

## Course Syllabus: DS 6371 Statistical Foundations for Data Science

Course Designers:	<a href="#">Dr. Bivin Sadler</a> and <a href="#">Dr. Monnie McGee</a>
Course Text:	Ramsey, F. L., and D. W. Schafer. <i>The Statistical Sleuth: A Course in Methods of Data Analysis</i> , 3rd ed. Boston, MA: Brooks/Cole, 2013, with associated website <a href="http://www.statisticalsleuth.com">www.statisticalsleuth.com</a> .
Other Materials:	ChatGPT account ... it can be paid or free.
Prerequisites:	A previous introductory statistics course and Bridge to Statistics
Midterm Date	Saturday, March 2 <sup>nd</sup> 2024 from 11am to 2pm CST on Zoom
Final Exam Date	Saturday, April 20 <sup>th</sup> 2024 from 11am to 2pm CST on Zoom

**The text is available as an electronic version from CengageBrain.com and is much less expensive this way!**

**All elements of the syllabus are subject to change by the instructor.**

### **Before taking this class, you should know**

- Statistical methods from an introductory statistics course: appropriate use of the mean and median, interpretation of box plots and histograms, use of simple linear regression, probability calculations using the normal distribution, applications of the central limit theorem, interpretation and calculation of confidence intervals for a population mean, and interpretation and calculation of a  $p$ -value for any hypothesis test. If you would like a review of any or all of these concepts please check out the free Bridge to Statistics Course that should be in your list of available courses. If you are at all concerned about being prepared for this course ... take the Bridge Course ... it will help tremendously.

### **Learning Objectives: The student will**

- Gain a better understanding of basic statistical methods covered in introductory statistics courses and problems that arise when these analytic methods are applied to real-life research problems
- Demonstrate understanding of the advantages and disadvantages of a given experimental design, particularly with respect to the type of conclusions that can be made
- Appropriately apply the methods discussed in the course to numerical and categorical data
- Communicate the findings of a statistical analysis in a clear, concise, and scientific manner

### **Course Expectations**

- As a data scientist in training, it is important to practice independently researching questions that come up before asking your instructor. It's amazing what five minutes on the internet can turn up! Please, train yourself to try to find an answer for yourself before asking someone else. Give yourself a time limit (e.g., I will search for one hour, and if I can't find anything, I will ask). It's better for you in the long run!! Good resources are StackOverflow, CrossValidated, Medium, etc.... or ... start your own blog with the answers you find!
- We will use both SAS OnDemand and RStudio (Posit) in this course. SAS is built for statistics and is primarily used as a teaching tool since it provides numerous tables, plots and charts we

will need with very little code. While this is not primarily a coding course, SAS is still widely used in industry thus having some experience with SAS is a nice side benefit. Please visit [https://www.sas.com/en\\_us/software/on-demand-for-academics.html](https://www.sas.com/en_us/software/on-demand-for-academics.html) and check out this [short video](#) for direction on how to access and use SAS OnDemand. We will also RStudio (Posit) which can be downloaded here: <https://posit.co/download/rstudio-desktop/>.

- Play around with SAS and R. They are very powerful software programs. Asynchronous videos regarding the use of SAS and R are available in 2DS, but there is no way that they can cover every scenario. Please examine the help documentation (available online or within the software itself) before asking your instructor a question about the use of SAS and R. And don't forget Google, StackOverflow, Cross Validated, generative AI (but see below) and blogs!!! You will likely learn several other tidbits in addition to what you were looking for!
- Consume and participate in all asynchronous material (videos and exercises) for the course.
- It is critical to complete the For Live Session Assignment (FLS) thoroughly and on time. This course is designed around the idea that if we build the pieces before class, we can put them together in Live Session. During the completion of the weekly FLS students will identify the material they have mastered and the questions that they need to ask to put clear up any misunderstanding. Live Session is designed to then harness that work to put all the pieces together. Without the FLS completed, Live Session time will not be optimized and students will likely struggle (and have in the past). However, once the FLS is completed thoroughly and on time, professors will be able to customize your Lives Session to your questions and learning style and create a truly transformative educational experience.

