

Fall 2020 Final Examination

Statistics 104

Due 17 December 2020, 5:00 PM

- The exam consists of 3 problems and 9 pages. The exam is worth 200 points, and the point value for each question is displayed.
- The exam emphasizes material in Units 4 - 9, but also covers material from Units 1 - 3.
- Your solutions are due at 5:00 pm, 17 December 2020. No late submissions will be accepted.
- Solutions must be uploaded to the course website. Submit 1) a PDF file produced from R Markdown and 2) the R Markdown file used to produce the PDF. Name the files with your first initial and last name; e.g. stat104_final_fall2020_k_mckeough.
- Before submitting the exam, read and (electronically) sign the statement on the second page confirming that you have worked independently.
- Be sure to read the questions carefully. Some parts of a problem statement may ask for more than one calculation.
- Some parts of a question may require the answer to an earlier part. If you cannot solve the earlier part, you can still receive partial credit for the later parts; make up a reasonable answer for the earlier part to use in subsequent parts of the problem.
- Show your work and explain your reasoning; the final answer is not as important as the process by which you arrived at that answer. We can more easily give partial credit if you have written out your steps clearly.
- You may use any course materials while working this exam, including the lecture slides, labs, lab notes, section materials, problem set solutions, and the *OpenIntro* textbooks. While access to non-course materials is also permitted, the exam has not been written to require any outside research.
- All your work must be your own. Collaboration is strictly forbidden, including any discussion about resolving technical issues related to knitting files, etc. This includes posting any questions about the exam or discussing the exam on the class Slack channels.
- Answers must be in your own words. Plagiarism is not acceptable and we will very likely detect if an answer has been copied from a website.
- Once the exam is released, the teaching staff will not be able to provide assistance with working through problems, or to answer questions about concepts covered in class.
- Office hours will be held for help with technical issues, such as files that may not knit, or R code that may not run. If you are experiencing technical issues, please contact the teaching staff via email (or private messages on Slack).
- Be sure to frequently save your work, in addition to saving copies on your local machine. Knitting errors and unexpected loss of work are not acceptable excuses for late submission.

PROBLEM 1: SHORT ANSWER (50 pts. total)

- a) (20 pts.) The coach of the Red Shirts, a Multi-dimensional League Baseball team, is recruiting new players. To do so, she is conducting a home run competition. Each batter gets 100 tries to hit an identically pitched ball over a fence that is 310 yards away; a ball hit over the fence is considered sufficient for making a home run. The coach would like to recruit the top 1% of all batters.
- i. A typical batter will hit 60% of the balls; of the balls they hit, suppose that the distance traveled is normally distributed with mean of 300 yards and standard deviation 50 yards. How many home runs should the coach set as a minimum limit for being recruited to the Red Shirts?
 - ii. J.D. Martian-ez hits the ball 80% of the time. When he hits the ball, the distance it travels is normally distributed with mean 295 yards and standard deviation of 60 yards. Based on the minimum limit set in part i., what is the probability that J.D. will make the team?
 - iii. The coach plans to host tryouts every week, allowing players to participate at most 10 times. J.D. plans on trying out until he makes the team. What is the expected number of times J.D. will have to try out to make the team?
 - iv. If the coach allows players to try out more than once, will she actually be identifying the top 1% of all batters? Explain your answer.
- b) (30 pts.) Suppose that a pharmaceutical company has asked you to work as a consultant on their COVID-19 vaccine project. They are planning to conduct a study to demonstrate that their vaccine candidate is more effective than the one Pfizer and BioNTech announced in early November 2020, which was stated to have efficacy of 90% against placebo in participants without prior evidence of SARS-CoV-2 infection. They currently plan to enroll 2,000 participants, randomizing half to receive their vaccine candidate and half to receive the vaccine candidate from Pfizer and BioNTech. They are interested in detecting a difference in proportions of at least 5% on the outcome of whether a patient becomes infected with SARS-CoV-2. The results of the trial will be analyzed with a two-sided test.

