SPC & Mahalanobis Distances

Prasanta Chandra Mahalanobis

Robert J McMahon April 24, 2022

BROADCOM

Definitions

- Statistical process control (SPC) is the use of statistical techniques to control a process or production method
- Distance metrics use a distance function that tells us the mathematically driven distance between various elements throughout a data set. Closer the distance, the more similar they are and vice-versa
- The Mahalanobis distance is a measure of the distance between a point (or a cluster centroid) and a distribution
- The Mahalanobis distance differs from the Euclidean distance per taking in account covariance or correlation of the metrics being measured
- This distance is an important aspect of SPC including, though not limited to, outlier detection

SPC Motivations (some but not all)

- Need a mathematically, rigorous, statistical process control (SPC) mechanism so machines can quickly detect and signal “defects” or unexpected changes
- Engineers can use SPC to make sure system or unit level test results are “equal” across developer commits, across nightly builds, across test rigs, etc.
- Supports multivariate analysis (as complex systems have multiple independent & dependent variables)
- Mahalanobis Distances are a distance metric used by SPC

WiFi possible metric examples

(Extend from mostly single traffic stream, peak average throughput)

- Multiple concurrent iperf traffic streams
- Individual throughputs
- Aggregate throughput
- Network Power
- Packet latency for latency sensitive flows
- Frame latency for realtime video flows
- Bufferbloat (size and duration)
- Write to read latency (TCP)
- Memory consumption (TCP congestion window, end/end bytes in flight)
- Energy consumed (e.g. using a Monsoon current meter)
- WiFi Mu/Su ratios, MCS, aggregation, etc.
- Airtime fairness metrics
- Packet loss
- Packets per second with small packets
- TCP connect times

Mahalanobis Distance: Matrix Math Formula

$$D^2 = \left(X_{p_1} - X_{p_2}\right)^T . C^{-1} . \left(X_{p_1} - X_{p_2}\right)$$

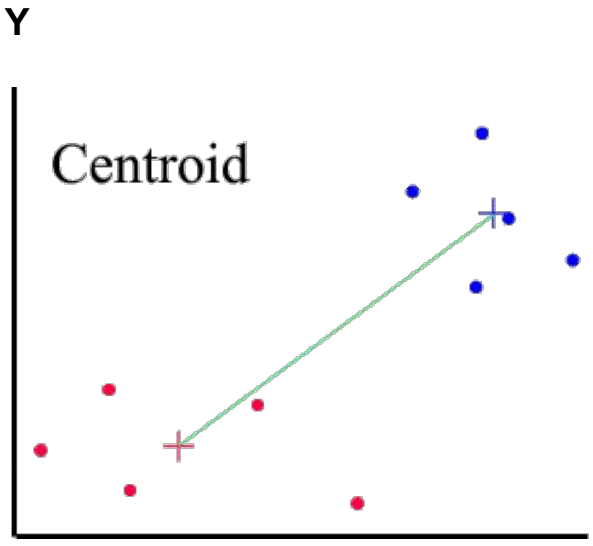
D = Mahalanobis distance
C = Covariance matrix
X_{p2} = Centroid or multivariate mean
X_{p1} = Observation
$(X_{p1} - X_{p2})$ = the vector of deviations of the observation from the multivariate mean

BROADCOM

Math problem more than a coding problem (python code two sample test)

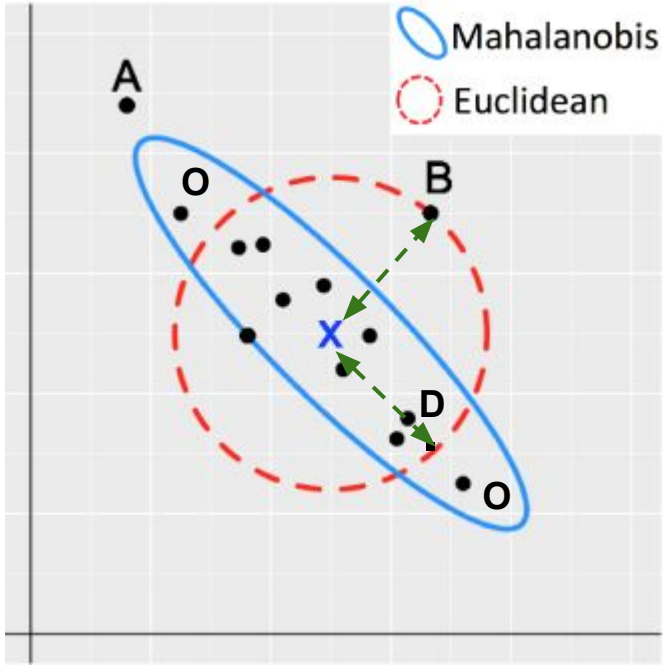
```
nx, p = X.shape
ny, _ = Y.shape
delta = np.mean(X, axis=0) - np.mean(Y, axis=0)
Sx = np.cov(X, rowvar=False)
Sy = np.cov(Y, rowvar=False)
S_pooled = ((nx-1)*Sx + (ny-1)*Sy)/(nx+ny-2)
t_squared = (nx*ny)/(nx+ny) * np.matmul(np.matmul(delta.transpose(),np.linalg.inv(S_pooled)), delta)
statistic = t_squared * (nx+ny-p-1)/(p*(nx+ny-2))
```


Euclidean vs Mahalanobis distances



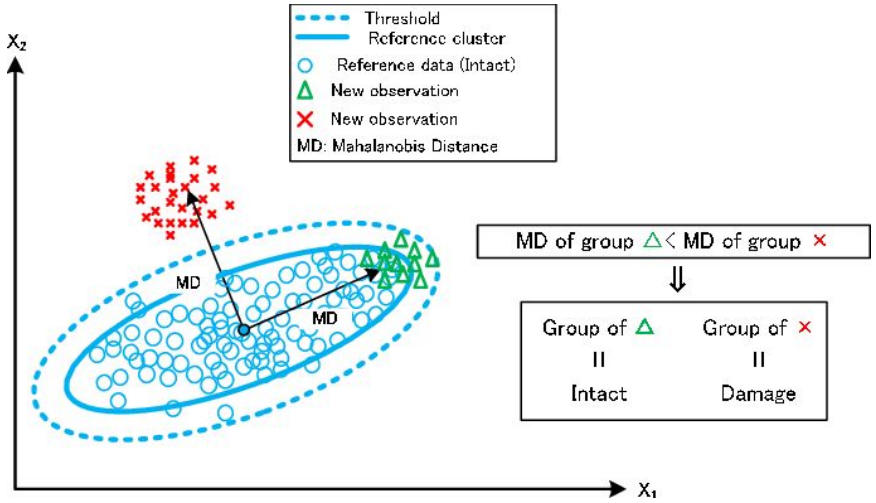
$$\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}.$$

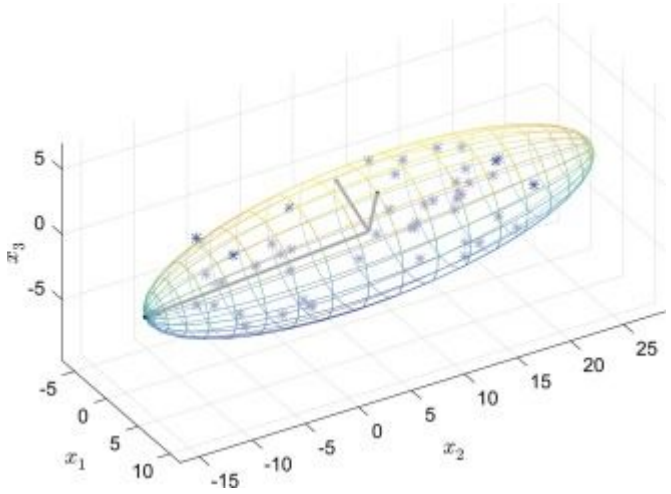
Is $\overline{DX} = \overline{BX}$? Is O an outlier?



BROADCOM

Visualizations (2D & 3D)



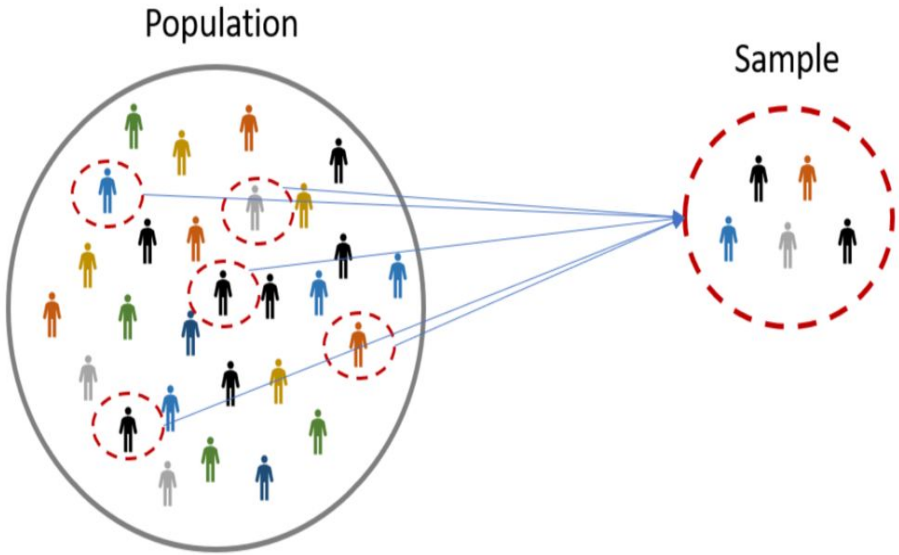


BROADCOM

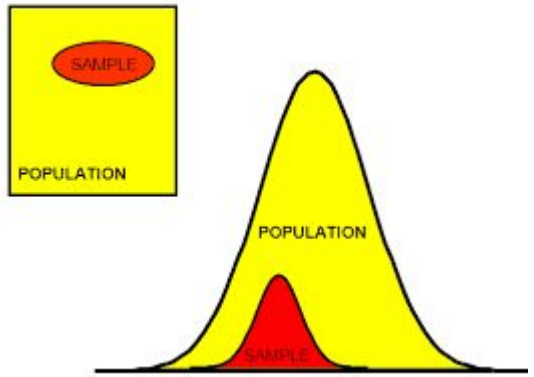
Stats review

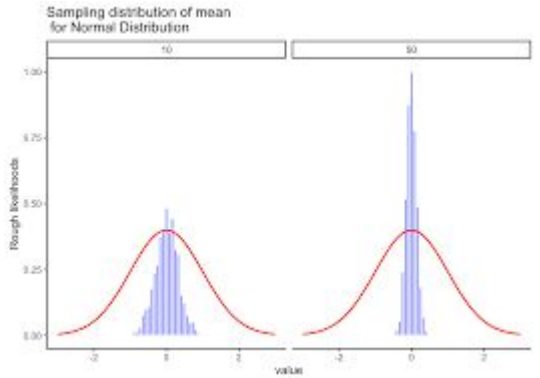
BROADCOM

Population vs Sample (or Subgroup)



