

# Statistical analysis of incomplete data: theory, techniques and software

A. M. Nikiforov

Preface and supplement to the Russian edition of Little R.J.A., Rubin D.B. *Statistical Analysis With Missing Data*. Moscow, Finansy i Statistika, pp. 3-5, 284-332 (1991) (in Russian)

## Preface

Statistical analysis with missing data is a problem known to almost any applied data analyst. The problems discussed in the book and in the supplement have big practical importance. Very often researches do not pay enough attention to the correct missing data analysis, for example they might try to get rid of the gaps in data as soon as possible with traditional primitive gap filling methods or excluding incomplete observations from the analysis. This leads to ineffective solutions, inconsistency, biased estimates and non-robustness of statistical inferences.

A low awareness of these specific features is reflected in the software for statistical analysis as well. Most of the statistical systems include just trivial complete observations methods and/or paired methods and their modifications.

A main object of the book is a sample of multivariate observations with some of the values missing. We represent an  $r$ -variate observation with missing data as a pair  $(X, M)$ , where  $X$  is the original  $r$ -variate vector of values, and  $M$  is the  $r$ -variate vector of gap indicators, with coordinates having the values “present” or “missing”. The random vector  $(X, M)$  has distribution  $P(X, M)$ . The goal is to provide statistical inferences on the distribution  $P(X)$  of the vector  $X$ , with a part of data missing.

This book is the first of a kind published in Russian. It surveys methods for handling missing data problems, and presents a likelihood-based theory for analysis with missing data. The problems related to the distribution

$P(X)$  covered in the book are estimation of mean and covariance matrix of multivariate normal distribution, variance, regression and factor analyses, contingency tables and log-linear model analyses, time series, robust estimation, data analysis with non-randomly missing data and so on. The expectation maximization algorithm, Bayes inference, and multiple imputation are also covered. Sample survey inference is discussed. A theory for analysis of problems based on likelihoods derived from statistical models for the data and the missing-data mechanism is presented, with application of the theory to a wide range of important missing-data problems.

The systematic approach adopted in the book is based on the modeling of the joint distribution of vectors  $X$  and  $M$ , i.e., the distribution  $P(X, M)$  and the development of algorithms for estimating the distribution parameters of  $P(X)$ , based on the method of maximum likelihood. Most of the attention is given to finding analysis methods when the least prior information about gap distribution is required, that is when the gap distribution can be ignored. For the parametric settings used in the book, this is the condition MAR (see chapter 5). The respective technique for the estimation problem is a generalization of maximum likelihood estimation for the case of missing data, in the supplement to the translation this generalization is called the *method of maximum marginal likelihood* (MMML).

Currently, the analysis methods for data with gaps are well developed only for parametric models, and only for the task of estimating unknown parameters. This is reflected in the content: basically, most of the book is dedicated to the building of the EM-algorithm for finding maximum marginal likelihood estimates for various models. The holes in the theory are partially filled in the translation supplement. In it, nonparametric criteria for testing homogeneity hypotheses of two or more samples, and independence of random variables when gaps are present are given. A problem related to discriminant analysis of incomplete data is also examined.

One of the conclusions that can be drawn from the supplement is the fact that the conditions required in nonparametric settings are much weaker than those in their parametric counterparts. In other words, nonparametric methods designed for incomplete data turn out to be robust with respect to the gap distribution or, more precisely, to the dependence of the gaps on variables in observation. For example, the conditions that allow the mentioned tests to be used are weaker (see chapters 5 and 6 of the supplement). There are other examples of such robustness related to the problems of nonparametric estimation, classification etc. On the other hand, the method

of complete observations (fully excluding incomplete observations from the analysis), gap filling methods and pair-wise methods require the execution of a fairly stronger MCAR condition (see chapter 5).

In the supplement there is also a useful theoretical justification to the estimation methods described in the book, testing of gap distribution hypotheses and other questions are discussed. We also hope that the reader will be interested in the program code that realizes an EM-algorithm for the multivariate normal distribution.

Without doubt, this book will be useful to any specialist who is in one or another way related to the problem of statistical analysis with missing data.

*A. M. Nikiforov*

## **Translation supplement**

### **Statistical analysis of incomplete data: theory, techniques and software**

## **1 Introduction**

This supplement is dedicated to those directions and problems of statistical analysis with missing data that are not analyzed or not analyzed in-depth enough in Little and Rubin's book. Their discussion, in our opinion, is appropriate within the scope of this book and will be interesting to the reader.

Basically, the theoretical foundation of methods developed in the book is reduced to the reference to the article [Rubin, 1976]. In this work, some invariance properties of three types of statistics (including likelihood ratio) were demonstrated under loosely formalized conditions. These properties and the reference to an analogy with the case of complete data obviously cannot replace proofs. For example, observations in a sample with gaps belong to different subspaces of the original sample space, which is not the case in traditional assumptions; the classic identifiability conditions need refinement etc. Therefore, the classical results are no longer valid with missing data, and the analysis techniques still lack formal justification.

To solve this problem, for particular tasks and models we need to define specific conditions related to the presence of gaps. Such an approach not only allows a valid theoretical justification of the results to be obtained, but

is also helpful methodically: it leads to looser conditions on gap distributions, reveals the new properties of various methods etc.

In the first few chapters of the supplement, generalizations of some classic results in the case of missing data are proposed. We analyze asymptotic properties of maximum likelihood estimates (maximum “marginal likelihood” estimates), the calculation of which with the EM-algorithm is the base subject of this book. Classification tasks and time series analysis are examined.

Approaches to constructing and using statistical tests with missing data for typical null hypotheses, such as the homogeneity of two or more samples and independence of random variables, are examined within the range of a very important part of mathematical and applied statistics – the hypothesis evaluation theory. This theory is only lightly touched on in the book. Criteria that require very weak conditions to be used, compared to MCAR and MAR conditions, are proposed (see chapter 5.3 and chapter 2.1 of the supplement for definitions). Discussed is the randomness evaluation problem (MCAR and MAR conditions), which isn’t examined in the book either.

Besides that, the appendix discusses some important qualities possessed by various filling methods. It is shown that the “local filling” method for gaps (see chapter 7 of the appendix) is free of the serious drawbacks typical for the primitive filling methods described in chapter 3 of the book.

Finally, the appendix examines the current state of software designed for applied statistical analysis of data with gaps. Ideas for construction of analysis methods for data with gaps within an all-around statistical system are proposed. Program code that realizes the EM-algorithm for multidimensional normal distribution is given.

## 2 Properties of maximum marginal likelihood estimates

### 2.1 Notation

Let  $X$  be measurements of  $r$  features of an object, some of which are missing according to the  $r$ -dimensional vector of gaps  $M$  with values either “*present*” or “*missing*”. We will call an  $r$ -dimensional observation with missing data a  $2r$ -dimensional vector  $(X, M)$ .  $P(X, M)$  will denote the distribution of the random vector (r.v.)  $(X, M)$ .  $E^{X, M}$  means below expectation with respect to the distribution  $P(X, M)$ .

There are  $2^r$  possible outcomes for the random vector  $M$ . Let  $\sum_m$  denote the sum over all possible outcomes of  $M$ . The formal definition of missing data mechanism is formalized as follows.

Condition Missing at Random (MAR):

$$p(m|x) = p(m|x_{obs}^m) \quad (1)$$

where  $p(m|x)$  is the conditional probability of  $x$  with a structure of gaps  $m$  given  $x$ ,  $x_{obs}^m$  is a part of data present according to  $m$ .

In general case

$$p(m) = \int_x p(m|x) dP^X$$

will denote the unconditional probability to observe the missing data pattern  $m$ . Condition Missing Completely at Random (MCAR) is:

$$p(m|x) = p(m)$$

## 2.2 Consistency of MML estimates

In this section we consider  $\{P_x\} = \{P(x|\theta)\}$  as a parametric family dominated by measure  $v = v_1 \dots v_r$ , with the respective family of probability density functions (pdf)  $\{f(x, \theta)\}$ . For every structure of missing data we have the respective marginal pdf

$$f_m(x, \theta) = \int f(x, \theta) dv_{i_1} \dots dv_{i_k}$$

Here integration is performed over the variables that are missing according to the pattern  $m$ .

