MA547 **Spring 2013**

HOMEWORK ASSIGNMENTS

Homework #5 **Assigned: 3/26/13** **Due: 4/9/13**

#1. Consider the data on vascetomy on myocardial infarction. The subjects were matched on two variables. However, the subjects were not matched for obesity and smoking history (see Rosenbaum pg 92). Suppose smoking history is a serious concern, but obesity is not, what conclusion might be reasonable? Suppose obesity a serious concern, but smoking history is not, what conclusion might be reasonable? Suppose both obesity and smoking history are serious concerns, what conclusions might be reasonable?

#2. (a) In a completely randomized experiment, let z_i denote the treatment indicators and r_i the responses, $i = 1, \ldots, n$. Show that the test statistic $t(\underset{\sim}{z}, \underset{\sim}{r}) = \sum_{i=1}^{n} z_i r_i$ is *effect increasing*.

(b) The Wilcoxon rank sum statistic is $t(\underset{\sim}{z}, \underset{\sim}{r}) = \sum_{i=1}^{n} z_i r_i$, where z_i are the treatment indicators and and r_i are the ranks. Show that it is *effect increasing*.

#3. Consider the data on genetic damage from alcohol (handout), and note that there are two covariates, age and sex. Pretend as if there is no matching. (a) Use logistic regression to obtain propensity scores. (b) Construct five strata, and identify which alcoholics and which controls belong to the strata. (c) Explain whether stratification on the propensity scores is a good idea. [5 points]

#4. (a) Suppose there are no covert biases. Then, the strata are homogeneous in the propensity score, $e(\underset{\sim}{x})$. That is, show that

$$P(Z, \underset{\sim}{X} \mid e(\underset{\sim}{x})) = P(\underset{\sim}{X} \mid e(\underset{\sim}{x}))P(Z \mid e(\underset{\sim}{x})).$$

Discuss the meaning of this equation.

(b) Considering a study free of overt biases but not covert biases, let $\underset{\sim}{z}$ represent the treatment indicators, and $\underset{\sim}{u}$ represent the covert bias. Starting with the lemma (proposition, stated in class) which shows the equivalence between the log odds ratio and the logit, show that

$$P(\underset{\sim}{z} \mid \underset{\sim}{m}) = \prod_{s=1}^{S} \left\{ \frac{\exp(\gamma \underset{\sim}{z}_s' \underset{\sim}{u}_s)}{\sum_{s=1}^{S} \exp(\gamma \underset{\sim}{z}_s' \underset{\sim}{u}_s)} \right\}, \quad m_s = \sum_{j=1}^{n_s} z_{sj}, s = 1, \ldots, S.$$

#5. (a) Use McNemar's test to perform a *sensitivity analysis* on the vascetomy data for the test of no treatment effect.

(b) Use Wilcoxon signed rank statistic to perfrom a *sensitivity analysis* on the chromosome data for the test of no treatment effect.

#6. Table 4.11 of Rosenbaum (2002) shows data on a study of the effects of the drug allopurinal on rash. The data are further stratified by race (black, white). Perform a sensitivity test of no treatment effect using the Mantel Haenszel procedure. [Use the normal approximations on pg. 130-131].

		Allopurinol	No Allopurinol
Black Males	Rash cass	2	20
	Non cases	20	445
Black Females	Rash cases	3	30
	Non cases	10	218
White Males	Rash cass	3	16
	Non cases	13	200
White Females	Rash cases	7	28
	Non cases	9	400

Homework #4 **Assigned: 3/12/13** **Due: 3/26/13**

#1. Ray's farm has 8,000 apple trees. A month before the harvest, Ray wants to estimate the total yield of his farm. He took a simple random sample of 80 trees, and poststratify the trees by age (years). He obtained the data on the number apples per tree in the table below.

age	n	avg	std
1-3	30	10	5
3-5	20	150	18
5-7	20	100	15
7-10	10	75	12

Obtain a 95% confidence interval for the total number of apples of all trees with and without the poststratification. [Assume SSB is twice SSW.]

#2. Suppose the design effect for a complex survey when the sample mean is the estimator is 0.50. For a simple random sample of size 100 the sample mean is 45 and the standard deviation is 9. Find a 95% confidence interval for the finite population mean for the complex survey. [Assume that the population is much bigger than the sample size.]

#3. (a) Find the mean and the variance of the Wilcoxon rank sum statistic under the hypothesis of no treatment effect. Hence, find an approximate formula for the p-value of the test.

(b) Table 4.15 of Rosenbaum (2002; see my data sheets) shows data on a study of 23 subjects who had eaten large quantities of fish contaminated with methylmercury. Assume that the study is based on a uniform randomized experiment with a single stratum. Use the procedure in (a) to test whether mercury in the blood of the exposed subjects is higher than in the controlled subjects.

#4. (a) Consider the Mantel-Haenszel test statistic for no treatment effect in matched binary data. Find an approximate formula for the p-value of the test.

(b) Table 4.11 of Rosenbaum (2002) shows data on a study of the effects of the drug allopurinal on rash.

		Allopurinol	No Allopurinol
Males	Rash cass	5	36
	Non cases	33	645
Females	Rash cases	10	58
	Non cases	19	518

Assume that the study is based on a uniform randomized experiment with two strata (male, female). Use the procedure in (a) to test whether allopurinal is a cause of rash.

#5. (a) Starting with the treatment-minus-control effect in Fisher's exact test, with $m = n/2$, show that the statistics $\sum_{i=1}^{n} z_i r_i + \sum_{i=1}^{n}(1 - z_i)(1 - r_i)$ and $\sum_{i=1}^{n} z_i r_i$ can be used as alternatives.

