Discrete Probabilities

Consider a set Ω of the possible individual outcomes of an experiment. In the roll of a die, Ω could be the set of six possible values. In the roll of three dice, Ω could be the set of $6^3 = 216$ combinations of values. In the sand-hill crane experiment, Ω could be the set of possible population sizes in year 7, or even the set of all possible sequences $y(n)$ of populations over 20 years, one sequence forming a single element ω in Ω. The set Ω is called the *universe*, because it considers all conceivable outcomes of interest.

The set $\mathcal{E}$ of *events* built on Ω is the set of all possible subsets of elements taken from Ω. For the roll of a die, the event set $\mathcal{E}$ is the following set of $2^6 = 64$ subsets:

$$\begin{aligned}\mathcal{E} \;=\; & \{\emptyset, \{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}, \{1,2\}, \{1,3\}, \{1,4\}, \{1,5\}, \{1,6\},\\ & \{2,3\}, \{2,4\}, \{2,5\}, \{2,6\}, \{3,4\}, \{3,5\}, \{3,6\}, \{4,5\}, \{4,6\}, \{5,6\},\\ & \{1,2,3\}, \{1,2,4\}, \{1,2,5\}, \{1,2,6\}, \{1,3,4\}, \{1,3,5\}, \{1,3,6\}, \{1,4,5\},\\ & \{1,4,6\}, \{1,5,6\}, \{2,3,4\}, \{2,3,5\}, \{2,3,6\}, \{2,4,5\}, \{2,4,6\}, \{2,5,6\},\\ & \{3,4,5\}, \{3,4,6\}, \{3,5,6\}, \{4,5,6\}, \{1,2,3,4\}, \{1,2,3,5\}, \{1,2,3,6\},\\ & \{1,2,4,5\}, \{1,2,4,6\}, \{1,2,5,6\}, \{1,3,4,5\}, \{1,3,4,6\}, \{1,3,5,6\}, \{1,4,5,6\},\\ & \{2,3,4,5\}, \{2,3,4,6\}, \{2,3,5,6\}, \{2,4,5,6\}, \{3,4,5,6\}, \{1,2,3,4,5\},\\ & \{1,2,3,4,6\}, \{1,2,3,5,6\}, \{1,2,4,5,6\}, \{1,3,4,5,6\}, \{2,3,4,5,6\}, \{1,2,3,4,5,6\}\}\end{aligned}$$

Conceptually, it is important to note that, say, $\{1,4\}$ above is really a shorthand for

$$\{\text{“the die has produced value } 1\text{”}, \text{“the die has produced value } 4\text{”}\}\ .$$

The first event, $\emptyset$, is the empty set (which is a subset of any set), and the next six events are called *singletons* because they comprise a single outcome each. The largest event is Ω itself (last in the list above). Note that events are *not* repetitions of outcomes. For instance, the event $\{1,4,6\}$ does not mean “first 1, then 4, then 6,” but rather “either 1 or 4 or 6.”

The Probability Function

A (discrete) probability function is a function P from the event set $\mathcal{E}$ to the real numbers that satisfies the following properties:

$$\begin{aligned}\mathcal{P}_1 &: P(E) \geq 0 \quad \text{for every} \quad E \in \mathcal{E}\\ \mathcal{P}_2 &: P(\Omega) = 1\\ \mathcal{P}_3 &: E \cap F = \emptyset \rightarrow P(E \cup F) = P(E) + P(F) \quad \text{for all} \quad E, F \in \mathcal{E}\ .\end{aligned}$$

A probability function can be viewed as a *measure* for sets in the event space $\mathcal{E}$, normalized so that the universe $\Omega \in \mathcal{E}$ has unit measure (property $\mathcal{P}_2$).

Property $\mathcal{P}_1$ states that measures are nonnegative, and property $\mathcal{P}_3$ reflects additivity: if we measure two separate (disjoint) sets, their sizes add up. For instance, the event $E = \{2, 4, 6\}$ has measure $1/2$ with the probabilities defined for a die roll. The event E is greater than the event $F = \{1, 3\}$, which has only measure $1/3$. Since E and F are disjoint, their union $E \cup F$ has measure $1/2 + 1/3 = 5/6$.

