YALE NATIONAL INITIATIVE
to strengthen teaching in public schools®

Curriculum Units by Fellows of the National Initiative
2005 Volume IV: Astronomy and Space Sciences

The Use of Astronomy to Teach Statistical Concepts

Curriculum Unit 05.04.09, published September 2005
by Michael Vasileff

Introduction

To me, the intent of the Yale Initiative is to produce modules that are not only integral to a curriculum unit, but add an example from a related, but different discipline. In this way, the subject matter comes alive with a topic of interest to the student. Mathematic textbooks are notorious for their "drill and kill" mentality where a new mathematical algorithm is introduced followed by pages of simple problems of rote solutions and then a collection of unrelated story problems. Few students have the perseverance to work through the tedium of problems with no grounding in actual experience, much less even more complex story problems on a variety of problems ranging from business to engineering. Our topic, Astronomy, is one I believe captures everyone's imagination, especially students in Florida. Whether it is a starry night, a space launch, or just letting our imagination run wild with the Universe, man has documented his fascination in all art forms for thousands of years.

I teach statistics to juniors and seniors in an average high school. Approximately 15% of our graduates continue their education at a four-year college, 45% attend a junior college and the remainder choose other options such as military service, trade schools or join the workforce. Consequently, I am trying to provide a glimpse into the science of astronomy (at a very basic level) for any post primary grade as well as add enrichment for those students taking statistics by using a few basic facts of astronomy as we study regression analysis.

Teachers are assigned a text and related materials. Each chapter, at least in my experience, covers some topic followed by specific vocabulary, discussion questions, problems and enrichment exercises. What I find lacking is a thread, tying all the chapters together, solving one intriguing problem that encompasses all the chapters in the text. For example, a statistic book usually begins with descriptive statistics and goes through describing data as nominal, ordinal, interval or ratio. Then it takes a look at types of samples such as *random* or *convenience* and then moves on to measures of location *(mean, mode, median*), *standard deviation* and *Z scores*; finally, ending the first half of the course with a little probability. The second semester is devoted to inferential statistics in which decision solutions are made and are expressed with some or no degree of confidence in whatever the problem issues are; i.e. does this medicine work? Each chapter has all sorts of interesting problems on medicine, crime, population, etc. I think this is good to an extent, but if we had some sort of parallel, non-graded set of exercises outside the text as an empowering topic, I think the course would

be much more enjoyable to the students and spark some self-motivation for further study.

Rational for this module

Every day, each of us make hundreds of decisions, some small, some very large. Students, as part of their academic and social growth, need to be taught structures and methodologies that will enable them to improve their decision-making capabilities.

To make good decisions, we need to have the necessary data/facts and analyze them for the best choice. This is what good science does. First, we observe phenomenon and then utilize hypothesis testing on what we perceive is reality. Thus, we have the null (H_0) hypothesis of no difference or equality and the alternate hypothesis (H_1) our view of reality, i.e. H_0 : : $_1$ = : $_2$ and $H_{1:}$: : $_1$ ≠ : $_2$ or you can use , or > . (To make this a bit more understandable, you might want to use an example in words. For instance, : $_1$ = average salaries of males, and : $_2$ = the average salaries of females. Our null hypothesis states that the average salaries of males equal the average salary of females. Our alternative hypothesis states that the average sales of males and females are not equal.) An appropriate set of statistics is utilized so we may determine whether the null hypothesis can be rejected. We do not "accept" the alternate hypothesis because we never know all the variables and facts. If, however, we can *reject* the null hypotheses, we, by default, accept that the alternate hypothesis is true.

While **regression analysis does not necessarily prove causality**, it is a pivotal statistical methodology for looking at two or more independent variables simultaneously and showing their relationship (strength, direction and interplay) to our phenomena of interest, the dependent variable. Regression Analysis is more logic than math. It is about understanding relationships and not just a rote set of algorithms with a magical answer that tells you what to do at the end of the arithmetic More often, the first level of analysis requires more layers before a truly insightful analysis can be ascertained.

Goals of this module

The goals of the module are twofold:

- 1. Make Statistics more enjoyable to the students by giving them a little "breathing room" from the text while teaching statistics.
- 2. Teach a little Astronomy and show students that all sciences can be interrelated and with a little imagination, the impossible is solvable.

Synopsis of topics to be covered

In this unit, we will cover two items of astronomy and how to use statistics to verify that these are facts and not mere opinions. The first topic is centered around the Titus-Bode Calculation that states that the planets in our solar system follow a simple relationship between their order and distance from the sun.

The second topic discusses the "Big Bang" theory, which states that the Universe expands according to a linear equation with the slope equal to the Hubble Law. Since this topic covers *time*, *distance* and *velocity*, with a little further analysis, we can calculate the age of the Universe.

Topic 1 – Our Solar System's Planets Distances from the Sun or the Titus-Bode Calculation (to be used during the study of inferential statistics; typically, second half of a statistics course).

We have all seen sequences in logic puzzles and exams. Some of these sequences help us determine the number of days in a month. For example, you can "play" a simple counting game to determine whether a month has 31 or 30 days by utilizing the knuckles and spaces between the knuckles of your hand. Start by labeling the first knuckle as January, February as the space between your first and second knuckle, March as the second knuckle and so. Since we know that February is the shortest month, and it is a space between the knuckles, it follows that all of the months that land on knuckles have 31 days and the months in between the knuckles have 30, except for February, which has 28 or 29 days. (Augustus Caesar insisted on 31 days in his month.)

