BBA182 Applied Statistics
Week 3 (2) Using numerical data to describe data

DR SUSANNE HANSEN SARAL

EMAIL: SUSANNE.SARAL@OKAN.EDU.TR

HTTPS://PIAZZA.COM/CLASS/IXRJ5MMOX1U2T8?CID=4#

WWW.KHANACADEMY.ORG

Using numerical measures to describe data

«Is the data in the sample centered or located around a specific value?»

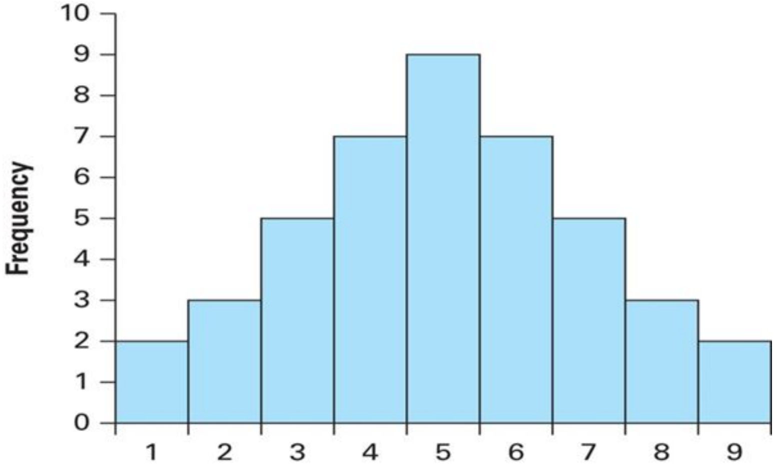
First question that business people, economists, corporate executives, etc. ask when presented with sample data.

Using numerical measures to describe data

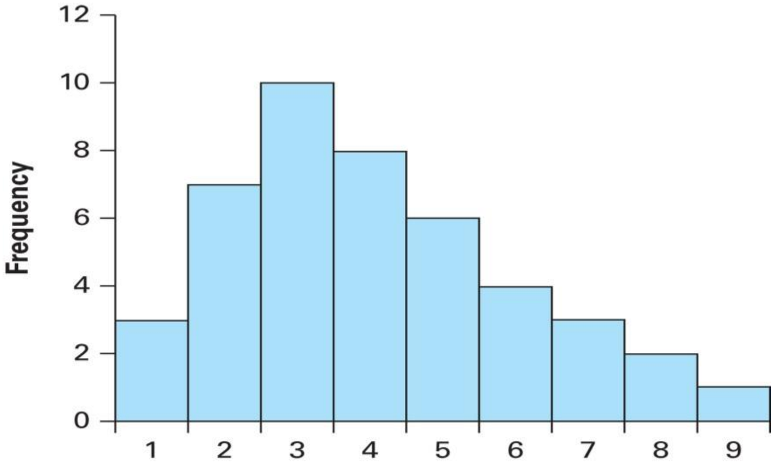
The **histogram** gives an idea whether the data is centered around a specific value.

The **histogram** provides a visual picture of how the data is distributed (symmetric, skewed, etc.)

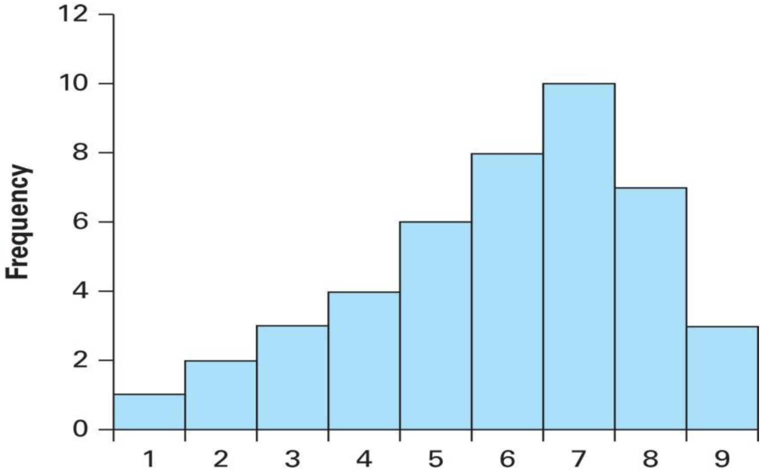
Is the data centered around a specific value?



