Resampling

Resampling is used when you are ignorant of the characteristics of the true sampling (t, F, etc) distribution.
These techniques allow the characteristics of the distribution to be estimated from the data.

- Jacknifing
- Bootstrapping
- Permutations

Jackknife

Jackknifing is used to estimate

- bias
- standard error

It functions by removing a single value from the observations, recalculating a statistic, and repeating for the original data and all one-less combinations

Algorithm

1. calculate statistic (mean, sd, etc)
2. recalculate n statistics removing each observation in turn
3. define $\phi_i = nS - (n-1)S_i$ where S is the statistic and S_i is the statistic calculated on a sample without observation i.
4. $\bar{\phi}$ is the jackknife estimate of the parametric statistic.
5. $\Sigma(\bar{\phi} - \phi_i)^2/(n-1)$ is an estimate of the variance associated with the jackknife estimate
6. SE defined as you might expect

Example: data

```
> library(DAAG)
> data(cuckoos)
> wrens <- cuckoos[cuckoos$species == "wren", ]
> wrens$length
```

```
 [1] 19.8 22.1 21.5 20.9 22.0 21.0 22.3 21.0 20.3 20.9 22.0 20.0 20.8 21.2 21.0
```

```
> wrens
```

```
    length breadth species  id
106   19.8    15.0    wren 224
107   22.1    16.0    wren 225
108   21.5    16.2    wren 226
109   20.9    15.7    wren 227
110   22.0    16.2    wren 228
111   21.0    15.5    wren 229
112   22.3    16.0    wren 230
113   21.0    15.9    wren 231
114   20.3    15.5    wren 232
115   20.9    15.9    wren 233
116   22.0    16.0    wren 234
117   20.0    15.7    wren 235
118   20.8    15.9    wren 236
119   21.2    16.0    wren 237
120   21.0    16.0    wren 238
```

Example: jackknife calcs

```
> m <- mean(wrens$length)
> mj <- NULL
> for (i in 1:length(wrens$length)) {
+     mj <- c(mj, mean(wrens$length[-i]))
+ }
> mj
```

```
 [1] 21.21429 21.05000 21.09286 21.13571 21.05714 21.12857 21.03571 21.12857
 [9] 21.17857 21.13571 21.05714 21.20000 21.14286 21.11429 21.12857
```

```
> m
```

```
[1] 21.12
```

Calcs 2

```
> n <- length(wrens$length)
> pseudovalues <- n * m - (n - 1) * mj
> pseudovalues
```

```
 [1] 19.8 22.1 21.5 20.9 22.0 21.0 22.3 21.0 20.3 20.9 22.0 20.0 20.8 21.2 21.0
```

```
> mhat <- mean(pseudovalues)
> mhat
```

```
[1] 21.12
```

```
> bias <- m - mhat
> bias
```

```
[1] -1.065814e-14
```

```
> sem.j <- sd(pseudovalues)/sqrt(n)
> sem.j
```

```
[1] 0.1947404
```

Examining each observation

Influence of particular samples: The *sample influence function* is the difference between the mean of the jackknifed statistic (mean(pseudovalues)) and the individual pseudovalues:

$$\bar{\phi} - \phi_i$$

Essentially, it allows identification of outliers for a particular statistic

```
> mhat - pseudovalues
```

```
 [1]  1.32 -0.98 -0.38  0.22 -0.88  0.12 -1.18  0.12  0.82  0.22 -0.88  1.12
[13]  0.32 -0.08  0.12
```

Bootstrapping: uses

Like Jackknifing, bootstrapping is also used for estimation of:

- bias
- sample statistics and confidence intervals around those statistics

Bootstrapping: details

- Bootstrapping works by resampling with replacement a complete-sized dataset, recalculating statistics and repeating 100s or 1000s of times.
- The mean of the statistic for all the bootstraps is the estimate for that statistic.

- The standard deviation of the bootstrap estimates is the standard error of the statistic
- Furthermore, the distribution of the bootstrapped statistic should be more or less normally distributed, so it is easy to calculate confidence intervals

Bootstrapping: example

```
> numboot <- 100
> boot.est <- NULL
> for (i in 1:numboot) {
+     boot.est <- c(boot.est, sd(sample(wrens$length, length(wrens$length),
+         replace = T)))
+ }
> mean(boot.est)

[1] 0.7321769

> sd(boot.est)

[1] 0.1092888

> mean(boot.est) + qt(0.975, 99) * sd(boot.est)

[1] 0.9490295

> mean(boot.est) - qt(0.975, 99) * sd(boot.est)

[1] 0.5153243
```

