Manual to ANATIS - English version   Jul-21-2012

The program collection supports three main tasks:
- statistics for distribution of points in a flat plane (foci), by using the domain method
- statistics for surface areas of delineated objects in the flat plane (contours)
- evaluation of relations between data of both types for one or a series of analogous planes

This manual is an extension to the help and information in the program itself.

TABLE OF CONTENTS

1   Points in a flat plane (foci)

1.1   Input

Input files for points are plain text files. Normally, such files have an extension (last part of the filename) .txt and can be made and edited by most database and word processor programs, and simple edit programs such as the Windows program Notepad.
- Each line starting with an at-sign (@) is considered a comment line, no further processing is done.
- The text of the first line not starting with a number, an equals sign (=) or a left square bracket ([) is taken as the unit of measure for the points; example: µm.
- All the other lines are considered to have data, two coordinate numbers, x and y, separated by a separator token. Fractional numbers have a decimal point, not a decimal comma. Allowed separators are , / : ; = \ ` | ~ or a tab character, a blank following a number, or a minus sign embedded in blanks (-). Multiple blanks preceding or trailing a number are ignored, as are all elements trailing the y coordinate. Missing numbers are taken as 0.
- As a supplement to these rules, one single separator line is allowed, a line starting with an equals sign (=) or a left square bracket ([). Then, the coordinates preceding this separator line are interpreted as the coordinates of the corner points of the flat plane, the coordinates following the separator line as the foci in the flat plane. If there is no separator line, all coordinates are of foci.
    For an example, see the file FocX.TXT.

The foci are numbered sequentially from 1 on, as are the resulting domains. The sequence can be made relevant, as explained later.

1.2   Operations

Calculation of domains[1] from foci. Domains are regions closest to the respective foci. The surface area of such a domain may be considered the 'region of influence of the focus', its inverse as a measure of the local density of the foci. The density distribution can be shown, in a separate window, but that might be a slow process depending on your computer system.

[1] *Reference: Hoofd et al. (1985). Variability of intercapillary distance estimated on histological sections of rat heart, in: Adv Exper Med Biol 191: 239–247.*

1.3   Statistics

Sampling of domain surface areas or domain perimeter lengths into a histogram. Special attention is paid to domains at the border of the field, i.e., sharing a border line with a field border line. Such a bordering domain line will be uncertain, since there might be foci outside the field being closer to some points inside the domain, contrary to the definition of domain border line.

As a consequence, three statistics cases are discerned:
 - Incorporate all domains, assuming that the end result will not be affected significantly.
 - Omit all border domains.
 - Omit half of the border domains. For a rectangular field, it has been demonstrated, that leaving out domains at two of the four sides is considered statistically the most correct[2].

'Field data' calculates mean, standard deviation (SD), median and standard deviation of the ^{10}log of the surface areas (^{10}logSD) as well as of the perimeter lengths. The term 'median' follows the existing literature but can be statistically incorrect, 'geometric mean' being the correct term, the Nth root of the product of the N individual values. It is the median of a presumed lognormal distribution[3].
From the surface areas, distances can be obtained by a variety of methods. Many methods are based on taking the radius of a circular area of the same size as the domain. So, the above calculations are also done for these radii. In addition, the mean surface area data are recalculated in terms of radius data. The ^{10}logSD can be taken as a measure of heterogeneity in focus density[1]. Some publications use the ^{10}logSD of surface areas, some of radii. The ^{10}logSD of the surface areas is twice that of the radii.

[2] *Reference: Egginton, S., Ross, H.F. (1989) Influence of muscle phenotype on local capillary supply. Adv. Exp. Med. Biol. 247: 281-291.*

[3] *Reference: van Haelst et al. (1985) Lognormal distribution of capillary domains in rat myocardium. J Physiol (London) 366: 114P.*

2   Objects in the flat plane (contours)

2.1   Input

Input files for contours are plain text also, similar to input files for points, but for the following differences and additions:
 - Data lines of coordinates are considered to be corner points of the contour. Consecutive corner points are connected by straight lines yielding a polygon in the program. The last corner point is connected to the first one, so that the starting point should not be given twice, at the end also.
 - Adjacent polygons are separated by a separator line, here, a line starting with the text [Delimiter] or at least starting with a left square bracket ([) or with an equals sign (=).
 - In case that the first polygon is by far the largest, it is considered to be the enveloping field of the flat plane section of the contours.

