5. ANOVA (Analysis of Variances) (anova)

- Some predictors are qualitative in nature, e.g. eye color
- Often described as categorical or factors
- Eye colors "blue" ⇒ Eye color is a factor
  "green" with 3 levels.
  "brown"

Data form

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

n_i for $i = 1 \cdots k$ is the total sample size within the level i and they can be the same or not.

② Stack observations into one column

Group	Original data	Re-indexed y'
1	y_{11}	y'_1
1	y_{12}	y'_2
$\vdots$	$\vdots$	$\vdots$
1	y_{1n_1}	y'_{n_1}
$\vdots$	$\vdots$	$\vdots$
k	y_{k1}	$y'_{n_1+n_2+\cdots+n_{k-1}+1}$
k	y_{k2}	$y'_{n_1+n_2+\cdots+n_{k-1}+\underline{2}}$
$\vdots$	$\vdots$	$\vdots$
k	y_{kn_k}	$y'_{n_1+n_2+\cdots+n_{k-1}+\underline{n_k}}$

Total # of observations / rows $n = \sum_{i=1}^{k} n_i$

Model formulation as regression

① $Y_{ij} = \mu_i + \epsilon_{ij}$ for $i = 1 \cdots k$ groups

(k) $j = 1 \cdots n_i$ sample units

μ_i : population mean for the i-th group

ϵ_{ij} : random errors for the j-th sample unit in the i-th group. $\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$

② An alternative form.

$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$ $(1 + k)$

△ $\mu = \frac{1}{K}\sum_{i=1}^{k} \mu_i$: grand population mean

$\alpha_i = \mu_i - \mu$: difference between i-th group mean and the grand mean.

△ The model has the constraint $\sum_{i=1}^{k} \alpha_i = 0$

(If no constraint of α's, parameters are not identifiable based on ② model only.

For example : $\alpha_i \to \alpha_i + 1$; $\mu \mapsto \mu - 1$

Qualitative predictors : factors

Regression parameters : effects

(Treated as fixed unknown parameters $\Rightarrow$ fixed-effect)

③ Stack all observations:

$$Y = \begin{pmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ \vdots \\ y_{k1} \\ \vdots \\ y_{kn_k} \end{pmatrix} = \underset{n\times k}{X} \; \underset{k\times 1}{\mu} + \underset{n\times 1}{\epsilon}$$

$\Downarrow$

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

$$X = \begin{pmatrix} 1_{n_1} & 0_{n_1} & \cdots & 0_{n_1} \\ 0_{n_2} & 1_{n_2} & \cdots & 0_{n_2} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{n_k} & 0_{n_k} & & 1_{n_k} \end{pmatrix} \qquad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{pmatrix} \qquad \epsilon = \begin{pmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1n_1} \\ \vdots \\ \epsilon_{k1} \\ \vdots \\ \epsilon_{kn_k} \end{pmatrix}$$

n rows $\times$ K columns

Because the model is written in a linear model form, assumptions on errors are satisfied. Conclusions we have derived can be applied: OLS, hypothesis test ...

[ANOVA 2]

△ One-way ANOVA (one factor as predictor)
Test

(i) $H_0: \mu_1 = \mu_2 = \cdots = \mu_K$ v.s. H_A: not all μ_i's are equal

(ii) $H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_K$ vs. H_A not all α_i's are equal

[2.1] By our discussion the general F-test

$$\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_K \end{pmatrix} \qquad H_0: A\mu = c$$

$$A = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 \\ 1 & 0 & -1 & \cdots & 0 \\ & & & & \\ 1 & 0 & 0 & \cdots & -1 \end{pmatrix} \qquad c = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$$

$$A\mu = \begin{pmatrix} \mu_1 - \mu_2 \\ \mu_1 - \mu_3 \\ \vdots \\ \mu_1 - \mu_{K-1} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}_{(k-1)\times 1}$$

We compare a full model ($\mu_1 \cdots \mu_K$) vs null model ($\mu_1 = \cdots = \mu_K$)

$$F_{stat} = \frac{(RSS_H - RSS_{Full}) / (df.H - df.Full)}{RSS_{Full} / df.Full}$$

$df._H = n-1$ $\qquad$ $df._{Full} = n-K$

[Step 1] Find RSS_{Full}

[Step 2] Find RSS_H under $H_0: \mu_1 = \mu_2 = \cdots = \mu_K$

[Step 1] Minimize Sum of squares $= \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \mu_i)^2$

$$\Rightarrow \hat{\mu}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} = \bar{Y}_{i\cdot}$$

$$\Rightarrow RSS_{Full} = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\cdot})^2$$