The event set is large, and the properties above imply that probabilities cannot be assigned arbitrarily to events. For instance, the empty set must be given probability zero. Since

$$\emptyset \cap E = \emptyset \quad \text{for any} \quad E \in \mathcal{E} \ ,$$

property $\mathcal{P}_3$ requires that

$$P(\emptyset \cup E) = P(\emptyset) + P(E) \ .$$

However,

$$\emptyset \cup E = E \ ,$$

so

$$P(E) = P(\emptyset) + P(E) \ ,$$

that is,

$$P(\emptyset) = 0 \ .$$

Independence and Conditional Probability

Two events E, F in $\mathcal{E}$ are said to be mutually *independent* if

$$P(E \cap F) = P(E)\, P(F) \ .$$

For instance, the events $E = \{1, 2, 3\}$ and $F = \{3, 5\}$ in the die roll are mutually independent:

$$P(E) = 1/2 \,, \ P(F) = 1/3 \quad \text{and} \quad P(E \cap F) = P(\{3\}) = 1/6$$

so that

$$P(E \cap F) = P(E)P(F) = 1/6 \ .$$

There is little intuition behind this definition of independence. To understand its importance, we introduce the notion of conditional probability:

Let F be an event of nonzero probability. Then, the *conditional probability of an event E given F* is defined as follows:

$$P(E \mid F) = \frac{P(E \cap F)}{P(F)} \ . \tag{1}$$

Note first that independence of E and F (assuming $P(F) > 0$) is equivalent to

$$P(E \mid F) = \frac{P(E \cap F)}{P(F)} = \frac{P(E)P(F)}{P(F)} = P(E) \ ,$$

that is,

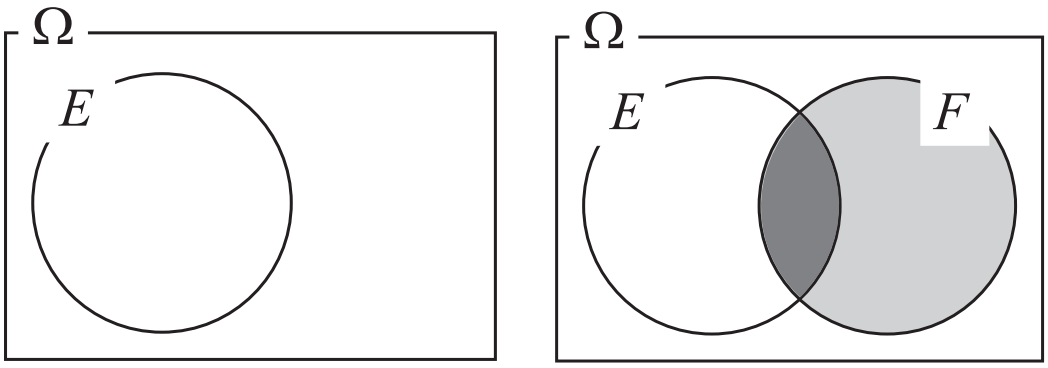
Any two events E, F in the event set $\mathcal{E}$ are mutually *independent* if

$$P(E \cap F) = P(E)\, P(F) \ .$$

If $P(F) > 0$, the two events E and F are mutually independent if and only if

$$P(E \mid F) = P(E) \ . \tag{2}$$

Both the notion of conditional probability and that of independence can be given a useful, intuitive meaning. Because of normalization ($\mathcal{P}(\Omega) = 1$), the probability of an event E is the fraction of universe Ω covered by E, as measured by $P(\cdot)$ (see the Venn diagram in Figure 1(a)). From the definition (1) we see that the conditional probability $P(E \mid F)$ measures the fraction of the area of F covered by the intersection $E \cap F$ (Figure 1 (b)). Thus, conditioning by F redefines the universe to be F, and excludes from consideration the part of event E (or of any other event) that is outside the new universe. In other words, $P(E \mid F)$ is the probability of the part of event E that is consistent with F, and re-normalized to the measure of F: Given that we know that F has occurred, what is the new probability of E?



(a) (b)

Figure 1: (a) The probability of an event E can be visualized as the fraction of the unit area of universe Ω that is covered by E. (b) The conditional probability of E given F redefines the new universe as F (both shaded areas), and only considers the area of the part of E in the new universe, that is, of $E \cap F$ (darker shading).

For example, suppose that we are interested in the probability of the event $E = \{4, 5, 6\}$ (either a 4, a 5, or a 6 is produced) in a single roll of a die. The (unconditional) probability of E is $1/2$.

We are subsequently told that the last roll has produced an odd number, so event $F = \{1, 3, 5\}$ has occurred (we have not seen the roll, so we do not know *which* odd number has come out). The conditional probability $P(E \mid F)$ measures the probability of E given that we know that event F has occurred. The two outcomes 4 and 6 in E are now inconsistent with F, and $E \cap F = \{5\}$ lists the only remaining possibility in favor of E. The new universe F collects all possible outcomes in light of the newly available information, and the probability of E *post facto* is

$$P(E \mid F) = \frac{P(E \cap F)}{P(F)} = \frac{P(\{5\})}{P(\{1, 3, 5\})} = \frac{1/6}{1/2} = 1/3 \; .$$

Thus, the fact that F has occurred has shrunk the probability of E from $1/2$ to $1/3$.

Of course, conditioning can also increase probabilities. For instance, with $E = \{4, 5, 6\}$ and $F = \{2, 4, 6\}$ we have $P(E) = 1/2$ and $P(E \mid F) = 2/3$ (verify this!).

