George Mason University

STAT 665 – Categorical Data Analysis

Fall 2021

Instructor: David Kepplinger

Version 1 (August 11, 2021)

Administrative

Course dates: Mondays, Aug 23 – Nov 29 (final exam period: Dec 13);
no class on Sept 6 (Labor day);
Fall break on Mon, Oct 11; **class will be held on Tuesday, Oct 12 instead**.

Important dates: Aug 30: last day to add
Sept 7: final drop deadline (no tuition penalty)
Sept 27: end of self-withdrawal period

Instructor: Dr. David Kepplinger (he/him/his)

Email: dkepplin@gmu.edu

Office: Room 1711, Nguyen Engineering Building (ENGR)

Office hours: Thursday, 2 – 3pm or by appointment.
In addition to in-person office hours, you will also have the option of joining the office hour via Zoom: https://gmu.zoom.us/j/93949244371?pwd=N1AwVDZrRnZLdnZpNHBGUFZpUlQwQT09.

Blackboard course page: https://mymasonportal.gmu.edu/ultra/courses/_435680_1

Class time: Mon, 7:20 – 10:00pm

The class is scheduled for face-to-face on-campus meetings. All learners taking courses with a face-to-face component are required to take the Safe Return to Campus Training prior to coming to campus. Training is available in Blackboard (https://mymason.gmu.edu). You are required to follow the university's public health and safety precautions and procedures outlined on the university Safe Return to Campus webpage (https://www2.gmu.edu/safe-return-campus). Importantly, everyone must complete the Mason COVID Health Check daily, seven days a week. The COVID Health Check system uses a color code system and you will receive either a Green, Yellow, or Red email response. Only if you receive a “green” notification you are permitted to attend courses with a face-to-face component. If you suspect that you are sick or have been directed to self-isolate, please quarantine or get testing. Faculty are allowed to ask you to show them that you have received a Green email and are thereby permitted to be in class. If the campus closes or class is canceled due to weather or other concern, learners should check Blackboard for updates on how to continue learning and information about any changes to events or assignments.

Communications The Blackboard site for this course is the primary channel of communication. Please check the Blackboard course regularly for updates! Information posted on the Blackboard site includes

- announcements,
- lecture notes,
- homeworks and assignments,
- changes to the posted office hours,
- handouts and readings.

Moreover, general topic forums and forums specific to assignments are available in the Discussion Board. Please check the Discussion Board regularly. You are strongly encouraged to post questions about assignments in the Discussion Board (you are likely not the only one with the same question). Consider the Discussion Board as an extra resource for getting help with assignments.

E-mail communication should be restricted to questions relating to sensitive, confidential information (such as grade concerns, personal circumstances requiring specific accommodations, etc.).

- E-mails will be returned within 2 business days and may not be returned on weekends/holidays.
- When you send an e-mail to me, please put `STAT 665` at the beginning of the subject line.
- E-mails related to this course must be sent and received via your Mason e-mail account. **E-mails sent from other e-mail accounts may not be answered.** (This is a university policy and part of your guaranteed rights under FERPA.)

Should you have concerns that you may not be able to fully participate or engage in any of the activities listed below, please do not hesitate to contact me either by e-mail or speak to me in person during office hours or after class. We can discuss alternative arrangements that suit your needs.

Course requirements

Prerequisites: STAT 654 or equivalent; working knowledge of R, intermediate knowledge of at least one statistical programming environment (e.g., SAS, R, STATA, Julia). Familiarity with parametric and nonparametric testing and confidence intervals for one-sample, two-sample, and ANOVA; familiarity with the statistical principles of modeling.

Corequisites: STAT 544 or equivalent (familiarity with calculus-based, applied probability theory).

Recommended readings: The main textbooks for this course are

- A. Agresti (2013). *Categorical Data Analysis*. 3rd edition. Wiley. [available online from the GMU library at https://wrlc-gm.primo.exlibrisgroup.com/permalink/01WRLC_GML/1giah39/alma9946905563404105]
- T. Hastie, R. Tibshirani, J. Friedman (2009). *The Elements of Statistical Learning*. 2nd edition. Springer. [available online for free at https://web.stanford.edu/~hastie/ElemStatLearn/]

A number of relevant articles will be posted in Blackboard as different topics are discussed. The following list comprises books which provide additional discussions on some of the topics covered in class, but these books are not required for this course:

- Julian J. Faraway (2016). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*, Second edition. Chapman and Hall/CRC.

Software requirements: This class will use R (version 4.0 or higher) to perform data manipulation, analysis, and visualization. For your assignments you are allowed to use any statistical *programming* environment where you can submit the complete code for reproducibility (e.g., R, SAS, Julia, Python, but not MS Excel or SPSS). The focus of using R (or any other statistical programming environment) in this course is on data analysis, implementing new methods, and managing code in a consistent and professional style, interpreting outputs, and ensuring reproducibility. The course will be taught using RStudio Desktop, accessible for free as download, but you can use any interface you are comfortable with. For details on accessing and using RStudio Desktop see Blackboard.

You will also be required to have a Git client on your computer. This course will be taught using GitKraken (https://www.gitkraken.com/) available free as download. Alternatively, you may choose to use a different interface for Git, such as Sourcetree, the UI built into RStudio Desktop, or the plain git console.

In case of campus closures, activities and assignments in this course will use the web-conferencing software Zoom. In addition to the requirements above, you are required to have a device with a functional camera and microphone.

Course description

Learning objectives: After successfully completing STAT 665, you will be able to

- recognize appropriate discrete probability distribution for categorical outcomes;
- identify valid models for analyzing data with categorical outcomes;
- conduct inference using common estimators for categorical outcomes;
- correctly interpret and report results from a categorical data analysis;
- adapt common estimators for categorical outcomes to specific problems and derive approximate confidence intervals;
- use common methods for categorical data analysis in statistical programming environments;
- implement variations of methods for categorical data analysis in statistical programming environments;

Main topics: You can expect the following topics to be covered in some detail.

- Classical and exact analyses methods for two- and three-way contingency tables.
- Generalized linear regression models (GLMs) for modeling the association between discrete outcomes and one or more explanatory variables.
- Loglinear models (comparable to ANOVA, but for categorical outcomes).
- Methods suited in settings where one or more assumptions of GLMs are violated (e.g., non-independent outcomes, over- or under-dispersion).
- Methods to handle large numbers of explanatory variables (high-dimensional data) or complex (even unknown) association structures.

Assessments and grading

Your grade in this course will be based on (roughly) weekly homework assignments of various types (40%), a written final project report (30%), an oral presentation (16%), and one set of presentation notes (14%). There are no tests for this course. However, you will be submitting a final project report due at the final exam period.

Written and oral communication are an integral part of any statistical project, and as such, grammar, style, and spelling are part of grading rubrics applied to all deliverables: homework assignments, reports, presentation notes and oral presentations. You are strongly encouraged to use the resources and tutoring offered by the writing center (https://writingcenter.gmu.edu).

All assignments in this course are designated as individual assignments, which are to be undertaken independently. You may discuss your ideas with others but everything you turn in must be your own work. You may not share analyses, graphs, and other materials. You are responsible for making sure that there is no reason to doubt that the work you hand in is your own. The following types of collaboration on individual assignments are not honor code violations:

- Working on assignments with someone who is at roughly the same stage of progress as you, provided both learners contribute in roughly equal quantity and quality (in particular, thinking) on whatever problem or problem parts they collaborate. This type of collaboration is actually encouraged!
- A moderate amount of asking, “How do I do this in R?” However, as you gain enough familiarity, you should get in the habit of using online help and trying logical possibilities, then asking for help only if these do not succeed after a reasonable try.
- Using SAS/R code found on the internet to conduct your analysis, if using proper attribution (clearly identifying all code snippets which are not your intellectual product).

The following types of collaboration on assignments **are** honor code violations:

- Working together with one learner the doer and one the follower.
- Any type of copying. In particular, splitting up a problem so that different learners do different parts is not authorized collaboration on homework. This also includes copying code from the internet without properly identifying the source.

Attendance: Attendance is expected and one oral presentation as well as presentation notes of one other oral presentation is part of your final grade. You will be able to choose the two dates of your oral presentation and your note taking at the beginning of the term and you need to be present on those days.

If you miss class, please get notes from your peers. You are responsible for material covered in class and announcements made during class.

Participation: Success in this course requires active participation in in-class activities, for which you will need to prepare in advance for each class period. Accordingly, you are expected to prepare for class period by

- reading the corresponding sections of the textbook to be covered in class,
- reviewing class materials posted in Blackboard to be covered in class,
- familiarizing yourself with the use of the covered methods and techniques in the statistical programming environment of your choice.

Homework assignments: There will be roughly weekly homework assignments throughout the term which will vary in length and content. Some involve in-class activities and continuing data analyses started in class, others involve solving exercises related to the material covered in class. Only 8 of your submissions will be grade, it is your responsibility to not submit a homework if you do not want it to be graded. Once the homework is submitted, you cannot withdraw that homework and it will count towards your final grade. When you have submitted 8 homeworks over the course of the semester, any following homeworks will not be graded or considered for the final grade. You will typically have 6 days to complete each homework. Due dates will be posted in Blackboard and your deliverables are submitted either on Blackboard or GitHub. Homework assignments will not be accepted late; late submissions will not count towards the 8 required submissions.

Oral presentation: You will give an oral presentation with visual aids on one specific method/concept extending ideas discussed in class the previous week. The presentation will be either based on one or more peer-reviewed journal articles or a book chapters which will be provided by me. You will be required to provide an in-depth discussion of the method/concept as well as an application. More details will be provided in class and on Blackboard.

Presentation notes: You will need to take notes for one presentation given by one of your peers. This will be on a different date than your own oral presentation, and assigned by me. You will have until the following Friday (4 days) to type up your notes and commit them to our shared GitHub repository. You may communicate with the presenter to ask for clarifications and the presenter may share the graphs and tables used for the presentation (if any). Once you push your notes to the shared repository, the presenter and I will give feedback and possibly request a revision within 4 days. You will have another 4 days to incorporate that feedback. Both the original submission and the final revision will be considered when assigning a grade. The revised version of the notes will be shared with everyone enrolled in the class. More details will be provided in class and on Blackboard.

Final project: The final project will involve discussion of one or more methods for categorical data analysis, analyzing a data set, and submitting a report and the fully functional SAS/R code via GitHub. You must ensure that the analysis and all results are reproducible. The final report and analysis code will not be accepted late. Rough topics for the final project will be provided and the specific deliverables are discussed individually with each learner.

You are expected to address your final project report with the same level of preparation and presentation that you would associate with a finished product on your job as an applied statistician. The report that you write for this course will be graded on both your analysis and your writing. Each report will be no more than 6 pages of text (typically 4–5 pages) plus a few more for tables, graphs, and charts. Reports must not include a long appendix of outputs from statistical software and no computer code.

Regrading policies: You have at most one week after a score is posted for an assignment to appeal the score. If you want parts of an assignment remarked, send me an email specifying the question/part and the reason for requesting a review of grading. If you do not notify me in writing of any issues with your score within that time, then the posted score stands (whether or not it is correct).

Policies and Classroom Climate

During classes and online you are encouraged to discuss and share ideas with your classmates (see above how this relates to the honor code). To facilitate a respectful and inclusive classroom climate, be open to explore and challenge each other's ideas without criticizing individuals. Diversity is a source of creativity and innovation and I ask that learners appreciate diverse perspectives, that they listen respectfully and let everyone speak. If you have concerns about the dynamics or classroom climate, please do not hesitate to bring them to my attention.

The School of Computing seeks to create a learning environment that fosters respect for people across identities. We welcome and value individuals and their differences, including gender expression and identity, race, economic status, sex, sexuality, ethnicity, national origin, first language, religion, age and ability. We encourage all members of the learning environment to engage with the material personally, but to also be open to exploring and learning from experiences different than their own.

Gender identity and pronoun use: If you wish, please share your name and gender pronouns with me and how best to address you in-person or via email. I use he/him/his for myself and you may address me as “David”, “Prof. Kepplinger” or “Dr. Kepplinger” in email and verbally.

Individual accommodations: Disability Services at George Mason University is committed to providing equitable access to learning opportunities for all learners by upholding the laws that ensure equal treatment of people with disabilities. If you are seeking accommodations for this class, please first visit http://ds.gmu.edu for detailed information about the Disability Services registration process. Then please discuss your approved accommodations with me. Disability Services is located in Student Union Building I (SUB I), Suite 2500. Email: ods@gmu.edu | Phone: (703) 993-2474.

Class etiquette: Class will start on time at 7:20 p.m. and end on time at 10:00 p.m., with a 10-minute break around 8:30 p.m. Although situations may arise making it impossible for you to arrive on time and/or requiring you to leave early, please remember that late arrivals and early departures can be quite disruptive to your classmates. So, please make arriving to class late or leaving early an exception, not a habit. **Regular attendance for the full period of each class is very important for this course!**

- Mute your phones during class, and keep them stowed away.
- You may eat during class, as long as it is done discreetly, quietly, and odorless.
- Immediately before or after class is not a good time to ask lengthy questions. Please come to office hours (or make an appointment) instead. Questions during class are welcomed and encouraged.

Netiquette: We will often communicate via discussion forums, GitHub issues, and other forms of online communication. To facilitate effective communication via these channels, please adhere to the following:

- *Be relevant and concise:* When posting a message to an online discussion, stick to the topic, make sure that you send enough information, and be concise.
- *Use accurate topic titles:* Each posting should include a topic title (a subject line) that lets the recipient know the posting’s content. This allows others to scan their online messages, read the more important messages first, and keep organized.
- *Read before posting:* Read posted questions/answers before asking a new question to avoid repeating points already made, asking questions already answered, or bringing up points

that have already been argued and either accepted, rejected, or exhausted. In addition, by "replying" to messages instead of starting a new message, a thread of communication can be kept going.

- *Be polite:* Avoid inflammatory messages and language. Do not send a message that ridicules someone else. Also, be careful when using humor or sarcasm, as most of it gets lost in the medium.
- *Review messages before submitting:* Think before you "speak" electronically. For the most part, electronic communication is a non-visual form of communication; therefore, people are unable to rely on facial expressions, tone of voice, or body language to interpret electronic messages. Misunderstandings can easily occur because of these factors.

Notice of mandatory reporting of sexual or interpersonal misconduct: As a faculty member, I am designated as a "Non-Confidential Employee," and must report all disclosures of sexual assault, sexual harassment, interpersonal violence, stalking, sexual exploitation, complicity, and retaliation to Mason's Title IX Coordinator per University Policy 1202. If you wish to speak with someone confidentially, please contact one of Mason's confidential resources, such as Student Support and Advocacy Center (SSAC) at 703-380-1434 or Counseling and Psychological Services (CAPS) at 703-993-2380. You may also seek assistance or support measures from Mason's Title IX Coordinator by calling 703-993-8730, or emailing titleix@gmu.edu.

Honor Code: The integrity of the University community is affected by the individual choices made by each of us. Mason has an Honor Code with clear guidelines regarding academic integrity; you are responsible to know your requirements for this course. All violations of these rules will be referred to the Honor Committee; I take the Honor Code seriously and so should you. No grade is important enough to justify academic misconduct. If you have any questions concerning the Honor Code and how it relates to this particular course, please contact me.

Some kinds of participation in online study sites violate the Mason Honor code: these include uploading of any of the course materials or exams; and uploading any of your own answers or finished work. Always consult your syllabus and me before using these sites.

Privacy: Your privacy is governed by the Family Educational Rights and Privacy Act (FERPA) and is an essential aspect of this course. You must use your MasonLive email account to receive important University information, including communications related to this class. I will not respond to messages sent from or send messages to a non-Mason email address.

All course materials posted to Blackboard or other course sites are private to this class; by federal law, any materials that identify specific learners (via their name, voice, or image) must not be shared with anyone not enrolled in this class.

- Videorecordings — whether made by me or learners — of class meetings that include audio, visual, or textual information from other learners are private and must not be shared outside the class.
- Live video conference meetings (e.g., Zoom) that include audio, textual, or visual information from other learners must be viewed privately and not shared with others in your household or recorded and shared outside the class.

Copyrights of course material: This course gives you access to presentations, handouts, and copyrighted material and articles. Please treat them accordingly. All material other than copyrighted material should be regarded as authored materials, which if used or referred to must be fully credited through reference to me, the course, and date. If used beyond citation, my permission is required.

TENTATIVE Class Schedule

Date	Topics covered	Textbook chapters†
M 08/23	Syllabus, expectations, and prerequisites	1
M 08/30	Principles of statistical inference for categorical data	1
~~M 09/06~~	~~Labor day (no class)~~	
M 09/13	Models for contingency tables	2
M 09/20	Inference for contingency tables	3
M 09/27	Basics of generalized linear models	4.1–4.6, 6.3, 6.5, 6.6
M 10/04	Models for multinomial responses	8
~~M 10/11~~	~~Fall break, class held on Tuesday~~	
T 10/12	Loglinear models	9
M 10/18	Marginal models	12
M 10/25	Generalized linear mixed effects models	13
M 11/01	Quasi-likelihood models and mixture models	4.7, 14
M 11/08	Model selection	7.5, ESL18.4
M 11/15	Classification – tree based methods	5.1, 15.2, ESL9.2, ESL15
M 11/22	Classification – boosting	ESL10
M 11/29	Discussion of final project topics	
M 12/13	**Final project due**	

† No prefix refers to the textbook “Categorical Data Analysis” (Agresti, 2013), and the prefix *ESL* refers to “The Elements of Statistical Learning” (Hastie, Tibshirani, Friedman, 2009).

Version History

Version 1 (August 11, 2021) Initial version.