**Individuals:** The objects that are described by a set of data. They may be people, animals, things, etc. (Also referred to as **Cases** or **Records**)

**Variables:** The characteristics recorded about each individual.

In a data table, the rows represent each individual/case/record while the columns represent the variables.

When presented with data, ask the key questions:

- **Who?** Who are the individuals? How many individuals?
- **What?** What are the variables? What units? (if applicable)
- **Why?** Do we hope to answer a specific question?

If possible, it is also helpful to know: **When? Where? How?**

**Categorical Variable:** Places an individual into one of several categories.

**Quantitative Variable:** A numerical variable for which arithmetic operations such as adding and averaging make sense.

The **Distribution** of a quantitative variable tells us what values the variable takes and how often it takes these values.

In addition to asking the W questions, you should always make a picture that displays your data.

- Categorical data can be displayed with a **Bar Graph** or **Pie Chart**
  - Show the count or the percent of individuals who fall into each category.

- Bar Graphs: Label axes and title the graph.
  - Scale the vertical axis according to counts.
  - Write category names at equally spaced intervals beneath the x-axis.

- Pie Charts: A central angle determines each piece of the pie.
  - You must include all categories that make up the whole.

- Quantitative data can be displayed with a **Dotplot**, **Stemplot**, **Boxplot**, or **Histogram**.

  **Histogram:**
  - Give each graph a title
  - Give each one of the axes a label
  - Make as neat as possible
  - Divide data values into equal-width piles (called bins)
  - Count number of values in each bin
  - Plot the bins on x-axis
  - Plot the bin counts on y-axis

  **Stemplots:**
  - Picture of Distribution
  - Generally used for smaller data sets
  - Group data like histograms
  - Still have original values (unlike histograms)
  - Two columns.
    - Left column: Stem
    - Right column: Leaf (includes only the final digit)

- **Dotplots:**
  - Pictures the distribution above a number line. Each dot represents an observation from the data. The mode is clearly shown.

**Exercise 1.12 (Pg. 22-23)**

---

How do we choose which measures of center and spread to use?

- The five number summary is usually better to use for skewed distributions or data with strong outliers.
- Use the mean and standard deviation for reasonably symmetric distributions that are free of outliers.
Mean: add the values and divide by \( n \). We call this \( \bar{x} \). \( n \) is the number of observations.

Median: the midpoint of the distribution

**Obtaining 1-Variable Statistics**
Enter the data into a List.

Calculate the 1-Variable Statistics (STAT CALC 1)
Specify which list your data is in, then press Enter

Median is resistant.
- Mean uses the actual values and will fluctuate on different values.
- In a symmetrical distribution, the mean and median are close together.
- In a skewed distribution, the mean is farther out in the tail than the median.

**Range:** difference between largest and smallest observation

**Quartiles:** Help us analyze the middle half of the data
- The first quartile is larger than 25\% of the observations.
- The third quartile is larger than 75\% of the observations
- The median is the second quartile
- The distance between the first and third quartiles \( \text{IQR} = Q_3 - Q_1 \)

How do we determine if an observation is truly an outlier?
- An outlier will fall more than \( 1.5 \times \text{IQR} \) above the third quartile or below the first quartile.
- An outlier is attributed to:
  - The measurement is recorded incorrectly.
  - The measurement is from a different population.
  - The measurement is correct, but rare.

A Five Number Summary consists of:
- Minimum
- Maximum
- First quartile
- Median
- Third quartile
- Offers a complete description of center and spread
- A boxplot is a good way of representing the 5 number summary

**Boxplots:**
- Regular boxplots show the median, first and third quartiles, max and min values
- Modified boxplots show outliers as points. The whiskers do not extend to the outliers.
Linear Transformations
How do they affect center and spread?

- A linear transformation changes the original variable into the new variable by using the equation: \( x_{\text{new}} = a + b \cdot x_{\text{original}} \)

Adding the constant \( a \) shifts all values of \( x \) upward or downward by same amount

Multiplying by the positive constant \( b \) changes the size of the unit of measurement.

Adding the same number, \( a \), to each observation will add \( a \) to measures of center and to quartiles but does Not change measures of spread.

Multiplying each observation by a positive number \( b \) will multiply both measures of center and measures of spread by \( b \).

