

INTRODUCTION TO PROBABILITY AND STATISTICS FOURTEENTH EDITION



Chapter 2

Describing Data with Numerical Measures

DESCRIBING DATA WITH NUMERICAL MEASURES

- Graphical methods may not always be sufficient for describing data.
- **Numerical measures** can be created for both **populations** and **samples**.
 - A **parameter** is a numerical descriptive measure calculated for a **population**.
 - Ex: mean μ , variance σ^2 .
 - A **statistic** is a numerical descriptive measure calculated for a **sample**.
 - Ex: sample mean \bar{X} , sample variance s^2 .



A decorative graphic on the left side of the slide. It features a series of vertical stripes in various shades of blue and white. Overlaid on these stripes are several circles of different sizes, also in shades of blue. The circles are arranged in a way that they appear to be floating or overlapping the stripes.

2.1 MEASURES OF CENTER

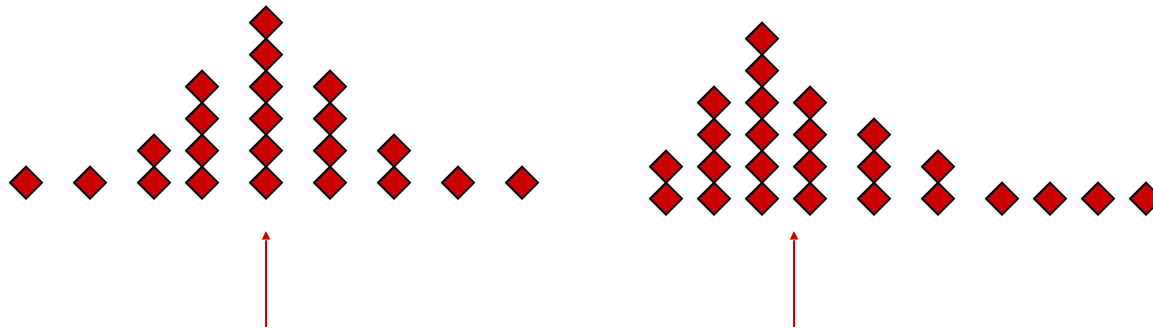
NUMERICAL MEASURES

- A **parameter** is a numerical descriptive measure calculated for a **population**.
 - N measurements, x_1, \dots, x_N
 - Ex: population mean μ , population variance σ^2
- A **statistic** is a numerical descriptive measure calculated for a **sample**.
 - n measurements, x_1, \dots, x_n
 - Ex: sample mean \bar{x} , sample variance s^2



MEASURES OF CENTER

- A measure along the horizontal axis of the data distribution that locates the center of the distribution.



ARITHMETIC MEAN OR AVERAGE

- The **mean** of a set of measurements is the sum of the measurements divided by the total number of measurements.

$$\bar{x} = \frac{\sum x_i}{n}$$

where n = number of measurements

$\sum x_i$ = sum of all the measurements



Notation

Sample mean: $\bar{x} = \frac{\sum x_i}{n}$

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



EXAMPLE

- The set: 2, 9, 11, 5, 6

$$\bar{x} = \frac{\sum x_i}{n} = \frac{2 + 9 + 11 + 5 + 6}{5} = \frac{33}{5} = 6.6$$

If we were able to enumerate the whole population, the **population mean** would be called μ (the Greek letter “mu”).



MEDIAN

- The **median** of a set of measurements is the middle measurement when the measurements are ranked from smallest to largest.
- The **position of the median** is

$$.5(n + 1)$$

once the measurements have been ordered.



EXAMPLE

- The set: 2, 4, 9, 8, 6, 5, 3 $n = 7$
- Sort: 2, 3, 4, 5, 6, 8, 9
- Position: $.5(n + 1) = .5(7 + 1) = 4^{\text{th}}$

Median = 4th largest measurement

- The set: 2, 4, 9, 8, 6, 5 $n = 6$
- Sort: 2, 4, 5, 6, 8, 9
- Position: $.5(n + 1) = .5(6 + 1) = 3.5^{\text{th}}$

Median = $(5 + 6)/2 = 5.5$ — average of the 3rd and 4th measurements



MODE

- The **mode** is the measurement which occurs most frequently.
- The set: 2, 4, 9, 8, 8, 5, 3
 - The mode is **8**, which occurs twice
- The set: 2, 2, 9, 8, 8, 5, 3
 - There are two modes—**8** and **2** (bimodal)
- The set: 2, 4, 9, 8, 5, 3
 - There is **no mode** (each value is unique).



EXAMPLE

The number of quarts of milk purchased by 25 households:

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

- Mean?

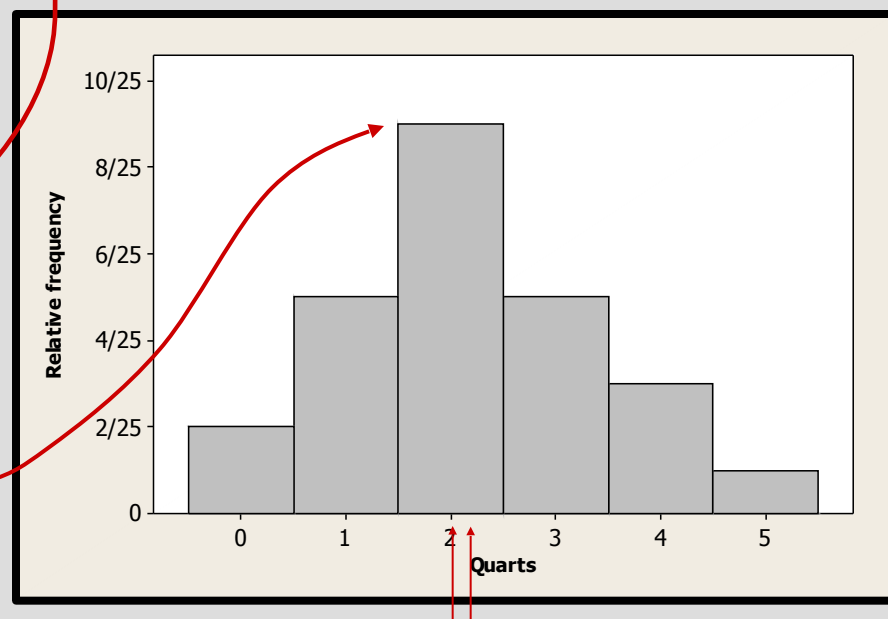
$$\bar{x} = \frac{\sum x_i}{n} = \frac{55}{25} = 2.2$$

- Median?

