e-PGPathshala
Subject : Computer Science
Paper: Data Analytics
Module No 11: CS/DA/11- Data Analysis Foundations – Bivariate and Multivariate analysis
Quadrant 1 – e-text

1.1 Introduction

In the previous module we discussed about univariate data and how it is handled in analysis. This module provides a comprehensive view of multivariate and how it may be used for analysis. In real world data comes with multiple features or attributes. So exploratory data analysis requires exploiting the associations, correlations among these attributes and the attributes with different scales also must be normalized for improving the computation. The aim of this chapter is to give an understanding of bivariate and multivariate data analysis and normalization of attributes.

1.2 Learning Outcomes

- To Understand the basics of bivariate analysis.
- To learn the fundamentals of multivariate analysis and its necessity.
- To gain knowledge about various data normalization methods.

1.3 Bivariate Analysis

Bivariate analysis is one of the simplest forms of quantitative (statistical) analysis. It involves the analysis of two variables (often denoted as X, Y), for the purpose of determining/understanding the association or dependence (empirical relationship) between them, if any. We thus restrict our attention to the two numeric attributes of interest, say X1 and X2 , with the data D represented as an n×2 matrix:

$$\mathbf{D} = \left(\begin{array}{cc} X_1 & X_2 \\ \hline x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{array}\right)$$

Geometrically, we can think of **D in two ways. It can be viewed as n points or vectors in 2-dimensional space** over the attributes *X_1 and X_2.*

With bivariate analysis, we are testing hypotheses of "association" and causality. In its simplest form, association simply refers to the extent to which it becomes easier to predict a value for the Dependent variable if we know a case's value on the independent variable.

A measure of association helps us to understand this relationship. These measures of association relate to how much better this prediction becomes with knowledge of the independent variable or how well an independent variable relates to the dependent variable. We have already discussed this in more abstract terms of "correlation". A measure of association often ranges between –1 and 1. Where the sign of the integer represents the "direction" of correlation (negative or positive relationships) and the distance away from 0 represents the degree or extent of correlation – the farther the number away from 0, the higher or "more perfect" the relationship is between the IV and DV.

Statistical significance relates to the generalisability of the relationship AND, more importantly, the likelihood the observed relationship occurred BY CHANCE. In political science, we typically consider a relationship significant if it has a significance level of .05 – In only 5/100 times will the pattern of observations for these two variables that we have measured occur by chance. Often significance levels, when n (total number of cases in a sample) is large, can approach .001 (only 1/1000 times will the observed association occur).

Measures of association and statistical significance that are used vary by the level of measurement of the variables analyzed.

1.4 Measures of location and dispersion

Common examples of measures of statistical **dispersion** are the variance, standard deviation and inter-quartile range. **Dispersion** is contrasted with **location** or central tendency, and together they are the most used properties of distributions.

a) Mean:

The bivariate mean is defined as the expected value of the vector random variable X

$$\mu = E\left[X\right] = E\left[\begin{pmatrix} X1 \\ X2 \end{pmatrix}\right] = \begin{pmatrix} \boldsymbol{E}\left[\boldsymbol{X1}\right] \\ \boldsymbol{E}\left[\boldsymbol{X2}\right] \end{pmatrix} = \begin{pmatrix} \mu1 \\ \mu2 \end{pmatrix}$$

b) Variance

We can compute the variance along each attribute, namely $\sigma1^{2}$ for X1 and $\sigma2^{2}$ for X2 using: $\sigma^{2} = \text{var}(X) = E[(X-\mu)^{2}]$

$$=\begin{cases}\Sigma(x \quad - \quad \mu)2f(x)\\ \int_{-\alpha}^{\alpha}(x \quad - \quad \mu)2f(x)d(x)\end{cases}$$

The *total variance is given as*: Var(D)= $\sigma1^{2} + \sigma2^{2}$

1.5 Measures of association

a) **Variance** is a measure of the variability or spread in a set of data. Mathematically, it is the average squared deviation from the mean score. We use the following formula to compute variance.

