1. Two-way tables.
2. Histograms.
3. mean, median, IQR, z score.
4. skew.
5. Boxplots.
7. Regression.
Simple data summaries

• For categorical data, two-way tables can be useful.

<table>
<thead>
<tr>
<th>What flavor of ice cream would you pick?</th>
<th>Chocolate</th>
<th>Vanilla</th>
<th>Neither</th>
</tr>
</thead>
<tbody>
<tr>
<td>Children</td>
<td>40</td>
<td>22</td>
<td>15</td>
</tr>
<tr>
<td>Teens</td>
<td>12</td>
<td>16</td>
<td>45</td>
</tr>
<tr>
<td>Adults</td>
<td>55</td>
<td>54</td>
<td>10</td>
</tr>
<tr>
<td>Total</td>
<td>107</td>
<td>92</td>
<td>70</td>
</tr>
</tbody>
</table>

• For quantitative data, histograms are useful.
• For a relative frequency histogram, the percentage of people in the bin is shown rather than the whole number.

The relative frequency histogram given next was constructed from data obtained from a random sample of 25 families. Each was asked the number of quarts of milk that had been purchased the previous week.

• Here, \( n = 25 \). \( 0.2 = 20\% \) of people in the sample had 3 quarts. The number of people with 3 quarts was \( 0.2 \times 25 = 5 \).

• The sizes of the bins can be adjusted and the look of the histogram can be influenced by the bin sizes.
• With histograms, look for symmetry, skew, bimodality, and outliers.

• The range = maximum observed value – minimum.
• For roughly symmetric data, the mean and sd are good summaries of the center and spread.
• When the data are skewed or there are serious outliers, the median and the IQR can be preferable.
3. mean, median, IQR, z score.

• The median is the middle in the sorted list of values. It is a value M where 50% of the observations are $\leq M$. Different software use different conventions, but we will use the convention that, if there is a range of possible medians, you take the middle of that range.

• For example, suppose data are $1, 3, 7, 7, 8, 9, 12, 14$.

• $M = 7.5$.

• Suppose 25% of the observations lie below a certain value $x$. Then $x$ is called the lower quartile (or 25th percentile).

• Similarly, if 25% of the observations are greater than $x$, then $x$ is called the upper quartile (or 75th percentile).

• The lower quartile can be calculated by finding the median $M$, and then determining the median of the values below $M$. Similarly the upper quartile is the median of the values greater than $M$. 
IQR and Five-Number Summary

• The difference between the quartiles is called the inter-quartile range (IQR), another measure of variability along with standard deviation.

• The five-number summary for the distribution of a quantitative variable consists of the minimum, lower quartile, median, upper quartile, and maximum.

• Technically the IQR is not the interval (25th percentile, 75th percentile), but the difference 75th percentile – 25th.

• Suppose data are 1, 3, 7, 7, 8, 9, 12, 14.

• M = 7.5, 25th percentile = 5, 75th percentile = 10.5. IQR = 5.5.
Many datasets are roughly symmetric and the histogram somewhat resembles the normal curve.

For such data, about 2/3 of observations are within 1 SD of the mean, and about 95% are within 2 SDs of the mean.

It can be useful to convert values to z scores. This simply means taking a value $x$ and standardizing it by subtracting the sample mean and then dividing by $s$.

IQ. mean = 100, $s = 15$. If $x = 125$, then $z=\frac{(125-100)}{15} = 1.33$.

About 95% of z-scores are between -2 and 2, for normal data.
4. Skew.

The mean and sd are very influenced by outliers and skew. However, the median and IQR are much more resistant to outliers and skew. In many cases the median and IQR will not change at all by the addition of one or two huge outliers.

For right skewed data, mean > median. For left skewed data, mean < median.
Geyser Eruptions

Example 6.1
Old Faithful Inter-Eruption Times

- How do the five-number summary and IQR differ for inter-eruption times between 1978 and 2003?
Old Faithful Inter-Eruption Times

<table>
<thead>
<tr>
<th></th>
<th>Minimum</th>
<th>Lower quartile</th>
<th>Median</th>
<th>Upper quartile</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>1978 times</td>
<td>42</td>
<td>58</td>
<td>75</td>
<td>81</td>
<td>95</td>
</tr>
<tr>
<td>2003 times</td>
<td>56</td>
<td>87</td>
<td>91</td>
<td>98</td>
<td>110</td>
</tr>
</tbody>
</table>

- 1978 IQR = 81 – 58 = 23
- 2003 IQR = 98 – 87 = 11
5. Boxplots.

![Boxplot diagram with annotations for Min, Q_lower, Med, Q_upper, and Max](image-url)
outliers on boxplots.

