Statistics 601 **November 25, 2018**

Final take-home exam— Due Dec 19, 2018

Final Exam Policy and Guidelines

1. The exam consists of problems similar to those that may later be encountered on the job. Submit only the electronic version of you report on canvas. There is no need to hand in the hard copy. The deadline for submission is **Dec 19 (Wednesday), 2018, 12:05pm.** You are allowed to update the submission anytime before the deadline. Late submission will not be accepted .

2. **Collaboration on any part of the problems is not permitted**. You need to work out the problems independently without discussing with other students, TA or instructor. The exam should be treated as confidential and you should ensure that no one else reads it until after the deadline of submission.

3. The exam has three problems: **the first two problems are required, and the third one is optional**. You will not lose any points for not working on the bonus problem. However, if your final scores are at the borderline, the bonus points may help you round up to a higher letter grade. Write a combined report for the problems on the exam. **Do not exceed page limits for each problem**, including summary. All major conclusions should be stated at the beginning in a summary intended for a scientifically literate reader who is not a statistician.

4. Following the summary, the report should describe the models fitted, the tests performed, and how these support the conclusions. The relevance of the models to the context under study is important. Writing should be concise, legible and thorough, while containing the appropriate amount of "hard" statistical information. Finding the balance between too much and too little formal statistical information is a key part of report preparation.

5. The report text must be in a font that produces no more than 56 lines of text per page (for example, 1.5 line spacing, standard width, and various 11pt fonts in Word; or various LaTeX formats using 11pt or 12pt). The captions for Figures and Tables should be informative.

6. A very brief Appendix is allowed to be included with each problem, and is not counted toward the page limit. In general, use caution in deciding what to include in an Appendix. Information critical to major analyses should appear in the main body of the report. Lengthy appendices beyond a few extra graphs and tables will receive much less attention than the main body of the report.

7. While it is necessary to demonstrate that you have mastered the computer system or statistical package, it is more important to demonstrate that you are not a slave to the computer by tailoring the computer output to the problem at hand. Below are suggestions along this line:
   (a) From the computer-generated analysis-of variance table, list only the parts that are relevant to your analysis.
   (b) If the model matrix is non-standard and cannot be generated by a model formula, as for example the additive skew-symmetric formula $\mathbb{E}(Y_{ij}) = \alpha_i - \alpha_j$, you need to explain what the structure of the matrix is.
   (c) Do not quote a p-value without stating the hypothesis under test and how the value supports your conclusion.
   (d) Do not analyze a model without explicitly stating or justifying the assumptions.
   (e) Parameters have a physical interpretation: do not pass up the opportunity to remind the reader what the physical interpretation of $\hat{\beta} = 0.684$ is in the context of the problem.

8. It is helpful to indicate what further analyses might have been helpful had there been more time or different software available.

9. Start early, work hard, and good luck!

Final Exam Problems:

1. **Calories (8 pages, 40 points)** Analyze the "common household food" dataset. Examine various aspects about calories of common household food. Any systematic variation if of interest. Write a report on your findings. Make sure you take out the last missing observation!

   The following questions give you a hint on questions to think about:

   (a) What variables to include in the model? Any multicollinearity in the predictors? How to address them?

   (b) Does the data follow the assumptions of regression? Do we need to remove any outlier or make necessary adjustments (i.e. transformation)?

   (c) Suppose we want to include exactly one nutritional measurement from each of the following categories

      i. Fats: Fat (in grams), saturated fat (in grams), monounsaturated fat (in grams), polyunsaturated fat (in grams)

      ii. Protein: Protein (in grams)

      iii. Carbohydrates: Carbohydrates (in grams)

      iv. Cholesterol: Cholesterol (in mg)

      v. Vitamins: Vitamin A (in IU), Vitamin A (in RE), Vitamin B1 (in mg), Vitamin B2 (in mg), Vitamin B3 (in mg), Vitamin C (in mg)

      vi. Minerals: Calcium (mg), Phosphorus (mg), Iron (mg), Potassium (mg), Sodium (in mg)

      vii. Weight: Weight (in grams)

   in to a regression model about calories. Specifically, we want

   $$Calories \sim \beta_0+\beta_1(Fats)+\beta_2(Protein)+\beta_3(Carbs)+\beta_4(Chol)+\beta_5(Vitamin)+\beta_6(Minerals)+\beta_7(Weight)$$

   Come up with a procedure to obtain the "best" model under this constraint. Here, "best" model is the model that has the most accurate prediction for a given value of X. You may use any procedure you wish (or you may come up with one of your own). What is your model?

   (d) To test your model's prediction power, consider the following food item in table 1. What are your predictive value and prediction interval?

2. **Audibility (12 pages. 60 points)** Analyze the "audibility" dataset.

   (a) Background: The researchers are interested in evaluating if EEGs (electro-encephalograms, or readings of brainwaves) can predict audibility of speech. The specific type of EEG responses used are called envelope following responses (EFRs). EFRs are recorded with scalp electrodes when speech is presented in an individual's ear and they represent neural activity that follows the periodicity in speech.

   (b) Experiment Design: In this experiment, we use 8 speech sounds (aka carriers) to elicit EFRs - /u/F1, /u/F2, /a/F1, /a/F2, /i/F1, /i/F2, sh and s. F1 and F2 refer to the first and second vowel formants, respectively. We use a range of stimuli because we are interested in evaluating audibility at low, mid and high frequencies (perceived as pitch). The F1 carriers are low frequency dominant, the F2 carriers are mid frequency dominant and the fricatives (sh and s) are high frequency dominant. EFRs are quantified by their amplitude (response amplitude) and phase (degrees). To evaluate if there is an EFR, we use two methods: (i) we compare its amplitude with noise amplitude using an F-test and (ii) we compare the inter-trial phase consistency using a Rayleigh test. P-values of < 0.05 are considered "detected" EFRs. We expect an EFR to be detected if the stimulus is audible. However, because EFRs reduce in amplitude as stimulus

Table 1: X value for Problem 1(d)

Fat	1.5 grams
Saturated fat	0 gram
Monounsaturated fat	0.5 grams
Polyunsaturated fat	0.5 grams
Protein	3 grams
Carbohydrates	26 grams
Cholesterol	0 mg
Vitamin A	1250 IU, 125 RE
Vitamin B1, Vitamin B2, Vitamin B3	0 mg
Vitamin C	30 mg
Calcium, Phosphorus	0 mg
Iron	1.8 mg
Potassium	95 mg
Sodium	85mg
Weight	33g

levels get quieter, we may not be able to detect EFRs at very quiet levels. Loudness, or sound pressure level (SPL), is measured in units of decibels (dBs). In each participant, the 8 stimuli were presented at 4 levels (20, 35, 50 and 65 dB SPL) ranging from inaudible/very quiet levels (20 dB SPL) to loud levels (65 dB SPL). Audibility for each stimulus and each level was quantified by sensation level expressed as dB SL. Positive SLs mean that the stimuli were audible and negative SLs mean that the stimuli were inaudible. You can interpret SL roughly as (SPL − threshold of Carrier), i.e., comparing SPL vs. the participant's known audible threshold for that sound. At any given stimulus level (e.g., 20 dB SPL), some of the 8 stimuli may be audible and some may not be.

Write a report of your analysis on the following questions:

(a) How accurately can the audibility of each speech stimulus be predicted using EFRs? We will mainly consider detectability, which is a binary outcome. Does accuracy differ between carriers or frequency groups?

(b) Is there a difference in performance between the F-test and the Rayleigh in predicting audibility?

(c) What is the minimum SL needed in order for the EFR to detect a response? How does the minimum SL vary by carrier or frequency group?

(d) What are the limitations of your methods? Pay special attention to the independence assumption. Any remedies?

Table 2: Code book for the audibility dataset

Variable	Meaning
Participants	unique participant identifier
Carrier	speech sounds, /u/F1, /u/F2, /a/F1, /a/F2, /i/F1, /i/F2, sh and s
SPL	controlled stimulus pressure level that was presented to the participant
SL	sensation level based on each participant?s known audible threshold for that sound
Rayleigh	p-value of the Rayleigh test
Ftest	p-value of the F-test

3. **Bonus problem (3 pages, additional 5 points)** The following data were collected as part of a high-school electro-chemical experiment by P. Ohtani. To obtain an observation, two metals i, j, were inserted into an electrolytic solution, and the voltage difference Y_{ij} between i and j recorded by a digital voltmeter. The voltage difference between i and j is, by definition, the negative of the difference between j and i, so each observation is recorded twice.

   (a) A circuit is a closed loop, $i_0, \cdots, i_{n-1}, i_n = i_0$ of length $n \geq 0$. Conservation of energy is a condition on the $k \times k$ matrix V to the effect that, on each circuit the sum is zero

$$V(i_0, i_1) + V(i_1, i_2) + \cdots + V(i_{n-1}, i_0) = 0.$$

   A matrix satisfying this condition is called conservative. Show that each conservative matrix is skew-symmetric. Deduce that the set of conservative matrices is a vector space, closed under vector-space operations. Exhibit a 3×3 skew-symmetric matrix that is not conservative. A skew-symmetric matrix of the form $V(i, j) = \alpha_i - \alpha_j$ is called additive. Prove that every additive matrix is conservative. Prove that every conservative matrix is additive. What is the dimension of the vector space of conservative 6×6 matrices?

   The following exercises refer to the linear model for the voltages in which $\mathbb{E}(Y_{ij}) = \alpha_i - \alpha_j$ is conservative. The data is as follows:

Electrolyte O

	Mg	Zn	Fe	Pb	Cu
Mg	0.0	0.414	0.807	0.876	1.291
Zn	−0.414	0.0	0.429	0.533	0.886
Fe	−0.807	−0.429	0.0	0.043	0.377
Pb	−0.876	−0.533	−0.043	0.0	0.271
Cu	−1.291	−0.886	−0.377	−0.271	0.0
Electrolyte A					
Mg	0.0	0.247	0.856	1.051	1.402
Zn	−0.247	0.0	0.434	0.521	0.867
Fe	−0.856	−0.434	0.0	0.058	0.443
Pb	−1.051	−0.521	−0.058	0.0	0.374
Cu	−1.402	−0.867	−0.443	−0.374	0.0
Electrolyte K					
Mg	0.0	0.443	0.895	0.973	1.281
Zn	−0.443	0.0	0.477	0.503	0.856
Fe	−0.895	−0.477	0.0	0.107	0.432
Pb	−0.973	−0.503	−0.107	0.0	0.392
Cu	−1.281	−0.856	−0.432	−0.392	0.0

   (b) For a single $k \times k$ table, obtain an expression for the least-squares estimate of $\boldsymbol{\alpha}$. Use this formula to compute $\hat{\boldsymbol{\alpha}}$ for each of the three electrolytes. Explain why $(\alpha_1, \cdots, \alpha_5)$ and $(\alpha_1, \cdots, \alpha_5) + (c, c, c, c, c)$ are equivalent as parameter points in the model.

   (c) Assess the evidence for and against the hypothesis that the vector of potentials is constant across electrolytes. That is to say, fit the linear model in which the potentials are constant across electrolytes, and compare the fit with the model in which α varies from one electrolyte to another. Obtain the relevant sums of squares, their degrees of freedom, and compute the appropriate F-statistic.

   (d) Discuss briefly the arguments for and against analysis of these data by linear models after transformation.