Equation (2) then shows that *for a conditioning event F with nonzero probability, independence of E and F means that the occurrence of F does not alter the probability of E.* Before or after we know that F has occurred, the probability of E remains the same. In the two examples above, the events E and F are not independent (we say that they are *dependent* on each other). Verify that the events $E = \{2, 3\}$ and $F = \{1, 3, 5\}$ are independent.

The following fact is easy to prove (and comforting!):

Conditional probabilities are indeed probabilities, that is, a measure on the original universe Ω, in that the following three properties hold for any F such that $P(F) > 0$:

$$\begin{aligned}
&P(E \mid F) \geq 0 \quad \text{for every} \quad E \in \mathcal{E} \\
&P(\Omega \mid F) = 1 \\
&E \cap E' = \emptyset \rightarrow P(E \cup E' \mid F) = P(E \mid F) + P(E' \mid F) \quad \text{for all} \quad E, E' \in \mathcal{E} \; .
\end{aligned}$$

Here are the proofs:

$$P(E \mid F) = \frac{P(E \cap F)}{P(F)} \geq 0$$

(immediate).

$$P(\Omega \mid F) = \frac{P(\Omega \cap F)}{P(F)} = \frac{P(F)}{P(F)} = 1$$

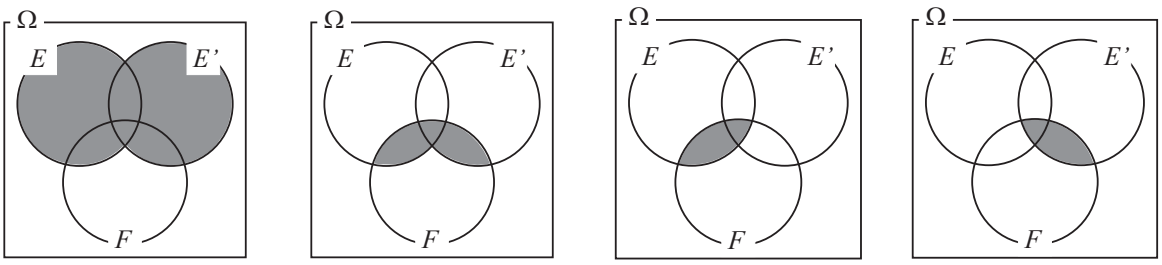
because F is a subset of Ω and so $\Omega \cap F = F$. Finally,

$$\begin{aligned}
P(E \cup E' \mid F) &= \frac{P((E \cup E') \cap F)}{P(F)} = \frac{P((E \cap F) \cup (E' \cap F))}{P(F)} = \frac{P(E \cap F) + P(E' \cap F)}{P(F)} \\
&= \frac{P(E \cap F)}{P(F)} + \frac{P(E' \cap F)}{P(F)} = P(E \mid F) + P(E' \mid F) \; .
\end{aligned}$$

In the second equality above we used De Morgan's law for the intersection with a union (see Figure 2):

$$(E \cup E') \cap F = (E \cap F) \cup (E' \cap F) \tag{3}$$

and in the third equality we used (i) the fact that if E and E' are disjoint so are the (smaller) sets $E \cap F$ and $E' \cap F$, and (ii) property $\mathcal{P}_3$ for probabilities.



$E \cup E'$ (a) $\quad (E \cup E') \cap F = (E \cap F) \cup (E' \cap F)$ (b) $\quad E \cap F$ (c) $\quad E' \cap F$ (d)

Figure 2: De Morgan's law (3) for the intersection with a union: The shaded area in (b) can be obtained in two ways: (i) first take the union (a) of E and E' and then intersect with F; or (ii) first intersect (c) E with F and (d) E' with F, and then take the union of the two.

Random Variables and Probability Distributions

A *discrete random variable* X is a way to (i) group outcomes into sets of outcomes to be considered equivalent for the purpose at hand, and to (ii) number these groups. Specifically:

> Any function from the universe Ω of outcomes to (a subset of) the integers
>
> $$X \ : \ \Omega \to \mathbb{Z}$$
>
> is a (discrete) random variable.

Translating outcomes (or rather groups of outcomes) into numbers is a crucial step towards providing statistical summaries of outcomes in experiments.

While there is no restriction on what function to use as a random variable, only some functions end up being useful random variables. For instance, a natural random variable to define for the roll of a die is the function X that assigns to the outcome ω =“the die has produced value x” the number $X(\omega) = x$ for $x = 1, 2, 3, 4, 5, 6$. The function that assigns the number 10 to every outcome is a random variable as well, but not a useful one.

