We use graphs to communicate data– pictographs, pie charts, bar graphs, line graphs, scatter diagrams.... The most common graphs are used to show *how many*...

1. 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 ↓

A pictograph is a special type of Bar graph.

(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.

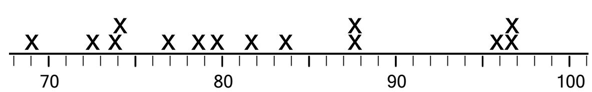
On the NSAT (National Science Achievement Test),
Ms. Smith's science class made the following scores:

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

& Mr. Jones' class earned these:

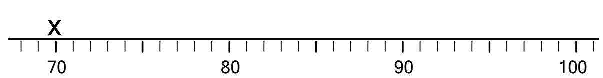
85	73	70	98	83	75	76
97	82	72	83	72	84	84

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



Ms. Smith's class scores on NSAT

➩ Make a line plot of Mr. Jones' class scores ⇦



4. Ms. Smith's class scores in a STEM-AND-LEAF diagram are shown below.

Add Jones' data to complete the back-to-back stem-and-leaf diagram below. Don't forget labels and legend. ⇦

➩ *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

Note: Four is considered too few classes when grouping data! So w e did NOT use ranges 70-79, 80-89, etc. To get more than 4 classes, we used shorter ranges.

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

Smith's Class		
9	6	
4 4 3	7	
9 7	7	
4 2 0	8	
8 8	8	
	9	
7 7 6	9	

⇦

Legend: ⇦

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

5. A FREQUENCY TABLE lists ranges of values for the data, and their frequencies– the number of data that fall in each range (class). Classify the data in a **combined** frequency table (ie *complete* what was started below). (Use classes that correspond to the stem-and-leaf diagram above.) Don't forget the title.

⇨

Scores on test	*Number of students*
65-69	1
70-74	7
75-79	4
80-84	

⇨

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.

USE the SPACE!!

⇨
⇨
⇨

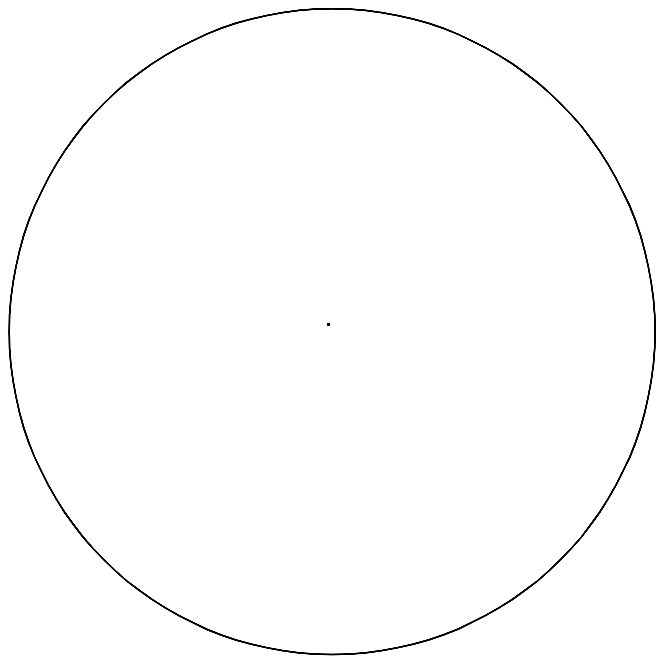
Discrete vs Continuous Data: Histograms and Bar Graphs both show frequencies of data grouped in categories, in summary form. The difference is that the HISTOGRAM **is used for** CONTINUOUS numeric data, and a BAR GRAPH is used when the possible data values are either non-numeric, or are SEPARATE VALUES, 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 ________________
...ARE 40"-49" TALL, 50"-59" TALL, ETC. at Elm St. Elementary: use a ________________

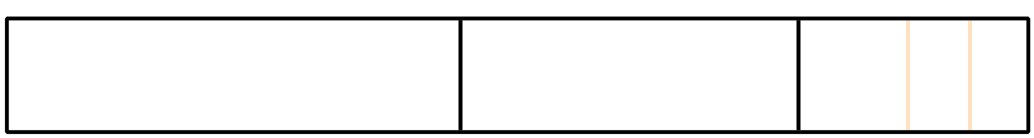
DETAILS: The areas of the rectangles (or bars) must be in proportion to the frequencies with which data falls into each category– if twice as much data, then twice the area. For histograms & frequency tables, we use equal ranges of values for each class (or category), except when there is a good reason 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.)

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, and to...
Label each sector/segment ($ amt or %)

ITEM	$$$	*degrees*
Rent	$300	$\frac{300}{900} \times 360^\circ =$
Food	100	
Books	50	
Tuition	400	
Clothing & misc	50	
Total		



A “Single-Bar” graph (not a bar graph) does the *same thing* as a pie chart, but in a bar rather than a “pie”.
Keeping that in mind, for the same data given above, complete this single-bar graph.
Don’t forget to communicate ALL the information.



