Dr. Tristan Whalen
OPRE/BUAN 6359.008 Fall 2023
Advanced Statistics for Data Science

The country is hungry for information; everything of a statistical character, or even a statistical appearance, is taken up with an eagerness that is almost pathetic; the community have not yet learned to be half skeptical and critical enough in respect to such statements.

—General Francis A. Walker, Superintendent of the 1870 census
(Quoted in Freedman, Pisani, and Purves *Statistics,* 4ed)

INSTRUCTOR EMAIL & OFFICE

Dr. Tristan Whalen tristan.whalen@utdallas.edu JSOM 3.420

I encourage you to email me throughout the semester with questions or concerns about the class. Please include **your course and section, 6359.008** in every email to me.

My Fall 2023 office hours at JSOM 3.420 (through December 7):

Mondays: Noon to 3:00pm (no appointment needed)
Tuesdays by appointment only
Wednesdays: 4:45pm to 6:00pm (no appointment needed)
Thursdays/Fridays by appointment
No office hours on holidays or school closures.

CLASSROOM LOCATION & TIMES

JSOM 1.107 Wednesdays 1:00pm – 3:45pm

CALENDAR & HOLIDAYS

Last day of classes: Thursday, Dec 7 Withdrawal dates: Sept 6 (no record), Nov 7 (withdrawal)
Thanksgiving Break: November 20-26

HARDWARE & SOFTWARE

>Regular and reliable access to **elearning**.
>**Microsoft Word** or similar software to compose solutions to the mini projects and paste the visuals created in R.
>We use **R** throughout the course. R is a powerful statistics and data software that is free (monetarily) to download and install: https://cran.r-project.org/
>After installing R, you will probably want to download and install **RStudio** Desktop. RStudio provides a more user-friendly interface on an already-installed R: https://posit.co/download/rstudio-desktop/
>You are not required to bring a laptop to class.

 # PREREQUISITES #

>Absolutely <u>NO</u> prior experience with R

We begin R "from scratch." With that said, I assume you have experience downloading, installing, and using software, typing, preparing documents, saving and loading files in folders on your computer, using elearning, copying and pasting graphs into a document, and things of that sort.

>Proficiency with routine mathematics.

Arithmetic with sums, products, powers, roots, fractions, formulas. Plotting and reading graphs. Reading tables of numbers. Equalities, inequalities, intervals. Functions.

>Acquainted with the ideas and concepts of calculus.

Primarily the concepts of "area under a curve," derivatives, and limits; hopefully you have seen The Fundamental Theorem of Calculus. Multivariable calculus will help (but is not necessary) for multiple regression. You will not have to do any calculus by hand—R will do that for you.

>Familiarity with undergraduate probability and statistics.

We cover all the following topics, but it will help if you already know about them: averages, standard deviations, histograms, medians, and boxplots. Basic probability calculations with multiplication, addition, and complements. (Think dice, coins, marbles, and cards.) Mutual exclusiveness and independence. Random variables and probability distributions, including some named distributions like the Binomial distributions and the Normal distributions.

 # TEXTBOOKS #

>*The Statistical Sleuth* (Ramsey, Schafer, 3rd edition). "Sleuth"

I don't like it and disagree with the authors on many topics. But some of the diagrams explain the statistical theory well, and every chapter includes examples and exercises with data already built-in to an R package. The methods they apply are not always valid on the data set, however.

>*The Book of R* (Davies, 1st edition) "Book of R"

Davies begins with the basics of R and builds up to the advanced statistical methods of this course. Start from the beginning, then start skipping ahead once you feel confident.

>*Statistics*, Freedman, Pisani, Purves, 4th edition. "Freedman"

The only statistics textbook I like. Explains the fundamentals better than any other book I know. Discusses deep concepts (inference) better than any graduate book I've seen. The only downside is that the book does not cover every required topic.

 # TOPIC LIST #

"Education never ends, Watson. It is a series of lessons with the greatest for the last."

