What Is Statistics?

Some Definitions of Statistics

This is a course primarily about statistics, but what exactly is *statistics*? In other words, what is this course about?[1] Here are some definitions of statistics from other people:

- a collection of procedures and principles for gaining information in order to make decisions when faced with uncertainty (J. Utts [**Utt05**]),
- a way of taming uncertainty, of turning raw data into arguments that can resolve profound questions (T. Amabile [**fMA89**]),
- the science of drawing conclusions from data with the aid of the mathematics of probability (S. Garfunkel [**fMA86**]),
- the explanation of variation in the context of what remains unexplained (D. Kaplan [**Kap09**]),
- the mathematics of the collection, organization, and interpretation of numerical data, especially the analysis of a population's characteristics by inference from sampling (American Heritage Dictionary [**AmH82**]).

While not exactly the same, these definitions highlight four key elements of statistics.

Data – the raw material

Data are the raw material for doing statistics. We will learn more about different types of data, how to collect data, and how to summarize data as we go along. This will be the primary focus of Chapter 1.

[1] As we will see, the words *statistic* and *statistics* get used in more than one way. More on that later.

Information – the goal

The goal of doing statistics is to gain some information or to make a decision. Statistics is useful because it helps us answer questions like the following:

- Which of two treatment plans leads to the best clinical outcomes?
- How strong is an I-beam constructed according to a particular design?
- Is my cereal company complying with regulations about the amount of cereal in its cereal boxes?

In this sense, statistics is a science – a method for obtaining new knowledge.

Uncertainty – the context

The tricky thing about statistics is the uncertainty involved. If we measure one box of cereal, how do we know that all the others are similarly filled? If every box of cereal were identical and every measurement perfectly exact, then one measurement would suffice. But the boxes may differ from one another, and even if we measure the same box multiple times, we may get different answers to the question *How much cereal is in the box?*

So we need to answer questions like *How many boxes should we measure?* and *How many times should we measure each box?* Even so, there is no answer to these questions that will give us absolute certainty. So we need to answer questions like *How sure do we need to be?*

Probability – the tool

In order to answer a question like *How sure do we need to be?*, we need some way of measuring our level of certainty. This is where mathematics enters into statistics. Probability is the area of mathematics that deals with reasoning about uncertainty. So before we can answer the statistical questions we just listed, we must first develop some skill in probability. Chapter 2 provides the foundation that we need.

Once we have developed the necessary tools to deal with uncertainty, we will be able to give good answers to our statistical questions. But before we do that, let's take a bird's eye view of the processes involved in a statistical study. We'll come back and fill in the details later.

A First Example: The Lady Tasting Tea

There is a famous story about a lady who claimed that tea with milk tasted different depending on whether the milk was added to the tea or the tea added to the milk. The story is famous because of the setting in which she made this claim. She was attending a party in Cambridge, England, in the 1920s. Also in attendance were a number of university dons and their wives. The scientists in attendance scoffed at the woman and her claim. What, after all, could be the difference?

All the scientists but one, that is. Rather than simply dismiss the woman's claim, he proposed that they decide how one should *test* the claim. The tenor of

the conversation changed at this suggestion, and the scientists began to discuss how the claim should be tested. Within a few minutes cups of tea with milk had been prepared and presented to the woman for tasting.

Let's take this simple example as a prototype for a statistical study. What steps are involved?

(1) Determine the question of interest.

Just what is it we want to know? It may take some effort to make a vague idea precise. The precise questions may not exactly correspond to our vague questions, and the very exercise of stating the question precisely may modify our question. Sometimes we cannot come up with any way to answer the question we really want to answer, so we have to live with some other question that is not exactly what we wanted but is something we can study and will (we hope) give us some information about our original question.

In our example this question seems fairly easy to state: Can the lady tell the difference between the two tea preparations? But we need to refine this question. For example, are we asking if she *always* correctly identifies cups of tea or merely if she does better than we could do ourselves (by guessing)?

(2) Determine the population.

Just who or what do we want to know about? Are we only interested in this one woman or women in general or only women who claim to be able to distinguish tea preparations?

(3) Select measurements.

We are going to need some data. We get our data by making some measurements. These might be physical measurements with some device (like a ruler or a scale). But there are other sorts of measurements too, like the answer to a question on a form. Sometimes it is tricky to figure out just what to measure. (How do we measure happiness or intelligence, for example?) Just how we do our measuring will have important consequences for the subsequent statistical analysis.

In our example, a measurement may consist of recording for a given cup of tea whether the woman's claim is correct or incorrect.

(4) Determine the sample.

Usually we cannot measure every individual in our population; we have to select some to measure. But how many and which ones? These are important questions that must be answered. Generally speaking, bigger is better, but it is also more expensive. Moreover, no size is large enough if the sample is selected inappropriately.

Suppose we gave the lady one cup of tea. If she correctly identifies the mixing procedure, will we be convinced of her claim? She might just be guessing; so we should probably have her taste more than one cup. Will we be convinced if she correctly identifies 5 cups? 10 cups? 50 cups?

What if she makes a mistake? If we present her with 10 cups and she correctly identifies 9 of the 10, what will we conclude? A success rate of 90% is, it seems, much better than just guessing, and anyone can make a mistake now and then. But what if she correctly identifies 8 out of 10? 80 out of 100?

And how should we prepare the cups? Should we make 5 each way? Does it matter if we tell the woman that there are 5 prepared each way? Should we flip a coin to decide even if that means we might end up with 3 prepared one way and 7 the other way? Do any of these differences matter?

(5) Make and record the measurements.

