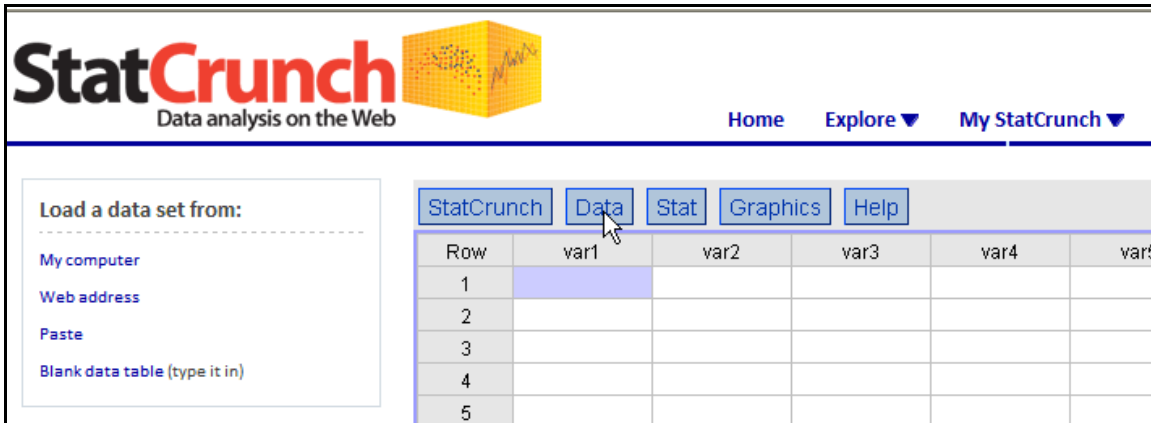
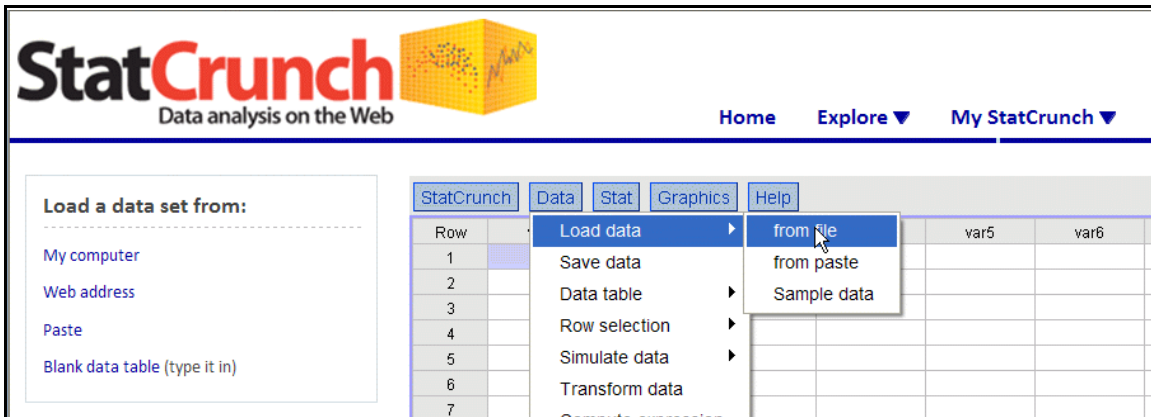


