DMAIC Training

Statistics Primer

QAI

Descriptive vs Inferential Statistics

- **Descriptive Statistics:**
  - Can be defined as those methods involving collection, presentation and characterization of a set of data in order to describe the various features of that data set appropriately
  - Example of descriptive statistics would be the average, median, mode, quartiles etc. of given data set
- **Inferential Statistics:**
  - These are the methods offered by statistics that make possible the estimation of a characteristic of a population or the making of a decision concerning a population based only on sample data

Terminology to familiarize

- **Population**
  - a complete set of data "N"
- **Sample**
  - a subset of data representing the population "n"
- **Mean**
  - the average of the population or sample set
  - Mean of a sample is denoted by X or μ
- **Variance**
  - the (corrected) mean of the squared deviations (It represents the spread of the data)
- **Standard Deviation**
  - is the (positive) square root of the variance
  - Standard Deviation of a sample is denoted by s or σ
- **Median**
  - The mid point
- **Range**
  - The difference between the maximum and the minimum
- **Mode**
  - The most frequently occurring data value
- **Histogram**
  - A frequency distribution for data values

Basic Statistics

- Average/Mean
- Median
- Mode
- Range: Minimum/ Maximum
- Sample Inter-quartile Range
- Sample Variance
- Standard Deviation
- Coefficient of Variation
- Percentile
- Box Plot
- Histogram
- Normal Standard Deviation

What to Describe?

- **What is the "location" or "center" of the Data? ("**measures of location**")**

- **How does the Data vary? ("**measures of variability**")**

Measures of Location

Mean

Median

Mode

Mean

- Another name for average
- If describing a population, denoted as μ, the Greek letter "mu"
- If describing a sample, called "x-bar"
- Appropriate for describing measurement data
- Seriously affected by unusual values called "outliers"

Calculating Sample Mean

Formula:

$$\overline{X} = \frac{\sum X_i}{n}$$

That is, add up all of the data points and divide by the number of data points

Data (of process defects): **2 8 3 4 1**

Sample Mean = (2 + 8 + 3 + 4 + 1) / 5 = 3.6

Do not round…Mean need not be a whole number

Median

- Another name for 50th percentile
- Appropriate for describing measurement Data
- "Robust to outliers", that is, not affected much by unusual values

Calculating Sample Median

- **Order data from smallest to largest:**

If **odd** number of data points, the median is the middle value

Data (# process defects): **2 8 3 4 1**

Ordered Data: **1 2 3 4 8**

Median

QAI

Calculating Sample Median

Order data from smallest to largest

If **even** number of data points, the median is the average of the two middle values

Data (# of process defects): **2 8 3 4 1 8**

Ordered Data: **1 2 3 4 8 8**

Median = (3 + 4) /2 = 3.5

Mode

- The value that occurs most frequently
- One data set can have many modes
- Appropriate for all types of data, but most useful for categorical data or discrete data with only a few number of possible values
- 2,3,3,4,5,2,3,4,5,3,4,3,3,4,5,2,
- 2,2,2,**3,3,3,3,3,3**,4,4,4,4,5,5,5

Measures of Variability

- Range
- Interquartile Range
- Variance and Standard Deviation
- Coefficient of Variation

All of these measures are appropriate for measurement data only

Range

- The difference between largest and smallest data point
- Highly affected by outliers
- Best for symmetric data with no outliers

R = largest obs. - smallest obs.

or, equivalently

$R = x_{max} - x_{min}$

QAI

Interquartile Range

- The difference between the "third quartile" (75th percentile) and the "first quartile" (25th percentile). So, the "middle-half" of the values
- IQR = Q3-Q1
- Robust to outliers or extreme observations
- Works well for skewed data

QAI

What is Standard Deviation?

- It is the typical (standard) difference (deviation) of an observation from the mean

- Think of it as the average distance a data point is from the mean.

NOTE : Standard deviation is an important measure from Six Sigma perspective and will be used in many examples as well as calculations.

Sample Variance

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}$$

xi - is individual data
x – is mean value of the sample
n – is the number of samples

QAI

Sample Standard Deviation

- **Formula for Std Deviation**

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

QAI

Measures of Variation - Some Comments

- Range is the simplest, but is very sensitive to outliers
- Variance units are the square of the original units
- Interquartile range is mainly used with skewed data (or data with outliers)
- We will use the standard deviation as a measure of variation often in this course

How to Calculate Percentile

Percentile value = Total number of readings * (decided percentile) + (decided percentile)

For example, the median is the 50th percentile

We want to determine the median of a dataset of 40 nos.

