Today’s plan:

- Section 4.1.4: Dispersion: Five-Number summary and Standard Deviation.
Once we know the central location of a data set, we want to know how close things are to the center.
Once we know the central location of a data set, we want to know how close things are to the center.

We’ll see two ways to measure dispersion of a data set.
• **five-number summary** (goes with the median)
➤ five-number summary (goes with the median)
➤ standard deviation (goes with the mean)
Five-Number Summary
Five-number Summary:
1. Min
2. Lower Quartile
3. Median
4. Upper Quartile
5. Max
Definition

- The **Min** is the smallest value in the whole data set.
- The **Max** is the largest value in the whole data set.
- The **Lower Quartile** is the median of the lower half.
- The **Upper Quartile** is the median of the upper half.
Definition

- **The Min** is the smallest value in the whole data set.
- **The Max** is the largest value in the whole data set.
Definition

- The **Min** is the smallest value in the whole data set.
- The **Max** is the largest value in the whole data set.
- The **Lower Quartile** is the median of the lower half.
Definition

- The **Min** is the smallest value in the whole data set.
- The **Max** is the largest value in the whole data set.
- The **Lower Quartile** is the median of the lower half.
- The **Upper Quartile** is the median of the upper half.
Example

The appraisals of the 10 houses are:

\[\$75K, \$96K, \$107K, \$110K, \$110K, \$118K, \$130K, \$135K, \$150K, \$520K\]
Example

The appraisals of the 10 houses are:

\[
[\$75K, \$96K, \$107K, \$110K, \$110K, \\
\$118K, \$130K, \$135K, \$150K, \$520K]
\]

Find the five-number summary.
Solution

We already found:

- the median, Med = $114K
Solution

We already found:

- the median, $\text{Med} = 114K$
- the lower half, $[75K, 96K, 107K, 110K, 110K]$
Solution

We already found:

- **the median, Med = $114K**
- **the lower half,**
  
  $[75K, 96K, 107K, 110K, 110K]$

- **the upper half**
  
  $[118K, 130K, 135K, 150K, 520K]$
We already found:

- the median, $\text{Med} = 114K$
- the lower half, $[75K, 96K, 107K, 110K, 110K]$
- the upper half $[118K, 130K, 135K, 150K, 520K]$

Since each half has size 5, their respective medians will be in the 3rd location.
Solution

Thus

- the lower quartile is $Q_1 = 107K$

- the upper quartile is $Q_3 = 135K$

- the lowest value is $Min = 75K$

- the highest value is $Max = 520K$

So the five-number summary is:

\[
Min = 75K, \quad Q_1 = 107K, \quad Med = 114K, \quad Q_3 = 135K, \quad Max = 520K
\]
Thus

- the lower quartile is $Q_1 = \$107K$
- the upper quartile is $Q_3 = \$135K$
Thus

- the lower quartile is $Q_1 = $107K
- the upper quartile is $Q_3 = $135K
- the lowest value is $Min = $75K
Thus

- the lower quartile is $Q_1 = $107K
- the upper quartile is $Q_3 = $135K
- the lowest value is $Min = $75K
- the highest value is $Max = $520K
Thus

- the lower quartile is $Q_1 = $107K
- the upper quartile is $Q_3 = $135K
- the lowest value is Min = $75K
- the highest value is Max = $520K

So the five-number summary is:

$[Min = $75K, \ Q_1 = $107K, \ Med = $114K, \ Q_3 = $135K, \ Max = $520K]$. 
The five-number summary can be visualized with a *boxplot* diagram, or *box-and-whiskers* diagram.
The box goes from the lower quartile to the upper quartile, with a mark at the median.
The box goes from the lower quartile to the upper quartile, with a mark at the median. Two whiskers extend from the box to the Min and Max.
Remarks:

- the left whisker spans the bottom 25%
Remarks:

- the left whisker spans the bottom 25% 
- the box spans the middle 50%
Remarks:

- the left whisker spans the bottom 25%
- the box spans the middle 50%
- the right whisker spans the top 25%
Remarks:

- the left whisker spans the bottom 25%
- the box spans the middle 50%
- the right whisker spans the top 25%
- each half of the box spans 25%
Example