The key insight for assigning probabilities to events in a manner that is consistent with the properties $\mathcal{P}_1$, $\mathcal{P}_2$, $\mathcal{P}_3$ is that every event that can be expressed in terms of a random variable can be built as a union of the disjoint sets

$$X^{-1}(x) = \{\omega \ : \ X(\omega) = x\} \quad \text{for} \quad x \in \mathcal{X} = X(\Omega) \ .$$

This idea leads to the notion of a probability distribution, which assigns probabilities to distinct values of a discrete random variable. Formally:

> Given a universe Ω and a random variable X on Ω, a (discrete) *probability distribution* is a function p from $\mathcal{X} = X(\Omega)$ to the set of real numbers between 0 and 1 defined by
>
> $$p(x) = P[X^{-1}(x)] \quad \text{for all} \quad x \in \mathcal{X} \ ,$$
>
> that is, the function that assigns probabilities to all events that map to each distinct value x of the random variable.

Since X is a function, the sets $X^{-1}(x)$ are disjoint subsets and their union covers all of Ω. Therefore, properties $\mathcal{P}_2$ and $\mathcal{P}_3$ imply that

$$\sum_{x \in \mathcal{X}} p(x) = 1 \ .$$

For instance, the probability distribution for the roll of a fair die is the constant function

$$p(x) = \frac{1}{6} \quad \text{for} \quad x = 1, 2, 3, 4, 5, 6 \ ,$$

and the probability of the event $\{1, 4, 6\}$ is $p(1) + p(4) + p(6) = 1/6 + 1/6 + 1/6 = 1/2$. This distribution is called the *uniform probability distribution* over the six possible values of x, and is sometimes denoted by $p_{U_6}(x)$. More generally,

$$p_{U_n}(k) = \frac{1}{n} \quad \text{for} \quad 1 \le k \le n \ \ \text{(or for } \ 0 \le k \le n-1 \ \ \text{if more convenient)}.$$

The fact that in this example there is a distinct value of the random variable for each outcome is a coincidence. A different random variable on the universe of die roll outcomes could be defined to assign a value of zero to each even outcome and a value of one to each odd outcome:

$$X('1') = X('3') = X('5') = 0 \quad \text{and} \quad X('2') = X('4') = X('6') = 1$$

where, again, $'i'$ means "the die has produced value i." This random variable, for a fair die, has probability distribution $p_{U_2}(k) = 1/2$ for $k = 0, 1$.

The single-attempt Russian roulette experiment has universe $\Omega = \{\text{“}survive,\text{”} \text{“}die\text{”}\}$, random variable $X(\text{“}die\text{”}) = 0$ and $X(\text{“}survive\text{”}) = 1$, and probability distribution $p(0) = q$ and $p(1) = p$ with $p + q = 1$.

A Russian roulette experiment with k attempts (on different birds) has a universe Ω with 2^k combinations of "survive" and "die," and a huge event space with 2^{2^k} sets.

Joint, Independent, and Conditional Distributions

The notions of independence and conditioning we introduced for events can be applied to probability distributions as well. This requires introducing the concept of a joint distribution.

Let $\mathcal{X} = X(\Omega)$ and $\mathcal{Y} = Y(\Omega)$ be the images of the universe Ω through two random variables X and Y. A *joint probability distribution* $p_{X,Y}(x, y)$ for X and Y is a probability distribution defined on $\mathcal{X} \times \mathcal{Y}$, the cross product[a] of $\mathcal{X}$ and $\mathcal{Y}$.

[a] The cross product of two sets $\mathcal{X}$ and $\mathcal{Y}$ is the set of all pairs (x, y) with $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.

Thus, there is essentially no difference between a probability distribution and a joint probability distribution, except for some more structure in the domain of the latter. However, the joint distribution completely determines the distributions of X and Y. This is because

$$\begin{aligned}\sum_{y\in\mathcal{Y}} p_{X,Y}(x,y) &= \sum_{y\in\mathcal{Y}} P(X = x \cap Y = y) = P(\bigcup_{y\in\mathcal{Y}}(X = x \cap Y = y)) \\ &= P(X = x \cap (\bigcup_{y\in\mathcal{Y}} Y = y)) = P((X = x) \cap \Omega) = P(X = x) = p_X(x)\ .\end{aligned}$$

The six equalities in this chain are justified by the following: (i) the definition of joint probability; (ii) the fact that different values of Y yield disjoint events, together with property $\mathcal{P}_3$; (iii) De Morgan's law; (iv) the fact that the union of all values for Y yields all of Ω; (v) the fact that $X = x$ is a subset of Ω; and (vi) the definition of p_X. Analogous reasoning holds for p_Y, so we can summarize as follows:

$$p_X(x) = \sum_{y\in\mathcal{Y}} p_{X,Y}(x,y) \tag{4}$$

$$p_Y(y) = \sum_{x\in\mathcal{X}} p_{X,Y}(x,y)\ . \tag{5}$$

If we arrange the joint probabilities $p_{X,Y}(x, y)$ into an array with one row for each value of x and one column for each value of y, then the probabilities $p_X(x)$ and $p_Y(y)$ are the sums in each row and in each column, respectively. Here is an example of a joint probability of random variable X with values $1, 2$ and random variable Y with values $1, 2, 3$:

$p_{X,Y}(x,y)$	1	2	3	$p_X(x)$
1	**0.2**	**0.1**	**0.25**	0.55
2	**0.1**	**0.15**	**0.2**	0.45
$p_Y(y)$	0.3	0.25	0.45	*1*

. (6)

The values in bold are the joint probabilities $p_{X,Y}(x, y)$. Since the values of $p_X(x)$ and $p_Y(y)$ are naturally written in the margins of this table, they are called the *marginal probability distributions*.