Pg. 53 Example 1.15

<table>
<thead>
<tr>
<th>Lakers Salaries</th>
<th>Salaries after 100K increase</th>
<th>Salaries after 10% increase</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \bar{x} = 4.14 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( s = 4.76 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Min = .3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q1 = 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M = 2.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q3 = 4.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Max = 17.1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Measuring Spread**  
Variance & Standard Deviation

**Variance** $s^2$

- The average of the squares of the deviations of the observations from their mean.
- Formula: 
  \[ s^2 = \frac{(x_1 - \bar{X})^2 + (x_2 - \bar{X})^2 + \ldots + (x_n - \bar{X})^2}{n - 1} \]
- Some of the deviations will be negative, so we square to rid them of negative values.
- The variance will be large if the observations are widely spread about their mean.
- $n - 1$ is called the degrees of freedom.

**Standard Deviation** $s$

- Measures spread by looking how far the observations are from the mean.
- Formula: 
  \[ s = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{X})^2} \]
- $s=0$ only when there is no spread. This only happens when the observations have the same value.
- $s > 0$ all other times. As the observations become more spread out about their mean, $s$ gets larger.
- $s$ is not resistant.
- $s$ has the same units of the original data.

**Example**

Page 37 Bonds Hr output

<table>
<thead>
<tr>
<th>16</th>
<th>25</th>
<th>24</th>
<th>19</th>
<th>33</th>
<th>25</th>
<th>34</th>
<th>46</th>
</tr>
</thead>
<tbody>
<tr>
<td>37</td>
<td>33</td>
<td>42</td>
<td>40</td>
<td>37</td>
<td>34</td>
<td>49</td>
<td>73</td>
</tr>
</tbody>
</table>

Find the standard deviation.

Omit the outlier of 73 homeruns and explain how the S.D. changed.

What does this tell us about standard deviation?
Density Curves and The Normal Distribution

The curve is a mathematical model for the distribution. A Density Curve is a curve that
- Is always on or above the horizontal axis, and
- Has area exactly 1 underneath it.
- The area under the curve, between any range of values, is the proportion of all observations that fall in that range.

The median of a density curve is the equal-areas point, the point that divides the area under the curve in half.
The mean of a density curve is the balance point, at which the curve would balance if made of solid material.

The median and the mean are the same for symmetrical distributions. The mean of a skewed curve is pulled away from the median in the direction of the tail.

A Normal distribution with mean, \( \mu \) and Standard deviation, \( \sigma \) will be notated: \( N(\mu, \sigma) \)

Example: The distribution of heights of adult males, in inches is approximately normal and can be modeled with: \( N(69, 2.5) \)
- What percent of men are shorter than 71.5 inches?
- Men who are 74 inches tall are in what percentile?
- What proportion of men are taller than 64 inches?
normalcdf(a, b, μ, σ)

Finds the percent of observations between a and b in a normal distribution whose mean is μ and whose standard deviation is σ.

This function can be used with 2 parameters instead of 4. In that case, the calculator assumes the standard normal distribution where μ = 0, and σ = 1.

[If you do this, make sure you enter z-scores, not actual observations]

normalcdf(z₁, z₂)

Finds the percent of observations between two z-scores.

Guess at each answer before using a calculator.

<table>
<thead>
<tr>
<th>Function</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>normalcdf(-1, 1)</td>
<td></td>
</tr>
<tr>
<td>normalcdf(0, 2)</td>
<td></td>
</tr>
<tr>
<td>normalcdf(-99, 0)</td>
<td></td>
</tr>
</tbody>
</table>

invNorm(p, μ, σ)

Finds the observation x, such that p is the proportion of data that fall below x in the normal distribution with mean μ, and standard deviation σ.

This function works with 1 parameter. It will assume the standard normal distribution, μ = 0 and σ = 1.

[If you do this, your calculator is producing a z-score]

invNorm(p)

Finds the z-score that represents the p<sup>th</sup> percentile.
Examining 2-variable Relationships
Scatterplots and Correlation

A response variable measures an outcome of a study. (Dependent variable)

An explanatory variable is the cause of the outcomes. (Independent variable)

A scatterplot will display the relation between two quantitative variables.

Describe the overall pattern of a scatterplot by the Form, Strength, Direction.
Look for unusual features such as outliers.

- **Form**: Straight, Curved, something exotic, no pattern, clusters, etc.
- **Strength**: How much scatter?
- **Direction**: Positive OR Negative?

Drawing Scatterplots:
- Scale both axes. The intervals must be uniform.
- If the scale does not begin at zero, use the “gap” symbol
- Label both axes.
- When given a grid, use a scale that uses the whole grid so that details can be seen.

**Correlation** measures the strength of the linear association between two quantitative variables.

- No categorical data. (Make sure you know the variables’ units and what they measure.
- Outliers drastically affect correlation. (It’s a good idea to report the correlations with and without the outlier) Correlation is NOT Resistant
- If you think the form is curved, a correlation would be misleading.
- Rescaling either axis will not affect correlation.
- Correlation ranges from -1 to +1.
- The sign indicates the direction of the association, positive or negative.
- Pg. 145 shows some variations of correlation.

