Ten common statistical mistakes to watch out for when writing or reviewing a manuscript

Authors: Tamar R Makin and Jean-Jaques Orban De Xivry

2024-10-07

Abstract

In this presentation I will identify 10 of the most common mistakes found within statistical papers. For each mistake, I will state what the problem is, how we can detect it, and how we can avoid making these mistakes.

Goal: After this presentation you should be able to peer review papers efficiently and effectively.

About the Authors

Tamar Makin- Israeli neuroscietnsit who studied at Hebrew University of Jerusalem. She is currently a professor and researcher at the University of Cambridge.

Jean-Jacques Orban de Xivry- Engineer and neuroscientist who studied at Université catholique de Louvain (Belgium). He is currently a professor and researcher at KU Leuven.

1. Absence of an Adequate Control Condition/Group

Problem: Studies that look at the effect of an experimental manipulation on a variable over time should be compared to a control over time

- ex. An experimental group that runs 3 miles a day on muscle mass of calves' vs a group that just walks 3 miles a day

Detection: Noticing that the control group does not account for key features of the task that are inherent to the manipulation

- ex. Having the experimental group be people aged 20-30 but the control people ages 40-50 (we can expect different muscle mass)

1. Absence of an Adequate Control Condition/Group

Solution: If you can't separate the effect of time from the effect of intervention, then the conclusions for the intervention should be presented as "tentative"

2. Interpreting comparisons between two effects without directly comparing them

Problem: When the treatment in the experimental group shows a strong effect compared to the control group, we automatically assume that the treatment is working

- ex. The correlation between group a=0.4 and group b=0.45 while the p-value for a is <0.05 and for group b is 0.08. Hence group A is statistically significant while B is not, but the correlation is moderate for both groups

2. Interpreting comparisons between two effects without directly comparing them

Detection: Can occur when a researcher “makes an inference without performing the necessary statistical analysis” (4 Makin and Orban)

Solution: Compare the groups directly.

- ANOVA can be useful for group comparisons
- Correlations in 2 groups can be compared to Monte Carlo simulations

3. Inflating the units of analysis

Problem: Experimental units get mixed up with the number of observations made

Definition of Experimental Unit is the number of independent values that are free to vary (number of subjects tested rather than number of observations made within each subject)

- ex. 1 person used for 3 observations (heart rate, height, and weight) only counts as 1 experimental unit

3. Inflating the units of analysis

Effect of problem: Results in a super high number of experimental units ->df increases ->threshold against statistical significance is judged decreases->easier to observe significant results-> hence higher likelihood of false positive

Detection: Be careful in group analysis and make sure “the unit of analysis should reflect the variance across subjects, not within subjects.” (5 Makin and Orban)

3. Inflating the units of Analysis

Solutions:

- Use a mixed effects linear model which allows us to put all data into a model and not violate the assumption of independence (can be complicated and misused)
- A simple solution is to calculate r (correlation) for each observation separately
- Average values across observations
- Calculate pre/post correlations and average the r values

4. Spurious (illegitimate) Correlations

Problem: Not following general assumptions (normality, independence, random selection, no outliers) leads to illegitimate correlations.

- Generally, arises from outliers inflating the correlation coefficient

4. Spurious (illegitimate) Correlations

Detection: Assess data plots for outliers and see if correlation coefficient makes sense.

Solutions: Make sure you satisfy the assumptions (normality, independence, random selection, no outliers)

5. Use of Small Samples

Problem: Small sample sizes only detects large effects (overestimates actual effect –> leads to large correlation value)

- Increases the chance of a false positive (Type 1 Error)
- Small sample sizes are not likely to be normally distributed (cannot use normality assumption)

5. Use of Small Samples

Detection: Examine the sample size and judge if it is appropriate; if there is a large effect, it may be considered suspicious depending on the sample size

Solutions: If possible, perform replications of the experiment so see if the outcome is the same. Also, make sure you have a control.

6. Circular Analysis

Problem: Researchers analyze data and pick out certain patters that describe their results; they then use that pattern to conduct further research which may lead to an incorrect finding

- Ex. Neuron firing rate due to manipulation: there were no significant effects initially, but the researchers observed an increased firing rate for some neurons and others less. This led them to split the neurons into a quick and slow response rate group. Any interpretation from this is misleading because maybe the neurons reacted by chance.

6. Circular Analysis

Problem: Dependencies are created between the dependent and independent variable

- Ex. Correlation reported between cell response post manipulation and the difference in cell response for pre and post manipulation. (notice that both are dependent upon the post manipulation measure, hence any analysis done is inaccurate)

6. Circular Analysis

Detection: Decide if the statistical test criteria are in favor of the hypothesis tested.

Solutions:

- Define the analysis criteria independently of the data (i.e. don’t pick the question based on the data)
- Run experiments again to see if same results arise

7. Flexibility of Analysis: P-hacking

Flexibility Definition: use of switched outcome parameters, adding covariates, using undetermined/erratic pre-processing pipeline/subject exclusion

Problem: When using flexibility, it increases the probability of getting a significant p-value.

- **Why?** Normative statistics relies on probabilities and when you run more then one test, there is an increased probability of getting a false positive (8 Makin and Orban)

7. Flexibility of Analysis: P-hacking

Detection: Make sure planned and executed trials were identical. This can be very hard to detect.

Solutions:

- Replicate the experiment to make sure results are similar.
- Check to make sure the experiment was well designed, executed, and analyzed.
- Be clear when you state pre-planned analysis vs the exploratory analysis since exploratory analysis cannot have strong conclusions

8. Failing to Correct for Multiple Comparisons

Problem: When researchers have multiple comparisons, there can be consequences for significant findings (type 1 errors)

- More factors-> greater number of tests/measures performed-> increasing probability of false positives

8. Failing to Correct for Multiple Comparisons

Detection: Look at the number of variables being tested and if only one of the variables are correlated with the dependent variable, the rest are likely to have been included to increase the chance of obtaining a significant result. Hence, we should not be conducting exploratory analysis on large sets of variables

Solution: State multiple comparison procedures and variables and see if signifigant results seem plausible with number of variables.

9. Over-Interpreting Non-signifigant Results

Problem: A non-significant p-value does not distinguish between a lack of effect due to objects absence or if there was an inappropriate experimental design. Hence, the p-value is insufficient evidence for no experimental result.

9. Over-Interpreting Non-signifigant Results

Detection: If a researcher states that a non-signifigant p-value indicates no effect was present

Solution:

- Do not interpret non-significant results, only describe them as non-significant
- Make sure to report magnitude of effect with the p-values and identify if approporiate experimental designs were used

10. Correlation and Causation

Problem: A correlation of 2 variables does not mean that one causes the other. Correlation cannot be used as evidence for cause and effect.

- Ex. Monthly ice cream sales and shark attacks (correlated but not causation)

10. Correlation and Causation

Detection: If a researcher reports an association that is not a direct cause of manipulation

Solutions: Explore relationships with a third variable to prove causation

Conclusion

One statistical mistake can snowball into other mistakes which will undermine the integrity of your research. It is our job as statisticians to look out for these errors and identify them to improve analyses. This way we can ensure our findings (and our peers findings) are accurate and meaningful.

References

Makin, T. R., & Orban de Xivry, J. J. (2019). Ten common statistical mistakes to watch out for when writing or reviewing a manuscript. eLife, 8, e48175. https://doi.org/10.7554/eLife.48175