One-way ANOVA

- One-way ANOVA can be used to compare the means11 of several groups
- E.g. compare the effectiveness of different brands of hand sanitizers at killing certain types of bacteria
- Two-sample *t*-test: 2 groups
- ANOVA: 2 or more groups
- ANOVA is a also a special case of multiple linear regression

Coding categorical predictors

Consider a predictor x with k levels. This can be represented using $k-1$ indicator variables:

$$x_{ij} = \begin{cases} 1 & \text{if } X \text{ for the } i\text{th obervation is level } j \\ 0 & \text{otherwise} \end{cases}$$

Where is the final level?

Set $x_{ij} = 0$ for $j = 1, 2, \ldots, k-1$.
This gives us the final level (level k).

Predictors that are categorical variables are called factors.

Their categories are called levels.

Example 4.9

We wish to predict happiness based on the type of pet, using a regression. How should we code the X matrix?

Owner	Pet (X)	Happiness (Y)
A	Rabbit	10
B	Cat	7
C	Cat	9
D	Dog	6
E	Dog	2

$k = 3$ levels

We need $k-1 = 2$ indicator variables.

Let $\underset{\text{cat}}{x_{i1}} = \begin{cases} 1 & \text{if owner } i \text{ has cat} \\ 0 & \text{otherwise} \end{cases}$

$\underset{\text{dog}}{x_{i2}} = \begin{cases} 1 & \text{if owner } i \text{ has dog} \\ 0 & \text{otherwise} \end{cases}$

$$X = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}, \quad Y = \begin{bmatrix} 10 \\ 7 \\ 9 \\ 6 \\ 2 \end{bmatrix}$$

(R assigns indicators in alphabetical order.) $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$ — β_0: intercept; β_1: coefficient for cat; β_2: coefficient for dog

Example 4.9 Solution

```
stats <- tribble(
    ~owner, ~pet, ~happiness,
    “A", "Rabbit", 10,
    “B", "Cat", 7,
    “C", "Cat", 9,
    “D", "Dog", 6,
    “E", "Dog", 2
)
stats$pet <- factor(stats$pet)
library(modelr)
model_matrix(stats, happiness~pet)
```

```
## # A tibble: 5 x 3
##  `(Intercept)` petDog petRabbit
##          <dbl>  <dbl>     <dbl>
## 1            1      0         1
## 2            1      0         0
## 3            1      0         0
## 4            1      1         0
## 5            1      1         0
```

$X =$

One-way layout

Consider independent observations in k groups

		from
Sample 1:	$y_{11}, y_{12}, \dots, y_{1n_1}$	$N(\mu_1, \sigma^2)$
Sample 2:	$y_{21}, y_{22}, \dots, y_{2n_2}$	$N(\mu_2, \sigma^2)$
$\vdots$	$\vdots \qquad \vdots$	
Sample k	$y_{k1}, y_{k2}, \dots, y_{kn_k}$	$N(\mu_k, \sigma^2)$

with

$$Y_{ij} \sim N(\mu_i, \sigma^2) \text{ for } j = 1, 2, \dots, n_i; \mathrm{i} = 1, 2, \dots, k$$

H_0: $\mu_1 = \mu_2 = \dots = \mu_k$

H_a: at least one of the μ_i are different

Testing for the same mean

Suppose we want to test

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k.$$

How can we perform this using a linear model?

In the ANOVA formulation:

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \quad \text{where } \varepsilon_{ij} \sim \text{iid } N(0, \sigma^2)$$

Rewrite μ_i in terms of the overall mean of all k groups:

$$\mu = \frac{1}{k}\sum_{i=1}^{k} \mu_i$$

So $\mu_i = \mu + \alpha_i$ difference between μ_i and the overall mean μ

called the effect of treatment/group i

Why not just do pairwise comparison with t-test?

We could do pairwise t-tests with α level of significance and reject $H_0: \mu_1 = \mu_2 = \cdots = \mu_k$ if any of the t-tests reject its H_0.

If we do this, then our results will be based on $\binom{k}{2}$ different tests. Observe that

$$\alpha^* = P\left(\text{reject } H_0 \text{ at least once in any of the } \binom{k}{2} \text{ tests} \mid H_0 \text{ true}\right)$$

$$= 1 - P(\text{not reject } H_0 \text{ in any of the test} \mid H_0 \text{ true})$$

$$= 1 - (1-\alpha)^{\binom{k}{2}}$$

If $\alpha = 0.05$, $k = 3$, then $\alpha^* = 0.143$.

$k = 4$, $\alpha^* = 0.265$

$k = 5$, $\alpha^* = 0.401$

$k = 6$, $\alpha^* = 0.537$

So we run into a high chance of Type I error if we do pairwise comparisons.

But, we could do some corrections to adjust for the error level. For example, we can use Bonferroni correction:

Set new α to be $\frac{\alpha}{k}$, so that the overall error $\leq \alpha$.

However, since are using a lower α for each of the pairwise test, we are actually loosing power.

Write as a multiple regression model

This model may also be formulated as a multiple linear regression model by considering the model formulation

$$M\colon \mu_i = \mu + \alpha_i,$$

where μ denotes the overall mean and α_i is a parameter specific to group i.