Once we have the design figured out, we have to do the legwork of data collection. This can be a time-consuming and tedious process. In the case of the lady tasting tea, the scientists decided to present her with ten cups of tea which were quickly prepared. A study of public opinion may require many thousands of phone calls or personal interviews. In a laboratory setting, each measurement might be the result of a carefully performed laboratory experiment.

(6) Organize the data.

Once the data have been collected, it is often necessary or useful to organize them. Data are typically stored in spreadsheets or in other formats that are convenient for processing with statistical packages. Very large data sets are often stored in databases.

Part of the organization of the data may involve producing graphical and numerical summaries of the data. We will discuss some of the most important of these kinds of summaries in Chapter 1. These summaries may give us initial insights into our questions or help us detect errors that may have occurred to this point.

(7) Draw conclusions from data.

Once the data have been collected, organized, and analyzed, we need to reach a conclusion. Do we believe the woman's claim? Or do we think she is merely guessing? How sure are we that this conclusion is correct?

Eventually we will learn a number of important and frequently used methods for drawing inferences from data. More importantly, we will learn the basic framework used for such procedures so that it should become easier and easier to learn new procedures as we become familiar with the framework.

(8) Produce a report.

Typically the results of a statistical study are reported in some manner. This may be as a refereed article in an academic journal, as an internal report to a company, or as a solution to a problem on a homework assignment. These reports may themselves be further distilled into press releases, newspaper articles, advertisements, and the like. The mark of a good report is that it provides the essential information about each of the steps of the study.

As we go along, we will learn some of the standard terminology and procedures that you are likely to see in basic statistical reports and will gain a framework for learning more.

At this point, you may be wondering who the innovative scientist was and what the results of the experiment were. The scientist was R. A. Fisher, who first described this situation as a pedagogical example in his 1925 book on statistical methodology [**Fis25**]. We'll return to this example in Sections 2.4.1 and 2.7.3.

Consider a model where the penetrances are

- $P(D|AA) = 0.01$,
- $P(D|Aa) = 0.05$,
- $P(D|aa) = 0.50$.

Q. Now consider an $AA \times aa$ cross. What is the probability that a child will have the disease?

A. In this case we know the child will have genotype Aa, so $\mathrm{P}(D) = \mathrm{P}(D \mid Aa) = 0.05$.

Q. What is the probability that a child will have the disease in an $AA \times Aa$ cross?

A. We can divide the event D into three mutually exclusive cases and sum their probabilities:

$$\begin{aligned} \mathrm{P}(D) &= \mathrm{P}(D \text{ and } AA) + \mathrm{P}(D \text{ and } Aa) + \mathrm{P}(D \text{ and } aa) \\ &= \mathrm{P}(AA) \cdot \mathrm{P}(D \mid AA) + \mathrm{P}(Aa) \cdot \mathrm{P}(D \mid Aa) + \mathrm{P}(aa) \cdot \mathrm{P}(D \mid aa) \\ &= (0.5)(0.01) + (0.5)(0.05) + (0)(0.5) = 0.03 \,. \end{aligned}$$

◁

2.3. Discrete Distributions

2.3.1. The Distribution of a Discrete Random Variable: pmfs and cdfs

Recall that the distribution of a variable in a data set described what values occurred and with what frequency. What we need now is a way to describe the distribution of a random variable. We do this somewhat differently for discrete and continuous random variables, so for the moment we will focus our attention on the discrete case.

One useful way to describe the distribution of a discrete random variable – especially one that has a finite range – is in a table like the one we used in Example 2.2.1 for the random variable X that counts the number of heads in four tosses of a fair coin:

value of X	0	1	2	3	4
probability	0.0625	0.2500	0.3750	0.2500	0.0625

Notice that the probability table allows us to assign to each possible outcome a probability. This means that the table is really describing a function. This function is called a **probability mass function** or **pmf**. A pmf can be any function that obeys the probability axioms.

Definition 2.3.1. Let $X : S \to \mathbb{R}$ be a random variable. The *probability mass function* (pmf) for X is a function $f : \mathbb{R} \to [0, 1]$ such that for all $x \in \mathbb{R}$,

$$f(x) = \mathrm{P}(X = x) \,.$$

□

We will write f_X for the pmf of X when we want to emphasize the random variable.

Lemma 2.3.2. *Let f be the pmf for a random variable $X : S \to \mathbb{R}$. Then*

(1) $f(x) \in [0,1]$ *for all* $x \in \mathbb{R}$, *and*

(2) $\displaystyle\sum_{s \in S} f(X(s)) = 1.$

Furthermore if g is a function such that

(1) $g(x) \in [0,1]$ *for all* $x \in \mathbb{R}$, *and*

(2) $\displaystyle\sum_{g(x) \neq 0} g(x) = 1 \in [0,1]$ *for all* $x \in \mathbb{R}$,

then g is a pmf for a random variable.

Proof. Exercise 2.38. □

As we will see, the pmf will be much more important to our study of random variables than the set S or the particular function mapping S to $\mathbb{R}$.

It is possible to write down an explicit formula for the pmf of the random variable in Example 2.2.1, namely

$$f_X(x) = \begin{cases} \frac{4!}{16x!(4-x)!} & \text{if } x \in \{0,1,2,3,4\}, \\ 0 & \text{otherwise.} \end{cases}$$

You can easily check that the values given by this formula match those in the table above. Sometimes it is not so easy to write down a formula for the pmf of a random variable, but the function exists nonetheless, since we can always define it by

$$f_X(x) = \mathrm{P}(X = x)\,. \tag{2.4}$$

Figure 2.6 shows three different ways that we can plot the pmf from Example 2.2.1. In the first plot, the non-zero values of this function are represented by dots placed according to the Cartesian coordinate scheme. In the second plot, vertical lines are used instead. In the third plot, lines are drawn connecting the dots. It is important to remember that although these lines help us see the shape of the distribution, the value of the pmf is zero between the dots. All of these plots can be made using `xyplot()` by setting the `type` argument to `'p'`, `'h'` (for histogram-like), or `c('p','l')`, respectively.

