2015-2016 Spring Semester
Material and Energy Balance
Statistical Concepts Applied to Measurement and Sampling

Assist. Prof. Dr. Murat Alkan

01.03.2016

3rd Week

The Reasons for Statistical Approach

- Engineers prepare material and energy balances for a variety of reasons and at various stages in the life of a process or plant.
- At the design stage, planners make various assumptions concerning efficiency of reactions, heat loss rates, etc., and they compute theoretical balances to set the process flow rates, temperatures, efficiencies, equipment sizes and so on.
- The balances are exact but theoretical, and satisfy all equations precisely. They might also conduct sensitivity analyses, which investigate how sensitive the conclusions are to the various assumptions made.
- Throughout the design stage, designers must estimate the errors associated with the material property and equipment specifications.

The Reasons for Statistical Approach

- Once a plant or process is built and operating, managers may wish to optimize its control or economic performance.
- This means the operator needs to compute material and/or energy balances on the real system using actual measurements of flow rates, chemical compositions, temperatures, and the like.
- Then, using those real-world balances, the operator can adjust various process factors to improve performance.
- Any recommendation for a process change based on plant measurement must recognize that errors can occur when taking and analyzing samples.

The Reasons for Statistical Approach

- Real-world material balances require samples from heterogeneous materials, indirect measurements or calculations of flow rates, temperature measurements under extreme environmental situations, and sometimes-arduous chemical analyses, all of which have some degree of error associated with them.
- Therefore, actual balances may have considerable uncertainty in the results.
- For example the presence of unaccounted-for or trace elements in a material mixture can give unknown or unacceptable errors in the material and energy balances.
- Finally, in process analysis, we often take repeated samples in preparation for an improvement campaign. We should take enough samples to get representative results so we can make a valid statistical analysis, or use the results to develop a process model.

Basic Statistical Concepts

- Copper is an impurity in flat-rolled steel. Steelmakers are concerned with the copper content because if it is too high, the steel is more difficult to process, leading to a higher cost to the producer.

Copper impurities in 50 batches of steel, mass percent

0.205	0.143	0.113	0.219	0.173
0.224	0.154	0.064	0.137	0.172
0.155	0.123	0.131	0.129	0.173
0.125	0.176	0.197	0.145	0.138
0.179	0.178	0.170	0.236	0.207
0.143	0.133	0.167	0.190	0.169
0.237	0.180	0.167	0.120	0.150
0.179	0.246	0.114	0.159	0.210
0.155	0.154	0.141	0.175	0.166
0.150	0.150	0.120	0.179	0.149

Basic Statistical Concepts

- The engineer is usually not interested in the sample for its own sake.
- The company probably doesn't care about the specific fact that the tenth batch contained 0.172 %Cu.
- Rather, they are more likely to be interested in demonstrating that the copper content falls within a desired composition range.
- In addition, they may want to search for a statistically significant relationship between the tensile strength of the steel and the %Cu.
- Statisticians use the words *population* or *process* to refer to the collection of all objects similar enough to the sample that the measurements of the sample objects should give a good idea about the (unknown) measurements of objects not in the sample.
- The collection of every batch of steel made by your company would be a population.

Basic Statistical Concepts

- In real life, one never knows everything about the population or process.
- The steel company may want to develop a correlation between tensile strength and copper content on each batch of steel rather than running tensile strength tests on all batches.
- The more important function of statistics is the process of making statements about the population using only the information in a sample. This process is called inferential statistics.
- Some questions that the company might try to answer based on this sample of 50 measurements are:
  - Can they estimate the largest %Cu likely to occur in a month's worth of (say) 240 batches?
  - If the steelmaking process is changed, and another 50 batches are analyzed, how do they know if the change has significantly affected the copper content?

Basic Statistical Concepts

Arithmetic Values

- Average-Mean (ortalama)
- Median (orta değer)
- Mode (mod)
- Range (aralık)
- Variance (değişim)
- Standard Deviation (standart sapma)

Graphs

- Scatter diagram (dağınıklık/ilişki)
- Stem-leaf (ağaç-yaprak)
- Bar (çubuk)
- Pie (pasta/daire)
- Probability (olasılık)
- Histogram

- The three most important summary statistics of a population are the **mean**, **standard deviation**, and **variance**.

Mean, Median

- The first of our summary statistics, the *mean*, is just another word for the ordinary arithmetic average. Add up all the numbers, and divide by how many of them there are.

$$Sample\ Mean = \mu = \bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

- The *median* is supposed to be the number "in the middle of the data", or the number that separates the bottom 50% of the data from the top 50% of the data. The numbers must be ranged ascending or descending order.
- If there is (2n-1) data (odd), the median value equals the (n)th data
- If there is (2n) data (even), the median value equals

$$Median = \frac{x_n + x_{n+1}}{2}$$

Mode, Range

- The *mode* is the value that appears most often in a set of data. In other words, it is the value that is most likely to be sampled. If no number is repeated, then there is no mode for the list.
- In arithmetic, the *range* is just the difference between the largest and smallest values. However, in statistics, this concept of range has a more complex meaning.
- The range is the size of the smallest interval which contains all the data and provides an indication of statistical dispersion. It is measured in the same units as the data. Since it only depends on two of the observations, it is most useful in representing the dispersion of small data sets.

Variance and Standard Deviation

- The other common type of summary statistic is a measure of *dispersion*, which quantifies how spread out the set of numbers is from its center. The most common statistical measures of dispersion are *standard deviation* and *variance*.

$$Population\ Variance = \sigma^2 = S^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$Sample\ Variance = \sigma^2 = S^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$Population\ Standard\ Deviation = \sigma = S = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$Sample\ Standard\ Deviation = \sigma = S = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Histograms and Frequency Distributions

- When test results are presented in a journal article or at a conference, the author often uses a *frequency distribution* or a *histogram* to display the measured variable.
- A frequency distribution is a table summarizing the different values occurring in the variable by subdividing the total range of values into sub-ranges (called *bins*), and counting the number of values in each sub-range. Sometimes, instead of reporting the number of measurements in a given range, researchers report the *relative frequency*, that is, the percentage of the sample that falls in each bin.
- The next step is to pick the ranges for a frequency distribution. Statisticians use a particular kind of column chart called a *histogram* to represent frequency distributions and relative frequency distributions.

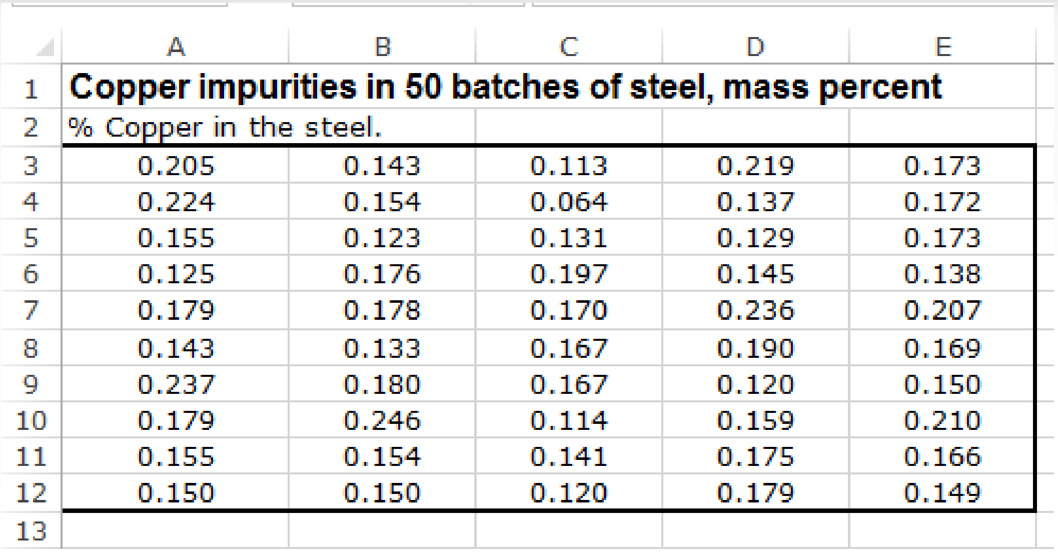
Let apply these parameters into an example

Copper impurities in 50 batches of steel, mass percent

0.205	0.143	0.113	0.219	0.173
0.224	0.154	0.064	0.137	0.172
0.155	0.123	0.131	0.129	0.173
0.125	0.176	0.197	0.145	0.138
0.179	0.178	0.170	0.236	0.207
0.143	0.133	0.167	0.190	0.169
0.237	0.180	0.167	0.120	0.150
0.179	0.246	0.114	0.159	0.210
0.155	0.154	0.141	0.175	0.166
0.150	0.150	0.120	0.179	0.149

Minimum, Maximum, Mean(Average), Median, Mode, Range
Variance, Standard Deviation, Frequency distribution and Histogram

We can use Microsoft$^{\text{TM}}$ Excel® program for calculation of statistical values.



