

DATA: LIST OF n NUMBERS/MEASUREMENTS

$x_1, x_2, x_3, \dots, x_n$

EXAMPLE: Table 3 Entrance Examination Scores of 100 Entering Freshmen

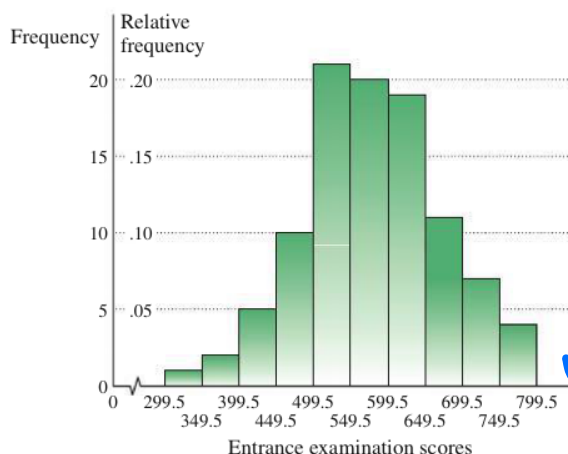
762	451	602	440	570	553	367	520	454	653
433	508	520	603	532	673	480	592	565	662
712	415	595	580	643	542	470	743	608	503
566	493	635	780	537	622	463	613	502	577
618	581	644	605	588	695	517	537	552	682
340	537	370	745	605	673	487	412	613	470
548	627	576	637	787	507	566	628	676	750
442	591	735	523	518	612	589	648	662	512
663	588	627	584	672	533	738	455	512	622
544	462	730	576	588	705	695	541	537	563

OUR GOAL IS TO EXTRACT USEFUL INFORMATION FROM THE DATA.

(DESCRIBE / SUMMARIZE)

TWO METHODS: I VISUALLY
II NUMERICALLY

I. VISUALLY: FREQUENCY / RELATIVE FREQUENCY HISTOGRAM



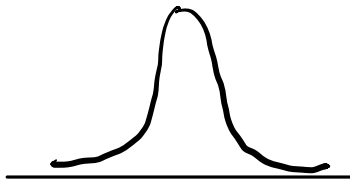
"CATEGORIES" = INTERVALS.

HISTOGRAM:
X-AXIS IN
LINE

(VS. BAR GRAPH)

NOW WE CAN SEE THE DISTRIBUTION OF MEASUREMENTS.

WORDS TO DESCRIBE DISTRIBUTIONS:

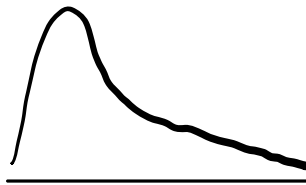


SYMMETRIC

e.g. SAT SCORES

HEIGHTS OF
ADULT MALES

WEIGHTS OF APPLES
PRODUCED BY 1 TREE

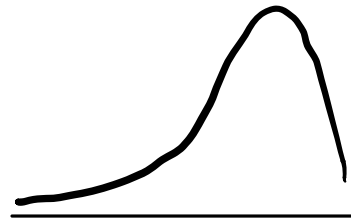


RIGHT-SKEWED

e.g. HOUSEHOLD
INCOME

NUMBER OF
SIBLINGS

SIZE OF NYC APTS



LEFT-SKEWED

e.g. LIFETIME OF HUMANS

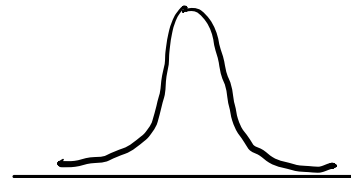


UNIFORM

e.g. # TIMES EACH FACE APPEARS
ON 1000 ROLLS OF A DIE.