Generalization of the maximum likelihood estimates to the case with missing data leads to the method of *maximum marginal likelihood* (MMML):

$$\theta(X, M) = \operatorname{argmax}_{\theta} f_M(X|\theta), \quad (2)$$

For example, if we deal with a sample of  $n$  independent observations, MMLE is

$$\operatorname{argmax} f_{m_1}(x_1|\theta) \dots f_{m_n}(x_n|\theta).$$

The identifiability condition is in our case:

*Condition I.* For any  $\theta \neq \theta_0$  there exist such a pattern  $m$  that  $f_m(x|\theta) \neq f_m(x|\theta_0)$  on  $C(\theta)$  such that

$$\int_{C(\theta)} p(m|x_{obs}) f(x|\theta_0) dv > 0.$$

The examples when this condition is true are:

1. Condition *I* holds for the multivariate normal distribution if the probability to observe any pairwise combination of variables is positive.

2. It is well-known that estimation of finite mixture distributions can be treated as the analysis with missing data (the class number is never observed and can be thought of as a missing variable). Condition *I* holds if the respective family of mixture distributions is identifiable (Teicher, 1963).

3. The probability of the complete observations is positive, MCAR is true, and  $P_\theta \neq P_{\theta_0}$  if  $\theta \neq \theta_0$ . This is the most common condition.

Nikiforov (1987) proves that MMML provides the same asymptotic properties as the usual ML method does. The required conditions are generalizations of classic identifiability and regularity conditions for *iid* observations.

**Theorem 1.** Under certain weak regularity assumptions and identifiability condition *I*, MMLE is strongly consistent for a sample of independent identically distributed random vectors from  $P(X, M)$ , if the condition MAR holds.

### 2.3 Asymptotic normality and effectiveness of MML estimates

In this section we derive the Cramer-Rao inequality for the case of missing data, i.e. we show the lower boundary of covariance matrices of estimates for the regular case, with estimates resulting only from the observed data. The traditional form does not apply here, since the pdf argument in  $f_M(x, \theta)$  is random (it varies depending on  $M$ ). However, the following statement

$$E^{X,M}(\hat{\theta} - \theta_0)(\hat{\theta} - \theta_0)^T \geq (J + D(\theta_0))I^{-1}(\theta_0)(J + D(\theta_0)) + d(\theta_0)d(\theta_0)^T \quad (3)$$

is still true for the case of missing data if we define the information matrix as

$$I(\theta) = \sum_m \int D_\theta \ln f_m(x, \theta) (D_\theta \ln f_m(x, \theta))^T p(m|x_{obs}) f(x, \theta) dv \quad (4)$$

i.e., the following theorem is true.

**Theorem 2.** Let  $f(x|\theta)$  satisfy certain regularity and differentiability conditions, the condition MAR holds, and  $I(\theta)$  be positively defined and continuous in  $\theta$ . Then Cramer-Rao inequality (3) is true for the case of missing data with information matrix  $I(\theta)$ .

**Theorem 3** further states that if  $f(x|\theta)$  is three times differentiable and satisfy other regularity conditions, then MMLE is asymptotically normal with covariance matrix  $I(t)$ , and of all solutions of the MMLE equation, there exist only one consistent solution, starting from some  $n$ , with  $P(X, M)$ -probability.

### 3 Classification of observations with missing data

We apply the same technique of replacing an original (full) distribution function by the marginal distribution functions for the problem of classification of observations with missing data.

Let us observe random vectors, with some data missing, and the distribution of the vector depends on the class we draw the current vector from, i.e.  $X$  is a random vector from one of the distributions  $P_1(X), \dots, P_k(X)$ . Further we consider these distributions as known, and derive the optimal classification rule.

The classification function in our case is the mapping  $d(X, M) = d(X_{obs}) \rightarrow (1, \dots, k)$ , i.e. classification of vectors into one of  $k$  classes using the part of  $X$  which is observed according to the missingness pattern  $M$ . The problem is to minimize the average risk

$$R(d) = \sum_j p_j E_j c(d(X, M)|j),$$

where  $c(i|j)$  is the risk (loss) matrix, and  $E_j$  is the expectation function for the distribution  $P(X, M)$  in the class  $j$ .

**Theorem 4.** Let the condition MAR hold, and the distribution of missingness pattern  $M$  does not depend on the class:

$$p(m|x, j) = p(m|x_{obs}). \quad (5)$$

Then the decision function based on the marginal pdf provides minimal risk:

$$D(X, M) = \operatorname{argmin}_i \sum_j p_j c(i|j) f_M(X|j).$$

Using this result we show that the gap filling proposed by many authors (for example, [Kennedy, Chien, 1982]) cannot improve the classification accuracy. Moreover, in many cases the filling procedure leads to average loss inflation.

The well-known change-point detection can be also considered as a classification problem, where the method performance is described often as the time spent from the process change point to the moment when the detection signal is issued (with a given intensity of false alarms). It is clear, that the optimal handling of missing data is very important in many practical applications of this problem, whether it is quality control, earthquake prediction, medical patient monitoring etc. We should expect that the technique discussed in this section is also optimal for the change-point detection.

## 4 Time series analysis with missing data

This section uses the technique of replacing an original (full) distribution function by the marginal distribution functions for the problem of time series analysis with missing data, when observations are no longer independent.

Let  $X_1^T = (X_1, \dots, X_T)$  be a sample from a time series with the probability density function (pdf)

$$f(x_1^T | x_{-\infty}^0, \theta) = \frac{f(x_{-\infty}^T | \theta)}{f(x_{-\infty}^0 | \theta)}. \quad (6)$$

If some observations from the series are missing and the missing-data mechanism is ignorable in the sense that the "missing at random" condition holds, the use of marginal distributions is justified. Thus, instead of pdf (6) the pdf

$$f_*(x_1^T | x_{-\infty}^0, \theta) = \frac{f_*(x_{-\infty}^T | \theta)}{f_*(x_{-\infty}^0 | \theta)} \quad (7)$$

applies, where the symbol  $*$  denotes integration over those elements of the time series whose values are not observed.

Consider the first order autoregression process

$$X_t = \mu + \alpha X_{t-1} + \varepsilon_t, \quad \alpha \in (-1, 1),$$

where  $\varepsilon_t$  are assumed hereafter to be random variables distributed independently with zero mean and variance  $\sigma^2$ . To specify the models completely, the normality of  $\varepsilon_t$  is also assumed. Let  $\theta_1 = (\mu, \alpha, \sigma)$ .

Let  $k - 1 \geq 0$  be the number of missing values between  $X_t$  and  $X_{t-k}$ . To specify pdf (7) for the AR (1) model we need the marginal pdf of  $X_t$  given  $X_{t-k}$ .

Let  $g(\rho, k)$  denote the sum of the geometric progression:

$$g(\rho, k) = 1 + \rho + \dots + \rho^k = (1 - \rho^{k+1})/(1 - \rho).$$

**Proposition.** The marginal distribution of  $X_t$  given  $X_{t-k}$  for autoregression process AR (1),  $k = 1, 2, \dots$ , is normal  $N(\mu(k), \sigma^2(k))$ , where  $\mu(k) = \alpha^k X_{t-k} + g(\alpha, k) \mu$  and  $\sigma^2(k) = \sigma^2 g(\alpha^2, k)$ .