Copyright ©2013 Pearson Education, publishing as Prentice Hall

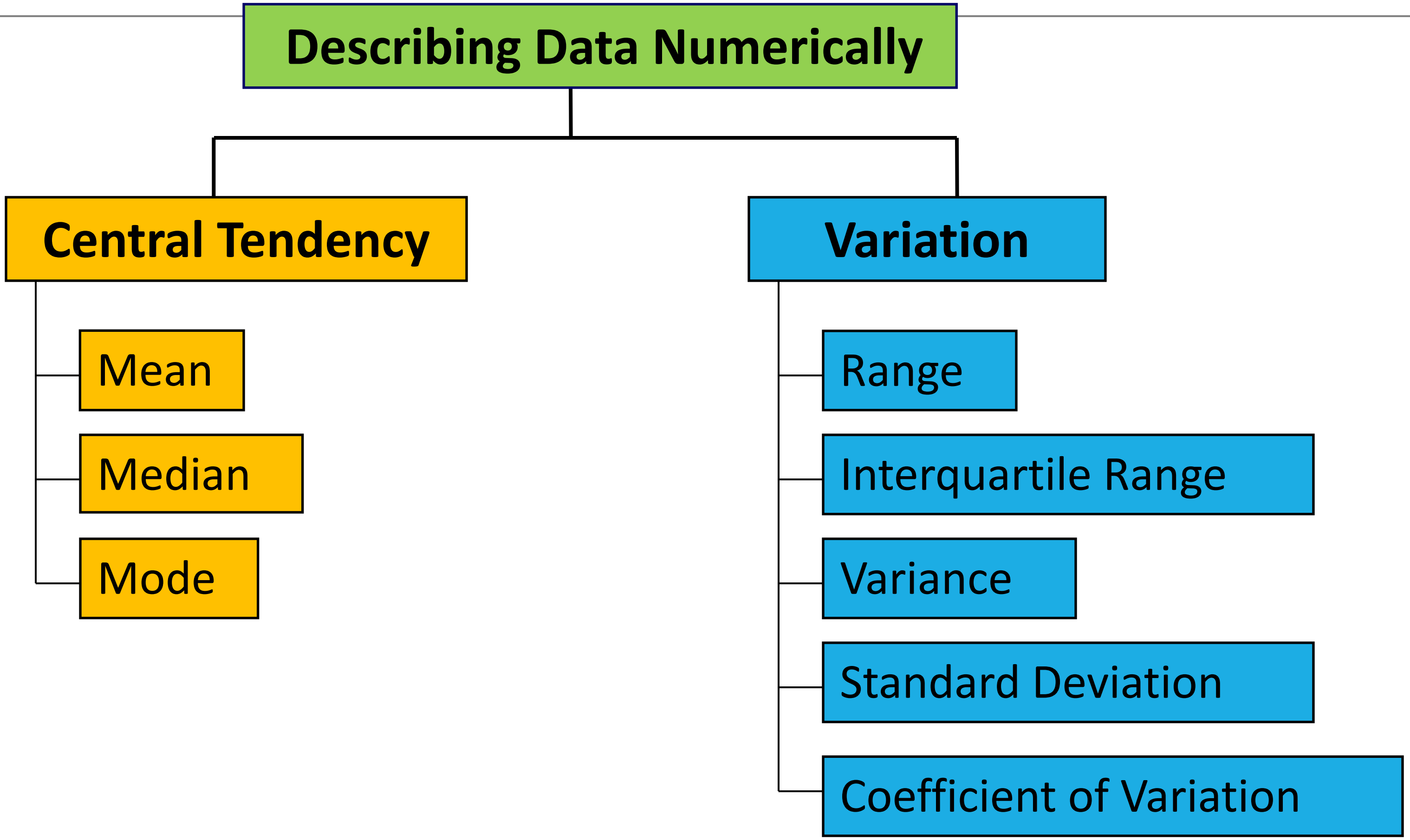


Copyright ©2013 Pearson Education, publishing as Prentice Hall



Copyright ©2013 Pearson Education, publishing as Prentice Hall

Numerical measures to describe data



Describing Data Numerically

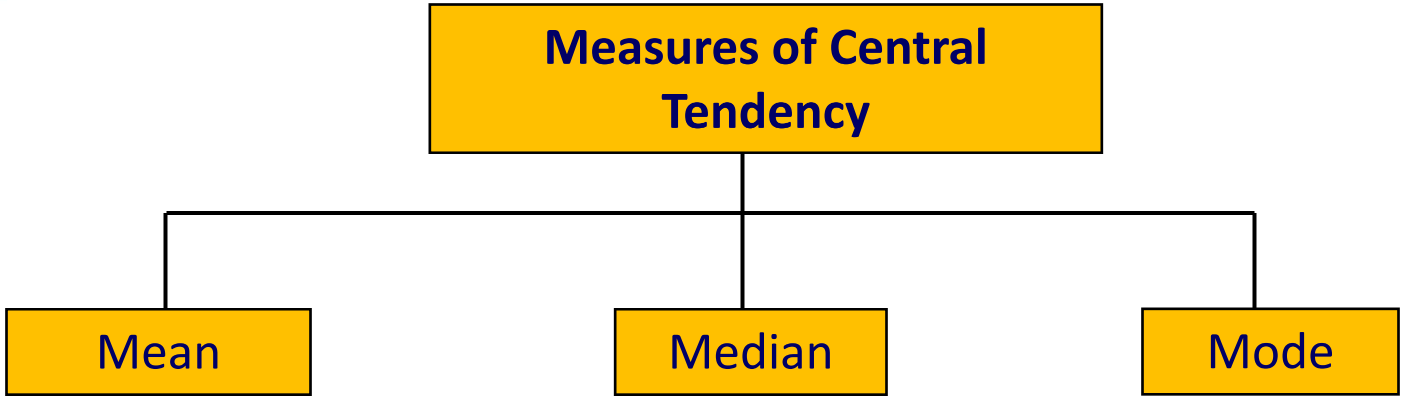
Central Tendency

- Mean
- Median
- Mode

Variation

- Range
- Interquartile Range
- Variance
- Standard Deviation
- Coefficient of Variation

Measures of the center of the data set



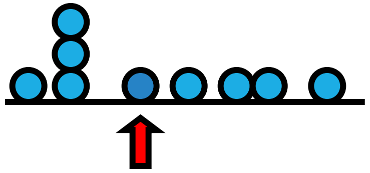
Measures of Central Tendency

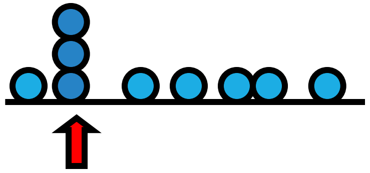
Mean

Median

Mode

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$





Arithmetic average of the data

Midpoint of ranked/ordered values in the data

Most frequently observed value in the data

(if one exists)

Mean

Population mean, μ

The mean is the most common measure of the center of a data set

- For a population of N values:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

Population values

Population size

Mean

Sample Mean, $\bar{x}$

- For a sample of **n** values:

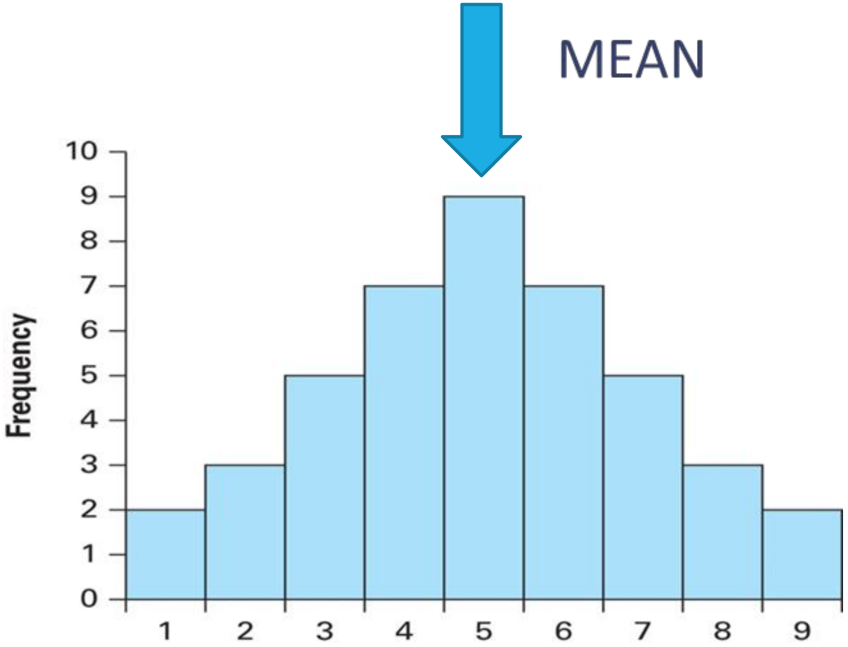
$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Observed values

Sample size

The Mean
symmetry and unimodal
ribution

WHEN WE HAVE A ***SYMMETRIC*** DISTRIBUTION WITH ONE MODE, THEN THE MEAN REPRESENTS THE **MIDDLE VALUE** IN A DATA SET.



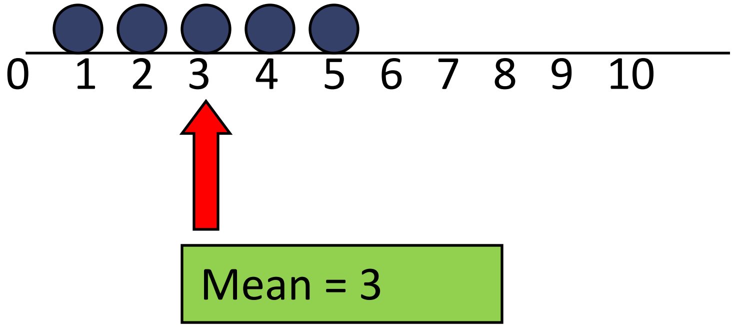
Copyright ©2013 Pearson Education, publishing as Prentice Hall

Mean

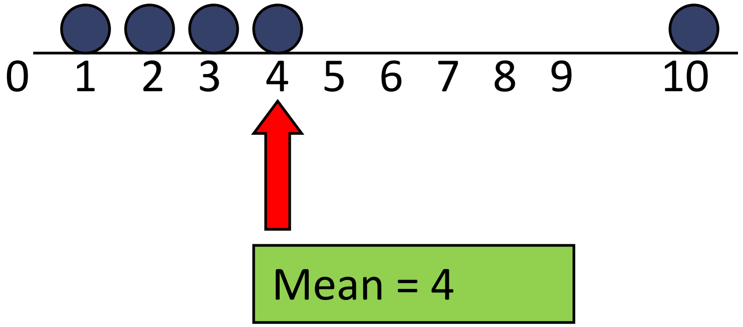
(continued)

The most common measure for the center of a data set

Affected by **extreme values (outliers)**



Mean = 3



Mean = 4

$$\frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$

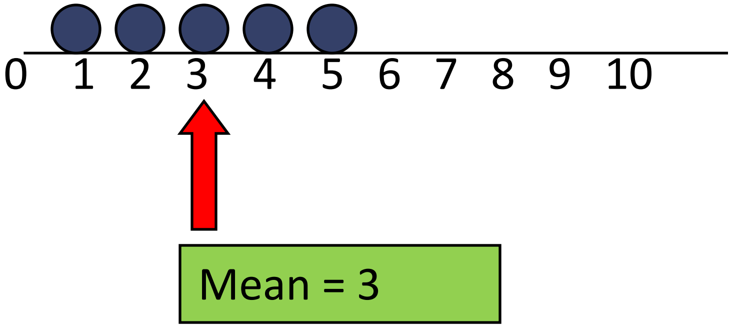
$$\frac{1+2+3+4+10}{5} = \frac{20}{5} = 4$$

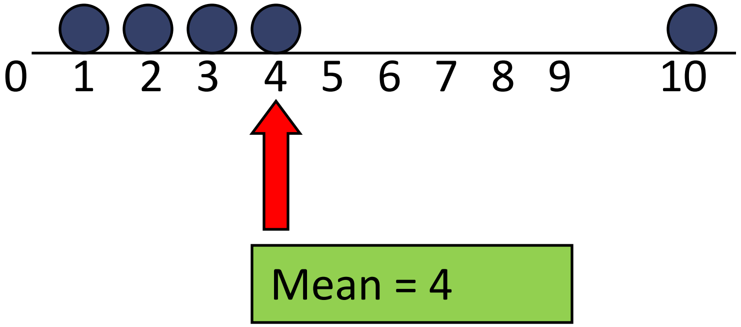
Mean

(continued)

The most common measure for the center of a data set

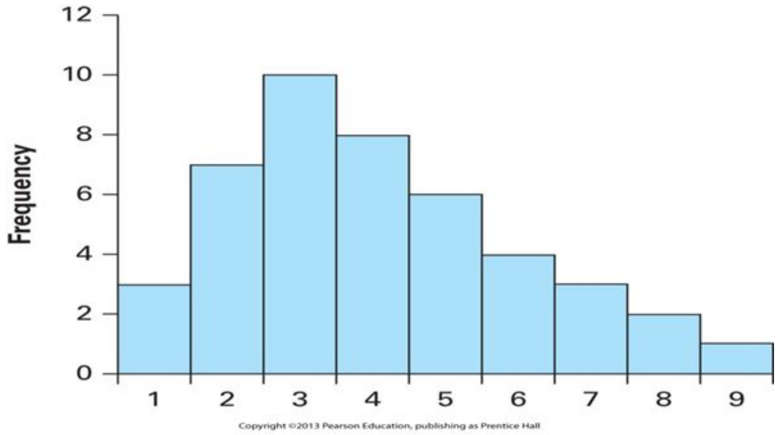
Affected by **extreme values (outliers)**





Skewed distribution

- An outlier will distort the picture of the data.
- It will inflate or deflate the mean, depending on the value of the outlier
- This creates a skewed distribution.

