The Empirical Rule

ASQ 904 SEPTEMBER MEETING

TAKEN FROM DR. DON WHEELER'S PAPER BY THE SAME NAME

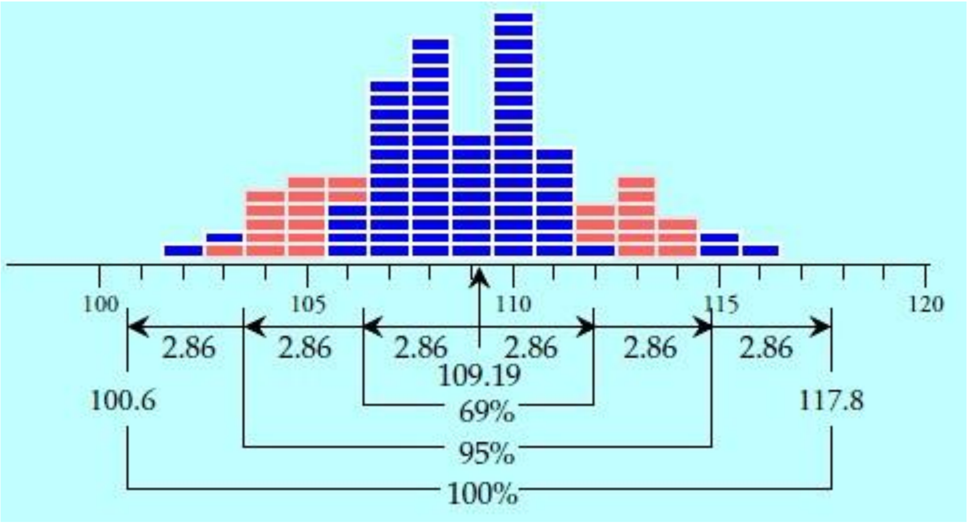
What is the Empirical Rule

- The Empirical Rule simply converts the average and standard deviation into comprehensible percentages.
- These are:
- • Part one: Roughly 60 percent to 75 percent of the data will be found within the interval defined by the average plus or minus the standard deviation statistic.
- • Part two: Usually 90 percent to 98 percent of the data will be found within the interval defined by the average plus or minus two standard deviations.
- • Part three: Approximately 99 percent to 100 percent of the data will be found within the interval defined by the average plus or minus three standard deviations.

Can it really be that simple

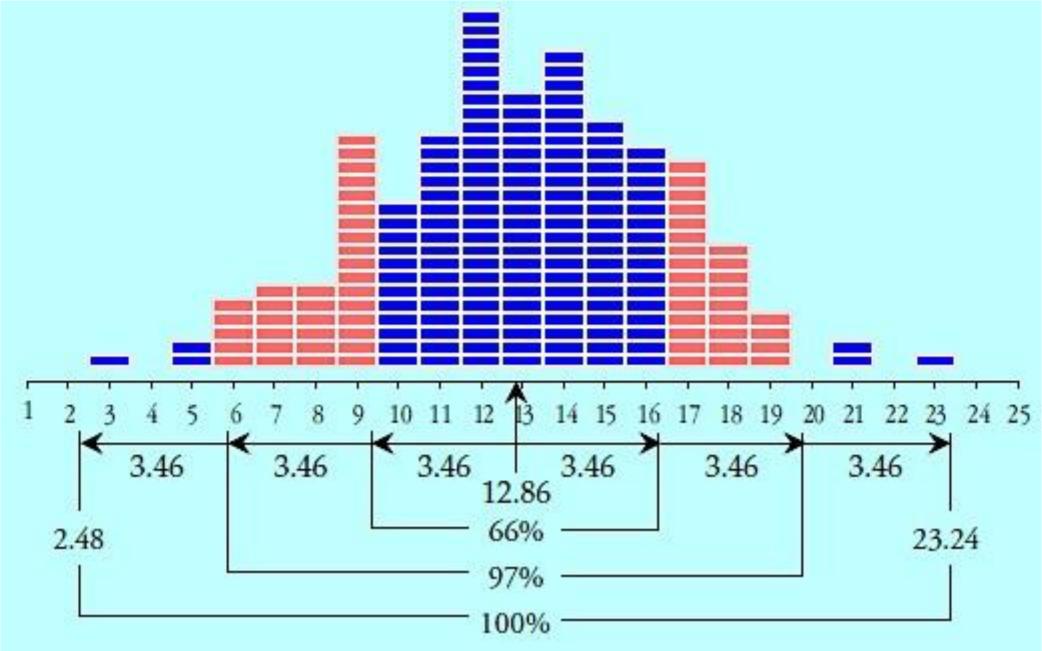
- Don’t we need to fit some probability distribution model first? Like making sure it is a normal distribution?
- We’re going to look at some example distributions to check that out.

Wire length data



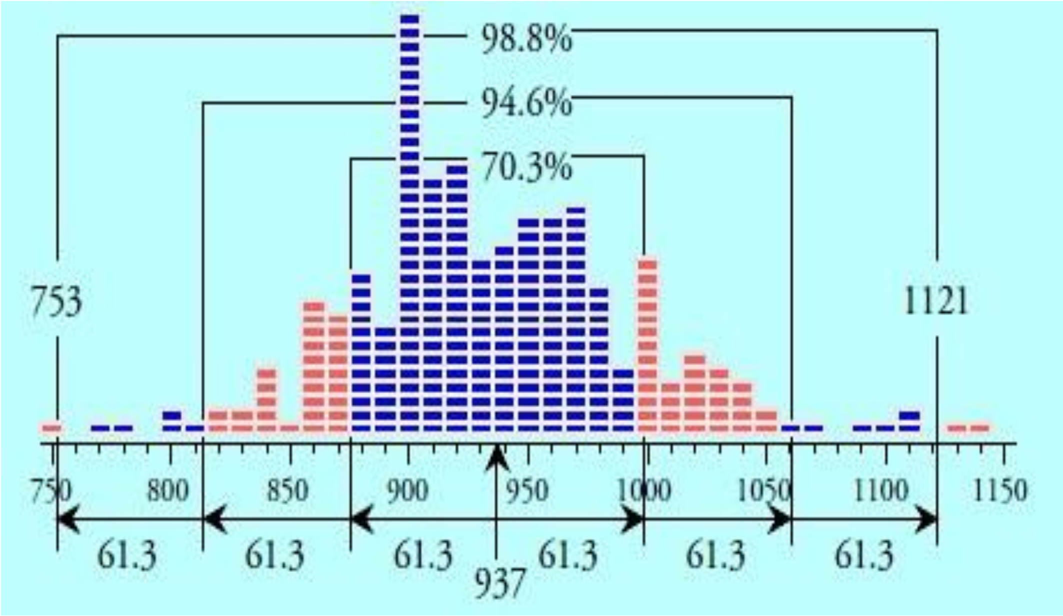
Here are 100 wire lengths with a mean of 109.19 and standard deviation of 2.86. You can see that 1,2,and 3 sigma intervals contain 69%, 95% and 100% respectively. Looks pretty normal, don't you think?

