One-way layout and ANOVA: An Example

one factor exp't (qualitative or quantitative)

Reflectance data in pulp experiment: each of four operators made five pulp sheets; reflectance was read for each sheet using a brightness tester.

Randomization : assignment of 20 containers of pulp to operators and order of reading.

* response : reflectance
* (treatment) factor : operator (qualitative)
  4 levels – A, B, C, D
* Exp'tal unit : container of pulp
  20 EUs

Q: What is the source of variation in the 5 observations?

Table 1: Reflectance Data, Pulp Experiment

Operator A	Operator B	Operator C	Operator D
59.8	59.8	60.7	61.0
60.0	60.2	60.7	60.8
60.8	60.4	60.5	60.6
60.8	59.9	60.9	60.5
59.8	60.0	60.3	60.5

Initial Data Analysis : Box plots

One-Way ANOVA

cf.

homogeneous → EU1 · · · EU20

completely randomized → r ABCD r

20! / 5!5!5!5!

* Each treatment repeats 5 times ← replicates

What if measure 5 points on a sheet? cf.

Objective : determine if there are "differences" among operators in making sheets and reading brightness.

μ_i : mean of ith group

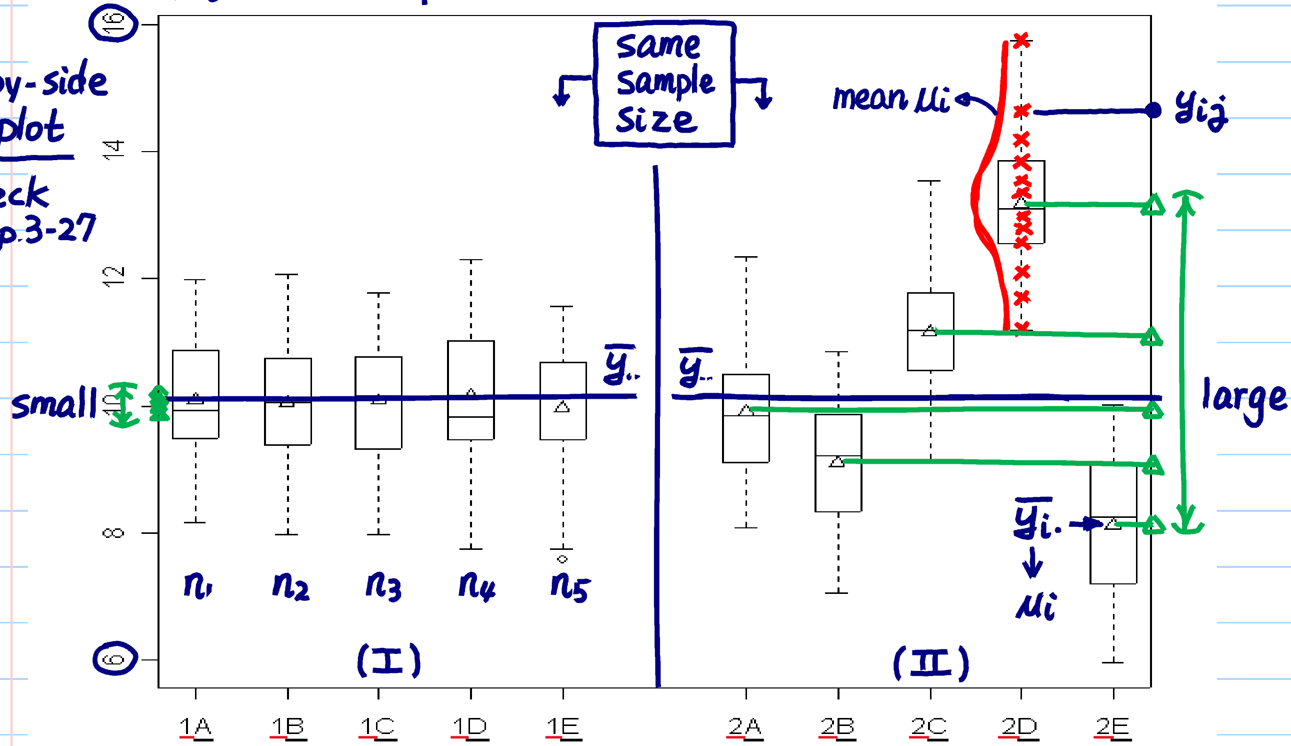
Interested in whether μ_i's are different, i.e., whether $\mu_1=\mu_2=\cdots=\mu_5$ holds.

$\sum_i n_i(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2$ — Why?

large ⇒ significant mean diff.

Small ⇒ insignificant mean diff.

side-by-side box plot
check LNp.3-27



between group variation → $\sum_i n_i(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2/df_1$

within group variation → $\sum_{i,j}(y_{ij}-\bar{y}_{i\cdot})^2/df_2$

large ⇒ significant mean diff.

small ⇒ insignificant mean diff.

