Most answers are in orange.

1. We use graphs to communicate data—pictographs, pie charts, bar graphs, line graphs....
The **PICTOGRAPH** is perhaps the most fun since it offers a creative opportunity in the choice of symbol.
For example, in charting the popularity of songs I chose the symbol ♪....

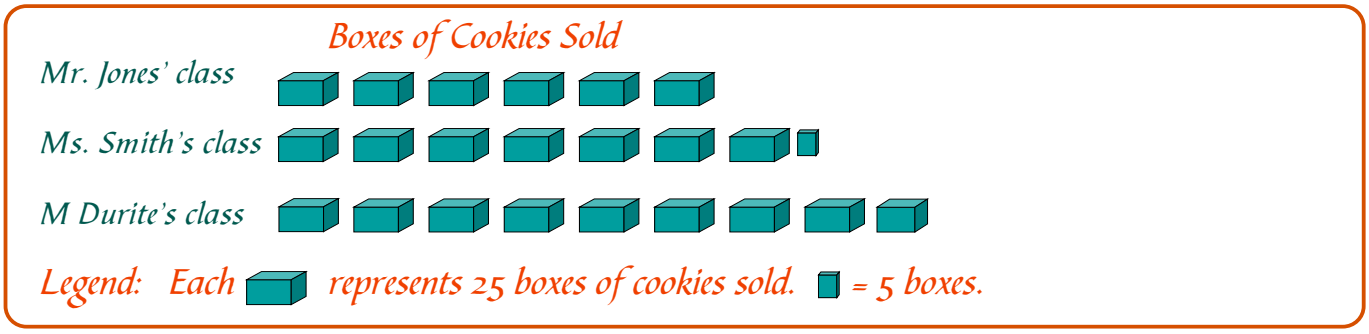
FAVORITE SONGS OF 40 KINDERGARTNERS

If You're Happy (and)	♪ ♪ ♪ ♪ ♪ ♪ ♪ ♪
Three Blind Mice	♪ ♪ ♪
Old MacDonald	♪ ♪ ♪ ♪
The Hokey Pokey	♪ ♪ ♪ ♪ ♪

Notice the ← TITLE and LEGEND

(Each ♪ represents 2 votes)

2. Picto-graph the cookie drive: Mr. Jones' class sold 150 boxes; Ms. Smith's, 180; M. Durite's, 220.



Would a pictograph be appropriate for displaying the following sets of data? (Hint: ask yourself how?)

➪ YES NO* Ages, in years, of members of a Bonsai class: 32, 21, 92, 26, 37, 48, 72, 41, 32, 16, 37, 48

YES NO Number of visits to the dentist for 20 kindergartners: 0,0,0,0,0,0,0,0,1,1,1,1,1,2,2,3,4,5,7,9

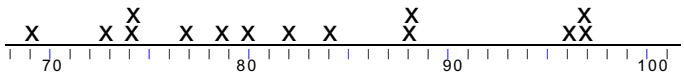
YES NO Eye color (brown, blue, green, gray) of the 275 students at Elm St. School.

* what would the categories be ?

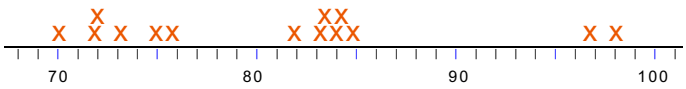
On the NSAT (National Science Achievement Test),
Ms. Smith's science class made the following scores: & Mr. Jones' class earned these:

69	73	74	74	77	79	80	85	73	70	98	83	75	76
82	84	88	88	96	97	97	97	82	72	83	72	84	84

3. Here is a **LINE PLOT** of Ms. Smith's class scores



Make a line plot of Mr. Jones' class scores



4. Ms. Smith's class scores in a **STEM-AND-LEAF** diagram.

Scores of Ms. Smith's 8th-graders on the National Science Achievement Test

6	9
7	3 4 4
7	7 9
8	0 2 4
8	8 8
9	
9	6 7 7

Legend: 7 | 7 9 *represents scores of* 77 & 79

Classify *all* the data in a back-to-back stem-and-leaf diagram.

Scores of two 8th-grade Science Classes on the National Science Achievement Test

Ms. Smith		Mr Jones
9	6	
4 4 3	7	0 2 2 3
9 7	7	5 6
4 2 0	8	2 3 3 4 4
8 8	8	5
	9	
7 7 6	9	7 8

Legend:

9 7 | 7 | 5 6

Represents scores of 75 & 76 by Mr. Jones' students, and 77 & 79 in Ms Smith's class.

