One More Formula for the Variance

By Sabiou Inoua*

Abstract This short paper establishes one more formula for the variance. Consider a random variable *X* whose possible values are $x_1, ..., x_n$, with probabilities $p_1, ..., p_n$ of occurring, respectively. Pick two of these possible values successively (each x_i having the probability p_i of being chosen). Compute the difference between the two chosen values. Square the difference. Claim: you are expected to get (twice) the variance of *X*. This formula makes the variance appear an even more natural measure of dispersion than usually thought.

Let *X* be a discrete random variable whose possible values are $x_1, ..., x_n$, with probabilities of occurring $p_1, ..., p_n$, respectively. It is natural to summarize the dispersion in *X* by considering the typical difference $(x_i - x_j)^2,\ i,j = 1,..., n$, between the possible values of *X* (the squaring being done to avoid positive and negative differences from cancelling each other out). The probability of observing x_i and x_j successively being $p_i p_j$ (assuming independence), one is led to consider the average squared difference $\sum_{i,j} p_i p_j (x_i - x_j)^2$. There is only one problem with this quantity however; it contains twice the same information:

$$\sum_{i,j} p_i p_j (x_i - x_j)^2 = \sum_{i<j} p_i p_j (x_i - x_j)^2 + \sum_{i=j} p_i p_j (x_i - x_j)^2 + \sum_{i>j} p_i p_j (x_i - x_j)^2 = 2\sum_{i<j} p_i p_j (x_i - x_j)^2.$$

Therefore, a better (because non-redundant) measure of dispersion is $\sum_{i<j} p_i p_j (x_i - x_j)^2$.

Claim: this coincides with the usual notion of variance.

Proposition: The variance of a discrete random variable *X* whose possible values are $x_1, ..., x_n$ with probabilities $p_1, ..., p_n$ equals

$$\sigma^2 = \sum_{i<j} p_i p_j (x_i - x_j)^2 \qquad (1)$$

Proof: In all what follows, μ refers to the expected value of *X*;

$$2\sum_{i<j} p_i p_j (x_i - x_j)^2 = \sum_{i,j} p_i p_j (x_i - x_j)^2 = \sum_{i,j} p_i p_j (x^2{}_i - 2x_i x_j + x_j{}^2) = \sum_j p_j \sum_i (p_i x^2{}_i - 2p_i x_i x_j + p_i x_j{}^2)$$
$$= \sum_j p_j (\sum_i p_i x^2{}_i - 2\mu x_j + x_j{}^2) = \sum_i p_i x^2{}_i - 2\mu^2 + \sum_j p_j x_j{}^2 = 2(\sum_i p_i x^2{}_i - \mu^2) = 2\sigma^2.$$

Proposition: the sample variance of a discrete variable *X* taking on values $x_1, ..., x_n$ equals

$$s^2 = n^{-2}\sum_{i<j} (x_i - x_j)^2 \qquad (2)$$

.

Proof: This follows immediately from (1) by letting $p_i = p_j = 1/n$.

---

* inouasabiou@gmail.com

The formula (1) remains true for continuous variables in the following sense:

Proposition: The variance of a continuous random variable *X* whose support is D and density function f equals

$$\sigma^2 = \iint_{x<x'} (x - x')^2 f(x)f(x')dxdx' \tag{3}$$

where the shorthand $\mathrm{x} < \mathrm{x}'$ means "integration over the region $\{(x, x') \in D \times D : x < x'\}$".

Proof:

$$2\iint_{x<x'} (x - x')^2 f(x)f(x')dxdx' = \iint_{D\times D} (x - x')^2 f(x)f(x')dxdx' = \iint_{D\times D} (x^2 - 2xx' + x'^2)f(x)f(x')dxdx'$$
$$= \iint_{D\times D} x^2 f(x)f(x')dxdx' - 2\iint_{D\times D} xx' f(x)f(x')dxdx' + \iint_{D\times D} x'^2 f(x)f(x')dxdx'$$
$$= \int_D x^2 f(x)dx \int_D f(x')dx' - 2\int_D xf(x)dx \int_D x' f(x')dx' + \int_D x'^2 f(x')dx' \int_D f(x)dx = 2[E(X^2) - \mu^2] = 2\sigma^2.$$

It is interesting to notice that these formulas could have been more compactly established by showing that if X_1 and X_2 are two random variables independent but identically distributed as X, then:

$$E(X_1 - X_2)^2 = E(X_1^2 - 2X_1X_2 + X_2^2) = E(X_1^2) - 2E(X_1)E(X_2) + E(X_2^2) = 2[E(X^2) - \mu^2] = 2\sigma^2.$$