Describing 1-Variable Statistics Toolkit #2

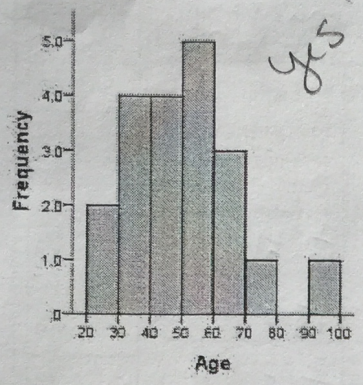
G___

Name: _____

Period: A1 A2 A3 B1 B3

Types of Displays for 1-Variable Quantitative Data (3 of these are NOT for quantitative data!)

Histogram



yes!

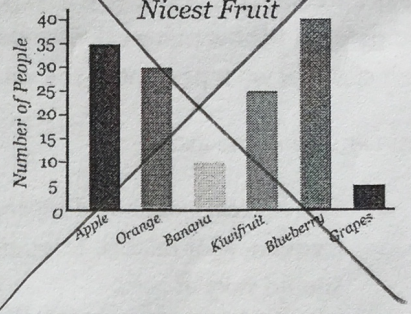
Pictograph

misleading sometimes

Year	Number of tv sets sold
1990	
1995	
2000	
2005	
2008	

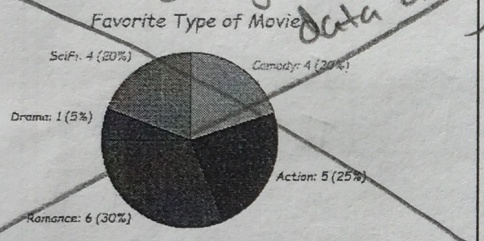
Bar Graph

categorical data only



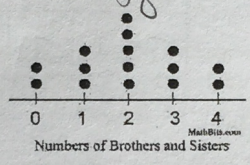
Pie Chart

Categorical data only



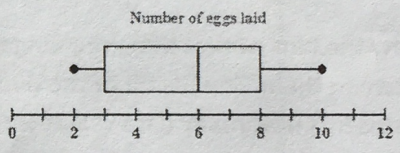
Dot Plot

yes!



Box Plot

yes!



Vocabulary (Give an example of each.)

Population: A collection of objects or group of people about whom information is gathered.

All the beanie babies in the world

Quantitative Data: The data values are measurements or counts.

The price Ms. Shanahan could get for each beanie baby on ebay

Categorical Data: The data is not numerical or it can be divided into categories.

Animal type

Sample: A subset (group) of a given population.

The tub of beanie babies at Mrs Shanahans House

Individual: The objects described by a set of data (including people).

Each beanie baby in Ms. Shanahan's sample

Variable: Any characteristics (including measurements) of the individuals.

For example: price or animal type

Statistic (with a lower-case "s"): A numerical fact or calculation that is used to describe a data set.

The mean price of ms. Shanahan's beanie babies or median, range etc.

Census: The process of measuring every member of the population.

If we could record the price or animal type for every beanie baby in the world

Distribution: A listing or display of all the possible values (or intervals) of the data and how often they occur.

All the values that represent prices of Ms. Shanahan's BBs: in a list, table, or data display.

Making a Histogram

Single-variable numerical data can be represented graphically with a histogram. Follow these steps!

EXAMPLE: Consider the data at right, a list of the heights of 18 Math 3 students.

