

Chapter 5 The Space Shuttle Challenger: Binary Response Data

5.1 Introduction On January 28th, 1986 the National Aeronautics and Space Administration (NASA) space shuttle program launched its 25th shuttle flight from Kennedy Space Center in Florida. Seventy-three seconds into the flight, a fuel tank on the space shuttle Challenger exploded, killing all seven astronauts on board.

Take a few minutes to view the BBC video in the link at the bottom of the page. Several investigations were held, reports to President Reagan and videos of the event are also available at the [Kennedy Space Center](http://www.nasa.gov/ksp) website.

Investigations showed that a small O-ring seal in the right solid rocket booster failed to isolate the fuel supply. Figure 5.2 below shows a diagram of the 149.16 foot long rocket booster. Because of its size, the rocket booster was created and shipped in three sections.

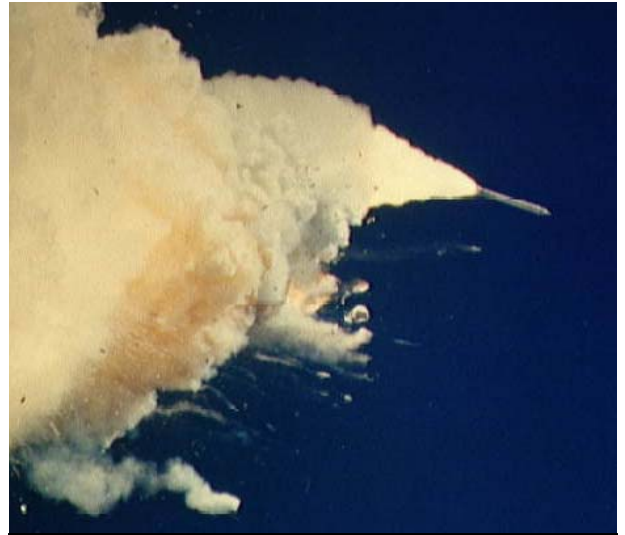


Figure 5.1 One of the solid rocket boosters of the Space Shuttle Challenger 76 seconds after ignition.