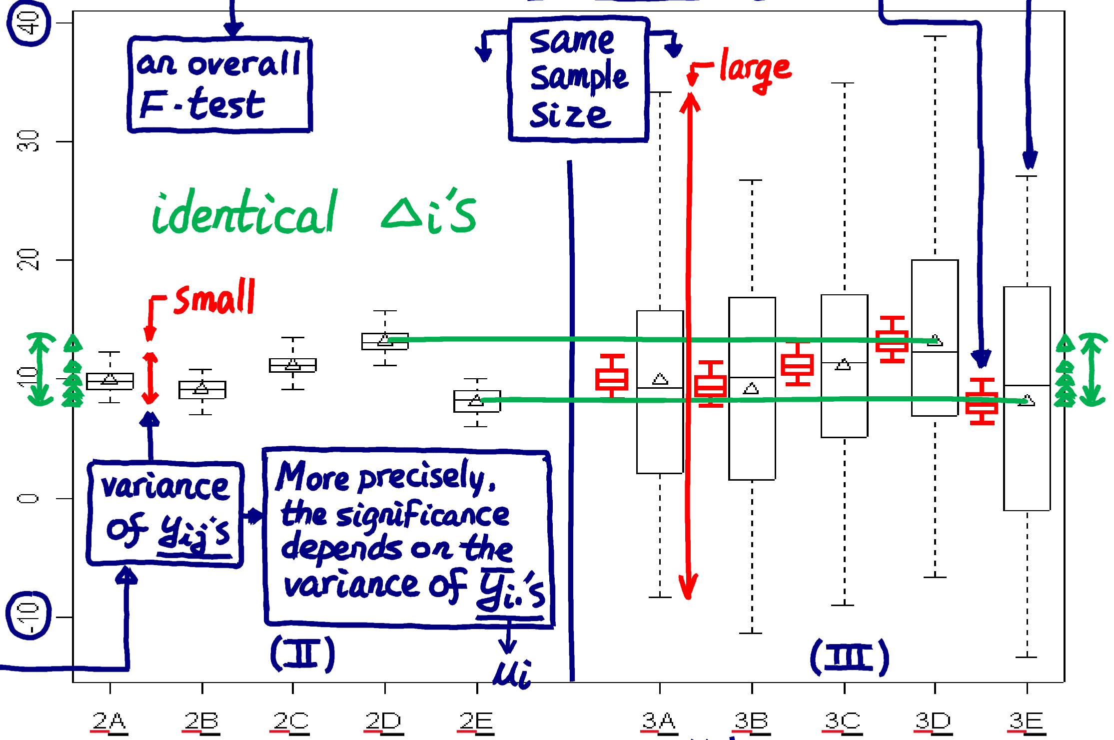
sampling dist. of $\bar{y}_{i\cdot}$

population dist. of y_{ij}

an overall F-test

Alternative issue: sample size

If II has a small sample size, but III has a large sample size, we might get insignificant in II, but significant in III



cf. population distribution of data (e.g., $Y_1,\cdots,Y_n \overset{i.i.d.}{\sim} F$)

sampling distribution of $\hat\theta$ (e.g., $\hat\theta \xrightarrow{d}$ normal, $\hat\theta \xrightarrow{P} \theta$, $Var(\hat\theta)\to 0$)

Model and ANOVA

$H_0: \mu_1=\mu_2=\cdots=\mu_k \Rightarrow X_\omega=\mathbf{1}$

or $\tau_1=\tau_2=\cdots=\tau_k \quad X_\Omega=(*)$

If $n_1=n_2=\cdots=n_k$ (balanced one-way) ⇒ orthogonality under sum coding

Model :

$$y_{ij}=\eta+\tau_i+\epsilon_{ij},\quad i=1,\ldots,k;\ j=1,\ldots,n_i,$$

μ_i ; $\eta = \frac{\mu_1+\cdots+\mu_k}{k}$; $\tau_i = \mu_i-\eta$; $N\equiv n_1+\cdots+n_k$

where y_{ij} = jth observation with treatment i,

τ_i = ith treatment effect,

ϵ_{ij} = error, independent $N(0,\sigma^2)$. $= y_{ij}-\mu_i$

Recall conceptual model (LNp.1-2) for one qualitative factor

$y_{ij} \overset{indep.}{\sim} N(\mu_i,\sigma^2)$

# of levels → k parameters

$d_1, d_2, \cdots, d_k$: known functions of the factor

$$\underbrace{\begin{bmatrix}\mu_1\\ \vdots\\ \mu_1\\ \mu_2\\ \vdots\\ \mu_2\\ \vdots\\ \mu_k\\ \vdots\\ \mu_k\end{bmatrix}}_{=\mu} \leftarrow \underbrace{\begin{bmatrix}y_{11}\\ \vdots\\ y_{1n_1}\\ y_{21}\\ \vdots\\ y_{2n_2}\\ \vdots\\ y_{k1}\\ \vdots\\ y_{kn_k}\end{bmatrix}}_{=Y} = \eta\underbrace{\begin{bmatrix}1\\ \vdots\\ 1\\ 1\\ \vdots\\ 1\\ \vdots\\ 1\\ \vdots\\ 1\end{bmatrix}}_{=d_0} + \tau_1\underbrace{\begin{bmatrix}1\\ \vdots\\ 1\\ 0\\ \vdots\\ 0\\ \vdots\\ 0\\ \vdots\\ 0\end{bmatrix}}_{=d_1} + \tau_2\underbrace{\begin{bmatrix}0\\ \vdots\\ 0\\ 1\\ \vdots\\ 1\\ \vdots\\ 0\\ \vdots\\ 0\end{bmatrix}}_{=d_2} + \cdots + \tau_k\underbrace{\begin{bmatrix}0\\ \vdots\\ 0\\ 0\\ \vdots\\ 0\\ \vdots\\ 1\\ \vdots\\ 1\end{bmatrix}}_{=d_k} + [\epsilon_{ij}]$$