Previous dose-testing trials have been conducted to determine the lowest possible dose of virus sufficient for infecting most of those exposed; this will be the dosage used in the current study. Participants in the study will be healthy volunteers ages 18-30 years who consent to being deliberately exposed to the SARS-CoV-2 virus in a controlled setting and have screened negative for risk factors associated with severe COVID-19. Following exposure, any participants who test positive for the virus will be immediately treated with an antiviral drug. Studies that involve deliberate exposure of participants to an infectious disease are commonly referred to as challenge trials.

You have been asked to conduct a simulation study to assess whether the study team has chosen an appropriate sample size. Based on discussions with the study team, you decide it seems feasible that they have identified a more effective vaccine and assign probability 0.75 to the alternative hypothesis.

- i. Conduct a simulation incorporating your belief about the efficacy of the new vaccine candidate; be sure to clearly comment your code where necessary. In one paragraph, briefly explain the general logic and organization of your simulation approach.

- ii. Based on the simulation results, comment on whether the sample size is appropriate. If you believe the sample size is not appropriate, suggest a new sample size and explain your reasoning.
- iii. Suppose the study team conducts the trial with the sample size you recommend in part ii. and obtains statistically significant results at the $\alpha = 0.05$ level. Compute the estimated probability that the alternative hypothesis is true, given the observed data. Explain why this estimated probability changes if you have less confidence in the effectiveness of the new vaccine candidate.
- iv. Suppose the study team conducts the trial and obtains statistically significant evidence that the new vaccine candidate has efficacy greater than 90%, with p -value = 0.03. A news outlet reports on the research finding, claiming that based on the p -value, the new vaccine is substantially better at protecting against COVID-19 than the Pfizer/BioNTech vaccine.

In language accessible to an audience without a statistics background, explain why the news outlet's interpretation of the p -value is flawed and provide an accurate interpretation. Limit your answer to no more than seven sentences.

- v. The vaccine data Moderna announced on 16 Nov 2020 are from a field study, where individuals were randomized to receive either the vaccine candidate or placebo then instructed to return to their daily routine and use an online app daily to screen themselves for COVID-19 symptoms. Both the study participants and the staff administering shots were blinded to which shots contained vaccine or placebo. The *New York Times* press release mentioned that Moderna "slowed enrollment in September to ensure diversity among participants, and ultimately included 37 percent from communities of color, and 42 percent from populations considered at high risk because they were over 65 or had conditions like diabetes, obesity or heart disease."

Comment on whether the results from Moderna's field study or the hypothetical challenge trial would be more informative for understanding whether a particular vaccine candidate is effective for protecting college undergraduates from SARS-CoV-2 infection while they are living on campus. Be sure to fully explain your reasoning. Limit your answer to no more than six sentences.

PROBLEM 2: MCAS (70 pts. total)

The Massachusetts Comprehensive Assessment System (MCAS) is a standardized exam administered to students in the state of Massachusetts since 1993. Under law, students educated with public funds are required to participate in statewide testing. Students take exams according to their grade level in subjects such as Mathematics and English Language Arts; passing grades on the Grade 10 MCAS are required for high school graduation. Exam results are used to check student progress as well as measure school and district performance.

However, research has shown that variation in standardized test scores is often explained by factors such as race and socioeconomic class; students of color and students from economically disadvantaged backgrounds tend to score lower on standardized tests as a result of having less access to educational resources than their peers. For example, a *New York Times* article from October 2020 reported the results of a study demonstrating that in the United States, students performed worse on standardized tests for every additional day that was 80 degrees Fahrenheit or higher—the association was observed for Black and Hispanic students, in addition to students with lower family income, but not for white students.

In this problem, you will use data from the 2018 Grade 10 Mathematics MCAS to investigate evidence of achievement gaps in standardized test score at the school level. The mcas data contains information on 355 schools in Massachusetts. Scores on the MCAS are classified into levels: warning/failing, needs improvement, proficient, and advanced. Students who score in the “proficient” category are said to “demonstrate a solid understanding of challenging subject matter”, while those in the “advanced” category are said to “demonstrate a comprehensive and in-depth understanding of rigorous subject matter”; the variable PA_perc represents the percentage of students at a school who score at the proficient or advanced level.