Bead Board Example



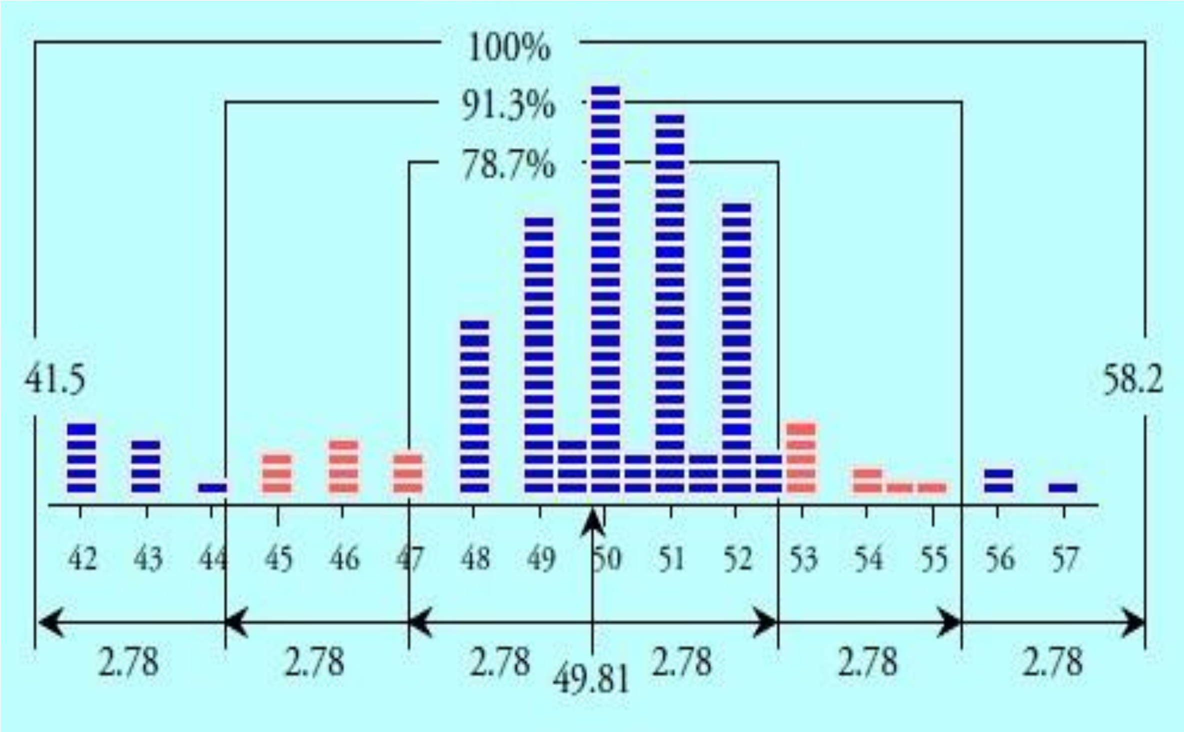
Here are 200 data from a bead board or Quincunx. They have a mean of 12.86 and sigma of 3.46. Note their 1,2, and 3 sigma intervals contain 66%, 97% and 100% respectively. Again reasonably normal looking data.

Batch Weights



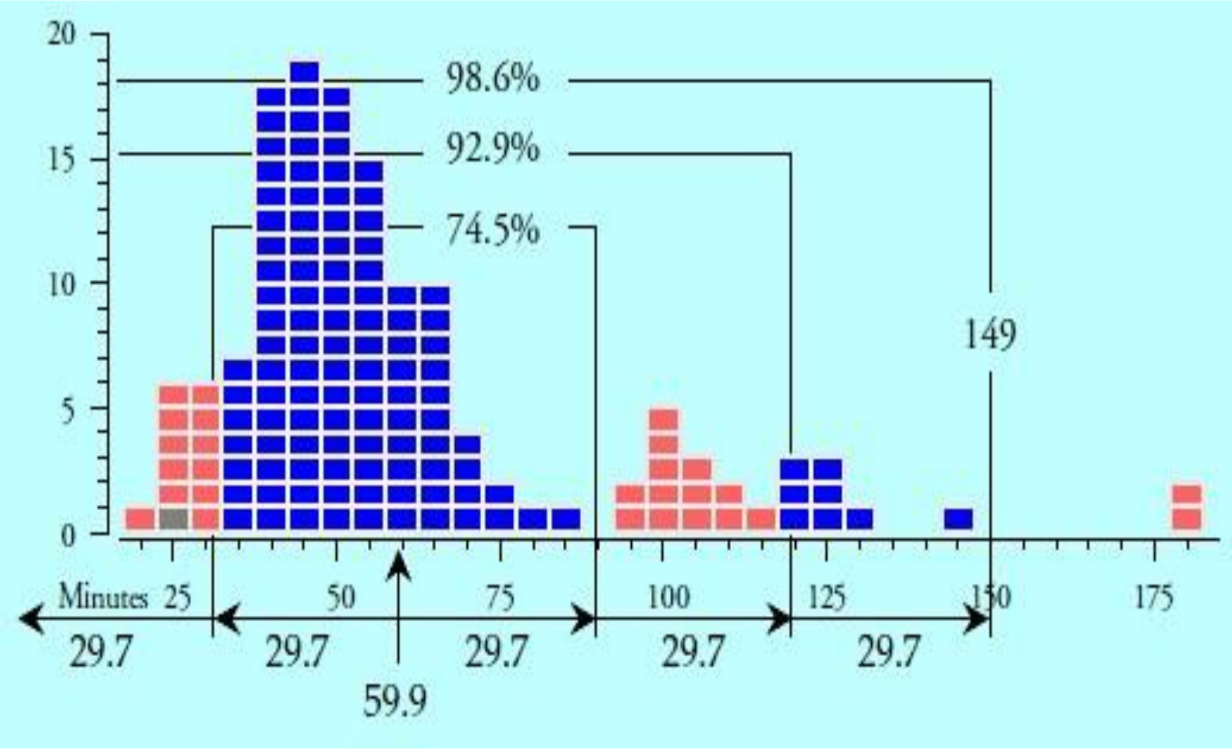
Here are 259 batch weight data, with a mean of 937 and sigma of 61.3. Note their 1,2 and 3 sigma intervals have 70.3%, 94.6% and 98.8% respectively. Does not look very normal!