In 1766, a German physicist, Johann Titus, predicted the mean (average) distance of each planet from the sun by using a relatively simple mathematical progression of numbers. The number progression is as follows: 0 (Mercury), 3 (Venus), 6 (Earth), 12 (Mars), 24 (Asteroid Belt), 48 (Jupiter), 96 (Saturn), 192 (Uranus), 384 (Neptune) and 768 (Pluto). Now add 4 to each number in the progression and then divide the result by 10. The third number in the progression (Earth) has a value of 1 and all other numbers (planets) are ratios of that 1, thus determining their mean distance from the sun. His discovery didn't garner much attention until Johann Bode, another German astronomer, published the calculation in the late 1770's. This set of ratios became known as the "Titus-Bode Law." Utilizing this expression, Bode predicted the existence of another planet between Mars and Jupiter. At the time of the development of the Titus-Bode Law, only six planets were known. However, this prediction was later borne out by the discovery of the asteroid belt in 1801. It took more than 100 years after Bode to discover the remaining planets: Uranus (1871), Neptune (1846) and Pluto (1930).

The question we are looking to answer is *how accurate is the Titus-Bode Law*? The work sheet and chart below offer a tabular and graphic representation of this law as well as one measure of its accuracy. To create this table, the planet names are in column 1, with Titus's numbers of 0,3,6 etc. in the second column. The third column is the constant 4 to be added to each of Titus's numbers; we add the two columns in column four. We divide by 10 and we get column five. Column five is then copied to column six as we wish to compare Titus's number to our current knowledge, which is in column seven. Just as we use one foot as a unit of length, astronomy has a standard unit of measure, the astronomical unit or AU. An AU is the average distance

between the Earth and the Sun, which is approximately 1.4960 x 10^8 km. To convert this distance to our American unit of miles, multiply the 1.4960 x 10^8 by .621 (1 km = .621 miles) for 92,901,600 miles.

At this point, you might want to give the students a little refresher in metric to English conversion. We start with 1.4960 x 10^8 km (kilometers or thousand meters) which translates into 149,600,000 km or 149,600,000,000 meters. If you have a meter stick available, have the students note that there are about 39.4 inches per meter (or have them look it up in the dictionary). We then multiply the 149,600,000,000 times 39.4 for a product of 5,894,240,000,000 inches. Since there are 12 inches in a foot we divide by 12 to get 491,186,667,000 feet, and since there are 5,280 feet in a mile we divide again and get 93,027,777 miles. Thus, one AU is approximately 93 million miles –; the distance from the Sun to the Earth.

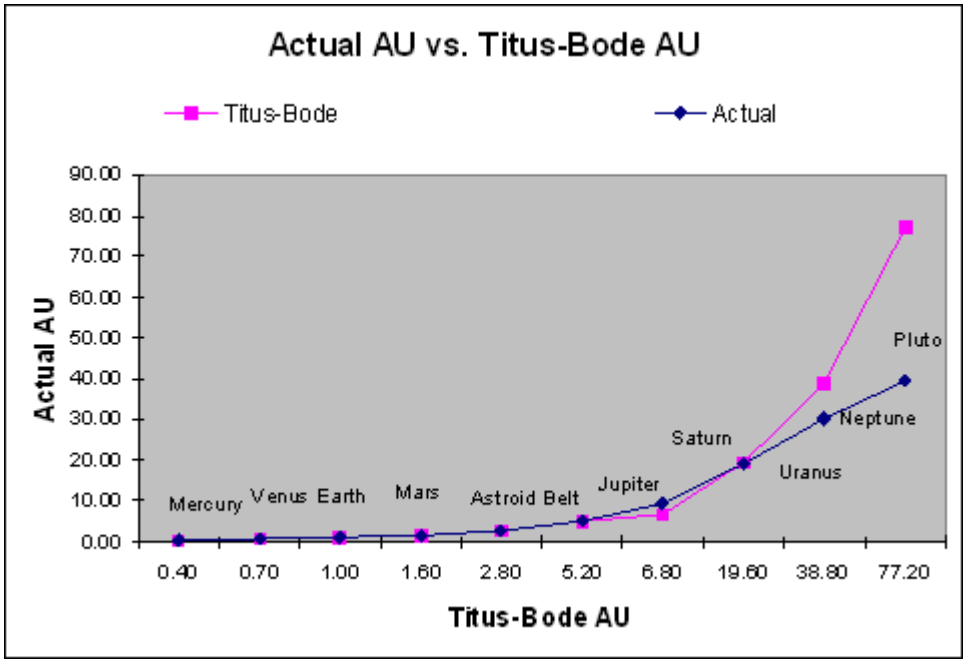
Titus-Bode Law - Estimations of the mean distance from the sun using Earth as a unit of 1

					Distance to Sun in Astronomical Units		
Planets		**+4**	**Total**	**Div by 10**	**Titus-Bode**	**Actual**	**Difference**
Mercury	0	4	4	0.40	0.40	0.39	-0.01
Venus	3	4	7	0.70	0.70	0.72	0.02
Earth	6	4	10	1.00	1.00	1.00	0.00
Mars	12	4	16	1.60	1.60	1.52	-0.08
Asteroid Belt	24	4	28	2.80	2.80	2.80	0.00
Jupiter	48	4	52	5.20	5.20	5.20	0.00
Saturn	96	4	100	10.00	10.00	9.54	-0.46
Uranus	192	4	196	19.60	19.60	19.18	-0.42
Neptune	384	4	388	38.80	38.80	30.06	-8.74
Pluto	768	4	772	77.20	77.20	39.44	-37.76

Taking the difference between the predictions of Titus-Bode and what is know today, we see that overall, the Titus-Bode Law does a fairly good job of predicting a planet's mean distance from the sun.

We can also use another statistical concept to illustrate the accuracy of the Titus-Bode Law predictions –; the scatter or XY diagram. A scatter diagram graphically illustrates the relationship between two variables or items of interests. In this case, that would be Titus-Bode calculation of mean distance and the accepted calculation of mean distance.

