# Week 11: Statistical Consulting

## MATH-516 Applied Statistics

Linda Mhalla

2025-05-05

Section 1

# Introduction

# The collaborative role of a statistician

- Statistics is an inherently collaborative discipline Ben-Zvi, 2007

→ To have a real-world impact and maximize 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, 2014, p.10), graduates of statistical science undergraduate programs,

  *"Should demonstrate ability to **collaborate** in teams and ...be able to **communicate** complex statistical methods"*

# The collaborative role of a statistician

| University | Department | Code | Class | Degree | Clients (classified by the authors) | Format (classified by the authors) | Date of retrieval | Link |
|---|---|---|---|---|---|---|---|---|
| Stanford University | Statistics | STATS 390 | Consulting workshop | PhD | Academic | Drop-in | 2022-01-15 | https://explorecourses.stanford.edu/search?view=catalog&filter-coursestatus-Active=on&q=STATS%20390:%20Consulting%20Workshop&academicYear=20202021 |
| Stanford University | Biomedical data science | BIODS 232 | Consulting workshop on biomedical data science | MS | Academic | Workshop | 2022-01-15 | https://explorecourses.stanford.edu/search?view=catalog&filter-coursestatus-Active=on&page=0&catalog=&academicYear=20202021&q=BIODS+232&collapse= |
| University of California Berkeley | Statistics | STATS 272 | Statistical consulting | MS/PhD | Academic | Drop-in | 2022-01-15 | http://guide.berkeley.edu/courses/stat/ |
| University of California Los Angeles | Statistics | STATS 140XP | Practice of statistical consulting | BS | Academic/ community | Drop-in | 2022-01-15 | https://registrar.ucla.edu/academics/course-descriptions?search=consulting |
| University of California Los Angeles | Biostatistics | BIOSTAT 402A | Principles of biostatistical consulting | PhD | None | Case studies | 2022-01-15 | https://registrar.ucla.edu/academics/course-descriptions?search=consulting |
| University of California Los Angeles | Biostatistics | BIOSTAT 409 | Doctoral statistical consulting seminar | PhD | Academic | Workshop | 2022-01-15 | https://registrar.ucla.edu/academics/course-descriptions?search=consulting |
| University of California Los Angeles | Statistics | STATS 291XP | Service learning for graduate statistical consulting | PhD | None | Case studies | 2022-01-15 | https://registrar.ucla.edu/academics/course-descriptions?search=consulting |
| University of Washington | Statistics | STATS 598 | Techniques of statistical consulting | PhD | None | Workshop | 2022-01-15 | https://stat.uw.edu/academics/course-catalog/stat-598 |

Figure 1: Statistical consulting courses in US universities

# The collaborative role of a statistician

| University | Department | Code | Class | Degree | Clients (classified by the authors) | Format (classified by the authors) | Date of retrieval | Link |
|---|---|---|---|---|---|---|---|---|
| Columbia University | Statistics | STATS GR6105 | Statistical consulting | PhD | Academic | Drop-in | 2022-01-15 | http://www.columbia.edu/cu/bulletin/uwb/#/cu/bulletin/uwb/subj/STAT/GR6105–20213-001 |
| Harvard University | Statistics | STAT 305R | Statistical consulting | PhD | Academic | Drop-in | 2022-01-15 | https://courses.my.harvard.edu/psp/courses/EMPLOYEE/EMPL/h/?tab=HU_CLASS_SEARCH&SearchReqJSON=%7B%22ExcludeBracketed%22%3Atrue%2C%22SaveRecent%22%3Atrue%2C%22Facets%22%3A%5B%5D%2C%22PageNumber%22%3A1%2C%22SortOrder%22%3A%5B%22SCORE%22%5D%2C%22TopN%22%3A%22%22%2C%22PageSize%22%3A%22%22%2C%22SearchText%22%3A%22statistical%20consulting%22%7D |
| Johns Hopkins | Biostatistics | 140.643.01 | Practice of statistical consulting | MS | Academic | Workshop | 2022-02-22 | https://www.jhsph.edu/courses/course/30415/2020/140.643.01/practice-of-statistical-consulting |
| University of Iowa | Statistics | STAT 6220 | Statistical consulting | MS | Academic | Long-term project with authentic client | 2022-02-22 | https://catalog.registrar.uiowa.edu/courses/stat/ |
| New York University | Applied statistics, social science, and humanities | APSTA-GE 2401 | Statistical consulting research seminar | MS | Academic | Workshop | 2022-02-23 | https://steinhardt.nyu.edu/courses/statistical-consulting-research-seminar |
| Wright State University | Applied statistics | STT 7910/7920 | Statistical consulting | MS | Academic | Drop-in | 2022-02-23 | https://science-math.wright.edu/statistical-consulting-center/statistical-consulting-course |
| NJIT | Mathematical sciences | MATH 664 | Methods for statistical consulting | MS | Academic | Case studies | 2022-02-23 | https://math.njit.edu/sites/math/files/Math_664-SP21.pdf |

