Chapter 2

Descriptive Statistics

Introduction

The first step in data analysis is to describe the data in some concise manner.

Descriptive statistics that involve numeric or graphic display are crucial in capturing and conveying the final results of studies in publications.

Features of good numeric or graphic form of data summarization:

- Self-contained
- Understandable without reading the text
- Clearly labeled of attributes with well-defined terms
- Indicate principal trends in data

Measures of Location

It is easy to lose track of the overall picture when there are too many sample points.

Data summarization is important before any inferences can be made about the population from which the sample points have been obtained.

Measure of location is a type of measure useful for data summarization that defines the center or middle of the sample.

Defining the Middle: The Arithmetic Mean

Table 2.1 Sample of birthweights (g) of live-born infants born at a private hospital in San Diego, California, during a 1-week period

i	x_i	i	x_i	i	x_i	i	x_i
1	3265	6	3323	11	2581	16	2759
2	3260	7	3649	12	2841	17	3248
3	3245	8	3200	13	3609	18	3314
4	3484	9	3031	14	2838	19	3101
5	4146	10	2069	15	3541	20	2834

Arithmetic mean: the sum of all the observations divided by the number of observations.

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

Limitation: Oversensitive to extreme values; in which case, it may not be representative of the location of the majority of sample points.

Arithmetic Mean Explained

Sigma ∑ is a summation sign.

$\sum_{i=1}^{n} x_i$ implies $(x_1 + x_2 + \ldots + x_n)$

- If a and b are integers where $a \leq b$, then $\sum_{i=a}^{b} x_i$ meaning $(x_a + x_{a+1} + \ldots + x_b)$
- If a = b, then $\sum_{i=a}^{b} x_i = x_a$
- If c is some constant, then $\sum_{i=1}^{n} cx_i = c\left(\sum_{i=1}^{n} x_i\right)$

EXAMPLE 2.3

If $x_1 = 2 \quad x_2 = 5 \quad x_3 = -4$

find $\sum_{i=1}^{3} x_i \quad \sum_{i=2}^{3} x_i \quad \sum_{i=1}^{3} x_i^2 \quad \sum_{i=1}^{3} 2x_i$

Solution:

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

$$\sum_{i=1}^{3} x_i^2 = 4+25+16=45 \quad \sum_{i=1}^{3} 2x_i = 2\sum_{i=1}^{3} x_i = 6$$

EXAMPLE 2.4 What is the arithmetic mean for the sample of birthweights in Table 2.1?

TABLE 2.1 Sample of birthweights (g) of live-born infants born at a private hospital in San Diego, California, during a 1-week period

i	x_i	i	x_i	i	x_i	i	x_i
1	3265	6	3323	11	2581	16	2759
2	3260	7	3649	12	2841	17	3248
3	3245	8	3200	13	3609	18	3314
4	3484	9	3031	14	2838	19	3101
5	4146	10	2069	15	3541	20	2834

$$\bar{x} = (3265 + 3260 + \cdots + 2834)/20 = 3166.9 \text{ g}$$

Median

Sample median is

- $\left(\frac{n+1}{2}\right)$ th the largest observation if n is odd
- Average of the $\left(\frac{n}{2}\right)$ th and the $\left(\frac{n}{2}+1\right)$ th observation if n is even

Example: Calculating the median:

a) 22,15,18,13,35,22

Order the sample as follows:

13,15,18,22,22,35

Since n=6 even then the Median

$$\frac{X_{\frac{6}{2}}+X_{(\frac{6}{2}+1)}}{2}=\frac{18+22}{2}=\frac{40}{2}=20$$

