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.

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*, $Med = \$114K$
- *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* $[\$118K, \$130K, \$135K, \$150K, \$520K]$

Solution

We already found:

- *the median,* $Med = \$114K$
- *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* $Q1 = \$107K$

Solution

Thus

- *the lower quartile is* $Q1 = \$107K$
- *the upper quartile is* $Q3 = \$135K$

Solution

Thus

- *the lower quartile is* $Q1 = \$107K$
- *the upper quartile is* $Q3 = \$135K$
- *the lowest value is* $Min = \$75K$

Solution

Thus

- *the lower quartile is* $Q1 = \$107K$
- *the upper quartile is* $Q3 = \$135K$
- *the lowest value is* $Min = \$75K$
- *the highest value is* $Max = \$520K$

Solution

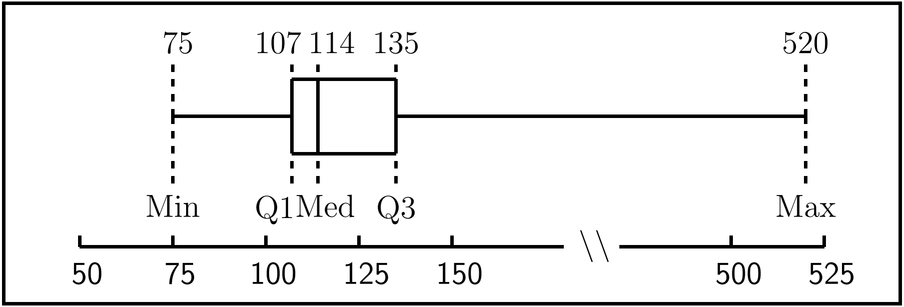
Thus

- *the lower quartile is* $Q1 = \$107K$
- *the upper quartile is* $Q3 = \$135K$
- *the lowest value is* $Min = \$75K$
- *the highest value is* $Max = \$520K$

So the five-number summary is:

$$[Min = \$75K, \quad Q1 = \$107K, \quad Med = \$114K, \quad Q3 = \$135K, \quad Max = \$520K].$$

The five-number summary can be visualized with a **boxplot** diagram, or *box-and-whiskers* diagram.


75
107
114
135
520
Min
Q1
Med
Q3
Max
50
75
100
125
150
500
525

- 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

Age	22	25	26	27	28	29	30	32	35	39
Freq.	3	4	3	5	4	6	5	4	5	2

Example

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

Age	22	25	26	27	28	29	30	32	35	39
Freq.	3	4	3	5	4	6	5	4	5	2

Find the five-number summary and draw the boxplot.

Age	22	25	26	27	28	29	30	32	35	39
Freq.	3	4	3	5	4	6	5	4	5	2
Cum. Freq	3	7	10	15	19	25	30	34	39	41

- The size is $n = 41$, so the median is in location

- The size is $n = 41$, so the median is in location $\dfrac{41+1}{2} = 21$.

- The size is $n = 41$, so the median is in location $\dfrac{41+1}{2} = 21$.
- The lower half has size 20, so the lower quartile is the average of the values at locations 10 and 11:

$$Q1 = \frac{26+27}{2} = 26.5$$

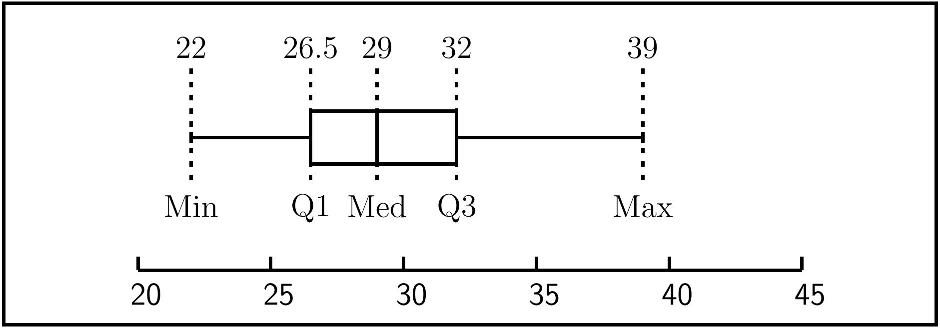
- The upper half also has size 20, so the upper quartile is the average of the values at locations 10 and 11 of the upper half.

- 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:

$$Q3 = \frac{32 + 32}{2} = 32$$

Five-number summary:

$$[\mathit{Min} = 22, \quad \mathit{Q1} = 26.5, \quad \mathit{Med} = 29, \quad \mathit{Q3} = 32, \quad \mathit{Max} = 39]$$



Remark: Outliers can be drawn separated from the rest of the data set.

