Titanic: A Statistical Exploration

Sandra L. Takis

The tremendously popular movie *Titanic* has rejuvenated interest in the *Titanic* and its passengers. Students are particularly captivated by the story and by the people involved. Consequently, when I was preparing to explore categorical data and the chi-square distribution with my class, I decided to use the available data about the *Titanic's* passengers to interest students in these topics. This article describes the activities that I incorporated into my statistics class and gives additional resources for collecting information about the *Titanic*.

Analyzing Categorical Data

A topic in the descriptive-statistics strand of the Advanced Placement curriculum is analyzing categorical data using conditional and marginal frequencies. In this analysis, students examined the distributions of specific outcomes for different groupings of the population, comparing proportions rather than counts of the data. To perform this type of analysis using a real-world example, we examined the overall population of *Titanic* passengers and survivors, as shown in Table 1.

Table 1

***Titanic* Passengers and Survivors by Class and Age or Gender**

Passenger Category	Number of Passengers	Number of Survivors
Children, first class	6	6
Children, second class	24	24
Children, third class	79	27
Women, first class	144	140
Women, second class	93	80
Women, third class	165	76
Men, first class	175	57
Men, second class	168	14
Men, third class	462	75

Source: His Majesty's Stationery Office, 1912

The first two questions that the students explored were the following:

- Was a difference in the survival rates related to the class of passenger?
- Was a difference in the survival rates related to the gender or age of the passenger?

Both of these questions are interesting not only from a statistical perspective but also from a historical perspective. The makers of the movie *Titanic* imply that lower-class passengers were treated unfairly. This analysis of categorical data allows students to determine whether that portrayal is accurate.

To examine these issues, my students started by developing relative frequencies of survival and loss for first-class, second-class and third-class passengers and similar rates for men, women and children. Figures 1 and 2 illustrate these data. After a preliminary graphical analysis, students likely will conclude that survival depended on both class and gender or age. Approximately 60 percent of the first-class passengers survived, compared with approximately 40 percent of the second-class passengers and approximately 25 percent of the third-class passengers. Women had the highest survival rate at approximately 75 percent; children were next at approximately 50 percent; and finally men at approximately 20 percent.

Figure 1

Survival Rates by Class

Because two factors appeared to influence survival, we examined survival data further. Specifically, we examined whether the differences in survival rate by class were caused by the very different distributions of men and women in each class. Many more men than women were in second and third class, whereas the numbers of men and women in first class were more similar. To determine the effect that the gender distribution had on the survival rates of the first-class, second-class and third-class passengers, we examined the data in more detail. The histogram in Figure 3 illustrates the disaggregated data on survival rates by both class and gender or age.

This analysis helped students examine the effect of class and gender or age together and to think about some of the shortfalls of examining aggregate data. For example, although the overall rate of survival for second-class passengers was higher than for third-class passengers, the rate was lower for second-class adult males than for third-class adult males.

Although this situation does not represent the complete reversal of the relationship known as *Simpson's paradox*, it did show the students that they need to be cautious about making conclusions based on aggregate data. Simpson's paradox occurs when an association found between two variables that are examined at disaggregate levels is reversed when the variables are examined at an aggregate level. In this situation, Simpson's paradox would occur if the overall survival rates were higher for second-class passengers than for third-class passengers even though the individual survival rates for second-class adult males, adult females and children were lower than those of the third-class adult males, adult females and children, respectively. Such a reversal occurs when differences exist in the distribution of each group—here, adult males, adult females and children—for each aggregate group—here, second- and third-class passengers.

From the analysis, the students more clearly realized that gender played a larger role than class. Several specific questions made the students think about the differences in the overall survival rates. A few examples follow:

- What proportion of the first-class passengers were adult males, adult females or children?
- How did this proportion affect the overall survival rate of the first-class passengers?
- How were the gender distributions different for second- and third-class passengers, and how would these distributions affect the overall survival rate of these groups?

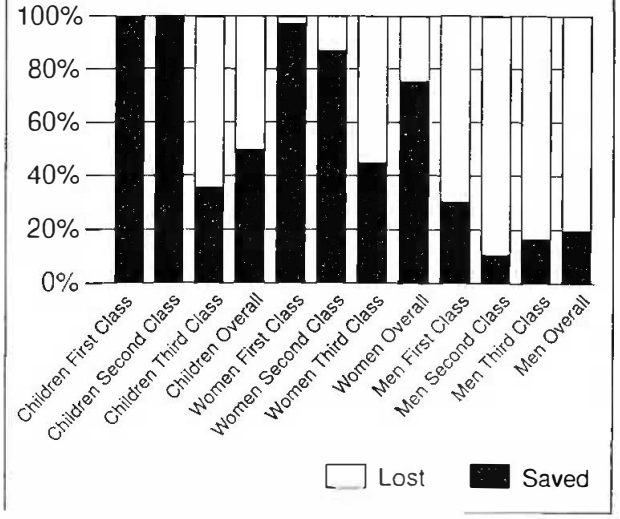
In the future, I plan to ask the students to write a summary of the relationship between passenger class and survival rates from a data-analysis perspective. To follow up this report, students could examine the formal investigation report of the disaster prepared for the British Parliament, as well as other reports prepared at the time.

