This project makes use of a probability calculator to understand and visualize the concept of a distribution for both discrete and continuous data.

First we must do some work to get things ready:

1. Start R
2. Load the pmg package: `library(pmg)`
3. Minimize the big R window, so that only pmg shows.
4. Find and click on the `Commands` tab near the center of the pmg window.
5. This tab allows us to type commands. You need to type (or copy and paste) the following exactly as typed

   ```
   source("http://www.math.csi.cuny.edu/verzani/classes/MTH113/YN.R")
   probabilityCalculator()
   ```

   After this, click the `evaluate` button.

This opens up the "Probability Calculutor" (PC).

1 The binomial

At the top of the PC you can adjust the law using a popup box. Change it to "Binomial". You can now enter in values for the `size` (n) and `prob` (p). If you enter in 4 and 0.5 you can imagine modeling the probabilities for the number of heads in 4 coin tosses.

Enter in 4 and 0.5, and then in the `Result` area put in 2 for the `Value`. Click on `update` and you should have (Figure 1):

- A graph of the specified binomial distribution with red indicating the values of k which are being added up to compute the probability.
- The result appearing below the Value area. In this case $p = 0.6875$ which is $P(X \leq 2)$ where X is the number of heads in 4 coin tosses.

Question 1.1 *Change the part `Cumulative` to "to right" and click update. What is the result? How does it relate the previous value of* 0.6875*? Can you tell if the answer is* $P(X > 2)$ *or* $P(X \geq 2)$*? (What is the distinction?)*

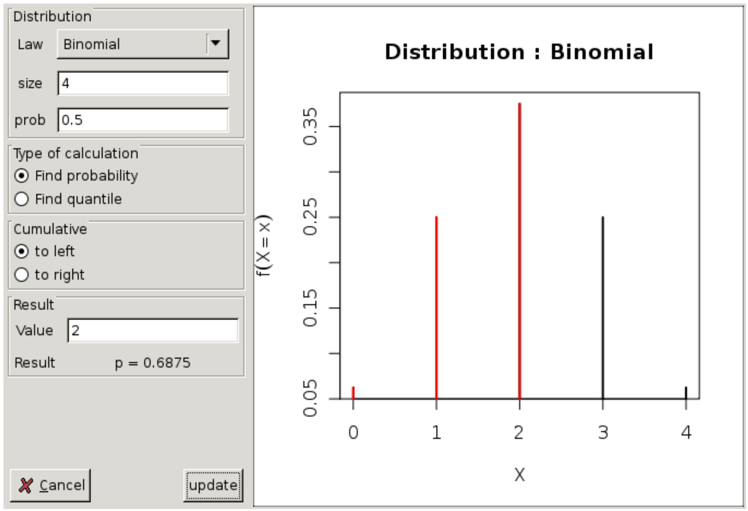

Figure 1: Probability Calculator showing binomial distribution with $n = 4$ and $p = 0.5$ for $k = 2$. The probability computed is $P(X \leq 2)$.

Question 1.2 *If you toss a coin 100 times, what is the probability of getting 40 or fewer heads? (This is found with* $n = 100, p = -.5$ *and* $k = 40$.

Question 1.3 *A baseball player has historically had a* $p = 0.300$ *batting average. If they get 450 at bats this year, what is the probability there average will be 0.333 or more? (That means they get 150 or more hits?)*

Question 1.4 *A survey of 100 people is taken of Staten Islanders to find out if their opinion of Rudy Guiliani is different from the rest of the countries. Suppose 50% of the country supports Rudy. What is the probability that in the survey 60 or more people will support Rudy?*

2 Normal distribution

The normal distribution is a continuous distribution. For these, the probability that a randomly chosen number is less than x is given by the area under some graph called the density. Different densities yield different probabilities for the random number.

The normal distribution refers to a whole family of distributions which are characterized by their mean and standard deviation. In PC these are specified with `mean` and `sd`. To see, lets find the probability a standard normal random variable is less than 0.75 (Figure2).

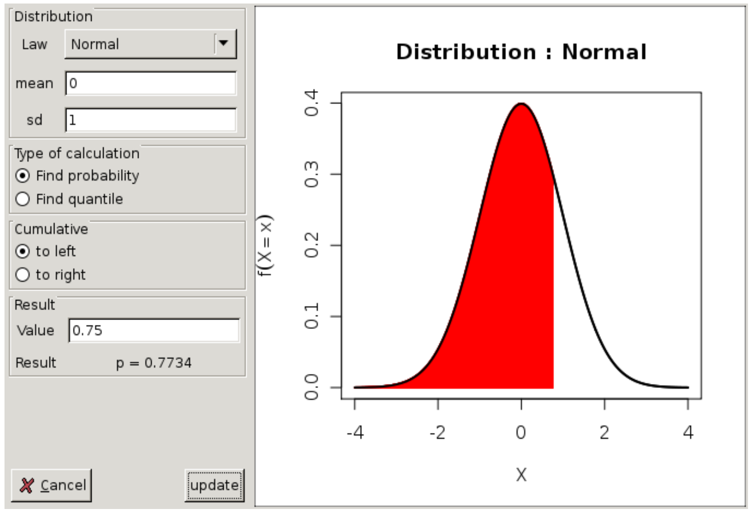

Figure 2: The probability a standard normal (a mean of 0 and standard deviation of 1) is less than 0.75 is found to be $p = 0.7734$.