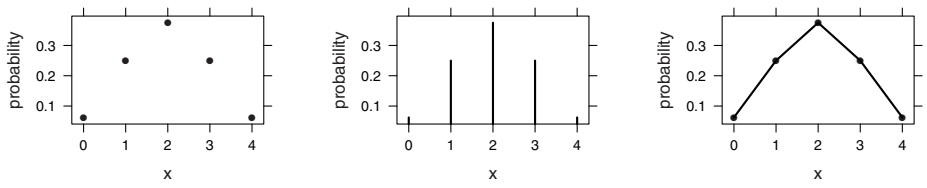

Figure 2.6. Graphs of a pmf.

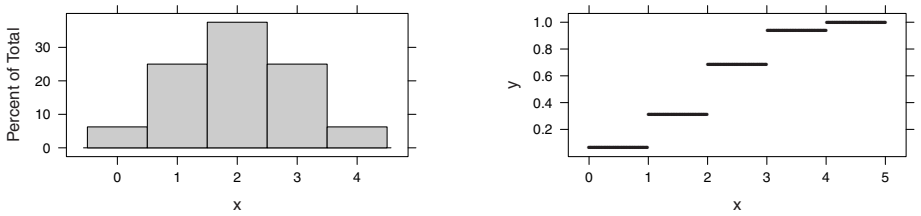

Figure 2.7. A probability histogram (left) and the graph of a cdf (right).

pmf-plot

```
# define pmf (vectorized),
# checking to be sure input is in { 0, 1, 2, 3, 4 }
> f <- function(x) {
+     sapply(x, function(x) {
+         if ( ! ( x %in% 0:4 ) ) { return(0) }
+         return( factorial(4) / ( 16 * factorial(x) * factorial(4-x) ) )
+     })
+ }
> f(0:6)
[1] 0.0625 0.2500 0.3750 0.2500 0.0625 0.0000 0.0000
> probplot1 <- xyplot(f(0:4)~0:4, xlab="x", ylab="probability")
> probplot2 <- xyplot(f(0:4)~0:4, xlab="x", ylab="probability",
+                     type="h")
> probplot3 <- xyplot(f(0:4)~0:4, xlab="x", ylab="probability",
+                     type=c("l","p"))
```

Another useful picture is a **probability histogram**. A probability histogram is made very much like the density histograms we made from data. Typically vertical bars are centered at the possible values of the random variable, but this is not required and for random variables with many possible values, it may be useful to combine some of the bins. In any case, the area of each bar represents the probability that the random variable takes on a value covered by the base of the rectangle. (If we choose boundaries that are not possible values of the random variable, we do not need to worry about whether the left or right endpoint belongs to the bin.) An example appears in Figure 2.7.

There is yet one more important way to describe the distribution of a discrete random variable, with a **cumulative distribution function** (**cdf**).

Definition 2.3.3 (Cumulative Distribution Function)**.** The *cumulative distribution function* F_X of a random variable X is defined by

$$F_X(x) = \mathrm{P}(X \leq x) \; .$$

□

The graph of a cdf of a discrete random variable is a step function. See Figure 2.7. There is, of course, a connection between the pmf and cdf of a given random variable, namely

$$F_X(x) = \sum_{w \le x} f_X(w)\,, \text{ and}$$
$$f_X(x) = F(x) - F(x^-)\,,$$

where $F(x^-) = \max\{F(w) \mid w < x\}$.[3] The notation here is more challenging than the idea. To get the cdf from the pmf, we simply add up the probabilities for all possible values up to and including x. To get the pmf from the cdf, we look at how much the cdf has increased since its "last change".

Example 2.3.1. Suppose that the cdf for a discrete random variable is given by

$$F(x) = \begin{cases} 1 - \frac{1}{\lfloor x+1 \rfloor} & \text{if } x > 0, \\ 0 & \text{otherwise .} \end{cases}$$

Q. What is the pmf for this random variable?

A. We begin by calculating $f(3)$:

$$f(3) = F(3) - F(3^-) = F(3) - F(2) = \left(1 - \frac{1}{4}\right) - \left(1 - \frac{1}{3}\right) = \frac{1}{12}\,.$$

More generally,

$$f(x) = \begin{cases} \frac{1}{x(x+1)} & \text{if } x \text{ is a positive integer,} \\ 0 & \text{otherwise.} \end{cases}$$ ◁

In the next section we turn our attention to two important examples of discrete distributions.

2.3.2. The Binomial and Negative Binomial Distributions

In Example 2.2.1 the random variable X counted the number of heads in four tosses of a fair coin. This is an example of an important family of distributions called the **binomial distributions**. A binomial random variable arises in a situation where our random process can be divided up into a sequence of smaller random processes called trials and

(1) the number of trials (usually denoted n) is specified in advance,

(2) there are two outcomes (traditionally called success and failure) for each trial,

(3) the probability of success (frequently denoted p or π) is the same in each trial, and

(4) each trial is independent of the other trials.