Correlation is not a complete description of 2-variable data. Give the means and st. deviations of both x and y.

Correlation, \( r \) is an average of the products of the standardized x values and the standardized y values.

\[
 r = \frac{1}{n-1} \sum \left( \frac{x_i - \bar{X}}{s_x} \right) \left( \frac{y_i - \bar{Y}}{s_y} \right)
\]
Examining 2-variable Relationships
Least Squares Regression

A **regression line** is a straight line that describes how a response variable \( y \) changes as an explanatory variable \( x \) changes.
- It is often used to predict the value of \( y \) for a given \( x \).
- The **least-squares regression line** serves as a mathematical model for the data.
- It is the line that best fits the data. abbreviated: LSRL
- Its equation is in the form \( y = mx + b \), however statisticians use the form:

\[
\hat{y} = a + bx
\]

“\( \hat{y} \)” or “predicted value of \( y \)”

The **least-squares regression line** of \( y \) on \( x \) is the line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible. When a LSRL has the equation \( \hat{y} = a + bx \), define and be able to interpret the variables \( a \) and \( b \).
- The slope of a regression line (\( b \)) is usually important for the interpretation of the data. The slope is the rate of change, the amount of change in \( \hat{y} \) when \( x \) increases by 1.
- The intercept of the regression line (\( a \)) is the value of \( \hat{y} \) when \( x=0 \). It is statistically meaningful only when \( x \) can actually take values close to zero.

To draw a LSRL on a scatterplot, plot two points (\( x, \hat{y} \)) that have some separation, then connect them with a straightedge. It is helpful to know that any LSRL will pass through the point \(( \bar{x}, \bar{y} )\)

The slope and intercept can be calculated by hand:

\[
\text{Slope: } b = r \frac{s_y}{s_x} \quad \text{Intercept: } a = \bar{y} - b\bar{x}
\]

Calculate slope and intercept using these formulas and the summary statistics for the data on Yankees attendance vs. years since 1995.

<table>
<thead>
<tr>
<th>2-Var Stats</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \bar{x} = 6 )</td>
</tr>
<tr>
<td>( s_x = 3.894440482 )</td>
</tr>
<tr>
<td>( \bar{y} = 3272469.462 )</td>
</tr>
<tr>
<td>( s_y = 760237.4194 )</td>
</tr>
</tbody>
</table>

\( r = .9632226091 \)
Examining 2-variable Relationships

A **residual** is the difference between an observed value and the value predicted by the LSRL.

\[
\text{Residual} = \text{Observed } y - \text{Predicted } y
\]

\[
\text{Residual} = y - \hat{y}
\]

Follow these steps to calculate residuals with a calculator:
1. STAT → Edit → Enter the data into L1 and L2.
2. STAT → CALC → 8. LinReg(a + bx) → L1, L2, Y1
3. STAT → Edit → Enter one of the following formulas into L3: \(L2 - Y1(L1)\)
   Or \(\text{RESID}\)

The sum of the residuals always equals zero.

A **residual plot** is a scatterplot of the regression residuals against the explanatory variable. Residual plots help us assess the fit of a regression line.

If the regression line captures the overall relationship between \(x\) and \(y\), the residuals should have no systematic pattern.
- A curved pattern in the residuals shows that the relationship is not linear.
- Increased spread for larger \(x\) values indicates that prediction of \(y\) will be less accurate for larger \(x\).

\[
\text{SST: Total sample variation of the observations about the mean.}
\]
\[
\text{SSE: Remaining “unexplained” sample variability after fitting the regression line.}
\]

\[
\text{SST} = \sum (y - \bar{y})^2 \quad \text{SSE} = \sum (y - \hat{y})^2 \quad r^2 = \frac{\text{SST} - \text{SSE}}{\text{SST}}
\]

The **coefficient of determination**, \(r^2\), is the fraction of the variation in the values of \(y\) that is explained by least-squares regression of \(y\) on \(x\).

<table>
<thead>
<tr>
<th>(x)</th>
<th>(y)</th>
<th>((y - \bar{y})^2)</th>
<th>((y - \hat{y})^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>24</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

An **outlier** is an observation that lies outside the overall pattern of the other observations.

An observation is **influential** for a statistical calculation if removing it would markedly change the result of the calculation. Points that are outliers in the \(x\) direction of a scatterplot are often influential for the least-squares regression line.