Box Plots

In more traditional statistics, a data set with no clear outliers can be summarized with the mean and standard deviation. A recent trend has been toward using five numbers to summarize the data set — the minimum value, the first or lower quartile, the median, the third or upper quartile, and the maximum value. This 5-number summary is a part of an "Exploratory Data Analysis" trend in statistics. As with the traditional approach, these five statistics together give a good picture of the center and spread or variability of the set of data. The 5-number summary includes:

	TI-83 Symbol	Other Common Symbols
Minimum value	minX	MIN
First quartile	Q_1	Q_L (for lower quartile)
Median	Med	Q_M
Third quartile	Q_3	Q_U (for upper quartile)
Maximum value	maxX	MAX

By way of review, a common way to compute the **median** is to sort the data set (in ascending or descending order), and then use the formula $\frac{n+1}{2}$ or $\frac{1}{2}(n+1)$ to calculate the position of the middle number. If the result of this calculation is whole (which would be true whenever n is odd), the data value in that position (from either end) is the median. If the result is midway between two whole numbers (true whenever n is even), use the data values in those two positions and average them to find the median.

A similar way is to use the formula $c = \frac{n}{2}$ to calculate the position of the middle number. Here, if the result is whole, use the value halfway between the data values in the c and c + 1 positions. If the result is not whole, round c to the nearest whole number and find the value in that position.

These 2 methods always produce identical results.

Quartiles divide the data set into quarters, or 4 equal parts. The median is the middle or second quartile. Roughly half the data is less than the median, and roughly half the data is greater than the median. As we have seen, the median may or may not be an actual data value.

Roughly one-quarter (25%) of the data is less than the first (or lower) quartile, with the rest of the data (three-quarters or 75%) greater than Q1. Roughly three-quarters (75%) of the data is less than the third (or upper) quartile, with the rest of the data (one-quarter or 25%) greater than Q3.

To calculate the lower quartile, there are several commonly used techniques.
We're outlining the method that consistently gives the same results as the TI calculator.

First, arrange or sort the data in ascending order. For the lower quartile, use the median of the observations to the left of the median, and for the upper quartile, use the median of the observations to the right of the median.

Example:

(1) Draw a complete box plot for this scenario: A policeman records the speeds of vehicles for 30 minutes of rural highway traffic.

44	54	50	46	45	49	55	44	42	55	51	52
54	44	60	44	59	41	44	47	51	42	43	

Solution:

First, sort or arrange the data values in order from least to greatest:

41	42	42	43	44	44	44	44	44	45	45	47
49	50	51	51	52	54	54	55	55	59	60	

Since n = 23, there's one middle number, and it's in the 12^{th} position.
Once we know the median (Med = 47), the data set is divided into 2 halves.

41	42	42	43	44	44	44	44	44	45	45	**47**
49	50	51	51	52	54	54	55	55	59	60	

For the lower quartile (Q_1), find the median of the 11 numbers to the left of 47:

41	42	42	43	44	44	44	44	44	45	45

The first quartile would then be the 6^{th} value, which is 44.

For the upper quartile (Q_3), figure out the median of the 11 numbers to the right of 47:

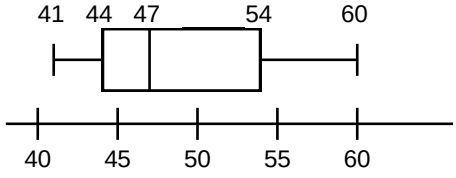
49	50	51	51	52	54	54	55	55	59	60

The third quartile would then be the 6^{th} value past the median, which is 54.

The minimum value, first quartile, median, third quartile, and maximum value are now shown on the original data set.

41	42	42	43	44	44	44	44	44	45	45	47
49	50	51	51	52	54	54	55	55	59	60	

And here's the box plot:



NOTE: A number line is highly recommended for scaling purposes so that the box plot lengths are proportional.

