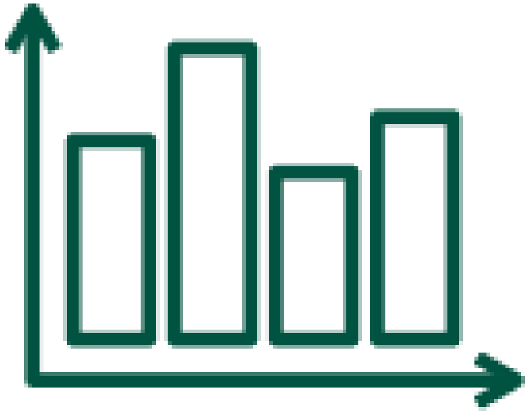
Food Safety and Inspection Service
U.S. DEPARTMENT OF AGRICULTURE

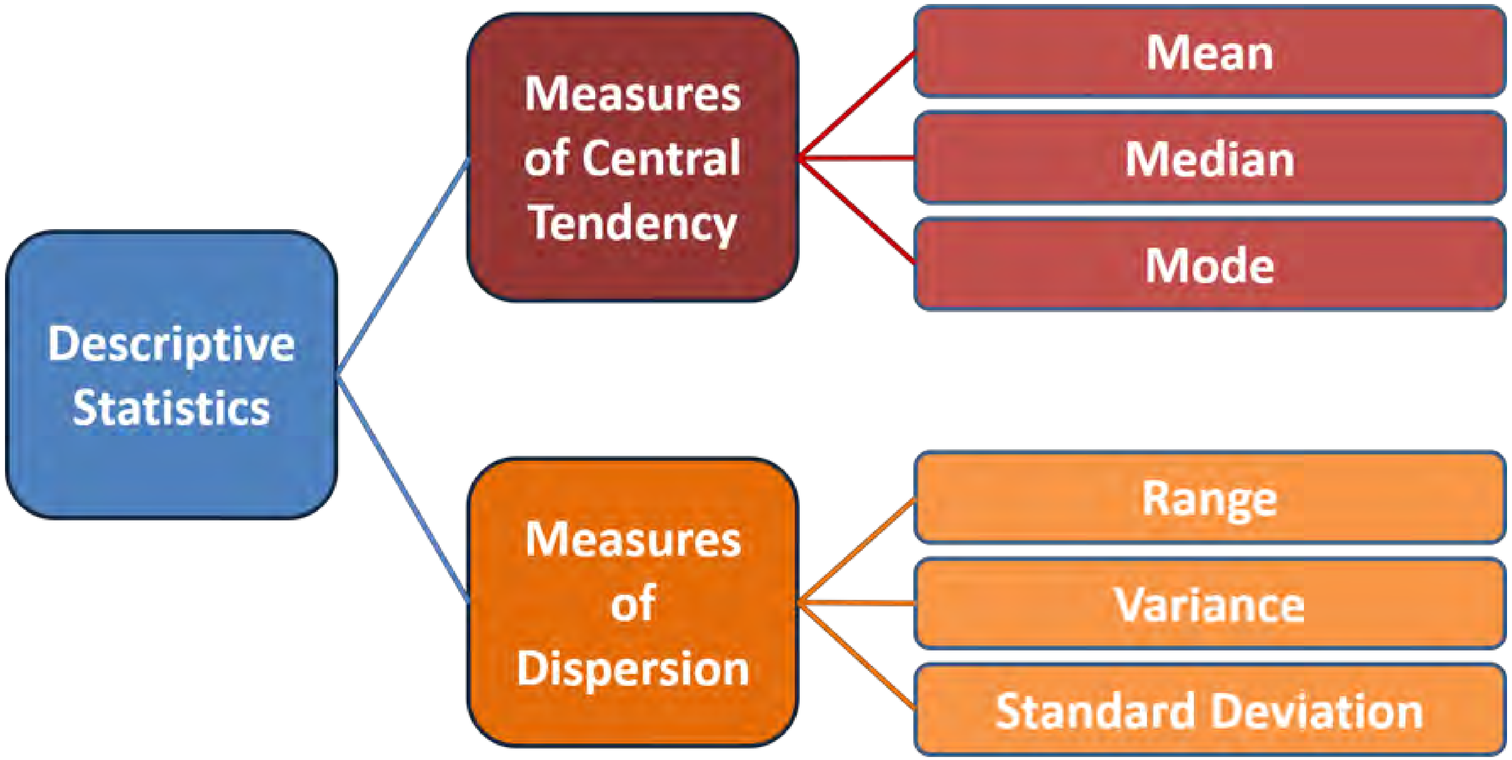
Statistics and their role in evaluating an Establishments process control procedures

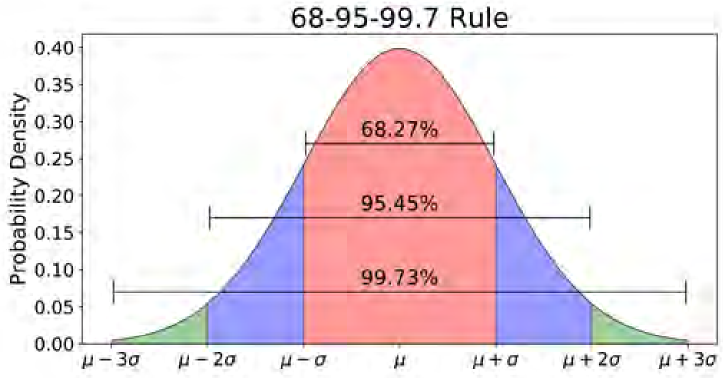


Topics

- Basic Tools/Measurements
- Introduction
- Measurements
- Normal Distribution
- Process Control Charts

