Chapter 4 – Discrete Random Variables

A *random variable* is a function that assigns a real number to each outcome of an experiment – shorthand abbreviation here is 'RV'. Four examples of RV's include randomly selecting a Loyola student on campus and recording (a) his/her hair color (nominal), (b) opinion of the statement "Ice cream is good for breakfast" on a Likert scale (ordinal), (c) the number of courses in which s/he is currently enrolled (discrete), and (d) his/her *exact* height (continuous). Our focus here is on *discrete* RVs and properties associated with them.

Example – Randomly draw n = 3 cards with replacement from a fair deck of cards and count the number of diamonds. In the last chapter, we saw that if X is the RV that counts the number of diamonds, then X takes on the values x = 0, 1, 2, and 3 with respective probabilities 0.421875, 0.421875, 0.140625 and 0.015625. Note that getting X = 0 is equivalent to the event {NNN}, where D = a diamond and N = a non-diamond. Getting X = 1 is equivalent to the events {DNN, NDN, NND}, and so on for the other values of X.

Definitions – The *probability mass function* (pmf) is the function defined for discrete RV's:

$$f_X(x) = f(x) = Pr(X = x)$$

Probabilities are accumulated using the cumulative *distribution function* (cdf):

$$F_X(x) = F(x) = Pr(X \leq x)$$

For the above example, the pmf is:

$$f(0) = 0.421875$$
$$f(1) = 0.421875$$
$$f(2) = 0.140625$$
$$f(3) = 0.015625$$
$$f(x) = 0, \text{ o.w.}$$

'o.w.' stands for 'otherwise' – that is, for other values of x.

Also, the cdf is:

$$\begin{aligned} F(x) &= 0 && \text{for } x < 0, \\ F(x) &= 0.421875 && \text{for } 0 \le x < 1, \\ F(x) &= 0.843750 && \text{for } 1 \le x < 2, \\ F(x) &= 0.984375 && \text{for } 2 \le x < 3, \\ F(x) &= 1 && \text{for } x \ge 3. \end{aligned}$$

Note that whereas pmf's have the 'spike shape' as in Fig. 4.1 (p.125) and Fig. 4.3 (p.127), cdf's have the 'step shape' as in Fig. 4.2 (p.126). Also the Basic Properties of these functions given on p. 127 are important and should be noted.

<u>Another example - waiting for a Diamond</u> – As in Ex. 4.4 on p.126, let X be *the number of non-diamonds (N) before our first diamond (D)* in repeated draws with replacement from a fair deck of cards. Note that X can take on the values x = 0, 1, 2 … (without any upper bound). The associated events and probabilities are:

x	Event	f(x)		
0	D	¼	= 1/4	= 0.2500
1	ND	¾ * ¼	= 3/16	= 0.1875
2	NND	$(¾)^2$ * ¼	= 9/64	= 0.1406
3	NNND	$(¾)^3$ * ¼	= 27/256	= 0.1055
4	NNNND	$(¾)^4$ * ¼	= 81/1024	= 0.0791
…	…	…		
k	N … ND	$(¾)^k$ * ¼		
…	…	…		

In general, the formula is $f(x) = \pi(1 - \pi)^x$ for x = 0, 1, 2, …

How can we show this is a valid pmf in the sense of property (i) on p.127: that is, that $\sum_{x=0}^{\infty} f(x) = 1$**?**

In *probability histograms* (theoretical histograms), areas do not relate to frequencies, but rather to probabilities. See e.g. Figure 4.5 on p.129 The Law of Large Numbers reminds us that as sample sizes get larger and larger, empirical histograms from data look more and more like the probability histogram.

The *50^{th} percentile* or *median* is the smallest value of x such that $F(x) \geq ½$; similarly, for r between 0 and 100, the *r^{th} percentile* is the smallest value of x such that $F(x) \geq r/100$. Though rarely used, the *mode* is the value of x where the pmf f(x) reaches its maximum.

In a manner very similar to our definition for a mean for grouped data, we define the (theoretical) *mean* μ_X – also called the *expected value* $E(X)$ – in this way:

$$\mu_X = E(X) = \Sigma\, x_k f(x_k)$$

Note that the x's here are weighted by the respective probabilities; also, be sure to heed the warning on p.133. For example, for choosing n = 3 cards with replacement from a fair deck of cards and counting the number of diamonds (X), the mean is

$$\mu_X = E(X) = (0)*(0.421875) + (1)*(0.421875) + (2)*(0.140625) + (3)*(0.015625) = 0.75 = ¾$$

Sometimes we omit writing the X subscript of μ when it is obvious.

Note also in passing that

$$E(X^2) = (0)^2*(0.421875) + (1)^2*(0.421875) + (2)^2*(0.140625) + (3)^2*(0.015625) = 1.125$$

This latter calculation exemplifies Section 4.5 on pp. 133-4.

The *variance* of the RV X is

$$\sigma_X^2 = Var(X) = E\{(X - \mu)^2\} = \Sigma\, (x_k - \mu)^2 f(x_k)$$

The standard deviation is not surprisingly the positive square root of the variance, denoted σ (subscript of X usually suppressed here too).

We have the very important next theorem: let RV X have mean μ and variance σ^2, then

$$\sigma^2 = E(X^2) - \mu^2 = \Sigma\, x_k^2\, f(x_k) - \mu^2$$

Returning to the above Diamond example,

$$\sigma^2 = 1.125 - (0.75)^2 = 0.5625 = 9/16.$$

Some Results (X is a RV, a and b are constants):

1. **E(aX + b) = a E(X) + b**
2. **Var(aX + b) = a^2 Var(X)**
3. **SD(aX + b) = |a| SD(X)**

<u>Discrete Uniform RVs</u> – take many forms, and the most popular is to consider a Box containing 'tickets' with the integers 1, 2, … up to n on the tickets, sampling with replacement from the box, and letting X be the number on a randomly selected ticket. Then, we associate the probability f(x) = 1/n with each of the values x = 1, 2, … n. It is then rather straightforward to show that the mean and variance of X are:

$$\mu = \frac{n+1}{2} \qquad \text{and} \qquad \sigma^2 = \frac{n^2-1}{12}$$

<u>Useful finite sums</u>:

$$1 + 2 + \ldots + n = (1/2)*n(n+1)$$
$$1^2 + 2^2 + \ldots + n^2 = (1/6)*n(n+1)(2n+1)$$

<u>Useful infinite sum (-1 < x < 1)</u>:

$$1 + x + x^2 + x^3 + \ldots = \frac{1}{1-x}$$