Spring 2020 Midterm Examination

Statistics 104

Due 09 March 2020, 4:00 PM

- The exam consists of 4 problems and 8 pages. The exam is worth 200 points, and the point value for each question is displayed.
- The exam covers material from Units 1, 2, and 3.
- Your solutions are due at 4:00 pm, 09 March 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_midterm_spring2020_m_parzen.
- Submit a printed, single-sided copy of your PDF by 5:00 pm, 09 March 2020. The printed PDF must not differ from the one submitted to the course website. You may use either of the two dropboxes labeled "Stat 104: Take-Home Exam"; the dropboxes are outside the Statistics 300 suite on the 3^{rd} floor of the Science Center.
- Before submitting the printed copy of the exam, read and 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, section materials, problem set solutions, and the text *Introductory Statistics for the Life and Biomedical Sciences* (*OI Biostat*). 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.
- 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 come to office hours or contact the teaching staff via email.

PROBLEM 1: SHORT ANSWER *(46 pts. total)*

a) *(14 pts.)* Under a ketogenic diet, consumption of carbohydrates is severely limited while fat intake occurs in high amounts, with as much as 90% of daily calories coming from fats. This diet triggers the body to begin breaking down stored fat in the liver to produce energy rather than metabolizing circulating blood sugar.

Some studies have suggested that ketogenic diets might affect bone health. A study was conducted to examine bone health in athletes adhering to a ketogenic diet versus a traditional high-carbohydrate diet while undergoing intense training for upcoming competitions. Prior to the study, all participants were on a high-carbohydrate diet. Participants were allocated into diet groups based on preference; about half the participants chose to try a ketogenic diet. For four weeks, the study participants followed study diets developed by the research team; the two treatment diets provided equal amounts of calories and differed only in the ratio of carbohydrate versus fat intake.

The researchers assessed bone health by measuring the concentration of biomarkers related to bone formation and bone breakdown. Researchers found that while the biomarkers of participants in the high-carbohydrate group showed little change over the course of the study, those of participants on the ketogenic diet showed signs indicative of impaired bone health: higher concentration of biomarkers related to bone breakdown and lower concentration of biomarkers related to bone growth.

i. In the paper reporting the results of the analysis, the authors claim "We have shown that a 4-week ketogenic diet in athletes has negative effects on bone health during periods of high intensity exercise". Write a short response to the publication editor explaining clearly why the claim is potentially misleading. Be sure to use language accessible to a general audience without a statistics background. Limit your answer to at most five sentences.

ii. Suppose that a friend of yours has been on a ketogenic diet for three months while training for a marathon. While he found it challenging initially to adapt to the high-fat diet, he is now accustomed to it and thinks the diet regimen has helped boost his energy levels. Your friend reads about the study and wonders whether the results imply that it might be advisable to switch to a high-carbohydrate diet for the remaining month before the marathon takes place.

You are not convinced the study results suggest that your friend should consider stopping the ketogenic diet. Explain two reasons supporting your viewpoint. Limit your answer to at most eight sentences.

b) *(8 pts.)* Fasting blood glucose (FBG) levels can be used to diagnose whether someone has Type 2 diabetes. Among non-diabetic individuals, FBG levels are approximately normally distributed with mean 100 mg/dL and standard deviation 12 mg/dL.

i. What FBG level is needed to be in the upper 1% of the distribution for non-diabetic individuals?

ii. An individual is classified as pre-diabetic if two consecutive FBG measurements are between 100 and 125 mg/dL. Calculate the probability of a non-diabetic being classified as pre-diabetic, under the assumption that these two consecutive measurements are independent.

c) *(6 pts.)* Consider the scenario of repeatedly rolling a 6-sided die. Starting with roll $i = 1$, let R_i denote the result of roll i. If $R_i > i$, the die is rolled again; if $R_i \leq i$, the die is not rolled again. Let N denote the number of times the die is rolled.

Calculate $P(N > 3)$.

d) *(18 pts.)* Suppose that at the beginning of March, a new strain of coronavirus, Coronavirus Disease 2020 (COVID-20), emerges and begins to spread through Harvard College undergraduates. The symptoms of COVID-20 are similar to the flu and other respiratory illnesses. It is thought that the infection rate of COVID-20 will be similar to that of COVID-19 in Wuhan, but that the disease will be much less deadly.

