Chapter 3

Distributions of random variables

3.1 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 3.1. Variables such as SAT scores and heights of US adult males closely follow the normal distribution.

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.

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

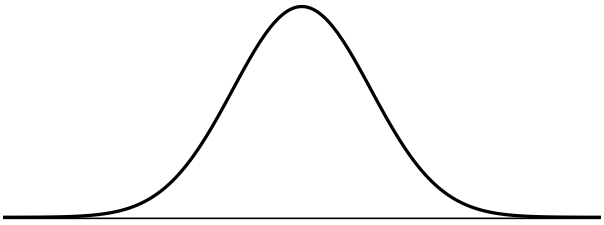


Figure 3.1: A normal curve.

3.1.1 Normal distribution model

The normal distribution model 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 3.2 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 3.3 shows these distributions on the same axis.

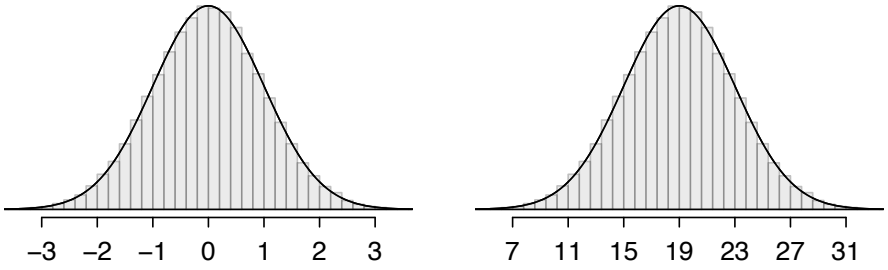


Figure 3.2: Both curves represent the normal distribution, however, they differ in their center and spread. The normal distribution with mean 0 and standard deviation 1 is called the **standard normal distribution**.

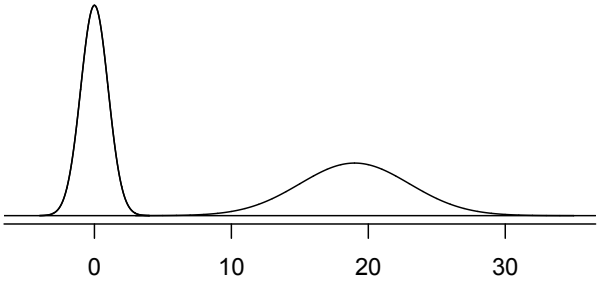


Figure 3.3: The normal models shown in Figure 3.2 but plotted together and on the same scale.

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 3.3 can 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**.

$N(\mu, \sigma)$
Normal dist. with mean μ & st. dev. σ

⊙ **Exercise 3.1** Write down the short-hand for a normal distribution with (a) mean 5 and standard deviation 3, (b) mean -100 and standard deviation 10, and (c) mean 2 and standard deviation 9.[2]

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

	SAT	ACT
Mean	1500	21
SD	300	5

Table 3.4: Mean and standard deviation for the SAT and ACT.

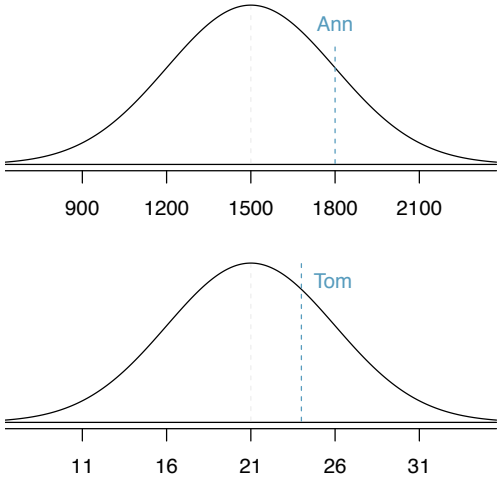


Figure 3.5: Ann's and Tom's scores shown with the distributions of SAT and ACT scores.

3.1.2 Standardizing with Z scores

Example 3.2 Table 3.4 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 1800 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: $1500 + 300 = 1800$. Tom is 0.6 standard deviations above the mean on the ACT: $21 + 0.6 \times 5 = 24$. In Figure 3.5, we can see that Ann tends to do better with respect to everyone else than Tom did, so her score was better.

Z
Z score, the standardized observation

Example 3.2 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 $\mu_{SAT} = 1500$, $\sigma_{SAT} = 300$, and $x_{Ann} = 1800$, we find Ann's Z score:

$$Z_{Ann} = \frac{x_{Ann} - \mu_{SAT}}{\sigma_{SAT}} = \frac{1800 - 1500}{300} = 1$$

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

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

⊙ **Exercise 3.3** Use Tom's ACT score, 24, along with the ACT mean and standard deviation to compute his Z score.[3]

Observations above the mean always have positive Z scores while those below the mean have negative Z scores. If an observation is equal to the mean (e.g. SAT score of 1500), then the Z score is 0.

⊙ **Exercise 3.4** 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]

⊙ **Exercise 3.5** Head lengths of brushtail possums follow a nearly 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. One 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.

⊙ **Exercise 3.6** Which of the observations in Exercise 3.5 is more unusual?[6]

3.1.3 Normal probability table

● **Example 3.7** Ann from Example 3.2 earned a score of 1800 on her SAT with a corresponding $Z = 1$. She would like to know what percentile she falls in among all SAT test-takers.

Ann's **percentile** is the percentage of people who earned a lower SAT score than Ann. We shade the area representing those individuals in Figure 3.6. The total area under the normal curve is always equal to 1, and the proportion of people who scored below Ann on the SAT is equal to the *area* shaded in Figure 3.6: 0.8413. In other words, Ann is in the 84^{th} percentile of SAT takers.

We can use the normal model to find percentiles. A **normal probability table**, which lists Z scores and corresponding percentiles, can be used to identify a percentile based on the Z score (and vice versa). Statistical software can also be used.

[3] $Z_{Tom} = \frac{x_{Tom} - \mu_{ACT}}{\sigma_{ACT}} = \frac{24-21}{5} = 0.6$

[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.

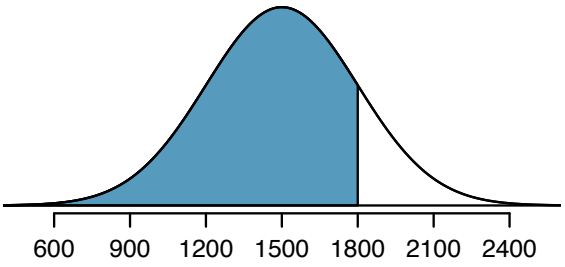


Figure 3.6: The normal model for SAT scores, shading the area of those individuals who scored below Ann.

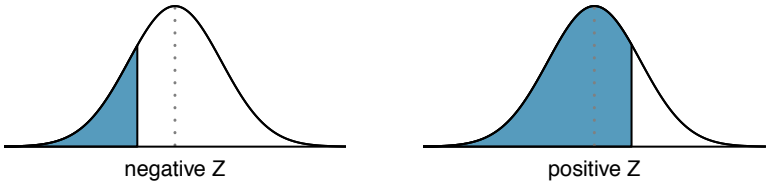


Figure 3.7: The area to the left of Z represents the percentile of the observation.

