Chapter 4

Approximating a Sampling Distribution

At the end of the last chapter, we saw how tedious it is to find the sampling distribution of $U$ even when there are only 20 possible assignments. We also experienced the limit of my comfort zone: 252 possible assignments. For studies like Dawn’s (184,756 possible assignments) and especially Sara’s ($1.075 \times 10^{23}$ possible assignments) there are way too many possible assignments to seek an exact answer. Fortunately, there is an extremely simple way to obtain a good approximation—subject to the caveats given below—to a sampling distribution regardless of how large the number of possible assignments.

4.1 Two Computer Simulation Experiments

Let’s return to Dawn’s study. Our goal is to create a table for Dawn that is analogous to Table 3.7 on page 65 for Kynn’s study; i.e. we want to determine the value of $u = \bar{x} - \bar{y}$ for every one of the 184,756 possible assignments. This is too big of a job for me!

Instead of looking at all possible assignments, we look at some of them. We do this with a computer simulation experiment. I wrote a computer program that selected 10,000 assignments for Dawn’s study. For each selection, the program selected one assignment at random from the collection of 184,756 possible assignments. (You can visualize this as using our randomizer website 10,000 times.) For each of the 10,000 simulated assignments, I determined its value of $u = \bar{x} - \bar{y}$. My results are summarized in Table 4.1.

Suppose that we want to know $P(U = 0)$. By definition, it is the proportion of the (184,756) possible assignments that would yield $u = 0$. We do not know this proportion because we have not looked at all possible assignments. But we have—with the help of my computer—looked at 10,000 assignments; of these 10,000 assignments, 732 gave $u = 0$ (see Table 4.1). The relative frequency of $u = 0$ in the assignments we have examined is an intuitively obvious approximation to the relative frequency of $u = 0$ among all possible assignments. The relative frequency of $u = 0$ among all possible assignments is, by definition, $P(U = 0)$. To summarize, our approximation of the unknown $P(U = 0)$ is the relative frequency of assignments that gave $u = 0$; which is, from our table, 0.0732.

The above argument for $P(U = 0)$ can be extended to $P(U = u)$ for any of the possible
values \( u \). But it will actually be more interesting to us to approximate \( P(U = u) \). In particular, recall that Dawn’s actual \( u \) was 2.2. As we will see in Chapter 5, we will be interested in one or more of the probabilities given in Table 4.2. Let me give you some details on how the answers in this table were obtained.

To obtain \( P(U \geq 2.2) \) we must sum the relative frequencies for the values 2.2, 2.4, 2.6, 2.8, 3.0 and 3.2. From Table 4.1, we obtain

\[
0.0093 + 0.0054 + 0.0023 + 0.0017 + 0.0008 = 0.0198.
\]

For \( P(U \leq 2.2) \) note that this value is \( 1 - \text{r.f.} (U > 2.2) = 1 - 0.0105 = 0.9895 \). Finally, \( P(|U| \geq 2.2) \) is the sum of two relative frequencies: \( (U \geq 2.2) \) and \( (U \leq -2.2) \). The first of these has been found to equal 0.0198. The second of these is

\[
0.0078 + 0.0045 + 0.0025 + 0.0016 + 0.0004 + 0.0003 + 0.0001 = 0.0173.
\]
Table 4.3: Selected probabilities of interest for Sara’s CRD and their approximations.

<table>
<thead>
<tr>
<th>Probability of Interest</th>
<th>Its Approximation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(U \geq 8.700)$</td>
<td>r.f. $(U \geq 8.700) = 0.0903$</td>
</tr>
<tr>
<td>$P(U \leq 8.700)$</td>
<td>r.f. $(U \leq 8.700) = 0.9107$</td>
</tr>
<tr>
<td>$P(</td>
<td>U</td>
</tr>
</tbody>
</table>

Adding these we get,

$$r.f. (U \geq 2.2) + r.f. (U \leq -2.2) = 0.0198 + 0.0173 = 0.0371.$$

Next, I performed a computer simulation experiment for Sara’s CRD. As stated earlier, there are more than $10^{23}$ different assignments for a balanced study with $n = 80$ total trials. Trying to enumerate all of these would be ridiculous, so we will use a computer simulation with 10,000 runs.

