ENTE PER LE NUOVE TECNOLOGIE,
L'ENERGIA E L'AMBIENTE

Dipartimento Ambiente

BASIC DISTRIBUTION FREE IDENTIFICATION TESTS FOR SMALL SIZE SAMPLES OF ENVIRONMENTAL DATA

A.G. FEDERICO, F. MUSMECI

ENEA - Dipartimento Ambiente
Centro Ricerche Casaccia, Roma

RT/AMB/97/29

Testo pervenuto nel dicembre 1997

Abstract

Testing two or more data sets for the hypothesis that they are sampled from the same population is often required in environmental data analysis. Typically the available samples have a small number of data and often then assumption of normal distributions is not realistic. On the other hand the diffusion of our days powerful Personal Computers opens new possible opportunities based on a massive use of the CPU resources. The paper reviews the problem introducing the feasibility of two non parametric approaches based on intrinsic equiprobability properties of the data samples. The first one is based on a full resampling while the second is based on a bootstrap approach. A easy to use program is presented. A case study is given based on the Chernobyl children contamination data.

[NON PARAMETRIC TEST, BOOTSTRAP, STATISTICAL SOFTWARE]

Riassunto

Nell'analisi di dati ambientali ricorre spesso il caso di dover sottoporre a test l'ipotesi di provenienza di due, o più, insiemi di dati dalla stessa popolazione. Tipicamente i dati disponibili sono pochi e spesso l'ipotesi di provenienza da distribuzioni normali non è sostenibile. D'altra parte la diffusione odierna di Personal Computer fornisce nuove possibili soluzioni basate sull'uso intensivo delle risorse della CPU. Il rapporto analizza il problema e presenta la possibilità di utilizzo di due test non parametrici basati sulle proprietà intrinseche di equiprobabilità dei campioni. Il primo è basato su una tecnica di ricampionamento esaustivo mentre il secondo su un approccio di tipo bootstrap. E' presentato un programma di semplice utilizzo e un caso di studio basato su dati di contaminazione di bambini a Chernobyl.

Index

"Basic distribution free identification tests for small size samples of environmental data"

1.Foreword

When analyzing environmental data the identification is usually defined as the problem to find a statistically based answer to the assignability of (at least) two samples to the same population. This may be the case of a pattern recognition of environmental events, the study of the effects of a treatment in agriculture or in medicine or the acknowledgment of a sustainable trend in an ecosystem with respect to a model. The problem class is not particularly sophisticated but sometimes a wrong approach may be totally misleading due to the data scarcity or to trivial errors in the statistical hypotheses. The theory of identification tests and the generic decision theory is very well stated in the bayesian approach even for multivariate data[1] but a number of items must be verified to assure the computability of the decision statistic and of the related questions. First of all the joint normality of the variables is requested for the more widespread identification tests statistics but unfortunately the assessment of this hypothesis remains very cumbersome. Even the marginal normality, a non-sufficient condition for the global property, is frequently not checked with accuracy, or the data are not enough to reach a believable conclusion. Furthermore in several cases neither the so called well behavior of the data is so certain, that is to say that strange and subjective property of the data that comforts everyone in assuming the data to be normally distributed even if there is not any experimental evidence for it nor there is any normality test overcome or even some non successful trial exists. But we are intended to face the case where the data are so few that no serious distribution test is allowed and cues exist of ill data behavior, like a pronounced asymmetry or a neat data clamping like the data being positive with an average value very close to zero with respect to the sample variance. In these situation, due to the ignorance of the probability density functions, any variable transformation to achieve better distributions may be hazardous. Nevertheless much people refers to tests as Student-t , Chi square or Fisher that not only requires normality but are demonstrated to have poor robustness against these kinds of data failures.

In the sequel the following notations will be adopted:

The data are assumed to be multivariate and multigrouped in the sense that any item, belonging to a pre defined group, is measured by a set of determinations or variables. Therefore any sample, $\mathbf{X}_k$, k=1, ...K, is defined in K n_k x m matrices with entries $x_{i,j,k}$ where m is the dimensionality of the data vector and n_k the k-th group dimension. We have the row index i=1, ... n_k, the column index j=1, ... m and k the group index with $n = \Sigma_k n_k$. The Centroid K x m matrix $\mathbf{C}$ is given by the expected values of the data in the k-th group and its best estimate is the average K x m matrix $\underline{\mathbf{X}}$, computed by the averaging of the row entries with respect to k in $\mathbf{X}_k$. The covariance matrices Σ_k (m x m) are the expected values of the group normalized sum of cross products of the deviations from the means. The sample covariance matrices $\mathbf{W}_k$ (m x m) are estimated averaging these products over k and normalizing to n_k -1. The normalization of the sample covariance matrices with respect to the product of the row/column elements of the diagonal gives the sample correlation matrices P_k, with ones on the diagonal.

