Incomplete Data Analysis, Assignment 1, Exercise 1+2

Miriam Fischer

17 10 2020

Exercise 1

1a: Correct answer: ii, 0.3

The correct answer is ii. If ALQ is missing completely at random, the missingness of ALQ must be independent of the value of ALQ itself. That means, whether ALQ is missing does not depend on whether ALQ is (would have been) Yes or No. Therefore, the probability for missingness given the variable is Yes must be the same as it is for given it is No. As the probability of ALQ missing given ALQ=Yes has 0.3, so must e the probability of missing given ALQ=No.

1b: Correct answer: ii, P(ALQ missing) is independent of Yes No value of ALQ after adjusting for gender

The correct answer is ii. Missing at random means that the missingness is allowed to depend on other variables (in our case gender). However, if we stratify on gender, the probability of missingness then must be independent of ALQ, given gender.

1c: Correct answer: iii, We do not have enough information

The correct answer is iii. We know that $P(ALQ = missing \mid men) = 0.1$. Further we only have one variable which we condition on, with only two possible events (which are exclusive). The law of total probability gives us $P(A) = \sum_n P(A \mid B_n) * P(B_n)$.

Here, it corresponds to $P(ALQ = missing) = P(ALQ = missing \mid Male) * P(Male) + P(ALQ = missing \mid Female) * P(Female)$. In order to calculate $P(ALQ = missing \mid Female)$ we must know the proportions of male to females in our dataset, and the total probability of missing values. However, these information is not given. So we do not have enough information to answer the question (however, the information we need we can find in the data).

Exercise 2: Correct answer Biggest: N' = 90 , smallest: N' = 0

We have $N = 100$ and $x_1, \ldots, x_{10}$ variables. Each of the ten variables x_i contains $100 * 0.1 = 10$ missing values. Under a complete case analysis, we only consider observations which have data in all of the ten variables. The largest sample we can get is if all ten variables have their missing values in the same observation. We know that each variable has 10 missing values, so if all variables have their missing values at same observations, there are 10 observations with missing data (in this case, they would not have any data). Therefore, the largest possible subsample is $100 - 10 = 90$.

Regarding the smallest possible subset, all variables have their missing values at different observations. For an observation to not be included in the complete case analysis, it is already enough if only one variable has a missing value. For simplicity, we assume the missingness in the variables is ordered, e.g. variable x_1 has missing values for observations 1 to 10, variable x_2 has missing values for observations 11 to 20, and so forth. We can then see that the the smallest possible subset is actually 0, that means if we are unlucky, the complete case analysis would remove all our data!

Information to Git-repository

The git-repository for Exercise 3,4 can be found at the following URL:

https://github.com/miribella/MF_missing_data_assignment1.git