The chart below is a XY diagram of our raw data –; Astronomical Units calculated using Titus-Bode and current accepted AU measurements, with a line drawn to connect the points or observations.



As we look at this chart, we can see that there are two somewhat bent, but almost identical lines. How strong is the relationship between these measurement systems? A statistician by the name of Carl Pearson studied this type of phenomenon and he figured out a way to calculate both the strength and direction of a linear relationship. Since we have a fairly linear relationship, we can use Pearson's product moment correlation coefficient, or *r*. To do this, he developed the following very famous formula.

$$r = \frac{\sum\left(x-\bar{x}\right)\left(y-\bar{y}\right)}{(n-1)SxSy}$$

Where

- r = correlation coefficient
- x = individual observations of a variable (Titus-Bode)
- $\bar{x}$ = the average of the X variables
- y = individual observation of another variable (current measurements)
- $\bar{y}$ = the average of the Y variables
- Sx = the standard deviation of the X variables
- Sy = the standard deviation of the Y variables

Before we attack this foreboding looking equation, it might be a good idea to review the elements that make up the equation. To do this, we need to review some of the basics of descriptive statistics. Statistics is divided into two main branches, descriptive and inferential. In descriptive, we merely describe and plot the data with no value judgments; the data is used to describe a phenomena of interest. In inferential statistics, value judgments are made. In this example, Titus-Bode vs. Actual, we will hypothesize, or make an educated guess, that there is little or no real (*significant*) difference between the two groups of data.

The first item that we will look at in this equation is the *average or mean* which is denoted by a bar over the variable such as $\bar{x}$ (verbalized as "xbar"). The mean is the sum of all the variables divided by the total number of variables. The standard statistical equation for this is Σ x/n. The Greek letter sigma (Σ) in the numerator translates to "sum all"; the denominator tells us to divide by the number of observations (n). For

example, if we took the height in inches of three students in a class of 25 and got 60, 65 and 70, we would add the heights together and get 195 and divide by 3 to get an average height of 65 inches.

The mean or average probably is the widely used of the statistical procedure; and it is also the most abused. Data is greatly influenced by a few or even a single much larger or smaller observation, with the result not accurately describing our collection of observations. Instead, we may want to use the *median* to describe the data. The *median* is the middle value of an ordered set of observations. In our case, the median would be 65. This is just like the median in a highway –; the middle between the two lanes going in opposite directions. If there is an even number of observations (say 6, 10, 14, 100), the middle two numbers are averaged; if the number of observations is odd, then the middle value is used. There is a final descriptive statistic called the *mode*, which refers to the most often, or most common, occurring value. In our example there is no mode. These three measures (*mean, mode, and median*) are often referred to as *measures of central tendency* as they describe the center of the data (or where the data tends to cluster) in a given data set.

An *observation* or data point is part of a set or group of data. For an observation to be meaningful, it needs a reference point and a way of measuring the relationship of that observation to that reference point. We call this reference point is the *mean* and our measurement is the *Standard Deviation*. Standard refers to the norm or what is expected (in statistics, this is typically the mean) and deviation is the difference or numeric distance that an individual data point is from the standard. To do this, we need to find the standard or mean, $\bar{x}$ which we already have calculated as 65. By subtracting 65 from 60, 65 and 70 we get differences of -5,0 and +5. If we add these differences together we get 0. This doesn't do us much good. This zero problem exists anytime we get a group of data, find the mean and then subtract the mean from each item and add the differences. (Your might want to ask student to verify or "prove" this to themselves with a very simple data set.) To avoid this problem and maintain the differences, we square the differences. This is called the *variance*. The -5 squared is 25, 0 squared is still 0 and 5 squared is 25. We can now add theses numbers and get 50. But this 50 doesn't tells us very much so now we need to find the average difference. To do this, we divide by the number of variables. In our case n = 3 so that 50/3 approximately equals 16.7. But what does this 16.7 *square inches* mean? We now need to turn this number back into something that is meaningful (the original unit of measurement, or inches), so we take the square root of our 16.7 square inches, which is approximately 4.1 inches –; this is the average difference of each variable from the mean.

The statistical formula for what we have just done is:

$$S=\sqrt{\frac{\sum\left(x-\bar{x}\right)^2}{(n-1)}}$$

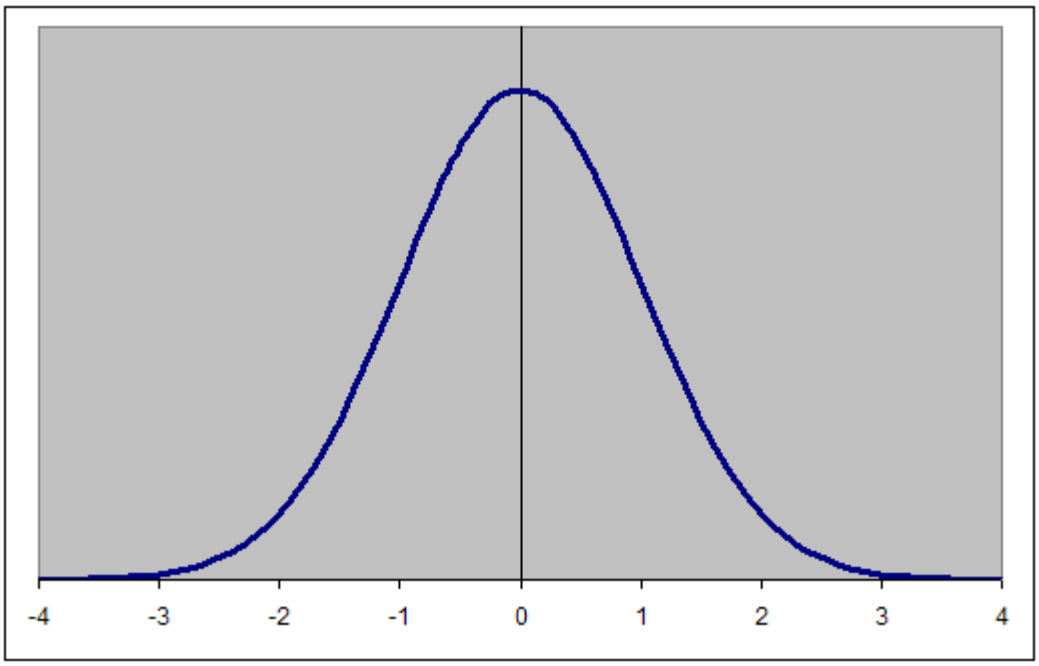
where s = the standard deviation and

$$\frac{\sum\left(x-\bar{x}\right)^2}{(n-1)}$$ = the variance

For purposes of illustration we use 3 for the denominator. In actual practice if we take a sample of data we must subtract one from the number of observations or n-1 as the denominator. Statisticians do this because of what is *called the error of the mean*. That is, if we take many samples of data from the whole population, each mean will be slightly different assuming we take an unbiased sample based on some random selection. Also note that the English "S". This implies a sample or a part of the population and not a census or including all

members of the population. Since standard deviation is a measure of how close data is together and there is always an element of error in using samples the denominator is decreased by one. The purpose of this is to be conservative in claiming how close the data is. The smaller the denominator, the larger the quotient. Statistics always wants to error on the conservative side of analysis. A larger standard deviation will have a flatter curve while a smaller standard deviation will have a steeper curve indicating that the data is grouped much closer. If every element of a population is included in our data we use the lower case Greek sigma, F in place of s and we use N instead of n-1. In this case, we have considered only a small sample of every element, so there is an error of the mean but to simplify calculations, we used the value of N or 3.

In nature most data, when graphed, follows what the statisticians call a normal distribution or bell shaped curve. The bell shaped curve indicates that the majority of the data fall under the center of curve. The chart on the next page illustrates this normal distribution but with a slight "twist" –; it has a mean (center) of 0. The numbers along the base of the curve, ranging from -4 to + 4 represent *standard deviations.* Now we have combined our reference point (the mean) and our measurement of the distance from an observation to the reference point (the standard deviation).



This is a *standard normal distribution*. The centerline at 0 represents the mean, mode and median which in normal distribution are the same (in addition, it divides the data in two, with 50% on one side of the line and 50% on the other). Data having a wider range would produce a flatter curve while data with a more narrow range would produce a more steep slope. In our example, the zero represents 65, the + 1 would represent 65 + 4 (standard deviation) or 69, and the -1 would represent 65-4 or 61. The higher and lower the standard deviation marks go the fewer the observations having that value; that is, at a -3, there would be very short people and at + 3, the people would be very tall.

The French mathematician Abraham de Moivre discovered the formula for this curve using calculus. He determined that if you add or subtract one standard deviation to the mean you capture about 34% of the area under the curve for a total of 68% of the data. A repeat of this addition and subtraction, gives two standard deviations yielding about 95 percent of the area under the curve and three standard deviations is about 99 percent of the area. Since the ends of the graph, called the *tails,* go on forever (infinity), there is no limit to

the number of standard deviations one can calculate; therefore, 100% of the observations are under the curve. Since the exact calculation of the percentages are very complicated, we use a z or t score table to calculate the location (in percentage) of an individual score expressed in standard deviations. The calculation of a z-score is as follows:

$$z = \frac{\left(x - \bar{x}\right)}{S}$$

Where

- x = the observation of interest (a specific point)
- $\bar{x}$ = the average of the observations (*mean*); and,
- S = the standard deviation of the variable *x*

Most every statistics text has a table for Z. By looking up the value in the columns of the table, the body of the table gives you the percentile of a given *x*.

So far, this discussion has been about one variable. When the discussion turns to two variables, an x and a y, there is *covariance* (the extent to which two variables vary from their means). This covariance becomes important when we want to calculate the strength and direction (correlation) between two variables or measurement systems. Covariance is mathematically expressed as:

$\text{Cov}(x,y) = \Sigma (x - \bar{x}) (y - \bar{y}) / N-1$

Using our Titus-Bode and Actual AU data, we illustrate this principle.

	Mercury	Venus	Earth	Mars	Astroid	Jupiter	Saturn	Uranius	Neptune	Pluto	Average
Titus-Bode	0.40	0.70	1.00	1.60	2.80	5.20	6.80	19.60	38.80	77.20	15.41
Actual	0.39	0.72	1.00	1.52	2.80	5.20	9.54	19.18	30.06	39.44	10.99
$(x - \bar{x})$											Totals
Tituis-Bode	-10.59	-10.29	-9.99	-9.39	-8.19	-5.79	-4.19	8.62	27.82	66.22	44.25
$(y - \bar{y})$											
Actual	-10.60	-10.27	-9.99	-9.47	-8.19	-5.79	-1.45	8.20	19.08	28.46	0.00
Product	112.15	105.58	99.70	88.83	66.99	33.47	6.05	70.60	530.57	1884.15	2998.08
N-1 or 10 Planet Positions -1											9.00
Covariance of (x,y)											333.12

The products of the differences are multiplied and the sum of the products is then divided by (N-1) since we are using a sample and not the population and want the average. However, this doesn't really mean much or tell us a lot because we have not standardized it. To do this, we must consider the standard deviations of the two variables (and brings us back to our original express of the *correlation coefficient*, or *r*).

The table below is an insert from an Excel worksheet that manually calculates the standard deviations of our two variables (Titus-Bode or x, and Actual or y). I often have student do a similar worksheet manually to get the basic idea of how the calculations are done manually before having them work calculators and spreadsheets.

