

Chapter 3

Approximating a Sampling Distribution

Table 3.1: Heights of the rectangles in the probability histogram of the sampling distribution of the test statistic for Fisher’s test for the Ballerina study.

x	$P(X = x)$	Height of rectangle $P(X = x)/0.08$
-0.40	0.0009	0.01125
-0.32	0.0081	0.10125
-0.24	0.0387	0.48375
-0.16	0.1127	1.40875
-0.08	0.2104	2.63000
0.00	0.2584	3.23000
0.08	0.2104	2.63000
0.16	0.1127	1.40875
0.24	0.0387	0.48375
0.32	0.0081	0.10125
0.40	0.0009	0.01125
Total	1.0000	

3.1 Study Suggestions

Chapter 1 introduced the CRD as a device for conducting a comparative study of two treatments. Chapter 2 introduced hypothesis testing as a technique for deciding whether the treatments have an identical effect. A hypothesis test yields a number, the P-value, which quantifies the debate between the Skeptic and the Advocate. A practical problem has arisen, however; the P-value can be difficult to compute. In fact, if a study has a large number of subjects, the P-value can be impossible to compute, even with an electronic computer equipped with any of the popular existing statistical software packages. Thus, Chapter 3 addresses the problem of finding an easy way of obtaining an approximate P-value.

The two approximation methods introduced in Chapter 3 are much easier to understand if a person has learned to represent a sampling distribution with a picture—the probability histogram. Given a sampling distribution, make sure you can draw its probability histogram by following the three steps in the key extract on page 80 of the text. In particular, practice obtaining the height of a value’s rectangle by dividing the probability of the value by δ . (Remember, after sorting, or ordering, the possible values of the test statistic from smallest to largest, δ equals the difference between any two successive values.)

In addition, given a probability histogram, be certain that you can create its sampling distribution. The centers of the bases of the rectangles of the probability histogram correspond to the possible values of the test statistic, and the area of a rectangle equals the probability of its value (center).

Some examples of these techniques follow.

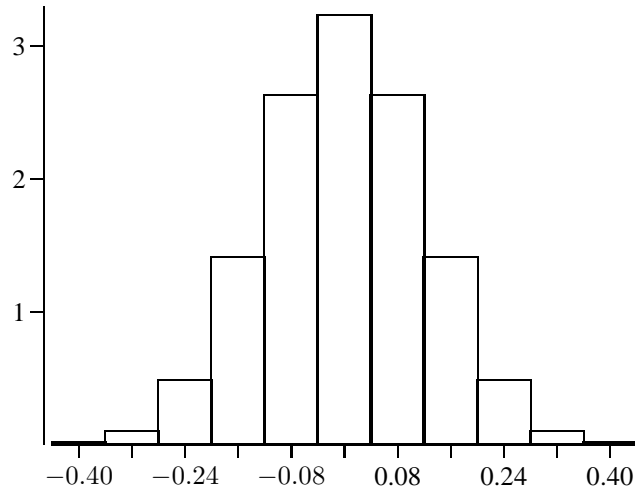
Consider the Ballerina study introduced in Chapter 1 of the text. The sampling distribution for the Ballerina study is presented in Table 3.1. In order to obtain its probability histogram, first we must determine the value of δ , the (constant) difference between any two of the ordered possible values of the test statistic. From Table 3.1, clearly $\delta = 0.08$. (Alternatively, you can remember that

$$\delta = n/(n_1 n_2).$$

For the current study, this formula gives $\delta = 50/[25(25)] = 0.08$.)

Second, we determine the height of each rectangle. This is tedious, but basically simple. Each rectangle is centered at a possible value of x ; the height of the rectangle is the probability of that value divided by δ . The heights are given in Table 3.1.

Figure 3.1: Probability histogram of the sampling distribution of the test statistic for Fisher's test for the Ballerina study.



You do not need to verify all the heights, but check enough to make sure you know how to compute them.

Finally, we draw the probability histogram, as shown in Figure 3.1.

