Chapter 3

Measures of Central Tendency and Location

Outline of Chapter 3

1. Notations and Symbols
2. Measures of Central Tendency
   - Mean
   - Median
   - Mode
3. Measures of Location
   - Percentiles
   - Quartiles
   - Deciles

Reference: Chapter 6-7 of Elementary Statistics by ACS

Part 1

Notations and Symbols

Summation

The symbol

$$\sum_{i=1}^{n} X_i$$

is equal to $\mathbf{X_1+X_2+\ldots+X_n}$,
where $\mathbf{X_i}$ is the value of the variable for the i^{th} observation, **i** is the index of the summation, **1** is the lower limit of the summation, **n** is the upper limit of the summation.
We read $\sum_{i=1}^{n} X_i$ as "summation of X sub i, where i is from 1 to n".

Index set is the collection of consecutive integers from lower limit to the upper limit of the summation.

Summation

The terms of the sum are determined by successively replacing the index in the term of the summation by the elements in the index set.

Examples:

$$\sum_{i=1}^{5} X_i = X_1 + X_2 + X_3 + X_4 + X_5$$

$$\sum_{i=3}^{5} X_i^i = X_3^3 + X_4^4 + X_5^5$$

Example

The weights in pounds of the five students are as follows: 110, 90, 105, 120, and 115. Express the formula for the total weight of these students using the summation notation.

Solution: Let X = weight of a student in pounds
$n = 5$ students

X_1 represents the weight of the first student = 110 lbs.
X_2 represents the weight of the second student = 90 lbs.
X_3 represents the weight of the third student = 105 lbs.
X_4 represents the weight of the fourth student= 120 lbs.
X_5 represents the weight of the fifth student = 115 lbs.

Example

We express the formula for the total weight of the 5 students in a compact manner using the summation notation as $\sum_{i=1}^{5} X_i$. This notation means

$$\sum_{i=1}^{5} X_i = X_1 + X_2 + X_3 + X_4 + X_5 = 110 + 90 + 105 + 120 + 115$$
$$= 540\ lbs$$

Some Notes on the Summation

1. The index may be any letter, but the letters i, j, and k are the most common. For example,

$$\sum_{i=1}^{n} X_i = \sum_{j=1}^{n} X_j$$

2. The lower limit of the summation may start with any integer smaller than the upper limit. For example,

$$\sum_{i=2}^{5} X_i = X_2 + X_3 + X_4 + X_5$$

3. The index will not necessarily appear as a subscript in the term of the summation. For example,

$$\sum_{i=1}^{5} i = 1 + 2 + 3 + 4 + 5$$

Example

1. Find $\sum_{i=1}^{3} X_i$ when $X_1 = 2, X_2 = 5, X_3 = -2$
2. Find $\sum_{j=1}^{3} X_j^2$ when $X_1 = 2, X_2 = 5, X_3 = -2$
3. Evaluate $\sum_{x=1}^{4} X$.
4. Find $\sum_{i=1}^{3}(X_i-1)$ when $X_1 = 3, X_2 = 5, X_3 = 7$
5. Find $\left(\sum_{i=1}^{3} X_i\right)\left(\sum_{i=1}^{3} Y_i\right)$ when $X_1 = 2, X_2 = 1, X_3 = 4, Y_1 = 1, Y_2 = 3, Y_3 = -1$

Solution:

$$\sum_{i=1}^{3} X_i = X_1 + X_2 + X_3 = 2 + 5 + (-2) = 5$$

$$\sum_{j=1}^{3} X_j^2 = X_1^2 + X_2^2 + X_3^2 = 2^2 + 5^2 + (-2)^2 = 33$$

$$\sum_{x=1}^{4} X = 1 + 2 + 3 + 4 = 10$$

Assignment (For Practice; Not for submission)

i	1	2	3	4	5
X_i	2	3	4	5	6
Y_i	1	4	2	3	5

Given the data above, find the following:

$$\sum_{i=1}^{5} X_i^2 \,, \qquad \sum_{i=1}^{5} (X_i - Y_i)^2 \,, \qquad \left(\sum_{i=1}^{5} X_i \right)^2 , \qquad \sum_{i=3}^{5} (X_i - 1)$$

Σ Assignment (For Practice; Not for submission)

Write the expansion of the following summation:

a) $\sum_{\mathrm{i}=3}^{6}\frac{\mathrm{Z_i}}{\mathrm{Y_i}}$

b) $\frac{\sum_{\mathrm{i}=3}^{6}\mathrm{Z_i}}{\sum_{\mathrm{i}=3}^{6}\mathrm{Y_i}}$

c) $\sum_{\mathrm{k}=0}^{3}\left(\mathrm{Y_j^k}-\mathrm{Y_k}\right)$

d) $\sum_{\mathrm{i}=2}^{4}\sqrt{\mathrm{i}}$

e) $\sum_{\mathrm{j}=1}^{3}3X_j^j$

Additional Notes on the Summation

(1) $$\sum_{i=1}^{n} X_i^2 \neq \left(\sum_{i=1}^{n} X_i\right)^2$$

(2) $$\sum_{i=1}^{n} (X_i + Y_i)^2 \neq \sum_{i=1}^{n} X_i^2 + \sum_{i=1}^{n} Y_i^2$$

(3) $$\sum_{i=1}^{n} X_i Y_i \neq \left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)$$

(4) $$\sum_{i=1}^{n} \frac{X_i}{Y_i} \neq \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} Y_i}$$

(5) $$\sum_{i=1}^{n} \sqrt{X_i} \neq \sqrt{\sum_{i}^{n} X_i}$$

Some Properties of the Summation

1. The summation of the sum (or difference) of two or more terms equals the sum (or difference) of the individual summations. For example,

$$\sum_{i=1}^{n}(X_i + Y_i + Z_i) = \sum_{i=1}^{n} X_i + \sum_{i=1}^{n} Y_i + \sum_{i=1}^{n} Z_i$$

$$\sum_{i=1}^{n}(X_i - Y_i - Z_i) = \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} Z_i$$

2. The summation of the product of a constant, c, with X_i equals the product of the constant with the summation of X_i, i.e.,

$$\sum_{i=1}^{n} cX_i = c\sum_{i=1}^{n} X_i$$

3. The summation of a constant, c, with index set = {1,2,…,n}, equals the product of n and c, i.e.,

$$\sum_{i=1}^{n} c = nc$$

Σ Exercise

Define $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$. Show the following using the properties of summation:

$$n\bar{X} = \sum_{i=1}^{n} X_i$$

Solution:
In showing that the above equation holds,

(1) we can work with the LHS of the equation and arrive at the desired RHS; or
(2) we can work with the RHS of the equation and arrive at the desired LHS; or
(3) we can work with both sides of the equation and arrive at an identity for the LHS and the RHS.