$$m = 2$$

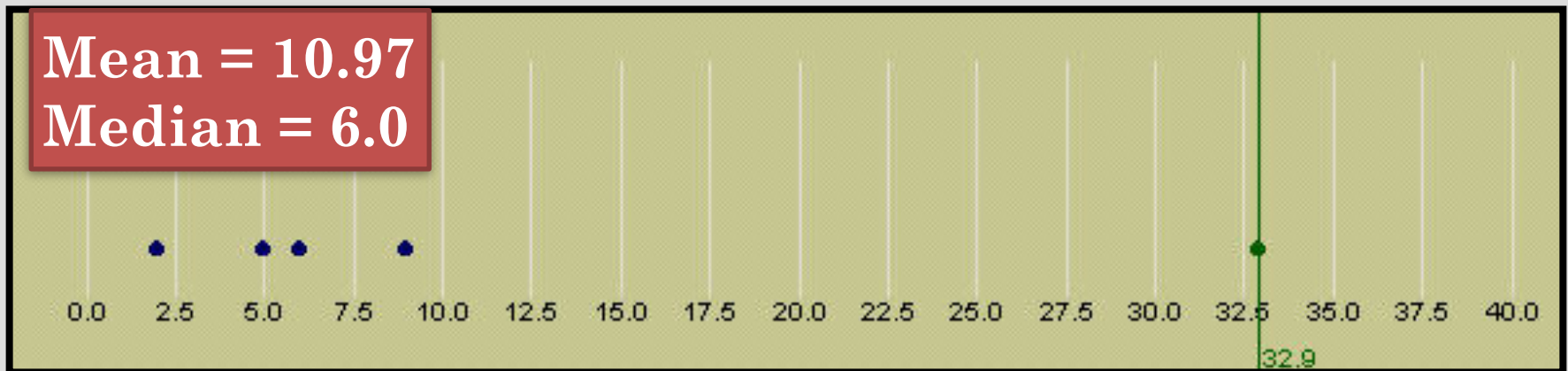
- Mode? (Highest peak)

$$\text{mode} = 2$$



EXTREME VALUES

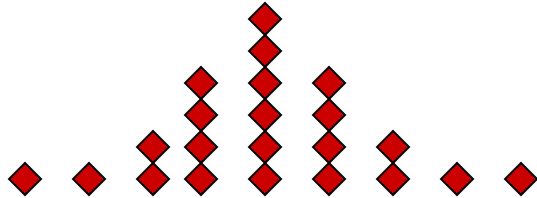
- The mean is more easily affected by extremely large or small values than the median.



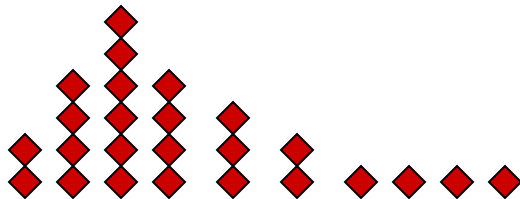
- The median is often used as a measure of center when the distribution is skewed.



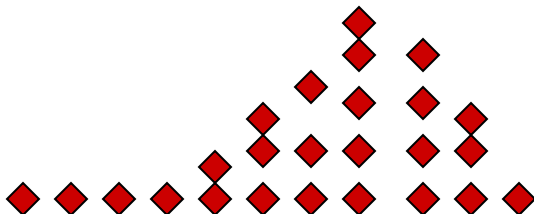
EXTREME VALUES



Symmetric: Mean = Median



Skewed right: Mean > Median



Skewed left: Mean < Median



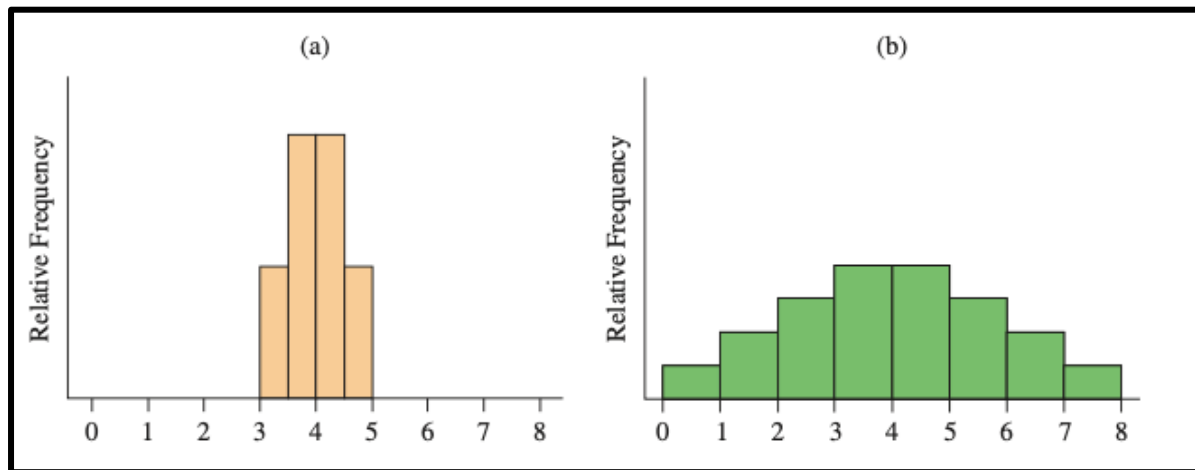
The slide features a dark blue background. On the left side, there are several vertical stripes of varying shades of blue and white. Overlaid on these stripes are several circles of different sizes, also in shades of blue. The largest circle is positioned near the top left, with smaller circles arranged in a descending pattern below it.

2.2 MEASURES OF VARIABILITY

MEASURES OF VARIABILITY

- A measure along the horizontal axis of the data distribution that describes the **spread** of the distribution from the center.

Fig 2.5 Variability or dispersion of data



THE RANGE



- The **range, R** , of a set of n measurements is the difference between the largest and smallest measurements.
- **Example:** A botanist records the number of petals on 5 flowers:

5, 12, 6, 8, 14

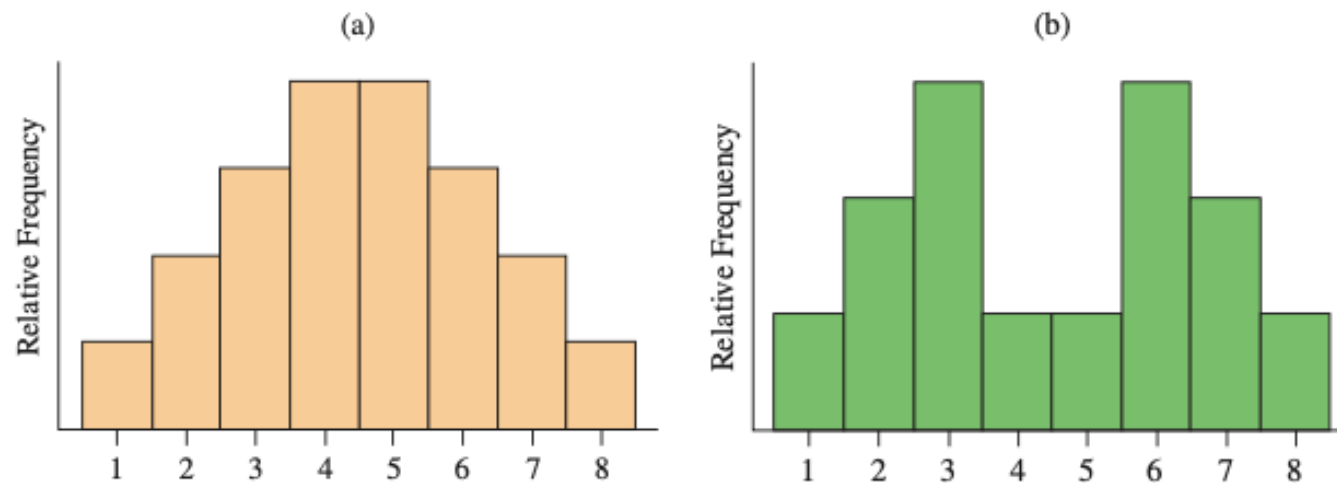
- The range is

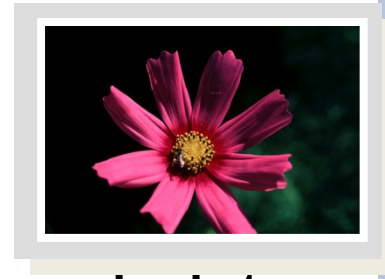
$$R = 14 - 5 = 9.$$

• Quick and easy, but only uses 2 of the 5 measurements.



