



# An Exercise for Illustrating the Logic of Hypothesis Testing

Leigh Lawton  
University of St. Thomas

*Journal of Statistics Education* Volume 17, Number 2 (2009), [www.amstat.org/publications/jse/v17n2/lawton.html](http://www.amstat.org/publications/jse/v17n2/lawton.html)

Copyright © 2009 by Leigh Lawton, all rights reserved. This text may be freely shared among individuals, but it may not be republished in any medium without express written consent from the author and advance notification of the editor.

---

**Key Words:** Problem-based learning; Chi-square; P-value.

## Abstract

Hypothesis testing is one of the more difficult concepts for students to master in a basic, undergraduate statistics course. Students often are puzzled as to why statisticians simply don't calculate the probability that a hypothesis is true. This article presents an exercise that forces students to lay out on their own a procedure for testing a hypothesis. The result is that the students develop a better understanding for the rationale and process of hypothesis testing. As a consequence, they improve their ability to grasp the meaning of a p-value and to interpret the results of a significance test.

## 1. Introduction

We present an exercise designed to engage students in the process of conducting an hypothesis test. While we developed the exercise to demonstrate the *general process* of hypothesis testing, it also could be used as an illustration of a chi-square goodness of fit test and a reinforcement of the logic of hypothesis testing. Since our experience has been with using the exercise for introducing hypothesis testing, this will be our focus, but we believe that it could be used equally effectively for introducing chi-square testing.

The relevance of hypothesis testing has been questioned over the past few years. In 1994, Jacob Cohen encouraged psychologists to drop significance testing altogether and rely completely on confidence intervals. Since that time, scholars have argued back and forth over the relevance of hypothesis testing. (See, for example, [Batanero 1997](#); [Dahl 1999](#)). Regardless of one's opinion of the place hypothesis testing *should* have in an undergraduate statistics course, it is almost impossible to find a basic textbook that does not devote considerable coverage to the topic and it is commonly taught in basic statistics courses. It is the aim of this article to present an effective approach for communicating the logic of hypothesis testing.

Hypothesis testing often is a difficult topic for students to master. As [Bart Holland \(2007\)](#) states, "Many students encountering classical hypothesis testing for the first time consider it an abstract procedure that is initially hard to grasp, and possibly even counterintuitive." One of the issues that many students struggle with is the 'reverse logic' employed in testing. Most of us would be far more satisfied if we could state the probability that the null hypothesis (or the alternative hypothesis) is true or false. Instead, of course, we have the abstract concept of the p-value that tells

us the probability of obtaining a sample result at least as extreme as that which we find in a particular case *if the null hypothesis is true.*

We have found that student understanding of hypothesis testing can be improved through the use of a problem-based learning (PBL) exercise. There is a substantial body of literature indicating that PBL has important advantages over more traditional pedagogies in producing sustainable learning outcomes. While PBL originated in the medical field, many disciplines have adopted this approach in an attempt to produce deeper understanding and better retention of important principles. The claimed benefits of PBL relative to more traditional forms of instruction include: 1) higher levels of skill development and comprehension (see, for example, [Albanese and Mitchell 1993](#); [Rhem 1998](#)); 2) greater engagement and satisfaction among students (see, for example, [de Vries, Schmidt, and deGraaf 1989](#); [Albanese and Mitchell 1993](#)); and 3) a more effective transference of knowledge and skills from the classroom to the "real" world. ([Gallagher, Stepien, and Rosenthal, 1992](#))

A number of authors have suggested exercises to illustrate hypothesis testing. (See, for example, [Bates 1991](#) and [Eckert 1994](#).) While these exercises may be effective demonstrations of hypothesis testing, they do not really challenge students to work out the logic of hypothesis testing on their own and as a result they do not capitalize on the possibilities offered by true PBL. The exercises are not presented as problems and the students are not encouraged to develop solutions on their own.

