# Multivariate Analysis using SPSS

## Bivariate & Multivariate Analysis - Correlation & Regression

Dr. M. Kamakshaiah
Assistant Professor - Business Analytics,
GSIB - GITAM (Deemed to be University)
kamakshaiah.musunuru@gitam.edu

## Bivariate Correlation

### Correlation coefficient

- Cos $\theta$
- Karl Pearson's r

### Significance test

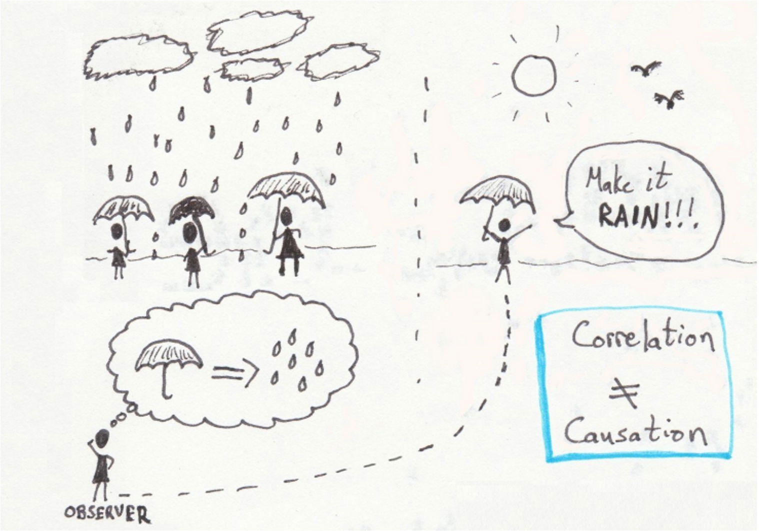
- Spearman's rank correlation
- Kendall's $\tau$
- Significance Test
- Goodman and Kruskal's $\gamma$

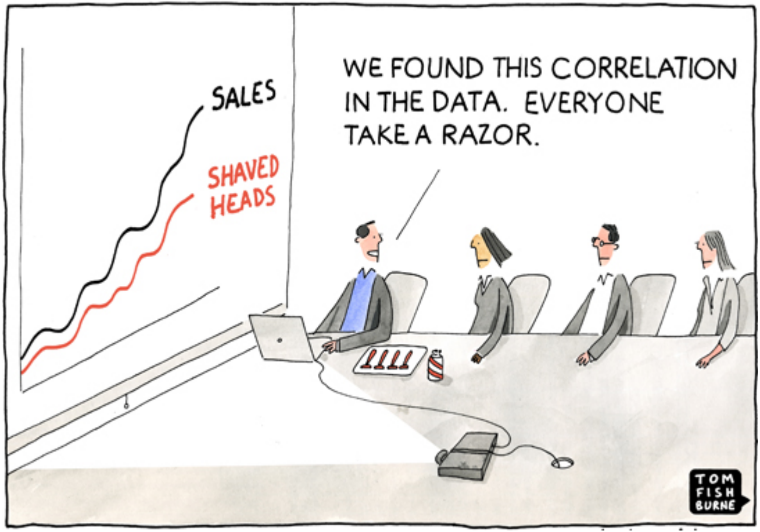
## Canonical Correlation

- CCA - Methodology

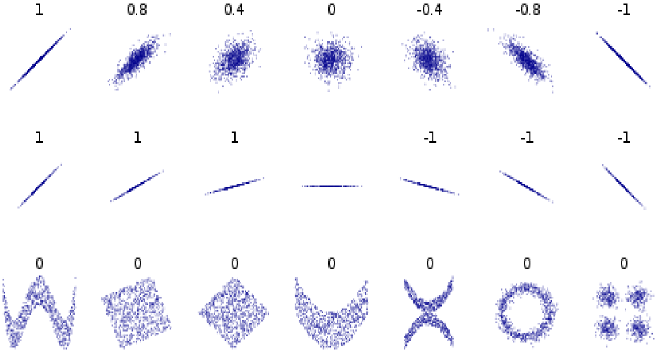
# Correlation - Relationship

In statistics, dependence or association is any statistical relationship, whether causal or not, between two random variables or bivariate data. *Correlation* is any of a broad class of statistical relationships involving dependence, though in common usage it most often refers to how close two variables are to having a linear relationship with each other.


Make it RAIN!!!
Correlation ≠ Causation
OBSERVER


WE FOUND THIS CORRELATION IN THE DATA. EVERYONE TAKE A RAZOR.
SALES
SHAVED HEADS
TOM FISHBURNE

# Patterns




Exercise: Use Excel spreadsheet to make visuals

# Correlation coefficient

A correlation *coefficient* is a numerical measure of some type of correlation, meaning a statistical relationship between two variables.
The variables may be two columns of a given data set of observations, often called a sample, or two *components* of a multivariate random variable with a known distribution.

