# Chapter 6:
# Descriptive Statistics

MAS291 - STATISTICS & PROBABILITY

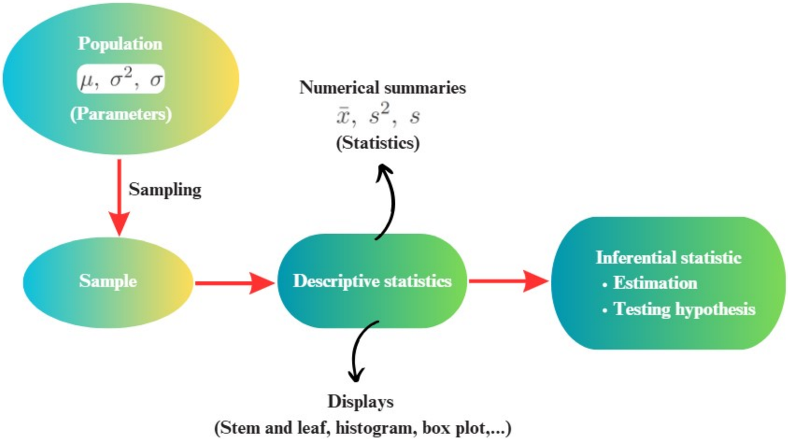
Ly Anh Duong

duongla3@fe.edu.vn

# Table of Contents

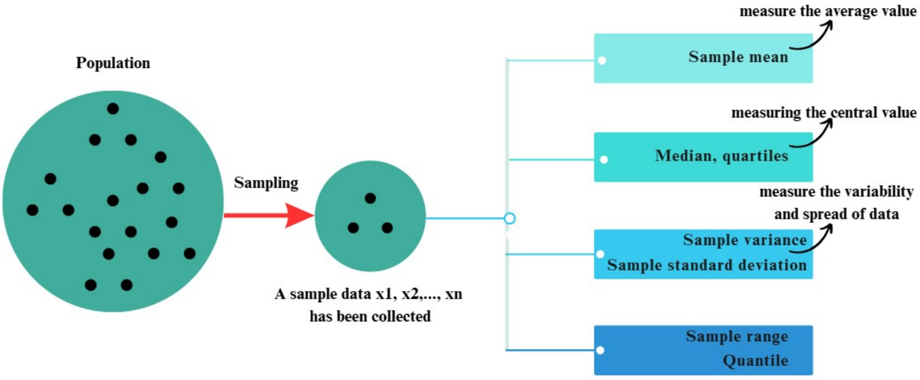
# Introduction

## 1 Numerical Summaries of Data



# Descriptive statistics

## 1 Numerical Summaries of Data



If the $n$ observations in a sample are denoted by $x_1, x_2, ..., x_n$ then

- The **sample mean** is $\overline{x} = \dfrac{x_1 + x_2 + ... + x_n}{n}$
- The **sample variance** is $s^2 = \dfrac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$
- The **sample range** is $r = \max(x_i) - \min(x_i)$

When the population is finite and consists of $N$ values then

- The **population mean** is $\mu = \dfrac{x_1 + x_2 + ... + x_N}{N}$
- The **population variance** is $\sigma^2 = \dfrac{\sum_{i=1}^{N}(x_i - \overline{x})}{N}$

# Sample mean



If the $n$ observations in a sample are denoted by $x_1, x_2, ..., x_n$ then the **sample mean** is

$$\overline{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{\sum\limits_{i=1}^{n} x_i}{n}$$

**Example:** Let's consider the eight observations on pull-off force collected from the prototype engine connectors from Chapter 1. The eight observations are $x_1 = 12.6, x_2 = 12.9, x_3 = 13.4,\ x_4 = 12.3, x_5 = 13.6, x_6 = 13.5, x_7 = 12.6$, and $x_8 = 13.1$. The sample mean is

$$\begin{aligned}\bar{x} &= \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{8} x_i}{8} \\ &= \frac{12.6 + 12.9 + \cdots + 13.1}{8} = \frac{104}{8} = 13.0 \text{ pounds}\end{aligned}$$

A physical interpretation of the sample mean as a measure of location is shown in the dot diagram of the pull-off force data. See Figure 6.1. Notice that the sample mean $\bar{x} = 13.0$ can be thought of as a balance point. That is, if each observation represents 1 pound of mass placed at the point on the $x$-axis, a fulcrum located at $\bar{x}$ would balance this system of weights exactly.

