Organizing Categorical Data

Know how to create a graphical representation of data with two variables (data in a two way table), (10, 6a; 09, 1a). Know how to describe differences in categorical data? (10, 6b; 09, 1b)

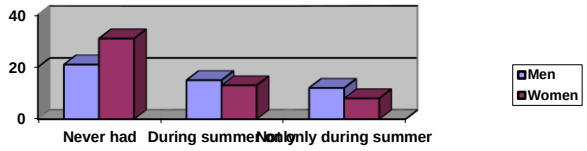
A simple random sample of 100 high school seniors was selected from a large school district. The gender of each student was recorded, and each student was asked the following questions.

1. Have you ever had a part-time job?
2. If you answered yes to the previous question, was your part-time job in the summer only?

The responses are summarized in the table below.

Job Experience	Male	Female	Total
Never had part-time job	21	31	52
Had part-time job during summer only	15	13	28
Had part-time job but not only during summer	12	8	20
Total	**48**	**52**	**100**

(a) Construct a graphical display that represents the association between gender and job experience for the students in the sample.

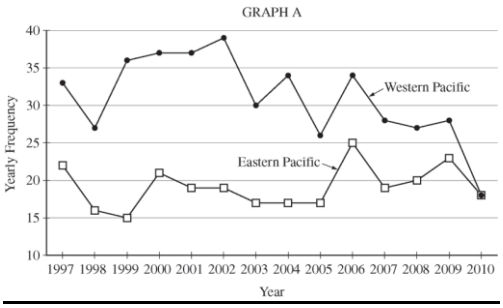


(b) Write a few sentences summarizing what the display in part (a) reveals about the association between gender and job experience for the students in the sample.

The females in the sample were more likely than males to have never had a part-time job and were less likely than males to have had a part-time job during summer only or to have had a part-time job not only during summer. Based on thedata in the graph, there appears to be a relationship between job experience and gender.

Know how to analyze timeplots and compare the center and spread of timeplots, (13, 6ab)

Tropical storms in the Pacific Ocean with sustained winds that exceed 74 miles per hour are called typhoons. Graph A below displays the number of recorded typhoons in two regions of the Pacific Ocean—the Eastern Pacific and the Western Pacific—for the years from 1997 to 2010.



Compare the distributions of yearly frequencies of typhoons for the two regions of the Pacific Ocean for the years from 1997 to 2010.

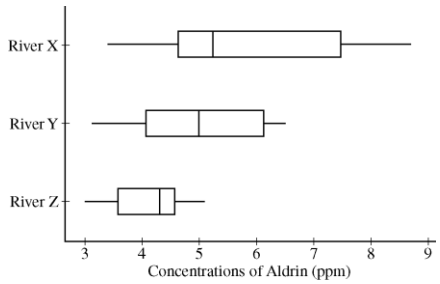
The Western Pacific Ocean had more typhoons than the Eastern Pacific Ocean in all but one of these years. The average seems to have been about 31 typhoons per year in the Western Pacific Ocean, which is higher than the average of about 19 typhoons per year in the Eastern Pacific Ocean. The Western Pacific Ocean also saw more variability (in number of typhoons per year) than the Eastern Pacific Ocean; for example, the range of the frequencies for the Western Pacific is about 21 typhoons and only 10 typhoons for the Eastern Pacific.

QUANTITATIVE DATA

Know how to compare distributions (Shape, center, spread)? Make sure you use comparative words (larger, wider, etc), (12, 3a; 11b, 1b; 10b, 1a; 08, 1a; 08b, 1a; 07, 1b; 06, 1a; 05, 1a;05b, 1a; 04b, 5a)

As a part of the United States Department of Agriculture's Super Dump cleanup efforts in the early 1990s, various sites in the country were targeted for cleanup. Three of the targeted sites—River X, River Y, and River Z—had become contaminated with pesticides because they were located near abandoned pesticide dump sites. Measurements of the concentration of aldrin (a commonly used pesticide) were taken at twenty randomly selected locations in each river near the dump sites.

The boxplots shown below display the five-number summaries for the concentrations, in parts per million (ppm) of aldrin, for the twenty locations that were sampled in each of the three rivers.



Compare the distributions of the concentration of aldrin among the three rivers.

