Vocabulary: *Samples and Populations*

Concept	Example
Different types of data • **Categorical data** results when the question asked in a survey or sample can be answered with a **non-numerical** answer. For example if we are investigating the visitor parking situation on a college campus then three issues that might come up are: location (commuter or central), payment method (free, metered. attendant), and type (on the street, in ramps, in lots). The three issues (location, payment method, and type) are called **attributes** of each parking spot. For each parking spot the investigator decides on the appropriate answer for "Location?" and may choose the answer "Central." In response to the question "Payment method?" the investigator may choose "metered." In response to "Type" the investigator might choose "ramp." **The answers are all non-numerical.** The **non-numerical values of one attribute** would go on the horizontal axis of a bar graph that summarizes information about that particular attribute. For example, "free," "metered,"	1. *Which of the following questions will result in **numerical data** and which in **categorical** data?* *a. Do you have any brothers and sisters?* *b. How many children live in your house?* *c. Do you live with both mother and father, mother only, father only, mother and stepfather, father and stepmother, grandparent, other?* *d. What grade are you in?* *e. How many years have you been in school?* *f. How many students are in your class?* *g. How long is the school day?* *h. Which day is your favorite day of the week?* *i. How far do you travel to school?* *j. How do you travel to school?* *k. What was your score on the last mathematics test you took?* *l. What was your grade in the last mathematics class you took?* a. A person might answer "yes." Categorical. b. A person might answer "3." Numerical. c. A person might answer "mother only." Categorical. d. A person might answer "7." Categorical. (This one is tricky. "7" is the *name* of the grade, which makes one think that the data is "categorical." But some might argue that "7" is a measurement of how many years of education, in which case one might think that the data is numerical. Depends on how you think that "7" is going to be used, as a category or as a measurement.) e. A person might answer "12." Numerical. f. A person might answer "32." Numerical. g. A person might answer "6.5 hours." Numerical. h. A person might answer "Friday." Categorical. i. A person might answer "3.1 miles." Numerical.

and "attendant" would be the values for the attribute "Payment method." • **Numerical Data** results when the question asked in a survey or sample is answered with **a number or measurement.** For example in the campus parking situation the **attribute** may be "Cost to park all day." Then for each parking spot the investigator asks "How much to park all day?" and records a numerical answer. These **numerical values of the attribute** "Cost to park all day" are arranged in a numerical scale on the horizontal axis of a bar graph summarizing information about daily costs.	j. A person might answer "by bus." Categorical. k. A person might answer "87%." Numerical. l. A person might answer "B." Categorical.
A Distribution of Data is just a data set, from which we want to extract information about the set *as a whole*. Students learn to summarize a data distribution by giving measures of center and variability, by making graphical displays, and by describing a graphical display. **Measures of Center** are **median, mean and mode** (See *Data Distributions*). We only use these measures with numerical data. (*Mode* can also be used with categorical data.)	2. *The following might be coach ticket prices paid for a flight from Lansing to Chicago by different passengers on the same flight. Make an appropriate graph and describe the distribution of ticket prices. (Note: there are many attributes that could be recorded about each ticket including some non-numerical data about gender of passenger, when the ticket was purchased, whether the passenger is a frequent flyer. Here we are only interested* ***in numerical data*** *about price.)*

Measures of Variability indicate how clustered together or spread out the individual data values are. The measures known to students are **range** (See *Data Distributions*) and **interquartile range** (see box plots below). Students may also use the existence of clusters, gaps and outliers to describe how clustered together or spread out the data set is. We only use these measures with numerical data.

Causes of Variability
Some variability might stem from error inherent in the process (such as repeatedly measuring the height of an 8th grade female student and getting slightly different answers), and some from natural variation (such as measuring the heights of several 8th grade female students).

