Visual Evidence: Box & Whisker Plots from 5-Number Summaries

Draft 2010-06-30
rob rucker

Numbers as Visual Evidence: Box & Whisker Plots from Five Number Summaries

The purpose of this tutorial is to show you how to analyze and then display a batch of numbers, with clarity and insight. That is, this tutorial will show how to turn a batch of numbers into *graphical evidence*. This capability allows you to add value in your role as business or technical analyst.

To tell or show something about a batch of numbers, we can present the whole batch, or just a few well chosen numbers. Depending on how bunched up or regular the data set is, these few numbers might tell us everything we want to know but, if there are *outliers*, we would also like a way to explicitly show them to supplement the few summary values. For example, the arithmetic *average* of the batch {1, 2, 3, 4, 1000} is 202, while the *median* (middle number) is 3. This simply shows that what you want to emphasize is crucial. In this case, it is easy to see that '1000' would be considered an *outlier* relative to the bulk of the other numbers and should be explicitly noted. For larger batches of numbers though, it's harder to pick out these unusual numbers, these outliers. We need some tools to extract out and present those key features of a batch, preferably with a visual emphasis. That's the intent of the two tools in this tutorial: five number summaries to extract out the key characteristics of a batch, and box & whisker plots to display them.

Below is a generic box & whisker plot, an example of visual evidence. The rectangle's length represents the bulk of the data set (50%) while the *adjacent* points delimit the data values exclusive of outliers. Outliers are specially marked, in this case, with * and •. These unusual values, relative to the bulk of the data set, probably merit additional investigation, but might be overlooked without this explicit visual presentation. Emphasizing these unusual values would be a contribution.

This (general) diagram graphically summarizes a batch of numbers, large or small, using 5-number summary values, plus a few derived numbers and a rectangle. This is essentially John Tukey's *Box and Whisker Plot.*

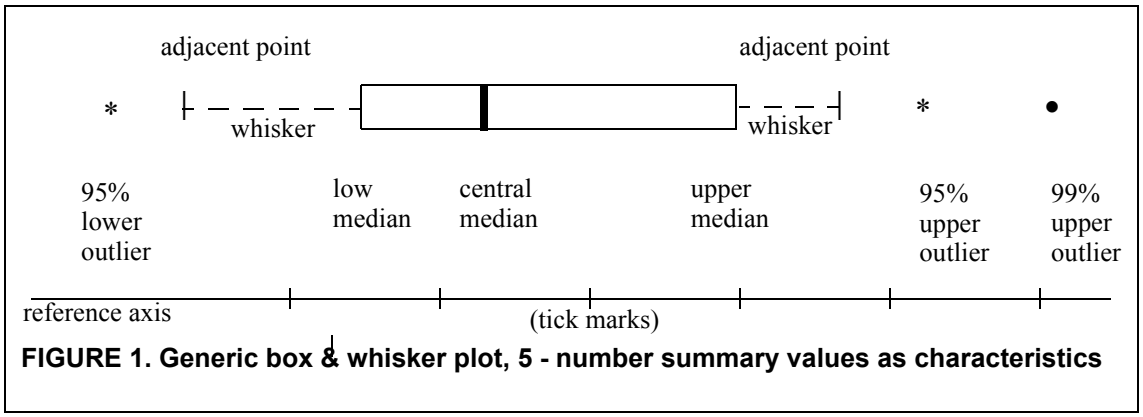

FIGURE 1. Generic box & whisker plot, 5 - number summary values as characteristics

The techniques developed in this tutorial are ways to appreciate what the numbers in a batch may be trying to tell us and, ways to make those insights apparent to our clients. *It's a good idea to keep in mind that your clients are busy people who want to cut to the chase as soon as possible, with an insightful, no-nonsense summary of a given batch of numbers.*

This tutorial is inspired by, and makes considerable use of, John Tukey's work *Exploratory Data Analysis* (1977) and I highly recommend studying his material. The current master of evidence presentation is Edward Tufte (see references). You might also want to look at the tutorial *EDA Introduction* by Rucker on the *milagrosoft.com* web site for an even more basic introduction.

Here are the major topics in this tutorial:

- Medians of batches of numbers
- 5-number summaries of batches
- Batch control limits for determining outliers
- Box and whisker plots
- Comparison of batches using Box and Whisker plots
- Time series graphs of batches using Box and Whisker plots
- Trend lines superposed on times series Box and Whisker plot

Cut to the Chase

Note: This is the start of a series of *how-to* sections. If you are a consumer of the type of graphical evidence described above, then you may want to stop here. If you want more insight into what you are seeing, or might want to be a producer of such evidence, then read on!

Very high level description of the analysis process

Let me summarize and anticipate the construction of the 5-number summary and box plot tools that are coming next (you aren't expected to understand all of this now, just flow along for the moment). All of the terms, steps, and options discussed below will be explained within this tutorial.

First: Starting with a batch of numbers, you will get their 5-number summary. (Those 5 summary numbers consist of the two extreme values plus three interior medians, to be explained later). That is, a 5 number summary will simply be 5 numbers. The names of them are as follows:

{low extreme, low median, central median, upper median, upper extreme}.

Next: Check for and mark unusual values in your batch called *outliers.*The test for outliers involves some arithmetic combinations of the summary values and an appeal to an analogy with the standard Normal Distribution.

Next: Using the summary values, and derived numbers related to outliers, construct a box and whisker plot

Optionally: If you have more than one data set, you can compare and contrast their representative box and whisker plots. (As an aside, the essence of evidence presentation is a compare and contrast phase - box plots support this very well).

Optionally: If you have data sets (batches) that you track over time, you could place corresponding box and whisker plots on a coordinate system with a time axis and so come up with a time series consisting of box plots! (This kind of time series will tell a lot more than simply plotting single points over a time axis since, now you have some idea of the *distribution* of the data sets at each time interval).