[Step 2] Under H_0 $\mu_1 = \mu_2 = \cdots = \mu_K = \mu_H$

Minimize Sum of squares $= \sum_{i=1}^{K} \sum_{j=1}^{n_i} (Y_{ij} - \mu_H)^2$

$$\Rightarrow \hat{\mu}_H = \frac{1}{n} \sum_{i=1}^{K} \sum_{j=1}^{n_i} Y_{ij} = \bar{Y}_{\cdot\cdot}$$

$$\Rightarrow RSS_H = \sum_{i=1}^{k} \sum_{j=1}^{n_j} (Y_{ij} - \hat{\mu}_H)^2$$

$$= \sum_{i=1}^{K} \sum_{j=1}^{n_j} (Y_{ij} - \bar{Y}_{\cdot\cdot})^2$$

$$F_{stat} = \frac{(RSS_H - RSS_{Full}) / (K-1)}{RSS_{Full} / (n-K)} \sim F_{K-1,\, n-K}$$

under $H_0: \mu_1 = \mu_2 = \cdots = \mu_K$

[2.2] Partition of Sum of squares

(1) Total sum of squares (Recall R^2. Notes Nov 2]

no covariate information

best estimation is all sample mean

$$SS_{Total} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2 \quad \text{with } df_H = n-1$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \hat{\mu}_H)^2$$

(2) Sum of squares of error

$$SS_{Error} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \hat{\mu}_i)^2$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2 \qquad df_F = n-K$$

(RSS from Full model regression)

We have $E\left(\frac{SS_{Error}}{n-k}\right) = \sigma^2$ (Notes Sep 28 . Page 3)

(3) Between groups / treatments sum of squares

$$SS_{Between} = SS_{Total} - SS_{Error}$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\hat{\mu}_i - \hat{\mu}_H)^2$$

$$= \sum_{i=1}^{k} \underline{\sum_{j=1}^{n_i}} (\underline{\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}})^2$$

does not depend on j

$$= \sum_{i=1}^{k} n_i (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$$

degrees of freedom $df._H - df._{Full} = K-1$

Lemma 1 Recall by Notes Oct 26. Page 3

$$\underbrace{\| Y - \hat{Y}_H \|^2}_{SS_{Total}} = \underbrace{\| Y - \hat{Y} \|^2}_{SS_{Error}} + \underbrace{\| \hat{Y} - \hat{Y}_H \|^2}_{SS_{Between}}$$

In summary, ANOVA table

Source	Sum of Squares (SS)	Degrees of freedom (df)
Between groups	$SS_{Between} = \sum_{i=1}^{k} n_i (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$	$K-1$
Within group (Error)	$SS_{within} = \sum_{i=1}^{k} \underline{\sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\cdot})^2}$ ~ variance of each group	$n-k$
Total	$SS_{Total} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{\cdot\cdot})^2$	$n-1$

F-test can be done based on the above table

Balanced design: all sample sizes are equal

$n_1 = n_2 = \cdots = n_K = n_B$

$n = n_B \times K$

$$SS_{Between} = \sum_{i=1}^{k} n_i (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$$

$$= n_B \times \sum_{i=1}^{k} (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$$

anova(lm.null, lm.full)

Example on One-Way ANOVA

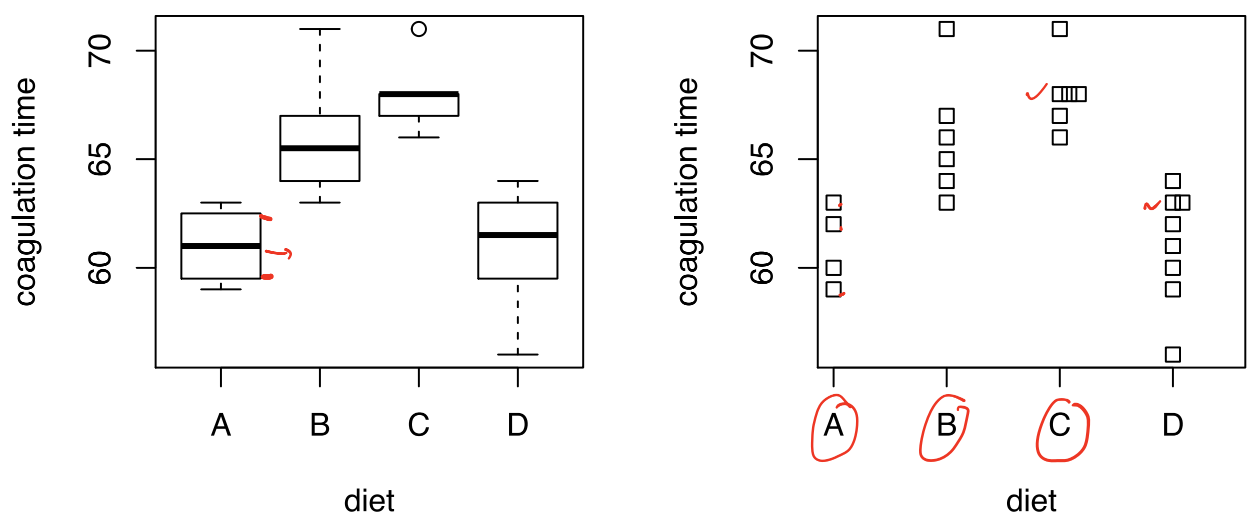
- 24 animals were randomly assigned to four different diets and
- The blood coagulation time was measured. Box et al. (1978).

```
library(faraway)
data(coagulation, package="faraway")
head(coagulation)
```