Var(X) = Σ (X_i - X)^2 / N = Σ x_i^2 / N
where,
N is the number of scores in a set of scores,
X is the mean of the N scores,
X_i is the ith raw score in the set of scores,
x_i is the ith deviation score in the set of scores,
Var(X) is the variance of all the scores in the set

b) **Covariance** is a measure of the extent to which corresponding elements from two sets of ordered data move in the same direction. We use the following formula to compute covariance.

Cov(X, Y) = Σ (X_i - X) (Y_i - Y) / N = Σ $x_i y_i$ / N
where,
N is the number of scores in each set of data,
X is the mean of the N scores in the first data set
X_i is the $i$the raw score in the first set of scores
x_i is the ith deviation score in the first set of scores
Y is the mean of the N scores in the second data set
Y_i is the $i$the raw score in the second set of scores
y_i is the ith deviation score in the second set of scores
Cov(X, Y) is the covariance of corresponding scores in the two sets of data

c) **Variance-Covariance Matrix:** Variance and covariance are often displayed together in a variance-covariance matrix, (aka, a covariance matrix). The variances appear along the diagonal and covariances appear in the off-diagonal elements, as shown below.

$$\mathbf{V} \quad = \quad \begin{bmatrix} \Sigma\, x_1^2 / \mathrm{N} & \Sigma\, x_1\, x_2 / \mathrm{N} & \ldots & \Sigma\, x_1\, x_c / \mathrm{N} \\ \Sigma\, x_2\, x_1 / \mathrm{N} & \Sigma\, x_2^2 / \mathrm{N} & \ldots & \Sigma\, x_2\, x_c / \mathrm{N} \end{bmatrix}$$

$$\begin{bmatrix} \cdots & \cdots & \cdots & \cdots \\ \Sigma\, x_c\, x_1 / N & \Sigma\, x_c\, x_2 / N & \cdots & \Sigma\, x_c^2 / N \end{bmatrix}$$

where,

V is a c x c variance-covariance matrix

N is the number of scores in each of the c data sets

x_i is a deviation score from the *ith* data set

$\Sigma\, x_i^2 / N$ is the variance of elements from the *i*th data set

$\Sigma\, x_i\, x_j / N$ is the covariance for elements from the *i*th and *j*th data sets

Example: Steps for creating Variance-Covariance Matrix:

The table below displays scores on math, English, and art tests for 5 students. Note that data from the table is represented in matrix **A**, where each column in the matrix shows scores on a test and each row shows scores for a student.

Student	Math	English	Art
1	90	60	90
2	90	90	30
3	60	60	60
4	60	60	90
5	30	30	30

$$\Rightarrow \underset{\textbf{A}}{\begin{bmatrix} 90 & 60 & 90 \\ 90 & 90 & 30 \\ 60 & 60 & 60 \\ 60 & 60 & 90 \\ 30 & 30 & 30 \end{bmatrix}}$$

Given the data represented in matrix **A**, compute the variance of each test and the covariance between the tests.

Solution

The solution involves a three-step process.

Step 1: Transform the *raw* scores in matrix **A** to *deviation* scores in matrix **a**, using the transformation formula.

a = **A** - **11'A** (1 / n)

where

1 is an *5* x *1* column vector of ones

a is an *5* x *3* matrix of *deviation* scores: $a_{11}, a_{12}, \ldots, a_{53}$

A is an *5* x *3* matrix of *raw* scores: A_{11}, A_{12}, . . . , A_{53}
n is the number of rows in matrix **A**

$$\mathbf{a} = \begin{bmatrix} 90 & 60 & 90 \\ 90 & 90 & 30 \\ 60 & 60 & 60 \\ 60 & 60 & 90 \\ 30 & 30 & 30 \end{bmatrix} - \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 90 & 60 & 90 \\ 90 & 90 & 30 \\ 60 & 60 & 60 \\ 60 & 60 & 90 \\ 30 & 30 & 30 \end{bmatrix} (1/5)$$

