Using R

A basic data type in R is a vector. As an example:

Declaring a vector and manipulating it

```
>x<-c(1,5,10,45)    # declare a vector
>x                  # display the vector
>mean(x)
>1/x
>sum(x)
>sd(x)
>length(x)
>range(x)
```

Other nice ways to declare vectors

```
>x=seq(0,20)
>x=seq(0,20,by=2)
>x=rep(5,10)
>x=rep(c(1,2,3),10)
>x=sample(20)
```

Random samples of your vec, but can be dangerous

Using R

Accessing vector elements

```
x=seq(50,60)
x[2]
x[2:5]
x[c(2,4,8)]
x[-2]  [-2] = excludes [2]
x[-c(2,4,8)]
```

Declaring and accessing matrices

```
x=seq(1,10)
y=matrix(x,nrow=2,ncol=5)
y[3,5] Throws error b/c no 3rd row
y[,3]
y[2,]
t(y)    # Matrix transpose
y%*%t(y)   #  %*% is the matrix multiplication operator
```

Using R

Data frames

Different vectors can be joined together into one object with the constraint that there is the same number of elements per vector.

```
z=factor(c("Ctrl","Ctrl","A","A","B","B"))
x=c(5,3,4,NA,10,4)
y=c(TRUE,TRUE,FALSE, TRUE,FALSE,TRUE)
d=data.frame(labels=z,heights=x,outcome=y)
```

Missing data

The `NA` construct can be used to specify missing data:

```
x=c(5,3,4,NA)
is.na(x)
mean(x)
mean(x, na.rm=TRUE)
```

Using R

The apply functions

There are three main functions: apply,sapply, tapply

```
# apply:  Apply function to all rows/columns of a matrix
x<-seq(1,10)
y<-matrix(x,nrow=2,ncol=5)
apply(y,1,mean)
apply(y,2,mean)

# sapply:  Apply function to each element of a list or dataframe
x <- list(a = 1:10, beta = exp(-3:3), logic = c(TRUE,FALSE,FALSE,TRUE))
sapply(x,mean)

# tapply: Apply a function to each set of elements with the
#              same level of a factor
z<-factor(c("Ctrl","Ctrl","A","A","B","B"))
x<-c(5,3,4,NA,10,4)
tapply(x,z,mean,na.rm=TRUE)
```

1 = keep dim 1 so mean of rows

2 = keep dim 2 so mean of cols

Extended example

Part 1: Hardy-Weinberg in practice **looking at 4,014 loci**

- Read in genotype data from a file (4,014 SNPs in 60 individuals)
- Exploratory plot of heterozygosity vs. allele frequency.
  - Recall that the Hardy-Weinberg expected proportion of heterozygotes H as a function of allele frequency p is: $H = 2p(1-p)$
- Formal test of Hardy-Weinberg proportions using a χ^2-test for each SNP

Part 2: Finding a quantitative trait locus via association mapping

- Read in phenotype data from a file (fasting glucose in units of mmol/L)
- Test for each SNP whether genotype is correlated ("associated") with phenotypic trait value using a linear model framework (and the lm function)
- Find which SNP has the association and visualize its effect in a boxplot

Data source

- 2.3 million SNPs genotyped on Hapmap CEU founders (60 individuals). Data downloaded in plink format from plink website.
- SNPs from chromosome 2 and LD pruning undertaken (`--indep-pairwise 50 5 0.2`) to make the data set smaller