Titus-Bode	x	$\bar{x}$	$(x-\bar{x})$	$(x-\bar{x})^2$	Actual y	$\bar{y}$	$(y-\bar{y})$	$(y-\bar{y})^2$
Mercury	0.40	15.41	-15.01	225.30	0.39	11.02	-10.63	112.89
Venus	0.70	15.41	-14.71	216.38	0.72	11.02	-10.30	105.99
Earth	1.00	15.41	-14.41	207.65	1.00	11.02	-10.02	100.30
Mars	1.60	15.41	-13.81	190.72	1.52	11.02	-9.50	90.16
Astroid	2.80	15.41	-12.61	159.01	2.80	11.02	-8.22	67.49
Jupiter	5.20	15.41	-10.21	104.24	5.20	11.02	-5.82	33.81
Saturn	6.80	15.41	-8.61	74.13	9.54	11.02	-1.48	2.18
Uranius	19.60	15.41	4.19	17.56	19.48	11.02	8.47	71.66
Neptune	38.80	15.41	23.39	547.09	30.06	11.02	19.05	362.71
Pluto	77.20	15.41	61.79	3818.00	39.44	11.02	28.43	807.98
average	15.41				11.02			
total				5560.09				1755.16
/n-1 or 9				617.79				195.02
sq root and standard deviation				24.86				13.96

The product of the two standard deviations is (24.86)(13.96) = 347. Dividing the *covariance* by the product of the standard deviations or, 333.12/347 yields and **r** = + .96

A correlation coefficient can only take on values between -1 to +1. This value gives us the strength (the number: the higher the number, the stronger the relationship) and direction (the + or -) of the relationship between our two variables or measurement systems. In our example, r = +.96 , a strong and positive relationship. Had our ratio turned out to be a + 1, there would be perfect correlation, meaning that the measurement systems were identical, and the a graph would show the two lines atop each other. As the ratio goes from + 1 to 0 there is less strength between the two until at 0 there is no relationship. Below 0, the strength of the relationship is reversed. In such a case, one variable increase as the other decreases. For example, from birth to about age 20, we get taller; this is a positive relationship. If the coefficient is zero, then there is no correlation; the scatter diagram looks like a circle of dots (an example might be finding the relationship between brown eyed people and I.Q). A negative correlation is illustrated by the fact that after the age of 40 we tend to shrink or as we age passed 20, our height decreases.

Calculating *r* can be done many ways. Using a TI83 hand calculator, press Clear and then clear memory by pressing the "gold" key and then the "mem" key located at the + sign, press 7 (reset), press 1 (all ram),press 2 and then press the enter key twice. Press the "stat" key and then the "edit" key. This brings up a spreadsheet. Key the x variables in column 1 and the y variables in column 2. Variables may be entered in other columns, but the calculator uses these two columns as the default. Press "gold" key and then the "quit" key to get a blank screen. You must now activate "diagnosticOn" . To do this, press the "gold" key and then the "catalog" or 0 key. Either scroll down to "diagnosticOn" or, since you are in the alpha mode, press the "d" or "x^{-1} " key to get to the d's and then scroll down to the "diagnosticOn" key and press enter. Press the "stat" key, use the right arrow to go to "calc" at the top of the screen and enter 4 or "LinearReg (ax+b). Voila, you should see y = ax + b, a = .539, b = 2.67, r^2 = 92 and r = .96. So if you wish to graph the best-fit line to our data, you have y = .539x + 2.67. Our "r" is the same as we manually calculated above. But we now have r^2 which should be R^2 . This is the square of little r or .96*.96 or .92. R^2 , though not needed here, is the measure of the amount of variability in one variable explained by another variable. In English this means that little r tells us that there is a very strong relationship between the two variables 96% and R^2 tells us 92% of Titus-Bode explains Actual astronomical units.

In Excel, the calculation is much easier. Go to "Tools" on the toolbar and open up "Add-ins" to see if "Analysis

Toolpack" is installed. If not, check the box and it will self install on most current releases of EXCEL. Enter the two columns of variables on a worksheet and grab them in a rectangle with the mouse. Go to "Tools" and press "Data Analsis", highlight "Correlation Coefficient" and answer the questions. You might also use the "Descriptive Statistics" option. Set up and the first time use can be frustrating, but once used, subsequent uses are quick and easy. On all these automated systems, patience is rewarded.

Topic 2 – Edwin Hubble and the expanding Universe (to be used after the principles of multiple regression have been studied).

Most of us associate Hubble with the telescope, which has given us many marvelous discoveries and pictures of the Universe. The telescope, launched in 1990, is 600 kilometers above the earth's surface. It is a fitting tribute to a great scientist, Edward Hubble, who spent his live studying distant galaxies. One of his major observations was that galaxies are in three basic formations. *Spiral Galaxies* are a flat disk with spiral arms; *Elliptical Galaxies* are egg shaped; and the final grouping, *Irregular Galaxies* which are linear or triangular. A collection of his pictures showing these groupings was published posthumously by one of his associates. His more famous discovery is that galaxies are moving away from each other. This measurement of the movement of galaxies away from each other, known as the Hubble Law, will be the second and final topic of this module.

The planet Earth is in a solar system that is part of the Milky Way galaxy. There are other planets, moons, comets, asteroids, etc. in other solar systems, which are in other galaxies. And, to paraphrase Carl Sagan, there are billions and billions of them. By using techniques such as parallax, light, and ray analysis, Hubble was able to determine *distances* and *velocities* of various galaxies.

Hubble found that there is a direct relationship between the *residual velocity* (the speed at which galaxies are moving away from each other) and the *distance* from each other. This can be plotted on a XY diagram, with *Y= velocity*. Based on his calculations, the Universe is expanding. This expansion is often compared to raisin bread rising. As the bread rises, the raisins move further apart from each other. Galaxies (the raisins) are constantly moving away from each other (the rising bread). Another way to illustrate this phenomenon is to blow up a balloon a little bit and put some dots on in to form a square. Now blow up the balloon even further and this is an illustration of our expanding Universe.

