Random variables, probability distributions, binomial random variable

Example 1: Consider the experiment of flipping a fair coin three times. The number of tails that appear is noted as a discrete random variable:

\[ X = \text{number of tails that appear in 3 flips of a fair coin.} \]

There are 8 possible outcomes of the experiment: namely the sample space consists of

\[ \Omega = \{ \text{HHH, HHT, HTH, THH, TTT, THT, TTH, TTH} \} \]

where

\[ X = 0, 1, 1, 2, 1, 3, 2, 2 \]

are the corresponding values taken by the random variable \( X \).

Definition: A random variable is a function from outcomes \( \omega \in \Omega \) to numbers*. For the above example with \( \omega = \text{HTT} \), \( X(\omega) = X(\text{HTT}) = 2 \) counts the number of tails in the three coin flips. A discrete random variable is one which only takes a finite or countable set of values as opposed to a continuous random variable which can take any real number value in an interval of real numbers. (There are uncountably many real numbers in an interval of positive length.)

* (Technically one should really say a measurable function which roughly means that \( X \) behaves nicely on the sets for which one can assign a probability.)

a) What are the possible values that \( X \) takes on and what are the probabilities of \( X \) taking a particular value? From the above we see that the possible values of \( X \) are the 4 values

\[ X = 0, 1, 2, 3. \]

Said differently the sample space is a disjoint union of the 4 events \( \{ X = j \} \) for short (\( j = 0,1,2,3 \) ) where what is meant by the event \( \{ X = j \} \) is:

\[ \{ \omega: X(\omega) = j \} j = 0,1,2,3 \]

Specifically in our example:

\[ \{ X = 0 \} = \{ \text{HHH} \} , \]
\[ \{ X = 1 \} = \{ \text{THH, HTH, HHT} \} , \]
\[ \{ X = 2 \} = \{ \text{TTH, HTT, THT} \} , \]
\[ \{ X = 3 \} = \{ \text{TTT} \} . \]

Since for a fair coin we assume that each element of the sample space is equally likely (with probability 1/8) we find that the probabilities for the various values of \( X \), called the probability distribution \( f(x) = f_X(x) = P(X = x) \) of \( X \) or for \( X \) discrete, also commonly called the probability mass function (or pmf) \( f(x) = p_X(x) = P(X = x) \) can be summarized in the following table listing the possible values beside the probability of that value:

<table>
<thead>
<tr>
<th>( x )</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p(x) )</td>
<td>1/8</td>
<td>3/8</td>
<td>3/8</td>
<td>1/8</td>
</tr>
</tbody>
</table>

We can say that this pmf places mass 3/8 on the value \( x = 2 \).

The numbers (=”masses” or probabilities) for a pmf should be between 0 and 1. The total mass (i.e. total probability ) must add up to 1.

To use the example in the text, the function \( f(x) = \frac{x-2}{2} \) for \( x = 1,2,3,4 \) could not serve as a pmf since this gives \( f(1) < 0 \) which is not allowed but \( f(x) = \frac{x-1}{6} \) for \( x = 1,2,3,4 \) would work since
these numbers are non-negative and add to 1. Similarly  \( h(x) = \frac{x^2}{25} \) for \( x = 0,1,2,3,4 \) fails as a pmf since the sum of these values is 6/5 not 1 but  \( h(x) = \frac{x^2}{30} \) for \( x = 0,1,2,3,4 \) would work.

b) What is the expected value of the random variable \( X \) ?

We recall from the end of chapter 3 that this is the sum of the values times their probabilities :

\[
E[X] = 0 \cdot (1/8) + 1 \cdot (3/8) + 2 \cdot (3/8) + 3 \cdot (1/8) = 12/8 = 3/2
\]

This is not surprising since for a fair coin on average half the time we expect to see tails, and the answer is half of the \( n=3 \) flips = \( np \) where \( p=1/2 \) is the probability of a tail occurring.

**Formula for the expected value** of a discrete r.v. : Given a pmf  for a discrete random variable \( X \), its expected value is given by the formula :

\[
E[X] = \sum_{x} x \cdot p_X(x) \quad \text{where the sum is over all possible values of the random variable.}
\]

c) What is the cumulative distribution function (c.d.f.) for the above pmf ? The cdf is denoted  \( F(x) = F_X(x) \) (when we want to emphasize the particular random variable \( X \) involved), and is defined to be the sum of the probabilities  \( F(x) = F_X(x) = \sum_{j \leq x} p_X(x) \) over all values less than or equal to \( x \). For the above example we have

\[
\begin{array}{|c|c|c|c|c|}
\hline
x & 0 & 1 & 2 & 3 \\
\hline
p(x) & 1/8 & 3/8 & 3/8 & 1/8 \\
\hline
F(x) & 1/8 & 4/8 & 7/8 & 1 \\
\hline
\end{array}
\]