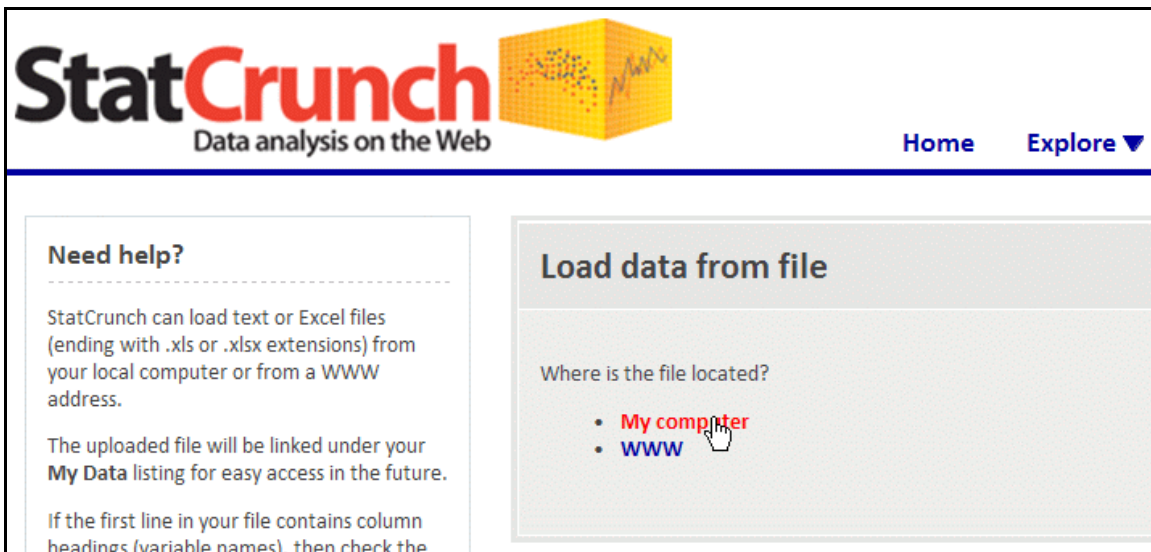
#1 START STATCRUNCH AND CLICK ON DATA



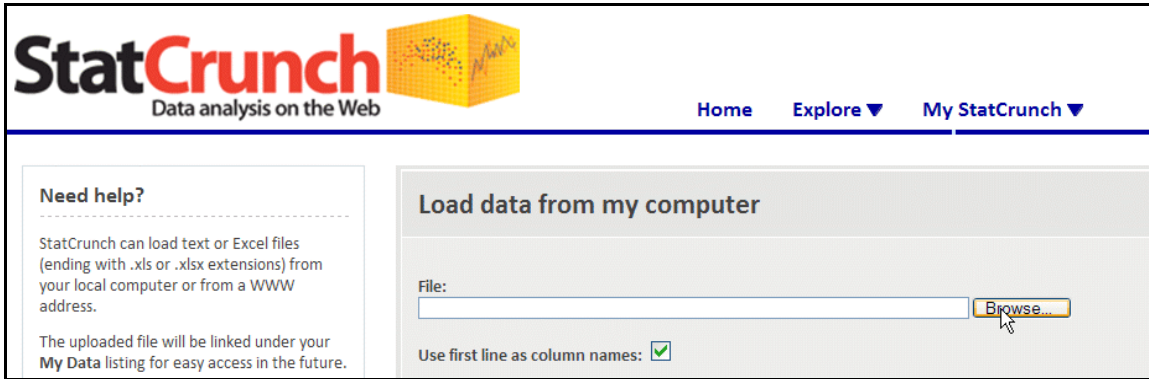
#2 THEN CLICK ON LOAD DATA / FROM FILE



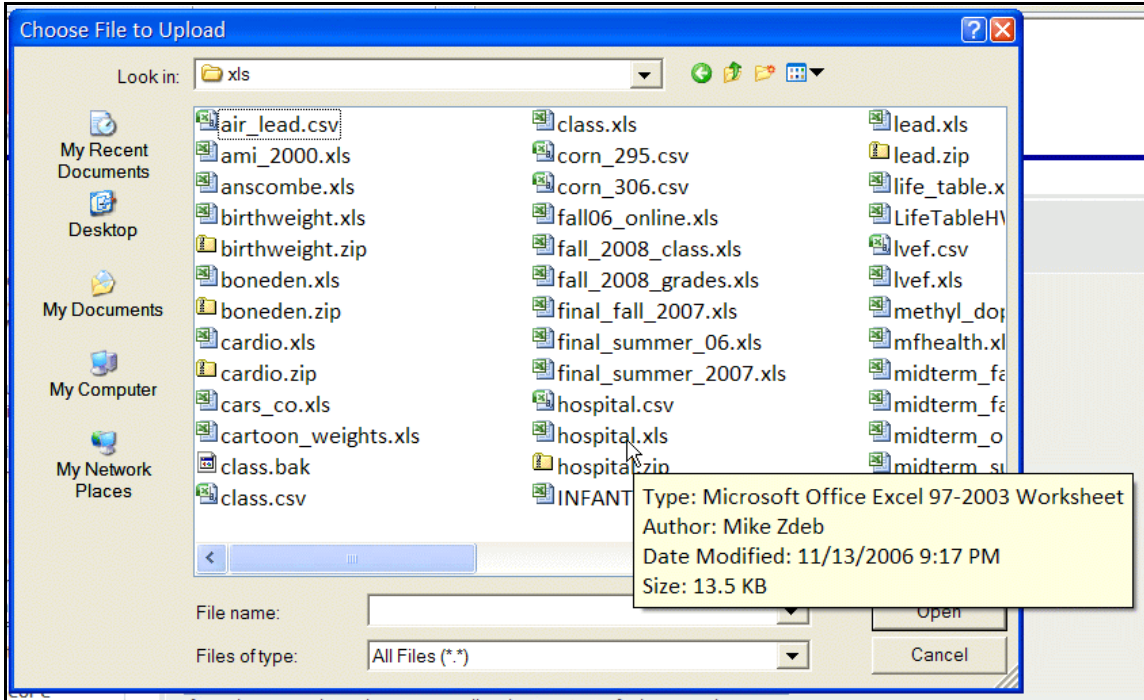
#3 CLICK ON MY COMPUTER



