A new method of multiple imputation for completely (or almost completely) missing data

ARKADY BOLOTIN
Epidemiology Department
Ben-Gurion University of the Negev
Beersheba, Israel
arkadyv@bgu.ac.il

One of the important questions the researcher must answer assessing data quality while preparing information for a data mining procedure is whether missing observations in the dataset are missing at random, and whether some form of imputation is needed. If all (or almost all) observations of a variable are missing, they cannot be classified as missing at random. Therefore, most known methods of imputation of missing values cannot be applied to this variable. This paper studies a particular way for creating imputations in datasets containing completely (or almost completely) missing variables. As it is shown in the paper, if no external data are available, the maximum entropy distribution is the only reasonable probability distribution for producing proper imputation in case of such variables. Two examples of real-life epidemiological studies demonstrate this approach.

Keywords: missing variables, non-random missingness; maximum entropy distributions.

1. INTRODUCTION

Before any data mining algorithm can be applied, a target set must be first cleaned from the observations with missing data. However, such cleaning may threaten the very goal of extracting patterns from data if the missing values create a biased sample. Improper handling of missing values will distort the patterns uncovering because, until proven otherwise, the researcher must assume that data with missing values differ in analytically important ways from data where values are present. That is, removing missing values is not so much the problem of reduced sample size as it is the possibility that the remaining dataset is biased.
In order to choose a proper strategy of dealing with missing values, it is important to understand why the data are missing.

Missing completely at random (MCAR) exists when missing values are randomly distributed across all observation $i = 1, 2, \dots, N$ in the dataset (here, the terminology is after [17] [22]). In the case of MCAR, probability of not observing a value x_i of variable X is unrelated to the value x_i or to the value of any other variables in the dataset (i.e. the probability is random).

Missing at random (MAR) is a condition, which exists when missing values are not randomly distributed across all observations but are randomly distributed within one or more subsamples. In the case of MAR, missingness of x_i does not depend on the value x_i *after controlling for another variable* (i.e. within subsamples, missingness is random).

Missing not at random (MNAR) (or ***not missing at random*** (NMAR)) is a situation when missing values x_i are distributed randomly neither across all observations in the dataset, nor within one or more subsamples. In the case of MNAR, missingness of x_i is not random but has an order in it (it is so-called *non-ignorable* missingness [10]).

Among different types of missingness encountering in data, MNAR is not unusual especially in the form of completely or almost completely missing variables.

A variable X is ***completely missing*** if values x_i of X are missing for all observations $i = 1, 2, \dots, N$ in

the dataset. Since all values x_i are missing, they obviously cannot be classed as missing at random (that is why this is the case of MNAR – missingness of x_i has the order). If, for example, the dataset contains an additive variable $A = X + W$ combining values x_i and w_i, then completely missing values x_i will cause the biased mean for A.

Let values x_i of a variable X be observed for O observations and missing for the rest $N - O$ observations in the dataset. We will call the variable X as ***almost completely missing*** if $O \ll N$ and missing values x_i happen in one or more continuous sequences of observations i, that is, if values x_i are completely missing within one or more subsamples. Again, here the missing values x_i cannot be classified as missing at random, and therefore this is the case of MNAR too. Clearly, in this case, the mean for the non-missing values x_i will not be an unbiased estimate of the mean X that we would have obtained with complete data.

So, when we have a completely missing or almost completely missing variable, we have a problem. In this case, the only way to obtain an unbiased estimate of parameters is to model missingness. In other words, we would need to write a model that imputes missing values, that is, fills in missing observations with plausible values. That model could then be incorporated into a more complex model for estimating missing values. However, to write such model is not a trivial task to do. That is because the common methods of imputations of missing values such as *regression substitution* [15] or *maximum likelihood estimation* (MLE) [18] [26] will not work in the case of completely (or almost completely) missing variables.

For instance, regression substitution predicts what missing values x_i of a variable X should be based on values z_i, w_i, …, v_i of other variables $Z, W, \dots, V$ in the dataset that are not missing $x_i = \beta_z z_i + \beta_w w_i + \cdots \beta_v v_i + \varepsilon_i$. But if the variable X is completely missing, then such analysis cannot be performed since the regression equation will be undetermined, and there will be no data to recover coefficients β_z, $\beta_w, \dots, \beta_v$. If the variable X is almost completely missing, then in general we cannot exclude the possibility that either there will be not enough data to estimate unique values for coefficients β_z, $\beta_w, \dots, \beta_v$, or the regression equation will have no solution or many solutions.

As to the MLE method, it assumes missing values x_i are MAR, which does not hold true when the variable X is completely missing or almost completely missing [18].

Given that the common methods of filling in missing observations will not work, we may try to simulate missing values. That is to say, we would try to replace every missing value x_i by one or more simulated versions of it. Then we would analyze the simulated complete dataset (or datasets) by standards methods, and if we have more than one imputation we would combine the results to produce estimates and confidence intervals that would incorporate missing data uncertainty. The main problem here is how to simulate missing values, to be exact, how to draw from the probability distribution in order to produce proper imputation.