My simulation study yielded 723 distinct values of $u$! This is way too many to present in a table as I did for Dawn’s study. (The simulation study for Dawn, recall, yielded 35 distinct values for $u$ and that was unwieldy.)

Recall that for Sara’s data

$$\bar{x} = 106.875 \text{ and } \bar{y} = 98.175, \text{ giving } u = 8.700.$$

Table 4.3 presents information that will be needed in Chapter 5.

4.2 How Good are These Approximations?

Tables 4.2 and 4.3 present six unknown probabilities and their respective approximations based on simulation experiments with 10,000 runs. Lacking knowledge of the exact probabilities I cannot say exactly how good any of these approximations are. What I can say, however, is that each of them is very likely to be very close to the exact (unknown) probability it is approximating. How can I know this? Well, we will see how later in this course when we learn about confidence intervals, so you will need to be patient.

And, of course, the terms very likely and very close are quite vague. Here is what we will do for now. First, the expression very close will be replaced by a specific number, call it $h$, that is computed from our simulation results. In words, it is very likely that the simulation study approximation is within $h$ of its exact probability. (If this is confusing, see the numerical examples below.)

Next, the expression very likely will be replaced by nearly certain. I know what you are thinking: nearly certain also is vague, but I hope that you feel that it is somehow more encouraging than very likely. As we will learn later (for those who can’t stand the suspense!) nearly certain
corresponds to what we will call being 99.73% confident. (Alas, I really can’t tell you exactly what this means until we study confidence intervals.)

Here is what we do. Let \( m \) denote the number of runs in our simulation experiment; recall that \( m = 10,000 \) for our two studies. Let \( r \) denote any (unknown) probability that interests us. Let \( \hat{r} \) denote the relative frequency approximation of \( r \). For example, in Dawn’s study \( \hat{r} = 0.0198 \) is our approximation to the unknown \( r = P(U \geq 2.2) \). The nearly certain interval for \( r \) is given by the following:

\[
\hat{r} \pm 3 \sqrt{\frac{\hat{r}(1 - \hat{r})}{m}}
\]  

(4.1)

Another way to say this is that we are nearly certain that \( \hat{r} \) is within \( h \) of \( r \), where

\[
h = 3 \sqrt{\frac{\hat{r}(1 - \hat{r})}{m}}.
\]

We will evaluate this interval for each of our six approximations. You don’t need to verify these computations, but you will be asked to do similar computations for homework.

- For \( \hat{r} = 0.0198 \), the nearly certain interval for \( r \) is
  
  \[
  0.0198 \pm 3 \sqrt{\frac{0.0198(0.9802)}{10000}} = 0.0198 \pm 0.0042 = [0.0156, 0.0240].
  \]

- For \( \hat{r} = 0.9895 \), the nearly certain interval for \( r \) is
  
  \[
  0.9895 \pm 3 \sqrt{\frac{0.9895(0.0105)}{10000}} = 0.9895 \pm 0.0031 = [0.9864, 0.9926].
  \]

- For \( \hat{r} = 0.0371 \), the nearly certain interval for \( r \) is
  
  \[
  0.0371 \pm 3 \sqrt{\frac{0.0371(0.9629)}{10000}} = 0.0371 \pm 0.0057 = [0.0314, 0.0428].
  \]

- For \( \hat{r} = 0.0903 \), the nearly certain interval for \( r \) is
  
  \[
  0.0903 \pm 3 \sqrt{\frac{0.0903(0.9097)}{10000}} = 0.0903 \pm 0.0086 = [0.0817, 0.0989].
  \]

- For \( \hat{r} = 0.9107 \), the nearly certain interval for \( r \) is
  
  \[
  0.9107 \pm 3 \sqrt{\frac{0.9107(0.0893)}{10000}} = 0.9107 \pm 0.0086 = [0.9021, 0.9193].
  \]

- For \( \hat{r} = 0.1824 \), the nearly certain interval for \( r \) is
  
  \[
  0.1824 \pm 3 \sqrt{\frac{0.1824(0.8176)}{10000}} = 0.1824 \pm 0.0116 = [0.1708, 0.1940].
  \]
There is a surprising feature to Formula 4.1. The feature is hidden, so it is easy to overlook. The surprising feature is that the total number of possible assignments (184,756 for Dawn, more than $10^{23}$ for Sara) does not appear in the formula! All that matters is the number of runs, $m$, in the simulation experiment. In my experience, many people believe that precision is a function of the percentage of objects examined. For our current problem, the percentage of assignments examined does not matter; it’s the number of assignments examined that matters.