Note that the pmf equals 0 for any value other than the allowed values. The cdf is defined for all real \( x \). If we let  

\[
F(x-) = \lim_{y \to x; y<x} F(x)
\]

denote the left hand limit of the cdf \( F \) as \( y \) approaches \( x \) for values \( y<x \). Then  

\[
p(x) = F(x) - F(x-) 
\]

That is, the pmf is the value by which the cdf jumps. These jumps only occur at one of the possible values of the pmf. A graph of the above step function cdf looks like :

<table>
<thead>
<tr>
<th>y=F(x) y=7/8 --------o</th>
</tr>
</thead>
<tbody>
<tr>
<td>y=1/2 ----------o</td>
</tr>
<tr>
<td>-------------------</td>
</tr>
<tr>
<td>y=1/8-------o</td>
</tr>
<tr>
<td>y=0 ---o-------------------</td>
</tr>
</tbody>
</table>

0 1 2 3 x

d) From the above table we can read off the probability that \( X \) takes a value less than 2 as
$$P(X < 2) = F(1) = p(0) + p(1) = 1/8 + 3/8 = 4/8 = 1/2.$$  
Note: If we would have said: less than or equal to 2 we would add 3/8 to the above answer to give 7/8.

**Binomial random variables:** The random variable $X$ above is an example of a binomial random variable with $n=3$ and $p = 1/2$ which means:

$$X = \text{number of successes in } n \text{ independent identical trials each having success with probability } p$$

Thus $X$ equals the number of successes in the larger experiment consisting of $n$ independent identical repetitions (or trials) of a smaller experiment having only two possible values for the outcomes: success (= tail) which occurs with probability $p$ and failure (= head) which occurs with probability $(1-p)$.

(Here the smaller experiment is to flip a fair coin once). Such **independent identical trials having only two possible outcomes** are called **Bernoulli trials** with **parameter** $p$.

**Example 2:** If we roll a fair six sided die $n=30$ times and let $X = \text{number of times a 3 is rolled}$

(with probability of “success” $p=1/6$) Then

a) The probability of rolling a 3 exactly 7 times out of 30 is

$$P(X = 7) = (30 \choose 7) \left( \frac{1}{6} \right)^7 \left( \frac{5}{6} \right)^{23}.$$  
This follows since there are $(30 \choose 7)$ ways where 7 of the 30 slots are filled with success and for the other 23 of the 30 failure occurs. The answer follows since failure occurs with probability $(5/6)$ each time we roll and using the fact that all 30 rolls are independent so that the probabilities multiply (by definition of independence).

b) Since the 3 is rolled with probability $p=1/6$ this means that in 30 rolls roughly $1/6$ of the time a 3 occurs so we'd guess that the expected number of times we roll a 3 in 30 rolls of the fair die is

$$E[X] = np = 30 \left( \frac{1}{6} \right) = 5.$$  
This turns out to be correct.

A binomial random variable $X$ has **parameters** $n = \text{number of identical trials} \text{ and } p = \text{success probability}$

The pmf **probability distribution for a binomial random variable with parameters } n \text{ and } p \text{ is denoted**

$$b(k ; n, p) = P(X=k) = P(k \text{ successes in } n \text{ independent identical trials }) = \binom{n}{k} p^k (1-p)^{n-k}$$

since by a counting argument $\binom{n}{k}$ is the number of sequences having exactly $k$ successes out of $n$ while the probability of any one of these must be $p^k (1-p)^{n-k}$ using the independence property of the trials.

The **cumulative distribution function of a binomial random variable**, the cdf, not the probability
distribution or pmf, is what is given in the back of the book in Table 1 and is denoted $B(k; n, p) = P(X \leq k) = P(\text{less than or equal to } k \text{ successes in } n \text{ independent identical trials})$

$$= \sum_{j=0}^{k} \binom{n}{j} p^j (1-p)^{n-j}$$

Define the random variable $X_j = 1$ if success (tail) occurs on the $j^{th}$ trial (with probability $p$) and $X_j = 0$ if failure occurs on the $j^{th}$ trial (with probability $(1-p)$).

Such a 2-valued random variable is called a Bernoulli random variable with parameter $p$. The pmf for such a Bernoulli random variable looks like

<table>
<thead>
<tr>
<th>$x$</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p(x)$</td>
<td>$1-p$</td>
<td>$p$</td>
</tr>
</tbody>
</table>

Note that the expected value of a Bernoulli random variable is just the success probability $p$:

$$E[X_j] = 0 \cdot (1-p) + 1 \cdot p = p.$$ 

Then we can write the binomial random variable $X$ as the sum of $n$ i.i.d. 0 or 1 valued Bernoulli random variables where i.i.d. is an abbreviation for independent, identically distributed (i.i.d.).

$$X = \sum_{j=1}^{n} X_j = \text{number of } 1\text{'s (successes) in } n \text{ i.i.d. trials}$$

I.I.D. sometimes goes by the name “memoryless” in some engineering contexts.

