EDUR 7130
Presentation 4b

2. Graphical and Tabular Displays

2a. Frequency Display

A **frequency display** or **frequency distribution** shows a count of each category for a given variable.

Example

If sex of people in this chat session is the variable of interest, then one would just count the number of males and females and show the frequency.

For instance (fictional data below):

Sex	Frequency	Percent
F	8	80
M	2	20

This example shows that there is a count of 8 females and 2 males.

What is the mode for the sex distribution above?

Female represents the mode since it is the most frequent category.

Could we calculate a mean or median for these data (sex distribution in this course)? Recall that the raw data here could look like this:

F, F, M, F, F, F, M, F, F

No, data must have natural rank to calculate mean and median, and sex (male/female) does not have that natural rank (and numbers assigned to categories), so only mode applies as a measure of central tendency as this variable is currently configured.

Here's a frequency display constructed by the statistics program Stata for course averages in educational research one semester:

Final Course Grade Average	Freq.	Percent	Cum.
64	1	3.70	3.70
65	1	3.70	7.41
67	1	3.70	11.11
69	1	3.70	14.81
72	2	7.41	22.22
74	2	7.41	29.63
75	2	7.41	37.04
78	1	3.70	40.74
79	1	3.70	44.44
81	1	3.70	48.15
82	1	3.70	51.85
83	1	3.70	55.56
84	2	7.41	62.96
85	2	7.41	70.37
86	1	3.70	74.07
87	1	3.70	77.78
88	1	3.70	81.48
91	1	3.70	85.19
92	1	3.70	88.89
93	2	7.41	96.30
96	1	3.70	100.00
Total	27	100.00	

http://www.bwgriffin.com/gsu/courses/edur7130/images/1_freq_grades.jpg

How many students scored between 80 and 89, inclusive?

10 students.

What percentage of students scored 90 or above?

Since there are 27 total students, and 5 scored 90 or above, the percentage is about 18.5% (5/27 = .185).

Here is another example of a frequency display. Suppose scores on a test have this frequency distribution:

Scores	Frequency
95	6
94	2
93	1
92	9

What is the mode for these test scores?

Mo = 92 (most frequent score).

What is the median for these data?

To find median, you must first sort the scores in order, then find the middle scores. In this example, the scores laid out would be as follows:

92, 92, 92, 92, 92, 92, 92, 92, >>92 93<< 94, 94, 95, 95, 95, 95, 95, 95

Since there are 18 scores in total, the middle scores would be scores 9th and 10th which I have marked with >> and << symbols above. Here the median would be the mean of (92+93)/2 = 92.5

2b. Stem-and-Leaf

A stem-and-leaf display retains the raw numbers but displays them in such a way as to reveal frequency by showing lengths similar to bar charts and histograms (which we will review below). Consider this frequency display:

Scores	Frequency
97	1
96	2
95	6
94	2
93	1
92	9

A stem-and-leaf display might group the numbers into units of 5 (90 to 94, 95 to 99), like this:

9 | 555555667
9 | 222222222344

Which were the more common scores, those between 90 and 94, or those equal to or over 95?

Those in the range of 90 to 94

Looking at the stem and leaf display, we can see this quickly by looking at the raw numbers, the leaf. The longer the leaf, the more scores reported.

Here is another example of mean final course grades:

```
. stem Grade, digits(2)
Stem-and-leaf plot for Grade (Final Course Grade Average)

  6* | 4579
  7* | 22445589
  8* | 1234455678
  9* | 12336
```

http://www.bwgriffin.com/gsu/courses/edur7130/images/4_stem_leaf_grouped.jpg

How many people had a mean final course grade of 85?

Two people had an average of 85.

Sometimes folks have trouble reading a stem-and-leaf display. Here's how, first, look at this row

8*|1234455678

What does the "8*" represent here?

It represents scores the "80s"

Now, what does this represent: 8*|1

Score of 81

And this? 8*|12

Scores of 81 and 82.

