Chapter 4

Distributions of random variables

In this chapter, we discuss statistical distributions that frequently arise in the context of data analysis or statistical inference. We start with the normal distribution in the first section, which is used frequently in later chapters of this book. The remaining sections will occasionally be referenced but may be considered optional for the content in this book.

For videos, slides, and other resources, please visit
www.openintro.org/os

4.1 Normal distribution

Among all the distributions we see in practice, one is overwhelmingly the most common. The symmetric, unimodal, bell curve is ubiquitous throughout statistics. Indeed it is so common, that people often know it as the **normal curve** or **normal distribution**,[1] shown in Figure 4.1. Variables such as SAT scores and heights of US adult males closely follow the normal distribution.

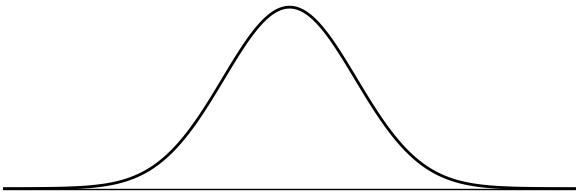


Figure 4.1: A normal curve.

NORMAL DISTRIBUTION FACTS

Many variables are nearly normal, but none are exactly normal. Thus the normal distribution, while not perfect for any single problem, is very useful for a variety of problems. We will use it in data exploration and to solve important problems in statistics.

4.1.1 Normal distribution model

The **normal distribution** always describes a symmetric, unimodal, bell-shaped curve. However, these curves can look different depending on the details of the model. Specifically, the normal distribution model can be adjusted using two parameters: mean and standard deviation. As you can probably guess, changing the mean shifts the bell curve to the left or right, while changing the standard deviation stretches or constricts the curve. Figure 4.2 shows the normal distribution with mean 0 and standard deviation 1 in the left panel and the normal distributions with mean 19 and standard deviation 4 in the right panel. Figure 4.3 shows these distributions on the same axis.

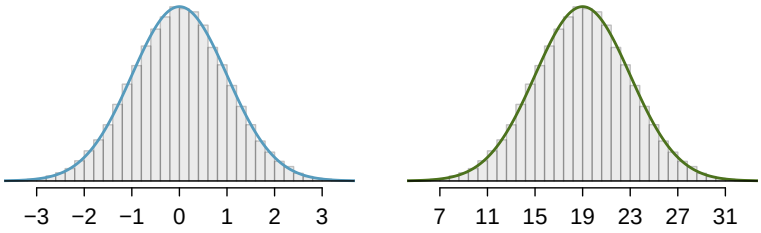


Figure 4.2: Both curves represent the normal distribution. However, they differ in their center and spread.

If a normal distribution has mean μ and standard deviation σ, we may write the distribution as $N(\mu, \sigma)$. The two distributions in Figure 4.3 may be written as

$$N(\mu = 0, \sigma = 1) \quad \text{and} \quad N(\mu = 19, \sigma = 4)$$

Because the mean and standard deviation describe a normal distribution exactly, they are called the distribution's **parameters**. The normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$ is called the **standard normal distribution**.

[1] It is also introduced as the Gaussian distribution after Frederic Gauss, the first person to formalize its mathematical expression.

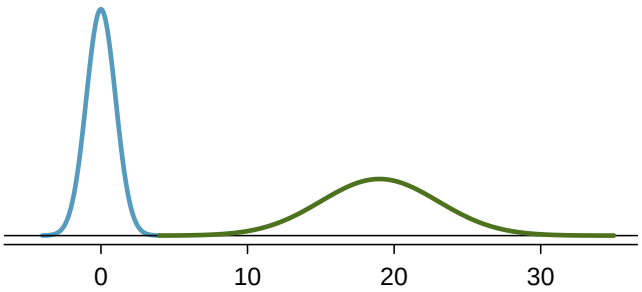


Figure 4.3: The normal distributions shown in Figure 4.2 but plotted together and on the same scale.

GUIDED PRACTICE 4.1

Write down the short-hand for a normal distribution with[2]
(a) mean 5 and standard deviation 3,
(b) mean -100 and standard deviation 10, and
(c) mean 2 and standard deviation 9.

4.1.2 Standardizing with Z-scores

We often want to put data onto a standardized scale, which can make comparisons more reasonable.

EXAMPLE 4.2

Table 4.4 shows the mean and standard deviation for total scores on the SAT and ACT. The distribution of SAT and ACT scores are both nearly normal. Suppose Ann scored 1300 on her SAT and Tom scored 24 on his ACT. Who performed better?

We use the standard deviation as a guide. Ann is 1 standard deviation above average on the SAT: $1100 + 200 = 1300$. Tom is 0.5 standard deviations above the mean on the ACT: $21 + 0.5 \times 6 = 24$. In Figure 4.5, we can see that Ann tends to do better with respect to everyone else than Tom did, so her score was better.

	SAT	ACT
Mean	1100	21
SD	200	6

Figure 4.4: Mean and standard deviation for the SAT and ACT.

Example 4.2 used a standardization technique called a Z-score, a method most commonly employed for nearly normal observations but that may be used with any distribution. The **Z-score** of an observation is defined as the number of standard deviations it falls above or below the mean. If the observation is one standard deviation above the mean, its Z-score is 1. If it is 1.5 standard deviations *below* the mean, then its Z-score is -1.5. If x is an observation from a distribution $N(\mu, \sigma)$, we define the Z-score mathematically as

$$Z = \frac{x - \mu}{\sigma}$$

Using $\mu_{SAT} = 1100$, $\sigma_{SAT} = 200$, and $x_{\text{Ann}} = 1300$, we find Ann's Z-score:

$$Z_{\text{Ann}} = \frac{x_{\text{Ann}} - \mu_{\text{SAT}}}{\sigma_{\text{SAT}}} = \frac{1300 - 1100}{200} = 1$$

[2] (a) $N(\mu = 5, \sigma = 3)$. (b) $N(\mu = -100, \sigma = 10)$. (c) $N(\mu = 2, \sigma = 9)$.

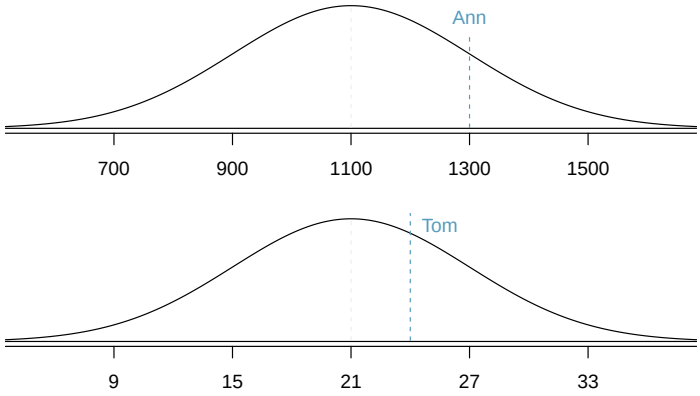


Figure 4.5: Ann's and Tom's scores shown against the SAT and ACT distributions.

THE Z-SCORE

The Z-score of an observation is the number of standard deviations it falls above or below the mean. We compute the Z-score for an observation x that follows a distribution with mean μ and standard deviation σ using

$$Z = \frac{x - \mu}{\sigma}$$

GUIDED PRACTICE 4.3

Use Tom's ACT score, 24, along with the ACT mean and standard deviation to find his Z-score.[3]

Observations above the mean always have positive Z-scores, while those below the mean always have negative Z-scores. If an observation is equal to the mean, such as an SAT score of 1100, then the Z-score is 0.

GUIDED PRACTICE 4.4

Let X represent a random variable from $N(\mu = 3, \sigma = 2)$, and suppose we observe $x = 5.19$.
(a) Find the Z-score of x.
(b) Use the Z-score to determine how many standard deviations above or below the mean x falls.[4]

GUIDED PRACTICE 4.5

Head lengths of brushtail possums follow a normal distribution with mean 92.6 mm and standard deviation 3.6 mm. Compute the Z-scores for possums with head lengths of 95.4 mm and 85.8 mm.[5]

We can use Z-scores to roughly identify which observations are more unusual than others. An observation x_1 is said to be more unusual than another observation x_2 if the absolute value of its Z-score is larger than the absolute value of the other observation's Z-score: $|Z_1| > |Z_2|$. This technique is especially insightful when a distribution is symmetric.

GUIDED PRACTICE 4.6

Which of the observations in Guided Practice 4.5 is more unusual?[6]

[3] $Z_{Tom} = \frac{x_{\text{Tom}} - \mu_{\text{ACT}}}{\sigma_{\text{ACT}}} = \frac{24-21}{6} = 0.5$

[4] (a) Its Z-score is given by $Z = \frac{x-\mu}{\sigma} = \frac{5.19-3}{2} = 2.19/2 = 1.095$. (b) The observation x is 1.095 standard deviations *above* the mean. We know it must be above the mean since Z is positive.

[5] For $x_1 = 95.4$ mm: $Z_1 = \frac{x_1-\mu}{\sigma} = \frac{95.4-92.6}{3.6} = 0.78$. For $x_2 = 85.8$ mm: $Z_2 = \frac{85.8-92.6}{3.6} = -1.89$.

[6] Because the *absolute value* of Z-score for the second observation is larger than that of the first, the second observation has a more unusual head length.

4.1.3 Finding tail areas

It's very useful in statistics to be able to identify tail areas of distributions. For instance, what fraction of people have an SAT score below Ann's score of 1300? This is the same as the **percentile** Ann is at, which is the percentage of cases that have lower scores than Ann. We can visualize such a tail area like the curve and shading shown in Figure 4.6.