A basic property of expectations which we will show when we define the joint probability distribution for two or more random variables in section 5.10 is that the expectation of a sum is the sum of the expectations, and more generally expectation is a linear operation meaning that for any constants $a, b$ and random variables $X$ and $Y$ one has: (true whether or not $X$ and $Y$ are independent)

$$E[aX + bY] = aE[X] + bE[Y].$$

But since all the Bernoulli random variables are identically distributed we find for the expected value of a binomial random variable:

$$\mu = E[X] = \sum_{j=1}^{n} E[X_j] = np$$

The text gives an alternate derivation of this based on the definition of expected value and the properties of binomial coefficients.

Example 3: Suppose 90% of the cars on Wisconsin highways get over 20 miles per gallon (mpg).

a) What is the probability that in a sample of $n=15$ cars exactly 10 of these get over 20 mpg?

The exact answer we have seen is

$$b(10; 15, .9) = \binom{15}{10} (.9)^{10} (.1)^5 = .0104708 \text{ to 7 place accuracy.}$$

We could attempt to compute this directly with the help of a calculator or we could employ Table 1 at the back of the book, making note that the pmf for the binomial can be found from the cdf in Table 1 via

$$b(10; 15, .9) = B(10; 15, .9) - B(9; 15, .9) = .0127 - .0022 = .0105.$$ 

which agrees with direct calculation to 4 places.

b) What is the expected number of Wisconsin cars that get over 20 mpg in a sample of size $n=15$ cars?

This is just the expected value $np$ of a binomial random variable with parameters $n=15$ and $p=.9$ or
Example 4: Suppose a part has a one in a hundred chance of failing (probability \( p = .01 \)) and we sample from 100 independent parts. What is the probability that none of the 100 parts fails? This is a binomial probability \( P(X=100) \) with \( n = 100, \ p = .01 \), mean \( np = 1 \) and \( k = 100 \), (so number of failures is \( n - k = 0 \)). Assuming independence this probability is

\[
( .99 )^{100} = \left( 1 - \frac{1}{100} \right)^{100} = .3660323 \text{ to 7 places.}
\]

If we had to estimate this crudely without using a calculator we might recall from calculus that

\[
\left( 1 - \frac{x}{n} \right)^n \approx e^{-x} \text{ for large } n
\]

(which will be important for us when we discuss Poisson random variables soon). So with \( n = 100, \lambda = 1 \), we would guess \( e^{-1} \approx .367879 \) or roughly \( 1/3 \).

The above approximation follows by Taylor series expansion of the logarithm (with \( \lambda = x \))

\[
\left( 1 - \frac{x}{n} \right)^n = e^{n \log \left( 1 - \frac{x}{n} \right)}.
\]

The above is a special case of the **Poisson approximation to the binomial probability** (in the above example, the probability that a Poisson random variable with mean \( \lambda = 1 \) takes the value 0). The Poisson approximation (which we’ll study in section 4.6) holds when the probability \( p \) of a success is small and the number of trials \( n \) is large in the limiting situation where \( n \to \infty \) while \( p \) shrinks to 0 so that the mean number of successes \( np \) of the binomial stays fixed: or \( p = \lambda / n \).

**The peak of the binomial distribution** essentially occurs at the mean \( np \): If we examine the binomial probabilities \( P(X = k) = b(k; n, p) = \binom{n}{k} p^k (1 - p)^{n-k} \) from the definition of the binomial coefficients we find that for a given \( n \) and \( p \) the ratio for two consecutive values of \( k \) is

\[
\frac{P(X = k + 1)}{P(X = k)} = \frac{\binom{n}{k+1} p^{k+1} (1 - p)^{n-k-1}}{\binom{n}{k} p^k (1 - p)^{n-k}} = \frac{p(n-k)}{(1-p)(k+1)}.
\]

This ratio will be greater than or equal to 1 provided \( p(n-k) \geq (1-p)(k+1) \) which can be simplified to

\[
k + 1 \leq (n+1) \frac{p}{1-p} ,
\]

meaning that the binomial probabilities will increase as \( k \) increases up until the greatest integer \( k+1 \) which is less than or equal to \( (n+1)p \) (denoted \( \lfloor (n+1)p \rfloor \)) and note that with \( p < 1 \) this integer where the probability histogram peaks is within 1 from the mean \( np \) of the distribution. Thus the mean \( np \) is essentially the peak of the binomial probability distribution.

