Chapter 1

An introduction to Stat 673

1.1 What is statistical computing?

One definition of statistics is that it is the science of collecting, analyzing and interpreting data in the presence of uncertainty.

Thus statistical computing is when we use computers in carrying out the following tasks.

1. Data visualization and analysis.
2. Monte Carlo experiments.
4. Mathematical statistics (e.g., computer algebra).

Stat 673 considers all of these topics.

1.2 An overview

- Statistics!
• General computing topics.
  
e.g., Windows, Mac, UNIX.

• Statistical software packages:
  
  – R.
  
  – SAS.

• Other software packages and topics as time permits.

1.3 Learning a new statistical package

There is too much to learn about each software package! In this course it will also be important to learn the skills involved to teach yourself:

• Learn how to use the build-in help in each package.

• Learn how to find online help on the web.

• Be able to ask the right questions (to me and your fellow students).

Q: Why do we look at multiple statistical packages? (rather than just one)

A:

Typically for each package you should be able to do the following essential tasks.

• Start and quit the program.

• Load, import, manipulate and save data.

• Produce custom graphics and figures.

• Obtain numerical summaries.
• Be able to handle probability distributions.
• Generate pseudo-random numbers.
• Carry out elementary statistical analyzes, e.g., confidence intervals and tests.
• Perform a simple linear regression or analysis of variance with full diagnostics.
• Design simple experiments.

1.4 A brief history of statistical computing

Taken from the “ASA Statistical Computing Section: a history” (?).

• 1950s and 1960s: "The involvement of statisticians in research, development, and use of computing facilities became increasingly central to much activity in the profession."

• mid 1960s: "Much statistical software existed, of varied styles and aimed at varying applications. Early software may seem crude from our present perspective, just as the computers it ran on seem amazingly primitive now. But these early efforts had already changed statisticians’ views of their activities profoundly and permanently."

• "In the early 1970s statistical computing was still working toward a clear identity and recognized respectability, especially among the more mathematically inclined. Theorems, or even crisp results, are rare, and some of the most important developments, such as the interactive exploration of data, cannot be adequately captured in written form."

• "Certain basic areas in computing emerged early: simulation, algorithms, computerized data analysis, and software. These remain the major areas in statistical computing today."

• "The 1970s was the era of the central mainframe computer, and most statistical packages were large batch programs (remember punched
cards). User interfaces were unfriendly, the methods were often obscure, and the answers were sometimes even wrong.”

• "In the 1980s, work in algorithms was much more diverse and included traditional research in linear methods as well as the new computer-intensive methods for data analysis."

• "We also explored ways to use computers effectively to analyze data. The area discussed most, with a total of seven sessions, was data management, including how to manage large, complex data sets, how to assure the quality of research data, and how to store data efficiently. Some of the most popular sessions were on the use of computers for graphics, from simple line-printer plots to interactive, dynamic graphics systems. In addition, there were talks on using computers in specific areas of analysis, including biased estimation, missing-data analysis, survival analysis, regression, ANOVA and the analysis of frequency data.”

1.5 Issues in statistical computing

Computer based statistics should aim to have

- stable and accurate calculations;
- reproducible results;
- understandable output.

Good statistical computer packages should

- have all the above features;
- be easy to use and/or learn (or at least have good help resources);
- be able to be extended or customized.
1.6 Statistical formulas

Suppose we have data \{X_i : i = 1, \ldots, n\}. Consider the equation for the sample variance:

\[ s^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n - 1}, \]

where \( \bar{X} \) is the sample mean. An equally valid equation for the variance is

\[ s^2 = \frac{\sum_{i=1}^{n} X_i^2 - n \bar{X}^2}{n - 1}, \]

or

\[ s^2 = \sum_{i=1}^{n} \left( \frac{X_i}{\sqrt{n - 1}} \right)^2 - \left( \frac{n}{n - 1} \right) \bar{X}^2. \]

Which equation do we use?

- Which expression when programmed into a computer is more accurate?
- Which equation is quickest to evaluate?

Exercise: How many definitions are there for the sample median or more generally a sample quantile?

1.7 Computer representations of numbers

A computer represents all data as binary digits or bits (the phrase is due to the famous statistician John W. Tukey) - ones and zeroes.

\[ \ldots 1010100100100 \ldots \]

All numbers in a computer are stored as a finite number of bits.

- Typically integers (guaranteed to be of maximum magnitude) are stored exactly.
• **Real** numbers are always stored to a given accuracy or **floating point**.

The point: **be careful** in certain calculations (especially those involving numbers of vastly different magnitudes) and **testing for equality**.

### 1.8 Data management and storage issues

Be organized. Here are some helpful hints.

- Make folders for your data.
- Keep a description of your dataset (including how the data is represented).
- Make backups.
- Sometimes you need to keep your data private – is **security** an issue?

### 1.9 General computing introduction

#### 1.9.1 Stat systems

We have the following systems in the Statistics department (only available to stat grads).

- A range of linux servers (e.g., `mordor.stat.ohio-state.edu`)
- PCs and Macs in the the student lab CH 341.
- Some PCs in stat student offices.
1.9.2 OIT Systems

For more information on the OSU computing centers go to http://scc.osu.edu.

- For this class we shall use the PCs and Macs in the EA cluster (EA 265 – EA 295).

All PCs run Windows XP and have SSH (secure shell) and X-terminal emulation software to access the linux machines. The Macs run OS X and have Terminal and X11 access to linux machines.

1.9.3 Getting help

For Stat 673 support email me at pfc@stat.osu.edu. For machine and connectivity problems check out:

- Stat dept computer support at http://www.stat.osu.edu/current/support.html.

You may also be able to solve your problem by searching for help on the internet; for example try a search engine such as http://www.google.com.

When emailing for help:

- Restrict yourself to one problem per email.
- Summarise the problem in the subject line of the email.
- Try to be as accurate about the problem as possible – give details of what steps you followed before you encountered the problem.
References