For an example, see the file CntX.TXT.
The contours are numbered sequentially from 1. An enveloping field polygon has no number. Contours should not overlap; and they need not cover the entire field.

AnaTis decides between foci or contour dataset depending on the number of separator lines. Two or more separator lines implies a contour dataset, else, a foci dataset.

2.2   Statistics

Sampling of contour surface areas into a histogram. From the original material, the user should know if the complete outline of a contour were known, and AnaTis does not perform any test for a contour to be at the border of the field. If desired, two input contour files can be made, one containing complete contours only and one also incorporating the contours with a partly uncertain outline, and results compared.

'Field data' calculates mean, standard deviation (SD), 'median' and standard deviation of the ^{10}log of the surface areas (^{10}logSD) as well as of the perimeter lengths. As for domains, the ^{10}logSD value can be taken as heterogeneity in contour size, but, different from domains, there is not a statistical foundation here.

3   Relations between foci en contours

3.1   Basic processing

Relations between foci and contours are derived from overlaps between contours and domains[4]. Such overlaps are polygons too, resulting in the following output possibilities.

- Overlap corner points, numbered sequentially from 1 on; this number follows the text [Delimiter]. Corner points files are written or saved virtually similar to input files for contours.

- Surface area values, for each overlap, numbered sequentially for area number, overlapping domain number followed by respective fraction of the domain surface area, and overlapping contour number followed by respective fraction of the contour surface area.

- Contour overlaps, for each contour sequentially, the number of overlapping domains (n), the sum of fractions of respective domain surface area (f), and this fractions sum divided by the surface area of the contour. The sum of fractions of respective domain surface area (f) can be interpreted as the net number of foci belonging to a contour taking into account, that a focus can belong to several contours, and consequently counts less than 100% for each individual contour – as explained below. The last number on the output line is helpful for relating the number of foci to the contour size.
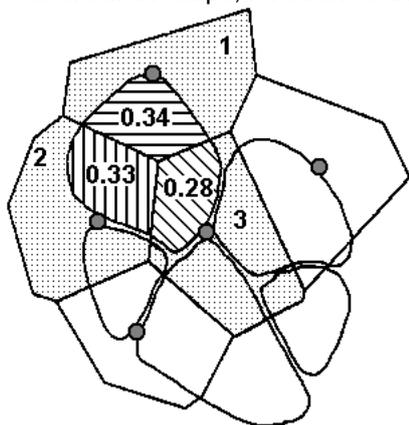
The derivation of n and f for contours is represented in the adjacent figure. The contour with hatchings has overlaps with n=3 domains, numbered 1, 2 and 3. The overlap surface area with domain 1 is 34% of the surface area of domain 1, for domain 2 this value is 33%, and for domain 3 it is 28%, resulting in a summed value of f= 0.34+0.33+0.28= 0.95.

- Domain overlaps, for each domain sequentially, the number of overlapping contours (n), the sum of fractions of respective contour surface area (f), and this fractions sum divided by the surface area of the domain. The sum of fractions of respective contour surface area (f) can be interpreted as the net number of contours belonging to a focus.

The derivation of n en f for domains is represented in the adjacent figure. The domain with hatchings has overlaps with n=5 contours, indicated A-E. The overlap surface area with contour A is 6% of the surface area of contour A, 29% for contour B, 40% for contour C, 21% for contour D and 24% for contour E, resulting in a summed value of f= 0.06+0.29+0.40+0.21+0.24= 1.20.
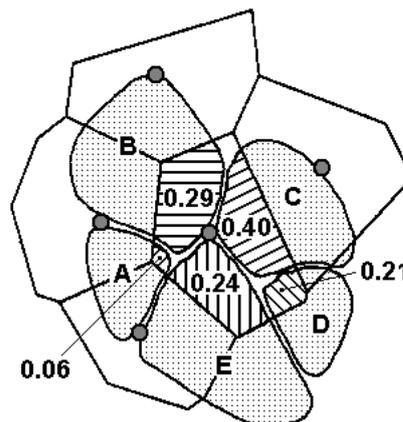
[4] *Reference: Egginton et al. (1987) Morphometric analysis of sparse capillary networks. Adv. Exp. Med. Biol. 215: 1-12.*

