STAT 528 - Advanced Regression Analysis II

Aster models

Daniel J. Eck
Department of Statistics
University of Illinois

Learning Objectives Today

- aster model example
- aster analysis

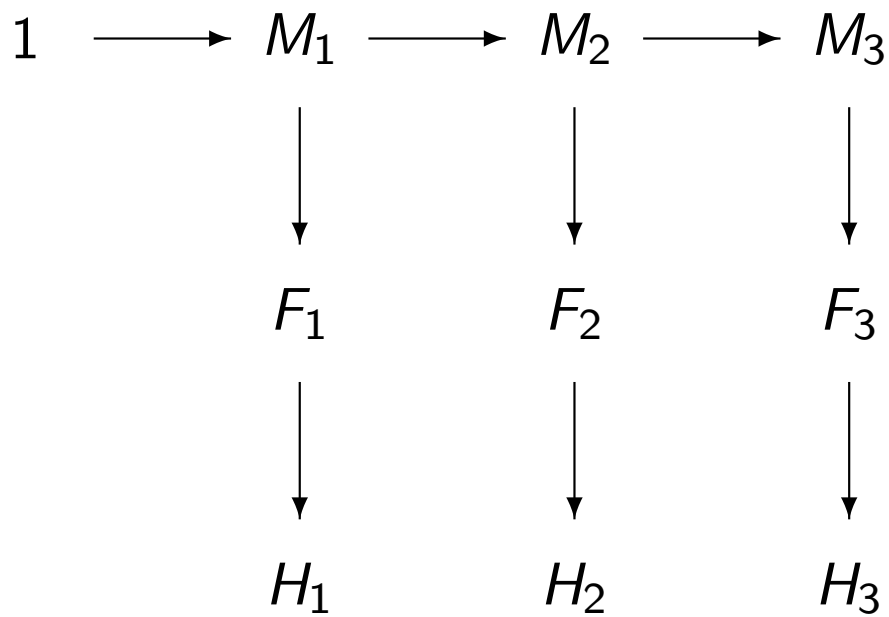
The variables under consideration:

- `nsloc` north-south location of each individual in the experimental plot
- `ewloc` east-west location of each individual in the experimental plot
- `pop` the ancestral population of each individual

Each individual was grown from seed taken from a surviving population in a prairie remnant in western Minnesota near the Echinacea Project field site.

Darwinian fitness (our best surrogate of Darwinian fitness) is total flower head count over the years of data collection.

The aster graph for *Echinacea angustifolia* (aster plants)



We load in necessary packages:

```
library(tidyverse)
library(ggplot2)
library(aster)
library(aster2)
```

Initial data processing

Here is a brief look at the data:

```
data("echinacea")
names(echinacea)
```

```
##  [1] "redata"        "repred"        "regroup"       "recode"
##  [5] "families"      "redelta"       "initial"       "response.name"
##  [9] "pred"          "group"         "code"
```

```
head(echinacea$redata)
```

```
##           pop ewloc nsloc varb resp id
## 1.ld02   NWLF    -8   -11 ld02    0  1
## 2.ld02 Eriley    -8   -10 ld02    1  2
## 3.ld02   NWLF    -8    -9 ld02    0  3
## 4.ld02    SPP    -8    -8 ld02    0  4
## 5.ld02    SPP    -8    -7 ld02    0  5
## 6.ld02 Eriley    -8    -6 ld02    1  6
```

```
echinacea$redata %>% filter(id == 1)
```

```
##           pop ewloc nsloc   varb resp id
## 1.ld02   NWLF    -8   -11   ld02    0  1
## 1.ld03   NWLF    -8   -11   ld03    0  1
## 1.ld04   NWLF    -8   -11   ld04    0  1
## 1.fl02   NWLF    -8   -11   fl02    0  1
## 1.fl03   NWLF    -8   -11   fl03    0  1
## 1.fl04   NWLF    -8   -11   fl04    0  1
## 1.hdct02 NWLF    -8   -11 hdct02    0  1
## 1.hdct03 NWLF    -8   -11 hdct03    0  1
## 1.hdct04 NWLF    -8   -11 hdct04    0  1
```

```
echinacea$redata %>% filter(id == 6)
```

```
##             pop ewloc nsloc   varb resp id
## 6.ld02   Eriley    -8    -6   ld02    1  6
## 6.ld03   Eriley    -8    -6   ld03    1  6
## 6.ld04   Eriley    -8    -6   ld04    1  6
## 6.fl02   Eriley    -8    -6   fl02    0  6
## 6.fl03   Eriley    -8    -6   fl03    0  6
## 6.fl04   Eriley    -8    -6   fl04    1  6
## 6.hdct02 Eriley    -8    -6 hdct02    0  6
## 6.hdct03 Eriley    -8    -6 hdct03    0  6
## 6.hdct04 Eriley    -8    -6 hdct04    1  6
```

