STAT 554: HW #4

due April 7, 2008

Please review the instructions given on the homework web page (HW #4 section) concerning the presentation of homework solutions. In order to maximize credit, you should *provide some sort of justification for your choice of inference procedure for each part.* (You may earn appreciably less than full credit for correct answers which are not adequately justified.) The various parts are each worth 2 points, and I'll count your best 10 of 11 parts to obtain a score out of 20 for the assignment. (For this assignment you may want to try some of the extra credit parts (in case you miss some things from the regular parts).)

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 *for each data set you should give at least one plot to support your conclusions*). But you *can* include extra 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. On the other hand, don't be *too* skimpy — I do want justification for your answers, and of course if you get something wrong then you might get more partial credit if you've shown more work and/or explanation. **Please limit your presentation to 12 pages** (or 6 sheets if you use both sides of the paper). *Extra credit parts should be handed in separately, not attached to the rest of your solutions* — they don't count with regard to the 6 sheet limit. (In general, it'd be a good idea to make full use of the 12 page limit. That is, if you have room, show more than one plot per data set (but if you don't have room for every plot that you looked at, if you indicate what you learned from the ones that you don't show me). In a nutshell, you should try to convince me that you've become comfortable with the material — that you can properly assess the situation and then correctly execute the proper procedure(s). In some cases you can explain any points of uncertainty that you encountered, and indicate to me why you made the choices that you did. Just be sure to give me a single final answer — not two or more answers — for each part. For each part, clearly indicate the p-value or estimate that you select, and clearly indicate the test or estimation method used to obtain your answer.) In some cases, several methods give about the same result, and in practice this means that the choice of method may not be so crucial. However, for this assignment, you should carefully select a method based on the information available, and you may be penalized if you get the "right answer" using a method that may generally be a bad choice given somewhat similar situations involving other data. (If data values are rather close to being symmetric, many point estimation methods can yield nearly the same value. Generally, I'm giving you fairly "nice" data sets and problems, and in some cases several methods give about the same result — but you should choose to report that your answer came from the method that you think is the best choice for similar cases for which the results may differ a bit more form method to method.)

In general, this assignment covers material contained in the first unit of the class notes, but do not use the permutation test or the one-sample normal scores test (unless explicitly instructed to) do so. Unless I specify otherwise, ignore values equal to $\xi^\star$ when doing the sign test and the signed-rank test. When using the normal approximation for the signed-rank test, use midranks if ties occur. For small sample sizes, the normal approximation should be avoided and let's use the conservative approach to break the ties, so that the value of the test statistic will be an integer and the exact sampling distribution can be used to obtain a p-value.

In general, for hypothesis tests, you are to report the smallest p-value that can be obtained from a valid method (from those that I've gone over in class). In the case that no method is guaranteed to be perfectly accurate, you should use a method that should be reasonably accurate, and certainly not appreciably anticonservative. Similarly, for confidence intervals you should use a method that provides as short of a confidence interval as is possible while also being a valid method (having an actual coverage probability as least as large as the nominal confidence coefficient). When providing point estimates, make use of the recommendations that I gave in class to choose an appropriate estimator. As always, round p-values to two (or perhaps one in some cases) significant digits. For confidence interval endpoints and point estimates, a good rule of thumb is to round to the place indicated by the second significant digit of the estimated standard error of the associated estimator. (For example, for the sample mean, the estimated standard error is $s/\sqrt{n}$, and if the value of this estimated standard error is 1.37, you would round the point estimate to the nearest tenth.) If the individual data points are rounded to the nearest integer, I may round the estimate to the nearest integer — breaking the rule of thumb, but being careful to not express more

precision than is warrented. For a confidence interval based on the sample mean, the confidence bounds can also be rounded using the place of the second significant digit of the estimated standard error or the associated point estimator. (Except for the sample mean and trimmed means, you may not know how to estimate the standard error. In such cases (say with the sample median or an M-estimator), you can use the estimated standard error of a trimmed mean as a substitute.)

Pay attention to whether I'm asking a question about the mean, median, or perhaps a treatment effect. Also, pay attention to whether I want a one-sided test or a two-sided test. If I'm just asking if the mean or median differs from some specified value without indicating a direction, then it's a two-sided hypothesis and you'll need to do a two-tailed test. If I'm just asking if there's a difference without mentioning the mean or the median, then I'm looking for evidence of a treatment effect. I also mean that you should focus on whether or not there is a treatment effect if I just ask whether the treatment affects the subjects or specimens. Note that one can have matched pairs data and still be interested in the mean or median (and if so, once you form the differences, you just proceed as if the differences are your sample of obervations that you're to use to make an inference about the mean or median (of the distribution underlying the observed differences)). So just because you see matched pairs data, it doesn't mean that you're always to just simply test for a treatment effect. (*Note*: Under the null hypothesis of no treatment effect you've got symmetry, but if you're estimating the mean or median of the difference distribution then you don't necessarily have that there was no treatment effect and so the distribution that you're dealing with could be skewed. If we believe that there is no treatment effect, then it really doesn't make sense to estimate the mean or median (because it'd be 0). Now in the case of a null hypothesis of the mean or median being 0 with matched pairs differences, one doesn't necessarily have symmetry even then, because there could be a treatment effect that created skewness but left the mean or median equal to 0. (However, at times I might be inclined to think that it would be odd to have a treatment effect that left the mean equal to 0, and so I might lean towards symmetry under such a null hypothesis. There are no good hard and fast rules about this though. I guess it's safe to say that I wouldn't make an assumption of symmetry if the data gave a clear indication of skewness. It's just that I might lean towards symmetry a bit more than I would otherwise. (In general, for tests and confidence intervals of a distribution mean, I tend to lean towards skewness in cases that aren't strongly supportive of symmetry or at least very close approximate symmetry. For point estimates, exact symmetry isn't so crucial when you're working with a small sample size, in which case one should should try to properly trade off between variance and bias — and for appreciably heavy-tailed distributions, mild asymmetry may be of secondary concern.)))