The *1* in the bottom right cell emphasizes the fact that the elements in the table add up to one, and so does each of the two marginal distributions (check this!).

We say that the random variables X and Y are *independent* if the events for every pair of values for X and Y are independent, that is, if

$$p_{X,Y}(x, y) = p_X(x)\, p_Y(y) \ .$$

It is immediate to verify that the joint probability $p_{X,Y}(x, y)$ in table (6) does not satisfy this property. For instance, $p_{X,Y}(1, 1) = 0.2$, but $p_X(1)p_Y(1) = 0.55 \times 0.3 = 0.165$, and one counterexample is enough. This also suggests how to build a joint probability distribution under the assumption of independence: first specify the marginals, and then multiply them together to obtain the joint. With the marginals in table (6), we obtain the following joint distribution:

$p_{X,Y}(x, y)$	1	2	3	$p_X(x)$
1	**0.165**	**0.1375**	**0.2475**	0.55
2	**0.135**	**0.1125**	**0.2025**	0.45
$p_Y(y)$	0.3	0.25	0.45	*1*

. (7)

Suppose that

$$p_Y(y) > 0 \quad \text{for all} \quad y \in \mathcal{Y} \ .$$

Then, the *conditional probability distribution* of random variable X given random variable Y is the following function of x and y:

$$p_{X|Y}(x \mid y) = \frac{p_{X,Y}(x, y)}{p_Y(y)} \ . \tag{8}$$

Finally, the same reasoning used for the analogous result for events shows the following:

If $p_Y(y)$ is nonzero for all y, the two random variables X and Y are mutually independent if and only if

$$p_{X|Y}(x \mid y) = p_X(x) \quad \text{for all} \quad x \in \mathcal{X} \text{ and } y \in \mathcal{Y} \ .$$

The classical example of independent random variables is the repeated Bernoulli trial. The *Bernoulli trial* is a binary event that has a probability p to occur, and a probability $q = 1 - p$ of not occurring. The *Bernoulli random variable* B_1 associates the value $B_1 = 1$ to occurrence, and $B_1 = 0$ to non-occurrence, so the probability distribution for the Bernoulli random variable is

$$p_{B_1}(0) = q \quad \text{and} \quad p_{B_1}(1) = p \ . \tag{9}$$

The subscript 1 refers to the fact that there is a single trial. Suppose now that the Bernoulli trial is repeated n times (think of determining for each of n sand-hill cranes if they will survive winter or

not). There are 2^n possible combinations of values for the n random variables, each corresponding to a different vector $(x_1, \ldots, x_n)$ with entries equal to either one or zero:

$$\begin{array}{l} 0\ldots00 \\ 0\ldots01 \\ 0\ldots10 \\ 0\ldots11 \\ \vdots \\ 1\ldots11 \end{array}$$

Each of these vectors could be assigned arbitrary probabilities $p(x_1, \ldots, x_n)$ as long as they are non-negative and they add up to one. If, on the other hand, the trials are assumed to be independent and have the same distribution (9), then

$$p_{B_n}(x_1, \ldots, x_n) = p_{B_1}(x_1) \cdot \ldots \cdot p_{B_1}(x_n)$$

so a vector $(x_1, \ldots, x_n)$ that has k ones and therefore $n-k$ zeros has probability $p^k q^{n-k}$. For instance, with $n = 3$ and $p = 0.7$ we have $q = 1 - 0.7 = 0.3$, and the probability distribution for the trice-repeated Bernoulli trial is

$$\begin{array}{rcl} p_{B_3}(0,0,0) &=& q^3 = 0.027 \\ p_{B_3}(0,0,1) &=& pq^2 = 0.063 \\ p_{B_3}(0,1,0) &=& pq^2 = 0.063 \\ p_{B_3}(0,1,1) &=& p^2q = 0.147 \\ p_{B_3}(1,0,0) &=& pq^2 = 0.063 \\ p_{B_3}(1,0,1) &=& p^2q = 0.147 \\ p_{B_3}(1,1,0) &=& p^2q = 0.147 \\ p_{B_3}(1,1,1) &=& p^3 = 0.343\ . \end{array} \tag{10}$$

Check that these numbers add up to 1, so that this is a valid probability distribution.

Incidentally, this construction in the case $n = 2$ is analogous to the procedure used to construct (7): first fill in the marginals, then multiply them to obtain the joint distribution:

$p_{X,Y}(x,y)$	0	1	$p_X(x)$
0	$\boldsymbol{q^2}$	$\boldsymbol{pq}$	q
1	$\boldsymbol{pq}$	$\boldsymbol{p^2}$	p
$p_Y(y)$	q	p	*1*

.