Cam shaft bearing diameters



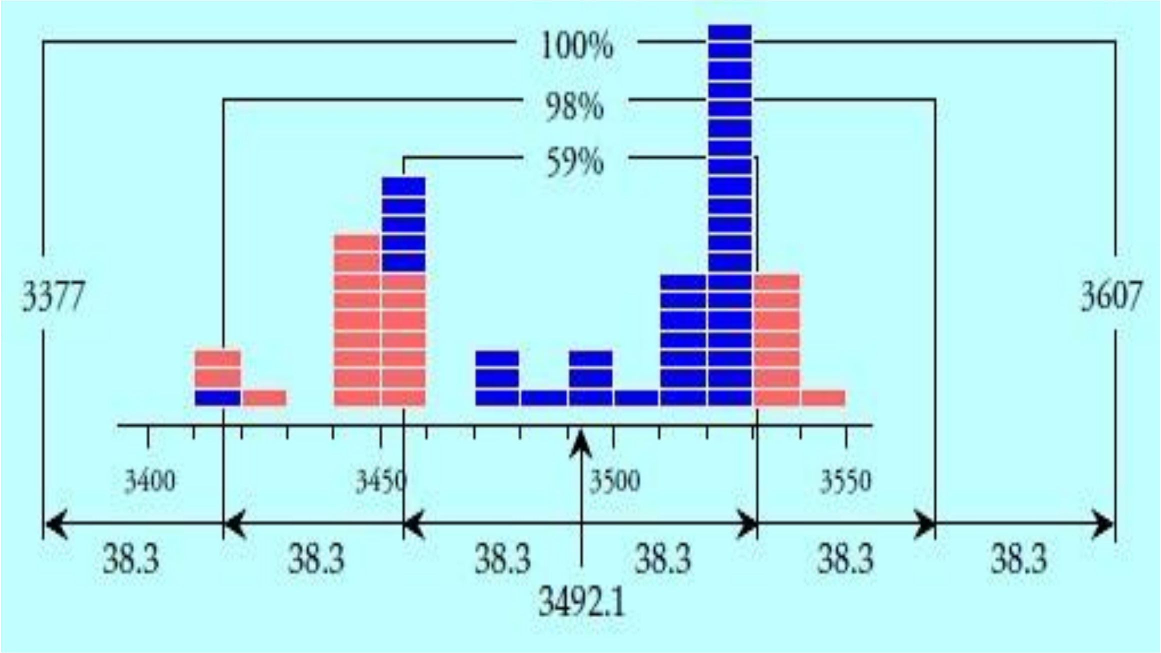
150 cam shaft bearing diameters with a mean of 49.81 and sigma of 2.78. Now we have 78.7%, 91.3% and 100% for the 1,2, and 3 sigma intervals. Definitely not normal.

Hot metal delivery time



Definitely not normal (time data rarely are) but the 141 time values have a mean of 59.9 minutes and sigma of 29.7. The three intervals yield 74.5%, 92.9% and 98.6% respectively

Last, Creel Yield data

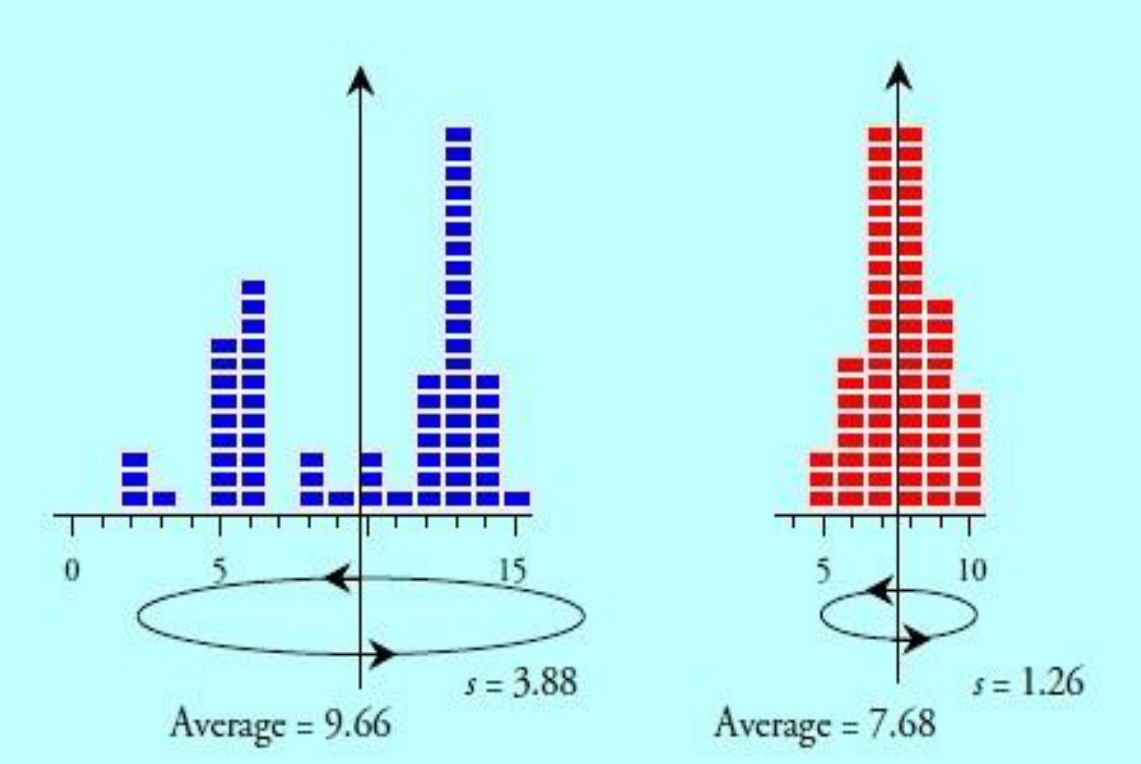


Last, 68 values from Creel Yields(?)with a mean of 3492.1 and sigma of 38.3. Yet the empirical rule still is close at 59%, 98% and 100% for the three intervals. Absolutely not normal.

