Statistical consulting: a crash course

Linda Mhalla[1]

[1]Institute of Mathematics, EPFL

MATH-516 Applied Statistics

Why is this course essential?

- Statistics is an inherently collaborative discipline (Ben-Zvi, 2007)
  $\rightarrow$ To have a real-world impact and maximise the impact of their work, applied statisticians collaborate with domain experts who own the data, originate the problems to be solved, and make decisions
- According to the American Statistical Association (ASA), graduates of statistical science undergraduate programs,

  "Should demonstrate ability to collaborate in teams and . . . be able to communicate complex statistical methods" (ASA, 2014, p.10)

- Collaborative skills are essential for professional statisticians. According to ASA,

  "Professional development is important to statisticians because it helps them advance their careers, remaining competitive and marketable . . . Factors such as communication, leadership, and influence skills . . . are vital to the impact of individual contributions and the visibility of our profession" (ASA, 2012, p.1)

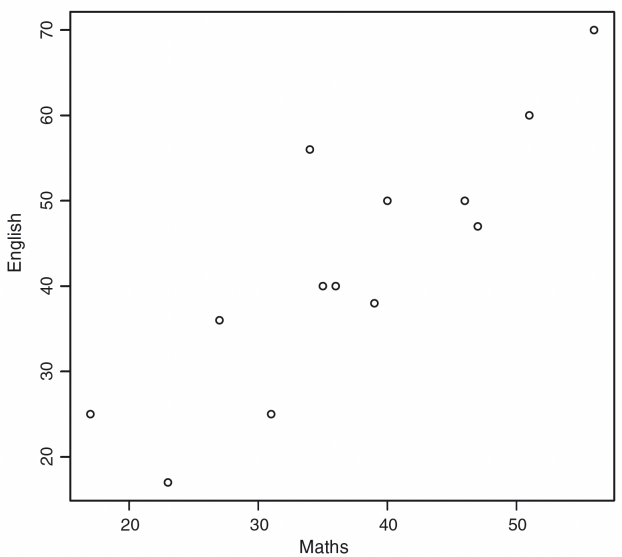
See Vance and Smith (2019) for a detailed description of interdisciplinary collaboration skills

Introductory example: Students' scores

Opening statement (Taplin, 2007)

I am a teacher's aide asked to teach remedial mathematics and I want to have the children taught remedial English also. Looking at their English scores, you can see that poor English skills cause the low mathematics scores as the children do not understand the questions . . .

Student	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13
Mathematics	23, 34, 17, 47, 35, 51, 46, 39, 31, 56, 40, 27, 36
English	17, 56, 25, 47, 40, 60, 50, 38, 25, 70, 50, 36, 40



Correlation coefficient is 0.86

There are two sets of the children's scores: one on an English test, and one on a mathematics test. My friend graphed the scores for me and worked out the correlation coefficient. But, I do not know what it means. How can I use these figures to convince the Principal that the children's English skills need to be improved also?

Introductory example: Students' scores

Possible questions from the consultant

- What do the scores represent?
- Are the children chosen from the same class?
  $\rightarrow$ We need to understand if there is an environmental factor that might affect both the English and math scores
- If yes (children from the same class), how were they chosen? Randomly?
  $\rightarrow$ We need understand whether there is a component (e.g. weak students) that affect both the English scores and the math scores and yields hence the observed high correlation
- When were the tests given? Were they given at the same time?
  $\rightarrow$ We need to check if both score are comparable in terms of individual's competences/abilities
- What is the children's first language?

Aspects of statistical consulting

Statistical consulting

- is not a simple data-analytic exercise
  $\rightarrow$ The role of the consultant neither starts from a single tidy dataset nor finishes with the analysis of that data
- requires skills combining the knowledge of statistical techniques (and the theory behind them) with **generic skills such as communication** and statistics-related skills such as recognizing which statistical techniques are appropriate
- extends into communication of questions to the client, business research, formulation of the solution approach, and business pitch of the findings

Outline

This lecture will not focus on the technical aspects of statistical consulting (statistical methodology, technical report writing, and presentation) but rather on the **non-technical aspects** including

- structure of a consulting session
- effective communication with clients
- (some) ethical questions

Structure of a consulting session

- **Beginning:** The client introduces the context and content of the problem
- **Middle:** The consultant asks questions and restate what the client said
- **End:** Discussion of next steps, allocation of responsibilities, and summary

Middle of the session: understand and identify the problem

The process of problem elicitation and formulation is based on **understanding the context**

Middle of the session: understand and identify the problem

The process of problem elicitation and formulation is based on **understanding the context**

Tips:

- Before any statistical treatment, address the following questions:
  - What is the main question to be addressed from the client's perspective? Are there related previous studies by the client or other investigators in the same field?
  - Can it be measured? What are the measurements and variables? Are there variations in the measurement process?
  - Where, when, and how will you get the data?
  - What do you think the data are telling you? Have similar data been analyzed by the client or others?
- Convey to the client your understanding of the problem and ask for corrections or additions as needed so that you have a full understanding of the relevant points, e.g., "Did I understand you correctly if I assume that you..."

