# 4.3: Binomial Distributions

**Learning Objectives**

- Define the binomial random variable
- Construct binomial distributions
- Develop and use the probability distribution function for binomial random variables
- Provide and use alternative formulas for expected value and variance of binomial random variables
- Assess the necessity of independent trials

## Binomial Random Variable

Suppose we had a weighted coin which only came up heads one-sixth of the time. If we flipped this coin $10$ times, what is the probability that exactly $2$ of those flips would be heads? In order to answer this question, we would need to recognize that each of the coin flips are independent and that the probability of getting heads is the same each time. The answer turns out to be about $29\%$; we will demonstrate how to compute this soon.

Now suppose we were rolling a fair die $10$ times. What is the probability that exactly $2$ of those rolls would be six? The astute reader may notice that the answer is the same: $29\%$. Why is that? Well, we either roll a six or we don't. Rolling a six is analogous to flipping heads and not rolling a six is analogous to flipping tails. Since each trial is independent and the probability of obtaining the outcome of interest is $\frac{1}{6}$, the two scenarios are the same from the perspective of probability.

In fact, we can be much more general. Suppose in a population consisting of millions of people, exactly one in six of them support some political candidate. If we randomly and independently selected $10$ people, what is the probability that $2$ of them would be supporters of the candidate? Suppose a factory produces light bulbs and one-sixth of them are dysfunctional. If an inspector were to randomly and independently select $10$ of them, what is the probability that exactly $2$ of them would be dysfunctional? Suppose an archer hits the bulls-eye with probability $\frac{1}{6}$ every time she shoots. If she takes $10$ shots, what is the probability that she gets exactly $2$ bulls-eyes? The answers to all of these questions are the same: $29\%$. It is clear that some of the details of the situations are irrelevant; all that matters is that a trial is repeated $10$ times, each trial is independent, and the probability of the outcome of interest occurring is $\frac{1}{6}$ each time. These sorts of situations are the object of our discussion: binomial random variables.

Binomial distributions are the probability distributions for a particular type of discrete random variable: the binomial random variable. With binomial random variables, we are considering a single random experiment repeated, identically and independently, a fixed number of times. We call each repetition a trial and indicate the number of trials with $n$. As the adjective "binomial" indicates, we group the outcomes into two categories: successes and failures. The probability of a success on any given trial is denoted $p$, while the probability of a failure on any given trial is denoted $q$. Note that since we have only two categories covering the entire sample space, $q = 1 - p$. We define the **binomial random variable $X$** as the number of successes throughout all $n$ trials. Every trial may fail, in which case, $X = 0$. On the other hand, every trial may be a success, in which case, $X = n$. In most cases, some trials will succeed while others fail. As such, $X$ takes on any number in the set $\{0,1,2,3,\ldots,n\}$. In the examples discussed above, $n = 10$, $p = \frac{1}{6}$, $q = \frac{5}{6}$, and we were asking what is $P(X = 2)$.

Consider an example to help solidify these ideas. In a previous text exercise, we considered tossing a fair die three times and determined the probability of getting one or more throws landing with one face up. We can understand this situation as a binomial random variable. Our underlying random experiment is rolling a fair die. We fix the number of trials to $3$. The trials are identical because we are similarly rolling the same die each time. The trials are independent because the outcomes of previous rolls do not affect current or future rolls. Since we are interested in rolling ones, we define that as a success. Rolling any other value ($2$, $3$, $4$, $5$, or $6$) constitutes a failure. We can easily compute the probabilities of success and failure on any individual trial; $p = \frac{1}{6}$ and $q = \frac{5}{6}$. We define our binomial random variable $X$ to be the number of ones rolled in $3$ tosses of a fair die. The possible values for $X$ are $0$, $1$, $2$, and $3$. Recall that our interest in random variables lies in their probability distributions. We will now address constructing a binomial random variable's probability distribution.