0.5 * 40 + 0.5 = 20.5

Thus the median will be the reading between the 20^{th} and 21^{st} reading

The third quartile will be 0.75 * 40 + 0.75 = 30.75

That is the extrapolated reading between the 30^{th} and the 31^{st} reading

Boxplot - a 5 Number Summary

- Smallest Observation (Min)
- Q_1
- Q_2 (median)
- Q_3
- Largest Observation (Max)

Boxplot Example

- Smallest observation = 3.20
- Q_1 = 43.645
- Q_2 (median) = 60.345
- Q_3 = 84.96
- Largest observation = 124.27

Creating a Boxplot

- Create a scale covering the smallest to largest values
- Mark the location of the five numbers
- Draw a rectangle beginning at Q1 and ending at Q3
- Draw a line in the box representing Q2, the median
- Draw lines from the ends of the box to the smallest and largest values
- Some software packages that create boxplots include an algorithm to detect outliers. They will plot points considered to be outliers individually

Boxplot Example

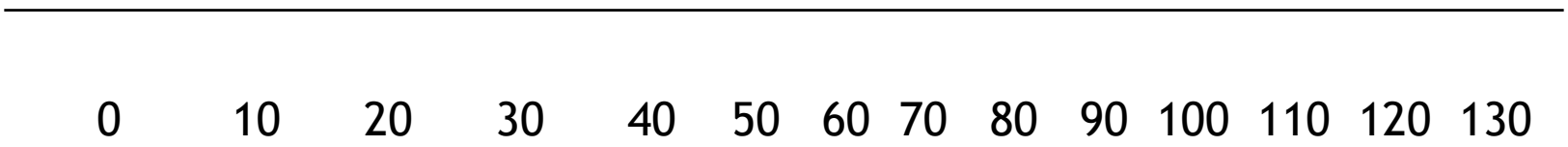
Min = 3.20

Q_2 = 43.35

Max = 124.27

Q_1 = 33.645

Q_3 = 84.96



QAI

Boxplot Interpretation

- The box represents the middle 50% of the data, i.e., IQR = length of box
- The difference of the ends of the whiskers is the range (if there are no outliers)
- Outliers are marked by an * by most software packages
- Boxplots are useful for comparing two or more samples
  - Compare center (median line)
  - Compare variation (length of box or whiskers)

Histograms

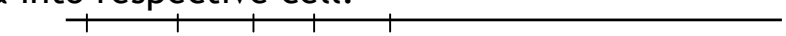
- Building Histograms

More Than	Up to	Cell Size
0	1	1
1	2	2
2	4	3
4	8	4
8	16	5
16	32	6
32	64	7
64	128	8
128	256	9
256	512	10

Based on number of data points, select # cells from table

- Determine cell interval.
- Building the limit histogram axis.
- Building histogram by placing data into respective cell.

$$\frac{\text{Max - Min}}{\text{\# Cells}}$$



Lowest Limit = Min - (0.5) x Cell Size

QAI

Histograms (cont)

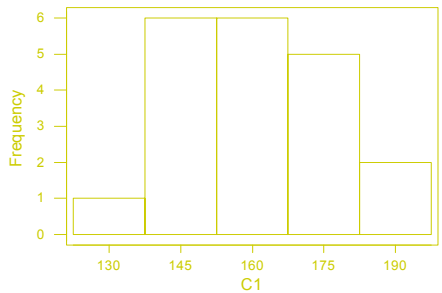
- **Guideline for Forming the Cell Intervals:**
  - Use intervals of equal length with midpoints at convenient numbers
  - For small data set, use a small number of intervals
  - For a large data set, use more intervals
  - Make sure that a data point can only be located in one cell
- **Histograms:**
  - What: Graph of frequency of occurrence of an event
  - Why: Used to breakdown the distribution by number of occurrences per event to further define the data
  - How: For each event, or group of data, plot number of times the event occurs. If needed, add cumulative percent. From left to right add sequentially each event percentage

Histogram: Example

- A Histogram is a bar graph that shows the number (relative frequency) of each data point within a cell or interval. Each bar covers the interval and is centered at the mid point. Using minitab and the following data, create a histogram.

