Key points Identifying and Managing Missing Data and Outliers in Clinical and Health Sciences Research Hybrid Seminar

Summarized by Nur Aazifah by Ilham, CRU HSAAS
Reviewed by by Prof Karuthan , UCSI

What is missing data?

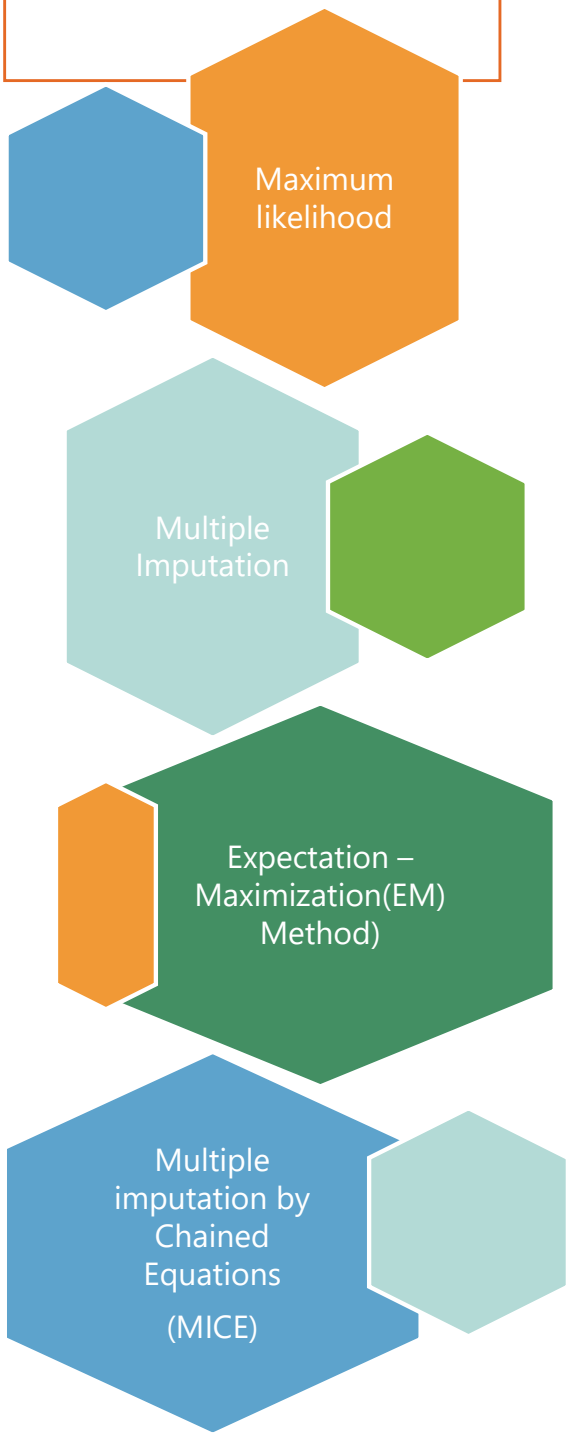
Missing data, or missing values, occur when we do not have data for certain variables or participants. It can be due to incomplete data entry, equipment malfunctions, lost files, etc. Missing data are errors because your data does not represent the true values of what you set out to measure. The reason for the missing data is important to consider because it helps you determine the type of missing data and what you need to do about it.

What are the types of missing data? There are three types which are summarized in the table below.

Missing Completely At Random (MCAR)	Missing at Random (MAR)	Missing not at Random (MNAR)
There is **no relationship** between the **missingness of the data and any values, observed or missing**. Those missing data points are a **random subset of the data**. There is nothing systematic going on that makes some data more likely to be missing than others $P(M \mid X_{miss}, X_{obs}) = P(M)$ M = missing indicator (1=missing, 0=non-missing) X_{miss} = missing value X_{obs} = observed value Example: The machine was not working, the lab sample was not placed in cold storage Implication: Dropping missing values will reduce power and precision	There is a **systematic relationship between the propensity of missing values and the observed data, but not the missing data**. Whether an observation is missing has nothing to do with the missing values, but it does have to do with the values of an individual's observed variables. $P(M \mid X_{miss}, X_{obs}) = P(M \mid X_{obs})$ Examples: 1) If men are more likely to tell you their weight than women, weight is MAR 2) Older people are more likely to answer a question on income than the younger ones	There **is a relationship between the propensity of a value to be missing and its values.** $P(M \mid X_{miss}, X_{obs}) = P(M \mid X_{miss}, X_{obs})$ It is called "**non-ignorable**" because the missing data mechanism itself must be modelled. There are no easy ways of dealing with this type of missing data. **Data collection process needs to be checked when missing not at random**. Examples: 1) People with the lowest education are missing on education. 2) The sickest people are most likely to drop out of the study.