For \( p = 1/2 \) the binomial probability distribution histogram will be symmetric about the mean \( np = n/2 \) reflecting the symmetry property of the binomial coefficients that \( \binom{n}{k} = \binom{n}{n-k} \). For \( p \) much less than \( 1/2 \) the peak will be on the left but the tail of the distribution will extend well to the right of the mean. We say such a distribution is **positively skewed** or **skewed to the right**. Similarly for \( p \) much greater than \( 1/2 \) the distribution is **skewed to the left** or **negatively skewed**.
**Sampling with replacement** from a finite population (with only success or failure as outcomes) corresponds to a binomial random variable. Alternately we can regard this as sampling with or without replacement from an infinite population: The coin flip or dice roll experiment could be done infinitely many times if we had an immortal coin or dice so we can think of an infinite conceptual population of outcomes of the coin flip or dice roll experiments. The binomial random variable corresponds to either sampling from an infinite conceptual population or alternately to sampling with replacement from a finite population. If we put the card back randomly into the 52 card deck then the next time we randomly sample we still have one chance in 52 \( \left( \frac{1}{52} \right) \) of getting the ace of spades. The binomial probability of getting 23 draws of an ace of spades from 500 draws of a randomly shuffled 52 card deck reflects these unchanging probabilities for sampling with replacement.

What is the **analogue of a binomial random variable** when we sample e.g. successes and failures from a finite population **without replacement**? Now the probabilities change so \( p \) does not stay the same. We have already encountered such situations in some of the counting problems we have seen:

**Hyper-geometric distribution**: Sampling without replacement from a finite population. For the random variable \( X = \text{number of defectives in a random sample of size } n \) taken from a population of size \( N \) having \( a \) defectives total the probability that \( X = m \) of selecting exactly \( m \) defective units is by our basic counting principle equal to the so-called **hyper-geometric** probability:

\[
P(X = m) = \binom{a}{m} \binom{N-a}{n-m} \binom{N}{n}
\]

for \( m = 0, 1, 2, \ldots, n \).

The fraction \( p = \frac{a}{N} \) of marked or different or defective items acts like the probability of success for a binomial random variable. The **mean of the hyper-geometric distribution** can be shown to equal

\[
E[X] = \mu = n \cdot \frac{a}{N}.
\]

In the limit as \( N \to \infty \) the hyper-geometric probabilities tend toward the corresponding binomial probabilities and in this limit sampling with or without replacement makes no difference as long as our sample size \( n \) is small compared to the population size \( N \).

**Example 5**: A committee of size \( n=4 \) persons is to be selected from a group of \( N=20 \) city government officials of which 14 are in the Rotary club and \( a=6 \) are not.

a) What is the probability that a randomly selected committee will contain exactly 1 Rotary club members?

This is the hyper-geometric probability

\[
\binom{6}{3} \binom{14}{1} \binom{20}{4}.
\]

b) What is the mean number of non-Rotary club members in the committee of size 4?

This is the mean of a hyper-geometric random variable so

\[
E[X] = \mu = n \cdot \frac{a}{N} = 4 \cdot \frac{6}{20} = \frac{6}{5}.
\]

**Example 6**: What is the expected value of the random variable \( Y = X^2 \), the square of the number of
tails in 3 flips of a fair coin? The pmf table for this random variable $Y$ looks like

<table>
<thead>
<tr>
<th>$y$</th>
<th>0</th>
<th>1</th>
<th>4</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p(y)$</td>
<td>$\frac{1}{8}$</td>
<td>$\frac{3}{8}$</td>
<td>$\frac{3}{8}$</td>
<td>$\frac{1}{8}$</td>
</tr>
</tbody>
</table>

so the expected value is $0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 4 \cdot \frac{3}{8} + 9 \cdot \frac{1}{8} = 24 \cdot \frac{8}{8} = 3$. This example is not particularly illuminating however since the pmf for $Y$ is identical to the one for $X$. Better is the following:

**Example 7**: Consider the random variable

$X = \text{the number of tails in 4 flips of a fair coin subtract 2}$

with pmf given by the table

<table>
<thead>
<tr>
<th>$x$</th>
<th>-2</th>
<th>-1</th>
<th>0</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p(x)$</td>
<td>$\frac{1}{16}$</td>
<td>$\frac{1}{4}$</td>
<td>$\frac{3}{8}$</td>
<td>$\frac{1}{4}$</td>
<td>$\frac{1}{16}$</td>
</tr>
</tbody>
</table>

