One-way analysis of variance

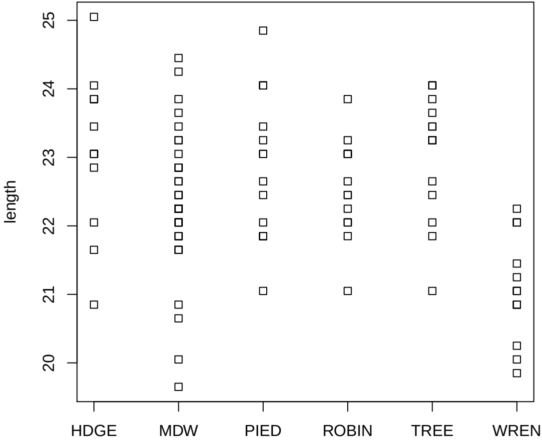
Math 463, Spring 2017, University of Oregon

David A. Levin

University of Oregon

May 8, 2017

```
> birds = read.table(
+     url("http://pages.uoregon.edu/dlevin/DATA/cuckoo.txt"),
+     header=T)
> stripchart(length~species, data=birds, vertical=T)
```



Model

Let $\boldsymbol{x}$ be the vector indicating variety.

- $$\mathbb{E}[\, Y_i \mid x_i] = \mu_j \quad \text{if } x_i = \text{species}\, j$$
- Define "dummy" variables δ_j for $j = 2, 3, \ldots, r$ (where r is the number of species).

$$\delta_{i,j} = \begin{cases} 1 & \text{if } i\text{th data point is from species } j \\ 0 & \text{otherwise} \end{cases} .$$

- Using these variables, we obtain linear model:

$$\mathbb{E}[\, Y_i \mid x_i] = \beta_0 + \sum_{j=2}^{r} \beta_j \delta_{i,j}$$

- Note the coefficient β_j is the difference $\mu_j - \mu_1$ between the expected response in species j and species 1.

Note that when a variable is a factor (a categorical variable), putting it into a linear model creates many dummy variables automatically:

```
> fit1 = lm(length~species, data=birds)
> model.matrix(fit1)[1:20,]
```

```
   (Intercept) speciesMDW speciesPIED speciesROBIN speciesTREE speciesWREN
1            1          1           0            0           0           0
2            1          1           0            0           0           0
3            1          1           0            0           0           0
4            1          1           0            0           0           0
5            1          1           0            0           0           0
6            1          1           0            0           0           0
7            1          1           0            0           0           0
8            1          1           0            0           0           0
9            1          1           0            0           0           0
10           1          1           0            0           0           0
11           1          1           0            0           0           0
12           1          1           0            0           0           0
13           1          1           0            0           0           0
14           1          1           0            0           0           0
15           1          1           0            0           0           0
16           1          1           0            0           0           0
17           1          1           0            0           0           0
18           1          1           0            0           0           0
19           1          1           0            0           0           0
20           1          1           0            0           0           0
```

To test the hypothesis that all the $\beta_2 = \cdots = \beta_r = 0$ (and so that all the species have the same mean), do a F-test comparing the sub-model

$$\mathbb{E}[\, Y_i \mid x_i] = \beta_0$$

to the "full" model above.

```
>   fit2 = lm(length~1, data=birds)
>   anova(fit2,fit1)
```

```
Analysis of Variance Table

Model 1: length ~ 1
Model 2: length ~ species
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1    119 137.188
2    114  94.248  5     42.94 10.388 3.152e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Power

- Note that to calculate power, one needs to compute, for $\boldsymbol{\theta} = \mathbb{E}[\,\boldsymbol{Y} \mid \boldsymbol{x}]$,

$$\|\Pi_W \boldsymbol{\theta}\|^2,$$

where $W = \mathcal{L}(\mathbf{1}, \delta_1, \ldots, \delta_e) \cap \mathcal{L}(\mathbf{1})^{\perp}$. (The non-centrality parameter of the F-statistic is $\|\Pi_W \boldsymbol{\theta}\|^2/\sigma^2$.)

- As seen before,

$$\|\Pi_W(\boldsymbol{\theta})\|^2 = \sum_{j=2}^{r} \beta_j^2 \|\delta_j^{\perp}\|^2 + 2 \sum_{2 \le j < k \le r} \beta_j \beta_k \langle \delta_j^{\perp}, \delta_k^{\perp} \rangle,$$

where $\boldsymbol{z}^{\perp} = \boldsymbol{z} - \bar{z}\mathbf{1}$ is the projection of $\boldsymbol{z}$ on the othogonal complement of $\mathcal{L}(\mathbf{1})$.