3.2   Statistics

Statistics involving overlaps are for sampled data only, where individual cases have to be added to these sampled data first. Sampled data can contain multiple fields but also selection of a single field is possible. There is no built-in limit for sampled data; it is limited by computer capacity only.

When there are different contour types, these can be assigned to different contour input files, each combined consecutively with their respective domains – which might be the same domains - and stored into the sampled data, under different label names. Later, the labels of the desired type can be selected together. Note, that this presupposes an adequate naming strategy!

When there are foci of different types, there is no possibility to assign these to different files since all foci are needed to calculate the domains. However, domains can be stored in groups under different label names, e.g., domains 1-7 under label Lbl_1, 8-27 under label Lbl_2 and 28-41 under label Lbl_3, through the 'Calculations – Add – Selection...' menu entry. Order the foci accordingly in the input file.

4   Sampled data

4.1   Basics

Three types of sampled data are discerned and handled separately. Calculated results are represented both graphically and in tabulated form.

- Contour surface areas;
- Domain surface areas;
- Overlaps between contours and domains.

4.2 Statistics

For the first two sampled data types, the same calculations can be done as described above for single cases, except output of corner points etcetera.

From overlap data, relations between foci and contours can be derived, assuming a linear relation between either the surface areas or the perimeter lengths of either contours or domains and the number of other type 'belonging to them', as described above. Below, this is detailed for surface area A but it is handled identically for perimeter length P. The linear relation is written as:

$$y = y^0 + b\,(A - A^0)$$

where A is the domain surface area or contour surface area, and y is either the number of overlaps (n) or the summed overlap fraction (f). The parameters y^0 and A^0 are midpoint values (the 'centre' of the line) and b is the slope. Standard deviations SD for the estimates of y^0 and A^0 are calculated as well as the standard error SE for the slope b. The above equation is preferred above the more common expression:

$$y = a + b\,A$$

since it allows for calculation of confidence bands also – see below.

Note: y^0 and A^0 are equal to the averaged values of y and A respectively but their SD's are different!

Linear relations can be determined in several ways, depending on the initial assumption for the variables involved. Because of that, some different types of linear relations are calculated.

4.3 Linear relation for totals n

Commonly, linear relations are determined through the least-squares method assuming one variable to be independent and the other one dependent. The sum of the squared distances of the dependent variable to the line is minimized. Traditionally, in the literature the surface area A is taken as the independent variable. This type of fit is represented as Type I.

One might argue, that this type of fit is not suitable, since it expects any possible value for the dependent variable, also fractional numbers, whereas the n only have whole numbers equal to or larger than 0. (The n is not a continuous variable.)

This problem can be circumvented by choosing the n as the independent variable, since that is considered fixed, without deviation. This type of fit is represented as Type II.

The output allows for calculation of confidence limits for data and for estimates, by means of the following formulas[5]:

[5] *Reference: Documenta Geigy, Wissenschaftliche Tabellen (1960). Geigy AG, Pharmazeutische Abteilung, Redaktion Konrad Diem, p.178a*