Basic Tools and Measurements





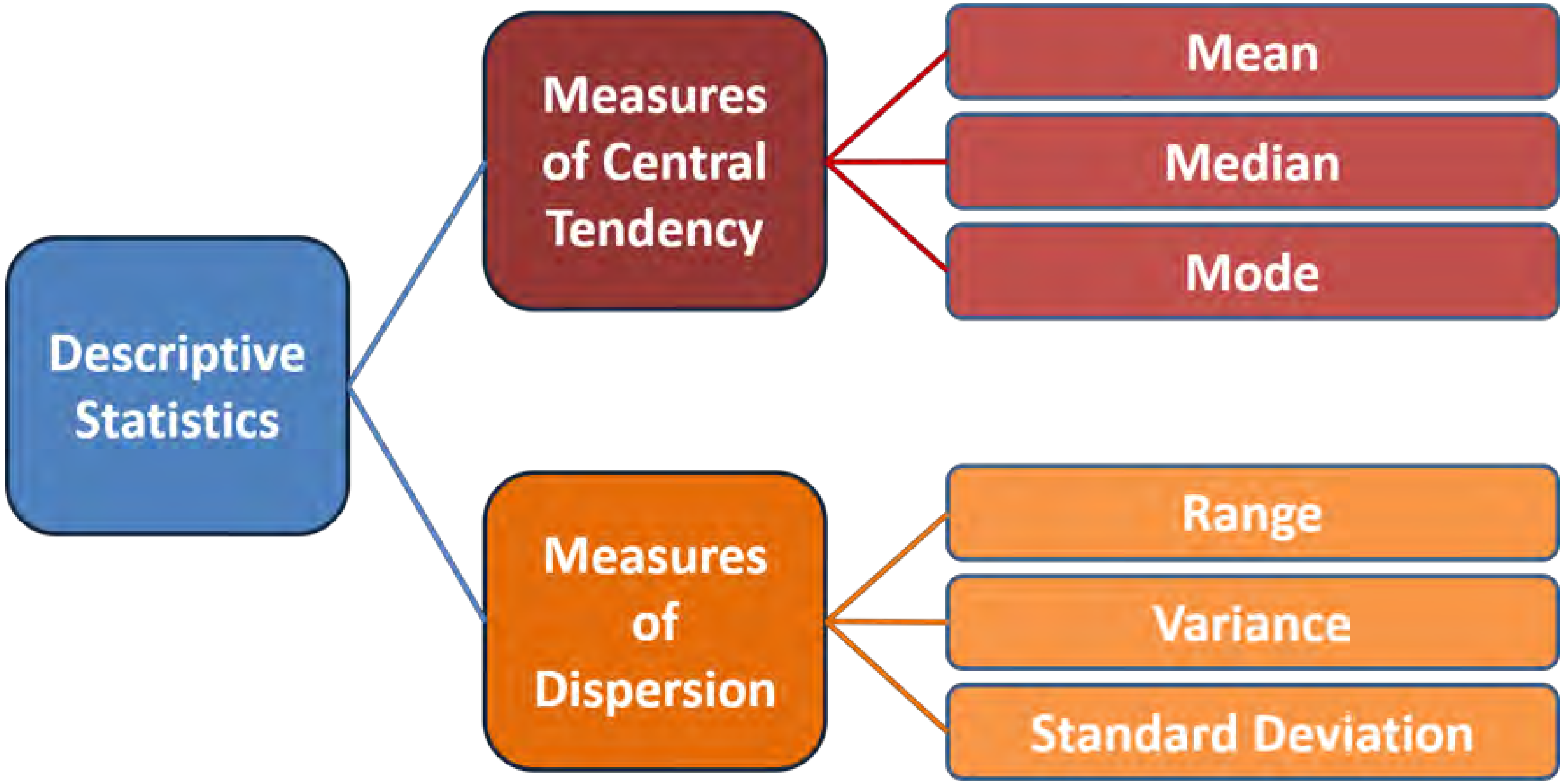
$$\log_{10} x$$

Statistics

Statistics at its core is a way to understand data

Suppose we have a rather large data set (Ex: 1,000 values). Is there a way we can describe this dataset without giving all these points?

Descriptive statistics are measures that quantitatively describes or summarizes a dataset



Statistics 1

- A measure of central tendency provides information about the typical value or center point of a dataset

- These measures are calculated primarily for comparison purposes
  - Which class did better on the quiz?
  - Did production line A produce more final product then line B?

- The most used measures of central tendency are Mean, Median, and Mode

Order of Operations

Parentheses
(and other grouping symbols)

Exponents
(and roots)

Multiplication or Division

Addition or Subtraction

- Please
- Excuse
- My Dear
- Aunt Sally

$(2+3-4)^2 2\times6\div2$ What did you get??

Order of Operations 2

$(2+3-4)^2 \times 6 \div 2 = ??$

PEMDAS

P $(2+3-4) = 1$

E $(1)^2 = 1$

M $1 \times 6 = 6$

D $6 \div 2 = 3$

3 is the answer!

Mean

- Mean (often called the Average)
- The sum of the values divided by the number of values
- The most frequently used measure of these three
- Ex: We have seven values of
  5, 4, 5, 6, 7, 5, 3. Then, the mean is

  (5+4+5+6+7+5+3)/7=35/7=**5**

Median

- The middle value of the dataset after ordering the dataset by size
- Splits the data into two equally-sized groups

- Half of the data is larger than this value, and half of the data is smaller than this value

Ex: We have seven values of
5, 4, 5, 6, 7, 5, 3. Then, the median is

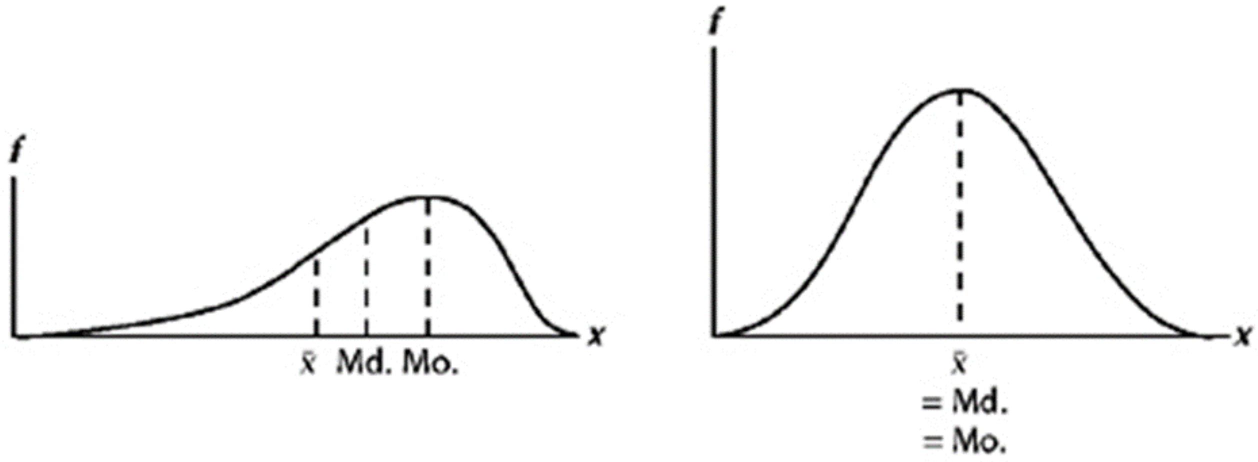
3, 4, 5, **5**, 5, 6, 7

Mode

- The value that occurs most frequently
- Ex: We have seven values of
  5, 4, 5, 6, 7, 5, 3.
- Then, the mode is 5 as it occurs the most

Why do we need different measures?

