MA206 - Probability & Statistics
Homework Assignment 1

Branch CQPA Comparison

General:

Thus far in MA206 we have studied methods of describing and visualizing data. A basic application of statistics is the use of a sample to describe the behavior of a population. In this assignment you will look at the same set of data while applying different filters in an effort to describe or explain a certain behavior. Your class will examine the academic, military and physical performance of firsties who chose certain branches. The accompanying EXCEL spreadsheet contains data for the USMA class of 2008.

Administrative Details:

1. The homework is an individual assignment and is worth 40 points.
2. The homework is due NLT 1600 hours on 23 January 2009. Place your completed assignment in the box located outside your instructor's office or bring to class that day.
3. The homework must consist of a cover sheet, a short write-up (2-4 pages), and the appropriate documentation. Brown bombers are not required, although staples are.
4. All EXCEL model building must be submitted electronically to your instructor prior to the homework due date. Your EXCEL file will NOT be graded. All of your answers (including graphs) need to be contained in the homework write-up. Name your EXCEL file as follows:
   SECTION_LASTNAME_FIRSTNAME_HW1.xls (i.e. B9_Patton_Hannible_HW1.xls)
   The instructor's goal is not to open any EXCEL files: if you're in big trouble and your instructor wants to throw you a couple partial credit points, your instructor may open your EXCEL file.
5. You are required to document any assistance that you received in accordance with *Documentation of Written Work*. While collaboration is allowed, you must specify in detail the level of help you receive. You may lose points for excessive collaboration. Remember the policy addendum to DWW!
6. If you receive an EXCEL model electronically from a classmate, you must specify who you received the model from and what portion(s) of that model you used in your assignment. Be very specific! You are STRONGLY discouraged from sharing electronic files!

Part I: Data Analysis (20 Points)

Follow the steps below using the data provided to analyze and describe the differences, if any, between cadets who branched Infantry and the rest of the class in terms of final CQPA. CQPA (Cumulative Quality Point Average) represents a cumulative grade of a cadet's performance including academic, military and physical performance. Keep in mind that real life data does not always look perfect. Some of the entries for branch or firstie CQPA may be missing because the cadet failed at some point and did not graduate, or studied overseas for a semester and at the time of the drawing of this data, did not have a CQPA yet or possibly left the academy for some reason sometime between their plebe year and firstie year. Other entries will have a CEER score but be missing all other information. These cadets left the academy sometime between showing up on R day and the end of their plebe year. You should delete these entries in order to clean the data.

1. Develop <u>a table</u> of descriptive statistics for CQPA for three different filters: the entire class of 2008, firsties who branched Infantry, and all firsties that did not branch Infantry. Your descriptive statistics should include at a minimum: sample size, maximum, minimum, mean, median, and standard deviation. Ensure that you clean the data by removing observations that are no longer members of the class of 2008.

2. Discuss the meaning of each of the descriptive statistics calculated above. What do these statistics tell you about the metrics for the class as a whole, firsties who branched Infantry and firsties who did not branch Infantry?

3. One might think that since Infantry is so sought after, the average CQPA of cadets choosing that branch would be higher than the rest of their classmates or of the class as a whole. Evaluate and discuss the validity of this claim citing evidence from your table of descriptive statistics to support your position.

Part II: Modeling the EDF (20 Points)

Follow the steps below to select an appropriate distribution to model the CEER Score for the class of 2008. The CEER score is assigned during application to West Point. It is a score that takes into account your high school class ranking, your high school GPA and your standardized test scores (SAT or ACT). This score evaluates your academic potential here at the academy and is used to create your whole candidate score (WCS) that allows the admissions department to quantitatively assess and compare applicants.

1. Create an Empirical Distribution Function for the CEER Scores for the class of 2008 that <u>entered</u> USMA (i.e. include those observations that are no longer members of the class of 2008 or don't have branch assignments).

2. Assume that one of the following two functions would best model the data you've been provided:

$$F(x) = 1 - e^{-\lambda x} \qquad G(x) = 1 - e^{-(\lambda x)^{\kappa}}$$

   Complete the following tasks and provide the requested information:

   - Fit *F(x)* onto your EDF and ***provide the value of*** λ that minimizes the SSE. What is the minimized sum of squared error when using *F(x)*?

   - Fit *G(x)* onto your EDF and ***provide the values of*** λ ***and*** κ that minimize the SSE. What is the minimized sum of squared error when using *G(x)*?

   - Provide one graph of the EDF with the optimal *F(x)* and *G(x)* curves superimposed on the EDF.

- Label graph axes and include legend

3. Based on your graphs and your answers to the questions above, which function fits better and why? Your answer should be in mathematical terms. Provide numerical and analytical support.

4. You learn that you were in the 70th percentile of CEER Scores. Estimate your CEER Score and explain your work. Interpret what this percentile means with respect to the rest of CEER Scores.