# Solutions to sheet 5

It is easiest to think of the medians as just some statistic whose distribution we would like to determine, at least to the extent of drawing a fairly accurate histogram. The fact that the medians are generated by samples of size 11 from the standard normal is irrelevant to most of the problem. The method should apply to any statistic from an experiment that can be simulated in Minitab.

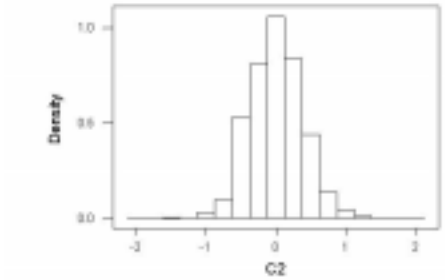
I decided to carry out all experimentation, to generate samples of medians (in column C2), with the following macro:

```
gmacro
hw5
erase c1-c2  ## clean out any junk already there
do k99 =1:???  ## replace ??? by various values of choice
    Random 11 c1;
            Normal 0 1.
            let c2(k99) = median(c1)
 enddo
            ## histogram commands could be placed here
endmacro
```

Also I experimented with the histogram command on the Graph menu to find out how to get the cut points where I wanted them. I could then paste from **Window>History** into the small window brought up by **Edit >Command line editor**, or into the macro. Following queries from several folks, I decided to center the big bar at zero, rather than make it run from 0 to1/4. (It makes little difference to the problem.) The central bar then corresponds to the bin from –1/8 to 1/8. The command

```
Histogram C2;
Density;
MidPoint -2:2/0.25;
Bar;
ScFrame;
ScAnnotation.
```

produced the sort of histogram that I wanted:




I experimented by choosing ??? as 1000. The central bar seemed to have height close to 1 in all the experiments. Of course the histograms looked slightly different for each new experiment.

The aim of the problem was to produce a histogram that represented the true theoretical distribution of the median, M, from a sample of 11 independent standard normal random variables. That is, ideally I would like the central bar to have area

$p = P\{-1/8 < M \le 1/8\}$.

The total area under the histogram should equal 1, the sum of the probabilities. For a bar of width 1/4 to have area p, its height would have to be h = 4p. A height close to 1 gives a value of p close to 1/4. My rough guess at p is 1/4.

Of course, we don't know p exactly, which is why we are simulating. Minitab draws a histogram by counting the numbers of medians lying in each bin range, then drawing bars whose areas are equal to proportions. For example, for a set of n medians, suppose X of them lie between –1/8 and 1/8. The bar for that bin would have area X/n, and height H= 4X/n. The value of X would change from one simulation to the next; it is a random variable. The mechanism that generates X is like n tosses of a coin that lands heads with probability p. The random variable X has a Bin(n, p) distribution. The mean of X equals np, and the variance equals np(1–p).

Binomial distributions are well approximated by normal distributions with the same mean and standard deviation. The count X has an approximate normal distribution with mean np and standard deviation $\sqrt{np(1-p)}$. The height H = (4/n)X has an approximate normal distribution with mean (4/n)np = 4p = h and standard deviation

$$\sigma = (4/n)\sqrt{np(1-p)} = 4\sqrt{p(1-p)/n}$$

From the normal approximation we know that

$$P\{h-2\sigma < H < h + 2\sigma\} \approx 95\%$$

If we make $2\sigma < h/20$ we will have a 95% chance of getting our central bar within 5% of its true value, that is, we will be reasonably sure (probability 0.95) that the height H lies within 5% of the true height h.

How big do we need n to be to ensure that $2\sigma < h/20$? We need

$$2 \times 4\sqrt{p(1-p)/n} < 4p/20$$

or,

$$1600\, p(1-p)/n < p^2$$

or,

$$1600\frac{1-p}{p} < n$$

A value of p close to 1/4 gives an n bigger than about 4800. (Now draw it.)

Once you generate a column of medians, something like

```
let c13 = mean(c2>-1/8 and c2<= 1/8)
```

will find the proportion of observations that contribute to the central bar of the histogram.

Using the quicker method described on the ***email correspondence*** page, I generated 100,000 medians, and got the more precise estimate .267 for p. The (estimated) standard error is about 0.0014, so we can probably believe the first two decimal places in the estimate for p.