Population & Sample distributions





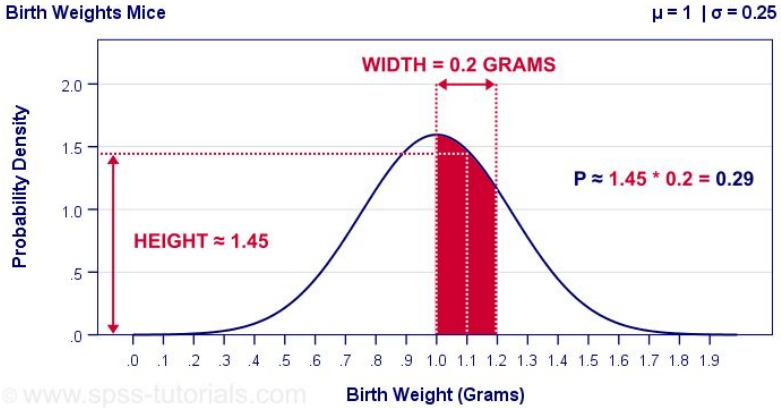
BROADCOM

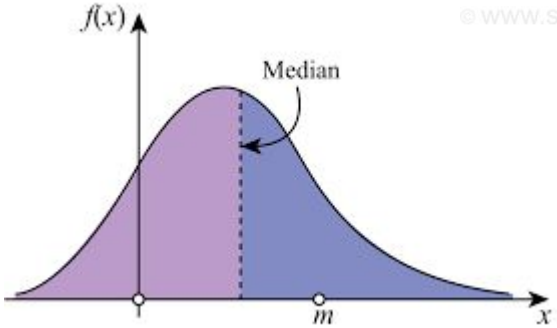
Single Variate

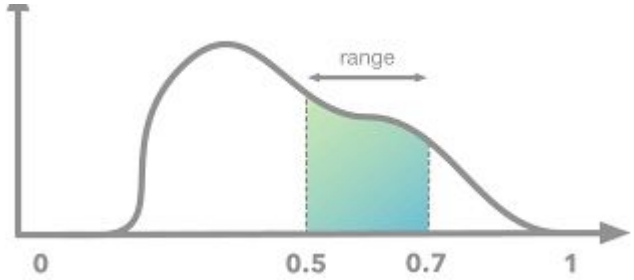
BROADCOM

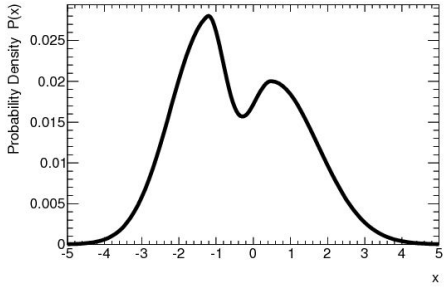
Probability Density Function (PDF)

1. A function of a continuous random variable, whose integral across an interval gives the probability that the value of the variable lies within the same interval.



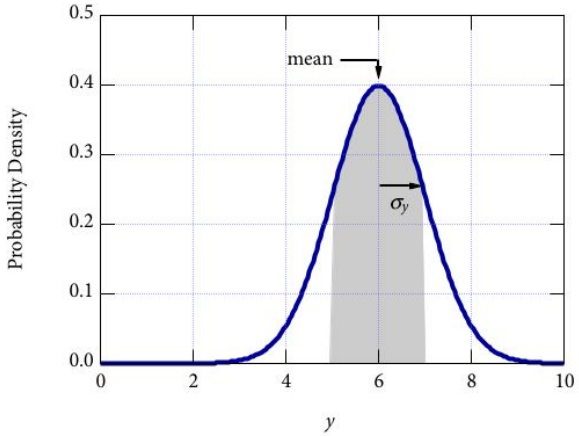


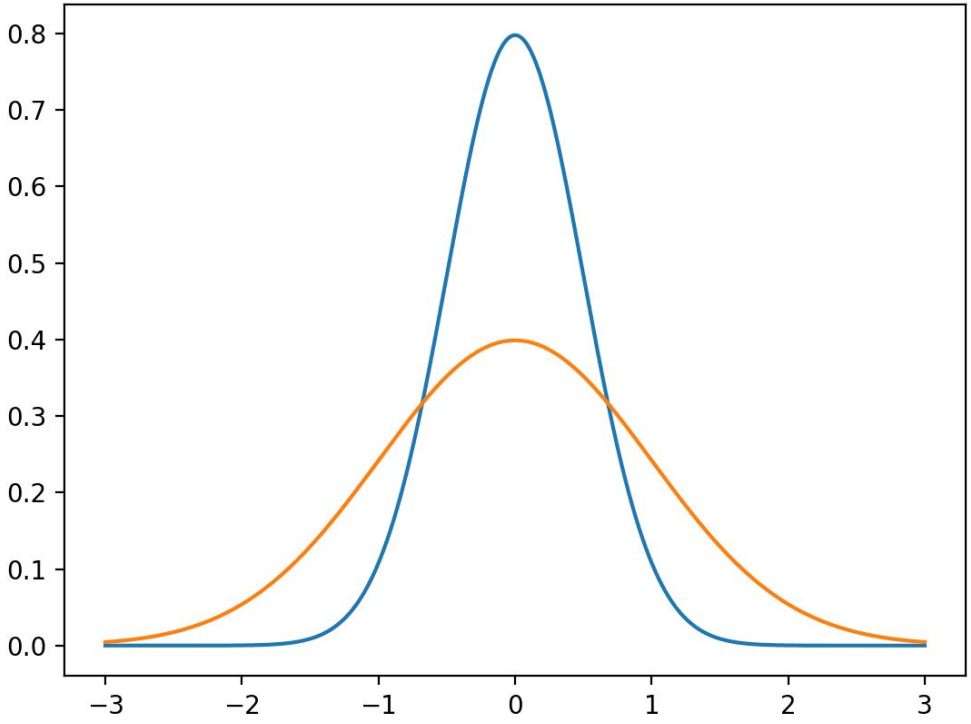




BROADCOM

Gaussian or normal PDF (parametric - mean and variance)



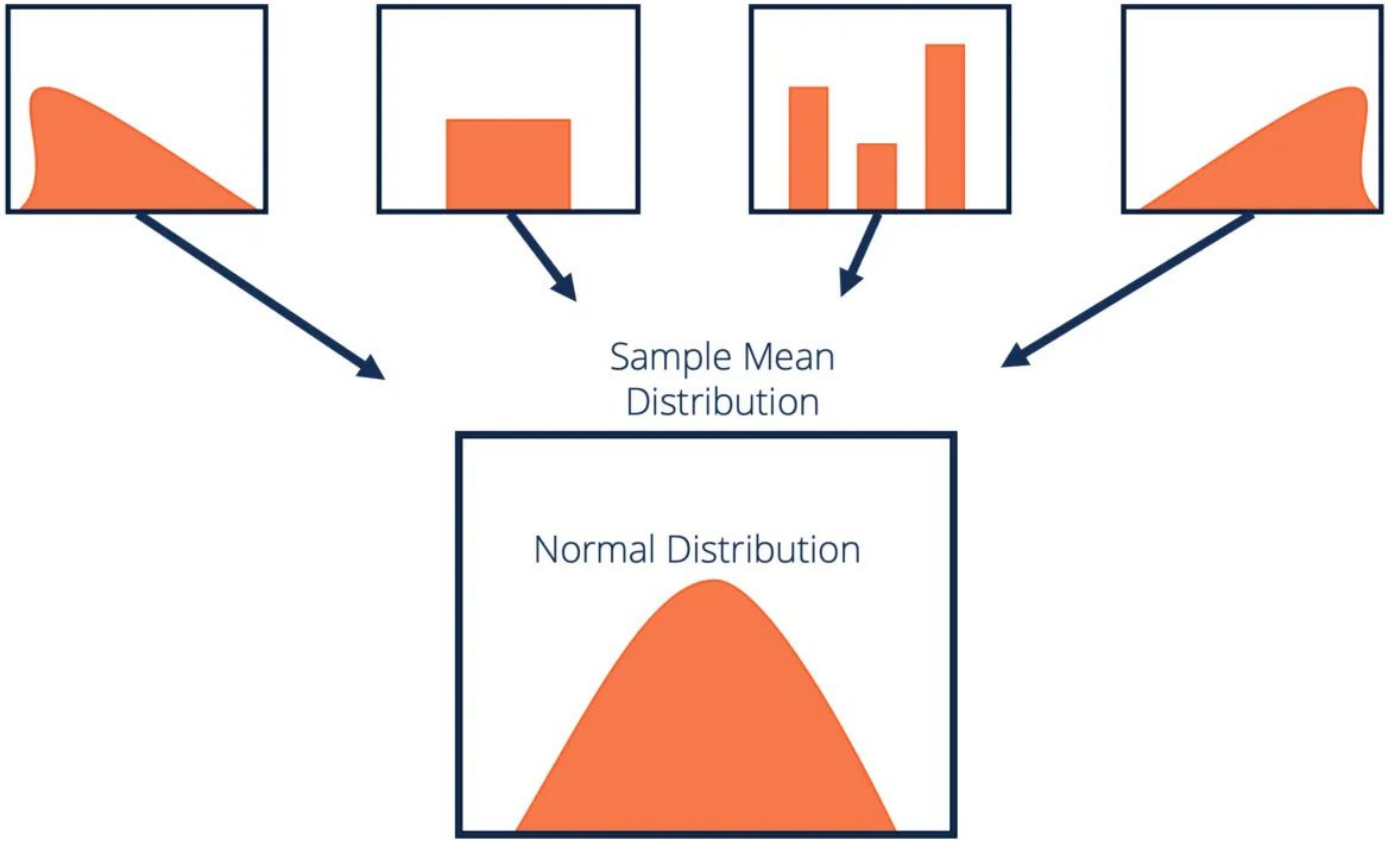


Reminders:

- The Empirical Rule states that 99.7% of data observed following a normal distribution lies within 3 standard deviations of the mean.
- Variance is the standard deviation squared

BROADCOM

Central Limit Theorem

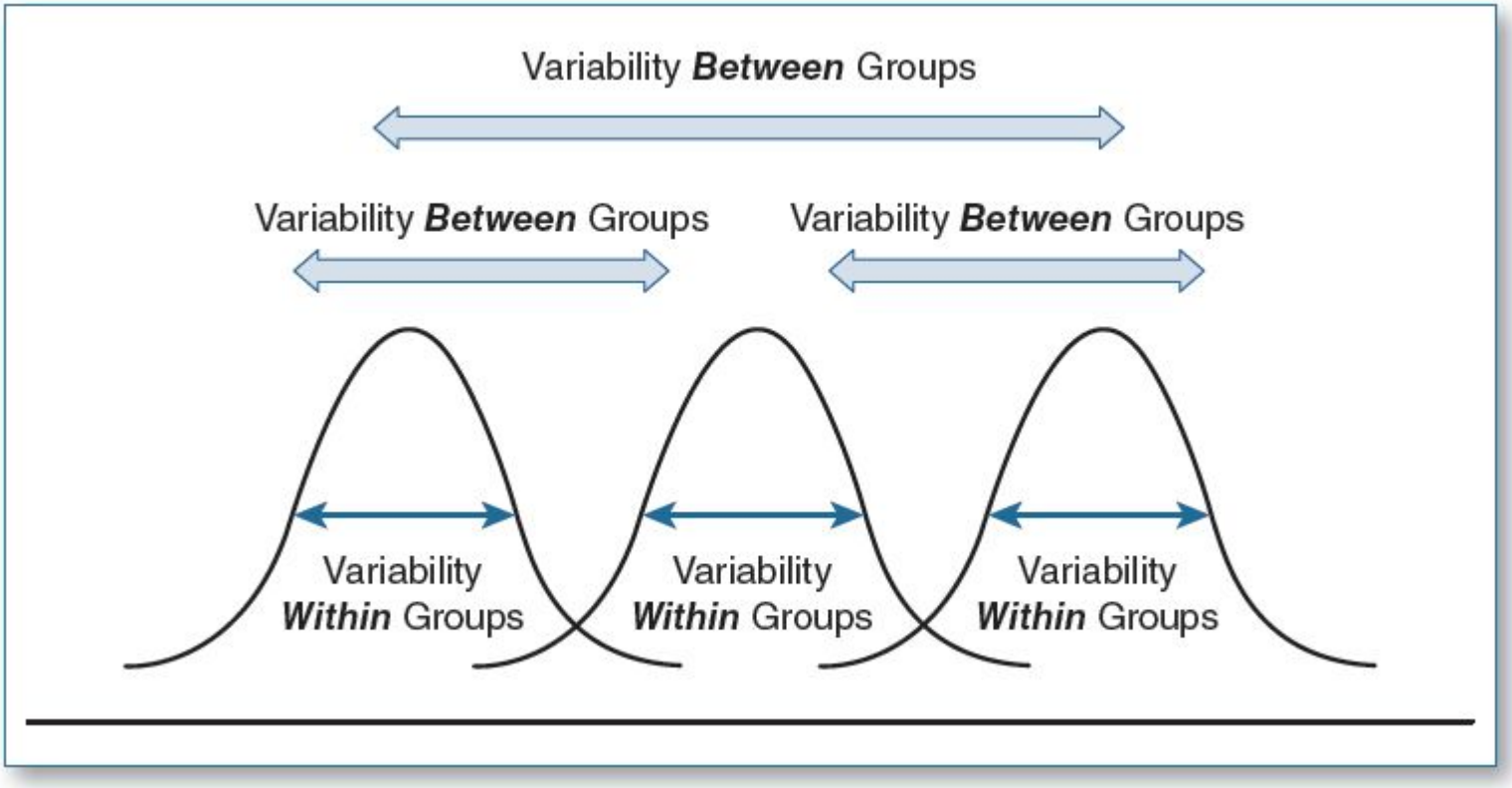


BROADCOM

Variance can signal process mean or variance shift (ANOVA & MANOVA)

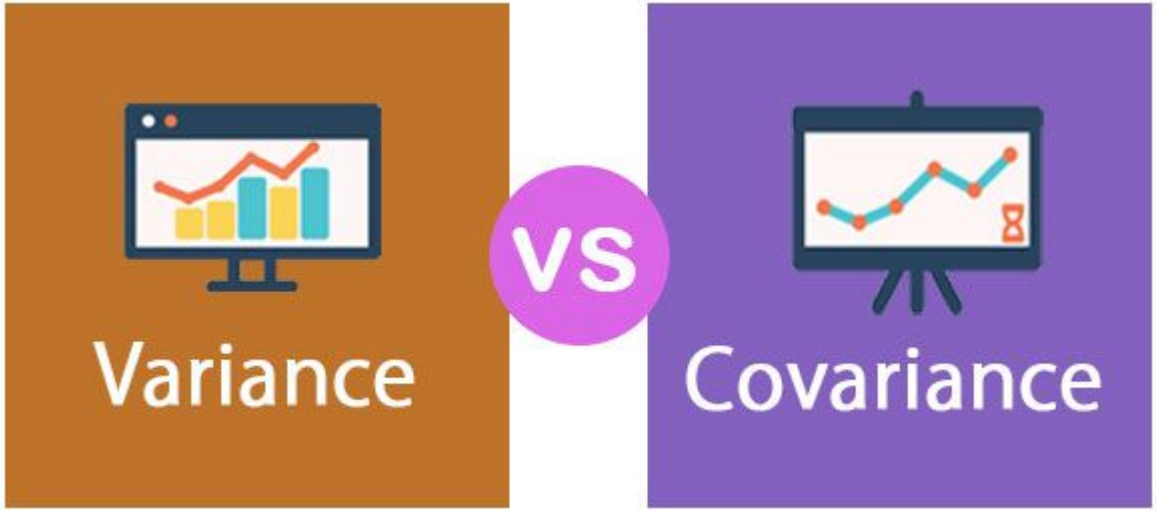
BROADCOM

Detect mean or variance change using only variances



BROADCOM

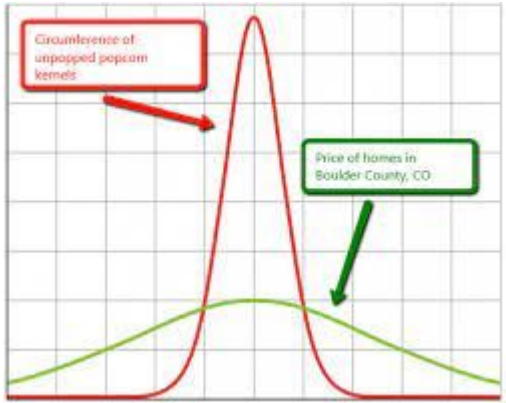
Variance-covariance



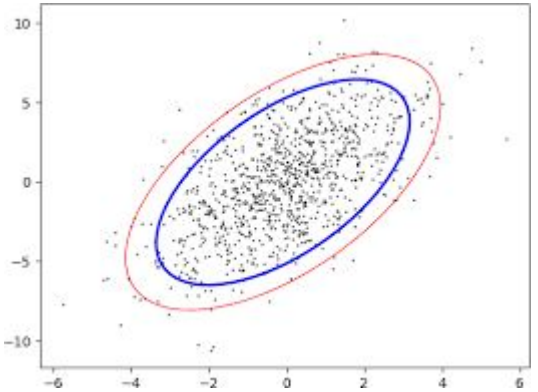
BROADCOM

Bell vs ellipse (assumes Gaussians, thanks for CLT)

