

Homework 13 - handed out Weds, 5 Dec 2007, with corrected value in 2b answers will be posted ca Thursday evening (6 Dec)

Most (or all) parts of these problems can be done either in SAS or by hand. You are free to use the computer, but you should understand how to do them by hand. Parts of problem 4 are tedious. I suggest you make sure you know how do the calculations then look at the answers to see the point I'm trying to make..

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 ($6 \times 7200 = 43200$) 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.
 - (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. $P[\text{child is male}]$ 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 families with 0, 1, ... 6 girls. The estimated $P[\text{girl}]$ is 0.485. I've also included the expected number of families for most of the categories.

Number of Boys	Number of Girls	Obs. # Families	Exp. # Families
0	6	113	
1	5	606	597.03
2	4	1577	1584.91
3	3	2239	2243.93
4	2	1691	1787.05
5	1	822	759.03
6	0	152	134.33

- (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 0 boys, assuming the data are described by a binomial distribution with $\pi = .485$. 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 $\pi = 0.485$? 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, $\sum_i \frac{(O_i - E_i)^2}{E_i}$ for the 6 categories with expected counts in the table (i.e. all but 0 boys and 6 girls) = 12.894

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. Each women was asked about their consumption of alcoholic 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

- Test whether drinking is associated with breast cancer.
 - Estimate the odds ratio (as odds of breast cancer in heavy drinkers to odds in light drinkers)
 - Estimate the standard error of the log odds ratio
 - Calculate a 95% confidence interval for the odds ratio.
4. This question is intended to demonstrate some properties of odds ratios, differences in proportions, and some characteristics of case-control studies. Answers to some the following questions may be 'obvious' but I'm asking them to make a point. These are based on the study in the previous question.
- Assume that study was done on women ranging from 20-50 years old in the San Fransisco area. It is tempting to consider the 459 women study as a random sample of women from the population of 20-50 year old women in San Fransisco. If you made that assumption, please estimate the incidence of breast cancer in that population. Note: Incidence = $P[\text{women has breast cancer}]$.
 - The national incidence of breast cancer in women 20-50 years old is approximately $1/54 = 0.0185 = 1.85\%$. Why is your estimate from part a) so very different?
 - Now, calculate the incidence of breast cancer in the two subgroups (heavy drinkers and light drinkers). What is difference in breast cancer incidence?
 - Calculate the incidence of heavy drinking in the cases, i.e. $P[\text{heavy drinking among the cases}]$, the incidence of heavy drinking among the controls, and the difference in those incidence rates.
 - Finally, estimate the s.e. of the difference in incidence of heavy drinking. Note: don't assume that the rates are the same, so think confidence interval form of the s.e.
 - Imagine a study based on a random sample of San Fransisco women between the ages of 20 and 50. The following data represent what might be found in a simple random sample of 8499 women.

	light	heavy	Total
Case	96	61	157
Control	4254	4088	8342

Based on these data, I estimate:

the incidence of breast cancer = 0.0185

the incidence of breast cancer separately in light and heavy drinkers:

0.0221 in light drinkers and 0.0147 in heavy drinkers

and the difference in incidence rates = 0.0074

the incidence of heavy drinking in cases and controls = 0.388 and 0.490
and the difference in the incidence of heavy drinking = -0.101
the s.e. of the difference in incidence of heavy drinking = 0.0392
the odds ratio (as odds of breast cancer in heavy drinkers / odds in light) = 0.661
and the s.e. of the log odds ratio = 0.165

The quantities from the random sample are all valid estimates of the corresponding population quantities. The case-control data provides valid estimates of some quantities but not others. What can be appropriately estimated from the case-control study?

- (g) The s.e.'s (of either the difference in incidence of heavy drinking or log odds ratio) are smaller in the study using a random sample of 8500 individuals than they are in the case-control study of 459 individuals. However, the s.e.'s from a case-control study of 8500 individuals would be in the neighborhood of:

s.e. of diff. in incidence of heavy drinking: $0.011 = 1.1\%$

s.e. of log odds ratio: 0.046

Using this information and all the other comparisons we've made, summarize the advantages and disadvantages of a case-control design to study factors associated with a rare event.