BBA182 Applied Statistics
Week 4 (1)Measures of variation

DR SUSANNE HANSEN SARAL

EMAIL: SUSANNE.SARAL@OKAN.EDU.TR

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

WWW.KHANACADEMY.ORG

Numerical measures to describe data

- **Describing Data Numerically**
  - **Central Tendency**
    - Mean
    - Median
    - Mode
  - **Variation**
    - Quartile
    - Range
    - Interquartile Range
    - Variance
    - Standard Deviation
    - Coefficient of Variation

terquatile range, IQR

Alternative way to calculate the IQR

Khan Academy

Interquartile range (IQR)

Practice finding the interquartile range (IQR) of a data set.

Get 5 questions correct in a row

A family of bears is going to a movie. The following data points represent the ticket and candy price (in dollars) for each bear.

Sort the data from least to greatest.

4	4	3	0	$7\frac{1}{2}$	$2\frac{1}{2}$	$1\frac{1}{2}$	7	4

Find the interquartile range (IQR) of the data set.

______ dollars

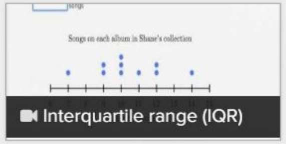
Answer

Check Answer

Show me how

I'd like a hint

Stuck? Watch a video.



Show scratchpad

Report a mistake in this question

Five-Number Summary of a data set

In **describing** numerical data, statisticians often refer to the **five-number summary.** It refers to five the **descriptive measures we have looked at**:

minimum value
first quartile
median
third quartile
maximum value

$$\boxed{\text{minimum} < Q_1 < \text{median} < Q_3 < \text{maximum}}$$

It gives us a good idea where the data is located and how it is spread in the data set

Five-Number Summary: Example

Sample **Ranked** Data:	6	7	8	9	10	11	11	12	13	14

minimum < Q_1 < median < Q_3 < maximum

6 < 7.75 < 10.5 < 12.25 < 14

Exercise

Consider the data given below:

110 125 99 115 119 95 110 132 85

a. Compute the mean.
b. Compute the median.
c. What is the mode?
d. What is the shape of the distribution?
e. What is the lower quartile, Q1?
f. What is the upper quartile, Q3?
g. Indicate the five number summary

Exercise

Consider the data given below.

85 95 99 *110* *110* *115* *119* *125* *132*

a. Compute the mean. 110

b. Compute the median. 110

c. What is the mode? 110

d. What is the shape of the distribution? Symmetric, because mean = median=mode

e. What is the lower quartile, Q1? 97

f. What is the upper quartile, Q3? 122

g. Indicate the five number summary $85 < 97 < 110 < 122 < 132$

Five number summary and Boxplots

Boxplot is created from the five-number summary

A **boxplot** is a graph for **numerical data** that describes **the shape of a distribution,** in terms of the 5 number summary.

It visualizes the **spread** of the data in the data set.

Five number summary and Boxplots

Boxplot is created from the five-number summary

The central box shows the middle half of the data from Q_1 to Q_3, (middle 50% of the data) with a line drawn at the median

Two lines extend from the box. One line is the line from Q_1 to the minimum value, the other is the line from Q_3 to the maximum value

A **boxplot** is a graph for numerical data that describes **the shape of a distribution,** like the histogram

Five number summary and boxplot

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

Minimum number = 1

Maximum number = 5

Q_1 = 1

$Q_{2=}$ 2.5

Median = 2

Five number summary: 1 = 1 < 2 < 2.5 < 5 (plot a dot chart, then boxplot)