## Course Coverage

The course will cover chapters 1–12 of the text. Some of the material may seem like a review to those who have had a previous course in statistics. That assessment may be somewhat true; however, we will pay more attention to sample size calculation and experimental design than a first course typically does. Furthermore, we will concentrate on understanding WHY a particular technique is appropriate and HOW to interpret the results. There is a lot in this course for everyone.

**Grading:** Midterm exam (25 percent, Points/Scale 0-100), final exam (25 percent, Points/Scale 0-100), homework (10 percent, Points/Scale 0-100), project (10 percent, Points/Scale 0-100), live session attendance (10 percent, Points/Scale 0-100), asynchronous video questions (Students Must Complete 100% of the asynch material on time (by the end of that unit) to pass the course), “for live session” questions (20 percent, Points/Scale 0-100)

**Homework (10 percent, Points/Scale 0-100):** Doing work and practicing the methods outside of class is critical to obtaining success in this course. The strategy of this course is to study the methods, identify questions, answer the questions and practice more to understand the methods on a deep level. To this end, there are homework problems on our GitHub site. The solutions to these problems are also on this site. The strategy is to do these problems on your own, check yourself against the solution and then address the TA and the professor with any questions that were generated from your study. This puts the onus on the student to make time each week to do the homework completely and frees up more time for the professor and the TA to address student's questions in a personal and meaningful way.

All data sets for case studies and exercises are available (or will be when assigned) on the 2DS website or the GitHub Repo. Datasets for problems assigned out of the book can also be found at the textbook website: <http://www.statisticalsleuth.com/>

**Project (10 percent, Points/Scale 0-100):** In lieu of typical homework for weeks 13 and 14, a group mini-project will be assigned.

**Live Session Attendance (10 percent, Points/Scale 0-100)** Note: **Participation** includes **attending every live session from start to finish** (the scheduled hour and a half). This includes sharing your video at all times to facilitate discussion and the full classroom setting. For a 1980s explanation of the dilemma not being physically present causes, see the following [link](#) from the 1980s cinema hit *Real Genius*! For those of you who have seen the movie, you will enjoy the whole montage! Those who have never seen it, you are in for a treat (watch the whole movie!). Please pay close attention to 2:45, 3:10, and 4:35.

**Asynchronous Videos and Questions (Must complete 100% on time to pass the course):** The course is structured with both synchronous and asynchronous sessions. In order to participate fully in the synchronous sessions, you must complete ALL of the material for the asynchronous sessions each week. This includes the concept check questions and activities that occur between videos and during the asynchronous material.

These are questions asked during the asynchronous videos that require an answer before going on to the next part of the video. Completion ***of all*** of the asynchronous video questions is required. In some open-ended questions, everyone can see each other's answers. This is meant to generate discussion and will be very helpful as a tool to help master the material!

**“For Live Session” Questions (20 percent, Points/Scale 0-100):** The “for live session” assignment is a set of questions and activities that the student is to work on and provide a response to ***before*** live session.

- The “for live session” assignment is due by noon the day of live session.
- It is important to note that the goal of the activities is to become familiar with the methods, ideas, and implementation involved in that activity so that we can efficiently iron out all the details in live sessions.
- Analogy: You are building the pieces of a puzzle in the “for live session” assignment, and we are putting them together to see the big picture in live session.
- It is ***not*** expected that the students have all the correct answers. The expectation is that each student spends the allotted time (indicated next to the activity) on each activity so that we can discuss the details in live session. In short, the requirement is to give genuine effort. If a student is stumped and works on a problem for the allotted time without much perceived success, that student should have several specific and well-written questions so that the professor can address those in live session.
- Again, if you max out the indicated time without finishing the activity and you don't have more time to finish, simply write up what you have learned by that time and **record the questions that were generated**, and we will address those in live session!
- We want to develop the questions before live session so that we can use the live session time to effectively answer any questions and master the material! (Build the pieces outside of live, session and put them together to see the big picture in live session and in the homework.)
- The FLS is critically important to learning in this class and must be completed thoroughly and in a timely manner in order to receive feedback before live session.