	A	B	C	D	E
1	**Copper impurities in 50 batches of steel, mass percent**				
2	% Copper in the steel.				
3	0.205	0.143	0.113	0.219	0.173
4	0.224	0.154	0.064	0.137	0.172
5	0.155	0.123	0.131	0.129	0.173
6	0.125	0.176	0.197	0.145	0.138
7	0.179	0.178	0.170	0.236	0.207
8	0.143	0.133	0.167	0.190	0.169
9	0.237	0.180	0.167	0.120	0.150
10	0.179	0.246	0.114	0.159	0.210
11	0.155	0.154	0.141	0.175	0.166
12	0.150	0.150	0.120	0.179	0.149
13					

To find the minimum value
"=MIN(A3:E12)"

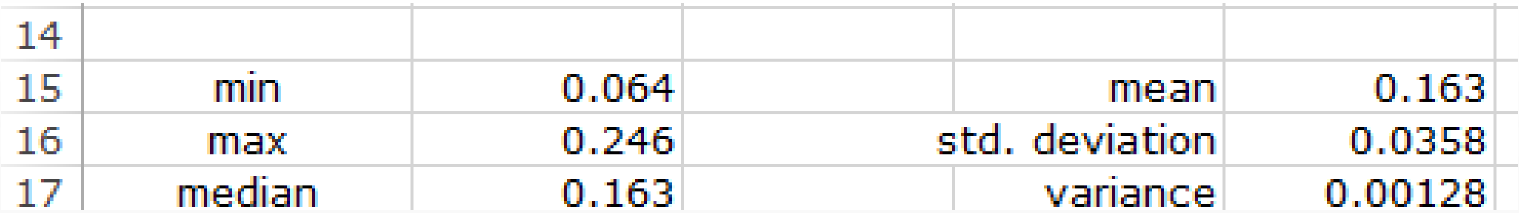
To find the maximum value
"=MAX(A3:E12)"

To find the mean value
"=AVERAGE(A3:E12)"

To find the variance
"=VAR(A3:E12)"

To find the median value
"=MEDIAN(A3:E12)"

To find the standard deviation
"=STDEV(A3:E12)"



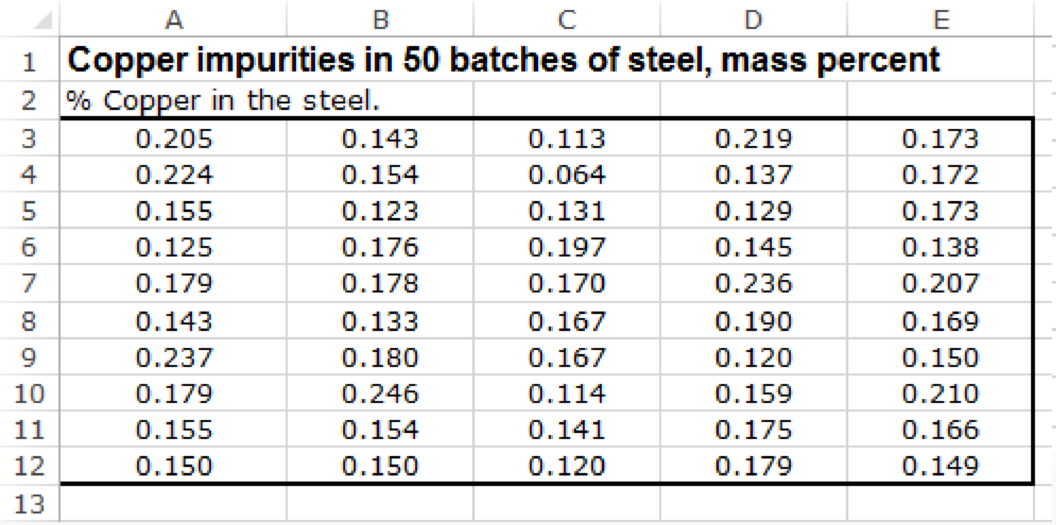
14					
15	min	0.064		mean	0.163
16	max	0.246		std. deviation	0.0358
17	median	0.163		variance	0.00128

For Frequency distributions we should decide sub-ranges. So we need to find number of bins.

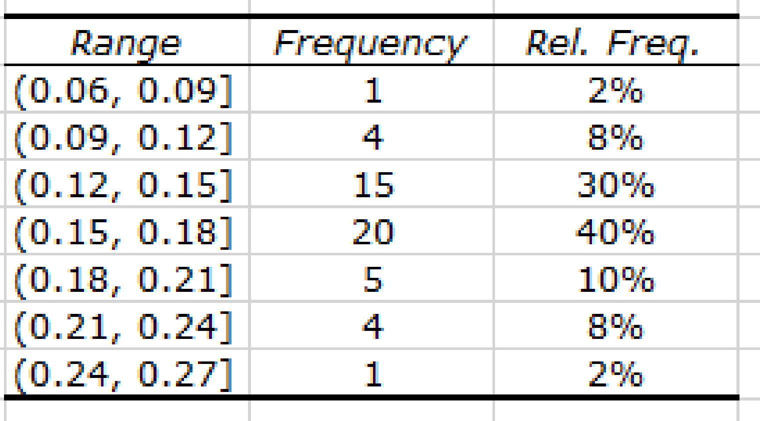
For this example, max:0.246 min:0.064 range:0.182 #of sample:50

$$\#\ of\ bins = \log_2 \#\ of\ sample = \log_2 50 = 5.64 \approx 6$$

$$sub - range = \frac{range}{\#\ of\ bins} = \frac{0.182}{6} \approx 0.030$$

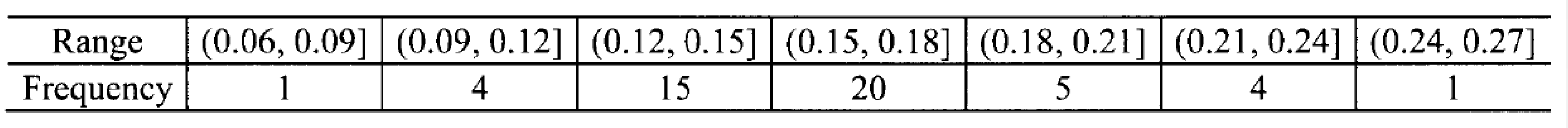


	A	B	C	D	E
1	**Copper impurities in 50 batches of steel, mass percent**				
2	% Copper in the steel.				
3	0.205	0.143	0.113	0.219	0.173
4	0.224	0.154	0.064	0.137	0.172
5	0.155	0.123	0.131	0.129	0.173
6	0.125	0.176	0.197	0.145	0.138
7	0.179	0.178	0.170	0.236	0.207
8	0.143	0.133	0.167	0.190	0.169
9	0.237	0.180	0.167	0.120	0.150
10	0.179	0.246	0.114	0.159	0.210
11	0.155	0.154	0.141	0.175	0.166
12	0.150	0.150	0.120	0.179	0.149
13					

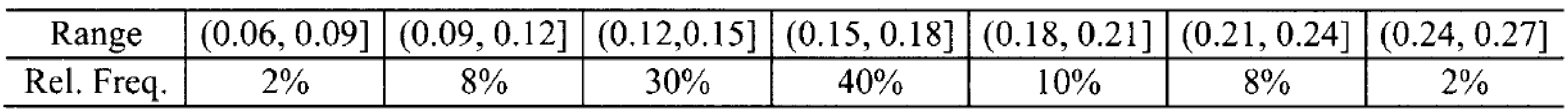


Range	*Frequency*	*Rel. Freq.*
(0.06, 0.09]	1	2%
(0.09, 0.12]	4	8%
(0.12, 0.15]	15	30%
(0.15, 0.18]	20	40%
(0.18, 0.21]	5	10%
(0.21, 0.24]	4	8%
(0.24, 0.27]	1	2%

