Statistics 500 - Homework #4, Fall 2020
Due: in class on Wednesday, 09/18/2020

Reading Assignment: Statistical Sleuth, Chapters 5 and 6. Review the conceptual exercises at the end of Chapter 5. Solutions to the conceptual exercises are given at the end of Chapter 5.

1. Suppose Y ~ Poisson(μ). Justify why $X = \sqrt{Y}$ is a variance-stabilizing transformation.

2. Consider a dataset about survival times (in days) of guinea pigs that were randomly assigned either to a control group or to a treatment group that received a dose of tubercle bacilli (a bacterium that causes tuberculosis). These data are shown below and are posted online as **guinea_pigs.txt.** Partial SAS code is given as guinea_pigs_partial.sas.

Survival Times (in days) of Guinea Pigs

Received Bacilli (n=58)				**Controls (n=64)**			
76	164	225	373	18	149	341	576
93	166	244	373	36	160	355	590
97	168	253	376	50	165	367	603
107	178	256	397	52	166	380	607
108	179	259	398	86	167	382	608
113	181	265	406	87	167	421	621
114	181	268	459	89	173	421	634
119	183	270	466	91	178	432	637
136	185	283	592	102	189	446	638
138	194	289	598	108	209	455	641
139	198	291		114	212	463	650
152	212	311		114	216	474	663
154	213	315		115	273	505	685
154	216	326		118	278	545	688
160	220	326		119	279	546	725
164	225	361		120	292	569	735

a. Perform the following two tests of the hypothesis that the distributions of survival times are the same for controls and the guinea pigs infected with tubercle bacilli against the **one-sided** alternative that infection with tubercle Bacilli tends to decrease survival times. Report your results (p-value, and conclusion).
(i) two-sample t-test based on a model that assumes two independent random samples from normal distributions with homogeneous variances.
(ii) permutation test based on computing t-statistics for each of 20,000 new random assignments of guinea pigs to treatment groups.

b. Perform the Wilcoxon rank-sum test of the hypothesis that the distributions of survival times are the same for controls and the guinea pigs infected with tubercle bacilli against the **one-sided** alternative that infection with tubercle Bacilli tends to decrease survival times. Report
(i) The sum the ranks computed from the combined data set, for each treatment group.
(ii) The p-value.
(iii) Your conclusion in the context of this study.
(iv) Compare with results from parts a(i) and a(ii).

c. Generate box plots for these data. Does the result from the Wilcoxon test in part (b) agree with what is suggested by the box plots? Justify your answer.

d. A two-sample t-test would be the most powerful test for detecting a difference in mean survival times if the data for the controls and the data for patients subjected to group therapy were independent random samples from normal distributions with homogeneous variances. Assuming independent samples from normal distributions, check the assumptions of homogeneous variances using
   (i) The ratio of standard deviations
   (ii) The folded F- test
   (iii) The Brown-Forsythe test
   State your conclusion for (i), (ii) and (iii).

e. Use the SAS package to construct a normal probability plot for each of the two samples. Examine the plots to assess the hypothesis that guinea pig survival times have a normal distribution for both the controls and the guinea pigs infected with tubercle Bacilli. What information do these plots provide about possible departures from normality.

f. Report the p-value for the Shapiro-Wilk test of normality for each sample. State your conclusion.

g. Which test in part (d) would you feel more comfortable in applying to these data to test the homogeneity of variance assumption? Provide some justification for your decision.

3. In a study of energy conservation in single family homes, 20 homes were randomly selected from homes built in a housing development in southern England. Ten out of the 20 houses were randomly selected and were constructed with standard levels of insulation. The other 10 houses were constructed with extra insulation. Each house was heated with a gas furnace and energy consumption was monitored for 8 years. The data shown below give annual gas consumption in MWh (megawatt hour) for each of the 20 houses.

Standard Insulation	**Extra Insulation**
16.8	15.1
16.2	13.9
16.0	15.9
28.6	17.2
6.6	15.2
19.3	13.8
11.0	11.3
14.7	13.2
20.2	18.8
21.0	14.0

a. Was this an experiment or an observational study? Justify your answer.

b. Construct a 95% confidence interval for the mean difference in annual energy use between standard and extra insulation. Interpret the confidence interval.

c. Based on the confidence interval you constructed in part (b), would you reject the null hypothesis that the mean annual energy use are the same for standard insulation and extra insulation? Justify your answer.

d. A future study is being designed to estimate the difference in mean annual energy use for two new types of insulation. Assuming that for each new type of insulation, the variation in energy use among houses will be similar to the variation observed in the current study. Determine the number of houses needed from each type of insulation for the future study so that the 95% confidence interval of the difference in mean energy use for the two types of insulation is no wider than 1 MWh. (For t-distribution with df larger than 100, use the corresponding Z-quantiles as approximate t-quantiles.)

e. Suppose that you want to use SAS to check your answer for part (d). Finish the following SAS code to get the number of houses needed from each type of insulation for the future study so that the 95% confidence interval of the difference in mean energy use for the two types of insulation is no wider than 1 MWh.

```
proc power;
     twosamplemeans
     alpha =
     meandiff =
     stddev=
     npergroup =
     power =              ;
run;
```

f. The analysis of parts (b-e) are based on 2-sample t-test. Assess the model assumptions for 2-sample t-test for this dataset.

g. Regardless of your answers to part (f), perform a 2-sample t-test assuming unequal variance (according to slide 207) comparing the standard and extra insulation. State the null and alternative hypotheses, compute the test statistic, give the p-value, and state your conclusion in the context of the study.

h. Regardless of your answers to part (f), perform the Wilcoxon rank-sum test comparing the standard and extra insulation. State the null and alternative hypotheses, compute the test statistic, give the p-value, and state your conclusion in the context of the study.