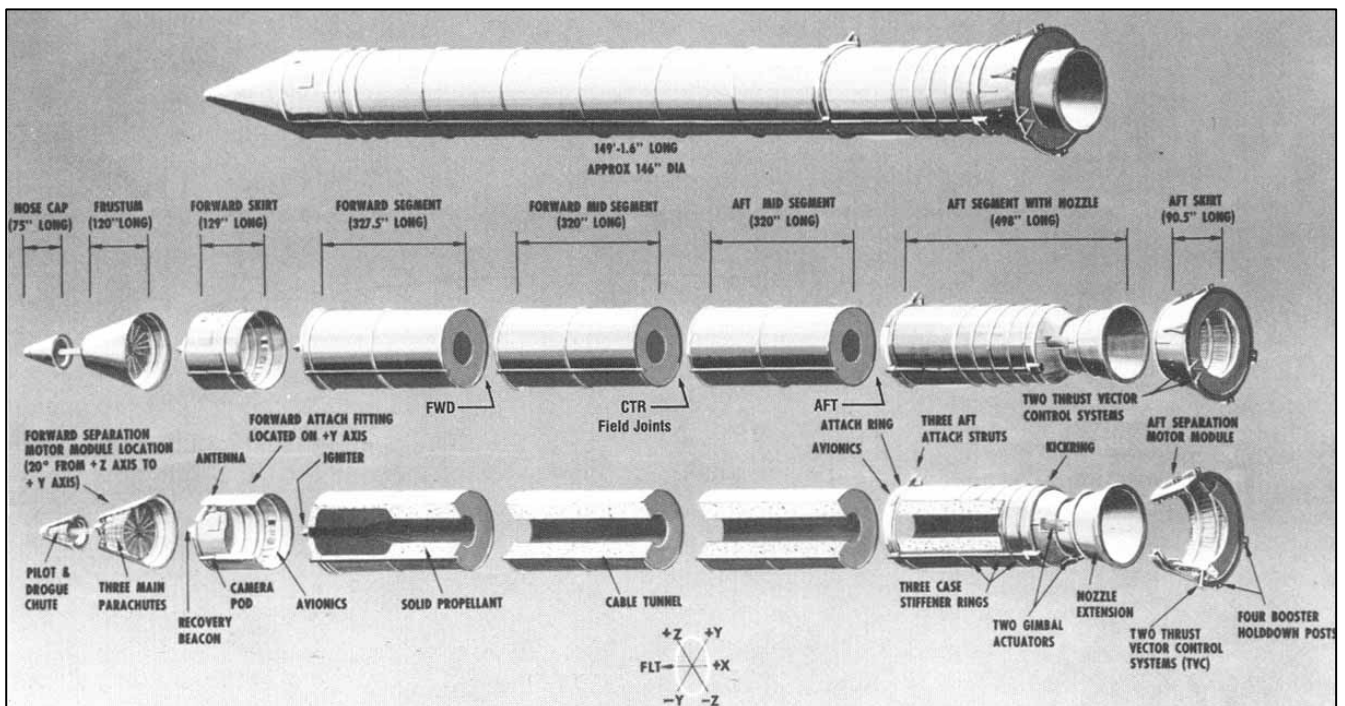


Figure 5.2 diagram of space shuttle Challenger rocket booster

Video Link: http://news.bbc.co.uk/onthisday/hi/dates/stories/january/28/newsid_2506000/2506161.stm

Figure 5.3 shows the placement of two O-rings, 0.28 inches in width but 37 feet in diameter, that were used to seal each of these sections. When the O-rings failed to seal properly, 5800°F gasses escaped and burned a hole through the side of the rocket booster.

The Report of the Presidential Commission on the Space Shuttle Challenger Accident, also known as the Rogers' Commission Report, states, "O-ring resiliency is directly related to its temperature.. A warm O-ring that has been compressed will return to its original shape much quicker than will a cold O-ring when compression is relieved. ... A compressed O-ring at 75 degrees Fahrenheit is five times more responsive in returning to its uncompressed shape than a cold O-ring at 30 degrees Fahrenheit.... At the cold launch temperature experienced, the O-ring would be very slow in returning to its normal rounded shape. ... It would remain in its compressed position in the O-ring channel and not provide a space between itself and the upstream channel wall. Thus, it is probable the O-ring would not... seal the gap in time to preclude joint failure due to blow-by and erosion from hot combustion gases... Of 21 launches with ambient temperatures of 61 degrees Fahrenheit or greater, only four showed signs of O-ring thermal distress; i.e., erosion or blow-by and soot. Each of the launches below 61. degrees Fahrenheit resulted in one or more O-rings showing signs of thermal distress."

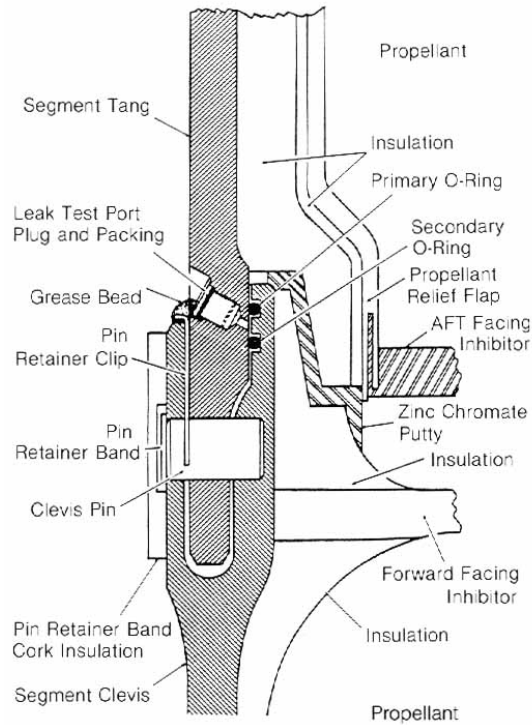


Figure 5.3 Solid Rocket Motor cross section shows positions of O-rings. Putty, which also was part of problem, lines the joint on the side toward the propellant.

A lamentable aspect of this disaster was that the problem with the O-rings was already understood by some engineers. In February of 1984, the Marshall Configuration Control Board sent a memo about their concern of the O-ring erosion that occurred on STS 41-B (space shuttle flight #10). These messages continued to increase in intensity as evidenced by the 1985 internal memo from Thiokol Corp., the company that designed the O-ring. Employees from Thiokol wrote the following to their Vice President of Engineering, "This letter is written to insure that management is fully aware of the seriousness of the current O-Ring erosion problem in the SRM joints from an engineering standpoint..."

With the temperature on January 28th expected to be 31° F, Thiokol Corp. recommended against the launch. However, this flight was getting significant publicity because a high school teacher, Christa McAuliffe, was on the flight. The flight had already been delayed several times and there was no quick solution to the O-ring concern. The engineers were overruled and the decision was made to go ahead with the flight.

The presidential investigation states that, "The decision to launch the Challenger was flawed. Those who made that decision were unaware of the recent history of problems

concerning the O-rings and the joint and were unaware of the initial written recommendation of the contractor advising against the launch at temperatures below 53 degrees Fahrenheit and the continuing opposition of the engineers at Thiokol after the management reversed its position. They did not have a clear understanding of Rockwell's concern that it was not safe to launch because of ice on the pad. If the decision makers had known all of the facts, it is highly unlikely that they would have decided to launch 51-L on January 28, 1986."