Variance



Covariance



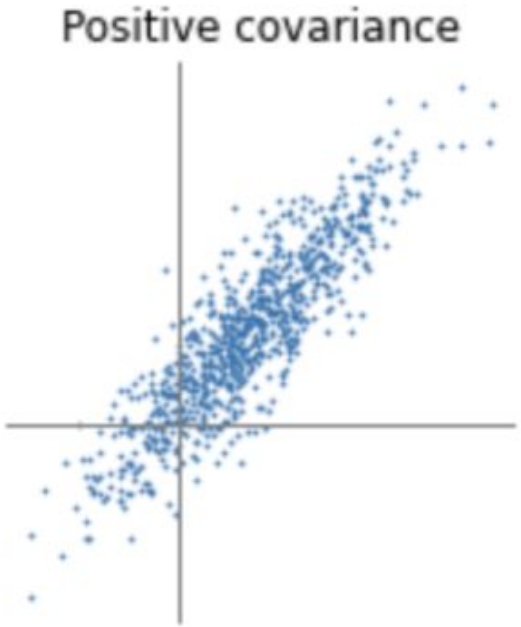
BROADCOM

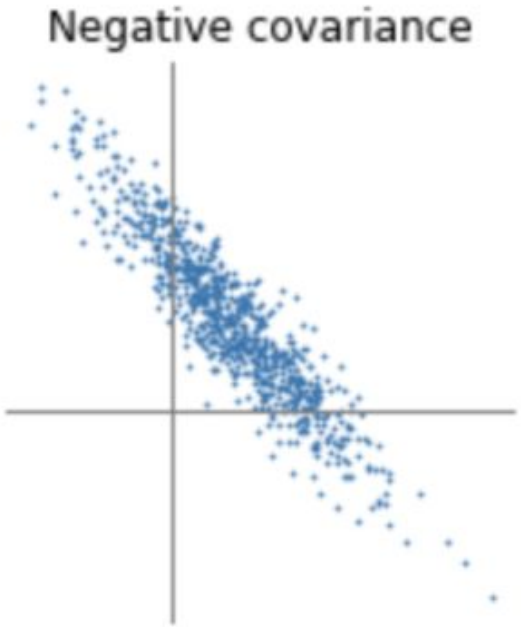
Covariance of two variables

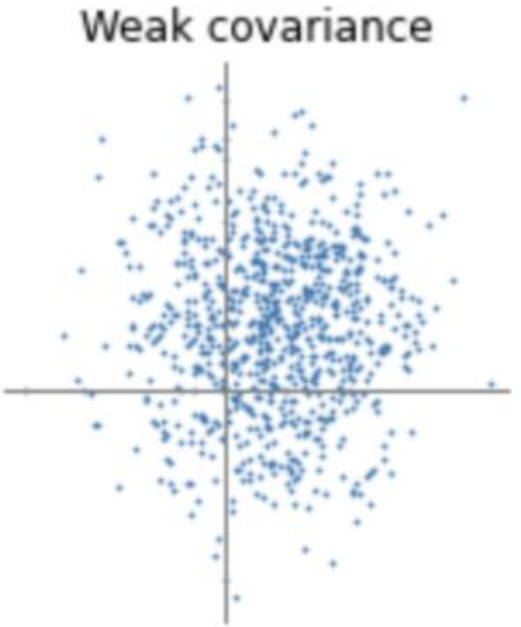
(units are product of each measurement unit, e.g. Throughput * Current = bits/sec * milliamps)

$$\mathrm{cov}_{xy} = \frac{\sum_{i=1}^{n} (x_i - x)(y_i - y)}{n}$$

Positive, negative and weak





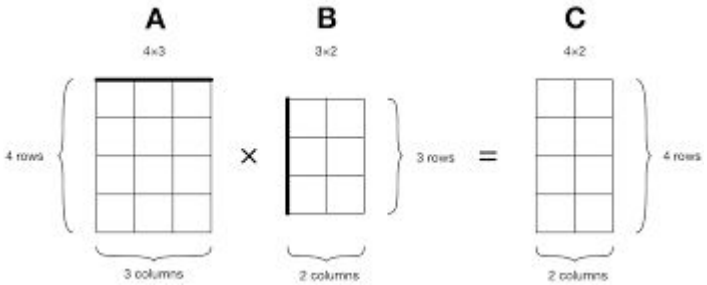


BROADCOM

Linear algebra review

$$D^2 = \left(X_{p_1} - X_{p_2}\right)^T . C^{-1} . \left(X_{p_1} - X_{p_2}\right)$$

Matrix multiply (MxN * NxT = MxT)



$$\begin{bmatrix} a & b & c \\ d & e & f \end{bmatrix} \times \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix} = \begin{bmatrix} 1a + 2b + 3c & 4a + 5b + 6c \\ 1d + 2e + 3f & 4d + 5e + 6f \end{bmatrix}$$

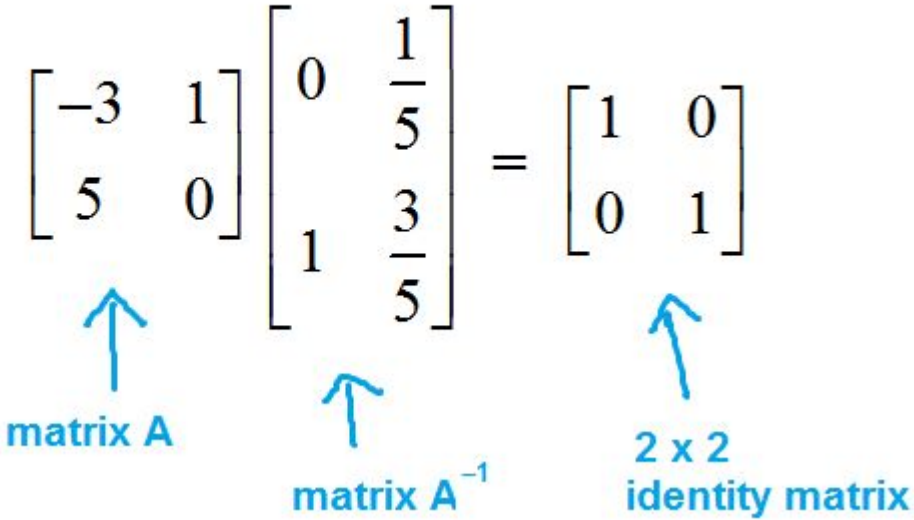
BROADCOM

Column Transpose

BROADCOM

Matrix inverse

The product of A and its inverse is the identity:



Covariance Matrix

Some intuition

More elaborately,

Σ (covariance matrix) is a measure of how the variables are dispersed around the mean (the diagonal elements) and how they co-vary with other variables (the off-diagonal) elements. The more the dispersion the farther apart they are from the mean and the more they co-vary (in absolute value) with the other variables the stronger is the tendency for them to 'move together' (in the same or opposite direction depending on the sign of the covariance).

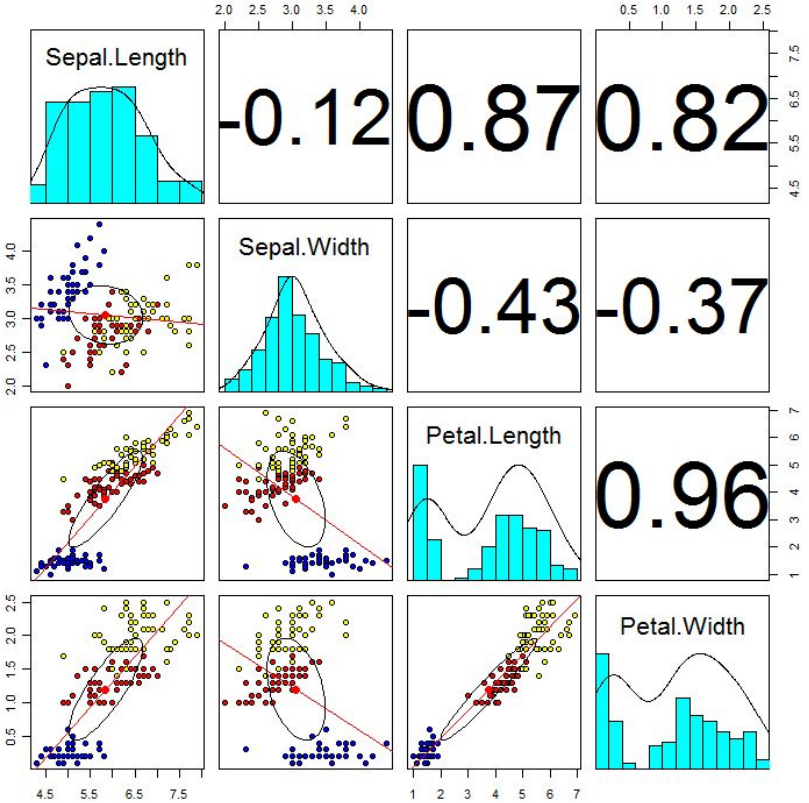
Similarly,

Σ−1 (inverse covariance matrix) is a measure of how tightly clustered the variables are around the mean

Variance-Covariance Matrix (Symmetric and Hermitian)

$$\begin{bmatrix} V_a & C_{a,b} & C_{a,c} & C_{a,d} & C_{a,e} \\ C_{a,b} & V_b & C_{b,c} & C_{b,d} & C_{b,e} \\ C_{a,c} & C_{b,c} & V_c & C_{c,d} & C_{c,e} \\ C_{a,d} & C_{b,d} & C_{c,d} & V_d & C_{d,e} \\ C_{a,e} & C_{b,e} & C_{c,e} & C_{d,e} & V_e \end{bmatrix}$$