—Sherlock Holmes (from Arthur Conan Doyle's *The Red Circle*)

#	Topic	Additional References
01	Experiments vs. Observations *RCTs, blinding, observational studies, confounding factors*	Freedman Ch.01-02
02	Shape, Center, and Spread *Histograms, the average of a list, the SD of a list, z-scores, empirical rule, Cheybshev inequality, quantiles*	Freedman Ch.03-04
03	Normal Approximation to Data *Properties of normal curves, compare to data histograms, qq-plots, skew, kurtosis*	Freedman Ch.05
04	Analyzing Variance with Sums of Squares! *Descriptive ANOVA: within-group, between-group, and total sum of squares; coefficient of determination; comparing one-average and multiple-average summaries*	Sleuth Ch.05
05	Scatterplots, Correlation, and Regression *Summarizing the relationship (or lack thereof) between two variables, using and interpreting a regression line, regression effect, warning about extrapolation, association is not causation*	Freedman Ch.08-10
06	Residuals and the Least Squares Method *RMS-error, coefficient of determination, log transforms, regression planes, interpreting coefficients, residual plots*	Freedman Ch.11-12
07	Multiple Regression Model Construction *Indicators, interactions, squared terms, variable selection*	Sleuth Ch.09-12
08	Probability: from Basics to Bayes' Rule *Interpretations, multiplication principle and independence, addition rule and mutual exclusiveness, complement rule; conditional probability, Bayes' rule*	Freedman Ch.13-14 Whalen 3341 lessons 01-02
09	Random Variables *Probability distributions and histograms, expected value and standard error*	Freedman Ch.15-17 Whalen 3341 lessons 03-04
10	Normal Approximation to Probability *The Central Limit Theorem (CLT)*	Freedman Ch.18 Whalen 3341 lesson 10
11	Classic Statistical Inference *Random sampling, z-intervals, z-tests, Student t-tests*	Freedman Ch.19-21, 23-24, 26 Whalen 3341 lessons 11-12
12	*F*-tests and Friends *Two-sample z-test, two-sample t-test, F-test (ANOVA)*	Sleuth Ch.02-03, 05
13	Inference from Regression on Random Samples *Standard errors and t-statistics for coefficients, F-statistic for entire regression summary, whether they apply*	Sleuth Ch.07-12

 # ABOUT THE ORDER OF TOPICS #

Lesson 01 establishes the difference between descriptive and inferential statistics.

Lessons 02-07 are about descriptive statistics and apply to practically any data set.

- *We do not distinguish between population and sample until after probability. In particular, we use the "SD of a list," defined to be the root-mean-square deviation from average of the list. These are mathematically and practically correct for descriptive statistics.*
- *We include ANOVA and regression before probability for theoretical, practical, and pedagogical reasons—*
    - *Theoretical: sums of squares and the least squares method (with all the related calculations) do not require any probability assumptions; it's just geometry and calculus.*
    - *Practical: regression models give useful insight into observational data and powerful descriptive summaries that don't have anything to do with probability.*
    - *Pedagogical: there's plenty to cover in regression before getting bogged down with probability.*

Lessons 08-10 are about probability, necessary for understanding statistical inference.

- *We adhere to this convention: expected value and standard error apply to random variables. By contrast, average and standard deviation apply to lists of numbers (data).*
- *The relationship: take a list of numbers. Draw one at random. The average of the list equals the expected value of the number you will draw, and the SD of the list equals the SE of the number you will draw.*

Lessons 11-13 are about inferential statistics and apply only to data gathered by a probability method.

COURSEBOOK DESCRIPTION

This course uses statistical methods to analyze data from observational studies and experimental designs to communicate results to a business audience. The course mandates prior knowledge of fundamental statistical concepts such as measures of central location, standard deviations, histograms, the normal and t-distributions* (knowledge of calculus is not required). The course also emphasizes interpretation and inference, as well as computation using a statistical software package such as R or STATA.

**Actually, we will introduce t-distributions as "new" to this course. Coursebook needs updating.*

 # ADVICE FOR SUCCESS #

>Attend every class and sit as close to the front as possible.
This is true for any college course, regardless of the presentation skill of the instructor. If you care about your grade and GPA, at a minimum you should never skip class. Attend every session, put away your smartphone, and actively participate (even just listening). These actions train your brain to take the course seriously, and they encourage the presenter, thus improving the quality of the class.

>Build time into your weekly schedule for the course outside of class time.
This is true for any college course. Set aside at least 3 hours per week outside of class session, depending on your skill level. As a starting point, I suggest 1.5 hours per week for reading the textbooks, 1.5 hours per week for reviewing notes and practicing R, and 2.5 hours per week for working the homework. You can adjust based on your individual abilities. If study and practice are not planned into your schedule, then they will not happen, and the exams will be very hard.

>Email the instructor early and often.
Do not let your first email to the instructor be about a low grade, and certainly do not let your first contact with the instructor be after the last exam. Instead, email the instructor as soon as you have any question, concern, or confusion. Make your first email about a course topic and not about a grade or what will be on a test.

