Lesson 4
Probability, Probability Distributions, and the Normal Distribution

Outline of the Lesson

You already know quite a bit about the subject matter of this lesson, since each of the first three lessons has contained a section on connections to probability. This is not to say you are a probability expert, nor is becoming an expert the goal of this lesson. (To become an expert would require mastery of the material in several complete college-level and graduate-level courses; the discipline is large and complex.) However, you are well on your way to understanding probability in the context that is necessary for our purpose in this course. The main reason we wish to study probability is to help you understand the logic underlying inferential statistics, the subject matter of the second part of this course. The primary purpose of this lesson is to develop the tools and understanding you will use in that study of inferential statistics. Some instructors may, as a secondary purpose, choose to further enhance the depth of your understanding of the fascinating and important topic of probability.

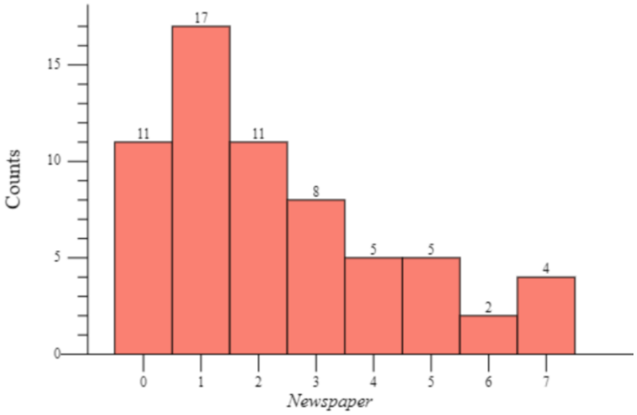
4.1 – Probability

Probability is intimately related to proportion. For example, if 12 people in a class of 80 are smokers, then the proportion of smokers in the class is 12/80 = 0.15 = 15%. If I put the names of the people on index cards, thoroughly shuffle the cards, and pick a card at random, the probability that I have picked a smoker is 12 out of 80, so 12/80 = 0.15 = 15%. (It is more usual to write 15% when talking of proportions, and 0.15 when talking of probabilities, but either is correct in either setting.)

Now consider an experiment in which I select a person at random from the class, note if they are a smoker, then put their card back in the deck and reshuffle before picking again. Each time I pick, the probability of choosing a smoker is 0.15 or 15%. Suppose I do this 100 times – would I expect to get a smoker exactly 15 times, since the probability of choosing a smoker is 15%? The answer is no – due to the randomness the number might be somewhat fewer or somewhat more than 15, but I would expect it to be reasonably close to 15. (I would be surprised if I got a smoker every time, for example!) Moreover, if I repeat the experiment 1000 times, I would expect the proportion of smokers chosen to be even closer to 15%. For random experiments, probability is related to long-term proportions.

4.2 – Probability Distributions

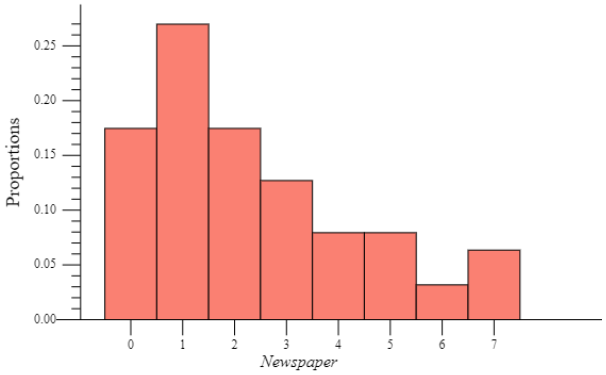
The concept of a probability distribution is closely related to the concepts of frequency tables and histograms which we studied in Lessons 1 and 2. As an example of the connection, consider this histogram:



This is data from the first-day survey described in previous lessons. The 63 students were asked, “How many times a week do you read a newspaper?” In that class, 11 of the 63 students (17.46%) answered 0 to the question, 17 (26.98%) answered 1, and so on. If we turn this discussion of proportions into a discussion of probabilities, we can list the probabilities for the possible answers in a table:

Answer	Probability of selecting a student who gave this answer
0	.1746
1	.2698
2	.1746
3	.1270
4	.0794
5	.0794
6	.0317
7	.0635

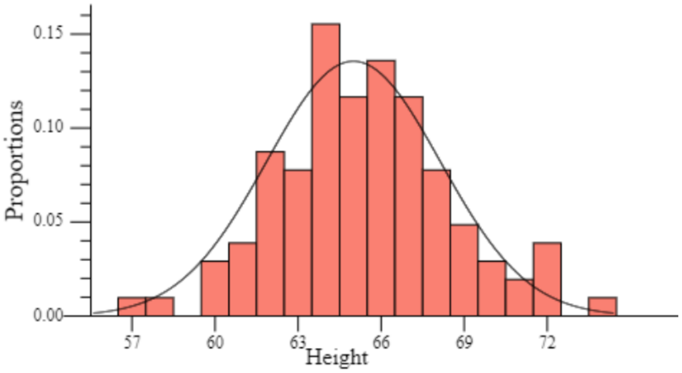
This is a *probability distribution*. It lists all the possible values of a *random variable*, along with the probability for each value. Now, if we graph this probability distribution, we get a histogram that looks identical to the original histogram, except that the heights of the rectangles are proportions – that is, probabilities – instead of frequencies. Here is that graph:



Connections to area

As we discussed back in Lesson 2, in addition to the connection between proportion and probability, there is an additional connection to area. The area of the rectangle for the answer "1 time a week" is 26.98% of the entire area of the histogram – which exactly corresponds to the probability of selecting a student who answered "1 time a week." Similarly, the total area for the rectangles for the answers 1, 2, and 3 is 57.14% of the entire area, which matches the probability of selecting a student who gave one of those answers.

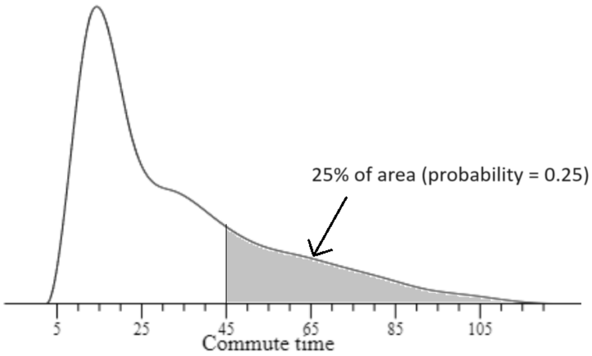
There is one more part of Lesson 2 to remind you of, namely the use of smooth curves to approximate histograms, especially when the underlying variable is continuous. For example, here is a graph showing a histogram for heights overlaid with a smooth curve.



As we now know, this graph can be viewed as the graph of a probability distribution for the heights of the particular group of people represented by the histogram.

As we will see, for graphs of probability distributions represented by smooth curves, the connection between area and probability helps us reason about various probabilities – it translates the relatively abstract notion of probability into the very concrete notion of area. The following exercise illustrates some of the ways we will make use of this connection.

Exercise 1[1]: Here is the smooth curve representing a histogram for the commuting time of a group of individuals:



The graph is a probability distribution, so that the total area under the graph is 1 (that is, 100%). The area to the right of 45 minutes is 0.25, or 25% of the total area. This means that 25% of the people in the survey have commutes longer than 45 minutes. Equivalently, the probability of randomly selecting an individual with a commute longer than 45 minutes is 0.25.