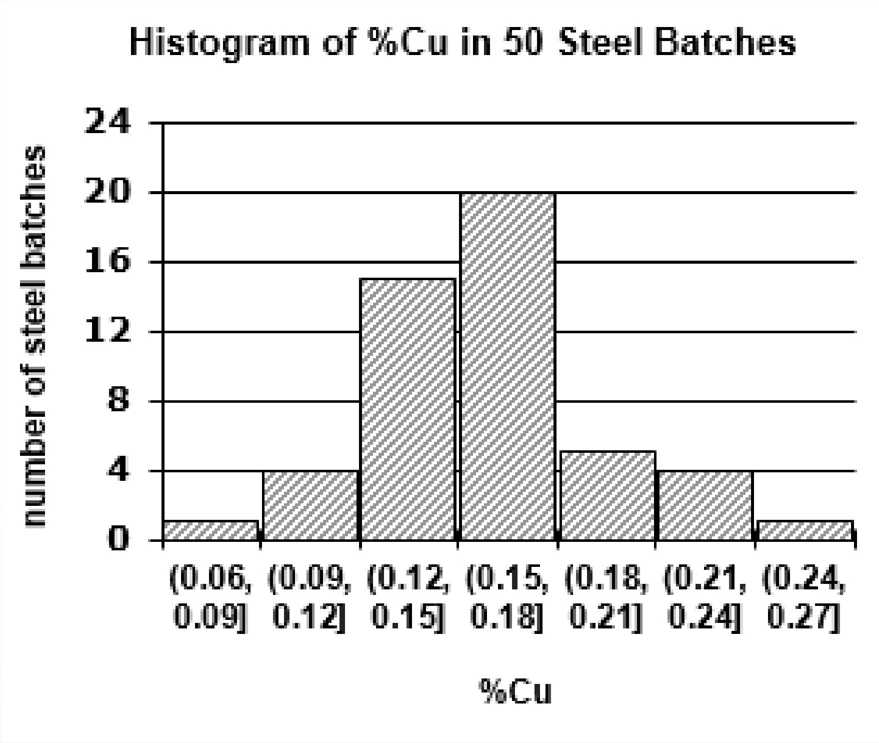
From Frequency distributions we can plot histogram diagram.

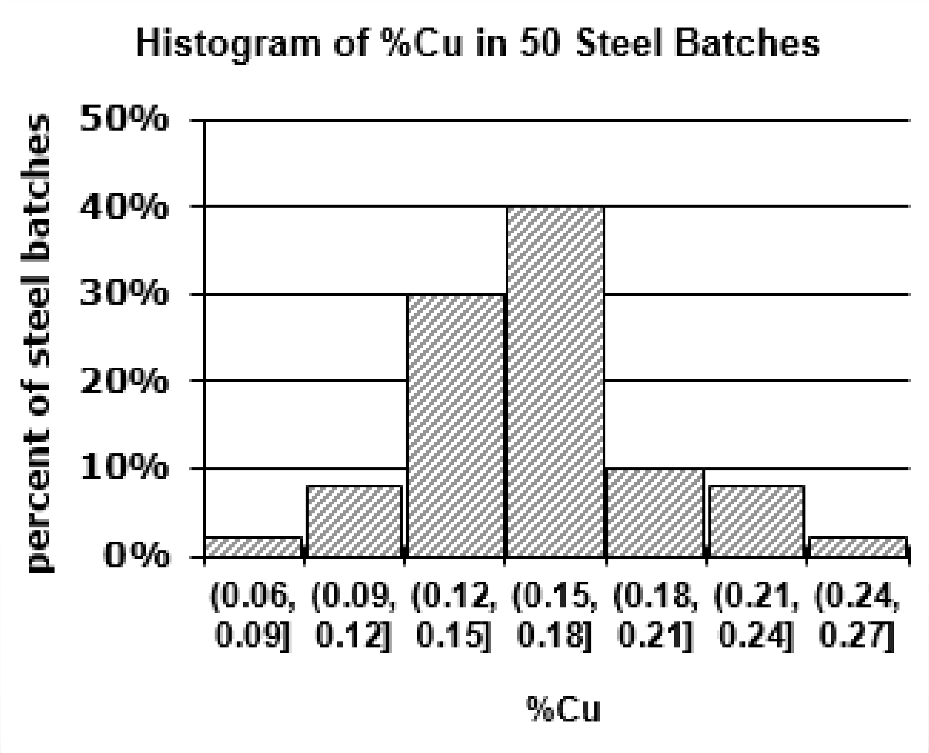


Range	(0.06, 0.09]	(0.09, 0.12]	(0.12, 0.15]	(0.15, 0.18]	(0.18, 0.21]	(0.21, 0.24]	(0.24, 0.27]
Frequency	1	4	15	20	5	4	1



Range	(0.06, 0.09]	(0.09, 0.12]	(0.12,0.15]	(0.15, 0.18]	(0.18, 0.21]	(0.21, 0.24]	(0.24, 0.27]
Rel. Freq.	2%	8%	30%	40%	10%	8%	2%





Median and Percentile

- The *median* separates the bottom 50% of the data from the top 50% of the data (ascending sort).
- If we want to consider lower 40% (40th percentile),
- Or, if we want to know, how many sample readings are under the any given number (for this example 0.179)
- how do we calculate these?

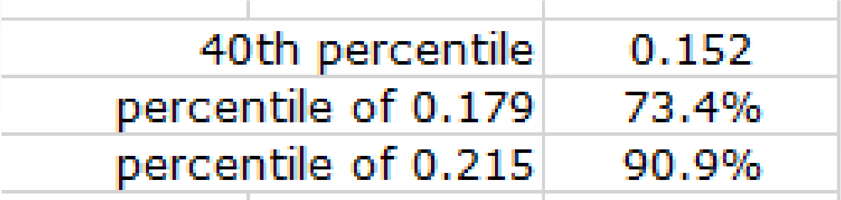
We can use **PERCENTILE** function and **PERCENTRANK** function.

for the %Cu data is

=PERCENTILE(A3:E12, 0.4) = 0.152.

=PERCENTRANK(A3:E12, 0.179, 3) = 73.4%

Median and Percentile



40th percentile	0.152
percentile of 0.179	73.4%
percentile of 0.215	90.9%

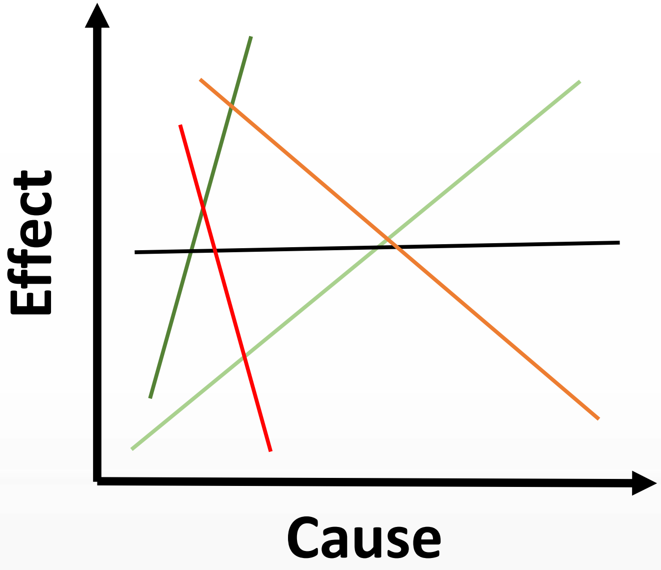
Looking at the actual numbers, place 20 contains 0.150, while place 21 contains 0.154. Thus, the 40th percentile falls between these two numbers, to separate the lower 20 from the upper 30 measurements.

There are 36 measurements in the %Cu dataset smaller than 0.179, and *36/50*= 72%.

PERCENTRANK still works even if the value given to it is not in the array. For example, the %Cu data contains no measurement of 0.215, but the formula =PERCENTRANK(A3:E12, 0.215) still returns a percentage between the percentages for 0.210 and 0.219, both of which are in the dataset.