## Probability Distributions of Binomial Random Variables

We first build our intuition by constructing the probability distribution of our binomial random variable $X$ : the number of ones rolled in $3$ tosses of a fair die. When determining the probability for a particular value of a random variable, we generally considered all of the outcomes in the sample space and proceeded from there. Considering all three trials based on the values landing up would result in $6^3 = 216$ different outcomes. We can simplify our analysis by considering all three trials based on successes and failures; that is, rolling a one is considered a success and rolling anything other than a one is considered a failure. In this case, we only have $2^3 = 8$ considerations. We shall use S to indicate a trial with success and F to indicate a trial with a failure, and represent the possibility of a successful trial followed by failures on the second and third trials as SFF. This procedure is illustrated for several, but not all, possible outcomes in the figure below.

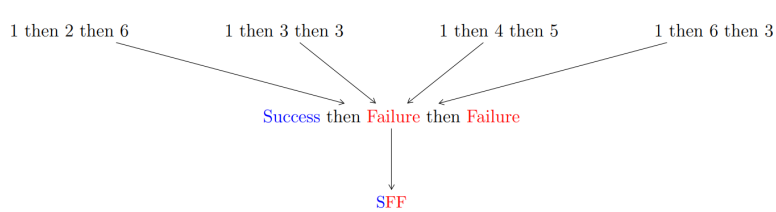


Figure $4.3.1$: Outcomes of three rolls in succession understood in terms of successes and failures

The following figure groups the $8$ possibilities by value of $X$ and helps us build the probability distribution.

| | | | |
|---|---|---|---|
| $X=0$ | FFF | | |
| $X=1$ | SFF | FSF | FFS |
| $X=2$ | SSF | SFS | FSS |
| $X=3$ | SSS | | |

Table 4.3.1: Initial probability distribution for the random variable $X$

| $X=x_j$ | $P(X=x_j)$ |
|---|---|
| 0 | $P(\text{FFF})$ |
| 1 | $P(\text{SFF or FSF or FFS})$ |
| 2 | $P(\text{SSF or SFS or FSS})$ |
| 3 | $P(\text{SSS})$ |

Each trial occurs in sequence and is identical and independent; we can use both our addition and multiplication rules for probabilities to determine our probabilities. Remember that $P(\text{S})=p=\frac{1}{6}$ and $P(\text{F})=q=\frac{5}{6}$.

Table 4.3.2: Probability distribution for the random variable $X$

| $X=x_j$ | $P(X=x_j)$ | |
|---|---|---|
| 1 | $\begin{aligned} P(\text{SFF or FSF or FFS}) &= P(\text{SFF})+P(\text{FSF})+P(\text{FFS}) \\ &= P(\text{S})\cdot P(\text{F})\cdot P(\text{F})+P(\text{F})\cdot P(\text{S})\cdot P(\text{F})+P(\text{F})\cdot P(\text{F})\cdot P(\text{S}) \\ &= 3P(\text{S})\cdot P(\text{F})^2 \\ &= 3pq^2 \end{aligned}$ | $3\cdot\frac{1}{6}\cdot\left(\frac{5}{6}\right)^2=\frac{75}{216}\approx 34.7222\%$ |
| 3 | $\begin{aligned} P(\text{SSS}) &= P(\text{S})\cdot P(\text{S})\cdot P(\text{S}) \\ &= P(\text{S})^3 \\ &= p^3 \end{aligned}$ | $\left(\frac{1}{6}\right)^3=\frac{1}{216}\approx 0.4630\%$ |
| 2 | $\begin{aligned} P(\text{SSF or SFS or FSS}) &= P(\text{SSF})+P(\text{SFS})+P(\text{FSS}) \\ &= P(\text{S})\cdot P(\text{S})\cdot P(\text{F})+P(\text{S})\cdot P(\text{F})\cdot P(\text{S})+P(\text{F})\cdot P(\text{S})\cdot P(\text{S}) \\ &= 3P(\text{S})^2\cdot P(\text{F}) \\ &= 3p^2q \end{aligned}$ | $3\cdot\left(\frac{1}{6}\right)^2\cdot\frac{5}{6}=\frac{15}{216}\approx 6.9444\%$ |
| 0 | $\begin{aligned} P(\text{FFF}) &= P(\text{F})\cdot P(\text{F})\cdot P(\text{F}) \\ &= P(\text{F})^3 \\ &= q^3 \end{aligned}$ | $\left(\frac{5}{6}\right)^3=\frac{125}{216}\approx 57.8704\%$ |

**? Text Exercise 4.3.1**

Consider the random variable $Y$ that counts the number of heads in 4 flips of a fair coin. Verify that the random variable $Y$ is a binomial random variable and construct its probability distribution.