NOTE: Using fewer than 5 classes is generally considered unacceptable.

Mr. Jones' science class earned these scores:

85	73	70	98	83	75	76
97	82	72	83	72	84	84

& Ms. Smith's class earned these:

69	73	74	74	77	79	80
82	84	88	88	96	97	97

5. A **FREQUENCY TABLE** lists ranges of values for the data, and their frequencies– the number of data that fall in each range . Classify the data in a **combined** frequency table.
(Use classes that correspond to the stem-and-leaf diagram above.)

Notice there is ALWAYS a TITLE!

Scores of 28 students in Mr. Jones & Ms. Smith's classes on the NSAT

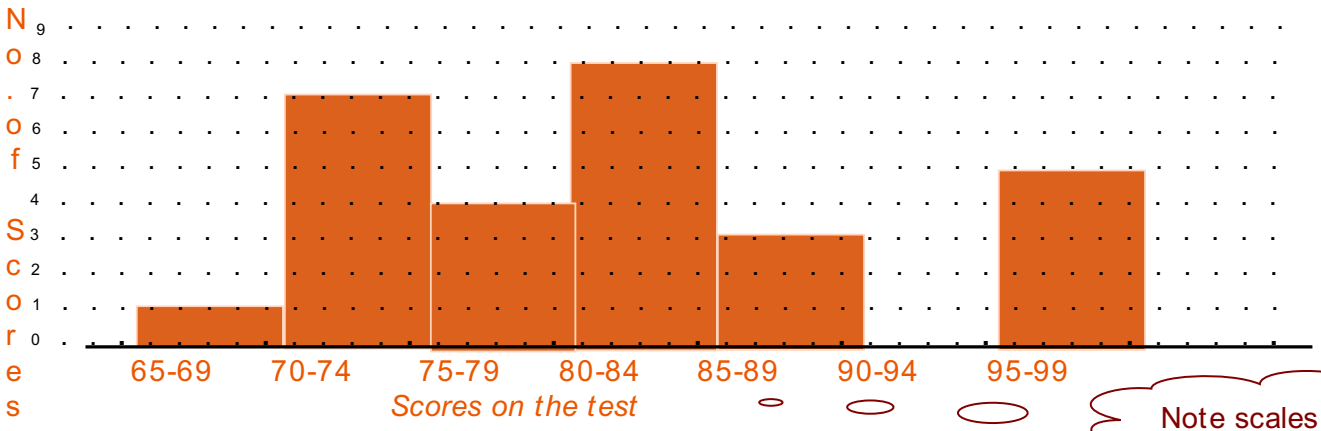
Scores on test	*Number of students*
65-69	1
70-74	7
75-79	4
80-84	8
85-89	3
90-94	0
95-99	5

Note informative headings, not generic jargon words

It would be misleading (and thus wrong !) to omit the empty category.

6. Show the combined data in a **HISTOGRAM** using classes which correspond to those in problem (5).
A histogram uses adjacent rectangles on a Cartesian coordinate system to display data.
The horizontal or x-axis displays the values of the data; the vertical or y-axis the frequencies.

Scores of 28 students in Mr. Jones & Ms. Smith's classes on the NSAT



Note scales & labels ...on both axes !

Discrete vs Continuous Data: Histograms and Bar Graphs both show frequencies of data grouped in categories, in summary form. The difference is that the HISTOGRAM is used for CONTINUOUS numeric data, and a BAR GRAPH is used when the possible data values are either non-numeric, or are SEPARATE VALUES, rather than a continuous range. (Most often, measurements are continuous.) For example:
We show distributions of trees BY HEIGHT via a histogram, since tree HEIGHTS cover a continuous range of values.
We show distribution of trees BY TYPE (oak, sycamore, manzanita) on a bar graph.

⇨ State whether we should use a histogram, or a bar graph to show HOW MANY CHILDREN...
...MADE THE HONOR ROLL IN CLASSROOM 1, IN CLASSROOM 2, ETC. at Elm St. Elementary:
use a Bar graph since the categories are the distinct, separate classrooms!
...ARE 40"-49" TALL, 50"-59" TALL, ETC. at Elm St. Elementary: use a Histogram the categories are ranges of heights!