55
60
61
62
63
64
64
65
65
66
67
69
70
70
71
72
72
76

Sum: 1192

1. Determine how many classes or **bins** (bars) to use (you should have 5-10 bins).

   Bin width: 4

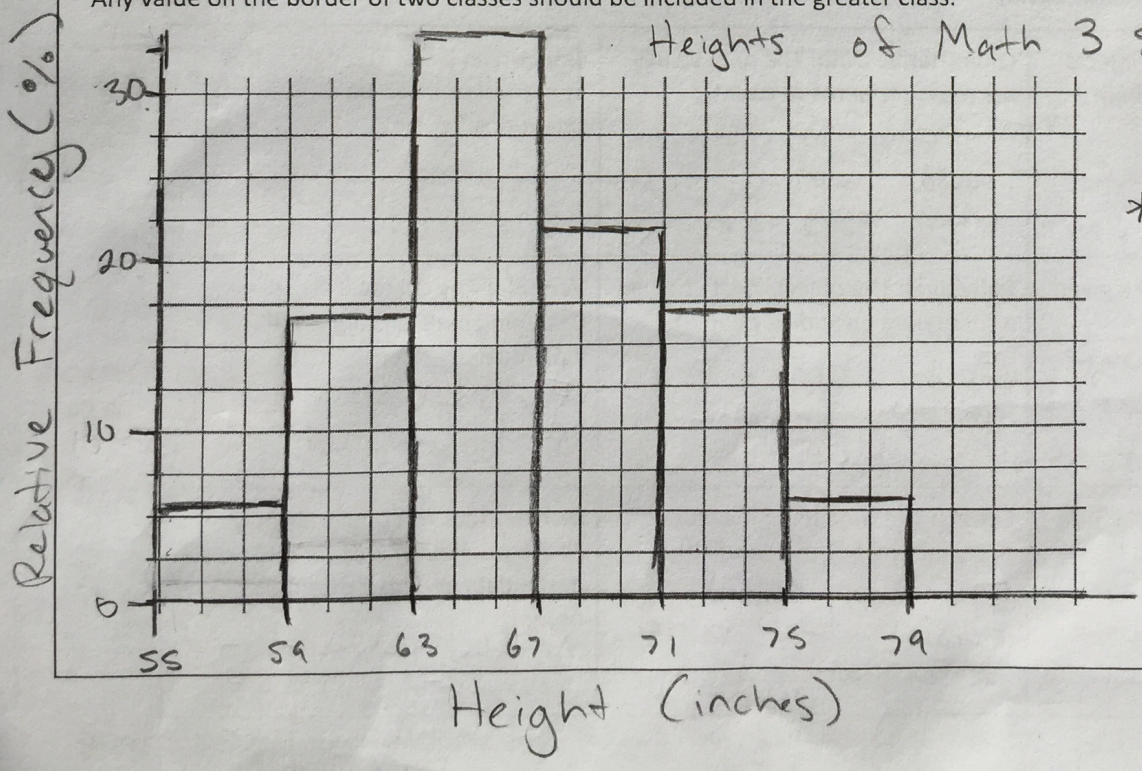
   Range is 21 80
   21 + 1 → 22 makes 6 bins

2. Make a frequency table to organize your data and determine how many values will be in each class (or what percent of the values).

Class (Range of values for a bin)	Frequency (How many values in this class)	Relative Frequency (% of values in this class)
55 to < 59	1	$\frac{1}{18} = 6\%$
59 to < 63	3	$\frac{3}{18} = 17\%$
63 to < 67	6	$\frac{6}{18} = 33\%$
67 to < 71	4	$\frac{4}{18} = 22\%$
71 to < 75	3	$\frac{3}{18} = 17\%$
75 to < 79	1	$\frac{1}{18} = 6\%$

3. Draw and label the axes.
   - The vertical axis should represent the frequency (or relative frequency) and should start at zero.
   - The horizontal axis should represent the variable (height, in this case) and doesn't have to start at zero.
4. Draw the bins so that there are no gaps between them. The height of the bins represents how many data values (or %) are in each class.
5. Title the histogram appropriately.

* Any value on the border of two classes should be included in the greater class.



* Reminders
- No gaps!
- Frequency starts at zero
- horizontal axis does not need to start at zero

MATH NOTES

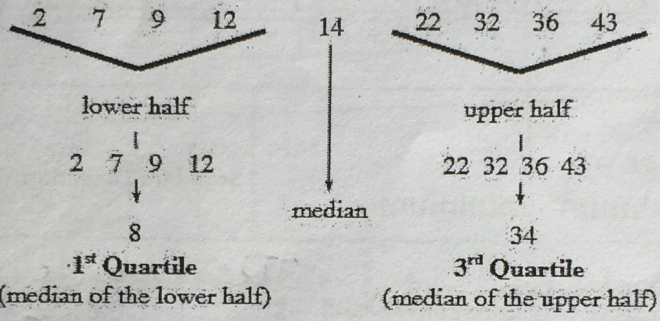
METHODS AND MEANINGS

Interquartile Range and Boxplots

Quartiles are points that divide a data set into four equal parts (and thus, the use of the prefix "quar" as in "quarter"). One of these points is the median. The **first quartile (Q1)** is the median of the lower half, and the **third quartile (Q3)** is the median of the upper half.

