Statistical Consulting Short Course, May 1998

Murray K. Clayton

University of Wisconsin—Madison

clayton@stat.wisc.edu

April 1998

An experience of mine as a graduate student:

- although we study a lot of statistics, it is easy for a client to ask a question that we do not know the answer to.

For most of us our training focused on:

- Theory and Methods
- Analyzing Messy Data
- Interacting with Clients
- (Data Management)

For most consultants, their time is spent:

- Interacting with Clients
- Data Management
- Analyzing Messy Data
- Methods
- Theory

Outline:

- Definitions, Roles of a Consultant
- Bibliography
- Interacting with the client
  - Your goals, their goals
  - Identifying and clarifying goals
  - Effective communication
  - Timing and style

Outline (continued):

- Dealing with problems
- Data analysis
- Experimental design
- Report writing
- Ethics
- Take-home problem

Outline (continued):

- Tomorrow:
  - Discussion of take-home problem
  - building a helicopter – a simulated experiment

Definitions:

- "Consultant" = Statistician
- "Client" = Scientist, Economist, Business Researcher

What are the roles that a consultant plays?

- Police officer
- Firefighter
- Technician
- Repair worker
- Teacher
- Ideally: Colleague and collaborator

The Ideal Relationship:

In the ideal relationship between statistical consultant and client, the two parties function together as a team. They work as colleagues and collaborators, each bringing to the project their skills and expertise to accomplish, in the end, what neither could do alone.

(Both care about project: if project fails, both are unhappy.)

- Some Reference Texts for Your Science: Statistics
  - Snedecor and Cochran
  - Milliken and Johnson
  - Encyclopedia of Statistical Sciences
  - Freedman, Pisani, Purves
- Articles and Books on Consulting (Human Relations)
  - Boen and Zahn
  - Chatfield (book and Stat. Sci. article)
  - Hunter (1981a)
  - Joiner (Encycl. Stat. Sci.)

Interacting with the client: your goals

First:

- To determine the client's goals. (What's the Question?)

Then:

- If analyzing data: how was the experiment done?
- If designing an experiment: what are the constraints and possibilities?

Why do you want to know about goals?

- There are often several approaches to a particular problem. Choosing the most appropriate one depends on the goals and constraints of the project. The more deeply you understand these, the better your advice will be, and the more valuable your contribution will be.
- Often the choice of an analysis involves judgment — yours and the client's.

Determining the client's goals:

(Don't ask: "What are your goals?")

Determining the client's goals:

You need to understand their research problem.

Take your time, and ask A LOT of questions.

Take your time and ask

A LOT of questions.

Stephen Covey: 7 Habits of Highly Effective People

- Seek first to understand, and then to be understood

If your first meeting is $1\frac{1}{2}$ hours long, you should spend at least 45 to 60 minutes discussing the science involved.

Asking questions about the science:

- usually pleases and impresses the client
- helps you understand the goals, constraints, and issues
- demonstrates that you are interested
- helps you make a contribution — makes you more a part of the team

You want to understand the science well enough that you can:

- talk intelligently about it
- understand the process of data collection, randomization
- understand the constraints
- make a contribution
- understand the impact of the results

A good consulting statistician:

- has a good scientific sense
- is eager and willing to learn
- understands the scientific aims and methods
- has a genuine concern for the project and its outcome.

In the ideal relationship we have:

- the client — expert in the science, sympathetic to the statistical issues
- the statistician — expert in the statistics, sympathetic to the scientific issues

Together they accomplish what neither one would alone, and they both learn in the process.

The impact that you can have is highly satisfying.

The science you learn is fascinating, and is part of what makes the job rewarding.

You need to understand the science —Getting a science background:

- Journals in the field
- Science News
- Encyclopedias
- Be a good learner
- On site: visit the lab, go to lab/staff meetings (even when statistics is not on the agenda).

Your aim is to get a feel for the topic — not to become an expert.

Effective Communication

Use the “playback” technique: in your own words try to explain to the client what the science is about.

- “Let me see if I understand . . . ”

Effective Communication

- Draw pictures . . .
- Use analogies.

Effective Communication

- Limit your use of statistical *and* scientific jargon.

Effective Communication

- Listen.
- Interrupt rarely.
- Don't be afraid to ask questions about the science.
- Don't get sidetracked unnecessarily.
- Take notes – but don't get bogged down with them.
- Pause and think.

Identifying and clarifying goals

Short term goals

- Need to submit an abstract by Friday
- Want to know how to respond to a reviewer's comments
- Want to know best way to construct a particular plot

Medium term goals

- Design, analyze or write up a specific experiment

Long term goals

- Design a series of experiments
  - preliminary and pilot experiments
  - main experiment
  - observational studies

Sometimes the client is not certain of the goals (junior member of team, inexperienced researcher, etc.).