The ages of the police officers in the Clearview Police Department are

<table>
<thead>
<tr>
<th>Age</th>
<th>22</th>
<th>25</th>
<th>26</th>
<th>27</th>
<th>28</th>
<th>29</th>
<th>30</th>
<th>32</th>
<th>35</th>
<th>39</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freq.</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>2</td>
</tr>
</tbody>
</table>
Example

The ages of the police officers in the Clearview Police Department are

<table>
<thead>
<tr>
<th>Age</th>
<th>22</th>
<th>25</th>
<th>26</th>
<th>27</th>
<th>28</th>
<th>29</th>
<th>30</th>
<th>32</th>
<th>35</th>
<th>39</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freq.</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>2</td>
</tr>
</tbody>
</table>

Find the five-number summary and draw the boxplot.
<table>
<thead>
<tr>
<th>Age</th>
<th>22</th>
<th>25</th>
<th>26</th>
<th>27</th>
<th>28</th>
<th>29</th>
<th>30</th>
<th>32</th>
<th>35</th>
<th>39</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freq.</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Cum. Freq</td>
<td>3</td>
<td>7</td>
<td>10</td>
<td>15</td>
<td>19</td>
<td>25</td>
<td>30</td>
<td>34</td>
<td>39</td>
<td>41</td>
</tr>
</tbody>
</table>
The size is \( n = 41 \), so the median is in location
The size is \( n = 41 \), so the median is in location \( \frac{41 + 1}{2} = 21 \).
The size is $n = 41$, so the median is in location \( \frac{41 + 1}{2} = 21 \).

The lower half has size 20, so the lower quartile is the average of the values at locations 10 and 11:

\[
Q_1 = \frac{26 + 27}{2} = 26.5
\]
The upper half also has size 20, so the upper quartile is the average of the values at locations 10 and 11 of the upper half.
The upper half also has size 20, so the upper quartile is the average of the values at locations 10 and 11 of the upper half.

Since the median is at location 21, the third quartile is the average of the values at locations 31 and 32 of the whole data set:

\[ Q_3 = \frac{32 + 32}{2} = 32 \]
Five-number summary:

\[ \begin{align*}
\text{Min} &= 22, & \text{Q1} &= 26.5, & \text{Med} &= 29, & \text{Q3} &= 32, & \text{Max} &= 39
\end{align*} \]
Remark: Outliers can be drawn separated from the rest of the data set.
Example

The appraisals of the 10 houses are:

Example

The appraisals of the 10 houses are:


Find the five-number summary with outliers separated.
Boxplots and five-number summaries are useful when comparing two data sets.
Example

Waiting times at two car washes:

Acme Car Wash:

\[ \text{Min} = 1, \quad Q1 = 5, \quad Med = 8, \quad Q3 = 9, \quad Max = 12 \]

Kleen Car Wash:

\[ \text{Min} = 3, \quad Q1 = 4, \quad Med = 5, \quad Q3 = 8, \quad Max = 20 \]

(Times are in minutes.)
Example

Draw the boxplots together, and compare them.
Solution

Here are the boxplots:

- **Acme**
- **Kleen**
The Min and Max tell us:
Solution

The Min and Max tell us:

- everyone at Kleen has to wait at least 3 minutes, and some people have a very long wait.
Solution

*The Min and Max tell us:*

- everyone at Kleen has to wait at least 3 minutes, and some people have a very long wait.
- at Acme, some have a tiny wait and everyone gets started in $\leq 12$ minutes.
Solution

The Min and Max tell us:

- everyone at Kleen has to wait at least 3 minutes, and some people have a very long wait.
- at Acme, some have a tiny wait and everyone gets started in $\leq 12$ minutes.

Acme seems better.
Solution

But, the Median tells us:
• half of the customers of Acme wait $\geq 8$ minutes for service
Solution

But, the Median tells us:

- half of the customers of Acme wait $\geq 8$ minutes for service
- at Kleen half of them start in $\leq 5$ minutes
Solution

But, the Median tells us:

- half of the customers of Acme wait \( \geq 8 \) minutes for service
- at Kleen half of them start in \( \leq 5 \) minutes

Now Kleen seems better.
Which is better? There’s no simple answer
Which is better? There’s no simple answer

If you don’t mind waiting a little, Acme is better, since there are no long waits.
Which is better? There’s no simple answer

If you don’t mind waiting a little, Acme is better, since there are no long waits.

