A Brief Review of Probability, Random Variables, and Some Important Distributions

“A pinch of probability is worth a pound of perhaps.”

James Thurber
(1984-1961)

A.1 Probability

A sample space, often denoted by the Greek capital letter Ω, is a list of all possible outcomes of an event or an experiment. An event is one particular outcome in the sample space. For example, the sample space for one flip of a two-sided coin is the set consisting of “heads” and “tails.” Mathematically, \( Ω = \{\text{heads}, \text{tails}\} \), or more succinctly, \( Ω = \{H, T\} \), where “H” stands for the event of the coin coming up heads. For the roll of one six-sided die, \( Ω \) would be the set of integers between one and six, inclusive, so \( Ω = \{1, 2, 3, 4, 5, 6\} \).

For a continuous measurement that could take on positive values only, say the weight of a person picked at random, then the sample space might be the positive, nonzero part of the real line: \( Ω = (0, \infty) \).

A probability measure on the sample space \( Ω \) is a function \( P \) from subsets of \( Ω \) to the real numbers that satisfies the following axioms. The notation \( P(A) \) stands for “the probability that event \( A \) occurs.”

**Probability Axioms**

1. For each event \( A \in Ω \), \( 0 \leq P(A) \leq 1 \).
2. \( P(Ω) = 1 \).
3. If two events \( A \) and \( B \) are disjoint then \( P(A \cup B) = P(A) + P(B) \).

The first axiom says that probability of each event in the sample space occurring must be between zero and one (inclusive). The second axiom says that the total probability in the sample space is equal to 1. The third axiom says that the probability of one of two events occurring that do not have
anything in common is the sum of the probabilities of the two individual events. More about this axiom shortly.

Under the assumption that every event in $\Omega$ is equally likely to occur, the probability that one of the events occurs is defined as the number of that type of event in $\Omega$ divided by the total number of events in $\Omega$. Using the notation $N(A)$ to mean “the number of events of type $A$ in $\Omega$,” then

$$\mathbb{P}(A) = \frac{N(A)}{\text{total number of outcomes in } \Omega}.$$  

(A.1)

So, for example, for a “fair” coin, the probability of getting heads is one-half, since $N(H) = 1$ and there are a total of two events in $\Omega$. On the other hand, the probability of getting heads on a two-headed coin is 1 (i.e., the coin is guaranteed to come up heads), since $N(H) = 2$. Of course, this is an idealized model of coin flipping since the event “the coin lands on its side” is not in the sample space.

### A.1.1 Some Basic Rules for Calculating Probabilities

#### Union of Disjoint Events

Axiom 3 defines a rule for calculating the probability of the union of two disjoint events. **Union** means together, so the probability of the union of two events $A$ and $B$ means the probability that either $A$ or $B$ occurs. It is written as $A \cup B$. **Disjoint** means that $A$ and $B$ cannot happen at the same time.

![Fig. A.1. A simple Venn diagram representing two disjoint events, A and B.](image)

In the classic Venn diagram in Figure A.1, the box represents the entire sample space, and the circles are the events $A$ and $B$. The box has a total area of 1, corresponding to the probability that something in the sample space must
happen (Axiom 2), and the size of the circles correspond to \( \mathbb{P}(A) \) and \( \mathbb{P}(B) \). Since the events are disjoint, the circles do not overlap and, because they do not overlap, the probability of \( A \cup B \) occurring is simply the probability that \( A \) occurs plus the probability that \( B \) occurs:

\[
\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B).
\]

This is sometimes referred to as the **addition rule**. A single role of a six-sided die provides a simple example to illustrate this rule. To determine the probability of rolling either a one or a two, since the event “roll a 1” is disjoint from the event “roll a 2” (they can’t both happen at the same time):

\[
\mathbb{P}(\text{“roll a 1 or roll a 2”}) = \mathbb{P}(\text{“roll a 1”} \cup \text{“roll a 2”}) = \mathbb{P}(\text{“roll a 1”}) + \mathbb{P}(\text{“roll a 2”}) = 1/6 + 1/6 = 1/3.
\]

Note that this rule applies for any number of disjoint events, not just two. For example, the addition rule can be used to determine the probability of rolling either a one, two, or three on the die. It is

\[
\mathbb{P}(\text{“roll a 1, 2, or 3”}) = \mathbb{P}(\text{“roll a 1 or 2”} \cup \text{“roll a 3”}) = \mathbb{P}(\text{“roll a 1”}) + \mathbb{P}(\text{“roll a 2”}) + \mathbb{P}(\text{“roll a 3”}) = 1/6 + 1/6 + 1/6 = 1/2.
\]

**Complementary Events**

Using the Axioms, additional useful rules for calculating probabilities can be defined. For example, consider an event \( A \) for which it is easy to calculate \( \mathbb{P}(A) \), but perhaps what is of interest is calculating the probability that anything other than \( A \) happens. In other words, the event of interest is “not \( A \),” which is denoted \( A^c \) for the compliment of \( A \). Rather than calculating \( \mathbb{P}(A^c) \) directly from the other events in the sample space, if \( \mathbb{P}(A) \) is easy to calculate, then the rule \( \mathbb{P}(\text{not } A) = \mathbb{P}(A^c) = 1 - \mathbb{P}(A) \) makes it easy to also calculate the compliment of event \( A \).

This rule follows directly from Axioms 1 and 2, and it is visually evident in the Venn diagram of Figure A.2. In the figure, the box represents \( \Omega \), and the circle is the event \( A \). As before, the box has a total probability of 1 and the size of the circle corresponds to \( \mathbb{P}(A) \), so it should be clear that \( \mathbb{P}(A^c) + \mathbb{P}(A) = 1 \), from which it follows that \( \mathbb{P}(A^c) = 1 - \mathbb{P}(A) \).

A simple example illustrates the utility of this rule. Consider the problem of calculating the probability of not throwing a four on one roll of a
fair die. Since \( P(\text{“roll a 4”}) = 1/6 \), it follows that \( P(\text{“not rolling a 4”}) = 1 - P(\text{“roll a 4”}) = 1 - 1/6 = 5/6 \). This is a fairly trivial example, but this rule can be very useful when calculating the probability of “not A” is hard.

