

Topic #2. Numerical Measures of Numerical Data Sets

Cheng Peng

Contents

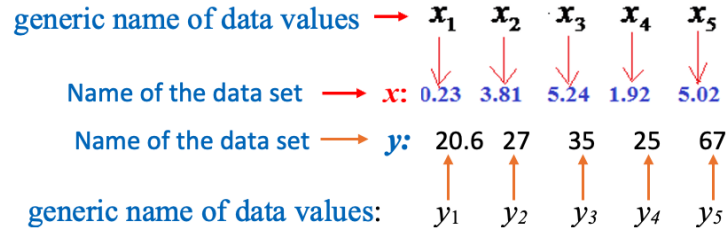
1	Introduction	1
2	Notations Using Greek Letters for Parameters	1
2.1	Big Sigma (Σ) Notation	2
2.2	Notations for Parameters and Statistics	2
3	Measures of Center	3
3.1	Mean	3
3.2	Median	3
3.3	Mode	4
3.4	Relationship between Mean, Median, and Mode	4
4	Measures of Variation	4
4.1	Variance	4
4.2	Standard Deviation	5
4.3	Steps for Calculating Variance	5
5	Measures of Location	5
5.1	z-score	5
5.2	Percentile	6
5.3	Applications of Numeric Measures	7
5.3.1	Five Number Summary	7
5.3.2	Box-plot	8
5.3.3	Inter-quartile Range (IQR)	9
6	Use of Technology	9
6.1	Example: Descriptive Statical Analysis of Course Grade	9
6.2	Exercises	11

1 Introduction

This note focuses on using numerical measures to characterize numerical data sets. The numerical measures are used to describe the features such as mean, variance, and percentiles of a given numerical data set. These numeric measures are classified into three categories: central tendency, variation, and locations.

2 Notations Using Greek Letters for Parameters

Every data set has a name. For example, the set of heights of a group of WCU students is a data set that can be named **h** or **height**. We can give each data value has a “generic name” such as h_1, h_2, \dots, h_5 , etc. The following figure gives other examples of generic names of values in different data sets.



2.1 Big Sigma (Σ) Notation

These “generic names” were used to make compact formulas in some numeric measures. The sum of all data values in the data set with the name x (in the above figure) is given by the following **big sigma** notation.

$$\sum_{i=1}^5 x_i = x_1 + x_2 + x_3 + x_4 + x_5$$

$$= 0.23 + 3.81 + 5.24 + 1.92 + 5.02 = 16.22$$

Example 1: Consider the following two data sets with names x and y , we want to take the product of the corresponding values and sum up the product of the corresponding values. The following is the **big sigma** notation of the **sum of the cross-product**.

x	1.1	1.2	1.3	1.4	1.5	1.6
y	5	4	3	2	1	0

$$\sum_{i=1}^4 x_i y_i = x_1 y_1 + x_2 y_2 + x_3 y_3 + x_4 y_4$$

$$= 1.1 \times 5 + 1.2 \times 4 + 1.3 \times 3 + 1.4 \times 2 = 17$$

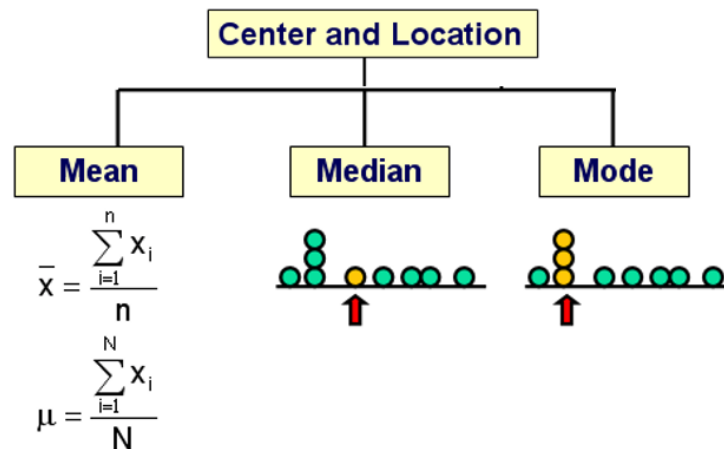
2.2 Notations for Parameters and Statistics

We use Greek letters to denote the population parameters of populations and English letters to denote statistics from the samples.