A normal probability table is given in Appendix B.1 on page 407 and abbreviated in Table 3.8. We use this table to identify the percentile corresponding to any particular Z score. For instance, the percentile of $Z = 0.43$ is shown in row 0.4 and column 0.03 in Table 3.8: 0.6664, or the 66.64^{th} percentile. Generally, we round Z to two decimals, identify the proper row in the normal probability table up through the first decimal, and then determine the column representing the second decimal value. The intersection of this row and column is the percentile of the observation.

We can also find the Z score associated with a percentile. For example, to identify Z for the 80^{th} percentile, we look for the value closest to 0.8000 in the middle portion of the table: 0.7995. We determine the Z score for the 80^{th} percentile by combining the row and column Z values: 0.84.

⊙ **Exercise 3.8** Determine the proportion of SAT test takers who scored better than Ann on the SAT.[7]

3.1.4 Normal probability examples

Cumulative SAT scores are approximated well by a normal model, $N(\mu = 1500, \sigma = 300)$.

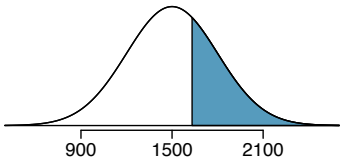
● **Example 3.9** Shannon is a randomly selected SAT taker, and nothing is known about Shannon's SAT aptitude. What is the probability Shannon scores at least 1630 on her SATs?

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 1630, so we shade this upper tail:

[7] If 84% had lower scores than Ann, the number of people who had better scores must be 16%. (Generally ties are ignored when the normal model, or any other continuous distribution, is used.)

	Second decimal place of Z									
Z	0.00	0.01	0.02	*0.03*	**0.04**	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	*0.6664*	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	**0.7995**	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

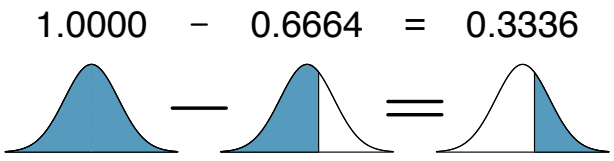
Table 3.8: A section of the normal probability table. The percentile for a normal random variable with $Z = 0.43$ has been *highlighted*, and the percentile closest to 0.8000 has also been **highlighted**.



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 $\mu = 1500$, $\sigma = 300$, and the cutoff value $x = 1630$, the Z score is computed as

$$Z = \frac{x - \mu}{\sigma} = \frac{1630 - 1500}{300} = \frac{130}{300} = 0.43$$

We look up the percentile of $Z = 0.43$ in the normal probability table shown in Table 3.8 or in Appendix B.1 on page 407, which yields 0.6664. However, the percentile describes those who had a Z score *lower* than 0.43. To find the area *above* $Z = 0.43$, we compute one minus the area of the lower tail:



The probability Shannon scores at least 1630 on the SAT is 0.3336.

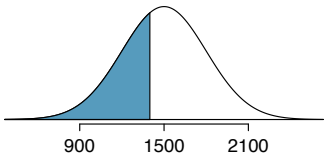
TIP: always draw a picture first, and find the Z score second
For any normal probability situation, *always always always* draw and label the normal curve and shade the area of interest first. The picture will provide an estimate of the probability.

After drawing a figure to represent the situation, identify the Z score for the observation of interest.

⊙ **Exercise 3.10** If the probability of Shannon scoring at least 1630 is 0.3336, then what is the probability she scores less than 1630? Draw the normal curve representing this exercise, shading the lower region instead of the upper one.[8]

● **Example 3.11** Edward earned a 1400 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 1400. These are the scores to the left of 1400.



Identifying the mean $\mu = 1500$, the standard deviation $\sigma = 300$, and the cutoff for the tail area $x = 1400$ makes it easy to compute the Z score:

$$Z = \frac{x - \mu}{\sigma} = \frac{1400 - 1500}{300} = -0.33$$

Using the normal probability table, identify the row of -0.3 and column of 0.03, which corresponds to the probability 0.3707. Edward is at the 37^{th} percentile.

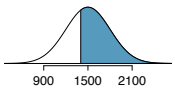
⊙ **Exercise 3.12** Use the results of Example 3.11 to compute the proportion of SAT takers who did better than Edward. Also draw a new picture.[9]

TIP: areas to the right
The normal probability table in most books gives the area to the left. If you would like the area to the right, first find the area to the left and then subtract this amount from one.

⊙ **Exercise 3.13** Stuart earned an SAT score of 2100. Draw a picture for each part. (a) What is his percentile? (b) What percent of SAT takers did better than Stuart?[10]

[8] We found the probability in Example 3.9: 0.6664. A picture for this exercise is represented by the shaded area below "0.6664" in Example 3.9.

[9] If Edward did better than 37% of SAT takers, then about 63% must have done better than him.



[10] Numerical answers: (a) 0.9772. (b) 0.0228.

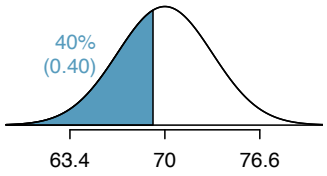
Based on a sample of 100 men,[11] the heights of male adults between the ages 20 and 62 in the US is nearly normal with mean 70.0" and standard deviation 3.3".

⊙ **Exercise 3.14** Mike is 5'7" and Jim is 6'4". (a) What is Mike's height percentile? (b) What is Jim's height percentile? Also draw one picture for each part.[12]

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

● **Example 3.15** Erik's height is at the 40^{th} percentile. How tall is he?

As always, first draw the picture.



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.

Because the percentile is below 50%, we know Z will be negative. Looking in the negative part of the normal probability table, we search for the probability *inside* the table closest to 0.4000. We find that 0.4000 falls in row −0.2 and between columns 0.05 and 0.06. Since it falls closer to 0.05, we take this one: $Z = -0.25$.

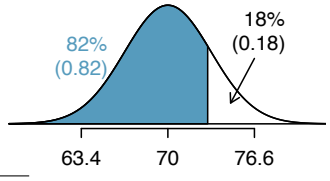
Knowing $Z_{Erik} = -0.25$ and the population parameters $\mu = 70$ and $\sigma = 3.3$ inches, the Z score formula can be set up to determine Erik's unknown height, labeled x_{Erik}:

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

Solving for x_{Erik} yields the height 69.18 inches. That is, Erik is about 5'9" (this is notation for 5-feet, 9-inches).

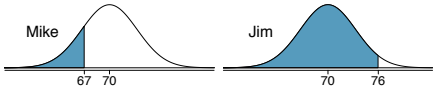
● **Example 3.16** What is the adult male height at the 82^{nd} percentile?

Again, we draw the figure first.



[11] This sample was taken from the USDA Food Commodity Intake Database.

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



Next, we want to find the Z score at the 82^{nd} percentile, which will be a positive value. Looking in the Z table, we find Z falls in row 0.9 and the nearest column is 0.02, i.e. $Z = 0.92$. Finally, the height x is found using the Z score formula with the known mean μ, standard deviation σ, and Z score $Z = 0.92$:

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

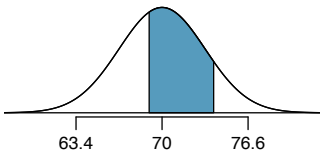
This yields 73.04 inches or about 6'1" as the height at the 82^{nd} percentile.