So, it seems that even if the engineers did comprehend the severity of the problem, they were unable to properly communicate the results.

Data for the 24 shuttle launches prior to the ill-fated Challenger flight are given below. Flight 4 has a missing data point because the rockets were lost at sea.

Table 5.1: O-Ring Failures of 24 Space Shuttle Launches

Flight Number	Date	Temperature (F)	O-Ring Failures
1	04/12/81	66	0
2	11/12/81	70	1
3	03/22/82	69	0
4	06/27/82	80	*
5	11/11/82	68	0
6	04/04/83	67	0
7	06/18/83	72	0
8	08/30/83	73	0
9	11/28/83	70	0
10	02/03/84	57	1
11	04/06/84	63	1
12	08/30/84	70	1
13	10/05/84	78	0
14	11/08/84	67	0
15	01/24/85	53	1
16	04/12/85	67	0
17	04/29/85	75	0
18	06/17/85	70	0
19	07/29/85	81	0
20	08/27/85	76	0
21	10/03/85	79	0
22	10/30/85	75	1
23	11/26/85	76	0
24	01/12/86	58	1

1) Based on the description of the Challenger disaster, identify which variable is the explanatory variable and which is the response variable.

2) Imagine you were an engineer working for Thiokol Corp. Try creating a few graphs of the data. Is it obvious that temperature impacts success of the O-rings? Share with the class any charts/graphs that you have created that show the relationship between temperature and O-ring failure.

Note that this data is in a slightly different format than you may have seen before. In this example we want to determine if ambient temperature (quantitative data) describes success or failure of the O-rings (categorical data). Most introductory classes focus on sample data where the response is quantitative. However, there are many situations where the responses only fit into categories, such as names, ranks, yes/no, or pass/fail. It is no longer appropriate to focus on means, but on the percent of responses that fall into each category.

[Key Concept: Analysis of categorical response variables requires using proportions instead of means.]

The type of data collected usually determines which type of statistical analysis should be conducted.

Table 5.2: Appropriate techniques for various data types

		Explanatory Variable	
		Categorical	Quantitative
Response Variable	Categorical	Chi-Square test z-test for proportions Fishers exact test	Logistic Regression
	Quantitative	t-test for means ANOVA	Regression

3) Based on Table 5.2, what type of analysis should be conducted on this data set?

In this chapter we will discuss multiple techniques to analyze data with categorical response variables. Later chapters will use logistic regression to analyze the data in Table 5.1. Before we study logistic regression, we will review data analysis techniques when both the explanatory and response variables are categorical.

5.2 Data Analysis for Tables of Categorical Data

To analyze Table 5.1 data in this section, we need to take the quantitative temperature measurements and convert them to categorical temperature measurements. Remember that any time you lose detail, you are also losing information that may be helpful in drawing conclusions with your data. However, some loss of information may make the

analysis much simpler to understand and analyze. This again shows that there is often not just one “best” technique to analyze a data set, just guidelines to ensure that our statistical model is appropriate.

5.2.1 Developing categories

4) Using Thiokol’s recommendation, create a 2-by-2 table where the response is whether or not an O-ring failed and the explanatory variable whether the temperature is 53 degrees or less vs. higher than 53 degrees. Do you see any problems with conducting a statistical analysis of this data using 53 degrees as the breaking point for temperature? HINT: How would you estimate variability of the 53 and under group?

It is not clear how the temperature variables should be categorized. However, there are *two wrong techniques*: 1) Do not look at the data and try to find where the failures occur to determine the categorization and 2) we should ensure there are enough observations in each group to obtain reliable information.

[Key Concept: Looking at the data to identify how to categorize variables will cause false positives in your analysis. In essence, you will be much more likely to conclude there is a difference between groups when there really is not any difference.]

Looking at the temperatures (not the response), there are gaps between 63 and 66 as well as between 70 and 72. Since the 70/72 gap will more equally divide the observations, you may choose to use 71 degrees as the breaking point, though several other choices for breaking points would also be reasonable.

Table 5.3 below is a 2-by-2 table where the response is whether an O-ring succeeded or failed and the explanatory variable whether the temperature is 71 degrees or less vs. higher than 71 degrees.

Table 5.3: 2-by-2 Table of O-Ring Successes for 24 Space Shuttle Flights

	C1-T	C2	C3
		Low Temp	High Temp
1	Success	8	8
2	Failure	6	1

5) Create a 2-by-3 table where the response is whether an O-ring failed and the explanatory variable whether the temperature is below 65 degrees, 66 to 71 degrees and higher than 71 degrees. Place the response variables in the rows.

