Notes:

Box Plots

Key Learning Points

1. Describe the importance of box plots.
2. Explain how to create and interpret box plots.
3. Utilize box plots in improvement projects.

What is a Box Plot?

Box Plots provide a graphic summary of the pattern of variation in a set of data. They are especially useful when working with small sets of data.

Few Points of Data Needed

Box plots can be used to assess as few as 8 data points.

Shows Outliers

Box plots easily identify which observations may be outliers.

Shows Comparisons

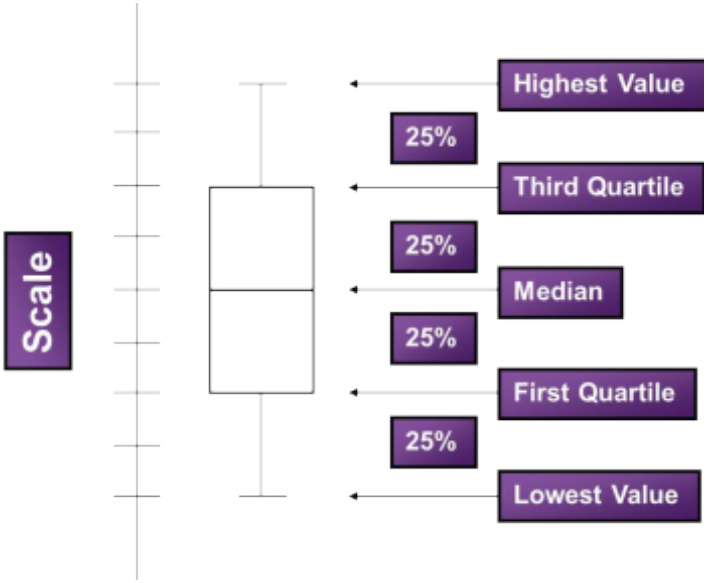
Box plots offer a quick side by side comparison of multiple sample sets.

Basic Box Plot

A basic box plot is a five number summary of a set of data consisting of:

- Lowest value
- First Quartile (the value below which 25% of the data lies)
- Median (the middle value)
- Third Quartile (the value below which 75% of the data lie)
- Highest value
- The five values divide the set of data into four equal groups
- Shows empirical distribution of data-it simply displays the pattern of the data and does not assume normal distribution
- Allows team to see some aspects of the pattern of variation in the data

The five values of the box plot divide the data into four groups of equal depth. Note, however, that the symmetry of the box plot to the left may be deceiving. Also note that three of the five values are calculated and may not actually be observed data values.



Characteristics

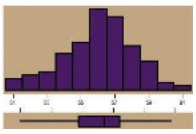
- The Lowest and Highest values refer to the lowest and highest data point value.
- To determine the quartiles and the median we need to know the number (n) of observations (data values) in the data set.
- You must arrange the data values in order from lowest/smallest to highest/ largest. This is referred to as an ordered data set.

Notes:

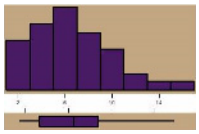
Common Histogram Patterns And Corresponding Box Plots

Unfortunately, the ability to identify and classify the specific pattern of variation in a set of data is somewhat limited when using box plots. The histogram simply displays more information than the box plot— the relative number of observations in each of 6 to 12 cells with a histogram is known, compared to a five-number summary of the data with a basic box plot.
The figure below shows common histogram patterns with corresponding box plots. The box plots do not do a very good job of distinguishing the patterns. At best they make a statement about the width of the variation and the presence of various forms of asymmetry.

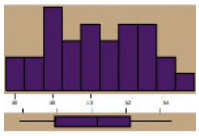
Bell Shaped



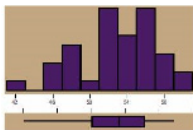
Skewed



Plateau



Isolated Peaked



Example: Is The Copier Down Again?