Exercise 3.17 (a) What is the 95^{th} percentile for SAT scores? (b) What is the 97.5^{th} percentile of the male heights? As always with normal probability problems, first draw a picture.[13]

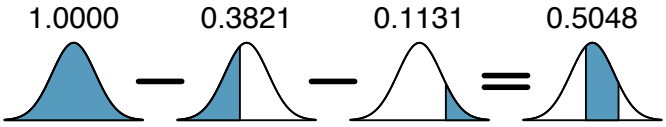
Exercise 3.18 (a) What is the probability that a randomly selected male adult is at least 6'2" (74 inches)? (b) What is the probability that a male adult is shorter than 5'9" (69 inches)?[14]

Example 3.19 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 Exercise 3.18, 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.

Exercise 3.20 What percent of SAT takers get between 1500 and 2000?[15]

Exercise 3.21 What percent of adult males are between 5'5" and 5'7"?[16]

[13] Remember: draw a picture first, then find the Z score. (We leave the pictures to you.) The Z score can be found by using the percentiles and the normal probability table. (a) We look for 0.95 in the probability portion (middle part) of the normal probability table, which leads us to row 1.6 and (about) column 0.05, i.e. $Z_{95} = 1.65$. Knowing $Z_{95} = 1.65$, $\mu = 1500$, and $\sigma = 300$, we setup the Z score formula: $1.65 = \frac{x_{95} - 1500}{300}$. We solve for x_{95}: $x_{95} = 1995$. (b) Similarly, we find $Z_{97.5} = 1.96$, again setup the Z score formula for the heights, and calculate $x_{97.5} = 76.5$.

[14] Numerical answers: (a) 0.1131. (b) 0.3821.

[15] This is an abbreviated solution. (Be sure to draw a figure!) First find the percent who get below 1500 and the percent that get above 2000: $Z_{1500} = 0.00 \rightarrow 0.5000$ (area below), $Z_{2000} = 1.67 \rightarrow 0.0475$ (area above). Final answer: $1.0000 - 0.5000 - 0.0475 = 0.4525$.

[16] 5'5" is 65 inches. 5'7" is 67 inches. Numerical solution: $1.000 - 0.0649 - 0.8183 = 0.1168$, i.e. 11.68%.

3.1.5 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.

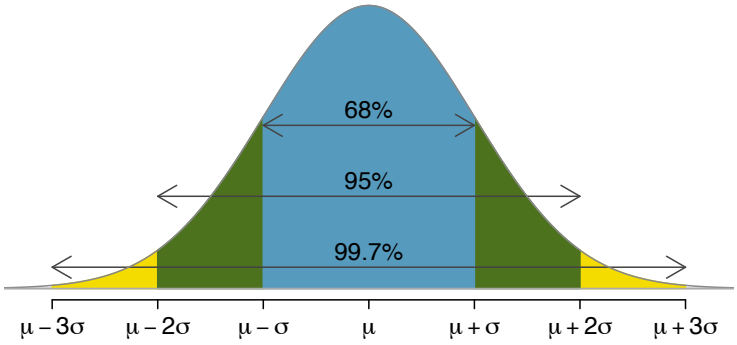


Figure 3.9: Probabilities for falling within 1, 2, and 3 standard deviations of the mean in a normal distribution.

⊙ **Exercise 3.22** Use the Z 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 $Z = 2$.[17]

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-30,000. For 5 and 6 standard deviations, it is about 1-in-3.5 million and 1-in-1 billion, respectively.

⊙ **Exercise 3.23** SAT scores closely follow the normal model with mean $\mu = 1500$ and standard deviation $\sigma = 300$. (a) About what percent of test takers score 900 to 2100? (b) What percent score between 1500 and 2100?[18]

3.2 Evaluating the normal approximation

Many processes can be well approximated by the normal distribution. We have already seen two good examples: SAT scores and the heights of US adult males. While using a normal model can be extremely convenient and helpful, it is important to remember normality is

[17] First draw the pictures. To find the area between $Z = -1$ and $Z = 1$, use the normal probability table to determine the areas below $Z = -1$ and above $Z = 1$. Next verify the area between $Z = -1$ and $Z = 1$ is about 0.68. Repeat this for $Z = -2$ to $Z = 2$ and also for $Z = -3$ to $Z = 3$.

[18] (a) 900 and 2100 represent two standard deviations above and below the mean, which means about 95% of test takers will score between 900 and 2100. (b) Since the normal model is symmetric, then half of the test takers from part (a) ($\frac{95\%}{2} = 47.5\%$ of all test takers) will score 900 to 1500 while 47.5% score between 1500 and 2100.

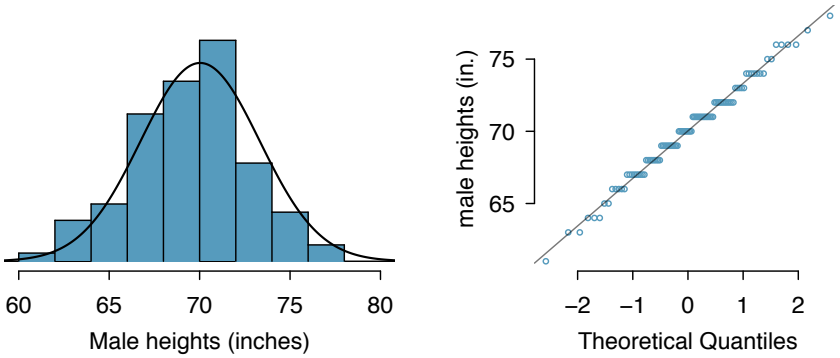


Figure 3.10: A sample of 100 male heights. The observations are rounded to the nearest whole inch, explaining why the points appear to jump in increments in the normal probability plot.

always an approximation. Testing the appropriateness of the normal assumption is a key step in many data analyses.

3.2.1 Normal probability plot

Example 3.15 suggests the distribution of heights of US males is well approximated by the normal model. We are interested in proceeding under the assumption that the data are normally distributed, but first we must check to see if this is reasonable.

There are two visual methods for checking the assumption of normality, which can be implemented and interpreted quickly. The first is a simple histogram with the best fitting normal curve overlaid on the plot, as shown in the left panel of Figure 3.10. The sample mean $\bar{x}$ and standard deviation s are used as the parameters of the best fitting normal curve. The closer this curve fits the histogram, the more reasonable the normal model assumption. Another more common method is examining a **normal probability plot**.[19], shown in the right panel of Figure 3.10. The closer the points are to a perfect straight line, the more confident we can be that the data follow the normal model. We outline the construction of the normal probability plot in Section 3.2.2

● **Example 3.24** Three data sets of 40, 100, and 400 samples were simulated from a normal distribution, and the histograms and normal probability plots of the data sets are shown in Figure 3.11. These will provide a benchmark for what to look for in plots of real data.

The left panels show the histogram (top) and normal probability plot (bottom) for the simulated data set with 40 observations. The data set is too small to really see clear structure in the histogram. The normal probability plot also reflects this, where there are some deviations from the line. However, these deviations are not strong.

The middle panels show diagnostic plots for the data set with 100 simulated observations. The histogram shows more normality and the normal probability plot shows a better fit. While there is one observation that deviates noticeably from the line, it is not particularly extreme.

[19] Also commonly called a **quantile-quantile plot**.