5.2.2 Odds and Rates

6) Table 5.3 can be used to calculate the rate (also called proportion) of failures for the below 71 degrees group. Six failures out of 14 flights equates to a failure rate of 42.86%. Calculate the rate of failures for the above 71 degrees group. Does it appear that more O-ring damage occurred at lower temperatures?

7) Create a segmented bar graph to visualize the difference in these rates. [In Minitab, enter Table 5.3. Graph>>Bar Chart>>Select “Values from a table” and the Two-way table Stack option. O.K. Graph variables: C2 C3, Row Labels: C1. Select Columns are outermost categories and Stack innermost category. In Bar Chart Options select the Default, Show Y as Percent, and Within Categories at level 1. O.K. You may also choose to show the actual percentages on the graph by selecting Label, Data Labels and Use y-value labels.] Notice that the explanatory variable is on the X axis and the response variable is on the Y axis.

8) Calculate the rates and a segmented bar graph for the table in question 5).

At this point it seems reasonable to question why the O-rings were not considered a higher risk. 42.86 % of cold temperature launches and 11.11% of warm temperature launches had O-ring damage. These previous launches did not result in the same disaster as the Challenger launch because the O-rings only showed “minor” damage. It wasn’t enough for gas to escape, only an indicator that the O-rings may not be as resilient as expected. Some investigators compared this to observing a major bridge beam showing cracks, but not enough for the bridge to collapse. At what point should bridge engineers consider cracks in a beam a “minor” issue because the bridge is still standing?

Two calculations that are common with this type of data are relative risk and odds ratio.

$$\text{Relative Risk} = \frac{\text{proportion of successes in group1}}{\text{proportion of successes in group2}}$$

$$\text{oddsratio} = \frac{\text{odds of success in group1}}{\text{odds of success in group2}} = \frac{\frac{\text{number of successes in group1}}{\text{number of failures in group1}}}{\frac{\text{number of successes in group2}}{\text{number of failures in group2}}}$$

In relative risk (and odds ratio) calculations, the group that has the lowest proportion (or lowest odds) is typically considered *group 2* in the denominator.

9) Calculate the relative risk and odds ratio for the data in Table 5.3.

10) Use the table in Question 5) to calculate the odds ratio of O-ring failure when the temperature is lower than 65 degrees (group the columns to represent 65 degrees or higher).

11) Explain the statement from the presidential report in terms relative risk. “A compressed O-ring at 75 degrees Fahrenheit is five times more responsive in returning to its uncompressed shape than a cold O-ring at 30 degrees Fahrenheit.”

5.2.3 Simulations: How likely is it that the observed sample would occur by chance?

In the context of this study, we are trying to determine if we are confident that high and low temperatures in fact have a different rate of O-ring failures. To formulate this into a hypothesis test we can write:

$$H_0: p_L \leq p_H \quad \text{vs.} \quad H_A: p_L > p_H$$

where p_L is the true rate of failure (% of failure) for the population of all possible low temperature launches and p_H is the true rate of failure (% of failure) for the population of all possible high temperature launches.

In order to draw conclusions about this hypothesis test, we will need to determine a p-value. A **p-value** is the likelihood of observing differences in the sample rates at least this extreme when there really is no difference in population failure rates. In this study a p-value determines if:

- 1) failure rates are equally likely in both temperature groups
- 2) a total of 7 failures were observed, and
- 3) a total of 14 low temperature and 9 high temperature launches took place,

then, how likely would it be that 6 or more of the failures occurred in the low temperature group.

We can simulate taking samples under conditions 1, 2 and 3 above many times and observe the percentage of times a sample results in 6 or more failures in the low temperature group.

12) Take 23 index cards to represent this sample. On 16 of the cards write “success”, on 7 of the cards write “failure”. Shuffle the cards and randomly select 14 cards. These 14 cards can represent the Low Temperature category. Did you observe 6 or more failures in this simulation?

13) Repeat this process 9 more times recording the number of failures you sampled in the Low Temperature group. Comparing your answers with the rest of the class, does it seem likely that 6 or more failures would occur in the Low Temperature group simply due to chance alone?

While simulations can be done by hand with cards, this process is very time consuming. In addition, a large number of simulations are needed to get a true feel of the likelihood of an outcome (many statisticians suggest 10,000 simulations).

14) Instead of randomly sampling 14 out of 23 cards 10,000 times, we will use computer programs to conduct a simulation.

Conduct the following in Minitab: In the first column of the worksheet write 16 0's (representing successes) and 7 1's (representing failures). Click anywhere on the session window (the top half of Minitab) and then select Editor>>Enable Commands. A Minitab prompt, MTB >, will show in the window. Use the pull down tabs to select the following:

```
Calc>>Random Data>>Sample from Columns  
Sample 14 rows from column(s): C1  
Store samples in C3
```

Do not check sample with replacement (Why is this necessary?)]

