PROBABILITIES

Independent random variables
We can extend the definition of independence from events to random variables.

Definition 17.1 Let $X$ and $Y$ be random variables defined on the same probability space. We say that $X$ and $Y$ are independent, written $X \perp Y$, when

$$\forall x \in \text{Val}(X) \quad \forall y \in \text{Val}(Y) \quad [X = x] \perp [Y = y]$$

Pairwise and mutual independence of three or more random variables are defined analogously.

Proposition 17.2 (Independence of indicator variables) Let $A, B$ be two events in the probability space $(\Omega, \Pr)$. Then $I_A \perp I_B$ iff $A \perp B$

Proof: If $I_A \perp I_B$ then, by definition, $[I_A = 1] \perp [I_B = 1]$. Hence $A \perp B$.

Conversely, if $A \perp B$ then $\overline{A} \perp \overline{B}$, $A \perp \overline{B}$, and $A \perp \overline{A}$ by Proposition 14.12 part (iv), see Lecture 14. These cover all the cases needed for the definition of $I_A \perp I_B$. ■

This proposition can be extended to mutual independence of arbitrary number of random variables (we omit the statement and proof). When we defined random variables with binomial distribution we considered $n$ Bernoulli trials “performed independently”. This is the same as saying the the $n$ Bernoulli random variables corresponding to the $n$ trials are mutually independent.

Random variables that are not independent are quite common, see next.

Example 17.3 A fair coin is tossed twice. Let $X_H$ be random variable that returns the number of heads observed and $X_T$ the random variable that returns the number of tails observed. Are $X_H$ and $X_T$ independent?

Solution: Intuitively, the random variables are not independent, for example $X_H = 1$ forces $X_T = 1$. Let’s verify this in detail. The probability space consists of outcomes $\{HH, HT, TH, TT\}$ each with probability $1/4$. Then,

$$\Pr[X_T = 1 \mid X_H = 1] = \Pr[\{X_T = 1 \cap X_H = 1\}] / \Pr[X_H = 1] = (1/2) / (1/2) = 1$$
But \( \Pr[X_T = 1] = 1/2. \)

The probability space in the previous example can also be used to show that, in general, the expectation of a product of random variables does not behave as nicely as the expectation of a sum.

**Example 17.4**. Show that \( \mathbb{E}[X_HX_T] \neq \mathbb{E}[X_H] \mathbb{E}[X_T] \) where \( X_H \) and \( X_T \) are defined in Example 17.3.

**Solution:** \( X_H \) and \( X_T \) are both binomial random variables with parameters \( n = 2 \) and \( p = 1/2 \). We have computed the expectation of a binomial random variable in Example 16.17 (Lecture 16). Thus, \( \mathbb{E}[X_H] = \mathbb{E}[X_T] = (2)(1/2) = 1. \)

Now, for \( X_HX_T \), recall that the space of outcomes is \( \{HH,HT,TH,TT\} \). We observe that \( \text{Val}(X_HX_T) = \{0,1\} \). Now \( \Pr[X_HX_T = 0] = \Pr[\{HH,TT\}] = 1/4 + 1/4 = 1/2 \) and similarly \( \Pr[X_HX_T = 1] = \Pr[\{HT,TH\}] = 1/4 + 1/4 = 1/2. \) Therefore \( X_HX_T \) is actually Bernoulli with \( p = 1/2 \) and we have \( \mathbb{E}[X_HX_T] = 1/2 \neq 1 = \mathbb{E}[X_H] \mathbb{E}[X_T]. \)

The following important result is related to the previous two examples.