**Answer**

The underlying random experiment is the flipping of the fair coin, which is to be repeated a fixed number of times; $n=4$. We are flipping the same coin in a similar fashion, meaning our trials are identical. We have independent trials because the outcome of one trial does not affect any of the other trials. We are counting the number of heads; in a successful trial, heads land face up, and alternatively, landing a tail would be a failure. Since we are using a fair coin, we have $P(\text{S})=P(\text{F})=\frac{1}{2}$ or $p=q=\frac{1}{2}$ confirming that $Y$ is a binomial random variable. Since we have 4 trials, the set of possible values for $Y$ is $\{0,1,2,3,4\}$. To construct the probability distribution, we consider the possible outcomes of all 4 trials in terms of successes and failures.

| | | | | | | |
|---|---|---|---|---|---|---|
| $Y=0$ | FFFF | | | | | |
| $Y=1$ | SFFF | FSFF | FFSF | FFFS | | |
| $Y=2$ | SSFF | SFSF | SFFS | FSSF | FSFS | FFSS |
| $Y=3$ | SSSF | SSFS | SFSS | FSSS | | |
| $Y=4$ | SSSS | | | | | |

Table 4.3.3 Initial probability distribution for the random variable $Y$

| $Y=y_j$ | $P(Y=y_j)$ |
|---|---|
| 0 | $P(\text{FFFF})$ |
| 1 | $P(\text{SFFF or FSFF or FFSF or FFFS})$ |
| 2 | $P(\text{SSFF or SFSF or SFFS or FSSF or FSFS or FFSS})$ |

| $Y = y_j$ | $P(Y = y_j)$ |
|---|---|
| 3 | $P(\text{SSSF or SSFS or SFSS or FSSS})$ |
| 4 | $P(\text{SSSS})$ |

Again, each trial occurs in sequence and is identical and independent; we use both our addition and multiplication rule for probabilities to determine our probabilities.

