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

cf.

homogeneous → EU1 · · · EU20

completely randomized → r → A B C D ← r

20! / 5!5!5!5!

★ Each treatment repeats 5 times ← replicates

What if measure 5 points on a sheet? cf.

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

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

Initial Data Analysis : Box plots

One-Way ANOVA

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

μ_i : mean of ith group

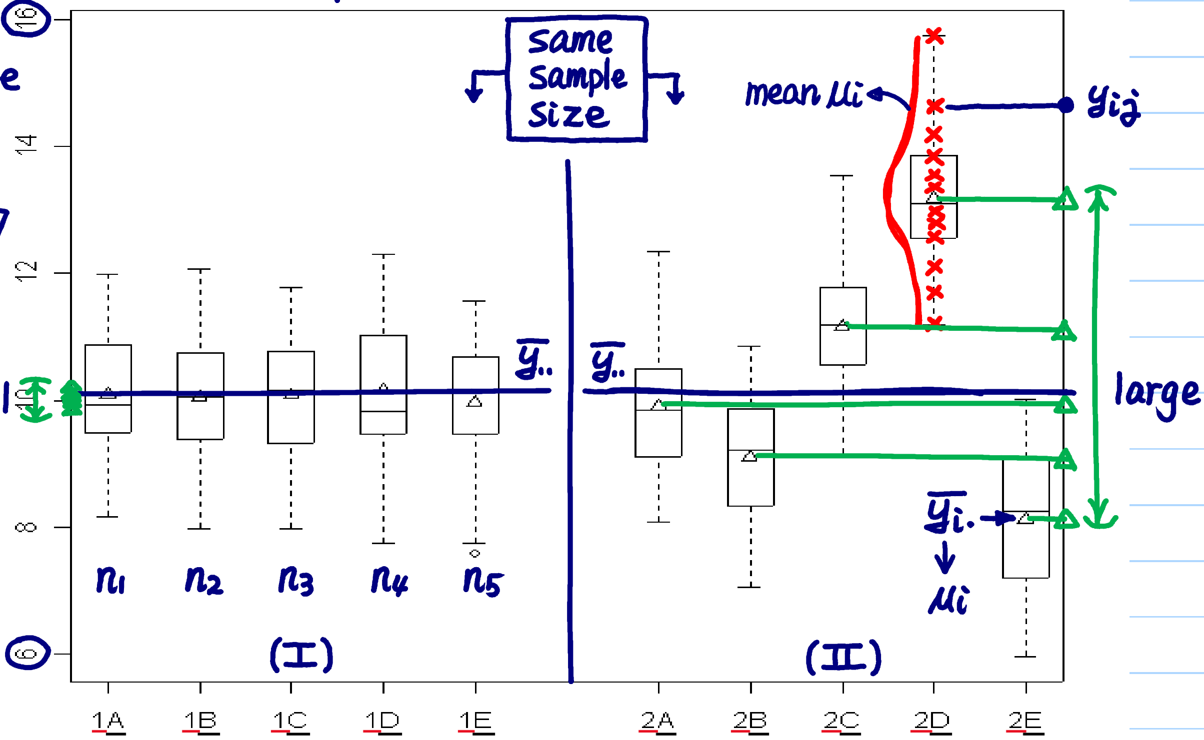
Interested in whether μ_i's are different, i.e., whether $\mu_1=\mu_2=\cdots=\mu_5$ holds.

$\sum_i n_i(\bar{y}_{i.}-\bar{y}_{..})^2$ — large ⇒ significant mean diff. — Small ⇒ insignificant mean diff.

Why?