**Proposition 17.5** If \( X \perp Y \) then \( \mathbb{E}[XY] = \mathbb{E}[X] \mathbb{E}[Y] \).

**Proof:** Let’s first try to apply directly one of the definitions of expectation.

\[
\mathbb{E}[X] \mathbb{E}[Y] = \left( \sum_{w \in \Omega} X(w) \Pr[w] \right) \left( \sum_{w' \in \Omega} Y(w') \Pr[w'] \right) = \sum_{w \in \Omega} \sum_{w' \in \Omega} X(w)Y(w') \Pr[w] \Pr[w']
\]

Now, it’s not clear what to do next, since it is quite possible to have different outcomes \( w, w', w'' \) such that \( X(w)Y(w') = X(w'')Y(w'') \). We will follow a different strategy.

For each \( x \in \text{Val}(X) \) and each \( y \in \text{Val}(Y) \) let \( Z_{xy} \) be the indicator variable of the event \( [X = x] \cap [Y = y] \). Now comes a nice trick worth remembering. Observe that:

\[
XY = \sum_{x \in \text{Val}(X)} \sum_{y \in \text{Val}(Y)} xy Z_{xy}
\]

(It is quite worthwhile to think carefully why the above holds!) With this we can use linearity of expectation. First note that by the independence assumption \( \Pr[[X = x] \cap [Y = y]] = \Pr[X = x] \Pr[Y = y] \) hence \( \mathbb{E}[Z_{xy}] = \Pr[X = x] \Pr[Y = y] \). Now we have:

\[
\mathbb{E}[XY] = \sum_{x \in \text{Val}(X)} \sum_{y \in \text{Val}(Y)} xy \mathbb{E}[Z_{xy}] = \sum_{x \in \text{Val}(X)} \sum_{y \in \text{Val}(Y)} xy \Pr[X = x] \Pr[Y = y]
\]

\[
= \left( \sum_{x \in \text{Val}(X)} x \Pr[X = x] \right) \cdot \left( \sum_{y \in \text{Val}(Y)} y \Pr[Y = y] \right) = \mathbb{E}[X] \mathbb{E}[Y]
\]

2
Note 17.6 The previous proposition generalizes to more than three random variables: expectation distributes over the product of an arbitrary number of mutually independent random variables. The proof of this result is also a generalization of the proof above: consider the indicator variables of the events \[ X_1 = x_1 \cap \cdots \cap X_n = x_n. \]

We already knew that \( X_H \) and \( X_T \) in Examples 17.3 and 17.4 are not independent and \( \text{E}[X_H X_T] \neq \text{E}[X_H] \text{E}[X_T] \) provides, by the contrapositive of the proposition above, another proof of this. However, beware (!), the two are not equivalent. (Prove this!)

Variance
We are interested in better ways of calculating how a random variable deviates from its mean (expectation). This mean, \( \text{E}[X] \), is often denoted by \( \mu = \text{E}[X] \).

If we just take \( \text{E}[X - \mu] \) we get, by the linearity of expectation, \( \text{E}[X] - \text{E}[\mu] = \mu - \mu = 0 \) because the expectation of a constant is that same constant!

This is not very informative. While calculating the deviations from the mean we do not want the positive and the negative deviations to cancel out each other. This suggests that we should take the expectation of the absolute value of \( X - \mu \). But working with absolute values is messy. It turns out that squaring \( X - \mu \) is more useful. This leads to the following definition.

Definition 17.7 The variance of a random variable \( X \) is defined as

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

where \( \mu = \text{E}[X] \).

The standard deviation of a random variable \( X \) is

\[
\sigma(X) = \sqrt{\text{Var}[X]}
\]

and when the variable is clear from the context we use just \( \sigma \) for the standard deviation and \( \sigma^2 \) for the variance.

The standard deviation, by undoing the squaring in the variance, brings us back to values of similar magnitude as the those the random variable returns (and certainly with the same units of measurement). When the random variable is understood from the context, its mean is often denoted, as we saw, by \( \mu \) and its variance by \( \sigma^2 \).
Proposition 17.8

\[ \text{Var}[X] = E[X^2] - E[X]^2 \]

**Proof:** Using linearity of expectation and the fact that \( E[X] \) is a constant we have:

\[
\]

\[ \blacksquare \]

**Example 17.9** What is the variance of the number shown when a fair die is rolled? In other words, what is \( \text{Var}[D] \) where \( D \) was defined in Example 16.5 (Lecture 16).