Optionally: If you do a time series boxplot display, you could then overlay a *trend line.* To calculate this line you could summarize each batch by using their central medians and now you have a standard least squares problem. The result is a trend line, useful for short range predictions.

And yes, we can do all this using just a few math ideas involving counting, some basic arithmetic, and rectangles with whiskers.

Below is the big picture for this tutorial. Just as an example, assume these data sets (batches) are the percentage rates of return for three stock portfolios: portA, portB, and portC. The bulk of the data is represented by the rectangles while *outliers* are indicated by symbols such as * for unusual data values and • for very unusual numbers.

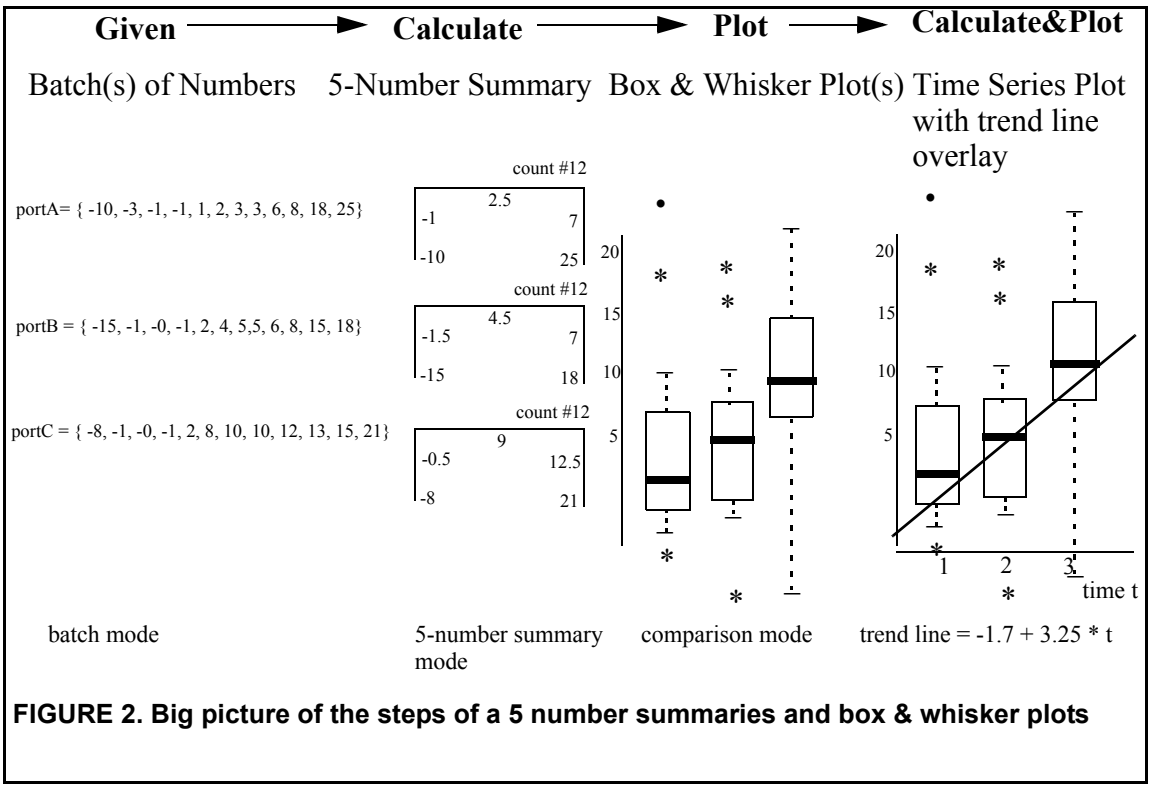

FIGURE 2. Big picture of the steps of a 5 number summaries and box & whisker plots

Learning the steps by using Portfolio A (portA)

Ok, let's start: you are confronted with a batch of numbers, such as the data set denoted by 'portA' above. What do you do? Here are the steps to follow:

1. First sort the numbers in order. This is already done.

portA= {-10, -3, -1, -1, 1, 2, 3, 3, 6, 8, 18, 25}

Find the three interior medians: central median, low median, and upper median

**Note: the batches in these first examples all have an even number of elements. Of course

there will be batches with an odd number. To calculate the medians for those, refer to section "Finding the 5-Number Summary for a batch with an odd count" on page 15 in this tutorial.

Ok, the batch is already sorted, so now I need to find three medians. What is the *central median?* That is the middle number of the whole batch. How do I find it? Count into the batch, from either end, until you reach the middle number or a location between two numbers of the batch. In this case, counting in from either end (that is, first, second, third, etc....) gets me to the location of 6 ½th. That is, the *index* of the middle number would be at 6 ½. Since there is no actual number at this index, that is, there is no number in the batch between values 2 and 3, we take the average of these two as the central median = (2 + 3)/2 = 2.5. This is the *central median* value and it is at index 6 ½.

> Hint: an easy way to find any of these median indexes is to take half of the (count of whatever batch you are working with, plus 1). That is, index = (count + 1)/2. So for a count of 12, I would get a median index of: (12 +1)/2 = 6 ½. For a batch with 6 elements the median index would be (6+1)/2 = 3 ½. For a batch with 9 elements, its median index would be (9+1)/2 = 5, (and there would be a batch number at that index).

Now, the *central median,* being at the 50% mark, partitions the batch into two parts, a lower batch of numbers {-10, -3, -1, -1, 1, 2}, and an upper batch of numbers {3, 3, 6, 9, 18, 25}. Next find the median of the low batch, that is, take the first 6 numbers of the batch and find *their* median. That value lies at an index of 3 ½. Again, there is no number of the batch at that location, so we take the average of the numbers lying on either side, namely -1 and -1 or (-1 + -1)/2 = -1. I call this the *low median* which is at index 3 ½.