Often your most important contribution will be to help identify and clarify the goals of the study.

Determining goals

- Don't try to grasp everything at once.
- Do ask "Why?" frequently (with interest, not doubt).
- Be flexible:
  - from specifics to the general
  - from the general to the specifics

Determining goals

- Try to develop an outline
  - Major sections = major goals
  - Subsections = minor goals
  - Subsubsections = individual steps along the way

Focus more attention on initial stages, but have a sense of how each step leads to the next.

Fragments of an Example Outline

Main goal: to determine the amounts of galactose in typical diets of children with inability to metabolize galactose.

1. Run preliminary assays on diets produced in lab to evaluate assay methods.
   (a) Construct typical diets.
   (b) Begin investigating variability in lab assay.
   (c) Determine costs of lab assay.
2. Run pilot study to determine major sources of variation.
   (a) Identify population of interest.
   (b) Guess major sources of variation.
   (c) Determine feasibility of gathering dietary information.
   (d) Determine sampling plan.
   (e) Implement pilot, and keep notes on problems.
   (f) Analyze pilot, and determine need for additional pilot work.
3. Implement main study to determine dietary quantities

Timing

Often it will be your responsibility to keep things going.

- Be prepared for meetings.
- Keep a regular schedule for meetings.
- Often frequent (every 7 – 14 days) and short ($\frac{1}{2}$ – $1\frac{1}{2}$ hr) meetings are more effective
- Make it clear who is responsible for the tasks to be done

Take the time you need — do not feel the need to give instant answers, but do be willing to make preliminary suggestions.

Style — Consulting Involves Teaching

The client may know very little statistics, but remember that they are experts in their field, just as you are an expert in yours. So:

- Each member of the team should be a good teacher, and a good student. Be sure that *they* understand *you*, at least at an intuitive level.

- Be sure that *you* understand *them*, at least at an intuitive level.

  (Ask, don't assume.)

Teaching (cont'd):

- Determine the client's statistics background, and adjust to that.

  Adjust:

  - your style of presentation
  - the statistical approach you use

Use clear intuitive explanations.

Working as a Team

- Visit the lab. Attend meetings even when statistics is not being discussed. (The more you know, the better you'll do. The more visible you are, the more impact you will have.)
- Contribute to the science — get involved at every step. (Show a clear interest.)
- When writing a report or manuscript, read *all* of it, and write more than just the statistics section.

Working as a Team: What the Client Should Expect

- A good statistician will work hard on the project but there are limits to what can be learned from any given study.
- Some experiments fail. Extract what information you can from them, and move on.
- The statistician might not have an immediate answer – they may have to check references, etc. before providing suggestions.

In a collaborative situation, these issues amount to trust (on their part) and responsibility (on your part).

Problems:

- Poorly designed experiments
- Weak clients
- Messengers
- Uncooperative clients

Poorly designed experiments

- Be sympathetic — never ridicule. (It's you and me against adversity.)
- Explain what the difficulties are.
- Explain what *can* be done. (plots, use as pilot info, discuss assumptions)

Weak clients

- Be patient.
- Use simple analyses.
- Use unsophisticated explanations (Freedman, Pisani, Purves).

Messengers — subordinates of the “real” boss.

$\boxed{\text{modify or delete??}}$

- Broaden communications.
- Invite the boss along.
- Call the boss.
- Be cautious — try to understand the messenger’s perspective.

Uncooperative clients

modify or delete??

Try to identify source of conflict — usually fear, ego, external pressures.

- If fear, external pressures
  – be sympathetic and patient.
- Ego

Recognize that some people are naturally uncooperative. You can't force everyone in the world to like you or respect you, however much you think you deserve it.

Data analysis

- Why were the data gathered?
- How were the data gathered?
- Scrutinize and plot the data.
- Model the data as needed.

Why were the data gathered?

- Goals (Are there primary and secondary goals?)
- What sort of results were expected?
- To what extent is there interest in finding patterns and results not anticipated in the original plan?

How were the data gathered?

- From a designed experiment? or from an observational study?
- Was randomization used? How?
- What is/are the experimental unit(s)?
- Are there hidden confounding or blocking conditions?

Scrutinize and plot the data.

- Does there seem to be support for the hypotheses of interest?
- Do unexpected patterns arise leading to new or revised hypotheses?
- Are there suspicious patterns (impacting on assumptions)?
- Is there evidence of data entry errors?

“Statistical tests are used to tell you how excited you should get about the patterns you see in the plots.”

Fitting the Data as Needed — The Art of Statistics.