Side note:

In Stata, the software that produced this stem-and-leaf, note that

* indicates scores between 0-4, and . indicates scores between 5-9, for example:

8*|01234
8.|56789

If all are lumped together, then it defaults to 8*

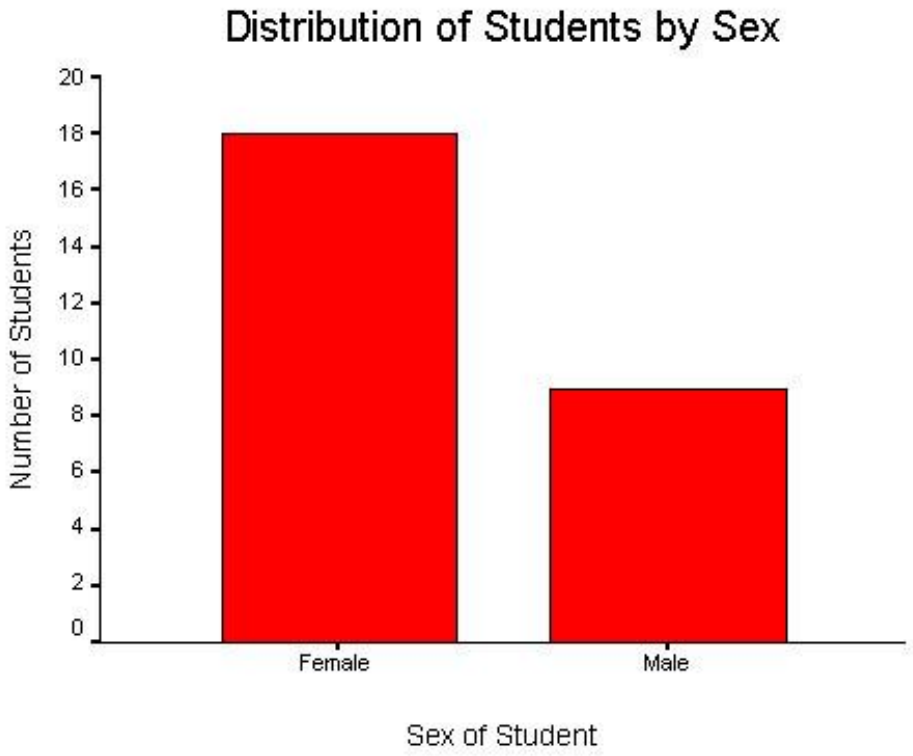
2c. Bar Graph and Histogram

Bar Graph: appropriate for a variable that is qualitative (e.g. showing distribution by sex, frequency of car type, etc.), so the order in which categories of the variable is presented does not matter.

Histogram: appropriate for quantitative variable (e.g. show counts for test scores), so categories should be sorted and presented in order.

Another difference is whether the bars may touch or show a gap between them.

Here is an example of a **bar graph** (bar chart) that shows counts (frequency) of sex distribution within a class:



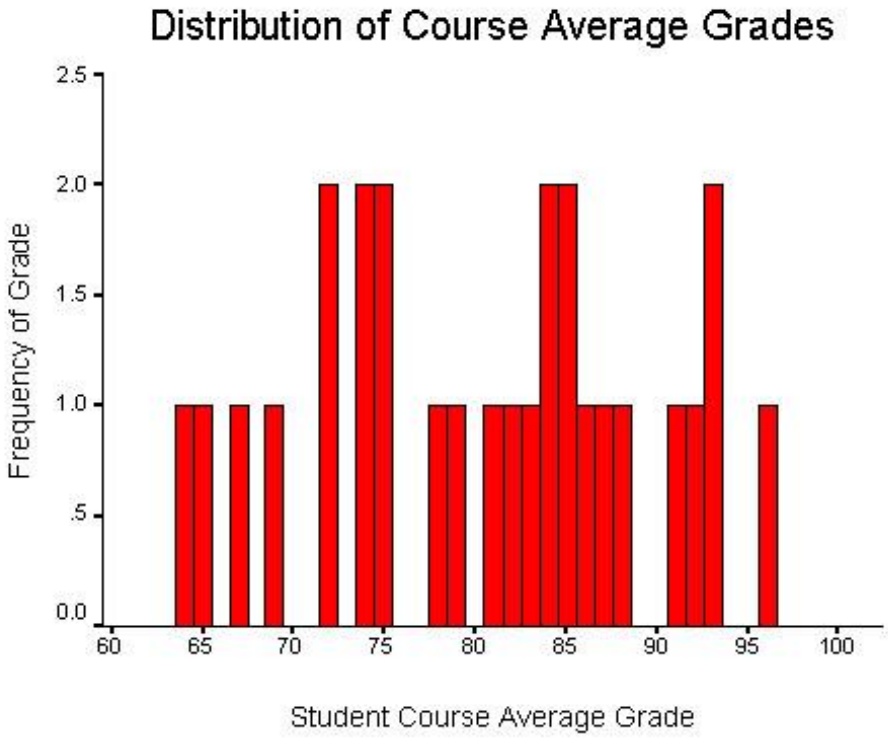
http://www.bwgriffin.com/gsu/courses/edur7130/images/5_bar_chart_sex.jpg

How many males and how many females were present?