Now take the upper 6 numbers and find *their* median. This number lies at an index of 3 ½ (counting in from the right hand side of the batch). I call this the *upper median* and it too, is at index = 3 ½.

Find the final two numbers of the 5-number summary

This is too easy, since the last two numbers are taken to be the extremes, that is, the numbers at the ends of the batch, at index=1. These will be -10 and 25. They are called the *extremes*. Ok, here is the final 5-number summary.

The 5-number summary for portfolio A (portA)

5-Number Summary for portA

median name	**index**	count #12		
central median	6 ½	2.5		
lower/upper medians	3 ½	-1	7	Median Spread (spread) = 7 - (-1) = 8
extremes	1	-10	25	

Finding the outliers

Note: It will be easier to follow this discussion by reference to the figure below, "Control limit analogies for the box & whisker plot of portA" on page 7.

To detect *outliers*, a set of "control limits" values can be set up. Batch numbers beyond these limits deserve additional scrutiny since they may represent important anomalies/opportunities. There are a set of 4 such control limit numbers, two above the upper median and two below the lower median, as in the diagram below. The question is, how to set these limits so as to detect 'out of control' values but not flag everything? That's where a *comparison/reference* distribution comes in, and that's where the *Normal* distribution (often called the *Bell Curve*) comes in. On the right of the diagram is a plot of the standard Normal distribution with distinctive critical values marked. My plan is to argue, by analogy, that the 95%, and 99% control limits calculated for the Normal distribution, can equate roughly to equivalent control limits for a batch of data and so be used to distinguish unusual or *outlier* data values.

The analogy to the Normal Curve distribution

(More theory is at "Batch Calculation Theory (optional material)" on page 13). From your standard Normal distribution statistics tables, it is rare to find values beyond 2 standard deviations (since ±2 standard deviations contain 95% of the data and so correspond to 95% control limits). *So, finding the rough equivalent to 2 standard deviations away from the central median for a batch of numbers, would be a way to set up 95% control limits for that batch.*

Since only 1% of the Normal distribution data lies beyond 3 standard deviations, then ± 3 standard deviations provide 99% control limits. *So, finding the rough equivalent to 3 standard deviations away from the central median for a batch of numbers, would be a way to set up a 99% control limits for that batch.*

How to calculate the 95% and 99% control limits for a batch

To find the outliers I need to calculate some numbers I will call *control limits*. Values in my batch that are above the upper control limits or below the lower control limits, will be characterized as *outliers*, and marked as such.

99% upper control limit = upper median + 2 * median spread = 7 +2* 8 = 15

95% upper control limit = upper median + median spread = 7 + 8 = 23

95% lower control limit = lower median - median spread = -1 - 8 = -9

99% lower control limit = lower median - 2 * median spread = -1 - 2* 8 = -17

The two upper control limits (95% and 99%) are located above the upper median at plus one median spread value and plus two median spread values. That is, the 95% control limit value is 7+8=15 and the 99% control limit value is at 7+8+8 = 23. Note that these numbers are not part of the batch, but are simply guides as to which, if any, batch values are anomalies (outliers). Similarly, the lower 95% control limit is the lower median minus one median spread, -1 -8 =-9, and the 99% lower control limit is the value -1 -8 -8 = -17.

How to draw the box plots

Now, with the 5-Number Summaries and the control limits, here is the final step, actually drawing the box plots. (Check out the previous diagrams, especially the first generic box

plot for suggested style).

Step 0. Draw an axis with just a few tick marks. The idea is not to distract from your box plot display by cluttering up your diagram. (No ‘chartjunk’ allowed!). Maybe 4 or 5 tick marks would be sufficient.

Step 1. Draw a long thin rectangle that stretches from the Lower Median to the Upper Median consonant with your tick marks.

Step 2. Draw a cross bar at the Central Median value

Step 3. Draw a ‘whisker’ from each end of the box to its adjacent value.

Note: an adjacent value is the last number in the batch that is not an outlier.

Step 4. Note the control values. They are not actually plotted but are guides to the plotting of batch values that do fall within their regions.

Step 5. Mark and label the outliers that are between the 95% and 99% control limits in one manner and those values that exceed the 99% control limits differently. See the examples for symbols you might use. You can use any symbols you like, just be consistent.

That’s it!

