

## Statlab Exam 1998: Outline Solutions

1. (a) Stem & Leaf Plot of Inter-Earthquake Intervals

```
0 |
0 |13344444458889
1 |02344569          Stem Unit: 100 days
2 |0125689          Leaf Unit: 10 days
3 |0233678
4 |03456
5 |6678
6 |0479
7 |123468
8 |349
9 |4
10 |
11 |
12 |
13 |45
14 |
15 |
16 |2
17 |
18 |
19 |0
20 |
```

The distribution is highly skewed (positively):

Median–Lower Quartile  $\simeq$  200 days; Upper Quartile–Median  $\simeq$  335 days.

There is perhaps a sharp drop in density around 1000 days but with several outliers above that value.

- (b) Sample median = 331.5 days; sample mean = 437 days; Est. mode = 0 days.
- (c) Clearly a skewed distribution taking only non-negative values. Simplest to use a continuous distribution, in particular exponential (since shape looks similar to S&L plot). Possibly a mixture of exponential + (mainly) another distribution with CDF nearly 1 at 1000 days.
- (d) Estimate unknown parameters of  $F$  from data (e.g. mean 437 for exponential distn.). Then compare proportion of observed values in given range with theoretical values from CDF of fitted  $F$ . Could apply a formal  $\chi^2$  test, but need to adjust (somehow!) for fitting parameter(s) and to note that expected count above 5 years would be very small.

### Constructive Criticism

- Have lost time ordering of data in above analysis; need to check IID assumptions [*I haven't covered time series etc., but students might suggest dividing into say 3 roughly equal time periods, or plotting  $Interval_i$  against  $Interval_{i+1}$* ].
- Would like more data, e.g. to check agreement of  $\Pr(\text{Interval} > 5 \text{ years})$ . This is obviously impossible because of time. However, one could perhaps include less severe earthquakes, and stratify for severity.
- Other useful explanatory variables: geographical centre of earthquake, magnitude on Richter scale (longer intervals may lead to more severe earthquakes).

- ...etc.

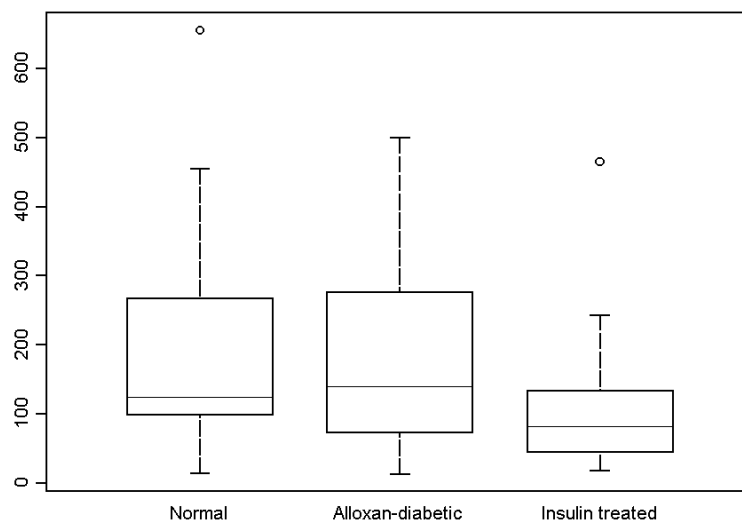
2. (a) Five number summaries are:

Normal: (14, 98, 124.5, 267.5, 655)

A-diabetic: (13, 73, 139.5, 276, 499)

I. treated: (18, 46, 82, 133, 465)

(some students may use slightly different formulae for quartiles). The following figure shows a boxplot of the data:



Distribution in all three groups is positively skewed. Plots are largely similar for Normal & Alloxan-diabetic, but both median & IQR for Insulin treated group are about half values in other groups. Suggests analysing the data on a transformed (square root?) scale.