About 9 males and 18 females in that course.

Histogram is similar to a bar chart except that with the histogram is developed for quantitative data and as a result the bars may touch (but not always).

The key is that the variable plotted will be quantitative for histogram, and qualitative for a bar chart. For example, here is a histogram for end of course grades:



http://www.bwgriffin.com/gsu/courses/edur7130/images/7_histogram_grades.jpg

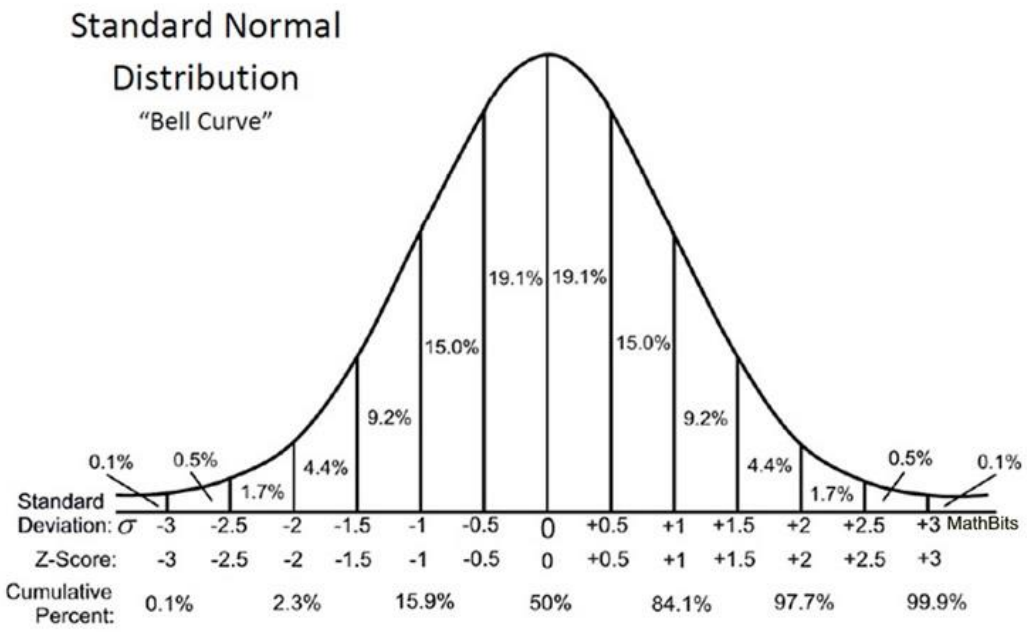
How many averaged an 84 in this course?

Two students averaged 84.

What was the highest average score?

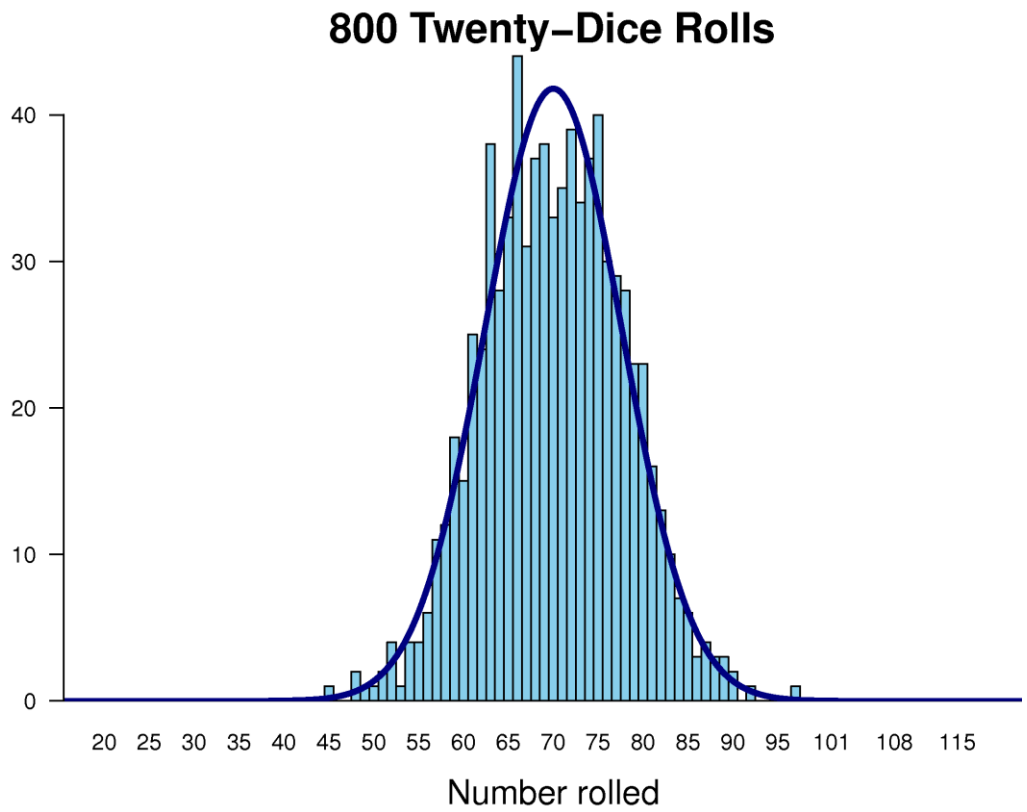
Top is 96.

