Introduction to Probability and Statistics Twelfth Edition

Chapter 2
Describing Data with Numerical Measures

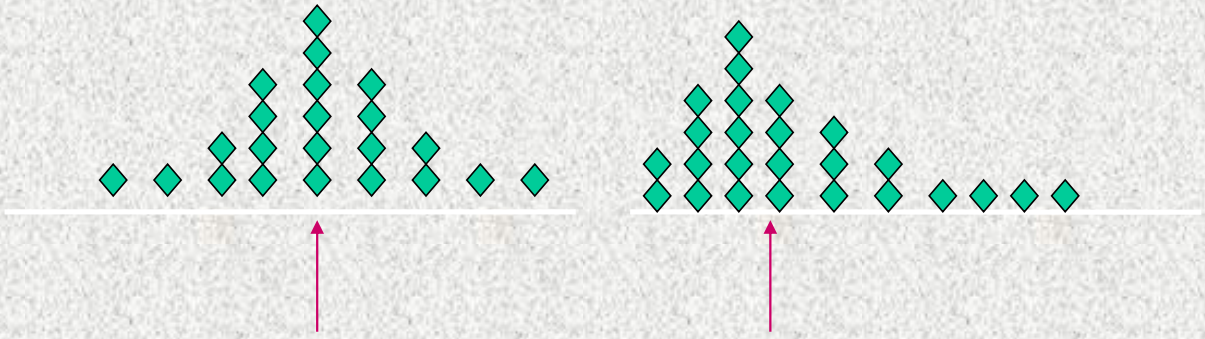
Some graphic screen captures from *Seeing Statistics* ®
Some images © 2001-(current year) www.arttoday.com

Describing Data with Numerical Measures

- **Previous lecture --** 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**.
  - A **statistic** is a numerical descriptive measure calculated for a **sample**.

Measures of Center

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



- The 3 measures are mean, median and mode.

Arithmetic Mean or Average

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

$$\text{Population Mean}: \quad \mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

$$\text{Sample Mean}: \quad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

where n = a sample size
and N = a population size.

Example 1

- A sample consists of 2, 9, 11, 5, and 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 **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 2

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

Median = 4^{th} largest measurement = 5

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

Median = (5 + 6)/2 = 5.5 — average of the 3^{rd} and 4^{th} 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 3

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?**

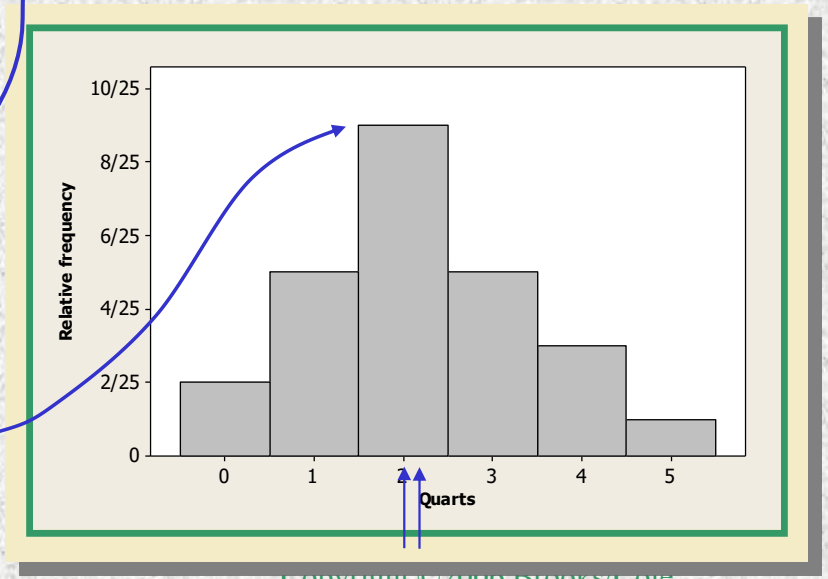
$$\bar{x} = \frac{\sum x_i}{n} = \frac{55}{25} = 2.2$$