The shape of the data can be skewed, could have outliers, multiple modes



The data on the right could be represented by either the mean, median or mode, but the skew on the left dataset makes the median a better choice

Why do we need different measures? 2

Ex:

Suppose we had seven values of
5, 4, 5, 6, 7, 5, 17. Then, the mean is

(5+4+5+6+7+5+**17**)/7=**49**/7=**7**

However, the median and the mode of this set are still 5. For this dataset, they would be better indicators of the central tendency than the mean.
Here, 17 is what we call an outlier

Population

A population is the entire group you would like to draw conclusions about. Suppose you want to know the average male height in the U.S. Then, your population is all males in the U.S.

We will likely never know the true population mean:
Impractical to measure the whole population
For the example above, the size would be roughly 150 million observations! (This would be our N)

The Population

Size = N

Mean = ?

Random Sample

- Instead, we can take a (random) sample that we think represents the population
- For the example, we could measure the height of the first 1000 (n) males from different malls in major cities
- A bad sample would be taking the height of 1000 basketball players
- Thus, we can make the distinction between a population measure and a sample measure
- Population mean vs Sample mean

Population Mean

$$\mu = \frac{x_1 + x_2 + \ \dots + x_N}{N}$$

μ is the population mean for the population that is of size N

Sample Mean:

$$\bar{x} = \frac{x_1 + x_2 + \ \dots + x_n}{n}$$

$\bar{x}$ is the sample mean for the sample that is of size n

Calculate Mean, Median, and Mode for the two Data Columns

Post-Hide Removal (PH)
Pre-Chill (PC)

	PH	PC
Mean		
Median		
Mode		

Let's take a few minutes so that you can do some practice calculations

Post Hide (PH)	Pre Chill (PC)
20	45
18	45
24	44
19	43
22	48
11	44
23	40
24	42
20	41

Post Hide Removal Data

20
18
24
19
22
11
23
24
20

20		20		11		11
18		18		18		18
24		24		19		19
19		19		20		20
22		22		20		20
11		11		22		22
23		23		23		23
24		24		24		24
20		20		24		24
181		20.11111		20		20
Sum		Mean		Mode		Median

Post Chill Data

45
45
44
43
48
44
40
42
41

45		45		40		40
45		45		41		41
44		44		42		42
43		43		43		43
48		48		44		44
44		44		44		44
40		40		45		45
42		42		45		45
41		41		48		48
392		43.55556		44		44
Sum		Mean		Mode		Median

Variance

Measure of how much a dataset is spread out

$$\frac{(x_1-\mu)^2 + (x_2-\mu)^2 + \dots + (x_N-\mu)^2}{N}$$

Ex: We have seven values of 5, 4, 5, 6, 7, 5, 3, and we know the mean is 5. Then, the variance is

([(5−5)]^2 +[(4−5)]^2+[(5−5)]^2+(6−5)^2+(7−5)^2+(5−5)^2+(3−5)^2)/7

=(0+1+0+1+4+0+4)/7=10/7**=~1.43**

Sample Size

- We can use data to determine sampling and sampling size.
- To determine sampling and sample size, we first need a unit of measurement. To find that unit of measurement, we ask: What is the sample?
  - One day's production
  - One hour of production
  - One lot
- What is identified in the HACCP Plan

- Sampling is done within the defined unit

Sample Size 2

Using a percentage of some total production
500 combos of trim are produced in 1 day
If you wanted to sample 3% of a day's production:
$500 \times 0.03 = \mathbf{15}$ combos would need to be sampled

Using the square root of some total production
$\sqrt{500} = \sim 22.36$, so we would sample 23 combos (Round up!)

Normal Distribution

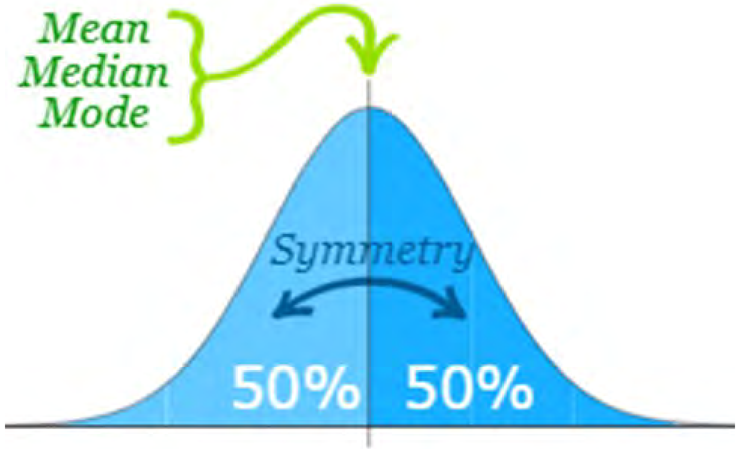
A normal distribution is one where the data is spread symmetrically about the mean

Half of the values are less than the mean

Half of the values are greater than the mean

In this distribution, the mean, median, and mode are all equal

“Bell Curve”



Many things closely follow a normal distribution:

Height/Weight

Blood Pressure

IQ scores

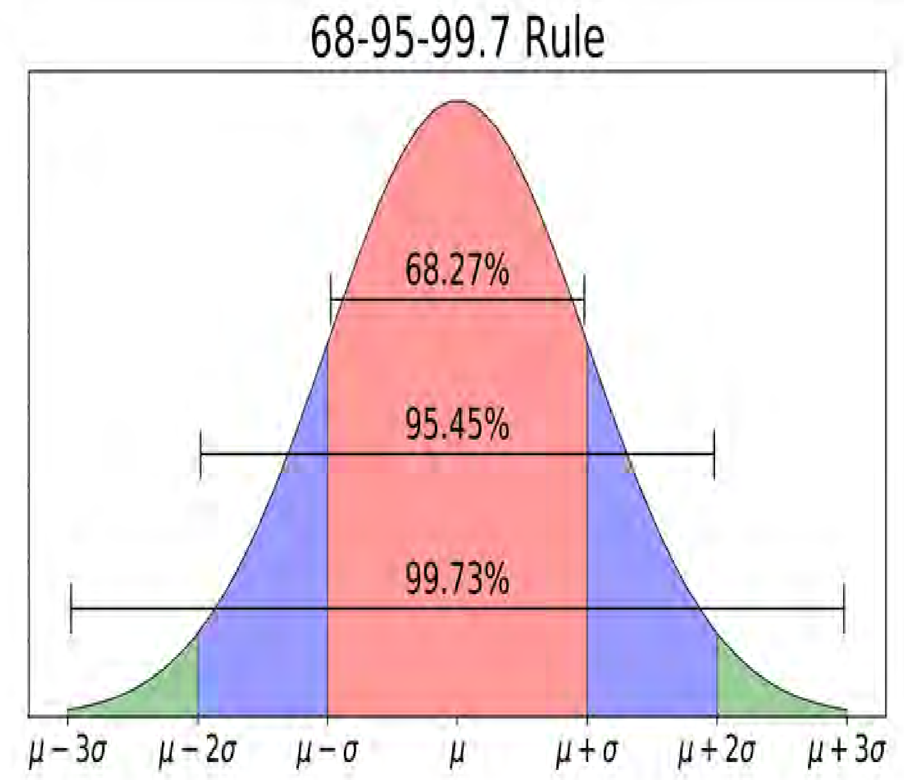
Normal Distribution 2

Properties of the Normal Distribution:

~68% of values are within 1 standard deviation of the mean ($\mu \pm 1\sigma$)

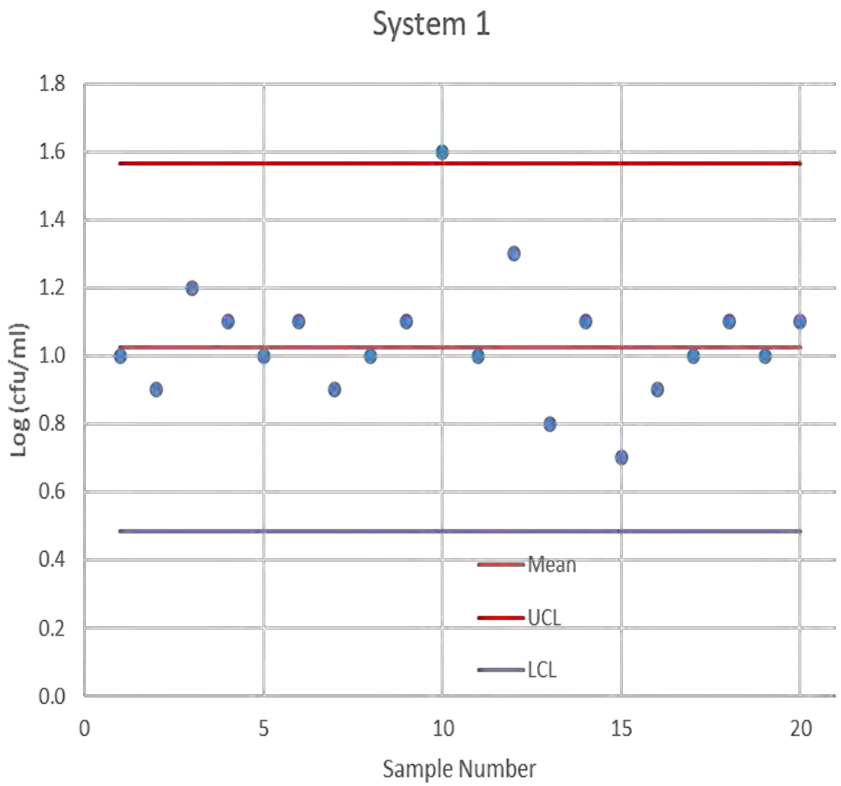
~95% of values are within 2 standard deviations of the mean ($\mu \pm 2\sigma$)

~99.7% of values are within 3 standard deviations of the mean ($\mu \pm 3\sigma$)



Process Control Charts

- Process control charts (Shewhart charts) are a statistical process control tool used to study how a process changes over time
- A control chart will have
  - a central line for the mean
  - an upper control limit (UCL) line
  - a lower control limit (LCL) line
- These values and lines are determined from historical data.



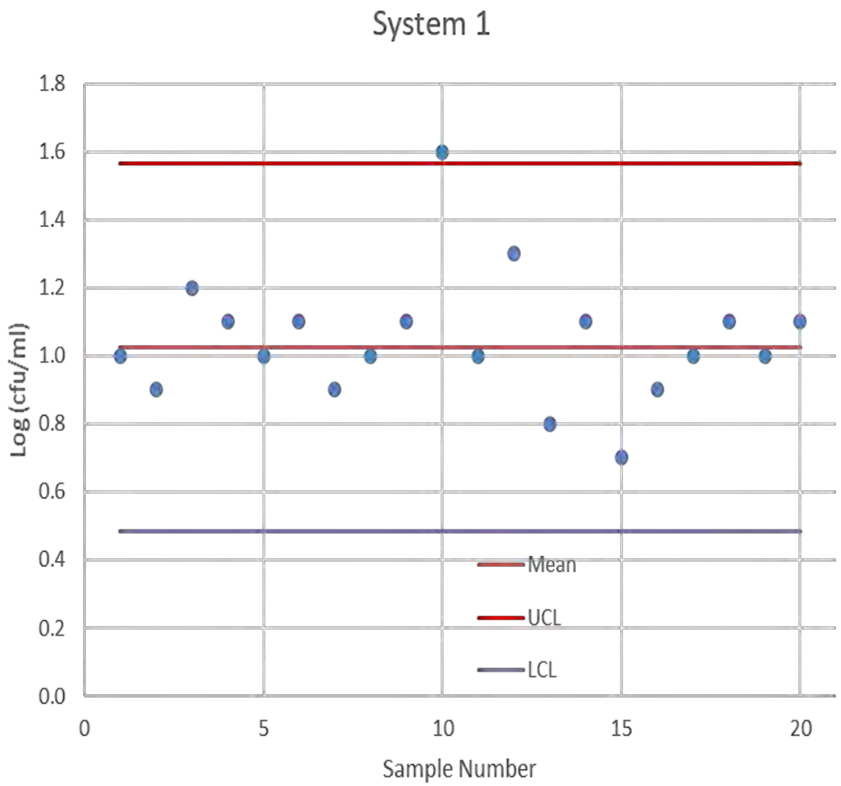
Upper and Lower Control Limits

- Many establishments may already have historical microbiological sampling data that must be used to set the UCL and LCL.
- If an establishment does not have historical sampling data, an establishment may choose to use the values cited in the FSIS baseline study, FSIS data resources, and FSIS guidance documents for its own control limit values until it has sufficient data to conduct its own SPC evaluation.

- Once the establishment collects sufficient data, FSIS baseline data are no longer the sole data source, instead, the establishment must use the data collected from its own sampling programs to conduct SPC analysis.

