STA221

Neil Montgomery

Last edited: 2017-03-05 14:32

One factor notation, models

"Balanced" case with equal sample size n for each of k levels for $N = nk$ total.

Levels:	1	2	...	i	...	k
	y_{11}	y_{21}	$\ldots$	y_{i1}	$\ldots$	y_{k1}
	y_{12}	y_{22}	$\ldots$	y_{i2}	$\ldots$	y_{k2}
	$\vdots$	$\vdots$		$\vdots$		$\vdots$
	y_{1n}	y_{2n}	$\ldots$	y_{in}	$\ldots$	y_{kn}
Sample average:	$\overline{y}_1$	$\overline{y}_2$	$\ldots$	$\overline{y}_i$	$\ldots$	$\overline{y}_k$

Grand overall average: $\overline{\overline{y}}$

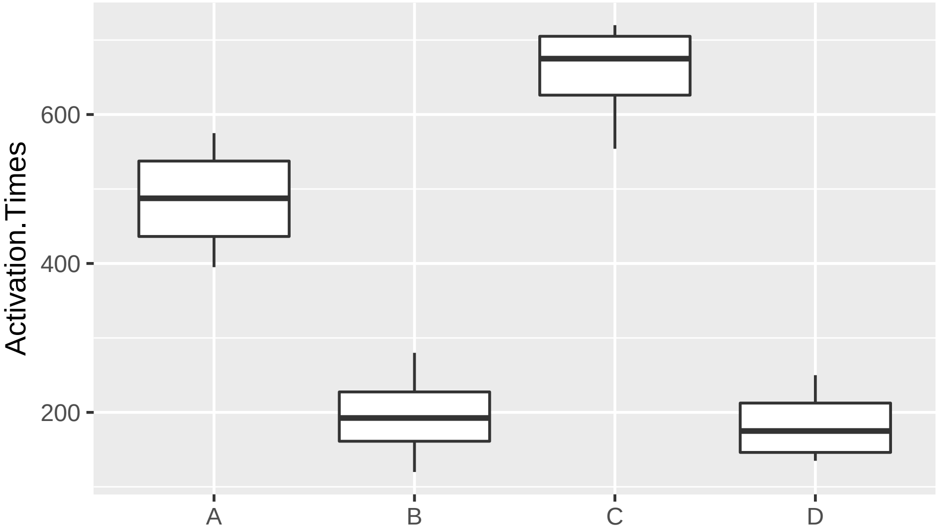
Models:

$$y_{ij} = \mu_i + \varepsilon_{ij}, \qquad \varepsilon_{ij} \text{ i.i.d. } N(0, \sigma^2)$$

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \qquad \sum \alpha_i = 0 \qquad \varepsilon_{ij} \text{ i.i.d. } N(0, \sigma^2)$$

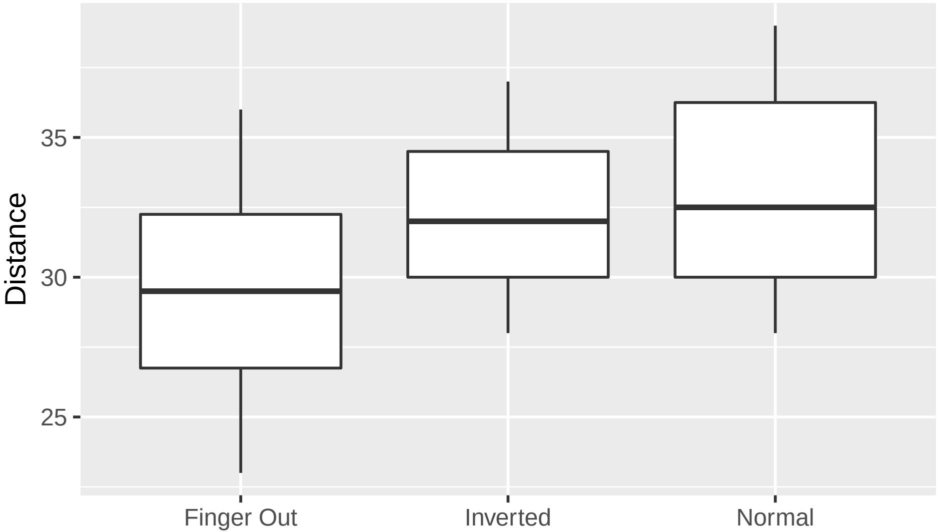
groups that are clearly different

From Q26.7 “Activating baking yeast”.



groups that aren't all that different

From Q26.8 "Frisbee throws".