- **Median?**

$$m = 2$$

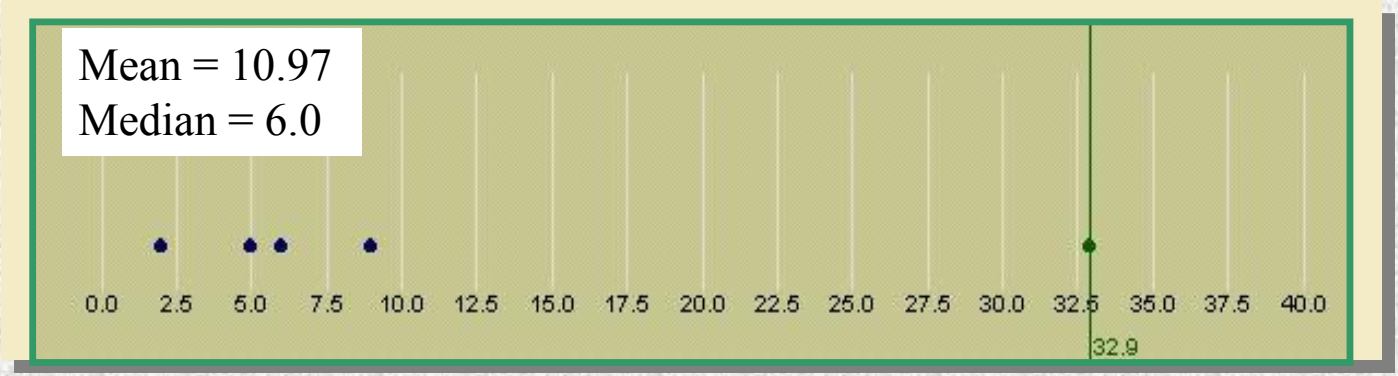
- **Mode? (Highest peak)**

$$\text{mode} = 2$$



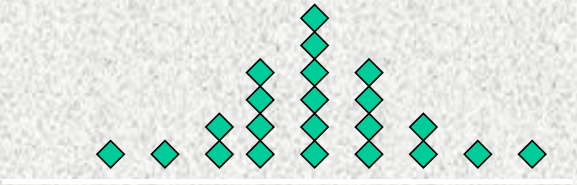
Extreme Values

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

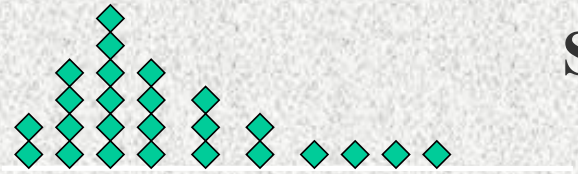


- The median is often used as a measure of center **when the distribution is skewed.**

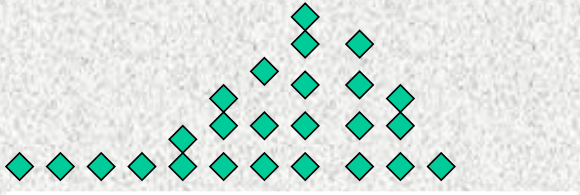
Extreme Values



Symmetric: Mean = Median



Skewed right: Mean > Median



Skewed left: Mean < Median

Class Activity

1. **Time on Task:** In a psychological experiment, the time on task was recorded for 10 subjects under a 5-minute time constraint. These measurements are in seconds: 14 15 16 17 18 19 20 20 20 30

17	19	20	20	14	20	18	15	16	30

   a) Find the average time on task. 18.9

   b) Find the median time on task. 18.5

   c) If you were writing a report to describe these data, which measure of central tendency would you use? Explain.

Measures of Variability

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

The Range

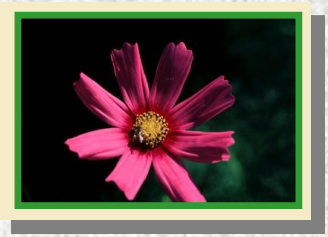
- The **range, R,** of a set of *n* measurements is the difference between the largest and smallest measurements.
- **Example:** A botanist records the number of petals on 5 flowers:

5, 12, 6, 8, 14

- The range is **R = 14 – 5 = 9.**

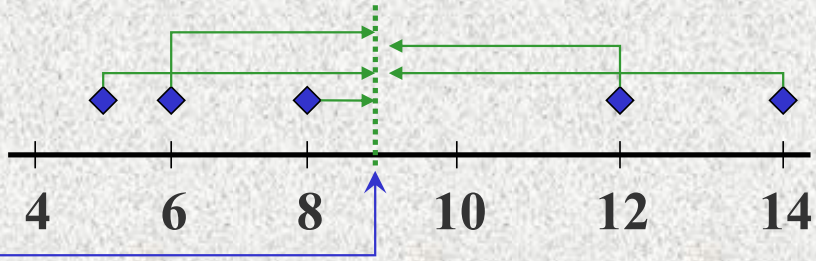
•Quick and easy, but only uses 2 of the 5 measurements.

