ST 790 Navigating the PhD program and beyond: perspectives, skills, and strategies

Jonathan P Williams

North Carolina State University

Fall 2024

Transitioning from coursework to research

After passing the qualifying exam:

- $\rightarrow$ Developing your dissertation research is the most important aspect of your graduate studies
- $\rightarrow$ Insofar as receiving passing grades, courses are no longer the highest priority
- $\rightarrow$ Future employers will evaluate you based on the quality of your dissertation research

Note:

While many people with a PhD degree in statistics are choosing to work in industry, the purpose of a PhD degree in statistics is to train you as a researcher

A PhD is *not* a professional degree (e.g., Medical Doctor)

Transitioning from coursework to research

Timeline of next steps:

1. Narrow down your areas/types of potential research interest
   $\rightarrow$ Will overview areas later

2. Find 1-2 PhD advisors
   $\rightarrow$ Begin working on a first project
   $\rightarrow$ Might spend 6-12 months on background reading

3. Schedule written preliminary exam
   $\rightarrow$ Within $\approx$ 18 months of beginning research
   $\rightarrow$ Assemble your PhD committee
   $\rightarrow$ $\approx$ 5 faculty members, mostly from your department
   $\rightarrow$ Your advisor(s) are your PhD committee chair(s)

Transitioning from coursework to research

Timeline of next steps (continued):

4. Complete $\approx 75\%$ of dissertation research
   - $\rightarrow$ Schedule oral preliminary exam with committee
   - $\rightarrow$ Present what you have already accomplished
   - $\rightarrow$ Propose what the remaining 25% will look like

5. Complete $\approx 99\%$ of dissertation research
   - $\rightarrow$ Schedule oral final defense with committee
   - $\rightarrow$ Present your dissertation work
   - $\rightarrow$ Argue it is substantial enough to earn your PhD degree

6. Submit your dissertation manuscript to the university
   - $\rightarrow$ Ask senior students for the university-compliant .tex file

Transitioning from coursework to research

Types of statistics research:

- → Theoretical or mathematical statistics
- → Machine learning or statistical learning
- → Statistics methodology
- → Applied statistics
- → Computational statistics
- → Statistical software

Note:

This list does not include statistical applications or collaborative research published in domain science journals

Transitioning from coursework to research

Theoretical or mathematical statistics:

- → Investigations of theoretical or mathematical properties of estimators or computational tools
- → Formulations/justifications for a paradigm of statistical inference. E.g., frequentist, Bayesian, fiducial
- → etc.
- → No immediate applications necessary

Top journals include:

Annals of Statistics (AoS)
Bernoulli

Transitioning from coursework to research

Machine learning or statistical learning:

- $\rightarrow$ Use data to train algorithms to perform tasks
- $\rightarrow$ Particular emphasis on prediction problems/tasks
- $\rightarrow$ Algorithm development
- $\rightarrow$ Theoretical and empirical performance metrics/evalaution
- $\rightarrow$ Unsupervised learning

Top journals include:

Journal of Machine Learning Research (JMLR)
Many prestigious conference proceedings (e.g., NeurIPS, ICML)

Transitioning from coursework to research

Statistics methodology (most common type):

- $\rightarrow$ Propose a new estimator/approach for making inference on population quantity of interest
- $\rightarrow$ Simulation study to investigate empirical properties of the proposed method
- $\rightarrow$ Formulate and prove theorems to guarantee consistency or other optimality properties of the proposed method, under certain assumptions
- $\rightarrow$ "Real data" implementations and proof of concept

Top journals include:

Journal of the Royal Statistical Society: Series B (JRSS B)
Journal of the American Stat Assoc: Theory and Methods
Biometrika

Transitioning from coursework to research

Applied statistics:

- → Method development/evaluation motivated by a real data set and/or questions of interest with considerable practical relevance in some application
- → Not necessarily methodologically novel
- → Illustration of important aspects of existing methods
- → Important case studies or comparisons

Top journals include:

Journal of the American Stat Assoc: Appl and Case Studies
Annals of Applied Statistics (AoAS)
Journal of the Royal Statistical Society: Series C (JRSS C)

Transitioning from coursework to research

Computational statistics:

- $\rightarrow$ Algorithms for implementation of estimation routines
- $\rightarrow$ Issues relating to computational efficiency versus statistical efficiency
- $\rightarrow$ Theoretical properties of algorithmic convergence

Top journals include:

Journal of Computational and Graphical Statistics (JCGS)

Transitioning from coursework to research

Statistical software:

- → R package development
- → Open-source statistical software development, more generally
- → Demonstration/comparison of existing software

Top journals include:

Journal of Statistical Software

Transitioning from coursework to research

Areas of statistics research:

. . . very many.