The commonest hypotheses are the independence of the data samples; the omoschedasticy of the samples i.e. the uniqueness of the covariance, that allows to get a better estimate of Σ, $\mathbf{W}$, by the so called pooling, i.e. the weighted average over the groups of the cross products and the multivariate normality of the data. The latter assumption, even if it should be carefully justified and tested as possible, gives a strong advantage in the mathematical normalization of the probability functions and of the test statistics. One further assumption, especially useful in absence of prior knowledge of

the probability functions is the absence of correlation between the variables. This is almost always a very weak hypothesis and may bring to erroneous conclusions.

A statistical distribution regulates the between groups relationships. The sample statistic is defined by the m-dimensional mean of the means vector $\underline{\mathbf{X}}$ and by the means covariance matrix Φ (mxm) estimated by the between groups sample covariance matrix **B** (m x m) computed from the averaged (to K) sum of cross products of deviation of the means normalized to K-1.

2.Parametric vs. non parametric tests

Here is the case only to recall that the parametric approach to a statistical test, or to a generic statistical problem, requires the definition of the probability density functions under the form $f(\mathbf{x} \mid \theta)$ being **x** a data vector of dimensions m and θ a higher dimension vector of the parameters that define completely the probability functions that must have predefined functional forms. In the bayesian approach both the vectors can be considered random variables. When a data sample **X** is collected the Bayes theorem gives the opportunity to compute the a posteriori density function of the parameters, given the data, $g(\theta \mid \mathbf{X})$, as a function of the before available distribution, the a priori density $g(\theta)$. The conditioning operator is the likelihood of the data that is, in force of data independence, the composition, i.e. the product of the probabilities of the data given the parameters $L(\mathbf{X} \mid \theta) = \Pi_i f(\mathbf{x}_i \mid \theta)$. L of course is not a probability because does not meet the axioms not summing to 1 in mutual exclusive events nor diverging to infinity because the limitation of the alternatives.

The more widespread statistical tests may be defined in relation to the expected value of θ, in many cases the hypothesis $E(\theta)=\theta_0$ must be simply tested against the alternative $E(\theta)=\theta_1$. Composite hypotheses tests may always be redirected to the former and simpler case but their properties must be restated. The test algorithm is developed in a way to partition the event space of **x** in two complementary regions, the acceptance ***A*** and the rejection area ***R*** so that the test decision will be assumed depending on the occurrence of **x** in one of the two regions. For each observed sample, or for any sample sufficient statistic, two related probabilities are computed :

- The false rejection probability $P_{f.rj} = \alpha = P(\mathbf{x} \in \boldsymbol{R} \mid E(\theta)=\theta_0)$, called the type I error probability being $1-\alpha$ the confidence level of the test;
- The false acceptance probability $P_{f.id} = \beta = P(\mathbf{x} \in \boldsymbol{A} \mid E(\theta)=\theta_1)$, called the type II error probability, being $1-\beta$ the power of the test.

The region ***R*** is chosen in the way that minimizes, given ***A***, the false rejection probability $P_{f.id}$, or, that is the same, to maximize the power of the test β. It is intuitive that the power of the test β must be a non decreasing function of the sample dimension n and that α, β are in countertendency. The sampling theory of the identification tests relies on the Neyman-Pearson lemma that states that the most powerful choice of the region ***R*** is the one that satisfies the condition that the likelihood ratio be:

$$L(\mathbf{X} \mid E(\theta)=\theta_0) / L(\mathbf{X} \mid E(\theta)=\theta_1) < b(\alpha)$$

being b a positive number determined by α. But the bayesian approach valorizes the knowledge of the a priori information about the distribution of the random variable θ, so that the optimum region is no more determined by the likelihoods only, but by the a posteriori probabilities in the sense of Bayes theorem, $g(\theta \mid \mathbf{X})$. Whichever probability is prevailing, either for $E(\theta)=\theta_0$ or for $E(\theta)=\theta_1$, it determines if the sample **x** is in the acceptance or in the rejection region by means of suitable cost

functions that must be integrated over the parameter space with the a posteriori densities. It can be shown that the Bayesian decision gives a test algorithm that is the most powerful[2] as those that are implemented in the sampling approach following the Neyman-Pearson lemma.