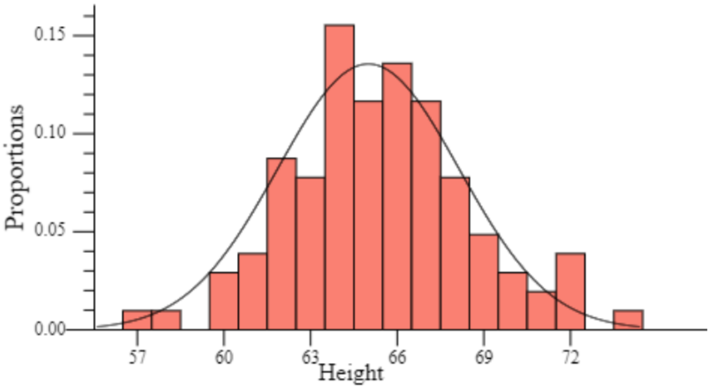
a. What is the probability for a commute under 45 minutes?
b. The researchers report that the probability of a commute less than 15 minutes is 0.22. Shade the area of the graph that corresponds to this probability.
c. What is the probability for a commute between 15 and 45 minutes?

Comment. There is a subtlety in the answers to this exercise that you may or may not have noticed. When we say the area to the right of 45 minutes is 0.25, we can write P(*commute* > 45) = 0.25. Then in the solution to part (a) we said that the probability of a commute under 45 minutes is 0.75, P(*commute* < 45) = 0.75. Should we not have said that the probability of a commute ***45 minutes or less*** is 0.75, that is P(*commute* ≤ 45) = 0.75? If the commutes are rounded to the nearest integer, this is certainly true. But for continuous distributions such as this, the probability of a commute exactly 45 minutes is 0, so we are correct using either < or ≤. In the various examples that follow, we will do this without further comment.

[1] Solutions to the exercises may be found at the end of the lesson.

4.3 – Introduction to the Normal Distribution and Applications

Consider again this histogram, which shows the heights for a group of adult females.



The histogram is mound-shaped, which is emphasized in this graph by overlaying a smooth curve on the histogram. The histogram has some irregularities, but the smooth curve gives a pretty good approximation to the histogram.

Now imagine including a larger and larger group of adult females in the histogram. If you did this, the irregularities in the histogram would begin to disappear, and the resulting histogram would come closer and closer to exactly matching the smooth curve. This particular type of smooth curve occurs very frequently. As a result, it has been studied extensively. It is called a *normal curve* or a *normal distribution*, and it has well-studied and well-documented properties. If the probability distribution for a particular variable is approximated well by a normal curve, we say the variable is **approximately normal.**

It turns out that adult female heights are approximately normal, with mean 65 inches and standard deviation 3.5 inches. Knowing that adult female heights are approximated well by a normal curve gives a lot of information about those heights. This information, for the normal distribution, makes more precise what we already know about mound-shaped distributions in general. For example, in any mound-shaped distribution, approximately 95% of the data lies within two standard deviations on either side of the mean. In a normal distribution, we can make that more precise in one of two ways:

- 95.44% of the data lies within 2 standard deviations of the mean
- 95% of the data lies within 1.96 standard deviations of the mean

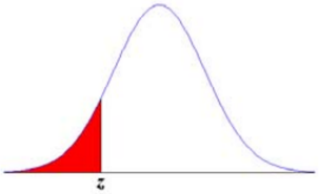
Using Table A, part 1 – the standard normal distribution

We begin our study with the so-called **standard normal distribution.** This is a normal distribution whose mean is 0 and whose standard deviation is 1, which we can abbreviate as N(0, 1) – N for “normal,” 0 for the mean, and 1 for the standard deviation.

When referring to the standard normal distribution, it is customary to use the variable *z* as the variable on the horizontal axis. Statisticians know how to calculate the area to the left of any given value of *z*, and the results of these calculations have been summarized in Table A, a link to which is provided here:

Table A

Here is a copy of the first part of that table, with two entries highlighted Notice that the table “reports the area under the standard normal curve to the left of the value *z*.”



The table reports the area under the standard normal curve to the left of the value *z*.

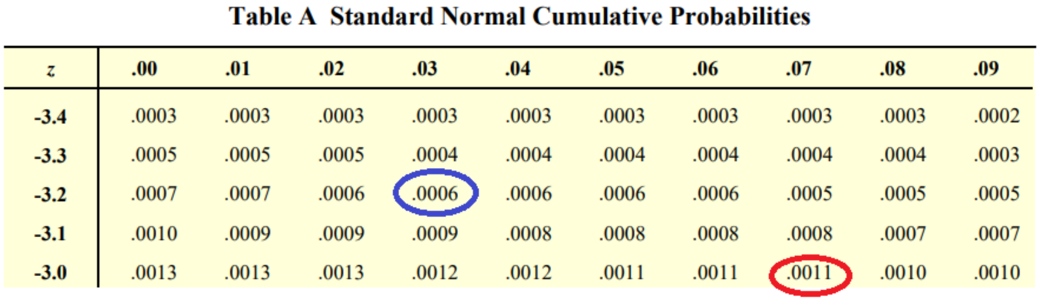


Table A Standard Normal Cumulative Probabilities

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010

Here is how the interpret what you see in the table. First, notice that the rows of the table are labeled by entries in the column headed by **z.** The entries are, from the top to bottom, s –3.4, –3.3, –3.2, and so on. Second, notice that the columns of the table are likewise labeled, with entries from left to right of .00, .01, .02, and so on. Now consider the entry circled in blue. It is in the row labeled –3.2 and the column labeled .03. This tells us that this entry corresponds to the *z* value of –3.23 (–3.2 + .03). The number itself (0.0006) is the area to the left of that *z* value. That is, this entry tells us the following:

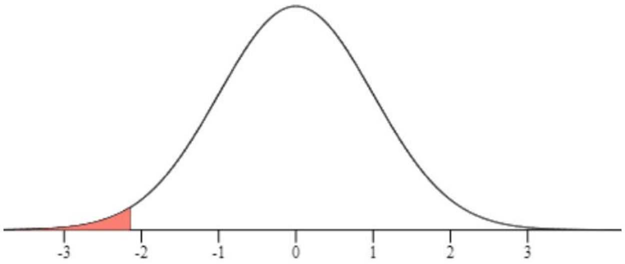
- The area to the left of $z = –3.23$ is 0.0006.
- In terms of percentages, this says that 0.06% of the total area lies to the left of the *z* value –3.23.
- Equivalently, we can state that the probability of having a *z* value less than –3.23 is 0.0006.
- In symbols, we can write this as $P(z < –3.23) = 0.0006$.
- we could also write $P(z \leq\ -3.23) = 0.0006$. As we commented earlier, for continuous distributions the probability $P(z < value)$ and $P(z \leq value)$ are identical.

Exercise 2: Give an interpretation for the other highlighted entry, circled in red.

