

ANOVA III. Application to Categorical Predictors

Miaoyan Wang

Department of Statistics
UW Madison

Two representations of multiple samples comparison

- Data table may look like:

Factor	Treatment(Trt) 1	Trt 2	...	Trt k
Observation	y_{11}	y_{21}	\cdots	y_{k1}
	y_{12}	y_{22}	\cdots	y_{k2}
	\vdots	\vdots	\vdots	\vdots
	y_{1n_1}	y_{2n_2}	\cdots	y_{kn_k}
mean	$\bar{y}_{1\cdot}$	$\bar{y}_{2\cdot}$	\cdots	$\bar{y}_{k\cdot}$

- Or more often,

Trt	originally indexed y	re-indexed y'
1	y_{11}	$y'_{\mathbf{1}}$
\vdots	\vdots	\vdots
1	y_{1n_1}	$y'_{\mathbf{n_1}}$
\vdots	\vdots	\cdots
k	y_{k1}	$y'_{\mathbf{n_1+\dots+n_{k-1}+1}}$
\cdots	\cdots	\cdots
k	y_{kn_k}	$y'_{\mathbf{n_1+\dots+n_{k-1}+n_k}}$

Model Formulation

Consider a linear model with a K -level categorical predictor:

$$Y'_n = \beta_1 \mathbb{1}_{Trt1} + \cdots + \beta_k \mathbb{1}_{Trtk} + \varepsilon_n, \quad \text{for } n = 1, \dots, (n_1 + \dots + n_k).$$

Equivalently, let

$$Y_{ij} = \mu_i + \varepsilon_{ij},$$

where

- $j = 1, \dots, n_i$ indexes sample unit and $i = 1, \dots, k$ indexes treatment levels.
- Y_{ij} is the response variable of the j th unit receiving the i th level of treatment (TrT).
- μ_i is the population mean for the i th treatment level
- ε_{ij} is a random error for the j th sample unit and the i th treatment level. We assume $\varepsilon_{ij} \sim \text{iid } N(0, \sigma^2)$.
- The model parameters are $\mu_1, \dots, \mu_k, \sigma^2$ estimated by $\bar{Y}_{1.}, \dots, \bar{Y}_{k.}$, and MSE, respectively.

Alternative Model Formulation

Let

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad \text{with constraints} \quad \sum_{i=1}^k \alpha_i = 0$$

where

- $j = 1, \dots, n_i$ indexes sample unit and $i = 1, \dots, k$ indexes treatment levels.
- μ is the grand population mean $\mu = \frac{1}{k} \sum_{i=1}^k \mu_i$.
- $\alpha_i = \mu_i - \mu$ is the difference between the i th treatment mean and the grand mean (note: $\sum_{i=1}^k \alpha_i = 0$).
- The random errors are assumed to follow

$$\varepsilon_{ij} \sim \text{iid } N(0, \sigma^2).$$

- The model parameters are $\mu, \alpha_1, \dots, \alpha_k, \sigma^2$ estimated by $\bar{Y}_{..}, \bar{Y}_{1.} - \bar{Y}_{..}, \dots, \bar{Y}_{k.} - \bar{Y}_{..}$, and MSE, respectively.

Test for Overall Trt Effect

- A one-way ANOVA analysis begins with testing

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \text{ vs } H_A : \text{not all } \mu_i\text{'s are equal.}$$

- Or equivalently, test

$$H_0 : \alpha_i = 0 \text{ for all } i \text{ vs } H_A : \text{not all } \alpha_i\text{'s are zero.}$$

- The approach is to partition the total sum of squares, construct an ANOVA table, and perform an F test.
- Notation regarding the means:
 - ▶ $\bar{Y}_{..}$ is the grand mean,
 - ▶ $\bar{Y}_{i.}$ is the i th group (Trt) mean.

Partition of Total Sum of Squares

- Partition the total sum of squares (SS):

$$SST = SSR + SSE,$$

where

- ▶ Total SS:

$$SST = \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2 \text{ on df} = N - 1.$$

If balanced design: $\sum_{i=1}^k \sum_{j=1}^n (Y_{ij} - \bar{Y}_{..})^2$ on $\text{df} = kn - 1$.

- ▶ Between-Trt SS:

$$SSE = \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{Y}_{i.} - \bar{Y}_{..})^2 \text{ on df} = k - 1.$$

If balanced design: $n \sum_{i=1}^k (\bar{Y}_{i.} - \bar{Y}_{..})^2$ on $\text{df} = k - 1$.

