4.3 Binomial distribution

The **binomial distribution** is used to describe the number of successes in a fixed number of trials. This is different from the geometric distribution, which described the number of trials we must wait before we observe a success.

4.3.1 The binomial distribution

Let's again imagine ourselves back at the insurance agency where 70% of individuals do not exceed their deductible.

EXAMPLE 4.28

Suppose the insurance agency is considering a random sample of four individuals they insure. What is the chance exactly one of them will exceed the deductible and the other three will not? Let's call the four people Ariana (A), Brittany (B), Carlton (C), and Damian (D) for convenience.

Let's consider a scenario where one person exceeds the deductible:

(E)

$$\begin{aligned} &P(A = \texttt{exceed},\ B = \texttt{not},\ C = \texttt{not},\ D = \texttt{not}) \\ &\quad = P(A = \texttt{exceed})\ P(B = \texttt{not})\ P(C = \texttt{not})\ P(D = \texttt{not}) \\ &\quad = (0.3)(0.7)(0.7)(0.7) \\ &\quad = (0.7)^3(0.3)^1 \\ &\quad = 0.103 \end{aligned}$$

But there are three other scenarios: Brittany, Carlton, or Damian could have been the one to exceed the deductible. In each of these cases, the probability is again $(0.7)^3(0.3)^1$. These four scenarios exhaust all the possible ways that exactly one of these four people could have exceeded the deductible, so the total probability is $4 \times (0.7)^3(0.3)^1 = 0.412$.

GUIDED PRACTICE 4.29

(G) Verify that the scenario where Brittany is the only one exceed the deductible has probability $(0.7)^3(0.3)^1$. [20]

The scenario outlined in Example 4.28 is an example of a binomial distribution scenario. The **binomial distribution** describes the probability of having exactly k successes in n independent Bernoulli trials with probability of a success p (in Example 4.28, $n = 4$, $k = 3$, $p = 0.7$). 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 Example 4.28.

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

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

The first component of this equation is the number of ways to arrange the $k = 3$ successes among the $n = 4$ trials. The second component is the probability of any of the four (equally probable) scenarios.

[20] $P(A = \texttt{not},\ B = \texttt{exceed},\ C = \texttt{not},\ D = \texttt{not}) = (0.7)(0.3)(0.7)(0.7) = (0.7)^3(0.3)^1$.

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 $P(\text{single scenario})$.

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**.[21] 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 $k = 3$ successes in $n = 4$ trials:

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

This result is exactly what we found by carefully thinking of each possible scenario in Example 4.28.

Substituting n choose k for the number of scenarios and $p^k(1-p)^{n-k}$ for the single scenario probability 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

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

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

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

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.

---

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

EXAMPLE 4.30

What is the probability that 3 of 8 randomly selected individuals will have exceeded the insurance deductible, i.e. that 5 of 8 will not exceed the deductible? Recall that 70% of individuals will not exceed the deductible.

We would like to apply the binomial model, so we check the 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 $k = 5$ successes in $n = 8$ trials (recall that a success is an individual who does *not* exceed the deductible, and the probability of a success is $p = 0.7$. So the probability that 5 of 8 will not exceed the deductible and 3 will exceed the deductible is given by

$$\binom{8}{5}(0.7)^5(1-0.7)^{8-5} = \frac{8!}{5!(8-5)!}(0.7)^5(1-0.7)^{8-5}$$
$$= \frac{8!}{5!3!}(0.7)^5(0.3)^3$$

Dealing with the factorial part:

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

Using $(0.7)^5(0.3)^3 \approx 0.00454$, the final probability is about $56 \times 0.00454 \approx 0.254$.

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. As the last stage use software or the formulas to determine the probability, then interpret the results.

If you must do calculations by hand, it's often useful to cancel out as many terms as possible in the top and bottom of the binomial coefficient.

GUIDED PRACTICE 4.31

If we randomly sampled 40 case files from the insurance agency discussed earlier, how many of the cases would you expect to not have exceeded the deductible in a given year? What is the standard deviation of the number that would not have exceeded the deductible?[22]

GUIDED PRACTICE 4.32

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?[23]

[22] We are asked to determine the expected number (the mean) and the standard deviation, both of which can be directly computed from the formulas: $\mu = np = 40 \times 0.7 = 28$ and $\sigma = \sqrt{np(1-p)} = \sqrt{40 \times 0.7 \times 0.3} = 2.9$. Because very roughly 95% of observations fall within 2 standard deviations of the mean (see Section 2.1.4), we would probably observe at least 22 but fewer than 34 individuals in our sample who would not exceed the deductible.

