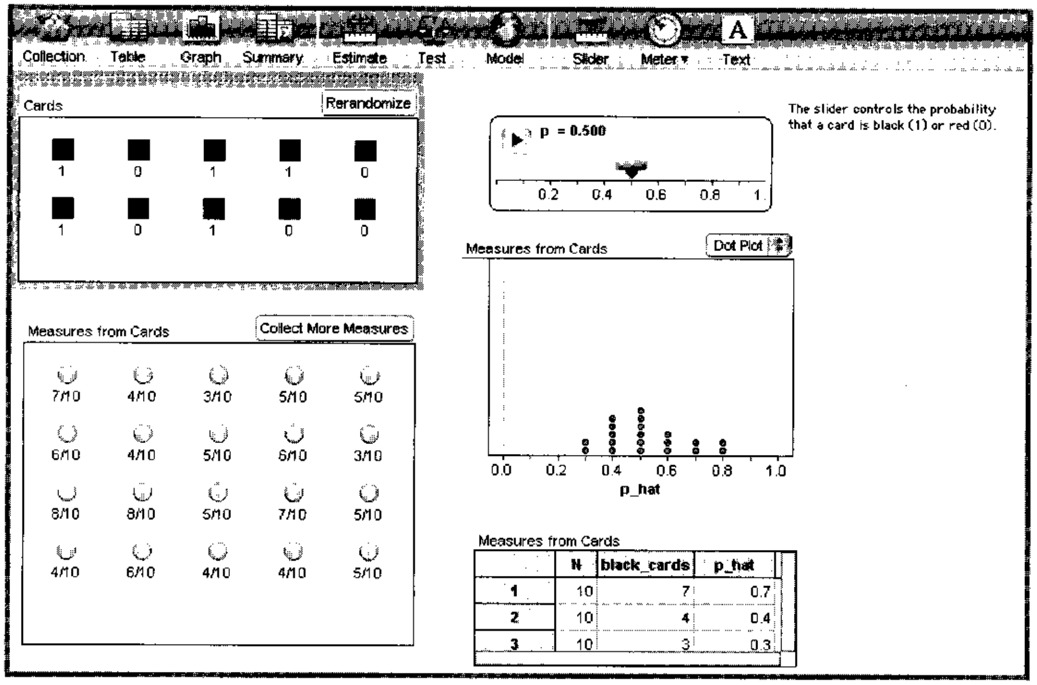
Demo 13: Building the Binomial Distribution

Constructing the binomial distribution by resampling • How the distribution depends on the population proportion

The binomial distribution is at the root of statistical situations having to do with proportions. It can be confusing to remember just what it's a distribution *of.* The conceptual problem is in the layering of the situation: You pull a card from a deck of red and black cards—where's the distribution? The answer is that you have to draw *n* cards (with replacement) and count the black ones; and *then you have to repeat that process.* The distribution is the distribution of *numbers* (or, equivalently, proportions) of blacks in those repeated series of draws.

This demo tries to clarify this "layering" and to give you some mental images of the distribution.



The slider controls the probability that a card is black (1) or red (0).

Measures from Cards

	N	black_cards	p_hat
1	10	7	0.7
2	10	4	0.4
3	10	3	0.3

What To Do

▹ Open **Building Binomial.ftm.** It will look like the illustration.

A collection, **Cards,** shows ten red and black cards, coded 0 and 1. The probability that a card is black is controlled by the slider **p** top center. Below it are the results of 20 sets of 10 cards. You can see a graph of those results (shown as proportions) and the top of a table as well. Note that we are not simulating the *whole* deck here; the **Cards** collection is already a sample, drawn randomly from the deck.

We begin by focusing only on the cards and the slider.

▹ Click the **Rerandomize** button in **Cards** repeatedly. See what happens.

▹ Drag the slider; see what happens (nothing will change below).

You should see that the slider controls the probability that a card is black, but that the number of black cards varies from set to set. That is, even though the probability might be 0.500, there aren't always exactly 5 cards.

So now we want to collect these numbers of black cards to see how they vary.

▹ Reset the **p** slider at 0.5.

▹ Click the **Collect More Measures** button on the lower (**Measures from Cards**) collection to see what happens. You will see that, when the collecting is done, the last green ball corresponds to the current **Cards** collection.

Fathom empties the measures collection and then collects 20 new measures, rerandomizing the data collection each time. The "flying balls" show how each new measure (which appears as a green ball with a fraction below it) represents an entire new set of cards. Note that when it starts over, Fathom rescales the graph.