**Discussion Boards:** Students are not required to post in discussion boards, unless specified by the professor.

**Midterm Exam (25 percent, Points/Scale 0-100):** There will be a midterm exam in week 8 of the course. It will cover concept and hand-calculation questions, as well as a data analysis question. Please clear your schedule now! We will have a review for the exam during live session 8. The midterm exam will be held on a Saturday for three hours. It is a proctored zoom session.

**Final Exam (25 percent, Points/Scale 0-100):** A final exam will be given in week 15 of the course. It will be comprehensive, containing concept and hand-calculation questions, as well as a data analysis question. It is a cumulative exam and is similar in format to the midterm. We will have a review for the exam during live session 15. The final exam will be held on a Saturday for three hours. It is a proctored zoom session.

**Oral Exam: (Pass / Fail ... Pass is required to pass the course.):** There will be an oral exam at the end of the semester that covers all of the material in the course that is pass / fail. Students are given a list of 20-25 questions to study and 2 or 3 of these questions will be asked (with follow ups questions based on student responses) during the 20 minute exam. If a student does not pass the oral exam, they are given the opportunity to keep taking the exam until they pass it. There is a mandatory 2 day waiting time between retakes to allow for studying and growth. Students must pass the oral exam to pass the class.

**Rescheduling Exams:** Life happens. Should you need to reschedule an exam, please give notice to your live session instructor at least 24 hours prior to the live session in which the exam review is discussed (unit 8). The notice should be given via email. You and your instructor will discuss the best course of action given your circumstances. Retakes of exams will not be allowed, and a missed exam cannot be made up if notification is received AFTER the exam has taken place.

### **Flow of the Class (This will help with the what, when, and where!)**

1. Complete the asynchronous videos and concept-check questions for unit X. (The “X” is for the unit number.)
2. Get the 6371 For Live Session Unit X.pptx slide deck off from our GitHub repository.
3. This slide deck will have questions on some of the slides. These are the “for live session” questions that make up the “for live session” assignment. Complete the answers to these questions with at least one slide per question. The idea is that you will present some or all of these in a breakout session during the live session. Make sure to add at least four takeaways on one slide and any questions you have on the last slide.
4. Submit the slide deck to the **For Live Session Assignment: Unit X** assignment on 2DS by 1pm, Central Time, the day of your live session. This is the absolute latest that they can be turned in without a penalty. It will be counterproductive to work on the slides during or after the live session; therefore, any slides turned in after the deadline will receive a maximum score of 70. Turning in the slides by noon, Central Time, allows for the professor to review your slides before the live session in order to customize and tweak the live session to the particular questions each section has—which is good for the whole class!
5. Come and participate in live session! Fun!
7. Begin the process for the next unit (unit X + 1!).

### **Course Grading Policy**

This course consists of a number of assignments and projects that are to be completed throughout the term. It is expected that

all students will put forth the effort required to earn an 'A' letter grade for this course. Assignment grades will be determined using evaluation rubrics. You are responsible for reviewing the rubrics and raising questions or concerns related to the assignments, their rubrics, and their grading prior to the submission of each assignment. Questions regarding the grading of any assignments should be directed to the course instructor as soon as possible and in accordance with any regrading policy instituted by the instructor.

The final grade for the course will be calculated on the bases of the earned cumulative percentage and the grade received for each of the components of the cumulative percentage. This course is not graded on a curve. The required cumulative percentage needed to earn each letter grade is given in Table 1.

Table 1: Cumulative Percentage Required to Reach Each Letter Grade

Cumulative Percentage	Earned Grade
[100 – 93]	A
(93 – 90]	A-
(90 – 88]	B+
(88 – 83]	B
(83 – 80]	B-
(80 – 78]	C+
(78 – 73]	C
(73 – 70]	C-
(70 – 60]	D
< 60	F

The cumulative percentage for the course is determined by the course assignment components with their corresponding percentages defined in Table 2.

