Morphometric data analysis

Types of data (variables)

1. Measurement variables
   - continuous (e.g. measurements)
   - discontinuous (discrete) (e.g. counts)

   Ranked variables

   Attributes (e.g. short, long)

2. Raw data

   Derived data: ratios, percentages, indices, transformations, etc.

Accuracy

4 is not 4.0 is not 4.00
(e.g. the number of teeth, 4 is the exact number; but for a measurement 4 is not the same as 4.00)

Descriptive statistics

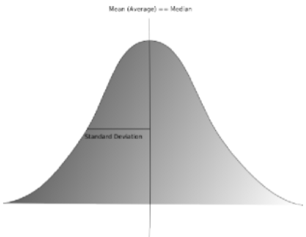
measurements : mean, standard deviation, range, etc.
meristics : median, mode, quartiles, range, etc.

Standard deviation: $\surd \sum(x\text{-mean})^2 / n\text{-}1$

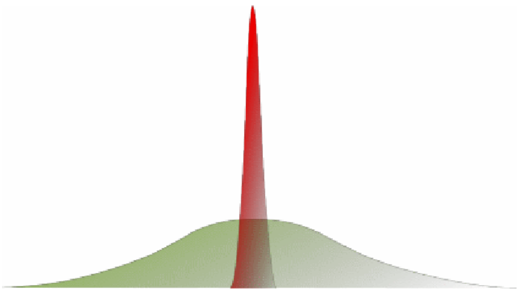
Normal distribution and its consequences

   95 % of data fall within the range of the mean + or – two SD

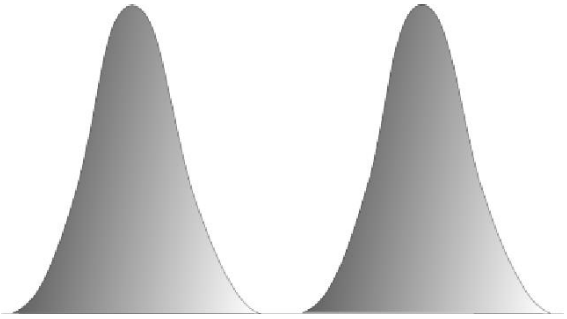
   departures : skewness and kurtosis



In principle, many statistical analyses may only be used for data with a normal distribution !!!



same mean different SD



same SD, different mean

Anova : assumptions

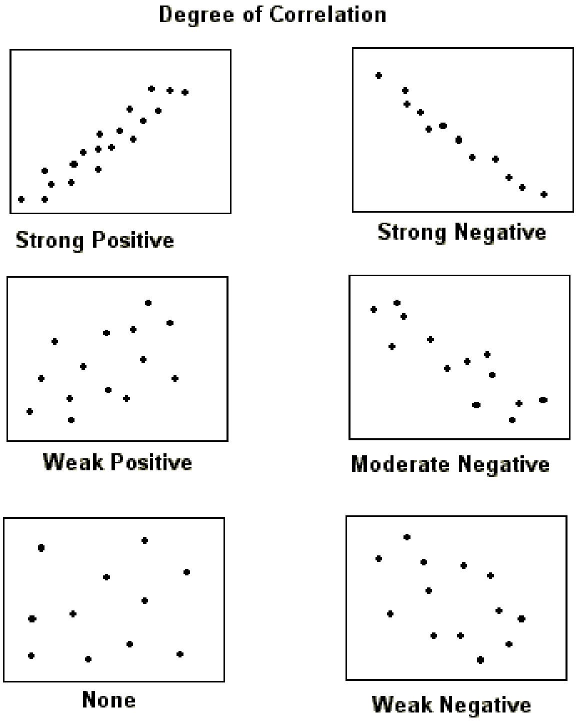
random sampling (choice of specimens)
independence = random error (measuring)
normal distribution
homogeneity of variances (homoscedasticity)
Manova = similar but in a multidimensional space

Correlation and regression

Two variates are related

correlation: the intensity of the relationship

regression: expresses the form of the relationship (predictive)



Allometry

$Y=bX^{\alpha}$
$\log Y = \log b + \alpha \log X$ ($\log b$ = intercept; α =slope)

Non-parametric tests (univariate)

Kruskal-Wallis, Mann-Whitney U, Wilcoxon, Kolmogorov-Smirnov, Wald-Wolfowitz, etc.
Used as a substitute for parametric tests (t-test, F-test etc, anova). They are also called distribution-free tests. Data do not have to be normally distributed. Can even be used on ranked data.

Multivariate data analysis

One dimension, Two dimensions, Three dimensions, Higher dimensions (n)

- No a priori distinction of subgroups: Principal Component Analysis (PCA), Cluster Analysis (CA)
    - PCA maximise variability and form linear combinations; PC components are linear combinations of observed variables; method for data reduction
  - CA groups cases (clusters); it calculates distances between individuals and/or clusters in various ways (options)

- A priori distinction of subgroups: Discriminant Analysis (DA), Canonical Variates Analysis (CVA), but statistical assumptions may be problematic.
  - CVA (=multiple DA): objective is to calculate maximal distances between two or more a priori defined groups of individuals; it maximises the ratio of between-group (among) to within-group variance.
  - DA involves two groups only and can also be used for placing an individual in one of the groups after the initial analysis.