TIMES EACH DIGIT APPEARS IN
FIRST 10 MILLION DIGITS OF π

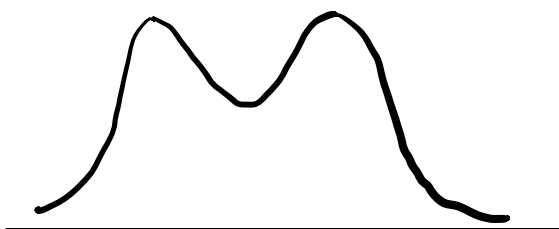
NORMAL NUMBERS



UNIMODAL

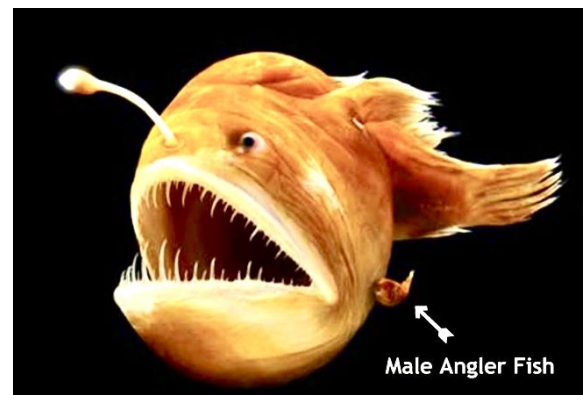
e.g. RESTING HEART RATE FOR
HUMANS

PSI AT WHICH A PARTICULAR TYPE
OF BICYCLE WHEEL TUBE POPS



BIMODAL

e.g. SIZE OF ANGLER FISH



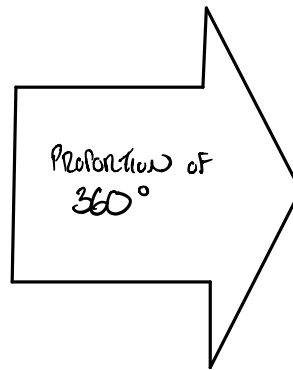
Pie Charts

PERCENTS

Type A 35%

Type B 18%

Type C 47%

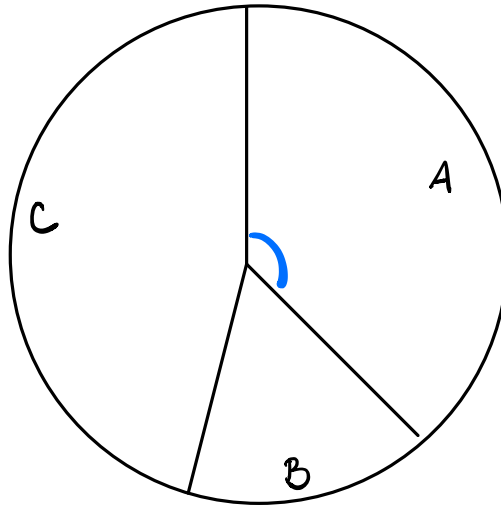


CENTRAL ANGLE

$$.35 \times 360 = 126^\circ$$

$$.18 \times 360 = 64.8^\circ$$

$$.47 \times 360 = 169.2^\circ$$



CENTRAL ANGLE

$$= \text{Proportion} \times 360^\circ$$

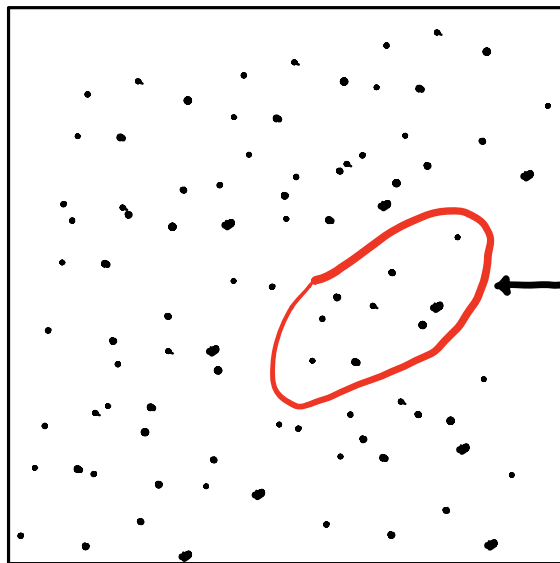
DATA VISUALIZATION

REDDIT.COM/r/DATA IS BEAUTIFUL

II. NUMERICALLY : MEASURES OF CENTER & VARIATION

DATA: $x_1, x_2, x_3, \dots, x_n$

2 SOURCES:



POPULATION DATA ← WANT TO STUDY

DESIRABLE
HARD TO ACQUIRE

SAMPLE DATA

NOT IDEAL
EASIER TO ACQUIRE

WE CAN USE SAMPLE
DATA TO MAKE ESTIMATES
FOR THE POPULATION
(INFERENTIAL STATISTICS)

3 MEASURES OF CENTER

x-bar

1. SAMPLE MEAN $\bar{x}$

POPULATION MEAN μ

MU

$$\left. \begin{array}{l} \text{SAMPLE MEAN } \bar{x} \\ \text{POPULATION MEAN } \mu \end{array} \right\} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n}$$

SIGMA NOTATION

END VALUE
FOR INDEX

eg.