*Proof.*  $X_t$  given  $X_{t-k}$  is normal as a linear combination of normal r.v., thus it is sufficient to calculate its two first moments. Taking the expectation of AR (1) recursion for  $X_{t-k+1}, \dots, X_t$ , we get

$$E(X_t | X_{t-k}, \theta_1) = \alpha^k X_{t-k} + \mu (1 + \alpha + \dots + \alpha^{k-1}) = \mu(k). \quad (8)$$

□

We apply proposition 1 to derive the EM-algorithm for the maximum likelihood estimation of  $\theta_1$  in AR (1) model. The E step provides the estimates of  $E(X_t | X_{t_1}, X_{t_2})$  and  $\text{var}(X_t | X_{t_1}, X_{t_2})$  for  $t = t_1 + 1, \dots, t_2 - 1$ , and  $\text{cov}(X_t, X_{t-1} | X_{t_1}, X_{t_2})$  for  $t = t_1 + 2, \dots, t_2 - 1$  ( $\theta_1$  is skipped in notation in this section). Little and Rubin describe the M step and derive expressions (9) for the case  $k_1 = k_2 = 1$  (see below). The simple closed-form solutions of this section complete the description of the algorithm for the general case.

The pdf of  $X_t$  given  $X_{t_1}$  and  $X_{t_2}$  is

$$f(x_t | x_{t_1}, x_{t_2}) = \frac{f(x_t, x_{t_2} | x_{t_1})}{f(x_{t_2} | x_{t_1})} = \frac{f(x_{t_2} | x_t) f(x_t | x_{t_1})}{f(x_{t_2} | x_{t_1})}.$$

Due to normality of  $X_t$  the parameters required are the coefficients  $M$  and  $S$  at  $(X_t - M)^2/S$  in the argument of the exponent:

$$E(X_t | X_{t_1}, X_{t_2}) = (\alpha^{k_1} G_2 X_{t_1} + \alpha^{k_2} G_1 X_{t_2}) / G_0 + (1 - \alpha^{k_0}) g(\alpha, k_1) g(\alpha, k_2) \mu / g(\alpha, 2k_0) \quad (9a)$$

$$\text{var}(X_t | X_{t_1}, X_{t_2}) = G_1 G_2 \sigma^2 / G_0, \quad (9b)$$

where  $G_i = g(\alpha^2, k_i)$ ,  $i = 0, 1, 2$ , with  $k_0 = k_1 + k_2 = t_2 - t_1$ .

In conclusion, we discuss the problem of initialization of time series with missing data. It is natural to consider the "observations" before time  $t = 1$  as missing. Then all that we have is to calculate the respective marginal distribution. For example, suppose that we observe the AR (1) process from  $t = 1$ . Then the marginal density of  $X_1$  is (8) with  $k = \infty$ :

$$f(x_1 | \theta_1) = \exp(-0.5(x_1 - \mu/(1 - \alpha))^2/\sigma_\infty^2)/(2\pi\sigma_\infty^2)^{1/2}$$

where  $\sigma_\infty^2 = \sigma^2/(1 - \alpha^2)$  follows from proposition 1. This is equation from (Box and Jenkins, 1970) for "exact" likelihood function derived from different considerations.

## 5 Conditional permutation homogeneity tests for multivariate samples with non-randomly missing data

This section formulates the tests and results for a two-sample problem. Then we describe their analogues for multiple samples and discuss the respective computational methods.

### 5.1 Description of tests

Let us consider the two-sample problem

$$H_0 : F = G$$

with the general alternative

$$H_1 : F \neq G.$$

We derive the homogeneity tests with statistics of Smirnov and omega-squared types for the case of missing data. In univariate case, these tests are distribution-free within the class of continuous distribution functions (DF). Direct multivariate generalizations lose this property due to the dependence between elements of the random vector. A natural and simple solution is building the tests that are conditional on the pooled sample. This approach has been proposed by (Sen P.K., Chatterjee S.K., 1964). Bickel (1969) proved

that this conditional test is consistent and distribution-free after he applied this principle the multivariate generalizations of Smirnov two-sample test.

It appears that there exist multivariate generalizations of Smirnov and omega-squared tests to the case of missing data such that they are consistent and distribution-free under very weak conditions on the distribution of missing data, compared to MCAR and MAR conditions.

Let

$$A_l = \{(X_1, M_1), \dots, (X_l, M_l)\},$$

$$B_n = \{(X_1, M_1), \dots, (X_n, M_n)\},$$

be two independent  $r$ -variate samples with missing data. We build empirical distribution functions (EDFs) for all observed missing data patterns. Thus, we get two sets of EFDs:

$$\{F_1^{l_1}(x), \dots, F_s^{l_s}(x)\},$$

$$\{G_1^{n_1}(x), \dots, G_s^{n_s}(x)\},$$

where  $F_i^{l_i}$  is the  $r_i$ -variate EDF for  $i$ -th missing data pattern.

The Smirnov type test is based on the statistic

$$D = \sup_x \left| \sum_i l_i F_i(x)/l - \sum_j n_j G_j(x)/n \right|. \quad (10)$$

Note that the sums in the equation above are not distribution functions if  $s > 1$  or  $t > 1$ , and, generally, they are even not consistent estimates of the linear combinations of the respective marginal distributions, if MCAR does not hold.

We take supremum in (10) over  $x \in R^r$ .  $D$  is compared to the critical value  $c(\alpha, \{X, M\})$ , where  $\{X, M\}$  is the pooled sample. Exactly,

$$\Phi(D, \{X, M\}) = (1, \text{ if } D > c; w, \text{ if } D = c; 0, \text{ if } D < c), \quad (11)$$

where  $\Phi$  is the critical function, and  $c$  and  $w$  are the values providing the desired significance level  $\alpha$ .

The distribution of  $D$  under  $H_0$  is defined on the set of permutations, conditional on the pooled sample  $\{X, M\}$ :

$$P(D \leq d | \{X, M\}) = \binom{n+l}{l}^{-1} \sum_i I\{D_i \leq d\}, \quad (12)$$

where  $I$  is the event indicator function, and  $D_i$  is the test statistic value for  $i$ -th of all possible variants to draw a sample of size  $l$  (without replacement) from the pooled sample.

**Theorem 5.** Let distributions of missing data are equal in the two samples:

$$p(m|x, 1) = p(m|x, 2) = p(m|x) \quad (13)$$

where  $p(m|x, i)$  is the probability of an observation with a missing data pattern  $m$ , if  $X = x$ , and arbitrary in other respects. Then the conditional Smirnov test (10), (11) is consistent and distribution-free against general alternative.

Observations with different number of present variables in have different "weight" in  $D$ . More general type of the statistic  $D$ :

$$D = \sup_x \left| \sum a_i F_i - \sum b_j G_j \right|. \quad (14)$$

can compensate this with a proper selection of coefficients  $a$  and  $b$ , for example  $a = l_i r_i / l$ , where  $r_i$  is the number of observed variables.

Smirnov statistic  $D$  provides other generalizations for the case with missing data, for example:

$$D = \max_i \sup_x |a_i F_i - b_j G_j|. \quad (15)$$

where maximum is taken over the patterns  $m$  in the pooled sample.

The statistic of omega-square type is for the problem at hand:

$$\int \left[ \sum_i a_i F_i(x) - \sum_j b_j G_j(x) \right]^2 dW(x), \quad (16)$$

where  $W(x)$  is a weight function (not necessarily DF).

In a multisample case these numerous options are multiplied by several methods to design statistics for  $k$  samples. One of the most known methods generalizes Smirnov test to:

$$D = \sup_x \sum_j^k d_j \left( \sum_i^{S_j} a_{ij} F_{ij}(x) - \sum_j^k c_j \sum_i^{S_j} a_{ij} F_{ij}(x) / \sum_j^k c_j \sum_i^{S_j} a_{ij} \right)^2, \quad (17)$$

where  $S_j$  is the number of  $m$  patterns in  $j$ -th sample,  $F_{ij}$  is  $r_{ij}$ -dimensional EDF built from  $l_{ij}$  observations with  $i$ -th structure,  $a$ ,  $c$  and  $d$  are positive numbers. Theorem formulations and proofs are pretty straightforward.

It follows from (12) that the conditional test considered in this section are distribution-free, since conditionally on a given pooled sample, the distribution

$$P^{X,M} \times \dots \times P^{X,M}(D \leq d|\{X, M\})$$

does not depend on  $P^{X,M}$  under  $H_0$ .

