種原庫資料之統計、分析與應用

台灣大學 農藝系
劉 清

KINDS OF DESCRIPTORS

1. General information

   Variety or accession number
   Collection information (collector, date, location, etc.)
   Nomenclature information (scientific and common)
   Origin (country, state or province, and precise lacality)
   Storage information (location, date placed in storage, etc.)
   Distribution information

2. Environmental information

   Altitute
   Latitute
   Longitude
   Climatic
   Ecological
   Soil description

3. Organismic information

   Morpho-agronomic
   Genetic

Physiological
Biochemical
Environmental damage (cold, drought, wind or other susceptibility)
Pest and disease information (bacterial, fungal, viral, insect, nematode, etc.)
Rejuvenation (viability of seeds, time in storage)

4. Use information

Including breeding, genetics, performance tests, etc.)

5. Food characteristics

6. Industrial characteristics

7. Bibliographic information

KINDS OF DESCRIPTORS

1. General information

   Variety or accession number
   Collection information (collector, date, location, etc.)
   Nomenclature information (scientific and common)
   Origin (country, state or province, and precise lacality)
   Storage information (location, date placed in storage, etc.)
   Distribution information

2. Environmental information

   Altitute
   Latitute
   Longitude
   Climatic
   Ecological
   Soil description

3. Organismic information

   Morpho-agronomic
   Genetic
   Physiological
   Biochemical
   Environmental damage (cold, drought, wind or other susceptibility)
   Pest and disease information (bacterial, fungal, viral, insect, nematode, etc.)
   Rejuvenation (viability of seeds, time in storage)

4. Use information

   Including breeding, genetics, performance tests, etc.)

5. Food characteristics

6. Industrial characteristics

7. Bibliographic information

What Analysis Should You Use?

DEFINITIONS

NORMAL refers to data that are well approximated by a normal (Gaussian) distribution.

NOT NORMAL refers to quantitative data that are not normally distributed.

CATEGORICAL refers to nominal data, such as male/female or brown/blue/black.

QUANTITATIVE refers to data that are numeric such as height, batting average, number of people per household, etc.

QUALITATIVE refers to data that describe attributes such as hair color, socioeconomic class, sex, etc.

ASSOCIATED refers to variables where knowledge of one helps predict the other.

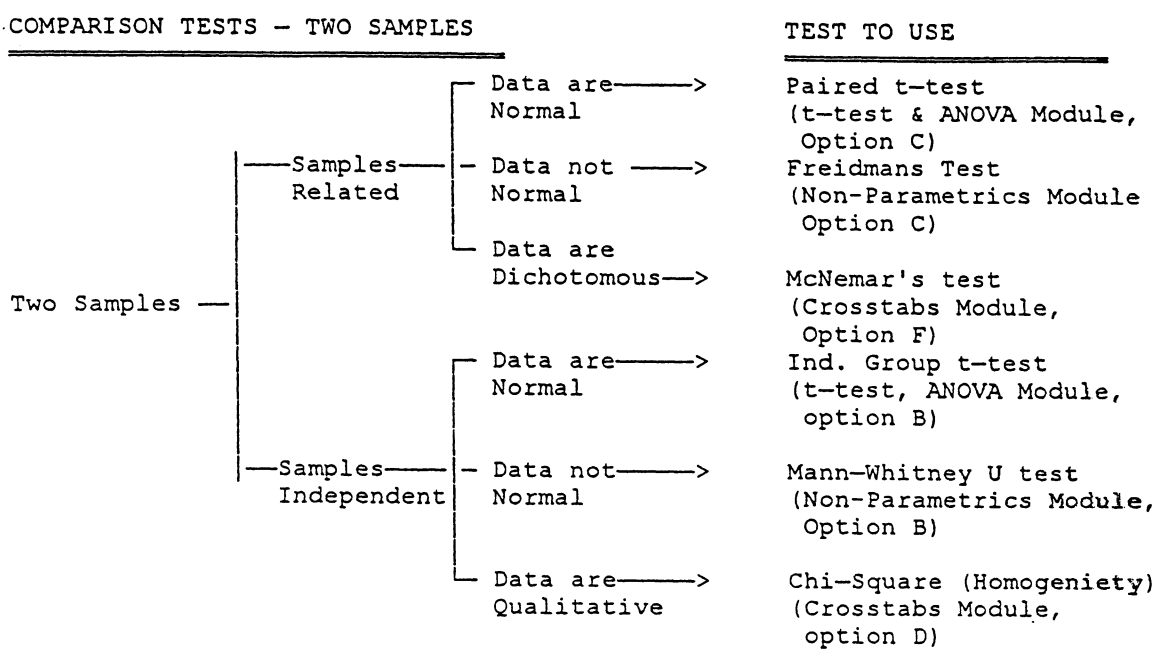
INDEPENDENT refers to variables where knowledge of one does not help predict others. Usually, samples from unrelated populations.

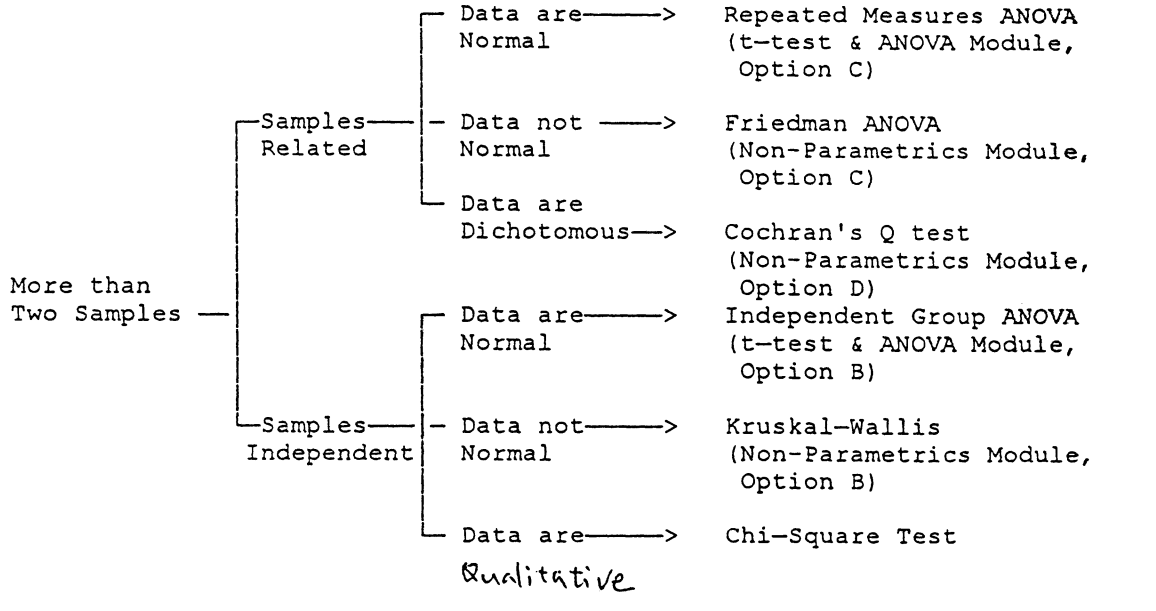
RELATED refers to samples where multiple measures are taken on the same or related entities. For example, before after weights for a diet, or heights of twins.

DICHOTOMOUS refers to data that are categorical and can take on only one of two possible states. For example, yes,/no or on/off. VARIABLE refers to the observed measure, such as height, hair color, etc.

DESCRIPTIVE STATISTICS & GRAPHS		PROCEDURES TO USE
One Sample	Data is Normal ——>	Mean, S.D., Box Plot, 5 number summary Histogram, Conf. Interval (Stat Module, B, C, & E)
	Data not Normal ——>	Median, Box Plot Histogram, 5 number summary (Stat Module, B & E)
	Data is Categorical—>	Frequencies, Pictogram (Crosstabs Module, B)
	Observations-> Over Time	Time Series Plot (Stat Module, option G)

DESCRIPTIVE STATISTICS & GRAPHS		PROCEDURES TO USE
Two Samples (Related)	Data are Normal ——>	Pearson's Corr. Coeff. & X-Y Scatterplot (Stat Module, option F & Regression Module option B & D)
	Data not Normal ——>	Spearmans Corr. Coeff. & X-Y Scatterplot (Stat Module, option F & Regression Module, option D)
	Data are Qualitative ——>	Crosstabulations and 3-D Bar Chart (Crosstabs Module, options D & E)


COMPARISON TESTS – TWO SAMPLES
TEST TO USE
Two Samples
Samples Related
Data are Normal
Paired t-test (t-test & ANOVA Module, Option C)
Data not Normal
Freidmans Test (Non-Parametrics Module Option C)
Data are Dichotomous
McNemar's test (Crosstabs Module, Option F)
Samples Independent
Data are Normal
Ind. Group t-test (t-test, ANOVA Module, option B)
Data not Normal
Mann-Whitney U test (Non-Parametrics Module, Option B)
Data are Qualitative
Chi-Square (Homogeniety) (Crosstabs Module, option D)


COMPARING MORE THAN TWO SAMPLES
TEST TO USE
More than Two Samples
Samples Related
Data are Normal
Repeated Measures ANOVA (t-test & ANOVA Module, Option C)
Data not Normal
Friedman ANOVA (Non-Parametrics Module, Option C)
Data are Dichotomous
Cochran's Q test (Non-Parametrics Module, Option D)
Samples Independent
Data are Normal
Independent Group ANOVA (t-test & ANOVA Module, Option B)
Data not Normal
Kruskal-Wallis (Non-Parametrics Module, Option B)
Data are Qualitative
Chi-Square Test

TESTING ASSOCIATION BETWEEN TWO VARIABLES		PROCEDURE TO USE
Regression	Two Samples Related — Data are Normal →	Pearson Correlation Simple Linear (Regression Module Option B or D)
	Two Samples Related — Data not Normal →	Spearman Correlation (Regression Module, option D)
(Independence)	Two Samples Related — Data are Qualitative →	Chi-Square (Crosstabs Module, Option D)
	Two Samples Related — Data mixed Normal, Not Normal →	Spearman Correlation (Regression Module, option D)

MORE THAN TWO ASSOCIATED VARIABLES		PROCEDURE TO USE
More than 2 Samples Related	Data are Normal →	Multiple Regression (Regression Module, Option C)
	Data not Normal →	Kendall partial rank-correlation (N.A.)
	Data are Qualitative →	Discriminant Analysis (N.A.)

OVERVIEW OF ATTRIBUTES AND VALUES

Attributes	Values
1. Presence of independent and dependent variables	a. One set of variables: no distinction between independent and dependent variables
	b. Two sets of variables: one set of independent and one set of dependent varoables
	c. More than one set of variables, but not b
2. Type of research problem	a. Description (Exploration, Estimation)
	b. Confirmation (Testing)
3. Measurement level	a. Binary
	b. Counts/Frequencies
	c. Nominal
	d. Ordinal
	e. Metric (Interval)
4. Number of variables	a. One
	b. Two
	c. Three
	d. Four or five
	e. Six or more

APPENDIX

COMPLETE EXAMPLE OF MATCHES BETWEEN ANALYSIS PROBLEMS AND ANALYSIS METHODS

This appendix contains the tables with matches between analysis problems and analysis methods that were also included in chapter 6. The matches presented in that chapter were based on the opinions of the 20 respondents interviewed in the study that is fully reported in chapter 3. As a consequence of incomplete information, for several analysis problems no matching analysis method could be found. The tables included in this appendix contain matches between every type of analysis problem and at least one suitable analysis method. The information required to construct these tables, has been gained from handbooks and consultation with statisticians. These completed tables are presented as an example to show, that in principle solutions are possible for every type of analysis problem.

Table A.1 Appropriate methods for the class of problems with two or more symmetrical or three or more asymmetrical sets of variables

Analysis problem	Analysis method
Two symmetrical sets of variables	Canonical correlation analysis, correlation coefficients, loglinear models
Three or more symmetrical sets of variables	Generalized canonical correlation, factor analysis
Nominal independent variables + covariates + dependent variables	Analysis of variance with covariates, loglinear models
Independent variables + intervening and/or latent variables + dependent variables	Structural modeling (LISREL)

Table A.2 Appropriate methods for the class of problems with both dependent and independent variables: one dependent and one independent variable

Attributes				Analysis method
Number of variables		Measurement level		
Dep.var.	Indep.var.	Dep.var.	Indep.var.	
One	One	Non-metric	Non-metric	Nonparametric tests, χ^2, logistic analysis of variance, loglinear analysis
One	One	Non-metric	Binary	Mann-Whitney test, χ^2
One	One	Non-metric	Metric	Logistic regression analysis
One	One	Binary	Non-metric	χ^2, Cramér's coefficient C
One	One	Binary	Metric	χ^2, logistic regression
One	One	Counts	Non-metric	χ^2
One	One	Counts	Metric	Regression
One	One	Nominal	Non-metric	Cramér's coefficient C
One	One	Nominal	Metric	(Logistic) regression, loglikelihood
One	One	Ordinal	Binary	Wilcoxon's two sample test, χ^2, Mann-Whitney test, Kolmogorov-Smirnov test
One	One	Ordinal	Counts	Nonparametric correlation
One	One	Ordinal	Nominal	Kruskal-Wallis test, analysis of variance
One	One	Ordinal	Ordinal	Nonparametric correlation, Spearman's rho, Kendall's tau
One	One	Ordinal	Metric	Nonparametric correlation, isotonic regression
One	One	Metric	Binary	T-test, correlation
One	One	Metric	Counts	Regression
One	One	Metric	Nominal	Analysis of variance
One	One	Metric	Ordinal	Nonparametric correlation
One	One	Metric	Metric	Regression analysis, correlation

Table A.3 Appropriate methods for the class of problems with both dependent and independent variables: more than one dependent and/or more than one independent variable

Attributes				Analysis method
Number of variables		Measurement level		
Dep.var(s)	Indep.var(s)	Dep.var(s)	Indep.var(s)	
One	More	Non-metric	Non-metric	Logistic analysis of variance, loglinear analysis, canonical correlation analysis
One	More	Non-metric	Metric	Logistic regression analysis
One	More	Nominal	Metric	Discriminant analysis
One	More	Metric	Non-metric	Analysis of variance
One	More	Metric	Metric	Multiple regression analysis
More	One	Non-metric	Non-metric	One-way multivariate analysis of variance with dummy variables
More	One	Non-metric	Metric	Multivariate regression with dummy variables
More	One	Metric	Non-metric	Multivariate analysis of variance
More	One	Metric	Metric	Multivariate regression analysis
More	More	Non-metric	Non-metric	Multivariate analysis of variance with dummy variables
More	More	Non-metric	Metric	Multivariate multiple regression with dummy variables
More	More	Metric	Non-metric	Multivariate analysis of variance
More	More	Metric	Metric	Multivariate multiple regression analysis, canonical correlation analysis, redundancy analysis

Table A.4 Appropriate methods for the class of problems without distinction between dependent and independent variables

Number of variables	Measurement level	Analysis methods for	
		Description	Confirmation
One	Binary	Frequencies, proportions, counts	Nonparametric tests, binomial test
	Counts	Frequencies, mean, median, mode	Nonparametric tests, t-test
	Nominal	Frequencies, proportions, mode	Nonparametric tests, χ^2
	Ordinal	Median, mode	Nonparametric tests, binomial test
	Metric	Mean, median, mode, variance, kurtosis	Confidence intervals, t-test
Two	Binary	Tetrachoric correlation, crosstabulations	Nonparametric tests
	Counts	Crosstabulations	Nonparametric tests, χ^2
	Nominal	Contingency tables, correspondence analysis	Loglinear analysis, χ^2
	Ordinal	Kendall's tau, Spearman's rho	Nonparametric tests
	Metric	Correlation coefficient, scatter plot	Correlation coefficient
More than two	Binary	Rasch models, Guttman scaling	Loglinear analysis
	Counts	Crosstabulations	Factor analysis
	Ordinal	Nonlinear principal components analysis, multidimensional scaling	Factor analysis
	Metric	Principal components analysis, one-dimensional scaling, cluster analysis	Factor analysis, F-test
Three	Nominal	Latent class analysis, contingency tables, correspondence analysis, homogeneity analysis	Loglinear analysis
Four or five	Nominal	Latent class analysis, homogeneity analysis	Loglinear analysis
Six or more	Nominal	Homogeneity analysis	Loglinear analysis

Level of measurement	One-sample case (Chap. 4)	Two-sample case		k-sample case		Measures of association (Chap. 9)
		Related or matched samples (Chap. 5)	Independent samples (Chap. 6)	Related samples (Chap. 7)	Independent samples (Chap. 8)	
Nominal or categorical	Binomial test (4.1) Chi-square goodness-of-fit test (4.2)	McNemar change test (5.1)	Fisher exact test for 2 × 2 tables (6.1) Chi-square test for $r \times 2$ tables (6.2)	Cochran Q test (7.1)	Chi-square test for $r \times k$ tables (8.1)	Cramer coefficient, C (9.1) Phi coefficient, r_ϕ (9.2) The kappa coefficient of agreement, K (9.8) Asymmetrical association, the lambda statistic, L_B (9.10)
Ordinal or ordered	Kolmogorov-Smirnov one-sample test, $D_{m,n}$ (4.3) One-sample runs test (4.4) Change-point test (4.5)	Sign test (5.2) Wilcoxon signed ranks test, T^+ (5.3)	Median test (6.3) Wilcoxon-Mann-Whitn test, W_s (6.4) Robust rank-order test, U(6.5) Komogorov-Smirnov two-sample test, D_m (6.6) Siegel-Tukey test for scale differences (6.8	Friedman two-way analysis of variance by ranks, F_r (7.2) Page test for ordered alternatives, L (7.3)	Extension of the median test (8.2) Kruskal-Wallis one-way analysis of variance, KW (8.3) Jonckheere test for ordered alternatives J (8.4)	Spearman rank-order correlation coefficient, r_s (9.3) Kendall rank-order correlation coefficient, T (9.4) Kendall partial rank-order correlation coefficient, $T_{xy.z}$ (9.5) Kendall coefficient of concordance, W (9.6) Kendall coefficient of agreement, u (9.7) Correlation between k judges and a criterion, T_C (9.7.4)
Interval		Permutation test for paired replicates (5.4)	Permutation test for two independent samples (6.7) Moses rank-like test fo scale differences (6.9			Gamma statistic, G (9.9) Somer's index of asymmetric association, d_{BA} (9.11)

Note: Each column lists, cumulatively downward, the tests applicable for the given level of measurement. For example, in the case of k related samples, when the variables are ordered, both the Friedman two-way analysis of variance and the Cochran Q test are applicable. However, see text for a discussion of appropriateness of a particular test to a given type of data. The numbers in parentheses refer to chapter sections.

KEY TO MULTIVARIATE ANALYSES

1. The objects were sampled from more than one population	-> 2
The objects were sampled from a single population	-> 3
2. The main purpose of the analysis is to determine if the samples could have been drawn from a single statistical population; i.e., are the mean vectors of the populations equal?	-> multivariate analysis of variance
The main purpose of the analysis is to find linear combinations of the variables that maximize differences among preexisting populations; i.e., one wishes to sort the objects into their appropriate populations with minimal error	-> discriminant analysis
The main purpose of the analysis is to sort a previously unpartitioned heterogeneous collection of objects into a series of sets; i.e., one wishes to identify sets and allocate objects to these sets simultaneously	-> cluster analysis
The main purpose of the analysis is to arranges the objects graphically in few dimensions, while retaining maximal fidelity to the original interobject relationships	-> nonmetric scaling
3. The variables can be logically divided into two (or more) sets and one wishes to establish maximal linear relationships among these sets	-> multiple regression and correlation, canonical correlation
The variables logically belong to a homogeneous set	->4
4. The main purpose of the analysis is to describe parsimoniously the total variance in a sample in a few dimensions; i.e., one wishes to reduce the dimensionality of the original data while minimizing loss of information. These few dimensions are the linear combinations of the original variables that successively account for the major independent patterns of variation in the sample	-> principle components analysis
The main purpose of the analysis is to resolve the intercorrelations among variables into their putative underlying causes; i.e., one wishes to reproduce only the intercorrelations among variables rather than their total variances	-> factor analysis