We can improve the precision of a nearly certain interval by making it narrower, which we can achieve by increasing the value of $m$. Personally, I think that the precision of the nearly certain interval with $m = 10,000$ runs is fine unless $\hat{r}$ is very close to 0. (See more on this below.) In order to show you why I feel this way, I did another simulation experiment for Dawn’s study, this time with $m = 90,000$ runs. Combining my two experiments I have $m = 100,000$ runs and I will use the results of the 100,000 runs to recalculate two of my nearly certain intervals for Dawn’s study.

As an example, suppose I am interested in $r = P(U \geq 2.2)$. Of the 100,000 runs, 1,879 assignments gave $u \geq 2.2$. Thus, the approximate probability is

$$\hat{r} = \frac{1879}{100000} = 0.01879.$$  

The nearly certain interval for the exact probability is

$$0.01879 \pm 3\sqrt{\frac{0.01879(0.98121)}{100000}} = 0.01879 \pm 0.00129 = [0.01750, 0.2008].$$  

Note that the value of $h$ for this interval is 0.00129, compared to $h = 0.0042$ for our simulation with 10,000 runs. Also, of the 100,000 runs, 1,792 assignments gave $u \leq -2.2$. Thus, the approximation for $P(|U| \geq 2.2)$ is

$$0.01879 + 0.01792 = 0.03671.$$  

The nearly certain interval for the exact probability is

$$0.03671 \pm 3\sqrt{\frac{0.03671(0.96329)}{100000}} = 0.03671 \pm 0.00178 = [0.03493, 0.03849].$$  

In my experience, the increase in precision does not justify running my computer 10 times longer than required for 10,000 runs. (These runs do take time and consume electricity!)

### 4.3 A Warning about Simulation Experiments

The six nearly certain intervals above for the six probabilities of interest suggest that with a simulation experiment consisting of 10,000 runs we can obtain a very precise approximation to an exact probability. There is a caveat I need to mention.

With a 10,000 run simulation experiment, if $\hat{r}$ is $\leq 0.0050$ or $\geq 0.9950$ then you should not compute the nearly certain interval. For other values of $m \geq 1000$, do not compute the nearly certain interval if $\hat{r}$ is $\leq 50/m$ or $\geq 1 - (50/m)$. If you ignore my directive and go ahead and
calculate the nearly certain interval, it might indeed contain the exact probability of interest. (And if it does, we won’t know it.) The difficulty is that we can no longer trust the modifier *nearly certain*. As an extreme example, suppose that \( \hat{r} = 0 \), which means that the event of interest never occurred in the simulation experiment. If you plug this value into Formula 4.1 the nearly certain interval becomes the single number 0. Whereas the probability of an event \( A \) can be very very small, as long as \( A \) contains at least one of the possible assignments, it can never be 0.

As we will learn later in these notes, if \( \hat{r} \) does equal 0, we can be nearly certain that \( r \leq 5.92/m \). There are other results for very small non-zero values of \( \hat{r} \) and we will learn how to find them later.

In many applications—though not all, as we shall see— if \( \hat{r} \leq 0.0050 \), then the researcher is happy with the conclusion that \( r \) is very small and is not very concerned with having a precise nearly certain interval.

### 4.4 Computing

I performed the simulation studies reported in this chapter by using the statistical software package **Minitab**. I could find no website that allows me to perform a *full-blown* simulation study, by which I mean a simulation that gives me all \( m \) (recall, usually \( m = 10,000 \)) values of the target statistic, in the current situation, \( u \). Later I will show you a website that does allow us to perform a simulation study, but it gives only two relative frequencies as output. It turns out that these are two very useful relative frequencies, so the website will be valuable to us.