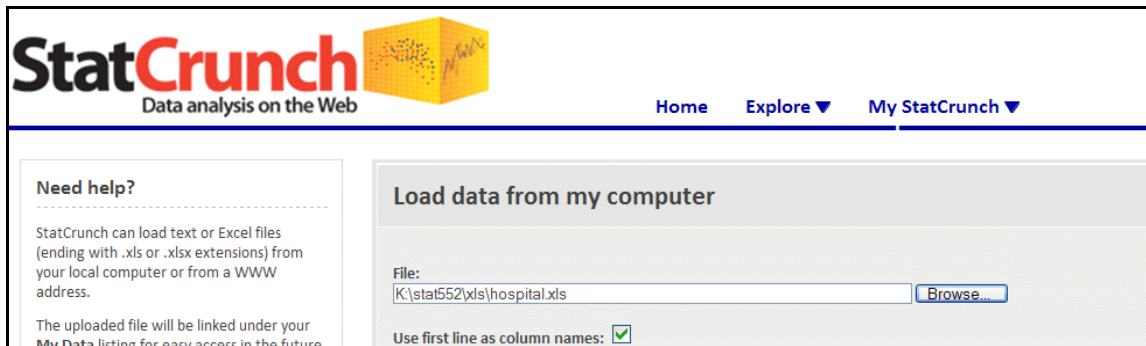
#4 CLICK ON BROWSE



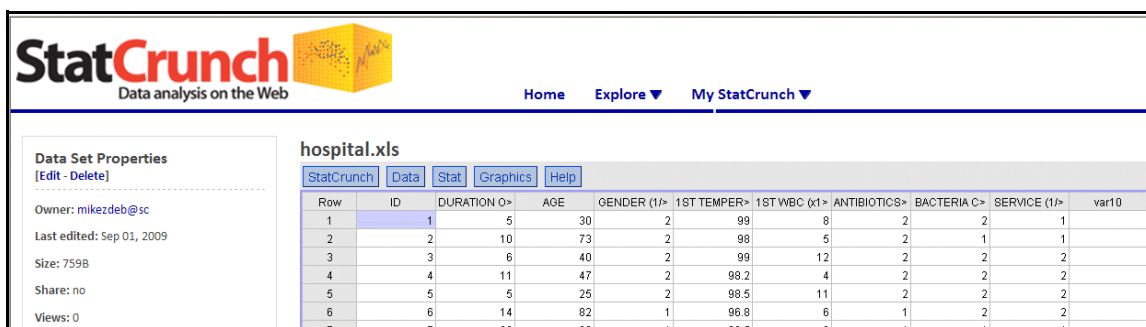
#5 YOU SHOULD SEE A WINDOW THAT YOU CAN USE TO FIND THE FILE HOSPITAL.XLS ON YOUR COMPUTER ... DOUBLE CLICK ON THE FILE NAME



