11 Discrete Random variables

In this section we define the expectation and variance of a discrete random variable and deduce some basic properties.

Proposition 11.1. *Let X be a discrete random variable on a sample space S. Then*

$$\sum_{r \in Range(X)} \mathbb{P}(X = r) = 1$$

Proof The fact that the range of X is either finite or countably infinite means that we can use Kolmogorov's second and third axioms to deduce that

$$\sum_{r \in Range(X)} \mathbb{P}(X = r) = P(S) = 1$$

●

This proposition is a useful to check we have calculated the probability mass function of X correctly.

Definition The *expectation* of a discrete random variable X is

$$\mathrm{E}(X) = \sum_{r \in Range(X)} r\, \mathbb{P}(X = r)$$

Thus the expectation of X is a 'weighted average' of the values taken by X, where each value $r \in Range(X)$ is weighted by the probability it occurs.

Sometimes it is useful to consider functions of random variables. If X is a random variable on S and $g : \mathbb{R} \to \mathbb{R}$ then $Y = g(X)$ is a new random variable on X which maps S into $\mathbb{R}$ by $Y(s) = g(X(s))$ for all outcomes $s \in S$. For example $Y = 2X - 7$ or $Z = X^2$ are both new random variables on S.

Proposition 11.2 (Properties of expectation). *Let X be a discrete random variable.*

   (a) *Suppose $b, c \in \mathbb{R}$. Then $\mathrm{E}(bX + c) = b\mathrm{E}(X) + c$.*

   (b) *If $m, M \in \mathbb{R}$ and $m \leq X(s) \leq M$ for all $s \in S$ then*

$$m \leq \mathrm{E}(X) \leq M.$$

(c) If there exists $k \in \mathbb{R}$ with $\mathbb{P}(X = k - t) = \mathbb{P}(X = k + t)$ for all $t \in \mathbb{R}$ then $\mathrm{E}(X) = k$.

Proof (a) We have

$$\mathrm{E}(bX + c) = \sum_{r \in Range(X)} (br + c)\mathbb{P}(X = r)$$

$$= \sum_{r \in Range(X)} br\mathbb{P}(X = r) + \sum_{r \in Range(X)} c\mathbb{P}(X = r)$$

$$= b \sum_{r \in Range(X)} r\mathbb{P}(X = r) + c \sum_{r \in Range(X)} \mathbb{P}(X = r)$$

$$= b\mathrm{E}(X) + c$$

where the fourth equality uses the definition of $\mathrm{E}(X)$ for the first term and Proposition 11.1 for the second term.

(b) See Exercises 7, Q3.

(c) We have

$$\mathrm{E}(X) = \sum_{r \in Range(X)} r\mathbb{P}(X = r)$$

$$= k\mathbb{P}(X = k) + \sum_{\substack{t > 0 \\ k - t \in Range(X)}} [(k - t)\mathbb{P}(X = k - t) + (k + t)\mathbb{P}(X = k + t)]$$

$$= k\mathbb{P}(X = k) + \sum_{\substack{t > 0 \\ k - t \in Range(X)}} [k\mathbb{P}(X = k - t) + k\mathbb{P}(X = k + t)]$$

$$= k \sum_{r \in Range(X)} \mathbb{P}(X = r)$$

$$= k$$

where:

- the second and fourth equalities follow by putting $r = k - t$ when $r < k$ and $r = k + t$ when $r > k$;

- the third equality holds because $\mathbb{P}(X = k - t) = \mathbb{P}(X = k + t)$ so the terms involving t cancel;

- the fifth equality uses Proposition 11.1.

2

●

Definition Let X be a discrete random variable with $E(X) = \mu$. Then the *variance* of X is

$$\text{Var}(X) = \sum_{r \in Range(X)} [r - \mu]^2 \, \mathbb{P}(X = r).$$

Thus $\text{Var}(X)$ is the expected value of $[X - \mu]^2$ i.e. the square of the distance from X to $E(X)$. It measures how the probability distribution of X is concentrated about its expectation. A small variance means X is sharply concentrated and a large variance means X is spread out.

