MATH11400 Statistics 1 2010–11

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

Problem Sheet 2

Remember: when online, you can access the Statistics 1 data sets from an **R** console by typing
`load(url("http://www.stats.bris.ac.uk/%7Emapjg/Teach/Stats1/stats1.RData"))`

1.* Let $\{x_1, \ldots, x_n\}$ be a data set of real numbers and let $y_i = ax_i + b$, for $i = 1, \ldots, n$.

(a) Let $\overline{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$ and $s_x^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$. Show that $\overline{y} = a\overline{x} + b$ and $s_y^2 = a^2 s_x^2$.

(b) Find expressions for the median, interquartile range and trimmed mean of $\{y_i\}$ in terms of those of $\{x_i\}$.

(c) Let x denote temperature in degrees centigrade and let y denote temperature in degrees Fahrenheit, so $y = 1.8x + 32$. Assume the $\{x_i\}$ data set has mean 68.1, median 68.9, variance 3.2 and IQR 7.7. Calculate the corresponding quantities for the $\{y_i\}$ data.

2. Having loaded the Statistics 1 data set into **R**, use the command `stem(us.temp,scale=4)` to produce a stem-and-leaf plot of the dataset `us.temp`. The data gives the mean January temperatures for 60 U.S. metropolitan areas. Comment on any unusual pattern in the data and try to find a plausible explanation.

3. Boxplots are most useful for comparing more than one sample. The built-in data set `InsectSprays` in **R** gives the number of insects found on plants subjected to 6 different treatments labelled A-F. Type the following in **R**:

```
data(InsectSprays)
help(InsectSprays)
InsectSprays
boxplot(count ~ spray, data = InsectSprays)
```

The `help` command gives some background information about the data, and the command `InsectSprays` on its own prints out the data. For this data set, the `boxplot` command produces a separate boxplot (on common axes) for each of the treatments. Use this plot to compare the different treatments. Calculate the mean and variance for each of the treatment types and see if you come to the same conclusions. (It is good practice working out how to do this in **R**).

4.* Make a new version of the `iridium` data set, excluding the apparent outliers, by typing `ir2<-iridium[-c(1,2,3,4,8)]`. Create a histogram and stem and leaf plot of this new data set. Now make similar plots for an artifical sample made by generating the same number (22) of observations from a normal distribution (e.g. `data<-rnorm(22)`). Visually compare the plots for the real data and the artificial data. Repeat for several more artificial data sets created, independently, in the same way.

You can repeat a similar exercise with the `storm.claims` data set, comparing it with an artificial sample of 19 observations drawn from an exponential distribution, created using `data<-rexp(19)`.

Do not hand in all the results from your experiments, but instead just select the stem and leaf plots of the 2 real data sets and of up to 4 artificial deta sets (in total), together with brief comments on your conclusions from these visual comparisons.