Class Notes 3-1-2019

Approximation Theorems, Chapter 4

- The text on pages 141-142 (Fig 4.1 and 4.3, exclude Fig 4.2) gives various graphical examples of how the binomial distribution appears to be approximated very well by a normal distribution. This is due to the fact that the binomial random variable S_n can be viewed as a sum of independent identically distributed (iid) Bernoulli random variables $X_1, \ldots, X_n$,
i.e., $S_n = X_1 + \ldots + X_n$. This approximate normal behavior is the result of the central limit theorem (CLT) which makes a very general statement about the distribution of such sum of independent random variables. I have posted a very general version on the class web site for anyone adventurous enough to dig into it.

  The CLT in its general form is very important in many areas of science and it often provides a rationale why so many measurable phenomena are (approximately) normally distributed.

- **The CLT for Binomial Random Variables (Moivre-Laplace CLT):**
Let $0 < p < 1$ be fixed and suppose that $S_n \sim \text{Bin}(n, p)$. Then for any fixed $-\infty \le a \le b \le \infty$ we have

$$\lim_{n\to\infty} P\left(a \le \frac{S_n - np}{\sqrt{np(1-p)}} \le b\right) = P(a \le Z \le b) = \Phi(b) - \Phi(a)$$

  or equivalently

$$\lim_{n\to\infty} P\left(a \le \frac{\frac{S_n}{n} - p}{\sqrt{p(1-p)/n}} \le b\right) = P(a \le Z \le b) = \Phi(b) - \Phi(a)$$

  The text gives a somewhat incomplete proof of the CLT in the binomial case, using Stirling's approximation

$$n! \sim \left(\frac{n}{e}\right)^n \sqrt{2\pi n} \quad \text{as } n \to \infty$$

  to approximate the $\binom{n}{k}$ term in $P(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k}$.

- **Rule of Thumb:**
The normal approximation will fail if p gets too close to 0 or 1 and n is not sufficiently large. A quick rule of thumb for the approximation to be "reliable" is: $np(1-p) > 10$. This is motivated by the following considerations. We know that S_n has values in $\{0, 1, 2, \ldots, n\}$. If $S_n \approx \mathcal{N}(np, np(1-p))$ one would expect that the approximating normal distribution should mostly fall between 0 and n, the range of S_n. Thus one would want to have

$$0 \le np - k\sqrt{np(1-p)} \quad \text{and} \quad np + k\sqrt{np(1-p)} \le n \tag{1}$$

  where k might be chosen as $k = 3$, for example. That would ensure that at least 99.74% of that normal distribution falls within the range of S_n. If we choose $k = \sqrt{10} = 3.16$ then that percentage rises to 99.84%.
The two inequalities in (**??**) translate to

$$\frac{np(1-p)}{\max(p^2, (1-p)^2)} \ge k^2$$

and since $\max(p^2,(1-p)^2) \le 1$ we see that

$$np(1-p) \ge k^2 \quad \Rightarrow \quad \frac{np(1-p)}{\max(p^2,(1-p)^2)} \ge k^2$$

Thus our rule of thumb $np(1-p) > k^2$ (with $k^2 = 10$ or 9) implies that most of the approximating normal distribution falls within the range $\{0,1,2,\ldots,n\}$ of S_n.

- **A Normal Approximation Example:**
  Suppose we flip a fair coin $n = 10,000$ times. What is the chance that the number of heads is between 4900 and 5075?
  Let S be the number of observed heads then $S \sim \text{Bin}(n = 10000, p = .5)$ and $E(S) = 5000$ and $\text{var}(S) = 2500$. Thus

$$\begin{aligned} P(4900 \le S \le 5075) &= P\left(\frac{4900-5000}{\sqrt{2500}} \le \frac{S-5000}{\sqrt{2500}} \le \frac{5075-5000}{\sqrt{2500}}\right) \\ &\approx P(-2 \le Z \le 1.5) = \Phi(1.5) - \Phi(-2) = \Phi(1.5) - (1-\Phi(2)) \\ &\approx .9332 - (1-.9772) = .9104 \end{aligned}$$

  The second $\approx$ acknowledges the rounding error of the Table in Appendix E. From R we get `pnorm(1.5)-pnorm(-2) = 0.9104427`

