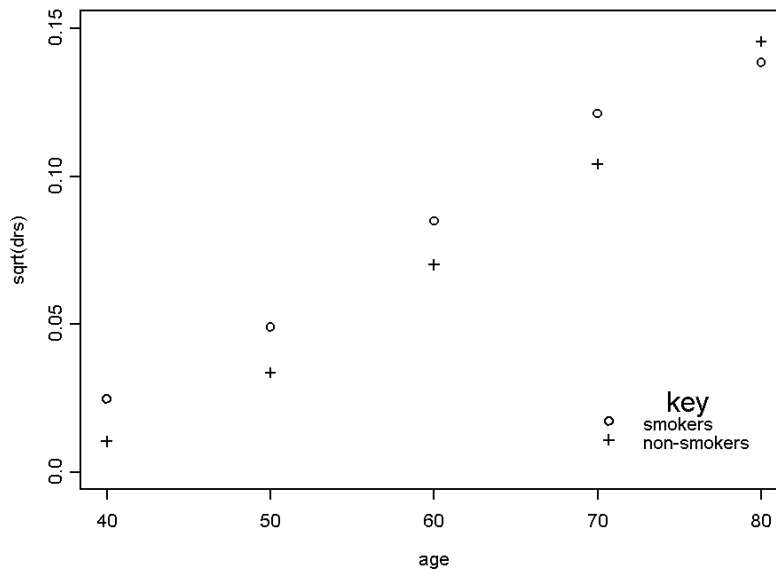


Statlab Exam 1999: Outline Solutions

1. Summaries (many different plots possible!):

Age	35–44	45–54	55–64	65–74	75–84
Death rate (non-smokers)	0.00010	0.00112	0.00490	0.01083	0.02120
Death rate (smokers)	0.00061	0.00240	0.00720	0.01468	0.01918



The death rate for 75–84 year olds is higher for non-smokers than for smokers, but it would clearly be inappropriate to recommend a 75 year old to continue (or take up) smoking:

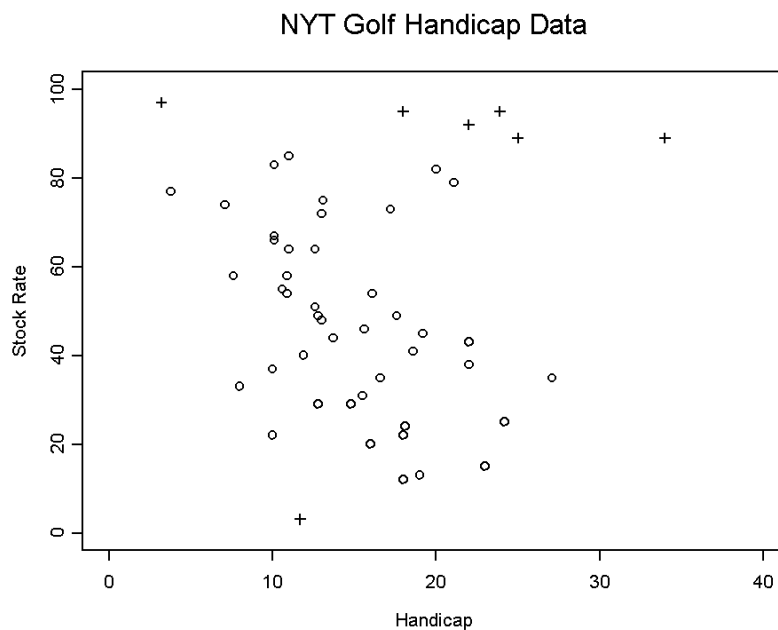
- (a) No evidence of causation, e.g. if smokers prone to CHD have tended to die earlier than non-smokers prone to CHD, then the remaining 75 year-olds will tend to show a higher CHD death rate amongst the non-smokers than amongst the smokers.
- (b) Deaths from other possibly smoking-related causes (e.g. emphysema) aren't considered.
- (c) Advice on smoking should take into account all aspects of smoking-related diseases, including quality of life.
- (d) Many effects of smoking are cumulative and long-term; changing the smoking habits of a lifetime may make little difference at age 75.