Standard **Normal Curve** or "Bell Curve" is derived from a histogram, but often with the histogram bars removed. See examples below.



Source: https://mathbitsnotebook.com/Algebra2/Statistics/STstandardNormalDistribution.html

The example below shows a normal curve superimposed on a histogram. The histogram is typical of distributions that are approximately normal – close enough to use the normal distribution tables to calculate percentages and proportions.

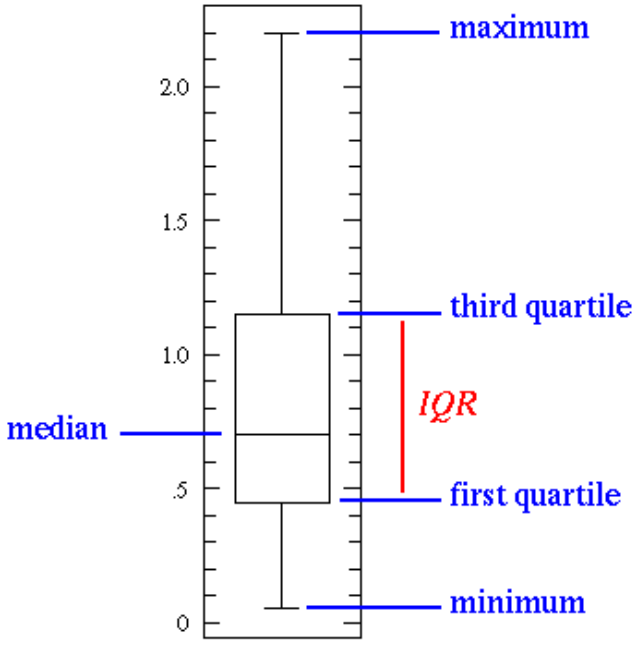


Source: https://scientificgems.wordpress.com/2015/11/03/mathematics-in-action-probability/

