# Flavors of missingness

If the data are not MCAR, listwise deletion can severely bias estimates of means, regression coefficients and correlations.

## MCAR - Missing Completely at Random

*The probability of being missing is the same for all cases.*

For example, a random sample of a population, where each member has the same chance of being included in the sample. The (unobserved) data of members in the population that were not included in the sample are MCAR.

In this case we can:

* Delete observations if you are losing less than 5% of the data.
* Impute your data

A picture containing graphical user interface

Description automatically generated

## MAR - Missing At Random

Graphical user interface, text, application

Description automatically generated

In this case we can:

* Careful imputing the data
* Don’t delete observations with missing values.

Then we can apply a statistical test

Text

Description automatically generated with medium confidence

Text

Description automatically generated

## MNAR - Missing Not At Random.

It means that the probability of being missing varies for reasons that are unknown to us.

Graphical user interface, text, application, email

Description automatically generated

Strategies to handle MNAR are:

* To find more data about the causes for the missingness
* To perform what-if analyses to see how sensitive the results are under various scenarios.

# Missing values

A picture containing graphical user interface

Description automatically generated

Text

Description automatically generated with medium confidence

## Sume R rules

|  |  |
| --- | --- |
| Graphical user interface, text, application, chat or text message  Description automatically generated | Graphical user interface, text, application, chat or text message  Description automatically generated |

## Vectorized functions

These functions can also work with data frames.

Graphical user interface, application

Description automatically generated

## Summary functions

|  |  |
| --- | --- |
| Shows the missing in each column | Shows the missing in each row. |
| Text  Description automatically generated with medium confidence | Table  Description automatically generated with low confidence |

We can also to add column using group\_by the dplyr.

Text

Description automatically generated

|  |  |
| --- | --- |
| Classify each column according to the number of missing values | Classify each row according to the number of missing values |
| Text  Description automatically generated with medium confidence | |

Calculates the number of missing values in a specified variable for a repeating span. This is really useful in time series data, to look for weekly (7 day) patterns of missingness. It works with the group\_by operator from dplyr.

Text

Description automatically generated with medium confidence

Calculates the number of "runs" or "streaks" of missingness. This is useful to find unusual patterns of missingness, for example, you might find a repeating pattern of 5 complete and 5 missing. It works with the group\_by operator from dplyr.

Text

Description automatically generated

## Visualize missing values

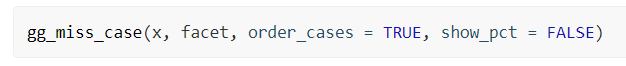
Graphical user interface, application

Description automatically generated

Chart, waterfall chart

Description automatically generated

This isn’t the real number of the variables.



Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

An upset plot of the airquality dataset shows there are only missing values in Ozone and Solar-dot-R, with 35 in only Ozone, 5 in Solar-dot-R, and in both Ozone and Solar-dot-R, there are 2 missing cases.

Chart

Description automatically generated with low confidence

The plot shows a grid that presents all combinations of missing (red) and observed (blue) values present in different variables. The bars to the right of the grid denote the percentage of the observations with the corresponding pattern, while the bars on top show the missing percentage for each variable. From the bottom row, we see that for roughly 84% of the observations there are no missing data in any variable. The second row from the bottom tells us that almost 9% of the observations have a missing value only for total cholesterol.

Chart

Description automatically generated

This doesn’t faces

Chart, histogram

Description automatically generated

Soporta facet y funciona practicamente como un histograma de missing

Chart, bar chart

Description automatically generated

We can make a t-test visually

Chart

Description automatically generated

Let’s make an ANOVA in a plot

Chart

Description automatically generated

## Searching for and replacing missing values

Searching

Text

Description automatically generated

Replacing

Text

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

A picture containing logo

Description automatically generated

## Filling down missing values

Graphical user interface, text, application

Description automatically generated

Table

Description automatically generated with medium confidence

A picture containing graphical user interface

Description automatically generated

## Tools to explore missing data dependence

### Bind shadow

Graphical user interface, website

Description automatically generated

Text

Description automatically generated

### Density plot

Chart

Description automatically generated

### Boxplot

Graphical user interface, chart, application

Description automatically generated

### Scarlett plot

Chart, scatter chart

Description automatically generated

Graphical user interface, chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart

Description automatically generated

### Checking two missing columns at once

Graphical user interface, chart

Description automatically generated

## Imputing the missing data

### Importance

1. What if these are the people whose cholesterol was so high that it maxed out the measuring device, and so it was not recorded? We would certainly not want to ignore these cases. **(It hides importance relationships)**
2. **Which of the two models is better?** We don’t know because the two models were trained on two different data samples (a different number of observations were removed) and **we cannot use the adjusted R-squared** to answer this.

Graphical user interface, text

Description automatically generated

### Bad imputation

#### Lower value

Graphical user interface, text, application

Description automatically generated

To track imputed values

Text

Description automatically generated

To visualize missings for two variables, we need to add a label that identifies whether there is a missing value in a column. The function add\_label\_shadow does this for us. We have now recreated the same figure as geom\_miss\_point!

Chart, scatter chart

Description automatically generated

#### Mean imputation

Graphical user interface, text, application, chat or text message

Description automatically generated

Text

Description automatically generated

Chart, line chart, scatter chart

Description automatically generated

### Good imputation

#### Imputation with lineal models

Graphical user interface, text, application

Description automatically generated

Text

Description automatically generated

Chart, scatter chart

Description automatically generated

##### Comparing imputation models

Text

Description automatically generated

Calendar

Description automatically generated with medium confidence

Graphical user interface, chart, scatter chart

Description automatically generated

Graphical user interface, chart, scatter chart

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Chart, scatter chart

Description automatically generated

### Is the imputation working?

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

# Multiple imputation

## Reasons to use it

* It solves the problem of ***“too small” standard errors,*** as it provides a mechanism for dealing with the inherent uncertainty of the imputations themselves
* It is able to deal with both high-confidence and low-confidence situations equally well.
* It separates the solution of the missing data problem from the solution of the complete-data problem.
* It uses observed data as donors to fill the missing data

Our level of confidence in a particular imputed value is expressed as the variation across the ***m*** completed datasets.

## Practique example

Graphical user interface

Description automatically generated with medium confidence

# Likelihood-based

[The lavaan Project (ugent.be)](https://lavaan.ugent.be/tutorial/index.html)