For example, imagine you want to determine the chance of getting a four on at least one of three rolls of a fair die. That would be complicated to calculate directly (since there are \( 6^3 = 216 \) events in the sample space). However, recognizing that the compliment to this is the probability of getting no fours on all three rolls of the die simplifies the problem. On one roll, the probability of not getting a four is \( 5/6 \). Using another concept (independent events) to be defined shortly, it turns out that not getting a four on three rolls is just \( 5/6 \times 5/6 \times 5/6 \). So, the probability of getting a four on at least one of three rolls is \( 1 - (5/6)^3 = 0.42 \).

To put this in a biosurveillance context, imagine that during an outbreak the probability of successful detection (PSD) on any particular day using some early event detection algorithm is \( p \). Assuming detection is independent between days, and constant at \( p \), then the probability of failing to detect the outbreak on any one day is \( 1 - p \), and the failure to detect the outbreak for \( k \) days is \( (1 - p)^k \). Thus, the probability of successfully detecting the outbreak over all \( k \) days is \( 1 - (1 - p)^k \).

**Union of Events in General**

There is a more general rule for events that may or may not be disjoint. The left Venn diagram of Figure A.3 illustrates the union of two events \( A \) and \( B \) that are not disjoint. The event “\( A \) and \( B \)” is the intersection of the two events, denoted \( A \cap B \), and is depicted by the dark shaded region in the right
Venn diagram of Figure A.3. Four possible events can now happen: $A$, $B$, “$A$ and $B$,” and “neither $A$ or $B$.” The general rule for calculating $A \cup B$ is:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \quad (A.3)$$

This rule says that, in general, the probability of the union of two events is the sum of their individual probabilities minus the probability of their intersection. The way to think about this is to see that the sum of their individual probabilities counts the intersection area twice, that area must be subtracted to make everything add up correctly.

Note that the previous rule for disjoint events is really a special case of the general rule. When events are disjoint, the probability of their intersection is zero (i.e., $P(A \cap B) = 0$), so the expression in Equation A.3 above simply reduces to the expression in Equation A.2. Also, note that the expression for the union of two events can be generalized to cases with more than two events.

**Intersection of Independent Events**

A very useful rule deals with the probability of the intersection of two independent events. Intuitively, two events are **independent** if knowing the outcome for one of the events provides no information about the outcome of the other event. For example, in two independent flips of a fair coin, knowing that a heads occurred on the first flip provides no information about what will happen on the second flip. Probabilistically, the condition that two events are independent is defined as follows.

**Definition: Independence.** Two events are independent if

$$P(A \cap B) = P(A) \times P(B).$$
Hence, under the condition of independence, the probability that the event “A and B” occurs (that is, the shaded region of the diagram on the right side of Figure A.3) can be calculated as the probability that event A occurs times the probability that event B occurs. Going back to the example of rolling a die three times, if the rolls are independent then the probability of not getting a four on three rolls of the die is simply the probability of not getting a four on the first roll of the die times the probability of not getting a four on the second roll of the die times the probability of not getting a four on the third roll of the die: \( \frac{5}{6} \times \frac{5}{6} \times \frac{5}{6} \).

### A.1.2 Conditional Probability

If two events are dependent, then knowing the outcome of one event does provide information about the outcome of the other event. The notation is \( \mathbb{P}(A \mid B) \), which is read “the probability of A given B,” means the probability that A will occur given that B has occurred. Assuming \( \mathbb{P}(B) \neq 0 \), it is defined as

\[
\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}. \tag{A.4}
\]

The idea behind this definition is that, given the event B occurred, the relevant sample space becomes the events in B rather than \( \Omega \) and the relevant probability measure is now over B. What Equation A.4 is doing is re-normalizing the probabilities from the entire sample space \( \Omega \) to B, since it is now known that the event must be one of those in B.

Note that, when A and B are independent, \( \mathbb{P}(A \mid B) = \mathbb{P}(A) \). Why is this true? If A and B are independent, then \( \mathbb{P}(A \cap B) = \mathbb{P}(A) \times \mathbb{P}(B) \), so substitute this result in Equation (A.4) to get the result. The interpretation, of course, is that if A and B are independent, then knowing what happened with event B provides no additional information about A.

A simple example to illustrate conditional probability is determining the probability of rolling a total of four on two dice. First imagine two fair dice rolled simultaneously so that their outcomes are independent. Then the probability of rolling a total of four is \( \frac{3}{36} \) (there are three ways to get a total of four out of 36 possible two-dice outcomes—write out the sample space if this is not obvious).

But now, consider that it is known that the total of the two rolls is four and the question is to determine the probability that the first die came up a two. Let A be the event that the first die shows a two and let B be the event that the total of the two dice is four. Then the intersection of A and B has probability \( \frac{1}{36} \) (there is only one way to get a total of four with the first die showing a two) and \( \mathbb{P}(B) = \frac{3}{36} \). So, \( \mathbb{P}(A \mid B) = (\frac{1}{36})/(\frac{3}{36}) = \frac{1}{3} \).

To check this result, note that the event B consists of only three events from \( \Omega \): \{1, 3\}, \{3, 1\}, and \{2, 2\}. Once the information is provided that B
has occurred, the relevant probability calculation is then based on only these three possible outcomes. If the question is the probability that the first roll is a two, then there is only one event out of three in $B$ where that occurs, and thus the conditional probability is $1/3$.

**Intersection of Events in General**

Equation A.4 can be turned around to define a more general way to calculate the intersection of two events. Let $A$ and $B$ be events and assume $\mathbb{P}(B) \neq 0$. Then

$$\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B)\mathbb{P}(B).$$

(A.5)

This is sometimes referred to as the *multiplication rule*.

Equation A.5 is often useful in finding the probabilities of intersections, such as the dark area in Figure A.3. For example, imagine a hat that contains the names of six people, four men and two women. Two names will be drawn at random from the hat and the question is to determine the probability that two men’s names are drawn. Let $M_1$ and $M_2$ denote the events of drawing a man’s name on the first and second tries, respectively. From the multiplication rule, it follows that:

$$\mathbb{P}(M_1 \cap M_2) = \mathbb{P}(M_1)\mathbb{P}(M_2 \mid M_1) = \frac{4}{6} \times \frac{3}{5} = \frac{6}{15}.$$

If this is not obvious, write out all the possible unique pairs that can occur and then count how many of them consist of two men.

