Due: 5pm Friday, 8 Mar

1. Analysis of boy/girl ratios in families has always been of great interest to sociologists and evolutionary geneticists. Apparently, there are wonderful historical records kept in Germany and predecessor states. One study looked at 7200 19th century families in what is now Germany with 6 children in each family. Those families had 43,200 children, 20,937 were girls and 22,263 were boys.

   (a) Assume that all of the children (6*7200 = 43200) can be viewed as a sample from a single population with proportion of females equal to π. Test the hypothesis that $\pi = 0.5$. Give the P-value and give a one-sentence conclusion.

   (b) Estimate π and calculate a 95% confidence interval. Tell me which interval you are reporting (but you can choose).

   (c) As part of the same study, 612 families with 12 children each were surveyed. Of these 7,344 children, 3534 were girls and 3810 were boys. Test whether the proportion of female children is different in 6-child families and 12-child families. Calculate a confidence interval for the difference.

   (d) If some families tended to 'maleness' and others to 'femaleness', i.e. P[child is male] varies between families, are the tests and confidence intervals reasonable? Explain why or why not.

2. The previous question seems to rule out 0.5 as a possible value for p, but does not necessarily rule out the binomial distribution for the number of female children in each family. Here is the detailed table of numbers of 6-children familes with 0, 1, ... 6 girls. The estimated P[girl] is 0.48465 = 20,937/43,200. I've also included the expected number of familes for most of the categories.

Number of Girls	Number of Boys	Obs. # Families	Exp. # Familes
0	6	152	134.88
1	5	822	761.05
2	4	1691	1789.31
3	3	2239	2243.65
4	2	1577	1582.51
5	1	606	595.30
6	0	113	

   (a) Let X represent the number of female children in a randomly chosen six-child family. Find the expected number of families (out of 7200 six-child families) with 6 girls and 0 boys, assuming the data are described by a binomial distribution with $\pi = 0.48465$. Remember that if frequencies follow a binomial distribution, the probability that $X = x$ in a 6 child family is $f(x) = \frac{6!}{x!(6-x)!}\pi^x(1-\pi)^{(6-x)}$.

   (b) Are the frequencies (number of families) in the above table consistent with a binomial distribution with $\pi = 0.48465$? Carry out a chi-square goodness-of-fit test. Calculate

the test statistic and approximate P-value and give a one-sentence conclusion. To save you some time, $\Sigma_i \frac{(O_i - E_i)^2}{E_i}$ for the 6 categories with expected counts in the table (i.e. all but 0 boys and 6 girls) = 12.6775

3. The following data come from a study of the association between drinking (alcoholic drinks) and breast cancer in medium weight women. Breast cancer is rare, so these data were collected using a retrospective (also called a case-control) study design. The investigators identified 159 women with breast cancer ('cases') using hospital records and another 300 'control' women without breast cancer. The control women were chosen to have similar ages and live in the same geographic area as the cases, but there is no specific matching of cases to controls. Each women was asked about their consumption of alcohoic drinks. We will compare two groups: those that drank less than 1 drink per month (light) and those that consume 1 or more drink per day (heavy). The data are:

	light	heavy	Total
Case	97	62	159
Control	153	147	300

   (a) Test whether drinking is associated with breast cancer.

   (b) Estimate the odds ratio (as odds of breast cancer in heavy drinkers to odds in light drinkers)

   (c) Estimate the standard error of the log odds ratio

   (d) Calculate a 95% confidence interval for the odds ratio.

4. This problem uses one version of a classic data story. Students in two departments (Statistics and Physics) at two colleges (Big U and Little U) took a standardized science exam. The number of students passing and the number failing are recorded. The data are here:

	Statistics		Physics	
College	Pass	Fail	Pass	Fail
Big U	12	6	8	14
Little U	20	16	1	3

   (a) Test the null hypothesis that the proportion of **Physics** students passing the exam is the same at Big U and Little U Colleges. Report which test you used and the p-value.

   (b) Construct the contingency table for the two departments combined. I.e. imagine you were given the data without being told which department a student belonged to. Which College has the higher percentage of passes if you are not told the department? (Description only, no test needed).

   (c) Notice that the aggregated data (ignoring department) gives a different impression of which college has the higher pass rate than do the two department-specific tables. Can you explain why?
   Hint: look at the pass rate for the two departments. Then look at the proportion of statistics dept. students at the two colleges.

Note: My answer will discuss this phenomenon in more detail and provide links to further information.

5. The data in valve.csv are from a study of valve failure in a nuclear reactor. Valves in this reactor are classified by system, operator type, valve type, head size, and operational mode. The response is the number of *Failures* within the specified time *Time*, measured in years. Note that Time is not the same for all valves. You are interested in identifying levels of each factor with high failure rate (# failures / year). Because valves only come in certain configurations, you use an additive model with all five explanatory factors (system, operator type, valve type, head size, and operational mode).

   (a) Estimate the effect associated with each level of the five explanatory factors. Which levels of which factors have a failure rate more than 5 times the rate of the most reliable level of each factor. Report those levels and the relative failure rate.
   For example, if the failure rates for three levels (A, B, and C) of a factor are: A: 0.2943/yr, B: 0.91/hr, and C: 1.53/yr. You would report C: "5.2 times".

   (b) Test the null hypothesis of no difference in failure rate between systems. Report your test statistic and p-value.

   (c) Is there evidence of overdispersion in these data? Explain why or why not.

   (d) Test the null hypothesis of no difference in failure rate between systems while accounting for the overdispersion. Report your test statistic and p-value.