By symmetry we find that the mean $E[X] = 0$. Now let us compute the mean of $Y = X^2$ in two ways: First we look at the pmf for $Y$ directly

<table>
<thead>
<tr>
<th>$y$</th>
<th>4</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p(y)$</td>
<td>$\frac{2}{16}$</td>
<td>$\frac{2}{4}$</td>
<td>$\frac{3}{8}$</td>
</tr>
</tbody>
</table>

Giving $E[X^2] = 4 \cdot \frac{2}{16} + 1 \cdot \frac{2}{4} + 0 \cdot \frac{3}{8} = 1$. Notice we have lost 2 columns due to the fact that the value of $x$ is getting squared and so the probabilities for the two columns (-1) and (1) of the table for $X$ must be combined in their squared value of 1 for $Y$ and similarly for columns (-2) and 2 which combine into the value 4 in the $Y$ table. Now we claim we could have arrived at the expectation using the probabilities of the original table for $X$ by writing

<table>
<thead>
<tr>
<th>$x$</th>
<th>-2</th>
<th>-1</th>
<th>0</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x^2$</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>$p(x)$</td>
<td>$\frac{1}{16}$</td>
<td>$\frac{1}{4}$</td>
<td>$\frac{3}{8}$</td>
<td>$\frac{1}{4}$</td>
<td>$\frac{1}{16}$</td>
</tr>
</tbody>
</table>

Then we have $E[X^2] = \sum x^2 p(x) = 4 \cdot \frac{1}{16} + 1 \cdot \frac{1}{4} + 0 \cdot \frac{3}{8} + 1 \cdot \frac{1}{4} + 4 \cdot \frac{1}{16} = 1$ as before. Note that in the above table there are now two values of $x$ which give rise to the value 4 for $x^2$. Adding these probabilities gives the corresponding probability of the value 4 in the table for $Y = X^2$.

**Expected value of a function $h(x)$ of a random variable $X$**: More generally for any random variable $h(X)$ which is a function $h(x)$ of a random variable $X$ one can show that the expected value of a function $h(x)$ of a random variable $X$ can be computed directly from the pmf for the original random variable (rather than from the pmf of the function $h(X)$) via the formula:

$E[h(X)] = \sum_x h(x) \cdot p_X(x)$ is the expected value of the random variable $Y = h(X)$

**Definition of the variance of a random variable**:

The variance of $X$ (whose square root is the standard deviation $\sigma$) is defined as

$V(X) = \sigma^2 = E[(X - \mu)^2] = \sum_x (x - \mu)^2 \cdot p(x)$
which can be shown to equal
\[ E[X^2] - \mu^2 \] where here \( \mu = E[X] \).

To see this note
\[
\sum_x (x-\mu)^2 p(x) = \sum_x x^2 p(x) - 2\mu \sum_x x p(x) + \mu^2 \sum_x p(x) = E[X^2] - 2\mu^2 + \mu^2 \quad \text{since}
\]
\[
\sum_x p(x) = 1 \quad \text{and using the definition of the mean} \quad \mu = \sum_x x p(x). 
\]
In the above example this reduces to \( E[X^2] = 1 \) since \( \mu = E[X] = 0 \) there.

**Variance of a binomial random variable**: one can show that the variance of a binomial random variable with parameters \( n \) and \( p \) is \( \sigma^2 = np(1-p) \).

The simplest way to show this is to establish that variance of a sum of independent random variables is the sum of the individual variances (we are not yet ready to prove this fact). Then one can regard the binomial random variable as a sum of \( n \) Bernoulli random variables each having variance \( p(1-p) \). The later is fairly clear since the Bernoulli random variable takes only the two values 1 with probability \( p \) and 0 with probability \( 1-p \) so since its mean \( \mu = p \) its variance is \( \sigma^2 = 1^2\cdot p - \mu^2 = p - p^2 = p(1-p) \).

The variance of a hyper-geometric random variable is similar if we write \( p = a/N \):
\[
\sigma^2 = n \left[ \frac{a}{N} \left(1 - \frac{a}{N}\right) \right] \left( \frac{N-n}{N-1} \right),
\]
which with \( p = a/N \) is the same as the variance of a binomial random variable \( np(1-p) \) except for the additional finite population correction factor \( \frac{N-n}{N-1} \).