"A STATISTIC" can be a value calculated from data that represents some characteristic of the data. Among statistics commonly used to describe the **"average"** (or "typical") **value** of a set of numeric 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 dividing line between the top half and the bottom half. The value in the* ***middle.***
(The average of the two values in the middle, when the number of data is even.)

Using an ordered stem-and-leaf diagram we can easily ascertain the median.

⇨ The median of Ms. Smith's science class scores is:

⇨ The median of Mr. Jones' science class scores is:

⇨ The median of the combined data is:

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

NOTE— 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
b. Suppose everyone has $80, except one person has only $60, and another has $100 . Each gets
c. Suppose everyone has $80, except one person has $500. Each gets

11. The **MEAN** is the arithmetic average: MEAN = (sum of data)/(number of data)
Compute the mean of these data (test scores): 70 80 82 83 87 90 91 97.
⇨

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

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?)

13. Following are mini-histograms for very simplistic sets of grouped data– kept very simple for clarity and easy computation, so we learn something about the mean and the median.

For each distribution, find the median & the mean:

a	f
50	1
70	8
90	1

b	f
50	3
70	6
90	3

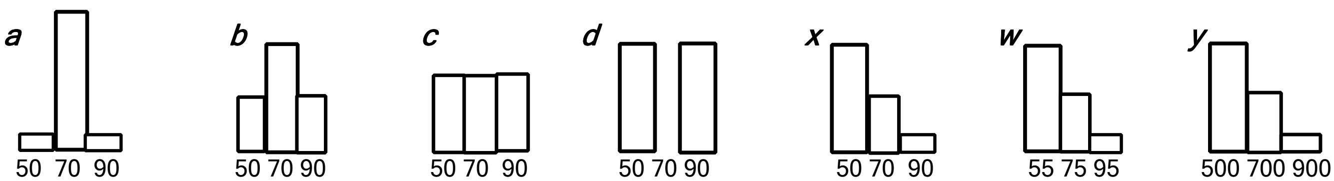
c	f
50	4
70	4
90	4

d	f
50	6
70	0
90	6

x	f
50	6
70	3
90	1

w	f
55	6
75	3
95	1

y	f
500	6
700	3
900	1



⇨ *mean:* $\bar{a} = 70$ $\bar{b} = 70$ $\bar{c} =$ $\bar{d} =$ $\bar{x} = 60$ $\bar{w} =$ $\bar{y} =$

median: $\tilde{a} = 70$ $\tilde{b} = 70$ $\tilde{c} =$ $\tilde{d} =$ $\tilde{x} \approx 57$ $\tilde{w} =$ $\tilde{y} =$
⇨

14. The first four distributions are symmetric.
⇨ When the distribution is symmetric, how do the mean & median compare ?

The fifth (x) distribution is asymmetric, "**skewed to the right**" (the "tail" is on the right).
⇨ Where is the mean, relative to the median?

Note: When data is skewed, it is because some values are exceptionally high (skewed right) or low (skewed left). A few exceptionally high values have a profound effect on the mean, but no effect on the median, as we will see on the next page. Thus data skewed right will have a mean higher than median, because the mean is pulled high by the asymetrically high values (high values that are not balanced by corresponding low values).

15. The values in the *W* distribution are ***each 5 more*** *than* those in the *X* distribution ($w_1 = 5 + x_1$, et cetera).

⇨ How do the means for *W* & *X* compare? [Does this make sense?]

⇨ IF EVERY VALUE IN A SET OF DATA IS INCREASED BY AN AMOUNT "A" (ie. AMOUNT A IS ADDED), THEN THE MEAN IS

16. The values in the *Y* distribution are ***each 10 times*** those in the *X* distribution ($y_1 = 10x_1$, $y_2 = 10x_2$, etc).

⇨ How does the mean for *Y* compare with the mean for *X*? [Does this make sense?]

IF EVERY VALUE IN A SAMPLE OR POPULATION IS MULTIPLIED BY A FACTOR "R", THEN THE MEAN IS
⇨

17. Suppose the example *X* data above is doubled, then increased by 3. What is the new mean?
⇨

At right are the selling prices of the 30 single-family homes sold in Northridge in January, 1995 (1 yr. after...).

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

18. Find the median.

⇨

19. Is the mean higher or lower than the median?
Find it.

⇨

20. Find the mode. (OK, the “modal class”)

⇨

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? Why?

⇨

Statistics that tell us the “typical” value of a set of data:

Mode = most frequent value

Median = the middle score [the *(n+1)/2th*] $\tilde{x}$

Mean = the evenly distributed total; also the balancing point of the distribution $\bar{x}$ or μ

What’s the difference between $\bar{x}$ and μ ?

Statisticians use $\bar{x}$ to refer to the mean of a *sample* (taken from a population),
and μ for the mean of the *entire population* .

Measures of Central Tendency: Among statistics commonly used to describe the amount of "variation", or spread, in a set of data... are t he STANDARD DEVIATION, the RANGE and the INTERQUARTILE RANGE.

