Data Analysis
Part 2

Interquartile Range

- The lower quartile, Q_1, is the value such that one-fourth of the observations fall below it and three-fourths fall above it.
- The middle quartile is the median half the observations fall below it and half above it.
- The third quartile, Q_3, is the value such that three-fourths of the observations fall below it and one-fourth above it.
- The **interquartile range** (IQR) is the difference between the third quartile and the first quartile.
- IQR = $Q_3 - Q_1$
- The larger the IQR value, the greater the spread of the data.
- To find the IQR, the data are ranked in ascending order; $Q1$ is located at rank $0.25(n+1)$, where n is the number of data points in the sample, and $Q3$ is located at rank $0.75(n+1)$.

- Random sample of 20 observations on the welding time (in minutes) of an operation gives the following values. Find IQR

  22 25 18 20 21 17 19 26 18 23 20 21 26 19 20 18 17 22 24 22

Solution

- location of $Q_1 = 0.25(n+1) = 0.25(20+1) = 5.25$
- location of $Q_3 = 0.75(n+1) = 0.75(20+1) = 15.75$

Q_1's location = 5.25 $\quad$ Q_3's location = 15.75

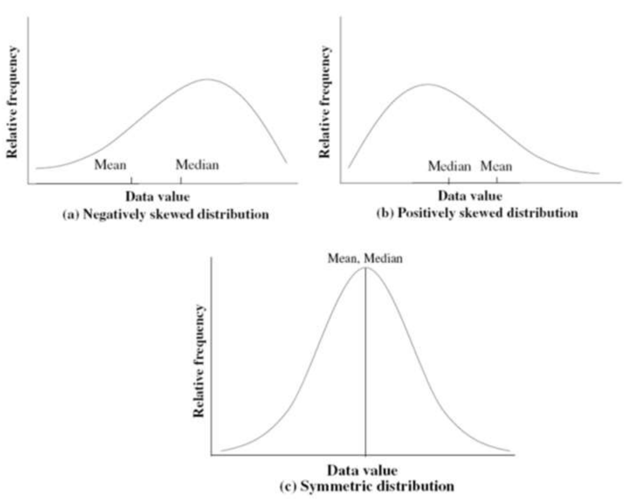
Rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Data Value	1.7	1.7	1.8	1.8	1.8	1.9	1.9	2.0	2.0	2.0	2.1	2.1	2.2	2.2	2.2	2.3	2.4	2.5	2.6	2.6

$Q_1 = 1.825$ $\quad$ $Q_3 = 2.275$

- Linear interpolation yields a Q_1 of 1.825 and a Q_3 of 2.275.
- The interquartile range is then IQR Q_3 - Q_1= 2 275 - 825 = 0 45 minute

Skewness

- The **skewness coefficient** describes the asymmetry of the data set about the mean.
- Part (a) is a negatively skewed distribution (skewed to the left),
- part (b) is positively skewed (skewed to the right), and
- part (c) is symmetric about the mean.



(a) Negatively skewed distribution
(b) Positively skewed distribution
(c) Symmetric distribution

Skewness

- The skewness coefficient is zero for a symmetric distribution, because [as shown in part (c)] the
- mean and the median are equal.
- For a positively skewed distribution, the mean is greater than the median because a few values are large compared to the others; the skewness coefficient will be a positive number.
- If a distribution is negatively skewed, the mean is less than the median because the outliers are very small compared to the other values, and the skewness coefficient will be negative.
- The skewness coefficient indicates the degree to which a distribution deviates from symmetry.
- The skewness coefficient is calculated as follows:

$$\text{Karl Pearson's Coefficient of skewness} = \frac{3(Mean - Median)}{Standard\ Deviation}$$

$$\text{Bowley's Coefficient of skewness} = \frac{Q_3 + Q_1 - 2Median}{Q_3 - Q_1}$$