### A.2 Random Variables

A *random variable* is a variable whose value is subject to variations due to chance (i.e. it’s value is *stochastic*). Compared to other mathematical variables that are typically fixed values (though perhaps unknown), a random variable by definition does not have a single, fixed value (even if unknown). Rather, a random variable can take on a set of possible different values (which may be finite or may be infinite). A random variables is either discrete, where it can take on any of a specified list of specific values, or continuous, where it can take on a value in an interval or a set of intervals.

For discrete random variables, the *probability mass function* (or pmf) maps the possible values of a random variable to their associated probabilities. For example, let $Y$ denote a discrete random variable that can take on one of two possible outcomes, $y_0$ and $y_1$. Then for some $p$, $0 \leq p \leq 1$, the probability mass function is:

$$\mathbb{P}(Y = y_i) = \begin{cases} p, & \text{if } i = 0 \\ 1 - p, & \text{if } i = 1. \end{cases}$$

(A.6)
What Equation A.6 says is that the probability random variable $Y$ takes on the value $y_0$ is $p$, and the probability that $Y$ takes on value $y_1$ is $1 - p$. Since, per probability Axiom 2 the total probability must add up to 1, this also implies that the probability that $Y$ take on any value other than $y_0$ or $y_1$ is zero. In Equation A.6, $Y$ is an example of a Bernoulli random variable. It is often the convention for a Bernoulli random variable to define $y_0 = 1$ and $y_1 = 0$.

Note the convention. Capital Roman letters are typically used to denote random variables while small Roman letters are used to denote the values the random variable can take on. Thus, in Equation A.6 $Y$ is a random variable whose value is not known. All that is known is that it can take on one of two values, $y_0$ or $y_1$.

To make this idea concrete, consider another random variable $X$ that represents the outcome of a fair die. Prior to rolling the die $X$ can take on any one of six values: 1, 2, 3, 4, 5, or 6. Let $x_i$ denote that after rolling the die it came up with value $i$. Then the probability mass function is

$$\mathbb{P}(X = x_i) = \begin{cases} 
1/6, & \text{if } i = 1 \\
1/6, & \text{if } i = 2 \\
1/6, & \text{if } i = 3 \\
1/6, & \text{if } i = 4 \\
1/6, & \text{if } i = 5 \\
1/6, & \text{if } i = 6.
\end{cases}$$

From this, one can write probability statements like $\mathbb{P}(X = 3) = 1/6$ to represent the statement “the probability of rolling a 3 with a fair, six-sided die is one-sixth.” More generally, $\mathbb{P}(X = x)$ is the mathematical statement “the probability that random variable $X$ takes on value $x$.”

Continuous random variables have probability density functions. For a continuous random variable $Y$, a probability density function (pdf) is a non-negative function $f_Y(y)$ defined on the real line having the property that for any set $A$ of real numbers

$$
\mathbb{P}(Y \in A) = \int_A f_Y(y)dy. \tag{A.8}
$$

The notation $f_Y(y)$ means that the probability density function $f$ evaluated at the point $y$ is for the random variable $Y$. When the random variable is understood, the notation can be abbreviated to $f(y)$.

Probability density functions must satisfy the property that

$$
\mathbb{P}\{Y \in (-\infty, \infty)\} = \int_{-\infty}^{\infty} f(y)dy = 1 \tag{A.9}
$$

and the probability that $Y$ falls in some interval between two values $a$ and $b$ is
Both discrete and continuous random variables have cumulative distribution functions. For random variable $Y$, the cumulative distribution function (cdf) is $F_Y(y) = \mathbb{P}(Y \leq y)$. That is, the cdf is a function that, for every value $y$ on the real line (i.e., for every $y \in \mathbb{R}$), it gives the cumulative probability that $Y$ is less than or equal to $y$. If $Y$ is a discrete random variable, then the cdf is

$$F_Y(y) = \sum_{y_i \leq y} \mathbb{P}(Y = y_i),$$

and if $Y$ is a continuous random variable, then the cdf is

$$F_Y(y) = \mathbb{P}(Y \leq y) = \int_{-\infty}^{y} f_Y(y)dy.$$ 

### A.2.1 Expected Value

The expected value of a random variable can be thought of as the average of a very large number (infinite, really) of observations of the random variable. Denote the expected value of a random variable $Y$ as $\mathbb{E}(Y)$.

Calculation of the expected value of $Y$ for discrete random variables is straightforward: it is the sum of the products of each possible outcome times the probability of the outcome. That is, if $Y$ can take on outcomes $y_1, y_2, y_3, \ldots$, then

$$\mathbb{E}(Y) = \sum_{i=1}^{\infty} y_i \mathbb{P}(Y = y_i). \quad (A.11)$$

What is the expected value? Simply stated, it is the average value of a random variable. However, the word “average” here refers to a theoretical quantity that is different from the sample average defined in Chapter 4. While it is useful intuition to think about the expected value as taking the average of a large number of observations, calculating an expected value does not require a sample of data like the calculation for the sample mean calculation does.

Returning to the Bernoulli random variable $Y$ in Equation A.6, the expected value of $Y$ in that case is

$$\mathbb{E}(Y) = \sum_{i=0}^{1} y_i \mathbb{P}(Y = y_i) = y_0 \times p + y_1 \times (1 - p) = y_1 + p(y_0 - y_1).$$

As previously mentioned, with Bernoulli random variables it is often the convention that $y_0 = 1$ and $y_1 = 0$. Using this convention, $\mathbb{E}(Y) = p$.

Similarly, calculating the expected value of $Y$ in the die example of Equation A.7 is straightforward. It is simply the sum of each possible outcome (1, 2, 3, 4, 5, 6).
2, 3, 4, 5, and 6) times their individual probabilities of occurrence which, in the case of a fair die, is always 1/6. That is:

\[
\mathbb{E}(Y) = \sum_{i=1}^{6} i \mathbb{P}(X = i)
\]

\[
= \sum_{i=1}^{6} (i \times \frac{1}{6})
\]

\[
= \frac{1}{6} + \frac{2}{6} + \frac{3}{6} + \frac{4}{6} + \frac{5}{6} + \frac{6}{6} = \frac{21}{6} = 3.5.
\]

The result: \( \mathbb{E}(Y) = 3.5 \), which can be interpreted as, if a fair die was rolled an infinite number of times, then the resulting average of all those rolls would be 3.5. Note that the expected value of \( Y \) does not have to be an integer, or even one of the possible outcomes.

