

Chapter 3 Descriptive Statistics: Numerical Measures

3-1 In this chapter, we present numerical measures for summarizing data, we start by developing numerical summary measures of data sets consisting only a single variable, i.e., measures apply to a data set are actually applying to a single variable.

Sample statistics, population parameters & point estimator

If the measures are calculated for data from a sample, they are called **sample statistics**.

If the measures are calculated for data from a population, they are called **population parameters**.

A sample statistic is referred to as a **point estimator** for the corresponding population parameter.

Any set of data can be characterized by its location and its variability.

Measure of Location:

Mean, median, mode, percentile, quartile

Mean: the average of all data values

- Sample mean is denoted as \bar{x}
- population mean is denoted as μ
- The sample mean is a point estimator for the population mean

Suppose we have n observations x_i in our sample, $i=1,2,\dots,n \implies$

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

- The denominator is the number of observations in the sample
- The numerator is the sum of all the n observations

Suppose we have N observations x_i in our population, $i=1,2,\dots,N \implies$

$$\mu = \frac{\sum_{i=1}^N x_i}{N}$$

Ex. Suppose we want to learn about the average rent for a one bed apartment in a small town. We randomly sample 5 apartments. Their rents are

Apt. ID	Rent(\$)
1	525
2	400
3	475
4	500
5	600

Let x be the rent, then $\bar{x} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5} = \500

If the small town has only 5 one bed apartment, then it's not a sample, instead it is the whole population, and what we get is μ , the population parameter.

- Mean is sensitive to outliers/ extreme values: some extreme values can inflate it a lot.
- Is sample mean always smaller than population mean?

Median: put the data in the data set in an ascending order, the median of a list of data is the one in the middle.

- For an odd number of observations, the median is the value in the middle
Ex. we have a list of data: 1, 2, 3, 4, 5, 6, 9
Median=4

- For an even number of observations, the median is the average of the middle two values.
Ex. a list of data: 1, 2, 3, 4, 5, 6
Median=(3+4)/2=3.5

- Median is not affected by extreme values, so whenever a data set has extreme values, the median is the preferred as a measure of central location.

Ex. we have a list of data: 1, 2, 3, 4, 5, 6, 99999

Ex. In summarizing the annual income, a few rich people like Bill Gates will inflate the mean a lot. To appropriately reflect the national income level, median is usually used.

Mode: the mode of a data set is the value that occurs with the greatest frequency, or the most frequently occurring value of a data set

- Two or more values can occur the most → there may be more than one modes in the data set.

Ex. we have a data list: 1, 2, 2, 4, 5, 6, 1, 1, 2, 5, 6

Both 1 and 2 appears 3 times, so there are 2 modes, 1 and 2

Ex. we have a data list: 1, 2, 9, 1, 0, 9, 2

1, 2 and 9 appear twice, so there are 3 modes, 1, 2 and 9

Both 1 and 2 appears 3 times, so there are 2 modes, 1 and 2

- If the data set has exactly two modes, it's called bimodal
- If the data set has more than two modes, it's called multimodal

Percentile: a percentile provides information about how the data spread over the interval from the smallest value to the largest value.

The ***p*th percentile** of a data set is a value such that **at least** *p* percent of the data less than or equal to this value and **at least** (100-*p*) percent of the data greater than or equal to this value.

Calculate percentiles

1. Arrange the data in an ascending order
2. Compute index *i*, the position of *p*th percentile

$$i = \frac{p}{100} n$$

3. if i is not an integer, round up, the p th percentile is the value in the i th position
4. if i is an integer, then the p th percentile is the average of the values in the i th and $i+1$ th positions.

Why at least? This is because there may be repeated values in the data set.

Ex. we have a data list: 1, 2, 4, 4, 4, 5, 5, 6, 7, 8

What's the 30th percentile?

The 30th percentile is given by 4. By definition, we need a value such that at least 30% of the data are less than or equal to this value. And at least 70% of the data are greater than or equal to this value. Let's check! Because of the repeated values, there are 50% observations having a value ≤ 4 , which is $>30\%$. There are 80% observations ≥ 4 , which is $>(100-30)\%$

Ex. $n=12$, data list: 1, 2, 4, 4, 6, 7, 8, 9, 10, 12, 16, 18

25th percentile=? 30th percentile=? 50th percentile=? 60th percentile=?

