Histograms

Distribution and Causal Analysis Tools

The Histogram is a useful tool for breaking out process data into regions or bins for determining frequencies of certain events or categories of data. These charts can help to show the most frequent causes of problems, distribution of process data, or provide other process improvement information.

Usually a histogram is performed over a fixed time interval of performance indicator results. The histogram should only be performed after a trend analysis has been completed using a control chart. Mixing process results from two statistically different time intervals will limit the usefulness' of the histogram. The time interval should be chosen to correspond to either:

- A statistically stable time period (no significant trends)
- Data points which have been identified as a significant trend.

This consideration of significant trends is important as the frequency distribution in each region or bin may differ due to the process changes which occurred to cause the significant trend. Thus the arbitrary choice of "Fiscal Year to Date" or "Calendar Year to Date" or "The past two years" may not be appropriate.

A histogram is generally shown as a bar chart. The natural order of the categories is maintained when performing a histogram. Note that a Pareto chart sorts the categories from most common to least common.

Important Note:

If you are intending to use a bar chart or histogram to find trends, you are reading the wrong section of this primer. Go back to Guidelines for Statistical Process Control if you are trying to find trends. Histograms and Pareto charts should only be made after you have completed the initial trend analysis and are trying to determine what to do to improve the process. Although the histogram and Pareto charts are useful to analyze process data, the time sequence that the process data occurred in is lost in this analysis. The control chart maintains the time sequence of the data.

Example of a Histogram in Action

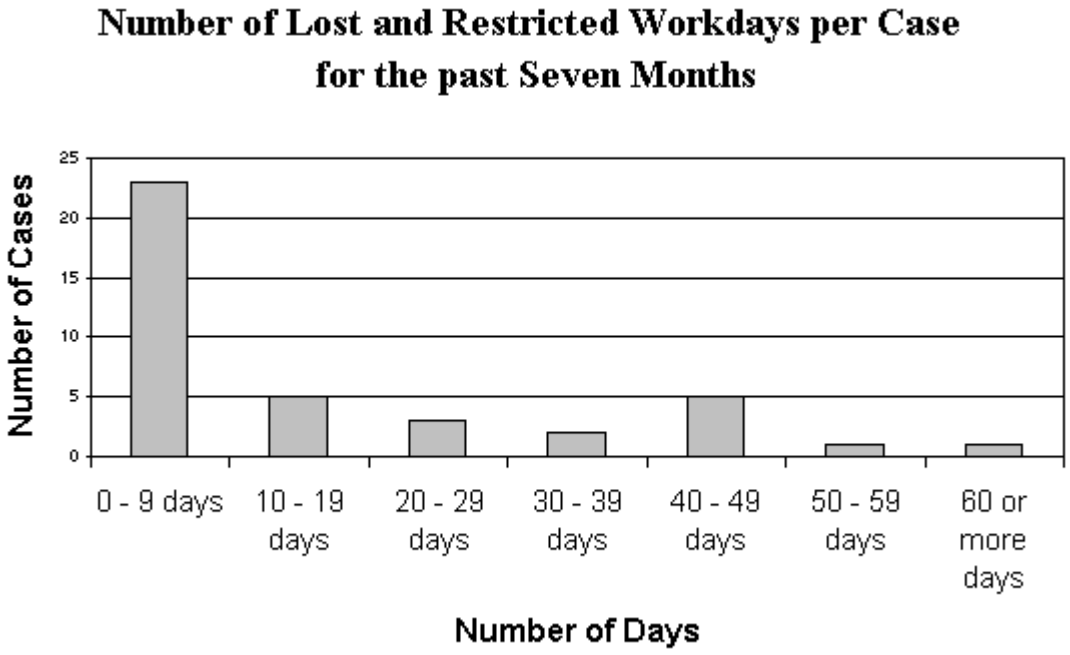
Let us assume we have a performance indicator for Lost or Restricted Workdays due to occupational illness and injuries. Further, we are concerned because the past seven months have been above the baseline average (an indication of a significant increase), and we need to determine the source of this trend. The time period for histogram analyses will be the past seven months.

A histogram analysis could be used to look at the distribution of the number of days in each injury and illness case. The histogram will show if we have a problem with a large

number of cases with a small number of days each, or a small number of cases with a large number of days each.

