Biostatistics 615/815 - Lecture 2
Introduction to C++ Programming

Hyun Min Kang

September 8th, 2011

BIOSTAT615/815 - Objectives

1. Equip the ability to IMPLEMENT computational/statistical IDEAS into working SOFTWARE PROGRAMs
2. Learn COMPUTATIONAL COST management in developing statistical methods.
3. Understand NUMERICAL and RANDOMIZED ALGORITHMS for statistical inference

Algorithm SingOldMacDonald

Data: $animals[1 \cdots n]$, $noises[1 \cdots n]$
Result: An "Old MacDonald" Song with $animals$ and $noises$
for $i = 1$ **to** n **do**
  Sing "Old MacDonald had a farm, E I E I O";
  Sing "And on that farm he had some $animals[i]$, E I E I O";
  Sing "With a $noises[i]$ $noises[i]$ here, and a $noises[i]$ $noises[i]$ there";
  Sing "Here a $noise[i]$, there a $noise[i]$, everywhere a $noise[i]$ $noise[i]$";
  for $j = i - 1$ **downto** 1 **do**
    Sing "$noise[j]$ $noise[j]$ here, $noise[j]$ $noise[j]$ there";
    Sing "Here a $noise[j]$, there a $noise[j]$, everywhere a $noise[j]$ $noise[j]$";
  end
  Sing "Old MacDonald had a farm, E I E I O.";
end

(adapted from Jeff Erickson(UIUC)'s class notes)

Key Idea of Insertion Sort

- For k-th step, assume that elements $a[1], \cdots, a[k-1]$ are already sorted in order.
- Locate $a[k]$ between index $1, \cdots, k$ so that $a[1], \cdots, a[k]$ are in order
- Move the focus to $k+1$-th element and repeat the same step

Algorithm InsertionSort

Data: An unsorted list $A[1 \cdots n]$
Result: The list $A[1 \cdots n]$ is sorted
for $j = 2$ **to** n **do**
    $key = A[j]$;
    $i = j - 1$;
    while $i > 0$ *and* $A[i] > key$ **do**
        $A[i+1] = A[i]$;
        $i = i - 1$;
    end
    $A[i+1] = key$;
end

Recap - Tower of Hanoi Problem

Problem

Input
- A (leftmost) tower with n disks, ordered by size, smallest to largest
- Two empty towers

Output Move all the disks to the rightmost tower in the original order

Condition
- One disk can be moved at a time.
- A disk cannot be moved on top of a smaller disk.

Key Idea - Think Recursively

- Move the other $n-1$ disks from the leftmost to the middle tower
- Move the largest disk to the rightmost tower
- Move the other $n-1$ disks from the middle to the rightmost tower

A Recursive Algorithm for the Tower of Hanoi Problem

Algorithm $\textsc{TowerOfHanoi}$

Data: n : # disks, (s, i, d) : source, intermediate, destination towers
Result: n disks are moved from s to d

if $n == 0$ **then**
  do nothing;
else
  $\textsc{TowerOfHanoi}(n-1, s, d, i)$;
  move disk n from s to d;
  $\textsc{TowerOfHanoi}(n-1, i, s, d)$;
end

Today's and Next Lectures

Today

- Basic Data Types
- Control Structures
- Pointers and Functions

Next few lectures

- The class does NOT focus on teaching programming language itself
- Expect to spend time to be familar to programming languages yourself
  - ✓ Online reference : *http://www.cplusplus.com/doc/tutorial/*
  - ✓ Offline reference : C++ Primer Plus, 5th Edition
- VERY important to practice writing code on your own.
- Utilize office hours or after-class minutes for detailed questions in practice

Example C++ Development Environment

1. UNIX / gcc environment
   - Instructor's preference
   - UNIX environment will be commonly used in large-scale data analysis, so it would be good to be familiar with it.
   - Ways to set up UNIX environment
     - Install Linux (e.g. Ubuntu) locally to your computer
     - Download and install Xcode in Mac OS X, and use terminal to access UNIX interface.
     - Install Cygwin to a windows machine (mimics UNIX environment)
     - Connect to U-M login service via SSH using PuTTY or similar software (Refer to *http://www.itd.umich.edu/login/* for details)
   - Learning Unix-cultured editors such as vi or emacs is also recommended.