One of the keys to effective PBL is using a "good" problem. [Lohman \(2002\)](#) contends that a good PBL problem should have three "structural features". One, the exact nature of the problem should be unclear and the information needed to solve the problem should be incomplete. Two, there should be more than one way to solve the problem. And three, the problem should not have a single right answer. We have constructed an exercise that satisfies these criteria and that effectively engages the students in the learning process.

## 2. The Problem and Exercise

What follows is a detailed description of the problem that is presented to the students, the questions that are asked of the students, and, finally, the lessons that the students are encouraged to take away from the exercise. The elements of the exercise presented in boxes are the items that actually are presented to the students. The sentences not in boxes are comments or directions to the instructor for working through the exercise. This exercise has been used successfully in introductory statistics courses for undergraduate students as well as for MBA students. It is presented after sampling distributions and confidence intervals have been covered. It is used to *introduce* the concepts of hypothesis testing, so it is the students' very first exposure to the topic. The entire exercise generally takes approximately one hour.

### 2.1 The Problem

The problem that we use to illustrate the hypothesis testing process involves a goodness of fit test because, in our experience, students find this type of problem to be the most intuitively accessible.

Stella Stat has been running a small-time gambling operation on her campus for several months. Stella sells each of the numbers 1 through 5 for \$1.00 (collecting a total of \$5.00) for each spin of a wheel. Then she spins the **Wheel of Destiny**. The person who holds the number where the spinner comes to rest gets \$4.75. (Stella keeps 25¢ per spin for running the game and supplying the beer and pretzels.)

Stella just purchased a new spinner, the critical piece of equipment for the game. Before she begins using this spinner, she wants to make certain that it is, in fact, *fair* — that is, she doesn't want some numbers to come up too often and others, not often enough. (Given the nature of the game, Stella has no incentive to cheat and she wants the game to be as fair as possible.)

Stella comes to you, her statistical guru, and asks you to verify that the new spinner is fit to use. *Describe a procedure for deciding whether the spinner is fair.*

Spinner site: [http://nlvm.usu.edu/en/nav/frames\\_asid\\_186\\_g\\_1\\_t\\_1.html?open=activities](http://nlvm.usu.edu/en/nav/frames_asid_186_g_1_t_1.html?open=activities)

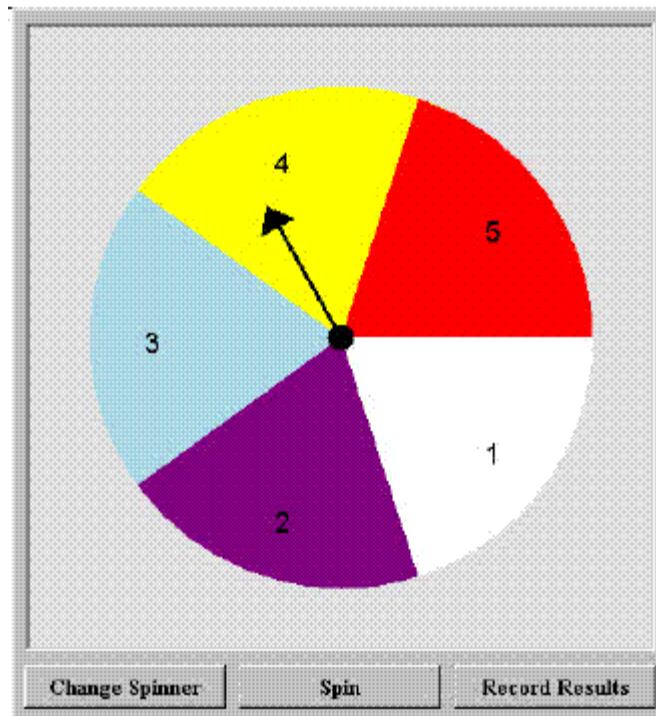


Figure 1. The Wheel of Destiny: Is It Fair?

