Chapter 1

Probability

You're playing Garen, peacefully farming in the top lane when the enemy Yasuo decides to go in for a trade with you. You're both at half health, but it's not a problem– you're Garen! Executing pro-level mechanics, you hit E, spinning around your opponent, confident that you will this duel. With dismay, you watch as Yasuo's attacks are critical strikes, not once, twice, but three times in a row, giving him the kill just before you were about to kill him. All you needed was for one of his attacks to be not a critical strike, and you would have won– but that didn't happen. Contemplating the grey screen in front of you, you decide to check Yasuo's items. He had 40% crit chance. You wonder... what is the chance that I would have won that fight? In other words, what's the chance that Yasuo gets anything other than three critical strikes in a row?

In the field of statistics, we use a specific definition of chance called probability. Defining what exactly "probability" means turns out to be a tricky question, but the good news is that regardless of how you define it, probability has to follow a few common-sense rules. For this problem, it's helpful to think about probability as a long-run frequency (or percentage) of events.

1.1 Introduction

Before we start learning about the math behind probability, let's get down some basic concepts and terminology.

Probability deals with the outcomes of uncertain or random events. We call these random events **random variables**. In other areas of math, you've seen variables, usually called x, to stand for a number. Well, *random* variables stand for a number whose value is uncertain. We'll usually refer to these random variables using the capital letter X.

Here are some examples of some random variables:

- We flip a coin. The random variable X is the outcome of the coin flip, whether it lands Heads or Tails. If it lands Heads, we say $X = \text{Heads}$, and if it lands Tails, we say $X = \text{Tails}$.
- We don't know whether it will rain tomorrow, and if it does rain, how much. The random variable X is the amount of rain tomorrow. If it rains $\frac{1}{2}$ an inch, we say that $X = \frac{1}{2}$. If it doesn't rain at all, we say $X = 0$.
- We play a game of League. Sadly, we're not certain that we'll win, so whether we win or not can be thought of as a random variable X. If we win, $X = \text{We win}$, and if we lose, $X = \text{We lose}$.

We use probability because we want to talk about how likely it is that a random variable X ends up taking on different values: we want to know how probable it is that the coin will land Heads, or that there will be a heavy rain tomorrow, or that we'll win this game of League. We're going to use the capital letter P to write about the probability of a random variable taking on some value. That is,

$$P(X = x) = \text{Probability that the random variable } X \text{ is equal to the outcome } x.$$

Some examples:

- If X is the outcome of a normal coin toss, then $P(X = \text{Heads}) = \frac{1}{2}$.
- If X is the amount of rain tomorrow, and the weather channel says that there's a "20% chance of rain", then we think that $P(X = 0) = 0.8$.
- If you win 60% of the competitive League games that you play, and X refers to whether you win the next game, we might say that $P(X = \text{You win}) = 0.6$.

One more thing: We're often interested in whether a random variable X will end up having *one of several* values. For instance, what's the probability that a 6-sided die will land 1 *or* 2? To talk about situations like this, we need the concept of an **event**. An event describes several possible outcomes of a random variable, and the probability that the event happens is the probability that at least one of the outcomes described by the event happens. In the die example, we want to know about the event $E = \{\text{Die lands 1 or 2}\}$. To refer to the probability of events, we use:

$$P(E) = P(\text{One of the outcomes described by the event } E \text{ happens}).$$

In the next section, we'll begin talking about how to calculate the probabilities of events. As a final note, you should keep in mind that there are certain rules probabilities have to follow. These match our intuitions about how we should measure uncertainty. Here are the rules:

1. All probabilities are between 0 and 1.

2. For any event E,
$$P(E \text{ or not-}E) = 1.$$
By not-E, we mean that "none of the outcomes in E" happens. So Rule 2 says: We're sure that either the event E will happen, or it won't. After all, the random variable has to end up taking *some* value!

3. If two events A and B are incompatible, then
$$P(A \text{ or } B) = P(A) + P(B).$$
By *incompatible*, we mean that if one of the outcomes described by A happens, we know that *none* of the outcomes described by B happens. Here are some examples of incompatible events:

   - $A = \{\text{Die lands 1 or 2}\}$ and $B = \{\text{Die lands 3 or 4}\}$.
   - $A = \{\text{You win}\}$ and $B = \{\text{You lose}\}$.
   - $A = \{\text{It rains between 0 and 1 inch}\}$ and $B = \{\text{It rains between 3 and 10 inches}\}$.

Now let's see how to actually calculate some probabilities.

1.2 Counting

We can compute probabilities for both continuous values (such as the amount of rain that falls tomorrow) and discrete ones (the number of faces on a die in the next roll). But for the sake of simplicity, in this chapter we'll focus on probabilities where the random variable can only take on finitely many values.

Suppose our random variable X can only be one of N values. For instance, if X is the face of the die which appears on the next roll, there are $N = 6$ possible values. Now suppose we have an event E that describes some possible feature of the outcome of X. How do we compute the probability of the event E happening, $P(E)$?

You can probably guess. We just count up all the possibilities described by the event E, and divide this number by the total number of possibilities:

$$P(E) = \frac{\text{Number of cases described by event } E}{\text{Total number of values } X \text{ can take}}$$

If X is a die roll and E is the event that the die roll is even, this gives us

$$P(E) = \frac{\text{Number of possible even die rolls}}{\text{Total number of possible die rolls}} = \frac{3}{6} = \frac{1}{2}.$$

In this case, it was easy to count up all the different possibilities described the event E. But often it is more complicated. Think of some other events whose probability we might be interested in calculating:

- Seeing exactly 19 heads in 100 flips of a fair coin.
- Drawing a flush (all cards of the same suit) from a standard deck of cards.
- ...More interesting examples

Take the case of getting 19 heads in 100 coin flips. Using the same approach we took to compute the probability of a die landing on an even number, we could list all the possible sequences of 100 coin flips, and count how many of these show exactly 19 heads. Unfortunately, this probably isn't worth our time, since there is an *astronomical* number of possible sequences of 100 coin flips! It would be nice to learn some shortcuts that allow us to count in cases where we can't simply list all of the possibilities. And luckily, these shortcuts do exist.

In the next sections we'll look at shortcuts for two different kinds of counting: counting *with replacement* and counting *without replacement*.

1.2.1 Counting with replacement

Imagine drawing balls from a bag that contains 10 yellow balls and 10 red balls. What is the probability of getting two yellow balls on the first two draws? It depends!