[23] 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, or those friends might make a pact to quit together.

GUIDED PRACTICE 4.33

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[24]

(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?

GUIDED PRACTICE 4.34

What is the probability that at least 2 of your 4 smoking friends will develop a severe lung condition in their lifetimes?[25]

GUIDED PRACTICE 4.35

Suppose you have 7 friends who are smokers and they can be treated as a random sample of smokers.[26]

(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.

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

GUIDED PRACTICE 4.36

Why is it true that $\binom{n}{0} = 1$ and $\binom{n}{n} = 1$ for any number n?[27]

GUIDED PRACTICE 4.37

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?[28]

[24] 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) using the binomial formula: $P(0) = \binom{4}{0}(0.3)^0(0.7)^4 = 1 \times 1 \times 0.7^4 = 0.2401$, $P(1) = \binom{4}{1}(0.3)^1(0.7)^3 = 0.4116$. Note: $0! = 1$. 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.

[25] The complement (no more than one will develop a severe lung condition) as computed in Guided Practice 4.33 as 0.6517, so we compute one minus this value: 0.3483.

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

[27] 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.)

[28] 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:

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

4.3.2 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 4.38

Approximately 15% 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 42 of the 400 participants smoke cigarettes. If the true proportion of smokers in the community was really 15%, what is the probability of observing 42 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.

(E) The question posed is equivalent to asking, what is the probability of observing $k = 0$, 1, 2, ..., or 42 smokers in a sample of $n = 400$ when $p = 0.15$? We can compute these 43 different probabilities and add them together to find the answer:

$$\begin{aligned} &P(k = 0 \text{ or } k = 1 \text{ or } \cdots \text{ or } k = 42) \\ &\quad = P(k = 0) + P(k = 1) + \cdots + P(k = 42) \\ &\quad = 0.0054 \end{aligned}$$

If the true proportion of smokers in the community is $p = 0.15$, then the probability of observing 42 or fewer smokers in a sample of $n = 400$ is 0.0054.

The computations in Example 4.38 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.

GUIDED PRACTICE 4.39

(G) Here we consider the binomial model when the probability of a success is $p = 0.10$. Figure 4.9 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?[29]

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 4.38.

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

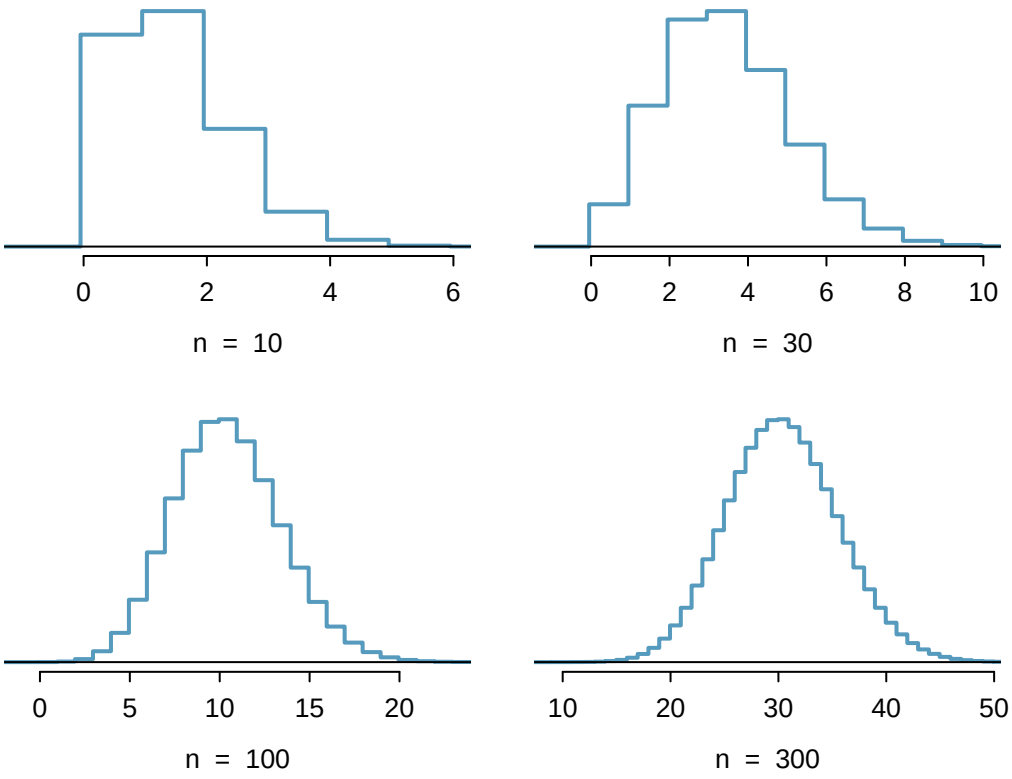

Figure 4.9: Hollow histograms of samples from the binomial model when $p = 0.10$. The sample sizes for the four plots are $n = 10$, 30, 100, and 300, respectively.