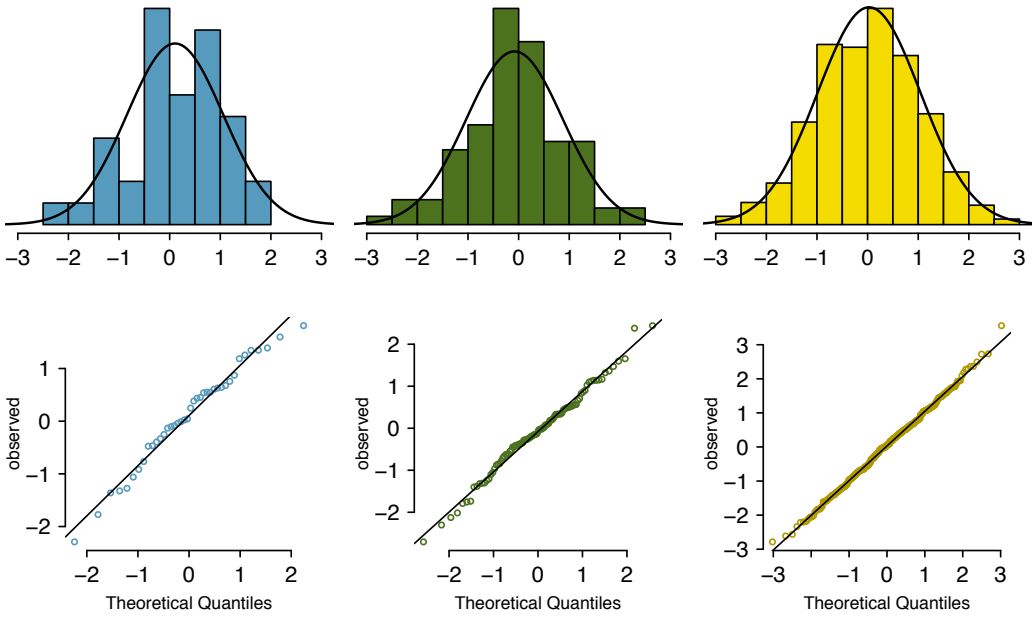


Figure 3.11: Histograms and normal probability plots for three simulated normal data sets; $n = 40$ (left), $n = 100$ (middle), $n = 400$ (right).

The data set with 400 observations has a histogram that greatly resembles the normal distribution, while the normal probability plot is nearly a perfect straight line. Again in the normal probability plot there is one observation (the largest) that deviates slightly from the line. If that observation had deviated 3 times further from the line, it would be of much greater concern in a real data set. Apparent outliers can occur in normally distributed data but they are rare.

Notice the histograms look more normal as the sample size increases, and the normal probability plot becomes straighter and more stable.

Example 3.25 Are NBA player heights normally distributed? Consider all 435 NBA players from the 2008-9 season presented in Figure 3.12.[20]

We first create a histogram and normal probability plot of the NBA player heights. The histogram in the left panel is slightly left skewed, which contrasts with the symmetric normal distribution. The points in the normal probability plot do not appear to closely follow a straight line but show what appears to be a "wave". We can compare these characteristics to the sample of 400 normally distributed observations in Example 3.24 and see that they represent much stronger deviations from the normal model. NBA player heights do not appear to come from a normal distribution.

[20]These data were collected from http://www.nba.com.

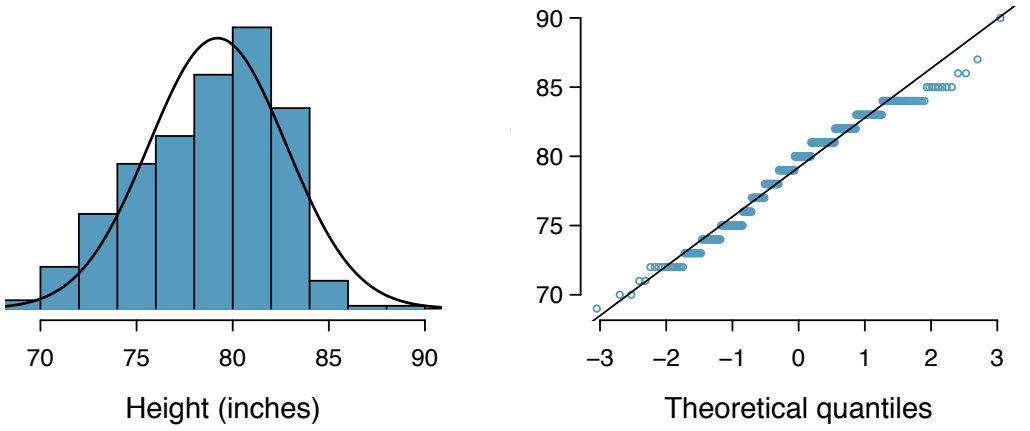


Figure 3.12: Histogram and normal probability plot for the NBA heights from the 2008-9 season.

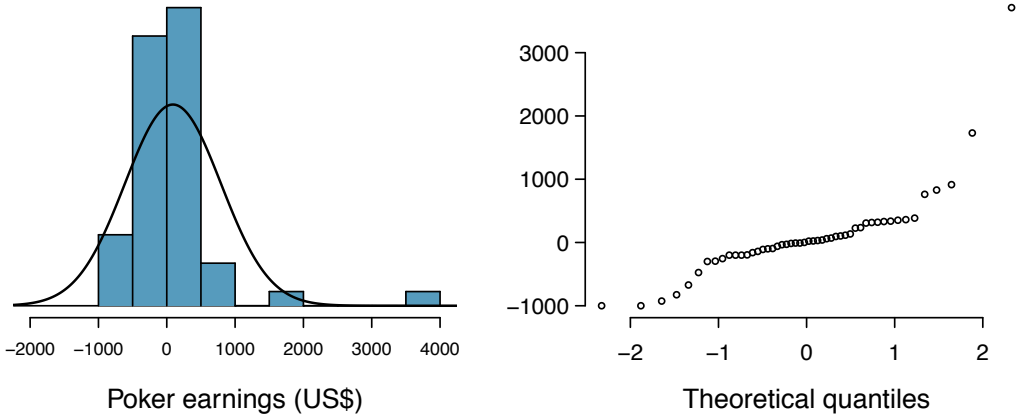


Figure 3.13: A histogram of poker data with the best fitting normal plot and a normal probability plot.

Example 3.26 Can we approximate poker winnings by a normal distribution? We consider the poker winnings of an individual over 50 days. A histogram and normal probability plot of these data are shown in Figure 3.13.

The data are very strongly right skewed in the histogram, which corresponds to the very strong deviations on the upper right component of the normal probability plot. If we compare these results to the sample of 40 normal observations in Example 3.24, it is apparent that these data show very strong deviations from the normal model.