An Information Management Department was reviewing complaints that copiers kept failing. The response time of the repair person was a major controllable factor in the availability of the copiers. The copiers were serviced by two different firms, under different maintenance contracts. A study was conducted to determine which

Notes:

supplier provided the fastest and most consistent response to a repair call. The team defined response time as the "time between the call for repair and the arrival of the repair person." They then pieced together the data by comparing various records and log books on a random sample of visits by each supplier.

Contractor Response Data

The team limited the data collection to a sample of ten visits for each contractor because the data were difficult to generate.

To construct the box plots, the team first recopied the data for each contractor, putting it in order from the smallest value to the largest value. This ordered dataset is shown below.

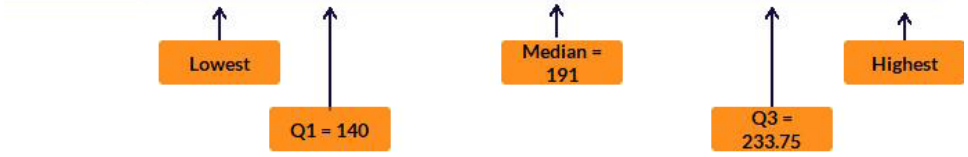
		1	2	3	4	5	6	7	8	9	10
Contractor	A	45	183	207	226	243	251	268	272	290	345
	B	102	125	145	156	190	192	215	228	251	279

Conctractor A Response Data

		1	2	3	4	5	6	7	8	9	10
Contractor	A	45	183	207	226	243	251	268	272	290	345

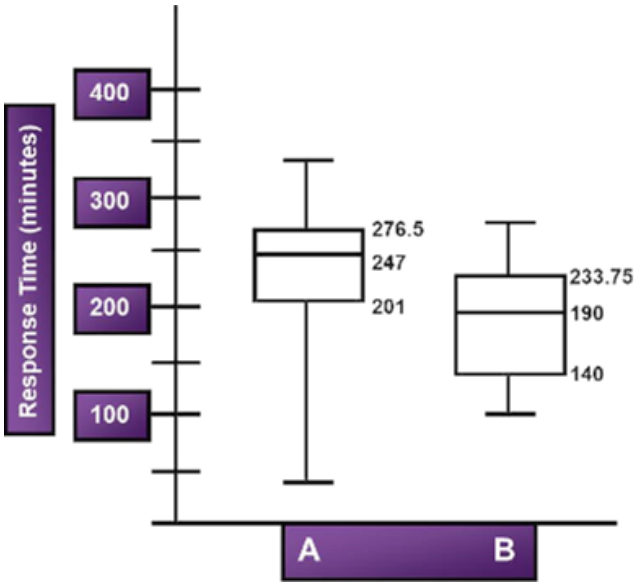
Conctractor B Response Data

		1	2	3	4	5	6	7	8	9	10
Contractor	B	102	125	145	156	190	192	215	228	251	279



Notes:

Contractor Response Box Plot



Contractor Response Conclusion

The team concluded that Contractor B was providing superior service because the box plots clearly showed that B's response was usually shorter and more consistent than A's. The team made a number of specific observations from the box plots:

- The median response time of B was 191 minutes, compared to 247 minutes for A.
- The variability in A's response time was greater than B's.
- Even though A had the shortest response time, only about 25 percent of A's response times were less than 201 minutes, while 75 percent of B's were less than 233.75 minutes.
- Nearly 25 percent of A's response times were longer than B's longest response time.

The box plot gave the team a "picture" of the data and enabled them to see some aspects of the pattern of variation, even though the number of data points was small.

Schematic Box Plot

In a schematic box plot, the box and median have the same definitions as a basic box plot. The whiskers, however, are different. In the schematic box plot, the whiskers are shown as dashed lines extending from the edge of the box to points called adjacent values.

To determine the adjacent values, first compute the inter-quartile range (IQR).

- The IQR is simply the length of the box—i.e., the difference between the third and first quartiles.