At this point we divide the class into groups of three to five students, have the students log in to the spinner site, and give the groups a short time to reflect on how they will tackle the problem. It does not take the students long to decide that they should spin the spinner to see if the outcomes look fair. (Note: If the students do not have access to computers, this exercise can be run with a real, homemade spinner or simply with the picture of a spinner.)

We intervene here and ask the students to spin the device 50 times and to record the results. (If we use only the picture, we present them with the results of 50 spins like those shown below.)

Fifty spins yield the following:

Number	Number of observations
1	10
2	12
3	6
4	7
5	<u>15</u>
Total	50

*All right, now you have the data. Is the spinner fair, or not?*

*Given the available data, determine what you should do to make a judgment as to whether the new spinner is fair.*

Figure 2. The results of 50 spins of the wheel

Either on their own or with a little prompting, the students will deduce that if the spinner is fair, all numbers are equally likely to occur. They realize that, given a total of 50 spins, each number should come up about 10 times, but

they also recognize that sampling error is sure to be present so they cannot reasonably expect identical frequencies of the numbers 1-5.

We intervene once again to provide the students with a measure of the departure of the sample results from 'fair.' Students are comfortable with the idea of looking at the difference between the observed and expected values. The idea of squaring this difference and then dividing by the expected value is not so intuitively obvious, but most will relate the squaring operation to what is done when calculating a standard deviation. As with the standard deviations, unless we square the differences or use absolute values, positive differences will be exactly offset by negative differences.

Number	Number of observations	'Expected' number of observations	Difference between observed and expected	Difference squared	Sq. diff/ exp. num of obs
1	10	10	0	0	.00
2	12	10	2	4	.40
3	6	10	-4	16	1.60
4	7	10	-3	9	.90
5	<u>15</u>	<u>10</u>	<u>-5</u>	<u>25</u>	<u>2.50</u>
<i>Total</i>	<i>50</i>	<i>50</i>	<i>0</i>		<i>5.40</i>

Chi-square is calculated as follows:

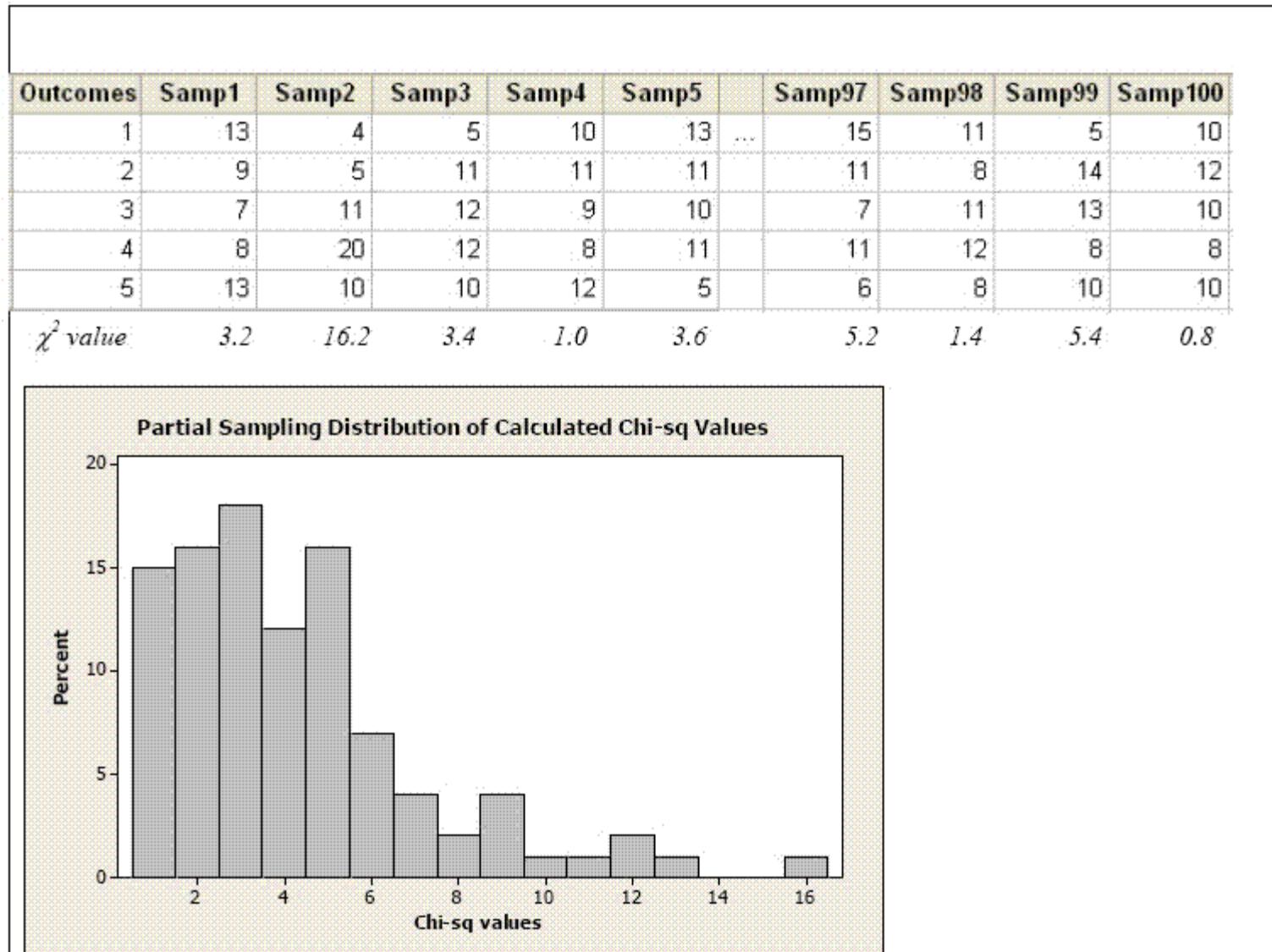
$$\chi^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$

*This is a measure of the magnitude of the discrepancy between what we actually observe and what we could anticipate from a fair spinner.*

a) Explain in common sense terms the conditions under which the calculated  $\chi^2$  value would be large and the conditions under which it would be small.  
 b) What is the range of possible values? Smallest? \_\_\_\_\_ Largest? \_\_\_\_\_  
 c) Is this calculated  $\chi^2$  value (5.40) large or small?  
 d) Is the spinner fair, or not?

