IRONWOOD INSIGHTS group, LLC

Graphic Display of Data: Box Plots

Box plots, also known as box and whisker plots, have been around for over 50 years. Many of today's modern researchers don't include box plots in their analytical toolkit even though the box plot chart provides a valuable way to graphically view and summarize large amounts of data.

They can be particularly useful when trying to evaluate the shape and variability of response distributions from several groups of respondents or on several different products. In this paper we will explore the uses of box plots along with their strengths and weaknesses compared to other data summarization tools.

Introduction

In today's world of big data, it is essential to have efficient methods of summarizing and displaying large amounts of data. Today's modern data analysts frequently use various descriptive statistics, such as means, medians, variances, and standard deviations. If comparing results across groups or products, it is easiest to utilize the summary of cross-tabulation tables. However, when there are many products or groups to evaluate, it can sometimes be extremely time consuming. In these cases, graphic summaries of data can be extremely efficient and helpful.

A box plot is one of the most useful graphic displays that is still used today. These simple plots communicate a great deal of information about the central tendency, variability, and shape of the distribution of responses. When placed side by side, box plots can be used effectively to compare results across variables, products, or groups.

Components of a Box Plot

The elements of a box plot can best be illustrated with an example. A frequency distribution of 420 responses obtained using an 11-point scale to provide an overall rating of a specific product is shown in Table One, along with the summary statistics. A box plot is shown in **Figure 1**.

The vertical axis identifies the scale of the plot. In the plot itself, the box extends from the 25th percentile to the 75th percentile. The median is shown as the horizontal line between the 25th and 75th percentiles and the mean is shown as an "X" in the same range (5.007 in this case). The length of the box is equal to a measure of variability known as the interquartile range. For normally distributed data, the interquartile range usually runs between 1.33 – 1.66 times the standard deviation. In this example, the box extends from 3 through 7 indicating that the middle 50% of the responses fall into a two-point range.

A quick check shows that:

Interquartile range / standard deviation => 4 / 2.548 = 1.57

So, our data appears to behave normally.

The vertical lines extending from the box are called “whiskers” which is why the box plot is sometimes also referred to as a box and whisker plot. If the data is normally distributed, each whisker extends somewhere between the length of the box (4 scale points in our example) or to the most extreme observation in that direction. Our whiskers extend 3 scale points in both directions.

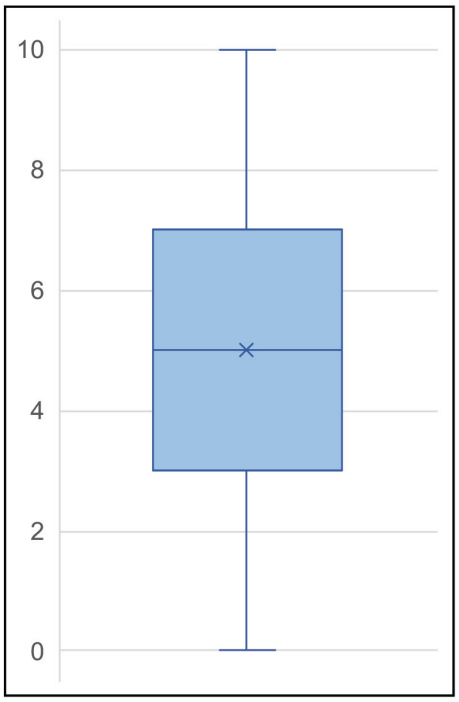
[Table 1]

FREQUENCY DISTRIBUTION

Response Alternatives	Frequency
0	17
1	20
2	26
3	39
4	51
5	57
6	53
7	38
8	27
9	21
10	16
Total	365
Mean	5.005
Median	5.00
75th Percentile	7.00
25th Percentile	3.00
Standard Deviation	2.548

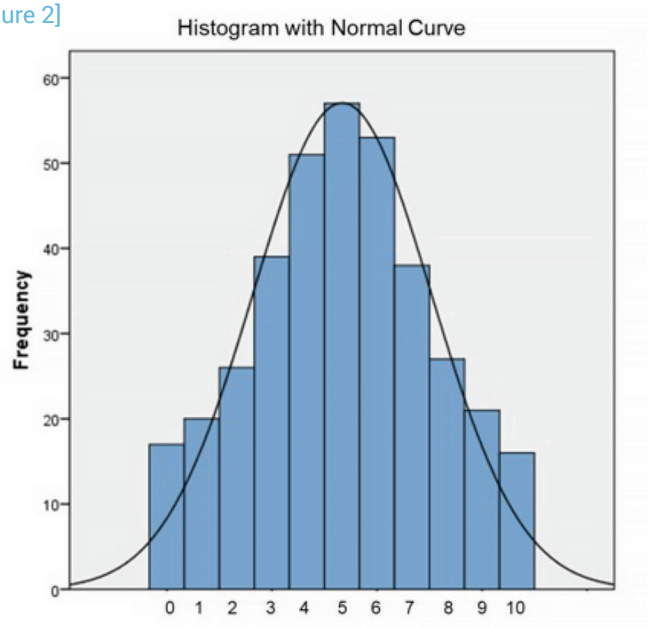
[Figure 1]

BOX PLOT



The box plot conveys information regarding the central tendency (the mean and the median) and variability (the range and the interquartile range). You can also determine the level of skewness (or asymmetry) in the data by examining the relative position of the mean and the median, by comparing the lengths of the whiskers, and by noting the location and number of outliers present in the data. From our box plot in **Figure 1** we can see that the data appears to be normally distributed. We can also see this from the histogram in **Figure 2**.