(b) For the McNemar's test statistic show that

$$\sum_{s=1}^{S}\sum_{i=1}^{2} z_{si} r_{si} = \sum_{s=1}^{S} r_{s2} + \sum_{s \in D} z_{s1}(r_{s1} - r_{s2}),$$

where D is the set of discordant pairs.

Homework #3 **Assigned: 2/26/13** **Due: 3/12/13**

#1. Sharon Chapter 3, #10.

#2. Sharon Chapter 3, # 6.

#3. A random sample of size 211 units is taken from a population with six strata. Some data are provided in the table below.

Stratum	N_h	S_h	c_h
1	300	3000	5
2	25	2000	85
3	60	9000	8
4	20	2000	99
5	75	12000	8
6	130	1000	40

(a) Find the sample sizes within strata under proportional allocation, Neyman allocation and optimal allocation.

(b) What is the overall sample size if the sample mean is required to be within 650 of the population mean with 95% confidence?

Homework #2 **Assigned: 2/5/13** **Due: 2/19/13**

#1. To estimate how many library books need rebinding, a librarian uses a random number table to select 100 locations on the library shelves; each location has at least one book. He then walks to each location, looks at the books in that location, and carefully checks all books which are at least 10 years old to see whether the books need rebinding or not.

(a.) Describe the sampling frame, sampling units and the observation units.

(b.) Describe the target population and sampled population. Are there any sources of bias? Describe them.

(c.) What is the population quantity of interest?

#2. A teacher wants to study the *relationship* between gender (male, female) and geography (local, foreign) of the students in her class. Set up a two-item questionnaire and discuss how you will present and analyze the data.

#3. (a) A simple random sample of size n is taken from a population of size N. Show that the probability there are $k \leq n$ *specific* individuals in the sample.is $\frac{n!}{N!}/\frac{(n-k)!}{(N-k)!}$.

(b) Let S^2 denote the population variance. Show that

$$S^2 = \frac{1}{N}\left\{\sum_{i=1}^{N} y_i^2 - 2\frac{\sum_{i<j} y_i y_j}{N-1}\right\},$$

where $y_1, \ldots, y_N$ are the unknown population values.

#4. Toyota manufactured 1000 2010 RAV4s (four cylinders). A random sample of 100 2010 RAV4s was taken and two characteristics were measured, the gas consumption on city raods and whether a RAV4 has a Pacific-Blue Metallic (PBM) body. The average and standard deviation (miles per gallon) are 22 mpg and 4 mpg respectively, and there are 10% cars with PBM bodies.

(a) Find 95% confidence intervals for the finite population mean consumption and the finite population percent of cars with PBM bodies.

(b) Find 95% confidence intervals for the coefficient of variation of the gas consumption and the proportion of 2010 RAV4ss with PBM bodies.

#5. A scientist wants to find out the average weight of aligators in 100 lakes in Florida. He took a random sample of 5 lakes and weighed one aligator from each lake. He used the age and size of each lake and the number of aligators in each lake to obtain sample weights. The weights (lbs) of the aligators are 120, 80, 90 ,95, 110 and the sample

weights are 30, 10, 15, 20, 25. Obtain an approximate 95% confidence interval of the average weight of the aligators in the 100 lakes. Compare your confidence interval with the one which ignores the sampling weights.

#6. (a) Exercise 2.19 in Sharon's book.

(b) Exercise 2.24 in Sharon's book.

Homework #1 **Assigned: 1/22/13** **Due: 2/5/13**

#1. When a group of people is randomly separated out into two groups, there could be differences in some characteristics (e.g., covariates) across the two groups. This is called a *covariate imbalance*. A group of 100 people is randomly divided into two groups of 40 and 60. Is there evidence for a covariate imbalance in each of the following cases?

(a) The number of males in the first group is 20 and the number of males in the second group is 45.

(b) The average weight of people in the first group is 120 lbs with a standard deviaton of 20 lbs and the average weight of people in the second group is 140 lbs with a standard deviation of 25 lbs.

#2. A random sample of n objects is drawn from a population of N objects without replacement. Let $Z_i = 1$ if the i^{th} unit is in the sample and $Z_i = 0$ otherwise. What is the probability mass function of Z_i? Write down $E(Z_i)$ and $\text{Var}(Z_i)$. Find $\text{cor}(Z_i, Z_j)$ for $i \neq j$.

3. In comparing Ford and Toyota, an investigator wishes to consider the percentages of defective vehicles they manufacture. She selected three different sizes of cars, small, medium and large, and obtained the data in Table 1 for some cars made in 2006.

Table 1: Classification of cars manufactured by Ford and Toyota by small and medium sizes

	small		medium		large		total	
	d	n	d	n	d	n	d	n
Ford	50	100	80	200	10	25	140	325
Toyota	300	600	50	100	3	15	355	715
		700		300		40		1040

NOTE: d is the number of defective cars of a total of n cars.

Compare Ford and Toyota by using the percentages based on the totals, and explain what is wrong with this. Based on these data, describe the best two percentages you can use to compare Ford and Toyota.

4. In our last annual department's picnic, there were 50 attendees and many different kinds of foods. Sometime after we ate lunch, 15 of us became sick. A smart

statistician at the picnic wanted to find out the reason. He found that 12 of the sick attendees ate Fried Rice, and 3 did not, and 5 of the attendees, who did not get sick, ate the Fried Rice. What information do you need to conclude that Fried Rice *caused* the attendees to get sick? What is the name of this study? Present the result from a statistical test that may be appropriate.