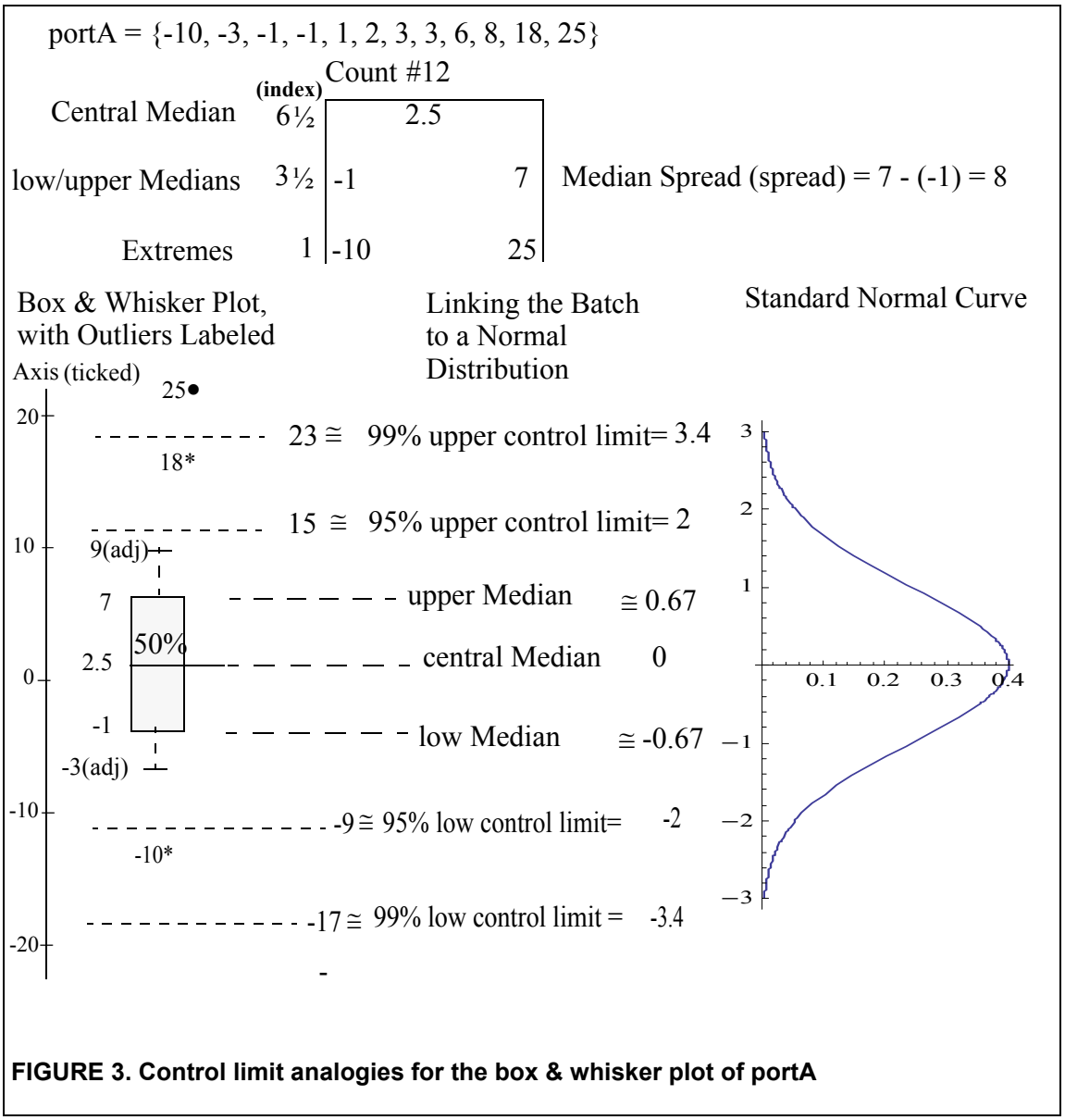

FIGURE 3. Control limit analogies for the box & whisker plot of portA

Connection with textbook names for these characteristics

The difference between the two extremes can be called the *range* of the data set= 25 - (-10) = 35

The difference of the upper median and the lower median is often called the *Inter-quartile Range* (IQR), 7 - (-1) =8 since these two medians divide the batch into a lower quarter of the data and an upper quarter of the data.

This median spread is also called the *Hinge Spread* in Tukey's original presentation. My lower median is Tukey's lower Hinge, while my upper median is called the upper Hinge by Tukey.

A 'Tukey' *step* is 1.5 * Hinge Spread, although I have simply used the median spread as my step.

Examples:

Using the ideas of the above section puts us in a good position to analyze some interesting data sets and perhaps discover some surprises. Let me start with an analysis of the populations of the worlds largest cities.

City Populations - the World's Largest Urban Areas

This data is from *2005 NY Times Almanac* p.480, sorted by millions, with Tokio at 36.2 and Seoul at 9.2 million. First look at the sorted population numbers below (these are the older names for some cities such as Tokyo and Bombay (mumbai)):

9.2	seoul
9.4	chicago
10	paris
10.9	moscow
11.1	beijing
11.3	istanbul
11.3	tianjin
11.4	osaka
12.4	rio
12.6	manila
12.7	shanghai
12.9	la
13.1	cairo
14.6	buenosaires
16.2	karachi
16.8	calcutta
17.	lagos
17.5	jakarta
17.9	dhaka
19.7	ny
20	saopaolo
20.6	mexcity
20.9	delhi
22.6	mumbai
36.2	tokio

The 5-number summary and associated box & whisker plot

Notice that Tokio is unusual even among these large population centers. Why this is so might merit a follow-on investigation? Using the city populations, here are the detailed steps leading to 5-number summary, identification of potential outliers, and an associated box and whisker plot. Note that this is an odd-numbered batch, so the calculation of medians follows the discussion in "Finding the 5-Number Summary for a batch with an odd count" on page 15. What this means is that to find the lower median, average the two medians obtained by including and then excluding the central median value. Same plan for the upper median.

count = 25

central median index = (25+1)/2 = 13

central median = 13.1 (the index of '13' picks out Cairo as the median population value)

low median = (Median[{9.2, . . . , 12.9, 13.1}] + Median[{9.2, . . ., 12.9}])/2 = (11.3 + 11.3)/

2 = 11.3

low median = 11.3

upper median = 17.9 (using the averaging of medians as in the lower median calculation)

median spread= 17.9 - 11.3 = 6.6

upper 99% limit =17.9 + 2* 6.6 = 31.1

(upper) 95% limit = 17.9 + 1* 6.6 = 24.5

(lower) 95% limit = 11.3 - 1* 6.6 = 4.7

lower 99% limit = 17.9 - 2* 6.6 = -1.9

Using these values we identify:

lower adjacent value = 9.2 (simply the lowest value we have, corresponding to Seoul)

This means we have no lower outliers. That is, no batch value is less than 4.7.

upper adjacent value = 22.6 (Mumbai (Bombay))

outlier value = 36.2 (Tokio, which is the largest value we have and the only outlier since it exceeds the 99% control limit, a very unusual value)

count # 25

	(index)			
Central Median	13		13.1	
low/upper Medians	7	11.3		17.9
Extremes	1	9.2		36.2

median spread = 17.9 - 11.3 = 6.6

FIGURE 4. 5-Number Summary of the City Populations batch

The next figure shows a computer generated box and whisker plot with an outlier data point indicated. No surprise here, it is Tokio.