The descriptions of the variables are as follows.

Variable	Description
PA_perc	percentage of students scoring proficient or advanced
class_size	average class size
math_class_size	average math class size
student_teacher_ratio	student to teacher ratio
attendance_rate	average percentage of days attended across students
number_of_students	total number of students who took the exam
largest_minority	largest minority group among the student body
white_less50	coded TRUE if less than 50% of students are white
exp_per_pupil	average expenditures per pupil in USD
econ_dis	percentage of economically disadvantaged students

Some additional background on certain variables:

- Whether a student is economically disadvantaged is a proxy measure for student family income; a student is considered economically disadvantaged if they are participating in one or more of the following programs: the Supplemental Nutrition Assistance Program (SNAP), the Transitional Assistance for Families with Dependent Children (TAFDC), the Department of Children and Families’ (DCF) foster care program, MassHealth (Medicaid).
- Expenditures per pupil is a proxy measure for the amount of funding available to a school district. This variable is measured by district (i.e., constant for schools in the same district).

Use these data to answer the following questions.

- a) (14 pts.) Explore the data.
- Briefly summarize features of the sampled schools, focusing on the percentage of students scoring at the proficient/advanced levels, the average expenditures per student, and the percentage of economically disadvantaged students. Reference appropriate numerical and graphical summaries as needed.
 - How many schools have a student body that is greater than (or equal to) 50% white? Of the schools where less than 50% of students are white, describe the distribution of `largest_minority`.
 - Create a single graphical summary showing the relationship between percentage of students scoring at the proficient/advanced levels, percentage of economically disadvantaged students, and whether a school has a student body that is less than 50% white. Provide an informative description of what you see.
- b) (10 pts.) Examine the association between racial demographics and the percentage of economically disadvantaged students.
- Conduct a formal analysis comparing the percentage of economically disadvantaged students between schools where less than 50% of the students are white and schools where 50% or more of the students are white. Summarize the results, including reporting and interpreting an appropriate confidence interval. Check any necessary assumptions.
 - Among schools where less than 50% of the students are white, conduct a formal analysis comparing the percentage of economically disadvantaged students between schools where the largest minority group is African American versus those where the largest minority group is Hispanic. Summarize the results and check any necessary assumptions.
- c) (30 pts.) Use a modeling approach to estimate the association between the percentage of students scoring at the proficient/advanced levels and whether a school's student body is less than 50% white, adjusting for the following potential confounders.
- Why it would not be advisable to include both `class_size` and `math_class_size` in a model predicting `PA_perc`? Explain your answer.
 - Would you recommend including `number_of_students` in a model predicting `PA_perc`? Explain your answer.
 - Fit a model predicting `PA_perc` from `white_less50`, `math_class_size`, `attendance_rate`, and `student_teacher_ratio`. Interpret the slope coefficient for `white_less50`.
 - Add `econ_dis` to the model from part iii. and interpret the slope coefficient for `white_less50`. Explain the difference in the interpretation of the slope coefficient for `white_less50` in this model versus the model from part iii., using terms accessible to someone who has not taken a statistics course.
 - Is the model from part iii. or the model from part iv. preferable for understanding the relationship between the percentage of students scoring at the proficient/advanced levels and whether a school's student body is less than 50% white? Explain your answer.

