MA181 INTRODUCTION TO STATISTICAL MODELLING
THE PROBABILITY GENERATING FUNCTION

Definition Suppose X is a discrete random variable that can take only non-negative integer values, i.e.X has range R that is a subset of $\{0, 1, 2, \ldots\}$; X might be the number of piglets in a litter or the number of letters you receive in the post today. In fact, the large majority of discrete random variables met in practice have a range of this type. If $p(x)$ is the probability function of X, consider the function $G(s)$ defined by

$$G(s) = \sum_{x \in R} s^x p(x) = E(s^X), \tag{1}$$

where the real variable s is a mathematical variable, as opposed to a random one, and, in a context such as this, is described as a *dummy variable* since it has no interpretation in terms of the problem being discussed. In order to see the use to which $G(s)$ might be put, consider a single example.

Example 1 A builder has completed a small development of three houses, and, from all the available evidence, the number X of then that he will sell within six months has the distribution given by

x	0	1	2	3
$p(x)$	$\frac{1}{16}$	$\frac{5}{16}$	$\frac{7}{16}$	$\frac{3}{16}$

Since X is a discrete random variable with an appropriate range, the function $G(s)$, defined at (1), is given by

$$\begin{aligned} G(s) &= \sum_{x=0}^{3} s^x p(x) \\ &= (s^0 + 5s^1 + 7s^2 + 3s^3)/16 \\ &= (1+s)^2(1+3s)/16. \end{aligned}$$

Thus $G(s)$ can be expressed as a factorised polynomial in S. Since a basic grounding in mathematics equips us to handle such expressions with relative ease, we should not be surprised to learn that, within this function, we have a convenient method for manipulating the distribution of X and deriving its properties.

Note that, for any distribution, $G(1) = 1$.

Inversion of $G(s)$ So much for the definition of $G(s)$; suppose now we are given a $G(s)$ like, for instance, the one derived in Example 1. By expanding this as a polynomial or power series in s, and picking out the coefficients, we can reconstruct the probability function of X; the value of $p(x)$ is found as the coefficient of s^x. For this reason, the function $G(s)$, defined at (1), is called the *probability generating function* (pgf) of X.

Example 2 If $G(s) = (4-3s)^{-1}$ is the pgf of a random variable, what is its distribution? Note that $G(1) = 1$. Expanding $G(s)$ as a power series in s, we obtain

$$G(s) = \frac{1}{4}\left[1 - \left(\frac{3s}{4}\right)\right]^{-1} = \frac{1}{4}\left[1 + \left(\frac{3s}{4}\right) + \left(\frac{3s}{4}\right)^2 + \ldots\right],$$

all the coefficients of which lie in $[0, 1]$. Calling the random variable X, we can find its probability function by extracting these coefficients. The result can be summarised by the formula

$$p(x) = \frac{1}{4}\left(\frac{3}{4}\right)^x, \; x = 0, 1, 2, \ldots$$

Moments Consider next the derivatives of $G(s)$ with respect to s. The first derivative is, from (1),

$$G'(s) = \frac{\partial G(s)}{\partial s} = \sum_x s^{x-1} x p(x).$$

Setting $s = 1$ in this expression leads to

$$G'(1) = \sum_x x p(x) = E(X). \tag{2}$$

Hence, if we know the pgf of a distribution, we can derive its mean by the use of this result. Note carefully the meaning of $G'(1)$: the pgf is first differentiated with respect to s, and only after that is s set equal to one.

Differentiating $G(s)$ a second time leads to

$$G''(s) = \sum_x s^{x-2} x(x-1) p(x),$$

so that, on putting $s = 1$, we find

$$G''(1) = \sum_x x(x-1)p(x) = E[X(X-1)].$$

This is the second factorial moment of W, denoted by $\mu_{[2]};$, from which the variance of X can be readily derived by using

$$\text{var}(X) = E[X(X-1)] + E(X) - [E(X)]^2 = G''(1) + G'(1) - [G'(1)]^2. \tag{3}$$

Continuing this process, if we differentiate $G(s)$ r times and then set $s = 1$, we find

$$G^{(r)}(1) = E[X(X-1)(X-2)\dots(X-r+1)] = \mu'_{[r]},$$

which is the rth factorial moment of X. The rth moment of X about the origin, or about its mean, can be derived by writing down a suitable function of that moment in terms of the first r factorial moments, as we have just seen for the variance. (Note that $\mu'_{[1]} = E(X)$.)

Example 3 Let us return to example 1, where we found that

$$G(s) = (1+s)^2(1+3s)/16$$

Differentiating with respect to s yields

$$G'(s) = [3(1+s)^2 + 2(1+s)(1+3s)]/16 = (1+s)(5+9s)/16.$$

Hence, from (2), the expected number of houses the builder will sell is

$$E(X) = G'(1) = \frac{7}{4}.$$

Differentiating $G(s)$ a second time leads to

$$G''(s) = [9(1+s) + (5+9s)]/16 = (7+9s)/;16.$$

Hence $G''(s) = 2$ and, using (3), the variance of X is found to be

$$\text{var}(X) = G''(1) + G'(1) - [G'(1)]^2 = 2 + \frac{7}{4} - \left(\frac{7}{4}\right)^2 = \frac{11}{16}.$$

Linear functions of random variables Suppose $G_X(s)$ is the pgf of the random variable X and we wish to know the pgf $G_Y(s)$ of the linear function $Y = a + bX$. From (1), we have

$$G_Y(s) = E(s^Y) = E(s^{a+bX} = s^a E[(s^b)^X] = s^a G_X(s^b)$$

so that $G_Y(s)$ can be found directly from $G_X(s)$.