We can see the proportion of individuals that survive each year.

```
## M1
echinacea$redata %>% filter(varb == "ld02") %>% pull(resp) %>% table()
```

```
## .
##   0   1
## 158 412
```

```
## M2
echinacea$redata %>% filter(id %in% (echinacea$redata %>%
                                      filter(varb == "ld02" & resp == 1) %>%
                                      pull(id)) & varb == "ld03") %>%
  pull(resp) %>% table()
```

```
## .
##   0   1
##  20 392
```

```
## M3
echinacea$redata %>% filter(id %in% (echinacea$redata %>%
                                      filter(varb == "ld03" & resp == 1) %>%
                                      pull(id)) & varb == "ld04") %>%
  pull(resp) %>% table()
```

```
## .
##   0   1
##  14 378
```

We can see the proportion of individuals that flower each year.

```
## F1
echinacea$redata %>% filter(id %in% (echinacea$redata %>%
                                      filter(varb == "ld02" & resp == 1) %>%
                                      pull(id)) & varb == "fl02") %>%
  pull(resp) %>% table()
```

```
## .
##   0   1
## 253 159
```

```
## F2
echinacea$redata %>% filter(id %in% (echinacea$redata %>%
                                      filter(varb == "ld03" & resp == 1) %>%
                                      pull(id)) & varb == "fl03") %>%
  pull(resp) %>% table()
```

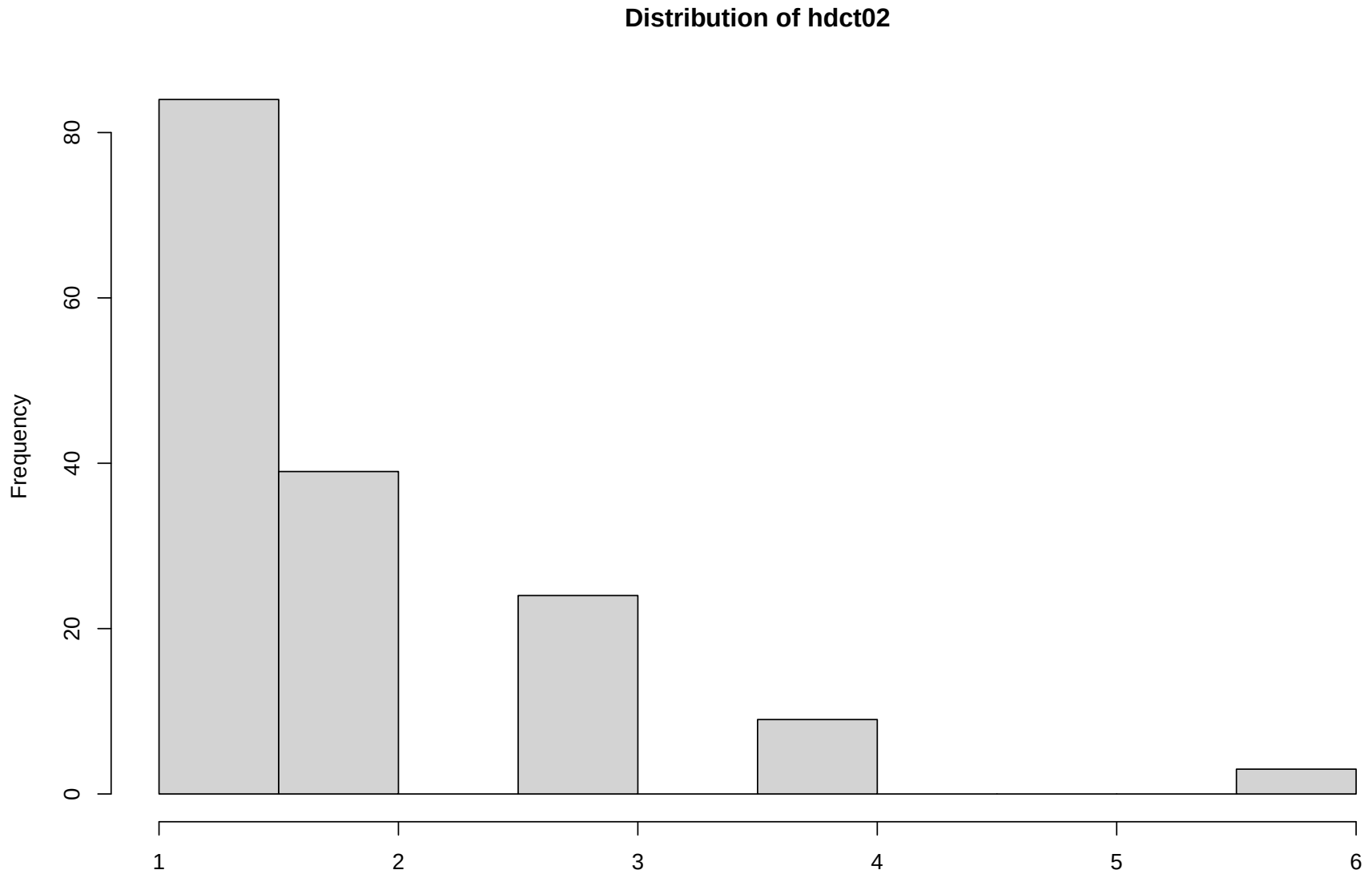
```
## .
##   0   1
## 266 126
```

```
## F3
echinacea$redata %>% filter(id %in% (echinacea$redata %>%
                                      filter(varb == "ld04" & resp == 1) %>%
                                      pull(id)) & varb == "fl04") %>%
  pull(resp) %>% table()
```