I have one other comment about our simulation studies. Some of you may have been wondering about this issue. Note that I have *never claimed* that my simulation examines \( m \) different assignments. Indeed, each run of my simulation selects an assignment at random without caring about which assignments have already been examined. I do this for the following reasons.

1. A program that keeps track of the assignments already examined would be much more difficult to write, it would require more computer storage space and would require much more computer time to execute. (If you have any experience writing programs, you probably agree that these three claims are believable.)

2. The *possibility* that my simulation experiment might examine some assignments more than once creates no bias in my answers; it simply makes my answers a little bit less efficient; i.e., the nearly certain interval would change ever so slightly—becoming narrower—if I guaranteed \( m \) different assignments are selected.

3. I don’t want you to waste time worrying about the validity of item 2 immediately above; we will revisit this issue when we discuss population-based inference in *Part II* of these *Course Notes*.

4. In view of items 1 and 2 in this list, if you want a more precise approximation to an unknown probability, increasing the value of \( m \) is a much smarter choice than writing a new computer program.
4.5 Summary

In Chapter 3 we learned about the exact sampling distribution for the test statistic $U$. In Chapter 5 we will learn why we need this sampling distribution to analyze our data with a statistical test of hypotheses. In this current chapter we learned an important technique for approximating a sampling distribution: a computer simulation experiment.

The sampling distribution calculates the value of $u$ for all possible assignments. A computer simulation experiment calculates the value of $u$ for some possible assignments. The number of assignments examined by a computer simulation experiment is called its number of runs and is denoted by $m$. In these notes, $m$ will usually be 10,000; on some occasions it will equal 100,000; and on a few rare occasions it will equal 1,000. The assignment for any run is selected at random from the collection of all possible assignments, for example, by using the randomizer website. As will be discussed later in these notes, you should realize that a computer simulation experiment need not select $m$ distinct assignments.

After performing a computer simulation experiment, the analyst creates a listing of all values of $u$ that were obtained in the experiment and the relative frequency of occurrence of each value of $u$. An example of such a table is given in Table 4.1 on page 76 for Dawn’s study of her cat. We may use this table to approximate the exact and unknown $P(U = u)$ for all possible values $u$ of the test statistics. The approximation is simply the relative frequency of occurrence of $(U = u)$ in the computer simulation experiment.

In our applications of these ideas in Chapter 5 and later in these notes, we will usually be interested in events that are more complicated than simply $(U = u)$. To this end, let $A$ denote any event which involves the value of the test statistic $U$. Let $r$ denote the exact probability of the event $A$. Let $\hat{r}$ be the relative frequency of the event $A$ in the computer simulation experiment. We view $\hat{r}$ as our approximation of $r$.

For $m \geq 1000$, if $\hat{r}$ satisfies the following inequality

$$50/m < \hat{r} < 1 - (50/m),$$

then one can compute the nearly certain interval for $r$, given below:

$$\hat{r} \pm 3\sqrt{\frac{\hat{r}(1 - \hat{r})}{m}}.$$

We are nearly certain that this interval of numbers includes the true, unknown, value of $r$. An equivalent way to interpret this result is:

We are nearly certain that $\hat{r}$ is within $h$ of $r$, where

$$h = 3\sqrt{\frac{\hat{r}(1 - \hat{r})}{m}}.$$

The meaning of the modifier nearly certain will be explored later in these notes. Also, the situations in which $\hat{r} \leq 50/m$ or $\hat{r} \geq 1 - (50/m)$ will be considered later in these notes.
4.6 Practice Problem

1. In the dart game 301 the object is to be the first player to score exactly 301 points. In each round a player throws three darts and the total score of the three darts is added to the player’s score at the end of the previous round. If the new total is greater than 301, the player’s score reverts to the total at the end of the previous round. Thus, for example, if a player reaches a total of 300 at the end of a round, then the player will need exactly one point on the next round to win; any larger score will be ignored. Doug performed a balanced CRD with \( n = 40 \) trials to compare his personal darts (treatment 1) to bar darts (treatment 2). The response was the number of rounds Doug required to score exactly 301. The sorted responses with Doug’s darts are:

\[
12 \quad 13 \quad 14 \quad 14 \quad 15 \quad 15 \quad 17 \quad 18 \quad 18 \quad 19 \\
19 \quad 19 \quad 20 \quad 20 \quad 21 \quad 21 \quad 22 \quad 23 \quad 25 \quad 27
\]

