San Jose State University
Department of Mechanical Engineering

ME 130 Applied Engineering Analysis

Instructor: Tai-Ran Hsu, Ph.D.

Chapter 10

Introduction to Statistics and Application in Engineering Analysis

CONDENSED VERSION

2018 version

Chapter Outline

- **What is Statistics?**
- **The scope of statistics**
- **Histogram and statistical data sets (Self learn)**
- **Common terminologies in statistics**
- **Normal distribution curve and normal distribution function**
- **Statistical quality control**
- **Statistical process control**
- **The control charts**
  - the 3-σ control charts
  - the sample range R-charts

Statistics – What is it?

Statistics is the science of decision making in a world full of uncertainties

Uncertainties in engineering and technology:

- Academic performance of this class
- In design engineering: uncertainties in design methodologies, material properties, fabrication techniques
- Quality of the products
- Market and sales of new and existing products

The Scope of Statistics

- **Collecting:**
  Data relating to certain events or physical phenomena Most datasets involve numbers
- **Organizing:**
  All collected data will be arranged in logical and chronicle order for viewing and analyses
  Datasets are normally organized in either ascending order or descending order.
- **Summarizing:** Summarizing the data to offer an overview of the situation
- **Presenting:** Develop a comprehensive way to present the dataset
- **Analyzing:** To analyze the dataset for the intended applications

Common Terminologies in Statistical Analysis:

The **histograms**, the **mode** of datasets; The **mean** and **mediam** of datasets; The **variance** and **Standard deviation** of datasets; The **Normal distribution** and **Normal distribution functions.**

The Mode of Statistical Dataset:

Statistical dataset can usually represented by the mode of the set

The mode of the dataset is represented by the number that appear in the dataset most frequently:

For instances: The set 2, 2, 5, 7, **9**, **9**, **9**, 10, 10 11, 12, 18 has mode **9**

The set:
1.75, 1.83, 1.85, 1.95, 1.97, 2.03, 2.03, 2.06, 2.13, 2.15, 2.15, 2.25, 2.35, 2.70, 2.70

has a triple mode of: 2.03, 2.15 and 2.70, as each of these numbers each appear twice in the set

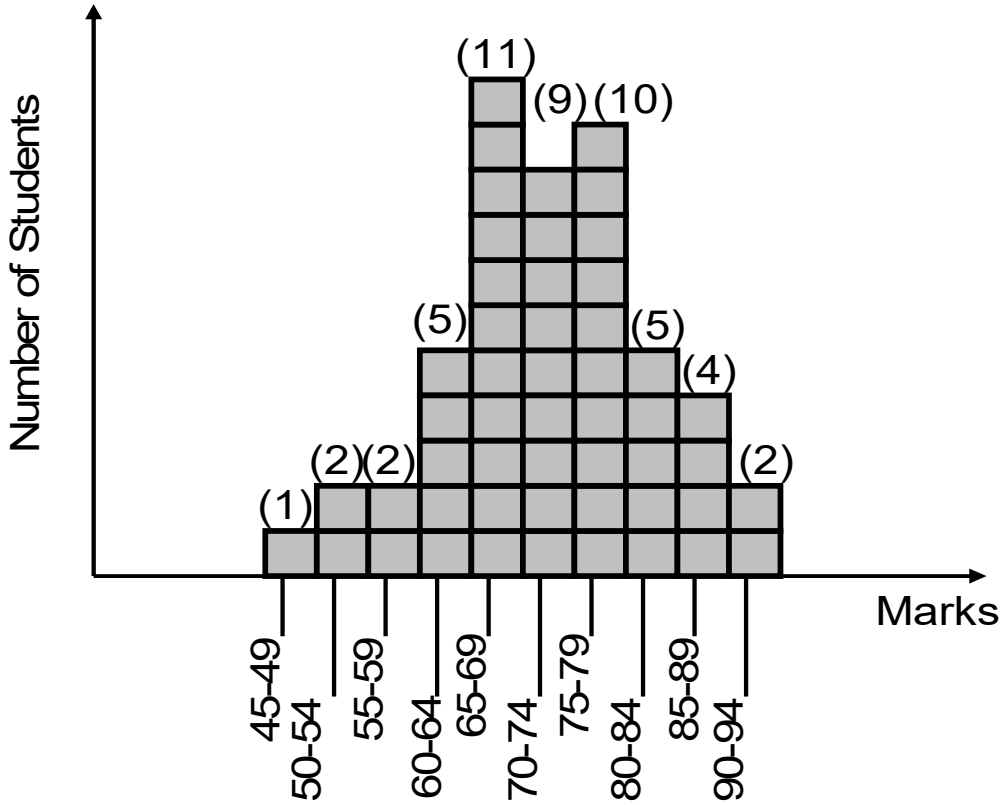
The set 3, 5, 8, 10, 12, 15 has no mode

Example on histogram:

The marks that students in the ME 130 class in Spring 2003 are tabulated in 10 intervals as:

Test scores	45-49	50-54	55-59	60-64	65-69	70-74	75-79	80-84	85-89	90-94
Frequency	1	2	2	5	11	9	10	5	4	2

The corresponding histogram, or the "frequency distribution" of the student's marks are:



Terminologies in Statistics for Engineering Analysis

The Mean

- The "Mean" of a a dataset is the arithmetic average of the data in the set
- It is a good way to represent the "Central tendency" of the set
- Mathematically, we may express the "Mean" of a dataset in the following way:

Given the dataset of: $x_1, x_2, x_3, \ldots\ldots\ldots\ldots\ldots, x_n$

with n = total number of data in the set

The arithmetic mean of the set may be computed by the following expression:

$$\bar{x} = \frac{Summation\ of\ all\ data}{Total\ number\ of\ data} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Advantage of using "Mean" in statistical analysis

- It includes ALL data in the set
- It always exists
- It is usually reliable in representing the "central tendency" of the data set

Disadvantage of using the "Mean" is that it loses its sense of representing the central tendency when a few out-ranged data are present in the set.
For example, the "Mean" of a set of 2,3,5,7,9,11,13 is 7.14, which is a close number representing the "central" value of the set.

This value becomes 15.71 if the last data of the set of 7 data become 73, i.e.: 2,3,5,7,9,11,73,
- Not a good representation of the "central tendency" of the data set.

The Median

In cases in which data in the set shows significant amount of "Out-ranged" data, the Median – meaning the "central data" is used to show the "central tendency" of the set.

For example, the same data set in the previous example with 7 data: 2,3,5,**7**,9,11, 73

We may take the data in the "central" of the set, i.e., **7** to be the Median representing the central tendency of the dataset.

The "central" data is readily identified in a set with odd number of data.

For the set with even number of data, the Median is the average of the two central data.

For example, the Median of the dataset: 5, 9, *11, 14*, 16, 19 is (11+14)/2 = 12.5

Like "Mean," the "median" of a dataset exists at all times. It is often a better way to express the "central tendency" of datasets with out-ranged data, such as the real estate price in Santa Clara Valley in California, in which significant number of substantially out-ranged house prices exist.

Deviation and Standard Deviation (σ)

Because the "Mean" of a dataset indicates its "central tendency," it is often required to measure how some data in the set "deviates" from its mean value $\bar{x}$

We may express such deviation of each data in the set as:

$x_1 - \bar{x}$

$x_2 - \bar{x}$

.

.

$x_n - \bar{x}$

The total variation of individual data from the mean is:

$$\sum\left[\left(x_1-\bar{x}\right)+\left(x_2-\bar{x}\right)+.....+\left(x_n-\bar{x}\right)\right]=0$$

In reality, of course, the total deviation of individual data to its mean CANNOT be zero, as the above mathematical expression shows.
We thus need to derive another mathematical expression that will not result zero in the summation

We realize the reason the total summation of individual deviation to be zero in the expression: $\sum\left[\left(x_1-\bar{x}\right)+\left(x_2-\bar{x}\right)+.....+\left(x_n-\bar{x}\right)\right]=0$

is because the first half of the content in the summation carry –ve signs, which cancel that of the second half with +ve signs. The sum of these two groups of variations is thus ZERO

In view the physical values for the deviation should not carry +ve or –ve signs, we can avoid recurrence of the above summation being zero by the following modification:

$$\sqrt{\sum\left[(x_1-\bar{x})^2+(x_2-\bar{x})^2+.....(x_n-\bar{x})^2\right]}=\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

We will observe that no term in the above expression may result in –ve value and thus avoid a zero total deviation of the dataset.

From which we define the "**Standard Deviation**" (**σ**) of a dataset to be:

$$\sigma=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}} \qquad (10.2)$$

The "Variance" which indicates the deviation of the dataset is:

$$\sigma^2=\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1} \qquad (10.3)$$

Example 10.4

To determine the standard deviation (σ) and the variance (σ^2) of the dataset:

