Introduction to Statistics and Data Science using *eStat*

Chapter 4 Data Summary Using Tables and Measures

4.3 Summary Measure for Quantitative Variable - Measure of Dispersion -

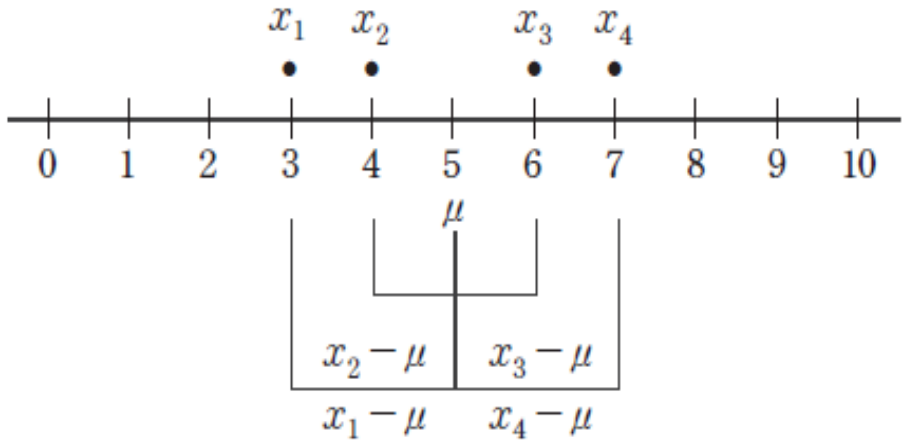
Jung Jin Lee
Professor of Soongsil University, Korea
Visiting Professor of ADA University, Azerbaijan

4.3 Summary Measures for Quantitative Variable

4.3.2 Measure of Dispersion

- Measuring the degree of data dispersion in numerical values
  => variance or standard deviation
  range, and inter-quartile range

- **Variance** is the average of the squared distances from data to the mean,
  - If data are spread widely around mean, variance will increase
  - If data is concentrated around the mean, variance will be small



$$\sigma^2 = \frac{(-2)^2+(-1)^2+1^2+2^2}{4} = 2.5$$

4.3 Summary Measures for Quantitative Variable

Population variance $\sigma^2 = \dfrac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}$ (N: number of population data)

Sample variance $S^2 = \dfrac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$ (n : number of sample data)

✓ **There are important reasons for using n-1 instead n when calculating the sample variance (Refer Chapter 6)**

4.3 Summary Measures for Quantitative Variable

- **Standard deviation** is the square root of the variance.
  - The standard deviation of the population is denoted as σ.
  - The standard deviation of the sample is denoted as s.

Population standard deviation	$\sigma = \sqrt{\sigma^2}$
Sample standard deviation	$s = \sqrt{s^2}$

- Variance is not easy to interpret because it is the mean of the squared distance.
- Standard deviation is the square root of the variance, which allows it to be interpreted as a measure of the average distance from each value to the mean.

4.3 Summary Measures for Quantitative Variable

[Example 4.3.5] Calculate mean and standard deviation from sample data
5, 6, 3, 7, 9, 4, 8.

<Answer>

- Note that this data is sample.

- $\overline{x} = \frac{5+6+3+7+9+4+8}{7} = 6$
- $s^2 = \frac{(5-6)^2+(6-6)^2+(3-6)^2+(7-6)^2+(9-6)^2+(4-6)^2+(8-6)^2}{7-1} = \frac{28}{6} = 4.6$
- $s = \sqrt{s^2} = \sqrt{4.667} = 2.16$

- 『eStat』 calculates both population and sample standard deviation

4.3 Summary Measures for Quantitative Variable

- **Coefficient of variation** is division of the standard deviation by its mean
  =>to compare several sets of data in different units

Population Coefficient of Variation	$C = \frac{\sigma}{\mu} \times 100$	(unit %)
Sample Coefficient of Variation표본)	$C = \frac{s}{\bar{x}} \times 100$	(unit %)

[Ex 4.3.6] The average weekly sales of a company was 1.36 billion dollar and the standard deviation was 0.28 billion dollar. When the same data was made in monthly sales, the average was 5.4.4 billion dollar and the standard deviation was 0.5 billion dollar. Calculate and compare the coefficient of variation .

