5 The binomial distribution

Say you roll a dice five times, what is the chance you get exactly three sixes? There are three parts to working out the probability. First consider the chance of getting three sixes in the first three rolls. This is $(1/6)^3$. Next consider the change of getting not-a-six in the next two rolls. This is $(5/6)^2$. Finally, we are not just interested in getting three sixes followed by two not-sixes: we want to count all the ways to get exactly three sixes out of five rolls. This means we also need to count the number of ways of choosing which three of the five rolls gives a six. Putting all this together and using R to denote the number of sixes

$$p(R=3) = \left(\begin{array}{c} 5 \\ 3 \end{array} \right) \left(\frac{1}{6}\right)^3 \left(\frac{5}{6}\right)^2 = 0.032 \tag{1}$$

This is an example of the **binomial distribution**; the binomial distribution is important because it describes the result of multiple independent trials. In a **binomial experiment**

- There are n identical trials.
- Each trial has one of two outcomes, which we call success, S, and failure, F.
- The trials are independent.
- The random variable of interest, say R, is the total number of successes.

Lets call the chance of success for an individual trial p and the probability of failure $q = 1 - p$. By reasoning similar to that used in the example above, it is easy to see

$$p_R(r) = \left(\begin{array}{c} n \\ r \end{array} \right) p^r q^{n-r} \tag{2}$$

Examples are plot in Fig. 1.

Say a student is doing a multiple choice vocabulary test in a language that is very different from the one they speak or any they have never learned, so they guess all the questions. If there are four options for each question and fifteen questions, lets calculate $p(5)$, the chance they get five correct:

$$p(R=5) = \left(\begin{array}{c} 15 \\ 5 \end{array} \right) (0.25)^5 (0.75)^{10} = 0.165 \tag{3}$$

Hopefully you will have seen the binomial coefficient before in the expansion of polynomials:

$$(q+p)^n = q^n + nq^{n-1}p + \left(\begin{array}{c} n \\ 2 \end{array} \right) q^{n-2}p^2 + \ldots + p^n = \sum_{r=0}^{n} \left(\begin{array}{c} n \\ r \end{array} \right) p^r q^{n-r} \tag{4}$$

Applied to our case, where $q = 1 - p$, the left hand side is one; the right hand side is a sum over the probabilities.

$$1 = \sum_{r=0}^{n} \left(\begin{array}{c} n \\ r \end{array} \right) p^r q^{n-r} = \sum_{r=0}^{n} p(R=r) \tag{5}$$

This is what you would expect, the probabilities should add to one. However, this formula leads to a very neat trick, first:

$$1 = \sum_{r=0}^{n} \left(\begin{array}{c} n \\ r \end{array} \right) p^r q^{n-r} \tag{6}$$

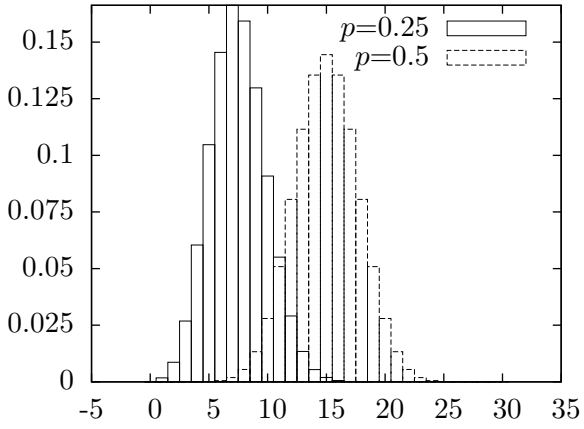

Figure 1: Binomial histograms. Here the value of $p_R(r)$ is plotted where R is a binomial with $n = 30$. In the left distribution $p = 0.25$ and in the right $p = 0.5$.

Now differentiate both sides with p, so take d/dp. The left side is a constant so it differentiates to zero and we have

$$0 = \sum_{r=0}^{n} \binom{n}{r} r p^{r-1} q^{n-r} - \sum_{r=0}^{n} \binom{n}{r} p^r (n-r) q^{n-r-1} \tag{7}$$

where we have used the product rule, remembering that $q = 1 - p$. Next we multiply and divide by either p or q, it will be clear why in a few lines time:

$$0 = \frac{1}{p} \sum_{r=0}^{n} \binom{n}{r} r p^r q^{n-r} - \frac{1}{q} \sum_{r=0}^{n} \binom{n}{r} p^r (n-r) q^{n-r} \tag{8}$$