Fig 2.6
Distributions with equal range and unequal variability

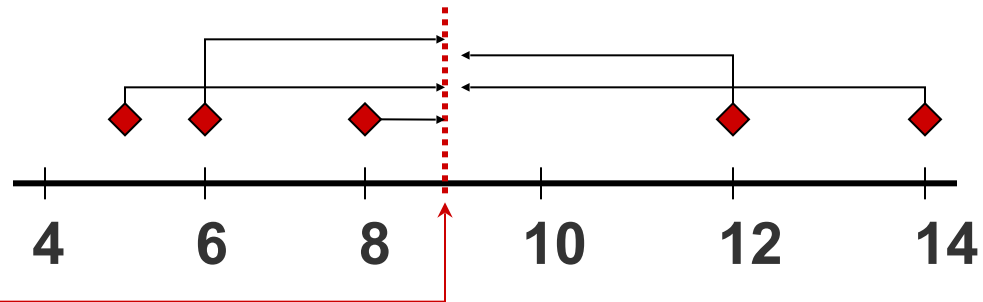




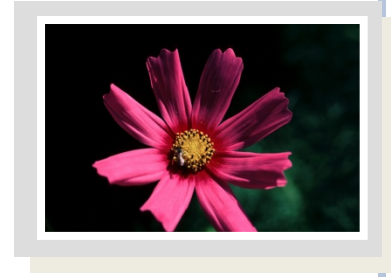
THE VARIANCE

- The **variance** is measure of variability that uses all the measurements. It measures the average deviation of the measurements about their mean.
- **Flower petals:** 5, 12, 6, 8, 14

$$\bar{x} = \frac{45}{5} = 9$$



THE VARIANCE



- The **variance of a population** of N measurements is the average of the squared deviations of the measurements about their mean μ .

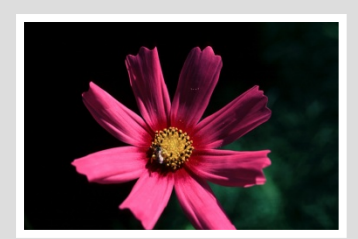
$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

- The **variance of a sample** of n measurements is the sum of the squared deviations of the measurements about their mean, divided by $(n - 1)$.

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$



THE STANDARD DEVIATION



- In calculating the variance, we squared all of the deviations, and in doing so changed the scale of the measurements.
- To return this measure of variability to the original units of measure, we calculate the **standard deviation**, the positive square root of the variance.

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

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



SOME NOTES

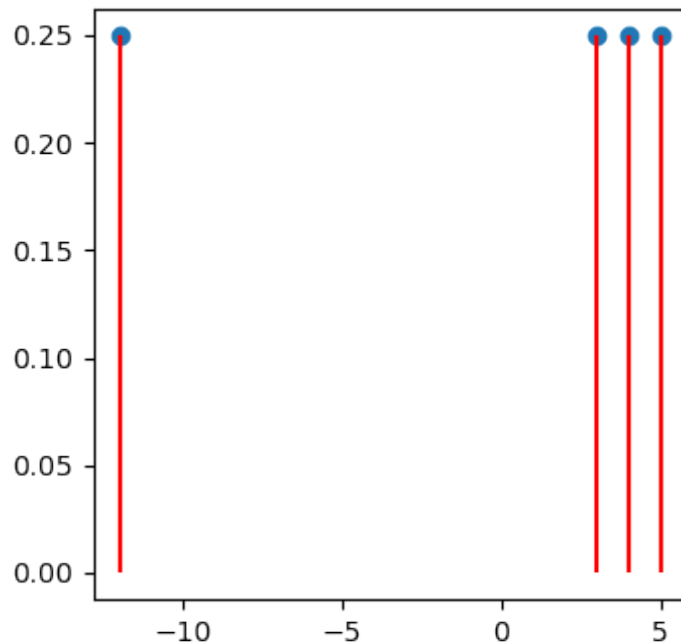
- The value of s is **ALWAYS** positive.
- The larger the value of s^2 or s , the larger the variability of the data set.
- **Why divide by $n - 1$?**
 - The sample standard deviation s is often used to estimate the population standard deviation σ .
 - Dividing by $n - 1$ gives us a better estimate of σ .



ILLUSTRATION: WHY DIVIDED BY $(n - 1)$?

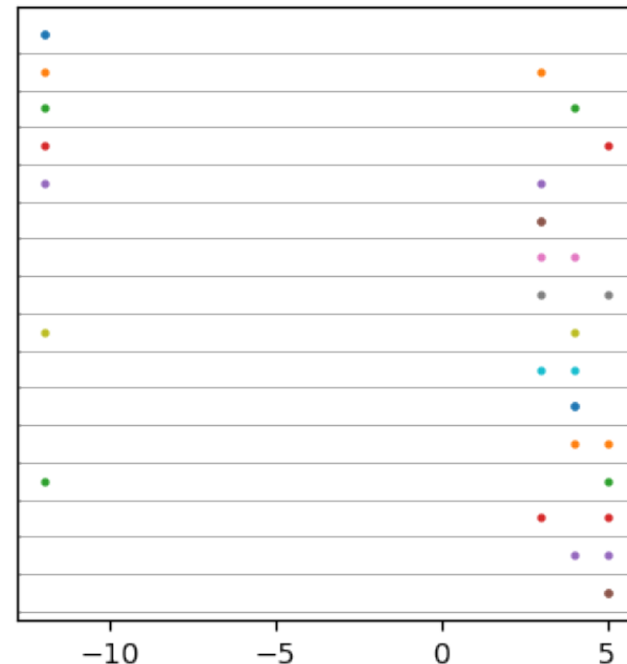
Link Ch 7: Sampling distribution

Population: $\{-12, 3, 4, 5\}$, $N = 4$



$$\mu = 0, \sigma^2 = 48.5$$

Samples of $n = 2$



16 possible outcomes

See 4.3 mn rules

<https://youtu.be/cR0HJYt9u6g>



Which is better?

