ANLY511 Take-Home Final

Section II (Dr. Touyz)

December 9, 2017

200 points in seven problems plus a Bonus problem with 20 points. This is the take-home portion of the exam. You may use your notes, your books, all material on the course website, and your computer or any computer in the departmental computer lab. You may also use official documentation for **R**, built-in or on `https://cran.r-project.org/`, but no other material on the Internet. Provide proper attribution for all such sources. You may not use any human help, not in person and not in any electronic form or otherwise, except whatever help is provided by me.

Return your solutions by **Tuesday, 12/19/2017, 11:59PM,** by e-mail *as a single pdf file*, or hand in a paper copy. If you choose to hand in a paper version, be sure to keep a copy.

Unless stated otherwise, a complete solution consists of all code (suitably edited and commented), a new seed value before each simulation, plots as required, numerical results as required, and text explaining your solution and giving the conclusion. **R** Markdown is the best format.

Please do not hand in printouts of data that are provided to you - I will take off points if you do that.

Part I: Probability theory. Problems 1-4 are 20 points each.

1. Consider a continuous random variable with density $f(x) = c(A^2 - x^2)$ on the interval $[-A, A]$.

a) Find c such that this is actually a probability distribution.

b) Find the approximate distribution of the sample mean $\bar{X}$ for sample size n.

2. Consider data X_i from an Poisson distribution with rate λ. Suppose you are not observing the X_i, but instead "zero-inflated" data

$$Y_i = \begin{cases} 0 & (probability = p) \\ X_i & (probability = 1 - p) \end{cases}$$

where $0 < p < 1$ is given and known. *This could happen when the data are counted with equipment that malfunctions with probability p and returns a value of 0 in those cases.*

a) Write down the pmf of the Y_i in terms of the parameters p and λ.

b) Suppose we want to estimate λ from the Y_i and $p = \frac{1}{3}$ is known. Set up the likelihood function for the sample $0, 0, 1, 1, 2, 2, 2, 4, 4, 5$ and find the maximum likelihood estimate for λ to two decimal digits. *Use any method you want, Calculus or otherwise.*

3. Given a sample $S = \{x_1, x_2, \ldots, x_n\}$ of odd size n where all values in the sample are different. Let m be the sample median. Let X be the median of a bootstrap sample from S. This is therefore a random variable. Is it always true that $E(X) = m$? Explore this question with simulations and then give a careful theoretical answer.

4. Let $X \sim N(0, 1)$ and let $Y = (X^{-1})|X > 1$. Find $E(Y)$ and $var(Y)$ to two decimal digits by simulation. Use statistical arguments to explain why you are convinced that the answers are accurate. Use as few simulations as you can to achieve this accuracy, with explanation.

Part II: Statistic. Problems 5-7 are worth 40 points each.

For full credit for problems involving tests, you are required to set up the null and alternative hypothesis, define the test statistic and give its value, and state your conclusion.

Please use only the `NCBirths2004.csv` data from the Canvas website,

5. Use the `NCBirths2004` data. Are the age of the mother and the gestation (length of pregnancy) independent?

Investigate this in two ways: (a) with a permutation test, (b) with a χ^2 test, after combining suitable cells. Summarize your conclusions.

6. Use the `NCBirths2004` data. Are median birth weights different for smoking and non-smoking mothers? If so, what can you say about the difference?

Explore this question graphically and with suitable statistical procedures (test and confidence interval), and state your conclusion.

7. Use the `NCBirths2004` data. Find 95% confidence intervals for the weight gains from gestation week k to gestation week $k + 1$, for $k = 37$, 38, 39, 40, 41. Be sure to use reliable procedures in each case.

State the procedure(s) you are using and explain your choice.

Bonus question (20 points)

Use the `NCBirths2004` data. Is there evidence that tobacco use by the mother shortens the gestation length? Propose a suitable statistical approach, carry it out, and state your conclusion.