CSE 2320 Experimentation Assessment Project

Design Submission: November 18, 2010
Final Submission: December 7, 2010

Goal:

Demonstration of "the ability to design and conduct experiments, analyze and interpret data" (ABET outcome b) based on the following narrative:

In the selection problem, the *k*th largest value in an unordered array with *n* values (`int`s) must be determined. Two common ways to perform this task are to: 1) apply PARTITION (Notes 8.B) or 2) apply a MSD-radix sort, but discarding all but one of the "bins" at each digit position ("multi-way" radix partitioning). The third phase of each counting sort will reveal the bin that will receive the *k*th largest value.

Your task is to ***compare*** the time and space required by the two methods and provide general principles for choosing which method should be used in a given situation. The number of keys (*n*) should be varied drastically, but the range of the random keys should always be 0 . . . 1,999,999,999. When using the MSD-radix sort, there are a number of radix options that may be tried.

Requirements:

The following requirements (with weights for the two submissions) are to be satisfied by submitting a preliminary report (parts 1, 2, and 3) by 9:15 a.m. on November 18 (graded by November 30) and a final report (parts 1, 2, 4, 5, 6) by 9:15 a.m. on December 7. Both submissions are to be sent as e-mail attachments to `huawang2007@mavs.uta.edu`.

1. Proposed solution and background information. (10+5=15%) You may assume your reader has a copy of Sedgewick, so this section should be short.

2. Testable hypotheses. (10+5=15%) These should be relevant to the task of comparing the methods. Some preliminary executions will be useful for formulating hypotheses and (1.).

3. Java code for collecting performance statistics. (30+0=30%) This will not be long, but should still follow the expectations for code in the course syllabus.

4. Description of the collected data. (0+10=10%) Be sure that someone reading your report gets a good overview.

5. Collected data as tables or graphs. (0+15=15%) Since you will need to work with fairly large tables, summarized data will be useful.

6. Conclusions with support from the data. (0+15=15%) Consideration of errors and discussion of possible additional work.

Getting Started:

1. The package `java.util.Date` is convenient for capturing elapsed time for sections of code. The following code times `Arrays.sort()`:

```
Date start=new Date();
Arrays.sort(arr);
Date stop=new Date();
double seconds=(stop.getTime()- start.getTime())/1000.0;
System.out.format("Arrays.sort for %d ints took %f seconds\n",n,seconds);
```

`System.nanoTime()` may also be used.

2. Pseudorandom numbers in the desired range should be generated using the class `java.util.Random`.

Grading Rubric:

C1: Questions and related background show that student clearly understands the issues to examine.

5: Summary indicates that student understands the issues and what should be explored.
3: Summary indicates that student should learn something from their experiments.
1: Not clear that student will be performing an organized experiment.

C2: Student has decomposed the problem into one or more experimentally testable hypotheses.

5: Hypotheses indicate that student performed initial work leading to testable hypotheses for the given problem.
3: Hypotheses indicate that initial work was hastily performed.
1: Hypotheses are flawed and may not be testable.

C3: The components (e.g. code for 2320) have been implemented, tested, and could be used by others.

5: Components are designed to be useful for anyone performing related experiments.
3: Components are insufficient for performing necessary experiments.
1: Components are incorrect.

C4: The nature of the collected data is accurately described.

5: Quantity of data is appropriate, but not overwhelming, for drawing conclusions.
3: Quantity of data is barely sufficient for drawing conclusions.
1: Experimental set-up precludes obtaining data.

C5: Data is presented appropriately in tables or graphical forms.

5: Trends in the data are clearly identifiable.
3: Trends appear, but are not obvious.
1: Data is very incomplete.

C6: Conclusion(s) regarding hypotheses are presented with support from the data. Possible errors are noted. Remaining issues are discussed, along with additional experimental work that could be performed.

5: Conclusions for all hypotheses, with support from data.
3: Conclusions are related to hypotheses, but questionable support from data.
1: Conclusions related to hypotheses are lacking.