The sorted responses with the bar darts are:

\[
13 \quad 15 \quad 16 \quad 16 \quad 17 \quad 17 \quad 17 \quad 18 \quad 19 \quad 21 \\
21 \quad 22 \quad 23 \quad 25 \quad 26 \quad 26 \quad 27 \quad 27 \quad 28 \quad 30
\]

I obtained the following summary statistics for Doug’s data:

\[
\bar{x} = 18.60 \text{ and } \bar{y} = 21.20, \text{ giving } u = 18.60 - 21.20 = -2.60.
\]

Note that smaller response values are preferred.

I performed a simulation experiment with 10,000 runs. Each run yielded a possible observed value \( u \) of the test statistic \( U \). I won’t show you all of my results, but I will tell you the following two relative frequencies:

\[
\text{r.f. } (U \leq -2.60) = 0.0426; \text{ and r.f. } (U \geq 2.60) = 0.0418.
\]

Use these results to answer the questions below.

(a) What is our approximation for \( r = P(U \leq -2.60) \)?

(b) Compute the nearly certain interval for \( r \) in (a).

(c) What is our approximation for \( r = P(|U| \geq 2.60) \)?

(d) Compute the nearly certain interval for \( r \) in (c).
4.7 Solution to Practice Problem

1. (a) The approximation is $\hat{r} = 0.0426$.
   (b) The nearly certain interval is
   
   $$0.0426 \pm 3 \sqrt{\frac{0.0426 \cdot (0.9574)}{10,000}} = 0.0426 \pm 0.0061 = [0.0365, 0.0487].$$

   (c) The approximation is
   
   $$\hat{r} = 0.0426 + 0.0418 = 0.0844.$$

   (d) The nearly certain interval is
   
   $$0.0844 \pm 3 \sqrt{\frac{0.0844 \cdot (0.9156)}{10,000}} = 0.0844 \pm 0.0083 = [0.0761, 0.0927].$$
4.8 Homework Problems

1. In the Chapter 1 Homework you learned about Reggie’s study of darts. Reggie performed a balanced CRD with a total of \( n = 30 \) trials. Dot plots of his data are presented in Figure 1.4. It can be shown that the means for Reggie’s data are:

\[
\bar{x} = 201.533 \text{ and } \bar{y} = 188, \text{ giving } u = 201.533 - 188 = 13.533.
\]

I performed a simulation experiment with 10,000 runs. Each run yielded a possible observed value \( u \) of the test statistic \( U \). I won’t show you all of my results, but I will tell you the following two relative frequencies:

\[
r.f. (U \geq 13.533) = 0.0058; \text{ and r.f. } (U \leq -13.533) = 0.0039.
\]

Use these results to answer the questions below.

(a) What is our approximation for \( r = P(U \geq 13.533) \)?

(b) Compute the nearly certain interval for \( r \) in (a).

(c) What is our approximation for \( r = P(|U| \geq 13.533) \)?

(d) Compute the nearly certain interval for \( r \) in (c).

2. In the Chapter 1 Homework you learned about Brian’s study of running. Brian performed a balanced CRD with a total of \( n = 20 \) trials. Dot plots of his data are presented in Figure 1.3. It can be shown that the means for Brian’s data are:

\[
\bar{x} = 333.0 \text{ and } \bar{y} = 319.5, \text{ giving } u = 333.0 - 319.5 = 13.5.
\]

I performed a simulation experiment with 100,000 runs. Each run yielded a possible observed value \( u \) of the test statistic \( U \). I won’t show you all of my results, but I will tell you the following two frequencies:

\[
\text{freq. } (U \geq 13.5) = 56; \text{ and freq. } (U \leq -13.5) = 45.
\]

Use these results to answer the questions below. Note that the number of runs is 100,000, not the usual 10,000.

(a) What is our approximation for \( r = P(U \geq 13.5) \)?

(b) Compute the nearly certain interval for \( r \) in (a).

(c) What is our approximation for \( r = P(|U| \geq 13.5) \)?

(d) Compute the nearly certain interval for \( r \) in (c).