Statistics and Science

• One definition of science: a process for uncertainty reduction
  – If so, then statistics is an essential underpinning of science (one of its core languages)
• Statistics is a necessary (though not sufficient) tool to turn data (the results of measurement) into something more…

What is…?

• Statistics: the science of learning from data in the presence of uncertainty
  – The science of measuring (estimating), controlling, and communicating uncertainty
• Data Analysis:
  – Inspecting, transforming, and modeling data
  – Discovering useful information
  – Generalizing beyond the data at hand
  – Supporting decision-making

…in the Real World

• The real world of data analysis is full of problems, non-ideal behaviors, and errors
  – Problems with the data
  – Information problems (too much, not enough, or the wrong data)
  – Problems with the model used to fit the data
  – Problems with the analysis
• In my experience, most students are unprepared to deal with the messy real world of data analysis

…in the Real World

• Problems with the data:
  – Outliers (flyers): something unexpected has influenced some of the data
  – Sampling (improper range of data, the wrong data, high leverage and influential data)
  – Sample size (not enough data points, or too much data)
  – Data uncertainty (and worse, when you don’t know what the uncertainty is)
  – Unknown context (conditions of the experiment), or unknown context uncertainty
• Problems with the analysis:
  – Poorly defined goals – how do you judge success?
  – The data doesn’t meet the needs of the analysis objectives
  – The analysis makes unwarranted assumptions
  – The model is wrong: do you have a way of judging this?
  – The data doesn’t support your conclusion (e.g., choosing one model over another)
The Knowledge Hierarchy

Chain of Increasing Value (the Knowledge Hierarchy)

- **Data**: A collection of numbers, with known context and uncertainty estimates
- **Information**: The right data, at the right time, in the right context, organized for access
- **Knowledge**: Interpretation of the information based on an understanding (that is, a model) of cause and effect
- **Decision**: Acting on the knowledge for some benefit

Knowledge Hierarchy Example
(from Semiconductor Lithography)

- **Data**: The measured dimensions are $56.2 \pm 0.9$ nm, $54.4 \pm 0.8$ nm, etc.
- **Information**: Transistor gate size varies systematically by $3.5$ nm across the slit
- **Knowledge**: A $-75$ nm focal tilt adjustment would reduce the systematic CD variation across the slit from $3.5$ nm to $1.8$ nm
- **Decision**: Let’s make the focus tilt adjustment before the next lot is run (we think it will have a positive, noticeable impact on yield)

Moving Up the Knowledge Hierarchy

- Moving up the knowledge hierarchy means moving from the specific (data) to the general (a model)
  - This is what data analysis is all about
- Given a set of data, are you extracting the most knowledge possible? Are you making the best possible decisions?
- How can you plan an experiment (data collection and analysis) to promote good decision making?

Five Steps in the Decision Making Process

- **Preparation** (planning) – be prepared to turn data into information, with a specific model and decision in mind.
- **Measurement** – generate the data, minimize the uncertainty, and keep track of the context.
- **Analysis** – using a model, turn information into knowledge (i.e., assign a probable cause to what is observed).
- **Decision** – with an estimate of the uncertainty in the analysis results, perform a risk/benefit analysis and make a decision.
- **Post-mortem** – have we learned anything from this experience that we can use to do things better next time?

Note: moving up the knowledge hierarchy is an iterative process, with feedback loops. Don’t let the linear “hierarchy” nomenclature fool you into thinking that this is a linear process.

Philosophy of Science

- Science is about model building – we judge our science on an axis of low to high accuracy, not the axis of true or false
- Theory choice involves judgments based on:
  - Accuracy of predictions compared to ex post facto measurements
  - Coherence with the rest of science
  - Usability of the models (often equates to simplicity)
- Data analysis is the foundation upon which science rests

Lecture 1: What have we learned?

- Define “statistics” and “data analysis”
- What are some of the problems affecting data in the real world?
- What are some of the problems affecting data analysis in the real world?
- What are the four levels of the knowledge hierarchy?
- What are the five steps in the decision making process?