The expected value of \( Y \) for continuous random variables uses calculus to do the calculation analogous to discrete case:

\[
\mathbb{E}(Y) = \int_{-\infty}^{\infty} y f(y) \text{d}y.
\]

The calculations for the continuous case will become clearer in the next section, once specific pdfs are introduced.

### A.2.2 Variance

The variability of a random variable is measured via its variance (and standard deviation). The general definition for the variance in Equation A.12 is a bit more complicated than the one for the expected value–in fact, it uses the expected value within it:

\[
\text{Var}(Y) = \mathbb{E} \left( [Y - \mathbb{E}(Y)]^2 \right).
\]

For a discrete random variable, substituting the definition of the expectation in the outermost expression gives:

\[
\text{Var}(Y) = \sum_{i=1}^{\infty} (y_i - \mathbb{E}(Y))^2 \mathbb{P}(Y = y_i).
\]

Equation A.13 is not as complicated as it may first appear. Starting inside the left set of parentheses, the expression \( y_i - \mathbb{E}(Y) \) is simply the difference between each possible value of the random variable and the random variable's expected value. It’s simply how far each observation is from the expected value of \( Y \). These differences are all then squared to make everything positive, multiplied by the probability that \( y_i \) occurs, and summed over all possible values of \( y_i \).
A.2 Random Variables

So, the variance is just the average squared distance of a random variable from its expected value. The larger the variance, the more one should expect to see observations far from the expected value. The smaller the variance the more such observations are likely to be closer the expected value and thus closer together.

Returning to Chapter 4 and definition of the sample variance, the above description should seem very similar. That is because the sample variance is calculating the variance for a sample of data while this is calculating an equivalent quantity for a random variable.

Equations A.12 and A.13 are mathematically equivalent to

\[
\text{Var}(Y) = \mathbb{E}(Y^2) - (\mathbb{E}(Y))^2, \tag{A.14}
\]

which is sometimes easier to use to calculate the variance. For example, using Equation A.14, the variance of the Bernoulli random variable \( Y \) in Equation A.6 is calculated as follows. First,

\[
\mathbb{E}(Y^2) = \sum_{i=0}^{1} y_i^2 \mathbb{P}(Y = y_i)
= y_0^2 \times p + y_1^2 \times (1 - p) = y_1^2 + p(y_0^2 - y_1^2).
\]

Then,

\[
\text{Var}(Y) = \mathbb{E}(Y^2) - (\mathbb{E}(Y))^2
= y_1^2 + p(y_0^2 - y_1^2) - [y_1 + p(y_0 - y_1)]^2
= y_1^2 + p(y_0^2 - y_1^2) - y_1^2 - 2py_1(y_0 - y_1) - p^2(y_0 - y_1)^2
= p(y_0^2 - y_1^2) - 2py_1(y_0 - y_1) - p^2(y_0 - y_1)^2
= p(y_0 - y_1)^2 - p^2(y_0 - y_1)^2
= p(1 - p)(y_0 - y_1)^2.
\]

Using the convention that \( y_0 = 1 \) and \( y_1 = 0 \), the variance of a Bernoulli random variable is \( \text{Var}(Y) = p(1 - p) \).

To give a numeric example, the variance of \( Y \) in the die example of Equation A.7 is calculated as follows.

\[
\mathbb{E}(Y^2) = \sum_{i=0}^{6} i^2 \mathbb{P}(Y = i)
= 1^2/6 + 2^2/6 + 3^2/6 + 4^2/6 + 5^2/6 + 6^2/6 = 91/6.
\]

Then,

\[
\text{Var}(Y) = \mathbb{E}(Y^2) - (\mathbb{E}(Y))^2
= 91/6 - (21/6)^2
= 546/36 - 441/36 = 105/36 \approx 2.92.
\]
For a continuous random variable, the variance is calculated as

\[
\text{Var}(Y) = \mathbb{E}[(Y - \mathbb{E}(Y))^2] = \mathbb{E}(Y^2) - [\mathbb{E}(Y)]^2 = \int_{-\infty}^{\infty} y^2 f(y)dy - \left(\int_{-\infty}^{\infty} y f(y)dy\right)^2.
\]

As with descriptive statistics, the standard deviation of a random variable is simply the square root of the variance. For example, in the die example, the standard deviation is \(\sqrt{2.92} = 1.71\).

### A.2.3 Covariance and Correlation

The sample covariance was defined in Equation 4.11 in Chapter 4. The equivalent concept for random variables is

\[
\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))].
\]

A mathematically equivalent expression to Equation A.15 is

\[
\text{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).
\]

For any two random variables \(X\) and \(Y\),

\[
\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y),
\]

and

\[
\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y) - 2\text{Cov}(X, Y).
\]

However, if \(X\) and \(Y\) are independent then \(\text{Cov}(X, Y) = 0\).

Covariance is a measure of both the strength and direction of the linear relationship between \(X\) and \(Y\). However, it is often much easier to work with the correlation, \(\rho\), defined as

\[
\rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}},
\]

because \(-1 \leq \rho \leq 1\).

### A.3 Some Important Probability Distributions

#### A.3.1 Discrete Distributions

While there are many important and useful probability distributions for discrete random variables, this section describes three: the binomial, Poisson, and negative binomial.
Binomial

In an experiment with \( n \) independent trials, each of which results in either a “success” with probability \( p \) or a “failure” with probability \( 1 - p \), the number of successes follows a **binomial distribution** with parameters \( n \) and \( p \), often abbreviated \( \text{Bin}(n,p) \). For example, imagine flipping a coin \( n \) times, where the random variable \( Y \) is the number of heads observed out of the \( n \) flips. As long as the probability of getting a heads is constant for each flip, with probability \( p \), the \( Y \) has a binomial distribution. The shorthand notation is \( Y \sim \text{Bin}(n,p) \).

