Tip!

Before you begin:
Print out a hard copy of this document to use as you follow along.

Welcome to "Show Me"!

"Show me" is a series of activities that I am developing to accompany my courses. Each is designed to be: hands on, fun, and "real world" relevant to the course materials to follow. Thus, ideally, "show me" activities would be completed ***BEFORE*** the start of the associated course unit and, therefore, no pre-requisites would be required for their completion.

Help me out, please. If I've developed these "show me" correctly, all that is required is that you follow along. Thus, the challenge for me is to be sure that I've left no detail out. The challenge for you is to slow down and follow each step, word for word. Your feedback is key to its success. Please tell me what works, what doesn't work, and what additions you recommend.

Introduction and Goals:

Data in the form of numbers (eg – text, excel, or other) are everywhere. You may be asked to produce and communicate a summary. Some basic skills are involved. The goal of this "show me" is to give you experience with a selection of these skills, including:

(1) Accessing data from the internet (downloading, saving, utilizing),
(2) Understanding the structure and definitions associated with data (data dictionaries),
(3) Using an online statistical software tool to obtain data summaries (http://www.lock5stat), and
(4) Interpreting data summaries

Data

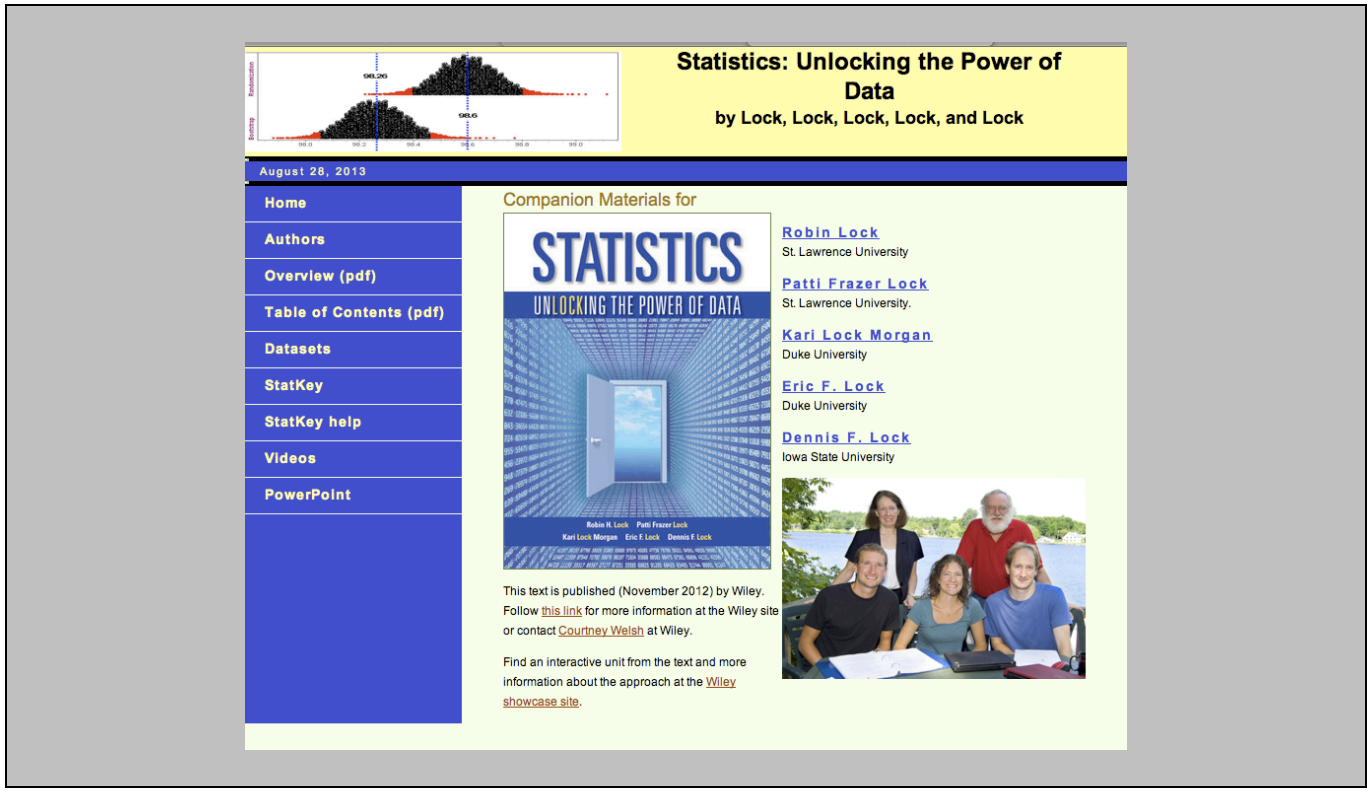
Source:
Lock RH, Lock PF, Morgan KL, Lock EF and Lock DF. *Statistics: Unlocking the Power of Data* New York, John Wiley, 2013.
Note – it is NOT necessary to purchase this book.

This "show me" utilizes a data set called **StudentSurvey**, that is freely available for download from the Lock, Lock, Morgan, Lock and Lock book. It is available in multiple formats. We will be working with the excel version, **StudentSurvey.xls.**

StudentSurvey contains information on 362 students in an introductory statistics class who completed an instructor administrated survey. For each student, 17 pieces of information (characteristic) were obtained (1st is their identification, 2nd is their gender, and so on). The 17 characteristics are called variables. In statistical parlance we say *"the sample size is 362" or "n=362"* and "*the number of variables is 17"*.

Activity #1. Launch lock5stat and look around.

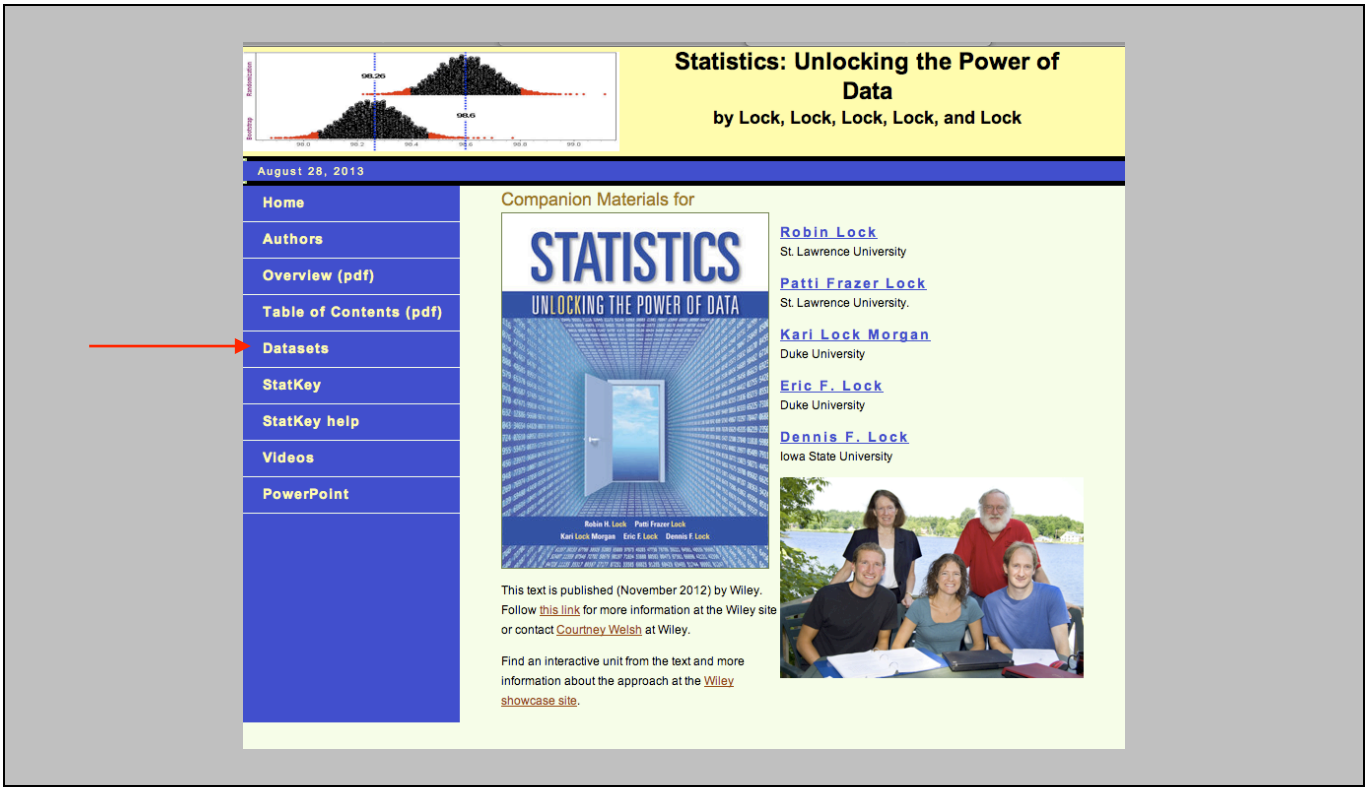
__1. Using whatever browser you have, launch http://www.lock5stat.com. A home page will appear:



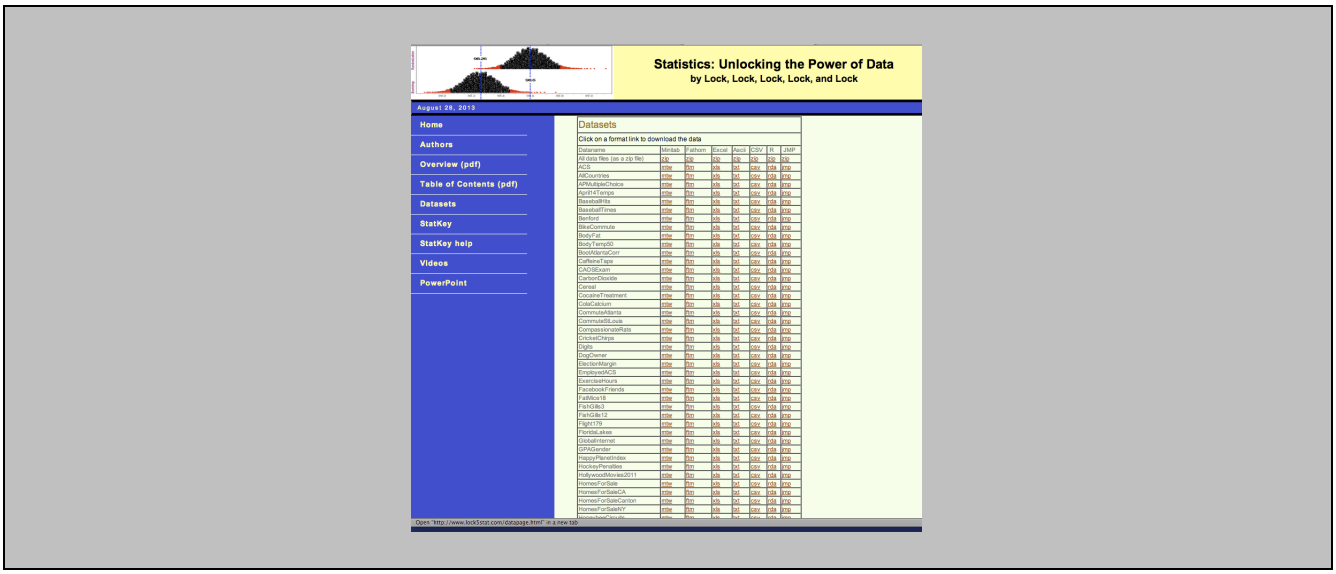
__2. Spend some time browsing this site.

Activity #2. Access the excel data set StudentSurvey.xls and download it to your computer.

__1. From the home page of http://www.lock5stat, click at left on the button, **Datasets**.



You will see a long table of data sets. They are listed in alphabetic order by dataset name.



__3. Scroll down to locate the row for **StudentSurvey**.
To download this data to your computer, click on the entry with the extension **.xls**.

__4. Your browser will provide some sort of drop down menu, select Download File. **Note – Don't panic, here. The options in your drop down menu might be slightly different, depending on whether you are a PC user or a MAC user. Play with the possibilities here.**

__5. Download **StudentSurvey.xls** to your desktop, or to some other location, that is easy for you to remember.

Activity #3. Familiarize yourself with the structure of the data.

Why? This is important because some pieces of information will be in the form of names. Others will be numbers. Still others will be numbers that are used to represent pieces of information that are actually names. These distinctions make a difference in producing summaries.

Take a look at the following table. It summarizes the structure of the information in **StudentSurvey**. This kind of documentation goes by various names, including ***coding manual*** or ***data dictionary***.

Variable Name	Label	Coding/Remarks
Year	Year on school	First Year, Sophomore, Junior, or Senior
Gender	Student gender	M for male, F for female
Smoke	Does the student smoke	yes or no
Award	Preferred award	Academy, Nobel, or Olympic
HigherSAT	Which SAT is higher	Math or Verbal
Exercise	Hours of exercise per week	Entered as whole number (integer) eg; 10
TV	Hours of TV viewing per week	Entered as whole number (integer) eg; 1
Height	Height in inches	Entered as whole number (integer) eg; 71
Weight	Weight in pounds	Entered as whole number (integer) eg; 180
Siblings	Number of siblings	Entered as whole number (integer) eg; 4
Birth	Birth order	1=oldest, 2=2nd oldest, etc
VerbalSAT	Verbal SAT score	Entered as whole number (integer) eg; 540
MathSAT	Math SAT score	Entered as whole number (integer) eg; 670
SAT	Combined Verbal + Math SAT	= (verbal SAT) + (math SAT)
GPA	Colleg grade point average on a 4-point scale	eg; 3.13
Pulse	Pulse rate (beats per minute)	Entered as whole number (integer) eg; 54
Piercings	Number of body piercings	Entered as whole number (integer) eg; 0

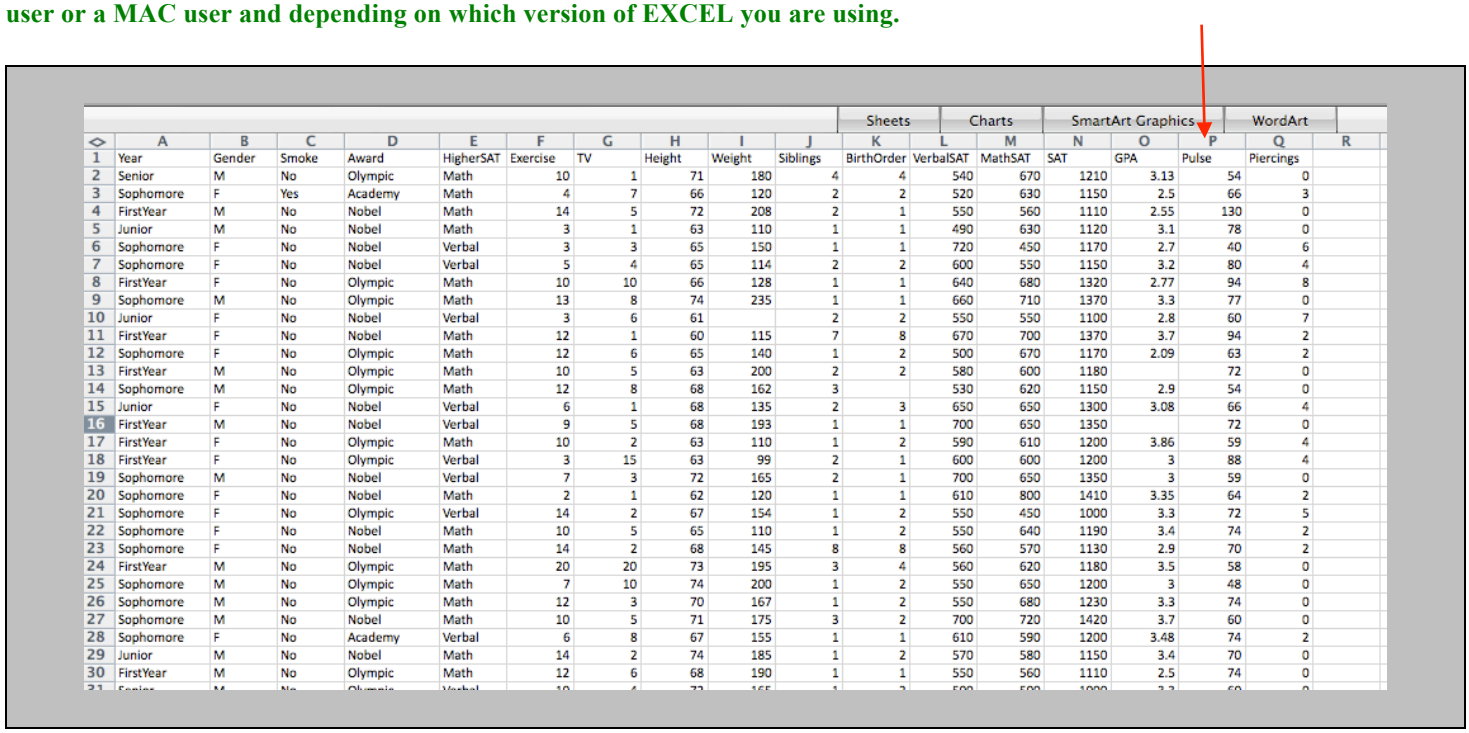
"Show me": If I were to show you the raw data, you might see that for student #1 we have:

Year	Gender	Smoke	Award	HigherSAT	Exercise	TV	Height	..
Senior	M	No	Olympic	Math	10	1	71	…

This student: is a senior, is male, does not smoke, would prefer to win an Olympic gold medal, scored higher on the Math SAT (compared to the Verbal SAT), exercises 10 hours per week, watches television for 1 hour per week, is 71 inches tall, and so on….

Activity #4. In a separate window that you will keep open, launch excel and open the file StudentSurvey.xls.