4.4 Boundaries for Type I - n

Here, A is the independent variable so that limits are for n as calculated for given A:

$$n = n^0 + b\,(A - A^0) \pm t_{2\alpha}\,\sigma_n$$

$$\sigma_n^2 = \left(1 + \frac{1}{N}\right) SD_n^2 + SE_b^2\,(A - A^0)^2 \qquad \text{for tolerance bounds}$$

$$\sigma_n^2 = \frac{1}{N} SD_n^2 + SE_b^2\,(A - A^0)^2 \qquad \text{for prediction bounds}$$

where N is the number of measurements and $t_{2\alpha}$ is a statistical factor depending on the confidence percentage and the degree of freedom, N−2. For example, for a 5% percentage and a large number N, $t_{2\alpha}=1.97$. Then, the tolerance bounds border the area where 95% of the measurements is expected; the prediction bounds limit the predicted value with 95% confidence. All this, assuming n not to be restricted to whole numbers. The adjacent graph is for the examples CntX.TXT combined with FocX.TXT, number of domains n vertically against contour surface area A horizontally. The straight line is for all 59 contours (N−2= 57: $t_{2\alpha}$= 2.0025), the bounding curved lines are vertical distances from the line, in the n-direction.
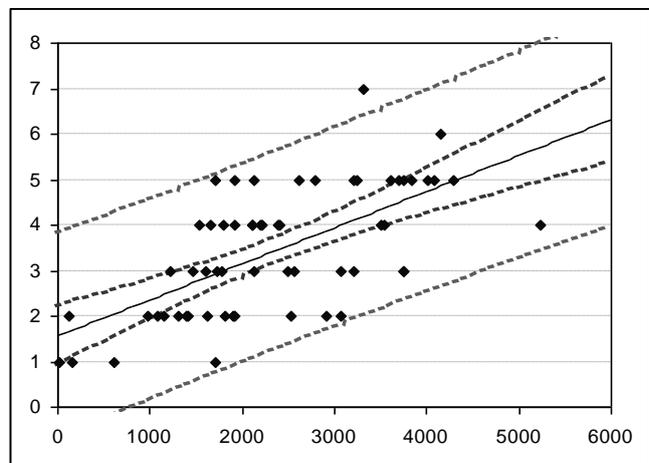
In the representation $n = a + b\,A$ the standard error in the estimate for a is:

$$SE_a^2 = \frac{1}{N} SD_n^2 + SE_b^2 (A^o)^2$$

since a is the value of n at A=0. For the current example, $a \pm SE_a = 1.566 \pm 0.327$. Note, that this is significantly different from 0.

4.5   Boundaries for Type II - n

Now, n is the independent variable so that confidence limits are in A as calculated for given n:

$$A = A^o + \frac{n-n^o}{b} \pm t_{2\alpha}\, \sigma_A$$

$$\sigma_A^2 = (1+\frac{1}{N})\, SD_A^2 + \frac{SE_b^2}{b^4} (n-n^o)^2 \qquad \text{for tolerance bounds}$$

$$\sigma_A^2 = \frac{1}{N}\, SD_A^2 + \frac{SE_b^2}{b^4} (n-n^o)^2 \qquad \text{for prediction bounds}$$

The adjacent graph is for the examples CntX.TXT + FocX.TXT, once more, for a 5% percentage. Results are distinctly different from the Type I results. The bounding curved lines are horizontal distances from the line, in the A-direction.

In the representation $n = a + b\,A$ the standard error in the estimate for a is:

$$SE_a^2 = \frac{b^2}{N}\, SD_A^2 + SE_b^2 (A^o)^2$$

For the current example, $a \pm SE_a = -1.112 \pm 0.754$, not significantly different from 0.

4.6   Linear relation for summed fractions f

The common treatment, A independent and f dependent variable, is denoted Type I here also.

In this case, the argument against the common treatment is, that neither variable will be truly independent, both exhibiting dispersion. The most often applied solution is, assuming this dispersion to be equally important for both variables (f, A equivalent), equal in the normalized values, a form of *Deming regression*[6]. The resulting slope is equal to the geometric mean of the slopes of the Types I en II of the preceding sections[7]. This linear relation is applied for the calculations denoted Type III.

