

Examining Categorical Data in Contingency Tables

- A. Categorical Data
 - 1. categorical, nominal, or frequency data
 - 2. frequency tables
 - 3. likelihood and odds
- B. What Is A Contingency Table?
 - 1. cross-tabulation
 - 2. relationships between variables
 - 3. likelihood ratio and relative risk
 - 4. conditioning
- C. Correcting Calculations For Group Sizes
 - 1. proportions and percentages

A. Categorical Data

Haven't you always wanted to take a cruise on one the world's truly premier luxury oceanliners? Sounds like a good spring break, right? I'm making your wish come true. Tomorrow morning when you wake up, the date will be April 12th, 1912, and you will be cruising across the Atlantic from Southampton, England, via Cherbourg, France, and Queenstown, Ireland, to New York City on the maiden voyage of one of the most spectacular cruise ships ever built -- the RMS Titanic. Sorry! Too late to get off now unless you are a VERY good swimmer.

If you remember your history (or saw the movie), then you know what's in store for you. In two days your ship will strike an iceberg and sink in about two hours in the freezing waters of the North Atlantic. What are your chances of surviving this catastrophe? Here are some data that might be of interest to you at this point.

Survived	
No	Yes
1490	711

This table shows the number of people who died in the wreck versus the number who survived. These are called categorical data, nominal data, or frequency data. In this type of data, subjects are classified into categories: male/female, Democrat/Republican, lived/died, brown eyes/blue eyes/hazel eyes/green eyes, etc. The easiest way to determine that you have categorical data is to examine one subject--we'll call him Fred. What kind of information did you get from Fred? In this case, Fred was placed into a category: he either survived or he did not.

Categorical data are summarized in frequency tables like the one above, in which we see that 1490 "Fred's" did not survive the sinking of the Titanic, while 711 "Fred's" did survive. These numbers are not your raw data. They are not what you get from one subject (Fred). They are a summarization of your raw data. Because categorical data are often summarized in frequency tables, they are often called frequency data, but the frequencies in such a table are not your data. They are a summary of your data. Your data are obtained from individual subjects, such as Fred.

Whenever you are presented with data such as these, there are two questions you should ask immediately. Where did these data come from? And are the data accurate?

These are the data collected by the British Board of Trade during their investigation of the sinking. There is, in fact, some debate over the accuracy of these data. Some sources maintain that there were 23 additional male crew members aboard, all of whom perished, and that one first class passenger listed as surviving in the data above actually did not survive. Therefore, these sources would amend the frequency table above as follows:

Survived		
	No	Yes
	1514	710

Which is correct? No one knows the answer to that, and therein lies your first lesson in statistics: Data are not always accurate. Sometimes accurate data are difficult or impossible to come by. However, in this case, there is very little doubt that very nearly 2,200 people (give or take) were aboard the Titanic, and about 1,500 of them did not survive the passage. (For detailed discussion, see the Encyclopedia Titanica at <http://www.encyclopedia-titanica.org/>.)

So, given that at the very best you are in for a rough few days, what are your chances of coming out of it alive? There are two numbers often calculated from such data that are given as answers to this question. The first is called the likelihood of survival, calculated by dividing the number of people who survived by the total number of people at risk (i.e., aboard the ship when it struck the iceberg). The second is called the odds of survival, calculated by dividing the number of people who survived by the number who did not survive. The following table gives these statistics (which you should be sure you can calculate for yourself, so break out your calculator).

	Board of Trade data	revised data
likelihood of survival	.323	.319
odds of survival	.477	.469

As you can see, it doesn't much matter which data we use. The results come out to be very similar. So from here on, we will use the Board of Trade data.

Here are some questions for you to attempt to answer about likelihood and odds. You shouldn't have much trouble with them if you understand how these statistics are calculated.

- 1) What value of likelihood would indicate a 50:50 (even) chance of survival? (I.e., if an equal number of people survived vs. died, what would be the likelihood of survival?)
- 2) What value of likelihood would indicate no chance of survival?
- 3) What value of likelihood would indicate that survival is a sure thing?
- 4) What are the smallest and largest values of likelihood that are possible?
- 5) In questions 1-4, change the word "likelihood" to "odds" and answer the questions again.
- 6) Which of these statistics is the easiest to make sense of?
- 7) Which of these statistics gives you the most information?
- 8) Which of these statistics is basically a proportion of people on board who survived? What do you get if you multiply this number by 100?