Example

The appraisals of the 10 houses are:

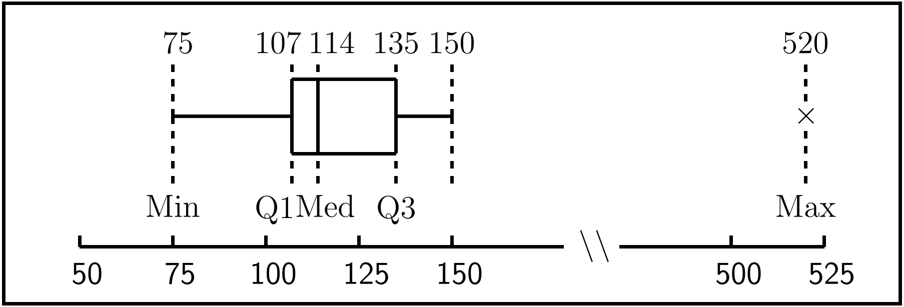
$$[\$75K, \$96K, \$107K, \$110K, \$110K, \$118K, \$130K, \$135K, \$150K, \$520K]$$

Example

The appraisals of the 10 houses are:

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

Find the five-number summary with outliers separated.


75
107
114
135
150
520
Min
Q1
Med
Q3
Max
50
75
100
125
150
500
525

Boxplots and five-number summaries are useful when comparing two data sets.

Example

Waiting times at two car washes:
Acme Car Wash:

$$[Min = 1, \;\; Q1 = 5, \;\; Med = 8, \;\; Q3 = 9, \;\; Max = 12]$$

Kleen Car Wash:

$$[Min = 3, \;\; Q1 = 4, \;\; Med = 5, \;\; Q3 = 8, \;\; Max = 20]$$

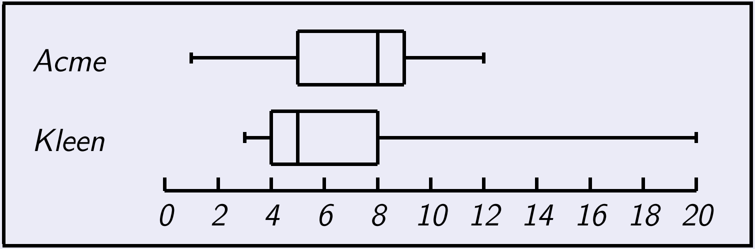
(Times are in minutes.)

Example

Draw the boxplots together, and compare them.

Solution

Here are the boxplots:



Solution

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* ≤ 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 ≤ 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* ≤ 5 *minutes*

Solution

But, the Median tells us:

- *half of the customers of Acme wait* ≥ 8 *minutes for service*
- *at Kleen half of them start in* ≤ 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:)

(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:)

(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 $x_i - \bar{x}$.

(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:)

Here are deviations $x_i - \bar{x}$:

Day	x_i (sales)	$x_i - \bar{x}$ (deviation)
Sunday	2,548.00	846.86
Monday	1,225.00	-476.14
Tuesday	1,732.00	30.86
Wednesday	1,871.00	169.86
Thursday	975.00	-726.14
Friday	2,218.00	516.86
Saturday	1,339.00	-362.14
Total	11,908.00	0.02
Average	1,701.14	0.00

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

$$\mathrm{var}(x) = \frac{1}{n} \sum (x_i - \bar{x})^2$$

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

$$\sigma(x) = \sqrt{\mathrm{var}(x)}$$

Example

Find the variance and standard deviation for the Home Town Pharmacy daily sales data set.

Day	x (sales)	$x - \bar{x}$	$(x - \bar{x})^2$
Sunday	2,548.00	846.86	717171.8596
Monday	1,225.00	-476.14	226709.2996
Tuesday	1,732.00	30.86	952.3396
Wednesday	1,871.00	169.86	28852.4196
Thursday	975.00	-726.14	527279.2996
Friday	2,218.00	516.86	267144.2596
Saturday	1,339.00	-362.14	131145.3796
Total	11,908.00	0.02	1899254.8572
Average	1,701.14	0.00	271322.1224571

- the **variance** is
  $\mathrm{var}(x) = 271322.1224571$

- the **variance** is
  $\mathrm{var}(x) = 271322.1224571$
- 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

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

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*

x	f	$x \cdot f$	$(x-\mu)$	$(x-\mu)^2$	$(x-\mu)^2 \cdot f$
4	1	4	-10.64	113.2096	113.2096
5	1	5	-9.64	92.9296	92.9296
8	2	16	-6.64	44.0896	88.1792
9	2	18	-5.64	31.8096	63.6192
10	3	30	-4.64	21.5296	64.5888
11	5	55	-3.64	13.2496	66.2480
12	9	108	-2.64	6.9696	62.7264
13	12	156	-1.64	2.6896	32.2752
14	11	154	-0.64	0.4096	4.5056
15	13	195	0.36	0.1296	1.6848
16	9	144	1.36	1.8496	16.6464
17	8	136	2.36	5.5696	44.5568
18	7	126	3.36	11.2896	79.0272
19	5	95	4.36	19.0096	95.0480
20	3	60	5.36	28.7296	86.1888
21	2	42	6.36	40.4496	80.8992
22	1	22	7.36	54.1696	54.1696
25	1	25	10.36	107.3296	107.3296
Tot.	95	1391			1067.6432
Ave.		14.64			11.2383

So the standard deviation is $\sigma = \sqrt{11.2383} = 3.35$.

To find the **Standard Deviation** σ

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 σ is small, the data points are crowded close to μ

For “nice” data sets, about $\frac{2}{3}$ of the data set is located within one standard deviation of the average.

- if σ is small, the data points are crowded close to μ
- if σ is large, the data points are scattered.