Next, we consider the *reverse* of the above example. In particular, if you are given a probability histogram, the value of x , and the alternative, then you should be able to compute the P-value. For example, Figure 3.2 is the probability histogram for an unbalanced study. Note that the height of each rectangle is printed above it. We shall use this hypothetical example to illustrate several computations.

1. Suppose that $x = 0.25$ and the alternative is $p_1 > p_2$. The P-value equals

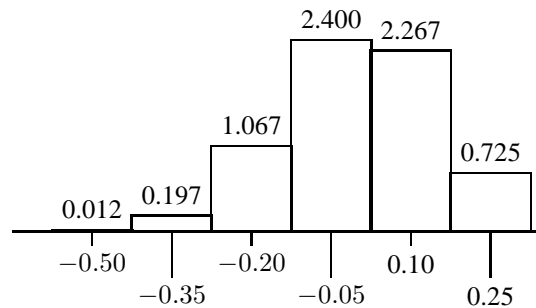
$$P(X \geq x) = P(X \geq 0.25).$$

But 0.25 is the largest possible value of x ; thus,

$$P(X \geq 0.25) = P(X = 0.25).$$

This latter probability equals the area of the rectangle centered at 0.25. Because the area of a rectangle is its base multiplied by its height,

Figure 3.2: Probability histogram of the sampling distribution of the test statistic for Fisher's test for an unnamed unbalanced study.



the P-value equals,

$$0.15(0.725) = 0.1088 \text{ (rounded).}$$

2. Suppose that $x = 0.10$ and the alternative is $p_1 > p_2$. The P-value equals $P(X \geq 0.10)$. This is the sum of two numbers: the areas of the rectangles centered at 0.10 and 0.25. Thus, the P-value equals,

$$0.15(2.267) + 0.1088 = 0.3400 + 0.1088 = 0.4488.$$

3. Suppose that $x = -0.20$ and the alternative is $p_1 < p_2$. The P-value equals $P(X \leq -0.20)$. This is the sum of three numbers: the areas of the rectangles centered at -0.20 , -0.35 , and -0.50 . Thus, the P-value equals,

$$0.15(1.067 + 0.197 + 0.012) = 0.15(1.276) = 0.1914.$$

The remainder of Chapter 3 focuses on two ways to approximate a sampling distribution, which will allow us to compute an approximate P-value. The first method is a simulation experiment.

Consider again the Ballerina study. It can be shown that there are over 126 trillion possible assignments of subjects (spins) to treatments (directions).

Table 3.2: Results of a simulation experiment with 10,000 runs for the Ballerina study.

x	Freq. of x	Rel. Freq. of x	$P(X = x)$
-0.40	9	0.0009	0.0009
-0.32	72	0.0072	0.0081
-0.24	383	0.0383	0.0387
-0.16	1137	0.1137	0.1127
-0.08	2169	0.2169	0.2104
0.00	2591	0.2591	0.2584
0.08	2022	0.2022	0.2104
0.16	1140	0.1140	0.1127
0.24	383	0.0383	0.0387
0.32	89	0.0089	0.0081
0.40	5	0.0005	0.0009
Total	10,000		

Whereas the exact sampling distribution is obtained by considering all the possible assignments, the idea behind a simulation experiment is that we can approximate the sampling distribution by looking at only some of the assignments. Examining 10,000 assignments usually gives an excellent approximation to the sampling distribution.

Each **run** of a simulation experiment selects an assignment at random from the collection of all possible assignments. Then the run computes the value of x that would be given by the selected assignment (remembering the assumption that the Skeptic is correct).

I performed a simulation experiment for the Ballerina study with 10,000 runs; my results are in Table 3.2. The first column of the table lists the different values of x that I obtained in my experiment. The second column presents the frequency of occurrence of each value of x . Because there are 10,000 runs, each frequency in the second column is divided by 10,000 to yield the relative frequencies of occurrence, which are presented in the third column of the table. The fourth column presents the exact probabilities. A careful inspection shows that each number in the third column is very close to the adjacent number in the fourth column; thus, the relative frequencies provide an excellent approximation to the sampling distribution.