(2) Draw a complete box plot for the following test scores for 28 Math 1101 students on Exam II (already sorted):

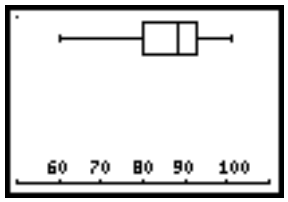
60	62	63	71	75	79	79	81	83	84	85	86	86	88
89	90	91	92	93	93	93	93	94	95	97	98	101	101

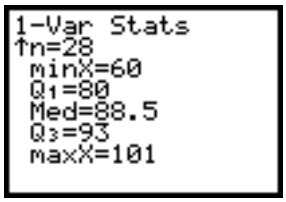
Since $\frac{28+1}{2} = 14.5$, the median is midway between the 14th and 15th values, or $\frac{88+89}{2} = 88.5$. The first quartile is in the $\frac{14+1}{2} = 7.5$ position, so we average 79 (the 7th value) and 81 (the 8th value) to find $Q_1 = 80$; and similarly, we average 93 and 93 to find $Q_3 = 93$; the minimum and maximum values are clearly 60 and 101.

We now show the 5-number summary on the original data set, and the given box plot image is from the TI display, with some text drawn on the horizontal axis. All of the 5 number summary statistics (and then some!) appear when the calculator carries out the 1-variable statistical calculation command.

60 62 63 71 75 79 79 81 83 84 85 86 86 88 89 90 91 92 93 93 93 93 94 95 97 98 101 101

80 88.5 93





Our two examples cover the odd and even n cases, and the procedure we've discussed works for every possible scenario, matching with what the TI calculator produces.

Exercises:

For #1-7, make a box plot for each of the given data sets. Use a number line with a consistent and appropriate scale, and carefully label each of the 5 statistics on your plot.

1. Employee salaries at A & B Co.

$24,000	$17,500	$21,000	$32,000	$20,000
$27,900	$30,850	$18,400	$26,500	$145,250

2. Rainfall totals (in inches) by month for Lomalinda, Colombia (South America)

Jan	Feb	Mar	Apr	May	June	July	Aug	Sept	Oct	Nov	Dec
0.3	1.5	12.9	26.6	11.5	12.6	21.7	18.1	11.4	11.9	8.8	3.2

3. Ages of Academy Award-winning actresses from 1942 to 2003

34	24	29	41	30	34	34	33	28	38
45	24	26	48	41	27	40	38	28	27
31	37	30	24	34	60	**61**	**26**	35	34
34	26	37	42	41	35	31	41	33	30
74	33	49	38	61	21	41	26	80	42
29	33	35	45	49	39	34	26	26	33

Double award given in 1968

4. Sales figures for ladies athletic shoes for the month

$6, 6\frac{1}{2}, 5, 7, 7\frac{1}{2}, 8\frac{1}{2}, 8, 5\frac{1}{2}, 6, 6, 7\frac{1}{2}, 8, 7, 5\frac{1}{2}, 6, 5\frac{1}{2}, 5, 7\frac{1}{2}, 7, 9, 8\frac{1}{2}, 7$

5. The number of consecutive hours a light bulb will last before it burns out

402	405	409	389	456	423	432	440	425	436
421	423	408	411	434	448	392	400	439	430

6. Cattle weights

748 lb	485 lb	807 lb	1,023 lb	761 lb
765 lb	934 lb	579 lb	865 lb	1,064 lb

7. Hank Aaron's and Babe Ruth's home run totals by season

Season	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Aaron	13	27	26	44	30	39	40	34	45	44	24	32
Ruth	0	4	3	2	11	29	54	59	35	41	46	25
Season	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	
Aaron	44	39	29	44	38	47	34	40	20	12	10	
Ruth	47	60	54	46	49	46	41	34	22	6		

Hank Aaron played for the Milwaukee Braves, Atlanta Braves, and Milwaukee Brewers from 1954-1976, and Babe Ruth played for the Boston Red Sox and the New York Yankees from 1914-1935.