__1. You should see the following. **Note – Yours might not look exactly the same, depending on whether you are a PC user or a MAC user and depending on which version of EXCEL you are using.**



Sheets | Charts | SmartArt Graphics | WordArt

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Year	Gender	Smoke	Award	HigherSAT	Exercise	TV	Height	Weight	Siblings	BirthOrder	VerbalSAT	MathSAT	SAT	GPA	Pulse	Piercings	
2	Senior	M	No	Olympic	Math	10	1	71	180	4	4	540	670	1210	3.13	54	0	
3	Sophomore	F	Yes	Academy	Math	4	7	66	120	2	2	520	630	1150	2.5	66	3	
4	FirstYear	M	No	Nobel	Math	14	5	72	208	2	1	550	560	1110	2.55	130	0	
5	Junior	M	No	Nobel	Math	3	1	63	110	1	1	490	630	1120	3.1	78	0	
6	Sophomore	F	No	Nobel	Verbal	3	3	65	150	1	1	720	450	1170	2.7	40	6	
7	Sophomore	F	No	Nobel	Verbal	5	4	65	114	2	2	600	550	1150	3.2	80	4	
8	FirstYear	F	No	Olympic	Math	10	10	66	128	1	1	640	680	1320	2.77	94	8	
9	Sophomore	M	No	Olympic	Math	13	8	74	235	1	1	660	710	1370	3.3	77	0	
10	Junior	F	No	Nobel	Verbal	3	6	61		2	2	550	550	1100	2.8	60	7	
11	FirstYear	F	No	Nobel	Math	12	1	60	115	7	8	670	700	1370	3.7	94	2	
12	Sophomore	F	No	Olympic	Math	12	6	65	140	1	2	500	670	1170	2.09	63	2	
13	FirstYear	M	No	Olympic	Math	10	5	63	200	2	2	580	600	1180		72	0	
14	Sophomore	M	No	Olympic	Math	12	8	68	162	3		530	620	1150	2.9	54	0	
15	Junior	F	No	Nobel	Verbal	6	1	68	135	2	3	650	650	1300	3.08	66	4	
16	FirstYear	M	No	Nobel	Verbal	9	5	68	193	1	1	700	650	1350		72	0	
17	FirstYear	F	No	Olympic	Math	10	2	63	110	1	2	590	610	1200	3.86	59	4	
18	FirstYear	F	No	Olympic	Verbal	3	15	63	99	2	1	600	600	1200	3	88	4	
19	Sophomore	M	No	Nobel	Verbal	7	3	72	165	2	1	700	650	1350	3	59	0	
20	Sophomore	F	No	Nobel	Math	2	1	62	120	1	1	610	800	1410	3.35	64	2	
21	Sophomore	F	No	Olympic	Verbal	14	2	67	154	1	2	550	450	1000	3.3	72	5	
22	Sophomore	F	No	Nobel	Math	10	5	65	110	1	2	550	640	1190	3.4	74	2	
23	Sophomore	F	No	Nobel	Math	14	2	68	145	8	8	560	570	1130	2.9	70	2	
24	FirstYear	M	No	Olympic	Math	20	20	73	195	3	4	560	620	1180	3.5	58	0	
25	Sophomore	M	No	Olympic	Math	7	10	74	200	1	2	550	650	1200	3	48	0	
26	Sophomore	M	No	Olympic	Math	12	3	70	167	1	2	550	680	1230	3.3	74	0	
27	Sophomore	M	No	Nobel	Math	10	5	71	175	3	2	700	720	1420	3.7	60	0	
28	Sophomore	F	No	Academy	Verbal	6	8	67	155	1	1	610	590	1200	3.48	74	2	
29	Junior	M	No	Nobel	Math	14	2	74	185	1	2	570	580	1150	3.4	70	0	
30	FirstYear	M	No	Olympic	Math	12	6	68	190	1	1	550	560	1110	2.5	74	0	

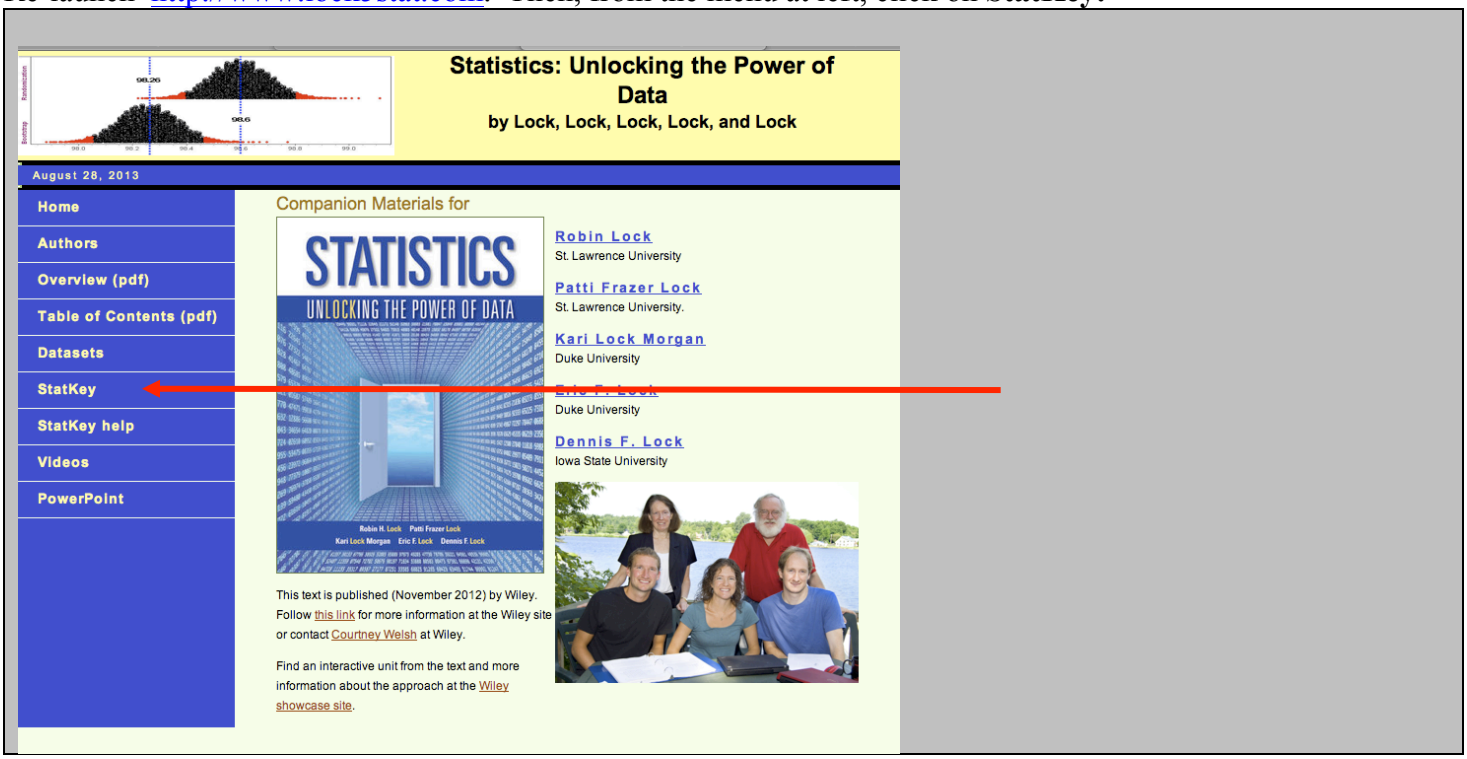
__2. Position your cursor over the "P" that is at the top of and that refers to column "P" (Pulse).
__ Select this entire column of data by clicking on the "P" that is the column heading
__ Next, from the main menu in Excel, click **EDIT > COPY** to save this entire column for pasting elsewhere.

You will do the pasting into StatKey later.
Note – StatKey is a tool within www.lock5stat.com. It is accessed from the homepage menu bar at left.

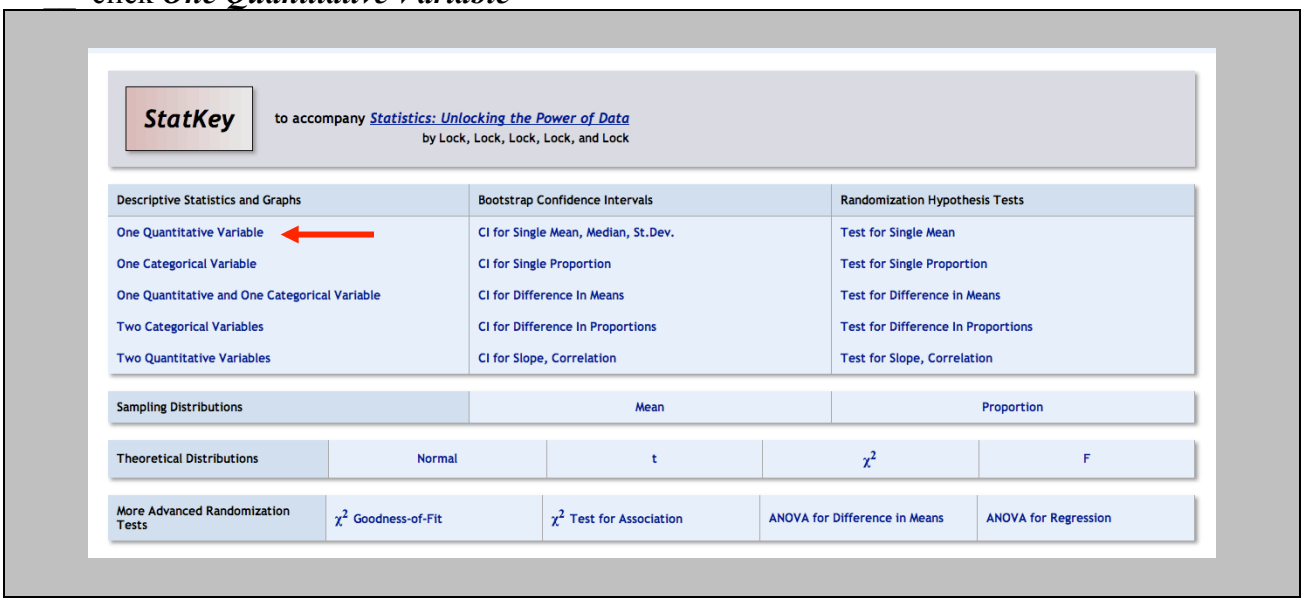
__3. **Minimize Excel** window
__ but ***do NOT exit*** Excel. We will be coming back to this.