Table 3.3: Results of a simulation experiment with 10,000 runs for the Ballerina study.

x	Rel. Freq. of x	Rel. Freq. of $\leq x$	Rel. Freq. of $\geq x$
-0.40	0.0009	0.0009	1.0000
-0.32	0.0072	0.0081	0.9991
-0.24	0.0383	0.0464	0.9919
-0.16	0.1137	0.1601	0.9536
-0.08	0.2169	0.3770	0.8399
0.00	0.2591	0.6361	0.6230
0.08	0.2022	0.8383	0.3639
0.16	0.1140	0.9523	0.1617
0.24	0.0383	0.9906	0.0477
0.32	0.0089	0.9995	0.0094
0.40	0.0005	1.0000	0.0005

It will be convenient to have cumulative sums of the relative frequencies; these are presented in Table 3.3. This table can be used to obtain an approximate P-value. Recall for the Ballerina study that $x = -0.24$ and Julie chose the second alternative. Thus, Julie's P-value is $P(X \leq -0.24)$ which can be approximated by the relative frequency of $x \leq -0.24$. My approximation of Julie's P-value is 0.0464, a good approximation of the exact P-value, 0.0477.

For a second example, Table 3.4 presents the results of a simulation experiment for the Crohn's study. Recall that $x = 0.27$ and we chose the first alternative. The exact P-value equals $P(X \geq 0.27)$ which can be approximated by the relative frequency of $x \geq 0.27$. My approximation of the P-value is 0.0193, an excellent approximation of the exact P-value, 0.0198.

Section 3.3 of the text introduces the important ideas of the center and spread of a sampling distribution. The four probability histograms on page 89 of the text have a single peak (one or two rectangles wide) and are symmetric or nearly symmetric. These histograms differ most notably in their amounts of spread. Of the myriad ways in which one *might* measure spread, statisticians choose the standard deviation because, as demonstrated in Section 3.4, it helps us solve our problem, namely, obtaining an approximation to the sampling distribution.

Table 3.4: Results of a simulation experiment with 10,000 runs for the Crohn's study.

x	Rel. Freq. of x	Rel. Freq. of $\leq x$	Rel. Freq. of $\geq x$
-0.46	0.0002	0.0002	1.0000
-0.41	0.0005	0.0007	0.9998
-0.35	0.0027	0.0034	0.9993
-0.29	0.0094	0.0128	0.9966
-0.24	0.0289	0.0417	0.9872
-0.18	0.0593	0.1010	0.9583
-0.12	0.1178	0.2188	0.8990
-0.07	0.1540	0.3728	0.7812
-0.01	0.1893	0.5621	0.6272
0.05	0.1724	0.7345	0.4379
0.10	0.1287	0.8632	0.2655
0.16	0.0830	0.9462	0.1368
0.21	0.0345	0.9807	0.0538
0.27	0.0143	0.9950	0.0193
0.33	0.0039	0.9989	0.0050
0.38	0.0010	0.9999	0.0011
0.44	0.0001	1.0000	0.0001

Throughout the book, test statistics and other random variables will be standardized. Let X be a random variable with population mean denoted by μ , and population standard deviation denoted by σ . The standardized version of X is denoted by Z , and is given by the equation

$$Z = \frac{X - \mu}{\sigma}.$$

If X is the test statistic for Fisher's test, the application of interest to us in Chapter 3, then $\mu = 0$, and the standardized version of X is

$$Z = \frac{X}{\sigma}.$$

Some of my students are confused by Z . Just remember that X is a rule that takes the outcome of a chance mechanism and turns it into a number, and Z is a rule that takes the number created by X and turns it into another number. For example, in the Infidelity study, the chance mechanism of assigning subjects to treatments yields (after responses are collected) the table below.

Version	S	F	Total
1	7	3	10
2	4	6	10
Total	11	9	20

The test statistic takes this table as input and yields the number

$$\hat{p}_1 - \hat{p}_2 = 0.70 - 0.40 = 0.30.$$

Since $\sigma = 0.2283$ for the Infidelity study, Z takes the number produced by X , 0.30, and uses it to obtain

$$0.30/0.2283 = 1.31.$$

As illustrated for the Infidelity study in Tables 2.2 and 2.3 on pages 93 and 94 of the text, knowing the sampling distribution for X is equivalent to knowing the sampling distribution for Z . Thus, the P-value, which in Chapter 2 is expressed in terms of X , can just as well be expressed in terms of Z . Carrying this process one step further, the P-value can be obtained from the probability histogram for Z , as illustrated on pages 94–96 of the text.

