# Chapter 3

# Review of Probability

*A review of the portions of probability useful for understanding experimental design and analysis.*

The material in this section is intended as a review of the topic of probability as covered in the prerequisite course (36-201 at CMU). The material in gray boxes is beyond what you may have previously learned, but may help the more mathematically minded reader to get a deeper understanding of the topic. You need not memorize any formulas or even have a firm understanding of this material at the start of the class. But I do recommend that you at least skim through the material early in the semester. Later, you can use this chapter to review concepts that arise as the class progresses.

For the earliest course material, you should have a basic idea of what a random variable and a probability distribution are, and how a probability distribution defines event probabilities. You also need to have an understanding of the concepts of parameter, population, mean, variance, standard deviation, and correlation.

## 3.1 Definition(s) of probability

We could choose one of several technical definitions for **probability**, but for our purposes it refers to an assessment of the likelihood of the various possible outcomes in an experiment or some other situation with a "random" outcome.

Note that in probability theory the term "outcome" is used in a more general

sense than the outcome vs. explanatory variable terminology that is used in the rest of this book. In probability theory the term "outcome" applies not only to the "outcome variables" of experiments but also to "explanatory variables" if their values are not fixed. For example, the dose of a drug is normally fixed by the experimenter, so it is not an outcome in probability theory, but the age of a randomly chosen subject, even if it serves as an explanatory variable in an experiment, is not "fixed" by the experimenter, and thus can be an "outcome" under probability theory.

The collection of all possible outcomes of a particular random experiment (or other well defined random situation) is called the **sample space**, usually abbreviated as **S** or $\Omega$ (omega). The outcomes in this set (list) must be exhaustive (cover all possible outcomes) and mutually exclusive (non-overlapping), and should be as simple as possible.

For a simple example consider an experiment consisting of the tossing of a six sided die. One possible outcome is that the die lands with the side with one dot facing up. I will abbreviate this outcome as 1du (one dot up), and use similar abbreviations for the other five possible outcomes (assuming it can't land on an edge or corner). Now the sample space is the set {1du, 2du, 3du, 4du, 5du, 6du}. We use the term **event** to represent any subset of the sample space. For example {1du}, {1du, 5du}, and {1du, 3du, 5du}, are three possible events, and most people would call the third event "odd side up". One way to think about events is that they can be defined before the experiment is carried out, and they either occur or do not occur when the experiment is carried out. In probability theory we learn to compute the chance that events like "odd side up" will occur based on assumptions about things like the probabilities of the elementary outcomes in the sample space.

Note that the "true" outcome of most experiments is not a number, but a physical situation, e.g., "3 dots up" or "the subject chose the blue toy". For convenience sake, we often "map" the physical outcomes of an experiment to integers or real numbers, e.g., instead of referring to the outcomes 1du to 6du, we can refer to the numbers 1 to 6. Technically, this mapping is called a **random variable**, but more commonly and informally we refer to the unknown numeric outcome itself (before the experiment is run) as a "random variable". Random variables commonly are represented as upper case English letters towards the end of the alphabet, such as Z, Y or X. Sometimes the lower case equivalents are used to represent the actual outcomes after the experiment is run.

Random variables are maps from the sample space to the real numbers, but they need not be one-to-one maps. For example, in the die experiment we could map all of the outcomes in the set {1du, 3du, 5du} to the number 0 and all of the outcomes in the set {2du, 4du, 6du} to the number 1, and call this random variable Y. If we call the random variable that maps to 1 through 6 as X, then random variable Y could also be thought of as a map from X to Y where the odd numbers of X map to 0 in Y and the even numbers to 1. Often the term **transformation** is used when we create a new random variable out of an old one in this way. It should now be obvious that many, many different random variables can be defined/invented for a given experiment.

A few more basic definitions are worth learning at this point. A random variable that takes on only the numbers 0 and 1 is commonly referred to as an **indicator (random) variable**. It is usually named to match the set that corresponds to the number 1. So in the previous example, random variable Y is an indicator for even outcomes. For any random variable, the term **support** is used to refer to the set of possible real numbers defined by the mapping from the physical experimental outcomes to the numbers. Therefore, for random variables we use the term "event" to represent any subset of the support.

Ignoring certain technical issues, probability theory is used to take a basic set of assigned (or assumed) probabilities and use those probabilities (possibly with additional assumptions about something called independence) to compute the probabilities of various more complex events.

**The core of probability theory is making predictions about the chances of occurrence of events based on a set of assumptions about the underlying probability processes.**

One way to think about probability is that it quantifies how much we can know when we cannot know something exactly. Probability theory is deductive, in the sense that it involves making assumptions about a random (not completely predictable) process, and then deriving valid statements about what is likely to happen based on mathematical principles. For this course a fairly small number of probability definitions, concepts, and skills will suffice.

For those students who are unsatisfied with the loose definition of probability above, here is a brief descriptions of three different approaches to probability, although it is not necessary to understand this material to continue through the chapter. If you want even more detail, I recommend *Comparative Statistical Inference* by Vic Barnett.

Valid probability statements do not claim what events will happen, but rather which are likely to happen. The starting point is sometimes a judgment that certain events are a priori equally likely. Then using only the additional assumption that the occurrence of one event has no bearing on the occurrence of another separate event (called the assumption of independence), the likelihood of various complex combinations of events can be worked out through logic and mathematics. This approach has logical consistency, but cannot be applied to situations where it is unreasonable to assume equally likely outcomes and independence.

A second approach to probability is to define the probability of an outcome as the limit of the long-term fraction of times that outcome occurs in an ever-larger number of independent trials. This allows us to work with basic events that are not equally likely, but has a disadvantage that probabilities are assigned through observation. Nevertheless this approach is sufficient for our purposes, which are mostly to figure out what would happen if certain probabilities are assigned to some events.

A third approach is subjective probability, where the probabilities of various events are our subjective (but consistent) assignments of probability. This has the advantage that events that only occur once, such as the next presidential election, can be studied probabilistically. Despite the seemingly bizarre premise, this is a valid and useful approach which may give different answers for different people who have different beliefs, but still helps calculate your rational but personal probability of future uncertain events, given your prior beliefs.

Regardless of which definition of probability you use, the calculations we need are basically the same. First we need to note that probability applies to some well-defined unknown or future situation in which some outcome will occur, the list of possible outcomes is well defined, and the exact outcome is unknown. If the

outcome is categorical or discrete quantitative (see section 2.3), then each possible outcome gets a probability in the form of a number between 0 and 1 such that the sum of all of the probabilities is 1. This indicates that impossible outcomes are assigned probability zero, but assigning a probability zero to an event does not necessarily mean that that outcome is impossible (see below). (Note that a probability is technically written as a number from 0 to 1, but is often converted to a percent from 0% to 100%. In case you have forgotten, to convert to a percent multiply by 100, e.g., 0.25 is 25% and 0.5 is 50% and 0.975 is 97.5%.)

**Every valid probability must be a number between 0 and 1 (or a percent between 0% and 100%).**

We will need to distinguish two types of random variables. Discrete random variables correspond to the categorical variables plus the discrete quantitative variables of chapter 2. Their support is a (finite or infinite) list of numeric outcomes, each of which has a non-zero probability. (Here we will loosely use the term "support" not only for the numeric outcomes of the random variable mapping, but also for the sample space when we do not explicitly map an outcome to a number.) Examples of discrete random variables include the result of a coin toss (the support using curly brace set notation is {H,T}), the number of tosses out of 5 that are heads ($\{0, 1, 2, 3, 4, 5\}$), the color of a random person's eyes ({blue, brown, green, other}), and the number of coin tosses until a head is obtained ($\{1, 2, 3, 4, 5, \ldots\}$). Note that the last example has an infinite sized support.

Continuous random variables correspond to the continuous quantitative variables of chapter 2. Their support is a continuous range of real numbers (or rarely several disconnected ranges) with no gaps. When working with continuous random variables in probability theory we think as if there is no rounding, and each value has an infinite number of decimal places. In practice we can only measure things to a certain number of decimal places, actual measurement of the continuous variable "length" might be 3.14, 3.15, etc., which does have gaps. But we approximate this with a continuous random variable rather than a discrete random variable because more precise measurement is possible in theory.

A strange aspect of working with continuous random variables is that each particular outcome in the support has probability zero, while none is actually impossible. The reason each outcome value has probability zero is that otherwise

the probabilities of all of the events would add up to more than 1. So for continuous random variables we usually work with intervals of outcomes to say, e.g, that the probability that an outcome is between 3.14 and 3.15 might be 0.02 while each real number in that range, e.g., $\pi$ (exactly), has zero probability. Examples of continuous random variables include ages, times, weights, lengths, etc. All of these can theoretically be measured to an infinite number of decimal places.

It is also possible for a random variable to be a mixture of discrete and continuous random variables, e.g., if an experiment is to flip a coin and report 0 if it is heads and the time it was in the air if it is tails, then this variable is a mixture of the discrete and continuous types because the outcome "0" has a non-zero (positive) probability, while all positive numbers have a zero probability (though intervals between two positive numbers would have probability greater than zero.)

## 3.2 Probability mass functions and density functions

.

