Data Analysis & Graphics Using R – Solutions to Exercises (March 21, 2004)

Preliminaries

```
> library(DAAG)
```

Exercise 1
This question has been reworded
For the `possum` data set, use `hist(possum$age)` to draw a histogram of possum ages. Where are the breaks, i.e., the class boundaries between successive rectangles. Repeat the exercise, this time specifying

```
hist(possum$age, breaks=c(0,1.5,3,4.5,6,7.5,9))
```

Use

```
table(cut(possum$age, breaks=c(0,1.5,3,4.5,6,7.5,9))
```

to obtain the table of counts. In which interval are possums with `age=3` included; in `(1.5,3]` or in `(3,4.5]`. List the values of `age` that are included in each successive interval. Explain why setting `breaks=c(0,1.5,3,4.5,6,7.5,9)` leads to a histogram that is misleading.

For convenience, we place the two histograms side by side.

```
> par(mfrow = c(1, 2))
> data(possum)
> hist(possum$age, main = "Breaks at 0, 1, ..., 9")
> hist(possum$age, breaks = c(0, 1.5, 3, 4.5, 6, 7.5, 9), main = "Breaks at 0, 1.5, ...")
> par(mfrow = c(1, 1))
```

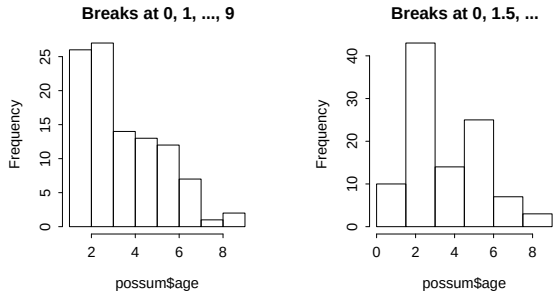

Figure 1: Histograms, with different choices of breaks. The choice for the graph on the right has bizarre consequences.

The second graph demonstrates an unsatisfactory and misleading choice of breaks. This is most easily seen by tabulating the frequencies for the two graphs, thus:

```
> table(cut(possum$age, breaks = 0:9))

(0,1] (1,2] (2,3] (3,4] (4,5] (5,6] (6,7] (7,8] (8,9]
   10    16    27    14    13    12     7     1     2

> table(cut(possum$age, breaks = c(0, 1.5, 3, 4.5, 6, 7.5, 9)))
```

```
(0,1.5] (1.5,3] (3,4.5] (4.5,6] (6,7.5] (7.5,9]
     10      43      14      25       7       3
```

For the second graph, the ages fall into groups 1, (2,3), 4, (5,6), 7, and (8,9). The category (0, 1.5] catches just 1 year olds, the category (1.5, 3] catches ages 2 and 3, and so on. Thus the histogram that has breaks at intervals of 1.5 years is highly misleading.

Exercise 2
Now try `plot(density(possum$age, na.rm=T))`. Which is the most useful representation of the distribution of data – one or other histogram, or the density plot? What are the benefits and disadvantages in each case?

The density plot treats all possible choices of breaks equally. It does however require the choice of a bandwidth that determines how smooth the resulting density will be. By default, this is chosen automatically. On the whole, a density plot is less likely to be seriously misleading.

A problem with the simple form of density plot is that it has a non-zero density for ages less than one. This can be fixed by changing the code to, e.g.

```
> plot(density(possum$age, na.rm = T, from = 0.5), main = "")
```

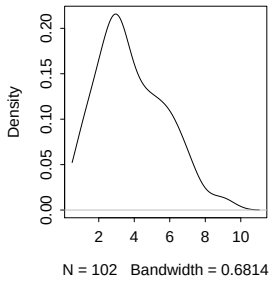

Figure 2: Density plot of possum ages

Exercise 3
Examine the help for the function `mean()`, and use it to learn about the trimmed mean. For the total lengths of female possums, calculate the mean, the median, and the 10% trimmed mean. How does the 10% trimmed mean differ from the mean for these data? Under what circumstances will the trimmed mean differ substantially from the mean?

```
> fossum <- possum[possum$sex == "f", ]
> mean(fossum$totlngth)

[1] 87.90698

> median(fossum$totlngth)

[1] 88.5
```

```
> mean(fossum$totlngth, trim = 0.1)

[1] 88.04286
```

To get an idea of the shape of the distribution, type in:

```
> plot(density(fossum$totlngth), main = "")
```

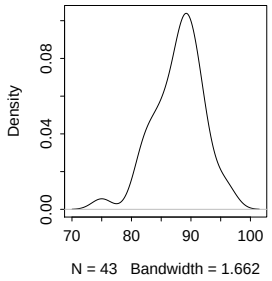

Figure 3: Density plot of female possum lengths.

The distribution is negatively skewed, i.e., it has a tail to the left. As a result, the mean is substantially less than the mean. Removal of the smallest and largest 10% of values leads to a distribution that is more nearly symmetric. The mean is then similar to the median. (Note that trimming the same amount off both tails of the distribution does not affect the median.)

The trimmed mean will differ substantially from the mean when the distribution is positively or negatively skewed.

