Some weak convergence under missing data

Alexis BOULIN

4 juin 2021

Let us describe our statistical experiment. We denote by $(\Omega, \mathcal{A}, \mathbb{P})$ a probability space. We observe a independent and identically distributed sample $(I_i, I_i X_i)_{i=1}^n$ from the random vector (I, X) where I_i is an indicator function who take the value 1 if we observe the data at i or 0 otherwise. We suppose that $p_X = \mathbb{P}(I = 1) > 0$. In our experiment, we want to estimate the expectancy of a certain measurable transformation of X. In other words, we want to estimate the following quantity $\mathbb{P}_X f := \int f d\mathbb{P}_X$ where $\mathbb{P}_X$ denote the marginal distribution of X.

In order to estimate within the missing data framework, we can't introduce the usual empirical measure but rather a weighted empirical measure given by :

$$\mathbb{P}_n = \frac{1}{np_X} \sum_{i=1}^n \delta_{\{(X_i, I_i)\}} \tag{1}$$

Where $\delta_{\{x\}}$ is the diract measure on x. Nevertheless, this empirical measure is only valid if the probability of missing if known, which can be never the case in practice. When p_X is unknown, we may consider the following empirical measure :

$$\tilde{\mathbb{P}}_n = \sum_{i=1}^n \delta_{\{(X_i, I_i, \sum_{i=1}^n I_i)\}} \tag{2}$$

We may note, for brievity $\tilde{n} = \sum_{i=1}^n I_i$ which is the number of data that were really observed in the sample. We gonna study the weak convergence of the estimator (1) where the analysis of the second is given in a second section.

1 Weak convergence under missing data with known probability of missing

In this section, we will study the convergence of the first estimator with a proof using characteristic function. We introduce the following notation :

$$\mathbb{P}_i = \delta_{\{(X_i, I_i)\}}$$

In order to write $\mathbb{P}_n = \frac{1}{np_X} \sum_{i=1}^n \mathbb{P}_i$. To estimate the desired quantity $\mathbb{P}_X f$, we may introduce the following function :

$$\begin{aligned} h \colon \mathbb{R} \times \{0, 1\} &\to \mathbb{R} \\ (x, y) &\mapsto f(x) y \end{aligned}$$

where f is a measurable function. We are now able to define our estimator.

Definition 1. *The estimator of the quantity $\mathbb{P}_X f$ is given by :*

$$\mathbb{P}_n g = \frac{1}{np_X} \sum_{i=1}^n f(X_i) I_i \tag{3}$$

We need some condition on the distribution of the vector (I, X) in order to guarantee the convergence and the wean convergence of our estimator. These are the following :

Condition 1. *1. We suppose that the pair (I,X) are independent.*

2. We suppose that $\mathbb{E}[f(X)]$ and $\mathbb{E}[f(X)^2]$ are finite.

Under these conditions, we can state the following :

Theorem 1. *Under the condition 1, we have :*

$$\mathbb{P}_n g \longrightarrow \mathbb{P} f \quad a.s. \quad n \to \infty \tag{4}$$

Theorem 2. *Under the framework given by conditions 1, we may have the following weak convergence :*

$$\sqrt{n}(\mathbb{P}_n g - \mathbb{P} f) \rightsquigarrow \mathcal{N}(0, \sigma_g^2) \tag{5}$$

Where $\sigma_g^2 = (p_X^{-1}\mathbb{P} f^2) - (\mathbb{P} f)^2$

Démonstration. In order to prove this statement, we have made use of characteristic functions. We denote by ϕ_X the characteristic function of X. By definition, we have for all $t \in \mathbb{R}$

$$\phi_{\sqrt{n}(\mathbb{P}_n f - \mathbb{P} f)}(t) = \mathbb{E}[e^{it\sqrt{n}(\mathbb{P}_n f - \mathbb{P} f)}] = \mathbb{E}[e^{\frac{it}{\sqrt{n}p_X}\sum_{i=1}^n (f(X_i)I_i - p_X \mathbb{P}_f)}] = \Pi_{i=1}^n \phi_{\mathbb{P}_i f - p_X \mathbb{P} f}(\frac{t}{\sqrt{n}p_X})$$

Where we use independency in the last inequality. We can compute $\mathbb{E}[\mathbb{P}_i f - p_X \mathbb{P}_X f] = 0$ and $var[\mathbb{P}_i f - p_X \mathbb{P}_X f] = Var[\mathbb{P}_i f] = p_X(\mathbb{P}_X f^2 - p_X(\mathbb{P}_X f)^2) \leq \infty$. We can now use a taylor expansion to obtain :

$$\begin{aligned}
\phi_{\mathbb{P}_i f - p_X \mathbb{P} f}(\frac{t}{\sqrt{n}p_X}) &= 1 + i\mathbb{E}[\mathbb{P}_i f - p_X \mathbb{P} f]\frac{t}{\sqrt{n}p_X} - \frac{t^2}{2np_X^2} var[\mathbb{P}_i f - p_X \mathbb{P}_X f] + o(\frac{1}{n}) \\
&= 1 - \frac{t^2}{2np_X^2} p_X(\mathbb{P}_X f^2 - p_X(\mathbb{P}_X f)^2) + o(\frac{1}{n}) \\
&= 1 - \frac{t^2}{2n}(p_X^{-1}\mathbb{P}_X f^2 - (\mathbb{P}_X f)^2) + o(\frac{1}{n})
\end{aligned}$$