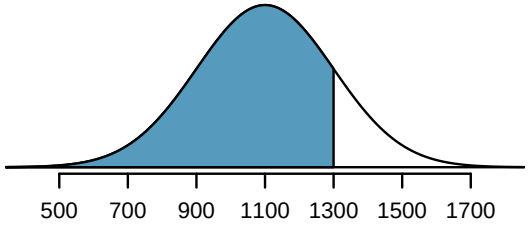


Figure 4.6: The area to the left of Z represents the fraction of people who scored lower than Ann.

There are many techniques for doing this, and we'll discuss three of the options.

1. The most common approach in practice is to use statistical software. For example, in the program **R**, we could find the area shown in Figure 4.6 using the following command, which takes in the Z-score and returns the lower tail area:

   ```
   > pnorm(1)
   [1] 0.8413447
   ```

   According to this calculation, the region shaded that is below 1300 represents the proportion 0.841 (84.1%) of SAT test takers who had Z-scores below $Z = 1$. More generally, we can also specify the cutoff explicitly if we also note the mean and standard deviation:

   ```
   > pnorm(1300, mean = 1100, sd = 200)
   [1] 0.8413447
   ```

   There are many other software options, such as Python or SAS; even spreadsheet programs such as Excel and Google Sheets support these calculations.

2. A common strategy in classrooms is to use a graphing calculator, such as a TI or Casio calculator. These calculators require a series of button presses that are less concisely described. You can find instructions on using these calculators for finding tail areas of a normal distribution in the OpenIntro video library:

   www.openintro.org/videos

3. The last option for finding tail areas is to use what's called a **probability table**; these are occasionally used in classrooms but rarely in practice. Appendix C.1 contains such a table and a guide for how to use it.

We will solve normal distribution problems in this section by always first finding the Z-score. The reason is that we will encounter close parallels called test statistics beginning in Chapter 5; these are, in many instances, an equivalent of a Z-score.

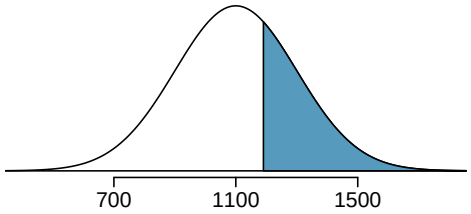
4.1.4 Normal probability examples

Cumulative SAT scores are approximated well by a normal model, $N(\mu = 1100, \sigma = 200)$.

EXAMPLE 4.7

Shannon is a randomly selected SAT taker, and nothing is known about Shannon's SAT aptitude. What is the probability Shannon scores at least 1190 on her SATs?

First, always draw and label a picture of the normal distribution. (Drawings need not be exact to be useful.) We are interested in the chance she scores above 1190, so we shade this upper tail:

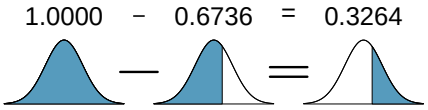


(E)

The picture shows the mean and the values at 2 standard deviations above and below the mean. The simplest way to find the shaded area under the curve makes use of the Z-score of the cutoff value. With $\mu = 1100$, $\sigma = 200$, and the cutoff value $x = 1190$, the Z-score is computed as

$$Z = \frac{x - \mu}{\sigma} = \frac{1190 - 1100}{200} = \frac{90}{200} = 0.45$$

Using statistical software (or another preferred method), we can find the area left of $Z = 0.45$ as 0.6736. To find the area *above* $Z = 0.45$, we compute one minus the area of the lower tail:



The probability Shannon scores at least 1190 on the SAT is 0.3264.

ALWAYS DRAW A PICTURE FIRST, AND FIND THE Z-SCORE SECOND

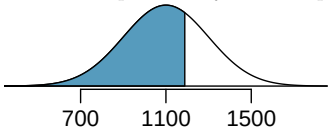
For any normal probability situation, *always always always* draw and label the normal curve and shade the area of interest first. The picture will provide an estimate of the probability. After drawing a figure to represent the situation, identify the Z-score for the value of interest.

GUIDED PRACTICE 4.8

(G)

If the probability of Shannon scoring at least 1190 is 0.3264, then what is the probability she scores less than 1190? Draw the normal curve representing this exercise, shading the lower region instead of the upper one.[7]

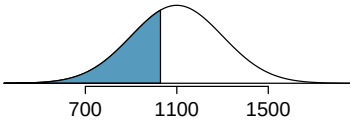
[7] We found this probability in Example 4.7: 0.6736.



EXAMPLE 4.9

Edward earned a 1030 on his SAT. What is his percentile?

First, a picture is needed. Edward's percentile is the proportion of people who do not get as high as a 1030. These are the scores to the left of 1030.



Identifying the mean $\mu = 1100$, the standard deviation $\sigma = 200$, and the cutoff for the tail area $x = 1030$ makes it easy to compute the Z-score:

$$Z = \frac{x - \mu}{\sigma} = \frac{1030 - 1100}{200} = -0.35$$

Using statistical software, we get a tail area of 0.3632. Edward is at the 36^{th} percentile.

GUIDED PRACTICE 4.10

Use the results of Example 4.9 to compute the proportion of SAT takers who did better than Edward. Also draw a new picture.[8]

FINDING AREAS TO THE RIGHT

Many software programs return the area to the left when given a Z-score. If you would like the area to the right, first find the area to the left and then subtract this amount from one.

GUIDED PRACTICE 4.11

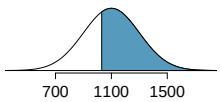
Stuart earned an SAT score of 1500. Draw a picture for each part.
(a) What is his percentile?
(b) What percent of SAT takers did better than Stuart?[9]

Based on a sample of 100 men, the heights of male adults in the US is nearly normal with mean 70.0" and standard deviation 3.3".

GUIDED PRACTICE 4.12

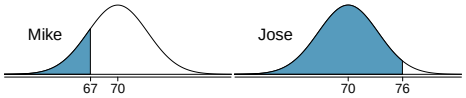
Mike is 5'7" and Jose is 6'4", and they both live in the US.
(a) What is Mike's height percentile?
(b) What is Jose's height percentile?
Also draw one picture for each part.[10]

[8] If Edward did better than 36% of SAT takers, then about 64% must have done better than him.



[9] We leave the drawings to you. (a) $Z = \frac{1500-1100}{200} = 2 \rightarrow 0.9772$. (b) $1 - 0.9772 = 0.0228$.

[10] First put the heights into inches: 67 and 76 inches. Figures are shown below.
(a) $Z_{\text{Mike}} = \frac{67-70}{3.3} = -0.91 \rightarrow 0.1814$. (b) $Z_{\text{Jose}} = \frac{76-70}{3.3} = 1.82 \rightarrow 0.9656$.

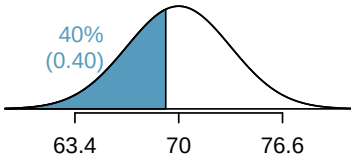


The last several problems have focused on finding the percentile (lower tail) or the upper tail for a particular observation. What if you would like to know the observation corresponding to a particular percentile?

EXAMPLE 4.13

Erik's height is at the 40^{th} percentile. How tall is he?

As always, first draw the picture.



(E) In this case, the lower tail probability is known (0.40), which can be shaded on the diagram. We want to find the observation that corresponds to this value. As a first step in this direction, we determine the Z-score associated with the 40^{th} percentile. Using software, we can obtain the corresponding Z-score of about -0.25.

Knowing $Z_{Erik} = -0.25$ and the population parameters $\mu = 70$ and $\sigma = 3.3$ inches, the Z-score formula can be set up to determine Erik's unknown height, labeled x_{Erik}:

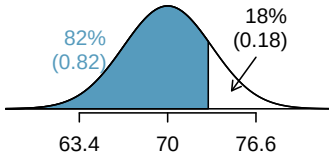
$$-0.25 = Z_{\text{Erik}} = \frac{x_{\text{Erik}} - \mu}{\sigma} = \frac{x_{\text{Erik}} - 70}{3.3}$$

Solving for x_{Erik} yields a height of 69.18 inches. That is, Erik is about 5'9".

EXAMPLE 4.14

What is the adult male height at the 82^{nd} percentile?

Again, we draw the figure first.



(E)

Next, we want to find the Z-score at the 82^{nd} percentile, which will be a positive value and can be found using software as $Z = 0.92$. Finally, the height x is found using the Z-score formula with the known mean μ, standard deviation σ, and Z-score $Z = 0.92$:

$$0.92 = Z = \frac{x - \mu}{\sigma} = \frac{x - 70}{3.3}$$

This yields 73.04 inches or about 6'1" as the height at the 82^{nd} percentile.

GUIDED PRACTICE 4.15

(G) The SAT scores follow $N(1100, 200)$.[11]
(a) What is the 95^{th} percentile for SAT scores?
(b) What is the 97.5^{th} percentile for SAT scores?

[11] Short answers: (a) $Z_{95} = 1.6449 \rightarrow 1429$ SAT score. (b) $Z_{97.5} = 1.96 \rightarrow 1492$ SAT score.

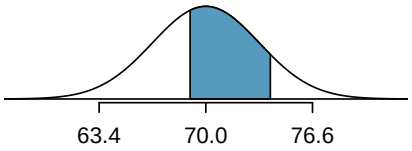
GUIDED PRACTICE 4.16

Adult male heights follow $N(70.0", 3.3")$.[12]
(a) What is the probability that a randomly selected male adult is at least 6'2" (74 inches)?
(b) What is the probability that a male adult is shorter than 5'9" (69 inches)?