Usually, we work on the side with more complex expression.

In this particular example, we will work on the LHS.

$$n\bar{X} = n\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) \qquad \text{Definition of } \bar{X}$$

$$= \sum_{i=1}^{n} X_i$$

Part 2

Measures of Central Tendency

Basic Concepts

A **summary measure** is a single value that we compute from a collection of measurements in order to describe one of the collection's particular characteristics.

A **measure of central tendency** is a single value that can be used to represent all the other values in the collection.

Notes:

- Some people refer to this measure as the “average”.
- This measure tells us where the “center" of the distribution lies.
- The use of this measure will also facilitate the comparison of two or more collections of measurements.

Arithmetic Mean

- The **arithmetic mean** is the sum of all the values in the collection divided by the total number of elements in the collection.
- The **population mean** for a finite population with **N** elements, denoted by the lowercase Greek letter **μ**, is

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

where X_i is the measure taken from the ith element of the population.

- The **sample mean** for a sample with **n** elements, denoted by $\bar{\mathbf{X}}$ is

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

where X_i is the measure taken from the ith element of the sample.

Example: Arithmetic Mean

Five judges give their scores on the performance of a gymnast as follows: 8, 9, 9, 9, and 10. Find the mean score of the gymnast.

Solution: We compute for the population mean. Let X_i be the score given by the i^{th} judge in the population. Add the scores given by the 5 judges.

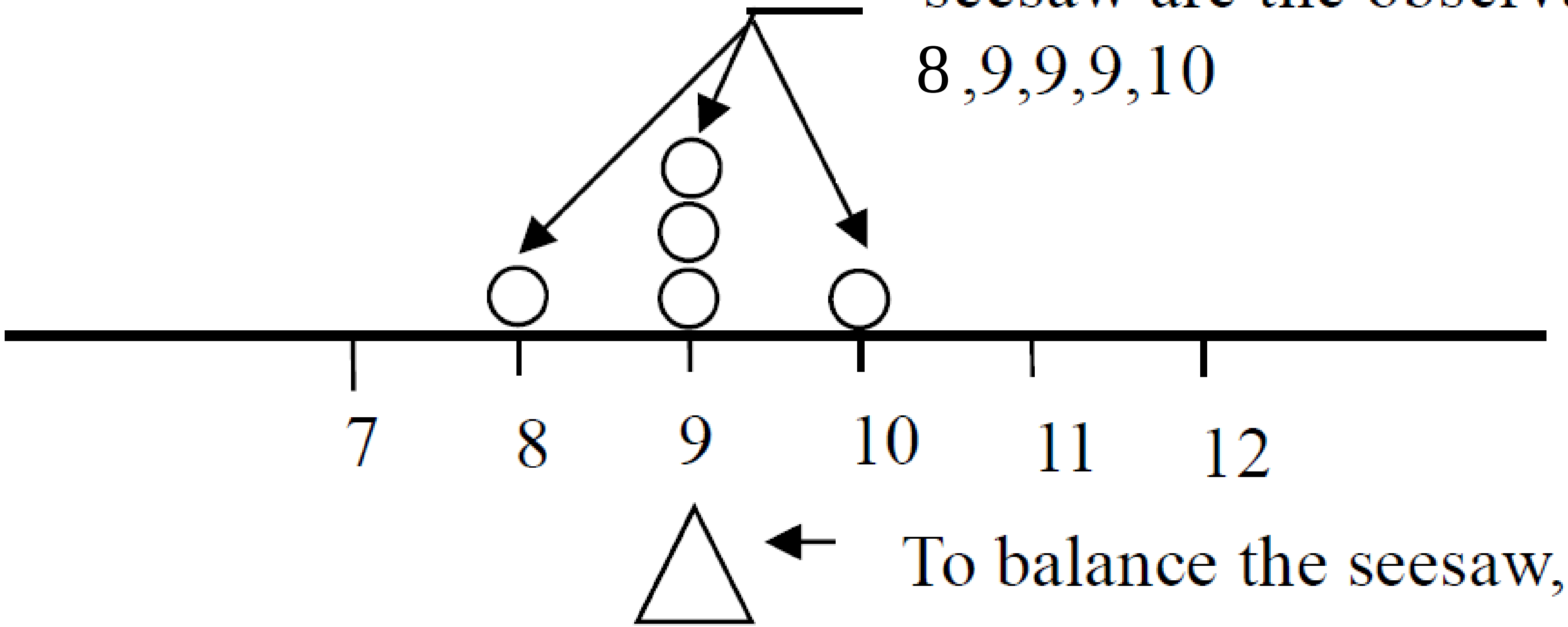
$$\sum_{i=1}^{5} X_i = 8 + 9 + 9 + 9 + 10 = 45$$

Then divide the sum of the scores by the number of judges. We have N=5 judges and get

$$\mu = \frac{45}{5} = 9$$

The mean score of the gymnast is 9.

The Mean as a “Center” of Mass



The loads that we now place on the seesaw are the observations, $X_{i,}$: 8 ,9,9,9,10

To balance the seesaw, the fulcrum must be at what value of μ?

What would happen this time if the last measurement had been 1000 instead of 10?

Outliers

Outliers are data values that are markedly different from the rest of the data items.