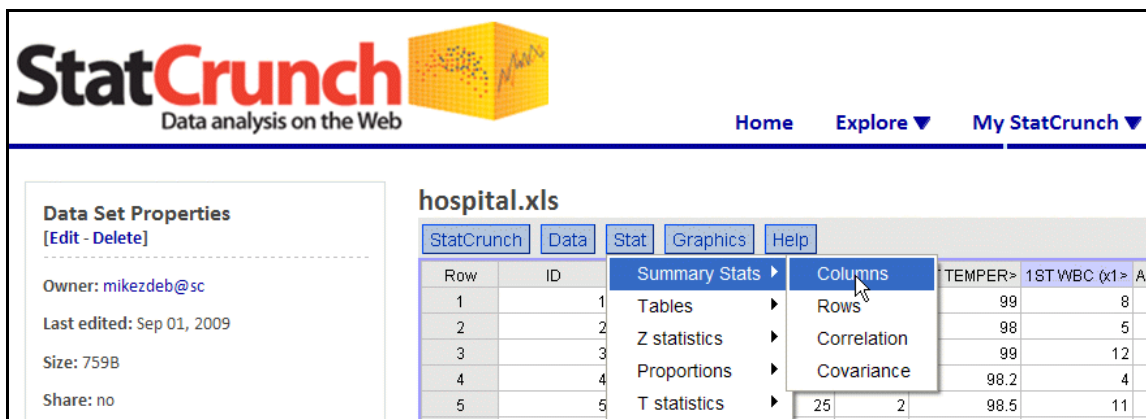
#6 THE NAME SHOULD APPEAR IN THE FILE WINDOW IN STATCRUNCH ... SCROLL TO THE BOTTOM OF THE PAGE AND CLICK ON LOAD FILE



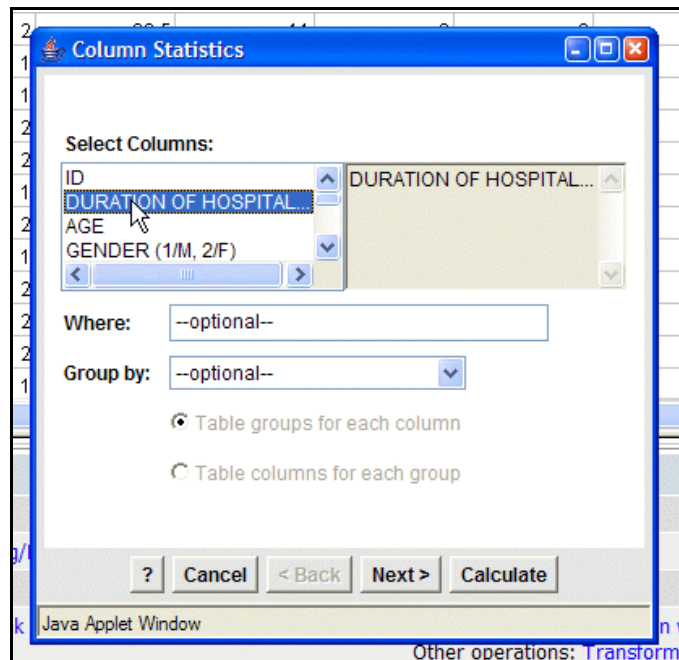
#7 YOU SHOULD SEE THE DATA IN THE STATCRUNCH SPREADSHEET (WHEN YOU SEE A COLUMN HEADING THAT ENDS WITH A ">" THAT MEANS THAT THE LABEL IS LONGER THAN WHAT CURRENTLY APPEARS ON THE SCREEN)