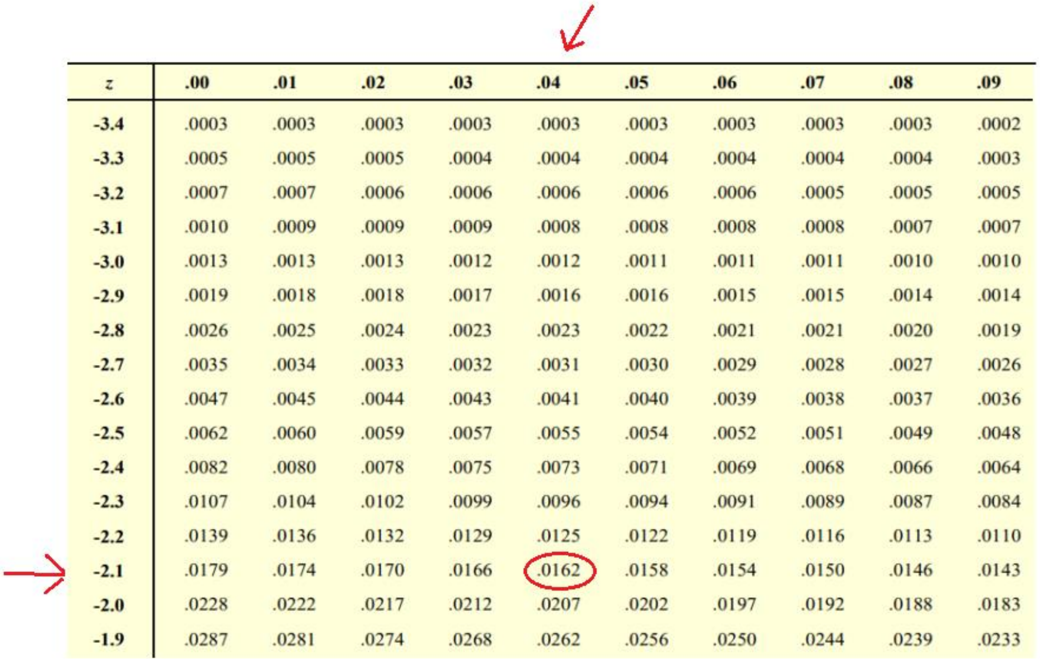
Now that we know how to interpret what we see in the table, we can use that knowledge to answer questions about the standard normal distribution.

Example. What proportion of all possible z values are less than -2.14? That is, calculate $P(z < -2.14)$. (Equivalently, calculate $P(z \leq -2.14)$.

Solution. Here is a graph of the area we wish to calculate:



The key to solving the problem is realizing that –2.14 can be viewed as consisting of two parts: –2.1, and .04. Using Table A, we locate the column labeled “***z***”, and go down that column to the entry –2.1. Next, we locate the column labeled “.04”. See the figure below:



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233

If we go across row “–2.1” and down column “.04” we see the number 0.0162. This is the answer. The probability that a z value is less than –2.14 is 0.0162 or 1.62%.

Example. Calculate $P(z > 1.87)$ (or, equivalently, $P(z \geq 1.87)$.

Solution. Here is a graph of the area we wish to calculate:

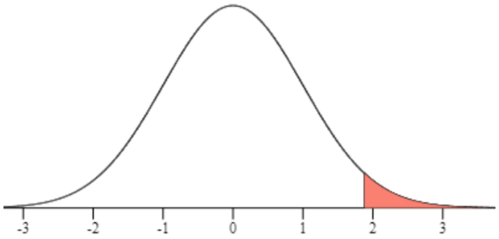
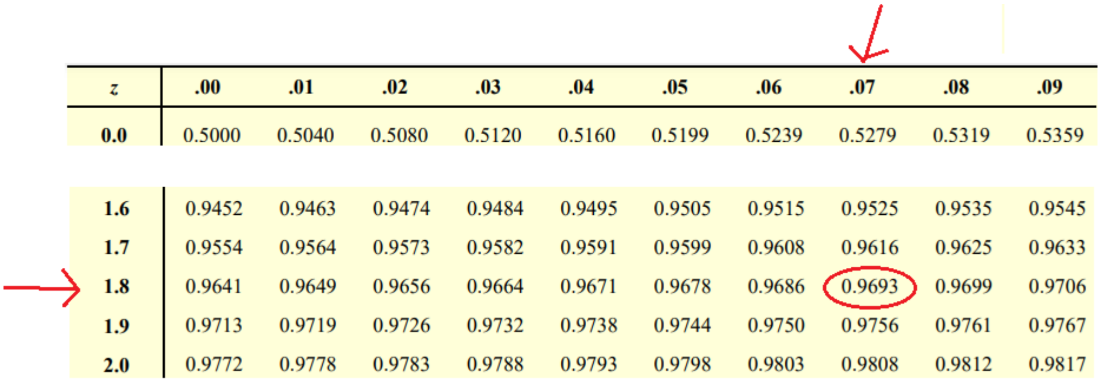


Table A gives probabilities that *z* is less than some given value. To solve our current problem, we need to first calculate $P(z < 1.87)$, then subtract that from 1 (that is, from 100%) to find the requested probability.

Since 1.87 consists of 1.8 and .07, we look in the row labeled 1.8 and the column labeled .07, on the second page of the table, as indicated below.

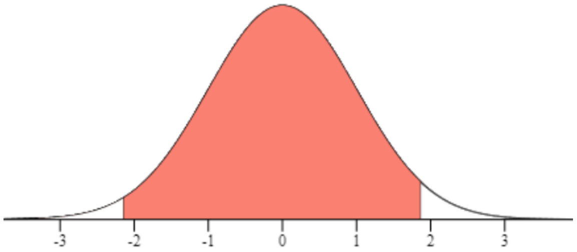


z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817

The circled entry shows that $P(z < 1.87) = .9693$, so $P(z > 1.87) = 1 - .9693 = .0307$, or 3.07%.

Example. Find the area between -2.14 and 1.87. That is, calculate $P(-2.14 < z < 1.87)$.

Solution. Here is a graph of the area we wish to calculate:

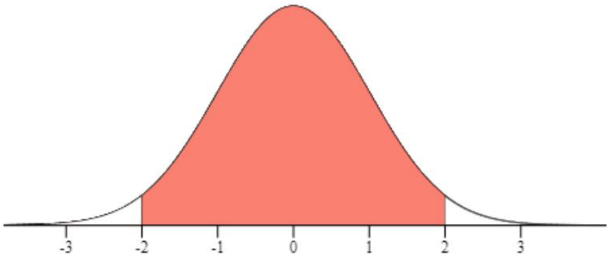


One strategy is to calculate the area to the left of the larger z score, then subtract the area to the left of the smaller z score. The remaining proportion/area will be to the left of the larger z score and to the right of the smaller z score; that is, it will be between the two z scores.

In the second example, we found the proportion to the left of 1.87 is .9693.
In the first example, we found the proportion to the left of -2.14 is .0162.
Subtracting, we obtain $P(-2.14 < z < 1.87) = .9693 - .0162 = .9531$ or 95.31%.

Example. Show that 95.44% of the data lies within 2 standard deviations of the mean, as we previously stated.

Solution. Since the mean is 0 and the standard deviation is 1 for the standard normal distribution, the *z* values that are within 2 standard deviations of the mean are the values between –2 and 2. So we are asked to calculate P(–2 < *z* < 2). Here is the area we need:



Using Table A, we find:

P(*z* < 2) = 0.9772
P(*z* < –2) = 0.0228
So P(–2 < z < 2) = 0.9772 – 0.0228 = 0.9544, or 95.44%

Exercise 3: We also stated earlier that 95% of the data lies for a normal distribution lies within 1.96 standard deviations of the mean, that is P(–1.96 < z < 1.96) = 0.95. Verify this using Table A.

Exercise 4: Calculate the following for the standard normal distribution.

a. The area to the left of *z* = –0.63.
b. The area to the right of *z* = –0.63
c. P(*z* > 1.27).
d. P(1.2 < *z* < 2.3).

The app at the following link provides additional practice using Table A to calculate probabilities for the standard normal distribution.

