MA552, Distribution-Free and Robust Statistical Methods, Fall 2009

HOMEWORK ASSIGNMENTS

Homework #1 **Assigned: 9/09/09** **Due: 9/16/09**

1. Let X have a continuous distribution symmetric about 0. Show that $|X|$ is independent of its sign S (S is positive if $X > 0$ and S is negative if $X < 0$). Let $X_i,\ i = 1, \ldots, n$ be independent from a continuous distribution symmetric about 0, and let $\underset{\sim}{S}$ be the vector of signs. Find the distribution of $\underset{\sim}{S}$.

2. A random sample of size 25 young adults is taken from a population and their weights (lbs) are recorded as follows.

120, 121, 123, 115, 118, 123, 150, 156, 140, 145, 146, 130, 131, 142, 143, 148, 136, 138, 137, 143.5, 132.2, 121.3, 120.2, 124.4, 119.5

A 95% confidence interval is needed for the population median. Find

a. a small sample interval;

b. a large sample interval;

c. an interval assuming the distribution is normal.

How close are the intervals in (a), (b) and (c)?

Homework #2 **Assigned: 9/16/09** **Due: 9/30/09**

1. Let $X_1 \ldots, X_n \overset{iid}{\sim} F(x)$, where $F(x)$ is a continuous distribution function. Let R_i be the rank of $X_i,\ i = 1, \ldots, n$. Find the $\text{cov}(R_i,\ R_j)$, $i, j = 1, \ldots, n$.

#2. Graph the Parzen-Rosenblatt kernel density estimate for the sodium data in Table 1.1.1 of Higgins' text book. Compare your answer from the SAS software to the one you obtain directly from the kernel density estimator.

#3. A random sample of 10 students was studied to test whether walking one hour a day for six weeks can reduce obesity. Their body mass indicies were measured just before and after the program. The body mass indices (wgt (kg)/ squared hgt (m)) are shown below.

Before	120	130	140	135	136	120	150	128	129	131
After	121	131	124	130	132	121	130	127	127	130

Use the binomial (sign) and the permutation tests to assess the effectiveness of the six-week program.

#4. Let $X_1, \ldots, X_n \overset{iid}{\sim} F(x)$, where $F(x)$ is a continuous distribution.

(a) For distinct obsevations denote the ranks by R_i. However, suppose the first $k < n$ observations are tied; denote their ranks by $\tilde{R}_i$. Show that

$$E(\tilde{R}_i) = E(R_i), \text{Var}(\tilde{R}_i) = \text{Var}(R_i) - k(k^2 - 1)/12n.$$

(b) Suppose the observations are symmetric about θ as in the Wilcoxon signed-rank statistic SR_+. Let $W_{ij} = (X_i + X_j)/2,\ i \leq j = 1, \ldots, n$, denote the Walsh averages, and let T denote the number of W_{ij} that are larger than θ. Show that (i) W_{ij} are dependent and nonidentical random variables, and (ii) $SR_+ = T$.

#5. Consider the sodium data in Table 1.1.1. Suppose the population has a scale-location exponential distribution. Find the asymptotic power (large sample size) of the (binomial) sign test at $\mu = 75.8$.

Homework #3 **Assigned: 9/30/09** **Due: 10/07/09**

#1. The table below gives Tom's 10 blood pressure readings (mmHg) he took last week

```
120, 135, 110, 140, 135, 141, 142, 121, 110, 125
```

Find a 80% confidence of the median of all meaurements Tom could have taken last week using a procedure which assumes the underlying distribution is continuous and symmetric, and the sample is random. Compare your answer with a procedure that does not use the assumption of symmetry. Discuss how you might use the binomial (sign) test to find a 80% confidence interval for the median?

#2. Consider one-sample test statistics under the null hypothesis of the location parameter and distinct observations. Show that the permutation and the Wilcoxon signed-rank statistics both have symmetric distributions.

#3. Consider the data set in Exercise 4, page 73, of Higgins. Use the two-sample permutation test and Wilcoxon rank-sum test to assess the difference of crayfish in the the two sections of the stream in Kansas. What differences do you see in your answers when the Student's t statistic is used? [You can use Proc npar1way on SAS.]

