Section 4.1.3: Median and Mean.

We've seen some graphical presentations of data sets. Now let's give numerical summaries of them.

First: **central location** of the data set. What are the data points centered around?

First: **central location** of the data set. What are the data points centered around?

- **median**
- **mean** or **average**

Median

Median $=$ midpoint.

Definition

The **median** of a data set is the value of the variable with half the points below it and half above it.

Definition

The **median** of a data set is the value of the variable with half the points below it and half above it.

- The points below the median are the lower half

Definition

The **median** of a data set is the value of the variable with half the points below it and half above it.

- The points below the median are the lower half
- The points above the median are the upper half

Example

Weekly pharmacy sales:

S	M	T	W	R	F	S
\$2,548,	\$1,225,	\$1,732,	\$1,871,	\$975,	\$2,218,	\$1,339.

Example

Weekly pharmacy sales:

S	M	T	W	R	F	S
\$2,548,	\$1,225,	\$1,732,	\$1,871,	\$975,	\$2,218,	\$1,339.

Find the **median** of the daily sales.

Solution

First sort the data:

$$[\$975; \$1,225; \$1,339; \color{red}{\$1,732}; \$1,871; \$2,218; \$2,548].$$

Solution

First sort the data:

$$[\$975; \$1,225; \$1,339; \$1,732; \$1,871; \$2,218; \$2,548].$$

The point in the middle, $\$1,732$, *is the median.*

Solution

First sort the data:

$$[\$975; \$1,225; \$1,339; \$1,732; \$1,871; \$2,218; \$2,548].$$

The point in the middle, $\$1,732$*, is the* ***median****.*

- *The* ***lower half*** *is* $[\$975, \$1,225, \$1,339]$

Solution

First sort the data:

$$[\$975; \$1,225; \$1,339; \$1,732; \$1,871; \$2,218; \$2,548].$$

The point in the middle, $\$1,732$*, is the median.*

- *The lower half is* $[\$975, \$1,225, \$1,339]$
- *The upper half is* $[\$1,871, \$2,218, \$2,548]$

Example

The prices of 10 houses in a city block, when sorted, are:

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

Find the median.

Solution

- *This time there's no point in the middle, since the size of the data set is even.*

Solution

- *This time there's no point in the middle, since the size of the data set is even.*
- *So, we take the* **two** *points in the middle and average them.*

Solution

That is, the median for this data set is:

$$Median = \frac{\$110K + \$118K}{2} = \$114K$$

Solution

That is, the median for this data set is:

$$Median = \frac{\$110K + \$118K}{2} = \$114K$$

- *The lower half is* $[\$75K, \$96K, \$107K, \$110K, \$110K]$

Solution

That is, the median for this data set is:

$$Median = \frac{\$110K + \$118K}{2} = \$114K$$

- *The lower half is* $[\$75K, \$96K, \$107K, \$110K, \$110K]$
- *The upper half is* $[\$118K, \$130K, \$135K, \$150K, \$520K]$.

Median

To find the median of a data set of size n:

1. Sort the data set.

Median

To find the median of a data set of size n:

1. Sort the data set.
2. If n is *odd*, take the point at location $\dfrac{n+1}{2}$.

Median

To find the median of a data set of size n:

1. Sort the data set.
2. If n is *odd*, take the point at location $\dfrac{n+1}{2}$.

   If n is *even*, take the average of the points at locations

$$\frac{n}{2} \text{ and } \frac{n}{2}+1$$

It's fine if values get repeated.

It's fine if values get repeated.

Example

Find the median, lower half and upper half of the data set:

$$[3, 5, 5, 6, 6, 6, 6, 8, 11]$$

Solution

- *The data set is already sorted, and has size $n = 9$ (odd)*

Solution

- *The data set is already sorted, and has size $n = 9$ (odd)*
- *So the* *median* *is at location* $\frac{9+1}{2} = 5$*, which is the second 6....*

Solution

The *lower half* *is* $[3, 5, 5, 6]$ *and the* *upper half* *is* $[6, 6, 8, 11]$.