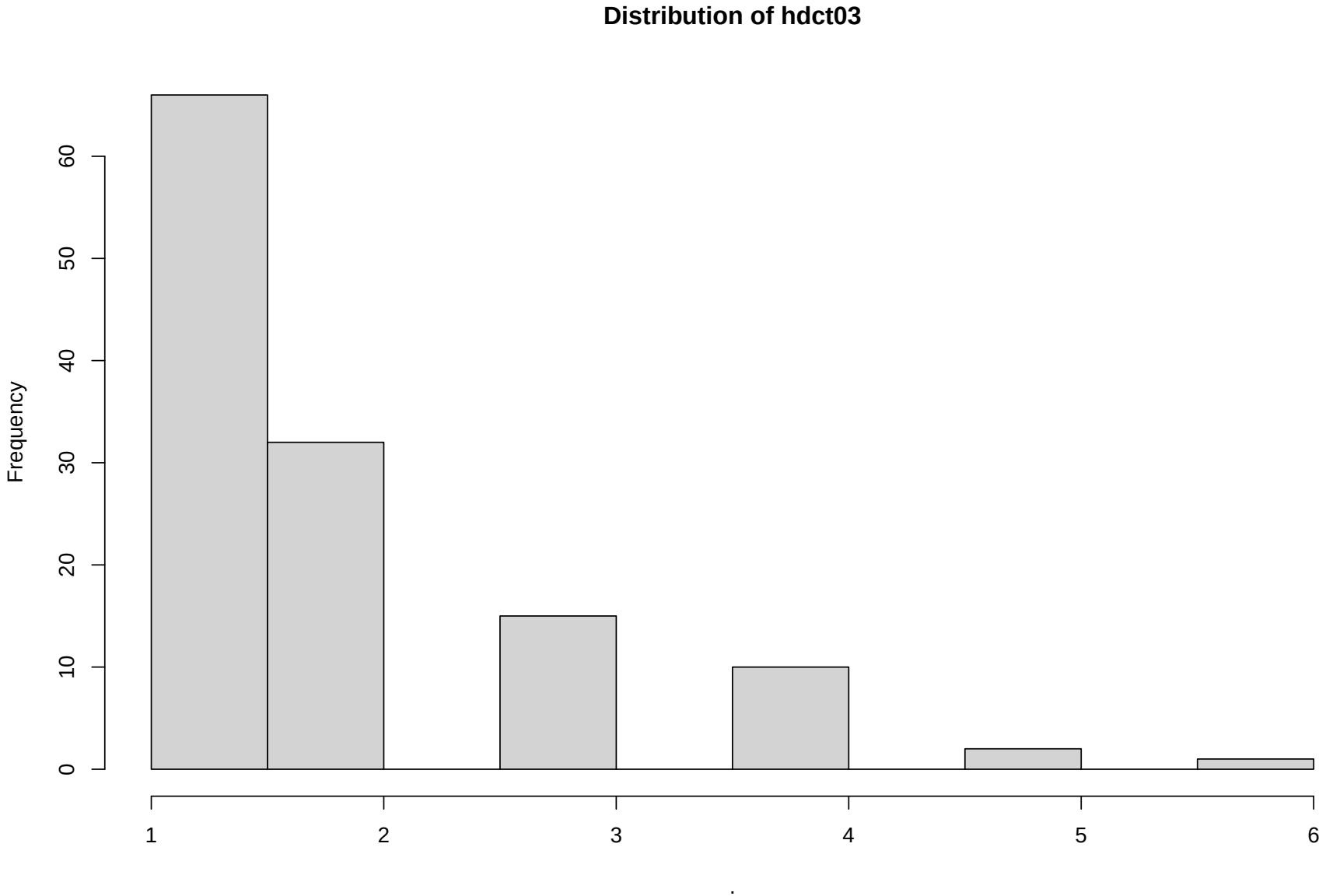
```
## .
##   0   1
## 162 216
```

We can see the distribution of head counts each year.

