Mean and Standard Deviation of Binomial Distribution

Binomial distribution comes from a series of Bernoulli trials (an experiment with two outcomes, "success" and "failure", where we translate "success" as the number 1, and "failure" as zero). The basic question: How many successes can we expect in N trials? We assume that on any one trial, the probability of success = p and of failure = 1-p, and also that the successive trials are independent.

TWO questions arise:
what is the mean and SD of each individual trial?
what is the expected value and standard error of a SUM of trials? How many successes can we expect if we play the game N times?

INDIVIDUAL TRIALS

To model the individual trials, set up a box model with tickets labelled 0 and 1, for "success" and "failure" For a fair coin, there will be the same number of tickets labelled 0 and 1, and it does not matter if there is only one ticket of each variety or one million of each variety. You can find the mean and standard deviation by direct computation (at least if you pick the model with two tickets):

Mean of box model = 1 + 0 / 2 = 0.5
Sum of squared deviations from mean = (square (1 - 0.5)) + (square (0 - 0.5))
= (square 0.5) + (square -0.5) = .25 + .25 = 0.5
Variance (mean squared deviation) = 0.5 / 2 = .25
Standard deviation = (sqrt .25) = .5

Satisfy yourself that we would get the same result for mean and SD if the contents of the box were [ 0 1 0 1]

It is convenient to have a formula capable of handling unfair coins, but this can lead to awkward computation problems -- if the probability of success is 0.6725, we need a box with 10,000 tickets, 6,725 labeled "1" and the other 3,275 labelled "0". And what if we are interested not in the fact of success (0 or 1) but the payoff from success (failure loses 5 dollars, success wins 2 dollars). What is the mean and SD of this box? Direct calculation, clearly, would take a long time.

Generalize by setting up a box with B total numbers; let A of the numbers be "large" (+ 2 in the example above), and B - A of the numbers be "small" (- 5 in the example above). The choices should reflect the percentages given: the probability of success on any one trial p = A / B and the probability of failure (1 - p) = q = B - A/B
Let the value of the "large" number = L, and the value of the "small" number = S.

Since there are A large numbers, the sum of the large numbers will equal A * L, and since there are B - A small numbers, the sum of the small numbers will equal (B - A) * S
The mean of the box will be: [A*L + (B - A) * S ] / B = [A/B] * L + [(B - A) / B] * S = p * L + (1 - p) * S.

For the binomial distribution, L = 1 and S = 0, so the mean of the box will be p*1 + (1-p)*0 = p

The formula for the SD is harder to derive, but easy to apply:
SD (of box with two values) = (L - S) * (sqrt [p * (1 - p)])

The derivation is on the reverse; working through it helps keep it in mind (but the derivation won't be "on the test"): 
Note first that the deviations of the "large" numbers from the mean of \( pL - (1-p) S \) will be:

\[
L - [p L + (1-p) S] = L - p L - (1-p) S = (1 - p) (L - S)
\]

and the deviations of the small numbers will be:

\[
S - [p L + (1 - p) S] = S - p L - (1-p) S = p (S - L)
\]

So the sum of squared deviations will be:

\[
A \ (1-p)^2 (L - S)^2 + (B - A) \ p^2 (S - L)^2
\]

and the variance (mean squared deviation) will be that divided by the number of numbers in the box, \( B \). But \( A/B = p \) and \( (B - A)/B = 1-p \), so the variance is:

\[
p \ (1-p)^2 (L - S)^2 + (1-p) \ p^2 (S - L)^2
\]

Factor out \( p(1-p) \) and note that the square of \( L - S \) = the square of \( S - L \), so:

\[
p \ (1-p) \ [(1-p) (L - S)^2 + p (L - S)^2]
\]

and the final result is that the variance is:

\[
p \ (1-p) \ [(L - S)^2 - p (L - S)^2 + p (L - S)^2] = p \ (1-p) (L - S)^2
\]

The SD of the box is the square root of the above, giving us the formula we wanted to derive:

\[
SD \ (of \ box \ with \ two \ values) = (L - S) \ast (sqrt \ [p \ast \ (1 - p)])
\]
EXPECTED VALUE and STANDARD ERROR of SUMS

We are often interested not in a single draw from the box, but the results of a series of draws: the sum of our winnings after we bet on many rolls of the dice or spins of a roulette wheel or the appreciation of a portfolio of stocks. It is very important to realize that the properties of the SUMS are very different from the properties of a SINGLE DRAW.