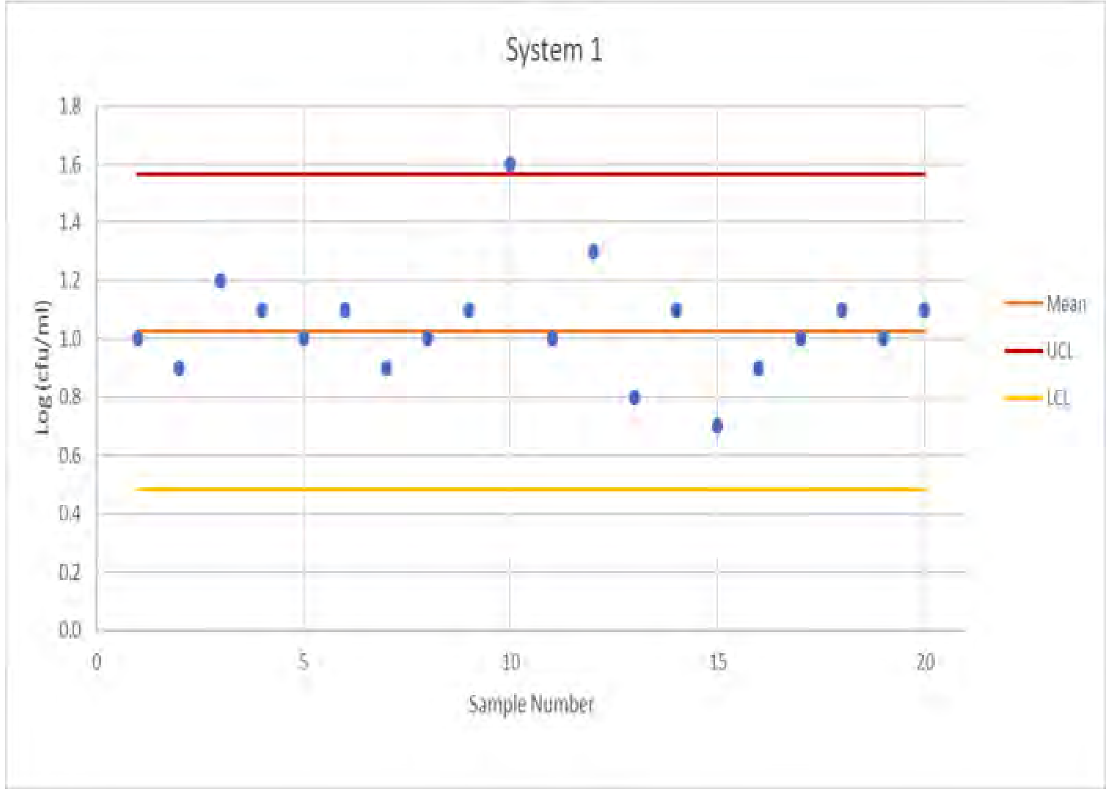
Process Control Charts 2

- You can draw conclusions about whether a process is in control or not by comparing current data to these lines.



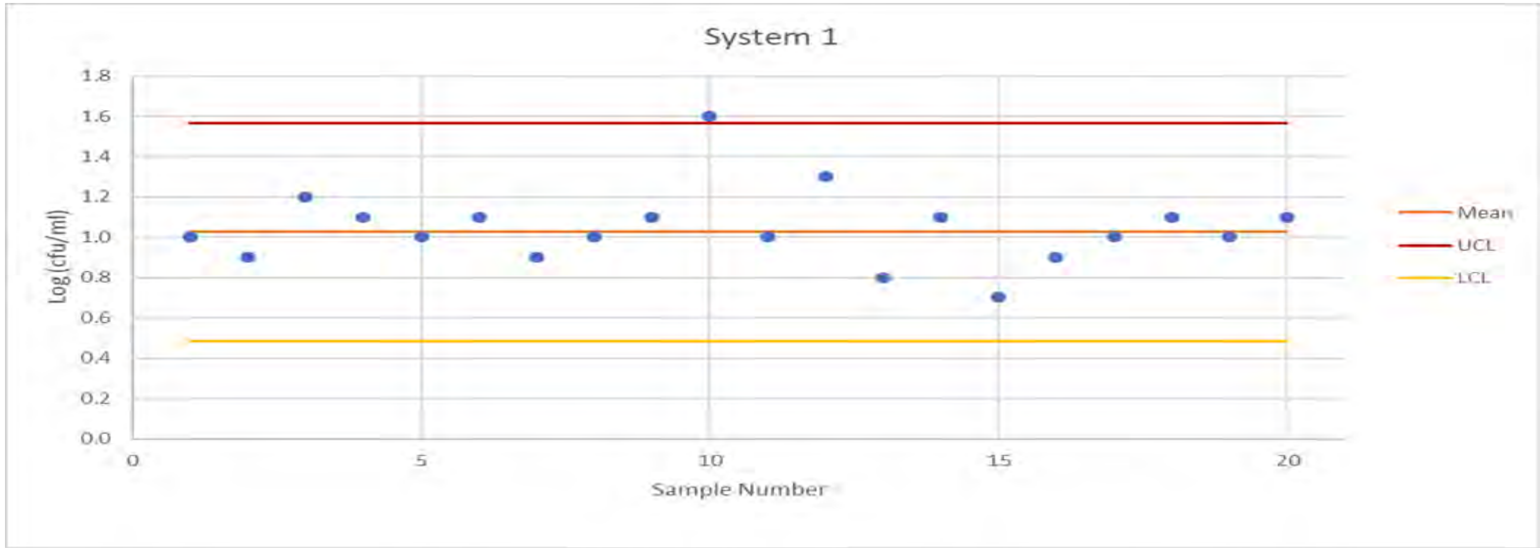
Process Control Charts in Excel

Sample Numbe	Log (cfu/ml	Mean	UCL	LCL
1	1.0	1.03	1.57	0.48
2	0.9	1.03	1.57	0.48
3	1.2	1.03	1.57	0.48
4	1.1	1.03	1.57	0.48
5	1.0	1.03	1.57	0.48
6	1.1	1.03	1.57	0.48
7	0.7	1.03	1.57	0.48
8	0.9	1.03	1.57	0.48
9	1.1	1.03	1.57	0.48
10	1.4	1.03	1.57	0.48
11	1.0	1.03	1.57	0.48
12	1.3	1.03	1.57	0.48
13	0.8	1.03	1.57	0.48
14	1.1	1.03	1.57	0.48
15	0.7	1.03	1.57	0.48
16	0.9	1.03	1.57	0.48
17	1.0	1.03	1.57	0.48
18	1.1	1.03	1.57	0.48
19	1.0	1.03	1.57	0.48
20	1.2	1.03	1.57	0.48