- Since the mean is the “center of mass” then its value is gravely affected by outliers. An outlier will pull the value of the arithmetic mean in its direction and away from the location of majority of the observations.
- With the presence of outliers, the mean might not be a suitable measure of central tendency because it may not be a good representative of the observations in the collection.

Illustration: Outliers

a. Let us consider the monthly salaries of a sample of 5 employees: P9,500.00, P10,200.00, P9,000.00, P10,500.00, and P11,000.00. Find the mean.

   Solution: $\bar{X} = \frac{9{,}500+10{,}200+9{,}000+10{,}500+11{,}000}{5} = P10{,}040/month$

b. Suppose the monthly salary of the fifth employee is P60,000.00 instead of P11, 000.00. What will be the mean salary?

   Solution: $\bar{X} = \frac{9{,}500+10{,}200+9{,}000+10{,}500+60{,}000}{5} = P19{,}840/month$

Is the value we got in (b) still a good representative of the values in the collection?

Properties of the Mean

- The mean is the “center of mass”.
- It uses all the observed values in the calculation.
- It may or may not be an actual observed value in the data set.
- We may treat its formula algebraically.
- Its value is gravely affected by outliers.
- The mean of a finite collection always exists and is unique.
- Data values should be measured using **at least an interval** scale.

Mathematical Properties of the Mean

1. The sum of the deviations of the observed values from the mean is zero. That is

$$\sum_{i=1}^{n}(X_i - \bar{X}) = 0$$

2. The term

$$\sum_{i=1}^{n}(X_i - c)^2$$

   where c is a constant, is minimum when the value of c is equal to $\bar{X}$

3. If we add (or subtract) a constant c to all original observations, then the mean of the new observations will increase by the same amount constant c. That is $\bar{X}_{new} = \bar{X}_{old} + c$.

4. If we multiply (or divide) a constant c to all original observations, then the mean of the new observations is the original mean multiplied by the constant c. That is $\bar{X}_{new} = \bar{X}_{old} \times c$.

NOTE: The above properties also apply to the population mean μ.

Example

Let us consider the ages in years of a sample of 5 children: 8, 9, 5, 7, and 6. The mean age of the children is 7 years old, denoted by $\overline{X}_{original}$.

In 5 years, the ages of these children will increase by 5. Get the mean of the new dataset.

If we multiply a constant c=5 to each of the ages, get the mean of the new dataset.

Modifications of Mean

- **Weighted Mean**
  - Used when observations are not of equal importance
  - If we assign a weight to each observation, where i=1,2,…,n, and n is the number of observations in the sample, then the weighted sample mean is given by

$$\bar{X}_W = \frac{\sum_{i=1}^{n} W_i X_i}{\sum_{i=1}^{n} W_i}$$

- **Combined Mean (or the Mean of Means)**
  - Suppose that k finite populations having $N_1, N_2, \dots, N_k$ measurements, respectively, having means $\mu_1, \mu_2, \dots, \mu_k$. The combined population mean, μ_C, if we combine the measurements of all the populations, is $\mu_C = \frac{\sum_{i=1}^{k} N_i \mu_i}{\sum_{i=1}^{k} N_i}$. The **combined sample mean,** $\bar{X}_C$, is defined similarly as $\frac{\sum_{i=1}^{k} n_i \bar{X}_i}{\sum_{i=1}^{k} n_i}$.

Modifications of Mean

- **Trimmed Mean**
  - Objective: remove the influence of possible outliers
  - Choose α (the proportion of observations that will be deleted), $0 < \alpha < 1$.
  - To find $\left(\frac{\alpha}{2}\right)100\%$-trimmed mean for a given dataset, we first arrange the data into an array. Then, we remove $\left(\frac{\alpha}{2}\right)100\%$ of the observations in both the lower and upper ends of the array. We then calculate the arithmetic mean for the remaining observations.

Example

Find the 10% trimmed mean of the following dataset:

{0, 10, 13, 14, 15, 15, 16, 18, 20, 21, 21, 22, 23, 24, 25, 25, 25, 25, 25, 30}

Median

- The **median** divides the array into two equal parts.
- Steps in Finding the Median

  Step 1: Arrange the observations in an array (in ascending order). We denote $X_{(i)}$ as the i^th observation in the array, where $i = 1,2,\dots,n$.

  Step 2: Determine the median, Md.

$$Md = \begin{cases} X_{\left(\frac{n+1}{2}\right)} & n\ is\ odd \\ \dfrac{X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}}{2} & n\ is\ even \end{cases}$$

Example: Median

The following are the total receipts of 7 mining companies (in million pesos):

1.3 6.6 10.5 12.6 50.7 4.7 7.3

Solution:

Array:	1.3	4.7	6.6	7.3	10.5	12.6	50.7
Notation:	$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$	$X_{(5)}$	$X_{(6)}$	$X_{(7)}$

Since $n = 7$ odd, $Md = X_{\left(\frac{n+1}{2}\right)} = X_{\left(\frac{7+1}{2}\right)} = X_{(4)} = 7.3$ million pesos

Question: How many observations are to the left of the median? How many observations are to the right of the median?

A median of 7.3 million pesos indicates that companies with total receipt of less than 7.3 million pesos belong in the lower half of the array; whereas, companies with total receipt greater than 7.3 million pesos belong in the upper half of the array.

Interpretation: Median

- The exact interpretation of the median is as follows:

  “**At least half** of the observations are **less than or equal** to the median and at the same time **at least half** of the observations are **greater than or equal** to the median."

- This general interpretation can handle all types of data sets, including those with tied values in the middle of the array.

- Example:

  $n = 12$ Array: 3, 4, 4, 4, 4, 5, 5, 5, 5, 7, 8, 9 Median = 5

- How many are less than 5? How many are greater than 5?

- A median of 5 means that at least half of the observations are less than or equal to 5 and at the same time at least half are greater than or equal to 5.

- The interpretation will simplify to “half of the observations are less than the median and half are greater than the median" **if the median is not one of the observed values**, that is, n is even and there are no ties.