	Parameter (Always about population)	Statistics (Always about sample)
Definition	A number that describes the population characteristic.	A number that describes the sample characteristic.
Example	The average of PA population is 40 as of 2020. Parameter value: $\mu = 40$	A random sample of 5000 Pennsylvanians was collected and found the average age to be 39.5 Sample value: $\bar{x} = 39.5$

3 Measures of Center

Three measures are used as the center of a given numeric data set.



3.1 Mean

The mean of a given data set is defined as the average of all data values. The big sigma notations of sample and population means are given by

The **Mean** is the average of data values

Sample mean ← n = Sample Size

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

Population mean ← N = Population Size

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

Remark: the mean can be affected significantly by outliers (extreme values). For example,

3.2 Median

The middle value of a sorted data set is called the median of the data set.

- If a data set has an **odd number** of data values, there is a unique “middle” value in the **sorted data** set.
- If a data set has an **even number** of data values, there will be two **middle** values in the **sorted data** set, in this case, the **average** of the two “middle** values is defined to be the median.

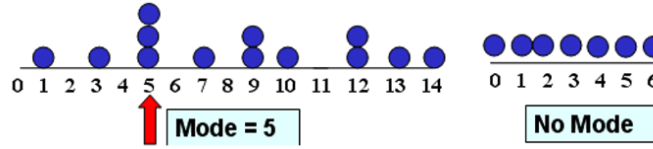
For example,

- $\{2, 6, 7\} \rightarrow \text{median} = 6$
- $\{1, 2, 6, 7\} \rightarrow \text{median} = (2 + 6) / 2 = 4.$

3.3 Mode

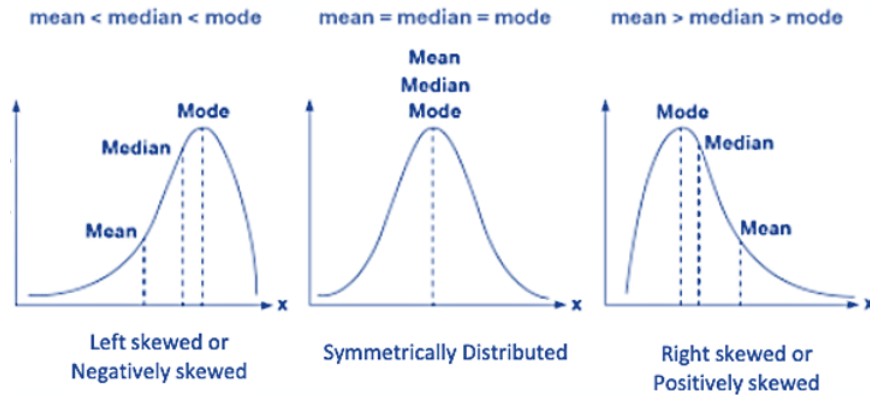
The mode(s) is (are) the data value(s) with highest frequency.

- If there is only one mode, the data set is unimodal.
- If there are two modes, the data set is bimodal.
- If there are more than two modes, the data is multi-modal.



3.4 Relationship between Mean, Median, and Mode

The relationship between mean, median, and mode is dependent on the shape of the distribution. The following figure illustrates this relationship.



4 Measures of Variation

Measures of variation are used to characterize the shape of the distribution. There are some different measures used in different situations. We only introduce the variance and the standard deviation in this course. We will also briefly introduce IQR in the applications of numerical measures.

4.1 Variance

Since the definitions of sample and population variances are different, we need to choose an appropriate formula based on whether the data set is a population or a sample. This information is provided to you before you select a formula to calculate the variances. The exact definitions using big sigma notation are given below

- Population Variance

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N} = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_N - \mu)^2}{N}$$

- Sample Variance

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}$$

We can see the only difference is in the denominator of the two definitions.

