

Assignment 1

Due January 29 at 11:00 AM

Problem 1: (20 points) *Smoking.syd, Smoking.xls*

A study published last year suggested that male fertility (defined by the sperm counts in fertile adults) was lower in rural areas of the US than in urban areas. The authors of the study hypothesized that exposure to pesticides and herbicides in rural areas decreased male fertility. The following table is a realistic subsample from the study.

Site	N	[Sperm x106 / ml]		Smoking (%)
		Avg	SD	
Columbia, MO	87	58.7	49.5	13.8%
Los Angeles	63	80.8	61.3	3.2%

- a) Calculate 95% confidence intervals for the mean sperm count for men in Columbia, MO and Los Angeles. Compare the intervals. What do you conclude about the differences? Do the confidence intervals suggest they are significantly different at the 5% level?
 - b) Calculate a 2-sample t-test to test the hypothesis that true average sperm counts are different (use a two-sided alternative). What do you conclude about the differences?
 - c) Compare and contrast your results in a) and b). How does the confidence interval differ from the t-test? Do they make different assumptions?
 - d) Test the hypothesis that the proportion of smokers in Columbia, MO is higher than LA using the standard uncorrected Pearson chi Square.
 - e) Calculate likelihood ratio Chi Square (G), the Yates corrected Chi Square, and Fisher's exact test.
-
- f) Calculate the Odds Ratio and build a 90% interval for the observed odds ratio.
(Optional – extra credit)

Problem 2: (10 points) *Teen Pregnancy.syd, Teen Pregnancy.xls*

Data from the “Across American” survey was published by *teenpregnancy.org*. The surveys were conducted between December 2001 and May 2002 with leaders from each state. Information was gathered from a wide variety of sources including representatives from the health, education, welfare, workforce, and social services; and statewide non-profit and private organizations addressing teen pregnancy prevention. Data was collected on the number of teens (\log_teen_pop , natural log of number of adolescents aged 15-19), proportion of children under 6 living under the federal poverty line ($poverty, \%$), proportion of children under 18 whose parents are high-school dropouts ($dropout, \%$), and the pregnancy rate ($pregnancy, \%$).

- a) Calculate correlation between poverty and pregnancy. Is the relationship significant? Is it strong?
- b) Develop a multiple regression model predicting pregnancy rate. What is the best model and why?
- c) Compare and contrast your results in a) and b). What is the role of poverty in teen pregnancy?

Problem 3: (10 points) *Whale.syd, Whale.xls*

Whale Melon is type of fat found in the head of whales. It is thought that the melon is important in the ability of whales to navigate. Data collected from stranded (dead) whales was used to investigate the physiological properties of whale melon, blubber, and muscle.

Evaluate the hypothesis that whale melon is a different density than blubber and muscle.

- a) Perform an ANOVA on the data. Evaluate your residuals for evidence of non-normality and/or variance problems.
- b) Perform a Levene's test on the residuals. Levene's test is an ANOVA on the absolute value of the residuals. What do you conclude?
- c) Perform a Kruskal-Wallis test on the same data. Is the analysis appropriate? Why or Why not? Does it confirm or refute your results from a)?

Problem 4: (10 points) *SD Temps (many sites).syd, SD Temps (many sites).xls*

Weather data has been collected in San Diego County for more than 100 years. It makes sense to look at the data to see if there is evidence of climate change (warming).

- a) Model climate change as a function of site and year. Is there evidence that temperatures are increasing?
- b) Compare the average temperature in San Diego (Lindberg Field) to La Mesa using a paired t-test (Paired by Year). *SD Temps (2 sites) A.syd, SD Temps (2 sites) A.xls*
- c) Compare the average temperature in San Diego (Lindberg Field) to La Mesa using a independent samples (2 group) t-test. *SD Temps (2 sites) B.syd, SD Temps (2 sites) B.xls*
- d) Compare b) and c). What are the mathematical differences (compare SS, df etc)?