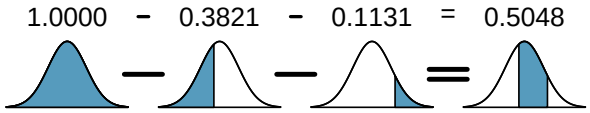
EXAMPLE 4.17

What is the probability that a random adult male is between 5'9" and 6'2"?

These heights correspond to 69 inches and 74 inches. First, draw the figure. The area of interest is no longer an upper or lower tail.



The total area under the curve is 1. If we find the area of the two tails that are not shaded (from Guided Practice 4.16, these areas are 0.3821 and 0.1131), then we can find the middle area:



That is, the probability of being between 5'9" and 6'2" is 0.5048.

GUIDED PRACTICE 4.18

SAT scores follow $N(1100, 200)$. What percent of SAT takers get between 1100 and 1400?[13]

GUIDED PRACTICE 4.19

Adult male heights follow $N(70.0", 3.3")$. What percent of adult males are between 5'5" and 5'7"?[14]

[12] Short answers: (a) $Z = 1.21 \rightarrow 0.8869$, then subtract this value from 1 to get 0.1131. (b) $Z = -0.30 \rightarrow 0.3821$.

[13] This is an abbreviated solution. (Be sure to draw a figure!) First find the percent who get below 1100 and the percent that get above 1400: $Z_{1100} = 0.00 \rightarrow 0.5000$ (area below), $Z_{1400} = 1.5 \rightarrow 0.0668$ (area above). Final answer: $1.0000 - 0.5000 - 0.0668 = 0.4332$.

[14] 5'5" is 65 inches ($Z = -1.52$). 5'7" is 67 inches ($Z = -0.91$). Numerical solution: $1.000 - 0.0643 - 0.8186 = 0.1171$, i.e. 11.71%.

4.1.5 68-95-99.7 rule

Here, we present a useful rule of thumb for the probability of falling within 1, 2, and 3 standard deviations of the mean in the normal distribution. This will be useful in a wide range of practical settings, especially when trying to make a quick estimate without a calculator or Z-table.

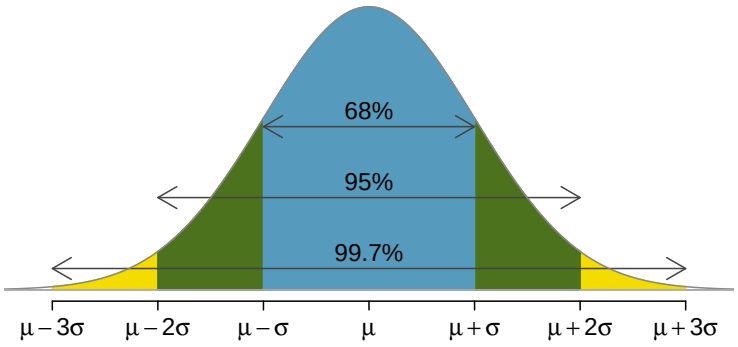


Figure 4.7: Probabilities for falling within 1, 2, and 3 standard deviations of the mean in a normal distribution.

GUIDED PRACTICE 4.20

Use software, a calculator, or a probability table to confirm that about 68%, 95%, and 99.7% of observations fall within 1, 2, and 3, standard deviations of the mean in the normal distribution, respectively. For instance, first find the area that falls between $Z = -1$ and $Z = 1$, which should have an area of about 0.68. Similarly there should be an area of about 0.95 between $Z = -2$ and $Z = 2$.[15]

It is possible for a normal random variable to fall 4, 5, or even more standard deviations from the mean. However, these occurrences are very rare if the data are nearly normal. The probability of being further than 4 standard deviations from the mean is about 1-in-15,000. For 5 and 6 standard deviations, it is about 1-in-2 million and 1-in-500 million, respectively.

GUIDED PRACTICE 4.21

SAT scores closely follow the normal model with mean $\mu = 1100$ and standard deviation $\sigma = 200$.[16]
(a) About what percent of test takers score 700 to 1500?
(b) What percent score between 1100 and 1500?

[15] First draw the pictures. Using software, we get 0.6827 within 1 standard deviation, 0.9545 within 2 standard deviations, and 0.9973 within 3 standard deviations.

[16] (a) 700 and 1500 represent two standard deviations below and above the mean, which means about 95% of test takers will score between 700 and 1500. (b) We found that 700 to 1500 represents about 95% of test takers. These test takers would be evenly split by the center of the distribution, 1100, so $\frac{95\%}{2} = 47.5\%$ of all test takers score between 1100 and 1500.

Exercises

4.1 Area under the curve, Part I. What percent of a standard normal distribution $N(\mu = 0, \sigma = 1)$ is found in each region? Be sure to draw a graph.

(a) $Z < -1.35$ (b) $Z > 1.48$ (c) $-0.4 < Z < 1.5$ (d) $|Z| > 2$

4.2 Area under the curve, Part II. What percent of a standard normal distribution $N(\mu = 0, \sigma = 1)$ is found in each region? Be sure to draw a graph.

(a) $Z > -1.13$ (b) $Z < 0.18$ (c) $Z > 8$ (d) $|Z| < 0.5$

4.3 GRE scores, Part I. Sophia who took the Graduate Record Examination (GRE) scored 160 on the Verbal Reasoning section and 157 on the Quantitative Reasoning section. The mean score for Verbal Reasoning section for all test takers was 151 with a standard deviation of 7, and the mean score for the Quantitative Reasoning was 153 with a standard deviation of 7.67. Suppose that both distributions are nearly normal.

(a) Write down the short-hand for these two normal distributions.
(b) What is Sophia's Z-score on the Verbal Reasoning section? On the Quantitative Reasoning section? Draw a standard normal distribution curve and mark these two Z-scores.
(c) What do these Z-scores tell you?
(d) Relative to others, which section did she do better on?
(e) Find her percentile scores for the two exams.
(f) What percent of the test takers did better than her on the Verbal Reasoning section? On the Quantitative Reasoning section?
(g) Explain why simply comparing raw scores from the two sections could lead to an incorrect conclusion as to which section a student did better on.
(h) If the distributions of the scores on these exams are not nearly normal, would your answers to parts (b) - (f) change? Explain your reasoning.

4.4 Triathlon times, Part I. In triathlons, it is common for racers to be placed into age and gender groups. Friends Leo and Mary both completed the Hermosa Beach Triathlon, where Leo competed in the *Men, Ages 30 - 34* group while Mary competed in the *Women, Ages 25 - 29* group. Leo completed the race in 1:22:28 (4948 seconds), while Mary completed the race in 1:31:53 (5513 seconds). Obviously Leo finished faster, but they are curious about how they did within their respective groups. Can you help them? Here is some information on the performance of their groups:

- The finishing times of the *Men, Ages 30 - 34* group has a mean of 4313 seconds with a standard deviation of 583 seconds.
- The finishing times of the *Women, Ages 25 - 29* group has a mean of 5261 seconds with a standard deviation of 807 seconds.
- The distributions of finishing times for both groups are approximately Normal.

Remember: a better performance corresponds to a faster finish.

(a) Write down the short-hand for these two normal distributions.
(b) What are the Z-scores for Leo's and Mary's finishing times? What do these Z-scores tell you?
(c) Did Leo or Mary rank better in their respective groups? Explain your reasoning.
(d) What percent of the triathletes did Leo finish faster than in his group?
(e) What percent of the triathletes did Mary finish faster than in her group?
(f) If the distributions of finishing times are not nearly normal, would your answers to parts (b) - (e) change? Explain your reasoning.

4.5 GRE scores, Part II. In Exercise 4.3 we saw two distributions for GRE scores: $N(\mu = 151, \sigma = 7)$ for the verbal part of the exam and $N(\mu = 153, \sigma = 7.67)$ for the quantitative part. Use this information to compute each of the following:

(a) The score of a student who scored in the 80^{th} percentile on the Quantitative Reasoning section.
(b) The score of a student who scored worse than 70% of the test takers in the Verbal Reasoning section.

4.6 Triathlon times, Part II. In Exercise 4.4 we saw two distributions for triathlon times: $N(\mu = 4313, \sigma = 583)$ for *Men, Ages 30 - 34* and $N(\mu = 5261, \sigma = 807)$ for the *Women, Ages 25 - 29* group. Times are listed in seconds. Use this information to compute each of the following:

(a) The cutoff time for the fastest 5% of athletes in the men's group, i.e. those who took the shortest 5% of time to finish.

(b) The cutoff time for the slowest 10% of athletes in the women's group.

4.7 LA weather, Part I. The average daily high temperature in June in LA is 77°F with a standard deviation of 5°F. Suppose that the temperatures in June closely follow a normal distribution.

(a) What is the probability of observing an 83°F temperature or higher in LA during a randomly chosen day in June?

(b) How cool are the coldest 10% of the days (days with lowest high temperature) during June in LA?

4.8 CAPM. The Capital Asset Pricing Model (CAPM) is a financial model that assumes returns on a portfolio are normally distributed. Suppose a portfolio has an average annual return of 14.7% (i.e. an average gain of 14.7%) with a standard deviation of 33%. A return of 0% means the value of the portfolio doesn't change, a negative return means that the portfolio loses money, and a positive return means that the portfolio gains money.

(a) What percent of years does this portfolio lose money, i.e. have a return less than 0%?

