SIROMATH

consultants in mathematics, statistics, planning, computing and operations research.

SIROMATH PTY. LTD.
(INCORPORATED IN N.S.W.)

8/154 HAMPDEN ROAD, NEDLANDS W.A. 6009
☐ *TEL.:* (09) 386 8211 ☐ *FAX:* (09) 386 8559

Our Reference:

Your Reference:

MULTIVARIATE STATISTICS

A Short Course

Dr Geoff Riley

C.A.L.M.
April 27 & 28, 1988.

Other Offices at:
☐ LEVEL 5, 156 PACIFIC HIGHWAY, ST. LEONARDS, N.S.W. 2065. (02) 436 0500
☐ 1 LIARDET STREET, PORT MELBOURNE 3207. (03) 699 9777
☐ McMILLANS ROAD, BERRIMAH N.T. 5789. (089) 84 5315

1. OVERVIEW

Many fields of research involve collection and interpretation of a number of variables, all of which are recorded for each of a collection of objects.

e.g.
- concentrations of large numbers of chemical elements or compounds (the variates) for each of a number of soil samples (the objects);
- measurements of a variety of physical dimensions (the variates) on animals (the objects) sampled from a population (e.g. dolphins of a particular species off northern Australia);
- species abundance scores for a selection of sites covering a gradation of habitats.

Multivariate techniques attempt to be more informative than would be possible with one variable at a time analyses. This usually involves regarding the measurements recorded for an object as constituting a "point" in some multivariate population (or a choice of such populations), and asking questions about such things as the distribution of such points, or matrices of pairwise distances between them.

Some categories of multivariate technique are:

(a) Generalisations of Univariate Ideas
- modelling distributions of points
- constructing informative summary statistics
- fitting explanatory models and conducting tests of null hypotheses (e.g. that two species differ in their distributions of a set of measurements)
- constructing confidence intervals for population parameters

(b) Assigning new individuals to predefined groups (discriminant analysis)
- need "training" data with grouping known
- generally involves multivariate normal model for each groups population distribution

(c) Generating Interesting Combinations of Variables
- usually "linear" combinations
- different combinations for different purposes (e.g. principal components,canonical variates)

(d) Clustering
- no predefined groups
- generally no modelling involved
- many, many variations of technique

(e) Ordination
- finding low dimensional configurations of points whose pairwise distances reflect the "distances" between the original data points

- once again, many variations of technique (e.g. many distance measures)

Course Plan:

2. Univariate Data Summaries and Presentations - A Review
3. Probability Distributions
4. Choosing a Data Transform - with Two Case Studies
5. Multivariate Data - Summary & Presentation
6. The Multivariate Normal Model
7. Changes of Coordinates; Principal Components
8. Sampling Distributions
9. Hypothesis Testing
10. Multivariate Analysis of Variance
11. Group Discrimination; Canonical Variates
12. Allocation and Atypicality
13. Clustering Techniques
14. Ordination Techniques

2. UNIVARIATE DATA SUMMARIES & PRESENTATIONS - A REVIEW

A. Statistics to quantify important properties of collections of numbers.

Suppose Y1,Y2,... Yn is a sampled collection of n values (e.g. n=191 values of SiO2 percent - see copy of SAS output on later page).

Important features of a data collection are:

Location - where it is centred

mean $\bar{Y}$ = (Y1+Y2+... Yn)/n

median = middle value when values put in ascending order (mean of middle two values when n is even).

The median is less affected by "skewness" or outlying values:

Scatter or Variability

(i) Sample Variance

$$s^2 = \{(Y1-\bar{Y})^2 + (Y2-\bar{Y})^2 + \ldots \quad (Yn-\bar{Y})^2 \} / (n-1)$$

(mean square deviation from middle)

(ii) Standard Deviation

s = square root of sample variance

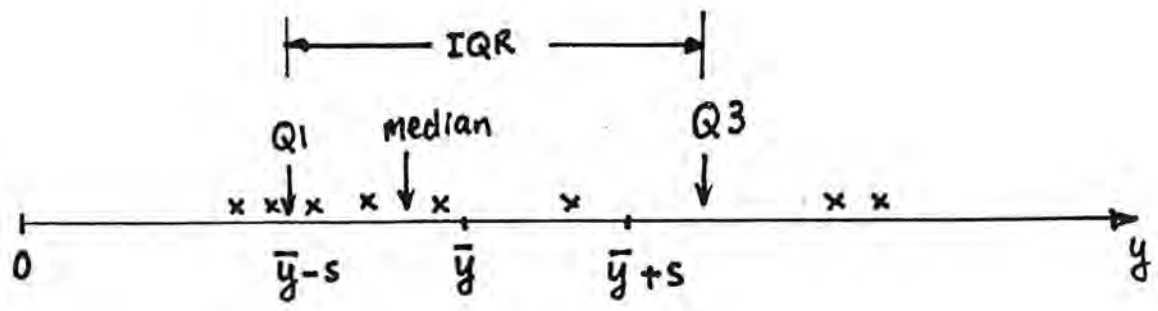
(iii) Range

= max value - min value

(iv) Interquartile Range ("Q3-Q1")

= upper quartile - lower quartile,

where the quartiles are defined so that 25% of data values fall above and below them.



Points to Note

(a) The mean and standard deviation are examples of "moment" statistics - based on averages of various powers of the data values.

(b) The median and interquartile range are examples of summaries based on "quantiles" or "percentiles".

e.g. SAS PROC UNIVARIATE also outputs the 99%,95%,90%,10%,5% and 1% quantiles - the values those percentages of the way through the ordered data.

(c) Quantile based statistics offer "robustness" - less sensitivity to atypical or wrong data (but not so versatile for statistical inference).

(d) If all data values are "relocated" (e.g. replacing Y1,...Y191 by

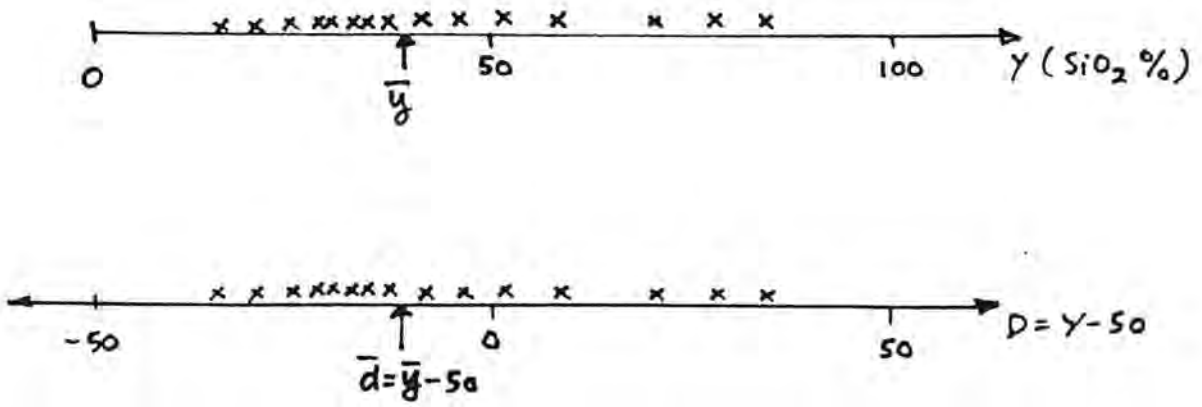
$$D1=Y1-50,$$
$$D2=Y2-50,$$
$$\vdots$$
$$D191=Y191-50),$$

then:

(i) The mean and median are related in the same way (so for example $\bar{D} = \bar{Y} - 50$).

(ii) The standard deviation and IQ range are unaffected.

Picture:



(e) If all data values are rescaled (e.g. replacing Y1,... Y191 (MgO percentages) by

$$X1=10000*Y1,$$
$$X2=10000*Y2,$$
$$\vdots$$
$$X191=10000*Y191$$

(MgO parts per million)),

then:

(i) the mean and median
(ii) the standard deviation and IQ range

are all transformed in the same way:

$$\bar{x} = 10\,000\,\bar{y} \quad , \quad s_x = 10\,000\,s_y$$

(f) Z1, Z2, ... Zn defined by

$$Z1 = (Y1-\bar{Y})/s$$
$$Z2 = (Y2-\bar{Y})/s$$
$$\vdots$$
$$Zn = (Yn-\bar{Y})/s$$

are "<u>standardised</u>" in that

$$\bar{Z} = 0$$
$$\& \quad s = 1.$$

Standardisation is an important concept:

- to enable description of many practical data distributions using a few standard models (particularly the standard normal).

- in some multivariate contexts (where it can be useful to relate each variable back to a standard scale and location).

Skewness

The classical measure (based on moments):

$$= [\text{mean of } (Yi-\bar{Y})^3 / s^3] \text{ times } [\ n/(n-1).(n-2)\ ]$$

Various forms of "interquartile skewness":

e.g. $(\ Q3-\text{median}\ )/(\text{median} - Q1) - 1$

Notes:

(i) Cases:

+ve skewness:
- most bunched at left
- mean > median
- Q3-median > median-Q1

-ve skewness:
- most bunched at right
- mean < median
- Q3-median < median-Q1

(ii) Skewness isn't affected by relocation or rescaling, but can be affected by use of other transformations - see later.

(iii) Pictures are better for judging skewness in data distributions.

In comparison with means and standard deviations, the actual numerical values of skewness measures are used little in model fitting or inference.

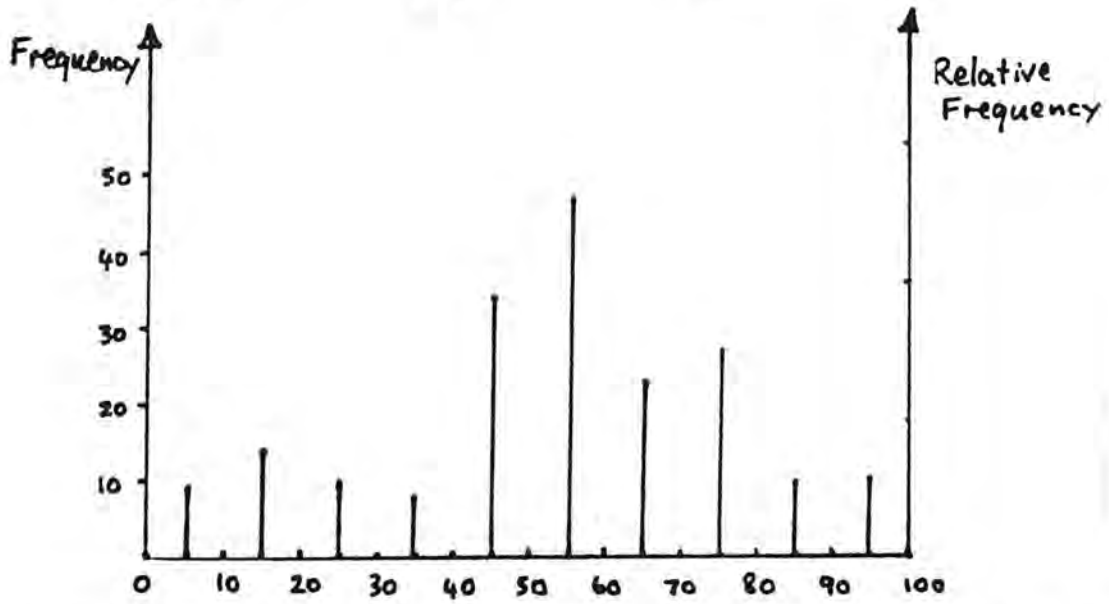
The direction (sign) of skewness is of importance in determining subsequent treatment.

B. Graphical Presentations

(i) Histograms or Bar Charts

e.g. for SiO2 percentages:

Class Interval	Fregency	Relative Freq.(%)
0.0- 9.9	9	4.7
10.0-19.9	14	7.3
20.0-29.9	9	4.7
30.0-39.9	8	4.2
40.0-49.9	34	17.8
50.0-59.9	47	24.6
60.0-69.9	23	12.0
70.0-79.9	27	14.1
80.0-89.9	10	5.2
90.0-99.9	10	5.2
total	191	100



When data sets aren't too large "stem and leaf" plots are a useful form of sideways bar chart.

Recipe:

- Create a "stem" listing the most significant digits of the data values vertically from high down to low.

- Create "leaves" by listing the next significant digit of each data value next to the stem position corresponding to the appropriate leading digits. Put the leaf values in ascending order.

e.g. The positioning of the highest value 99 for SiO2 (see over).

stem posn — leaf value

Stem and leaf plots allow:

- graphical appreciation of the shape of a data distribution (as with histograms);
- reading off of numerical values.

STEM/LEAF	#
9 556667789	9
9 3	1
8 59	2
8 00122224	8
7 559	3
7 00111112222233333344444444	25
6 55555667888	11
6 0011222334444	13
5 5555666777888886899	19
5 000111111122222223333334444	27
4 55556677888889999999	20
4 0011222333444	13
3 56789	5
3 034	3
2 66889	5
2 1222	4
1 556699	6
1 00111223 4	9
0 57899	5
0 133	3
----+----+----+----+----+--	
MULTIPLY STEM.LEAF BY 10**+01	

Another histogram

rounded value = 14

VARIABLE=SIO2PCT

MOMENTS

N	191	SUM WGTS	191
MEAN	52.9764	SUM	10118.5
STD DEV	22.4416	VARIANCE	503.627
SKEWNESS	-0.334216	KURTOSIS	-0.183267

QUANTILES(DEF=4)

100% MAX	98.8	99%	96.432
75% Q3	69.8	95%	93.68
50% MED	53.5	90%	80.42
25% Q1	43.1	10%	15.14
0% MIN	1.1	5%	9.96
		1%	2.648
RANGE	97.7		
Q3-Q1	26.7		
MODE	48.6		

(ii) Box Plots:

- convenient, quick pictorial representation of the main features of data spread

- show quartiles, median (the "box")

- show range of data (the "whiskers" extend out to the maximum & minimum value, provided they are within 1.5 interquartile ranges of the upper or lower quartile)

- outlying values are highlighted (plotted separately beyond the whiskers - which, in SAS at least, are limited to a length of 1.5*IQR when outliers exist)

- are good for showing skewness

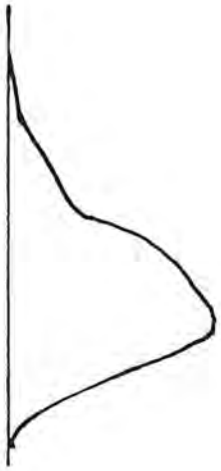
- are good for univariate comparisons between groups.

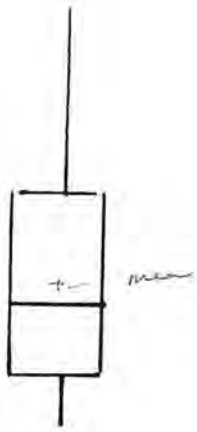
Some Cases:

stem&leaf

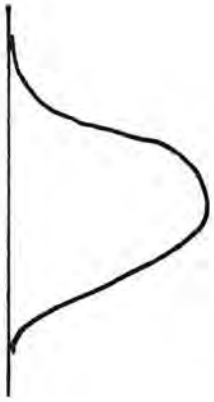
boxplot

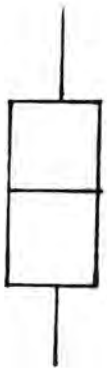
+ve
skewness



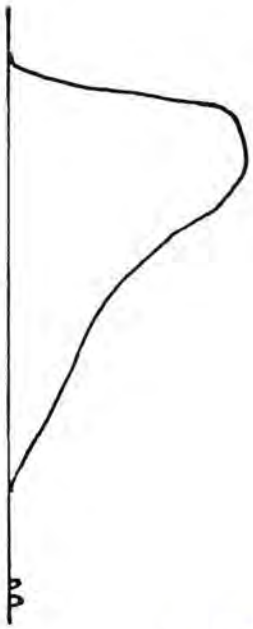


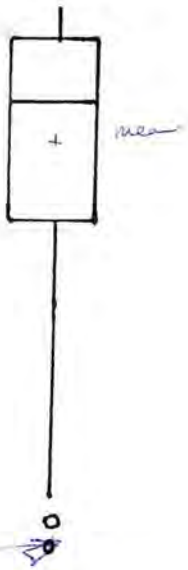
no
skewness





-ve
skewness

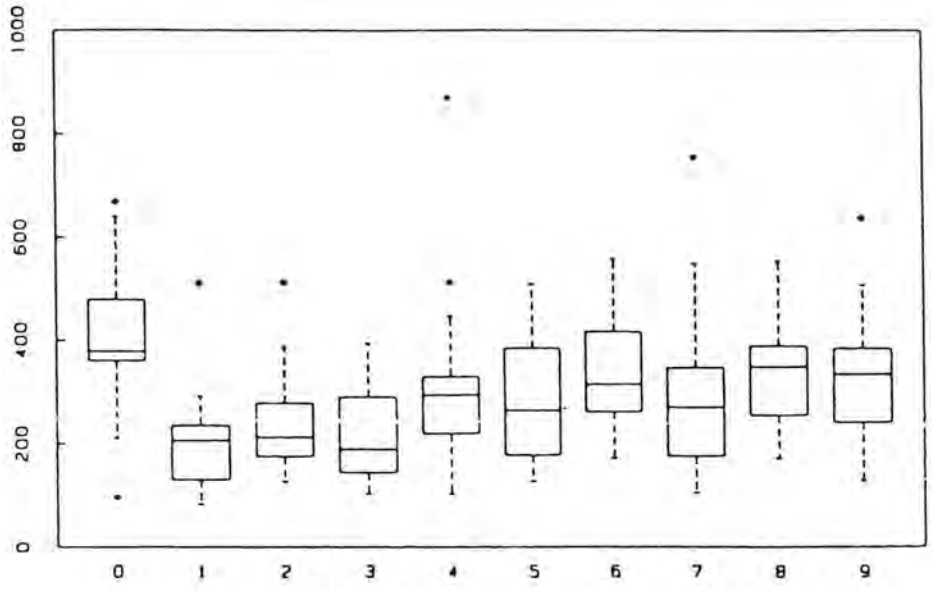




if Max, Min exceeds
1.5 inter-quartile range

Example Multiple Group Box Plots:

(Done by "S" - available for Unix & VMS systems)



(iii) Normal Probability Plots

- very useful when want to work out appropriate transforms
- straight when the data has a normal distribution
- more on this later.

straight line - data has normal distribution

3. PROBABILITY DISTRIBUTIONS

- basic to the mathematical description (modelling) of the spread of possibilities (variability) in a "population"

- viewed via samples

- for infinite populations, they describe the limiting values of proportions within samples as sample size gets very large (assuming samples are representative).

Example 1

Population = a herd of 100 cows
(finite)

Probability that a cow chosen at random has yield between a and b

= the proportion out of 100 with yields in that interval

e.g. Prob(12<Y<22) = 75/100 (75%) when a histogram of the values for the finite population is as follows:

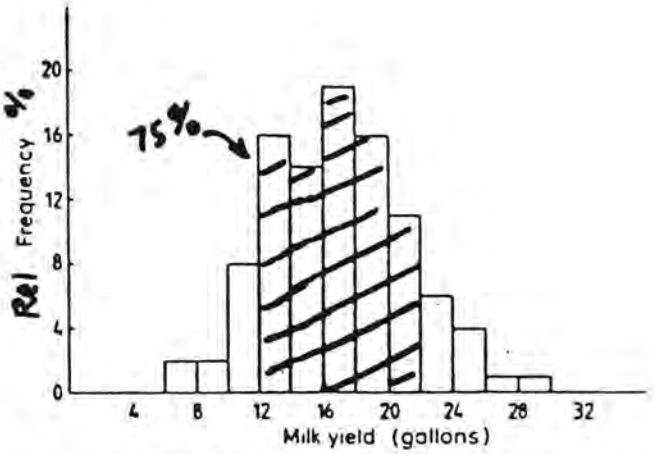
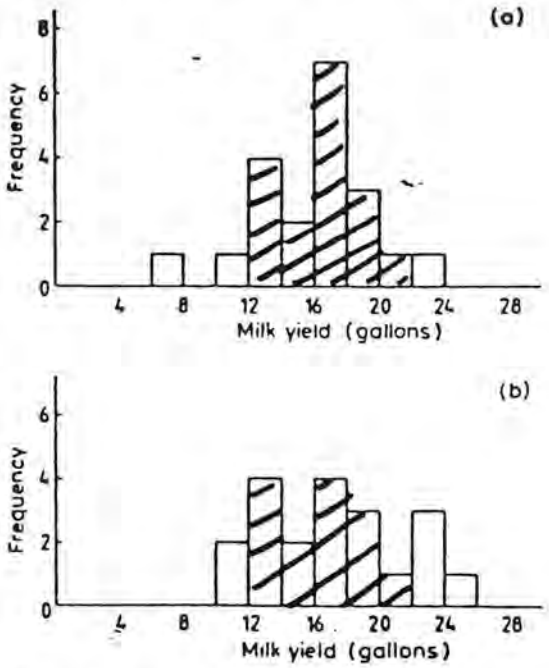


Figure 2.1 Histogram of the data in Table 2.1. Grouping the yields into intervals of width 2 gallons per block.

Samples of 20 from the population produce approximations to the population probability distribution.

Proportions between any pair of limits can be calculated for the samples, and compared to the corresponding population proportion (probability). e.g. the sample proportions between 12 and 22 for two samples of 20 are:

17/20 (85%) for sample (a)
14/20 (70%) for sample (b).



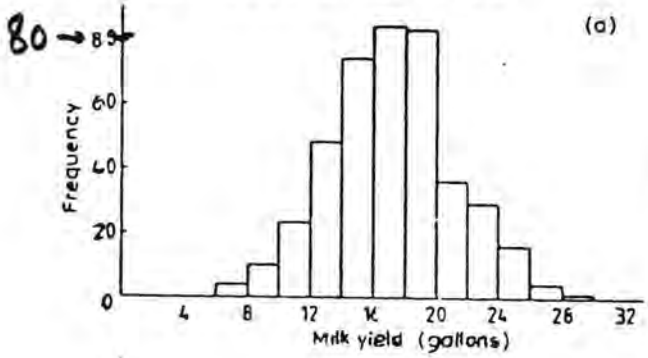
Distributions for two samples of twenty from the data of Table 2.1.

Example 2

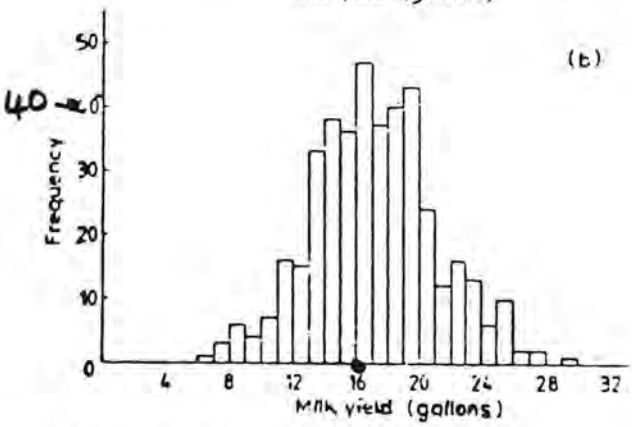
Population = all milking cows in a region (say 100,000+). The population is effectively infinite. Consider taking increasingly large samples (presumed representative) from it:

e.g. sample of 400:

(a) interval width = 2 gallons



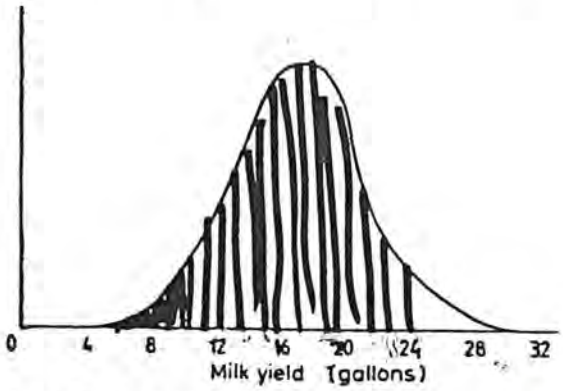
(b) interval width = 1 gallon
(same sample)



N.B. Halving interval widths halves frequencies (on average) for similar midpoints.

Limiting Case:

Very large sample with very small interval width:



Population distribution for milk yield data.

f = probability density function

Generally find:

relative frequency $\approx$ f(interval midpt).(int. width),

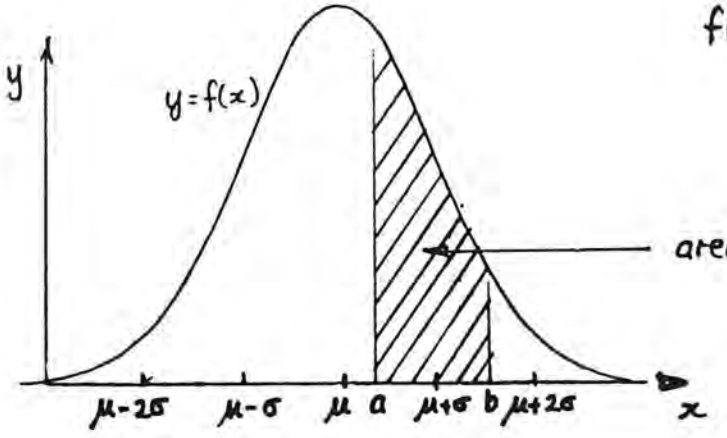
with a better and better approximation with bigger and bigger samples, where:

the function f(x) , called the probability density function characterises the underlying population probability distribution:

Prob(a < X < b) = area under the curve with equation y=f(x) between limits a and b.

e.g. For a fitted "normal" f(x) for milk yields, the area under the curve between a=12 and b=22 is 0.73 (73%). This is an estimate of the proportion of milk yields in the region between these limits.

Normal Distribution Formula:



location

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

spread

area = Prob (a < X < b)

μ = mid point mean

σ = sds

Can standardize any normal distr by subtracting the mean

The Standard Normal Distribution

If X is chosen randomly from a normally distributed population (i.e. is a normal "random variable") then

$$Z = (X - \mu)/\sigma$$

is also normal and has $\mu = 0$, $\sigma = 1$ (the standard normal distribution:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}).$$

X

$X - \mu$
(location parameter = 0)

$(X-\mu)/\sigma$
(scale parameter = 1)

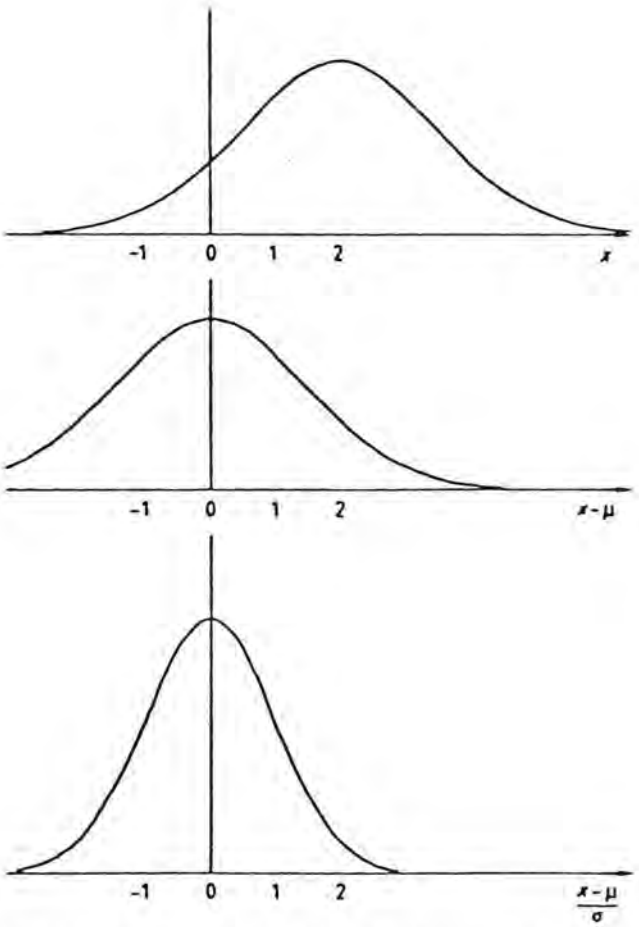


Figure 2.7 Distribution of x, $x - \mu$, $(x - \mu)/\sigma$ where x is normally distributed with mean μ and variance σ^2.

Reasons for Importance of the Normal Model for Population Distributions

- much of statistical theory (e.g. t and F tests) is based on normality assumptions;

- using standard methods with non-normal data can produce misleading results;

- the normal distribution turns out to be at least approximately appropriate to lots of populations of practical interest;

- even when it's not, transformations can often make it so;

- the multivariate normal distribution (with the necessity that individual variables be normally distributed) is essentially the only model distribution for which a decent range of model specific statistical methods (e.g. multivariate analysis of variance) have been worked out;

- multivariate methods not necessarily assuming normality (e.g. clustering, principal coordinate analysis, canonical variate analysis) often work best in normal situations anyway.

Population Means and Variances

Define:

E(X) = the average value of X over the population from which it is chosen

(= the value approximated by sample means from large representative samples).

N.B. The E stands for "expected value".

For the normal, $E(X) = \mu$ (the location parameter).

Thus the standard normal has zero population mean. μ is used very generally to denote population means.

Var(X) = the average value of $(X-E(X))^2$ over the population from which X is chosen.

(approximated by variances of large representative samples)

For X normal, $Var(X) = \sigma^2$.

Thus σ, the scale (of scatter) parameter is simply the standard deviation of the population.

Important Characteristics of the Normal

mean = median
symmetric, zero skewness
68.3% of population within 1 standard deviation of mean
95.4% of population within 2 standard deviations of mean
99.7% of population within 3 standard deviations of mean

Normal Probabilities and Percentiles in General

$F(z) = \text{Prob}(Z \leq z)$

is available in tables and statistical packages (e.g. PROBNORM(Z) in SAS).

(N.B. When using tables need to note:

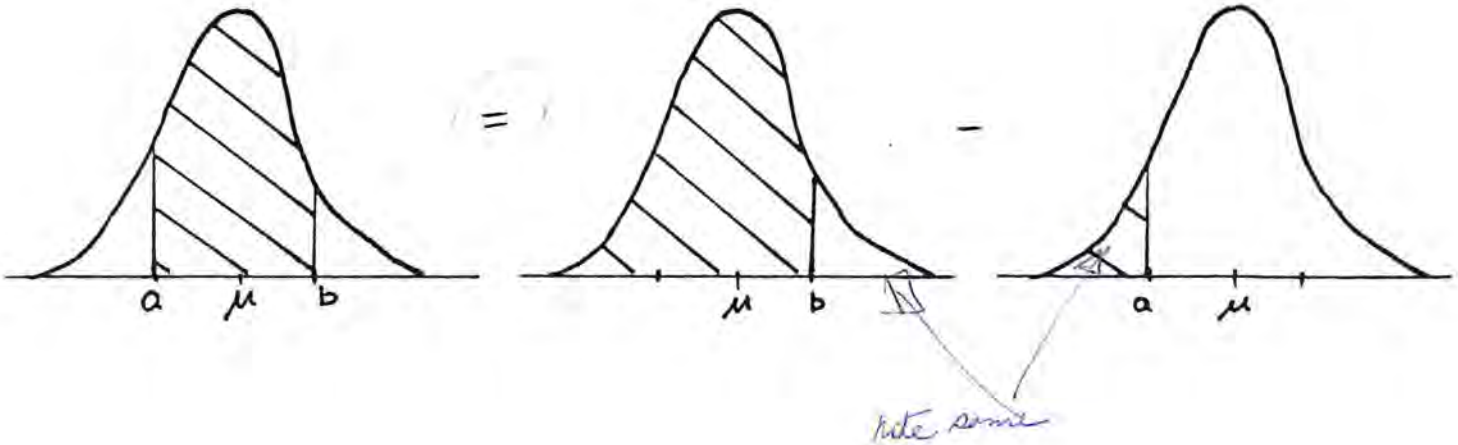
PROBNORM(-Z)=1-PROBNORM(Z).)

To calculate P=Prob(a<X<b) in general, where X is normal, (say with μ=30, σ=5), use:

$$P = \text{PROBNORM}((b-\mu)/\sigma) - \text{PROBNORM}((a-\mu)/\sigma).$$

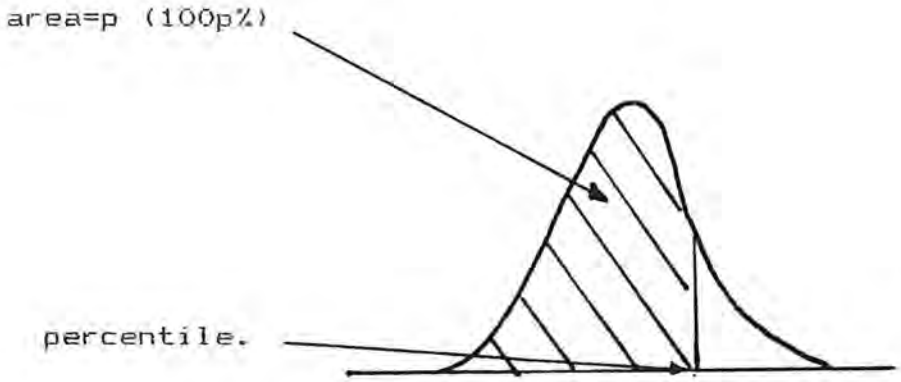
N.B. $(b-\mu)/\sigma$ is just the number of standard deviations from the mean to the interval endpoint.

Picture:



Percentiles

The 100p% percentile of a population distribution is the value exceeding 100p% of population values.



e.g. the well known number 1.96 for the standard normal distribution is its 97.5% point.

95% of a normal distribution lies within 1.96 st. deviations of its mean:

Normal percentiles are available in tables or in most statistical packages. e.g. in SAS:

```
PERCENTILE=PROBIT(P),
```

for the standard normal case, or:

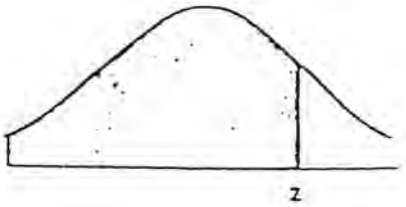
```
PERCENTILE = MU + SIGMA*PROBIT(P),
```

for the general normal case.

N.B. These functions are needed to construct normal probability plots, using statistical packages.

CUMULATIVE PROBABILITIES FOR THE NORMAL DISTRIBUTION

$$F(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{1}{2}t^2} dt$$



z	·00	·01	·02	·03	·04	·05	·06	·07	·08	·09
·0	·5000	·5040	·5080	·5120	·5160	·5199	·5239	·5279	·5319	·5359
·1	·5398	·5438	·5478	·5517	·5557	·5596	·5636	·5675	·5714	·5753
·2	·5793	·5832	·5871	·5910	·5948	·5987	·6026	·6064	·6103	·6141
·3	·6179	·6217	·6255	·6293	·6331	·6368	·6406	·6443	·6480	·6517
·4	·6554	·6591	·6628	·6664	·6700	·6736	·6772	·6808	·6844	·6879
·5	·6915	·6950	·6985	·7019	·7054	·7088	·7123	·7157	·7190	·7224
·6	·7257	·7291	·7324	·7357	·7389	·7422	·7454	·7486	·7517	·7549
·7	·7580	·7611	·7642	·7673	·7704	·7734	·7764	·7794	·7823	·7852
·8	·7881	·7910	·7939	·7967	·7995	·8023	·8051	·8078	·8106	·8133
·9	·8159	·8186	·8212	·8238	·8264	·8289	·8315	·8340	·8365	·8389
1·0	·8413	·8438	·8461	·8485	·8506	·8531	·8554	·8577	·8599	·8621
1·1	·8643	·8665	·8686	·8708	·8729	·8749	·8770	·8790	·8810	·8830
1·2	·8849	·8869	·8888	·8907	·8925	·8944	·8962	·8980	·8997	·9015
1·3	·9032	·9049	·9066	·9082	·9099	·9115	·9131	·9147	·9162	·9177
1·4	·9192	·9207	·9222	·9236	·9251	·9265	·9279	·9292	·9306	·9319
1·5	·9332	·9345	·9357	·9370	·9382	·9394	·9406	·9418	·9429	·9441
1·6	·9452	·9463	·9474	·9484	·9495	·9505	·9515	·9525	·9535	·9545
1·7	·9554	·9564	·9573	·9582	·9591	·9599	·9608	·9616	·9625	·9633
1·8	·9641	·9649	·9656	·9664	·9671	·9678	·9686	·9693	·9699	·9706
1·9	·9713	·9719	·9726	·9732	·9738	·9744	·9750	·9756	·9761	·9767
2·0	·9772	·9778	·9783	·9788	·9793	·9798	·9803	·9808	·9812	·9817
2·1	·9821	·9826	·9830	·9834	·9838	·9842	·9846	·9850	·9854	·9857
2·2	·9861	·9864	·9868	·9871	·9875	·9878	·9881	·9884	·9887	·9890
2·3	·9893	·9896	·9898	·9901	·9904	·9906	·9909	·9911	·9913	·9916
2·4	·9918	·9920	·9922	·9925	·9927	·9929	·9931	·9932	·9934	·9936
2·5	·9938	·9940	·9941	·9943	·9945	·9946	·9948	·9949	·9951	·9952
2·6	·9953	·9955	·9956	·9957	·9959	·9960	·9961	·9962	·9963	·9964
2·7	·9965	·9966	·9967	·9968	·9969	·9970	·9971	·9972	·9973	·9974
2·8	·9974	·9975	·9976	·9977	·9977	·9978	·9979	·9979	·9980	·9981
2·9	·9981	·9982	·9982	·9983	·9984	·9984	·9985	·9985	·9986	·9986
3·0	·9987	·9987	·9987	·9988	·9988	·9989	·9989	·9989	·9990	·9990
3·1	·9990	·9991	·9991	·9991	·9992	·9992	·9992	·9992	·9993	·9993
3·2	·9993	·9993	·9994	·9994	·9994	·9994	·9994	·9995	·9995	·9995
3·3	·9995	·9995	·9995	·9996	·9996	·9996	·9996	·9996	·9996	·9997
3·4	·9997	·9997	·9997	·9997	·9997	·9997	·9997	·9997	·9997	·9998

PERCENTAGE POINTS

2 [1—F(z)]	·1	·05	·02	·01	·002	·001
1—F(z)	·05	·025	·01	·005	·001	·0005
z	1·645	1·960	2·326	2·576	3·090	3·291

CHOOSING A DATA TRANSFORM - WITH TWO CASE STUDIES

1. Concentrating on the distribution of a single sampled variable.

- finding transformations to normalise single univariate data distributions

- one variate at a time in a multivariate cyclone data study

- multivariate distributions also normalised in many cases

2. Where our modelling includes the dependence of Y mean values on the values of an X variable

- the regression modelling situation

- e.g. want to transform Y to Z so that

$$Z = aX + b + \text{error}$$

, with "error"s randomly sampled from a single normal population of mean zero.

Traditional Approaches:

A. Appeal to Theory

e.g. the "central limit theorem" implies Z=logY is a good transformation to try when Y1,.. Yn is a sample from a population of particle sizes.

B. Appeal to Experience

e.g. the log transform has a long history of use in geostatistical studies of geochemical data.

More Modern Approaches:

C. Use of Normal Probability Plots

-traditionally done by hand on normal probability paper

-now done by computer (e.g. SAS PROC UNIVARIATE)

The Normal Probability Plot Idea:

Take the Y values (or residuals from a regression analysis, as appropriate) and put them in increasing order.

If we call these values Y(1),Y(2) ... Y(n) then we have

$$Y(1) < Y(2) < \ldots < Y(n).$$

If the Y's were from a standard normal distribution, then we'd have

$$Y(i) = \text{mean} + (\text{st.dev'n}) * Z(i),$$

with the Z(i)'s being ordered values sampled from a standard normal distribution.

This is a straight line relationship.

Although we can't know the actual Z's for our sample, we can calculate "expected" ordered values for a sample of n points from a normal distribution, to a good approximation. Here "expected" means the average of the ith ordered value over a very large number of samples of n values from a standard normal distribution.

SAS PROC UNIVARIATE uses

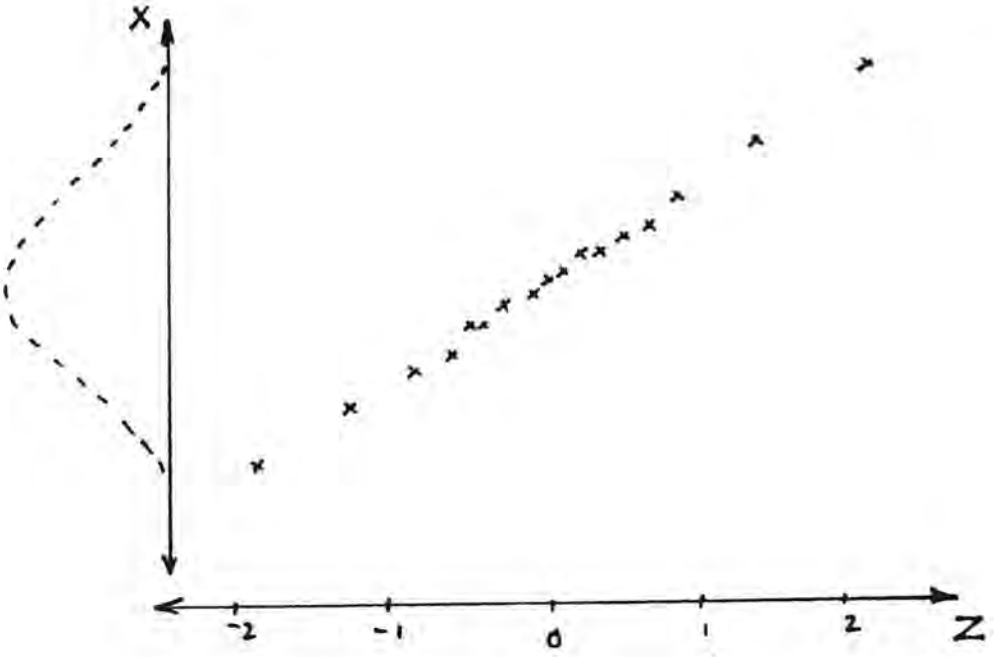
$$E(Z(i)) \quad \text{PROBIT}((i-0.375)/(n+0.25))$$

(the $100i/(n+1)$ % percentile is also reasonable).

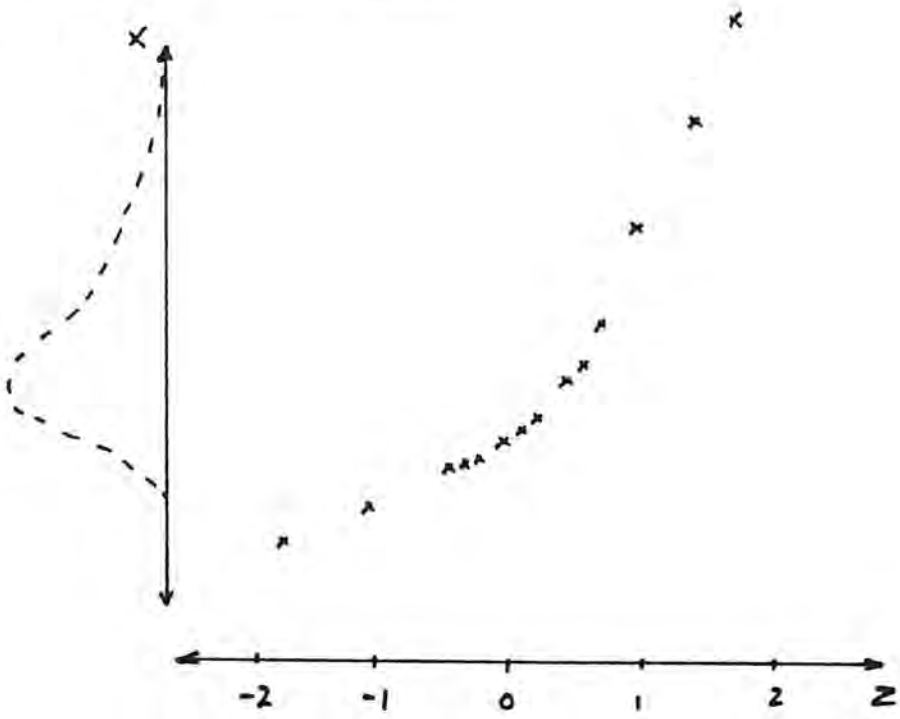
The Normal Probability Plot (or "QQ Plot") for the Y's is obtained by plotting

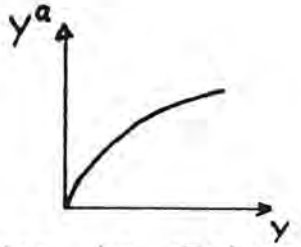
$$Y(i) \text{ versus } E(Z(i)).$$

If the data is already approximately normal then should get a plot like:



The positively skewed case:





Transformations to try:

Power function

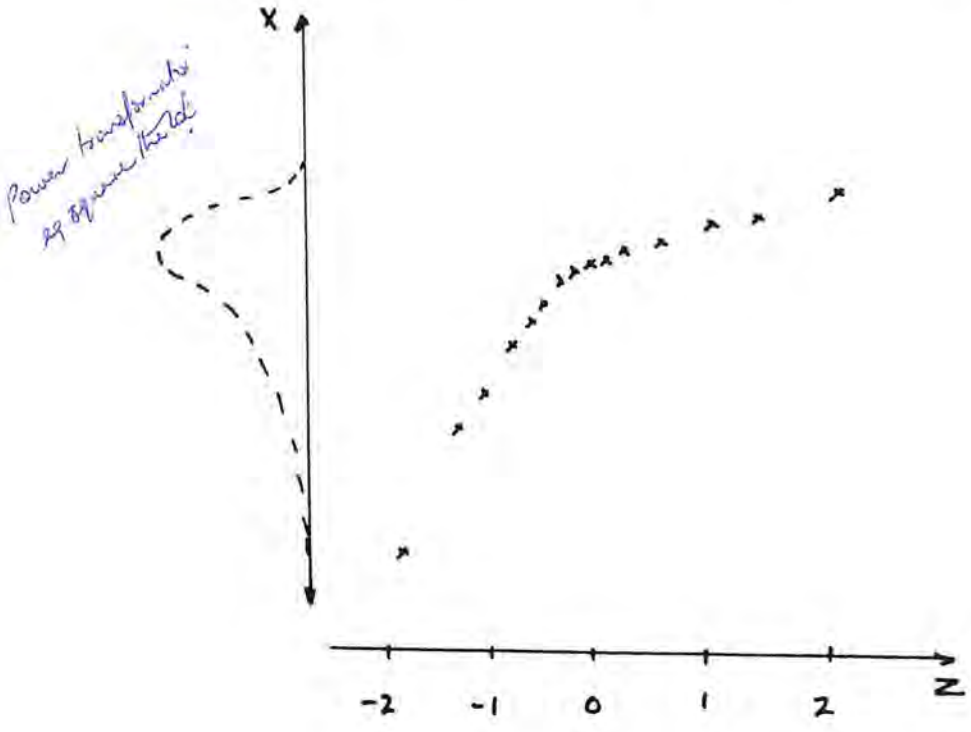
Y^a with a<1 or logY

(decrease distances between high values relative to distances between low values).

Experiment until linear QQ plot is obtained.

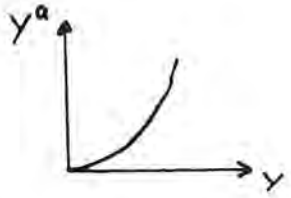
The negatively skewed case:

(can occur if "overtransform" positively skewed data).



Transformations to try:

Try Y^a with a > 1 to give greater relative separation of high values. [N.B. shape of Z versus Y plot is that of required function].



D. Another Method: Inspecting Mean-variance Relationships

- useful only where there is systematic variation in mean levels of the variable under study according to the values of others

- based on the requirement that variances not depend on means on the transformed scale.

The power or log transforms will work well when residual standard deviations (the standard deviations of the error terms) are related to mean values by a power law:

If

$$sd(m) = k.m^b ,$$

then try:

logY, if b=1 (scatter proportional to mean);

Y^a , with a=1-b otherwise.

N.B. A common special case occurs with count data having a Poisson distribution, in which case the scatter (the standard dev'n) of Y values is proportional to the square root of their means (b=0.5), and the square root transformation (a=0.5) will be appropriate.

Graphical implementation:

- Do an initial regression model fit to obtain fitted values, FV(explanatory variables), and residuals, Y - FV.

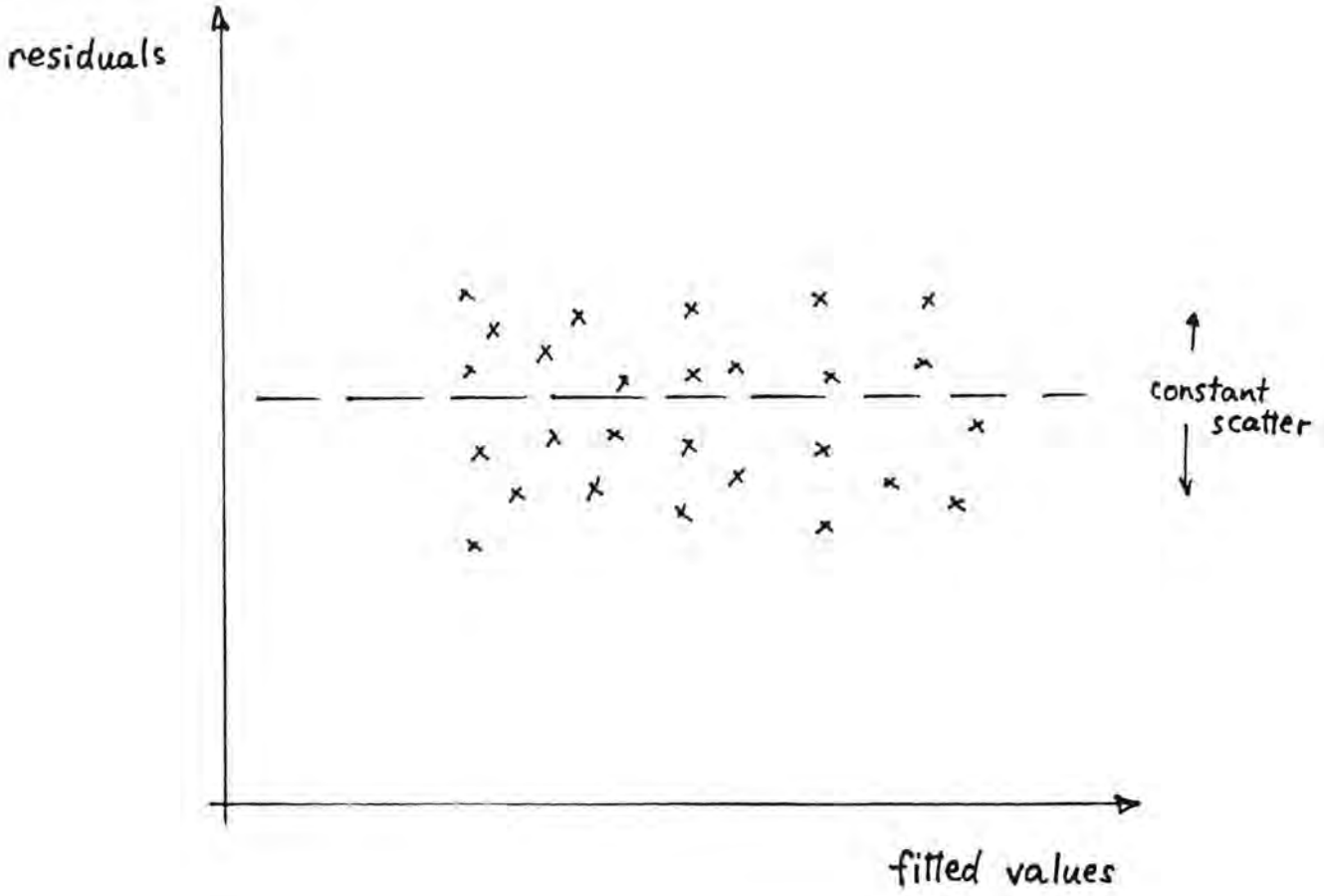
- Graph residuals against fitted values.

- Assess the residual standard deviation versus mean value relation "by eye".

- Make a guess at a transformation and repeat previous steps.

- Stop when scatter appears not to vary with fitted values.

Picture:



E. Maximum Likelihood Approach

Concentrate once again on the "Box-Cox" family of transformations (1964 paper):

$$Y \rightarrow \begin{cases} Y^a & \text{for } a \neq 0, \\ \log Y & \text{for } a = 0. \end{cases}$$

- applies to positive variables
- capable of fixing skewness of distribution
- capable of fixing inhomogeneity of variance
- have proven useful in many practical situations

One reference: Draper & Smith, "Applied Regression Analysis", 2nd edition.

Idea of Maximum Likelihood in General:

- Need a parametric model for data probabilities. (e.g. that $Y1^a$, $Y2^a$, ... Yn^a are a statistically independent random sample from a normal population with mean μ and standard deviation σ)

- Apply a numerical optimisation procedure to choose the parameter values (μ, σ ,a above) for which the model's probability for the observed data is greatest.

In the example given this amounts to maximising:

$$f(y_1, \ldots, y_n \mid \mu, \sigma, a) = \left(\frac{a \dot{y}^{a-1}}{\sqrt{2\pi}\,\sigma}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i^a - \mu}{\sigma}\right)^2}$$

, where $\dot{y}$ denotes the geometric mean of the data values:

$$\dot{y} = (y1.y2.y3. \ldots yn)$$

Practical Implementation:

For data from a single skew distribution or a regression situation, maximising the likelihood function boils down to the following recipe (GENSTAT, GLIM or SAS macro):

(a) Standardise the Yi's to have unit geometric mean (divide by $\dot{Y}$).

(b) Search for the minimum of

$$LLK(a) = n.\log(\text{residual SS}/n),$$

where the residual sum of squares is from a regression of

$$Z = Y^{a}/a$$

on the explanatory variables; or is simply

$$\sum_{i=1}^{n} (z_i - \bar{z})^2$$

when there are no explanatory variables.

N.B. (i) Minimising the "log-likelihood function" LLK(a) can be done by minimising the residual SS itself. LLK(a) is used because it is easier to pick out the minimum from a plot of LLK(a) versus a (a few points can be sufficient for this), and confidence limits for a can be constructed by cutting LLK(a) a certain distance above its minimum.

(ii) Standard procedures for doing the optimisation are not included in statistical packages in general, but macros can & have been programmed to automate the steps required.

CASE STUDY ONE

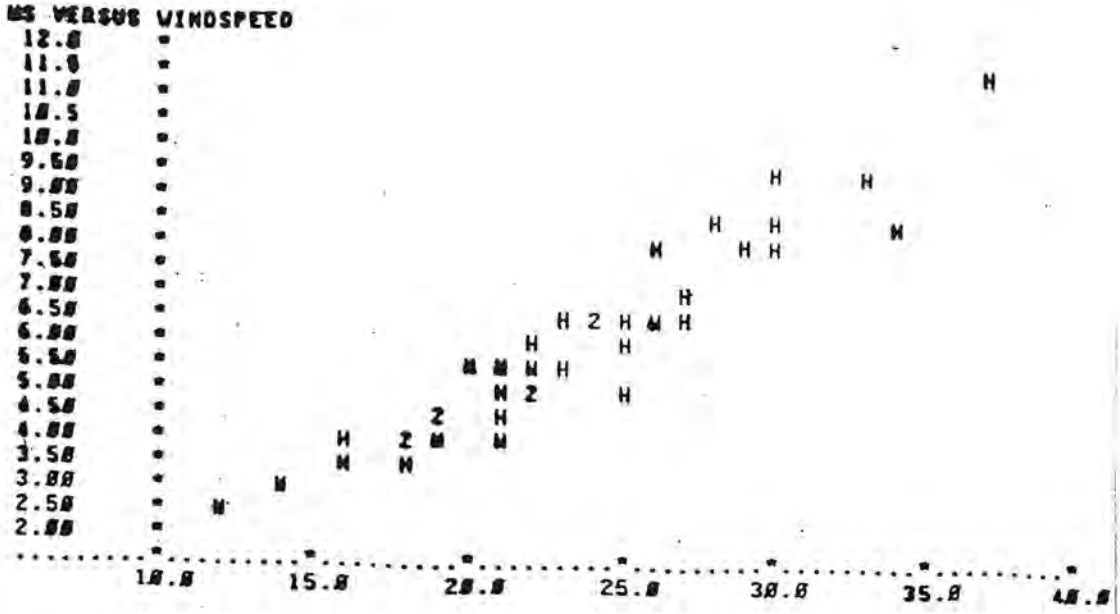
Tasks:

1. Model joint distribution of cyclone-induced wave, wind and current variables (from 38 data points "hindcast" from historical cyclones that came within 200km of a certain structure).

2. Develop procedure for simulating joint wave, wind and current occurrences.

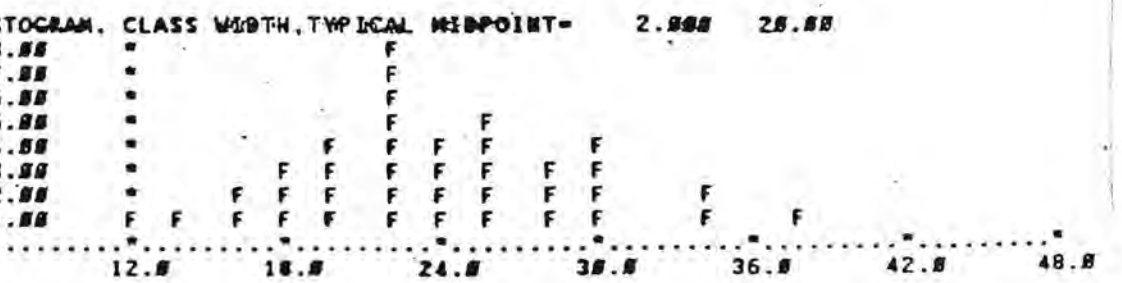
3. Use this as a risk analysis tool.

The Data:



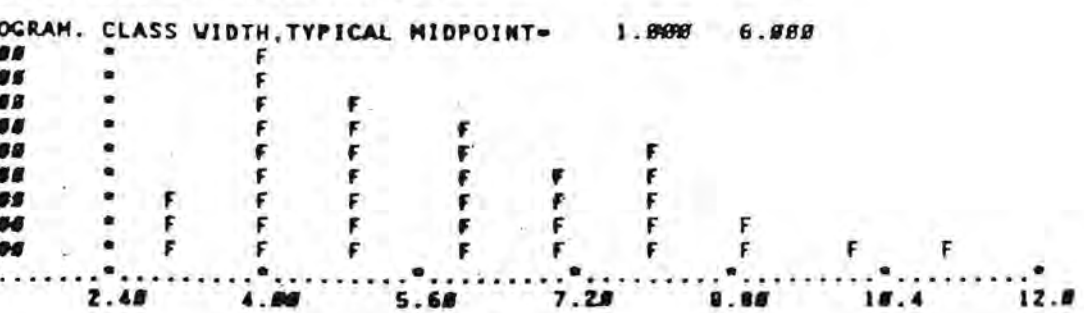
WIND SPEED

MIN, MAX = 12.00 37.00
MEAN, ST DEVN = 23.34 5.542



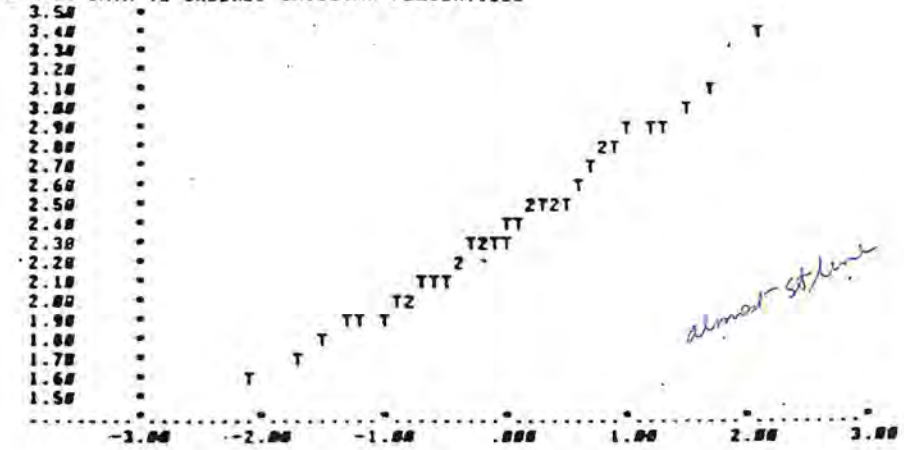
SIGNIFICANT WAVE HEIGHT

MIN, MAX = 2.700 11.40
MEAN, ST DEVN = 5.911 2.021



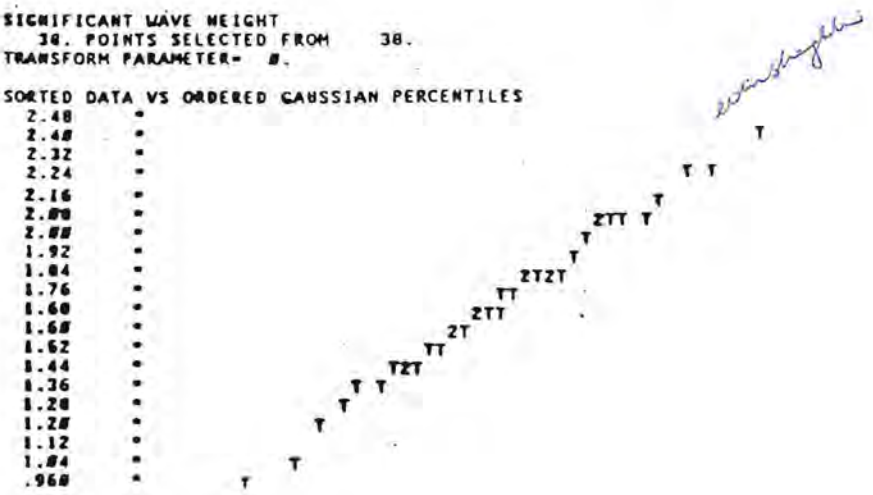
QQ Plots for Transforms of Hs

$H_s \quad (a=1)$



$H_s^{\frac{1}{2}} \quad (a=\frac{1}{2})$

ie √ of H_s



$\log H_s \quad (a=0)$

ie Log H_s

QQ Plots for Transforms of Windspeed

```
WINDSPEED
   30. POINTS SELECTED FROM    30.
TRANSFORM PARAMETER-   1.000

SORTED DATA VS ORDERED GAUSSIAN PERCENTILES
  40.0     *                                                        T
  38.4     *                                                      T T
  36.8     *                                                    T TT
  35.2     *                                                   TT
  33.6     *                                                  TT
  32.0     *                                               T2TT
  30.4     *                                              2
  28.8     *                                          2T2T
  27.2     *                                      TT2T
  25.6     *                                   2T
  24.0     *                                T TT
  22.4     *                             T T
  20.8     *                          T
  19.2     *                    T
  17.6     *
  16.0     *
  14.4     *
  12.8     *
  11.2     *
  9.60     *
  8.00     *
     ------*---------*---------*---------*---------*---------*---------*
          -3.00     -2.00     -1.00      .000      1.00      2.00      3.00
```

a = 1

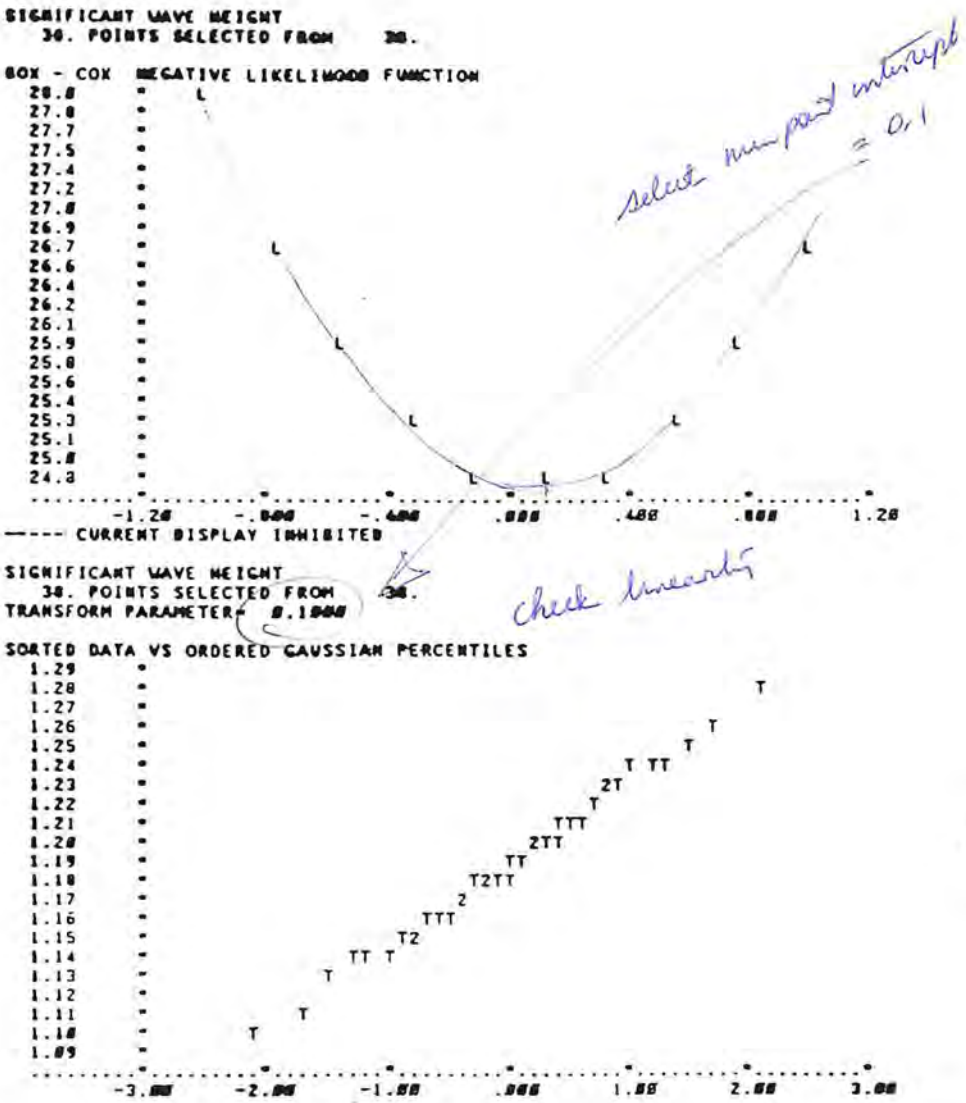
```
WINDSPEED
   30. POINTS SELECTED FROM    30.
TRANSFORM PARAMETER-   0.

SORTED DATA VS ORDERED GAUSSIAN PERCENTILES
  3.60     *                                                        T
  3.54     *                                                      T
  3.48     *                                                    T
  3.42     *                                                  T TT
  3.36     *                                                 TT
  3.30     *                                               TT
  3.24     *                                            T2TT
  3.18     *                                           2
  3.12     *                                       2T2T
  3.06     *                                    T2T
  3.00     *                                   T
  2.94     *                                2T
  2.88     *                             T TT
  2.82     *
  2.76     *                          T T
  2.70     *
  2.64     *                       T
  2.58     *
  2.52     *
  2.46     *                 T
     ------*---------*---------*---------*---------*---------*---------*
          -3.00     -2.00     -1.00      .000      1.00      2.00      3.00
```

a = 0

over transformed

Box-Cox Results for Hs



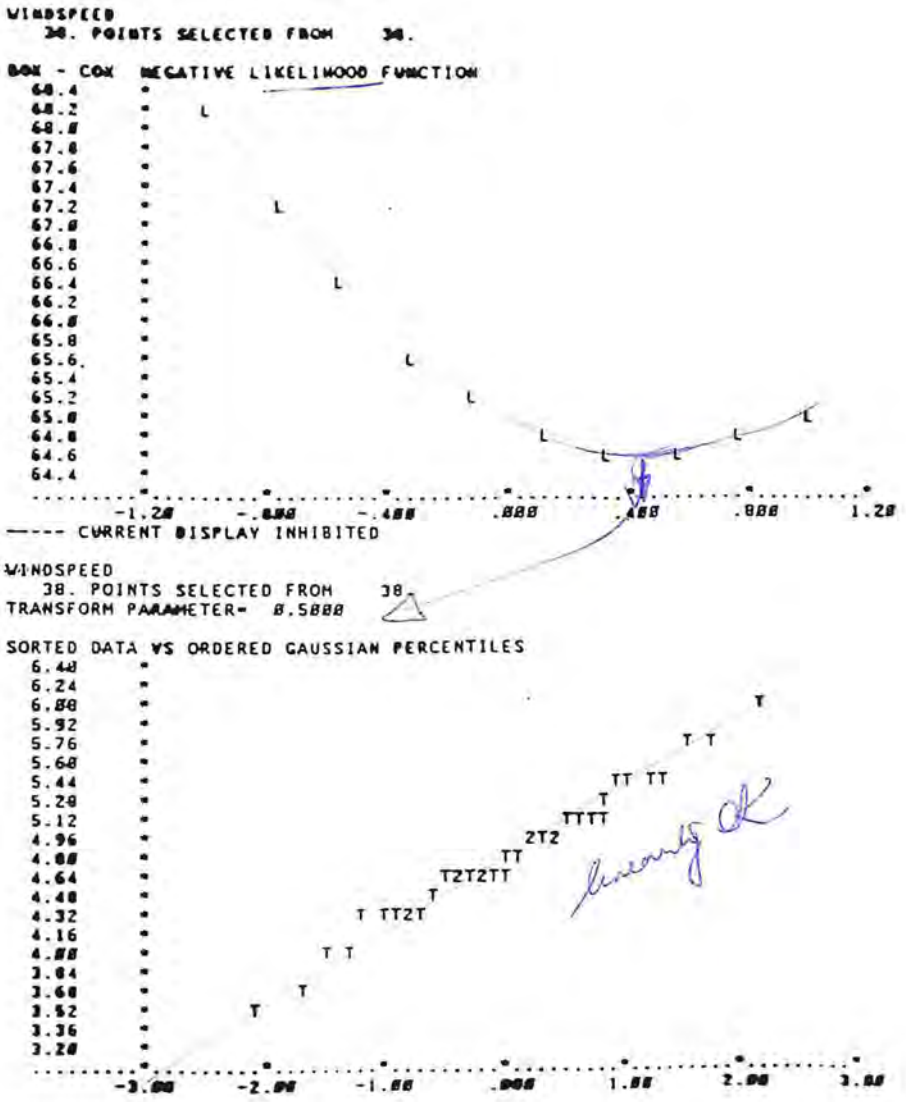
select min pool interval

= 0.1

check linearity

a = 0.1

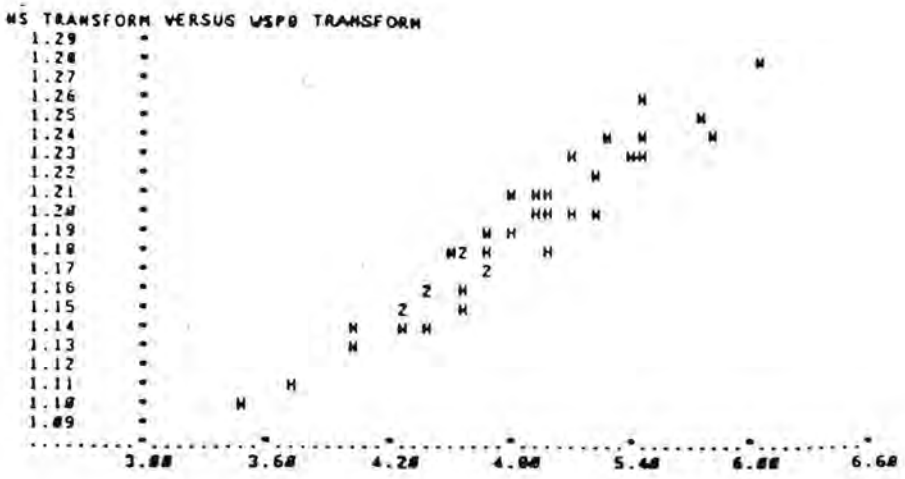
Box-Cox Results for Windspeed



$a = 0.5$

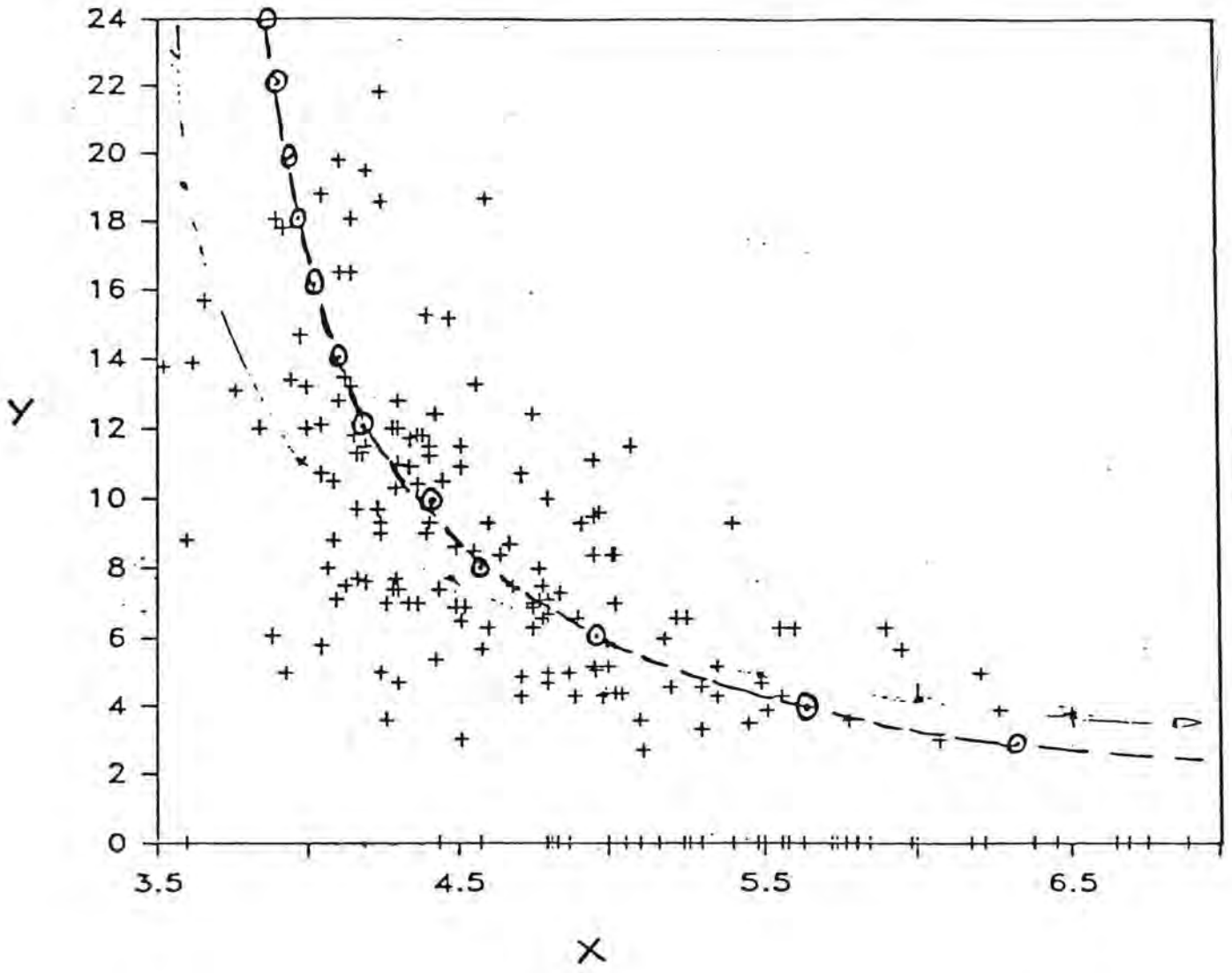
Maximum likelihood approach

Crossplot of Transformed Data

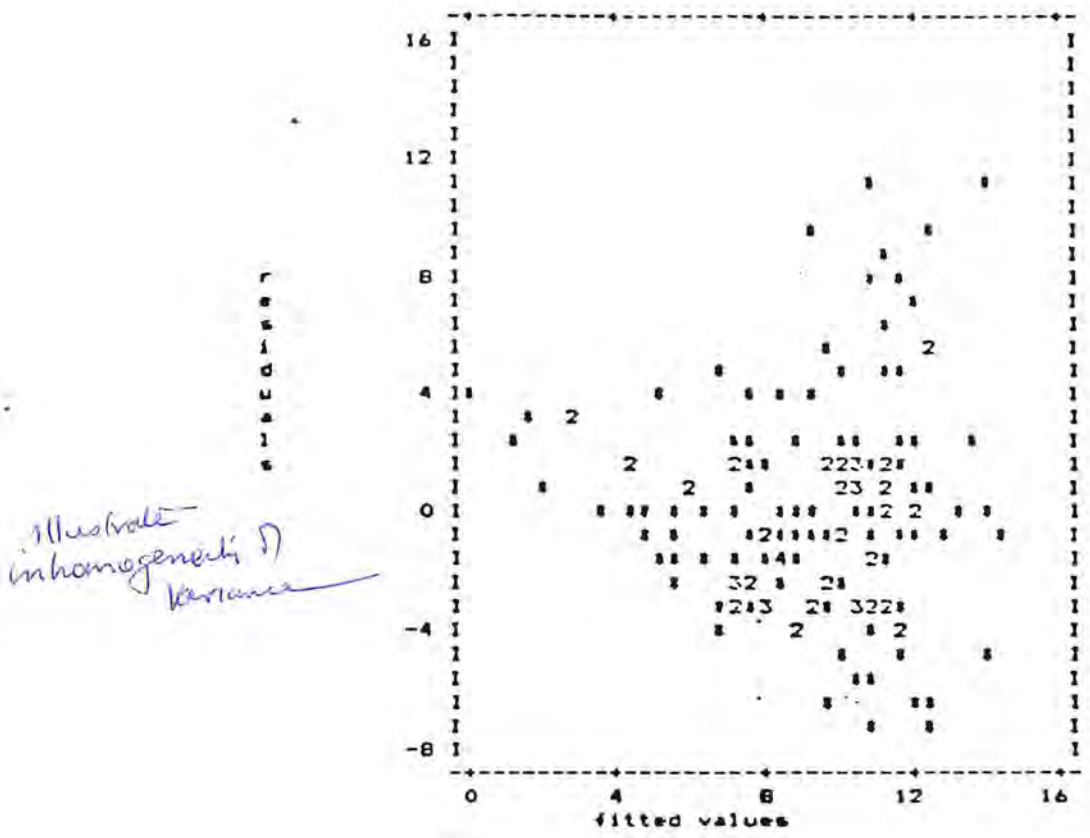


CASE STUDY TWO

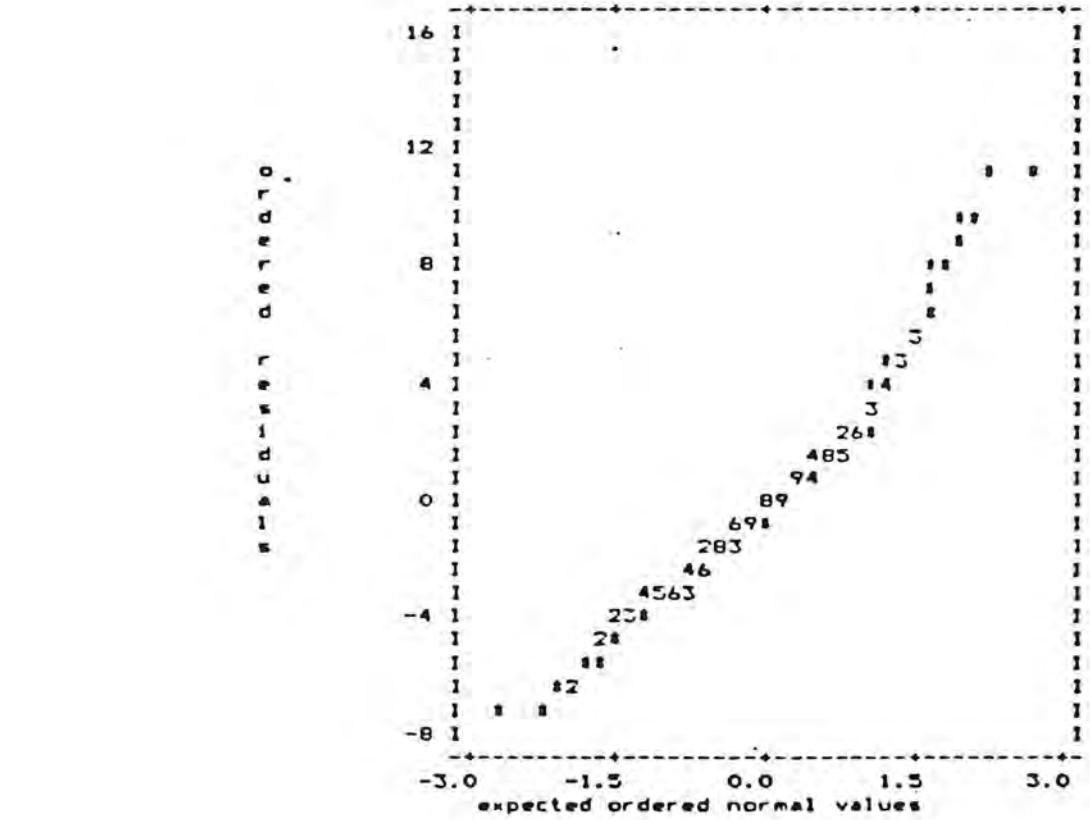
Aim: To model the dependence of Y on X in the following data from a soil rehabilition area:



Linear Regression on X:

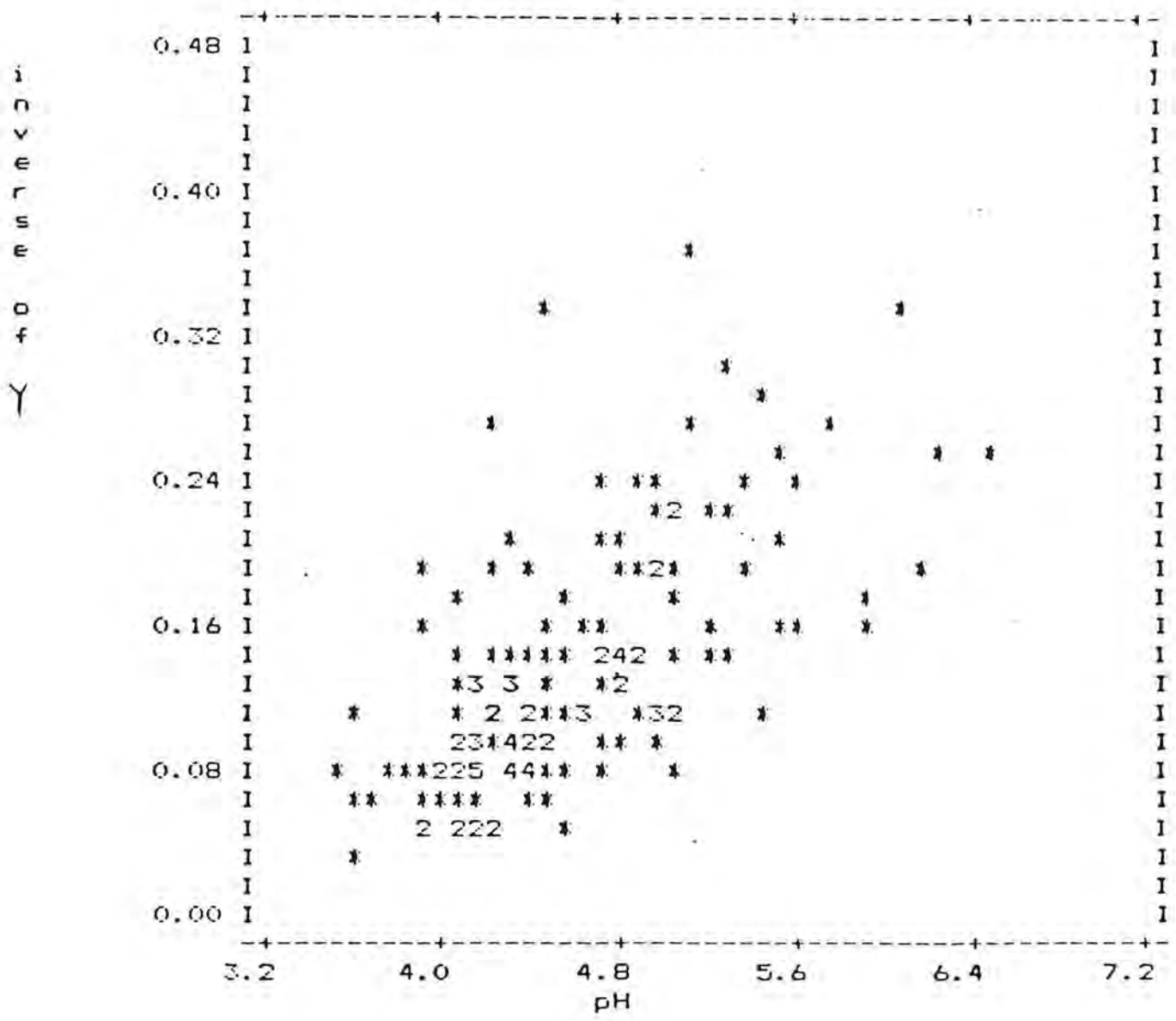


Illustrate inhomogeneity of variance

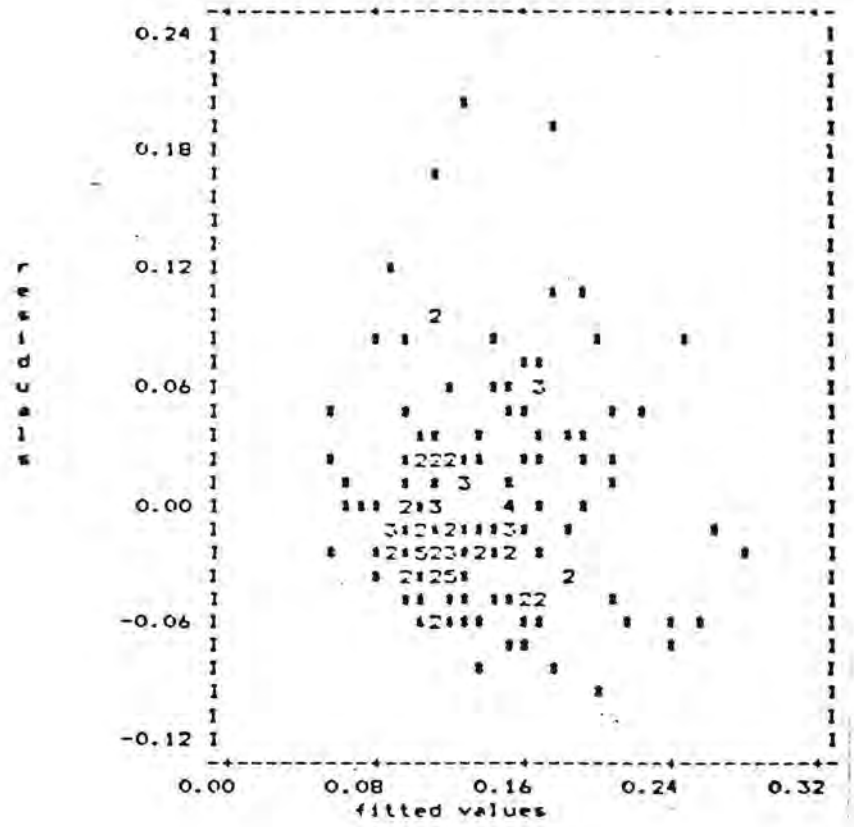


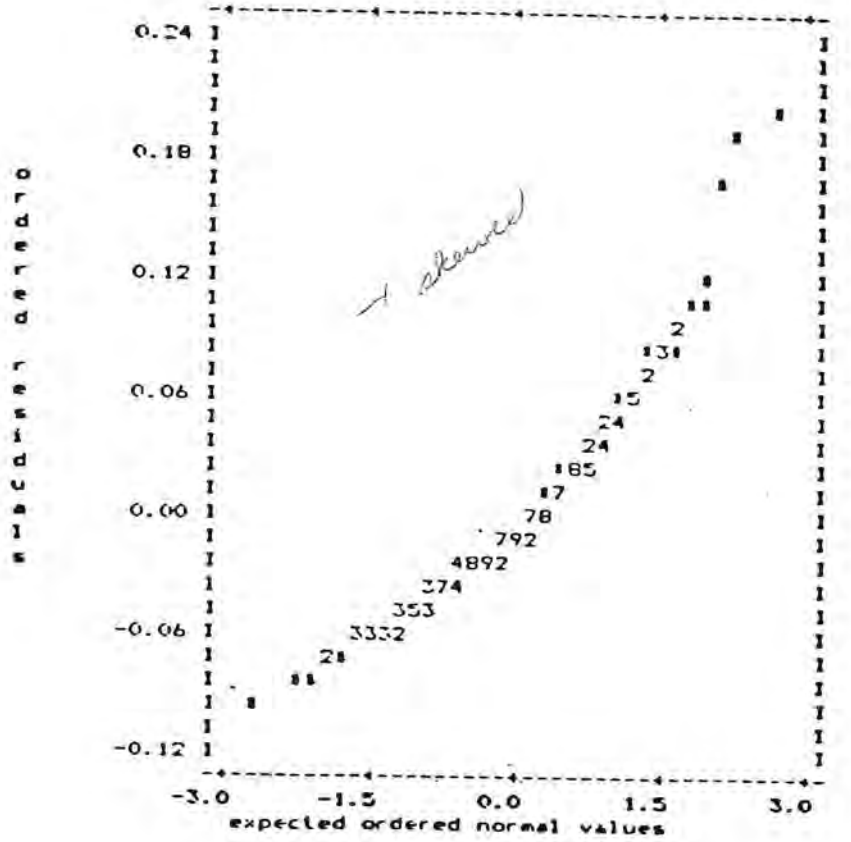
1/Y (a= -1) versus X

hyperbolic rel ship



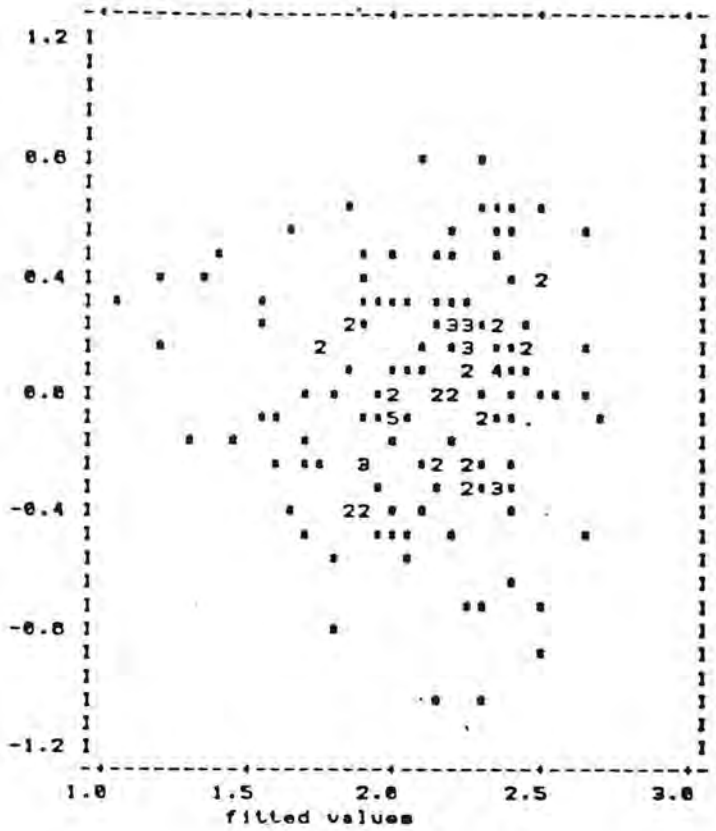
Residuals from Linear Regression of 1/Y on X

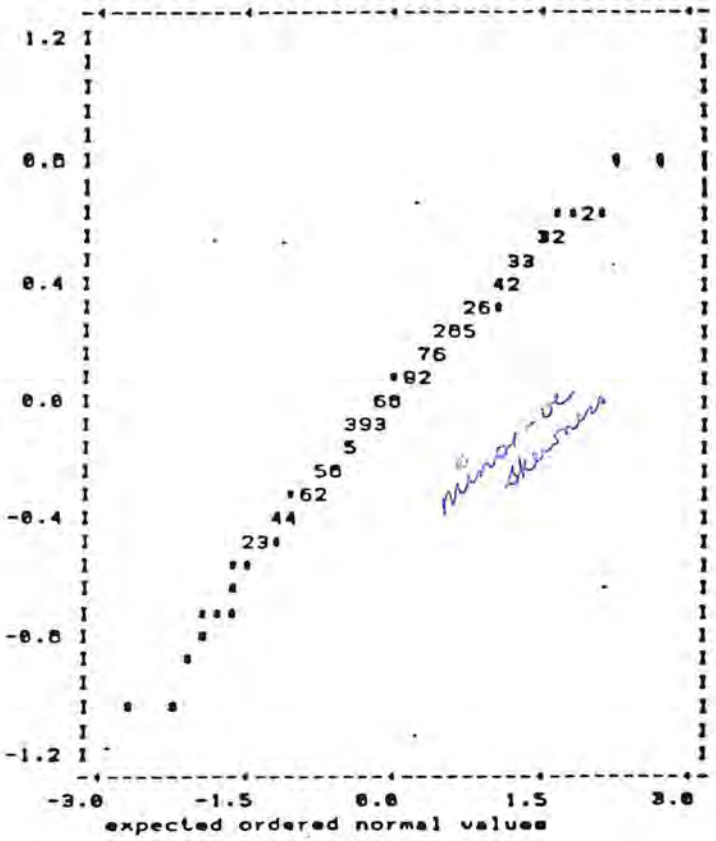




log-linear regression on X

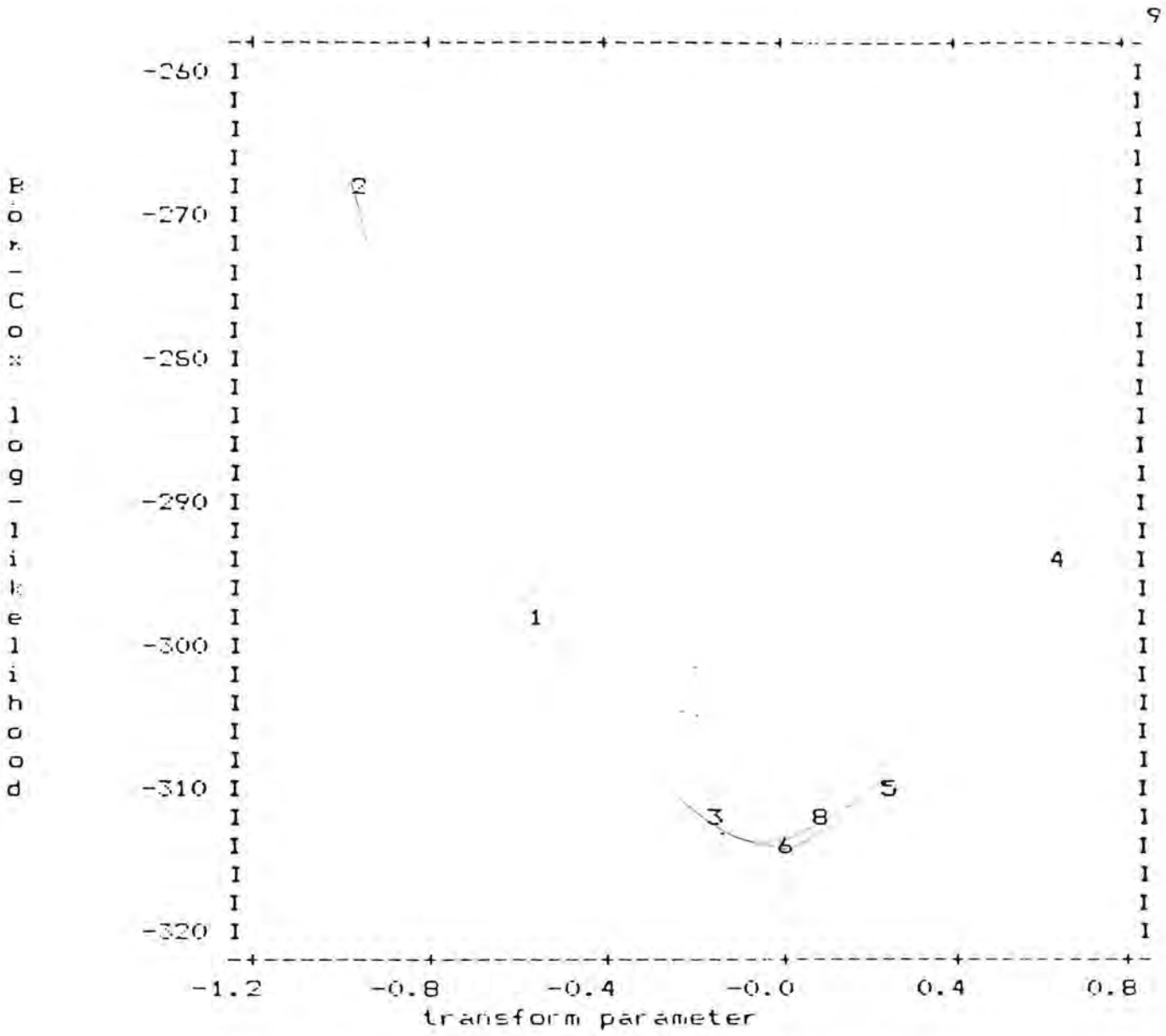
Residuals from Linear Regression of logY on X





Box-Cox Choice of Transform (using GENSTAT)

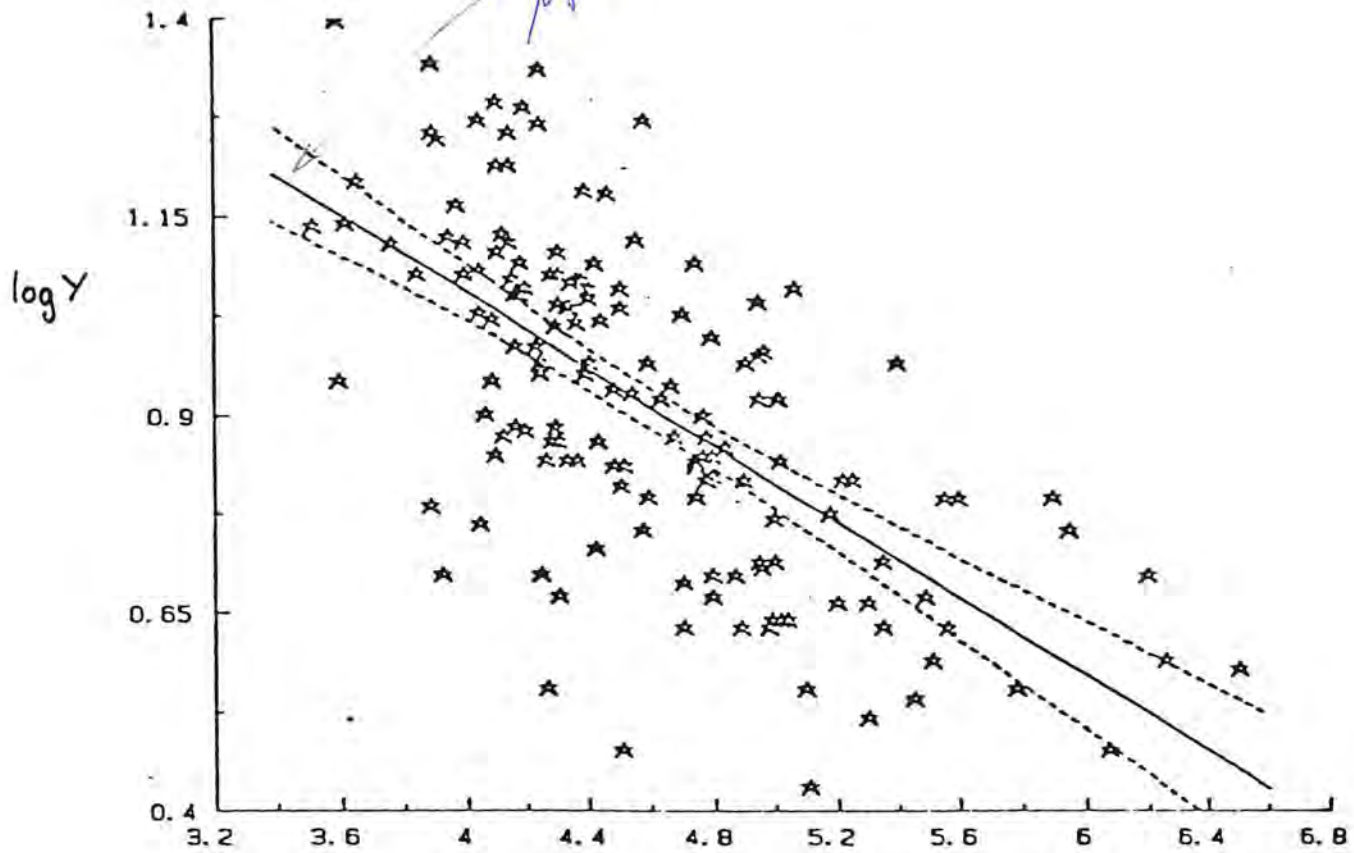
max likelihood

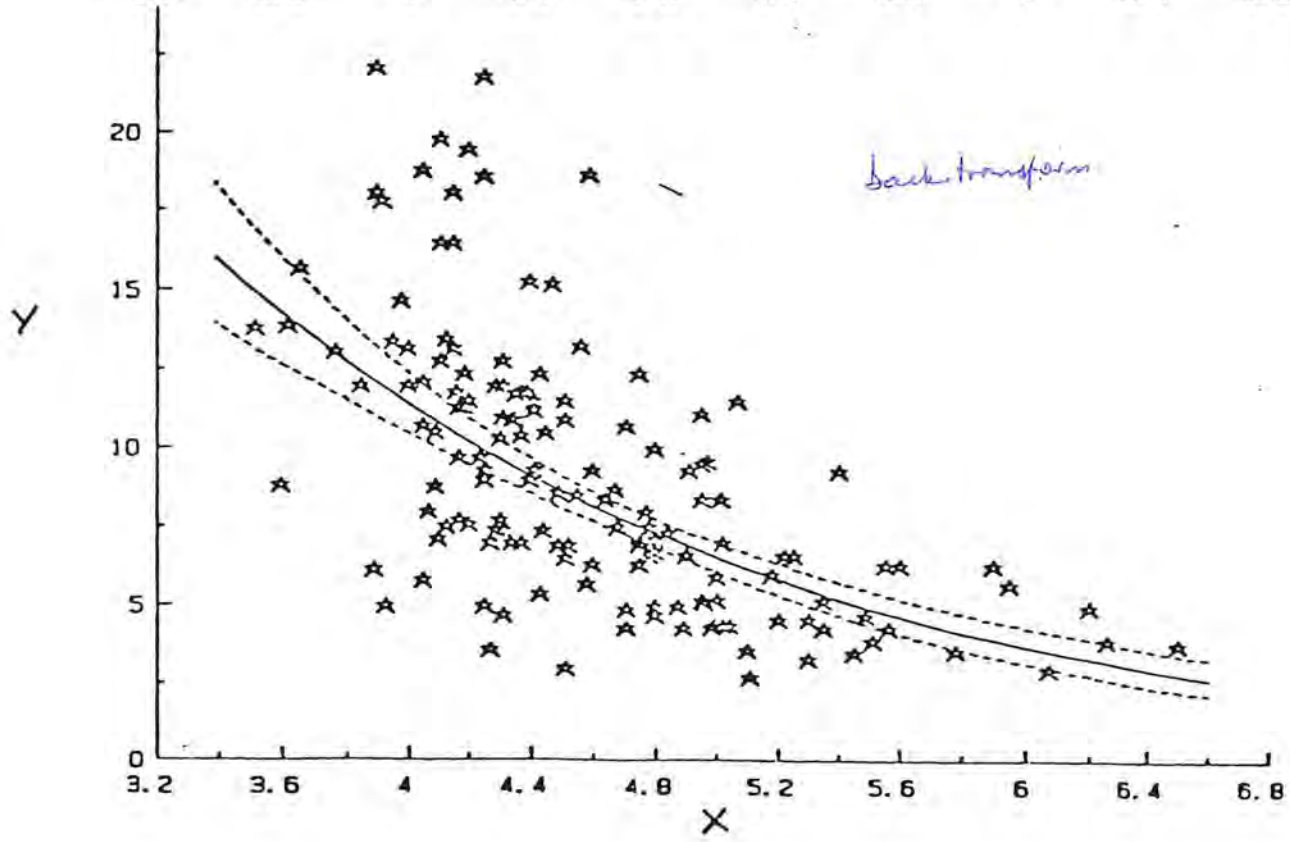


evalno	lvals	llkvals
1	-6.0000E -1	-2.9716E 2
2	-1.0000E 0	-2.6746E 2
3	-2.0000E -1	-3.1180E 2
4	6.0000E -1	-2.9383E 2
5	2.0000E -1	-3.1063E 2
6	-2.9599E -2	-3.1324E 2
7	-1.4960E -1	-3.1253E 2
8	9.0401E -2	-3.1253E 2
9	*	*
10	*	*
11	-2.9531E -2	-3.1324E 2

optimum = 0.0

The Final Model

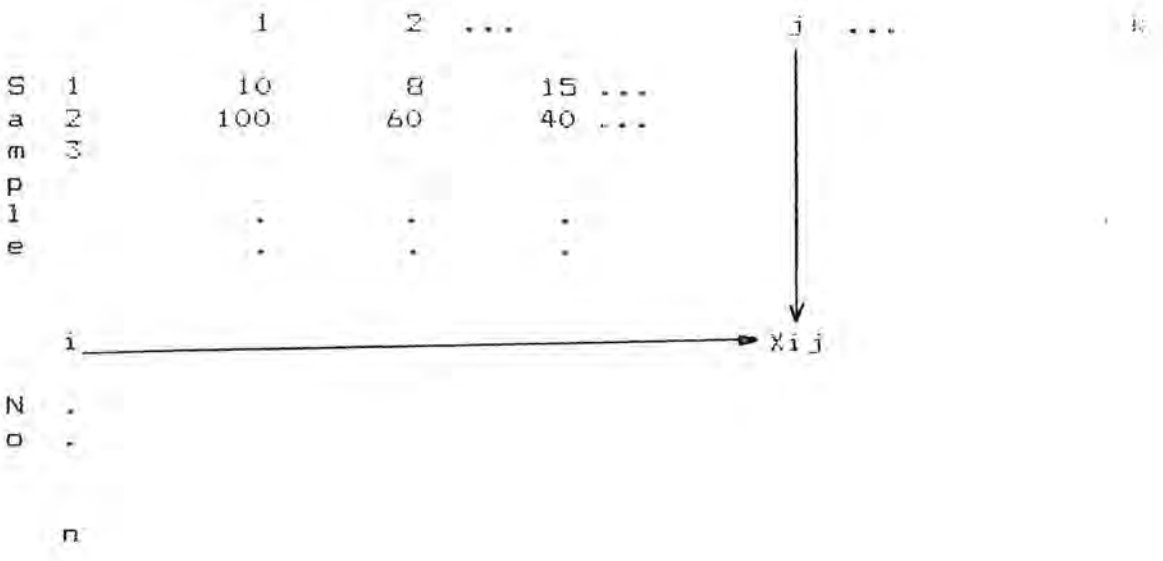




5. MULTIVARIATE DATA - SUMMARY AND PRESENTATION

Setting:

Data matrix:



k possibly related measurements on n samples.

Notation:

Xij = the value of the jth random variable when measured on the ith individual.

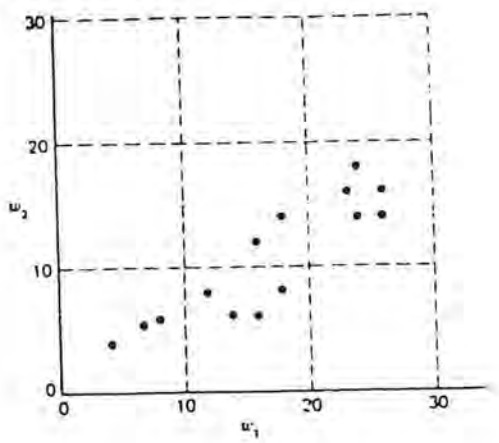
The array of variable values for an individual

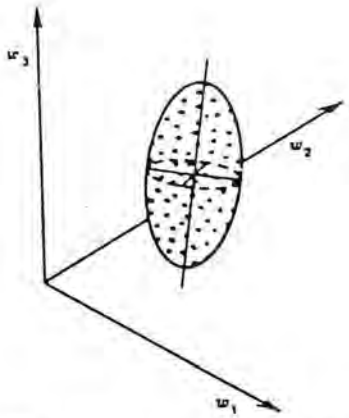
e.g. (total length, alar extent, beak+head length, humerus length, keel of sternum length)

= (156, 245, 31.6, 18.5, 20.5)

is often referred to (and thought of) as a (randomly chosen) "point" from one (or more) populations of such arrays in k dimensional "space".

In two and three dimensions, scatter plots can be used to give pictures of this view of multidimensional data:





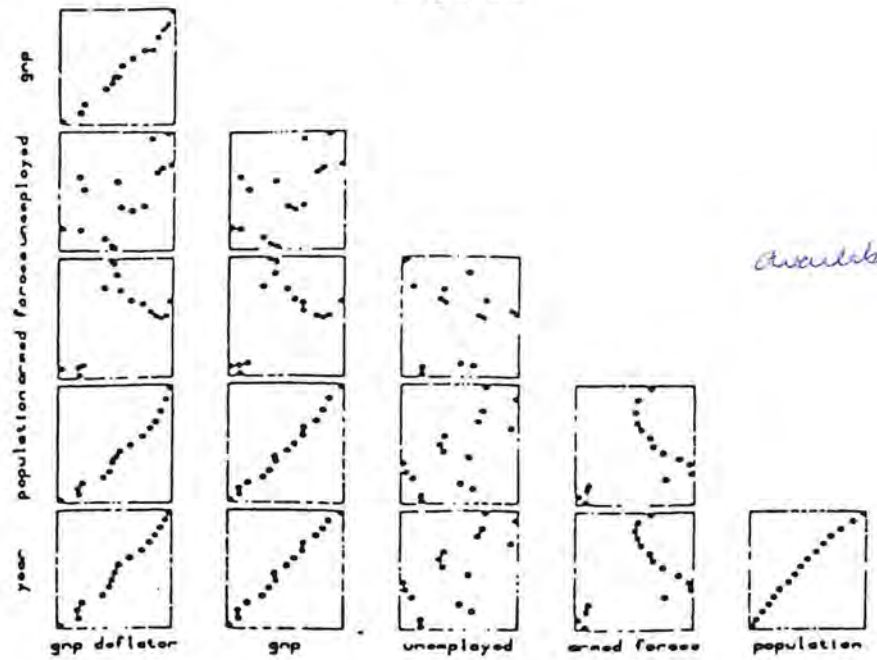
Plot of trivariate observations on a scatter diagram in three dimensions.

Even when there are more than two variables, two variable at a time scatter plots are important to understanding your data. But it is easy to be overwhelmed with the number of such plots when k is large:

k	No. of pairs out of k (k(k+1)/2)
3	3
4	6
5	10
6	15
7	21
8	28

The "draughtsman's layout" (available form "S") is one way to ease the confusion:

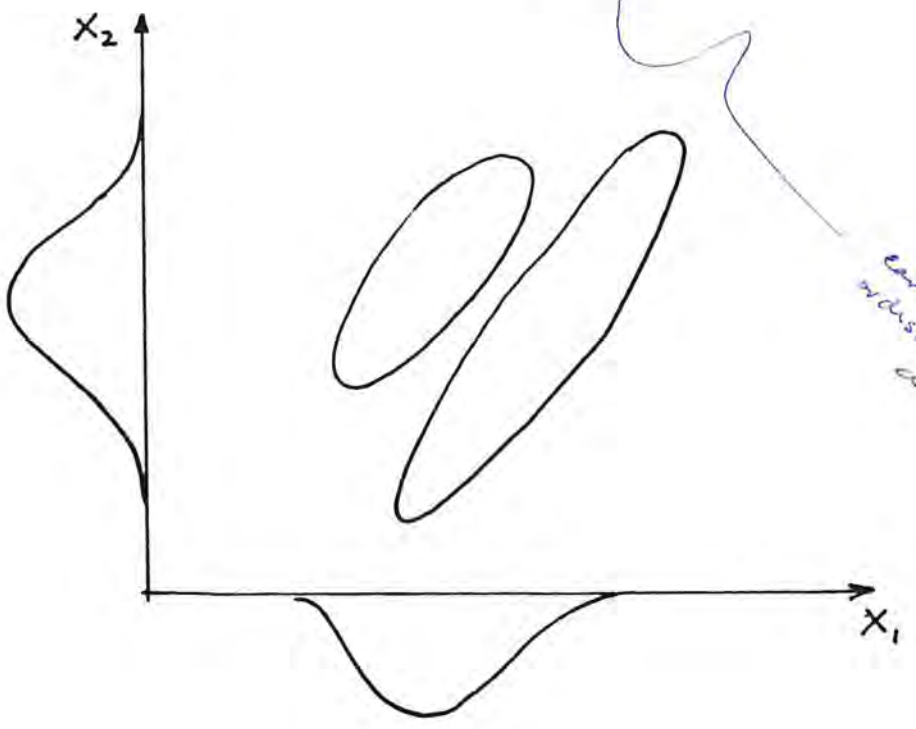
Longley Data



available in "S" package

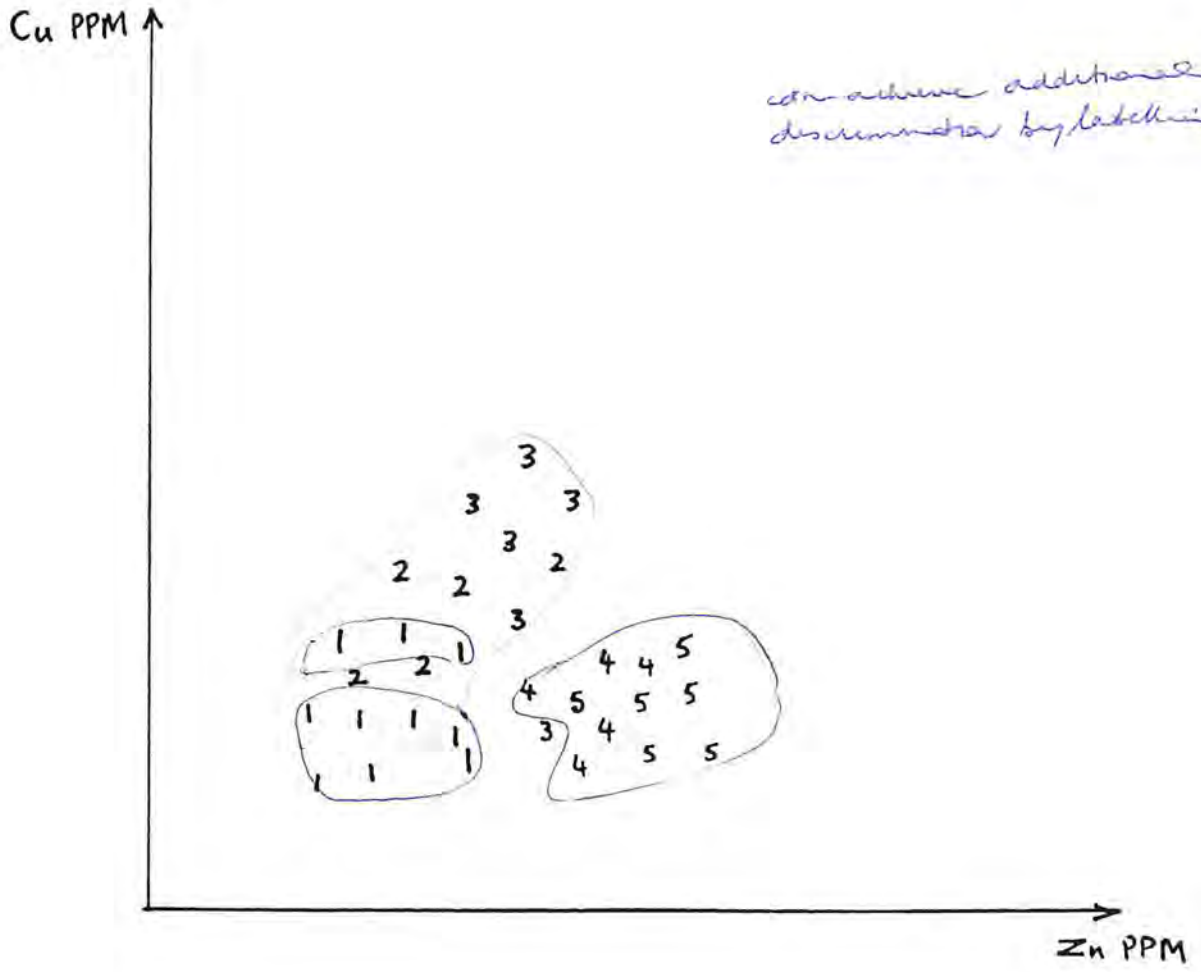
There are many ways in which two at a time scatter plots can fail to give a full picture.

Such possibilities are illustrated by the following case in which single variable summaries or plots would fail to reveal an important split into groups in two dimensions:



To avoid similar traps in two at a time plots, schemes are needed to picture 3 or more variables in ways allowing visual discrimination of data groupings or other patterns.

Three variables can be displayed on a single bivariate plot by using the 3rd variable to code the symbols used (PLOT Y*X=Z in SAS). e.g.:

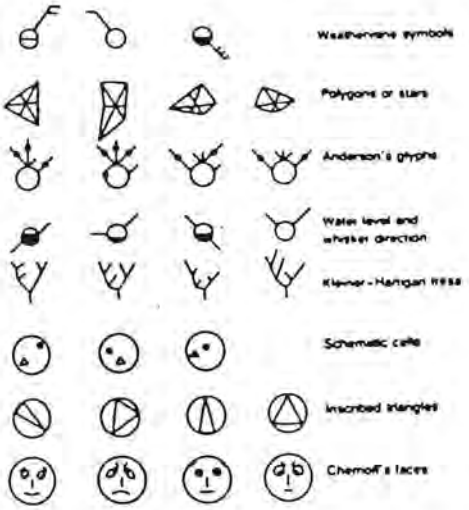


Legend:

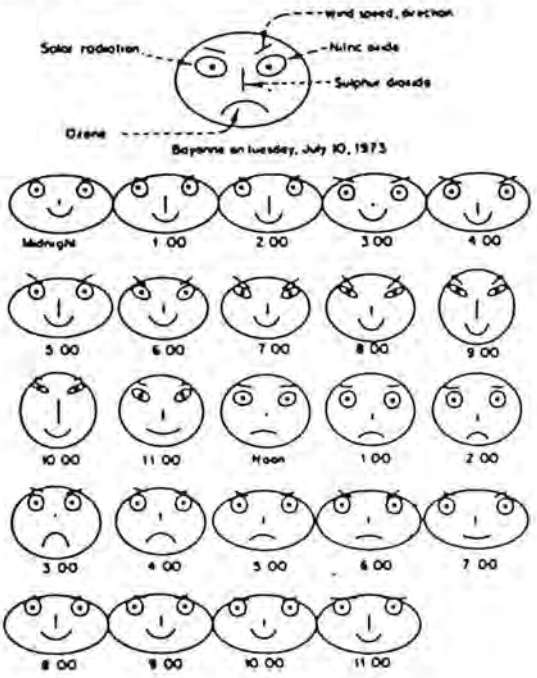
Point Symbol (Colour)	Pb PPM
1	0 - 5
2	5 - 25
3	25 - 125
4	125 - 625
5	625+

There is a range of more sophisticated pictorial codings available for multivariate data:

Several composite symbols into which to code "back" variates



Example use of Chernoff's faces:



Using such symbols information about more than one extra variable can be coded in the symbols for points on what is basically a bivariate plot.

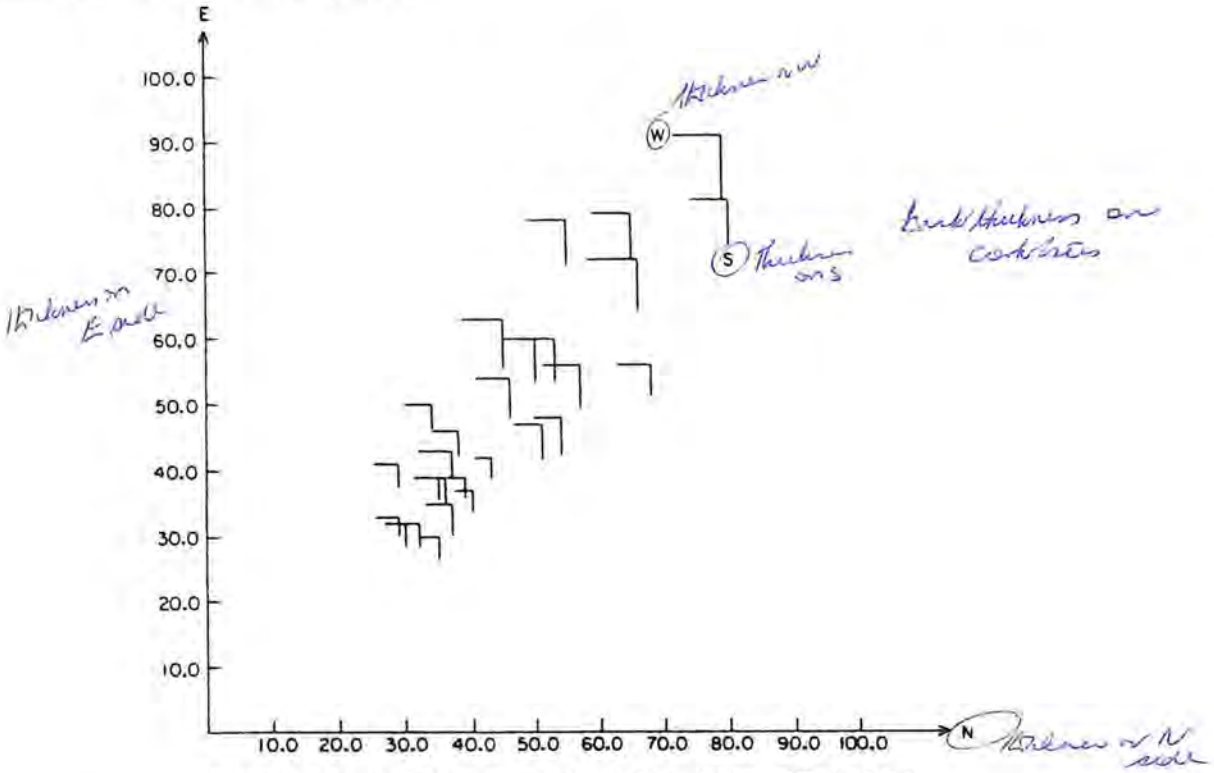
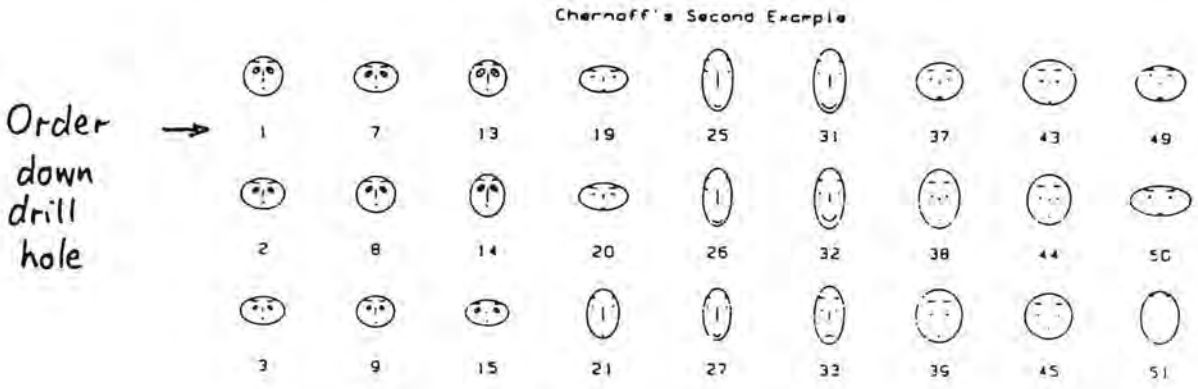
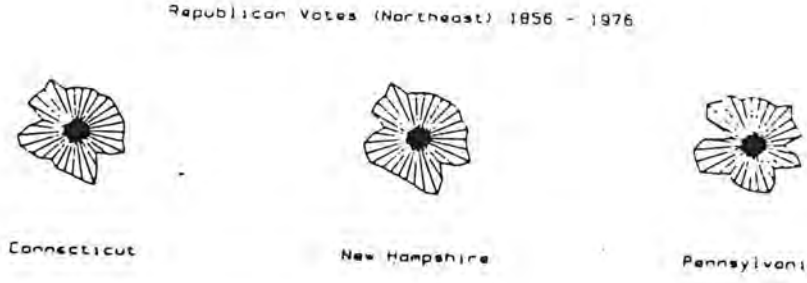


Figure 1.7.3 A glyph representation of the cork data of Table 1.4.1.

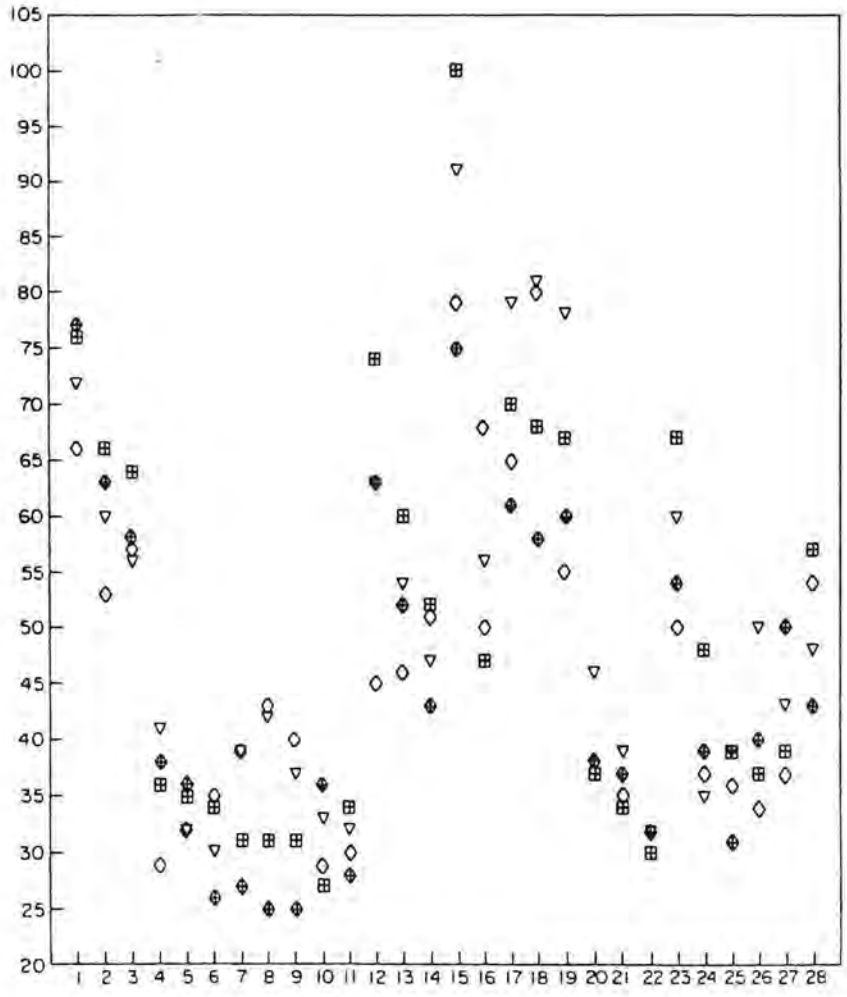
Where data has been collected from different positions on a map, such symbols could be pasted or directly plotted onto the map to get a picture of spatial variation in multivariate "signature".

Chernoff's faces and star symbols are available form the statistical package S (under Unix or VMS). e.g. the following output:





A final data display/exploration idea to mention is that of plotting all variables at once on the same axis (standardising in some way first if different measurement scales are involved), using sample number or (say) distance along a transect for the x-variable. e.g.:



Consecutive univariate representation (after Pearson, 1956). ▽ = N, ◇ = E, ⊞ = S, ⊕ = W.

Summary Statistics:

The main non-univariate summary statistics are bivariate - measuring the degree of linear association between two variable. They are:

1. Covariance (cov(X,Y))

= population mean value of (X-meanX).(Y-meanY).

The sample covariance used as an estimator of the population quantity is:

= (the sum of $\frac{(Xi-\bar{X}).(Yi-\bar{Y})}{(n-1)}$)

as for sample variance

2. Correlation

Covariances have dimensions: (scale of X).(scale of Y).

As well as reflecting the degree of linear association, they are determined by the extent of X and Y scatter. By contrast, correlations are dimensionless quantities reflecting only the degree of linear association:

correlation X,Y = covariance(X,Y) / $\sigma_x . \sigma_y$

population st. deviations

Alternatively, can regard the correlation of X and Y as the covariance of the standardised forms of these variables:

$(X - \mu_x)/\sigma_x$ & $(Y - \mu_y)/\sigma_y$.

unit standard deviations

Correlations (often denoted ρ) always lie between -1 and 1 (inclusive).

The meaning of high correlation:

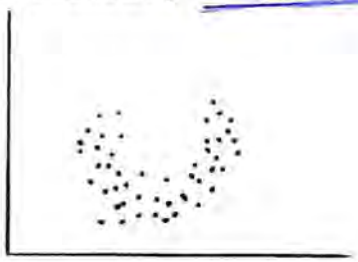
(a) ρ is close to 1 when $Y-\mu_y$ is very nearly equal to $a(X-\mu_x)$, with $a>0$. If $\rho=1$ then $Y=\mu_y+a(X-\mu_x)$ exactly.

i.e. ρ is near one when positive X deviations from the mean are closely associated with positive Y deviations, in a nearly linear relation.

(b) ρ is close to -1 when $Y-\mu_y$ is very nearly equal to $a(X-\mu_x)$ with $a<0$.

The Meaning of Near Zero Correlations:

Associations don't exist, or are non-linear in nature. e.g.:



Unless checks are made for nonlinearities in relationships, correlations shouldn't always be trusted (transformations can be important here).

Estimating Correlations:

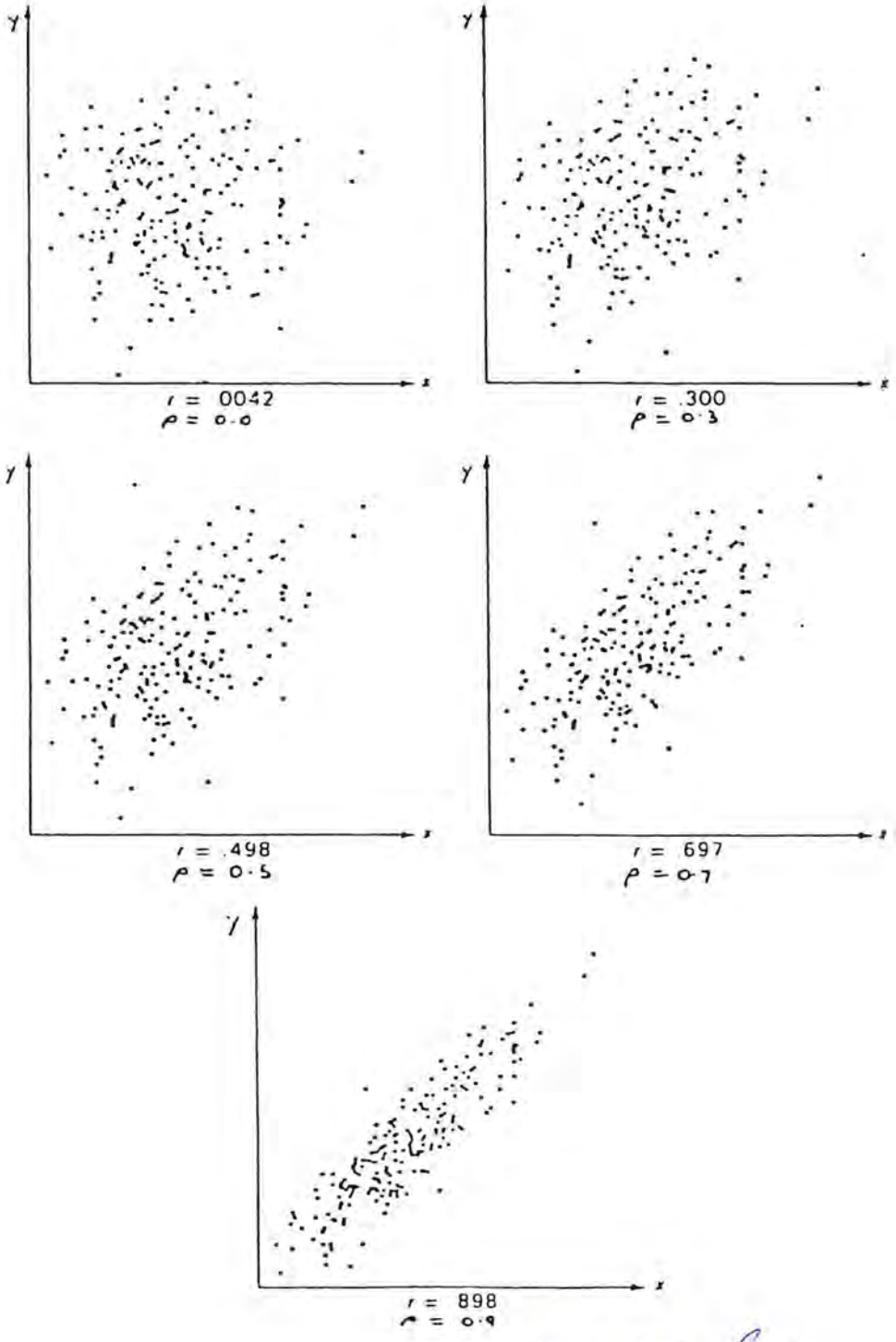
$$r = s_{xy} / s_x s_y$$

$$= \frac{\text{sum of } (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\text{sum } (X_i - \bar{X})^2 \text{ sum } (Y_i - \bar{Y})^2}}$$

The following page illustrates a series of correlations in cases (bivariate normal) where associations are linear, to the extent that they exist at all.

Example Values of r_s With n=100

("random" computer selections from bivariate populations)



for straight line $r_s = 1$

(perfect)

6. THE MULTIVARATE NORMAL MODEL

This is essentially the only model for the "joint" distribution of multivariate data for which a decent, practically useable statistical theory exists.

One requirement for this model to apply is that each single variable be normally distributed:

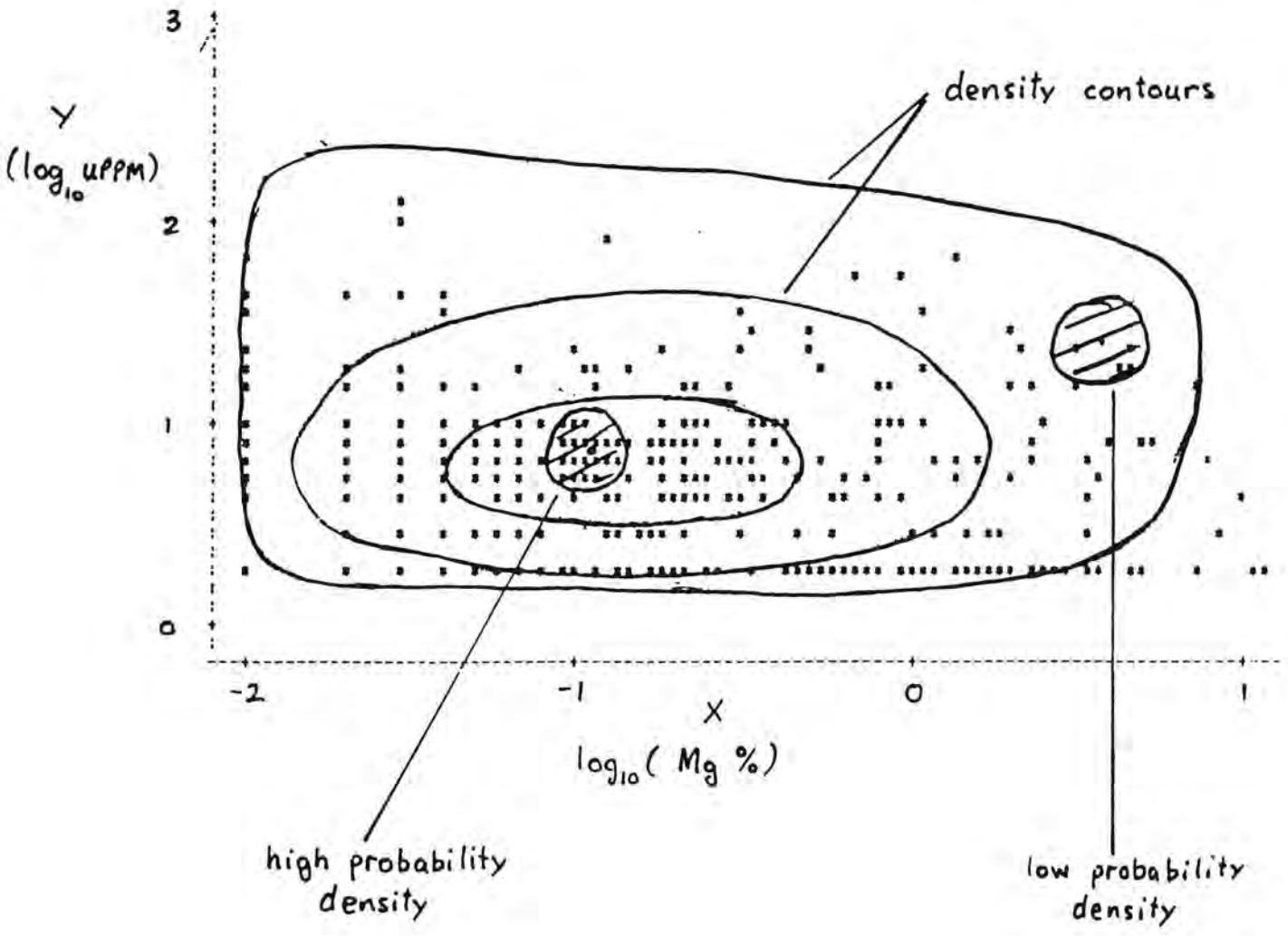
$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\tfrac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

square of standardised distance from mean

To understand the multivariate normal the idea of a "multivariate probability density" is worth knowing. In the bivariate case,

$f(x,y)$=population proportion of points per unit xy-area

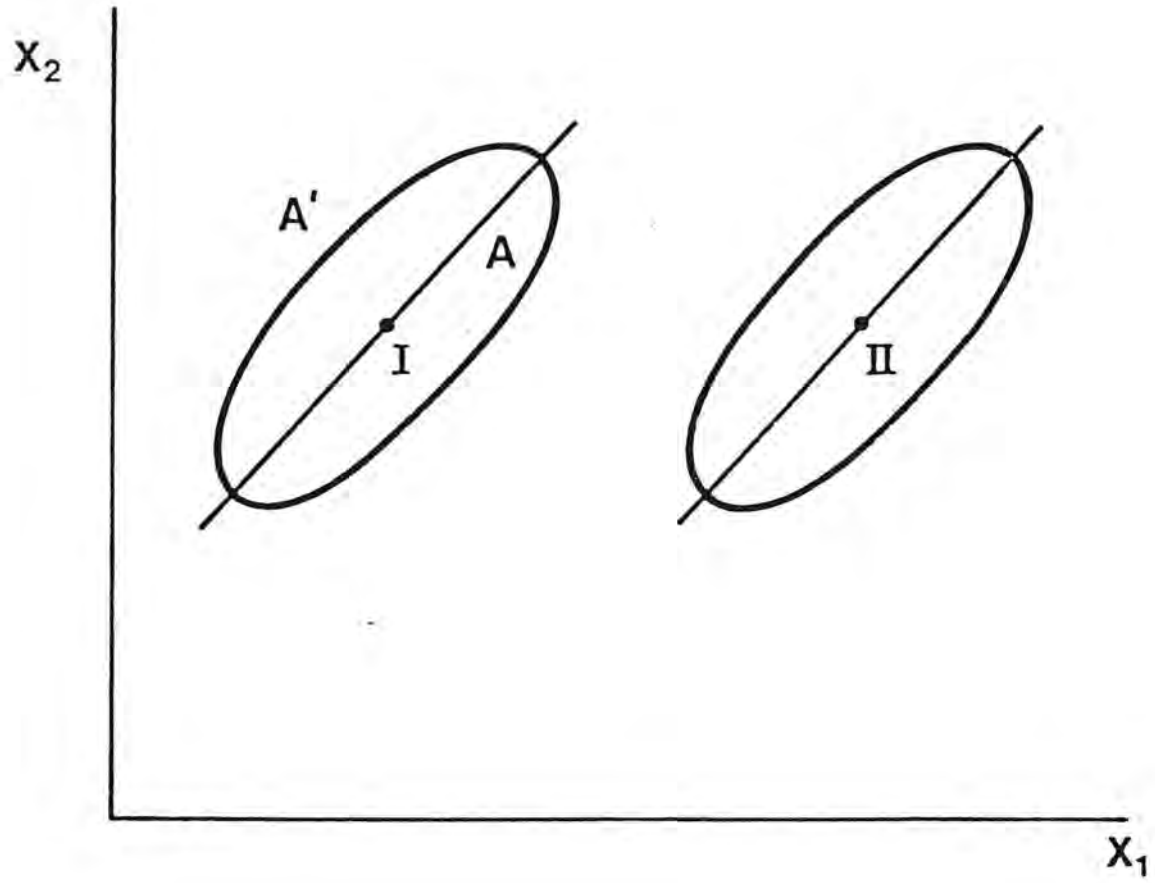
$$\approx \frac{\left(\dfrac{\text{no. sample pts in a small xy-area}}{\text{total no. of sample points}}\right)}{\text{area of small xy-region}}$$

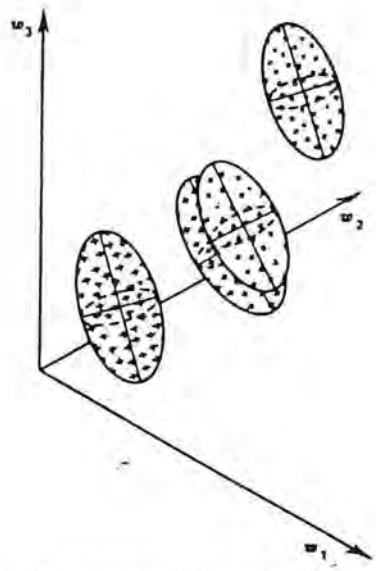


In more than two dimensions the idea remains the same:

f(x1,x2,x3 ...,xk) = prop'n pop'n in small neighbourhood
volume of neighbourhood

The multivariate normal model can be characterised by the shape of its probability density contours. They are always ellipsoidal.





Four ellipsoids representing clouds of observations

For a multivariate normal population all the probability density ellipsoids have the same straight line axes:

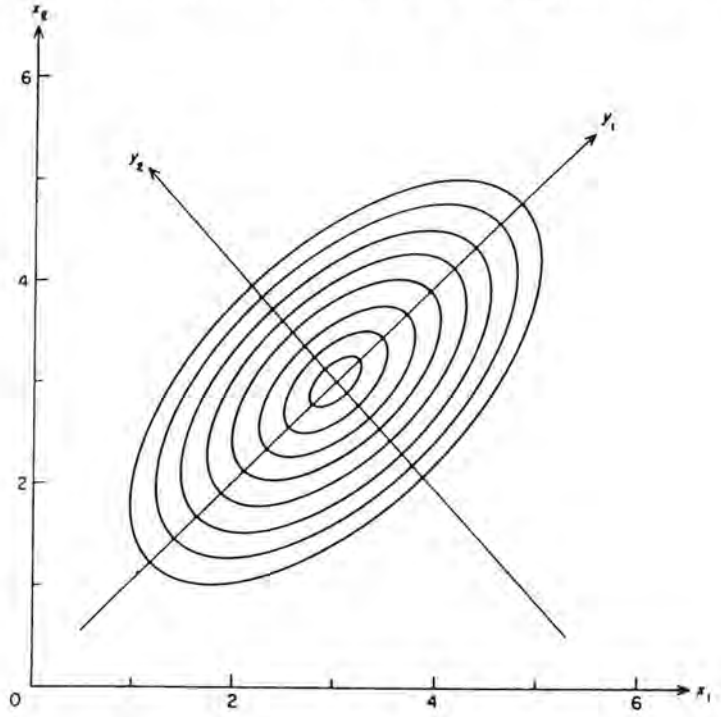


Figure 2.5.1 Ellipses of equal concentration for the bivariate normal distribution, showing the principal components y_1 *and* y_2*, where* $\mu' = (3, 3)$, $\sigma_{11} = 3$, $\sigma_{12} = 1$, $\sigma_{22} = 3$.

As a consequence of the above, all associations between variables are "linear":

Best predictor of X2 from X1 = a0 + a1.X1

Best predictor of X3 from X1, X2 = a0 + a1.X1 + a2.X2

A good way to pictorially represent a bi- or tri-variate normal distribution is by drawing the probability density contour (or surface) which includes 95% of the relevant population.

Mahalanobis Distance

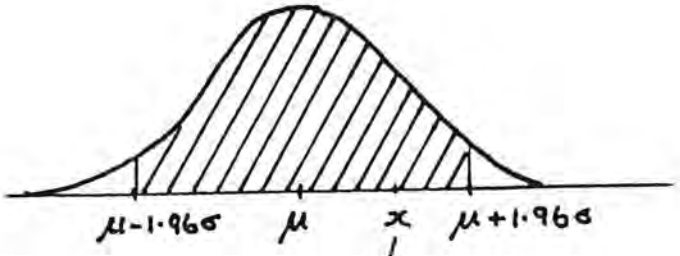
In the one dimensional case, the Mahalanobis distance of value X from the population mean is just:

$$D = |(X - \mu)/\sigma|$$

(ordinary distance scaled to adjust for the X scatter implied by σ).

The 95% probability interval with equal probability densities at its end points is always given by:

$$D < 1.96$$



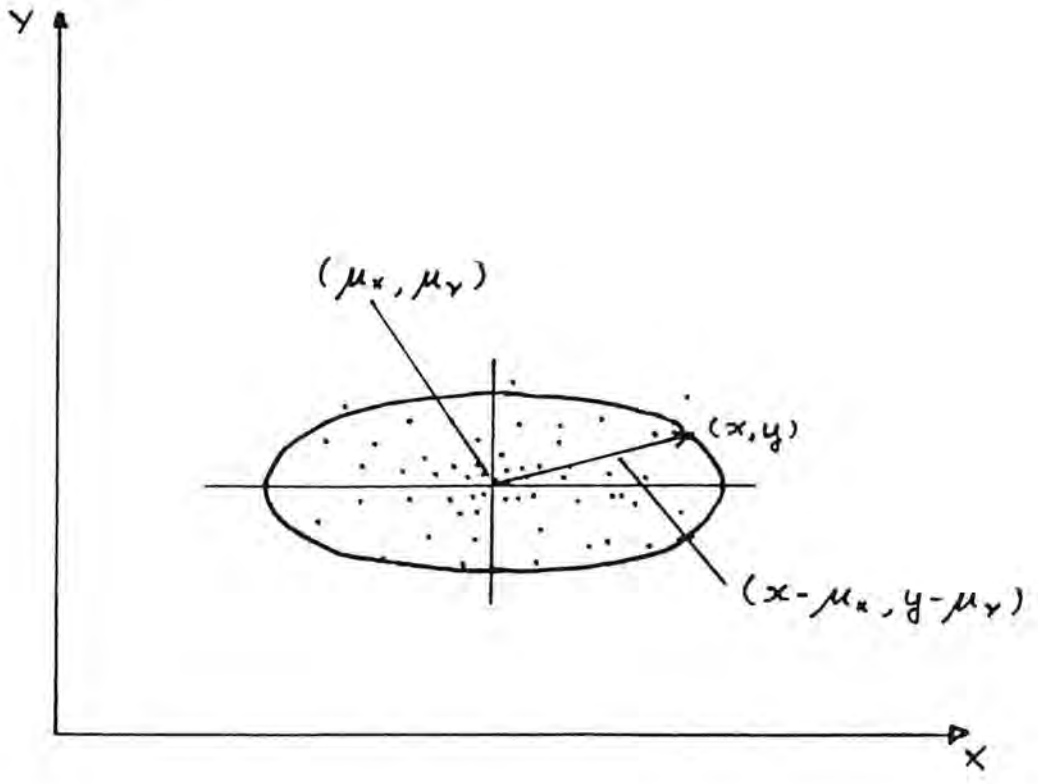
$$D = \left| \frac{x-\mu}{\sigma} \right|$$

= number of standard deviations from the mean

< 1.96

Simplest Bivariate Case - Uncorrelated X,Y

In this case, the probability ellipses have axes parallel to the X,Y axes (no relation between X and Y values).



The probability density is:

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} \cdot \exp\left(-\tfrac{1}{2}\left(\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right)\right)$$

$$= \text{constant}.\exp(-D^2/2),$$

where the Mahalanobis distance

$$D = \sqrt{\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2}$$

is basically ordinary Euclidean distance standardised to allow for the scale of X and Y scatter implied by σ_x and σ_y .

Because the probability density is a function of D alone, the probability density ellipses are given by equations of the form

$$D = c$$

i.e. $$\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2 = c^2$$

If σ_x and σ_y are both 1 (with zero correlation) then D is just ordinary Euclidean distance and the probability density contours become circles.

In two dimensions, the 95% probability ellipse is always given by:

$$D^2 = 5.99 \quad (\text{c.f. } 1.96^2 \text{ for 1 dim'n})$$

The 50% probability ellipse in two dimensions is always given by:

$$D^2 = 1.39.$$

In fact for any specified probability p, the 100p% probability ellipse will always be expressible as:

$$D = d_p ,$$

where d_p depends only on p (not any of the distribution parameters).

This is the meaning of the standardisation implicit in Mahalanobis distances.

The General Case

Always have:

(a) $f(x1,x2, \ldots xk) = \text{constant}.\exp(-D^2/2)$,

where the constant is determined by variances and covariances (high spread means lower densities), and D is the Mahalanobis distance of $(x1,x2, \ldots xk)$ from $(\mu1,\mu2,\ldots\mu k)$.

(b) D^2 is standardised for the variances and covariances of the variables. e.g. when all variables are uncorrelated,

$$D^2 = \left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 + \ldots + \left(\frac{x_k-\mu_k}{\sigma_k}\right)^2 .$$

Mathematical Aside:

In matrix terms, D^2 equals

$$(x_1-\mu_1, \ldots, x_k-\mu_k) \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1k} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2k} \\ \vdots & & \ddots & \\ \sigma_{k1} & \sigma_{k2} & & \sigma_{kk} \end{pmatrix}^{-1} \begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \\ \vdots \\ x_k-\mu_k \end{pmatrix}$$

(inverse)

(variance-covariance matrix V)

(or in shorthand:

$$(\underset{\sim}{x}-\underset{\sim}{\mu})^t V^{-1} (\underset{\sim}{x}-\underset{\sim}{\mu}) .$$

Note the analogy with

$$(x-\mu)(\sigma^2)^{-1}(x-\mu) = ((x-\mu)/\sigma)^2 .$$

(c) For each k (no. of variates), the probability ellipses containing 5,10,15,20 ... ,90,95,99, .. percent of the population

are all ellipsoids of equal Mahalanobis distance d, with the distances for given percentages p being fully determined by p and k.

e.g.:

Example Mahalanobis distances containing specified population proportions

k \ p	.50	.75	.90	.95	.99
1	0.67	1.15	1.65	1.96	2.57
2	1.18	1.66	2.15	2.45	3.03
3	1.54	2.03	2.50	2.79	3.36
4	1.83	2.32	2.79	3.08	3.65
5	2.09	2.57	3.04	3.33	3.89

N.B. These are the square roots of the 100p% percentile of the chi-squared distribution with k "degrees of freedom".

Thus the standardisation implicit in Mahalanobis distances means that probability ellipsoids can be worked out from standard tables (or percentile functions in statistical packages).

Mahalonobis Distances in Practice?

They are important conceptually. What good are they in practice?

(a) Calculating Them

If you have a group of observations on X1, ... Xk which might've come from a single multivariate normal population (after some transformation?), then you can do:

(i) Estimate each variable's mean and variance in the usual way.

(ii) Estimate covariances and variances as explained in section 5:

$$sij = \frac{\text{sum of } (xi-\bar{x}).(yi-\bar{y})}{(n-1).}$$

(iii) Estimate individual D's, using:

$$\hat{D}^2 = (\underset{\sim}{x} - \underset{\sim}{\bar{x}})^1 S^{-1} (\underset{\sim}{x} - \underset{\sim}{\bar{x}}) .$$

This would generally require a bit of programming (not a standard item in statistical packages). For instance, in SAS, a combination of PROCS CORR and MATRIX could be used (SAS PROC CANDISC will output Mahalanobis distances in a discriminant analysis situation; PROC DISCRIM is based on them).

stat properties of D

(b) Using them to test for the adequacy of the normal assumptions

A collection $D1^2, D2^2, \ldots Dn^2$ of Mahalanobis distances^2 estimated as above should be approximately chi-squared with k degrees of freedom (the approximation comes in because $\bar{x}$ and S must be used in place of the unknown population mean and covariance matrix).

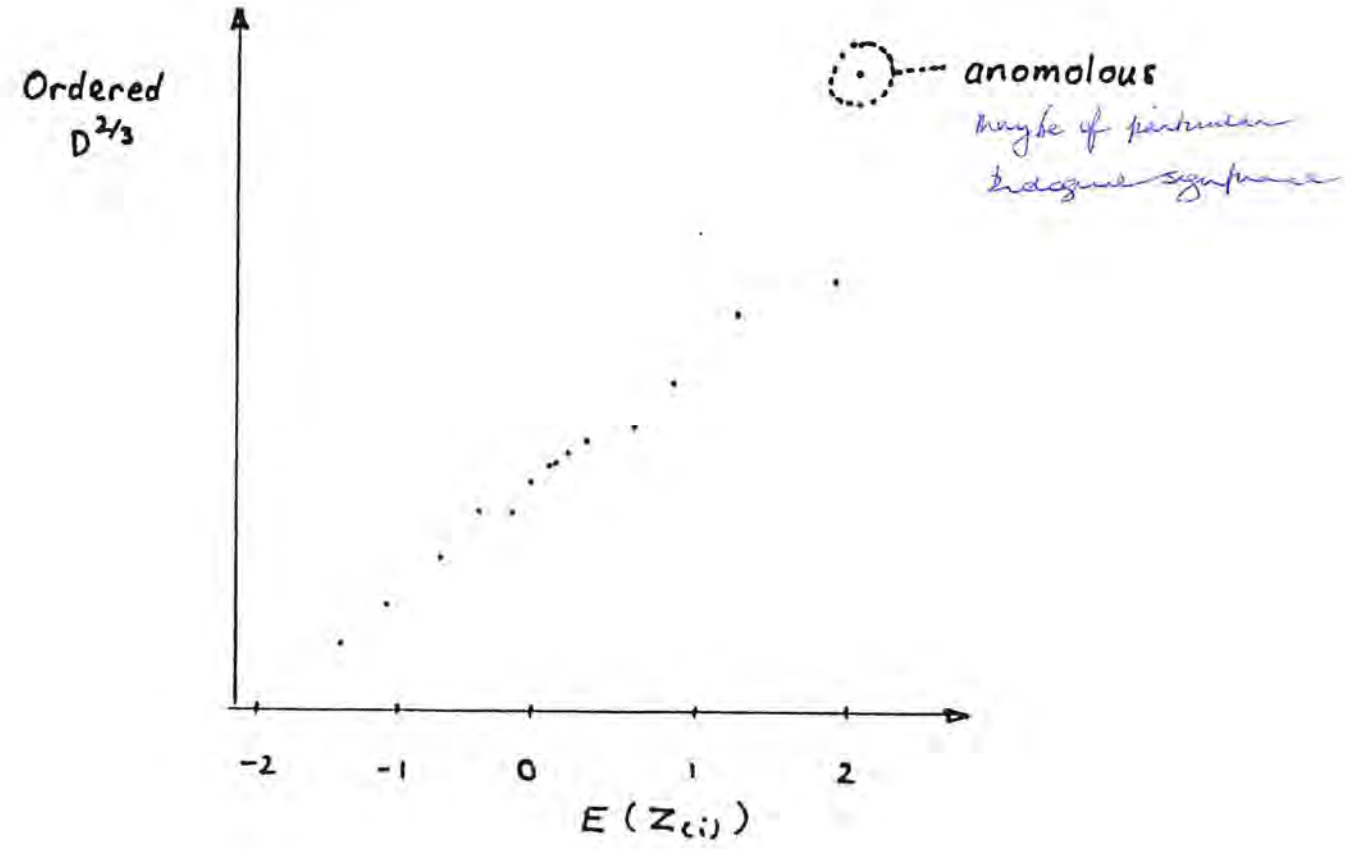
A chi-squared probability plot (ordered D^2 s versus expected ordered chi-squares) should therefore be approximately linear under multivariate normal assumptions. A graphical test for normality can therefore be done by inspecting such a plot.

A simpler, and in some ways better idea, is to use the fact that the cube roots of the Mahalanobis distances should be approximately normally distributed. This can be checked using a standard normal QQ plot.

N.B. These tests only work on the distance structure of the data. Other tests have been worked out for angles made with radii drawn from the mean point (see the book by GNANADESIKAN for details).

(c) Checking for anomolous values

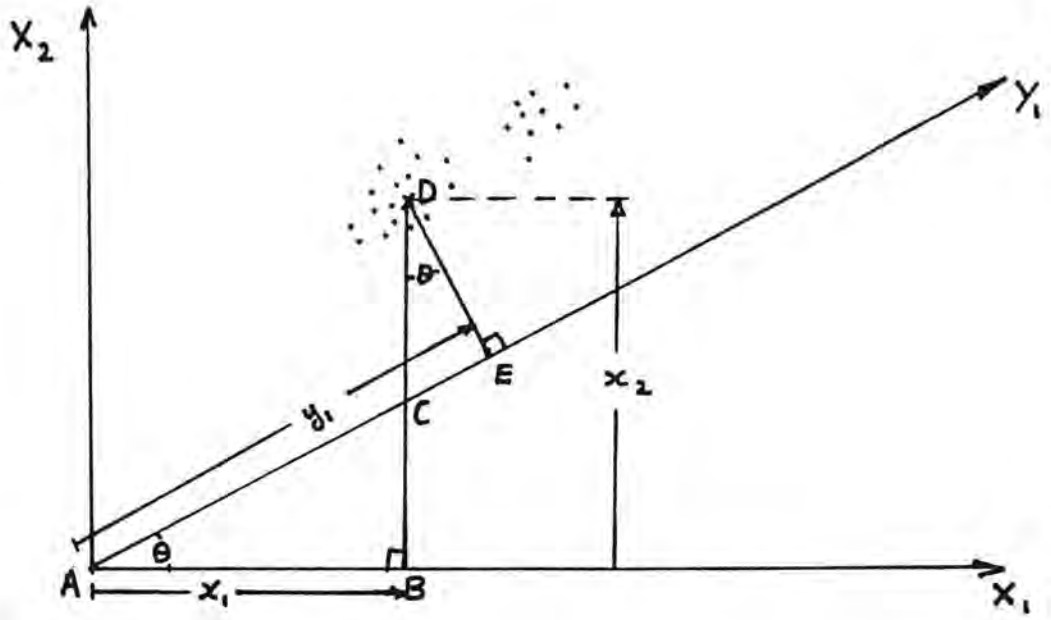
Outlying values (with distances larger than is likely from the mean point) will stick out on QQ plots done as above:



7. CHANGES OF COORDINATES; PRINCIPAL COMPONENTS

A. Linear Changes of Coordinates

Consider:



By a bit of trigonometry, can work out that:

$$y1 = \cos\theta\, x1 + \sin\theta\, x2.$$

[$y1 = AC + CE$
$= AC + CD\sin\theta$
$= AC + (x2 - BC)\sin\theta$
$= AC + x2\sin\theta - AC\sin^2\theta$
$= AC\cos^2\theta + x2\sin$
$= AC\cos\theta.\cos\theta + x2\sin\theta$
$= x1\cos\theta + x2\sin\theta$]

Thus a new coordinate axis rotated from the X1 axis produces coordinates given by an expression of the form:

$$y1 = a1.x1 + a2.x2$$

(where, in the example, a1 and a2 are $\cos\theta$ and $\sin\theta$ respectively).

In general, any expression of the form

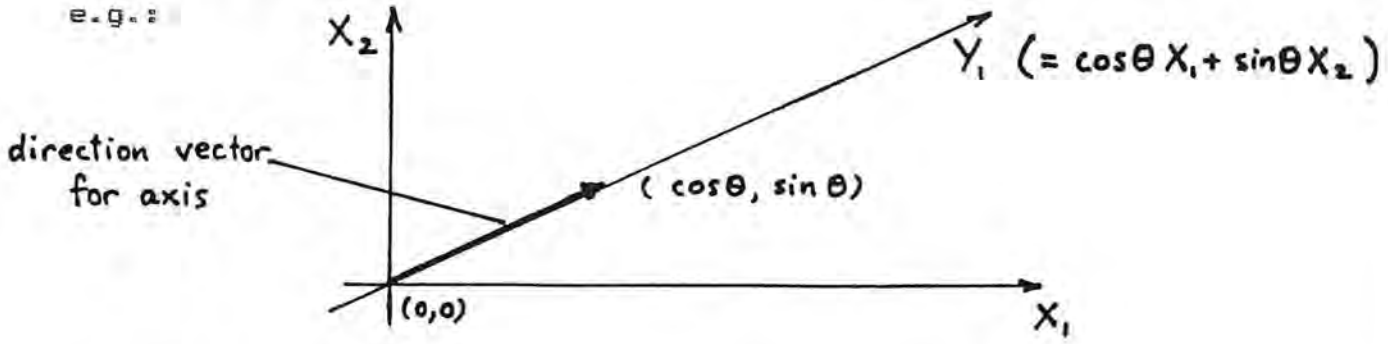
$$yi = a1.x1 + a2.x2 + \ldots + ak.xk$$

(termed a "linear combination" of x1,x2, ... xk) is a (possibly rescaled) measurement along a rotated coordinate axis in k-dimensional space (obtained from dropping a perpendicular for (x1,x2, ... xk) to that axis).

The axis fulfilling the above is simply that in the direction of the vector joining the origin (0,0,... 0) to

(a1,a2,...ak).

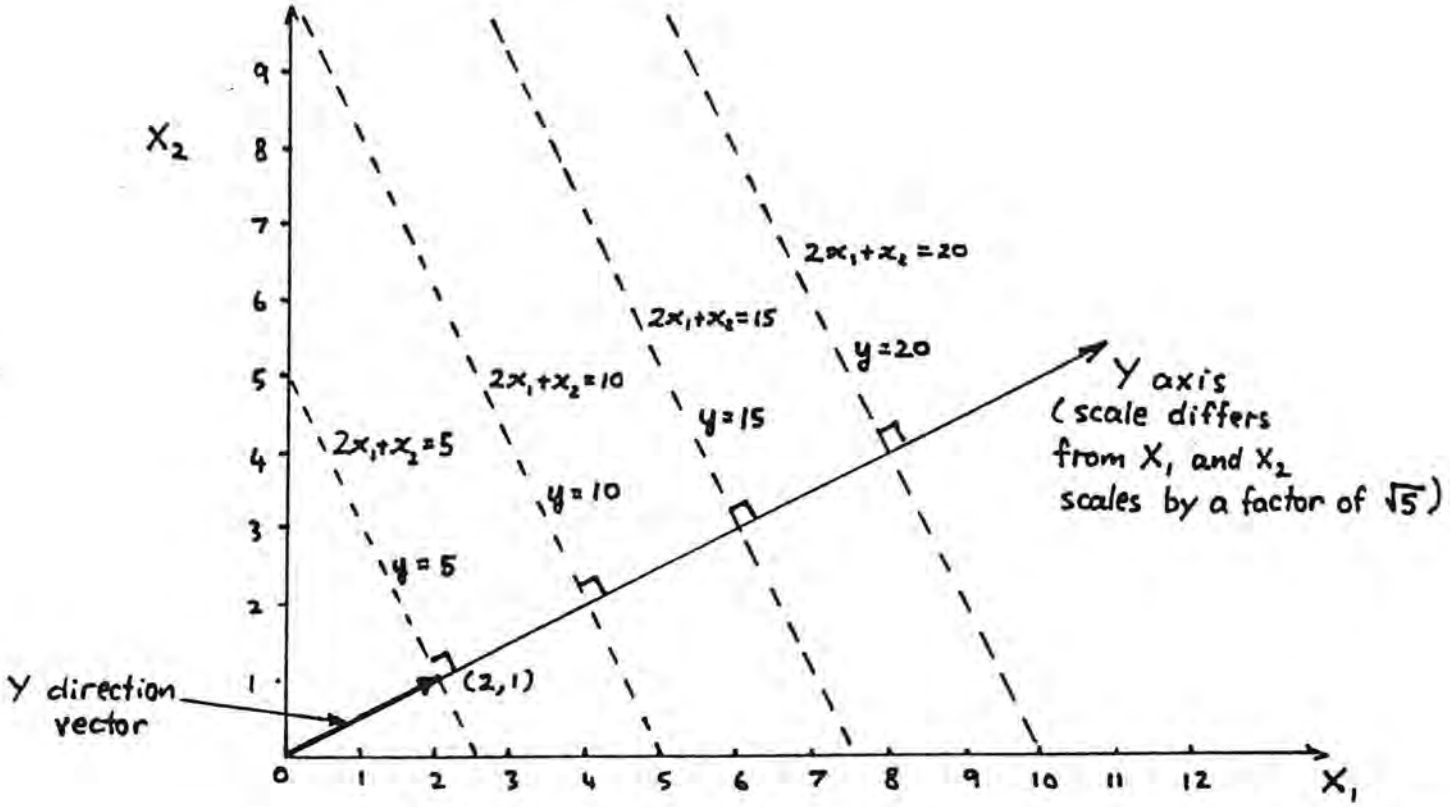
e.g.:



A concrete example:

Consider y = 2x1 + x2

$(= \sqrt{5} \cdot (\frac{2}{\sqrt{5}} x_1 + \frac{1}{\sqrt{5}} x_2)$)



Much of classical multivariate analysis, particularly that involving the multivariate normal distribution, revolves around the identification of new coordinate axes along which important data features are highlighted.

As we have just seen, this is equivalent to looking for linear combinations of the original variables (possibly transformed) which best capture a particular data feature.

e.g. for a morphometric data set,

Y = 4.log(length 1) - 3.sqrt(length 2) - 2.(length 3)

might conceivably be found to be best at separating two closely related species.

i.e. the axis in the (4,-3,-2) direction in (log(length 1), sqrt(length 2), length 3) "space" may be such that perpendicular projections onto it keep the groups separate.

By appropriately choosing r linear combinations (axes)

$$
\begin{aligned}
Y1 &= a11.X1 + a12.X2 + \ldots + a1k.Xk \\
Y2 &= a21.X1 + a22.X2 + \ldots + a2k.Xk \\
Y3 &= a31.X1 + a32.X2 + \ldots + a3k.Xk \\
&\vdots \\
Yr &= ar1.X1 + ar2.X2 + \ldots + ark.Xk
\end{aligned}
$$

the data can be represented in a lower number of dimensions than originally. Usually hope to get r down to two or three without losing the information of importance in the data.

Two methods of axis choice to be discussed are:

- Principal Components (e.g. from SAS PROC PRINCOMP)

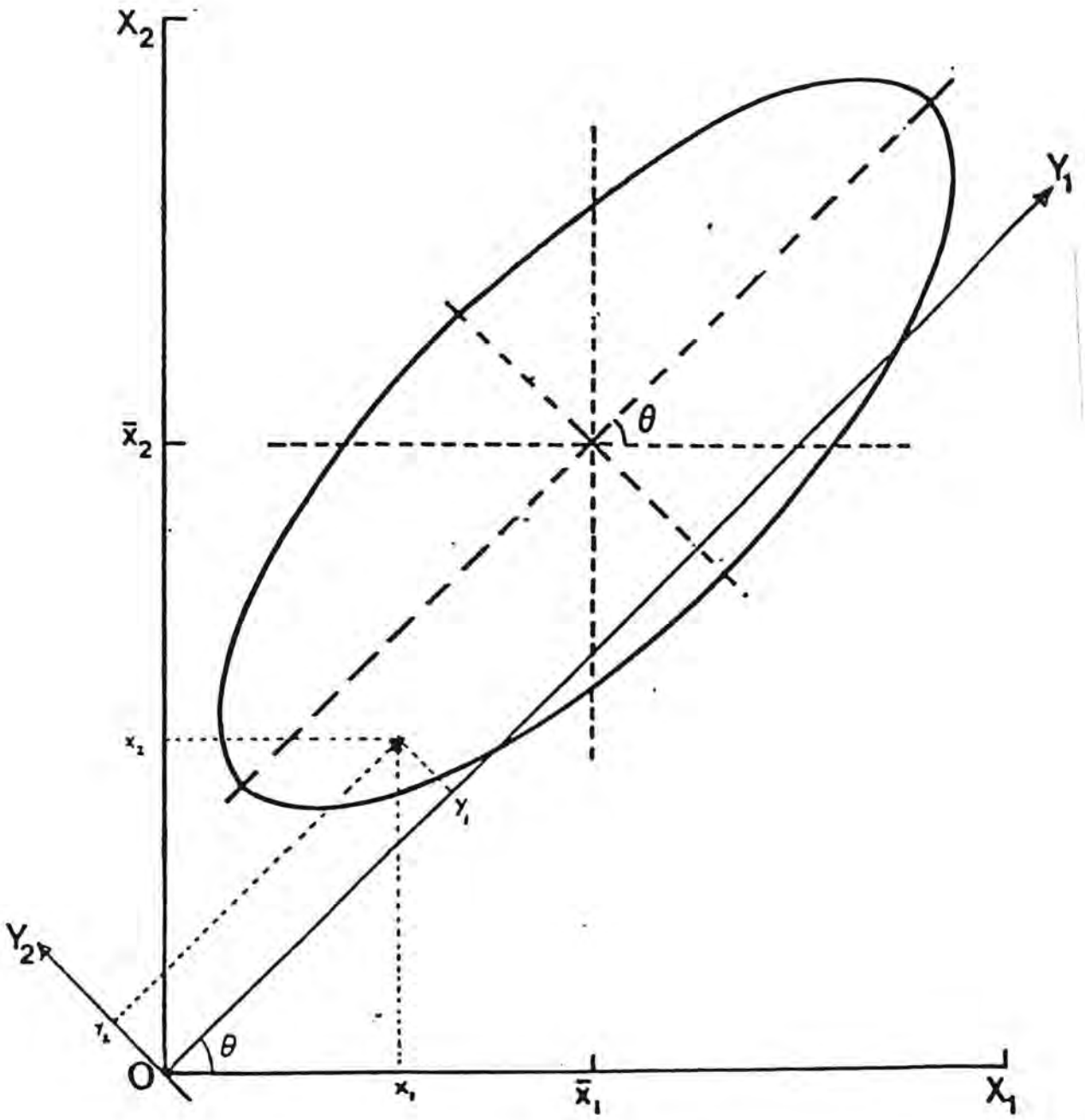
- Canonical Variates (e.g. from SAS PROC CANDISC)

B. Principal Components

In two dimensions, the idea is:

- choose Y1 so that the variance in the Y1 values is maximised

- choose Y2 orthogonal (perpendicular) to Y1.



Points to Note:

(a) As well as being orthogonal (as axes in two dimensional space), Y1 and Y2 (PRIN1 and PRIN2 in SAS PROC PRINCOMP output) are also statistically uncorrelated (as random variables).

(b) When (X1,X2) have a bivariate normal distribution then (Y1,Y2) align themselves with the principal axes of the probability ellipses, in decreasing order of axis length. They are the "natural" axes from which to view the normal model.

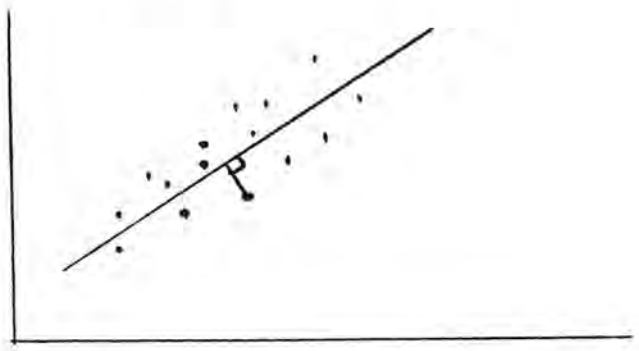
(c) Principal components are obtained in statistical packages by matrix algebra. The coefficient vectors (a11,a12) and (a21,a22) are "eigenvectors" of the sample variance-covariance matrix

$$S = \begin{pmatrix} s_1^2 & s_{12} \\ s_{21} & s_2^2 \end{pmatrix}$$

(d) Usually, these eigenvectors are constrained to have unit length ($a1^2 + a2^2 = 1$), so that the Y1 and Y2 axis measurements are simple perpendicular perpendicular projections, without any rescaling.

(e) With the above constraint, the variances of the uncorrelated variables are equal to the so-called "eigenvalues" (the name by which they are often labelled in output from statistical packages e.g. SAS PROC PRINCOMP).

(f) The line through the mean point parallel to the Y1 axis is the "line of best perpendicular fit", giving the minimum sum of squares of perpendicular distances:



(g) It often makes sense to do principal components on standardised versions of the variates (otherwise the variables of largest variance will tend to dominate the principal components).

(h) In practice, principal components on standardised variates can be obtained by submitting the correlation matrix to the eigenvalue analysis, rather than the covariance matrix (this is the PROC PRINCOMP option).

More Than Two Dimensions

- Start as in the two dimensional case, choosing Y1 of maximum variance, then Y2 of maximum variance orthogonal to Y1.

- Then choose Y3 of maximum variance orthogonal to the Y1,Y2 plane.

- Choose Y4 of maximum variance orthogonal to Y1,Y2 and Y3.

.
.
.

- Choose Yk of maximum variance orthogonal to Y1,Y2 ... Yk.

Points (a)-(h) for the two dimensional case apply in general. Point (f) can be expanded:

- The plane passing through the mean point and parallel to both Y1 and Y2 is the plane of best perpendicular fit.

- In general, the r-dimensional subspace through (x1,x2,...xk) paralel to all of Y1,Y2,...Yr is the r-dimensional subspace of best perpendicular fit.

In choosing the number of dimensions r needed to retain the main features of the data, the main tool is the decreasing sequence of eigenvalues (variances) associated with the principal components. Their sum can be shown mathematically to equal the sum of the variances of the original variables X1,X2...Xk (the "trace" of the variance-covariance matrix).

The relative importance of each principal component can be judged by expressing it's variance as a percentage of the total over all components - the "percentage of variance explained". Cumulative percentages are often also calculated. Rules such as requiring 90% of variance explained are sometimes used to choose r so that the first r principal can be considered to contain all worthwhile information.

When the data is multivariate normal, statistical tests can be conducted to aid decision making. e.g. a test of the hypothesis that all eignenvalues past the rth are equal (so if one is judged unimportant, so must the rest be). See MARDIA, KENT & BIBBY for details.

Principal components can be used simply as an exploratory tool, without any necessity for model assumptions (though transformations to, at least roughly, remove any gross distributional asymmetries would still be recommended).

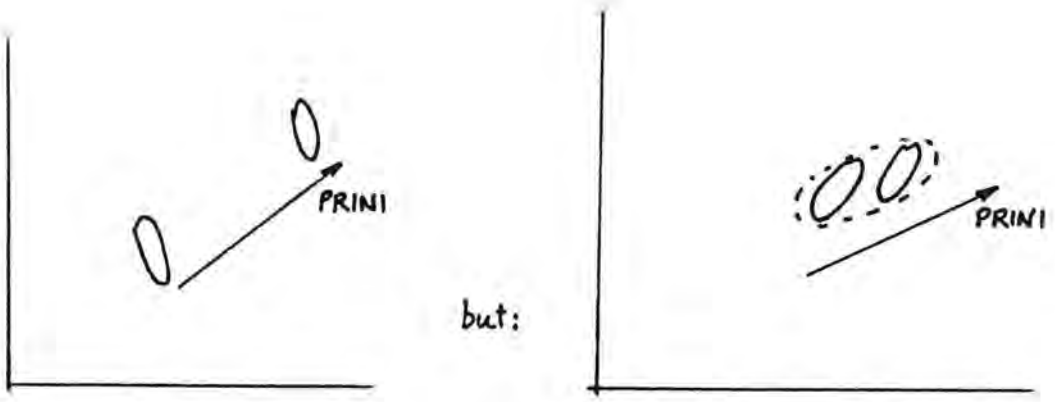
Size and Shape

When the original data consists of sizes of various parts of individuals (e.g. total lateral shoot length, trunk circumference

and height for trees), the first principal component often gives all the original variables roughly equal positive weights (coefficients). This means it can be interpreted as some kind of overall size measure. In this case the other principal components must have both negative and positive weights (a consequence of orthogonality), and therefore are contrasts between weighted sums over different subsets of the original variables. These components sometimes have sensible interpretations as as "shape" measures.

Discovering Groups

Sometimes principal components may help , but not always:



CASE STUDY THREE - PRINCIPAL COMPONENT ANALYSIS
(from the SAS User's Guide: Stat's)

The data consists of crime rates per 100,000 people in seven categories, for 50 U.S states.

The default form of principal component analysis in SAS uses does an eigenanalysis of the correlation matrix (the covariance matrix of the standardised variates). The eigenvectors provide coefficients for standardised variates.

The first principal component measures overall crime rate. The second can be interpreted as measuring the preponderance of property crime over violent crime.

The plot of PRIN2 vs PRIN1 with points labelled (an example of an ordination) provides interesting information on regional trends (e.g high preponderance of violent crime in S.E. states, vice-versa in New England states, low overall crime rates in the Dakotas, high in California and Nevada).

Crime Rates: Example 2

The data below give crime rates per 100,000 people in seven categories for each of the fifty states. Since there are seven variables, it is impossible to plot all the variables simultaneously. Principal components can be used to summarize the data in two or three dimensions and help to visualize the data.

```
DATA CRIME;
   TITLE 'CRIME RATES PER 100,000 POPULATION BY STATE';
   INPUT STATE $1-15 MURDER RAPE ROBBERY ASSAULT BURGLARY LARCENY AUTO;
   CARDS;
ALABAMA         14.2 25.2  96.8 278.3 1135.5 1881.9 280.7
ALASKA          10.8 51.6  96.8 284.0 1331.7 3369.8 753.3
ARIZONA          9.5 34.2 138.2 312.3 2346.1 4467.4 439.5
ARKANSAS         8.8 27.6  83.2 203.4  972.6 1862.1 183.4
CALIFORNIA      11.5 49.4 287.0 358.0 2139.4 3499.8 663.5
COLORADO         6.3 42.0 170.7 292.9 1935.2 3903.2 477.1
CONNECTICUT      4.2 16.8 129.5 131.8 1346.0 2620.7 593.2
DELAWARE         6.0 24.9 157.0 194.2 1682.6 3678.4 467.0
FLORIDA         10.2 39.6 187.9 449.1 1859.9 3840.5 351.4
GEORGIA         11.7 31.1 140.5 256.5 1351.1 2170.2 297.9
HAWAII           7.2 25.5 128.0  64.1 1911.5 3920.4 489.4
IDAHO            5.5 19.4  39.6 172.5 1050.8 2599.6 237.6
ILLINOIS         9.9 21.8 211.3 209.0 1085.0 2828.5 528.6
INDIANA          7.4 26.5 123.2 153.5 1086.2 2498.7 377.4
IOWA             2.3 10.6  41.2  89.8  812.5 2685.1 219.9
KANSAS           6.6 22.0 100.7 180.5 1270.4 2739.3 244.3
KENTUCKY        10.1 19.1  81.1 123.3  872.2 1662.1 245.4
LOUISIANA       15.5 30.9 142.9 335.5 1165.5 2469.9 337.7

VERMONT          1.4 15.9  30.8 101.2 1348.2 2201.0 265.2
VIRGINIA         9.0 23.3  92.1 165.7  986.2 2521.2 226.7
WASHINGTON       4.3 39.6 106.2 224.8 1605.6 3386.9 360.3
WEST VIRGINIA    6.0 13.2  42.2  90.9  597.4 1341.7 163.3
WISCONSIN        2.8 12.9  52.2  63.7  846.9 2614.2 220.7
WYOMING          5.4 21.9  39.7 173.9  811.6 2772.2 282.0
;
PROC PRINCOMP OUT=CRIMCOMP;
```

Output 28.4 Results of Principal Component Analysis: PROC PRINCOMP

CRIME RATES PER 100,000 POPULATION BY STATE 1

PRINCIPAL COMPONENT ANALYSIS

50 OBSERVATIONS
7 VARIABLES

SIMPLE STATISTICS

	MURDER	RAPE	ROBBERY	ASSAULT	BURGLARY	LARCENY	AUTO
MEAN	7.44400	25.7340	124.092	211.300	1291.90	2671.29	377.526
ST DEV	3.86677	10.7596	88.349	100.253	432.46	725.91	193.394

CORRELATIONS

	MURDER	RAPE	ROBBERY	ASSAULT	BURGLARY	LARCENY	AUTO
MURDER	1.0000	0.6012	0.4837	0.6486	0.3858	0.1019	0.0688
RAPE	0.6012	1.0000	0.5919	0.7403	0.7121	0.6140	0.3489
ROBBERY	0.4837	0.5919	1.0000	0.5571	0.6372	0.4467	0.5907
ASSAULT	0.6486	0.7403	0.5571	1.0000	0.6229	0.4044	0.2758
BURGLARY	0.3858	0.7121	0.6372	0.6229	1.0000	0.7921	0.5580
LARCENY	0.1019	0.6140	0.4467	0.4044	0.7921	1.0000	0.4442
AUTO	0.0688	0.3489	0.5907	0.2758	0.5580	0.4442	1.0000

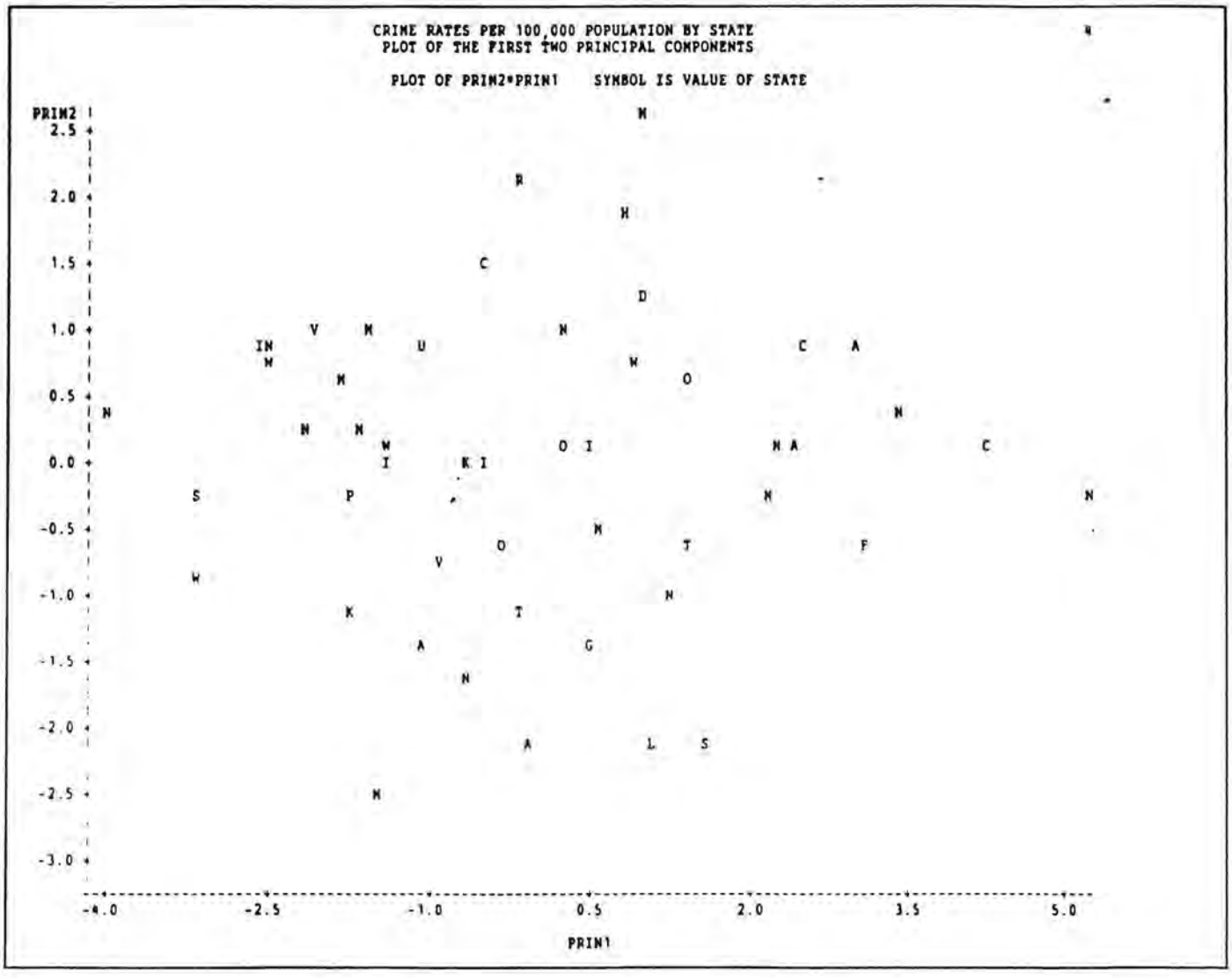
	EIGENVALUE	DIFFERENCE	PROPORTION	CUMULATIVE
PRIN1	4.11496	2.87624	0.587851	0.58785
PRIN2	1.23872	0.51291	0.176960	0.76481
PRIN3	0.72582	0.40938	0.103688	0.86850
PRIN4	0.31643	0.05846	0.045205	0.91370
PRIN5	0.25797	0.03593	0.036853	0.95056
PRIN6	0.22204	0.09798	0.031720	0.98228
PRIN7	0.12406	.	0.017722	1.00000

EIGENVECTORS

	PRIN1	PRIN2	PRIN3	PRIN4	PRIN5	PRIN6	PRIN7
MURDER	0.300279	-.629174	0.178245	-.232114	0.538123	0.259117	0.267593
RAPE	0.431759	-.169435	-.244198	0.062216	0.188471	-.773271	-.296485
ROBBERY	0.396875	0.042247	0.495861	-.557989	-.519977	-.114385	-.003903
ASSAULT	0.396652	-.343528	-.069510	0.629804	-.506651	0.172363	0.191745
BURGLARY	0.440157	0.203341	-.209895	-.057555	0.101033	0.535987	-.648117
LARCENY	0.357360	0.402319	-.539231	-.234890	0.030099	0.039406	0.601690
AUTO	0.295177	0.502421	0.568384	0.419238	0.369753	-.057298	0.147046

```
PROC PLOT;
   PLOT PRIN2*PRIN1=STATE;
   TITLE2 'PLOT OF THE FIRST TWO PRINCIPAL COMPONENTS';
PROC PLOT;
   PLOT PRIN3*PRIN1=STATE;
   TITLE2 'PLOT OF THE FIRST AND THIRD PRINCIPAL COMPONENTS';
```

Output 28.6 Plots of Principal Components: PROC PLOT



8. SAMPLING DISTRIBUTIONS

Given simple random sampling resulting in n independent measurements X1, X2 ... Xn what can we expect from $\bar{X}$ and s as estimators of the common population mean and variance?

How accurate are they? Do they have "bias" to one side or other of the true values? If the random sampling could be repeated say 100 times how scattered would the 100 $\bar{X}$ and s values be? How precise (repeatable)would they be?

These are all questions about the "sampling distributions" of $\bar{X}$ and s . The distributions that would be observed if the sampling could be repeated indefinitely? The distributions for the population of sample values?

If the common population distribution for the Xi's is normal, then these questions can be answered precisely.

Even if it is not, certain basic facts still hold:

The sampling distribution for $\bar{X}$ has mean

$$E(\bar{X}) = \mu$$

and variance

$$var(\bar{X}) = \sigma^2/n$$

i.e. standard deviation = $\sigma / \sqrt{n}$

Example:

Suppose the underlying distribution for the Xi's is uniform on [0,1] (i.e. the probability density f(x)=1 for x between 0 and 1). Consider the means of independent random samples of n=4 such numbers.

e.g. 1st time:

$\bar{x}$ = (0.022+0.541+0.165+0.678)/4
= 0.351

Repeat 20 times:

Sample				$\bar{x}$
0.022	0.541	0.165	0.678	0.351
0.020	0.266	0.138	0.274	0.774
0.644	0.114	0.805	0.119	0.420
0.824	0.741	0.344	0.925	0.708
0.920	0.164	0.105	0.621	0.452
0.231	0.800	0.426	0.862	0.580
0.742	0.903	0.545	0.981	0.729
0.873	0.823	0.641	0.502	0.710
0.821	0.588	0.232	0.538	0.545
0.632	0.588	0.846	0.902	0.742
0.641	0.276	0.423	0.914	0.624
0.989	0.185	0.173	0.236	0.396
0.793	0.887	0.960	0.961	0.900
0.718	0.439	0.364	0.848	0.592
0.188	0.959	0.555	0.707	0.602
0.721	0.475	0.525	0.241	0.490
0.288	0.119	0.008	0.306	0.180
0.321	0.537	0.573	0.823	0.563
0.163	0.367	0.303	0.910	0.436
0.173	0.784	0.240	0.192	0.347

mean of $\bar{x}$'s = 0.528

standard deviation of $\bar{x}$'s
= 0.187

Stem and leaf plot for $\bar{x}$'s:

```
0.0
0.1  7 8
0.2
0.3  5 5
0.4  0 2 4 5 9
0.5  4 6 8 9
0.6  0 2
0.7  1 1 3 4
0.8
0.9  0
```

How well does the standard deviation of these $\bar{x}$'s match the theoretical value $\sigma/\sqrt{n}$? The population standard deviation for a uniform distribution can be worked out mathematically (by calculating an integral). It is 1/ 12 (0.28).

So the sampling distribution for x's from samples of size 4 is:

$0.28/\sqrt{4} = 0.14$

This compares well enough with the estimate obtained from

our 20 example $\bar{x}$'s. If we'd generated more $\bar{x}$'s in the same way, the agreement would've been closer.

Note also the reasonable agreement between the mean 0.528 and the assertion that the expected value of $\bar{X}$ is 0.5.

The following exhibit from the "Minitab" handbook uses computer random sampling from a normal distribution (μ=42, σ=6) to illustrate the sampling distribution of means of independent random samples of size 9. The mean and standard deviation of the 100 $\bar{x}$'s are even closer to the theoretical values for the sampling distribution ($E(\bar{X})$=42, st. dev'n $\bar{X} = 6/\sqrt{9} = 2$).

```
RANDOM 100 C1-C9;
  NORMAL MU=42 SIGMA=6.
RMEAN C1-C9 INTO C10
HISTOGRAM  C10

Histogram of C10   N = 100

Midpoint   Count
      36       2  **
      39       4  ****
      40      21  *********************
      41      12  ************
      42      18  ******************
      43      20  ********************
      44       9  *********
      45      11  ***********
      46       1  *
      47       2  **

DESCRIBE C10

               N      MEAN    MEDIAN    TRMEAN     STDEV   SEMEAN
C10          100    42.101    41.852    42.069     1.936    0.194

               MIN       MAX        Q1        Q3
C10         36.114    47.427    40.328    43.352
```

Bias

For $\bar{X}$ considered as an estimator of μ, the bias is defined: $E(\bar{X})-\mu$.

In general, if some formula calculated on observed random values is used as an estimator of some population parameter, then the bias is defined as

E(estimator) - true value.

If the bias is zero (as it is for $\bar{X}$) then the estimator is called unbiased. This means that the average of the estimator over sampling occasions equals the population quantity it is trying to estimate. Unbiasedness (at least approximately) is clearly a desirable property.

Other examples of unbiased estimators are:

(a) s^2 as an estimator of σ^2;

(b) the least squares estimators $\hat{a}$, $\hat{b}$ for the slope and intercept when the simple linear regression model $Y = aX+b+error$ applies.

(c) the sample covariance $s_{xy} = \Sigma (X-\bar{X})(Y-\bar{Y}) / (n-1)$ as an

estimator of a population covariance.

N.B. (a) is the reason that n-1 is used in the denominator of s^2 rather than n. It is related to the fact that there are n-1 "degrees of freedom" in the terms $(Xi-\bar{X})$, i=1...n (the sum of these terms always equals zero).

Standard Errors - Precision

When a statistic is used as an estimator of a population parameter, (an estimate of) the standard deviation of the statistic's sampling distribution is often quoted; and referred to as its standard error. In the case when a positive quantity is measured, a relative standard error (S.E./estimate) may be quoted.

Assuming the estimation procedure is accurate (unbiased), then the standard error provides a tolerance within which the true parameter value will lie most of the time.

sampling distribution for $\bar{X}$

Standard errors are a measure of the precision (repeatability) of an estimator.

The t Distribution

This sampling distribution is important quite generally for assessing estimates of "location" parameters (e.g. for a single distribution, or the slope and intercept provided by the least squares fit of straight line regression model) when the random variations in the data are statistically independent and normally distributed.

It applies to quantities of the form $(\hat{a} - a)/s.e.(\hat{a})$ where $\hat{a}$ is the least squares estimate of some coefficient a in linear (in location parameters) regression model (e.g. a coefficient estimated in SAS PROC REG or GLM); and $s.e.(\hat{a})$ is calculated

from a residual sum of squares (divided by its degrees of freedom).

The degrees of freedom of the residual SS alter the form of the t distribution (flatter and wider for small d.f.)

For very large d.f. (so the residual mean square estimates the residual variance very precisely) , the t distribution is very close to the standard normal distribution.

Example:

As an estimator of μ, the mean $\bar{X}$ of a independent random sample has its standard error estimated by $s/\sqrt{n}$.

If the population sampled is normal, then the quantity $(\bar{X}-\mu)/(s/\sqrt{n})$ has a t sampling distribution with n-1 degrees of freedom.

For example, the Minitab handbook gives the following histogram for 100 sample t values with 8 degrees of freedom (the case where the population mean was estimated by means $\bar{x}$ of 9 independent sample values):

```
MSTDEV C1-C9 INTO C11
LET C13 = (C10-42)/(C11/3)
MHISTOGRAM C13

Mistogram of C13   N = 100

Midpoint   Count
      -4       1  *
      -3       0
      -2       4  ****
      -1      28  ****************************
       0      33  *********************************
       1      24  ************************
       2       6  ******
       3       3  ***
       4       1  *

DESCRIBE C13

               N     MEAN   MEDIAN   TRMEAN    STDEV   SEMEAN
C13          100    0.061   -0.070    0.036    1.201    0.120

              MIN      MAX       Q1       Q3
C13        -4.027    4.260   -0.731    0.860
```

This is a wider distribution than the standard normal distribution which applies to $(\bar{X}-\mu)/(\sigma/\sqrt{n})$.

The Chi-Squared Distribution

This the sampling distribution which allows us to make inferences about the variance of a normal population , or the variance of the deviations about a regression model. It also arises as the sampling distribution of popular test statistics for tables of counts, and for tests on Mahalonobis distances and other multivariate quantities.

Let SS denote a sum of squares of random quantities each of which is normally distributed with mean zero and variance .

e.g. $SS = \Sigma (X-\bar{X})^2$
or $SS = \Sigma (Y-\hat{Y})^2$
or $SS = \Sigma (obs-exp)^2 / exp$ ("observed" and "expected" cell counts, under some hypothesis)

Let df denote the number of "degrees of freedom" in SS.

Then the sampling distribution of SS/σ^2 is the "chi-squared" (χ^2) distribution with df degrees of freedom.

Notes:

(a) Starting with n d.f. in X1, ... , Xn , subtracting X takes away one d.f. because of the constraint

$$\Sigma (X-\bar{X}) = 0.$$

(b) With straight line regression residuals, there are two constraints:

$$\Sigma (Y-\hat{Y}) = 0$$
$$\Sigma (Y-\hat{Y})X = 0;$$

so the d.f. are reduced to n-2.

(c) In both cases, the χ^2 variate can be rewritten in the form:

$$\text{d.f. } s^2/\sigma^2$$

(d) If the inverse covariance matrix used in its calculation were known exactly, then the Mahalanobis D^2 discussed in section 6 would have a chi-squared sampling distribution with d.f. equal to the number of variables involved. In practice, the inverse covariance used in calculating D will have been estimated from data and the chi-squared distribution will only be an approximation.

(e) The expected value and variance of a chi-squared distribution are

$$E(\chi^2) = \text{d.f.}$$
$$Var(\chi^2) = 2\text{d.f.}$$

(f) The following are the 95% chi-squared percentiles for d.f.=1..10.

1	2	3	4	5	6	7	8	9	10
3.84	5.99	7.81	9.49	11.1	12.6	14.1	15.5	16.9	18.3

Example Application

Manganese assays on a "standard" sample produced the following results:

78, 90, 100, 95, 102, 85 ppm.

Are these consistent with the assay company's claim that the population standard deviation for lab variability is no more than 5ppm?

Soln:

$(n-1)s^2$ ($\Sigma(x-\bar{x})^2$) = 433.83

$(n-1)s^2/\sigma^2$ = 433.83/25
= 17.35.

Now, for a χ^2 quantity with n-1=5 d.f., the mean value over samples is 5; 11.1 is exceeded in only 5% of samples and 16.7 is exceeded with probability 0.5%. Therefore the probability of obtaining the ratio 17.35 by chance if σ were 5 is very small (even smaller for σ less than 5). The assay company's claim seems implausible.

The F Distribution

This can be defined as the distribution applying to a ratio

$$ (\chi^2_{df_1} /df1) / (\chi^2_{df_2} /df2) $$

where $\chi^2_{df_1}$ and $\chi^2_{df_2}$ are statistically independent chi-squared random quantities.

Examples:

(a) Given independent random samples X1 ... Xm and Y1 ... Yn from separate normal populations

$$ (s_x^2 / \sigma_x^2) / (s_y^2 / \sigma_y^2) $$

has an F distribution with m-1 and n-1 degreees of freedom.

(b) Ratios of "analysis of variance" mean squares.

9. HYPOTHESIS TESTING

Setting:

Desire to test a conservative hypothesis which is judged worthy of being accepted in the absence of good evidence to the contrary.

Call this the "null hypothesis" H0.

The appropriate test depends on the "alternative hypothesis", usually a composite of all the deviations from H0 in the direction worth detecting.

Detection of deviations relies on construction (theory) ad calculation (practice) of an appropriate test statistic.

Requirements of Test Statistic:

(a) should be a numerical measure of a meaningful distance, discrepancy or discordance between data and what would be expected under H0 (e.g. Mahalanobis D^2 when H0 is that a new data point comes from a known multivariate normal population);

(b) its sampling distribution should be known;

(c) it should be appropriate to H1 i.e. to the type of departures you wish to detect.

---

e.g. Suppose X1 to Xn are independent random samples form a normal population with unknown parameters μ and σ .

Let H0 be: $\mu = 10$
& H1 be: $\mu \neq 10$

Then an appropriate indicator of disagreement with H0 is

$$D = (\bar{X} - 10) / (s/\sqrt{n})$$

Suppose d is the observed value of D. Then the null distribution probability

$$p = P(|D| \geq |d| \mid H0)$$

can be calculated from the known sampling distribution of D given H0 (the t distribution).

It is referred to as the "p-value" for the observed value of the test statistic.

A small p value means that sampling fluctuations would be unlikely to have produced a D value as large as that observed purely by chance, if H0 were true.

e.g. p=0.01 means that a value has been observed which ought to have occurred only once in a hundred times by chance.

Small p values provide evidence for rejecting H0 in favour of H1.

Example (cont):

Suppose n=10 and independent random sampling gives:

10.1, 10.2, 10.2, 10.1, 10.3, 10.0, 10.2, 10.0, 10.1, 9.9.

Then $\bar{x}$ = 10.11 and s = 0.12, so the standard error of x is $s/\sqrt{10}$ = 0.038.

To compare $\bar{x}$ with 10 on the scale appropriate to its precision (repeatability) calculate:

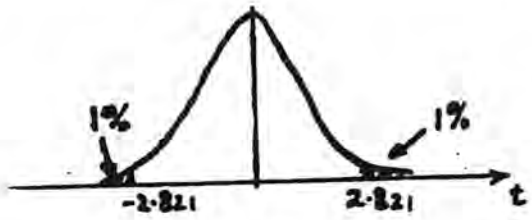
$$d = (10.11-10)/(0.038) = 2.89.$$

Now $D = (\bar{X}-\mu)/(s/\sqrt{n})$ has the t with 9 degrees of freedom as its "null distribution" (sampling distribution under H0).

Therefore, we can look up tables and get an approximation to the p-value:

p = Pr(|t| > 2.89 ; =10, normal distn, independent sampling)

Find t (0.99)=2.821, so p<2% :



H0 is unlikely to have produced the given sample.

Decision Making

In many ways it would be best to present the results of an hypothesis testing exercise as follows:

- the estimate or test statistic obtained
- the p-value for the test statistic

It would then be up to the recipient of the report to make the decisions as to statistical and practical significance.

Such decisions would also often be aided by a confidence interval giving a range of parameter values reasonably consistent with the

data.

By tradition (e.g. scientific journals requiring a "significant" result before publishing), and because it is sometimes genuinely necessary, the decision making process is incorporated in the hypothesis testing.

This is done by specifying in advance the minimum p-value α which under which H0 will be deemed acceptable (α = 0.05 is popular).

Presetting a minimum p-value implies:

(a) Defining a critical region for the test statistic - such that any value falling in this region leads to rejection of H0.

(b) Having a preset risk (100α %) of rejecting H0 when it is true. This risk is called the significance level of the test procedure.

Example:

Laboratory assay precision is asserted to be such that the relative standard deviation of a single assay is:

$$\sigma = 0.15\%$$

Reassay results were collected by splitting 16 soil samples in half and submitting subsamples of each half in different batches.

The data:

assay	reassay	diff (x).
58.8	58.9	-0.1
61.5	61.7	-0.2
54.9	55.8	-0.9
47.7	48.5	-0.8
60.0	60.5	-0.5
58.7	57.3	(1.4)
63.0	63.2	-0.2
61.2	60.9	0.3
55.8	56.1	-0.3
60.1	60.1	0.0
59.9	60.7	-0.8
62.9	63.7	-0.8
50.6	50.8	-0.2
53.0	54.0	-1.0
61.8	62.0	-0.2
59.1	59.1	0.0

$\bar{x}$ = -0.269 (-0.38)
s = 0.585 (0.393)
s^2 = 0.342 (0.155)

[N.B. Why are so many of the reassays increases on the assays?

$$s.e.(\bar{x}) = 0.393/\sqrt{15} \quad = 0.102$$

"paired t" = -3.74 (omitting one unusual value)]

How should we test the claimed precision at the 5% significance level?

First some theory:

Var(assay-reassay)
= var(assay) + var(reassay)
= $2\sigma^2$
= 2(0.15)
= 0.045 (under H0).

To compare s^2 (the sample variance of the differences) with 0.045 (the population value under H0), take the ratio:

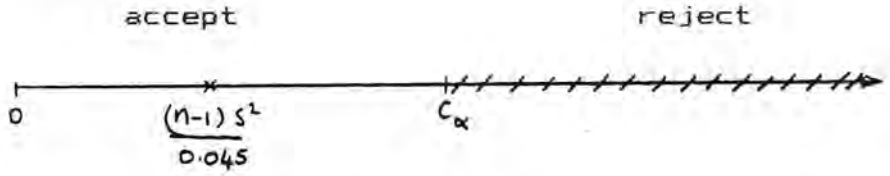
$$= (n-1)s^2/0.045.$$

The sampling distribution of this under H0 is chi-squared with n-1 degrees of freedom.

To construct an appropriate critical region, note that the appropriate H1 (for detecting greater imprecision than claimed) is

H1: $\sigma^2_{diff} > 0.045$

In line with this, we need a critical region of the form:



With such a region, the minimum acceptable p-value (of getting a chi-squared as big as the sample value just by chance) occurs when the sample value equals the critical level c_α (see picture).

To make this minimal acceptable p-value (or significance level) equal 0.05 means choosing c_α such that

$$P(\chi^2_{n-1} > c) = 0.05$$

i.e. must choose c_α to be the 95% percentile of the χ^2 distribution.

From tables, this is $c_\alpha = 25.0$.

Thus (at last), at the 5% level of significance, the procedure for testing H0 is:

- accept H0 if $15s^2/0.045 > 25$;

- otherwise reject it & conclude the data is not consistent with the claimed level of precision.

The result:

$$15s^2/0.045 = (15)(0.342)/0.045$$
$$= 114$$

Thus reject HO (would get same result leaving out 1.4 difference).

N.B. The p-value for the observed test statistic is less than 0.005. It would be very unlikely to occur just by chance if HO were true.

A Catalogue of Test Statistics

(a) inference on μ, σ known:

$$(\bar{x}-\mu)/(\sigma/\sqrt{n})$$ (st. normal)

(b) inference on μ, σ unknown:

$$(\bar{x}-\mu)/(s/\sqrt{n})$$ (t_{n-1} dist'n)

(c) inference on $\mu_x-\mu_y$, variances known:

$$((\bar{x}-\bar{y})-(\mu_x-\mu_y))/\sqrt{\frac{\sigma_x^2}{m}+\frac{\sigma_y^2}{n}}$$ (st. normal)

(d) inference on $\mu_x-\mu_y$, unknown common variance:

$$((\bar{x}-\bar{y})-(\mu_x-\mu_y))/s_p\sqrt{\frac{1}{m}+\frac{1}{n}},$$

with $$s_p^2 = \left(\sum_{i=1}^{m}(x_i-\bar{x})^2+\sum_{i=1}^{n}(y_i-\bar{y})^2\right)/(m+n-2)$$

(t_{m+n-2})

(e) inference on $\mu_x-\mu_y$, unknown & possibly different variances

$$((\bar{x}-\bar{y})-(\mu_x-\mu_y))/\sqrt{\frac{s_x^2}{m}+\frac{s_y^2}{n}}$$ (approx. t with d.f. given by Satterthwaite approximation)

(f) inference on σ^2

$$(n-1)s^2/\sigma^2$$ (χ^2_{n-1})

(g) inference on σ_x^2/σ_y^2

$$(s_x^2/s_y^2)/(\sigma_x^2/\sigma_y^2)$$ ($F_{m-1,n-1}$)

(h) coefficients a,b etc. in a general linear regression model

$$(\hat{a}-a)/\underbrace{\text{s.e.}(\hat{a})}_{\text{from computer}}$$ ($t_{\text{res. d.f.}}$)

(i) significance of groups of parameters in a regression or analysis of variance model

$$\frac{MS_{source}}{MS_{error}}$$ ($F_{\text{source d.f., error d.f.}}$)

(j) the proportion p of individuals of a certain type (assuming independent random sampling)

$$(\hat{p}-p)/\sqrt{\hat{p}(1-\hat{p})/n}$$ (approx normal if neither np or n(1-p) is small)

10. MULTIVARIATE ANALYSIS OF VARIANCE

When there is more than one variable measured per plot in the design of an experiment, multivariate analyses of variance can be applied as a direct generalisations of the usual univariate analyses. They allow tests of treatment effects to take into account all variates at once.

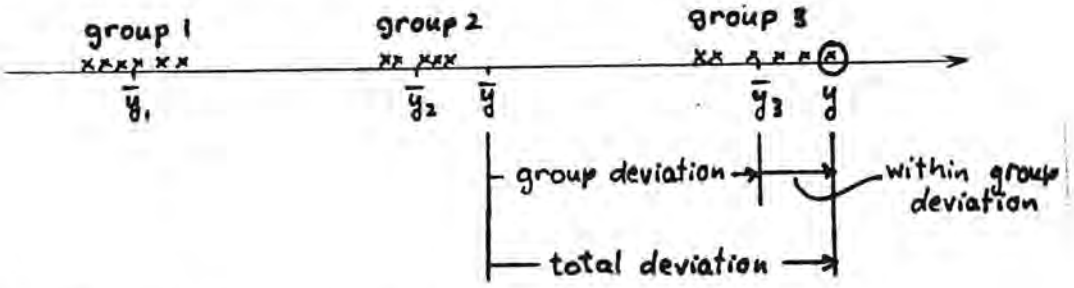
The simplest case to deal with (and the only one in this course) is the simple one-way classification, in which the individuals are classified into a number of known groups (e.g. according to experimental treatment applied).

Review of the Univariate Analysis

In the one-way classification the deviation of any single observation from the overall mean can be split up into two components:

total dev'n = group dev'n + within group dev'n

picture:



A mathematical theorem (like Pythagoras):

Total SS (sum of squares of total dev'ns)

= Between SS (sum of squares over all points of group components of deviations)

+ Within SS (of within group deviations)

Analytical Idea:

Use the relative sizes of the between and within group sums of squares to establish a measure of group separation.

e.g between SS/total SS or between SS/within SS.

In the univariate case, it is more usual to think in terms of an F ratio, where each SS is replaced by a mean square:

MS = SS/DF ("degrees of freedom")

effective number of independent terms in SS

[N.B. In the one-way ANOVA, the number of degrees of freedom between groups is the number of groups minus 1, and the within groups DF is the sum over groups of (no. in group -1). The total DF is one less than the number of data points.]

F ratios much bigger than one indicate a significant difference between groups (detectable above the variability evident within groups).

F ratios have tabulated percentiles showing what one should expect by chance if the population means for the groups are in fact all the same. Values which should be exceeded by chance less than 5% of the time usually lead to rejection of the hypothesis that there are no group differences.

For F tables to be reliable, it is necessary that the distribution within each group be approximately normal, and that all groups have the same within group variance.

The multivariate case:

For each variate Xj, we know from the above that the sum of squares of total deviations equals the sum of between and within sums of squares of deviations.

The same identity works for products of deviations of two variates:

total deviations:

(var j - overall mean j).(var j'- overall mean j')

between group deviations:

(grp mean j - overall mean j).(grp mean j' -overall mean j')

within group deviations:

(var j - group mean j).(var j'- group mean j')

All of this information can be summarised in matrices of total, between group and within groups sums of squares and products (of deviations). Denote these T, B and W.

Then W divided by the within groups degrees of freedom estimates the within group variance-covariance matrix (which must be assumed common to all groups for the analysis to be valid as described below).

If the group population means were all the same then T divided by the number of data points minus one would also estimate this variance-covariance matrix.

We need some way to compare W and T=B+W. The customary way is to use "determinants". These can probably be best understood in terms of the eigenvalues of the sums of squares and products matrices. Divided by their degrees of freedom, these ssp matrices are covariance matrices, and hence could be subjected to a principal component analysis -- wherein eigenvalues are obtained as maximal variances orthogonal to variance already extracted.

For a covariance matrix, the determinant is simply the product of the principal component variances (eigenvalues).

Likewise, the determinant of an ssp matrix is the product of all its eigenvalues. It provides an overall measure of the multivariate scatter summarised in the matrix.

"Wilk's Lambda" compares the relative scales of within and total deviations using the ratio of det(W) to det(T):

$$\Lambda = \det(W)/\det(T).$$

Under the null hypothesis that there are no differences in group means, this statistic has a known distribution, the

$$\Lambda\,(k, n-ngps, ngps-1) \text{ distribution.}$$

no. vars error df hypothesis df

Wilk's lambda always falls between zero and one. Values of near zero indicate that total deviations are large relative to within group deviations, meaning that population group means are likely to be different. Values near one mean the between group components of total deviations are generally small and unlikely to be significant.

To test significance of Λ ,Bartlett's approximation is mostly used. This states that $-\ln\Lambda$ times a constant is approximately chi-squared with degrees of freedom equal to the number of variates times the number of hypothesis degrees of freedom. The constant is:

$$\text{error df} - \tfrac{1}{2}(\text{no.vars} - \text{hyp. df} + 1)$$

The approximation is most valid when the number of error degrees of freedom is large. For 2 variates, or for 1 or 2 hypothesis df (2 or 3 groups in a one way classification), Wilk's lambda can be transformed to have an F distribution exactly, given multivaraite normal assumptions. See MARDIA, KENT & BIBBY for details of:

$$\frac{1-\Lambda(k,m,1)}{\Lambda(k,m,1)} \sim \frac{k}{m-k+1} F_{k,\,m-k+1}$$

$$\frac{1-\Lambda(1,m,n)}{\Lambda(1,m,n)} \sim \frac{n}{m} F_{n,m}$$

$$\frac{1-\sqrt{\Lambda(k,m,2)}}{\sqrt{\Lambda(k,m,2)}} \sim \frac{k}{m-k+1} F_{2k,\,2(m-k+1)}$$

$$\frac{1-\sqrt{\Lambda(2,m,n)}}{\sqrt{\Lambda(2,m,n)}} \sim \frac{n}{m-1} F_{2n,\,2(m-1)}$$

Small Example (3 small groups of anteater skull measurements):

Means:

	x1	x2	x3	
Minas Graes, Brazil	2.097	2.100	1.625	(6 skulls)
Matto Grosso, Brazil	2.080	2.087	1.617	(4 skulls)
Santa Cruz, Bolivia	2.099	2.102	1.643	(3 skulls)

SSP Matrices:

$$B\ (2\ d.f.) = \begin{pmatrix} 8060 & 6233 & 7498 \\ 6233 & 4820 & 5859 \\ 7498 & 5859 & 11844 \end{pmatrix}$$

$$W\ (10\ d.f.) = \begin{pmatrix} 63423 & 62418 & 76157 \\ 62418 & 63528 & 76127 \\ 76157 & 76127 & 109673 \end{pmatrix}$$

$$T\ (12\ d.f) = \begin{pmatrix} 71483 & 68651 & 83655 \\ 68651 & 63528 & 76127 \\ 83655 & 76127 & 121517 \end{pmatrix}$$

Λ (3, 10, 2) = det(W)/det(T) = 0.6014

(3: dimension; 10: W d.f.; 2: B d.f.)

From page 10-3,

$$\frac{(10-3+1)}{3} \cdot (1-\Lambda^{1/2})/\Lambda^{1/2} \sim F_{6,16}$$

The value of this test statistic is 0.772 ($p>0.5$, n.s.).

N.B. The univariate $F_{2,10}$'s (e.g. (8060/2)/(63423/10) for x1) aren't significant either.

The following example of a two-way MANOVA illustrates the fact that MANOVA can be used for data from a variety of structured experiments or sampling designs. Using SAS, this can be done by using the MANOVA statement in PROC ANOVA or PROC GLM (see page 498 of the SAS/STATS manual for an example). MANOVA offers most advantage of univariate ANOVA's when the group separations are clear in multivariate space but not variable by variable.

Two-Way MANOVA Example:

Example 12.7.1 (Morrison, 1976, p. 190) We wish to compare the weight losses of male and female rats ($r = 2$ sexes) under $c = 3$ drugs where $n = 4$ rats of each sex are assigned at random to each drug. Weight losses are observed for the first and second weeks ($p = 2$) and the data is given in Table 12.7.1. We wish to compare the effects of the drugs, the effect of sex, and whether there is any interaction.

Table 12.7.1 Weight losses (in grams) for the first and second weeks for rats of each sex under drugs A, B, and C (Morrison, 1976, p. 190)

	Drug			
Sex	A	B	C	Row sums
Male	(5, 6)	(7, 6)	(21, 15)	(33, 27)
	(5, 4)	(7, 7)	(14, 11)	(26, 22)
	(9, 9)	(9, 12)	(17, 12)	(35, 33)
	(7, 6)	(6, 8)	(12, 10)	(25, 24)
Column sums	(26, 25)	(29, 33)	(64, 48)	(119, 106)
Female	(7, 10)	(10, 13)	(16, 12)	(33, 35)
	(6, 6)	(8, 7)	(14, 9)	(28, 22)
	(9, 7)	(7, 6)	(14, 8)	(30, 21)
	(8, 10)	(6, 9)	(10, 5)	(24, 24)
Column sums	(30, 33)	(31, 35)	(54, 34)	(115, 102)
Treatment sums	(56, 58)	(60, 68)	(118, 82)	
Grand total		(234, 208)		

Table 12.7.2 MANOVA table for the data in Table 12.7.1

		SSP matrix **A**		
Source	d.f.	a_{11}	a_{12}	a_{22}
Sex (**R**)	1	0.667	0.667	0.667
Drugs (**C**)	2	301.0	97.5	36.333
Interaction (**I**)	2	14.333	21.333	32.333
Residual (**W**)	18	94.5	76.5	114.0
Total (**T**)	23	410.5	196.0	183.333

We first test for interaction. We construct the MANOVA table (Table 12.7.2), using the totals in Table 12.7.1. From Table 12.7.2, we find that

$$|\mathbf{W}| = 4920.75, \qquad |\mathbf{W}+\mathbf{I}| = 6281.42.$$

Hence,

$$\Lambda = |\mathbf{W}|/|\mathbf{W}+\mathbf{I}| = 0.7834 \sim \Lambda(2, 18, 2).$$

From (12.3.7),

$$(17/2)(1-\sqrt{0.7834})/\sqrt{0.7834} = 1.10 \sim F_{4,34}.$$

This is clearly not significant so we conclude there are no interactions and proceed to test for main effects.

First, for drugs, $|\mathbf{W}+\mathbf{C}| = 29\,180.83$. Therefore

$$\Lambda = 4920.75/(29\,180.83) = 0.1686 \sim \Lambda(2, 18, 2).$$

From (12.3.7),

$$(17/2)(1-\sqrt{0.1686})/\sqrt{0.1686} = 12.20 \sim F_{4,34}.$$

This is significant at 0.1%, so we conclude that there are very highly significant differences between drugs.

Finally, for sex, $|\mathbf{W}+\mathbf{R}| = 4957.75$. Thus

$$\Lambda = 4920.75/4957.75 = 0.9925 \sim \Lambda(2, 18, 1).$$

Again from (12.3.7), the observed value of $F_{2,34}$ is 0.06. This is not significant so we conclude there are no differences in weight loss between the sexes.

11. GROUP DISCRIMINATION; CANONICAL VARIATES

- To set up canonical variates, you need multivariate data from each of a number of known (or hypothesised) groups (not necessary for these to include all of your data points, especially if some are of uncertain classification).

- Canonical variates give a low dimensional representation (on canonical variate axes) which tries to maintain as much of the multivariate group separation as possible.

- Best when the variance-covariance matrices for the groups are all the same (transformations help here).

- Makes use of a "pooled" within group variance-covariance matrix and a between group mean variance-covariance matrix (the W and B of section 10 divided by their degrees of freedom) as the basis of its calculations (these matrices provide the analogues of the univariate within and between group mean squares).

The Canonical Variates Recipe
(one description, matrix algebra in practice)

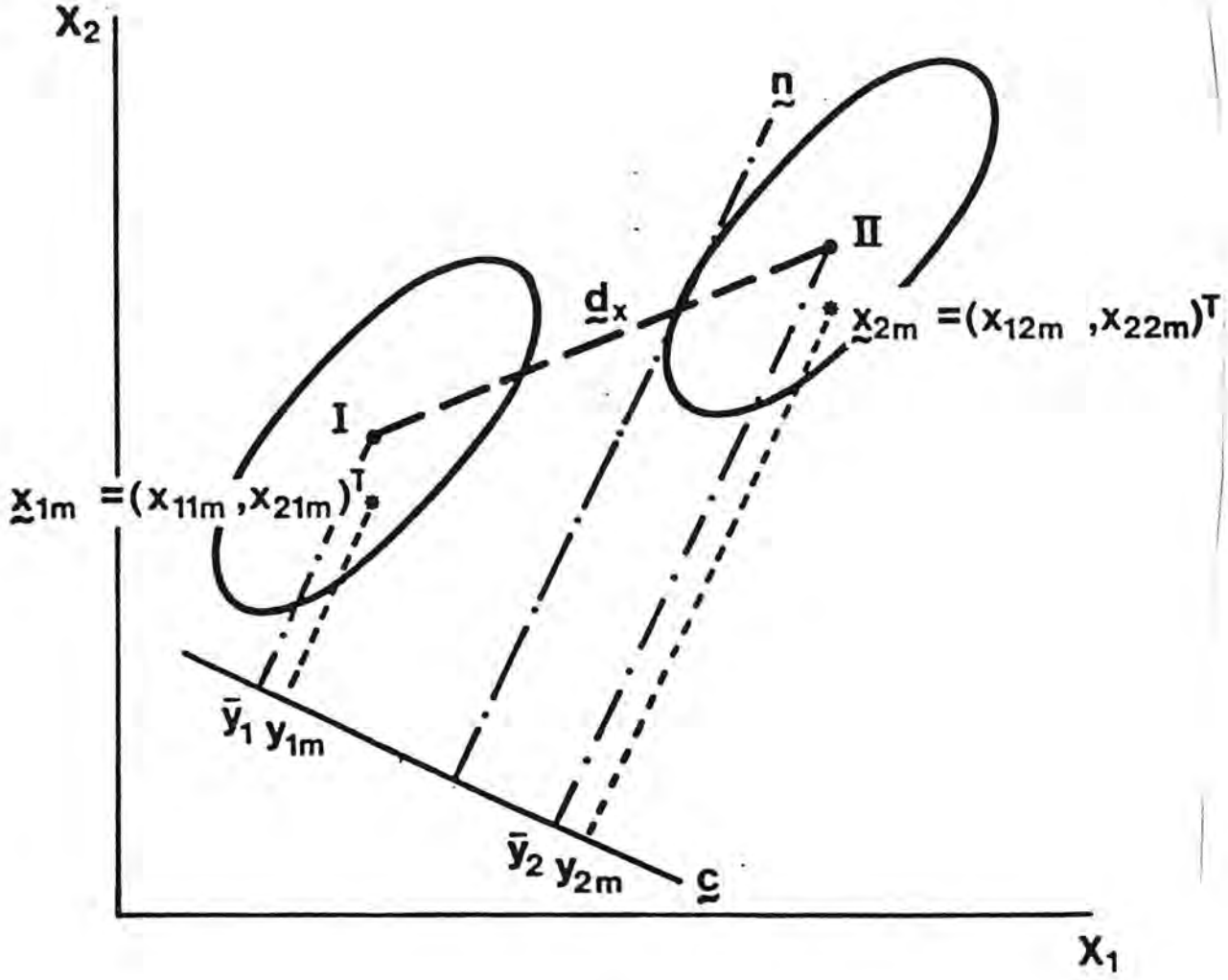
(a) Choose the first canonical variate axis Y1 (↔ linear combination of X1,X2, ... Xk) so that the y1 values obtained have the maximum possible F ratio (between MS/within MS) for group separation.

(b) Choose subsequent canonical variates Y2,Y3 ... Yr so that they are:

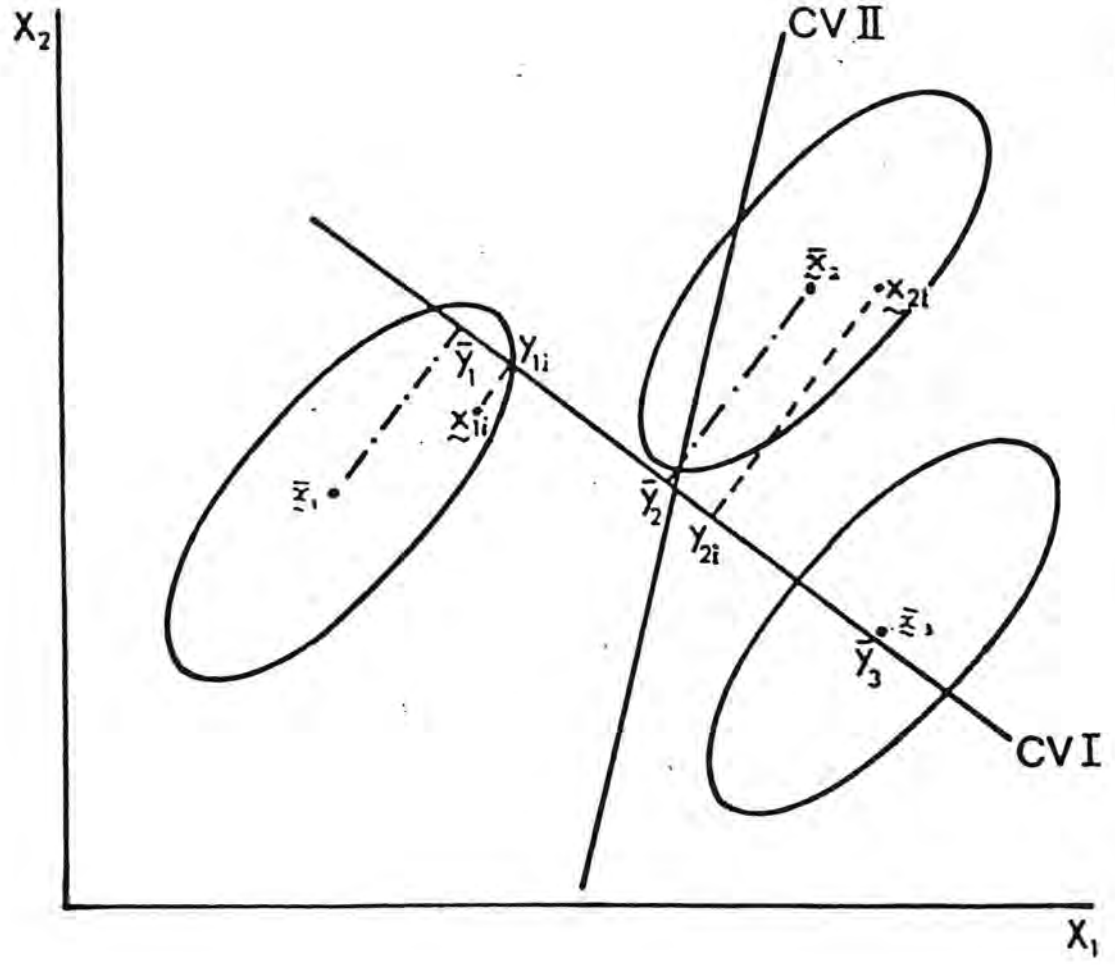
- uncorrelated with previous ones (when correlations are determined on a pooled within groups basis or overall)

- have maximum F ratio for group separation under this constraint.

Picture for Two Groups:



Picture with Three Groups:



Points to Note:

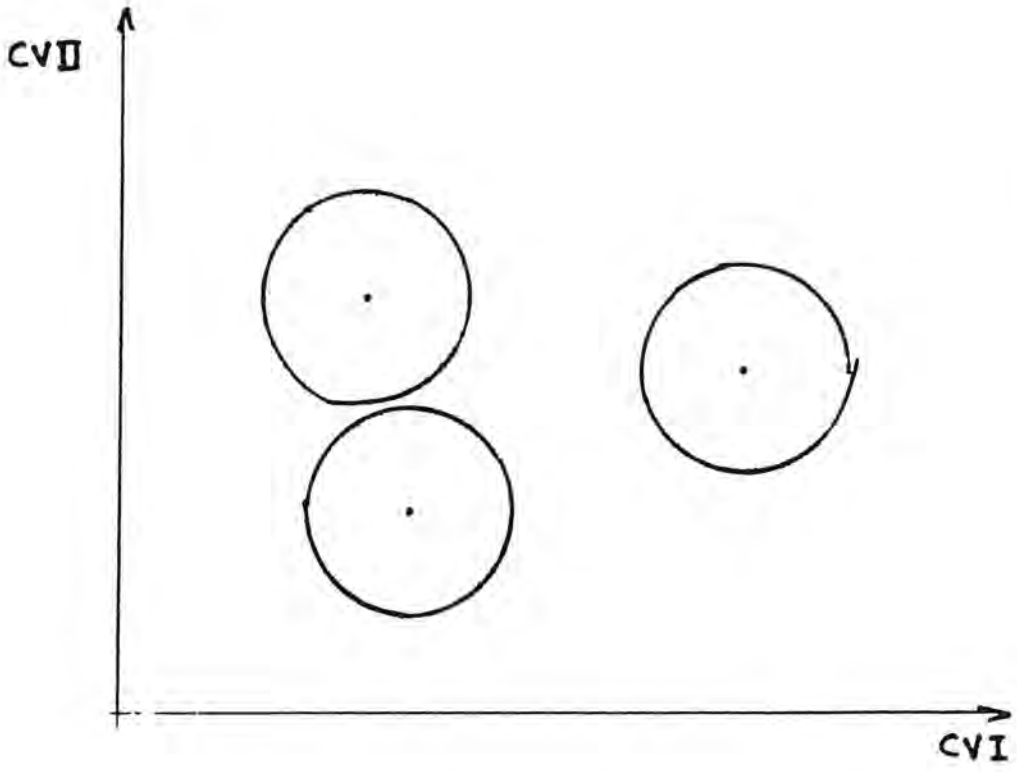
(a) Canonical variate axes are not necessarily perpendicular to each other.

(b) The number of canonical variates which can be obtained (r) is limited by the number of groups g: $r < g-1$. Also must have $r < k$, the original number of variables.

(c) Canonical variates are usually scaled so that their within group standard deviation (the pooled version) is unity (SAS PROC CANDISC does this).

(d) If the canonical variate scores are plotted, using recangular axes, and the constancy of within group variances and covariances is a valid assumption, then each group should have spherical probability contours or surfaces, all of the same standard radius. e.g. circles in two dimensions.

Picture:



In this picture, closeness in simple Euclidean distance determines which group a new point is most likely to have come from.

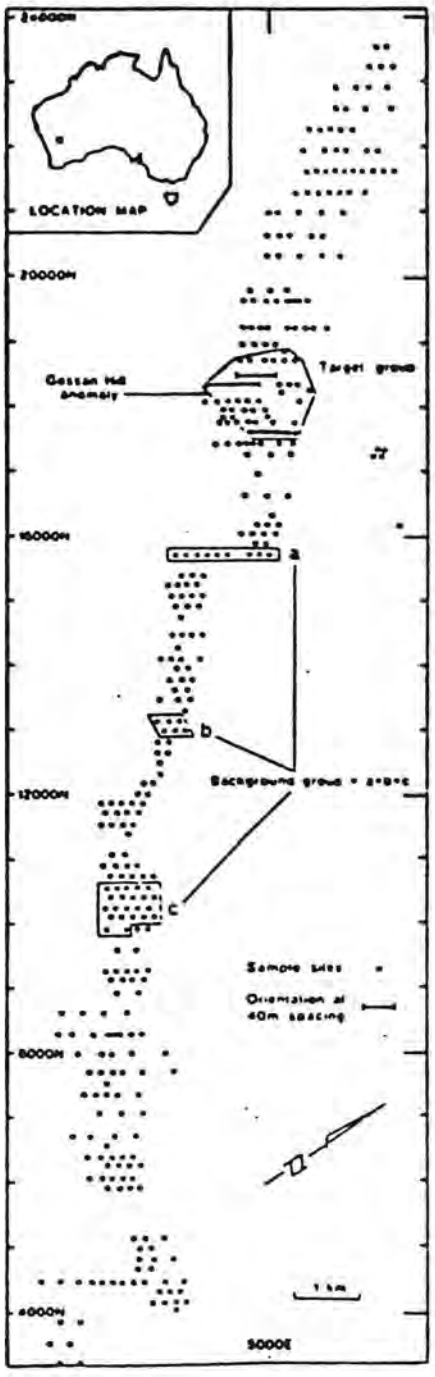
A Concrete Example

(from work on exploration geochemistry data by Smith, Campbell & Perdrix, CSIRO WA).

- Golden Grove W.A.

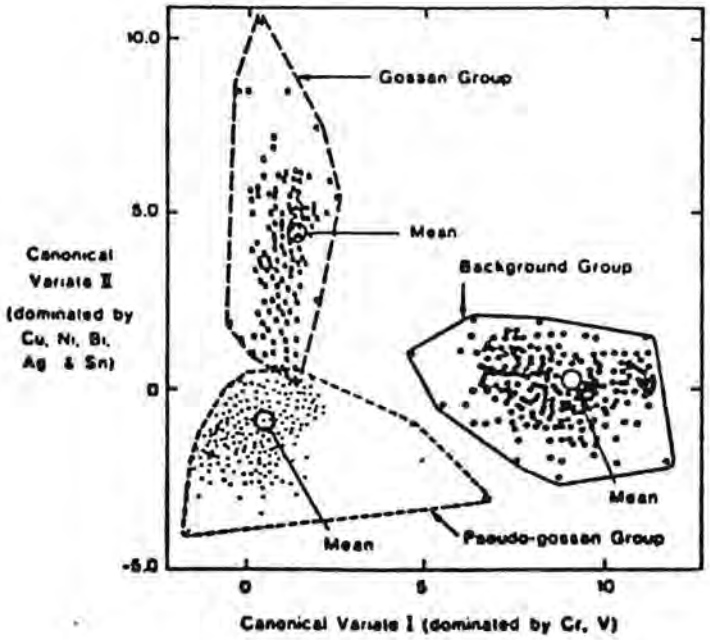
- 3 groups identified: gossans, pseudo-gossans and background.

sample locations:

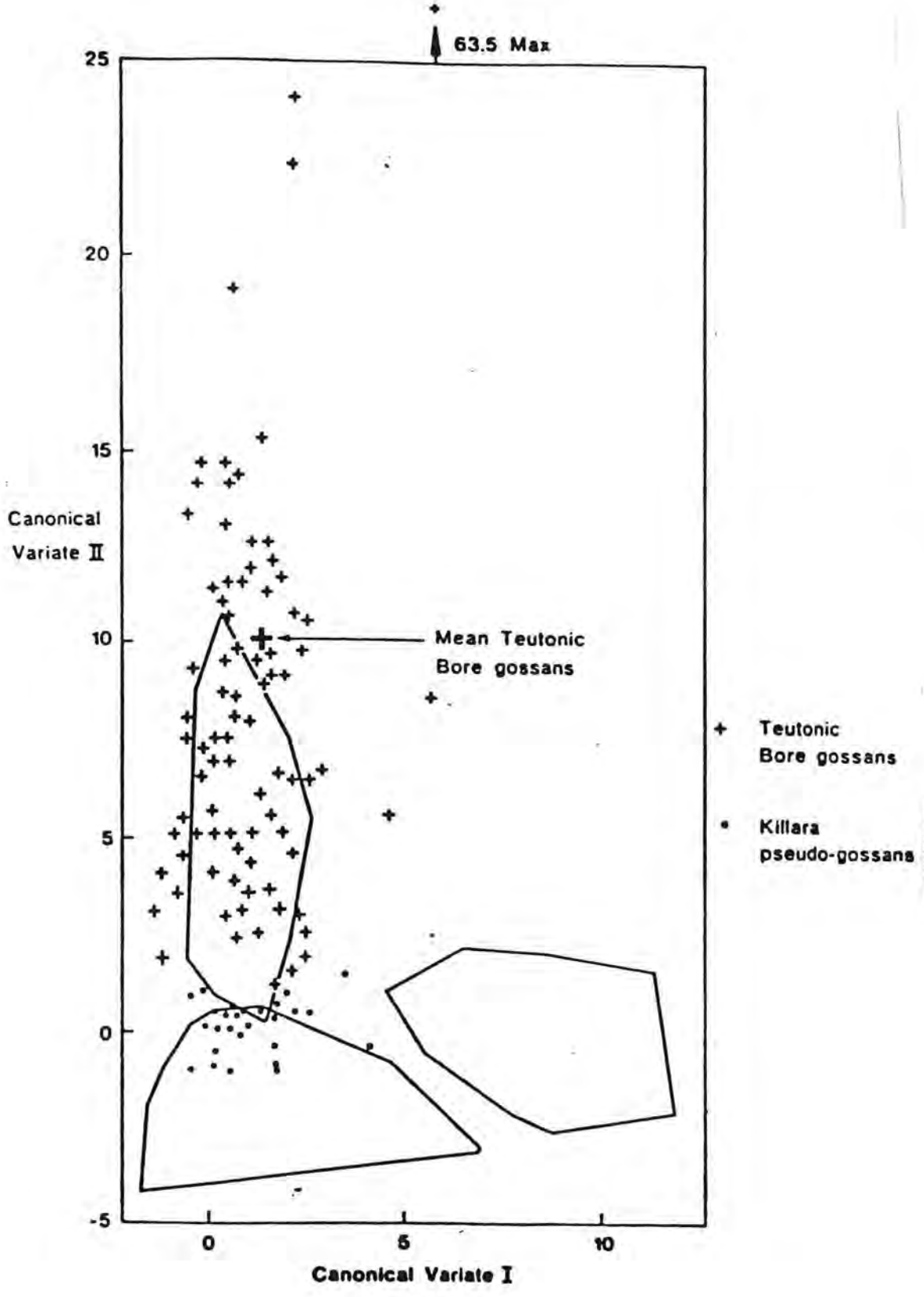


stratified sample

CV II vs CV I , showing the points used in estimating the canonical variate analyses:



CVII vs CVI on some new samples (from Teutonic Bore & Killara):



EXAMPLE

Analysis of Iris Data Using PROC CANDISC

The iris data published by Fisher (1936) have been widely used for examples in discriminant analysis and cluster analysis. The sepal length, sepal width, petal length, and petal width were measured in millimeters on 50 iris specimens from each of three species, *Iris setosa, I. versicolor,* and *I. virginica*. The following example is a canonical discriminant analysis, creating an output data set containing scores on the canonical variables, and plotting the canonical variables.

```
DATA IRIS;
   TITLE 'FISHER (1936) IRIS DATA';
   INPUT SEPALLEN SEPALWID PETALLEN PETALWID SPEC_NO @@;
   IF SPEC_NO=1 THEN SPECIES='SETOSA    ';
   IF SPEC_NO=2 THEN SPECIES='VERSICOLOR';
   IF SPEC_NO=3 THEN SPECIES='VIRGINICA ';
   LABEL SEPALLEN='SEPAL LENGTH IN MM.'
         SEPALWID='SEPAL WIDTH  IN MM.'
         PETALLEN='PETAL LENGTH IN MM.'
         PETALWID='PETAL WIDTH  IN MM.';
   CARDS;
50 33 14 02 1 64 28 56 22 3 65 28 46 15 2
67 31 56 24 3 63 28 51 15 3 46 34 14 03 1
69 31 51 23 3 62 22 45 15 2 59 32 48 18 2
46 36 10 02 1 61 30 46 14 2 60 27 51 16 2
65 30 52 20 3 56 25 39 11 2 65 30 55 18 3
58 27 51 19 3 68 32 59 23 3 51 33 17 05 1
57 28 45 13 2 62 34 54 23 3 77 38 67
63 33 47 16 2 67 33 57 25 3
49 25 45 17 3 55 35
70 32 47 14
48 34
                                        56 22 3
                                  51 35 14 03 1
                            .. 20 3 61 29 47 14 2
                        .. 31 49 15 2 64 27 53 19 3
                  41 3 55 25 40 13 2 48 34 16 02 1
   . 50 14 01 1 45 23 13 03 1 57 25 50 20 3
 57 38 17 03 1 51 38 15 03 1 55 23 40 13 2
 66 30 44 14 2 68 28 48 14 2 54 34 17 02 1
 51 37 15 04 1 52 35 15 02 1 58 28 51 24 3
 67 30 50 17 2 63 33 60 25 3 53 37 15 02 1
;
PROC CANDISC ALL OUT=DISC;
   CLASSES SPECIES;
   VAR SEPALLEN SEPALWID PETALLEN PETALWID;
PROC PLOT;
   PLOT CAN2*CAN1=SPEC_NO;
   TITLE2 'PLOT OF CANONICAL DISCRIMINANT FUNCTIONS';
```

Output 13.1 Iris Data: PROC CANDISC

FISHER (1936) IRIS DATA 1

CANONICAL DISCRIMINANT ANALYSIS

150 OBSERVATIONS	149 DF TOTAL
4 VARIABLES	147 DF WITHIN CLASSES
3 CLASSES	2 DF BETWEEN CLASSES

SPECIES	FREQUENCY	WEIGHT	PROPORTION
SETOSA	50	50	0.333333
VERSICOLOR	50	50	0.333333
VIRGINICA	50	50	0.333333

⓫ TOTAL-SAMPLE SSCP MATRIX

VARIABLE	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SEPALLEN	10216.83	-632.267	18987.3	7692.433
SEPALWID	-632.267	2830.693	-4911.88	-1812.43
PETALLEN	18987.3	-4911.88	46432.54	19304.58
PETALWID	7692.433	-1812.43	19304.58	8656.993

⓬ BETWEEN-CLASS SSCP MATRIX

VARIABLE	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SEPALLEN	6321.213	-1995.27	16524.84	7127.933
SEPALWID	-1995.27	1134.493	-5723.96	-2293.27
PETALLEN	16524.84	-5723.96	43710.28	18677.4
PETALWID	7127.933	-2293.27	18677.4	8041.333

⓭ POOLED WITHIN-CLASS SSCP MATRIX

VARIABLE	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SEPALLEN	3895.62	1363	2462.46	564.5
SEPALWID	1363	1696.2	812.08	480.84
PETALLEN	2462.46	812.08	2722.26	627.18
PETALWID	564.5	480.84	627.18	615.66

UNIVARIATE STATISTICS

VARIABLE	❶ MEAN	❷ TOTAL STD	❸ WITHIN STD	❹ BETWEEN STD	❺ R-SQUARED	❻ RSQ/(1-RSQ)	F	❼ PROB > F
SEPALLEN	58.43333333	8.28066128	5.14789436	7.95060585	0.618706	1.623	119.265	0.0001
SEPALWID	30.57333333	4.35866285	3.39687732	3.36822406	0.400783	0.669	49.160	0.0001
PETALLEN	37.58000000	17.65298233	4.30338469	20.90700361	0.941372	16.057	1180.161	0.0
PETALWID	11.99333333	7.62237669	2.04650025	8.96734818	0.928883	13.061	960.007	0.0

AVERAGE R-SQUARED: UNWEIGHTED = 0.7224358 WEIGHTED BY VARIANCE = 0.8689444

FISHER (1936) IRIS DATA 2

CANONICAL DISCRIMINANT ANALYSIS

❽ CLASS MEANS

SPECIES	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SETOSA	50.06000000	34.28000000	14.62000000	2.46000000
VERSICOLOR	59.36000000	27.70000000	42.60000000	13.26000000
VIRGINICA	65.88000000	29.74000000	55.52000000	20.26000000

❾ TOTAL-STANDARDIZED CLASS MEANS

SPECIES	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SETOSA	-1.0112	0.8504	-1.3006	-1.2507
VERSICOLOR	0.1119	-0.6592	0.2844	0.1662
VIRGINICA	0.8993	-0.1912	1.0163	1.0845

(continued on next page)

(continued from previous page)

⑩ WITHIN-STANDARDIZED CLASS MEANS

SPECIES	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SETOSA	-1.6266	1.0912	-5.3354	-4.6584
VERSICOLOR	0.1800	-0.8459	1.1665	0.6189
VIRGINICA	1.4465	-0.2453	4.1689	4.0394

⑭ TOTAL-SAMPLE COVARIANCE MATRIX

VARIABLE	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SEPALLEN	68.5693512	-4.2434004	127.4315436	51.6270694
SEPALWID	-4.2434004	18.9979418	-32.9656376	-12.1639374
PETALLEN	127.4315436	-32.9656376	311.6277852	129.5609396
PETALWID	51.6270694	-12.1639374	129.5609396	58.1006264

⑮ BETWEEN-CLASS COVARIANCE MATRIX

VARIABLE	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SEPALLEN	63.2121333	-19.9526667	165.2484000	71.2793333
SEPALWID	-19.9526667	11.3449333	-57.2396000	-22.9326667
PETALLEN	165.2484000	-57.2396000	437.1028000	186.7740000
PETALWID	71.2793333	-22.9326667	186.7740000	80.4133333

⑯ POOLED WITHIN-CLASS COVARIANCE MATRIX

VARIABLE	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SEPALLEN	26.50081633	9.27210884	16.75142857	3.84013605
SEPALWID	9.27210884	11.53877551	5.52435374	3.27102041
PETALLEN	16.75142857	5.52435374	18.51877551	4.26653061
PETALWID	3.84013605	3.27102041	4.26653061	4.18816327

⑰ TOTAL-SAMPLE CORRELATION MATRIX

VARIABLE	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SEPALLEN	1.0000	-0.1176	0.8718	0.8179
SEPALWID	-0.1176	1.0000	-0.4284	-0.3661
PETALLEN	0.8718	-0.4284	1.0000	0.9629
PETALWID	0.8179	-0.3661	0.9629	1.0000

⑱ BETWEEN-CLASS CORRELATION MATRIX

VARIABLE	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SEPALLEN	1.0000	-0.7451	0.9941	0.9998
SEPALWID	-0.7451	1.0000	-0.8128	-0.7593
PETALLEN	0.9941	-0.8128	1.0000	0.9962
PETALWID	0.9998	-0.7593	0.9962	1.0000

⑲ POOLED WITHIN-CLASS CORRELATION MATRIX

VARIABLE	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SEPALLEN	1.0000	0.5302	0.7562	0.3645
SEPALWID	0.5302	1.0000	0.3779	0.4705
PETALLEN	0.7562	0.3779	1.0000	0.4845
PETALWID	0.3645	0.4705	0.4845	1.0000

⑳ MAHALANOBIS DISTANCES BETWEEN CLASSES

SPECIES	SETOSA	VERSICOLOR	VIRGINICA
SETOSA	.	9.4797	13.3935
VERSICOLOR	9.4797	.	4.1474
VIRGINICA	13.3935	4.1474	.

FISHER (1936) IRIS DATA 4

CANONICAL DISCRIMINANT ANALYSIS

F STATISTICS, NDF=4 DDF=144

SPECIES	SETOSA	VERSICOLOR	VIRGINICA
SETOSA	.	550.19	1098.27
VERSICOLOR	550.19	.	105.31
VIRGINICA	1098.27	105.31	.

PROB > MAHALANOBIS DISTANCE

SPECIES	SETOSA	VERSICOLOR	VIRGINICA
SETOSA	.	0.0	0.0
VERSICOLOR	0.0	.	0.0001
VIRGINICA	0.0	0.0001	.

㉑	CANONICAL CORRELATION	㉒ ADJUSTED CANONICAL CORRELATION	㉓ APPROX STANDARD ERROR	㉕ SQUARED CANONICAL CORRELATION	㉔ EIGENVALUES OF INV(E)*H = CANRSQ/(1-CANRSQ) EIGENVALUE	DIFFERENCE	PROPORTION	CUMULATIVE
1	0.984821	0.984508	0.002468	0.969872	32.1919	31.9065	0.9912	0.9912
2	0.471197	0.461445	0.063734	0.222027	0.2854	.	0.0088	1.0000

TESTS OF H0: THE CANONICAL CORRELATION IN THE CURRENT ROW AND ALL THAT FOLLOW ARE ZERO

	㉖ LIKELIHOOD RATIO	㉗ F	NUM DF	㉘ DEN DF	PR > F
1	0.02343863	199.1453	8	288	0.0
2	0.77797337	13.7939	3	145	0.0001

㉙ MULTIVARIATE TEST STATISTICS AND F APPROXIMATIONS
S=2 M=0.5 N=71.5

STATISTIC	VALUE	F	NUM DF	DEN DF	PR > F
WILKS' LAMBDA	0.02343863	199.145	8	288	0.0
PILLAI'S TRACE	1.191899	53.466	8	290	0.0001
HOTELLING-LAWLEY TRACE	32.47732	580.532	8	286	0.0
ROY'S GREATEST ROOT	32.19193	1166.957	4	145	0.0

NOTE: F STATISTIC FOR ROY'S GREATEST ROOT IS AN UPPER BOUND
F STATISTIC FOR WILKS' LAMBDA IS EXACT

FISHER (1936) IRIS DATA 5

CANONICAL DISCRIMINANT ANALYSIS

㉚ TOTAL CANONICAL STRUCTURE

	CAN1	CAN2
SEPALLEN	0.7919	0.2176
SEPALWID	-0.5308	0.7580
PETALLEN	0.9850	0.0460
PETALWID	0.9728	0.2229

㉛ BETWEEN CANONICAL STRUCTURE

	CAN1	CAN2

(continued on next page)

(continued from previous page)

SEPALLEN	0.9915	0.1303
SEPALWID	-0.8257	0.5642
PETALLEN	0.9998	0.0224
PETALWID	0.9940	0.1090

32 WITHIN CANONICAL STRUCTURE

	CAN1	CAN2
SEPALLEN	0.2226	0.3108
SEPALWID	-0.1190	0.8637
PETALLEN	0.7061	0.1677
PETALWID	0.6332	0.7372

33 STANDARDIZED CANONICAL COEFFICIENTS

	CAN1	CAN2
SEPALLEN	-0.6868	0.0200
SEPALWID	-0.6688	0.9434
PETALLEN	3.8858	-1.6451
PETALWID	2.1422	2.1641

34 RAW CANONICAL COEFFICIENTS

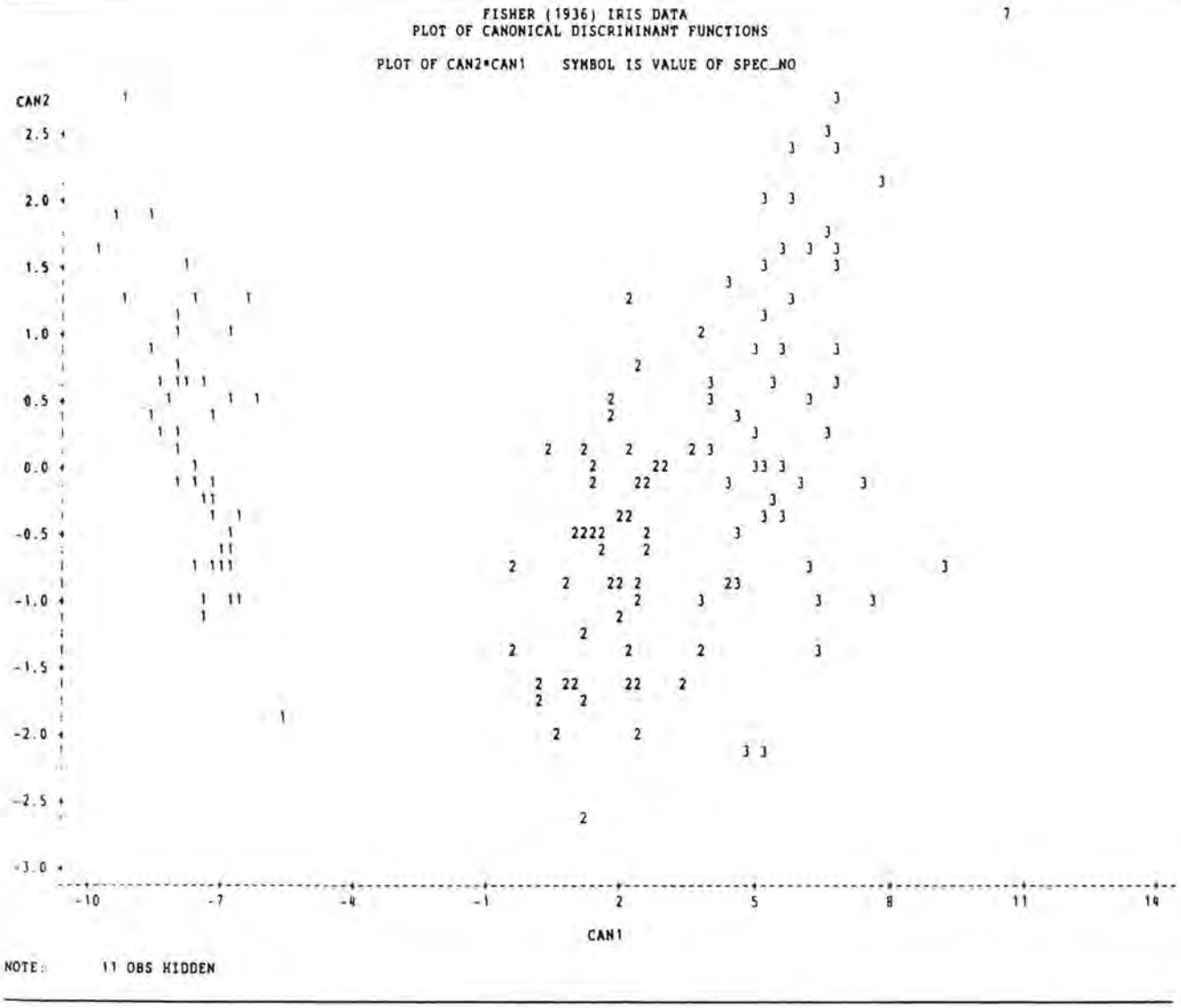
	CAN1	CAN2
SEPALLEN	-.0829377642	0.0024102149
SEPALWID	-.1534473068	0.2164521235
PETALLEN	0.2201211656	-.0931921210
PETALWID	0.2810460309	0.2839187853

35 CANONICAL DISCRIMINANT ANALYSIS

CLASS MEANS ON CANONICAL VARIABLES

SPECIES	CAN1	CAN2
SETOSA 1	-7.6076	0.2151
VERSICOLOR 2	1.8250	-0.7279
VIRGINICA 3	5.7826	0.5128

Output 13.2 Iris Data: PROC PLOT



12. ALLOCATION AND ATYPICALITY

(In SAS: PROC DISCRIM)

Basic Ideas:

- Need "training" samples from each of a number of groups.

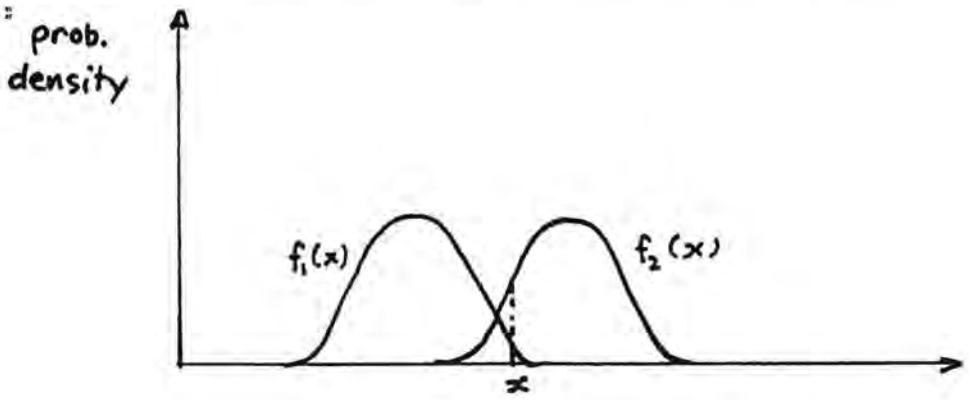
- Estimate variances and covariances within each group.

- Optionally assume these constant over groups and use pooled within groups variance-covariance matrix for all groups.

- Allocate new samples to groups according to a calculation (using the multivariate normal model) of their probabilities of being in each group (assuming they are in one of the predefined groups).

With common covariance matrices, this calculation is one of checking which group mean is closest (according to the common Mahalanobis distance).

In one dimension:



Prob(in group 1, assuming in 1 or 2) = $\dfrac{f_1(x)}{f_1(x)+f_2(x)}$

determined by the Mahalanobis distances, for common σ (=0.5 when they are equal).

The group with the biggest probability is the one to which x is allocated.

What if a point is not in reality from any of the training groups? e.g.:

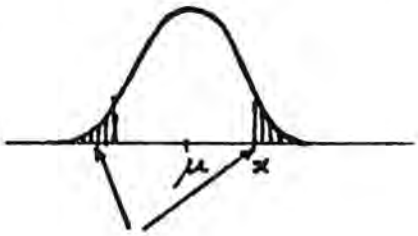
$$f_2(x) / (f_1(x) + f_2(x))$$

,the group 2 "membership probability" can be large without valid implication of actual membership (since the assumption of being in at least one of the groups is invalid).

Atypicality indices guard against this.

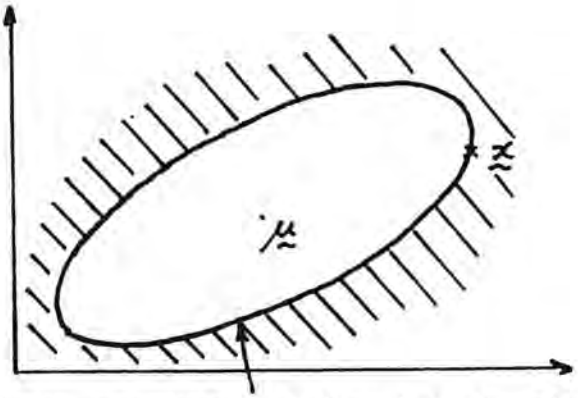
They use the probability of a point being as far away as it is from the group mean:

In one dimension:



typicality index (low not typical)

In two:

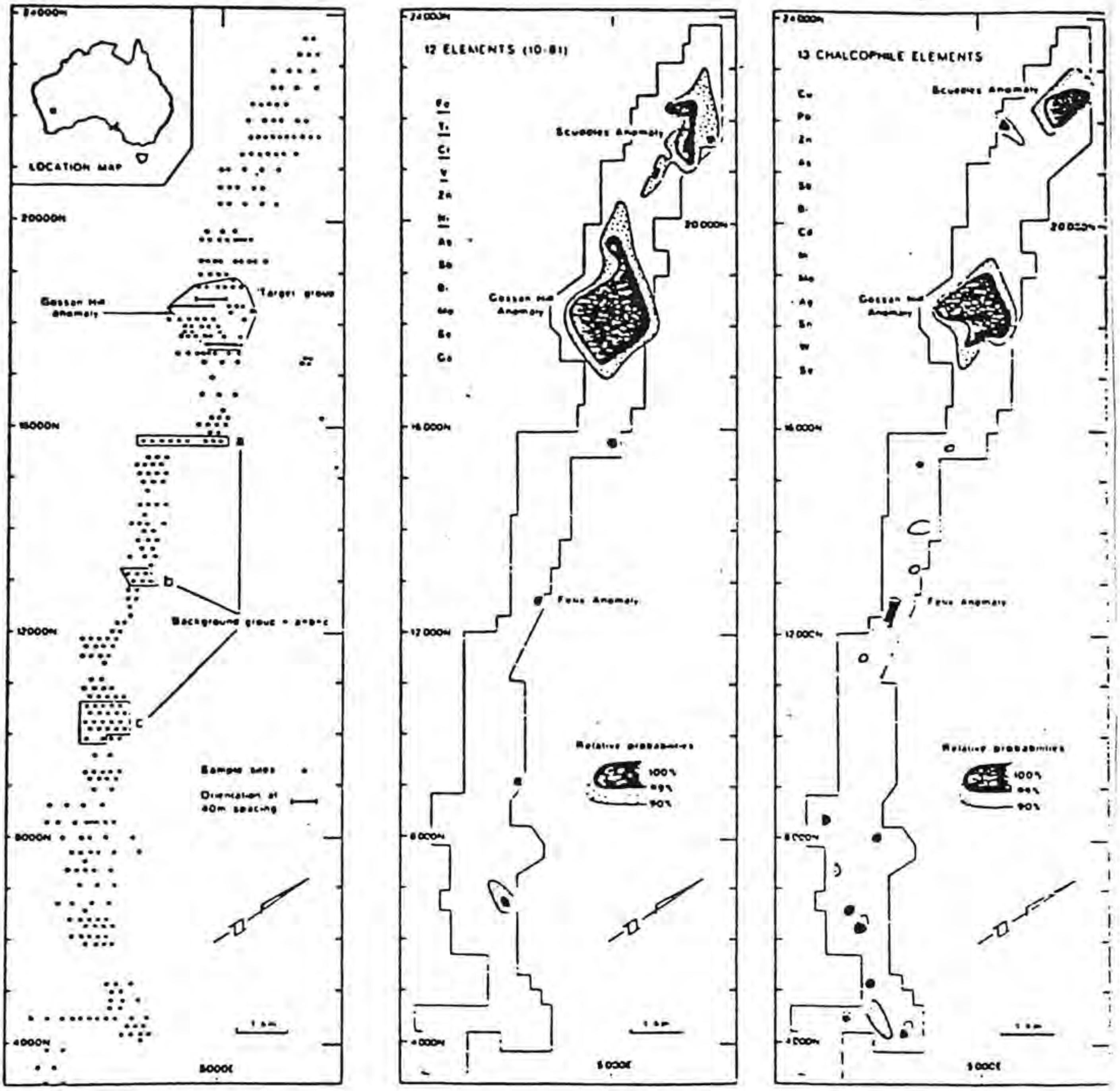


density contour through x (also Mahalanobis distance contour)

Typicality index = probability of being outside contour

(calculable from chi-squared distribution for D^2).

Results:


LOCATION MAP
Gossan Hill Anomaly
Target group
Background group = a+b+c
Sample sites
Orientation at 40m spacing
24000N
20000N
16000N
12000N
8000N
4000N
5000E
12 ELEMENTS (10-81)
Fe
Ti
Cr
V
Zn
Ni
As
Sb
Bi
Mo
Se
Co
Scuddles Anomaly
Gossan Hill Anomaly
Relative probabilities
100%
99%
90%
13 CHALCOPHILE ELEMENTS
Cu
Pb
Zn
As
Sb
Bi
Cd
In
Mo
Ag
Sn
W
Se

Another Example (from SAS/STATS manual, pp329-333):

Four remotely sensed spectral variables used to classify crops. First do analysis on training data to set up Mahalanobis distance functions (assuming constant within crop varaince and covariance). Then apply to a test data set. N.B. SAS doesn't seem to give access to the individual D values to allow calculation of atypicality probabilities.

Remote-Sensing Data on Crops: Example 2

In the example below, the observations are grouped into five crops: clover, corn, cotton, soybeans, and sugar beets. Four measures called X1-X4 make up the descriptive variables. The first PROC DISCRIM statement creates a calibration data set using the OUT= option. The second DISCRIM statement uses the information in the calibration data set to classify a test data set. Note that the values of the identification variable, XVALUES, are obtained by rereading the X1-X4 fields in the data lines as one character variable.

```
DATA CROPS;
   TITLE 'REMOTE SENSING DATA ON FIVE CROPS';
   INPUT CROP $ 1-10 X1-X4 XVALUES $ 11-21;
   CARDS;
CORN       16 27 31 33
CORN       15 23 30 30
CORN       16 27 27 26
CORN       18 20 25 23
CORN       15 15 31 32
CORN       15 32 32 15
CORN       12 15 16 73
SOYBEANS   20 23 23 25
SOYBEANS   24 24 25 32
SOYBEANS   21 25 23 24
SOYBEANS   27 45 24 12
SOYBEANS   12 13 15 42
SOYBEANS   22 32 31 43
```

```
COTTON     31 32 33 34
COTTON     29 24 26 28
COTTON     34 32 28 45
COTTON     26 25 23 24
COTTON     53 48 75 26
COTTON     34 35 25 78
SUGARBEETS22 23 25 42
SUGARBEETS25 25 24 26
SUGARBEETS34 25 16 52
SUGARBEETS54 23 21 54
SUGARBEETS25 43 32 15
SUGARBEETS26 54  2 54
CLOVER     12 45 32 54
CLOVER     24 58 25 34
CLOVER     87 54 61 21
CLOVER     51 31 31 16
CLOVER     96 48 54 62
CLOVER     31 31 11 11
CLOVER     56 13 13 71
CLOVER     32 13 27 32
CLOVER     36 26 54 32
CLOVER     53 08 06 54
CLOVER     32 32 62 16
;
PROC DISCRIM DATA=CROPS POOL=YES LIST OUT=CROPCAL;
  CLASS CROP;
  ID XVALUES;
  VAR X1-X4;
  TITLE2 'CLASSIFICATION OF CROP DATA';

DATA TEST;
   INPUT CROP $ 1-10 X1-X4 XVALUES $ 11-21;
   CARDS;
CORN       16 27 31 33
SOYBEANS   21 25 23 24
COTTON     29 24 26 28
SUGARBEETS54 23 21 54
CLOVER     32 32 62 16
;
PROC DISCRIM DATA=CROPCAL TESTDATA=TEST TESTLIST;
  CLASS CROP;
  TESTCLASS CROP;
  TESTID XVALUES;
  VAR X1-X4;
  TITLE2 'CLASSIFICATION OF TEST DATA';
```

Output 16.2 Remote Sensing Data on Five Crops: PROC DISCRIM

REMOTE SENSING DATA ON FIVE CROPS 1
CLASSIFICATION OF CROP DATA

DISCRIMINANT ANALYSIS

CROP	FREQUENCY	PRIOR PROBABILITY
CLOVER	11	0.20000000
CORN	7	0.20000000
COTTON	6	0.20000000
SOYBEANS	6	0.20000000
SUGARBEETS	6	0.20000000
TOTAL	36	1.00000000

REMOTE SENSING DATA ON FIVE CROPS 2
CLASSIFICATION OF CROP DATA

DISCRIMINANT ANALYSIS POOLED COVARIANCE MATRIX INFORMATION

COVARIANCE MATRIX RANK	NATURAL LOG OF DETERMINANT OF THE COVARIANCE MATRIX
4	21.30189392

REMOTE SENSING DATA ON FIVE CROPS 3
CLASSIFICATION OF CROP DATA

DISCRIMINANT ANALYSIS PAIRWISE SQUARED GENERALIZED DISTANCES BETWEEN GROUPS

$$D^2(I|J) = (\bar{X}_I - \bar{X}_J)' \, COV^{-1} \, (\bar{X}_I - \bar{X}_J)$$

GENERALIZED SQUARED DISTANCE TO CROP

FROM CROP	CLOVER	CORN	COTTON	SOYBEANS	SUGARBEETS
CLOVER	0.00000000	4.25308108	0.86616669	2.58313162	1.48909745
CORN	4.25308108	0.00000000	1.88446483	0.73030740	2.89042690
COTTON	0.86616669	1.88446483	0.00000000	1.43466961	1.29555784
SOYBEANS	2.58313162	0.73030740	1.43466961	0.00000000	1.07646391
SUGARBEETS	1.48909745	2.89042690	1.29555784	1.07646391	0.00000000

⑩ REMOTE SENSING DATA ON FIVE CROPS 4
CLASSIFICATION OF CROP DATA

DISCRIMINANT ANALYSIS LINEAR DISCRIMINANT FUNCTION

$$CONSTANT = -.5 \, \bar{X}_J' \, COV^{-1} \, \bar{X}_J \qquad COEFFICIENT\ VECTOR = COV^{-1} \, \bar{X}_J$$

CROP

	CLOVER	CORN	COTTON	SOYBEANS	SUGARBEETS
CONSTANT	-9.79894962	-6.08308779	-9.67360774	-5.49083759	-8.01003003
X1	0.08907263	-0.04180494	0.02462407	0.00003693	0.04244951
X2	0.17378658	0.11970448	0.17595574	0.15896277	0.20987506
X3	0.11899303	0.16510688	0.15880134	0.10622011	0.06540371
X4	0.15637491	0.16768459	0.18361917	0.14132806	0.16407580

REMOTE SENSING DATA ON FIVE CROPS 5
CLASSIFICATION OF CROP DATA

DISCRIMINANT ANALYSIS CLASSIFICATION RESULTS FOR CALIBRATION DATA: WORK.CROPS

GENERALIZED SQUARED DISTANCE FUNCTION:

$$D^2_J(X) = (X-\bar{X}_J)' \, COV^{-1} \, (X-\bar{X}_J)$$

POSTERIOR PROBABILITY OF MEMBERSHIP IN EACH CROP:

$$PR(J|X) = EXP(-.5\, D^2_J(X)) \,/\, SUM_K \, EXP(-.5\, D^2_K(X))$$

XVALUES	FROM CROP	CLASSIFIED INTO CROP		POSTERIOR PROBABILITY OF MEMBERSHIP IN CROP: CLOVER	CORN	COTTON	SOYBEANS	SUGARBEETS
16 27 31 33	CORN	CORN		0.0541	0.3855	0.1956	0.2653	0.0995
15 23 30 30	CORN	CORN		0.0466	0.4341	0.1579	0.2811	0.0802
16 27 27 26	CORN	SOYBEANS	*	0.0591	0.3236	0.1506	0.3390	0.1277
18 20 25 23	CORN	SOYBEANS	*	0.0637	0.3460	0.1198	0.3645	0.1060
15 15 31 32	CORN	CORN		0.0360	0.5535	0.1317	0.2342	0.0447
15 32 32 15	CORN	SOYBEANS	*	0.0583	0.3091	0.1450	0.3762	0.1112
12 15 16 73	CORN	CORN		0.0274	0.4964	0.2044	0.1521	0.1197
20 23 23 25	SOYBEANS	SOYBEANS		0.0807	0.2672	0.1307	0.3675	0.1539
24 24 25 32	SOYBEANS	SOYBEANS		0.1091	0.2407	0.1794	0.3009	0.1698
21 25 23 24	SOYBEANS	SOYBEANS		0.0900	0.2320	0.1336	0.3696	0.1748
27 45 24 12	SOYBEANS	SUGARBEETS	*	0.1452	0.0530	0.1148	0.3075	0.3795
12 13 15 42	SOYBEANS	CORN	*	0.0330	0.4487	0.1014	0.3052	0.1117
22 32 31 43	SOYBEANS	COTTON	*	0.0898	0.2494	0.2929	0.2063	0.1616
31 32 33 34	COTTON	COTTON		0.1806	0.1530	0.2795	0.2078	0.1791
29 24 26 28	COTTON	SOYBEANS	*	0.1601	0.1838	0.1780	0.2967	0.1814
34 32 28 45	COTTON	COTTON		0.2021	0.1040	0.2851	0.1609	0.2480
26 25 23 24	COTTON	SOYBEANS	*	0.1318	0.1767	0.1418	0.3469	0.2028
53 48 75 26	COTTON	COTTON		0.3407	0.0432	0.5660	0.0288	0.0214
34 35 25 78	COTTON	COTTON		0.1389	0.0768	0.4300	0.0669	0.2875
22 23 25 42	SUGARBEETS	CORN	*	0.0870	0.2947	0.2132	0.2503	0.1548
25 25 24 26	SUGARBEETS	SOYBEANS	*	0.1219	0.1994	0.1537	0.3359	0.1891
34 25 16 52	SUGARBEETS	SUGARBEETS		0.1869	0.0874	0.1949	0.1731	0.3576
54 23 21 54	SUGARBEETS	CLOVER	*	0.4743	0.0232	0.1749	0.0694	0.2582
25 43 32 15	SUGARBEETS	SOYBEANS	*	0.1398	0.1104	0.1868	0.3144	0.2486
26 54 2 54	SUGARBEETS	SUGARBEETS		0.0483	0.0072	0.0543	0.0688	0.8214
12 45 32 54	CLOVER	COTTON	*	0.0406	0.2453	0.3648	0.1570	0.1923
24 58 25 34	CLOVER	SUGARBEETS	*	0.0977	0.0351	0.1826	0.1579	0.5267
87 54 61 21	CLOVER	CLOVER		0.8835	0.0005	0.0831	0.0043	0.0287
51 31 31 16	CLOVER	CLOVER		0.5211	0.0253	0.1254	0.1380	0.1902
96 48 54 62	CLOVER	CLOVER		0.8649	0.0002	0.1039	0.0012	0.0297
31 31 11 11	CLOVER	SUGARBEETS	*	0.1566	0.0392	0.0538	0.3424	0.4080
56 13 13 71	CLOVER	CLOVER		0.4657	0.0253	0.1707	0.0567	0.2817
32 13 27 32	CLOVER	SOYBEANS	*	0.1731	0.2665	0.1797	0.2687	0.1121
36 26 54 32	CLOVER	COTTON	*	0.1717	0.2693	0.4152	0.1091	0.0347
53 08 06 54	CLOVER	CLOVER		0.4433	0.0279	0.0929	0.1073	0.3287
32 32 62 16	CLOVER	COTTON	*	0.1378	0.3183	0.3885	0.1313	0.0240

* MISCLASSIFIED OBSERVATION

REMOTE SENSING DATA ON FIVE CROPS 6
CLASSIFICATION OF CROP DATA

DISCRIMINANT ANALYSIS CLASSIFICATION SUMMARY FOR CALIBRATION DATA: WORK.CROPS

GENERALIZED SQUARED DISTANCE FUNCTION:

$$D^2_J(X) = (X-\bar{X}_J)' \, COV^{-1} \, (X-\bar{X}_J)$$

POSTERIOR PROBABILITY OF MEMBERSHIP IN EACH CROP:

$$PR(J|X) = EXP(-.5\, D^2_J(X)) \,/\, SUM_K \, EXP(-.5\, D^2_K(X))$$

NUMBER OF OBSERVATIONS AND PERCENTS CLASSIFIED INTO CROP:

FROM CROP	CLOVER	CORN	COTTON	SOYBEANS	SUGARBEETS	TOTAL
CLOVER	5	0	3	1	2	11
	45.45	0.00	27.27	9.09	18.18	100.00
CORN	0	4	0	3	0	7
	0.00	57.14	0.00	42.86	0.00	100.00
COTTON	0	0	4	2	0	6
	0.00	0.00	66.67	33.33	0.00	100.00
SOYBEANS	0	1	1	3	1	6
	0.00	16.67	16.67	50.00	16.67	100.00
SUGARBEETS	1	1	0	2	2	6
	16.67	16.67	0.00	33.33	33.33	100.00

(continued on next page)

(continued from previous page)

TOTAL	6	6	8	11	5	36
PERCENT	16.67	16.67	22.22	30.56	13.89	100.00
PRIORS	0.2000	0.2000	0.2000	0.2000	0.2000	

DISCRIMINANT ANALYSIS CLASSIFICATION RESULTS FOR TEST DATA: WORK.TEST

GENERALIZED SQUARED DISTANCE FUNCTION:

$$D^2_J(X) = (X-\bar{X}_J)' \, COV^{-1} \, (X-\bar{X}_J)$$

POSTERIOR PROBABILITY OF MEMBERSHIP IN EACH CROP:

$$PR(J|X) = EXP(-.5\, D^2_J(X)) \,/\, SUM_K \, EXP(-.5\, D^2_K(X))$$

POSTERIOR PROBABILITY OF MEMBERSHIP IN CROP:

XVALUES	FROM CROP	CLASSIFIED INTO CROP		CLOVER	CORN	COTTON	SOYBEANS	SUGARBEETS
16 27 31 33	CORN	CORN		0.0541	0.3855	0.1956	0.2653	0.0995
21 25 23 24	SOYBEANS	SOYBEANS		0.0900	0.2320	0.1336	0.3696	0.1748
29 24 26 28	COTTON	SOYBEANS	*	0.1601	0.1838	0.1780	0.2967	0.1814
54 23 21 54	SUGARBEETS	CLOVER	*	0.4743	0.0232	0.1749	0.0694	0.2582
32 32 62 16	CLOVER	COTTON	*	0.1378	0.3183	0.3885	0.1313	0.0240

* MISCLASSIFIED OBSERVATION

DISCRIMINANT ANALYSIS CLASSIFICATION SUMMARY FOR TEST DATA: WORK.TEST

GENERALIZED SQUARED DISTANCE FUNCTION:

$$D^2_J(X) = (X-\bar{X}_J)' \, COV^{-1} \, (X-\bar{X}_J)$$

POSTERIOR PROBABILITY OF MEMBERSHIP IN EACH CROP:

$$PR(J|X) = EXP(-.5\, D^2_J(X)) \,/\, SUM_K \, EXP(-.5\, D^2_K(X))$$

NUMBER OF OBSERVATIONS AND PERCENTS CLASSIFIED INTO CROP:

FROM CROP	CLOVER	CORN	COTTON	SOYBEANS	SUGARBEETS	TOTAL
CLOVER	0 0.00	0 0.00	1 100.00	0 0.00	0 0.00	1 100.00
CORN	0 0.00	1 100.00	0 0.00	0 0.00	0 0.00	1 100.00
COTTON	0 0.00	0 0.00	0 0.00	1 100.00	0 0.00	1 100.00
SOYBEANS	0 0.00	0 0.00	0 0.00	1 100.00	0 0.00	1 100.00
SUGARBEETS	1 100.00	0 0.00	0 0.00	0 0.00	0 0.00	1 100.00
TOTAL	1	1	1	2	0	5
PERCENT	20.00	20.00	20.00	40.00	0.00	100.00
PRIORS	0.2000	0.2000	0.2000	0.2000	0.2000	