MAT 2379, Introduction to Biostatistics

Chapter 6. Discrete Random Variables
6.1. Definition

The term "random" is not easy to define mathematically but is an important part of human activities, and in particular of the statistical analysis of data. A **random variable** X is a variable whose numerical value x is determined by chance.

We denote random variables with capital letters X, Y, Z, etc. and their values with small letters x, y, z, etc.

Random variables can be of two types: **discrete** (when they assume only a finite number of values) and **continuous** (when they assume an infinite number of values in some interval).

Examples of discrete random variables:
1. The sex of a newborn human infant; 2. The number of worker bees in a honeybee society; 3. The number of students enrolled in a graduate course; 4. The blood type of an individual; 5. The number of genes A in an offspring of two heterozygous individuals Aa;

Examples of continuous random variables:
1. The weight (or height) of an individual; 2. The blood pressure (or temperature) of a patient; 3. The weight gain of a woman during pregnancy.

When working with a discrete random variable X, one can not predict exactly the value that this variable will take. The best thing that one can do is to estimate the probability that the random variable X will take a particular value x.

The function f which gives these probabilities is called **the probability mass function** of X:

$$f(x) = P(X = x)$$

Note that if x is an impossible value for X, then $f(x) = 0$. Moreover, since $f(x)$ are probabilities, we should always have $0 \leq f(x) \leq 1$ and

$$\sum_x f(x) = 1$$

where the sum is taken over all possible values x for X.

Example 1. Childhood lead poisoning is a public health concern in the US. In a certain population one child in 10 has a high blood lead level (i.e. the probability that a randomly selected child in that population has high blood lead level is 0.1). Consider a randomly chosen group of 3 children from that population. Let X denote the number of children in the chosen group that have a high blood lead level.

X is a discrete random variable since it can take only the values $0, 1, 2$ or 3. To answer probability questions regarding X, we should first find its density function $f(x)$.

In class we will draw the tree diagram associated with this example.

From the tree diagram we see that the path probabilities are:

$$f(0) = P(X = 0) = (0.9)^3 = 0.729$$

$$f(1) = P(X = 1) = 3 \cdot (0.1) \cdot (0.9)^2 = 0.243$$
$$f(2) = P(X = 2) = 3 \cdot (0.1)^2 \cdot (0.9) = 0.027$$
$$f(3) = P(X = 3) = (0.1)^3 = 0.001$$

We can also draw a table which summarizes all this information.

x	0	1	2	3
$f(x)$	0.729	0.243	0.027	0.001

The probability that **at least** 2 children in the group have high blood lead level is

$$P(X \geq 2) = P(X = 2) + P(X = 3) = 0.027 + 0.001 = 0.028$$

The probability that **at most** 2 children in the group have high blood lead level is

$$P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2) = 0.729 + 0.243 + 0.027 = 0.999$$

Example 2. The random variable X gives the number of persons per day who are seeking emergency room treatment unnecessarily in a small hospital. It can take the values $0, 1, 2, 3, 4, 5$. The following table gives the observed values of X for 365 days:

value	0	1	2	3	4	5
frequency	297	25	18	14	8	3

The respective probabilities for the values $0, 1, 2, 3, 4, 5$ are given by the table below:

x	0	1	2	3	4	5
$f(x)$	0.81	0.07	0.05	0.04	0.02	0.01

What is the probability that in a randomly chosen day, there will be at least 4 persons seeking emergency room treatment unnecessarily? (in class)

Expectation

The "mathematical expectation" of a random variable X is best understood if we think of it as the *average* value for the variable X.

Whereas the exact value that the random variable X will take is unknown (since it is random), its *expected* or (average) value denote by $E(X)$ is no longer random and can be calculated by means of the following formula:

$$\mu = E(X) = \sum_x x f(x)$$

where the sum is taken over all possible values x for the variable X.

Example 1. (continued) We calculate the expected value $E(X)$ in the case of the random variable X given in Example 1.

$$E(X) = 0 \cdot (0.729) + 1 \cdot (0.243) + 2 \cdot (0.027) + 3 \cdot (0.001) = 0.3$$

The significance of this result is that if we repeatedly draw groups of 3 children from the group, then the average number of children with high blood lead level per group will be 0.3.