Table 4.3.4 Probability distribution for the random variable $Y$

| $Y = y_j$ | $P(Y = y_j)$ | |
|---|---|---|
| 0 | $P(\text{FFFF}) = P(\text{F}) \cdot P(\text{F}) \cdot P(\text{F}) \cdot P(\text{F}) \cdot$ $= P(\text{F})^4$ $= q^4$ | $\left(\frac{1}{2}\right)^4 = \frac{1}{16} = 6.25\%$ |
| 1 | $P(\text{SFFF or FSFF or FFSF or FFFS}) = P(\text{SFFF}) + P(\text{FSFF}) + P(\text{FFSF}) + P(\text{FFFS})$ $= P(\text{S}) \cdot P(\text{F}) \cdot P(\text{F}) \cdot P(\text{F}) + P(\text{F}) \cdot P(\text{S}) \cdot P(\text{F}) \cdot P(\text{F})$ $+ P(\text{F}) \cdot P(\text{F}) \cdot P(\text{S}) \cdot P(\text{F}) + P(\text{F}) \cdot P(\text{F}) \cdot P(\text{F}) \cdot P(\text{S})$ $= 4P(\text{S}) \cdot P(\text{F})^3$ $= 4pq^3$ | $4 \cdot \left(\frac{1}{2}\right) \cdot \left(\frac{1}{2}\right)^3 = \frac{4}{16} = 25\%$ |
| 2 | $P(\text{SSFF or SFSF or SFFS or FSSF or FSFS or FFSS}) = P(\text{SSFF}) + P(\text{SFSF})$ $+ P(\text{SFFS}) + P(\text{FSSF})$ $+ P(\text{FSFS}) + P(\text{FFSS})$ $= P(\text{S}) \cdot P(\text{S}) \cdot P(\text{F}) \cdot P(\text{F}) + P(\text{S}) \cdot P(\text{F}) \cdot P(\text{F}) \cdot P(\text{S})$ $+ P(\text{S}) \cdot P(\text{F}) \cdot P(\text{F}) \cdot P(\text{S}) + P(\text{F}) \cdot P(\text{S}) \cdot P(\text{S}) \cdot P(\text{F})$ $+ P(\text{F}) \cdot P(\text{S}) \cdot P(\text{F}) \cdot P(\text{S}) + P(\text{F}) \cdot P(\text{F}) \cdot P(\text{S}) \cdot P(\text{S})$ $= 6P(\text{S})^2 \cdot P(\text{F})^2$ $= 6p^2q^2$ | $6 \cdot \left(\frac{1}{2}\right)^2 \cdot \left(\frac{1}{2}\right)^2 = \frac{6}{16} = 37.5\%$ |
| 3 | $P(\text{SSSF or SSFS or SFSS or FSSS}) = P(\text{SSSF}) + P(\text{SSFS}) + P(\text{SFSS}) + P(\text{FSSS})$ $= P(\text{S}) \cdot P(\text{S}) \cdot P(\text{S}) \cdot P(\text{F}) + P(\text{S}) \cdot P(\text{S}) \cdot P(\text{F}) \cdot P(\text{S})$ $+ P(\text{S}) \cdot P(\text{F}) \cdot P(\text{S}) \cdot P(\text{S}) + P(\text{F}) \cdot P(\text{S}) \cdot P(\text{S}) \cdot P(\text{S})$ $= 4P(\text{S})^3 \cdot P(\text{F})$ $= 4p^3q$ | $4 \cdot \left(\frac{1}{2}\right)^3 \cdot \left(\frac{1}{2}\right) = \frac{4}{16} = 25\%$ |
| 4 | $P(\text{SSSS}) = P(\text{S}) \cdot P(\text{S}) \cdot P(\text{S}) \cdot P(\text{S}) \cdot$ $= P(\text{S})^4$ $= p^4$ | $\left(\frac{1}{2}\right)^4 = \frac{1}{16} = 6.25\%$ |

Hopefully, we have noticed some patterns after building probability distributions for two binomial random variables. Let us formulate the patterns in the context of a general binomial random variable $X$ with $n$ trials and probability of success on any individual trial $p$. Recall that $q$, the probability of failure, is $1 - p$.

When we consider the possible outcomes of all trials in terms of successes and failures, the probabilities depend on the number of successes and failures, not on the order in which those successes and failures appear. Each event in $X = x_j$ has the same probability. If there are $x_j$ successes, meaning we have $n - x_j$ failures, the probability of each event in $X = x_j$ is $p^{x_j} q^{n-x_j}$. All that is left to do is count the number of such events for any particular value $x_j$.

To count the number of ways that $x_j$ successes can be assigned to the $n$ trials, we can use combinations: ${}_nC_{x_j}$. We have ${}_nC_{x_j}$ many events in $X = x_j$ each with a probability of $p^{x_j} q^{n-x_j}$. Putting this all together, we arrive at a function that returns the probability of our binomial random variable. We call this the **probability distribution function for a binomial random variable $X$.**

$$P(X = x_j) = {}_nC_{x_j}\, p^{x_j} q^{n-x_j}$$

Check that the formula works by using it on the preceding example.

**? Text Exercise 4.3.2**

1. A virtual education company produces short multiple-choice quizzes for each content module. They currently have 5 questions with 2 options for each question. One school that uses this product worries about students passing these quizzes without learning the content. Determine the probability of a student obtaining an A or a B (obtaining at least an 80%,) on such a quiz by literally randomly guessing on each question.

