MATH11400 Statistics 1 2010–11

Homepage http://www.stats.bris.ac.uk/%7Emapjg/Teach/Stats1/

Solution Sheet 2

1. (a).

$$\begin{aligned} \bar{y} &= \textstyle\sum_{i=1}^n y_i/n \\ &= \textstyle\sum_{i=1}^n (ax_i + b)/n \\ &= a\textstyle\sum_{i=1}^n x_i/n + \sum_{i=1}^n b/n \\ &= a\bar{x} + b \end{aligned}$$

$$\begin{aligned} s_y^2 &= \textstyle\sum_{i=1}^n (y_i - \bar{y})^2/(n-1) \\ &= \textstyle\sum_{i=1}^n ((ax_i + b) - (a\bar{x} + b))^2/(n-1) \\ &= a^2 \textstyle\sum_{i=1}^n (x_i - \bar{x})^2/(n-1) \\ &= a^2 s_x^2 \end{aligned}$$

(b). The linear transformation from x to $y = ax + b$ will either preserve the order of the observations when they are ranked in increasing size (if $a > 0$) or reverse it (if $a < 0$). Thus the middle ranked y observation will be the one corresponding to the middle ranked x observation; the top and bottom 10% of the y observations will be exactly those values corresponding to the top and bottom 10% of the x observations (though top and bottom may swap if $a < 0$), and so on.

Thus the median, the quartiles and the trimmed means of the $\{y_i\}$ will be exactly ($a\times$ the value of the corresponding quantity for the $\{x_i\}) + b$. Since the IQR is the difference between the two quartiles, the b will cancel as in the claculation for s_y^2 above, and the IQR for the $\{y_i\}$ will be just ($a\times$ the IQR for the $\{x_i\}$).

(c). Here $a = 1.8$ and $b = 32$, so the $\{y_i\}$ observations will have mean $= 1.8 \times 68.1 + 32 = 154.58$; median $= 1.8 \times 68.9 + 32 = 156.02$; variance $= 1.8^2 \times 3.2 = 10.368$; IQR $= 1.8 \times 7.7 = 13.86$.

2. The data are of the form $-11.1, -6.6, -5.0, -5.0, -5.0, -4.4, -4.4, \ldots$ etc., with multiple data values at slightly unusual decimal values that seem to be just over 0.5 apart. Thus it displays the kind of clustering or granularity mentioned in Section 1.2 of the notes.

   The clue is in the previous question, and the fact that the data represents temperatures. If you guess that the data x is in units of degrees Celsius, then the corresponding y values in degrees Fahrenheit satisfy $y = 1.8x+32$, giving, after slight rounding, values $12, 20, 23, 23, 23, 24, 24, \ldots$. Thus the data were probably recorded in Fahrenheit, rounded to the nearest integer value, and then later transformed to degrees Celsius, acquiring in the process a spurious air of accuracy to an extra decimal point.

   Of course, the quite irregular shape of the whole distribution is worth some interpretation. But these are data from just 60 arbitrary cities, there is no reason to expect conformity to any standard distribution. From the extremes, one imagines that cities in both Alaska and in the South (or perhaps Hawaii) were included.

3. The boxplot resulting from the commands in the question should look like the one below. The treatment types appear to split into two disjoint groups, {A,B,F} and {C,D,E}. Within each group, the treatments are quite comparable – for example, they appear roughly similar in terms of location and spread. However, there are substantial differences between the groups, both in terms of location and spread.

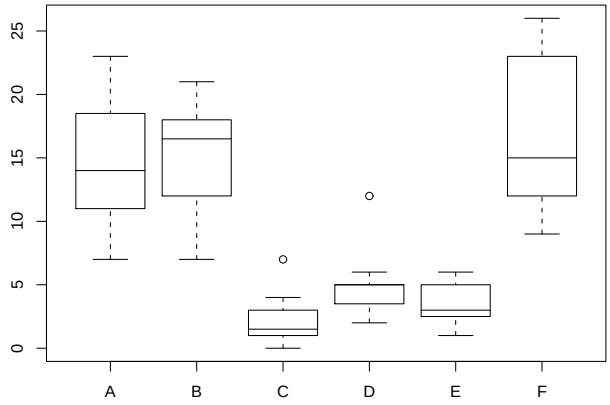

Figure 1: Box plots of the insect spray data

There are a variety of ways to compute the treatment means and variances in R. The way shown below works by reforming the data into a matrix, with one column of data corresponding to each treatment and then computing the column means and column variances using the `apply` command together with the `mean` and `var` commands. The standard deviations are then found as the square roots of the variances. As expected, the numerical summaries show exactly the same type of similarities within the groups {A,B,F}and {C,D,E}, and consistent differences between the two groups, as the boxplots.

```
> attach(InsectSprays)
> spraymat<-matrix(InsectSprays[,1], ncol=6)
> apply(spraymat, 2, mean)
[1] 14.500000 15.333333  2.083333  4.916667  3.500000 16.666667
> apply(spraymat, 2, var)
[1] 22.272727 18.242424  3.901515  6.265152  3.000000 38.606061
> sqrt(apply(spraymat, 2, var))
[1] 4.719399 4.271115 1.975225 2.503028 1.732051 6.213378
```

4. Here is a transcript of my runs in **R**:

```
> ir2<-iridium[-c(1,2,3,4,8)]
> stem(ir2)

  The decimal point is at the |

  159 | 123
  159 | 55556678
  160 | 00112234
  160 | 68
  161 | 1

> stem(rnorm(22))

  The decimal point is at the |
```

```
  -2 | 60
  -1 | 83
  -0 | 7554432211
   0 | 38
   1 | 223567

> stem(rnorm(22))

  The decimal point is at the |

  -2 | 3
  -1 | 6332
  -0 | 8666310
   0 | 13337
   1 | 00129

> stem(storm.claims)

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

  0 | 00000112233334
  0 | 5588
  1 | 2

> stem(rexp(19))

  The decimal point is at the |

  0 | 1223457899
  1 | 011248
  2 | 3
  3 | 8
  4 | 0

> stem(rexp(19))

  The decimal point is at the |

  0 | 011223
  0 | 5567
  1 | 22
  1 | 6689
  2 | 02
  2 | 5
```

Of course, when using random numbers, you get different 'data' each time, so your output may look different in detail. Probably the most important conclusions to be drawn from the visual comparisons are:

(a) With small sample sizes ($n = 22$ or 19), there is considerable variation between samples

in the appearance of stem and leaf plots and histograms: some look very similar in broad terms to the true density functions from which the data are generated, other less so.

(b) The real data sets give plots that qualitatively lie within the range of variation in the simulated data sets (allowing for the differences in units).

(c) But on the basis of small samples, you cannot expect to conclude that a particular statistical model generated the data (in the sense that an infintely large sample would exactly conform to the true density function).