$$\mathbf{a} = \begin{bmatrix} 90 & 60 & 90 \\ 90 & 90 & 30 \\ 60 & 60 & 60 \\ 60 & 60 & 90 \\ 30 & 30 & 30 \end{bmatrix} - \begin{bmatrix} 66 & 60 & 60 \\ 66 & 60 & 60 \\ 66 & 60 & 60 \\ 66 & 60 & 60 \\ 66 & 60 & 60 \end{bmatrix} = \begin{bmatrix} 24 & 0 & 30 \\ 24 & 30 & -30 \\ -6 & 0 & 0 \\ -6 & 0 & 30 \\ -36 & -30 & -30 \end{bmatrix}$$

Step 2: Compute **a'a** - the deviation score sums of squares matrix, as shown below.

$$\mathbf{a}'\,\mathbf{a} = \begin{bmatrix} 24 & 24 & -6 & -6 & -36 \\ 0 & 30 & 0 & 0 & -30 \\ 30 & -30 & 0 & 30 & -30 \end{bmatrix} \begin{bmatrix} 24 & 0 & 30 \\ 24 & 30 & -30 \\ -6 & 0 & 0 \\ -6 & 0 & 30 \\ -36 & -30 & -30 \end{bmatrix} = \begin{bmatrix} 2520 & 1800 & 900 \\ 1800 & 1800 & 0 \\ 900 & 0 & 3600 \end{bmatrix}$$

Step 3: And finally, to create the variance-covariance matrix, we divide each element in the deviation sum of squares matrix by *n*, as shown below:

$$\mathbf{V} = \mathbf{a}'\,\mathbf{a} / n = \begin{bmatrix} 2520/5 & 1800/5 & 900/5 \\ 1800/5 & 1800/5 & 0/5 \\ 900/5 & 0/5 & 3600/5 \end{bmatrix} = \begin{bmatrix} 504 & 360 & 180 \\ 360 & 360 & 0 \\ 180 & 0 & 720 \end{bmatrix}$$

We can interpret the variance and covariance statistics in matrix **V** to understand how the various test scores vary and covary.

- Shown in red along the diagonal, we see the variance of scores for each test. The art test has the biggest variance (720); and the English test, the smallest (360). So we can say that art test scores are more variable than English test scores.
- The covariance is displayed in black in the off-diagonal elements of matrix **V**.
  - The covariance between math and English is positive (360), and the covariance between math and art is positive (180). This means the scores tend to covary in a positive way. As scores on math go up, scores on art and English also tend to go up; and vice versa.
  - The covariance between English and art, however, is zero. This means there tends to be no predictable relationship between the movement of English and art scores.

If the covariance between any tests had been negative, it would have meant that the test scores on those tests tend to move in opposite directions. That is, students with relatively high scores on the first test would tend to have relatively low scores on the second test.

d) Correlation

The strength of the linear association between two variables is quantified by the ***correlation coefficient***.