Homework #4 **Assigned: 10/07/09** **Due: 10/21/09**

#1. Show that the the Mann-Whitney test statistic is symmetric under the hypothesis that the combined data from the two samples is a random sample from the same continuous distribution. Use this result to show that the Wilcoxon rank-sum statistic formed from each sample (i.e., W_1 and W_2) is also symmetric. Find the points of symmetry for the Mann-Whitney test and the Wilcoxon tests (W_1 and W_2) corresponding to the two samples.

#2. Consider the data set in Exercise 4, page 73, of Higgins. Obtain the Hodges-Lehmann estimate of the shift parameter. Obtain an asymptotic 90% confidence interval of the shift parameter. How would the Hodges-Lehmann estimate and the confidence interval change if the observations 16 and 15 were both 15 and the observations 11 and 14 were both 14?

#3. For the two-sample problem with shift alternatives, show that the asymptotic relative efficiency (ARE) of the Wilcoxon rank-sum test to the t test for (a) a double exponential population is 1.5 and (b) an exponential population is 3.0.

#4. For a sample of size N from an exponential distribution that the scores are

$$E_{(i)} = \sum_{k=1}^{i} \frac{1}{N-k+1}, \quad i = 1, \ldots, N.$$

[Hint: For $X_1, \ldots, X_N \overset{iid}{\sim} \exp(1)$, let $U_i = X_{(i)} - X_{(i-1)}$ where $X_{(i)}$ are the order statistics and $U_1 = X_{(1)}$. It can be shown that the U_i are independent and $U_i \sim \exp(N-i+1)$.]

Show that the normal scores are symmetric [i.e., for n even $E(Z_{(i)}) = -E(Z_{(n-i+1)})$ and for n odd this relation still holds with $E(Z_{(n+1)/2}) = 0$].

Consider the data set in Exercise 4, page 73, of Higgins. In HW 3 you have used the two-sample permutation test, Wilcoxon rank-sum test and the Student's t test to assess the difference of crayfish in the the two sections of the stream in Kansas. Now use permutation tests based on VW scores and exponential scores to assess the difference.

Explain how you might construct a two-sample permutation t test for this application.

#5. To study weight gain in laboratory animals, two groups of animals were administered a standard feed and a fortified feed over a six-week period. The weight gain (g) of each animal in the two groups is shown below.

```
Standard:  8.9,  3.0,  8.2,  5.0, 3.9,  2.2,  5.7,  3.2, 9.6, 3.1,  8.8

Fortified: 5.7, 12.0, 10.1, 13.7, 6.8, 11.9, 11.7, 10.4, 7.3, 5.3, 11.8
```

Use the Kolomogorov-Smirnov test to assess any difference between the two groups. Use the Siegel-Tukey test to assess the difference in scales. Also, use a permutation test based on the interquatile range to assess difference in scales.

Homework #5 **Assigned: 10/21/09** **Due: 11/04/09**

#1. An experiment was conducted to assess the difference of three toxins on the liver of a certain species of trout. Eighteen trouts were randomized into the three toxins as in a completely randomized design. Unfortunately, three trouts died. The data are the amounts of deterioration (in standard units) of the liver of each sacrificed trout after 5 weeks.

1	2	3
28	33	18
23	36	21
14	34	20
27	29	22
	31	24
	34	

Obtain a table of p-values based on the F test and the Kruskal-Wallis test under the following scenarios.

a. The three dead trouts are ignored.

b. The median of the deteriorations of each toxin with the dead trouts is imputed for deteriorations of the dead trouts.

c. Within each group with the dead trouts, the deteriorations of dead trouts are imputed by drawing randomly from the others within each toxin.[Just do one draw!]

#2. Consider the data in Exercise 1. Repeat the analyses in (a) using two permutation tests, one based on observed data and the other based on ranks. What do you see if you apply the three two-sample Wilcoxon rank-sum tests?

#3. In a completely randomized design, let $R_{ij}, i = 1, \ldots, k,\ j = 1, \ldots, n_i$ denote the ranks of the observations. Also, let $R_i = \sum_{j=1}^{n_i} R_{ij},\ i = 1, \ldots, k$. Show that

$$\text{cov}(R_i, R_{i'}) = -n_i n_{i'} \sigma_R^2/(N-1), i \neq i' = 1, \ldots, k.$$

Let A_{ij} denote scores assigned to the observations (ranks are a special case), and let $\bar{A}_i = \sum_{j=1}^{n_i} A_{ij}/n_i$. Find $\text{var}(\bar{A}_i - \bar{A}_{i'}),\ i \neq i' = 1, \ldots, k$.

