

## Input data summary

It is important to understand the nature of the data you are working with: how large is the sample, how appropriate is the sample for asking the science questions you have, how well did you sample across a gradient of values for each covariate, what correlations exist among your covariates, is there any evidence of outliers or errors in the dataset, etc. To do this, you'll need to work with the data in a stats package to obtain summary statistics and to view some plots of the covariates.

In R, we could easily import a text file containing the data and view some summary stats as follows.

Start with a text file that contains:

```
area      freq  sex  weight length time  status one
Treatment 192.78 Female 24.5 108 118 0 1
Control   191.19 Female 27.5 112 15 0 1
Treatment 192.57 Female 32.5 112.5 1 1
Treatment 192.79 Female 31 113 16 0 1
Treatment 192.68 Female 26.5 113 26 0 1
...
Control   192.71 Male 38.6 131 130 0 1
Treatment 192.63 Male 45.1 132 52 0 1
Treatment 192.37 Male 45.6 133 35 0 1
Control   192.79 Male 40 133 119 0 1
Treatment 192.4 Male 45.5 134 105 0 1
```

In R, input the file, create numeric versions of the character variables (*area* & *sex*), & summarize

```
fawns <- read.table(file="fawns.txt", sep="\t", header=TRUE)
# Create numeric versions of 'area' & 'sex' that are 0/1
fawns$AREA <- as.numeric(fawns$area) - 1 # female=0, male=1
fawns$SEX <- as.numeric(fawns$sex) - 1 # control=0, trtmt=1
```

```
summary(fawns)
```

```
> summary(fawns)
   area      freq      sex      weight      length
Control :59  Min.   :191.1  Female:57  Min.   :22.80  Min.   :108.0
Treatment:56 1st Qu.:192.0  Male  :58  1st Qu.:31.90 1st Qu.:120.0
              Median :192.4              Median :33.60  Median :124.0
              Mean   :192.2              Mean   :34.38  Mean   :123.2
              3rd Qu.:192.7              3rd Qu.:36.60 3rd Qu.:127.0
              Max.   :192.8              Max.   :72.00  Max.   :135.5

   time      status      one      AREA      SEX
Min.   : 7.00  Min.   :0.0000  Min.   :1  Min.   :0.000  Min.   :0.0000
1st Qu.:27.75 1st Qu.:0.0000  1st Qu.:1  1st Qu.:0.000 1st Qu.:0.0000
Median :59.50 Median :0.0000  Median :1  Median :0.000 Median :1.0000
Mean   :61.50 Mean   :0.4087  Mean   :1  Mean   :0.487  Mean   :0.5043
3rd Qu.:98.25 3rd Qu.:1.0000  3rd Qu.:1  3rd Qu.:1.000 3rd Qu.:1.0000
Max.   :130.00 Max.   :1.0000  Max.   :1  Max.   :1.000  Max.   :1.0000
NA's   : 47.00
```

Note: there appears to be one very heavy fawn; is it an error in the data?

Next, you might want to look at simple cross-tabulations for fate versus sex or area

```
> a=table(fawns$sex,fawns$status)
> a
```

	0	1
Female	28	29
Male	40	18

```
>prop.table(a,1) # build a table of proportions by row (dimension 1 of table)
```

	0	1
Female	0.4912281	0.5087719
Male	0.6896552	0.3103448

```
> b=table(fawns$area,fawns$status)
```

```
> b
```

	0	1
Control	36	23
Treatment	32	24

```
> prop.table(b,1)
```

	0	1
Control	0.6101695	0.3898305
Treatment	0.5714286	0.4285714

Next, you might want to take a look at plots of pairs of covariates to gain understanding of the sample in hand.

```
f=subset(fawns,select=c(status,AREA,SEX,weight,length))
pairs(f)
```