On January 17, 1929, Hubble presented his paper, "A relation between distance and radial velocity among extra-galactic nebulae". A portion of his data is listed in the table below. In calculating *r*, the correlation between the variables of *distance and velocity*, we find a fairly robust measure of association, .79. We can also set this up as multiple regression problem since we know that this is a linear relationship. (The charts below show the results of such an analysis. The regression calculation results in the following value for a straight line y= 454x –; 40, where x = distance.)

Galaxy	Distance	Velocity
S. Mag	0.032	170
L. Mag	0.034	290
NGC 6822	0.214	-130
598	0.263	-70
221	0.275	-185
224	0.275	-220
5057	0.45	200
4736	0.5	290
5194	0.5	270
4449	0.63	200
4214	0.8	300
3031	0.9	-30
3627	0.9	650
4826	0.9	150
5236	0.9	500
1068	1	920
5055	1.1	450
7331	1.1	500
4258	1.4	500
4151	1.7	960
4382	2	500
4472	2	850
4486	2	800
4649	2	1090

Distance is in units of 106 parsec
Velocity is in km/sec

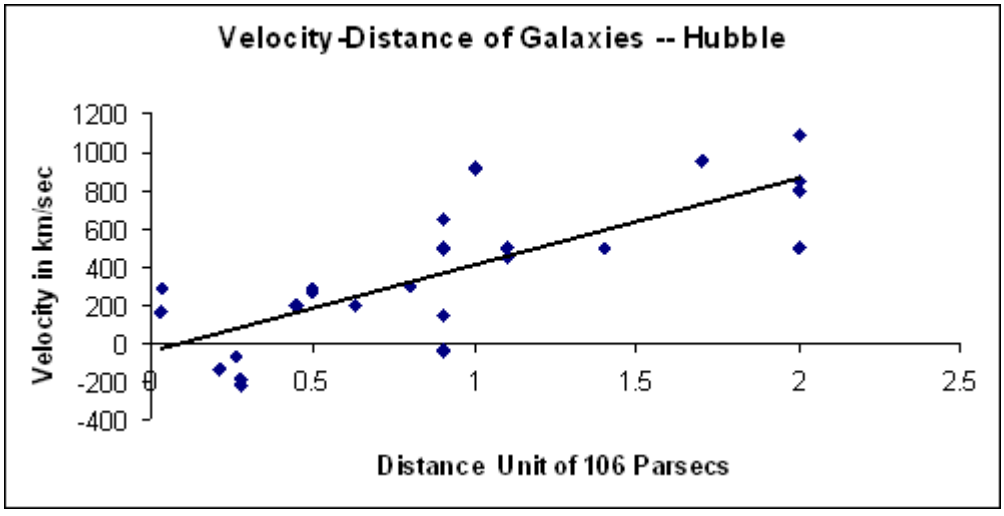
Excel Output

	Column 1	*Column 2*
Column 1	1	
Column 2	0.789639	1

The TI83 gives the following valu
r = .79
r^2 = .62
y = 454x -40

Calculating *r* is done in the same way that it was done for the Titus-Bode problem. If you use both the TI-83 and Excel, you see that the calculations are off a bit but this is due to the programming of algorithms; also, the hardware of each computer is always a little different; this difference is both machine and copyright induced.

Here we see that Hubble had a very good, but not excellent *r*. However, the *r* does show an irrefutable relationship. If we graph the x,y coordinates in Excel and draw a trend line (or regression line), the following chart results.



Again we see that there is a fair amount of variation of the actual observations versus the trend line, but there is definitely a linear pattern.

Hubble established the validity of his thoughts through the scientific method. As always happens, a host of other scientists retested his conclusions. However, there are skeptics, because proving a given theory on such a grand or galactic scale is always very difficult and unknown errors can creep in since we can never have perfect information and cannot conduct a controlled experiment, especially if it concerns the Universe. We have to work with what is out there with all its variables. In addition, other scientists may have a different opinion and wish to promulgate their theory over another's.

Another factor is improved science. By today's standards, the equipment available to Hubble in the 1920's is less accurate or sensitive. We now have coordinated telescopes around the world so that the combined observations result in a theoretical lens the diameter of the earth. There are all manner of radio telescopes. Since 1990, we have the Hubble telescope orbiting Earth. Obviously, then, our measurements since the 1920's are far superior.

Below is a second set of data (and the calculation of *r*) published by Kelsey and provided by an astronomer by the name of Abell, in the 1950's.