- A data value that is more than $1.5 \times \text{IQR}$ above the upper quartile or below the lower quartile is considered an outlier.
- When these occur, the whiskers on a boxplot extend out to the farthest value not considered an outlier and outliers are represented by a dot or an asterisk.
Cancer Pamphlet Reading Levels

- Short et al. (1995) compared reading levels of cancer patients and readability levels of cancer pamphlets. What is the:
  - Median reading level?
  - Mean reading level?
- Are the data skewed one way or the other?

<table>
<thead>
<tr>
<th>Pamphlets’ readability levels</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count (number of pamphlets)</td>
<td>3</td>
<td>3</td>
<td>8</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>30</td>
</tr>
</tbody>
</table>
6
12
15
18

Mean = 9.800
median = 9

(n=30)

Skewed a bit to the right

Mean to the right of median
6. Scatterplots and Correlation

Suppose we collected data on the relationship between the time it takes a student to take a test and the resulting score.

<table>
<thead>
<tr>
<th>Time</th>
<th>30</th>
<th>41</th>
<th>41</th>
<th>43</th>
<th>47</th>
<th>48</th>
<th>51</th>
<th>54</th>
<th>54</th>
<th>56</th>
<th>56</th>
<th>57</th>
<th>58</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score</td>
<td>100</td>
<td>84</td>
<td>94</td>
<td>90</td>
<td>88</td>
<td>99</td>
<td>85</td>
<td>84</td>
<td>94</td>
<td>100</td>
<td>65</td>
<td>64</td>
<td>65</td>
</tr>
<tr>
<td>Time</td>
<td>58</td>
<td>60</td>
<td>61</td>
<td>61</td>
<td>62</td>
<td>63</td>
<td>64</td>
<td>66</td>
<td>66</td>
<td>69</td>
<td>72</td>
<td>78</td>
<td>79</td>
</tr>
<tr>
<td>Score</td>
<td>83</td>
<td>85</td>
<td>86</td>
<td>92</td>
<td>74</td>
<td>73</td>
<td>75</td>
<td>53</td>
<td>91</td>
<td>85</td>
<td>62</td>
<td>68</td>
<td>72</td>
</tr>
</tbody>
</table>
Scatterplot

Put explanatory variable on the horizontal axis.

Put response variable on the vertical axis.
Describing Scatterplots

• When we describe data in a scatterplot, we describe the
  • Direction (positive or negative)
  • Form (linear or not)
  • Strength (strong-moderate-weak, we will let correlation help us decide)
  • Unusual Observations
Correlation

- **Correlation** measures the strength and direction of a **linear** association between two **quantitative** variables.
- Correlation is a number between -1 and 1.
- With positive correlation one variable increases, on average, as the other increases.
- With negative correlation one variable decreases, on average, as the other increases.
- The closer it is to either -1 or 1 the closer the points fit to a line.
- The correlation for the test data is -0.56.
## Correlation Guidelines

<table>
<thead>
<tr>
<th>Correlation Value</th>
<th>Strength of Association</th>
<th>What this means</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.7 to 1.0</td>
<td>Strong</td>
<td>The points will appear to be nearly a straight line</td>
</tr>
<tr>
<td>0.3 to 0.7</td>
<td>Moderate</td>
<td>When looking at the graph the increasing/decreasing pattern will be clear, but there is considerable scatter.</td>
</tr>
<tr>
<td>0.1 to 0.3</td>
<td>Weak</td>
<td>With some effort you will be able to see a slightly increasing/decreasing pattern</td>
</tr>
<tr>
<td>0 to 0.1</td>
<td>None</td>
<td>No discernible increasing/decreasing pattern</td>
</tr>
</tbody>
</table>

**Same Strength Results with Negative Correlations**
Back to the test data

Actually the last three people to finish the test had scores of 93, 93, and 97.

What does this do to the correlation?
Influential Observations

• The correlation changed from -0.56 (a fairly moderate negative correlation) to -0.12 (a weak negative correlation).

• Points that are far to the left or right and not in the overall direction of the scatterplot can greatly change the correlation. (influential observations)
Correlation

• **Correlation** measures the strength and direction of a linear association between two quantitative variables.
  
  • \(-1 \leq r \leq 1\)
  
  • Correlation makes no distinction between explanatory and response variables.
  
  • Correlation has no units. Any linear change in units of \(x\) or \(y\), meaning adding and multiplying every observation by some number, does not influence \(r\).
  
