Objectives. The goal of the course project is to two-fold. First, it is a sort of introduction to the research process, i.e., reading existing literature, thinking about new ways to approach an interesting problem, carrying out the work, and presenting your findings in a written report. Second, I expect that the results of these course projects will be, at least, very good starts towards a journal publication and/or a PhD dissertation.

General Requirements. Students first select a research topic of interest. A list of potential topics is given below, but students are welcome to select a topic that is not in the list, subject to instructor approval. Roughly, after identifying a problem of interest, students will read some existing literature with the intention of finding a direction that would benefit from a new approach and, if possible, provide some first steps towards this new approach. Not all projects fit with this general structure, but the expectation is that some progress towards something new will be made. The instructor will be available to provide guidance at each step.

Specific Requirements. There are three specific requirements for the project. The first two are just natural steps along the way, designed to give the instructor an opportunity to provide some guidance.

1. A proposal document that simply lists your name(s), the general topic to be considered, and a few key books/papers that have been identified as relevant.

2. A progress report that describes how you have narrowed down your focus and, if possible, what specific problem you plan to address in the final report.

3. A final written report. Students are encouraged to prepare their report in \texttt{\LaTeX}, but this is not required; a formal bibliography with proper references is required, however. The report should introduce the problem, explaining why it is important and/or interesting, review the existing literature on the problem, identifying a gap that is to be filled, and present the new ideas/results to fill the identified gap.

Timeline. (Subject to minor changes.)

\textit{October 3rd:} Proposal document due.
\textit{November 6th:} Progress report due.
\textit{December 8th:} Final written report due.
Below is a rough list of project topics. They are arranged into three categories, but likely all projects would involve elements from each category.

**APPLIED/METHODOLOGICAL PROBLEMS.**

- Survival analysis and other problems in reliability theory often involve censored data. I am almost done with a first attempt at IMs for censored data, but there would be lots of possible follow-ups/improvements to this.

- Time-series, longitudinal, and spatial problems involve dependent data, and the interesting problem would be to make inference on those parameters that help to describe the dependence structure.

- Discrete-data problems are particularly challenging for all approaches to statistical inference. Analysis of contingency tables, log-linear models, logistic regression, etc are all very important and yet to be addressed directly from an IM perspective.

- Generalized linear models (GLMs) are widely used in applied statistics and all the inference is based on asymptotic approximations. It would be interesting to develop some specific (and then general) IM tools for inference in GLMs.

- Problems that involve measurement error and/or instrumental variables seem fairly challenging. The point is that the measurement error (or other kinds of unobservables) introduces some bias that the usual least squares estimators will miss, even asymptotically. Moreover, many of these problems fall within the Gleason–Hwang class where the usual methods for constructing interval estimates will fail, so perhaps new ideas are needed.

- Classification and clustering are important problems that I don’t understand very well. For example, what are the relevant assertions in clustering? These and other IM-based machine learning tools would be a very interesting and useful contribution.

- Another class of problems often overlooked in introductory courses are those with non-trivial parameter constraints. Some work on IMs for constrained problems is available, and it would be interesting to explore the use of these techniques in other problems. Interestingly, non-trivial constraints often pop up unexpectedly in certain problems (e.g., Stein’s paradox) so a better understanding is needed in general.

- Meta-analysis concerns the combination of separate analyses of several data sets concerning the same underlying problem. There’s a couple different routes one can take in this case. First, one can combined the data from individual studies before constructing the IM, which would boil down to some conditioning operations; second, one can try to combine after constructing an IM from the individual analyses, and “Dempster’s rule of combination” could be useful, or maybe something else.
• Prediction is a fundamentally important problem in statistics, but often doesn’t get much attention in statistics theory courses. A general IM approach for prediction is available but the emphasis is on independent models. Of course, prediction in dependent-data problems is potentially even more important, so it would be interesting to suitably extend the currently available IM approach.

• IMs for inference on covariance matrices or other kinds of non-vector parameters? What about functional data analysis?

• Network problems have received a lot of attention recently, but my understanding is that inference is not yet well developed in this context.

• There are a number of problems that involve some kind of “model selection,” e.g., variable selection in regression, order selection in mixture models, order selection in autoregressive models, etc. Some basic tools for these problems are now available, and applying these to other contexts would be interesting and useful; “model selection” is discussed again below.

• Nonparametric problems are those that involve an infinite-dimensional parameter, e.g., a density function, regression function, etc. Certain nonparametric problems are relatively easy, such as cases like the “one-sample test,” but otherwise this area is completely open.

• . . .

Computational Problems.