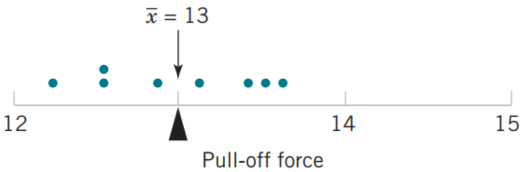



**FIGURE 6.1**

**Dot diagram showing the sample mean as a balance point for a system of weights.**

# Sample median

1. The value that lies in the middle of the data when the data set is ordered.
2. Measures the center of an ordered data set by dividing it into two equal parts.
3. If the data set has an (L$= \dfrac{n+1}{2}$)
   - even number of entries: median is the mean of the two middle data entries.
   - odd number of entries: median is the middle data entry.

# Sample median - Example



The female students in an undergraduate engineering core course at ASU self-reported their heights to the nearest inch. The data follow. Calculate the sample median of height.

62 64 61 67 65 68 61 65 60 65 64 63 59 68 64 66 68 69 65 67 62 66 68 67 66 65 69 65 69 65 67 67 65 63 64 67 65

**Solution**

Sort the data

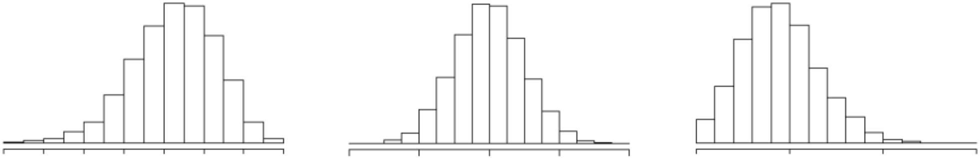
59 60 61 61 62 62 63 63 64 64 64 64 65 65 65 65 65 65 65 65 65 66 66 66 67 67 67 67 67 67 68 68 68 68 69 69 69

n = 37 is odd and (n + 1)/2 = 19

The 19th smallest observation = 65 is the sample median

# Mean vs Median

## 1 Numerical Summaries of Data



Mean < Median
Left-skewed
Negative skew

Median = Mean
Symmetric

Median < Mean
Right-skewed
Positive skew

# Sample mode



- The data entry that occurs with the greatest frequency.
- If no entry is repeated the data set has no mode.
- If two entries occur with the same greatest frequency, each entry is a mode (bimodal).

**Example:** At a political debate a sample of audience members was asked to name the political party to which they belong. Their responses are shown in the table. What is the mode of the responses?

| Political Party | Frequency |
|---|---|
| Democrat | 35 |
| Republican | 60 |
| Other | 25 |
| Did not respond | 8 |

# Sample Variance and Standard Deviation



The **sample variance** is

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

The **sample standard deviation**, s, is the positive square root of the sample variance.

**Example:** The female students in an undergraduate engineering core course at ASU self-reported their heights to the nearest inch. The data follow. Calculate the sample variance and the standard deviation.

62 64 61 67 65 68 61 65 60 65 64 63 59 68 64 66 68 69 65 67 62 66 68 67 66 65 69 65 69 65 67 67 65 63 64 67 65

# Sample Variance and Standard Deviation



The **sample variance** is

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

The **sample standard deviation**, s, is the positive square root of the sample variance.

**Example:** The female students in an undergraduate engineering core course at ASU self-reported their heights to the nearest inch. The data follow. Calculate the sample variance and the standard deviation.