[Figure 2]



Interpreting Unusual Box Plots

Departures from a normal distribution alter the appearance of the box plot. To illustrate this, we show frequency distributions and descriptive statistics in **Table 2**, histograms in **Figure 3**, and box plots in **Figure 4**.

We are going to look at four examples of non-normal distributions: (A) skewed, (B) peaked, (C) flat, and (D) bimodal. Each of the four data sets include 100 responses on a 5-point scale.

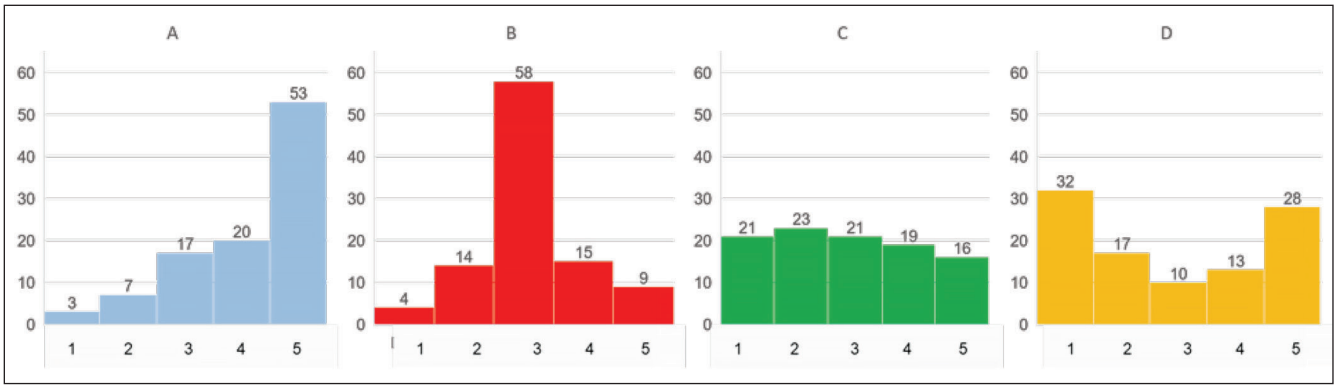
[Table 2]

EXAMPLES OF NON-NORMAL DISTRIBUTIONS

	Frequencies			
Response Alternatives	A Skewed	B Peaked	C Flat	D Bimodal
1	3	4	21	32
2	7	14	23	17
3	17	58	21	10
4	20	15	19	13
5	53	9	16	28
Mean	4.13	3.11	2.86	2.88
Median	5	3	3	3
Standard Deviation	1.110	0.893	1.371	1.639
25th Percentile	3	3	2	1
75th Percentile	5	3	4	5

[Figure 3]

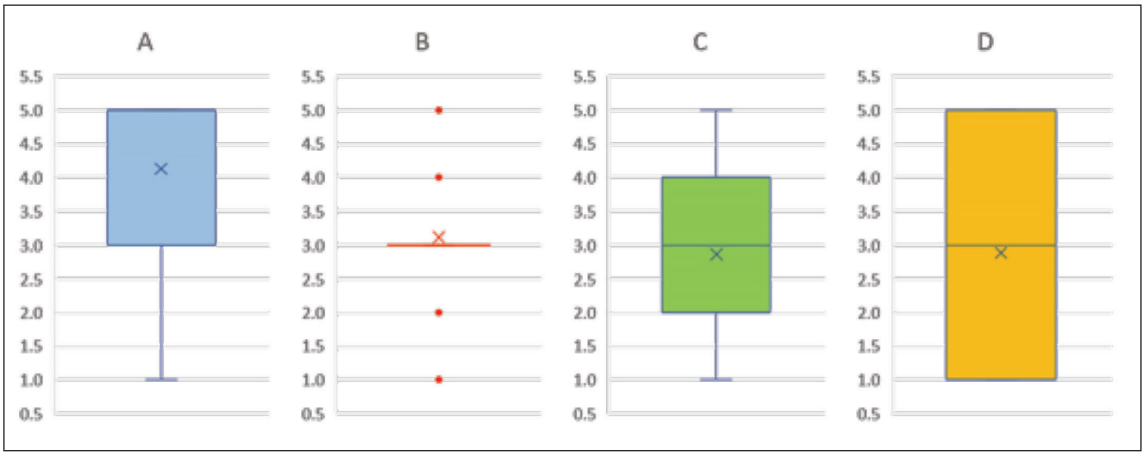
HISTOGRAMS OF NON-NORMAL DISTRIBUTIONS



When the data are skewed, (example A), the median may occur at or near one end of the box plot and the whiskers will be of unequal length. As we can clearly see in the histogram above, the data are so skewed that the median is at the highest value of 5. On the box plot for example A below we see that this skewness results in the disappearance of the upper whisker.

