Random Variables

A *random variable* arises when we assign a numeric value to each elementary event that might occur. For example, if each elementary event is the result of a series of three tosses of a fair coin, then X = "the number of Heads" is a random variable. Associated with any random variable is its *probability distribution* (sometimes called its *density function*), which indicates the likelihood that each possible value is assumed. For example, $Pr(X=0) = 1/8$, $Pr(X=1) = 3/8$, $Pr(X=2) = 3/8$, and $Pr(X=3) = 1/8$.

The *cumulative distribution function* indicates the likelihood that the random variable is less-than-or-equal-to any particular value. For example, $Pr(X \leq x)$ is 0 for $x < 0$, 1/8 for $0 \leq x < 1$, 1/2 for $1 \leq x < 2$, 7/8 for $2 \leq x < 3$, and 1 for all $x \geq 3$.

Two random variables X and Y are *independent* if all events of the form "$X \leq x$" and "$Y \leq y$" are independent events. (Basically, X and Y are independent if knowing the value of one provides no information concerning the value of the other.)

The *expected value* of X is the average value of X, weighted by the likelihood of its various possible values. Symbolically,

$$E[X] = \sum_x x \cdot Pr(X = x)$$

where the sum is over all values taken by X with positive probability. Multiplying a random variable by any constant simply multiplies the expectation by the same constant, and adding a constant just shifts the expectation:

$$E[kX+c] = k \cdot E[X]+c .$$

For any event A, the *conditional expectation* of X given A is defined as $E[X|A] = \sum_x x \cdot Pr(X=x \mid A)$. A useful way to break down some calculations is via $E[X] = E[X|A] \cdot Pr(A) + E[X|A^c] \cdot Pr(A^c)$.

The expected value of the sum of several random variables is equal to the sum of their expectations, e.g.,

$$E[X+Y] = E[X]+ E[Y] .$$

On the other hand, the expected value of the product of two random variables is not necessarily the product of the expected values. For example, if they tend to be "large" at the same time, and "small" at the same time, $E[XY] > E[X] \cdot E[Y]$, while if one tends to be large when the other is small, $E[XY] < E[X] \cdot E[Y]$. However, in the special case in which X and Y are independent, equality *does* hold: $E[XY] = E[X] \cdot E[Y]$.

The *variance* of X is the expected value of the squared difference between X and its expected value:

$$\text{Var}[X] = E[(X-E[X])^2] = E[X^2] - (E[X])^2 .$$

(The second equation is the result of a bit of algebra: $E[(X-E[X])^2] = E[X^2 - 2 \cdot X \cdot E[X] + (E[X])^2] = E[X^2] - 2 \cdot E[X] \cdot E[X] + (E[X])^2$.) Variance comes in squared units (and adding a constant to a random variable, while shifting its values, doesn't affect its variance), so

$$\text{Var}[kX+c] = k^2 \cdot \text{Var}[X] .$$

What of the variance of the sum of two random variables? If you work through the algebra, you'll find that

$$\text{Var}[X+Y] = \text{Var}[X] + \text{Var}[Y] + 2 \cdot (E[XY] - E[X] \cdot E[Y]) .$$

This means that variances add when the random variables are independent, but not necessarily in other cases. The *covariance* of two random variables is $\text{Cov}[X,Y] = E[(X-E[X]) \cdot (Y-E[Y])] = E[XY] - E[X] \cdot E[Y]$. We can restate the previous equation as

$$\text{Var}[X+Y] = \text{Var}[X] + \text{Var}[Y] + 2 \cdot \text{Cov}[X,Y] .$$

Note that the covariance of a random variable with itself is just the variance of that random variable.

While variance is usually easier to work with when doing computations, it is somewhat difficult to interpret because it is expressed in squared units. For this reason, the *standard deviation* of a random variable is defined as the square-root of its variance. A practical interpretation is that the standard deviation of X indicates roughly how far from E[X] you'd expect the actual value of X to be.

Similarly, covariance is frequently "de-scaled," yielding the *correlation* between two random variables:

$$\text{Corr}(X,Y) = \text{Cov}[X,Y] / (\text{StdDev}(X) \cdot \text{StdDev}(Y)) .$$

The correlation between two random variables will always lie between -1 and 1, and is a dimensionless measure of the strength of the linear relationship between the two variables.

Continuous Random Variables

A ***continuous random variable*** is a random variable which can take any value in some interval. A continuous random variable is characterized by its *probability density function*, a graph which has a total area of 1 beneath it: The probability of the random variable taking values in any interval is simply the area under the curve over that interval.

The ***normal distribution***: This most-familiar of continuous probability distributions has the classic "bell" shape (see the left-hand graph below). The peak occurs at the mean of the distribution, i.e., at the expected value of the normally-distributed random variable with this distribution, and the standard deviation (the square root of the variance) indicates the spread of the bell, with roughly 68% of the area within 1 standard deviation of the peak.
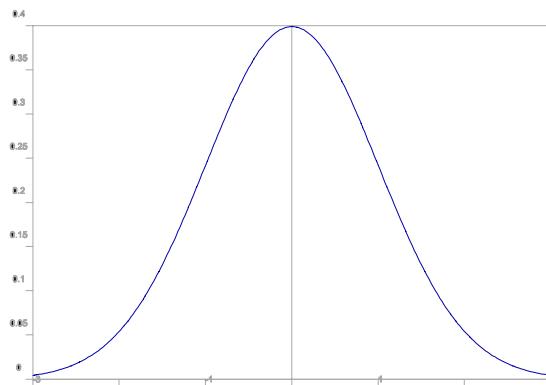
The normal distribution arises so frequently in applications due to an amazing fact: If you take a bunch of independent random variables (with comparable variances) and add them, the result will be roughly normally distributed, *no matter what the distributions of the separate variables might be*. (This is known as the "Central Limit Theorem".) Many interesting quantities (ranging from IQ scores in a demographically-homogeneous group, to demand for a retail product, to lengths of shoelaces) are actually a composite of many separate random variables, and hence are roughly normally distributed.

If X is normal, and Y = aX+b, then Y is also normal, with E[Y] = a·E[X] + b and StdDev[Y] = a·StdDev[X] . If X and Y are normal (independent or not), then X+Y and X-Y = X+(-Y) are also normal (intuition: the sum of two bunches is a bunch). Any normally-distributed random variable can be transformed into a "standard" normal random variable (with mean 0 and standard deviation 1) by subtracting off its mean and dividing by its standard deviation. Hence, a single tabulation of the cumulative distribution for a standard normal random variable can be used to do probabilistic calculations for *any* normally-distributed random variable.
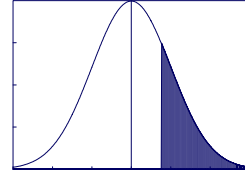
Right-Tail Probabilities of the Normal Distribution

		+0.01	+0.02	+0.03	+0.04	+0.05	+0.06	+0.07	+0.08	+0.09	+0.10
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641	0.4602
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247	0.4207
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859	0.3821
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483	0.3446
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121	0.3085
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776	0.2743
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451	0.2420
0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148	0.2119
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867	0.1841
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611	0.1587
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379	0.1357
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170	0.1151
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985	0.0968
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823	0.0808
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681	0.0668
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559	0.0548
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455	0.0446
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367	0.0359
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294	0.0287
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233	0.0228
2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183	0.0179
2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143	0.0139
2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110	0.0107
2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084	0.0082
2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064	0.0062
2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048	0.0047
2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036	0.0035
2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026	0.0026
2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019	0.0019
2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014	0.0013