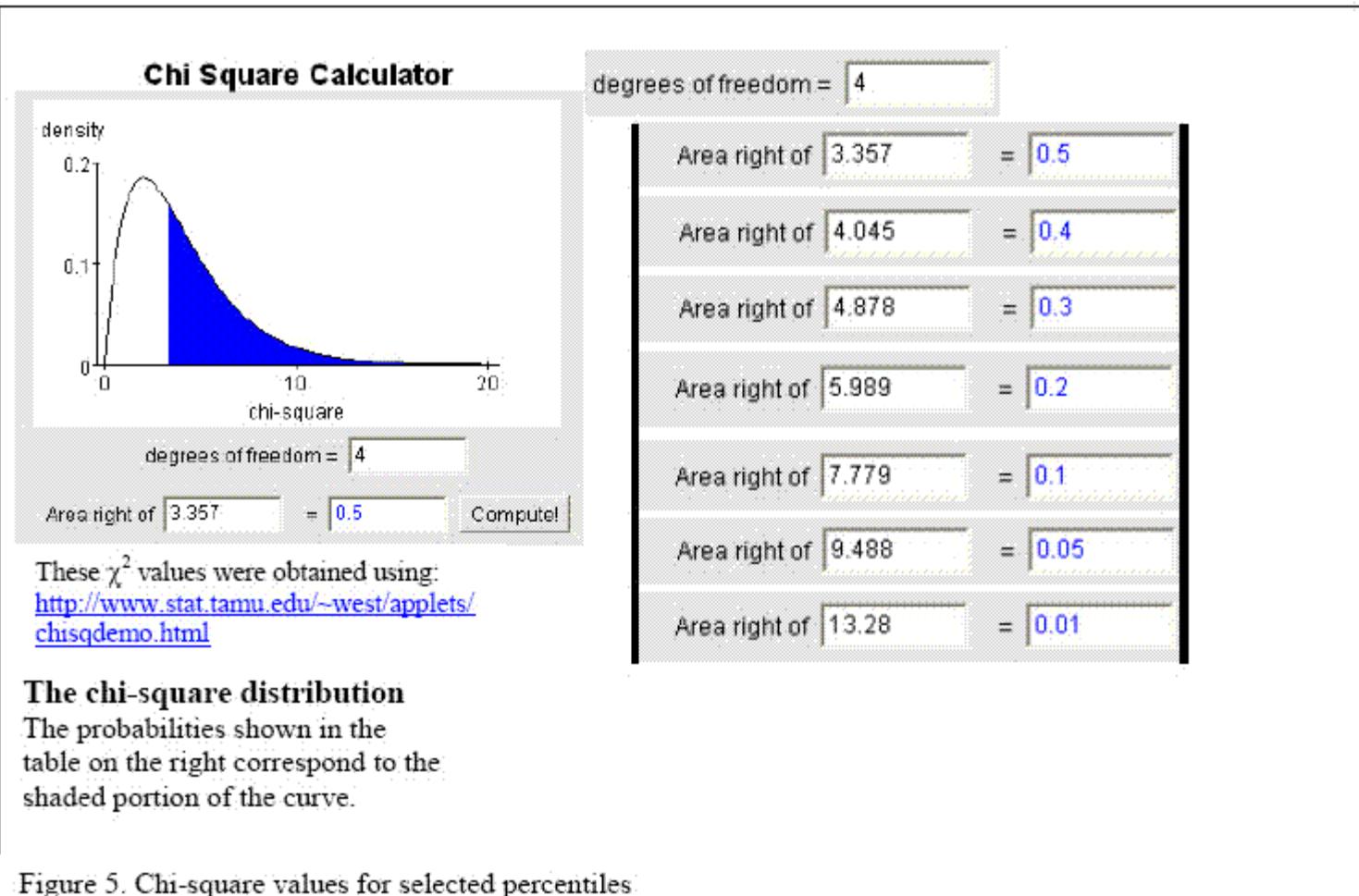
Figure 3. A measure of the wheel's "fairness"

Most will see that this number alone has limited usefulness; they will recognize that they need some external standard against which they can compare their results if they are to draw a conclusion as to whether the spinner is fair. To maintain the intuitive appeal of the exercise, we now present the students with the results of a Monte Carlo simulation of a fair spinner. We use the Minitab random number generator to produce 100 samples with each sample consisting of 50 spins. We explain to them that these results were generated using a uniform distribution where each number 1-5 has an equal chance of occurring. We have good reason to believe that sampling from the Minitab random number generator produces results that closely approximate a uniform distribution. We calculate the value of  $\chi^2$  for each sample of 50 spins.



This Monte Carlo simulation demonstrates that we can create a probability distribution of the  $\chi^2$  statistics that we calculated from our randomly drawn samples for our fair spinner analog. We should note, however, that this example produces just a *partial* sampling distribution (based on only 100 samples). In 1875 Friedrich Helmert, a German scientist, determined that the true, theoretically *complete* chi-square distribution for the conditions we face would look like the graph shown below.

Figure 4. A Monte Carlo simulation of chi-square values



### The chi-square distribution

The probabilities shown in the table on the right correspond to the shaded portion of the curve.