If, after each time you draw a ball, you put it back in the bag, you'll have a $\frac{10}{20}$ chance of drawing a yellow each time, since there will always be 10 yellows and 10 reds in the bag. This an example of **counting with replacement** (since you replace the ball after each draw).

How do we compute probabilities in cases like this? We probably won't want to list out all of the possible draws of 10, and it turns out there is a formula that allows us to compute this much more easily. First, we need to count all possible sequences of 10 draws. Notice that, on each draw, there are two possibilities: Yellow or Red. Thus, on two draws, there are 2×2 possibilities; on three draws, $2 \times 2 \times 2$ possibilities; and so on. So if we draw 10 times with replacement, there are 2^{10} possible sequences of Y or R values.

Now that we know how many possible outcomes there are, we just need to count the number of outcomes in which exactly 4 yellows are drawn. Let's break this counting problem down into a few steps:

1. There are 10 possibilities for the location of the first yellow draw. Then, there are 9 possibilities for the location of the second yellow draw, since one of the locations was already taken for the first draw. Following the same reasoning, the are 8 possibilities for the third yellow draw, and 7 for the fourth. Putting this together, there are $10 \times 9 \times 8 \times 7$ ways to distribute four balls over 10 draws.

2. The mathematical shorthand for $10 \times 9 \times 8 \times \cdots \times 2 \times 1$ is written $10!$, pronounced "10 factorial". Notice that we can write $10 \times 9 \times 8 \times 7$ more succinctly as:

$$10 \times 9 \times 8 \times 7 = \frac{10 \times 9 \times \cdots \times 2 \times 1}{6 \times 5 \times \cdots \times 2 \times 1} = \frac{10!}{6!}.$$

3. But, this way of distributing the balls counts some sequences several times! In particular, there are $4 \times 3 \times 2 \times 1 = 4!$ double-countings. (More explanation.) In order to correct for this, and get the final number of sequences of 10 draws with exactly 4 yellows, we need to divide $\frac{10!}{6!}$ by 4!. This gives us a final answer of $\frac{10!}{6!4!}$ sequences of 10 draws-with-replacement that have exactly 4 yellows.

This formula is just one example of a very useful quantity called the **binomial coefficient**. Suppose we have n objects and we want to count the ways we can select just k of them. Then the binomial coefficient, written $\binom{n}{k}$ and pronounced "n choose k", tells us this number. The formula is:

$$\binom{n}{k} = \frac{n!}{(n-k)!\,k!}$$

Now that we have the total number of possible draws of 10 from this bag of colored balls, and the total number of possible draws with exactly 4 yellows, we can finally compute the probability we want:

$$P(\text{Get exactly 4 yellows in 10 draws}) = \frac{\binom{10}{4}}{2^{10}}$$

In fact, using the same reasoning, we can compute the probability of exactly k yellows in 10 draws for any k between 0 and 10:

$$P(\text{Get exactly } k \text{ yellows in 10 draws}) = \frac{\binom{10}{k}}{2^{10}}$$

By the way, this is our first example of a **probability mass function**, or **PMF**. A PMF is a function that tells us the probability of each of the possible outcomes that a discrete random quantity can take. (In this case, the random quantity is "number of yellow draws in 10 draws-with replacement from the bag".) Figure 1.1 plots the PMF this example; from the plot we can easily tell that getting 5 yellow draws has the highest probability. We'll gradually introduce other examples of PMFs, and discuss them directly in a later section. In the next section, we'll see how to calculate probabilities when we sample items without replacing them.

1.2.2 Counting without replacement

As we said before, the probabilities of different sequences of draws from our bag of balls depend on whether the balls are replaced each time we take one

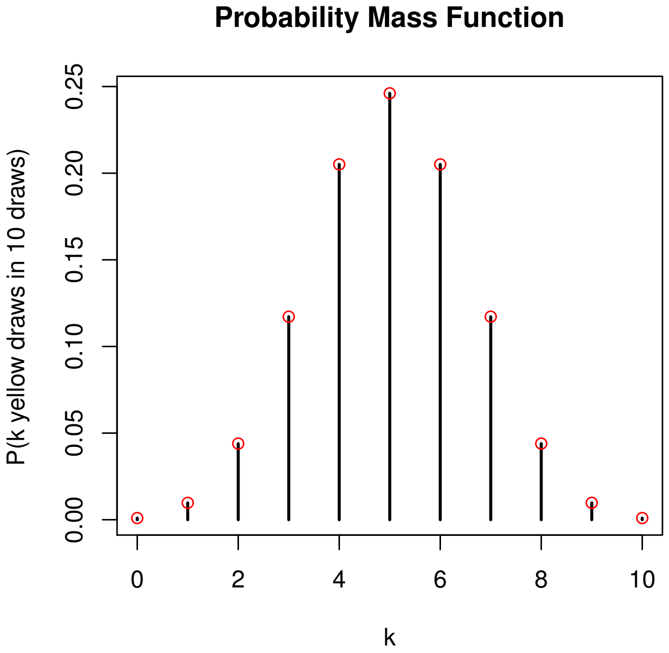


Figure 1.1: PMF for the number of yellow balls in a draw of 10 from our bag of colored balls. Horizontal axis is the number of yellow balls in 10 draws. The red points show the probabilities of getting k yellow balls, for every possible value of k.

out. Let's introduce **counting without replacement** by taking a look at what happens when the balls aren't replaced.

Suppose, to make the example slightly more complicated, that there are 20 balls in the bag – 10 yellow, 10 red – and we only draw 10, this time without replacement. What is the probability that we get exactly 4 yellows? Our job is a little easier now that we know how to count combinations using binomial coefficients. We can immediately see that there are $\binom{20}{10}$ possible sequences that we can draw, that is, ways of drawing 10 different balls from a total of 20.

How many possible outcomes are compatible with the event we're interested in, getting exactly 4 yellow? We can reframe this as the number of ways of getting 4 yellows from the 10 possible yellows in the bag, and 6 reds from the 10 possible reds in the bag. Again using binomial coefficients to count these possible draws, the number of sets of 10 balls with exactly 4 yellows turns out to be $\binom{10}{4} \times \binom{10}{6}$. That's because there are $\binom{10}{4}$ ways to choose 4 Yellows from the 10 Yellows in the bag, and $\binom{10}{6}$ ways for the remaining 6 balls in our draw to be taken from the 10 Reds in the bag.