A **probability mass function** (pmf) is just a full description of the possible outcomes and their probabilities for some discrete random variable. In some situations it is written in simple list form, e.g.,

$$f(x) = \begin{cases} 0.25 & \text{if} \quad x = 1 \\ 0.35 & \text{if} \quad x = 2 \\ 0.40 & \text{if} \quad x = 3 \end{cases}$$

where f(x) is the probability that random variable X takes on value x, with f(x)=0 implied for all other x values. We can see that this is a valid probability distribution because each probability is between 0 and 1 and the sum of all of the probabilities is 1.00. In other cases we can use a formula for f(x), e.g.

$$f(x) = \left( \frac{4!}{(4-x)!\ x!} \right) p^x (1-p)^{4-x} \text{ for } x = 0, 1, 2, 3, 4$$

which is the so-called binomial distribution with parameters 4 and p.

It is not necessary to understand the mathematics of this formula for this course, but if you want to try you will need to know that the exclamation mark symbol is pronounced "factorial" and $r!$ represents the product of all the integers from 1 to $r$. As an exception, $0! = 1$.

This particular pmf represents the probability distribution for getting $x$ "successes" out of 4 "trials" when each trial has a success probability of $p$ independently. This formula is a shortcut for the five different possible outcome values. If you prefer you can calculate out the five different probabilities and use the first form for the pmf. Another example is the so-called geometric distribution, which represents the outcome for an experiment in which we count the number of independent trials until the first success is seen. The pmf is:

$$f(x) = p(1-p)^{x-1} \text{ for } x = 1, 2, 3, \ldots$$

and it can be shown that this is a valid distribution with the sum of this infinitely long series equal to 1.00 for any value of p between 0 and 1. This pmf cannot be written in the list form. (Again the mathematical details are optional.)

By definition a random variable takes on numeric values (i.e., it maps real experimental outcomes to numbers). Therefore it is easy and natural to think about the pmf of any discrete continuous experimental variable, whether it is explanatory or outcome. For categorical experimental variables, we do not need to assign numbers to the categories, but we always can do that, and then it is easy to consider that variable as a random variable with a finite pmf. Of course, for nominal categorical variables the order of the assigned numbers is meaningless, and for ordinal categorical variables it is most convenient to use consecutive integers for the assigned numeric values.

**Probability mass functions apply to discrete outcomes. A pmf is just a list of all possible outcomes for a given experiment and the probabilities for each outcome.**

For continuous random variables, we use a somewhat different method for summarizing all of the information in a probability distribution. This is the **probability density function** (pdf), usually represented as "f(x)", which does not represent probabilities directly but from which the probability that the outcome falls in a certain range can be calculated using integration from calculus. (If you don't remember integration from calculus, don't worry, it is OK to skip over the details.)

One of the simplest pdf's is that of the uniform distribution, where all real numbers between a and b are equally likely and numbers less than $a$ or greater than $b$ are impossible. The pdf is:

$$f(x) = 1/(b-a) \text{ for } a \leq x \leq b$$

The general probability formula for any continuous random variable is

$$\Pr(t \leq X \leq u) = \int_t^u f(x)dx.$$

In this formula $\int \cdot \, dx$ means that we must use calculus to carry out integration.

Note that we use capital X for the random variable in the probability statement because this refers to the potential outcome of an experiment that has not yet been conducted, while the formulas for pdf and pmf use lower case x because they represent calculations done for each of several possible outcomes of the experiment. Also note that, in the pdf *but not* the pmf, we could replace either or both $\leq$ signs with $<$ signs because the probability that the outcome is *exactly* equal to $t$ or $u$ (to an infinite number of decimal places) is zero.

So for the continuous uniform distribution, for any $a \leq t \leq u \leq b$,

$$\Pr(t \leq X \leq u) = \int_t^u \frac{1}{b-a}dx = \frac{u-t}{b-a}.$$

You can check that this always gives a number between 0 and 1, and the probability of any individual outcome (where u=t) is zero, while the

probability that the outcome is some number between a and b is 1 (u=a, t=b). You can also see that, e.g., the probability that X is in the middle third of the interval from $a$ to $b$ is $\frac{1}{3}$, etc.

Of course, there are many interesting and useful continuous distributions other than the continuous uniform distribution. Some other examples are given below. Each is fully characterized by its probability density function.

### 3.2.1 Reading a pdf

In general, we often look at a plot of the probability density function, f(x), vs. the possible outcome values, x. This plot is high in the regions of likely outcomes and low in less likely regions. The well-known standard Gaussian distribution (see 3.2) has a bell-shaped graph centered at zero with about two thirds of its area between x = -1 and x = +1 and about 95% between x = -2 and x = +2. But a pdf can have many different shapes.

It is worth understanding that many pdf's come in "families" of similarly shaped curves. These various curves are named or "indexed" by one or more numbers called parameters (but there are other uses of the term parameter; see section 3.5). For example that family of Gaussian (also called Normal) distributions is indexed by the mean and variance (or standard deviation) of the distribution. The t-distributions, which are all centered at 0, are indexed by a single parameter called the degrees of freedom. The chi-square family of distributions is also indexed by a single degree of freedom value. The F distributions are indexed by two degrees of freedom numbers designated numerator and denominator degrees of freedom.

In this course we will not do any integration. We will use tables or a computer program to calculate probabilities for continuous random variables. We don't even need to know the formula of the pdf because the most commonly used formulas are known to the computer by name. Sometimes we will need to specify degrees of freedom or other parameters so that the computer will know which pdf of a family of pdf's to use.

Despite our heavy reliance on the computer, getting a feel for the idea of a probability density function is critical to the level of understanding of data analysis

and interpretation required in this course. At a minimum you should realize that a pdf is a curve with outcome values on the horizontal axis and the vertical height of the curve tells which values are likely and which are not. The total area under the curve is 1.0, and the under the curve between any two "x" values is the probability that the outcome will fall between those values.

**For continuous random variables, we calculate the probability that the outcome falls in some interval, not that the outcome exactly equals some value. This calculation is normally done by a computer program which uses integral calculus on a "probability density function."**

## 3.3 Probability calculations

This section reviews the most basic probability calculations. It is worthwhile, but not essential to become familiar with these calculations. For many readers, the boxed material may be sufficient. You won't need to memorize any of these formulas for this course.

Remember that in probability theory we don't worry about where probability assignments (a pmf or pdf) come from. Instead we are concerned with how to calculate other probabilities given the assigned probabilities. Let's start with calculation of the probability of a "complex" or "compound" event that is constructed from the simple events of a discrete random variable.

For example, if we have a discrete random variable that is the number of correct answers that a student gets on a test of 5 questions, i.e. integers in the set $\{0, 1, 2, 3, 4, 5\}$, then we could be interested in the probability that the student gets an even number of questions correct, or less than 2, or more than 3, or between 3 and 4, etc. All of these probabilities are for outcomes that are subsets of the sample space of all 6 possible "elementary" outcomes, and all of these are the union (joining together) of some of the 6 possible "elementary" outcomes. In the case of any complex outcome that can be written as the union of some other disjoint (non-overlapping) outcomes, the probability of the complex outcome is the sum of the probabilities of the disjoint outcomes. To complete this example look at Table 3.1 which shows assigned probabilities for the elementary outcomes of the random variable we will call T (the test outcome) and for several complex events.

| Event | Probability | Calculation |
|---|---|---|
| T=0 | 0.10 | Assigned |
| T=1 | 0.26 | Assigned |
| T=2 | 0.14 | Assigned |
| T=3 | 0.21 | Assigned |
| T=4 | 0.24 | Assigned |
| T=5 | 0.05 | Assigned |
| T$\in \{0, 2, 4\}$ | 0.48 | 0.10+0.14+0.24 |
| T<2 | 0.36 | 0.10+0.26 |
| T$\leq$2 | 0.50 | 0.10+0.26+0.14 |
| T$\leq$4 | 0.29 | 0.24+0.05 |
| T$\geq$0 | 1.00 | 0.10+0.26+0.14+0.21+0.24+0.05 |

Table 3.1: Disjoint Addition Rule

You should think of the probability of a complex event such as T<2, usually written as Pr(T<2) or P(T<2), as being the chance that, when we carry out a random experiment (e.g., test a student), the outcome will be any one of the outcomes in the defined set (0 or 1 in this case). Note that (implicitly) outcomes not mentioned are impossible, e.g., Pr(T=17) = 0. Also something must happen: Pr(T$\geq$0) = 1.00 or Pr(T $\in \{0, 1, 2, 3, 4, 5\}$) = 1.00. It is also true that the probability that nothing happens is zero: Pr($T \in \phi$) = 0, where $\phi$ means the "empty set".

**Calculate the probability that any of several non-overlapping events occur in a single experiment by adding the probabilities of the individual events.**

The addition rule for disjoint unions is really a special case of the general rule for the probability that the outcome of an experiment will fall in a set that is the union of two other sets. Using the above 5-question test example, we can define event E as the set $\{T : 1 \leq T \leq 3\}$ read as all values of outcome T such that 1 is less than or equal to T and T is less than or equal to 3. Of course E = $\{1, 2, 3\}$. Now define F = $\{T : 2 \leq T \leq 4\}$ or F = $\{2, 3, 4\}$. The union of these sets, written $E \cup F$ is equal to the set of outcomes $\{1, 2, 3, 4\}$. To find Pr($E \cup F$) we could try

adding Pr(E) + Pr(F), but we would be double counting the elementary events in common to the two sets, namely {2} and {3}, so the correct solution is to add first, and then subtract for the double counting. We define the intersection of two sets as the elements that they have in common, and use notation like $E \cap F = \{2, 3\}$ or, in situations where there is no chance of confusion, just $EF = \{2, 3\}$. Then the rule for the probability of the union of two sets is:

$$\Pr(E \cup F) = \Pr(E) + \Pr(F) - \Pr(E \cap F).$$

For our example, Pr(E F) = 0.61 + 0.59 - 0.35 = 0.85, which matches the direct calculation $\Pr(\{1, 2, 3, 4\}) = 0.26 + 0.14 + 0.21 + 0.24$. It is worth pointing out again that if we get a result for a probability that is not between 0 and 1, we are sure that we have made a mistake!