Note: We have animation turned on in order to show this process more slowly. You can, of course, turn it off when you're ready. But if the process is too fast, here's a suggestion: Change the measures collection to collect only one measure at a time, without emptying the collection. Then you can look at each **Cards** collection and see what new measure appears.

▹ Click on a green ball in the collection to see where it is in the graph. Then select points in the graph to see which balls they correspond to (for example, drag a rectangle around all the 0.5's).

▹ Drag the attribute **black_cards** from the case table (at the bottom) to the horizontal axis of the graph, replacing **p_hat.** See how the two graphs show the same distribution. Put **p_hat** back when you're done.

▹ Explore what happens when you change **p** and then press **Collect More Measures.**

Questions

1 As you change **p,** how does the distribution change?

2 As you re-collect measures (leaving **p** constant), how does the distribution change?

3 If you set **p** to 1.0, what will the distribution look like? Why?

The distribution of **black_cards** is called a *binomial* distribution. The "bi" part comes from "two." There are two possible outcomes for each card: red and black. It could be heads or tails, success or failure, male or female, east or west, or whatever. The probability of each choice can range from 0 to 1; if you know one probability (p), the probability of the other (often q) is 1 minus the probability of the first (that is, $q = 1 - p$).

The distribution shows the proportions you get for one of these choices when you draw N cards repeatedly. These proportions (often called p-hat, or $\hat{p}$—the sample proportion) tend to be close to the true probability, but there's variation about that "expected" number.

If you do a random walk repeatedly, the distribution of ending positions is also binomial. And if you have a large enough sample size, the binomial distribution looks very much like a normal distribution (given certain conditions; see Demo 31, "Why np > 10 Is a Good Rule of Thumb").

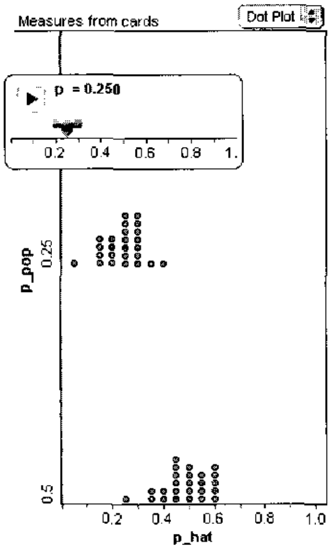
Onward!

Now we'll study more explicitly how the distribution depends on the value of **p**.

▹ Open **Building Binomial part 2.ftm.** It will resemble the previous file, but there are now 20 cases in the **Cards** collection instead of 10, and 25 measures (green balls) instead of 20. We have also made the graph bigger and made changes to the way the measures collection collects. (So, *don't* click **Collect More Measures** until we tell you to!)

▹ Play with the slider to see what it does to the **Cards.** When you're ready, set it at 0.25 (by editing the number in the slider).

▹ Now click the **Collect More Measures** button. Fathom quickly collects more measures, *adding* them to the collection—instead of replacing them—and adding the points to the plot.



Your graph should look something like the one in the illustration, split to show how the distributions are different. We have scaled the axis to show the range 0 to 1. The attribute **p_pop** (probability for the population) is the value of **p**, the slider, at the time the measures were collected.

▹ Set the slider to three more values and collect measures each time so that you have a total of five different probabilities. An interesting set of values for the slider is {0.1, 0.25, 0.5, 0.8, 0.95}.

▹ Change the graph to a histogram, an Ntigram, and a box plot using the pop-up menu in the graph to see other representations of this distribution.

More Questions

4 How does the center of the distribution depend on **p_pop**?

5 How does the spread of the distribution (for example, the width of boxes in the box plot) depend on **p_pop**?

6 Why should the spread be smaller when **p_pop** is close to 1 or 0? **Sol**

7 Why is it that before, all the proportions in the graph were 0.40, 0.50, 0.60, and so on, while now we have 0.45, 0.55, and the like?

Extensions

8 Select the graph and choose **Plot Value** from the **Graph** menu. Use the formula editor to specify a value to plot; Fathom will plot that value for each part of the graph. The simplest is **mean( ),** but you could, for example, plot the 10th percentile with **percentile(10,).** (Leaving the blank after the comma tells Fathom to use the attribute on the axis.) You can also plot multiple values by choosing **Plot Value** again. Edit an existing formula by double-clicking the formula at the bottom of the graph.

9 We have just explored how the distribution depends on **p_pop.** Open **Building Binomial part 3.ftm** and see how it depends on the sample size, **N.** To change the sample size, select the **Cards** collection and then choose **New Cases** from the **Collection** menu. (It starts out with 10.) An interesting set of values for *N* is {10, 40, 100, 200, 1000}. This is closely related to Demo 12, "How Random Walks Go as Root *N*."