Change the law to "Normal" and keep the mean and standard deviation at 0 and 1. These parameters are what we mean by the standard normal. Change the value to 0.75, click update, and the resulting graph should look like the figure. The computer found that $p = 0.7734$. This means if a random number has a standard normal distribution then it will be less than 0.75 roughly 77 times out of 100.

Question 2.1 *Again for the standard normal, find the probability a random number is* ***more*** *than* 0.75*. Relate this to* $p = 0.7734$.

Question 2.2 *In class we discussed that the probability a normal is more than 1 standard deviation from the mean is roughly* $0.16 = (1 - .68)/2$*. Verify this, by using the PC to compute the probability a standard normal is more than 1. Write down the answer to 4 decimal points.*

Question 2.3 *Change the parameters to* $\mu = 10$ *and* $\sigma = 15$*. Now find the probability a number is bigger than 25. (Why 25?). Relate to the last question.*

Question 2.4 *IQ scores are well approximated by a normal distribution with a mean of 100 and standard deviation of 16. Find the probability a randomly chosen person has an IQ less than 90.*

Question 2.5 *SAT math scores are well approximated by a normal distribution with a mean of 460 and a standard deviation of 70. Find the probability a randomly chosen person scored more than 600.*

Question 2.6 *Heights of full-grown American males are well approximated by the normal distribution with a mean of 70 inches (5 feet 10 inches) with a standard deviation of 3 inches. Assuming this, find the probability a randomly chosen person is more than 6 feet 6 inches. Less than 5 feet 3 inches.*

3 Quantiles

The probability question is roughly:

> If a random number has some distribution, find the probability it is less than x.

The quantile question is the reverse:

> If a random number has some distribution, find the value of x so that is has some specified probability (for example 0.5 which gives the median.)

The median is a quantile, as the probability a value is less than the median should be 1/2 (why?). So are Q_1 and Q_3.

To find Q_1 for a standard normal distribution we do the following: change the mean and standard deviation to 0 and 1; **select "Find quantile"** for the type of calculation; enter a probability (between 0 and 1) for the "Value". For Q_1 this value if 0.25. Clicking update produces an answer of -0.6745.

Question 3.1 *Find Q_3 for the standard normal.*

Question 3.2 *Keep the mean at 0 but change the standard deviation to 15. Now find Q_3. It should change. Verify that it is related to 15 and how?*

Question 3.3 *If math SAT scores are normally distributed with mean 460 and standard deviation 70, find the score which indicates the top 5% (95% of the area is to the left of this value.)*

Question 3.4 *If gestation times for adults is 270 days on average with a standard deviation of 7 days, find the time for the shortest 5% of gestation times assuming a normal distribution.*

4 The binomial and the normal

In graphs of the binomial when n is large, one sees a distribution that is basically bell shaped like the normal distribution. This is no trick of the eye, the binomial distribution is known to be approximately normal when n is large enough (unless p is really close to 0 or 1). This allows one to compute the binomial probabilities using the normal. Which before computers was a big deal to get numbers, and now is very important for theoretical reasons.

The basic idea is this. If X is binomial with n and p then it has a mean $\mu = np$ and standard deviation $\sigma = \sqrt{np(1-p)}$. For notation's sake let Z be the standard normal then

$$P(X \leq x) \approx P(Z \leq \frac{x-\mu}{\sigma})$$

That is the probability a binomial is less than x is approximately the probability that a standard normal is less than the z-score of x.

Let's verify this.

If we toss a coin 100 times and let X count the number of heads we have that X is binomial with mean $\mu = 100(1/2) = 50$ and $\sigma = \sqrt{100(1/2)(1-1/2)} = 5$.

Let $x = 45$. For the binomial you can compute with PC that $P(X \leq 45) = 0.1841$.

Comparing to the normal, we have a z-score of $(45-50)/5 = -1$ and PC gives a probability of 0.1587. Not exactly the same, but approximate.

Let's do another. Suppose we have a survey of 1,000 people to see if their opinions have changed. Before we knew that 75% were supporting a position. What is the probability that in the survey 725 or fewer support (less than 72.5%). If we use the binomial model this is found with `size=1000`, `prob=0.75` and `Value=725`. Using PC gives `p = 0.03782`.

The normal approxmation needs us to find a z score of 725 for $\mu = 1000(.75)$ and $\sigma = \sqrt{1000(.75)(1-.75)}$. These values can be found with your calculator or within the Commands tab with these commands:

```
> mu = 1000 * 0.75
> mu

[1] 750

> sdev = sqrt(1000 * 0.75 * (1 - 0.75))
> sdev

[1] 13.69306
```

```
> (725 - mu)/sdev
[1] -1.825742
```

Putting this z-score into the PC for the standard normal gives the approximate value of $p = 0.3394$.

Question 4.1 *Use the normal to approximate the probability a* 0.300 *average baseball player gets 150 or more hits in 450 at bats. (The binomial is* $n = 450$ *and* $p = 0.300$*. If a player gets 150 or more hits, his batting average for the year is* 0.333 *or better.*