{5 9 11 14 19} with n = 5

We computer the mean to be: $\bar{x} = 11.6$

We will then compute $(x_i - \bar{x})$

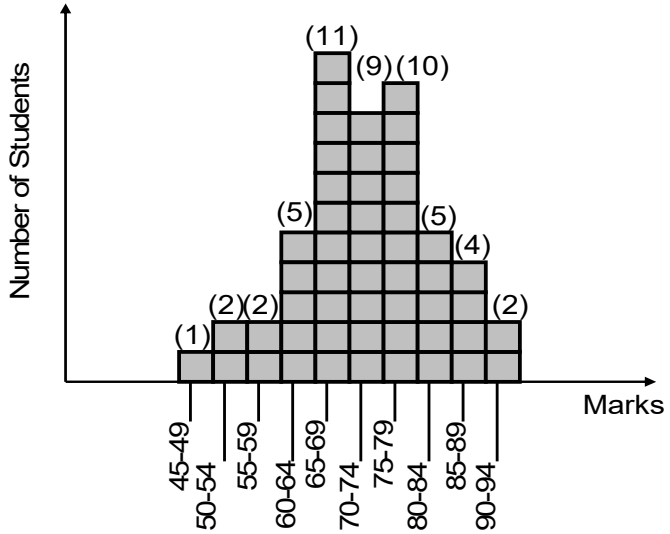
$$x_1 - \bar{x} = 5 - 11.6 = -6.6$$
$$x_2 - \bar{x} = 9 - 11.6 = -2.6$$
$$x_3 - \bar{x} = 11 - 11.6 = -0.6$$
$$x_4 - \bar{x} = 14 - 11.6 = 2.4$$
$$x_5 - \bar{x} = 19 - 11.6 = 7.4$$

Leading to: $$\sigma = \sqrt{\frac{(5-11.6)^2 + (9-11.6)^2 + (11-11.6)^2 + (14-11.6)^2 + (19-11.6)^2}{5-1}} = 5.27$$

The variance of the dataset is: $\sigma^2 = (5.27)^2 = 27.8$

The Normal Distribution Curve

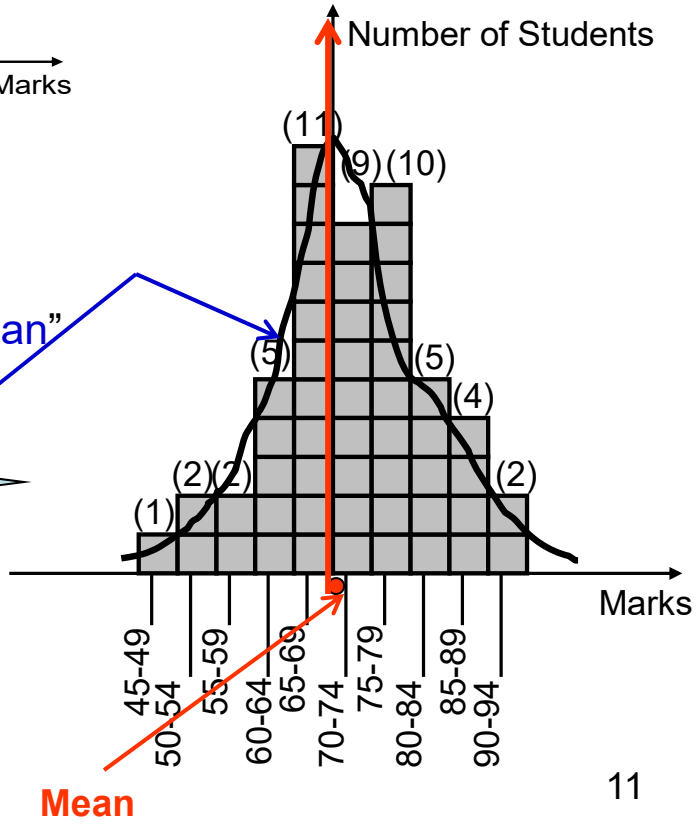
We have shown the mark distribution of 52 students in a class of Engineering Analysis in a histogram as:



If we show the above histogram by "shifting" the vertical axis from the left edge (@zero mark) to the "mean" value of the marks shown in the horizontal axis, we will have the same histogram but with a distribution shown as

The solid curve links the peak values of mark intervals in the histogram

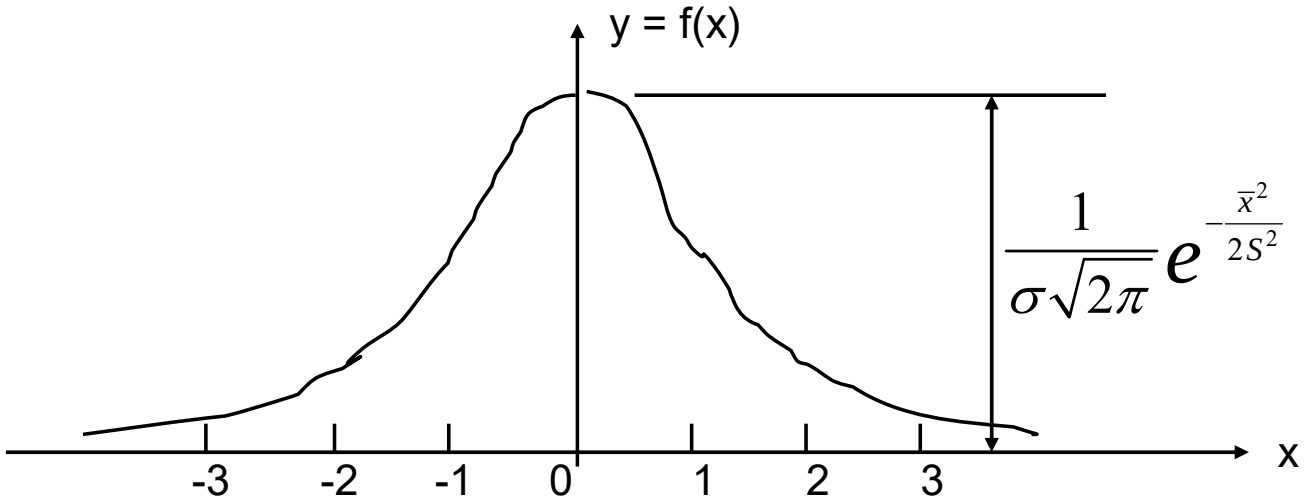
The histogram shown by a solid line curve with its population located at the "MEAN" is called **NORMAL distribution** of a statistical dataset