## 5.2 Computational techniques

The task of calculating the P-value  $Q$  for the conditional permutational tests described above can be represented as a problem of estimating of the average of the finite population with the number of elements  $C(l, n)$  (in the  $k$ -sample case,  $(n_1 + \dots + n_k)!/(n_1! \dots n_k!)$ , where  $n_1, \dots, n_k$  are the sample volumes, should be used instead of  $C(l, n)$ ).

When the test is used practically, the finiteness of  $C(l, n)$  can be ignored (in cases where  $C(l, n)$  is small,  $Q$  can be easily calculated by the method of full search), while we come to the task of estimating the parameter  $Q$  of binomial distribution.

Unfortunately, an effective Smirnov-type [Chernomordik, 1980]  $Q$  calculation scheme doesn't allow generalizing over to the multidimensional case, and even to the one-dimensional case for omega-square tests. This is also true for other methods, in which problems of  $k$ -sample homogeneity are analyzed using a model of random wandering, since in the multidimensional case, it is very difficult to establish the connection between the wandering trajectory and the statistic value. The natural (and acceptable from the practical perspective) method is the method of statistical simulations (Monte-Carlo), in which the number of simulations with the statistic having a value more or equal to the observed value is used as the estimate of  $Q$ . The precision of the estimate of  $Q$  is naturally characterized by an interval around the levels 95%, 99%, etc. For its calculation, the binomial approximation shown above can be used, and for a large enough number of trials, the Poisson approximation of binomial distribution (for  $Q$  close to 0 or 1) or the normal approximation can be used.

As with the sample volume growth and especially with multidimensionality of observations, the calculation time grows very quickly, it is reasonable to recalculate the estimate of  $Q$  and the corresponding confidence interval and to output these values to the screen every single or every few trials, so that the user can stop the process (for example, increasing the precision of

the estimate of  $Q$  is pointless if the current 99% confidence interval for  $Q$  is equal to (0.3,.8).

Note that the interruption of iterations when the given precision (interval length) is given will, on average, lead to an optimistic confidence level and  $Q$  estimate, biased to the value 0 or 1. In practice, this effect can be ignored, assuming that the user is under the influence of many outside factors when he cancels the calculation, so that precision sample and the canceling moment are random.

Generally speaking, the accuracy of calculation of  $Q$  via the Monte-Carlo method isn't significant for the consistency properties of the criterion and independence from the distribution and is important, basically, just for its power. This is a conclusion drawn from the following statement.

Let the significance level  $0 < \alpha < 1$  be defined. Call the approximate P-value with  $K$  runs, its  $Q_K$  estimate given by the Monte Carlo method after  $K$  trials. (Note. The significance level of the approximate test is the same as the significance level of the exact criterion because  $Q_K$  is an unbiased estimate of  $Q$  for any  $K > 0$ ).

**Theorem 6.** Under the conditions of theorem 5, the approximate tests of the level  $\alpha > 0$  for  $K > 0$  trials with statistics (14)-(17) are consistent.

Theorem 5 can be considered a specific case of this result. It corresponds to theorem 6 when  $K = C(l, n)$ . The generalization of the theorem 6 statement for tests with statistics (14) – (17) and for other statements from section 5.1 is fairly lucid and is skipped.

## 6 Analysis of contingency tables and testing of independence of random variables

Independence tests within the nonparametric framework also allow very weak assumptions. We start from independence testing of two variables in the two-way contingency table.

Let the experiment include receiving two-dimensional observations with  $(X, M)$  gaps, where  $X = (x_1, x_2)$  and  $M = (m_1, m_2)$ , where  $x_1$  takes the values  $i \in \{1, \dots, R\}$  and  $x_2$  the values  $k \in \{1, \dots, C\}$ . Let the distribution  $P$  satisfy one of the following conditions (function difference may, as before, be given by argument difference):

$$p(m|x_1, x_2) = p(m_1|x_1)p(m_2|x_2) \quad (18a)$$

or

$$p(m|x_1, x_2) = p(m_1|x_2)p(m_2|x_1) \quad (18b)$$

In this example, the MAR condition is one of the specific cases of (18b).

Let the cell probabilities in the contingency table  $R \times C$  equal  $p_{ij}^0, i = 1, \dots, R, j = 1, \dots, C$  when gaps are absent. Then, when gaps are present and if (18a) is true (the case (18b) has a similar analysis), cell probabilities of the conjugate table constructed using complete observations equal

$$Ap_{ij}^0 p_i^1 p_j^2 \quad (19)$$

where  $A$  is the normalizing constant and  $p_i^1$  and  $p_j^2$  are the probabilities  $p(m_1 = \text{"present"} | x_1 = i)$  and  $p(m_2 = \text{"present"} | x_2 = j)$ . During this section it will be assumed that  $p_i^1 > 0$  and  $p_j^2 > 0$  for all  $i$  and  $j$ .

From (19) one can conclude that when gaps satisfy (18a) or (18b), the independence hypothesis can be checked using complete observations with simple likelihood ratio test or chi-squares without any changes.

To prove this, we assume that independence assumption is true, i.e.,  $p_{ij}^0 = p_{i+}^0 p_{+j}^0$ , where  $p_{i+}^0 = \sum_j p_{ij}^0$  and  $p_{+j}^0 = \sum_i p_{ij}^0$ . After simple calculations we get  $p_{ij} = p_{i+} p_{+j}$ , where probabilities  $p$  are same as  $p^0$ , but are calculated from a table formed by complete observations. On the other hand, the tests specified retain their consistency against the general dependence alternative  $p_{ij}^0 \neq p_{i+}^0 p_{+j}^0$ , because this inequality leads  $p_{ij} \neq p_{i+} p_{+j}$  to be true for any  $p_i^1 > 0, p_j^2 > 0, p_{ij}^0, \sum p_{ij}^0 = 1, i = 1, \dots, R, j = 1, \dots, C$ . To prove this, it is enough to show that the system of equations relative to  $RC + 1$  unknowns  $p_{ij}^0$  and  $A$ :

$$p_{kl}^0 = A \sum_{i,j} p_i^1 p_j^2 p_{kj}^0 p_{il}^0, \sum_{i,j} p_{ij}^0 = 1,$$

has only one solution  $p_{ij}^0 = p_{i+}^0 p_{+j}^0$ . The computations associated with this proof are skipped here.

There are uniformly most powerful non-biased conditional tests known for checking independence in  $2 \times 2$  tables (equivalent to the hypothesis  $\Theta = p_{11}^0 p_{22}^0 / (p_{12}^0 p_{21}^0) = 1$ ). These are based on hypergeometrical distribution. Conducting a proof analogous to [Leman, 1975, pages 163-165] will yield that,

when gaps are present, those same tests calculated using complete observations are optimal, if (18a) or (18b) is true. Moreover, non-biased tests for checking different hypotheses about  $\Theta$  remain uniformly more powerful for alternatives indicated in [Leman, 1975, page 153] if when gaps of types (18a) or (18b) are present these tests will also only be calculated using observations with both factors present.

The results listed above reflect the circumstance that observations with one  $X$  factor component don't carry data on the dependence of the two factors if condition (18) is true.

Similar effects will be observed for a loglinear analysis of  $k$ -factor conjugate tables. For example, for checking overall independence (hypotheses about the absence of a higher level interaction in a loglinear model),

$$p(m|x) = p(m|x_1, \dots, x_k) = \prod_i p(m_i|x_{j(i)}), \quad (20)$$

is analogous to (18a) and (18b), where now  $m$  and  $x$  are random  $k$ -dimensional vectors and a  $(i(1), \dots, i(k))$  is an arbitrary permutation of the  $(1, \dots, k)$  set.