When we have a frequency table, to find a median we can use the cumulative frequency row.

When we have a frequency table, to find a median we can use the cumulative frequency row.

Example

25 point quiz results:

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

Find the median.

Solution

- *From the end of the cumulative frequency row we see* $n = 95$.

Solution

- *From the end of the cumulative frequency row we see* $n = 95$.
- *Since* n *is odd, the median is at location* $\dfrac{95+1}{2} = 48$

Solution

- *Location 48 isn't immediately clear, but....*

Solution

- *Location 48 isn't immediately clear, but....*
- *the last point with value 14 is at location 46, and the last point with value 15 is at location 59.*

Solution

- *Location 48 isn't immediately clear, but....*
- *the last point with value 14 is at location 46, and the last point with value 15 is at location 59.*
- *So the point at location 48 must be a 15.*

Solution

So we have:

$$Median = 15$$

Last step in Computing Median was:

To find the point at **location** k:

1. Find the first cumulative frequency that's greater than or equal to k.
2. The value for that cumulative frequency is the value at location k.

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

1. Find the ages at locations 10,11,21,31, and 32.

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

1. Find the ages at locations 10,11,21,31, and 32.
2. Find the median age.

Solution

First we add a cumulative frequency row:

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

Solution

First we add a cumulative frequency row:

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

- *To find the point at location 10, we look for the first cumulative frequency that is* ≥ 10.

Solution

- *The first cumulative frequency* ≥ 10 *is 10 itself, at age 26, so the point at location 10 is 26.*

Solution

- *The first cumulative frequency* ≥ 10 *is 10 itself, at age 26, so the point at location 10 is 26.*
- *The first cumulative frequency* ≥ 11 *is....*

Solution

- *The first cumulative frequency* ≥ 10 *is 10 itself, at age 26, so the point at location 10 is 26.*
- *The first cumulative frequency* ≥ 11 *is....*
- *...15, at age 27, so the point at location 11 is 27.*

Solution

- *the age at location 21 is*

Solution

- *the age at location 21 is 29*

Solution

- *the age at location 21 is 29*
- *the age at location 31 is*

Solution

- *the age at location 21 is 29*
- *the age at location 31 is 32*

Solution

- *the age at location 21 is 29*
- *the age at location 31 is 32*
- *the age at location 32 is*

Solution

- *the age at location 21 is 29*
- *the age at location 31 is 32*
- *the age at location 32 is 32*

Solution (Continued)

To find the median, note the size of the data set is $n = 41$, which is odd.

Solution (Continued)

To find the median, note the size of the data set is $n = 41$, which is odd.

So the median is at location

$$\frac{41+1}{2} = 21.$$

Solution (Continued)

To find the median, note the size of the data set is $n = 41$, which is odd.

So the median is at location $\dfrac{41+1}{2} = 21$. Thus,

$$Median\ age = 29.$$

Mean

Definition

The **mean** or **average** of a data set is equal to: the sum of all data points, divided by the size of the data set. The mean is usually denoted by $\bar{x}$, or μ.

Example

Pharmacy sales:

S	M	T	W	R	F	S
\$2,548,	\$1,225,	\$1,732,	\$1,871,	\$975,	\$2,218,	\$1,339.

Find the **average** of the daily sales.

Solution

1. *add up the numbers and divide by 7*

$$\frac{2,548 + 1,225 + 1,732 + 1,871 + 975 + 2,218 + 1,339}{7}$$

Solution

1. *add up the numbers and divide by 7*

$$\frac{2,548 + 1,225 + 1,732 + 1,871 + 975 + 2,218 + 1,339}{7}$$

So

$$\bar{x} = \mu = \frac{\$11,908}{7} = \$1,701.14$$

Mean from Frequency Table (or Bar Graph)

Quick notation aside: The symbol $\sum$ just means “add them all up”, so

$$\sum f_i$$

means the sum of all f_i’s and

Quick notation aside: The symbol $\sum$ just means "add them all up", so

$$\sum f_i$$

means the sum of all f_i's and

$$\sum (x_i \cdot f_i)$$

means the sum of all products $x_i \cdot f_i$.