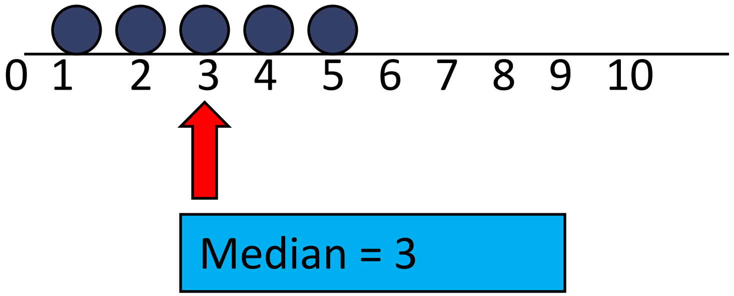


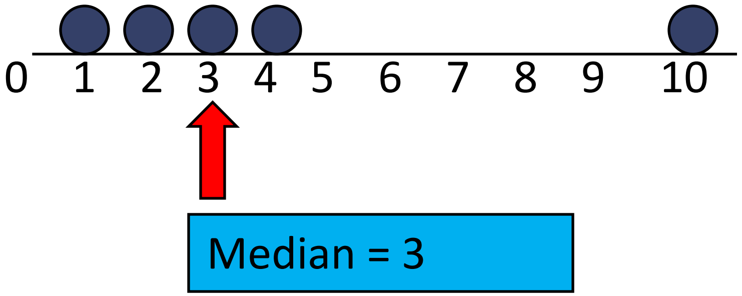
Copyright ©2013 Pearson Education, publishing as Prentice Hall

In this case we may want to use a different measure of the data center

Median

In an **ordered** list of data, the median is the “middle” number (50% above, 50% below)





Not affected by outliers

Finding the Median

The **location** of the median:

- If the number of values is **odd** (uneven), the median is the **middle** number

- 17 6 25 -5 13 9 33

For this data set: -17 -5 6 9 13 25 33

Finding the Median

The **location** of the median:

If the number of values is **even**, the median is the two middle numbers divided by 2

inding the median

Determine the median of the following data set:

17 5 3 11 12 8 25 3

OKAN ÜNİVERSİTESİ
İSTANBUL

inding the median

Determine the median of the following data set:

17 5 3 11 12 8 25 3

3 3 5 8 11 12 17 25

Median: 8 +11 = 19/ 2 = 9.5

Mode

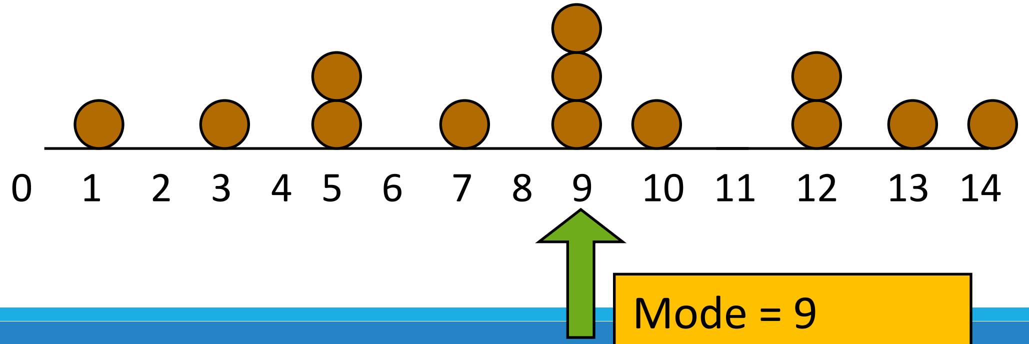
Value that occurs most often in the data set