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 two equal groups, in order, QUARTILE MARKS divide the data into four equal groups. These quartile marks are referred to as the **first** and **third quartile** marks (Q_1 & Q_3), and the **second quartile mark**, which is also the median. The interquartile range (IQR) is the distance from Q_1 to Q_3.

22. Compute the *range* and *interquartile range* of the scores in Mr. Jones' class.

⇨ Range = maximum data value − minimum data value = 28

⇨ InterQuartile Range = Q_3 − Q_1 = 11

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

↑ ↑ ↑

The standard deviation is a bit more complicated....

STANDARD DEVIATION = sq. root of (average square distance from the mean) = $\sqrt{\dfrac{\Sigma(x-\mu)^2}{n}}$ ← μ *or* $\bar{x}$; ← n, *or* n − 1 for a sample

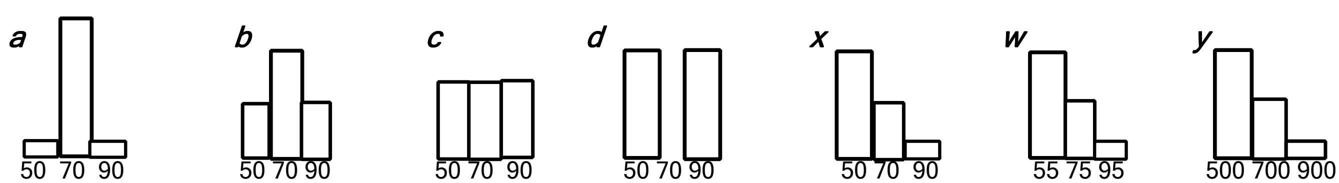
23. Compute the s.d. (s or σ) for the data: 1 2 3 4 5 (First find the mean!)

x	x - mean distance	(distance)^2
1		
2		
3		
4		
5		
Total:		

⇨ Would the data: 1 1 3 5 5 have the same standard deviation as the data above?

24. Consider the simple grouped data examples we used to explore the mean and median— now compute the standard deviations:

a	f	b	f	c	f	d	f	x	f	w	f	y	f
50	1	50	3	50	4	50	6	50	6	55	6	500	6
70	8	70	6	70	4	70	0	70	3	75	3	700	3
90	1	90	3	90	4	90	6	90	1	95	1	900	1



	a	b	c	d	x	w	y
mean:	$\bar{a}$ = 70	$\bar{b}$ = 70	$\bar{c}$ = 70	$\bar{d}$ = 70	$\bar{x}$ = 60	$\bar{w}$ = 65	$\bar{y}$ = 600
⇨	sd ≐ 8.96	**sd ≈**	sd ≐ 16.3	sd ≈ 20	sd ≐ 13.4	sd ≐ 13.4	sd ≐ 134
	s ≐ 9.4		s ≐ 17.1	s ≐ 20.9	s ≐ 14.1	s ≐ 14.1	s ≐ 141

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

⇨

⇨

26. What kind of data would have a standard deviation 0 ? Can you find data with a negative standard deviation ?

⇨

⇨

27. The values in the W distribution are each 5 more than those in the X distribution, as noted earlier.

⇨ How does the standard deviation for W compare with the std. deviation for X? Does this make sense?

⇨

28. The values in the Y distribution are 10 times those in the X distribution, as noted earlier.

⇨ How does the standard deviation for Y compare with the std. deviation for X? Does this make sense?

⇨

29. **If an amount "a" is ADDED to every value in a sample or population, then**

⇨ **the mean is**

⇨ **the standard deviation is**

If every value in a sample or population is MULTIPLIED by a factor "r", then

⇨ **the mean is**

⇨ **the standard deviation is**

Statistics that describe the amount of "spread" in the data:

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)

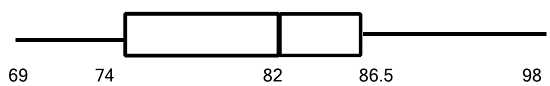
Smith:	69	70	72	72	73	73	74	74	75	76	77	79	80	82
Jones:	82	83	83	84	84	84	85	88	88	96	97	97	97	98

30. Find the median and the quartile marks of the combined data *of #3*.
(The QUARTILE MARKS consist of the 1st, Q_1, and the 3rd, Q_3, along with the 2nd quartile mark, Q_2, which is also called the median, . These three values divide the data into quarters, thus the name.)

⇨

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.)*



Notice the numbers must be proportionally spaced, as they would be on the number line.

31. What are the IQRs- the INTERQUARTILE RANGES ($Q_3 - Q_1$) of the Classes in #4?
In space below, draw *two separate* BOX PLOTS for BOTH sets of scores (Jones & Smith) on the NSAT (see #3).
To make a Box Plot:
0: Find the quartile marks, and the interquartile range (to see if there are outliers). Find the min & max.
1: Draw a box across the middle 50%- thus from Q_1 to Q_3, divided at the median.
2: Determine any OUTLIERS- data more than 1½ IQRs outside of interval from Q_1 to Q_3.
3: Draw “whiskers” from the box outward to the highest and lowest data that are not outliers.
4: Add asterisks to the line plot for any outliers. (Label, with values at all important points.)

⇨

⇨