The Normal Distribution Function

Normal distribution function is a mathematical expression for the distribution of statistical datasets that are commonly happen in real world – a "BELL" shape histogram with vertical axis normalized and located at the "mean" of the data set

So, it has a great value in Statistical ANALYSES of many real-world cases



The NORMAL DISTRIBUTION FUNCTION has a mathematical expression of:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left[\frac{x-\bar{x}}{\sigma}\right]^2} \qquad (10.5)$$

where σ = the standard deviation of the data set given in Equation (10.2), and $\bar{x}$ is the mean of the set

Properties of Normal Distribution Curve

With the help of the mathematical expression of "normal distribution function," we are able to come up with the following important but interesting properties from mathematical analysis:

1)The data distribution is SYMMETRICAL about the mean

2)The percentage (%) of ALL data included are:

68.26% with the mean ± one standard derivation (σ)
94.4% with the mean ± 2σ
99.73% with the mean ± 3σ

Example 10.5:

A company produced 150,000 tires
The average life measured from 15,000 (i.e. 10%) tires is: 42,000 miles, with a standard deviation σ = 3000 miles

From the properties of "normal distribution curve," which we assume it fits the measured tire lives, we will come to the following observations:

68.26% cars had tire life of: ± σ = 42000 ± 3000 miles,
94.40% cars had tire life of: ± 2σ = 42000 ± 6000 miles, and
99.73% cars had tire life of : 3σ = 42000 9000 miles.

Statistical Quality Control

A SIMPLE FACT:

Cost and **Reliability** are two fundamental requirements for market success of any product

- Quality assurance is the key to Reliability of any product

Principal causes of poor quality of products:

- Poor design in setting dimensions, tolerances, surface finishing, improper selection of materials, etc.
- Manufacturing and fabrication processes relating to improper machining, assembly, testing and inspection.
- Improper conditions of machine tools and fabrication process control.
- Poor workmanship in all above production processes.

Why statistical quality control?

- The best assurance for quality of products is by **thorough inspection of EVERY** piece of product produced by a company.
- **Quality inspection of products cost: time and money.** Thorough inspection of EVERY piece of product produced by MASS PRODUCTION time will cost a lot of money and time – an Impractical solution to industry **– Thorough inspection on selected SAMPLE product is!!**
- **Questions on this practice are: (1)** How samples of products are to be selected **(2)** How many samples for quality inspection are required to give enough confidence of the quality of the ENTIRE BATCH of the product? **(3) How should each sample product be inspectgedd?** **(4)** What criteria (standards) to use for pass/fail of each sample product selected for quality inspection? **(5)** Most importantly, how one can relate the result of inspection obtained from the selected sample products to the quality assurance of the entire batch of the same products produced by the company in mass production??

STATISTICAL QUALITY CONTROL METHOD OFFERS SOLUTION TO THE ABOVE PROBLEMS

Statistical Process Control

We used the statement:

'conduct **thorough** INSPECTION of the manufacturing, and TESTING on the finished products" for "perfect" quality assurance of products'

The question is "How thorough" should we conduct those inspections and testing to ensure good quality in typical MASS PRODUCTION environment?"

If we focus our attention on the inspection and testing of the FINISHED product, it will be ideal if we can inspect EVERY PIECE of the finished product – practically impossible proposition!!

A question is then: "How many PIECES of the product should we pick up for inspection, and also how many TESTING points we should select on each piece in order to achieve credible quality control of the batch of the products from a mass production process?"

Here comes the value of statistical method for determining:

- the number of samples for quality assurance, and
- the number of inspection on each sample

for quality assurance of products in MASS PRODUCTION environment

It is called **Statistical Process Control**

Statistical Process Control-the use of control charts

In general, **Statistical process control (SPC)** is an effective method of monitoring a process through the use of **control charts**.