Proposition 11.3 (Properties of variance)**.** *Let X be a discrete random variable. Then*

(a) *$Var(X) \geq 0$.*

(b) *If $b, c \in \mathbb{R}$ then $\text{Var}(bX + c) = b^2 \, \text{Var}(X)$.*

Proof (a) Let $E(X) = \mu$. Then

$$\text{Var}(X) = \sum_{r \in Range(X)} [r - \mu]^2 \, \mathbb{P}(X = r) \geq 0$$

since $[r - \mu]^2 \geq 0$ and $\mathbb{P}(X = r) \geq 0$ for all $r \in RangeX$.

(b) We have $E(bX + c) = b\mu + c$ by Proposition 11.2. Thus

$$
\begin{aligned}
\text{Var}(bX + c) &= \sum_{r \in Range(X)} [(br + c) - (b\mu + c)]^2 \, \mathbb{P}(X = r) \\
&= b^2 \sum_{r \in Range(X)} [r - \mu]^2 \, \mathbb{P}(X = r) \\
&= b^2 \, \text{Var}(X)
\end{aligned}
$$

●

We close this section by deriving a useful formula for calculating variance.

Proposition 11.4. *Suppose X is a discrete random variable. Then $\text{Var}(X) = E(X^2) - E(X)^2$.*

3

Proof Let $\mathrm{E}(X) = \mu$. Then

$$
\begin{aligned}
\mathrm{Var}(X) &= \sum_{r \in Range(X)} [r - \mu]^2 \, \mathbb{P}(X = r) \\
&= \sum_{r \in Range(X)} [r^2 - 2r\mu + \mu^2] \, \mathbb{P}(X = r) \\
&= \sum_{r \in Range(X)} r^2 \, \mathbb{P}(X = r) - 2\mu \sum_{r \in Range(X)} r \, \mathbb{P}(X = r) + \mu^2 \sum_{r \in Range(X)} \mathbb{P}(X = r)
\end{aligned}
$$

We have:

- $\sum r^2 \, \mathbb{P}(X = r) = \mathrm{E}(X^2)$;

- $\sum r \, \mathbb{P}(X = r) = \mathrm{E}(X)$;

- $\sum \mathbb{P}(X = r) = 1$, by Proposition 11.1.

Hence
$$
\mathrm{Var}(X) = \mathrm{E}(X^2) - 2\mu\mathrm{E}(X) + \mu^2 = \mathrm{E}(X^2) - \mathrm{E}(X)^2
$$

since $\mathrm{E}(X) = \mu$. ●

12 Some Common Discrete Probability Distributions

Definition Given two discrete random variables X and Y we say that they have *the same probability distribution* if they have the same range and the same probability mass function i.e. $\mathbb{P}(X = r) = \mathbb{P}(Y = r)$ for all values r in the range. We use the notation $X \sim Y$ to mean X and Y have the same probability distribution.

Certain probability distributions occur so often that it is convenient to give them special names. In this section we study a few of these. In each case we will determine the probability mass function, expectation and variance, as well as describing the sort of situation in which the distribution occurs.

4

12.1 Bernoulli distribution

Suppose that a trial (experiment) with two outcomes is performed. Such a trial is called a *Bernoulli trial* and the outcomes are usually called *success* and *failure* with $\mathbb{P}(\text{success}) = p$, and hence $\mathbb{P}(\text{failure}) = 1 - p$. We define a random variable X by putting $X(\text{success}) = 1$ and $X(\text{fail}) = 0$. Then the probability mass function of X is

$$\begin{array}{c|cc} n & 0 & 1 \\ \hline P(X = n) & 1 - p & p \end{array}$$

We say that X has the *Bernoulli distribution* and write $X \sim \text{Bernoulli}(p)$.

Lemma 12.1. *If* $X \sim Bernoulli(p)$ *then*

   (a) $\mathrm{E}(X) = p$.