DETAILS: The areas of the rectangles (or bars) must be in proportion to the frequencies with which data falls into each category– if twice as much data, then twice the area. For histograms & frequency tables, we use equal ranges of values for each class (or category), except when there is a good reason to do otherwise.
(Bar graphs may be displayed sideways, with the variable of interest on the vertical axis & frequencies on the horizontal axis. Histograms are generally drawn with frequencies on the vertical axis and the data– always numeric– on horizontal axis.)

(ADDENDUM TO PAGE 2—THIS IS NOT REQUIRED MATERIAL!)

Here are a few additional examples of data for which you may decide whether to use a simple bar graph, or a histogram.

To display each of the following, identify whether you would use a histogram or bar graph (& why):

a. Heights of students at M Powers Jr. High School were measured.
HEIGHT IS A CONTINUOUS VARIABLE. USE A HISTOGRAM.

b. The same students were asked to select a snack food for the cafeteria.

Popcorn	78 votes
Pretzels	35 votes
Peanuts	62 votes
Phruit	44 votes
Phried Chips	86 votes

VARIABLE ↗ Frequency↗

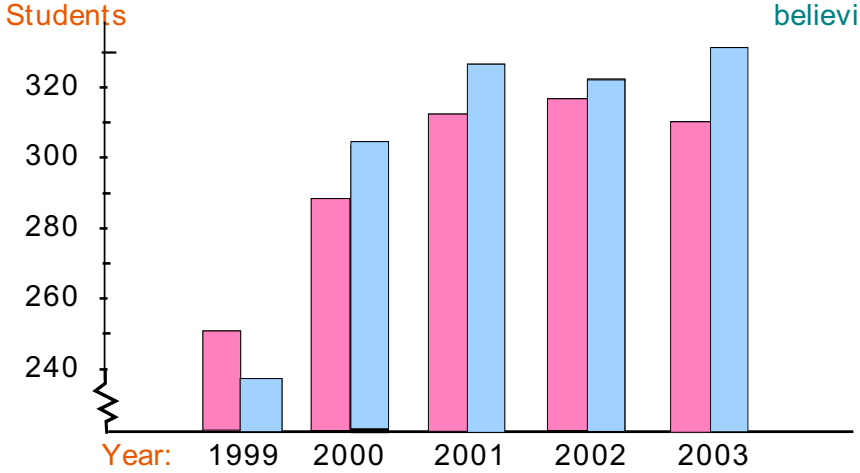
The VARIABLE here is not even numeric, let alone continuous. Histograms are used for variables which have a continuous possible range of values. So if we choose to use a bar graph to display the number of votes for each of these separate categories, the bars should be separated, not contiguous. A histogram would be inappropriate. USE A SIMPLE BAR GRAPH.

c. Over the five years starting in 1999, Elm St. Elementary has had the following Fall enrollments:

1999:	248 girls	234 boys
2000:	287 girls	299 boys
2001:	310 girls	325 boys
2002:	315 girls	320 boys
2003:	307 girls	330 boys

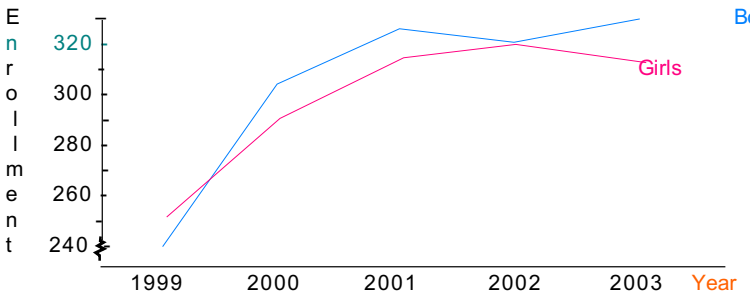
Each school year is a separate, distinct category. We are counting the number enrolled for each separate school year, so a simple bar graph is appropriate. However, because we have a PAIR of enrollment figures for each year, we make the artistic decision to cluster the two bars for each year, believing this will make the graph more readable.

[⇦ Paired Bar graphs are used fairly often, but are NOT required material in this course.]