Correlation matrix visualization



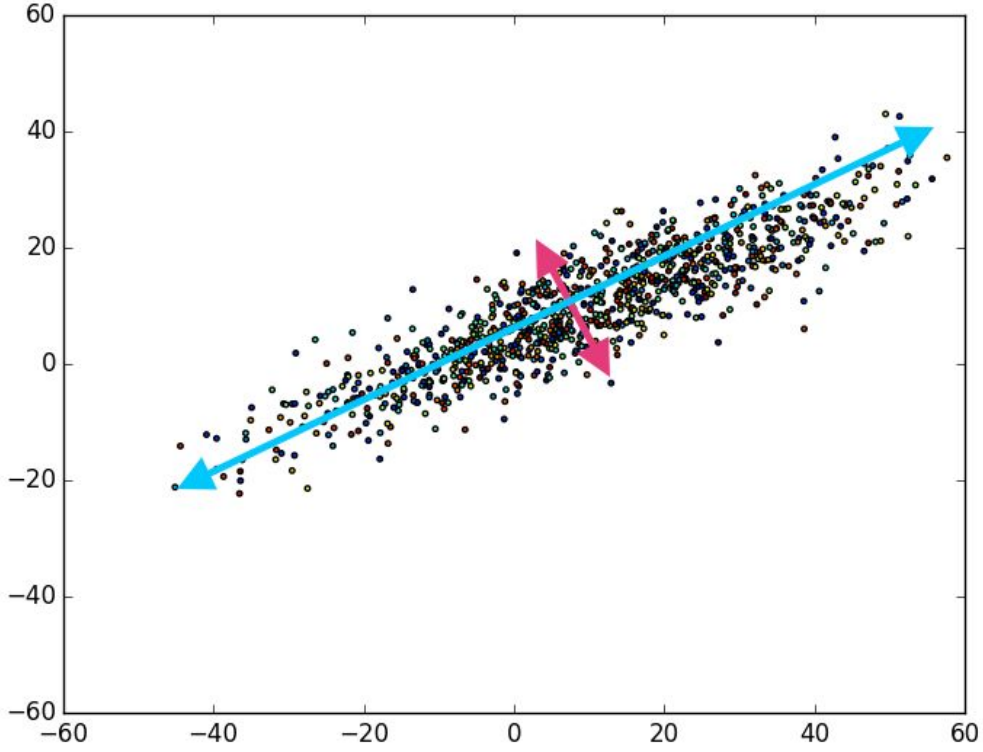
BROADCOM

Linear Transformation

**(e.g. to remove correlation or maximize variance or,
for LDA, maximize between group variance and minimize within group variance)**

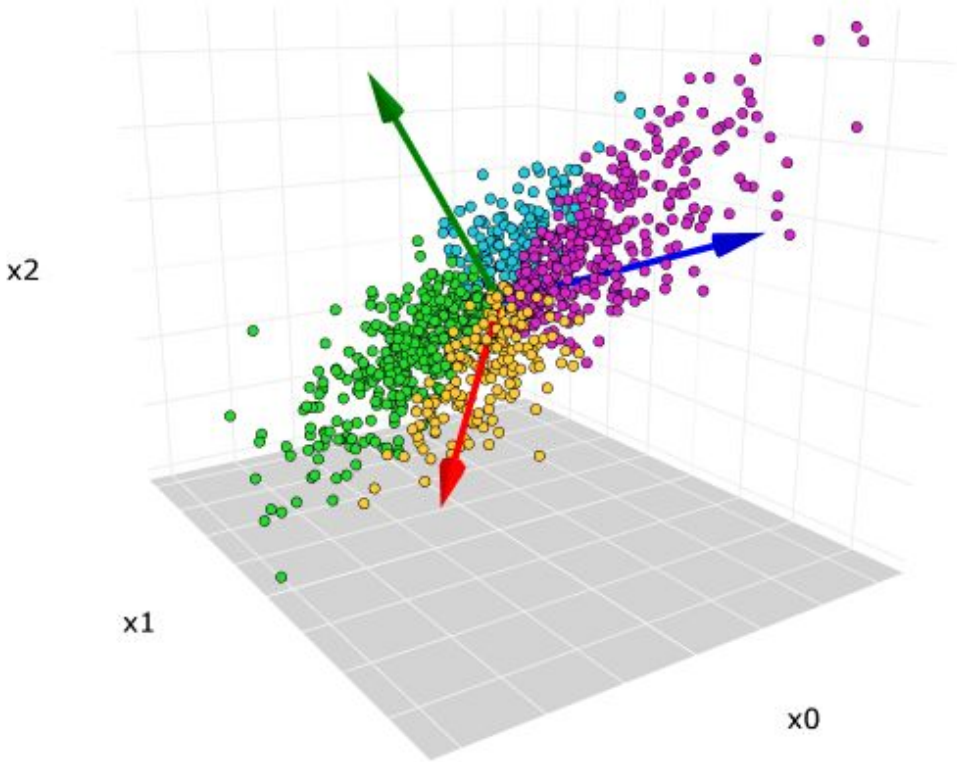
BROADCOM

Eigenvalues and Eigenvectors (2D)



BROADCOM

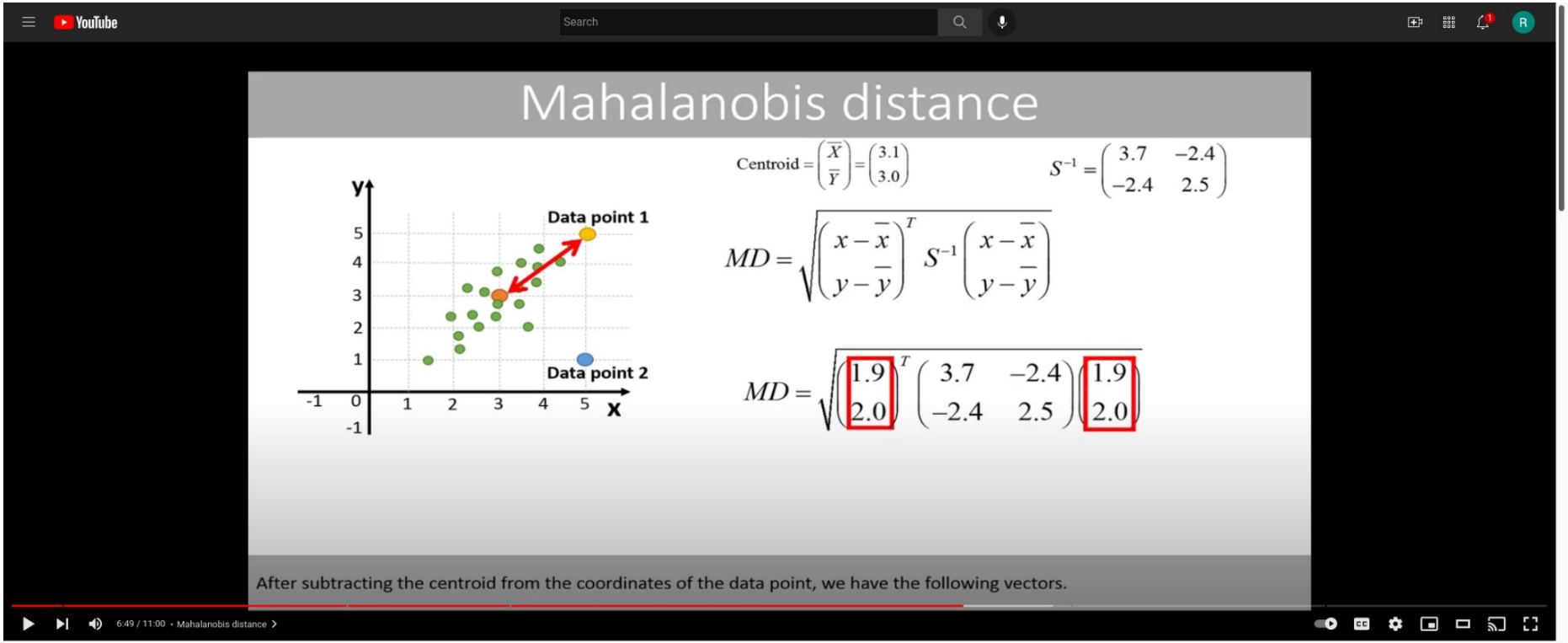
Eigenvalues & Eigenvectors (3D)



Finally - Putting it all together

BROADCOM

Example calculation



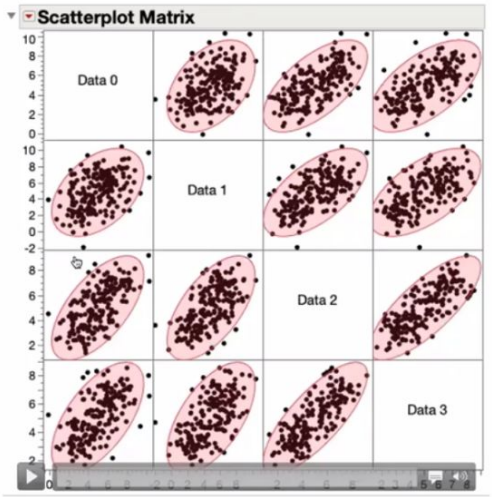
BROADCOM

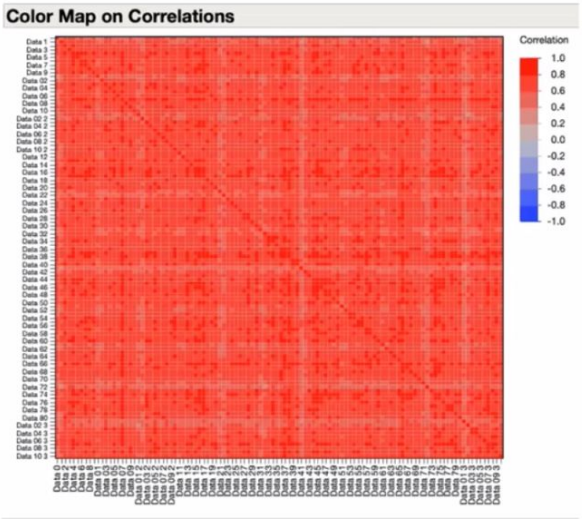
Moving visualization (starts at 8 min)

Community | Discussions | File Exchange | JSL Cookbook | JMP Wish List | JMP Blogs | Sign In

Model Driven Multivariate Control Charts

Key Question: Can I detect when relationships between process variables change





(view in My Videos)

Note: Q&A is included at ~ Times 16:28, 22:00, 23:56, 27:36, 40:24, 43:00 and 45:38.

