Empirical Methods for Computer Science (CS 5970) Homework 3 Solutions

December 20, 2008

Question 1

In this question, we will look at dealing with censoring of samples and at the use of the single sample bootstrap test. Implement a matlab function that shows the power of three distinct statistical tests as a function of the assumed mean of the underlying distribution. The three tests are:

- Bootstrap test that explicitly acknowledges censored values in the sampling process
- Bootstrap test that does not acknowledge censored values (i.e., censored elements are removed from the sample before the bootstrap test is performed).
- t-test (remove censored elements from the test before performing the test).

The following functions are provided:

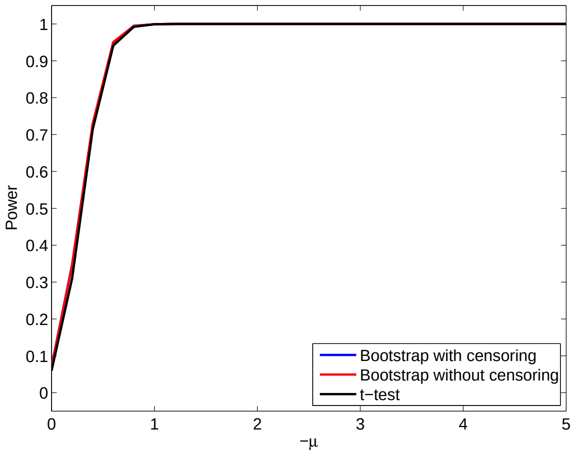
- randn_bogotify($N, Q, prob$) produces an NxQ matrix of samples from a standard normal distribution, except: with probability $prob$, the samples are set to NaN.

- rand_multi_bogotify$(N, Q, prob)$ produces an NxQ matrix of samples from some distribution. As above, with probability $prob$, the samples are set to NaN.
- mean_safe(x): if x is a matrix, computes the mean of the columns of x; if a vector, computes the mean of the elements in the vector. In both cases, samples that are NaN are removed from the mean.

Make the following assumptions:

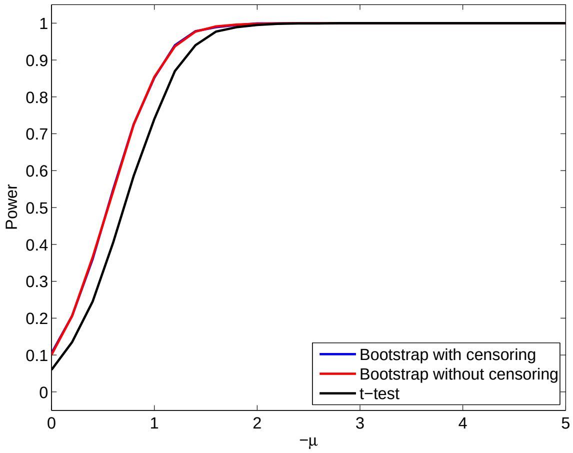
- $N = 7$ or $N = 30$: number of items in the sample.
- $Q = 1000$: number of times that we will take N samples and perform the statistical tests in order to estimate the power of the tests.
- $M = 1000$: the number resamples that we will take for the bootstrap tests.
- All tests are right-tailed.
- $\mu = [-5...0]$: The assumed mean of the underlying distribution for all three tests. (Note that $\mu = 0$ corresponds to the case in which there is no difference between the null and alternative hypothesis distributions.
- $\alpha = 0.05$.

1. (10pts) For the normal distribution, $N = 30$, and $prob = 0$, show power for each of the tests as a function of the assumed mean. Discuss the similarities/differences.



With such a large N and with a normal distribution, we expect that the three tests should behave the same.

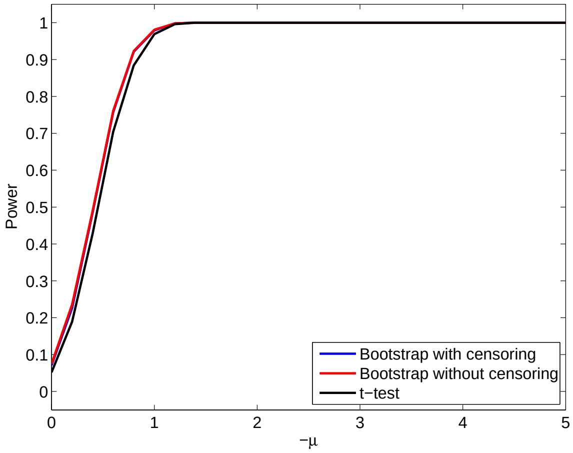
2. (10pts) For the normal distribution, $N = 7$, and $prob = 0$, show power for each of the tests as a function of the assumed mean. Discuss the similarities/differences.