Numerical

- Find Skewness coefficient of following data

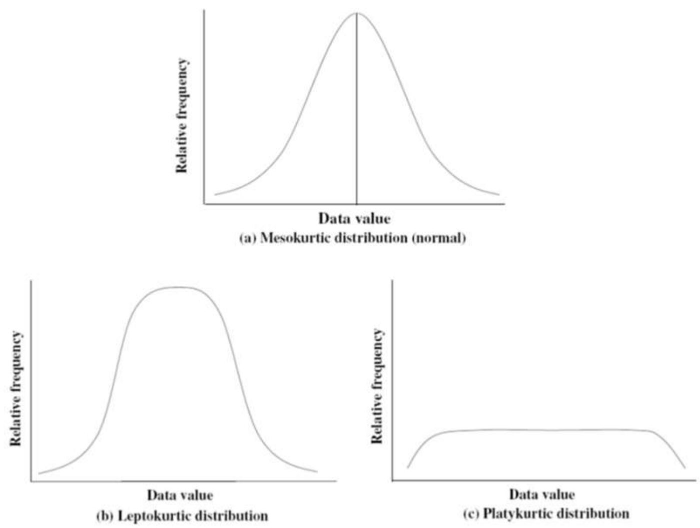
22 25 18 20 21 17 19 26 18 23

Mean	20.9		
Median	20.5		
Q1	18	location	2.75
Q3	23.5	location	8.25
SD	2.91376		
Karl Pearson CS	**0.411839**		
Bowley's CS	**0.090909**		

Since the skewness coefficient will be a positive number, the distribution is positively skewed.

Kurtosis

- **Kurtosis** is a measure of the peakedness of the data set. It is also viewed as a measure of the "heaviness" of the tails of a distribution.
- The **kurtosis coefficient** is a relative measure. For normal distributions (discussed in depth later), the kurtosis coefficient is zero.
- Figure shows a normal distribution (*mesokurtic*), a distribution that is more peaked than the normal (*leptokurtic*), and one that is less peaked than the normal (*platykurtic*).



Kurtosis

- For a leptokurtic distribution, the kurtosis coefficient is greater than zero.
- The more pronounced the peakedness, the larger the value of the kurtosis coefficient.
- For platykurtic distributions, the kurtosis coefficient is less than zero.
- The kurtosis coefficient is given by

$$\gamma_2 = \frac{\mu_4}{\mu_2^2}$$

Second Moment, $\mu_2 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})^2}{N}$ Fourth Moment, $\mu_4 = \frac{\sum_{i=1}^{N}(X_I - \bar{X})^4}{N}$

- Find Kurtosis coefficient of following data

22 25 18 20 21 17 19 26 18 23

			1	2	3	4	5	6	7	8	9	10
			17	18	18	19	20	21	22	23	25	26
		Mean	20.9									
		(Xi-mean)	-3.9	-2.9	-2.9	-1.9	-0.9	0.1	1.1	2.1	4.1	5.1
		(Xi-mean)^2	15.21	8.41	8.41	3.61	0.81	0.01	1.21	4.41	16.81	26.01
Second moment		**Variance**	**8.49**									
		Std Dev	2.91376									
		(Xi-mean)^4	231.3441	70.72	70.72	13.03	0.65	0.0001	1.46	19.44	282.57	676.52
Fourth moment			**136.6497**									
	Kurtosis Coefficient		**0.06213**									

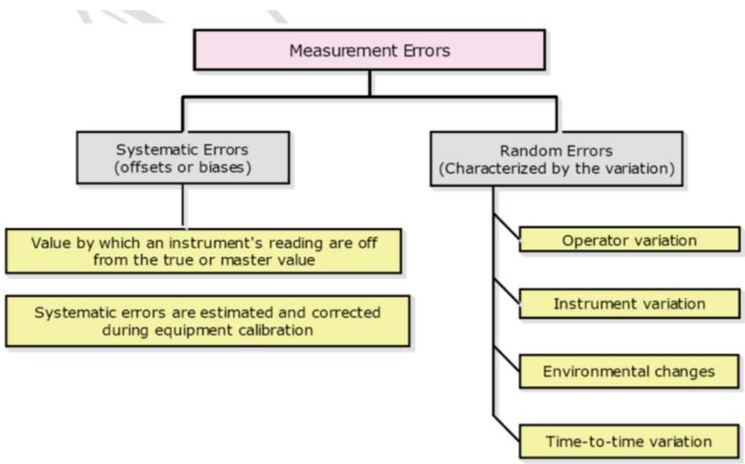
Measurement System Analysis using Gage R&R