Example C++ Development Environment

1. UNIX / gcc environment
2. Windows / Microsoft Visual C++
3. Windows / Borland C++ Builder
4. Mac OS X / Xcode development environment

Getting Started with C++

Writing helloWorld.cpp

```
#include <iostream> // import input/output handling library
int main(int argc, char** argv) {
  std::cout << "Hello, World" << std::endl;
  return 0; // program exits normally
}
```

Compiling helloWorld.cpp

```
user@host:~/$ g++ -o helloWorld helloWorld.cpp
```

Running helloWorld

```
user@host:~/$ ./helloWorld
Hello, World
```

How helloWorld works

main() : First function to be called

```
// type of return value is integer
// return value of main() function is program exit code
// 0 is normal exit code and the others are abnormal exit codes
int
// name (identifier) of function is 'main'
// 'main' is a special function, invoked at the beginning of a program.
main
(                // function arguments are surrounded by parentheses
  int argc,      // number of command line arguments
  char** argv    // list of command line arguments - will explain later
)
{
  // ... function body goes here
  return 0; // return normal exit code
}
```

How helloWorld works

Using iostream to output strings to console

```
// includes standard library for handling I/Os (inputs/outputs)
// std::cout and std::endl cannot be recognized without including <iostream>
#include <iostream>
int main (int argc, char** argv) {
  std::cout // standard output stream - messages are printed to console.
    << // insertion operator : appends the next value to the output stream
    "Hello, World" // string appended to std::cout via operator<<
    << // insertion operator : appends the next value to the output stream
    std::endl; // end-of-line appended to std::cout via operator<<
  return 0;
}
```

Implementing TowerOfHanoi Algorithm in C++

towerOfHanoi.cpp

```
#include <iostream>
#include <cstdlib>
// recursive function of towerOfHanoi algorithm
void towerOfHanoi(int n, int s, int i, int d) {
  if ( n > 0 ) {
    towerOfHanoi(n-1,s,d,i); // recursively move n-1 disks from s to i
    // Move n-th disk from s to d
    std::cout << "Disk " << n << " : " << s << " -> " << d << std::endl;
    towerOfHanoi(n-1,i,s,d); // recursively move n-1 disks from i to d
  }
}
// main function
int main(int argc, char** argv) {
  int nDisks = atoi(argv[1]); // convert input argument to integer
  towerOfHanoi(nDisks, 1, 2, 3); // run TowerOfHanoi(n=nDisks, s=1, i=2, d=3)
  return 0;
}
```

Running TowerOfHanoi Implementation

Running towerOfHanoi

```
user@host:~/$ ./towerOfHanoi 3
Disk 1 : 1 -> 3
Disk 2 : 1 -> 2
Disk 1 : 3 -> 2
Disk 3 : 1 -> 3
Disk 1 : 2 -> 1
Disk 2 : 2 -> 3
Disk 1 : 1 -> 3
```

Homework 0

- Implement the following two programs and send the output screenshots to the instructor (hmkang at umich dot edu) by E-mail
  1. HelloWorld.cpp
  2. TowerOfHanoi.cpp
- Briefly describe your operating system and C++ development environment with your submission
- This homework will not be graded, but mandatory to submit for everyone who wants to take the class for credit
- No due date, but homework 0 must be submitted prior to submitting any other homework.

Declaring Variables

Variable Declaration and Assignment

```
int foo; // declare a variable
foo = 5; // assign a value to a variable.
int foo = 5; // declararion + assignment
```

Variable Names

```
int poodle; // valid
int Poodle; // valid and distinct from poodle
int my_stars3; // valid to include underscores and digits
int 4ever;  // invalid because it starts with a digit
int double; // invalid because double is C++ keyword
int honky-tonk; // invalid -- no hyphens allowed
```

Basic Digital Units

bit: A single binary digit number which can represent either 0 or 1

byte: A collection of 8 bits which can represent $256 (= 2^8)$ unique numbers. One character can typically be stored within one byte.

word: An ambiguous term for the natural unit of data in each processor. Typically, a word corresponds to the number of bits to represent a memory address. In 32-bit address scheme which can represent up to 4 gigabytes, 32 bits (4 bytes) are spent to represent a memory address. In 64-bit address scheme, up to 18 exabytes can be represented by using 64 bits (8 bytes) to represent a memory address.