If you’re willing to risk a long wait, in hope of a really short wait, Kleen is better.
Standard Deviation
When using the mean to measure the center, we use the standard deviation to measure dispersion.
When using the mean to measure the center, we use the standard deviation to measure dispersion.

Think of standard deviation as measuring how far from the average the data points tend to be.
(Wrong way:)

1. take the deviation of each data point from the average
2. average those deviations

The deviation of a point $x_i$ from the average $\bar{x}$ is just $x_i - \bar{x}$
(Wrong way:)

1. take the deviation of each data point from the average
(Wrong way:)

1. take the deviation of each data point from the average
2. average those deviations
(Wrong way:)

1. take the deviation of each data point from the average
2. average those deviations

The deviation of a point $x_i$ from the average $\bar{x}$ is just

$$x_i - \bar{x}$$
(Wrong way:)

Example: Weekly Sales of Home Town Pharmacy:

S M T W R F S

$2,548, $1,225, $1,732, $1,871, $975, $2,218, $1,339.

Find the average of 

\( x_i - \bar{x} \).

We have already found the average:

\( \bar{x} = 1701.14 \).
(Wrong way:)

Example

Weekly Sales of Home Town Pharmacy:

<table>
<thead>
<tr>
<th>S</th>
<th>M</th>
<th>T</th>
<th>W</th>
<th>R</th>
<th>F</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>$2,548</td>
<td>$1,225</td>
<td>$1,732</td>
<td>$1,871</td>
<td>$975</td>
<td>$2,218</td>
<td>$1,339</td>
</tr>
</tbody>
</table>

Find the average of $x_i - \bar{x}$. 
(Wrong way:)

Example

Weekly Sales of Home Town Pharmacy:

<table>
<thead>
<tr>
<th></th>
<th>S</th>
<th>M</th>
<th>T</th>
<th>W</th>
<th>R</th>
<th>F</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$2,548</td>
<td>$1,225</td>
<td>$1,732</td>
<td>$1,871</td>
<td>$975</td>
<td>$2,218</td>
<td>$1,339</td>
</tr>
</tbody>
</table>

Find the average of \( x_i - \bar{x} \).

We have already found the average: \( \bar{x} = 1701.14 \).
(Wrong way:)
Here are deviations $x_i - \bar{x}$:

<table>
<thead>
<tr>
<th>Day</th>
<th>$x_i$ (sales)</th>
<th>$x_i - \bar{x}$ (deviation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sunday</td>
<td>2,548.00</td>
<td>846.86</td>
</tr>
<tr>
<td>Monday</td>
<td>1,225.00</td>
<td>-476.14</td>
</tr>
<tr>
<td>Tuesday</td>
<td>1,732.00</td>
<td>30.86</td>
</tr>
<tr>
<td>Wednesday</td>
<td>1,871.00</td>
<td>169.86</td>
</tr>
<tr>
<td>Thursday</td>
<td>975.00</td>
<td>-726.14</td>
</tr>
<tr>
<td>Friday</td>
<td>2,218.00</td>
<td>516.86</td>
</tr>
<tr>
<td>Saturday</td>
<td>1,339.00</td>
<td>-362.14</td>
</tr>
<tr>
<td>Total</td>
<td>11,908.00</td>
<td>0.02</td>
</tr>
<tr>
<td>Average</td>
<td>1,701.14</td>
<td>0.00</td>
</tr>
</tbody>
</table>
(Wrong way:)

- Deviations are like distances, but with a sign
(Wrong way:)

- Deviations are like distances, but with a sign
- Positive deviation $\Rightarrow x_i$ is to the right of $\bar{x}$
(Wrong way:)

- Deviations are like distances, but with a sign
- Positive deviation $\Rightarrow x_i$ is to the right of $\bar{x}$
- Negative deviation $\Rightarrow x_i$ is to the left of $\bar{x}$
(Wrong way:)
The average of those deviations:

\[
\frac{846.86 - 476.14 + 30.86 + 169.86 - 726.14 + 516.86 - 362.14}{7} = 0.00
\]
(Wrong way:)
The average of those deviations:

\[
\frac{846.86 - 476.14 + 30.86 + 169.86 - 726.14 + 516.86 - 362.14}{7} = 0.00
\]

This is going to happen with any data set! Average deviation from the mean is a **useless measure of dispersion**.
(Right way:)

- However, if we square all deviations, they will turn all positive
(Right way:)

- However, if we square all deviations, they will turn all positive
- We can then average those squared deviations
(Right way:)

- However, if we square all deviations, they will turn all positive.
- We can then average those squared deviations.
- That is called the **variance**.
**Definition**

The **variance** $\text{var}(x)$ of a data set $x$ is the average of the squared deviations from the mean $\bar{x}$:

$$\text{var}(x) = \frac{1}{n} \sum (x_i - \bar{x})^2$$
To compensate for the squaring, we take the square root of the variance.
To compensate for the squaring, we take the square root of the variance.

