Now we turn our attention back to studies with a *quantitative* response variable. Previously we used a two-sample *t*-test to compare two groups with a quantitative response. Now we’ll learn a new procedure, called **analysis of variance (ANOVA)**, for comparing *several* groups with a quantitative response. We will concentrate on two things:

- Recognizing situations for which this procedure is appropriate
- Interpreting computer output for this procedure

Keep in mind that a p-value and confidence interval mean the same things they’ve always meant.

**Example 23-1: Employment Discrimination**

In a study of whether physical handicaps affect perceptions of employment qualifications, researchers showed videotapes of an employer interviewing an applicant, but both men were actors following a set script. In the videotapes the “applicant” appeared with different handicaps (none, amputee, crutches, hearing, wheelchair). A group of seventy students were randomly assigned to watch a videotape and evaluate the applicant’s qualifications on a 100-point scale. The study aimed to assess whether the data provide evidence that the mean qualification scores differ significantly among the various handicaps presented.

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

b) Identify the explanatory and the response variable. For each one, indicate whether it is quantitative or categorical.

Analysis of variance (ANOVA) is a widely used technique that applies to situations like this where the explanatory variable is categorical and the response variable is quantitative. If the categorical explanatory variable was also binary, then we could use a two-sample *t*-test to analyze the data. But in this situation we have five categories to compare, and we need a test that will compare all five simultaneously. Before we get to the test, though, we start with graphical and numerical summaries of the data.

c) Examine dotplots and boxplots of the qualification ratings among the five groups (**HandicapApply.mtw**). Then examine summary statistics. Do the groups appear to differ with regard to qualification ratings? Explain.
d) State the null hypothesis (of “no effect”) to be tested, using appropriate symbols. Also describe what the symbols mean.

The key idea of ANOVA is to compare variability between groups to variability within groups. The results are displayed in an ANOVA table that culminates with a test statistic denoted by F.

e) Use Minitab (Stat > Anova > Oneway) to calculate the ANOVA table, F-statistic, and p-value.

f) Would you reject the null hypothesis at the $\alpha = .05$ significance level? Explain.

g) Summarize the conclusion that you would draw from this study and ANOVA output.

h) How would you expect the F-statistic and p-value to change if the sample means were further apart? How would your conclusion change? Explain.

i) How would you expect the F-statistic and p-value to change if the qualification ratings in each group were further apart? How would your conclusion change? Explain.
The **technical conditions** required for the validity of the ANOVA procedure and $F$-test are:
- that the data can be regarded as independent random samples from the populations or as arising from random assignment to treatment groups
- that the underlying populations follow normal distributions
- that the standard deviations of those populations are the same for all groups

To check the normality condition, examine graphical displays of the sample data. To check the equal standard deviations condition, determine whether the ratio of the largest to the smallest sample standard deviation is less than 2.

j) Comment on whether the technical conditions for ANOVA appear to be satisfied for these data.

When an ANOVA $F$-test leads to rejecting the null hypothesis that all groups have the same population mean, the natural question to ask is “which groups’ means differ?” We cannot simply perform a bunch of two-sample $t$-tests, because the probability of making an error goes up for each test that we perform. We need a procedure that will look for differences between pairs of groups with a global error rate of whatever we set the significance level $\alpha$ to be.

There are many different **multiple comparison** procedures. We’ll learn to implement the *Tukey* method using Minitab. This procedure gives a confidence interval for the difference in group population means for all pairs of groups. The confidence level applies to all of the intervals *simultaneously*. The intervals that do not include the value zero suggest a significant difference between the group population means for that pair of groups.

**Example 23-2: Employment Discrimination (cont.)**
For this discrimination study, the sample mean qualification ratings, listed in order from smallest to largest, are:
- Hearing Amputee
- None Wheelchair Crutches

<table>
<thead>
<tr>
<th></th>
<th>Sample mean rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hearing Amputee</td>
<td>40.50</td>
</tr>
<tr>
<td>None</td>
<td>44.29</td>
</tr>
<tr>
<td>Wheelchair</td>
<td>49.00</td>
</tr>
<tr>
<td>Crutches</td>
<td>53.43</td>
</tr>
<tr>
<td></td>
<td>59.21</td>
</tr>
</tbody>
</table>

a) To produce the Tukey intervals, start with **Stat > ANOVA > One-way...**, but select the “Comparisons” option and check the box for the Tukey procedure with a family error rate of 5%. Look through all ten of the confidence intervals that Minitab produces. Record the confidence intervals that do not include zero. Also note which pairs of groups these are.
b) Which pairs of handicap types have significantly different (at the $\alpha = .05$ level) mean qualification ratings? Explain how you can tell this from the Tukey confidence intervals.

c) In the listing of sample means above, underline the groups whose mean qualification ratings do not differ significantly from each other.

**Example 23-3: Hot dogs**

A study compared sodium contents and calories of hot dogs classified as beef, poultry, and meat. The data can be found in the Minitab worksheet *HotDogs.mtw*.

a) Produce comparative dotplots to compare the distributions of calorie amounts among the three types of hot dogs. Then do the same for comparing distributions of sodium contents. Comment on what the graphs reveal.

b) Use Minitab to calculate summary statistics; record them in the table below:

<table>
<thead>
<tr>
<th></th>
<th>Beef</th>
<th>Poultry</th>
<th>Meat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample size</td>
<td>20</td>
<td>17</td>
<td>17</td>
</tr>
<tr>
<td>Sample mean (calories)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample SD (calories)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample mean (sodium)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample SD (sodium)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

c) Which response variable – calories or sodium – do you expect to produce a larger value for the ANOVA F-statistic? Which do you expect to produce a smaller p-value? Explain.
d) Use Minitab to produce output for testing whether the sample data provide evidence of a difference in mean calorie amounts among these three types of hot dogs. State the hypotheses, and report the test statistic and P-value. Also summarize your conclusion.

e) Repeat d) for the sodium contents of the hot dogs.

f) Do you need to rethink your answer to c) in light of the output?

g) If you found a significance difference among the groups for either response variable, use the Tukey procedure to decide which types of hot dogs differ significantly (at the .05 level) from which others. Summarize your findings.

h) Comment on whether the technical conditions for the ANOVA procedure appear to be satisfied with these data.