Five number summary and boxplot

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

Minimum number = 1

Maximum number = 120

$Q_1 = 1$

$Q_2 = 2.5$

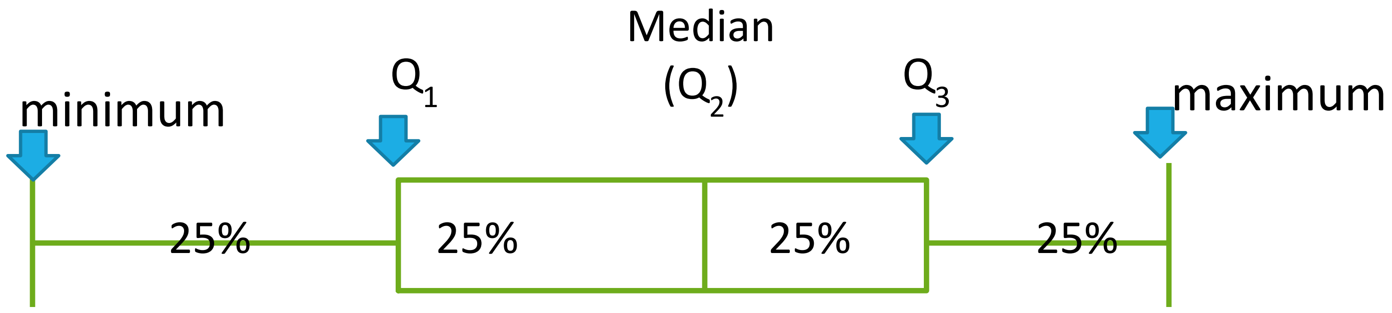
Median = 2

Five number summary: $\mathbf{1 = 1 < 2 < 2.5 < 120}$

Boxplot

The plot can be oriented horizontally or vertically

Example:



12	30	45	57	70

Gilotti's Pizza Sales in $100s

Table 2.2 Gilotti's Pizzeria Sales (in $100s)

Location 1	Location 2	Location 3	Location 4
6	1	2	22
8	19	3	20
10	2	25	10
12	18	20	13
14	11	22	12
9	10	19	10
11	3	25	11
7	17	20	9
13	4	22	10
11	17	26	8

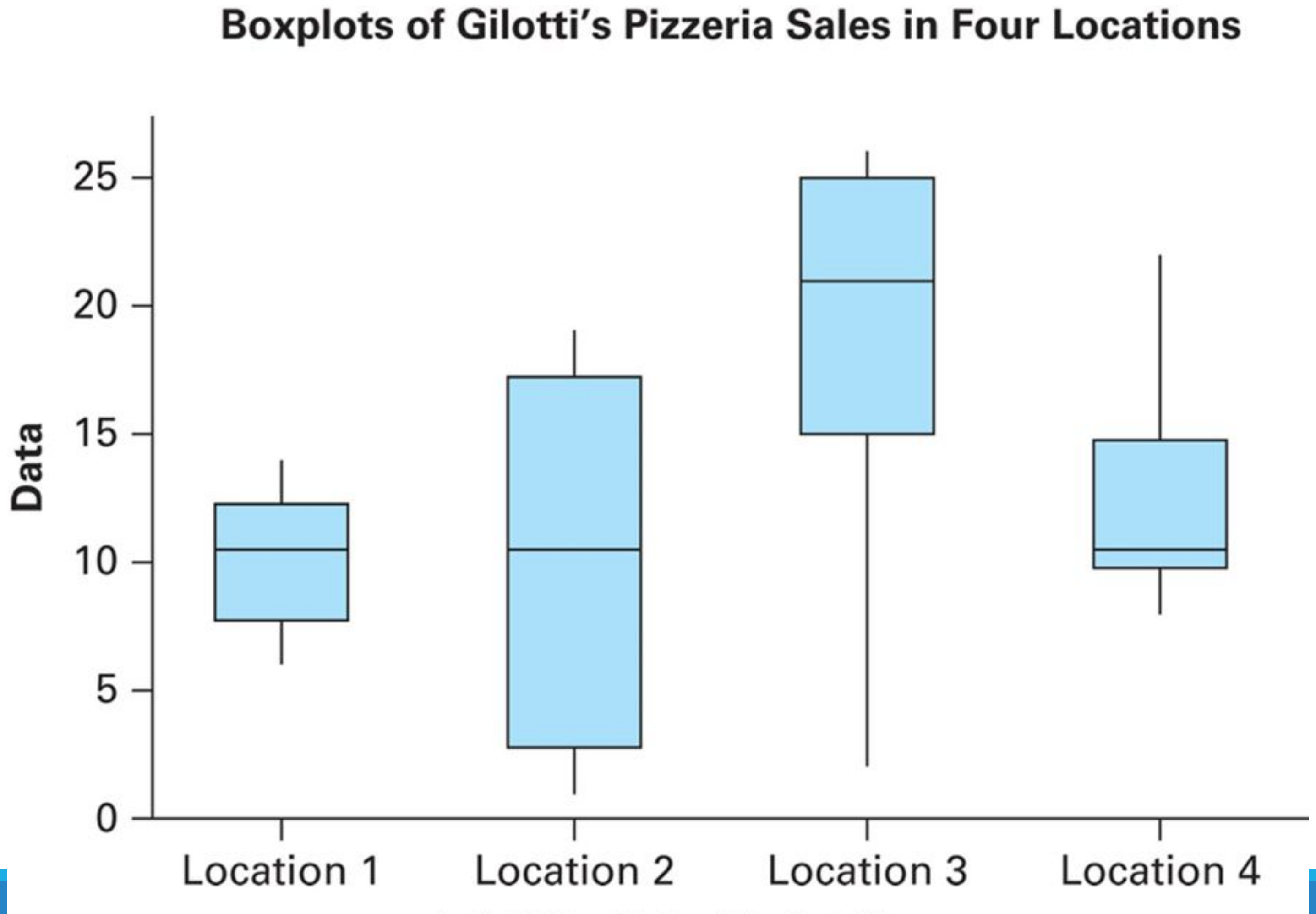
Gilotti's Pizza Sales

What are the shapes of the distribution of the four data set?