No	x_1	x_2	\bar{x}	$\frac{\sum (x_i - \bar{x})^2}{n - 1}$	$\frac{\sum (x_i - \bar{x})^2}{n}$
1	-12	-12	-12.0	0.0	0.0
2	-12	3	-4.5	112.5	56.25
3	-12	4	-4.0	128.0	64.0
4	-12	5	-3.5	144.5	72.25
5	3	-12	-4.5	112.5	56.25
6	3	3	3.0	0.0	0.0
7	3	4	3.5	0.5	0.25
8	3	5	4.0	2.0	1.0
9	4	-12	-4.0	128.0	64.0
10	4	3	3.5	0.5	0.25
11	4	4	4.0	0.0	0.0
12	4	5	4.5	0.5	0.25
13	5	-12	-3.5	144.5	72.25
14	5	3	4.0	2.0	1.0
15	5	4	4.5	0.5	0.25
16	5	5	5.0	0.0	0.0
			Average	48.5	24.25

Recall $\sigma^2 = 48.5$



TWO WAYS TO CALCULATE THE SAMPLE VARIANCE



Use the Definition Formula:

	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
	5	-4	16
	12	3	9
	6	-3	9
	8	-1	1
	14	5	25
Sum	45	0	60

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

$$= \frac{60}{4} = 15$$

$$s = \sqrt{s^2} = \sqrt{15} = 3.87$$

TWO WAYS TO CALCULATE THE SAMPLE VARIANCE



Use the Computational Formula:

	x_i	x_i^2
	5	25
	12	144
	6	36
	8	64
	14	196
Sum	45	465

$$s^2 = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n - 1}$$

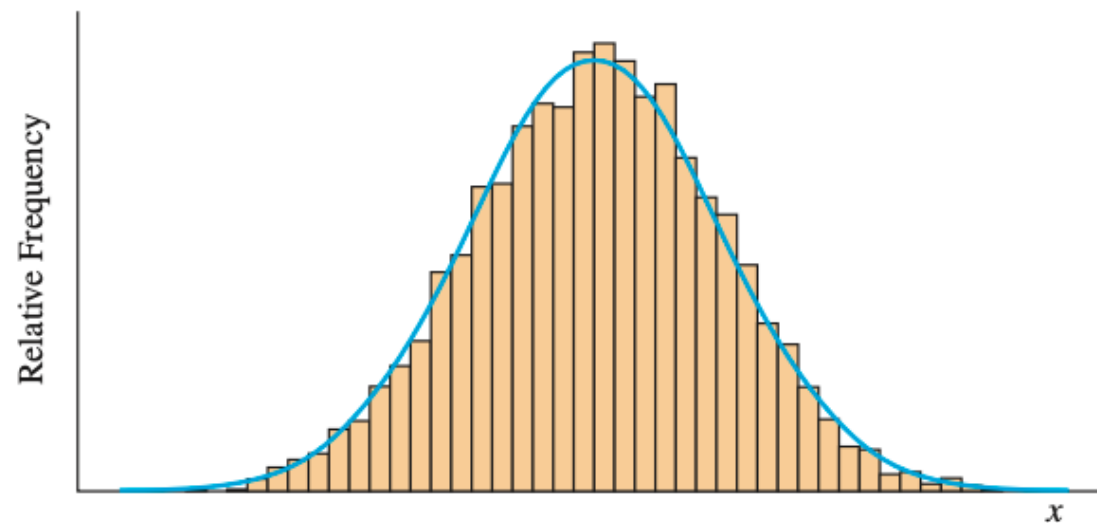
$$= \frac{465 - \frac{45^2}{5}}{4} = 15$$

$$s = \sqrt{s^2} = \sqrt{15} = 3.87$$

The left side of the slide features a series of vertical stripes in various shades of blue, ranging from light to dark. Overlaid on these stripes are several circles of different sizes, also in shades of blue, creating a modern, abstract design.

2.3 UNDERSTANDING AND INTERPRETING THE STANDARD DEVIATION

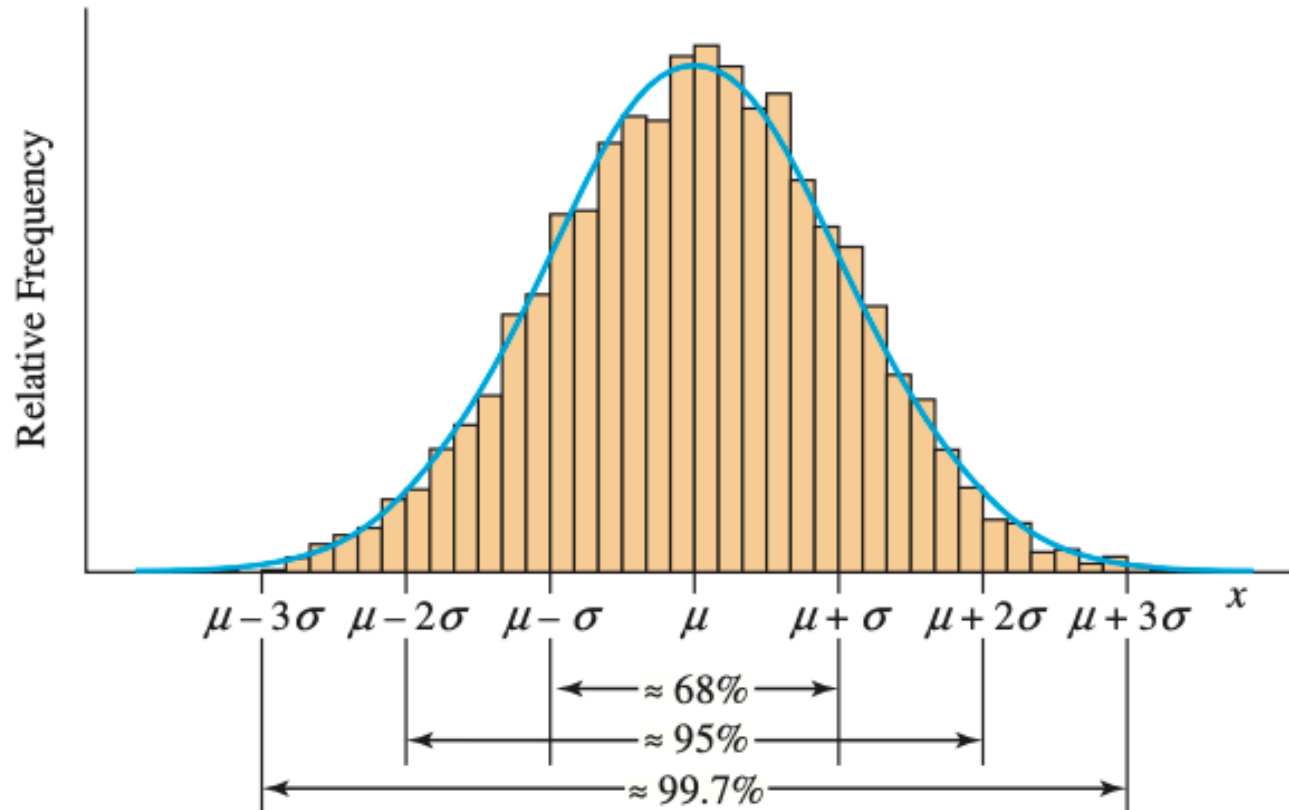
Fig 2.9 Mound-shaped distribution



The **normal distribution** will be discussed in detail in Chapter 6.



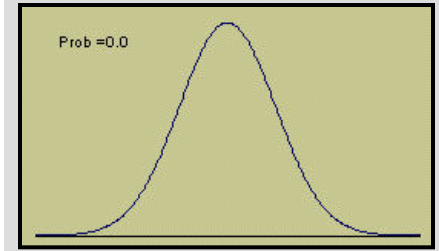
Fig 2.10 Illustrating the Empirical Rule



The Empirical Rule works very well for data that “pile up” in the familiar **mound shape** shown in Figure 2.9*.