Why the Emperical Rule works

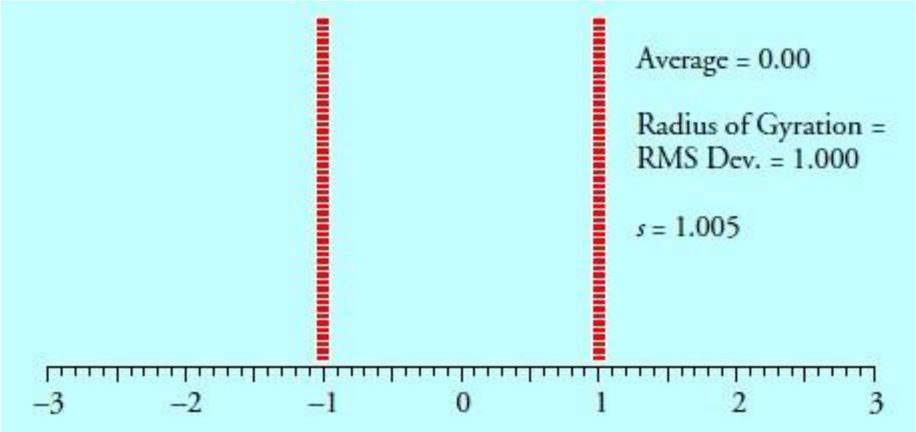
- You don’t need to make the data “fit” a specific probability model
  - While occasionally distributions violate rule 1, they almost always follow rules 2 and 3
- This is because of what the statistics represent
  - The mean or x-bar can be thought of as the center of mass of the data. The balance point at which the weight is equal on either side
  - The standard deviation is effectively the square root of the rotational inertia of the histogram. SAY WHAT?

Rotational Inertia



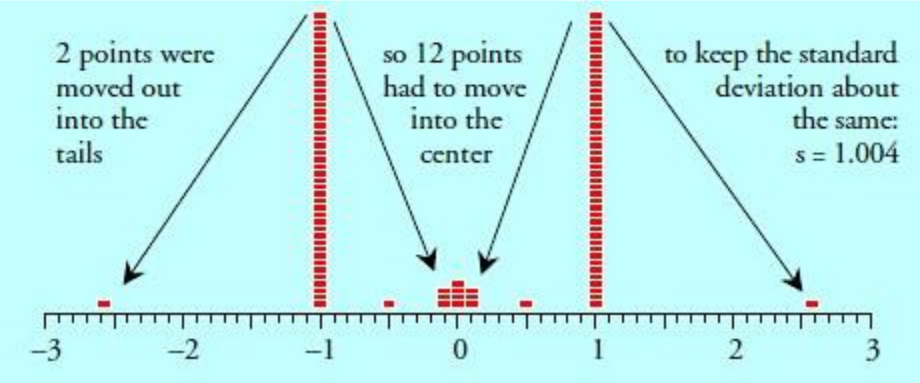
Think about the two histograms on the left as spinning about the axis of the mean. They both have the same mass, but the one on the left will spin with more inertia due the distribution of the mass (think standard deviation)

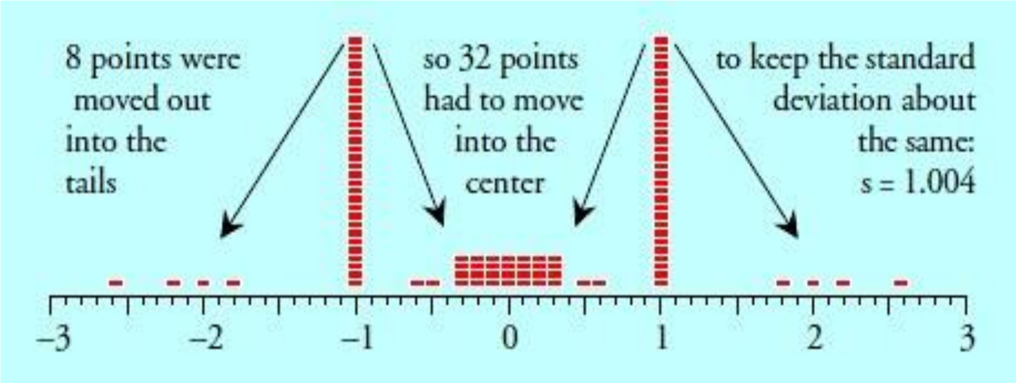
Deriving the Emperical Rule mathematically

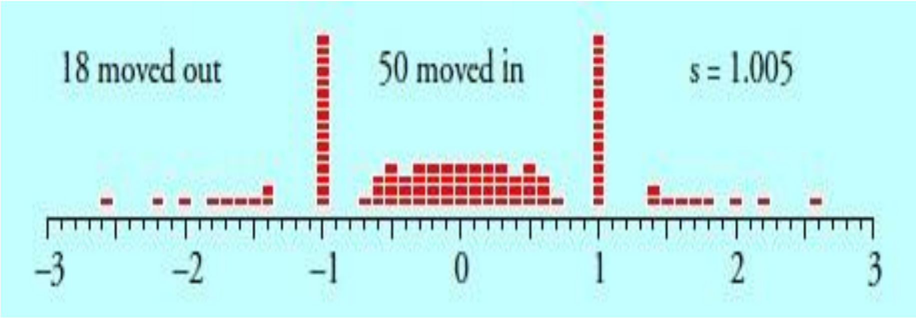


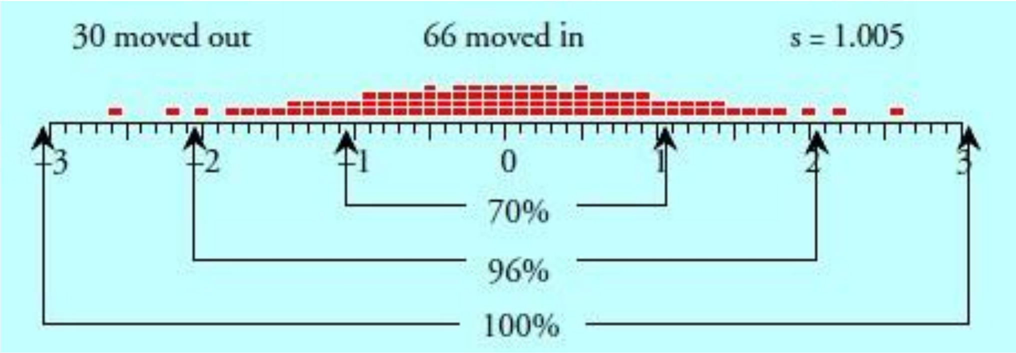
Dr. Wheeler starts with a theoretical distribution of 100 values, 50 at 1 and 50 at minus 1. Therefore the mean is 0 and the global standard deviation is $\sqrt{100/99}$ =1.005. He then procedes to demonstrate that if you move points out to the extremes, a significantly larger amounts must be moved towards the center to maintain the standard deviation of 1.005