Table 2.3 Gilotti's Pizzeria Sales

Variable	Mean	Min.	Q_1	Median	Q_3	Max.	IQR	Range
Location 1	10.1	6.0	7.75	10.5	12.25	14.0	4.5	8.0
Location 2	10.2	1.0	2.75	10.5	17.25	19.0	14.5	18.0
Location 3	18.4	2.0	15.00	21.0	25.00	26.0	10.0	24.0
Location 4	12.5	8.0	9.75	10.5	14.75	22.0	5.0	14.0

Gilotti's Pizza Sales - boxplot

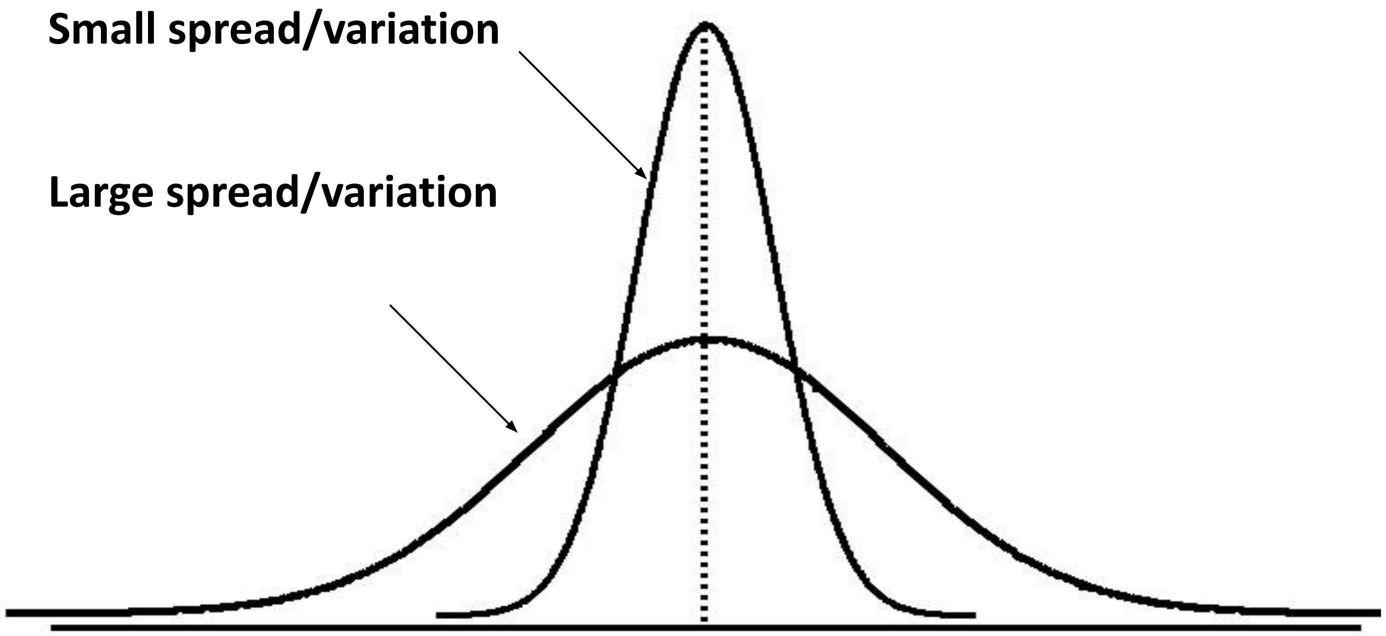


Gilotti's Pizza Sales in $100s

Table 2.2 Gilotti's Pizzeria Sales (in $100s)

Location 1	Location 2	Location 3	Location 4
6	1	2	22
8	19	3	20
10	2	25	10
12	18	20	13
14	11	22	12
9	10	19	10
11	3	25	11
7	17	20	9
13	4	22	10
11	17	26	8

Measuring variation in a data set that follows a *normal distribution*



Measuring variation in a data set

Data set 1 : 23 19 21 18 24 21 23 **Mean:** 21.3

Data set 2 : 23 35 19 7 21 24 22 **Mean**: 21.6

Which of these two data sets has the highest spread/variation? Why?

