STAT 554: HW #5
due April 21, 2008

Please review the instructions from HW #4 concerning presentation of homework solutions. Each part is worth 5 points and I'll sum your best five (of six) scores to get your overall score for assignment. I'll answer any *general* questions that you have, but *don't expect me to tell you what procedures to use* for the various parts. I want to see whether or not you can come up with the proper methods on your own.

You should provide some sort of justification for your choice of inference procedure for each of the parts. (Failure to do so may result in a deduction of points even though you report the right answer.) Please be concise (but not too concise). You don't have to provide me with every plot that you looked at. In most cases, it will be sufficient for you to merely state what you learned from each plot that you looked at. But you *can* include a few plots in your solutions if you want to draw on them to point something out to me that cannot be clearly summarized in a sentence or two. I'm not establishing a limit to the number of pages, but if you go with a long solution, it is crucial that it is well-organized and that your answers are clearly marked. (But even a long solution should not include too many plots. And *you should* **not** *give me any plot that isn't on one page.* (That is, don't give me a plot for which part of the plot is on one page and the rest of the plot is on another page.) Still, throughout the entire assignment, I think some plots are appropriate.)

This assignment covers material up through the end of Unit 2 of the class notes, but I don't expect you to try (and please don't use) any permutation tests, any of the tests based on trimmed means (except for the part for which you are specifically instructed to use a test based on trimmed means), or Fung's test from the Appendices. It *may* be the case (but it is not necessarily the case) that one or more of the problems is best dealt with using material from Unit 1. For each test of hypotheses, accurately report the smallest p-value that can be obtained without using a procedure which might be subject to an appreciably inflated type I error rate for the situation under consideration. Similarly, for the confidence intervals report the shortest interval which can be obtained using a valid method (and if no method seems ideal, select the method which should have the most accurate coverage probability), and for point estimates use the method which should be the most accurate given that the data is not terribly misleading about the general shapes of the underlying distributions. Select one answer for each part (as opposed to indicating that the answer is either *this* or *that*), and be sure to *clearly mark your final answers.* With regard to the number of significant digits used in reporting answers, rely on guidelines given previously unless it is explicitly stated to do something else for a particular part. (Typically, point estimates and confidence bounds should not indicate more accuracy than the individual data values do. For example, if the data values were rounded to the nearest tenth, it typically isn't sensible to give estimates using values rounded to the nearest hundreth.)

For Fisher's exact test (the two-sample median test), do not ignore values equal to the median of the combined sample, but instead count them as part of the "lower portion" of the combined sample. (Note: When using a hypergeometric distribution to obtain an exact p-value, it's not always the case that for a two-tailed test you can simply double a one-tail probability. In general, for a two-tailed Fisher's exact test, one sums all of the null probabilities less than or equal to the null probability associated with the observed outcome.) For the W-W runs test, break ties in such a way as to maximize the resulting p-value, and then use the distribution derived for the case of no ties. For the W-M-W test, break ties conservatively to arrive at an integer value for the test statistic if you have small sample sizes and need to use the exact null distribution. But otherwise, if both sample sizes are at least 10, and you use the normal approximation, use midranks to handle ties. *Be sure to do the nonparametric tests as I've instructed here* — I don't intend to make a special effort to check your results if you ignore the instructions and do them differently.

Pay close attention to what is being requested in each part. The data sets are available on the course web site, but I can send you the data by e-mail if you send me an e-mail request for it. (Note: Throughout this assignment, you are to make inferences about the distribution(s) from which the observed data arose (and it will be assumed that we are dealing with random samples resulting from iid random variables (from either one or two different distributions)).)

1) Consider the Chaetocnema data.

(a) Does this data provide statistically significant evidence that the distributions of widths are not the same? Respond to this query by reporting the smallest p-value which can be obtained from a valid test.

(b) Now address the general two sample problem by reporting the p-value that results from an appropriate

test based on trimmed means, trimming one observation off each end of each ordered sample. (Since I didn't spend much time on this in class, I'll add a few comments here. The first test encountered in Unit 2 based on the difference in sample trimmed means is often a decent possibility for the general two sample problem, especially if the sample sizes are the same. If the null hypothesis of identical distributions is true, and we have two samples of equal size from the common distribution, and I trim the two samples the same way to get a difference in trimmed means, then the difference in trimmed means should not be very large — where the assessment of "very large" is made using the appropriate standard error estimate and null sampling distribution of the standardized difference. Note that for the general two sample problem, if the null hypothesis is true, then the distribution variances are equal, and so we don't need the unequal variance version of the two sample test based on the difference in trimmed means.) It should be noted that the requested test for this part is not necessarily the best test to use — basically, I'm just interested in seeing if you can properly execute the procedure based on the trimmed means.

2) Consider the EDV data. Let μ_n denote the mean of the distribution of the EDV values for persons with normal cardiac function, and let μ_c denote the mean of the distribution of the EDV values for persons with constrictive pericarditis.

(a) Does this data provide statistically significant evidence that $\mu_c < \mu_n$.

(b) Give a 95% confidence interval for $\mu_n - \mu_c$.

3) Consider the M & M data. Does this data provide statistically significant evidence indicating that the underlying distributions differ? Respond to this query by reporting the smallest p-value which can be obtained from a valid test.

4) Consider the IQ data related to an investigation of the effect of birth order. Is there statistically significant evidence that birth order affects IQ? Report the smallest p-value which can be obtained using an appropriate test.