We say that a test is *correct* if its power $1-\beta > \alpha$, so assuring that the probability to reject the identification $E(\theta)=\theta_0$ is greater when $E(\theta)=\theta_1$. The test is *consistent* if the power $1-\beta$ tends to 1 as n increases. The relative *efficiency* of two tests is assessed by comparing the numbers of samples n_1, n_2, that assure the same power in the consistency hypothesis.

If the form of probability densities are not known nor any suitable hypothesis on it is acceptable, the parametric approach fails because the likelihoods and the priors are no more computable. We fall in the domain of non parametric approaches or, more correctly, of distribution free statistic. In this domain all the algorithmic development of a test may be questioned and a test statistic may be very difficult to define. Several times suitable functional transformations are applied to the data to try to have some practicable hypotesis on the probabilities at hand. Sometimes logarithmic transformations that transform multiplicative in additive random contributions may, even intuitively, better approximate the condition of the central limit theorem and then the normality. A special case is when we have very small sized samples without any knowledge on the distributions. May be that in these situations the only sufficient statistic are the data themselves and we have to refer to their basic properties to develop a test. One of these items will be discussed next.

3.The permutation test basic approach

One interesting property of independent data constituting a sample from the k-th population in a given experiment is that being k the number of data in the k-th group we can resample to create a lower size group with $k'<k$ data in $q=k!/[k'!(k-k')!]$ different ways. Each new sample is equiprobable with probability 1/q. This assumpion is by no means restrictive.

Suppose we have two data sets (k=2) with n_1 and n_2 data and that the question is to test if they belong to the same population. If the assumption is true we can imagine a unique sample with $n=n_1+n_2$ data that we can resample attributing to group 1 or 2 the n data in

$$n!/(n_1!n_2!)$$

possible ways with equal probability. If what we are interested to test is the position of the sample, that is to say that we assume equal moments for the unknown distributions for all the moment orders >1 and we try to check the identification of the group means $\underline{x}_1$ and $\underline{x}_2$, and if we want a confidence level $1-\alpha$ for the identification test, we can assume the test statistic as:

$$d = |\underline{x}_1 - \underline{x}_2|$$

and reject the identification only if the resampled groups fraction with $d'>|\underline{x}_1 - \underline{x}_2|$ are less than or equal to α. For the multivariate case, assumed the absence of correlation among the variables, to avoid the lack of scale homogeneity it is advisable to normalize the variables with respect to their standard deviations or to assume instead of d a statistic based on the well known Mahalanobis distance D^3. Alternatively the test may be decomposed in a sequential test with the same confidence level α, each for every variable, so that the overall test power is given by the complement to one of a $P_{f.id}$ obtained as a compound multiplicative probability $\beta_1\ \beta_2 \ldots\beta_m$. The latter option is unavoidable if the data matrix is sparse, that is to say if some variable measures are missing.

This kind of tests is called traditionally "permutation test" and is only based onto a very basic symmetry property of the data consisting in the equal probability of each combination of the available data from group1 and2 that, supposed the population to be the same, can be rearranged in two groups of dimensions n_1 and $n_2=n-n_1$ in $n!/ (n_1!n_2!)$ equal probability combinations. The reader must consider that no assumption was done on the form of probability distributions of the data, therefore the permutation tests are certainly distribution free (non parametric as someone says) and have a largely wider field of application with respect to the tests (Student-t, Fisher, Chi-square) that require normality. Unfortunately the computer enumeration of the permutations becomes very soon unpracticable as n increases (cfr. § 7). In some cases, for higher values of n, a bootstrap resampling method may be adopted to approximate the test statistics and to reduce the computation time.

But there is another topic advantage when using this test: it is possible to evaluate the power β and then the false identification probability $P_{f.id}$. This allows the evaluation of the efficiency of the test. Take the case of mean position test where we can assume that the null hypothesis is d=0 to be tested against the alternative d= D, being the same the variance and the higher order moments. Assuming that the alternative hypothesis is true translate the group1 exactly of D in a way that the modified data centroid 1 is exactly positioned on centroid 2. This way we will achieve a new $n=(n_1+n_2)$ dimensional group of population 2 data. The power of test,β , is the fraction of the $n!/ n_1!n_2!$ combinations of the n available data in a $n_1$dimensional subgroup that, falling in the previously defined acceptance region of the test, give rise to a false identification. It was demonstrated[2] that the efficiency of the permutation test is asymptotically the same as the efficiency of the Student-t test on the means of normal variables. The same work shows that as n_1 increases the probability to accept the alternative hypothesis, whichever may be, reduces to zero. Therefore the test is consistent.

There is a class of non parametric test that are of widespread application in literature and are based on the variable ranking. Even if they are less efficient than those based on the permutations of the original data, they are much less time consuming on the computer. Therefore we can consider that the application domain of the tests here presented should be restricted only to the case of very low size of the sample and very clear rejection of their normality.

