MATH11400 Statistics 1 2010–11

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

Problem Sheet 1

Remember: when online, you can access the Statistics 1 data sets from an **R** console by typing

```
load(url("http://www.maths.bris.ac.uk/%7Emapjg/Teach/Stats1/stats1.RData"))
```

1.* In an experiment to investigate the heat of sublimation of iridium, the following 27 measurements were made, listed across the rows in the order they were taken. The data is contained in the Statistics 1 data set `iridium`.

136.6	145.2	151.5	162.7	159.1	159.8	160.8	173.9	160.1
160.4	161.1	160.6	160.2	159.5	160.3	159.2	159.3	159.6
160.0	160.2	160.1	160.0	159.7	159.5	159.5	159.6	159.5

(a) Use the **R** commands `stem`, `hist`, `boxplot`, and `plot` to make a stem-and-leaf plot, a histogram, a boxplot and a plot of the observations in the order they were taken. Print off your plots and comment on the overall pattern of the data and any striking features.

(b) Use the **R** commands `median` and `mean` to find the median and the mean.

When there are outliers, it may help to 'trim' the extreme values, so the mean of the remaining 'trimmed' observations gives a better indication of the centre of the distribution. For a Δ% trimmed mean from n ordered observations, let k denotes the integer part of $n\Delta/100$; then discard the k smallest values and the k largest values, and compute the mean of the remaining $n - 2k$ values. Use `?mean` in **R** to see how to compute a trimmed mean in **R**. Compute the 10% and 20% trimmed means for the iridium data set.

Compare how well the these four measures represent the centre of this data set.

(c) Use the **R** commands `var` and `sd` to find the sample variance and standard deviation, use the **R** commands `fivenum` and `summary` to find the hinges and the sample quartiles, and use the **R** command `IQR` to find the interquartile range (but see comments on 'Hinges and Quartiles' overleaf). Again, compare how these values represent the spread of the data. [Care! In an 'ideal' data set, would the variance, standard deviation and inter-quartile range be the same?]

(d) What conclusions do you draw from your plots and numerical summaries? What effect do the outliers have on the numerical and graphical summaries? What would the corresponding results look like if the outliers were removed?

2. Construct an **R** function for calculating an empirical distribution function by typing in the following instructions (note that the **R** prompt will switch from `>` to `+` while it is waiting for the command to be completed):

```
plot.edf <- function(x){
n <- length(x)
plot(sort(x),(1:n)/(n+1),type='s',xlab='data',ylab='empirical cdf')
}
```

Having loaded the Statistics 1 data sets, produce an empirical distribution function (edf) plot of the iridium data by typing the command `plot.edf(iridium)` and comment on how the shape of the edf relates to the data.

3. 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.

4. 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.

5. 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**).

Hinges and Quartiles

The lower hinge is 'the median of the set of values $\leq$ the sample median' and the upper hinge is 'the median of the set of values $\geq$ the sample median'. Hinges were introduced by Tukey as a simple alternative to quartiles, since sources disagreed on exactly how quartiles should be defined.

Loosely speaking, 'quartiles' are values that divide a dataset into four equal parts – a quarter of the data values are greater than the upper quartile, a quarter are between the upper quartile and the median, a quarter are between the median and the lower quartile, and a quarter are less than the lower quartile.

Given a dataset with n data values $x_1, x_2, \ldots, x_{n-1}, x_n$, denote the ordered values (the order statistics) by $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n-1)} \leq x_{(n)}$. This suggests Q_1 should be roughly the $n/4$th observation (ordered in increasing size). If $n/4$ is not an integer, then Q_1 should lie between $x_{(i)}$ and $x_{(i+1)}$ where $i = [n/4]$ (the integer part of $n/4$). The methods actually used to compute, say, Q_1 are more complicated than this, but most have the following common basis: Set $r = (n + 1 - 2a)/4 + a$, set $i = [r]$ (the integer part of r) and set $\gamma = r - i$. Then the required value is $Q_1 = (1 - \gamma)x_{(i)} + \gamma x_{(i+1)}$, i.e. the value that lies γ of the way between $x_{(i)}$ and $x_{(i+1)}$. Similarly for Q_3, set $s = 3(n + 1 - 2a)/4 + a$, set $j = [s]$ and set $\gamma = s - j$. Then the required value is $Q_3 = (1 - \gamma)x_{(j)} + \gamma x_{(j+1)}$.

Where methods differ is in the value of a – some have used $a = 0$ (Minitab), some $a = 1$ (Excel), some $a = 1/2$ (S-plus). Rice suggests using $a = 0$ or $a = 1/2$. **R** uses $a = 1/2$ in some places and $a = 3/8$ in other places.

The differences between using different values of a, or indeed the differences between the hinges and the quartiles, have no real practical importance in terms of interpretation, and they are negligible numerically in larger data sets.