Approximate Distances & Velocities of Clusters of Galaxies

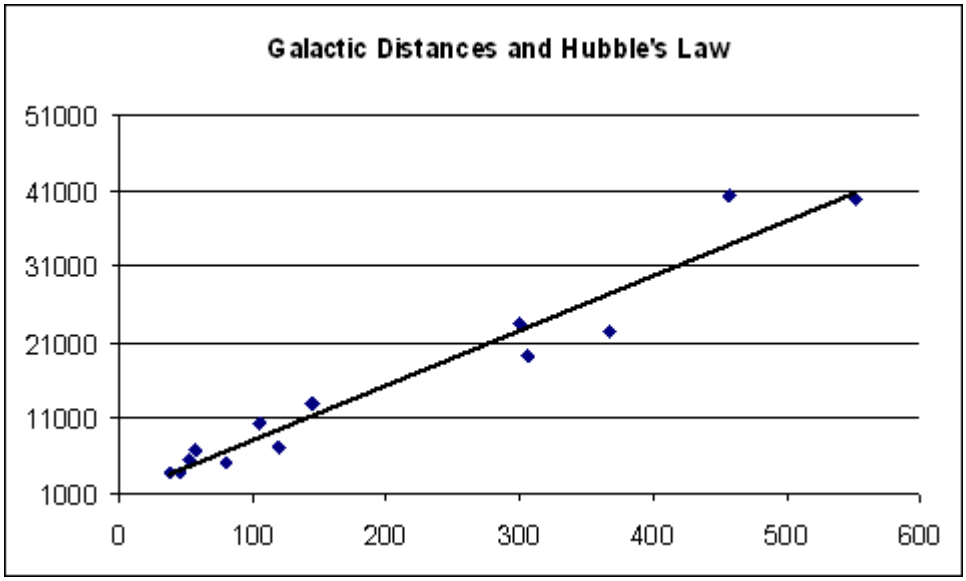
	Distance Mpc	Velosity KM/SEC
Pegasus I	38	3810
	46	3860
Perseus	53	5430
	80	4960
Coma	58	6657
	120	7200
Hercules	105	10400
Pegasus II	145	12800
Gemini	300	23400
	368	22400
Leo	307	19200
Urs Major II	457	40400
	552	40000

From Excel

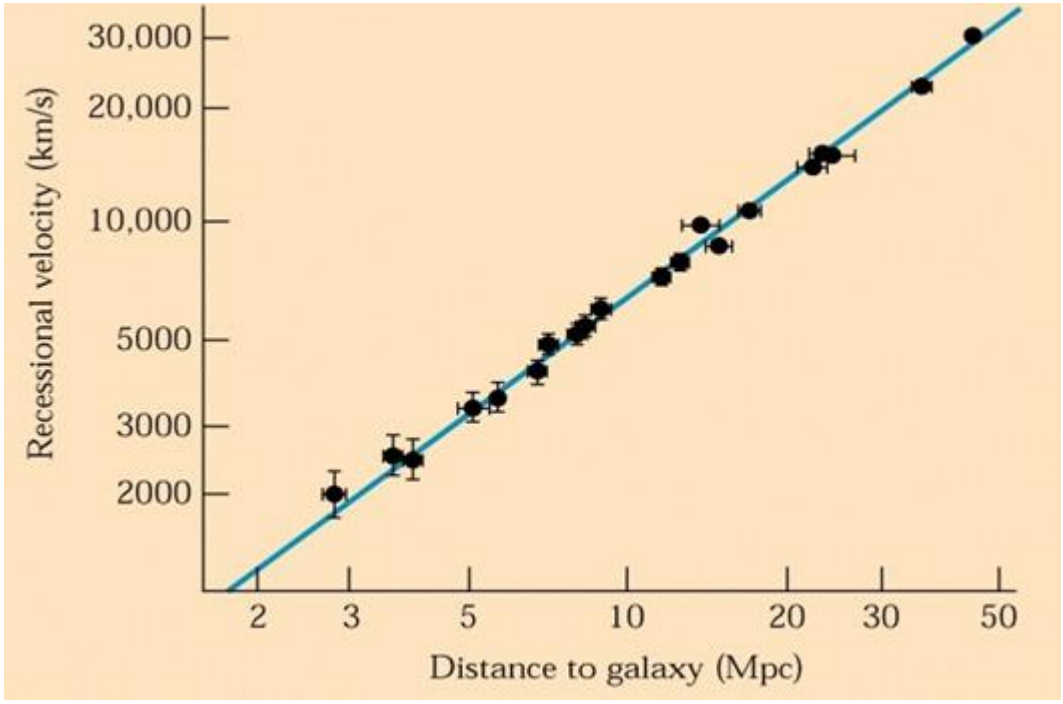
	Column 1	*Column 2*
Column 1	1	
Column 2	0.974406	1

The TI83 gives the following valu
r = .97
r^2 = .95
y = 72x + 788

Using the same relation as postulated by Hubble, but with more accurate data, he achieves an *r* of .97, a much stronger relationship than Hubble found. Also, when we examine the results of the regression analysis, we see that the *slop*e (the rise over the run) of the line is much less (72 compared with 454), resulting in a flatter line. Note that the data is much closer together, showing that the Hubble Law is re-verified. With the Hubble data, the slope of the line had a value of 454 (the value of Y when distance equals 0), with negative values. In the Abell data, all of our numbers are positive. This number, called the regression coefficient, gives us a much more realistic H_0 (72 rather than 454).



NASA continues these studies and its most current cart is reproduced below. (I was unable to find the input data for several reasons. Astronomical data are collected through very complicated measurements of light and rays. These are converted to data and graphed by the scientists and are typically not published in papers. Although we do not have the value of r, we can see that we have almost a perfect relationship as almost all of the data points are on the line.



For our final discussion, we can extrapolate another useful piece analysis from the Hubble Law. Not only is the Hubble Law a great piece of scientific discovery explaining our ever-expanding Universe, but also gives us data, which we can use to determine the age of the Universe.

The age of the Earth and other bodies in the Universe are determined in a variety of different ways, with the Hubble Law serving as a cross check. Using the "Big Bang" theory of the creation of the Universe, we assume that the "Big Bang" starts from a reference point equal to zero. Using Hubble's Law, we can calculate the time

of the "Big Bang" and thereby the age of the Universe.