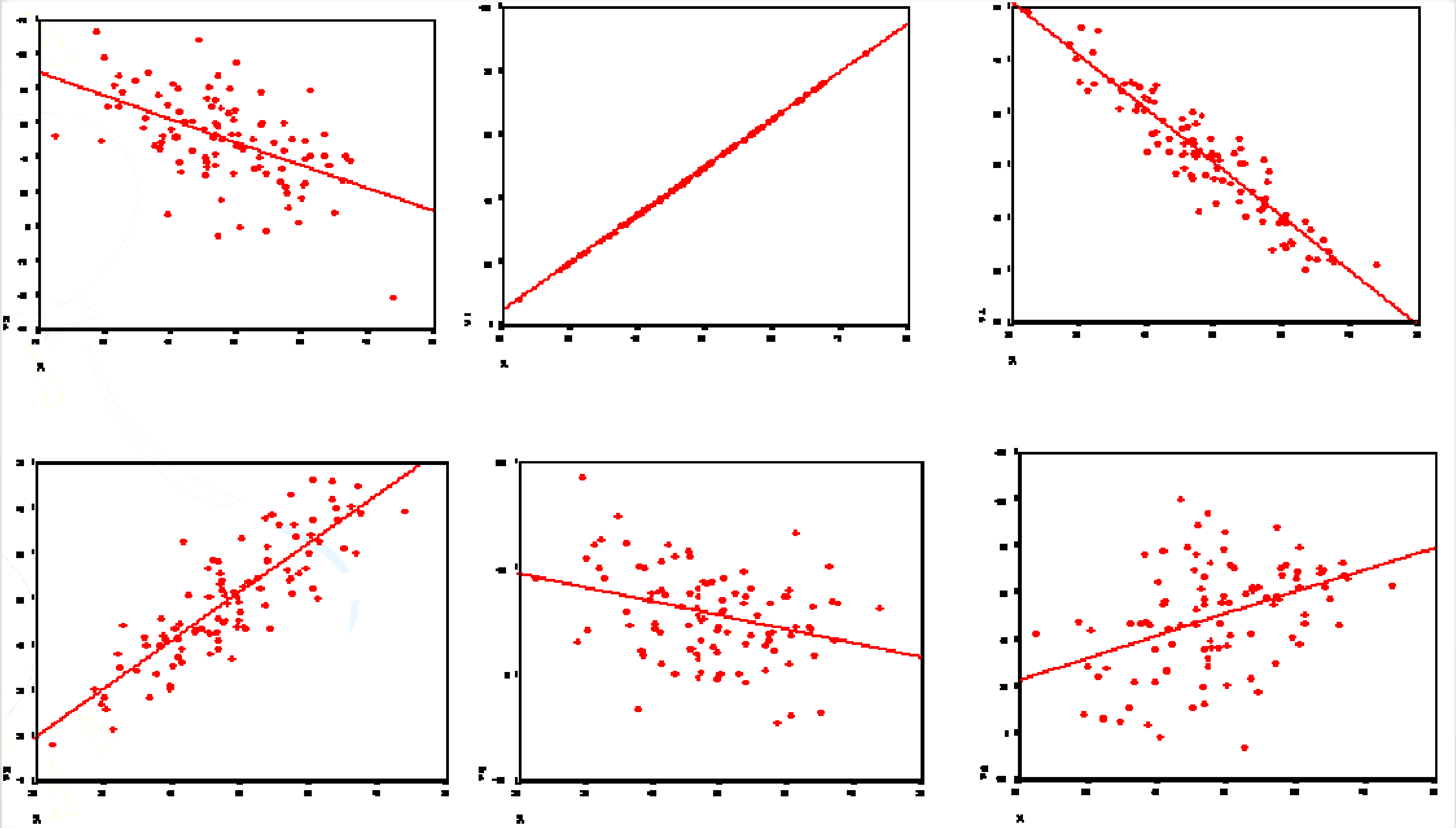
Scatter diagram

- A scatter diagram is a visualization of the relationship between two variables measured on the same set of individuals. This relationship may or may not depend on any cause and effect relation.
- The linear line of this relation may be negative or positive tendency. Or there may not be any relation between two variables.



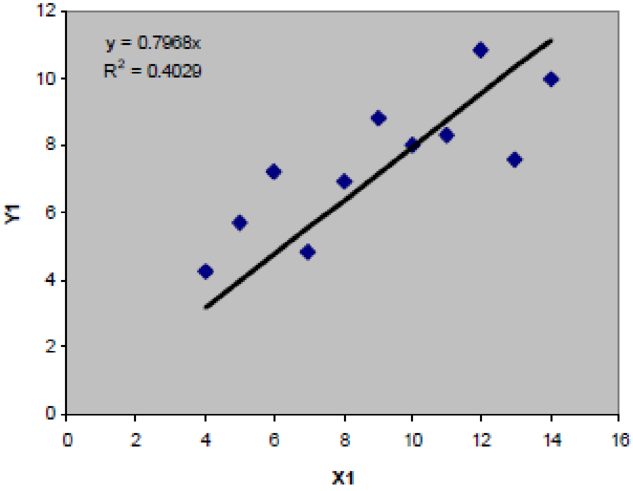
Strongly Positive relation
Positive relation
No relation
Negative relation
Strongly Negative relation

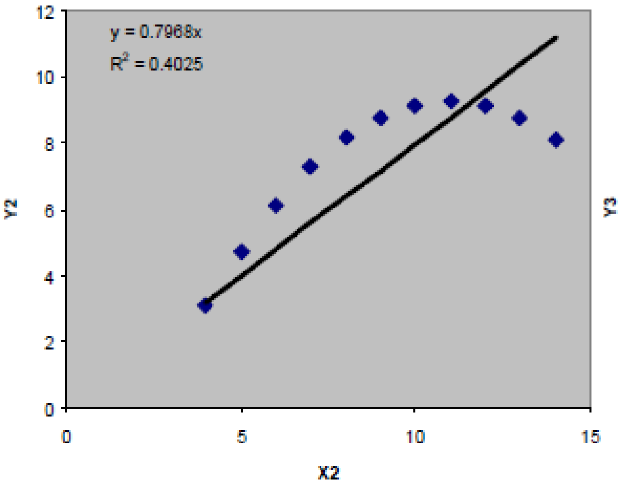
Scatter diagram

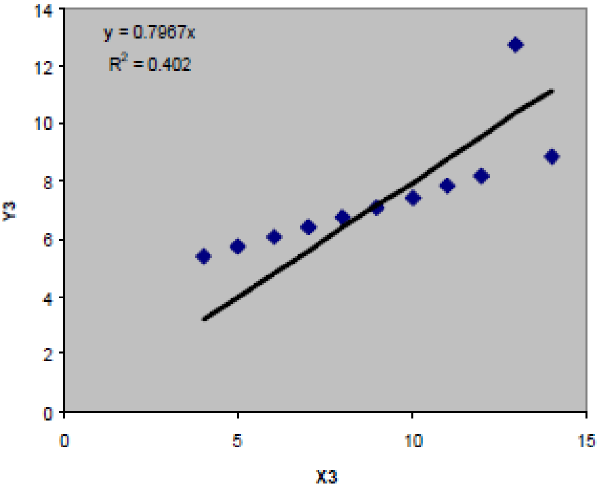


Scatter diagram

- A scatter diagram doesn’t always give the rights for statistical analysis. Sometimes it can mislead us.