Here are the "major areas" of research in our department:

`https://statistics.sciences.ncsu.edu/research/research-areas/`

Transitioning from coursework to research

Things to consider in choosing an advisor:

- → Type/area of research focus
  - → But be careful not to overemphasize this one...
- → Personal compatibility
  - → It is difficult to work with someone that you find difficult to interact with
  - → You'll meet $\approx$ weekly for the next 4 years
  - → You'll eventually need a strong letter of recommendation from them; so it's important they like you, as well

Transitioning from coursework to research

Things to consider in choosing an advisor (continued):

- → Their work ethic and intensity of expectations
  - → If you only want to work 30-40 hours per week, then you're never going to impress your advisor if she/he works around the clock
  - → Look for an advisor with a likeminded attitude about work–life balance
- → Feedback from current advisees
  - → So long as $n > 1$, this is perhaps the best calibrated source of information for a glimpse into what your experience with a potential advisor might be like

Transitioning from coursework to research

Things to consider in choosing an advisor (continued):

- → Advisor's network
  - → Do their students tend to get jobs in careers you are aiming for?
  - → Some faculty send almost all students to industry
  - → Some have better connections in academia or industry
- → Resources available from the potential advisor
  - → Can they fund you as an RA?
  - → Do they have funds for you to travel to present your research?
  - → Do they work with collaborators in domain sciences of interest to you?

Transitioning from coursework to research

Things to consider in choosing an advisor (continued):

- → Amount of interaction you need
  - → Some advisors meet with each student for 30 min/week
  - → Some advisors are willing to meet 4-5 hours/week
  - → In part, depends on how many other students are advised
    - → The number of students a faculty member chooses to advise in a given year gives an indication of how carefully they choose to think about research problems
    - → Also indicates how active the faculty member is

Transitioning from coursework to research

Things to consider in choosing an advisor (continued):

- → You are exclusively your own best advocate for you
  - → Don't expect that your advisor will make you aware of all that you need to be aware of
  - → Don't expect your advisor to always be correct
  - → Don't expect your advisor to always know best
  - → But you need to be able to trust their judgement
  - → Your advisor is as human as you are, proceed as such

Transitioning from coursework to research

Things to consider in choosing **to be an adult**:

- → Whatever choices you make:
  - → Sometimes you will have to work more hours in a day/week/month/year/etc. than you want to
  - → Oftentimes you will have to do work you don't want to
  - → Your work should be about more than how it benefits you; we live in a society
  - → Aiming for purpose, satisfaction, and fulfillment is more sustainable than aiming to feel happy, on any given day

Monte Carlo simulation studies versus mathematical proofs

A typical framework for statistical research is as follows.

1. Begins with a population and questions of interest
2. Population features are formulated and quantified in relation to the questions of interest
3. Data relevant to the population features of interest are collected
4. Statistics (i.e., functions of the data) are formulated to use the data to make inference on the the population features of interest in a manner that is optimal in some way
   $\rightarrow$ e.g., least biased, most efficient, most powerful, most consistent, etc.

Monte Carlo simulation studies versus mathematical proofs

Research might be done to choose or formulate an estimator

As a research statistician, much of the work is to establish the properties of the chosen/formulated estimator

This work can be approached in a few ways:

$\rightarrow$ Gold standard: properties established by mathematical proof

$\rightarrow$ Simulation studies:

- $\rightarrow$ Helps to develop intuition for proofs
- $\rightarrow$ Drives intuition for reformulating/adjusting estimator
- $\rightarrow$ Can be used if proof is too complicated
- $\rightarrow$ Support arguments used in proof
- $\rightarrow$ To demonstrates concepts or strange phenomena

Monte Carlo simulation studies versus mathematical proofs

Consider a simple example:

$\rightarrow$ Population of measurements $\sim$ normal$(\mu, 1)$
$\rightarrow$ Unknown population feature μ
$\rightarrow$ Perhaps use a sample mean or median to make inference on μ

What are the properties of the sample mean $\overline{X}_n$ for estimating μ?

What are the properties of the sample median M_n for estimating μ?

Monte Carlo simulation studies versus mathematical proofs

Theorem

The same mean of iid normal$(\mu, 1)$ data follows the normal$(\mu, 1/n)$ distribution.