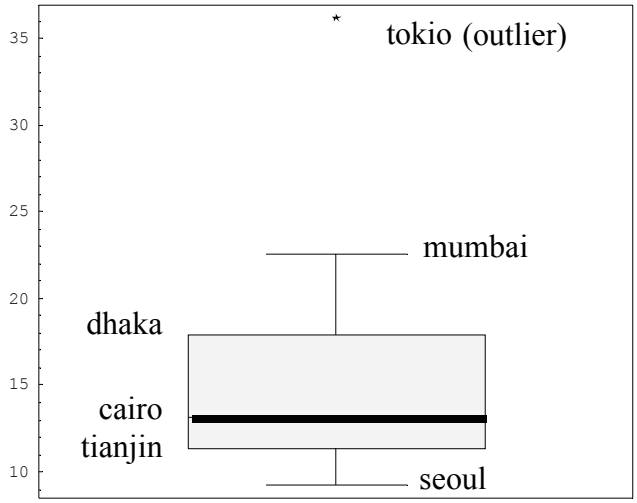

FIGURE 5. Box and Whisker Plot of World's Largest City Populations

The box and whisker plot for the populations shows 'whiskers' extending to values that are adjacent. There is one outlier value, shown in a distinguished way as a '*'. This outlier is an "outside value" that lies *beyond* the upper 95% limit and the upper 99% limit. Generally such values require some explaining. In our case the outlier is the population of Tokio which turns out to be very unusual, *when compared against the bulk of city populations*, that is, the middle 50%.

Example: Temperature comparisons using box plots

Box plots are particularly good at exploratory comparison studies. Below, I have copied off the mean temperatures over a 30 year period for three cities of interest. The vectors below show these temperatures, starting in January and going through December. For example, the average Albany February temperature over 30 years (1971 - 2000), is 25 degrees Fahrenheit. For Phoenix, it is 58 degree Fahrenheit. Looking at the box plots, the median temperature for Albany, over the 30 year period, is approximately 47° while that of Phoenix is 73°, very close to Tampa Florida's 74°. (Note that I am taking medians of a set of averages)!

Three City Average Temperatures

albany = {22,25,35,47,58,66,71,89,61,49,39,28}, five numbers = {22,31.5,48.,63.5,89}

phoenix = {54,58,63,70,79,89,93,91,86,75,62,54}, five numbers = {54,60.,72.5,87.5,93}

tampa = {61,63,67,72,78,82,83,83,82,76,69,63}, five numbers = {61,65.,74.,82.,83}

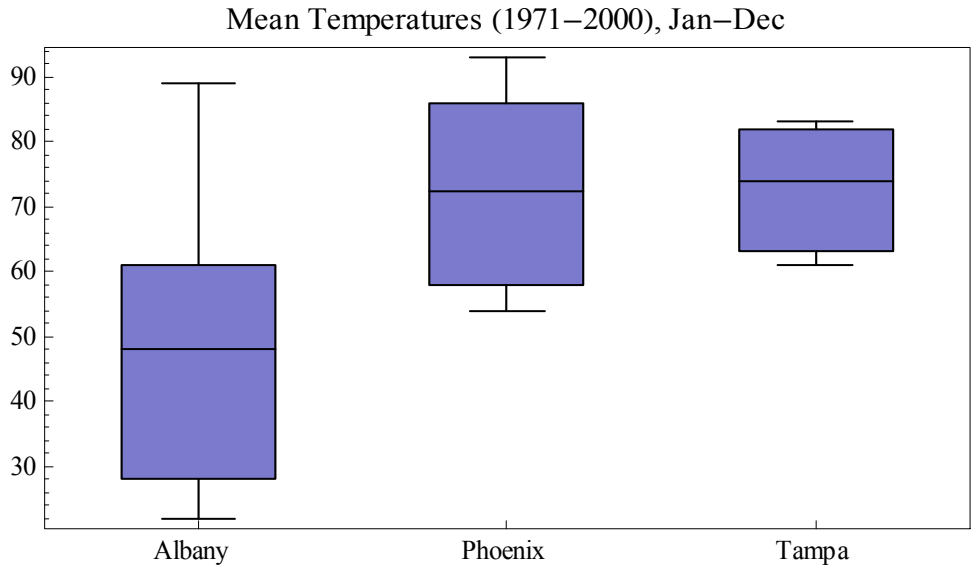

FIGURE 6. Temperature Comparison Box and Whisker Plots

Example: Using Box Plots to Show Time Series Data

Box plots can be used to show how time-based batches of data track. As a hypothetical example, suppose we are managing, say, 9 comparable retail stores in Phoenix and, we have monthly revenue data from all 9 each month, over a 4 month period. How might we display this 'time series' of data values? Of the many ways you might come up with, let me show how a series of box plots could be used to show this time progression.

Here are the 4 data sets corresponding to the 9 stores over a 4 month period. I have used a home-grown computer program to calculate 5-number summaries and then used the *Mathematica* math package to plot these data sets. If you are interested, I have an easy tutorial on trend analysis on the *milagrosoft.com* site called *Business Trend Analysis*.

The 4-Month Revenue Streams (1000s)

month1 = {10, 12, 9, 11, 13, 15, 18, 10, 12}; 5-num->{9.00,10.00,12.00,13.50,18.00}

month2 = {8, 14, 11, 14, 13, 17, 20, 12, 15}; 5-num->{8.00,11.80,14.00,15.50,20.00}

month3 = {10, 14, 12, 12, 16, 19, 25, 12, 19}; 5-num->{10.00,12.00,14.00,19.00,25.00}

month4 = {6, 15, 15, 16, 13, 22, 19, 18, 26}; 5-num->{6.00,14.50,16.00,19.70,26.00}

The Time Series Display of the Revenue Data as BoxPlots

The plot below shows a time series using box plots together with distinguished outliers. It turns out that these are barely outliers, but even so they would probably bear close examination. Note that the raw data above is unsorted.