**Answer**

We can understand this situation as a binomial random variable. We have a random experiment of a student randomly guessing on a multiple-choice question. The experiment is repeated 5 times because there are 5 questions. Since the student is randomly guessing on each question, the trials are identical and independent, with a probability of success at $50\%$. Let $X$ be the binomial random variable that counts the number of correct guesses on these 5 question multiple-choice quizzes. This means a student needs 4 or 5 correct answers to obtain an A or a B. We need to find $P(X=4 \text{ or } 5$. We can use the probability distribution function to find the answer. Recall that $n=5$ and $p=0.5$. Thus,

$$P(X=4)={}_5C_4\cdot\left(\frac{1}{2}\right)^4\cdot\left(\frac{1}{2}\right)^{5-4}=5\cdot\left(\frac{1}{16}\right)\cdot\left(\frac{1}{2}\right)=\frac{5}{32}$$
$$P(X=5)={}_5C_5\cdot\left(\frac{1}{2}\right)^5\cdot\left(\frac{1}{2}\right)^{5-5}=1\cdot\left(\frac{1}{32}\right)\cdot\left(\frac{1}{2}\right)^0=\frac{1}{32}$$

Thus, $P(X=4 \text{ or } 5)=\frac{5}{32}+\frac{1}{32}=\frac{6}{32}=18.75\%$.

2. Given the analysis in the first part of this text exercise, the virtual education company has decided to increase the number of options on each question while keeping the number of questions fixed at 5. They are considering using 3 or 4 options. Determine the probability that a student randomly guessing on a quiz will obtain an A or a B under both options.

**Answer**

Changing the number of options does not change the number of questions necessary, but it does change the probability of success on any given question. Let $Y$ be the binomial random variable counting the number of correct guesses when there are 3 options on each question and $Z$ be for 4 options. Thus, we are interested in $P(Y=4 \text{ or } 5)$ and $P(Z=4 \text{ or } 5)$. When there are 3 options, the probability of success, $p_Y$ is only $\frac{1}{3}$. Similarly, when there are 4 options, the probability of success, $p_Z$ is only $\frac{1}{4}$.

$$P(Y=4)={}_5C_4\cdot\left(\frac{1}{3}\right)^4\cdot\left(\frac{2}{3}\right)^{5-4}=5\cdot\left(\frac{1}{81}\right)\cdot\left(\frac{2}{3}\right)=\frac{10}{243}$$
$$P(Y=5)={}_5C_5\cdot\left(\frac{1}{3}\right)^5\cdot\left(\frac{2}{3}\right)^{5-5}=1\cdot\left(\frac{1}{243}\right)\cdot\left(\frac{2}{3}\right)^0=\frac{1}{243}$$
$$P(Z=4)={}_5C_4\cdot\left(\frac{1}{4}\right)^4\cdot\left(\frac{3}{4}\right)^{5-4}=5\cdot\left(\frac{1}{256}\right)\cdot\left(\frac{3}{4}\right)=\frac{15}{1024}$$
$$P(Z=5)={}_5C_5\cdot\left(\frac{1}{4}\right)^5\cdot\left(\frac{3}{4}\right)^{5-5}=1\cdot\left(\frac{1}{1024}\right)\cdot\left(\frac{3}{4}\right)^0=\frac{1}{1024}$$

Thus, $P(Y=4 \text{ or } 5)=\frac{10}{243}+\frac{1}{243}=\frac{11}{243}\approx 4.5267\%$ and $P(Z=4 \text{ or } 5)=\frac{15}{1024}+\frac{1}{1024}=\frac{16}{1024}\approx 1.5625\%$. Increasing the number of options significantly reduces the chances of a student obtaining an A or a B on a quiz by randomly selecting answers. We go from nearly $19\%$ to just below $5\%$ to just over $1\%$.

## Expected Value, Variance, and Standard Deviation of Binomial Random Variables

Remember that binomial random variables are just a particular type of discrete random variable. That means everything we know about discrete random variables applies to binomial random variables. Binomial random variables have some very nice properties that make the calculations of expected value and variance much easier. Note that the formulas we develop here in this section only apply to binomial random variables and not all discrete random variables.

**? Text Exercise 4.3.3**

Using the definitions of expected value, variance, and standard deviation provided in the section on discrete random variables, determine these measures of centrality and dispersion for the binomial random variables: $X$ being the number of ones rolled in 3 tosses of a fair die and $Y$ being the number of heads in 4 flips of a fair coin.