Notes:

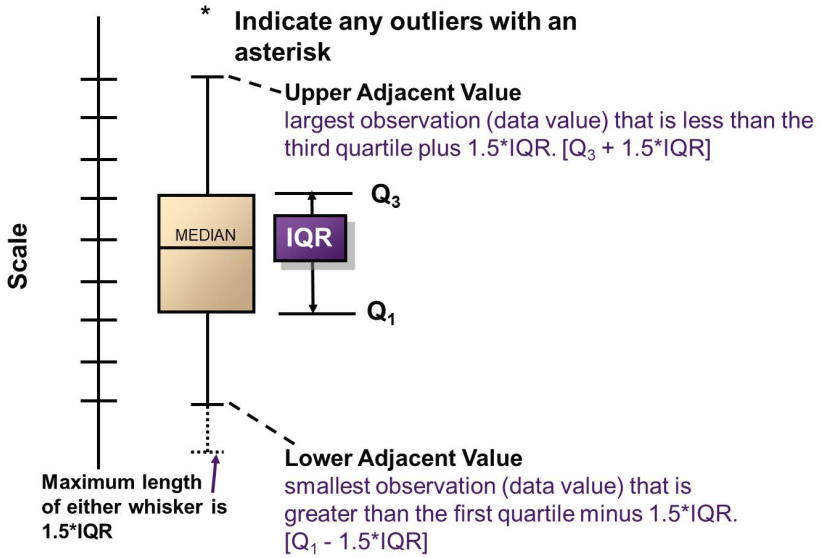
- The upper adjacent value is the highest value in the dataset that is not an outlier.
- The lower adjacent value is the lowest value in the dataset that is not an outlier.

Points that lay more than 1.5 times the IQR beyond the first or third quartiles are called outliers. Outliers are like the extreme values discussed in the skewed and isolated peak distributions in the discussion on histograms.

Teams will want to investigate these outliers to see if they are the result of measurement error, or more importantly, some unusual set of conditions in the process.

The values for terminating the whiskers are now limited and are called the lower adjacent value and the upper adjacent value, instead of the highest and lowest value.

Observations beyond these values are called outliers.



Example: Hospice vs. Hospital Patient Costs

Observation: Hospice patients' acute care costs appear to be less than hospital patients' acute care costs. There is less variation among hospice patients (less spread). There are more significant outliers among hospital patients than there are among hospice patients.

Hospice and Hospital Costs

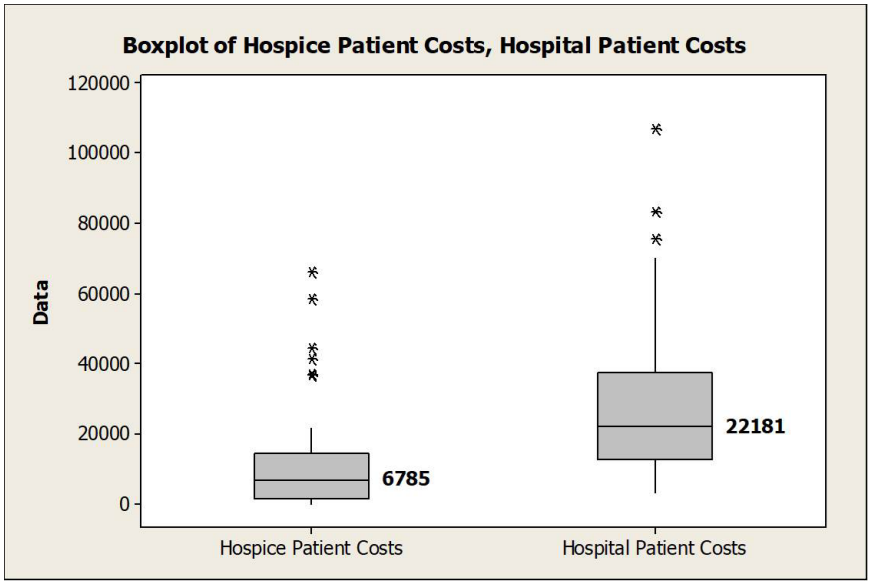
The median acute care costs for patients in hospice is $6,785 compared with the median acute care costs of $22,181 for patients who remain in the hospital. The inter-quartile range (where 50% of the data lays) for hospice data is $13,077 compared with the IQR of $24,860 for hospital patients. The length of span for the whiskers for hospital patients is much longer than for hospice patients, indicating

Notes:

there is less variability in cost for hospice patients than for hospital patients. Although hospice patients have more outliers than hospital patients, the highest cost for a hospice patient is $66,435. This is half of the highest cost of $107,182 for a hospital patient.

