Part VII
Data Summary – Traps for the Unwary

For background, see SPDM: Section 3.

1 Multi-way Tables

	Small (<2cm)			Large (>=2cm)			Total	
	open	ultrasound		open	ultrasound		open	ultrasound
yes	81	234	yes	192	55	yes	273	289
no	6	36	no	71	25	no	77	61
Success rate	93%	87%		73%	69%		78%	83%

Table 1: Outcomes for two different types of surgery for kidney stones. The overall success rates (78% for open surgery as opposed to 83% for ultrasound) favor ultrasound. Comparison of the success rates for each size of stone separately favors, in each case, open surgery.

Exercise 1
Table 1 illustrates the potential hazards of adding a multiway table over one of its margins. Data are from a study (Charig, 1986) that compared outcomes for two different types of surgery for kidney stones; A: `open`, which used open surgery, and B: `ultrasound`, which used a small incision, with the stone destroyed by ultrasound. The data can be entered into R, thus:

```
> stones <- array(c(81, 6, 234, 36, 192, 71, 55, 25), dim = c(2,
+     2, 2), dimnames = list(Sucess = c("yes", "no"), Method = c("open",
+     "ultrasound"), Size = c("<2cm", ">=2cm")))
```

(a) Determine the success rate that is obtained from combining the data for the two different sizes of stone. Also determine the success rates for the two different stone sizes separately.

(b) Use the following code to give a visual representation of the information in the three-way table:

```
mosaicplot(stones, sort=3:1)
   # Re-ordering the margins gives a more interpretable plot.
```

Annotate the graph to show the success rates?

(c) Observe that the overall rate is, for open surgery, biased toward the open surgery outcome for large stones, while for ultrasound it is biased toward the outcome for small stones. What are the implications for the interpretation of these data?

[Without additional information, the results are impossible to interpret. Different surgeons will have preferred different surgery types, and the prior condition of patients will have affected the choice of surgery type. The consequences of unsuccessful surgery may have been less serious than for ultrasound than for open surgery.]

The relative success rates for the two different types of surgery, for the two stone sizes separately, can be calculated thus:

```
> stones[1, , ]/(stones[1, , ] + stones[2, , ])
```

To perform the same calculation after adding over the two stone sizes (the third dimension of the table), do

```
> stones2 <- stones[, , 1] + stones[, , 2]
> stones2[1, ]/(stones2[1, ] + stones2[2, ])
```

1.1 Which multi-way table? It can be important!

Each year the National Highway Traffic Safety Administration in the USA collects, using a random sampling method, data from all police-reported crashes in which there is a harmful event (people or property), and from which at least one vehicle is towed. The data in Table 2 summarize data in the data frame `nass9702cor`.

The data frame is a subset of the 1997-2002 data, restricted to front seat occupants and in various other ways. They are a corrected version of the data analyzed in Meyer and Finney (2005).[1] The web page `http://www.maths.anu.edu.au/~johnm/courses/dm/math3346/data/` has an R image file (`nass9702cor.RData`) that holds these data. For convenience, a subset of columns that will be used in this laboratory has been extracted into `nass_cds`, held in the image file **nass_cds.RData**, again available from the web page.

```
> load("nass9702cor.RData")
```

Here are the details of the extraction from `nass9702cor` into `nass_cds`.

```
> nass_cds <- nass9702cor[, c("dvcat", "national", "dead", "airbag",
+     "seatbelt", "frontal", "male", "age.of.o", "yearacc")]
> nass_cds$dead <- 2 - nass_cds$dead
> names(nass_cds)[8] <- "ageOFocc"
> table(nass_cds$seatbelt)

    0     1
 7644 18573

> nass_cds$seatbelt <- factor(nass_cds$seatbelt, labels = c("none",
+     "belted"))
> nass_cds$airbag <- factor(nass_cds$airbag, labels = c("none",
+     "airbag"))
```

The data are a sample. The use of a complex sampling scheme has the consequence that the sampling fraction differs between observations. Each point has to be multiplied by the relevant sampling fraction, in order to get a proper estimate of its contribution to the total number of accidents. The column `national` (*national inflation factor*) gives the relevant multiplier.