**Answer**

These are the same random variables that we have been using throughout this section. We can utilize the probability distributions that we have already created.

Table 4.3.5 Table of computation for the random variable $X$

| $X=x_j$ | $P(X=x_j)$ | $x_j\cdot P(X=x_j)$ | $(x_j-\mu)^2\cdot P(X=x_j)$ |
|---|---|---|---|
| 0 | $\frac{125}{216}$ | $0\cdot\frac{125}{216}=0$ | $\left(0-\frac{1}{2}\right)^2\cdot\frac{125}{216}=\frac{1}{4}\cdot\frac{125}{216}=\frac{125}{864}$ |
| 1 | $\frac{75}{216}$ | $1\cdot\frac{75}{216}=\frac{75}{216}$ | $\left(1-\frac{1}{2}\right)^2\cdot\frac{75}{216}=\frac{1}{4}\cdot\frac{75}{216}=\frac{75}{864}$ |
| 2 | $\frac{15}{216}$ | $2\cdot\frac{15}{216}=\frac{30}{216}$ | $\left(2-\frac{1}{2}\right)^2\cdot\frac{15}{216}=\frac{9}{4}\cdot\frac{15}{216}=\frac{135}{864}$ |

| $X=x_j$ | $P(X=x_j)$ | $x_j \cdot P(X=x_j)$ | $(x_j-\mu)^2 \cdot P(X=x_j)$ |
|---|---|---|---|
| $3$ | $\frac{1}{216}$ | $3 \cdot \frac{1}{216}=\frac{3}{216}$ | $\left(3-\frac{1}{2}\right)^2 \cdot \frac{1}{216}=\frac{25}{4} \cdot \frac{1}{216}=\frac{25}{864}$ |
| $\mu=\mathrm{E}(X)=0+\frac{75}{216}+\frac{30}{216}+\frac{3}{216}=\frac{108}{216}=\frac{1}{2}$ | | | |
| $\sigma^2=\mathrm{Var}(X)=\frac{125}{864}+\frac{75}{864}+\frac{135}{864}+\frac{25}{864}=\frac{360}{864}=\frac{5}{12} \approx 0.4167$ | | | |
| $\sigma=\sqrt{\mathrm{Var}(X)}=\sqrt{\frac{5}{12}} \approx 0.6455$ | | | |

Table $4.3.6$ Table of computation for the random variable $Y$

| $Y=y_j$ | $P(Y=y_j)$ | $y_j \cdot P(Y=y_j)$ | $(y_j-\mu)^2 \cdot P(Y=y_j)$ |
|---|---|---|---|
| $0$ | $\frac{1}{16}$ | $0 \cdot \frac{1}{16}=0$ | $(0-2)^2 \cdot \frac{1}{16}=4 \cdot \frac{1}{16}=\frac{1}{4}$ |
| $1$ | $\frac{1}{4}$ | $1 \cdot \frac{1}{4}=\frac{1}{4}$ | $(1-2)^2 \cdot \frac{1}{4}=1 \cdot \frac{1}{4}=\frac{1}{4}$ |
| $2$ | $\frac{3}{8}$ | $2 \cdot \frac{3}{8}=\frac{3}{4}$ | $(2-2)^2 \cdot \frac{3}{8}=0 \cdot \frac{3}{8}=0$ |
| $3$ | $\frac{1}{4}$ | $3 \cdot \frac{1}{4}=\frac{3}{4}$ | $(3-2)^2 \cdot \frac{1}{4}=1 \cdot \frac{1}{4}=\frac{1}{4}$ |
| $4$ | $\frac{1}{16}$ | $4 \cdot \frac{1}{16}=\frac{1}{4}$ | $(4-2)^2 \cdot \frac{1}{16}=4 \cdot \frac{1}{16}=\frac{1}{4}$ |
| $\mu=\mathrm{E}(Y)=0+\frac{1}{4}+\cdots+\frac{1}{4}=\frac{8}{4}=2$ | | | |
| $\sigma^2=\mathrm{Var}(X)=\frac{1}{4}+\frac{1}{4}+\ldots+\frac{1}{4}=\frac{4}{4}=1$ | | | |
| $\sigma=\sqrt{\mathrm{Var}(X)}=\sqrt{1}=1$ | | | |