The main question

The main question is $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ versus the negation (equivalently: all the $\alpha_i = 0$.)

In other words "is the variation among all the y_{ij} due to the factor variable, or just due to random chance?". The analysis even follows this logic.

The variation among the y_{ij} is quantified as:

$$(N-1) \cdot s_y^2 = \sum_{i=1}^{k} \sum_{j=1}^{n} \left(y_{ij} - \overline{\overline{y}}\right)^2$$

We will eventually split this up into the "factor" part and the "random chance" part (like done in regression).

some gory details

Build up from the inside out. For any i and j fixed:

$$(y_{ij} - \overline{\overline{y}})^2 = (y_{ij} - \bar{y}_i + \bar{y}_i - \overline{\overline{y}})^2$$

some gory details

Build up from the inside out. For any i and j fixed:

$$
\begin{aligned}
\left(y_{ij} - \overline{\overline{y}}\right)^2 &= \left(y_{ij} - \bar{y}_i + \bar{y}_i - \overline{\overline{y}}\right)^2 \\
&= \left(y_{ij} - \bar{y}_i\right)^2 + \left(\bar{y}_i - \overline{\overline{y}}\right)^2 + 2\left(y_{ij} - \bar{y}_i\right)\left(\bar{y}_i - \overline{\overline{y}}\right)
\end{aligned}
$$

some gory details

Build up from the inside out. For any i and j fixed:

$$\begin{aligned}\left(y_{ij} - \bar{\bar{y}}\right)^2 &= \left(y_{ij} - \bar{y}_i + \bar{y}_i - \bar{\bar{y}}\right)^2 \\ &= \left(y_{ij} - \bar{y}_i\right)^2 + \left(\bar{y}_i - \bar{\bar{y}}\right)^2 + 2\left(y_{ij} - \bar{y}_i\right)\left(\bar{y}_i - \bar{\bar{y}}\right)\end{aligned}$$

Next, for any fixed i, sum from $j = 1$ to n to get:

$$\sum_{j=1}^{n}\left(y_{ij} - \bar{\bar{y}}\right)^2 = \sum_{j=1}^{n}\left(y_{ij} - \bar{y}_i\right)^2 + \sum_{j=1}^{n}\left(\bar{y}_i - \bar{\bar{y}}\right)^2 + 2\left(\bar{y}_i - \bar{\bar{y}}\right)\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_i\right)$$

some gory details

Build up from the inside out. For any i and j fixed:

$$\begin{aligned}\left(y_{ij} - \overline{\overline{y}}\right)^2 &= \left(y_{ij} - \bar{y}_i + \bar{y}_i - \overline{\overline{y}}\right)^2 \\ &= \left(y_{ij} - \bar{y}_i\right)^2 + \left(\bar{y}_i - \overline{\overline{y}}\right)^2 + 2\left(y_{ij} - \bar{y}_i\right)\left(\bar{y}_i - \overline{\overline{y}}\right)\end{aligned}$$

Next, for any fixed i, sum from $j = 1$ to n to get:

$$\sum_{j=1}^{n} \left(y_{ij} - \overline{\overline{y}}\right)^2 = \sum_{j=1}^{n} \left(y_{ij} - \bar{y}_i\right)^2 + \sum_{j=1}^{n} \left(\bar{y}_i - \overline{\overline{y}}\right)^2 + 2\left(\bar{y}_i - \overline{\overline{y}}\right) \sum_{j=1}^{n} \left(y_{ij} - \bar{y}_i\right)$$

The term on the right hand side is always 0!

some gory details

Build up from the inside out. For any i and j fixed:

$$\begin{aligned}\left(y_{ij} - \bar{\bar{y}}\right)^2 &= \left(y_{ij} - \bar{y}_i + \bar{y}_i - \bar{\bar{y}}\right)^2 \\ &= \left(y_{ij} - \bar{y}_i\right)^2 + \left(\bar{y}_i - \bar{\bar{y}}\right)^2 + 2\left(y_{ij} - \bar{y}_i\right)\left(\bar{y}_i - \bar{\bar{y}}\right)\end{aligned}$$

Next, for any fixed i, sum from $j = 1$ to n to get:

$$\sum_{j=1}^{n}\left(y_{ij} - \bar{\bar{y}}\right)^2 = \sum_{j=1}^{n}\left(y_{ij} - \bar{y}_i\right)^2 + \sum_{j=1}^{n}\left(\bar{y}_i - \bar{\bar{y}}\right)^2 + 2\left(\bar{y}_i - \bar{\bar{y}}\right)\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_i\right)$$

The term on the right hand side is always 0!