As a matter of fact, most existing commercially available software procedures that can run simulated imputation (PROC MI and PROC MIANALYZE in SAS [14], mi impute and mi estimate in Stata [4] [5], or NORM and PAN in S-PLUS [23] [24] – just to mention a few) assume that the missing data are MAR. But what is more important, they all generate imputed values on the basis of existing data – just as regression substitution or MLE algorithm does – and then add an error component drawn randomly (to introduce the necessary level of uncertainty into the imputed value: this is known as "random imputation") (see, for example, [13] [16] [25]). However, as we said before, this will not work in the case of completely or almost completely missing variables, for there will be no data or not enough data to impute values.

This paper studies a particular way for creating imputations in datasets with completely or almost completely missing variables.

2. ENTROPY OF THE OBSERVATION

Assume that x_{i1} is the arbitrary value imputed for the missing observation x_i of the variable X, and $\Pr(x_i = x_{i1})$ is the probability that x_i would be equal to x_{i1} if x_i was not missing. Similarly, $\Pr(x_i = x_{i2})$ is the probability that x_i would be equal to x_{i2} if x_i was not missing, and so on, thus $\sum_{k=1}^{M} \Pr(x_i = x_{ik}) = 1$. If the variable X is continuous, than the last constraint should be $\int p(x_i)dx_i = 1$. Of course, proceeding from our background knowledge about the variable X and its plausible relationship with other variables in the dataset, we can also put some additional constraints on the probabilities $\Pr(x_{ik})$ or on the probability density function $p(x_i)$.

Let us define *the entropy of the observation* x_i as

$$H(x_i) = -\sum_{k=1}^{M} \Pr(x_{ik}) \cdot \log \Pr(x_{ik}) \quad , \qquad (1)$$

(here $\Pr(x_{ik}) \log \Pr(x_{ik}) = 0$ whenever $\Pr(x_{ik}) = 0$), or as

$$H(x_i) = -\int p(x_i) \cdot \log p(x_i)\ dx_i \qquad (2)$$

when X is continuous (again, here $p(x_i) \log p(x_i) = 0$ whenever $p(x_i) = 0$).
If the variable X had no missing values at all, the observation x_i would be known to be equal to its only possible value, say, x_{im}. In that case, the entropy of the observation x_i would be equal to zero:

$$\begin{aligned} H(x_i) = &-\underbrace{\Pr(x_{im})}_{=1} \log \Pr(x_{im}) \\ &-\sum_{k \neq m}^{M} \underbrace{\Pr(x_{ik})}_{=0} \log \Pr(x_{ik}) \qquad (3) \\ &= 0 \ . \end{aligned}$$

If the variable X had missing at random values, the observation x_i – if missing – would be known (from regression substitution, for example) to be situated within the confidence interval, say, from x_{il} to x_{iu}. For the sake of simplicity, let us assume that the probabilities $\Pr(x_{ik})$ are zero outside the interval $[x_{il}; x_{iu}]$ in which they are equal to each other and to $\Pr(x_{ik}) = [1 + (u - l)]^{-1}$. Then, the entropy of the observation x_i would be:

$$\begin{aligned} H(x_i) = &-\sum_{k=l}^{u} \Pr(x_{ik}) \log \Pr(x_{ik}) \\ &= \log[1 + (u - l)] \quad . \end{aligned} \qquad (4)$$

As it can be readily seen, the less is the difference $(u - l)$ (which determines the interval $[x_{il}; x_{iu}]$ containing all plausible values for x_i), the nearer to zero is the entropy of the observation x_i.

In contrast, if the variable X is completely (almost completely) missing, then the observation x_i is totally unknown. Suppose we have no background knowledge about the variable X, so that the only limitation on the probabilities $\Pr(x_{ik})$ is the constraint $\sum_{k=1}^{M} \Pr(x_{ik}) = 1$. Then, all values x_{ik} we might impute instead of missing x_i would be equally plausible, which means that all the probabilities $\Pr(x_{ik})$ would be the same $\Pr(x_{ik}) = M^{-1}$. In that case, the entropy of the observation x_i would be equal to its *maximum value*:

$$\begin{aligned} H(x_i) = &-\sum_{k=1}^{M} \Pr(x_{ik}) \log \Pr(x_{ik}) \\ &= \log M \quad . \end{aligned} \qquad (5)$$

Hence, if we merely added additional limitations on the probabilities $\Pr(x_{ik})$, we would find the entropy $H(x_i)$ equal to its *maximum value allowed* by those limitations, i.e. by our background knowledge about the variable X:

$$\log[1 + (u - l)] < H(x_i) < \log M \quad . \qquad (6)$$

Now, imagine for a moment that in order to produce imputations, we choose a probability distribution $\Pr'(x_{ik})$ whose entropy $H'(x_i)$ is lower than the entropy $H(x_i)$ allowed by all the constraints we could derive from the background knowledge

about the variable X. Evidently, this would mean either making up values of the variable X, or filling in them by using data otherwise *external* to the analyzed dataset (as, for example, guessing ethnicity completely missing in the study dataset based on the population census data associated with the patient's address). Then again, to choose a probability distribution $\Pr''(x_{ik})$ with the entropy $H''(x_i)$ higher than the entropy $H(x_i)$ allowed by the constraints would mean to ignore those constraints.

Thus, if no external data are available, *the maximum entropy distribution* is the only reasonable probability distribution for producing proper imputation in case of completely or almost completely missing variables.

