Statistics 2

Dr Oliver
Mathematics

Binomial distribution

This distribution is suitable when

- there is a fixed number of trials,
- there are only two outcomes (those that we regard as 'success' and 'failure'),
- the trials are independent,
- the probabilities of success and failure are constant.

If all of those conditions are met then we can model our random variable X as a binomial distribution and we use the notation $X \sim \mathrm{B}(n, p)$ where n is the number of trials and p is the probability of success. For any integer $0 \leq r \leq n$ the probability of achieving exactly r successes from n trials is

$$\mathrm{P}(X = r) = \binom{n}{r} p^r (1-p)^{n-r}.$$

You should be able to use both your calculator and the tables in your formula book in order to calculate binomial probabilities. The expectation and the variance of $X \sim \mathrm{B}(n, p)$ are found by

$$\mathrm{E}(X) = np$$
$$\mathrm{Var}(X) = np(1-p).$$

Poisson distribution

This distribution is suitable when

- events occur singly in space and time,
- the events are independent,
- the events occur at a constant rate, i.e., the average number of occurrences in an interval of time or space is proportional to the length of the interval.

If all of those conditions are met then we can model our random variable X as a Poisson distribution and we use the notation $X \sim \mathrm{Po}(\lambda)$ where λ is the constant rate of occurrence. For any integer $r \geq 0$, the probability of achieving exactly r outcomes is given by

$$\mathrm{P}(X = r) = \frac{e^{-r}\lambda^r}{r!}.$$

The expectation and the variance of $X \sim \mathrm{Po}(\lambda)$ are found by

$$\mathrm{E}(X) = \mathrm{Var}(X) = \lambda.$$

Approximating binomial by Poisson

If $X \sim \mathrm{B}(n, p)$ where n is 'large' and p is 'small' then X can be approximated by the Poisson distribution $\mathrm{Po}(np)$.

- If p is close to zero then $\mathrm{Var}(X) = np(1-p)$ will be close to np and so the mean and the variance of X will be similar and this is needed as a Poisson distribution has the same mean and variance.
- As a general rule of thumb, use $np \leq 10$ as a test for whether the Poisson distribution is likely to be a suitable approximation.
- If p is close to one then working with the related binomial distribution $Y \sim \mathrm{B}(n, 1-p)$, where we count failures rather than successes, gives a binomial distribution that can be approximated by the Poisson distribution $\mathrm{Po}(n(1-p))$.

Probability density functions

A function $\mathrm{f}(x)$, defined on the whole of the real line, is a *probability density function* if

- $\mathrm{f}(x) \geq 0$ for all $x \in \mathbb{R}$,
- $\int_{-\infty}^{\infty} \mathrm{f}(x)\,\mathrm{d}x = 1$.

If the random variable X has a probability density function $\mathrm{f}(x)$ then

$$\mathrm{P}(a < X < b) = \int_a^b \mathrm{f}(x)\,\mathrm{d}x.$$

We can calculate the mean and the variance of X using

$$\begin{aligned}\mathrm{E}(X) &= \int_{-\infty}^{\infty} x\,\mathrm{f}(x)\,\mathrm{d}x \\ \mathrm{Var}(X) &= \mathrm{E}(X^2) - [\mathrm{E}(X)]^2 \\ &= \int_{-\infty}^{\infty} x^2\,\mathrm{f}(x)\,\mathrm{d}x - \left[\int_{-\infty}^{\infty} x\,\mathrm{f}(x)\,\mathrm{d}x\right]^2.\end{aligned}$$

Cumulative distribution function

Given a probability density function $\mathrm{f}(x)$, we define the *cumulative distribution function* by

$$\mathrm{F}(x) = \int_{-\infty}^{x} \mathrm{f}(t)\,\mathrm{d}t.$$

Modes and quartiles

Given a continuous random variable X with probability density function $\mathrm{f}(x)$ and cumulative distribution function $\mathrm{F}(x)$,

- the mode is the value of the random variable X where it is 'most dense', i.e., where the probability density function $\mathrm{f}(x)$ has its maximum value.
- the lower quartile, Q_1, satisfies $\mathrm{F}(Q_1) = 0.25$,
- the median, Q_2, satisfies $\mathrm{F}(Q_2) = 0.5$,
- the upper quartile, Q_3, satisfies $\mathrm{F}(Q_3) = 0.75$,

Continuous uniform distribution

Given $a < b$, the *continuous uniform distribution* or *rectangular distribution* is defined to have a probability density function

$$\mathrm{f}(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

and it is denoted by $X \sim \mathrm{U}[a, b]$.

- $\mathrm{E}(X) = \frac{1}{2}(a+b)$,
- $\mathrm{Var}(X) = \frac{1}{12}(b-a)^2$,
- $\mathrm{F}(x) = \frac{x-a}{b-a}$, $a < x < b$.

Continuity correction

When approximating a discrete distribution (binomial or Poisson) by a continuous distribution (normal) then it is conventional to apply the *continuity correction* in order to improve the accuracy of the answer.

- Approximate $\mathrm{P}(X \leq n)$ by $\mathrm{P}(Y < n + 0.5)$,
- Approximate $\mathrm{P}(X \geq n)$ by $\mathrm{P}(Y \geq n - 0.5)$.

Approximating binomial by normal

If $X \sim \mathrm{B}(n, p)$ where n is 'large' and p is 'close' to 0.5 then X can be approximated by the normal distribution $\mathrm{N}(np, npq)$. Of the two conditions, it is $p \approx 0.5$ that is, in general, the more important.

Approximating Poisson by normal

If $X \sim \mathrm{Po}(\lambda)$ where λ is 'large' then X can be approximated by the normal distribution $\mathrm{N}(\lambda, \lambda)$.

Hypothesis testing

The *null hypothesis*, H_0, is the hypothesis that we assume to be true about a population's parameter until such time as there is evidence to the contrary; the alternative hypothesis, H_1, is a statement about a population's parameter if we reject the null hypothesis. The *level of significance* of the test is the (generally) small probability of something happening if the null hypothesis holds. The *critical region* is the set of all values whose probability of occurring is at most that set by the level of significance. The *critical values* are the extreme ends of the critical region. It may be that, for discrete distributions, the probability of being in the critical region is less than the level of significance. If the test gives a significant result, i.e., the probability of achieving such an outcome under the null hypothesis is less than the level of significance then we reject H_0: there is evidence at this level of significance to reject the null hypothesis (although the conclusion should be stated in terms of the original problem). If the test does not give a significant result, i.e., the probability of achieving such an outcome under the null hypothesis is not less than the level of significance then we fail to reject H_0: there is no evidence at this level of significance to reject the null hypothesis (again, the conclusion should be stated in terms of the original problem).

One-tailed test

H_0: the population parameter $\lambda = p$.
H_1: the population parameter $\lambda < p$ **or** $\lambda > p$.
In a one-tailed test there will be a single critical region and a single critical value.

Two-tailed test

H_0: the population parameter $\lambda = p$.
H_1: the population parameter $\lambda \neq p$.
In a two-tailed test there will generally be two critical regions and two critical values (one for each region). The critical regions are determined by halving the level of significance, e.g., if the level of significance is 10% then the usual approach is to allow for 5% at each tail.