  • Correlation is not resistant to outliers. It is sensitive.
Calculating correlation, \( r \):

\[ \rho = \text{rho} = \text{correlation of the population.} \]

Suppose there are \( N \) people in the population, \( X = \text{temperature}, Y = \text{heart rate} \),

the mean and sd of temp in the pop. are \( \mu_x \) and \( \sigma_x \),

and the pop. mean and sd of heart rate are \( \mu_y \) and \( \sigma_y \).

\[
\rho = \frac{1}{N} \sum_{i=1}^{N} \left( \frac{x_i - \mu_x}{\sigma_x} \right) \left( \frac{y_i - \mu_y}{\sigma_y} \right).
\]

Given a sample of size \( n \), we estimate \( \rho \) using

\[
r = \frac{1}{n-1} \sum_{i=1}^{n} \left( \frac{x_i - \overline{x}}{s_x} \right) \left( \frac{y_i - \overline{y}}{s_y} \right).
\]
## Temperature and Heart Rate

<table>
<thead>
<tr>
<th>Tmp</th>
<th>98.3</th>
<th>98.2</th>
<th>98.7</th>
<th>98.5</th>
<th>97.0</th>
<th>98.8</th>
<th>98.5</th>
<th>98.7</th>
<th>99.3</th>
<th>97.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>HR</td>
<td>72</td>
<td>69</td>
<td>72</td>
<td>71</td>
<td>80</td>
<td>81</td>
<td>68</td>
<td>82</td>
<td>68</td>
<td>65</td>
</tr>
<tr>
<td>Tmp</td>
<td>98.2</td>
<td>99.9</td>
<td>98.6</td>
<td>98.6</td>
<td>97.8</td>
<td>98.4</td>
<td>98.7</td>
<td>97.4</td>
<td>96.7</td>
<td>98.0</td>
</tr>
<tr>
<td>HR</td>
<td>71</td>
<td>79</td>
<td>86</td>
<td>82</td>
<td>58</td>
<td>84</td>
<td>73</td>
<td>57</td>
<td>62</td>
<td>89</td>
</tr>
</tbody>
</table>
Temperature and Heart Rate

$r = 0.378$
Heart Rate and Body Temp

- Another group studied the relationship between heart rate and body temperature with 130 healthy adults
- Predicted Heart Rate = $-166.3 + 2.44(Temp)$
- $r = 0.257$
7. Regression.

- Fitting a line to a scatterplot.
- Unless the points are perfectly linearly alligned, there will not be a single line that goes through every point.
- We want a line that gets as close as possible to all the points.
Regression.

• We want a line that minimizes the vertical distances between the line and the points
  • These distances are called **residuals**.
  • The line we will find actually minimizes the sum of the squares of the residuals.
• This is called a **least-squares regression line**.
Are Dinner Plates Getting Larger?
Growing Plates?

• There are many recent articles and TV reports about the obesity problem.
• One reason some have given is that the size of dinner plates are increasing.
• Are these black circles the same size, or is one larger than the other?
Growing Plates?

• They appear to be the same size for many, but the one on the right is about 20% larger than the left.

• This suggests that people will put more food on larger dinner plates without knowing it.

• There is name for this phenomenon: Delboeuf illusion
Growing Plates?

- Researchers gathered data to investigate the claim that dinner plates are growing
- American dinner plates sold on eBay on March 30, 2010 (Van Ittersum and Wansink, 2011)
- Year manufactured and diameter are given.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>10</td>
<td>10.75</td>
<td>10.125</td>
<td>10</td>
<td>10.625</td>
<td>10.75</td>
<td>10.625</td>
<td>10</td>
<td>10.5</td>
<td>10.125</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>10.375</td>
<td>10.75</td>
<td>10.375</td>
<td>11</td>
<td>10.75</td>
<td>10.125</td>
<td>11.5</td>
<td>11</td>
<td>11.125</td>
<td>11</td>
</tr>
</tbody>
</table>
Growing Plates?

• Both year (explanatory variable) and diameter in inches (response variable) are quantitative.
• Each dot represents one plate in this scatterplot.
Growing Plates?

- The association appears to be roughly linear
- The least squares regression line is added.
The regression equation is $\hat{y} = a + bx$:

- $a$ is the $y$-intercept
- $b$ is the slope
- $x$ is a value of the explanatory variable
- $\hat{y}$ is the predicted value for the response variable
- For a specific value of $x$, the corresponding distance $y - \hat{y}$ (or actual – predicted) is a residual
Regression Line

- The least squares line for the dinner plate data is
  \[ \hat{y} = -14.8 + 0.0128x \]
- Or diameter = \(-14.8 + 0.0128\text{(year)}\)
- This allows us to predict plate diameter for a particular year.
\[ \hat{y} = -14.8 + 0.0128x \]