(b) What is the cutoff for the highest 15% of annual returns with this portfolio?

4.9 LA weather, Part II. Exercise 4.7 states that average daily high temperature in June in LA is 77°F with a standard deviation of 5°F, and it can be assumed that they to follow a normal distribution. We use the following equation to convert °F (Fahrenheit) to °C (Celsius):

$$C = (F - 32) \times \frac{5}{9}.$$

(a) Write the probability model for the distribution of temperature in °C in June in LA.

(b) What is the probability of observing a 28°C (which roughly corresponds to 83°F) temperature or higher in June in LA? Calculate using the °C model from part (a).

(c) Did you get the same answer or different answers in part (b) of this question and part (a) of Exercise 4.7? Are you surprised? Explain.

(d) Estimate the IQR of the temperatures (in °C) in June in LA.

4.10 Find the SD. Cholesterol levels for women aged 20 to 34 follow an approximately normal distribution with mean 185 milligrams per deciliter (mg/dl). Women with cholesterol levels above 220 mg/dl are considered to have high cholesterol and about 18.5% of women fall into this category. What is the standard deviation of the distribution of cholesterol levels for women aged 20 to 34?

4.2 Geometric distribution

How long should we expect to flip a coin until it turns up `heads`? Or how many times should we expect to roll a die until we get a `1`? These questions can be answered using the geometric distribution. We first formalize each trial – such as a single coin flip or die toss – using the Bernoulli distribution, and then we combine these with our tools from probability (Chapter 3) to construct the geometric distribution.

4.2.1 Bernoulli distribution

Many health insurance plans in the United States have a deductible, where the insured individual is responsible for costs up to the deductible, and then the costs above the deductible are shared between the individual and insurance company for the remainder of the year.

Suppose a health insurance company found that 70% of the people they insure stay below their deductible in any given year. Each of these people can be thought of as a **trial**. We label a person a **success** if her healthcare costs do not exceed the deductible. We label a person a **failure** if she does exceed her deductible in the year. Because 70% of the individuals will not hit their deductible, we denote the **probability of a success** as $p = 0.7$. The probability of a failure is sometimes denoted with $q = 1 - p$, which would be 0.3 for the insurance example.

When an individual trial only has two possible outcomes, often labeled as `success` or `failure`, it is called a **Bernoulli random variable**. We chose to label a person who does not hit her deductible as a "success" and all others as "failures". However, we could just as easily have reversed these labels. The mathematical framework we will build does not depend on which outcome is labeled a success and which a failure, as long as we are consistent.

Bernoulli random variables are often denoted as `1` for a success and `0` for a failure. In addition to being convenient in entering data, it is also mathematically handy. Suppose we observe ten trials:

`1 1 1 0 1 0 0 1 1 0`

Then the **sample proportion**, $\hat{p}$, is the sample mean of these observations:

$$\hat{p} = \frac{\text{\# of successes}}{\text{\# of trials}} = \frac{1+1+1+0+1+0+0+1+1+0}{10} = 0.6$$

This mathematical inquiry of Bernoulli random variables can be extended even further. Because `0` and `1` are numerical outcomes, we can define the mean and standard deviation of a Bernoulli random variable. (See Exercises 4.15 and 4.16.)

BERNOULLI RANDOM VARIABLE

If X is a random variable that takes value 1 with probability of success p and 0 with probability $1 - p$, then X is a Bernoulli random variable with mean and standard deviation

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

In general, it is useful to think about a Bernoulli random variable as a random process with only two outcomes: a success or failure. Then we build our mathematical framework using the numerical labels `1` and `0` for successes and failures, respectively.

4.2.2 Geometric distribution

The **geometric distribution** is used to describe how many trials it takes to observe a success. Let's first look at an example.

EXAMPLE 4.22

Suppose we are working at the insurance company and need to find a case where the person did not exceed her (or his) deductible as a case study. If the probability a person will not exceed her deductible is 0.7 and we are drawing people at random, what are the chances that the first person will not have exceeded her deductible, i.e. be a success? The second person? The third? What about we pull $n-1$ cases before we find the first success, i.e. the first success is the n^{th} person? (If the first success is the fifth person, then we say $n = 5$.)

(E) The probability of stopping after the first person is just the chance the first person will not hit her (or his) deductible: 0.7. The probability the second person is the first to hit her deductible is

$$\begin{aligned} &P(\text{second person is the first to not hit deductible}) \\ &\quad = P(\text{the first will, the second won't}) = (0.3)(0.7) = 0.21 \end{aligned}$$

Likewise, the probability it will be the third case is $(0.3)(0.3)(0.7) = 0.063$.

If the first success is on the n^{th} person, then there are $n-1$ failures and finally 1 success, which corresponds to the probability $(0.3)^{n-1}(0.7)$. This is the same as $(1-0.7)^{n-1}(0.7)$.

Example 4.22 illustrates what the **geometric distribution**, which describes the waiting time until a success for **independent and identically distributed (iid)** Bernoulli random variables. In this case, the *independence* aspect just means the individuals in the example don't affect each other, and *identical* means they each have the same probability of success.

The geometric distribution from Example 4.22 is shown in Figure 4.8. In general, the probabilities for a geometric distribution decrease **exponentially** fast.

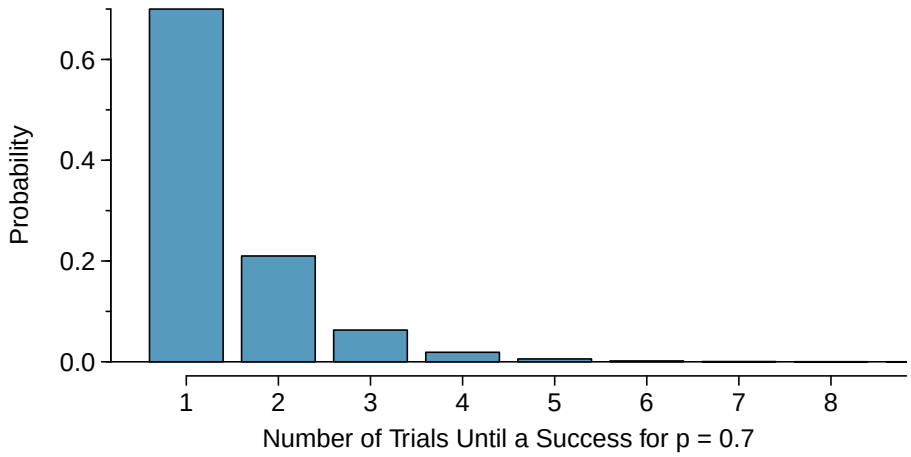


Figure 4.8: The geometric distribution when the probability of success is $p = 0.7$.

While this text will not derive the formulas for the mean (expected) number of trials needed to find the first success or the standard deviation or variance of this distribution, we present general formulas for each.

GEOMETRIC DISTRIBUTION

If the probability of a success in one trial is p and the probability of a failure is $1-p$, then the probability of finding the first success in the n^{th} trial is given by

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

The mean (i.e. expected value), variance, and standard deviation of this wait time are given by

$$\mu = \frac{1}{p} \qquad \sigma^2 = \frac{1-p}{p^2} \qquad \sigma = \sqrt{\frac{1-p}{p^2}}$$

It is no accident that we use the symbol μ for both the mean and expected value. The mean and the expected value are one and the same.

It takes, on average, $1/p$ trials to get a success under the geometric distribution. This mathematical result is consistent with what we would expect intuitively. If the probability of a success is high (e.g. 0.8), then we don't usually wait very long for a success: $1/0.8 = 1.25$ trials on average. If the probability of a success is low (e.g. 0.1), then we would expect to view many trials before we see a success: $1/0.1 = 10$ trials.

GUIDED PRACTICE 4.23

G The probability that a particular case would not exceed their deductible is said to be 0.7. If we were to examine cases until we found one that where the person did not hit her deductible, how many cases should we expect to check?[17]

EXAMPLE 4.24

What is the chance that we would find the first success within the first 3 cases?

This is the chance it is the first ($n = 1$), second ($n = 2$), or third ($n = 3$) case is the first success, which are three disjoint outcomes. Because the individuals in the sample are randomly sampled from a large population, they are independent. We compute the probability of each case and add the separate results:

E

$$\begin{aligned} &P(n = 1, 2, \text{ or } 3) \\ &\quad = P(n = 1) + P(n = 2) + P(n = 3) \\ &\quad = (0.3)^{1-1}(0.7) + (0.3)^{2-1}(0.7) + (0.3)^{3-1}(0.7) \\ &\quad = 0.973 \end{aligned}$$

There is a probability of 0.973 that we would find a successful case within 3 cases.

GUIDED PRACTICE 4.25

G Determine a more clever way to solve Example 4.24. Show that you get the same result.[18]

[17] We would expect to see about $1/0.7 \approx 1.43$ individuals to find the first success.

[18] First find the probability of the complement: $P(\text{no success in first 3 trials}) = 0.3^3 = 0.027$. Next, compute one minus this probability: $1 - P(\text{no success in 3 trials}) = 1 - 0.027 = 0.973$.

EXAMPLE 4.26

Suppose a car insurer has determined that 88% of its drivers will not exceed their deductible in a given year. If someone at the company were to randomly draw driver files until they found one that had not exceeded their deductible, what is the expected number of drivers the insurance employee must check? What is the standard deviation of the number of driver files that must be drawn?

