Lecture 7

The binomial distribution, and a normal approximation

Consider n independent trials, each succeeds with probability p and fails with probability $1 - p$. A common problem that arises is to know what the chances are that we have exactly k successes [and hence also exactly $n - k$ failures]. The following addresses that problem.

Theorem 1. *If $k = 0 \ldots, n$, then*

$$P(k \text{ successes}) = \binom{n}{k} p^k (1-p)^{n-k}. \tag{5}$$

Otherwise the probability is zero.

Definition 1. The distribution (5) [of probabilities] is called the *binomial distribution* with parameters n and p.

Proof. Let S_i denote the event that the ith trial leads to a success. Then,

$$P(k \text{ successes}) = P(S_1 \cap \cdots \cap S_k \cap S_{k+1}^c \cap \cdots S_n^c) + \cdots$$

where we are summing over all possible ways of distributing k successes and $n - k$ failures in n spots. By independence, each of these probabilities is $p^k(1-p)^{n-k}$. The number of probabilities summed is the number of ways we can distributed k successes and $n - k$ failures into n slots. That is, $\binom{n}{k}$. Therefore, the result follows. $\square$

Example 1 (Coin tossing and sex of children, p. 83 of Pitman)**.** This example is in two parts:

(1) Find the probability of getting four or more heads in six tosses of a fair coin.

If $P(k)$ denotes the probability of getting exactly k heads in six tosses, then $P(k) = \binom{6}{k}0.5^k(1-0.5)^{6-k} = \binom{6}{k}0.5^6$. Therefore,

$$P(4) + P(5) + P(6) = \binom{6}{4}0.5^6 + \binom{6}{5}0.5^6 + \binom{6}{6}0.5^6 = 22 \times 0.5^6.$$

(2) What is the probability that among five families, each with six children, at least 3 of the families have four or more girls?

This is a binomial problem; each trial corresponds to one family having children [there are $n = 5$ trials]; each success corresponds to a family having four or more girls [by the previous part, $p = 22 \times 0.5^6 = \frac{11}{32}$]. Therefore, the answer is

$$\begin{aligned} &P(3 \text{ successes}) + P(4 \text{ successes}) + P(5 \text{ successes}) \\ &\quad = \binom{5}{3}\left(\frac{11}{32}\right)^3\left(\frac{21}{32}\right)^2 + \binom{5}{4}\left(\frac{11}{32}\right)^4\left(\frac{21}{32}\right)^1 + \binom{5}{5}\left(\frac{11}{32}\right)^5\left(\frac{21}{32}\right)^0 \\ &\quad = 10\left(\frac{11}{32}\right)^3\left(\frac{21}{32}\right)^2 + 5\left(\frac{11}{32}\right)^4\left(\frac{21}{32}\right) + \left(\frac{11}{32}\right)^5. \end{aligned}$$

Here is a generalization of the preceding example: It follows from (5) that for every $a \le b$ with $0 \le a, b \le n$,

$$P\left\{\text{No. of successes is somewhere between } a \text{ and } b\right\} = \sum_{j=a}^{b}\binom{n}{j}p^j(1-p)^{n-j}.$$

If n is large, then it is frequently onerous to compute the preceding sum. Later on, we will see the following remarkable fact:

Theorem 2 (The De Moivre–Laplace central limit theorem)**.** *Suppose $0 < p < 1$ is fixed. Then, as $n \to \infty$,*

$$P\left\{\textit{Between } a \textit{ and } b \textit{ successes}\right\} \approx \Phi\left(\frac{b-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a-np}{\sqrt{np(1-p)}}\right),$$

where Φ [read as capital "Phi"] is the "standard normal c.d.f [cumulative distribution function],"

$$\Phi(z) := \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,dz \qquad \textit{for all } -\infty < z < \infty.$$

Next we learn to use this theorem; we will learn to understand its actual meaning later on.

Normal distributions

Given two numbers $-\infty < \mu < \infty$ and $\sigma > 0$, the *normal curve* [with parameters μ and σ] is described by the function

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(x-\mu)^2/(2\sigma^2)} \qquad \text{for } -\infty < x < \infty.$$

For mainly historical reasons, special emphasis is paid to the case that $\mu = 0$ and $\sigma = 1$. In that case, we have a socalled *standard normal curve*, and reserve the symbol ϕ [small "phi"]. Note that the function Φ that we saw earlier can be computed from ϕ as follows:

$$\Phi(z) = \int_{-\infty}^{z} \phi(x)\,dx \qquad \text{for all } -\infty < z < \infty.$$