A **Description of a Distribution** should include

- **Summary statistics** (mean, median or mode, range, interquartile range)
- **Shape** of the distribution (symmetric, uniform or rectangular, skewed to right or left, clusters and gaps, outliers)

Price ($)	Number of passengers
180	4
200	3
210	5
220	2
240	8
260	1
270	2
310	7
320	1
330	1
350	2
380	4
390	6
410	6
690	2

The distribution below shows 3 **clusters** or peaks of data. There is a group of "cheap" tickets around $240 to $260. There is another cluster around $300. There is a third group, of "expensive" tickets, around $400. There are small gaps between these groups, and a larger more **significant gap** between the **outliers** at $690 and the rest of the data. The distribution seems to balance around $300, so the **mean** must be around $300 perhaps a little higher because of the outliers. So a typical ticket price is about $300 with a wide **range** from $180 to $690.

Types of Graphical Displays

- **Line Plot** In a line plot, each case is represented as a dot (or an "∉") positioned over a labeled number line. A grouped dot plot shows data grouped in equal-width intervals. Every individual

*3. Make a **histogram** of the data about prices of plane tickets in example 2.*

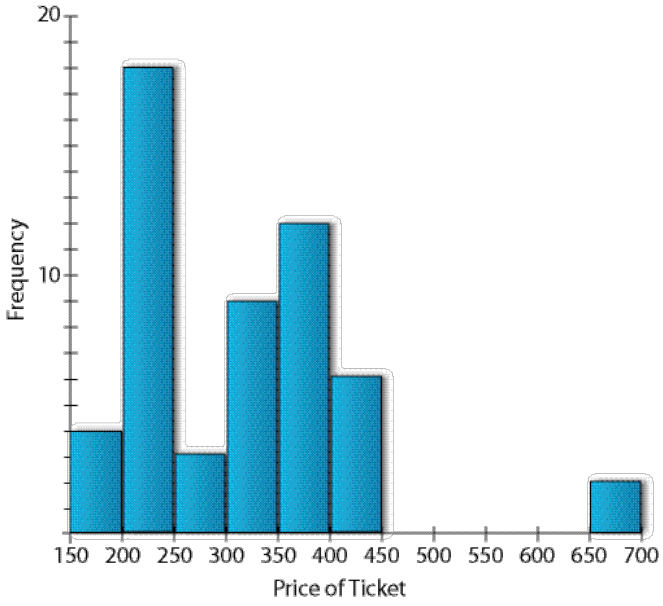
The first thing that needs to be decided is the size of each interval. Generally we try to choose a scale and intervals so that there will be 8 – 10 intervals in all. The goal is to try to show the general shape without getting overwhelmed with details.

data value can be read. We can see if the distribution is approximately symmetric or skewed. We can see clusters and outliers and range.

- **Histogram** The data set is ordered and grouped into i**ntervals**. The size of the bar over each interval shows the frequency of data values in that interval along the range of data values (how often data values in that interval occurred); frequencies may be displayed as counts or as percentages. Individual data values cannot be read. Generally we try to choose a scale and intervals so that the data set is divided into 8 – 10 intervals. With fewer intervals we may lose details about gaps. With more intervals we may lose information about clusters. We can see if the distribution is approximately symmetric or skewed. We can see clusters, outliers and gaps, and range. We can approximate the mean by looking for the balance point of the distribution.

- **Box Plot** The data set is ordered and divided into **quartiles**. The **1st quartile** is the value that divides the bottom 25% of the data from the next 25%. The **second quartile is the median**. The **3rd quartile** is the value that divides the top 25% of

Price ($)	Number of passengers
150-200	4
200 - 250	18
250 - 300	3
300 - 350	9
350 - 400	12
400- 450	6
450-500	0
500- 550	0
550 – 600	0
600 – 650	0
650 – 700	2



Notice that the data value "$200" (which occurs 3 times) is placed in the interval 200 – 250, not in the interval 150 – 200. This is a convention for making histograms so that there is no confusion about where data values are placed.