The Variance

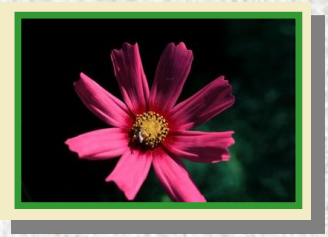


- The **variance** is a measure of variability that uses all the measurements. It measures the average deviation of the measurements about their mean.
- **Flower petals:** **5, 12, 6, 8, 14**

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



The Variance



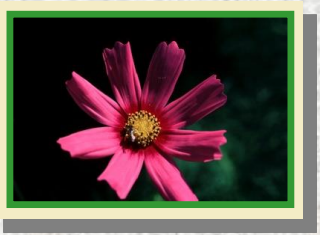
- The **variance of a population** of N measurements is the average of the squared deviations of the measurements about their mean μ.

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

- 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

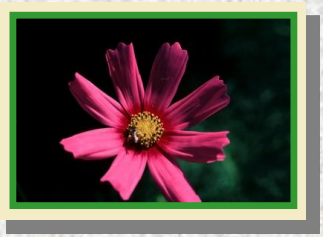


- In calculating the variance, we squared all of the deviations, and in doing so changed the scale of the measurements.
- To return this measure of variability to the original units of measure, we calculate the **standard deviation**, the positive square root of the variance.

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

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

Two Ways to Calculate the Sample Variance



	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

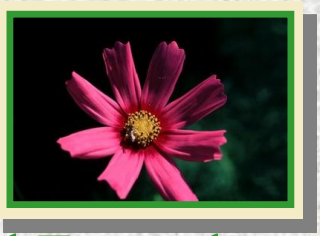
Use the Definition Formula:

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

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

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

Two Ways to Calculate the Sample Variance



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

Use the Calculational Formula:

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

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

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

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 variance s^2 is often used to estimate the population variance σ^2. Dividing by n –1 gives us a better estimate of σ^2 .

Class Activity

2. You are given $n = 6$ measurements:

3	10	5	6	5	1

a) Calculate the range.

b) Calculate the sample mean.

c) Calculate the sample variance and standard deviation.

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.**

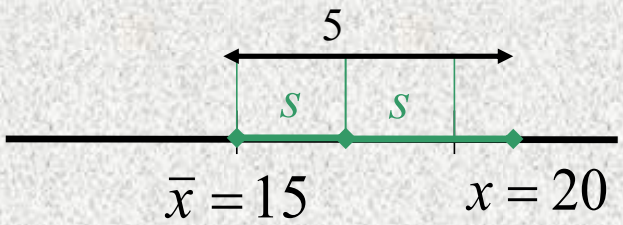
$$z\text{ - score} = \frac{x - \bar{x}}{s}$$

z-Scores

- Suppose that the mean and standard deviation of the quiz scores based on a total of 25 points are 15 and 2, respectively. If a student has a score of 20, what is his ***z*-score?**

$$z\text{-score} = \frac{x - \bar{x}}{s} = \frac{20 - 15}{2} = 2.5$$

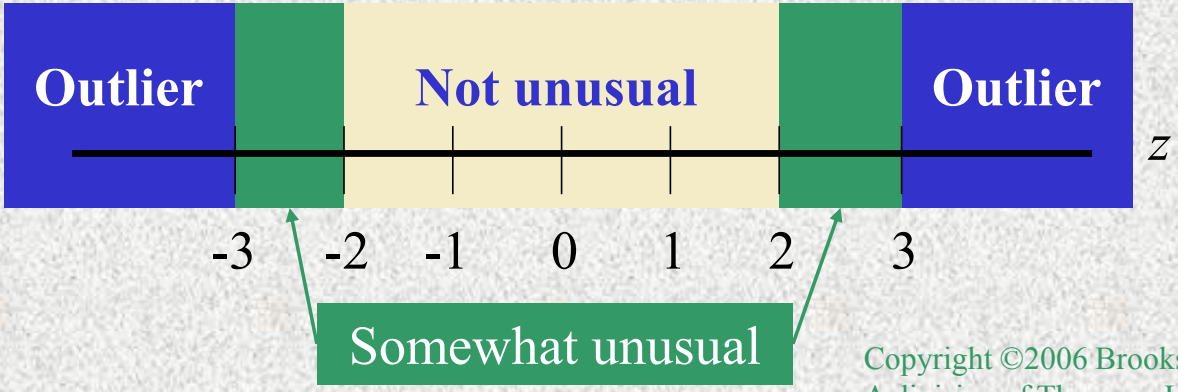
Given $s = 2$.