B. What Is A Contingency Table?

So far, our investigation of your possible fate has been fairly coarse. We have looked at overall survival rate. That is, so far the only thing we have assumed we know about you is that you are on the Titanic when it sinks. Suppose we know something else about you, let's say gender. Does that help us revise our estimate of your likelihood of survival? Here are the data from the Board of Trade.

	Survived	
Sex	No	Yes
Male	1364	367
Female	126	344

This table is a little more complex than the previous one. It shows survival broken down by gender (or "contingent upon" gender). Such a table is called a contingency table or a cross-tabulation.

We now see that there were $1364 + 367 = 1731$ men aboard, of whom 367 survived. So if you're a man (male) aboard the Titanic when it sinks, your likelihood of survival is $367/1731 = .212$, or 21.2%. In other words, 21.2% of the male passengers are destined to survive.

There are fewer women on board, $126 + 344 = 470$, according to the Board of Trade data. Of those, most (344) will survive, for a survival rate (likelihood of survival) of $344/470 = .732$, or 73.2%. You are much better off if you are female, and not surprisingly so, since the policy for putting people into lifeboats is woman and children first.

In this case, we can say that there is "a relationship between gender and survival." Other ways to say the same thing: survival is related to gender, survival is associated with gender, survival is dependent upon gender, there is an effect of gender on survival (although cause-and-effect is not implied). A variable is something we record information about, and the information we record can take on different values. Gender is a variable, since recording information about a person's gender can result in two values: male and female. Survival is also a variable. (What are its possible values?) Two variables are related when knowing the value of one of them gives us information about the value of the other. Here, knowing your gender gives us quite a bit of information about your chances of survival. Therefore, survival is related to gender.

In fact, you are $.732/.212 = 3.45$ times as likely to survive if you are female. This statistic is called a likelihood ratio, because it is the ratio of two likelihoods, and it is a reasonable way to compare likelihoods from two distinct groups. It is also called relative risk, although in this case it would be the risk of surviving, which seems a strange thing to be called "risk."

Some more questions:

- 1) Gender is called a categorical variable (also a nominal variable), since its values place the people about whom it is observed into categories. What are some other ways your fellow passengers aboard the Titanic might be categorized?
- 2) Is survival a categorical variable? Why or why not?
- 3) How are the "groups" mentioned in the last paragraph related to the values of a variable?
- 4) What likelihood ratio would indicate no relationship between gender and survival? (Hint: No relationship would mean that gender gives us no information about survival, or in other words that the likelihood of survival is the same for men and women. What would be the likelihood ratio if the likelihood of surviving is the same for men and women?)
- 5) Suppose the likelihood ratio were close to the number you arrived at as your answer to question 4 but was not quite equal to it. Would you then say there is a relationship between gender and survival?

If we do this the other way around and calculate the likelihood of perishing, we find that it is .788 for men and .268 for women. (Do it!) This gives a likelihood ratio of $.788/.268 = 2.94$. In other words, a man is almost three times as likely to perish in the sinking as is a woman. Put another way, the risk of a man dying is nearly three times as great as the risk of a woman dying. The term "relative risk" makes a lot more sense when the calculation is done this way.

To put this number in some perspective, if you are in your early thirties, are male, started smoking when you were a teenager, and smoke a pack of cigarettes a day, you are about 8.7 times as likely to die of lung cancer than is a similar

person who is a nonsmoker (<http://www.smokefree.gov/smokersrisk/about.asp>). In the media, these numbers are more often reported as percentages, as in "you are 33% more likely to die of blah if you blah blah blah." To turn these percentages into a likelihood ratio, you should divide them by 100 and then add 1. So 33% more likely would correspond to a likelihood ratio of 1.33. (This calculation can be done somewhat differently, which would give a somewhat different answer.)

