22S:166 Computing in Statistics
Fall 2011
Final Exam

You may submit your answers either as:

- a plain text file:
  Number the questions; include all R and SAS code and the output requested, and put your name as the first line of the file.
- or, for 3 percentage points extra credit, as an .Rnw file and PDF file:
  embed R code into the .Rnw file, and
  just copy and paste SAS code and output into a `verbatim` environment.

Run R (and possibly LaTeX) on the Linux machines. You may run SAS wherever is most convenient for you, including the Virtual Desktop.

1. R and R documentation

   (a) Create an R function to accept a numeric argument and to calculate:

   $$f(x) = \frac{1}{2x^2}$$

   Submit your code.

   (b) Read the documentation of the function `integrate` that is in the R package `stats`. Then use the `integrate` function to numerically evaluate the integral of your function over the interval (0.01,5). Submit your code and output.

2. SAS 1

   Two datasets on the course webpage represent a tiny part of a bank's database. They are:

   ```
   accounts.dat

   Variables:
        Account number
        Customer number
        Date opened
   ```

   and

   ```
   checks.dat

   Variables:
        Account number
        Check number
        Date
        Amount
   ```

   (a) Are the two tables represented by these datasets in first normal form? Submit a yes-or-no answer, and a brief explanation.

(b) Read both of these datasets into SAS. Add a variable to the `checks` dataset called `type`, with the value "Chk" in every record.

(c) Combine the two datasets, and print the result. It should have these records in it, although you don't have to use the same variable names.

Obs	acctno	custno	dateopen	checkno	date	amt	type
1	1001	2016	10/15/10	101	07/15/11	25.00	Chk
2	1001	2016	10/15/10	102	08/03/11	58.17	Chk
3	1001	2016	10/15/10	103	08/04/11	10.00	Chk
4	1002	3178	01/15/11	101	07/15/11	177.50	Chk
5	1003	4207	02/28/11	101	08/05/11	281.98	Chk

Submit all of your SAS code, and also the listing of the final dataset.

3. SAS 2

A dataset called `bed.dat` is posted under Datasets on the course web page. Read the documentation in `bed.info`.

(a) Write a SAS data step that:

i. Reads the `bed.dat` data file into SAS.

ii. Creates a new variable called `pctchg` that represents the percent change in beds per 1000 population from 1980 to 1986. Note that percent change is calculated as

$$pctchg = \frac{new - old}{old} \times 100$$

Submit the SAS code for the data step.

(b) Write a `tabulate` procedure step to produce the following table:

	N	Mean	StdDev
Beds 1980	51	4.56	1.01
Beds 1986	51	4.23	1.11
Percent change	51	-7.55	7.49

Please match your row labels and formatting for numbers to my example. If you are unable to get the three rows to appear in the table, include just the row for percent change.

Submit your SAS code and the table it outputs.

(c) Now write a SAS macro to produce tables like the one above. The macro should always output three rows, one for each variable. However, it should allow the user to specify when calling the macro which summary statistics should appear in the table.

Write two calls to your macro, which produce the two tables below:

	N	Mean	StdDev
Beds 1980	51	4.56	1.01
Beds 1986	51	4.23	1.11
Percent change	51	-7.55	7.49

	N	Min	Median	Max
Beds 1980	51	2.70	4.50	7.40
Beds 1986	51	2.40	4.20	7.70
Percent change	51	-19.15	-8.33	22.22

Submit your SAS code for defining and calling the macro, as well as the output (the two tables).

4. The jackknife, and more

   (a) Now read the `bed` dataset (same one used in the SAS problem) into a data frame in R. Submit your code.

   (b) Use the `sample` function to create a vector called `beds10` by drawing a random sample of size 10 (without replacement) from the column of the dataset representing beds in 1980. Include in your code whatever is required so that, if I run it, I will get exactly the same sample values that you did. Submit your code, and also display the vector `beds10`.

   (c) The vector `beds10` is a random sample of the variable beds per 1000 population from the population of all states and District of Columbia in 1980. Use the jackknife procedure on your sample to obtain a bias-corrected estimate of the population median. Submit your code and the numeric result.

   (d) Since this is one of those rare occasions on which you actually have measurements for the entire population, compute the actual population median. Show the R code and numeric output.

   (e) If your jackknife estimate was not equal to the true median, does this prove that the jackknife unbiased estimator is actually biased in this case? Submit a yes-or-no answer, and a brief explanation.