[3] Technically, if X can take on infinitely many values, we may need to use the supremum (sup) in place of the maximum (max). The supremum of a bounded, non-empty set is the smallest number that is at least as large as all numbers in the set. This is the maximum when the maximum exists. See Exercise 2.39.

The binomial random variable counts the *number of successes*. There are actually many different binomial distributions, one for each positive integer n and probability π. Collectively, we refer to these distributions as the **binomial family**. Members of the family are distinguished by the values of the **parameters** n and π but are otherwise very similar. We will use the following notation to succinctly describe binomial random variables.

Notation 2.3.4. If X is a binomial random variable with parameters n and π, we will write $X \sim \mathsf{Binom}(n, \pi)$.

We would like to have a general formula for the pmf of a binomial distribution:

$$f_X(x; n, \pi) = \mathrm{P}(X = x) = ??? \ .$$

Notice the use of the semi-colon in this equation. The semi-colon separates the parameters of the distribution from the independent variable of the pmf.

The tools we developed in Section 2.2 make deriving a formula for the pmf of a binomial random variable straightforward.

Theorem 2.3.5 (pmf for Binomial Distributions)**.** *Let $X \sim \mathsf{Binom}(n, \pi)$. Then the pmf for X is given by*

$$f_X(x; n, \pi) = \binom{n}{x} \pi^x (1-\pi)^{n-x} \ .$$

Proof. For a fixed n, there are 2^n possible outcomes (see Example 2.2.8). These outcomes are not equally likely unless $\pi = 0.5$, but we can determine the probability of any particular outcome. For example, if $n = 4$, then

$$\mathrm{P}(SSFS) = \pi \cdot \pi \cdot (1-\pi) \cdot \pi = \pi^3 (1-\pi)^1 \ .$$

More generally, any outcome with x successes in n trials will have probability $\pi^x (1-\pi)^{n-x}$, and the number of such outcomes is $\binom{n}{x}$ since we must select x of the n trials to be successful. So

$$\mathrm{P}(X = x) = \underbrace{\binom{n}{x}}_{\substack{\text{number of outcomes} \\ \text{with } X = x}} \cdot \underbrace{\pi^x (1-\pi)^{n-x}}_{\substack{\text{probability of} \\ \text{each outcome}}} \ . \qquad \square$$

The cdf for a binomial random variable cannot be expressed simply in closed form, but R offers functions to compute both the pmf and the cdf, as well as a function that will make random draws from a binomial distribution. These functions are summarized in Box 2.3.

Example 2.3.2. Here are some example uses of the functions in Box 2.3.

binom-demo01

```
> randomData <- rbinom(n=20,size=4,prob=0.5)
> randomData
 [1] 3 1 2 2 3 1 3 4 2 2 0 1 1 2 2 3 2 3 4 2
> table(randomData)
randomData
0 1 2 3 4
1 4 8 5 2
```

Box 2.3. Working with Binom(`size`, `prob`) in R

The following functions are available in R for working with a binomial random variable $X \sim$ Binom(`size`, `prob`):

function (& arguments)	explanation
`dbinom(x,size,prob)`	returns $P(X =$ `x`$)$ (the pmf)
`pbinom(q,size,prob)`	returns $P(X \leq$ `q`$)$ (the cdf)
`qbinom(p,size,prob)`	returns smallest x such that $P(X \leq x) \geq$ `p`
`rbinom(n,size,prob)`	makes `n` random draws of the random variable X and returns them in a vector
`set.seed(seed)`	sets the seed for the random number generator; see `?set.seed` for details

```
> dbinom(0:4,size=4,prob=0.5)          # matches earlier example
[1] 0.0625 0.2500 0.3750 0.2500 0.0625
> dbinom(0:4,size=4,prob=0.5) * 20     # pretty close to our table above
[1] 1.25 5.00 7.50 5.00 1.25
> pbinom(0:4,size=4,prob=0.5)          # same as cumsum(dbinom(...))
[1] 0.0625 0.3125 0.6875 0.9375 1.0000                                    ◁
```

It is important to note that

- R uses `size` for the number of trials (n), `n` for the number of random draws, and `prob` for the probability of success (π). `prob` and `size` can be abbreviated to `p` and `s` if desired, but most often we will simply use them without names and in the required order.
- `pbinom()` gives the cdf not the pmf. Reasons for this naming convention will become clearer later.
- There are similar functions in R for many of the distributions we will encounter, and they all follow a similar naming scheme. We simply replace `binom` with the R-name for a different distribution.

Example 2.3.3.

Q. Free Throw Freddie is a good free throw shooter. Over the last few seasons he has made 80% of his free throws. Let's assume that each of Freddie's shots is independent of the others[4] and that he has an 80% probability of making each. At the end of each practice, Freddie shoots 20 free throws to keep sharp. What is the probability that he makes all 20? At least 15? Exactly 16 (80% of 20)?

A. We will model this situation as a binomial random variable. Let X be the number of made free throws in 20 attempts. Then $X \sim$ Binom$(20, 0.8)$. We'll let

[4] How one might gather data to test this assumption of independence is an interesting question. But it will have to wait for another day.

