Activity 2

1 Synopsis.

What we are set about to do. This activity will take us through a journey exploring the central limit theorem. We will learn about the *quincunx*, a device designed to demonstrate convergence to the normal distribution.

We will not construct a physical quincunx; that would demand far greater manual dexterity and time than your instructor has. Rather, we will *simulate* a quincunx or, if you wish, create a virtual one. On our way, you will tackle a simple programming assignment.

What you need to know. In order to benefit from this activity you need to know

1. Some rudiments of R. The introduction in the computer practice session already held may well be enough.
2. The central limit theorem. Go back to your book or notes and be sure to understand its meaning and implications.
3. The notion of convergence in distribution. Again, go back to your book or notes. You should also remind yourself of the Tchebycheff inequality.

2 Context.

2.1 The central limit theorem (CLT).

There are many varieties of the central limit theorem. In its simplest form, the CLT tells that if $X_1, X_2, \ldots, X_n$ are independent random variables with common mean m and variance σ^2, then

$$\frac{X_1 + X_2 + \cdots + X_n}{n} \longrightarrow N(m, \sigma^2/n) \tag{1}$$

as n grows "large". An alternative, more formal statement, would be:

$$\lim_{n\to\infty} \sqrt{n}\left(\frac{\overline{X} - m}{\sigma}\right) \xrightarrow{d} N(0,1) \tag{2}$$

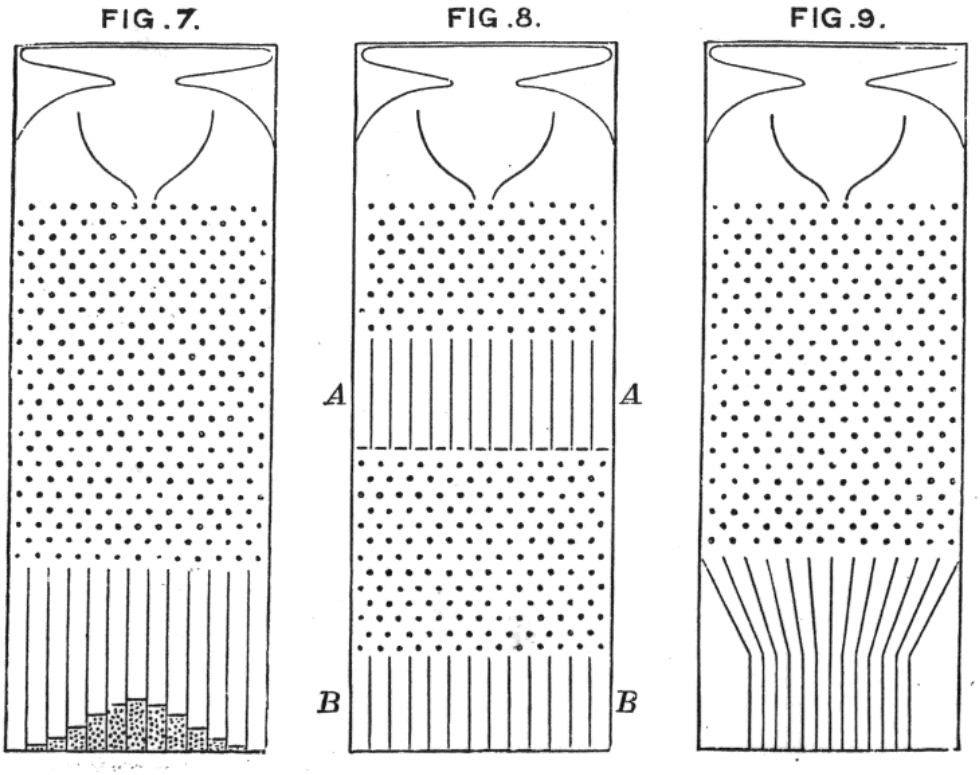
You have seen that proven in class for the case of binary random variables (de Moivre's theorem), but it holds much more generally. In fact, convergence to normality does not require equal means or variances, nor independence: it even holds under mild dependence of $X_1, X_2, \ldots$.

Now, the CLT is just that: a theorem. It is true, period. If you accept the axioms of probability, everything follows. It remains to be seen if in the real world magnitudes that we think well described by a sum of many independent random variables end up having an approximate normal distribution.

2.2 Galton's *quincunx*.

Francis Galton (1822–1911) was a men of great curiosity and ingenuity. Around 1870–1875 he was greatly interested in the normal distribution (remember that some theoretical breakthroughs had already been achieved by de Moivre in 1730; for an account of the discovery of the normal distribution, see [2]).

In trying to show how the CLT worked, he invented the so-called *quincunx*, represented in three different variations in the picture next.



Let's focus on the leftmost panel, under the "FIG. 7" heading. A large number of balls are thrown from the top. As they fall down, they repeatedly hit the pegs on their way falling with equal probability to each of the two possible sides. As a result, when the ball reaches the bottom it has taken a number of steps to either left or right, which can be thought of as random, with probability $\frac{1}{2}$ for each side.

