3.4 Binomial distribution (special topic)

Example 3.37 Suppose we randomly selected four individuals to participate in the "shock" study. What is the chance exactly one of them will be a success? Let's call the four people Allen (A), Brittany (B), Caroline (C), and Damian (D) for convenience. Also, suppose 35% of people are successes as in the previous version of this example.

Let's consider a scenario where one person refuses:

$$\begin{aligned} &P(A = \texttt{refuse},\ B = \texttt{shock},\ C = \texttt{shock},\ D = \texttt{shock}) \\ &\quad = P(A = \texttt{refuse})\ P(B = \texttt{shock})\ P(C = \texttt{shock})\ P(D = \texttt{shock}) \\ &\quad = (0.35)(0.65)(0.65)(0.65) = (0.35)^1(0.65)^3 = 0.096 \end{aligned}$$

But there are three other scenarios: Brittany, Caroline, or Damian could have been the one to refuse. In each of these cases, the probability is again $(0.35)^1(0.65)^3$. These four scenarios exhaust all the possible ways that exactly one of these four people could refuse to administer the most severe shock, so the total probability is $4 \times (0.35)^1(0.65)^3 = 0.38$.

Exercise 3.38 Verify that the scenario where Brittany is the only one to refuse to give the most severe shock has probability $(0.35)^1(0.65)^3$.[29]

3.4.1 The binomial distribution

The scenario outlined in Example 3.37 is a special case of what is called the binomial distribution. The **binomial distribution** describes the probability of having exactly k successes in n independent Bernoulli trials with probability of a success p (in Example 3.37, $n = 4$, $k = 1$, $p = 0.35$). We would like to determine the probabilities associated with the binomial distribution more generally, i.e. we want a formula where we can use n, k, and p to obtain the probability. To do this, we reexamine each part of the example.

There were four individuals who could have been the one to refuse, and each of these four scenarios had the same probability. Thus, we could identify the final probability as

$$[\#\text{ of scenarios}] \times P(\text{single scenario}) \tag{3.39}$$

The first component of this equation is the number of ways to arrange the $k = 1$ successes among the $n = 4$ trials. The second component is the probability of any of the four (equally probable) scenarios.

Consider $P(\text{single scenario})$ under the general case of k successes and $n - k$ failures in the n trials. In any such scenario, we apply the Multiplication Rule for independent events:

$$p^k(1-p)^{n-k}$$

This is our general formula for $P(\text{single scenario})$.

[29] $P(A = \texttt{shock},\ B = \texttt{refuse},\ C = \texttt{shock},\ D = \texttt{shock}) = (0.65)(0.35)(0.65)(0.65) = (0.35)^1(0.65)^3$.

Secondly, we introduce a general formula for the number of ways to choose k successes in n trials, i.e. arrange k successes and $n-k$ failures:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

The quantity $\binom{n}{k}$ is read **n choose k**.[30] The exclamation point notation (e.g. $k!$) denotes a **factorial** expression.

$$\begin{aligned}
0! &= 1 \\
1! &= 1 \\
2! &= 2 \times 1 = 2 \\
3! &= 3 \times 2 \times 1 = 6 \\
4! &= 4 \times 3 \times 2 \times 1 = 24 \\
&\vdots \\
n! &= n \times (n-1) \times ... \times 3 \times 2 \times 1
\end{aligned}$$

Using the formula, we can compute the number of ways to choose $k = 1$ successes in $n = 4$ trials:

$$\binom{4}{1} = \frac{4!}{1!(4-1)!} = \frac{4!}{1!3!} = \frac{4 \times 3 \times 2 \times 1}{(1)(3 \times 2 \times 1)} = 4$$

This result is exactly what we found by carefully thinking of each possible scenario in Example 3.37.

Substituting n choose k for the number of scenarios and $p^k(1-p)^{n-k}$ for the single scenario probability in Equation (3.39) yields the general binomial formula.