#8 CLICK ON STAT / SUMMARY STATS / COLUMNS



#9 CLICK ON THE VARIABLE DURATION ... THEN CLICK ON CALCULATE

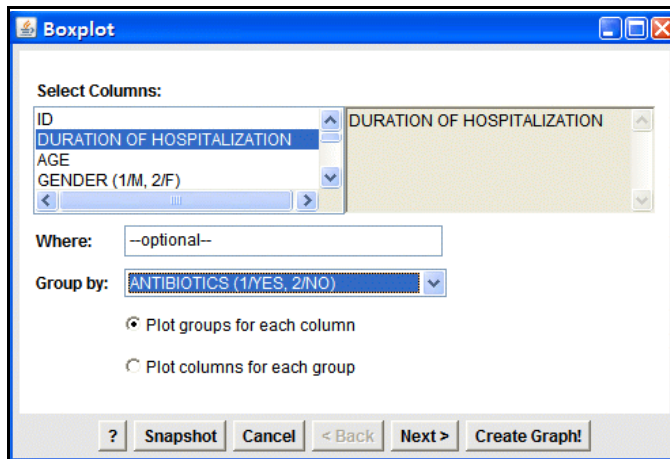
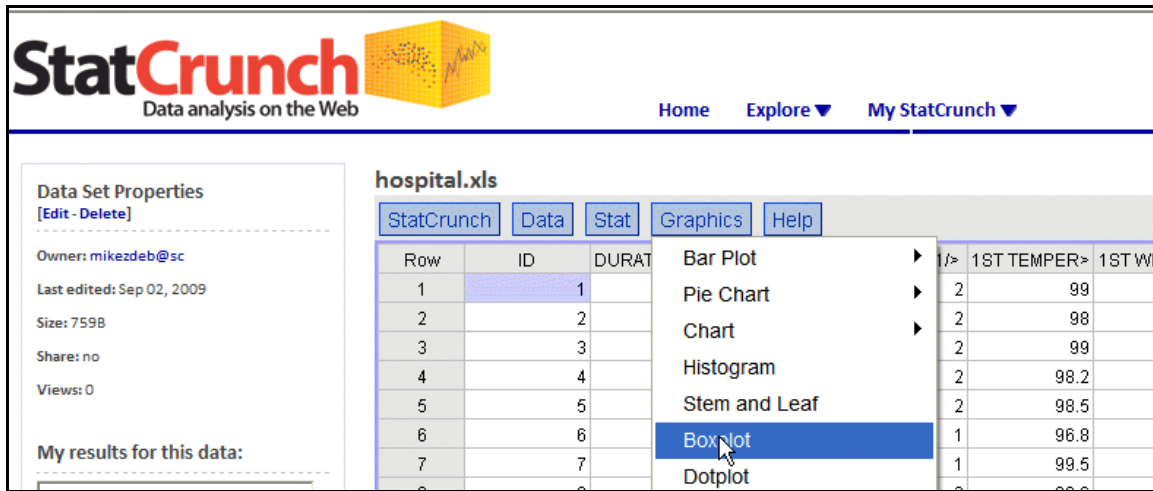


#10 YOU SHOULD SEE SUMMARY STATISTICS

Column	n	Mean	Variance	Std. Dev.	Std. Err.	Median	Range	Min	Max	Q1	Q3
DURATION OF HOSPITALIZATION	25	8.6	32.666668	5.715476	1.1430953	8	27	3	30	5	11

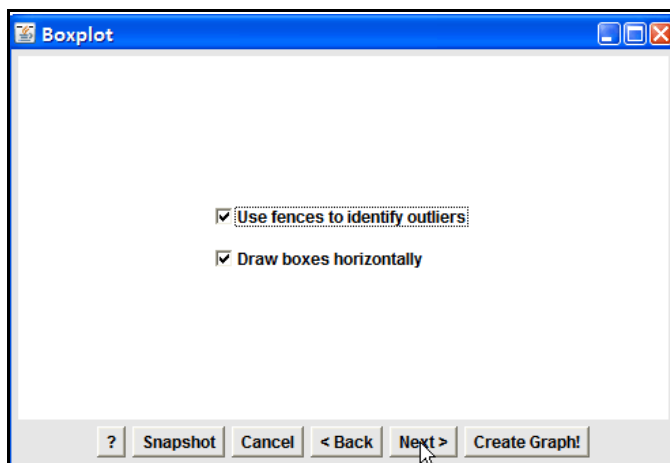
#11 TO CREATE GRAPHICS, CLICK ON THE GRAPHICS TAB AND YOU WILL SEE THAT YOU CAN CREATE STEM-AND-LEAF, BOX PLOTS, ETC. (AS YOU DID IN THE INTRODUCTORY EXERCISE) ... WHAT METHOD WOULD YOU CHOOSE ? LET'S TRY A BOX PLOT