And, thanks to the fundamental theorem of calculus, ϕ can also be computed from Φ as follows:

$$\Phi'(z) = \phi(z) \qquad \text{for all } -\infty < z < \infty.$$

Note that areas under the normal curve with parameters μ and σ can always be transformed to [other] areas under the standard normal curve. In fact a change of variables tells us that for all $-\infty \le a \le b \le \infty$,

$$\begin{aligned} \int_a^b f(x)\,dx &= \int_a^b \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(x-\mu)^2/(2\sigma^2)}\,dx \\ &= \int_{(a-\mu)/\sigma}^{(b-\mu)/\sigma} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\,dz \qquad [z = (x-\mu)/\sigma] \\ &= \int_{(a-\mu)/\sigma}^{(b-\mu)/\sigma} \phi(z)\,dz \\ &= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right). \end{aligned} \tag{6}$$

In other words, the De Moivre–Laplace central limit theorem tells us that *if n is large, then the binomial probability of having between a and b successes is approximately equal to the area between a and b under the normal curve with parameters $\mu := np$ and $\sigma := \sqrt{np(1-p)}$.*

Equation (6) is called *standardization*. It tells us that in order to know how to compute areas under a normal curve, it suffices to know how to compute areas under a standard normal curve. Unfortunately, a theorem of Liouville tells us that $\Phi(z)$ cannot be computed [in terms of other "nice" functions]. However, it is possible to check that $\Phi(-\infty) = 0$, $\Phi(0) = \frac{1}{2}$, and $\Phi(\infty) = 1$. [Of course, by $\Phi(\pm\infty)$ we really mean $\lim_{z\to\pm\infty} \Phi(z)$!]

Clearly, $\Phi(-\infty) = 0$. And since the function ϕ is symmetric, $\Phi(\infty) = 2\Phi(0)$ [plot ϕ!]. Therefore, it suffices to prove the following:

Theorem 3. $\Phi(\infty) = 1$.

Proof. Note that

$$\begin{aligned}
\left[\Phi(\infty)\right]^2 &= \int_{-\infty}^{\infty} \frac{e^{-x^2/2}}{\sqrt{2\pi}}\,dx \cdot \int_{-\infty}^{\infty} \frac{e^{-y^2/2}}{\sqrt{2\pi}}\,dy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{e^{-(x^2+y^2)/2}}{2\pi}\,dx\,dy \\
&= \int_0^{2\pi}\int_0^{\infty} \frac{e^{-r^2/2}}{2\pi}\,r\,dr\,d\theta.
\end{aligned}$$

After a change of variables [$u = r^2/2$], the inside integral is seen to be

$$\int_0^{\infty} \frac{e^{-u}}{2\pi}\,du = \frac{1}{2\pi}.$$

Plug this into the outside integral to see that $[\Phi(\infty)]^2 = 1$, whence $\Phi(\infty) = 1$ [since $\Phi(\infty) \geq 0$]. $\square$

As I mentioned earlier, $\Phi(z)$ cannot be computed exactly for any value of z other than $z = 0, \pm\infty$. Therefore, people have approximated and tabulated $\Phi(z)$ for various choices of z, using standard methods used for approximating integrals; see Appendix 5 of your text.

Here are some consequences of that table [check!!]:

$$\Phi(0.09) \approx 0.5359, \quad \Phi(0.90) \approx 0.8159, \quad \Phi(3.35) \approx 0.9996.$$

And because ϕ is symmetric, $\Phi(-z) = 1 - \Phi(z)$. Therefore [check!!],

$$\Phi(-0.09) = 1 - \Phi(0.91) \approx 1 - 0.8186 = 0.1814, \quad \text{etc.}$$

Example 2. *A certain population is comprised of half men and half women. In a random sample of 10,000 what is the chance that the percentage of the men in the sample is somewhere between 49% and 51%?*

The exact answer to this question is computed from a binomial distribution with $n = 10,000$ and $p = 0.5$. We are asked to compute

$$P\left\{\text{between 4900 and 5,100 men}\right\} = \sum_{j=4900}^{5100} \binom{10000}{j} \left(\frac{1}{2}\right)^j \left(1 - \frac{1}{2}\right)^{10000-j}.$$

Because $np = 5000$ and $\sqrt{np(1-p)} = \sqrt{50}$, the normal approximation [Theorem 2] yields the following which turns out to be a quite good approximation:

$$\begin{aligned}
P\{\text{between 4900 and 5100 men}\} &\approx \Phi\left(\frac{5100-5000}{50}\right) - \Phi\left(\frac{4900-5000}{50}\right) \\
&= \Phi(2) - \Phi(-2) = \Phi(2) - \{1 - \Phi(2)\} \\
&= 2\Phi(2) - 1 \\
&\approx (2 \times 0.9772) - 1 = 0.9544.
\end{aligned}$$

In other words, the chances are approximately 95.44% that the percentage of men in the sample is somewhere between 49% and 51%. This phenomena is generally referred to as the *law of large numbers*: In a large sample, the probability is nearly one that the percentage of the men in the sample is quite close to the percentage of men in the population [i.e., with high probability, random sampling works well for large sample sizes].