The two bootstrap tests perform comparably. However, they have more power than the t-test. This is likely due to the fact that at N=7, the sample is so small that it is not very representative of the underlying distribution. One red flag: when $\mu = 0$ (when both the null and alternative hypotheses are the same), the power is above 0.1. *For this case, we would hope that our power was the same as α.*

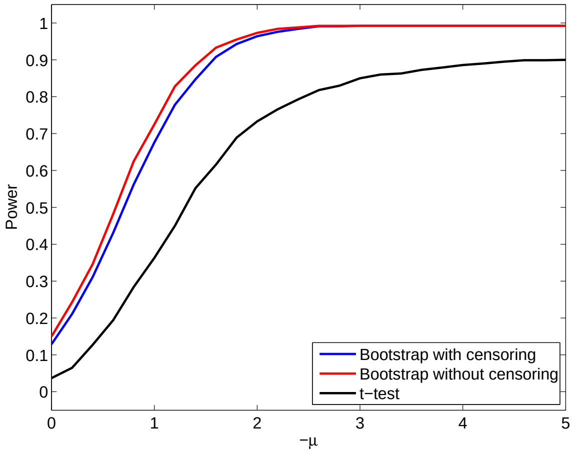
3. (10pts) For the normal distribution, $N = 30$ and $N = 7$, and $prob = 0.5$, show power for each of the tests as a function of the assumed mean. Discuss the similarities/differences.

   $N = 30$*:*



As with the large sample case above, there is little difference between the three tests. However, we are starting to see a little separation between the t-test and the bootstrap tests. This is due to the fact that our effective sample size is $N = 15$ *with the censoring.*

$N = 7$*:*

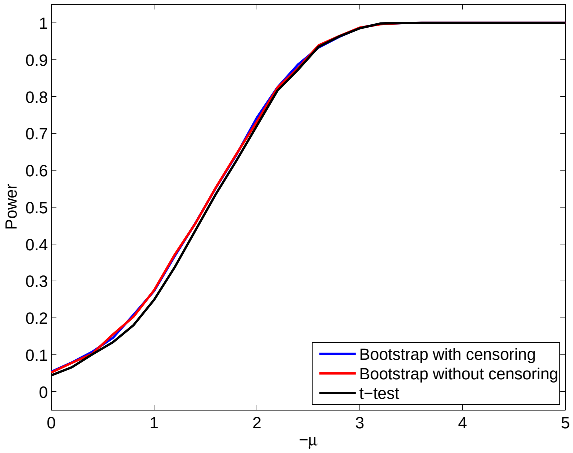


Our effective sample size is $N = 3.5$*. This exacerbates the difference between the t-test and the bootstrap tests. Note that we are starting to see a little bit of separation between bootstrap with censoring and without.*

4. (10pts) For the non-normal distribution, $N = 30$ and $N = 7$, and $prob = 0$, show power for each of the tests as a function of the assumed

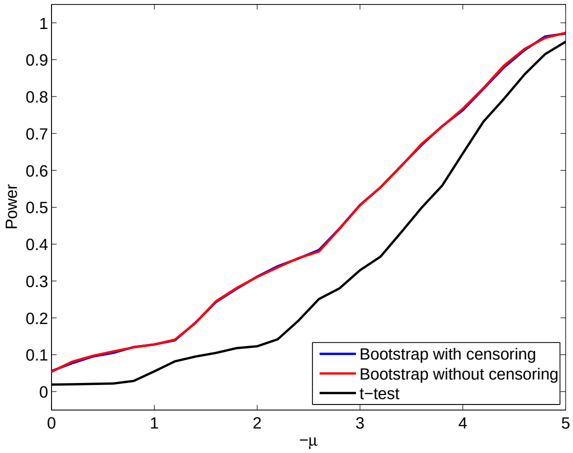
mean. Discuss the similarities/differences. Compare to the normal case.

$N = 30$:



There is little difference between the three tests (unexpectedly). However, the tests require a larger separation between the means before the power asymptotes. This is due to the fact that the standard deviation of this distribution is wider than the normal distribution from above.