Notice C3 now has 14 rows of data and the session window now has a small amount of computer code used to run this function. This line of code “Sample 14 c1 c3.” will do the same as the pull down commands. Next, in the pull down menu select:

```
Calc>>Column Statistics, select Sum Input variable: C3.
```

15) How many 1’s (failures) did you find in your computer generated sample? Repeat this process 9 more times. How does this compare to your index card simulation? Would you expect to get exactly the sample number of samples with 6 or more failures? Why or why not?

To repeat this process thousands of times do the following:

In Minitab it is necessary to write a macro. Open up Notepad and type the following code:

```
Sample 14 C1 C3.  
Let C5(k1) = sum(C3)  
Let k1 = k1+1
```

Save the file as “oring sim.mtb”. It is important to include the quotation marks and the .mtb in order for Minitab to recognize the file.

In the Minitab prompt, type `MTB > let k1=1`

File>>Other Files>>Run an Exec, execute the file one time and select “oring sim.mtb”.

Verify that C3 contains 14 observations and C5 is the total number of failures (1’s) in C3.

Repeat the simulation 999 more times.

16) Create a histogram of the 1000 simulations and calculate the number of times 6 or more failures occurred. What is the p-value (i.e. the probability of observing 6 or more failures) in the Low Temperature group? [In Minitab, select Graph >> Histogram >> Simple. O.K. Graph variables: C5. At the prompt type: `let c6=(c5>=6)` on the next line type `MTB > sum c6`]

Using simulations to approximate p-values has many advantages. Often simple programs or macros can be written to simulate thousands of samples. In addition there are no distributional assumptions that need to be made to ensure that the p-value is accurate. Computer programs can often be modified to fit a variety of situations, where theoretical assumptions can be somewhat rigid. However, simulations only provide approximate p-values based on the samples we happened to draw in our simulation study. Increasing the number of simulations will help to improve the accuracy of the p-value. 10,000 simulations usually provides accurate p-values and these programs can be written and run in just a few minutes. While simulations to determine p-values are quickly gaining in popularity, there are also exact tests based on theoretical distributions.

5.2.4 Fisher's Exact Test: How likely is it that the observed sample would occur by chance?

Students who have taken a combinatorics or discrete mathematics course may have already considered that there are ways to count the likelihood of a certain number of observations falling within a certain group. **Fisher's exact test** uses the hypergeometric distribution to calculate exact probabilities in 2-by-2 tables. More details on how to derive the hypergeometric distribution and Fisher's exact test is provided in the appendix.

The hypergeometric distribution is given by the following formula:

$$f(x) = P(X = x) = \frac{\binom{M}{x} \binom{N-M}{n-x}}{\binom{N}{n}} \quad \text{where} \quad \binom{M}{x} = \frac{M!}{x!(M-x)!} \quad (5.2.1)$$

M choose x , written $\binom{M}{x}$, is the number of ways that x items can be chosen from a group

of size M . For example, $\binom{7}{6} = \left(\frac{7!}{6!(7-6)!}\right) = \left(\frac{7*6*5*4*3*2*1}{6*5*4*3*2*1(1)}\right) = 7$ counts the number of distinct groups of 6 items that can be selected from a group of 7 items. The table below identifies the location of each element of the hypergeometric probability calculation.

Table 5.4: Finding appropriate elements of the Hypergeometric distribution for 2x2 table.

	Low Temp	High Temp	
Success	8	8	16
Failure	6 = x	1	7 = M
	14 = n	9	23 = N

In the space shuttle example, the p-value is the probability that 6 or more failures occur in the Low Temperature group, assuming temperature is unrelated to the probability of o-ring failure

$$p\text{-value} = P(X = 6) + P(X = 7)$$

where

$P(X = 6) = \{\text{number of ways to assign 6 failures to the Low Temp group}\} * \{\text{number of ways to assign 8 successes to the Low Temp group}\} / \{\text{number of ways to assign 14 temperatures to the Low Temp group}\}.$

$$= \frac{\binom{7}{6} \binom{16}{8}}{\binom{23}{14}} = \frac{\left(\frac{7!}{6!(7-6)!}\right) \left(\frac{16!}{8!(16-8)!}\right)}{\left(\frac{23!}{14!(23-14)!}\right)} = \frac{\left(\frac{7}{1}\right) \left(\frac{16*15*14*13*12*11*10*9}{8*7*6*5*4*3*2*1}\right)}{\left(\frac{23*22*21*20*19*18*17*16*15}{9*8*7*6*5*4*3*2*1}\right)} = .11024$$

Software is often used to calculate probabilities for the hypergeometric distribution. [In Minitab Calc>> Probability Distributions>>Hypergeometric, select Probability, Input N, M and n. Select Input constant: 6.]