Table 2: Grade Components and Weightings of the Cumulative Percentage

Percentage of Cumulative Percentage	Component
Must complete 100% on time to pass the course.	Asynchronous Video Response Questions and Discussions ( <i>Must be completed before live session in the week assigned.</i> )
20%	For Live Session Assignments
10%	Homework
25%	Midterm
10%	Project
10%	Live Session Attendance
25%	Final Exam

### Use of Generative AI:

I strongly believe that generative AI will be a necessary tool we will all use in the future to many tasks in our jobs and in our lives in general. With that said, I also feel strongly that it is, by far, much more valuable if we are experts in what we are asking it to do. At this point it becomes a collaborator; otherwise, it is just dangerous.

**For this reason, output from generative AI should not be copied and pasted into any assignment (HW, FLS, Exam, etc.)**

However, after you have worked hard to understand the concepts in the asynch and FLS, we will cover the opportunities and challenges presented by generative AI relative to each unit in that unit's live session.

In summary, my goal for you is to understand these methods and concepts completely so that you can wield them and generative AI in a useful, knowledgeable, confident and powerful way. Diligently

working through the material first will enable this ability and the bold directive above will be strictly enforced in class to ensure this outcome.

## Honor Code

### The Honor Code

When you signed your letter of intent to enroll in the program, you initialed the following statement:

“I have read and agree to abide by the [SMU Honor Code](#).”

Please know that the Honor Code is taken very seriously.

Honor code violations include academic sabotage, cheating, fabrication, facilitating academic dishonesty, and plagiarism among others.

### Collaborative Work

Data science is a collaborative subject.

Most professors encourage collaborative work except when explicitly prohibited (usually on quizzes and exams).

Collaboration means helping one other, not copying answers from one another.

Students who turn in exactly the same answers to the same homework will share the grade assigned (i.e., if two students have the same answers, and the grade on the assignment is a 90, then each student will receive a 45).

Some instructors may impose stricter penalties.

Okay	Not Okay
Discussing problems (or even solutions) with other students  Telling another student how you did something in SAS (e.g., which PROC and how to use it) or sending “snippets” of code (e.g., an example of how to use a PROC) to help them out	Sending your answers to a student or receiving them from another student  Sharing anything, including ideas, when expressly forbidden, such as on exams  Passing off anyone else’s work or explanations as your own  <b>Using solutions from prior terms to inform your own work</b>

### Plagiarism

Here is an example of plagiarism:

A regression is a statistical analysis assessing the association between two variables. It is used to find the relationship between two variables.

The following is NOT plagiarism:

"A regression is a statistical analysis assessing the association between two variables. It is used to find the relationship between two variables" (<https://www.easycalculation.com/statistics/learn-regression.php>).

The difference is in the punctuation and the attribution.

Note that one can self-plagiarize. If you are using something that you wrote (e.g., a blog or a previously published article), please reference yourself.

## Consequences

Plagiarism, sabotage, fabrication, and cheating carry high penalties.

Instructors may choose to fail the student on the assignment, give a zero for the assignment, fail the student for the course, and/or bring the student before the Honor Council, at which point expulsion is an option.

The worst penalty from breaking the honor code is the fact that either the person or their peers (or both) miss out on learning the material; they miss out on growth.

The obvious takeaway is . . .

Have a blast in this course learning the material. We learn through making mistakes and then talking through and correcting those mistakes.

## Weekly Breakdown of Course Including Required Reading

<i>Unit</i>	<i>Topic</i>	<i>Reading from Statistical Sleuth</i>	<i>Additional Reading</i>
1	<i>Drawing Statistical Conclusions</i>	1	Brewer, L. M., and Cochran, J. J. (June 2013), "August's Child Is...Favoured by Fortune." Significance, 20-24. Cavagnaro, C. E. (Fall 2006), "Glide Testing: A Paired Samples Experiment." STATS, 46, 3-6.
2	<i>Inference Using T-Distributions</i>	2	
3	<i>Data Screening, Assumptions, and Transformations</i>	3	
4	<i>Type I, Type II Error, and Power</i>	<i>Asynch Only</i>	
5	<i>Alternatives to the T-Tools</i>	4	
6	<i>Comparisons Among Several Samples</i>	5	
7	<i>Linear Combinations and the Multiple Comparison Problem</i>	6	
8	<i>Midterm Exam Week</i>	1-6	