- In both categories, other analyses exist. For many the matrix algebra is similar, but they have limited use in taxonomy at the moment

There is a considerable robustness in these analyses to violations of the assumptions (see above). But quite importantly, PCA is essentially assumption free if tests of significance are not performed. So a perfect tool in taxonomy for exploring the data set.

Principal Component Analysis

The goal of PCA is to summarise a multivariate dataset as accurately as possible using a only a few components (principal axes)

Steps:
1. To find linear composites (axes) of maximal variance within a point swarm (e.g. specimens with their respective values of various measurements).
2. To find out which variables have the largest contributions (loadings) to these axes
3. To present graphically the new location of the points (e.g. specimens in the case of taxonomy) on these axes.

PCA axes are in the direction of the greatest overall variance among individuals. Groups will only be apparent if the distances among groups are making a large contribution to this overall variance.
The axes are orthogonal (perpendicular) to one another.
The first axis is defined so as to go through the longest part of the point swarm (= the greatest amount of linear variation). The second axis accounts for the maximum amount of residual linear variation; PC3, PC4, etc. (see illustrations in ppt).
The idea is to obtain derived variables of the original variables (e.g. measurements or meristics in taxonomy) that express a large proportion of the total variance of the data with a smaller number of variables. Obviously one loses information, but most probably this information is of no importance to our research question.

Separate analysis of meristics and measurements is recommended for fish taxonomy
meristics : raw data, correlation matrix
measurements : log^{10} transformation by preference, or percentages (in exceptional cases may produces better results), covariance by preference, or correlation matrix

correlation matrix : variables standardised to zero mean and unit SD; must be used if variables are heterogenic (e.g. various types of counts).

covariance matrix : variables only standardised to zero mean; better used if variables are expressed in the same measurement unit

If loadings of variables on PCI are of the same magnitude and the same sign (clearly the case for log-transformed measurements), than PC1 = size vector, but subsequent axes (PC2, PC3, ...) can still contain allometric size.

Short overview on various multivariate data analysis (not complete)

(Be aware that there is some discussion about the details of these methods; terminology varies considerably among textbooks)
(Ordination = process of producing a small number of variables than the original ones, that can be used to describe the relationship between a group of objects (cases)

Principal Component Analysis (PCA) : most commonly used ordination method; by far the best method for exploratory, descriptive multivariate data analysis without statistical inference (in all other methods, one should, in principal, first test the assumptions); used to reduce a large number of variables to a few principal components (axes); results can be used to visualise the location of the cases (e.g. specimens examined) in a two dimensional space. No a priori groups assigned before starting analysis (See above for further details). In principal, no extra rotation of axes is done.
Remark: PCA was originally designed to explore variance within one homogenous sample but is currently used to explore pooled groups.

Factor Analysis (FA) : mostly, PCA is regarded is a special case of FA. Several methods of rotation of the calculated axes can be done, e.g. Varimax. Not used in taxonomy anymore

Discriminant Analysis (DA) (= Discriminant Function Analysis): multi group (ordination) method. On two samples often referred to as a Linear DA; on more than two samples referred to as a Canonical Analysis or a Canonical Variates Analysis (CVA) (be aware that in some statistical programmes, a Canonical Analysis is in fact a Canonical Correlation Analysis and not some kind of DA; this can be confusing). DA is basically quite similar to PCA but starts from a priori groups defined before the analysis starts (see also above). In principle no need to standardise data (covariance or correlation matrix) as is done in PCA. Various options possible depending on software used. Possibility to assign an individual (not in the analysis) to one of the groups after the analysis.

Cluster Analysis (CA): not an ordination method; no a priori groupings made; very large scale of possible options. Two main methods: (1) hierarchical techniques producing dendrograms (trees) and (2) partitioning techniques (K-means in Statistica or Past) allocating the cases to a predetermined number of groups.

Canonical Correlation Analysis (CCA): used for finding relationships, correlations, between two sets of variables (not of cases like in DA), e.g. a set of environmental characters and a set of gene frequencies for a group of cases. Ordination method. Called Canonical Analysis is some software.

Multidimensional scaling (MDS): ordination method; does not analyse the data but some measure of distance between a number of cases; constructs a kind of ‘map’ for these distances. Sometimes used as an alternative to CA. There is metric and non-metric MDS. Non-metric MDS compares the rank order of the new calculated distances with the original distances through a measure called ‘stress’.

Principal Coordinates Analysis (PCOA) : ordination method; also used not on measurements but on some measure of distance; similar to MDS and producing similar results, but other kinds of assumptions and calculations (PCA type of approach with eigenvalues). Sometimes regarded as a special case of MDS.

Correspondence Analysis (COR) (= reciprocal averaging) : ordination method often used in ecology on data of abundance (frequencies); especially useful for plotting of variables and cases in one plot. Can be used for presence/absence data.

Other kinds of tests

Often used is Mantel’s test to statistically compare two distance or similarity matrices

PCA : L'analyse en composantes principales (ACP)
FA : L'analyse factorielle
CoA : L'analyse des correspondances
CA : Méthodes de classification hierarchique
MDS : Le positionnement multidimensionnel
DA : L'analyse discriminante
CCA : L'analyse canonique des corrélations
CVA : L’analyse de variance canonique