MA181 INTRODUCTION TO STATISTICAL MODELLING
BINOMIAL DISTRIBUTION

Bernoulli distribution Let X be a random variable with probability function

$$p_X(x) = \begin{cases} \pi, & x = 1, \\ 1 - \pi, & x = 0. \end{cases}$$

Then X follows a *Bernoulli distribution.*

Examples

1. The toss of a coin: $x = 1$ if a head shows, $x = 0$ if a tail.
2. The birth of a baby: $x = 1$ if a girl, $x = 0$ if a boy.
3. Testing items from a factory: $x = 1$ if defective, $x = 0$ if good.
4. Generally: $x = 1$ is called success, $x = 0$ failure.

If $X_1, X_2, \ldots, X_n$ $(n \geq 2)$ are independent and identically distributed (iid) random variables following a Bernoulli distribution, then they constitute a sequence of *Bernoulli trials*.

Binomial distribution Let X_1 and X_2 be a sequence of two Bernoulli trials and let $Y = X_1 + X_2$. What is $P(Y = y)$? Since Y can take only the three values 0,1 and 2, we have

$$P(Y = 0) = P(X_1 = 0 \text{ and } X_2 = 0) = (1 - \pi)^2,$$

$$P(Y = 1) = P[(X_1 = 0 \text{ and } X_2 = 1) \text{ or } (X_1 = 1 \text{ and } X_2 = 0)] = 2\pi(1-\pi),$$

$$P(Y = 2) = P(X_1 = 1 \text{ and } X_2 = 1) = \pi^2.$$

Generally, let $Y = X_1 + X_2 + \ldots + X_n$. Then

$$P(Y = 0) = P(X_1 = 0, X_2 = 0, \ldots X_n = 0) = (1 - \pi)^n,$$

$$P(Y = n) = P(X_1 = 1, X_2 = 1, \ldots, X_n = 1) = \pi^n \text{ and}$$

$$\begin{aligned} P(Y=y) &= P[\text{a particular sequence of } y1\text{'s and } (n-y)0\text{'s}] \times \\ & \quad \text{Number of such sequences} \\ &= \binom{n}{y} \pi^y (1-\pi)^{n-y}, \ y = 1, 2, \ldots, n-1. \end{aligned}$$

The random variable Y is said to follow a *binomial distribution* since the terms of its probability function derive from the binomial expansion of $[\pi + (1-\pi)]^n$. If 0! is set, by convention, to one, then the probability function of Y can be written as

$$P_Y(y) = \binom{n}{y} \pi^y (1-\pi)^{n-y}, \ y = 0, 1, \ldots, n,$$

since this then gives the correct probabilities for the cases $y = 0$ and $y = n$.

Notation If Y has this probability function, then we write $Y \sim b(n\pi)$.

Distribution function The cumulative distribution function $F_Y(y) = P(Y \leq y)$ is given by

$$F_Y(y) = \sum_{r=0}^{y} p_Y(r) = \sum_{r=0}^{y} \binom{n}{r} \pi^r (1-\pi)^r,$$

which cannot be simplified further.

Example The probability that a child is born with an inherited disease (cystic fibrosis), given that both parents are normal carriers of the associated gene, is $\frac{1}{4}$. If Y is the number of affected children in a family of six children, then $Y \sim b\left(6, \frac{1}{4}\right)$. Hence, the probability if two affected children is given by

$$P_Y(2) = \binom{6}{2} = \left(\frac{1}{4}\right)^2 \left(\frac{3}{4}\right)^4 = 15 \times \frac{3^4}{4^6} = 0.2966.$$

Further,

$$
\begin{aligned}
P(Y \leq 2) &= p_Y(0) + p_y(1) + p_Y(2) \\
&= \begin{pmatrix} 6 \\ 0 \end{pmatrix} \left(\frac{1}{4}\right)^0 \left(\frac{3}{4}\right)^6 + \begin{pmatrix} 6 \\ 1 \end{pmatrix} \left(\frac{1}{4}\right)^1 \left(\frac{3}{4}\right)^5 + \begin{pmatrix} 6 \\ 2 \end{pmatrix} \left(\frac{1}{4}\right)^2 \left(\frac{3}{4}\right)^4 \\
&= 0.1780 + 0.3560 + 0.2966 = 0.8306.
\end{aligned}
$$