Applying the Chi-Square Test of Independence to an Entire Population

A second activity in which I used the *Titanic* data was in introducing the chi-square (χ^2) test of independence. This test can be used to examine

Figure 2
Survival Rates by Gender/Age

Figure 3
Survival Rates by Gender/Age and Class



whether a relationship exists between two categorical variables. In this example, we applied the test to the entire population of *Titanic* passengers to determine whether a statistically significant relationship existed between passenger class and survival. The χ^2 test can be applied to data from independent samples from several populations where the populations act as one of the two categorical variables or to data from a single sample where subjects are classified by two categorical variables. For further discussion, see Moore (1995, 535–37).

Table 2 summarizes survival and loss by passenger class. The crucial concept in the test of independence is understanding what the data would look like if no relationship existed between the variables, in this example, survival rate and passenger class. First, if survival did not depend on class, the overall survival rate should stay constant regardless of the passenger class. The overall survival rate was 37.9 percent, and the overall loss rate was 62.1 percent. These rates were applied to the total number of passengers for each class to determine the expected number of survivals and losses in each class. Students needed to understand that they were applying the survival and loss rates to the total number of passengers in each category and were consequently assuming that survival and class are independent.

Table 2

Survival by Class: Summary and Percents

Class	Survived	Lost	Total
First	203	122	325
Second	118	167	285
Third	178	528	706
Total	499	817	1,316
Percent	37.9%	62.1%	

After we calculated the expected values for each category, we calculated the chi-square statistic for the sample data. The formula for this calculation is

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

The calculation is similar to that of standard deviation in that we measured how far the observed value was from the expected outcome, just as we measured how far individual data points are from the mean when calculating standard deviation. This difference is squared to eliminate negative values and is divided by the expected outcome to standardize the measure. The χ^2-test statistic is the sum of these measures for each of the categories. Table 3 shows the calculation of the chi-square statistic for the population data.

Table 3

Calculation of Expected Outcome and Chi-Square-Test Statistic

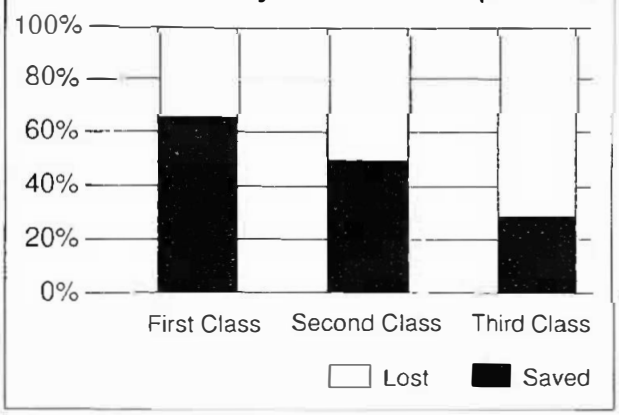
Passenger Category	Observed Data	Expected Outcome	Differences between Observed and Expected	Difference Squared and Divided by the Expected
First class survived	203	$\frac{499}{1316} \cdot 325 = 123.23$	$203 - 123.23 = 79.77$	$\frac{79.77^2}{123.23} = 51.64$
First class lost	122	$\frac{817}{1316} \cdot 325 = 201.77$	$122 - 201.77 = -79.77$	$\frac{(-79.77)^2}{201.77} = 31.54$
Second class survived	118	$\frac{499}{1316} \cdot 285 = 108.07$	$118 - 108.07 = 9.93$	$\frac{9.93^2}{108.07} = 0.91$
Second class lost	167	$\frac{817}{1316} \cdot 285 = 176.93$	$167 - 176.93 = -9.93$	$\frac{(-9.93)^2}{176.93} = 0.56$
Third class survived	178	$\frac{499}{1316} \cdot 706 = 267.70$	$178 - 267.70 = -89.70$	$\frac{(-89.70)^2}{267.70} = 30.06$
Third class lost	528	$\frac{817}{1316} \cdot 706 = 438.30$	$528 - 438.30 = 89.70$	$\frac{89.70^2}{438.30} = 18.36$
Total	1316	1316		$\chi^2 = 133.07$

As in other tests of significance, we compared the calculated measure with a standardized distribution of measures to determine whether the statistic's measured differences were caused by random variation or whether these differences were too large to be caused by chance. We made this determination in two ways. One approach was to compare the calculated measure with a cutoff value in the chi-square distribution. The distribution represented variation in the chi-square value that would occur randomly when no relationship existed between our two variables. Most random variation will result in small chi-square scores. We set a cutoff, or critical, value at a point above which very few randomly generated scores occurred. We set our critical value at 5.99, using a significance level of 5 percent and degrees of freedom of 2. Degrees of freedom for a test of independence equals (number of rows -1) times (number of columns -1). In this example, it is $(3-1)(2-1)=2$; therefore, at a 5 percent level of significance, the critical chi-square score is 5.99. On the basis of this cutoff value, we saw that the population showed more variation from the expected value than would generally occur by chance. We therefore concluded that a statistically significant relationship existed between survival rate and class.