Practice using Table A

Using Table A, part 2 – other normal distributions

We begin with a quick review of the concept of a *z*-score, first described in Lesson 2. This is a paraphrase of how that lesson described the topic:

> For mound-shaped distribution, knowing how many standard deviations separate a piece of data from the mean gives a good indication of approximately where the data lies. For example, data which is more than two standard deviations away from the mean is either in the smallest 2½% or the largest 2½% of the data, since approximately 95% of the data is closer than two standard deviations. Because of this, it is useful to calculate the separation between a piece of data and the mean, in terms of standard deviations. The result of this calculation is called the *z*-score for the data. The *z*-score shows not only how far away from the mean the data lies, but also in which direction. The score is positive if the data is larger than the mean, negative if the data is smaller than the mean). So, for example:

A *z*-score of 3 indicates the data is 3 standard deviations above the mean.
A *z*-score of –2 indicates the data is 2 standard deviations below the mean.

The calculation for the *z*-score is simple: 1) calculate the distance from the data item to the mean, and 2) find how many standard deviations this difference represents, by dividing by the standard deviation. Using *x* to stand for the data item, we have:

$$z = \frac{x - mean}{standard\ deviation}$$

The key idea, then, will be to translate a statement about a probability for a variable (described as *x* in the formula) into a statement about that variable's *z* score. Here is a simple example. We illustrate the process in the following examples that deal with adult female heights, assumed to be normal with mean 65" and standard deviation 3.5"

Example. What proportion of adult females are shorter than 60"? That is, calculate $P(height < 60)$**.**

Solution. The first step is to calculate the *z* score for a height of 60", $z = \frac{60-65}{3.5} = -1.43$. Now since the *z* score for height 60 is –1.43, the following statements are equivalent:

- A particular person's height is less than 60"
- A particular person's *z* score for their height is less than –1.43

That is, we can covert the desired result, $P(height < 60)$ to the equivalent result, $P(z < -1.43)$. Using Table A, we see that the answer is 0.0764.

Likewise, in each of the following examples, the first step is to convert the given height or heights to corresponding *z* scores.

Example. Calculate $P(height > 71)$.

Solution: $z = \frac{71-65}{3.5} = 1.71$, so $P(height > 71)$ is equivalent to $P(z > 1.71)$. From Table A, the area to the left of this *z* score is 0.9564. So the area to the right is $1 - 0.9564 = 0.0436$.

Example. Calculate$P(62 < height < 70)$.

Solution. We calculate *z* scores for 62 and for 70, obtaining –0.86 and 1.43. This converts our desired calculation to the equivalent $P(-0.86 < z < 1.43)$. From Table A,

$P(z < -0.86) = 0.1949$
$P(z < 1.43) = 0.9236$

Therefore, $P(-.86 < z < 1.43) = .9236 - .1949 = .7287$.

The app at the following link provides further practice using Table A to solve probability problems for various normal distributions. In each problem, the first step will be calculating the corresponding *z* score for the data item involved.

Practice using Table A, part 2

4.4 – Using Technology to Calculate Probabilities for Normal Distributions

In the previous section, you learned how to use Table A to calculate various probabilities, and you learned that probabilities correspond to areas. In this section, we indicate how you can use the online calculator provided by the author of these lessons to assist in the calculations. Here again is a link to that calculator:

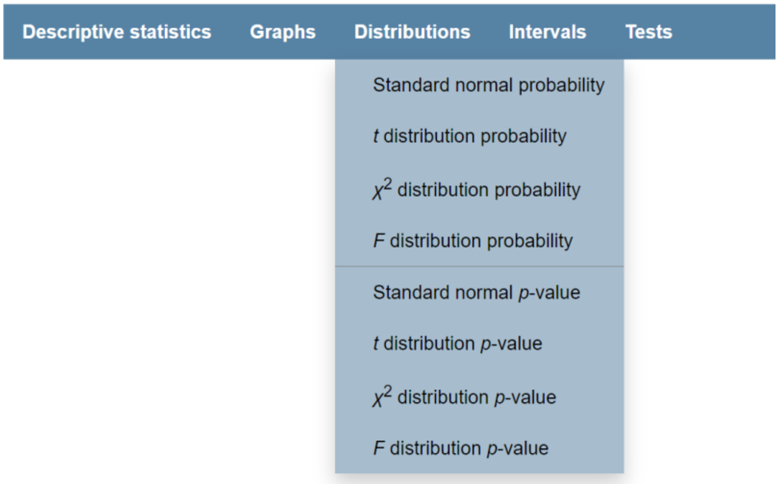
Statistical calculator

Calculating probabilities for the standard normal distribution

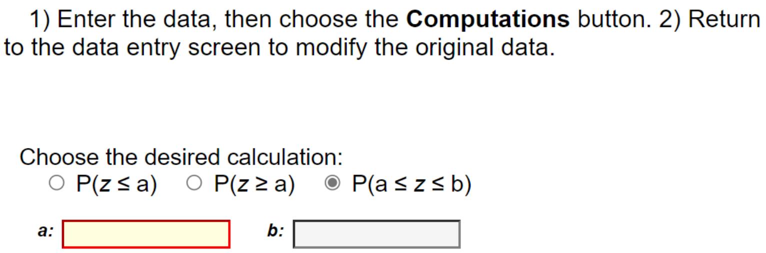
We begin by calculating probabilities based on *z* scores, using the problems from the first three examples:

$P(z < -2.14)$, or equivalently $P(z \leq -2.14)$
$P(z \geq 1.87)$, or equivalently $P(z > 1.87)$
$P(-2.14 \leq z \leq 1.87)$, or equivalently $P(-2.14 < z < 1.87)$

To solve these problems, using menu option *Distributions*, generating the following submenu:



We will use submenu option *Standard normal probability* for all our work in this section of the lesson. Choosing that option generates this screen:



1) Enter the data, then choose the **Computations** button. 2) Return to the data entry screen to modify the original data.

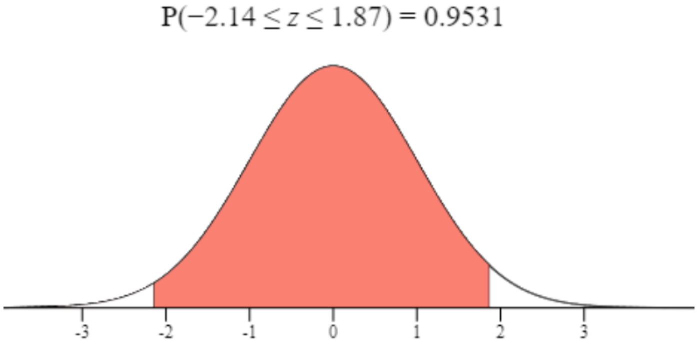
Choose the desired calculation:
○ P(z ≤ a) ○ P(z ≥ a) ● P(a ≤ z ≤ b)

a: ***b:***

The radio button allows us to choose which area we want – to the left of a particular value, to the right of a particular value, or between two values. Because the default is the latter, we begin with that example.

Example. Calculate $P(-2.14 \leq z \leq 1.87)$

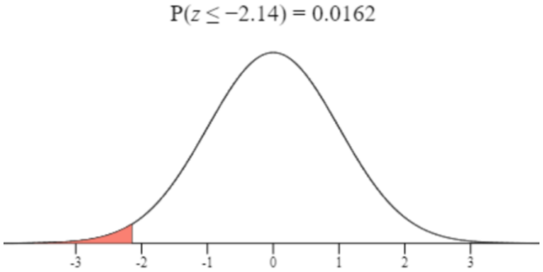
Solution: Simply enter –2.14 for a and 1.87 for b, then click computations, obtaining this result:



In addition to the desired probability (0.9531 or 95.31%) the calculator provides a graph of the corresponding area. Observe this is the same answer we obtained using Table A.