It is important to realize that at this point I have achieved nothing of value. Computing probabilities for Z is no easier than computing probabilities for X . The pictures on pages 94–96 of the text reveal, however, that the probability histograms of Z for the four studies considered look very similar. This similarity suggests that one picture might provide a good approximation to the sampling distribution of Z for any of these four studies, and perhaps for other studies as well. The one picture we use is the standard normal curve.

The two-step algorithm on page 102 of the text provides a simple and quick way to use the standard normal curve to obtain an approximate P-value for Fisher's test. Page 102 is a milestone in the text; it took us 102 pages to introduce and solve our first category of real problems. It was a difficult journey—for the number of ideas, their complexity, and the sheer length of the trip. You now have a solid foundation, however, for the remainder of the book, having been introduced to many of the important ideas and concepts of Statistics.

3.2 Solutions to Odd-Numbered Exercises

Solutions for Section 3.3

- $m_1 = 35, m_2 = 15, n_1 = 25, n_2 = 25,$ and $n = 50$; thus, $\sigma = 0.1309$.
- $m_1 = 50, m_2 = 15, n_1 = 34, n_2 = 31,$ and $n = 65$; thus, $\sigma = 0.1054$.
- $m_1 = 28, m_2 = 22, n_1 = 25, n_2 = 25,$ and $n = 50$; thus, $\sigma = 0.1418$.

Solutions for Section 3.4

- The values $x = 0.272$ and $\sigma = 0.1525$ yield $z = 0.272/0.1525 = 1.78$. The alternate formula yields

$$z = \frac{\sqrt{43}[14(14) - 8(8)]}{\sqrt{22(22)(22)(22)}} = 1.79.$$

(The two formulas will give the same answer, except for round-off error.) The approximate P-value for the third alternative is twice the area to the right of $|z|$; for $z = 1.79$ the area equals $2(0.0367) = 0.0734$; for $z = 1.78$ the area equals $2(0.0375) = 0.0750$.

- The values $x = 0.237$ and $\sigma = 0.1054$ yield $z = 0.237/0.1054 = 2.25$. The approximate P-value for the third alternative is twice the area to the right of 2.25 under the standard normal curve. This area equals $2(0.0122) = 0.0244$.
- The values $x = -0.255$ and $\sigma = 0.0904$ yield $z = -0.255/0.0904 = -2.82$. The approximate P-value for the second alternative is the area to the right of 2.82 under the standard normal curve. This area equals 0.0024.
- The values $x = 0.08$ and $\sigma = 0.1047$ yield $z = 0.08/0.1047 = 0.76$. The approximate P-value for the third alternative is twice the area to the right of 0.76 under the standard normal curve. This area is equal to $2(0.2236) = 0.4472$.
- The standardized version of the test statistic is

$$z = \frac{\sqrt{99}[36(2) - 14(48)]}{\sqrt{50(50)(84)(16)}} = -3.26.$$

The approximate P-value for the third alternative is twice the area to the right of 3.26 under the standard normal curve. This area is equal to $2(0.0006) = 0.0012$.

3.3 Exam Questions

- I performed a simulation experiment with 1,000 runs. Each run yielded an observed value of the test statistic for a Fisher's test. The results of the simulation experiment are in the table below.

Observed value of test statistic	Frequency
-0.48	1
-0.39	4
-0.30	25
-0.21	73
-0.12	186
-0.03	270
0.06	229
0.15	135
0.24	63
0.33	12
0.42	1
0.51	1
Total	1,000

Use the results of this simulation experiment to obtain an approximate P-value for the first alternative if $x = 0.33$.

