Chapter Objectives:

At the end of this chapter you should be able to:

1) Calculate appropriate numerical summaries of quantitative data to describe center (median, mean, quartiles) and spread (range, interquartile range, standard deviation) [the use of software will be emphasized!]
2) Describe the characteristics of various numerical summaries with emphasis on the affects of outliers
3) Interpret the values of the numerical summaries for a particular data set.
4) Match graphical displays of quantitative data to the values of the summary statistics.
5) Apply graphical and numerical procedures to compare 2 or more sets of data

Throughout the course we will emphasize the paradigm "Think, Show, Tell". The above objectives fit into this paradigm as follows:

"Think" about what numerical summaries of center and spread are appropriate for the data at hand; calculate the values of the numerical summaries to "show" the center and spread.
"Tell" what characteristics of the data are conveyed by the values of the numerical summaries.

Reading Assignment:
Text: Chapter 5.

The Big Picture

Slices of frozen pizza in 4 markets: Denver, Baltimore, Dallas, Chicago

![Pizza Slice Prices](image)

5-number summary: $\text{min} \ Q_1 \ \text{median} \ Q_3 \ \text{maximum}$

1.55 2.51 2.65 2.78 3.40
approximately symmetric so we can use the mean and standard deviation:

$$\bar{y} = 2.66, \ s = 0.24$$

Is $1.55 an outlier or is it just the cheapest pizza slice in these 4 markets?

**Boxplots: a picture of the 5-number summary**

**SUMMARY OF BOX PLOT CONSTRUCTION**

1) draw a single number axis spanning the extent of the data;
2) construct a rectangle (the box) with ends located at $Q_1$ and $Q_3$; mark the location of the median (usually with a "+")
3) fences are determined by moving a distance 1.5(IQR) from each end of the box;
   - lower fence: $Q_1 - 1.5*IQR$
   - upper fence: $Q_3 + 1.5*IQR$
   lines are drawn from each end of the box to the most extreme data values within the fences.
4) include outliers by displaying each data value beyond the fences with a special symbol, like "*"

Different software programs use different symbols in part 4)

**EXAMPLE:** boxplot of 138 pulses

**INTERPRETING BOX PLOTS**

1) Examine length of box; since the ends of the box are at $Q_1$ and $Q_3$, the length of the box is $Q_3 - Q_1 = IQR$; the IQR is a useful indicator of variability and useful for comparing the variability of 2 or more sets of measurements
2) Lengths of lines can be useful indicators of skewness
3) Measurements beyond the fences: fewer than 5% of the measurements should fall beyond the fences, even for very skewed measurements; measurements beyond the fences are possible outliers:
   a. incorrect (observed, recorded, entered)
   b. different population
   c. rare event

**EXAMPLE** 2004 major league baseball salaries

<table>
<thead>
<tr>
<th>Minimum</th>
<th>Q1</th>
<th>Median</th>
<th>Q3</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>300,000</td>
<td>326,200</td>
<td>787,500</td>
<td>3,000,000</td>
<td></td>
</tr>
<tr>
<td>21,726,881</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

IQR = $3,000,000 - $326,200 = $2,673,800 (range of middle 50%)

$1.5*IQR = $4,010,700$

outliers: $Q_1 - 1.5*IQR = $326,200 - $4,010,700$ (effectively $0$)

$Q_3 + 1.5*IQR = $3,000,000 + $4,010,700 = $7,010,700$ (85 outliers)
EXAMPLE (ages of rock concert goers who died from being crushed, 1999-2000)

Comparing Distributions

DataDesk Histograms

annual hurricane frequency

1944–1969

1970–2000
Comparing Groups with Boxplots

EXAMPLE heights of ST101 students by gender

EXAMPLE Pizza Prices in 4 markets
Re-expressing Data to Improve Symmetry

Metric tons of CO2 emissions per 1000 citizens for 175 countries

<table>
<thead>
<tr>
<th>Metric Tons CO2 per 1000 citizens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Countries</td>
</tr>
<tr>
<td>120</td>
</tr>
<tr>
<td>100</td>
</tr>
<tr>
<td>90</td>
</tr>
<tr>
<td>80</td>
</tr>
<tr>
<td>70</td>
</tr>
<tr>
<td>60</td>
</tr>
<tr>
<td>50</td>
</tr>
<tr>
<td>40</td>
</tr>
<tr>
<td>30</td>
</tr>
<tr>
<td>20</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>0</td>
</tr>
</tbody>
</table>

mean = 4.19  median = 1.93

It can be difficult to decide what we mean by the "center" of a skewed distribution

**SOLUTION:** re-express, or transform, the data.

Frequently-used transformations: \( \log(y) \), square root(\( y \)) = \( y^{0.5} \), other powers \( y^\lambda \)
1) the log transformation

Metric Tons CO2 per 1000 citizens

- Mean = 0.160
- Median = 0.287

2) square root transformation ($y^{1/2}$)

Metric Tons CO2 per 1000 citizens

- Mean = 1.66
- Median = 1.39
3) cube root transformation ($y^{1/3}$)

Metric Tons CO2 per 1000 Citizens

Number of Countries

0 0.5 1 1.5 2 2.5 3 3.5
0 10 20 30 40

mean = 1.32  median = 1.24

4) $y^{1/4}$

Metric Tons CO2 per 1000 Citizens

Number of Countries

0 0.5 1 1.5 2 2.5 3
0 10 20 30 40

mean = 1.20  median = 1.18
Interpretation

1) In the logarithm re-expression, what does the value 1.2 actually indicate about the country's CO2 emissions?

2) In the square root re-expression, what does the value 2.5 actually indicate about the country's CO2 emissions?

3) In the y^{1/4} re-expression, what does the value 1.2 actually indicate about the country's CO2 emissions?