

AMA 2634

Introduction to Statistics

2024 Fall

Lecture 3

Measure of Central Tendency



A **statistic** is a characteristic or measure obtained by using the data values from a sample.

A **parameter** is a characteristic or measure obtained by using all the data values for a specific population.



What Do We Mean By **Average**?

- Mean
- Median
- Mode
- Midrange
- Weighted Mean

Measures of Central Tendency: Mean



The **mean** is the quotient of the sum of the values and the total number of values.

The symbol $\bar{X}$ is used for sample mean (sample size n).

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{\sum_{i=1}^n X_i}{n}$$

For a population with size N , the Greek letter μ (mu) is used for the mean.

$$\mu = \frac{X_1 + X_2 + \cdots + X_N}{N} = \frac{\sum_{i=1}^N X_i}{N}$$

Example: Days Off per Year



The data represent the number of days off per year for a sample of individuals selected from nine different countries. Find the mean.

20, 26, 40, 36, 23, 42, 35, 24, 30

$$\bar{X} = \frac{20 + 26 + 40 + 36 + 23 + 42 + 35 + 24 + 30}{9} = \frac{276}{9} = 30.7$$

The mean number of days off is 30.7 years.



- The mean for grouped data is calculated by multiplying the frequencies and midpoints of the classes.

$$\bar{X} = \frac{\sum f \cdot X_m}{n}$$

Example: Miles Run



Below is a frequency distribution of miles run per week. Find the mean.

Class Boundaries	Frequency
5.5 - 10.5	1
10.5 - 15.5	2
15.5 - 20.5	3
20.5 - 25.5	5
25.5 - 30.5	4
30.5 - 35.5	3
35.5 - 40.5	2
	$\Sigma f = 20$

Example: Miles Run



Class	Frequency, f	Midpoint, X_m	$f \cdot X_m$
5.5 - 10.5	1	8	8
10.5 - 15.5	2	13	26
15.5 - 20.5	3	18	54
20.5 - 25.5	5	23	115
25.5 - 30.5	4	28	112
30.5 - 35.5	3	33	99
35.5 - 40.5	2	38	76
	$\Sigma f = 20$		$\Sigma f \cdot X_m = 490$

$$\bar{X} = \frac{\sum f \cdot X_m}{n} = \frac{490}{20} = \boxed{24.5 \text{ miles}}$$

Given a set of data $X_1, \dots, X_n$, the sample mean $\bar{X}$ of $X_1, \dots, X_n$ satisfies

$$\bar{X} = \operatorname{argmin}_a \frac{1}{n} \sum_{i=1}^n (X_i - a)^2$$

Equivalently

$$\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \leq \frac{1}{n} \sum_{i=1}^n (X_i - a)^2$$

for any real number a .



The **median** is the midpoint of the data array. The symbol for the median is MD.

The median will be one of the data values if there is an odd number of values.

The median will be the average of two data values if there is an even number of values.

More robust (less affected by outliers) than the mean.

Example: Hotel Rooms



The number of rooms in the seven hotels in downtown Pittsburgh is 713, 300, 618, 595, 311, 401, and 292. Find the median.

Sort in ascending order.

292, 300, 311, 401, 595, 618, 713



Select the middle value.

MD = 401

The median is 401 rooms.

Example: Tornadoes in the U.S.



The number of tornadoes that have occurred in the United States over an 8-year period follows. Find the median.

684, 764, 656, 702, 856, 1133, 1132, 1303

Find the average of the two middle values.

656, 684, 702, 764, 856, 1132, 1133, 1303
 ↑ ↑

$$\text{MD} = \frac{764 + 856}{2} = \frac{1620}{2} = 810$$

The median number of tornadoes is 810.

Given a set of data $X_1, \dots, X_n$, the sample median MD of $X_1, \dots, X_n$ satisfies

$$MD = \operatorname{argmin}_a \frac{1}{n} \sum_{i=1}^n |X_i - a|$$

Equivalently

$$\frac{1}{n} \sum_{i=1}^n |X_i - MD| \leq \frac{1}{n} \sum_{i=1}^n |X_i - a|$$

for any real number a .