How to handle missing data?

There are many way to handle missing data such **case deletion**, **single imputation** and more complex method of imputation such as :



Listwise deletion

The cases with a missing value on at least one of the variables are deleted. This restricts analyses to individuals with full observed data. This is not good as this assumes missing is MCAR. Hence it leads to biased and under-powered results

Pairwise deletion

The statistical procedure uses case that contain some missing data. The procedure cannot include a particular variable when it has a missing value, but it can still use the case when analysing other variables with non-missing values It allows more data to be analysed. However, there will be an issue when correlations are computed using different subset of cases.

Single Imputation

1. Using sample **mean:** the missing value as the average of the observed values.
2. **Regression prediction**: the missing value is imputed based on a trend value
3. "Hot-Deck": find **cases with the same** observed values on other variables and then choose one randomly
4. Predictive matching: this combines methods (2) and (3)
5. Use a dummy variable, with values of 0 and 1 for missing and non-missing and use this variable in the model.
6. **Intention to treat (ITT)**: last observation carried forward
7. **Modified intention to treat (mITT)**: only consider cases with at least two observations

Note that the goal of missing value analysis is not to get correct predictors of missing values, but to obtain accurate parameter estimates.

Maximum Likelihood

- Maximum likelihood (ML) is a statistical approach used to estimate the values of model parameters that make the observed data most probable under the chosen model. These estimated values are called ML estimates and are used to make inferences about the population from which the data were collected.
- ML estimates are obtained by maximizing the likelihood function, of the model parameters that measures how well the model fits the observed data.
- The likelihood function is a probability density function that describes the probability of observing the data given the model parameters.
- This method uses observed values but considers the missing values as well.
- Cases are usually weighted by the inverse probability of response.
- ML approach is often used to deal with attrition.

Multiple Imputation

- Missing values are filled multiple times by creating multiple “complete” data sets.
- Analyses are done separately on each data set and then the results are combined across the data sets
- It takes into account the uncertainty in the imputations.
- Total variance as a function of the within-imputed variance and between-imputed variance are used
- Once the missing values are imputed, the same data can be used for many analyses

EM-Method

- Each iteration consists of an E step and an M step.
- The E step finds the conditional expectation of the "missing" data, given the observed values and current estimates of the parameters.
- These expectations are then substituted for the "missing" data.
- In the M step, maximum likelihood estimates of the parameters are computed as though the missing data had been filled in.
- "Missing" is enclosed in quotation marks because the missing values are not directly filled in, but the functions of them are used in the log-likelihood.

MICE

- a.k.a “Fully conditional specification” or “Sequential regression multiple imputation”
- It fits model of each variable, conditional on all other variables
- Model used depends on type of variable (continuous/binary/ ordinal)

RECRUS Res.Newsl.Sept 2023;Vol. 3, Issues 24:Page 781.CSTR:https://cstr.cn/33294.43.032423.RECRUS.9

OUTLIERS

WHAT ARE OUTLIERS?

- In data analytics, outliers are values within a dataset that vary greatly from the others —they are either much larger, or much smaller.
- Outliers may indicate variabilities in a measurement, experimental errors, or a novelty.
- Outlier can mislead the regression result as it can pull the regression line towards itself.
- **In data analysis, outliers can cause anomalies in the results obtained.**
- **This means that they require some special attention and, in some cases, will need to be removed to analyze data effectively.**
- **There are two main reasons why giving outliers special attention is a necessary in data analytics process:**
    - Outliers may have a negative effect on the result of an analysis
    - Outliers—or their behavior—may be the information that a data analyst requires from the analysis

How do outliers end up in datasets?