Progression of the example





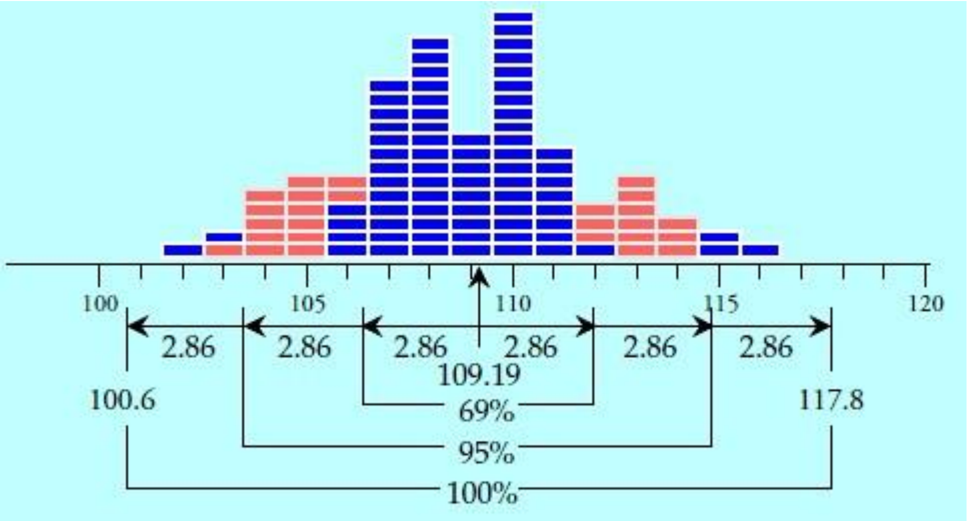


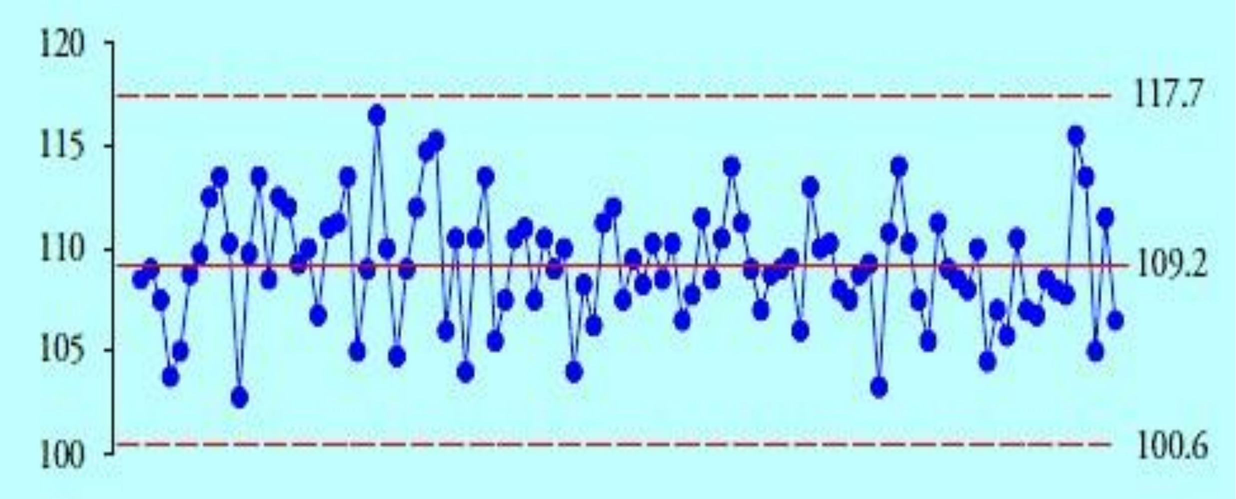


Comparing global standard deviation with control chart sigma

- One thing the histograms DON'T tell you is anything about the time ordered values of the data.
- While the overall standard deviation may describe a large group of numbers, the within subgroup standard deviation is much better indicator of whether a process is stable or not. In all of the cases shown above, we will not compare the histogram with the process chart to see the differences.
- Regardless of what you may have heard or read, Walter Shewhart developed the chart to be used <u>without regard to the parent distribution</u>.