The **mode** is the value that occurs most often in a data set.

It is sometimes said to be the most typical case.

There may be no mode, one mode (unimodal), two modes (bimodal), or many modes (multimodal).

Example: NFL Signing Bonuses



Find the mode of the signing bonuses of eight NFL players for a specific year. The bonuses in millions of dollars are

18.0, 14.0, 34.5, 10, 11.3, 10, 12.4, 10

You may find it easier to sort first.

10, 10, 10, 11.3, 12.4, 14.0, 18.0, 34.5



Select the value that occurs the most.

The mode is 10 million dollars.



Find the mode for the number of branches that six banks have.

401, 344, 209, 201, 227, 353

Since each value occurs only once, there is no mode.

Note: Do not say that the mode is zero. That would be incorrect, because in some data, such as temperature, zero can be an actual value.

Example: Licensed Nuclear Reactors



The data show the number of licensed nuclear reactors in the United States for a recent 15-year period. Find the mode.

104 104 104 104 104 107 109 109 109 110
109 111 112 111 109

104 and 109 both occur the most. The data set is said to be bimodal.

The modes are 104 and 109.

Example: Miles Run per Week



Find the modal class for the frequency distribution of miles that 20 runners ran in one week.

Class	Frequency
5.5 – 10.5	1
10.5 – 15.5	2
15.5 – 20.5	3
20.5 – 25.5	5
25.5 – 30.5	4
30.5 – 35.5	3
35.5 – 40.5	2

The modal class is
20.5 – 25.5.

The mode, the midpoint
of the modal class, is
23 miles per week.



The **midrange** is the average of the lowest and highest values in a data set.

$$MR = \frac{Lowest + Highest}{2}$$

Example: Water-Line Breaks



In the last two winter seasons, the city of Brownsville, Minnesota, reported these numbers of water-line breaks per month. Find the midrange.

2, 3, 6, 8, 4, 1

$$\text{MR} = \frac{1+8}{2} = \frac{9}{2} = 4.5$$

The midrange is 4.5.



Find the **weighted mean** of a variable by multiplying each value by its corresponding weight and dividing the sum of the products by the sum of the weights.

$$\bar{X} = \frac{w_1X_1 + w_2X_2 + \cdots + w_nX_n}{w_1 + w_2 + \cdots + w_n} = \frac{\sum_{i=1}^n w_iX_i}{\sum_{i=1}^n w_i}$$

Example: Grade Point Average



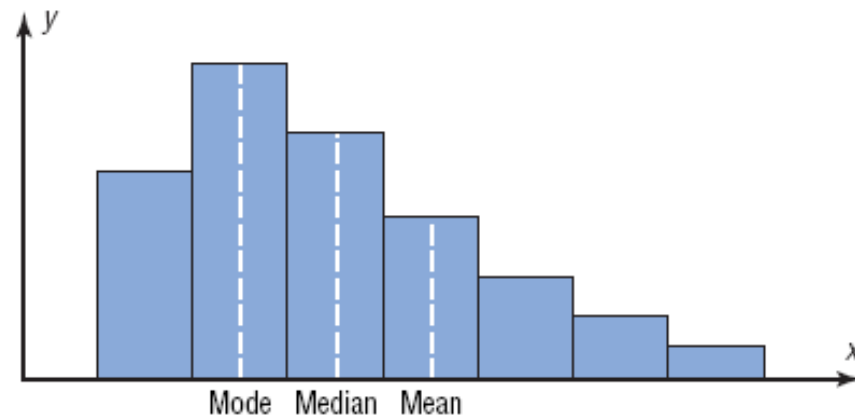
A student received the following grades. Find the corresponding GPA.