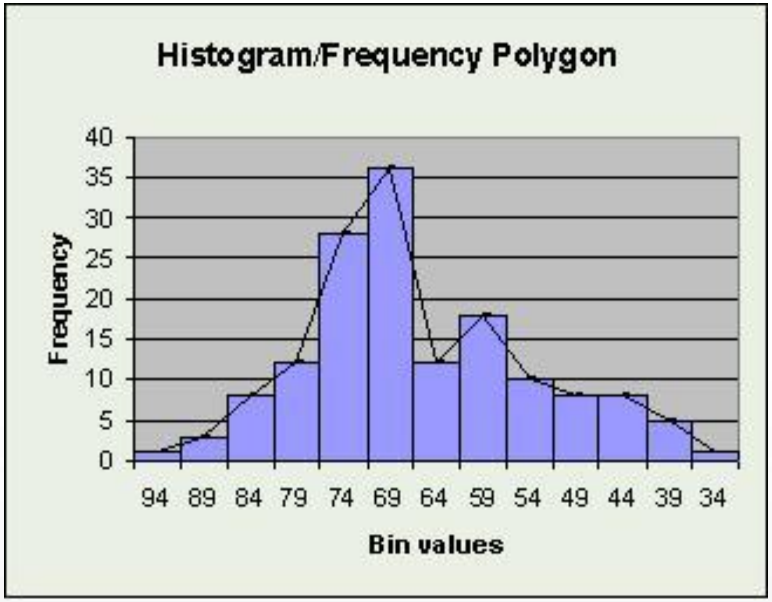






Frequency Polygon

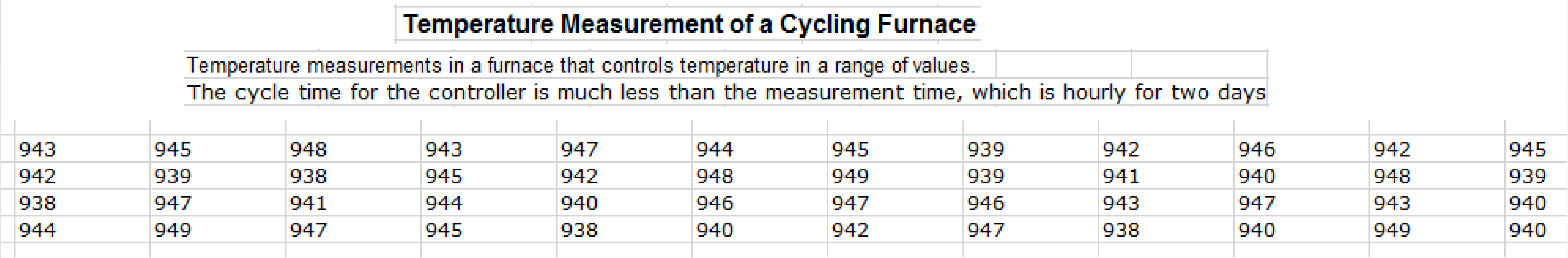
- The frequency polygons derived from a histogram by connecting the mid-points of the tops of the rectangles in the histogram.
- The line connecting the centers of histogram rectangles is called frequency polygon.
- A special type of frequency polygon is the Normal Distribution Curve



Distribution of Random Values

- Statisticians use the word distribution in two senses. The first use is qualitative, and describes the features of a dataset displayed in a histogram. The second use of the word distribution is more precise. Here, a distribution is a mathematical formula whose plotted shape or pattern serves as an idealized model for the shape or pattern seen in a histogram of a sample variable.
- The mathematical formulas of these idealized distributions will let us calculate the probability that a given measurement falls into a given range of values. For our purposes, we can use a very informal definition of probability or likelihood: the probability or likelihood of a random event is the percentage of time that event is expected to occur.
- We will describe two of the most important mathematical distributions for engineering purposes, the *uniform* and the *normal* distributions.

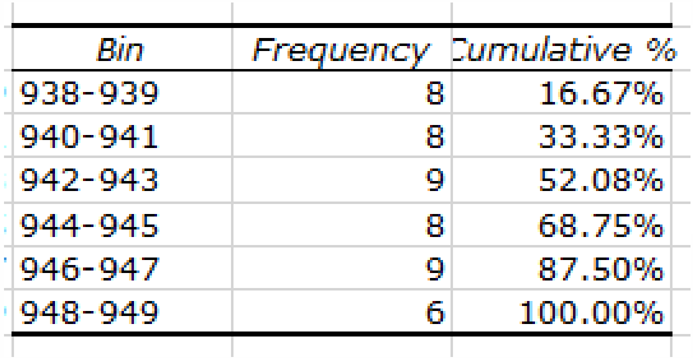
The Uniform Distribution



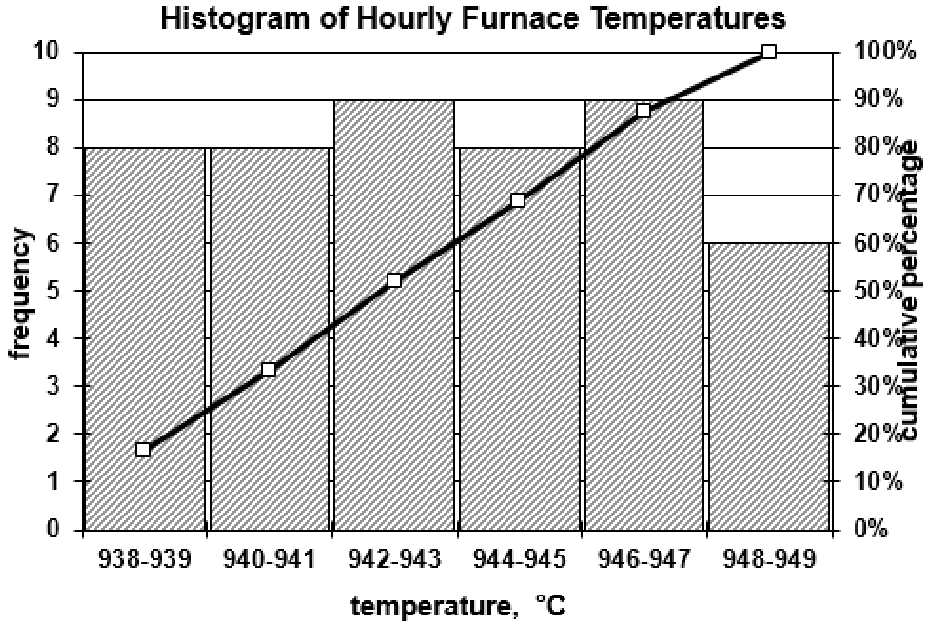
Temperature Measurement of a Cycling Furnace

Temperature measurements in a furnace that controls temperature in a range of values.
The cycle time for the controller is much less than the measurement time, which is hourly for two days

943	945	948	943	947	944	945	939	942	946	942	945
942	939	938	945	942	948	949	939	941	940	948	939
938	947	941	944	940	946	947	946	943	947	943	940
944	949	947	945	938	940	942	947	938	940	949	940

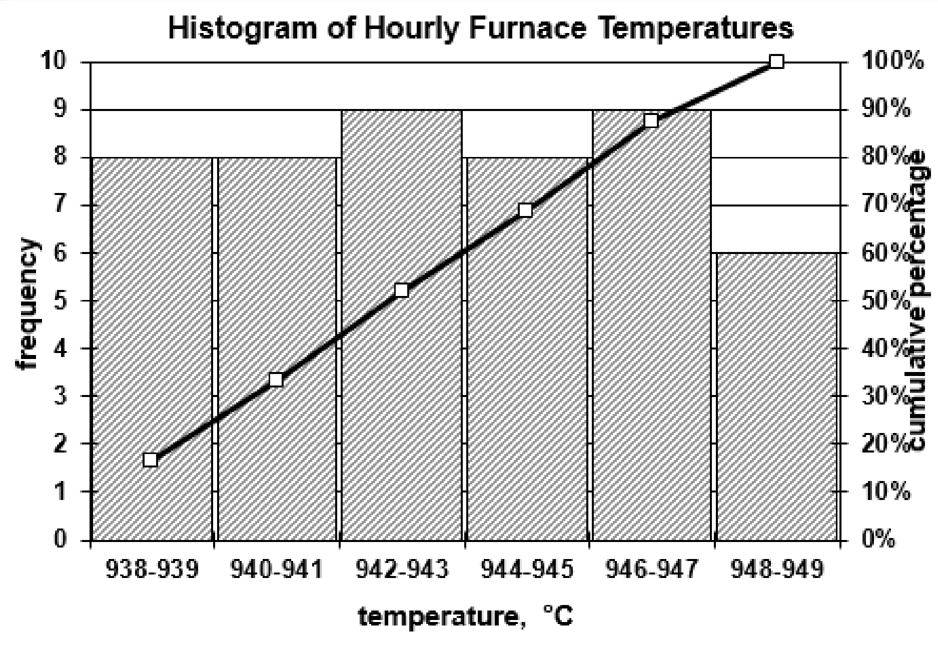


Bin	*Frequency*	*Cumulative %*
938-939	8	16.67%
940-941	8	33.33%
942-943	9	52.08%
944-945	8	68.75%
946-947	9	87.50%
948-949	6	100.00%



The *uniform distribution* is the mathematical distribution that models situations like this, where a randomly chosen member of the population is equally likely to be found in any part of the distribution.

The Uniform Distribution



This line called the *cumulative percentage line*, which is based on the Cumulative % column in the frequency distribution.

