

BST 140.651**Problem Set 6****Due in the biostat main office on November 29th**

Problem 1. The following data show the results of caries surveys in five towns and also the fluoride content of the drinking water.

Area	Surrey and Essex	Slough	Harwick	Burnham	West Meres	Total
Fluoride p.p.m.	0.15	0.9	2.0	3.5	5.8	
Number children with with caries	243	83	60	31	39	456
Number children with caries free teeth	16	36	32	31	12	127
Number examined	259	119	92	62	51	583

The data refer to samples of children aged 12-14 only.

- Do a significance test to determine whether the proportions of children caries free varies from area to area. What does this test reveal about the effect of the fluoride content of water?
- Interpret the data and/or display it in any way you think will make it more interpretable.
- Briefly discuss the limitations of this type of study for studying the effect of fluoride.

Problem 2. Test your computer's random normal generator. Simulate 10,000 random normal deviates and test whether or not they appear to be normal with a Chi-squared test. Explain your steps and interpret your results.

Problem 3. Retinitis pigmentosa is a disease which manifests itself via different genetic modes of inheritance. Cases have been documented with a dominant, recessive, and sex-linked form of inheritance. It has been conjectured that the form of inheritance is related to the ethnic origin of the individual. Cases of the disease have been surveyed in an English and Swiss population with the following results: out of 125 English cases, 46 had sex-linked disease, 25 had recessive disease and 54 had dominant disease; out of the 110 Swiss cases, one had sex-linked disease, 99 had recessive disease, and 10 had dominant disease. Based on these data is there a significant association between ethnic origin and genetic type? Analyze and interpret (in words) this data. (10 points)

Problem 4. A small study was done to compare how well students with different majors do in an introductory statistics course. Seven majors were found: biology, psychology, sociology, business, education, meteorology and economics. At the end of the course, the

students were given a special test to measure their understanding of basic statistics. Then a series of t-tests were performed to compare every pair of majors. Thus, biology and psychology majors were compared, biology and sociology majors, psychology and sociology majors, etc., for a total of 21 t-tests.

Simulate this study assuming all majors do about the same. Assume there are 20 students in each major, and that scores on the test have a normal distribution with mean $\mu = 12$ and $s = 2$. Use the computer to generate random test scores that are normally distributed for biology majors, then do it a second time to get a sample for psychology majors and so on, for 7 samples (one for each major).

- List the 21 pairs of majors and perform the 21 t -tests.
- In how many of the tests did you reject the null hypothesis at $\alpha = 0.10$?
- Since this study was simulated, the true situation is known – there aren't any differences. But you probably did find at least one pair of majors where there was a significant difference. This illustrates the “hazards” of doing a lot of comparisons. Try to think of some other situations where one might do a lot of statistical tests. For example, suppose a pharmaceutical firm had 16 possible new drugs which they wanted to try out in hopes that at least one was better than the present best competing brand. What are the consequences of doing a lot of statistical tests?

Problem 5. In a study of the association between cigarette smoking and lung cancer, 1,357 male lung cancer patients were compared with 1,357 controls in terms of their cigarette consumption as follows:

	Cigarette Consumption Daily						Total
	0	1–	5–	15–	25–	50+	
Lung cancer patients	7	49	516	445	299	41	1,357
Controls	61	91	615	408	162	20	1,357

Compute the odds ratio and log odds ratio in each of the 5 smoking groups compared with non-smokers. Find confidence intervals and graphically display. Comment and interpret. Can relative risks be estimated. Why or why not.

Problem 6. In a retrospective study of the possible effect of blood group on the incidence of peptic ulcers, Woolf (1955) obtained data from three cities. The table gives for each city data for blood groups 0 and A only. In each city, blood group is recorded for peptic ulcer subjects and for a control series of individuals not having peptic ulcer.

	Peptic Ulcer		Control	
	Group 0	Group A	Group 0	Group A
London	911	579	4578	4219
Manchester	361	246	4532	3775
Newcastle	396	219	6598	5261

- a. Compute the odds ratio for each city with a confidence interval. Interpret.
- b. Suppose that it is required to estimate $P(\text{ulcer}|A) - P(\text{ulcer}|0)$. What further information is needed to do this from the current data?

Problem 7. Prove that the odds ratio is greater than 1 if and only if the relative risk is greater than 1.

Problem 8. The odds ratio approximates the relative risk when the disease is rare. But how rare? Suppose the relative risk is 2 for disease (B) and risk factor (A). Also, suppose 20% of the population has been exposed to the risk factor.

- a. Suppose the probability of having the disease among those without the risk factor is d . Derive an expression for the odds ratio in terms of d . Calculate the odds ratio for $d = 1/10, 1/100, 1/1000, 1/10,000$ and $1/100,000$. Make a plot of the odds ratio versus d . What do you conclude? Does the odds ratio approximate the relative risk?
- b. What do you think would be the effect on the approximation if the prevalence of the risk factor changed, e.g. increased? decreased? Suppose the relative risk changed; increased, decreased?

Problem 9. Suppose that a case-control (retrospective) study was conducted in a certain school district. School children with emotional disturbances requiring psychological care were compared with presumably normal children on a number of antecedent characteristics. Suppose it was found that one quarter of the emotionally disturbed children versus one tenth of the normal controls had lost (by death, divorce, or separation) at least one parent before age 5.

Suppose that a new retrospective study is being conducted in a different community. From a survey of the available school records, we estimate that the proportion of normal children who have lost a parent was .30.

- a. Suppose based on the first study that we think the odds ratio that applies in the new community is the same as that found in the first study. We would like a 90% chance of detecting an association. Perform a sample size calculation for a two-sided test at level $\alpha = .05$. How many cases and controls would you recommend?
- b. Suppose the proportion of normal children who had lost a parent was less than .3, say .10. How would this affect your sample sizes? Suppose it was much larger? Explain.