- The sampling distribution of the test statistic for Fisher's test for an unbalanced CRD is given by the following table.

x	$P(X = x)$	$P(X \leq x)$	$P(X \geq x)$
-0.7	0.0186	0.0186	1.0000
-0.4	0.1632	0.1818	0.9814
-0.1	0.3916	0.5734	0.8182
0.2	0.3263	0.8998	0.4266
0.5	0.0932	0.9930	0.1002
0.8	0.0070	1.0000	0.0070

A researcher draws the probability histogram of this sampling distribution. What is the height of the rectangle centered at -0.4 ?

- Refer to the previous question. A 1000 run simulation experiment was conducted in which

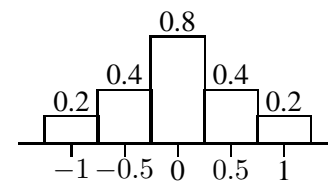
each run yielded a value of the test statistic for Fisher's test. The frequencies for the six different values of the test statistic are:

372 327 177 101 20 3

Match these frequencies with the six values of the test statistic. (Hint: The frequencies and test statistic values match exactly as one would expect based on the long-run relative frequency interpretation of probability.)

4. Figure 3.3 on page 85 of the text describes a single run of a simulation experiment for the Colloquium study. Use the information in the top box of the figure to determine (according to the Skeptic) the value of x if all odd-numbered subjects (1, 3, 5, ...) are assigned to treatment 1, and all even-numbered subjects (2, 4, 6, ...) are assigned to treatment 2.
5. Bob enjoys playing Yahtzee (a game in which a person tosses five dice). Bob hates it when all five dice have different numbers showing, and wants to compute the probability that this hated event occurs. Unfortunately, Bob is not very good at computing probabilities. Thus, Bob programs his computer to simulate 10,000 tosses of five dice. In exactly 932 of his simulated tosses, the hated event occurs. What should Bob conclude about the probability that the hated event occurs? Explain your answer.
6. Elaine performs a balanced CRD with 50 subjects. She obtains a total of 28 successes, with 18 of the successes occurring on the second treatment. Use the standard normal curve to obtain the approximate P-value for Fisher's test and the second alternative ($<$) for Elaine's data.
7. Refer to the previous question.
 - (a) Use the standard normal curve to obtain the approximate P-value for Fisher's test and the first alternative ($>$) for Elaine's data.
 - (b) Use the standard normal curve to obtain the approximate P-value for Fisher's test and the third alternative (\neq) for Elaine's data.

8. Find the area under the standard normal curve between 0.72 and 1.63.
9. Al performs a CRD with 50 subjects on the first treatment, 25 subjects on the second treatment, and obtains a total of 12 successes—10 on the first treatment, and two on the second treatment. Use the standard normal curve to approximate the P-value for the first alternative ($>$).
10. Use the following probability histogram to compute $P(X = 0.50)$.



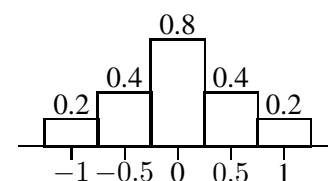
11. A random variable X has the following sampling distribution:

x	$P(X = x)$
1.0	0.2
1.5	0.3
2.0	0.2
2.5	0.1
3.0	0.2
Total	1.0

Construct the probability histogram for X . (Remember to label the possible values of X on the horizontal axis and to specify the heights of the rectangles.)

12. Cliff performs a CRD and obtains the probability histogram, pictured below, of the test statistic for Fisher's test.

Given that $X = -0.5$, find the exact P-value for the second alternative ($<$).



13. For an unbalanced CRD, the possible values of the test statistic for Fisher's test are: -0.24 , -0.10 , 0.04 , 0.18 , 0.32 , 0.46 , and 0.60 . A researcher draws the probability histogram of the sampling distribution. Given $P(X = 0.04) = 0.3507$, find the height of the rectangle centered at 0.04 .
14. Sam performs an unbalanced CRD with a dichotomous response and two treatments on 20 subjects and obtains a total of 10 successes.
- True or false? The probability histogram of the test statistic for Fisher's test is symmetric about the point 0.
15. A simulation experiment with 100 runs is performed. Thirty-five runs yield a value of the random variable X that is larger than 0.25. What is the simulation approximation of $P(X \leq 0.25)$?
16. Bob performs CRD and obtains the following data.