"S" FOR
SUM
(+)

$$\sum_{i=1}^5 \frac{1}{2^i} = \frac{1}{2^1} + \frac{1}{2^2} + \frac{1}{2^3} + \frac{1}{2^4} + \frac{1}{2^5} = \frac{63}{64}$$

INDEX
START
VALUE

TEMPLATE FOR OBJECTS
TO ADD

INDEX INCREASES BY 1

The mean of 4 numbers is 90. If the mean of the first three numbers is 88, find the fourth number.

Wrong: $\frac{88 + 92}{2} = 90 \Rightarrow 4^{\text{th}} \# 92$ 

$$\frac{x_1 + x_2 + x_3 + x_4}{4} = 90$$

$$\frac{x_1 + x_2 + x_3}{3} = 88$$

$$x_1 + x_2 + x_3 = 3 \cdot 88 = 264$$

$$\frac{264 + x_4}{4} = 90$$

$$264 + x_4 = 4 \cdot 90 = 360$$

$$x_4 = 96$$

Example. Suppose I buy 20 gallons of gas at an average price of \$2.40/gallon, and you buy 10 gallons of gas at an average price of \$2.10/gallon. Together, what is the average price per gallon that we've paid for gas?

WEIGHTED AVERAGES:

GIVEN n #'S THE AVERAGE (MEAN) IS

$$\frac{x_1 + x_2 + x_3 + \dots + x_n}{n} = \frac{x_1}{n} + \frac{x_2}{n} + \frac{x_3}{n} + \dots + \frac{x_n}{n}$$

$$= \frac{1}{n}x_1 + \frac{1}{n}x_2 + \frac{1}{n}x_3 + \dots + \frac{1}{n}x_n$$

ADD UP TO 1. ALL THE SAME.

MORE GENERALLY, A **WEIGHTED AVERAGE OF n #'S** IS

$$\alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \dots + \alpha_n x_n$$

ADD UP TO 1. ALL POSITIVE.

Example. Suppose you have a homework grade of 90, and quiz grade of 85, and an exam grade of 80. If homework counts for 20% of your grade, the quiz counts for 35% of your grade, and the exam counts for 45% of your grade, calculate your average for the class.

$$.2 \times 90 + .35 \times 85 + .45 \times 80 = 83.75$$

Example. It costs a shipping company \$8.75 to ship a small package overnight. Suppose the shipping company charges a flat rate of \$19 to ship a small package overnight, and if the package is late they refund the full amount. If 93% of all packages are delivered on time, what is the average profit per package that the shipping company earns?

Two Possible Profits: $+19 - 8.75 = 10.25$ ON TIME

$+19 - 8.75 - 19 = -8.75$ LATE

100 PACKAGES:
$$\frac{93(10.25) + 7(-8.75)}{100}$$

$$.93(10.25) + .07(-8.75) = 8.92$$

\$8.92



ACTUARIAL SCIENCE

2. MEDIAN

Arrange the data in order from least to greatest.

If the number of measurements is odd, then the median is the number in the middle position.

e.g. 1 1 3 3 4
 ↑
 MEDIAN

If the number of measurements is even, then the median is the mean of the two numbers that share/straddle the middle position.

e.g. 1 1 3 5 5 6
 └───┘
 $\frac{3+5}{2} = 4$
 ↑
 MEDIAN

Note: AN EQUAL NUMBER OF MEASUREMENT LIE TO THE RIGHT/LEFT OF THE MEDIAN WHEN ARRANGED IN ORDER.

"HALF THE DATA IS LESS THAN THE MEDIAN, &
HALF THE DATA IS GREATER THAN THE MEDIAN."

GOOD WHEN YOU DON'T WANT THE "CENTER" TO BE INFLUENCED BY EXTREME VALUES

35,000 42,000 46,000 48,000 59,000 63,000

MEAN 48,833

MEDIAN 47,000 (HALF ABOVE, HALF BELOW)

35,000 42,000 46,000 48,000 59,000 63,000,000

MEAN 10,538,333

MEDIAN 47,000 (HALF ABOVE, HALF BELOW)

Blood cholesterol levels. Find the mean and median for the data in the following table.