For $n > 2$ this is still doable in principle, but now the table becomes three-dimensional for $n = 3$, and so forth, so the different format used above for the computation in the general case is preferable.

The *binomial random variable* b_n counts the number of occurrences of a 1 in the vector $(x_1, \ldots, x_n)$ of outcomes in a repeated Bernoulli trial. The number of occurrences is between zero and n. For instance, in the example above b_3 is either 0, 1, 2, or 3.

Since different random variable values correspond to disjoint events, property $\mathcal{P}_3$ of probabilities implies that the probability that there are k ones in $(x_1, \ldots, x_n)$ is the sum of the probabilities $p_{B_n}(x_1, \ldots, x_n)$ for all the vectors with k ones. In the example,

$$\begin{aligned} p_{b_3}(0) &= p_{B_3}(0,0,0) = q^3 = 0.027 \\ p_{b_3}(1) &= p_{B_3}(0,0,1) + p_{B_3}(0,1,0) + p_{B_3}(1,0,0) = 3pq^2 = 0.189 \\ p_{b_3}(2) &= p_{B_3}(0,1,1) + p_{B_3}(1,0,1) + p_{B_3}(1,1,0) = 3p^2q = 0.441 \\ p_{b_3}(3) &= p_{B_3}(1,1,1) = p^3 = 0.343 \end{aligned}$$

(again, these add up to one, this time quite obviously).

More generally, since the number of binary vectors $(x_1, \ldots, x_n)$ with k ones is

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!} = \frac{n(n-1)\cdot\ldots\cdot(n-k+1)}{k(k-1)\cdot\ldots\cdot 2}$$

we have the following formula for the probability distribution of a binomial random variable b_n with parameter p:

$$p_{b_n}(k) = \binom{n}{k} p^k q^{n-k} \quad \text{for} \quad 0 \le k \le n\ .$$

The formula for the expansion of the power of a binomial[1] shows immediately that these probabilities add up to one:

$$\sum_{k=0}^{n} p_{b_n}(k) = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} = (p+q)^n = 1^n = 1\ .$$

Figure 3 shows plots of the binomial distribution for different values of the parameters n and p.

Moments

Often a summary of the distribution of a random variable is more useful than the full distribution itself. The following are the most widely used summaries:

The *mean* and the *centered moments* of a probability distribution $p(x)$ are defined as follows:

$$\begin{aligned} \text{mean} \quad &: \quad m_X = \sum_{x\in\mathcal{X}} x p(x) \\ i\text{-th centered moment} \quad &: \quad \mu_X(i) = \sum_{x\in\mathcal{X}} (x - m_X)^i p(x) \quad \text{for} \quad i \ge 2\ . \end{aligned}$$

[1] This is where the name "binomial" random variable comes from.

