Statistics 215a – 8/30/04 – D.R. Brillinger

Statistics.

Part of the methodology of science

Concerns: data collection, data analysis, data reduction, data modeling and inference from data

Primitive concept: a datum

Exploratory data analysis (EDA).

Procedures for analyzing data
Techniques for interpreting results of such procedures
Ways of planning data gathering to make analysis easier/more precise/more accurate
Results of mathematical statistics applicable to analyzing data

Data mining.

EDA with little or no human interaction using computational feasible techniques
Process of seeking interesting/valuable information within large databases

Stem-and-leaf displays.

stem(x, scale, width)

Test scores: 44, 63, 60, 66, 68, 76, 72, 74, 70, 70, 76, 78, 84, 84, 86, 86, 88, 90, 92, 96

"The decimal point is 1 digit(s) to the right of the |

```
4 | 4
5 |
6 | 3 0 6 8
7 | 6 2 4 0 0 6 8
8 | 4 4 6 6 8
9 | 0 2 6"
```

E.g. 44 --> 4 | 4

stem - leading digit(s}
leaf - following digit (no rounding)
units - decimal place
scale / interval width / number of rows per stem

Highlights:

- symmetry / asymmetry / skewness / tails
- range
- outliers
- concentrations / clumps
- gaps /coarseness / granularity / patterns
- summaries (center, spread, mode(s), ...)

Advantages:

- both numerical and graphical information
- sorts
- can prepare by hand
- surprises

Difficulties:

- line overflow
- outliers (Splus better)
- programming

Programming decisions:

no. lines? 10*log10(n), 2*sqrt(n), 1 + 2*log2(n)

lines per stem? e.g. 2, 5, 10

+- 0 values