Finally, sum from $i = 1$ to k and rearrange:

$$\sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij} - \bar{\bar{y}}\right)^2 = \sum_{i=1}^{k} n\left(\bar{y}_i - \bar{\bar{y}}\right)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_i\right)^2$$

more details

$$\sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij}-\overline{\overline{y}}\right)^2 = \sum_{i=1}^{k} n\left(\bar{y}_i-\overline{\overline{y}}\right)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij}-\bar{y}_i\right)^2$$

$$SS_{Total} \quad = \quad SS_T \quad + \quad SS_E$$

more details

$$\sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij}-\overline{\overline{y}}\right)^2 = \sum_{i=1}^{k} n\left(\bar{y}_i-\overline{\overline{y}}\right)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij}-\bar{y}_i\right)^2$$

$$SS_{Total} \quad = \quad SS_T \quad + \quad SS_E$$

Holding SS_{Total} fixed, what would it mean for one or the other of SS_T and SS_E to be large?

more details

$$\sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij}-\overline{\overline{y}}\right)^2 = \sum_{i=1}^{k} n\left(\bar{y}_i-\overline{\overline{y}}\right)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij}-\bar{y}_i\right)^2$$
$$SS_{Total} \quad = \quad SS_T \quad + \quad SS_E$$

Holding SS_{Total} fixed, what would it mean for one or the other of SS_T and SS_E to be large?

It turns out we'll look at a ratio of SS_T and SS_E to make our final decision.

more details

$$\sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij}-\overline{\overline{y}}\right)^2 = \sum_{i=1}^{k} n\left(\bar{y}_i-\overline{\overline{y}}\right)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij}-\bar{y}_i\right)^2$$

$$SS_{Total} \quad = \quad SS_T \quad + \quad SS_E$$

Holding SS_{Total} fixed, what would it mean for one or the other of SS_T and SS_E to be large?

It turns out we'll look at a ratio of SS_T and SS_E to make our final decision.

From which family of distributions will SS_T and SS_E come from?

the F distributions

Call (updated notation to match book):

$$MS_T = \frac{SS_T}{k-1} \qquad \text{and} \qquad MS_E = \frac{SS_E}{N-k}$$

the F distributions

Call (updated notation to match book):

$$MS_T = \frac{SS_T}{k-1} \qquad \text{and} \qquad MS_E = \frac{SS_E}{N-k}$$

These are called "mean squares", and the ratio of mean squares will follow what is called an F distribution, with $k-1$ and $N-k$ "degrees of freedom".

the F distributions

Call (updated notation to match book):

$$MS_T = \frac{SS_T}{k-1} \qquad \text{and} \qquad MS_E = \frac{SS_E}{N-k}$$

These are called "mean squares", and the ratio of mean squares will follow what is called an F distribution, with $k-1$ and $N-k$ "degrees of freedom".

When the null hypothesis is true, $\frac{MS_T}{MS_E}$ lives near 1, and large values of this ratio give small p-values.

putting it all together

All this information is concisely displayed in what is called the "analysis of variance" table (or ANOVA table, or AOV table). Here's the table for the Yeast example:

```
##             Df Sum Sq Mean Sq F value      Pr(>F)
## Recipe       3 638968  212989   44.74 0.000000864
## Residuals   12  57128    4761
```

putting it all together

All this information is concisely displayed in what is called the "analysis of variance" table (or ANOVA table, or AOV table). Here's the table for the Yeast example:

```
##              Df Sum Sq Mean Sq F value      Pr(>F)
## Recipe        3 638968  212989   44.74 0.000000864
## Residuals    12  57128    4761
```

And for the Frisbee example:

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Grip          2  58.58   29.29   2.045  0.154
## Residuals    21 300.75   14.32
```