Other variables than those included in `nass_cds` might be investigated – those extracted into `nass_cds` are enough for present purposes.

seatbelt	airbag	dead	total	Prop_dead
none	none	24067	1366089	0.01762
belted	none	15609	4118833	0.00379
none	airbag	13760	885635	0.01554
belted	airbag	12159	5762975	0.00211

Table 2: Number of fatalities, by use of seatbelt and presence of airbag. Data are for front-seat occupants.

The following gives two classifications of the data – a simple classification according to airbag availability, and a slightly more detailed classification a/c use or otherwise of a functioning seatbelt.

[1] A SAS transport file from which the data in `nass9702cor` is derived, is available from the web page `http://www.stat.uga.edu/~mmeyer/airbags.htm`

```
> total <- with(nass_cds, as.data.frame(xtabs(national ~ airbag)))
> dead <- with(nass_cds, as.data.frame(xtabs(national * dead ~
+     airbag)))
> cbind(dead[, 1, drop = FALSE], dead = dead[, 2], total = total[,
+     2], Prop = dead[, 2]/total[, 2])

  airbag     dead   total        Prop
1   none 39676.02 5484922 0.007233652
2 airbag 25919.11 6648610 0.003898425

> total <- with(nass_cds, as.data.frame(xtabs(national ~ seatbelt +
+     airbag)))
> dead <- with(nass_cds, as.data.frame(xtabs(national * dead ~
+     seatbelt + airbag)))
> cbind(dead[, 1:2], dead = dead[, 3], total = total[, 3], Prop = dead[,
+     3]/total[, 3])

  seatbelt airbag     dead     total        Prop
1     none   none 24066.65 1366088.6 0.017617199
2   belted   none 15609.36 4118833.4 0.003789753
3     none airbag 13759.94  885635.3 0.015536805
4   belted airbag 12159.17 5762974.8 0.002109877
```

Exercise 2
The following generates the classification a/c airbag use, but adds a calculation of the estimate of the change in the number of deaths from the presence of an airbag. The relevant baseline expected proportion is given by the proportion of deaths for the corresponding cell of the table when an airbag was not installed.

```
> total <- with(nass_cds, as.data.frame(xtabs(national ~ airbag)))
> dead <- with(nass_cds, as.data.frame(xtabs(national * dead ~
+     airbag)))
> airbagAlone <- cbind(dead[, 1, drop = FALSE], dead = dead[, 2],
+     total = total[, 2], Prop = dead[, 2]/total[, 2])
> with(airbagAlone, dead[airbag == "none"] - total[airbag == "airbag"] *
+     Prop[airbag == "none"])

[1] -8417.715
```

Note the result. Why is it not a whole number? This figure (here, negative, i.e. airbags seem beneficial) will be termed the number of excess deaths due to use of airbags.
Repeat the calculation, now taking account of whether or not seatbelts were deployed. Compare the apaprent risk of airbag versus no airbag, for for each of the levels of `seatbelt` equal to `none` and `belted` separately. What is now the total number of excess deaths?

Exercise 3
At the end of this exercise is a listing of the function `bagrisk()`. Enter the function and run it with the default arguments, i.e. type

```
> bagrisk()
```

Compare the output with that obtained in Exercise 2 when the classification was a/c seatbelt (and airbag), and check that the output agrees.
Now do the following calculations, in turn:

(a) Classify according to `dvcat` as well as `seatbelt`. All you need do is add `dvcat` to the first argument to `bagrisk()`. What is now the total number of excess deaths?
[The categories are 0-9 kph, 10-24 kph, 25-39 kph, 40-54 kph, and 55+ kph]

(b) Classify according to `dvcat`, `seatbelt` and `frontal`, and repeat the calculations. What is now the total number of excess deaths?