The probability mass function of a binomial random variable with parameters \( n \) and \( p \) is

\[
P(Y = y) = p(y) = \binom{n}{y} p^y (1 - p)^{n-y}, \quad y = 0, 1, 2, \ldots, n \tag{A.18}
\]

where \( \binom{n}{y} = \frac{n!}{(n-y)! m!} \) equals the number of different groups of \( y \) objects that can be chosen from the set of \( n \) objects. For the parameters, \( n \) is a positive, non-zero integer and \( 0 < p < 1 \). For \( n = 1 \) the distribution reduces to the Bernoulli. The expected value of a binomial random variable is \( E(Y) = np \) and the variance is \( \text{Var}(Y) = np(1 - p) \).

Figure A.4 shows a variety of binomial probability mass functions for select values of \( n \) and \( p \).

An example should help illustrate the application of this distribution. Imagine that a fair coin is flipped four times, each flip being independent of the others. What is the probability of obtaining two heads and two tails? The solution is as follows: Let \( Y \) denote the number of heads that occur on four flips of the coin, so that \( Y \) is a binomial random variable with parameters \( n = 4 \) and \( p = 1/2 \). Then by Equation A.18,

\[
P(Y = 2) = \binom{4}{2} \left( \frac{1}{2} \right)^2 \left( \frac{1}{2} \right)^2 = 4 \cdot 3 \cdot 2 \cdot 1 \cdot 2 \cdot 1 \cdot \frac{1}{2}^4 = \frac{3}{8}.
\]

To verify this manually, write out all the possible combinations of heads and tails that can occur on four flips of a coin. There are \( 2 \times 2 \times 2 \times 2 = 16 \) possible ways, and six of them have two heads and two tails. For this example, \( E(Y) = np = 4 \times 1/2 = 2 \), so if this experiment was conducted a large number of times, on average two heads will be observed. The variance is \( \text{Var}(Y) = np(1 - p) = 4 \times 1/2 \times 1/2 = 1 \). So, the variance is 1 and thus the standard deviation is also 1.

Poisson

The **Poisson distribution** is often useful for modeling the probability of a given number of events occurring in a fixed interval of time or space. To apply, the
Fig. A.4. Illustrative binomial probability mass functions for combinations of $n = 10, 20$ and $p = 0.1, 0.5, 0.8$. 
events occur with a constant average rate and independently of the time since
the last event.

The probability mass function for the Poisson distribution, with parameter
\( \lambda \), is

\[
P(Y = y) = \frac{\lambda^y e^{-\lambda}}{y!},
\]

(A.19)

for \( \lambda > 0 \) and \( y = 1, 2, 3, \ldots \), and where \( e \) is the base of the natural logarithm
\( (e = 2.7182\ldots) \). The notation is \( \text{Pois}(\lambda) \) and for \( Y \sim \text{Pois}(\lambda) \)
\( \mathbb{E}(Y) = \mathbb{V}(Y) = \lambda \).

Figure A.5 shows a variety of Poisson probability mass functions for select
values of \( \lambda \).

To illustrate the application of the Poisson distribution, imagine it is rea-
sonable to assume that during the summer when there is no outbreak the
daily ILI syndrome counts \( Y \) follow a Poisson distribution with mean of 5 per
day. That is, \( Y \sim \text{Pois}(5) \). Given this, the probability that a daily account
greater than 10 is observed is

\[
P(Y > 10) = 1 - \sum_{y=0}^{10} \frac{5^y e^{-5}}{y!} = 0.014,
\]

(A.20)

so there is only a 1.4 percent chance of observing an ILI daily count of 11 or
larger.

**Negative Binomial**

With the Poisson distribution it must be that \( \mathbb{E}(Y) = \mathbb{V}(Y) \). However,
biosurveillance data is often *overdispersed*, meaning \( \mathbb{E}(Y) > \mathbb{V}(Y) \). A dis-
tribution for count data that allows for overdispersion is the *negative binomial*.

The negative binomial distribution has two parameters, \( r \) and \( p \), and its
pmf is

\[
P(Y = y) = \binom{y + r - 1}{r - 1} p^r (1 - p)^y,
\]

(A.21)

where \( r > 0, 0 < p < 1 \), and \( y = 0, 1, 2, \ldots \).

The negative binomial is a probability distribution on the number of fail-
ures (\( y \)) until observing \( r \) “successes,” where each “trial” has a Bernoulli dis-
tribution with probability of success \( p \). For \( Y \sim \text{NBin}(r, p) \), \( \mathbb{E}(Y) = r(1-p)/p \)
and \( \mathbb{V}(Y) = r(1-p)/p^2 \).

An alternative parameterization of the binomial pmf, setting \( p = r/(r+l) \)
with \( l = r(1-p)/p \), is
Fig. A.5. Illustrative Poisson probability mass functions for \( \lambda = 0.1, 0.5, 1, 5, 10, 15 \). (Note the change in \( y \)-axis scale between the left and right columns of pmfs.)
\[ \mathbb{P}(Y = y) = \binom{y + r - 1}{r - 1} \left( \frac{r}{r + l} \right)^r \left( \frac{1 - \frac{r}{r + l}}{r + l} \right)^y \]  
\[ = \frac{\Gamma(y + r)}{y! \Gamma(r)} \left( \frac{r}{r + l} \right)^r \left( 1 - \frac{r}{r + l} \right)^y \]  
\[ = \frac{I^y}{y!} \left( \frac{r}{r + l} \right)^r \frac{\Gamma(y + r)}{\Gamma(r)(r + l)^y}, \]  
where the second line follows because \((y - 1)! = \Gamma(y)\) and the third line is just an algebraic rearrangement of the terms from the second line.

With this parameterization, as \(r \to \infty\), \(\left( \frac{r}{r + l} \right)^r \to e^{-l}\). This follows from the fact that \(\lim_{n \to \infty} \left( 1 + \frac{x}{n} \right)^n = e^x\). Further, if \(r \to \infty\) and \(p \to 1\) in such a way that \(r(1 - p)\) converges to a constant, then \(\frac{\Gamma(y + r)}{\Gamma(r)(r + l)^y} \to 1\). Under these conditions the limiting distribution of the negative binomial is the Poisson:

\[ \lim f(y) = \frac{ly e^{-l}}{y!}. \]

Figure A.6 shows a variety of negative binomial probability mass functions, using the definition in Equation A.21, for select values of \(r\) and \(p\).

