Data analysis errors – cheat sheet
Day 1

In this cheat sheet, we have summarized the most common mistakes in data analysis. Read the list thoroughly and prepare to be skeptical with the findings presented in the court. After the Plaintiff and Defendant show their results and arguments in each round, it is your job as a Data Expert to uncover possible errors in the analysis. In each round, you must prepare 2-3 comments, either based on this list or it can also be any other idea. Try to be impartial and comment on both sides.

Cognitive biases

1. Confirmation bias: "Confirmation bias is the tendency to seek out or interpret data to confirm beliefs you already hold. (...) Confirmation bias pressures you to ignore the negative signs and focus on the positive."[1]
2. Selection bias: "Selection Bias is when the group chosen to be analyzed is not representative of the population you are trying to draw conclusions about."[2] It can be either due to convenience (easy to reach or measure, already have the data on them) or self-selection (the group of people who opt in to be analyzed can have different characteristics than the whole population).
3. Survivorship bias: "Survivorship bias is the tendency to draw conclusions based on things that have survived some selection process and to ignore things that did not survive."[3]

Analysis errors

4. Improper outlier detection: Outliers can greatly affect statistical analysis and distort the results. Analysts should handle them properly: investigate the cases and remove if needed. For example, imagine that in the dataset the highest wage is doubled: it increases the mean wage, but the median wage is unaffected.
5. Lack of statistical significance: To test an assumption and draw conclusions with statistical validity, we need to conduct hypothesis testing. If the p-value of the test is small enough (say, less than 5%), we can reject the null hypothesis, and infer that the relationship is not risen by chance. Although, if the p-value is high it means the result lacks statistical significance, hence we cannot draw conclusions with high certainty. Small sample size can also lead to insignificant results.
6. Use of the wrong metric: Is it the right metric to answer the question it hand, or just a proxy - a "close enough" measurement. Be aware of the limitations of the data.
7. Use the wrong benchmark for comparison: Finding a right benchmark data is difficult, but essential for drawing valid conclusions. The absence of an adequate control group leads to inaccurate results. Several factors should be considered when selecting an appropriate benchmark, where the control group should differ from the experimental group in only one characteristic which is the variable of interest (x). This enables us to study the effect of one variable at a time. If the control group differs in more than one

[1] David (2021), p.4
[2] David (2021), p.7
[3] David (2021), p.10

characteristic, it is impossible to determine which changes in the outcome (y) are due to the variable of interest (x) as opposed to being due to some other observed (or unobserved) variable.

8. Overall vs. groups: Create overall statistics that describe the whole sample is a good first step, although it can be also misleading. It does not always show you nuanced patterns about the underlying data. Once data is grouped by a relevant characteristic (for example country, gender, age, education), the statistics for each group may vary: different patterns can emerge in different groups, and the relationship in groups can even be the opposite direction as in the overall data (this is called Simpson's paradox).[4]

Interpretation errors

9. Misinterpret statistical concepts: Incorrect interpretation leads to the misunderstanding of the underlying phenomenon. Analysts should use correct wording when talking about statistical concepts such as: biases, mean, distribution, standard error, null hypotheses, t-test, p-value, significance, coefficients, and else.
10. Relative vs. absolute differences: “When choosing between reporting a relative change or absolute change, take a second and think if you are choosing the type of change that best represents what is actually happening or are you selecting the more sensational number. The best practice is to provide both numbers. (...) To get a clearer idea if an absolute or relative change is significant compare it to other changes that are related to it."[5] Is it small or big? Is it a relevant difference?

[4] David (2021), p.19
[5] David (2021), p.32