#12 CLICK ON GRAPHICS / BOX PLOT



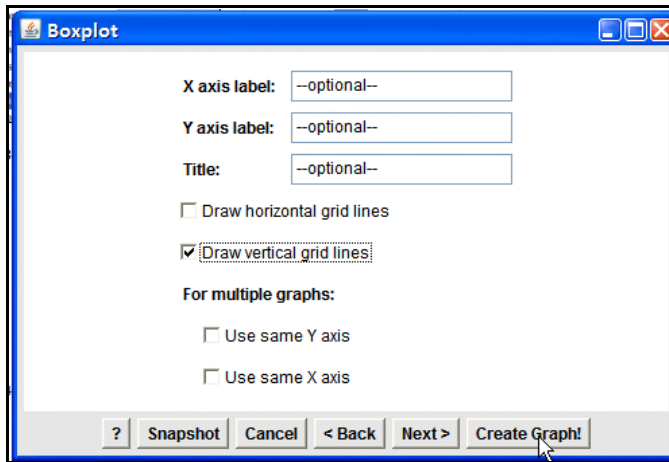
#13 SELECT DURATION ... AS THE COLUMN VARIABLE AND SELECT ANTIBIOTICS ... AS A GROUP BY VARIABLE

CLICK ON NEXT

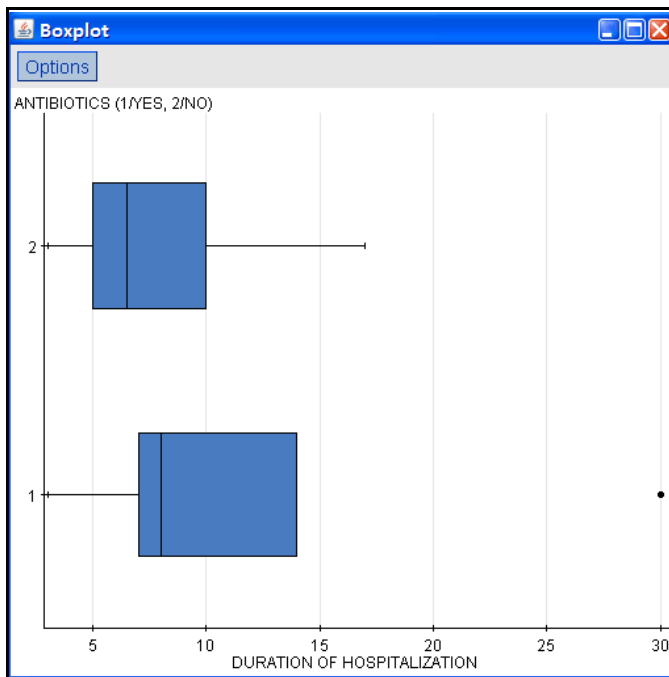


#14 CHECK BOTH BOXES (USE FENCES, DRAW HORIZONTALLY)

CLICK ON NEXT



#15 CHECK THE BOX FOR DRAWING VERTICAL GRID LINES  
CLICK ON CREATE GRAPH



QUESTION ...

WHAT DOES THIS BOX PLOT SAY TO YOU?

ROSNER SUGGESTED EITHER GRAPHIC OR NUMERIC METHOD TO ANSWER THE QUESTION

IF YOU CHOOSE ANTIBIOTICS ... AS A GROUP BY VARIABLE WHEN COMPUTING SUMMARY STATISTICS, YOU GET THE ANALYSIS SHOWN IN THE TABLE BELOW THE BOX PLOT

DOES THE BOX PLOT SHOW YOU ANYTHING ABOUT YOUR DATA THAT MIGHT NOT HAVE BEEN AS OBVIOUS IN THE SUMMARY TABLE?

Summary statistics for DURATION OF HOSPITALIZATION:  
Group by: ANTIBIOTICS (1/YES, 2/NO)

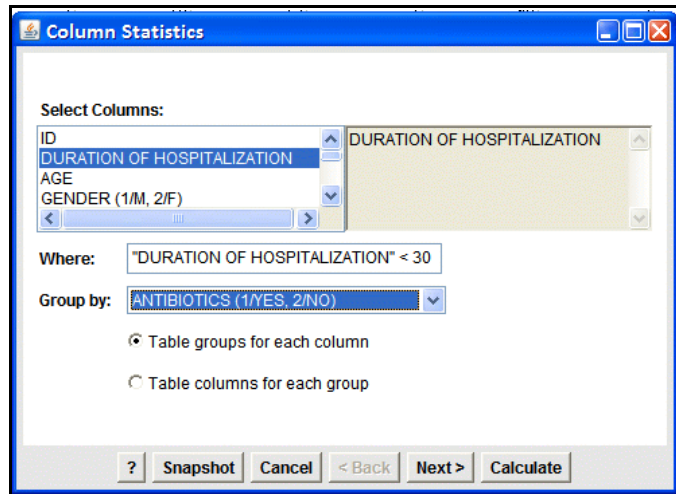
ANTIBIOTICS (1/YES, 2/NO)	n	Mean	Variance	Std. Dev.	Std. Err.	Median	Range	Min	Max	Q1	Q3
1	7	11.571428	77.61905	8.810167	3.3299303	8	27	3	30	7	14
2	18	7.4444447	13.6732025	3.6977293	0.8715632	6.5	14	3	17	5	10