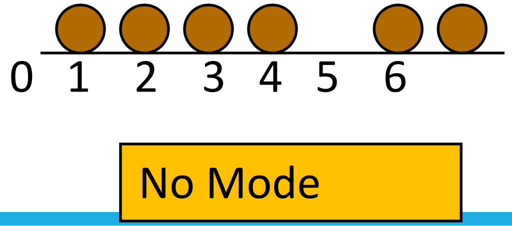
Not affected by **outliers**

Used for either **numerical or categorical** data

There may be no mode

There may be several modes, uni-modal, bi-modal, multimodal



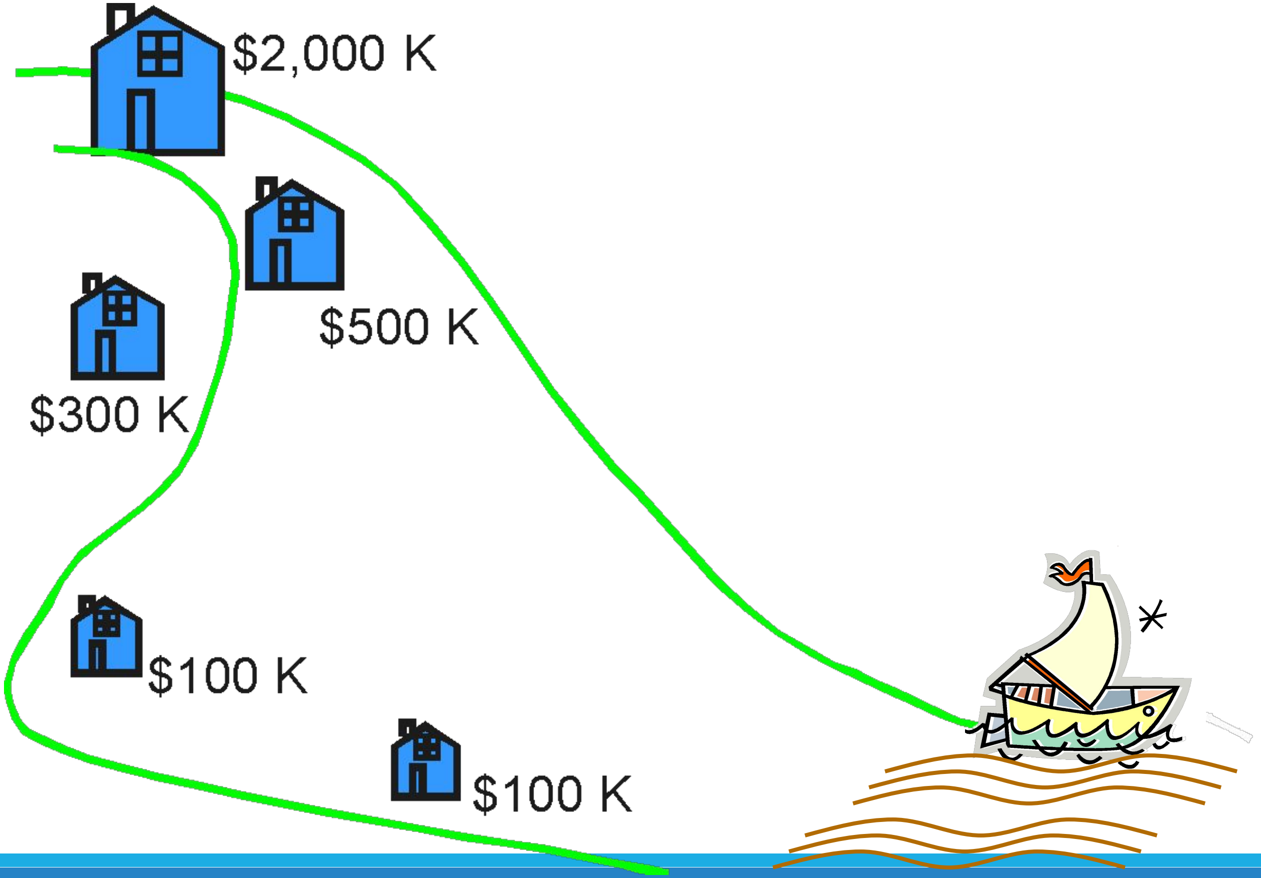


Measures of the center summary data

Five houses on a hill by the beach

House Prices:

$2,000,000
500,000
300,000
100,000
100,000



Measures of the center summary data

House Prices:

$2,000,000
500,000
300,000
100,000
100,000

Sum 3,000,000

What is the mean house price?

What is the median house price?

What is the modal house price?

Measures of the center - summary

House Prices:

$2,000,000
500,000
300,000
100,000
100,000

Sum 3,000,000

Mean: ($3,000,000/5)

= **$600,000**

Median: middle value of ranked data
= **$300,000**

Mode: most frequent house price
= **$100,000**

When is which measure of the center the "best"?

- **Mean** is generally used, **unless outliers** exist. If there are outliers the mean does not represent the center well.

- Then **median** is used when outliers exist in the data set.

  - Example: Median home prices may be reported for a region – less sensitive to outliers

Shape of a Distribution
Describe the shape of a distribution

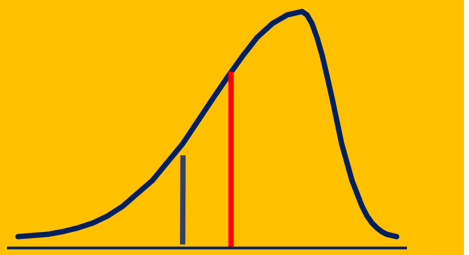
Describes how data is distributed

The presence or not of outliers in a data set, influence the shape of a distribution

- Symmetric or skewed

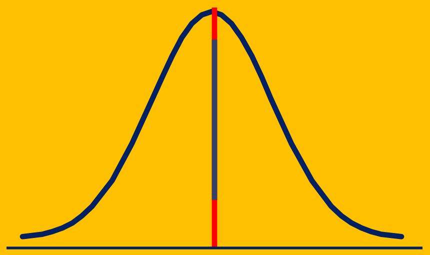
Left-Skewed

Mean < Median



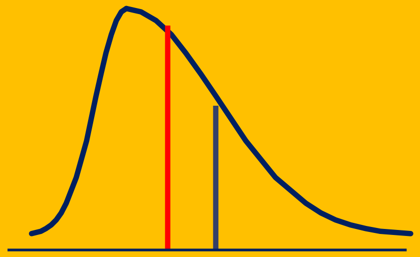
Symmetric

Mean = Median=Mode

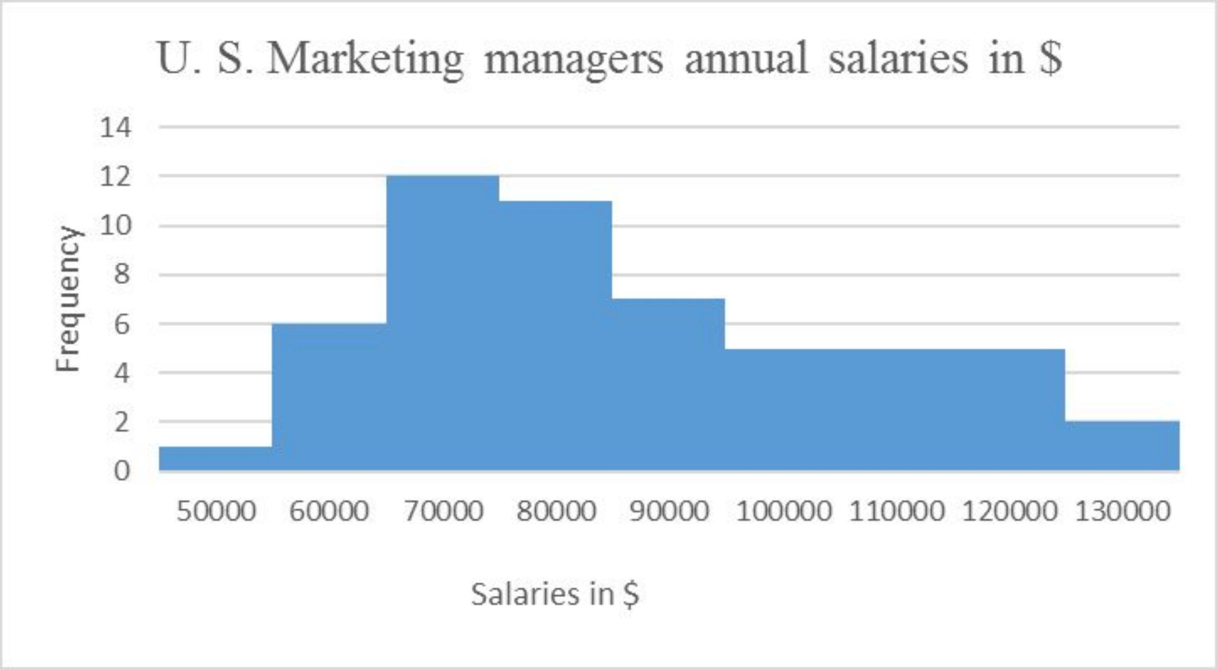


Right-Skewed

Median < Mean



Histogram of annual salaries (in $) for a sample of U.S. marketing managers:



- Describe the shape of this histogram (of the distribution)

- **Without doing calculations**. Do you expect the mean salary to be higher or lower than the median salary?

