LECTURE 19: 13.5 IMPUTATION METHODS

Unreasonable but Commonly-Used Imputation Methods:

1. Hot-Deck Imputation – data are ordered in some way and a missing value is substituted by an observed value of the same variable in the same dataset.

(a) Sequential Hot-Deck Imputation – the missing value is substituted by the previous observed value of the same variable.

Example. In our example, the missing value will be imputed by the value 98.

Individual	HW1	HW2	HW3	EXAM1	EXAM2	GRADE
1	100	90	100	100	100	A
2	94	95	97	97	94	A
3	100	85	**98**	98	95	A
4	95	83	.	97	100	A
5	94	84	94	97	95	A
6	91	85	88	91	89	B
7	97	85	84	98	77	B
8	86	72	82	94	94	B
9	86	77	84	95	89	B
10	85	77	86	88	96	B

(b) Random Hot-Deck Imputation – the missing value is substituted by a randomly chosen observed value of the same variable.

Example In our example, the missing value will be imputed by a randomly chosen value 82. It may be wiser to chose at random a value from among the non-missing values only for A students. Then the missing value will be imputed by, say, 97.

Individual	HW1	HW2	HW3	EXAM1	EXAM2	GRADE
1	100	90	100	100	100	A
2	94	95	97	97	94	A
3	100	85	98	98	95	A
4	95	83	.	97	100	A
5	94	84	94	97	95	A
6	91	85	88	91	89	B
7	97	85	84	98	77	B
8	86	72	**82**	94	94	B
9	86	77	84	95	89	B
10	85	77	86	88	96	B

Hot-deck imputation is widely used by the U.S. Census Bureau.

2. Cold Deck Imputation – imputed values are from a previous survey of the same or similar population.

Example. The instructor taught STAT 108 the previous semester. She finds that a person who got very similar scores on the first two homeworks received 93 for homework 3, so she imputes the missing value by 93.

Note on the name origin: The name *hot-deck* is from the days when computer programs were prepared on punched cards. The deck of cards containing the data set being analyzed was warmed by the card reader, so the term *hot deck* was used to refer to imputations made using the same data set. In *cold-deck* imputation, the imputed values are from another data set not the one running through the computer, so the deck is *cold*.

A word of caution: Both hot-deck and cold-deck imputation procedures are unreasonable in the sense that they may result in very messy data set, with pregnant men, and women with prostate cancer.

3. Multiple Imputation – each missing value is imputed some fixed number of times m ($m > 1$), and then each imputed data set is analyzed separately. Typically, the same imputation method is used each time. The different results give a measure of the additional variance due to the imputation.

This method is applicable when a large number of observations are missing.

Example. In our example, let two observations for hw3 be missing.

Individual	HW1	HW2	HW3	EXAM1	EXAM2	GRADE
1	100	90	100	100	100	A
2	94	95	97	97	94	A
3	100	85	98	98	95	A
4	95	83	.	97	100	A
5	94	84	94	97	95	A
6	91	85	88	91	89	B
7	97	85	84	98	77	B
8	86	72	82	94	94	B
9	86	77	.	95	89	B
10	85	77	86	88	96	B

Suppose we use the random hot-deck method to impute both values. For subject 4, the value may be imputed by 100, 97, 98 or 94. For subject 9, the missing value may be imputed by 88, 84, 82, or 86. There is a total of (4)(4)=16 imputed datasets.

For pure illustrative purposes, suppose we would like to estimate the mean score on hw3 in the population. The 16 imputed data sets are

Imputed Data Sets															
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
97	97	97	97	97	97	97	97	97	97	97	97	97	97	97	97
98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98
100	**100**	**100**	**100**	**97**	**97**	**97**	**97**	**98**	**98**	**98**	**98**	**94**	**94**	**94**	**94**
94	94	94	94	94	94	94	94	94	94	94	94	94	94	94	94
88	88	88	88	88	88	88	88	88	88	88	88	88	88	88	88
84	84	84	84	84	84	84	84	84	84	84	84	84	84	84	84
82	82	82	82	82	82	82	82	82	82	82	82	82	82	82	82
88	**84**	**82**	**86**	**88**	**84**	**82**	**86**	**88**	**84**	**82**	**86**	**88**	**84**	**82**	**86**
86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86

Suppose we pick at random three of the 16 data sets, that is, *m=3*. Let the chosen data set be 3, 9, and 14.

	3	9	14
	100	100	100
	97	97	97
	98	98	98
	100	**98**	**94**
	94	94	94
	88	88	88
	84	84	84
	82	82	82
	82	**88**	**84**
	86	86	86
Mean	91.1	91.5	90.7
SE	7.4603	6.6207	6.6341

In every imputed data set, the mean $\bar{x}_i, i = 1,..., m$, is different. The overall mean of the *m* realizations of the imputation is

$$\bar{x} = \frac{\bar{x}_1 + ... + \bar{x}_m}{m} = \frac{91.1 + 91.5 + 90.7}{3} = 91.1.$$

The estimated variance <u>within</u> the realizations is computed as

$$s_w^2 = \frac{s_1^2 + ... + s_m^2}{m} = \frac{(7.4603)^2 + (6.6207)^2 + (6.6341)^2}{3} = 47.8337.$$

The estimated variance <u>between</u> the realizations is found according to the formula

$$s_b^2 = \left(1 + \frac{1}{m}\right)\frac{\sum_{i=1}^{m}(\bar{x}_i - \bar{x})^2}{m-1} = \left(1 + \frac{1}{3}\right)\frac{(91.1-91.1)^2 + (91.5-91.1)^2 + (90.7-91.1)^2}{3-1}$$

$$= \left(\frac{4}{3}\right)(0.16) = 0.2133.$$

The overall variance and standard error of the estimated mean is given by

$$Var(\bar{x}) = s_w^2 + s_b^2 = 47.8337 + 0.2133 = 48.0470,$$

$$SE(\bar{x}) = \sqrt{Var(\bar{x})} = \sqrt{48.0470} = 6.9316.$$

Note that in this example the overall estimated standard error (6.9316) is not much different from the estimated standard errors of individual realizations (7.4603, 6.6207, and 6.6341). It means that multiple imputation is not really necessary in this case (the variability due to imputations is very small compared to the variability within the imputed datasets).