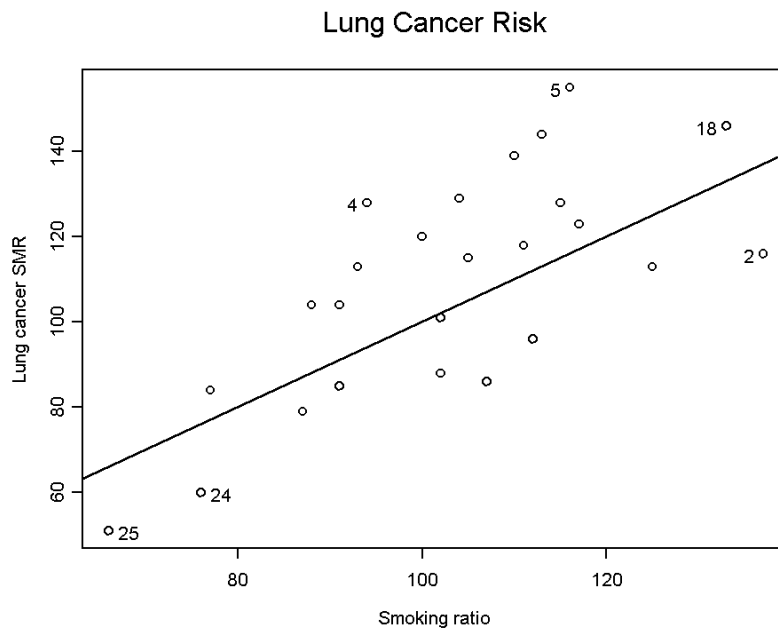
(b) Formal calculation gives: means = (181.8, 112.9), variances = (20981, 11191),  $t = 1.68$  on 35 d.f. Accept  $H_0$  at 5% level (alternatively, p-value = 0.106).

(c) **Constructive Criticism**

- Boxplots strongly suggest a difference between alloxan-diabetic & insulin treated mice. Result of t-test appears misleading: get more data (see below) or transform, or (preferably) both.
- Underlying assumptions of t-test are violated. Variances are roughly proportional to the means for the 2 groups, which suggests trying a t-test after a square-root transformation  
*[I've suggested choosing transformations based on e.g. roughly equalising distance between 5%–50% and 50%–95% points, but haven't discussed mean  $\propto$  variance or s.d. etc.]*
- Need more information about experiment (were diabetic mice randomized to treatment groups; were they housed separately etc.)

- What is thought to regulate the amount of ‘bovine serum albumin produced’? How is it measured? How does the amount matter (why are distributions similar for normal and placebo-treated mice, but different for insulin-treated)? Are the results relevant to human diabetes?
- Possibly useful explanatory variables: age, sex, weight

3. (a) Clearly a positive association in general:



(b) Apart from group 2 (miners & quarrymen), points seem to lie around a line  $\hat{y} - 100 = c(x - 100)$  for  $c$  nearly 2 rather than  $c = 1$ .

Following occupational groups are indicated on diagram:

- 2 Miners and quarrymen (large  $x$ ;  $y \ll \hat{y}$ ),
- 4 Glass and ceramics makers ( $y \gg \hat{y}$ ),
- 5 Furnace, forge, foundry, rolling mill workers (possibly  $y \gg \hat{y}$ ),
- 18 Labourers not included elsewhere (large  $x$ ),
- 24 Administrators and managers (low  $x$ ),
- 25 Professional, technical workers, artists (low  $x$ ).

The two professional groups (24 & 25) respectively have the second lowest and lowest values for both  $x$  and  $y$ .

(c) **Constructive criticism**

- Lung cancer SMR for miners & quarrymen is unusually low given their smoking ratio. Need to look for possible explanations—e.g. do miners with damaged lungs tend to die of (or to be diagnosed as dying of) pneumoconiosis etc. instead? do they retire early? (do such cases appear in the data?)

- Explanatory information for individuals, such as family history of cancer, are presumably impossible to collect for this large study. However, explanatory variables related to the occupational groups may be important, such as the degree of dust + fume exposure, and the salary level (e.g. wealthier people might have more health checks & better care).
  - How was the ‘smoking ratio’ obtained? If by a sample of workers in that group, then how was the sample chosen? Might people have lied about their smoking habits? How were cigar & pipe smokers treated?
  - Some of the groupings don’t seem appropriate for this study—e.g. ‘food, drink and tobacco workers’. Food workers would presumably be barred from smoking at work for hygiene reasons, whereas one might expect workers with tobacco to smoke more.
  - Would also like information on rate of death from other smoking-related diseases such as throat cancer, heart disease etc.
  - It might be important to distinguish between the smoking ratio being high because there are more smokers, or because the smokers in that group tend to smoke more.
-