Stat 400, section 3.5, Hypergeometric and Negative Binomial Distributions

Notes by Tim Pilachowski

Background for hypergeometric probability distributions:

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

The number of trials, n, in an experiment is fixed in advance.
There are exactly two events/outcomes for each trial, usually labeled success (S) and failure (F).
Trials are independent from one trial to the next, i.e. the outcome of one trial doesn't affect the next.
$P(S) = p$ must be the same for each trial. [$P(F) = 1 - p$ is often designated as q].

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 B ("A Math 220 class, taught in the Fall of 2010 at UMCP, had the following grade distribution.") and Example C ("2 out of every 90 spark plugs produced is defective.") we needed to confirm that the sample size, n, was at most 5% of the population size, N, so the experiment could be treated as though it were a binomial experiment.

But in practice, selections are usually made "without replacement" in situations where our rule of thimb does not apply. 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 2.3.

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. Unlike a binomial calculation, we'll use population size in finding hypergeometric probabilities.

3.4 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?

Is $n \leq 5\%$ of N ?

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):

N = the number of items in the population (In the block example above, $N = 50$ blocks total.)
M = the number of items in the population that are classified as good/successes (In the block example above, $M = 30$ blue blocks.)
$N - M$ = the number of items in the population that are classified as bad/failures (In the block example above, $N - M = 50 - 30 = 20$ yellow blocks.)
n = the number of items in the sample (In the block example above, $n = 4$ blocks picked.)
x = the number of items in the sample that are classified as good/successes (In the block example part a) above, $x = 2$ blue blocks picked.)
$n - x$ = the number of items in the sample that are classified as bad/failures (In the block example part a) above, $n - x = 2$ yellow blocks picked.)

If we define random variable X = number of good/successes, then using the combination formula gives us,

$$P(X = x) = h(x; n, M, N) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}, \qquad \max(0, \quad n - N + M) \le x \le \min(n, M).$$

Note how (as in the Examples of section 2.3) the numbers add up.

The mean, variance and standard deviation of a hypergeometric random variable X are,

$$E(X) = np = n\left(\frac{M}{N}\right), \quad V(X) = \left(\frac{N-n}{N-1}\right)npq = \left(\frac{N-n}{N-1}\right)n\left(\frac{M}{N}\right)\left(1-\frac{M}{N}\right), \quad \sigma_X = \sqrt{V(X)}.$$

3.4 Example A-2 continued. 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”. Find the mean, variance and standard deviation.

3.4 Example C revisited. 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 360 spark plugs. Define random variable X = number of good spark plugs.

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

p = the probability that a spark plug is good =

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

$N - M$ = 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? *answer*: ≈ 0.3027

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

$n - x$ = 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? *answer*: ≈ 0.9328

x =

$n - x$ =

c) What is the probability that at least 2 spark plugs are defective? *answer*: ≈ 0.0672

x =

$n - x$ =

d) In a sample of 20 spark plugs, what is the expected number of number of good spark plugs? *answer*: ≈ 19.5556

e) In a sample of 20 spark plugs, what are the variance and standard deviation for X = number of good spark plugs? *answers*: ≈ 0.4116, ≈ 0.6415

Background for negative binomial probability distributions:

Three out of four conditions for a **negative binomial** experiment are the same as for a binomial experiment:

Trials are independent from one trial to the next, i.e. the outcome of one trial doesn't affect the next.
There are exactly two events/outcomes for each trial, usually labeled success (S) and failure (F).
$P(S) = p$ must be the same for each trial. [$P(F) = 1 - p$ is often designated as q].

The difference is that, instead of the number of trials, n, being fixed in advance, the trials are conducted until a total of r successes have been observed. That is, for a negative binomial random variable, the number of successes, r, is fixed in advance and the number of trials, n, is random.

Example B. You test electrical components until you find four that work (event S = success). Define X = number of failures (event F = failure) that precede 4 successes. If $P(S) = p = 0.9$, what is the probability that there will be exactly 2 Fs before the 4^{th} S?

What are the outcomes for this experiment that meet the requirement "exactly 2 Fs before the 4^{th} S" ?

What is the probability calculation for each of these outcomes?

What is the probability that there will be exactly 2 Fs before the 4^{th} S?

The process used above can be generalized to any negative binomial probability for which we define random variable X = number of failures observed prior to a specified number of successes.

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

r = the desired number of successes (In Example B above, $r = 4$.)
p = the probability of success, $P(S)$ (In Example B above, $p = 0.9$.)
x = the number of failures before r successes are achieved (In Example B above, $x = 2$.)

A negative binomial experiment will consist of a total of $(r + x)$ trials, with the last trial being a success, and the $(r + x - 1)$ preceding trials consisting of $(r - 1)$ successes and x failures.

Using the combination formula gives us,

$$P(X = x) = nb(x; r, p) = \binom{x+r-1}{r-1} p^{r-1} * (1-p)^x * p = \binom{x+r-1}{r-1} p^r * (1-p)^x, \qquad x = 0, 1, 2, 3, \ldots.$$

Notes→

The mean, variance and standard deviation of a negative binomial random variable X are,

$$E(X)=\frac{r(1-p)}{p}=\frac{rq}{p},\quad V(X)=\frac{r(1-p)}{p^2}=\frac{rq}{p^2},\quad \sigma_X=\sqrt{V(X)}.$$

Example B continued. You test electrical components until you find four that work (event S = success). Define X = number of failures (event F = failure) that precede 4 successes. If $P(S) = p = 0.9$, what are the expected value, variance and standard deviation for this experiment?

Example B variation. You test electrical components until you find a dozen that work. If P(component working) = $p = 0.9$, what is the probability that there will be at most 1 failure before the 12^{th} success? What are the expected number of failures, variance and standard deviation for this experiment?
answers: ≈ 0.6213, ≈ 1.33, ≈ 1.4815, ≈ 1.2172