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}

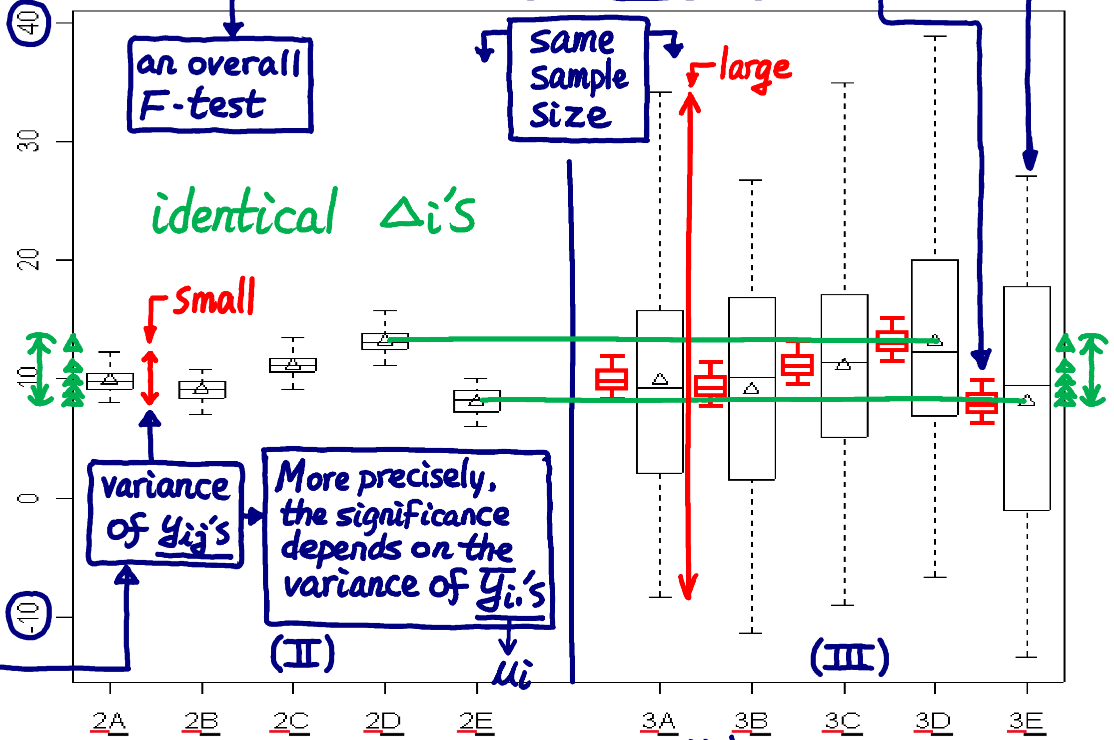
an overall F-test

same sample size

large

small

identical Δ_i's



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.

variance of y_{ij}'s

More precisely, the significance depends on the variance of $\bar{y}_{i\cdot}$'s

μ_i

(II) (III)

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

μ_i

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

$\eta=\dfrac{\mu_1+\cdots+\mu_k}{k}$ $\quad \tau_i=\mu_i-\eta$ $\quad N=n_1+\cdots+n_k$

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 (LN p.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

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, LN p.5-10~11)

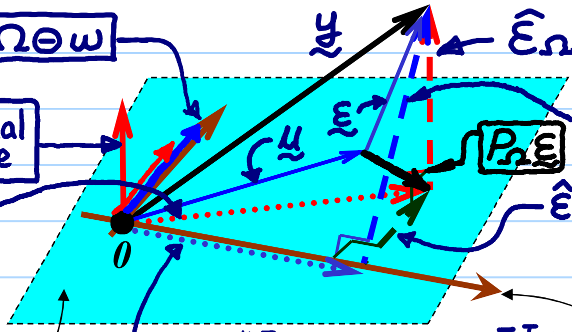
$$\begin{bmatrix}\mu_1\\ \vdots\\ \mu_1\\ \mu_2\\ \vdots\\ \mu_2\\ \vdots\\ \mu_k\\ \vdots\\ \mu_k\end{bmatrix} \leftarrow \begin{bmatrix}y_{11}\\ \vdots\\ y_{1n_1}\\ y_{21}\\ \vdots\\ y_{2n_2}\\ \vdots\\ y_{k1}\\ \vdots\\ y_{kn_k}\end{bmatrix} = \eta\begin{bmatrix}1\\ \vdots\\ 1\\ 1\\ \vdots\\ 1\\ \vdots\\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

μ, Y, $d_0 = d_1 + d_2 + \cdots + d_k$

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

$\boxed{\hat{\underline{y}}_\Omega \xrightarrow{\text{estimate}} \underline{\mu}}$ $\boxed{W_1 = \Omega \ominus \omega}$

residual space



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

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

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

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

$P_\Omega = X(X^TX)^{-1}X^T$ where $X = [d_1\ d_2 \cdots d_k]$

(exercise) $P_\Omega = ?$

Ω (dim = k)

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

ω (dim = 1) ← span{$\underline{1}$}

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

Model fit:

$$y_{ij} = \hat\eta + \hat\tau_i + r_{ij}$$
$$= \bar y_{\cdot\cdot} + (\bar y_{i\cdot} - \bar y_{\cdot\cdot}) + (y_{ij} - \bar y_{i\cdot}),$$

$\|\underline{y}\|^2 - \|\bar y_{\cdot\cdot}\underline{1}\|^2 = \|\hat\varepsilon_\omega\|^2 = \|\underline{a}\|^2$

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

estimate of $\mu_i - \eta$

exercise

where "." means average over the particular subscript.

ANOVA Decomposition :

residuals under ω

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

residuals under Ω

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

(source of variation) →	CTSS	RegrSS	RSS
(variance decomposition) →	Var(Y)	Var[E(Y|X)]	E[Var(Y|X)]