This project again makes use of a probability calculator. Our goal today is to understand two things:

- The relationship between the normal distribution and the binomial (The so-called normal approximation to the binomial)
- The concept of a quantile

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 Finding μ and σ.

For the binomial distribution we need to compute μ and σ from n and p. We have formulas $\mu = np$ and $\sigma = \sqrt{np(1-p)}$. These can be computed on the calculator, but we should be able to do so using the computer.

To do so we use the `Commands` tab. Click on this tab and enter

```
n = 100
p = 0.5
n*p
sqrt(n*p*(1-p))
```

Click on evaluate. Your output should look like

```
> n = 100; p = 0.5
> n * p
```

```
[1] 50

> sqrt(n * p * (1 - p))

[1] 5
```

If you want to change n and p, simply click on edit, then change the values and again click evaluate.

Question 1.1 *If* $n = 14,000$ *and* $p = 18/35$ *find* μ *and* σ.

Question 1.2 *If* $n = 82$ *and* $p = 1/2$ *find* μ *and* σ.

2 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 $\sigma = \sqrt{np(1-p)}$. The key relationship is

> Binomial probabilities for X, such as $P(X \leq k)$ are answered using the normal distribution with mean $\mu = np$ and $\sigma = \sqrt{np(1-p)}$.

If we were to use the table in the book, we need to relate this to the standard normal. For notation's sake let Z be the standard normal then we have

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

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 can use the 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 2.1 *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.*

You can also just change the parameters for μ and σ with the normal distribution, rather than finding the z-score.

Question 2.2 *Suppose it is known that 60% of all the gasoline stations in America charge more than $2.50 per gallon of gasoline. A survey of 250 randomly chosen stations finds that 160 charge more than $2.50. Find the probability that such a survey would find 160 or more if it was done again? Find both the binomial value, and the normal approximation.*

Question 2.3 *The 2000 census found that 23% of New Yorkers live below the poverty line. Has this changed since then. Suppose a survey of 12,000 find that 22.5% of those surveyed were below the poverty line. Would this be an unusally small amount given 23% is the true proportion? We can answer this by finding the probability another survey of 12,000 would find* $2,700 = (0.225)(12,000)$ *or fewer below the poverty line.*

Find this probability two ways. First using the exact binomial distribution. The second time using the normal approximation.

3 Quantiles

The "probability question" is roughly:

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

For instance, if birthweights are normally distributed with a mean of 7.5 pounds and a 1 pound standard deviation, find the probability a randomly chosen birth is less than 5.5 pounds.

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.)

For instance, find a value so that a randomly chosen birth has a 5% chance of being more than this value.

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.*