Note that it is fairly obvious that $\Pr A \cap B = \Pr B \cap A$ because $A \cap B = B \cap A$, i.e., the two events are equivalent sets. Also note that there is a complicated general formula for the probability of the union of three or more events, but you can just apply the two event formula, above, multiple times to get the same answer.

**If two events overlap, calculate the probability that either event occurs as the sum of the individual event probabilities minus the probability of the overlap.**

Another useful rule is based on the idea that something in the sample space must happen and on the definition of the complement of a set. The complement of a set, say E, is written $E^c$ and is a set made of all of the elements of the sample space that are not in set E. Using the set E above, $E^c = \{0, 4, 5\}$. The rule is:

$$\Pr(E^c) = 1 - \Pr(E).$$

In our example, $\Pr\{0, 4, 5\} = 1 - \Pr\{1, 2, 3\} = 1 - 0.61 = 0.39$.

**Calculate the probability that an event will *not* occur as 1 minus the probability that it will occur.**

Another important concept is **conditional probability**. At its core, conditional probability means reducing the pertinent sample space. For instance we might want to calculate the probability that a random student gets an odd number of questions correct while ignoring those students who score over 4 points. This is usually described as finding the probability of an odd number given $T \leq 4$. The notation is Pr(T is odd$|T \leq 4$) , where the vertical bar is pronounced "given". (The word "given" in a probability statement is usually a clue that conditional probability is being used.) For this example we are excluding the 5% of students who score a perfect 5 on the test. Our new sample space must be "renormalized" so that its probabilities add up to 100%. We can do this by replacing each probability by the old probability divided by the probability of the reduced sample space, which in this case is (1-0.05)=0.95. Because the old probabilities of the elementary outcomes in the new set of interest, $\{0, 1, 2, 3, 4\}$, add up to 0.95, if we divide each by 0.95 (making it bigger), we get a new set of 5 (instead of 6) probabilities that add up to 1.00. We can then use these new probabilities to find that the probability of interest is $0.26/0.95 + 0.21/0.95 = 0.495$.

Or we can use a new probability rule:

$$\Pr(E|F) = \frac{\Pr(E \cap F)}{\Pr(F)}.$$

In our current example, we have

$$\begin{aligned} \Pr\left(T \in \{1,3,5\} | T \leq 4\right) &= \frac{\Pr(T \in \{1,3,5\} \cap T \leq 4)}{\Pr(T \leq 4)} \\ &= \frac{\Pr(T) \in \{1,3\}}{1 - \Pr(T = 5)} = \frac{0.26 + 0.21}{0.95} = 0.495 \end{aligned}$$

**If we have partial knowledge of an outcome or are only interested in some selected outcomes, the appropriate calculations require use of the conditional probability formulas, which are based on using a new, smaller sample space.**

The next set of probability concepts relates to **independence** of events. (Sometimes students confuse disjoint and independent; be sure to keep these concepts

separate.) Two events, say E and F, are independent if the probability that event E happens, Pr(E), is the same whether or not we condition on event F happening. That is $\Pr(E) = \Pr(E|F)$. If this is true then it is also true that $\Pr(F) = \Pr(F|E)$. We use the term **marginal probability** to distinguish a probability like Pr(E) that is not conditional on some other probability. The marginal probability of E is the probability of E *ignoring* the outcome of F (or any other event). The main idea behind independence and its definition is that knowledge of whether or not F occurred does not change what we know about whether or not E will occur. It is in this sense that they are independent of each other.

Note that independence of E and F also means that $\Pr(E \cap F) = \Pr(E)\Pr(F)$, i.e., the probability that two independent events both occur is the product of the individual (marginal) probabilities.

Continuing with our five-question test example, let event A be the event that the test score, T, is greater than or equal to 3, i.e., A=$\{3, 4, 5\}$, and let B be the event that T is even. Using the union rule (for disjoint elements or sets) Pr(A) = 0.21 + 0.24 + 0.05 = 0.50, and Pr(B) = 0.10 + 0.14 + 0.24 = 0.48. From the conditional probability formula

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{\Pr(T = 4)}{\Pr(B)} = \frac{0.24}{0.48} = 0.50$$

and

$$\Pr(B|A) = \frac{\Pr(B \cap A)}{\Pr(A)} = \frac{\Pr(T = 4)}{\Pr(A)} = \frac{0.24}{0.50} = 0.48.$$

Since $\Pr(A|B) = \Pr(A)$ and $\Pr(B|A) = \Pr(B)$, events A and B are independent. We therefore can calculate that Pr(AB) = Pr(T=4) = Pr(A) Pr(B) = 0.50 (0.48) = 0.24 (which we happened to already know in this example).

If A and B are independent events, then we can calculate the probability of their intersection as the product of the marginal probabilities. If they are not independent, then we can calculate the probability of the intersection from an equation that is a rearrangement of the conditional probability formula:

$$\Pr(A \cap B) = \Pr(A|B)\Pr(B) \text{ or } \Pr(A \cap B) = \Pr(B|A)\Pr(A).$$

For our example, one calculation we can make is

$$\begin{aligned}\Pr(\text{T is even} \cap T < 2) &= \Pr(\text{T is even}|T < 2)\Pr(T < 2) \\ &= [0.10/(0.10 + 0.26)] \cdot (0.10 + 0.26) = 0.10.\end{aligned}$$

Although this is not the easiest way to calculate $\Pr(\text{T is even}|T < 2)$ for this problem, the small bag of tricks described in the chapter come in very handy for making certain calculations when only certain pieces of information are conveniently obtained.

A contrasting example is to define event G=$\{0, 2, 4\}$, and let H=$\{2, 3, 4\}$. Then G$\cap$H=$\{2, 4\}$. We can see that Pr(G)=0.48 and Pr(H)=0.59 and $\Pr(G \cap H)$=0.38. From the conditional probability formula

$$\Pr(G|H) = \frac{\Pr(G \cap H)}{\Pr(H)} = \frac{0.38}{0.59} = 0.644.$$

So, if we have no knowledge of the random outcome, we should say there is a 48% chance that T is even. But if we have the partial outcome that T is between 2 and 4 inclusive, then we revise our probability estimate to a 64.4% chance that T is even. Because these probabilities differ, we can say that event G is *not* independent of event H. We can "check" our conclusion by verifying that the probability of G$\cap$H (0.38) is *not* the product of the marginal probabilities, $0.48 \cdot 0.59 = 0.2832$.

Independence also applies to random variables. Two random variables are independent if knowledge of the outcome of one does not change the (conditional) probability of the other. In technical terms, if $\Pr(X|Y = y) = \Pr(X)$ for all values of $y$, then $X$ and $Y$ are independent random variables. If two random variables are independent, and if you consider any event that is a subset of the $X$ outcomes and any other event that is a subset of the $Y$ outcomes, these events will be independent.

**At an intuitive level, events are independent if knowledge that one event has or has not occurred does not provide new information about the probability of the other event. Random variables are independent if knowledge of the outcome of one does not provide new information about the probabilities of the various outcomes of the other. In most experiments it is reasonable to assume that the outcome for any one subject is independent of the outcome of any other subject. If two events are independent, the probability that both occur is the product of the individual probabilities.**

## 3.4 Populations and samples

In the context of experiments, observational studies, and surveys, we make our actual measurements on individual **observational units** . These are commonly people (subjects, participants, etc.) in the social sciences, but can also be schools, social groups, economic entities, archaeological sites, etc. (In some complicated situations we may make measurements at multiple levels, e.g., school size and students' test scores, which makes the definition of experimental units more complex.)

We use the term **population** to refer to the entire set of actual or potential observational units. So for a study of working memory, we might define the population as all U.S. adults, as all past present and future human adults, or we can use some other definition. In the case of, say, the U.S. census, the population is reasonably well defined (although there are problems, referred to in the census literature as "undercount") and is large, but finite. For experiments, the definition of population is often not clearly defined, although such a definition can be very important. See section 8.3 for more details. Often we consider such a population to be theoretically infinite, with no practical upper limit on the number of potential subjects we could test.

For most studies (other than a census), only a subset of all of the possible experimental units of the population are actually selected for study, and this is called the **sample** (not to be confused with sample space). An important part of the understanding of the idea of a sample is to realize that each experiment is conducted on a particular sample, but might have been conducted on many other different samples. For theoretically correct inference, the sample should be

randomly selected from the population. If this is not true, we call the sample a **convenience sample**, and we lose many of the theoretical properties required for correct inference.

Even though we must use samples in science, it is very important to remember that we are interested in learning about populations, not samples. Inference from samples to populations is the goal of statistical analysis.