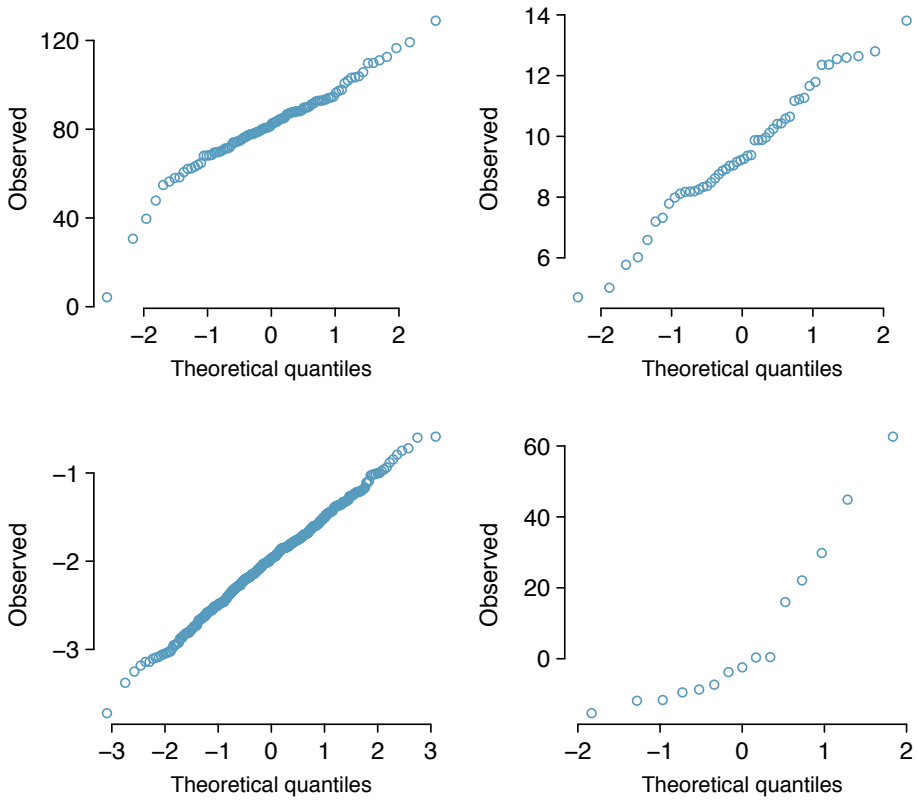


Figure 3.14: Four normal probability plots for Exercise 3.27.

⊙ **Exercise 3.27** Determine which data sets represented in Figure 3.14 plausibly come from a nearly normal distribution. Are you confident in all of your conclusions? There are 100 (top left), 50 (top right), 500 (bottom left), and 15 points (bottom right) in the four plots.[21]

⊙ **Exercise 3.28** Figure 3.15 shows normal probability plots for two distributions that are skewed. One distribution is skewed to the low end (left skewed) and the other to the high end (right skewed). Which is which?[22]

[21] Answers may vary a little. The top-left plot shows some deviations in the smallest values in the data set; specifically, the left tail of the data set has some outliers we should be wary of. The top-right and bottom-left plots do not show any obvious or extreme deviations from the lines for their respective sample sizes, so a normal model would be reasonable for these data sets. The bottom-right plot has a consistent curvature that suggests it is not from the normal distribution. If we examine just the vertical coordinates of these observations, we see that there is a lot of data between -20 and 0, and then about five observations scattered between 0 and 70. This describes a distribution that has a strong right skew.

[22] Examine where the points fall along the vertical axis. In the first plot, most points are near the low end with fewer observations scattered along the high end; this describes a distribution that is skewed to the high end. The second plot shows the opposite features, and this distribution is skewed to the low end.

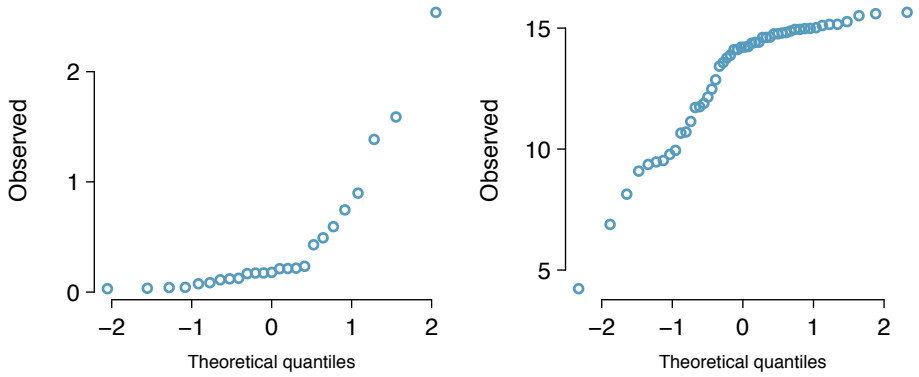


Figure 3.15: Normal probability plots for Exercise 3.28.

3.2.2 Constructing a normal probability plot (special topic)

We construct a normal probability plot for the heights of a sample of 100 men as follows:

(1) Order the observations.

(2) Determine the percentile of each observation in the ordered data set.

(3) Identify the Z score corresponding to each percentile.

(4) Create a scatterplot of the observations (vertical) against the Z scores (horizontal).

If the observations are normally distributed, then their Z scores will approximately correspond to their percentiles and thus to the z_i in Table 3.16.

Observation i	1	2	3	$\cdots$	100
x_i	61	63	63	$\cdots$	78
Percentile	0.99%	1.98%	2.97%	$\cdots$	99.01%
z_i	−2.33	−2.06	−1.89	$\cdots$	2.33

Table 3.16: Construction details for a normal probability plot of 100 men's heights. The first observation is assumed to be at the 0.99^{th} percentile, and the z_i corresponding to a lower tail of 0.0099 is −2.33. To create the plot based on this table, plot each pair of points, (z_i, x_i).

Caution: z_i correspond to percentiles
The z_i in Table 3.16 are *not* the Z scores of the observations but only correspond to the percentiles of the observations.

Because of the complexity of these calculations, normal probability plots are generally created using statistical software.

3.3 Geometric distribution (special topic)

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 2) to construct the geometric distribution.

3.3.1 Bernoulli distribution

Stanley Milgram began a series of experiments in 1963 to estimate what proportion of people would willingly obey an authority and give severe shocks to a stranger. Milgram found that about 65% of people would obey the authority and give such shocks. Over the years, additional research suggested this number is approximately consistent across communities and time.[23]

Each person in Milgram's experiment can be thought of as a **trial**. We label a person a **success** if she refuses to administer the worst shock. A person is labeled a **failure** if she administers the worst shock. Because only 35% of individuals refused to administer the most severe shock, we denote the **probability of a success** with $p = 0.35$. The probability of a failure is sometimes denoted with $q = 1 - p$.

Thus, `success` or `failure` is recorded for each person in the study. When an individual trial only has two possible outcomes, it is called a **Bernoulli random variable**.

Bernoulli random variable, descriptive
A Bernoulli random variable has exactly two possible outcomes. We typically label one of these outcomes a "success" and the other outcome a "failure". We may also denote a success by `1` and a failure by `0`.

TIP: "success" need not be something positive
We chose to label a person who refuses to administer the worst shock 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:

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

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

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

[23] Find further information on Milgram's experiment at www.cnr.berkeley.edu/ucce50/ag-labor/7article/article35.htm.

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.[24]

Bernoulli random variable, mathematical
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

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

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.

3.3.2 Geometric distribution

Example 3.29 Dr. Smith wants to repeat Milgram's experiments but she only wants to sample people until she finds someone who will not inflict the worst shock.[25] If the probability a person will *not* give the most severe shock is still 0.35 and the subjects are independent, what are the chances that she will stop the study after the first person? The second person? The third? What about if it takes her $n - 1$ individuals who will administer the worst shock before finding her first success, i.e. the first success is on the n^{th} person? (If the first success is the fifth person, then we say $n = 5$.)

The probability of stopping after the first person is just the chance the first person will not administer the worst shock: $1 - 0.65 = 0.35$. The probability it will be the second person is

$$\begin{aligned} &P(\text{second person is the first to not administer the worst shock}) \\ &\quad = P(\text{the first will, the second won't}) = (0.65)(0.35) = 0.228 \end{aligned}$$

Likewise, the probability it will be the third person is $(0.65)(0.65)(0.35) = 0.148$.