- The conclusions drawn from the statistical methods depend on the accuracy of data.
- If the measuring instrument and the measurement method are not capable of making accurate and repeatable measurements, the data can have significant measurement error.
- In such cases, the conclusions drawn from the data are inaccurate and misleading.
- It is critical to assess the accuracy of the measurement process at the start of the study.
- Inaccurate measurements may lead to false signals on control charts.
- In the presence of significant error in the measurement process, a capable process may be confused with an incapable process.
- Several factors affect the
  - reliability of measurements including:
  - differences in measurement procedures,
  - differences among operators,
  - instrument repeatability and reproducibility, and
  - instrument calibration and resolution.

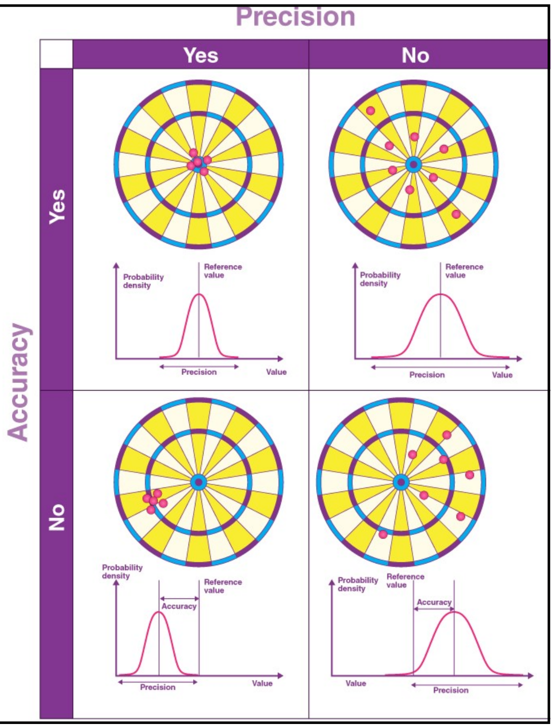
Terms

- **Gage**: Is any device used to obtain measurements
- **Part**: Item that is subject to measurement
- **Trial**: Set of measurements taken on apart
- **Measurement System** : Complete process of obtaining measurements.
- **Location:** Bias, Stability, linearity
- **Gage Bias** is the difference between average observed value and a reference value
- **Stability:** Measurement system should be in statistical control i.e. stable. This means the variation in measurements are due to common causes and special causes.
- **Gage linearity** tell you how accurate the measurements are across the expected range of measurements

Measurement Errors



- Measurement system errors can be divided into two categories: **accuracy and precision**.
- ***Accuracy*** *is the difference between the average of measurements made on a part and the true value of that part.*
- ***Precision*** *is the degree to which repeated measurements tend to agree with each other.*
- It is getting consistent results repeatedly.
- **Accuracy** refers to long-term average of measurements while **precision** refers to long-term variation



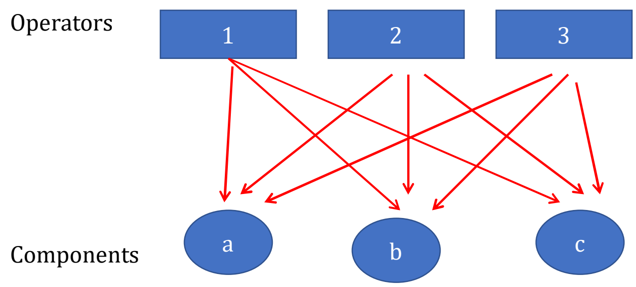
- When measurements are made, some of the observed variability is due to the part or the product being measured, and some variability can be attributed to measurement or gage variability.
- The total variability can be stated as: $\sigma^2_{Total} = \sigma^2_{product} + \sigma^2_{measurement}$

where, σ^2_{Total} is total variation, $\sigma^2_{product}$ is the component of variation due to the product or the part, $\sigma^2_{measurement}$ is the component of variation due to measurement error.