We end up with the probability

$$P(\text{Get exactly 4 yellows in a draw of size 10, from a bag with 10/20 yellows}) = \frac{\binom{10}{4}\binom{10}{6}}{\binom{20}{10}}.$$

Using this reasoning, we can get a more general formula for the probability of k "successes" if we draw n objects, without replacement, from a group with N total objects and K total successes. (In our example, "successes" are yellow balls, and $N = 20$ balls total in the bag.) The formula is:

$$P(k \text{ successes}; N, K, n) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

This is our second example of a PMF. And the random quantity associated with this PMF has a name: it's a **hypergeometric** random variable, written HG(N, K, n). N, K, and n are called **parameters** of the distribution.

1.2.3 Application: Counting teams

In 5 vs. 5 League of Legends game, how many possible teams are there? There are 141 possible characters. And, since no character can be chosen twice, we know this is a straightforward use of the combination formula for choosing 10 unique objects from a set of 141:

$$\text{Number of possible 5v5 teams} = \binom{141}{10} = \frac{141!}{(141-10)!10!} \approx 6.2 \times 10^{14}.$$

Let's look at a slightly more challenging counting problem. Suppose there are 5 champions you know how to play very well, and suppose that you choose your character last. This means that 9 characters will be chosen by the other players, and according the default banning scheme (?), 10 characters will be banned. That means there will be 19 characters that you can't choose. Assume each character has an equal chance of being chosen or banned before it's your turn. What is the probability that you get to use one of the 5 characters you're good at?

One way to answer this question is to first compute the probability that you won't get to use *any* of your 5 good characters. Then, since we know from Section 1.1 that

$$P(\text{All 5 characters are taken}) + P(\text{At least one of your 5 characters is available}) = 1,$$

we'll be able to calculate the probability we're interested in.

Remembering what we learned in Section 1.2.2, we can recognize this as a problem involving the hypergeometric distribution. We have:

- $N = 141$ total characters;
- $K = 5$ "successes" (i.e. the 5 characters we're good at);
- $n = 19$ characters chosen from the 141 total without replacement,

giving

$$P(\text{All 5 characters are taken}) = \frac{\binom{5}{5}\binom{141-5}{19-5}}{\binom{141}{19}} \approx 2.7 \times 10^{-5},$$

where the "≈" symbol means "approximately equals". 2.7×10^{-5} is a very small number. So, you can be confident that you'll get to play one of your best characters.[1]

1.3 Bernoulli trials

So far, we've calculated probabilities by counting the number of outcomes that are compatible with the event we're interested in, and divided by the number of total possible outcomes. Underlying this method is the assumption that each possible outcome is *equally probable.* For instance, it works for calculating the probability of getting heads on the toss of a fair coin. But what do we do in the case of a *bent* coin, which we don't think will land

[1] Before you count on this calculation, however, remember that we assumed that each character had an equal chance of being chosen or banned. But this probably isn't the case, since we expect that some characters will be chosen or banned more often than others. (Specifics?) So if your 5 characters have a high chance of getting picked or banned before it's your turn to choose, you may have a significantly lower probability of getting to play one of your best characters than this calculation suggests!

heads 50% of the time? Enumerating possibilities no longer works. And of course, the case of the bent coin is conceptually similar to that of crit strikes, another situation where outcomes are not equiprobable.

To model situations like this, we introduce a very simple kind of random variable called a *Bernoulli trial*, which can only take two values: 0 and 1. We will write $X \sim \mathrm{Ber}(p)$, pronounced "X follows a Bernoulli distribution with parameter p". This means that the probability of X taking a value of 1 is p, and the probability of X taking a value of 0 is $1-p$. We can write this succinctly in the form of a PMF:

$$P(X=k) = p^k(1-p)^{1-k}, \text{ for } k \text{ equal to 0 or 1.}$$

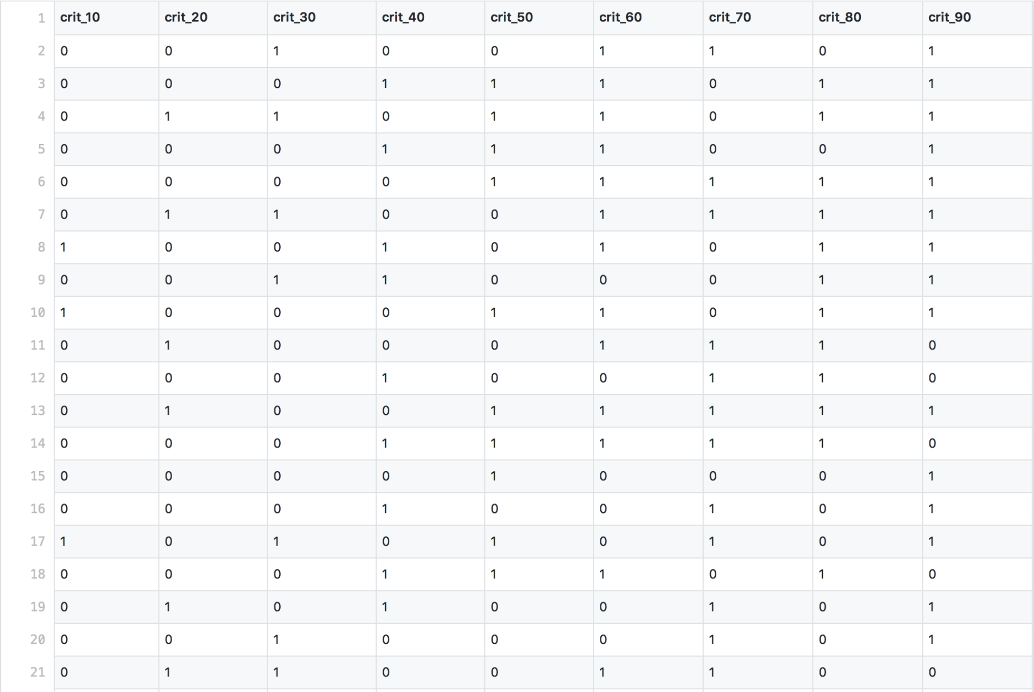
Here are some examples of quantities which can be modeled as Bernoulli trials:

- A coin landing heads ($p = 0.5$ if the coin is fair), coding Heads as 1 and Tails as 0
- A candidate being elected
- A crit strike happening
- You failing an exam
- You winning a lottery

1.3.1 Application: Bernoulli model for crit strikes

Figure 1.2 shows some data collected on crit strike occurrences. Each column contains data collected for each attack in a single game at a particular *nominal crit chance*, that is, the crit probability reported by the LoL interface (??). The 1's indicate a crit and 0's indicate not-crit, with the data points being collected in sequence.