3. VARIANCE ESTIMATE

Assume that in the study dataset the dependent (analyzed) variable Y is completely observed whereas the independent (analyzing) variable X might be affected by missing at random values. Given the event $X = x_i$, the mean Y takes the following form (in matrix notation) $E(\mathbf{Y}|\mathbf{X}) = f(\mathbf{X})$, where $\mathbf{Y} = \{y_i\}_{i=1}^{N}$ and $\mathbf{X} = \{x_i\}_{i=1}^{N}$ are vectors of observations of the variables Y and X, respectively (in both vectors the subscript i indexes a particular observation).

If the variable X is completely missing, the expression $E(\mathbf{Y}|\mathbf{X}) = f(\mathbf{X})$ has no sense since the components x_i of the covariate vector $\mathbf{X}$ do not exist. Therefore, in that case, instead of each observation x_i we may try to use its plausible expected value $E(x_i)$:

$$x_i \longmapsto E(x_i) = \sum_{k=1}^{M} x_{ik} \Pr(x_{ik}) \quad , \tag{7}$$

or – if X is a continuous variable –

$$x_i \longmapsto E(x_i) = \int_{-\infty}^{\infty} x_i p(x_i) dx_i \quad , \tag{8}$$

where $\Pr(x_{ik})$ (or $p(x_i)$) is the probability distribution (density function) with maximum entropy defining all plausible values for the observation x_i. The problem is, however, that due to the lack of relevant background knowledge about the variable X, the constraints we may derive from that knowledge would be only approximate, i.e. falling within some non-zero intervals. Therefore, to account for this uncertainty in the constraints, it might be better to use the imputed sample means $\langle x_i \rangle_{(m)}$ for the observations x_i

$$x_i \longmapsto \langle x_i \rangle_{(m)} = \frac{1}{m} \sum_{k=1}^{m} x_{ik} \quad , \tag{9}$$

where each $\langle x_i \rangle_{(m)}$ is calculated on the set of m imputed values x_{ik} identically distributed with distribution function $\mathrm{F}(x_i)$

$$\begin{aligned} \mathrm{F}(x_i) &= \sum_{x_{ik} \le x_i} \Pr(x_i = x_{ik}) \\ &= \sum_{x_{ik} \le x_i} \mathrm{p}(x_{ik}) \quad , \end{aligned} \tag{10}$$

or – if X is a continuous variable –

$$\mathrm{F}(x_i) = \int_{-\infty}^{x_i} p(x)\, dx \quad , \tag{11}$$

where $\mathrm{p}(x_{ik})$ (or $p(x)$) is the probability distribution (density function) that maximizes the entropy of the observation x_i with respect to the approximate constraints prescribed.

In general, the replacement for an arbitrary function of the observation x_i, $f(x_i)$, would be the following imputed point estimator $\langle f(x_i) \rangle_{(m)}$:

$$f(x_i) \longmapsto \langle f(x_i) \rangle_{(m)} = \frac{1}{m} \sum_{k=1}^{m} f(x_{ik}) \quad . \tag{12}$$

The uncertainty in $f(x_i)$ can be then calculated using the formula for the variance of $f(x_i)$ over the set of the m imputed values x_{ik}:

$$\mathrm{Var}[f(x_i)]_{(m)} = \frac{1}{m-1}\sum_{k=1}^{m}\left[f(x_{ik}) - \langle f(x_i)\rangle_{(m)}\right]^2 \quad . \tag{13}$$

Consequently, the average of the uncertainty $\mathrm{Var}[f(x_i)]_{(m)}$ over the distribution of the N observations x_i would be

$$\langle \mathrm{Var}[f(x)]_{(m)}\rangle_{(N)} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{Var}[f(x_i)]_{(m)} \quad . \tag{14}$$

On the other hand, the average of the estimator $\langle f(x_i)\rangle_{(m)}$ along with its variance over the observations x_i would be

$$\langle\langle f(x)\rangle_{(m)}\rangle_{(N)} = \frac{1}{N}\sum_{i=1}^{N}\langle f(x_i)\rangle_{(m)} \tag{15}$$

and

$$\mathrm{Var}\left[f(x)_{(m)}\right]_{(N)} = \frac{1}{N-1}\sum_{i=1}^{N}\left\{\langle f(x_i)\rangle_{(m)} - \langle\langle f(x)\rangle_{(m)}\rangle_{(N)}\right\}^2 \quad , \tag{16}$$

respectively. Therefore, the combined variance estimate $\mathrm{Var}[f(x)]$ for the arbitrary function $f(x)$ would be

$$\mathrm{Var}[f(x)] = \mathrm{Var}\left[f(x)_{(m)}\right]_{(N)} + \langle \mathrm{Var}[f(x)]_{(m)}\rangle_{(N)} \quad . \tag{17}$$

Theoretical example #1.