[6] *Reference: Deming, W. E. (1943). Statistical adjustment of data. Wiley, NY (Dover Publications edition, 1985). ISBN 0-486-64685-8.*
[7] *Reference: personal derivation, available on request.*

4.7   Boundaries for Type I - f

Again, A is the independent variable and limits are for f as calculated for given A:

$$f = f^o + b (A-A^o) \pm t_{2\alpha}\, \sigma_f$$

$$\sigma_f^2 = (1+\frac{1}{N})\, SD_f^2 + SE_b^2 (A-A^o)^2 \qquad \text{for tolerance bounds}$$

$$\sigma_f^2 = \frac{1}{N}\, SD_f^2 + SE_b^2 (A-A^o)^2 \qquad \text{for prediction bounds}$$

The adjacent graph is for the examples CntX.TXT + FocX.TXT, fractions sum f vertically against contour surface area A horizontally; again, for a 5% percentage, and bounding curved lines vertical distances from the line, in the f-direction.

For the representation $f = a + b\,A$:

$$SE_a^2 = \frac{1}{N}\, SD_f^2 + SE_b^2 (A^o)^2$$

results $a \pm SE_a = 0.1167 \pm 0.0359$. This is not significantly different from 0 at a 5% level (though it would have been at a 10% level).
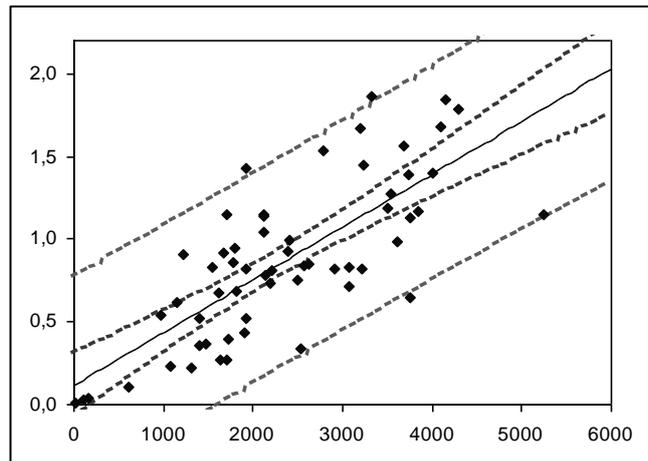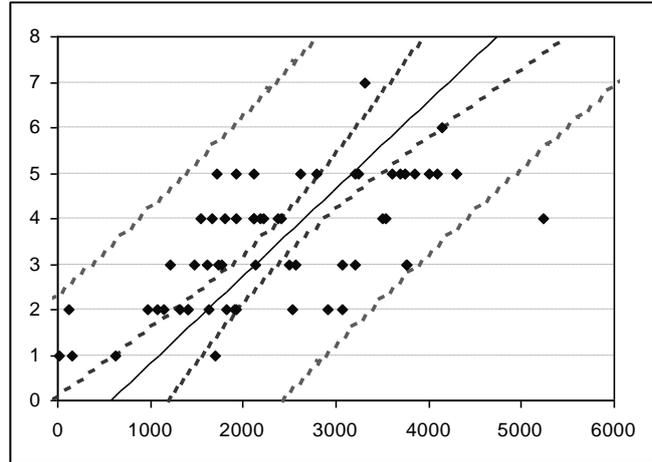
4.8 Boundaries for Type III - f

In this case, there is no independent variable and both A and f are not exact but subject to uncertainty. Values calculated for a chosen value A_c for the surface area A then yield, simultaneously:

$$f = f^0 + b\,(A_c - A^0) \pm t_{2\alpha}\,\sigma_f$$

$$A = A_c \mp t_{2\alpha}\,\sigma_A$$

$$\left.\begin{array}{l} \sigma_f^2 = (1+\dfrac{1}{N})\,SD_f^2 + SE_b^2\,(A_c - A^0)^2 \\[2ex] \sigma_A^2 = (1+\dfrac{1}{N})\,SD_A^2 + \dfrac{SE_b^2}{b^2}\,(A_c - A^0)^2 \end{array}\right\} \text{ for tolerance bounds}$$

$$\left.\begin{array}{l} \sigma_f^2 = \dfrac{1}{N}\,SD_f^2 + SE_b^2\,(A_c - A^0)^2 \\[2ex] \sigma_A^2 = \dfrac{1}{N}\,SD_A^2 + \dfrac{SE_b^2}{b^2}\,(A_c - A^0)^2 \end{array}\right\} \text{ for prediction bounds}$$