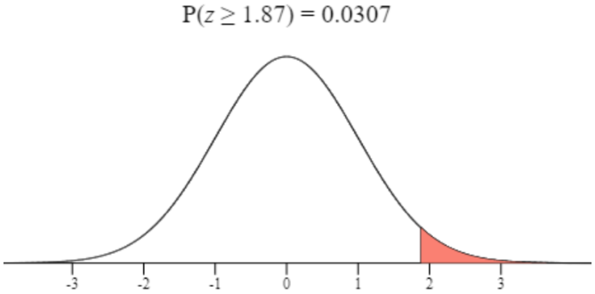
Example. Calculate $P(z \leq -2.14)$

Solution. Click on the radio button labeled $P(z \leq a)$, then enter –2.14 for a and click computations. Here is the resulting calculation with corresponding graph of the area.



Example. Calculate $P(z \geq 1.87)$

Solution. Click on the radio button labeled $P(z \geq a)$, enter 1.87 for a, and click computations, obtaining:



Calculating probabilities for other normal distributions

In practice, the probabilities we wish to calculate arise from application areas, where the distribution is (at least approximately) normal but the mean and standard deviation are not 0 and 1, respectively, as they are for the standard normal distribution. Just as when we use Table A for the calculation, when we use the provided calculator the first step is to calculate the *z* score.

Comment. When we were using Table A, we rounded our *z* scores to two decimal places. Why? Because Table A calculated areas to the left of *z* scores given as numbers with two decimal places. The online calculator has no such restriction; however, in these lessons we routinely round intermediate results to four places. Our final answers will be somewhat more accurate than those obtained using only two places. We could of course enter our numbers without rounding beyond what the calculation of the *z* score yields, but this would not improve our accuracy significantly.

For example, for the first example below our calculator shows *z* to be –1.428571429. Entering *z* rounded to two places gives 0.0764 as the answer; to four places gives the slightly more accurate 0.0766; entering the entire value –1.428571429 also gives 0.0766.

Examples. For adult female heights (normal with mean 65” and standard deviation 3.5”), calculate:

$P(height \leq 60)$
$P(height \geq 71)$
$P(62 \leq height \leq 70)$

Solution. Just as we did when using Table A, the first step in each problem is to calculate the *z* score for the heights. This converts the probability about heights to a probability about z scores, which we solve in the calculator just as we did for the previous three examples. The table below shows the results.

Original	**In terms of *z***	**Resulting probability**
$P(height \leq 60)$	$P(z \leq -1.4286)$	0.0766 = 7.66%
$P(height \geq 71)$	$P(z \geq 1.7143)$	0.0432 = 4.32%
$P(62 \leq height \leq 70)$	$P(-0.8571 \leq z \leq 1.4286)$	0.7277 = 72.77%

The app at the following link provides additional practice using the calculator for a variety of normal distributions. When you calculate the *z* scores, round to four decimal places.

Probabilities for normal distributions

4.5 – More About the Normal Distribution

Building intervals that contain a certain proportion of the data

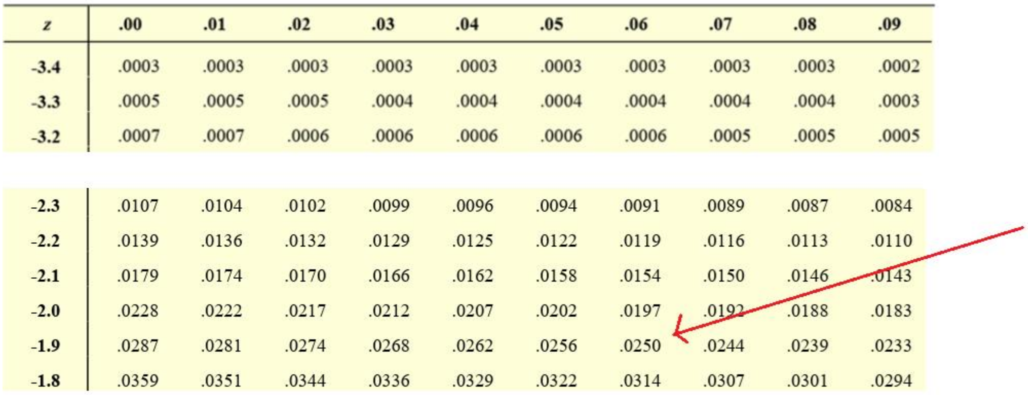
It is fairly easy, given a particular *z*-score (let’s call it *z**), to calculate what proportion of the data has a *z*-score between – *z** and + *z**. For example, consider *z** = 2. From Table A we can gather these facts:

1. Area to the left of –2 is 0.0228
2. Area to the left of +2 is 0.9772

From these facts, there are at least two ways to calculate the area between –2 and +2. First, we can subtract 0.9772 – 0.0228 = 0.9544. Second, we can use either fact #1 (and symmetry) or fact #2 to determine that the area to the right of +2 is also 0.0228; so the area *not* between –2 and + 2 is 0.0228 + 0.0228 = 0.0456, and the area between –2 and +2 is 1 – 0.0456 = 0.9544.

A more complicated, but ultimately more important, question is this: given a percentage, how can we determine a *z** with the property that that percentage of the data lies between – *z** and *z**? For example, what must *z** be if we want 95% of the data to lie between – *z** and *z**? To answer this, we take a clue from the second strategy used in the previous calculation. If we want 95% to lie between, then 5% must *not* be between, and of this amount 2.5% will be to the left of – *z**. Since 2.5% is 0.0250, we look in Table A for a *z*-score with an entry of 0.0250. As shown here, –1.96 has that property. That is how we determined that 95% of the data has a *z*-score between –1.96 and +1.96.

Table A Standard Normal Cumulative Probabilities



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294

Now let's take this one step further. It turns out that adult female heights are approximately normal, with mean 65 inches and standard deviation 3.5 inches. Fill in the blanks: 95% of adult females are between _____ inches tall and _______ inches tall. To do this, we must convert the *z* scores –1.96 and 1.96 into inches. We learned how to do this in Lesson 2, and in fact we observed a general pattern *x* = mean + *z* times standard deviation. So we can say that 95% of adult female heights are between ___65 – 1.96(3.5)___ inches and ___65 + 1.96(3.5)___ inches, that is between ___58.14___ inches and ___71.86___ inches.

Exercise 5: Fill in the blanks relating to a normal distribution, and specifically to adult female heights.

e. ______ % of the data has a *z*-score between –1.8 and 1.8.
f. 90% of the data has a *z*-score between _____ and _____.
g. 90% of the adult women are between _____ inches and _____ inches tall.

"Unusual" observations

Lesson 2 introduced you to the concept of "unusual" observation. (We use the quotes because this is not a technical, widely-used term. However, we will shortly describe a more technical way to describe this idea, at which point we will drop the quotes.)

Following is the discussion, and an exercise, adapted from Lesson 2.

When you see an adult male walk into the room, you can instinctively judge his height, perhaps identifying him as "about average height" or as "unusually tall" or perhaps as "unusually short." Using the Empirical Rule, we can quantify this idea, not only for heights but for any variable whose distribution is approximately mound-shaped. Of course, "unusual" is a vague notion, but in the framework of statistics one possible way to think of it is this: *Let us agree, for the current discussion, that the middle 95% of the data is* ***not*** *unusual*. In other words, anything within two standard deviations of the mean is not unusual; anything *not* within two standard deviations *is* unusual. This exercise gives you some practice thinking about these ideas.