Consider, for example, the simplest case of the function $f(x) = x$. Assume that the imputed values x_{ik}, replacing values x_i of the completely missing variable X, are obtained from the continuous uniform distribution (maximum entropy distribution):

$$p(x_{ik}) = \begin{cases} (a_i + b_i)^{-1} & , \quad a_i \le x_{ik} \le b_i \\ 0 & , \ x_{ik} < a_i \ \text{ or } \ x_{ik} > b_i \end{cases} \quad (k = 1, \dots, m) \quad , \tag{18}$$

where the maximum boundary b_i can be written down through the interval $[a_i, b_i]$ length, Δb_i: $b_i = a_i + \Delta b_i$. According to (12), (13) and (15), we have

$$\langle x_i\rangle_{(m)} = \frac{1}{m}\sum_{k=1}^{m} x_{ik} = \bar{x}_\iota \underbrace{=}_{m\to\infty} a_i + \frac{1}{2}\Delta b_i \quad , \tag{19}$$

$$\mathrm{Var}(x_i)_{(m)} = \frac{1}{m-1}\sum_{k=1}^{m}(x_{ik} - \bar{x}_\iota)^2 = \underbrace{=}_{m\to\infty} \frac{1}{12}{\Delta b_i}^2 \quad , \tag{20}$$

$$\langle \mathrm{Var}(x)_{(m)}\rangle_{(N)} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{12}{\Delta b_i}^2 = \frac{1}{12}\,\overline{\Delta b^2} \quad . \tag{21}$$

As

$$\langle\langle x\rangle_{(m)}\rangle_{(N)} = \frac{1}{N}\sum_{i=1}^{N}\bar{x}_\iota = \bar{a} + \frac{1}{2}\overline{\Delta b} \tag{22}$$

and

$$\mathrm{Var}\left[x_{(m)}\right]_{(N)} = \frac{1}{N-1}\sum_{i=1}^{N}\left[\left(a_i + \frac{1}{2}\Delta b_i\right) - \left(\bar{a} + \frac{1}{2}\overline{\Delta b}\right)\right]^2 = {S_a}^2 + \frac{1}{4}{S_{\Delta b}}^2 \quad , \tag{23}$$

where ${S_a}^2 = \overline{a^2} + \bar{a}^2$ and ${S_{\Delta b}}^2 = \overline{\Delta b^2} + \overline{\Delta b}^2$, we finally get

$$\mathrm{Var}(x) = {S_a}^2 + \frac{1}{4}{S_{\Delta b}}^2 + \frac{1}{12}\,\overline{\Delta b^2} \quad . \tag{24}$$

The last expression can be simplified if we assume that all the intervals $[a_i, b_i]$ are of the same length Δb; in that case, we will have

$$\mathrm{Var}(x) = {S_a}^2\left(1 + \frac{\Delta b^2}{12\,{S_a}^2}\right) \quad . \tag{25}$$

Although the variance estimate (17) is logically straightforward, we can introduce another variance estimate – the symmetric inversion of (17) with respect to the m and N sets:

$$\mathrm{Var}[f(x)]_+ = \mathrm{Var}\big[f(x)_{(N)}\big]_{(m)} + \langle \mathrm{Var}[f(x)]_{(N)}\rangle_{(m)} \quad , \tag{26}$$

where $\mathrm{Var}\big[f(x)_{(N)}\big]_{(m)}$ represents so-called the between-imputation variability B (here the terms' names follow [1] [9])

$$\mathrm{B} = \frac{1}{m-1}\sum_{k=1}^{m}\big\{\langle f(x_k)\rangle_{(N)} - \langle\langle f(x)\rangle_{(N)}\rangle_{(m)}\big\}^2 \quad , \tag{27}$$

and $\langle \mathrm{Var}[f(x)]_{(N)}\rangle_{(m)}$ represents the within-imputation variability W

$$\mathrm{W} = \frac{1}{m}\sum_{k=1}^{m}\mathrm{Var}[f(x_k)]_{(N)} \quad . \tag{28}$$

Let us observe how $\mathrm{Var}[f(x)]_+$ is related to $\mathrm{Var}[f(x)]$.

Theoretical example #2.

For that purpose, let us again consider the previous case of the simplest function, $f(x) = x$. Assuming now that

$$p(\overline{x_k}) = \begin{cases} (a+b)^{-1} \ , & a \le \overline{x_k} \le b \\ 0 \ , & \overline{x_k} < a \ \text{ or } \ \overline{x_k} > b \end{cases} \quad (k = 1, \dots, m) \quad , \tag{29}$$

where $\overline{x_k} = \frac{1}{N}\sum_{i=1}^{N} x_{ki}$ and $b = a + \Delta b$, we immediately obtain these results

$$\langle\langle x\rangle_{(N)}\rangle_{(m)} = \frac{1}{m}\sum_{k=1}^{m}\overline{x_k} = \overline{x} \underset{m\to\infty}{=} a + \frac{1}{2}\Delta b \quad , \tag{30}$$

$$\mathrm{B} = \frac{1}{m-1}\sum_{k=1}^{m}(\overline{x_k} - \overline{x})^2 \underset{m\to\infty}{=} \frac{1}{12}\,\Delta b^2 \quad . \tag{31}$$

On the other hand, comparing $\overline{x_k}$ and (30), we infer that if $a = m^{-1}\sum_k a_k$ and $\Delta b = m^{-1}\sum_k \Delta b_k$, then $\frac{1}{N}\sum_{i=1}^{N} x_{ki} = a_k + \frac{1}{2}\Delta b_k$, which gives us this

$$\mathrm{Var}(x_k)_{(N)} = \frac{1}{N-1}\sum_{i=1}^{N}\left[x_{ki} - \left(a_k + \frac{1}{2}\Delta b_k\right)\right]^2 = {S_{a_k}}^2 + \frac{1}{4}{S_{\Delta b_k}}^2 \quad , \tag{32}$$

and afterward, this

$$\mathrm{W} = \frac{1}{m}\sum_{k=1}^{m}\left({S_{a_k}}^2 + \frac{1}{4}{S_{\Delta b_k}}^2\right) = \overline{{S_a}^2} + \frac{1}{4}\overline{{S_{\Delta b}}^2} \quad . \tag{33}$$