What can we say about these Bernoulli random variables, given the data in Figure 1.2? One very simple model is to assume that the X_i's are **identically distributed** as $\mathrm{Ber}(p)$, where X_i is the random variable in the $i$the trial, and p is an unknown probability parameter. As an example, let our X_i's be crit events when the nominal crit chance is 40%, so p refers to the true probability of a crit when the nominal chance is 40%.



1	crit_10	crit_20	crit_30	crit_40	crit_50	crit_60	crit_70	crit_80	crit_90
2	0	0	1	0	0	1	1	0	1
3	0	0	0	1	1	1	0	1	1
4	0	1	1	0	1	1	0	1	1
5	0	0	0	1	1	1	0	0	1
6	0	0	0	0	1	1	1	1	1
7	0	1	1	0	0	1	1	1	1
8	1	0	0	1	0	1	0	1	1
9	0	0	1	1	0	0	0	1	1
10	1	0	0	0	1	1	0	1	1
11	0	1	0	0	0	1	1	1	0
12	0	0	0	1	0	0	1	1	0
13	0	1	0	0	1	1	1	1	1
14	0	0	0	1	1	1	1	1	0
15	0	0	0	0	1	0	0	0	1
16	0	0	0	1	0	0	1	0	1
17	1	0	1	0	1	0	1	0	1
18	0	0	0	1	1	1	0	1	0
19	0	1	0	1	0	0	1	0	1
20	0	0	1	0	0	0	1	0	1
21	0	1	1	0	0	1	1	0	0

Figure 1.2: Crit strike data. Each column corresponds to a sequence of attacks in a single game, under a fixed nominal crit chance. 1 indicates that the attack was a crit, and 0 indicates that it wasn't.

Now, we want to **estimate** the unknown parameter p using the data in Figure 1.2. This means that we want to make our best guess about p using just the observed data. Under our assumption that the each X_i has a Ber(p) distribution, a reasonable way to estimate p is to simply use the proportion of observed attacks in this column that were crits. There are 20 data points in column `crit-40`, and 9 of these are 1's. We will call our estimate $\hat{p}$, pronounced "p hat" – this "hat" symbol is a standard way of denoting an estimate. We end up with:

$$\begin{aligned}\widehat{p} &= \text{Estimated crit probability when nominal crit chance is } 40\% \\ &= \frac{9}{20} \\ &= 0.45.\end{aligned}$$

But our ultimate goal is to understand crit smoothing, which means we want to know about the probability of a crit strike *given what's already happened in the game*. To do this, we'll first need to learn about *conditional probability*, which we turn to in the next section.

1.4 Joint probability, conditional probability, and dependence

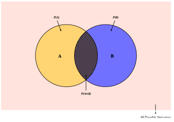
1.4.1 Joint and conditional probability

We often want to know the probability that some event happens, *given* some other information. For instance, what is the probability that it will rain tomorrow, given that the sky looks cloudy tonight? Intuitively, this probability will be different than the probability that it rains in the absence of information about cloudiness. We say there is some *dependence* between the events {Rains tomorrow} and {Cloudy tonight}.

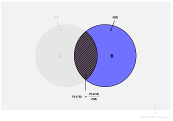
To formalize the idea of dependence, we first need to define *joint probability*, which is the probability of multiple events happening together. We will denote the joint probability of two events A and B using the $\cap$ symbol:

$$P(A \cap B) = \text{Probability that } A \text{ and } B \text{ both happen.}$$

We've actually already seen examples of joint probability. For example, in Section 1.2.1, we computed the probability that the first draw from the bag was yellow *and* the second draw was yellow. If A and B are events for a discrete random variable with equally probable outcomes, we can compute the joint probability of A and B simply by counting all the outcomes compatible with both A and B, using the techniques of Section 1.2.



(a) Joint probability



(b) Conditional probability

The notion of joint probability leads to a natural definition of *conditional* probability. FIGURE NAME illustrates the principles behind the definition. The two circles represent all the cases compatible with A and B, respectively.

Now, suppose we want to compute the probability that A happens, given that we already know that B has happened. First, we can get rid of all the worlds that are not compatible with B happening. Then, we can take the area of the overlap between the two events (which measures the probability of both A and B happening), and divide by the area of all the cases compatible with B.

The formal definition of the probability of A conditional on B, also pronounced "A given B", is:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}. \tag{1.1}$$

1.4.2 Dependence

Now that we've formalized the idea of conditional probability, we can formally introduce probabilistic dependence. As we saw earlier, an event A depends on an event B if its probability *given* B is different from its probability without conditioning on B. Formally, we say that A and B are **dependent** if

$$P(A \mid B) \neq P(A),$$

and **independent** if

$$P(A \mid B) = P(A).$$

Combining these definitions with Equation (1.1), we get a new formula for calculating joint probability:

$$P(A \cap B) = P(A \mid B)P(B) = P(B \mid A)P(A).$$

1.4.3 Binomial random variables

We can combine what we have learned about counting, Bernoulli trials, and (in)dependence to define another important distribution: the *binomial* distribution. Suppose we want to know the probabilities associated with not just a single Bernoulli trial, but multiple independent Bernoulli trials – for instance, multiple flips of a (possibly biased) coin.

In particular, we want to know the probability of getting k 1's in n independent Bernoulli trials. To model this situation, we say the random variables $X_1, \ldots, X_n$ are **independently and identically distributed (i.i.d)** as Ber(p) – abbreviated $X_1, \ldots, X_n \overset{i.i.d}{\sim}$ Ber(p). We have just seen what we mean by "independently distributed". By "identically distributed", we mean that each variable X_i has the same Bernoulli distribution, the Bernoulli distribution with probability parameter p.