In order to utilize Hubble's Law (or the value of H_o), we must start out with one of the most basic equations in Astronomy. This is the formula for Velocity:

$V = H_o d$

Where

- V = velocity of expansion
- H_o = Hubble's constant
- d = distance

In order to find the age of the Universe we must solve for T (time) which will require a little bit of algebra. If we divide both sides of the equation by H_o and we get $d = V/H_o$. The standard distance formula is d = TR or distance equals time (T) multiplied by rate (R) of speed. Velocity however is *directed* speed. Substituting Velocity (V) for Rate of Speed (R), we get d = TV. By substituting for times x velocity for distance, we get $TV = V/H_o$. Multiplying both sides of the equation by 1/V, we get the following:

$$\left(\frac{1}{V}\right)\left(\frac{TV}{1}\right) = \left(\frac{1}{V}\right)\left(\frac{V}{H_0}\right) \text{ the V's cancel and } T = \frac{1}{H_0}$$

Thus, time is the *reciprocal* of the Hubble Constant H_o. In order to calculate T, we have to remember that H_o is in Mega Parsecs or M/pc. A parsec is the distance light travels in 3.26 years in km or 3.09×10^{13} km. Converting kiloparsec to megaparsec yields 3.09×10^{19} megaparsecs. One calendar year equals 3.156×10^{7} seconds. The age of the Universe is calculated as follows:

$$T = \frac{1}{H_o} x \frac{Mpc/s}{km} x \frac{3.09x10^{19}\,km}{1Mps} x \frac{1year}{3.156x10^{7}\,s}$$

using the original H_o of 454 we get T = 2,156,580,292; using the current H_o of 72, we get T = 13,591,436,840 or approximately 14 billion years.

Try recalculating the above values using a calculator having parentheses and then use other values of H_o, which are available on a variety of internet sites. The values I've seen are all in the 70's and the goal is to get within 10% accuracy (but that's a topic for another session).

Have fun!!!

Bibliography

Freeman, Roger A. and Kaufmann, William J. III. *Universe* (New York , NY.: W. H. Freeman and Company, 2005, 7 th Edition).

Hubble, Edwin. *A relationship between Distance and Radial Velocity among Extra-Galactic Nebulae* (from the proceedings of the

National Academy of Sciences Volumne15: March 15, 1929: Number 3

Kelsey, Linda J, Hoff, Darrel B., Neff, Johns, *Astronomy: Activities and Experiments* (University of Iowa Press, 1975)

Kleinbaum, David G. *Applied Regression Analysis and other Multivariable Methods* (Boston, MA.:PWS-KENT Publishing Company, 1988, 2 nd Edition).

Koo, David, C and Kron, Richard G. A Deep Redshift Survey of Field Galaxies, Space and Comments on the reality of the Butcher-Oemler Effect, Telescope Science Institute 3700 San Martin Drive Baltimore, MD , September 1987.

(Get this 15 page article only if you want an advanced discussion on the redshift.)

Yates, Daniels S. *The Practice of Statistics* (New York, NY.: W. H. Freeman and Company, 2003, 2 nd Edition)

On-line sources

---

http://articles.adsabs.harvard.edu/full/gif/seri/PASP./0108//0001074.000.html

(On page 2 of this site is a graph depicting the H_0 values over time. I found this most interesting. All science seems to improve with age and here is yet another example.)

http://astrosun2.astro.cornell.edu/academics/courses//astro201/bodes_law.htm

(An excellent descri8ptyion of Bode's law along with calculations and graphs.)

http://calculators.stat.ucla.edu/

(for those without a calculator or statistics program)

http://ciencies.udg.es/w3/EGarcia/humor.html

(a little bit of humor always helps)

http://en.wikipedia.org/wiki/

(This site is a listing of may galaxies and specific information about each one. It is very interesting and includes space photographs.)

http://hyperphysics.phy-astr.gsu.edu/hbase/astro/hubble.html

(If you want a graphic of the raisin bread analogy along with some calculations, this is a great site.)

http://hyperphysics.phy-astr.gsu.edu/hbase/astro/para.html

(Explains astronomical units of measure, Parallax or how we can tell the distance of stars, and gives examples of these measurements.)

http://imagine.gsfc.nasa.gov/

(This is the Goddard Space Flight Center web site. It contains a wealth of information about NASA and space in general. I highly recommend it as a general resource for any questions or just pleasure in learning about Astronomy.)

http://mathbits.com/MathBits/TISection/Statistics2/correlation.htm

(This is a secondary site for explanations about r ,R^2 , correlation coefficient and other aspects of correlation. It might be a helpful aid in reading about these relationships from another source.

http://news.nationalgeographic.com/news/2001/02/0216_Pluto.html

(Published in 2001 discusses whether or not Pluto is a planet as it is mostly ice and only a fraction of the size of Earth)

http://software.isixsigma.com/st/hypothesis_testing/

(excellent anthology of statistical concepts explanations)

http://visualstatistics.net/index.html

(for those students who like a visual an audio presentation of basic statistics)

http://www.astro.lsa.umich.edu/Course/MMSS/neargal.html.

(This is a University of Michigan publication that shows the distances to nearby galaxies. There are graphs, sketches and actual space photos of the galaxies. It also includes an excellent description of the Hubble Law and associated calculations.)

http://www.astronomynotes.com/galaxy/s7.htm

(Discusses distances to the galaxies and gives an excellent description of red and blue shift.)

http://www.cs.csubak.edu/Physics/phys110/NotesCh16.html

(Hubble's major achievements of classification of galaxies, properties of galaxies, redshift discussion and Hubble's law are clearly explained.)

http://www.geocities.com/jurgenshestani/hubble.html?200514

(This is another set of measurements to calculate H_0. I didn't use these numbers, but you may wish to include them or have the students work with them as another example.)

http://www.ipac.caltech.edu/H0kp/

(This is an index of galaxies which have been observed by the Hubble Telescope)

http://www.psychstat.smsu.edu/introbook/sbk00.htm

(a very complete on-line statistical text)

http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm

(This Yale publication is an excellent discussion of linear regression. It also discusses least-square and residuals.)

http://www.supernovae.net/galmod.htm

(For anyone wanting a list of galaxies, this is the place to go.)

Unless otherwise noted, all science facts and information in this unit are taken and combined from Freedman (2005).