Now we learn how to conduct statistical inference when comparing results between two groups, which greatly expands the kinds of questions that we can investigate.

Example 17-1: Dolphin Therapy?
Swimming with dolphins can certainly be fun, but is it also therapeutic for patients suffering from clinical depression? To investigate this possibility, researchers recruited 30 subjects aged 18-65 with a clinical diagnosis of mild to moderate depression. Subjects were required to discontinue use of any antidepressant drugs or psychotherapy four weeks prior to the experiment, and throughout the experiment. These 30 subjects went to an island off the coast of Honduras, where they were randomly assigned to one of two treatment groups. Both groups engaged in the same amount of swimming and snorkeling each day, but one group (the animal care program) did so in the presence of bottlenose dolphins and the other group (outdoor nature program) did not. At the end of two weeks, each subject’s level of depression was evaluated, as it had been at the beginning of the study (Antonioli and Reveley, 2005).

a) Identify the observational units, explanatory variable, and response variable in this study. Also classify the variables as categorical or quantitative.

b) Is this an experiment or an observational study? Explain.

c) Why did the researchers include a comparison group in this study? Why didn’t they just see how many patients showed substantial improvement when given the dolphin therapy?

The researchers found that 10 of 15 subjects in the dolphin therapy group showed substantial improvement, compared to 3 of 15 subjects in the control group.

d) Organize this information into a 2×2 table of counts:

<table>
<thead>
<tr>
<th></th>
<th>Dolphin therapy</th>
<th>Control group</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Showed substantial improvement</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Did not show substantial improvement</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

e) Produce a segmented bar graph to display these data. Calculate the conditional proportion who improved in each group. Also calculate the difference in these two conditional proportions (dolphin group – control group).
Earlier we studied descriptive (graphical, numerical) analyses of two-way table data. Now we will consider how to draw **inferences** about the population from sample data in a two-way table. As with the toddlers/toys study, the key question is whether the observed results would have been unlikely to occur by chance alone.

f) Suppose for the moment that dolphin therapy is **not** effective. Then is it possible that the data would have turned out in favor of the research conjecture as much as the observed data actually did? Explain briefly.

While it is possible that the dolphin therapy is not more effective and the researchers were unlucky and happened to “draw” more of the subjects who were going to improve into the dolphin therapy group, the key question is whether it is **probable**. But if 13 of the 30 people were going to improve regardless of whether they swam with dolphins or not (the **null model**), we would have expected about 6 or 7 of them to end up in each group. The key question is how unlikely it would be for a 10/3 split, or one even more extreme, to occur by this random assignment process alone.

g) Use the “Dolphin Therapy” applet to **simulate** the process of randomly assigning these 30 subjects into the 2 groups, assuming that swimming with dolphins has **no effect** on whether or not the person improves. Conduct 1000 repetitions of this random assignment. What proportion of these 1000 random assignments produced a result as (or more) extreme as the actual research study (10 improvers in the dolphin group)? (This proportion is an **approximate p-value**.)

h) Based on your answer to g), does the simulation analysis suggest that the observed result would be surprising if there were really no effect of the dolphin therapy? Explain.

i) Based on your answers to g) and h), summarize your conclusion about whether dolphin study is effective for relieving depression symptoms. Also describe the reasoning process that leads to your conclusion. Also address the issue of whether a cause/effect conclusion is justified.

- Notice that the reasoning process here is the same as with the study about infant toy preferences.
- Another advantage of random assignment is that it enables statisticians to determine whether an observed effect is **statistically significant**, meaning unlikely to occur by chance variation alone if there were no effect of the explanatory variable on the response.
This analysis is called simulating a **randomization test**.
- We replicate the random assignment process a large number of times, assuming that there’s really no effect of the explanatory variable on the response.
- If the observed result turns out to occur rarely in the simulation, then the observed result is statistically significant and provides strong evidence that the explanatory variable does cause a change in the response.
- Next we’ll study more conventional procedures, based on the normal distribution, of hypothesis tests and confidence intervals for comparing proportions between two groups.

**Example 17-2: AZT for HIV**

One of the first studies aimed at preventing maternal transmission of AIDS to infants gave the drug AZT to pregnant, HIV-infected women in 1993. Roughly half of the women were randomly assigned to receive the drug AZT, and the others received a placebo. The HIV-infection status came to be known for 363 babies, 180 from the AZT group and 183 from the placebo group. Of the 180 babies whose mothers had received AZT, 13 were HIV-infected, compared to 40 of the 183 babies in the placebo group.

a) Identify the explanatory and response variables in this study. Are they both binary?

b) Organize these data into a two-way table, with the explanatory variable in columns and the response in rows.

c) What proportion of the women given AZT had HIV+ babies? What about for the women in the control group? Use appropriate symbols. Does this difference appear to be large?

We will learn a procedure for testing whether two sample proportions differ significantly:

- The null hypothesis is $H_0$: $\pi_1 = \pi_2$
- The alternative hypothesis is $H_a$: $\pi_1 > \pi_2$ or $H_a$: $\pi_1 < \pi_2$ or $H_a$: $\pi_1 \neq \pi_2$.
- The test statistic is $z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_c (1 - \hat{p}_c) \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$.