USING MEASURES OF **CENTER** AND **SPREAD**: THE EMPIRICAL RULE



Given a distribution of measurements that is approximately mound-shaped:

✓ The interval $\mu \pm \sigma$ contains approximately 68% of the measurements.

✓ The interval $\mu \pm 2\sigma$ contains approximately 95% of the measurements.

✓ The interval $\mu \pm 3\sigma$ contains approximately 99.7% of the measurements.

EXAMPLE

The ages of 50 tenured faculty at a state university.

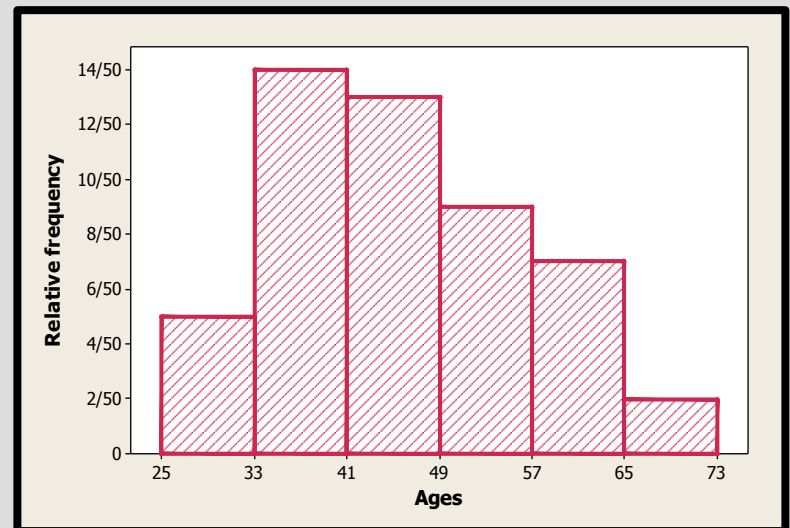


- 34 48 **70** 63 52 52 35 50 37 43 53 43 52 44
- 42 31 36 48 43 **26** 58 62 49 34 48 53 39 45
- 34 59 34 66 40 59 36 41 35 36 62 34 38 28
- 43 50 30 43 32 44 58 53

$$\bar{x} = 44.9$$

$$s = 10.73$$

Shape? **Skewed right**



k	$\bar{x} \pm ks$	Interval	Proportion in Interval	Empirical Rule
1	44.9 \pm 10.73	34.17 to 55.63	31/50 (.62)	\approx .68
2	44.9 \pm 21.46	23.44 to 66.36	49/50 (.98)	\approx .95
3	44.9 \pm 32.19	12.71 to 77.09	50/50 (1.00)	\approx .997

•Do they agree with the Empirical Rule?

•Why or why not?

•No. Not very well.

• The data distribution is not very mound-shaped, but skewed right.

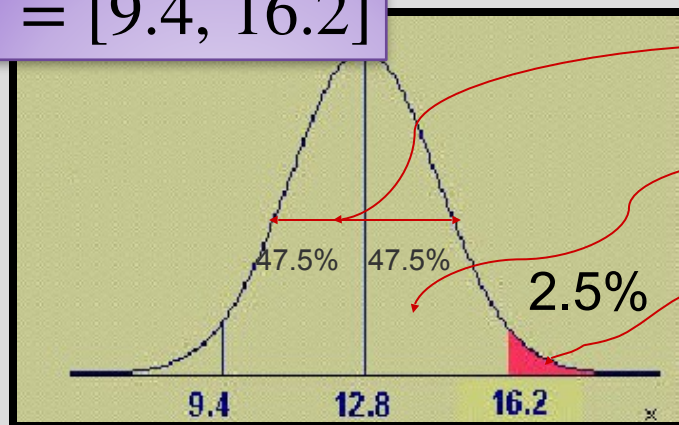


EXAMPLE



The length of time for a worker to complete a specified operation averages 12.8 minutes with a standard deviation of 1.7 minutes. If the distribution of times is approximately mound-shaped, what proportion of workers will take longer than 16.2 minutes to complete the task?

$$\bar{x} \pm 2s = [9.4, 16.2]$$

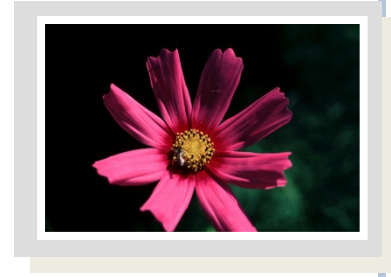


95% between 9.4 and 16.2

47.5% between 12.8 and 16.2

(50-47.5)% = 2.5% above 16.2

APPROXIMATING s



- From the Empirical Rule, we know that
$$R \approx 4-6 s$$
- To approximate the standard deviation of a set of measurements, we can use:

$$s \approx R / 4$$

or $s \approx R / 6$ for a large data set.

■ **Table 2.7 Divisor for the Range Approximation of s**

Number of Measurements	Divide the Range by
5	2.5
10	3
25	4

APPROXIMATING s



The ages of 50 tenured faculty at state university.

- 34 48 **70** 63 52 52 35 50 37 43 53 43 52 44
- 42 31 36 48 43 **26** 58 62 49 34 48 53 39 45
- 34 59 34 66 40 59 36 41 35 36 62 34 38 28
- 43 50 30 43 32 44 58 53

$$R = 70 - 26 = 44$$

$$s \approx R / 4 = 44 / 4 = 11$$

$$\text{Actual } s = 10.73$$



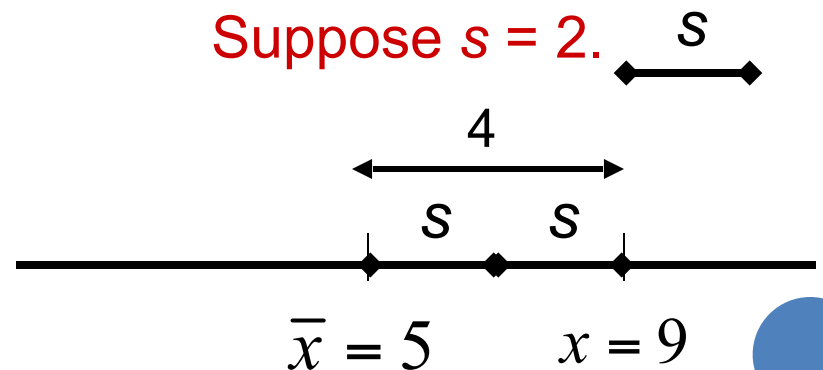
A decorative graphic on the left side of the slide. It features a series of vertical stripes in various shades of blue and white. Overlaid on these stripes are several circles of different sizes, also in shades of blue. The circles are arranged in a way that they appear to be floating or overlapping the stripes.

2.4 MEASURES OF RELATIVE STANDING

MEASURES OF RELATIVE STANDING

- Where does one particular measurement stand in relation to the other measurements in the data set?
- How many standard deviations away from the mean does the measurement lie? This is measured by the **z-score**.

