Cross Validated

# How does the formula for generating correlated random variables work?

Asked 8 years, 9 months ago Modified 5 years, 5 months ago Viewed 23k times

28

If we have 2 normal, uncorrelated random variables $X_1, X_2$ then we can create 2 correlated random variables with the formula

$$Y = \rho X_1 + \sqrt{1-\rho^2} X_2$$

and then $Y$ will have a correlation $\rho$ with $X_1$.

Can someone explain where this formula comes from?

correlation normal-distribution covariance

Share Cite Improve this question Follow

edited Mar 12, 2015 at 22:17
D.W.
6,578 4 43 71

asked Mar 12, 2015 at 12:06
Lanza
539 1 4 10

1 An extensive discussion of this and related issues appears in my answer at stats.stackexchange.com/a/71303. Among other things, it makes plain that (1) the Normality assumption is irrelevant and (2) you need to make additional assumptions: the variances of $X_1$ and $X_2$ must be equal in order for the correlation of $Y$ with $X_1$ to be $\rho$. – whuber ♦ Mar 12, 2015 at 13:42

Very interesting link. I'm not sure I understand what you mean by normality being irrelevant. If $X_1$ or $X_2$ is not normal, and it becomes harder to control the density of $Y$ through the Kaiser-Dickman algorithm. This is the whole reason for specialized algorithms to generate non-normal correlated data (e.g., Headrick, 2002; Ruscio & Kaczetow, 2008; Vale & Maurelli, 1983) For example, imagine your goal is to generate $X$~normal, $Y$~uniform, with $\rho$=.5. Using $X_2$~uniform results in a $Y$ that is not uniform ($Y$ ends up being a linear combination of a normal and uniform). – Anthony Mar 12, 2015 at 17:25

@Anthony The question only asks about *correlation*, which is purely a function of first and second moments. The answer does not depend on any other properties of the distributions. What you are discussing is a different subject altogether. – whuber ♦ Mar 12, 2015 at 23:51

## 3 Answers

Sorted by: Highest score (default)

24

Suppose you want to find a linear combination of $X_1$ and $X_2$ such that

$$\operatorname{corr}(\alpha X_1 + \beta X_2, X_1) = \rho$$

Notice that if you multiply both $\alpha$ and $\beta$ by the same (non-zero) constant, the correlation will not change. Thus, we're going to add a condition to preserve variance: $\operatorname{var}(\alpha X_1 + \beta X_2) = \operatorname{var}(X_1)$

This is equivalent to

$$\rho = \frac{\operatorname{cov}(\alpha X_1 + \beta X_2, X_1)}{\sqrt{\operatorname{var}(\alpha X_1 + \beta X_2)\operatorname{var}(X_1)}} = \frac{\overbrace{\alpha \operatorname{cov}(X_1, X_1)}^{=\operatorname{var}(X_1)} + \overbrace{\beta \operatorname{cov}(X_2, X_1)}^{=0}}{\sqrt{\operatorname{var}(\alpha X_1 + \beta X_2)\operatorname{var}(X_1)}}$$
$$= \alpha \sqrt{\frac{\operatorname{var}(X_1)}{\alpha^2 \operatorname{var}(X_1) + \beta^2 \operatorname{var}(X_2)}}$$

Assuming **both random variables have the same variance** (this is a crucial assumption!) ( $\operatorname{var}(X_1) = \operatorname{var}(X_2)$), we get

$$\rho \sqrt{\alpha^2 + \beta^2} = \alpha$$

There are many solutions to this equation, so it's time to recall variance-preserving condition:

$$\operatorname{var}(X_1) = \operatorname{var}(\alpha X_1 + \beta X_2) = \alpha^2 \operatorname{var}(X_1) + \beta^2 \operatorname{var}(X_2) \Rightarrow \alpha^2 + \beta^2 = 1$$

And this leads us to

$$\alpha = \rho$$
$$\beta = \pm\sqrt{1 - \rho^2}$$

**UPD**. Regarding the second question: yes, this is known as whitening.

Share Cite Improve this answer Follow

edited Jun 25, 2018 at 7:29
Stupidient
3 3

answered Mar 12, 2015 at 12:51
Artem Sobolev
2,826 14 18

9

The equation is a simplified bivariate form of Cholesky decomposition. This simplified equation is sometimes called the Kaiser-Dickman algorithm (Kaiser & Dickman, 1962).

Note that $X_1$ and $X_2$ must have the same variance for this algorithm to work properly. Also, the algorithm is typically used with normal variables. If $X_1$ or $X_2$ are not normal, $Y$ might not have the same distributional form as $X_2$.

**References:**

Kaiser, H. F., & Dickman, K. (1962). Sample and population score matrices and sample correlation matrices from an arbitrary population correlation matrix. *Psychometrika, 27*(2), 179-182.

Share Cite Improve this answer Follow

edited Mar 13, 2015 at 14:23

answered Mar 12, 2015 at 12:31
Anthony
1,602 14 24

2 I suppose you don't need standardized normal variables, just having the same variance should be enough. – Artem Sobolev Mar 12, 2015 at 12:50

2 No, the distribution of $Y$ is **not** a *mixture* distribution as you claim. – Dilip Sarwate Mar 12, 2015 at 14:39

Point taken, @Dilip Sarwate. If either $X_1$ or $X_2$ is nonnormal, then $Y$ becomes a linear combination of two variables that might not result in the desired distribution. This is the reason for specialized algorithms (instead of Kaiser-Dickman) for generated non-normal correlated data. – Anthony Mar 12, 2015 at 17:27

4

Correlation coefficient is the $\cos$ between two series if they are treated as vectors (with $n^{th}$ data point being $n^{th}$ dimension of a vector). The above formula simply creates a decomposition of a vector into its $\cos\theta$, $sin\theta$ components (with respect to $X_1, X_2$).
if $\rho = cos\theta$ , then $\sqrt{1-\rho^2} = \pm sin\theta$.

Because if $X_1, X_2$ are uncorrelated, the angle between them is a right angle (ie, they can be considered as orthogonal, albeit non-normalized, basis vectors ).

Share Cite Improve this answer Follow

edited Mar 23, 2015 at 2:20

answered Mar 23, 2015 at 0:46
Dmitry Rubanovich
171 4

2 Welcome to our site! I believe your post will get more attention if you mark up the mathematical expressions using $\TeX$: enclose them between dollar signs. There's help available when you're editing. – whuber ♦ Mar 23, 2015 at 0:53