Elm St. Elementary Enrollments
By Year and Gender 1999-2003

Boys

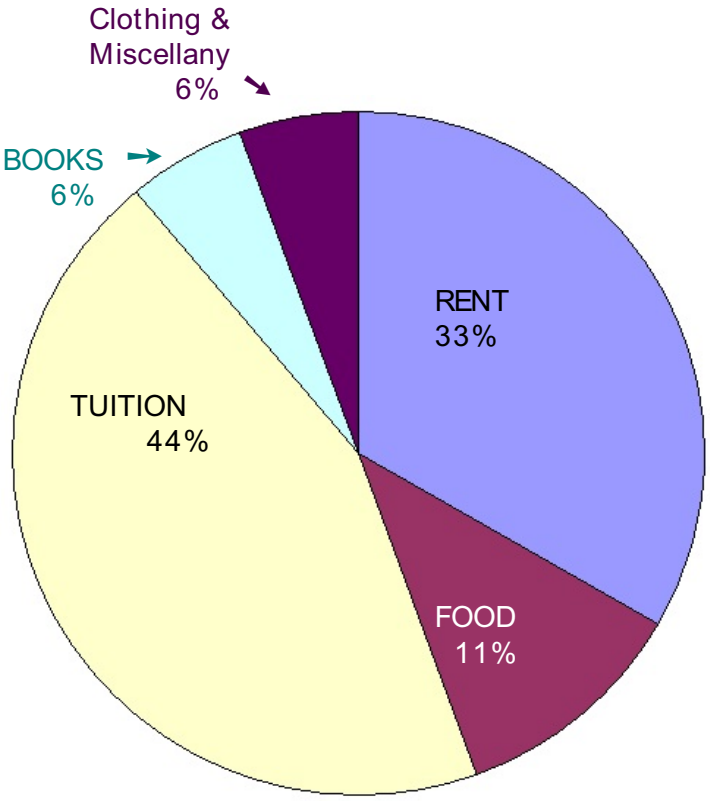


Here the above data is shown in a "Paired line graph" (also NOT required!) We formed the line graphs by connecting the midpoint of the top of each bar in the paired bar graph above. Line graphs are sometimes called frequency polygons, although technically a frequency polygon must drop to the baseline at each end. .

7. Gina spent the following amounts every month on the average while attending United University in '96. Illustrate the proportions with a **PIE CHART (CIRCLE GRAPH)**.
Don't forget titles and legends.
Label each sector/segment ($ amt or %)

ITEM	$$$	*degrees*
Rent	$300	$^{300}/_{900}$ x 360° = 120°
Food	100	$^{100}/_{900}$ x 360° = 40°
Books	50	et cetera
Tuition	400	
Clothing & misc	50	
Total	900	

We need the total to find what PART each expenditure is of the whole pie.



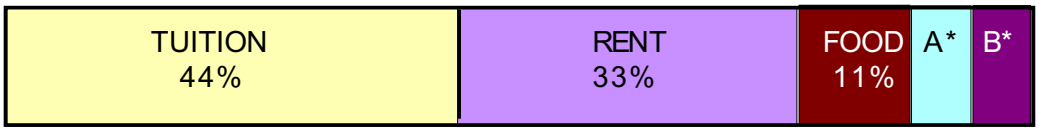
Monthly expenses of Gina, a college student at United University in 1996, out of $900.

NOTES:
It is also common practice to place the data in the pie chart.
That is, in lieu the percentages (44%, 33%, etc.) we might state the amounts ($400, $300, etc.)
When percentages are stated, somewhere the total should be expressed—
here it is given in the title, "...out of $900"

It is **NOT** APPROPRIATE to state the degrees of central angles for the pie sectors in the pie chart!
No one looking at the graph cares about the number of degrees. That is something you compute so that you can allot the appropriate portion of the circle. It's part of your scratch work.

Here is a BAR CHART (just like a pie chart only it's not round... they ought to call it a sheet-cake chart!)

Monthly expenses of college student Gina at United University, out of $900.



A* : BOOKS, 6%
B* : Clothing and Miscellany, 6%

[Since it was too hard to draw the arrows needed here, we coded the chart with "A" and "B".]

"**A STATISTIC**" can be a value calculated from data that represents some characteristic of the data. Among statistics commonly used to describe the **"average" ("typical", "central") value** of a set of data are the **MEAN**, the **MEDIAN** and the **MODE**.

8. The **MEDIAN** of data is the value at which 50% of the data consists of higher values, and 50% lower.
– In short: *The value in the middle.*
(Or the average of the two values in the middle*, when the number of data is even.)