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 $(0.65)^{n-1}(0.35)$. This is the same as $(1 - 0.35)^{n-1}(0.35)$.

---

[24] If p is the true probability of a success, then the mean of a Bernoulli random variable X is given by

$$\begin{aligned} \mu = E[X] &= P(X = 0) \times 0 + P(X = 1) \times 1 \\ &= (1 - p) \times 0 + p \times 1 = 0 + p = p \end{aligned}$$

Similarly, the variance of X can be computed:

$$\begin{aligned} \sigma^2 &= P(X = 0)(0 - p)^2 + P(X = 1)(1 - p)^2 \\ &= (1 - p)p^2 + p(1 - p)^2 = p(1 - p) \end{aligned}$$

The standard deviation is $\sigma = \sqrt{p(1-p)}$.

[25] This is hypothetical since, in reality, this sort of study probably would not be permitted any longer under current ethical standards.

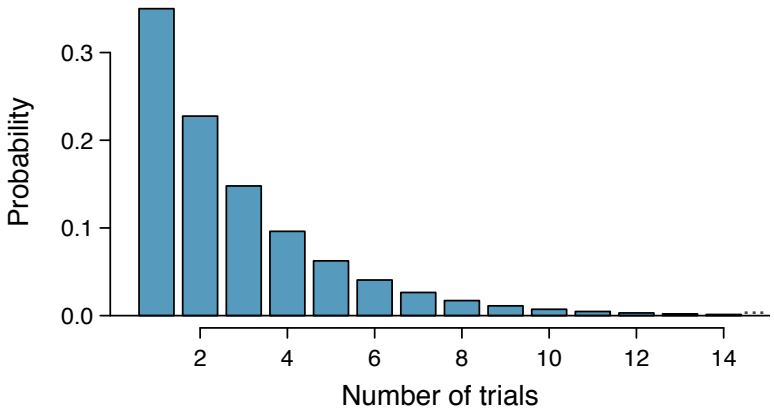


Figure 3.17: The geometric distribution when the probability of success is $p = 0.35$.

Example 3.29 illustrates what is called 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 3.29 is shown in Figure 3.17. In general, the probabilities for a geometric distribution decrease **exponentially** fast.

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

$$(1-p)^{n-1}p \tag{3.30}$$

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

$$\mu = \frac{1}{p} \qquad \sigma^2 = \frac{1-p}{p^2} \qquad \sigma = \sqrt{\frac{1-p}{p^2}} \tag{3.31}$$

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.

The left side of Equation (3.31) says that, on average, it takes $1/p$ trials to get a success. 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.

⊙ **Exercise 3.32** The probability that an individual would refuse to administer the worst shock is said to be about 0.35. If we were to examine individuals until we found one that did not administer the shock, how many people should we expect to check? The first expression in Equation (3.31) may be useful.[26]

● **Example 3.33** What is the chance that Dr. Smith will find the first success within the first 4 people?

This is the chance it is the first ($n = 1$), second ($n = 2$), third ($n = 3$), or fourth ($n = 4$) person as the first success, which are four 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:

$$\begin{aligned} &P(n = 1, 2, 3, \text{ or } 4) \\ &\quad = P(n = 1) + P(n = 2) + P(n = 3) + P(n = 4) \\ &\quad = (0.65)^{1-1}(0.35) + (0.65)^{2-1}(0.35) + (0.65)^{3-1}(0.35) + (0.65)^{4-1}(0.35) \\ &\quad = 0.82 \end{aligned}$$

There is an 82% chance that she will end the study within 4 people.

⊙ **Exercise 3.34** Determine a more clever way to solve Example 3.33. Show that you get the same result.[27]

● **Example 3.35** Suppose in one region it was found that the proportion of people who would administer the worst shock was "only" 55%. If people were randomly selected from this region, what is the expected number of people who must be checked before one was found that would be deemed a success? What is the standard deviation of this waiting time?

A success is when someone will **not** inflict the worst shock, which has probability $p = 1 - 0.55 = 0.45$ for this region. The expected number of people to be checked is $1/p = 1/0.45 = 2.22$ and the standard deviation is $\sqrt{(1-p)/p^2} = 1.65$.

⊙ **Exercise 3.36** Using the results from Example 3.35, $\mu = 2.22$ and $\sigma = 1.65$, would it be appropriate to use the normal model to find what proportion of experiments would end in 3 or fewer trials?[28]

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.

---

[26] We would expect to see about $1/0.35 = 2.86$ individuals to find the first success.

[27] First find the probability of the complement: $P(\text{no success in first 4 trials}) = 0.65^4 = 0.18$. Next, compute one minus this probability: $1 - P(\text{no success in 4 trials}) = 1 - 0.18 = 0.82$.

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

3.4 Binomial distribution (special topic)

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

Let's consider a scenario where one person refuses:

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

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

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

3.4.1 The binomial distribution

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Binomial distribution

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

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

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

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

TIP: Is it binomial? Four conditions to check.

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

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

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

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

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

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

Dealing with the factorial part:

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

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

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

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

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

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

---

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

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

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

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

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

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

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

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

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

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

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

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

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

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

3.4.2 Normal approximation to the binomial distribution

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

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

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

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

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

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

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

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

Normal approximation of the binomial distribution

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

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

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

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

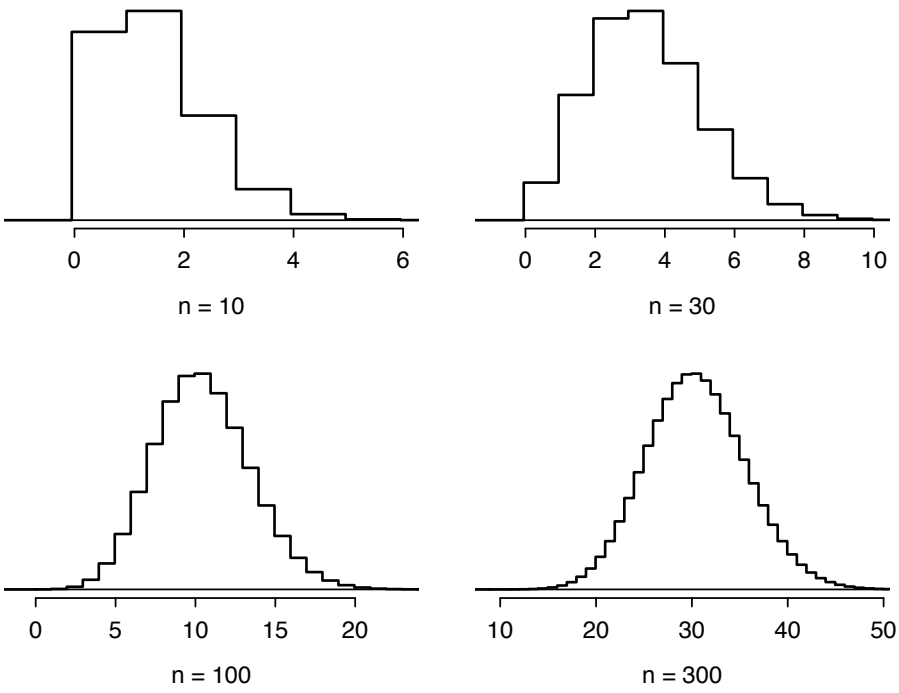


Figure 3.18: Hollow histograms of samples from the binomial model when $p = 0.10$. The sample sizes for the four plots are $n = 10$, 30, 100, and 300, respectively.