**NOTE: THESE ARE ALL "SUGGESTIONS" AND YOU NEED NOT LIMIT YOUR APPROACH TO ANSWERING PROBLEMS 2.1 THROUGH 2.3 TO WHAT YOU SEE IN THESE NOTES**

#16 THIS IS ONE MORE USE OF STATCRUNCH THAT MIGHT NOT BE OBVIOUS ... IN THE BOX PLOT IT LOOKS AS IF THERE MIGHT BE AN 'OUTLIER' ... THE PERSON WITH A 30 DAY HOSPITAL STAY

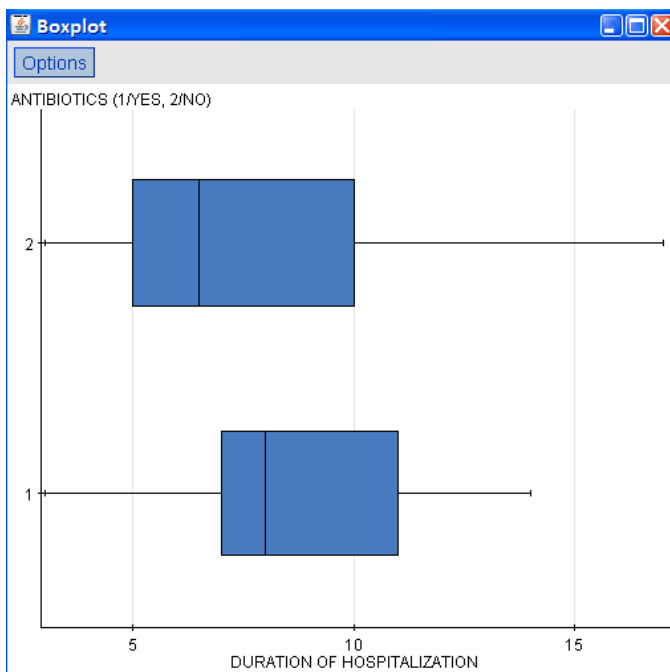
WHAT DO THE SUMMARY STATISTICS AND BOX PLOT LOOK LIKE WITHOUT THAT OUTLIER?

YOU CAN USE A WHERE STATEMENT TO EXCLUDE THE OUTLIER



NOTE: IF THE COLUMN NAME HAS SPACES IN IT, YOU MUST ENCLOSE THE COLUMN NAME IN DOUBLE QUOTES IN THE WHERE BOX

ANTIBIOTICS (1/YES, 2/NO)	n	Mean	Variance	Std. Dev.	Std. Err.	Median	Range	Min	Max	Q1	Q3
1	6	8.5	13.9	3.7282703	1.52206	8	11	3	14	7	11
2	18	7.4444447	13.6732025	3.6977293	0.8715632	6.5	14	3	17	5	10



NOTICE THAT WITHOUT THAT ONE OBSERVATION, THE DIFFERENCE IN MEANS IS NOT AS GREAT

ALSO, THERE IS A LOT MORE OVERLAP IN THE DISTRIBUTION OF THE DATA AS SHOWN IN THE BOX PLOT

**NOTE: ONCE AGAIN ... THESE ARE ALL "SUGGESTIONS" AND YOU NEED NOT LIMIT YOUR APPROACH TO ANSWERING THESE TO WHAT YOU SEE IN THESE NOTES (AND YOU MIGHT NOT HAVE FIGURED OUT HOW TO USE A WHERE STATEMENT ON YOUR OWN ... THOUGH IT MIGHT HAVE OCCURRED TO YOU THAT IT WOULD BE WORTH A LOOK WITHOUT THE OUTLIER)**