9	<i>Linear Correlation and Simple Linear Regression</i>	<i>7.1 -7.3 And 7.5.4</i>	
10	<i>Quantifying Uncertainty: Confidence, Prediction, and Calibration Intervals</i>	<i>7.4 – 7.5</i>	
11	<i>Regression Diagnostics and Model Refinement</i>	<i>8</i>	
12	<i>Multiple Linear Regression</i>	<i>9</i>	
13	<i>Inferences for Multiple Linear Regression</i>	<i>10</i>	
14	<i>Model Selection and Validation</i>	<i>11–12</i>	
15	<i>Review and Final Exam</i>	<i>1–12</i>	

### **Best Practices for Success in MSDS 6371 (And Other Courses Also)**

**Attendance.** Take responsibility for your commitment. Attendance means not only being there for synchronous sessions but also participating in asynchronous work.

**Citizenship.** You need to be actively engaged to succeed in this class. Talking on cell phones, texting, “Facebooking,” tweeting, or leisure web browsing are prohibited in class. These are considered to be disruptive (not to mention rude).

**Integrity.** A lot of the graded work occurs outside of class, so honesty and integrity are expected in your submissions. Evidence of academic dishonesty will minimally result in zeroes for all involved parties and perhaps University-level disciplinary action. Don’t risk your academic career.

**Humility.** Don’t get lost! Ask questions in class. If something isn’t clear to you, it probably isn’t clear to others either. Questions may arise because your professor hasn’t made a connection clear or has inadvertently left out an important point. Your question gives the professor a chance to explain more clearly. Don’t be proud or shy.

**Organization.** Don’t procrastinate! This is a technology-driven course. Count on your computer failing or your wireless connection breaking the night before a due date. Start early, and give yourself a chance to succeed.

**Deadlines.** You will generally have a week to complete an assignment. Due dates and times will be clearly indicated. Late submissions will be penalized, but it is much better to turn in work late than not at all (or to turn in incomplete/sloppy work). Work turned in after solutions have been posted to the course website will receive no credit.

**Getting help.** If questions arise while doing assignments/exams, do your best to resolve these questions before the assignment is due, first by taking time to seek answers yourself, next by asking questions on the wall, and finally via email to your instructor or other students. **I encourage you and expect you to seek help.** For questions during exams, please email the live session instructor directly.

**Collaboration.** The formation of study groups and collaboration with your fellow students in tackling the assignments is encouraged. Working together in groups on homework is permitted, even encouraged.

**However, every student should write up and complete his or her homework independently.**

**Students who chose to turn in exactly the same work will share the grade assigned.** Talking about problems with other people does help in learning, but just copying the solutions from one another doesn’t help!



**Looks do matter!** All assignments must be NEATLY executed and organized. You risk a zero on any assignment submitted in a sloppy manner. See submission guidelines for more detail. You are a graduate student and students are expected to submit thorough, organized and well thought out responses.

### **Submission guidelines for assignments**

- Your name must be at the top of the first page and on each successive page.
- Submit solutions in problem order.
- Use an easy-to-read variable-width font (try Ariel, Helvetica, and Geneva fonts—this document is in Helvetica, 11 point) with a minimum of 11-point font.
- Relevant SAS/R code and output from the SAS/R console must be included inline at the appropriate point using Courier New (or other fixed-width) font, in 10-point size. **Inclusion of irrelevant code or output will be penalized (aka SAS DUMP). Do not include a printout of the data as well (e.g., data/datalines statements in SAS).**
- Any graphics from SAS/R must be electronically cut and pasted inline at the appropriate point of the write-up. You can use Word to resize the graphics appropriately.
- Any mathematical notation must be provided with appropriate use of subscripts, superscripts, and symbols. Use MS Equation or another equation editor if you submit your work in Word.

