PROBLEMS FOR STAT 261, DUE FEBRUARY 2, 2007

The "Central Limit Theorem" for Bernoulli trials tells us that for any $0 < p < 1$, letting S_n be the number of successes in n independent trials, each of which has probability p of success (so probability $q = 1 - p$ of failure), and $a < b$ any real numbers,

$$\lim_{n\to\infty} P\left\{a \leq \frac{S_n - pn}{\sqrt{pqn}} \leq b\right\} = \int_a^b \phi(x)dx,$$

where $\phi(x)$ is the *standard normal density* defined by

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Prove this for the case $p = \frac{1}{2}$. Use the following steps:

a) We may as well assume that n is even. Let

$$p_k = P\big\{S_n = \frac{k}{2} + \frac{n}{2}\big\} = 2^{-n}\binom{n}{\frac{n}{2}+\frac{k}{2}}$$

for k an even integer between $-n$ and n. Show that

$$\begin{aligned} p_k &= p_0 \cdot \frac{\frac{n}{2}\big(\frac{n}{2}-1\big)\cdot\dots\cdot\big(\frac{n}{2}-(\frac{|k|}{2}-1)\big)}{\big(\frac{n}{2}+1\big)\big(\frac{n}{2}+2\big)\cdot\dots\cdot\big(\frac{n}{2}+\frac{|k|}{2}\big)} \\ &= p_0 \cdot \frac{\big(1-\frac{2}{n}\big)\big(1-\frac{4}{n}\big)\cdot\dots\cdot\big(1-\frac{|k|-2}{n}\big)}{\big(1+\frac{2}{n}\big)\big(1+\frac{4}{n}\big)\cdot\dots\cdot\big(1+\frac{|k|}{n}\big)}. \end{aligned}$$

Take $k \geq 0$. We have

$$\begin{aligned} \frac{p_{k+2}}{p_k} &= \frac{\binom{n}{\frac{n}{2}+\frac{k}{2}+1}}{\binom{n}{\frac{n}{2}+\frac{k}{2}}} \\ &= \frac{\frac{n!}{(\frac{n}{2}+\frac{k}{2}+1)!(\frac{n}{2}-\frac{k}{2}-1)!}}{\frac{n!}{(\frac{n}{2}+\frac{k}{2})!(\frac{n}{2}-\frac{k}{2})!}} \\ &= \frac{\big(\frac{n}{2}-\frac{k}{2}\big)!}{\big(\frac{n}{2}+\frac{k}{2}\big)!}\frac{\big(\frac{n}{2}+\frac{k}{2}+1\big)!}{\big(\frac{n}{2}-\frac{k}{2}-1\big)!} \\ &= \frac{\big(\frac{n}{2}-\frac{k}{2}\big)}{\big(\frac{n}{2}+\frac{k}{2}+1\big)}. \end{aligned}$$

Since $p_k = p_0 \prod_{i=1}^{k} \frac{p_i}{p_{i-1}}$ the first statement follows. The second statement follows by dividing numerator and denominator by $(\frac{n}{2})^{k/2}$. The case of negative k is done the same way.

b) Using calculus, show that for $-\frac{1}{2} \leq x \leq \frac{1}{2}$,

$$x - x^2 \leq \log(1+x) \leq x.$$

(Hint: Take the difference, and then find the maximum or minimum.) The logarithm here is the natural logarithm.

Let $f(x) = \log(1+x) - x$. Then

$$f'(x) = \frac{1}{1+x} - 1 = \frac{-x}{1+x}.$$

Thus, the only local extremum is at $x = 0$, where $f(x)$. Since $f(-1/2) < 0$ and $f(1/2) < 0$, it follows that $f(x) \leq 0$ everywhere on the interval $[-1/2, 1/2]$. Similarly, if we let $g(x) = \log(1+x) - x + x^2$, we get

$$g'(x) = 2x - 1 + \frac{1}{1+x} = \frac{2x^2 + x - 1}{1+x}.$$

This is 0 only at $x = \frac{1}{2}$ and $x = -1$. Thus, the extrema are at $-\frac{1}{2}$ and $\frac{1}{2}$. Since $g(-\frac{1}{2})$ and $g(\frac{1}{2})$ are both positive, it follows that $g(x) \geq 0$ everywhere on the interval $[-1/2, 1/2]$.

c) Show that this implies that for $-n/2 < k < n/2$,

$$p_0 e^{-k^2/2n} \cdot e^{-|k|^3/4n^2} \leq p_k \leq p_0 e^{-k^2/2n} \cdot e^{|k|^3/4n^2}$$

You may use the formula $1 + 2 + \cdots + j = j(j+1)/2$.