*4. Make a **box plot** showing the distribution of prices of plane tickets in example 2.*

To make a boxplot we start with ALL the data values in order. There are 54 pieces of data:

180, 180, 180, 180, 200, 200, 200, 210, 210, 210, 210, 210, 220, **220**, 240, 240, 240, 240, 240, 240, 240, 240, 260, 270, 270, 310, 310,|| 310, 310, 310, 310, 310, 320,

the data from the rest. (Sometimes we have to average two adjacent data values to find a quartile marker.) The data below the 1st quartile is shown as a line or "whisker" connecting the minimum to the 1st quartile. The data between the 1st and 3rd quartiles is shown as a box, with the position of the median marked. And the data above the 3rd quartile is shown as a line or "whisker" connecting the 3rd quartile to the maximum. We can see if the distribution is approximately symmetric or skewed. We can see range and outliers. We can see the median, and we can tell how closely clustered the middle 50% of the data is. Gaps cannot be seen. The difference between the 1st and 3rd quartiles tells us the range of the middle 50% of the data, the **interquartile range (IQR)**. This is a measure of variability. (A small IQR indicates that the middle 50% of the data was closely clustered, giving us a small range for what we might call "typical." A large IQR indicates that the middle 50% is widely spread, indicating that it is difficult to say what is "typical" for the data set.)

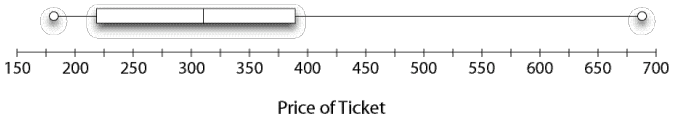
- **Scatter Plot** The relationship between two different attributes is explored by plotting

330, 350, 350, 380, 380, 380, 380, **390**, 390, 390, 390, 390, 390, 410, 410, 410, 410, 410, 410, 690. 690.

Then we divide the list into 4 groups, each of which holds about 25% of the data. First we find the **median**, in this case the value in the middle between the 27th and 28th data values. This is between 310 and 310, so 310 is the median value. (Marked with a double vertical line above. See *Data Distributions* for more about medians.) Next we divide the first 27 pieces of data into 2 equal groups; the marker that makes this division is the 14th data value, or 220 (bolded above). This **value or marker is called the 1st quartile.** Lastly we divide the upper 27 values into 2 equal groups; the value that does this is 390, bolded above. This **value or marker is called the 3rd quartile.** Note: the median is also the 2nd quartile.

The **5-number summary of this data** is:
Minimum = 180
1st quartile = 220
Median = 310
3rd quartile = 390.
Maximum = 690

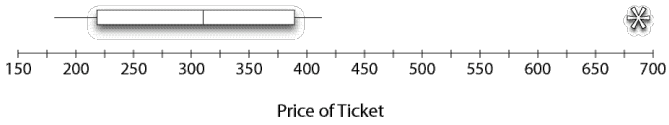
From this summary we make a **boxplot**. The labeled scale is necessary so we can read the box plot.



Sometimes we make the box plot show whether extreme values are **outliers** or not. **The rule is: if a value is 1.5 IQR's above the 3rd quartile or 1.5 IQR's below the 1st quartile then the value is considered unusual enough to be an outlier.** $690 dollars definitely LOOKS far from the other data values.
IQR = 3rd quartile – 1st quartile
= 390 – 220
= 170.
1.5 x 170 = 255.
390 + 255 = 645.
The values at 690 are more than 1.5 IQR's above 390, so they are outliers. We show this as:

numerical values of two attributes on a Cartesian coordinate system. If a pattern appears then we see that there is some relationship that will permit us to predict the value of one variable given the value of another.

Note: students have been calling any data value that seemed distant from the rest of the distribution an "**outlier**." The IQR (interquartile range) gives a measure for how distant a data value has to be. If a data value is 1.5(IQR) above the 3rd quartile or 1.5(IQR) below the 1st quartile then it is considered an outlier. Sometimes the outliers are marked with asterisks instead of connecting them to the rest of the data by extending the "whisker."