## 3.5 Parameters describing distributions

As mentioned above, the probability distribution of a random variable (pmf for a discrete random variable or pdf for a continuous random variable) completely describes its behavior in terms of the chances that various events will occur. It is also useful to work with certain fixed quantities that either completely characterize a distribution within a family of distributions or otherwise convey useful information about a distribution. These are called **parameters**. Parameters are fixed quantities that characterize theoretical probability distributions. (I am using the term "theoretical distribution" to focus on the fact that we are assuming a particular mathematical form for the pmf or pdf.)

The term parameter may be somewhat confusing because it is used in several slightly different ways. Parameters may refer to the fixed constants that appear in a pdf or pmf. Note that these are somewhat arbitrary because the pdf or pmf may often be rewritten (technically, re-parameterized) in several equivalent forms. For example, the binomial distribution is most commonly written in terms of a probability, but can just as well be written in terms of odds.

Another related use of the term parameter is for a summary measure of a particular (theoretical) probability distribution. These are most commonly in the form of **expected values**. Expected values can be thought of as long-run averages of a random variable or some computed quantity that includes the random variable. For discrete random variables, the expected value is just a probability weighted average, i.e., the **population mean**. For example, if a random variable takes on (only) the values 2 and 10 with probabilities 5/6 and 1/6 respectively, then the expected value of that random variable is 2(5/6)+10(1/6)=20/6. To be a bit more concrete, if someone throws a die each day and gives you \$10 if 5 comes up and \$2 otherwise, then over $n$ days, where $n$ is a large number, you will end up with very close to $\$\frac{20 \cdot n}{6}$, or about \$3.67(n).

The notation for expected value is $E[\cdot]$ or $E(\cdot)$ where, e.g., $E[X]$ is read as "expected value of X" and represents the population mean of $X$. Other parameters such as variance, skewness and kurtosis are also expected values, but of expressions involving $X$ rather than of $X$ itself.

The more general formula for expected value is

$$E[g(X)] = \sum_{i=1}^{k} g(x_i)p_i = \sum_{i=1}^{k} g(x_i)f(x_i)$$

where $E[\cdot]$ or $E(\cdot)$ represents "expected value", $g(X)$ is any function of the random variable $X$, $k$ (which may be infinity) is the number of values of $X$ with non-zero probability, the $x_i$ values are the different values of $X$, and the $p_i$ values (or equivalently, $f(x_i)$) are the corresponding probabilities. Note that it is possible to define $g(X) = X$, i.e., $g(x_i) = x_i$, to find $E(X)$ itself.

The corresponding formula for expected value of a continuous random variable is

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx.$$

Of course if the support is smaller than the entire real line, the pdf is zero outside of the support, and it is equivalent to write the integration limits as only over the support.

To help you think about this concept, consider a discrete random variable, say $W$, with values -2, -1, and 3 with probabilities 0.5, 0.3, 0.2 respectively. $E(W) = -2(0.5) - 1(0.3) + 3(0.2) = -0.7$. What is $E(W^2)$? This is equivalent to letting $g(W) = W^2$ and finding $E(g(W)) = E(W^2)$. Just calculate $W^2$ for each $W$ and take the weighted average: $E(W^2) = 4(0.5) + 1(0.3) + 9(0.2) = 4.1$. It is also equivalent to define, say, $U = W^2$. Then we can express $f(U)$ as $U$ has values 4, 1, and 9 with probabilities 0.5, 0.3, and 0.2 respectively. Then $E(U) = 4(0.5) + 1(0.3) + 9(0.2) = 4.1$, which is the same answer.

Different parameters are generated by using different forms of $g(x)$.

| **Name** | **Definition** | **Symbol** |
|---|---|---|
| mean | $E[X]$ | $\mu$ |
| variance | $E[(X-\mu)^2]$ | $\sigma^2$ |
| standard deviation | $\sqrt{\sigma^2}$ | $\sigma$ |
| skewness | $E[(X-\mu)^3]/\sigma^3$ | $\gamma_1$ |
| kurtosis | $E[(X-\mu)^4]/\sigma^4 - 3$ | $\gamma_2$ |

Table 3.2: Common parameters and their definitions as expected values.

You will need to become familiar with several parameters that are used to characterize theoretical population distributions. Technically, many of these are defined using the expected value formula (optional material) with the expressions shown in table 3.2. You only need to become familiar with the names and symbols and their general meanings, not the "Definition" column. Note that the symbols shown are the most commonly used ones, but you should not assume that these symbol always represents the corresponding parameters or vice versa.

### 3.5.1 Central tendency: mean and median

The **central tendency** refers to ways of specifying where the "middle" of a probability distribution lies. Examples include the mean and median parameters. The mean (expected value) of a random variable can be thought of as the "balance point" of the distribution if the pdf is cut out of cardboard. Or if the outcome is some monetary payout, the mean is the appropriate amount to bet to come out even in the long term. Another interpretation of mean is the "fair distribution of outcome" in the sense that if we sample many values and think of them as one outcome per subject, the mean is result of a fair redistribution of whatever the outcome represents among all of the subjects. On the other hand, the median is the value that splits the distribution in half so that there is a 50/50 chance of a random value from the distribution occurring above or below the median.

The median has a more technical definition that applies even in some less common situations such as when a distribution does not have a single unique median. The median is any $m$ such that $\mathrm{P}(X \leq m) \geq \frac{1}{2}$ and $\mathrm{P}(X \geq m) \geq \frac{1}{2}$.

### 3.5.2 Spread: variance and standard deviation

The **spread** of a distribution most commonly refers to the variance or standard deviation parameter, although other quantities such as interquartile range are also measures of spread.

The **population variance** is the mean squared distance of any value from the mean of the distribution, but you only need to think of it as a measure of spread on a different scale from standard deviation. The **standard deviation** is defined as the square root of the variance. It is not as useful in statistical formulas and derivations as the variance, but it has several other useful properties, so both variance and standard deviation are commonly calculated in practice. The standard deviation is in the same units as the original measurement from which it is derived. For each theoretical distribution, the intervals $[\mu-\sigma,\ \mu+\sigma]$, $[\mu-2\sigma,\ \mu+2\sigma]$, and $[\mu-3\sigma,\ \mu+3\sigma]$ include fixed known amounts of the probability. It is worth memorizing that *for Gaussian distributions only* these fractions are 0.683, 0.954, and 0.997 respectively. (I usually think of this as approximately 2/3, 95% and 99.7%.) Also exactly 95% of the Gaussian distribution is in $[\mu-1.96\sigma,\ \mu+1.96\sigma]$

When the standard deviation of repeated measurements is proportional to the mean, then instead of using standard deviation, it often makes more sense to measure variability in terms of the **coefficient of variation**, which is the s.d. divided by the mean.

There is a special statistical theorem (called Chebyshev's inequality) that applies to *any* shaped distribution and that states that at least $\left(1 - \frac{1}{k^2}\right) \times 100\%$ of the values are within $k$ standard deviations from the mean. For example, the interval $[\mu{-}1.41\sigma,\ \mu{+}1.41\sigma]$ holds at least 50% of the values, $[\mu{-}2\sigma,\ \mu{+}2\sigma]$ holds at least 75% of the values, and $[\mu{-}3\sigma,\ \mu{+}3\sigma]$ holds at least 89% of the values.

### 3.5.3 Skewness and kurtosis

The **population skewness** of a distribution is a measure of asymmetry (zero is symmetric) and the population kurtosis is a measure of peakedness or flatness compared to a Gaussian distribution, which has $\gamma_2 = 0$. If a distribution is "pulled out" towards higher values (to the right), then it has positive **skewness**. If it is pulled out toward lower values, then it has negative skewness. A symmetric distribution, e.g., the Gaussian distribution, has zero skewness.

The **population kurtosis** of a distribution measures how far away a distribution is from a Gaussian distribution in terms of peakedness vs. flatness. Compared to a Gaussian distribution, a distribution with negative kurtosis has "rounder shoulders" and "thin tails", while a distribution with a positive kurtosis has more a more sharply shaped peak and "fat tails".

### 3.5.4 Miscellaneous comments on distribution parameters

Mean, variance, skewness and kurtosis are called **moment** estimators. They are respectively the $1^{st}$ through $4^{th}$ (central) moments. Even simpler are the non-central moments: the $r^{th}$ non-central moment of $X$ is the expected value of $X^r$. There are formulas for calculating central moments from non-central moments. E.g., $\sigma^2 = E(X^2) - E(X)^2$.

It is important to realize that for any particular distribution (but not family of distributions) each parameter is a fixed constant. Also, you will recognize that

these parameter names are the same as the names of statistics that can be calculated for and used as descriptions of **samples** rather than probability distributions (see next chapter). The prefix "population" is sometimes used as a reminder that we are talking about the fixed numbers for a given probability distribution rather than the corresponding sample values.

It is worth knowing that any formula applied to one or more parameters creates a new parameter. For example, if $\mu_1$ and $\mu_2$ are parameters for some population, say, the mean dexterity with the subjects' dominant and non-dominant hands, then $\log(\mu_1)$, $\mu_2^2$, $\mu_1 - \mu_2$ and $(\mu_1 + \mu_2)/2$ are also parameters.

In addition to the parameters in the above table, which are the most common descriptive parameters that can be calculated for any distribution, fixed constants in a pmf or pdf, such as degrees of freedom (see below) or the $n$ in the binomial distribution are also (somewhat loosely) called parameters.