In the general case of testing independence of  $k$  random values that are not necessary discrete, with a finite number of gradations, the situation remains. That is, the independence of random values when there are no gaps,  $P^X(x) = P_1(x_1)P_2(x_2), \dots, P_k(x_k)$ , leads to independence in complete observations, if (20) is true :

$$p(x|m^*) = p_1(x_1|m^*) \dots p_k(x_k|m^*),$$

where  $m^*$  is the gap structure corresponding to the complete observation. Therefore, when (20) is true, random value independence for complete observations can be checked using the same nonparametric tests (for example, Spearman or Kendall's range tests for  $k = 2$ ) used for data without gaps. For completeness, it is noted that this kind of approach for  $k > 2$  might significantly lower the efficiency for alternatives that include partial dependence (meaning dependence inside a variable subset within the observation) and for a bigger gap percentage.

## 7 Methods for gap filling and their properties. Local filling

Generally speaking, gap-filling methods are characterized with the following two fundamental drawbacks:

1. Usually, gap filling algorithm parameters are calculated using existing data, which causes dependence between observations. Of course, this kind of artificial dependence does not arise if filling with constant or random values independent of existing observations in the sample or the values are chosen using the *cold deck* method (see section 4.5 of the book) is being conducted. In practice these methods have little value. Dependence can also be avoided by dividing the initial sample into two sub-samples and calculating fill values (for example, mid-sample values) for the sub-sample being analyzed using the observation values in the second sub-sample. This kind of approach comes with the loss of information for filling missing values.

2. The distribution of data after filling will differ from the true distribution, even if dependence, explained above, is ignored. This is especially obvious in simple filling methods – mid-sample, by regression, etc. (see section 3.4 of the book or [Basilevsky et al., 1985]). Therefore, filling with the sample mean  $\bar{X}_j, j = 1, \dots, r$  using existing values will result in a mixture distribution, one of the components of which is the true distribution based on existing values (corresponding to complete observations), and the other components are distributions based on incomplete observations with different gap structures and degenerated in

$$\{x_j = \bar{X}_j, j \in S_j\},$$

where  $S_i$  - the set of characteristics with gaps of the structure  $i$ , where  $i = 1, \dots, N$ , and  $N$  is the number of observed gap structures. Different variations of regression filling methods will also have the main component of this type of method lead to the mix of true and degenerated distributions with degeneration on the hyperplanes that contain the predicted values.

The analysis of such “complete” data using standard methods is inadequate and leads to drawbacks similar to those discussed in section 3.4: inconsistency and biased parameter estimates. The estimate quality worsens as the gap percentage increases. Analogous problems (inconsistency, distortion of the nominal importance value) are characteristic of statistical tests for hypothesis checking used for filled data.

It is reasonable to call the method described below the *local* gap filling. It is similar in its nature to one of the filling with hot deck imputation methods used for estimating the average of a one-dimensional variable in the final population with the method of the nearest neighbor. A simple example of the method of local filling is described here.

Let  $R^r$  denote the  $r$ -dimensional Euclidian sample space. Assume that the probability of a complete observation is more than 0. Let  $F(x), x \in R^r$ , be the distribution function of a random vector  $X$  (corresponding to  $P^X$ ), and  $(X_1, M_1), \dots, (X_n, M_n)$  is a sample of independent observations with gaps. Suppose that the distribution  $P$  is dominated by a measure of  $R^r$ , so that a density  $f$  exists relative to that measure.

Suppose that the  $i$ 'th observation of  $X_i$  lacks the variables  $X_{i,mis}$  but has  $X_{i,obs}$ . The calculation for the Euclidian distances between the  $i$ 'th and all the complete observations of  $X_j$  in the subspace corresponding to the variables existent in  $X_i$  is:

$$d(X_{i,obs}^{M_i}, X_j^{M_i}) \quad (21)$$

Let  $I_i = j_1, \dots, j_k$  be the subset of the indexes with minimal distance value (21). If  $I_i$  only includes one object  $j_1$ , then we take  $X_{i,mis} = X_{j_1,mis}$ . If  $k > 1$ , then the index  $j$  is selected randomly from  $I_i$  and  $X_{i,mis} = X_{j,mis}$  is assumed.

Denote an empirical distribution function built on a sample filled using this method by  $\tilde{F}_n(x)$ .

**Theorem 7.** For  $n \rightarrow \infty$  and the condition MCAR

$$\sup_x |\tilde{F}_n(x) - F(x)| \rightarrow 0 \quad (22)$$

This statement means that, for an unlimited sample volume growth, local filling provides the true distribution of the filled sample.

This conclusion results, in particular, in consistency of the estimates derived from  $\tilde{F}_n$  if they are continuous in  $F$  in uniform metrics [Borovkov, 1984, p. 26] and consistent with complete data (which include many reasonable estimates) and if local filling is used.

The simple method described above can be generalized in several directions. First of all, many different distance calculations can be used – Mahalanobis, Hemming, Kolmogorov metrics, their combinations, variable weighing, nonmetric distances etc.

Secondly, variations not limited to complete object imputation are possible, for example, filling using observations with a growing number of variables present with “accumulation” of values (this variation uses the information present in the sample more “evenly”).

In the general situation, where the gap distribution may not follow the MCAR condition, the density

$$f_0^*(x_{mis}^m | x_{obs}^m) f_m^*(X_{obs}^m),$$

corresponds to the observations with structure  $m$  for filling using the described method. This assumes the existence of the corresponding densities, where  $f_0^*$  is the density conditional to the presence of a complete observation and  $f_m^*$  is the density conditional to the presence of an observation with the structure  $m$ . The resulting marginal distribution of the filled sample has the density

$$\sum_m p(m) f_0^*(x_{mis}^m | x_{obs}^m) f_m^*(x_{obs}^m).$$

For the MCAR condition,  $f_m^*(X_{obs}^m) = f_m(X_{obs}^m)$  for all  $m$  (the right hand value of this equation contains the marginal distribution density of  $P^X$ ), which is consistent with the result (22).

Examined next is the widely known algorithm ZET [Zagoruyko, Yelkina, Timerkayev, 1976]. Externally it is similar to the local filling method. But from a mathematician’s point of view, it can’t be considered satisfactory. It is impossible to even come close to conducting a thorough inspection of the algorithm’s properties, since it is a sequence of fairly complex heuristic procedures. Instead of that the work of the mechanisms will be demonstrated on simple examples that will generally bring to distortions of the original distribution when the gaps are filled with the ZET algorithm. The issues discussed below should also be considered when constructing local filling methods.

The version of the algorithm being discussed in more detail is the latest modification – ZETM [Zagoruyko et al., 1986, chapter 2]. In this algorithm, the gaps are filled with a value that is a linear combination (weighted average) of regression estimates of the missing value. The estimates are calculated using a prediction submatrix of the original table “object-property”. The submatrix is small in size (in the example that is used in the algorithm description, it ranges from 3x3 to 10x10 in size). The exact calculation

method isn't stated in [Zagoruyko et al., 1986, chapter 2], but any reasonable explanation (is the regression used simple or multiple, etc.) leads to the conclusion that the algorithm uses the following mechanisms.

1. If the filling calculation uses more than one object, then the averaging of the values being predicted can lead to unpleasant consequences, even if the number of such objects is small. An incomplete sample of two-dimensional vectors  $(x, y)$  is examined here. A part of the objects in this sample is complete and another part contains all the  $x$ -values but has gaps in  $y$ -values. The filling will be conducted using the following "two nearest neighbors" method: two complex objects  $i_1$  and  $i_2$  with minimal distances  $|x_{i_1} - x_i|, |x_{i_2} - x_i|$  will be imputed for the  $i$ 'th object with the gap  $y$  and the gap will be filled with the value  $\hat{y}_i = (y_{i_1} + y_{i_2})/2$ . Then, as  $n \rightarrow \infty$  for distributions continuous in  $x$ , the conditional variance of  $y$  fillings will be half the true conditional variance of  $y$  for the given  $x$ . This means that if the dependence between  $x$  and  $y$  isn't very strong, the variance bias will be noticeable even with a fairly small fraction of gaps. The property (22) will only generally be true when  $y = y(x)$ .