### **University Policies**

- **Disability Accommodations:**  
Students needing academic accommodations for a disability must first register with Disability Accommodations & Success Strategies (DASS). Students can call 214-768-1470 or visit <http://www.smu.edu/Provost/SASP/DASS> to begin the process. Once approved and registered, students will submit a DASS Accommodation Letter to faculty through the electronic portal DASS Link and then communicate directly with each instructor to make appropriate arrangements. Please note that accommodations are not retroactive and require advance notice to implement.
- **Religious Observance:**  
Religiously observant students wishing to be absent on holidays that require missing class should notify their professors in writing at the beginning of the semester and should discuss with them, in advance, acceptable ways of making up any work missed because of the absence.  
<https://www.smu.edu/StudentAffairs/ChaplainandReligiousLife/ReligiousHolidays>.
- **Excused Absences for University Extracurricular Activities:**  
Students participating in an officially sanctioned, scheduled University extracurricular activity should be given the opportunity to make up class assignments or other graded assignments missed as a result of their participation. It is the responsibility of the student to make arrangements with the instructor prior to any missed scheduled examination or other missed assignment for making up the work. (See 2020–2021 SMU Undergraduate Catalog under “Enrollment and Academic Records/Excused Absences.”)
- **Student Academic Success Programs:**

Students needing assistance with writing assignments for SMU courses may schedule an appointment with the Writing Center through Canvas. Students wishing support with subject-specific tutoring or success strategies should contact SASP, Loyd All Sports Center, Suite 202; 214-768-3648; <https://www.smu.edu/sasp>.

- **Caring Community Connections (CCC) program:**

This is a resource for anyone in the SMU community to refer students of concern to the Office of the Dean of Students. Faculty play a critical role in identifying students who are experiencing challenges, as you may be the first to notice a change in behavior such as class attendance or performance. The online referral form can be found at [smu.edu/deanofstudentsccc](https://www.smu.edu/deanofstudentsccc). After a referral is submitted, students will be contacted to discuss the concern, strategize options, and be connected to appropriate resources. Additionally, should you have concerns about students and are unclear about what to do, please see the CCC Reference Guide, or contact the Office of the Dean of Students at 214-768-4564.

- **Pregnant and Parenting Students:**

- Accommodations for pregnant and parenting students: Under Title IX, students who are pregnant or parenting may request academic adjustments by contacting Elsie Johnson ([elsiej@smu.edu](mailto:elsiej@smu.edu)) in the Office of the Dean of Students, or by calling 214-768-4564. Students seeking assistance must schedule an appointment with their professors as early as possible, present a letter from the Office of the Dean of Students, and make appropriate arrangements. Please note that academic adjustments are not retroactive and, when feasible, require advance notice to implement.

- **Covid-19 Attendance Statement:**

Students who are experiencing COVID-19 symptoms or who have been notified through contact tracing of potential exposure and need to self-quarantine or isolate must follow the protocols laid out in SMU's Contact Tracing Protocol. To ensure academic continuity, students in these situations will not be penalized and will be provided appropriate modifications to assignments, deadlines, and testing. Please also note that SMUFlex classes might, in rare circumstances, go remote for two-week periods to accommodate COVID-related issues. To ensure these necessary accommodations, affected students must

- Provide as much advance notification as possible to the instructor about a change in circumstances. Students must notify their instructor about a potential absence as well as plans for a return to class. For cases in which students test positive for COVID-19, they should fill out a CCC form at this link:  
<https://www.smu.edu/StudentAffairs/officeofthedeanofstudents/StudentSupport/CCCProgram>
- 
- Communicate promptly with the instructor to establish, as necessary, alternative assignments and/or changes to deadlines and exams. Students are then responsible for meeting the expectations laid out in these alternative arrangements.
- Continue participation in class via Zoom, as health circumstances permit. Attend class regularly, when not in a situation outlined above, in accordance with safety measures laid out by SMU CAN in the Pledge to Protect (including wearing masks, maintaining social distancing, and cleaning personal space after class). In-person participation in SMUFlex classes is required on students' assigned red/blue rotation days except in cases when students are experiencing illness, are in self-quarantine, or in isolation.
- Students facing multiple or extended COVID-19-related absences or illness can work with the Office of the Dean of Students to consider options such as fully remote learning or medical withdrawal.

This policy, aligned with the SMU Honor Code and the SMU Pledge to Protect, relies on mutual trust and respect between students and faculty to ensure safety, academic integrity, and instructional continuity.

## **CALENDAR**

**<http://www.smu.edu/EnrollmentServices/Registrar/AcademicCalendarsCourseCatalogs/AcademicCalendars>**