**Solution:** Using the formula given by the Proposition 17.8

\[
E[D^2] - E[D]^2 = \frac{1}{6}(1 + 4 + 9 + 16 + 25 + 36) - \left( \frac{1}{6} (1 + 2 + 3 + 4 + 5 + 6) \right)^2 = \frac{35}{12} \approx 2.92
\]

\[ \blacksquare \]

**Example 17.10** What is the variance of a Bernoulli random variable \( X \) with parameter (probability of success) \( p \)?

**Solution:** Note that for Bernoulli random variables we have \( X^2 = X \). Hence \( \text{Var}[X] = E[X^2] - E[X]^2 = p - p^2 = p(1 - p) \).

\[ \blacksquare \]

Next we would like to compute the variance of a random variable with binomial distribution. We recall that these can be seen as the sum of independent identically distributed Bernoulli random variables. In general, variance does not have the linearity property, for two reasons. One is easy to handle: \( \text{Var}[cX] = c^2 \text{Var}[X] \). The other is that, again in general, \( \text{Var}[X + Y] \neq \text{Var}[X] + \text{Var}[Y] \). When the random variables are independent, however, nicer things happen (recall Proposition 17.5).

**Proposition 17.11** If \( X_1, \ldots, X_n \) are pairwise independent then

\[ \text{Var}[X_1 + \cdots + X_n] = \text{Var}[X_1] + \cdots + \text{Var}[X_n] \]
**Proof:** To keep the notation lighter we just sketch this for three random variables. Using the formula for variance given in Proposition [17.8](#) and using linearity of expectation, you can see that (verify!):

\[
\text{Var}[X + Y + Z] = \text{Var}[X] + \text{Var}[Y] + \text{Var}[Z] + \\
2(\text{E}[XY] - \text{E}[X]\text{E}[Y]) + 2(\text{E}[YZ] - \text{E}[Y]\text{E}[Z]) + 2(\text{E}[ZX] - \text{E}[Z]\text{E}[X])
\]

Now the result follows from Proposition [17.5](#). ■

**Note 17.12** It is interesting, and sometimes quite useful, that the fact that variance distributes over sums requires only pairwise, rather than mutual, independence. Note also that even that is not strictly necessary, as the proof uses only equalities of the form \(\text{E}[X_iX_j] = \text{E}[X_i]\text{E}[X_j]\). This property has a name, we say that \(X_i\) and \(X_j\) are **uncorrelated**. Thus, variance distributes over sums of pairwise uncorrelated random variables.

**Example 17.13** What is the variance of a binomial random variable \(B\) with parameters \(n\) and \(p\)?

**Solution:** As we have seen in Example 16.17 (Lecture 16) \(B = I_{S_1} + \cdots + I_{S_n}\) where \(S_i\) is the event “success in trial \(i\)” . Importantly, these events, hence their indicator random variables, are mutually independent so we can apply the previous proposition. As we have seen in Example 17.10 each of the \(I_{S_i}\) has variance \(p(1 - p)\). Therefore \(\text{Var}[B] = np(1 - p)\). ■

**Example 17.14** Recall the Hat Check Problem (Example 16.19, Lecture 16). What is the variance of the random variable \(X\) that returns the number of Al Capone associates who get their own hat back? Equivalently, let \(S\) be a set of \(n\) distinct real numbers. We sort them by producing a random permutation of the elements of \(S\). What is the variance of the random variable that returns the number of elements of \(S\) that end up in the right position?

**Solution:** Recall, from the solution of Example 16.19 that \(X = X_1 + \cdots + X_n\) where \(X_k = I_{E_k}\), the indicator random variable of the event \(E_k\), “\(a_k\) occurs in position \(k\)”.