	Median Acute Care Costs	Interquartile Range	Highest Cost
Hospice	$6,785	$13,077	$66,435
Hospital	$22,181	$24,860	$107,182

Hospice and Hospital Costs Schematic Box Plot



Sample Size = 688, Source SMS

Hospice and Hospital Costs Conclusion

While the table of data clearly shows that Hospice costs are lower, the schematic box plot displays this in a much more striking way. The entire IQR of the Hospice costs is below the lower end of the IQR of Hospital costs.

Steps in Constructing Box Plots

1. Collect the raw data, and convert it to an ordered dataset by arranging the values from the lowest to the highest.
2. Decide on the type of box plot you wish to construct.
   a. Begin with a basic box plot.
   b. If some whiskers are greater than 1.5 times the length of the box, convert all to schematic box plots.
   c. If other quartiles are of interest, redefine the box plot.

Notes:

3. Calculate the appropriate numerical summaries.
4. Draw and label the horizontal axis.
    a. Provide labels and a caption to identify each box plot.
5. Draw and label the vertical axis.
    a. Begin labeling the axis with a multiple of 1, 2, or 5 that is smaller than the lowest value in any box plot.
    b. Continue labeling the axis until you reach a multiple of 1, 2, or 5 that is larger than the highest value in any box plot.
    c. Provide a caption to describe the variable and its unit of measure.
6. Draw the box plots.
    a. Refer back to the image "Basic Box Plot" for an example of a basic box plot.
    b. Use dashed whiskers for schematic box plots. The dashed whiskers should extend to the highest and lowest values in the dataset that are not outliers.These points are called adjacent values.
    c. Use asterisks to indicate all of the outliers on schematic box plots and the highest and lowest values on box plots which use other quartiles.
7. Show a key to define the parts of the box plot.
8. Title the chart, and show nominal values and limits (if applicable).
9. Analyze the pattern of variation.
10. Develop a plausible and relevant explanation for the pattern, and determine your next steps.

Interpretation

Values in any set of data will vary, and the variation will display some pattern. Box plots are a good way to compare patterns of variation among datasets. In comparing patterns of variation with box plots, look for:

- Differences in the location of the median
- Differences in the amount of variation
- The presence or absence of outlier points
- Symmetry or asymmetry in the data

Quantifying Patterns of Variation

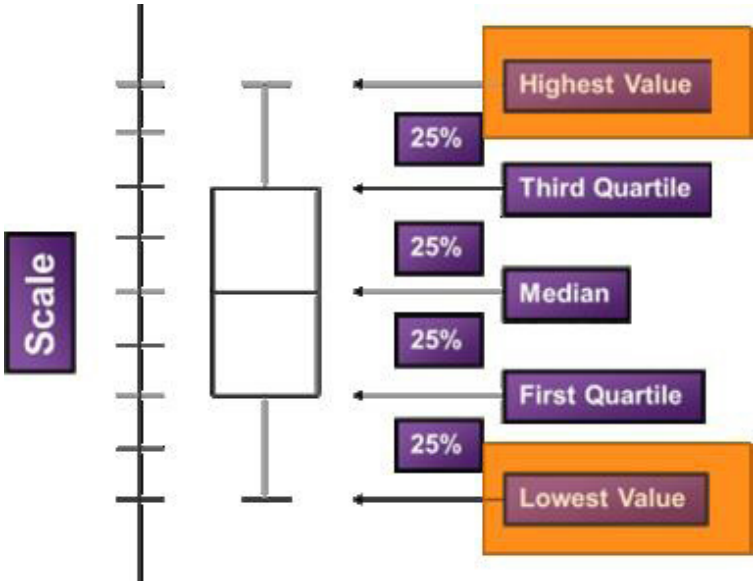
While the box plot gives less information about the shape of the pattern of variation than the histogram, it does provide more quantitative information about the pattern. For instance, you can say things such as: "50 percent of the observations lay

Notes:

between X and Y," or "25 percent of the measured values were greater than X."