Class exercise

Eleven economists were asked to predict the percentage growth in the Consumer Price Index over the next year.

Their forecasts were as follows:

3.6 3.1 3.9 3.7 3.5 1.0 3.7 3.4 3.0 3.7 3.4

a) Compute the mean, median and the mode
b) Are there any outliers in the data set that may influence the value of the mean?
c) If there are outliers, how do they affect the shape of the data distribution?

Solution to class exercise

Mean: 36/11 = 3.27 rounded up to 3.3

Median: 3.5

Mode: 3.7

Outlier: 1.0

How does the outlier affect the shape of the distribution?

It decreases the average of the data set and distorts the picture of the histogram.

The shape is skewed to the left.

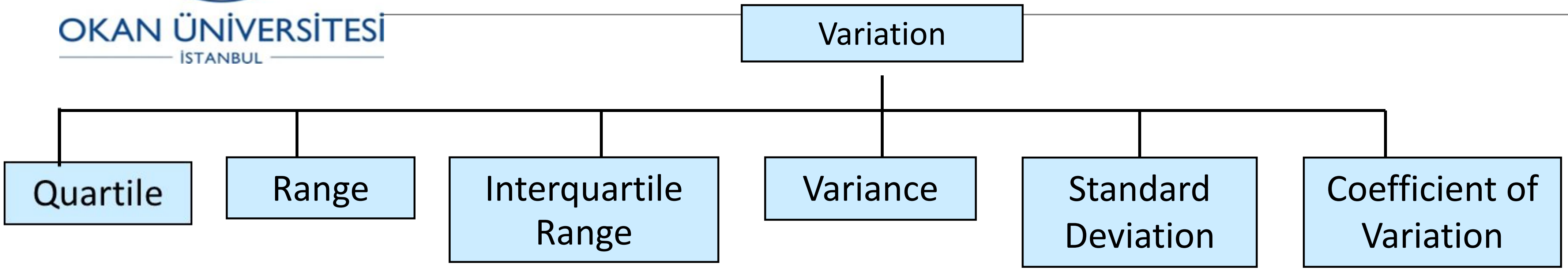
Measures of variability

The three measures of data center do not provide complete and sufficient description of the data.

Next to knowing how data is located around **a specific value** (mean, median or mode), we need information on **how far the data is spread from that specific value**, most often from the mean.

The **measure of variability** will provide us with this information.

Measures of Variability



Variation

Quartile

Range

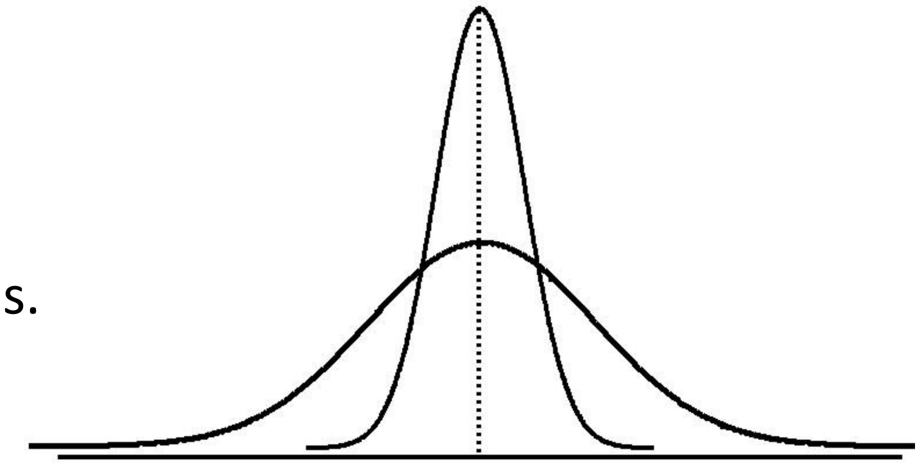
Interquartile Range

Variance

Standard Deviation

Coefficient of Variation

- Measures of variation give information about the spread or variability of the data values.



Same center, different variation

Quartiles

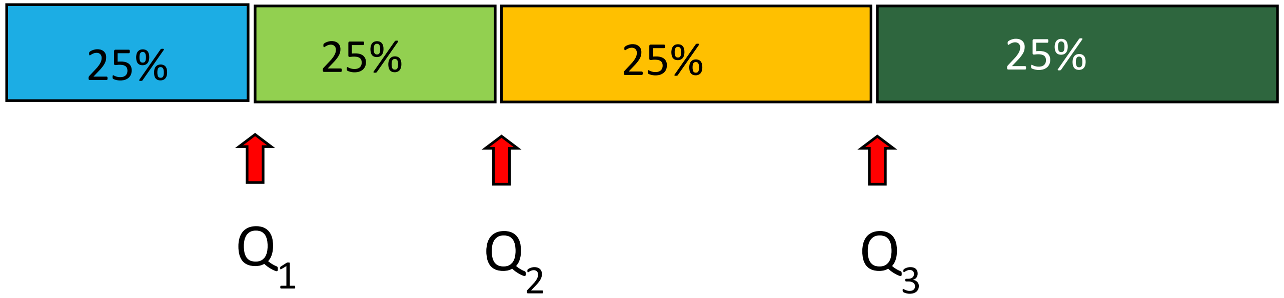
Quartiles are descriptive measures that separate large data set into four quarters.

The **first quartile** ($\boldsymbol{Q_1}$) separates approximately the smallest 25 % of the data from the remaining largest 75 % of the data.