In Example 13.10 (Lecture 13) we have computed \(\text{Pr}[E_k] = 1/n\) and in Example 16.19 we have computed \(\text{E}[X] = 1\).

Now, to compute \(\text{Var}[X]\) we cannot, unfortunately, apply Proposition 17.11 because, as we saw in Example 14.13, the events \(E_1, \ldots, E_n\) are not pairwise independent.

However, we try to calculate directly

\[
\text{E}[X^2] = \sum_{i=1}^{n} \text{E}[X_i^2] + 2 \sum_{i<j} \text{E}[X_i \cdot X_j]
\]
As noted before $X_i^2 = X_i$. This extends a bit, in that $X_i \cdot X_j$ turns out also to be a Bernoulli random variable, in fact the indicator variable of the event $E_i \cap E_j$!

Now, for $i \neq j$ we have, by the Chain Rule

$$\Pr[E_i \cap E_j] = \Pr[E_i] \Pr[E_j | E_i] = \frac{1}{n} \cdot \frac{1}{n-1}$$

(can also be computed as $(n-2)!/n!$.) Therefore

$$E[X^2] = \sum_{i=1}^{n} \frac{1}{n} + 2 \left( \frac{n(n-1)}{2} \right) \left( \frac{1}{n(n-1)} \right)$$

$$= n \cdot \frac{1}{n} + 1$$

$$= 2$$

Now $\text{Var}[X] = E[X^2] - E[X]^2 = 2 - 1^2 = 1$. ■

SUPPLEMENT: Probability Bounds
(this material is not required for homework or exams)

Tail bounds

In probabilistic analysis of algorithms we are often interested in how likely it is that a value taken by a random variable presents a “large deviation” from the mean (expectation). Such values are said to belong to the “tail” of the random variable’s distribution. The following result gives a bound on the probability of a large deviation, what we usually call a tail bound.

**Proposition 17.15 (Markov’s Inequality)** Let $X$ be an random variable that takes only non-negative values and let $a > 0$ be a constant. Then

$$\Pr[ X \geq a ] \leq \frac{\mu}{a}$$

where $\mu = E[X]$ is the expectation (mean) of $X$.

**Solution:** Follows from

$$E[X] = \sum_{x \in \text{Val}(X)} x \Pr[X = x] \geq \sum_{x \in \text{Val}(X)} x \Pr[X = x] \geq a \sum_{x \geq a} \Pr[X = x] = a \Pr[X \geq a]$$

■
Of course, for $a \leq \mu$ the statement is useless. Think rather of $a = 100\mu$. Then the inequality tells us that the probability that $X$ returns values bigger than 100 times its mean is very small. In view of this discussion, you may find the following formulation of Markov’s Inequality more suggestive. For $c > 0$, we have

$$\Pr[ X \geq c \mu ] \leq \frac{1}{c}$$

Note that this form requires $E[X] > 0$, since $\Pr[X \geq 0] = 1$ for a random variable that takes only nonnegative values. This is not a significant restriction since the only way a nonnegative random variable can have expectation 0 is when it is identically zero.

**Example 17.16** Suppose we flip a fair coin $n$ times. Using Markov’s Inequality bound the probability of obtaining at least $3n/4$ heads.

**Solution:** Let $X$ be the random variable denoting the total number of heads in $n$ flips of a fair coin. This variable has a binomial distribution with parameters $n$ and 1/2. We know that $\mu = E[X] = n/2$. Applying the above inequality we get

$$\Pr[ X \geq \frac{3n}{4}] = \Pr[ X \geq \frac{3}{2} \cdot \frac{n}{2}] \leq \frac{1}{\frac{3}{2}} = \frac{2}{3}$$

How good is this bound?…

**Example 17.17** Suppose we roll a fair die. Using Markov’s Inequality bound the probability of obtaining a number greater than or equal to 7.