The plotted function is the *cumulative distribution function (cdf)*, usually denoted *F(x)*, and is an idealization of the cumulative percentage line.

While the cumulative percentage for a bin is the fraction of the data that lie in or below that bin, F(x) is defined as the probability that a randomly chosen element of the population will be ≤ x.

The cdf — the idealized cumulative percentage line — should be a straight line. To see which straight line, we just have to think about what it is supposed to mean.

The Uniform Distribution

If the smallest possible temperature is 937.5 °C, then F(937.5) would be the percentage of the data that was smaller than 937.5 °C, which is 0%.

Similarly, the highest possible temperature is 949.5 °C, so F(949.5) would be the percentage of data that was smaller than 949.5, which is 100%, or 1.

Some algebra shows you that the formula of that function is
$F(x) = (100/\text{range})(x - 937.5)$.

Let’s consider finding the probability between 940 °C and 948 °C.
≤ 948 °C is $(100/12)(948 - 937.5) = 87.50\ \%$
≤ 940 °C is $(100/12)(940 - 937.5) = 20.83\ \%$

$$\Pr(940 \leq x \leq 948) = \Pr(x \leq 948) - \Pr(x \leq 940) = F(948) - F(940) = 66.67\ \%$$

$$\Pr(x < 940 \text{ or } x > 948) = 1 - \Pr(940 \leq x \leq 948) = 100\% - 66.67\% = 33.33\%$$

The Uniform Distribution

The second function we will use is the *probability density function (pdf)*, denoted as *f(x)*. The pdf is the derivative of the cdf, so for the temperature example:

$$f(x) = \frac{d}{dx}\left(\frac{x - 937.5}{12}\right) = \frac{1}{12}$$

$$Pr(940 \leq x \leq 948) = F(948) - F(940) = \int_{940}^{948} f(x)dx = \int_{940}^{948} \frac{1}{12}dx = 0.667$$

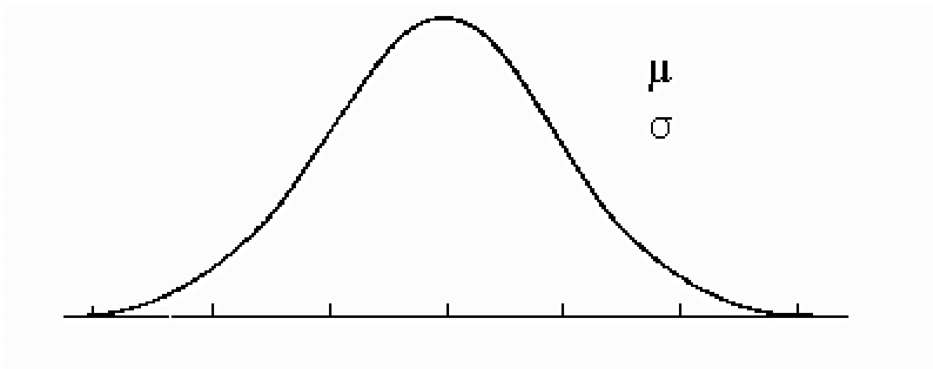
The Normal Distribution

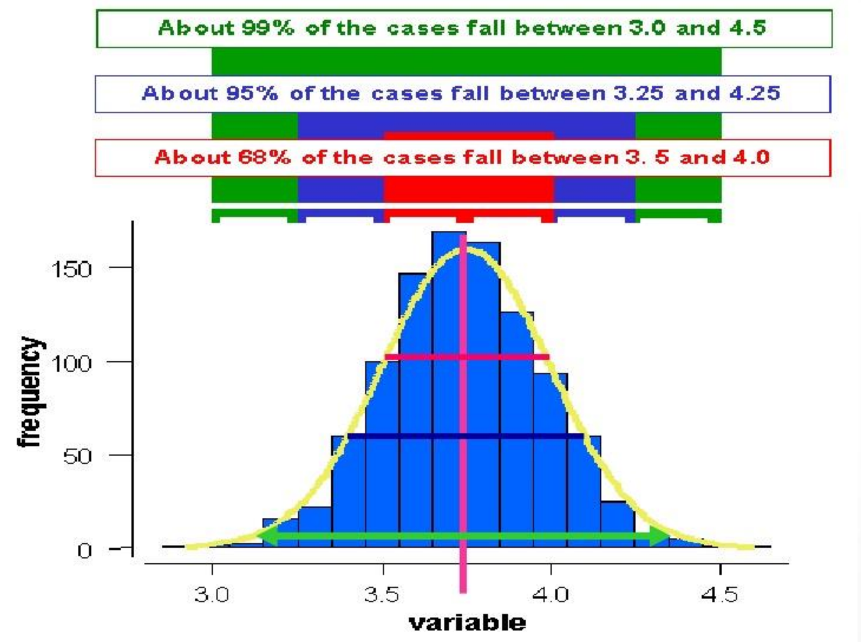
- The most important statistical distribution is the **normal distribution**, often called the "bell curve".
- The normal distribution is a useful model in any situation where the measurements are symmetric around some center point, and values near the center are more likely than values at the extremes.
- Normal distributions can have any mean. Since the (ideal) normal distribution is symmetric, the mean will always lie at the center of the distribution.
- Normal distributions can also have any standard deviation. The greater the standard deviation, the greater the average distance from the center. In terms of a normal curve, that means that the standard deviation governs how wide the central "hump" is.
- The larger the standard deviation, the wider the central hump.

The Normal Distribution

- A normal distribution pdf (*probability density function*)with mean μ and standard deviation σ is:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$





Characteristics of Normal Distribution Curve

- It is *bell* shaped, *continuous* curve.
- It is *symmetrical* can be divided into two equal halves vertically.
- The *tails never touch the base line* but extended to infinity in either direction.
- The mean, median and mode values coincide
- It is described by two parameters: **arithmetic mean** determine the location of the center of the curve and **standard deviation** represents the scatter around the mean.

In symmetrical NDC,

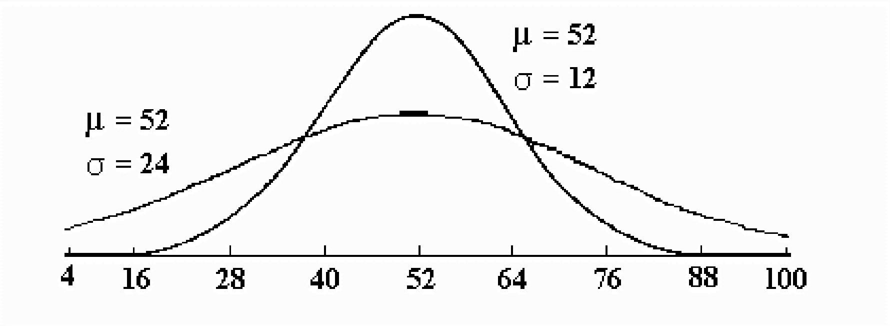
Average(mean) = Mode = Median

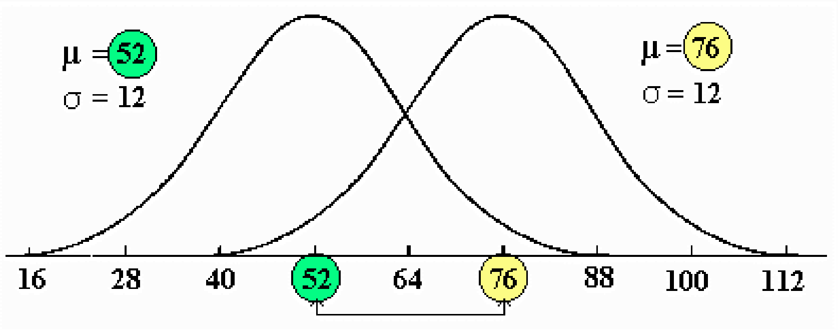
65 % of values in between $\mu \pm \sigma$

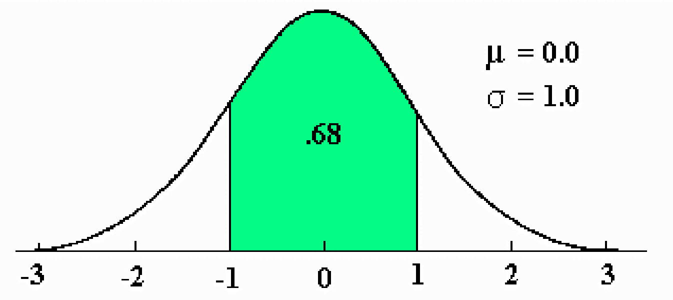
95 % of values in between $\mu \pm 2\sigma$