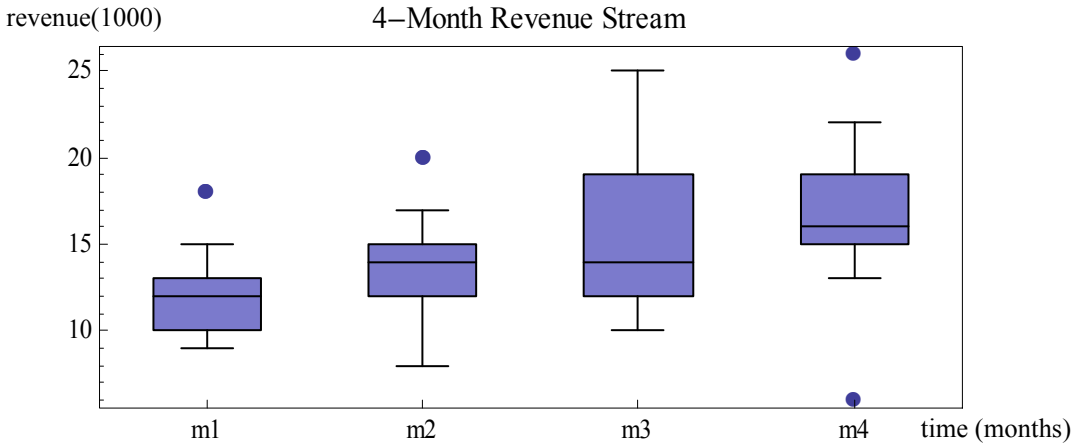

FIGURE 7. Revenue from 9 retail stores over a 4-month period

Fitting a Trend Line to the Box Plot Time Series

Now this gets more interesting. What value should I use to represent each data set? I am going to use the *central median* for this, getting a series of values {12, 14, 14, 16}. Doing a least squares linear fit to the data results in the trend line:

revenue = 11 + 1.2 * monthNumber.

So for month 5, we would get a *median* revenue = 11 + 1.2 * 5 = 17

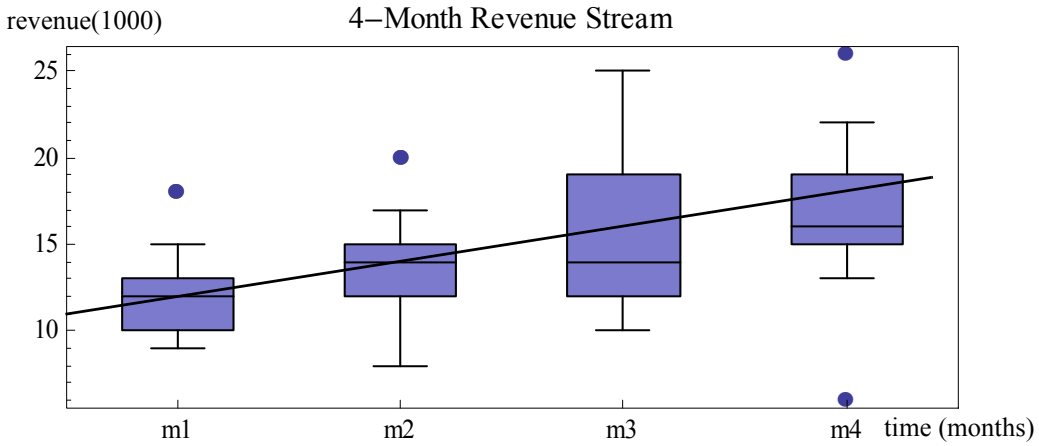

FIGURE 8. Revenue from 9 retail stores over a 4-month period, trend line superposed

Example: A batch with an odd number of elements:

revenue = {-20, 2, 4, 5, 7, 10, 12, 16, 20, 91, 93}

Count [revenue] = 11

Central Median Index = (11 + 1)/2 = 6 which is the median rank or index. This is the index at which a batch value will be found for these odd numbered batches, '10', in this case.

(Central) Median = 10 (since there was a number, 10, in the batch at that index)

Note: since this is an odd count batch we need to use the procedure explained in "Finding the 5-Number Summary for a batch with an odd count" on page 15.

low Median = (Median[{-20, 2, 4, 5, 7, 10}]+Median[{-20, 2,4,5,7}])/2 = (4.5 + 4)/2 = 4.25

upper Median= (Median[{10, 12, 16, 20, 91, 93}]+ Median[{12,16,20,91,93}])/2= 18 + 20 = 19

Median Spread = upper Median - low Median = 19 - 4.25 = 14.75

lowExtreme = -20

upperExtreme = 93

Median Spread =(upper median - low median) = 19 - 4.25 = 14.75 (this includes the 'bulk of the data', and in fact, 50% of the data, this is also called the InterQuartile Range IQR)

Lower 95 Limit = (low median - median spread) = 4.25 -14.75 = -10.5 (values ought to be greater that this almost all the time, this would be very roughly comparable to a *' -2 sigma' limit*)

Lower 99 Limit =(low median - 2* median spread)= -25.25 (numbers ought to be really greater than this almost all of the time - this is roughly comparable to a *' -3 sigma' limit*)

Upper 95 Limit = (upper median + median spread)= 19 + 14.75 = 33.75 (numbers in the batch beyond this value suggest anomalies, this is roughly comparable to a *' +2 sigma' limit*)

Upper 99 Limit = upper median + 2 *median spread = 19+ 2* 14.75= 48.5 (this is roughly comparable to a *' +3 sigma' limit*)

Calculating Adjacent Values (values bounding all of the 'good data in the batch)

These are the numbers in the batch, that represent the largest and smallest values in the batch that are not exceptional. That is, all of the 'good' data lies within these batch numbers, inclusive.