Proof. If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{normal}(\mu, 1)$, then each X_i/n has a moment generating function of the form $m_{X_i/n}(t) = e^{\mu t/n + (t/n)^2/2}$. By independence,

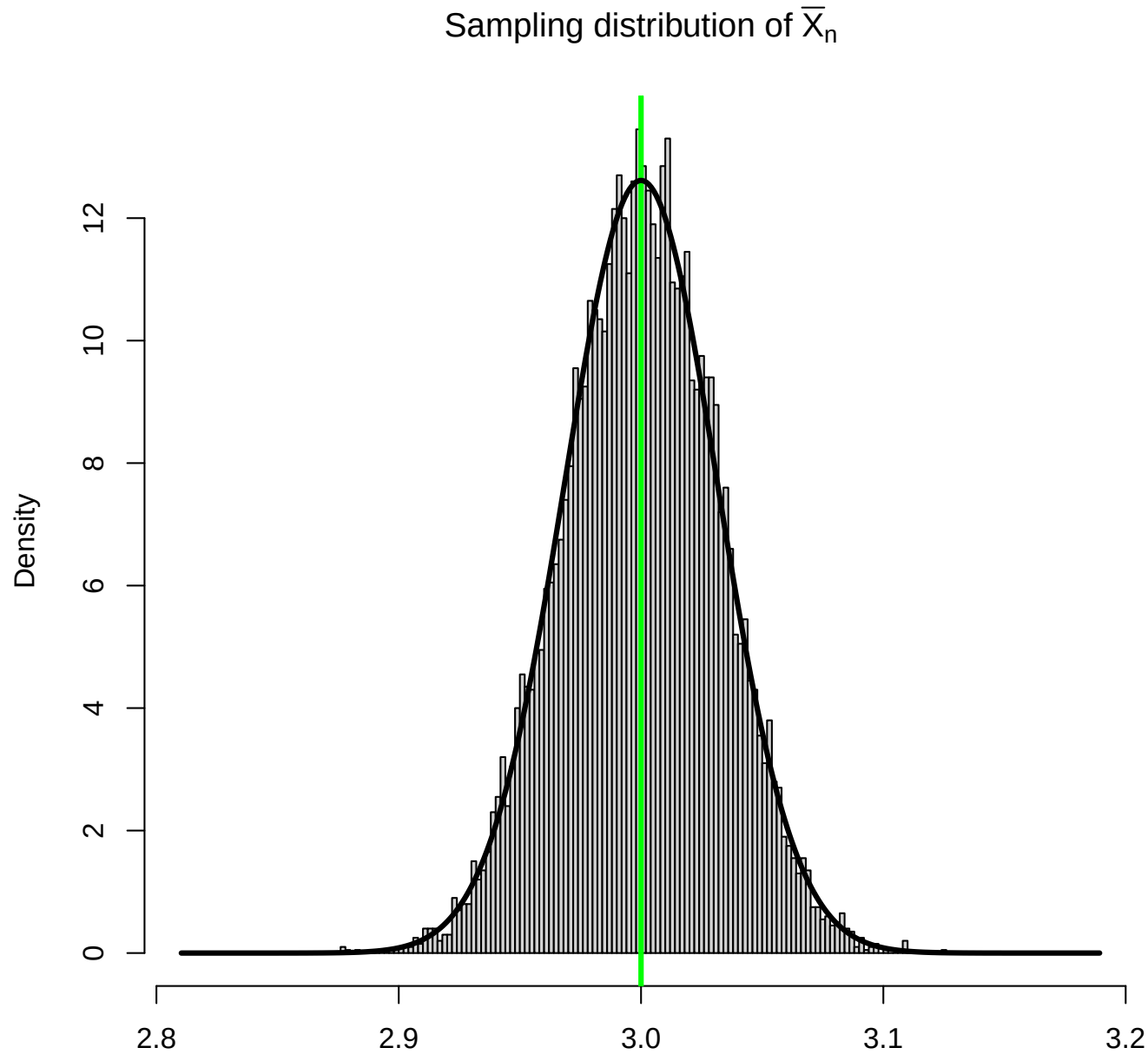
$$m_{\overline{X}_n}(t) = \prod m_{X_i/n}(t) = e^{\mu t + (1/n)t^2/2},$$

so that $\overline{X}_n \sim \text{normal}(\mu, 1/n)$. ■

Monte Carlo simulation studies versus mathematical proofs

```r
library(latex2exp)

mu = 3
sigma = 1
n = 30

# Simulate a large number of data sets and least squares estimators
num_sims = 300
x_bar_vec = rep( NA, n=num_sims)
for(k in 1:num_sims){
  y = rnorm( n, mean=mu, sd=sigma)
  x_bar_vec[k] = mean(y)
}

# Plot the sampling distributions of the estimator
upper = mu + 6*sigma/sqrt(n)
lower = mu - 6*sigma/sqrt(n)
grid = seq( lower, upper, by=.01)

hist( x_bar_vec, freq=F, main=TeX(r'(Sampling distribution of $\bar{X}_{n}$)'),
      xlab=NULL, xlim=c(lower,upper), breaks=floor(sqrt(num_sims)))
abline( v=mu, col="green", lwd=3)

lines(grid, dnorm( grid, mean=mu, sd=sigma/sqrt(n)), lwd=3)
```