Technical note: For some distributions, parameters such as the mean or variance may be infinite.

**Parameters such as (population) mean and (population) variance are fixed quantities that characterize a given probability distribution. The (population) skewness characterizes symmetry, and (population) kurtosis characterizes symmetric deviations from Normality. Corresponding sample statistics can be thought of as sample estimates of the population quantities.**

### 3.5.5 Examples

As a review of the concepts of theoretical population distributions (in the continuous random variable case) let's consider a few examples.

Figure 3.1 shows five different pdf's representing the (population) probability distributions of five different continuous random variables. By the rules of pdf's, the area under each of the five curves equals exactly 1.0, because that represents

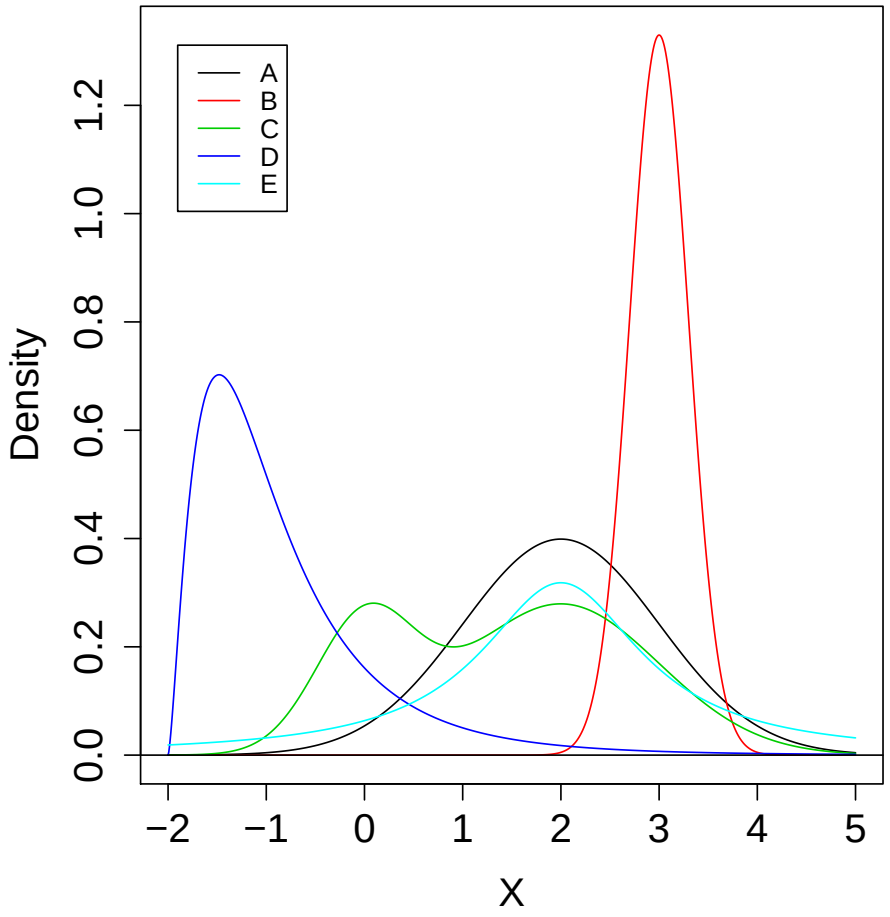


Figure 3.1: Various probability density function

the probability that a random outcome from a distribution is between -infinity and +infinity. (The area shown, between -2 and +5 is slightly less than 1.0 for each distribution because there is a small chance that these variables could have an outcome outside of the range shown.) You can see that distribution **A** is a unimodal (one peak) symmetric distribution, centered around 2.0. Although you cannot see it by eye, it has the perfect bell-shape of a Gaussian distribution. Distribution **B** is also Gaussian in shape, has a different central tendency (shifted higher or rightward), and has a smaller spread. Distribution **C** is bimodal (two peaks) so it cannot be a Gaussian distribution. Distribution **D** has the lowest center and is asymmetric (skewed to the right), so it cannot be Gaussian. Distribution **E** appears similar to a Gaussian distribution, but while symmetric and roughly bell-shaped, it has "tails" that are too fat to be a true bell-shaped, Gaussian distribution.

So far we have been talking about the parameters of a given, known, theoretical probability distribution. A slightly different context for the use of the term parameter is in respect to a real world population, either finite (but usually large) or infinite. As two examples, consider the height of all people living on the earth at 3:57 AM GMT on September 10, 2007, or the birth weights of all of the Sprague-Dawley breed of rats that could possibly be bred. The former is clearly finite, but large. The latter is perhaps technically finite due to limited resources, but may also be thought of as (practically) infinite. Each of these must follow some true distribution with fixed parameters, but these are practically unknowable. The best we can do with experimental data is to make an estimate of the fixed, true, unknowable parameter value. For this reason, I call parameters in this context "secrets of nature" to remind you that they are not random and they are not practically knowable.

## 3.6 Multivariate distributions: joint, conditional, and marginal

The concepts of this section are fundamentals of probability, but for the typical user of statistical methods, only a passing knowledge is required. More detail is given here for the interested reader.

So far we have looked at the distribution of a single random variable at a time. Now we proceed to look at the **joint distribution** of two (or more) random variables. First consider the case of two categorical random variables. As an

example, consider the population of all cars produced in the world in 2006. (I'm just making up the numbers here.) This is a large finite population from which we might sample cars to do a fuel efficiency experiment. If we focus on the categorical variable "origin" with levels "US","Japanese", and "Other", and the categorical variable "size" with categorical variable "Small", "Medium" and "Large", then table 3.3 would represent the joint distribution of origin and size in this population.

| origin / size | Small | Medium | Large | Total |
|---|---|---|---|---|
| US | 0.05 | 0.10 | 0.15 | |
| Japanese | 0.20 | 0.10 | 0.05 | |
| Other | 0.15 | 0.15 | 0.05 | |
| Total | | | | 1.00 |

Table 3.3: Joint distribution of car origin and size.

These numbers come from categorizing all cars, then dividing the total in each combination of categories by the total cars produced in the world in 2006, so they are "relative frequencies". But because we are considering this the whole population of interest, it is better to consider these numbers to be the probabilities of a (joint) pmf. Note that the total of all of the probabilities is 1.00. Reading this table we can see, e.g., that 20% of all 2006 cars were small Japanese cars, or equivalently, the probability that a randomly chosen 2006 car is a small Japanese car is 0.20.

The joint distribution of $X$ and $Y$ is summarized in the joint pmf, which can be tabular or in formula form, but in either case is similar to the one variable pmf of section 3.2 except that it defines a probability for each combination of levels of $X$ and $Y$.

This idea of a joint distribution, in which probabilities are given for the combination of levels of two categorical random variables, is easily extended to three or more categorical variables.

**The joint distribution of a pair of categorical random variables represents the probabilities of combinations of levels of the two individual random variables.**

| origin / size | Small | Medium | Large | Total |
|---|---|---|---|---|
| US | 0.05 | 0.10 | 0.15 | 0.30 |
| Japanese | 0.20 | 0.10 | 0.05 | 0.35 |
| Other | 0.15 | 0.15 | 0.05 | 0.35 |
| Total | 0.40 | 0.35 | 0.25 | (1.00) |

Table 3.4: Marginal distributions of car origin and size.

Table 3.4 adds the obvious margins to the previous table, by adding the rows and columns and putting the sums in the margins (labeled "Total"). Note that both the right vertical and bottom horizontal margins add to 1.00, and so they each represent a probability distribution, in this case of origin and size respectively. These distributions are called the **marginal distributions** and each represents the pmf of one of the variable *ignoring* the other variable. That is, a marginal distribution is the distribution of any particular variable when we don't pay any attention to the other variable(s). If we had only studied car origins, we would have found the population distribution to be 30% US, 35% Japanese and 35% other.

It is important to understand that every variable we measure is marginal with respect to all of the other variables that we could measure on the same units or subjects, and which we do not in any way control (or in other words, which we let vary freely).

**The marginal distribution of any variable with respect to any other variable(s) is just the distribution of that variable ignoring the other variable(s).**

The third and final definition for describing distributions of multiple characteristics of a population of units or subjects is the **conditional distribution** which relates to conditional probability (see page 31). As shown in table 3.5, the conditional distribution refers to fixing the level of one variable, then "re-normalizing" to find the probability level of the other variable when we only focus on or consider those units or subjects that meeting the condition of interest.

So if we focus on Japanese cars only (technically, we condition on cars be-

| origin / size | Small | Medium | Large | Total |
|---|---|---|---|---|
| US | 0.167 | 0.333 | 0.400 | 1.000 |
| Japanese | 0.571 | 0.286 | 0.143 | 1.000 |
| Other | 0.429 | 0.429 | 0.142 | 1.000 |

Table 3.5: Conditional distributions of car size given its origin.

ing Japanese) we see that 57.1% of those cars are small, which is very different from either the marginal probability of a car being small (0.40) or the joint probability of a car being small and Japanese (0.20). The formal notation here is Pr(size=small|origin=Japanese) = 0.571, which is read "the probability of a car being small given that the car is Japanese equals 0.571".

It is important to realize that there is another set of conditional distributions for this example that we have not looked at. As an exercise, try to find the conditional distributions of "origin" given "size", which differ from the distributions of "size" given "origin" of table 3.5.