Effect of Outliers on Median

Let us consider the dataset in the example on the monthly salaries of 5 employees (in pesos): 9,500 10,200 9,000 10,500 60,000. The mean monthly salary is P19,840. Let us now determine the median.

Array: 9000, 9500, 10200, 10500, 60000

$X_{(1)}$ $X_{(2)}$ $X_{(3)}$ $X_{(4)}$ $X_{(5)}$

$$Md = X_{\left(\frac{n+1}{2}\right)} = X_{\left(\frac{5+1}{2}\right)} = X_{(3)} = 10{,}200$$

The median monthly salary is P10,200.

Which is a better measure of central tendency?

Characteristics: Median

- The median is the **“center” of the array**.
- The median is also a **measure of position/location**. An observation whose value is smaller than the median belongs in the lower half of the array while an observation whose value is higher than the median belongs in the upper half of the array.
- Unlike the mean, it **uses only the middle value/s** in the array for its computation.
- Unlike the mean, the median is **not affected by outliers** (observations whose values are extremely different from the others in the data set).
- Unlike the mean, the median is **not amenable to algebraic manipulation**.
- Unlike the mean, the median is still **interpretable** when the level of measurement is as low as **ordinal**.
- The median **will always exist** and is **unique**.

Mode

- The **mode** is the observed value that occurs with the greatest frequency in a data set.
- Example:
  Determine the mode of the number of absences of 20 students in Stat 101:

0	0	0	0	0
0	0	0	0	1
1	1	1	2	3
3	3	4	5	5

Characteristics: Mode

- The mode is the “center” in the sense that it is the **most typical value** in a set of observations.
- Outliers **do not** affect the mode.
- **The mode will not always exist**; and if it does**, it may not be unique**. A data set is said to be unimodal if there is only one mode, bimodal if there are two modes, trimodal if there are three modes, and so on.
- The value of the mode is always **one of the observed values** in the data set.
- We can get the mode for both quantitative and qualitative types of data; that is, the mode is **interpretable** even if the level of measurement is as low as **nominal**.

Characteristics: Mode

- The mode is generally not as useful a measure of central tendency as the mean and the median when the data consist of only a few numbers. For example, for the numbers 7, 12, 18, 22, 31, 31, the mode is 31 since it appears twice and all other numbers only once. But 31 cannot be considered a good measure of central tendency for these data since it is in fact at the extreme high end of the values and its frequency exceeds the frequency of the other values by only 1.

Part 3

Measures of Location (or Position)

Measures of Position

- A **measure of position** provides information on the percentage of observations in the collection whose values are less than or equal to it. It indicates the relative position of an observation in the array.
- A measure of position is also referred to as a **fractile** or **quantile**.

Percentile

- The **percentiles** divide the ordered observations into 100 equal parts.
- There are 99 percentiles, denoted by $P_1, P_2, \dots, P_{99}$. The kth percentile, denoted by P_k, is the value such that at least k% of the observations are less than or equal to it and at least (100-k)% are greater than or equal to it, where k=1,2,…,99.

- Illustration:

  P_1 or the 1st percentile is a value such that at least 1% of the observations are less than or equal to it and at least 99% of the observations are greater than or equal to it.

  P_{25} or the 25th percentile is a value such that at least 25% of the observations are less than or equal to it and at least 75% of the observations are greater than or equal to it.

Percentile Score vs Percentage Score

- **Percentage score** is

(total score/total number of points)100%.

Example:

Total no. of points = 120

Score of Juan = 90

Percentage score of Juan = (90/120) x 100% = 75%

- $\boldsymbol{\alpha}(\mathbf{100\%})$ **Percentile** score indicates that at least $\boldsymbol{\alpha}(\mathbf{100\%})$ of all scores in the collection are **less than or equal** to the individual's score while at least $(\mathbf{1}-\boldsymbol{\alpha})(\mathbf{100\%})$ are **greater than or equal** to the individual's score, $0 < \alpha < 1$.

Percentile Score vs Percentage Score

- Example:

  Juan's percentile score for his section is 95.5. This means that the scores of at least 95.5% of all students in his section are less than or equal to Juan's score and the scores of at least 4.5% are greater than or equal to Juan's score.

  Juan's percentile score for his school is 40. This means that the scores of at least 40% of all students who took the same test in his school are less than or equal to Juan's score and the scores of at least 60% are greater than or equal to Juan's score.

Interpretation: Percentile Score

- Juan's **percentile score** for their section is **95.5**. This means that the scores of **at least 95.5%** of all students in his section are **less than or equal** to Juan's score and the scores of **at least 4.5%** are **greater than or equal** to Juan's score.

- Saying that **at least 95.5%** of all students in his section are **less than or equal** to Juan's score is **equivalent** to saying that **at most 4.5%** are **greater than** his score. So if there are 50 students in Juan's class, the number of students whose scores are greater than Juan's score will not exceed $(50)(.045) = 2.25$ or 2.

Interpretation: Percentile Score

- Saying that **at least 4.5%** are **greater than or equal** to Juan's score is **equivalent** to saying that **at most 95.5%** are **less than his score**. So if there are 50 students in Juan's class, the number of students whose scores are less than Juan's score will not exceed $(50)(.955) = 47.75$ or 47.

Interpretation of P_k

- P_k will be an interpolated value **if $\frac{nk}{100}$ is an integer**. If the values used in the interpolation are not tied values then P_k will not be one of the observations. **In such a case, the interpretation of P_k will simplify as follows**:

  "$k\%$ of observations are less than P_k. Likewise, $(100-k)\%$ are greater than P_k." [analogous to the median. Why?]

  That is, nk/100 observations are less than P_k so that the remaining $n-\left(\frac{nk}{100}\right)=n\left(1-\frac{k}{100}\right)$ are greater than Pk.

Interpolating of P_k

Some Methods for Interpolating P_k

- empirical distribution number with averaging
- weighted average estimate
- observation numbered closed to $nk/100$
- empirical distribution number estimate
- Tukey's method

Note: Before using any software to determine the percentile, make sure you know the method used by the software.