4.Normality assumption and tests

This item is indeed a very delicate one. Many authors have shown that the identification tests based on the normality assumption are not enough robust against the probability density alterations. On the other side the normality assumption is fundamental to have a good heritage of algorithmic developments, of test statistics and of mathematical options. In the multivariate statistic any non normal approach is almost impracticable. These considerations lead to an often too simplistic assumption of normality and to a relatively too strong confidence on the central limit theorem implications about statistically complex processes that however, several times, do not meet the normality conditions, mainly when the samples are few. There are cases, like those of signed variables, in which the symmetry violation should advise the experimenter to not risk the normality, or there may be that the data scarcity and the process ignorance do not allow to verify firmly the normality. Furthermore we do not have any test for the joint normality of multivariate data and the normality assumption on the marginal probability density functions is merely a necessary, non sufficient condition for the global normality. The only theorem we can trust on states that to have multivariate joint normality all the linear transformations of the data, ideally all the projections on an arbitrary axis in the event space , must be normal. To verify this property of the data samples is computationally unfeasible.

There are many sound non parametric methods to get good estimates of unknown probability density functions, in recent years, in particular, good fortune arose to the multinormal approximation to an unknown function inspired by the good mathematical quality of the gaussian functions. The algorithmic and the multivariate extensions remain however quite cumbersome.

What to do? The normality test on the marginal distribution is essential. We know that the random residuals of any functional approximation must be in turn normal with zero mean and equal variance so that their sum of squares may be distributed as a chi-square with as many degrees of freedom as the histogram boxes are. After this very popular test some good non parametric tests are available, mainly the Kolmogorov-Smirnov test that is based[4] on the observation that, as the number of observation increases, the absolute value of the maximal deviation of the sample value of the distribution function from the model value, multiplied by $\sqrt{n}$, has a (limit) Kolmogorov distribution. The test on the upper deviation may be managed in the usual way by the definition of a suitable confidence level $1-\alpha$.

However the most practical procedure for testing the identity of the two population means is to use sequentially the normal approach (the t-test), given α and then, if the sample dimensions are not too high, the permutation test with the same choice for α. The subordinate probabilities of false identification and the power of the test should be computed in both the approaches. The results comparison may evidence substantial differences in the test parameters. The main cue for the contradiction may not be else than an hypothesis violation on the probability functions assumptions. Therefore the two test do not reinforce each other and the permutation test results must be assumed as correct against the normal tests, because of the distribution free nature of the former. That will be the sustainable conclusion. As a matter of fact, even if expensive, the method of confrontation between the results of a normal approach to a distribution free counterpart, if it exists, is the best evidentiation of the lack of robustness of the normal tests and therefore of the non normality of the data and may be advantageously substituted to the normality tests.

Suppose we have two small samples drawn from an exponential distribution that is strongly skewed. We want to test if the difference of the sample means is significant at 95% confidence level, that is to say if the two samples are coming from the same population against the hypothesis that they belong to two moved away distributions. Sample #1 will be extracted from a distribution that has the test acceptance region, at 95% confidence level, defined for $\underline{x}_2<0$. This will be accomplished simply assigning n_1 high enough to assume a normal distribution for the mean and solving the Student t-test inequality to find $\underline{x}_1$. Suppose $n_2=1$ to have the density function of the mean still exponential. The analytical form of the density function is very simple, i.e. $\lambda e^{-\lambda x}$ for $x>0$ and 0 elsewhere, with mean λ^{-1} and variance λ^{-2}. We can assume $\lambda=1$ without loss of generality. Immediately follows $\beta=0$. Suppose now the two samples coming from normal distributions with unit mean and variance (the same parameters of the given exponential distributions). It readily follows $\beta=15{,}9\%$, being β the erf(-1). An experimental test using the permutations with the following parameters:

Probability density functions: exponential with unit variance.
Sample #1: $n_1=10$, mean = -.86
Sample #2: $n_2= 1$, mean = 1;

the data are :
Sample #1: (-1.5, -1.2, -1.3, -.6, -.7, -1.6, 2.4, -1.6, -1.7, -1.5)
Sample #2: (1);

the results are:
% of resampled means differences exceeding $|\underline{x}_1 - \underline{x}_2| = 9\%$ (this results depends from the low number of permutations available. Repeating the sample generation this value converges to 2%);
sampleβ=0

that, following the above calculations, strongly confirm the non applicability of the normal hypothesis.