17) Use Equation 5.2.1 to show that $P(X=7) = 0.013999$, check your answer with computer software.

[In Minitab, the Cumulative probability tab can be used to find $P(X \geq 6)$, but note that since these are discrete outcomes, $P(X \geq 6) = 1 - P(X \leq 5)$.]

Thus the exact p-value for this problem is $P(X \geq 6) = 0.124243$. If there truly was no difference between failure rates at high or low temperatures (which would mean our null hypothesis H_0 was true), random sampling would produce this outcome (6 or more failures in Low Temp) 12.42% of the time. Since this outcome occurs more than 1 in 10 times, we do not have significant evidence that H_0 should be rejected. Thus we fail to reject H_0 .

18) Using the data in Table 5.4, find the probability of 1 or fewer failures in the High Temp group. Explain why this number looks familiar.

19) Calculate the Fisher's exact test assuming Low Temp was determined as lower than 65 degrees. If you had chosen this as your breaking point what would you conclude? Why is it wrong to search for a breaking point that is more likely to give a small p-value?

Before leaving the topic of categorical data tables, it is important to mention two other tests frequently used with this type of data; the Chi-Square Test and the 2-sample z-test for proportions. You may recall in previous statistics courses that both the chi-square test and the 2-sample z-test required that certain assumptions were met before any analysis was done. For example, large enough sample sizes are needed and the proportions should not be too close to 0% or 100%. These assumptions are necessary because both the chi-square and 2-sample z-tests are approximate tests. Large enough sample sizes are needed before the distributions of the test statistics start resembling either the chi-square or the normal distributions.

Fisher's exact test gives an exact p-value for any hypothesis test with 2-by-2 data. However, in some situations confidence intervals are needed. If the sample size is large, often confidence intervals may be more easily obtained using a z-test for equal proportions or a chi-squared test. Also, as shown in the next section, the chi-square test can be easily extended to more than two rows and more than two columns.

[Key Concept: Fisher's exact test is based on the hypergeometric distribution and provides exact p-values for any sample sizes and any proportions. Fisher's exact test does not have any distributional assumptions.]

[Key Concept: In a 2-by-2 table, Fisher's exact test, a simulation study, the chi-square test, and the 2-sample z-test for proportions are all addressing the same question: Does our sample provide enough evidence to conclude that the observed difference in proportions is not just due to chance?]

5.3 Chi-Square Test: How likely is it that the observed sample would occur by chance?

The steps needed to conduct a chi-square test are also the same as previous hypothesis tests:

- 1) **State the null and alternative hypothesis:** related to the space shuttle problem
 $H_0: p_L \leq p_H$ vs. $H_A: p_L > p_H$
- 2) **Graph the data and check technical assumptions:** A table of expected counts (labeled exp below) are calculated by (row total*column total/table total). To use the chi-square test, the sample size needs to be large enough that the expected counts of each cell are at least 5.
- 3) **Calculate the test statistic:** Expected counts (exp below) are calculated by (row total*column total/table total)

$$X^2 = \sum \frac{(obs - exp)^2}{exp}$$

- 4) **Calculate the p-value:** Using statistical software or the chi-square table with (# rows-1)*(# columns-1) degrees of freedom.
- 5) **State your conclusions.**

Cocaine Example: To reduce the level of addiction to cocaine, a standard drug, Disipramine, is often given to patients in treatment centers. However, new research has shown that many patients who start using cocaine again after leaving a treatment center have severe depression symptoms. The following study was done to determine if an antidepressant, called Lithium, could be used to reduce the likelihood of patients using cocaine again. 72 patients were selected for this study. 24 patients were randomly assigned to each of three treatment groups: a disipramine group, a lithium group, and a placebo group. After 6 months these patients were tested to determine if they were successful in stopping their cocaine habit. The data from the study is located below:

<i>Observed</i>	Success	Failure	
Disipramine	14	10	24
Lithium	6	18	24
Placebo	4	20	24
	24	48	72