Using the identical distribution of the sample, we then have :

$$\begin{aligned}
\phi_{\sqrt{n}(\mathbb{P}_n f - \mathbb{P} f)}(t) &= (\phi_{\mathbb{P}_1 f - p_X \mathbb{P} f}(\frac{t}{\sqrt{n}p_X}))^n \\
&= exp\{nlog(1 - \frac{t^2}{2n}(p_X^{-1}\mathbb{P}_X f^2 - (\mathbb{P}_X f)^2) + o(\frac{1}{n}))\} \\
&= exp\{-\frac{t^2}{2}(p_X^{-1}\mathbb{P}_X f^2 - (\mathbb{P}_X f)^2) + o(1)\}
\end{aligned}$$

tending n to infinity give the statement of the theorem. □

2 Weak convergence under missing data with unknown probability of missing

In this section, we will tackle the more realistic case with unknown probability of missing. As said if the first section, we now estimate the number of missing data by counting them in the sample. We introduce the following function :

$$\begin{aligned}
h\colon \mathbb{R} \times \{0,1\} \times \{0,1\ldots,n\} &\to \mathbb{R} \\
(x,y,v) &\mapsto \frac{f(x)y}{v}
\end{aligned}$$

This notation give us all tools to define our estimator

Definition 2. *The estimator of the quantity $\mathbb{P}_X$ is given by;*

$$\tilde{\mathbb{P}}_n h = \frac{1}{\sum_{i=1}^n I_i} \sum_{i=1}^n f(X_i) I_i \tag{6}$$

Theorem 3. *Under the condition 1, we have :*

$$\tilde{\mathbb{P}}_n g \longrightarrow \mathbb{P} f \quad a.s. \quad n \to \infty \tag{7}$$

Theorem 4. *Under the framework given by conditions 1, we may have the following weak convergence :*

$$\sqrt{n}(\tilde{\mathbb{P}}_n g - \mathbb{P}f) \rightsquigarrow \mathcal{N}(0, \sigma_h^2) \tag{8}$$

Where $\sigma_h^2 = p_X^{-1}\left(\mathbb{P}f^2 - (\mathbb{P}f)^2\right)$

Démonstration. Before starting the proof, we may introduce the following notations :

$$\begin{aligned} h_1 &= f(X_i)I_i \\ f_1 &= I_i \\ X_n(f) &= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} f(I_i, X_i) - \mathbb{E}[f(I_i, X_i)]\right) \end{aligned}$$

We know, by the central limit theorem that :

$$X_n(f) \rightsquigarrow \mathcal{N}(0, \mathbb{P}(f - \mathbb{P}f)^2) \quad n \to \infty$$

Furthermore, we have that :

$$\tilde{\mathbb{P}}_n h = \frac{p_X \mathbb{P}_X f + n^{-1/2} X_n(g_1)}{p_X + n^{-1/2} X_n(f_1)}$$

We claim that :

$$\sqrt{n}(\tilde{\mathbb{P}}_n h - \mathbb{P}_X f) = p_X^{-1} X_n(h_1 - f_1 \mathbb{P}_X f) + o_p(1) \tag{9}$$

Indeed, using the equation below gives :

$$\begin{aligned} p_X(\tilde{\mathbb{P}}_n h - \mathbb{P}_X f) &= n^{-1/2} X_n(h_1) - n^{-1/2} X_n(f_1)\tilde{\mathbb{P}}_n h \\ &= n^{-1/2} X_n(h_1 - f_1 \mathbb{P}_X f) - n^{-1/2} X_n(h_1)(\tilde{\mathbb{P}}_n h - \mathbb{P}_X f) \end{aligned}$$

Then we have :

$$\sqrt{n}(\tilde{\mathbb{P}}_n h - \mathbb{P}_X f) = p_X^{-1} X_n(h_1 - f_1 \mathbb{P}_X f) - p_X^{-1} X_n(f_1)(\tilde{\mathbb{P}}_n h - \mathbb{P}_X f)$$

By the CLT $X_n(f_1)$ weakly converge to a centered gaussian law with variance $\mathbb{P}(f_1 - \mathbb{P}f_1)^2$. Furthermore, we know by theorem 3 that $\tilde{\mathbb{P}}_n h - \mathbb{P}_X f \to 0 \quad a.s.$ Applying Slutsky theorem gives the claim. It suffice now to notice that $X_n(h_1 - f_1 \mathbb{P}_X f)$ weakly converge to a centered gaussian with variance $p_X^{-1}\mathbb{P}_X(f - \mathbb{P}_X f)^2$ and apply again Slutsky's theorem to conclude. □