   (b) $\mathrm{Var}(X) = p(1 - p)$

Proof (a) Using the definition of $\mathrm{E}(X)$ we have

$$\mathrm{E}(X) = 0 \times (1 - p) + 1 \times p = p$$

(b) Using Proposition 11.4 we have

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2 = [0^2 \times (1 - p) + 1^2 \times p] - p^2 = p(1 - p)$$

$$\bullet$$

12.2 Binomial distribution

Suppose we perform n independent Bernoulli trials. Let X be the number of trials which result in success. Then X has the *binomial distribution*. We write $X \sim \text{Bin}(n, p)$. The range of X is $\{0, 1, 2, \ldots, n\}$ and X has probability mass function

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1 - p)^{n-k} \quad \text{for } 0 \leq k \leq n.$$

Lemma 12.2. *If* $X \sim Bin(n, p)$ *then*

(a) $\mathrm{E}(X) = np$.

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

Proof (a) Using the definition of $\mathrm{E}(X)$ we have

$$
\begin{aligned}
\mathrm{E}(X) &= \sum_{k \in Range(X)} k\, \mathbb{P}(X = k) \\
&= \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} \\
&= \sum_{k=1}^{n} k \frac{n!}{k!\,(n-k)!} p^k (1-p)^{n-k} \\
&= np \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!\,(n-k)!} p^{k-1}(1-p)^{n-k}
\end{aligned}
$$

where the third equality uses the fact that the $k = 0$ term in the summation is zero. We can now put $m = n - 1$ and $i = k - 1$ in the last equality to obtain

$$
\begin{aligned}
\mathrm{E}(X) &= np \sum_{i=0}^{m} \frac{m!}{i!\,(m-i)!} p^i (1-p)^{m-i} \\
&= np \sum_{i=0}^{m} \binom{m}{i} p^i (1-p)^{m-i} \\
&= np[p + (1-p)]^m \\
&= np
\end{aligned}
$$

where the third equality uses the Binomial Theorem.

(b) See Exercises 8. $\qquad\bullet$

12.3 Hypergeometric distribution

Suppose that we have B objects, R of which are red and the remaining $B - R$ of which are white. We make a selection of n objects without replacement. Let X be the number of red objects among the n objects chosen. We say that X has the *hypergeometric distribution* and write $X \sim \mathrm{Hg}(n, R, B)$. The

random variable X takes values in $\{0, 1, 2, \ldots, n\}$ and has probability mass function

$$\mathbb{P}(X = k) = \frac{\binom{R}{k}\binom{B-R}{n-k}}{\binom{B}{n}} \quad \text{for } 0 \le k \le n.$$

This follows from our results on sampling. There are $\binom{R}{k}$ ways of choosing the k red objects; for each of these there are $\binom{B-R}{n-k}$ ways of choosing the $n-k$ white objects; and the sample space has size $\binom{B}{n}$.

It can be shown that:

Lemma 12.3. *If $X \sim Hg(n, R, B)$ then*

(a) $\mathrm{E}(X) = n\left(\frac{R}{B}\right)$.

(b) $\mathrm{Var}(X) = n\left(\frac{R}{B}\right)\left(1 - \frac{R}{B}\right)\left(\frac{B-n}{B-1}\right)$

It is perhaps interesting to compare this with the analogous random variable Y we obtain when the sampling is done **with** replacement. We have $Y \sim \mathrm{Bin}(n, \frac{R}{B})$. The expectation of Y is the same as that of X but the variance differs by a factor of $\left(\frac{B-n}{B-1}\right)$; this is close to 1 when B is very large compared to n.

12.4   Geometric distribution

Suppose we perform an unlimited number of Bernoulli trials and let X be the number of trials up to and including the first success. In this case X has the *geometric distribution*. We write $X \sim \mathrm{Geom}(p)$. The random variable X has range $\{1, 2, 3, \ldots\}$ and has probability mass function

$$\mathbb{P}(X = k) = p(1 - p)^{k-1} \quad \text{for } k \ge 1.$$

Lemma 12.4. *If $X \sim Geom(p)$ then*

(a) $\mathrm{E}(X) = \frac{1}{p}$.

