Data Analysis

Thomas Lumley, tlumley@u
Office: HSB F656
Office Hours: Monday 9-10, or by email, or probably other times if you can find me

This course is designed to accompany the advanced applied methodology sequence 570–1–2 in Statistics and Biostatistics. In many ways the applied qualifiying exam is more a test of your data analysis skills than of your knowledge of applied methodology. The data analysis sequence will provide practise in data analysis and report writing, together with some lectures on relevant techniques and ideas.

The basic distinction between the examples presented in 570–1–2 and those in this course is the intent of the analysis. In 570–1–2 the focus is on the methods, and a good example is one that illustrates the power of the methods. In data analysis the focus is on the scientific question, and any method that answers the question is appropriate.

The course will involve roughly weekly written assignments and class discussion. As the assignments are discussed as an important part of the class time the deadlines are not negotiable. There will also be a term project, due at the end of the quarter.

Grading is CR/NC for this couse. A CR requires handing in a reasonable effort at the term project and all but one weekly homework.

Week 1

Due Thursday morning 9am (email of PDF, PostScript or plain text is acceptable; not Word documents)

1. Send me an email so I can make a class list

2. The Anscombe quartet of data sets are

"x1"	"x2"	"x3"	"x4"	"y1"	"y2"	"y3"	"y4"
10	10	10	8	8.04	9.14	7.46	6.58
8	8	8	8	6.95	8.14	6.77	5.76
13	13	13	8	7.58	8.74	12.74	7.71
9	9	9	8	8.81	8.77	7.11	8.84
11	11	11	8	8.33	9.26	7.81	8.47
14	14	14	8	9.96	8.1	8.84	7.04
6	6	6	8	7.24	6.13	6.08	5.25
4	4	4	19	4.26	3.1	5.39	12.5
12	12	12	8	10.84	9.13	8.15	5.56
7	7	7	8	4.82	7.26	6.42	7.91
5	5	5	8	5.68	4.74	5.73	6.89

   These are clearly artificial, which is atypical for this class. Give a suitable confidence interval for the mean of Y when $X = 10$ for each pair (x1, y1), (x2, y2), (x3, y3), (x4, y4) and explain why.

3. The 2001 applied exam dataset looked at relationships between exercise and mortality in the elderly.

   (a) It is known that exercise reduces blood pressure and that reducing blood pressure is beneficial in this population. Give reasons why adjusting the exercise–mortality relationship for blood pressure would be appropriate and reasons why it would be inappropriate.

   (b) An important concern is that higher levels of exercise may be a result of good health rather than a cause. Can you think of any way to get evidence about the direction of causation in these data?