$x = 20$ lies $z = 2.5$ standard deviations above the mean.

z-Scores

- From the Empirical Rule (*You are not responsible for it.)*
  - About 95% of measurements lie within 2 standard deviations of the mean.
  - 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.**

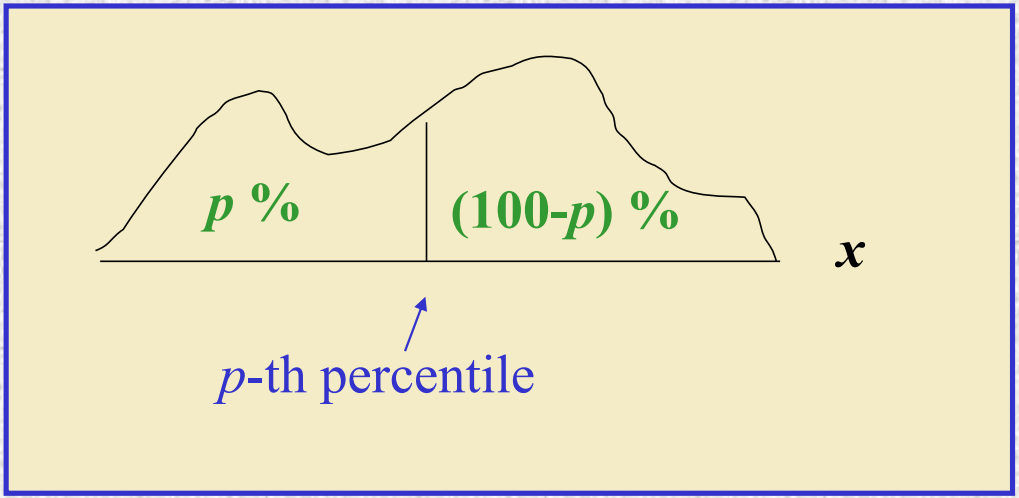


Class Activity

5. A sample of 200 students took a math test. The mean score is 70 points and standard deviation is 4 points. The distribution of test scores is unknown.
   a) Miss A earns 81 points and 90 points for Miss B Compute their *z*-scores. Interpret the results.
   b) Are these two scores outliers?
   c) If Miss C receives a *z*-score of -1.5, then what is her test score?

Measures of Relative Standing

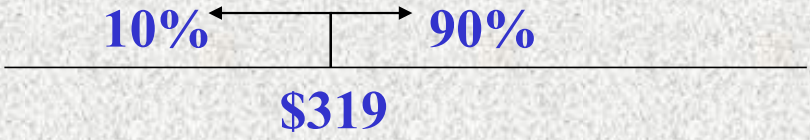
- How many measurements lie below the measurement of interest? This is measured by the **p^{th} percentile, where $0 < p < 100$.**



Example 6

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

BUREAU OF LABOR STATISTICS



$319 is the 10th percentile.

50th Percentile ≡ Median

25th Percentile ≡ Lower Quartile ≡ 1st Quartile (Q_1)

75th Percentile ≡ Upper Quartile ≡ 3rd Quartile (Q_3)

Quartiles and the IQR

- The **lower quartile (first quartile), Q_1,** is the value of variable *x* that is larger than 25% and less than 75% of the ordered measurements.
- The **upper quartile (third quartile), Q_3,** is the value of *x* that 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,**

$$IQR = Q_3 - Q_1$$

Calculating p^{th} percentile

- The ***p*th percentile,** can be calculated as follows:
  - ✓ Firstly, arrange all measurements in ascending order.
  - ✓ Then find the **position of *p*th percentile**

$$\text{position of } p^{th} \text{ percentile} = \frac{p}{100}(n+1)$$

  - ✓ If the position is not an integer, find the ***p*th percentile** by interpolation.

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 7

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

Position of $Q_1 = .25(18 + 1) = 4.75$

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 7

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

Position of $Q_1 = .25(18 + 1) = 4.75$

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

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

✓and

$$IQR = Q_3 - Q_1 = 74.25 - 65 = 9.25$$

Class Activity

6. 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

Find the 40th and 85th percentiles?