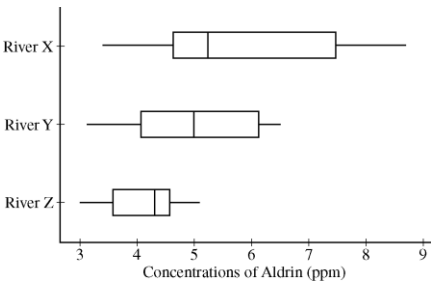
When comparing distributions, you would need 3 things

1) Talk about shape, center, and spread
2) Use COMPARATIVE language for ALL THREE
3) Put an answer in context

We can see when looking at the two rivers that they all have **different** shapes. River X is skewed right, River Y is approximately symmetric, and River z is skewed to the left. River X has the **largest** median, while river Z has the **smallest** median. River X also has the **largest** variability, while river Z has the **smallest** variability.

Know how to create (including labels) and interpret a boxplot, including how to find the range, interquartile range, 25th/75th quartile, and how to determine which way, if, it's skewed (09b, 1a; 04, 1ab; 04b, 5a; 03, 1ab; M7, 29; M2, 14)

The five number summary shows the…
– Min, Q1, Med, Q3, Max
A boxplot is a visual representation of a five number summary, and can be used to show/determine skewness. Outliers are marked with stars



Looking at the graphs on the left, we could say that 75% of the observations from river X are greater than 4.8 ppm.

We could also say that 50% of the observations from river X are greater than the largest measurement from river Z. This is because the median of river X is higher than the maximum for river Z.

Know how to make and interpret a stemplot or back to back stemplot, including labels and a legend, (13, 1; 10b, 1b; 07b, 1a; M7, 7). Know how to describe a distribution (shape, center, spread) when looking at a stemplot, boxplot, dotplot, and other types of graphical representations of quantitative data. What is a gap, and what are clusters? Remember to state where they are in the distribution and what that means in context of a problem, (10b, 1c;07b, 1c)

The Better Business Council of a large city has concluded that students in the city's schools are not learning enough about economics to function in the modern world. These findings were based on test results from a random sample of 20 twelfth-grade students who completed a 46-question multiple-choice test on basic economic concepts.

0|8 9
1|0 2 3 4
1|6 6 7 8 8 9 9
2|
2|
3|3 **1|2 = 12 questions correct**
3|5 6 8
4|1 3 4

Use the stemplot to describe the main features of this score distribution.

Normally, we would talk about shape center and spread. This is a special case, where center and spread are replaced by the fact that there is a gap in the distribution. If you say it is bimodal, that is enough to cover the shape, but you also need to note that it is bimodal with a gap. You also should note where the data are.

Answer: The distribution of grades is bimodal with a gap in the 20's. There are two clusters of students. One group got between 8 and 29 questions correct, while the other got between 33 and 44 correct.

Know how to create a dotplot, (08b, 1a)

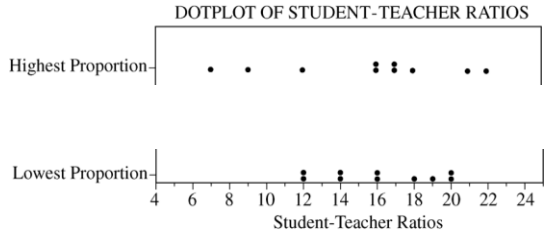
A certain state's education commissioner released a new report card for all the public schools in that state. This report card provides a new tool for comparing schools across the state. One of the key measures that can be computed from the report card is the student-to-teacher ratio, which is the number of students enrolled in a given school divided by the number of teachers at that school. The data below give the student-to-teacher ratios.

Ratios in the 10 Schools with Highest Proportion of Students Meeting Standards
7 21 18 22 9 16 12 17 17 16

Ratios in the 10 Schools with Lowest Proportion of Students Meeting Standards
14 16 18 20 12 14 16 12 20 19

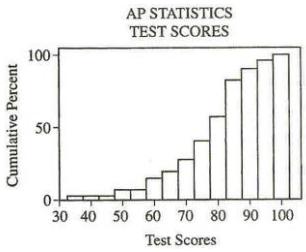
Dotplots are used when you have a small amount of data, and, like a stemplot, can display every piece of data collected

(a) Display a dotplot for each group to compare the distribution of student-to-teacher ratios in the top 10 schools with the distribution in the bottom 10 schools. Comment on the similarities and differences between the two distributions.



They **both** appear to be approximately symmetric and also appear to **both** have a median around 15 pupils per teacher. The biggest difference appears to be that the schools with the highest proportion of students meeting standard have a **larger** variability in student-teacher ratios that those with the lowest proportion.