It is interesting and useful to note that an equivalent alternative to specifying the complete joint distribution of two categorical (or quantitative) random variables is to specify the marginal distribution of one variable, and the conditional distributions for the second variable at each level of the first variable. For example, you can reconstruct the joint distribution for the cars example from the marginal distribution of "origin" and the three conditional distributions of "size given origin". This leads to another way to think about marginal distributions as the distribution of one variable *averaged over* the distribution of the other.

**The distribution of a random variable conditional on a particular level of another random variable is the distribution of the first variable when the second variable is fixed to the particular level.**

The concepts of joint, marginal and conditional distributions transfer directly to two continuous distributions, or one continuous and one joint distribution, but the details will not be given here. Suffice it to say the the joint pdf of two continuous random variables, say $X$ and $Y$ is a formula with both $x$s and $y$s in it.

### 3.6.1 Covariance and Correlation

For two quantitative variables, the basic parameters describing the strength of their relationship are **covariance** and **correlation**. For both, larger absolute values indicate a stronger relationship, and positive numbers indicate a direct relationship while negative numbers indicate an indirect relationship. For both, a value of zero is called uncorrelated. Covariance depends on the scale of measurement, while correlation does not. For this reason, correlation is easier to understand, and we will focus on that here, although if you look at the gray box below, you will see that covariance is used as in intermediate in the calculation of correlation. (Note that here we are concerned with the "population" or "theoretical" correlation. The sample version is covered in the EDA chapter.)

Correlation describes both the strength and direction of the (linear) relationship between two variables. Correlations run from -1.0 to +1.0. A negative correlation indicates an "inverse" relationship such that population units that are low for one variable tend to be high for the other (and vice versa), while a positive correlation indicates a "direct" relationship such that population units that are low in one variable tend to be low in the other (also high with high). A zero correlation (also called **uncorrelated**) indicates that the "best fit straight line" (see the chapter on Regression) for a plot of $X$ vs. $Y$ is horizontal, suggesting no relationship between the two random variables. Technically, independence of two variables (see above) implies that they are uncorrelated, but the reverse is not necessarily true.

For a correlation of +1.0 or -1.0, $Y$ can be perfectly predicted from $X$ with no error (and vice versa) using a linear equation. For example if $X$ is temperature of a rat in degrees C and $Y$ is temperature in degrees F, then $Y = 9/5 * C + 32$, exactly, and the correlation is +1.0. And if $X$ is height in feet of a person from the floor of a room with an 8 foot ceiling and $Y$ is distance from the top of the head to the ceiling, then $Y = 8 - X$, exactly, and the correlation is -1.0. For other variables like height and weight, the correlation is positive, but less than 1.0. And for variables like $IQ$ and length of the index finger, the correlation is presumably 0.0.

It should be obvious that the correlation of any variable with itself is 1.0. Let us represent the population correlation between random variable $X_i$ and random variable $X_j$ as $\rho_{i,j}$. Because the correlation of $X$ with $Y$ is the same as $Y$ with $X$, it is true that $\rho_{i,j} = \rho_{j,i}$. We can compactly represent the relationships between multiple variables with a **correlation matrix** which shows all of the pairwise correlations in a square table of numbers (square matrix). An example is given in table 3.6 for the case of 4 variables. As with all correlations matrices, the matrix is symmetric with a row of ones on the main diagonal. For some actual population and variables, we could put numbers instead of symbols in the matrix, and then make statements about which variables are directly vs. inversely vs. not correlated, and something about the strengths of the correlations.

| Variable | $X_1$ | $X_2$ | $X_3$ | $X_4$ |
|---|---|---|---|---|
| $X_1$ | 1 | $\rho_{1,2}$ | $\rho_{1,3}$ | $\rho_{1,4}$ |
| $X_2$ | $\rho_{2,1}$ | 1 | $\rho_{2,3}$ | $\rho_{2,4}$ |
| $X_3$ | $\rho_{3,1}$ | $\rho_{3,2}$ | 1 | $\rho_{3,4}$ |
| $X_4$ | $\rho_{4,1}$ | $\rho_{4,2}$ | $\rho_{4,3}$ | 1 |

Table 3.6: Population correlation matrix for four variables.

There are several ways to measure "correlation" for categorical variables and choosing among them can be a source of controversy that we will not cover here. But for quantitative random variables covariance and correlation are mathematically straightforward.

The population covariance of two quantitative random variables, say $X$ and $Y$, is calculated by computing the expected value (population mean) of the quantity $(X - \mu_X)(Y - \mu_Y)$ where $\mu_X$ is the population mean of $X$ and $\mu_Y$ is the population mean of $Y$ across all combinations of $X$ and $Y$. For continuous random variables this is the double integral

$$\text{Cov}_{X,Y} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y) f(x, y) dx dy$$

where $f(x, y)$ is the joint pdf of $X$ and $Y$.

For discrete random variables we have the simpler form

$$\text{Cov}_{X,Y} = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} (x - \mu_X)(y - \mu_Y) f(x, y)$$

where $f(x, y)$ is the joint pmf, and $\mathcal{X}$ and $\mathcal{Y}$ are the respective supports of $X$ and $Y$.

As an example consider a population consisting of all of the chickens of a particular breed (that only lives 4 years) belonging to a large multi-farm poultry company in January of 2007. For each chicken in this population we have $X$ equal to the number of eggs laid in the first week of January and $Y$ equal to the age of the chicken in years. The joint pmf of $X$ and $Y$ is given in table 3.7. As usual, the joint pmf gives the probabilities that a random subject will fall into each combination of categories from the two variables.

We can calculate the (marginal) mean number of eggs from the marginal distribution of eggs as $\mu_X = 0(0.35) + 1(0.40) + 2(0.25) = 0.90$ and the mean age as $\mu_Y = 1(0.25) + 2(0.40) + 3(0.20) + 4(0.15) = 2.25$ years.

The calculation steps for the covariance are shown in table 3.8. The population covariance of $X$ and $Y$ is 0.075 (exactly). The (weird) units are "egg years".

Population correlation can be calculated from population covariance and the two individual standard deviations using the formula

$$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sigma_x \sigma_y}.$$

In this case $\sigma_X^2 = (0-0.9)^2(0.35)+(1-0.9)^2(0.40)+(2-0.9)^2(0.25) = 0.59$. Using a similar calculation for $\sigma_Y^2$ and taking square roots to get standard deviation from variance, we get

$$\rho_{X,Y} = \frac{0.075}{0.7681 \cdot 0.9937} = 0.0983$$

which indicates a weak positive correlation: older hens lay more eggs.

| Y (year) / X (eggs) | 0 | 1 | 2 | Margin |
|---|---|---|---|---|
| 1 | 0.10 | 0.10 | 0.05 | 0.25 |
| 2 | 0.15 | 0.15 | 0.10 | 0.40 |
| 3 | 0.05 | 0.10 | 0.05 | 0.20 |
| 4 | 0.05 | 0.05 | 0.05 | 0.15 |
| Margin | 0.35 | 0.40 | 0.25 | 1.00 |

Table 3.7: Chicken example: joint population pmf.

| X | Y | X-0.90 | Y-2.25 | Pr | Pr·(X-0.90)(Y-2.25) |
|---|---|---|---|---|---|
| 0 | 1 | -0.90 | -1.25 | 0.10 | 0.11250 |
| 1 | 1 | 0.10 | -1.25 | 0.10 | -0.00125 |
| 2 | 1 | 1.10 | -1.25 | 0.05 | -0.06875 |
| 0 | 2 | -0.90 | -0.25 | 0.15 | 0.03375 |
| 1 | 2 | 0.10 | -0.25 | 0.15 | -0.00375 |
| 2 | 2 | 1.10 | -0.25 | 0.10 | -0.02750 |
| 0 | 3 | -0.90 | 0.75 | 0.05 | -0.03375 |
| 1 | 3 | 0.10 | 0.75 | 0.10 | 0.00750 |
| 2 | 3 | 1.10 | 0.75 | 0.05 | 0.04125 |
| 0 | 4 | -0.90 | 1.75 | 0.05 | -0.07875 |
| 1 | 4 | 0.10 | 1.75 | 0.05 | 0.00875 |
| 2 | 4 | 1.10 | 1.75 | 0.05 | 0.09625 |
| | | | Total | 1.00 | 0.07500 |

Table 3.8: Covariance calculation for chicken example.

**In a nutshell: When dealing with two (or more) random variables simultaneously it is helpful to think about joint vs. marginal vs. conditional distributions. This has to do with what is fixed vs. what is free to vary, and what adds up to 100%. The parameter that describes the strength of relationship between two random variables is the correlation, which ranges from -1 to +1.**

## 3.7 Key application: sampling distributions

In this course we will generally be concerned with analyzing a **simple random sample** of size $n$ which indicates that we randomly and independently choose $n$ subjects from a large or infinite population for our experiment. (For practical issues, see section 8.3.) Then we make one or more measurements, which are the realizations of some random variable. Often we combine these values into one or more **statistics**. A statistic is defined as any formula or "recipe" that can be explicitly calculated from observed data. Note that the formula for a statistic must not include unknown parameters. *When thinking about a statistics always remember that this is only one of many possible values that we could have gotten for this statistic, based on the random nature of the sampling.*

If we think about random variable $X$ for a sample of size $n$ it is useful to consider this a multivariate situation, i.e., the outcome of the random trial is $X_1$ through $X_n$ and there is a probability distribution for this multivariate outcome. If we have simple random sampling, this $n$-fold pmf or pdf is calculable from the distribution of the original random variable and the laws of probability with independence. Technically we say that $X_1$ through $X_n$ are **iid** which stands for independent and identically distributed, which indicates that distribution of the outcome for, say, the third subject, is the same as for any other subject and is independent of (does not depend on the outcome of) the outcome for every other subject.