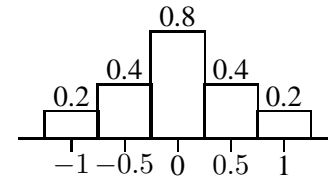
Treatment	S	F	Total
1	12	8	20
2	9	11	20
Total	21	19	40

He then decides to perform a simulation experiment to obtain an approximate P-value. The first run of the experiment yields the following table:

Treatment	S	F	Total
1	15	5	20
2	13	7	20
Total	28	12	40

What should Bob do next?

17. Find the area under the standard normal curve to the right of -0.73 .
18. Cliff performs a CRD and obtains the probability histogram, pictured below, of the test statistic for Fisher's test.



Given that $X = 1$, find the exact P-value for the third alternative (\neq).

19. A balanced CRD with two treatments and a dichotomous response yields 73 successes and 47 failures. Compute the standard deviation of the test statistic for Fisher's test.
20. The possible values of the random variable X are 3, 4, 5, 6, and 7. The sampling distribution of X has a mean of 5 and a standard deviation of 0.5. What are the possible values of the standardized version of X ?
21. A balanced CRD with two treatments and a dichotomous response yields 55 successes and 35 failures. If 25 of the successes occur on the first treatment, compute the standard normal curve approximation to the P-value for Fisher's test for the second alternative ($<$).
22. A balanced CRD with two treatments and a dichotomous response yields 28 successes and 52 failures. If 18 of the successes occur on the first treatment, compute the standard normal curve approximation to the P-value for Fisher's test for the third alternative (\neq).
23. A controlled comparative study yields the data presented in the table below.

Treatment	S	F	Total
1	7	13	20
2	1	9	10
Total	8	22	30

A 1,000 run simulation experiment yields the following results:

x	Frequency
-0.50	7
-0.35	49
-0.20	184
-0.05	337
0.10	302
0.25	101
0.40	20
Total	1,000

- (a) Use the above data and simulation results to approximate the P-value for Fisher's test with the first alternative ($>$).
- (b) Use the above data and simulation results to approximate the P-value for Fisher's test with the third alternative (\neq).
- (c) Remember that sampling distributions are computed and simulation experiments are performed on the assumption that the Skeptic is correct. With this in mind, you can see that the simulation experiment did not yield all possible values of the test statistic. List all possible values of the test statistic (according to the Skeptic) that are not represented in the simulation experiment.

3.4 Solutions to Exam Questions

1. The P-value equals $P(X \geq 0.33)$. This probability is approximated by the relative frequency of simulated values that are greater than or equal to 0.33. The relative frequency equals 0.014.
2. $\delta = 0.3$, so the height equals $0.1632/0.3 = 0.544$.
3. The correspondence is:

Value	-0.7	-0.4	-0.1	0.2	0.5	0.8
Freq.	20	177	372	327	101	3

4. The 2×2 table is below.

Treatment	<i>S</i>	<i>F</i>	Total
1	6	8	14
2	2	12	14
Total	8	20	28

Thus, $x = 6/14 - 2/14 = 4/14 = 2/7 = 0.286$.

5. The relative frequency of occurrence of the hated event is 0.0932. Thus, the probability of the hated event is approximately (not exactly!) equal to 0.0932 by an application of the long-run relative frequency interpretation of probability.

6. Elaine's 2×2 table is below.

Treatment	<i>S</i>	<i>F</i>	Total
1	10	15	25
2	18	7	25
Total	28	22	50

Thus, $x = -0.32$ and $\sigma = 0.14182$, giving $z = -0.32/0.14182 = -2.26$. The approximate P-value for the second alternative is 0.0119.

