



HINDUSTAN SARVA VIDHYAPEETH

WE MAXIMIZE YOUR VALUE

HINDUSTAN SARVA VIDHYAPEETH

An Internationally Reputed And Recognised Autonomous Institution

Run by A Registered Not-For-Profit Autonomous Organisation Under Indian Govt. Act 2 of 1882.

INDUSTRY INTEGRATED EDUCATION PROGRAMME

Ph.D. COURSE WORK SYLLABUS

QUANTITATIVE TECHNIQUES

Quantitative Techniques:

Quantitative techniques: quantitative techniques are those statistical and operation research techniques which help in the decision making process especially concerning business and industry. Those techniques which provide the decision maker a systematic means of analysis based on the quantitative data in formulating policies for achieving pre-determined goals.

Classification: Quantitative techniques can be of two types

1. Statistical techniques; those techniques which are used in conducting the statistical inquiry concerning certain phenomenon.
2. Programming techniques: these are the model building techniques used by decision maker.

Role of quantitative techniques: this technique greatly helps in handling the many complex problems.

- The role can be understood under following heads:
- They provide a tool for scientific analysis.
- They provide solution for various business problems. They enable proper use of resources.
- They help in minimizing waiting and service costs.
- They enable the management to decide when to buy and how much to buy. They assist in choosing an optimum strategy.
- They render greater help in optimum resource allocation. They facilitate the process of decision making.

Chi square test (χ^2 test):

Chi-square is a statistical test commonly used to compare observed data with data we would expect to obtain according to a specific hypothesis. For example, if, according to Mendel's laws, you expected 10 of 20 offspring from a cross to be male and the actual observed number was 8 males, then you might want to know about the "goodness to fit" between the observed and expected. Were the deviations (differences between observed and expected) the result of chance, or were they due to other factors. How much deviation can occur before you, the investigator, must conclude that something other than chance is at work, causing the observed to differ from the expected. The chi-square test is always testing what scientists call the null hypothesis, which states that there is no significant difference between the expected and observed result. It is calculated as:

$$X^2 = \sum (o - e)^2 / e$$
 Where o refers to the observed frequencies and e refers to the expected frequencies.

Example:

Suppose that we flip a coin 20 times and record the frequency of occurrence of heads and tails. We know from the laws of probability that we should expect 10 heads and 10 tails. We also know that because of sampling error we could easily come up with 9 heads and 11 tails or 12 heads and 8 tails. Let us suppose our coin-flipping experiment yielded 12 heads and 8 tails. We would enter our expected frequencies (10 - 10) and our observed frequencies (12 - 8) in a table.

	Observed	Expected	(fo-fe)	(fo-fe) ²	(fo-fe) ² /fe
Heads	12	10	2	4	0.4
Tails	8	10	-2	4	0.4
	20	20			0.8

The calculation of χ^2 in a one-way classification is very straight forward. The expected frequency in a category ("heads") is subtracted from the observed frequency, and the difference is squared, and the square is divided by its expected frequency. This is repeated for the remaining categories, and as the formula for χ^2 indicates, these results are summed for all categories. How does a calculated χ^2 of 0.8 tell us if our observed results of 12 heads and 8 tails represent a significant deviation from an expected 10-10 split? The shape of the chi square sampling distribution depends upon the number of degrees of freedom. The degree of freedom for a one-way classification χ^2 is $r - 1$, where r is the number of levels. In our problem above $r = 2$, so there would obviously be 1 degree of freedom. From our statistical reference tables, a χ^2 of 3.84 or greater is needed for χ^2 to be significant at the .05 level, so we conclude that our χ^2 of 0.8 in the coin-flipping experiment could have happened by sampling error and the deviations between the observed and expected frequencies are not significant. We would expect any data set yielding a calculated χ^2 value less than 3.84 with one degree of freedom at least 5% of the time due to chance alone. Therefore, the observed difference is not statistically significant at the .05 level.

Correlation:

Correlation is a measure of the relation between two or more variables. The correlation analysis involves various methods and techniques used for studying and measuring the extent of the relationship between two variables. So correlation analysis is a statistical procedure by which we can determine the degree of association or relationship between two or more variables.

Coefficient of correlation:

Coefficient of correlation is a measure of such a tendency, i.e. the degree to which the two variables are interrelated is measured by a coefficient which is called the coefficient of correlation.

Properties of correlation coefficient:

- The coefficient of correlation lies between +1 and -1, i.e. $-1 \leq r \leq +1$.
- The coefficient of correlation is independent of change of origin and scale of the variable x and y .
- The coefficient of correlation is the geometric mean of two regression coefficient. $R = b_{xy} \cdot b_{yx}$

- The degree of relationship between the two variables is symmetric, $r_{xy} = r_{yx}$.

Importance of correlation:

- Most of the variables show some kind of relationship, with the help of correlation analysis we can measure in one figure the degree of relationship between these variables.
- Once we know that two variables are closely related, we can estimate the value of one variable given the value of another.
- Correlation analysis contributes to the understanding of economic behaviour.

The various types of correlation are as follow:

- Positive correlation: if both the variables are varying in the same direction i. E. If as one variable increasing the other also increasing on an average and if one is decreasing then other also decreasing, then the correlation is said to be positive.
- Negative correlation: if both the variables are varying in the opposite direction i. E. If as one variable increasing the other is decreasing then the correlation is said to be negative correlation.
- Simple correlation: when only two variables are studied it is a problem of simple correlation.
- Multiple or partial correlation: when three or more variables are studied then it is the problem of either multiple or partial correlation.
- Linear correlation: if the amount of change in one variable tends to bear constant ratio to the amount of change in the other variable then the correlation is said to be linear.
- Non- linear (curvilinear) correlation: correlation is said to be non- linear(curvilinear) if the amount of change in one variable does not bear a constant ratio to the amount of change in the other variable.

The various methods of ascertain whether two variables are correlated or not are: Scatter diagram method Graphic method Karl Pearson's coefficient of correlation

Scattered diagram method:

Scattered diagram is a graphical method of showing the correlation between the two variables x and y. The value of x and y is plotted on x- axis and y- axis by choosing suitable scale. Thus corresponding to every ordered pair (x_i, y_j) there corresponds a point or a dot in the coordinate plane. The diagram of dots or points so obtained is called a scattered diagram. The scattered diagram may indicate both degree and the type of correlation.

Graphical method:

In this method individual values of the two variables are plotted on graph paper. We thus obtain two curves, one for x variable and another for y variable; by examining the direction and closeness of the two curves drawn we can infer whether or not the variables are related. If both the curves drawn on the graph are moving in the same direction correlation is said to be positive. And if the curves are moving in the opposite directions correlation is said to be negative.

Karl Pearson's coefficient of correlation: Karl Pearson is a mathematical method of measuring correlation. It is most commonly used method. Karl Pearson's coefficient of correlation is denoted by symbol r. The formula for computing r is:

$R = \text{covariance of } x \text{ and } y / (\text{standard deviation of } x)(\text{standard deviation of } y)$

Example:

Find the correlation of given data.

Age	Minimum stopping at 40 kp
Months	Meters
9	28.4
15	29.3
24	37.6
30	36.2
38	36.5
46	35.3
53	36.2
60	44.1
64	44.8
76	47.2

$X\text{-bar} = 415/10 = 41.5$ y

$\text{Bar} = 376.6/10 = 37.7$

$R = 10 \times 16713.3 - 415 \times 375.6 / r = \{(10 \times 21623 - 4152) (10 \times 14457.72 - 375.62)\} 11259 / (44005 \times 3501.84) r = 11259 / 124.14$
 $r = 0.91$

X	Y	X ²	Y ²	XY	
	9	28.4	81	806.56	255.6
	15	29.3	225	858.49	439.5
	24	37.6	576	1413.76	902.4
	30	36.2	900	1310.44	1086
	38	36.5	1444	1332.25	1387
	46	35.3	2116	1246.09	1623.8
	53	36.2	2809	1310.44	1918.6
	60	44.1	3600	1944.81	2646
	64	44.8	4096	2007.04	2867.2
	76	47.2	5776	2227.84	3587.2
TOTAL	415	375.6	21623	14457.72	16713.3

$$r = \frac{n \sum XY - \sum X \sum Y}{\sqrt{[n \sum X^2 - (\sum X)^2][n \sum Y^2 - (\sum Y)^2]}}$$

Regression analysis:

A statistical measure that attempts to determine the strength of the relationship between one dependent variable (usually denoted by y) and a series of other changing variables (known as independent variables) Regression analysis includes any techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables. More specifically, regression analysis helps one understand how the typical value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are held fixed. Most commonly, regression analysis estimates the conditional expectation of the dependent variable given the independent variables — that is, the average value of the dependent variable when the independent variables are held fixed.

When to use regression analysis

- Regression analysis provides estimates of values of the dependent variable from values of the independent variables. The device used to accomplish this estimation procedure is the regression line. The regression line describes the average relationship existing between x and y variables i.e. it displays mean values of x for given values of y ,
- A second goal of regression analysis is to obtain a measure of the error involved in using the regression line as a basis for estimation. For this purpose the standard error of estimate is calculated. This is a measure of the scatter or spread of the observed values of y around the corresponding values estimated from the regression line.

Regression equation:

Regression equation, also known as estimating equations, is algebraic expressions of the regression lines.

Elements of a regression equation:

The regression equation is written as $y = a + bx + e$

Y is the value of the dependent variable (y), what is being predicted or explained A or α , a constant; equals the value of y when the value of $x=0$

B or β , the coefficient of x ; the slope of the regression line; how much y changes for each one-unit change in x .

X is the value of the independent variable (x), what is predicting or explaining the value of y

E is the error term; the error in predicting the value of y , given the value of x (it is not displayed in most regression equations).

For example, say we know what the average speed is of cars on the freeway when we have 2 highway patrols deployed (average speed=75 mph) or 10 highway patrols deployed (average speed=35 mph).

But what will be the average speed of cars on the freeway when we deploy 5 highway patrols?

Average speed on freeway (y)	Number of patrol cars deployed (x)
75	2
35	10

From our known data, we can use the regression formula (calculations not shown) to compute the values of and obtain the following equation: $y = 85 + (-5)x$, where Y is the average speed of cars on the freeway $A=85$, or the average speed when $x=0$ $B=(-5)$, the impact on y of each additional patrol car deployed X is the number of patrol cars deployed That is, the average speed of cars on the freeway when there are no highway patrol's working ($x=0$) will be 85 mph. For each additional highway patrol car working, the average speed will drop by 5 mph. For five patrols ($x=5$), $y = 85 + (-5)(5) = 85 - 25 = 60$ mph There may be some variations on how regression equations are written in the literature. For example, you may sometimes see the dependent variable term (y) written with a little "hat" (^) on it, or called y -hat. This refers to the predicted value of y . The plain y refers to observed values of y in the data set used to calculate the regression equation. You may see the symbols for alpha (α) and beta (β) written in Greek letters, or you may see them written in English letters. The coefficient of the independent variable may have a subscript, as may the term for x , for example, b_1x_1 (this is common in multiple regression).

Steps in linear regression:

1. State the hypothesis.
2. State the null hypothesis
3. Gather the data.
4. Compute the regression equation
5. Examine tests of statistical significant and measures of association
6. Relate statistical findings to the hypothesis. Accept or reject the null hypothesis.
7. Reject, accept or revise the original hypothesis. Make suggestions for research design and management aspects of the problem.

Example: the motor pool wants to know if it costs more to maintain cars that are driven more often. Hypothesis: maintenance costs are affected by car mileage null hypothesis: there is no relationship between mileage and maintenance costs.

Dependent variable: y is the cost in dollars of yearly maintenance on a motor vehicle Independent variable: x is the yearly mileage on the same motor vehicle

Data are gathered on each car in the motor pool, regarding number of miles driven in a given year, and maintenance costs for that year. Here is a sample of the data collected.

Car number	Miles driven (x)	Repair costs (y)
1	80,000	1,200
2	29,000	150
3	53,000	650
4	13,000	200
5	45,000	325

The regression equation is computed as (computations not shown): $y = 50 + .03x$ For example, if $x=50,000$ then $y = 50 + .03(50,000) = 1,550$

$A=50$ or the cost of maintenance when $x=0$; if there is no mileage on the car, then the yearly cost of maintenance= 50

$B=.03$ the value that y increases for each unit increase in x ; for each extra mile driven (x), the cost of yearly maintenance increases by $.03$

$S.E.B = .0005$; the value of b divided by $S.E.B=60.0$; the t -table indicates that the b coefficient of x is statistically significant (it is related to y)

$R^2=.90$ we can explain 90% of the variance in repair costs for different vehicles if we know the vehicle mileage for each car

Conclusion: reject the null hypothesis of no relationship and accept the research hypothesis, that mileage affects repair costs.

Discriminant analysis

Discriminant analysis may be used for two objectives: either we want to assess the adequacy of classification, given the group memberships of the objects under study; or we wish to assign objects to one of a number of (known) groups of objects. Discriminant analysis may thus have a descriptive or a predictive objective.

In both cases, some group assignments must be known before carrying out the discriminant analysis. Such group assignments, or labelling, may be arrived at in any way. Hence discriminant analysis can be employed as a useful complement to cluster analysis (in order to judge the results of the latter) or principal components analysis. Alternatively, in star-galaxy separation, for instance, using digitized images, the analyst may define group (stars, galaxies) membership visually for a conveniently small training set or design set. Methods implemented in this area are multiple discriminant analysis, Fisher's linear discriminant analysis, and k -nearest neighbours discriminant analysis.

Multiple discriminant analysis:

(MDA) is also termed discriminant factor analysis and canonical discriminant analysis. It adopts a similar perspective to PCA: the rows of the data matrix to be examined constitute points in a multidimensional space, as also do the group mean vectors. Discriminating axes are determined in this space, in such a way that optimal separation of the predefined groups is attained. As with PCA, the problem becomes mathematically the Eigen reduction of a real, symmetric matrix. The eigenvalues represent the discriminating power of the associated eigenvectors. The n_y groups lie in a space of dimension at most $n_y - 1$. This will be the number of discriminant axes or factors obtainable in the most common practical case when $n > m > n_y$ (where n is the number of rows, and m the number of columns of the input data matrix).

Linear discriminant analysis:

It is the 2-group case of MDA. It optimally separates two groups, using the Mahalanobis metric or generalized distance. It also gives the same linear separating decision surface as Bayesian maximum likelihood discrimination in the case of equal class covariance matrices.

Purpose:

The main purpose of a discriminant function analysis is to predict group membership based on a linear combination of the interval variables. The procedure begins with a set of observations where both group membership and the values of the interval variables are known. The end result of the procedure is a model that allows prediction of group membership when only the interval variables are known. A second purpose of discriminant function analysis is an understanding of the data set, as a careful examination of the prediction model that results from the procedure can give insight into the relationship between group membership and the variables used to predict group membership.

Examples:

For example, a graduate admissions committee might divide a set of past graduate students into two groups: students who finished the program in five years or less and those who did not. Discriminant function analysis could be used to predict successful completion of the graduate program based on GRE score and undergraduate grade point average. Examination of the prediction model might provide insights into how each predictor individually and in combination predicted completion or noncompletion of a graduate program. Another example might predict whether patients recovered from a coma or not based on combinations of demographic and treatment variables. The predictor variables might include age, sex, general health, time between incident and arrival at hospital, various interventions, etc. In this case the creation of the prediction model would allow a medical practitioner to assess the chance of recovery based on observed variables. The prediction model might also give insight into how the variables interact in predicting recovery.

Factor analysis

Factor analysis is a statistical method used to describe variability among observed variables in terms of a potentially lower number of unobserved variables called factors. In other words, it is possible, for example, that variations in three or four observed variables mainly reflect the variations in a single unobserved variable, or in a reduced number of unobserved variables. Factor analysis searches for such joint variations in response to unobserved latent variables. The observed variables are modeled as linear combinations of the potential factors, plus "error" terms. The information gained about the interdependencies between observed variables can be used later to reduce the set of variables in a dataset. Factor analysis can be applied in order to explore a content area, structure a domain, map unknown concepts, classify or reduce data, illuminate causal nexuses, screen or transform data, define relationships, test hypotheses, formulate theories, control variables, or make inferences. Our consideration of these various overlapping usages will be related to several aspects of scientific method: induction and deduction; description and inference; causation, explanation, and classification; and theory.

Uses of Factor Analysis

This section will outline factor analysis applications relevant to various scientific and policy concerns. Many of the uses described below overlap. My aim is not to avoid redundancy but explicitly to relate factor analysis to the diverse interests of readers.

Interdependency and Pattern Delineation

If a scientist has a table of data--say, un votes, personality characteristics, or answers to a questionnaire--and if he suspects that these data are interrelated in a complex fashion, then factor analysis may be used to untangle the linear relationships into their separate patterns. Each pattern will appear as a factor delineating a distinct cluster of interrelated data.

Parsimony or Data Reduction

Factor analysis can be useful for reducing a mass of information to an economical description. For example, data on fifty characteristics for 300 nations are unwieldy to handle, descriptively or analytically. The management, analysis, and understanding of such data are facilitated by reducing them to their common factor patterns. These factors concentrate and index the dispersed information in the original data and can therefore replace the fifty characteristics without much loss of information. Nations can be more easily discussed and compared on economic development, size, and politics dimensions, for example, than on the hundreds of characteristics each dimension involves.

Structure

Factor analysis may be employed to discover the basic structure of a domain. As a case in point, a scientist may want to uncover the primary independent lines or dimensions--such as size, leadership, and age--of variation in group characteristics and behavior. Data collected on a large sample of groups and factor analyzed can help disclose this structure.

Classification or description

Factor analysis is a tool for developing an empirical typology.⁷ it can be used to group interdependent variables into descriptive categories, such as ideology, revolution, liberal voting, and authoritarianism. It can be used to classify nation profiles into types with similar characteristics or behavior. Or it can be used on data matrices of a transaction type or a social-choice type to show how individuals, social groups, or nations cluster on their transactions with or choices of each other.

Scaling

A scientist often wishes to develop a scale on which individuals, groups, or nations can be rated and compared. The scale may refer to such phenomena as political participation, voting behavior, or conflict. A problem in developing a scale is to weight the characteristics being combined. Factor analysis offers a solution by dividing the characteristics into independent sources of variation

(factors). Each factor then represents a scale based on the empirical relationships among the characteristics. As additional findings, the factor analysis will give the weights to employ for each characteristic when combining them into the scales. The factor score results (see section 4.5 below) are actually such scales, developed by summing characteristics times these weights.

Hypothesis testing

Hypotheses abound regarding dimensions of attitude, personality, group, social behavior, voting, and conflict. Since the meaning usually associated with "dimension" is that of a cluster or group of highly inter-correlated characteristics or behavior, factor analysis may be used to test for their empirical existence. Which characteristics or behavior should, by theory, be related to which dimensions can be postulated in advance and statistical tests of significance can be applied to the factor analysis results.

Besides those relating to dimensions, there are other kinds of hypotheses that may be tested. To illustrate: if the concern is with a relationship between economic development and instability, holding other things constant, a factor analysis can be done of economic and instability variables along with other variables that may affect (hide, mediate, depress) their relationship. The resulting factors can be so defined (rotated) that the first several factors involve the mediating measures (to the maximum allowed by the empirical relationships). A remaining independent factor can be calculated to best define the postulated relationships between the economic and instability measures. The magnitude of involvement of both variables in this pattern enables the scientist to see whether an economic development-instability pattern actually exists when other things are held constant.

Data Transformation

Factor analysis can be used to transform data to meet the assumptions of other techniques. For instance, application of the multiple regression technique assumes (if tests of significance are to be applied to the regression coefficients) that predictors--the so-called independent variables—are statistically unrelated (EZEKIEL and fox, 1959, pp. 283-84). If the predictor variables are correlated in violation of the assumption, factor analysis can be employed to reduce them to a smaller set of uncorrelated factor scores. The scores may be used in the regression analysis in place of the original variables, with the knowledge that the meaningful variation in the original data has not been lost. Likewise; a large number of dependent variables also can be reduced through factor analysis.

Exploration

In a new domain of scientific interest like peace research, the complex interrelations of phenomena have undergone little systematic investigation. The unknown domain may be explored through factor analysis. It can reduce complex interrelationships to a relatively simple linear expression and it can uncover unsuspected, perhaps startling, relationships. Usually the social scientist is unable to manipulate variables in a laboratory but must deal with the manifold complexity of behaviors in their social setting. Factor analysis thus fulfills some functions of the laboratory and enables the scientist to untangle interrelationships, to separate different sources of variation, and to partial out or control for undesirable influences on the variables of concern.⁹

Mapping

Besides facilitating exploration, factor analysis also enables a scientist to map the social terrain. By mapping I mean the systematic attempt to chart major empirical concepts and sources of variation. These concepts may then be used to describe a domain or to serve as inputs to further research. Some social domains, such as international relations, family life, and public administration, have yet to be charted. In some other areas, however, such as personality, abilities, attitudes, and cognitive meaning, considerable mapping has been done.

Cluster analysis:

Cluster analysis' is a class of statistical techniques that can be applied to data that exhibit —natural groupings. Cluster analysis sorts through the raw data and groups them into clusters. A cluster is a group of relatively homogeneous cases or observations. Objects in a cluster are similar to each other. They are also dissimilar to objects outside the cluster, particularly objects in other clusters. Cluster analysis, like factor analysis and multi-dimensional scaling, is an interdependence technique: it makes no distinction between dependent and independent variables. The entire set of interdependent relationships is examined. It is similar to multi-dimensional scaling in that both examine inter-object similarity by examining the complete set of interdependent relationships. The difference is that multidimensional scaling identifies underlying dimensions, while cluster analysis identifies clusters. Cluster analysis is the obverse of factor analysis. Whereas factor analysis reduces the number of variables by grouping them into a smaller set of factors, cluster analysis reduces the number of observations or cases by grouping them into a smaller set of clusters.

Types of clustering:

Hierarchical algorithms find successive clusters using previously established clusters. These algorithms usually are either agglomerative ("bottom-up") or divisive ("topdown"). Agglomerative algorithms begin with each element as a separate cluster and merge them into successively larger clusters. Divisive algorithms begin with the whole set and proceed to divide it into successively smaller clusters.

Partition

Algorithms typically determine all clusters at once, but can also be used as divisive algorithms in the hierarchical clustering. Density-based clustering algorithms are devised to discover arbitrary-shaped clusters. In this approach, a cluster is regarded as a region in which the density of data objects exceeds a threshold. DB scan and optics are two typical algorithms of this kind. Subspace clustering methods look for clusters that can only be seen in a particular projection (subspace, manifold) of the data. These methods thus can ignore irrelevant attributes. The general problem is also known as correlation clustering while the special case of axis-parallel subspaces is also known as two-way clustering, co-clustering or bi-clustering: in these methods not only the objects are clustered but also the features of the objects, i.e., if the data is represented in a data matrix, the rows and columns are clustered simultaneously. They usually do not however work with arbitrary feature combinations as in general subspace methods. But this special case deserves attention due to its applications in

bioinformatics. Many clustering algorithms require the specification of the number of clusters to produce in the input data set, prior to execution of the algorithm. Barring knowledge of the proper value beforehand, the appropriate value must be determined, a problem on its own for which a number of techniques have been developed.

Clustering procedures:

There are several types of clustering methods:

1. Non-hierarchical clustering (also called k-means clustering)

First determine a cluster center, and then group all objects that are within a certain distance

Examples:

- Sequential threshold method - first determine a cluster center, then group all objects that are within a predetermined threshold from the center - one cluster is created at a time
- Parallel threshold method - simultaneously several cluster centers are determined, then objects that are within a predetermined threshold from the centers are grouped
- Optimizing partitioning method - first a non-hierarchical procedure is run, then objects are reassigned so as to optimize an overall criterion.

2. Hierarchical clustering

Objects are organized into a hierarchical structure as part of the procedure

Examples:

- Divisive clustering - start by treating all objects as if they are part of a single large cluster, then divide the cluster into smaller and smaller clusters
- Agglomerative clustering - start by treating each object as a separate cluster, then group them into bigger and bigger clusters

Examples:

- Centroid methods - clusters are generated that maximize the distance between the centers of clusters (a centroid is the mean value for all the objects in the cluster)
- Variance methods - clusters are generated that minimize the within-cluster variance

Example:

- Ward's procedure - clusters are generated that minimize the squared Euclidean distance to the center mean
- Linkage methods - cluster objects based on the distance between them
- Single linkage method - cluster objects based on the minimum distance between them (also called the nearest neighbor rule)
- Complete linkage method - cluster objects based on the maximum distance between them (also called the furthest neighbor rule)

- Average linkage method - cluster objects based on the average distance between all pairs of objects (one member of the pair must be from a different cluster)

Multidimensional scaling (MDS):

Multidimensional scaling (MDS) is a set of related statistical techniques often used in information visualization for exploring similarities or dissimilarities in data. MDS is a special case of ordination. An MDS algorithm starts with a matrix of item-item similarities, and then assigns a location to each item in n-dimensional space, where n is specified a priori.

General purpose:

Multidimensional scaling (MDS) can be considered to be an alternative to factor analysis (see factor analysis). In general, the goal of the analysis is to detect meaningful underlying dimensions that allow the researcher to explain observed similarities or dissimilarities (distances) between the investigated objects. In factor analysis, the similarities between objects (e.g., variables) are expressed in the correlation matrix. With MDS, you can analyze any kind of similarity or dissimilarity matrix, in addition to correlation matrices.

Types:

MDS algorithms fall into a taxonomy, depending on the meaning of the input matrix:

Classical Multidimensional Scaling

Also known as Torgerson Scaling or Torgerson-Gower Scaling - takes an input matrix giving dissimilarities between pairs of items and outputs a coordinate matrix whose configuration minimizes a loss function called strain

Metric multidimensional scaling

A superset of classical MDS is generalizes the optimization procedure to a variety of loss functions and input matrices of known distances with weights and so on. A useful loss function in this context is called stress which is often minimized using a procedure called stress memorization.

Non-metric multidimensional scaling

In contrast to metric MDS, non-metric MDS finds both a non-parametric monotonic relationship between the dissimilarities in the item-item matrix and the Euclidean distances between items, and the location of each item in the low-dimensional space. The relationship is typically found using isotonic regression. Louisguttman's smallest space analysis (SSA) is an example of a non-metric MDS procedure.

Generalized multidimensional scaling

An extension of metric multidimensional scaling the target space is an arbitrary smooth non-Euclidean space. In case when the dissimilarities are distances on a surface and the target space is another surface, GMDS allows finding the minimum-distortion embedding of one surface into another.

Procedure for MDS research:

There are several steps in conducting MDS research:

1. **Formulating the problem** - what variables do you want to compare? How many variables do you want to compare? More than 20 are often considered cumbersome. Fewer than 8 (4 pairs) will not give valid results. What purpose is the study to be used for?
2. **Obtaining input data** - respondents are asked a series of questions. For each product pair they are asked to rate similarity (usually on a 7 point likert scale from very similar to very dissimilar). The first question could be for coke/Pepsi for example, the next for coke/hires root beer, the next for Pepsi/dr pepper, the next for dr pepper/hires root beer, etc. The number of questions is a function of the number of brands and can be calculated as $q = n(n - 1) / 2$ where q is the number of questions and n is the number of brands. This approach is referred to as the —perception data: direct approach. There are two other approaches. There is the —perception data: derived approach in which products are decomposed into attributes which are rated on a semantic differential scale. The other is the —preference data approach in which respondents are asked their preference rather than similarity.
3. **Running the MDS statistical program** - software for running the procedure is available in many software for statistics. Often there is a choice between metric MDS (which deals with interval or ratio level data), and nonmetric MDS (which deals with ordinal data).
4. **Decide number of dimensions** - the researcher must decide on the number of dimensions they want the computer to create. The more dimensions, the better the statistical fit, but the more difficult it is to interpret the results.
5. **Mapping the results and defining the dimensions** - the statistical program (or a related module) will map the results. The map will plot each product (usually in two dimensional spaces). The proximity of products to each other indicates either how similar they are or how preferred they are, depending on which approach was used. The dimensions must be labeled by the researcher. This requires subjective judgment and is often very challenging. [vague] the results must be interpreted (see perceptual mapping) [vague]
6. **Test the results for reliability and validity** - compute r-squared to determine what proportion of variance of the scaled data can be accounted for by the MDS procedure. An r-square of 0.6 is considered the minimum acceptable level. [citation needed] an r-square of 0.8 is considered good for metric scaling and .9 is considered good for non-metric scaling. Other possible tests are

kruskal's stress, split data tests, data stability tests (i.e., eliminating one brand), and test-retest reliability.

7. **Report the results comprehensively** - along with the mapping, at least distance measure (e.g. sorensen index, jaccard index) and reliability (e.g., stress value) should be given. It is also very advisable to give the algorithm (e.g., kruskal, mather) which is often defined by the program used (sometimes replacing the algorithm report), if you have given a start configuration or had a random choice, the number of runs, the assessment of dimensionality, the montecarlo method results, the number of iterations, the assessment of stability, and the proportional variance of each axis (r-square).

T-test:

The t-test (or student's t-test) gives an indication of the separateness of two sets of measurements, and is thus used to check whether two sets of measures are essentially different (and usually that an experimental effect has been demonstrated). The typical way of doing this is with the null hypothesis that means of the two sets of measures are equal.

The t-test assumes:

- A normal distribution (parametric data)
- Underlying variances are equal (if not, use welch's test)

It is used when there is random assignment and only two sets of measurement to compare. There are two main types of t-test:

- Independent-measures t-test: when samples are not matched. Matched
- Pair t-test: when samples appear in pairs (eg. Before-and-after).

A single-sample t-test compares a sample against a known figure, for example where measures of a manufactured item are compared against the required standard.

Calculation:

The value of t may be calculated using packages such as SPSS. The actual calculation for two groups is:

$$T = \frac{\text{experimental effect}}{\text{variability}}$$
$$= \frac{\text{difference between group means}}{\text{Standard error of difference between group means}}$$

Interpretation

The resultant t-value is then looked up in a t-table to determine the probability that a significant difference between the two sets of measures exists and hence what can be claimed about the efficacy of the experimental treatment.

Uses

Among the most frequently used t-tests are:

- A one-sample location test of whether the mean of a normally distributed population has a value specified in a null hypothesis.
- A two sample location test of the null hypothesis that the means of two normally distributed populations are equal. All such tests are usually called student's t-tests, though strictly speaking that name should only be used if the variances of the two populations are also assumed to be equal; the form of the test used when this assumption is dropped is sometimes called welch's t-test. These tests are often referred to as "unpaired" or "independent samples" t-tests, as they are typically applied when the statistical units underlying the two samples being compared are non-overlapping.
- A test of the null hypothesis that the difference between two responses measured on the same statistical unit has a mean value of zero. For example, suppose we measure the size of a cancer patient's tumor before and after a treatment. If the treatment is effective, we expect the tumor size for many of the patients to be smaller following the treatment. This is often referred to as the "paired" or "repeated measures" t-test see paired difference test
- A test of whether the slope of a regression line differs significantly from 0.

Calculation:

Explicit expressions that can be used to carry out various t-tests are given below. In each case, the formula for a test statistic that either exactly follows or closely approximates a t-distribution under the null hypothesis is given. Also, the appropriate degrees of freedom are given in each case. Each of these statistics can be used to carry out either a one-tailed test or a two-tailed test. Once a t value is determined, a p-value can be found using a table of values from student's t-distribution. If the calculated p-value is below the threshold chosen for statistical significance (usually the 0.10, the 0.05, or 0.01 level) then the null hypothesis is rejected in favor of the alternative hypothesis

Independent one-sample t-test

In testing the null hypothesis that the population means is equal to a specified value μ_0 , one uses the statistic

Where s is the sample standard deviation of the sample and n is the sample size. The degrees of freedom used in this test is $n - 1$.

Independent two-sample t-test:

Equal sample sizes, equal variance

This test is only used when both:

- The two sample sizes (that is, the number, n , of participants of each group) are equal;
- It can be assumed that the two distributions have the same variance.

Violations of these assumptions are discussed below. The t statistic to test whether the means are different can be calculated as follows:

Here s_{x_1, x_2} the grand standard deviation (or pooled standard deviation), 1 = group one, 2 = group two. The denominator of t is the standard error of the difference between two means. For significance testing, the degrees of freedom for this test are $2n - 2$ where n is the number of participants in each group.

$$t = \frac{X_1 - X_2}{S_{X_1 X_2} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Where

$$S_{X_1 X_2} = \sqrt{\frac{1}{2} (S_{X_1}^2 + S_{X_2}^2)}$$

Unequal sample sizes, equal variance

This test is used only when it can be assumed that the two distributions have the same variance. (When this assumption is violated, see below.) The t statistic to test whether the means are different can be calculated as follows:

Note that the formulae above are generalizations of the case where both samples have equal sizes (substitute n for n_1 and n_2).

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{X_1 X_2} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Where

$$S_{X_1 X_2} = \sqrt{\frac{(n_1 - 1)S_{X_1}^2 + (n_2 - 1)S_{X_2}^2}{n_1 + n_2 - 2}}$$

F test:

The f-distribution is formed by the ratio of two independent chi-square variables divided by their respective degrees of freedom. F-test (Snedecor and Cochran, 1983) is used to test if the standard deviations of two populations are equal. This test can be a two-tailed test or a one-tailed test. The two-tailed version tests against the alternative that the standard deviations are not equal. The one-tailed version only tests in one direction that is the standard deviation from the first population is either greater than or less than (but not both) the second population standard deviation. The choice is determined by the problem. For example, if we are testing a new process, we may only be interested in knowing if the new process is less variable than the old process. Since f is formed by chi-square, many of the chi-square properties carry over to the f distribution.

- The f-values are all non-negative
- The distribution is non-symmetric

- The mean is approximately 1
- There are two independent degrees of freedom, one for the numerator, and one for the denominator.
- There are many different f distributions, one for each pair of degrees of freedom.

The f-test is designed to test if two population variances are equal. It does this by comparing the ratio of two variances. So, if the variances are equal, the ratio of the variances will be 1.

$$F = \frac{S_1^2}{S_2^2}$$

If the null hypothesis is true, then the f test-statistic given above can be simplified (dramatically). This ratio of sample variances will be test statistic used. If the null hypothesis is false, then we will reject the null hypothesis that the ratio was equal to 1 and our assumption that they were equal.

There are several different f-tables. Each one has a different level of significance. So, find the correct level of significance first, and then look up the numerator degrees of freedom and the denominator degrees of freedom to find the critical value.

Example:

As an example, assume we want to see if a method (method 1) for measuring the arsenic concentration in soil is significantly more precise than a second method (method 2). Each method was tested ten times, with, yielding the following values:

Method	Mean (ppm)	Standard deviation (ppm)
1	6.7	0.8
2	8.2	1.2

A method is more precise if its standard deviation is lower than that of the other method. So we want to test the null hypothesis $H_0: \sigma_2^2 = \sigma_1^2$, against the alternate hypothesis $H_a: \sigma_2^2 > \sigma_1^2$. Since $s_2 > s_1$, $f_{calc.} = s_2^2/s_1^2 = 1.22/0.82 = 2.25$. The tabulated value for D.O.F. $v = 9$ in each case, and a 1-tailed, 95% confidence level is $f_{9,9} = 3.179$. In this case, $f_{calc.} < f_{9,9}$, so we accept the null hypothesis that the two standard deviations are equal, and we are 95% confident that any difference in the sample standard deviations is due to random error. We use a 1-tailed test in this case because the only information we are interested in is whether method 1 is more precise than method 2.

Z test:

Z-test is a statistical test where normal distribution is applied and is basically used for dealing with problems relating to large samples when $n \geq 30$.

Where n = sample size

The z-test compares sample and population means to determine if there is a significant difference. It requires a simple random sample from a population with a normal distribution and where the mean is known.

Calculation

The z measure is calculated as:

$$Z = (x - m) / se$$

Where x is the mean sample to be standardized, m (μ) is the population mean and se is the standard error of the mean.

$$Se = s / \sqrt{n}$$

Where s is the population standard deviation and n is the sample size. The z value is then looked up in a z-table. A negative z value means it is below the population mean (the sign is ignored in the lookup table).

Discussion

The z-test is typically with standardized tests, checking whether the scores from a particular sample are within or outside the standard test performance. The z value indicates the number of standard deviation units of the sample from the population mean.

Z-test's for different purposes

There are different types of z-test each for different purpose. Some of the popular types are outlined below:

1. Z test for single proportion is used to test a hypothesis on a specific value of the population proportion. Statistically speaking, we test the null hypothesis $h_0: p = p_0$ against the alternative hypothesis $h_1: p \neq p_0$ where p is the population proportion and p_0 is a specific value of the population proportion we would like to test for acceptance. The example on tea drinkers explained above requires this test. In that example, $p_0 = 0.5$. Notice that in this particular example, proportion refers to the proportion of tea drinkers.
2. Z test for difference of proportions is used to test the hypothesis that two populations have the same proportion. For example suppose one is interested to test if there is any significant difference in the habit of tea drinking between male and female citizens of a town. In such a situation, ztest for difference of proportions can be applied. One would have to obtain two independent samples from the town- one from males and the other from females and determine the proportion of tea drinkers in each sample in order to perform this test.
3. Z -test for single mean is used to test a hypothesis on a specific value of the population mean. Statistically speaking, we test the null hypothesis $h_0: \mu = \mu_0$ against the alternative hypothesis $h_1: \mu \neq \mu_0$ where μ is the population mean and μ_0 is a specific value of the population that we would like to test for acceptance. Unlike the t-test for single mean, this test is used if $n \geq 30$ and population standard deviation is known.
4. Z test for single variance is used to test a hypothesis on a specific value of the variance. Statistically speaking, we test the null hypothesis $h_0: \sigma = \sigma_0$ against $h_1: \sigma \neq \sigma_0$ where σ is the population mean and σ_0 is a specific value of the population variance that we would like to test for acceptance. In other words, this test enables us to test if the given sample has been drawn from a population with specific variance σ_0 . Unlike the chi square test for single variance, this test is used if $n \geq 30$.
5. Z-test for testing equality of variance is used to test the hypothesis of equality of two population variances when the sample size of each sample is 30 or larger.