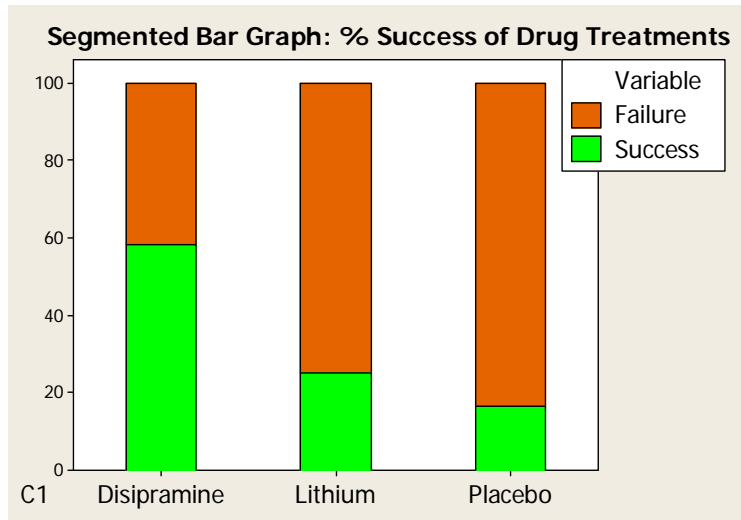
Our hypothesis tests for this study are:

H_0 : All treatment groups have the same success rates. This can also be written as

$H_0: p_D = p_L = p_P$ or H_0 : The type of treatment is independent of success rate.

H_a : At least one treatment group has a different success rate.

To visualize this data: [In Minitab, Graph >> Bar Charts select values from a table and Two-way table Stack. Graph variables: Success Failure, Row Labels C1. Select Rows are outermost categories and columns are innermost and select Stack innermost category. In Bar Chart – Options select Show Y as a Percent and Within categories at level 1. OK. OK]



A table of expected counts is created assuming the null hypothesis is true. In other words, if the treatment groups make no difference, we would expect the success rates of all groups to be the same. Note that all expected counts are greater than 5, so it is appropriate to use the chi-square test.

<i>Expected</i>	Success	Failure	
Desipramine	8	16	24
Lithium	8	16	24
Placebo	8	16	24
	24	48	72

row total*col total/overall total

$$8 = 24 * 24 / 72$$

$$16 = 24 * 48 / 72$$

The Chi-square statistic is then calculated:

$$\frac{(14 - 8)^2}{8} + \frac{(6 - 8)^2}{8} + \frac{(4 - 8)^2}{8} + \frac{(10 - 16)^2}{16} + \frac{(18 - 16)^2}{16} + \frac{(20 - 16)^2}{16}$$

$$= 4.5 + .5 + 2 + 2.25 + .25 + 1 = 10.5$$

The p-value is found by using software or the chi-square table to determine the likelihood of observing a sample statistic this large assuming the null hypothesis is true. This table has three rows and two columns of data. Thus $(\# \text{ rows} - 1) * (\# \text{ columns} - 1) = 2 * 1 = 2$ degrees of freedom are used in looking up the p-value. [In Minitab: Stat >> tables >> Chi-Square Test we find the p-value = 0.005]

20) Create a segmented bar graph of the space shuttle data. Create a table of expected values and explain why a chi-square test is not appropriate. [In Minitab, Stat >> Tables >> Chi-Square Test select Low Temp and High Temp. O.K.] Explain why the p-value for this test is not reliable.

Chapter 5B: A Closer Look at Binary Response Data

The chi-square test is used in several situations:

- 1) **Test for differences in proportions (homogeneity of proportions):** determines if samples from different populations fall into categories at the same rate? The hypothesis test related to the space shuttle problem, $H_0: p_L \leq p_H$ vs. $H_A: p_L > p_H$, fits this situation. If there are only two categories from two populations, this test is equivalent to the z-test for two proportions.
- 2) **Test for independence:** determines if two categorical variables within a population are independent. This tests for an association between the rows and columns of a categorical data table. Population proportions or odds can be tested here. It is not necessary to determine which variable is the explanatory variable and which is the response variable.
- 3) **Tests for goodness of fit:** determines if the proportions observed are equivalent to some hypothetical proportions. This approach assumes some theoretical distribution determines the probability that a sampled unit should randomly fall into a certain category.

The calculation involved in each of these 3 tests are very similar, however the type of question asked will impact the conclusions that can be drawn.

Math and Statistics Example: A study was conducted by introductory statistics students identifying whether majors in the sciences are as likely to take a statistics course as majors in the humanities. They sampled 50 senior students and asked if they have taken (or are scheduled to take) a statistics course before they graduate. Their data is provided below:

	Yes	No
Humanities	23	27
Science Major	32	18

Note that the way the data is collected also impacts the type of test conducted. In **tests of homogeneity**, the intent is to look at two distinct populations, so two samples were taken separately and then placed in categories. In **tests of independence**, samples are taken at one time from one population, then the data is classified into two categories within the one population.

Note that the data tables and the calculations in tests for homogeneity and tests for independence are exactly the same; the only difference is in the statement of the hypothesis and the conclusions.