Empirical Distribution Number with Averaging (EDNA)

1. Arrange the n observations in the collection in an array. Denote the i^th ordered observation by $X_{(i)}$.
2. Compute for $\frac{nk}{100}$.
3. Determine P_k using the given formula:

$$P_k = \begin{cases} X_{\left(\left[\!\left[\frac{nk}{100}\right]\!\right]+1\right)} & nk/100 \ \ is\ not\ an\ integer \\ \dfrac{X_{\left(\frac{nk}{100}\right)} + X_{\left(\frac{nk}{100}+1\right)}}{2} & nk/100 \ \ is\ an\ integer \end{cases}$$

Empirical Distribution Number with Averaging

Example:
The following are the total receipts of seven mining companies (in million pesos): 4.6, 1.3, 7.3, 6.6, 10.5, 50.7, and 12.6. Determine the 75th percentile.

Solution: Arrange the data in an array (lowest to highest).

Array:	1.3	4.6	6.6	7.3	10.5	12.6	50.7
Notation:	$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$	$X_{(5)}$	$X_{(6)}$	$X_{(7)}$

Find P_{75}. Thus k=75 and $\frac{nk}{100} = \frac{7(75)}{100} = 5.25$ in not an integer.

$P_{75} = X_{\left(\left\llbracket\frac{nk}{100}\right\rrbracket+1\right)} = X_{(\llbracket 5.25\rrbracket+1)} = X_{(5+1)} = X_{(6)} = 12.6$

How many observations are less than or equal to 12.6? Greater than or equal to 12.6?

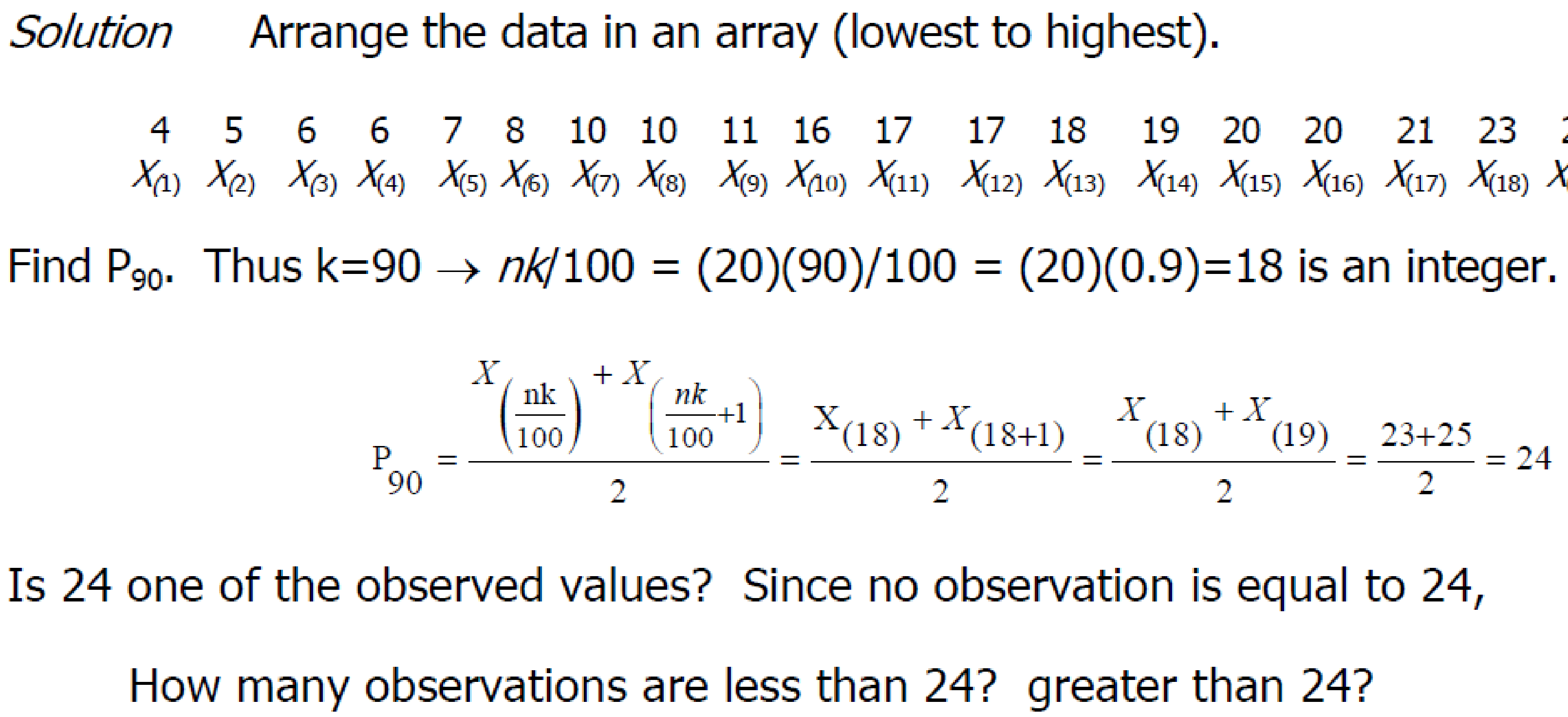
Empirical Distribution Number with Averaging

Example:

The following are the number of years of operation of 20 mining companies:

4, 5, 6, 6, 7, 8, 10, 10, 11, 16, 17, 17, 18, 19, 20, 20, 21, 23, 25, 30

Determine the 90th percentile.



Solution Arrange the data in an array (lowest to highest).

