Homework 4 *Probability & Statistics for DS & AI, Michele Guindani, 2022*

Topics: Estimation & Testing of Hypothesis — Due on August 24, 2022

Note: Attach your code with your submission.

1. **Statistical Inference: Estimation** [4+6+10=20 points]

   (a) The observations on number of hours of daily TV watching for the 10 subjects in the 2018 GSS who identified themselves as Islamic were $0, 0, 1, 1, 1, 2, 2, 3, 3, 4$.

      i. Construct and interpret a 95% confidence interval for the population mean.

      ii. Suppose the observation of 4 was incorrectly recorded as 24 . What would you obtain for the 95% confidence interval? What does this suggest about potential effects of outliers on confidence intervals for means?

   (b) The **Anorexia data** file results for the cognitive behavioral and family therapies and the control group. Using data for the 17 girls who received the family therapy:

      i. Conduct a descriptive statistical analysis using graphs and numerical summaries.

      ii. Construct a 95% confidence interval for the population mean change in weight.

      iii. Construct a 95% confidence interval for the difference between the population mean weight changes for the family therapy and the control. Interpret.

   (c) The **Houses data** for 100 home sales in Gainesville, Florida, several variables, including the selling price in thousands of dollars and whether the house is new (1 = yes, 0 = no). Prepare a short report in which, stating all assumptions including the relative importance of each, you conduct descriptive and inferential statistical analyses to compare the selling prices for new and older homes.

2. **Statistical Inference: Hypothesis Testing** [4+6+10=20 points]

   (a) Introducing notation for a parameter, state the following hypotheses in terms of the parameter values and indicate whether it is a null hypothesis or an alternative hypothesis.

      i. The proportion of all adults in the UK who favor legalized gambling equals 0.50.

      ii. The correlation for Australian adults between smoking (number of cigarettes per day) and blood pressure is positive.

      iii. The mean grade point average this year of all college graduates in the U.S. is the same for females and males.

   (b) You want to know whether adults in your country think the ideal number of children is equal to 2 , on the average, or higher or lower than that.

      i. Defining notation, state H_0 and H_a for investigating this.

      ii. Software shows these results for responses in a recent GSS to the question, "What do you think is the ideal number of children to have?":
      Test of $\mu = 2.0$ vs $\mu \neq 2.0$

Variable	n	Mean	StDev	SE Mean	T	P-value
Children	1302	2.490	0.850	0.0236	20.80	0.0000

      Show how to obtain the test statistic from other values reported. Explain what the P-value represents, and interpret its value.

(c) Ideally, results of a statistical analysis should not depend greatly on a single observation. In a sensitivity study, we re-do the analysis after deleting an outlier from the data set or changing its value to a more typical value and checking whether results change much. For the **Anorexia data** the weight change of 20.9 pounds for the *cb* group was a severe outlier. Suppose this observation was actually 2.9 pounds but recorded incorrectly. Find the P-value for testing $H_0 : \mu_1 = \mu_2$ against $H_a : \mu_1 \neq \mu_2$ with and without that observation. Summarize its influence.