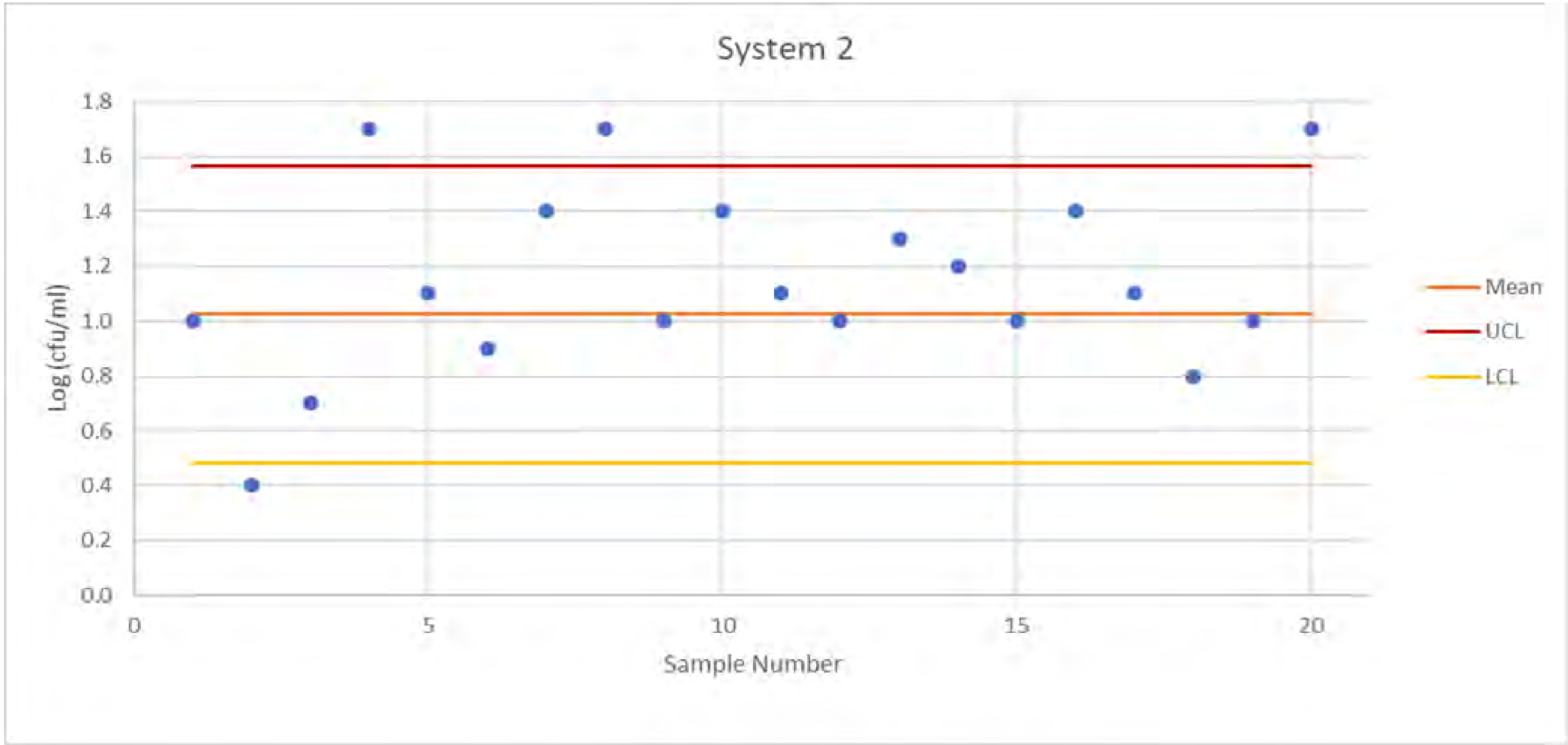


Process Control Charts 3

The control chart for System 1 depicts a pattern of test results that would be seen in a well-controlled system.
In a well-controlled system, the majority of test results will be clustered around a central value.
Note: Even in a well-controlled system, there is some frequency of isolated results above the acceptable level



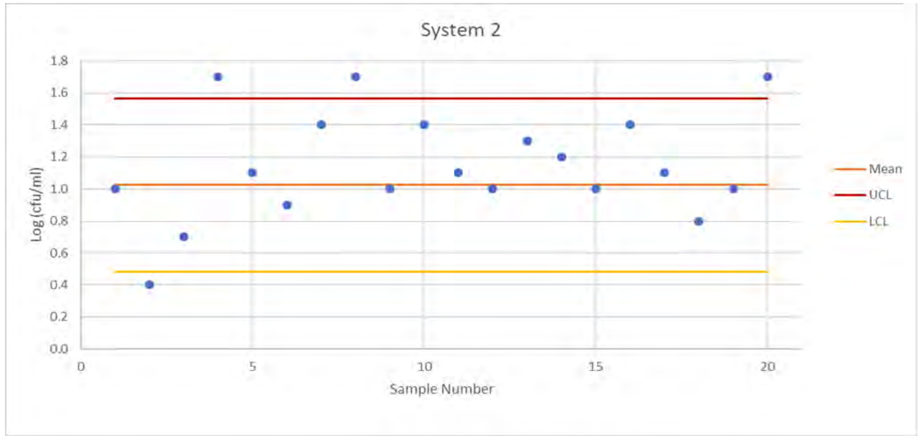
Process Control Charts 4



What are your thoughts regarding this System?

Process Control Charts 5

This chart depicts a loss of process control due to excess variability. This is reflected in both an increased number of results above the UCL, and an increase in scatter points directly below it as well.