Blood Cholesterol Levels (milligrams per deciliter)		
Midpoint	Interval	Frequency
159.5	149.5–169.5	4
179.5	169.5–189.5	11
199.5	189.5–209.5	15
219.5	209.5–229.5	25
239.5	229.5–249.5	13
259.5	249.5–269.5	7
279.5	269.5–289.5	3
299.5	289.5–309.5	2
(Total 80)		

WE ASSUME ALL MEASUREMENTS IN EACH CLASS/BIN ARE EQUAL TO THE MIDPOINT OF THAT CLASS/BIN'S INTERVAL

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n} \sum m_k f_k$$

MIDPOINT OF EACH CLASS
×
FREQUENCY OF EACH CLASS

$$= \frac{1}{80} \left(4(159.5) + 11(179.5) + 15(199.5) + 25(219.5) + 13(239.5) + 7(259.5) + 3(279.5) + 2(299.5) \right)$$

$$= \frac{4}{80}(159.5) + \frac{11}{80}(179.5) + \frac{15}{80}(199.5) + \frac{25}{80}(219.5) + \frac{13}{80}(239.5) + \frac{7}{80}(259.5) + \frac{3}{80}(279.5) + \frac{2}{80}(299.5)$$

RELATIVE FREQUENCY OF MIDPOINTS
 $\left(\frac{\text{FREQUENCY}}{n} \right)$

3. Mode : Most frequently occurring measurement(s).

Yes, there may be multiple modes (ties allowed).

Blood cholesterol levels. Find the mean and median for the data in the following table.

Blood Cholesterol Levels (milligrams per deciliter)		
Midpoint	Interval	Frequency
159.5	149.5–169.5	4
179.5	169.5–189.5	11
199.5	189.5–209.5	15
219.5	209.5–229.5	25
239.5	229.5–249.5	13
259.5	249.5–269.5	7
279.5	269.5–289.5	3
299.5	289.5–309.5	2

MODAL CLASS

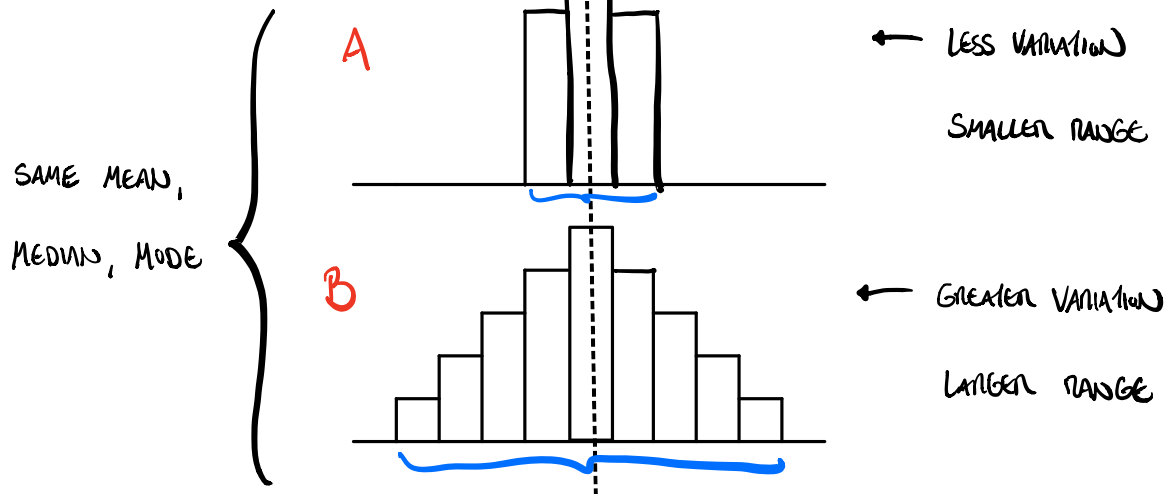
[209.5, 229.5]

