Stat/Biostat 572
Statistical Methodology
Project

Ken Rice

UW Biostatistics

March 27, 2017

Overview: About 572

- Welcome!

- Course website on Canvas
- PDF syllabus with additional info
- Lecture slides will be posted – and many other resources

Ken's office hours: Tuesday 11am-12pm in HSB F656.

Overview: Goals of 572

- Learn about research interests of faculty (and you)
- Learn how to read a paper carefully
- Learn what goes into a paper ... or should
- Learn how to give a good presentation
- Learn what you need in order to get started on research

Overview: Stat & Biostat

- All of you are (should be!) enrolled in Biost/Stat 572
- Course expectations and grading are the same for all students
- Stat students also take the Stat Methods Prelim, which happens later and is evaluated by a separate committee

Overview: Learning objectives

- Identify the components of an effective presentation.
- Give an effective research presentation to a statistical audience
- Perform a review of the statistical literature
- Read a statistical research paper
- Write a statistical research paper

Overview: Grading

- Class attendance and participation: 10% (This is compulsory – so do attend, ask questions, etc)
- Draft of Intro & Motivation: 5%
- Intro presentation: 10%
- Update presentation: 10%
- Final presentation: 25%
- Final written report: 40%

Overview: Dates & deadlines

Three deadlines:

- Thursday, March 30: e-mail me your paper selection.
- Tuesday, April 25: draft of Intro & Motivation due. E-mail it to me, and bring a hard copy to class
- Thursday, June 1: final written report due in class.

Plus three presentations:

- Intro presentation: weeks 2 & 3
- Update presentation: weeks 5 (half) 6 and 7 (half)
- Final presentation: weeks 9 & 10

Other sessions will include lectures, exercises about writing and editing, and discussion of other people's talks (seminars). Not every session runs until 3.20pm.

Overview: Paper selection

- Read the list of papers on the course site, and pick one
- Or, find your own that has not been previously studied
- E-mail me your selection by this Thursday
- First come first served!

Some of you have done this, some have not. Either is okay.

Overview: How to choose a paper

- Will this lead to your dissertation work?
- Are you considering working with the faculty member who suggested the paper?
- What will this paper selection entail, in terms of:
  - replicating numerical results?
  - working out analytical results?
- Does it interest you?
- Is it fairly recent?

Overview: Draft intro & motivation

- No more than three pages
- What problem is being studied, and why?
- Due Tuesday, April 25
- Upload to course site, and bring a hard copy to class

Overview: Final report

- Due Thursday, June 1 in class
- Should be $\leq$ 20 pages, plus any appendices you need
- Appendices may contain longer derivations, R code, etc. But not your explanations/reproduction/criticism of the original paper

Overview: Final report

Reports will be assessed as follows:

- **Introduction, Motivation, and Background Literature:** What problem is being studied, and why?
- **Methods:** What are the details of the proposed method?
- **Implementation and Derivation:** A correct implementation of the method, as well as simulations and data analysis, is expected. Key results should be derived
- **Conclusions and critique:** Did the paper stimulate other work? What are the paper's shortcomings?
- **Presentation:** Is the writing style effective? Are tables, graphs, and proofs clear?

Overview: Intro talks

- Weeks 2 & 3
- 20 minutes long, plus time for questions
- Should answer the following questions:
  - What is this paper about?
  - Why is that an important problem?
  - How is this problem typically solved in the literature, i.e. what was the state-of-the-art before this paper?
  - Very briefly, what approach is taken in this paper?

Overview: Update talks

- Weeks 5 & 6 & 7
- Two options;
  - 10 minutes, graphics-only explanation of a key idea
  - Up to 3 equations, if you need them – but try not to
  - This is a challenge!

  - 20 minutes, on a topic you found difficult
  - Explain it, fully
  - More material to cover, but words okay

Overview: Final talks

- Weeks 9, & 10
- 25 minutes long, plus time for questions
- Assume the audience has not seen your paper before, but is familiar with 570s methodology

For later talks, students will be responsible for introducing, timing nad questioning speakers.

Overview: Evaluating talks

- **Presentation:** Was the talk clear and audible? Was the overall style effective? Were equations, graphs and tables clear and effective? Was the pace good? Did the speaker use the time well? Did was the speaker enthusiastic? Any distracting behavior or mannerisms?
- **Organization:** Was the talk well organized? Good balance of introductory and more advanced material?
- **Content:** Was the problem addressed by the paper well motivated? Were the methods clearly explained? Was the information accurate? Did the speaker convey a deep understanding of the method?