Here are some recent examples:

- 1) A study published in the Journal of the American Medical Association (2011, v.306, p.1549) and widely reported in the media found a 17% increase in risk among men of developing prostate cancer after taking 400 IU of vitamin E daily for three years. In other words, the relative risk, or likelihood ratio, of developing prostate cancer between this group and the placebo group was 1.17.
- 2) A study in the International Journal of Cancer (2008, August 14th) reported that people taking antibiotics (compared to those not taking antibiotics) had a 14% increased risk of developing breast cancer (relative risk = 1.14) and a 79% increased risk of developing lung cancer (relative risk = 1.79).
- 3) A study published in the Journal of the American Medical Association (2006, January 11th) reported that people who were obese earlier in life (body mass index of 30 or more) but who otherwise engaged in low risk behaviors had 4.2 times the risk for hospitalization with coronary heart disease by age 65 than people who were of normal weight (BMI < 25). In other words, the likelihood ratio was 4.2, and the percentage increase in risk was 320%.

But back to your ocean voyage. We know something else about you. You're not a member of the crew. This doesn't make much difference if you're a man, since male members of the crew will die at about the same rate as other male passengers, 77.7% (or 670 deaths out of 862 male crew members). If you're a woman, on the other hand, you might consider joining the crew, since only 3 of 23 female crew members will die (13.0%). Looking only at ticket-holding passengers then, the survival statistics contingent upon gender are as follows:

	Survived	
Sex	No	Yes
Male	694	175
Female	123	324

From this we can calculate (do it!) that 79.9% of ticket-holding male passengers will die, while 27.5% of ticket-holding female passengers will die. The risk of dying is 2.90 times greater for the men as compared to the women. These calculations are called "conditional" because they are conditioned on the fact that you are not a crew member. (I.e., only ticket-holding passengers were included in the calculations.)

Questions:

- 1) If we consider only men (calculations conditioned on gender), and divide them into two groups, members of the crew and ticket-holding passengers (variable 1), and then subsequently divide them into two groups by survival (yes and no, variable 2), is there a relationship between these two variables? What is the likelihood ratio?
- 2) Answer question 1 for women.

We also know that you're not a child. (Are you?) Let's look at survival rates for children versus adults (again ignoring the crew).

Age	Survived		Sum
	No	Yes	
Child	52	57	109
Adult	765	442	1207
Sum	817	499	1316

I have added the marginal sums to this contingency table to make the calculations go a bit faster. We now see that 52 of 109 children will die (47.7%), while 765 of 1207 adults will die (63.4%), not as much of a difference as one might have hoped for, given the "women and children first" policy.

Let's condition your risk of dying on the fact that you are an adult, ticket-holding passenger and look at the risk contingent upon gender.

Sex	Survived		Sum
	No	Yes	
Male	659	146	805
Female	106	296	402
Sum	765	442	1207

Calculating the relative risk of dying for adult, male, ticket-holding passengers compared to adult, female, ticket-holding passengers (do it!), we find it is 3.10, still about three, but a little worse for you if you're male.

One of the richest men in the world, John Jacob Astor, is traveling with you. All his money will do him no good. He is destined to perish. In general, however, passengers traveling on a first class ticket will fare better than those traveling in second or third class. Let's look at survival by class of ticket held. (Crew will be excluded since they do not hold tickets, but children will be included.)

Class	Survived		Sum
	No	Yes	
1st	122	203	325
2nd	167	118	285
3rd	528	178	706
Sum	817	499	1316

Once again, the marginal sums are included to spare you some computational work. We see that the risk of dying for a first class passenger is $122/325 = .375$, for a second class passenger is $167/285 = .586$, and for a third class passenger is $528/706 = .748$. If you are a second class passenger, you are $.586/.375 = 1.56$ times as likely to die as a first class passenger. If you are a third class passenger, you are 1.99 times as likely to die (or twice as likely) as a first class passenger. Several questions arise.