2. The feature selection is related to another drawback, which is easiest to illustrate in an example of independent three-dimensional binary vectors  $(x, y, z), i = 1, 2, \dots, n$ , which can be expressed in the form of a three-factor contingency table  $2 \times 2 \times 2$ . Let the distribution be concentrated in 3 points:  $(0,0,0), (0,1,1)$  and  $(1,0,1)$ , each of which has a mass of  $1/3$ . Let a  $p$ -th fraction of the observations contain gaps in the variable  $z$ . The MCAR condition is still considered true. After filling the gaps using the local method (imputating the closest object and randomly picking the values if the number of objects with the minimal distance is more than one), but using only the variable  $x$  and ignoring  $y$ , the filled sample will include a fraction of observations that have impossible values:  $(0,0,1)$  and  $(0,1,0)$ . In the limit  $n \rightarrow \infty$  this fraction equals  $p/3$ .

The fact that in ZETM the columns are selected depending on their distance to each other [Zagoruyko and others, 1986, page 20] might reduce such an effect, but obviously cannot eliminate it (the case of strict linear dependence is an exception). For instance, in the given example the columns  $x$  and  $y$  are at the same distance from column  $z$  in euclidian metrics, which are used in ZETM.

3. Another possible source of distortions has the same nature. It is this way of finding filling values. Let there be a gap of the  $j$ -th variable of the  $i$ -th object,  $x_{ij}$ , that needs filling. If the fillings are calculated using the subsets

of observation that have the variable  $j$  and this subset is formulated independently from the other analogous subsets generated for the filling of gaps in other properties, then this approach can also lead to "outliers", objects with an unnatural value combination and other distortions. This approach was realized in the ZET algorithm.

4. The ZETM algorithm has an iterative mode for calculating new filling values with the values calculated in previous steps taken into consideration [Zagoruyko and others, 1986, pages 21, 115]. This can lead to additional artificial dependence between objects in the sample and amplify "centering tendencies", especially if the gap percentage is high.

Chapter 4.5.1 of the book mentions two more approaches to gap filling. Both of them generally don't satisfy the property (22). The first method (section e, see also [Titterington, Jiang, 1983, Little, Smith, 1987]) adds a random number, generated using the distribution conditional by the existing values to the parameter value equal to its current estimate of  $\hat{\Theta}$ , to the fill value for the gap, calculated using a regression equation. The fill value distribution will "fit" the true distribution to the distribution of the selected parametric model with the parameter value  $\hat{\Theta}$ . The complex method isn't satisfactory either: it is obvious that for a different distribution of deviations from the regression for different values of "independent" variables addition to the regression prediction of randomly picked residuals can noticeably distort the original distribution.

The variation close to one of the methods in [Little, Smith, 1987], the combination of regression and local filling seems more acceptable. In this method, the residual from the regression for the closest (in the space of known variables) complete observation is added to the regression prediction. The properties of such a method would be close to the above described simple local filling method and, in particular, will satisfy (22). The question which of these two approaches is more favorable (and in what conditions) remains unanswered.

When it comes to properties of the filling methods described in chapter 4.5.3 of the book, they are close to the properties of the simple local filling method described above. Note that the methods in part 4 of the book are meant to solve a specific problem, the estimation of a scalar variable property, so the situation is simpler here. Specifically, it is not necessary to carry out imputation along the whole subset of "covariables"  $x_i$ , unlike local filling in the general multidimensional case.

In conclusion of the chapter the influence of gap filling on the "non-

probabilistic” data analysis methods will be discussed and some alternative approaches will be proposed. These methods include cluster analysis methods, multidimensional scaling and other methods (for example, data visualization). Their use usually isn’t based on a probability model, so it’s pointless to attempt to characterize their properties in statistical consistence terms, (no) tests and parameter estimation shifts, stability and effectiveness. Nevertheless, filling for these methods also distorts data nature and output character. Here, if gaps aren’t dependent on property values, then filling with averages, by regression or analogous methods leads to artificial growth of the fraction of object with property values in the sample center or on corresponding hyperplanes. Classes in cluster analysis will be artificially compact after gap filling using in-group averages or using the ZETM algorithm. Also, the degree of distortion rises with gap fraction growth. Therefore, it is desirable to look for gap treatment methods unrelated to their filling (and at their absence use the local filling method) in analysis methods.

Non-probability gapped data analysis methods without filling include the approach described in chapter 5 of E. Diday’s and coauthor’s book “Data analysis methods” designed for ”template-based” cluster analysis (a generalization of the ISODATA algorithm). For analysis methods based on an object distance matrix (hierarchical cluster analysis, multidimensional scaling) it is acceptable to “fill” those distance components that are impossible to calculate because of observation gaps, meaning to add the midsampleal distance in the addition  $S_i$  to the full space or that same distance multiplied by a variable proportional to  $d_i$  to the distance  $d_i$  calculated in the subspace  $S_i$  possibly biggest for every pair of size objects, instead of filling the gaps. In multidimensional scaling, when gaps are present, it is natural to minimize the sum  $\sum p_i(\hat{d}_i - d_i)^2$  for all  $i = 1, \dots, n(n - 1)/2$  pairs of objects, where  $d_i$  is the initial and  $d_i$  is the modeled distance for the  $i$ -th pair, when a monotonously growing as the number of properties participating in the calculation of the distance  $d_i$  increases is input into  $p_i$ . All these approaches need further investigation.

## 8 Analyzing distribution of gaps and testing the randomness

Gap randomness conditions (MAR and MCAR) are an important requirement for usability of most known incomplete data analysis methods, including those described in Little and Rubin's book. Nonetheless, there are currently only a handful of specific methods for checking gap randomness, such as the simple method of comparing one-dimensional distributions, mentioned in the book (section 3.2) and described with more detail in [Little, Smith, 1987], or the MCAR condition checking method for independent variables in the task of analyzing a linear regression model in [Simonoff, 1988].

Nevertheless, it's possible to construct useful multidimensional tests to test MCAR and MAR. Checking MAR is basically only possible when the originally missing values become known, meaning by conducting more expensive or destructive measuring or by using data about an object attained some time after the study has been conducted, etc. (but not through filling the gaps with some method in which the MAR or MCAR condition itself is considered adequate).

Despite the fact that the new methods are correct under the MAR condition and do not require the stronger MCAR condition, checking the MCAR condition is also important since simple methods of treating incomplete data (for example, analysis of complete observations or local filling methods discussed in this supplement) that are generally only acceptable when MCAR is carried out, apparently won't be used in applications for a fairly long time. New methods (described, in particular, in this book) have fairly high calculation requirements. Obtaining additional observations and conducting full observation analysis might be cheaper than treating the original sample with gaps. Besides, for many hypothesis-checking tasks, no methods that work with MAR are developed. This includes, in particular, traditional hypothesis testing problems with the distribution assumed to be normal (regression, correlation, discriminant analysis and others).

First, the tests for testing the MAR condition using restored data will be discussed. The null hypothesis:

$$H'_0 : p(m|x) = p(m|x_{obs}, x_{mis}) = p(m|x_{obs}) \quad (23)$$

Let observations with various gap structures  $m_1, \dots, m_s$  be present in a random sample of i.i.d.  $r$ -dimensional observations with gaps  $(X_1, M_1), \dots, X \in$

$R^r$ . Then it can be concluded from (23) that for every structure  $m_i, i = 1, \dots, s$ , the distribution of variables absent according to  $m_i$  (with the distribution function  $F_i^1$ ) is a marginal distribution of the initial distribution  $P^X$ :

$$F_i(x_{mis}) = F(x_{mis}, \infty), \quad (24)$$

where a conditional entry in the right hand side means that the argument of  $F$  is  $x = (x_{mis}, x_{obs})$  with the values of variables belonging to  $x_{obs}$  equal to  $+\infty$  (given that the set  $m_1, \dots, m_s$  excludes the structure that corresponds to the complete observation).

Let  $s$  empirical distributions functions  $\{F_1^*, \dots, F_s^*\}$  for  $s$  structures be built based on the gap values. This way, the  $i$ -th EDF  $F_i^*$  is defined in the subspace of variables absent according to  $m_i$ . Also, built are the  $F^n(x)$ , the  $r$ -dimensional EDF of the “reconstructed” sample  $X_1, \dots, X_n$ .