5. Each of the graphical displays in examples 2, 3 and 4 show the same distribution of prices. For each of these graphs answer the following questions:

a. *Does the graph show **individual values**?*
b. *Can we see **gaps**?*
c. *Does graph show **overall pattern**?*
d. *Does graph show **outliers**?*
e. *Does graph show a **measure of variability**?*
f. *Does the graph show a **typical** price?*

a.

- The line plot shows every individual value.
- The individual data values are blurred into the interval groups (or quartile groups) in the histogram (or boxplot).

b.

- We can see the gaps in the line plot (for example at $320 and $360 and $420 - $680).
- The only gap that is visible on the histogram is the large gap between $450 and $650. Smaller gaps are obscured by the interval groups. This is not a bad thing; small gaps are not necessarily significant to the over all pattern or shape of the distribution.
- Boxplots generally show NO gaps, unless an outlier exists (as in the second boxplot above).

c.

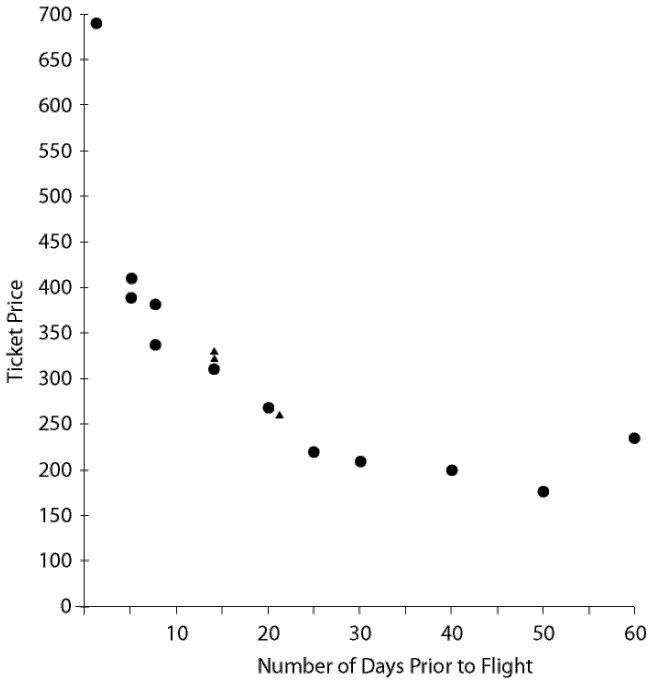
- The overall pattern visible on the line plot is that there are 3 clusters of data; the cheaper tickets between $180 and $240, the middle cluster at about $310, and the expensive tickets between $380 and $410.
- The overall pattern visible on the histogram shows a very large cluster between $200 and $250, and another large cluster between $300 and $450.
- The overall pattern on the boxplot is that typical ticket prices fall between $220 and $390.

Note: These overall pattern statements are slightly different, but not contradictory. The line plot gives more detail, but sometimes more detail obscures

	pattern. The boxplot gives least detail, but does show where most of the data values lie. d. All of the graphs show that $690 is unusually high. The boxplot gives us a way to measure how unusual (more than 1.5 IQR's above the 3rd quartile). e. Two ways to *measure* variability are "range" and "IQR." *Range* is a very rough measure of variability, easily influenced by outliers. *IQR* is a more reliable measure of variability since it measures how closely clustered the middle of the data distribution is. If students use "range" to measure variability they should also check for outliers, and they should use clusters to back up their statements about variability. • The line plot gives the Range, $690 - $180 = $510, with data clustered as noted in part c. • The histogram gives an approximate range, $700 - $150 = $550, with data clustered as noted in part c. • The boxplot gives the range as exactly $510; without the $690 as an outlier the range would be $410 - $180 = $230. AND it gives the IQR as $390 - $220 = $170. *This gives us the most information about variability*; we can say that almost all the ticket prices were between $410 and $180, with typical prices between $220 and $390. f. "Typical" price can be measured by using *mean, median, mode or clusters*. • The lineplot seems to balance (like a teeter-totter) at the $310 cluster, maybe higher because of the $690 outliers. The mean is *approximately* $310. • The histogram seems to balance at the $300 - $350 interval. The mean is in this interval. • The boxplot shows us that the median is exactly at $310. We might also say that the typical prices are from $220 to $390 (the "box"). Because of the outliers at $690 we would expect that the mean would be higher than the median. Note: we can make a statement about a measure of center, or a typical ticket price, from each type of graph, but only the boxplot gives us an exact measure of center at a glance and tells us where the middle half of the data lie.

