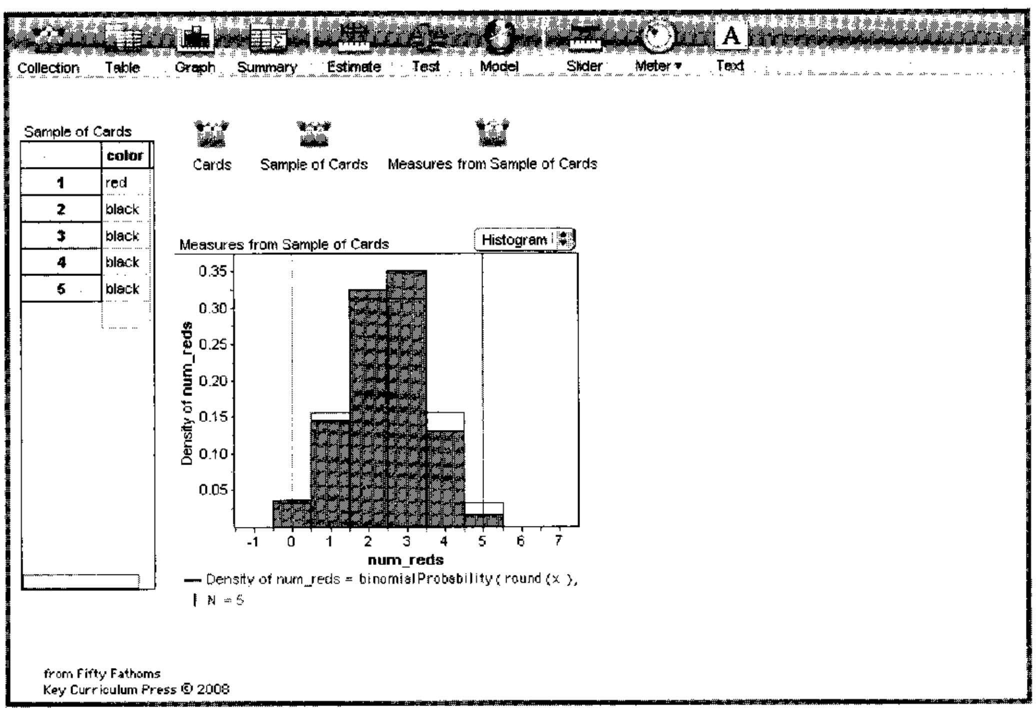
Demo 49: Sampling Without Replacement and the Hypergeometric Distribution

How distributions change when the sample is large compared to the population

Most statistical tests and classroom problems that have to do with sampling assume that you are sampling with replacement. In real life, however, you often sample *without* replacement. For example, if you're taking a survey, you probably won't survey the same person twice. Usually, this doesn't matter a lot. If the population you're sampling from is much larger than the sample, the sample will have about the same characteristics whichever way you sample. But there are some situations where it does matter: where the population is small (for example, a classroom) or the sample is a large fraction of the population it is drawn from.

One such setting is where you draw cards from a deck and don't put them back. We'll use a standard 52-card deck in this demo, simplified so that the cards are only **red** or **black**—26 of each. We'll draw samples of various sizes, counting how many **red**s there are. Then (of course) we'll build up sampling distributions of those numbers.



What To Do

- ▹ Open **Hypergeometric.ftm.** It will look something like the illustration.

Here you can see the "source" collection, **Cards; Sample of Cards** is a sample of five cards from that collection—you can see the values in the table at the left. **Measures from Sample of Cards** is a collection of 200 measures—the number of **red**s in the sample of 5, called **num_reds.**

You can see the binomial probability function plotted on the histogram of those measures, as well as a vertical gold line at the sample size—in this case, 5. That is the maximum possible value for **num_reds;** if you draw five cards, you can't get six red ones. The histogram hangs over only because the bins are wide.

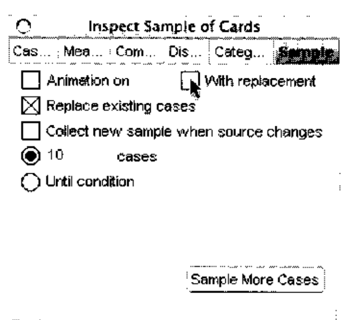
- ▹ Let's get a new set of measures. Click once on the **Measures from Sample of Cards** collection to select it. Then choose **Collect More Measures** from the **Collection** menu (or press its shortcut, **⌘+Y** on the Mac or **Control+Y** in Windows). After a moment, the graph updates with new values.

▹ Repeat as necessary so that you're sure you understand what's going on.

Note: Unlike in some earlier demos, we have omitted the part where you just redraw samples a few times to understand that part of this three-layer process. If you need to see it, make a graph, graph the data from one sample, and do some resampling. You might even turn on animation. We would have set that up, but we'll need the screen space, as you will see later.

You should see that the binomial function matches the histogram pretty well.

▹ Let's increase the sample size. Double-click the sample collection (the middle one) to open its inspector. It will resemble the one in the illustration. Note how the **With replacement** box is unchecked—for one of the only times in this book.



▹ Change the number sampled from 5 to 10, as shown. Click **Sample More Cases** and see how the table updates to show ten values.

▹ Close the inspector or move it aside so that you can see the graph.

▹ Again, collect more measures (select the measures collection and choose **Collect More Measures** from the **Collection** menu). See how the graph changes, especially compared with the theoretical curve. Repeat as necessary to get the feel for it.

▹ Do the same for sample sizes of 20 and 30.

By the time you get to $n = 30$, the function should not fit the data very well.

Questions

1 When we had a sample size of 5, there were two bars with the same theoretical value. Now there's only one. What made the difference?

2 At $n = 30$, how would you characterize the difference between the data from the samples and the theoretical curve? **Sol**

3 Why does that difference occur? Why does drawing without replacement change the distribution of reds you would get?

4 If we sampled 52 cards instead of 30, what would the distribution look like? **Sol**

Onward!

▹ Close the inspector if it is open (for screen space).

▹ Choose **Show Hidden Objects** from the **Object** menu. A new graph appears in the lower right. It has a *hypergeometric* curve on it.

▹ Repeat as needed, even changing sample size, to convince yourself that this distribution fits the data better than the regular binomial curve.

▹ If you wish, return to a sample size of 5, and see that the two distributions are nearly identical in this case.

Thus this situation produces a different distribution, more bunched together than a regular binomial distribution. If you look at the arguments of the function that is plotted in the new graph, you can see that Fathom has to know the population size (52 cards) and the number of "successes" (26 reds) to calculate the hypergeometric probability function.

Extension

Develop a rule of thumb for when you should use the hypergeometric distribution instead of the binomial distribution. That is, when we did a sample size of 5, it looked okay, but 30 looked much different. For this, it would help to put the two functions on the same graph. You might also want to change the proportions of successes and the size of the population.

Challenges

5 Make a Fathom document to simulate card counting in this scenario:

Suppose you shuffle a deck and deal out 21 cards, face up. Then you bet $1 (at even odds) on whether the next card is red or black. If you play 100 hands, using the obvious strategy, how much do you win on average? **Sol**

6 Explain clearly why this is different from playing the same game, but with flipping a coin.

7 The same as the first task, but with four decks in play. This is why multiple decks are used at casinos.

8 Start with a complete deck. (You could use **DeckOfCards.ftm**, which comes with the full version of Fathom, as a starting point). This time, you deal out 31 cards, then you can bet $1 or $2 on a hand of blackjack against a dealer. (This demands complicated formulas.) No hitting, but if you get a blackjack you win outright, and you lose a push. Can you win in the long run by counting face cards?