An example should make this clear. Consider a simple random sample of size $n = 3$ from a population of animals. The random variable we will observe is gender, and we will call this $X$ in general and $X_1$, $X_2$ and $X_3$ in particular. Lets say that we know the parameter that represent the true probability that an animal is male is equal to 0.4. Then the probability that an animal is female is 0.6. We can work out the multivariate pmf case by case as is shown in table 3.7. For example, the

| $X_1$ | $X_2$ | $X_3$ | Probability |
|---|---|---|---|
| F | F | F | 0.216 |
| M | F | F | 0.144 |
| F | M | F | 0.144 |
| F | F | M | 0.144 |
| F | M | M | 0.096 |
| M | F | M | 0.096 |
| M | M | F | 0.096 |
| M | M | M | 0.064 |
| Total | | | |

Table 3.9: Multivariate pmf for animal gender.

chance that the outcome is FMF in that order is (0.6)(0.4)(0.6)=0.144.

Using this multivariate pmf, we can easily calculate the pmf for derived random variables (statistics) such as $Y$=the number of females in the sample: Pr(Y=0)=0.064, Pr(Y=1)=0.288, Pr(Y=2)=0.432, and Pr(Y=3)=0.216.

Now think carefully about what we just did. We found the probability distribution of random variable $Y$, the number of females in a sample of size three. This is called the **sampling distribution** of $Y$, which refers to the fact that $Y$ is a random quantity which varies from sample to sample over many possible samples (or experimental runs) that *could* be carried out if we had enough resources. We can find the sampling distribution of various sample quantities constructed from the data of a random sample. These quantities are **sample statistics**, and can take many different forms. Among these are the sample versions of mean, variance, standard deviation, etc. Quantities such as the sample mean or sample standard deviation (see section 4.2) are often used as estimates of the corresponding population parameters. The sampling distribution of a sample statistic is then the key way to evaluate *how good of an estimate* a sample statistic is. In addition, we use various sample statistics and their sampling distributions to make probabilistic conclusions about statistical hypotheses, usually in the form of statements about population parameters.

**Much of the statistical analysis of experiments is grounded in calculation of a sample statistic, computation of its sampling distribution (using a computer), and using the sampling distribution to draw inferences about statistical hypotheses.**

## 3.8 Central limit theorem

The Gaussian (also called bell-shaped or Normal) distribution is a very common one. The central limit theorem (CLT) explains why many real-world variables follow a Gaussian distribution.

It is worth reviewing here what "follows a particular distribution" really means. A random variable follows a particular distribution if the observed probability of each outcome for a discrete random variable or the the observed probabilities of a reasonable set of intervals for a continuous random variable are well approximated by the corresponding probabilities of some named distribution (see Common Distributions, below). Roughly, this means that a histogram of the actual random outcomes is quite similar to the theoretical histogram of potential outcomes defined by the pmf (if discrete) or pdf (if continuous). For example, for any Gaussian distribution with mean $\mu$ and standard deviation $\sigma$, we expect 2.3% of values to fall below $\mu - 2\sigma$, 13.6% to fall between $\mu$–$2\sigma$ and $\mu$–$\sigma$, 34.1% between $\mu$–$\sigma$ and $\mu$, 34.1% between $\mu$ and $\mu$+$\sigma$, 13.6% between $\mu$+$\sigma$ and $\mu$+$2\sigma$, and 2.3% above $\mu$+$2\sigma$. In practice we would check a finer set of divisions and/or compare the shapes of the actual and theoretical distributions either using histograms or a special tool called the quantile-quantile plot.

In non-mathematical language, the "CLT" says that *whatever* the pmf or pdf of a variable is, if we randomly sample a "large" number (say $k$) of independent values from that random variable, the sum or mean of those $k$ values, if collected repeatedly, will have a Normal distribution. It takes some extra thought to understand what is going on here. The process I am describing here takes a sample of (independent) outcomes, e.g., the weights of all of the rats chosen for an experiment, and calculates the mean weight (or sum of weights). Then we consider the less practical process of repeating the whole experiment many, many times (taking a new sample of rats each time). If we would do this, the CLT says that a histogram of all of these mean weights across all of these experiments would show

a Gaussian shape, even if the histogram of the individual weights of any one experiment were not following a Gaussian distribution. By the way, the distribution of the means across many experiments is usually called the "sampling distribution of the mean".

For practical purposes, a number as small as 20 (observations per experiment) can be considered "large" when invoking the CLT if the original distribution is not very bizarre in shape and if we only want a reasonable approximation to a Gaussian curve. And for almost all original distributions, the larger $k$ is, the closer the distribution of the means or sums are to a Gaussian shape.

It is usually fairly easy to find the mean and variance of the sampling distribution (see section 3.7) of a statistic of interest (mean or otherwise), but finding the *shape* of this sampling distribution is more difficult. The Central Limit Theorem lets us predict the (approximate) shape of the sampling distribution for sums or means. And this additional shape information is usually all that is needed to construct valid confidence intervals and/or p-values.

But wait, there's more! The central limit theorem also applies to the sum or mean of many *different* independent random variables as long as none of them strongly dominates the others. So we can invoke the CLT as an explanation for why many real-world variables happen to have a Gaussian distribution. It is because they are the result of many small independent effects. For example, the weight of 12-week-old rats varies around the mean weight of 12-week-old rats due to a variety of genetic factors, differences in food availability, differences in exercise, differences in health, and a variety of other environmental factors, each of which adds or subtracts a little bit relative to the overall mean.

See one of the theoretical statistics texts listed in the bibliography for a proof of the CLT.

**The Central Limit Theorem is the explanation why many real-world random variables tend to have a Gaussian distribution. It is also the justification for assuming that if we could repeat an experiment many times, any sample mean that we calculate once per experiment would follow a Gaussian distribution over the many experiments.**

## 3.9 Common distributions

A brief description of several useful and commonly used probability distributions is given here. The casual reader will want to just skim this material, then use it as reference material as needed.

The two types of distributions are discrete and continuous (see above), which are fully characterized by their pmf or pdf respectively. In the notation section of each distribution we use "$X \sim$" to mean "X is distributed as".

What does it mean for a random variable to follow a certain distribution? It means that the pdf or pmf of that distribution fully describes the probabilities of events for that random variable. Note that each of the named distributions described below are a family of related individual distributions from which a specific distribution must be specified using an index or pointer into the family usually called a parameter (or sometimes using 2 parameters). For a theoretical discussion, where we assume a particular distribution and then investigate what properties follow, the pdf or pmf is all we need.

For data analysis, we usually need to choose a theoretical distribution that we think will well approximate our measurement for the population from which our sample was drawn. This can be done using information about what assumptions lead to each distribution, looking at the support and shape of the sample distribution, and using prior knowledge of similar measurements. Usually we choose a family of distributions, then use statistical techniques to estimate the parameter that chooses the particular distribution that best matches our data. Also, after carrying out a statistical test that assumes a particular family of distributions, we use model checking, such as residual analysis, to verify that our choice was a good one.

### 3.9.1 Binomial distribution

The **binomial distribution** is a discrete distribution that represents the number of successes in $n$ independent trials, each of which has success probability $p$. All of the (infinite) different values of $n$ and $p$ define a whole family of different binomial distributions. The outcome of a random variable that follows a binomial distribution is a whole number from 0 to $n$ (i.e., n+1 different possible values). If $n = 1$, the special name **Bernoulli distribution** may be used. If random variable $X$ follows a Bernoulli distribution with parameter $p$, then stating that $\Pr(X = 1) = p$

and $\Pr(X = 0) = 1 - p$ fully defines the distribution of $X$.

If we let $X$ represent the random outcome of a binomial random variable with parameters $n$ and $p$, and let $x$ represent any particular outcome (as a whole number from 0 to $n$), then the pmf of a binomial distribution tells us the probability that the outcome will be $x$:

$$\Pr(X = x) = f(x) = \left(\frac{n!}{(n-x)!\,x!}\right) p^x (1-p)^{n-x}.$$

As a reminder, the exclamation mark symbol is pronounced "factorial" and $r!$ represents the product of all the integers from 1 to $r$. As an exception, $0! = 1$.

The true, theoretical mean of a binomial distribution is $np$ and the variance is $np(1-p)$. These refer to the ideal for an infinite population. For a sample, the sample mean and variance will be similar to the theoretical values, and the larger the sample, the more sure we are that the sample mean and variance will be very close to the theoretical values.

As an example, if you buy a lottery ticket for a daily lottery choosing your lucky number each of 5 different days in a lottery with a 1/500 chance of winning each time, then knowing that these chances are independent, we could call the number of times (out of 5) that you win $Y$, and state that $Y$ is distributed according to a binomial distribution with $n = 5$ and $p = 0.002$. We now know that if many people each independently buy 5 lottery tickets they will each have an outcome between 0 and 5, and the mean of all of those outcomes will be (close to) $np = 5(0.002) = 0.01$ and the variance will be (close to) $np(1-p) = 5(0.002)(0.998) = 0.00998$ (with sd=$\sqrt{0.0098} = 0.0999$.)