(E)

In this example, a success is again when someone will not exceed the insurance deductible, which has probability $p = 0.88$. The expected number of people to be checked is $1/p = 1/0.88 = 1.14$ and the standard deviation is $\sqrt{(1-p)/p^2} = 0.39$.

GUIDED PRACTICE 4.27

(G) Using the results from Example 4.26, $\mu = 1.14$ and $\sigma = 0.39$, would it be appropriate to use the normal model to find what proportion of experiments would end in 3 or fewer trials?[19]

The independence assumption is crucial to the geometric distribution's accurate description of a scenario. Mathematically, we can see that to construct the probability of the success on the n^{th} trial, we had to use the Multiplication Rule for Independent Processes. It is no simple task to generalize the geometric model for dependent trials.

[19] No. The geometric distribution is always right skewed and can never be well-approximated by the normal model.

Exercises

4.11 Is it Bernoulli? Determine if each trial can be considered an independent Bernoulli trial for the following situations.

(a) Cards dealt in a hand of poker.

(b) Outcome of each roll of a die.

4.12 With and without replacement. In the following situations assume that half of the specified population is male and the other half is female.

(a) Suppose you're sampling from a room with 10 people. What is the probability of sampling two females in a row when sampling with replacement? What is the probability when sampling without replacement?

(b) Now suppose you're sampling from a stadium with 10,000 people. What is the probability of sampling two females in a row when sampling with replacement? What is the probability when sampling without replacement?

(c) We often treat individuals who are sampled from a large population as independent. Using your findings from parts (a) and (b), explain whether or not this assumption is reasonable.

4.13 Eye color, Part I. A husband and wife both have brown eyes but carry genes that make it possible for their children to have brown eyes (probability 0.75), blue eyes (0.125), or green eyes (0.125).

(a) What is the probability the first blue-eyed child they have is their third child? Assume that the eye colors of the children are independent of each other.

(b) On average, how many children would such a pair of parents have before having a blue-eyed child? What is the standard deviation of the number of children they would expect to have until the first blue-eyed child?

4.14 Defective rate. A machine that produces a special type of transistor (a component of computers) has a 2% defective rate. The production is considered a random process where each transistor is independent of the others.

(a) What is the probability that the 10^{th} transistor produced is the first with a defect?

(b) What is the probability that the machine produces no defective transistors in a batch of 100?

(c) On average, how many transistors would you expect to be produced before the first with a defect? What is the standard deviation?

(d) Another machine that also produces transistors has a 5% defective rate where each transistor is produced independent of the others. On average how many transistors would you expect to be produced with this machine before the first with a defect? What is the standard deviation?

(e) Based on your answers to parts (c) and (d), how does increasing the probability of an event affect the mean and standard deviation of the wait time until success?

4.15 Bernoulli, the mean. Use the probability rules from Section 3.4 to derive the mean of a Bernoulli random variable, i.e. a random variable X that takes value 1 with probability p and value 0 with probability $1 - p$. That is, compute the expected value of a generic Bernoulli random variable.

4.16 Bernoulli, the standard deviation. Use the probability rules from Section 3.4 to derive the standard deviation of a Bernoulli random variable, i.e. a random variable X that takes value 1 with probability p and value 0 with probability $1 - p$. That is, compute the square root of the variance of a generic Bernoulli random variable.

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 to 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

(E)

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

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

(G) 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

(G) 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.

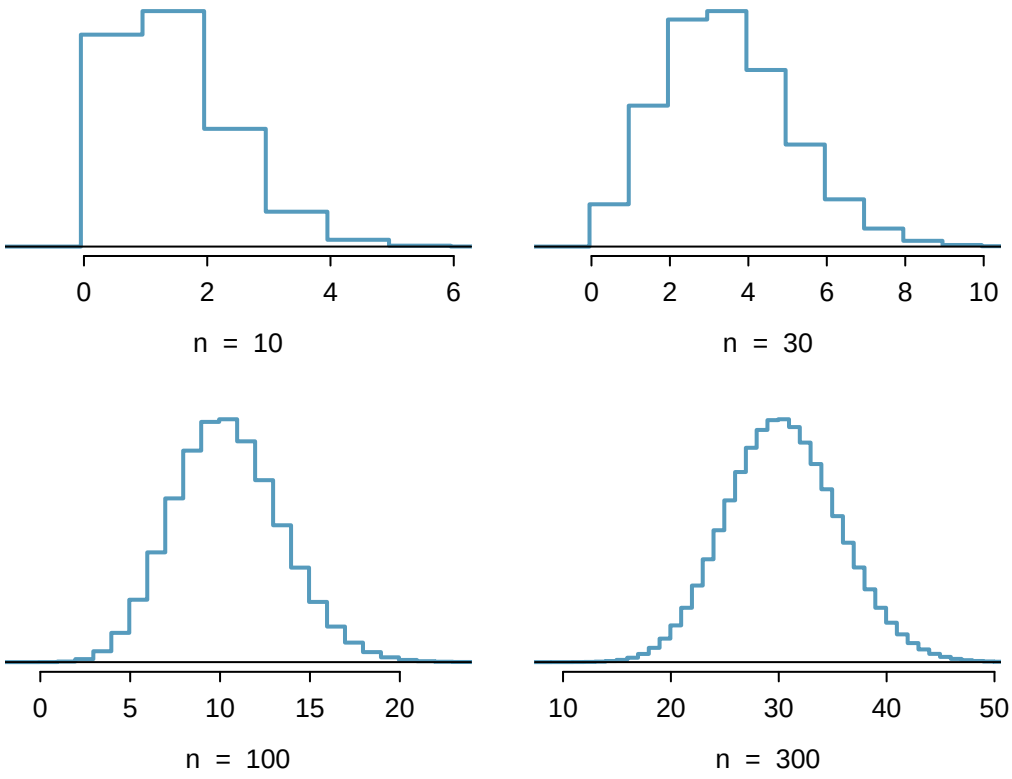


Figure 4.9: Hollow histograms of samples from the binomial model when $p = 0.10$. The sample sizes for the four plots are $n = 10$, 30, 100, and 300, respectively.

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.

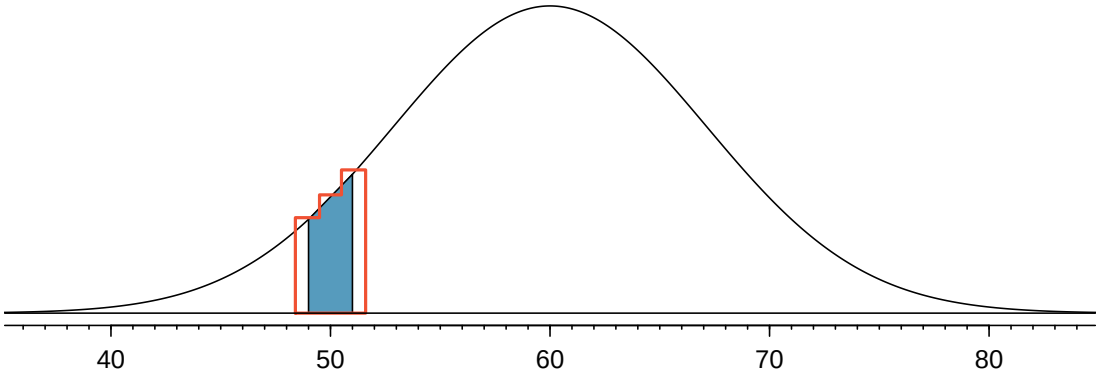


Figure 4.10: A normal curve with the area between 49 and 51 shaded. The outlined area represents the exact binomial probability.

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.

Exercises

4.17 Underage drinking, Part I. Data collected by the Substance Abuse and Mental Health Services Administration (SAMSHA) suggests that 69.7% of 18-20 year olds consumed alcoholic beverages in any given year.[31]

(a) Suppose a random sample of ten 18-20 year olds is taken. Is the use of the binomial distribution appropriate for calculating the probability that exactly six consumed alcoholic beverages? Explain.

(b) Calculate the probability that exactly 6 out of 10 randomly sampled 18- 20 year olds consumed an alcoholic drink.

(c) What is the probability that exactly four out of ten 18-20 year olds have *not* consumed an alcoholic beverage?

(d) What is the probability that at most 2 out of 5 randomly sampled 18-20 year olds have consumed alcoholic beverages?

(e) What is the probability that at least 1 out of 5 randomly sampled 18-20 year olds have consumed alcoholic beverages?

4.18 Chickenpox, Part I. Boston Children's Hospital estimates that 90% of Americans have had chickenpox by the time they reach adulthood.[32]

(a) Suppose we take a random sample of 100 American adults. Is the use of the binomial distribution appropriate for calculating the probability that exactly 97 out of 100 randomly sampled American adults had chickenpox during childhood? Explain.

(b) Calculate the probability that exactly 97 out of 100 randomly sampled American adults had chickenpox during childhood.

(c) What is the probability that exactly 3 out of a new sample of 100 American adults have *not* had chickenpox in their childhood?

(d) What is the probability that at least 1 out of 10 randomly sampled American adults have had chickenpox?

(e) What is the probability that at most 3 out of 10 randomly sampled American adults have *not* had chickenpox?

4.19 Underage drinking, Part II. We learned in Exercise 4.17 that about 70% of 18-20 year olds consumed alcoholic beverages in any given year. We now consider a random sample of fifty 18-20 year olds.

(a) How many people would you expect to have consumed alcoholic beverages? And with what standard deviation?

(b) Would you be surprised if there were 45 or more people who have consumed alcoholic beverages?

(c) What is the probability that 45 or more people in this sample have consumed alcoholic beverages? How does this probability relate to your answer to part (b)?