Lower Adjacent value - the value in the data set that is closest to, but still on or inside the Lower 95 Limit value. In our case this is the value just inside -17.13 and is '2'.

Upper Adjacent Value - the value in the data set that is closest to but still on or inside the Upper 95 Limit. This value is the one that is just inside of 40.38, which is '20'.

Outlier values - are those lying between the 95 and 99 Limits. Are there any values lying between [-25.25 and -10.5]? or between [33.75 and 48.5]? Yes there is -20 in the lower set of values.

Out values should be labeled individually.

Far Out values - are those lying beyond the Outer Control 99% Limits.

Are there any values lying below -25.25? No. Are there any above 48.5? Yes, 91 and 93.

Far out values are very suspect and should be distinctively labeled.

Batch Calculation Theory (optional material)

Using the Cut to the Chase data set 'portA', I will put in all the detailed steps for the various calculations. There are some useful ideas in this section but the reader probably has enough detail already, except perhaps for the section on how to treat odd number batches. See

"Finding the 5-Number Summary for a batch with an odd count" on page 15.

portA = {-10, -3, -1, -1, 1, 2, 3, 3, 6, 8, 18, 25}

Index of Central Median= (Count+1)/2 =6.5

Central Median = Number at that index (or the average of the 2 numbers on either side of that index, as in this case) = (2 + 3)/2 = 2.5

Index of Lower Median = (6+1)/2 = 3.5

Lower Median = Number at that index (or the average of the 2 numbers on either side at that index) = (-1 + -1)/2 = -1

Index of Upper Median = (6+1)/2 = 3.5

Upper Median = Number at that index (or the average of the 2 numbers on either side at that index)= (6+8)/2 = 7

Median Spread = 7 - (-1) = 8

upper 95% control limit = 7 + 8= 15(upper median + median spread)

upper 99% control limit = 7 + 2* 8 = 23 (upper median + 2* median spread)

low 95% control limit = -1 -8= -9 (low median - median spread)

low 99% control limit = -1- 2* 8 = -17 (low median - 2* median spread)

Where do the Batch Control Limit Numbers Come From?

I am going to give some plausible arguments as to why the batch data control limits can be usefully approximated by starting at the upper or low medians and adding or subtracting a multiple of the Median Spread. I'll do this by first showing how the control limits for the Normal distribution are calculated and then equating these limits to the corresponding batch control limits. This is admittedly just a very rough estimate, since we don't know what the real underlying distribution of the batch data is, all we have is just a batch of numbers, but, the Normal is probably the best approximation we can use.

The Normal Control Limits

Notice that for the Normal distribution, the middle 50% of the data is between [-.67, +.67] (I got these value from tables of the standard Normal Distribution). So,

lower median = -0.67 where 25% of the data lies below this value

upper median = 0.67 where 25% of the data lies above this value.

The *Normal Median Spread (spread)* is then this difference = 1.34.

Now, using the 'spread' as a convenient factor, I can calculate:

upper 95% control limit = upper median + Normal Spread = 0.67 + 1.34 ≅ 2 which means the amount of data between -2 and +2 is 95% and so *the upper median plus the median spread is an approximate 95% control limit.* (This is the analogy I will use for the batch limits)

upper 99% control limit = upper median + 2 * Normal Spread = .67 + 2.68 ≅ 3.4

so, the upper median plus the median spread is an approximate 99% control limit

So, I am going to use this as the 99% control limit, although it is a bit more inclusive than this. This means that ± 3.4 includes 99+% of the data values.

The Analogous Batch Control Limits

For the batch, the Median Spread also describes the middle 50% of the data set. In analogy with the Normal, the *Batch Median Spread* could be used to add or subtract from the upper and low medians. This means that:

upper 95% Batch control limit = upper Batch median + Batch Median Spread, is an approximate 95% control limit.

upper 99 Batch control limit = upper Batch median +2 * Batch Median Spread, is an approximate 99% control limit.

and the low control limits would be given by:

low 95% Batch control limit = low Batch median - Batch Median Spread, is an approximate 95% control limit.

low 99% Batch control limit = low Batch median - 2 * Batch Median Spread, is an approximate 99% control limit.

Finding the 5-Number Summary for a batch with an odd count

Now suppose the count of the number of elements in the batch is odd, that is, the batch count is odd. Below is such a vector, portX.

portX = {2, 4, 5, 7, 10, 12, 16, 20, 20}

Median = Single middle value for a batch with an odd number of elements.

The location of this number is at an index of (number of elements +1)/2 = (9 + 1) / 2 = 5

Or, to look at this another way, there are nine numbers, so if I count in from either end I get to a middle index of 5, that is the index for the median. There is a number at this index, namely 10. So 10 is the median value.

Count[portX] = 9 (this is the batch count)

Index of median = 5

Central Median[portX] = 10

Finding the Upper and Low Medians for an Odd Count Batch is a Little Tricky:

To find the upper median, construct one group of data that holds the central median value and all the values larger than it, then construct a second group with all the values above the central median. For, example, for the upper median, first group the data sets {10,12,16,20,20} and then {12,16,20, 20}. Now find the median of each group and average that to find the upper median value. This would be (16 + 18)/2 = 17.

upper median [portX] = (Median [{10,12,16,20,20}]+Median[{12,16,20,20}])/2= 16+18 =17

For the low median this would be the average of the medians of the two groups:

{2,4,5,7,10} and {2,4,5,7}

low median [portX] = (Median [{2,4,5,7,10}] + Median[{2,4,5,7}])/2 = 5+4.5 = 4.75

low Extreme[portX] = 2

upper Extreme[portX] = 20

After calculating these summary numbers, the remainder of the calculations is the same as the even numbered batches.