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

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

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

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

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

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

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

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

3.4.3 The normal approximation breaks down on small intervals

Caution: The normal approximation may fail 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 69, 70, or 71 smokers in 400 when $p = 0.20$. With such a large sample, we might be tempted to apply the normal approximation and use the range 69 to 71. However, we would find that the binomial solution and the normal approximation notably differ:

Binomial: 0.0703 Normal: 0.0476

We can identify the cause of this discrepancy using Figure 3.19, 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.

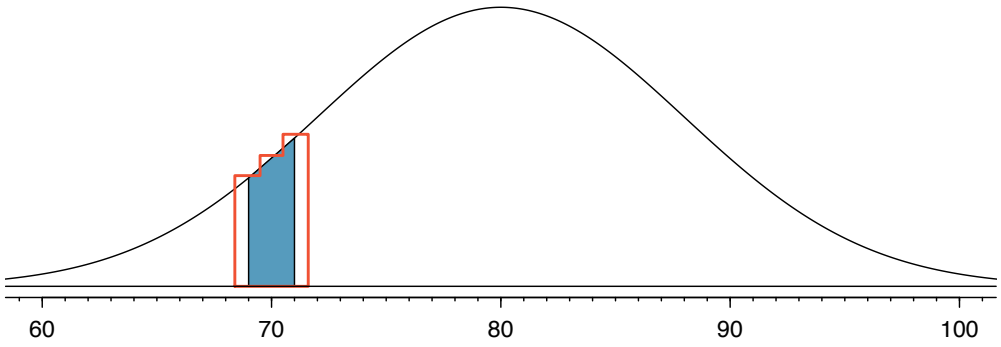


Figure 3.19: A normal curve with the area between 69 and 71 shaded. The outlined area represents the exact binomial probability.

TIP: Improving the accuracy of the normal approximation to 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. While it is possible to apply it when computing a tail area, the benefit of the modification usually disappears since the total interval is typically quite wide.

3.5 More discrete distributions (special topic)

3.5.1 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 3.54** 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?

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.[40]

TIP: 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.

⊙ **Exercise 3.55** 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.[41]

● **Example 3.56** 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.

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 Table 3.20. 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.

⊙ **Exercise 3.57** Each sequence in Table 3.20 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.[42]

[40] See a similar guide for the binomial distribution on page 138.
[41] 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.
[42] 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	$\overset{1}{S}$	$\overset{2}{S}$	$\overset{3}{S}$	$\overset{4}{S}$
2	F	$\overset{1}{S}$	F	$\overset{2}{S}$	$\overset{3}{S}$	$\overset{4}{S}$
3	F	$\overset{1}{S}$	$\overset{2}{S}$	F	$\overset{3}{S}$	$\overset{4}{S}$
4	F	$\overset{1}{S}$	$\overset{2}{S}$	$\overset{3}{S}$	F	$\overset{4}{S}$
5	$\overset{1}{S}$	F	F	$\overset{2}{S}$	$\overset{3}{S}$	$\overset{4}{S}$
6	$\overset{1}{S}$	F	$\overset{2}{S}$	F	$\overset{3}{S}$	$\overset{4}{S}$
7	$\overset{1}{S}$	F	$\overset{2}{S}$	$\overset{3}{S}$	F	$\overset{4}{S}$
8	$\overset{1}{S}$	$\overset{2}{S}$	F	F	$\overset{3}{S}$	$\overset{4}{S}$
9	$\overset{1}{S}$	$\overset{2}{S}$	F	$\overset{3}{S}$	F	$\overset{4}{S}$
10	$\overset{1}{S}$	$\overset{2}{S}$	$\overset{3}{S}$	F	F	$\overset{4}{S}$

Table 3.20: 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 Table 3.20. 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 Exercise 3.57 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 3.4. 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 138.

Negative binomial distribution

The negative binomial distribution describes the probability of observing the k^{th} success on the n^{th} trial:

$$P(\text{the } k^{th} \text{ success on the } n^{th} \text{ trial}) = \binom{n-1}{k-1} p^k (1-p)^{n-k} \qquad (3.58)$$

where p is the probability an individual trial is a success. All trials are assumed to be independent.

Example 3.59 Show using Equation (3.58) that the probability Brian kicks his fourth successful field goal on the sixth attempt is 0.164.

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 $n = 6$.

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

Exercise 3.60 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?[43]

Exercise 3.61 Assume Brian's kick attempts are independent. What is the probability that Brian will kick his fourth field goal within 5 attempts?[44]

[43] 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.

[44] 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 Equation (3.58) to compute the probability of $n = 4$ tries and $n = 5$ tries, then add those probabilities together:

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

TIP: 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.

⊙ **Exercise 3.62** 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.[45]

(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?

3.5.2 Poisson distribution

● **Example 3.63** 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?

A histogram of the number of occurrences of AMI on 365 days[46] for NYC is shown in Figure 3.21. 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.

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

- 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 short unit of time for a fixed population if the individuals within the population are independent.

[45] 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.

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

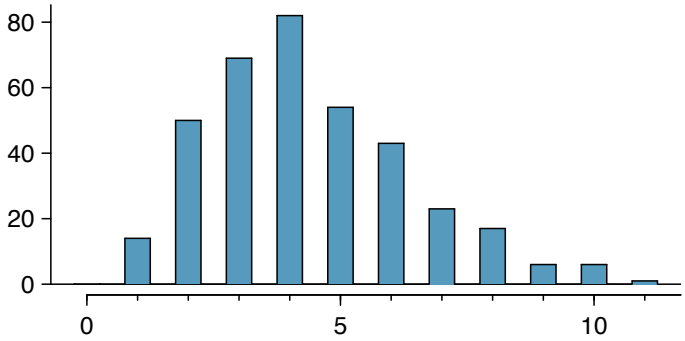


Figure 3.21: A histogram of the number of occurrences of AMI on 365 separate days in NYC.

The histogram in Figure 3.21 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 3.63, 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 rare 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 rare events in a single unit of time.

λ
Rate for the Poisson dist.

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

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

where k may take a value 0, 1, 2, and so on, and $k!$ represents k-factorial, as described on page 138. 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.

TIP: Is it Poisson?
A random variable may follow a Poisson distribution if the event being considered is rare, the population is large, and the events occur independently of each other.

Even when rare 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 `dayOfTheWeek` forms the foundation of some more advanced methods that fall in the realm of **generalized linear models**. In Chapters 7 and 8, we will discuss a foundation of linear models.

3.6 Exercises

3.6.1 Normal distribution

3.1 Area under the curve, 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) $|Z| > 2$

3.2 Area under the curve, 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) $|Z| < 0.5$

3.3 Scores on the GRE, Part I. A college senior who took the Graduate Record Examination exam scored 620 on the Verbal Reasoning section and 670 on the Quantitative Reasoning section. The mean score for Verbal Reasoning section was 462 with a standard deviation of 119, and the mean score for the Quantitative Reasoning was 584 with a standard deviation of 151. Suppose that both distributions are nearly normal.

(a) Write down the short-hand for these two normal distributions.

(b) What is her 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 her raw scores from the two sections would lead to the incorrect conclusion that she did better on the Quantitative Reasoning section.

(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.

3.4 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.

3.5 GRE scores, Part II. In Exercise 3.3 we saw two distributions for GRE scores: $N(\mu = 462, \sigma = 119)$ for the verbal part of the exam and $N(\mu = 584, \sigma = 151)$ 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.