At the bottom, balls are collected in containers. As they accumulate in the different containers, they approximate the profile of a gaussian bell curve —the density of a normal distribution.

Now, it is clear that for this to work out, the spacing of the pegs have to be very precise —if the ball knocks any of them slightly sideways, then the probability of falling to one side will be greater than one half—. So the whole thing has to be constructed to very exact specifications. Your instructor will provide a more detailed description in class, perhaps showing you one of the many animations available in the Web. For an example, point your Java-enabled browser to http://www.jcu.edu/math/isep/quincunx/quincunx.html or look any of the videos available: http://www.youtube.com/watch?v=9tTHST1sLV8 or http://www.youtube.com/watch?v=xDIyAOBa_yU for instance[1].

3 Questions.

Using the above information, the hints provided by your instructor in the classroom and the additional information in the Appendix, you should be ready to answer the following:

1. Simulate 1000, then 10000 flips of a regular coin (equal probability of heads and tails). Compute the relative frequency of heads as a function of the number of throws.
   (a) What number c does that relative frequency tend to?
   (b) Can you state a number of throws such that *with certainty* you will have a relative frequency of tails in the interval $c \pm 0.01$?
   (c) Can you state a number of throws such that *with probability at least* 0.99 (or any other probability of your choice) the relative frequency of tails falls in the interval $c \pm 0.01$? (HINT: Remember the Tchebycheff bound, or use the fact that a binomial with large n and $np > 18$ is close to a normal.)
   (d) Looking back at your answer to questions 1b and 1c, what does it mean exactly "tend to" in question 1a? An ordinary limit? (Remember from your first course in Calculus what a limit is!) Explain why it cannot be the same thing. (HINT: Look back to question (b).)

2. Write a small program in R simulating a quincunx. It is not hard at all! Use what we did in Seminar 2 as a head start and follow the hints below.
   (a) Fix the number of balls you are going to drop to something "large", for instance 50.000.
   (b) Fix the number of levels of the quincunx to a not too small even integer, for instance 40. At each level, the balls can take an step right or left, so after hitting 40 pegs the total number of steps can vary from -40 to +40. It is quite easy to "simulate" a random step left or right with probabilities (0.5, 0.5); remember Seminar 2.
   (c) Define and fill with zeroes a `bins` vector of length 81 ($= 2 \times 40 + 1$). It is as simple as this:

   ```
   > bins <- rep(0,81)
   ```

   (d) When a simulated ball reaches the bottom of the quincunx, check how many steps right it has taken and "count" it in one element of `bins`. Balls having taken as many steps right as steps left will be counted in `bins[41]` —the central element of the vector—. Those having taken k steps right more than to the left, in `bins[41+k]`. k may of course be negative.
   (e) After the `bins` vector is filled, you may plot its values with an instruction such as:

```
> plot(bins,type="l")
```

The frequency polygon you get should approximate "almost" a bell-shaped density. "Almost", because there is a problem, yet: even-numbered entries of `bins` are all zero (if you drop through 40 levels, you cannot end in an even-numbered bin of `bins`; see why?). You can take every second element of `bins` with an instruction such as this:

```
> bins <- bins[ seq(from=1,to=length(bins),by=2) ]
```

and now the plot will come out right.

3. *(Optional, for private study; you do not have to submit anything on this)* Yoy may be wondering: "OK, so in problem 1 the distribution of

$$\frac{(Z/n) - \frac{1}{2}}{\sqrt{pq/n}}$$

approaches a $N(0,1)$. But, how well, how fast?"

It turns out that the answer to this question is given by the so called Berry-Esseen theorem, and results which improve on it. You may look in any probability text of medium to advanced level, or in the Wikipedia (http://en.wikipedia.org/wiki/Berry-Esseen_theorem).

A Background information

A.1 Use of if.

Sometimes we want our code to follow a different path according to some condition. For instance, when simulating our quincunx we would like a (virtual) ball to go to one side or the other according to whether a random number is greater or smaller than 0.5. The following code shows one way of doing that:

```
> StepsRight <- 0
> if ( runif(1) > 0.5 ) {
     StepsRight <- StepsRight + 1
  }  else {
     StepsRight <- StepsRight - 1
   }
```

Aside from the brief in-class introduction to R, you may benefit from using any of the available introductory books: among them, [3], [1], [4] and the freely available documents in http://cran.r-project.org.

Notes

[1]Links visited on February, 28, 2012. Google for "quincunx videos" if the location has changed.

References

[1] P. Dalgaard. *Introductory Statistics with R.* Statistics and Computing. Springer-Verlag, 2002. Signatura: 519.682 DAL.

[2] S.M. Stigler. *The History of Statistics*. Belknap Press, 1986.

[3] M.D. Ugarte, A.F. Militino, and A.T. Arnholt. *Probability and Statistics with R*. CRC Press, 2008.

[4] W.N. Venables and B.D. Ripley. *Modern Applied Statistics with S-Plus*. Springer-Verlag, New York, third edition, 1999.