Highest and Lowest Values

Indicated by the ends of the whiskers on the basic box plot, and by asterisks on other types of box plots. These are the maximum and minimum values in the set of data. If these values are about the same distance from the median, then the distribution is roughly symmetrical. Caution: If the data is from a sample, these are the highest and lowest values in the sample only—larger and smaller values may exist in the population, but were not selected in the sample.

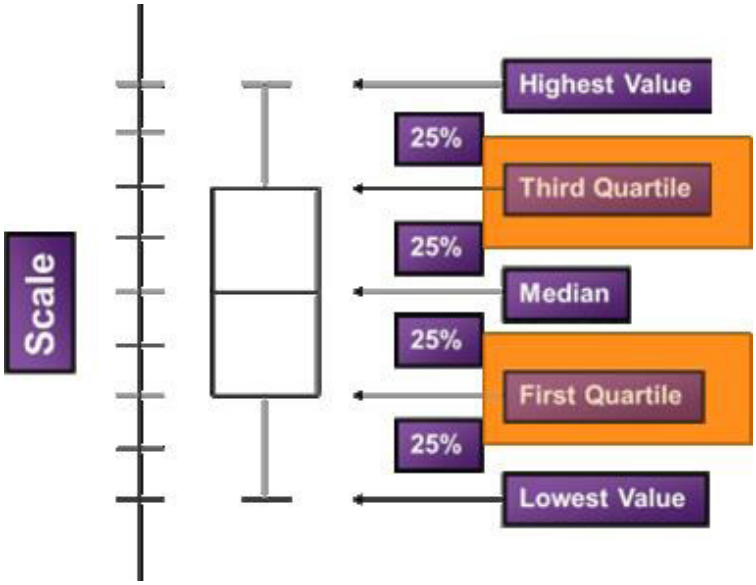


Quartiles

Indicated by the ends of the box in standard box plots. About 25 percent of the observations lay below the first quartile and 25 percent above the third quartile. About 75 percent of the observations lay above the first quartile and 75 percent below the third quartile. About 50 percent of the observations lay between the first and third quartiles.

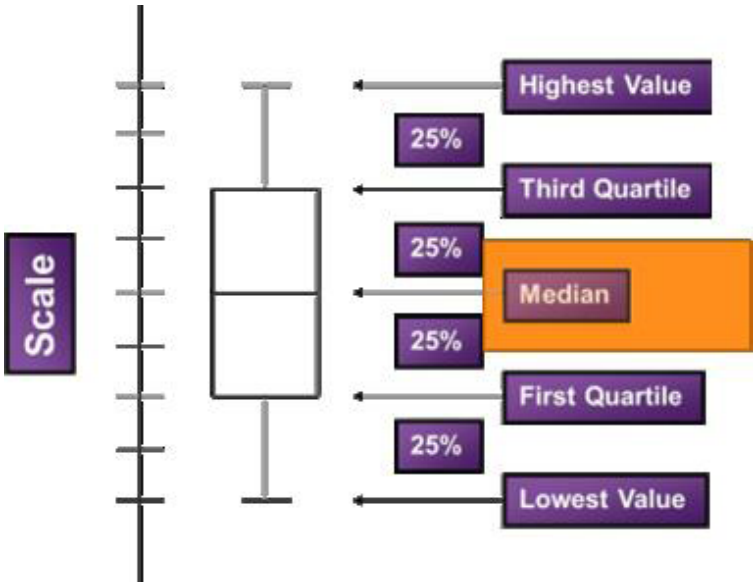
Notes:

Notes:



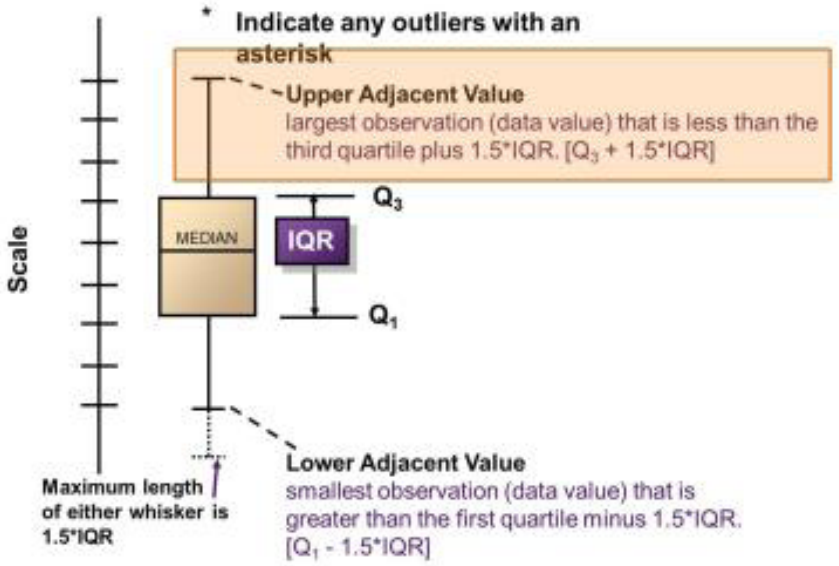
Median

Indicated by a horizontal line dividing the box in standard box plots. Half of the observations lay above and half lay below the median. If the median lays near the center of the box, then the distribution is roughly symmetrical.



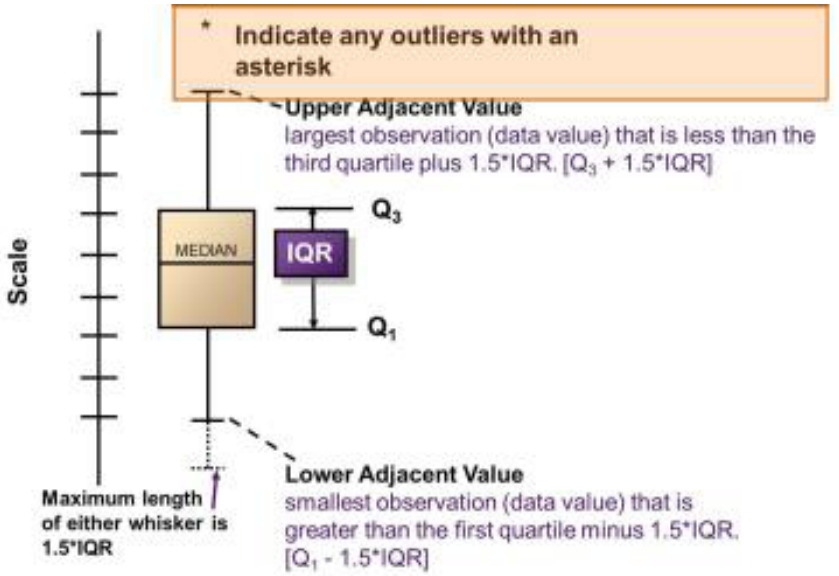
Adjacent Values

Indicated by the ends of dashed whiskers on the schematic box plot. The upper adjacent value is the largest observation that is less than or equal to the third quartile plus 1.5 x IQR (IQR = the length of the box). The lower adjacent value is the smallest observation that is greater than or equal to the first quartile minus 1.5 x IQR. While these values lay some distance from the center of the distribution, they are not unusually far away.