6. Make a scatter plot showing how the ticket price relates to when it was purchased. Is there an overall pattern?

Price ($)	Number of passengers	# days prior to flight for date of purchase
180	4	50
200	3	40
210	5	30
220	2	25
240	8	60
260	1	21
270	2	20
310	7	14
320	1	14
330	1	14
340	2	7
380	4	7
390	6	5
410	6	5
690	2	1



	Note: multiple values are shown here as large dots, individual values as smaller triangles. In general we can see that as the time increases the price of the ticket decreases, though there are exceptions. The group of people who bought their ticket 60 days before the flight seem to have missed out on all opportunities for reductions in price.
Samples are subsets of the entire data set or population, selected in some way. **Sampling Methods** • **Convenience**: members of the sample are chosen in a way that is convenient to the person collecting the data. • **Voluntary**: members of the sample are self-selected. • **Systematic**: members of the sample are selected according to some formula (every hundredth person in the phone book, for example) • **Random**: members of the sample are chosen by some random device that removes both the members of the sample and the person doing the sampling from the process. (By removing both members of the sample and person doing the sample from the process any systematic bias, intended or not, is eliminated.) **Why Random?** A Random sample MAY resemble the population in important ways, such as measures of center or	*7. What type of **sampling method** was used for each of the following samples?* *To find out what proportion of passenger cars on the road are SUV's or larger* *a. Note all the cars that you see from the school bus window on the way to school and find the proportion which are SUV's or larger.* *b. Note the type for every 10th car at a busy intersection, and find the proportion that are SUV's.* *c. Get a list of vehicle registration numbers from the Department of Motor Vehicles and use a random number generator to select numbers at random.* a. This is a **convenience sample**. You may have a very biased sample because of the time of day and the route that the school bus takes. Perhaps the route is through a rural area where many pickup trucks are used. b. This is a **systematic sampling** method. Bias, in the form of input from the person doing the sampling, has been removed. But there is still potential bias in the choice of intersection. And this kind of sampling method makes it impossible to have a sample of consecutive cars. That is, not EVERY sample is possible. c. This is a **random sample.** Every registration number has an equal chance of being chosen, and every sample of a particular size has the same chance of being chosen. That is, it is *possible* that consecutive numbers will be chosen, *possible* that every 10th number will be chosen, *possible* that every car chosen is an SUV, and *possible* that every car chosen is NOT and SUV. Over the long haul, unusual (that is not representative) samples will be chosen less frequently than typical (that is, representative) samples.

variability. But a random sample MAY be very unusual and very unlike the population from which it was drawn. The former will give us good predictions about the population, and the latter will be misleading. However, over the long haul, large random samples will give us good predictions about a population, and we can even say how confident we are that our predictions (of a population mean for example) are quite accurate. This is because, while a single random sample is unpredictable, if we took many random samples of the same size and calculated the mean of each sample, we would find that 95% of all random sample means cluster measurably close to the population mean. So, for any ONE random sample we can say that we are 95% confident in the prediction it gives us about the population mean. (Predictions are usually given as intervals.) 5% of our random samples will have means that will fall far enough from the population mean to be poor predictors. The problem is that we never know whether the one sample we have collected is like the 95% "usual" samples, or like the 5% "unusual" samples. Statisticians have developed the language of confidence intervals to state their predictions from random samples. If we worked with non-random samples (however carefully representative they are) then all the probabilities about how sample means will cluster round a population mean will not be valid.