Middle of the session: understand and identify the problem

The process of problem elicitation and formulation is based on understanding the context and **generating hypotheses** in conjunction with subject matter knowledge (client's knowledge and expertise)

Middle of the session: understand and identify the problem

The process of problem elicitation and formulation is based on understanding the context and **generating hypotheses** in conjunction with subject matter knowledge (client's knowledge and expertise)

Tips:

- Critical thinking is the key!
  Objectively review all information on the issue to develop an informed opinion that leads to a judgment and a precise statement of the problem
- Avoid confusing the problem with a predefined solution.
  Beware of overconfident clients that might constrain you to address the symptom of the problem rather than its root causes

$\rightarrow$ Acknowledge the fact that defining the problem is a demanding process

Communication tips

- Establish a common language/vocabulary. Avoid using technical terms that might be misunderstood by the client and risk loosing their attention
  - statistical significance vs biological significance
  - (linear) correlation vs dependence vs causation

Communication tips

- Establish a common language/vocabulary. Avoid using technical terms that might be misunderstood by the client and risk loosing their attention
  - statistical significance vs biological significance
  - (linear) correlation vs dependence vs causation
- Quite often, the client is unsure about the desired output
  - a client asks for a power analysis but did not think of how to analyse the data once collected

$\rightarrow$ design your questions to guide them in clarifying their expectations

Communication tips

- Establish a common language/vocabulary. Avoid using technical terms that might be misunderstood by the client and risk loosing their attention
  - statistical significance vs biological significance
  - (linear) correlation vs dependence vs causation
- Quite often, the client is unsure about the desired output
  - a client asks for a power analysis but did not think of how to analyse the data once collected

  $\rightarrow$ design your questions to guide them in clarifying their expectations
- Clarify accountability: what is expected from each party and do they agree to accept the responsibility?

Communication tips

- Establish a common language/vocabulary. Avoid using technical terms that might be misunderstood by the client and risk loosing their attention
  - statistical significance vs biological significance
  - (linear) correlation vs dependence vs causation
- Quite often, the client is unsure about the desired output
  - a client asks for a power analysis but did not think of how to analyse the data once collected

  $\rightarrow$ design your questions to guide them in clarifying their expectations
- Clarify accountability: what is expected from each party and do they agree to accept the responsibility?

After the analysis

- Communicate effectively the analysis and conclusions to the client by means of a report and a presentation
- Critically discuss the conclusion, pointing out possible pitfalls

Ethical questions

- Confidentiality of the data
- Transfer of intellectual property
- Authorship? Fees?

Closing example: Potting mix

Opening statement (Taplin, 2007)

I am an avid gardener but have limited time to spend on gardening. As such, I am interested in knowing which potting mix is best for getting the best growth in seedlings grown from seed. I have done some research and have come up with the three most suitable potting mixes. How should I set up this experiment and how do I gauge which potting mix is best?

Closing example: Potting mix

Possible questions from the consultant

- What is the measure you are interested in when comparing the mixes?
- Will you be using the mixes for different seeds?
  $\rightarrow$ This would add a new factor to consider in the experiment
- When and where will you be planting the seeds?
  $\rightarrow$ The information we are looking for here is the number of pots the gardener has and whether he/she can control the weather/environment for all the pots/outcomes of the experiment, i.e., can the experimental error be controlled, e.g. use of a greenhouse?
  $\rightarrow$ What is the sampling unit? Are you using a pot (one measurement per unit) or a block/cluster of pots (multiple measurements per unit)?

Conclusion: Design your experiment to isolate and identify environmental, between pot, and temporal variability through blocking, randomization, and incorporating time (round 1, 2, 3...) into your model

Project

Statement A clinical dietitian wants to compare two different diets, A and B, for diabetic patients. She hypothesizes that diet A (Group 1) will be better than diet B (Group 2), in terms of lower blood glucose. She plans to get a random sample of diabetic patients and randomly assign them to one of the two diets. At the end of the experiment, which lasts 6 weeks, a fasting blood glucose test will be conducted on each patient. She also expects that the average difference in blood glucose measure between the two groups will be about 10 mg/dl. The dietitian wants to know the number of subjects needed in each group assuming equal sized groups.

Task: Write a report answering the dietitian request, to the best of your ability. Keeping in mind that the dietitian has no background in statistics, please explain the most important technical terms, as well as the hypotheses on which you base your results.

Hints

After a thorough research in the matter, we found out that

- the standard deviation of blood glucose is typically around 15 mg/dl
- clinical dietitians usually consider a significance level at 5% and a power level at 80%

ASA (2012). Asa board statement on continuing professional development: Its importance for statisticians and the role of the asa. Technical report, American Statistical Association.

ASA (2014). Curriculum guidelines for undergraduate programs in statistical science. Technical report, American Statistical Association.

Ben-Zvi, D. (2007). Using Wiki to Promote Collaborative Learning in Statistics Education. *Technology Innovations in Statistics Education*, 1(1).

Taplin, R. (2007). Enhancing statistical education by using role-plays of consultations. *Journal of the Royal Statistical Society Series A*, 170(2):267–300.

Vance, E. A. and Smith, H. S. (2019). The asccr frame for learning essential collaboration skills. *Journal of Statistics Education*, 27(3):265–274.