Exercise 4
Assuming that the variability in egg length for the cuckoo eggs data is the same for all host birds, obtain an estimate of the pooled standard deviation as a way of summarizing this variability. [Hint: Remember to divide the appropriate sums of squares by the number of degrees of freedom remaining after estimating the six different means.]

```
> data(cuckoos)
> sapply(cuckoos, is.factor)

 length breadth species      id
  FALSE   FALSE    TRUE   FALSE

> specnam <- levels(cuckoos$species)
> ss <- 0
> ndf <- 0
> for (nam in specnam) {
+     lgth <- cuckoos$length[cuckoos$species == nam]
+     ss <- ss + sum((lgth - mean(lgth))^2)
+     ndf <- ndf + length(lgth) - 1
+ }
> sqrt(ss/ndf)

[1] 0.9051987
```

A more cryptic solution is:

```
> diffs <- unlist(sapply(split(cuckoos$length, cuckoos$species),
+     function(x) x - mean(x)))
> df <- unlist(sapply(split(cuckoos$length, cuckoos$species), function(x) length(x) -
+     1))
> sqrt(sum(diffs^2)/sum(df))
```

Exercise 5
Plot a histogram of the `earconch` measurements for the `possum` data. The distribution should appear *bimodal* (two peaks). This is a simple indication of clustering, possibly due to sex differences. Obtain side-by-side boxplots of the male and female `earconch` measurements. How do these measurement distributions differ? Can you predict what the corresponding histograms would look like? Plot them to check your answer.

```
> par(mfrow = c(1, 2))
> hist(possum$earconch, main = "")
> boxplot(split(possum$earconch, possum$sex))
> par(mfrow = c(1, 1))
```

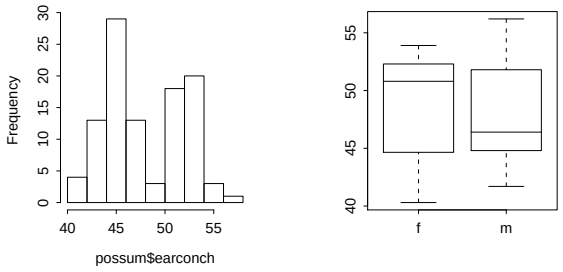

Figure 4: The left panel shows a histogram of possum ear conch measurements. The right panel shows side by side boxplots of the measurements, one for each sex.

Note: The following gives side by side histograms:

```
> par(mfrow = c(1, 2))
> hist(possum$earconch[possum$sex == "f"], border = "red", main = "")
> hist(possum$earconch[possum$sex == "m"], border = "blue", main = "")
> par(mfrow = c(1, 1))
```

The histograms make it clear that sex differences are not the whole of the explanation for the bimodality.

Note: We note various possible alternative plots.

Density plots are an alternative to histograms, with the further advantange that it is easy to overlay them. Alternatives 1 & 2 obtain overlaid density plots:

```
> "Alternative 1: Overlaid density plots"
> fden <- density(possum$earconch[possum$sex == "f"])
> mden <- density(possum$earconch[possum$sex == "m"])
> xlim <- range(c(fden$x, mden$x))
> ylim <- range(c(fden$y, mden$y))
> plot(fden, col = "red", xlim = xlim, ylim = ylim, main = "")
> lines(mden, col = "blue", lty = 2)
```

```
> library(lattice)
> "Alternative 2: Overlaid density plots, using the lattice package"
> print(densityplot(~earconch, data = possum, groups = sex), main = "")
```

Alternatives 3 and 4 give alternative forms of histogram plot.

```
> "Alternative 3: Overlaid histograms, using regular graphics"
> fhist <- hist(possum$earconch[possum$sex == "f"], plot = F, breaks = seq(from = 40,
+     to = 58, by = 2))
> mhist <- hist(possum$earconch[possum$sex == "m"], plot = F, breaks = seq(from = 40,
+     to = 58, by = 2))
> ylim <- range(fhist$density, mhist$density)
> plot(fhist, freq = F, xlim = c(40, 58), ylim = ylim, border = "red",
+     main = "")
> lines(mhist, freq = F, border = "blue", lty = 2)
```

Note the use of `border="red"` to get the histogram for females outlined in red. The parameter setting `col="red"` gives a histogram with the rectangles filled in red.

Unfortunately, `histogram()` in the lattice package ignores the parameter `groups`. With `histogram()`, we are limited to side by side histograms:

```
> "Alternative 4: Side by side histograms, using the lattice package"
> print(histogram(~earconch | sex, data = possum), main = "")
```

Both for density plots and for histograms, do we really want the separate total areas to be scaled to 1, as happens with the setting `freq=FALSE`, rather than to the total frequencies in the respective populations? This will depend on the specific application.

Exercise 6
Install the package *Devore5*, available from the CRAN sites. Then gain access to data on tomato yields by typing

```
library(Devore5)
data(ex10.22)
tomatoes <- ex10.22
```

This data frame gives tomato yields at four levels of salinity, as measured by electrical conductivity (`EC`, in nmhos/cm).

(a) Obtain a scatterplot of `yield` against `EC`.

(b) Obtain side-by-side boxplots of `yield` for each level of `EC`.

(c) The third column of the data frame is a factor representing the four different levels of `EC`. Comment upon whether the yield data are more effectively analyzed using `EC` as a quantitative or qualitative factor.

```
> library(Devore5)
> data(ex10.22)
> tomatoes <- ex10.22
> plot(yield ~ EC, data = tomatoes)
> boxplot(split(tomatoes$yield, tomatoes$EC))
```

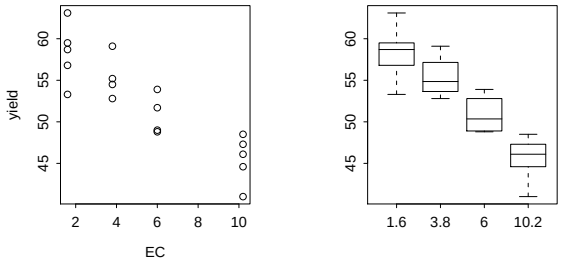

Figure 5: The left panel plots `yield` against `EC`. The right panel shows boxplots of `yield` for each distinct value of `EC`.

The data are more effectively analyzed using `EC` as a quantitative factor. Treating `EC` as a factor would ignore the linear or near linear dependence of `yield` on `EC`.