Using an ordered stem-and-leaf diagram, for instance, we can easily ascertain the median.
The median of Ms. Smith's science class scores is: 81
The median of Mr. Jones' science class scores is: (average of 82 & 83)* = 82.5
The median of the combined data is: (Because there are 28 data, average of 82 & 82)* = 82
*We counted in to the middle, using the ordered stem-and-leaf diagram in #4.

9. The **MODE** is the most frequent value in the data. Find the mode of the combined class scores.

Since 84 and 97 each occur three times in the combined data, and no other value occurs that many times, we say the data is bimodal, and the modes are 84 and 97.
If there are two equally most-frequently-occurring values, we say the distribution is bimodal, and has two modes. If there are three, then we have to call it trimodal. (Four? Let's not go there!)

10. Suppose ten people in a room decide to pool and split their wealth evenly: they will each throw all their money into a pot, then divide the money among themselves equally.
⇨ a. If every person has $80 to start, what will they each end up with? Each gets $80
b. Suppose everyone has $80, except one person has only $60, and another has $100 . Each gets $80
c. Suppose everyone has $80, except one person has $500. Each gets $122
(Nine get an extra $42 each, one loses $378)

11. Compute the **MEAN** of data: 70 80 82 83 87 90 91 97..

The sum of: 70 80 82 83 87 90 91 97 is 680 ... The mean is 680/8 = 85

12. Suppose the person above who scored 70 had instead given up and received a score of 0.
What would the class mean have been?

The mean for the class would have been $\frac{\text{total pts.}}{n} = \frac{(680-70)}{8} = \frac{610}{8} = \frac{305}{4} = 76.25$
(Quite a bit lower... the effect of an **extreme** value on the mean.)

The 16 students in another class achieved a mean of 75. What is the combined mean for the two classes?
(NOT 80!)

The otherclass of 16 students averaged 75, so their total must be 75•16 = 1200

Combined mean $= \frac{\text{total of all points}}{\text{Number of students}} = \frac{1200 + 680}{16+8} = \frac{1880}{24} = \frac{235}{3} \doteq 78.33$

We could also have said this: Combined mean = (⅔)(75) + (⅓)(85) ≐ 78.33 ❖

NOTE it is WRONG to "average" the two numbers, 75 and 85, as if they should have equal weight!
The 2nd class has *twice as many students*, so should *weigh twice as heavily* in the computation of the mean!
In fact, we could have obtained the correct result by computing: (1•85 + 2•75) ÷ 3 ⇨ same as ❖ above.

Suppose two students in your class earn 100 on a test, five earn 85, and thirteen earn 70.
Would the mean score for those students on that test be (100+85+70)/3 = 255/3 = 85? (If not, then what?)

NO, the mean would be $\frac{\text{Total ALL Points}}{20} = 76.75$

13. For the following simplistic sets of grouped data— find the median & the mean:

a	f
50	1
70	8
90	1

b	f
50	3
70	6
90	3

c	f
50	4
70	4
90	4

d	f
50	6
70	0
90	6

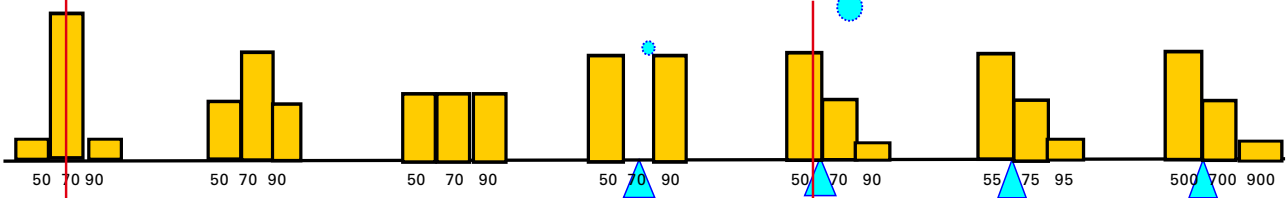
x	f
50	6
70	3
90	1

y	f
55	6
75	3
95	1

z	f
500	6
700	3
900	1

the Mean is the balancing point of the distribution!

Median is dividing line between top and bottom halves



	a	b	c	d	x	y	z
mean:	70	70	70	70	60	65	600
median:	70	70	70	70	~57	~62	~570

14. The first four distributions are symmetric.
When the distribution is symmetric, the mean & median are the same, at the middle of the distribution.
The fifth (**x**) distribution is asymmetric, "skewed to the right".
Where is the mean, relative to the median? The mean is above the median. The mean is "**pulled to the right**" **by** the one **extreme value, 90.** The median is unaffected by one or a few very large or very small values.