4.20 Chickenpox, Part II. We learned in Exercise 4.18 that about 90% of American adults had chickenpox before adulthood. We now consider a random sample of 120 American adults.

(a) How many people in this sample would you expect to have had chickenpox in their childhood? And with what standard deviation?

(b) Would you be surprised if there were 105 people who have had chickenpox in their childhood?

(c) What is the probability that 105 or fewer people in this sample have had chickenpox in their childhood? How does this probability relate to your answer to part (b)?

4.21 Game of dreidel. A dreidel is a four-sided spinning top with the Hebrew letters *nun*, *gimel*, *hei*, and *shin*, one on each side. Each side is equally likely to come up in a single spin of the dreidel. Suppose you spin a dreidel three times. Calculate the probability of getting

(a) at least one *nun*?

(b) exactly 2 *nuns*?

(c) exactly 1 *hei*?

(d) at most 2 *gimels*?

Photo by Staccabees, cropped (http://flic.kr/p/7gLZTf) CC BY 2.0 license

[31] SAMHSA, Office of Applied Studies, National Survey on Drug Use and Health, 2007 and 2008.
[32] Boston Children's Hospital, Chickenpox summary page, referenced April 29, 2021.

4.22 Arachnophobia. A Gallup Poll found that 7% of teenagers (ages 13 to 17) suffer from arachnophobia and are extremely afraid of spiders. At a summer camp there are 10 teenagers sleeping in each tent. Assume that these 10 teenagers are independent of each other.[33]

(a) Calculate the probability that at least one of them suffers from arachnophobia.

(b) Calculate the probability that exactly 2 of them suffer from arachnophobia.

(c) Calculate the probability that at most 1 of them suffers from arachnophobia.

(d) If the camp counselor wants to make sure no more than 1 teenager in each tent is afraid of spiders, does it seem reasonable for him to randomly assign teenagers to tents?

4.23 Eye color, Part II. Exercise 4.13 introduces a husband and wife with brown eyes who have 0.75 probability of having children with brown eyes, 0.125 probability of having children with blue eyes, and 0.125 probability of having children with green eyes.

(a) What is the probability that their first child will have green eyes and the second will not?

(b) What is the probability that exactly one of their two children will have green eyes?

(c) If they have six children, what is the probability that exactly two will have green eyes?

(d) If they have six children, what is the probability that at least one will have green eyes?

(e) What is the probability that the first green eyed child will be the 4^{th} child?

(f) Would it be considered unusual if only 2 out of their 6 children had brown eyes?

4.24 Sickle cell anemia. Sickle cell anemia is a genetic blood disorder where red blood cells lose their flexibility and assume an abnormal, rigid, "sickle" shape, which results in a risk of various complications. If both parents are carriers of the disease, then a child has a 25% chance of having the disease, 50% chance of being a carrier, and 25% chance of neither having the disease nor being a carrier. If two parents who are carriers of the disease have 3 children, what is the probability that

(a) two will have the disease?

(b) none will have the disease?

(c) at least one will neither have the disease nor be a carrier?

(d) the first child with the disease will the be 3^{rd} child?

4.25 Exploring permutations. The formula for the number of ways to arrange n objects is $n! = n \times (n-1) \times \cdots \times 2 \times 1$. This exercise walks you through the derivation of this formula for a couple of special cases.

A small company has five employees: Anna, Ben, Carl, Damian, and Eddy. There are five parking spots in a row at the company, none of which are assigned, and each day the employees pull into a random parking spot. That is, all possible orderings of the cars in the row of spots are equally likely.

(a) On a given day, what is the probability that the employees park in alphabetical order?

(b) If the alphabetical order has an equal chance of occurring relative to all other possible orderings, how many ways must there be to arrange the five cars?

(c) Now consider a sample of 8 employees instead. How many possible ways are there to order these 8 employees' cars?

4.26 Male children. While it is often assumed that the probabilities of having a boy or a girl are the same, the actual probability of having a boy is slightly higher at 0.51. Suppose a couple plans to have 3 kids.

(a) Use the binomial model to calculate the probability that two of them will be boys.

(b) Write out all possible orderings of 3 children, 2 of whom are boys. Use these scenarios to calculate the same probability from part (a) but using the addition rule for disjoint outcomes. Confirm that your answers from parts (a) and (b) match.

(c) If we wanted to calculate the probability that a couple who plans to have 8 kids will have 3 boys, briefly describe why the approach from part (b) would be more tedious than the approach from part (a).

[33] Gallup Poll, What Frightens America's Youth?, March 29, 2005.

4.4 Negative binomial distribution

The geometric distribution describes the probability of observing the first success on the n^{th} trial. The **negative binomial distribution** is more general: it describes the probability of observing the k^{th} success on the n^{th} trial.

EXAMPLE 4.42

Each day a high school football coach tells his star kicker, Brian, that he can go home after he successfully kicks four 35 yard field goals. Suppose we say each kick has a probability p of being successful. If p is small – e.g. close to 0.1 – would we expect Brian to need many attempts before he successfully kicks his fourth field goal?

(E)

We are waiting for the fourth success ($k = 4$). If the probability of a success (p) is small, then the number of attempts (n) will probably be large. This means that Brian is more likely to need many attempts before he gets $k = 4$ successes. To put this another way, the probability of n being small is low.

To identify a negative binomial case, we check 4 conditions. The first three are common to the binomial distribution.

IS IT NEGATIVE BINOMIAL? FOUR CONDITIONS TO CHECK

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

GUIDED PRACTICE 4.43

(G) Suppose Brian is very diligent in his attempts and he makes each 35 yard field goal with probability $p = 0.8$. Take a guess at how many attempts he would need before making his fourth kick.[34]

EXAMPLE 4.44

In yesterday's practice, it took Brian only 6 tries to get his fourth field goal. Write out each of the possible sequence of kicks.

(E)

Because it took Brian six tries to get the fourth success, we know the last kick must have been a success. That leaves three successful kicks and two unsuccessful kicks (we label these as failures) that make up the first five attempts. There are ten possible sequences of these first five kicks, which are shown in Figure 4.11. If Brian achieved his fourth success ($k = 4$) on his sixth attempt ($n = 6$), then his order of successes and failures must be one of these ten possible sequences.

GUIDED PRACTICE 4.45

(G) Each sequence in Figure 4.11 has exactly two failures and four successes with the last attempt always being a success. If the probability of a success is $p = 0.8$, find the probability of the first sequence.[35]

[34] One possible answer: since he is likely to make each field goal attempt, it will take him at least 4 attempts but probably not more than 6 or 7.

[35] The first sequence: $0.2 \times 0.2 \times 0.8 \times 0.8 \times 0.8 \times 0.8 = 0.0164$.

	Kick Attempt					
	1	2	3	4	5	6
1	F	F	^{1}S	^{2}S	^{3}S	^{4}S
2	F	^{1}S	F	^{2}S	^{3}S	^{4}S
3	F	^{1}S	^{2}S	F	^{3}S	^{4}S
4	F	^{1}S	^{2}S	^{3}S	F	^{4}S
5	^{1}S	F	F	^{2}S	^{3}S	^{4}S
6	^{1}S	F	^{2}S	F	^{3}S	^{4}S
7	^{1}S	F	^{2}S	^{3}S	F	^{4}S
8	^{1}S	^{2}S	F	F	^{3}S	^{4}S
9	^{1}S	^{2}S	F	^{3}S	F	^{4}S
10	^{1}S	^{2}S	^{3}S	F	F	^{4}S

Figure 4.11: The ten possible sequences when the fourth successful kick is on the sixth attempt.

If the probability Brian kicks a 35 yard field goal is $p = 0.8$, what is the probability it takes Brian exactly six tries to get his fourth successful kick? We can write this as

$$\begin{aligned} &P(\text{it takes Brian six tries to make four field goals}) \\ &\quad = P(\text{Brian makes three of his first five field goals, and he makes the sixth one}) \\ &\quad = P(1^{st} \text{ sequence OR } 2^{nd} \text{ sequence OR ... OR } 10^{th} \text{ sequence}) \end{aligned}$$

where the sequences are from Figure 4.11. We can break down this last probability into the sum of ten disjoint possibilities:

$$\begin{aligned} &P(1^{st} \text{ sequence OR } 2^{nd} \text{ sequence OR ... OR } 10^{th} \text{ sequence}) \\ &\quad = P(1^{st} \text{ sequence}) + P(2^{nd} \text{ sequence}) + \cdots + P(10^{th} \text{ sequence}) \end{aligned}$$

The probability of the first sequence was identified in Guided Practice 4.45 as 0.0164, and each of the other sequences have the same probability. Since each of the ten sequence has the same probability, the total probability is ten times that of any individual sequence.

The way to compute this negative binomial probability is similar to how the binomial problems were solved in Section 4.3. The probability is broken into two pieces:

$$\begin{aligned} &P(\text{it takes Brian six tries to make four field goals}) \\ &= [\text{Number of possible sequences}] \times P(\text{Single sequence}) \end{aligned}$$

Each part is examined separately, then we multiply to get the final result.