Variance

The expected value of the squared difference between X and $E[X]$ is called the **variance**. (Try to explain why we have to consider the squared difference and not the difference itself.)
The variance is denoted with σ^2:

$$\sigma^2 = E[(X-\mu)^2] = \sum_x (x-\mu)^2 f(x)$$

where the sum is taken over all possible values x for X.
The square root of the variance σ is called the **standard deviation**.

Example 1. (continued) We calculate the variance and the standard deviation for the variable X.

$$\sigma^2 = (0-0.3)^2 \cdot (0.729) + (1-0.3)^2 \cdot (0.243) + (2-0.3)^2 \cdot (0.027) + (3-0.3)^2 \cdot (0.001) = 0.27$$

$$\sigma = \sqrt{0.27} \approx 0.5196$$

Computational shortcut for calculating the variance:

$$\sigma^2 = E[X^2] - (E[X])^2$$

Example 1. (continued) First calculate

$$E[X^2] = \sum_x x^2 f(x) = 0^2 \cdot (0.729) + 1^2 \cdot (0.243) + 2^2 \cdot (0.027) + 3^2 \cdot (0.001) = 0.36$$

Then use

$$\sigma^2 = E[X^2] - (E[X])^2 = 0.36 - (0.3)^2 = 0.36 - 0.09 = 0.27$$

Cumulative Distribution Function

The cumulative distribution function $F(x)$ gives the probability that X takes values smaller or equal to x:

$$F(x) := P(X \le x) = \sum_{y \le x} f(y)$$

Example 1. We calculate the cumulative distribution function for the random variable X.

$$F(0) = f(0) = 0.729$$

$$F(1) = f(0) + f(1) = 0.729 + 0.243 = 0.972$$

$$F(2) = f(0) + f(1) + f(2) = 0.729 + 0.243 + 0.027 = 0.999$$

$$F(3) = f(0) + f(1) + f(2) + f(3) = 0.729 + 0.243 + 0.027 + 0.001 = 1$$

We can also draw a table which contains all these values.

6.2 Binomial Distribution

Binomial distribution appears in the following situation:

1. we have an experiment consisting of n identical and independent trials;
2. each trial can be classified as a "success" or a "failure";
3. for each trial, the probability of a "success" is p;
4. we are interested in the total number X of "successes".

The random variable X is said to have a **binomial distribution** with n trials and p probability of success. The probability density function is given by:

$$f(x) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x}, \qquad x = 0, 1, 2, \ldots, n$$

The expectation and variance are:

$$E(X) = np \quad \text{and} \quad \text{Var}(X) = np(1-p).$$

The *binomial coefficient* is calculated by the formula:

$$\left(\begin{array}{c} n \\ k \end{array} \right) = \frac{n!}{k!(n-k)!}$$

The values of the binomial coefficient are given in Table 17.1 of the textbook for $n \leq 20$.

Example 3. The probability of germination of a beet seed is 0.8 and the germination of a seed is called a "success". We plant 3 seeds and we assume that the germination of one seed is independent of the germination of another seed. Let X be the random variable which gives the total number of germinated seeds. Then X has a binomial distribution with 3 trials and 0.8 probability of success. The possible values for X are $0, 1, 2, 3$. To calculate the probability that X takes one of these values, we use a tree diagram (in class). Its density function is:

$$f(x) = \left(\begin{array}{c} 3 \\ x \end{array} \right) (0.8)^x (0.2)^{3-x}, \qquad x = 0, 1, 2, 3$$

Example 4. In Canada, 85% of the population has Rh positive blood. We take a random sample of 6 persons and count the number of persons with RH positive blood. The drawing of the tree diagram in this case becomes cumbersome. X has a binomial distribution with 6 trials and 0.85 probability of success. The density function of X is given by

$$f(x) = P(X = x) = \left(\begin{array}{c} 6 \\ x \end{array} \right) (0.85)^x (0.15)^{6-x}, \qquad x = 0, 1, 2, 3, 4, 5, 6$$

For instance,

$$f(2) = P(X = 2) = \left(\begin{array}{c} 6 \\ 2 \end{array} \right) (0.85)^2 (0.15)^4 \approx 0.0055$$

The expected value and the variance of X are:

$$E(X) = 6 \cdot (0.85) = 5.1, \quad \text{Var}(X) = 6 \cdot (0.85) \cdot (0.15) = 0.765$$