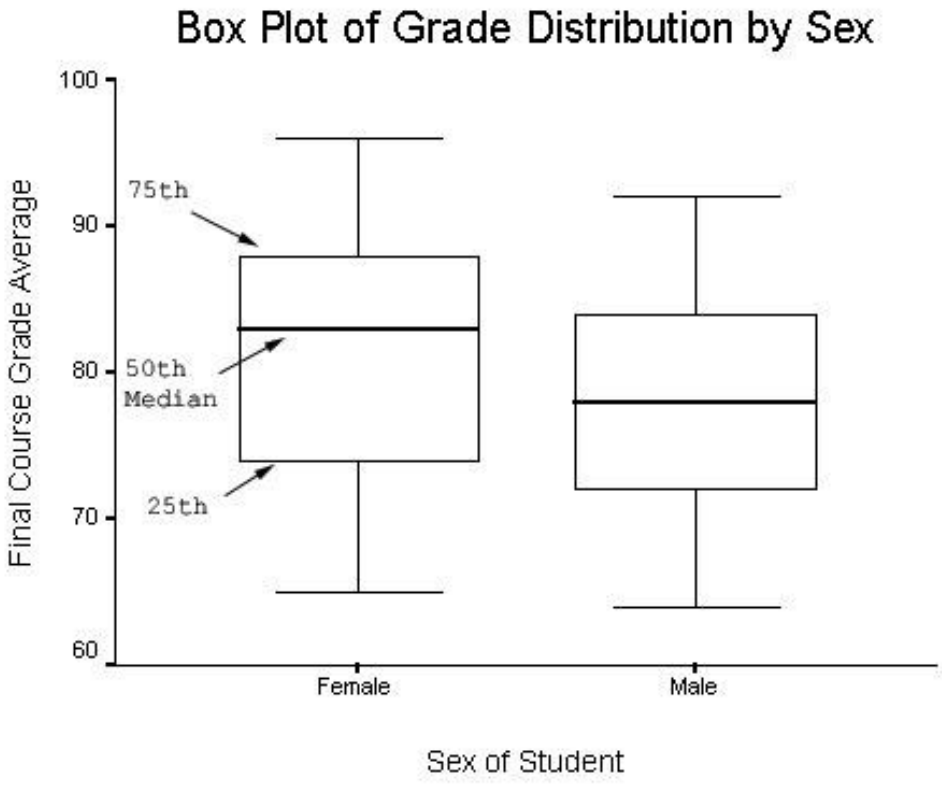
2d. Box Plots (or Box and Whisker displays)

Box plots are used to show range of scores and quartiles. The two whiskers display the extreme minimum and maximum scores, the box shows the 25th and 75th percentile, and the thick link inside the box shows the 50th percentile, which is the median.

Examples



Source: http://www.physics.csbsju.edu/stats/box2.html



Approximately, what was the score for the 50th percentile for females (i.e., what was the median score for females)?

The median score is denoted as the thick line in the middle of the box (on the graphic is 50th percentile or median mark denoted). For females this line appears to align with a score of about 83 or 84.

Females at the 25th percentile had a score of about what?

About 74.

About what percentage of females scored between about 87 (or 88) and 73 (or 74)?

About 50%, how was this determined?

The box shows the 25 to 75th percentile, and 75 – 25 = 50, or 50% of all scores.

What appears to be the median score for males?

About 77 or 78.

Males at the 25th percentile scored about what in the course, and males at the 75th scores about what?

Males at the 25th percentile seem to have a course average of about 72 or 73. Males at the 75th percentile seemed to score about 83 or 84.

What do the whiskers on the box-plot represent?

The whiskers represent the upper and lower ranges for scores. Sometimes the upper whisker represents the 90th percentile and the lower the 10th percentile. With some software, the upper whisker indicates the highest score obtained while the lower whisker indicates the lowest score obtained.

For males, the score corresponding with the lowest whisker is about what?

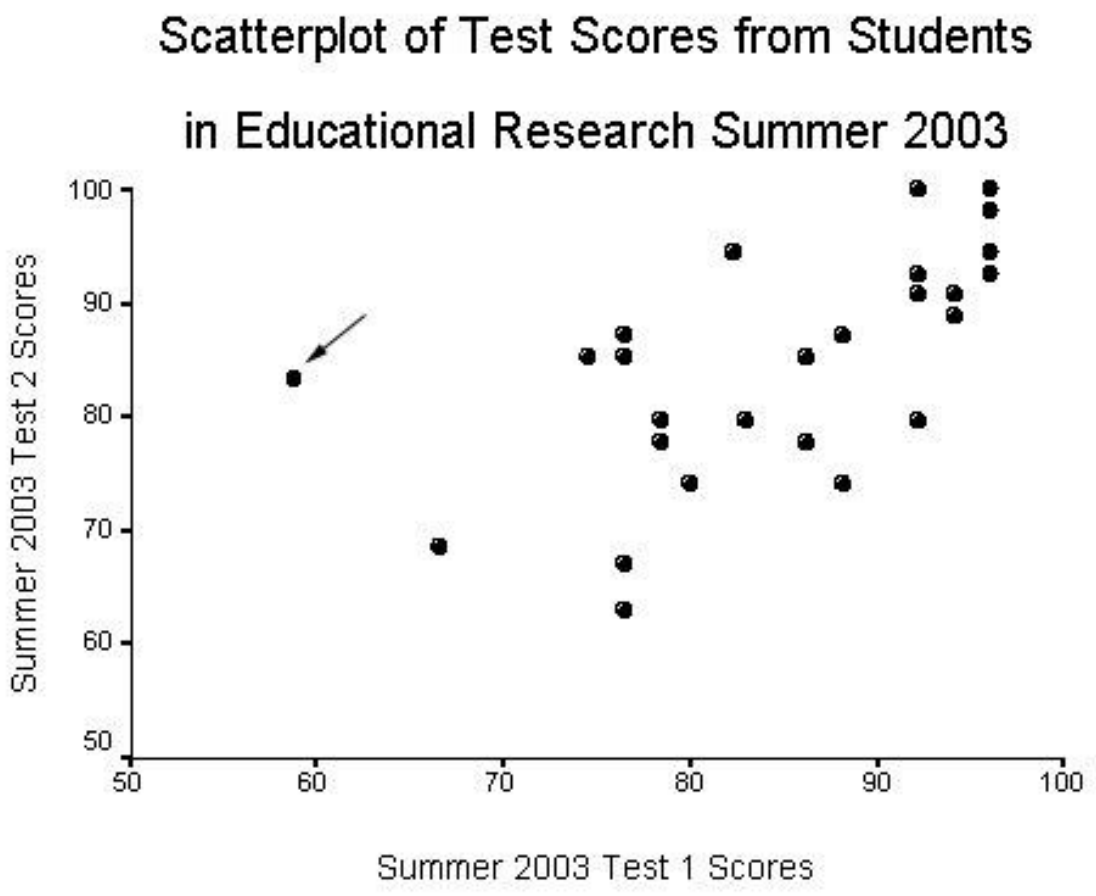
The lowest whisker shows males scoring at about 63 or 64.