4.2 Standard Deviation

Once the variance is calculated, we simply take the square root to obtain the standard deviation

- Population standard deviation

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}} = \sqrt{\frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_N - \mu)^2}{N}}$$

- Sample Standard Deviation

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}} = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}}$$

4.3 Steps for Calculating Variance

The following are steps for calculating the variance of a data set.

Step 1: Compute the sample mean \bar{x} .

Step 2: For each sample value x , compute the difference $x - \bar{x}$. This quantity is called a deviation.

Step 3: Square the deviations, to obtain quantities $(x - \bar{x})^2$.

Step 4: Sum the squared deviations, obtaining $\sum (x - \bar{x})^2$.

Step 5: Divide the sum obtained in Step 4 by $n - 1$ to obtain the sample variance s^2 .

Example 2 The following table illustrates how to use the above steps to calculate the variance of a small sample toy data set: $A = \{1, 4, 7\}$.

X (dataset)	Deviation: $X - \bar{X}$	Squared Deviation: $(X - \bar{X})^2$
1	$1 - 4 = -3$	$(-3)^2 = 9$
4	$4 - 4 = 0$	$0^2 = 0$
7	$7 - 4 = 3$	$3^2 = 9$
$\bar{X} = 12/3$		$s^2 = \frac{18}{3 - 1} = 9$

Based on the above table, we can see that the standard deviation is $\sqrt{9} = 3$.

5 Measures of Location

Two important types of measures of location will be introduced in this course: z-score and percentiles.

5.1 z-score

A Z-score of a value of a **sample** data set is a standardized score that is defined by

$$z = \frac{x - \bar{x}}{s}.$$

We can easily adjust the above formula for a population as

$$z = \frac{x - \mu}{\sigma}.$$

Example 3: We still use the same sample toy data, $A = \{1, 4, 7\}$, used in **Example 2** to illustrate how to find z-scores of corresponding data values.

Solution: We know from **Example 2** that $\bar{x} = 4$ and $s = 3$. Therefore, the z-scores of the corresponding data values are calculated in the following.

$$x_1 = 1 \rightarrow z_1 = \frac{1-4}{3} = -1$$

$$x_2 = 4 \rightarrow z_2 = \frac{4-4}{3} = 0$$

$$x_3 = 7 \rightarrow z_3 = \frac{7-4}{3} = 1$$

That is, the standardized set of z-scores is $\{-1, 0, 1\}$. Note this is a set of sample z-scores. We can easily verify that the mean and standard deviation of the above three z-scores are 0 and 1 respectively.

5.2 Percentile

A **percentile** indicates the percentage of scores that fall below a particular value.

Example 4 Consider the following PSAT percentile table.

PSAT PERCENTILES		
PSAT SECTION SCORE	EVIDENCE-BASED READING & WRITING PERCENTILE	MATH PERCENTILE
760	99+	99+
750	99+	99
740	99	98
730	99	97
720	98	97
710	97	96
700	96	95
690	95	94
680	94	93
670	93	92
660	91	91
650	90	90
640	88	89
630	86	87
620	84	85

If you took PSAT and scored 640 in the MATH section, according to the above table, your MATH percentile is 89% meaning that 89% of all examinees scored below 640 in the PSAT. This also means that you did better than 89% of your peers on the PSAT.

Steps for Calculating Percentiles

Assume that we have a data set $\{x_1, x_2, \dots, x_n\}$. we want to find k -th percentile, denoted by P_k .

Step 1: Sort the data in ascending order: $\{x_{(1)}, x_{(2)}, \dots, x_{(n)}\}$.

Step 2: Calculate the rough location of the k -th percentile

$$L = \frac{k}{100} \times n.$$

Step 3: The k^{th} percentile is obtained depending on the form of L .