Figure 2: More can be found in Shilane et al. (2022)

# The collaborative role of a statistician

- Collaborative skills are essential for professional statisticians. According to (ASA, 2012, p.1),

  *"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"*

See Vance & Smith for a detailed description of interdisciplinary collaboration skills

# Statistical consulting

Based on Vance & Smith

> *"Statistical consulting is "working cooperatively with clients to answer statistics or data questions.""*
> *"Statistical collaboration is "working cooperatively with domain experts to create solutions to research, business, and policy challenges and achieve research, business, and policy goals.""*

$\Rightarrow$ the main difference is the aspect of the problem. Both require, among others,

- verbal and written **communication** skills
- adapting **communication** techniques based on factors such as the client's cultural background

# 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 ...

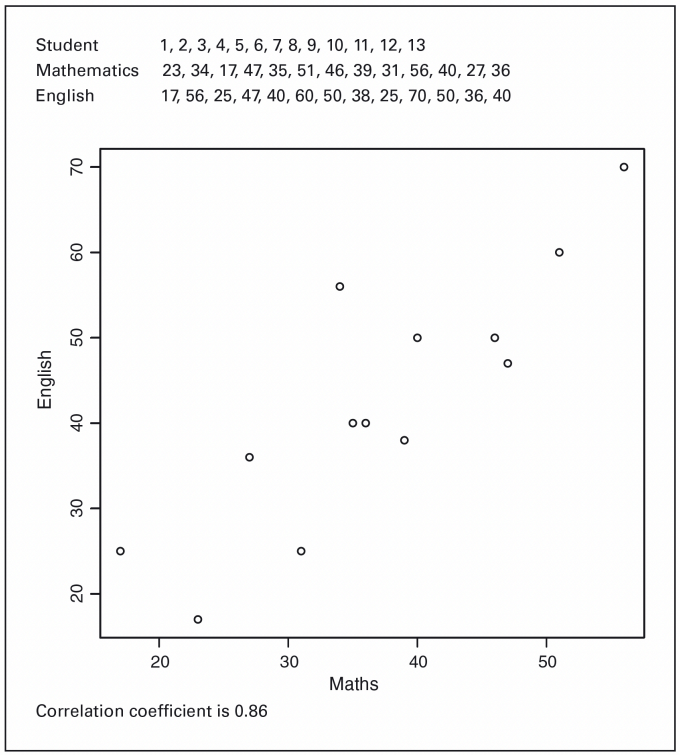


Figure 3: Children's scores

These 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?

# Students' scores (continued)

**Possible questions from the consultant:**

- What do the scores represent?
- Are the children 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 to 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?

**General comment**: The distinction between correlation and causation is important!

# 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 and what are their limitations
- 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

Section 2

# Non-technical statistical consulting skills

# Consulting session structure

**Beginning:** Introduction of the problem by the client

**Middle:** Deep dive into the problem details by the consultant: The consultant asks questions and restate what the client said

**End:** Summary and discussion of next steps and responsibilities

# 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/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)

## 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
- 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?

# Middle of the session: Analysis and post-analysis

Up to now, the statistical consultant has

- understood and refined the questions to be answered, and
- accessed and analysed the appropriate data.

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 potential pitfalls
- Present recommendations for decision making and implementation

# Ethical questions

- How is data confidentiality handled?
- Who owns the intellectual property developed during consulting?
- How are authorship and fees managed?

# 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?

# Consultant's questions on potting mix

- 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 they 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