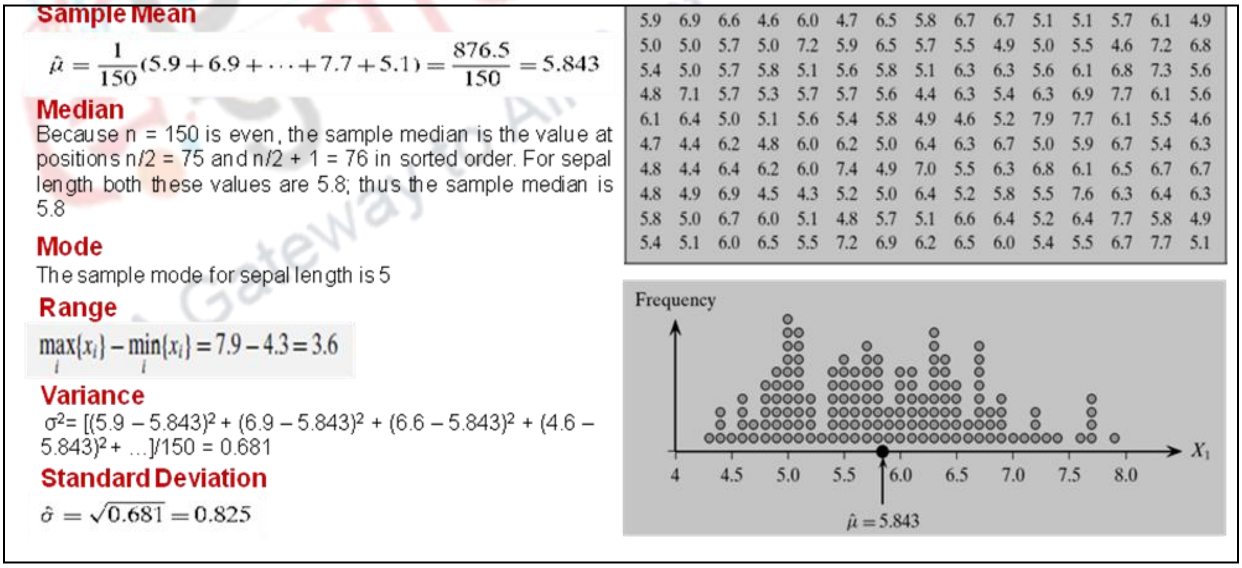
Given a set of observations (x_1, y_1), (x_2,y_2),...(x_n,y_n), the formula for computing the correlation coefficient is given by

$$r = \frac{1}{n-1}\Sigma\left(\frac{x-\overline{x}}{s_x}\right)\left(\frac{y-\overline{y}}{s_y}\right)$$

The correlation coefficient always takes a value between -1 and 1, with 1 or -1 indicating perfect correlation (all points would lie along a straight line in this case). A positive correlation indicates a positive association between the variables (increasing values in one variable correspond to increasing values in the other variable), while a negative correlation indicates a negative association between the variables (increasing values is one variable correspond to decreasing values in the other variable). A correlation value close to 0 indicates no association between the variables.

Since the formula for calculating the correlation coefficient standardizes the variables, changes in scale or units of measurement will not affect its value. For this reason, the correlation coefficient is often more useful than a graphical depiction in determining the strength of the association between two variables.

The following is a sample calculation of the above discussed measures for the IRIS data set:



Sample Mean

$$\hat{\mu} = \frac{1}{150}(5.9+6.9+\cdots+7.7+5.1) = \frac{876.5}{150} = 5.843$$

Median

Because n = 150 is even, the sample median is the value at positions n/2 = 75 and n/2 + 1 = 76 in sorted order. For sepal length both these values are 5.8; thus the sample median is 5.8