*8. Describe how to choose a **random sample**.*

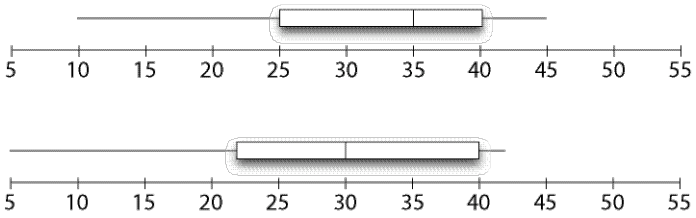
There are various ways to choose a random sample. All of them are variants of: number the "population" in some way, and then select numbers at random using some kind of device that is unbiased. For example:

- Suppose you wish to find out what the student body of a middle school thinks about a new lunch plan, then you might use student numbers to number the population, say 1 - 1000, and select numbers at random by throwing 3 ten sided dice (or one die 3 times). You will have to code the dice outcomes: say 10 = 0, 1 = 1, 2 =2 etc, so that 3-10-2 means that student 302 has been selected.
- Suppose you want to try out a new medication for high blood pressure, then you might ask for a list of all patients being treated at a hospital for high blood pressure, and number them, say 1 – 92. Now use a random number generator to choose a sample to receive the new medication. (In this kind of application the researcher usually chooses another random sample to receive a placebo, and maybe a third random sample to receive the standard treatment. When the results are reported we might compare the mean blood pressures for the three samples. If they are very different then we might attribute the difference to the treatment. We can then say that we are confident that a particular treatment lowers blood pressure, at least for the population of patients at that hospital.)
- Suppose you want to check the quality of an automotive part made at a factory, then you might find out how many are produced in a day, say 4000. Then use a random number generator to select 30 numbers between 1 and 4000 and remove these particular parts for testing as they are made. We might then calculate the mean strength (perhaps measured as a breaking weight in pounds per square inch) of the sample and say that this represents the mean strength of *all* the parts made that day. Of course, the sample MIGHT be very unusual quite by chance, and our prediction will be very wrong. But over the long haul we can use large samples chosen at random to make confident predictions about the population from which they were chosen.

Sample Size impact
Small samples of data will generally have larger variability than large samples, and will be less predictive about the population than large samples. Students found that random samples of size 30 had measures of center and variability, and shapes of distribution, that were USUALLY very like the population.

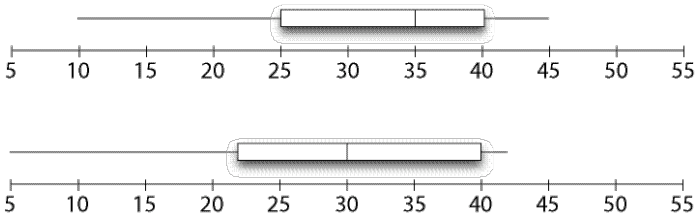
Sampling Distribution
This is different from the "distribution of a sample." The distribution of a sample implies that we have collected a set of data from a population (the parent set of data), and have made a graph or calculated some measures to see how the data values **in the sample** are spread out or clustered together, and to determine what is typical for that sample. However, a "sampling distribution" implies that we have collected **many samples of the same size**, and summarized each sample with a mean, and then we have used these **means of samples as a new data set**. A longer and more informative name for a "sampling distribution" would be a "distribution of the means of samples." Each sample is reduced to one value typical of the sample. Some of these sample means will be far from the mean of the population from which the samples were drawn. But most of the sample means will be clustered closely around the population mean. Students found that if all the samples had 30 members then the means of the samples clustered very closely round

9. Here are two samples.



Do you think it is likely that they were drawn from the same population?
The 5 number summaries for the 2 boxplots are {10, 25, 35, 40, 45} and {5, 22, 30, 40, 42}. The ranges are similar (35 and 37), the maximum and minimum values are about the same, the medians are 35 and 30, and the IQR's are 15 and 18. The graphs have similar shapes, skewed to the left. It is possible, though not certain that each of these samples came from the same population, since they seem to have similar measures of center and variability.