- vi. Check the modeling assumptions for the model chosen in part v. Summarize the results.
- d) (6 pts.) Formally assess whether the association between the percentage of students scoring at the proficient/advanced levels and the percentage of students who are economically disadvantaged differs by whether less than 50% of the students are white, after adjusting for math class size, attendance rate, and student-teacher ratio. Summarize the results.
- e) (4 pts.) Consider a school with the following features: an average math class size of 20 students, attendance rate of 93%, and student to teacher ratio of 11, where 60% of students are African American and 45% of students are economically disadvantaged. Suppose that in 2018, 90% of students at this school scored at the proficient/advanced level on the Mathematics Grade 10 MCAS. Based on the model from part c) iv., is this percentage of high scorers considered unusual for schools with the same features? Explain your answer.
- f) (6 pts.) The results from these analyses will be discussed at a future meeting of the Racial Imbalance Advisory Council, which advises the Massachusetts Commissioner of Education and the Board of Education on matters related to providing access to effective educational programs for all students in the state regardless of race or socioeconomic class.

Prepare a short statement, no more than ten sentences long, summarizing the main findings of the analyses with respect to understanding whether these data show indication of poverty and/or race-based achievement gaps in Grade 10 Mathematics MCAS scores. Be sure to use language that is accessible to a general audience. Reference previous numerical results as needed.

PROBLEM 3: TRIAGE (80 pts. total)

An important problem in emergency medicine is the prioritization of high-risk patients. Traditional triage algorithms classify patients into categories based on vital signs (such as heart rate and level of consciousness) in addition to the patient's reason for seeking medical care: red (life-threatening), orange (seriously ill), yellow (ill), green (needs assessment), and blue (minor complaints). However, studies suggest that this system may suffer from low specificity, such that too many patients who are actually at low risk are sorted into high-risk categories; this can lead to increased waiting times for patients who are in urgent need of care.

Recent studies have attempted to improve traditional triage algorithms. One study was conducted in Norway to investigate whether information from a set of routine blood tests administered to almost all patients admitted to an emergency department could improve the prediction of mortality risk. The primary outcome in the study was 30-day mortality; i.e., death within 30 days of admission to the emergency department. The eight blood tests examined were c-reactive protein, potassium, sodium, hemoglobin, creatinine, leukocyte count, albumin level, and lactate dehydrogenase; the results of these tests are typically available within 15 minutes.

Data from 4,545 individuals with complete observations are in `mort.Rdata`; individuals classified as blue (minor complaints) were excluded from the study sample. The descriptions of the variables are as follows.

Variable	Description
triage	triage rating upon admission to the emergency department
age	age, age in years, rounded to the lowest integer
sex	sex, coded female and male
crp	c-reactive protein level in nmol/L
k	potassium level in mmol/L
na	sodium level in mmol/L
hb	hemoglobin level in mmol/L
cre	creatinine level in $\mu\text{mol/L}$
leu	leukocyte count in 10^9 cells/L
alb	albumin level in g/L
ldh	lactate dehydrogenase level in units/L
mort30	coded 1 if died within 30 days, 0 otherwise

Use these data to answer the following questions.

a) (10 pts.) Explore the data.

- Briefly summarize features of the study participants, focusing on triage classification category, 30-day mortality, and the demographic variables age and sex. Reference appropriate numerical and graphical summaries as needed.
- For some blood test measures, both low and high values may indicate higher risk of mortality. Briefly explain why using such a measure as a predictor for mortality risk could potentially introduce a modeling issue and describe a possible method for handling the issue based on ideas discussed in the course.

Note: Proceed with the rest of the problem without implementing the described method.

- Compute and interpret the relative risk of 30-day mortality for a patient classified as red versus one classified as orange.

- b) (6 pts.) Conduct a formal analysis investigating whether there is an association between triage classification and 30-day mortality. Summarize the results in language accessible to someone who has not taken a statistics course. Be sure to check any necessary assumptions.
- c) (12 pts.) Investigate associations with the two demographic variables, age and sex. For each of the following parts, briefly justify your choice of analysis approach and check any necessary assumptions. Summarize the results.
 - i. Assess whether older individuals are at greater risk of 30-day mortality.
 - ii. Assess whether risk of 30-day-mortality is independent of sex.
 - iii. Assess whether triage classification is associated with age.
- d) (10 pts.) Fit a model estimating the association between 30-day mortality and triage classification, adjusting for age and sex.
 - i. Interpret the model coefficient for age.
 - ii. Interpret the model coefficient for triagegreen.
 - iii. Compare the estimated odds of 30-day mortality for a 55-year-old female categorized as orange to those of a 70-year-old male categorized as yellow. Are either more likely to die within 30 days than survive? Explain your answer.
 - iv. Is this model a better parsimonious model than a model with only triage classification as a predictor? Explain your answer.