Slightly more exotic application

```
ATCADH       ------atgt ctaccaccgg acagattatt cgatgcaaag ctgctgtggc
ATHADH       ------atgt ctaccaccgg acagattatt cgatgcaaag ctgctgtggc
PSADH1       ---atgtcga acacagttgg tcagatcatc aagtgcagag ctgcggttgc
PHADH1       atgtcgagca atactgctgg tcaagtcatt cgttgcaaag ctgcggttgc
VVU36586     ---atgtcag gcactgctgg tcaagtcatc tgctgcaaag ctgctgtggc
LEADH2       ---atgtcga ctactgtagg ccaagtcatt cgttgcaaag ctgctgtggc
FAADH        ---atgtcaa gtactgaggg aaaggtcata tgctgcagag ctgctgtggc
HVADH2       ------atgg cgaccgccgg gaaggtgatc aagtgcaaag cggcggtggc
HVADH3       ------atgg cgaccgctgg gaaggtgatc aagtgcaaag cggcggtggc
MZEADH1CM    ------atgg cgaccgcggg gaaggtgatc aagtgcaaag ctgcggtggc
ZMADH1FA     ------atgg cgaccgcggg gaaggtgatc aagtgcaaag ctgcggtggc
```

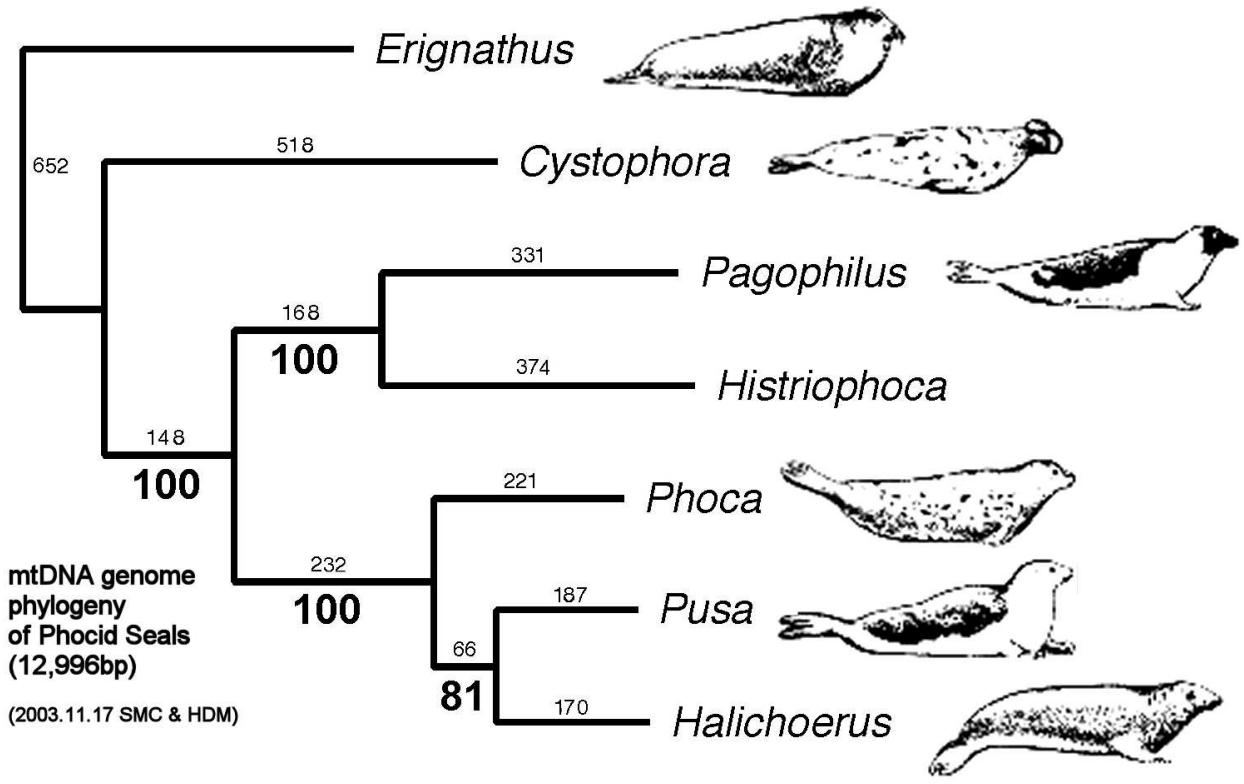
```
RICADH2A     ---------a tggcgacagg gaaggtgatc aagtgcaaag cggcggtggc
ZMADH2N      ------atgg cgacagcagg aaaggtgatc aagtgcagag gctgccgtga
HUMADH6      ------atga gtactacagg ccaagtcatc agatgcaaag cagccatact
ANADH3       ---------- -----atgtc tgtccccgaa gtgcaatggg cccaagtggt
EMEALCA      ---------- -----atgtg catccccact atgcaatggg cccaggtcgc
SCZADH       ---------- ---------- ---------- ---------- ----------
DNADH1       ---------- ---------- ---------- ---------- ----------
DMADH        ---------- ---------- ---------- ---------- ----------
```

Phylogenetic inference

- Model: evolutionary model
- Hypothesis: phylogenetic tree
- Data: previous slide

The unit of resampling in the data in this case is the variable (column)

Phylogenetic tree with bootstrap values



Randomization/Permutation Tests

These tests usually compare a single value to sample distribution generated by randomizing the data. The single value could be any statistic you could imagine.

