We now begin to consider studies with two variables, where the goal is to compare results between two groups.

Variables are classified not only by their type (categorical vs. quantitative), but also according to the role that they play in a statistical study.

- The variable whose effect you want to study is called the **explanatory variable**.
- The variable that measures the outcome of interest, that you suspect might be affected by the other, is the **response variable**.

- In an **observational study**, researchers observe individuals and measure variables of interest but do not attempt to influence responses.
  - The explanatory variable is not imposed by the researchers.
  - The goal is to *describe* the situation and perhaps discover *association* between variables.

**Example 9-1: Night Lights and Near-Sightedness**
A recent study classified children according to their eyesight (near-sighted, far-sighted, or normal vision) and according to how much light their parents used in their bedroom prior to the age of two (full light, night light, or no light). Researchers found that children who slept with more light were more likely to be near-sighted than children who slept with less light.

a) Identify the observational units in this study. Also identify and classify the explanatory and response variables.

   Observational units:

   Explanatory variable: Type:

   Response variable: Type:

b) Explain why this is an observational study.

c) Is it reasonable to conclude that sleeping with more light *causes* near-sightedness? If not, suggest an alternative explanation for the observed relationship.
• A **confounding variable** is one whose potential effects on a response variable cannot be distinguished from those of the explanatory variable.
  o A confounding variable is related to both the explanatory and response variable.
  o Because of the potential for confounding variables, one *cannot* legitimately draw **cause-and-effect** conclusions from observational studies.

d) Identify a confounding variable in this study, and explain how this confounding variable is related to both the explanatory and response variable.

---

**Example 9-2: Basketball Sell-Outs**
The 2008-09 Oklahoma City Thunder, a National Basketball Association team in its second year after moving from Seattle, found that their win-loss record was actually worse for home games with a sell-out crowd (3 wins and 15 losses) than for home games without have a sell-out crowd (12 wins and 11 losses). (These data were noted in the April 20, 2009 issue of *Sports Illustrated* in the Go Figure column.)

a) Identify the observational units in this study, and also the explanatory and response variables.

  Observational units:
  
  Explanatory:
  
  Response:

b) Organize the data into a 2×2 table of counts, with the explanatory variable groups in columns:

<table>
<thead>
<tr>
<th></th>
<th>Smaller crowd</th>
<th>Sell-out crowd</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Win</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Loss</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

c) Calculate the proportion of wins for each group. Identify them with appropriate symbols.

  • These are called the **conditional distributions** of the “game outcome” variable for the two different categories of the “crowd size” variable.
The relationship between two categorical variables can be displayed visually in a **segmented bar graph**. Each category of the explanatory variable has a bar with a height of 100%, but these bars contain segments whose length corresponds to the conditional proportions. (Excel calls this a 100% stacked column chart.)

d) Produce a segmented bar graph to display these conditional proportions.

e) Do these proportions, and the segmented bar graph, suggest an *association* (relationship) between the two variables? Explain.

f) Suggest a confounding variable that plausibly explains the observed relationship.

g) What would the segmented bar graph look like if there were *no association* between the two variables?

**Example 9-3: Winter Heart Attacks**

Studies conducted in New York City and Boston have noticed that more heart attacks occur in December and January than in all other months. Some people have tried to conclude that holiday stress and overindulgence causes the increased risk of heart attack.

a) Identify a variable that might be confounded with those of the month variable, providing an alternative explanation for the increased risk of heart attack in December and January.
A more recent study in Los Angeles that revealed a similar finding eliminated one potential confounding variable from consideration.

b) Identify another potentially confounding variable that still pertains to the Los Angeles study.

Example 9-4: 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. Part of the evidence against Gilbert was a statistical analysis of more than one thousand 8-hour shifts during the time Gilbert worked in the ICU (Cobb and Gelbach, 2005). Here are the data:

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

a) Identify the observational units and variables for the data in this table.

b) Were deaths more likely to occur on shifts that Gilbert was working than on shifts when she was not? Calculate and compare the relevant proportions from the two-way table above. Also produce a segmented bar graph to display these data.
c) Calculate the ratio of the proportion of Gilbert shifts with a death to the proportion of non-Gilbert shifts with a death.

- The ratio of conditional proportions is called the **relative risk**.

d) Do the data provide evidence of an association between whether or not Gilbert worked on a shift and whether or not a death occurred on the shift. Justify your answer.

e) Do these data allow you to conclude that Gilbert caused deaths? If not, identify a potential confounding variable.