>Make emails useful and courteous.
Read Dr. Whalen's email guide.

>Expect to make mistakes, get frustrated, and try repeatedly.
We learn as much when things go wrong as when they go right. ("Get messy! Make mistakes!" as Ms. Frizzle says.) All of us (students and instructors) will encounter errors and mistakes both in and out of R.

>This is not only about using R.
There is plenty of calculation, but R does all the tedious work in a flash. More importantly, we check whether the calculation is valid and decide how to interpret it. Many questions on homework and exams address concepts, such as whether the statistical methods are valid, or how to interpret the output. Beware of getting lost while looking through R tutorials.

>Take care of yourself.
This is true for any work. Your physical, mental, emotional, and spiritual health are interrelated. Get plenty of sleep each night on a regular schedule. Maintain a regular diet. Limit your time in social media. (I guarantee that, if you go at least 4 weeks without using any social media, you will notice an improvement in your physical and mental wellbeing. How often do you come away from social media sessions feeling refreshed, rejuvenated, and enlightened?) Spend time with friends and/or family in person. Engage in wholesome activities and entertainment.

GRADE CALCUATION

25% Online homework average	A	$X \geq 93\%$
10% Histogram mini project	A-	$90\% \leq X < 93\%$
10% Scatterplot mini project	B+	$87\% \leq X < 90\%$
10% Probability Simulation mini project	B	$83\% \leq X < 87\%$
15% In-Class Exam 1 (Regression Topics)	B-	$80\% \leq X < 83\%$
15% In-Class Exam 2 (CLT & Inference Topics)	C+	$75\% \leq X < 80\%$
15% Online Exam 3 (F-test & ANOVA)	C	$70\% \leq X < 75\%$
	F	$X < 70$

Online Homework: Sets of problems to be completed and submitted via elearning, usually within about one week from the date assigned. These may require the use of R.

Mini Projects: Each project is one long homework problem (usually about a single data set) with multiple parts. You will submit a one-page document that includes a nice plot or graph that you constructed in R, along with your R script. Each individual student must submit his or her own work.

In-Class Exams: Paper-based and given in class on assigned days. The in-class exams focus more on concepts than calculations (any R output will be provided on the exam). Closed notes and closed book. Questions may be multiple-choice or free response. Only a handheld calculator is required.

Online exam: Take online in elearning, at home during the assigned time frame. Timed (around 60 minutes), but open notes and open book. Requires the use of R.

Do you curve grades? When we finish grading an exam, I always look at the score distribution and summary. If the median or average are low (depending on the course), I always adjust the scores by a constant shift in the positive direction. I do not call this "curving." This is a statistics class, so note two things: if you do ask about curves, you ought to specify what curve you have in mind (there are lots: bell curves, exponential curves, logarithmic curves, Pareto curves,…). Also, you don't *really* want me to *curve* grades, because this might imply lowering your score, along with some others, to get the histogram to fit under a curve.

Is there extra credit? I build some bonus points into most of my assignments and exams, so make sure to complete all the work. Otherwise, please do not ask for extra credit, bonus projects, second attempts, and the like. These requests will be ignored to ensure fairness to the whole class. Instead, do the assigned work.

What if I have to miss an exam? Notify the instructor *immediately* if an emergency prevents you from taking an exam. We can usually work out a make-up exam with prompt notice. See more policies below.

IN-CLASS EXAM POLICIES

>Come to class regularly for announcements. You are responsible for missing an announcement made in class. Also check your UTD email regularly.
>Dates of exams will be announced in class at least the week prior.
>Take each exam during the allotted class time.
>Bring pencils, a handheld calculator, and your CometCard.
>Closed notes, closed book, closed everything, except your mind.
>The exam and scratch paper will be provided.

ONLINE EXAM POLICIES

>Come to class regularly for announcements. You are responsible for missing an announcement made in class. Also check your UTD email regularly. (This must be important.)
>The date of the exam will be announced in class.
>Take the online exam during the specified day at home in elearning.
>Open notes and open book, but you are expected to prepare in advance.
>Individual work only. No collaboration with anyone else. No communication about the quiz with anyone before, during, or after, except the instructor or the TA.

MAKE UP POLICIES

>If an emergency (such as you wake up sick) prevents you from taking an exam, notify the instructor immediately (usually an email).
>No make-ups will be offered for your own personal matters (such as family trips or weddings) or other non-emergency reasons.
>If you miss an exam and do not contact the instructor before the end of the exam day, you will receive a grade of zero for the missed quiz or exam.