Sample	Value
1	140
2	145
3	160
4	140
5	155
6	165
7	150
8	190
9	138
10	155

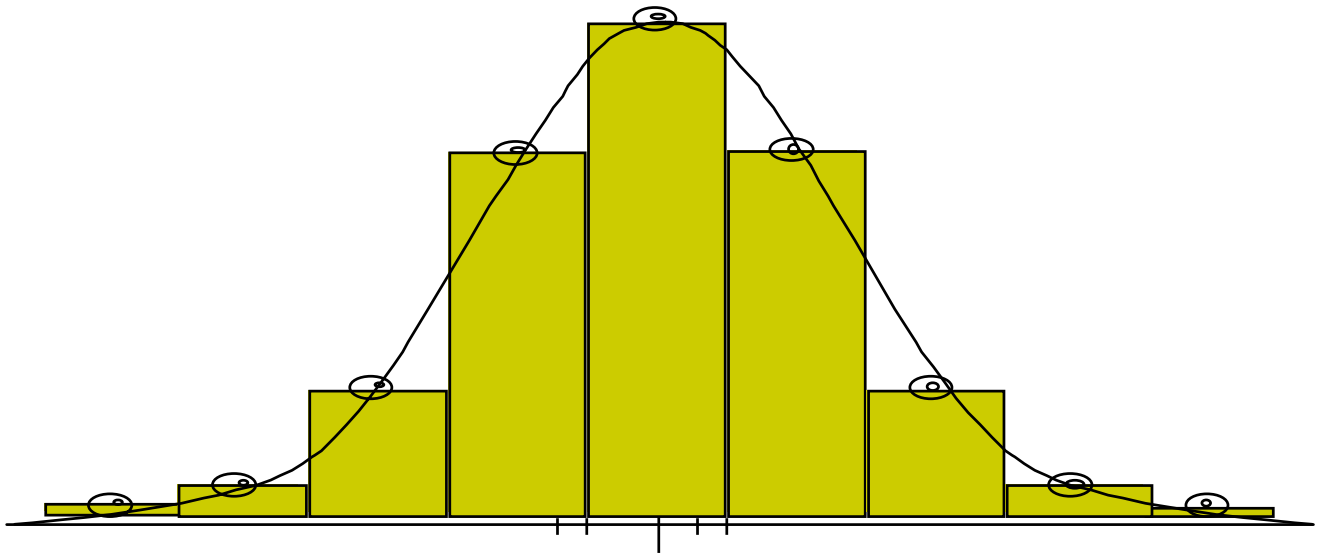
Sample	Value
11	153
12	145
13	170
14	175
15	175
16	180
17	135
18	157
19	170
20	185



QAI

Normal Curve

Smooth Line Connects the Center of Every Bar of the Histogram



The Characteristic of a normal curve is that Mean, median and the mode are the same.

QAI

Probability Distribution of Continuous Random Variable:

- **Properties of a Normal distribution:**
  - It is a bell shaped curve that is symmetrical
  - Its measure of central tendency (mean, median, mode) are all identical
  - The probability that a point is x distance below the average is the same as that of a point to be x distance greater than the average
  - Measures that we expect to find normally distributed would be absenteeism, abandon rate, service level

Features of a Normal Standard Distribution

- 100% of the area under the normal curve lies between ± infinity, we may calculate that area which lies beyond the specification limit. Both above the USL (Upper Specification Limit) or / and below the Lower Specification Limit (LSL) Doing so would reveal the random chance probability of creating a defect

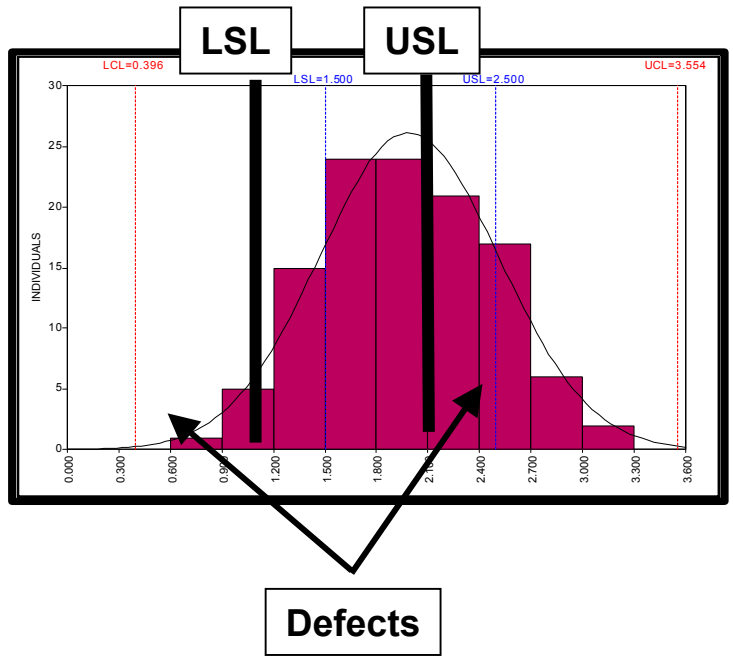
In order to use the normal table to

Accommodate any normal distribution, we need only to standardize our data:

$Z = (X - \mu) / \sigma$

The Z score, denoted by "Z", corresponds to the value at X, and is the

distance between X and the mean (μ) in units of the standard deviation



QAI