To test (24), nonparametric permutational tests are proposed. These are similar in nature to the tests in chapter 5, with statistics analogous to (14)-(17) (obviously, the given task doesn’t boil can not be reduced to testing  $s$  sample homogeneity since the samples are definitely heterogeneous based on their gap structure).

$$D = \sup \sum_j^s c_j [F_j^*(x) - F_j^n(x)]^2 \quad (25)$$

is a Smirnov-type statistic.  $F(x)$  is marginal EDF of the empirical distribution function  $F(x)$  in the subspace of variables absent according to  $m$ .  $c$  is some weight, for example,  $c$  can be the number of values absent for  $m$ , number of objects with the  $m$  structure, their product, etc. The distribution (25) is defined conditionally using the restored sample  $X_1, \dots, X_n$ . It isn’t difficult to construct tests with other statistics like (15), (16), omega-square type, etc.

The MCAR condition:

$$H_0^2 : p(m|x) = p(m)$$

means that the distribution of  $X$  is the same for every gap structure present in the sample  $(X_1, M_1), \dots$  (coincides with the distribution  $F(x)$ ).

Checking the MCAR condition using the original gapped sample is only

possible in relation to existing variables:

$$F_i^2(x_{obs}) = F(x_{obs}, \infty), i = 1, \dots, s \quad (26)$$

where the vector  $x$  with  $x_{mis} = (+\infty, \dots, +\infty)$  is now the  $F$  argument, and  $F_i^2$  is the DF of variables present according to  $m_i$  (it is assumed that the set  $m_1, \dots, m_s$  excludes the structure that corresponds to an observation fully lacking values). This way, it's principally impossible to uncover those deviations from MCAR, for which (26) is true, but MAR isn't, using the original sample.

Building simple permutational tests of the type (25) to test MCAR is difficult, so the case examined here has  $P^X$  belong to a parametric family of multidimensional distributions (normal, for example). Then the hypothesis is that  $s$  samples are drawn from the distributions that are marginal in respect to  $P^X$ , and the suitable criterion would be the generalization of the likelihood relation criterion for the case of MAR-type data with gaps (the "marginal likelihood" relation criterion). The steps taken will be: get the sum  $L$  of likelihood functions for every  $s$  sample, calculated separately using regular methods for data without gaps, then calculate the likelihood function  $L$  of the initial sample with gaps  $(X_1, M_1), \dots, (X_n, M_n)$  assuming that all  $X_i$  have distribution  $P^X$  (if  $P^X$  is the normal distribution, this can be done using an EM-algorithm for a multidimensional normal distribution. See chapter 8.2 of the book and the text of the program in chapter 10 of the supplement). Then the value  $2(\ln L_1 - \ln L_2)$  has a chi-square asymptotic distribution (the number freedom degrees depends on model type of  $P^X$ , and on  $r_i$ , the number of values present for the  $i$ -th gap structure). Incidentally, it is possible to construct similar (simpler actually) parametric tests for checking the MAR condition using data with rebuilt gap values, not only using permutational tests of type (25).

Checking the MCAR condition using a restored sample boils down to a simple problem about the homogeneity of  $s$   $r$ -dimensional samples, which can be solved using parametric tests, for example from [Anderson, 1963] or using nonparametric tests. Of course, if the restored values are known, it is possible to construct criteria to test not only MAR and MCAR, but other conditions, for example (18) and (20) from chapter 5 or conditions (5) and (13) used in theorems 4 and 5 (chapters 4 and 5).

Since gaps are random objects, they can themselves be a subject for a statistical study.

In an  $r$ -dimensional sample of the volume  $N$  the gaps generate a random  $r \times N$  matrix with element values “gap” or “no gap”. For independent observations with gaps there are  $N$  independent  $r$ -dimensional binary random vectors. A hypothesis can be put forward about the equal chance of a gap in the variables:  $p_{ij} = p_{ik}$ , where  $p_{ij}$  is the probability of a gap in the  $j$ -th variable in the  $i$ -th observation and where  $i = 1, \dots, N, j, k = 1, \dots, r$ . This hypothesis can be checked as in the equal gap distribution for different objects assumption, meaning  $p_{ij} = p_{rj}$  (the corresponding asymptotic tests can be found in [Fleiss, 1989, chapter 8,4] and cited works) or without [Orlov, 1982, Nikiforova, 1989]. The tests described in [Fleiss, 1989, chapter 13] can be used to test hypotheses about the presence of dependence between gaps in different variables, also without relying on the equal gap distribution assumption for different objects. According to the works stated above, other similar hypotheses and corresponding tests can be constructed.

## 9 Software for statistical analysis of incomplete data

### 9.1 State of the art

The main instruments of applied statistical data analysis are program packages, libraries and other software. Modern statistical software is basically on the level statistical software was in the 60's (this section doesn't discuss analysis methods from the reliability theory point of view etc., only taking into consideration the problems where the mechanism of gap origination is of no direct concern to the user). Almost all statistical software that takes the possibility of data gaps into consideration contain only simple methods, such as exclusion of incomplete observations, filling the gaps with averages, filling the gaps using regression or principle components, calculation of a covariation matrix and a vector of averages using pair methods, and others, meaning methods that were realized in the first few versions of SSP, IMSL and BMD (BMDP). As it was shown before (see chapter 5 and the supplement), these methods are, generally, unacceptable. Because of this, it is pointless to examine in detail the gapped data analysis methods realized currently in software related to applied statistics (there are a few hundred of them [see Silvestrov (1988)]). It is enough to direct the reader to a short review I. S. Enukov wrote in [Ayvazyan, Enukov, Meshalkin (1983)] about analysis with missing data

capabilities of these: 3rd and 4th versions of BDMP, SPSS, PPSA, OTEKS, PNP (extension of the SSP library) and DIAS. In these and practically every other applied statistical analysis software product only a part of the simple methods described or their modifications are implemented (excluding the OTEKS packet, which is based largely on the ZETM algorithm, discussed in chapter 7).

Nonetheless, the development of statistical software based on the new approaches discussed in this book has begun and, apparently, in a few years, many statistical software products will contain realizations of modern, theoretically substantiated methods.

One of the first general statistics software packages including the new methods discussed in this book, will be the latest version of the BDMP package, planned for release in the year 1990. This package is expected to realize, for example, analysis methods for many of the models related to normal multidimensional distribution and the structures generated from section 8.5<sup>1</sup>.

In Russia, in the Center of Statistical Research and Informatics (CSRI), software including modern incomplete data analysis methods is being developed. Some of these methods include those described in the book.

One software product developed by CSRI is the dialog statistical system DISAN, which is basically a specialized tool for analyzing "object-property" data tables with gaps. All the parts of this system are designed with the presence of gaps in mind. Gap randomness checking methods are realized in the system. The powerful on-screen editor treats every table as initially filled with gaps. Every gap can have one be in one of two conditions, "missing value" or "removed value". Missing or removed values are displayed as empty cells on the screen, so the user has no need to code for the gaps using a numerical value.

The system doesn't offer any filling methods that could be a source for distorted inferences. If some problems allows for several correct gap processing approaches, the system computes all of them (if this doesn't lead to large computational requirements). The system "helps" the user to avoid false inferences related to gap presence. For example, in regression analysis for the complete observation method, the program conducts a dispersion analysis of the regression, and for the EM algorithm, the system outputs only

---

<sup>1</sup>Professor P. J. A. Little, one of this book's authors, is one of the developers of the latest version of the BDMP package

corresponding estimates.

## 9.2 On missing data analysis for a general-purpose statistical system

This section describes a list methods for analysis with missing data that are currently expedient to place in general statistics packages based on their modern level and the advanced development of their substantiation. The set of functions described is similar to the set of methods realized in the statistical analysis dialog system for gapped data developed by the Center of Statistical Research and Informatics.