Activity #5. Launch the StatKey tool. Then, under *Descriptive Statistics and Graphs*, click *One Quantitative Variable*

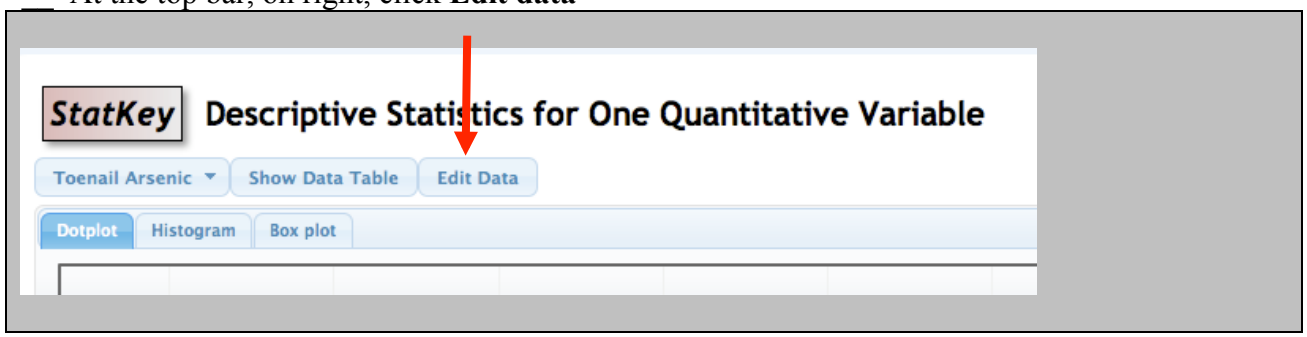
__1. Re-launch http://www.lock5stat.com. Then, from the menu at left, click on **StatKey:**



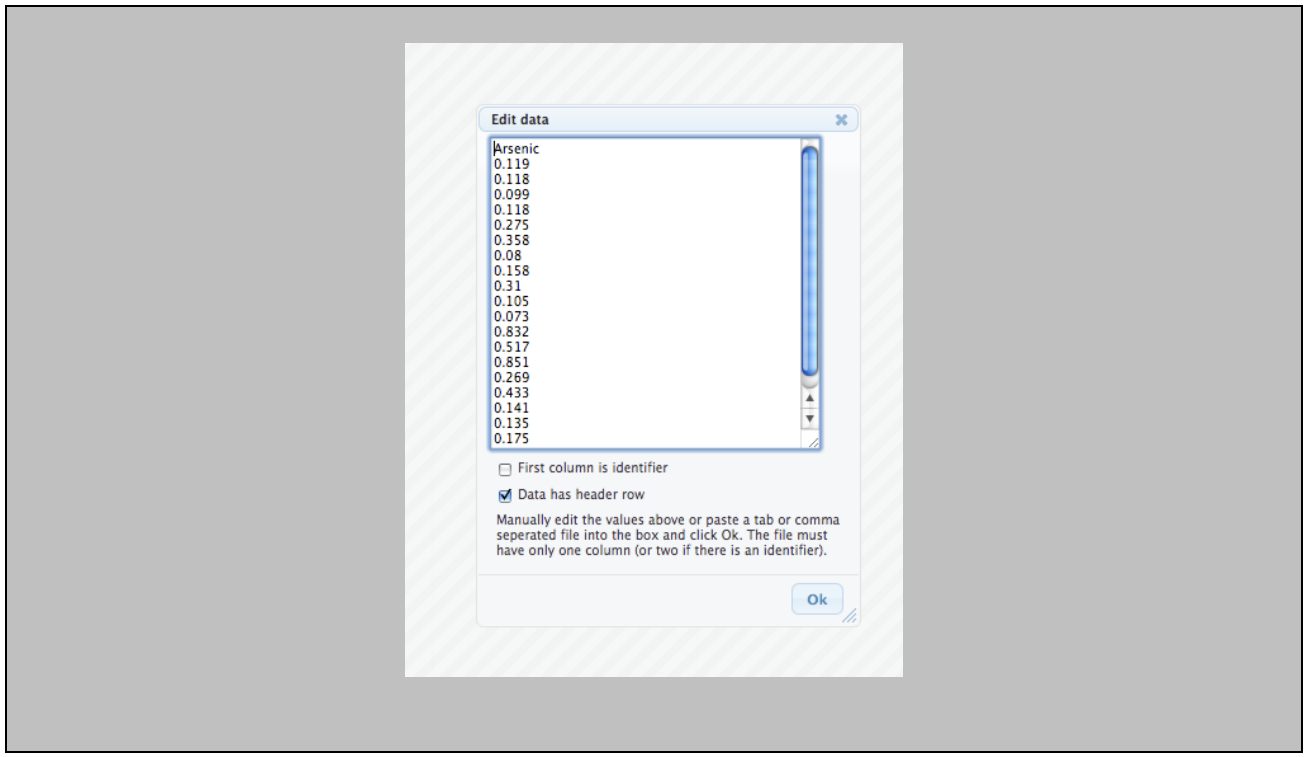
___2. From the selection of analysis options offered, under *Descriptive Statistics and Graphs*, __ click ***One Quantitative Variable***



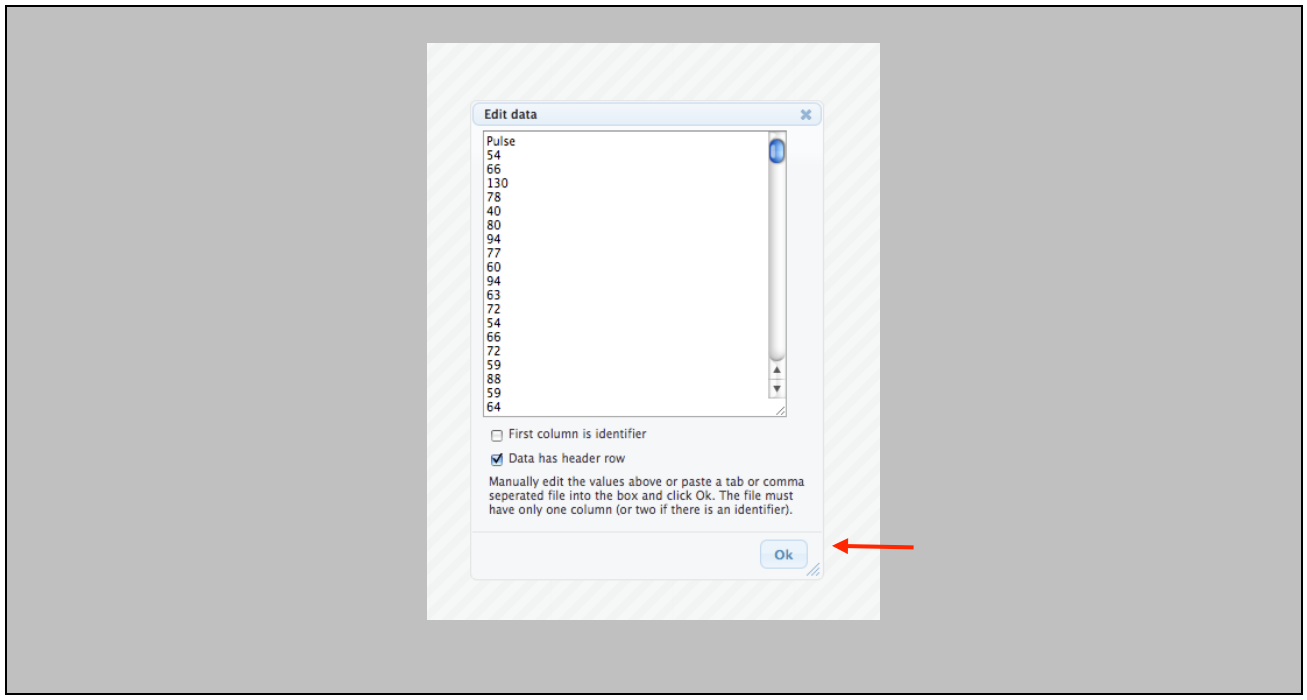
___3. Do the following to replace the default data (a sample of arsenic values) with the pulse data of interest.
__ At the top bar, on right, click **Edit data**



___4. Delete the default, arsenic, data by doing the following.
___ Using your cursor, position and drag to select all values of arsenic.
___ Click on the delete key on your keyboard.
___ ***Important*** – Do ***NOT*** click the ok button just yet.