- What is the predicted diameter for a plate manufactured in 2000?
  - \(-14.8 + 0.0128(2000) = 10.8\) in.
- What is the predicted diameter for a plate manufactured in 2001?
  - \(-14.8 + 0.0128(2001) = 10.8128\) in.
- How does this compare to our prediction for the year 2000?
  - 0.0128 larger
- Slope \(b = 0.0128\) means that using the regression line, diameters are predicted to increase by 0.0128 inches per year on average.
Slope

• Slope is the predicted change in the response variable for one-unit change in the explanatory variable.
• Both the slope and the correlation coefficient for this study were positive.
  • The slope is 0.0128
  • The correlation is 0.604
• The slope and correlation coefficient will always have the same sign.
y-intercept

• The y-intercept is where the regression line crosses the y-axis or the predicted response when the explanatory variable equals 0.

• We had a y-intercept of -14.8 in the dinner plate equation. What does this tell us about our dinner plate example?
  • Dinner plates in year 0 were -14.8 inches.

• How can it be negative?
  • The equation works well within the range of values given for the explanatory variable, but fails outside that range.

• Our equation should only be used to predict the size of dinner plates from about 1950 to 2010.
Extrapolation

- Predicting values for the response variable for values of the explanatory variable that are outside of the range of the original data is called *extrapolation*. 
\( r^2 \)

- While the intercept and slope have meaning in the context of year and diameter, remember that the correlation does not have units. It is just 0.604.
- The square of the correlation \( r^2 \) has meaning in terms of variation.
- \( r^2 = 0.604^2 = 0.365 \) or 36.5%
- 36.5% of the variation in plate size (the response variable) can be explained by its linear association with the year (the explanatory variable).
Slope of regression line.

- Suppose \( \hat{y} = a + bx \) is the regression line.
- The slope \( b \) of the regression line is \( b = r \frac{s_y}{s_x} \).
  
  This is usually the thing of primary interest to interpret, as the predicted increase in \( y \) for every unit increase in \( x \).

- Beware of assuming causation though, esp. with observational studies. Be wary of extrapolation too.

- The intercept \( a = \bar{y} - b \bar{x} \).

- The SD of the residuals is \( \sqrt{1 - r^2} s_y \).
  
  This is a good estimate of how much the regression predictions will typically be off by.
8. How well does the line fit?

- $r^2$ is a measure of fit. It indicates the amount of scatter around the best fitting line.
- Residual plots can indicate curvature, outliers, or heteroskedasticity.
- $\sqrt{1 - r^2} s_y$ is a very useful summary. It is a measure of how far off predictions would have been on average.
• Heteroskedasticity: when the variability in $y$ is not constant as $x$ varies.
How well does the line fit?

- $r^2$ is a measure of fit. It indicates the amount of scatter around the best fitting line.
- Residual plots can indicate curvature, outliers, or heteroskedasticity.
- $\sqrt{1 - r^2} s_y$ is useful as a measure of how far off predictions would have been on average.

Note that regression residuals have mean zero, whether the line fits well or poorly.
9. common problems with regression.

- a. Correlation is not causation. Especially with observational data.

Number of people who drowned by falling into a pool correlates with Films Nicolas Cage appeared in

Data sources: Centers for Disease Control & Prevention and Internet Movie Database
Common problems with regression.
Holmes and Willett (2004) reviewed all prospective studies on fat consumption and breast cancer with at least 200 cases of breast cancer. "Not one study reported a significant positive association with total fat intake.... Overall, no association was observed between intake of total, saturated, monounsaturated, or polyunsaturated fat and risk for breast cancer."

They also state "The dietary fat hypothesis is largely based on the observation that national per capita fat consumption is highly correlated with breast cancer mortality rates. However, per capita fat consumption is highly correlated with economic development. Also, low parity and late age at first birth, greater body fat, and lower levels of physical activity are more prevalent in Western countries, and would be expected to confound the association with dietary fat."
Common problems with regression.

• b. Extrapolation.

If the birthrate remains at 1.19 children per woman, South Korea could face natural extinction by 2750.

Common problems with regression.

- b. Extrapolation.
- Often researchers extrapolate from high doses to low.
Common problems with regression.

- b. Extrapolation.
  The relationship can be nonlinear though.
  Researchers also often extrapolate from animals to humans.
  Zaichkina et al. (2004) on hamsters
Common problems with regression.

- c. Curvature.
The best fitting line might fit poorly. Port et al. (2005).
Common problems with regression.

- c. Curvature.

The best fitting line might fit poorly. Wong et al. (2011).
Common problems with regression.

- d. Statistical significance.

Could the observed correlation just be due to chance alone?