Since I'm giving you some warnings/suggestions/hints/points of clarification here, I don't intend to give you much help with the specific problems of this assignment. I want to see what you can do on your own. Of course, you should feel free to discuss general things with me and get clarification on anything from the class notes or a previous assignment — it's just that I don't intend to tell you what to do for any particular part of this assignment.

1) Consider the data set of air-conditioning system failure times (which I will supply separately). (*Note*: The homework page of my 554 web site has a link to the data. Alternatively, I can e-mail you the data sets if you send me an e-mail request.)

(a) Give a point estimate for the mean of the distribution underlying the data.

(b) Is there statistically significant evidence that the mean of the underlying distribution is less than 100? Respond to this query by reporting the smallest p-value which results from an appropriate test. (*Note*: In general, if there is more than one test which is clearly valid, simply choose the smallest p-value that a valid test can produce. On the other hand, if there is some question as to the validity of all of the testing procedures that you're supposed to be considering, then you should place your emphasis on accuracy rather than the values of the resulting p-values. I want as honest of an assessment as is possible of the strength of the evidence against the null hypothesis.)

(c) Is there statistically significant evidence that the median of the underlying distribution is less than 100? Respond to this query by reporting the smallest p-value which results from an appropriate test.

(d) Give a 90% confidence interval for the median of the distribution underlying the data.

(e) Give a point estimate of the 90th percentile of the distribution underlying the data.

2) Consider the housing insulation data sets (which I will supply separately), and in particular consider

the third of the four data sets (the one where comparable households were paired). Give a 99% confidence interval for the mean energy saving which results from using extra insulation instead of standard insulation.

3) Consider the FEV1 data (which I will supply separately).
(a) Give a 99% confidence interval for the mean of the distribution underlying the data.
(b) Give a point estimate of the 90th percentile of the distribution underlying the data.

4) Given with this assignment are observations of a relatively large number of variables for a group of 20 control subjects in a study of the relationships between fitness characteristics and cholesterol levels. (The 20 control subjects are labeled Observation 18 through Observation 37 in the table of data for Group 2 (the control group). (I'm not giving you the data for the other 3 groups of subjects at this time.)) For this problem, you are to concern yourself with just one of the many variables, the one labeled X_{23} (which is a certain type of cholesterol measurement), and you are to use the 20 observations of this particuar cholesterol measurement to make inferences about the underlying population (for the corresponding cholesterol measurements of all people who would be eligible for the control group). Give a point estimate of the median of the distribution underlying the observed sample.

5) Consider the residual flame time data (which I will supply separately).
(a) Is there statistically significant evidence that the mean of the underlying distribution is less than 9.9? Respond to this query by reporting the smallest p-value which results from an appropriate test.
(b) Is there statistically significant evidence that the median of the underlying distribution is less than 9.9? Respond to this query by reporting the smallest p-value which results from an appropriate test.

Extra Credit

Remember that you are suppose to work the extra credit parts entirely on your own — don't get help from anyone, and don't compare answers with anyone. Each of the three items below is worth 0.5 point, and to earn full credit you need to show an adequate amount of work (so don't just give the answers without supporting work). *Remember to not attach the extra credit solutions to the rest of your solutions — staple the extra credit solutions you turn in together, but do not staple them to the rest of your solutions.*

A) For the air-conditioning system data, test $H_0 : \xi_{0.9} \leq 100$ against $H_1 : \xi_{0.9} > 100$ and report the p-value. (There are two good ways for you to go about this test. One way would be to modify the sign test. That is, let the test statistic be the number of observations exceeding 100, but note that the null distribution that you should use is not the same as you would use for the sign test since here we're considering a test about the 90th percentile and not the distribution median. Alternatively, you can recast the problem as a test about a population proportion and solve it using HW #1 techniques. Both ways result in the same p-value.)

B) For the housing insulation data set used in Problem 2, test the null hypothesis that the choice of insulation type has no effect on energy use against the general alternative, reporting the p-value which results from the approximate version of the normal scores test. (I'll give you a description and example.)

C) For the housing insulation data set used in Problem 2, test the null hypothesis that the choice of insulation type has no effect on energy use against the general alternative, reporting the p-value which results from a test based on a 12.5% trimmed mean.