<Answer>

- **The coefficient of variation in weekly sales is (0.28 / 1.36) × 100 = 20.6%.**
- **The coefficient of variable in monthly sales is (0.50 / 5.44) × 100 = 9.2%.**
- **The change in monthly sales is smaller than the change in weekly sales.**

4.3 Summary Measures for Quantitative Variable

- **Range = maximum - minimum**
  - easy to calculate,
  - not a good measure if there are extreme points.

- **p percentile : there are p% of observations less than(≤) this value**
  (100-p)% of observations above(≥) this value

- 25 percentile of the data is called the 1st quartile (Q1),
  50 percentile is called the 2nd quartile(Q2) or the median,
  75 percentile is called the 3rd quartile (Q3).

  Inter-quartile range (IQR) = Q3 - Q1

4.3 Summary Measures for Quantitative Variable

[Ex 4.3.7] For data 5, 6, 3, 7, 9, find the range and inter-quartile range.

<Answer>

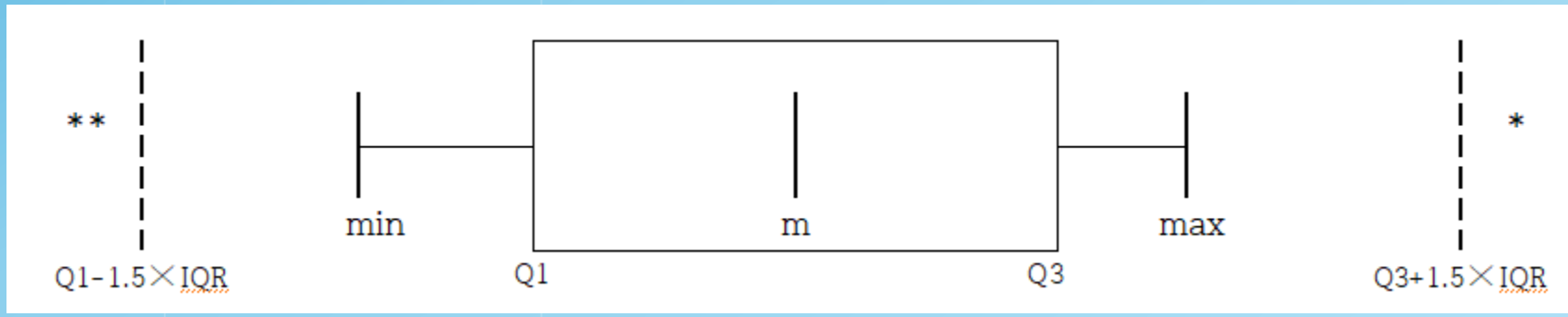
- Range = max(9) – min(3) = 6.
- Arrange data in ascending order.
  (3, 5, 6, 7, 9)
- Median is (5+1)/2 th data which is 6.
- Divide sorted data into two parts
  (3,5,6) (6,7,9) <= note that median 6 included in both parts
- Median of the (3,5,6) which is 5 is the 25 percentile (Q1).
- Median of the (6,7,9) which is 7 is the 75 percentile (Q3).
- IQR is Q3 - Q1 = 7 - 5 = 2.

4.3 Summary Measures for Quantitative Variable

- **Box-whiskers plot** is a method to show the quartiles of data.

  => mark Q1 and Q3 at a horizontal line and connects with a square box.
  displays median (Q2) at the location proportional to Q1 and Q3 in box.
  draw a dotted line at (Q3 + 1.5*IQR) and (Q1- 1.5*IQR)
  data below line (Q1- 1.5*IQR) and over line (Q3 + 1.5*IQR) are extremes
  after excluding extremes, find minimum, maximum of remaining data
  connect minimum to the box with a line.
  connect maximum to the box with a line

- Box graph shows symmetry of data, central location, degree of dispersion.