Mode

The sample mode for sepal length is 5

Range

$$\max_i\{x_i\} - \min_i\{x_i\} = 7.9 - 4.3 = 3.6$$

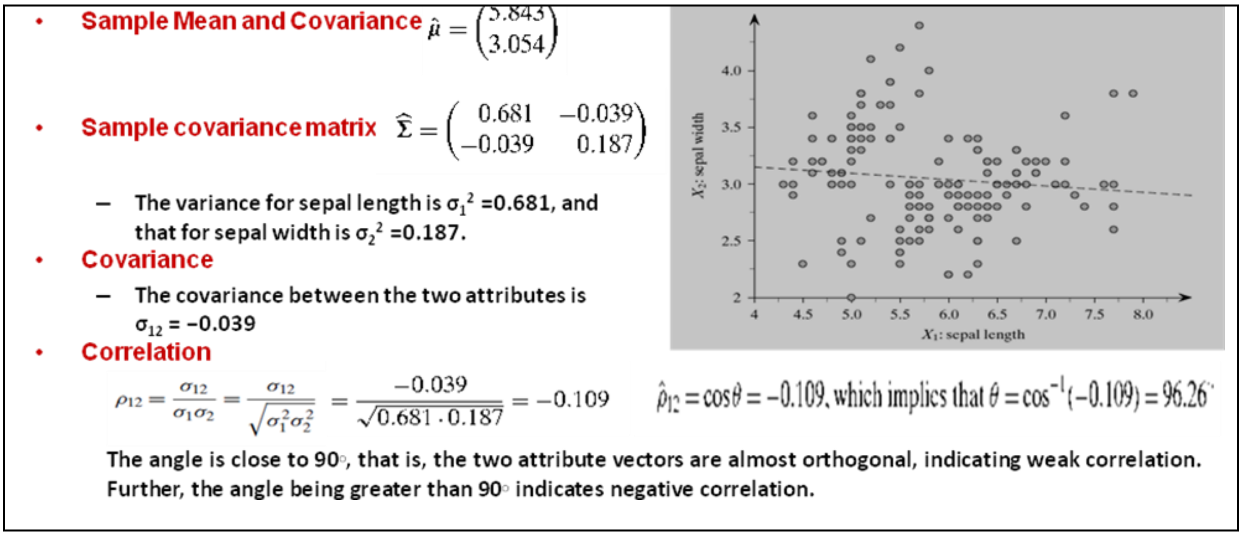
Variance

σ^2= [(5.9 – 5.843)^2 + (6.9 – 5.843)^2 + (6.6 – 5.843)^2 + (4.6 – 5.843)^2 + …]/150 = 0.681

Standard Deviation

$$\hat{\sigma} = \sqrt{0.681} = 0.825$$

5.9	6.9	6.6	4.6	6.0	4.7	6.5	5.8	6.7	6.7	5.1	5.1	5.7	6.1	4.9
5.0	5.0	5.7	5.0	7.2	5.9	6.5	5.7	5.5	4.9	5.0	5.5	4.6	7.2	6.8
5.4	5.0	5.7	5.8	5.1	5.6	5.8	5.1	6.3	6.3	5.6	6.1	6.8	7.3	5.6
4.8	7.1	5.7	5.3	5.7	5.7	5.6	4.4	6.3	5.4	6.3	6.9	7.7	6.1	5.6
6.1	6.4	5.0	5.1	5.6	5.4	5.8	4.9	4.6	5.2	7.9	7.7	6.1	5.5	4.6
4.7	4.4	6.2	4.8	6.0	6.2	5.0	6.4	6.3	6.7	5.0	5.9	6.7	5.4	6.3
4.8	4.4	6.4	6.2	6.0	7.4	4.9	7.0	5.5	6.3	6.8	6.1	6.5	6.7	6.7
4.8	4.9	6.9	4.5	4.3	5.2	5.0	6.4	5.2	5.8	5.5	7.6	6.3	6.4	6.3
5.8	5.0	6.7	6.0	5.1	4.8	5.7	5.1	6.6	6.4	5.2	6.4	7.7	5.8	4.9
5.4	5.1	6.0	6.5	5.5	7.2	6.9	6.2	6.5	6.0	5.4	5.5	6.7	7.7	5.1



- Sample Mean and Covariance $\hat{\mu} = \begin{pmatrix} 5.843 \\ 3.054 \end{pmatrix}$
- Sample covariance matrix $\hat{\Sigma} = \begin{pmatrix} 0.681 & -0.039 \\ -0.039 & 0.187 \end{pmatrix}$
  - The variance for sepal length is σ_1^2 =0.681, and that for sepal width is σ_2^2 =0.187.
- Covariance
  - The covariance between the two attributes is σ_{12} = −0.039
- Correlation

$$\rho_{12} = \frac{\sigma_{12}}{\sigma_1 \sigma_2} = \frac{\sigma_{12}}{\sqrt{\sigma_1^2 \sigma_2^2}} = \frac{-0.039}{\sqrt{0.681 \cdot 0.187}} = -0.109$$

$$\hat{\rho}_{12} = \cos\theta = -0.109, \text{ which implies that } \theta = \cos^{-1}(-0.109) = 96.26^{\circ}$$

The angle is close to 90°, that is, the two attribute vectors are almost orthogonal, indicating weak correlation. Further, the angle being greater than 90° indicates negative correlation.