1) Human error while manually entering data, such as a typo.
2) Intentional errors, such as dummy outliers included in a dataset to test detection methods.
3)Sampling errors that arise from extracting or mixing data from inaccurate sources.
4)Data processing errors that arise from data manipulation, or unintended mutations.
5)Measurement errors because of instrumental error.
6)Experimental errors, from the data extraction process/experiment planning/execution.
7)Natural outliers which occur “naturally” in the dataset, as opposed to being the result of an error otherwise listed. These naturally occurring errors are known as novelties.

What is the consideration before you remove the outliers?

It may seem natural to want to remove outliers as part of the data cleaning process. But sometimes it’s best—even necessary—to keep outliers in your dataset. Removing outliers solely due to their place in the extremes of your dataset may create inconsistencies in your results. These inconsistencies may lead to reduced statistical significance in an analysis.

WHAT ARE THE TYPE OF OUTLIER?

In general, they can be divided into two types:

•Univariate outlier

- ❑ It is an extreme value that relates to just one variable.
- ❑ Example : Sultan Kosen is currently the tallest person alive (2.51m).
- ❑ This case would be considered a univariate outlier as it's an extreme case of one factor : height.

• Multivariate outlier

- ❑ It is a combination of unusual values for at least 2 variables.
- ❑ Example: in a group of adults, one person is 2 m tall and weighs 100kg
- ❑ **If we consider his height alone, he may be in the 'usual' range.** And if we consider his weight alone**, he may be in the 'usual' range too.**
- ❑ However, when you consider these two observations in conjunction, you have an adult 2 m tall and weighs 100kg being **a 'surprising'** combination. That's a **multivariate outlier!**

How can you identify outliers?

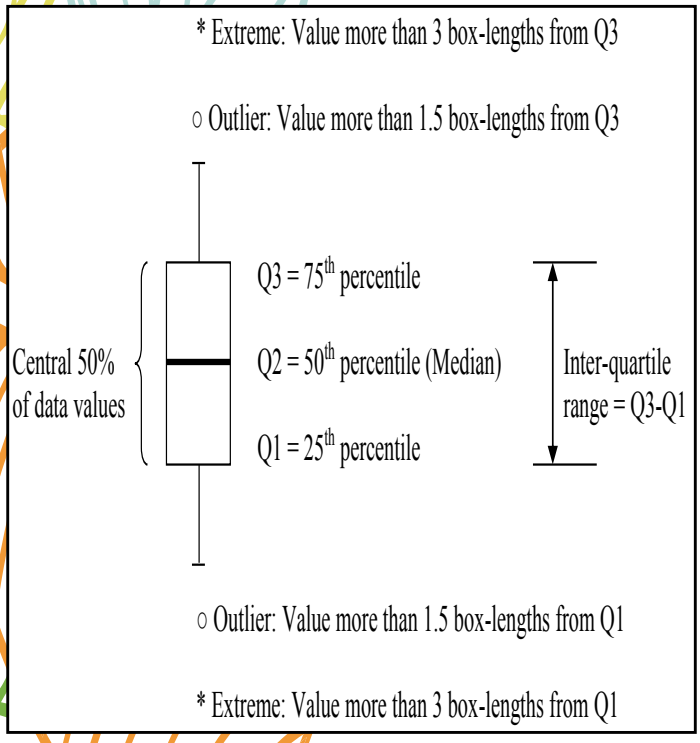
With small datasets, it can be easy to spot outliers manually