$P_{40} = 0.4(19) = 7.6$

$P_{85} = 0.85(19) = 16.15$

Using Measures of Center and Spread: The Box Plot

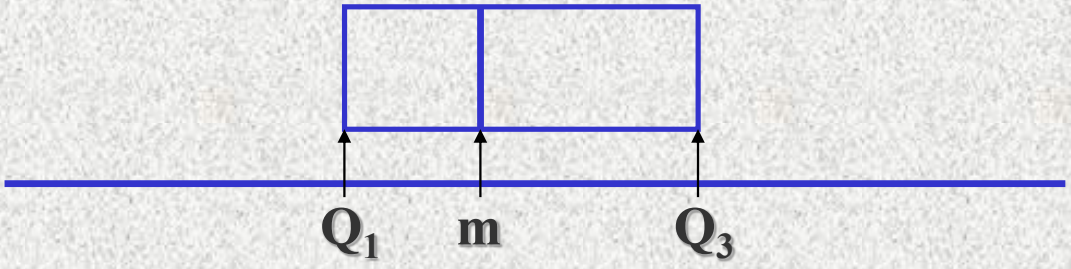
The Five-Number Summary:

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.**

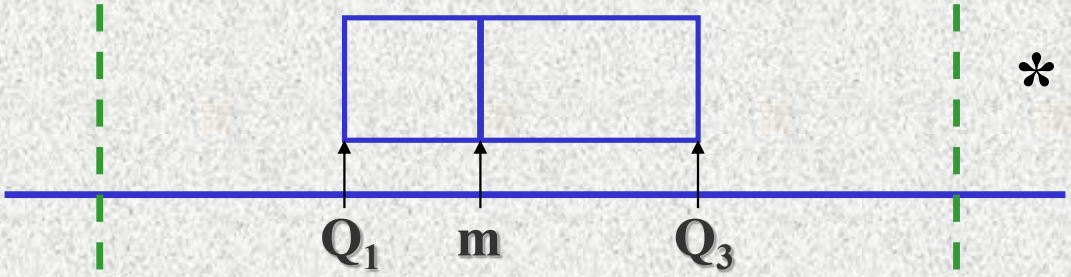
Constructing a Box Plot

- ✓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.



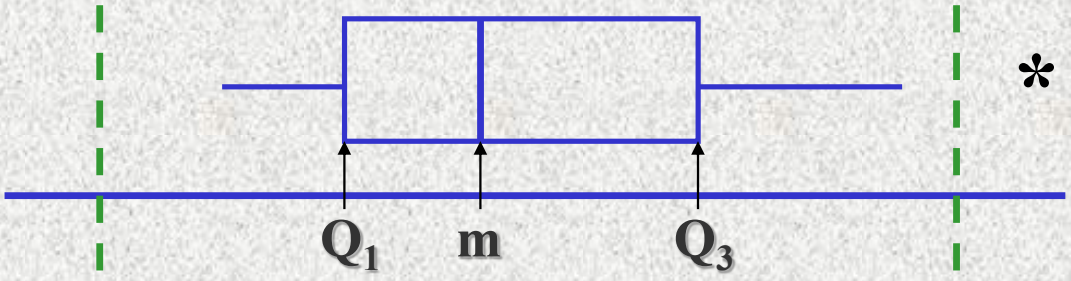
Constructing a Box Plot

- ✓Isolate outliers by calculating
  - ✓Lower fence: Q_1-1.5 IQR
  - ✓Upper fence: Q_3+1.5 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.