15. The values in the ***Y*** distribution are *each 5 more than* those in the ***X*** distribution (each $y_i = 5 + x_i$).
How do the means for ***Y*** & ***X*** compare? The mean for Y is 5 MORE THAN for X. Does this make sense?
If every value in a set of data is increased by an amount "a" (i.e. amount *a* is ADDED), then the mean is INCREASED BY THE SAME AMOUNT "a". (Or decreased, in the case where "a" is negative.) This makes sense if you consider that the mean is the balancing point, which will move with the values. It also makes sense in view of the arithmetic... every data contributes 5 more "to the pot", so when the pot is divided up, each will get 5 more.

16. The values in the ***Z*** distribution are *10 times* those in the ***X*** distribution ($z_i = 10 \cdot x_i$). How does the mean for Z compare with the mean for X? 10 TIMES AS GREAT. Does this make sense? Yes! As we note here:
If every value in a sample or population is MULTIPLIED by a factor "r", then the mean is MULTIPLIED BY THE SAME FACTOR "r".

17. Suppose the example ***X*** data above is doubled and then increased by 3. What is the new mean?
The mean for the new data is 123.
Doubling every value, then adding 3, the mean will be INCREASED IN THE SAME MANNER, to 2•60+3 = 123.

DSB6

At right are sales prices of the 30 single-family homes sold in Northridge in January, 1995.

$	f
125K	15
175K	11
250K	2
1000K	1
3000K	1

18. Find the median.
There are 30 data. The median is the value "in the middle"–
the average of the 15th and 16th data. The 15th can be viewed as 125K,
the 16th as 175K. The average of those two values is 150K.

19. Is the mean higher or lower than the median? Higher, because the data is skewed right.
Find it. mean = (15•125K + 11•175K + 2•250K + 1•1000K + 1•3000K) /30 = 276.66666K
The mean selling price of Northridge homes in that month was $276,667

20. Find the mode. 125K.

21. If you are interested in the price of housing in a particular area, which of the above "average" statistics would you want to know, to estimate the price of houses in that area?
The median, since it is a middle value, and is not affected by a few houses that are very expensive.
Why not the mode? In general, the mode is a rather "serendipitous" value, easily affected by minor differences in the data. (PS We don't really know the mode in this case, because the data was summarized.)

Mode = most frequent value
Median = the middle score [the *(n+1)/2th*]
Mean = the evenly distributed total; also the balancing point of the distribution

Measures of Central Tendency: Among statistics commonly used to describe the "variation", or spread, in a set of data— are the **STANDARD DEVIATION**, the **RANGE** and the **INTERQUARTILE RANGE ("IQR")**.
The **RANGE** is the total span or spread of the data, i.e. the highest value – the lowest value.
Just as the median divides the ordered data into halves, the **QUARTILE MARKS** divide the data into four equal groups. These quartile marks are referred to as the **first** and **third quartile** marks, and the **second quartile mark**, which is also the median. EG we find range, IQR and standard deviation of the scores in Mr. Jones' class.

22. range = span of the data = maximum data value – minimum data value = 98 – 70 = 28

Interquartile range = $Q_3 - Q_1$ = 84 – 73 = 11

70 72 72 73 75 76 82 | 83 83 84 84 85 97 98
Q_1 Q_3

23. Standard deviation = sq. root of (average square distance from the mean) = $\sqrt{\frac{\Sigma(x-\mu)^2}{n}}$ ← μ *or* $\bar{x}$

n – 1 for a sample

x	dist to mean (3)	squared...
1	2	$2^2 = 4$
2	1	$1^2 = 1$
3	0	$0^2 = 0$
4	1	$1^2 = 1$
5	2	$2^2 = 4$
		$\Sigma = 10$

s.d. = $\sqrt{\text{average sq. dist. to mean}}$

$= \sqrt{10/5} = \sqrt{2}$

For the data: 1 1 3 5 5 the mean is also 3, but the std. dev. is greater: $\sqrt{16/5} = \sqrt{3.2}$