62 64 61 67 65 68 61 65 60 65 64 63 59 68 64 66 68 69 65 67 62 66 68 67 66 65 69 65 69 65 67 67 65 63 64 67 65

**Solution**

The sample variance is $s^2 = 6.47$

The sample standard deviation is $s = 2.54$

# Example - Sample Variance



**EXAMPLE 6.2** | Sample Variance

Table 6.1 displays the quantities needed for calculating the sample variance and sample standard deviation for the pull-off force data. These data are plotted in Figure 6.2. The numerator of $s^2$ is

$$\sum_{i=1}^{8}(x_i - \overline{x})^2 = 1.60$$

so the sample variance is

$$s^2 = \frac{1.60}{8-1} = \frac{1.60}{7} = 0.2286 \text{ (pounds)}^2$$

and the sample standard deviation is

$$s = \sqrt{0.2286} = 0.48 \text{ pounds}$$

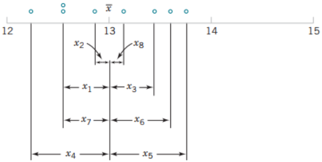



**FIGURE 6.2**

**How the sample variance measures variability through the deviations $x_i - \overline{x}$.**

**TABLE 6.1** **Calculation of Terms for the Sample Variance and Sample Standard Deviation**

| $i$ | $x_i$ | $x_i - \overline{x}$ | $(x_i - \overline{x})^2$ |
|---|---|---|---|
| 1 | 12.6 | −0.4 | 0.16 |
| 2 | 12.9 | −0.1 | 0.01 |
| 3 | 13.4 | 0.4 | 0.16 |
| 4 | 12.3 | −0.7 | 0.49 |
| 5 | 13.6 | 0.6 | 0.36 |
| 6 | 13.5 | 0.5 | 0.25 |
| 7 | 12.6 | −0.4 | 0.16 |
| 8 | 13.1 | 0.1 | 0.01 |
| Total | 104.0 | 0.0 | 1.60 |

## Computation for $s^2$



$$s^2 := \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1}$$

**Example:** Shortcut Calculation for $s^2$

$$s^2 := \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1} = \frac{1353.6 - \frac{104^2}{8}}{7} = 0.2286\ (pound)^2$$

and

$s = \sqrt{0.2286} = 0.48$ pounds

These results agree exactly with those obtained previously.

# Sample range

1 Numerical Summaries of Data

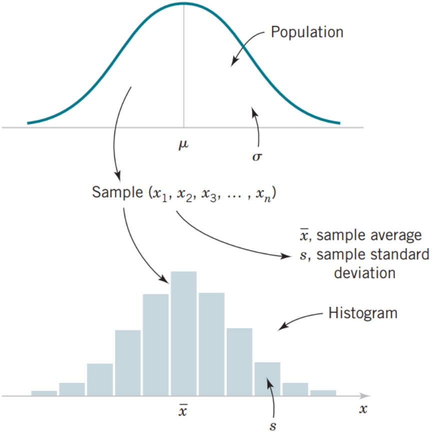
- The difference between the maximum and minimum data entries in the set.
- The data must be quantitative.
- If the $n$ observations in a sample are denoted by $x_1, x_2, \cdots, x_n$, the sample range is

$$r = \max(x_i) - \min(x_i)$$

Casio: 580VN
MENU $\Longrightarrow$ 6 $\Longrightarrow$ INPUT DATA
OPTION $\Longrightarrow$ 3

# Relationship between a population and a sample





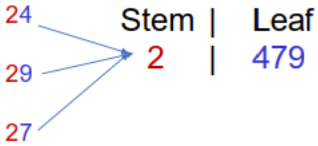
# Table of Contents

2 Stem-and-Leaf diagrams

# Stem-and-Leaf diagrams

## Steps to construct a stem-and-leaf diagram

(1) Divide each number xi into two parts: a stem, consisting of one or more of the leading digits, and a leaf, consisting of the remaining digit.
(2) List the stem values in a vertical column.
(3) Record the leaf for each observation beside its stem.
(4) Write the units for stems and leaves on the display.