Exercise 6: Adult male heights are mound-shaped, with a mean of 70 inches and a standard deviation of 4 inches. Use the empirical rule to fill in the blanks.

a. Approximately 95% are between ________ inches and ________ inches tall. Sketch a picture.
b. Anyone taller than 78 inches is in the tallest _____________%.
c. Anyone shorter than _____________ inches is in the shortest 2.5%.
d. If by "unusual" we mean unusually far away from the average of 70 inches, then both short people and tall people qualify as unusual. The 5% of most unusual people are either shorter than _________ inches or taller than ____________ inches.
e. If Sam is 79 inches tall, he is unusually tall because he falls in the top _________%. If Joe is 61 inches tall, he is unusually short because he falls in the bottom __________ %. Both these people are unusual; they are in the rarest ____________% of all adult male heights.
f. Sam is 79 inches tall and Bill is 81 inches tall. Both fall in the top 2.5% of heights using the empirical rule, but whose height would you say is more unusual, Sam's or Bill's?
g. Joe is 61 inches tall and Ted is 58.5 inches tall. Both fall in the bottom 2.5% of heights using the empirical rule, but whose height would you say is more unusual, Joe's or Ted's?

The previous exercise used the empirical rule to answer the questions. However, since adult male heights are actually not just "mound-shaped" but in fact approximately normal, we can use what we learned about normal distributions to provide more precise answers to the questions. For example:

Example. Rework part (b) of the exercise: Anyone taller than 78 inches is in the tallest ______%.

Solution. Using the empirical rule, we answered "2.5%." Using normality, we can calculate the *z*-score for a height of 78 inches, $z = 2$. Then we use Table A, which tells us that the area to the left of $z = 2$ is 0.9772, leaving an area of 0.0228 to the right. Our more precise answer is "2.28%."

Example. Rework part (e) of the exercise: If Sam is 79 inches tall, he is unusually tall because he falls in the top _________%. If Joe is 61 inches tall, he is unusually short because he falls in the bottom __________ %. Both these people are unusual; they are in the rarest ____________% of all adult male heights.

Solution. Using the empirical rule, we answered "2.5%", "2.5%", and "5%." (Sam's 79 inch height was to the right of the 78 inches that was 2 standard deviations above the mean, and so on.) With our current knowledge, we can do better than this. Using *z*-scores and Table A, we calculate $z = 2.25$ for Sam and $z = -2.25$ for Joe. Table A gives 0.9878 for $z = 2.25$, which means that an area of 0.0122 is to the right of $z = 2.25$. Sam is in the top 1.22%. Table A gives 0.0122 to the left of $z = -2.25$, so Joe is in the bottom 1.22%. Both these people are unusual, they are in the rarest 2.44% of all adult male heights (obtained by doubling the 1.22%).

Example. Rework part (f): Sam is 79 inches tall and Bill is 81 inches tall. Both fall in the top 2.5% of heights using the empirical rule, but whose height would you say is more unusual, Sam's or Bill's?

Solution. Obviously Bill's height is more unusual, but our new methods allow us to quantify this statement. Sam is in the tallest 1.22%. Using Table A in a similar way for Bill's *z* score of 2.75, we see that Bill is in the tallest 0.3%.

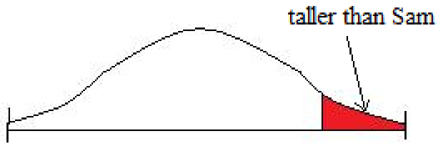
Example. Rework part (g): Joe is 61 inches tall and Ted is 58.5 inches tall. Both fall in the bottom 2.5% of heights using the empirical rule, but whose height would you say is more unusual, Joe's or Ted's?

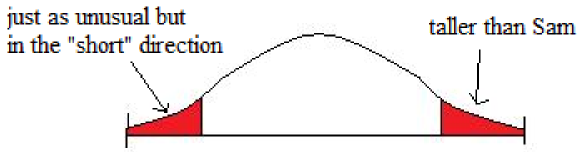
Solution. Similarly, Joe is in the shortest 1.22%, but Ted's *z*-score of –2.88 (rounded to two places) shows that Ted is in the shortest 0.2%.

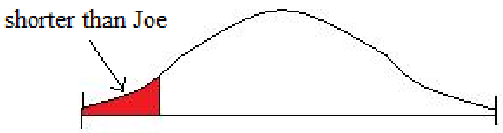
Note: In part (e) we thought of "unusual" in two ways. Sam is unusually tall, and Joe is unusually short. If we concentrate only on "tall," Sam is in the rarest 1.22%. If we concentrate only on "short," Joe is in the rarest 1.22%. But if we just think of unusual heights in general, where both tall and short are unusual, they are both in the rarest 2.44%.

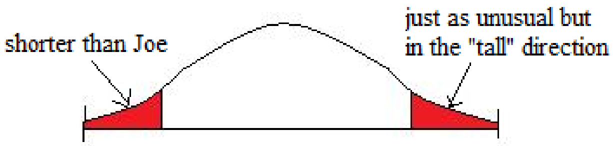
Similarly, measured in terms of "tall" Bill is in the rarest 0.3%; but just in terms of "unusual height" he is in the rarest 0.6%, because there are also short people who are just as far away from the mean as he is. Likewise, in terms of "short" Ted is in the rarest 0.2%, but he is in the rarest 0.4% in terms of "unusual height."

The following pictures illustrate these ideas.









The P-value

The calculations we have been doing will be extremely important as we proceed in the course, so much so that the results have been given a special name. Many texts use the notation "P-value." Other ways to write it are "*p*-value" or sometimes simply "*p*." The P-value is a measure of just how unusual the data is, in terms of what percentage of the data is even more unusual than the given data. It is customary to write the P-value as a decimal fraction (for example, 0.0122 rather than 1.22%).

To distinguish between the notions of "unusually tall" and the more generic "unusual height" we will speak of a one-tail P-value and a two-tail P-value. For example, for Sam, who was unusually tall, the one-tail P-value is 0.0122. There are only 1.22% of adult males taller than Sam. The two-tail P-value is 0.0244, because in addition to the 1.22% who are taller than Sam, there are an additional 1.22% who are further away from the mean but in the "short" direction. The picture above captures the distinction, and also illustrates the source of the "one-tail" and "two-tail" terminology. A similar analysis can be applied to Joe, but because Joe is shorter than average we think of the one-tail P-value in terms of unusually short, using the left tail as illustrated in the picture above.

Exercise 7: Use similar reasoning, along with the calculations we have already done, to give the one-tail and two-tail P-values for:

a. Joe
b. Bill
c. Ted

Comment: Observe that the more unusual the data, the smaller the P-value. Visually, this is true because the P-value measures the area even further out in the tail or tails than the given piece of data. The more unusual the data, the smaller that area is. This is important enough to state it again:

Small P-values indicate unusual data items.