### A.3.2 Continuous Distributions

As with discrete distributions, there are many important and useful probability distributions for continuous random variables. This section describes four: the uniform, normal, \(t\), and \(F\).

#### Uniform

The uniform distribution is a good place to start to illustrate the concept of continuous distributions and demonstrate their associated calculations. The probability density function for a random variable \(Y\) with a uniform distribution on the interval \([a, b]\), is

\[ f_Y(y) = \begin{cases} \frac{1}{b-a}, & \text{for } y \in [a, b] \text{ with } b > a \\ 0, & \text{otherwise} \end{cases} \]  
\[ (A.23) \]

The uniform probability density function is the continuous analog of the fair die pmf of Section A.1 (which was a discrete uniform distribution on the integers 1-6). The interpretation of this density is that every set of equal size that lies between \(a\) and \(b\) is equally likely, while nothing outside of the interval \([a, b]\) can occur. A graph of the density function is shown in Figure A.7.

To verify that Equation A.23 meets the definition of a probability density function per Equation A.9, note that
Fig. A.6. Illustrative negative binomial probability mass functions, using the pmf as defined in Equation A.21, for combinations of $r = 2, 5$ and $p = 0.3, 0.5, 0.7$. 
The probability density function for a uniform distribution is defined in Equation A.23.

$$f(y) = \frac{1}{b-a}$$

for $a \leq y \leq b$. The expected value of a uniform random variable is

$$E(y) = \frac{1}{b-a} \int_a^b y \, dy = \left. \frac{y^2}{2} \right|_a^b = \frac{b^2 - a^2}{2(b-a)} = \frac{b + a}{2}.$$

Then, by calculating $E(Y^2)$ as

$$E(Y^2) = \int_{-\infty}^{\infty} y^2 f_Y(y) \, dy = \frac{1}{b-a} \int_a^b y^2 \, dy = \left. \frac{y^3}{3} \right|_a^b = \frac{b^3 - a^3}{3(b-a)}$$

and so the variance is

$$\text{Var}(Y) = E(Y^2) - (E(Y))^2 = \frac{b^3 - a^3}{3(b-a)} - \left( \frac{b + a}{2} \right)^2.$$

Thus, for example, if $Y$ has a continuous uniform distribution on the interval $[0, 1]$, denoted as $Y \sim U(0, 1)$, then $E(Y) = \frac{1}{2}$ and $\text{Var}(Y) = \frac{1}{12}$.

**Normal**

The *normal distribution* is an important distribution in statistics; perhaps the most important distribution. Many natural physical phenomena and statistics
follow a normal distribution. This happens for a very good reason described more fully in Section A.3.3: the Central Limit Theorem (CLT).

The univariate normal distribution is described by two parameters: \( \mu \) (Greek letter “mu”) and \( \sigma \) (Greek letter “sigma”). For \( Y \sim N(\mu, \sigma^2) \), the probability density function is

\[
f_Y(y) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left( \frac{-(y - \mu)^2}{2\sigma^2} \right).
\]  

(A.24)

For the parameters of the normal distribution, \(-\infty < \mu < \infty \) and \( \sigma^2 > 0 \). For \( Y \sim N(\mu, \sigma^2) \), \( \mathbb{E}(Y) = \mu \) and \( \text{Var}(Y) = \sigma^2 \).

It is conventional to use the letter \( Z \) to represent a random variable from a “standard normal” distribution. A standard normal distribution has \( \mu = 0 \) and \( \sigma^2 = 1 \). The symbol \( \Phi(z) \) represents the cumulative probability that \( Z \) is less than or equal to some number \( z \): \( \Phi(z) = \mathbb{P}(Z \leq z) \).

The probability density function for the standard normal is plotted in Figure A.8 and Figure A.9 compares the density functions for the standard normal with a normal distribution \( \mu = 0 \) and \( \sigma^2 = 2 \), a \( N(0, 2) \).

Note that the normal distribution is symmetric about \( \mu \), meaning if its pdf is “folded it in half” along the vertical line at \( \mu \) then the two halves would line up exactly. This is evident in Figure A.8 where the left and right halves of the curve on either side of \( z = 0 \) look exactly alike. The normal distribution is also the canonical “bell-shaped curve.”

![Fig. A.8. The probability density function for a standard normal distribution.](image-url)
A.3 Some Important Probability Distributions

Fig. A.9. Comparison of the standard normal distribution density function, \(N(0,1)\), versus a \(N(0,2)\) pdf.

particular reference to the \(t\)-distribution described in the next section. Note that the name “normal” does not imply that random variables that follow another distribution are abnormal—they’re simply not normally distributed.

**Standardizing**

An observation from any normal distribution, \(N(\mu, \sigma^2)\), is standardized by subtracting off the mean and dividing by the standard deviation. That is, if \(Y\) comes from a \(N(\mu, \sigma^2)\) distribution, then \(Z = (Y - \mu) / \sigma\) has a \(N(0, 1)\) distribution. That is,

\[
\Pr(a < Y < b) = \Pr\left(\frac{a - \mu}{\sigma} < Z < \frac{b - \mu}{\sigma}\right) = \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right)
\]

This is very useful, as it allows us to relate any normal distribution to the standard normal. Table A.1 provides a detailed summary for the \(\Pr(-z \leq Z \leq z)\), when \(Z\) is a random variable with a standard normal distribution.

**Bivariate and Multivariate Normal Distributions**

The bivariate normal distribution is the *joint distribution* of two variables, \(Y_1\) and \(Y_2\), each of which individually is normally distributed. The bivariate
normal distribution is described by five parameters: the means of each variable, \( \mu_1 \) and \( \mu_2 \), the variances of each variable, \( \sigma_1^2 \) and \( \sigma_2^2 \), and the correlation between the two variables. The correlation is denoted as \( \rho \), \(-1 \leq \rho \leq 1\). The probability density function is

\[
f_{Y_1,Y_2}(y_1,y_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left(\frac{-z}{2(1-\rho^2)}\right),
\]

where

\[
z = \frac{(y_1 - \mu_1)^2}{\sigma_1^2} - \frac{2\rho(y_1 - \mu_1)(y_2 - \mu_2)}{\sigma_1\sigma_2} + \frac{(y_2 - \mu_2)^2}{\sigma_2^2}.
\]

If \( \rho = 0 \) then \( Y_1 \) and \( Y_2 \) are independent. If, in addition, \( \mu_1 = \mu_2 \) and \( \sigma_1 = \sigma_2 \) then \( Y_1 \) and \( Y_2 \) are iid.