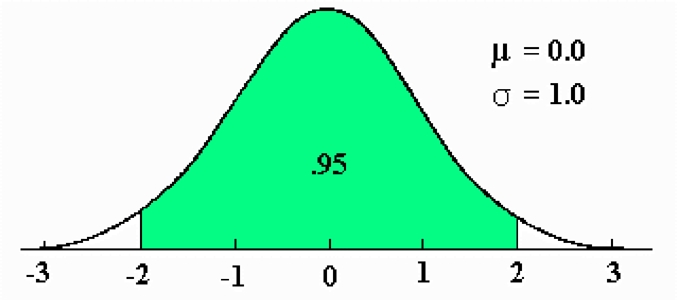
99 % of values in between $\mu \pm 3\sigma$

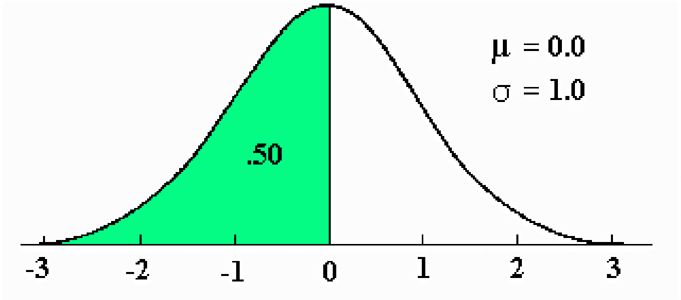
The Normal Distribution Curve



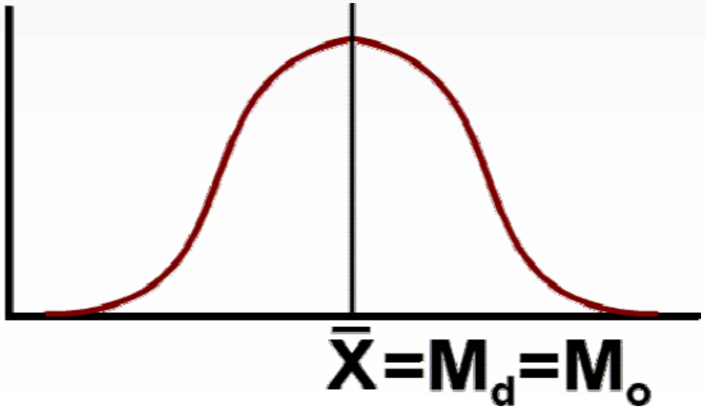




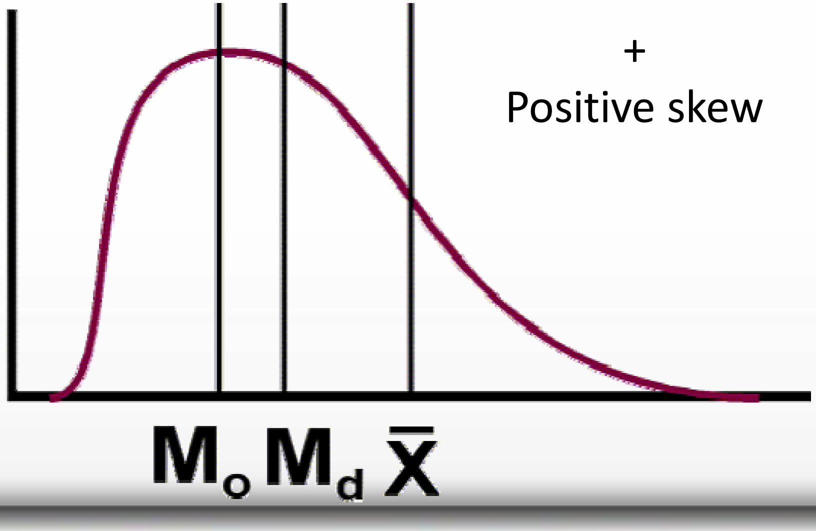




The Symmetry and Asymmetry of NDC

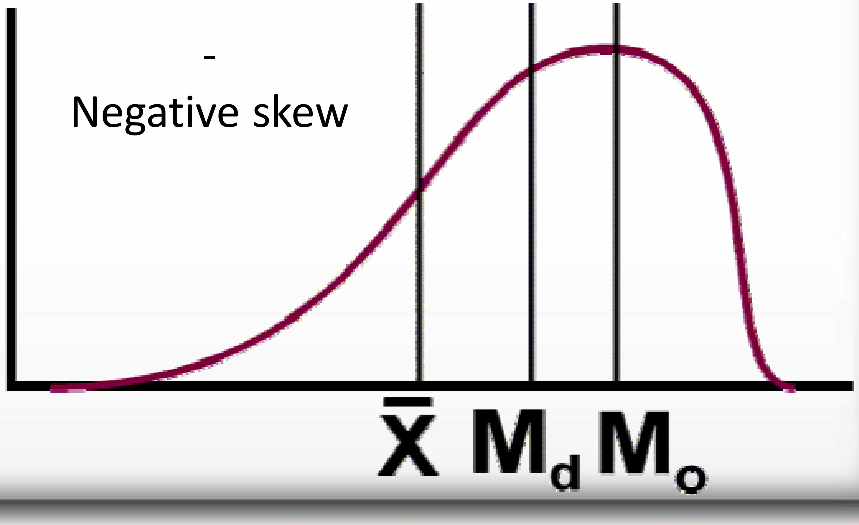


Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. The skewness value can be positive or negative, or even undefined. (M_o = mode, M_d = median, X = mean)



+

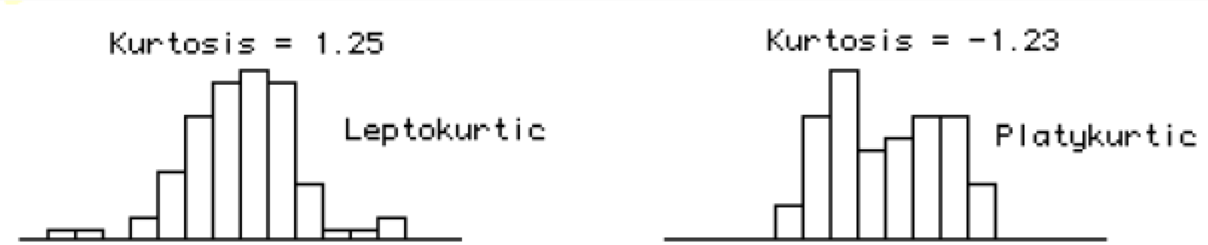
Positive skew

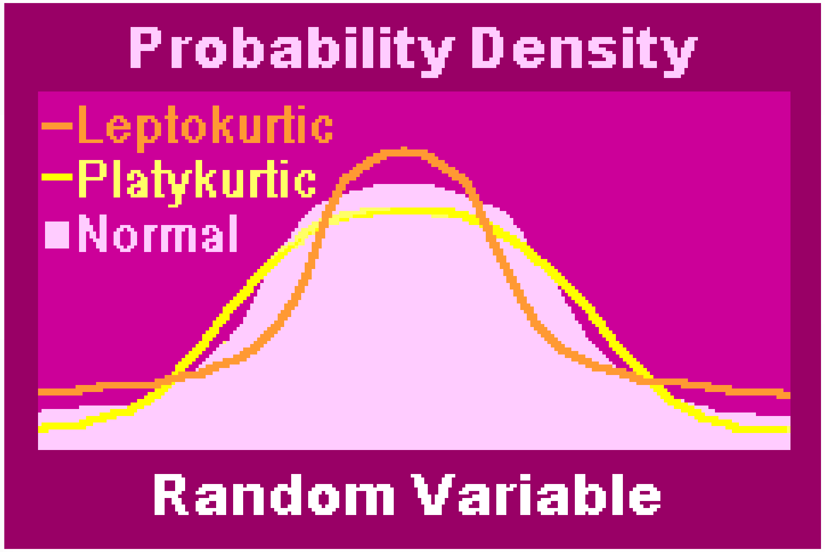


-

Negative skew

The Symmetry and Asymmetry of NDC





a leptokurtic distribution has fatter tails

a platykurtic distribution has thinner tails

- In a similar way to the concept of skewness, *kurtosis* is a descriptor of the shape of a probability distribution
- Distributions with zero excess kurtosis are called **mesokurtic.**
- A distribution with positive excess kurtosis is called **leptokurtic**.
- A distribution with negative excess kurtosis is called **platykurtic**.