4	5	6	6	7	8	10	10	11	16	17	17	18	19	20	20	21	23	25	30
$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$	$X_{(5)}$	$X_{(6)}$	$X_{(7)}$	$X_{(8)}$	$X_{(9)}$	$X_{(10)}$	$X_{(11)}$	$X_{(12)}$	$X_{(13)}$	$X_{(14)}$	$X_{(15)}$	$X_{(16)}$	$X_{(17)}$	$X_{(18)}$	$X_{(19)}$	$X_{(20)}$

Find P_{90}. Thus k=90 → $nk/100 = (20)(90)/100 = (20)(0.9) = 18$ is an integer.

$$P_{90} = \frac{X_{\left(\frac{\mathrm{nk}}{100}\right)} + X_{\left(\frac{nk}{100}+1\right)}}{2} = \frac{\mathrm{X}_{(18)} + X_{(18+1)}}{2} = \frac{X_{(18)} + X_{(19)}}{2} = \frac{23+25}{2} = 24$$

Is 24 one of the observed values? Since no observation is equal to 24,

How many observations are less than 24? greater than 24?

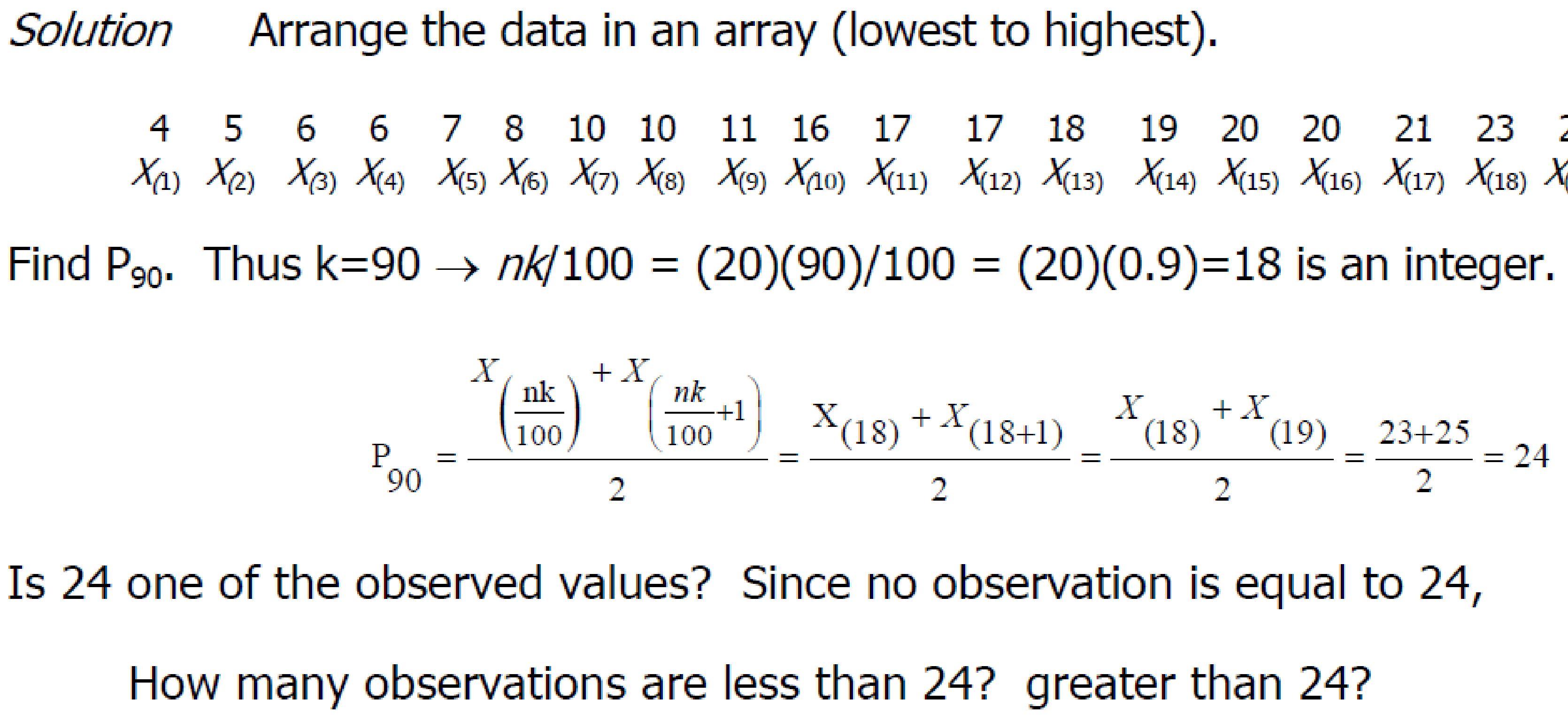
Empirical Distribution Number with Averaging

Example:

The following are the number of years of operation of 20 mining companies:

4, 5, 6, 6, 7, 8, 10, 10, 11, 16, 17, 17, 18, 19, 20, 20, 21, 23, 25, 30

Determine the 90th percentile.



Solution Arrange the data in an array (lowest to highest).

4	5	6	6	7	8	10	10	11	16	17	17	18	19	20	20	21	23	25	30
$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$	$X_{(5)}$	$X_{(6)}$	$X_{(7)}$	$X_{(8)}$	$X_{(9)}$	$X_{(10)}$	$X_{(11)}$	$X_{(12)}$	$X_{(13)}$	$X_{(14)}$	$X_{(15)}$	$X_{(16)}$	$X_{(17)}$	$X_{(18)}$	$X_{(19)}$	$X_{(20)}$

Find P_{90}. Thus k=90 → $nk/100 = (20)(90)/100 = (20)(0.9)=18$ is an integer.

$$P_{90} = \frac{X_{\left(\frac{nk}{100}\right)} + X_{\left(\frac{nk}{100}+1\right)}}{2} = \frac{X_{(18)} + X_{(18+1)}}{2} = \frac{X_{(18)} + X_{(19)}}{2} = \frac{23+25}{2} = 24$$

Is 24 one of the observed values? Since no observation is equal to 24,

How many observations are less than 24? greater than 24?