Course	Credits, w	Grade, X
English Composition	3	A (4 points)
Introduction to Psychology	3	C (2 points)
Biology	4	B (3 points)
Physical Education	2	D (1 point)

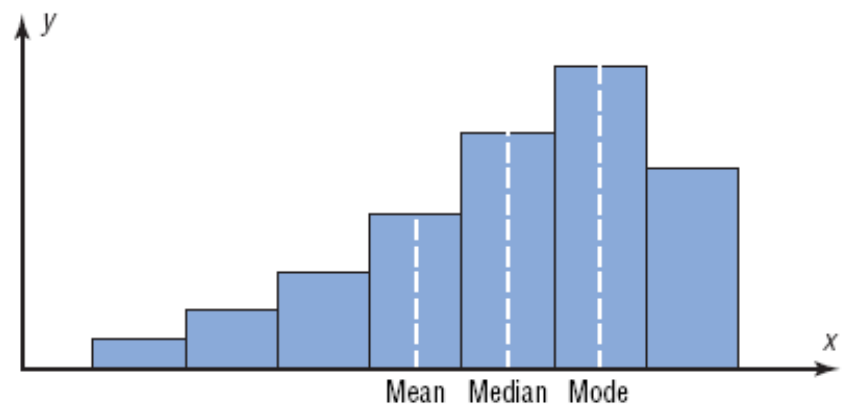
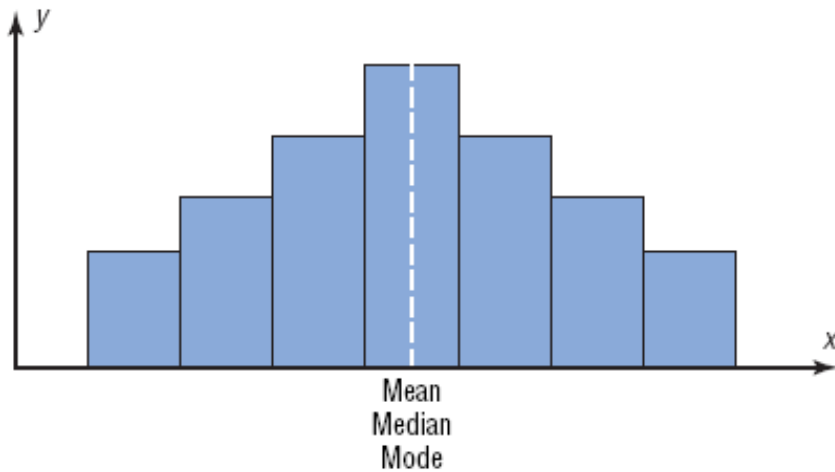
$$\bar{X} = \frac{\sum_{i=1}^4 w_i X_i}{\sum_{i=1}^4 w_i} = \frac{3 \cdot 4 + 3 \cdot 2 + 4 \cdot 3 + 2 \cdot 1}{3 + 3 + 4 + 2} = \frac{32}{12} = 2.7$$

The grade point average is 2.7.

Distributions



(a) Positively skewed or right-skewed



How Can We Measure Variability?

- Range
- Variance
- Standard Deviation
- Coefficient of Variation
- Chebyshev's Theorem
- Empirical Rule (Normal)



- The **range** is the difference between the highest and lowest values in a data set.

$$R = \textit{Highest} - \textit{Lowest}$$

Example: Outdoor Paint



Two experimental brands of outdoor paint are tested to see how long each will last before fading. Six cans of each brand constitute a small population. The results (in months) are shown. Find the mean and range of each group.

Brand A	Brand B
10	35
60	45
50	30
30	35
40	40
20	25

Example: Outdoor Paint



Brand A	Brand B
10	35
60	45
50	30
30	35
40	40
20	25

$$\text{Brand A: } \mu = \frac{\sum X}{N} = \frac{210}{6} = \boxed{35}$$
$$R = 60 - 10 = \boxed{50}$$

$$\text{Brand B: } \mu = \frac{\sum X}{N} = \frac{210}{6} = \boxed{35}$$
$$R = 45 - 25 = \boxed{20}$$

The average for both brands is the same, but the range for Brand A is much greater than the range for Brand B.

Which brand would you buy?



The **variance** is the average of the squares of the distance each value is from the mean.