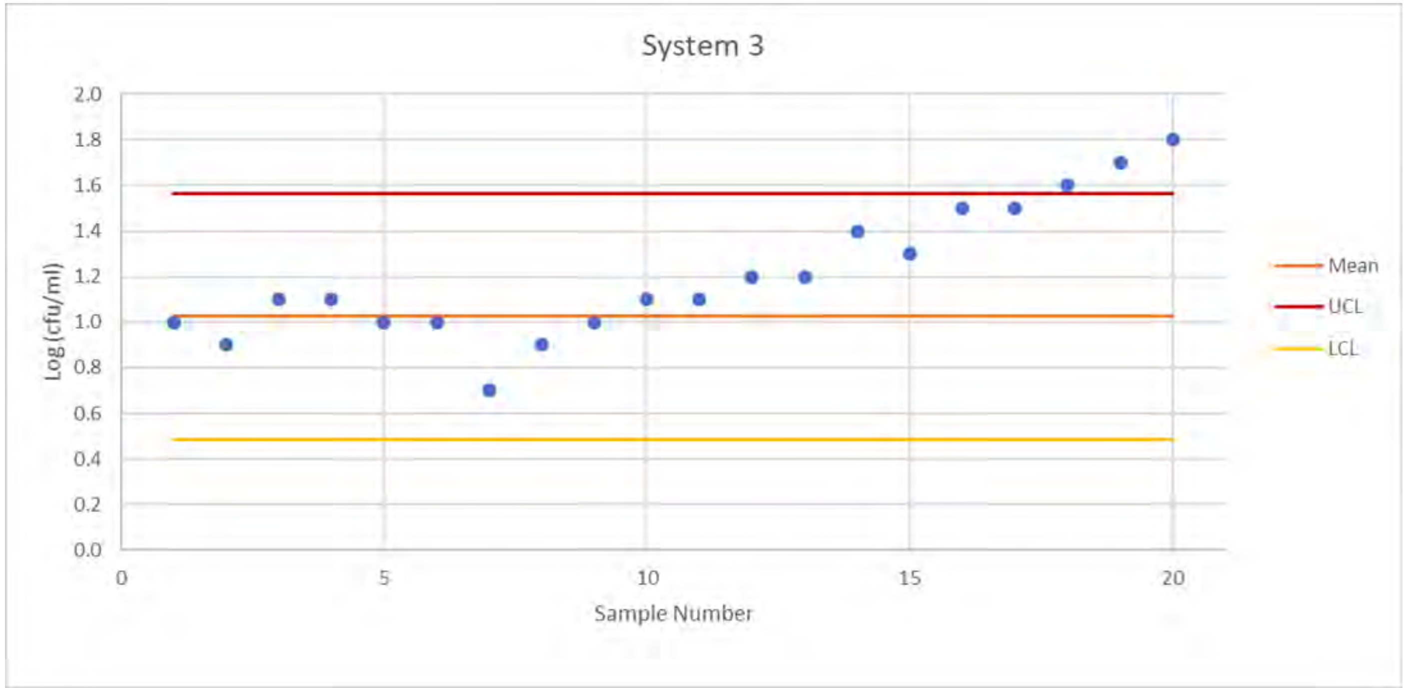


This chart suggests either a loss of control at a critical control point or the existence of another critical control point that had not been identified and controlled.

Process Control Charts 6

This chart depicts a situation where a component of the process is losing its effectiveness over time.

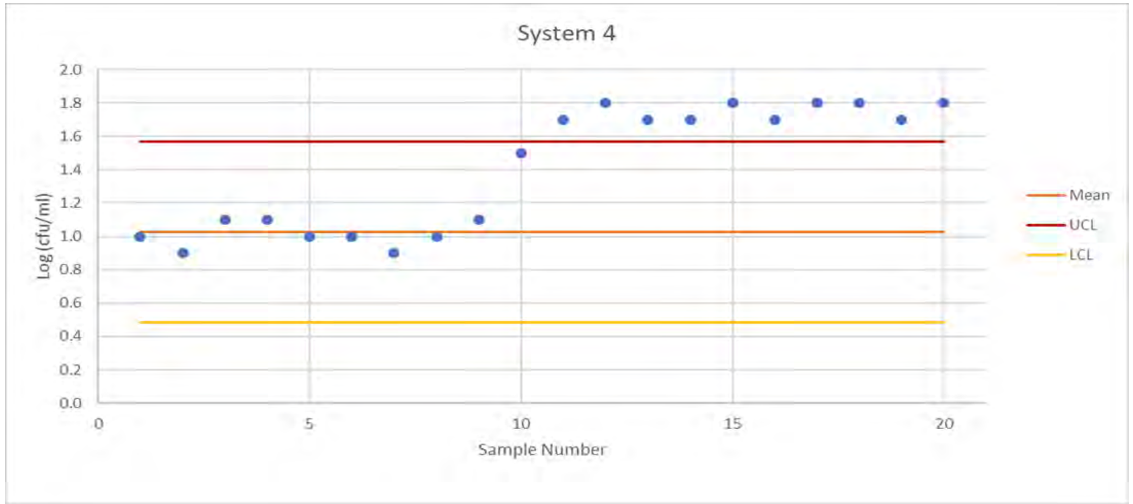
The loss of process control is apparent by the upward trend in the data points toward the UCL.



Process Control Charts 7

This chart depicts a catastrophic loss of process control.

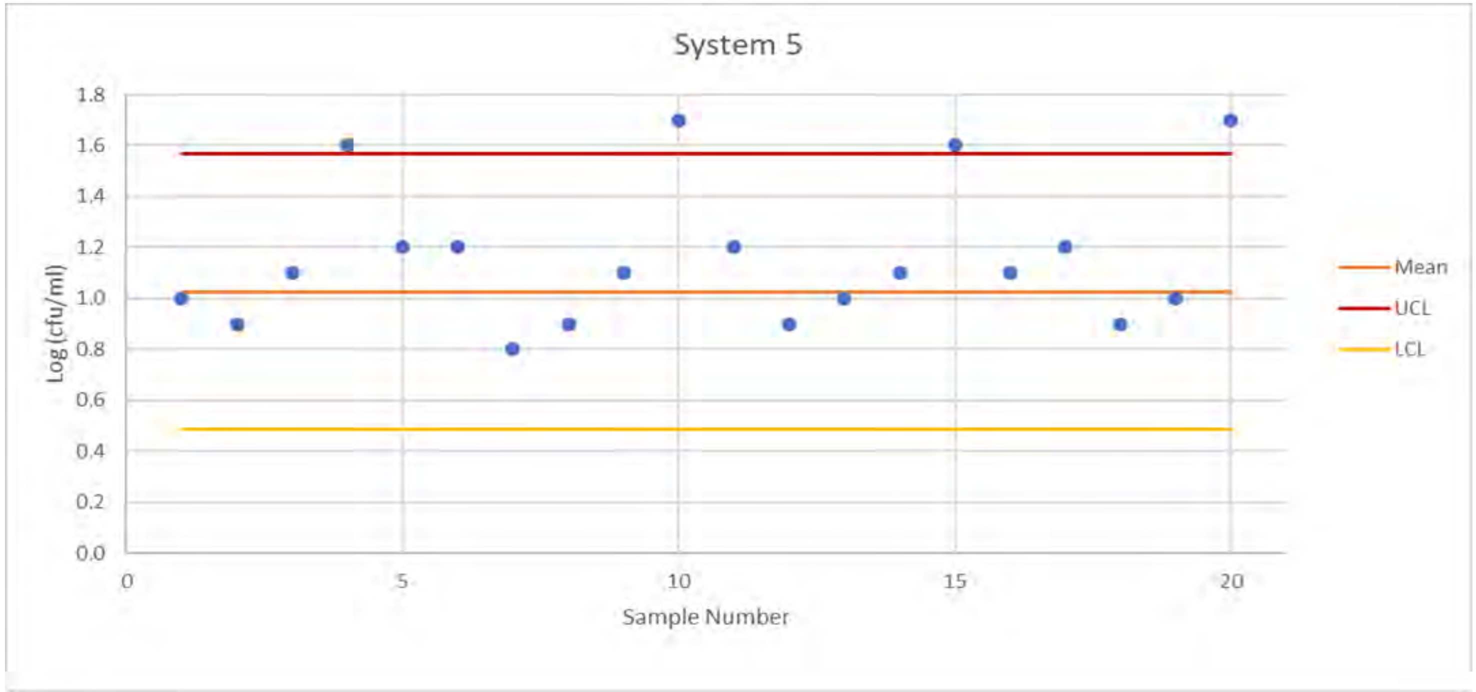
This pattern of test results would be encountered in a situation such as an abrupt failure of a key piece of equipment, such as an antimicrobial wash cabinet.