Know how to read and interpret a cumulative frequency or cumulative relative frequency plot and interpret points and slope in context, (06b, 1; M2, 27)



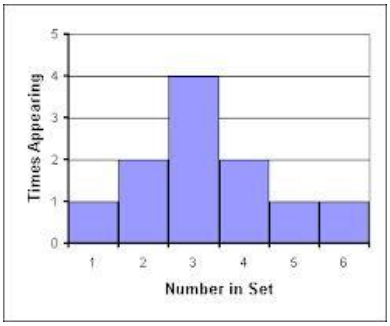
The figure above shows a cumulative relative frequency of 40 scores on a test given to an AP Statistics class. Which of the following conclusions can be made from the graph.

(A) There is greater variability in the lower 20 test scores than in the higher 20 test scores.
(B) The median test score is less than 50.
(C) Sixty percent of the students had test scores above 80.
(D) If the passing score is 70, most students did not pass the test.
(E) The horizontal nature of the graph for test scores of 60 and below indicates that those scores occurred most frequently

Cumulative plots show the numbers up to and including that number. For example, about 55% of students scored 80 or less. There were no students how scores in the 55-60 range, because there was not growth. The group with the most number of students is the 85-90 group, because it has the largest jump (about 30% of the population).

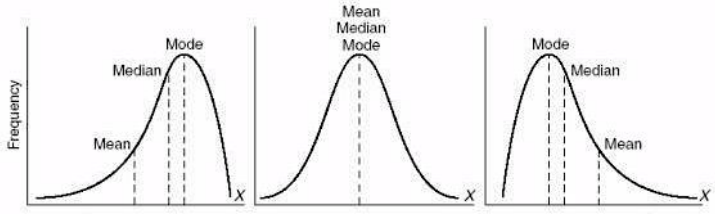
The answer to the above question would be A. Since the median is around 80, about 50% of students scored above 80. There is more variability in the bottom half, because the bottom 50% goes from 30 to 80, while the top 50% goes from 80 to 100. Since 70 has a cumulative frequency of about 25%, that means that 75% passed, so D is not correct. The horizontal nature below 60, means that there wasn't any growth, removing E as an option.

Know how to find the five number summary when looking at a histogram, (11b, 1a)



A Five Number Summary consists of the minimum, maximum, medium, first and third quartiles. On this example histogram, the minimum is 1 and the maximum is 6. From this graph, we can see that there were 11 observations. The median (6^{th} observation) would be 3, the first quartile (3^{rd} observation) would be 2, and the third quartile (9^{th} observation) would be 4.

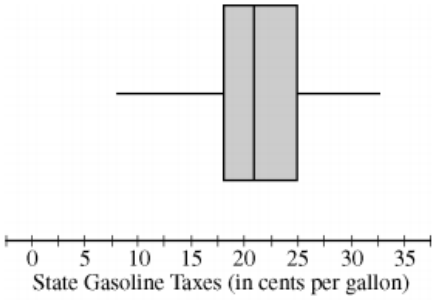
Know how skewness and symmetry affect the relationship between mean and median, (11b, 1c; 09, 6b; 09, 1c; 05, 1d; 05b, 1b)



A graph can be skewed left, skewed right or symmetric. When a graph has a right skew, the tail goes to the right. When the tail goes to the left, it is sad to have a left skew. When both sides are even, it is symmetric. The mean gets dragged in whichever direction the skew goes in. The median also is dragged in the direction of the tail, but not as far. When a distribution is symmetric, the median and the mean should be the same.

Know what effect adding a constant has on the center and spread of a set of data. Know what multiplying a constant has on the center and spread of a set of data, (09b, 1b;09, 1b; M2, 7)

As gasoline prices have increased in recent years, many drivers have expressed concern about the taxes they pay on gasoline for their cars. In the United States, gasoline taxes are imposed by both the federal government and by individual states. The boxplot below shows the distribution of the state gasoline taxes, in cents per gallon, for all 50 states on January 1, 2006.

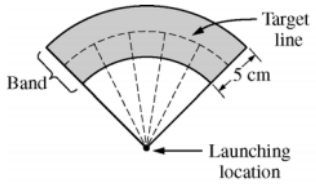


The 5 number summary for this graph (FRQ from 2009B #1b) is 7, 18, 21, 25 and 32. If, say, all numbers has 2 cents/gallon added to them, the 5 number summary would then be 9, 20, 23, 27, and 34. The median goes up by 2 but the range stays the same. If we doubles the price of gasoline, however, then the new 5 number summary is 14, 36, 42, 50, and 64. The range is now 50, which is 2 times the original range.