Overview: Reading the paper

In the written report and final talk, students will be expected to:

- Summarize main contributions and novelty of paper. This will require reading literature that preceded the paper in order to provide context, and papers that followed it in order to understand impact
- Understand all of the analytical work in the paper, including filling in all of the analytical arguments.
- Reproduce at least a subset of the simulation studies and data analyses in the paper
- Critique the paper

Overview: LaTeX

- Your written reports & slides should be in LaTeX
- If you haven't used Latex or Beamer before, get started quickly!
- Use BibTex for references
- The course site will have examples and tutorials. But we won't spend much time on these in class, you are expected to be able to acquire these skills without explicit instruction

If you do get stuck, ask.

Finding a paper

With modern tools, it's no longer hard to search the literature:

- Google Scholar
  - Full-text links available; if off campus use
    `http://scholar.google.com.offcampus.lib.washington.edu/`
  - More relevant than Google results
- PubMed: Most medical or medically-relevant research ends up here. Can search directly, or Google
- arXiv: Many papers show up 1/2 years ahead of publication – but it depends on field
- JSTOR: most reliable for older papers
- Web of Science: Thomson-Reuters citation database

Finding a paper

Some other resources;

- Google Books; worth a look, may not be reliable
- Real books; `lib.washington.edu` – some are online, pick up the squashed tree variety from HS library
- Researcher's websites, and their CVs
- Encyclopedia of Statistics, also of Biostatistics
- Wikipedia: ... very good for math-type results, and 'Biology 101'. Beyond this, quality of material is a lucky dip; some of it is very bad
- Conference abstracts – e.g. JSM, WNAR, ENAR
- `RSiteSearch()`, also 'Task Views' on CRAN

Reading the literature

This is harder. For most papers;

- Read the title, authors and abstract
- Maybe check the references, and citations
- Note some or all of these for future reference

Once you have found something proto-useful;

- Get hold of the whole paper
- Look through conclusions/discussion, introduction, all the figures/tables, and *possibly* some stuff in the middle
- If the paper is relevant, make a note of it (e.g. save the PDF)

Actually reading most scientific papers is non-trivial...

Reading the literature

For papers worth a thorough reading;

- Print them
- After reading 'front and back', go through to try to get an overall view of the paper; motivation, methods, results
- Make notes/highlight as you go; what terms are new to you? What results are important?
- Read through again, more carefully, trying to answer these questions – or making new ones. Consult other sources when needed
- Read *critically* – being aware of publication bias

One system* for *fully* digesting papers then suggests a third pass;

* Keshav (2007) ACM SIGCOMM Computer Communication Review 37:3

Reading the literature

The 'third pass'* attempts to reproduce the paper;

- Understand the proofs, and fill in the technical details
- Code the methods, and validate them – at least in 'toy' examples
- Critically evaluate all claims made in the paper – incorporating reasonable background knowledge of underlying science

This is generally very hard, slow work (10 weeks!); teaching yourself material from a high-level textbook may be required, in order to fully understand the material.

* Just getting the big picture may take > 2 thorough 'reads'. Doing the 'third pass' will involve reading the paper multiple times, though not sequentially

Reading the literature

Why isn't reading papers easier?

- Journals have limited space; expect terse exposition
- Science attracts those who opted out of 'writing' classes!
- Writing clearly about nuanced topics is just difficult – just try explaining *exactly* what a p-value is, and does
- It's called research because we *don't yet know* how it all fits together, or which parts are most important. Be aware the 570s moves from "being taught" or "searching for a correct answer" to asking questions like:
  - Why is it done this way?
  - Why isn't it done another way?

Also: it's good practice to keep dated, detailed notes of any calculations you carry out or thoughts you have when reading a paper. (In detail or otherwise)

Published Papers and Tech Reports

Increasingly, papers are posted online as tech reports before publication.

- Often the paper will change quite a bit between when it's a tech report and when it's published
- Where possible, read the published paper, and not a preliminary version
- This is crucial if you'll be:
  - e-mailing the author about the paper
  - reading it very carefully (like for 572)
  - citing it in your own work

Also note you can subscribe for a daily email digest from arXiv.