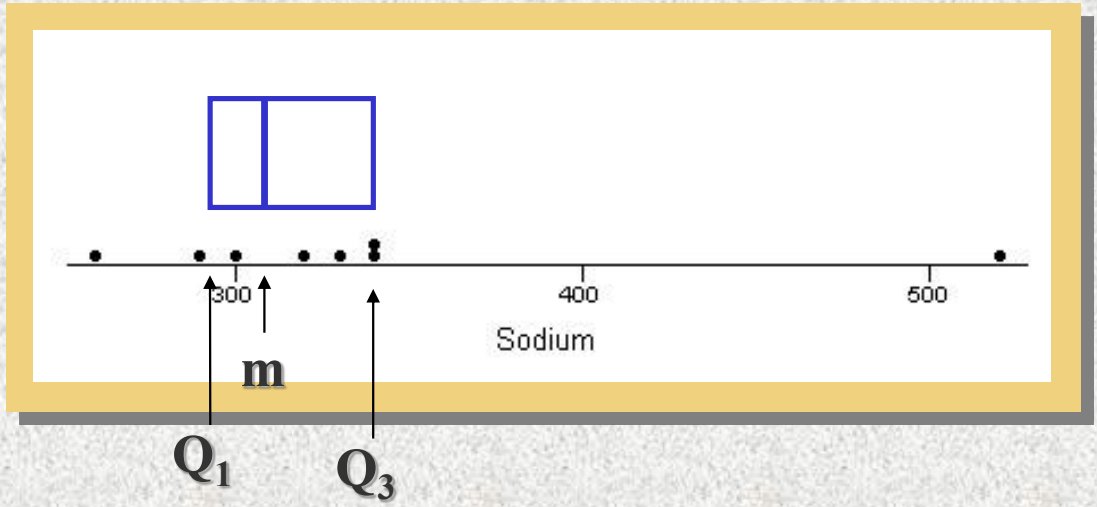


Example 8

The amounts of sodium per slice (in milligrams) for each of 8 brands of cheese:

260 290 300 320 330 340 340 520

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



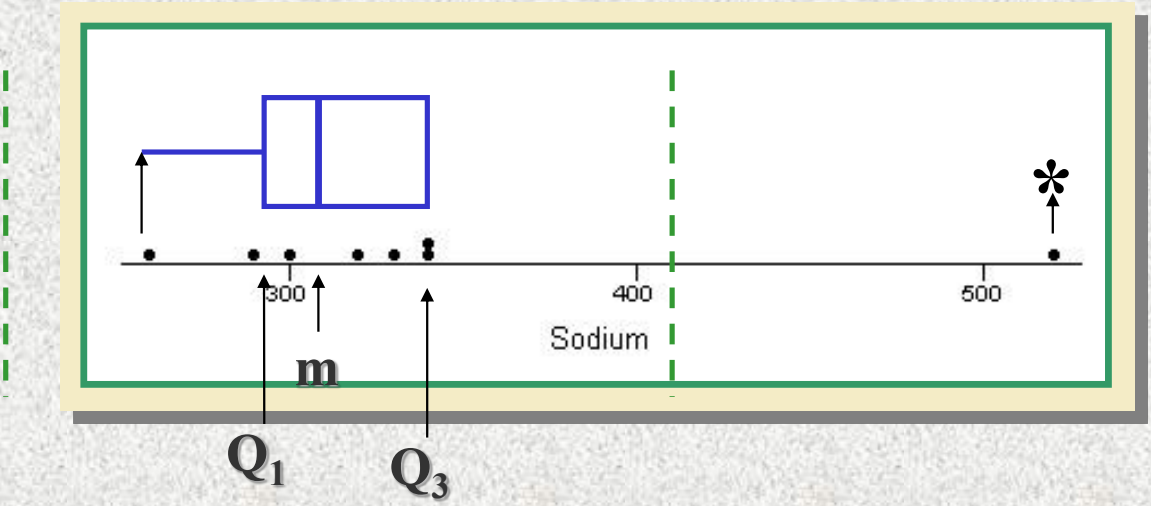
Example 8

IQR = 340-292.5 = 47.5

Lower fence = 292.5-1.5(47.5) = 221.25

Upper fence = 340 + 1.5(47.5) = 411.25

Outlier: $x = 520$



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

Class Activity

7. Construct a box plot for these data and identify any outliers.

   3, 9, 10, 2, 6, 7, 5, 8, 6, 6, 4, 9, 22

Class Activity

8. **Hamburger Meat:** The weight (in pounds) of the 27 packages of ground beef are listed here in order from smallest to largest. It is given that the sample mean is 1.05 and the standard deviation is 0.17.

0.75, 0.83, 0.87, 0.89, 0.89, 0.89, 0.92, 0.93, 0.96

0.96, 0.97, 0.98, 0.99, 1.06, 1.08, 1.08, 1.12, 1.12

1.14, 1.14, 1.17, 1.18, 1.18, 1.24, 1.28, 1.38, 1.41

The two largest packages of meat weight 1.38 and 1.41 pounds. Are these two packages unusually heavy? Explain.

HW, use boxplot, Due 9/5

Find outlier