The Impact of Missing Values on Different Measures of Uncertainty

Chantal Larose[1], Dipak K. Dey[2], Ofer Harel[2]

[1]*Eastern Connecticut State Univeristy.* [2]*University of Connecticut.*

Supplementary Material

This supplement contains the proofs for Theorems 1 through 4.

S1 Proof of Theorem 1

The component-wise distribution of missingness in y_{2i} is $r_{2i} \sim f(r_{2i}|y_{1i}, y_{2i}, \phi)$. Since we are under the MCAR mechanism, $r_{2i} \sim Bernoulli(\phi)$, where ϕ is the complement of the percent missing in the data. Parameters of y_1, y_2 and r (θ and ϕ respectively) have been suppressed in the following derivations.

The entropy of one record is

$$\begin{aligned} H(x_i) &= -\int_{y_{1i},y_{2i},r_{2i}} f(y_{1i}, y_{2i}, r_{2i}) ln f(y_{1i}, y_{2i}, r_{2i}) d(y_{1i}, y_{2i}, r_{2i}) \\ &= -\int_{y_{1i}}\int_{y_{2i}}\int_{r_{2i}} f(y_{1i}, y_{2i}, r_{2i}) ln f(y_{1i}, y_{2i}, r_{2i}) dr_{2i} dy_{2i} dy_{1i} \end{aligned}$$

To separate the joint distribution of y_{1i}, y_{2i}, and r_{2i}, we use $f(y_{1i}, y_{2i}, r_{2i}) = f(y_{1i})f(y_{2i}|y_{1i})f(r_{2i})$. Thus, the above entropy is reduced to

$$
\begin{aligned}
H(x_i) &= -\int_{y_{1i}}\int_{y_{2i}}\int_{r_{2i}} f(y_{1i})f(y_{2i}|y_{1i})f(r_{2i})ln\left[f(y_{1i})f(y_{2i}|y_{1i})f(r_{2i})\right]dr_{2i}dy_{2i}dy_{1i} \\
&= -\int_{y_{1i}}\int_{y_{2i}}\int_{r_{2i}} f(y_{1i})f(y_{2i}|y_{1i})f(r_{2i})\left[lnf(y_{1i}) + lnf(y_{2i}|y_{1i}) + lnf(r_{2i})\right]dr_{2i}dy_{2i}dy_{1i} \\
&= -\int_{y_{1i}}\int_{y_{2i}}\int_{r_{2i}} f(y_{1i})f(y_{2i}|y_{1i})f(r_{2i})lnf(y_{1i})dr_{2i}dy_{2i}dy_{1i} \\
&\quad -\int_{y_{1i}}\int_{y_{2i}}\int_{r_{2i}} f(y_{1i})f(y_{2i}|y_{1i})f(r_{2i})lnf(y_{2i}|y_{1i})dr_{2i}dy_{2i}dy_{1i} \\
&\quad -\int_{y_{1i}}\int_{y_{2i}}\int_{r_{2i}} f(y_{1i})f(y_{2i}|y_{1i})f(r_{2i})lnf(r_{2i})dr_{2i}dy_{2i}dy_{1i} \\
&= -\int_{y_{1i}}\int_{y_{2i}} f(y_{1i})f(y_{2i}|y_{1i})lnf(y_{1i})\left(\int_{r_{2i}} f(r_{2i})dr_{2i}\right)dy_{2i}dy_{1i} \\
&\quad -\int_{y_{1i}}\int_{y_{2i}} f(y_{1i})f(y_{2i}|y_{1i})lnf(y_{2i}|y_{1i})\left(\int_{r_{2i}} f(r_{2i})dr_{2i}\right)dy_{2i}dy_{1i} \\
&\quad -\int_{y_{1i}}\int_{y_{2i}} f(y_{1i})f(y_{2i}|y_{1i})\left(\int_{r_{2i}} f(r_{2i})lnf(r_{2i})dr_{2i}\right)dy_{2i}dy_{1i},
\end{aligned}
$$

The first and second integral equal one; the third is the entropy of the distribution of r_{2i}, denoted $H(r_{2i})$:

$$
\begin{aligned}
H(x_i) &= -\int_{y_{1i}}\int_{y_{2i}} f(y_{1i})f(y_{2i}|y_{1i})lnf(y_{1i})dy_{2i}dy_{1i} \\
&\quad -\int_{y_{1i}}\int_{y_{2i}} f(y_{1i})f(y_{2i}|y_{1i})lnf(y_{2i}|y_{1i})dy_{2i}dy_{1i} \\
&\quad + H(r_{2i})\int_{y_{1i}}\int_{y_{2i}} f(y_{1i})f(y_{2i}|y_{1i})dy_{2i}dy_{1i}.
\end{aligned}
$$