**Definition**

The **standard deviation** is

$$\sigma(x) = \sqrt{\text{var}(x)}$$
Example

Find the variance and standard deviation for the Home Town Pharmacy daily sales data set.
<table>
<thead>
<tr>
<th>Day</th>
<th>$x$ (sales)</th>
<th>$x - \bar{x}$</th>
<th>$(x - \bar{x})^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sunday</td>
<td>2,548.00</td>
<td>846.86</td>
<td>717171.8596</td>
</tr>
<tr>
<td>Monday</td>
<td>1,225.00</td>
<td>-476.14</td>
<td>226709.2996</td>
</tr>
<tr>
<td>Tuesday</td>
<td>1,732.00</td>
<td>30.86</td>
<td>952.3396</td>
</tr>
<tr>
<td>Wednesday</td>
<td>1,871.00</td>
<td>169.86</td>
<td>28852.4196</td>
</tr>
<tr>
<td>Thursday</td>
<td>975.00</td>
<td>-726.14</td>
<td>527279.2996</td>
</tr>
<tr>
<td>Friday</td>
<td>2,218.00</td>
<td>516.86</td>
<td>267144.2596</td>
</tr>
<tr>
<td>Saturday</td>
<td>1,339.00</td>
<td>-362.14</td>
<td>131145.3796</td>
</tr>
<tr>
<td>Total</td>
<td>11,908.00</td>
<td>0.02</td>
<td>1899254.8572</td>
</tr>
<tr>
<td>Average</td>
<td>1,701.14</td>
<td>0.00</td>
<td>271322.1224571</td>
</tr>
</tbody>
</table>
The variance is
$$\text{var}(x) = 271322.1224571$$
- the variance is
  \[ \text{var}(x) = 271322.1224571 \]
- the standard deviation is
  \[ \sigma(x) = \sqrt{271322.1224571} = 520.89 \]
What if we start with a frequency table or a histogram?
Example

Find the standard deviation for the Math 109 quizzes

<table>
<thead>
<tr>
<th>score</th>
<th>freq.</th>
<th>cum fr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>11</td>
<td>5</td>
<td>14</td>
</tr>
<tr>
<td>12</td>
<td>9</td>
<td>23</td>
</tr>
<tr>
<td>13</td>
<td>12</td>
<td>35</td>
</tr>
<tr>
<td>14</td>
<td>11</td>
<td>46</td>
</tr>
<tr>
<td>15</td>
<td>13</td>
<td>59</td>
</tr>
<tr>
<td>16</td>
<td>9</td>
<td>68</td>
</tr>
<tr>
<td>17</td>
<td>8</td>
<td>76</td>
</tr>
<tr>
<td>18</td>
<td>7</td>
<td>83</td>
</tr>
<tr>
<td>19</td>
<td>5</td>
<td>88</td>
</tr>
<tr>
<td>20</td>
<td>3</td>
<td>91</td>
</tr>
<tr>
<td>21</td>
<td>2</td>
<td>93</td>
</tr>
<tr>
<td>22</td>
<td>1</td>
<td>94</td>
</tr>
<tr>
<td>25</td>
<td>1</td>
<td>95</td>
</tr>
</tbody>
</table>
Solution