7. (a) 0.9881.
(b) 0.0238.
8. The area equals

$$0.2358 - 0.0516 = 0.1842.$$

9. Al's tables are below.

Treatment	<i>S</i>	<i>F</i>	Total
1	10	40	50
2	2	23	25
Total	12	63	75

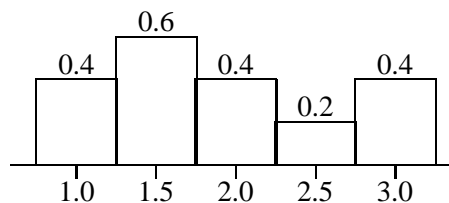
Treatment	<i>S</i>	<i>F</i>	Total
1	0.20	0.80	1.00
2	0.08	0.92	1.00

By subtraction, $x = 0.12$. Further,

$$\sigma = \sqrt{\frac{12(63)}{50(25)(74)}} = 0.0904.$$

Thus, $z = 0.12/0.0904 = 1.33$ and the approximate P-value equals 0.0918.

10. $0.5(0.4) = 0.2$.
11. The probability histogram for X is below.



12. $0.5(0.2) + 0.5(0.4) = 0.3$.

13. $0.3507/0.14 = 2.505$.
14. True.
15. 0.65.
16. Bob should check his method of simulation. It is obviously flawed since every simulation run must yield a total of 21 successes and 19 failures, just as the actual study did.
17. 0.7673.
18. $2(0.5)(0.2) = 0.2$.
19. $\sigma = 0.0895$.
20. The possible values are $-4, -2, 0, 2$, and 4 .
21. The success rates are $\hat{p}_1 = 0.556$ and $\hat{p}_2 = 0.667$, yielding $x = -0.111$. In addition,

$$\sigma = \sqrt{\frac{55(35)}{45(45)(89)}} = 0.1033.$$

Thus, $z = -0.111/0.1033 = -1.07$, and the approximate P-value is 0.1423.

22. The success rates are $\hat{p}_1 = 0.450$ and $\hat{p}_2 = 0.250$, yielding $x = 0.200$. In addition,

$$\sigma = \sqrt{\frac{28(52)}{40(40)(79)}} = 0.1073.$$

Thus, $z = 0.200/0.1073 = 1.86$, and the approximate P-value for the third alternative is $2(0.0314) = 0.0628$.

23. (a) The data yield $\hat{p}_1 = 0.35$ and $\hat{p}_2 = 0.10$. Thus, $x = 0.25$. The P-value equals $P(X \geq 0.25)$ which is approximated by the relative frequency of $(X \geq 0.25)$. This relative frequency equals

$$(101 + 20)/1000 = 0.121.$$

- (b) For the third alternative, the P-value equals

$$P(X \leq -0.25) + P(X \geq 0.25) = \\ P(X \leq -0.35) + P(X \geq 0.25),$$

which is approximated by

$$\text{Rel. freq. } (X \leq -0.35) +$$

$$\text{Rel. freq. } (X \geq 0.25),$$

This sum of relative frequencies equals

$$(49 + 7 + 101 + 20)/1000 = 0.177.$$

- (c) There are nine possible values of the test statistic because the “a” position in the contingency table can be filled with 0, 1, 2, 3, 4, 5, 6, 7, or 8. It is easy to check that if $a = 0$, then $x = -0.80$ and if $a = 1$, then $x = -0.65$. The other seven possible values are represented in the simulation experiment.

3.5 More Mathematics

I will continue the presentation of Section 2.5 of this guide. Recall that we have derived the hypergeometric formula which gives probabilities for the random variable A .

Of course,

$$1 = \sum_a P(A = a) =$$

$$\sum_a \frac{C(m_1, a)C(n - m_1, n_1 - a)}{C(n, n_1)}.$$

It is important to understand this identity in the following way. The function C has two arguments and appears three times in this formula. The arguments that vary with the summation are the second arguments in the numerator terms. The sum of the first arguments in the numerator equals the first argument in the denominator, and the sum of the second arguments in the numerator equals the second argument in the denominator.

