Name: _____

Biostatistics 1st year Comprehensive Examination: Applied Take-Home Exam

Due June 9th, 2016 by 5pm. Late exams will not be accepted.

Instructions:

1. ***This is exam is to be completed independently. Do not discuss your work with anyone else.***
2. There are 2 questions and 3 pages.
3. Answer each question to the best of your ability. Read the exam carefully.
4. Be as specific as possible and type up or Latex your answers.
5. *This is a take-home examination. You may consult books, notes, and papers. You may use the Internet as a research resource. However, you may not consult or discuss this exam with another human being, directly or indirectly, nor may you seek help from another individual on the internet (e.g., no posting questions to chat rooms or message boards).*
6. If you have any questions, please contact Professor Blume by email or by phone (cell: 615-545-2656). Texting is fine as well. Do not worry about being polite. Contact Professor Blume as needed; call for emergencies.
7. Turn in your exam by emailing it to Professor Blume at j.blume@vanderbilt.edu **AND** Amanda Harding at amanda.harding@vanderbilt.edu. Your exam is not submitted until Professor Blume or Ms. Harding confirm your exam was received. Alternatively, you may turn in a hard copy to either person by the deadline. *If you do not receive confirmation you should assume that your exam has not been received.*
8. ***Vanderbilt's academic honor code applies; adhere to the spirit of this code.***

Link to this exam:
https://dl.dropboxusercontent.com/u/25204698/Comps/Comp2016_1yr_ATH_Final.pdf

Question	Points	Score	Comments
1	100		
2	100		
Total			

1. Dr. Picardo's tri-fitness score is a continuous scalar summary measure of heart, lung, and circulatory health. Let S_i represent the i^{th} patient's score for $i = 1, \dots, n$. A random sample of n patients is collected, say $S_1, \dots, S_n$ and let $\overline{S}$ represent the sample mean. Let $z_{1-\alpha}$ represent the standard normal quantile with $1 - \alpha$ area to the left (lower tail).

   Assume that the distribution of scores is normal such that $S \sim N(\mu = 100, \sigma^2 = 100)$.

   a. Consider the interval $\left(\overline{S} \pm z_{0.975} * 10/\sqrt{n}\right)$. **Prove** that the probability that this interval contains μ is 0.95 for any n.

   b. What is the $1/8^{th}$ *likelihood support interval* for μ when $n = 7$? Use **numerical methods to find** the coverage probability of this interval. Be sure to retain a numerical accuracy of within ± 0.0025 for the coverage probability.

   c. **Explain** the relationship between the number of simulations, the accuracy of your numerical approximation, and the coverage probability.

   Now assume that the scores follow a shifted exponential distribution such that $S = U + 90$ where $U \sim Exp(1/10)$.

   d. Reconsider the interval $\left(\overline{S} \pm z_{0.975} * 10/\sqrt{n}\right)$. Use **numerical methods to find** the coverage probability of this interval when $n = 7$. Be sure to retain a numerical accuracy of within ± 0.0025 for the coverage probability.

   e. Make a **recommendation** for the smallest sample size n that would ensure the coverage probability is at least ≥ 0.94. Be sure to retain a numerical accuracy of within ± 0.0025 for the coverage probability. Justify your recommendation.

   f. **Propose an interval or procedure** that will result in good coverage properties when $n \geq 7$ and the variance is *unknown*. If you are not able to find a satisfactory interval or procedure, detail the methods you tried, the results, and the smallest sample size n at which those methods work.

   g. Suppose you are able to directly observe $U_1, \dots, U_n$. **Derive** the $1/8^{th}$ *likelihood support interval* for $\mu = E[U_i]$ when $n = 7$. Use numerical methods to find the coverage probability of this interval. Be sure to retain a numerical accuracy of within ± 0.0025 for the coverage probability. Suggest a way to make inference about $E[S_i]$ using this new interval.

2. Analysis of Prostate Cancer Data

Objective: Investigate if prostate-specific antigen (PSA) and/or Gleason score can be used to predict whether the tumor has penetrated the prostatic capsule. Questions to be answered are listed on the next page.