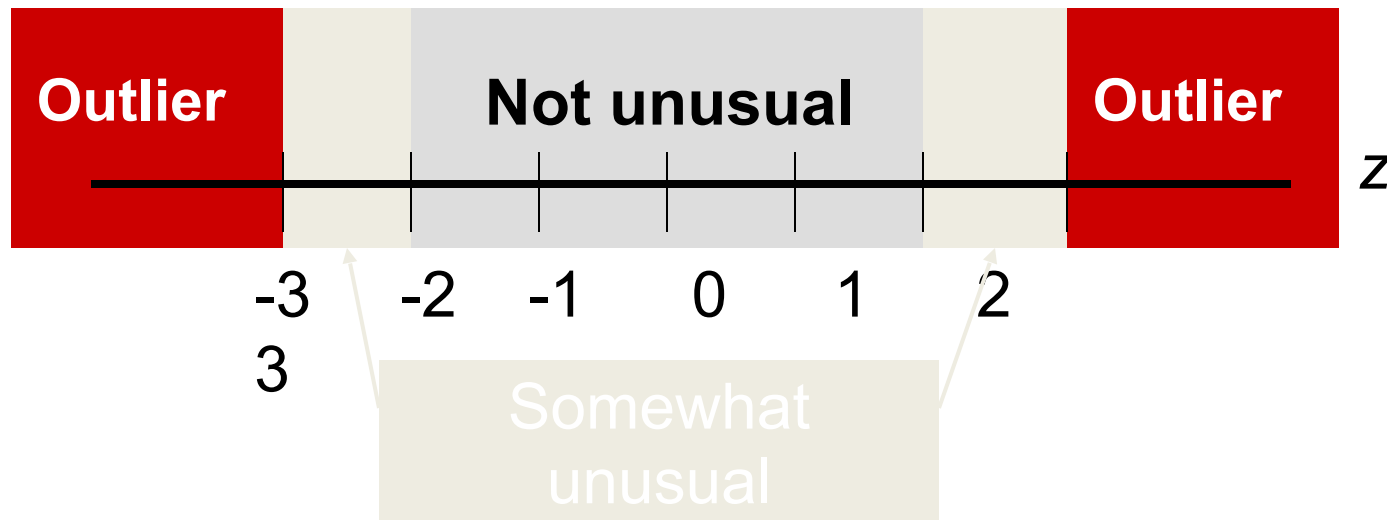
$$z - \text{score} = \frac{x - \bar{x}}{s}$$



$x = 9$ lies $z = 2$ std dev from the mean.

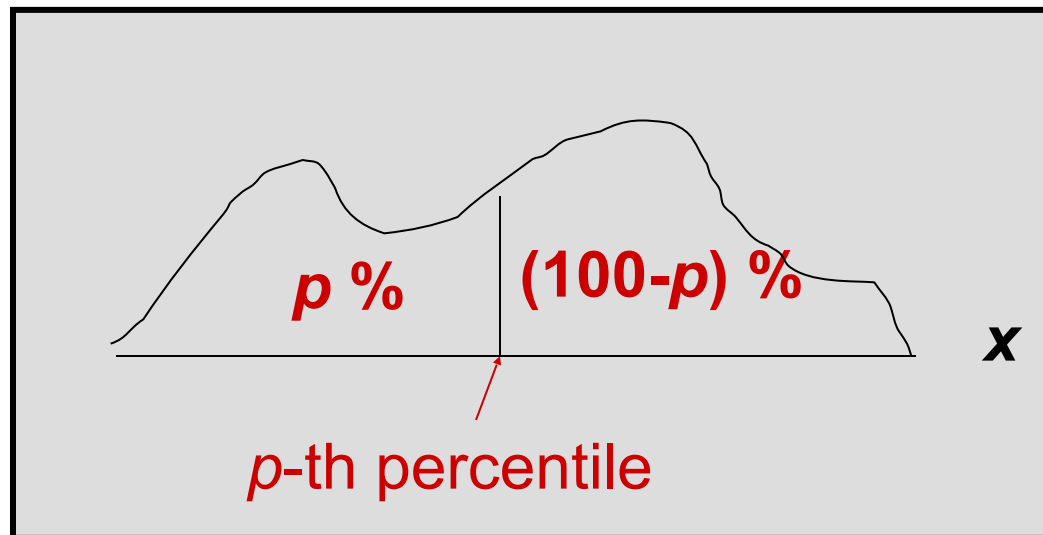
Z-SCORES

- From the Empirical Rule
 - About 99.7% of measurements lie within 3 standard deviations of the mean.
- z -scores between -2 and 2 are not unusual.
- z -scores should not be more than 3 in absolute value. z -scores larger than 3 in absolute value would indicate a possible **outlier**.



MEASURES OF RELATIVE STANDING

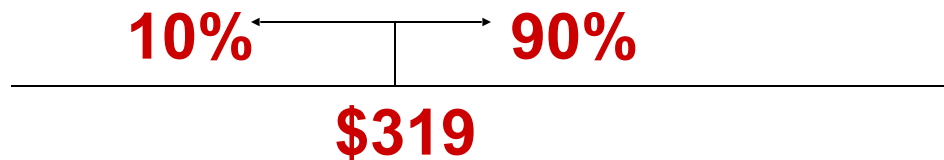
- How many measurements lie below the measurement of interest? This is measured by the p^{th} percentile.



EXAMPLES

- 90% of all men (16 and older) earn more than \$319 per week.

BUREAU OF LABOR STATISTICS



\$319 is the 10th percentile.

50th Percentile \equiv Median

25th Percentile \equiv Lower Quartile (Q_1)

75th Percentile \equiv Upper Quartile (Q_3)

QUARTILES AND THE IQR

- The **lower quartile** (Q_1) is the value of x which is larger than 25% and less than 75% of the ordered measurements.
- The **upper quartile** (Q_3) is the value of x which is larger than 75% and less than 25% of the ordered measurements.
- The range of the “middle 50%” of the measurements is the **interquartile range**,

$$\text{IQR} = Q_3 - Q_1$$



CALCULATING SAMPLE QUARTILES

- The **lower and upper quartiles** (Q_1 and Q_3), can be calculated as follows:
- The **position of Q_1** is $.25(n + 1)$
- The **position of Q_3** is $.75(n + 1)$

once the measurements have been ordered. If the positions are not integers, find the quartiles by interpolation.