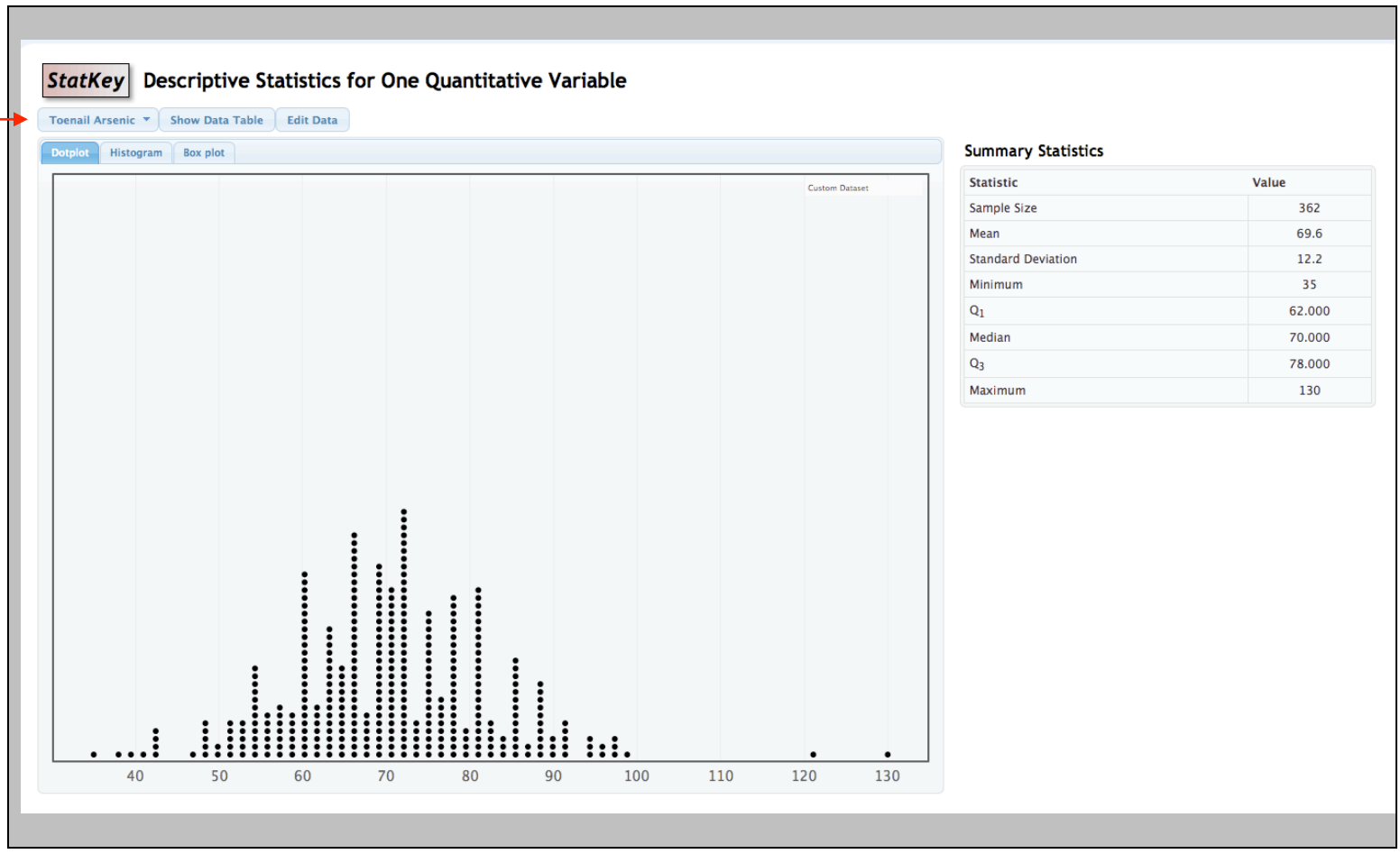
___5. Now you can paste in the pulse data values from your excel file by doing the following.
__ Position your cursor at the top of the now empty data box of StatKey.
__ Paste in your pulse data by using **EDIT > PASTE** located in the main tool bar of your browser.
__ BOX CHECK: Check to see that the box next to "Data has a header row" is checked.
__ Click on **OK** at bottom right.



Note – Unfortunately, StatKey doesn't show you the data set name Custom. Instead, it retains the data set name "Toenail Arsenic"

"Show me": **StatKey returns a dot plot summary of the distribution of the values of pulse among the n=362, together with some numerical summaries (mean, standard deviation, etc) at right.**

Activity #6. Interpret the summaries you just produced for the 362 observations of pulse.
Note – Unfortunately, StatKey doesn't show you the data set name Custom. Instead, it retains the data set name "Toenail Arsenic"



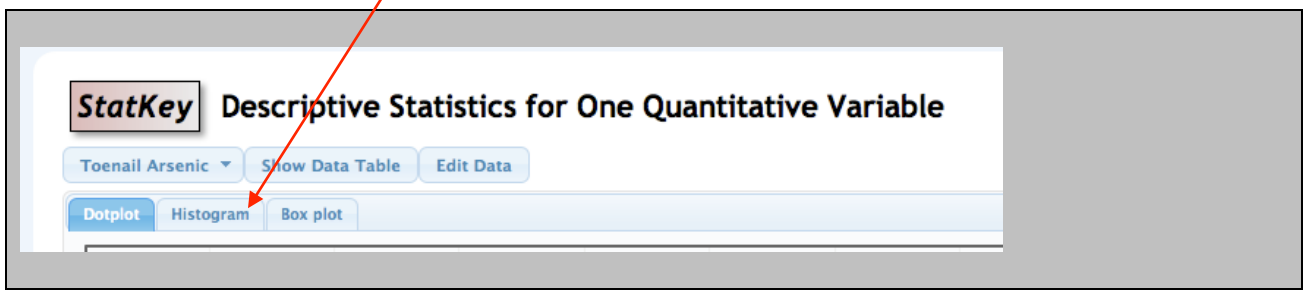
"Show me":

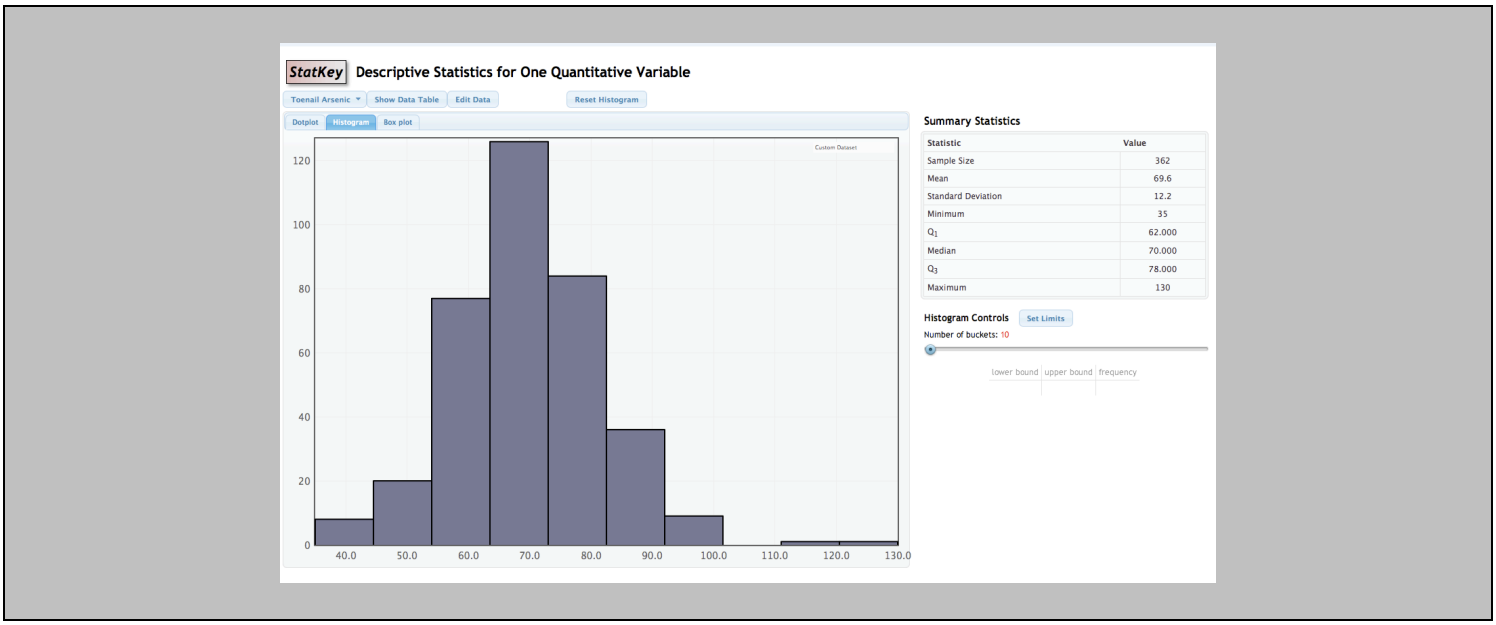
- The variable **pulse** is an example of a **quantitative variable** that is **continuous**. Its values are numbers and between any two numbers (eg 100 and 101), it is theoretically possible for an in-between value to occur (eg a person might have a pulse of 100.5)
- The scatter of values of pulse range from a low of 35 beats/min to a high of 130 beats/min.
- The pattern of scatter looks, for the most part, bell shaped. 50% are between 62 (**Q1**) and 78 (**Q3**) beats/min
- We see 2 very high values, one at 122 (approx) and the other at 130
- The **median** pulse is 70; 50% of the sample have pulse rates that are lower than 70 beats/min.
- The average (**mean**) pulse obtained by summing over the 362 pulses and dividing by 362 is 69.6.
- A typical scatter (**standard deviation**) of an individual pulse away from the average is 12.2 beats/min.
- Possible graphical summaries for **continuous** variables include: **dot plot, histogram, box plot**

"Show me": **StatKey also returns a histogram summary of the distribution of the values of pulse among the n=362, together with the same numerical summaries (mean, standard deviation, etc) at right.**

Activity #7. View the histogram graphical summary for the variable pulse.

