸の分散 \times y軸の分散}}$$

$$r = \frac{\dfrac{\sum (x_i \times y_i) - \bar{x} \times \sum y_i}{n-1}}{\sqrt{\dfrac{\sum (x_i \times x_i) - \bar{x} \times \sum x_i}{n-1} \times \dfrac{\sum (y_i \times y_i) - \bar{y} \times \sum y_i}{n-1}}}$$

分母子に(n - 1)を掛ければすっきりする。また、これが分散の分母がnでもn-1でも構わない理由になる。

$$r = \frac{\sum (x_i \times y_i) - \bar{x} \times \sum y_i}{\sqrt{(\sum (x_i \times x_i) - \bar{x} \times \sum x_i) \times (\sum (y_i \times y_i) - \bar{y} \times \sum y_i)}}$$

勿論、この式でも標準偏差と同様に1パスアクセスで計算できる。
計算手順を纏めると、次のようになる。

```
while データがある {
        Σx = Σx + xi
        Σy = Σx + yi
        Σxx = Σxx + xi * xi
        Σyy = Σyy + yi * yi
        Σxy = Σxy + xi * yi
        n = n + 1
}
xBar = Σx / n   // x軸平均
yBar = Σy / n   // y軸平均
sxx = Σxx - xBar * Σx
syy = Σyy - yBar * Σy
sxy = Σxy - xBar * Σy
xstdev = √(sxx / (n - 1))          // x軸標準偏差
ystdev = √(syy / (n - 1))          // y軸標準偏差
r = sxy / √(sxx * syy)   // 相関係数
```

最後に、相関の強さについてまとめておく。

1.0〜0.7	強い正の相関がある
0.7〜0.4	中程度の正の相関がある
0.4〜0.2	弱い正の相関がある
0.2〜-0.2	ほとんど相関がない
-0.2〜-0.4	弱い負の相関がある
-0.4〜-0.7	中程度の負の相関がある
-0.7〜-1.0	強い負の相関がある