COVID-20 is expected to infect about 20 out of every 1,000 undergraduate students per month for the rest of the semester, while the flu is expected to infect about 30 out of every 1,000 undergraduates per month for the rest of the semester. Assume that this rate remains the same for each of the remaining months in the semester (March, April, and May). There are currently 6,800 Harvard College undergraduates on campus.

You may assume that infection with COVID-20 is independent of infection with the flu.

i. You and your two roommates are concerned about becoming ill, either with flu or with COVID-20. Calculate the probability that all of you will successfully avoid contracting either flu or COVID-20 over the next month (i.e., by the end of March).

ii. Comment on two assumptions required to make the calculation in part i. and explain whether they are reasonable. Limit your answer to at most five sentences.

iii. Calculate the total expected number of COVID-20 cases among Harvard College undergraduates over the next three months.

iv. If more than 150 cases of COVID-20 are observed among the undergraduates during the next month, the University will go into shut-down mode and cancel class meetings. What is the probability the University goes into shut-down mode by the end of March?

PROBLEM 2: RETIREMENT PLANNING *(38 pts. total)*

Congratulations! You have graduated from college and are starting your first job. Financial experts advise everyone to start saving for retirement as soon as possible; starting early and increasing contributions gradually makes it much easier to have enough money to spend in retirement.

At age 22, you decide to start saving for retirement.

Write a simulation to investigate 40 years of saving for retirement, with the following guidelines:

- Add $2,000 to your retirement savings account in the first year.
- Increase the amount you contribute by 3% each subsequent year.
- Invest money in an 80/20 portfolio of stocks and bonds.
  - Invest 80% in the S&P500.
  - Investigate 20% in bonds.

Run the simulation for 1,000 replicates to model 40 years of time. The file `stocksbonds.csv` contains data on the historical yearly returns of the S&P500 from 1928 to 2018, along with data on the historical yearly returns of bonds during the same time period.

a) *(8 pts.)* Write the simulation according to the above guidelines; be sure to clearly comment your code. In one paragraph, briefly explain the general logic and organization of your simulation; in other words, provide a high-level overview of the steps in your simulation.

b) *(4 pts.)* Based on the simulation from part a), what is the 90^{th} percentile of returns after 40 years of investing?

c) *(4 pts.)* Someone mentions that the 80/20 scheme is too conservative and you should be 100% in stocks. Rerun the simulation under this scenario; what is the 90^{th} percentile of returns after 40 years of investing?

d) *(8 pts.)* You read some investment advice and learn about a common rule of thumb: one should invest their age (as a percentage) in bonds and the rest in stocks, increasing the percentage invested in bonds by 1% each year. A friend suggests that instead, you start at 100% in stocks at age 22 and go down 2% per year with the rest in bonds (e.g., in the second year, put 98% in stocks and 2% in bonds). Under this scenario, what is the 90^{th} percentile of returns after 40 years of investing?

e) *(6 pts.)* Compare the three investing schemes discussed in parts a) through d). Which investing scheme do you think is most advisable? Explain your answer, referencing results from the simulations as needed.

f) *(8 pts.)* Suppose you decide to invest 100% in the S&P500 starting at age 22 years. Once you reach 62 years old, you decide to cash out completely from the stock market and stop investing. Suppose that you spend $50,000 a year on living expenses and that these expenses increase by 3% each year. What is the estimated probability that you are left with no money 20 years later (i.e., when you are 82 years old)?

PROBLEM 3: GENETICS OF AUSTRALIAN CATTLE DOGS *(44 pts. total)*

Australian cattle dogs are known to have a high prevalence of congenital deafness. Deafness in both ears is referred to as bilateral deafness, while deafness in one ear is referred to as unilateral deafness.

Deafness in dogs is associated with the white spotting gene S that controls the expression of coat and eye pigmentation. The dominant allele S produces solid color, while the three recessive alleles contribute to increasing amounts of white in coat pigmentation: Irish spotting (s^i), piebald (s^p), and extreme white piebald (s^w). The s^p and s^w alleles are responsible for the distinctive Australian cattle dog coat pattern of white hair evenly speckled throughout either a predominantly red or black coat. The dogs are born with white coats, and the speckled pattern develops as they age.

While all Australian cattle dogs have some combination of the s^p and s^w alleles, the gene displays incomplete penetrance such that individuals show some variation in phenotype despite having the same genotype. Individuals with low penetrance of the alleles tend to have additional patterns on their coat, such as a dark "mask" around one or both eyes (in other words, a unilateral mask or a bilateral mask). High penetrance of the piebald alleles is associated with deafness.