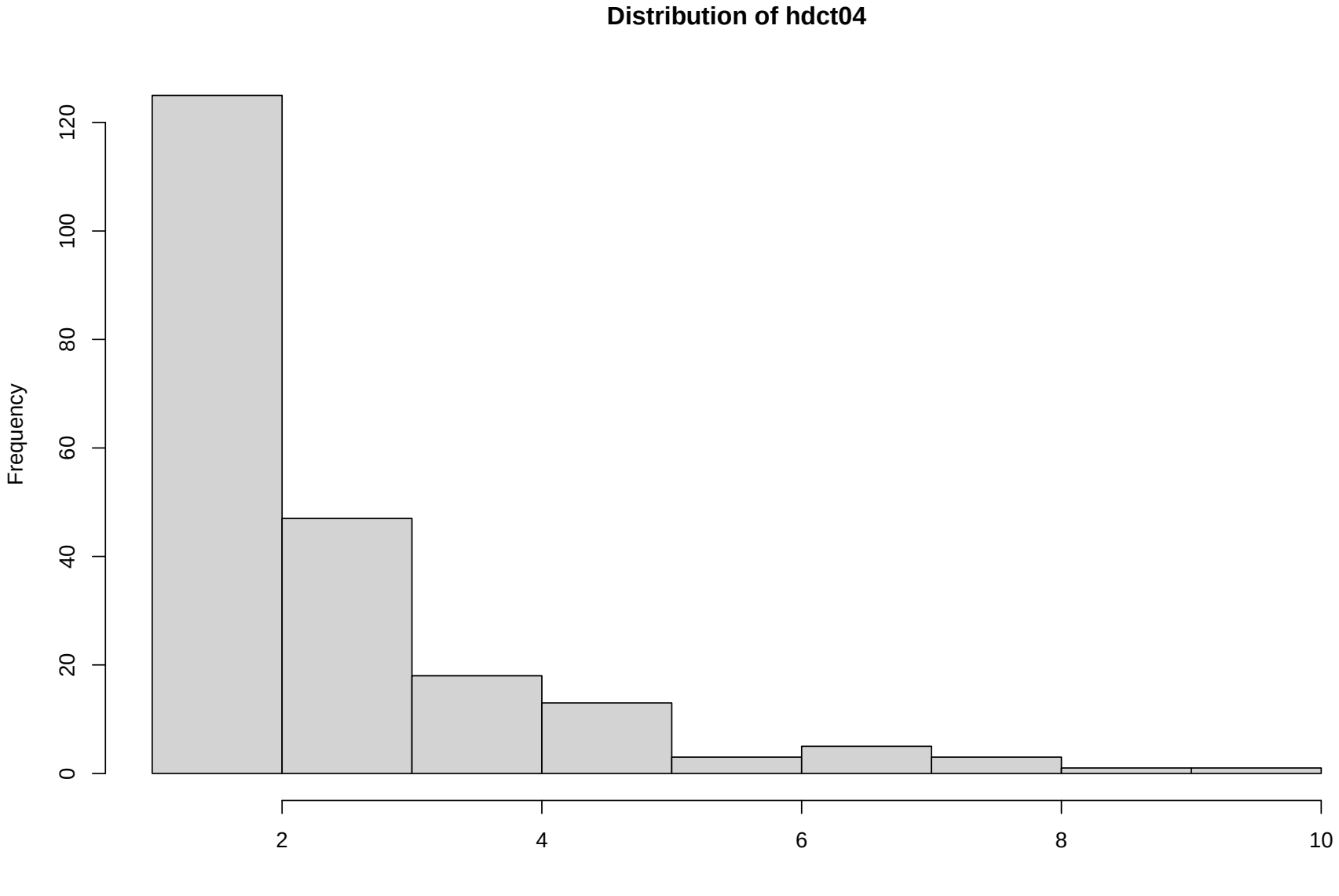
```
echinacea$redata %>% filter(id %in% (echinacea$redata %>%
                                      filter(varb == "fl02" & resp == 1) %>%
                                      pull(id)) & varb == "hdct02") %>%
  pull(resp) %>% hist(., main = "Distribution of hdct02")
```