Now, n is just a constant so we can go

$$\sum_{r=0}^{n} \binom{n}{r} n p^r q^{n-r} = n \sum_{r=0}^{n} \binom{n}{r} p^r q^{n-r} = n \tag{9}$$

Once we deal with this part we see we have two terms that look the same except for the $1/p$ or $1/q$ at the start,

$$0 = \frac{n}{q} - \left(\frac{1}{q} + \frac{1}{p}\right) \sum_{r=0}^{n} \binom{n}{r} p^r r q^{n-r} \tag{10}$$

However, be definition

$$\langle R \rangle = \sum_{r=0}^{n} p(R = r) r = \sum_{r=0}^{n} \binom{n}{r} p^r r q^{n-r} \tag{11}$$

so

$$\left(\frac{1}{p} + \frac{1}{q}\right) \langle R \rangle = \frac{n}{q} \tag{12}$$

It only remains to note

$$\frac{1}{p} + \frac{1}{q} = \frac{q+p}{pq} = \frac{1}{pq} \tag{13}$$

to derive

$$\langle R \rangle = pn \tag{14}$$

A similar argument based on differentiating twice gives the standard deviation:

$$\sigma^2 = pqn \tag{15}$$

Summary

- In a **binomial experiment**
  1. There are n identical trials.
  2. Each trial has one of two outcomes, which we call success, S, and failure, F.
  3. The trials are independent.
  4. The random variable of interest, say R, is the total number of successes.
- In a binomial experiment, if p is the chance of success for an individual trial, and $q = 1-p$ is the chance of failure, then the probability of r successes is given by

$$p_R(r) = \left(\begin{array}{c} n \\ r \end{array} \right) p^r q^{n-r} \tag{16}$$

- The mean is np and the variance is npq.
- The mean is derived using a fancy trick involving differenciating

$$1 = \sum_{r=0}^{n} \left(\begin{array}{c} n \\ r \end{array} \right) p^r q^{n-r} = \sum_{r=0}^{n} p(R = r) \tag{17}$$

  with respect to p.

Example question

Show the variance of the binomial distribution is npq.
solution This is a hard question. As ever we start with

$$1 = Z = \sum_{r=0}^{n} \left(\begin{array}{c} n \\ r \end{array} \right) p^r q^{n-r} \tag{18}$$

and we take the second derivative

$$\begin{aligned} \frac{d^2 Z}{dp^2} &= \sum_{r=0}^{n} \left(\begin{array}{c} n \\ r \end{array} \right) r(r-1) p^{r-2} q^{n-r} - 2 \sum_{r=0}^{n} \left(\begin{array}{c} n \\ r \end{array} \right) r(n-r) p^{r-1} q^{n-r-1} \\ &\quad + \sum_{r=0}^{n} \left(\begin{array}{c} n \\ r \end{array} \right) (n-r)(n-r-1) p^r q^{n-r-2} \end{aligned} \tag{19}$$

so

$$0 = \sum_{r=0}^{n} \left[\frac{1}{p^2}(r^2 - r) - \frac{2}{pq}(rn - r^2) + \frac{1}{q^2}(n^2 - 2nr - n + r^2 + r) \right] \left(\begin{array}{c} n \\ r \end{array} \right) p^r q^{n-r} \tag{20}$$

or, using $\langle R \rangle = np$,

$$0 = \frac{1}{p^2}\langle R^2 \rangle - \frac{1}{p^2} np - \frac{2n}{pq} np + \frac{2}{pq}\langle R^2 \rangle + (n^2 - n)\frac{1}{q^2} - (2n-1)\frac{1}{q^2} np + \frac{1}{q^2}\langle R^2 \rangle \tag{21}$$

Putting this together we get

$$0 = \left(\frac{1}{p} + \frac{1}{q} \right)^2 \langle R^2 \rangle - \frac{n}{p} - \frac{2n^2}{q} + \frac{n^2 - n}{q^2} - \frac{(2n^2 - n)p}{q^2} \tag{22}$$

Simplifying we get

$$0 = \frac{\langle R^2 \rangle}{p^2 q^2} - \frac{n^2}{q^2} - \frac{n}{p} - \frac{n}{q} \tag{23}$$

so

$$\langle R^2 \rangle = n^2 p^2 + npq \tag{24}$$

and finally

$$\sigma^2 = \langle R^2 \rangle - \langle R \rangle^2 = npq \tag{25}$$