Binomial distribution

Suppose the probability of a single trial being a success is p. Then the probability of observing exactly k successes in n independent trials is given by

$$\binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \tag{3.40}$$

Additionally, the mean, variance, and standard deviation of the number of observed successes are

$$\mu = np \qquad \sigma^2 = np(1-p) \qquad \sigma = \sqrt{np(1-p)} \tag{3.41}$$

TIP: Is it binomial? Four conditions to check.

(1) The trials are independent.
(2) The number of trials, n, is fixed.
(3) Each trial outcome can be classified as a *success* or *failure*.
(4) The probability of a success, p, is the same for each trial.

[30] Other notation for n choose k includes ${}_nC_k$, C_n^k, and $C(n,k)$.

Example 3.42 What is the probability that 3 of 8 randomly selected students will refuse to administer the worst shock, i.e. 5 of 8 will?

We would like to apply the binomial model, so we check our conditions. The number of trials is fixed ($n = 8$) (condition 2) and each trial outcome can be classified as a success or failure (condition 3). Because the sample is random, the trials are independent (condition 1) and the probability of a success is the same for each trial (condition 4).

In the outcome of interest, there are $k = 3$ successes in $n = 8$ trials, and the probability of a success is $p = 0.35$. So the probability that 3 of 8 will refuse is given by

$$\begin{aligned}\binom{8}{3}(0.35)^3(1-0.35)^{8-3} &= \frac{8!}{3!(8-3)!}(0.35)^3(1-0.35)^{8-3} \\ &= \frac{8!}{3!5!}(0.35)^3(0.65)^5\end{aligned}$$

Dealing with the factorial part:

$$\frac{8!}{3!5!} = \frac{8\times7\times6\times5\times4\times3\times2\times1}{(3\times2\times1)(5\times4\times3\times2\times1)} = \frac{8\times7\times6}{3\times2\times1} = 56$$

Using $(0.35)^3(0.65)^5 \approx 0.005$, the final probability is about $56 * 0.005 = 0.28$.

TIP: computing binomial probabilities
The first step in using the binomial model is to check that the model is appropriate. The second step is to identify n, p, and k. The final step is to apply the formulas and interpret the results.

TIP: computing n choose k
In general, it is useful to do some cancelation in the factorials immediately. Alternatively, many computer programs and calculators have built in functions to compute n choose k, factorials, and even entire binomial probabilities.

Exercise 3.43 If you ran a study and randomly sampled 40 students, how many would you expect to refuse to administer the worst shock? What is the standard deviation of the number of people who would refuse? Equation (3.41) may be useful.[31]

Exercise 3.44 The probability that a random smoker will develop a severe lung condition in his or her lifetime is about 0.3. If you have 4 friends who smoke, are the conditions for the binomial model satisfied?[32]

---

[31] We are asked to determine the expected number (the mean) and the standard deviation, both of which can be directly computed from the formulas in Equation (3.41): $\mu = np = 40 \times 0.35 = 14$ and $\sigma = \sqrt{np(1-p)} = \sqrt{40 \times 0.35 \times 0.65} = 3.02$. Because very roughly 95% of observations fall within 2 standard deviations of the mean (see Section 1.6.4), we would probably observe at least 8 but less than 20 individuals in our sample who would refuse to administer the shock.

[32] One possible answer: if the friends know each other, then the independence assumption is probably not satisfied. For example, acquaintances may have similar smoking habits.

⊙ **Exercise 3.45** Suppose these four friends do not know each other and we can treat them as if they were a random sample from the population. Is the binomial model appropriate? What is the probability that (a) none of them will develop a severe lung condition? (b) One will develop a severe lung condition? (c) That no more than one will develop a severe lung condition?[33]

⊙ **Exercise 3.46** What is the probability that at least 2 of your 4 smoking friends will develop a severe lung condition in their lifetimes?[34]

⊙ **Exercise 3.47** Suppose you have 7 friends who are smokers and they can be treated as a random sample of smokers. (a) How many would you expect to develop a severe lung condition, i.e. what is the mean? (b) What is the probability that at most 2 of your 7 friends will develop a severe lung condition.[35]

Below we consider the first term in the binomial probability, n choose k under some special scenarios.