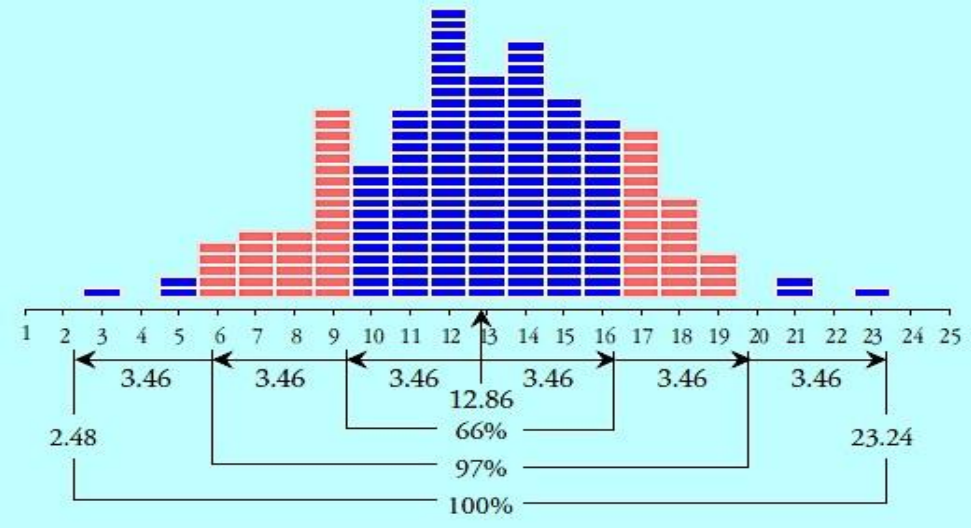
Wire Length Data

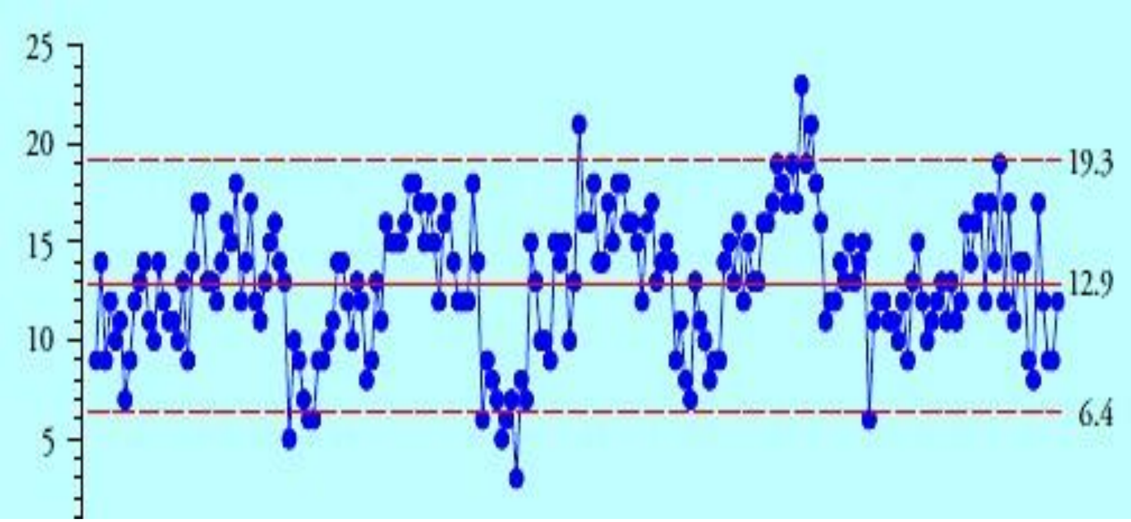




Note when the process is in control the global standard deviation and the within group sigma are very similar. This obvious came out of a stable process and is thus predictable

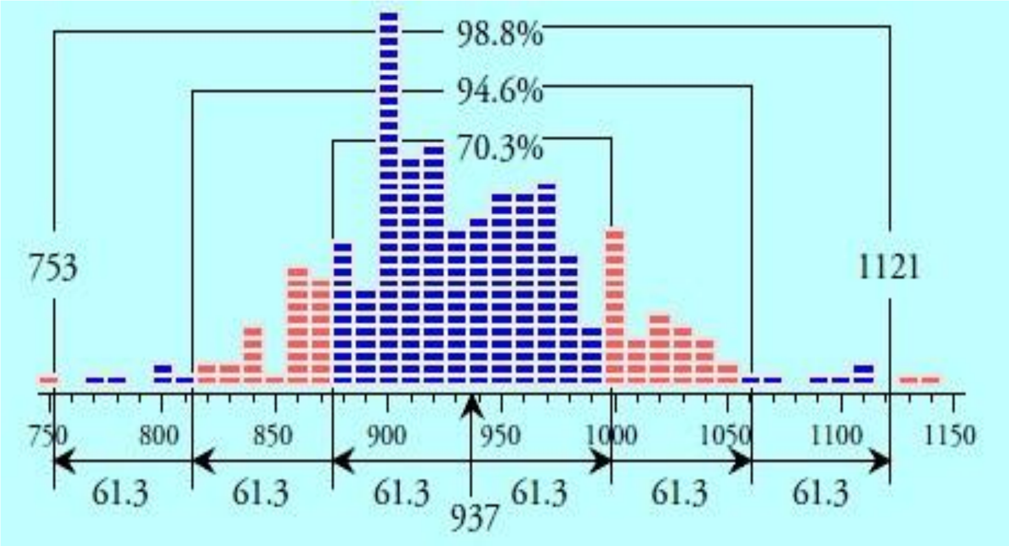
Bead Board Example

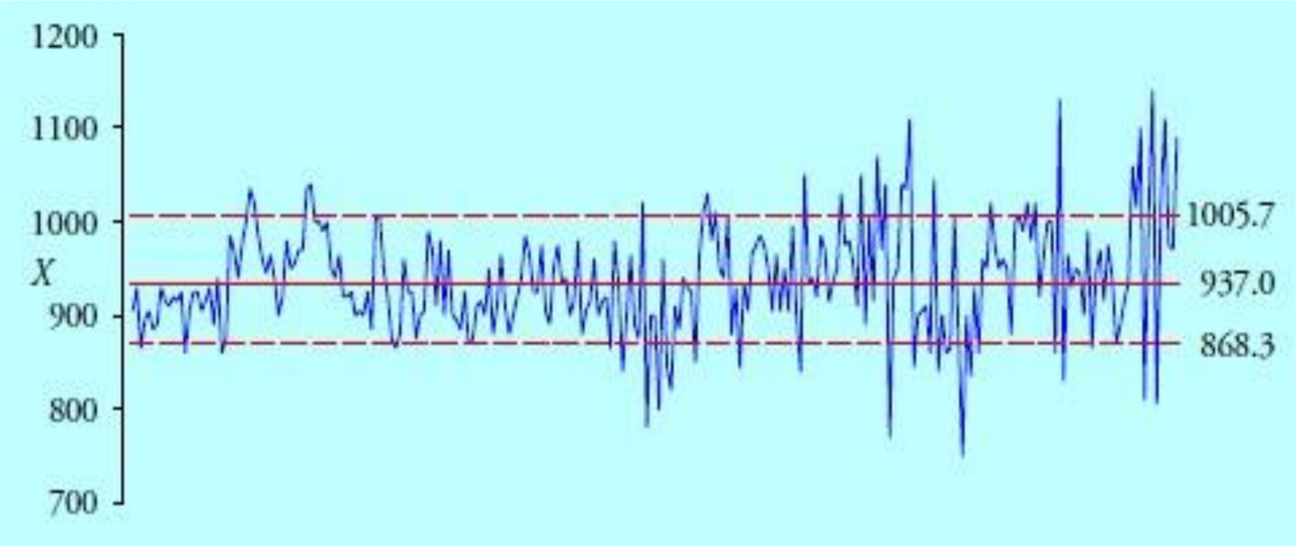




Here is a process where variation has been introduced. Note the out of control points on the chart. The within group sigma is 2.16 while the overall is 3.46. The low kurtosis or spread of the data is due to points out of control in the process.