Finally we get the expression symmetrical to (24)

$$\mathrm{Var}(x)_+ = \overline{{S_a}^2} + \frac{1}{4}\overline{{S_{\Delta b}}^2} + \frac{1}{12}\,\Delta b^2 \quad . \tag{34}$$

As it can be seen from the last two examples, the variance estimate $\mathrm{Var}[f(x)]$ posits that in a study dataset each observation of a completely (almost completely) missing variable contains not one but multiple imputed plausible versions of the missing value. In contrast, the variance estimate $\mathrm{Var}[f(x)]_+$ implies that every observation of the missing variable holds only one imputed value but the study dataset has not one but multiple versions so that each version of the dataset contains a different copy of the imputed value. Clearly, the variance estimate $\mathrm{Var}[f(x)]_+$ is computationally more preferable because that estimation suggests

using full, complete data sets on which to perform analyses, and these analyses can be performed by nearly any method or software package available.

4. MULTIPLE-IMPUTATION STATISTICS

In our practical examples, we use the multiple-imputation inference based on the variance estimate $\mathrm{Var}[f(x)]_+$. Here is a brief account of the statistics we employ in those examples.

The variance estimate $\mathrm{Var}[f(x)]_+$ in the form of (26) entails a big number of imputations m, which, of course, might be not practically achievable. So, we use the adjusted formula of the estimate $\mathrm{Var}[f(x)]_+$ allowing for a small m [20] [21]

$$\mathrm{Var}[f(x)]_+^* = \mathrm{W} + \left(1+\frac{1}{m}\right)\mathrm{B} \quad . \tag{35}$$

Suppose that the imputed point estimator

$$\begin{aligned}\langle f(x_k)\rangle_{(N)} &= \frac{1}{N}\sum_{i=1}^{N} f(x_{ik}) \\ &\equiv f(x_k) \quad (k=1,\dots,m)\end{aligned} \tag{36}$$

has the expected value $E[f(x_k)] = \epsilon$, and its average over the m complete data sets is

$$\langle f(x_k)\rangle_{(m)} = \frac{1}{m}\sum_{k=1}^{m} f(x_k) \equiv \overline{f(x)} \quad . \tag{37}$$

Then the ratio

$$t = \frac{\epsilon - \overline{f(x)}}{\sqrt{\mathrm{Var}[f(x)]_+^*}} \tag{38}$$

will be approximately distributed as a Student's t-distribution with v_m degrees of freedom

$$v_m = (m-1)\left(1+\frac{1}{\mathrm{R}}\right)^2 \quad , \tag{39}$$

where R is so-called the relative increase in variance due to nonresponse [21]

$$\mathrm{R} = \frac{\left(1+\frac{1}{m}\right)\mathrm{B}}{\mathrm{W}} \quad . \tag{40}$$

With a large value of m or a small value of R, the degrees of freedom v_m will be large. However, if v_m is much larger than the complete dataset degrees of freedom v_N, it is inappropriate. Therefore, instead of (39) we use the formula for adjusted degrees of freedom v_m^* [2] [3] [19]

$$v_m^* = \left(\frac{1}{v_m} + \frac{1}{\widehat{v_N}}\right) \quad , \tag{41}$$

where

$$\widehat{v_N} = \frac{v_N+1}{v_N+3}\, v_N \left[1 - \frac{\left(1+\frac{1}{m}\right)\mathrm{B}}{\mathrm{Var}[f(x)]_+^*}\right] \quad . \tag{42}$$

As we said, due to uncertainty in the constraints, we replace the plausible expected value $E[f(x_k)]$ of the estimator $f(x_k)$ with its imputed sample mean $\overline{f(x)}$. To evaluate suitability of such replacement we use the ratio E

$$\mathrm{E} = \frac{1}{\left(1+\frac{\Gamma}{m}\right)} \quad , \tag{43}$$

which gives the relative efficiency of using the finite m rather than an infinite number, in units of the variance estimate$\mathrm{Var}[f(x)]_+^*$, and where the statistic Γ is called the rate of missing information [2] [21]

$$\Gamma = \frac{\mathrm{R} + \frac{2}{(v_m+3)}}{\mathrm{R}+1} \quad . \tag{44}$$

In spite of being well-known in multiple imputation, the statistics R and Γ acquire a rather different meaning when a variable is missing completely (almost completely). Take for instance the theoretical example #2 that we have just considered above. In that case, the statistics R and Γ take the forms (if $\Delta b_k = \Delta b$ for all k)

$$\mathrm{R} \underset{m\to\infty}{=} \frac{\mathrm{B}}{\mathrm{W}} = \frac{\Delta b^2}{12\overline{S_a}^2} \tag{45}$$

And

$$\Gamma \underset{m\to\infty}{=} \frac{\Delta b^2}{\Delta b^2 + 12\overline{S_a}^2} \quad . \tag{46}$$

Obviously, these expressions, (45) and (46), cannot be categorized as the increase in variance due to nonresponse (missingness) or the rate of missing information because now both values Δb^2 and $\overline{S_a}^2$ are hypothetical outcomes of a "what-if" distribution of the completely missing variable x. Instead – if we take into consideration that the parameter a may be obtained from some non-probabilistic (deterministic) model, and the length Δb determines the entropy of x distribution, $\Delta b = e^{H(x)}$, – the statistics R and Γ can be called as *the increase in variance due to uncertainty* and *the rate of uncertainty*, respectively.