In this example, we can test if each type of major has an equal proportion of students taking a statistics course. However, since in this test of homogeneity 50 students were selected in each population, it is inappropriate to use this data to test if there are an equal proportion of science and humanities majors.

[KEY CONCEPT: In tests of independence it is reasonable to evaluate percentages in both categories, while tests of homogeneity can only test one category. In some situations, tests of homogeneity may be preferable because it is possible to obtain equal sample sizes from two populations. Equal sample sizes can increase the likelihood of rejecting the null hypothesis. Care must be taken to identify how the data was collected before any conclusions are drawn.]

1) Use the math and statistics student data to determine if humanities and science majors have the same proportion of students who take a statistics course. Assuming that students took care to collect a simple random sample of seniors from each division, what conclusions can be drawn?

2) **Donner Party Example:** In the winter of 1846-1847, members of the Donner party were trapped in the Sierra Nevada Mountains. The data shown in the following table is from www.utahcrossroads.org/DonnerParty/Statistics.htm.

	Male	Female
Survived	23	25
Died	32	9

Plot the data and conduct a chi-square test to determine if gender and survival are independent. Does the p-value show that there was an association between gender and survival rate? If there is an association, can you give a reasonable explanation? Can we conclude that females have stronger survival abilities than males?

3) **Rolling Die Example:** Your friend offers to play a simple game with you. He has a die and will roll it 30 times. Each time he rolls a 5 or 6 you need to pay him \$5. Each time he rolls a 1, 2, 3, or 4, he will pay you \$5. The results are below:

	1	2	3	4	5	6
observed	2	5	3	3	8	9
expected						

If you had agreed to play this game you would have owed your friend some money. Do you have any reason to suspect the die wasn't fair? Explain.

4) In this **goodness-of-fit test** there is no need to use the data to calculate an expected table. If we assume that the die is fair, we can easily determine how many times we would expect each outcome. Fill out the expected row in the rolling die table above, then conduct a chi-square test to determine if there is a significant difference between what was observed and what was expected. In goodness of fit tests such as this where all parameters are well defined, the **degrees of freedom is equal to the number of categories minus 1**. Clearly state your conclusions.

5) The **baby weight** data provides newborn weights of a simple random sample (SRS) of 135 infants born in the United States in 1995. Can we conclude that the population from which this sample came is normally distributed?

- First, find the 10th, 20th, 30th, ... percentiles of the standard normal distribution. For example the 10th percentile of the standard normal distribution is -1.28155.
- Second, standardize the sample data (subtract the mean and divide by the standard deviation).
- Third, place the sample data into categories based on the standard normal percentile groups. For example the first group would count the number of standardized sample observations that are less than or equal to -1.28155.
- Finally conduct a chi-square goodness-of-fit test and clearly state your conclusions.

In the goodness-of-fit tests we are testing the form of the distribution instead of particular parameter values. Before doing many types of hypothesis tests or confidence intervals it is appropriate to determine if the data fits certain model assumptions. Often this includes determining if the data were sampled from a normal distribution.

In the example above, the class sizes (and thus number of classes) were determined rather arbitrarily. The steps in conducting a goodness-of-fit tests are as follows:

- Group the observed Y's into k arbitrarily chosen classes. In the example above we used percentiles of the normal distribution to create classes, but other classes could also be used. Classes must be distinct so that each observation can only fall into one class.
- Calculate the expected frequencies based on some theoretical assumption.
- Calculate the chi-square statistic. If the observed frequencies are very different from the expected frequencies, the test statistic will be large and we can reject the theoretical model.

Note that we never prove that the data follow a normal (or any other theoretical) model. We only prove the model is wrong or fail to reject the model. Care should be taken not to state that we have proven any data follows a particular distribution.

6) Goodness-of-fit tests: parameters unknown Often experimenters believe that data follow some type of distribution, but little if any information exists about the distribution parameters, such as the mean and variance. It is reasonable to estimate any parameters from the sample data. However, *one degree of freedom is lost for each parameter estimated*. By reducing the degrees of freedom, we are less likely to reject the null hypothesis when the alternative is in fact true.

Test the *baby weight* data by classifying the data into groups 6-6.5 ounces, 6.500001 -7 ounces, etc...

- Find the observed frequency for each class.
- Find the expected frequencies for each class. Use the sample mean and variance for the population mean and variance respectively. For example, use the normal distribution to find the $P(6 < Y \leq 6.5)$ and then multiply this probability times the total sample size to find the expected frequency for the 6-6.5 ounce class. Class sizes do not need to be equal. The first and last classes can remain open-ended to accommodate an infinite range and still keep the expected frequencies above 5 in every cell.
- Calculate the test statistic and find the p-value using $6-1-2 = 3$ degrees of freedom.

7) Explain why the chi-square test is typically one-tailed.