- ▶ SS for error:

$$SSE = \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2 \text{ on df} = N - k.$$

If balanced design: $\sum_{i=1}^k \sum_{j=1}^n (Y_{ij} - \bar{Y}_{i.})^2$ on $\text{df} = k(n - 1)$.

- Recall that MSE is an estimate of the error variance σ^2 .

ANOVA Table

SS: Sum of squares

MS: Mean sum of squares

Within-group SS S_{Within} , Between-group SS $S_{Between}$

	SS	df	MS	E(MS)
Between Treatment SS (Model)	$S_{Between} = \sum_j n_j (y_{j.} - y_{..})^2$	$K - 1$	$S_{Between} / (K - 1)$	$\sigma^2 + (K - 1)^{-1} \sum_{j=1}^K n_j t_j^2$
Within Treatment SS (Error)	$S_{Within} = \sum_j \sum_{k=1}^{n_j} (y_{jk} - y_{j.})^2$	$n - K$	$S_{Within} / (n - K)$	σ^2
Total SS	$S_{Total} = \sum_j \sum_{k=1}^{n_j} (y_{jk} - y_{..})^2$	$n - 1$	$S_{Total} / (n - 1)$	

where $t_j = \mu_j - \mu$ and $\mu = n^{-1} \sum_j n_j \mu_j$. Q: simplify in a balanced design $n_1 = \dots = n_K$?

- In a general one-way ANOVA, under the $H_0 : \alpha_i = 0$ for all i ,

$$F = \frac{MSR}{MSE} \sim F_{K-1, N-K}.$$

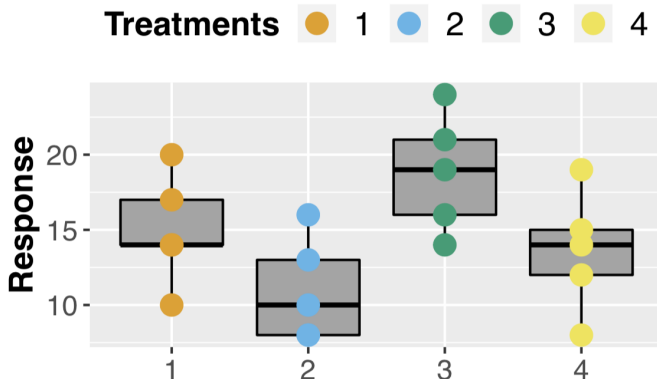
Example

```
library(ggplot2)
y <- c(14, 20, 10, 14, 17, 13, 8, 10, 16, 8, 16, 14, 24, 21, 19, 8, 14, 19, 12, 15)
trt <- factor(rep(1:4, each = 5))
cbind(y, trt)
```

```
##      y trt
## [1,] 14  1
## [2,] 20  1
## [3,] 10  1
## [4,] 14  1
## [5,] 17  1
## [6,] 13  2
## [7,]  8  2
## [8,] 10  2
## [9,] 16  2
## [10,]  8  2
## [11,] 16  3
## [12,] 14  3
## [13,] 24  3
## [14,] 21  3
## [15,] 19  3
## [16,]  8  4
## [17,] 14  4
## [18,] 19  4
## [19,] 12  4
## [20,] 15  4
```


Example

```
ggplot(data.frame(response=y, trt = trt), aes(x = trt, y = response)) +  
  geom_boxplot(aes(colour=factor(trt)), fill="#A4A4A4", color="black")+  
  geom_point(aes(color = factor(trt)), size = 4)+  
  scale_colour_manual(values = c("#E69F00", "#56B4E9", "#009E73", "#F0E442"))+  
  scale_fill_manual(values = c("#E69F00", "#56B4E9", "#009E73", "#F0E442"))+  
  labs(colour = "Treatments")+xlab("")+ylab("Response")+  
  theme(axis.text=element_text(size=12),  
        axis.title=element_text(size=14,face="bold"),  
        legend.text=element_text(size=14),  
        legend.title=element_text(size=14, face="bold"), legend.position = "top")
```



Example

```
anova(lm(y~trt))
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value  Pr(>F)
## trt         3  158.8   52.933   3.6506 0.03533 *
## Residuals  16  232.0   14.500
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Marginal evidence against the null hypothesis.

Note that

```
1-pf(3.6506 ,3, 16)
```

```
## [1] 0.03532877
```

```
pf(3.6506 ,3, 16, lower.tail = FALSE)
```

```
## [1] 0.03532877
```