EXAMPLE 1. OBESITY IN YOUTH

Consider a real epidemiological study whose objective was the extraction of information about the relationship between subjects' ages and their levels of physiologically active bloodstream substances (such as triglyceride and hemoglobin) in overweight and obese young persons[1]. In the study dataset, which is a collection of records of 91 teenagers – 60 girls and 31 boys, by some cause a person's age has been registered in whole years without the fractional part (by the way, this is the practice adopted in many epidemiological studies). Let a_i represent a person's age value for the i^{th} observation: $a_i = \lfloor a_i \rfloor + \{a_i\}$, where $\lfloor a_i \rfloor$ is the floor function: $\lfloor a_i \rfloor = \max\{n \in \mathbb{Z} | n \le a_i\}$ (n is the integer, $\mathbb{Z}$ is the set of integers), and $\{a_i\}$ is the fractional part of a_i. It follows then that in fact the study dataset contains one more variable whose values are completely missing – the variable $\{a\}$. As a result, the study dataset is biased: for example, the mean age computed from the study dataset $\overline{\lfloor a \rfloor}$ is not an unbiased estimate of the mean age $\overline{\lfloor a \rfloor} + \overline{\{a\}}$ that we would compute if we got complete data with non-missing $\{a\}$.

We use Pearson's product-moment correlation coefficient and the Spearman's rank correlation coefficient to study the relationship between subjects' ages and their levels of triglyceride and hemoglobin. The Table 1 contains the observed correlation coefficients:

Sex	Substance	Correlation	*r*	Sample size	*p*-value	95% confidence interval	
Girls	Triglyceride	Pearson	0.27	30	0.145	-0.10	0.58
		Spearman	0.31		0.098	-0.06	0.60
	Hemoglobin	Pearson	0.06	31	0.769	-0.31	0.40
		Spearman	0.03		0.856	-0.32	0.38
Boys	Triglyceride	Pearson	-0.06	18	0.828	-0.51	0.42
		Spearman	-0.10		0.707	-0.54	0.39
	Hemoglobin	Pearson	0.39	17	0.127	-0.12	0.73
		Spearman	0.28		0.273	-0.23	0.67

Table 1. Correlation between age and bloodstream substances

Since the 95% confidence intervals include zero in both substances and both sexes, it is likely that the corresponding population correlations ρ are not significantly different from zero, at the 95% level of confidence. However, it is also likely that that those results are biased because of the completely missing fractional part of the subject's age, the variable$\{a_i\}$. The only way to sort out this doubt is to impute $\{a_i\}$.

The sole constraint, which we can put on the probability density function $p(\{a_i\})$ determining the distribution of plausible values for $\{a_i\}$, stems directly from the nature of this variable:

$$p(\{a_i\}) = 0 \quad \text{if } \{a_i\} \notin [0;1) \quad , \tag{47}$$

and this is all we can possibly know about the missing variable $\{a_i\}$. Among all continuous distributions, subject to the constraint (47), only one – the uniform distribution $U(0, 1-\varepsilon)$

[1] The data for the example are by courtesy of Ms. Anat Altschuler, who collected and used them in her undergraduate thesis.

$$U(0,1-\varepsilon)=\begin{cases}1 & \text{for } x\in[0;\ 1-\varepsilon]\\ 0 & \text{otherwise}\end{cases} \quad (48)$$

(where ε is an infinitesimal) – has the maximum entropy [11] [12]. For that reason, we use this distribution (choosing for ε the value 2^{-32}) to fill in the variable $\{a_i\}$ with plausible values. Generating $m=100$ values for each observation i of $\{a_i\}$ and calculating correlation coefficients r on each of the 100 imputed complete datasets and then combining the results, we get the following table:

Sex	Substance	Correlation	Overall *r*	Complete-data observation	Overall *p*-value	Overall 95% confidence interval		Imputation statistics*	
								R	Γ
Girls	Triglyceride	Pearson	0.27	30	0.149	-0.12	0.59	0.030	0.093
		Spearman	0.30		0.104	-0.09	0.61	0.034	0.096
	Hemoglobin	Pearson	0.05	31	0.773	-0.33	0.42	0.026	0.088
		Spearman	0.03		0.858	-0.35	0.41	0.047	0.107
Boys	Triglyceride	Pearson	-0.06	18	0.825	-0.55	0.46	0.034	0.136
		Spearman	-0.09		0.720	-0.59	0.45	0.110	0.201
	Hemoglobin	Pearson	0.38	17	0.135	-0.18	0.75	0.038	0.145
		Spearman	0.26		0.307	-0.31	0.70	0.082	0.183

Table 2. Correlation between imputed age and bloodstream substances (* - Efficiency of the correlation estimate based on 100 imputations is not less than 0.99)

The results presented in this table concur with the conclusion we have already made – that likely there is no connection between the study subjects' ages and their levels of triglyceride and hemoglobin.