5.Software implementation generic overview

Because analysis of small data sets is often the case in the environmental data analysis, a ready-to-use computer program was developed for applying the introduced tests which are not commonly available to researchers in typical statistical packages. The software compares the means of two groups of observations of several variables, hypothesizing the coincidence of the higher moments, to test if they are belonging to the same population. To partially overcome the effects of the combinatorial explosion as the sample sizes increase, two options are available: full combinations or bootstrap resampling. Using the first option the comparison of the means is repeated among all the possible $n!/(n_1!n_2!)$ combinations. The original difference is then compared with the distribution of the values obtained with the generated samples. With the bootstrap approach only a given number (5000 in this program version) of random resampling is made of n_1 and n_2 sets out of the original set of n1+n2 observations. Then the comparison is again carried out. Instead of the implemented differences between means any different statistic could be computed with simple modifications of the software. The program prompts the user to select two files with data sets 1 and 2. Data should be available in ASCII format organized by observations (rows) and variables (columns).
The software is written using Visual Basic version 4 for a Windows 95 environment and it is available for download from the ENEA Environment Department WEB site at: http://wwwamb.casaccia.enea.it./anvas

5.1 Full combinations

The kernel for the "*full combinations*" option is an algorithm for generating all the possible combinations of n_1 and n_2 data out of a set of $n = n_1+n_2$ observations. The idea is to use a binary n digit number as a string of flags to indicate if an observation should be temporary regarded as drawn from the first or from the second set. The following example is drawn from two sets of n_1=3 (2.2; 2.1; 2.4; mean=2.3) and n_2=2 (1.3; 1.5; mean=1,4). The original sample corresponds to the string 11100, a general binary string like 11010 reassigns observations 1,2,4 to set 1 while observations number 3 and 5 to set 2. In this case 11010 the means would be 1.87 for set 1 and 1.3 for set 2 (difference 0.57). For n_1=3 and n_2=2 there is a total of 5!/(2!3!)= 10 possible different equiprobable resampled sets.
The algorithm works on the flags generating all the combinations with n_1 "1's and n_2 0's. The following table contains the whole sequence:

FLAGS	Notes
11100	The first 3 observations are from set 1, the last two from set 2 (orginal sets)
11010	Digit #3 was moved
10110	Digit #2 was moved (it had a 0 on the right position)
01110	Digit #1 was moved
11001	Digit #4 was moved and the other two 1's are collapsed to the first positions
10101	

01101
10011 Digit #3 was moved and the 1 on position 2 was collapsed to the first position
01011
00111 The first 2 observations are from set 2, the last three from set 1

The program flowchart is extended hereafter in nine steps:
1) Allocate $n= n_1+n_2$ flags;
2) Initialize the fist n_1 flags to 1 and the remaining n_2 to 0;
3) Assign $\underline{x}_2$ the expected mean value of the alternative population; compute statistics (means, differences, statistics and probabilities) for the two initial sets;
4) Being i the position of the least significant flag 1 with a 0 on its right set i-th digit to 0 and i+1-th digit to 1. If such position is not found then exit the procedure;
5) Set j to the count of flags1 found on the left of the moved flag;
6) Set the first j flags to 1 and the remaining i-j to 0;
7) Compute statistics (means, differences, error probabilities etc.) for the temporary sets counting the samples with statistic values higher than the initial sets;
8) Repeat from step 4. At the loop end compute and display test results and general statistics;
9) Run again the loop to compute the power of the test β.

The time required for the algorithm to run became prohibitive as $n!/ n_1!n_2!$ (the number of executions of the loop) increases. Given n, the worst case is for $n_1=n_2$., with $(2n_1)!/n_1!^2$ loops needed. This number grows exponentially as function of n as n became large. This could be shown using the Stirling's formula for the factorial:

$$n! \cong \sqrt{(2\pi)}\, n^{n+\frac{1}{2}} e^{-n}$$

$$(n)!/(n/2)!^2 \cong 2^{n+\frac{1}{2}} / \sqrt{(\pi n)}$$

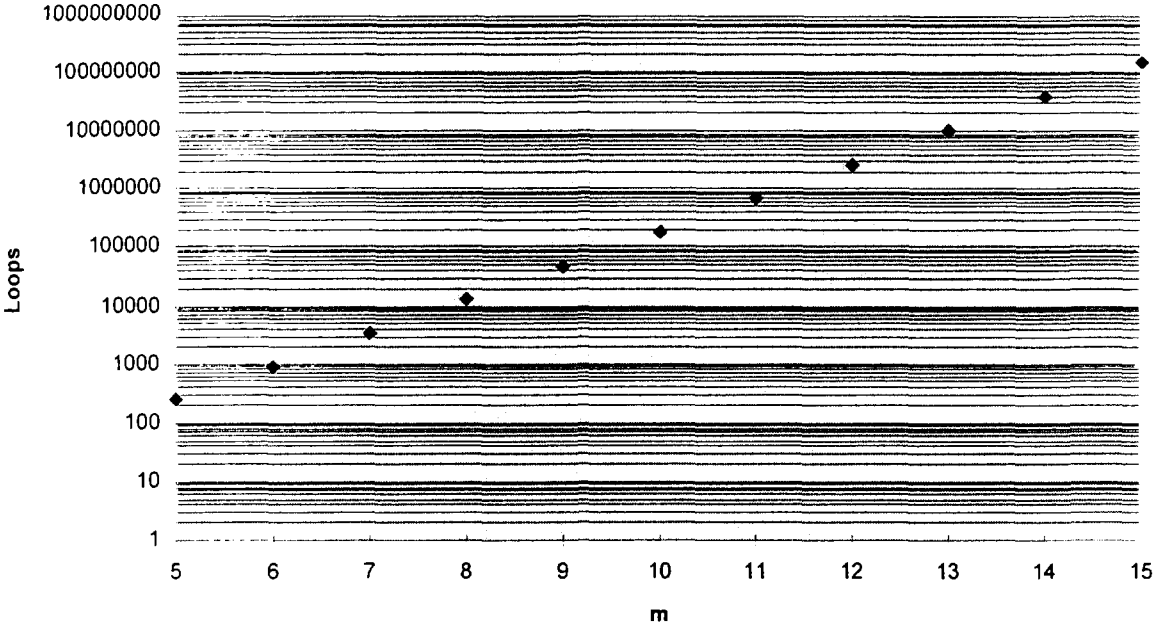

Fig. 5.1: Number of loops as function of m (n/2)

The time in seconds required for a typical 100 Mhz Pentium CPU (Compaq XL 5100) to execute the "*full combinations*" is shown in Fig 5.2 as a function of the number of the loops needed.

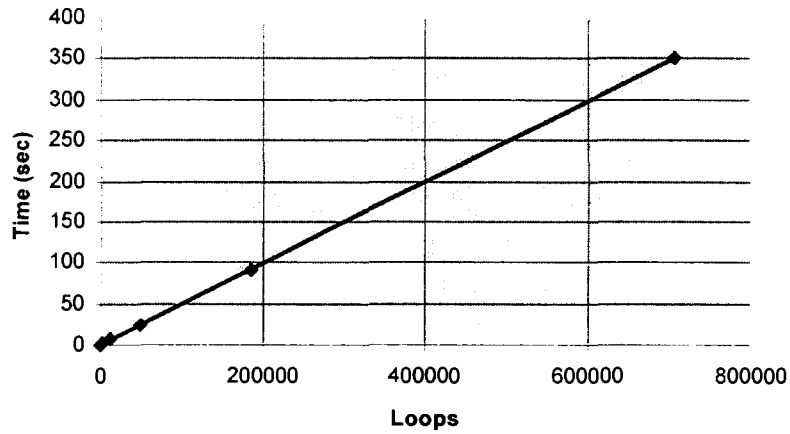

Fig 5.2: CPU time (typical) as function of the number of loops

Typically 5 sec. CPU time is required for each 10000 loops execution. With $n_1=n_2=9$ (48620 loops) the required time is 23 sec. For the same value of n and different n_1,n_2 the CPU time decays similarly to the binomial coefficients and is inversely proportional to the computer clock rate. An approximation can be given for the CPU time requested, in seconds, as function of the sample sizes n_1,n_2 and of S (the CPU clock in Mhz).

$$\text{Time} = 0.05\ (n_1+n_2)!/(n_1!n_2!) / S - 0.3.$$

After each step the program computes the statistics, multivariate or sequential, i.e. a β_j for each variable, cumulating the fraction of temporary group 1 samples with a statistic greater than that of the original sample. At the loops conclusion this fraction will be α^* and can be suitably compared with α, being the identification rejected if $\alpha^*<\alpha$. If at least one of the variables generates rejection, the identification hypothesis is discarded.

If the identification is assumed, $\alpha^*\geq\alpha$, the probability of false identification must be computed. Assumed $\underline{x}_2$ to be the mean value of the alternative population, a new sample is generated summing $\underline{x}_2$ to the temporary group 1 sample and running again the program. β_j is accumulated for each variable as the fraction of temporary n_1-dimensional samples, derived by the resampling of the cluster of the group 2 and of the modified group 1 data, that will be erroneously identified with group 1. The power of the multivariate test will be :

$$1-\beta = 1- \beta_1\ \beta_2 \ldots \beta_j \ldots \beta_m$$

being β the overall probability of false identification.