EXAMPLE



The prices (\$) of 18 brands of walking shoes:

40. 60 65 65 65 68 68 70 70
70 70 70 70 74 75 75 90 95

$$\text{Position of } Q_1 = .25(18 + 1) = 4.75$$

$$\text{Position of } Q_3 = .75(18 + 1) = 14.25$$

✓ Q_1 is 3/4 of the way between the 4th and 5th ordered measurements, or

$$Q_1 = 65 + .75(65 - 65) = 65.$$

EXAMPLE



The prices (\$) of 18 brands of walking shoes:

40. 60 65 65 65 68 68 70 70
70 70 70 70 74 75 75 90 95

$$\text{Position of } Q_1 = .25(18 + 1) = 4.75$$

$$\text{Position of } Q_3 = .75(18 + 1) = 14.25$$

✓ Q_3 is 1/4 of the way between the 14th and 15th ordered measurements, or

$$Q_3 = 75 + .25(75 - 74) = 74.25$$

✓ and


$$\text{IQR} = Q_3 - Q_1 = 74.25 - 65 = 9.25$$



USING MEASURES OF CENTER AND SPREAD: THE BOX PLOT

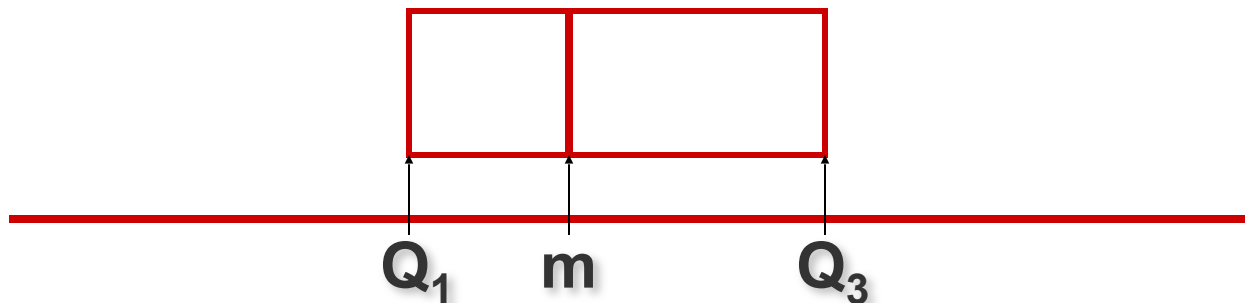
The Five-Number Summary:

Min	Q_1	Median	Q_3	Max
-----	-------	--------	-------	-----

- Divides the data into 4 sets containing an equal number of measurements.
 - A quick summary of the data distribution.
 - Use to form a **box plot** to describe the **shape** of the distribution and to detect **outliers**.
- 

CONSTRUCTING A BOX PLOT

- ✓ Calculate Q_1 , the median, Q_3 and IQR.
- ✓ Draw a horizontal line to represent the scale of measurement.
- ✓ Draw a box using Q_1 , the median, Q_3 .



CONSTRUCTING A BOX PLOT

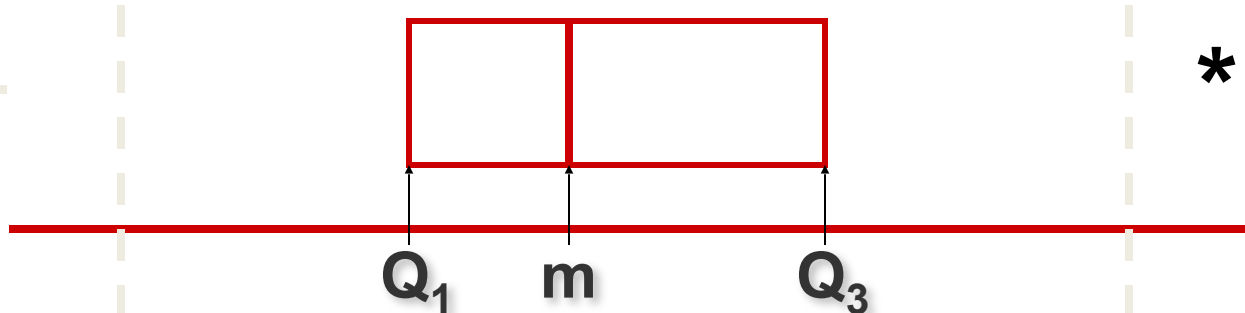
✓ Isolate outliers by calculating

✓ Lower fence: $Q_1 - 1.5 \text{ IQR}$

✓ Upper fence: $Q_3 + 1.5 \text{ IQR}$

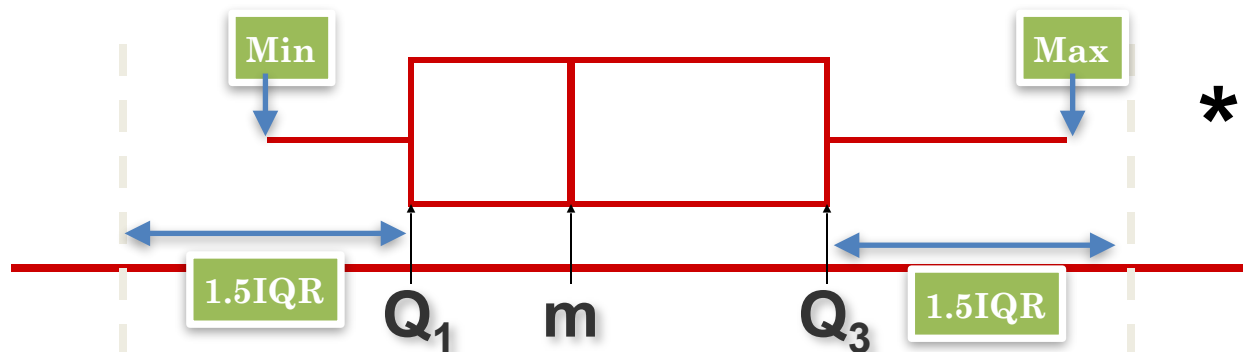
✓ Measurements beyond the upper or lower fence are outliers and are marked

(*)



CONSTRUCTING A BOX PLOT

✓ Draw “**whiskers**” connecting the largest and smallest measurements that are NOT outliers to the box.

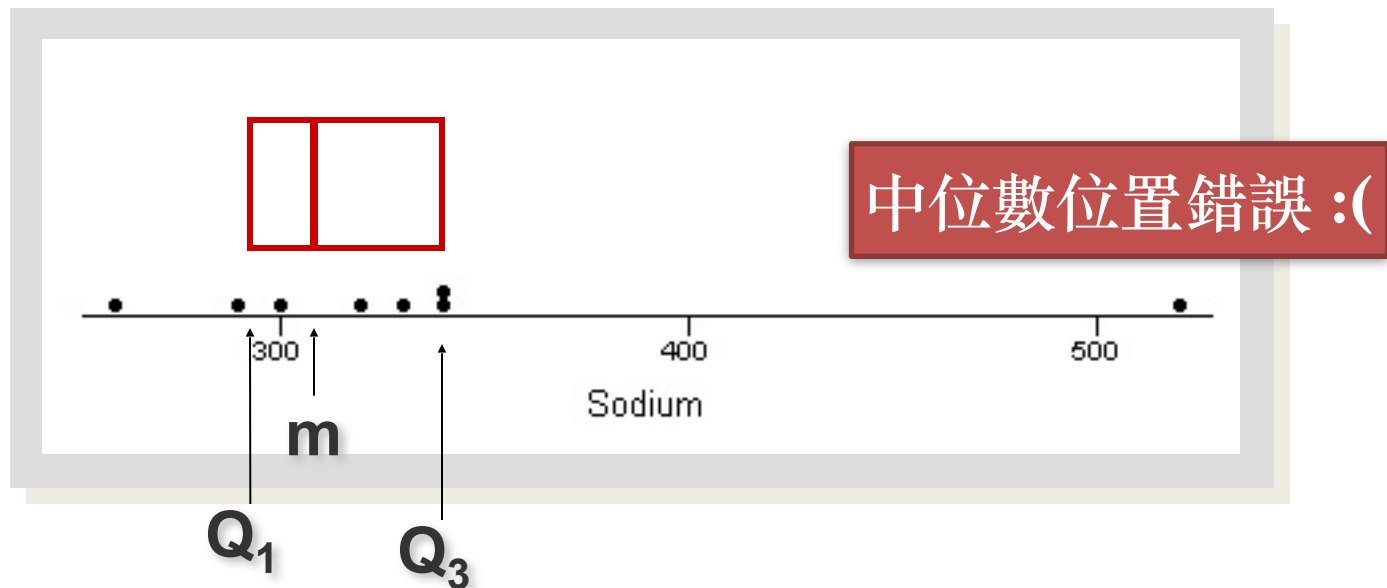
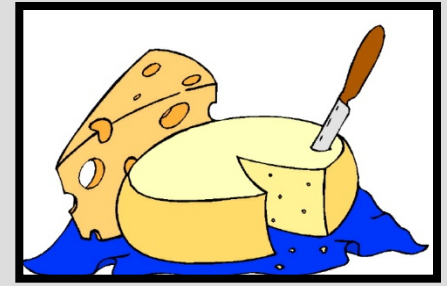