Suppose that 40% of Australian cattle dogs have black coats; these individuals are commonly referred to as "Blue Heelers" as opposed to "Red Heelers". Among Blue Heelers, 35% of individuals have bilateral masks and 25% have unilateral masks. About 50% of Red Heelers exhibit no eye masking and 10% have bilateral masks.

Let M represent the event that an Australian cattle dog has a facial mask, where M_2 represents a bilateral mask, M_1 represents a unilateral mask, and M_0 indicates lack of a mask.

a) *(4 pts.)* Calculate the probability an Australian cattle dog has a facial mask and a black coat.

b) *(4 pts.)* Calculate the prevalence of bilateral masks in Australian cattle dogs.

c) *(4 pts.)* Among Australian cattle dogs with bilateral facial masks, what is the probability of being a Red Heeler?

d) *(32 pts.)* Unilateral deafness occurs in Red Heelers with probability 0.15, in both dogs that either lack facial masking or exhibit a unilateral mask; for both unmasked and unilaterally masked Red Heelers, 60% of dogs are not deaf. The overall prevalence of bilaterally masked Australian cattle dogs with bilateral deafness and red coats is 1.2% and the overall prevalence of bilaterally masked Australian cattle dogs with unilateral deafness and red coats is 4.5%; these prevalences are the same for Australian cattle dogs with black coats. Among Blue Heelers with either no facial masking or a unilateral mask, the probability of unilateral deafness is 0.05 and the probability of bilateral deafness is 0.01.

   Let D represent the event that an Australian cattle dog is deaf (i.e., deaf in at least one ear), where D_2 represents bilateral deafness and D_1 represents unilateral deafness.

   i. What is the probability that an Australian cattle dog has a bilateral mask, no hearing deficits, and a red coat?

   ii. Calculate the proportion of bilaterally masked Blue Heelers without hearing deficits.

   iii. Compare the prevalence of deafness between Red Heelers and Blue Heelers.

   iv. If a dog is known to have no hearing deficits, what is the probability it is a Blue Heeler?

PROBLEM 4: SELF-MANAGEMENT ***(72 pts. total)***

Chronic conditions are diseases that last one year or more and require ongoing medical attention, such as diabetes and cancer. The growing number of patients with chronic conditions has significant economic implications; according to the Centers for Disease Control, diabetes costs the United States health care system and employers $237 billion annually.

A strategy for reducing the burden on health care systems is promoting patient self-management, in which patients proactively take action to maintain and improve their own health. For patients with a chronic condition, daily maintenance constitutes a form of prevention that can lead to better quality of life and reduce health care costs. In individuals with diabetes, for instances, adhering to recommended diet and exercise regimens can help reduce the risk of serious complications like heart disease or stroke.

Self-management is not appropriate for all patients. A study was conducted among 1,154 adult patients from different regions in the Netherlands to identify characteristics associated with activation for self-management. Study participants had a clinical diagnosis of one of four chronic conditions: Chronic Obstructive Pulmonary Disease (COPD), Chronic Heart Failure (CHF), Diabetes Mellitus Type II (DM-II), or Chronic Renal Disease (CRD).

The primary outcome, activation for self-management, was assessed with the Patient Activation Measure (PAM-13), a questionnaire which assesses self-reported knowledge, skills, and confidence for self-management. Score on the PAM-13 ranges from 0 to 100, with a higher score indicating a higher level of self-management activation. Scores can be classified into one of four levels: Level 1 ($\leq$ 47.0 points), Level 2 (47.1 – 55.1 points), Level 3 (55.2 – 67 points), and Level 4 ($\geq$ 67.1 points). A Level 1 patient demonstrates lack of knowledge and confidence to manage their condition. A Level 2 patient demonstrates a limited amount of knowledge and ability to set goals. Both Level 1 and Level 2 patients typically believe that their health is largely out of their control and are considered passive recipients of care. In contrast, a Level 3 patient is goal-oriented and actively working to achieve best practice behaviors. A Level 4 patient has successfully maintained new behavior over time.

In addition to the PAM-13, participants completed assessments measuring health status, anxiety and depression, and perceived social support.