• A common summary of the IM output is the “plausibility function” and a corresponding “plausibility region.” This region corresponds to a level set of the plausibility function which, except for in simple problems, can only be evaluated via Monte Carlo. Doing separate Monte Carlo runs for each candidate parameter value is not efficient—how to do it better? Importance sampling and related ideas would be useful, maybe also stochastic approximation.

• This is too vague, but it would be very interesting to develop some “black-box” techniques for IM computation. That is, a user inputs their data, model, and other relevant information and the software returns suitable output.

• . . .

Theoretical Problems.

• A Bayesian’s criticism of the IM approach might be that it’s not coherent in the betting sense while the Bayesian posterior distribution is. These coherence arguments, however, involve a sort of symmetry assumption which may be questionable. It would be interesting to see if relaxing the symmetry assumption can provide a coherence-like justification for the IM approach.
- Related to the previous point, certain Bayesians are attracted to the Bayesian approach because of the connections to formal decision theory. That is, as soon as a distribution for the quantity of interest is available, along with a measure of loss in taking a particular action, one can formulate the problem of finding the action which minimizes the expected loss. The IM approach does not work with probabilities, at least not in this more familiar Bayesian-like way, so is there a parallel to this kind of decision theory?

- Again, related to the previous two points, even if we don’t really understand all that well what probabilities mean in a given application, we are at least familiar with thinking in terms of probabilities. However, the belief and plausibility function output we get from the IM approach is less familiar. Is there a betting system in which these quantities can be viewed as prices, or upper/lower bounds thereof? Of primary interest here, I think, is the gap between belief and plausibility, the “don’t know” component, as this actually seems meaningful—and perhaps necessary—in understanding the kind of uncertainty in a statistical inference problem.

- Most of the classical statistical theory relies on “regularity conditions.” While many of the standard models satisfy these conditions, e.g., exponential families, these are not the only models one can encounter. A feature that distinguishes regular from non-regular problems is that sufficiency does not provide a fully satisfactory data reduction, so going all the way requires something extra, often a conditioning argument. The IM approach relies on a strategy that involves solving a differential equation, and there’s several applications of this now available. But I don’t really know all that much about differential equations, so it would be advantageous to understand when these can be set up and solved, and how to do so.

- IM marginalization is straightforward in certain nice problems, but there are important problems outside this class, such as the Behrens–Fisher problem. Our current strategy to overcome this is to find a conservative approximation. Understanding when this kind of approximation is available and how to get it is an important problem. There’s also some challenging marginalization problems that don’t seem to fit exactly in the situations we’ve studied; for example, if \( X_i \sim N(\theta_i, 1) \), \( i = 1, \ldots, n \), independent, and the goal is inference on \( \psi = \max_i \theta_i \), the largest of the means.

- In some cases, a combination of conditioning and marginalization is needed to reach a good solution. Inference on the heritability coefficient in a linear mixed-effects model is a good example. Any principled way to think about this iterative conditioning/marginalization process?

- High-dimensional problems all involve an implicit low-dimensional structure, otherwise, quality inference would be impossible. Typically, this low-dim structure is introduced via some form of regularization and, for example, estimation or model selection corresponds to a penalized optimization problem. However, this does not lead directly to “inference” in the sense we discussed in class, so this fundamentally important area is completely open. In the paper mentioned above, the regularization is incorporated via a “simultaneous validity” condition, which is like an IM version.
of the notion of multiplicity correction. This is actually for a low-dimensional problem, so there is lots of room for improvement and extension.

- in nonparametric problems, e.g., regression function estimation, it is common to impose certain constraints, such as smoothness. Usually this constraint is accommodated through a penalty function, but why not think of it actually as a sort of constraint? Perhaps the ideas for constrained IM can be adapted to this setting to provide new insights?

- It is common for there to be some partial prior information available in a given problem, i.e., we know something about the parameter but maybe not enough to develop a full prior distribution. Ideally, one could incorporate exactly the information available, nothing more and nothing less, but existing frameworks don’t allow for this. It would be possible, however, to describe this partial prior information via a belief/plausibility function, rather than a probability distribution, and incorporate this into the IM in a relatively easy way. This amounts to a generalization of the constrained IM setup discussed above, and seems quite interesting.

- Implementation of the IM approach requires a choice of predictive random set. We currently have some theory (and more intuition) that drives this choice, but this needs some further developments. We first need a satisfactory definition of “optimality” in this context, then an idea on how to achieve it. My hunch is that the best strategy is to use random level sets corresponding to the density function of the auxiliary variable, but this needs to be explored further. Also, there are cases where, at least formally, there is a family of predictive random sets indexed by the (interest) parameter space. How does this situation fit in with the general optimality theory?

- ...