ANLY-511
Final Exam (December 9, 2015)
Take Home Portion

200 points in five problems plus 20 Bonus 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, except whatever help is provided by me.

Return your solutions by Monday, 12/21/15, 11:59PM, by e-mail **as a single pdf file**, or hand in a paper copy, or fax it to me at 202.687.6067. If you choose to hand in a paper version, be sure to keep a copy for yourself.

Unless stated otherwise, a complete solution consists of all code (suitably edited and commented), plots as required, numerical results, and text explaining your solution and conclusion.

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

All data are in the R workspace `final.rdata` that is available on Blackboard.

1. (40) Consider the vector `problem1` that is in the workspace. It was obtained by sampling 1000 values of a random variable $Y = X_1 + X_2$, where $X_1 \sim N(0, \sigma^2)$ and $X_2 \sim Poisson(\lambda)$. Also, X_1 and X_2 are independent. Here σ is an unknown **whole** number between 1 and 10 and λ is another unknown **whole** number between 5 and 20. Find σ and λ exactly, and explain your work. Properties of these distributions may be found in appendix B and on pages 396-398 of the textbook. *Hint: What are $\mathcal{E}(Y)$ and $var(Y)$?*

2. (40) Consider the data frame `handle.times` that is available in the workspace. It contains times to process individual customer calls (in seconds) from a call center, grouped by the nature ("program") of the call. *For example, `Nevada` identifies calls from customers from the state of Nevada, `LD Care` identifies calls from customers who are interested in a product with this name, etc..*
a) Make a 95% confidence interval for the mean time for calls for the

program "Nevada". Explain the choice of your method.
b) You are interested in the question whether calls in the program "West Billing" on average take less than 540 seconds. Formulate suitable hypotheses and test them with a suitable method.
c) You are interested in the difference of the mean times for calls in the programs "LD Care" and "Wild Blue". Make a 90% bootstrap confidence interval for this difference and explain why you chose your method.

3. (40) Given data from a Poisson distribution with unknown rate λ, you have to decide whether $\lambda = 10$ or $\lambda > 10$, based on a sample $(x_1, x_2, \ldots, x_{20})$ of size $n = 20$. Here are two possible tests to do this:
(i) Use the test statistic $T_1 = \sum_i x_i$. Reject the null hypothesis $H_0 : \lambda = 10$ if $T_1 \geq c_1$ for a suitable c_1.
(ii) Use the test statistic $T_2 = \max_i x_i$. Reject the null hypothesis $H_0 : \lambda = 10$ if $T_2 \geq c_2$ for a suitable c_2.
Use simulations to answer the following questions:

a) Find the smallest c_1 and c_2 such each test has type I error < 0.05.
b) Suppose that actually $\lambda = 15$. What is the type II error probability for each test?
Bonus: Suppose that in fact $\lambda = 8$. Which of the two tests has the smallest probability of rejecting H_0?

4. (40) The Pearson 2 skewness of a distribution that comes from a random variable X is defined as

$$Sk_2 = 3\frac{\mathcal{E}(X) - \mathcal{M}(X)}{\sigma(X)}$$

where $\mathcal{M}(X)$ is the median and $\sigma(X)$ is the standard deviation.

a) Compute the plug-in estimate of the Pearson 2 skewness for the data in the vector `problem4`, available in the workspace. Explain the formula and R code that you are using.
b) Make a 90% confidence interval for the Pearson 2 skewness for the data, using the bootstrap.
Bonus: Estimate the bias of the plug-in estimator for the Pearson 2 skewness for this data set, using the bootstrap.

5. (40) Consider the Bayes network given in the figure below. The random variables A, B, C, D can have values 0 or 1. All dependency arrows run from left to right. The data frame `problem5` in the workspace contains 1000 observations from the network (1000 row vectors with four entries from $\{0, 1\}$, one for each of the variables).

a) Estimate $\mathcal{P}(A = 0)$ and $\mathcal{P}(A = 1)$.
b) Estimate $\mathcal{P}(B = 1|A = 1)$ and $\mathcal{P}(B = 1|A = 0)$ from the data.
c) Show with a test that B and C are indeed independent.
d) Are perhaps A and D also independent? Investigate with a suitable test.
e) Estimate $\mathcal{P}(D = 1|B = 1, C = 0)$ from the data.
Estimation means just that - come up with a number. No confidence intervals are needed.

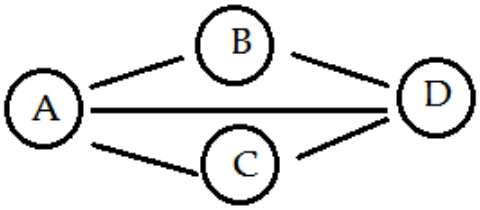

FIGURE 1. Bayes network for problem 5