Y_{ij}

```
##   coag diet
## 1   62    A
## 2   60    A
## 3   63    A
## 4   59    A
## 5   63    B
## 6   67    B
```

```
par(mfrow=c(1,2))
plot(coag ~ diet, coagulation,ylab="coagulation time")
stripchart(coag ~ diet, coagulation, vertical=TRUE, method="stack",
           xlab="diet",ylab="coagulation time")
```



```r
par(mfrow=c(1,1))
```

- Left: boxplot.
- Right: stripchart. (1-dim scatterplot, an alternative to boxplots when sample sizes are small.)
- Median and upper quartile of diet C are the same.
- There are ties in diets C and D.

ANOVA code version 1 $\mu_1 I(diet=A) + \mu_2 I(diet=B) + \mu_3 I(diet=C) + \mu_4 I(diet=D)$

Full

```r
lmodi <- lm(coag ~ diet -1, coagulation)
summary(lmodi)$coefficients
```

no intercept

```
##       Estimate Std. Error  t value     Pr(>|t|)
## dietA       61  1.1832160 51.55441 9.547815e-23
## dietB       66  0.9660918 68.31649 3.532325e-25
## dietC       68  0.9660918 70.38669 1.948886e-25
## dietD       61  0.8366600 72.90895 9.663048e-26
```

μ_1, μ_2, μ_3, μ_4

Null

```r
lmnull <- lm(coag ~ 1, coagulation)
anova(lmnull,lmodi)
```

$n=24$, $K=4$

SS_{Total}	$n-1$		
SS_{Within}	$n-k$	$k-1$	$SS_{Between}$

```
## Analysis of Variance Table
##
## Model 1: coag ~ 1
## Model 2: coag ~ diet - 1
##   Res.Df RSS Df Sum of Sq      F    Pr(>F)
## 1     23 340
## 2     20 112  3       228 13.571 4.658e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

SS_{Total}, $n-K$, SS_{Within}, $SS_{Between}$

- We see that there is indeed a difference in the levels.

interept

$\mu_1 + (\mu_2-\mu_1)\ I(\text{diet B}) + (\mu_3-\mu_1)\ I(\text{die C}) + (\mu_4-\mu_1)\ I(\text{diet D})$

ANOVA code version 2

```
lmod <- lm(coag ~ diet, coagulation)
summary(lmod)$coefficients
```

```
##                  Estimate Std. Error      t value     Pr(>|t|)
## (Intercept) 6.100000e+01   1.183216 5.155441e+01 9.547815e-23
## dietB       5.000000e+00   1.527525 3.273268e+00 3.802505e-03
## dietC       7.000000e+00   1.527525 4.582576e+00 1.805132e-04
## dietD       2.991428e-15   1.449138 2.064281e-15 1.000000e+00
```

μ_1

$\mu_2-\mu_1$

$\mu_3-\mu_1$

$\mu_4-\mu_1$

```
anova(lmod)
```

```
## Analysis of Variance Table
##
## Response: coag
##           Df Sum Sq Mean Sq F value    Pr(>F)
## diet       3    228    76.0  13.571 4.658e-05 ***
## Residuals 20    112     5.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$\text{Mean Sq} = \frac{\text{Sum Sq}}{\text{Df}}$

k-1

n-k

SS within

Df	Sum Sq
k-1	$SS_{Between}$
n-k	SS_{Within}

Note

```
anova(lmnull, lmod) #This is also ok.
```

```
## Analysis of Variance Table
##
## Model 1: coag ~ 1
## Model 2: coag ~ diet
##   Res.Df RSS Df Sum of Sq      F    Pr(>F)
## 1     23 340
## 2     20 112  3       228 13.571 4.658e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

model with -1

```
anova(lmodi) #This is incorrect

## Analysis of Variance Table
##
## Response: coag
##           Df Sum Sq Mean Sq F value    Pr(>F)
## diet       4  98532 24633.0  4398.8 < 2.2e-16 ***
## Residuals 20    112     5.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```