# Stem-and-Leaf diagrams - Example



**TABLE 6.2** Compressive Strength (in psi) of 80 Aluminum-Lithium Alloy Specimens

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 105 | 221 | 183 | 186 | 121 | 181 | 180 | 143 |
| 97 | 154 | 153 | 174 | 120 | 168 | 167 | 141 |
| 245 | 228 | 174 | 199 | 181 | 158 | 176 | 110 |
| 163 | 131 | 154 | 115 | 160 | 208 | 158 | 133 |
| 207 | 180 | 190 | 193 | 194 | 133 | 156 | 123 |
| 134 | 178 | 76 | 167 | 184 | 135 | 229 | 146 |
| 218 | 157 | 101 | 171 | 165 | 172 | 158 | 169 |
| 199 | 151 | 142 | 163 | 145 | 171 | 148 | 158 |
| 160 | 175 | 149 | 87 | 160 | 237 | 150 | 135 |
| 196 | 201 | 200 | 176 | 150 | 170 | 118 | 149 |

# Stem-and-Leaf diagrams - Example



Stem-and-leaf diagram for the compressive strength data in Table 6.2.

| Stem | Leaf | Frequency |
|---|---|---|
| 7 | 6 | 1 |
| 8 | 7 | 1 |
| 9 | 7 | 1 |
| 10 | 5 1 | 2 |
| 11 | 5 8 0 | 3 |
| 12 | 1 0 3 | 3 |
| 13 | 4 1 3 5 3 5 | 6 |
| 14 | 2 9 5 8 3 1 6 9 | 8 |
| 15 | 4 7 1 3 4 0 8 8 6 8 0 8 | 12 |
| 16 | 3 0 7 3 0 5 0 8 7 9 | 10 |
| 17 | 8 5 4 4 1 6 2 1 0 6 | 10 |
| 18 | 0 3 6 1 4 1 0 | 7 |
| 19 | 9 6 0 9 3 4 | 6 |
| 20 | 7 1 0 8 | 4 |
| 21 | 8 | 1 |
| 22 | 1 8 9 | 3 |
| 23 | 7 | 1 |
| 24 | 5 | 1 |

# Table of Contents

3 Frequency Distributions and Histograms

Frequency Distributions

- More compact than a stem-and-leaf diagram
- The range of the data is divided into intervals (class intervals, cells, bins)

A frequency histogram consists of columns, one for each bin, whose height is determined by the number of observations in the bin.

A relative frequency histogram has the same shape but a different vertical scale. Its column heights represent the proportion of all data that appeared in each bin.

# Histograms

3 Frequency Distributions and Histograms

- The histogram is a visual display of the frequency distribution.
- Histograms have a shape similar to the pmf or pdf of data, especially in large samples.
- Histograms are stable and reliable for large data sets, preferably of size 75 to 100 or more.

**Constructing a Histogram (Equal Bin Widths)**

(1) Label the bin (class interval) boundaries on a horizontal scale.
(2) Mark and label the vertical scale with the frequencies or the relative frequencies.
(3) Above each bin, draw a rectangle where height is equal to the frequency (or relative frequency) corresponding to that bin.

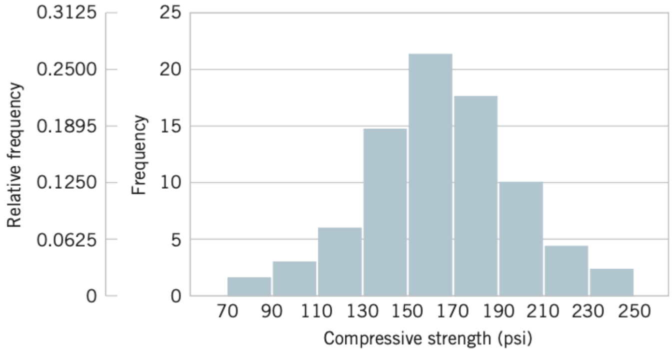


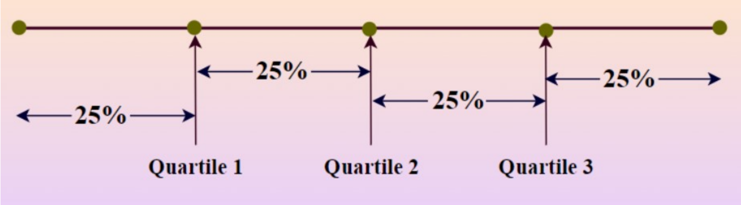

**FIGURE 6.7**

**Histogram of compressive strength for 80 aluminum-lithium alloy specimens.**

# Table of Contents

# Box Plots

The box plot is a graphical display that simultaneously describes several important features of a data set, such as center, spread, departure from symmetry, and identification of unusual observations or outliers.