$$d_0 = d_1 + d_2 + \cdots + d_k$$

regression expression

$Y = X\beta + \epsilon$

$X = [d_0\ d_1\ d_2 \cdots d_k]$ — (*)

⇒ k+1 parameters in β

⇒ over-parameterized

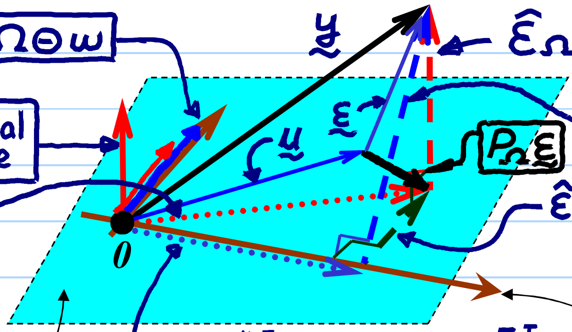
⇒ unidentifiable (LM, LNp.5-10~11)

$\mu = \eta d_0 + \tau_1 d_1 + \cdots + \tau_k d_k$

$\hat{\underline{y}}_\Omega \xrightarrow{\text{estimate}} \underline{\mu}$

$W_1 = \Omega \ominus \omega$

residual space



$\hat{\varepsilon}_\Omega = \underline{y} - \bar{\underline{y}}_{i\cdot} \equiv \underline{c}$

$\hat{\varepsilon}_\omega = \underline{y} - \bar{y}_{\cdot\cdot}\mathbb{1} \equiv \underline{a}$

$\hat{\varepsilon}_\omega - \hat{\varepsilon}_\Omega = \hat{\underline{y}}_\Omega - \hat{\underline{y}}_\omega = \bar{\underline{y}}_{i\cdot} - \bar{y}_{\cdot\cdot}\mathbb{1} \equiv \underline{b}$

$P_\Omega \underline{y} = \hat{\underline{y}}_\Omega = \begin{bmatrix} \bar{y}_{1\cdot} \\ \vdots \\ \bar{y}_{1\cdot} \\ \vdots \\ \bar{y}_{k\cdot} \\ \vdots \\ \bar{y}_{k\cdot} \end{bmatrix} \equiv \bar{\underline{y}}_{i\cdot}$

$P_\Omega = X(X^TX)^{-1}X^T$

where

$X = [\underline{d}_1\ \underline{d}_2 \cdots \underline{d}_k]$

(exercise) $P_\Omega = ?$

Ω (dim = k)

$\hat{\underline{y}}_\omega = \frac{\underline{y}^T\mathbb{1}}{\|\mathbb{1}\|^2}\mathbb{1} = \bar{y}_{\cdot\cdot}\mathbb{1} = \frac{\bar{\underline{y}}_{i\cdot}^T\mathbb{1}}{\|\mathbb{1}\|^2}\mathbb{1}$

ω (dim = 1) ← span{$\mathbb{1}$}

check whether $\mathbb{1} \perp \underline{b}$, $\underline{b} \perp \underline{c}$, $\mathbb{1} \perp \underline{c}$

Model fit:

$$\begin{aligned} y_{ij} &= \hat{\eta} + \hat{\tau}_i + r_{ij} \\ &= \bar{y}_{\cdot\cdot} + (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}) + (y_{ij} - \bar{y}_{i\cdot}), \end{aligned}$$

exercise

estimate of $\frac{\mu_1 + \cdots + \mu_k}{k} = \eta$

estimate of $\mu_i - \eta$

where "." means average over the particular subscript.

ANOVA Decomposition :

residuals under ω

$$\overbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{\cdot\cdot})^2}^{\|\underline{a}\|^2} = \overbrace{\sum_{i=1}^{k} n_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2}^{\|\underline{b}\|^2} + \overbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2}^{\|\underline{c}\|^2}.$$

residuals under Ω

overall F-test
$\frac{RSS_\omega - RSS_\Omega / df_\omega - df_\Omega}{RSS_\Omega / df_\Omega}$

	RSS_ω	$RSS_\omega - RSS_\Omega = \|P_{\Omega\ominus\omega}(\underline{y})\|^2$	RSS_Ω
(source of variation) →	CTSS	RegrSS	RSS
(variance decomposition) →	Var(Y)	Var[E(Y|X)]	E[Var(Y|X)]