Steps in Making the Example Histogram

Step	Example
Define the data	The Number of Lost or Restricted Work Days per Case
Define the time period for the data	Past seven months of cases
Tabulate the data	List the number of days in each case: 47, 1, 55, 30, 1, 3, 7, 14, 7, 66, 34, 6, 10, 5, 12, 5, 3, 9, 18, 45, 5, 8, 44, 42, 46, 6, 4, 24, 24, 34, 11, 2, 3, 13, 5, 5, 3, 4, 4, 1
Determine the Range of the data (minimun value and maximum value).	The data ranges from a minimum of 1 day to a maximum of 66 days per case
Decide on the number of bins, and the width of each bin (usually 5 - 7 equally spaced bins over the range of the entire data)	Use seven bins in ten day increments starting from zero (0-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60 or more). This gives a convenient grouping of the data in easy to work with increments, covering from 1 to 66.
Count the number of items in each bin	23 cases ranged from 0 to 9 days, 5 ranged from 10 to 19 days, 3 from 20 to 29 days, 2 from 30 to 39 days, 5 from 40 to 49 days, 1 from 50 to 59 days, and 1 involved more than 60 days.
Make a bar chart of the data using graph paper or a computer graphics routine	See example made using Excel spreadsheet below



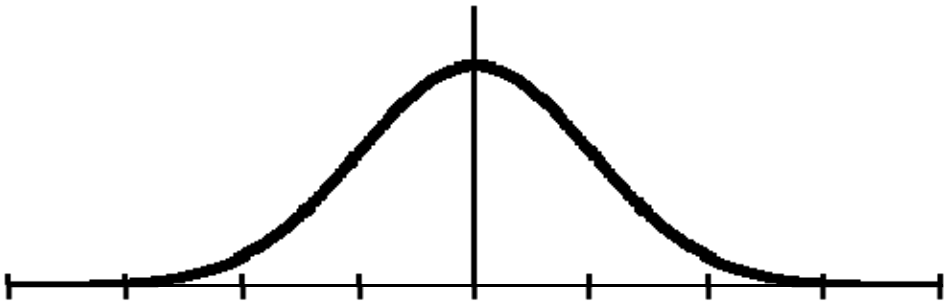
There were seven equal width intervals chosen for the histogram. The histogram shows that most cases involve less than ten days, with a tapering off of cases in the higher value bins. The exception is 40 to 49 day bin, which is higher than the previous two bins. This may be the source of our increase and these five cases may need to be looked at in detail. A comparison with the shape of a histogram of cases prior to the increase may be worthwhile.

Some examples of histogram analyses use unequal width bins. This should be avoided if at all possible. Most observers expect the bins to be of equal widths, and one can distort the data presentation (and affect the conclusions drawn) by manipulating interval widths.

Histogram Shapes

The shape of the histogram distribution can be used to diagnose the process that generated the data. This discussion assumes the bin intervals chosen were of equal width.

Generally, one should expect a generic histogram to follow a "bell-shaped" curve (or "Normal Distribution"). This type of distribution has one peak at the center of the distribution, and the data trails off towards the tails. A bell-shaped curve is shown below:



The example histogram differs from the bell-shaped curve in two respects.

- **SKEWING** First, the peak bar is at the far left hand side of the distribution (23 cases in the 0 to 9 day bin). This type of distribution is a "skewed" distribution (skewed to the left in this case). This form of skewed curve is expected when the data is near zero, and no negative numbers are allowed. The left hand side of the bell curve is in effect chopped off. Time between failure data is often skewed in this form also, and can follow a specialized curve known as the "exponential distribution".
- **BIMODAL** Second, there is a second peak in the histogram at the 40 to 49 day bin. This is known as a "bimodal" distribution. The distribution has two peaks (or "modes"). We may then decide that any case with more than 40 days should be looked at in more detail, as these six cases contribute 52% of the lost/restricted days for the seven months.

Determining the source of the second peak in a bimodal distribution will often provide insight into methods to improve the process. The second may be a result of a "special cause" in the distribution. For example, if one were to plot the body weights of a group of people, one would find two peaks. The first peak would correspond to the central tendency for female body weight, and the second peak would correspond to the central tendency for male body weight. Thus, there is a "special cause" impacting the data - the gender of the person.