Control charts enable the use of objective criteria for distinguishing background variation from events of significance based on statistical techniques

Variations in the process that may affect the quality of the end product or service can be detected and corrected, thus reducing waste as well as the likelihood that problems will be passed on to the customer. With its emphasis on early detection and prevention of problems

SPC can lead to a reduction in the time required to produce the product or service from end to end

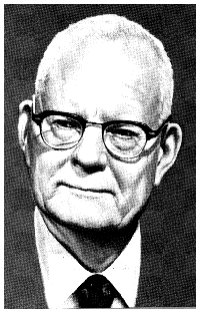
Process cycle time reductions coupled with improvements in yield have made SPC a valuable tool from both a cost reduction and a customer satisfaction standpoint

Statistical process control was pioneered by Walter A. Shewhart in the early 1920s. W. Edwards Deming later applied SPC methods in the United States during World War II, thereby successfully improving quality in the manufacture of munitions and other strategically important products.

Source: Wikipedia

Two Well-Recognized Pioneers in Statistical Quality Control for Mass-Produced Products

W. Edwards Deming



Born
October 14, 1900)
Sioux City, Iowa, USA

Died
December 20, 1993 (aged 93)
Washington DC, USA

Fields
Statistician

Alma mater
University of Wyoming
-BSEE (1921),
University of Colorado
-MS Math (1925),
Yale University
-PhD Physics (1928)

Genichi Taguchi

Born
January 1, 1924

Birth place:
Tokamachi, Japan

CitizenshipJapan

Fields
engineering, statistics

Institutions
Aoyama Gakuin University

Alma mater
Kyushu University **Known for**
Taguchi methods

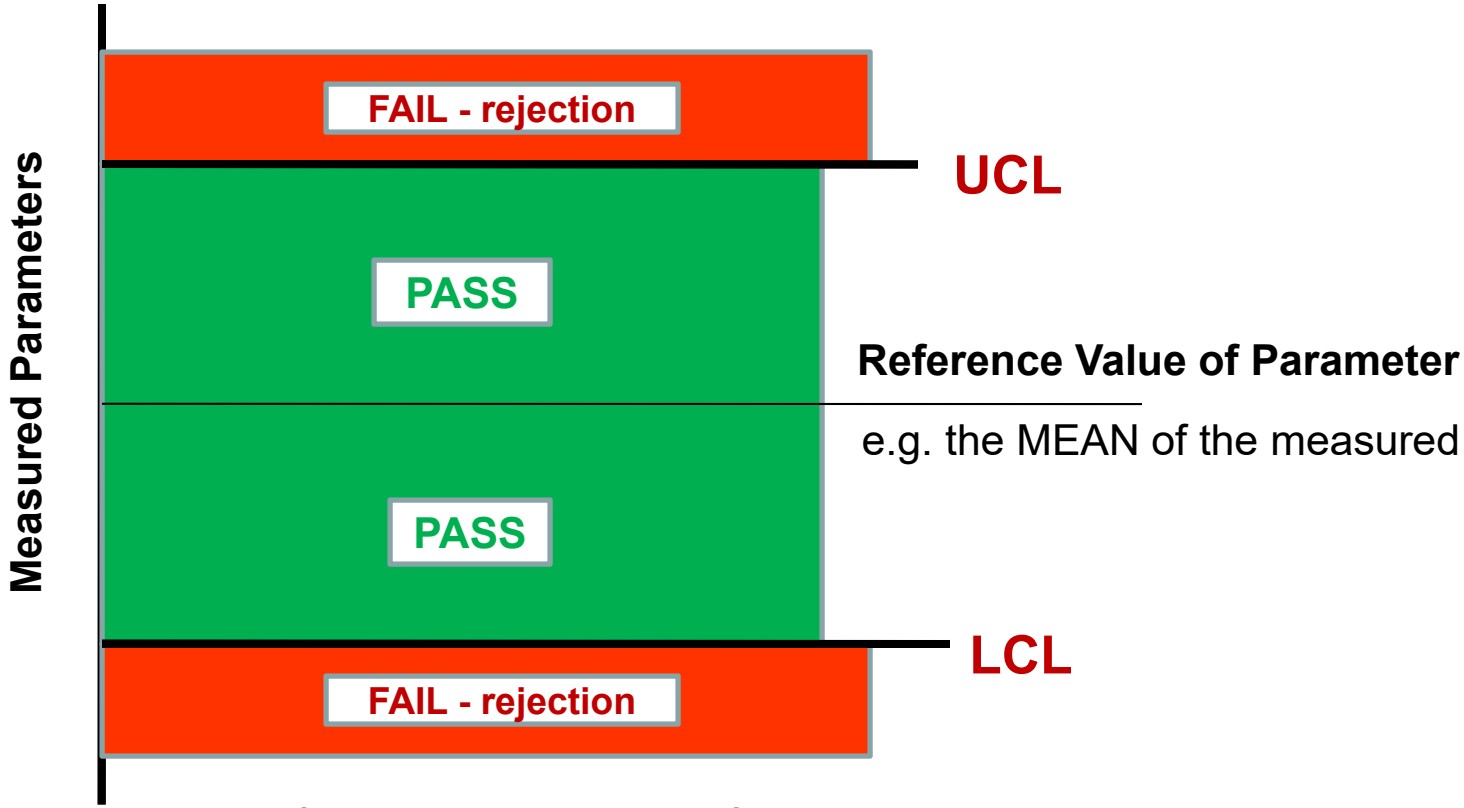
Influences
Matosaburo Masuyama

(Source: Wikipedia)