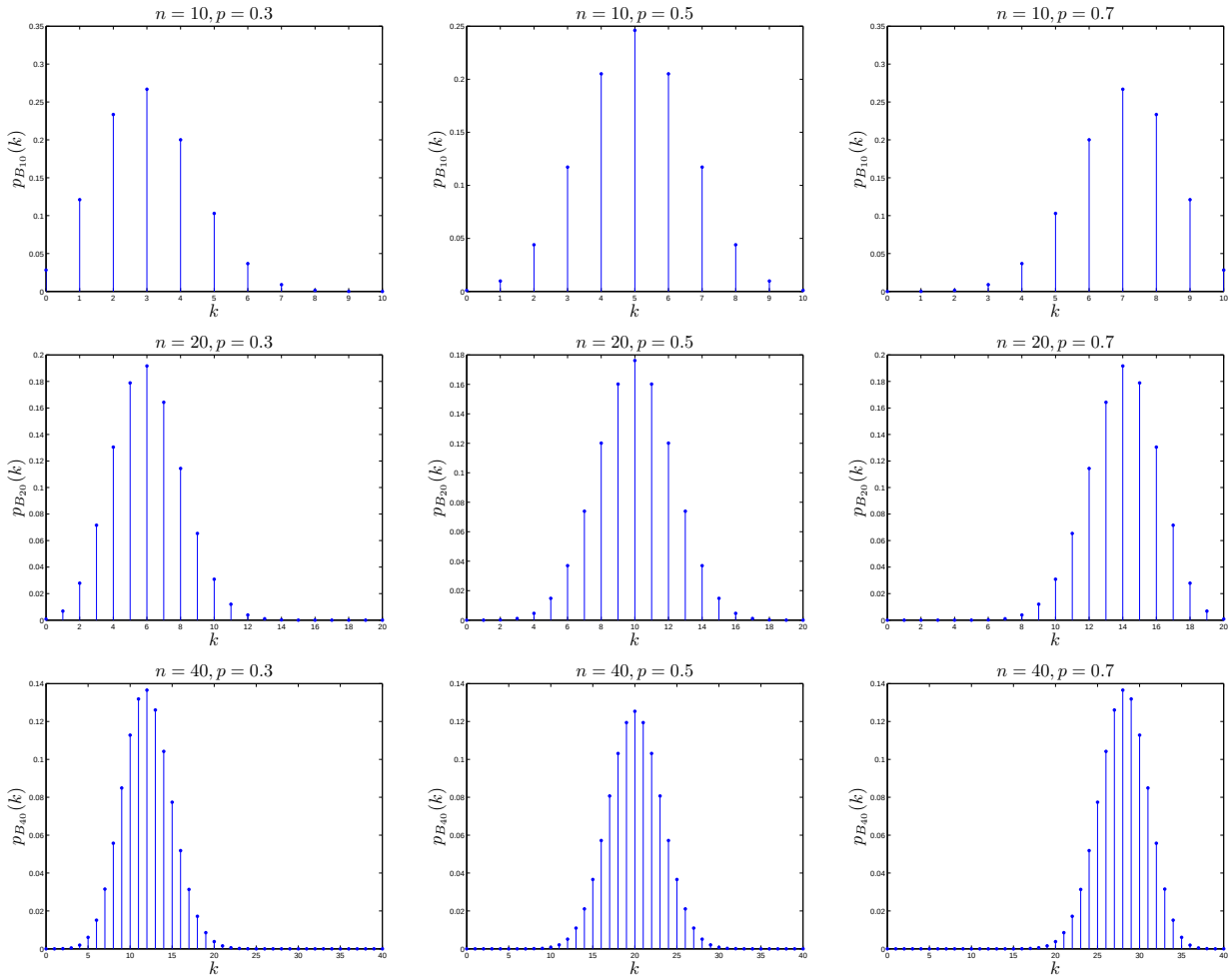


Figure 3: Binomial distributions for different values of the parameters n and p.

Since i is unbounded, there are infinitely many centered moments. It is possible to show that knowledge of the mean and of *all* the moments is equivalent to knowledge of the full distribution, so one can strike a trade-off between conciseness and completeness by providing varying numbers of moments.

The mean m_X of a random variable X is the single number that is closest to all possible outcomes of X in the following sense. Suppose that you are given a fixed guess g for the outcome of, say, the roll of a die. You can only guess once, and then you pay a penalty $(x-g)^2$ every time the outcome is x. This is called a *quadratic penalty*. Then, setting $g = m_X = 3.5$ minimizes your overall loss for a very large number of rolls. This is called the *quadratic loss*. Note that m_X is *not* the most likely outcome. In fact, an outcome of 3.5 when a die is rolled is not just unlikely, it is impossible.

The 2-nd centered moment is called the *variance* of the random variable, and is denoted with σ_X^2:

$$\sigma_X^2 = \mu_X(2) = \sum_{x\in\mathcal{X}} (x - m_X)^2 p(x)$$

and its square root,

$$\sigma_X = \sqrt{\sum_{x\in\mathcal{X}} (x - m_X)^2 p(x)}$$

is called the *standard deviation* of X.

In the guessing scenario above, the overall quadratic loss in n rolls is close to $n\sigma_X^2$ if you make the optimal guess $g = m_X$. This approximation becomes better and better as n tends to infinity. This is rather obvious: since the probability of x is $p(x)$, the number x will occur about $np(x)$ times in n rolls, and result in an overall penalty of $np(x)(x-m_X)^2$ to be imputed to the outcomes of x. Adding up this penalty over the six possible outcomes x yields $n\sigma_X^2$. Thus, σ_X^2 gives a measure of how far overall the mean m_X is from all the outcomes in the aggregate sense above. Because of this, variance and standard deviation are said to measure the *spread* of a distribution. In summary,

The mean minimizes the quadratic loss, and the variance measures the resulting minimal loss.

A normalized version of the third centered moment is sometimes used to measure the degree of asymmetry of a distribution, and is called the *skew* of the random variable, sometimes denoted by $\gamma_X(2)$:

$$\text{skew} \quad : \quad \gamma_X(2) = \frac{\mu_X(3)}{\sigma_X^3} = \frac{\sum_{x\in\mathcal{X}} (x-m_X)^3 p(x)}{\left[\sum_{x\in\mathcal{X}} (x-m_X)^2 p(x)\right]^{\frac{3}{2}}} \,.$$

Skew is positive if there is more mass on the right of the mean, and negative otherwise. Table 1 lists mean, variance, and skew for the uniform, Bernoulli, and binomial random variable. Note the

increasing magnitude of the skew for the latter two variables as p moves away from $1/2$ (either way). The uniform distribution is symmetric, and has therefore zero skew. The moments in this table are fairly straightforward to compute (try this!).