- We computed the average \( \mu = 14.64 \)
Solution

- We computed the average $\mu = 14.64$
- For convenience turn the frequency table into a vertical table
<table>
<thead>
<tr>
<th>x</th>
<th>f</th>
<th>x · f</th>
<th>(x − μ)</th>
<th>(x − μ)^2</th>
<th>(x − μ)^2 · f</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>1</td>
<td>4</td>
<td>-10.64</td>
<td>113.2096</td>
<td>113.2096</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>5</td>
<td>-9.64</td>
<td>92.9296</td>
<td>92.9296</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>16</td>
<td>-6.64</td>
<td>44.0896</td>
<td>88.1792</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>18</td>
<td>-5.64</td>
<td>31.8096</td>
<td>63.6192</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>30</td>
<td>-4.64</td>
<td>21.5296</td>
<td>64.5888</td>
</tr>
<tr>
<td>11</td>
<td>5</td>
<td>55</td>
<td>-3.64</td>
<td>13.2496</td>
<td>66.2480</td>
</tr>
<tr>
<td>12</td>
<td>9</td>
<td>108</td>
<td>-2.64</td>
<td>6.9696</td>
<td>62.7264</td>
</tr>
<tr>
<td>13</td>
<td>12</td>
<td>156</td>
<td>-1.64</td>
<td>2.6896</td>
<td>32.2752</td>
</tr>
<tr>
<td>14</td>
<td>11</td>
<td>154</td>
<td>-0.64</td>
<td>0.4096</td>
<td>4.5056</td>
</tr>
<tr>
<td>15</td>
<td>13</td>
<td>195</td>
<td>0.36</td>
<td>0.1296</td>
<td>1.6848</td>
</tr>
<tr>
<td>16</td>
<td>9</td>
<td>144</td>
<td>1.36</td>
<td>1.8496</td>
<td>16.6464</td>
</tr>
<tr>
<td>17</td>
<td>8</td>
<td>136</td>
<td>2.36</td>
<td>5.5696</td>
<td>44.5568</td>
</tr>
<tr>
<td>18</td>
<td>7</td>
<td>126</td>
<td>3.36</td>
<td>11.2896</td>
<td>79.0272</td>
</tr>
<tr>
<td>19</td>
<td>5</td>
<td>95</td>
<td>4.36</td>
<td>19.0096</td>
<td>95.0480</td>
</tr>
<tr>
<td>20</td>
<td>3</td>
<td>60</td>
<td>5.36</td>
<td>28.7296</td>
<td>86.1888</td>
</tr>
<tr>
<td>21</td>
<td>2</td>
<td>42</td>
<td>6.36</td>
<td>40.4496</td>
<td>80.8992</td>
</tr>
<tr>
<td>22</td>
<td>1</td>
<td>22</td>
<td>7.36</td>
<td>54.1696</td>
<td>54.1696</td>
</tr>
<tr>
<td>25</td>
<td>1</td>
<td>25</td>
<td>10.36</td>
<td>107.3296</td>
<td>107.3296</td>
</tr>
<tr>
<td>Tot.</td>
<td>95</td>
<td>1391</td>
<td></td>
<td>1067.6432</td>
<td></td>
</tr>
<tr>
<td>Ave.</td>
<td>14.64</td>
<td></td>
<td></td>
<td></td>
<td>11.2383</td>
</tr>
</tbody>
</table>
So the standard deviation is
\[ \sigma = \sqrt{11.2383} = 3.35. \]
To find the **Standard Deviation** $\sigma$

1. Compute the deviations $x_i - \mu$.
2. Square the deviations $(x_i - \mu)^2$.
3. Average the squared deviations to the variance
   \[ \text{var} = \frac{\sum (x_i - \mu)^2}{n}. \]
4. Take the square root of the variance
   \[ \sigma = \sqrt{\text{var}}. \]
Question

What does standard deviation mean in practice?
In the previous example:

- The average is $\mu = 14.64$
- the standard deviation is $\sigma = 3.35$
How many data points are within one standard deviation of the average?
How many data points are within one standard deviation of the average?

$$\mu - \sigma = 11.29 \text{ and } \mu + \sigma = 17.99$$
How many data points are within one standard deviation of the average?

\[ \mu - \sigma = 11.29 \text{ and } \mu + \sigma = 17.99 \]

Between these two values there are a total of

\[ 9 + 12 + 11 + 13 + 9 + 8 = 62 \]

data points (out of 95), i.e., about two thirds.
For “nice” data sets, about \( \frac{2}{3} \) of the data set is located within one standard deviation of the average.
For “nice” data sets, about $\frac{2}{3}$ of the data set is located within one standard deviation of the average.

- if $\sigma$ is small, the data points are crowded close to $\mu$
For “nice” data sets, about $\frac{2}{3}$ of the data set is located within one standard deviation of the average.

- if $\sigma$ is small, the data points are crowded close to $\mu$
- if $\sigma$ is large, the data points are scattered.