

Project # 1 (Due on September 27, 2018)

- For starting seed value, use the last 4-5 digits of your ID.
 - This is an individual project rather than a group project. You may discuss it with others but you must write your own code and answers. If the submitted report (including code and answer) is similar (either partially or fully) to someone else's, this will be considered evidence of academic dishonesty, and you will be referred to appropriate university authorities.
 - Your project report must be typed. It must consist of answers to the specific questions asked, appropriate justifications for the statistical methods used, annotated **SAS** code, and at most 6 pages of relevant parts of SAS output with relevant numbers highlighted. You may not get any credit if you just dump code and output. Be sure to justify the choice of any statistical method you use and mention and verify any inherent assumptions.
 - You are welcome to ask me or TA questions. However, first try to find the answer on your own. Don't be afraid to google! It is a necessary skill for developing expertise in any programming language.
 - Some useful links:
 - <https://stats.idre.ucla.edu/sas/library/>
 - <https://stats.idre.ucla.edu/sas/library/sas-libraryoverview-of-the-sas-language/>
 - <https://stats.idre.ucla.edu/sas/library/sas-libraryoverview-of-sas-procedures/>
 - <http://www.ssc.wisc.edu/sscc/pubs/4-8.htm>
 - https://documentation.sas.com/?cdcId=pgmsascdc&cdcVersion=9.4_3.4&docsetId=allprodsproc&docsetTarget=procedures.htm&locale=en
-

1. (15 points) An electric utilities company must choose between two technology options for generating electricity for its customers: coal or nuclear energy. To assess the attitudes of local citizens, a power company conducted a public opinion poll. Four sectors were considered: news media, coal miners' union, environmentalists and conservationists, and local groups. Fifty persons were randomly sampled from each sector and asked to give their opinions. The results of the survey appear in the datafile **COALNUC.DAT** posted on eLearning. A **SAS** datafile of the same data — **coalnuc.sas7bdat** — is also posted.
 - (a) (7 points) Does public opinion regarding the choice of technology options for generating electricity differ among the four groups? Answer this question by performing an appropriate test at 5% level of significance.
 - (b) (5 points) Does there appear to be more overall support for the coal option rather than the nuclear option? Answer this question by performing an appropriate test at 5% level of significance.
 - (c) (3 points) Construct a 95% confidence interval for the percentage of environmentalists and conservationists who support the nuclear option and interpret the result.

2. (30 points) At the time of data collection, M&Ms plain chocolate candies came in six different colors: brown, yellow, red, orange, green, and tan. According to the manufacturer (Mars Inc.), the color ratio in each large batch was 30% brown, 20% yellow, 20% red, 10% orange, 10% green, and 10% tan. To test this claim, a professor at Carleton College (Minnesota) had students count the colors of M&Ms found in “fun-size” bags of the candy. The results for 370 M&Ms are provided in the table below. They are also available in datafiles `M&M.DAT` and `m.m.sas7bdat` posted on eLearning.

Brown	Yellow	Red	Orange	Green	Tan	Total
84	79	75	49	36	47	370

- (a) (10 points) Find p -value for an appropriate test of hypothesis to determine whether the true percentages of colors produced differ from the manufacturer’s stated percentages. Use the null distribution discussed in the class.
- (b) (17 points) Repeat (a) but this time compute p -value using a Monte Carlo approach. How does this p -value compare with the one obtained in (a)?
- (c) (3 points) State conclusion at 5% level of significance.
3. (15 points) Consider a Harris Corporation/University of Florida study to determine whether a manufacturing process performed at a remote location can be established locally. Test devices (pilots) were set up at both the old and new locations and voltage readings on 30 production runs at each location were obtained. The data are available in `VOLTAGE.DAT` and `voltage.sas7bdat` files. Perform an appropriate analysis of data. Does it appear that the manufacturing process can be established locally?
4. (15 points) The files `VAPOR.DAT` and `vapor.sas7bdat` provide data on the theoretical (calculated) and experimental values of the vapor pressures for dibenzothiophene, a heterocyclic aromatic compound similar to those found in coal tar. If the theoretical model for vapor pressure is a good model of reality, the true mean difference between the experimental and calculated values of vapor pressure will be zero. Perform an appropriate analysis of these data to see if there is sufficient evidence to indicate that the mean difference differs from zero.
5. (25 points) We know how to construct a large-sample confidence interval for a population proportion p . How large n should be for this interval to have acceptable accuracy? Answer this question by computing the coverage probability of this interval using Monte Carlo simulation and examining how close the probability is to the nominal confidence level. Take the level of confidence to be 95% but use a variety of values for n and p , e.g., $n = 5, 10, 30, 50, 100$, and $p = 0.05, 0.1, 0.25, 0.5, 0.9, 0.95$. Summarize your results. Comment on any patterns you see in the results. Based on your findings, what n would you recommend for the use of this confidence interval? Would your answer depend on p ? Explain.