EXAMPLE

Amt of sodium in 8 brands of cheese:

260 290 300 320 330 340 340 520

$$Q_1 = 292.5 \quad m = 325 \quad Q_3 = 340$$



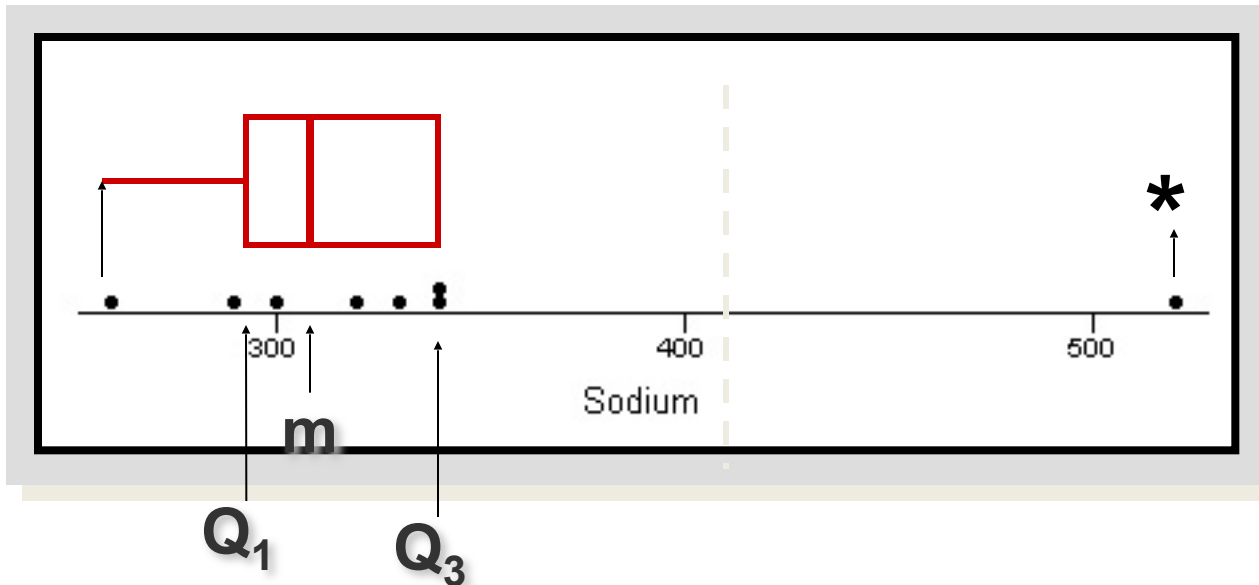
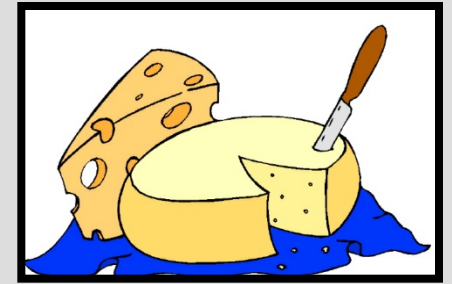
EXAMPLE

$$\text{IQR} = 340 - 292.5 = 47.5$$

$$\text{Lower fence} = 292.5 - 1.5(47.5) = 221.25$$

$$\text{Upper fence} = 340 + 1.5(47.5) = 411.25$$

Outlier: $x = 520$



STEPS IN DROWING A BOXPLOT

1. Calculate the m , Q_1 , Q_3 , and IQR
2. Draw to box enbraced by m , Q_1 , Q_3
3. Draw the lower and upper fence ($Q_1 - 1.5\text{IQR}$, $Q_3 + 1.5\text{IQR}$)
4. Identify outliers if there are any
5. Draw whiskers for the min and max values if they fall between the lower and upper fence



INTERPRETING BOX PLOTS

- ✓ Median line in center of box and whiskers of equal length—symmetric distribution
- ✓ Median line left of center and long right whisker—skewed right
- ✓ Median line right of center and long left whisker—skewed left



KEY CONCEPTS

I. Measures of Center

1. Arithmetic mean (mean) or average

a. Population: μ

b. Sample of size n

$$\bar{x} = \frac{\sum x_i}{n}$$

2. Median: **position** of \bar{x} in = $.5(n + 1)$

3. Mode

4. The median may preferred to the mean if the data are highly skewed.

II. Measures of Variability

1. Range: $R = \text{largest} - \text{smallest}$



KEY CONCEPTS

2. Variance

a. Population of N measurements:

b. Sample of n measurements:

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1}$$

3. Standard deviation

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

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

4. A rough approximation for s can be calculated as $s \approx R/4$.

The divisor can be adjusted depending on the sample size.



KEY CONCEPTS

III. Tchebysheff's Theorem and the Empirical Rule

1. Use Tchebysheff's Theorem for any data set, regardless of its shape or size.
 - a. At least $1-(1/k^2)$ of the measurements lie within k standard deviation of the mean.
 - b. This is only a lower bound; there may be more measurements in the interval.
2. The Empirical Rule can be used only for relatively mound-shaped data sets.
 - Approximately 68%, 95%, and 99.7% of the measurements are within one, two, and three standard deviations of the mean, respectively.



KEY CONCEPTS

IV. Measures of Relative Standing

1. Sample z -score:
2. p th percentile; $p\%$ of the measurements are smaller, and $(100 - p)\%$ are larger.
3. Lower quartile, Q_1 ; **position** of $Q_1 = .25(n + 1)$
4. Upper quartile, Q_3 ; **position** of $Q_3 = .75(n + 1)$
5. Interquartile range: $IQR = Q_3 - Q_1$

V. Box Plots

1. Box plots are used for detecting outliers and shapes of distributions.
2. Q_1 and Q_3 form the ends of the box. The median line is in the interior of the box.



KEY CONCEPTS

3. Upper and lower fences are used to find outliers.

a. **Lower fence:** $Q_1 - 1.5(\text{IQR})$

b. **Upper fence:** $Q_3 + 1.5(\text{IQR})$

4. **Whiskers** are connected to the smallest and largest measurements that are not outliers.

5. Skewed distributions usually have a long whisker in the direction of the skewness, and the median line is drawn away from the direction of the skewness.