The Control Charts

- **Control charts** involves a range (BOUND) of acceptance of a parameter that relates to the quality of a product

- This range is defined by: **Upper control limit (UCL)** and **Lower control limit (LCL)**

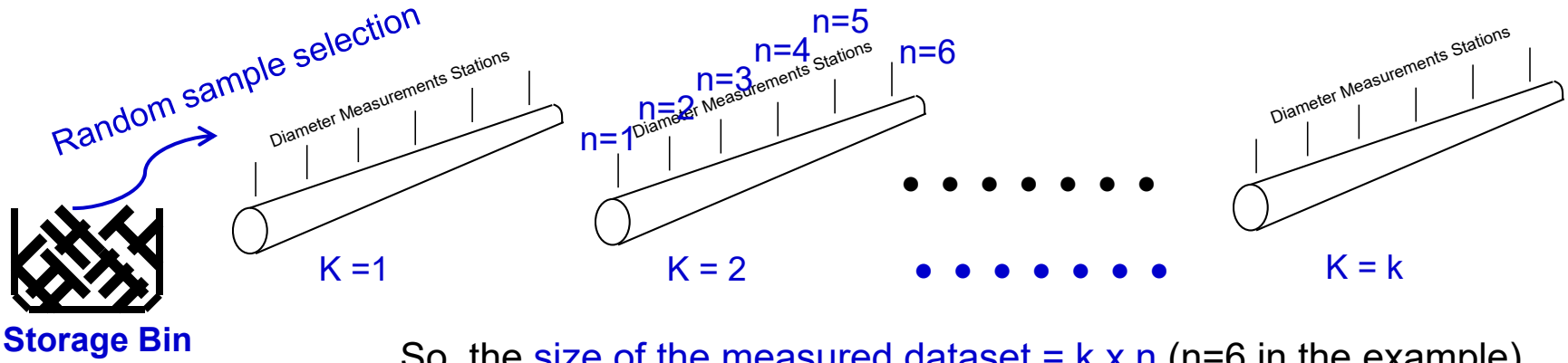


- Once established, all further measured data fall in the bounded region i.e., the green region are accepted
- Whenever a further measured data falls outside the bounded region (the red), the process is stopped for investigations on the causes for the failure

Construction of Control Charts

- The fundamental assumption is:
  All measured parameters (data) fall into a Normal (Bell shape) Curve
- So, the normal distribution function in Equation (10.5) can be used as the basis for mathematical derivations

- Control charts will be derived and constructed based on the MEASURED parameters (dataset) of:
  - Sample size = the number of selected sample from a batch = **k**
  - Number of measurements on EACH sample = **n**



So, the size of the measured dataset = k x n (n=6 in the example)

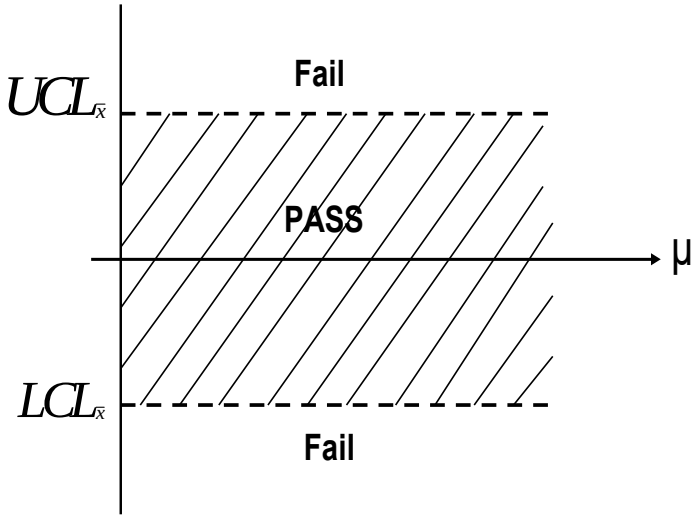
- One needs to compute: the MEAN (**μ**), and the STANDARD DEVIATION (**σ**)

The Three-Sigma Control Charts

- This is the simplest control chart of all
- It is constructed on the basis of the MEAN of the measured dataset
- k-number samples are randomly picked up from a storage bin with produced products
- Take n-measurements on each sample with a mean value $\bar{x}$ on the sample
- Compute the MEAN of the k x n measurement values = μ, and STANDARD DEVIATION = σ
- The upper and lower control limits for quality control can be determined by:

$$LCL_{\bar{x}} = \mu - \frac{3\sigma}{\sqrt{n}} \quad \text{and} \quad UCL_{\bar{x}} = \mu + \frac{3\sigma}{\sqrt{n}} \qquad (10.6)$$

- Graphical expression of 3-σ control charts:



- **The use of 3-σ control chart in quality control:**
  Once the chart is established with n measurements form each of the k-number of samples, the quality control engineer or technician will pick up samples from future productions and conduct same n-measurements with a calculated sample mean measured $\bar{x}$. The sample is accepted if this value falls within the bounds.
  If this value falls outside the bounds, the manufacturing process should be stopped, and the causes of failure Need to be investigated with remedial actions.