EXAMPLE 4.40

How can we use the normal approximation to estimate the probability of observing 42 or fewer smokers in a sample of 400, if the true proportion of smokers is $p = 0.15$?

Showing that the binomial model is reasonable was a suggested exercise in Example 4.38. We also verify that both np and $n(1-p)$ are at least 10:

(E)

$$np = 400 \times 0.15 = 60 \qquad n(1-p) = 400 \times 0.85 = 340$$

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:

$$\mu = np = 60 \qquad \sigma = \sqrt{np(1-p)} = 7.14$$

We want to find the probability of observing 42 or fewer smokers using this model.

GUIDED PRACTICE 4.41

(G) Use the normal model $N(\mu = 60, \sigma = 7.14)$ to estimate the probability of observing 42 or fewer smokers. Your answer should be approximately equal to the solution of Example 4.38: 0.0054. [30]

[30] Compute the Z-score first: $Z = \frac{42-60}{7.14} = -2.52$. The corresponding left tail area is 0.0059.

4.3.3 The normal approximation breaks down on small intervals

The normal approximation to the binomial distribution tends to perform poorly when estimating the probability of a small range of counts, even when the conditions are met.

Suppose we wanted to compute the probability of observing 49, 50, or 51 smokers in 400 when $p = 0.15$. With such a large sample, we might be tempted to apply the normal approximation and use the range 49 to 51. However, we would find that the binomial solution and the normal approximation notably differ:

Binomial: 0.0649 Normal: 0.0421

We can identify the cause of this discrepancy using Figure 4.10, which shows the areas representing the binomial probability (outlined) and normal approximation (shaded). Notice that the width of the area under the normal distribution is 0.5 units too slim on both sides of the interval.

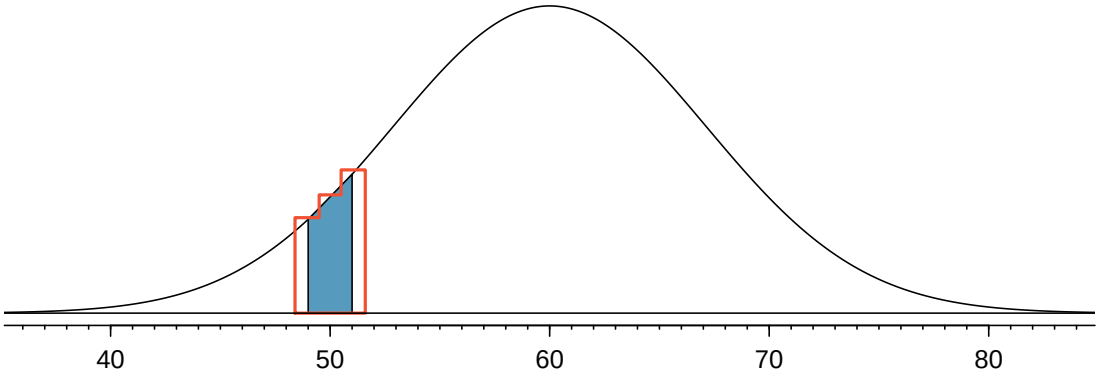

Figure 4.10: A normal curve with the area between 49 and 51 shaded. The outlined area represents the exact binomial probability.

IMPROVING THE NORMAL APPROXIMATION FOR THE BINOMIAL DISTRIBUTION

The normal approximation to the binomial distribution for intervals of values is usually improved if cutoff values are modified slightly. The cutoff values for the lower end of a shaded region should be reduced by 0.5, and the cutoff value for the upper end should be increased by 0.5.

The tip to add extra area when applying the normal approximation is most often useful when examining a range of observations. In the example above, the revised normal distribution estimate is 0.0633, much closer to the exact value of 0.0649. While it is possible to also apply this correction when computing a tail area, the benefit of the modification usually disappears since the total interval is typically quite wide.