to be interpreted as pairs $A = A_c - t_{2\alpha}\,\sigma_A$ associated with $f = f^0 + b\,(A_c - A^0) + t_{2\alpha}\,\sigma_f$ and $A = A_c + t_{2\alpha}\,\sigma_A$ associated with $f = f^0 + b\,(A_c - A^0) - t_{2\alpha}\,\sigma_f$. The adjacent graph is for the examples CntX.TXT + FocX.TXT, again, for a 5% percentage. Distances to the line now are neither horizontal nor vertical, but transverse to the line.

For the representation $f = a + b\,A$ the standard error in the estimate for a involves both midpoint uncertainties:

$$SE_a^2 = \frac{1}{N}\,SD_f^2 + \frac{b^2}{N}\,SD_A^2 + SE_b^2\,(A^0)^2$$

For the current example, $a \pm SE_a = -0.1257 \pm 0.102$.
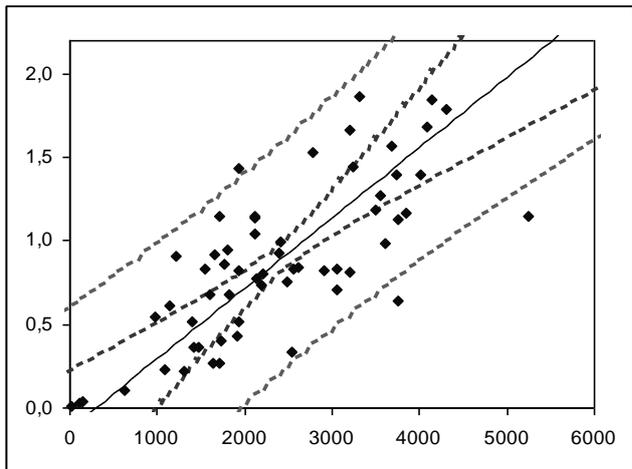
4.9 Direct proportionality

A contour with a surface area of 0 cannot be overlapped by a domain and vice versa, justifying the assumption that for A=0 always n=0 and f=0. If still a straight line, the corresponding equation is:

$$y = B\,A$$

without a y-axis cutoff value a. Also for this case, AnaTis calculates statistics. The corresponding confidence limits are:

Type	tolerance bounds	prediction bounds
I - n	$\sigma_n^2 = SD_n^2 + SE_B^2\,A^2$	$\sigma_n = SE_B\,A$
II - n	$\sigma_A^2 = SD_A^2 + \dfrac{SE_B^2}{B^4}\,n^2$	$\sigma_A = \dfrac{SE_B}{B^2}\,n$
I - f	$\sigma_f^2 = SD_f^2 + SE_B^2\,A_c^2$	$\sigma_f = SE_B\,A_c$
III - f	$\sigma_f^2 = SD_f^2 + SE_B^2\,A_c^2 \quad;\quad \sigma_A^2 = SD_A^2 + \dfrac{SE_B^2}{B^2}\,A_c^2$	$\sigma_f = SE_B\,A_c \quad;\quad \sigma_A = \dfrac{SE_B}{B}\,A_c$

5 Concluding remarks

In the tables and graphs, 'no border' stands for leaving out all cases where domains at the border were involved. So, for domains, all non-border domains are included. For contours, however, it implies leaving out all contours having an overlap with a border domain, even when the contour itself is complete within the field. These contours may be complete, but their overlap results are questionable. The columns '½ border' half include the border cases analogous to the histogram results. For histograms, this was on statistical grounds – for linear relations there is no statistical justification (yet); it is for comparison.

In AnaTis, graphical results in black incorporate all domains, including the border cases. Results for domains at the border omitted are in red. These results as well as the corresponding statistics can be very different from the all-domain results. For relations, a yellow line represents the results leaving out half of the border cases. These red and yellow lines are dissimilar from the lines shown in the graphs above; AnaTis does not show confidence lines nor direct proportionality lines. These can be calculated separately, from the tabulated data output of AnaTis.