- 1) Why do you think this relationship between class and survival exists?
- 2) There are more third class passengers than first and second combined. Why do you think this is? (Lesson: know your data!) Do you think this has an impact on the calculations? Or have we somehow corrected for the number of passengers in each class?
- 3) Do you think these figures would change if we looked only at men? Only at women? Only at adults? Only at children?
- 4) And now that you've answered question 3, another question has become obvious. Why are there so many more children in third class than in first and second class?
- 5) Now here's a tough question for you. Were there any children in the crew? Be careful how you answer!

Here are some figures that might help you to answer these questions. This is the complete Board of Trade data broken down by class, gender, age, and survival. The

last column gives the frequencies in each of these multivariate categories.

	Class	Sex	Age	Survived	Freq
1	1st	Male	Child	No	0
2	2nd	Male	Child	No	0
3	3rd	Male	Child	No	35
4	Crew	Male	Child	No	0
5	1st	Female	Child	No	0
6	2nd	Female	Child	No	0
7	3rd	Female	Child	No	17
8	Crew	Female	Child	No	0
9	1st	Male	Adult	No	118
10	2nd	Male	Adult	No	154
11	3rd	Male	Adult	No	387
12	Crew	Male	Adult	No	670
13	1st	Female	Adult	No	4
14	2nd	Female	Adult	No	13
15	3rd	Female	Adult	No	89
16	Crew	Female	Adult	No	3
17	1st	Male	Child	Yes	5
18	2nd	Male	Child	Yes	11
19	3rd	Male	Child	Yes	13
20	Crew	Male	Child	Yes	0
21	1st	Female	Child	Yes	1
22	2nd	Female	Child	Yes	13
23	3rd	Female	Child	Yes	14
24	Crew	Female	Child	Yes	0
25	1st	Male	Adult	Yes	57
26	2nd	Male	Adult	Yes	14
27	3rd	Male	Adult	Yes	75
28	Crew	Male	Adult	Yes	192
29	1st	Female	Adult	Yes	140
30	2nd	Female	Adult	Yes	80
31	3rd	Female	Adult	Yes	76
32	Crew	Female	Adult	Yes	20

C. Correcting Calculations For Group Sizes

It is customary (although not required) to set up a contingency table so that the explanatory variable, if there is one, is in the rows and the outcome variable is in the columns. The survival-by-class table above was set up this way.

Class	Survived	
	No	Yes
1st	122	203
2nd	167	118
3rd	528	178

If you want to see if there is a difference in survival by class, you cannot just look at numbers in one of the columns, for example, the survived=No column. This will work only if there are equal numbers of people in each of the rows. Since there are not, it is necessary to correct for that when looking for effects by calculating proportions (likelihoods) or percentages. In this case, it would make the most sense to calculate them across the rows, relative to the row sums.

Class	Survived	
	No	Yes
1st	0.375	0.625
2nd	0.586	0.414
3rd	0.748	0.252

Notice that across the rows of this table the proportions add up to one. (Why?) We can see clearly in this table what the effect is of class on survival without the complicating factor of there being more third class passengers than anyone else. Calculating a proportion (or percentage) is one way of correcting for differing group sizes.

Groups can then be compared by comparing the sizes of these proportions. A higher proportion of second class passengers died than first class passengers. We can see that in the difference between the proportions.

$$0.586 - 0.375 = 0.211$$

We can also see it in the ratio of the proportions (the likelihood ratio).

$$0.586 / 0.375 = 1.56$$

Question: In what fundamental way do these two calculations differ? (Hint: the difference between proportions of 0.212 and 0.001 is the same as above.)

Graphing.

Contingency Tables in R.

Exercises.

Here are some data collected from 592 statistics students at the University of Delaware (reported 1974). Three variables were recorded:

- hair color - in the rows of the table
- eye color - in the columns of the table
- sex (or gender) - represented here as two tables

Use these data to determine if any two of these variables are related.

, , Sex = Male

Hair	Eye			
	Brown	Blue	Hazel	Green
Black	32	11	10	3
Brown	53	50	25	15
Red	10	10	7	7
Blond	3	30	5	8

, , Sex = Female

Hair	Eye			
	Brown	Blue	Hazel	Green
Black	36	9	5	2
Brown	66	34	29	14
Red	16	7	7	7
Blond	4	64	5	8