5.2 The Bootstrap option

The bootstrap approach merely limits the number of combinations selecting randomly from the whole equiprobable combinations allowed the two sample of n_1 and n_2 observations. Then the

statistics of interest are computed. The process is repeated L times (L=5000 for this program version). It is clear that some of the configurations of the flags of the "*full combinations*" may be drawn more than once and that many other are fully neglected . Nevertheless it is empirically demonstrated that, if L is large enough, the final results will not be very different[5]. The idea of flags indicating the membership to set 1 or set 2 is retained also for the bootstrap approach. To select the temporary sample the algorithm works selecting n_1 positions in a string of length $n=n_1+n_2$. All the positions are set in a array (urn) and a drawing without replacement is curried out. As a position is drawn, the corresponding flag is set to 1 and the selected position is taken out from the possible next extraction.

```
Allocate n= n1+n2 flags(j)
Allocate n= n1+n2 positions(j)
set all flags to 0
set positions(j)=j
repeat for  i = 1 To n1
  k = Int(RND * (n - i + 1) + i)
  SWAP position(i), position(k)
  flags(position(i)) = 1
Next i
```

here:
RND is a uniform random number generator (0,1)
INT(x) is the integer part of x
SWAP a,b is the exchange between the value of "a" and "b"

The above algorithm produces random numbers with n_1 1's and n_2 0's. With $n_1 = n_2 = 9$ the bootstrap algorithm takes about 10 seconds to execute 5000 resamples on a 100 Mhz CPU. This means that with $n_1 > 9$ the bootstrap approach is faster and should be used for $n_1 > 10$ if a quick computation is required.

6.A case study, the Chernobyl children contamination

It is based on the work done by ENEA Environment Department on the genetic effects of the Ucrainian children exposed to the Chernobyl accident[6,7] . In this study an analysis of chromosomal aberrations (Acentric fragments, Dicentrics and Translocations counting) was carried out on the populations and a statistic test is needed to check if the increase in aberrations of the exposed population is statistically significative. The aberration variables are counts with low probabilities to be nonzero, therefore their statistical distributions are something of very different from the normals. Therefore the Student-t test may be suspected to give rise to errors. The data are few but their sizes are probably not so small that we are not allowed to invoke the beneficial effects of the central limit theorem.

The children were divided into 3 groups:

- CONTROL GROUP A (Control): 11 subjects coming from the Smolensk region (WBC < 70 Bq). It is the regional reference group that can be assumed as a welfare level group.
- GROUP B (Evacuated): 7 subjects from the uncontaminated area of Smolensk (Russian Federation) but living in Pripjat (Ukraine) at the moment of the accident and thus exposed to the initial "acute" dose (in comparison to the exposure of non-evacuated people) of ionizing

radiation; these children, evacuated to Smolensk 36 hours after the accident, showed an internal contamination in the range 0-128 Bq.

- GROUP C (Contaminated): 24 subjects from Novosybkov in the Brianskya region of the Russian Federation (ground contamination 148x1010 Bq/km2 as reported by IAEA maps), who exhibited an internal contamination in the range 780-30,000 Bq.

Supposing that the aberration background of the russian population is higher no control population group coming from abroad, let say Europe, is introduced in the test. We gained the following results:

	Observations	Mean 1	Mean 2	Mean 3
Group A	11	2.38	0.47	0.61
Group B	7	3.14	1.00	1.17
Group C	23	3.52	0.87	1.55

Tab 6.1: Number of observations and mean values for group A,B,C
and var #1:acentric; var#2:dicentric; var#3:translation chromosomal aberrations

Acentrics	Control	Evacuated	Contaminated
Control	0	-0.76	-1.14
Evacuated	0.76	0	-0.38
Contaminated	1.14	0.38	0

Dicentrics	Control	Evacuated	Contaminated
Control	0	-0.53	-0.40
Evacuated	0.53	0	0.13
Contaminated	0.40	-0.13	0

Traslocations	Control	Evacuated	Contaminated
Control	**0**	**-0.56**	**-0.94**
Evacuated	**0.56**	**0**	**-0.38**
Contaminated	**0.94**	**0.38**	**0**

Tab 6.2: Differences among the means of the variables

% exceeded	PERMUTATIONS (Bootstrap)			STUDENT-t		
Acentrics	**Control**	**Evacuated**	**Contaminated**	**Control**	**Evacuated**	**Contaminated**
Control		18.3	12.4		18.0	12.0
Evacuated			39.4			37.1
Contaminated						