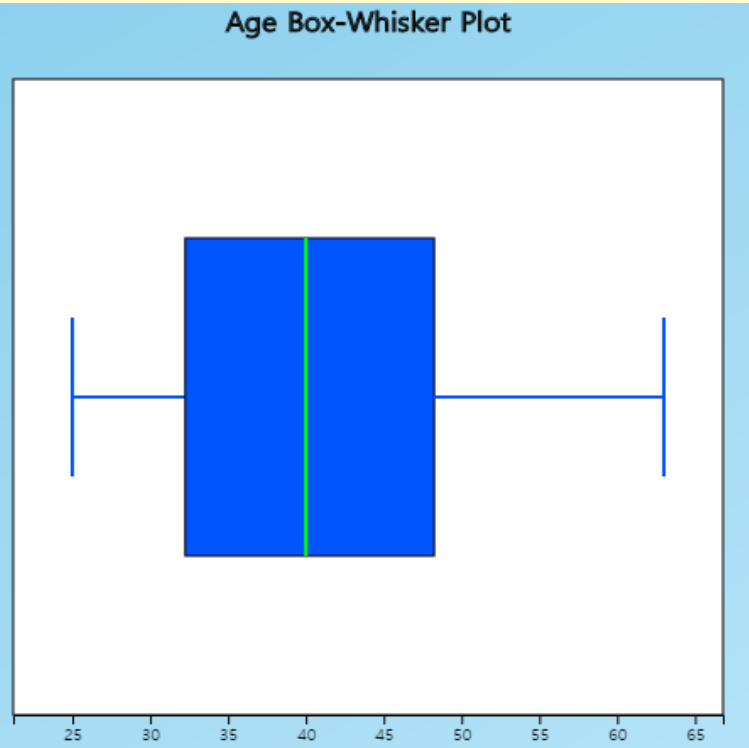


4.3 Summary Measures for Quantitative Variable

[Ex 4.3.8]] Using data 032Continous_TeacherAgeByGender.csv in『eStat』
1) Draw a box plot of age and examine median, range, quartiles and IQR.
2) Draw a box plot of age by gender and compare median, range, quartiles and IQR.

<Answer>

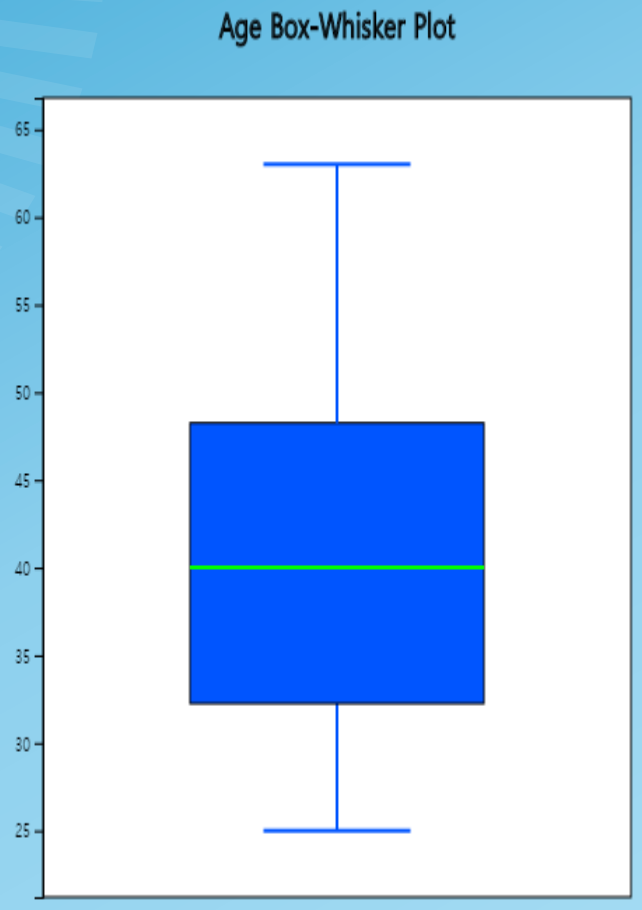
- **After loading the data in『eStat』, enter the value label of 'Gender' as 'Male' for 1 and 'Female' for 2 at [EditVar] button.**
- **Clicking on the box graph icon and then the 'age' variable**
- **Based on the median, we can see that the upper value is more scattered.**



