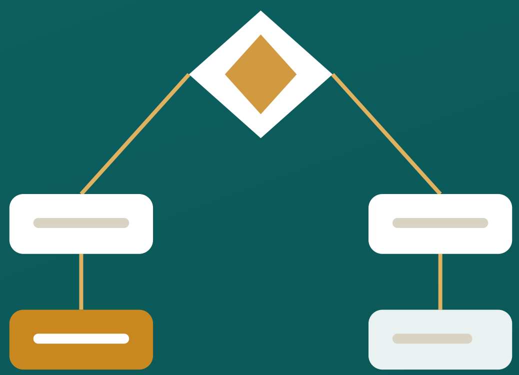
Dissertation Statistics

DISSERTATION STATISTICS FIELD GUIDE

Check Your Statistical Assumptions

What each common test assumes about your data, how to check it, and what to do when an assumption is violated, so your analysis holds up to scrutiny. A working guide from PhD statisticians.



Written by PhD statisticians dissertationstatisticshelp.org

START HERE

A test is only valid if your data meet what it assumes

Every common statistical test rests on assumptions about the data it is given. When those hold, the result can be trusted. When they are violated and ignored, the probability value can be misleading, and an examiner who spots it will rightly question the whole analysis. Checking assumptions is not a formality; it is what makes the answer real.

> **PRINCIPLE**
>
> Check assumptions before you interpret a result, not after. Run the diagnostic, decide whether the assumption holds, and only then read the test. If an assumption fails, you have clear, defensible options: transform the data, switch to a test that does not require it, or use a robust method. What you cannot do is ignore it.

The assumptions behind most analyses

1. **Normality**
   Many tests assume the outcome, or the residuals from a model, are roughly normally distributed. This matters most in small samples; in large ones, tests are fairly forgiving of mild departures.

2. **Homogeneity of variance**
   Tests that compare groups often assume the groups have similar spread. When variances differ sharply, the standard test can mislead, and an adjusted version is needed.

3. **Independence of observations**
   Most basic tests assume each data point is independent of the others. Repeated measures on the same person, or clustering within sites, breaks this and calls for a different model.

4. **Linearity and no severe multicollinearity**
   Regression assumes the relationships modelled are linear and that predictors are not so correlated with each other that their effects cannot be separated.

THE DIAGNOSTICS

How to check each assumption, and what failure looks like

For each assumption, there is a way to look at it and a sign that it has failed. Use the visual check and the formal test together; neither alone tells the whole story.

ASSUMPTION, HOW TO CHECK IT, AND THE WARNING SIGN

Assumption	How to check	Sign it is violated
Normality	Histogram, Q-Q plot, and a formal test such as Shapiro-Wilk	Clear skew or heavy tails, especially with a small sample
Homogeneity of variance	Levene's test and grouped box plots	Groups with markedly different spread
Independence	Study design, and residual plots over time or cluster	Repeated measures or clustering not modelled
Linearity	Scatter plots of outcome against each predictor	Curved rather than straight patterns
Multicollinearity	Variance inflation factor for each predictor	High inflation values, often above ten
Outliers and influence	Standardised residuals and influence statistics	A few points that dominate the result

WATCH OUT FOR

A formal normality test is very sensitive in large samples, flagging tiny, harmless departures, and weak in small ones, where it matters most. Read the Q-Q plot alongside it rather than trusting the probability value alone. The shape of the data tells you more than a single yes-or-no test.

PROPORTION THE RESPONSE

A violation matters in proportion to your sample size and the statistic

The defensible position is not "all assumptions must pass" but "this violation, at this sample size, threatens this inference". An examiner respects judgement here far more than reflexive transformation.

HOW SERIOUSLY TO TAKE EACH VIOLATION, WITH THE WORKING THRESHOLDS

Violation	When it is largely harmless	When it is serious, and the number to cite
Non-normality	Large n; the central limit theorem protects means and coefficients	Small n, or inference on individual cases; Shapiro-Wilk is over-powered at large n, so weight the Q-Q plot
Unequal variances	Equal group sizes absorb moderate differences	Unequal n with unequal variance; default to Welch's t and Welch ANOVA rather than testing first
Non-independence	Genuinely independent sampling	Any clustering or repeated measures; the design effect inflates real error, so model it (mixed model)
Multicollinearity	Predictors you do not interpret individually	VIF above ~5 warrants concern, above 10 is serious; near-zero tolerance says the same
Sphericity (repeated measures)	Mauchly's test non-significant	Mauchly significant: use Greenhouse-Geisser, or Huynh-Feldt when its epsilon exceeds ~0.75
Influential cases	High leverage but low influence	Cook's distance above 1 (or above 4/n as a screen) with large standardised residuals

PRINCIPLE

Distinguish an outlier (extreme on the outcome) from a high-leverage point (extreme on the predictors) from an influential point (one that actually moves the coefficients). Only the last forces a decision, and Cook's distance, not the raw residual, is what identifies it. Before transforming, ask whether a robust method answers the objection more cleanly: Welch's correction for variance, heteroscedasticity-consistent (HC3) standard errors or a percentile bootstrap for regression inference, Yuen's trimmed-means t-test for skew with outliers, or MM-estimation for influential cases. These hold the original scale, which keeps the result interpretable.

THE FIXES

What to do when an assumption fails

A violated assumption is a fork in the road, not a dead end. Each one has accepted remedies. Choose deliberately and report what you did and why.

WHEN AN ASSUMPTION FAILS, YOUR OPTIONS

The problem	Defensible responses
Outcome is not normal	Transform the variable, or use a non-parametric equivalent
Variances are unequal	Use a version that does not assume equal variance, such as Welch's
Observations are not independent	Use a mixed model or a repeated-measures method
Relationship is not linear	Add a polynomial term, transform, or model the non-linearity
Predictors are collinear	Drop or combine redundant predictors, or use a regularised model
Influential outliers	Investigate the cases, report with and without them, and justify

PRINCIPLE

Whatever you decide, report it. State which assumptions you checked, how, what you found, and what you did about any violation. A results section that documents its assumption checks is far harder to challenge than one that quietly hopes nobody asks.

BEFORE YOU TRUST THE RESULT

The assumptions checklist

Run your analysis through this before interpreting it. If every box is ticked, the test you ran is the right one for the data you have.

- [x] You know which assumptions your chosen test requires.
- [x] You checked normality with both a plot and, where useful, a formal test.
- [x] You checked that groups being compared have comparable variance.
- [x] You confirmed observations are independent, or used a model that allows for clustering.
- [x] For regression, you checked linearity and multicollinearity.
- [x] You inspected outliers and influential points rather than deleting them blindly.
- [x] For any violation, you chose a defensible remedy instead of ignoring it.
- [x] You documented every assumption check and decision in the write-up.

WHEN THE ANALYSIS HAS TO SURVIVE SCRUTINY

Have your assumptions checked and your analysis defended

If you want your data checked against the assumptions of every test, the right method chosen when one fails, and the whole thing documented so it holds up in your defense, our PhD statisticians work through it with you. You stay the author and understand each decision.

Get a quote for your statistics

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