Note: $\sigma^2_{measurement}$ is also referred as σ^2_{gage}

- The measurement system analysis is commonly known as ***Gage Repeatability and Reproducibility (Gage R&R)*** study.
- ***Repeatability*** of a measuring instrument refers to how well the instrument is repeatedly able to measure the same characteristic under the same condition.
- ***Reproducibility*** is the variation due to different operators using the same measuring instrument at different time periods, and different environmental conditions.
- The purpose of Measurement System Analysis or Gage R&R study is to determine the part of variation in the data resulting from the variation in the measurement system.

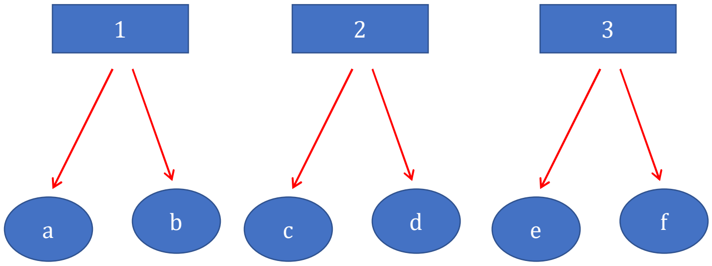
Crossed Gage R & R



Crossed GR&R is used in non-destructive scenarios—when parts are not destroyed during measurement and can be measured twice.

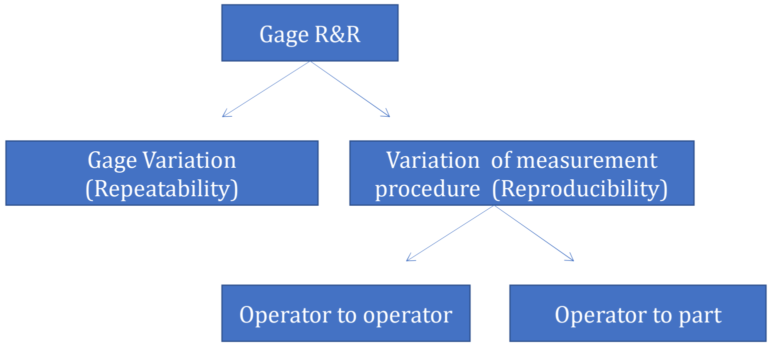
For example, when measuring the length of a part, the part is not changed during the measurement.

Nested Gage R&R Study

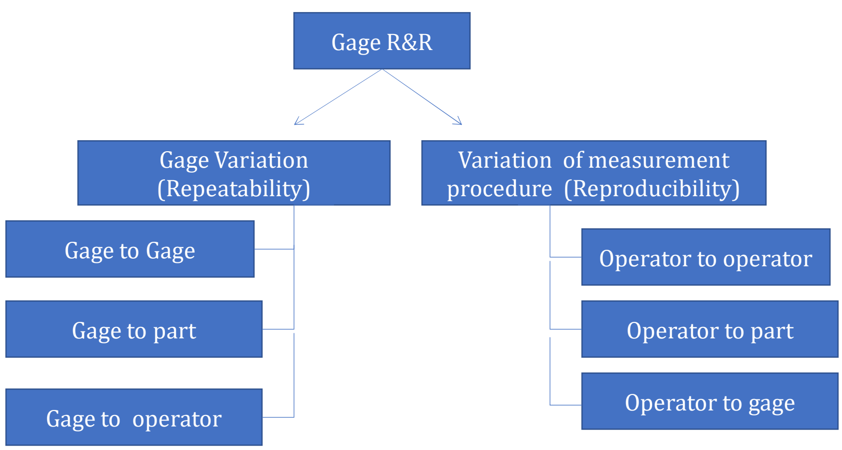


- Nested GR&R is used in destructive testing scenarios.
- For example, testing the amount of force required to open a bag of potato chips, heat-treating steel tubes, and testing the strength of a rope until it breaks.
- In these examples, the sample is destroyed during testing, making it impossible to test the sample again.
- A critical factor of a nested GR&R study is to identify a batch of material that is so close to the original it reasonably can be assumed the parts in the batch are the same part.
- The key to being able to run a destructive GR&R study is the assumption that a batch is homogeneous.

Standard (Crossed and Nested) Gage R&R Study



Extended Gage R&R Study



In many cases, the effects of operator and part are not sufficient to provide a complete understanding of the measurement system, so a third variable (typically gage) is added to the standard GR&R study.