- if L is a whole number, then the k^{th} percentile is the average of the number in position L and the number in position $L + 1$ in the sorted data set.
- if L is NOT a whole number, round it **up** to the next higher whole number. The k^{th} percentile is the number in the position corresponding to the **rounded-up** value.

Example 5: Consider the following data set

9, 13, 7, 7, 12, 15, 10, 10, 6, 19, 17, 10, 15, 9, 14, 12, 9, 13, 7,
7, 4, 8, 19, 5, 18, 20, 14, 1, 23, 10, 10, 7, 22, 9, 1

Find P_{40} and P_{55} percentiles respectively.

Solution we first sort the data in ascending order.

1, 1, 4, 5, 6, 7, 7, 7, 7, 7, 8, 9, 9, 9, 9, 9, 10, 10, 10, 10, 10, 12, 12,
13, 13, 14, 14, 15, 15, 17, 18, 19, 19, 20, 22, 23

For convenience, we have organized the sorted data values into the table below, including the corresponding physical location ID in parentheses.

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)
1	1	4	5	6	7	7	7	7	7	8	9	9	9	9	10	10
(19)	(20)	(21)	(22)	(23)	(24)	(25)	(26)	(27)	(28)	(29)	(30)	(31)	(32)	(33)	(34)	(35)
10	10	12	12	13	13	14	14	15	15	17	18	19	19	20	22	23

To find 40th percentile,

$$L = \frac{40}{100} \times 35 = 14.$$

Since $L = 14$ is an integer, the 40th percentile is the average of 14th and 15th data values in the sorted data set. That is, $(9 + 9)/2 = 9$.

To find 55th percentile,

$$L = \frac{55}{100} \times 35 = 19.25.$$

Since $L = 19.25$ is NOT an integer, we round up L to get **20**. The 55th percentile is the 20th data value in the sorted data set which is 10.

5.3 Applications of Numeric Measures

Three concepts based on the numeric measures will be introduced in the following.

5.3.1 Five Number Summary

The five number summary consists of the minimum, 25th, 50th, 75th percentiles, and the maximum. The 25th, 50th, and 75th percentiles are also called the first (Q_1), second (Q_2) and third quartiles (Q_3), respectively.

Example 6: We use the **length of CD** data to show the five-number-summary. The unit of data values is minute. The following is the sorted data.

20.5, 29, 32, 32, 32, 33, 36, 37, 38, 39, 39, 43, 47, 48, 49, 49, 49,
50, 50, 51, 51, 52, 52, 52, 53, 54, 54, 54, 56, 56, 57, 58, 60, 61, 62,
62, 69, 73, 74, 74.5

Solution, The minimum and maximum are 20.5 and 74.5 minutes. The quartiles are calculated in the following.

$Q_1 : L = (25/100) \times 40 = 10$, Q_1 is the average of the 10th and 11th data values $= (39 + 39)/2 = 39$.

$Q_2 : L = (50/100) \times 40 = 20$, Q_2 is the average of the 20th and 21st data values $= (51 + 51)/2 = 51$.

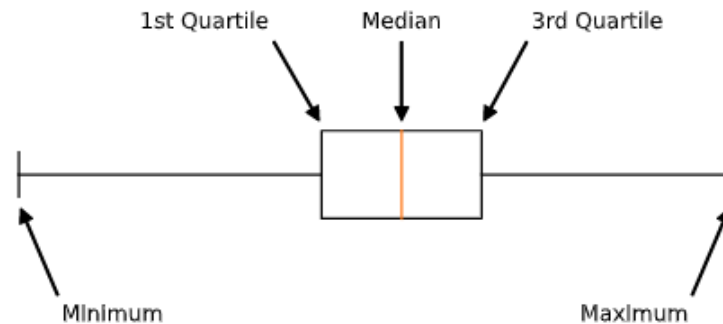
$Q_3 : L = (75/100) \times 40 = 30$, Q_3 is the average of the 30th and 31st data values $= (56 + 57)/2 = 56.5$.