**Solution:** Let $D$ be the random variable denoting the result of the roll of a die. We know that $E[D] = 3.5$. Using Markov’s Inequality we get $\Pr[X \geq 7] \leq 1/2$. In fact, $\Pr[X \geq 7] = 0$! As this result shows, Markov’s Inequality gives a bound that can be quite loose.

**Concentration bounds**

The concept of variance allows us to derive bounds on how well the values taken by a random variable concentrate around its mean (expectation). Such concentration bounds also give (by taking the complement) two-sided tail bounds. This following result is, in itself, enough reason to study variance.

**Proposition 17.18 (Chebyshev’s Inequality)** Let $X$ be a random variable. For any $a > 0$ we have

$$\Pr[ |X - \mu| \geq a ] \leq \frac{\sigma^2}{a^2}$$
Proof: Note that $\Pr[|X - \mu| \geq a] = \Pr[(X - \mu)^2 \geq a^2]$. This suggests that we apply Markov’s Inequality (Proposition 17.15) to the random variable $Y = (X - \mu)^2$ which takes nonnegative values. We obtain $\Pr[Y \geq a^2] \leq E[Y]/a^2$ but $E[Y] = E[(X - E[X])^2] = \text{Var}[X]$ by the definition of variance. Chebyshev’s Inequality follows.

Let’s put Chebyshev’s Inequality in an equivalent but perhaps more suggestive form. Recall that we denote by $\sigma = \sqrt{\text{Var}[X]}$ the standard deviation of $X$ and by $\mu = E[X]$ its mean. Assuming $\text{Var}[X] \neq 0$ we can take $a = c\sigma$ and we obtain

$$\Pr[X \leq \mu - c\sigma \text{ or } X \geq \mu + c\sigma] \leq \frac{1}{c^2}$$

We can read this as a two-sided tail bound. Equivalently we have the following concentration bound:

$$\Pr[\mu - c\sigma \leq X \leq \mu + c\sigma] > 1 - \frac{1}{c^2}$$

The next example shows that Chebyshev’s Inequality can derive a significantly better tail bound than the one given by Markov’s Inequality.

Example 17.19 Use Chebyshev’s Inequality to bound the probability of obtaining at least $3n/4$ heads in a sequence of $n$ fair coin flips.

Begin sol Let $X$ denote the random variable denoting the total number of heads that result during $n$ flips of a fair coin. This random variable has binomial distribution with parameters $n$ and $1/2$. Hence $E[X] = n/2$ and $\text{Var}[X] = n(1/2)(1 - 1/2) = n/4$.

Now $|X - n/2| \geq n/4$ iff $X - n/2 \geq n/4$ or $X - n/2 \leq -n/4$ iff $X \geq 3n/4$ or $X \leq n/4$. The last two events are disjoint therefore

$$\Pr[|X - n/2| \geq n/4] = \Pr[X \geq 3n/4] + \Pr[X \leq n/4]$$

The distribution of $X$ is “symmetric” around its mean, $n/2$, and we have $\Pr[X \geq 3n/4] = \Pr[X \leq n/4]$ (to see this, interchange heads and tails and perform the flips in reverse).

Now, applying Chebyshev’s Inequality

$$\Pr[X \geq 3n/4] = \frac{1}{2}\Pr[|X - n/2| \geq n/4] \leq \frac{1}{2} \cdot \frac{n/4}{(n/4)^2} = \frac{2}{n}$$

This is a much better tail bound than the $2/3$ that we obtained from Markov’s Inequality in Example 17.16. Moreover it supports the intuition that as $n$ gets larger, the number of heads concentrates around $n/2$.

Problem 17.20 Use Markov and (separately) Chebyshev’s Inequality to provide bounds to the probability that at least $k \geq 1$ of Al Capone’s associates get their own hat back (see Example 17.14).

Solve this!