The **second quartile** ($\boldsymbol{Q_2}$)**,** is the median, which separates the data set into two identical halves.

The **third quartile** ($\boldsymbol{Q_3}$) separates approximately the smallest 75 % of the data from the remaining largest 25 % of the data

Quartiles



- **The first quartile, Q_1,** is the value for which 25% of the observations in the data set are smaller and 75% are larger
- **The second quartile, Q_2** is the same as **the median** (50% are smaller, 50% are larger)
- Only 25% of the observations in the data set are greater than the **third quartile, Q_3**

How to calculate quartiles manually

Find a quartile by determining the value in the **appropriate position of the ranked data**, where

First quartile **position**: $Q_1 = 0.25(n+1)$

Second quartile **position**: $Q_2 = 0.50(n+1)$
(the median position)

Third quartile **position**: $Q_3 = 0.75(n+1)$

where n is the number of observed values

Quartiles

- Example: Find **the first** and third quartile

  14 12 16 21 11 17 22 16 18

$Q_1 = 0.25(n+1)$

Sample **Ranked** Data: 11 12 14 16 16 17 18 21 22

Q_1

(n = 9

1^{st} Quartile = the value located in the 0.25(n+1)th **ordered** position

1st Quartile = value located in the 0.25(9+1)th **ordered** position

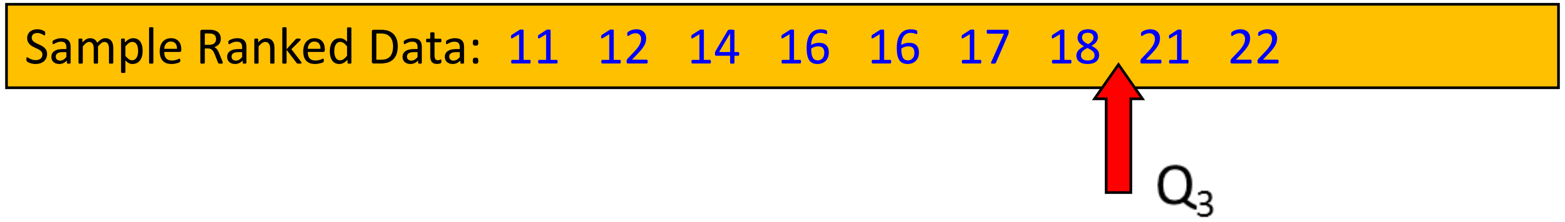
1st Quartile = value located in the 2.5th position

The value in the 2^{nd} position is 12 and the value in the 3rd position is 14. The value in the 2.5th position is 50 % of the distance between 12 and 14. The value of the first quartile therefore: 12 + 0.5(14-12) = **13**

Quartiles

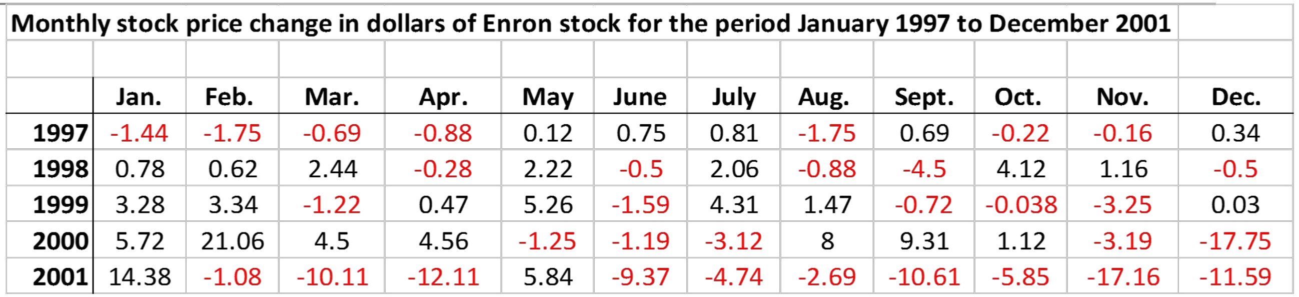
- Example: Find the first and **third quartile**

$Q_3 = 0.75(n+1)$



Sample Ranked Data: 11 12 14 16 16 17 18 21 22

Quartiles and Enron case



Monthly stock price change in dollars of Enron stock for the period January 1997 to December 2001

	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
1997	-1.44	-1.75	-0.69	-0.88	0.12	0.75	0.81	-1.75	0.69	-0.22	-0.16	0.34
1998	0.78	0.62	2.44	-0.28	2.22	-0.5	2.06	-0.88	-4.5	4.12	1.16	-0.5
1999	3.28	3.34	-1.22	0.47	5.26	-1.59	4.31	1.47	-0.72	-0.038	-3.25	0.03
2000	5.72	21.06	4.5	4.56	-1.25	-1.19	-3.12	8	9.31	1.12	-3.19	-17.75
2001	14.38	-1.08	-10.11	-12.11	5.84	-9.37	-4.74	-2.69	-10.61	-5.85	-17.16	-11.59

In the Enron data we had 60 data points.

There are 30 values to right and 30 values to left side of the median (Q_2):

(Q_1) = -$1.68 (between15th and 16th data points) - 75 % of the data is larger than -$ 1.68

(Q_2) = -$ 0.19 **median (between 30th and 31st points) -** 50 % of the data is smaller than -$.19 and 50 % of the data is larger than -$.19 .