Pull terms out of the integral over y_{2i}:

$$H(x_i) = -\int_{y_{1i}} f(y_{1i}) ln f(y_{1i}) \left(\int_{y_{2i}} f(y_{2i}|y_{1i}) dy_{2i} \right) dy_{1i}$$
$$- \int_{y_{1i}} f(y_{1i}) \left(\int_{y_{2i}} f(y_{2i}|y_{1i}) ln f(y_{2i}|y_{1i}) dy_{2i} \right) dy_{1i}$$
$$+ H(r_{2i}) \int_{y_{1i}} f(y_{1i}) \left(\int_{y_{2i}} f(y_{2i}|y_{1i}) dy_{2i} \right) dy_{1i}.$$

The first and third integral equal one; the second is the entropy of the distribution of $y_{2i}|y_{1i}$, denoted $H(y_{2i}|y_{1i})$:

$$H(x_i) = -\int_{y_{1i}} f(y_{1i}) ln f(y_{1i}) dy_{1i}$$
$$+ H(y_{2i}|y_{1i}) \int_{y_{1i}} f(y_{1i}) dy_{1i}$$
$$+ H(r_{2i}) \int_{y_{1i}} f(y_{1i}) dy_{1i}.$$

The second and third integrals equal one, while the third is the entropy of the distribution of y_{1i}, denoted $H(y_{1i})$:

$$H(x_i) = H(y_{1i}) + H(y_{2i}|y_{1i}) + H(r_{2i}), \forall i. \tag{S1.1}$$

Since the records are independent and identically distributed, we sum Equation S1.1 n:

$$H(x) = \sum_{i=1}^{n} (H(y_{1i}) + H(y_{2i}|y_{1i}) + H(r_{2i})) = nH(y_1) + nH(y_2|y_1) + nH(r_2). \tag{S1.2}$$

The above is the framework for entropy of an MCAR incomplete bivariate normal dataset.

The following are clear from our assumptions, introductory mathematical statistics, and textbook entropy derivations:

- $y_{1i} \sim N_1(\mu_1, \sigma_1^2)$, therefore $H(y_{1i}^o) = \frac{1}{2}ln(2\pi e\sigma_1^2), \forall i$, for fixed y_{i1}^o.
- $y_{2i}|y_{1i}^o \sim N_1(\mu_{2.1}, \sigma_{2.1}^2 = \sigma_2^2(1-\rho^2))$, therefore $H(y_{2i}|y_{1i}) = \frac{1}{2}ln(2\pi e\sigma_2^2(1-\rho^2)), \forall i$
- $r_{2i} \sim Bern(\phi)$, therefore $H(r_{2i}) = -(1-\phi)ln(1-\phi) - \phi ln(\phi), \forall i$.

Plug the above into Equation S1.2 to complete the proof.

S2 Proof of Theorem 2

Begin with entropy for bivariate normal data with Bernoulli missingness:

$$\frac{n}{2}ln(2\pi e\sigma_1^2) + \frac{n}{2}ln(2\pi e\sigma_2^2(1-\rho^2)) - n(1-\phi)ln(1-\phi) - n\phi ln(\phi).$$

To see whether $\lim_{\phi\to 0}[n(1-\phi)ln(1-\phi) + n\phi ln(\phi)] = 0$:

$$\lim_{\phi \to 0}[n(1-\phi)ln(1-\phi) + n\phi ln(\phi)] = n \lim_{\phi \to 0}[(1-\phi)ln(1-\phi)] + n \lim_{\phi \to 0}[\phi ln(\phi)]$$

$$= n \lim_{\phi \to 0}[(1-\phi)] \lim_{\phi \to 0}[ln(1-\phi)] + n \lim_{\phi \to 0}[\phi ln(\phi)]$$

$$= n \lim_{\phi \to 0}[ln(1-\phi)] + n \lim_{\phi \to 0}[\phi ln(\phi)],$$

since $\lim_{\phi \to 0}[(1-\phi)] = 1.$

$$n \lim_{\phi \to 0}[ln(1-\phi)] + n \lim_{\phi \to 0}[\phi ln(\phi)] = n \lim_{\phi \to 0}[\phi ln(\phi)] = 0.$$

S3 Proof of Theorem 3

This proof follows the same structure as the proof of Theorem 1.