Now, let \( \mathbf{Y} \) denote a vector of \( k \) observations, \( \mathbf{Y} = \{Y_1, Y_2, \ldots, Y_k\} \). \( \mathbf{Y} \sim \mathcal{N}(\mathbf{\mu}, \mathbf{\Sigma}) \) denotes that \( \mathbf{Y} \) has a multivariate normal distribution with mean vector \( \mathbf{\mu} \) and covariance matrix \( \mathbf{\Sigma} \). The multivariate normal pdf is

\[
f_{\mathbf{Y}}(y_1, \ldots, y_k) = \frac{1}{(2\pi)^{k/2} |\mathbf{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{y} - \mathbf{\mu})' \mathbf{\Sigma}^{-1} (\mathbf{y} - \mathbf{\mu})\right),
\]

where \( |\mathbf{\Sigma}| \) denotes the determinant of \( \mathbf{\Sigma} \).

For \( k = 2 \), Equation A.26 is equivalent to Equation A.25 with \( \mathbf{\mu} = \{\mu_1, \mu_2\} \) and

\[
\mathbf{\Sigma} = \begin{pmatrix}
\sigma_1^2 & \rho \sigma_1 \sigma_2 \\
\rho \sigma_2 \sigma_1 & \sigma_2^2
\end{pmatrix}.
\]
For $k=1$ Equation A.26 reduces to the univariate normal distribution with the pdf shown in Equation A.24 where $\Sigma$ is a $1 \times 1$ matrix (i.e., it’s a real number).

### t-Distribution

When standardizing a normally distributed observation, it is often the case that the population standard deviation must be estimated from data. Under such circumstances, when the sample variance is used in the standardization calculation in place of the population variance, the standardized value then follows a \textit{t-distribution} with $n - 1$ “degrees of freedom.”

The density function for the \textit{t}-distribution with $n$ degrees of freedom is

$$f(y) = \frac{\Gamma[(n+1)/2]}{\sqrt{n\pi} \Gamma(n/2)} \left(1 + \frac{y^2}{n}\right)^{-(n+1)/2}, \quad (A.27)$$

where $n > 0$.

The \textit{t}-distribution is symmetric and bell-shaped, like the normal distribution, but has “heavier tails,” meaning that it is more likely to produce values that fall farther from its mean (compared to the normal distribution). When standardizing, the heavier tails account for the extra uncertainty that is introduced into the standardized value because of the use of the sample variance.

Figure A.10 shows four different \textit{t}-distributions, corresponding to four different degrees of freedom, compared to a standard normal distribution. When looking at the tails of the curves, the curve with the largest tails (corresponding to “t(2)”) is a \textit{t}-distribution with 2 degrees of freedom; the curve with the next largest tails (corresponding to “t(3)”) is a \textit{t}-distribution with 3 degrees of freedom, followed by a \textit{t}-distribution with 5 degrees of freedom; and, the curve with the smallest tails is a \textit{t}-distribution with 10 degrees of freedom.

What this graph shows is that the fewer degrees of freedom, the heavier the tails. Conversely, the more degrees of freedom, the closer the \textit{t}-distribution gets to the normal. With an infinite number of degrees of freedom the \textit{t}-distribution is the normal distribution.

Table A.2 gives selected quantiles from the \textit{t}-distribution for various degrees of freedom. Most statistics textbooks provide more detailed tabulations, and various statistical software packages can also be used to look up the quantiles for any number of degrees of freedom and/or tail probabilities.

### Chi-square ($\chi^2$) Distribution

The $\chi^2$ ("chi-square") \textit{distribution} is another important statistical distribution. The sum of the squares of $n$ standard normal random variables follow a
χ^2 distribution with n “degrees of freedom.” That is, for \( Y = Z_1^2 + Z_2^2 + \cdots + Z_n^2 \), 
\( Y \sim \chi_n^2 \) with \( \mathbb{E}(Y) = n \) and \( \text{Var}(Y) = 2n \). Hence, the \( \chi_n^2 \) distribution is characterized by \( n \), the number of degrees of freedom.

The density function for \( Y \sim \chi_n^2 \) is

\[
f_Y(y) = \begin{cases} 
\frac{2^{-n/2} \Gamma((n/2))}{\Gamma(n/2)} y^{n/2 - 1} e^{-y/2}, & y > 0 \\
0, & \text{elsewhere}
\end{cases}
\]  
(A.28)

where \( n \) is a positive integer and \( \Gamma \) is the gamma function: \( \Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y} dy \).

As shown in Figure A.11, when the number of degrees of freedom is small the \( \chi^2 \) distribution is asymmetric. As the number of degrees of freedom increases, the resulting distribution becomes more symmetric (as a direct result of the Central Limit Theorem discussed in Section A.3.3).

Fig. A.10. Examples of the \( t \)-distribution for 2, 3, 5, and 10 degrees of freedom compared to the standard normal distribution.
A.3 Some Important Probability Distributions