$N = 7$:

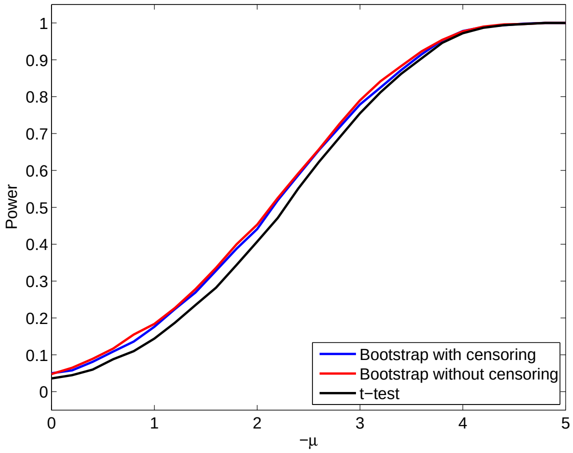


The power of the t-test is reduced. This is presumably due to the fact that the underlying distribution is far from normal. Note that the power for the bootstrap tests when $\mu = 0$ is approximately 0.05 (which is where we want it).

5. (10pts) For the non-normal distribution, $N = 30$ and $N = 7$, and $prob = 0.5$, show power for each of the tests as a function of the assumed mean. Discuss the similarities/differences. In the latter figure, explain the inflection point at $\mu = -2.5$

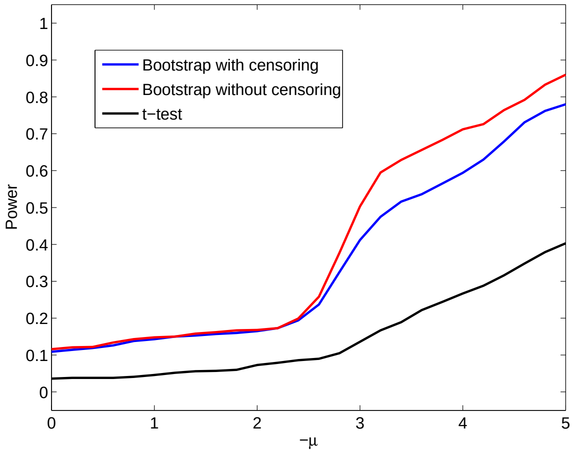
The effective sample size is halved for these cases, so the effects are amplified.

$N = 30$*:*



The t-test is slightly less powerful than the two bootstrap tests

$N = 7$*:*



The power curves have an odd inflection point at $\mu = -2.5$*. This is due to the fact that the distribution has two modes. With effective sample sizes of* $N = 3.5$*, it is easy for a sample to only include samples from* *only one mode. Once the means are separated enough, we can tell the difference between the high mode of the null hypothesis and the low mode of the alternative.*

6. (10pts) Would you rather use the version of the bootstrap test that does or does not acknowledge censoring?

   It is not entirely clear given the experiments that we have performed. In most cases, they perform the same. Where they do perform differently, the non-censoring test exhibits higher power. One downside to this is that this is true even when $\mu = 0$ (the case where power should be identical to α.

Question 2

dat_pre and **dat_post** are scores that are received for a set of 20 different tests. Each row of these matrices represents a set of samples corresponding to a single individual. Note, however, that individuals are not paired across the pre and post conditions. You may consider the scores to be drawn from continuous random variables for the purposes of our analysis (they are actually interval variables).

We wish to determine whether the post tests perform better than the pre tests.

1. (10pts) Use the appropriate t-test to compare the performance for each of the conditions. For which conditions is there a significant difference in performance (assume $\alpha = 0.15$)?

   Condition 5 ($p < 0.017$), condition 16 ($p < 0.11$), and condition 17 ($p < 0.06$) show a significant difference

2. (20pts) Using the bootstrap randomization method (your own implementation), for which conditions is there a significant difference? (again, assume $\alpha = 0.15$)

   - *Condition 3: $p < 0.136$.*

- *Condition 4:* $p < 0.135$.
- *Condition 5:* $p < 0.016$.
- *Condition 6:* $p < 0.142$.
- *Condition 14:* $p < 0.149$.
- *Condition 15:* $p < 0.122$.
- *Condition 16:* $p < 0.081$.
- *Condition 17:* $p < 0.049$.

3. (10pts) Under this experiment and analysis design, do we have a multiple comparisons problem?

   No: the different tests are independent of one-another.