MODE

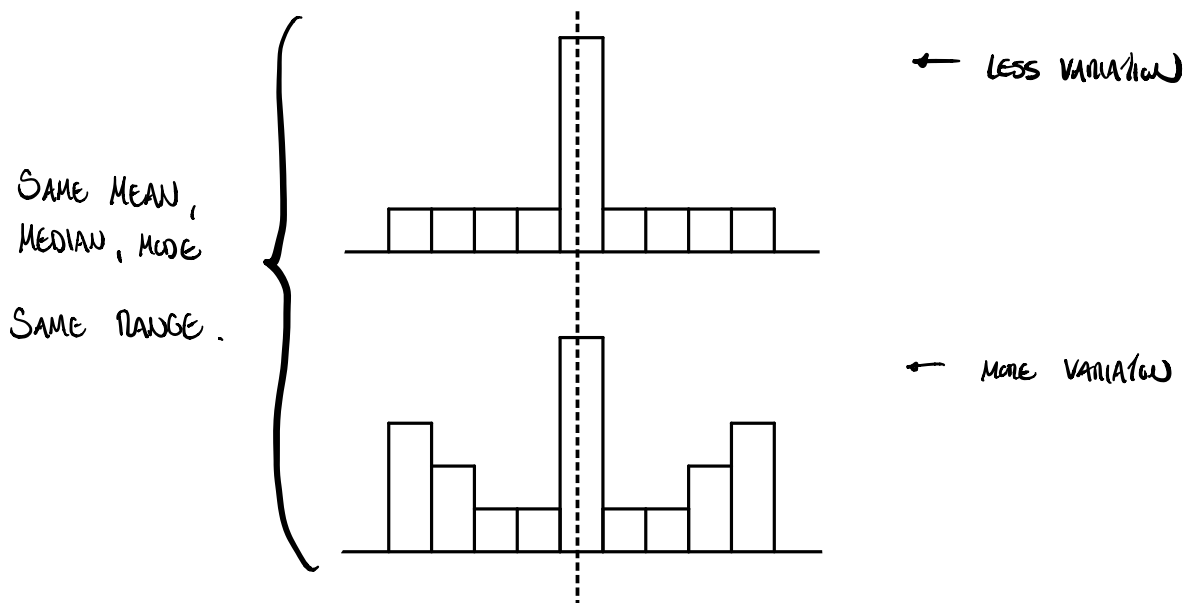
3 MEASURES OF VARIATION

SIZE OF FISH IN 2 LAKES

1. RANGE = MAX - MIN



WHAT ABOUT THE FOLLOWING DISTRIBUTIONS:



WE CAN MEASURE THE DIFFERENCE IN VARIATION HERE WITH

VARIANCE & STANDARD DEVIATION.

Variance σ^2 & Standard Deviation σ For Population

DATA: $x_1, x_2, x_3, \dots, x_n$

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

$$\text{Standard Deviation } \sigma = \sqrt{\sigma^2} \quad (\text{SQRT OF VARIANCE})$$

Variance s^2 & Standard Deviation s For Sample

DATA: $x_1, x_2, x_3, \dots, x_n$

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

$$\text{Standard Deviation } s = \sqrt{s^2} \quad (\text{SQRT OF VARIANCE})$$

s^2 & s ARE
GOOD ESTIMATES FOR
 σ^2 & σ

SMALLER DENOM. $\Rightarrow$ BIGGER FRACTION

$$\frac{1}{2} > \frac{1}{3}$$

e.g. DATA: 13 14 17 25 26 (MEAN = 19)

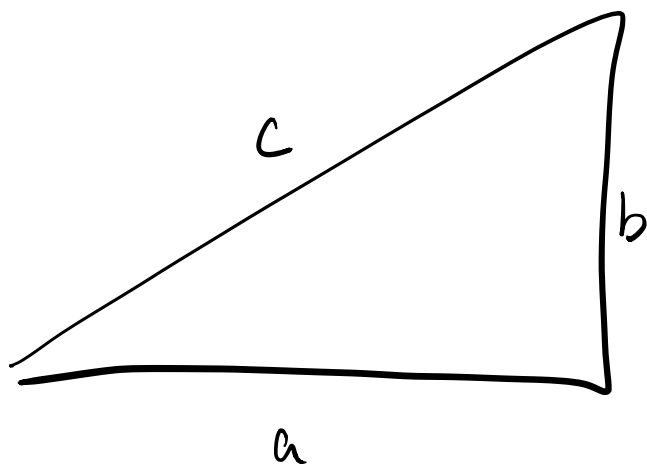
CALCULATE THE VARIANCE & STANDARD DEVIATION

ASSUMING THE DATA COMES FROM A (a) POPULATION

(b) SAMPLE

(c) HOW MANY MEASUREMENTS LIE WITHIN 1 STD. DEV. OF MEAN?