### Table A.2

<table>
<thead>
<tr>
<th>df</th>
<th>$t_{0.10}$</th>
<th>$t_{0.05}$</th>
<th>$t_{0.025}$</th>
<th>$t_{0.01}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.078</td>
<td>6.314</td>
<td>12.706</td>
<td>31.821</td>
</tr>
<tr>
<td>2</td>
<td>1.886</td>
<td>2.920</td>
<td>4.303</td>
<td>6.965</td>
</tr>
<tr>
<td>3</td>
<td>1.638</td>
<td>2.353</td>
<td>3.182</td>
<td>4.541</td>
</tr>
<tr>
<td>4</td>
<td>1.533</td>
<td>2.132</td>
<td>2.776</td>
<td>3.747</td>
</tr>
<tr>
<td>5</td>
<td>1.476</td>
<td>2.015</td>
<td>2.571</td>
<td>3.365</td>
</tr>
<tr>
<td>6</td>
<td>1.440</td>
<td>1.943</td>
<td>2.447</td>
<td>3.143</td>
</tr>
<tr>
<td>7</td>
<td>1.415</td>
<td>1.895</td>
<td>2.356</td>
<td>2.998</td>
</tr>
<tr>
<td>8</td>
<td>1.397</td>
<td>1.860</td>
<td>2.306</td>
<td>2.896</td>
</tr>
<tr>
<td>9</td>
<td>1.383</td>
<td>1.833</td>
<td>2.262</td>
<td>2.821</td>
</tr>
<tr>
<td>10</td>
<td>1.372</td>
<td>1.812</td>
<td>2.228</td>
<td>2.764</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>df</th>
<th>$t_{0.10}$</th>
<th>$t_{0.05}$</th>
<th>$t_{0.025}$</th>
<th>$t_{0.01}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>1.341</td>
<td>1.753</td>
<td>2.131</td>
<td>2.602</td>
</tr>
<tr>
<td>20</td>
<td>1.325</td>
<td>1.725</td>
<td>2.086</td>
<td>2.528</td>
</tr>
<tr>
<td>25</td>
<td>1.316</td>
<td>1.708</td>
<td>2.060</td>
<td>2.485</td>
</tr>
<tr>
<td>30</td>
<td>1.310</td>
<td>1.697</td>
<td>2.042</td>
<td>2.457</td>
</tr>
<tr>
<td>∞</td>
<td>1.282</td>
<td>1.645</td>
<td>1.960</td>
<td>2.326</td>
</tr>
</tbody>
</table>

Table A.2. Quantiles from the $t$-distribution for various degrees of freedom, where the quantiles displayed are for the probability that a random variable $T$ with a $t$-distribution is greater than $t_\alpha$: $P(T > t_\alpha) = \alpha$.

**Fig. A.11.** Some chi-square probability density functions for various degrees of freedom.

#### $F$ Distribution

The $F$ distribution is related to the $\chi^2$ distribution. Let $W$ and $X$ be independent $\chi^2$ random variables with $m$ and $n$ degrees of freedom respectively. Then the random variable

$$Y = \frac{W/m}{X/n}$$

has an $F$ distribution with $m$ and $n$ degrees of freedom. Note that the order of $m$ and $n$ is important. The density function for the $F$ distribution is
\[ f_Y(y) = \frac{\Gamma[(m+n)/2]}{\Gamma(m/2)\Gamma(n/2)} \left( \frac{m}{n} \right) y^{m/2-1} \left( 1 + \frac{m}{n} y \right)^{-(m+n)/2}. \]

\[ (A.29) \]

**Fig. A.12.** Some F distribution probability density functions for various degrees of freedom.

As shown in Figure A.12, a wide variety of probability density functions are possible with the F distribution depending on how the degrees of freedom (m and n) are varied.
A function of a set of random variables is itself a random variable with its own distribution referred to as the sampling distribution. For example, the sum of \(n\) iid normally distributed random variables, \(X_i \sim N(\mu, \sigma^2), i = 1, \ldots, n,\) is also normally distributed with mean 0 and variance \(n \times \sigma^2\). That is, if \(Y = X_1 + \cdots + X_n,\) then \(Y \sim N(\mu, \sigma^2).\) Similarly, for \(\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} X_i\) then \(\bar{Y} \sim N(\mu, \sigma^2/n).\)

The term standard error is another name for the standard deviation of a statistic. Thus, in the previous example, the standard error of \(\bar{Y}\) is \(\sigma/\sqrt{n}.\) Note that the standard error of the mean is smaller than the standard deviation of the individual observations, so the sampling distribution of the mean is narrower than the distribution of the individual observations. One consequence of this result is that as the sample size increases it becomes more likely that \(\bar{Y}\) will be close to \(\mu.\)

Note the use of capital letters in the foregoing are purposeful. Capital letters represent random variables while lower case letters represent numbers. Hence, \(\bar{Y}\) denotes the average of a set of random variables. Because the random variables are random, so is \(\bar{Y},\) and thus \(\bar{Y}\) has a probability density function, which is referred to as the sampling distribution of \(\bar{Y}.\) In contrast, \(\bar{y}\) is a sample average, which is a number, and hence cannot have a distribution.

It turns out that the sums and averages of normally distributed random variables are themselves distributed normally. Thus, for example, the sampling distribution of the mean of normally distributed random variables with mean \(\mu\) and variance \(\sigma^2\) is itself normally distributed with mean \(\mu\) and variance \(\sigma^2/n\) (and thus standard error \(\sigma/\sqrt{n}\)). Furthermore, even if the distribution of individual observations is not normally distributed, the sampling distribution will be approximately normal via the Central Limit Theorem.

The Central Limit Theorem (CLT) says that the distribution of sums of random variables tends towards a normal distribution as the sum gets large even if the random variables are not themselves normally distributed. Formally:

**Theorem A.1 (Central Limit Theorem).** Let \(X_1, X_2, \ldots, X_i, \ldots\) be a sequence of independent random variables having mean 0, variance \(\sigma^2\) and a common distribution function \(F.\) Then,

\[
\lim_{n \to \infty} P \left( \frac{X_n}{\sigma/\sqrt{n}} \leq x \right) = \Phi(x), -\infty < x < \infty.
\]

Furthermore, since the mean is just the sum of random variables renormalized by dividing by the number of random variables being summed, the mean also has an approximate normal distribution.
The Central Limit Theorem applies both to discrete and continuous random variables. Figure A.13 illustrates the CLT for data with a uniform distribution. In the figure, the upper left plot is a histogram resulting from 10,000 draws from a \( U(0, 1) \). The plot to the right is 10,000 sums of two draws from the same distribution. The bottom left plots 10,000 sums of five \( U(0, 1) \) random variables, and the bottom right is 10,000 sums of 30. Note how the sample average tends toward the normal distribution as the sample size gets large.

**Fig. A.13.** An illustration of the Central Limit Theorem applied to sums of uniformly distributed random variables. The upper left plot is a histogram resulting from 10,000 draws from a \( U(0, 1) \). The plot to the right is 10,000 sums of two draws from the same distribution. The bottom left plots 10,000 sums of five \( U(0, 1) \) random variables, and the bottom right is 10,000 sums of 30.