We consider only the case of k positive. We need to show that

$$\frac{-k^2}{2n} + \frac{k^3}{4n^2} \leq \log \frac{p_k}{p_0} \leq \frac{-k^2}{2n} + \frac{k^3}{4n^2}.$$

From part (a), we have that

$$\log \frac{p_k}{p_0} = \sum_{i=1}^{k/2-1} \log\left(1 - \frac{2i}{n}\right) - \sum_{i=1}^{k/2} \log\left(1 + \frac{2i}{n}\right)$$

Applying the upper bound from part (c) to the first sum, and the lower bound to the second sum, we get

$$\begin{aligned}
\log \frac{p_k}{p_0} &\leq \sum_{i=1}^{(k-1)/2} -\frac{2i}{n} - \sum_{i=1}^{k/2} \left[\frac{2i}{n} - \left(\frac{2i}{n}\right)^2\right] \\
&= -\frac{2}{n}\frac{1}{2}\left(\frac{k}{2} - 1\right)\left(\frac{k}{2}\right) - \frac{2}{n}\frac{1}{2}\left(\frac{k}{2}\right)\left(\frac{k}{2} + 1\right) + \frac{4}{n^2}\sum_{i=1}^{k/2} i^2 \\
&\leq -\frac{k^2}{2n} + \frac{k^3}{4n^2}
\end{aligned}$$

The lower bound goes the same way, using the lower bound from part (c) for the first sum, and the upper bound for the second sum.

d) The probability we want to estimate is

$$P\{\frac{n}{2}+\frac{a\sqrt{n}}{2} \leq S_n \leq \frac{n}{2}+\frac{b\sqrt{n}}{2}\}.$$

Write this as a sum of p_k's.

$$P\{\frac{n}{2}+\frac{a\sqrt{n}}{2} \leq S_n \leq \frac{n}{2}+\frac{b\sqrt{n}}{2}\} = \sum_{a\sqrt{n}\leq k\leq b\sqrt{n}} p_k.$$

e) We can ignore those p_k's in the sum (if any) for which $|k| \geq n/4$. Why?

The weak law of large numbers says that

$$0 = \lim_{n\to\infty} P\{\left|S_n - \frac{n}{2}\right| \geq \frac{n}{4}\} = \lim_{n\to\infty} \sum_{|k|\geq n/4} p_k.$$

f) Let $z = k/\sqrt{n}$. Show that

$$\frac{1}{p_0\sqrt{n}} P\{\frac{n}{2}+a\frac{\sqrt{n}}{2} \leq S_n \leq \frac{n}{2}+b\frac{\sqrt{n}}{2}\}$$

approximates a Riemann sum for the integral $\int_a^b e^{-z^2/2}dz$ with $\Delta z = 2/\sqrt{n}$, and so converges to this integral.

By the previous results,

$$\frac{p_k}{p_0} = f\left(\frac{k}{2}\Delta z\right) + g\left(\frac{k}{2}\Delta z\right),$$

where $f(z) = e^{-z^2/2}$ and

$$|g_n(z)| \leq \left(e^{|z|^3/4\sqrt{n}} - 1\right) e^{-z^2/2}.$$

We have

$$\begin{aligned}\frac{1}{p_0\sqrt{n}} P\{\frac{n}{2}+\frac{a\sqrt{n}}{2} \leq S_n \leq \frac{n}{2}+\frac{b\sqrt{n}}{2}\} &= n^{-1/2} \sum_{a\sqrt{n}/2\leq k\leq b\sqrt{n}/2} \frac{p_{2k}}{p_0} \\ &= \frac{1}{2} \sum_{a\leq k\Delta z\leq b} \left[(k\Delta z) + g_n(k\Delta z)\right](\Delta z).\end{aligned}$$