Average distance to the mean: Standard deviation

Most **commonly used measure of variability**

Measures the standard (average) distance of each individual data point from the mean.

Calculating the average distance to the mean

Our goal is to measure the standard distance of each single data in the data set from the mean.

1st step: Calculate the mean of the data set $\mu = \frac{\sum x_i}{N}$

2nd step: Calculate the standard distance from the mean is to determine distance from the mean for each individual score:

deviation score = X - μ

Where x is the value of each individual score and μ the population mean.

Calculating the average distance to the mean

Step 3: Once we have calculated the distance between each single score and the mean, we add up the those deviation scores. Our mean in this example is μ= 3.

Example: We have a set of 4 scores ($x_{1,}$ $x_{2,..}$ $x_{i,}$): 8, 1, 3, 0,

	X	X - μ	
1	8	5	(= 8 - 3)
2	1	-2	(= 1 - 3)
3	3	0	(= 3 - 3)
4	0	-3	(= 0 - 3)
Σ X	12	0	= ∑(X - μ)

Calculating the average distance to the mean

Notice that the deviation score adds up to *zero!*

This is not surprising because the **mean serves as balance point (middle point) for the distribution**. (!**Remember: In a symmetric distribution the mean and the median are identical**)

The distances of the single score above the mean equal the distances of the single scores below the mean.

Therefore the deviation score always adds up to zero.

Calculating the average distance to the mean

Step 3: The solution is to get rid of the + and – which causes the cancelling out effect. **We square each deviation score and sum them up**

	X	X - μ		$(x - \mu)^2$		
1	8	5	(= 8 - 3)	25		
2	1	-2	(= 1 - 3)	4		
3	3	0	(= 3 - 3)	0		
4	0	-3	(= 0 - 3)	9		
	12	0		38		

Population Variance, σ^2

Average of squared deviations from the mean

- Population variance:

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

Where:

μ = population mean

N = population size

x_i = i^{th} value of the variable x

OKAN ÜNİVERSİTESİ
İSTANBUL

Sample Variance, s^2

Average of squared deviations from the mean

- Sample variance:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Where:

$\bar{X}$ = arithmetic mean

n = sample size

X_i = i^{th} value of the variable X

Population Standard Deviation, σ

Most commonly used measure of variation in a population

Shows variation about the mean in a **symmetric** data set

Has the **same units as the original data,**

Example: If original data is in meters than the standard deviation will also be in meters.

- Population standard deviation:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

Sample Standard Deviation, s

Most commonly used measure of variation in a sample

Shows variation about the mean

Has the same units as the original data

- Sample standard deviation:

$$S = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Calculation Example: Sample Standard Deviation, s

$$S = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Sample Data (x_i) :

10	12	14	15	17	18	18	24

n = 8 Mean = $\bar{x}$ = 16

$$s = \sqrt{\frac{(10 - \bar{X})^2 + (12 - \bar{x})^2 + (14 - \bar{x})^2 + \cdots + (24 - \bar{x})^2}{n - 1}}$$

$$= \sqrt{\frac{(10 - 16)^2 + (12 - 16)^2 + (14 - 16)^2 + \cdots + (24 - 16)^2}{8 - 1}}$$

$$= \sqrt{\frac{130}{7}} = \boxed{4.3095}$$

A measure of the "average" distance about the mean

Class example

Calculating sample variance and standard deviation

Compute the variance, s^2, and standard deviation, s, of the following sample data:

6 8 7 10 3 5 9 8

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

...ass example (continued)

When we analyze the variance formula we, see that we need to calculate the sample mean, $\bar{X}$, first:

$$\bar{X} = \frac{6+8+7+10+3+5+9+8}{8} = \frac{56}{8} = 7$$

$$\boxed{s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

ass example (continued)

The mean = 7

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

$$S = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

6 8 7 10 3 5 9 8

Class example (continued)

Calculating the **sample variance**: **6 8 7 10 3 5 9 8**

$$s^2 = \frac{(6-7)^2 + (8-7)2 + (7-7)2 + (10-7)2 + (3-7)2 + (5-7)2 + (9-7)2 + (8-7)2}{8-1}$$

$$s^2 = \frac{1+1+0+9+16+4+4+1}{8-1}$$

$$s^2 = \frac{36}{8-1} = 5.14$$

Sample standard deviation, s = $\sqrt{5.14} = 2.27$ (average distance to the mean of 7)