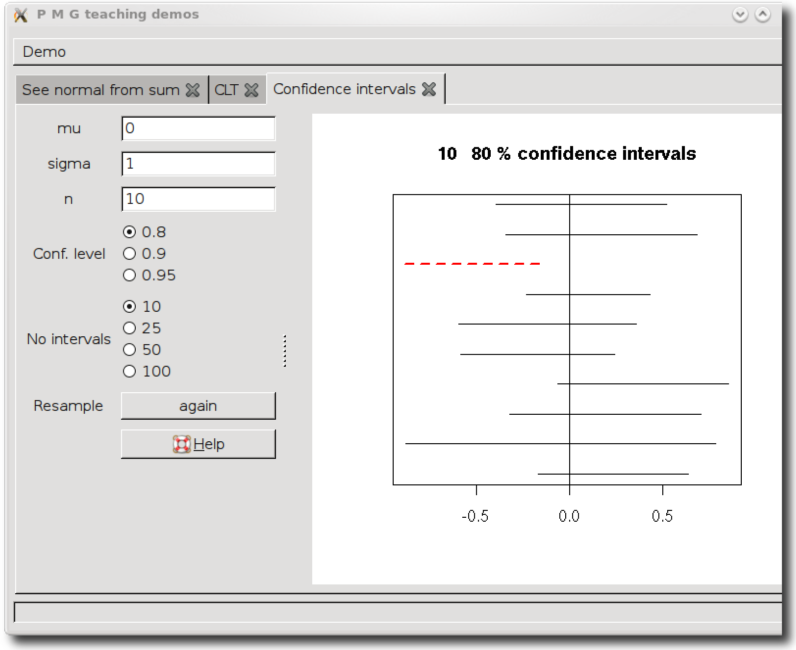

Figure 1: The teaching demos window after a few demos have been loaded. The one shown illustrates several simulated confidence intervals.

This project use 3 teaching demos for `pmg` to illustrate some important concepts in chapter 5 and 6 of the text: the binomial distribution, the central limit theorem, and a confidence interval.

First load `pmg` then under the `Plots` menu find the `teaching demos` item. Figure 1 shows the resulting window after a few demos are loaded.

1 The binomial

The binomial distribution describes the randomness found in counting the number of successes in n independent (Bernoulli) trials, each of which has success probability p. In math language

$$X = X_1 + X_2 + \cdots + X_n \quad \text{where} \quad p = P(X_i = 1) = 1 - P(X_i = 0).$$

This distribution can be used to describe a wide range of problems: the number of tagged dear in a capture/recapture sample; the number of surveyed people who agreed; The number of plain cheese pizzas sold out of 100; etc. The basic thing is we are *counting* the number of times something "good" happens out of n times (where n is fixed ahead of time).

To demonstrate, on the teaching demos window under the demo menu select the "Binomial distribution."

You will see 10 balls some are colored blue, others not. The 10 balls are the "n" trials, and a success is the ball has a "blue" color. The parameters n and p for the binomial determine a) the number of balls and b) how likely a given all is to be colored blue.

Click on click a few times and you should see that a histogram is being created recording the number of blue balls that have appeared in each sample.

1. Keep clicking until there are just 2 balls colored blue. The probability of this occuring is only 0.043. How many times did it take?

2. Describe the shape of your histogram.

The binomial distribution, as discussed in class, has a a formula to describe its distribution. The demo "construction of normal" can plot this. Open this demo, then change the value of `Law` to `Binomial` and `n for:` to 10, then click the `update` button.

You should see a spike graph of the distribution.

1. Does this look like that from your last simulation?

2. Go back to the binomial demo and keep clicking `click` only this time do 25 at a time. Do this a few times to gets lots of simulated data. Now does your graphic show data like that of the spike plot?

Again, as discussed in class. The binomial distribution can be difficult when n is large (atleast by hand). The normal distribution is a good approximation when n is large enough so that np and $n(1-p)$ are more than 10. From the "see normal from sum" demo we can see how well the normal distribution fits the binomial.

Change the `n for:` to 10, `prob` to 0.5 and click on "show normal law" and then `update`. The read normal distribution with mean np and standard deviation $\sqrt{np(1-p)}$ is drawn on top of the spike plot.