Therefore, the five-number-summary is given by

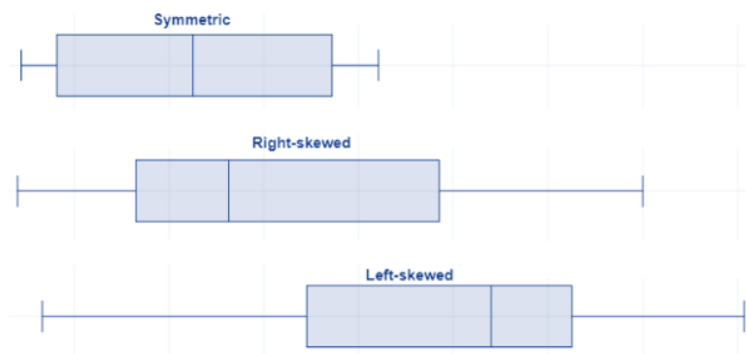
Min	Q1	Q2	Q3	Max
20.5	39	51	56.5	74.5

5.3.2 Box-plot

The box-plot is a geometric representation of the five-number-summary given in the following figure



Box-plots are used to describe the distribution of data. The following three box-plots represent three different types of distributions:

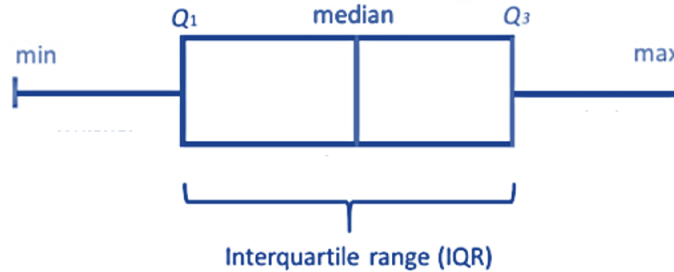


Example 7: (Length of CS Continued) The box-plot is given by



5.3.3 Inter-quartile Range (IQR)

The inter-quartile range of data is defined by $IQR = Q_3 - Q_1$. IQR is used to measure the variation of the data set. It is NOT sensitive to extremely large and small values since IQR is defined only based on the “middle 50% of data values”.



6 Use of Technology

We still use the same app that we used in the previous note to find various numerical measures to summarize a given data set. The app can be found at <https://wcupeng.shinyapps.io/DescriptiveStats/>. The next screenshot illustrates the measures of variations in the length of CD data. You can choose to find other numerical measures of the data.

ISLA: DESCRIPTIVE STATISTICS

Types of Descriptive Statistics


Numerical Measures

comma separated numerical data

20.5, 29, 32, 32, 32, 33, 36, 37, 38, 39, 39, 43, 47, 48, 49, 49, 49, 50, 50, 51, 51, 52, 52, 52, 53, 54, 54, 54, 56, 56, 57, 58, 60, 61, 62, 62, 69, 73, 74, 74.5

Measure Types

measures of center



Report bugs to C. Peng

Measures of Center

The data values are:
20.5, 29, 32, 32, 32, 33, 36, 37, 38, 39, 39, 43, 47, 48, 49, 49, 49, 50, 50, 51, 51, 52, 52, 52, 53, 54, 54, 54, 56, 56, 57, 58, 60, 61, 62, 62, 69, 73, 74, 74.5

The sorted data values are:
20.5, 29, 32, 32, 32, 33, 36, 37, 38, 39, 39, 43, 47, 48, 49, 49, 49, 50, 50, 51, 51, 52, 52, 52, 53, 54, 54, 54, 56, 56, 57, 58, 60, 61, 62, 62, 69, 73, 74, 74.5

1. Sample (population) mean

$\bar{x} = \sum_{i=1}^n \frac{x_i}{n} = 49.7$, and $\mu = \sum_{i=1}^n \frac{x_i}{n} = 49.7$ (if this data set is a population).

2. Median

The median of a given data set is the middle number of **sorted** data set. Based on this definition, the median of the given data set is: 51

3. Mode

A data value that appears most frequently (frequency > 1) is called the mode of the data. Based on this definition, a set of data may have one mode, more than one mode, or no mode at all. Using the above definition, this data set has 4 modes: 32, 49, 52, 54.

