

Events A_1, \dots, A_n are said to be mutually independent if for all subsets $S \subseteq \{1, \dots, n\}$, $p(\cap_{i \in S} A_i) = \prod p(A_i)$. (For example, flip a coin N times, then the events $\{A_i = i^{\text{th}} \text{ flip is heads}\}$ are mutually independent.)

Example: suppose events A, B , and C are pairwise independent, i.e., A and B are independent, B and C are independent, and A and C are independent. Note that this pairwise independence does not necessarily imply *mutual* independence of A, B , and C . To check that $p(\cap_{i \in S} A_i) = \prod_i p(A_i)$ for all subsets $S \subset \{A, B, C\}$ in this case means checking the non-trivial subsets with 2 or more elements: $\{A, B\}, \{A, C\}, \{B, C\}, \{A, B, C\}$. By assumption it follows for the first three, so the only one we need to check is $p(A, B, C) \stackrel{?}{=} p(A)p(B)p(C)$. But that this is not always the case can be seen by an explicit counterexample: consider tossing a fair coin three times, and consider the three events: A = the number of heads is even, B = the first two flips are the same, C = the second two flips are heads. It follows that $p(A) = p(B) = 1/2$, $p(C) = 1/4$, $p(A, B) = 1/4 = p(A)p(B)$, $p(A, C) = p(B, C) = 1/8 = p(A)p(C) = p(B)p(C)$; but $p(A, B, C) = 0 \neq p(A)p(B)p(C) = 1/16$.)

The complement of a set $A \subseteq S$ in S is denoted $\bar{A} = S - A$, i.e. the set of elements in S not contained in A . We can prove that an event A is independent of another event B if and only if A is independent of \bar{B} . To show this, first recall that if S can be written as the union of a set of non-intersecting subsets S_i : $S = \cup_i S_i$, $S_i \cap S_j = \phi$, then $p(A) = \sum_i p(A \cap S_i) = \sum_i p(A, S_i)$. The two sets $S_1 = B$, $S_2 = \bar{B}$ clearly satisfy these conditions, so we can write

$$p(A) = p(A, B) + p(A, \bar{B}) .$$

Note also that $p(B) + p(\bar{B}) = 1$. If A and B are independent, then by definition $p(A, B) = p(A)p(B)$ and substituting in the above results in $p(A, \bar{B}) = p(A)(1 - p(B)) = p(A)p(\bar{B})$, so A and \bar{B} are independent. In the opposite direction, if $p(A, \bar{B}) = p(A)p(\bar{B})$ then substitution in the above gives $p(A, B) = p(A)(1 - p(\bar{B})) = p(A)p(B)$, and A and B are independent.

Finally, note that the notions of “disjoint” and “independent” events are very different. Two events A, B are disjoint if their intersection is empty, whereas they are independent if $p(A, B) = p(A)p(B)$. Two events that are disjoint necessarily have $p(A, B) = p(A \cap B) = 0$, so if their independent probabilities are non-zero they are necessarily negatively correlated ($p(A, B) < p(A)p(B)$). For example, if we flip 2 coins, and event A = exactly 1 H, and event B = exactly 2 H, these are disjoint but not independent events: they’re negatively correlated since $p(A, B) = 0$ is less than $p(A)p(B) = (1/2)(1/4)$. Non-disjoint events can be positively or negatively correlated, or they can be independent. If we take event C = exactly 1T, then A and C are not disjoint (they’re equal): and they’re positively correlated since $p(A, C) = 1/2$ is greater than $p(A)p(C) = 1/4$. If we now flip 3 coins and let C = at least 1 H and at least one T, and D = at most 1 H. We see that $C \cap D = 1H$, and independence of events C, D follows from $p(C)p(D) = (6/8)(1/2) = 3/8 = p(C, D)$.

Random variables, mean and variance:

Suppose in a collection of people there are some number with height 6', and equal numbers with heights 5'11" and 6'1". The mean or average of this distribution is 6', as can be determined by summing the heights of all the people and dividing by the number of people, or equivalently by summing over distinct heights weighted by the fractional number of people with that height. Suppose for example, that the numbers in the above height categories are 5,30,5, then the latter calculation corresponds to $(1/8) \cdot 5'11" + (3/4) \cdot 6' + (1/8) \cdot 6'1" = 6'$. But the average gives only limited information about a distribution. Suppose there were instead only people with heights 5' and 7', and an equal number of each, then the average would still be 6' though these are very different distributions. It is useful to characterize the variation within the distribution from the mean. The average deviation from the mean gives zero due to equal positive and negative variations (as proven below), so the quantity known as the variance (or mean square deviation) is defined as the average of the *squares* of the differences between the values in the distribution and their mean. For the first distribution above, this gives the variance $V = \frac{1}{8}(-1")^2 + \frac{3}{4}(0")^2 + \frac{1}{8}(1")^2 = \frac{1}{4}(\text{inch})^2$, and for the second distribution the much larger result $V = \frac{1}{2}(-1')^2 + \frac{1}{2}(1')^2 = 1(\text{foot})^2$. The standard or r.m.s ("root mean square") deviation σ is defined as the square root of the variance, $\sigma = \sqrt{V}$. The above two distributions have $\sigma = (1/2 \text{ inch})$ and $\sigma = (1 \text{ foot})$ respectively.

More generally, a random variable is a function $X : S \rightarrow \mathbb{R}$, assigning some real number to each element of the probability space S . The average of this variable is determined by summing the values it can take weighted by the corresponding probability,

$$\langle X \rangle = \sum_{s \in S} p(s) X(s) .$$

(An alternate notation for this is $E[X] = \langle X \rangle$, for the "expectation value" of X .)

Example 1: roll two dice and let X be the sum of two numbers rolled. Thus $X(\{1, 1\}) = 2$, $X(\{1, 2\}) = X(\{2, 1\}) = 3$, ..., $X(\{6, 6\}) = 12$. The average of X is

$$\langle X \rangle = \frac{1}{36}2 + \frac{2}{36}3 + \frac{3}{36}4 + \frac{4}{36}5 + \frac{5}{36}6 + \frac{6}{36}7 + \frac{5}{36}8 + \frac{4}{36}9 + \frac{3}{36}10 + \frac{2}{36}11 + \frac{1}{36}12 = 7 .$$

Example 2: flip a coin 3 times, and let X be the number of tails. The average is

$$\langle X \rangle = \frac{1}{8}3 + \frac{3}{8}2 + \frac{3}{8}1 + \frac{1}{8}0 = \frac{3}{2} .$$

The expectation of the sum of two random variables X, Y satisfies $\langle X + Y \rangle = \langle X \rangle + \langle Y \rangle$. In general, they satisfy a "linearity of expectation" $\langle aX + bY \rangle = a\langle X \rangle + b\langle Y \rangle$ proven as follows: $\langle aX + bY \rangle = \sum_s p(s)(aX(s) + bY(s)) = a \sum_s p(s)X(s) + b \sum_s p(s)Y(s) = a\langle X \rangle + b\langle Y \rangle$. Thus an alternate way to calculate the mean of $X =$

$X_1 + X_2$ for the two dice rolls in example 1 above is to calculate the mean for a single die, $X_1 = (1+2+3+4+5+6)/6 = 21/6 = 7/2$, and so for two rolls $\langle X \rangle = \langle X_1 \rangle + \langle X_2 \rangle = 7/2 + 7/2 = 7$.

By definition, independent random variables X, Y satisfy $p(X=a \wedge Y=b) = p(X = a)p(Y = b)$ (i.e., the joint probability is the product of their independent probabilities, just as for independent events). For such variables, it follows that the expectation value of their product satisfies

$$\langle XY \rangle = \langle X \rangle \langle Y \rangle \quad (X, Y \text{ independent})$$

since $\sum_{r,s} p(r,s)X(r)Y(s) = \sum_{r,s} p(r)p(s)X(r)Y(s) = (\sum_r p(r)X(r))(\sum_s p(s)Y(s))$.

As indicated above, the average of the differences of a random variable from the mean vanishes: $\sum_{s \in S} p(s)(X(s) - \langle X \rangle) = \langle X \rangle - \langle X \rangle \sum_s p(s) = \langle X \rangle - \langle X \rangle = 0$. The variance of a probability distribution for a random variable is defined as the average of the squared differences from the mean,

$$V[X] = \sum_{s \in S} p(s)(X(s) - \langle X \rangle)^2. \quad (V1)$$

The variance satisfies the important relation

$$V[X] = \langle X^2 \rangle - \langle X \rangle^2, \quad (V2)$$

following directly from the definition above:

$$\begin{aligned} V[X] &= \sum_{s \in S} p(s)(X(s) - \langle X \rangle)^2 \\ &= \sum_s X^2(s)p(s) - 2\langle X \rangle \sum_s p(s)X(s) + \langle X \rangle^2 \sum_s p(s) \\ &= \langle X^2 \rangle - 2\langle X \rangle^2 + \langle X \rangle^2 = \langle X^2 \rangle - \langle X \rangle^2. \end{aligned}$$