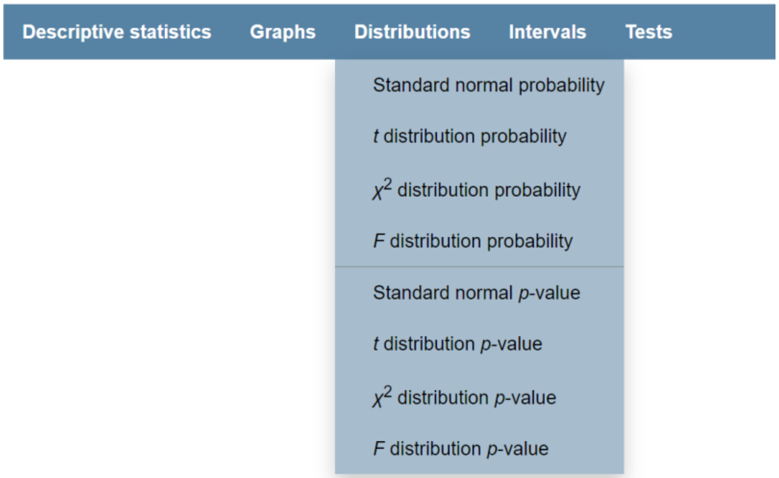
The smaller the P-value, the more unusual the data item.

Exercise 8: Find the one-tail and two-tail P-values for these *z*-scores. (For positive *z*-scores, the one-tail P-value to calculate is the right tail; for negative *z*-scores, the left tail.)

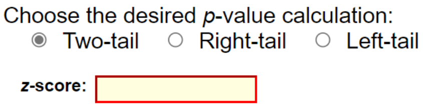
a. $z = 2.03$
b. $z = 1.27$
c. $z = –2.65$
d. $z = –0.17$

Using technology to calculate P-values

If you worked Exercise 8, you realize that calculating a P-value using Table A is not significantly more difficult than the calculations we did earlier in this lesson. The only added difficulty is realizing that a P-value is a probability. For example, consider exercise 8(a). We want the right tail P-value for $z = 2.03$. This is the same as $P(z \geq 2.03)$; and the two-tail P-value can be found as $P(z \leq -2.03) + P(z \geq 2.03)$, or as simply twice the one-tail value: $2 \cdot P(z \geq 2.03)$. However, the calculation of P-values will be so important in the second half of this course that the author has provided options specifically for these calculations. If we choose the *Distributions* menu option, it brings up this sub-menu:


Descriptive statistics
Graphs
Distributions
Intervals
Tests
Standard normal probability
t distribution probability
χ^2 distribution probability
F distribution probability
Standard normal *p*-value
t distribution *p*-value
χ^2 distribution *p*-value
F distribution *p*-value

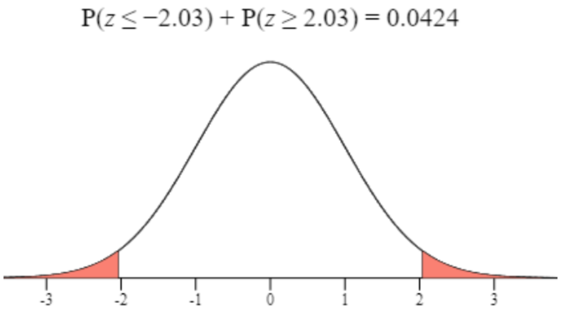
Choosing the submenu option *Standard normal p-value* yields this screen:



Choose the desired *p*-value calculation:
● Two-tail ○ Right-tail ○ Left-tail

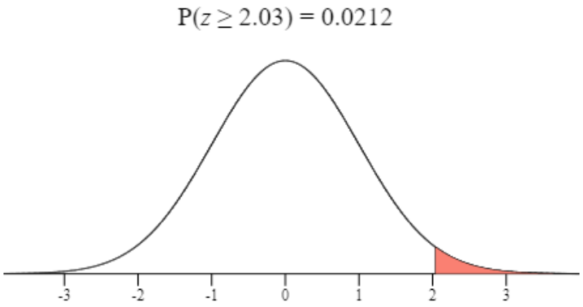
***z*-score:**

Entering 2.03 for *z* and clicking *Computations* gives the answer along with a graph illustrating that answer:



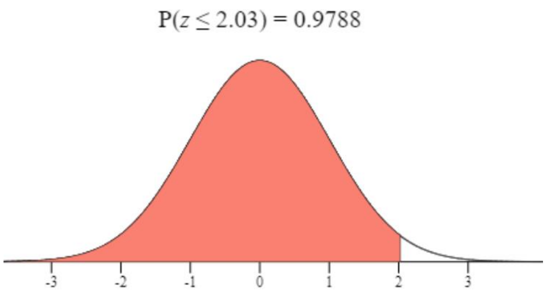
$P(z \leq -2.03) + P(z \geq 2.03) = 0.0424$

Choosing the Right-tail radio button similarly yields just the right-tail P-value:



$P(z \geq 2.03) = 0.0212$

Comment. Since P-values are generally used to measure how far out into the tail a value lies, for positive *z* scores we generally calculate right-tail P-values. However, we can calculate a left-tail P-value as shown here:



How should we interpret this result? Well, for example, if we are talking about heights a left tail P-value measure how unusual that height is in the "short person" direction. A person whose height has a *z* score of positive 2.03 is not at all unusually short – fully 97.88% of the people are shorter than that person.

Although it is certainly possible to do this calculation, in general we do not. If the question is about unusual shortness and we have a person whose *z* score is positive, we know they are not unusually short – they are in the top half of the population in terms of height. We could get the exact P-value, but generally we do not.

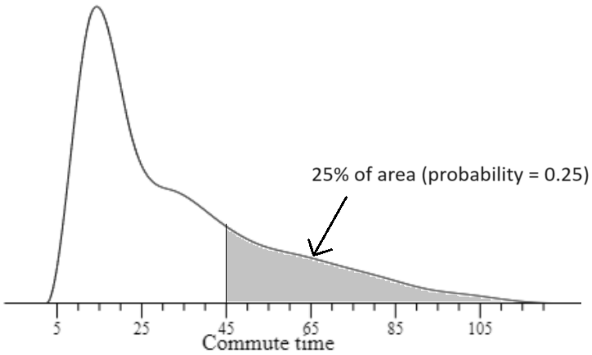
Comment. Because of this last observation, we will take the approach that if a one-tail P-value is asked for, unless specified otherwise, we will take it to mean a right-tail P-value for positive *z* scores and a left-tail P-value for negative *z* scores.

The app at the following link gives practice calculating P-values for a variety of normal distributions. As usual, the first step is calculating the *z* score – rounded to two places if you plan to use Table A, or to four places if you are using the author's calculator.

Normal distribution P-values

4.6 – Working with Other Continuous Distributions

Later in the course, we will be working extensively with other continuous distributions. Some, like the normal distribution, are symmetric. The notion of unusual observation and the corresponding concept of P-value are entirely similar for this type of distribution, although of course we cannot use Table A to determine the P-value. Other distributions of importance are skewed to the right, similar to this distribution we have already encountered for commuting time:



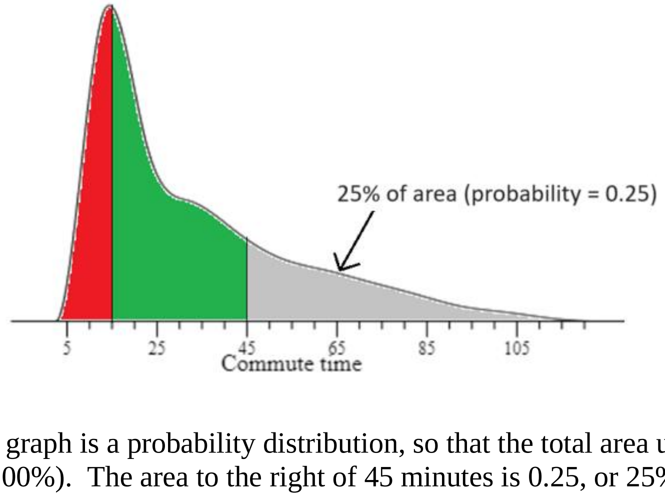
For these distributions, it turns out that we are most interested in a one-tail (right tail) P-value. In the context of this graph, this means that we are focusing on unusually long commuting times, and are not interested at all in unusually short commuting times.