Example

Let $\mathbf{x} = [1, 4, 1, 2, 0, -3]$. Then

$$\sum x_i = 1 + 4 + 1 + 2 + 0 + (-3) = 5$$

Example

Let $\mathbf{x} = [1, 4, 1, 2, 0, -3]$. Then

$$\sum x_i = 1 + 4 + 1 + 2 + 0 + (-3) = 5$$

Let $\mathbf{f} = [2, 0, -1, 1, 2, 0]$. Then

$$\sum (x_i \cdot f_i) = 1 \cdot 2 + 4 \cdot 0 + 1 \cdot (-1) + 2 \cdot 1 + 0 \cdot 2 + (-3) \cdot 0 = 3$$

Example

Compute the mean from a frequency table:

Value	x_1	x_2	$\cdots$	x_{m-1}	x_m
Freq.	f_1	f_2	$\cdots$	f_{m-1}	f_m

- Each value x_i has to be counted f_i times.

- Each value x_i has to be counted f_i times.
- But

$$\underbrace{x_i + x_i + \cdots + x_i}_{f_i} = x_i \cdot f_i$$

- Each value x_i has to be counted f_i times.
- But

$$\underbrace{x_i + x_i + \cdots + x_i}_{f_i} = x_i \cdot f_i$$

- Also, the size of the data set is the sum of all the frequencies.

Formula for the average:

$$\begin{aligned}
\bar{x} &= \mu \\
&= \frac{(x_1 \cdot f_1) + (x_2 \cdot f_2) + \cdots + (x_m \cdot f_m)}{f_1 + f_2 + \cdots + f_m} \\
&= \frac{\sum (x_i \cdot f_i)}{\sum f_i}
\end{aligned}$$

Example

Quiz scores:

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

Example

Quiz scores:

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

Find the mean.

Solution

- *Create a row with the products* $x_i \cdot f_i$

Solution

- *Create a row with the products* $x_i \cdot f_i$
- *Create a totals column*

Solution

- *Create a row with the products $x_i \cdot f_i$*
- *Create a totals column*

x	4	5	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	25	Tot.
f	1	1	2	2	3	5	9	12	11	13	9	8	7	5	3	2	1	1	95
$x \cdot f$	4	5	16	18	30	55	108	156	154	195	144	136	126	95	60	42	22	25	1391

Solution

- *Create a row with the products $x_i \cdot f_i$*
- *Create a totals column*

x	4	5	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	25	Tot.
f	1	1	2	2	3	5	9	12	11	13	9	8	7	5	3	2	1	1	95
$x \cdot f$	4	5	16	18	30	55	108	156	154	195	144	136	126	95	60	42	22	25	1391

$$\bar{x} = \mu = \frac{1391}{95} = 14.64$$

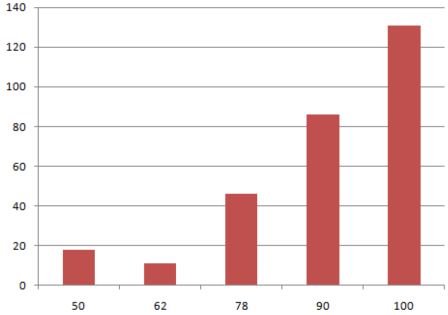
$\boxed{\text{Raw Data}}$

$\boxed{\text{Raw Data}} \longrightarrow \boxed{\text{Frequency Table}}$

$\boxed{\text{Raw Data}} \longrightarrow \boxed{\text{Frequency Table}} \longrightarrow \boxed{\begin{array}{l}\text{Median}\\ \text{Mean}\end{array}}$