___1. Easy. At the top bar, click **Histogram**





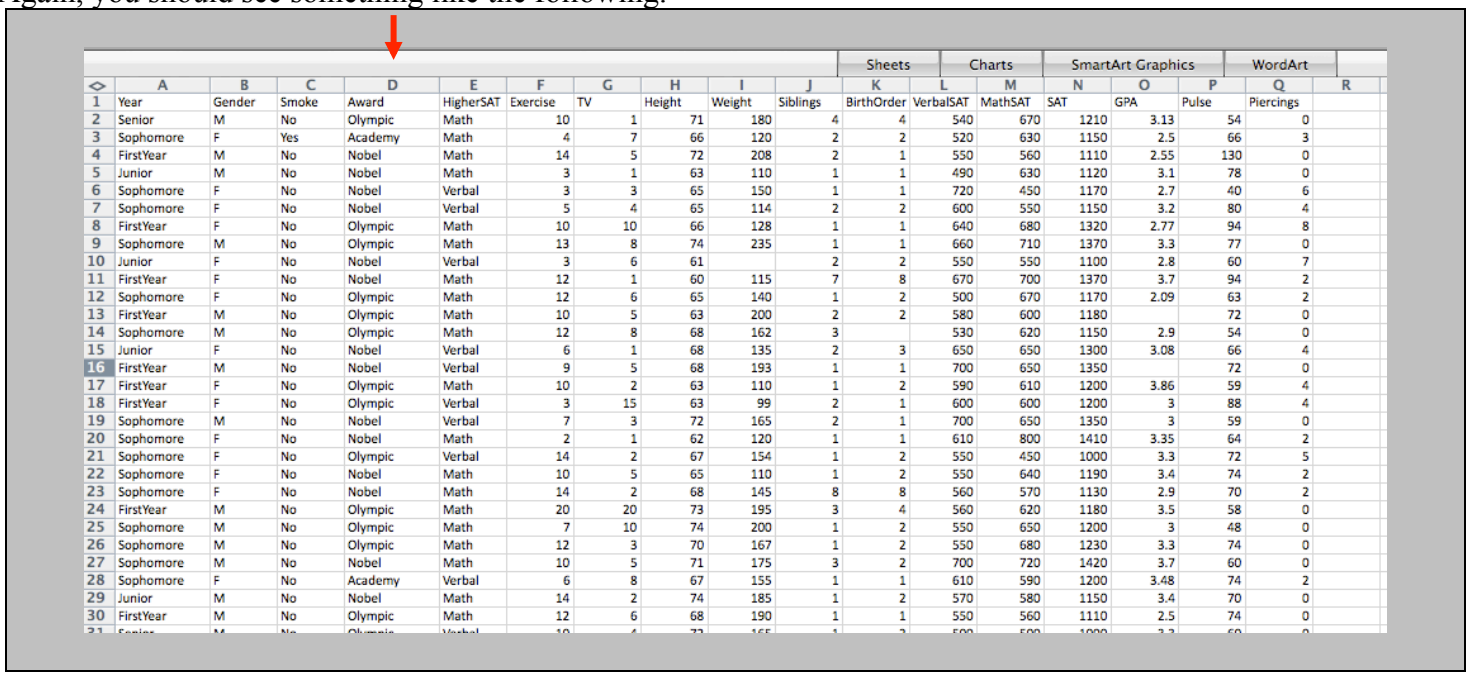
"Show me":

- This **histogram** is a graph of the values of the **continuous variable** pulse, grouped.
- Each interval of pulse spans 10 beats/min and there are 10 intervals.
- Most of the pulse values are between 60 and 90 beats/min
- There are no observations of pulse in the interval 100-110 beats/min.

Activities #8 and #9 pertain to the categorical variable, award. Tip - The summaries that are possible for any given variable are different, depending on the variable type.

Activity #8. Activate the EXCEL window that you minimized earlier.

__1. To do this, maximize the EXCEL window that you minimized previously.
Again, you should see something like the following.



Sheets | Charts | SmartArt Graphics | WordArt

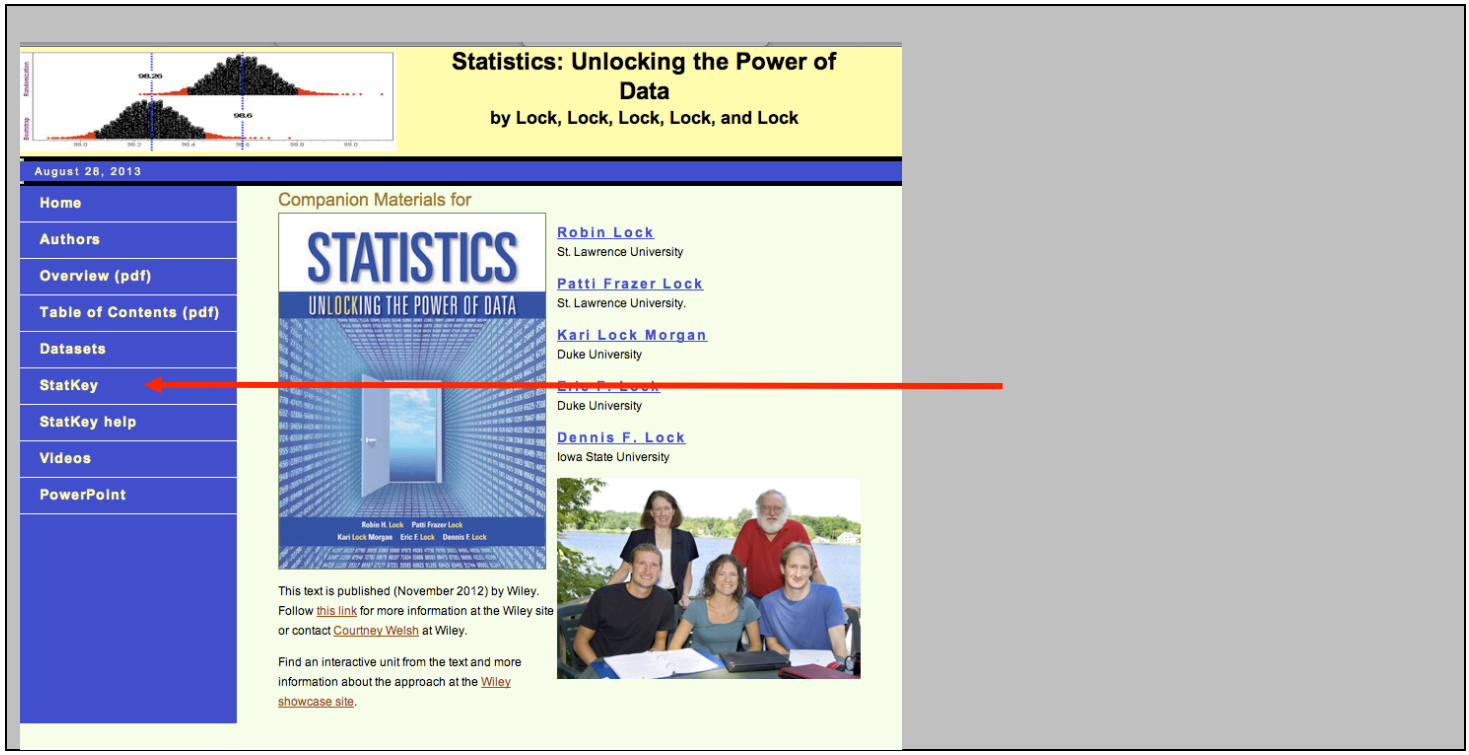
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Year	Gender	Smoke	Award	HigherSAT	Exercise	TV	Height	Weight	Siblings	BirthOrder	VerbalSAT	MathSAT	SAT	GPA	Pulse	Piercings	
2	Senior	M	No	Olympic	Math	10	1	71	180	4	4	540	670	1210	3.13	54	0	
3	Sophomore	F	Yes	Academy	Math	4	7	66	120	2	2	520	630	1150	2.5	66	3	
4	FirstYear	M	No	Nobel	Math	14	5	72	208	2	1	550	560	1110	2.55	130	0	
5	Junior	M	No	Nobel	Math	3	1	63	110	1	1	490	630	1120	3.1	78	0	
6	Sophomore	F	No	Nobel	Verbal	3	3	65	150	1	1	720	450	1170	2.7	40	6	
7	Sophomore	F	No	Nobel	Verbal	5	4	65	114	2	2	600	550	1150	3.2	80	4	
8	FirstYear	F	No	Olympic	Math	10	10	66	128	1	1	640	680	1320	2.77	94	8	
9	Sophomore	M	No	Olympic	Math	13	8	74	235	1	1	660	710	1370	3.3	77	0	
10	Junior	F	No	Nobel	Verbal	3	6	61		2	2	550	550	1100	2.8	60	7	
11	FirstYear	F	No	Nobel	Math	12	1	60	115	7	8	670	700	1370	3.7	94	2	
12	Sophomore	F	No	Olympic	Math	12	6	65	140	1	2	500	670	1170	2.09	63	2	
13	FirstYear	M	No	Olympic	Math	10	5	63	200	2	2	580	600	1180		72	0	
14	Sophomore	M	No	Olympic	Math	12	8	68	162	3		530	620	1150	2.9	54	0	
15	Junior	F	No	Nobel	Verbal	6	1	68	135	2	3	650	650	1300	3.08	66	4	
16	FirstYear	M	No	Nobel	Verbal	9	5	68	193	1	1	700	650	1350		72	0	
17	FirstYear	F	No	Olympic	Math	10	2	63	110	1	2	590	610	1200	3.86	59	4	
18	FirstYear	F	No	Olympic	Verbal	3	15	63	99	2	1	600	600	1200	3	88	4	
19	Sophomore	M	No	Nobel	Verbal	7	3	72	165	2	1	700	650	1350	3	59	0	
20	Sophomore	F	No	Nobel	Math	2	1	62	120	1	1	610	800	1410	3.35	64	2	
21	Sophomore	F	No	Olympic	Verbal	14	2	67	154	1	2	550	450	1000	3.3	72	5	
22	Sophomore	F	No	Nobel	Math	10	5	65	110	1	2	550	640	1190	3.4	74	2	
23	Sophomore	F	No	Nobel	Math	14	2	68	145	8	8	560	570	1130	2.9	70	2	
24	FirstYear	M	No	Olympic	Math	20	20	73	195	3	4	560	620	1180	3.5	58	0	
25	Sophomore	M	No	Olympic	Math	7	10	74	200	1	2	550	650	1200	3	48	0	
26	Sophomore	M	No	Olympic	Math	12	3	70	167	1	2	550	680	1230	3.3	74	0	
27	Sophomore	M	No	Nobel	Math	10	5	71	175	3	2	700	720	1420	3.7	60	0	
28	Sophomore	F	No	Academy	Verbal	6	8	67	155	1	1	610	590	1200	3.48	74	2	
29	Junior	M	No	Nobel	Math	14	2	74	185	1	2	570	580	1150	3.4	70	0	
30	FirstYear	M	No	Olympic	Math	12	6	68	190	1	1	550	560	1110	2.5	74	0	

__2. Select all of the data on **award** in column D:
___ Position your cursor over the "D" that is at the top of and that refers to column "D" (Award).
___ Click on the "D" that is the column heading.
___ Again, from the main menu in Excel, click **EDIT > COPY** to select this entire column of data.

