SCHOOL OF DATA
by Springboard

Data Science Career Track
The Art of Statistics, Chapters 12 and 13
Take-Away Notes

The **reproducibility crisis** (p.342) is the problem that many scientific papers are not replicated by peers, perhaps because they cannot be. The crisis has been attributed in part to poor statistical practice.

Statistics can be done poorly at **every stage of the PPDAC cycle** (p.344).

- P: Problem-identification can go wrong by choosing a problem that simply can't be solved with the information available.
- P: Planning can go wrong by, for example, choosing just a convenient sample.
- D: Data collection can go wrong by, for example, people dropping out of the study, and recruitment being slower than expected.
- A: Analysis can go wrong by simply making errors in code or in spreadsheets, or due to more blameworthy actions, such as interpreting 'non-significant' as meaning 'no effect'
- C: Conclusions can go wrong, perhaps most disastrously when multiple statistical tests are carried out, and then only the most significant one reported.

The deliberate fabrication of data is relatively rare. But **questionable research practices** (p.350) include **P-hacking** (p.351), which can be done through decisions regarding:

- the design of the experiment
- when to stop collecting data
- what data to exclude
- what factors to adjust for

- what groups to emphasize
- what outcome measures to focus on
- how to split continuous variables into groups
- how to handle missing data, and so on.

To tackle this, Professor Spiegelhalter offers 10 useful rules for **effective statistical practice** (p.379). These are:

1. Statistical methods should enable data to answer scientific questions
2. Signals always come with noise
3. Plan ahead, really ahead
4. Worry about data quality
5. Statistical analysis is more than a set of computations
6. Keep it simple
7. Provide assessments of variability
8. Check your assumptions
9. When possible, replicate!
10. Make your analysis reproducible.