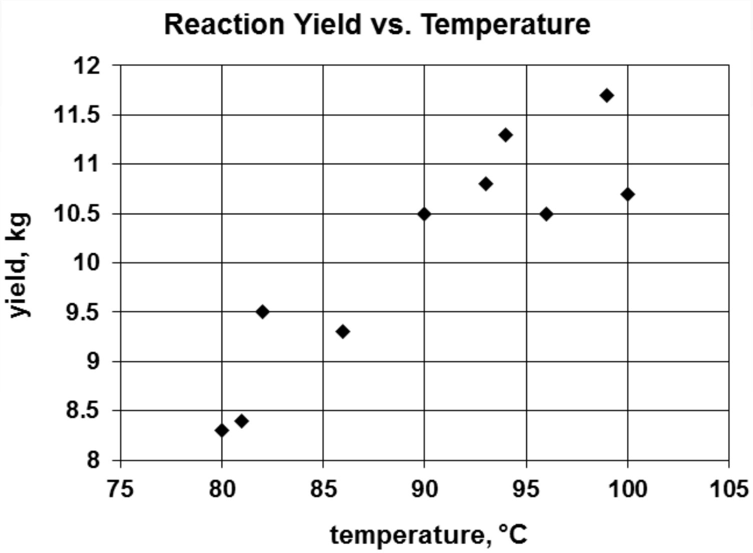
Curve Fitting

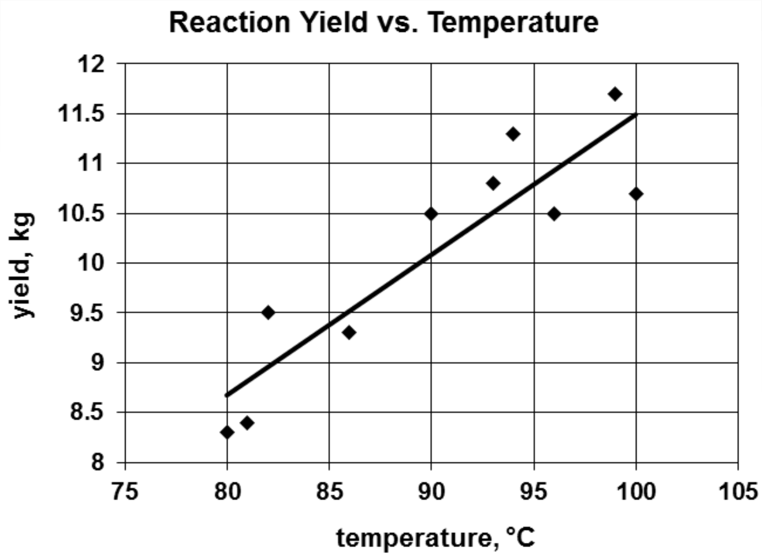
Yield of Reaction Product as a Function of Temperature										
Temperature (°C)	80	86	100	82	90	99	81	96	94	93
Yield (kg)	8.3	9.3	10.7	9.5	10.5	11.7	8.4	10.5	11.3	10.8

- We have measured yield at selected temperature values.
- Can we predict the yield values at different temperatures?
- If we can, how ?

- The measurements are one yield for each temperature.
- Rather than try to understand how a single variable is behaving, our goal is to determine if a functional relationship exists between two variables, and if so, what it is.
- How do we find a formula that predicts the yield for a given a process temperature?

Curve Fitting





- First, to use Excel to chart the data (x-y scatter plot), contains the points from table as well as the regression line for this dataset.
- By tilting the slope or shifting, the line can be closer to some points but makes it farther away from others.
- Although the regression line in this example doesn't pass through any points of the sample dataset, it seems to be a good "middle of the road" line.
- We will be seeking ways to find lines and curves that are at the *smallest possible average distance* from the measured points.

Curve Fitting

- The goal is *given a set of measurements for which we are trying to find a mathematical relationship or formula, we seek a formula that makes* ***the sum of the squared differences*** *between the measurements and the formula’s predicted values as small as possible*.
- Statisticians prefer to measure “how data varies from the center”, or “how it deviates from an ideal value” using the squared distance rather than absolute distance.
- "the sum of the squared differences between the measurements and the formula's predicted values", we usually abbreviate this as ***the sum of squares***.
- Although we will always let Excel calculate regression formulas for us, calculations to find the slope and intercept of a regression line are not difficult:

$$\text{Slope } m = \frac{n\sum_{i=1}^{n}(x_i y_i) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n}\left(x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$\text{Intercept } b = \frac{\sum_{i=1}^{n} y_i - m\sum_{i=1}^{n} x_i}{n}$$

$$\mathbf{y = m \cdot x + b}$$

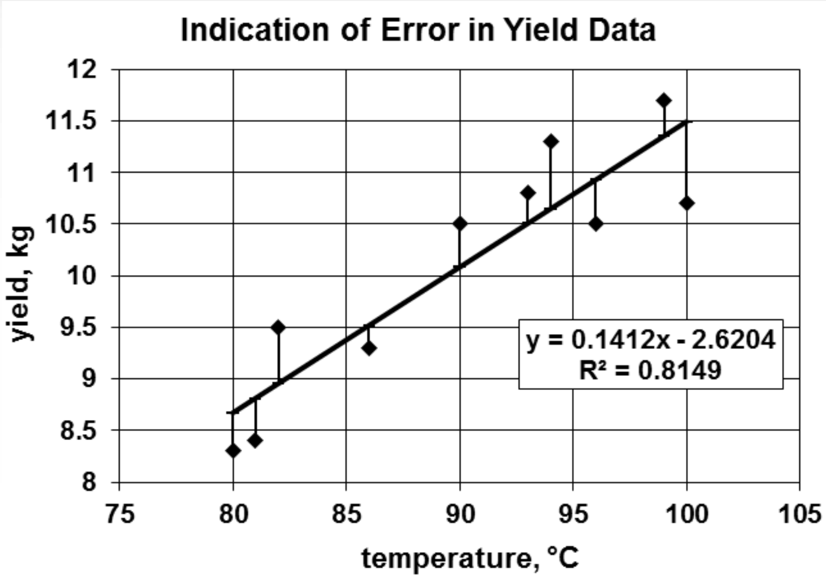
Results from Excel's SLOPE and INTERCEPT functions:

Exact slope = m = 0.14118 Exact Intercept = b = -2.6204

Yield, kg = 0.141 (temperature, °C) – 2.62

Yield of Reaction Product as a Function of Temperature										
Temperature (°C)	80	86	100	82	90	99	81	96	94	93
Yield (kg)	8.3	9.3	10.7	9.5	10.5	11.7	8.4	10.5	11.3	10.8
Yield from formula	8.67	9.52	11.50	8.96	10.09	11.36	8.82	10.93	10.65	10.51
Difference	-0.37	-0.22	-0.80	0.54	0.41	0.34	-0.42	-0.43	0.65	0.29
Difference squared	0.140	0.049	0.636	0.295	0.171	0.118	0.172	0.187	0.422	0.084

The sum of squared distances (SSD) = 2.28



- The error is defined to be the difference between the measured value and the true value.
- R^2 is the coefficient of determination. R^2 will always be a number between 0 and 1, and the closer it is to 1.
- We now introduce some terminology to explain what R^2 means and how to interpret it.

Calculation of R^2

- R^2, the coefficient of determination, is defined as:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Where SSR (*regression sum of squares*), SST(*total sum of squares*), SSE(*sum of squared errors*). The word "error“ is used here to indicate the difference between the "true" measured yield and the yield resulting from the regression equation.

SST is the sum of all the squared differences between individual yields y_i and their mean $\bar{y}$. SST is often described as the *total variation in y*.

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad SSE = \sum_{i=1}^{n} \big(y_i - (m \cdot x_i + b)\big)^2$$

$$SSR = SST - SSE$$

Next Week

- **Fundamentals of Material Balances with Applications to Non-Reacting Systems**

- **HW Assignments**