When the data are peaked, (example B), the box and the whiskers are relatively short, and as in our extreme case, they disappear entirely. As we can see from the data in Table Two, the middle 50% of the responses all have the same value (3), so the length of the box and the whiskers is zero.

[Figure 4]



When the data are flat, (example C), the distribution is displayed as having a relatively long box, and in our example, the whiskers are shorter than the box.

When the data are bimodal, (example D), the data result in a long box. In our case, the middle 50% of the responses range from 1 to 5, which is the entire range of the scale, so our whiskers vanish. We frequently experience these types of distributions when two sub-groups of individuals are combined.

As we can see from our examples, most types of non-normality have predictable effects on box plots and can be easily identified.

Uses of Box Plots

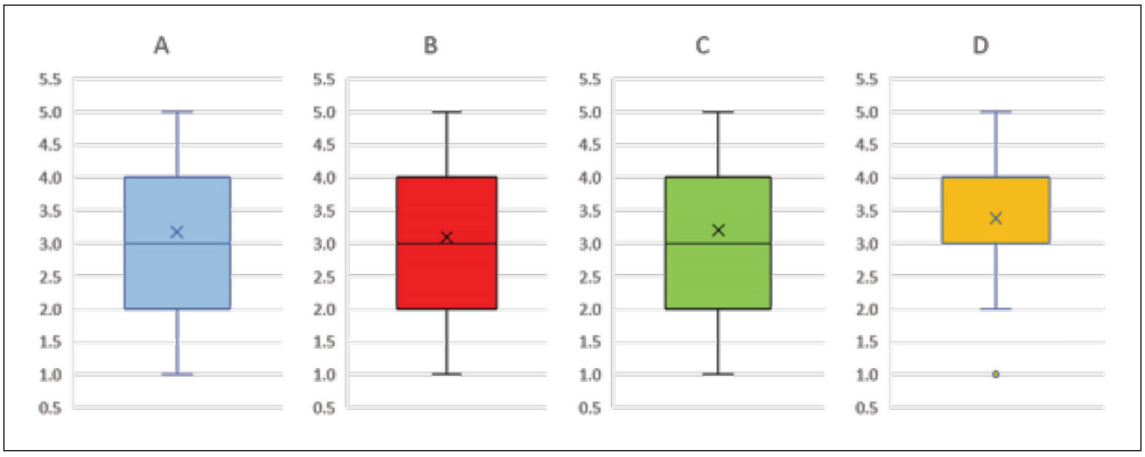
Since they contain so much information, box plots have several useful applications when trying to understand or explain the data. They can be used to graphically present descriptive information about the data – means, medians, measures of variability, and skewness. They can also aid when assessing the normality of the data when combined with histogram charts.

Side-by-side box plots are excellent tools for comparing responses to various questions or responses by different groups of respondents to the same question. We can easily compare the average response and the variability from one variable or group of respondents to another.

An example is shown in **Figure 5**. Here we see side-by-side plots of the overall ratings of four products. From the charts we can see that Brand D is rated highest, followed in order by C, A, and B. Also, we see that there is less variability in the Brand D data than in the data of the other brands.

Side-by-side box plots can also be used to examine relationships between two quantitative variables. Frequently in market research, respondents are asked to rate the overall quality of a product along with their likelihood of purchasing the product again in the future. Researchers can easily create separate plots of likelihood of future purchase for respondents at each level of quality. These box plots can then be placed side by side. This would reveal the future likelihood of purchase and how it changes at different levels or quality ratings as well as the variability associated with each level of perceived quality.

[Figure 5]



Box Plots vs. Other Types of Data Summaries

Box plots are effective supplements to cross-tabulations and other descriptive statistics. As shown in the preceding section, they are particularly useful when comparing responses among respondent segments or various distributions of responses. Since the mean, median (50th percentile), and the 25th and 75th percentiles are clearly shown, they provide more information than simple frequency tables or histogram charts. Also, unlike simple descriptive statistics, box plots quickly reveal if the data has any extreme data points or outliers.

Box plots are not without their limitations. Data distributions with different shapes can have similar plots. Also, the number of outliers is not shown. Both of these limitations can be overcome by examining the frequency tables when an unusually shaped box plot appears or by utilizing other graphical methods to display the data distribution.

Box plots are quite effective tools for displaying data in an efficient format. Such plots can greatly ease the burden of examining large quantities of data and comparing responses across variables or groups of respondents. The graph option in Excel allows anyone to create box plots quickly from the data.