# Correl. Coef. - Interpretations

1. They all assume values in the range from -1 to +1, where +1 indicates the strongest possible agreement and -1 the strongest possible disagreement.
2. The *Pearson product-moment correlation coefficient*, also known as *Pearson's r*, is a measure of the strength and direction of the linear relationship between two variables.
   - $-0.3 <= r <= 0.3$ (Weak)
   - $\pm 0.5 <= r <= \pm 0.7$ (Moderate)
   - $r > \pm 0.7$ (Strong)

[1] [2]

[1] Dr. Smith, Interpreting Correlation Coefficents, Available at `https://campus.fsu.edu`

[2] D. J. Rumsey, How to Interpret a Correlation Coefficient $r$. Retrieved from `http://www.dummies.com/education`

# Cos Function

$$\cos\theta = \frac{\mathbf{x}\cdot\mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|}$$

Note: SPSS can't help you!

# Karl Pearson's $r$

Defined as the covariance of the variables divided by the product of their standard deviations.

For Population -

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

For Sample -

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

# Cos $\theta$ vs. Karl Pearson's $r$

Excercise: make abstract or simulated data sets in Excel and try to vivisect Cos $\theta$ and Karl Pearson's $r$.

Note: SPSS can't help you!

Pearson's correlation coefficient follows Student's t-distribution with degrees of freedom n - 2.

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

$$r = \frac{t}{\sqrt{n-2+t^2}}$$

# Spearman's $\rho$

Spearman's rank correlation coefficient or Spearman's rho, often denoted by the Greek letter $\rho$ (rho) or as $r_s$, is a nonparametric measure of rank correlation (statistical dependence between the rankings of two variables).

$$\boxed{r_s = \rho_{\mathrm{rg}_X,\mathrm{rg}_Y} = \frac{\mathrm{cov}(\mathrm{rg}_X,\mathrm{rg}_Y)}{\sigma_{\mathrm{rg}_X}\sigma_{\mathrm{rg}_Y}}}$$

If all n ranks are distinct integers, it can be computed using the popular formula.

$$\boxed{r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}}$$

# Spearman's $\rho$

1. The Spearman correlation between two variables is equal to the Pearson correlation between the rank values of those two variables.
2. Pearson's correlation assesses linear relationships, Spearman's correlation assesses monotonic relationships.
3. If there are no repeated data values, a perfect Spearman correlation of $+1$ or 1 occurs.
4. the Spearman correlation between two variables will be high when observations have a similar rank.
5. Spearman's coefficient is appropriate for both continuous and discrete ordinal variables. [3]
6. Spearman's $\rho$ and Kendall's $\tau$ can be formulated as special cases of a more general correlation coefficient

[3] Lehman, Ann (2005). Jmp For Basic Univariate And Multivariate Statistics: A Step-by-step Guide. Cary, NC: SAS Press. p. 123. ISBN 1-59047-576-3.

Kendall rank correlation coefficient, commonly referred to as Kendall's $\tau$ coefficient, is a statistic used to measure the ordinal association between two measured quantities.

A tau test is a nonparametric hypothesis test for statistical dependence based on the $\tau$ coefficient.

The Kendall $\tau$ coefficient is defined as:

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{n(n-1)/2}$$

$$\tau = \frac{1}{n(n-1)} \sum_{i \neq j} \operatorname{sgn}(x_i - x_j) \operatorname{sgn}(y_i - y_j)$$

# Significance Test

For larger samples, it is common to use an approximation to the normal distribution, with mean zero and variance

$$\frac{2(2n+5)}{9n(n-1)}$$

# Goodman and Kruskal's $\gamma$

Goodman and Kruskal's gamma is a measure of rank correlation, i.e., the similarity of the orderings of the data when ranked by each of the quantities.

It measures the strength of association of the cross tabulated data when both variables are measured at the ordinal level.

$$G = \frac{N_s - N_d}{N_s + N_d}$$

$$t \approx G\sqrt{\frac{N_s + N_d}{n(1 - G^2)}}$$

# Canonical correlation

Canonical is the statistical term for analyzing latent variables (which are not directly observed) that represent multiple variables (which are directly observed).

1. Canonical-correlation analysis (CCA) is a way of inferring information from cross-covariance matrices.
2. If we have two vectors $X = (X1, ..., Xn)$ and $Y = (Y1, ..., Ym)$ of random variables, and there are correlations among the variables, then canonical-correlation analysis will find linear combinations of the $Xi$ and $Yj$ which have maximum correlation with each other.
3. Canonical correlation analysis requires the multivariate normal and homogeneity of variance assumption.
4. Canonical correlation analysis assumes a linear relationship between the canonical variates and each set of variables.
5. Similar to multivariate regression, canonical correlation analysis requires a large sample size.

# CCA - Methodology

Let $\Sigma_{XX} = cov(X, X)$ and $\Sigma_{YY} = cov(Y, Y)$, the parameter to maximize is as follows.

$$\rho = \frac{a^T \Sigma_{XY} b}{\sqrt{a^T \Sigma_{XX} a} \sqrt{b^T \Sigma_{YY} b}}$$

Objective function: $\mathbf{a}', \mathbf{b}' = \text{argmax}_{\mathbf{a},\mathbf{b}} \, \text{corr}(a^T X, b^T Y)$

# CCA in SPSS

There is no GUI based process in SPSS for CCA or DCA. Got to you use Syntax Window.

```
manova x1 to x6
/discrim all
/print = sig(eig dim).
execute.
```