1.4 Multivariate Analysis

Multivariate analysis is essentially the statistical process of simultaneously analyzing multiple independent (or predictor) variables with multiple dependent (outcome or criterion) variables. Using matrix algebra (most **multivariate analyses** are correlational).

1.4.1 Simpson's Paradox

When you compare a population with labeled subpopulations with another population (or "the same" at a different time), it's extremely likely that the two populations will have different proportions of their subpopulations. This is the heart of Simpson's paradox.

Simpson's paradox occurs when your sample is composed of separate classes with different mean values of a statistical value. In this case, if the class distribution within sample changes between two measures, the trend observed on average might be opposed to the trend observed in each of the two classes.

Example 1: [Source: https://en.wikipedia.org/wiki/Simpson%27s_paradox]

One of the best-known examples of Simpson's paradox is a study of gender bias among graduate school admissions to University of California, Berkeley. The admission figures for the fall of 1973 showed that men applying were more likely than women to be admitted, and the difference was so large that it was unlikely to be due to chance.

	Applicants	Admitted
Men	8442	44%
Women	4321	35%

But when examining the individual departments, it appeared that six out of 85 departments were significantly biased against men, whereas only four were significantly biased against women. In fact, the pooled and corrected data showed a "small but statistically significant bias in favor of women." The data from the six largest departments is listed below.

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%

The research concluded that women tended to apply to competitive departments with low rates of admission even among qualified applicants (such as in the English Department), whereas men tended to apply to less-competitive departments with high rates of admission among the qualified applicants (such as in engineering and chemistry). The conditions under which the admissions' frequency data from specific departments constitute a proper defense against charges of discrimination are formulated in the book Causality by Pearl.

Example 2: Let's say that a teacher has a class with 100 students, 10 among them come from a disadvantaged background and their average in year 1 is 80/100. All the other students are normal and have an average of 90/100. Thus the average of her class was at 89/100. Since disadvantaged student were feeling themselves well with that particular teacher they advised other disadvantaged students to sign in into the course and in the year 2 the same teacher had 50 students from a disadvantageous background and 50 normal students. Normal student average became 91 and disadvantaged ones - 81. However, the average of the class dropped to 86 and she gets a call from the dean asking her why did she become so bad at teaching.

Conclusion:

Always look for a meaningful split of your data into classes that might have different behavior. If not, you might obtain a correlation opposed to the real one. If your data don't look like a Gaussian, don't try to pretend it is a Gaussian and proceed anyway: instead try to split it into classes that look Gaussian.

The examples discussed above in general incidate that, in real world data collection, it is common to have many relevant variables (e.g. in market research surveys; typical surveys have ~200 variables). Typically researchers pore over many crosstabs, however, it can be difficult to make sense of these, and the crosstabs may be misleading. Multivariate analysis (MVA) can help summarize the data and also can reduce the chance of obtaining spurious results.