Section 3.6 of the text defines the mean of the sampling distribution of A as

$$\mu = \sum_a aP(A = a).$$

In order to obtain the value of μ , it is helpful to note that for $t \geq 1$,

$$C(s, t) = \frac{s}{t}C(s - 1, t - 1).$$

For later use note that for $t \geq 2$,

$$C(s, t) = \frac{s(s-1)}{t(t-1)} C(s-2, t-2).$$

Thus,

$$\begin{aligned} \mu &= \sum_a aP(A=a) = \\ &= \sum_a a \frac{C(m_1, a)C(n-m_1, n_1-a)}{C(n, n_1)}. \end{aligned}$$

The summand equals 0 if $a=0$. For $a \geq 1$ use the above identity to write this last sum as

$$\begin{aligned} \sum_a a \frac{m_1}{a} \frac{n_1}{n} \frac{C(m_1-1, a-1)C(n-m_1, n_1-a)}{C(n-1, n_1-1)} &= \\ &= \frac{m_1 n_1}{n}. \end{aligned}$$

Next, the variance of A equals

$$E[(A-\mu)^2] = E[A^2 - 2\mu A + \mu^2].$$

Since the operation of expectation is simply a special kind of summation, it inherits the linearity properties of summation. In particular, this last sum equals

$$\begin{aligned} E(A^2) - 2\mu E(A) + \mu^2 &= E(A^2) - 2\mu^2 + \mu^2 = \\ &= E(A^2) - \mu^2. \end{aligned}$$

Note also that

$$E[A(A-1)] = E(A^2) - \mu.$$

Thus,

$$\text{Var}(A) = E[A(A-1)] + \mu - \mu^2.$$

Finally,

$$\begin{aligned} E[A(A-1)] &= \sum_a a(a-1)P(A=a) = \\ &= \sum_a a(a-1) \frac{C(m_1, a)C(n-m_1, n_1-a)}{C(n, n_1)}. \end{aligned}$$

The summand equals 0 if $a=0$ or $a=1$. For $a \geq 2$ use the earlier identity to write this last sum as

$$\sum_a a(a-1) \frac{m_1(m_1-1)}{a(a-1)} \frac{n_1(n_1-1)}{n(n-1)} \times$$

$$\begin{aligned} &= \frac{C(m_1-2, a-2)C(n-m_1, n_1-a)}{C(n-2, n_1-2)} = \\ &= \frac{m_1(m_1-1)n_1(n_1-1)}{n(n-1)}. \end{aligned}$$

Substituting the above values into

$$\text{Var}(A) = E[A(A-1)] + \mu - \mu^2$$

and simplifying yields (details are left to the reader)

$$\text{Var}(A) = \frac{n_1 n_2 m_1 m_2}{n^2 (n-1)}.$$

All that remains is to translate these results for A to the results in the textbook for X . To this end, we need two results.

Let Y be any random variable and let c_1 and c_2 be any numbers.

$$E(c_1 + c_2 Y) = c_1 + c_2 E(Y).$$

Proof: The left hand side of the equation equals

$$\sum_y (c_1 + c_2 y P(Y=y)).$$

The result follows from the linearity of summation.

Let Y be any random variable and let c_1 and c_2 be any numbers.

$$\text{Var}(c_1 + c_2 Y) = c_2^2 \text{Var}(Y).$$

Proof: Let $W = c_1 + c_2 Y$, and let μ denote the mean of Y . By the definition of variance and the preceding result,

$$\begin{aligned} \text{Var}(W) &= E[(c_1 + c_2 Y - c_1 - c_2 \mu)^2] = \\ &= E[(c_2 Y - c_2 \mu)^2] = c_2^2 \text{Var}(Y), \end{aligned}$$

as desired.

As stated earlier, $X = c_1 + c_2 A$, with

$$c_1 = -\frac{m_1}{n_2}, \text{ and } c_2 = \frac{n}{n_1 n_2}.$$

Thus,

$$\begin{aligned} E(X) &= -\frac{m_1}{n_2} + \frac{n}{n_1 n_2} \frac{n_1 m_1}{n} = 0. \\ \text{Var}(X) &= \frac{n^2}{n_1^2 n_2^2} \frac{n_1 n_2 m_1 m_2}{n^2 (n-1)} = \\ &= \frac{m_1 m_2}{n_1 n_2 (n-1)}. \end{aligned}$$

These values for the mean and variance of X agree with the answers given in the text.