We pull $H(r_{2i}|y_{1i}^o)$ out of the integral over y_1 because entropy focuses on the distribution of $r_{2i}|y_{1i}^o$ and not the realized values. Since $H(r_{2i}|y_{1i}^o) = -(1-\phi^*)ln(1-\phi^*) - \phi^* ln(\phi^*)$, the entropy term is pulled out of the integral over y_{1i}^o, and y_{1i}^o is fixed. Applying the same steps as in the proof for Theorem 1, we obtain:

$$H(y_{1i}^o) + H(y_{2i}|y_{1i}^o) + H(r_{2i}|y_{1i}^o). \tag{S3.1}$$

The records are independent but not identically distributed, due to the realized values of y_{1i}^o impacting the value of ϕ^*. Therefore, sum Equation

S3.1, n times to account for the n records:

$$\sum_{i=1}^{n} \left(H(y_{1i}^{o}) + H(y_{2i}|y_{1i}^{o}) + H(r_{2i}|y_{1i}^{o}) \right)$$

$$= nH(y_{1i}^{o}) + nH(y_{2i}|y_{1i}^{o}) + \sum_{i=1}^{n} \left(H(r_{2i}|y_{1i}^{o}) \right). \tag{S3.2}$$

The above is a framework for entropy of an MAR incomplete bivariate normal dataset.

The following are clear:

- $y_{1i} \sim N_1(\mu_1, \sigma_1^2)$, therefore $H(y_{1i}^{o}) = \frac{1}{2}ln(2\pi e\sigma_1^2)$
- $y_{2i}|y_{1i}^{o} \sim N_1(\mu_{2.1}, \sigma_{2.1}^2 = \sigma_2^2(1-\rho^2))$, therefore $H(y_{2i}|y_{1i}^{o}) = \frac{1}{2}ln(2\pi e\sigma_2^2(1-\rho^2))$
- $r_{2i} \sim Bern(\phi^*)$, therefore $H(r_2) = -(1-\phi^*)ln(1-\phi^*) - \phi^* ln(\phi^*)$.

Plug in the above into Equation S3.2:

$$\frac{n}{2}ln(2\pi e\sigma_1^2) + \frac{n}{2}ln(2\pi e\sigma_2^2(1-\rho^2)) - \sum_{i=1}^{n} \left((1-\phi_i^*)ln(1-\phi_i^*) + \phi_i^* ln(\phi_i^*) \right)$$

$$= \frac{n}{2}ln(2\pi e\sigma_1^2) + \frac{n}{2}ln(2\pi e\sigma_2^2(1-\rho^2)) - \sum_{i=1}^{n}\{(1-\phi_i^*)ln(1-\phi_i^*)\} - \sum_{i=1}^{n}\{\phi_i^* ln(\phi_i^*)\},$$

which completes the proof.

S4 Proof of Theorem 4

We are looking at

$$\lim_{\phi^* \to 0} \left(\sum_{i=1}^{n} \{(1 - \phi_i^*) ln(1 - \phi_i^*)\} - \sum_{i=1}^{n} \{\phi_i^* ln(\phi_i^*)\} \right).$$

Note that $\phi^* = \dfrac{e^{\beta_0 + y_{1i}}}{1 + e^{\beta_0 + y_{1i}}}$, where y_{1i} are the values in the data set. We cannot let $\phi^* \to 0$, since the y_{1i} values are fixed. Instead, we examine the behavior of the only arbitrary parameter in ϕ^*, β_0.

If ϕ^* goes to zero, all elements ϕ_i^* go to zero; a requirement satisfied using properties of the logistic function. For a realized value of y_{1i}, ϕ_i^* goes to zero when β_0 goes to negative infinity. Therefore, we consider $\lim_{\beta_0 \to -\infty}$, which results in $\phi_i^* \to 0$ for all i.

Our limit equation is:

$$\lim_{\phi_i^* \to 0} \left(\sum_{i=1}^{n} \{(1 - \phi_i^*) ln(1 - \phi_i^*)\} - \sum_{i=1}^{n} \{\phi_i^* ln(\phi_i^*)\} \right),$$

which goes to zero following the same steps presented in the proof to Theorem 2.