Example 4 Let X follow the distribution of Example 2 and let $Y = 5+3X$. The the pgf $G_Y(s)$ of Y is given by

$$G_Y(s) = s^5 G_X(s^3) = \frac{1}{4}s^5\left[1 - \left(\frac{3s^3}{4}\right)\right]^{-1} = \frac{1}{4}s^5\left[1 + \frac{3}{4}s^3 + \left(\frac{3}{4}\right)^2 s^6 + \ldots\right].$$

Extracting the coefficient of s^y, we find

$$p_Y(y) = \frac{1}{4}\left(\frac{3}{4}\right)^{\frac{(y-5)}{3}}, \; y = 5, 8, 11, \ldots$$

Binomial distribution revisited Suppose the random variable Y follows a binomial $b)n\pi)$ distribution with probability function

$$p_Y(y) = \left(\begin{array}{c} n \\ y \end{array}\right)\pi^y(1-\pi)^{n-y}, \; y = 0, 1, \ldots, n.$$

The pgf of Y is given by

$$G_Y(s) = \sum_{y=0}^{N} s^y p_Y(y) = \sum_{y=0}^{n}\left(\begin{array}{c} n \\ y \end{array}\right)(\pi s)^y(1-\pi)^{n-y} = [(1-\pi) + \pi s]^n.$$

The mean and variance of Y can be found by differentiating this function. The first derivative with respect to s is

$$G'_Y(s) = n\pi[(1-\pi) + \pi s]^{n-1},$$

so that

$$E(Y) = G'_Y(1) = n\pi.$$

The second derivative of $G_Y(s)$ is given by

$$G''_Y(s) = n(n-1)\pi^2[(1-\pi)+\pi s]^{n-2}.$$

From (3), we therefore find the variance of Y to be

$$\begin{aligned} \text{var}(Y) &= G''_Y(1) + G'_Y(1) - [G'_Y(1)]^2 \\ &= n(n-1)\pi^2 + n\pi - (n\pi)^2 \\ &= n\pi(1-\pi). \end{aligned}$$

By differentiating $G_Y(s)$ twice more, we could go on to evaluate the skewness and kurtosis of the distribution, although the calculations become a little tedious.

Poisson distribution Under certain conditions, the binomial distribution converges to a different distribution as a limiting form. We have seen that the pgf of a binomial variate is

$$G_Y(s) = [(1-\pi)+\pi s]^n.$$

Consider the limit as $n \to \infty$ and $\pi \to 0$ in such a way as the product $n\pi$ remains constant. In view of the restriction, there is only a single limit at work here, and it is convenient, in this analysis, to set $n\pi = \mu$ ($n\pi$ is, after all, the mean of the binomial distribution), to write $\pi = \frac{\mu}{n}$, and to consider the limit of $G_Y(s)$ as $n \to \infty$. Then we have

$$\lim_{n\to\infty} G_Y(s) = \lim_{n\to\infty} \left[1 + \mu\frac{(s-1)}{n}\right]^n = e^{\mu(s-1)}. \tag{4}$$

If Y is a random variable with this pgf, its probability function can be found by expanding (4) as a power series in s. Since $e^{\mu s} = 1 + \frac{\mu s}{1!} + \frac{(\mu s)^2}{2!} + \ldots$, this leads to

$$p_Y(y) = \frac{e^{-\mu}\mu^y}{y!},\ y = 0, 1, 2, \ldots, \tag{5}$$

which is the probability function of the *Poisson distribution*, named after the French mathematician Siméon Denis Poisson (1781-1840). Note that the range of Y is infinitely large as we have taken the limit as

$n \to \infty$. And, in view of the way the distribution has been derived, it will be of no surprise to learn that it is particularly applicable in situations where a large number (n) of items or individuals are 'at risk ', each with a small probability (π) of producing an event, the variate Y measuring the total number of events over an interval of time or an area of space. If, for example, Y is the number of particle emissions per unit time from a piece of radioactive material, then n is not only very large but also unknown, and, because π (also unknown) is extremely small, the Poisson distribution fits data collected from such a source very closely.

The mean and variance of Y can be found by using (2) and (3). Thus, since

$$G'_Y(s) = \mu e^{\mu(s-1)} \text{ and } G''_Y(s) = \mu^2 e^{\mu(s-1)},$$

we have that

$$E(Y) = \mu \text{ and } \text{var}(Y) = \mu^2 + \mu - \mu^2 = \mu.$$

These values can also be derived by taking the appropriate limits of the mean and variance of the binomial distributions or, of course, from the probability function (5), using the appropriate summations. Being the mean value of Y, μ is a suitable symbol to use for the parameter of the distribution.