In the case of independent random variables X, Y , as defined above, the variance is additive:

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

To see this, use (V2) together with $\langle XY \rangle = \langle X \rangle \langle Y \rangle$:

$$\begin{aligned} V[X + Y] &= \langle (X + Y)^2 \rangle - (\langle X \rangle + \langle Y \rangle)^2 \\ &= \langle X^2 \rangle + 2\langle XY \rangle + \langle Y^2 \rangle - \langle X \rangle^2 - 2\langle X \rangle \langle Y \rangle - \langle Y \rangle^2 \\ &= \langle X^2 \rangle - \langle X \rangle^2 + \langle Y^2 \rangle - \langle Y \rangle^2 = V[X] + V[Y]. \end{aligned}$$

Example: again flip a coin 3 times, and let X be the number of tails. $\langle X^2 \rangle = \frac{3}{8}4 + \frac{3}{8}1 + \frac{1}{8}9 = 3$ so $V[X] = 3 - (3/2)^2 = 3/4$. If we let $X = X_1 + X_2 + X_3$, where X_i

is the number of tails (0 or 1) for the i^{th} roll, then the X_i are independent variables with $\langle X_i \rangle = 1/2$ and $\langle X_i^2 \rangle = (1/2) \cdot 1 + (1/2) \cdot 0 = 1/2$, so $V[X_i] = 1/2 - 1/4 = 1/4$ (or equivalently $V[X_i] = 1/2(1/2)^2 + 1/2(-1/2)^2 = 1/8 + 1/8 = 1/4$). For the three rolls, $V[X] = V[X_1] + V[X_2] + V[X_3] = 1/4 + 1/4 + 1/4 = 3/4$, confirming the result above.

A Bernoulli trial is a trial with two possible outcomes: “success” with probability p , and “failure” with probability $1 - p$. The probability of r successes in N trials is

$$\binom{N}{r} p^r (1-p)^{N-r} .$$

Note the correct overall normalization automatically follows from $\sum_{r=0}^N \binom{N}{r} p^r (1-p)^{N-r} = [p + (1-p)]^N = 1^N = 1$. The overall probability for r successes is a competition between $\binom{N}{r}$, which is maximum at $r \sim N/2$, and $p^r (1-p)^{N-r}$ with is largest for small r when $p < 1/2$ (or large r for $p > 1/2$).

In class, we considered the case of rolling a standard six-sided die, with a roll of 6 considered a success, so $p = 1/6$. (See figures on next page for $N = 1, 2, 4, 10, 40, 80, 160, 320$ trials.) For a larger number N of trials, the distribution of expected number of successes becomes more narrowly peaked and more symmetrical about a fractional distance $r = N/6$.

To analyze this in the framework outlined above, let the random variable $X_i = 1$ if the i^{th} trial is success. Then $\langle X_i \rangle = p$. Let $X = X_1 + X_2 + \dots + X_N$ count the total number of successes. Then it follows that the average satisfies

$$\langle X \rangle = \sum_i \langle X_i \rangle = Np . \tag{B1}$$

From $V[X_i] = \langle X_i^2 \rangle - \langle X_i \rangle^2 = p - p^2 = p(1-p)$, it follows that the variance satisfies

$$V[X] = \sum_i V[X_i] = Np(1-p) , \tag{B2}$$

and the standard deviation is $\sigma = \sqrt{V[X]} = \sqrt{Np(1-p)}$. This explains the observation that the probability gets more sharply peaked as the number of trials increases, since the width of the distribution (σ) divided by the average $\langle X \rangle$ behaves as $\sigma/\langle X \rangle \sim \sqrt{N}/N \sim 1/\sqrt{N}$, a decreasing function of N .

By the “central limit theorem” (not proven in class), many such distributions under fairly relaxed assumptions always tend for sufficiently large number of trials to a “gaussian” or “normal” distribution, of the form

$$P(x) \approx \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} . \tag{G}$$

This is properly normalized, with $\int_{-\infty}^{\infty} dx P(x) = 1$, and also has $\int_{-\infty}^{\infty} dx x P(x) = \mu$, $\int_{-\infty}^{\infty} dx x^2 P(x) = \sigma^2 + \mu^2$, so the above distribution has mean μ and variance σ^2 . Setting $\mu = Np$ and $\sigma = \sqrt{Np(1-p)}$ for $p = 1/6$ in (G) thus gives a good approximation to the distribution of successful rolls of 6 for large number of trials in the example above.