- In virtually every case, there will be several possible approaches — none of which will be perfect.
- You will often need to improvise.
- There is rarely a single “right” answer.
- The analysis should be guided by the needs of the client, and the way the data were gathered.
- A simpler analysis that serves the client’s needs is almost always better than a complicated one — easier to explain to colleagues, easier to justify.

Optimality: There are many factors that play a role in determining an "optimal" analysis — ease of understanding, clarity, simplicity, timeliness, and statistical optimality.

Modeling

- Start with simple models
- Model subsets of the data
- Work toward more complex models if necessary.
- Check that the models make sense and are appropriate.
- Involve the client in the decision-making.

Check the assumptions of each model you try.

- Based on the client's knowledge.
- Based on the way the experiment was run.
- Check residuals for each model you fit.
- Look for outliers and autocorrelation.

- Do the residuals suggest the need for a transformation or addition of new variables?
- Should a weighted analysis be used?
- Modify your models as needed, and recheck your modified models.

Experimental design

The experiment should be designed so that its analysis will be as straightforward as possible.

Experimental design

- Understand the goals (exploration or confirmation)
- Understand the constraints
- Use pilot studies
- Use simple designs
- Understand how data will be collected
- Block when necessary
- Stress the need for randomization
- Be cautious of possible missing data

Power — what good is it?

- Determining sample size
- Assessing the value of the experiment

Subsampling and split plot designs

Cochran and Cox: “The Experimental Unit (EU) is that group of material to which a treatment is applied in a single trial of the experiment.”

Example: An experiment was conducted to determine the effect of electrical stimulation on neural growth in the brains of rats. There were 3 treatments used:

1. Control — no stimulation
2. Low — 20 stimulations
3. High — 40 stimulations

The experiment was conducted as a completely randomized design, with 5 rats per group. After treatment, each rat was euthenized, the brain was removed, and four sections (thin slices) from the hippocampus were examined under the microscope. Each section was evaluated for evidence of neural growth.

The EU is the rat. The sections are *subsamples.* There are 60 data points, but only 15 EUs.

Model: $Y_{ijk} = \mu + \alpha_i + e_{ij} + \delta_{ijk}$

ANOVA Table:

Source	df	E(MS)
Trts	2	$\sigma_\delta^2 + 4\sigma_e^2 + 20\sum \alpha_i^2/2$
Rats	12	$\sigma_\delta^2 + 4\sigma_e^2$
Sections	45	σ_δ^2
Total	59	

$$Var(\bar{Y}_{1\cdot\cdot}) = \frac{\sigma_\delta^2 + 4\sigma_e^2}{20}$$

A CI for the mean of treatment 1 uses a t-distribution on 12 df.

In general, if there are s subsamples and n experimental units per treatment, then:

$$Var(\bar{Y}_{i\cdot\cdot}) = \frac{\sigma_\delta^2 + s\sigma_e^2}{ns}$$

Example: $\sigma_\delta^2 = 1$, $\sigma_e^2 = 4$, $n = 5$.

s	$Var(\bar{Y}_{i\cdot\cdot})$
1	1.00
2	0.90
4	0.85
10	0.82
100	0.80

If $\sigma_\delta^2 << \sigma_e^2$ a lot of subsampling does little good.

Useful trick: average over the subsamples before analysis.

Split-plot designs

These are designs with nested treatments.

Example: As before, 5 rats are randomized to each of the 3 stimulation treatments. There are four different stains that were used to help evaluate growth in the brain. For each rat, four sections of the brain are removed, and a different stain was used on each section, chosen at random.

The whole plot treatment is the stimulation level; the subplot treatment is the stain.

ANOVA Table:

Source	df
Stimulation level	2
Whole plot error	12
Stain	3
Stim × Stain	6
Subplot error	36
Total	59

Report writing

- Be brief — don't give a historical account of your efforts
- Describe analyses in enough detail that a colleague could reproduce the analysis
- Address the client and their audience — read the intended literature for style suggestions.
  - "There is strong evidence that the coconut oil diet increased plasma cholesterol compared to the control diet (p = 0.002)."
  - Not: "We reject the UMPU test of $H_0 : \mu_1 = \mu_2$ at $\alpha = 0.05$."

- Similarly, avoid statistical jargon — don't say:
  - "We assume $X_1, X_2, \ldots$ are iid $N(0, \sigma^2)$"
  - or: "We fit the model $X_{ijk} = \mu + \beta_k + \alpha_i + \tau_j + (\alpha\tau)_{ij} + \epsilon_{ijk}$"
- Avoid trivia: excessive digits, unneeded information
- Use plots and tables — use captions effectively

Ethics

(Ref: Sigma Xi: *Honor in Science*)

modify or delete??

- Your responsibilities to science and to statistics
- Your responsibilities to the client
- Your responsibilities as a co-author
- Confidentiality