⊙ **Exercise 3.48** Why is it true that $\binom{n}{0} = 1$ and $\binom{n}{n} = 1$ for any number n?[36]

⊙ **Exercise 3.49** How many ways can you arrange one success and $n-1$ failures in n trials? How many ways can you arrange $n-1$ successes and one failure in n trials?[37]

[33] To check if the binomial model is appropriate, we must verify the conditions. (i) Since we are supposing we can treat the friends as a random sample, they are independent. (ii) We have a fixed number of trials ($n = 4$). (iii) Each outcome is a success or failure. (iv) The probability of a success is the same for each trials since the individuals are like a random sample ($p = 0.3$ if we say a "success" is someone getting a lung condition, a morbid choice). Compute parts (a) and (b) from the binomial formula in Equation (3.40): $P(0) = \binom{4}{0}(0.3)^0(0.7)^4 = 1 \times 1 \times 0.7^4 = 0.2401$, $P(1) = \binom{4}{1}(0.3)^1(0.7)^3 = 0.4116$. Note: $0! = 1$, as shown on page 138. Part (c) can be computed as the sum of parts (a) and (b): $P(0)+P(1) = 0.2401+0.4116 = 0.6517$. That is, there is about a 65% chance that no more than one of your four smoking friends will develop a severe lung condition.

[34] The complement (no more than one will develop a severe lung condition) as computed in Exercise 3.45 as 0.6517, so we compute one minus this value: 0.3483.

[35] (a) $\mu = 0.3 \times 7 = 2.1$. (b) $P(0, 1, \text{or } 2 \text{ develop severe lung condition}) = P(k = 0) + P(k = 1) + P(k = 2) = 0.6471$.

[36] Frame these expressions into words. How many different ways are there to arrange 0 successes and n failures in n trials? (1 way.) How many different ways are there to arrange n successes and 0 failures in n trials? (1 way.)

[37] One success and $n-1$ failures: there are exactly n unique places we can put the success, so there are n ways to arrange one success and $n-1$ failures. A similar argument is used for the second question. Mathematically, we show these results by verifying the following two equations:

$$\binom{n}{1} = n, \qquad \binom{n}{n-1} = n$$

3.4.2 Normal approximation to the binomial distribution

The binomial formula is cumbersome when the sample size (n) is large, particularly when we consider a range of observations. In some cases we may use the normal distribution as an easier and faster way to estimate binomial probabilities.

Example 3.50 Approximately 20% of the US population smokes cigarettes. A local government believed their community had a lower smoker rate and commissioned a survey of 400 randomly selected individuals. The survey found that only 59 of the 400 participants smoke cigarettes. If the true proportion of smokers in the community was really 20%, what is the probability of observing 59 or fewer smokers in a sample of 400 people?

We leave the usual verification that the four conditions for the binomial model are valid as an exercise.

The question posed is equivalent to asking, what is the probability of observing $k = 0$, 1, ..., 58, or 59 smokers in a sample of $n = 400$ when $p = 0.20$? We can compute these 60 different probabilities and add them together to find the answer:

$$\begin{aligned} &P(k = 0 \text{ or } k = 1 \text{ or } \cdots \text{ or } k = 59) \\ &\quad = P(k = 0) + P(k = 1) + \cdots + P(k = 59) \\ &\quad = 0.0041 \end{aligned}$$

If the true proportion of smokers in the community is $p = 0.20$, then the probability of observing 59 or fewer smokers in a sample of $n = 400$ is less than 0.0041.

The computations in Example 3.50 are tedious and long. In general, we should avoid such work if an alternative method exists that is faster, easier, and still accurate. Recall that calculating probabilities of a range of values is much easier in the normal model. We might wonder, is it reasonable to use the normal model in place of the binomial distribution? Surprisingly, yes, if certain conditions are met.