Having computed the expected value, variance, and standard deviation for two binomial random variables using the definitions, we now present quicker and easier methods for computing the expected value and variance. Just as with the alternative formula for the variance of a discrete random variable, these formulas are derived from our original definitions through mathematical simplification and produce the same values as the original definitions. We will not provide the work for this mathematical simplification but will provide a little intuition before providing the formulas. For example, if $p=0.5$ and $n=10$, then we are repeating a trial 10 times with probability of success being $0.5$. We should expect, then, that half of the time, we will succeed. This means $E(X)=0.5 \cdot 10=5$. Similarly, if $p=0.9$ and $n=100$, we should expect to see success $90\%$ of the time, so $E(X)=0.9 \cdot 100=90$. In general, $E(X)=np$ for binomial distributions. For a binomial random variable $X$ with $n$ trials, probability of success on any individual trial $p$, and probability of failure on any individual trial $q$, we can compute the expected value and variance using the following formulas.

$$\mu=\mathrm{E}(X)=np$$
$$\sigma^2=\mathrm{Var}(X)=npq$$

**? Text Exercise $4.3.4$**

Using the above formulas, compute the expected value and variance for the random variables: $X$ being the number of ones rolled in 3 tosses of a fair die and $Y$ being the number of heads in 4 flips of a fair coin. Verify that the values match what was computed in the previous text exercise.

**Answer**

When considering the random variable $X$, we have that $n=3$, $p=\frac{1}{6}$, and $q=\frac{5}{6}$. We thus compute $\mu=\mathrm{E}(X)=3 \cdot \frac{1}{6}=\frac{1}{2}$ and $\sigma^2=\mathrm{Var}(X)=3 \cdot \frac{1}{6} \cdot \frac{5}{6}=\frac{5}{12}$. These values match what was computed in the previous exercise.

When considering the random variable $Y$, we have that $n=4$, $p=\frac{1}{2}$, and $q=\frac{1}{2}$. We thus compute $\mu=\mathrm{E}(Y)=4 \cdot \frac{1}{2}=2$ and $\sigma^2=\mathrm{Var}(X)=4 \cdot \frac{1}{2} \cdot \frac{1}{2}=1$. These values again match what was computed in the previous exercise.

## Necessity of Independent Trials

Binomial distributions are related to important distributions in inferential statistics, such as computing the probability of obtaining a sample with a particular proportion. Recall our discussion regarding obtaining a random sample from a large population and having $80\%$ of them be women. The probability of this happening was significantly less with a sample size of 20 as opposed to 10 ($0.4621\%$ vs $4.3945\%$). These probabilities were computed using the binomial distribution. Here, we treated our random experiment as selecting an individual from a large enough population composed of equal numbers of men and women. We considered selecting a woman a success and treated $p=q=\frac{1}{2}$. In the case of a sample of size of $10$, we noted that $80\%$ of 10 is $8$. And in the case of a sample size of $20$, we need $16$ women to get $80\%$. However, this fails to satisfy our definition of a binomial random variable because we do not have

the same probabilities of success and failure for each trial. When one person is chosen, that person is no longer eligible to be chosen for subsequent trials. We have fewer people to choose from and no longer equal numbers of men and women. Our trials are not independent.