We need to set $\underline{\alpha_1 = 0}$. Why?

$$\cancel{\mu} = \frac{1}{k}\sum_{i=1}^{k}\mu_i = \frac{1}{k}\sum_{i=1}^{k}(\mu+\alpha_i) = \cancel{\mu} + \frac{1}{k}\sum_{i=1}^{k}\alpha_i$$

$$\Rightarrow \quad 0 = \sum_{i=1}^{k}\alpha_i$$

There is a redundant parameter in this formulation, as α_i is a linear combination of the other α_i's.

So we set $\alpha_1 = 0$. This is equivalent to setting $\mu = \mu_1$.

Group 1: $y_{1j} = \mu_1 + \epsilon_{ij} = \mu \quad + \epsilon_{1j}$

Group 2: $y_{2j} = \mu_2 + \epsilon_{2j} = \mu + \alpha_2 \quad + \epsilon_{2j}$

Group 3: $y_{3j} = \mu_3 + \epsilon_{3j} = \mu \quad + \alpha_3 \quad + \epsilon_{2j}$

$\vdots$

Group k: $y_{kj} = \mu_k + \epsilon_{kj} = \mu \quad + \alpha_k + \epsilon_{kj}$

$$
\begin{array}{c}
\\ \text{Group 1} \\ \\ \\ \\ \text{Group 2} \\ \\ \\ \\ \\ \text{Group k} \\ \\ \\
\end{array}
\begin{bmatrix} y_{11}\\ y_{12}\\ \vdots\\ y_{1n_1}\\ y_{21}\\ y_{22}\\ \vdots\\ y_{2n_2}\\ \vdots\\ y_{k1}\\ y_{k2}\\ \vdots\\ y_{kn_k}\end{bmatrix}
=
\begin{array}{c}
\begin{array}{ccccc}\mu & \alpha_2 & \alpha_3 & \cdots & \alpha_k\end{array}\\
\begin{bmatrix}
1 & 0 & 0 & \cdots & 0\\
1 & 0 & 0 & \cdots & 0\\
\vdots & \vdots & \vdots & & \vdots\\
1 & 0 & 0 & \cdots & 0\\
1 & 1 & 0 & \cdots & 0\\
1 & 1 & 0 & \cdots & 0\\
\vdots & \vdots & \vdots & & \vdots\\
1 & 1 & 0 & \cdots & 0\\
\vdots & \vdots & \vdots & & \vdots\\
1 & 0 & 0 & \cdots & 1\\
1 & 0 & 0 & \cdots & 1\\
\vdots & \vdots & \vdots & & \vdots\\
1 & 0 & 0 & \cdots & 1
\end{bmatrix}
\end{array}
\begin{bmatrix}\mu\\ \alpha_2\\ \alpha_3\\ \vdots\\ \alpha_k\end{bmatrix}
+
\begin{bmatrix} \epsilon_{11}\\ \epsilon_{12}\\ \vdots\\ \epsilon_{1n_1}\\ \epsilon_{21}\\ \epsilon_{22}\\ \vdots\\ \epsilon_{2n_2}\\ \vdots\\ \epsilon_{k1}\\ \epsilon_{k2}\\ \vdots\\ \epsilon_{kn_k}\end{bmatrix}
$$

MLR model

$$\underset{(n\times 1)}{\boldsymbol{y}} = \begin{bmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ y_{22} \\ \vdots \\ y_{2n_2} \\ \vdots \\ y_{k1} \\ y_{k2} \\ \vdots \\ y_{kn_k} \end{bmatrix}, \underset{(n\times k)}{\boldsymbol{X}} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \cdots & 1 \\ 1 & 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \cdots & 1 \end{bmatrix}, \underset{(k\times 1)}{\boldsymbol{\beta}} = \begin{bmatrix} \mu \\ \alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_k \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2-\mu_1 \\ \mu_3-\mu_1 \\ \vdots \\ \mu_k-\mu_1 \end{bmatrix}$$

$$n = \sum_{i=1}^{k} n_i$$

New hypothesis

We can rewrite

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$$

as

$$H_0: \alpha_2 = \alpha_3 = \cdots = \alpha_k = 0$$

$$H_0: \quad \mu_1 = \mu_2 = \dots = \mu_k$$

$$\mu_1 - \mu_1 = \mu_2 - \mu_1 = \dots = \mu_k - \mu_1$$

$$0 = \alpha_2 = \dots = \alpha_k$$

Testing this H_0 can be done in the same way as the hypothesis test for several parameters in MLR. In our case, we have

full model: $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$

reduced model: $Y_{ij} = \mu + \varepsilon_{ij}$

ANOVA table

Source	SS	df	ms	F
Between Groups	SSG $=\sum_i n_i(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2$	$k-1$	$\frac{SSG}{k-1}$	$\frac{SSG}{SSE}\left(\frac{n-k}{k-1}\right)$
Within Groups	SSE $=\sum_{ij}(y_{ij} - \bar{y}_{i\bullet})^2$	$n-k$	$\frac{SSE}{n-k}$	
Total	SST $=\sum_{ij}(y_{ij} - \bar{y}_{\bullet\bullet})^2$	$n-1$		

where

$$n = \sum_{i=1}^{k} n_i,\ \bar{y}_{i.} = \frac{1}{n_i}\sum_{i=1}^{k} y_{ij} \text{ and } \bar{y}_{..} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}$$

The F-statistic

$$F = \frac{\sum_i n_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 \,/\, (k-1)}{\sum_{ij}\left(y_{ij} - \bar{y}_{i\cdot}\right)^2 / \,(n-k)},$$

and H_0 is rejected when $\underline{F \geq F_{k-1,n-k,\alpha}}$.