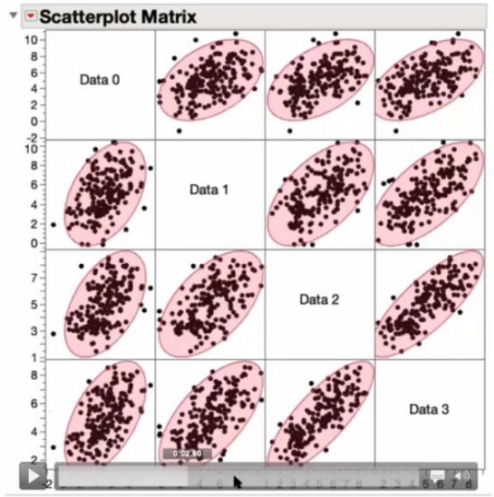
BROADCOM

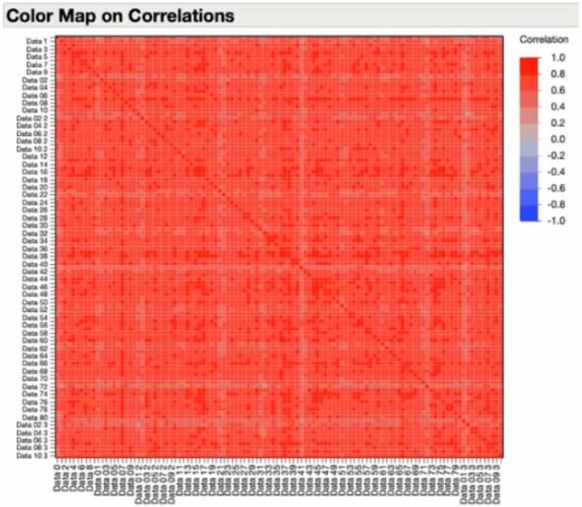
Moving visualization (starts at 8 min)

Community | Discussions | File Exchange | JSL Cookbook | JMP Wish List | JMP Blogs | Sign In

Model Driven Multivariate Control Charts

Key Question: Can I detect when relationships between process variables change





8:39 / 47:33

(view in My Videos)

Note: Q&A is included at ~ Times 16:28, 22:00, 23:56, 27:36, 40:24, 43:00 and 45:38.

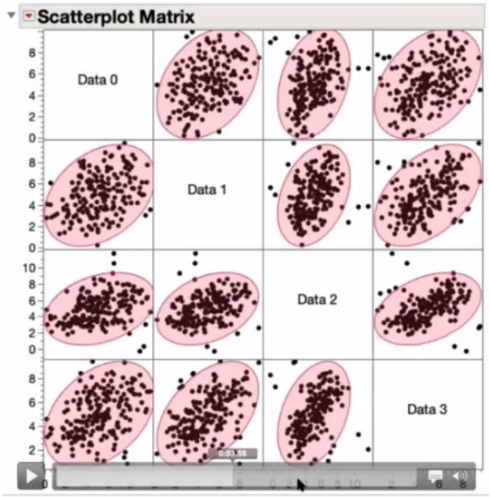
BROADCOM

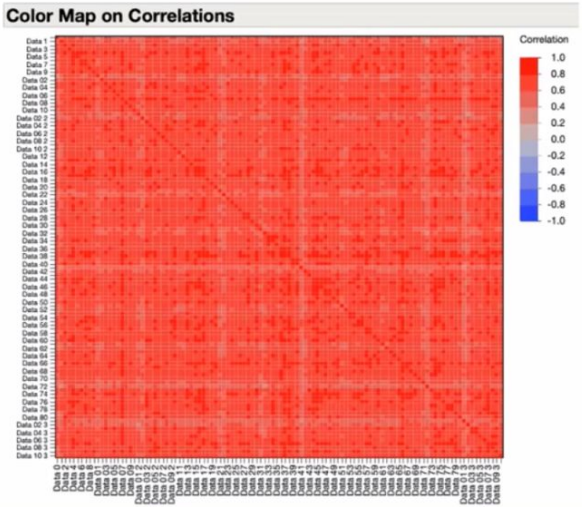
Moving visualization (starts at 8 min)

Community | Discussions | File Exchange | JSL Cookbook | JMP Wish List | JMP Blogs | Sign In

Model Driven Multivariate Control Charts

Key Question: Can I detect when relationships between process variables change





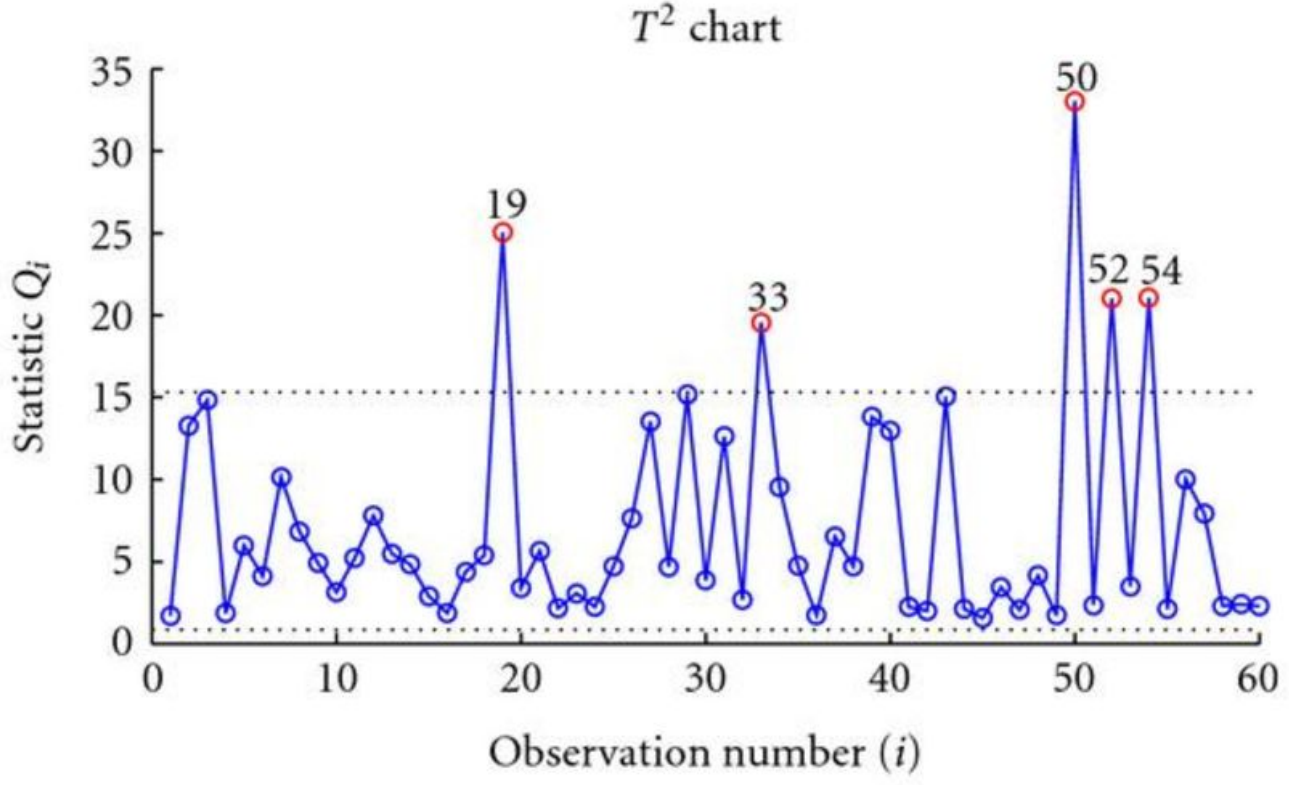
8:39 / 47:33

(view in My Videos)

Note: Q&A is included at ~ Times 16:28, 22:00, 23:56, 27:36, 40:24, 43:00 and 45:38.

BROADCOM

Hotelling T^2 control charts (presentation w/examples for another day)



BROADCOM

For another day

- Non parametric distributions
- Decomposing a multivariate signal into causes/correlations
- Multivariate control chart limits - how computed
- The f-distribution
- Linear discriminant analysis (LDA) and pooled covariance
- P-value and the null hypothesis
- P-value with large sample sizes

External links

- Types of distance in machine learning
- Elements of multivariate analysis
- PCA visualization
- Hotelling Python Code
- Leveraing and utilizing multivariate control charts
- Python pandas
- Hotelling and Excel
- Understanding the Covariance Matrix
- Covariance matrix estimation on condensed data
- Code examples
- Engineering Statistics Handbook

Backup slides

BROADCOM

Background

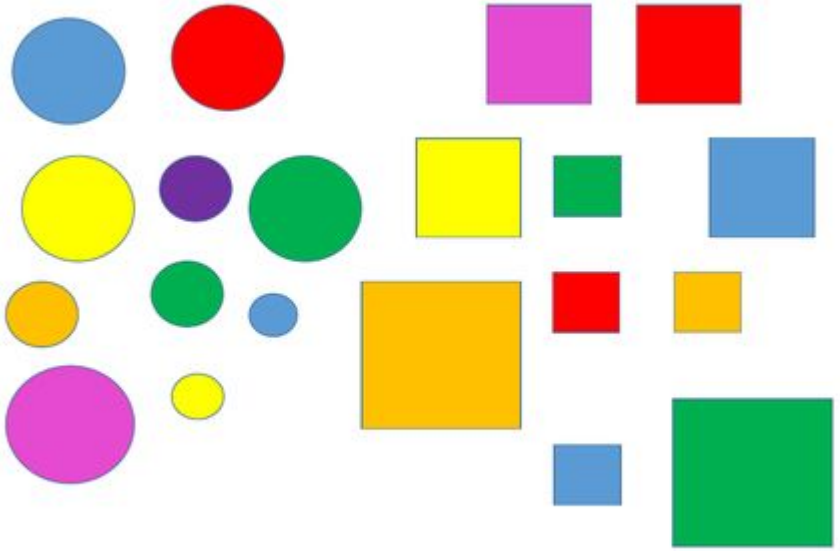
- Mahanabolis distance was introduced in 1936 by Prasanta Chandra Mahalanobis
- Was laborious to calculate until computers came to assist
- Mathematical calculations today use Excel, Python, R, etc.
- Need to define and use key performance indicators (KPIs) per domain expertise (e.g. WiFi metrics, end/end metrics)
- Used by many, many industries, e.g. make sure a widget constructed is the same regardless of which manufacturing line

Pass/Fail results are insufficient (SPC should reduce Type 1 and Type 2 errors)

		Reality: True	Reality: False
Measured or Perceived	True	Correct ☺	**Type 1 error** False Positive
	False	**Type 2 error** False Negative	Correct ☺

BROADCOM

Metric choice matters (shape, size, color?)



