Data Analysis, Statistics, Machine Learning

Leland Wilkinson

Adjunct Professor
UIC Computer Science
Chief Scientist
H2O.ai

leland.wilkinson@gmail.com

Analyzing

What Statistics is not

- mathematics
- machine learning
- computer science
- probability theory

Statistical reasoning is rational

- Statistics conditions conclusions
- Statistics factors out randomness

Wise words

- David Moore
- Stephen Stigler
- TFSI

Analyzing

David Moore

Look at your data

Recognize the importance of data production

Understand randomness and the role of the long run

Data take priority over any model

Statistics can be viewed as a branch of the liberal arts

Data beat anecdotes

Is this the right question?

Does the answer make sense?

Can you read a graph?

Do you have filters for quantitative nonsense?

Analyzing

- Stephen Stigler
  - Seven Pillars of Statistical Wisdom
    - **Aggregation**
      - You can gain knowledge by discarding information
    - **Information**
      - The amount of information in a data set is often proportional only to the square root of the number of observations, not the number itself (take that, Big Data!)
    - **Likelihood**
      - We calibrate information through the use of probability
    - **Comparisons**
      - Statistical comparisons do not need to be made with respect to an external (gold) standard
    - **Regression**
      - The basis of a paradox and the basis of inference, including Bayesian inference and causal reasoning
    - **Experiments**
      - Randomization plus combination yields the most trustworthy information
    - **Residuals**
      - Search for structure after subtracting structure

Analyzing

Task Force on Statistical Inference (Wilkinson and TFSI, 1999)

Design

Make clear at the outset what type of study you are doing. Do not cloak a study in one guise to try to give it the assumed reputation of another. For studies that have multiple goals, be sure to define and prioritize those goals.

Population

The interpretation of the results of any study depends on the characteristics of the population intended for analysis. Define the population (participants, stimuli, or studies) clearly. If control or comparison groups are part of the design, present how they are defined.

Sample

Describe the sampling procedures and emphasize any inclusion or exclusion criteria. If the sample is stratified (e.g., by site or gender) describe fully the method and rationale. Note the proposed sample size for each subgroup.

Analyzing

Task Force on Statistical Inference

Random assignment

For research involving causal inferences, the assignment of units to levels of the causal variable is critical. Random assignment (not to be confused with random selection) allows for the strongest possible causal inferences free of extraneous assumptions. If random assignment is planned, provide enough information to show that the process for making the actual assignments is random.

Nonrandom assignment

For some research questions, random assignment is not feasible. In such cases, we need to minimize effects of variables that affect the observed relationship between a causal variable and an outcome. Such variables are commonly called confounds or covariates. The researcher needs to attempt to determine the relevant covariates, measure them adequately, and adjust for their effects either by design or by analysis. If the effects of covariates are adjusted by analysis, the strong assumptions that are made must be explicitly stated and, to the extent possible, tested and justified.

Analyzing

Task Force on Statistical Inference

Variables

Explicitly define the variables in the study,show how they are related to the goals of the study, and explain how they are measured. The units of measurement of all variables, causal and outcome, should fit the language you use in the introduction and discussion sections of your report.

Procedure

Describe any anticipated sources of attrition due to noncompliance, dropout, death, or other factors. Indicate how such attrition may affect the generalizability of the results. Clearly describe the conditions under which measurements are taken (e.g., format, time, place, personnel who collected data). Describe the specific methods used to deal with experimenter bias, especially if you collected the data yourself.

Analyzing

Task Force on Statistical Inference

Power and sample size

Provide information on sample size and the process that led to sample size decisions. Document the effect sizes, sampling and measurement assumptions, as well as analytic procedures used in power calculations. Because power computations are most meaningful when done before data are collected and examined, it is important to show how effect-size estimates have been derived from previous research and theory in order to dispel suspicions that they might have been taken from data used in the study or, even worse, constructed to justify a particular sample size. Once the study is analyzed, confidence intervals replace calculated power in describing results.

Analyzing

Task Force on Statistical Inference

Complications

Before presenting results, report complications, protocol violations, and other unanticipated events in data collection. These include missing data, attrition, and nonresponse. Discuss analytic techniques devised to ameliorate these problems. Describe nonrepresentativeness statistically by reporting patterns and distributions of missing data and contaminations. Document how the actual analysis differs from the analysis planned before complications arose. The use of techniques to ensure that the reported results are not produced by anomalies in the data (e.g., outliers, points of high influence, nonrandom missing data, selection bias, attrition problems) should be a standard component of all analyses.

Simple analyses

The enormous variety of modern quantitative methods leaves researchers with the nontrivial task of matching analysis and design to the research question. Although complex designs and state-of-the-art methods are sometimes necessary to address research questions effectively, simpler classical approaches often can provide elegant and sufficient answers to important questions. Do not choose an analytic method to impress your readers or to deflect criticism. If the assumptions and strength of a simpler method are reasonable for your data and research problem, use it. Occam's razor applies to methods as well as to theories.

Analyzing

Task Force on Statistical Inference

Fisher (1935)

"Experimenters should remember that they and their colleagues usually know more about the kind of material they are dealing with than do the authors of text-books written without such personal experience, and that a more complex, or less intelligible, test is not likely to serve their purpose better, in any sense, than those of proved value in their own subject. "

Causality

Inferring causality from nonrandomized designs is a risky enterprise. Researchers using nonrandomized designs have an extra obligation to explain the logic behind covariates included in their designs and to alert the reader to plausible rival hypotheses that might explain their results. Even in randomized experiments, attributing causal effects to any one aspect of the treatment condition requires support from additional experimentation.

Analyzing

Task Force on Statistical Inference

Interpretation

When you interpret effects, think of credibility, generalizability, and robustness. Are the effects credible, given the results of previous studies and theory? Do the features of the design and analysis (e.g., sample quality, similarity of the design to designs of previous studies, similarity of the effects to those in previous studies) suggest the results are generalizable? Are the design and analytic methods robust enough to support strong conclusions?

Conclusions

Speculation may be appropriate, but use it sparingly and explicitly. Note the shortcomings of your study. Remember, however, that acknowledging limitations is for the purpose of qualifying results and avoiding pitfalls in future research. Confession should not have the goal of disarming criticism. Recommendations for future research should be thoughtful and grounded in present and previous findings. Gratuitous suggestions ("further research needs to be done ...") waste space. Do not interpret a single study's results as having importance independent of the effects reported elsewhere in the relevant literature.