Figure 5. Chi-square values for selected percentiles

Selected percentiles...	Monte Carlo partial sampling distribution	Chi-square distribution
50% of all $\chi^2$ values are greater than	3.6	3.357
40% of all $\chi^2$ values are greater than	4.2	4.045
30% of all $\chi^2$ values are greater than	4.8	4.878
20% of all $\chi^2$ values are greater than	5.6	5.989
10% of all $\chi^2$ values are greater than	8.2	7.779
5% of all $\chi^2$ values are greater than	10.4	9.488
1% of all $\chi^2$ values are greater than	13.4	13.28

(so, for example, 50% of all  $\chi^2$  values in the Monte Carlo partial distribution are greater than 3.6 while 50% of all  $\chi^2$  values in the chi-square distribution are greater than 3.357)

We ask you to accept that if we were to run the Monte Carlo simulation an infinite number of times (rather than only 100 as we did here), the  $\chi^2$  values for a 'fair' spinner would be equal to those shown by the chi-square distribution. This should not be too hard to believe given the similarity of the numbers shown in the table.

Figure 6. Comparison of our partial sampling distribution to the chi-square distribution

### Now that all the data are in, is the spinner fair, or not?

We show the students the probabilities that correspond to the sample statistic.

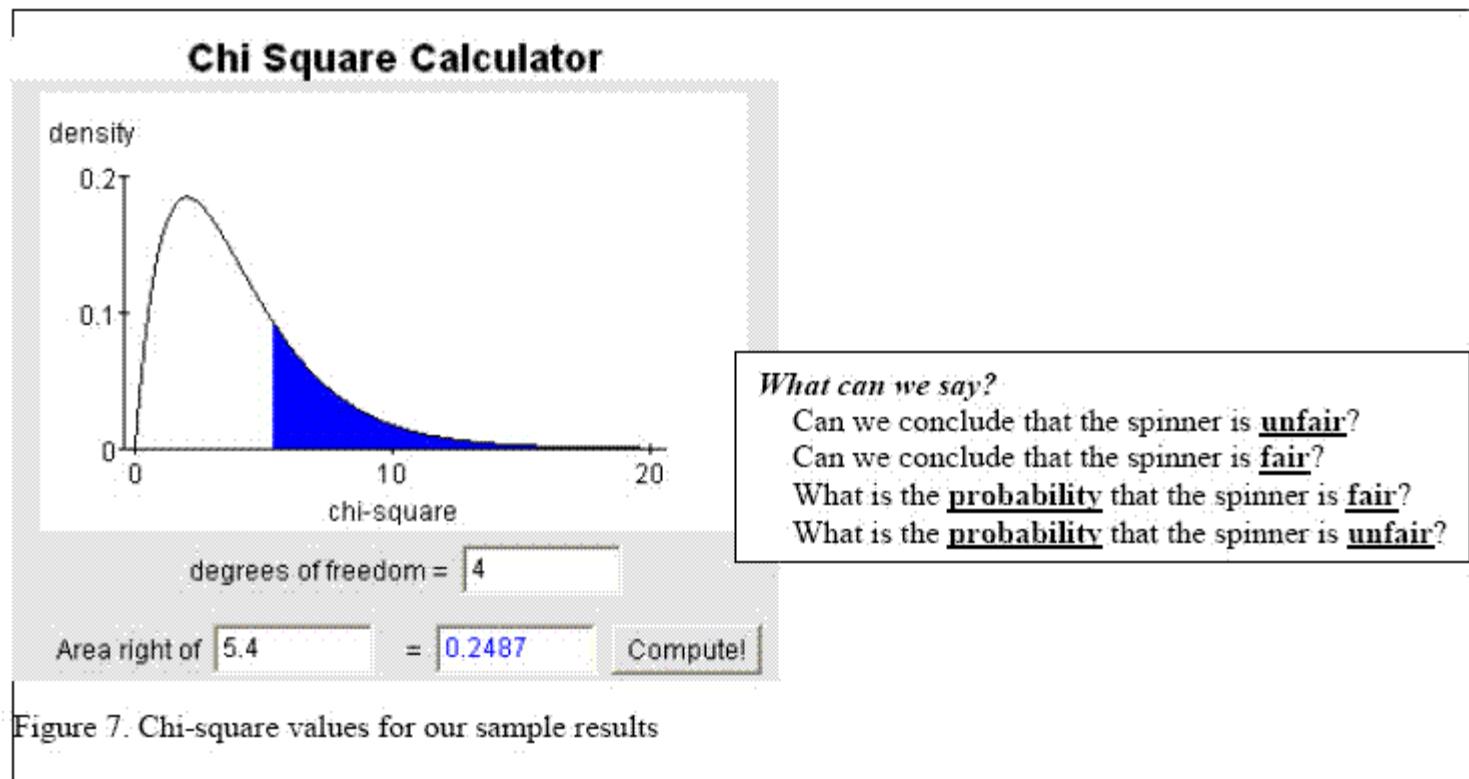


Figure 7. Chi-square values for our sample results

In the ensuing discussion, we make the points shown below.

## 2.2 Lessons to be Learned from the Wheel of Destiny

**Note the elements of the procedure that we intuitively follow:**

- We assume that the spinner is fair (i.e., we assume that the null hypothesis is true).