We have run into this issue previously in a text exercise. When populations are huge (when the difference between an event's probability and a conditional probability related to that event is relatively small) treating the events as if they were independent will result in a value which is approximately, not exactly, correct. Since it is often much easier to compute assuming independence, this is common practice when the error would be negligible. It is difficult to define exactly how large a population must be, in general, for the assumption of independence to be reasonable. For example, if there are $1,000,000$people, exactly half of which are women, and we randomly select $2$ individuals from this group, the probability that they are both women would be $\frac{500,000}{1,000,000} \cdot \frac{499,999}{999,999} \approx 0.24999975$. If we had assumed independence, that is, that each time we selected a person, there was a $50\%$ chance it was a woman, we would have obtained $\frac{1}{2} \cdot \frac{1}{2} = 0.25$. Notice the error we get from assuming independence is quite small. On the other hand, if the population size were $6$ and $3$ of them were women, the assumption of independence is much less reasonable. If we randomly select $2$ people from this group of $6$, the probability that they are both women is $\frac{3}{6} \cdot \frac{2}{5} = \frac{1}{5} = 0.2$. Simply saying there's a $50\%$ chance each time obtains an estimate of $0.25$. Notice the error is much larger than before. If we take our population size to be even smaller, the error gets larger. In summary, if the sample we are selecting is a tiny proportion of the population, then assuming independence introduces little error; however, if we assume independence when the sample is a significant proportion of the population, then we will have large errors in our estimates. The following exercise illustrates in more detail how much error there is in different population sizes.

**? Text Exercise 4.3.5**

1. Consider sampling $10$ people from a population composed of an equal number of men and women. We denote the outcome of such a sampling as a sequence of W and M. Determine $P(\text{WWWWWWWWMM})$for each of population size.

   a. $N = 50$
   b. $N = 100$
   c. $N = 200$
   d. $N = 1000$

**Answer**

a. $P(\text{WWWWWWWWMM}) = \frac{25}{50} \cdot \frac{24}{49} \cdot \frac{23}{48} \cdot \frac{22}{47} \cdot \frac{21}{46} \cdot \frac{20}{45} \cdot \frac{19}{44} \cdot \frac{18}{43} \cdot \frac{25}{42} \cdot \frac{24}{41} \approx 0.0702\%$
b. $P(\text{WWWWWWWWMM}) = \frac{50}{100} \cdot \frac{49}{99} \cdot \frac{48}{98} \cdot \frac{47}{97} \cdot \frac{46}{96} \cdot \frac{45}{95} \cdot \frac{44}{94} \cdot \frac{43}{93} \cdot \frac{25}{92} \cdot \frac{24}{91} \approx 0.08443\%$
c. $P(\text{WWWWWWWWMM}) = \frac{100}{200} \cdot \frac{99}{199} \cdot \frac{98}{198} \cdot \frac{97}{197} \cdot \frac{96}{196} \cdot \frac{95}{195} \cdot \frac{94}{194} \cdot \frac{93}{193} \cdot \frac{100}{192} \cdot \frac{99}{191} \approx 0.09117\%$
d. $P(\text{WWWWWWWWMM}) = \frac{500}{1000} \cdot \frac{499}{999} \cdot \frac{498}{998} \cdot \frac{497}{997} \cdot \frac{496}{996} \cdot \frac{495}{995} \cdot \frac{494}{994} \cdot \frac{493}{993} \cdot \frac{500}{992} \cdot \frac{499}{991} \approx 0.09638\%$

2. Determine the $P(\text{WWWWWWWWMM})$as if each selection were independent with $P(\text{W}) = \frac{1}{2}$ and $P(\text{M}) = \frac{1}{2}$.

**Answer**

$P(\text{WWWWWWWWMM}) = \left(\frac{1}{2}\right)^{10} \approx 0.09766\%$

3. Compare the value computed in each part of part 1 with the value computed in part 2 of this text exercise.

**Answer**

a. The difference is $0.02746\%$.
b. The difference is $0.01323\%$.
c. The difference is $0.0065\%$.
d. The difference is $0.0013\%$.

The difference in computations of these values is in the hundredths and thousandths of a percent and decreases as the population increases. We only dealt with population sizes up to $1000$. In general, our populations of interest will be much larger than that so that we would expect even smaller differences. The comparison between sample size and population size is really at play down deep. Without going into the details, we share a fairly common recommendation. If the sample size is more than $5\%$ of the population, we do not assume independence.

4.3: Binomial Distributions is shared under a Public Domain license and was authored, remixed, and/or curated by LibreTexts.

- **5.7: Binomial Distribution** by David Lane is licensed Public Domain. Original source: https://onlinestatbook.com.