In this example we can calculate $n! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$, and for x=2, $(n-x)! = 3! = 3 \cdot 2 \cdot 1 = 6$ and $x! = 2! = 2 \cdot 1 = 2$. So

$$\Pr(X = 2) = \left(\frac{120}{6 \cdot 2}\right) 0.002^2 (0.998)^3 = 0.0000398.$$

Roughly 4 out of 100,000 people will win twice in 5 days.

It is sometimes useful to know that with large $n$ a binomial random variable with parameter $p$ approximates a Normal distribution with mean $np$ and variance $np(1-p)$ (except that there are gaps in the binomial because it only takes on whole numbers).

Common notation is $X \sim \text{bin}(n, p)$.

### 3.9.2 Multinomial distribution

The **multinomial distribution** is a discrete distribution that can be used to model situations where a subject has $n$ trials each of which independently can result in one of $k$ different values which occur with probabilities $(p_1, p_2, \ldots, p_k)$, where $p_1 + p_2 + \ldots + p_k$=1. The outcome of a multinomial is a list of $k$ numbers adding up to $n$, each of which represents the number of times a particular value was achieved.

For random variable X following the multinomial distribution, the outcome is the list of values $(x_1, x_2, \ldots, x_k)$ and the pmf is:

$$\Pr(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) = \left(\frac{n!}{x_1! \cdot x_2! \cdots x_k!}\right) p_1^{x_1}\, p_2^{x_2} \cdots p_k^{x_k}.$$

For example, consider a kind of candy that comes in an opaque bag and has three colors (red, blue, and green) in different amounts in each bag. If 30% of the bags have red as the most common color, 20% have green, and 50% have blue, then we could imagine an experiment consisting of opening $n$ randomly chosen bags and recording for each bag which color was most common. Here $k = 3$ and $p_1 = 0.30$, $p_2 = 0.20$, and $p_3 = 0.50$. The outcome is three numbers, e.g., $x_1$=number of times (out of 2) that red was most common, $x_2$=number of times blue is most common, and $x_3$=number of times green is most common. If we choose n=2, one calculation we can make is

$$\Pr(x_1 = 1, x_2 = 1, x_3 = 0) = \left(\frac{2!}{1! \cdot 1! \cdot 0!}\right) 0.30^1\, 0.20^1\, 0.50^0 = 0.12$$

and the whole pmf can be represented in this tabular form (where "# of Reds" means number of bags where red was most common, etc.):

| $x_1$ (# of Reds) | $x_2$ (# of Blues) | $x_3$ (# of Greens) | Probability |
|---|---|---|---|
| 2 | 0 | 0 | 0.09 |
| 0 | 2 | 0 | 0.04 |
| 0 | 0 | 2 | 0.25 |
| 1 | 1 | 0 | 0.12 |
| 1 | 0 | 1 | 0.30 |
| 0 | 1 | 1 | 0.20 |

Common notation is $X \sim \mathrm{MN}(n, p_1, \ldots, p_k)$.

### 3.9.3 Poisson distribution

The **Poisson distribution** is a discrete distribution whose support is the non-negative integers $(0, 1, 2, \ldots)$. Many measurements that represent counts which have no theoretical upper limit, such as the number of times a subject clicks on a moving target on a computer screen in one minute, follow a Poisson distribution. A Poisson distribution is applicable when the chance of a countable event is proportional to the time (or distance, etc.) available, when the chances of events in non-overlapping intervals is independent, and when the chance of two events in a very short interval is essentially zero.

A Poisson distribution has one parameter, usually represented as $\lambda$ (lambda). The pmf is:

$$\Pr(X = x) = f(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

The mean is $\lambda$ and the variance is also $\lambda$. From the pmf, you can see that the probability of no events, $\Pr(X = 0)$, equals $e^{-\lambda}$.

If the data show a substantially larger variance than the mean, then a Poisson distribution is not appropriate. A common alternative is the **negative binomial distribution** which has the same support, but has two parameters often denoted $p$ and $r$. The negative binomial distribution can be thought of as the number of trials until the $r^{th}$ success when the probability of success is $p$ for each trial.

It is sometimes useful to know that with large $\lambda$ a Poisson random variable approximates a Normal distribution with mean $\lambda$ and standard deviation $\sqrt{\lambda}$ (except that there are gaps in the Poisson because it only takes on whole numbers).

Common notation is $X \sim \text{Pois}(\lambda)$.

### 3.9.4 Gaussian distribution

The **Gaussian or Normal distribution** is a continuous distribution with a symmetric, bell-shaped pdf curve as shown in Figure 3.2. The members of this family are characterized by two parameters, the mean and the variance (or standard deviation) usually written as $\mu$ and $\sigma^2$ (or $\sigma$). The support is all of the real numbers, but the "tails" are very thin, so the probability that $X$ is more than 4 or 5 standard deviations from the mean is extremely small. The pdf of the Normal distribution

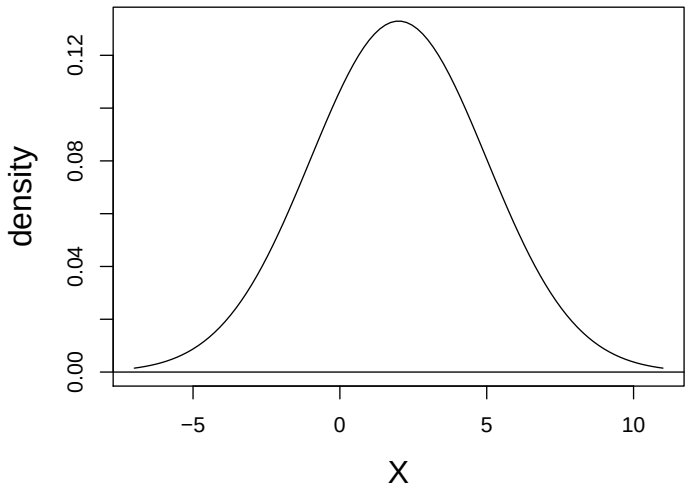


Figure 3.2: Gaussian bell-shaped probability density function

is:

$$f(x) = \frac{1}{\sqrt{2}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}.$$

Among the family of Normal distributions, the standard normal distribution, the one with $\mu = 0$ and $\sigma^2 = 1$ is special. It is the one for which you will find information about the probabilities of various intervals in textbooks. This is useful because the probability that the outcome will fall in, say, the interval from minus infinity to any arbitrary number $x$ for a non-standard normal distribution, say, $X$, with mean $\mu \neq 0$ and standard deviation $\sigma \neq 1$ is the same as the probability that the outcome of a standard normal random variable, usually called $Z$, will be less than $z = \frac{x-\mu}{\sigma}$, where the formula for $z$ is the "z-score" formula.

Of course, there is not really anything "normal" about the Normal distribution, so I always capitalize "Normal" or use Gaussian to remind you that we are just talking about a particular probability distribution, and not making any judgments about normal vs. abnormal. The Normal distribution is a very commonly used

distribution (see CLT, above). Also the Normal distribution is quite flexible in that the center and spread can be set to any values independently. On the other hand, every distribution that subjectively looks "bell-shaped" is not a Normal distribution. Some distributions are flatter than Normal, with "thin tails" (negative kurtosis). Some distributions are more "peaked" than a true Normal distribution and thus have "fatter tails" (called positive kurtosis). An example of this is the t-distribution (see below).

Common notation is $X \sim \mathrm{N}(\mu, \sigma^2)$.

### 3.9.5 t-distribution

The **t-distribution** is a continuous distribution with a symmetric, unimodal pdf centered at zero that has a single parameter called the "degrees of freedom" (df). In this context you can think of df as just an index or pointer which selects a single distribution out of a family of related distributions. For other ways to think about df see section 4.6. The support is all of the real numbers. The t-distributions have fatter tails than the normal distribution, but approach the shape of the normal distribution as the df increase. The t-distribution arises most commonly when evaluating how far a sample mean is from a population mean when the standard deviation of the sampling distribution is estimated from the data rather than known. It is the fact that the standard deviation is an estimate (i.e., a standard error) rather than the true value that causes the widening of the distribution from Normal to t.

Common notation is $X \sim \mathrm{t}_{df}$.

### 3.9.6 Chi-square distribution

A **chi-square distribution** is a continuous distribution with support on the positive real numbers whose family is indexed by a single "degrees of freedom" parameter. A chi-square distribution with df equal to $a$, commonly arises from the sum of squares of $a$ independent N(0,1) random variables. The mean is equal to the df and the variance is equal to twice the df.

Common notation is $X \sim \chi^2_{df}$.

### 3.9.7 F-distribution

The **F-distribution** is a continuous distribution with support on the positive real numbers. The family encompasses a large range of unimodal, asymmetric shapes determined by two parameters which are usually called numerator and denominator degrees of freedom. The F-distribution is very commonly used in analysis of experiments. If $X$ and $Y$ are two independent chi-square random variables with $r$ and $s$ df respectively, then $\frac{X/r}{Y/s}$ defines a new random variable that follows the F-distribution with $r$ and $s$ df. The mean is $\frac{s}{s-2}$ and the variance is a complicated function of $r$ and $s$.

Common notation is $X \sim \mathrm{F}(r, s)$.