Weighted Average Estimate

1. Arrange the n observations in the collection in an array. Denote the ith ordered observation by $X_{(i)}$
2. Compute for $\frac{(n+1)k}{100} = j + g$, where j is the integer part and g is the fractional part
3. Determine P_k by linear interpolation, using the given formula:

$$P_k = (1 - g)X_{(j)} + gX_{(j+1)} = X_{(j)} + g\left(X_{(j+1)} - X_{(j)}\right)$$

Weighted Average Estimate

Example:

The following are the total receipts of seven mining companies (in million pesos): 4.6, 1.3, 7.3, 6.6, 10.5, 50.7, and 12.6. Determine the 75th percentile.

Solution:

Arrange the data in an array (lowest to highest).

Array:	1.3	4.6	6.6	7.3	10.5	12.6	50.7
Notation:	$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$	$X_{(5)}$	$X_{(6)}$	$X_{(7)}$

Find P_{75}. Thus k=75 and $\frac{(n+1)k}{100} = \frac{(7+1)(75)}{100} = 6$. The integer part of 6 is j=6 and since there is no fractional part then g=0.

Weighted Average Estimate

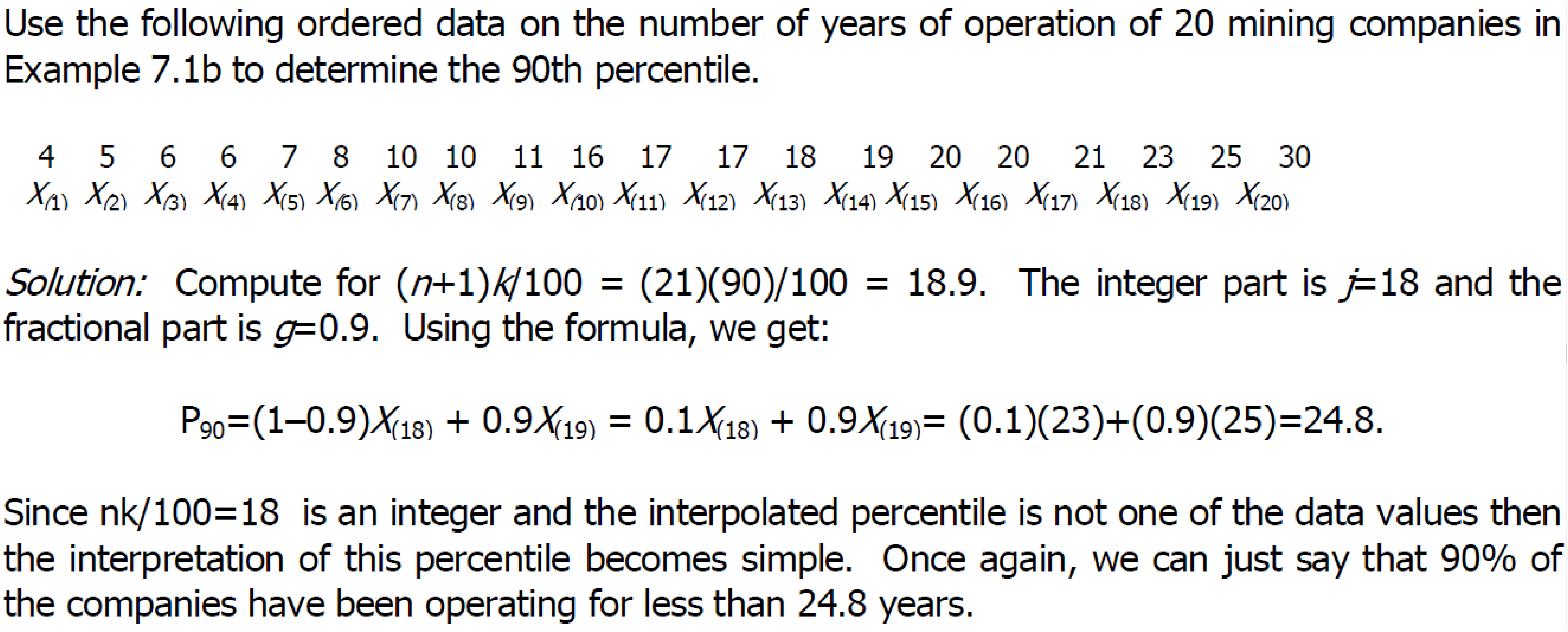
Using the formula, we get:

P$_{75}$ = $(1-0)X_{(6)}+0\cdot X_{(7)}=X_{(6)}=12.6$

- We notice that whenever $\frac{(n+1)k}{100}$ is an integer then the kth percentile is the observation. in the array occupying that position. In other words, $P_k=X_{\left(\frac{(n+1)k}{100}\right)}$.

Weighted Average Estimate

Example:



Use the following ordered data on the number of years of operation of 20 mining companies in Example 7.1b to determine the 90th percentile.

4	5	6	6	7	8	10	10	11	16	17	17	18	19	20	20	21	23	25	30
$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$	$X_{(5)}$	$X_{(6)}$	$X_{(7)}$	$X_{(8)}$	$X_{(9)}$	$X_{(10)}$	$X_{(11)}$	$X_{(12)}$	$X_{(13)}$	$X_{(14)}$	$X_{(15)}$	$X_{(16)}$	$X_{(17)}$	$X_{(18)}$	$X_{(19)}$	$X_{(20)}$

Solution: Compute for $(n+1)k/100 = (21)(90)/100 = 18.9$. The integer part is $j=18$ and the fractional part is $g=0.9$. Using the formula, we get:

$$P_{90}=(1–0.9)X_{(18)} + 0.9X_{(19)} = 0.1X_{(18)} + 0.9X_{(19)}= (0.1)(23)+(0.9)(25)=24.8.$$

Since nk/100=18 is an integer and the interpolated percentile is not one of the data values then the interpretation of this percentile becomes simple. Once again, we can just say that 90% of the companies have been operating for less than 24.8 years.