**GENERAL COMMENT ON GRAPHICS**

stem-and-leaf plots are really from an era when easy access to the personal computer did not exist

when John Tukey wrote his book *Exploratory Data Analysis* and introduced stem-and-leaf plots and box plots, it was the mid-1970s

if you had data, you had a pencil and paper or a mainframe computer to use and a stem-and-leaf plot was a good way to easily make a graph without trying to be 'too fancy'

notice that Rosner (page 28) says that one reason to use a stem-and-leaf plot is that bar graphs are difficult to construct ... that was true in another era, but not now

a stem-and-leaf plot is just a number-based histogram

yes, his reason #2 about seeing the actual values of the sample point still holds, but you can also see them if you just sort and print the data

we once used a very good statistical package called JMP (much more sophisticated than StatCrunch) for this course and it referred to stem-and-leaf plots as (I am paraphrasing) and anachronism (the old way to get a look at your data, supplanted by histograms and bar charts)

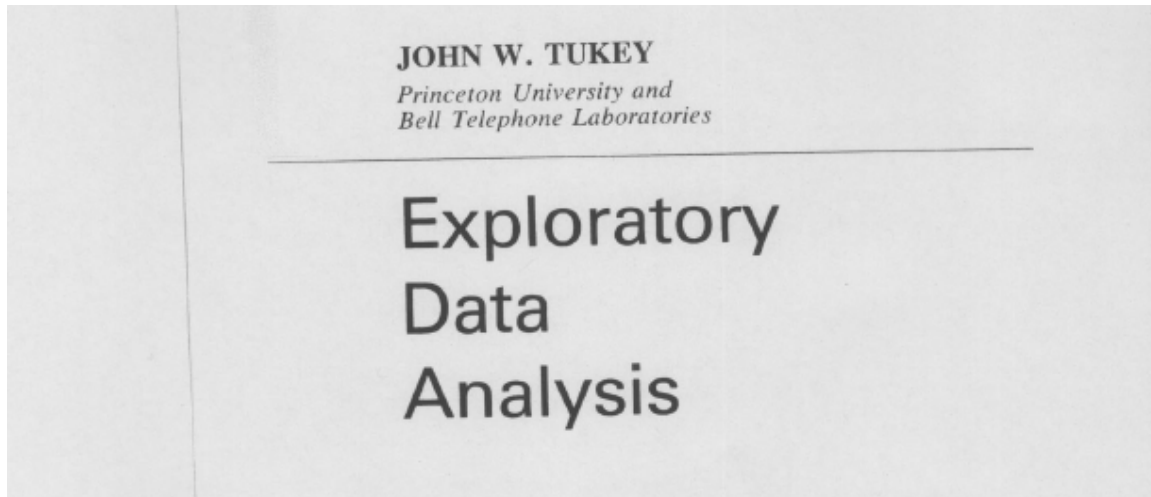
by the way, you will continue to find references in Rosner to using one method over another due to one method being hard to compute ... I call those comments 'anachronisms' since given access to computers and software, nothing in the topics we will cover in Rosner is hard to compute (even with an inexpensive, web-based package like StatCrunch, you can do all the difficult things cited by Rosner)

## GENERAL COMMENT ON OUTLIERS

in the book *Exploratory Data Analysis* came up with the rules for box plots

values outside the "fences" have come to be called outliers (Tukey called them "stray values" not "outliers") ... what's the basis of the rules for identifying "stray values" ... look below and read

the basis is that it is a "...useful rule of thumb..." (not very "mathematical")



#### 2D. Fences, and outside values

Hinges are for our convenience. They can--and will--serve various purposes for us. Their role in 5-number summaries is only the beginning.

When we look at some batches of values, we see certain values as apparently straying out far beyond the others. In other batches straying is not so obvious, but our suspicions are alerted. It is convenient to have a rule of

#### 44 exhibit 7(A)/2: Easy summaries

thumb that picks out certain values as "outside" or "far out". To do this, we set up appropriate "fences" and use "outside" and "far out" accordingly.

A useful rule of thumb runs as follows:

- ◊ "H-spread" = difference between values of hinges.
- ◊ "step" = 1.5 times H-spread.
- ◊ "inner fences" are 1 step outside hinges.
- ◊ "outer fences" are 2 steps outside hinges (and thus 1 step outside of inner fences).
- ◊ the value at each end closest to, but still inside, the inner fence is "adjacent".
- ◊ values between an inner fence and its neighboring outer fence are "outside".
- ◊ values beyond outer fences are "far out".