STAT 100 Chance and Data Analysis Sept 6, 2002

Today: Review of examples from last lecture drawing out the main points

1. Influence of Unexplained Variation (UV) on Interpretation of Data
2. UV can make temporary effects seem like permanent ones (illusions)
3. Graphing of Data is an essential first step in data analysis
4. Need for summary measures when UV present

1. Influence of Unexplained Variation on Interpretation of Data

Unexplained Variation (UV)

Weigh scale – why is UV a problem? – studying effects of changes

Data often used to study cause-effect relationships …

How can one prove that A causes B in the presence of UV?

Skip lunch example (or low carb, or low calorie, ….)

Need for more than one person

Need for control treatment

Need to have investigator assign treatment to prove causality

Smoking and Lung Cancer – link took a long time to prove

UV really complicates the interpretation of data.

2. UV can make temporary effects seem like permanent ones (illusions)

Row	Team	Points (Win=2, Loss=0, no ties)
1	2	16
2	3	10
3	5	8
4	1	6
5	4	0
1	2	14
2	1	8
3	4	8
4	3	6
5	5	4
1	3	12
2	5	10
3	1	8
4	4	8
5	2	2

All these five teams are of EQUAL Quality in terms of chance of winning!

Look again at sports league:

Leeds had 23 points, Derby County had 7. Does this mean Leeds is the better team?

We can study this using **Simulation.**

We can also do some mental experiments:

What is the effect on the spread of points in league standings of allowing ties?

What is the effect … of a home team advantage?

Our speculations about these questions can also be checked by simulation.
More next time.

3. Graphing of Data is an essential first step in data analysis

Recall Gas consumption data. The graph suggested the seasonality, and this led to interesting questions about the cause of the seasonality.

Note: this is an example of a scatter plot. Two "variables" are plotted for a data set in which the rows of the data are linked:

Var 1(Date)	Var 2(Miles per Km)
May 5 1999	6.5
May 12 1999	6.7
May 15 1999	5.9

Because one of the variables is "Time", this kind of data is called a time series.

Why is a time series difference from other kinds of data?

Consider a data set with height and weight of each member of a football team…

Would it be possible to detect seasonality without the graph?

Would fitting a straight line to the data help?

How about calulating the mean and standard deviation?

What IS the mean and standard deviation?

4. Need for summary measures when UV present

Mean, or Arithmetic Average, or Average

A “Middle” Value
Add them up and divide by the number of numbers
Mean of {1,5,6} = (1+5+6)/3

Standard Deviation (SD, or s)
A measure of spread of the numbers
Just square the deviations from the mean, add them up, divide by the number of numbers, and then take the square root!
SD of {1,5,6}:
Mean is 3
$(1-3)^2 = 4$
$(5-3)^2 = 4$
$(6-3)^2 = 9$
4+4+9 = 17
17/3 = 5.7 approx
square root of 5.7 = 2.4 approx
So SD of {1,5,6} is 2.4 – it is a “typical deviation” from the mean.

Would mean and SD have a role in describing league outcomes?

Assignment 1: Will be assigned on Monday Sept 9, and will be due on Monday Sept 16 at 4:30 pm at the Statistics Workshop (K 9516).

Homework from today (not to hand in yet): Read text pp 93-103.
It will defintiely include something about your small scale league simulation. Also, it will include some questions based on the text pp 93-103: Measuring the Effects of Social Innovations.