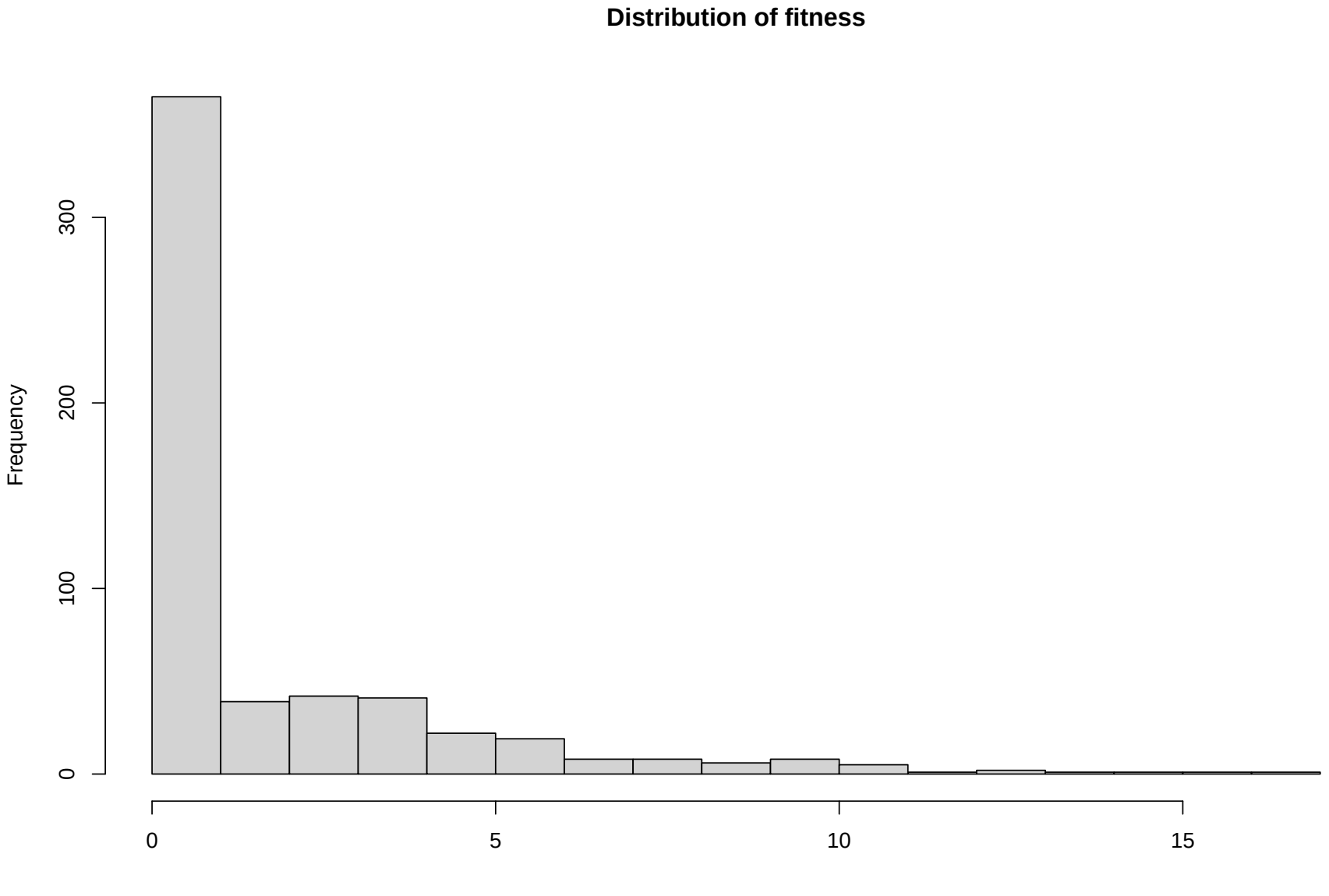
```
echinacea$redata %>% filter(id %in% (echinacea$redata %>%
                                      filter(varb == "fl03" & resp == 1) %>%
                                      pull(id)) & varb == "hdct03") %>%
  pull(resp) %>% hist(., main = "Distribution of hdct03")
```



```
echinacea$redata %>% filter(id %in% (echinacea$redata %>%
                                      filter(varb == "fl04" & resp == 1) %>%
                                      pull(id)) & varb == "hdct04") %>%
  pull(resp) %>% hist(., main = "Distribution of hdct04")
```



```
echinacea$redata %>% group_by(id) %>%
  filter(varb %in% c("hdct02","hdct03","hdct04")) %>%
  summarise(fitness = sum(resp)) %>%
  pull(fitness) %>% hist(., main = "Distribution of fitness", breaks = 20)
```