Scatterplot

A scatterplot or scattergram shows how data from two variables correspond.

The points on a scatterplot represent the combination of scores for a given individual for two variables.

Here's an example scatterplot showing scatter of scores between Test 1 and Test 2 in educational research:



Note the point highlighted by the arrow. What score did this student obtain on Test 1 and Test 2?

For Test 1 this student's score seems to be about 58 or 59, and for Test 2 this student increased his or her score to about 83 or 84.

Notice the four students on the extreme right – what were there scores for Tests 1 and 2?

For Test 1, all four scored the same thing – about 96 or 97. For Test 2, their scores seem to range between 93 and 100.

Does this scatterplot show a positive or negative relationship between these two sets of scores?

A positive relationship.

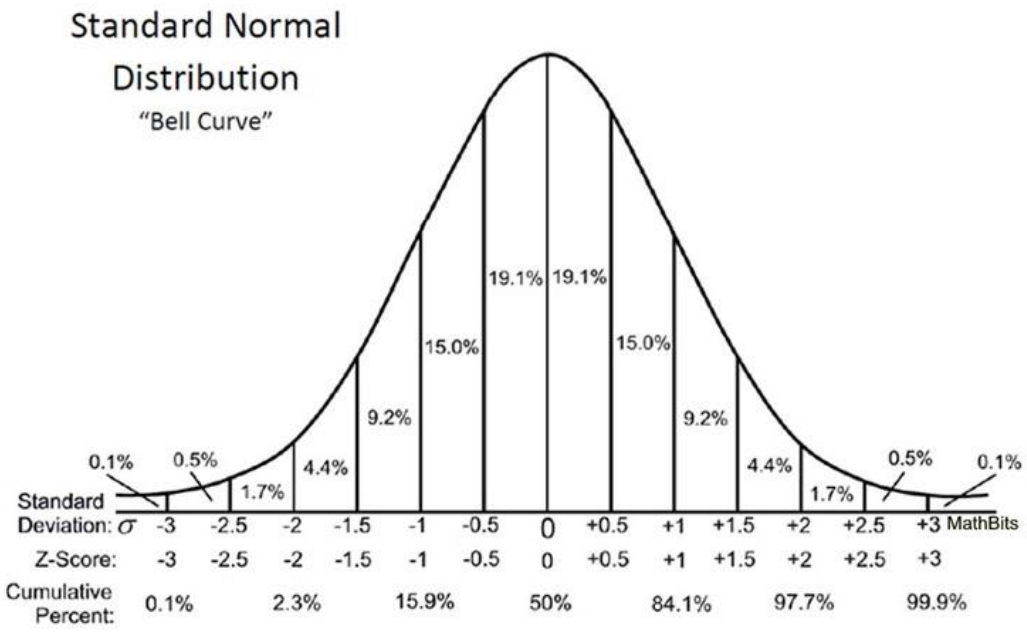
How can we determine that it appears to be a general trend toward a positive relationship?