For example, for 28, 26, 21, 24, 78, you can see that 78 is the outlier

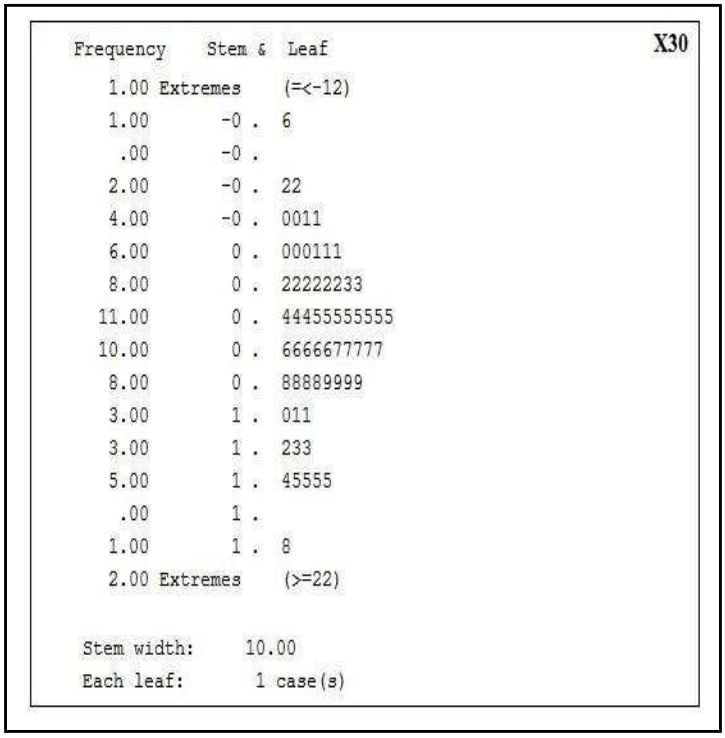
But when it comes to large datasets or other tools are required.

For univariate outlier

Box plot



Stem & Leaf



Frequency	Stem &	Leaf	X30
1.00	Extremes	(=<-12)	
1.00	-0 .	6	
.00	-0 .		
2.00	-0 .	22	
4.00	-0 .	0011	
6.00	0 .	000111	
8.00	0 .	22222233	
11.00	0 .	44455555555	
10.00	0 .	6666677777	
8.00	0 .	88889999	
3.00	1 .	011	
3.00	1 .	233	
5.00	1 .	45555	
.00	1 .		
1.00	1 .	8	
2.00	Extremes	(>=22)	

Stem width: 10.00
Each leaf: 1 case(s)

How about multivariate outlier?

There are many method to detect multivariate outlier such as regression method, distance-bases, density based and clustering bases. Here we will focus on distance-base.

Mahalanobis:

- Measure of how much a case's values on the IVs differ from the average of all other cases.
- This distance has a chi-square distribution with *p* (#IVs) degree of freedom.
- Threshold values are based on a 0.001 level of significance.

Cook's:

- Measure of how much the residuals of all cases would change if a particular case were excluded from the calculation of the regression coefficient.
- A large Cook's D indicates that excluding a case from computation of the regression statistics changes the coefficients substantially.

Leverage values:

- Measures the influence of a case on the fit of the regression.
- The centered leverage ranges from 0 to (n-1)/n, where n is the sample size.
- A value of 0 indicates that the case has no influence at all on the fit.
- The larger the leverage value the larger is the influence of the case on the fit.

What is robust method to handle outlier?

1) Winsorization
2) Robust Regression

Winsorization is the transformation of statistics by limiting extreme values in the statistical data to reduce the effect of possibly spurious outliers.

- A typical strategy is to set all outliers to a specified percentile of the data;
- Example, a 90% winsorization would see all data below the 5th percentile set to the 5th percentile, and data above the 95th percentile set to the 95th percentile.
- Winsorized estimators are usually more robust to outliers than their more standard forms, although there are alternatives, such as trimming, that will achieve a similar effect.

Robust regression seeks to overcome some limitation of traditional regression analysis. It reduce the impact of outlier, violation of distribution assumption and heterogeneity in variance.

Regression with outliers vs without outliers vs robust regression

With outlier n= 108

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	120.560	2.734		44.091	<.001
	Tryglyceride	2.915	.833	.322	3.501	<.001

a. Dependent Variable: Systolic BP

Without outlier n= 105

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	116.418	3.009		38.695	<.001
	Tryglyceride	4.724	1.016	.417	4.652	<.001

a. Dependent Variable: Systolic BP

$$Change\ in\ slope = 4.724 - 2.915 = 1.809$$

With Robust regression

Coefficients

	Value	Std. Error	t value
(Intercept)	119.400	2.824	42.274
TG	3.063	.860	3.561

rlm(formula = SBP ~ TG, data = dta, na.action = na.exclude, method = "MM", model = FALSE)
Residual standard error: 13.66880
Degrees of freedom: 106

$$Change\ in\ slope = 3.063 - 2.915 = 0.148$$

Recording of the hybrid webinar is available. If you are interested to watch the interesting session including SPSS demonstration by Prof Karuthan please contact cru at cru_hsass@upm.edu.my.