Dicentrics	Control	Evacuated	Contaminated	Control	Evacuated	Contaminated
Control		**19.1**	**19.0**		**16.7**	**15.9**
Evacuated			**34.5**			**40.0**
Contaminated						

Traslocations	Control	Evacuated	Contaminated	Control	Evacuated	Contaminated

Control	26.1	6.1	22.0	6.8
Evacuated		29.3		29.9
Contaminated				

Tab. 6.3 Tests results compared with canonical normal Student-t test

To estimate the multivariate power of the test we have first of all to define the alternative hypothesis that may be assumed with the worst case approach that gives rise to the theoretically lower value of the power. We assume therefore the alternative hypothesis to have a multivariate population exactly at the upper limits of the critical acceptance region, and we assume the variables correlation to be zero. Being in this case the mono-dimensional βs asymptotically 0.5, the multivariate expected value of β is 0.125 and the test power is 87.5%. Running the permutation test the sample values obtained are β_1=48.7; β_2=49.3 and β_3= 48.5. It follows β=11.6 and the sample power is 88.4% that is really a good approximation. Summing up the results of the case study, even if there are significative differences in the numerical values, they do not question the identification of all the three groups suggested by the authors working with the normal hypothesis. The sample sizes involved are at the limit under that the applicability of the asymptotic t-Student test is no longer justified.

7.Conclusions

The work further improves the caution due to the normality assumption when nothing is known about distributions of the samples, especially when the data size is small or very small and there are good reasons of deviation from normality of the data probability distributions. The Chernobyl case study first of all confirms that the exposure to the accident radiations of the children were not sufficient for a diagnosis of statistical significance in chromosomal aberrations with respect to the basic population of the former USSR. From the statistical point of view it shows that the convergence to the asymptotic normal properties of the equal mean tests is quite rapid, so that we can consider that sample sizes of the order of 10, those of the case test, are already enough to assume convergence and then the practicability of the classical parametric approach like in[6,7]. For very small sample sizes the paper offers a straightforward distribution-free approach that works only on the basic statistical properties of the data and is computationally advantageous. The permutation method offers at the same time the possibility to compute the power of the test and the probabilities of false identification with respect to any alternative hypothesis, (normally the worst case alternative), simply by resampling of the native data without any complicated integration of badly known density functions. The multidimensional problem is also approached and a procedure to compute the multivariate test power is given and the related multivariate false identification error. With reference to different tests, i.e. on variances, correlations, data independence or probability density, the permutation test remains a chance even if it has no normal asymptotic parametric counterpart as sample sizes increases. In these cases, being the computation time an increasing function of the sizes, the parametric tests must be implemented resorting to the bootstrap approach before stepping to the classical non-parametric methods (Kolmogorov-Smirnov, Wilcoxon, etc.) that, working on the ranks of the data, as it is very well known, are prone to lose information (entropy) carried by the data.

To face correctly a new problem of statistical testing for poor data samples of unknown distributions the final suggestion is to run both the tests, either the parametric or the permutation one. If differences are evidenced in the test parameters refer to the results of the permutation test only and discard the parametric approach. In both cases valorize the feature of the permutation test that

allows to get a sample value of the test power and of the false identification probability avoiding any futher analytical complication.

8.Acknowledgments

We are grateful to F. Mauro and to L. Padovani of the Environment Dept. of ENEA that posed the former question of testing some awkward data measured on the Chernobyl children and gave assistance in the test case data elaboration.

---

1) A. Federico, R.Picchia "Metodi innovativi nell'analisi di elementi grafici"
Quaderni di linguistica, Università della Calabria
2) B.Giardina, "Statistica non parametrica", Franco Angeli 1972
3) Mardia, Bibby, King, "Multivariate Analysis", Academic Press 1979 ...
4) V.S.Pugachev, "Probability theory and mathematical statistics", Pergamon Press, 1984
5) B.Efron:, R.Tibshipani, "An introduction to the bootstrap, Chapman Hall 1993
6) L. Padovani et Al. "Cytogenetic study in lymphocytes from children exposed to ionizing radiation after the Chernobyl accident", Mutation Research, 319 (1993).
7) L. Padovani et Al."Conventional cytogenetic and chromosome painting analyses of abberations in lymphocytes from children exposed to radiation fall-out after the chernobil accident", Mutation Research 1997 in press
8) A. Federico "Rappresentazione agli autovalori di matrici di distanze" Rapporto FUB 18 1984 2B2784

Edito dall' ENEA
Unità Comunicazione e Informazione
Lungotevere Grande Ammiraglio Thaon di Revel, 76 - 00196 Roma
Stampa: Centro Stampa Tecnografico - C. R. Frascati

Finito di stampare nel mese di gennaio 1998