T

	(index)	count #9		
Central Median	5		10	
low/upper Medians	3	4.75		17
Extremes	1	2		20

median spread = 17 - 4.5 = 12.5

FIGURE 9. 5-Number Summary of the portX batch

Extra Practice (Optional Material)

Finding the 5-Number Summary for a batch with an even count

Consider the (sorted) batch of values, 'hum1', below which could represent observations of something, say, number of different hummingbirds observed hourly at a feeding station in the Phoenix Desert Botanical Gardens. The entries in this observation vector are called its *elements*.

hum1 = {2, 4, 5, 7, 10, 12, 16, 20}

There are 8 numbers/elements here and they are already sorted (this is usually the first step to be taken).

Remember, we are trying to get a handle on the 'look' of the batch without duplicating all of the original numbers. We have just calculated a middle number, 8.5 which, by itself, would be a limited picture of the batch. That first median, what I call the *Central Median*, divided the batch into halves and so a reasonable thing to do would be to find the median of each of those remaining halves, that is, a *upper median* and a *low median*.

These are some new terms for these two new medians we are about to calculate:

- The *Upper Median* is the number that divides the upper half of the data set in half - this means that we find the median of the upper batch {10, 12, 16, 20}. This works out to be (12 + 16)/2 = 14. NOTE: Another name for this number is the *Upper Quartile*. This name points out that a *quarter* of the batch of numbers lie above the upper median.
- *Low Median* is the number that divides the lower half of the data set in half. - this means for our batch that we find the median of the batch {2, 4, 5, 7} which is (4+5)/2 = 4.5. NOTE: Another name for this number is the *Lower Quartile* which means that a quarter of the batch of numbers lie below this low median value.

So, in effect, we have divided the data set up into quarters. So, if you like, you can equate *Upper Median to* Upper Quartile, and *Low Median* to Lower Quartile.

- *Median Spread* is the difference of the Upper Median - Low Median. NOTE: This is also called the *Inter-Quartile Range* (IQR).
- *Extremes* - the lowest number in the batch and the largest number in the batch are called the *extremes*. They are at the ends of the batch and have 'index' of 1. In this case the extremes are 2 and 20. NOTE: familiar common names for these numbers are *Minimum and Maximum* and the difference between the two is commonly called the *Range*.

- *Count* - the count of a batch is the number of elements it contains. An indication that a count is being presented is to prefix a number with the symbol '#'. So, the count of 'hum1' is 8, which might be indicated as #8.

Here is what we know about the batch 'hum1' so far

Count[hum1] = 8

Central Median[hum1] = (7+ 10)/2 = 17/2 = 8.5; at index 4.5

Low Median[hum1] = Median[{2,4,5,7}] = (4+5)/2 = 4.5; at index 2.5

Upper Median[hum1] = Median[{10,12, 16, 20}] = (12 + 16)/2 = 14; at index 2.5

Low Extreme[hum1] = 2; at index 1

Upper Extreme[hum1] = 20; at index 1

Range = Upper Extreme - Lower Extreme = 18

Median Spread = Upper Median - Low Median = 9.5 (this difference is also called the *Interquartile Range (IQR)* since this is equivalent to Upper Quartile - Lower Quartile.

5-Number Summary

To concisely summarize the characteristics of the 'hum1' batch, it is useful to arrange these numbers in a semi-graphical format as below, called a 5-number summary.

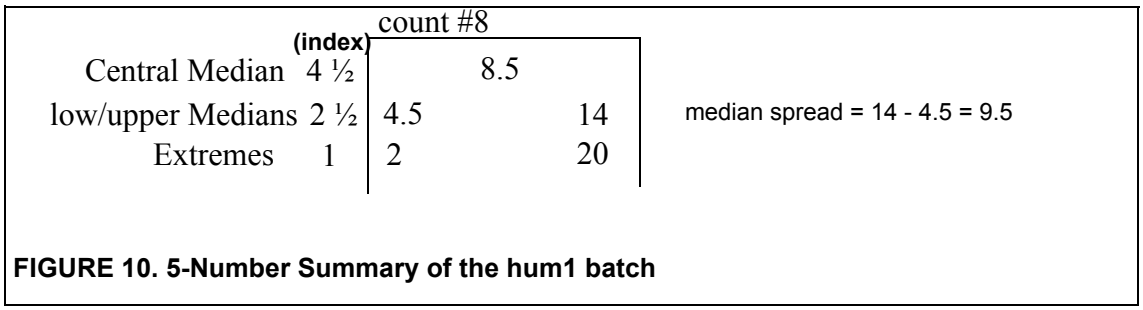

FIGURE 10. 5-Number Summary of the hum1 batch

Above is a standard semi-graphical way to show the 5-Number summary. The count of the batch is denoted by "#8", since there are 8 values in the batch. The central median *index* is 4.5 with a median value equal to the average of the numbers on either side of this index. the low and upper medians are at index 2.5, counting in from either end. the Lower Median value is 4.5 and the Upper Hinge value is 14. The extremes are 2 and 20.

Summary

Using the basic ideas presented above, you can gain a lot of insight with very little effort, by constructing box plots for various purposes, as indicated in the examples. These plots are very helpful when you want to compare multiple data sets, since drawing box plots side by side is a good way to compare the sets. (See Tufte's books, noted in the references sections, for graphical suggestions to make box plots even more effective).

References

Tukey, John. (1977), *Exploratory Data Analysis*, Prentice Hall.

Rucker, R. (2008) *Introduction to EDA*, milagrosoft.com.

References

Rucker, R (2008) *Business Trend Analysis*, milagrosoft.com

Rucker, R (2010) *Five Number Summary and Box Plots*, milagrosoft.com

Tufte, E. (1983) *Visual Display of Quantitative Data*, Graphics Press

Tufte, E. (1987) *Envisioning Data*, Graphics Press

Tufte, E. (2006) *Beautiful Evidence*, Graphics Press.