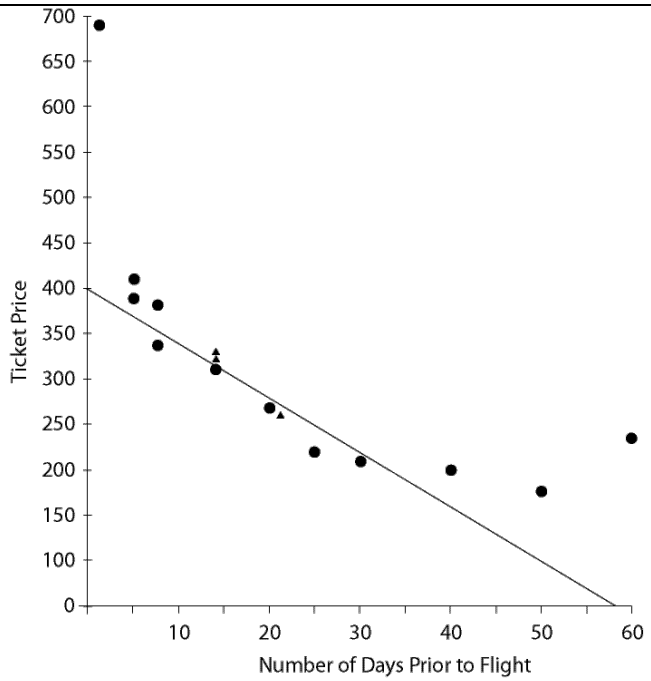
10. *Here are two samples drawn from the same populations.*



What conclusions can you draw about the population based on the samples?

Assuming that we have not been unlucky enough to have two unusual samples, then we can conclude that the population will resemble these samples: the median will be approximately 32, and at least 50% of the population will cluster between approximately 25 and 40.

the population mean, while the means of small samples were more widely spread. Since most samples size 30 (or greater) give good predictions about the population we can have confidence in predictions made on the basis of a single sample. Since means of small samples do not cluster closely around the mean of the population we have less confidence in predictions based on the basis of a single small sample. Note: we can also make sampling distributions with proportions deduced from samples**.**	
Covariation, Association, Linear model When we wish to see if there is a relationship between 2 variables we make a scatterplot. If there is a pattern or trend then we say there is a **relationship, or an association** between the variables. If the pattern is linear, increasing or decreasing, then we may try to fit a line to the pattern. This line is called a **linear mode**l. The purpose of the linear model is to assist in making predictions about pairs of data that are not in the data set.	*11. A **scatter plot** showing price of airplane ticket versus days prior to flight shows that there is a pattern.* *a. Fit **a linear model** to this pattern. What is the equation of your linear model?* *b. Use your linear model to predict the price of a ticket if you buy 45 days and 55 days before the flight date.* *c. Comment on how well your linear model fits the pattern.*



a. Students will have to use a straight edge (a piece of spaghetti works well) to try different positions for the linear model. The goal is to get as close as possible to as many pieces of data as possible. Obviously this requires compromising and perhaps not hitting any one piece of data exactly. A line that fits reasonably well is shown on the graph.
b. Reading from the graph, when the number of days is 45 the price should be $130; when the number of days is 55 the price should be $70.
c. A linear model makes no sense if you continue for larger and larger numbers of days; eventually the price would be a nonsensical negative number. In fact there seems to be a curve to this pattern, so perhaps a linear model was not the best predictive tool for this data set.