The **standard deviation** is the square root of the variance.

The standard deviation is a measure of how spread out your data are.



To determine the spread of the data.

To determine the consistency of a variable.

To determine the number of data values that fall within a specified interval in a distribution (Chebyshev's Theorem).

Used in inferential statistics.



- The **population variance** is

$$\sigma^2 = \frac{\sum_{i=1}^N (X_i - \mu)^2}{N}$$

- The **population standard deviation** is

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (X_i - \mu)^2}{N}}$$

Example: Outdoor Paint



Find the variance and standard deviation for the data set for Brand A paint. 10, 60, 50, 30, 40, 20

Months, X	μ	$X - \mu$	$(X - \mu)^2$
10	35	-25	625
60	35	25	625
50	35	15	225
30	35	-5	25
40	35	5	25
20	35	-15	225
			1750

$$\sigma^2 = \frac{\sum (X - \mu)^2}{n}$$

$$= \frac{1750}{6}$$

$$= \boxed{291.7}$$

$$\sigma = \sqrt{\frac{1750}{6}}$$

$$= \boxed{17.1}$$



- The **sample variance** is

$$s^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$$

- The **sample standard deviation** is

$$s = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}}$$

Example: European Auto Sales



Find the variance and standard deviation for the amount of European auto sales for a sample of 6 years. The data are in millions of dollars.

11.2, 11.9, 12.0, 12.8, 13.4, 14.3

X	X^2
11.2	125.44
11.9	141.61
12.0	144.00
12.8	163.84
13.4	179.56
14.3	204.49
<u>75.6</u>	<u>958.94</u>

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

$$s^2 = \frac{6(958.94) - (75.6)^2}{6(5)}$$

$$\begin{aligned} s^2 &= 1.28 \\ s &= 1.13 \end{aligned}$$

$$s^2 = (6 \cdot 958.94 - 75.6^2) / (6 \cdot 5)$$



The **coefficient of variation** is the standard deviation divided by the mean, expressed as a percentage.

$$CVAR = \frac{s}{\bar{X}} \cdot 100\%$$

Use *CVAR* to compare standard deviations when the units are different.

The mean of the number of sales of cars over a 3-month period is 87, and the standard deviation is 5. The mean of the commissions is \$5225, and the standard deviation is \$773. Compare the variations of the two.

$$CVar = \frac{5}{87} \cdot 100\% = 5.7\% \quad \text{Sales}$$

$$CVar = \frac{773}{5225} \cdot 100\% = 14.8\% \quad \text{Commissions}$$

Commissions vary more than sales.



The **Range Rule of Thumb** approximates the standard deviation as

$$s \approx \frac{\textit{Range}}{4}$$

when the distribution is unimodal and approximately symmetric.



Use $\bar{X} - 2s$ to approximate the lowest value and $\bar{X} + 2s$ to approximate the highest value in a data set.