- *Health Status*. Higher scores on the Short Form-12 Health Survey (SF-12) are indicative of better health status; possible scores range from 0-100 points. The survey consists of two summary scores on the 0-100 scale, one measuring physical health and one measuring mental health.
- *Anxiety and Depression*. Higher scores on the Hospital Anxiety and Depression Scale (HADS) indicates a higher state of anxiety or depression. Scores range from 0-21 points for each component, and a score of 11 or greater suggests the presence of an anxiety or depressive disorder.
- *Social Support*. Higher scores on the Multidimensional Scale of Perceived Support (MSPSS) indicates higher perceived support, such as from family, friends, and significant others. Scores range from 0-84 points.

Information was also collected on participant sociodemographics, such as level of financial distress and educational level. Educational level was divided into lower (primary school through vocational training), middle (secondary school), and high (college or university) education, according to the Dutch school system.

Data from the study are in the file `self_manage.Rdata`. The following table provides a list of the variables in the dataset and their descriptions.

Variable	Description
`sex`	sex, coded `female` or `male`
`height`	height in centimeters (cm)
`weight`	weight in kilograms (kg)
`financial`	financial distress, either `none`, `low`, or `high`
`age`	age in years
`bmi`	body mass index (kg/m^2)
`pam.score`	PAM-13 score, on 0-100 point scale
`pam.cat`	PAM level, from Level 1 to Level 4
`sf.total`	SF-12 total score, on 0-100 point scale
`sf.phys`	SF-12 score for physical health, on 0-100 point scale
`sf.ment`	SF-12 score for mental health, on 0-100 point scale
`has.depress`	HADS score for depression, on 0-21 point scale
`has.anxiety`	HADS score for anxiety, on 0-21 point scale
`supp.total`	MSPSS score for perceived social support, on 0-84 point scale
`disease`	chronic disease, either `DM-II`, `COPD`, `HF`, or `CRD`
`edu`	educational level, either `lower`, `middle`, or `higher`
`duration`	disease duration
`severity`	disease severity, either `mild`, `moderate`, or `severe`

Use the data to answer the following questions.

a) *(6 pts.)* Write an informative summary describing features of the study participants with respect to the variables age, sex, and type of chronic disease. Reference appropriate numerical and graphical summaries as needed. Limit your answer to no more than five sentences.

b) *(6 pts.)* The main measurement of interest is activation for self-management. Describe the distribution of activation for self-management within the study participants, both in terms of PAM-13 score and PAM level. Reference appropriate numerical and graphical summaries as needed. Limit your answer to no more than five sentences.

c) *(12 pts.)* Explore the relationship between activation for self-management and disease type.

   i. Create a plot to graphically show the association between disease type and PAM-13 score. Describe what you see.

   ii. Create a summary that shows how the distribution of PAM level differs between disease type. Describe what you see.

   iii. Do you find the summary from part i. or the summary from part ii. more informative with regards to understanding the relationship between activation for self-management and disease type? Explain your answer in no more than five sentences.

d) *(10 pts.)* Is PAM-13 score associated with perceived level of social support?

   i. Using graphical and numerical methods, investigate this question separately within disease types. Summarize your findings, with particular focus on whether the association between PAM-13 score and perceived level of social support seems to differ between types of chronic disease.

   ii. Do these data suggest that an increase in (perceived) social support leads to better capacity for self-management? Explain your answer.

e) *(8 pts.)* Investigate the relationship between PAM-13 score and educational level.

   i. With reference to appropriate numerical and graphical summaries, describe the association between PAM-13 score and educational level.

   ii. Propose one possible explanation for the trends observed in part i. Limit your answer to no more than five sentences.

f) *(14 pts.)* Investigate the relationship between PAM level, age, and educational level.

   i. Create a graphical summary that shows the association between age and PAM level. Describe what you see.

   ii. Create graphical summaries that show the association between age and PAM level when comparing individuals of the same educational level. Describe what you see.

   iii. A news outlet is interested in reporting on the study results. You have been asked to address the following question: "Do these data suggest that older individuals tend to have a lower level of activation for self-management?"

   Based on the findings from parts i. and ii., address the question in language accessible to a non-statistician. Limit your answer to no more than five sentences.

g) *(16 pts.)* Explore the relationship between depression and activation for self-management.

   i. From these data, compare the risk (i.e., probability) of being PAM Level 1 for individuals classified as having depressive disorder versus those not classified as having a depressive disorder.

   ii. Calculate the difference in mean PAM-13 score between individuals classified as having a depressive disorder versus those not classified as having a depressive disorder. In one sentence, report the calculation; use phrasing that is informative for a general audience.

   iii. There are some individuals missing responses for the HADS questionnaire. Does this missingness represent a potential source of bias for the calculations in parts i. and ii? Explain your answer.