(b) $\mathrm{Var}(X) = \frac{1-p}{p^2}$

Proof (a) We have

$$\mathrm{E}(X) = \sum_{k=1}^{\infty} kp(1-p)^{k-1} = \sum_{r=0}^{\infty}(r+1)\, p(1-p)^r \tag{1}$$

where the second equality follows by putting $r = k - 1$. We also have

$$(p - 1)\mathrm{E}(X) = \sum_{k=1}^{\infty} kp(1 - p)^k = \sum_{r=0}^{\infty} rp(1 - p)^r \qquad (2)$$

where the second equality follows by putting $r = k$ and using the fact that the $r = 0$ term in the last summation is zero. Subtracting (2) from (1) gives

$$[1 - (p - 1)]\mathrm{E}(X) = \sum_{r=0}^{\infty} p(1 - p)^r = p\sum_{r=0}^{\infty}(1 - p)^r \qquad (3)$$

Now $[1 - (p - 1)]\mathrm{E}(X) = p\mathrm{E}(X)$ and $\sum_{r=0}^{\infty}(1 - p)^r = \frac{1}{1-(1-p)} = \frac{1}{p}$ by the formula for the sum of a geometric progression. Hence we have $p\mathrm{E}(X) = p\frac{1}{p} = 1$, so $\mathrm{E}(X) = 1/p$.

(b) See Exercises 8.   ●

12.5   Poisson distribution

Suppose that incidents of some kind occur at random times but at an average rate λ incidents per unit time. Let X be the number of these incidents which occur in a fixed unit time interval. (For example we could take X to be the number of emissions from a radioactive source in one particular second, or the number of phone calls received on a particular day.) In this case X has a *Poisson distribution*. We write $X \sim \mathrm{Poisson}(\lambda)$. The random variable X has range $\{0, 1, 2, \ldots\}$. It can be show that the probability mass function for X is

$$\mathbb{P}(Y = k) = e^{-\lambda}\frac{\lambda^k}{k!} \quad \text{for } k \geq 0.$$

(See Remark at the end of this subsection for some justification of this.)

Lemma 12.5. *If $X \sim Poisson(\lambda)$ then*

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

  (b) $\mathrm{Var}(X) = \lambda$

Proof (a) Using the definition of $\mathrm{E}(X)$ we have

$$
\begin{aligned}
\mathrm{E}(X) &= \sum_{k \in Range(X)} k\, \mathbb{P}(X = k) \\
&= \sum_{k=0}^{\infty} k\, e^{-\lambda} \frac{\lambda^k}{k!} \\
&= e^{-\lambda} \sum_{k=1}^{\infty} k\, \frac{\lambda^k}{k!} \\
&= \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \\
&= \lambda e^{-\lambda} \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} \\
&= \lambda
\end{aligned}
$$

where:

- the third equality uses the fact that the $k = 0$ term in the summation is zero;

- the fifth equality uses the substitution $i = k - 1$;

- the last equality uses the Tayler expansion $e^{\lambda} = \sum_{i=0}^{\infty} \frac{\lambda^i}{i!}$ for any $\lambda \in \mathbb{R}$.

(b) See Exercises 8. ●

Remark The probability mass function for the Poisson distribution looks rather mysterious. It comes from considering a limit of probabilities coming from binomial distributions. Specifically, suppose that we divide the unit time interval we are interested in into n equal subintervals. If n is very large then the subintervals are very small so it is unlikely that there will be two or more incidents in any one of the subintervals. So lets assume that there is never more than 1 incident in a subinterval and that the probability that there is 1 incident is equal to λ/n. Let Y be the total number of incidents in the unit time interval when we make this assumption. Then $Y \sim \mathrm{Bin}(n, \lambda/n)$, so

$$
P(Y = k) = \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}.
$$

Of course the probability mass function of Y is not the same as that of X (since the Poisson distribution allows there to be 2 or more incidents in a subinterval). However if n is made larger and larger then the subintervals become smaller and smaller and the probability mass function of Y gives a better and better approximation to that of X. If we take the limit of the expression for $P(Y = k)$ as $n \to \infty$ then we get $e^{-\lambda}\frac{\lambda^k}{k!}$. This is the probability mass function of X. (We need techniques from Analysis to make this argument precise. These will be covered in the level 5 module Convergence and Continuity.)

12.6   Summary

The three most important discrete distribution are the binomial, geometric and Poisson distributions. Condense this section of the notes into half a page by writing your own summary of their properties (probability mass function, expectation, variance and when they occur). You could include the Bernoulli and Hypergeometric distributions also if you like.