Background: Prostate cancer is cancer that begins in tissues of the prostate gland. The prostate capsule is the membrane that surrounds the prostate gland. As prostate cancer advances, the disease may extend into the capsule (extraprostatic extension) or beyond (extracapsular extension) and into the seminal vesicles. Capsular penetration is a poor prognostic indicator, which accounts for a reduced survival expectancy and a higher progression rate following radical prostatectomy.

PSA is an enzyme produced in the epithelial cells of both benign and malignant tissue of the prostate. PSA is used as a tumor marker to determine the presence of prostate cancer because a greater prostatic volume, associated with prostate cancer, produces larger amount of prostate-specific antigen. PSA is measured in samples of blood.

After a prostate biopsy, a pathologist examines the samples of prostate cancer cells to see how the patterns, sizes, and shapes are different from healthy prostate cells. The pathologist assigns a Gleason grade to the most common pattern of prostate cancer cells and assigns a second Gleason grade to the second-most common pattern of prostate cancer cells. These two Gleason grades indicate the cancer's aggressiveness, which in turn indicates how quickly prostate cancer may extend out of the prostate gland. The *Gleason score* is the sum of the two Gleason grades: Gleason 1 + Gleason 2.

Table of variables in data set

Variable	Description	Codes/Values	Name
1	Identification Code	1 to 380	id
2	Tumor Penetration of Prostatic Capsule	0=No Penetration 1=Penetration	capsule
3	Age	Years	age
4	Race	1=White 2=Black	race
5	Results of the Digital Rectal Exam	1= No Nodule 2= Unilobar Nodule (Left) 3= Unilobar Nodule (right) 4=Bilobar Nodule	dpros
6	Detection of Capsular Involvement in Rectal Exam	1=No 2=Yes	dcaps
7	Prostatic Specific Antigen Value	mg/ml	psa
8	Tumor Volume Obtained from Ultrasound	cm^3	vol
9	Total Gleason Score	0 to 10	gleason

Investigative Questions:

1. Summarize the variables in the data set. Use figures and/or tables as necessary.

2. Consider PSA alone to predict capsular penetration.
   a. Do you recommend transforming *psa* prior to analysis? Explain.
   b. What is the odds ratio (provide a point estimate and confidence interval) associated with an increase of 10 mg/ml? Provide an interpretation.
   c. For what range of PSA values would you predict capsular penetration? How does this prediction perform?

3. Consider Gleason score alone to predict capsular penetration.
   a. The Gleason score takes on discrete, but ordinal, values. Should this affect your analysis? Explain.
   b. What is the odds ratio (provide a point estimate and confidence interval) associated with an increase of 1 unit in the score? Provide an interpretation.
   c. For what range of Gleason values would you predict capsular penetration? How does this prediction perform?

4. Consider PSA and Gleason score together to predict capsular penetration.
   a. What is the odds ratio (provide a point estimate and confidence interval) associated with an increase of 10 mg/ml in PSA? Provide an interpretation.
   b. Do these measures provide a better prediction model when used together? Explain.
   c. Provide an index measure (function of *psa* and *gleason*) and related cutpoint for predicting capsular penetration? How does this prediction perform?

5. Now consider the remaining covariables. Does controlling for all or a subset of these result in an improved prediction model? Explain. What model do you recommend?

6. A 68 year-old black man presents with a PSA of 17.1 mg/ml and a Gleason score of 9. His tumor volume measured at 35 cm^3. The digital rectal exam found a bilobar nodule with no detection of capsular involvement. Using your prediction model, what is the estimated probability of capsular penetration (provide a point estimate and confidence interval)? What would you predict for this subject? Compare your result to the subject with *id=89*.

Directions: Answers the questions and write a brief summary of your findings from the model you recommended in part (5) (summary should be less than two paragraphs). Put all code and output in an appendix, except when explicitly requested in a question.

Data:
CSV file: https://dl.dropboxusercontent.com/u/25204698/Comps/prostate_study.csv
STATA file: https://dl.dropboxusercontent.com/u/25204698/Comps/Prostate_study.dta