Students who scored lower on Test 1 tended to score lower on Test 2. Similarly, those who scored higher on Test 1 tended to score higher on Test 2. This is a positive relationship.

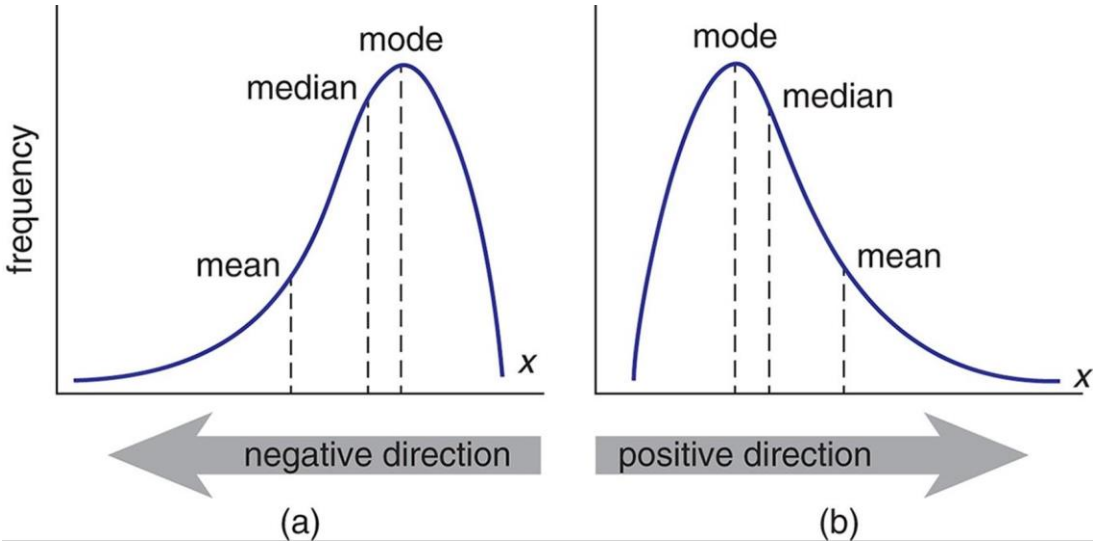
Also, if one were to draw a line through this graph to show the general pattern of points, the line would rise over the range of Test 1 values --- as Test 1 scores increase, the line would also increase.

Skew

Recall that the three measures of central tendency I covered include mean (M), median (Md), and mode (Mo). In a normal distribution the M, Md, and Mo are all the same.



Skew is created when a distribution has extreme scores all located in one tail of a distribution.



Source: http://schoolbag.info/physics/physics_math/71.html

More resources on normal and skewed distributions:
http://www.mathsisfun.com/data/skewness.html
http://www.mathsisfun.com/data/standard-normal-distribution.html

With skewed distributions, the mean will be pulled toward the skewed data, the median might show some pull, but not by much, and the mode will stay in the most frequent area.

Question: In a class of 20 students 17 students are 19, 20, 21, or 22 years of age, and 3 students are in their 40s and 50s, which measure of central tendency should be used? Is there likely to be a difference in the values obtained for the various measures of central tendency for these scores?

This distribution is not normal – it is skewed because of the few older students – their ages will skew the mean higher than normal. For skewed data, sometimes the median is preferred, but best to report all measures and let readers decide which measure they find most useful.

For example, here is a set of data that is skewed due to the large number:

1,1,1,1,1, 100

What is the M, Md, and Mo for this set of data? And how does the 100 affect these measures of central tendency?

M = (1+1+1+1+1+100) / 6
= 105 / 6
= 17.5

Note that the mean is affected by the one extreme score; the 100 score skews the distribution and makes the mean larger than it would be otherwise. The median is unaffected by the skew, it remains at Md = 1. The mode is also 1.

Try this example: Which measure of central tendency would one use to describe race/ethnicity (Black, Latino, etc.)?

Only mode works here since this variable is qualitative/nominal/categorical.

If the categories of the variable cannot be ranked, then the variable is nominal (qualitative) and therefore no rank exists. As a result, the median and mean are inappropriate measures of central tendencies since they assume rank among the categories, so only mode would work for qualitative variables. In short, median and mean require categories to be ranked, mode does not require rank.