6.1 Example: Descriptive Statistical Analysis of Course Grade

The following are grades of a weekly assignment:

4.98, 5.98, 6.48, 7.00, 7.49, 7.49, 7.98, 7.98, 7.99, 7.99, 8.00,
8.00, 8.00, 8.02, 8.02, 8.49, 8.49, 8.50, 8.50, 8.50, 8.50, 8.50,
8.50, 8.50, 8.51, 8.51, 8.51, 8.98, 8.99, 8.99, 8.99, 8.99, 9.00,
9.00, 9.00, 9.00, 9.00, 9.01, 9.01, 9.01, 9.01, 9.01, 9.01, 9.01,
9.02, 9.02, 9.02, 9.48, 9.51, 9.98, 9.98, 9.99, 9.99, 10.00, 10.00,
10.00, 10.00

The questions we may be interested in are:

1. What is the grade distribution (i.e., frequency or histograms).

ISLA: DESCRIPTIVE STATISTICS

Types of Descriptive Statistics

Table and Chart: Numerical Data

comma separated numeric raw data


4.98, 5.98, 6.48, 7.00, 7.49, 7.49, 7.98, 7.98, 7.9

Summary Types

Frequency Tables

Boundary [must be equally spaced!]

4.0, 5.0, 6.0, 7.0, 8.0, 9.0, 10.0



Report bugs to C. Peng

The input data values:

4.98, 5.98, 6.48, 7, 7.49, 7.49, 7.98, 7.98, 7.99, 7.99, 8, 8, 8, 8.02, 8.02, 8.49, 8.49, 8.5, 8.5, 8.5, 8.5, 8.5, 8.5, 8.51, 8.51, 8.51, 8.98, 8.99, 8.99, 8.99, 8.99, 9, 9, 9, 9, 9.01, 9.01, 9.01, 9.01, 9.01, 9.01, 9.01, 9.02, 9.02, 9.02, 9.02, 9.48, 9.51, 9.98, 9.98, 9.99, 9.99, 10, 10, 10, 10

The sorted input data values:

4.98, 5.98, 6.48, 7, 7.49, 7.49, 7.98, 7.98, 7.99, 7.99, 8, 8, 8, 8.02, 8.02, 8.49, 8.49, 8.5, 8.5, 8.5, 8.5, 8.5, 8.5, 8.5, 8.51, 8.51, 8.51, 8.98, 8.99, 8.99, 8.99, 8.99, 9, 9, 9, 9, 9.01, 9.01, 9.01, 9.01, 9.01, 9.01, 9.01, 9.02, 9.02, 9.02, 9.02, 9.48, 9.51, 9.98, 9.98, 9.99, 9.99, 10, 10, 10, 10

The class boundary is: 4.0, 5.0, 6.0, 7.0, 8.0, 9.0, 10.0

cut.data.freq	Freq	midpts	rel.freq	cum.freq	rel.cum.freq
[4,5]	1	4.50	0.02	1	0.02
(5,6]	1	5.50	0.02	2	0.04
(6,7]	2	6.50	0.04	4	0.07
(7,8]	9	7.50	0.16	13	0.23
(8,9]	24	8.50	0.42	37	0.65
(9,10]	20	9.50	0.35	57	1.00

2. we also may want to know the class average, median, variance and standard deviation, etc.

