111, section 8.4, Appendix E The Hypergeometric Distribution

Notes by Tim Pilachowski

Background:

Recall the definition of Bernoulli trials which make up a **binomial** experiment:

The number of trials in an experiment is fixed.

There are exactly two events/outcomes for each trial, usually labeled success and failure.

P(success) = p must be the same for each trial.
Therefore, P(failure) = $q = 1 - p$.

Success and failure must be independent from one trial to the next.

Also recall Examples A-1 and A-2. "Suppose that a box contains 3 blue blocks and 2 yellow blocks." Picking blocks ***without replacement*** was not a binomial experiment. Picking blocks ***with replacement*** was a binomial experiment.

When it came to Example G-2 ("A Math 220 class, taught in the Fall of 2010 at UMCP, had the following grade distribution.") andExample H ("2 out of every 90 spark plugs produced is defective.") we also needed to pick students and spark plugs "with replacement" to keep the trials independent.

That is, the probabilities for picking any one given student or spark plug needed to remain the same throughout the experiment.

But in practice, selections are usually made "without replacement". The same student is not counted twice; there is no need to test the same spark plug again; it is unlikely that a person wants to answer the same survey questions more than once.

So the question becomes, is there a way to calculate these probabilities, without having to go through a lot of rigamarole like extensive tree diagrams and asking "How many are left to pick from?" [rigamarole: (noun): a complex and sometimes ritualistic procedure]

The answer is "Yes", and we already have the tools we need, introduced back in section 7.4.

The resulting probability distribution is called a **hypergeometric probability distribution**.

The significant difference between a binomial probability and a hypergeometric probability is that binomial picks are done "with replacement" and hypergeometric picks are done "without replacement".

The conditional probabilities involved in picking "without replacement" need to be taken into account.

For a hypergeometric probability, we'll need to know the size of the population from which we're picking the sample.

Example A-2 revised. Suppose that a box contains 30 blue blocks and 20 yellow blocks. You pick four blocks without replacement. Define success as “picking a blue block”.

a) What is the probability of picking exactly two blue blocks?

The total number of ways to pick four blocks out of the fifty blocks is

The number of ways to pick two out of the thirty blue blocks is

If exactly two blocks are blue, then the other two must be yellow (“not blue”). The number of ways to pick two out of the twenty yellow blocks is

So, for random variable *X* = number of blue blocks picked, *P*(*X* = 2) =

b) What is the probability of picking at least two blue blocks?

The process used above can be generalized to any hypergeometric probability.

Using the variable letters in your text (other sources use other letters as the variables):

r = the number of items in the population (In the block example above, *r* = 50 blocks total.)

g = the number of items in the population that are classified as good/successes (In the block example above, *g* = 30 blue blocks.)

b = the number of items in the population that are classified as bad/failures (In the block example above, *b* = 20 yellow blocks.)

Note that *r* = *g* + *b*.

n = the number of items in the sample (In the block example above, *n* = 4 blocks picked.)

i = the number of items in the sample that are classified as good/successes (In the block example part a) above, *i* = 2 blue blocks picked.)

k = the number of items in the sample that are classified as bad/failures (In the block example part a) above, *k* = 2 yellow blocks picked.)

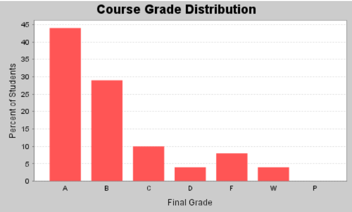
Note that *n* = *i* + *k*.

Your text defines random variable *X* = number of bad/failures. If we define *Y* = number of good/successes, then using the combination formula gives us,

$$P(X=k)=P(Y=i)=\frac{C(b,k)*C(g,i)}{C(r,n)}=\frac{\frac{b!}{k!(b-k)!}*\frac{g!}{i!(g-i)!}}{\frac{r!}{n!(r-n)!}}.$$

This text defines the random variable X in terms of “number of failures”, while most texts use the more traditional definition in terms of “number of successes.” Since multiplication is commutative, you can choose whichever is more convenient for the question asked (as we will in Examples G and H below). Just make sure that the numbers match up.

As observed in Lecture 7.5 on conditional probability, if the population is small (as in Example A) when sampling without replacement, the probabilities from one trial to the next change a lot. In a real statistical analysis, working with larger populations, it is common to sample without replacement on a regular basis. When the sample size is small in comparison to the size of the population, a binomial probability calculation can provide a pretty good estimate of the hypergeometric probability.



Example G-2 revisited. A Math 220 class, taught in the Fall of 2010 at UMCP, had the following grade distribution among 165 students on the roster. Define success as a grade of “C” or better. Pick 15 students at random *without replacement*. What is the probability that between 11 and 14 students have a grade of C or better?

r = the number of students in the class =

p = the probability of getting a C or better =

g = the number of students in the population that are classified as good/success =

b = the number of students in the population that are classified as bad/failure =

n = the number of students in the sample =

i = the number of students in the sample that are classified as good/success =

k = the number of students in the sample that are classified as bad/failure =

Let random variable Y = number of students that have a grade of C or better. Then,

P(between 11 and 14 students have a grade of C or better) =

Example H. From prior experience and testing, Shockingly Good, Inc. has determined that 2 out of every 90 spark plugs produced is defective. The company picks 20 spark plugs at random (*without replacement*) from a production line that has produced 1800 spark plugs. Define random variable X = number of defective spark plugs.

r = the number of spark plugs in the production run =

p = the probability that a spark plug is good =

g = the number of spark plugs in the population that are classified as good/success =

b = the number of spark plugs in the population that are classified as bad/failure =

n = the number of spark plugs in the sample =

a) What is the probability that exactly 1 spark plug is defective?

i = the number of spark plugs in the sample that are classified as good/success =

k = the number of spark plugs in the sample that are classified as bad/failure =

b) What is the probability that at most 1 spark plug is defective?

i =

k =

c) What is the probability that at least 2 spark plugs are defective?

i =

k =

d) In a sample of 20 spark plugs, what is the expected number of number of defective spark plugs? What is the expected number of number of good spark plugs?

e) In a sample of 20 spark plugs, what are the variance and standard deviation for X = number of defective spark plugs? [You don't have to memorize the variance formula.]