Now, we want to know the distribution of the *count* of n i.i.d. Bernoulli random variables; since each X_i is equal to 0 or 1, the count of X_i's equal to 1 can be written as the sum $\sum_{i=1}^{n} X_i$.[2] We'll denote the random variable of interest as $S = \sum_{i=1}^{n} X_i$. To derive a formula for $P(S = k)$, let's first consider the probability that the first k X_i's are equal to 1, and the rest are equal to 0. Notice that this is a joint probability, since we want to know the probability that X_1 *and* X_2 *and* $\ldots X_n$ each take on certain values together. But since we assumed that the X_i's are independent, this joint probability is easy to calculate; it's just the product of the individual probabilities

$$\begin{aligned}
&P(\{\text{First } k \text{ values are } 1\}) \times P(\{\text{Next } n-k \text{ values are } 0\}) \\
=&P(X_1 = 1) \times \cdots \times P(X_k = 1) \times P(X_{k+1} = 0) \times \cdots \times P(X_n = 0) \\
=&p^k(1-p)^{n-k}.
\end{aligned}$$

However, we don't want the probability that the *first* k values are equal to 1. We want to know the probability of exactly k values equalling 1, with no restriction on where the 1's fall in the sequence of Bernoulli trials. From Section 1.2 we know how to count the number of n-length sequences with exactly k 1's: it's $\binom{n}{k}$. And since each of these sequences will have the same probability – remember that the X_i's all have the same probability parameter – we can get the probability of exactly k 1's by multiplying the number of sequences with exactly k 1's by our expression above, which gives the probability of a *particular* sequence with exactly k 1's. That is,

$$P(S = k) = \binom{n}{k} p^k (1-p)^{n-k}, \text{ for values of } k \text{ between 0 and 10.} \tag{1.2}$$

We say that S is a **binomial random variable** with parameters n and p, or $S \sim \text{Bin}(n, p)$. Equation (1.2) gives its PMF.

[2] If you haven't seen this "summation notation" before, $\sum_{i=1}^{n} X_i$ is the same as $X_1 + X_2 + \cdots + X_n$.

Application: Binomial model for crit strikes

Remember the scenario we saw at the beginning of the chapter – your enemy hits you 3 times at a nominal crit chance of 40%, and crits all 3 times. What is the probability that exactly 3 of 3 attacks crit?

We now have the tools to make a first pass at an answer. Call the outcomes of the three attacks X_1, X_2, X_3. First, let's use the estimate of p we computed earlier, $\widehat{p} = 0.45$. Therefore we assume that the X_1, X_2, X_3 are distributed as $\mathrm{Ber}(0.45)$. Second, let's assume that the crit events are independent. Then the probability distribution of the number of successful crits $S = X_1 + X_2 + X_3$ follows a $\mathrm{Bin}(3, 0.45)$ distribution, and the probability of the observed outcome is given by

$$P(S = 3) = \binom{3}{3} 0.45^3 0.55^0 \approx 0.09.$$

So, if our model is right, you got unlucky, but not *wildly* unlucky – we'd see this outcome in a little under 10% of these situations.

However, we do have reason to believe our model is wrong. In particular, we don't think that crit events are independent, and so S won't really have a binomial distribution. Using what we know about dependence, we can consider more realistic models of crit strike probability.

Application: Markov model for crit strike

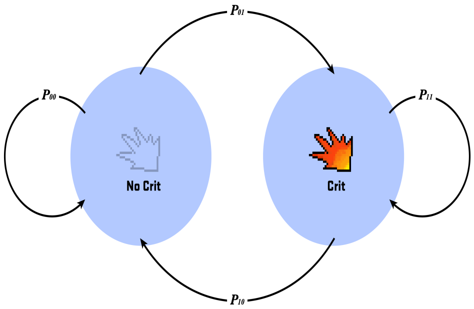


Figure 1.3: A simple Markov chain.

In contrast to the binomial model, a **Markov chain** is a model in which the outcomes of discrete events are dependent. In particular, if random variables $X_1, X_2, X_3, \dots$ are described by a Markov chain, then the outcome at time n depends (only) on the outcome at time $n-1$:

$$P(X_n = k \mid X_{n-1}, \dots, X_1) = P(X_n = k \mid X_{n-1}). \tag{1.3}$$

The probabilities referenced in Equation (1.3) are referred to as *transition probabilities*, since they tell us the probability of the Markov chain transitioning from a given value at time $n-1$ to a given value at time n.

Since we suspect that the chance of a crit strike depends on whether recent attacks have been crit strikes, let's model crit chances using a Markov chain with these unknown transition probabilities:

$$\begin{aligned}
p_{11} &= P(\text{Crit at attack } n \mid \text{Crit at attack } n-1)\\
p_{01} &= P(\text{Crit at attack } n \mid \text{No crit at attack } n-1)\\
p_{10} &= P(\text{No crit at attack } n \mid \text{Crit at attack } n-1)\\
p_{00} &= P(\text{No crit at attack } n \mid \text{No crit at attack } n-1)
\end{aligned}$$

Using a method similar to the one we used in Section 1.3.1, let's estimate these transition probabilities using the data in Figure 1.2. For simplicity, we'll again just consider the column `crit-40`. Let's estimate the probability of crit at attack n given crit at attack $n-1$, denoted $\widehat{p}_{11}$,

$$\widehat{p}_{11} = \frac{\text{Number of two consecutive crits}}{\text{Number of consecutive attack pairs where the first attack is a crit}} = \frac{2}{9}.$$

Using the same spirit, we can estimate the other transition probabilities. The summary of the calculation we use to get our estimates is in the so-called *transition matrix* displayed here:

$$\begin{pmatrix} \widehat{p}_{00} & \widehat{p}_{01} \\ \widehat{p}_{10} & \widehat{p}_{11} \end{pmatrix} = \begin{pmatrix} 7/10 & 3/10 \\ 7/9 & 2/9 \end{pmatrix}.$$

Given these estimates, we can estimate the probability of three crit strikes in a row, now using our Markov model instead of the binomial model from the previous example. So that we don't have to worry about how X_1 depends on previous attacks, let's first assume that X_1 is the crit outcome of the first attack in the game, and its nominal crit chance is 40%. Let's also assume that the (unconditional) nominal crit chance is correct, so $P(X_1 = 1) = 0.4$.

Then, our Markov estimate of the probability of three crits in a row at a nominal level of 40% is given by:

$$\begin{aligned}
&P(X_1 = 1 \cap X_2 = 1 \cap X_3 = 1) \\
&= P(X_1 = 1) \times P(X_2 = 1 \mid X_1 = 1) \times P(X_3 = 1 \mid X_2 = 1) \\
&= 0.4 \times \widehat{p}_{11} \times \widehat{p}_{11} \\
&= 0.4 \times \frac{2}{9} \times \frac{2}{9} \\
&\approx 0.05.
\end{aligned}$$

So under our Markov model, our estimated probability of seeing three crits in a row is somewhat lower (0.05) than our estimate under the binomial model (0.09).