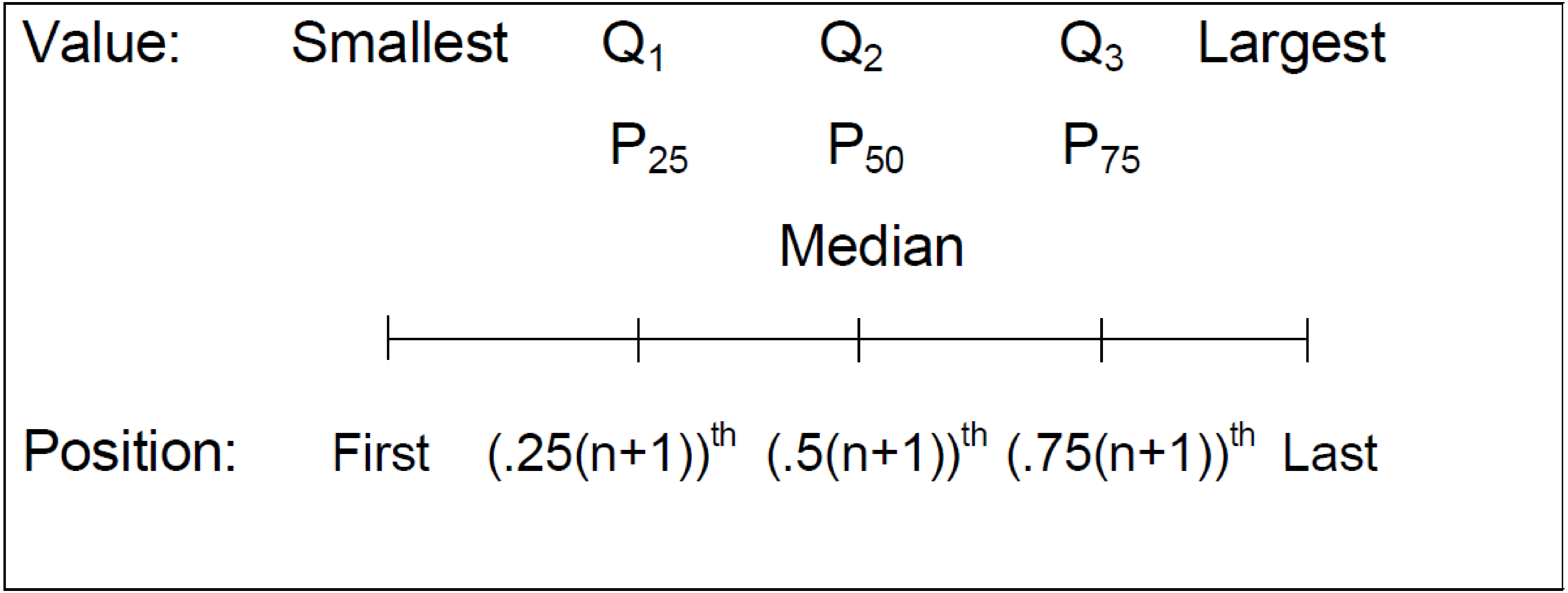
Quartiles

The **quartiles** divide the ordered observations into 4 equal parts.

There are 3 quartiles, denoted by Q1, Q2, and Q3:

- Q1 or the 1st quartile is a value such that at least 25% of the observations are less than or equal to it and at least 75% of the observations are greater than or equal to it.
- Q2 or the 2nd quartile is a value such that at least 50% of the observations are less than or equal to it and at least 50% of the observations are greater than or equal to it. In other words, Q2 is the **median**.
- Q3 or the 3rd quartile is a value such that at least 75% of the observations are less than or equal to it and at least 25% of the observations are greater than or equal to it.

Relationship of Quartiles, Percentiles and Median



Value:	Smallest	Q_1	Q_2	Q_3	Largest
		P_{25}	P_{50}	P_{75}	
			Median		
Position:	First	$(.25(n+1))^{th}$	$(.5(n+1))^{th}$	$(.75(n+1))^{th}$	Last

To determine the quartile, compute for the corresponding percentile.

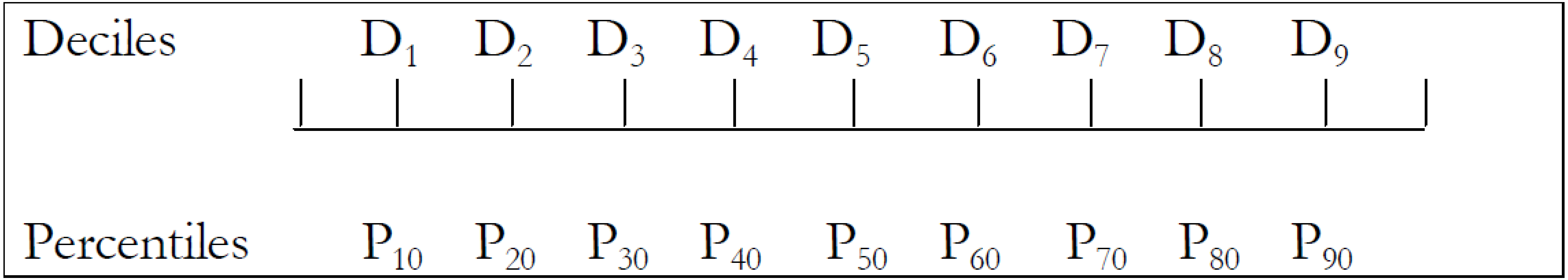
Deciles

The **deciles** divide the ordered observations into 10 equal parts.

There are 9 deciles, denoted by D1, D2,..., D9.

- D1 or the 1st decile is a value such that at least 10% of the observations are less than or equal to it and at least 90% of the observations are greater than or equal to it.
- D5 or the 5th decile is a value such that at least 50% of the observations are less than or equal to it and at least 50% of the observations are greater than or equal to it. In other words, D5 is the **median**.
- D9 or the 9th decile is a value such that at least 90% of the observations are less than or equal to it and at least 10% of the observations are greater than or equal to it.

Relationship of Deciles, Percentiles and Median



Deciles	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9
Percentiles	P_{10}	P_{20}	P_{30}	P_{40}	P_{50}	P_{60}	P_{70}	P_{80}	P_{90}

To determine the decile, compute for the corresponding percentile.