(Q_3) = $2.14 (between 45th and 46th data pots) - 25 % of the data is larger than $2.14

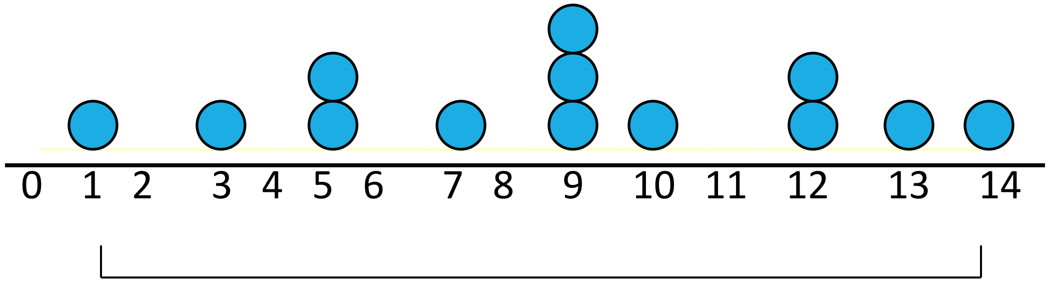
Range

Simplest measure of variation

Difference between the largest and the smallest observations:

$$\text{Range} = X_{\text{largest}} - X_{\text{smallest}}$$

Example:



Range = 14 - 1 = 13

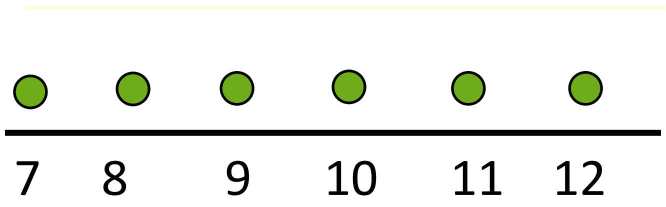
Range – Example Enron case

Range = Maximum value – minimum value

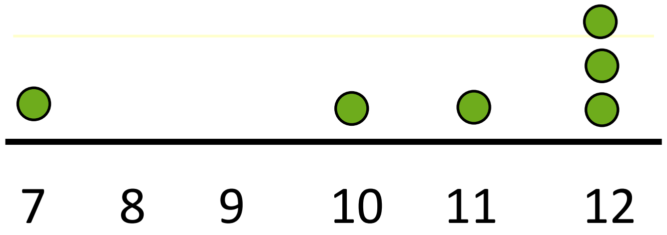
Enron data range = $21.06 – (-$17.75) = **$ 38.81**

Disadvantages of the Range

Ignores the way in which data is distributed



Range = 12 - 7 = 5



Range = 12 - 7 = 5

Disadvantages of the Range

Sensitive to outliers

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,5

Range = 5 - 1 = 4

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,120

Range = 120 - 1 = 119

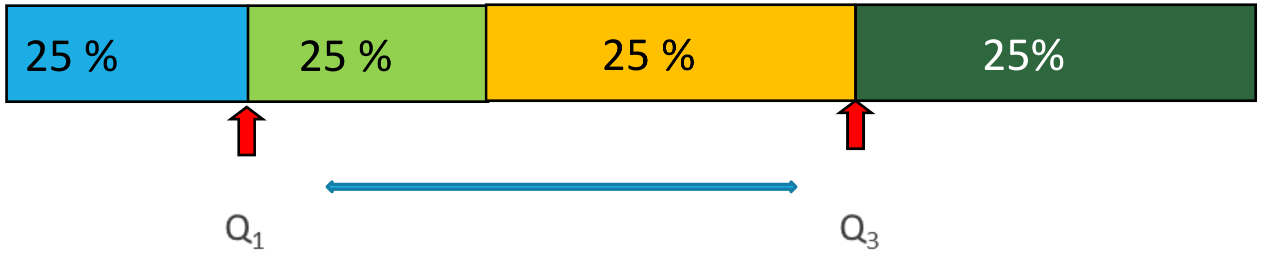
Range: short-comings as a good measure for iability

Because the range does not provide us with a lot of information about the spread of the data it is not a very good measure for variability.

Interquartile Range

We can eliminate some outlier problems by using the interquartile range and concentrate on the **middle 50 % of the data** in the data set

Eliminate high- and low-valued observations and calculate the range of the middle 50% of the data



Interquartile range

The Interquartile range, IQR = $Q_3 - Q_1$

Interquartile Range

The interquartile range (IQR) measures the spread of the data in the middle 50% of the data set

Defined as the difference between the observation at the **third quartile** and the observation at the **first quartile**

$$IQR = Q_3 - Q_1$$

Interquartile Range

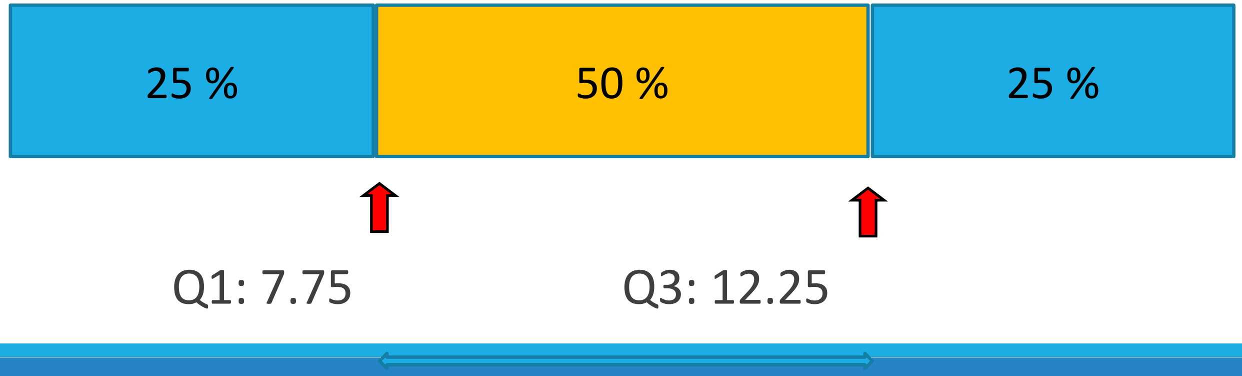
Raw data: 6 8 10 12 14 9 11 7 13 11 n = 10

Ranked data: 6 7 8 9 10 11 11 12 13 14

1. Quartile: 7.75

3. Quartile: 12.25

IQR = Q3 – Q1 = 12.25 – 7.75 = 4.5



25 % 50 % 25 %

Q1: 7.75 Q3: 12.25

Enron data: Interquartile range

Interquartile range: **IQR = Q_3 - Q_1**

$(Q_1) = -\$1.68$
$(Q_3) = \ \$2.14$

IQR : \$2.14 – (-\$ 1.68) = \$ 3.82

The middle 50 % of the Enron data has a spread of \$ 3.82 compared to the range of \$ 38. 81!