ISLA: DESCRIPTIVE STATISTICS

Types of Descriptive Statistics


Numerical Measures

comma separated numerical data

4.98, 5.98, 6.48, 7.00, 7.49, 7.49, 7.98, 7.98, 7.9

Measure Types

measures of center



Report bugs to C. Peng

Measures of Center

The data values are:

4.98, 5.98, 6.48, 7, 7.49, 7.49, 7.98, 7.98, 7.99, 7.99, 8, 8, 8, 8.02, 8.02, 8.49, 8.49, 8.5, 8.5, 8.5, 8.5, 8.5, 8.5, 8.51, 8.51, 8.51, 8.98, 8.99, 8.99, 8.99, 8.99, 9, 9, 9, 9, 9.01, 9.01, 9.01, 9.01, 9.01, 9.01, 9.01, 9.02, 9.02, 9.02, 9.02, 9.48, 9.51, 9.98, 9.98, 9.99, 9.99, 10, 10, 10, 10

The sorted data values are:

4.98, 5.98, 6.48, 7, 7.49, 7.49, 7.98, 7.98, 7.99, 7.99, 8, 8, 8, 8.02, 8.02, 8.49, 8.49, 8.5, 8.5, 8.5, 8.5, 8.5, 8.5, 8.5, 8.51, 8.51, 8.51, 8.98, 8.99, 8.99, 8.99, 8.99, 9, 9, 9, 9, 9.01, 9.01, 9.01, 9.01, 9.01, 9.01, 9.01, 9.02, 9.02, 9.02, 9.02, 9.48, 9.51, 9.98, 9.98, 9.99, 9.99, 10, 10, 10, 10

1. Sample (population) mean

$\bar{x} = \sum_{i=1}^n \frac{x_i}{n} = 8.639$, and $\mu = \sum_{i=1}^n \frac{x_i}{n} = 8.639$ (if this data set is a population).

2. Median

The median of a given data set is the middle number of **sorted** data set. Based on this definition, the median of the given data set is: 8.99

3. Mode

A data value that appears most frequently (frequency > 1) is called the mode of the data. Based on this definition, a set of data may have one mode, more than one mode, or no mode at all. Using the above definition, this data set has 2 modes: 8.5, 9.01.

Types of Descriptive Statistics


Numerical Measures

comma separated numerical data

4.98, 5.98, 6.48, 7.00, 7.49, 7.49, 7.98, 7.98, 7.9

Measure Types

measures of variation



Report bugs to C. Peng

Measures of Variation

The data values are:
4.98, 5.98, 6.48, 7, 7.49, 7.49, 7.98, 7.98, 7.99, 7.99, 8, 8, 8, 8.02, 8.02, 8.49, 8.49, 8.5, 8.5, 8.5, 8.5, 8.5, 8.5, 8.51, 8.51, 8.51, 8.98, 8.99, 8.99, 8.99, 9, 9, 9, 9, 9, 9.01, 9.01, 9.01, 9.01, 9.01, 9.01, 9.02, 9.02, 9.02, 9.02, 9.48, 9.51, 9.98, 9.98, 9.99, 9.99, 10, 10, 10, 10

The sorted data values are:
4.98, 5.98, 6.48, 7, 7.49, 7.49, 7.98, 7.98, 7.99, 7.99, 8, 8, 8, 8.02, 8.02, 8.49, 8.49, 8.5, 8.5, 8.5, 8.5, 8.5, 8.5, 8.51, 8.51, 8.51, 8.98, 8.99, 8.99, 8.99, 9, 9, 9, 9, 9, 9.01, 9.01, 9.01, 9.01, 9.01, 9.01, 9.02, 9.02, 9.02, 9.02, 9.48, 9.51, 9.98, 9.98, 9.99, 9.99, 10, 10, 10, 10

1. Sample (population) variance

$$s^2 = \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{n-1} = 0.949, \text{ and } \sigma^2 = \sum_{i=1}^n \frac{(x_i - \mu)^2}{n} = 0.933 \text{ (if this data set is a population)}$$

2. Sample (population) standard deviation

The standard deviation is the square root of variance. Therefore, the both standard deviations are: 8.99

$$s = \sqrt{s^2} = \sqrt{\sum_{i=1}^n \frac{(x_i - \bar{x})^2}{n-1}} = 0.974, \text{ and } \sigma = \sqrt{\sigma^2} = \sqrt{\sum_{i=1}^n \frac{(x_i - \mu)^2}{n}} = 0.966 \text{ (if this data set is a population)}$$

3. Inter – quartile range (IQR)

The inter-quartile range is defined to be the difference between the first and third quartiles. By the definition,
 $IQR = P_{75} - P_{25} = 9.01 - 8.02 = 0.99$.