Tables **(i)** If $Y \sim b(10, 0.45)$, then $P(Y \leq 3) = 0.2660$,

(ii) If $Y \sim b(16, 0.32)$, then $P(Y = 6) = P(Y \leq 6) = P(Y \leq 5) = 0.7743 = 0.5926 = 0.1817$,

(iii) If $Y \sim b(13, 0.18)$, then $P(\geq 4) = 1 - P(Y \leq 3) = 1 - 0.8061 = 0.1939$,

(iv) If $T \sim b(17, 0.403)$, then, by linear interpolation, $P(Y \leq 5) = 0.2639 + 0.3(0.2372 - 0.2639) = 0.2639 - 0.0080 = 0.2559$.

Properties 1. $\displaystyle\sum_{y=0}^{n} \begin{pmatrix} n \\ y \end{pmatrix} \pi^y (1-\pi)^{n-y} = [\pi + (1-\pi)]^n = 1$,

2. Let $Y' = n - Y$, where $Y \sim b(n\pi)$. Then

$$
\begin{aligned}
P(Y' = y') &= P(n - Y = y') = P(Y = n - y') \\
&= \begin{pmatrix} n \\ n - y' \end{pmatrix} \pi^{n-y'} (1-\pi)^{y'} \\
&= \begin{pmatrix} n \\ y' \end{pmatrix} (1-\pi)^{y'} \pi^{n-y'}, \ y' = 0, 1, \ldots, n.
\end{aligned}
$$

So $Y' \sim b(n, 1-\pi)$.
This result is often useful if $\pi > \frac{1}{2}$, for which tables are not generally available, since $p_Y(y) = p_{Y'}(n-y)$ and $P(Y \leq y) = P(Y' \geq n - y)$, where the success probability for the distribution of Y' is $1 - \pi$.

3. Suppose $\pi = \frac{1}{2}$. Then

$$
P_Y(y) = \begin{pmatrix} n \\ y \end{pmatrix} \left(\frac{1}{2}\right)^n = \begin{pmatrix} n \\ n - y \end{pmatrix} \left(\frac{1}{2}\right)^n = p_Y(n-y)
$$

for all values of Y. Hence the distribution is symmetric.

Tables (continued) **(v)** If $Y \sim b(14, 0.68)$, then $P(Y \leq 9) = P(Y' \geq 5) = 1 - P(Y' \leq 4)$, where $Y' \sim b(19, 0.32)$. So $P(Y \leq 9) = 1 - 0.5187 = 0.4813$.

Estimation Suppose a sequence of n Bernoulli trials yields y successes. Then the natural, and in many respects the best, estimate of π, the success probability, is the observed proportion of successes $\frac{y}{n}$.

Example (Multiple observations) The table below gives, in its second columns, the frequency distribution of the number Y of peas found in the pod of a four-seeded line of pea. A total of 269 pods were inspected.

Peas per pod y	observed frequency of pods	$\hat{p}_Y(y)$	Expected frequency of pods
0	16	0.0399	10.74
1	45	0.1976	53.15
2	100	0.3666	98.62
3	82	0.3023	81.33
4	26	0.0935	25.15
Total	269	0.9999	268.99

We will assume that $Y \sim b(4\pi)$ and estimate π by the average proportion of successes per pod, i.e. by

$$\hat{\pi} = \frac{16\left(\frac{0}{4}\right) + 45\left(\frac{1}{4}\right) + 82\left(\frac{3}{4}\right) + 26\left(\frac{4}{4}\right)}{269} = 0.5530.$$

Substituting the value into the probability function of Y yields the estimated probability function given by

$$\hat{p}_Y(y) = \begin{pmatrix} 4 \\ y \end{pmatrix} (0.5530)^y (0.4470)^{4-y}, \; y = 0, 1, 2, 3, 4.$$

The values of this function are shown in the third columns of the table. multiplying them by 269 gives the expected frequencies, for $y = 0, 1, 2, 3, 4$, which may be compared with the observed frequencies to determine how good a fit the binomial distribution is to the data. These values are shown in the last column of the table.