Use the formula for $\text{var}(\bar{A}_i - \bar{A}_{i'})$ to obtain the formula for the special case of ranks.

#4. Consider the data in exercise 1. Obtain the three multiple comparison rank-based procedures (Bonferroni's, Fisher's, Tukey's) using asymptotic and permutation methods.

#5. For the RCBD show that $(b-1)(k-1)MSE/\sigma^2 \sim \chi^2_{(b-1)(k-1)}$ always. Also, for the RCBD show that the permutation F-test is equivalent to rejecting the null hypothesis for large values of $\sum_{i=1}^{k} \bar{Y}_{i\cdot}^2$.

Homework #6 **Assigned: 11/4/09** **Due: 11/18/09**

#1. Consider the data set in Exercise 10, page 143, of Higgins.

(i) Test for agreement among the four judges' rankings. [Assume blocking by contestants, and note that ranks are scores.]

(ii) Test for differences among the six contestants. [Assume blocking by judges.]

(iii) Use Tukey's HSD multiple comparison procedure to find similar contestants. [See my notes for the appropriate form!]

#2. For testing $H_0 : \beta_1 = 0$ in the regression model $y_i = \beta_0 + \beta_1 x_i + e_i,\ e_i \overset{iid}{\sim} \text{Normal}(0, \sigma^2),\ i = 1, \ldots, n$, we have two test statistics,

$$t_{\text{slope}} = \hat{\beta}_1 \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{MSE}} \quad \text{and} \quad t_{\text{corr}} = r\sqrt{\frac{n-2}{1-r^2}}.$$

Show that $t_{\text{slope}} = t_{\text{corr}}$.

Consider the formula that relates r_s to D. Under the hypothesis of no association, we have shown that $r_s \overset{d}{=} -r_s$ (i.e., r_s is symmetric about 0). Now show that D is symmetric about a point, to be determined.

#3. The amount of a wear (inches) on a certain carbide-steel cutting tool depends on the time (minutes) it runs. The data below are for 9 similar tools cutting through the a unifrom piece of a metal.

Time	130	120	100	75	70	55	42	30	19
Wear	.09	.09	.07	.07	.06	.05	.04	.03	.02

Perform two permutation tests, one based on Pearson's correlation coefficient and the other on Spearman's rank correlation coefficient, to assess the corrrelation between time and tool wear. [10 points]

Homework #7 **Assigned: 11/18/09** **Due: 12/2/09**

#1. (a) Exercise 15, page 193, of Higgins.

(b) In Fisher's exact test show that the permutation distribution of X (the number of observation is the first row and first column) is the same whether permutations are done by rows or columns.

(c) Obtain a large sample test for no association in a 2×2 table based on the count, X, in the first row and first column.

#2. How many permutations are there in each of the 2×2 tables below?

```
(i)      A   B      (ii)      A  B        (iii)        A  B

     A   6   5           A    6  5  11            A    6  5  11

     B   7   2           B    7  2   9            B    7  2   9
           ....                    ...                .........
           20                      20                 13  7  20
           ...                     ...                .........
```

#3. Exercise 12, page 192, of Higgins.

#4. The conditioning in McNemar's can be removed as follows. On the first occasion, let $X_1 = 1$ if an individual responds 'yes' and 0 otherwise, and onthe second occasion, let $X_2 = 1$ if an individual responds 'yes' and 0 otherwise. Let $P(X_1 = 1) = p$, $P(X_2 = 1 \mid X_1 = 0) = \pi_o$ and $P(X_2 = 1 \mid X_1 = 1) = \pi_1$. Find $E(X_2 - X_1)$ and $Var(X_2 - X_1)$. [I have developed this much further!]

#5. In a random sample of 18 persons (age 20-20) in Worcester, body mass index (kgm^{-2}) was measured. The body mass indices are

20, 18, 30, 40, 25, 20, 21, 35, 28, 27, 20, 21, 22, 29, 28, 33, 35, 25.

Apply the bootstrap method to make inference about the population mean body mass index using both the empirical cdf and the Parzen-Rosenblatt cdf. Compare the bootstrap mean, variance, and mean square error. Also, give 95% confidence intervals based on the percentile method and the residual method.