1.5 Distribution of a discrete random variable

We've already seen several examples of probability mass functions (PMFs). They were

- Hypergeometric: If $X \sim \mathrm{HG}(N, K, n)$, then $P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$, for k between 0 and n.
- Bernoulli: If $X \sim \mathrm{Ber}(p)$, then $P(X = k) = p^k(1-p)^{1-k}$, for k equal to 0 or 1.
- Binomial: If $X \sim \mathrm{Bin}(n, p)$, then $P(X = k) = \binom{n}{k}p^k(1-p)^{n-k}$, for $k = 0, 1, \cdots, n$.

Now, we can give formal rules that a function f has to follow for it to qualify as a PMF:

- Its values are greater than or equal to 0.
- It takes on non-zero values for a finite or countably infinite set of outcomes. (Intuition for countability.)
- The sum of all values of f is 1.

(More discussion.)

Another important way of describing the probabilities associated with a random variable X is the **cumulative distribution function (CDF)** of X. Like the name suggests, the CDF at a point gives the probability that X takes on a value less than or equal to that point:

$$F(k) = \sum_{i=0}^{k} P(X = i).$$

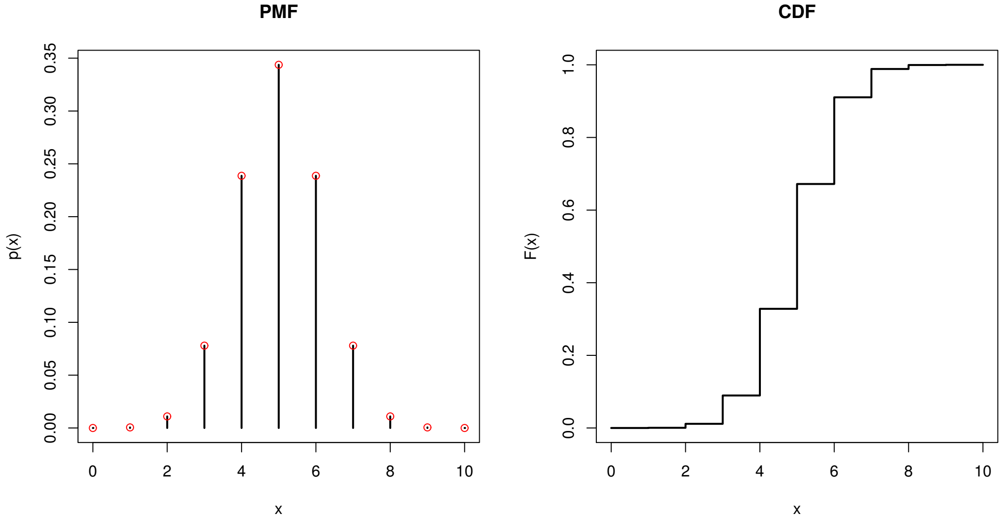


Figure 1.4: PMF and CDF for HG$(20, 10, 10)$ distribution

1.6 Application: Likelihood of the nominal crit hypothesis

Let's return to the problem of comparing the nominal crit chance to the observed crit data. In Section 1.3.1 we assumed that crit events at the nominal crit chance of 40% were identically distributed as Ber(p), and using the partial data shown we already got the estimated p to be 0.45.

Now let's look at a slightly different question: What is the true crit chance, *given that* the previous attack was a crit? For the n^{th} attack, write

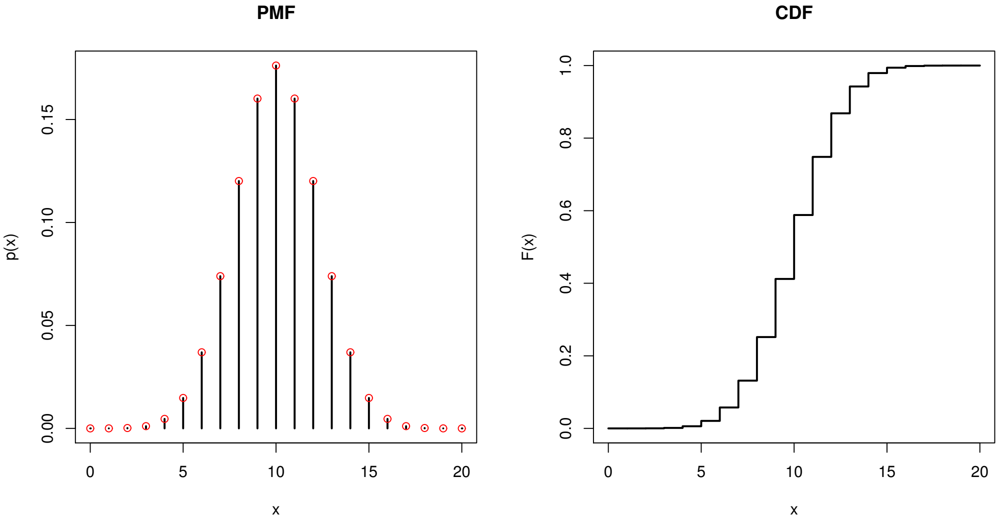


Figure 1.5: PMF and CDF for $\mathrm{Bin}(20, 0.5)$ distribution

this event as $(X_n = 1 \mid X_{n-1} = 1)$. This is a question about the conditional probability $P(X_n = 1 \mid X_{n-1} = 1)$. In particular, we want to know whether the nominal crit chance is accurate for this conditional probability.

Let's assume, as we did in Section 1.4.3, that the sequence of X_n's is a Markov chain. This means that we are only concerned about a single transition probability, since for any n, $p_{11} = P(X_n = 1 \mid X_{n-1} = 1)$. And, we already computed an estimate: $\hat{p}_{11} = \frac{2}{9}$. Now we want to know: Under the assumption that crit events are a Markov chain, how strong is the evidence against the claim that the conditional probability of a crit given that the previous attack was a crit is equal to the nominal crit chance of 0.4?