**EXAMPLE 8**: (like HW problem 4.20) A quality-control engineer wants to check whether (in accordance with specifications) 96% of the electronic components shipped by his company are in good working condition. To this end, he randomly selects 16 from each large lot ready to be shipped and passes the lot if the selected components are all in good working condition; otherwise, each of the components in the lot is checked. Find the probabilities that the quality control engineer will commit the error of
a) holding a lot for further inspection even though 96% of the components are in good working condition;
b) letting a lot pass through without further inspection even though only 92% of the components are in good working condition.
c) letting a lot pass through without further inspection even though only 85% of the components are in good working condition.

a) The probability of a given component being in working condition is 96% = 24/25. The probability of all 16 components being in working condition, assuming independence of components, \( (24/25)^{16} \approx .96^{16} \approx .5204 \) so the probability of not all passing (holding the lot for further inspection) is just the probability of the complement \( = 1 - .5204 = .4796 \)
b) The probability of a given component being in working condition is now 92%. So the probability of all 16 being OK (letting the lot pass through without further inspection) is \( .92^{16} = .2634 \) (about 26%)
c) Now we have \( .85^{16} = .0743 \) (or bit more than a 7% chance)

We recall the following binomial probability calculation for parts a) and b) to which we add the
EXAMPLE 3: Suppose 90% of the cars on Wisconsin highways get over 20 miles per gallon (mpg). For the binomial r.v. X= the number of vehicles that get over 20 mpg in our sample of size n=15 with p =.90

a) What is the probability that in a sample of n=15 cars exactly 10 of these get over 20 mpg ?

This is the probability \( P(X=10) = b(10 ; 15,.9) \)

The exact answer we have seen is

\[
\begin{align*}
\frac{15}{10} \times (.9)^{10} \times (.1)^{5} &= .0104708 \\
&\text{to 7 place accuracy.}
\end{align*}
\]

We could attempt to compute this directly with the help of a calculator or we could employ Table 1 at the back of the book, making note that the pmf for the binomial can be found from the cdf in Table 1 via

\[
\begin{align*}
b(10 ; 15,.9) &= B(10 ; 15,.9) - B(9 ; 15,.9) \\
&= .0127 - .0022 = .0105.
\end{align*}
\]

which agrees with direct calculation to 4 places.

b) What is the expected number of Wisconsin cars that get over 20 mpg in a sample of size n=15 cars ?

This is just the expected value np of a binomial random variable with parameters n=15 and p=.9 or

\[
\mu_X = np = 15(.9) = 13.5 \text{ cars}.
\]

c) What is the variance of the number of Wisconsin cars that get over 20 mpg in a sample of size n=15 cars ?

For a binomial r.v. we know that the variance is

\[
\sigma^2_X = np(1-p) = 13.5(.1) = 1.35 \text{ (cars)}^2
\]

The standard deviation is thus

\[
\sigma_X = \sqrt{1.35} = 1.162 \text{ cars}.
\]

To parts a) and b) of this hyper-geometric example which appeared in last week's notes we add the variance computation part c). (We didn't cover any of this in class until today10/1/07.)

EXAMPLE 5: A committee of size n=4 persons is to be selected from a group of N=20 city government officials of which 14 are in the Rotary club and a=6 are not.

a) What is the probability that a randomly selected committee will contain exactly 1 Rotary club member? (so that X = number of non-Rotary members on the committee =3 )

This is the hyper-geometric probability \( P(X=3) = \frac{\binom{6}{3} \times \binom{14}{1}}{\binom{20}{4}} = \frac{56}{969} = .05779 \)

b) What is the mean number of non-Rotary club members in the committee of size n=4 ?

This is the mean of a hyper-geometric random variable which is like that of a binomial where \( p = a/N \) so

\[
E[X] = \mu_X = n \cdot p = n \cdot \frac{a}{N} = 4 \cdot \frac{6}{20} = \frac{6}{5}.
\]

c) What is the variance of the number of non-Rotary club members in the committee of size n=4 ?
Taking \( p = a/N \), the hyper-geometric variance looks like that for a binomial r.v. except that now there is a correction factor of \( \frac{N-n}{N-1} \) since now we are sampling without replacement from a finite population. That is

\[
\sigma_x^2 = np(1-p) \frac{N-n}{N-1} = \left( \frac{6}{5} \right) \frac{1 - 6/20}{19} = \left( \frac{6}{5} \right) \frac{16}{475} = \frac{336}{255} = \frac{3 \cdot 7 \cdot 16}{25 \cdot 19}.
\]

Thus the standard deviation is

\[
\sigma_x = \sqrt{\frac{4}{5} \frac{21}{19}} = 0.84105 \text{ persons}.
\]

Note that the probability found in a) is rather rare since we expect to see \( 6/5 \) non-Rotary members and hence \( 4-(6/5) = 14/5 \) Rotary members on average in the committee of size 4 but there we had 3 non-Rotaries.