A second approach was to use a p-value rather than a critical value. The TI-83 calculator will perform the chi-square calculation using matrices and the statistical test functions. It will also supply a p-value from which our conclusion can be drawn. In this case, it calculates a p-value of 1.28×10^{-29}. We interpreted this result as meaning that very little chance, a 1.28×10^{-29} probability, exists that such extreme differences in survival between classes would occur when truly no difference exists in survival by class. The ability to calculate a p-value forges a strong link to the underlying probability concepts that give meaning to statistical tests.

Figure 4

Survival Rates by Class for Sample Data



Comparing Conclusions Drawn from Sample Data with Those Drawn from the Entire Population

A third activity for using the *Titanic* data was to have students answer the previously proposed questions using sample data. Because data on the entire population were available, the sample test that follows was not necessary to draw a conclusion. However, we used sample data from a known population so that students could compare their results with the known results and revisit such concepts as sampling error and variation in samples. The following is a synopsis of the students' process.

To study the questions of differing survival rates by class and gender, students broke up into small groups and developed a method for taking random samples of 100 passengers. Each group took a different approach to its sampling plan. Some groups randomly selected a starting point and systematically chose every 20th passenger. Others used their calculators to randomly generate numbers and counted off the list to find those passengers. When students had their samples, I asked them to create a contingency table by class and a relative-frequency histogram summarizing the survival rates. Table 4 and Figure 4 illustrate an example of sample results.

The students then calculated expected values and chi-square scores for each of their samples. Each group completed a test of independence on its sample data. Of the seven groups that performed the analysis in my class, only one did not conclude that a relationship existed between passenger class and survival rates. Such results gave the students an opportunity to discuss possible sources of error. The other students wanted to conclude that the group made an error either in its calculation or in its sampling technique. This group, however, when given the opportunity to defend

Table 4

Example of a Student Sample

Class	Survived	Lost	Total
First	12	6	18
Second	14	14	28
Third	15	39	54
Total	41	59	100

its work, could justify the work and give the alternate possibility; the conclusion was incorrect because of a type II error. In this situation, a type II error represents the outcome for which the null hypothesis—passenger class and survival rates are independent—is not rejected even though it is incorrect. As a follow-up to this activity, I asked the students to perform a similar analysis to determine whether survival and gender or age were related. The list we used had titles (Mr., Mrs. and so on) and identified children, so the students could easily classify individuals in the sample as adult male, adult female or child. Students wrote up their full analysis, including relative-frequency histograms, calculations and the final conclusion.

The overall process was exciting and motivating for both my students and me. The students were very interested in the sociological aspects of the study and enjoyed talking about the movie and fact versus fiction. Several resources on the Internet allow further examination of the *Titanic* disaster. I used the following sites to obtain information on the *Titanic*. These sites also have links to additional sites for more information.

Internet Sites

- www2.nexus.edu.au/TeachStud/titanic2/home: This site is called "The Titanic in the Classroom." It gives activities for using the *Titanic* for classroom activities.
- www.anesi.com/titanic.htm: This site is called "The Titanic Casualty Figures." It examines the casualty figures in this article and provides sources of investigations performed at the time of the disaster.

References

Anesi, C. "The Titanic Casualty Figures." Available online <http://www.anesi.com/titanic.htm.>

"British Parlimentary Papers, Shipping Casualties (Loss of the Steamship 'Titanic'), 1912, cmd. 6352." In *Report of a Formal Investigation into the Circumstances Attending the Foundering on the 15th April, 1912, of the British Steamship "Titanic," of Liverpool, after Striking Ice in or near Latitude 41B 46' N., Longitude 50E 14' W. North Atlantic Ocean, Whereby Loss of Life Ensued,* 42. London: His Majesty's Stationery Office, 1912.

Department of Education Training and Employment (South Australia). "The Titanic in the Classroom." Available online <http://www2.nexus.edu.au/TeachStud/titanic2/home.>

Moore, D. S. *The Basic Practice of Statistics*. New York: Freeman, 1995.

Reprinted with permission from The Mathematics Teacher, Volume 92, Number 8 (November 1999), pages 660–64, an NCTM publication. Minor changes have been made to spelling and punctuation to fit ATA style.

Dodecahedra

I have an unknown number of regular dodecahedra, which I cannot differentiate from one another. If I take red and blue paint and paint every face of each dodecahedron either red or blue, how many different dodecahedra (based on color) can I create?