We first identify the probability of a single sequence. One particular case is to first observe all the failures ($n - k$ of them) followed by the k successes:

$$\begin{aligned} &P(\text{Single sequence}) \\ &= P(n-k \text{ failures and then } k \text{ successes}) \\ &= (1-p)^{n-k}p^k \end{aligned}$$

We must also identify the number of sequences for the general case. Above, ten sequences were identified where the fourth success came on the sixth attempt. These sequences were identified by fixing the last observation as a success and looking for all the ways to arrange the other observations. In other words, how many ways could we arrange $k-1$ successes in $n-1$ trials? This can be found using the n choose k coefficient but for $n-1$ and $k-1$ instead:

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

This is the number of different ways we can order $k-1$ successes and $n-k$ failures in $n-1$ trials. If the factorial notation (the exclamation point) is unfamiliar, see page 150.

NEGATIVE BINOMIAL DISTRIBUTION

The negative binomial distribution describes the probability of observing the k^{th} success on the n^{th} trial, where all trials are independent:

$$P(\text{the } k^{th} \text{ success on the } n^{th} \text{ trial}) = \binom{n-1}{k-1} p^k (1-p)^{n-k}$$

The value p represents the probability that an individual trial is a success.

EXAMPLE 4.46

Show using the formula for the negative binomial distribution that the probability Brian kicks his fourth successful field goal on the sixth attempt is 0.164.

(E) The probability of a single success is $p = 0.8$, the number of successes is $k = 4$, and the number of necessary attempts under this scenario is $n = 6$.

$$\binom{n-1}{k-1} p^k (1-p)^{n-k} \;=\; \frac{5!}{3!2!}(0.8)^4(0.2)^2 \;=\; 10 \times 0.0164 \;=\; 0.164$$

GUIDED PRACTICE 4.47

(G) The negative binomial distribution requires that each kick attempt by Brian is independent. Do you think it is reasonable to suggest that each of Brian's kick attempts are independent?[36]

GUIDED PRACTICE 4.48

(G) Assume Brian's kick attempts are independent. What is the probability that Brian will kick his fourth field goal within 5 attempts?[37]

[36] Answers may vary. We cannot conclusively say they are or are not independent. However, many statistical reviews of athletic performance suggests such attempts are very nearly independent.

[37] If his fourth field goal ($k = 4$) is within five attempts, it either took him four or five tries ($n = 4$ or $n = 5$). We have $p = 0.8$ from earlier. Use the negative binomial distribution to compute the probability of $n = 4$ tries and $n = 5$ tries, then add those probabilities together:

$$P(n = 4 \text{ OR } n = 5) = P(n = 4) + P(n = 5)$$
$$= \binom{4-1}{4-1}0.8^4 + \binom{5-1}{4-1}(0.8)^4(1-0.8) = 1 \times 0.41 + 4 \times 0.082 = 0.41 + 0.33 = 0.74$$

BINOMIAL VERSUS NEGATIVE BINOMIAL

In the binomial case, we typically have a fixed number of trials and instead consider the number of successes. In the negative binomial case, we examine how many trials it takes to observe a fixed number of successes and require that the last observation be a success.

GUIDED PRACTICE 4.49

On 70% of days, a hospital admits at least one heart attack patient. On 30% of the days, no heart attack patients are admitted. Identify each case below as a binomial or negative binomial case, and compute the probability.[38]

(G)

(a) What is the probability the hospital will admit a heart attack patient on exactly three days this week?

(b) What is the probability the second day with a heart attack patient will be the fourth day of the week?

(c) What is the probability the fifth day of next month will be the first day with a heart attack patient?

[38] In each part, $p = 0.7$. (a) The number of days is fixed, so this is binomial. The parameters are $k = 3$ and $n = 7$: 0.097. (b) The last "success" (admitting a heart attack patient) is fixed to the last day, so we should apply the negative binomial distribution. The parameters are $k = 2$, $n = 4$: 0.132. (c) This problem is negative binomial with $k = 1$ and $n = 5$: 0.006. Note that the negative binomial case when $k = 1$ is the same as using the geometric distribution.

Exercises

4.27 Rolling a die. Calculate the following probabilities and indicate which probability distribution model is appropriate in each case. You roll a fair die 5 times. What is the probability of rolling

(a) the first 6 on the fifth roll?

(b) exactly three 6s?

(c) the third 6 on the fifth roll?

4.28 Playing darts. Calculate the following probabilities and indicate which probability distribution model is appropriate in each case. A very good darts player can hit the bull's eye (red circle in the center of the dart board) 65% of the time. What is the probability that he

(a) hits the bullseye for the 10^{th} time on the 15^{th} try?

(b) hits the bullseye 10 times in 15 tries?

(c) hits the first bullseye on the third try?

4.29 Sampling at school. For a sociology class project you are asked to conduct a survey on 20 students at your school. You decide to stand outside of your dorm's cafeteria and conduct the survey on a random sample of 20 students leaving the cafeteria after dinner one evening. Your dorm is comprised of 45% males and 55% females.

(a) Which probability model is most appropriate for calculating the probability that the 4^{th} person you survey is the 2^{nd} female? Explain.

(b) Compute the probability from part (a).

(c) The three possible scenarios that lead to 4^{th} person you survey being the 2^{nd} female are

$$\{M, M, F, F\}, \{M, F, M, F\}, \{F, M, M, F\}$$

One common feature among these scenarios is that the last trial is always female. In the first three trials there are 2 males and 1 female. Use the binomial coefficient to confirm that there are 3 ways of ordering 2 males and 1 female.

(d) Use the findings presented in part (c) to explain why the formula for the coefficient for the negative binomial is $\binom{n-1}{k-1}$ while the formula for the binomial coefficient is $\binom{n}{k}$.

4.30 Serving in volleyball. A not-so-skilled volleyball player has a 15% chance of making the serve, which involves hitting the ball so it passes over the net on a trajectory such that it will land in the opposing team's court. Suppose that her serves are independent of each other.

(a) What is the probability that on the 10^{th} try she will make her 3^{rd} successful serve?

(b) Suppose she has made two successful serves in nine attempts. What is the probability that her 10^{th} serve will be successful?

(c) Even though parts (a) and (b) discuss the same scenario, the probabilities you calculated should be different. Can you explain the reason for this discrepancy?

4.5 Poisson distribution

EXAMPLE 4.50

There are about 8 million individuals in New York City. How many individuals might we expect to be hospitalized for acute myocardial infarction (AMI), i.e. a heart attack, each day? According to historical records, the average number is about 4.4 individuals. However, we would also like to know the approximate distribution of counts. What would a histogram of the number of AMI occurrences each day look like if we recorded the daily counts over an entire year?

E

A histogram of the number of occurrences of AMI on 365 days for NYC is shown in Figure 4.12.[39] The sample mean (4.38) is similar to the historical average of 4.4. The sample standard deviation is about 2, and the histogram indicates that about 70% of the data fall between 2.4 and 6.4. The distribution's shape is unimodal and skewed to the right.

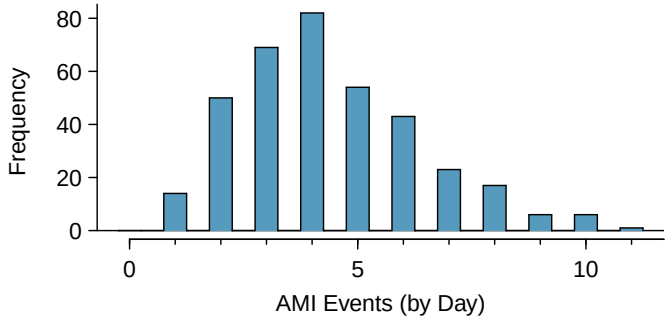


Figure 4.12: A histogram of the number of occurrences of AMI on 365 separate days in NYC.

The **Poisson distribution** is often useful for estimating the number of events in a large population over a unit of time. For instance, consider each of the following events:

- having a heart attack,
- getting married, and
- getting struck by lightning.

The Poisson distribution helps us describe the number of such events that will occur in a day for a fixed population if the individuals within the population are independent. The Poisson distribution could also be used over another unit of time, such as an hour or a week.

The histogram in Figure 4.12 approximates a Poisson distribution with rate equal to 4.4. The **rate** for a Poisson distribution is the average number of occurrences in a mostly-fixed population per unit of time. In Example 4.50, the time unit is a day, the population is all New York City residents, and the historical rate is 4.4. The parameter in the Poisson distribution is the rate – or how many events we expect to observe – and it is typically denoted by λ (the Greek letter *lambda*) or μ. Using the rate, we can describe the probability of observing exactly k events in a single unit of time.

[39] These data are simulated. In practice, we should check for an association between successive days.

POISSON DISTRIBUTION

Suppose we are watching for events and the number of observed events follows a Poisson distribution with rate λ. Then

$$P(\text{observe } k \text{ events}) = \frac{\lambda^k e^{-\lambda}}{k!}$$

where k may take a value 0, 1, 2, and so on, and $k!$ represents k-factorial, as described on page 150. The letter $e \approx 2.718$ is the base of the natural logarithm. The mean and standard deviation of this distribution are λ and $\sqrt{\lambda}$, respectively.

We will leave a rigorous set of conditions for the Poisson distribution to a later course. However, we offer a few simple guidelines that can be used for an initial evaluation of whether the Poisson model would be appropriate.

A random variable may follow a Poisson distribution if we are looking for the number of events, the population that generates such events is large, and the events occur independently of each other.

Even when events are not really independent – for instance, Saturdays and Sundays are especially popular for weddings – a Poisson model may sometimes still be reasonable if we allow it to have a different rate for different times. In the wedding example, the rate would be modeled as higher on weekends than on weekdays. The idea of modeling rates for a Poisson distribution against a second variable such as the day of week forms the foundation of some more advanced methods that fall in the realm of **generalized linear models**. In Chapters 8 and 9, we will discuss a foundation of linear models.