Requires domain knowledge as well as devices that can provide key metrics

BROADCOM

What's the difference between univariate, bivariate and multivariate descriptive statistics?

- **Univariate** statistics summarize only **one variable** at a time.
- **Bivariate** statistics compare **two variables**.
- **Multivariate** statistics compare **more than two variables**.

enjoy algorithms

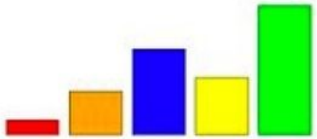
Univariate, Bivariate, Multivariate Analysis:

Data Analysis

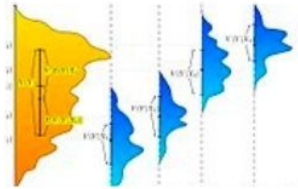
Distribution, Histogram, Pi-chart, Boxplot, Bargraph

Scatter plot, Chi-squared, ANOVA

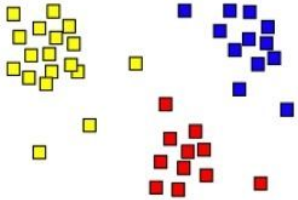
Clustering, PCA, MCA



Univariate

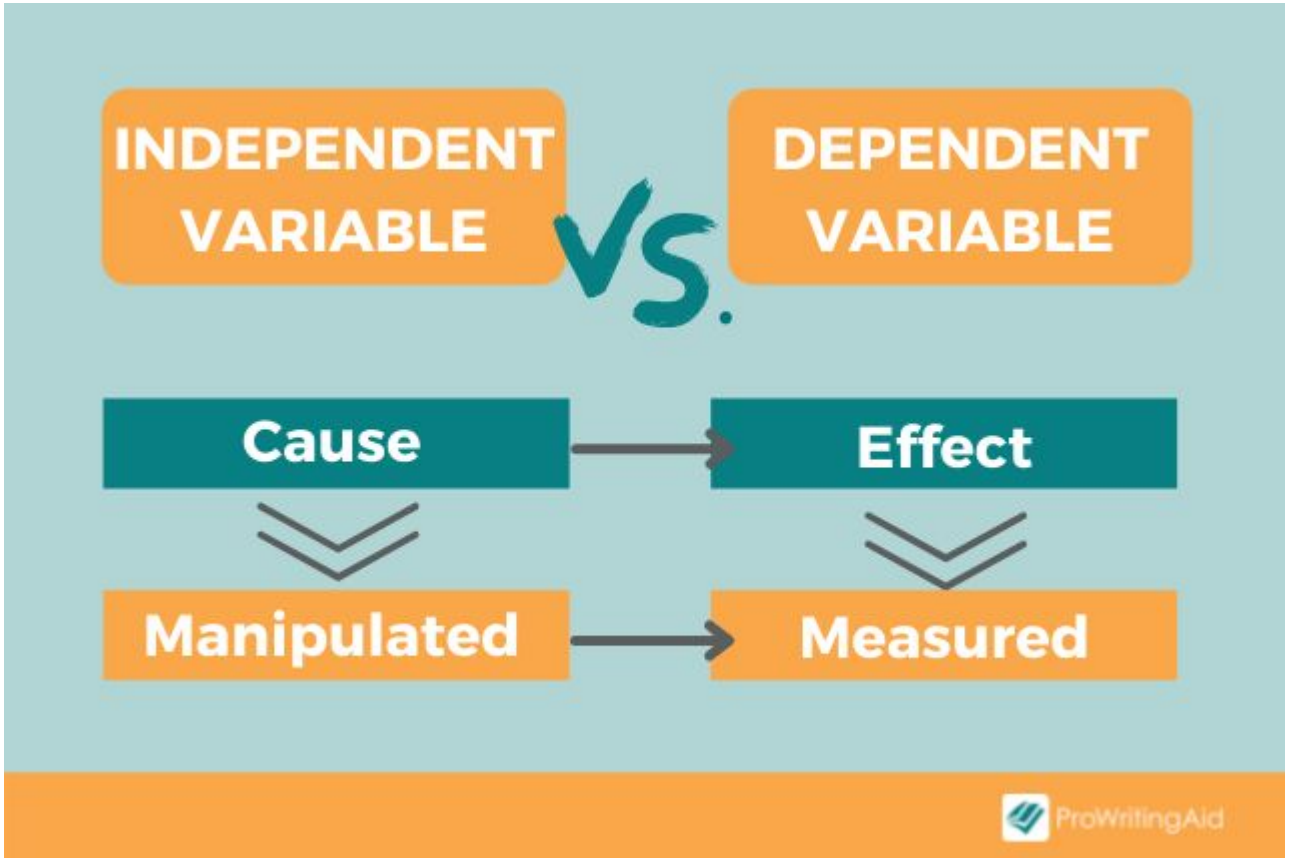


Bivariate



Multivariate

BROADCOM



BROADCOM

INDEPENDENT VARIABLES

Variables that is changed

Amount of water

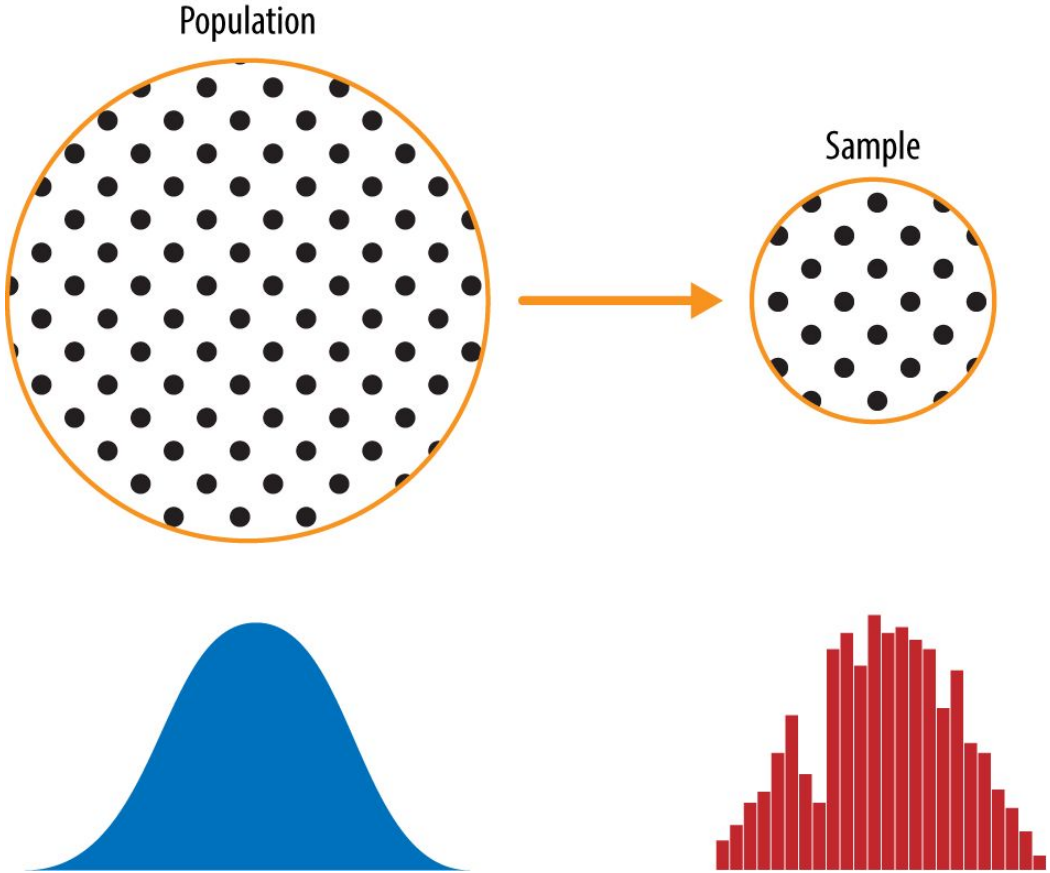
DEPENDENT VARIABLES

Variables affected by the change

Size of plant
number of leaves
living or dead?

