4 types of missing data:

Stat 576
5-29-25

1) Structurally missing - missing for a logical reason
   Example: "Age of youngest child", but the respondent has no children
2) Missing completely at random (MCAR) - whether or not the person has missing data is completely unrelated to any other information. Safe to omit
3) Missing at random (MAR) - missing data is systematically related to the observed data Impute
4) Missing not at random (MNAR) - missing data is related to other unobserved data Cannot impute

Imputation Methods

(1) Deductive: use existing info to deduce what the missing value must have been

(2) Cell mean: Say you have several categorical variables:

Gender \ Ethnicity	1	2	3
M	y_1, y_2, y_3	y_4, y_5	y_6, y_7
F	y_8, y_9, y_{10}	y_{11}, y_{12}	y_{13}, y_{14}

Suppose y_{10} is missing

The imputed value is the average of the nonmissing values in that cell

(3) Sequential: Same as (2), but the most recently observed value in that cell is used as the imputed value.

(4) Random: Similar to (2), but you randomly select an observed value in that cell, to use as your imputed value

(5) Nearest neighbor As in cluster analysis, you would create a distance measure for each pair of observations. For the observation with the missing value, copy that value from its nearest neighbor

(6) Regression Predict the missing value from the other variables, using multiple regression or logistic regression

For methods (2) thru (5), you could add a stochastic adjustment to the imputed value

For example, use Excel to generate a $U(0,1)$ observation. Take $\Phi^{-1}(u)$ to get a standard normal observation

Φ^{-1} = inverse of cumulative standard normal distribution

Multiply that by the desired σ to get an observation from $N(0, \sigma^2)$

Example of the regression method, using the judging grid on the next page:

Competitors	Judges A	B	C	Avg before imp	Rank before imp	imp. avg.
1	8	7.2	(6.95)	7.6	(3)	7.383 (3)
2	7.1	(7.275)	6.8	6.95	(4)	7.058 (4)
3	(9.075)	9.1	8.4	8.75	(2)	8.858 (1)
4	9.2	8.7	(8.3)	8.95	(1)	8.733 (2)
5	7.3	(7.225)	6.5	6.9	(5)	7.008 (5)
6	(6.625)	6.6	6.0	6.3	(6)	6.408 (6)
	7.9	7.9	6.925	7.575		

Model $y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}$

Using L.S., we can show that

$$\hat{\mu} = \bar{y}_{..}$$

$$\hat{\tau}_i = \bar{y}_{i.} - \bar{y}_{..}$$

$$\hat{\beta}_j = \bar{y}_{.j} - \bar{y}_{..}$$

$$\hat{y}_{ij} = \bar{y}_{..} + (\bar{y}_{i.} - \bar{y}_{..}) + (\bar{y}_{.j} - \bar{y}_{..})$$

$$= \bar{y}_{i.} + \bar{y}_{.j} - \bar{y}_{..}$$

Survey Design

-1 Decide what type of information is needed. If a hypothesis test is to be done, state the null & alternative hypotheses in advance

-2 Test questions

Either use questions from an established survey, or put together an expert panel to evaluate your questions

-3 Make the questions simple & clear

- 4 Make the questions as specific as possible, avoiding generalities

- 5 When reporting the answers, let the reader see the exact wording of the question

- 6 Avoid leading questions

- 7 Consider the number of choices in each answer

Likert scale: 5 to 7 ranked choices

1 2 3 4 5

- 8 Pay attention to question order + order of answers