Process Control Charts 8

This chart depicts conditions where there is the existence of an intermittent but reoccurring problem within the process. Note the repeating pattern of the test results over time

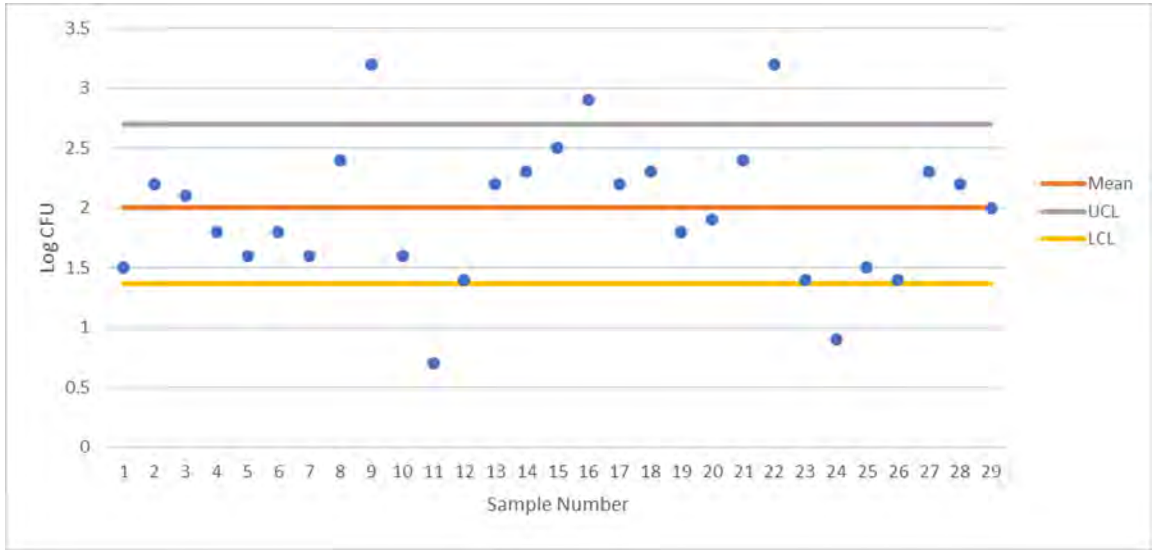
An example of a situation where this pattern may be observed is the dripping of condensation onto product as it travels down a conveyor belt.



Process Control Charts 9

For the PCC described above, what is being depicted in the system?

A. A gradual loss of control over time
B. Excessive Variability indicating a loss of control
C. A sudden loss of control due to an isolated incident
D. A sudden loss of control that remains out of control
E. The system is in control

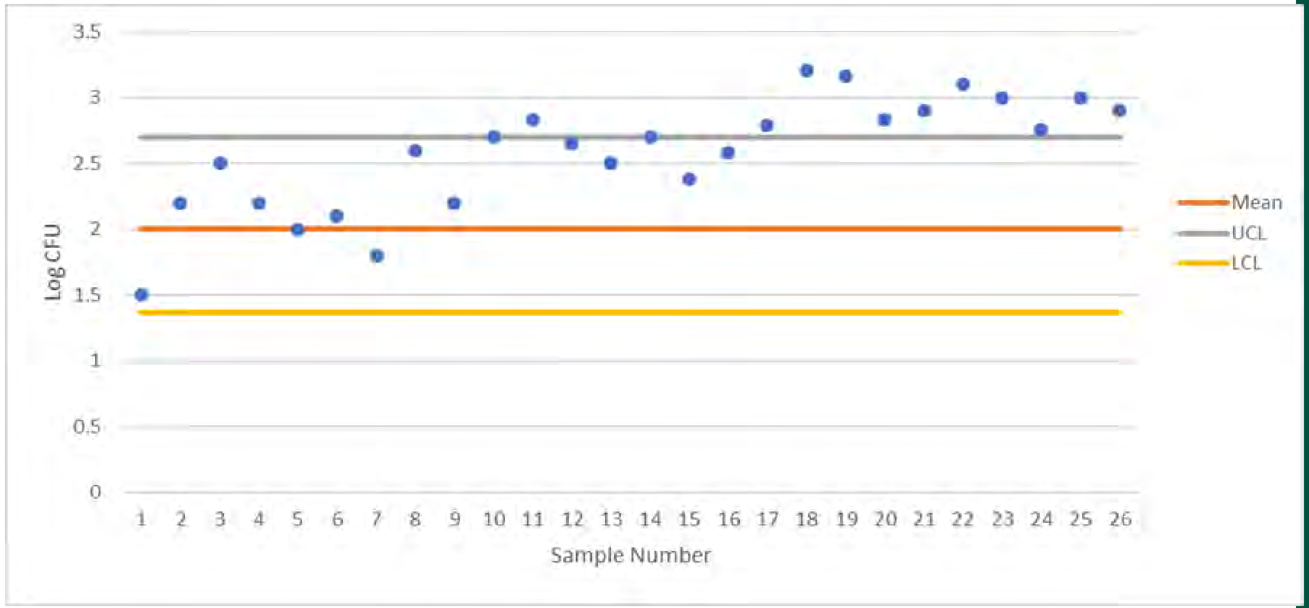


Excessive Variability indicating a loss of control

Process Control Charts 10

For the PCC described above, what is being depicted in the system?

A. A gradual loss of control over time
B. Excessive Variability indicating a loss of control
C. A sudden loss of control due to an isolated incident
D. A sudden loss of control that remains out of control
E. The system is in control



A gradual loss of control over time

$$\iiint \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} e^{-i\omega t} \frac{\Delta y}{\Delta x}$$

ANY QUESTIONS ??