Box 2.4. Working with NBinom(size, prob) in R

The following functions are available in R for working with a binomial random variable $X \sim$ NBinom(size, prob):

function (& arguments)	explanation
`rnbinom(n,size,prob)`	makes `n` random draws of the random variable X and returns them in a vector
`dnbinom(x,size,prob)`	returns $P(X =$ `x`$)$ (the pmf)
`qnbinom(p,size,prob)`	returns smallest x such that $P(X \leq x) \geq$ `p`
`pnbinom(q,size,prob)`	returns $P(X \leq$ `x`$)$ (the cdf)

R do the number crunching here, but you are welcome to check these using the formulas.

binom-freddy01

```
> dbinom(20,20,0.8)           # probability of making all 20
[1] 0.011529
> 1 - pbinom(14,20,0.8)       # probability of NOT making 14 or fewer
[1] 0.8042
> dbinom(16,20,0.8)           # probability of making exactly 16
[1] 0.2182
```

◁

The **negative binomial** random variables arise in a very similar situation to that of the binomial random variables. The difference is that instead of deciding in advance how many trials to perform and counting the number of successes, now we will decide how many successes there will be and repeat the trials until we have obtained the desired number of successes. The negative binomial random variable counts *the number of failures* that occur before getting the desired number of successes.[5]

Let X be negative binomial with parameters s (number of successes) and π (probability of success). We will denote this $X \sim$ NBinom(s, π). The R functions related to negative binomial distributions are similar to those for the binomial distributions (see Box 2.4). In R the number of successes is called `size` rather than s but can be abbreviated to `s`.

Example 2.3.4.

Q. Suppose you roll a pair of standard 6-sided dice until you get double sixes. How many rolls will it take? What is the probability that it will take you at least 20 rolls? At least 30? At least 40? At least 50?

A. We'll work out formulas for the negative binomial distributions shortly. For now, we'll let R do the work. Let $X \sim$ NBinom$(1, 1/36)$. Remember that X is

[5] Whether the negative binomial variable counts the number of *failures* or the number of *trials* varies in the literature, so when looking at information about this distribution, be sure to check which convention the author is using. The convention we are choosing matches what is done in R.

the number of failures. Let $Y = X + 1$ be the total number of rolls. We want to determine $\mathrm{P}(Y \geq 20) = \mathrm{P}(X \geq 19) = 1 - \mathrm{P}(X \leq 18)$ and the corresponding probabilities for the other questions. We can do them all in one step in R.

nbinom-first-example01

```
> 1-pnbinom(c(18,28,38,48),size=1,prob=1/36)
[1] 0.58552 0.44177 0.33332 0.25148
```

You could be rolling quite a while if you are waiting for double sixes. ◁

Now we want to derive the formula for the pmf of a negative binomial variable. The simplest case is when we stop after the first success, i.e., when $s = 1$. In this case

$$\begin{aligned}\mathrm{P}(X = x) &= \mathrm{P}(x \text{ failures followed by a success}) \\ &= (1-\pi)^x \pi \ .\end{aligned}$$

When $s = 1$, a negative binomial distribution is called a **geometric distribution** because the pmf forms a geometric series with ratio $(1 - \pi)$. The sum of the series is

$$\sum_{x=0}^{\infty} (1-\pi)^x \pi = \frac{\pi}{1-(1-\pi)} = \frac{\pi}{\pi} = 1 \ .$$

That's a good thing, since all pmfs are supposed to sum to 1. You can access the geometric distribution directly in R using `rgeom()`, `dgeom()`, and `pgeom()`.

Now let's try the case where $s > 1$. Let E be the event that there are x failures and $s - 1$ successes in the first $x + s - 1$ trials followed by a success in trial $x + s$. Then

$$\begin{aligned}\mathrm{P}(X = x) = \mathrm{P}(E) &= \binom{x+s-1}{x} (1-\pi)^x \pi^{s-1} \cdot \pi \\ &= \binom{x+s-1}{x} (1-\pi)^x \pi^s \ .\end{aligned}$$

This proves the following theorem. Notice that when $s = 1$ we get the same expression that we just derived for the geometric distribution.

Theorem 2.3.6. *Let* $X \sim \mathsf{NBinom}(s, \pi)$. *Then* $\mathrm{P}(X = x) = \binom{x+s-1}{x}(1-\pi)^x \pi^s$. □

2.4. Hypothesis Tests and p-Values

We have said that a fair coin is equally likely to be heads or tails when tossed. Based on this assumption, we determined that the probability of getting heads is 50%. But now suppose we have a coin and we do not know if it is a fair coin. How can we test it? Clearly we need to flip the coin and check the outcomes. But how many times do we flip the coin? And what decision do we make? If we flip the coin 100 times, we would expect roughly 50 heads and 50 tails, but we know that it is very likely we won't get *exactly* 50 of each. At what point would we become suspicious that the coin is *biased* (more likely to give one outcome than the other)?

2.4.1. The (Exact) Binomial Test

If we flip a coin n times and let X be the number of heads, then $X \sim \mathsf{Binom}(n, \pi)$ for some *unknown* value of π. We want to know whether or not $\pi = 0.50$. For example, suppose that $n = 100$ and we get $x = 40$ heads in our sample. What do we conclude? Is this consistent with a fair coin? Or is it sufficient evidence to suggest that the coin is biased?

Well, if it really is the case that $\pi = 0.50$, then $\mathrm{P}(X \leq 40) = 0.02844$, so we would only get 40 or fewer heads about 2.8% of the times that we did this test. In other words, getting only 40 heads is pretty unusual, but not extremely unusual. This gives us some evidence to suggest that the coin may be biased. After all, one of two things must be true. Either

- the coin is fair ($\pi = 0.50$) and we were just "unlucky" in our particular 100 tosses, or
- the coin is not fair, in which case the probability calculation we just did doesn't apply to the coin.

That in a nutshell is the logic of a **statistical hypothesis test**. We will learn a number of hypothesis tests, but they all follow the same basic four-step outline.