1.4.2 Multivariate Analysis methods

There are two general types of Multivariate analysis as given below:

a) **Analysis of dependence:-** If the variables are dependent on others, they are called analysis of dependence. ie,a category of multivariate statistical techniques; dependence methods explain or predict a dependent variable(s) on the basis of two or more independent variables

   E.g. Multiple and Partial Least Square(PLS) regression, Multiple Discriminant Analysis(MDA)

b) **Analysis of interdependence:-** If the variables are not dependent on others, they are called analysis of interdependence.ie, a category of multivariate statistical techniques; interdependence methods give meaning to a set of variables or seek to group things together

   E.g. Cluster analysis, factor analysis

1.4.3 Data normalization

When analyzing two or more attributes it is often necessary to normalize the values of the attributes, especially in those cases where the values are vastly different in scale.

a)**Range normalization:** Let *X be an attribute and let* $x_1, x_2, \ldots, x_n$ *be a random sample drawn from X. In range normalization each value is scaled by the sample range of X.*

$$x_i' = \frac{x_i - \min_i\{x_i\}}{\hat{r}} = \frac{x_i - \min_i\{x_i\}}{\max_i\{x_i\} - \min_i\{x_i\}}$$

The mean vector is often referred to as the centroid and the variance-covariance matrix as the dispersion or dispersion matrix. Also, the terms variance-covariance matrix and covariance matrix are used interchangeably. After transformation the new attribute takes on values in the range [0,1]

b)Standard score normalization: There are various techniques available for transforming indicators in pure, dimensionless numbers, a process called normalization. Standardization or z-scores is the most commonly used method. It converts (using equation given below) all indicators to a common scale with an average of zero and standard deviation of one.

$$x_i' = \frac{x_i - \hat{\mu}}{\hat{\sigma}}$$

where, μ is the sample mean and σ^2 is the sample variance of *X*. After transformation, the new attribute has mean μ′ = 0, and standard deviation σ′ = 1.

A Z-Score is a statistical measurement of a score's relationship to the mean in a group of scores. A Z-score of 0 means the score is the same as the mean. A Z-score can also be positive or negative, indicating whether it is above or below the mean and by how many standard deviations.

Case Study

DETERMINANTS OF TRUST IN THE INDONESIAN POTATO INDUSTRY:
A COMPARISON AMONG GROUPS OF POTATO FARMERS

Source: http://ageconsearch.umn.edu/bitstream/100699/2/Puspitawati%20E.pdf

This is a study about Indonesia's potato industry, presenting producers with new and profitable opportunities to participate in sales to the modern channels. However, few farmers are involved in the new channels. This study offers an analysis of three groups of potato farmers' perceptions of trust in their buyers. The aim is to understand the many different ways producers can enter modern chains and how different channels suit the individual characteristics of different producers. In this study 50 farmer field schools (FFS) producers, 60 Indofood suppliers, and 192 general potato farmers (GPF) in the largest potato producing area in Indonesia, West Java were surveyed. Using MANOVA and linear regression methods, the study reveals that flexibility and dependence are determinate factors of trust in the three groups. Particularly among the FFS producers, relative price and firm size are factors identified to increase the farmers'trust. Farmers contracting with Indofood establish the relationship with the firm in terms of reputation and flexibility. On the other hand, the GPF has more concerns about buyers offering price transparency and joint problem solving. This article provides a conceptual model and an empirical analysis of the buyer-seller relationship in the potato industry in Indonesia.

As the study objective is to compare the level of trust and its antecedents among the three farmer groups the independent variable is the farmer groups and the dependent variables are trust, its antecedents and the demographic variables. Multivariate analysis of variance (MANOVA) and post-hoc test were done in order to test the hypotheses that there is a significant difference in the level of trust, its antecedents and the demographic factors among the groups. Multivariate differences across groups were assessed using the Wilks‟ Lambda criterion (known as the U statistics). The test examines whether groups are somehow different without being concerned with whether they differ on at least one linear combination of dependent variable. Independent variables that determine trust like flexibility, price transparency, relative price, price quality ratio, communication, dependence, reputation, flexibility and joint problem solving were measured on a five-point likert scale Finally, the variables identified are modeled in a linear regression model to know which dependent variables influence trust.

Summary:

- In real world data comes with multiple features/attributes.
- Exploratory data analysis requires exploiting the associations, correlations among these attributes.
- Attributes with different scales must be normalized for improving the computation.