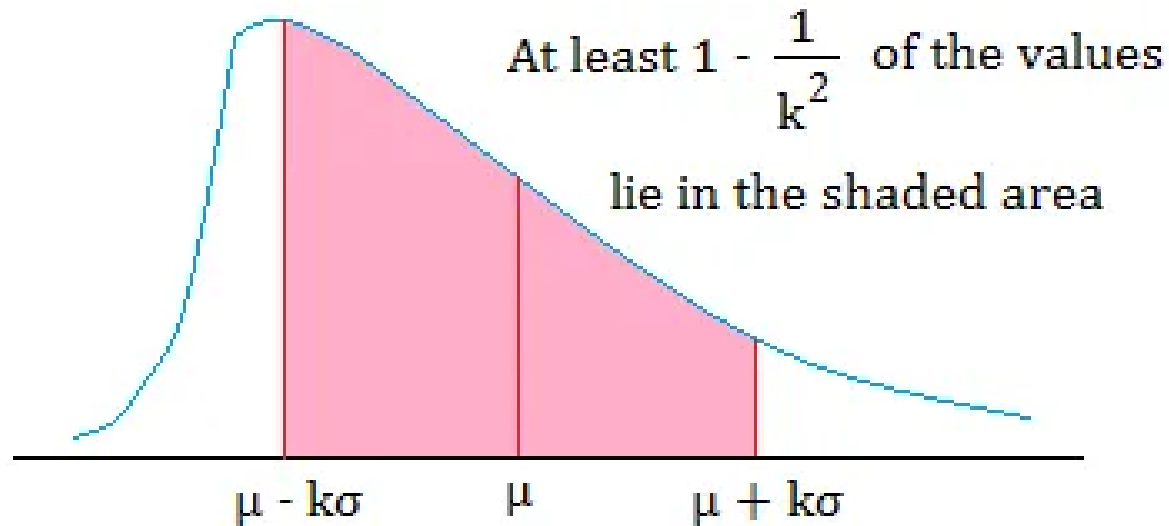
Example: $\bar{X} = 10$, $Range = 12$

$$s \approx \frac{12}{4} = 3 \quad \begin{array}{l} LOW \approx 10 - 2(3) = \boxed{4} \\ HIGH \approx 10 + 2(3) = \boxed{16} \end{array}$$



The proportion of values from any data set that fall within k standard deviations of the mean will be at least $1 - 1/k^2$, where k is a number greater than 1 (k is not necessarily an integer).

# of standard deviations, k	Minimum Proportion within k standard deviations	Minimum Percentage within k standard deviations
2	$1 - 1/4 = 3/4$	75%
3	$1 - 1/9 = 8/9$	88.89%
4	$1 - 1/16 = 15/16$	93.75%



Chebyshev's theorem



Markov's inequality states that for any real-valued random variable Y and any positive number a , we have

$$\Pr(|Y| \geq a) \leq \mathbb{E}[|Y|]/a$$

One way to prove Chebyshev's inequality is to apply Markov's inequality

$$\Pr(|X - \mu| \geq k\sigma) = \Pr((X - \mu)^2 \geq k^2 \sigma^2) \leq \frac{\mathbb{E}[(X - \mu)^2]}{k^2 \sigma^2} = \frac{\sigma^2}{k^2 \sigma^2} = \frac{1}{k^2}.$$

Example: Prices of Homes



The mean price of houses in a certain neighborhood is \$50,000, and the standard deviation is \$10,000. Find the price range for which at least 75% of the houses will sell.

Chebyshev's Theorem states that at least 75% of a data set will fall within 2 standard deviations of the mean.

$$50,000 - 2(10,000) = 30,000$$

$$50,000 + 2(10,000) = 70,000$$

At least 75% of all homes sold in the area will have a price range from \$30,000 and \$70,000.

Example: Travel Allowances



A survey of local companies found that the mean amount of travel allowance for executives was \$0.25 per mile. The standard deviation was 0.02. Using Chebyshev's theorem, find the minimum percentage of the data values that will fall between \$0.20 and \$0.30.

$$\begin{aligned} (.30 - .25) / .02 &= 2.5 \\ (.25 - .20) / .02 &= 2.5 \\ k &= 2.5 \end{aligned}$$