Data Types

Signed Integer Types

```
char foo;  // 8 bits (1 byte) : -128 <= foo <= 128
short foo; // 16 bits (2 bytes) : -32,768 <= foo <= 32,767
int foo;   // Mostly 32 bits (4 bytes) : -2,147,483,648 <= foo <= 2,147,483,647
long foo;  // 32 bits (4 bytes) : -2,147,483,648 <= foo <= 2,147,483,647
long long foo; // 64 bits
short foo = 0;  foo = foo - 1;   // foo is -1
```

Unsigned Integer Types

```
unsigned char foo; 8 bits (1 byte) : 0 <= foo <= 255
unsigned short foo; // 16 bits (2 bytes) : 0 <= foo <= 65,535
unsigned int foo;   // Mostly 32 bits (4 bytes) : 0 <= foo <= 4,294,967,295
unsigned long foo;  // 32 bits (4 bytes) : 0 <= foo <= 4,294,967,295
unsigned long long foo; // 64 bits
unsigned short foo = 0;  foo = foo - 1;   // foo is 65,535
```

Floating Point Numbers

Comparisons

Type	`float`	`double`	`long double`
Precision	Single	Double	Quadruple
Size	32 bits	64 bits	128 bits
(in most modern OS)	4 bytes	8 bytes	16 bytes
Sign	1 bit	1 bit	1 bit
Exponent	8 bits	11 bits	15 bits
Fraction	23 bits	52 bits	112 bits
(# decimal digits)	7.2	16	34
Minimum (>0)	1.2×10^{-38}	2.2×10^{-308}	3.4×10^{-4932}
Maximum	3.4×10^{38}	1.8×10^{308}	1.2×10^{4932}

Handling Floating Point Precision Carefully

precisionExample.cpp

```cpp
#include <iostream>
int main(int argc, char** argv) {
  float smallFloat = 1e-8; // a small value
  float largeFloat = 1.;   // difference in 8 (>7.2) decimal figures.
  std::cout << smallFloat << std::endl; // "1e-08" is printed
  smallFloat = smallFloat + largeFloat; // smallFloat becomes exactly 1
  smallFloat = smallFloat - largeFloat; // smallFloat becomes exactly 0
  std::cout << smallFloat << std::endl; // "0" is printed
  // similar thing happens for doubles (e.g. 1e-20 vs 1).
  return 0;
}
```

Basics of Arrays and Strings

Array

```
int A[] = {3,6,8}; // A[] can be replaced with A[3]
std::cout << "A[0] = " << A[0] << std::endl; // prints 3
std::cout << "A[1] = " << A[1] << std::endl; // prints 6
std::cout << "A[2] = " << A[2] << std::endl; // prints 8
```

String as an array of characters

```
char s[] = "Hello, world"; // or equivalently, char* s = "Hello, world"
std::cout << "s[0] = " << s[0] << std::endl; // prints 'H'
std::cout << "s[5] = " << s[5] << std::endl; // prints ','
std::cout << "s = " << s << std::endl; // prints "Hello, world"
```

Summary - Data Types and Precisions

- Each data type consumes different amount of memory
  - For example, 1GB can store a billion characters, and 125 million double precision floating point numbers
  - To store a human genome as character types, 3GB will be consumed, but 12GB will be needed if each nucleotide is represented as an integer type
- Precision is not unlimited.
  - Unexpected results may happen if the operations require too many significant digits.

Assignment and Arithmetic Operators

```
int a = 3, b = 2;  // valid
int c = a + b;     // addition : c == 5
int d = a - b;     // subtraction : d == 1
int e = a * b;     // multiplication : e == 6
int f = a / b;     // division (int) generates quotient : f == 1
int g = a + b * c; // precedence - add after multiply : g == 3 + 2 * 5 == 13
a = a + 2;         // a == 5
a += 2;            // a == 7
++a;               // a == 8
a = b = c = e;     // a == b == c == e == 6
```

Comparison Operators and Conditional Statements

```
int a = 2, b = 2, c = 3;
std::cout << (a == b) << std::endl; // prints 1 (true)
std::cout << (a == c) << std::endl; // prints 0 (false)
std::cout << (a != c) << std::endl; // prints 1 (true)
if ( a == b ) {  // conditional statment
  std::cout << "a and b are same" << std::endl;
}
else {
  std::cout << "a and b are different" << std::endl;
}
std::cout << "a and b are " << (a == b ? "same" : "different") << std::endl
  << "a is " << (a < b ? "less" : "not less") << " than b" << std::endl
  << "a is " << (a <= b ? "equal or less" : "greater") << " than b" << std::endl;
```