24. Consider our simple grouped data examples— compute standard deviation:

a	f
50	1
70	8
90	1

b	f
50	3
70	6
90	3

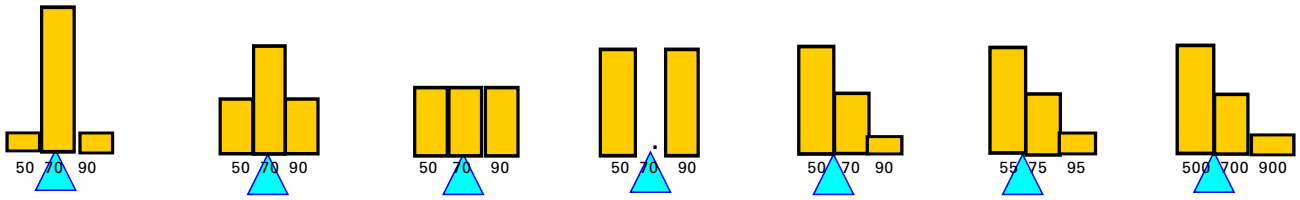
c	f
50	4
70	4
90	4

d	f
50	6
70	0
90	6

x	f
5	6
7	3
9	1

y	f
15	6
17	3
19	1

z	f
50	6
70	3
90	1



mean:	$\bar{a} = 70$	$\bar{b} = 70$	$\bar{c} = 70$	$\bar{d} = 70$	$\bar{x} = 60$	$\bar{y} = 65$	$\bar{z} = 600$
	sd ≈ 8.9	sd ≈ 14.1	sd ≈ 16.3	sd ≈ 20	sd ≈ 13.4	sd ≈ 13.4	sd ≈ 134
	s ≈ 9.4	s ≈ 14.8	s ≈ 17.1	s ≈ 20	s ≈ 14.1	s ≈ 14.1	s ≈ 141

Computed ⇦ using "n"

"s" is computed with "n– 1"

One example of how we do the computation:

In the first distribution, there is one value at 50, another at 90, and the rest at 70.
The mean of this symmetric distribution is clearly 70.
We sum the squared distances to the mean:

$(50 - 70)^2 = 20^2 = 400.$
$(70 - 70)^2 = 0 = 0$
$(90 - 70)^2 = 20^2 = 400.$
Total sq. distances = 800

...and divide by n*

The std. dev is either $\sqrt{\frac{800}{10}} \doteq 8.94$

or (n – 1)*

$\sqrt{\frac{800}{9}} \doteq 9.43$

For the second (b) distribution:

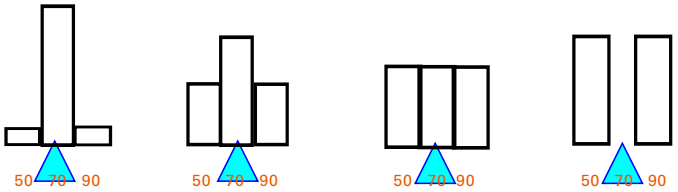
$3\cdot(50 - 70)^2 = 3\cdot400 = 1200$
$6\cdot(70 - 70)^2 = 6\cdot0 = 0$
$3\cdot(90 - 70)^2 = 3\cdot400 = 1200$
Total sq. distances = 2400

$\sqrt{\frac{2400}{12}} = \sqrt{200} \doteq 14.1$

3 + 6 + 3 ↗

*If computing for a whole population, we divide by n. If we are computing based on a sample, rather than from the entire population, we divide by (n–1), a correction factor necessary because the average variation in a sample is just slightly smaller than the variation in the population. **Our text uses n, so you can just use n!!!**

25. Would you say the first four distributions are "clustered around the mean" to the same degree? How is this reflected by the standard deviation?



Each distribution above has mean 70 & range 40.

In A, 80% of the distribution is at the mean, and only 20% at the extremes of the range. Most of the data is tightly clustered at the mean. In B, C,& D, the data is progressively **less clustered** at the mean, and **more dispersed** toward the extremes of the range. This is seen in the **standard deviations**, which **increase** from ~9 to 20, from left to right.

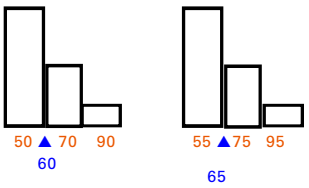
26. What kind of data would have a standard deviation of 0? ...a negative standard deviation?

Since the standard deviation is directly related to the sum of the squared distances* to the mean, it is not possible for the standard deviation to be negative.
The standard deviation can be zero only if each distance to the mean is zero, and that can occur only if every value in the distribution is the same number– all the data is the same value as the mean.
Here is an example: 70 70 70 70 70. Mean is 70. Std. Dev. is 0.