Aster analysis preliminaries

The variables that correspond to nodes of the graph are, in the order they are numbered in the graph

```
vars <- c("ld02", "ld03", "ld04", "fl02", "fl03",
                      "fl04", "hdct02", "hdct03", "hdct04")
```

The graphical structure is specified by a vector that gives for each node the index (not the name) of the predecessor node or zero if the predecessor is an initial node.

```
pred <- c(0, 1, 2, 1, 2, 3, 4, 5, 6)
```

This says the predecessor of the first node given by the `vars` vector is initial (because `pred[1] == 0`), the predecessor of the second node given by the `vars` vector is the first node given by the `vars` vector (because `pred[2] == 1`), and so forth.

```
foo <- rbind(vars, c("initial", vars)[pred + 1])
rownames(foo) <- c("successor", "predecessor")
foo
```

```
##             [,1]      [,2]   [,3]   [,4]   [,5]   [,6]   [,7]     [,8]
## successor   "ld02"    "ld03" "ld04" "fl02" "fl03" "fl04" "hdct02" "hdct03"
## predecessor "initial" "ld02" "ld03" "ld02" "ld03" "ld04" "fl02"   "fl03"
##             [,9]
## successor   "hdct04"
## predecessor "fl04"
```

That's right.

The last part of the specification of the graph is given by a corresponding vector of integers coding families (distributions). The default is to use the codes:

- 1 = Bernoulli
- 2 = Poisson
- 3 = zero-truncated Poisson

Optionally, the integer codes specify families given by an optional argument `famlist` to functions in the `aster` package, and this can specify other distributions besides those in the default coding.

```
fam <- c(1, 1, 1, 1, 1, 1, 3, 3, 3)
rbind(vars, fam)
```

```
##      [,1]   [,2]   [,3]   [,4]   [,5]   [,6]   [,7]     [,8]     [,9]
## vars "ld02" "ld03" "ld04" "fl02" "fl03" "fl04" "hdct02" "hdct03" "hdct04"
## fam  "1"    "1"    "1"    "1"    "1"    "1"    "3"      "3"      "3"
```

There is one more step before we can fit models.

The R function `aster` which fits aster models wants the data in long rather than wide format, the former having one line per node of the graph rather than one per individual.

```
## aster example already in long format
redata <- data.frame(echinacea$redata, root = 1)
head(redata)
```

```
##           pop ewloc nsloc varb resp id root
## 1.ld02   NWLF    -8   -11 ld02    0  1    1
## 2.ld02 Eriley    -8   -10 ld02    1  2    1
## 3.ld02   NWLF    -8    -9 ld02    0  3    1
## 4.ld02    SPP    -8    -8 ld02    0  4    1
## 5.ld02    SPP    -8    -7 ld02    0  5    1
## 6.ld02 Eriley    -8    -6 ld02    1  6    1
```

All of the variables in `echinacea` that are named in `vars` are gone. They are packed into the variable `resp`.

Which components of `resp` correspond to which components of `vars` is shown by the new variable `varb`.

```
levels(redata$varb)
```

```
## [1] "ld02"   "ld03"   "ld04"   "fl02"   "fl03"   "fl04"   "hdct02" "hdct03"
## [9] "hdct04"
```

Fitting aster models

We will now discuss fitting aster models.

Different families for different nodes of the graph means it makes no sense to have terms of the regression formula applying to different nodes.

In particular, it makes no sense to have one *intercept* for all nodes. To in effect get a different *intercept* for each node in the graph, include `varb` in the formula

$$\texttt{y} \sim \texttt{varb} + \texttt{...}$$

The categorical variable `varb` gets turned into as many dummy variables as there are nodes in the graph, one is dropped, and the *intercept* dummy variable.

Similar thinking says we want completely different regression coefficients of all kinds of predictors for each node of the graph.

That would lead us to formulas like

$$\texttt{y} \sim \texttt{varb + varb:(...)}$$

where ... is any other part of the formula.

We should not think of this formula as specifying *interaction* between `varb` and *everything else* but rather as specifying separate coefficients for *everything else* for each node of the graph.

That being said, formulas like this would likely yield too many regression coefficients to estimate well.

Maybe different for each kind of node (whatever that may mean) would be enough.

```
layer <- gsub("[0-9]", "", as.character(redata$varb))
redata <- data.frame(redata, layer = layer)
unique(layer)
```

```
## [1] "ld"   "fl"   "hdct"
```

Maybe

$$\texttt{y} \sim \texttt{varb + layer:(...)}$$

is good enough? But formulas like this would still yield too many regression coefficients to estimate well.