4.3 Summary Measures for Quantitative Variable

<Answer of Ex 4.3.8>

- **Click the [Descriptive Statistics] button in the graph options to display the basic statistics of the ages**
- **Select 'Vertical' from the options below the graph for a vertical box graph**

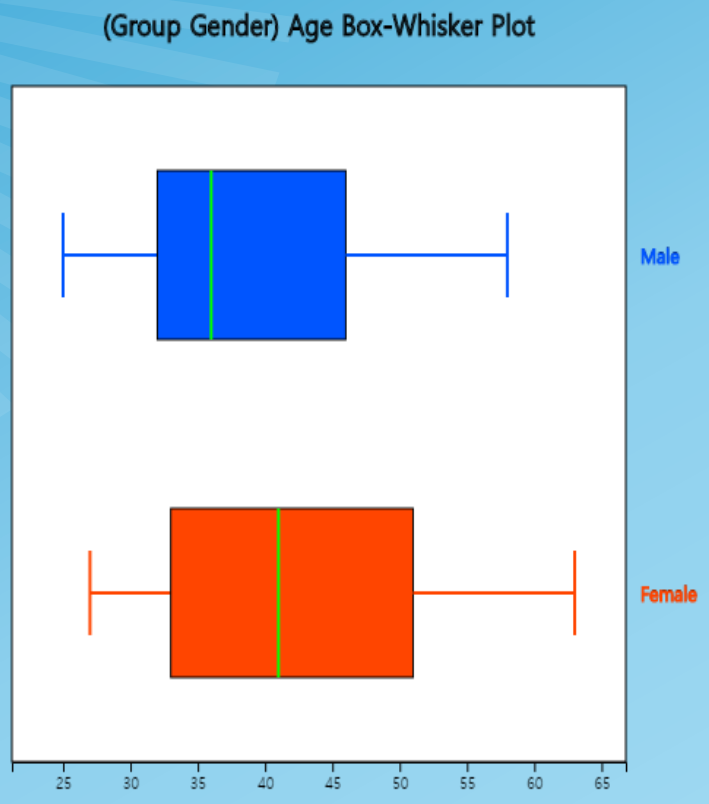


Descriptive Statistics	Analysis Var (Age)
Observation	30
Missing Observations	0
Mean	40.667
Variance (n)	116.822
Variance (n-1)	120.851
Std Dev (n)	10.808
Std Dev (n-1)	10.993
Minimum	25.000
1st Quartile	32.250
Median	40.000
3rd Quartile	48.250
Maximum	63.000
Range	38.000
Interquartile Range	16.000
Coefficient of Variation (n)	26.58 %
Coefficient of Variation (n-1)	27.03 %

4.3 Summary Measures for Quantitative Variable

<Answer of Ex 4.3.8>

- Click on a 'gender' variable with the 'age' variable selected, a horizontal box plot by gender appears.
- The dispersion of female teachers' ages is greater than that of male teachers.



Descriptive Statistics	Analysis Var (Age)	Group Name (Gender) 1 (Group 1)	Group Name (Gender) 2 (Group 2)
Observation	30	13	17
Missing Observations	0		
Mean	40.667	38.846	42.059
Variance (n)	116.822	106.592	120.173
Variance (n-1)	120.851	115.474	127.684
Std Dev (n)	10.808	10.324	10.962
Std Dev (n-1)	10.993	10.746	11.300
Minimum	25.000	25.000	27.000
1st Quartile	32.250	32.000	33.000
Median	40.000	36.000	41.000
3rd Quartile	48.250	46.000	51.000
Maximum	63.000	58.000	63.000
Range	38.000	33.000	36.000
Interquartile Range	16.000	14.000	18.000
Coefficient of Variation (n)	26.58 %	26.58 %	26.06 %
Coefficient of Variation (n-1)	27.03 %	27.66 %	26.87 %

Thank you