*which CANNOT be negative!

27. The values in the *Y* distribution are each 5 more than those in the X distribution (*ie*: $y_i = 5 + x_i$). How does the standard deviation for Y compare with the std. deviation for X? Does this make sense?

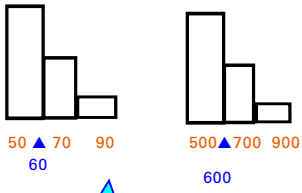
Consider the range of the data in the X distribution, 90 – 50 = 40, and the range of the Y distribution, 95 – 55 = 40.



Likewise, the standard deviation of the Y distribution is the same as that of the W . The data in both distributions is dispersed the same distances from the mean– 6 values 10 units below the mean, 3 values 10 units above the mean, and 1 value 30 units above the mean. Both lead to the same sum of squares: 6•100 + 3•100 + 1•900.

28. The values in the *Z* distribution are 10 times those in the *X* distribution ($z_i = 10 \cdot x_i$). How does the standard deviation for Z compare with the std. deviation for X? Does this make sense?

Again, consider the range– the Z distribution covers a range that is ten times as wide as that of the X distribution.
Furthermore, each value in the Z distribution is ten times as far from the mean as its counterpart in the X distibution.
Compare the sum of squares distances for X: $6 \cdot 10^2 + 3 \cdot 10^2 + 1 \cdot 30^2$.
With the sum of square distances for Z: $6 \cdot 100^2 + 3 \cdot 100^2 + 1 \cdot 300^2$
Conclusion: The std. dev. for Z is TEN TIMES the std. dev. for W.



As we see in # 28... all values & distances Are 10 times as great, So mean and std dev are too.

29. **If every value in a sample or population is MULTIPLIED by a factor "r", then the mean is multiplied by the same factor r, and the standard deviation is also multiplied by the same factor r.**

If an amount "a" is ADDED to every value in a sample or population, then the mean is increased by "a", and the standard deviation is UNCHANGED.
(All values move right, so mean moves right, too. But "spread" of data is unchanged.)

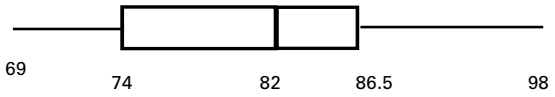
Range = highest value — the lowest value = the "width" of the data
Interquartile range = third quartile — first quartile = the "width of the middle 50%" of the data
Standard deviation = square root of average square distance from the mean (almost)

69 70 72 72 73 73 74 | 74 75 76 77 79 80 82 | 82 83 83 84 84 84 85 | 88 88 96 97 97 97 98

30. Find the median and the quartile marks of the combined data *of #3*.
(The **QUARTILE MARKS** consist of the first quartile,Q_1, and the third quartile, Q_3, along with the second quartile mark, which is also the median. These three values divide the data into quarters, thus the name.)
We align the data in order, and find the locations that divide the data into four equal groups.
These are: $Q_1 = (74+74)/2 = 74$
$Q_2 = (82+82)/2 = 82$ (median)
$Q_3 = (85+88)/2 = 86.5$

This is a box plot for data with median 82, Q_1=74, Q_3=86.5, lowest value 69, and highest value 98 (Of course this box plot needs a title.)

Scores of 28 students on the NSAT



Mr Jones' class scores, in order:
70 72 72 73 75 76 82 | 83 83 84 84 85 97 98
$Q_1 = 73$ $Q_2 = 82.5$ $Q_3 = 84$

Ms Smith's class scores, in order:
69 73 74 74 77 79 80 | 82 84 88 88 96 97 97
$Q_1 = 74$ $Q_2 = 81$ $Q_3 = 88$

31. What are the IQRs– the **INTERQUARTILE RANGES** ($Q_3 - Q_1$) of the Classes in #4? Draw **BOX PLOTS** for both.
Step 0: Find the FIVE-NUMBER SUMMARY (min, max, Q_1 & Q_2 & Q_3)
Step 1: Draw a box across the middle 50%– thus from Q_1 to Q_3, divided at the median.
Step 2: Determine any **OUTLIERS**– data more than 1½ IQRs outside of interval from Q_1 to Q_3.
Step 3: Draw "whiskers" from the box outward to the highest and lowest data that are not outliers.
Step 4: Add asterisks to the line plot for any outliers. (Label, with values at all important points.)