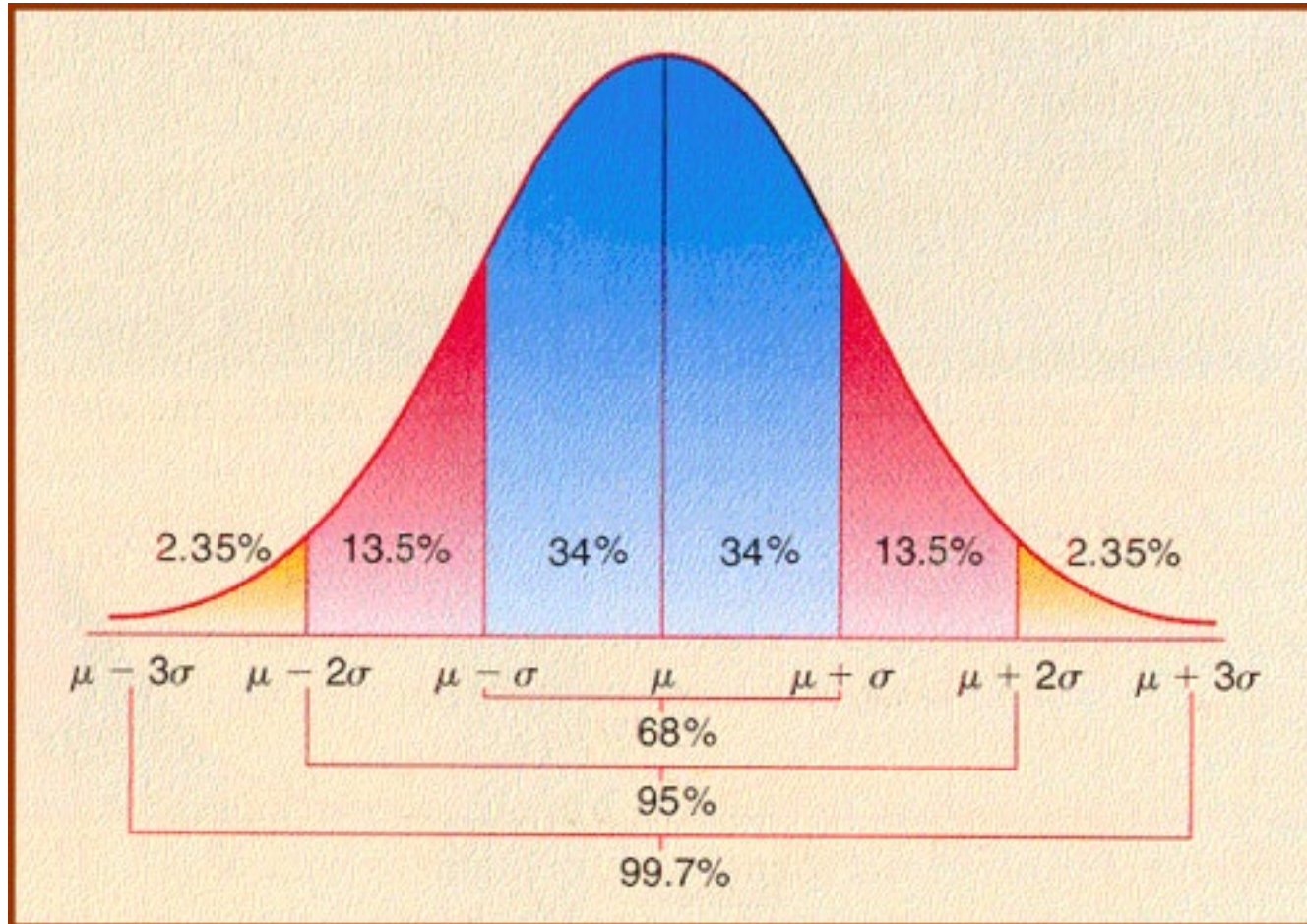
$$\begin{aligned} 1 - 1/k^2 &= 1 - 1/2.5^2 \\ &= 0.84 \end{aligned}$$

At least 84% of the data values will fall between \$0.20 and \$0.30.



The percentage of values from a data set that fall within k standard deviations of the mean in a normal (bell-shaped) distribution is listed below.

# of standard deviations, k	Proportion within k standard deviations
1	68%
2	95%
3	99.7%





z-score

Percentile

Quartile

Outlier



A **z-score** or **standard score** for a value is obtained by subtracting the mean from the value and dividing the result by the standard deviation.

$$z = \frac{X - \mu}{\sigma}$$

A z-score represents the number of standard deviations a value is above or below the mean.

Example: Test Scores



A student scored 65 on a calculus test that had a mean of 50 and a standard deviation of 10; she scored 30 on a history test with a mean of 25 and a standard deviation of 5. Compare her relative positions on the two tests.

$$z = \frac{X - \mu}{\sigma} = \frac{65 - 50}{10} = 1.5 \quad \text{Calculus}$$

$$z = \frac{X - \mu}{\sigma} = \frac{30 - 25}{5} = 1.0 \quad \text{History}$$

She has a higher relative position in the Calculus class.