Exercises

4.31 Customers at a coffee shop. A coffee shop serves an average of 75 customers per hour during the morning rush.

(a) Which distribution have we studied that is most appropriate for calculating the probability of a given number of customers arriving within one hour during this time of day?

(b) What are the mean and the standard deviation of the number of customers this coffee shop serves in one hour during this time of day?

(c) Would it be considered unusually low if only 60 customers showed up to this coffee shop in one hour during this time of day?

(d) Calculate the probability that this coffee shop serves 70 customers in one hour during this time of day.

4.32 Stenographer's typos. A very skilled court stenographer makes one typographical error (typo) per hour on average.

(a) What probability distribution is most appropriate for calculating the probability of a given number of typos this stenographer makes in an hour?

(b) What are the mean and the standard deviation of the number of typos this stenographer makes?

(c) Would it be considered unusual if this stenographer made 4 typos in a given hour?

(d) Calculate the probability that this stenographer makes at most 2 typos in a given hour.

4.33 How many cars show up? For Monday through Thursday when there isn't a holiday, the average number of vehicles that visit a particular retailer between 2pm and 3pm each afternoon is 6.5, and the number of cars that show up on any given day follows a Poisson distribution.

(a) What is the probability that exactly 5 cars will show up next Monday?

(b) What is the probability that 0, 1, or 2 cars will show up next Monday between 2pm and 3pm?

(c) There is an average of 11.7 people who visit during those same hours from vehicles. Is it likely that the number of people visiting by car during this hour is also Poisson? Explain.

4.34 Lost baggage. Occasionally an airline will lose a bag. Suppose a small airline has found it can reasonably model the number of bags lost each weekday using a Poisson model with a mean of 2.2 bags.

(a) What is the probability that the airline will lose no bags next Monday?

(b) What is the probability that the airline will lose 0, 1, or 2 bags on next Monday?

(c) Suppose the airline expands over the course of the next 3 years, doubling the number of flights it makes, and the CEO asks you if it's reasonable for them to continue using the Poisson model with a mean of 2.2. What is an appropriate recommendation? Explain.

Chapter exercises

4.35 Roulette winnings. In the game of roulette, a wheel is spun and you place bets on where it will stop. One popular bet is that it will stop on a red slot; such a bet has an 18/38 chance of winning. If it stops on red, you double the money you bet. If not, you lose the money you bet. Suppose you play 3 times, each time with a $1 bet. Let Y represent the total amount won or lost. Write a probability model for Y.

4.36 Speeding on the I-5, Part I. The distribution of passenger vehicle speeds traveling on the Interstate 5 Freeway (I-5) in California is nearly normal with a mean of 72.6 miles/hour and a standard deviation of 4.78 miles/hour.[40]

(a) What percent of passenger vehicles travel slower than 80 miles/hour?

(b) What percent of passenger vehicles travel between 60 and 80 miles/hour?

(c) How fast do the fastest 5% of passenger vehicles travel?

(d) The speed limit on this stretch of the I-5 is 70 miles/hour. Approximate what percentage of the passenger vehicles travel above the speed limit on this stretch of the I-5.

4.37 University admissions. Suppose a university announced that it admitted 2,500 students for the following year's freshman class. However, the university has dorm room spots for only 1,786 freshman students. If there is a 70% chance that an admitted student will decide to accept the offer and attend this university, what is the approximate probability that the university will not have enough dormitory room spots for the freshman class?

4.38 Speeding on the I-5, Part II. Exercise 4.36 states that the distribution of speeds of cars traveling on the Interstate 5 Freeway (I-5) in California is nearly normal with a mean of 72.6 miles/hour and a standard deviation of 4.78 miles/hour. The speed limit on this stretch of the I-5 is 70 miles/hour.

(a) A highway patrol officer is hidden on the side of the freeway. What is the probability that 5 cars pass and none are speeding? Assume that the speeds of the cars are independent of each other.

(b) On average, how many cars would the highway patrol officer expect to watch until the first car that is speeding? What is the standard deviation of the number of cars he would expect to watch?

4.39 Auto insurance premiums. Suppose a newspaper article states that the distribution of auto insurance premiums for residents of California is approximately normal with a mean of $1,650. The article also states that 25% of California residents pay more than $1,800.

(a) What is the Z-score that corresponds to the top 25% (or the 75^{th} percentile) of the standard normal distribution?

(b) What is the mean insurance cost? What is the cutoff for the 75th percentile?

(c) Identify the standard deviation of insurance premiums in California.

4.40 SAT scores. SAT scores (out of 1600) are distributed normally with a mean of 1100 and a standard deviation of 200. Suppose a school council awards a certificate of excellence to all students who score at least 1350 on the SAT, and suppose we pick one of the recognized students at random. What is the probability this student's score will be at least 1500? (The material covered in Section 3.2 on conditional probability would be useful for this question.)

4.41 Married women. The American Community Survey estimates that 47.1% of women ages 15 years and over are married.[41]

(a) We randomly select three women between these ages. What is the probability that the third woman selected is the only one who is married?

(b) What is the probability that all three randomly selected women are married?

(c) On average, how many women would you expect to sample before selecting a married woman? What is the standard deviation?

(d) If the proportion of married women was actually 30%, how many women would you expect to sample before selecting a married woman? What is the standard deviation?

(e) Based on your answers to parts (c) and (d), how does decreasing the probability of an event affect the mean and standard deviation of the wait time until success?

[40] S. Johnson and D. Murray. "Empirical Analysis of Truck and Automobile Speeds on Rural Interstates: Impact of Posted Speed Limits". In: *Transportation Research Board 89th Annual Meeting*. 2010.

[41] U.S. Census Bureau, 2010 American Community Survey, Marital Status.

4.42 Survey response rate. Pew Research reported that the typical response rate to their surveys is only 9%. If for a particular survey 15,000 households are contacted, what is the probability that at least 1,500 will agree to respond?[42]

4.43 Overweight baggage. Suppose weights of the checked baggage of airline passengers follow a nearly normal distribution with mean 45 pounds and standard deviation 3.2 pounds. Most airlines charge a fee for baggage that weigh in excess of 50 pounds. Determine what percent of airline passengers incur this fee.

4.44 Heights of 10 year olds, Part I. Heights of 10 year olds, regardless of gender, closely follow a normal distribution with mean 55 inches and standard deviation 6 inches.

(a) What is the probability that a randomly chosen 10 year old is shorter than 48 inches?

(b) What is the probability that a randomly chosen 10 year old is between 60 and 65 inches?

(c) If the tallest 10% of the class is considered "very tall", what is the height cutoff for "very tall"?

4.45 Buying books on Ebay. Suppose you're considering buying your expensive chemistry textbook on Ebay. Looking at past auctions suggests that the prices of this textbook follow an approximately normal distribution with mean \$89 and standard deviation \$15.

(a) What is the probability that a randomly selected auction for this book closes at more than \$100?

(b) Ebay allows you to set your maximum bid price so that if someone outbids you on an auction you can automatically outbid them, up to the maximum bid price you set. If you are only bidding on one auction, what are the advantages and disadvantages of setting a bid price too high or too low? What if you are bidding on multiple auctions?

(c) If you watched 10 auctions, roughly what percentile might you use for a maximum bid cutoff to be somewhat sure that you will win one of these ten auctions? Is it possible to find a cutoff point that will ensure that you win an auction?

(d) If you are willing to track up to ten auctions closely, about what price might you use as your maximum bid price if you want to be somewhat sure that you will buy one of these ten books?

4.46 Heights of 10 year olds, Part II. Heights of 10 year olds, regardless of gender, closely follow a normal distribution with mean 55 inches and standard deviation 6 inches.

(a) The height requirement for *Batman the Ride* at Six Flags Magic Mountain is 54 inches. What percent of 10 year olds cannot go on this ride?

(b) Suppose there are four 10 year olds. What is the chance that at least two of them will be able to ride *Batman the Ride*?

(c) Suppose you work at the park to help them better understand their customers' demographics, and you are counting people as they enter the park. What is the chance that the first 10 year old you see who can ride *Batman the Ride* is the 3rd 10 year old who enters the park?

(d) What is the chance that the fifth 10 year old you see who can ride *Batman the Ride* is the 12th 10 year old who enters the park?

4.47 Heights of 10 year olds, Part III. Heights of 10 year olds, regardless of gender, closely follow a normal distribution with mean 55 inches and standard deviation 6 inches.

(a) What fraction of 10 year olds are taller than 76 inches?

(b) If there are 2,000 10 year olds entering Six Flags Magic Mountain in a single day, then compute the expected number of 10 year olds who are at least 76 inches tall. (You may assume the heights of the 10-year olds are independent.)

(c) Using the binomial distribution, compute the probability that 0 of the 2,000 10 year olds will be at least 76 inches tall.

(d) The number of 10 year olds who enter Six Flags Magic Mountain and are at least 76 inches tall in a given day follows a Poisson distribution with mean equal to the value found in part (b). Use the Poisson distribution to identify the probability no 10 year old will enter the park who is 76 inches or taller.

4.48 Multiple choice quiz. In a multiple choice quiz there are 5 questions and 4 choices for each question (a, b, c, d). Robin has not studied for the quiz at all, and decides to randomly guess the answers. What is the probability that

(a) the first question she gets right is the 3^{rd} question?

(b) she gets exactly 3 or exactly 4 questions right?

(c) she gets the majority of the questions right?

[42] Pew Research Center, Assessing the Representativeness of Public Opinion Surveys, May 15, 2012.