Resources

See the course site for material on research, writing, and presenting

Resources relevant to your paper include;

- Doing a broader survey of the literature
- The instructor
- The faculty member who suggested the paper
- The paper's authors

Resources: Literature review (again)

Stat papers are typically written for an audience familiar with the relevant literature;

- If you are new to an area, expect to beef up your background before you can really get started
- Follow the reference trail in your paper; read the references
- Try textbooks, online tutorials, and review articles
- Perform a literature review; enter well-chosen search terms into PubMed/Web of Science, examine everything it spits out

Resources: Instructor

- I expect to read all the 572B papers – eventually – and am happy to chat with you about your paper in broad terms
- Also happy to talk about the expectations for this class, i.e. what sort of empirical and analytical results make sense given your paper
- Stop by my office hours, or email me

But: I am not an expert on each of your papers, and so cannot answer detailed questions about your paper. (Perhaps not quickly, perhaps not at all)

Resources: Other faculty

Faculty member who suggested the paper should have expertise in the area, if not specific familiarity with the paper.

- Make use of this resource when you have technical questions about the paper that you can't solve on your own
- But, do your homework before seeking faculty help
- Use but don't abuse;
  - Faculty time is valuable, and they are volunteering it to help you with this paper
  - Consider the possible use of this resource in making a paper selection!

Resources: Authors

You may need to e-mail the authors;

- To request software, or example datasets
- To ask for clarification about a simulation set-up
- To ask if there is an error in the paper

You should **minimize** the use of this resource – don't even think about emailing the authors before exhausting all other possible resources.

For example; first check CRAN, the authors' website(s), github, Cross-validated, me, other faculty and Google for code. Then try to implement it yourself. Then (maybe) email the authors.

Resources: Authors

If/when you do email;

- Do your homework first; faculty are busy people
- Sending a professional e-mail will increase your chance of getting a response;
  - State who you are—and use your UW e-mail address
  - Specify the paper
  - State a specific question
  - Sign off graciously
- Do **not**;
  - Send an open-ended question
  - Be vague
  - Include spam-like attachments, figures, or text

If you get any response, act appreciative – send a quick thank you; this will help with any future email!

If you don't hear back in 7 days: send a gentle reminder. If that doesn't work, drop it and let me know.

Resources: Authors

To: xavier@royalroads.edu
Re: Request for code, "Uniformly Unbiased Ultimate U-statistics"

Dear Professor X,

My name is John Smith, I am a 2nd year PhD student at University of Washington in the Department of Biostatistics.

I have read with interest your 2014 JASA paper "Uniformly Unbiased Ultimate U-statistics". I would like to apply the UUUU method as part of a course project, but have been unable to find software for UUUU online.

Would it be possible for you to send me a software implementation for UUUU, in particular the code used for Table 2 in your paper?

Thank you in advance

John Smith

Resources: Authors

```
To: xavier@royalroads.edu
Re: Request for code, "Uniformly Unbiased Ultimate U-statistics"

Dear Professor X,

My name is John Smith, I am a 2nd year PhD student at University of Washington
in the Department of Biostatistics. I have read with interest your 2014 JASA
paper "Uniformly Unbiased Ultimate U-statistics".

As part of a class project, I am trying to reproduce the results in your paper,
using the UUUU package available on CRAN. (I am using v2.17.3)

Unfortunately, I am unable to reproduce the results in Table 2 on pg 1839, so I
am writing to ask if you could answer these questions about Table 2;

1. When running UUUU, what value of "uuparam" parameter was used?
2. In Equation 10, you mention that each observation is drawn from
$N(0,\sigma^2)$, but do not specify $\sigma$. What value of $\sigma$ was used?

Thank you in advance

John Smith
```

Resources: Authors

Prior to your email, most researchers have had frustrating experiences, trying to help confused emailers, and expending lots of time for little gain;

- Research time is valuable, so they are wary about getting sucked into this again
- An e-mail with a clear and coherent request, and that indicates thought on the part of the sender, suggests fewer issues later
- An open-ended, unclear, or disorganized e-mail that indicates little thought on the part of the sender is a red flag. Authors may quite reasonably just ignore such emails

If requesting data, also be aware researchers may not be able to send it, due to HIPAA and related concerns.