3. We may also want to see the 5-number summary and box-plot

ISLA: DESCRIPTIVE STATISTICS

Types of Descriptive Statistics


Numerical Measures

comma separated numerical data

4.98, 5.98, 6.48, 7.00, 7.49, 7.49, 7.98, 7.98

Measure Types

5-number summary and boxplot



Report bugs to C. Peng

Five Number Summary and Boxplot

The data values are:
4.98, 5.98, 6.48, 7, 7.49, 7.49, 7.98, 7.98, 7.99, 7.99, 8, 8, 8, 8.02, 8.02, 8.49, 8.49, 8.5, 8.5, 8.5, 8.5, 8.5, 8.5, 8.51, 8.51, 8.51, 8.98, 8.99, 8.99, 8.99, 8.99, 9, 9, 9, 9, 9, 9.01, 9.01, 9.01, 9.01, 9.01, 9.01, 9.02, 9.02, 9.02, 9.02, 9.48, 9.51, 9.98, 9.98, 9.99, 9.99, 10, 10, 10, 10

1. Five Number Summary :

The five-number summary is use used to describe the shape of the distribution of a given numerical data. It consists of five numbers: minimum data value, first quartile, median, the third quartile, and the maximum data value.

The five-number summary of this given data set is:

stats	value
Min.	4.98
1st Qu.	8.02
Median	8.99
3rd Qu.	9.01
Max.	10.00

4. We can also standardize the grades by z-score transformation and find inter-quartile range (IQR).
See the screenshot in problem 2.

6.2 Exercises

The following two exercises are taken from the previous note. Please answer the following two questions manually before using ISLA app to check your answer.

1. Construct a histogram with 5 classes.
2. Find the five numer summary and the corresponding box-plot.

Problem 1. Following are 80 measurements of the iron-solution index of tin-plate specimens, designed to measure the corrosion resistance of tin-plated steel. The original data set has been sorted in an ascending order as:

14, 26, 28, 28, 28, 28, 30, 32, 34, 35, 36, 36, 37, 37, 40, 40,
 40, 41, 41, 41, 42, 42, 42, 43, 43, 43, 44, 44, 44, 44, 45, 45,
 45, 45, 45, 45, 46, 46, 46, 46, 47, 47, 47, 48, 49, 49, 49, 50,
 50, 50, 51, 52, 52, 52, 52, 52, 53, 53, 54, 54, 54, 54, 55, 55,
 55, 55, 55, 55, 55, 56, 56, 56, 56, 56, 56, 57, 57, 57, 57, 57,
 58, 58, 58, 58, 58, 59, 59, 60, 60, 60, 60, 61, 61, 61, 61, 61,
 62, 62, 62, 62, 62, 62, 63, 63, 63, 63, 63, 63, 64, 65, 66, 66,
 67, 68, 68, 69, 69, 70, 70, 70, 70, 70, 70, 71, 71, 72, 72, 72,
 73, 74, 74, 74, 76, 76, 77, 77, 79, 80, 81, 81, 83, 83, 84, 86,
 86, 86, 87, 89, 92, 95

Problem 2. From the 140 children whose urinary concentration of lead were investigated 40 were chosen who were aged at least 1 year but under 5 years. The following concentrations of copper were found.

0.70, 0.45, 0.72, 0.30, 1.16, 0.69, 0.83, 0.74, 1.24, 0.77, 0.65, 0.76, 0.42,
 0.94, 0.36, 0.98, 0.64, 0.90, 0.63, 0.55, 0.78, 0.10, 0.52, 0.42, 0.58, 0.62,
 1.12, 0.86, 0.74, 1.04, 0.65, 0.66, 0.81, 0.48, 0.85, 0.75, 0.73, 0.50, 0.34, 0.88