3.6 Triathlon times, Part II. In Exercise 3.4 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.

3.7 Temperatures in LA, 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 cold are the coldest 10% of the days during June in LA?

3.8 Portfolio returns. The Capital Asset Pricing Model 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?

3.9 Temperatures in LA, Part II. Exercise 3.7 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 3.7? Are you surprised? Explain.

3.10 Heights of 10 year olds. 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"?

(d) 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?

3.11 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 LA.

3.12 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.[47]

(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.

3.13 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.

3.14 Find the SD. Find the standard deviation of the distribution in the following situations.

(a) MENSA is an organization whose members have IQs in the top 2% of the population. IQs are normally distributed with mean 100, and the minimum IQ score required for admission to MENSA is 132.

(b) 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.

3.15 Buying books on Ebay. The textbook you need to buy for your chemistry class is expensive at the college bookstore, so you consider buying it on Ebay instead. A look at past auctions suggest that the prices of that chemistry textbook have 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?

[47] 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.

3.16 SAT scores. SAT scores (out of 2400) are distributed normally with a mean of 1500 and a standard deviation of 300. Suppose a school council awards a certificate of excellence to all students who score at least 1900 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 2100? (The material covered in Section 2.2 would be useful for this question.)

3.17 Scores on stats final, Part I. Below are final exam scores of 20 Introductory Statistics students.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
57,	66,	69,	71,	72,	73,	74,	77,	78,	78,	79,	79,	81,	81,	82,	83,	83,	88,	89,	94

The mean score is 77.7 points. with a standard deviation of 8.44 points. Use this information to determine if the scores approximately follow the 68-95-99.7% Rule.

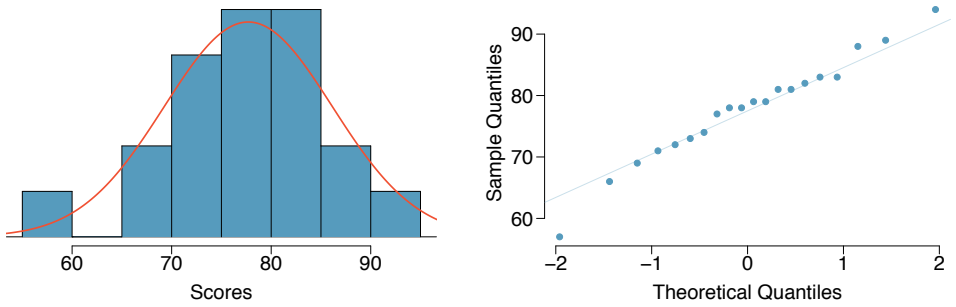
3.18 Heights of female college students, Part I. Below are heights of 25 female college students.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
54,	55,	56,	56,	57,	58,	58,	59,	60,	60,	60,	61,	61,	62,	62,	63,	63,	63,	64,	65,	65,	67,	67,	69,	73

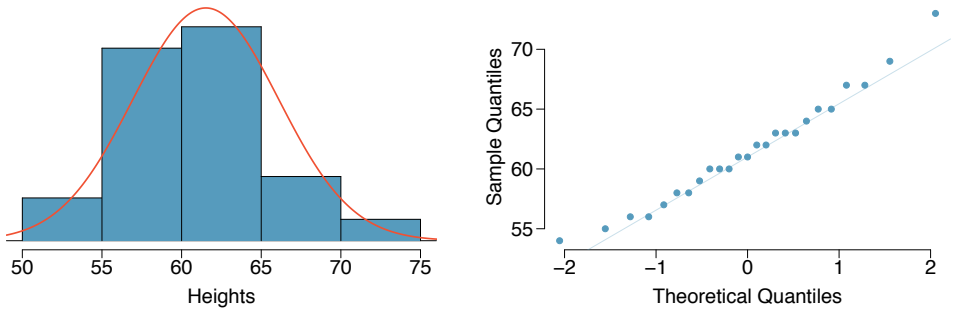
The mean height is 61.52 inches with a standard deviation of 4.58 inches. Use this information to determine if the heights approximately follow the 68-95-99.7% Rule.

3.6.2 Evaluating the Normal approximation

3.19 Scores on stats final, Part II. Exercise 3.17 lists the final exam scores of 20 Introductory Statistics students. Do these data appear to follow a normal distribution? Explain your reasoning using the graphs provided below.



3.20 Heights of female college students, Part II. Exercise 3.18 lists the heights of 25 female college students. Do these data appear to follow a normal distribution? Explain your reasoning using the graphs provided below.



3.6.3 Geometric distribution

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

(a) Cards dealt in a hand of poker.

(b) Outcome of each roll of a die.

3.22 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.

3.23 Married women. The 2010 American Community Survey estimates that 47.1% of women ages 15 years and over are married.[48]

(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?

3.24 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?

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

3.25 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?

3.26 Speeding on the I-5, Part II. Exercise 3.12 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?

3.6.4 Binomial distribution

3.27 Underage drinking, Part I. The Substance Abuse and Mental Health Services Administration estimated that 70% of 18-20 year olds consumed alcoholic beverages in 2008.[49]

(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 the 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?

3.28 Chickenpox, Part I. The National Vaccine Information Center estimates that 90% of Americans have had chickenpox by the time they reach adulthood.[50]

(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 had chickenpox before they reached adulthood? 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?

[49] SAMHSA, Office of Applied Studies, National Survey on Drug Use and Health, 2007 and 2008.
[50] National Vaccine Information Center, Chickenpox, The Disease & The Vaccine Fact Sheet.

3.29 Underage drinking, Part II. We learned in Exercise 3.27 that about 70% of 18-20 year olds consumed alcoholic beverages in 2008. 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)?

3.30 Chickenpox, Part II. We learned in Exercise 3.28 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)?

3.31 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?

3.32 Survey response rate. Pew Research reported in 2012 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?[51]

3.33 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[52]

(a) at least one *nun*?

(b) exactly 2 *nun*s?

(c) exactly 1 *hei*?

(d) at most 2 *gimel*s?

3.34 Arachnophobia. A 2005 Gallup Poll found that 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.[53]

(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?

[51] The Pew Research Center for the People and the Press, Assessing the Representativeness of Public Opinion Surveys, May 15, 2012.

[52] Photo by Staccabees on Flickr.

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

3.35 Eye color, Part II. Exercise 3.25 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?

3.36 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?

3.37 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.

3.38 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?

3.39 Exploring combinations. 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?

3.40 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 events. 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).

3.6.5 More discrete distributions

3.41 Identify the distribution. 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?

3.42 Darts. Calculate the following probabilities and indicate which probability distribution model is appropriate in each case. A very good darts player can hit the bullseye (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?

3.43 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 $\binom{n}{k}$.

3.44 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?

3.45 Customers at a coffee shop, Part I. A coffee shop serves an average of 75 customers per hour during the morning rush.

(a) Which distribution we have studied 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?

3.46 Stenographer's typos, Part I. 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?

3.47 Customers at a coffee shop, Part II. Exercise 3.45 gives the average number of customers visiting a particular coffee shop during the morning rush hour as 75. Calculate the probability that this coffee shop serves 70 customers in one hour during this time of day?

3.48 Stenographer's typos, Part II. Exercise 3.46 gives the average number of typos of a very skilled court stenographer as 1 per hour. Calculate the probability that this stenographer makes at most 2 typos in a given hour.