This is a Riemann sum which converges to the integral

$$\int_a^b \left[f(z) + g_n(z)\right] dz$$

For n large enough, $|z|^3/4\sqrt{n} < \frac{1}{2}$ for all $a \leq z \leq b$, so that by part (b) we then have $e^{|z|^3/4\sqrt{n}} \leq 2 \cdot |z|^3/4\sqrt{n}$ for all such z. Thus

$$\lim_{n\to\infty} \left|\int_a^b g_n(z)dz\right| \leq \int_a^b \frac{|z|^3}{2\sqrt{n}} e^{-z^2/2}dz \leq n^{-1/2} \int_0^\infty z^3 e^{-z^2/2}dz.$$

Since the integral on the right is finite, the entire right-hand side goes to 0 as $n \to \infty$. Thus, the sum converges to

$$\frac{1}{2}\int_a^b f(z)dz.$$

(Formally, we are exchanging a limit and an integral here, so we would need a little bit of careful analysis to make this formally correct. It can be done hands-on, but this would be covered by the *Dominated Convergence Theorem* — part of a second semester of probability.)

g) All that remains is to show that $\lim_{n\to\infty}\sqrt{n}p_0 = \sqrt{2/\pi}$. Assuming that $\int_{-\infty}^{\infty}\phi(z)dz = 1$, show that this is the only possible limit.

We know that

$$\lim n \to \infty \frac{1}{p_0\sqrt{n}} P\{\frac{n}{2} + \frac{a\sqrt{n}}{2} \le S_n \le \frac{n}{2} + \frac{b\sqrt{n}}{2}\} = \frac{1}{2}\int_a^b f(z)dz,$$

We know that for K very large, if we take $a = -K$ and $b = K$,

$$P\{\frac{n}{2} + \frac{a\sqrt{n}}{2} \le S_n \le \frac{n}{2} + \frac{b\sqrt{n}}{2}\} = P\left\{(S_n - E[S_n])^2 \le K\,\mathrm{Var}(S_n)\right\},$$

which is very close to 1, by Chebyshev's inequality. Thus, $p_0\sqrt{n}$ has a limit, it must be that

$$\lim_{n\to\infty}\frac{1}{p_0\sqrt{n}}\cdot 1 = \frac{1}{2}\int_{-\infty}^{\infty} f(z)dz.$$

Since $(2\pi)^{1/2}\phi(z) = f(z)$, it must be that

$$\lim_{n\to\infty} p_0\sqrt{n} = \sqrt{\frac{2}{\pi}}.$$

h) Now we need to show that $\lim_{n\to\infty}\sqrt{n}p_0$ exists. To do this, we need *Stirling's formula*:

$$\lim_{n\to\infty}\frac{n!}{\sqrt{n^{n+1/2}e^{-n}}} = c,$$

where c is a constant. Otherwise, just use this formula to complete the proof of the CLT. Show, by the way, that the constant c in Stirling's formula is $\sqrt{2\pi}$.

By definition,

$$p_0 = 2^{-n}\binom{n}{n/2} = 2^{-n}\frac{n!}{(n/2)!^2}.$$

Stirling's formula tells us that

$$\begin{aligned}\lim_{n\to\infty}\sqrt{n}p_0 &= \lim_{n\to\infty} 2^{-n}\frac{\sqrt{n}cn^{n+1/2}e^{-n}}{\left(c(n/2)^{n/2+1/2}e^{-n/2}\right)^2} \\ &= \lim_{n\to\infty}\frac{cn^{n+1}e^{-n}}{c^2 n^{n+1}2^{-n-1}e^{-n}} \\ &= \frac{2}{c}\end{aligned}$$

Since the only possible limit (by h) is $\sqrt{2/\pi}$, we have shown that $c = \sqrt{2\pi}$, and

$$\lim_{n\to\infty} P\left\{a \le \frac{S_n - n/2}{\sqrt{n}/2} \le b\right\} = \lim_{n\to\infty}\sqrt{n}p_0\frac{1}{2}\int_a^b f(z)dz = \frac{1}{\sqrt{2\pi}}\int_a^b e^{-z^2/2}dz.$$