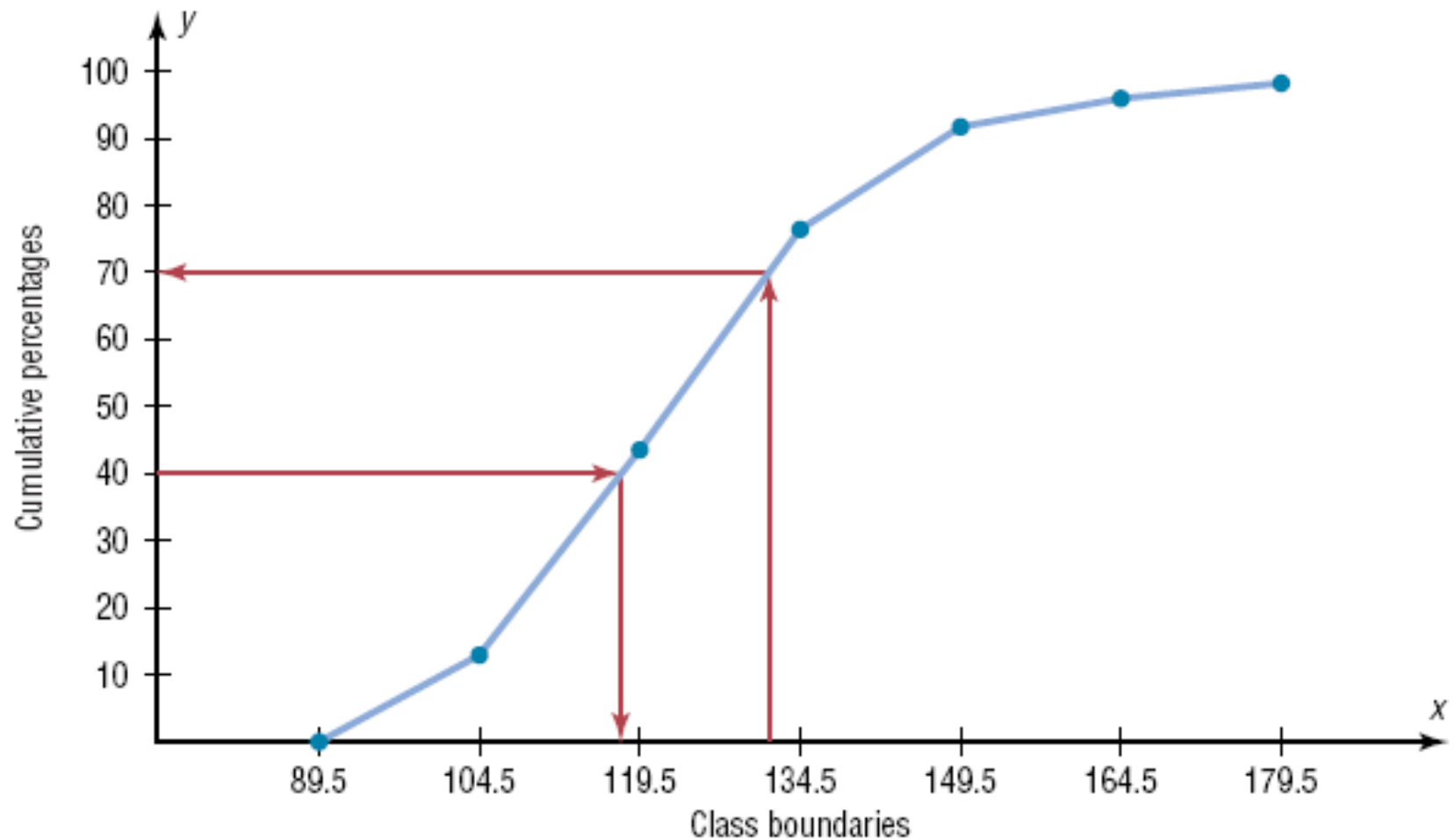
Percentiles separate the data set into 100 equal groups.