**Chebyshev’s inequality** says that the probability that a random variable \( X \) deviates from the mean value by at least a number \( k \) of standard deviations \( \sigma \) is bounded above by \( \frac{1}{k^2} \). That is

\[
P(|X-\mu_x| \geq k \sigma) \leq \frac{1}{k^2}.
\]

Equivalently for the complementary event one has

\[
P(|X-\mu_x| < k \sigma) \geq 1 - \frac{1}{k^2}.
\]

In the example of part a) the deviation from the mean is \( 3-(6/5) = 9/5 \) and so for this to equal \( k \sigma \) says \( k = \frac{9}{4} \sqrt{\frac{19}{21}} \approx 2.14 \) i.e the number 3 of non-Rotary members is greater than 2 standard deviations from the mean and Chebyshev says that the probability \( P(|X-\mu_x| \geq k \sigma) \) of seeing any observed value of \( X \) more than \( k=2 \) standard deviations from the mean (i.e in this example the probability of the event that \( X=3 \) or \( X=4 \)) is bounded by \( \frac{1}{k^2} < \frac{1}{4} \). In fact the actual value .05779 from part a) is quite a bit smaller than that bound. We will come back to Chebyshev’s inequality shortly.

**Remark on moments**: For positive integers \( k \) the expected value

\[
E[X^k] = \mu'_k \quad (\text{---Johnson’s notation})
\]

is called the \( k^{th} \) moment of a random variable \( X \) (what Johnson calls a moment about the origin).

Note \( \mu'_1 = \mu = E[X] \). The \( k^{th} \) moment about the mean is defined as

\[
E[(X-\mu)^k] = \mu_k \quad (\text{---Johnson’s notation})
\]

Thus the usual formula for the variance becomes

\[
\sigma^2 = \mu_2 = \mu'_2 - \mu^2 = E[X^2] - (E[X])^2
\]

It is also common to speak of \( k^{th} \) absolute moments \( E[|X|^k] \) which reduces to the usual \( k^{th} \) moment when \( k \) is an even power. One also speaks of the moment generating function when it exists which typically means the exponential moment:

\[
E[e^{\alpha X}]
\]

which is the Laplace transform of the distribution (when it exists and is finite for some \( \alpha \neq 0 \) which then implies that all the usual moments exist for all positive integers \( k \). This is a strong assumption. When it holds the usual moments can then be found by taking various derivatives of the generating function with respect to \( \alpha \) and evaluating these at \( \alpha = 0 \).
Example of a random variable having no mean or variance: For a general random variable $X$ there is no guarantee that the first moment or mean value even exists let alone the variance (second moments). The random variable defined by its probability distribution

$$P(X=k) = \frac{6}{\pi^2} \frac{1}{k^2} \quad k=1,2,3,...$$

is such an example since $E[X]=\sum_k k P(X=k) = \frac{6}{\pi^2} \sum_k \frac{1}{k}$ diverges.

The constant out in front was chosen so that the total probability which is the sum for all $k$ adds up to 1.

It is a famous result I believe due to Euler that

$$\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6} \quad \text{.}$$

Proof of Chebyshev’s inequality: Assuming the second moments exist, if we recall the definition of the variance but then restrict to only the terms in the sum where $|x-\mu| \geq k \sigma$ or equivalently where $(x-\mu)^2 \geq k^2 \sigma^2$, since all the terms are positive one has

$$\sigma^2 = \sum_x (x-\mu)^2 P(X=x) \geq \sum_{x:|x-\mu|^2 \geq k^2 \sigma^2} (x-\mu)^2 p(x) \geq \sum_{x:|x-\mu|^2 \geq k^2 \sigma^2} k^2 \sigma^2 p(x)$$

$$= k^2 \sigma^2 \sum_{x:|x-\mu| \geq k \sigma} p(x) = k^2 \sigma^2 P(|X-\mu| \geq k \sigma)$$

The last equality holds since summing the probabilities $p(x)=P(X=x)$ over only those values $x$ for which $|X-\mu| \geq k \sigma$ or equivalently $(x-\mu)^2 \geq k^2 \sigma^2$ is exactly what we mean by $P(|X-\mu| \geq k \sigma)$.

Upon dividing the above inequality $\sigma^2 \geq k^2 \sigma^2 P(|X-\mu| \geq k \sigma)$ by $k^2 \sigma^2$ we find

Chebyshev’s inequality: $P(|X-\mu_x| \geq k \sigma) \leq \frac{1}{k^2}$ or equivalently for the complementary event one has

$$P(|X-\mu_x| < k \sigma) \geq 1- \frac{1}{k^2}$$

Example 9 (Chebyshev’s inequality like HW 4.47) Show that for a sample of $n = 90,000$ flips of a fair coin, the probability is at least .9975 that the sample proportion

$$\hat{p} = \frac{X}{n} = \frac{X}{90000}$$

of heads will fall strictly between $\frac{42}{90}$ and $\frac{48}{90}$.