Monte Carlo simulation studies versus mathematical proofs



Scientific writing: general principles

Rough outline of a typical statistics publication:

- Section 1. Introduction
- Section 2. Methods
  - Subsection 2.1. Algorithms
- Section 3. Theoretical results
  - Subsection 3.1. Proofs
- Section 4. Empirical results
  - Subsection 4.1. Numerical illustrations
  - Subsection 4.2. Simulation studies
- Section 5. Real data analyses
- Section 6. Concluding remarks and future work
- Appendix A. Additional proofs
- Appendix B. Additional figures, tables, algorithms etc.

Scientific writing: general principles

Things to consider when writing a title and abstract

Scientific writing: general principles

Link to TeX: `https://en.wikipedia.org/wiki/TeX`

Link to Overleaf: `https://www.overleaf.com/`

Scientific writing: general principles

The role of mathematical notation in writing about mathematical and statistical ideas

`https://jonathanpw.github.io/ST790/Marron1999.pdf`

Scientific writing (and reading): literature reviews

Generally, 4 levels of depth to reading a statistics research article:

1. Title + abstract
2. Title + abstract + introduction
3. Full manuscript
4. Full manuscript + appendices + proof details

Scientific writing (and reading): literature reviews

How to approach learning about new topics?

→ Usually start with a key reference(s) from your advisor, a colleague, a collaborator, etc.

→ Forward and backward search of key articles

→ Keyword search in a repository (e.g., Google scholar)

→ Decide on the reliability of a found article:

- → Do the authors have established credibility on the topic?
- → Is the article published in a relevant journal?
- → Should you trust preprints less then publications?

Scientific writing (and reading): literature reviews

How to approach learning about new topics? (continued)

- → Reach out to authors
  - → Quick questions over email
  - → Non-quick questions over Zoom or meet for a coffee, e.g., at a conference if non-local.
  - → Most serious researchers enjoy having conversations about their work; I'm happy to talk about my work if anyone wants to come by my office
- → Find good literature review articles; usually titled:
  - → "Survey of . . . ", "Primer on . . . ", "Tutorial on . . . ", etc.
  - → Journal of the American Statistical Association: Reviews
  - → Statistical Science

Scientific writing (and reading): literature reviews

arXiv challenge

Start every workday by scrolling through all new submissions appearing in the Statistics topic section of arXiv:

`https://arxiv.org/list/stat/new`

$\rightarrow$ There are typically $\approx 30 - 40$ new articles each day

Scientific writing (and reading): literature reviews

arXiv challenge

→ As you scroll, read each article title and author list

→ Skim the abstract if:

- → the title sounds interesting
- → it's an author that you tend to appreciate

→ Of the abstracts read, if is compelling enough, open article:

- → maybe it's on a topic of interest
- → maybe it's relevant for a current/future literature review
- → maybe it's on a topic that you hadn't heard of

→ Of the articles opened, decide how much of them to read

- → recall the previously discussed levels of reading depth

Scientific writing (and reading): literature reviews

Things to consider in writing a literature review:

$\rightarrow$ How broad is the audience?

$\rightarrow$ Trying to establish credibility in an area?

$\rightarrow$ Trying to establish relevance of an idea?

$\rightarrow$ Trying to be informative?

$\rightarrow$ Does it have to be exhaustive?

$\rightarrow$ Scope versus depth of each article discussed in the literature review

Scientific writing (and reading): literature reviews

(assignment)

Scientific writing: academic publishing and the purpose of journals

Understanding the purpose of journals
How to navigate the peer-review process
Description of an editorial board. Editors, Associate Editors, Reviewers
Timeline of the publication process

Scientific writing: academic publishing and the purpose of journals

Open source research practices

Surveying articles on `https://arxiv.org`
Preprint repositories such as arXiv
Roles and purposes of preprints
Open source code repositories such as GitHub Making research papers and code available on personal academic websites
How these practices are tied with the reach and impact of your paper ASSIGNMENT: create your own website