- We collect sample data and make a calculation using the sample data (i.e., we calculate a number that captures the magnitude of the difference between our sample results and what's expected if the spinner is fair; we call this number our *sample statistic*).
- We select some external standard against which we can compare the results of our calculation (i.e., we find a distribution – in this case the  $\chi^2$  distribution – that provides a measure of what constitutes 'normal' variation between our observed results and what we would expect to see if the spinner is fair).
- We look at the value of our sample statistic to see how likely it would be to occur if the null hypothesis is true (i.e., we see where the 5.4 falls on the  $\chi^2$  distribution. In this case, .7513 of all  $\chi^2$  values lie below 5.4 – or .2487 of all  $\chi^2$  values lie above 5.4 – if the spinner is fair. We call this .2487 the **p-value**. It represents the probability of obtaining a sample statistic as high or higher than that found in our sample *if the spinner is fair*.)
- We make an arbitrary judgment about the point at which the discrepancy between our sample statistic and the external standard is too great for us to consider the difference to be a reflection of sampling error alone (i.e., we choose some 'a' like 0.10 or 0.05).
- We conclude that our null hypothesis is untrue (the spinner is unfair) if our p-value is smaller than our arbitrarily chosen a. If our p-value is larger than our a, we have not proven that our null hypothesis is true, but we don't have compelling evidence that it is false.

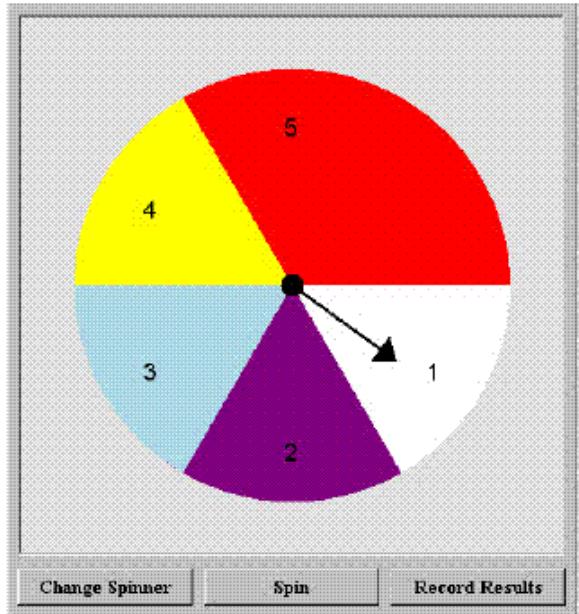
**Note where things stand when we have completed the procedure:**

- We can't say for certain whether the spinner is fair or unfair.
- We can't calculate the probability that the spinner is fair.
- We can't calculate the probability that the spinner is unfair.

