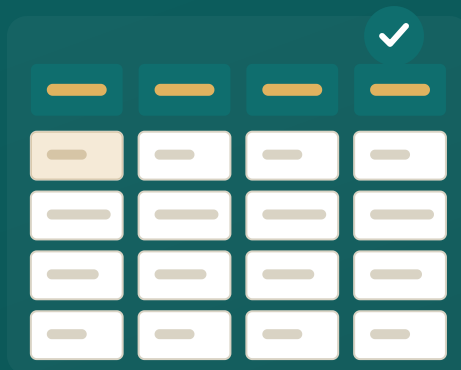


Handling Missing Data Without Wrecking Your Analysis

Why the reason data are missing matters more than how much is missing, which method fits your situation, and the default that quietly biases most dissertations. A working guide from PhD statisticians.



Written by PhD statisticians dissertationstatisticshelp.org

The mechanism behind the gaps decides what you can do

Almost every real dataset has missing values, and how you handle them can change your results entirely. The mistake is to reach for the easiest fix. The right method depends not on how much is missing but on why it is missing, and that question has to be answered before you choose anything.

PRINCIPLE

Diagnose the missingness mechanism first. Data can be missing completely at random, missing at random given other variables you measured, or missing not at random because the missing value itself drives whether it is recorded. Each demands a different response, and assuming the wrong one introduces bias you cannot see in the output.

The three mechanisms

1 Missing completely at random

The chance of a value being missing is unrelated to anything in your study. This is the most forgiving case, but it is also the rarest, and you should not assume it without checking.

2 Missing at random

Missingness depends on other variables you did measure, but not on the missing value itself. Modern methods are built for this case and can give unbiased results when it holds.

3 Missing not at random

The probability of being missing depends on the unobserved value, for example heavier drinkers skipping a question about drinking. This is the hardest case and usually needs sensitivity analysis and honest discussion.

Match the method to the mechanism

The popular shortcuts are popular because they are easy, not because they are sound. Use this to choose deliberately.

COMMON APPROACHES AND WHEN EACH IS DEFENSIBLE

Method	What it does	When it is reasonable
Listwise deletion	Drops any case with a missing value	Only when data are missing completely at random and the loss is small
Mean substitution	Fills gaps with the variable average	Rarely; it shrinks variance and distorts relationships
Multiple imputation	Creates several plausible filled datasets and pools results	The default for data missing at random; widely accepted
Maximum likelihood	Estimates the model using all available data	A strong choice for missing at random, common in structural models
Sensitivity analysis	Tests how conclusions change under different assumptions	Essential when data may be missing not at random

WATCH OUT FOR

Mean substitution and single-value fill-ins feel harmless but artificially reduce variability and weaken correlations, which biases standard errors and can manufacture or hide significance. They are the most common avoidable error in dissertation analysis. Prefer multiple imputation or maximum likelihood unless you have a specific reason not to.

PRINCIPLE

Multiple imputation fails quietly when the imputation model is built wrong. Three rules keep it honest. First, congeniality: the imputation model must be at least as rich as the analysis model, so include every analysis variable plus auxiliary variables that predict missingness or the missing values, and crucially include the outcome, imputing the dependent variable rather than dropping its cases. Omitting a variable from imputation that you later analyse biases its association toward null. Second, pool with Rubin's rules, which combine the estimates and add the between-imputation variance, never average the imputed datasets into one and analyse that. Third, use enough imputations: the old "three to five" is outdated, and a common rule is to set the number of imputations roughly equal to the percentage of cases that are incomplete. Maximum likelihood (full-information) reaches the same target in one step and is asymptotically equivalent, which is why it is preferred inside structural equation models.

Report missingness so a committee trusts your data

Examiners want to see that you understood your missing data, not that you hid it. State the extent, the likely mechanism, and what you did.

Missing data reporting template

Extent:	[how much is missing, by variable]
Pattern:	[is it concentrated in certain cases/variables]
Mechanism:	[your judgement: MCAR / MAR / MNAR, with reasoning]
Method:	[what you used, e.g. multiple imputation, number of imputations]
Software:	[tool and procedure]
Sensitivity:	[if MNAR is plausible, how you tested robustness]
In words:	[a short statement that conclusions held, or how they changed, under your handling of missing data]

PRINCIPLE

Decide your missing data strategy before you analyse, and write it into your plan. Choosing a method after seeing which one gives the cleanest result is a form of fishing that a careful examiner will recognise. A pre-specified, mechanism-appropriate approach is both better science and a far easier defense.

The missing data checklist

If every box is ticked, your handling of missing data will strengthen your analysis rather than quietly undermine it.

- ☒ You report how much data are missing, by variable.
- ☒ You form a reasoned judgement about the missingness mechanism.
- ☒ You avoid mean substitution and single-value fill-ins as a default.
- ☒ You use multiple imputation or maximum likelihood when data are missing at random.
- ☒ Your imputation model includes the outcome and auxiliary variables, with enough imputations, and pools by Rubin's rules.
- ☒ You run a sensitivity analysis where data may be missing not at random.
- ☒ You decided your method in advance, not after seeing results.
- ☒ You state the extent, mechanism, and method in your write-up.

WHEN THE DATA ARE NOT CLEAN

Have your missing data handled in a way that holds up

If you want your missingness diagnosed, the right method applied, and the whole approach reported so a committee accepts it, our PhD statisticians work through it with you. You stay the author and understand every decision made about your data.

Get a quote for your analysis

dissertationstatistics.org . Written by PhD statisticians . We work in writing, on your schedule.