Chapter 5 – Success/Failure Experiments

In this chapter, we'll discuss three discrete distributions related to so-called urn models, namely, the *Bernoulli*, *Binomial* and *Hyper-geometric* distributions.

A. Bernoulli Random Variables

Dichotomous experiments that record a '1' for a Success and a '0' for a Failure are given this name (after Jakob Bernoulli of Basel, CH). If the Success probability is denoted π – so the Failure probability is $1 - \pi$ – then the pmf is $f(x) = \pi^x(1 - \pi)^{1-x}$ for x = 0 and 1. (Also, f(x) = 0 for other values of x.) Further, the mean is $\mu = 0*(1-\pi) + 1*(\pi) = \pi$. Since $E\{X^2\} = 0^2*(1-\pi)+1^2*(\pi)=\pi$, the variance is $\sigma^2 = \pi(1-\pi)$. This RV and distribution are really important to help us understand Binomial RV's, the next topic. (Read through Examples 5.1-5.4.)

Box Model – Recall the concept from Chapter 1 of introducing a Box to represent the Population. For the above Bernoulli distribution, the Box would contain tickets with "0"s and "1"s on them, and the proportion of "1"s is π while the proportion of "0"s is $1-\pi$. For this distribution, we envisage taking *only one draw* from this Box.

B. Binomial Random Variables

Here, conditions are similar to the Bernoulli setting in the sense of a 'Success' or 'Failure' on each trial, but here we take n independent draws or trials (with replacement) and we count the number of successes (denoted K in the book). Provided the following rules are met, K has a Binomial distribution (K is a Binomial RV):

Rule 1. The number of Bernoulli trials n is predetermined.
Rule 2. The RV K is the number of Successes in the n trials.
Rule 3. The trials are independent.
Rule 4. The probability of a Success π is the same for each trial.

These are sometimes called the 'BINS' conditions.

<u>Box Model</u> – The Box here is the same as above, but now we make n random draws *with replacement* and count the number of Successes.

The pmf (probability function) for the Binomial distribution is

$$f(k) = \binom{n}{k}\pi^k(1 - \pi)^{n-k}, \text{ for } k = 0, 1, 2 \ldots n$$

where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$**, n! = n*(n-1)*(n-2)*…*2*1, and 0! = 1;** **see p. 150.**

To illustrate, recall drawing n = 3 cards independently and with replacement from a fair deck of cards and noting the number of Diamonds (K). We can use the above formula with π = ¼ to find the probabilities f(0), f(1), f(2) and f(3); for example $f(2) = 3*(1/4)^2(3/4)^1$**.**

Notation: Sometimes, we'll write K ~ Binomial(n,π); so for the above Diamond example, K ~ Binomial(3 , ¼)

To find the *mean* and *variance* of a Binomial RV, we note that

$K = X_1 + X_2 + \ldots + X_n$**, the sum of n independent Bernoulli RVs.**

Since for each X_s**,** $E(X_s) = \pi$ **and** $Var(X_s) = \pi(1-\pi)$**, it follows that**

$$\mu_K = \mu = n\pi \quad \text{and} \quad \sigma^2 = n\pi(1-\pi)$$

To illustrate, if we draw n = 48 cards with replacement from a fair deck of cards, and we count the number of diamonds (K), we expect $\mu = n\pi = 48*(1/4) = 12$ **diamonds, give or take** $\sigma = \sqrt{n\pi(1-\pi)} = \sqrt{48*(1/4)*(3/4)} = \sqrt{9} = 3$ **diamonds.**

C. Hypergeometric Random Variables

The Hypergeometric (HG) distribution is appropriate in the Binomial setting but where draws are made *without replacement*. The number of tickets in the Box (i.e., the population size) is N, the number of "1"s is $N_1 = \pi N$, and the number of "0"s is $N_0 = (1-\pi)*N$. The sample size is again n.

The pmf (probability function) for the HG distribution is

$$f(k) = \frac{\binom{N_1}{k} \times \binom{N_0}{n-k}}{\binom{N}{n}}, \text{ for } k = 0, 1, 2 \ldots n$$

The *mean* and *variance* of a HG variable are given by the following:

$$\mu = n\pi \qquad \text{and} \qquad \sigma^2 = \frac{N-n}{N-1} * n\pi(1-\pi)$$

Note that the mean of a HG variable is the same as that of a Binomial variable, but here the standard deviation (SD) is the same as the SD of the Binomial distribution multiplied by the so-called Small Population Reduction Factor (SPRF): $\text{SPRF} = \sqrt{\frac{N-n}{N-1}} \approx \sqrt{\frac{N-n}{N}} = \sqrt{1-\frac{n}{N}}$

When N is large with respect to n, the SPRF is nearly one. This gives the General Rule: A population of size N is considered *large* viz-a-viz the sample size n whenever $N \geq 20n$, in which case the sample size is no more than 5% of the population. In general, the SPRF is ignored whenever $N \geq 20n$; when this is not the case, it needs to be taken into account.

In class example if time: Exercise 5.19 on p.158.