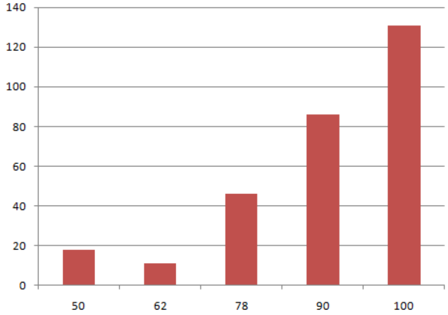
Raw Data $\longrightarrow$ Frequency Table $\longrightarrow$ Median
Mean

$\downarrow$



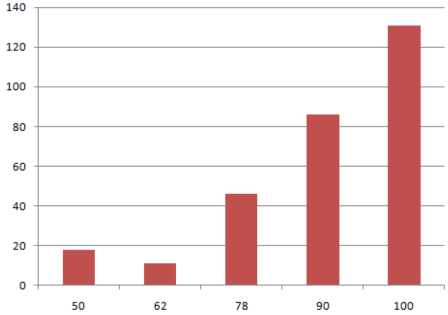
Raw Data $\longrightarrow$ Frequency Table $\longrightarrow$ Median Mean

$\downarrow$



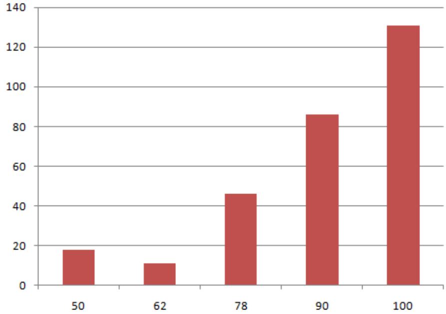
$\boxed{\text{Raw Data}} \longrightarrow \boxed{\text{Frequency Table}} \longrightarrow \boxed{\begin{array}{l}\text{Median}\\\text{Mean}\end{array}}$

$\updownarrow$



Raw Data $\longrightarrow$ Frequency Table $\longrightarrow$ Median
Mean

$\updownarrow$



We can recover median and mean from the bar graph!

Median v/s Mean

We have two “central locations” of a data set: median and mean.

We have two "central locations" of a data set: median and mean.

Question

Which is better?

We have two “central locations” of a data set: median and mean.

Question

Which is better?

Answer: Depends.

Question

Are the median and the mean even different?

Example

Pharmacy sales:

S	M	T	W	R	F	S
\$2,548,	\$1,225,	\$1,732,	\$1,871,	\$975,	\$2,218,	\$1,339.

Find the **median** and the **mean**.

Example

Pharmacy sales:

S	M	T	W	R	F	S
\$2,548,	\$1,225,	\$1,732,	\$1,871,	\$975,	\$2,218,	\$1,339.

Find the **median** and the **mean**.

Median=\$1,732 and
Mean=\$1,701.14.

So median and mean **don't have to be equal**, but often they are close.

Example

House prices in a city block:

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

Example

House prices in a city block:

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

Compare the mean and the median.

Solution

The median is $114,000.

Solution

The *median* *is $114,000.*
For the average we get:

$$\begin{aligned}\bar{x} &= \frac{75+96+107+110+110+118+130+135+150+520}{10}\\ &= 155.1\end{aligned}$$

Solution

The median is \$114,000.
For the average we get:

$$
\begin{aligned}
\bar{x} &= \frac{75+96+107+110+110+118+130+135+150+520}{10} \\
&= 155.1
\end{aligned}
$$

So the average is \$155,100. (Much higher!)

Remarks:

- All but one house is below average.

Remarks:

- All but one house is below average.
- The most expensive house is so expensive it brings the average up.

Remarks:

- All but one house is below average.
- The most expensive house is so expensive it brings the average up.
- This house is called an outlier.