A percentile rank for a datum represents the percentage of data values below the datum.

$$\textit{Percentile} = \frac{(\# \text{ of values below } X) + 0.5}{\text{total \# of values}} \cdot 100\%$$

$$c = \frac{n \cdot p}{100}$$

(if c is not a whole number, round it up to the next whole number and c is then the position of the ranked data (from small to large) corresponding to the p th percentile)



Example: Test Scores

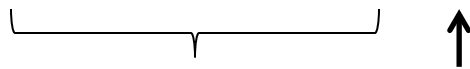


A teacher gives a 20-point test to 10 students. Find the percentile rank of a score of 12.

18, 15, 12, 6, 8, 2, 3, 5, 20, 10

Sort in ascending order.

2, 3, 5, 6, 8, 10, 12, 15, 18, 20


6 values

$$\text{Percentile} = \frac{(\# \text{ of values below } X) + 0.5}{\text{total \# of values}} \cdot 100\%$$

$$= \frac{6 + 0.5}{10} \cdot 100\%$$

$$= 65\%$$

A student whose score was 12 did better than 65% of the class.

Example: Test Scores



A teacher gives a 20-point test to 10 students. Find the value corresponding to the 25th percentile.

18, 15, 12, 6, 8, 2, 3, 5, 20, 10

Sort in ascending order.

2, 3, 5, 6, 8, 10, 12, 15, 18, 20



$$c = \frac{n \cdot p}{100} = \frac{10 \cdot 25}{100} = 2.5 \approx 3$$

The value 5 corresponds to the 25th percentile.



Deciles separate the data set into 10 equal groups. $D_1 = P_{10}$, $D_4 = P_{40}$

Quartiles separate the data set into 4 equal groups. $Q_1 = P_{25}$, $Q_2 = MD$, $Q_3 = P_{75}$

The **Interquartile Range**, $IQR = Q_3 - Q_1$.

- Finding Data Values Corresponding to Q_1 , Q_2 , and Q_3

Step 1 Arrange the data in order from lowest to highest.

Step 2 Find the median of the data values. This is the value for Q_2 .

Step 3 Find the median of the data values that fall below Q_2 . This is the value for Q_1 .

Step 4 Find the median of the data values that fall above Q_2 . This is the value for Q_3 .

Example: Quartiles



Find Q_1 , Q_2 , and Q_3 for the data set.

15, 13, 6, 5, 12, 50, 22, 18

Sort in ascending order.

5, 6, 12, 13, 15, 18, 22, 50

↑ ↑ ↑

$$Q_2 = \frac{13 + 15}{2} = \boxed{14}$$

$$Q_1 = \frac{6 + 12}{2} = \boxed{9}$$

$$Q_3 = \frac{18 + 22}{2} = \boxed{20}$$



An **outlier** is an extremely high or low data value when compared with the rest of the data values.

A data value less than $Q_1 - 1.5(IQR)$ or greater than $Q_3 + 1.5(IQR)$ can be considered an outlier.

The **Five-Number Summary** is composed of the following numbers:

Low, Q_1 , MD, Q_3 , High

The Five-Number Summary can be graphically represented using a **Boxplot**.

Constructing Boxplots

1. Find the five-number summary.
2. Draw a horizontal axis with a scale that includes the maximum and minimum data values.
3. Draw a box with vertical sides through Q_1 and Q_3 , *and draw a vertical line* through the median.
4. Draw a line from the minimum data value to the left side of the box and a line from the maximum data value to the right side of the box.

Example: Meteorites



The number of meteorites found in 10 U.S. states is shown. Construct a boxplot for the data.

89, 47, 164, 296, 30, 215, 138, 78, 48, 39

30, 39, 47, 48, 78, 89, 138, 164, 215, 296

↑ ↑ ↑ ↑ ↑
Low Q_1 MD Q_3 High

Five-Number Summary: 30-47-83.5-164-296