In aster models regression coefficients *for* a node of the graph also influence all *earlier* nodes of the graph (predecessor, predecessor of predecessor, predecessor of predecessor of predecessor, etc.)

So maybe it would be good enough to only have separate coefficients for the layer of the graph consisting of terminal nodes?

```
fit <- as.numeric(layer == "hdct")
redata <- data.frame(redata, fit = fit)
unique(fit)
```

```
## [1] 0 1
```

Maybe

$$\texttt{y} \sim \texttt{varb + fit:(...)}$$

is good enough.

We called the variable we just made up `fit` which is short for Darwinian fitness.

The regression coefficients in terms specified by ... have a direct relationship with expected Darwinian fitness (or a surrogate of Darwinian fitness).

And that is usually what is wanted in life history analysis.

We now fit our first aster model.

```
aout <- aster(resp ~ varb + layer : (nsloc + ewloc) +
                            fit : pop, pred, fam, varb, id, root, data = redata)
summary(aout)
```

```
##
## Call:
## aster.formula(formula = resp ~ varb + layer:(nsloc + ewloc) +
##     fit:pop, pred = pred, fam = fam, varvar = varb, idvar = id,
##     root = root, data = redata)
##
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -1.079946   0.241164  -4.478 7.53e-06 ***
## varbld03         1.769353   0.529200   3.343 0.000827 ***
## varbld04         4.217879   0.368282  11.453  < 2e-16 ***
## varbfl02         0.029302   0.315703   0.093 0.926050
## varbfl03        -0.319794   0.316120  -1.012 0.311720
## varbfl04        -0.314920   0.295018  -1.067 0.285763
## varbhdct02       1.350716   0.259254   5.210 1.89e-07 ***
## varbhdct03       1.372676   0.262255   5.234 1.66e-07 ***
## varbhdct04       1.880630   0.251000   7.493 6.75e-14 ***
## layerfl:nsloc    0.070102   0.014652   4.785 1.71e-06 ***
## layerhdct:nsloc -0.005804   0.005550  -1.046 0.295638
## layerld:nsloc    0.007165   0.005867   1.221 0.221957
## layerfl:ewloc    0.017977   0.014413   1.247 0.212294
## layerhdct:ewloc  0.007606   0.005561   1.368 0.171381
## layerld:ewloc   -0.004787   0.005919  -0.809 0.418635
## fit:popAA        0.129238   0.089129   1.450 0.147058
## fit:popEriley   -0.049561   0.071279  -0.695 0.486858
## fit:popLf       -0.033279   0.079573  -0.418 0.675789
## fit:popNessman  -0.186269   0.127787  -1.458 0.144936
## fit:popNWLF      0.021028   0.063600   0.331 0.740920
## fit:popSPP       0.149179   0.067716   2.203 0.027593 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

The regression coefficients are of little interest.

The main interest is in what model among those that have a scientific interpretation fits the best.

```
aout.smaller <- aster(resp ~ varb +
  fit : (nsloc + ewloc + pop),
  pred, fam, varb, id, root, data = redata)
aout.bigger <- aster(resp ~ varb +
  layer : (nsloc + ewloc + pop),
  pred, fam, varb, id, root, data = redata)
anova(aout.smaller, aout, aout.bigger)
```

```
## Analysis of Deviance Table
##
## Model 1: resp ~ varb + fit:(nsloc + ewloc + pop)
## Model 2: resp ~ varb + layer:(nsloc + ewloc) + fit:pop
## Model 3: resp ~ varb + layer:(nsloc + ewloc + pop)
##   Model Df Model Dev Df Deviance P(>|Chi|)
## 1       17   -2746.7
## 2       21   -2712.5  4   34.203 6.772e-07 ***
## 3       33   -2674.7 12   37.838 0.0001632 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Despite the largest model fitting the best, we choose the middle model because that one tells us something about fitness directly that the other one does not.

The argument for doing this is because we are interested in modeling fitness, and the distribution of fitness (actually best surrogate of fitness in their data) is not very different between the two models.

The distribution of other components of fitness (other than the final one) may differ quite a lot, but that was not the question of scientific interest.