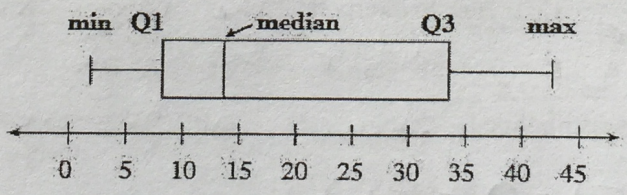
To find quartiles, the data set must be placed in order from smallest to largest. Note that if there are an odd number of data values, the median is not included in either half of the data set.

Suppose you have the data set: 22, 43, 14, 7, 2, 32, 9, 36, and 12.



The **interquartile range (IQR)** is the difference between the third and first quartiles. It is used to measure the spread (the variability) of the middle 50% of the data. The interquartile range is $34 - 8 = 26$.

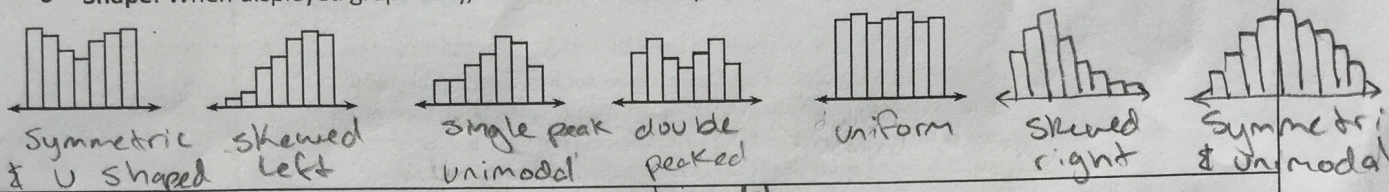
A **boxplot** (also known as a box-and-whisker plot) displays a five number summary of data: minimum, first quartile, median, third quartile, and maximum. The box contains "the middle half" of the data and visually displays how large the IQR is. The right segment represents the top 25% of the data and the left segment represents the bottom 25% of the data. A boxplot makes it easy to see where the data are spread out and where they are concentrated. The wider the box, the more the data are spread out.



5-Number Summary & Summary Statistics	
"5 number summary" includes • Minimum (minX) • 1st Quartile (Q1) • Median (Med) • 3rd Quartile (Q3) • Maximum (maxX)	Summary Statistics from your Calculator: • $\bar{x}$ = mean • Σx = sum of list (checksum!) • Σx^2 = Variance • Sx = Sample Standard Deviation • σx = Population Standard Deviation • n = number of values in list

Describing a Data Distribution

- **Center**: Where are the peaks? **Use:** mean or median
- **Shape**: When displayed graphically, how is the data shaped?



- **Variability**: How spread out is the data? **Use:** Standard Deviation or IQR
- **Outliers**: Are any values far away from all the others? **Remember:** use term apparent or possible or make calculations

Describing Spread

Interquartile Range:	**Range:**	**Standard Deviation:** $S_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$
IQR = Q3 - Q1	**Range = Maximum - Minimum**	

Range
In words: Difference between the min & the max
Common mistake: 100 - 0 = 100
Don't write, "The range goes from 0 to 100"
For the data set below, the range is 100. The data values vary from 0 to 100.

Example: How much change do you have in your pockets right now?

mean = 33.8 SD = 32.4

min. median max.
0, 0, 0, 0, 5, 25, 35, 48, 50, 50, 52, 75, 100
Q1 = 0 Q3 = 51

Interquartile Range (IQR)
In words: The difference between Q3 & Q1
How to interpret: The range for the middle 50% of the data
Best measure of variability when the distribution is skewed or has outliers.
In the data set above, the IQR is 51 - 0 = 51

Standard Deviation
In words: Its kind of like the average distance from each data point to the mean
Sample S.D. is more commonly used than Population S.D. because we rarely have data from the entire population
How to interpret: The amount of change in peoples pockets varies from the mean of ___ by about ___.
Best measure of variability when the distribution is symmetric.

The (context) typically vary from the mean of (mean) by about (SD).