b) 22,15,18,13,22

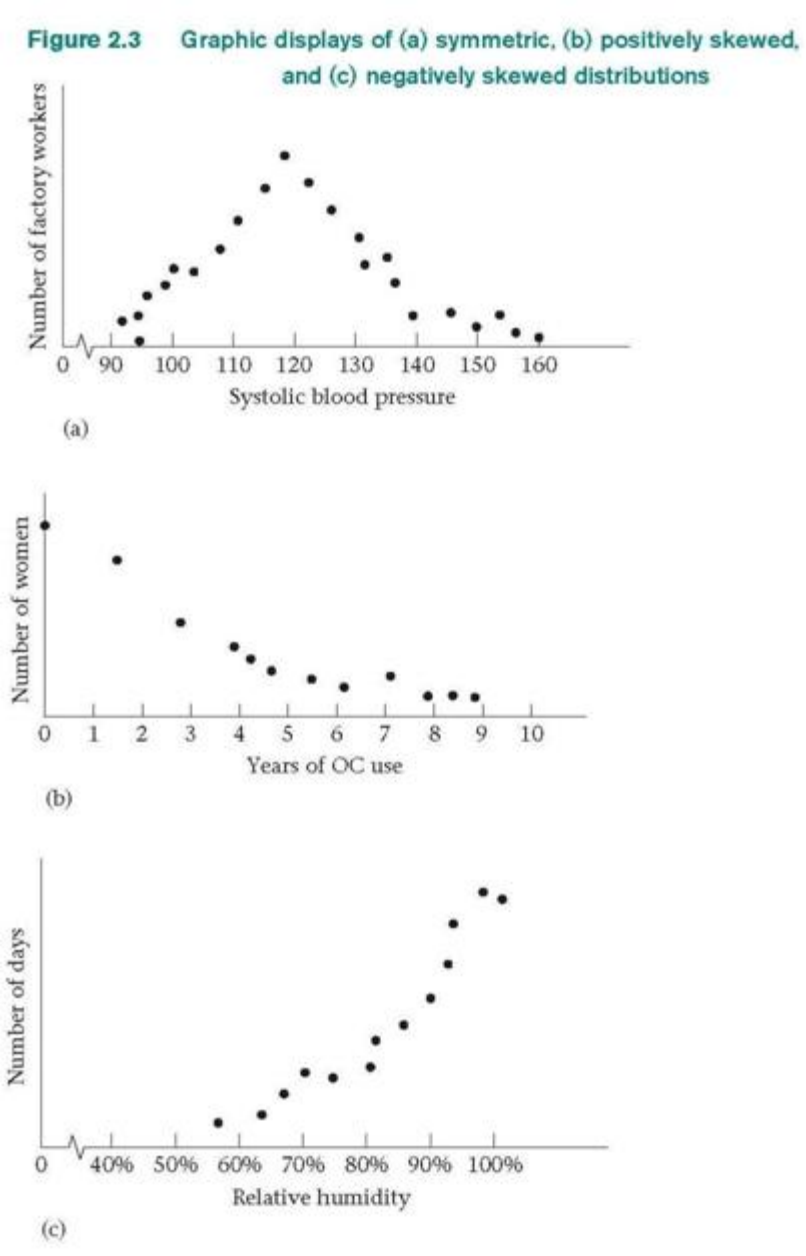
Order the sample as follows:

13,15,18,22,22

Since $n = 5$ odd then the Median

$$X_{(\frac{5+1}{2})} = X_{(3)} = 18$$

Comparing Mean and Median



For symmetric distributions, arithmetic mean is approximately the same as the median

For positively skewed distributions, arithmetic mean tends to be larger than the median

For negatively skewed distributions, the arithmetic mean tends to be smaller than the median

EXAMPLE 2.5

Compute the sample median for the sample in Table 2.1.

Solution: First, arrange the sample in ascending order:

2069, 2581, 2759, 2834, 2838, 2841, 3031, 3101, 3200, 3245, 3248, 3260, 3265, 3314, 3323, 3484, 3541, 3609, 3649, 4146

Because n is even,

$$\text{Sample median} = \text{average of the 10th and 11th largest observations} = (3245 + 3248)/2 = 3246.5 \text{ g}$$

EXAMPLE 2.6

Infectious Disease Consider the data set in Table 2.3, which consists of white-blood counts taken upon admission of all patients entering a small hospital in Allentown, Pennsylvania, on a given day. Compute the median white-blood count.