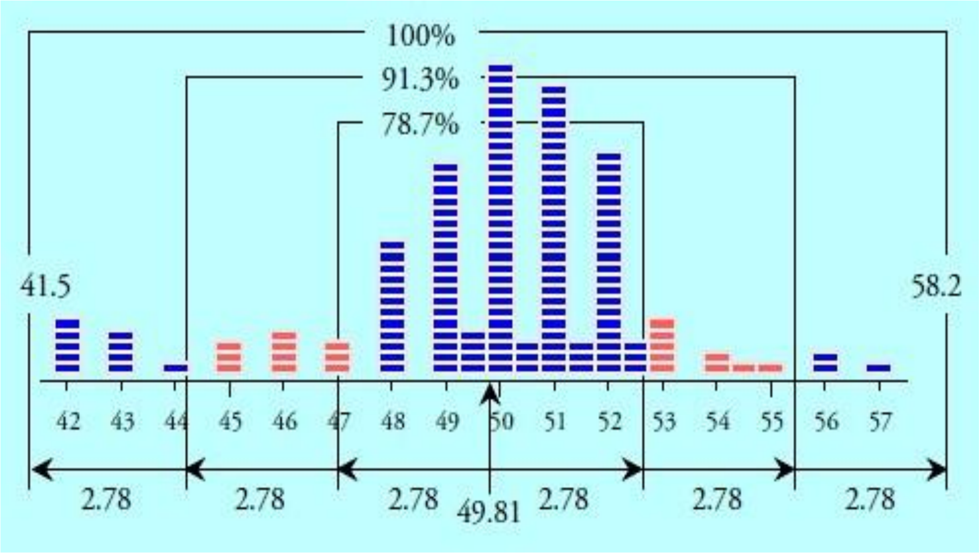
Batch Weights

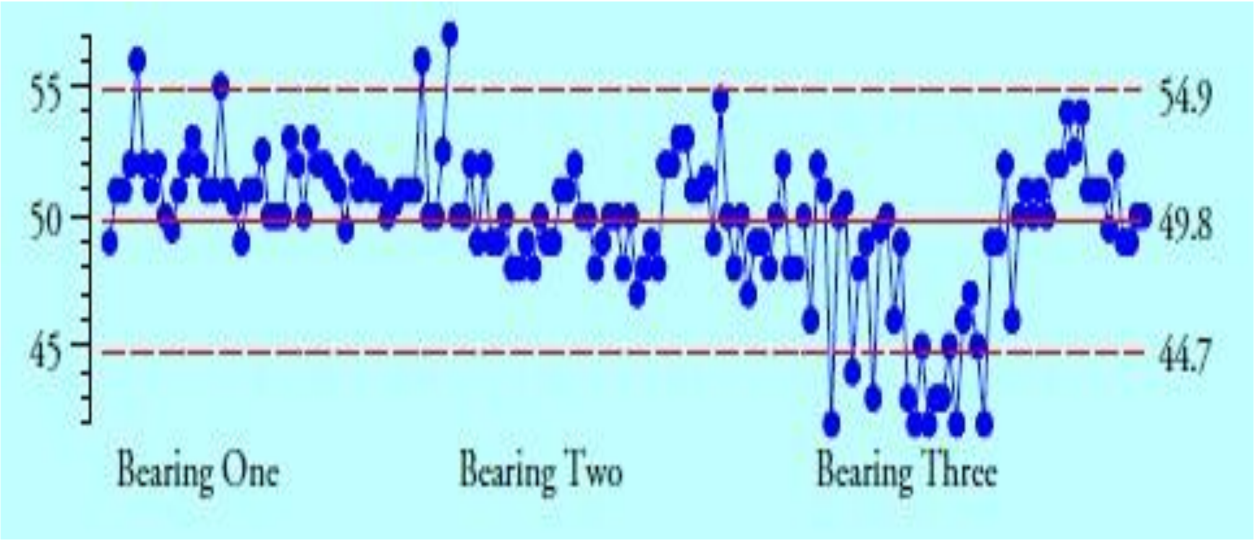




In this example almost 10% of the points in the control chart are out of control. The within group sigma is almost 1/3 of the overall. Also note that the process began close to a stable state and then began erratic behavior about halfway through the chart. Here is a classic example of what would be classified as a non-normal distribution that is actually a process run amok.