Loops

while loop

```
int i=0;  // initialize the key value
while( i < 10 ) { // evaluate the loop condition
  std::cout << "i = " << i << std::endl; // prints i=0 ... i=9
  ++i; // update the key value
}
```

for loop

```
for(int i=0; i < 10; ++i) { // initialize, evaluate, update
  std::cout << "i = " << i << std::endl; // prints i=0 ... i=9
}
```

Pointers

s = 0x3fb5						

Address	0x3fb5	0x3fb6	0x3fb7	0x3fb8	0x3fb9	0x3fba
Value	'H'	'E'	'L'	'L'	'O'	'\0'

Another `while` loop

```
char* s = "HELLO"; // array of {'H','E','L','L','O','\0'}
while ( *s != '\0' ) {  // *s access the character value pointed by s
  std::cout << *s << std::endl; // prints 'H','E','L','L','O' at each line
  ++s; // advancing the pointer by one; points to the next element
}
```

Pointers and Loops

while loop

```
char* s = "HELLO"; // array of {'H','E','L','L','O','\0'}
while ( *s != '\0' ) {
  std::cout << *s << std::endl; // prints 'H','E','L','L','O' at each line
  ++s; // advancing the pointer by one
}
```

for loop

```
// initialize array within for loop
for(char* s = "HELLO"; *s != '\0'; ++s) {
  std::cout << *s << std::endl; // prints 'H','E','L','L','O' at each line
}
```

Pointers are complicated, but important

```
int A[] = {3,6,8}; // A is a pointer to a constant address
int* p = A;        // p and A are containing the same address
std::cout << p[0] << std::endl;   // prints 3 because p[0] == A[0] == 3
std::cout << *p << std::endl;   // prints 3 because *p == p[0]
std::cout << p[2] << std::endl; // prints 8 because p[2] == A[2] == 8
std::cout << *(p+2) << std::endl;   // prints 8 because *(p+2) == p[2]
int b = 3;     // regular integer value
int* q = &b;  // the value of q is the address of b
b = 4;         // the value of b is changed
std::cout << *q << std::endl;  // *q == b == 4

char s[] = "Hello";
char *t = s;
std::cout << t << std::endl; // prints "Hello"
char *u = &s[3]; // &s[3] is equivalent to s + 3
std::cout << u << std::endl; // prints "lo"
```

Pointers and References

```
int a = 2;
int& ra = a;   // reference to a
int* pa = &a;  // pointer to a
int b = a;     // copy of a
++a;  // increment a
std::cout << a << std::endl;  // prints 3
std::cout << ra << std::endl; // prints 3
std::cout << *pa << std::endl; // prints 3
std::cout << b << std::endl; // prints 2
int* pb;        // valid, but what pb points to is undefined
int* pc = NULL; // valid, pc points to nothing
std::cout << *pc << std::endl; // Run-time error : pc cannot be dereferenced.
int& rb;        // invalid. reference must refer to something
int& rb = 2;    // invalid. reference must refer to a variable.
```

Summary so far

- Algorithms are computational steps
- `towerOfHanoi` utilizing recursions
- `insertionSort`
  - ✓ Simple but a slow sorting algorithm.
  - ✓ Loop invariant property
- Data types and floating-point precisions
- Operators, `if`, `for`, and `while` statements
- Arrays and strings
- Pointers and References
- At Home : Reading material for novice C++ users : *http://www.cplusplus.com/doc/tutorial/*

At Home : Write, Compile and Run..

The following list of programs

- `helloWorld.cpp`
- `towerOfHanoi.cpp`
- `callByValRef.cpp`
- `precisionExample.cpp`

How to ...

Write Notepad, Vim, Emacs, Eclipse, VisualStudio, etc

Compile `g++ -Wall -o [progName] [progName].cpp`
(Unix, Mac OS X, or Cygwin)

Run `./[progName] [list of arguments]`

Next Lecture

- Fisher's Exact Test
- More on C++ Programming
  - Standard Template Library
  - User-defined data types