Step 1: State the null and alternative hypotheses

A **hypothesis** is a statement that can be either true or false. A **statistical hypothesis** is a hypothesis about a parameter (or parameters) of some population or process. In this example, the statistical hypothesis we are testing is

- H_0: $\pi = 0.50$

where π is the probability of obtaining a head when we flip the coin. This is called the **null hypothesis**. In some ways it is like a straw man. We will collect evidence (data) against this hypothesis. If the evidence is strong enough, we will reject the null hypothesis in favor of an **alternative hypothesis**. In our coin tossing example, the alternative hypothesis is

- H_a: $\pi \neq 0.50$

because $\pi = 0.50$ when the coin is a fair coin but will be some different value if the coin is biased.

Step 2: Calculate a test statistic

A **statistic** is a number calculated from sample data. Mathematically, a statistic is simply a function that assigns a real number to a data set:

$$f : \mathrm{DATA} \to \mathbb{R} \, .$$

In our coin tossing example, we may wish to count the number of heads obtained in 100 tosses of the coin. So our statistic is the function that takes any sequence of heads and tails (the data) and returns the number of heads (a number).

If we have a particular data set in mind, we will also refer to the numerical output of this function applied to that data set as a statistic. If we use this number to test a statistical hypothesis, we will call it a **test statistic**. In our example, the number of heads is 40, and we could denote this test statistic as $x = 40$. A test statistic should be a number that measures in some way how consistent the data are with the null hypothesis. In this case, a number near 50 is in keeping with the null hypothesis. The farther x is from 50, the stronger the evidence against the null hypothesis.

Step 3: Compute the p-value

Now we need to evaluate the evidence that our test statistic provides. To do this requires yet another way of thinking about our test statistic. Assuming our sample was obtained in some random way, we can also think about a statistic as a random variable. A random process produces a data set, from which we calculate some number. Schematically,

$$\text{random process} \to \text{sample} \to \text{DATA} \to \text{statistic}\,.$$

The distribution of this kind of random variable is called its **sampling distribution**. To distinguish between these two views (a particular number vs. a random variable), we will often use capital letters to denote random variables and lowercase to indicate particular values. So our random variable in this case will be called X.

Now we can ask probability questions about our test statistic. The general form of the question is, *How unusual would my test statistic be if the null hypothesis were true?* To answer this question, it is important that we know something about the distribution of X when the null hypothesis is true. In this case, $X \sim \mathsf{Binom}(100, 0.5)$. So how unusual is it to get only 40 heads? If we assume that the null hypothesis is true (i.e., that the coin is fair), then

$$\mathrm{P}(X \le 40) = \texttt{pbinom(40,100,0.5)} = 0.0284\,,$$

and since the $\mathsf{Binom}(100, 0.5)$ is a symmetric distribution, we get the same probability for the other tail:

$$\mathrm{P}(X \ge 60) = \texttt{1 - pbinom(59,100,0.5)} = 0.0284\,.$$

So the probability of getting a test statistic at least as extreme (unusual) as 40 is 0.0568. This probability is called a **p-value**.

Step 4: Draw a conclusion

Drawing a conclusion from a p-value is a bit of a judgment call. Our p-value is 0.0568. This means that if we flipped 100 fair coins many times, between 5% and 6% of these times we would obtain fewer than 41 or more than 59 heads. So our result of 40 is a bit on the unusual side, but not extremely so. Our data provide some evidence to suggest that the coin may not be fair, but the evidence is far from conclusive. If we are really interested in the coin, we probably need to gather more data.

Other hypothesis tests will proceed in a similar fashion. The details of how to compute a test statistic and how to convert it into a p-value will change from test to test, but the interpretation of the p-value is always the same. On the other hand, this interpretation does involve some amount of judgment, whereas the computation of the p-value is more or less automatic (a computer will typically do it for us). A famous rule of thumb regarding p-values is that when the p-value is less than 0.05, then we have enough evidence to reject the null hypothesis. This is a useful rule of thumb, but it should not be taken too literally. A p-value of 0.049 is hardly different from a p-value of 0.051. Both indicate nearly the same strength of evidence against the null hypothesis even though one is less than 0.05 and the other greater. Furthermore, when interpreting p-values, we must take into consideration the consequences of making a mistake. We'll return to the topic of errors in a moment, but first let's do some more examples.

Example 2.4.1.

Q. Let's return to our example of the lady tasting tea. Suppose we decide to test whether the lady can tell the difference between tea poured into milk and milk poured into tea by preparing 10 cups of tea. We will flip a coin to decide how each is prepared. Then we present the ten cups to the lady and have her state which ones she thinks were prepared each way. If she gets 9 out of 10 correct, what do we conclude?

A. The null hypothesis is that she is just guessing, i.e.,

$$H_0 : \pi = 0.5 \ .$$

Under that assumption, $\mathrm{P}(X \geq 9) = 0.0107$, so the chances of getting 9 or 10 correct just by guessing is just over 1%. This test can be conducted easily in R:

binomtest-lady-tea01

```
> 1-pbinom(8,10,0.5);
[1] 0.010742
> binom.test(9,10);

        Exact binomial test

data:  9 and 10
number of successes = 9, number of trials = 10, p-value = 0.02148
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.55498 0.99747
sample estimates:
probability of success
                   0.9
```

There is more in this output than we have discussed to this point, but it is easy to find the p-value. Notice that it is twice the probability that we just calculated. This is because by default, `binom.test()` does a **two-sided test**. In this case the p-value computed by `binom.test()` is $\mathrm{P}(X \geq 9 \text{ or } X \leq 1)$. It is possible to have R compute a one-sided test instead:

binomtest-lady-tea02

```
> binom.test(9,10,alternative="greater");

        Exact binomial test

data:  9 and 10
number of successes = 9, number of trials = 10, p-value = 0.01074
alternative hypothesis: true probability of success is greater than 0.5
95 percent confidence interval:
 0.60584 1.00000
sample estimates:
probability of success
                   0.9
```