HOMEWORK POLICIES

>Come to class regularly for announcements. You are responsible for missing an announcement made in class. Also check UTD email regularly. (Look familiar?)
>Homework will be posted and announced at least one week prior to the due date.
>Submit all homework in elearning. You must complete all assigned work.
>No make-ups or extensions of homework for any reason.
>Submit only your own work.
>Evidence of academic dishonesty (such as copying work, sharing or getting exam material, or communicating about the exam in online platforms) will be referred to the appropriate dean's office. If you are found guilty, as far as this course is concerned, you will receive a score of zero on your submission.
>Read https://www.utdallas.edu/conduct/dishonesty
>Also see the student code of conduct: https://policy.utdallas.edu/utdsp5003

CLASSROOM POLICIES

>In general, please treat others the way you want them to treat you.
>Please show courtesy and charity to other students and the instructor. Focus on the lecture, put away smartphones and other devices, and raise your hand to contribute.
>Avoid leaving early and avoid arriving late.
>At the instructor's discretion, you may be asked to leave the classroom and/or receive a grade penalty for behavior that interferes with class.
>You are expected to attend class and participate regularly.
>Skipping class and not participating will bring you lots of stress and difficulties. Instead, always come to class, sit as close to the front as possible, and participate regularly (in class, in email, or in office hours).

Various lawyers have suggested that we mention that the content of this syllabus may change at the instructor's discretion.

GENERAL SYLLABUS STUFF

>>University restrictions about class recordings

Students are expected to follow appropriate University policies and maintain the security of passwords used to access recorded lectures. Unless the Office of Student AccessAbility has approved the student to record the instruction, **students are expressly prohibited from recording any part of this course**. The instructor's recordings may not be published, reproduced, or shared with those not in the class, or uploaded to other online environments except to implement an approved Office of Student AccessAbility accommodation. If the instructor or a UTD school/department/office plans any other uses for the recordings, consent of the students identifiable in the recordings is required prior to such use unless an exception is allowed by law. Failure to comply with these University requirements is a violation of the Student Code of Conduct.

>>University restrictions about class materials (lecture note files, instructor notes, solutions, etc.)

The materials posted by the instructor may be downloaded during the course; however, **these materials are for registered students' use only**. Classroom materials may not be reproduced or shared with anyone not in the class, or uploaded to other online environments except to implement an approved Office of Student AccessAbility accommodation. Failure to comply with these University requirements is a violation of the Student Code of Conduct.

>>University technical requirements and help

In addition to a confident level of computer and Internet literacy, certain minimum technical requirements must be met to enable a successful learning experience. Please review the important technical requirements on the Getting Started with eLearning webpage: https://ets.utdallas.edu/elearning/students/current/getting-started

This course can be accessed using your UT Dallas NetID account on the eLearning website. Please see the course access and navigation section of the Getting Started with eLearning webpage for more information. To become familiar with the eLearning tool, please see the Student eLearning Tutorials webpage.

UT Dallas provides eLearning technical support 24 hours a day, 7 days a week. The eLearning Support Center includes a toll-free telephone number for immediate assistance (1-866-588-3192), email request service, and an online chat service.

The University is committed to providing a reliable learning management system to all users. However, in the event of any unexpected server outage or any unusual technical difficulty which prevents students from completing a time sensitive assessment activity, the instructor will provide an appropriate accommodation based on the situation. Students should immediately report any problems to the instructor and contact the online eLearning Help Desk. The instructor and the eLearning Help Desk will work with the student to resolve any issues at the earliest possible time.

>>UT Dallas syllabus policies and procedures:

The information contained in the following link constitutes the university's policies and procedures segment of the course syllabus.

https://go.utdallas.edu/syllabus-policies

>>Comet Creed

This creed was voted on by the UT Dallas student body in 2014. It is a standard that Comets choose to live by and encourage others to do the same:

"As a Comet, I pledge honesty, integrity, and service in all that I do."

>>Academic Support Resources

The information contained in the following link lists the University's academic support resources for all students. Please go to Academic Support Resources webpage for these policies:
https://provost.utdallas.edu/syllabus-policies/#academic-support-resources

"That is the one eternal education: to be sure enough that something is true that you dare to tell it to a child."

—G. K. Chesterton (from the book *What's Wrong with the World*)