To formalize the problem, let's introduce two possible hypotheses (which both assume that $X_1, X_2, \ldots$ is a Markov chain):

- $H_0 : p_{11} = 0.4$
- $H_1 : p_{11} = \widehat{p}_{11} = \frac{2}{9}$

We want to know how well H_0 and H_1 fit the data. One way to measure how well a probability model fits the data is called the **likelihood**[3]. Let H be a hypothesis about a probability distribution. Then, the likelihood of H

[3] This can be a little confusing at first, because in everyday language we use "probabil-

tells us the probability of seeing the data that we have, *given* that H is the true model. We write it like this:

$$\mathcal{L}(H \mid X) = P(X \text{ when } H \text{ is true}),$$

pronounced "the likelihood of H given X".

Intuitively, if a model describes the actually probabilities with which a random quantity X takes different values, then the data we observe will tend to have higher probability under this model than under other (false) models. So the likelihood is a reasonable way of comparing how well the different models we're considering fit our data. For example,

- Examples of likelihood-based reasoning?

Returning to the crit chance problem, let's start by computing the likelihood of H_0. In this case our data is the observation that, among the 10 attacks which were followed by crit strikes, 2 of these were also crit strikes. And, because of our Markov assumption, we know that the event $(X_n \mid X_{n-1} = 1)$ is independent of all the previous crit strikes, since (by assumption) the only event that matters to the probability that a crit occurs at attack n is whether there was a crit at $n-1$. This means that we can re-write our models in terms of i.i.d Bernoulli random variables:

- $H_0 : (X_n \mid X_{n-1} = 1) \overset{i.i.d}{\sim} \text{Ber}(0.4)$
- $H_1 : (X_n \mid X_{n-1} = 1) \overset{i.i.d}{\sim} \text{Ber}(\frac{2}{9})$

Finally, notice that our data consist of *counts* of these i.i.d Bernoulli random variables. That means they follow a binomial distribution, and we can use what we know about this distribution to compute the likelihoods! As we did before, write the sum of these Bernoullis as S. Remember we use 20 rows of data in column `crit-40`, and notice there are 9 pairs of attacks which have the crit first. So after conditioning on $X_{n-1} = 1$, we only need to consider those 9 pairs of data. Each model gives a different distribution for S:

- $H_0 : S \sim \text{Bin}(9, 0.4)$
- $H_1 : S \sim \text{Bin}(9, \frac{2}{9})$

ity" and "likelihood" interchangeably. But in statistics, although likelihood is related to probability, they aren't synonyms: likelihood has its own technical definition.

We observed that $S = 2$, so the likelihoods turn out to be:

$$\begin{aligned}
\mathcal{L}(H_0 \mid \text{Crit data}) = P(S = 2 \mid p_{11} = 0.4) \quad &= \binom{9}{2} 0.4^2 (1 - 0.4)^7 \approx 0.16 \\
\mathcal{L}(H_1 \mid \text{Crit data}) = P(S = 2 \mid p_{11} = 2/9) \quad &= \binom{9}{2} (\frac{2}{9})^2 (1 - \frac{2}{9})^7 \approx 0.31.
\end{aligned}$$

Given these likelihoods, a natural way to compare how well the models fit the data is their **likelihood ratio**, which is, well, exactly what is sounds like:

$$\text{Likelihood rato of } H_0 \text{ and } H_1 = \frac{\mathcal{L}(H_1 \mid \text{Crit data})}{\mathcal{L}(H_0 \mid \text{Crit data})} = \frac{0.31}{0.16} \approx 2.$$

So, our best-fitting model in terms of likelihood - H_1 – gives about 2 times much probability to the observed data as does the model which assumes that the nominal crit chance gives the correct (conditional) crit probability.

1.7 Expectation, variance, and correlation

We've seen some uses of the probability distribution of a random variable. We often also want to have summaries of the distribution of a random variable. Some of the most important questions that we'd like to be able to answer include:

- What is the average value of the random variable?
- How spread out is the probability of the random variable?
- How strongly related are two different random variables?

Remember that in Chapter 1, we introduced an estimate of the "center" of a dataset, called the sample mean; an estimate of the "spread" of a dataset, called the sample variance; and an estimate of how strongly related two variables are, called the sample correlation. Because each of these measures is computed from a *sample* of random variables, we call them sample statistics.

In this chapter, we'll look at *population-level* versions of each of these quantities. You can think of these as the sample statistics we would get if we calculated them for an infinite number of draws from our probability distribution.

1.7.1 Expectation

The **expectation** of a random variable is a certain kind of average. For the rest of this section, we'll consider a discrete random variable X, taking values $0, 1, 2, 3, \ldots, N$. The expectation of X is defined as

$$\mathbb{E}\big[X\big] = \sum_{i=0}^{N} i \cdot P(X = i).$$

For example, what is the expectation of X when X is a Ber(0.5) random variable? Using the formula above, we get:

$$\begin{aligned}\mathbb{E}\big[X\big] &= \sum_{i=0}^{1} i \cdot P(X = i)\\ &= 0 \cdot P(X = 0) + 1 \cdot P(X = 1)\\ &= 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2}\\ &= \frac{1}{2}.\end{aligned}$$

We can also take the expectations of *functions* of a random variable. If g is a function, then $g(X)$ is also a random variable, that simply takes the value $g(x)$ whenever X takes the value x. For instance, we often want to measure how good the different possible values of a random variable are. Suppose we play a (not very fun) game where I win \$1 if a fair coin toss lands heads, and I lose \$2 if the coin lands tails. Then we can define a function g to record how good or bad each of these outcomes is:

$$\begin{aligned}g(0) &= -2\\ g(1) &= 1\end{aligned}$$

and the outcome X of the coin toss is again a Ber(0.5) random variable. Then we can compute how much we expect to win from this game:

$$\begin{aligned}
\mathbb{E}\big[g(X)\big] &= \sum_{i=0}^{1} g(i) \cdot P(X = i) \\
&= g(0) \cdot P(X = 0) + g(1) \cdot P(X = 1) \\
&= -2 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} \\
&= -\frac{1}{2}.
\end{aligned}$$

Remember we said that we can think of expectation as the sample mean calculated for on infinite number of data points from the same probability distribution. So, we should expect our sample mean to get closer and closer to the expectation as it's computed for more and more draws from the same distribution. And this is, in fact, what happens. Take a look at Figure **??**. This plots the sample means of VARIABLE, taken for larger and larger samples of games (?). We can see the sample mean is converging to a fixed number as the sample sizes increases. While we don't know the true expectation of this distribution, the law of large numbers assures us that this number is close to the true expectation. This is what we mean when we say "the expectation can be thought of as the sample mean taken for an infinitely large data set", and it's one reason that the sample mean is a reasonable estimate for the expectation.