For the remaining questions, consider the following three candidate models for predicting 30-day mortality:

- Model 1: triage classification
 - Model 2: the eight blood test measures, age, sex
 - Model 3: triage classification, the eight blood test measures, age, sex
- e) (14 pts.) Investigate the association between 30-day mortality and risk category, comparing the triage classification system based on vital signs versus the proposed system based on blood test measures and demographic information.
 - i. Compare how individuals who died within 30 days were classified under the existing triage system versus individuals who survived.
 - ii. Based on the predicted risks of 30-day mortality according to Model 2, categorize individuals into risk groups: green for less than or equal to 1% risk, yellow for greater than 1% and less than or equal to 10% risk, orange for greater than 10% and less than or equal to 25% risk, and red for greater than 25% risk.

Compare how individuals who died within 30 days were classified based on Model 2 versus individuals who survived.
 - iii. Based on the results from parts i. and ii., discuss which system seems more preferable for classifying patients. Explain your answer.

f) (14 pts.) Randomly split the data into a training set comprising 80% of the data and a testing set comprising 20% of the data. Fit the three candidate models on the training set then compute predicted risk probabilities for the test set.

- i. Describe Type I and Type II error in the context of these data.
- ii. Compute the Type I and Type II error rates on the test set, based on a 0.10 cutoff; i.e., consider patients with 30-day mortality risk greater than 0.10 to be predicted as dying within 30 days.

Based on these findings, which model is preferable for classifying patients? Summarize the findings and explain your reasoning.

- iii. Suppose the cutoff for predicting mortality within 30 days were lowered to 0.05; i.e., if a patient has predicted mortality risk greater than 0.05, they are predicted as dying within 30 days. How does this change in cutoff value affect Type I and Type II error rates? Explain your answer in language accessible to a non-statistician.

g) (8 pts.) The Brier score is a way to measure prediction accuracy based on comparing the predicted probability of an outcome to whether the outcome occurred or not:

$$\frac{1}{n} \sum_{i=1}^n (\hat{p}_i - y_i)^2,$$

where \hat{p} is the predicted probability of an event occurring, y_i is the observed response, and n is the sample size. Note that y_i can only take on two values: 1 if the event occurred and 0 if the event did not occur.

A Brier score of 0 represents the highest possible prediction accuracy, while a value of 1 represents complete inaccuracy.

For example, consider a sequence of three days where rain was predicted with probability 0.40, 0.90, and 0.80, respectively; it did not rain on the first day and it rained on the next two days. The Brier score for this set of predictions is:

$$\frac{1}{n} \sum_{i=1}^n (\hat{p}_i - y_i)^2 = \frac{1}{3} [(0.40 - 0)^2 + (0.90 - 1)^2 + (0.80 - 1)^2] = 0.07$$

- i. Using a 5-fold cross-validation approach, compute the average Brier score for each model. Based on this metric, which model is the most accurate?
- ii. A naive model for prediction predicts the risk probability for all cases as equal to the event prevalence; e.g., if 5% of individuals die within 30 days, then the naive model predicts that the risk of 30-day mortality is 0.05 for all individuals.

The average Brier score (from a 5-fold cross-validation approach) for the naive model applied to these data is 0.0467. Comment on whether the most accurate model from part i. is an improvement over the naive model.

h) (6 pts.) Based on the analyses from parts e) - g), discuss whether the information contributed by blood tests seem valuable for the risk classification of patients admitted to the emergency department. Reference previous numerical results as necessary. Limit your answer to no more than eight sentences.