We can say either of two things (depending upon our sample results):

- our sample results are reasonably consistent with results that we would observe if the spinner is fair (i.e., the null hypothesis does not appear to be unreasonable), or
- our sample results are *not* consistent with the results we should observe if the spinner is fair (i.e., the null hypothesis appears to be unreasonable).

### 3. Additional Exercises Using the Spinner



It is possible to expand upon this exercise. For example, the spinner can be altered so that regions differ in size (as shown here). Using this spinner will, of course, generate a very different set of outcomes and can be used to show that the resulting  $\chi^2$  values will be much larger. If students develop a sense for the meaning of the chi-square statistic, they should grasp with little difficulty that this ‘weighted’ spinner inevitably will produce a large chi-square value and that the probability of seeing such a big value is remote if the spinner really is fair – that is, they should accept the notion that the resulting p-value will be quite small.

### 4. Conclusion

Our experience with the exercise described above suggests that students grasp the reason why we test hypotheses as we do and consequently have an easier time interpreting the results of hypothesis testing if they develop the process of conducting a hypothesis test on their own rather than being *told* how tests are performed. When students begin to show signs of confusion as we present other types of hypothesis tests, we have found it helpful to refer them to the Wheel of Destiny example and to think about how and why we acted as we did.

### References

Albanese, M., and Mitchell, S. (1993), "Problem-Based Learning: A Review of the Literature on its Outcomes and Implementation Issues," *Academic Medicine*, 68(1), 52-81.

Batanero, C. (1997), "Should We Get Rid of Statistical Testing? The Significance Test Controversy," *ISI Newsletter*,

Bates, J. A. (1991), "Teaching Hypothesis Testing by Debunking a Demonstration of Telepathy," *Teaching of Psychology*, 18(2), April, 94-97.

Cohen, J. (1994). "The Earth is Round," *American Psychologist*, 49(12), 997-1003.

Dahl, H. (1999), "Teaching Hypothesis Testing. Can it Still be Useful?"

<http://www.stat.auckland.ac.nz/~iase/publications/5/dahl0745.pdf> (accessed February 24, 2008).

de Vries, M., Schmidt, H. G., and deGraaf, E. (1989), "Dutch Comparisons: Cognitive and Motivational Effects of Problem-Based Learning on Medical Students," in Schmidt, H. G., Lipkin, M., de Vries, M. W., and Greep, J. M. (eds), *New Directions for Medical Education: Problem-Based Learning and Community Oriented Medical Education*. New York: Springer-Verlag, 230-240.

Eckert, S. (1994), "Teaching Hypothesis Testing with Playing Cards: A Demonstration," *Journal of Statistics Education*, 2(1), <http://www.amstat.org/publications/jse/v2n1/eckert.html>.

Gallagher, S. A., Stepien, W. J., and Rosenthal, H. (1992), "The Effects of Problem-Based Learning on Problem Solving," *Gifted Child Quarterly*, 36(4), 195-200.

Holland, B. K. (2007), "A Classroom Demonstration of Hypothesis Testing," *Teaching Statistics*, 29(3), 71-73.

Lohman, M. C. (2002), "Cultivating Problem-Solving Skills through Problem-Based Approaches to Professional Development," *Human Resource Development Quarterly*, 13(3), 243-261.

Rhem, J. (1998), Problem-Based Learning: An Introduction. *National Teaching & Learning Forum*, 8(1), pp. 1-4.

---

Leigh Lawton  
Mail No. MCH 316  
University of St. Thomas  
St. Paul, MN 55105-1096  
651.962.5084  
[l9lawton@stthomas.edu](mailto:l9lawton@stthomas.edu)

---