1.7.2 Variance

Now that we have defined expectation, we can use this to answer our next question: How spread out is the probability of a random variable? To get a sense for what we mean, look at Figure **??**, which shows two different probability mass functions. The PMF on the left is clearly more spread out more than the one on the right, and so we would expect instances of the variable on the left to take on a wider random of values than the one on the right. To formalize what's different about these random variables, let's introduce **variance**:

$$\begin{aligned}
V(X) &= \sum_{i=0}^{N} (i - \mathbb{E}\big[X\big])^2 \cdot P(X = i) \\
&= \mathbb{E}\big\{\big[X - \mathbb{E}(X)\big]^2\big\}.
\end{aligned}$$

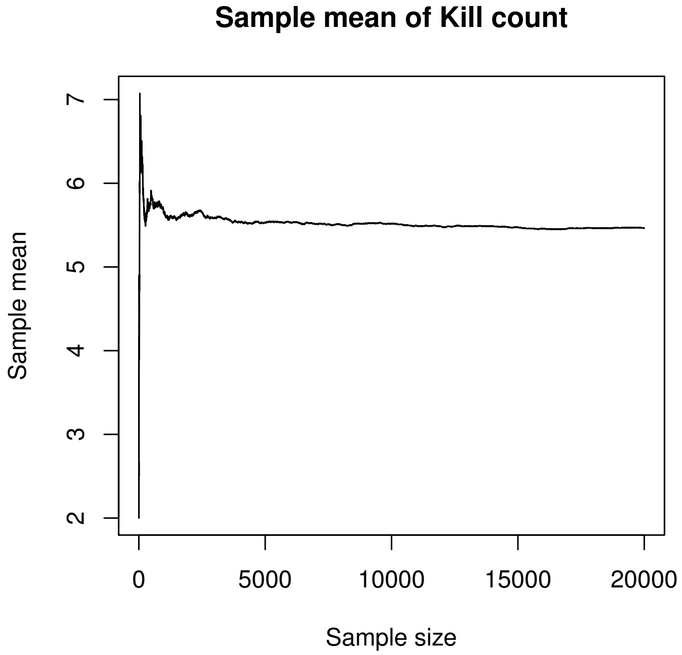


Figure 1.6: Law of large numbers for the mean kill count.

To make sense of this formula, think about the meaning of $[X - \mathbb{E}(X)]^2$. This is a measure of (squared) distance from the outcome taken by X to the expectation of X, $\mathbb{E}(X)$. Since the variance is the average value (expectation) of this squared distance, we can think of the variance as a measure of how far away the random variable X is from its expectation, on average.

1.7.3 Covariance and correlation

The last question we wanted to answer in this section was: How strongly related are two random variables? We've already looked at what it means for random quantities to be dependent, but now we want a measure of just *how* dependent they are. For instance, take a look at Figure 1.8. Intuitively, the quantities in on the x and y axes on the left figure are less strongly related than the two quantities on the right. Now we want to introduce one formal measure of the dependence between two random variables.

Define the **covariance** of X and Y as:

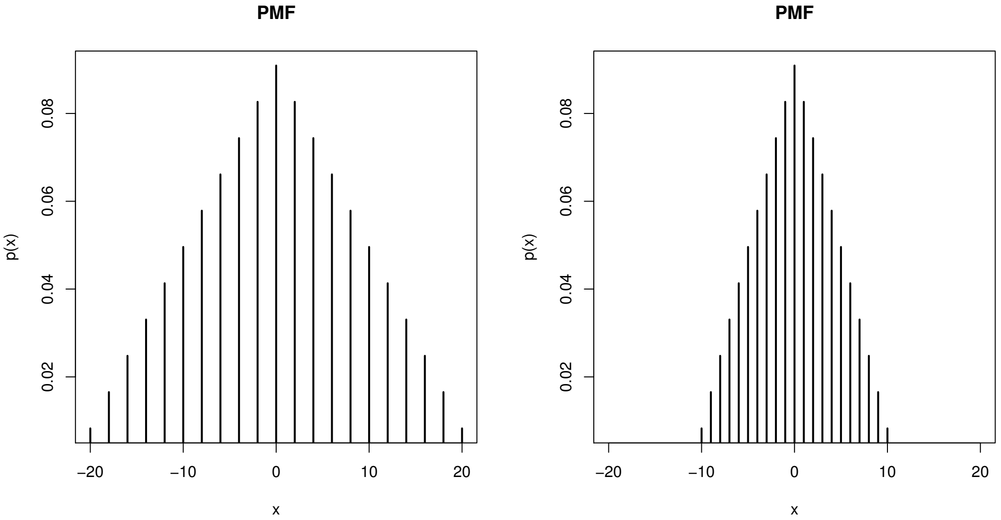
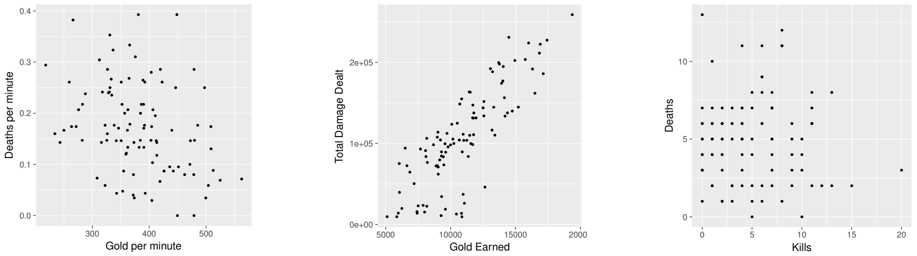


Figure 1.7: The PMFs of two discrete random variables. The variance of the random variable on the left is 80. The variance of the random variable on the right is 20.



(a) Correlation= −0.32 (b) Correlation= 0.82 (c) Correlation= −0.03

Figure 1.8: Correlation plots

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

The **correlation** between X and Y is a standardized version of the covariance:

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{V(X) \cdot V(Y)}}$$

While the covariance can be any number, the correlation must be between -1 and 1. Figure 1.8 shows datasets with moderate negative correlation; correlation that's close to 0; and strong positive correlation, to illustrate the kinds of relationships indicated by different values for correlation. (League of Legends commentary?)

As a last note, keep in mind that covariance and correlation measure the strength of the *linear* dependence between two random variables. It's possible that there is a strong relationship between two quantities, but that this relationship is more complicated than a linear one.

1.8 Conclusion