Constructive criticism

- Data for women or for different professions might show different patterns.
- It isn't clear how ex-smokers are treated—it will distort the picture if they are included among the non-smokers (particularly as some may have quit because of smoking-related illness). However, the high numbers of smokers suggest that ex-smokers have been included amongst the smokers.

- If one wanted to use the data to guide advice on smoking, then a minimum requirement would be to subdivide into groups like ‘never smoked’, ‘quit at least 5 years ago’, ‘current or recent smoker’. However, the more complicated the information collected, the harder it is to obtain reliable data.
- CHD is worsened by stress, whereas for some people stress is alleviated by smoking. There is no way to tell what data would have been obtained if such people had not smoked.
- Useful explanatory variables include: amount smoked, some measure of stress (e.g. recent number of hours worked per week), diet, exercise etc.

2. Simple plot (o = included point, + = ‘outlier’):



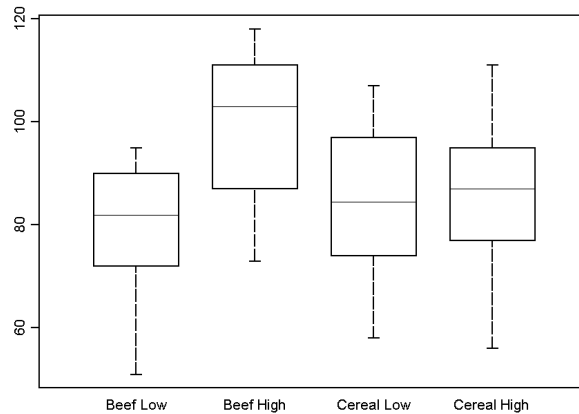
Constructive criticism

- The picture is very different (and complex) if the so-called ‘outliers’ are kept in. The six best performing companies were all omitted; four of their CEOs had handicaps worse than that of any of the CEOs for the 50% top included companies.
- The more possible associations are investigated, the more will appear by chance. The investigator is even quoted as saying that he investigated factors other than golfing ability.
- Nothing is known about companies whose CEOs do not play golf, or who declined to report their golf handicaps. Quite possibly CEOs of moderately successful companies may be particularly reluctant to report their own inability at golf.
- Even if there were a clear association between golfing ability and stock market performance, there is no evidence of causation in any respect!

- Other explanatory variables, such as age of CEO (which is presumably related to experience both in management and in golf) would be useful.
- More info is needed on how the ‘Stock Rating’ was calculated. For example, does it represent long-term performance or not?
- Some CEOs might play golf together, CEOs move between companies, companies merge, etc. All these facts make assumption like ‘independence’ implausible.

3. Summaries:

	Min	Lower Quartile	Median	Upper Quartile	Max
BL	51	72	82	90	95
BH	73	87	103	111	118
CL	58	74	84.5	97	107
CH	56	77	87	95	111



Rats fed the beef high protein diet clearly tend to gain most weight, those fed the beef low protein diet possibly least. There’s little difference between the effects of the two cereal based diets. Scatter is similar for all four groups (IQRs 18, 24, 23, 18).

Two sample t-test: $t=2.32$ on 38 d.f., reject null hypothesis of ‘no treatment difference’ at 5% level.

Normality & homoscedasticity assumptions plausible (variances 214 & 267).

Constructive criticism

- Inappropriate to combine the beef+cereal groups, since the effect of a high protein diet clearly differs between beef & cereal. The estimated ‘overall effect of a high protein diet’ is largely an artefact of the number and type of protein sources used, and the numbers of rats in the various groups.

- Rats in a given group should not all be kept together, lest (for example) they all suffer the same infections, disturbances etc. and distort the apparent effect of diet.
 - The rats may eat different amounts of food, particularly if several rats are kept together and some are dominant.
 - Possibly useful explanatory variables include sex, initial weight + age of rat.
 - Difficult to know how to compare the groups unless we are assured that (e.g.) the protein contents were similar between the two 'low protein' diets, and also between the two 'high protein' diets.
 - Weight gain is obviously the most easily objectively measurable summary of effect of diet, but is not necessarily the best indicator of 'good' or 'bad' diet.
 - Units aren't given (presumably grammes?)
-