Explain the dependence of the estimates of numbers of excess deaths on the choice of factors for the classification.

2 Weighting Effects – Example with a Continous Outcome

	baclofen	placebo
females	15	7
males	3	16

Table 3: Numbers of males and females on the two treatments, in a trial that investigated the effect of pentazocine on post-operative pain (VAS scores).

	min	mbac	mpl	fbac	fpl
2	10	1.76	1.76	2.18	2.55
3	30	1.31	1.65	3.48	4.15
4	50	0.05	0.67	3.13	3.66
5	70	−0.57	−0.25	3.03	2.05
6	90	−1.26	−0.50	2.08	0.61
7	110	−2.15	−2.22	1.60	0.34
8	130	−1.65	−2.18	1.38	0.67
9	150	−1.68	−2.86	1.76	0.76
10	170	−1.68	−3.23	1.06	0.39

Table 4: The table shows, separately for males and females, the effect of pentazocine on post-operative pain (average VAS scores), with (mbac and fbac) and without (mpl and fpl) preoperatively administered baclofen.
.

Exercise 4
The data in Table 3, in the data frame `gaba`, are from Gordon et al (1995).
[An image file that holds the data frame `gaba` is in the **data** subdirectory, and image files that hold the functions `plotGaba()` and `compareGaba()`, in the **functions** subdirectory, of the web page `http://www.maths.anu.edu.au/~johnm/courses/dm/math3346/`]
Table 4 has a tabular summary of the outcome of the trial to which Table 3 relates.

(a) What do you notice about the relative numbers on the two treatments?

(b) For each treatment, obtain overall weighted averages at each time point, using the numbers in Table 3 as weights. (These should be the numbers you would get if you divided the total over all patients on that treatment by the total number of patients.) This will give columns `avbac` and `avplac` that can be added the data frame.

(c) Plot `avbac` and `avplac` against time, on the same graph. On separate graphs, repeat the comparisons (a) for females alone and (b) for males alone. Which of these graphs make a correct and relevant comparison between baclofen and placebo (albeit both in the presence of pentazocine)?

3 Listing of the function `bagrisk()`

An image file that holds this function is available from my web page `http://www.maths.anu.edu.au/~johnm/courses/dm/math3346/functions`

```
> "bagrisk" <- function(form = national ~ seatbelt + airbag, data = nass_cds,
+     decpl = 6) {
+     funtxt <- deparse(form)
+     leftrt <- strsplit(funtxt, split = " ~ ")[[1]]
+     formdead <- formula(paste(leftrt[1], "*dead", " ~ ", leftrt[2],
+         sep = ""))
+     total <- with(nass_cds, as.data.frame(xtabs(form, data = data)))
+     dead <- with(nass_cds, as.data.frame(xtabs(formdead, data = data)))
+     nc <- match("Freq", names(total))
+     nway <- nc - 1
+     nair <- match("airbag", names(total))
+     none <- with(total, airbag == "none")
+     bag <- with(total, airbag == "airbag")
+     airdf <- cbind(dead[none, (1:nway)[-nair], drop = FALSE],
+         nobag_d = dead[none, nc], nobag_tot = total[none, nc],
+         bag_d = dead[bag, nc], bag_tot = total[bag, nc], nobagProp = dead[none,
+             nc]/total[none, nc], bagProp = dead[bag, nc]/total[bag,
+             nc])
+     airdf$xtra_d <- airdf$bag_d - airdf$bag_tot * airdf$nobagProp
+     printdf <- airdf
+     numcols <- c("nobag_d", "nobag_tot", "bag_d", "bag_tot",
+         "xtra_d")
+     fraccols <- c("nobagProp", "bagProp")
+     printdf[, numcols] <- round(printdf[, numcols])
+     printdf[, fraccols] <- round(printdf[, fraccols], decpl)
+     print(printdf)
+     invisible(airdf)
+ }
```