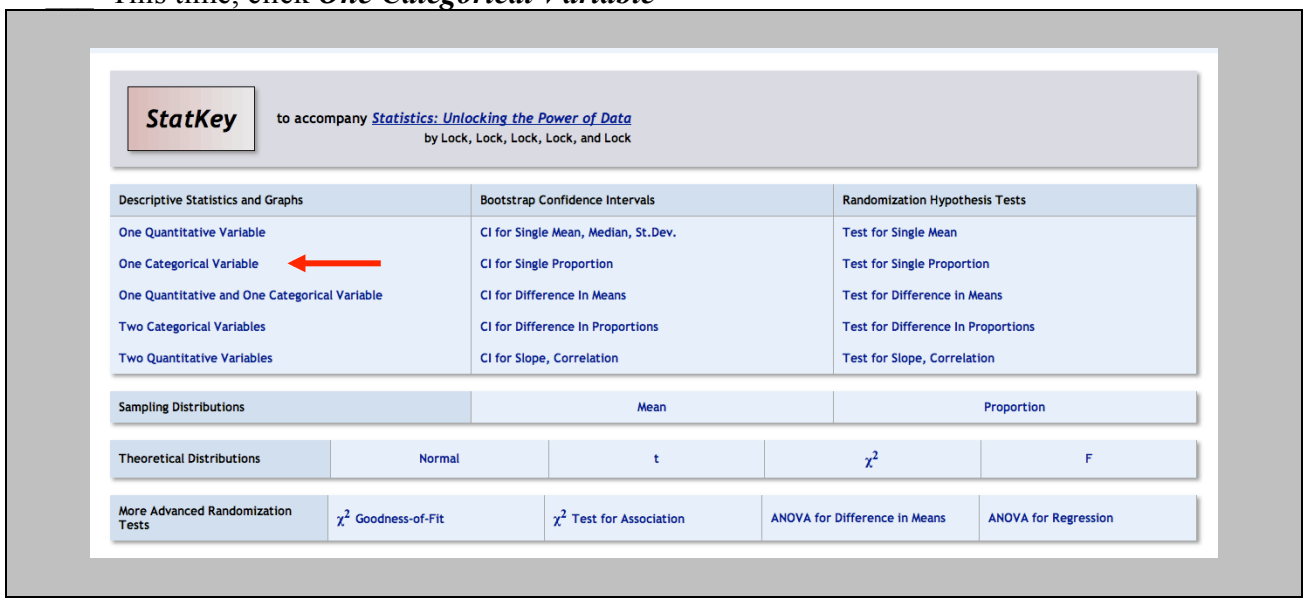
__3. **Minimize Excel** window.
Now it's your choice whether or not to exit Excel **….perhaps you want to look at some other data on your own?**

Activity #9. Activate the window that is at the internet site lock5stat.com and obtain summaries for the variable AWARD.

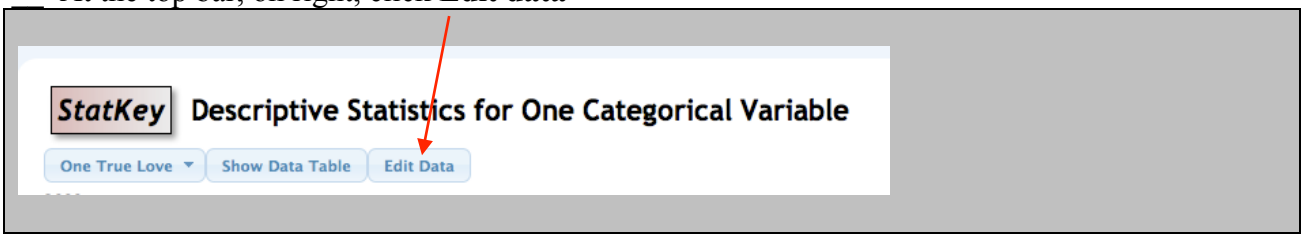
__1. Re-launch (or re-activate) http://www.lock5stat.com.
As before, from the menu at left, click on **StatKey:**



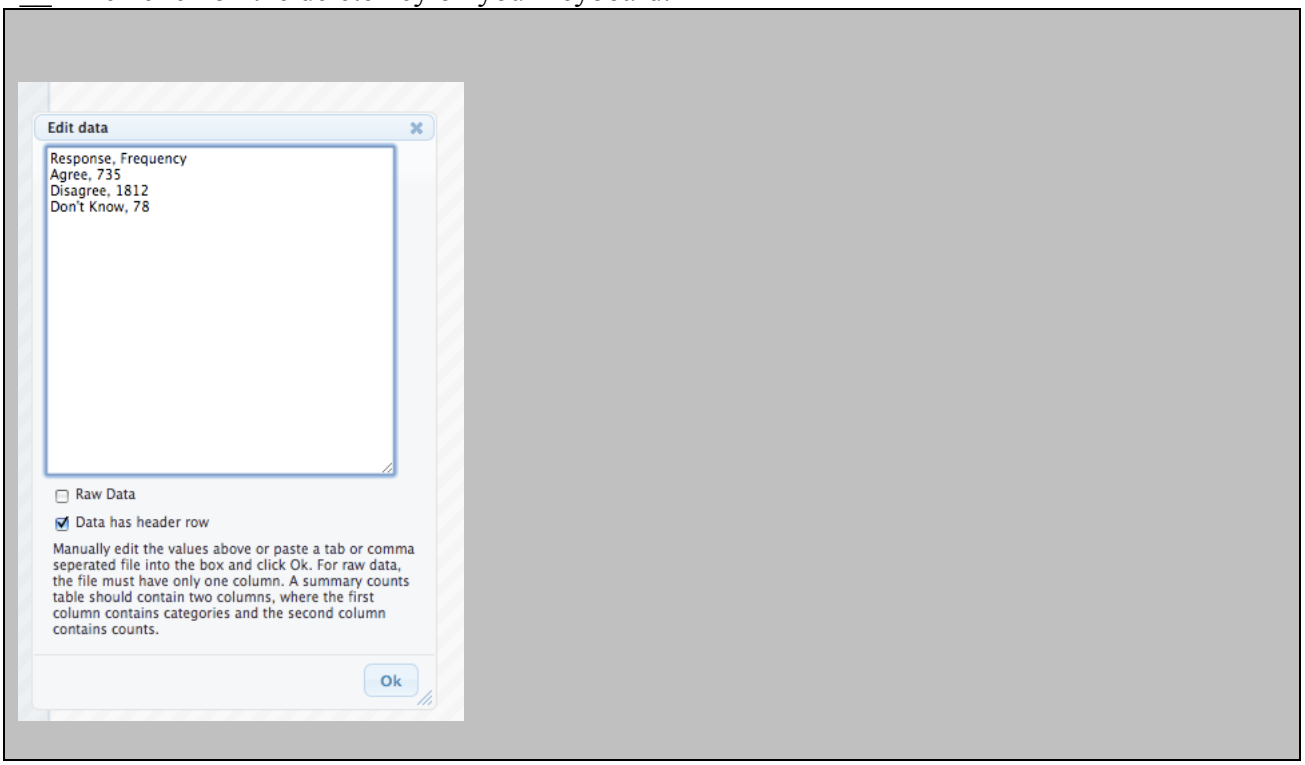
___2. From the menu choices, under *Descriptive Statistics and Graphs*,
___ This time, click ***One Categorical Variable***



___3. As before, replace the default data (values for "one true love") with the data on award
__ At the top bar, on right, click **Edit data**



___4. Delete the "One True Love" data that is there.
__ Using your cursor, position and drag to select all the data values.
__ Then click on the delete key on your keyboard.