Example 10.6

- Quality control on the IC chip by measuring the output voltage using 3-σ control chart.
- 5 samples randomly picked up from a storage bin with IC chips mass produced from a process
- 3 measurement of voltage output (mV) from each sample, recorded as follows:

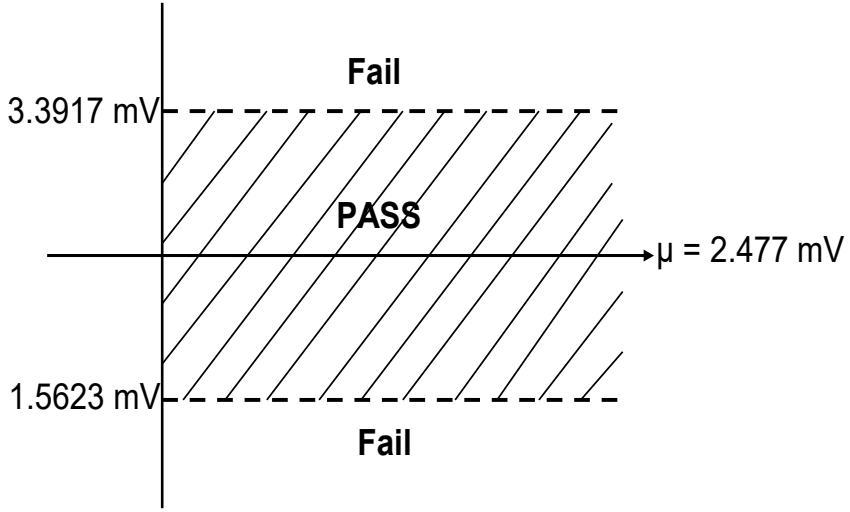
Sample 1:	2.25	3.16	1.80
Sample 2:	2.60	1.95	3.22
Sample 3:	1.75	3.06	2.45
Sample 4:	2.15	2.80	1.85
Sample 5:	3.15	2.76	2.20

k = 5
n = 3

The MEAN of the 15 measurements **μ = 2.477** mV by Equation (10.1), and the STANDARD DEVIATION **σ = 0.5281** mV by Equation (10.3)

The upper and lower control limits of the dataset are computed from Equation (10.6):

$$LCL_{\bar{x}} = \mu - \frac{3\sigma}{\sqrt{n}} = 2.477 - \frac{3x0.5281}{\sqrt{3}} = 1.5623 \quad \text{and} \quad UCL_{\bar{x}} = \mu + \frac{3\sigma}{\sqrt{n}} = 2.477 + \frac{3x0.5281}{\sqrt{3}} = 3.3917$$



Application of the control chart:

Average measured voltage output from any future randomly selected sample should fall within the bounds In the chart.

Any sample fails to have its average measured output fall outside the bounds will be rejected, and the process will be halted for further inspection.

Control Charts for Sample Range - The R-chart

Difference of 3-σ control charts and R- control charts:

- **3-σ charts:** Based on the MEAN of the k x n measured dataset
- **The R-charts:** Based on the "range," i.e. the difference of the MAX and MIN of the measured values
- The R-chart is established on "NORMAL distribution of measured parameters

Working Sheet for R-Cart

Sample	Measured Parameters	Sample Mean $\overline{X}$	R
k = 1	x_1 x_2 x_3 x_4 x_5.....x_n	$\overline{x}_{k=1}$	x_{max} - x_{min}
k = 2	• • • • • • • • • • • • • • • • • • •	$\overline{x}_{k=2}$	•
•	• • • • • • • • • • • • • • • • • • •	•	•
•	• • • • • • • • • • • • • • • • • • •	•	•
•	• • • • • • • • • • • • • • • • • • •	•	•
k = k	• • • • • • • • • • • • • • • • • • •	•	•
			•

MEAN VALUES: $\overline{X}$ $\quad \bar{R} = d_2\sigma$ or as computed

NOTE: The computed average sample range $\overline{R}$ obtained by computed from sample average = ($\bar{R} = d_2\sigma$) if the measured dataset fits perfectly with NORMAL distribution

Table 10.2 Factors for Estimating $\overline{R}$ and Lower and Upper Control Limits