EXAMPLE 2. HEART FAILURE EVENTS AND WEATHER

Let us consider a real observational study analyzing whether variations in day-to-day temperature T can predict daily frequency of heart failure events (such as congestive unspecified heart failure, left ventricle failure, and death from heart failure) on the example of the population living in the Northern District, Israel [6]. Despite the fact, that the study dataset (which comprises temperature and numbers of heart failure events registered every day over the period from January 1, 2000 to October 3, 2006 – overall 2468 observations) has no missing values for heart failure frequency, temperature reading was available only from May 10, 2004 to December 31, 2005 (601 observations).

Variables	*N*	Mean	Variance	Min Value	Max Value
Unspecified heart failure	2468	0.723	0.888	0	6
Left ventricle failure	2468	0.139	0.163	0	3
Death from heart failure	2468	0.297	0.307	0	4
Daily temperature *T*	580	22.759	45.770	6.4	33.9

Table 3. Summary statistics

Even within this interval the variable T has 21 missing observations (presumably MAR-type), but outside the interval, T is completely missing. As it follows from the table 3 presenting summary statistics of the study dataset (where *N* stands for the number of non-missing observations; Variance is for the variable variance) T is missing in 1,888 observations (i.e. the variable T is classed as almost completely missing). Therefore, any conventional analysis while ignoring the information about the heart failure events in these 1,888 observations will lose power and – more importantly – be potentially biased. The actual results of the Poisson regression of the number of daily occurrences of each heart failure event analyzed on the non-missing observations of T demonstrate this:

Analyzed Variable	*N*	Coef. β	Robust Std. Err.	*p*-value	95% confidence interval	
Unspecified heart failure	580	-0.0138	0.0073	0.060	-0.0282	0.0006
Left ventricle failure	580	-0.0389	0.0282	0.167	-0.0942	0.0163
Death from heart failure	580	-0.0148	0.0102	0.146	-0.0348	0.0052

Table 4. Poisson regression models on the observed temperature

(where Coef. β is the regression coefficient, Robust Std. Err. stands for the robust estimate of the standard error for β). As we can see, all the 95% confidence intervals include zero, therefore, it is likely that there is no connection between the heart failure events and temperature. However, we cannot rid ourselves of a strong suspicion that this

conclusion is wrong because of the biased study dataset.

Is it possible to preserve the information about the heart failure events in those 1,888 observations in the analysis? The answer is yes, and the solution is to impute missing values of the variable T. To do this, let us first inspect what we know about the variable T.

It is known that the daily temperature T can be viewed as a cyclostationary process [7]. This means, that, even if values T_i of different days $i \in [1; 2468]$ are statistically different, temperature of the days, which are divided by some regular intervals, will have identical statistics. Thus, we can describe the random process composed of daily temperatures as the set of interleaved stationary processes, each of each takes on a new value once for a certain period (a year, for example).

Mathematically, a cyclostationary process T_i can be expressed in an additive form [8]: $T_i = \widetilde{T}_\iota + \mathcal{T}_i$, where $\widetilde{T}_\iota$ is the deterministic cyclic process with period τ which can be represented by a Fourier series

$$\widetilde{T}_\iota = \sum_{n=1}^{L} \left[a_n \sin\left(\frac{2\pi}{\tau} ni\right) + b_n \cos\left(\frac{2\pi}{\tau} ni\right) \right] + a_0 \quad , \tag{49}$$

whilst $\mathcal{T}_i$ is the stochastic process described by some probability distribution with zero mean $\overline{\mathcal{T}_\iota}$ and non-zero variance $\overline{\mathcal{T}_\iota^2}$.

Fortunately for us, the deterministic function $\widetilde{T}_\iota$ can be estimated with the non-missing measurements of the daily temperature T_i presented in the study dataset. In fact, fitting the model of $\widetilde{T}_\iota$

$$\widetilde{T}_\iota = \sum_{n=1}^{4} \left(a_n \sin \omega_y ni + b_n \cos \omega_y ni \right) + a_0 \quad \left(\omega_y = \frac{2\pi}{365.4} \right) \tag{50}$$

on the observed values of T_i using linear regression, we get the result:

$$\widehat{\widetilde{T}}_\iota = 21.4697 - 0.5633 \cos \omega_y 3i - 1.0983 \cos \omega_y 2i - 3.7552 \sin \omega_y i - 7.9494 \cos \omega_y i \tag{51}$$

With regard to the stochastic process $\mathcal{T}_i$, all we know about its distribution is its mean $\overline{\mathcal{T}_i} = 0$ and variance $\overline{\mathcal{T}_i^2}$ which can be estimated by computing statistics of the residual $T_i - \widehat{\widetilde{T}}_i$:

Variable	*N*	Mean	Variance	Min Value	Max Value
Residual $T - \hat{T}$	580	0.0000	7.928	-7.0201	8.8515

Table 7. Summary statistics of the stochastic component of daily temperature

Among all distributions with the given mean and variance, only the normal distribution has the maximum entropy [12] [27] [28]. Therefore, it reasonable to assume that the distribution of the stochastic process $\mathcal{T}_i$ is normal:

$$\mathrm{F}(\mathcal{T}_i) = \Phi\left(\frac{\mathcal{T}_i - 0}{\sqrt{7.928}}\right) \quad . \tag{52}$$

For each of the 1,888 observations with missing T_i, we calculate one value of the estimated deterministic process $\widehat{\widetilde{T}}_\iota$ and create $m = 100$ imputations for the stochastic process $\mathcal{T}_i$; thus, we get:

Analyzed variable	Complete-data *N*	Overall Coef. β	Overall Robust Std. Err.	Overall *p*-value	Overall 95% confidence interval		Imputation statistics*	
							R	Γ
Unspecified heart failure	2468	-0.0242	0.0038	0.000	-0.0316	-0.0167	0.0886	0.0823
Left ventricle failure	2468	-0.0368	0.0091	0.000	-0.0546	-0.0190	0.1433	0.1265
Death from heart failure	2468	-0.0167	0.0056	0.003	-0.0277	-0.0058	0.0928	0.0859

Table 8. Poisson regression models on the imputed daily temperature.