TABLE 2.3 Sample of admission white-blood counts (× 1000) for all patients entering a hospital in Allentown, Pennsylvania, on a given day

i	x_i	i	x_i
1	7	6	3
2	35	7	10
3	5	8	12
4	9	9	8
5	8		

Solution: First, order the sample as follows: 3, 5, 7, 8, 8, 9, 10, 12, 35. Because n is odd, the sample median is given by the fifth largest point, which equals 8 or 8000 on the original scale.

Mode

Mode: the most frequently occurring value among all the observations in a sample.

Data distributions may have one or more modes.

One mode = unimodal

Two modes = bimodal

Three modes = trimodal and so on.

Example

TABLE 2.4 Sample of time intervals between successive menstrual periods (days) in college-age women

Value	Frequency	Value	Frequency	Value	Frequency
24	5	29	96	34	7
25	10	30	63	35	3
26	28	31	24	36	2
27	64	32	9	37	1
28	185	33	2	38	1

Mode is 28

EXAMPLE 2.8 Compute the mode of the distribution in Table 2.3.

Solution: The mode is 8000 because it occurs more frequently than any other white-blood count.

EXAMPLE 2.9 Compute the mode of the distribution in Table 2.1.

Solution: There is no mode, because all the values occur exactly once.

2.3 Some Properties of Arithmetic Mean

Original sample : $x_1, \ldots, x_n$

Translated sample : $x_1 + c, \ldots, x_n + c$ (where c is some constant)

Let $y_i = x_i + c$ $i = 1, \ldots, n$ then $\overline{y} = \overline{x} + c$

TABLE 2.6 Translated sample for the duration between successive menstrual periods in college-age women

Value	Frequency	Value	Frequency	Value	Frequency
–4	5	1	96	6	7
–3	10	2	63	7	3
–2	28	3	24	8	2
–1	64	4	9	9	1
0	185	5	2	10	1

Note: $\overline{y} = [(-4)(5) + (-3)(10) + \ldots + (10)(1)]/500 = 0.54$

$\overline{x} = \overline{y} + 28 = 0.54 + 28 = 28.54$ days

If the unit or scale changes, then using the **rescaled sample**

$$y_i = cx_i \qquad i = 1, \ldots, n$$

Arithmetic mean is then $\overline{y} = c\overline{x}$

Let $x_1, \ldots, x_n$ be the original sample of data.

Let $y_i = c_1 x_i + c_2$ $\quad i = 1, \ldots, n$ represent a transformed

sample obtained by multiplying each original sample point by a factor c_1 and then shifting over by a constant c_2

If $y_i = c_1 x_i + c_2$ $\quad i = 1, \ldots, n$

then $\overline{y} = c_1\overline{x} + c_2$

Example: Suppose that $\bar{x} = 55$, let $y = 5 + 3x$, terraforms the sample $x_1, x_2, \ldots, x_n$ *in to* $y_1, y_2, \ldots, y_n$,find $\bar{y}$.

Solution

$$\bar{y} = 5 + 3\bar{x}$$
$$\bar{y} = 5 + 3(55) = 170$$

EXAMPLE 2.13

If we have a sample of temperatures in °C with an arithmetic mean of 11.75°C, then what is the arithmetic mean in °F?

Solution: Let y_i denote the °F temperature that corresponds to a °C temperature of x_i. The required transformation to convert the data to °F would be

$$y_i = \frac{9}{5}x_i + 32, \quad i = 1, \ldots, n$$

so the arithmetic mean would be

$$\bar{y} = \frac{9}{5}(11.75) + 32 = 53.15°\text{F}$$

Key Terms

Descriptive statistics: methods used to organize and summarize observed data.

Measure of location: a numerical measure describing the center of a sample.

Mean: sum of observations divided by their number.
Median: middle value of ordered observations.

Mode: most frequently occurring value.

Quick Review Questions

- What are two characteristics of a good numerical or graphical summary?
- How is the arithmetic mean calculated, and what is its main limitation?
- How do you find the median when the sample size is odd? What changes when it is even?
- What do the terms unimodal and bimodal mean?
- For a positively skewed distribution, how does the mean usually compare with the median?
- If every observation is transformed by $y_i = 3x_i + 2$, what is the new sample mean in terms of $\bar{x}$?