- **Example with Hidden Binomial:**
  Suppose in a game we roll a fair die at each turn. When we roll a 1 or 2 we move out token 2 steps forward. When we roll a 3,4,5,6 we move the token 3 steps forward. Evaluate the chance that our token will have moved at least 320 steps in $n = 125$ rolls of the die.
  Let Y_n be the number of steps taken after n rolls. This is not a binomial random variable, but it can be related to one, namely the number of times S_n that we get a 1 or a 2. Then $S_n \sim \text{Bin}(n, p = 1/3)$ and we have $Y_n = S_n \cdot 2 + (n - S_n) \cdot 3 = 3 \cdot n - S_n$. Thus

$$\begin{aligned} P(Y_n \ge 320) &= P(3 \cdot n - S_n \ge 320) = P(S_n \le 375 - 320) = P\left(\frac{S_n - n \cdot \frac{1}{3}}{\sqrt{n \cdot \frac{1}{3} \cdot \frac{2}{3}}} \le \frac{55 - n \cdot \frac{1}{3}}{\sqrt{n \cdot \frac{1}{3} \cdot \frac{2}{3}}}\right) \\ &\approx \Phi\left(\frac{55 - 125/3}{\sqrt{250/9}}\right) = \Phi\left(\frac{165-125}{\sqrt{250}}\right) = \Phi(2.529822) = .9943 \end{aligned}$$

  R gives `pnorm(2.529822) = 0.994294`.

- **The Continuity Correction:**
  Since we are approximating the distribution of a discrete random variable (S_n) by the distribution of a continuous normal random variable we can run into trouble when approximating $P(k_1 \le S_n \le k_2)$ when $k_2 - k_1$ is small, in particular when $k_2 - k_1 = 0$. In that last case we would get

$$P(S_n = k) = P(k \le S_n \le k) \approx P\left(Z = \frac{k - np}{\sqrt{np(1-p)}}\right) = 0$$

  This does not mean that the CLT breaks down. In fact, $P(S_n = k) \to 0$ as $n \to \infty$, but the 0 approximation is not very satisfactory. The following trick makes for a better approximation. It is referred to as a continuity correction, correcting for the fact that we approximate the distribution of a discrete random variable (with probabilities on discrete points) by the distribution of a continuous random variable which calculates probabilities via areas and areas over discrete points are zero. In using the continuity correction we envision the probabilities of $P(S_n = k)$ as represented by areas of rectangles over $[k - .5, k + .5]$ with height $P(S_n = k)$. We then approximate the areas of such boxes by corresponding areas under the normal curve, i.e., we approximate areas with areas. This leads to much better results. Some examples will make this method clear.

Let $S_{72} \sim \text{Bin}(72, 1/6)$ count the number of sixes in 72 rolls of a fair die. Without continuity correction we would approximate $P(S_{72} = 6)$ by zero, when in fact it is 0.01990304 (via R). With continuity correction we proceed as follows

$$\begin{aligned} P(S_{72} = 6) &= P(5.5 \le S_{72} \le 6.5) \approx \Phi\left(\frac{6.5 - 72 \cdot \frac{1}{6}}{\sqrt{72 \cdot \frac{1}{6} \cdot \frac{5}{6}}}\right) - \Phi\left(\frac{5.5 - 72 \cdot \frac{1}{6}}{\sqrt{72 \cdot \frac{1}{6} \cdot \frac{5}{6}}}\right) \\ &= \Phi(-1.739253) - \Phi(-2.05548) = \Phi(2.05548) - \Phi(1.739253) \\ &= 0.98 - .9591 = 0.0209 \end{aligned}$$

or via R I get $\Phi(2.05548) - \Phi(1.739253) = 0.0210788$, which is certainly more reasonable than the zero without continuity correction. Note that $np(1-p) = 10$ almost satisfies the rule of thumb, by a hair. It certainly exceeds 9.

For this same situation let us use the continuity correction to approximate $P(12 \le X \le 18)$

$$\begin{aligned} P(12 \le X \le 18) &= P(11.5 \le X \le 18.5) \approx \Phi\left(\frac{18.5 - 72 \cdot \frac{1}{6}}{\sqrt{72 \cdot \frac{1}{6} \cdot \frac{5}{6}}}\right) - \Phi\left(\frac{11.5 - 72 \cdot \frac{1}{6}}{\sqrt{72 \cdot \frac{1}{6} \cdot \frac{5}{6}}}\right) \\ &= \Phi(2.0555) - (1 - \Phi(.1581)) = 0.5429 \end{aligned}$$

Without continuity correction we would have gotten

$$\begin{aligned} P(12 \le X \le 18) &\approx \Phi\left(\frac{18 - 72 \cdot \frac{1}{6}}{\sqrt{72 \cdot \frac{1}{6} \cdot \frac{5}{6}}}\right) - \Phi\left(\frac{12 - 72 \cdot \frac{1}{6}}{\sqrt{72 \cdot \frac{1}{6} \cdot \frac{5}{6}}}\right) \\ &= \Phi(6/\sqrt{10}) - \Phi(0) = 0.4711 \end{aligned}$$

The exact value via R is
$P(12 \le X \le 18) =$ `pbinom(18,72,1/6)-pbinom(11,72,1/6) = 0.5243775`.