The output indicates which kind of test is being done by reporting which of the following alternative hypotheses is being used:

$$\begin{aligned} H_a: &\quad \pi \neq 0.5 \quad \text{(two-sided alternative)}, \\ H_a: &\quad \pi < 0.5 \quad \text{(one-sided alternative)}, \\ H_a: &\quad \pi > 0.5 \quad \text{(one-sided alternative)}. \end{aligned}$$

The fact that R by default computes a two-sided p-value can serve as a reminder that one must offer additional justification to legitimately use a one-sided test. ◁

There is a certain attraction to one-sided tests. Since the p-value for a one-sided test is always smaller than the p-value for a two-sided test (unless the sample proportion is on the "wrong side" of the hypothesized proportion), a one-sided test appears to give stronger evidence against the null hypothesis. But it is not appropriate to use a one-sided test simply because you want a smaller p-value. There are differing opinions about when a one-sided test is appropriate. At one extreme are those who say one should *never* do a one-sided test. In any case one thing is certain: the decision to do a one-sided test must be something that can be *defended without referring to the data*. That is, it must be based on some *a priori* knowledge about the situation.

In the coin tossing example, it is clear that a two-sided alternative is the appropriate choice. We have no reason to expect the coin, if biased, to be biased in a particular direction. So without looking at data, we wouldn't even know which of the two possible one-sided tests to do. For the lady tasting tea, we can at least identify a difference between the two possible alternatives *before collecting any data*: in one case she is correct more often than expected and in the other case she is *wrong* too often. What will we conclude if she gets 9 out of 10 wrong? Should we consider that to be evidence that she can indeed tell the difference between the two tea preparations? The answer to that question will essentially answer the question of whether to use a one-sided or two-sided alternative.

Example 2.4.2.

Q. A children's game uses a die that has a picture of a ghost named Hugo on one side and numbers on the other sides. If the die is fair, the ghost should be rolled 1 time in 6. You test the die by rolling 50 times, and the ghost is rolled 16 times. Is there any reason to be concerned that the die is not fair?

A. We can perform a binomial test with the following hypotheses:

- $H_0 : \pi = 1/6$,
- $H_a : \pi \neq 1/6$.

Not having any prior knowledge about the die, a two-sided alternative is appropriate.

Now we need to convert our test statistic, $x = 16$, into a p-value. We can do this directly with `binom.test()`.

binomtest-hugo01

```
> binom.test(16,50,1/6)

        Exact binomial test

data:  16 and 50
number of successes = 16, number of trials = 50, p-value =
0.006943
alternative hypothesis: true probability of success is not equal to 0.16667
95 percent confidence interval:
 0.19520 0.46699
sample estimates:
probability of success
                  0.32
```

It is interesting to do this manually as well. Let's start with a one-sided test and compare to the results of `binom.test()`.

binomtest-hugo02

```
# one-sided test manually and using binom.test()
> 1-pbinom(15,50,1/6);
[1] 0.0057345
> binom.test(16,50,1/6,alternative="greater");

        Exact binomial test

data:  16 and 50
number of successes = 16, number of trials = 50, p-value =
0.005734
alternative hypothesis: true probability of success is greater than 0.16667
95 percent confidence interval:
 0.21210 1.00000
sample estimates:
probability of success
                  0.32
```

Obtaining a two-sided p-value is a bit more challenging this time. Since $X \sim \mathsf{Binom}(50, 1/6)$, the p-value should be $\mathrm{P}(X \geq 16) + \mathrm{P}(X \leq k)$ for some number k. But what number k do we use? The distribution of X is not symmetric in this case, and the mean ($50/6 = 8.33$) is not an integer, so it isn't so easy to simply take the "mirror image" like we did when the null hypothesis was $\pi = 0.5$. The usual solution is quite clever. We will add up $\mathrm{P}(X = x)$ for all values of x with

$\mathrm{P}(X = x) \leq \mathrm{P}(X = 16)$:

$$\text{p-value} = \sum_{\mathrm{P}(X=x) \leq \mathrm{P}(X=16)} \mathrm{P}(X = x) \, .$$

That is, we add the probabilities for all values that are at least as unusual as the value obtained from our data.

binomtest-hugo03

```
# finding the "other side" by inspection:
> dbinom(16,50,1/6);
[1] 0.0035458
> rbind(0:4,dbinom(0:4,50,1/6));
           [,1]      [,2]      [,3]    [,4]    [,5]
[1,] 0.00000000 1.0000000 2.0000000 3.00000 4.00000
[2,] 0.00010988 0.0010988 0.0053844 0.01723 0.04049
>
# this should match the p-value from binom.test()
> pbinom(1,50,1/6) + 1 - pbinom(15,50,1/6);
[1] 0.0069432
# letting R automate finding the interval too:
> probs <- dbinom(0:50,50,1/6); sum(probs[probs <= dbinom(16,50,1/6)])
[1] 0.0069432
```

Since $\mathrm{P}(X = x) \leq \mathrm{P}(X = 16)$ when $x = 0$ or $x = 1$, we get the left tail probability from `pbinom(1,50,1/6)`.