For example, the commuting time of 45 minutes has a P-value of 0.25, because 25% of the area is to the right of this data value. A time of 45 minutes is somewhat unusual, but not as unusual as a time whose P-value would be 0.05, for example.

Exercise 9: The probability of a commute less than 15 minutes is 0.22. Find the (right-tail) P-value for a commute of 15 minutes.

Solutions to Exercises

Exercise 1: Here is the smooth curve representing a histogram for the commuting time of a group of individuals:



The graph is a probability distribution, so that the total area under the graph is 1 (that is, 100%). The area to the right of 45 minutes is 0.25, or 25% of the total area. This means that 25% of the people in the survey have commutes longer than 45 minutes.

Equivalently, the probability of randomly selecting an individual with a commute longer than 45 minutes is 0.25.

a. What is the probability for a commute under 45 minutes? 1 - 0.25 = 0.75.

b. The researchers report that the probability of a commute less than 15 minutes is 0.22. Shade the area of the graph that corresponds to this probability. See the graph above, the area shaded in red.

c. What is the probability for a commute between 15 and 45 minutes? See the graph above, the area shaded in green. One way to calculate that area is to subtract the area to the left of 15 (the red part) from the area to the left of 45 (the red part plus the green part). This gives 0.75 – 0.22 = 0.53.

Exercise 2: Give an interpretation for the other highlighted entry, circled in red.

The area to the left of – 3.07 is 0.0011

0.11% of the area lies to the left of the *z* value – 3.07

$P(z < -3.07) = 0.0011$

Exercise 3: We also stated earlier that 95% of the data lies for a normal distribution lies within 1.96 standard deviations of the mean, that is $P(-1.96 < z < 1.96) = 0.95$. Verify this using Table A.

$P(z < 1.96) = 0.9750$

$P(z < -1.96) = 0.0250$

$P(-1.96 < z < 1.96) = 0.9750 - 0.0250 = 0.9500$

Exercise 4: Calculate the following for the standard normal distribution.

a. The area to the left of $z = -0.63$. 0.2643

b. The area to the right of $z = -0.63$ 1 – 0.2643 = 0.7357

c. $P(z > 1.27)$. 1 – 0.8980 = 0.1020

d. $P(1.2 < z < 2.3)$. 0.9893 – 0.8849 = 0.1044

Exercise 5: Fill in the blanks relating to a normal distribution, and specifically to adult female heights.

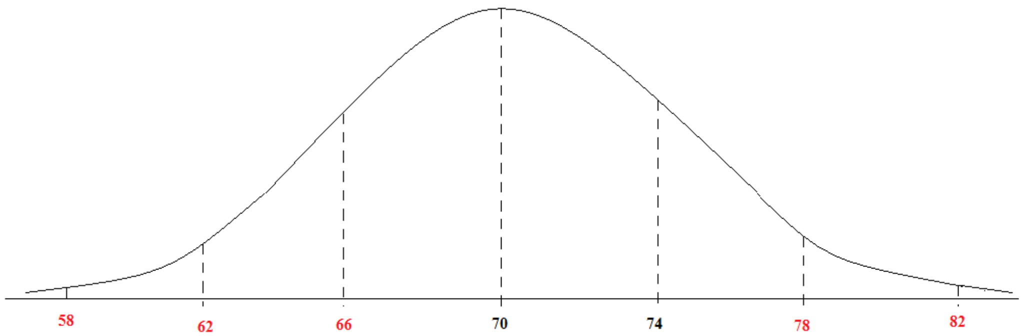
a. __92.82__% of the data has a *z*-score between –1.8 and 1.8.

b. 98% of the data has a *z*-score between __–2.33__ and __2.33__. (We look for 0.0100 in the table; the closest we find is 0.0099, so we use that.)

c. 98% of the adult women are between __65 – 2.33(3.5) = 56.845__ inches and __65 + 2.33(3.5) = 73.155__ inches tall.

Exercise 6: Adult male heights are mound-shaped, with a mean of 70 inches and a standard deviation of 4 inches. Use the empirical rule to fill in the blanks.

a. Approximately 95% are between __62__ inches and __78__ inches tall. Sketch a picture.



b. Anyone taller than 78 inches is in the tallest __2.5__ %.

c. Anyone shorter than __62__ inches is in the shortest 2.5%.

d. If by "unusual" we mean unusually far away from the average of 70 inches, then both short people and tall people qualify as unusual. The 5% of most unusual people are either shorter than __62__ inches or taller than __78__ inches.

e. If Sam is 79 inches tall, he is unusually tall because he falls in the top __2.5__ %. If Joe is 61 inches tall, he is unusually short because he falls in the bottom __2.5__ %. Both these people are unusual; they are in the rarest __5__ % of all adult male heights.

f. Sam is 79 inches tall and Bill is 81 inches tall. Both fall in the top 2.5% of heights using the empirical rule, but whose height would you say is more unusual, Sam's or Bill's? Bill's

g. Joe is 61 inches tall and Ted is 58.5 inches tall. Both fall in the bottom 2.5% of heights using the empirical rule, but whose height would you say is more unusual, Joe's or Ted's? Ted's

Exercise 7: Use similar reasoning, along with the calculations we have already done, to give the one-tail and two-tail P-values for:

a. Joe one-tail 0.0122, two-tail 0.0244 (in shortest 1.22%)

b. Bill one-tail 0.0030, two-tail 0.0060 (in tallest 0.3%)

c. Ted one-tail 0.0020, two-tail 0.0040 (in shortest 0.2%)

Exercise 8: Find the one-tail and two-tail P-values for these z-scores. (For positive z-scores, the one-tail P-value to calculate is the "upper" tail; for negative z-scores, the "lower" tail.)

a. $z = 2.03$ one-tail 0.0212, two-tail 0.0424

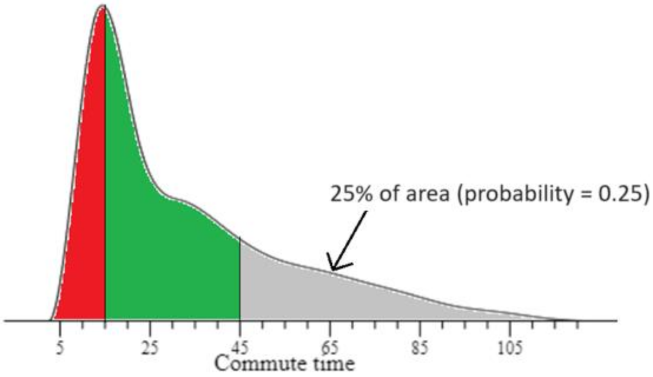
b. $z = 1.27$ one-tail 0.1020, two-tail 0.2040

c. $z = -2.65$ one-tail 0.0040, two-tail 0.0080

d. $z = -0.17$ one-tail 0.4325, two-tail 0.8650

Exercise 9: The probability of a commute less than 15 minutes is 0.22. Find the (right-tail) P-value for a commute of 15 minutes.

Here is the graph from the solution to Exercise 1.



The red area is 0.22, so the area to the right (green plus purple) is 0.78. The P-value is therefore 0.78. A commute time of 15 minutes is definitely not unusually long; a P-value of 0.78 is not at all small.