(d) HOW MANY MEASUREMENTS LIE WITHIN 2 STD. DEV. OF MEAN?



$$a^2 + b^2 = c^2$$

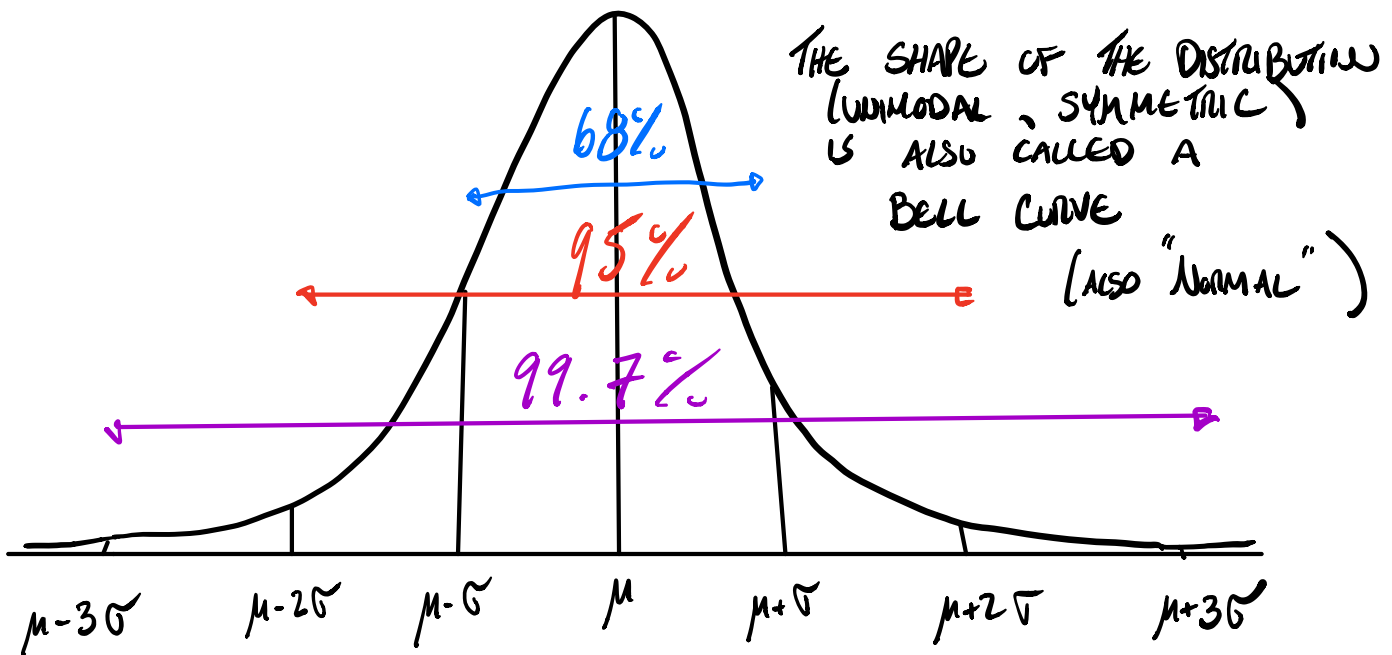
$$c = \sqrt{a^2 + b^2}$$

THE EMPIRICAL RULE :

APPROX. 68% OF DATA LIES WITHIN 1 STAND. DEV. OF MEAN

APPROX. 95% OF DATA LIES WITHIN 2 STAND. DEV. OF MEAN

APPROX. 99.7% OF DATA LIES WITHIN 3 STAND. DEV. OF MEAN



In Problems 11 and 12, find the standard deviation for each set of grouped sample data using formula (5) on page 525.

11. *no part*

Interval	Frequency
0.5–3.5	2
3.5–6.5	5
6.5–9.5	7
9.5–12.5	1

$$\frac{2(2) + 5(5) + 7(8) + 1(11)}{15} = \frac{96}{15} = 6.4$$

File #	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
2	2	-4.4	19.36
5	5	-1.4	1.96
7	8	1.6	2.56
1	11	4.6	21.16

$$s^2 = \frac{2(19.36) + 5(1.96) + 7(2.56) + 1(21.16)}{15 - 1}$$

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