Adding or subtracting a constant changes the center (mean, median), but it does NOT change the spread (average,IQR)

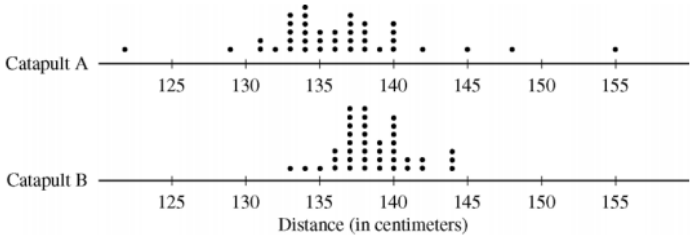
Multiplying or dividing a constant changes both the center and the spread.

Two parents have each built a toy catapult for use in a game at an elementary school fair. To play the game, students will attempt to launch Ping-Pong balls from the catapults so that the balls land within a 5-centimeter band. A target line will be drawn through the middle of the band, as shown in the figure below. All points on the target line are equidistant from the launching location.



If a ball lands within the shaded band, the student will win a prize.

The parents have constructed the two catapults according to slightly different plans. They want to test these catapults before building additional ones. Under identical conditions, the parents launch 40 Ping-Pong balls from each catapult and measure the distance that the ball travels before landing. Distances to the nearest centimeter are graphed in the dotplots below.



Know what the standard deviation or overall spread of a distribution will look like for a set of data that is more consistent than a different set of data, (06,1b)

In some data, the standard deviation of the set is greater than in other data. When data is further spread out, the standard deviation is greater. When the data is more consistent and close to the mean, the standard deviation is smaller. In the FRQ above, Catapult A has less consistent findings than catapult B, and so, has a greater standard deviation.

Know how to interpret center in the context of a problem, (06, 1c)

The center of the data can be interpreted as the mean, median, or mode, depending on the context of the problem. Median gives the number in the direct middle of the set, mode gives the number most often gotten, and mean is the average score of all the numbers. In the FRQ, the mean, median and mode of the data sets for Catapult B are all the same (the mean as 138.6, median as 138, and mode as between 137 and 138) making all the numbers useful and accurate in determining where the center of the data is. Sometimes center is not an appropriate number to use. That would occur in a problem where the center lies in a gap, such as a bimodal graph where the average lies in between the two clusters. In that case, average is not representative of anything "typical" about the data.

Know what a median is, and what it represents in the context of a problem, (M7, 1)

The median of a set of data is the number directly in the middle of the numbers. 50% of the data lies above the median, and 50% of the data lies below it. In the FRQ, the median is 138 because it is in the middle of all of the numbers given.

Know how to interpret standard deviation in the context of a problem, (07, 1a)

Standard deviation is the typical distance away from the average. So, let's say that a school has an average SAT Math score of 560 with a standard deviation of 60. That means

Typically, students scored about 60 points away from the average of 560.

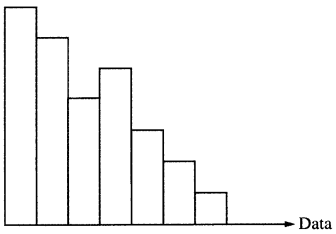
Or

The average deviation from 560 for students at this school was around 60 points.

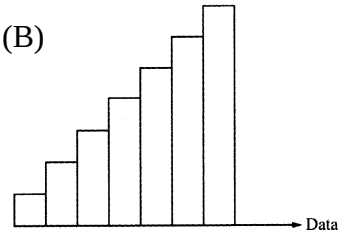
Know how to compare standard deviations when looking at two or graphical displays (If you have two histograms, know which one has a larger standard deviation), (M7, 15)

The histograms below represent the distribution of five different data sets, each containing 28 integers, from 1 through 7, inclusive. The horizontal and vertical scales are the same for all graphs. Which graph represents the data set with the largest standard deviation.

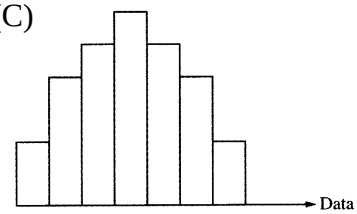
(A)



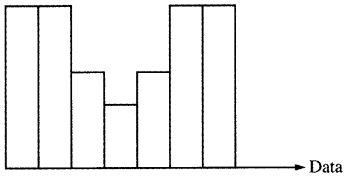
(B)



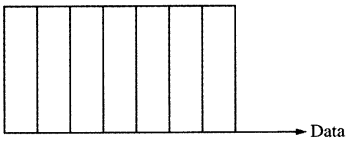
(C)



(D)



(E)



Standard deviation measures the typical distance from the average. So, if we are trying to find the graph that has the largest distance from the average, graph D is the best choice. D is the best, because it only has a few observations near the average, while most of the other observations are far away from the average.