Outliers

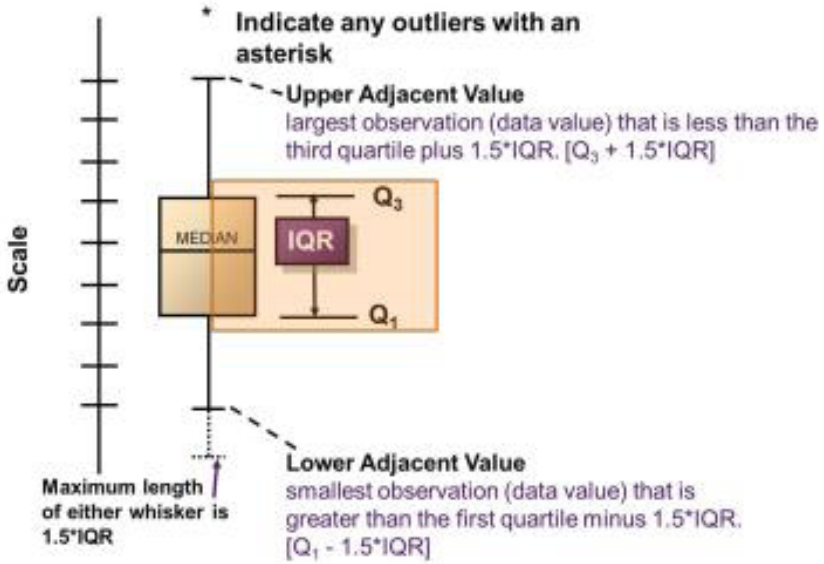
Indicated by asterisks on the schematic box plot. These are observations in the data that lay outside the adjacent values. These observations are unusually far from the center of the distribution, and the team should investigate to see if they are the result of measurement error or an unusual set of conditions in the process.



Notes:

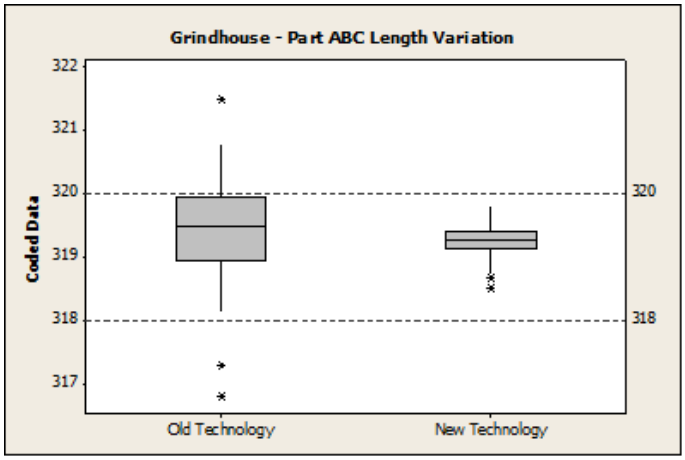
Other Quartiles

Indicated by box and whisker lengths that have been redefined to represent different quartile values. The specific definition of these nonstandard box plots should be shown in a key somewhere on the chart. The q(.X) quartile means that about X percent of the observation lays below the quartile boundary and (100 – X) percent lays above.



Specification or Target Limits

Indicated by solid or dashed horizontal lines across the chart.



When Should a Box Plot Be Used?

- Testing Theories and Identifying Root Causes
- Remedying the Cause and Holding the Gains

Notes:

- But Which Box Plot Should I Use?
  - Construct a basic box plot.
  - Are the whiskers on some of the boxes longer than 1.5 times the length of the box? If so, convert the whiskers on all the boxes to a schematic box plot to highlight any outlier points.
  - Are other quantiles of the data of special interest? If so, redefine the box plot and calculate appropriate values.

Pitfalls to Avoid

The primary motivation for the use of the box plot is to get around the need for 40 or more data points. If you use box plots, you must accept the fact that you will not be able to discern the various histogram patterns.

- Data must represent the full range of current conditions.
- The more data points, the better the interpretation you can make.
- Interpretation is based on theory and must be confirmed through analysis.

Notes: