eNote 03

TRANSFORMATION OF VARIABLES FOR COMPOSITIONAL DATA-ANALYSIS

BY

George F. Hart

Compositional data, percentage data, counts, ppm and similar proportional data often need to be transformed. Data-frames should have each variable checked for error and outliers prior to any analysis.

Error checking of original variables

The Index plot can be used to show outliers and aberrant values. This imaging tool plots a graph of each sample according to its row position in the data-frame against the variable. The Indian deltas data-frame is used in the following examples [see Hart, eNote02]. This data-frame had each delta added in sequence, thus if we look at the variation in a single variable over the whole sample sequence [e.g. plagioclase] the index plot [see figure below] shows the colors representing each delta in sequence. This provides an initial view of the distribution of the variable by delta. The important axis is the vertical axis which shows the distribution of the plagioclase: in the example there is one abnormalities in the Cauvery Delta data-frame. The R function identify() determined this was sample 24. This sample must be checked prior to further analysis to determine if an error was made. If an error is suspected either the sample should be completely excluded from the data-frame or, at least, the value of the variable for that sample should be excluded by marking it as NA.

The R code for the index plot below is:

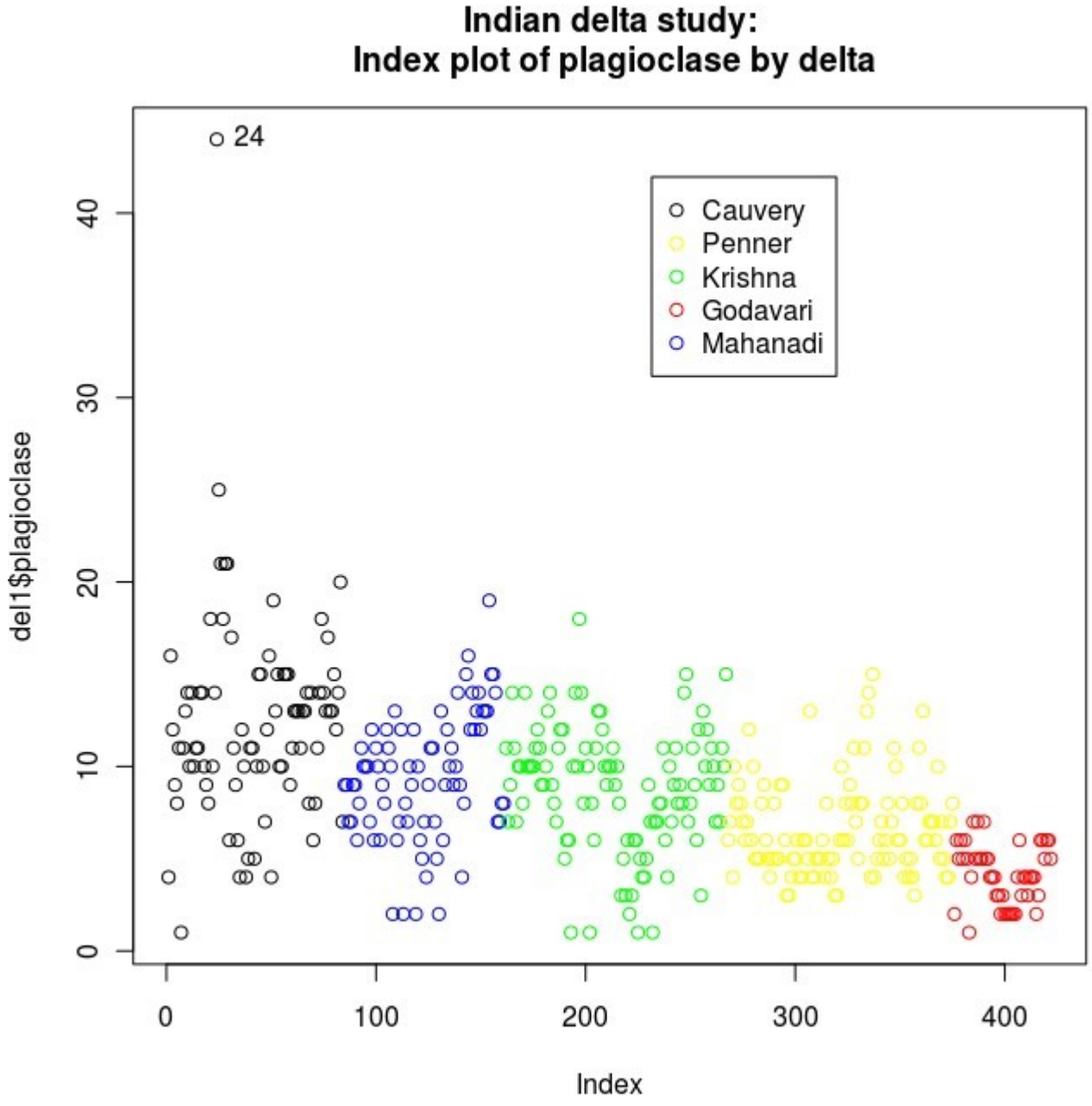
```
> my.colors<-c("black","yellow","green","red","blue") # set up a color array.

> plot(del1$pcchlorite,col=my.colors[del1$delta]) # plot the chlorite data.

> title(main="Indian delta study:\n Index plot of chlorite by delta") # add the title.

>legend(locator(1),c("Cauvery","Penner","Krishna","Godavari","Mahanadi"),pch=c(1,),col=my.color s) # add and position the legend using the locator.
```

The individual point of discrepancy is identified using: >identify(plagioclase,y=NULL) # place the pointer over the sample [see plagioclase below where point 24 is an outlier and possibly an error].



Tests for normality of variables

Prior to the application of any statistical procedure the raw variables must be examined and their distribution understood. Those variables that need transformation should be transformed with purpose in mind. To do otherwise can lead to a poor analysis of the data-frame. This is important for multivariate and multi-population analysis as well as univariate and bivariate analysis.

The assumption of normality is important because it is a requirement of many statistical tests. The normal distribution is a reasonable model for many variables in the natural sciences. The central limit theorem shows that as the sample size gets large, many of the sample summary statistics, such as the sample mean, behave as if they are from a normally distributed variable. Thus it is assumed that parametric tests or statistical models have associated errors that follow a normal distribution. EPA Report EM 1110-1-4014, 31-Jan-08 [p:F-1] points out: "*statistical tests for normality do not actually demonstrate normality but the lack of normality. They rely on the probability a given data set is normal (e.g., statistical software typically reports a "p value" for the hypothesis that the population distribution is normal). If the probability is low (e.g. p < 0.01), one rejects the assumption of normality, that is, one concludes, based upon weight of evidence, that the data set is not normal. However, if the assumption of normality is not rejected, then, strictly speaking, the statistical test is inconclusive; the data may or may not be normal. This constitutes an additional reason to visually examine the data set for normality and to decide whether to proceed with a statistical test that requires normality. In practice, if the assumption of normality is not rejected and graphical plots suggest normality, the statistical tests that rely upon normality are typically used*" , and "*The assumption of normality should not be rejected on the basis of a statistical test alone. In particular, when a large number of data are available, statistical tests for normality can be sensitive to very small (i.e., negligible) deviations in normality. Therefore, if a very large number of data are available, a statistical test may reject the assumption of normality when the data set, as shown using graphical methods, is essentially normal and the deviation from normality too small to be of practical significance.*"

In the initial analysis of variables a good procedure is to examine each variable in the data-frame for normality by the eye-ball method using box-plots of each variable, followed by a test for **skewness**, **kurtosis**, a **Q-Q plot** and calculation of a test for normality, especially the **Shapiro-Wilks test** [for sample sizes <=50] and the **D'Agnostino's test [for samples sizes 50 to 1,000]**. Examining these comprehensive results allows a better view of the distribution of each variable prior to performing any transformation. This approach to normality testing

combines graphical tests, and statistical tests. The Shapiro-Wilks test statistic is based on H_0 that the data are normally distributed. If the the sample size [n] is larger than 50 the D'Agnostino's test is calculated. The D'Agnostino statistic measures the linearity of the points on the normal probability plot. "*If the normal probability plot is approximately linear (the data follow a normal curve), the correlation coefficient will be relatively high. If the normal probability plot contains significant curves (the data do not follow a normal curve), the correlation coefficient will be relatively low*". [EPA, 2008] .

Skewness and kurtosis

A test for skewness and kurtosis may not be part of the R functions but it is easy to write in R [the code is from **Crawley, 2007**, p:285-286]. Press enter at the end of each input line. Remember the data-frame must be attached to use the variable directly.

Crawley's skewness function

```
skew<-function(x)
{m3<-sum((x-mean(x))^3)/length(x)
s3<-sqrt(var(x))^3
m3/s3}
>skew(clay)
-0.7940148
```

Crawley's kurtosis function

```
kurtosis<-function(x) {
m4<-sum((x-mean(x))^4)/length(x)
s4<-var(x)^2
m4/s4-3}
>kurtosis(clay)
0.5484766 # the clay kurtosis is neither platykurtic nor leptokurtic and thus kurtosis is normal.
```

Alternatively we can use the library(moments):

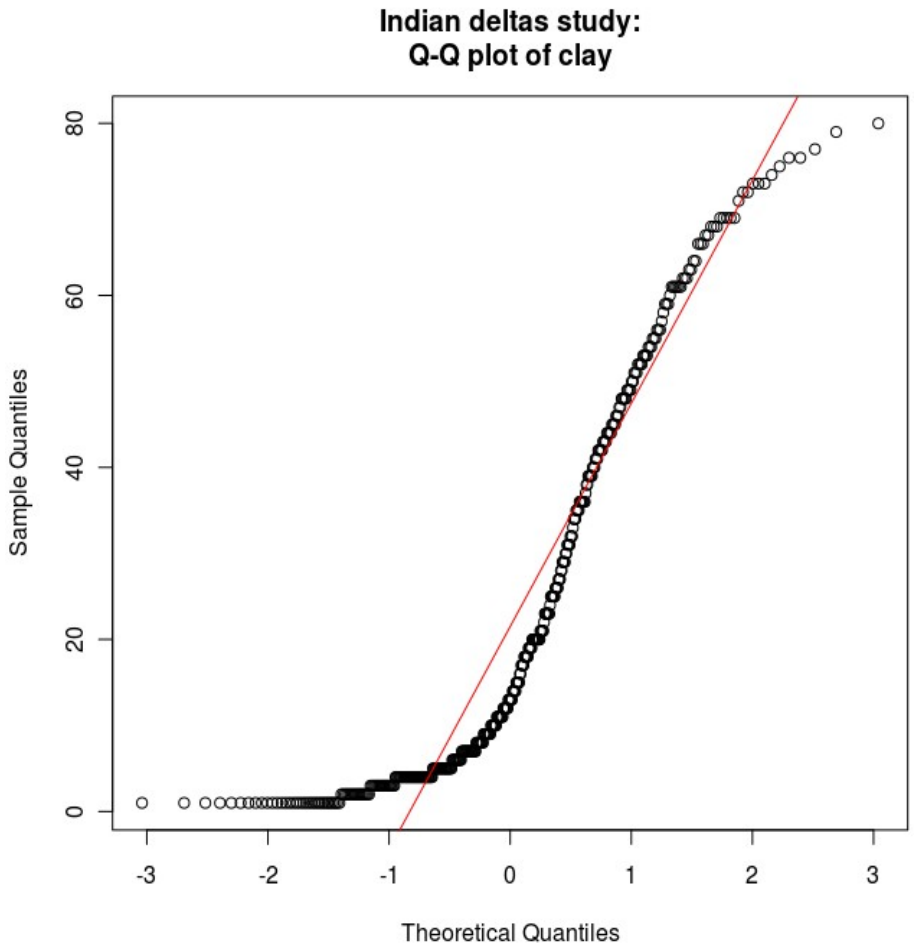
>skewness(clay,na.rm=TRUE), and
>kurtosis(clay,na.rm=TRUE) # Pearson's measure of kurtosis

Q-Q plot

We use the quartile-quartile plot to examine a variable against it's standard normal distribution [N(0,1)] to assess the goodness-of-fit. If the two distributions agree and lie on a straight line in which x = y the data has normal distribution. If the line is straight but x != y then a linear transformation can be applied. This is a common graphical method for assessing whether or not the distribution of a variable is Gaussian.

>qqnorm(clay,main="Indian deltas study:\n Q-Q plot of clay") # the data distribution.

>qqline(clay,col="red") #use a red line for the normal distribution.



Shapiro-Wilks test

The **Shapiro Wilk test** is used for sample sizes of 50 or less. The null hypothesis of the Shapiro-Wilk test is that the sample is taken from a normal distribution, thus P < 0.05 for **W**

rejects the hypothesis of normality. Samples which fail the test are difficult to analyze with parametric statistical methods. Many statistical studies indicate this is the most reliable test for non-normality for small to medium sized samples [Conover, 1999; Shapiro and Wilk, 1965; Royston, 1982a, 1982b, 1995]. Nevertheless the results are not clear evidence of normality or non-normality but should be assessed along with other tests.

```
>shapiro.test(clay)
```

>W=0.957, p-value < 2.2e-16 # for alpha 0.05 the p-value is very small so the NULL hypothesis that the the sample comes from a normal population, is rejected.

It must be remembered that using the Shapiro-Wilk test statistic is not the same as looking at confidence intervals. Confidence intervals pertain to values inside the probabilities, whereas the normality test pertains to values below p=0.05 or p=0.01 i.e a sample can fail the test at p=0.01 [1 in 100 the sample is not normal] but pass it at the higher probability of p=0.05 [1 in 20 that the sample is not normal].

D'Agnostino test

The D test is a skewness test and used as an alternative to the Shapiro-Wilks test when **n** is between 50 and 1000. The D'Agnostino statistic measures the linearity of the points on the normal probability plot. "*If the normal probability plot is approximately linear (the data follow a normal curve), the correlation coefficient will be relatively high. If the normal probability plot contains significant curves (the data do not follow a normal curve), the correlation coefficient will be relatively low*". [EPA, 2008] . The usage is:

```
>agostino.test(quartz, alternative = ("two.sided", "less","greater") #two.sided is the default.
> library(moments)
> agostino.test(quartz)
```

Typical output is:

D'Agostino skewness test data: quartz skew = -0.343, z = -1.879, p-value = 0.06032
alternative hypothesis: data have a skewness.

In this case the p-value is greater than 0.05 therefore the variable is accepted as normally distributed.

Kolmogorov-Smirnov Goodness-of-fit test

The K-S test is used to decide if a sample data set comes from a population with a particular distribution: see www.itl.nist.gov/div898/handbook/eda/section3/eda35g.htm.
An adaptation of the K-S test is the **Lilliefors test** which It is used to test the null hypothesis that the sample data came from a normally distributed population, when the null hypothesis does not specify the expected value and variance of the distribution: see http://en.wikipedia.org/wiki/Lilliefors_test.

TABLE OF TEST STATISTICS

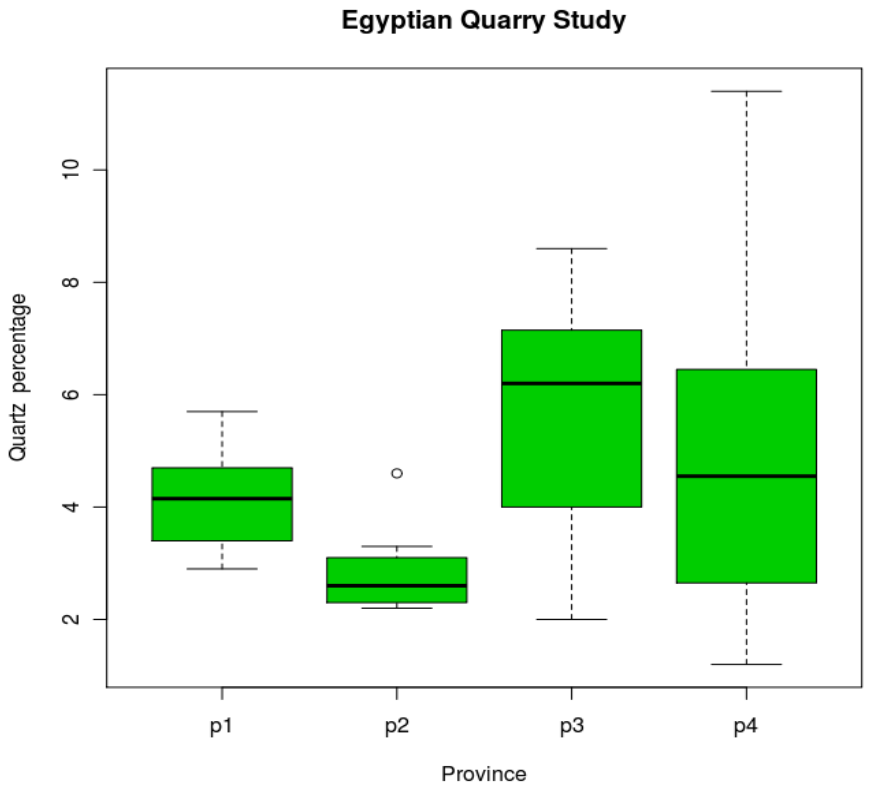
[Source EPA]

Test	Sample Size	Recommended Use
Shapiro-Wilk Test D'agnostino Test	# 50 < or = 1000	Highly recommended.
Filliben's Statistic	#100	Highly recommended but difficult to compute.
Geary's Test	> 50	Useful when tables for other tests are not available .
Studentized Range Test	# 1000	Highly recommended if the data are symmetric, tails of the data are not heavier than the normal distribution, and there are no extreme values.
Chi-Square Test	Large	Useful for grouped data and when the comparison distribution is known. May be used for other distributions besides the normal distribution

Box plot

In the initial analysis of variables a good procedure is to examine each variable in the data-frame for normality by the eye-ball method using box-plots of each variable. The eye-ball method using box plots can be used to quickly examine replicate samples to see how well they represent a sample taken from the single location. The box plot shows on a single chart the range, maximum and minimum value, median, skewness, critical limits and outliers. In the example below Province is designated a factor:

```
>ebentonites$Province<-factor(ebentonites$Province) # make Province a factor.
>boxplot(ebentonites$quartz~ebentonites$Province, main=' Egyptian Quarry Study',
ylab='Quartz percentage', xlab='Province')
```



P2 shows an outlier and the narrowest range, P1, P2 and P3 are all skewed, P4 is not skewed and shows the broadest range.

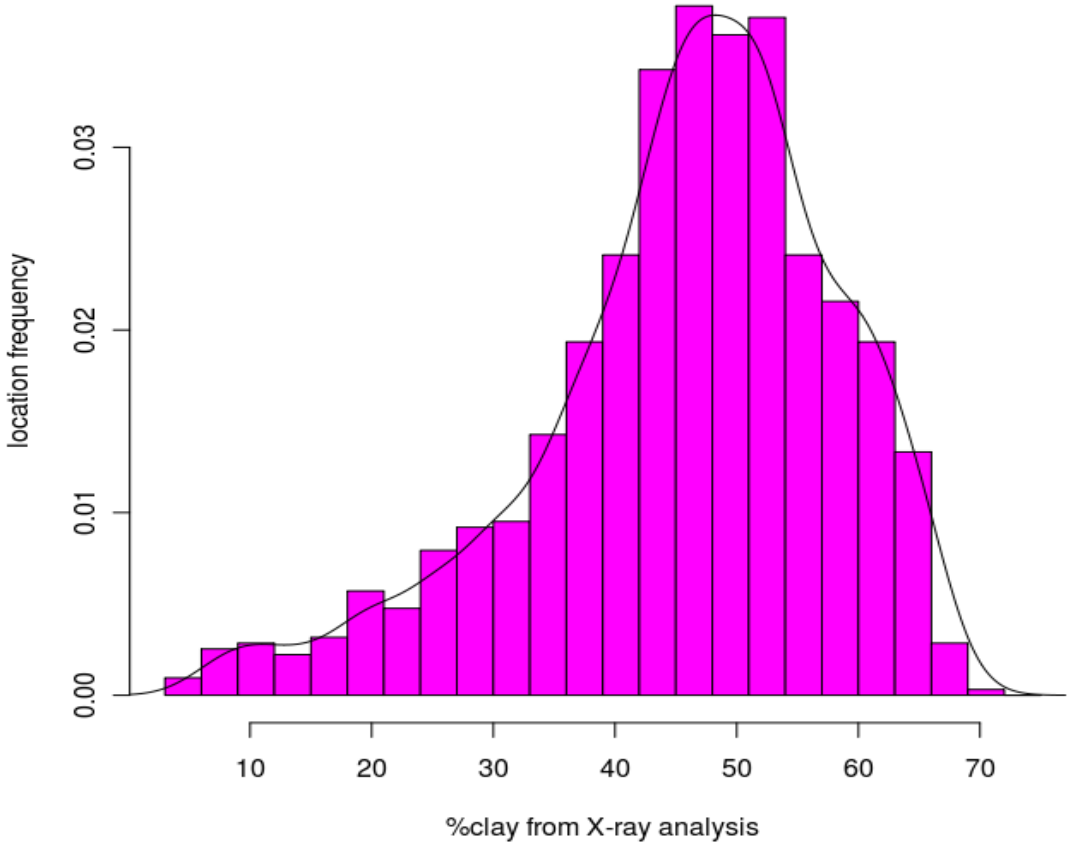
Histogram with density plot

In addition to plotting a simple histogram there is a function in the MASS library called **truehist()** which allows a density function [Gaussian] to be overlain on a histogram.

```
>truehist(clay, breaks=seq(3,75,3), xlab="% clay from Xray analysis", ylab="location frequency", col=22)
```

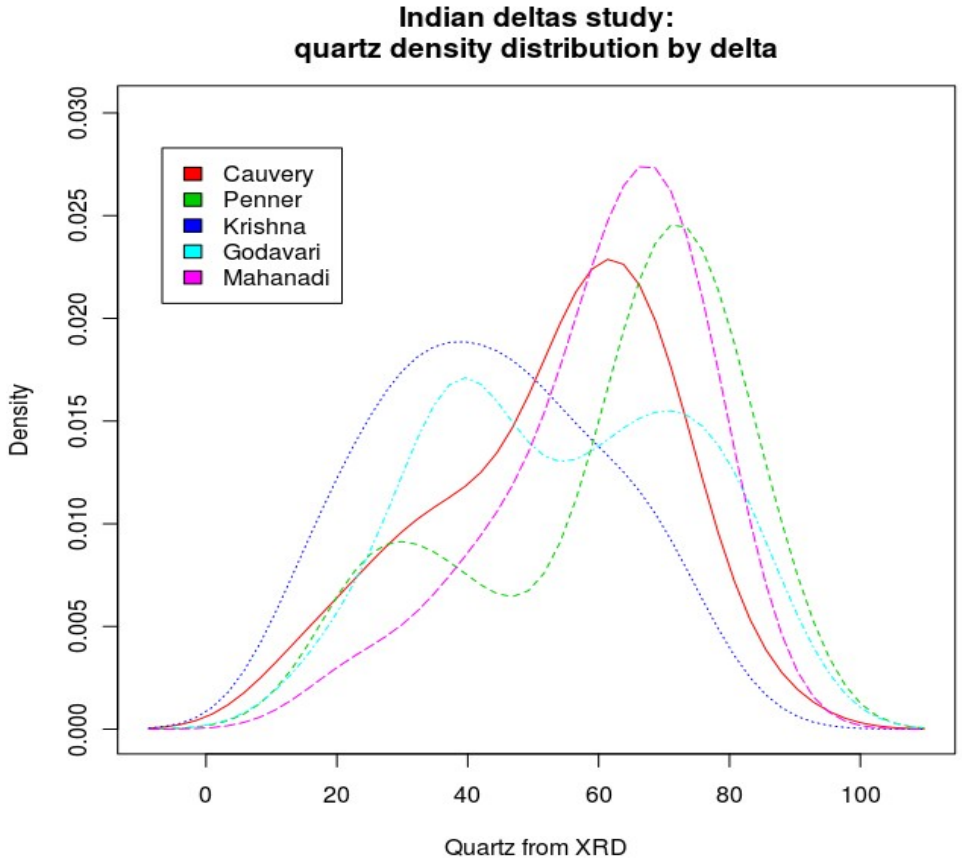
```
>lines(density(clay,n=1051,na.rm=T), col="red")
```

overlay a density function based on n=1051, removing all of the **0** values.



Density plot

```
>library(sm) # load the non-parametric smoothing lib)
> delta.f <- factor(delta, levels=
c("Cauvery","Penner","Krishna","Godavari","Mahanadi"),labels = c("Cauvery", "Penner",
"Krishna", "Godavari", "Mahanadi"))  # create value labels
> sm.density.compare(quartz,delta,xlab="Quartz from XRD")  # plot densities
> title(main="Indian delta study:\n quartz density distribution by delta") # add title
colfill<-c(2:(2+length(levels(delta.f))))  #specify the color fill
legend(locator(1), levels(delta.f), fill=colfill) # add legend via mouse click
```



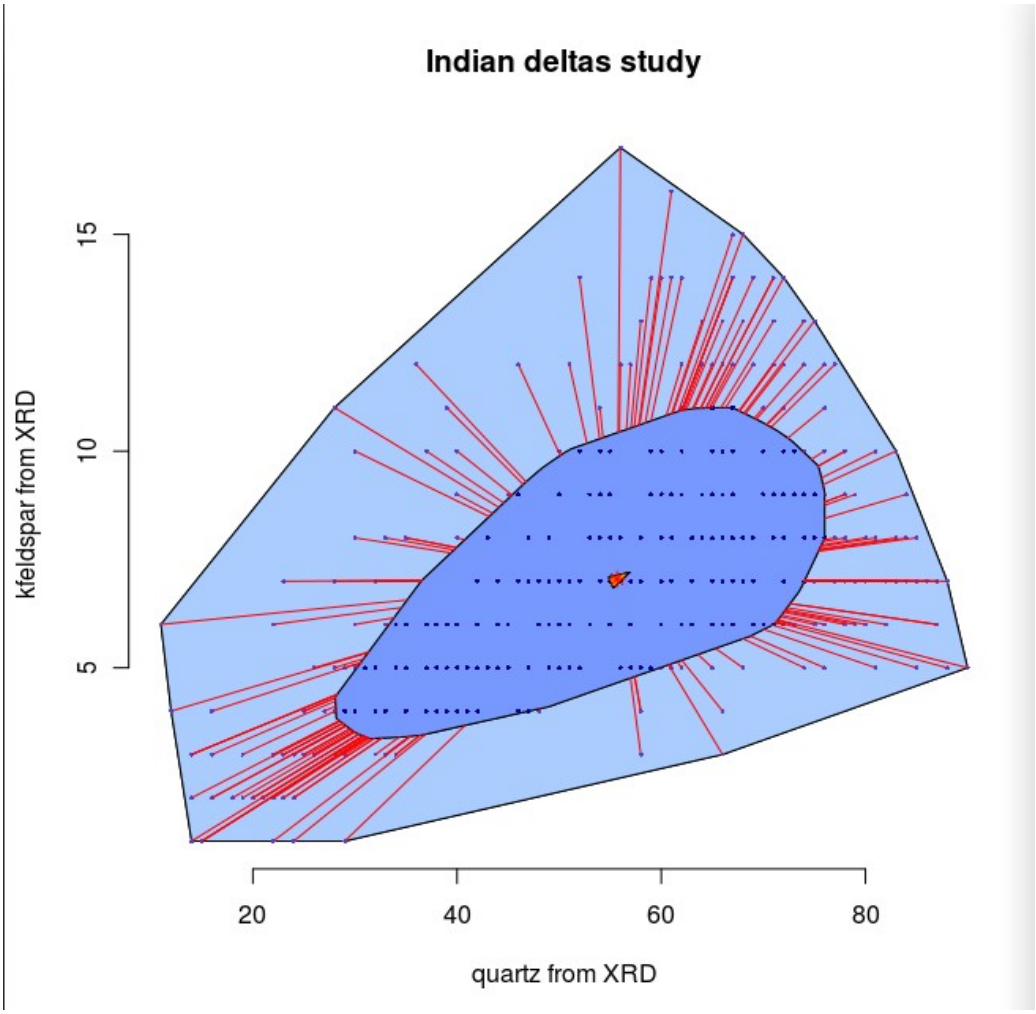
Bag plot

A bagplot is a bivariate version of a boxplot in which 50% of all points are in the central bag and the bivariate median is approximated. Outliers are displayed joined by lines to the median, and a fence separates points inside the bag from those outside the bag.

Using:

>library(aplpack)

> bagplot(quartz,kfeldspar, xlab="quartz from XRD", ylab="kfeldspar from XRD", main="Indian deltas study")



TRANSFORMATIONS

The main reason for transforming data is that many statistical tests assume that the variables are normally distributed. Unfortunately true normality is rare in geological data. Fundamentally, mathematical transformations should be applied only to variables measured on the **ratio scale** of measurement. The ordinal scale can be transformed by any increasingly monotonic function. The **interval scale** allows multiplication by a constant, and although moving from an interval to a ratio scale of measurement does not add much to the statistics that can be used it does allow transformations. Using the **ratio scale** opens up the power of mathematical techniques of analysis. Numerals from data on a ratio scale have all the properties of real numbers and can be raised to powers, have roots extracted, and reciprocal and logarithms taken. Thus if a variable is not normally distributed then it can be transformed to a form that is normally distributed and analyzed using the powerful procedures of Gaussian statistics and the Normal curve.

Transformations that are commonly used for improving the normality of variables do alter the data at a fundamental level, especially if the data set has many values around zero [exceptions are the square root and arcsine root transformations which work well for low counts and zero's]. However, if we accept that it is desirable to manipulate the data-frame prior to analysis, so that compositional values are switched from the sample space to the real number space, then the computationally simplest method that is constrained by mathematical theory should be employed.

Many scientists have doubts about applying simple unconstrained statistical analysis directly to compositional datasets (Aitchison, 2003) because, as Chayes noted: "A basic problem is "the sum of the covariances of each variable is exactly equal to its variance, and opposite in sign". Aitchison, 1982, 1986 and Davis, 2002, along with many others, have addressed this problem, and Davis notes "at the present time there is no completely satisfactory way of evaluating the strengths of correlations between variables in closed data sets."

Percentage and proportional data tend to be highly skewed and this can cause spurious correlations amongst variables and many scientists have doubts about applying simple unconstrained statistical analysis directly to compositional data-frames. As a result using untransformed variables, especially in multi-variate analysis, draws much criticism from statisticians because of the constant sum problem that occurs with proportional data-frames. ALL methods of solving the closure problem draw criticism unless stringent mathematical assumptions are met. Many Natural Science data-frames are constrained by this constant sum problem because they are formed from proportional counts, as in palaeobiology; or,

are semi-quantitative data on mineral composition expressed as a raw percentage or normalized to 100%, as in X-Ray diffraction.

On the other hand, there are numerous studies, especially in palaeontology, where interval and ratio data-frames based on proportional counts that sum to 100%, are used to draw interpretable and meaningful geological conclusions: unfortunately it seems that many of the statistical procedures applied were inappropriate for the data-frame. Nevertheless, good geological conclusions were derived and proved successful tools in geological analysis. The interpretation of such data demanded a lot of prior geological insight to derive conclusions from the results.

The work of Aichison [1981, 1882, 1983, 1984, 1986] changed the approach to the problem. Aitchison's argument that compositional data with N-variates constrained to a constant sum create a N-1 dimensional sample space [a simplex] led to his transformation of simplex space to real space using a log-ratio transform. The log-ratio transformation maps the k-dimensional simplex to the k-dimension real place. Aitchison's insight was in "*recognizing that it is the relative magnitudes and variations of components, rather than their absolute values, that provide the key to analyzing compositional data*" ..... *"Rock (1988, p. 203) described some of the problems that arise in treating compositional data with conventional statistical techniques: (a) trends and clusters on petrological ternary and principal components diagrams can have little or no geological significance; (b) skewness and leptokurtosis, properties of the shape of a distribution defined by a probability density function (e.g. the normal distribution), can be induced by closure; (c) dendrograms produced by cluster analysis can be severely biased; d) results from discriminant analysis are likely to be illusory; (e) any correlation coefficient will be affected to an unknown degree by spurious effects induced by the constant sum constraint; (f) the results of tests of significance will be intrinsically flawed since they arise from techniques applied to data for which they were never designed to be used."* [Pawlowsky-Glahn and Egozcue, 2006]

Xie et al [2003] note *"Although no universal consensus currently exists within the literature, Aitchison's log-ratio transformation is applied to geochemical datasets more frequently than any other and has been shown to reliably account for non-normality and closure (Verrucchi and Minissale 1995; Reyment and Savazzi 1999; Cullers 2000). Because the absolute magnitudes of compositional variables are ratios to a common sum, Aitchison proposed to use relative magnitudes by calculating the ratio of each compositional variable compared to a single variable that functions as a constant divisor. Further, by taking the logarithm of the ratios, the transformed values can vary over the entire real number range, rather than being restricted to the range from zero to 100%. Thus, this transformation makes the application of*

conventional statistical techniques more justifiable". Unfortunately, because of the large number of zero values that often occur in natural science datasets a log-ratio transformation does not solve the closure problem. Moreover, using one of the variables as a universal divisor eliminates that variable from analysis and this is itself a problem because the variable used often is one of sufficient value to be of interest. Zero's are not allowed in the Aitchison's log-ratio transformation, therefore zero's should be replaced by values for the lower limit of detection of the particular variable in the analysis. For example, in XRD compositional data this means the smallest amount of a given phase that can be identified. Connolly [2010] notes the detection limits are dependent upon the square root of the count time, which is dependent upon the "*counts produced by a phase of interest at a particular concentration."* Further, *"it is relatively easy to calculate the lower limit of detection for that phase."* The lower limits of detection will be different for each phase in each analysis.

Besides the log-ratio transformation Aitchison (1983) suggested the use of the additive log-ratio transformation

>alr(x) = [log(x1/xD) log(x2/xD) ... log(xD-1/xD)] , and, the centered log-ratio transformation

>clr(x) = [log(x1/g(x)) log(x2/g(x)) ...log(xD/g(x))]

where g(x) denotes the geometric mean of the D components of x.

Later, Egozcue et al [2003] introduce the isometric log-ratio.

Lee and Bacon-Stone have noted "the centered log-ratio transformation has the disadvantage that the covariance matrix formed based on such transformation is singular, while the operational problem of the additive log-ratio transformation is that a common divisor has to be chosen. As proven in the monograph of Aitchison (1986), the choice of common divisors would not affect the results of analysis due to scale invariance property, so the choice could be arbitrary, but the clear disadvantage of the additive log-ratio transformation is that the chosen common divisor could not be used in the analysis. Therefore, some geologists resist the log-ratio transformation and continue to analyze the data with the pathological approach (Aitchison, 2003). To allow consensus between statisticians and geologists, the selected common divisor for the log-ratio transformation should be of moderate abundance and relatively small variance."

Osborne [2002] in a very clear discussion of the issues of transformation covers the square root and the log transformation of variables and notes that for negatively skewed variables special pre-processing is necessary. For negatively skewed variables Osborne recommends first reflecting the distribution, adding a constant to bring it to 1.0, applying the transformation and then reflecting it again to restore the original order. "*To reflect, one multiplies a variable by -1, and then adds a constant to the distribution to bring the minimum value back above 1.0. Then, once the inverse transformation is complete, the ordering of*

the values will be identical to the original data.", Osborne, 2002.

The **square root transformation** has one important problem in that numbers between 0.00 and 0.99 behave differently from numbers of 1.00 and above. Prior to applying a square root to the numbers of a variable the values must be examined to see if they represent both groups. If both groups occur then add 1.00 to the value to make the number set consistent, if only one group occurs i.e. all values are between 0.00 and 0.99, or are 1.00 and above, then do nothing. Adding a constant only changes the mean, not the variance, kurtosis or skewness. Negative numbers cannot have their square root taken so a similar logic applies i.e. add a value to each variable number equivalent to the positive value of the largest negative number [if the largest negative is -22 then add +22 to all numbers in the variable. The square root transformation is effective for positively skewed distributions and can be used on count data.

The log transformation has a different set of problems, but it is effective for positively skewed distributions. Any number less than 1.00 or that is negative must have a value added that will bring its value to above 1.00 or above. The next problem is which base to choose? If ranges are large the base 10 is used, and if ranges are small the base 2 or base **e** [92.7182818] is preferable because they pull the extreme values in less severely.

Some scientists have suggested using the Freeman-Tukey transformation which is **arcsin(sqrt(x / n+1)) + arcsin(sqrt(x+1 / n+1))** in which the variance depends only on the denominator of the proportion in question...something that can be used to advantage. *"Arcsine or angular transformations have been used for many years to transform proportions to make them more suitable for statistical analysis. A problem with such transformations is that the arc sines do not bear any obvious relationship to the original proportions. For this reason, results expressed in arc sine units are difficult to interpret. A simple linear transformation of the arcsine transform has been used to produce values that are numerically close to the original percentage values over most of the percentage range while retaining all of the desirable statistical properties of the arcsine transform." [anon].*

One interesting aspect about percentage data is that it provides information on not only on how often a value occurs but also how often it did not occur. Conventional frequency data provides information only on the former. Because of this duality a percentage is binomially distributed with two outcome vectors. One vector is the presence number of the variable, the other is the absence number of the variable. If we use percentage data as a dependent variable in regression, the data is transformed to confine the projected value within 0-1, and

make it closer to a normal distribution. The model used is a general linear model [glm] for binomial outcome, in which the samples are weighted by the sample size and arc sine transformed to make the error distribution normal. This lead to the logit model. The logit transformation assumes the data is binomial, and the logistic model for p as a function of x is: $p = \exp(a+bx)/(1 + \exp(a+bx))$. This is obviously non-linear. To linearize it, consider the odds p/q (where q is 1-p): $p/q = \exp(a+bx)$. Or: $\ln(p/q) = a + bx$ in which ln(p/q) is the logit transformation of p .

R does not simply do a linear regression of ln(p/ q) against x. It also handles non-constant binomial variance , logit(p) going to -∞ and +∞, and differences between sample sizes using weighted regression.

Finally, Maarten, Cox and Jenkins have written the -betafit- function in R as another method for analyzing percentage data. This can fit a regression through the untransformed proportions by assuming that the proportions follow a beta distribution: providing for variance stabilization.

For **univariate** and **bivariate** analysis the following procedure for dealing with skewness and normality problems is used when analyzing compositional variables.

The percentage data is first converted to proportions by dividing them by 100. The square root of the proportion is computed, and then the inverse sine (arcsin function) derived for the resultant proportional values.

The sample statistics can be calculated, such as mean, variance, stand deviation and confidence intervals, and back transformed for interpretation by simply reversing the original transformation. To undo the arcsine, the sign function is used, and, to undo the square root the value was squared. This methodology aims, especially to take care of the closure problem. The interpretation of back-transformed statistics must be done with care as often the back transformed data are biased estimates of such values, as means, variances and regression coefficients. The confidence limits are always questionable whether the data is transformed or not! The arcsine root transformation used in proportional variable transformation attempts to remove the skewness. The variable values should be between 0.00 and 1.00 and therefore percentage data is divided by 100 prior to transformation.
An example in **R:** using the variables in columns 5 - 18, in a data-frame called **mo** the transformation is: `mosqrt<-sqrt(mo[,5:18]/100)`. This sets up a new data-frame containing the square root of the independent variables divided by 100. The data-frame is recalculate as the inverse sine [arcsine] of the variables, which in R is: `moarc<-asin(mosqrt)`. Each observation **must** be weighted by the number in the denominator of the proportion prior to

analyzing the variable motrans<-(mo[,19:32]*100). The standard deviation is the root-mean square error derived from the analysis of the transformed variable, which takes into account the weighting value. To back transform the observed effects and confidence limits, add the effect to the mean, take its sine, then square it and multiply by 100 [Hopkins, 2009].

For multivariate analysis where multi-variate normality is needed the Aitchison transformation is used. The procedure is applicable under conditions that involve log-ratio transformation to convert the vector from simplex space[Sd] to real number space [Rd-1], using the transformation of the predictor variable vector V1 → V2. Where V2 becomes:

$$[\log(X_1/X_d), \log(X_2/X_d), \log(X_3/X_d), \ln(X_4/X_d), \ldots\ldots \log(X_{d-1}/X_d),]$$

with Xd being any suitable variable in the data frame. An example in R in which feldspar is used as the divisor and the data-frame is called mo is:

```
>moP$Xa <= log(moP$X/moP$Feldspar)
```