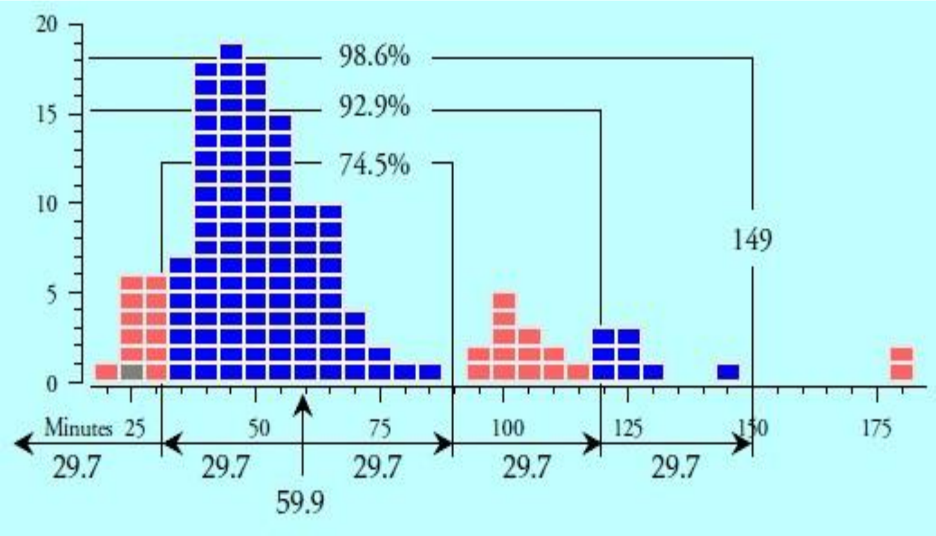
Cam shaft bearing diameters

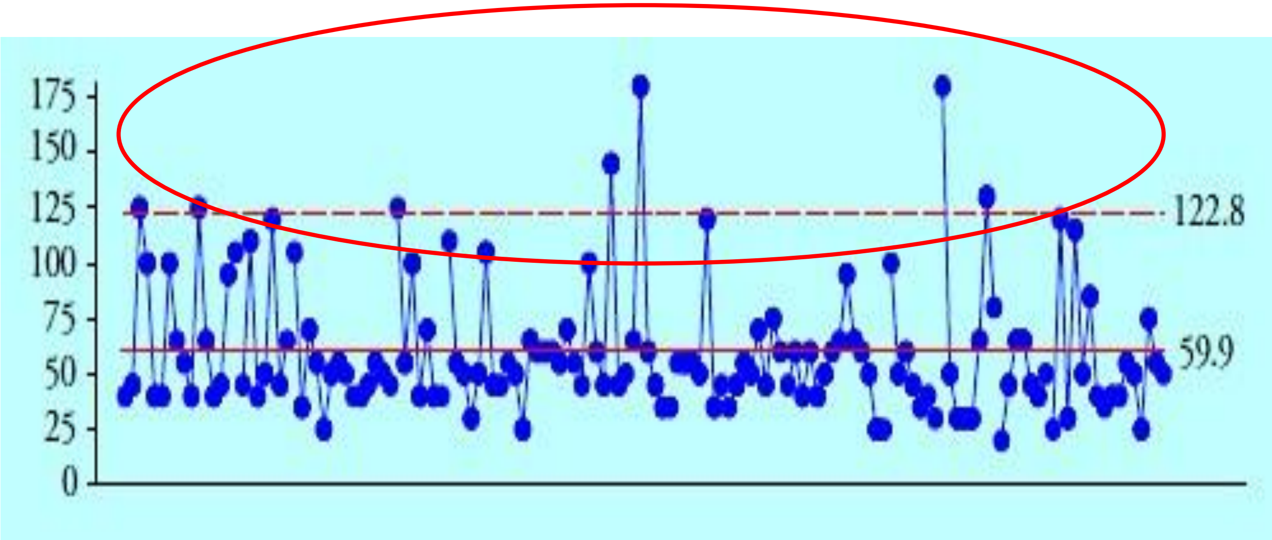




One of things you might suspect looking at this histogram is that it is a mixed process stream. As you can see by the chart combining the three bearing diameters into one distribution will never allow you to solve the problems associated with this. Again knowing where the data came from is vital to understanding the process.

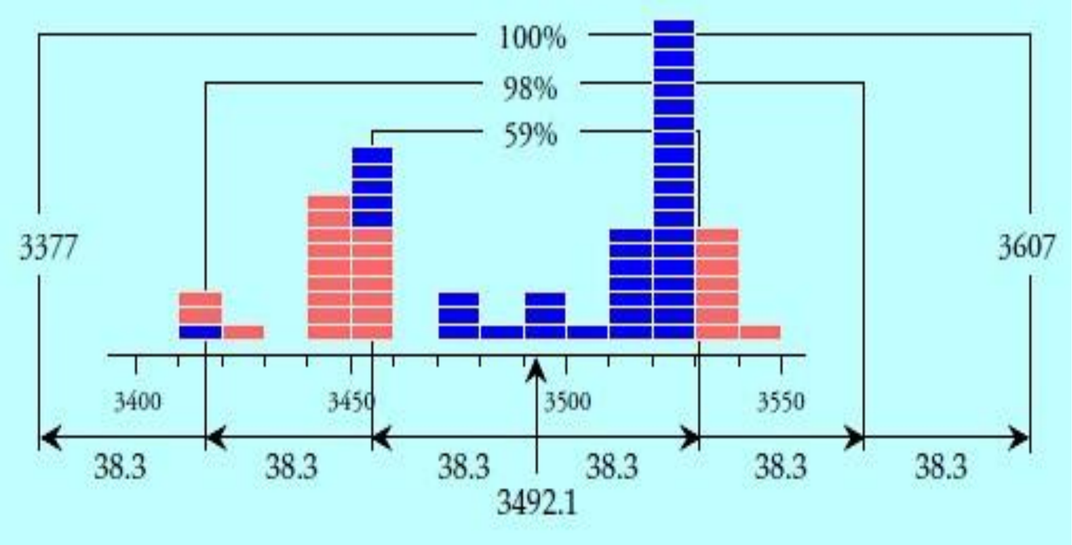
Hot metal delivery time

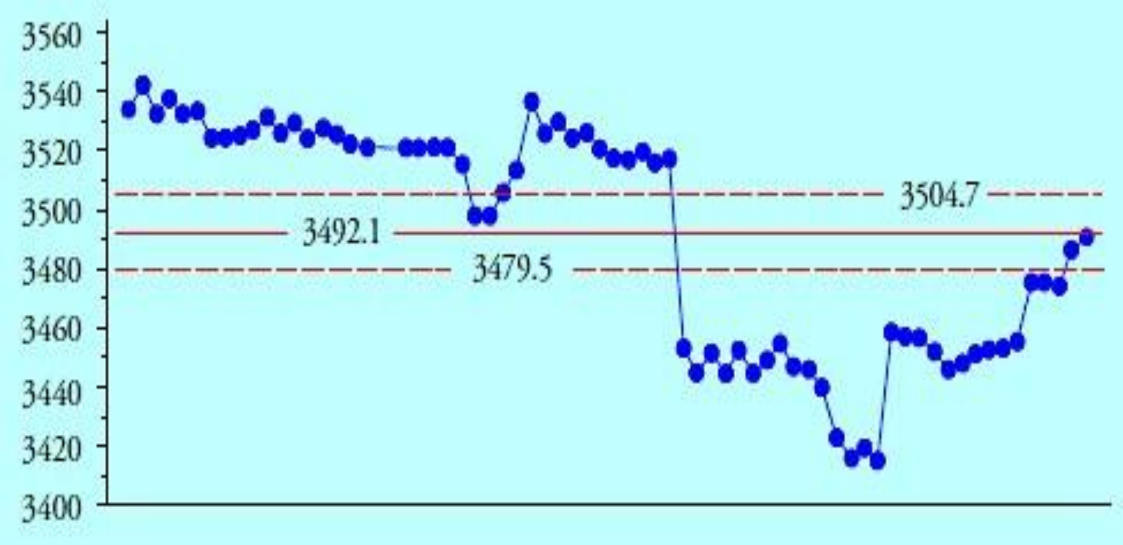




Almost a classic bi-modal distribution. Understanding the circled out of control and near limit points and resolving those issues would make this process predictable.

Last, Creel Yield data





A wandering process. A significant amount of work will need to be done to understand the wandering. Process instructions not followed, no control of process inputs, etc. We could absolutely not make predictions about where this process will be in the future.

Conclusions

- While the empirical rule helps us understand the histogram at a high level, computing the average and global standard deviation is the absolute limit of the information available. The within subgroup sigma is needed to troubleshoot the process and make it predictable.
- Note that 3 of the 5 processes showed 100% within the limits of the empirical rule yet were not from a stable process.
- Remember your data is not generated by some probability distribution, it is generated from a process. Thus, understanding that process is the key to creating predictable outcomes.