Chapter 3 continued
Describing distributions with numbers

Measuring center: the mean
The mean $\bar{x}$ of a set $x_1, \ldots, x_n$ of observations is
\[
\bar{x} = \frac{x_1 + \cdots + x_n}{n} = \frac{1}{n} \sum x_i.
\]

Example 1a: Mark McGuire's home run counts for his first 12 major league seasons are given in the accompanying Minitab worksheet. Use Minitab to find the mean number of home run counts.

b. McGuire was injured in 1993 and there was a baseball strike in 1994. Exclude these years as "not full seasons" and compute the mean again. How did the partial seasons affect the mean home run count?

Example 1 shows that the mean is sensitive to the influence of extreme observations. For this reason the mean is not a resistant measure of center.

Measuring center: The median
To find the median $M$ of a set of numbers, first sort the numbers, then find the number in the middle of the list.
• The median of 1, 2, 5, 5, 6 is 5
• The median of 3, 4, 10, 15, 20, 22 is the number halfway between 10 and 15, which is 12.5

Example 2: Find the median number of home runs Mark McGuire hit in his first 12 seasons. Then find the median excluding the two partial seasons. Do the outliers have a great effect on the median?

Comparing the mean and median
Because the median is not changed much by outliers, it is called a resistant measure of center. The last two examples show that low outliers pull the mean below the center. High outliers would pull the mean above the center.

Measuring spread: The quartiles
Usually the center of a distribution does not tell the whole story.
Example 3a: The correct amount of the active ingredient in a pill is 3 mg. In a batch of 1000 pills, the mean amount of active ingredient is 3 mg. Is it safe to sell the pills in this batch?
3b. The Census Bureau reports that in 1997 the median income of American households was $37,005. How many American households are above the poverty level?

We are interested in the *spread* or *variability* of incomes and drug potencies as well as their centers. In general a useful numerical description of a distribution contains the center and spread.

**The quartiles** $Q_1$ and $Q_3$.

To calculate the quartiles of a list $L$ of numbers,

1. Sort the observations
2. $Q_1$ is the median of the numbers to the left of the median of $L$.
3. $Q_3$ is the median of the numbers to the right of the median of $L$.

**Example 4:** Mark McGuire’s and Babe Ruth’s home run counts (arranged in order) are shown below. Find the quartiles of each list.

9, 9, 22, 32, 33, 39, 39, 42, 49, 52, 58

22, 25, 34, 35, 41, 41, 46, 46, 47, 49, 54, 54, 59, 60

**The five-number summary**
The five-number summary of a set of data is the following list of numbers:

<table>
<thead>
<tr>
<th>Minimum</th>
<th>$Q_1$</th>
<th>$M$</th>
<th>$Q_3$</th>
<th>Maximum</th>
</tr>
</thead>
</table>

**Box and whiskers plots**

- A box and whiskers plot is just a pictorial representation of the five-number summary.
- Box and whiskers plots are especially useful when we need to compare two distributions

**Example 5:** Draw side-by-side box-and-whiskers plots of Ruth’s and McGuire’s home run counts. Use these plots to compare the two distributions.