Here $X$ is the number of heads observed in 90000 flips is a binomial random variable. As such we know that $X$ is the sum of 90000 Bernoulli random variables (which take the value 1 if the corresponding flip came up heads 0 else). The sample proportion $\hat{p}=\bar{X}$ is thus the same as the sample mean of these Bernoulli random variables and the hat over the $p$ indicates that it is an estimator for the actual population mean which for the Bernoulli random variables is their expected value

$$p=E[X]=\frac{1}{2}$$, the actual probability of getting a head on the $j^{th}$ flip. (In general the sample mean $\hat{\mu}=\bar{X}$ is an estimator for the actual mean $\mu$.) We wish to show that

$$P(\frac{42}{90}<\hat{p}<\frac{48}{90}) = P(42000<X<48000) \geq .9975 = 1- \frac{1}{400}$$.
But the above is just a statement of Chebyshev’s inequality (complementary event form) with $X$ lying within $k = 20$ standard deviations of the mean which for the binomial random variable $X$ is

$$
\mu_X = E[X] = np = 90000(1/2) = 45000 \text{ heads on average}
$$

and standard deviation

$$
\sigma_X = \sqrt{np(1-p)} = \sqrt{300^2(1/2)^2} = 150 \text{ heads}
$$

where we note that the above inequality

$$
42000 < X < 48000
$$

can also be written

$$
|X - 45000| < 3000 \quad \text{where } 3000 = 20 \cdot 150 = k \sigma .
$$

This verifies the claim.

Now let us sketch the derivation of the mean and variance for a hyper-geometric r.v.:

As with the binomial r.v. for the hyper-geometric random variable $X$ with parameters $n, a, N$ :

$$
X = \text{number of special (“defective”) items in a sample of size } n
$$

(from a population of size $N$ having $a$ special or “defective” items)

we can write

$$
X = X_1 + X_2 + \ldots + X_n
$$

as a sum of Bernoulli random variables where

$$
X_j = 1 \text{ if the } j^{th} \text{ item in our sample is special} \quad \text{(defective)}
$$

$$
= 0 \text{ otherwise}
$$

We would not call these Bernoulli trials however because in this case these Bernoulli random variables are not independent.

The expected value of one of these Bernoulli’s is

$$
E[X_j] = P(X_j = 1) = p = \frac{a}{N}
$$

(= the value 1 times the probability of equaling 1)

(Being a random sample means that each of the $N$ items has probability $1/N$ of being selected and since there are $a$ of these $N$ which are special we find the above probability $p = a/N$ of selecting a special item)

Exactly as for binomial r.v.’s we use the fact that the expectation of a sum equals the sum of the expectations to write

$$
\mu_X = E[X] = np = n \cdot \frac{a}{N} .
$$

This gives the formula for the mean of a hyper-geometric. To get the variance of a hyper-geometric random variable we saw

$$
V[X] = \sigma_X^2 = E[X^2] - \mu_X^2
$$

holds for any random variable $X$. Now for $X = \sum_i X_i$ write

$$
E[X^2] = \sum_i E[X_i^2] + \sum_{j \neq k} E[X_j X_k].
$$

Since for a 0 or 1 valued r.v. we have $X_i^2 = X_i$ the first sum on the right equals $E[X] = np$. In the second sum there are $n(n-1)$ identical terms in the double sum over all $j$ and $k$ with $j \neq k$ since for each such term $X_j X_k$ is either 1 or 0 and equals 1 only when both $X_j$ and $X_k$ are equal to 1 so that
\[ E[X_j X_k] = P(X_j = 1 \text{ and } X_k = 1) = P(X_j = 1) \cdot P(X_k = 1 | X_j = 1) = \frac{a}{N} \cdot \frac{a-1}{N-1}. \]

This last calculation is entirely analogous to the probability of selecting two aces from a 52 card deck being \( \frac{4}{52} \cdot \frac{3}{51} \) since conditioned on having selected an ace for the first selection, for the second selection to be an ace there are now 3 aces remaining in a deck with 51 cards left in it. Putting it all together it is now just a question of algebra to show that the variance

\[ \sigma^2_X = np + n(n-1) \frac{a}{N} \frac{a-1}{N-1} - (np)^2 \]

reduces to the formula for the variance of a hyper-geometric random variable:

\[ \sigma^2_X = np(1-p) \left( \frac{N-n}{N-1} \right) \text{ where } p = \frac{a}{N} \]

when we substitute \( a=NP \) above. The finite population correction factor \( \frac{N-n}{N-1} \) is a correction to the formula for the variance of a binomial random variable due to sampling without replacement from a finite population (i.e. the hyper-geometric case).