The normal approximation says: the area to the left of some value under the red curve is about equal to the sum of the heights of the spikes to the left of some value. Our eye doesn't really see that, we see the "fit." That's what we will use to assess if n is large enough.

For our initial choice $n = 10$ and $p = 0.5$ so $np = 5$. Is this too small?

1. Compare the graphs of $n = 10$ and $p = 0.5$ to $n = 20$ and $p = 0.5$. in terms of how well the normal "fits" the binomial. Is there a big difference? No difference? ...

2. Crank up n to 100. Does the "fit" look good?

3. The binomial distribution in the graphic is approximated by a sample and when n is this large we need a bigger sample. Change the number of simulations to 5000 and repeat. Does this look better?

   [The actual difference can be computed with code like

```
> theDiff <- function(n, p) {
+     mean = n * p
+     sigma = sqrt(n * p * 1 - p)
+     x = (mean - 3 * sigma):(mean + 3 * sigma)
+     y = dnorm(x, mean, sigma) - dbinom(x, n, p)
+     max(abs(y))
+ }
> theDiff(n = 1000, p = 0.5)

[1] 0.01785013
```

   ]

2 Central Limit Theorem

The central limit theorem describes what happens to the distribution of $barX = (X_1 + X_2 + \cdots + X_n)/n$ when the X_i's are a random sample. We had this table in class

```
                                   n is
                          small        |     large
            normal    |                |
                      |  NORMAL        |  NORMAL
population is         |----------------------------------
                      |                |
            not normal|  ??????        |  NORMAL
```

| |

We want to verify this as best we can.

Look at the demo `See normal from sum`.

When the `Law` is binomial, we saw that the red normal curve tracks the binomial distribution well provided we use a big enough simulation to represent the binomial distribution. What about other distributions?

1. Select `Normal` for the law. Is there a big difference in "fit" between $n = 1$, $n = 10$, $n = 100$?

2. Select `Uniform` for the law. Is there a big difference in fit between $n = 1$, $n = 2$, $n = 5$,$n = 10$, $n = 100$?

3. Based on your experimenting what would you consider "large" for the uniform Law? [Large depends on the population (or law). When the population is long tailed or skewed, large can be, well, large. In this case large isn't so large.]

3 Confidence intervals

The `Confidence Intervals` demo displays a graphic that might be one of the most common in introductor statistics books. (See Moore and McCabe page 386.) Let's see if we can make any sense of it.

The graphic is trying to show that confidence intervals are random, as they depend on a random sample. Some are "good" and some are "bad" We can't control which are which, but by making the intervals long enough (a big margin of error) we can make more of them "good."

What is "good?" In this case the interval covers the population parameter, which is indicated with a vertical bar in the graphic. Intervals which are "bad" are marked in red.

Click the `again` button a few times to get different confidence intervals. You should see some are red and most aren't. In reality, as we don't know the parameter, we don't see a given interval as red or not. So when we say a interval is a 95% confidence interval we can't say that it actually contains the parameter. All we can say is that the process we used to find the interval gets it right with probability 0.95.

So, here's a question: What is the probability an interval is `red`? It is $p = 1-$ the confidence level. Now, another question: are the different intervals independent? (Yes). Finally, what is the distribution of the number of red intervals in a simulation of 10 intervals? (If you said binomial with parameters $n = 10$ and $p = 1-$ the confidence level pat your self on the back.)

Let's see what this means

1. Run the simulation 10 times. Keep track of the number of times there are no red intervals. How many did you get. (The expected number is 7).

2. Do 100 intervals now. How many miss? (The expected number is $100 \cdot (0.5)$.)

Let's look at the margin of error. For this simulation the value is $z\sigma/\sqrt{n}$ where $z = 1.96$ when 0.95 is the confidence level, 1.64 when 0.90 and 1.28 when 0.80.

1. Verify that when you change the confidence level to be *smaller* the margin of error gets bigger.

2. Verify that if you change `sigma` to be *smaller* than the margin of error gets smaller.

3. Verify that if you make n *bigger* that the margin of error gets smaller.