- t
- X^2
- F
- difference in means between groups
- variance among groups
- slopes of lines

Algorithm

- Calculate test statistic on original data
- randomly rearrange order between independent and dependent variables
- recalculate test-statistic
- repeat lots (100s) of times
- compare actual value to distribution of randomized values.

Example for two groups

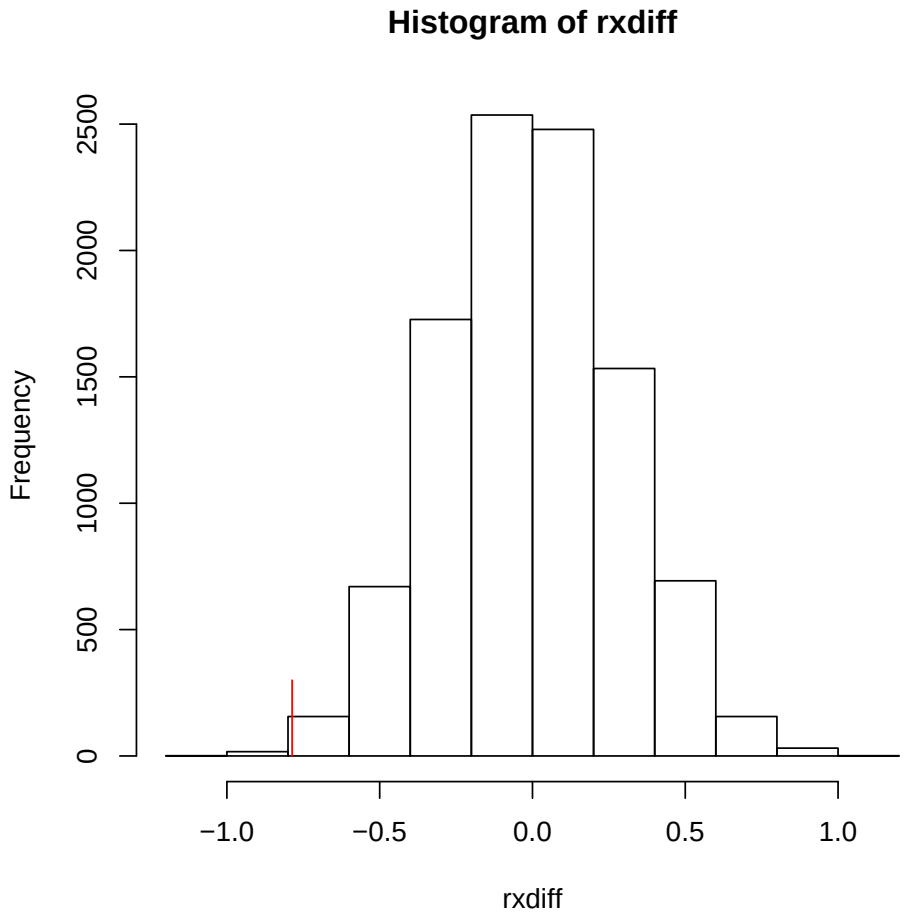
```
> mpipit <- cuckoos[cuckoos$species == "meadow.pipit", ]
> tpipit <- cuckoos[cuckoos$species == "tree.pipit", ]
> X <- mpipit$length
> Xn <- length(X)
> Y <- tpipit$length
> Yn <- length(Y)
> xdiff <- mean(X) - mean(Y)
> xdiff

[1] -0.7866667
```

```
> XY <- c(X, Y)
> rxdiff <- NULL
> for (i in 1:10000) {
+     randord <- sample(XY, (Xn + Yn), replace = F)
+     x <- randord[1:Xn]
+     y <- randord[c(-1:-Xn)]
+     rxdiff <- c(rxdiff, (mean(x) - mean(y)))
+ }
```

Example continued

```
> hist(rxdiff)
> points(c(xdiff, xdiff), c(0, 300), type = "l", col = "red")
> sum(rxdiff < xdiff)/length(rxdiff)
```

```
[1] 0.0018
```



Resample what?

The unit of resampling is important to define in any resampling study.

- Often just single observations
- sometimes it is only important to mix among groups