Based on these results, we conclude that it looks as if a connection between daily temperature T

and rate of heart failure events does exist but we just failed to find it in the first place, owing to missing T measurements.

REFERENCES

1. Allison, P.D. (2000), Multiple Imputation for Missing Data: A Cautionary Tale. – *Sociological Methods and Research*, 28, 301–309.
2. Barnard, J. and Rubin, D.B. (1999), Small-Sample Degrees of Freedom with Multiple Imputation. – *Biometrika*, 86, 948–955.
3. Barnard, J., and Meng, X.L. (1999), Applications of Multiple Imputation in Medical Studies: From AIDS to NHANES. – *Statistical Methods in Medical Research*, 8, 17–36.
4. Carlin, J. B., Greenwood, N. Li, P. and Coffey, C. (2003), Tools for analyzing multiple imputed datasets. – *Stata Journal* 3: 226–244.
5. Carlin, J. B., J. C. Galati, and P. Royston. (2008), A new framework for managing and analyzing multiply imputed data in Stata. – *Stata Journal* 8: 49–67.
6. Friger M., Novack, V., Bolotin, A. and Novack, L. (2010) Weather–mortality association in patients with chronic heart failure: 5-year longitudinal study across four different regions in Israel – *ISES-ISEE 2010 Proceedings*, Seoul, Korea (Forthcoming).
7. Gardner, W. A., Napolitano, A., and Paura, L. (2006), Cyclostationarity: Half a century of research. – *Signal Processing*, 86: 639–697. doi: 10.1016/ j.sigpro. 2005. 06.016.
8. Gardner, W.A. (2002), Two alternative philosophies for estimation of the parameters of time-series. – *IEEE Transactions on Information Theory*, 37 Issue: 1, 216-218. doi: 10.1109/18.61145.
9. Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004), *Bayesian Data Analysis*, 2nd ed. London: Chapman & Hall/CRC.
10. Graham, J. W. (2009), Missing data analysis: Making it work in the real world. – *Annual Review of Psychology*, 60: 549–576.
11. Guiasu, S. and Shenitzer, A. (1985), The principle of maximum entropy. – *The Mathematical Intelligencer*, 7(1), 42-48.
12. Harremoës P. and Topsøe F. (2001), *Maximum Entropy Fundamentals*. – *Entropy*, 3(3), 191-226.
13. Horton, N. J., and Kleinman, K. P. (2007), Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. – *American Statistician*, 61: 79–90.
14. Horton, N.J. and Lipsitz, S.R. (2001), Multiple Imputation in Practice: Comparison of Software Packages for Regression Models With Missing Variables. – *Journal of the American Statistical Association*, 55, 244–254.
15. Jones, M.P. (1996), Indicator and stratification methods for missing explanatory variables in multiple linear regression. – *Journal of the American Statistical Association*, 91,222-230.
16. Kenward, M. G., and Carpenter, J. R. (2007), Multiple imputation: Current perspectives. – *Statistical Methods in Medical Research*, 16: 199–218.
17. Little, R.J.A. and Rubin, D.B. (2002), *Statistical Analysis with Missing Data*, *2nd ed.*, New York: John Wiley & Sons, Inc.
18. Pampel, Fred C. (2000). *Logistic regression: A primer*. Sage Quantitative Applications in the Social Sciences Series #132. Thousand Oaks, CA: Sage Publications. (pp. 40-48 provide a good discussion of maximum likelihood estimation).
19. Reiter, J. P. (2007), Small-sample degrees of freedom for multi-component significance tests with multiple imputation for missing data. – *Biometrika*, 94: 502–508.
20. Rubin, D.B. (1976), Inference and Missing Data. – *Biometrika*, 63, 581–592.
21. Rubin, D.B. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley & Sons, Inc.
22. Rubin, D.B. (1996), Multiple Imputation After 18+ Years. – *Journal of the American Statistical Association*, 91, 473–489.
23. Schafer, J.L. (1997), *Analysis of Incomplete Multivariate Data*, New York: Chapman and Hall.
24. Schafer, J.L. (1999), *Software for multiple imputation.* – www.stat.psu.edu/~jls/ misoftwa.html.
25. Scheuren, F. (2005). Multiple imputation: How it began and continues. – *The American Statistician*, 59: 315-319.
26. Schneider T. (2001), Analysis of Incomplete Cli-

mate Data: Estimation of Mean Values and Covariance Matrices and Imputation of Missing Values. – *Journal of Climate*, 14: 853-871.

27. Uffink, J., (1995), Can the Maximum Entropy Principle be explained as a consistency requirement? – *Studies in History and Philosophy of Modern Physics*, 26B: 223-261.

28. Uffink, J., (1995), *Constraints in the Maximum Entropy Principle*, Preprint: University of Utrecht, Utrecht, Netherlands.