So what do these models say about the distribution of fitness?

```
## we will go over this later
pop <- levels(redata$pop)
nind <- length(unique(redata$id))
nnode <- nlevels(redata$varb)
npop <- length(pop)
amat <- array(0, c(nind, nnode, npop))
amat.ind <- array(as.character(redata$pop),
  c(nind, nnode, npop))
amat.node <- array(as.character(redata$varb),
  c(nind, nnode, npop))
amat.fit <- grepl("hdct", amat.node)
amat.fit <- array(amat.fit,
  c(nind, nnode, npop))
amat.pop <- array(pop, c(npop, nnode, nind))
amat.pop <- aperm(amat.pop)
amat[amat.pop == amat.ind & amat.fit] <- 1
pout <- predict(aout,  varvar = varb, idvar = id,
  root = root, se.fit = TRUE, amat = amat)
pout.bigger <- predict(aout.bigger, varvar = varb,
  idvar = id, root = root, se.fit = TRUE, amat = amat)
```

The first interesting thing about these *predictions* (actually point estimates of parameters with standard errors) is that the point estimates are exactly the same for the two models.

```
pout$fit
```

```
## [1]  81 171 112  31 286 218 167
```

```
pout.bigger$fit
```

```
## [1]  81 171 112  31 286 218 167
```

```
all.equal(pout$fit, pout.bigger$fit)
```

```
## [1] TRUE
```

And why is that? These are submodel canonical statistics (components of $M^T y$). Thus by the observed-equals-expected property of exponential families their MLE are equal to their observed values and hence equal to each other.

So that is certainly not a reason to prefer one model to the other. If the estimated means are exactly the same how about estimated asymptotic variances?

The asymptotic variance matrix of these canonical statistics is actually diagonal for each model.

The reason is that different populations of origin have different individuals in the sample, and only individuals from one population contribute to estimating one of these canonical statistics.

Thus it is enough to look at the asymptotic standard errors (all the covariances are zero).

```
pout$se.fit
```

```
## [1] 13.617532 19.984170 16.267065  8.524453 25.968492 22.227096 19.884556
```

```
pout.bigger$se.fit
```

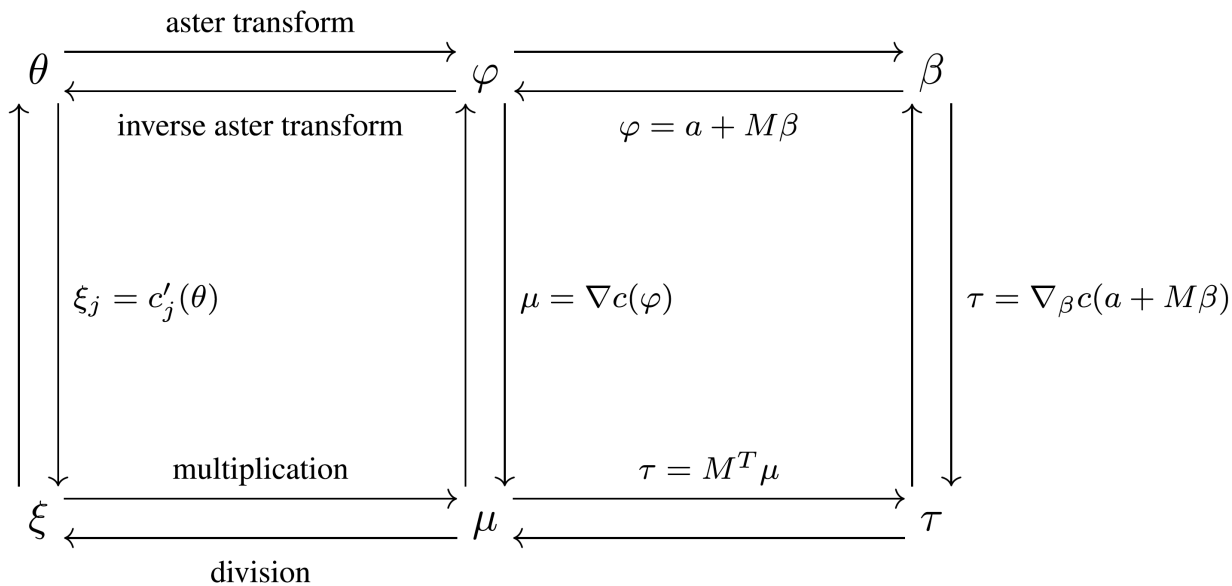
```
## [1] 14.521691 17.870387 14.513433  9.105173 27.857509 21.589790 21.642168
```

We see that they are not that different.

If we were interested in the effect of population on the different components of fitness, then the P-value 0.00016 does indicate that the model `aout.bigger` fits the data better.

The model `aout.bigger` has different population effects in different *layers* of the graph does show a statistically significant difference in the way the components of fitness combine to make up fitness in the various population of origin groups.

But if we are only interested in overall fitness rather than the separate components, then there is hardly any difference in the two models.


aster transform
θ
φ
β
inverse aster transform
$\varphi = a + M\beta$
$\xi_j = c'_j(\theta)$
$\mu = \nabla c(\varphi)$
$\tau = \nabla_\beta c(a + M\beta)$
multiplication
$\tau = M^T\mu$
ξ
μ
τ
division