

Formulæ for concepts that Tom is drawing

Thomas Levine

April 2, 2013

These formulæ are for the bivariate (rather than more-than-two) populations (rather than samples).

1 Covariance

Conceptually, it's this.

$$\sigma_{xy} = E \left[(x - E[x])(y - E[y]) \right]$$

More precisely, it could be this.

$$\sigma_{xy} = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})$$

As matrix arithmetic, it's written in a way that works for more than two variables. I'm not going to explain it, but here it is. (Σ is the covariance matrix.)

$$\Sigma = E \left[(\mathbf{X} - E[\mathbf{X}])(\mathbf{X} - E[\mathbf{X}])^T \right]$$

In R, it's this.

```
# Two variables  
cov(x,y)
```

```
# More variables  
var(iris[-5])  
cov(iris[-5])
```

2 Variance

Variance is the covariance of something with itself.

$$\sigma_x^2 = \sigma_{xx} = E \left[(x - E[x])^2 \right]$$

More precisely, like the covariance, it could be this.

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

As matrix arithmetic, it's the diagonal of the covariance. Again, I'm not really going to explain it.

$$\text{diag}(\Sigma)$$

In R, all of these would work.

```
# One variable
var(x)
cov(x,x)

# More variables
lapply(iris[-5],var)
diag(cov(iris[-5]))
```

3 Correlation

Correlation (specifically, the Pearson product-moment correlation coefficient—there are others) measures how much the variables linearly depend on each other.

For perfectly linear variables that move together, ρ_{xy} will be 1. For perfectly linear variables that move oppositely, ρ_{xy} will be -1 .

ρ_{xy} is just the covariance (σ_{xy}) divided by the product of the standard deviations ($\sigma_x \sigma_y$).

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

We can also factor out the N term and write it as the sum of the rectangles divided by the rectangle formed by the square roots of the sums of the squares.

$$\beta = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^N (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^N (y_i - \bar{y})^2}}$$

If you need to be convinced that σ_{xy} is no greater than $\sigma_x \sigma_y$, recall that the covariance includes negative rectangles and the variance does not;

4 Ordinary least-squares regression

Simple regression is a best fit line between two variables; we are looking for α and β .

$$\hat{y}_i = \alpha + \beta x_i$$

β is just the covariance divided by the variance.

$$\beta = \sigma_{xy} / \sigma_x^2$$

We can also factor out the N term and write it as the sum of the rectangles divided by the sum of the squares.

$$\begin{aligned}\beta &= \frac{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x}) (y_i - \bar{y})}{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2} \\ \beta &= \frac{\sum_{i=1}^N (x_i - \bar{x}) (y_i - \bar{y})}{\sum_{i=1}^N (x_i - \bar{x})^2}\end{aligned}$$

5 Credits

Some formulæ were lifted from these pages.

- <http://en.wikipedia.org/w/index.php?title=Covariance&action=edit>
- http://en.wikipedia.org/w/index.php?title=Pearson_product-moment_correlation_coefficient&action=edit§ion=3
- http://en.wikipedia.org/w/index.php?title=Ordinary_least_squares&action=edit