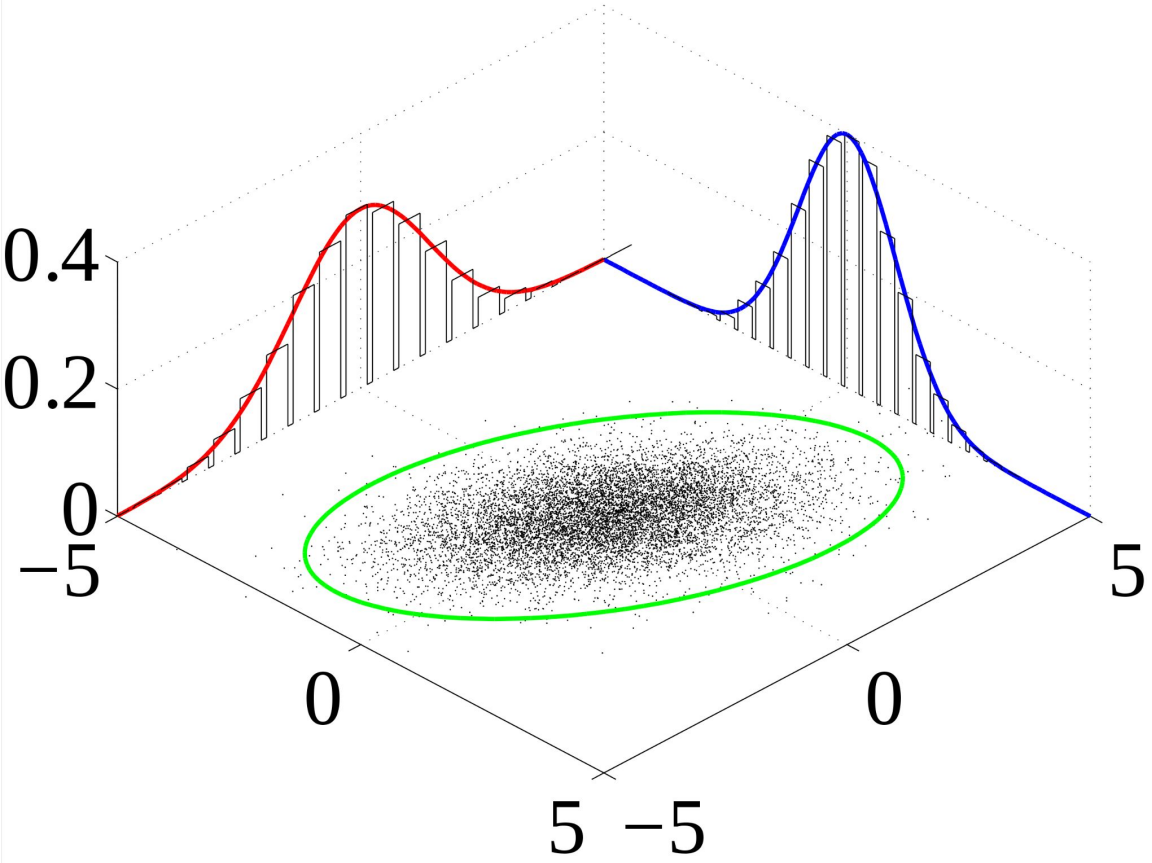
BROADCOM

Population vs Sample (or subgroup)



BROADCOM

Covariance (ellipse w/Gaussians)



BROADCOM

Covariance formula (units are metric A * metric B)

Covariance Formula

For Population

$$Cov(x,y) = \frac{\sum (x_i - \bar{x}) * (y_i - \bar{y})}{N}$$

For Sample

$$Cov(x,y) = \frac{\sum (x_i - \bar{x}) * (y_i - \bar{y})}{(N-1)}$$

Correlation Formula (normalize to 1, remove units)

@}

$$Cor(X,Y) = \frac{Cov(X,Y)}{s_x s_y}$$

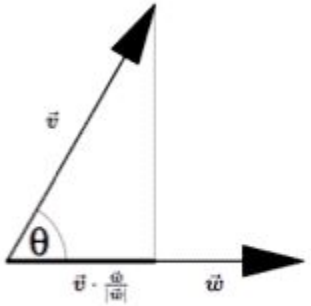
covariance

$$= \quad Cor(X,Y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 (y_i - \bar{y})^2}}$$

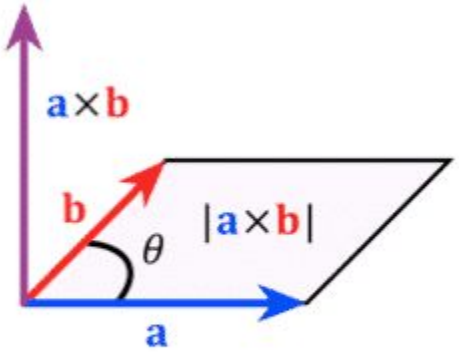
standard deviations

BROADCOM

“Distance” & “Sameness” for vectors



dot product

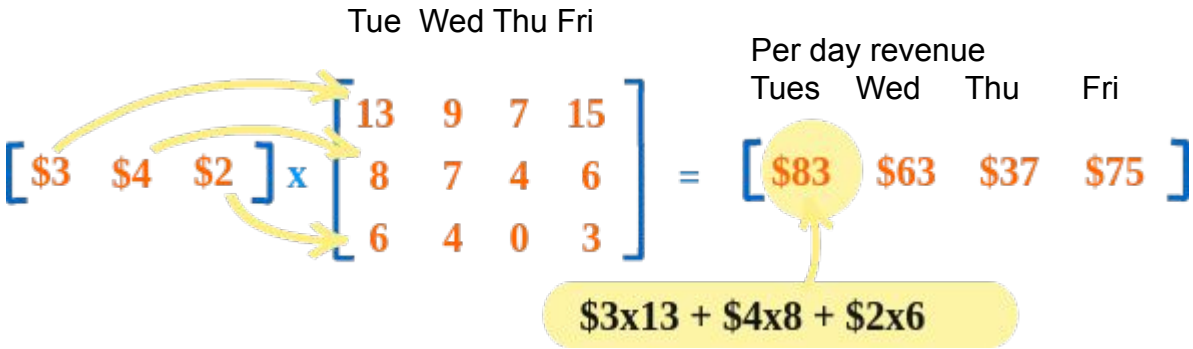


cross product

BROADCOM

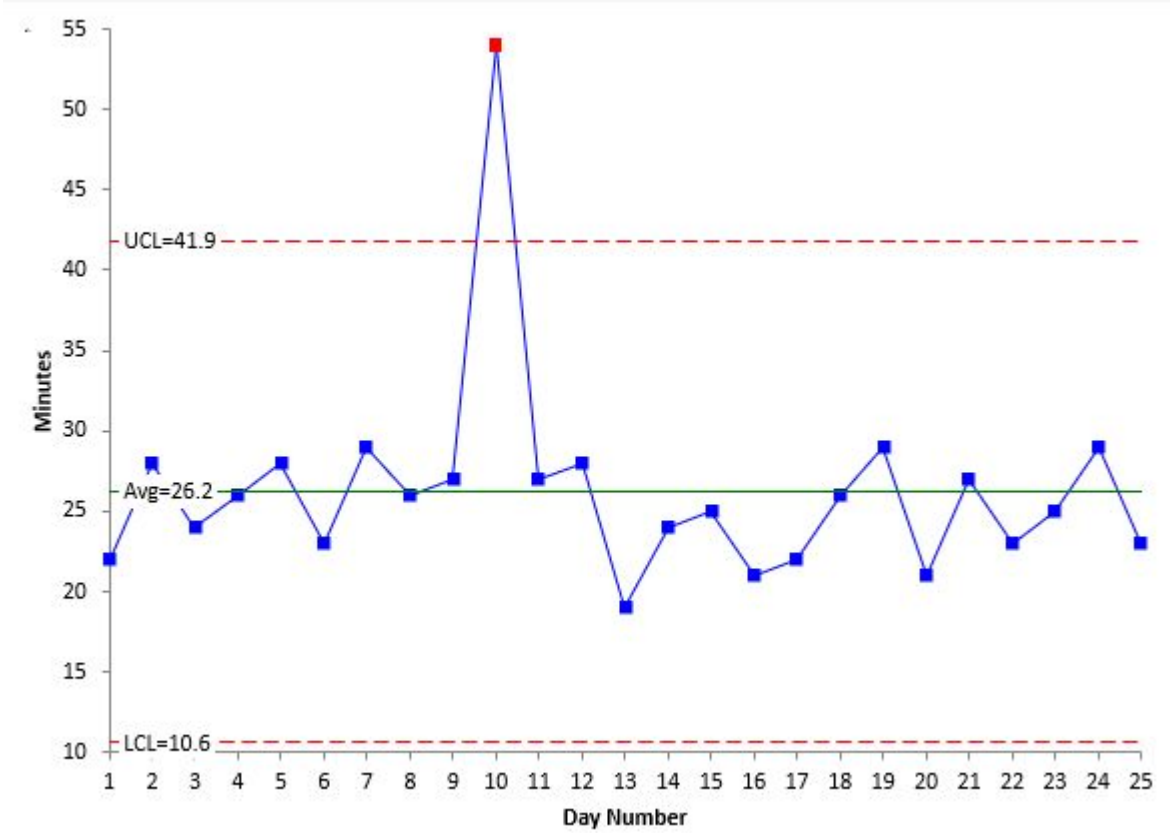
Matrix multiply

Apple = $3
Orange = $4
Pear = $2



BROADCOM

Use distance metric for SPC control chart



BROADCOM

Example Python code

```
import numpy as np
from sklearn import datasets
from scipy.stats import f
def TwoSampleT2Test(X, Y):
        nx, p = X.shape
        ny, _ = Y.shape
        delta = np.mean(X, axis=0) - np.mean(Y, axis=0)
        Sx = np.cov(X, rowvar=False)
        Sy = np.cov(Y, rowvar=False)
        S_pooled = ((nx-1)*Sx + (ny-1)*Sy)/(nx+ny-2)
        t_squared = (nx*ny)/(nx+ny) * np.matmul(np.matmul(delta.transpose(), np.linalg.inv(S_pooled)), delta)
        statistic = t_squared * (nx+ny-p-1)/(p*(nx+ny-2))
        F = f(p, nx+ny-p-1)
        p_value = 1 - F.cdf(statistic)
        print(f"Test statistic: {statistic}\nDegrees of freedom: {p} and {nx+ny-p-1}\np-value: {p_value}")
        return statistic, p_value
        iris = datasets.load_iris()
        versicolor = iris.data[iris.target==1, :2]
        virginica = iris.data[iris.target==2, :2]
        TwoSampleT2Test(versicolor, virginica)
        ## Test statistic: 15.82660099191812
        ## Degrees of freedom: 2 and 97
        ## p-value: 1.1259783253558808e-06
```


BROADCOM