___5. Paste the award data:

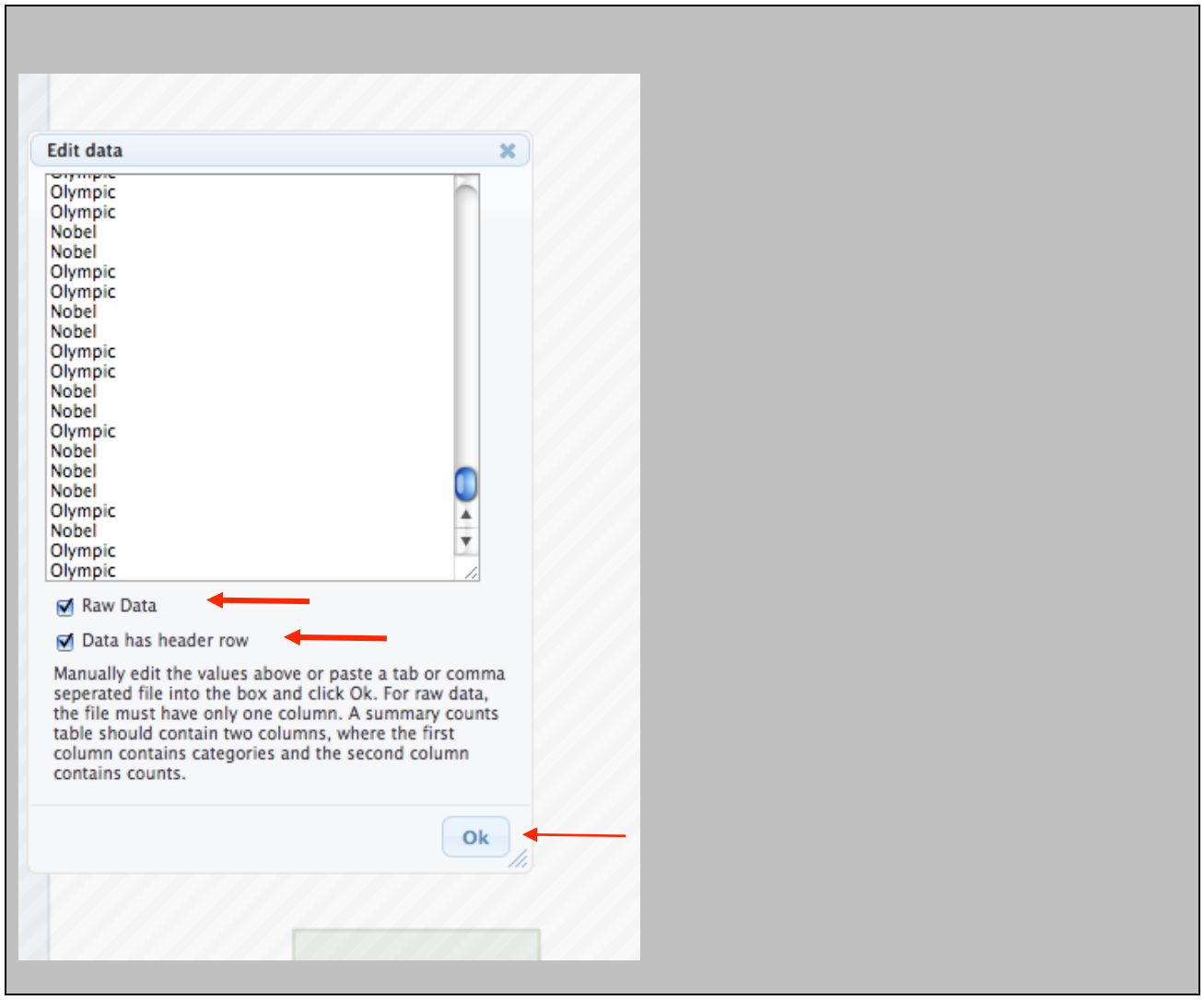
__ Position your cursor at the top of the now empty data box.

__ Paste in your award data by using **EDIT > PASTE** located in the main tool bar of your browser.

__ BOX CHECK: Check to see that the box next to "Data has a header row" is checked.

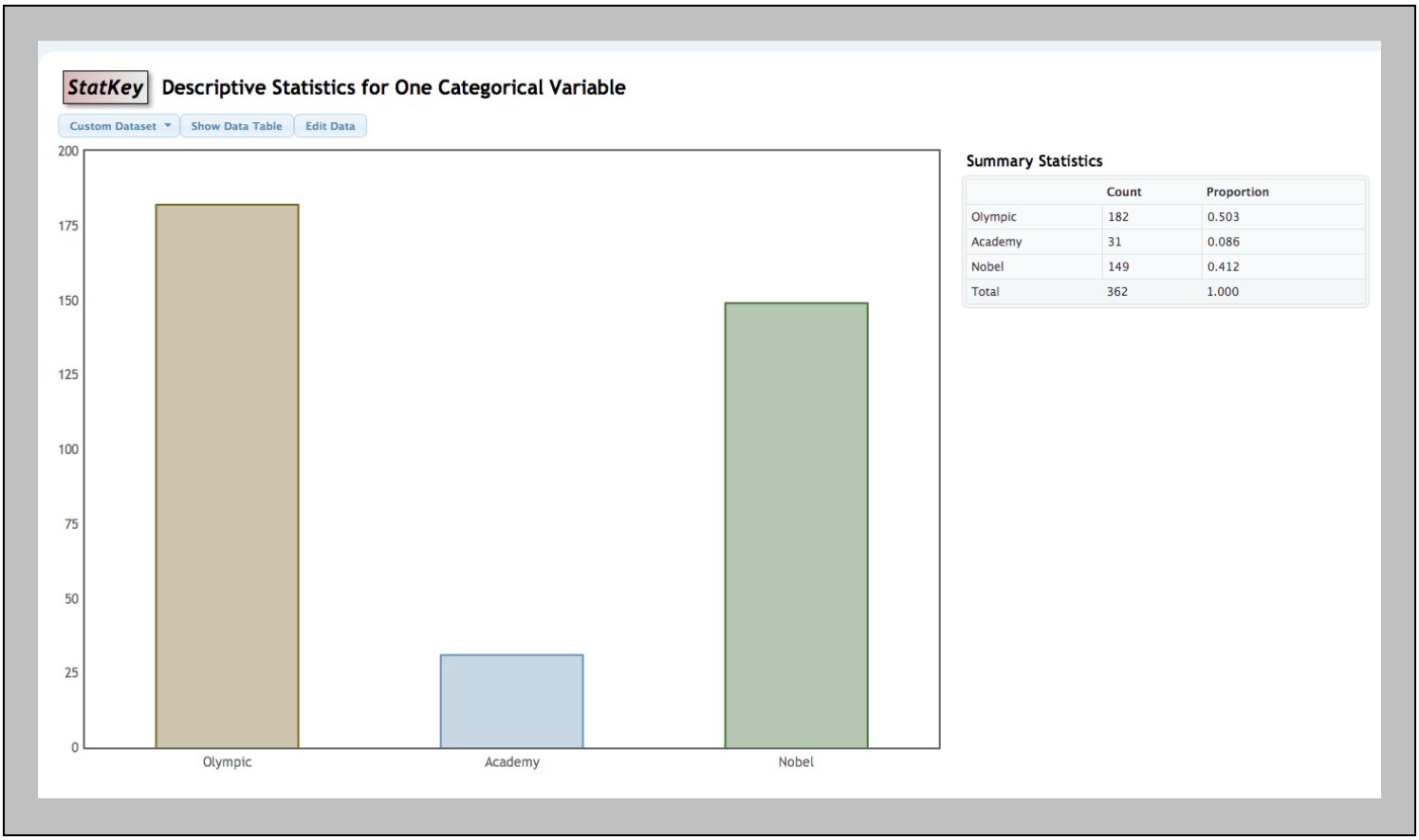
__ BOX CHECK: Check to see that the box next to "Raw Data" is checked.

__ Click on **OK** at bottom right.



"Show me":

StatKey returns a bar chart summary of the distribution of the values of award among the n=362, together with some numerical summaries (counts and proportions) at right.



"Show me":

- The variable **award** is an example of a **qualitative variable (a nominal one, actually)** and is **discrete** or **categorical**. Its values are names or labels and in this case are unordered. If these names were ordered (eg "mild", "moderate", "severe"), we would say this is ordinal data. But the values of award are not ordered, so this is an example of a **nominal categorical variable**.
- A defining feature of a **bar graph** (also called **bar chart**) is that, on the horizontal axis, the possible values ("olympic", "academy" and "nobel") are separated by spaces.
- The height of each bar is equal to the **count** or **frequency** of that response in the data set. For example, a count of 182 respondents said they would prefer to win an Olympic Medal.
- Division of the count by the total (362) gives the **proportion** or **relative frequency** of each response. For example, 182/362 is .503. This is saying that 50.3% of the respondents said they would prefer to win an Olympic Medal.
- The counts (**frequencies**) add up to 362 (the total sample size) and the proportions (**relative frequencies**) add up to 1 or 100% (everyone is accounted for).
- Possible graphical summaries for **categorical** variables include: **bar chart, pie chart** (I hate them).

"Show me":

Putting it all together:

Type of Variable	Appropriate Summary	NOT OKAY to do
Continuous: Measured as a number on a continuum. Between any two values, an intermediate values is theoretically possible. **Example:** **pulse** (beats/minute)	Graphical: dot plot, histogram, + some others Numerical: means, medians, standard deviations, percentiles	Graphical: bar chart + some others Numerical: counts/frequencies, proportions/relative frequencies + some others when the categorical variable is also ordinal
Categorical: Recorded as a group classification. The group classification can be as simple as a name with no order (nominal), a name with order, or a whole number. Between any two group classifications, it is NOT theoretically possible for an intermediate to occur. **Example:** **award** (academy, nobel, olympic)	Graphical: bar chart + some others Numerical: counts/frequencies, proportions/relative frequencies + some others when the categorical variable is also ordinal	Graphical: dot plot, histogram, + some others Numerical: means, medians, standard deviations, percentiles