Exercise 3.51 Here we consider the binomial model when the probability of a success is $p = 0.10$. Figure 3.18 shows four hollow histograms for simulated samples from the binomial distribution using four different sample sizes: $n = 10, 30, 100, 300$. What happens to the shape of the distributions as the sample size increases? What distribution does the last hollow histogram resemble?[38]

Normal approximation of the binomial distribution

The binomial distribution with probability of success p is nearly normal when the sample size n is sufficiently large that np and $n(1 - p)$ are both at least 10. The approximate normal distribution has parameters corresponding to the mean and standard deviation of the binomial distribution:

$$\mu = np \qquad \sigma = \sqrt{np(1-p)}$$

The normal approximation may be used when computing the range of many possible successes. For instance, we may apply the normal distribution to the setting of Example 3.50.

[38] The distribution is transformed from a blocky and skewed distribution into one that rather resembles the normal distribution in last hollow histogram

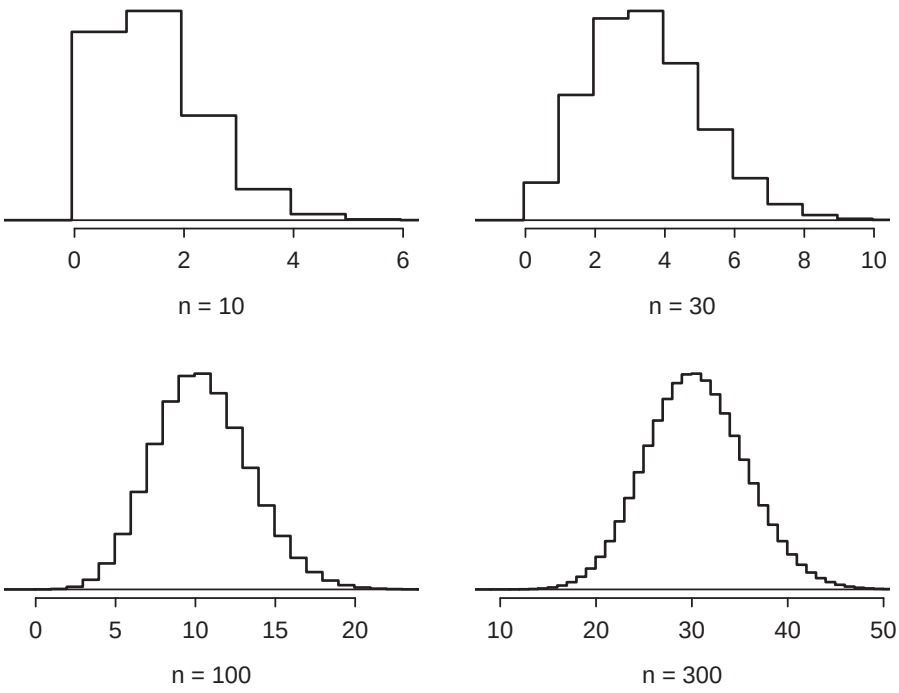

Figure 3.18: Hollow histograms of samples from the binomial model when $p = 0.10$. The sample sizes for the four plots are $n = 10$, 30, 100, and 300, respectively.

Example 3.52 How can we use the normal approximation to estimate the probability of observing 59 or fewer smokers in a sample of 400, if the true proportion of smokers is $p = 0.20$?

Showing that the binomial model is reasonable was a suggested exercise in Example 3.50. We also verify that both np and $n(1-p)$ are at least 10:

$$np = 400 \times 0.20 = 80 \qquad n(1-p) = 400 \times 0.8 = 320$$

With these conditions checked, we may use the normal approximation in place of the binomial distribution using the mean and standard deviation from the binomial model:

$$\mu = np = 80 \qquad \sigma = \sqrt{np(1-p)} = 8$$

We want to find the probability of observing fewer than 59 smokers using this model.

Exercise 3.53 Use the normal model $N(\mu = 80, \sigma = 8)$ to estimate the probability of observing fewer than 59 smokers. Your answer should be approximately equal to the solution of Example 3.50: 0.0041.[39]

[39] Compute the Z score first: $Z = \frac{59-80}{8} = -2.63$. The corresponding left tail area is 0.0043.