Random variable	Distribution	m_X	σ_X^2	$\gamma_X(2)$
Uniform	$p_{U_n}(k) = \frac{1}{n}$ for $1 \leq k \leq n$	$\frac{n}{2}$	$\frac{n^2-1}{12}$	0
Bernoulli	$p_{B_1}(0) = q$ and $p_{B_1}(1) = p$	p	pq	$\frac{q-p}{\sqrt{pq}}$
Binomial	$p_{b_n}(k) = \binom{n}{k} p^k q^{n-k}$ for $0 \leq k \leq n$	np	npq	$\frac{q-p}{\sqrt{npq}}$

Table 1: Distribution, mean, variance, and skew for three random variables with parameters n and p (as applicable), and with $p + q = 1$.

Note that Table (1) lists only the moments for the single-trial Bernoulli distribution p_{B_1}. While in principle the moments can be computed for the general distribution p_{B_n} wit n repetitions, these moments have little significance. For instance, the distribution for p_{B_3} was given in (10). To compute, say, the mean, one would have to define a scalar random variable that assigns a single number to each of the eight outcomes $(0, 0, 0), \ldots, (1, 1, 1)$. This number is not likely to represent a quantity of interest, so the mean would be meaningless.

It is not a coincidence that the mean and variance of the binomial distribution on n points are n times those for the single-trial Bernoulli variable. The binomial random variable is merely the sum of n independent Bernoulli variables. The mean is a linear operation, so the mean of the sum is in any case the sum of the means. For two variables:

$$\begin{aligned}
m_{X+Y} &= \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} (x+y) p_{X,Y}(x,y) \\
&= \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} x p_{X,Y}(x,y) + \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} y p_{X,Y}(x,y) \\
&= \sum_{x\in\mathcal{X}} x \sum_{y\in\mathcal{Y}} p_{X,Y}(x,y) + \sum_{y\in\mathcal{Y}} y \sum_{y\in\mathcal{Y}} p_{X,Y}(x,y) \\
&= \sum_{x\in\mathcal{X}} x\, p_X(x) + \sum_{y\in\mathcal{Y}} y\, p_Y(y) \\
&= m_X + m_Y \ .
\end{aligned}$$

The equality in the next-to-last line uses the fact that summing a joint distribution of two variables over one of the variables yields the marginal distribution of the other (equations (4) and (5)).

The variance is nonlinear, so additivity usually does not hold. It does hold, however, for independent random variables, that is, when the joint distribution $p_{X_1,\ldots,X_n}(x_1, \ldots, x_n)$ can be factored into the product $p_{X_1}(x_1) \ldots p_{X_n}(x_n)$ of the marginal distributions. The proof of this result for two variables is long but straightforward:

$$
\begin{aligned}
\sigma^2_{X+Y} &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}(x+y-m_{X+Y})^2 p_{X,Y}(x,y) \\
&= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}(x+y-m_X-m_Y)^2 p_{X,Y}(x,y) \\
&= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}((x-m_X)+(y-m_Y))^2 p_{X,Y}(x,y) \\
&= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}(x-m_X)^2 p_X(x)p_Y(y) \;+\; \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}(y-m_Y)^2 p_X(x)p_Y(y) \\
&\quad + \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} 2(x-m_X)(y-m_Y)p_X(x)p_Y(y) \\
&= \sum_{x\in\mathcal{X}}(x-m_X)^2 p_X(x)\sum_{y\in\mathcal{Y}} p_Y(y) \;+\; \sum_{y\in\mathcal{Y}}(y-m_Y)^2 p_Y(y)\sum_{x\in\mathcal{X}} p_X(x) \\
&\quad + 2\sum_{x\in\mathcal{X}}(x-m_X)p_X(x)\sum_{y\in\mathcal{Y}}(y-m_Y)p_Y(y) \\
&= \sum_{x\in\mathcal{X}}(x-m_X)^2 p_X(x) \;+\; \sum_{y\in\mathcal{Y}}(y-m_Y)^2 p_Y(y) \\
&\quad + 2\left(\sum_{x\in\mathcal{X}} x p_X(x) - m_X\sum_{x\in\mathcal{X}} p_X(x)\right)\left(\sum_{y\in\mathcal{Y}} y p_Y(y) - m_Y\sum_{y\in\mathcal{Y}} p_Y(y)\right) \\
&= \sigma^2_X \;+\; \sigma^2_Y \;+\; 2(m_X - m_X\cdot 1)\,(m_Y - m_Y\cdot 1) \\
&= \sigma^2_X \;+\; \sigma^2_Y\,.
\end{aligned}
$$

For more variables, just repeat these arguments (or, as mathematicians would say, use induction). In summary,

For *any* random variables $X_1, \ldots, X_n$,

$$m_{X_1+\ldots+X_n} = m_{X_1} + \ldots + m_{X_n}\,.$$

For *independent* random variables $X_1, \ldots, X_n$,

$$\sigma^2_{X_1+\ldots+X_n} = \sigma^2_{X_1} + \ldots + \sigma^2_{X_n}\,.$$