Already in ascending order.

$$i = (25/100) \times 12 = 3$$

$$25^{\text{th}} \text{ percentile} = (4+4)/2 = 4$$

Check: at least 25% values ≤ 4 $4/12 = 33.3\% > 25\%$
 at least 75% values ≥ 4 $10/12 = 83.3\% > 75\%$

$$i = (30/100) \times 12 \approx 4$$

$$30^{\text{th}} \text{ percentile} = 4$$

Check: at least 30% values ≤ 4 $4/12 = 33.3\% > 30\%$
 at least 70% values ≥ 4 $10/12 = 83.3\% > 70\%$

$$i = (50/100) \times 12 = 6$$

$$50^{\text{th}} \text{ percentile} = (7+8)/2 = 7.5$$

Check: at least 50% values ≤ 7.5 $6/12 = 50\% \geq 50\%$
 at least 50% values ≥ 7.5 $6/12 = 50\% \geq 50\%$

$$i = (60/100) \times 12 \approx 8$$

$$60^{\text{th}} \text{ percentile} = 9$$

Check: at least 60% values ≤ 9 $8/12 = 66.7\% > 60\%$
 at least 40% values ≥ 9 $5/12 = 41.7\% > 40\%$

Quartiles: quartiles are specific percentiles

First quartile = 25th percentile

Second quartile = 50th percentile = median

Third quartile = 75th percentile

Ex. cont. third quartile = ?

$$i = (75/100) \times 12 = 9$$

$$\text{Third quartile} = (10+12)/2 = 11$$

Check: at least 75% values ≤ 11 $9/12=75\%$ $\geq 75\%$
at least 25% values ≥ 11 $3/12=25\%$ $> 25\%$

3-2 Measure of Variability:

Range, interquartile range, variance, standard deviation, coefficient of variation

It is often desirable to consider measures of variability (dispersion), as well as measures of location.

For example, in choosing supplier A or supplier B we might consider not only the average delivery time for each, but also the variability in delivery time for each supplier

Range: the range of a data set is the difference between the largest and the smallest data values.

Ex. -1, 9, 4, 0, 5, 11, 2, 10
→ -1, 0, 2, 4, 5, 9, 10, 11
Range = $11 - (-1) = 12$

It's very sensitive to the smallest and the largest values.

Ex: in the above example, range = 12, but if the largest value is 10000, then the range is 10001

Interquartile range (IQR) = the third quartile – the first quartile

In the previous example, interquartile range = $9.5 - 1 = 8.5$

Third quartile = 75% percentile; first quartile = 25% percentile →

It is the range for the middle 50% of the data. It overcomes the sensitivity to extreme data values.

Variance: the variance is a measure of variability using all the data.

The variance is the average of the squared difference between each data value and the mean.

Deviation about the mean: the difference between each data value (x_i) and the mean (\bar{x} for a sample, μ for a population)

The formulas for the variance are

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

$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \mu)^2}{N} \text{ for a population}$$

Given a data set: calculate the mean → deviation about the mean → squared deviation about the mean → average

The **standard deviation** of a data set is the positive square root of the variance.

$$s = \sqrt{s^2} \text{ for a sample ; } \sigma = \sqrt{\sigma^2} \text{ for a population}$$

Ex: (note the negative sign!) Five companies' profits

profit (\$1000)	Mean profit(\$1000)	Deviation about the mean	Squared deviation about the mean
10	-1	9	81
6	-1	7	49
-6	-1	-5	25
-12	-1	-11	121
-3	-1	-2	4
Total	-5	0	280

$$\rightarrow s^2 = 280 / (5 - 1) = 70$$

$$\rightarrow s = \sqrt{s^2} = 8.37 (\$1000)$$

- Can variance be negative?
- In what cases variance will be 0? 0 variance means no variability in the data set, all constant
- Can deviation about the mean be negative?
- Is sample variance always smaller than the population variance?

Coefficient of variation (CV): coefficient of variation measures how large the standard deviation is relative to the mean. It's defined as the ratio of the standard deviation to the mean. It is often expressed as a percent \rightarrow CV is given by

$$CV = \left(\frac{s}{\bar{x}} \times 100 \right) \% \text{ for a sample; } CV = \left(\frac{\sigma}{\mu} \times 100 \right) \% \text{ for a population}$$

- Note that the coefficient of variation is a dimensionless number. It's unit free. It allows comparison of the variation of populations/samples that have significantly different mean values.
- According to the above formula, CV, s, and \bar{x} , given any two of the three, we should know the third one.