Observation

Outliers distort the average, and make it not a very accurate measurement of the center.

The median would stay the same if the value in the last data point was \$160K or \$2M.

The median would stay the same if the value in the last data point was \$160K or \$2M.

Observation

The median is immune to outliers.

Not the mean! If the value of the last house was

- \$160K, the average would drop to \$119.1K

Not the mean! If the value of the last house was

- \$160K, the average would drop to \$119.1K
- \$2M, the average would jump up to \$299.1K

Not the mean! If the value of the last house was

- $160K, the average would drop to $119.1K
- $2M, the average would jump up to $299.1K

Observation

The mean is highly sensitive to outliers.

Example

Suppose someone wants to buy, and can afford a \$120K house.

Example

Suppose someone wants to buy, and can afford a $120K house.

- The median, $114K, shows they can afford at least half the houses.

Example

Suppose someone wants to buy, and can afford a \$120K house.

- The median, \$114K, shows they can afford at least half the houses.
- But if they just see the average, \$155.1K, they may get scared away.

Here the median is a better central locator than the mean.

Example

In every airplane there are three of each instrument (speedometer, altimeter, etc.)

Example

In every airplane there are three of each instrument (speedometer, altimeter, etc.)

- If 2 show different values, the third shows which one is broken

Example

In every airplane there are three of each instrument (speedometer, altimeter, etc.)

- If 2 show different values, the third shows which one is broken
- Actual speed (high, etc.) is determined as a **median** of 3 measurements.

Example

- If one instrument is broken, it may significantly affect the average

Example

- If one instrument is broken, it may significantly affect the average
- But the median should not change much.

Example

- In exam statistics, the **median** is more important, since it's not affected by low outliers (e.g., people who skipped the exam).

Example

- In exam statistics, the **median** is more important, since it's not affected by low outliers (e.g., people who skipped the exam).
- If you got above the median, then you're in the top half.

Example

- In exam statistics, the **median** is more important, since it's not affected by low outliers (e.g., people who skipped the exam).
- If you got above the median, then you're in the top half.
- Can't say the same about the mean.

There are also situations when outliers are significant.

Example

Suppose the pharmacy had a big influx on Sunday, and instead of \$2,548, sales were \$12,000.

Example

Suppose the pharmacy had a big influx on Sunday, and instead of \$2,548, sales were \$12,000.

- the median doesn't change

Example

Suppose the pharmacy had a big influx on Sunday, and instead of \$2,548, sales were \$12,000.

- the median doesn't change
- the average jumps from \$1,701.14 to \$3,051.43.

Example

Suppose the pharmacy had a big influx on Sunday, and instead of \$2,548, sales were \$12,000.

- the median doesn't change
- the average jumps from \$1,701.14 to \$3,051.43.

For the owner, this outlier is very significant, and should certainly be reflected in the central location indicator.

Here it is more appropriate to use the mean as central locater, rather than the median.

Example

- Now suppose there were mistakes on the sales figures for Sunday, Thursday and Saturday.

Example

- Now suppose there were mistakes on the sales figures for Sunday, Thursday and Saturday. Each of those days, the sales figures were \$350 higher:

Old sales:
$[\$2,548, \$1,225, \$1,732, \$1,871, \$975, \$2,218, \$1,339]$
Corrected sales:
$[\$2,898, \$1,225, \$1,732, \$1,871, \$1,325, \$2,218, \$1,689]$.

Example

- Some of the values changed, but none of them moved across the old median.

Example

- Some of the values changed, but none of them moved across the old median.
- So the median remains the same, \$1,732.

Example

- Some of the values changed, but none of them moved across the old median.
- So the median remains the same, $1,732.
- On the other hand, the average goes up by $150 to $1,851.14.

Observation

The median is insensitive to small changes in the data set, whereas the average is sensitive. (Either can be good or bad, depending.)

Next time: Section 4.1.4: Dispersion: Standard Deviation, Five-Number Summary.