Note: This situation is based on a game the author once played with his children. Basing his strategy on the expected number of ghosts that would be rolled, he lost badly and became suspicious of the die. In fact, the die had *two* ghosts (on opposite sides). ◁

2.4.2. Types of Error and Statistical Power

When we carry out a hypothesis test, there are two kinds of mistakes we could make. It could be that the null hypothesis is true but that we reject it (because the p-value is small by random chance). This is called **type I error**. If we decide in advance what amount of type I error we can live with, that amount is called the significance level of the test. Usually it is denoted by α. You may see the result of a hypothesis test reported in terms of α instead of with a p-value. In our example above, we could say that "our results were not significant at the $\alpha = 0.05$ level" (or that they were significant at the $\alpha = 0.10$ level). This style of reporting used to be especially common when calculating p-values was more cumbersome than it is with today's computers and is equivalent to saying whether our p-value was above (not significant) or below (significant) our pre-specified threshold α. If we pick a significance level α in advance and the null hypothesis is true, then the probability of type I error is α.

On the other hand, it could be that the null hypothesis is false but that we do not reject it. This is called **type II error**. The probability of type II error is usually denoted β, but it is not as straightforward to calculate. This is because the probability of type II error depends on two things:

	reject H_0	don't reject H_0
H_0 is true	type I error	☺
H_0 is false	☺	type II error

Figure 2.8. Types of error when conducting hypothesis tests.

- the value of α that will be used and
- "just how wrong" the null hypothesis is.

Suppose that we choose $\alpha = 0.05$. Then we will reject the null hypothesis if the number of heads in our sample is less than 40 or greater than 60:

binomtest-power01

```
> qbinom(0.025,100,0.5)       # find q with pbinom(q,100,0.5) >= 0.025
[1] 40
> pbinom(39:40,100,0.5)       # double checking
[1] 0.017600 0.028444
```

On the other hand, if the coin is biased but $40 \leq x \leq 60$, then we will make a type II error. The values of our test statistic that lead to rejection of H_0 ($[0, 39] \cup [61, 100]$ in our example) are called the **rejection region** for the test. The boundary values of the rejection region are called **critical values**. (Note: This has nothing to do with critical values from calculus.)

If our coin is biased so that heads actually occurs 95% of the time, (i.e., $\pi = 0.95$ is our particular alternative), then we will be very likely to reject the null hypothesis and β will be small. In fact, we can easily calculate it:

binomtest-power02

```
> pbinom(60,100,0.95) - pbinom(39,100,0.95);
[1] 6.2386e-26
```

The chances of making the wrong decision – type II error – are very small in this case.

But if the bias is smaller and the coin comes up heads 55% of the time instead, then the probability of making a type II error is quite large:

binomtest-power03

```
> pbinom(60,100,0.55) - pbinom(39,100,0.55);
[1] 0.8648
```

This shows that 100 coin tosses isn't very likely to catch a coin with only a modest bias (differing from a fair coin by 5% or less).

Similarly, for any particular alternative value of π, we can calculate the probability of making a type II error. The **power** of the test against that alternative is $1 - \beta$, which is the probability that we will make the correct decision (reject the null hypothesis) when that alternative is true. That is, β and power are really functions of π_a, the probability of getting heads in some particular alternative. These functions can be plotted (by calculating a number of values and "connecting the dots").

binomtest-power05

```
> p <- seq(0,1,by=0.02);
> power <- 1 - ( pbinom(60,100,p) - pbinom(39,100,p) );
```

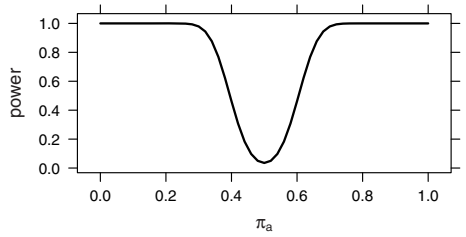

Figure 2.9. The power of a binomial test depends on π_a, the alternative value of π.

```
> myplot <- xyplot(power~p,ylab="power",xlab=expression(pi[a]),
+             type='l', lwd=2);
>
```

Figure 2.9 shows that the power is very low when π_a is near 0.5 but is nearly 1 when π_a is sufficiently far from 0.5.

Another way to explore power is to fix π_a and let n vary. This can help us determine how many coins to toss based on our desired power (see Figure 2.10).

binomtest-power06

```
> p <- rep(c(0.52,0.55,0.60), each=2000);
> plab <- paste("alt prob =", as.character(p));
> n <- rep(1:2000,times=3);
> critical <- qbinom(0.025,size=n,prob=p);
> power <- 1 - ( pbinom(n-critical+1,n,p) - pbinom(critical-1,n,p) );
> myplot <- xyplot(power~n|plab,ylab="power",xlab="number of coin tosses",
+             ylim=c(0,1.1), type='l', lwd=2);
```

A test is said to be **under-powered** if we collect too little data to have much chance of detecting an effect of some desired magnitude and **over-powered** if we collect more data than were necessary. The design of a good statistical study will include a **power analysis** that attempts to determine a reasonable sample size given some assumptions about effect size.

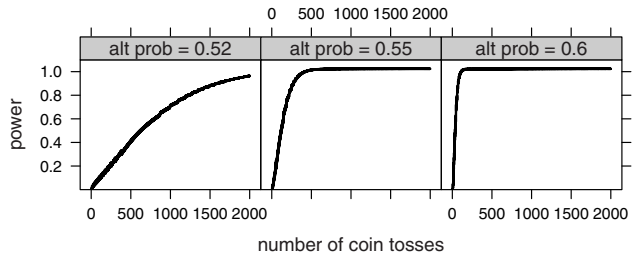

Figure 2.10. Power curves show how power depends on sample size for various values of π_a.