For every method one or more acceptable ways of processing the gaps and also the corresponding conditions on the gap distribution are named. The following abbreviations are used: ACO - analysis of complete observations, PM - pair method. If nothing is mentioned, the gap distribution condition is the MCAR condition for ACO and PM and the MAR condition for the EM algorithm.

1. Working with data. Input and editing of gapped data. Possibility of deleting values and restoring "erased" values. Deletion by condition. Standard data manipulation methods<sup>2</sup>.

2. Testing gap randomness and studying gap distribution. Checking the MAR and MCAR conditions using conditional multidimensional permutational tests for Kolmogorov-Smirnov type statistics, omega-squares or using the likelihood ratio test for a multidimensional normal distribution. Checking hypotheses about equal gap probability in variables and about gap dependence between different variables.

3. One-dimensional random value statistics. Calculation of sample characteristics (arithmetic mean, variance, variation coefficient, range), their drawbacks and their confidence intervals. Normality test. Nonparametric one-dimensional analysis (calculation of median, percentiles, interquartile range, mode). Robust characteristic estimates. Histogram construction and distribution function estimation. Parzen density estimate.

---

<sup>2</sup>It is stated above that gap filling isn't directly included in the dialog system functions to discourage their application by the user (which can lead to false inferences). Nonetheless, using the table editor capabilities (and, for example, the one-dimensional statistics or the regression analysis sections), it is easily possible to conduct filling using the sample average, regression, regression with the addition of random upsets and other methods.

Homogeneity test of two independent samples using Kramer-Welch, Student and generalized Smirnov tests. Effect detection (checking homogeneity of two dependent samples) using Student, nonparametric sign, Smirnov distribution and other tests.

Gap processing method - gap exclusion from the sample (for checking the homogeneity of two matched samples - exclusion of pairs missing at least one observation). Gap distribution conditions depend on analysis type: for most parametric methods, ignoring the gaps is only acceptable for MCAR (or MAR, which is equivalent in the one-dimensional case with independent observations in the sample). Most nonparametric methods have weaker requirements. For example, the equality of the gap probabilities relative to the point  $F^{-1}(1/2)$  is enough for point and confidence estimation of the median, and equal gap distribution (condition (13)) in the samples is enough for homogeneity test.

4. Homogeneity test for two independent multidimensional samples using Hotelling and Bennett tests [see Anderson (1963)] (ACO gap processing method) and likelihood ratios (the calculations for this test are based on the EM-algorithm).

Multidimensional normal distribution value hypotheses checking using Hotelling (ACO) tests and likelihood ratios (the corresponding variation of the EM-algorithm is used). Homogeneity test of two dependent multidimensional samples using these tests (every observation pair's value difference is only calculated for simultaneously present variables, otherwise the difference is considered a gap).

Homogeneity of several multidimensional samples using nonparametric tests (see chapter 5 of the supplement, condition (13)).

5. Estimation of matrices of paired (ACO, PM, EM-algorithm) and partial (ACO, EM-algorithm) correlation coefficients, Spearman and Kendall's coefficients (PM). Checking of correlation coefficient value hypotheses: paired (ACO, PM) and partial (ACO). Independence test for two random variables using Kendall and Spearman's correlation coefficients (conditions (18)).

6. Regression analysis. Multiple linear regression, nonlinear regression, nonparametric regression and multifactor variance analysis (estimation - ACO and EM-algorithm, hypothesis testing - ACO).

7. Classification. Linear discriminant analysis (ACO and EM-algorithm). Cluster analysis (gap processing method described in section 7). Mixture analysis (EM-algorithm).  $k$  nearest neighbors.

8. Dimensionality reduction and data visualization. Factor analysis and

principal component analysis. Multivariate scaling (missing data processing described in section 7).

## 10 Appendix. The program of the EM-algorithm for multivariate normal distribution

The program of the EM-algorithm for multivariate normal distribution

### References

Aivazyan, S.A., I.S.Yenyukov and L.D.Meshalkin (1983) *Prikladnaja statistika* [Applied statistics]. Basics of modeling and initial data processing, 1983, 471 p. *Finansy i statistika*. Moscow (In Russian).

Borovkov, A.A. (1984), *Mathematical statistics*, Moscow (In Russian) [translation in English: Borovkov, A.A. (1998), *Mathematical statistics*, Gordon and Breach Science Publishers].

Anderson T.W. (1963) *An introduction to multivariate statistical analysis*. New York, Wiley.

Basilevski A., Sabourin D., Hum D., Anderson A. (1985) Missing data estimators in the general linear model: an evaluation of simulated data as an experimental design. *Commun. Statist.-Simula. Computa.* 14 (2), pp. 371-394.

Bickel P.I. (1969) A distribution free version of the Smirnov two sample test in the p-variate case. *Ann. Math. Statist.* 40, pp. 1-23.

Box, G. E .P. and Jenkins, G. M. (1970) *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco.

Chernomordik, O.M. (1980) On a non-parametric test for homogeneity of several samples. *Theory Probab. Applic.*, 25, 197-200.

Clarke, M.R.B. (1982), The Gauss-Jordan Sweep operator with detection of collinearity, *Applied Statistics*, vol. 31, p. 166 - 168.

Cox, D.R. and Hinkley, D. (1974) *Theoretical Statistics*. Chapman and Hall, London.

David F.N., Fix E. (1960). Rank correlation and regression in a non-normal surface. *Proc. Fourth Berkeley Symp. Math. Statist. Prob.* 6, pp.177-197. Univ. of California Press.

Diday, E. et collaborateurs: *Optimisation en classification automatique*. Rocquencourt, INRIA, 1979.

- Fleiss J.L., Statistical analysis for rates and proportions, New York, Wiley, 1981.
- Hajek A. and Sidak F.G. (1967) Theory of rank tests. Academia. Prague.
- Kennedy D.P., Chien Y.T. (1982) Optimal estimation for full space classification of incomplete data, in Pattern Recognition and Image Processing, Proceedings. Las Vegas, pp. 152-154.
- Krzysko M. (1983). The discriminant analysis of multivariate time series. IEEE Trans. on Inform. Theory, vol. 29, p. 612-614.
- Lehmann E. (1959) Testing Statistical Hypotheses, Springer, New York.
- Little R.J.A., Smith P.J. (1987) Editing and imputation for quantitative survey data. Journal of American Statistical Association, vol. 82, No.397, pp.58-68.
- Nikiforov A.M. (1987) Pattern recognition with unsupervised classification and statistical analysis with missing data. Ph.D. Dissertation, MIPT, 144 p. (in Russian)
- Nikiforov A.M. (1989) Statistical analysis with randomly missing data. The Fifth International Conference on Probability Theory and Statistics. Vilnius, p.98-99 (in Russian).
- Nikiforova G.V. (1989) Nonparametric hypothesis tests of random binary matrices in asymptotic of growing number of parameters. The Fifth International Conference on Probability Theory and Statistics. Vilnius, p.100-101 (in Russian).
- Orlov A.I.. (1982) Paired comparisons in Kolmogorov asymptotic. Expert estimates in control problems, Moscow, pp. 58-66 (in Russian)
- Patrick E. A. Fundamentals of Pattern Recognition, Prentice Hall, 1972.
- Rubin D.B. (1976) Inference and missing data. Biometrika, vol. 63, pp.581-592.
- Sen P.K., Chatterjee S.K. (1964). Nonparametric tests for the bivariate two sample location problem. Calcutta Statist. Ass. Bull., vol. 13, pp.18-58.
- Simonoff J.S. (1988) Regression diagnostics to detect nonrandom missingness in linear regression. Technometrics, vol. 30, No. 2, pp. 205-214.
- Titterton D.M., Jiang J.M. (1983) Recursive estimation procedures for missing data problems. Biometrika, vol. 70, pp. 613-624.
- Zagoruiko N.G., V.N. Elkina and V.S. Temirkaev (1976) ZET - a gap filling algorithm in experimental data tables, Computing Systems, vol. 67, Novosibirsk: Nauka, pp. 3-28.