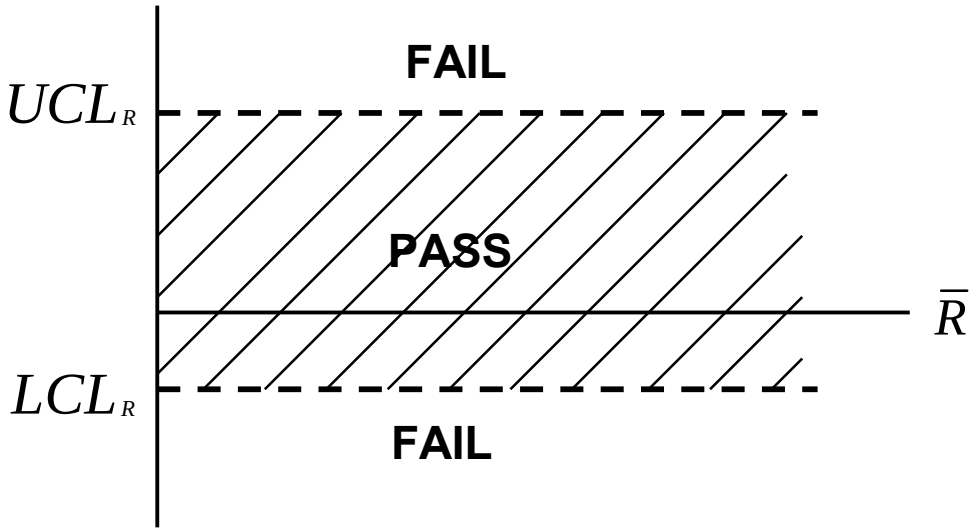
(Ref: Rosenkrantz, W. A. "Introduction to Probability and Statistics for Scientists and Engineers," McGraw-Hill, New York)

No. of Measurements On Each Sample, n	Factor, d_2	Coefficient, D_1	Coefficient, D_2
2	1.128	0	3.69
3	1.693	0	4.36
4	2.059	0	4.70
5	2.326	0	4.92
6	2.534	0	5.08
7	2.704	0.20	5.20
8	2.847	0.39	5.31
9	2.970	0.55	5.39
10	3.075	0.69	5.47
11	3.173	0.81	5.53
12	3.258	0.92	5.59
13	3.336	1.03	5.65
14	3.407	1.12	5.69
15	3.472	1.21	5.74

The Lower control limit: $LCL_R = D_1\sigma$ and the Upper control limit: $UCL_R = D_2\sigma$

where σ = standard deviation of the k x n dataset of the measured parameter

The R-Chart on Sample Range:



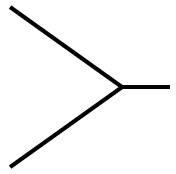
- The application of the R-chart for quality control is similar to that with the 3-σ charts:

- The "**range**" of the measured parameter of any sample after the R-chart is established with values outside the bounds will result in the rejection of the sample. The manufacturing process will be stopped for an investigation on the causes for the inferior quality.

Example 10.7

Use the R-chart for quality control in a process of IC-chip manufacturing described in Example 10.6. The measurements of the IC chip's output voltage at 3 leads on each of the 5 samples are tabulated below:

Sample 1:	2.25	3.16	1.80
Sample 2:	2.60	1.95	3.22
Sample 3:	1.75	3.06	2.45
Sample 4:	2.15	2.80	1.85
Sample 5:	3.15	2.76	2.20



k = 5
n = 3

The working chart for the R-range is:

Sample	Measured Voltages (mV)			Mean value, $\overline{X}$	Sample Range
k = 1	2.25	3.16	1.80	2.4033	1.36
2	2.60	1.95	3.22	2.5900	1.27
3	1.75	3.06	2.45	2.4200	1.31
4	2.15	2.80	1.85	2.2667	0.95
5	3.15	2.76	2.20	2.7033	0.95
Total k = 5	Total n = 3			Mean, μ = 2.477	$\overline{R}$ = 1.168 (from dataset)

The STANDARD DEVIATION σ for the dataset of 15 is calculated using Equation (10.2) to be σ = 0.5281

With n = 3, we find the coefficients d_2, D_1 and D_2 from Table 10.2 to be:

$d_2 = 1.693$ $D_1 = 0$ $D_2 = 4.36$

From which, we will calculate:

- The average of "sample range"

$$\overline{R} = d_2\sigma = 1.693\,x\,0.5281 = 0.8941$$

which is not the same value as computed from the dataset (= 1.168). This is because The dataset is **not** a good fit to the NORMAL distribution curve. A larger sample size may improve this situation

- The Lower control limit: $LCL_R = D_1\sigma = 0 \times 0.5281 = 0$, and

- The Upper control limit: $UCL_R = D_2\sigma = 4.36 \times 0.5281 = 2.3025$

- **The R-chart is:**