The box model is, in a sense, our model of "reality" -- the dice are fair, and may be represented by 36 tickets with 1 twelve, 2 elevens, and so on. The box model can also be regarded as our model of the population -- for example, a population of 100 million voters, 55 percent of whom will vote Democratic, can be modeled by a box model with 55 tickets marked 1 and 45 tickets marked zero. But we do not directly know reality or the population; what we know is the results of a few (or a few thousand) rolls of the dice or telephone calls to the voters we have polled. In order to draw sensible conclusions from the experiments conducted on the dice, or from the poll results, we must know how the results relate to reality, or how the sample relates to the population.

We will summarize experiments or polling samples with statistics, and the key statistics are based on sums (means, standard deviations, percentages). The properties of sums of many draws from a box are therefore critical for interpreting statistics.

To stress the difference between reality and the statistics summarizing reality, we talk of the mean and standard deviation of the population, but of the expected value and standard error of statistics based on a sample of that population.

The key relationships are straightforward:

\[
\begin{align*}
\text{EV of sum of } N \text{ draws} &= N \times \text{(mean of the box model)} = N \times p \text{ for the binomial.} \\
\text{Var of sum of } N \text{ draws} &= N \times \text{(variance of the box model)} = N \times p \times q \text{ for the binomial.} \\
\text{SE of sum of } N \text{ draws} &= \sqrt{\text{variance of the box model}} = \sqrt{N \times p \times q} \text{ for the binomial.}
\end{align*}
\]

Demonstration for the binomial case, considering \(N\) draws from a box with six tickets labeled zero and four tickets labeled 1. Mean of the box = 0.4 and SD of the box = \(\sqrt{0.24} = .4899\)

\[
\begin{align*}
\text{(bind b (list 0 0 0 0 0 0 1 1 1 1))} \\
\text{(bind sums nil)} &; \text{variable created as an empty list} \\
\text{(dotimes (i 10000) (push (sum (draw N b) sums)))} &; \text{loop which sums } N \text{ draws from the box and pushes it onto} \\
\text{(mean sums)} &; \text{the variable sums} \\
\text{(sd sums)}
\end{align*}
\]

Do this procedure for \(N = 25\), when the formulas tell you to expect mean of sums = 25 \(\times 0.4 = 10\) and SE of sums = \((\sqrt{25}) \times (\sqrt{0.6 \times 0.4}) = 2.4495\)

(My run got (mean sums) = 9.9934 and SE of sums = (sd sums) = 2.4395)

Note that you use (SD sums) to get what the text calls the "standard error" of the sums.

Do it again for \(N = 100\), when the formulas tell you to expect mean of sums = 10 \(\times 0.4 = 40\) and SE of sums = \((\sqrt{100}) \times (\sqrt{0.6 \times 0.4}) = 10 \times (\sqrt{0.24}) = 4.8990\)

Remember to start by setting the SUMS variable to an empty list again: (bind sums nil) (dotimes (i 10000) (push (sum (draw 100 b)) sums)).

My run got (mean sums) = 40.0194 and (sd sums) = 4.8663
If you still don't believe the formula works, try it for N = 225, 625, 900 ...
Also see chapter 17 review, problem 1, for a similar exercise.

Or consider the basic logic for two draws from the box:

You will have TWO random variables, representing the results of the two draws, which we can label X1 and X2. Each of course will have the expected value of the box, and the SD for each will be the SD of the box. For a binomial distribution, the EV of X1 is p and the EV of X2 is p, so the EV of X1 + X2 = 2p

For the variance, note that the variance of the box is pq, and that variances are additive IF the draws are independent (this is an important condition), so Var (X1 + X2) = 2 * p q
Take the square root of this to get SE (X1 + X2) = (sqrt 2) * (sqrt p q)

A more general proof is for two variables, and can be extended to any number of INDEPENDENT draws.

EV (X1 + X2) = EV (X1) + EV (X2)

Var (X1 + X2) = (square (X1 + X2 - EV(X1 + X2))) / 2
= (square (X1dev + X2dev)) if we define X1dev = X1 - mean X1 and
= (square X1dev) / 2 + (square X2dev) / 2 + (2 * X1dev * X2dev) / 2
= Var X1 + Var X2 + 2 * Cov (X1 X2)

The importance of the assumption of independence is shown here: if two variables are independent, their covariance is zero. If they are NOT independent, their covariance will be either positive or negative, and we must be very careful in drawing conclusions about the variance of the sums.