The quartiles $q_1$, $q_2$, $q_3$ of the sample data are the values at positions: $0.25(n+1)$, $0.5(n+1)$, $0.75(n+1)$, respectively.
The interquartile (IQR) is defined by $IQR = q_3 q_1$

# Box Plots

## 4 Box Plots

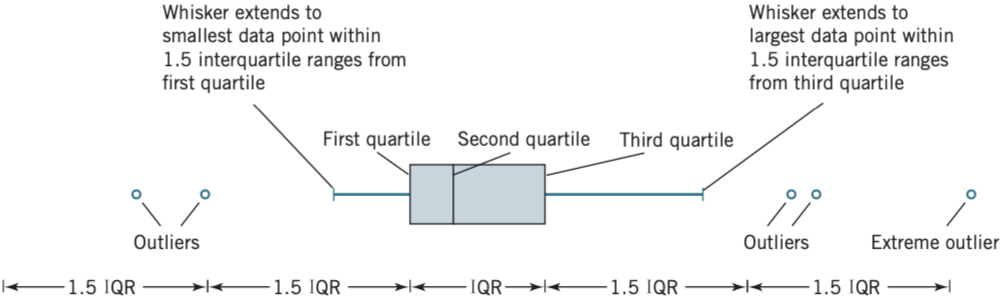



**FIGURE 6.13**

**Description of a box plot.**

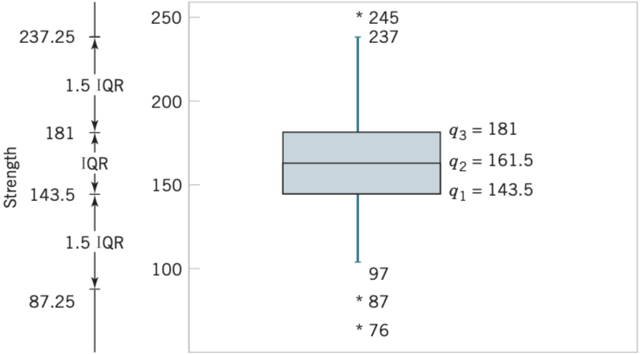



**FIGURE 6.14**

**Box plot for compressive strength data in Table 6.2.**

# Table of Contents

A time series or time sequence is a data set in which the observations are recorded in the order in which they occur.

A time series plot

- the vertical axis denotes the observed value
- the horizontal axis denotes the time

In a time series plot, we often see

- trends
- cycles
- or other broad features of the data

# Time Sequence Plots

## 5 Time Sequence Plots

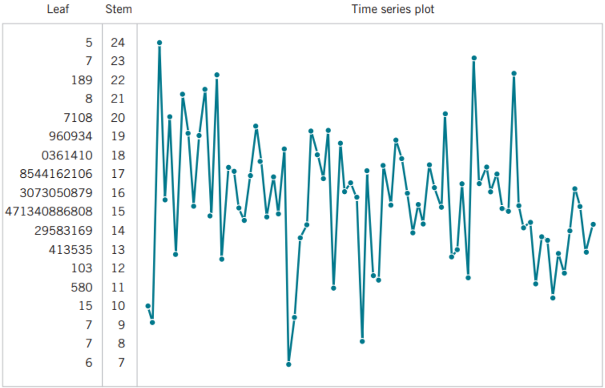



**FIGURE 6.17**

**A digidot plot of the compressive strength data in Table 6.2.**

# Q&A

*Thank you for listening!*