- The $p$-value is the area under the standard normal curve more extreme than the test statistic, in the direction indicated by the alternative hypothesis.
- The test decision is to reject $H_0$ whenever $p$-value $\leq \alpha$ (significance level).
- The technical conditions required for this procedure to be valid are:
  - independent random samples from the populations, or randomly assigned groups
  - at least 10 “successes” and at least 10 “failures” in each group
d) State the null and alternative hypotheses for the AZT study, in symbols and in words.

e) What proportion of the mothers (for the two groups combined) had HIV+ babies?

f) Calculate the test statistic and \( p \)-value.

g) Use Minitab (Stat> Basic Statistics> 2 Proportions…) or the Theory-Based Inference applet to confirm these calculations.

h) What conclusion would you draw from this test? Explain how this conclusion follows from your test result.

i) Can you legitimately draw a cause-and-effect conclusion between AZT and the baby’s HIV status? [Hint: Was this an observational study or a randomized experiment?]

Note that the hypothesis test assesses how much evidence the data provide that AZT and placebo produce different results, but it does not say anything about how different the results are. To address this question, we can estimate the difference \( \pi_1 - \pi_2 \) with a confidence interval:

\[
(\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}.
\]
j) Produce a 95% confidence interval for estimating the difference in HIV+ rates between the two groups $\pi_{\text{plac}} - \pi_{\text{AZT}}$.

k) Interpret this interval. Pay particular attention to whether it includes zero or is entirely positive or entirely negative.

l) What would be different if you instead found a 95% confidence interval for $\pi_{\text{AZT}} - \pi_{\text{plac}}$?

Example 17-3: Murderous Nurse?
For several years in the 1990s, Kristen Gilbert worked as a nurse in the intensive care unit (ICU) of the Veteran’s Administration hospital in Northampton, Massachusetts. Over the course of her time there, other nurses came to suspect that she was killing patients by injecting them with the heart stimulant epinephrine. Gilbert was eventually arrested and charged with these murders. Part of the evidence presented against Gilbert at her murder trial was a statistical analysis of more than one thousand 8-hour shifts during the time Gilbert worked in the ICU. The following table displays these data:

<table>
<thead>
<tr>
<th></th>
<th>Gilbert working on shift</th>
<th>Gilbert not working on shift</th>
</tr>
</thead>
<tbody>
<tr>
<td>Death occurred on shift</td>
<td>40</td>
<td>34</td>
</tr>
<tr>
<td>Death did not occur on shift</td>
<td>217</td>
<td>1350</td>
</tr>
</tbody>
</table>

a) Identify the observational units, explanatory variable, and response variable for these data.

b) Calculate the conditional proportion of death occurring in each group.

c) Calculate the relative risk of death occurring on a Gilbert shift vs. a non-Gilbert shift. Also write a sentence or two interpreting this relative risk.

d) Do you think that Gilbert and her attorneys would be pleased with your calculations thus far? Explain.
One argument that Gilbert’s attorneys could make in her defense is a “luck of the draw” argument. They could claim that such a big difference in conditional proportions of death, and such a large value for the relative risk, could have happened just by random chance, even if there were no connection between Gilbert and the deaths.

e) Calculate the test statistic (z-score) for a two-proportion z-test to compare the proportions of shifts in which death occurred between the two groups. Show how to calculate this test statistic (z-score) by hand. Also interpret the value of the test statistic.

f) Is the observed difference in sample proportions statistically significant at the $\alpha = .001$ significance level? Explain how you know.

g) Respond to a jury member, who has probably never taken a statistics course, who asks “what is the p-value here, and what is it the probability of?”

h) Does your analysis provide very strong evidence against “luck of the draw” as an explanation for the observed association between Gilbert and death as revealed in the table of data? Explain.

i) Does your analysis provide strong evidence that Gilbert caused the deaths? If so, explain why. If not, identify a potential confounding variable and explain how it could be confounding (and therefore provide an alternative explanation for the observed association between Gilbert and death).
Example 17-4: Pet CPR
A recent survey of pet owners, found that 53% of cat owners and 63% of dog owners said that they would perform CPR on their pets in the event of a medical emergency.

a) Are these numbers parameters or statistics? What symbols would we use to represent these numbers?

b) State the appropriate null and alternative hypotheses for testing whether the difference between 53% and 63% is statistically significant in this context. Use appropriate symbols, and also explain what the symbols represent in this context.

c) What additional information would you need in order to conduct a test of these hypotheses?

Suppose for now that the sample sizes had been 100 in each group.

d) Calculate the z-test statistic by hand. Also determine the p-value. Would you reject the null hypothesis at the .05 significance level?

e) Determine by hand a 95% confidence interval for comparing the two proportions. Also write a sentence or two interpreting this interval.

f) Are the test decision and confidence interval consistent with each other? Explain how you can tell.

g) Now suppose that the sample sizes had been 500 in each group. Calculate the z-test statistic, p-value, and confidence interval. Feel free to use software.

h) Summarize what this example reveals about the role of sample size in determining whether a difference between two sample proportions is statistically significant.