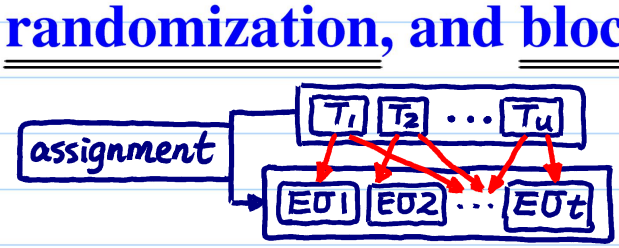
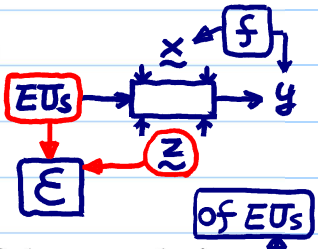
Fundamental Principles : Replication, randomization, and blocking

multiple observations (y) per treatment

Replication





- Each treatment is applied to units that are representative of the population (example : measurements of 3 units vs. 3 repeated measurements of 1 unit).
  - Q: What are the "sources of variation" in ε ?
- Replication vs Repetition (i.e., repeated measurements).
- Enable the estimation of experimental error. Use sample standard deviation.
  - "true" error variance (Recall. test for lack of fit , LM, LNp 6-6~11)
- Decrease variance of estimates and increase the power to detect significant differences : for independent y_i's,

$\bar{y}$ → estimator of $\mu_{T_j} = E(y_{T_j})$

$$Var\left(\frac{1}{N}\sum_{i=1}^{N} y_i\right) = \frac{1}{N} Var(y_1).$$

Var(ε)

# of replicates

Replicates and Experimental Errors

exp'ters often resist to perform replicates Why ?

replicate: replication of same treatment

$$y = f(X_1, X_2, \ldots, X_m) + \epsilon$$
$$= \hat{f}(X_1, X_2, \ldots, X_m) + \hat{\epsilon}$$

lack of fit

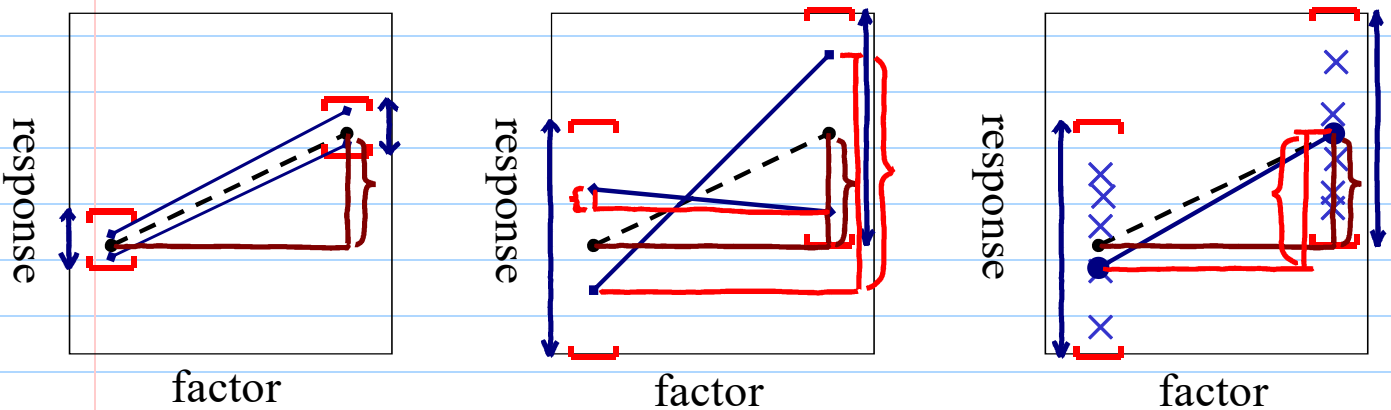
over-fitting

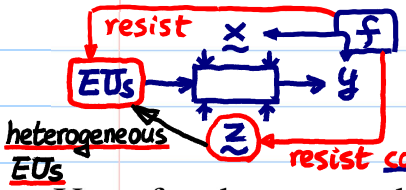
$\frac{\hat{\beta}}{s.e.(\hat{\beta})} \propto \frac{1}{\sqrt{n}} \frac{\hat{\sigma}}{\text{var est'or}}$

sample size

- Q: Why do we need to understand the magnitude of exp'tal error?

We need to know Var(ε) so that we can judge whether an effect is (statistically) significant relative to the error.





Randomization

for uncontrollable, unknown, nonmeasurable z, EUs

Use of a chance mechanism (e.g., random number generators) to assign treatments to units or to run order. It has the following advantages. ↑ LNp.1-17

- Protect against latent variables or "lurking" variables (give an example).
- Reduce influence of subjective bias in treatment assignments (e.g., clinical trials).
  ↳ single-blind, double-blind, ...
- Ensure validity of statistical inference (This is more technical; will not be discussed in the book. See Chapter 4 of "*Statistics for Experimenters*" by Box, Hunter, Hunter for discussion on randomization distribution.)

population → without replacement → EUs; ss with → i.i.d. → ε

true model: $Y = X\beta + Z\gamma + \varepsilon$

fitted model: $Y = X\beta + \varepsilon'$

$Z\gamma = HZ\gamma + (I-H)Z\gamma$

bias $\hat{\beta} \Rightarrow E(\hat{\beta}) = \beta + (X^TX)^{-1}X^TZ\gamma$

$X(X^TX)^{-1}X^T$ ← hat matrix

If $X \perp Z$ ($X^TZ = 0$) $\Rightarrow HZ\gamma = 0$ $\Rightarrow E(\hat{\beta}) = \beta$

Effect Aliasing/Confounding ← collinearity

混淆(污染)

design matrix / planning matrix

confounded

run order	A	B	C	Operator
1	low	low	low	Peter
2	low	low	high	Peter
3	low	high	low	Peter
4	low	high	high	Peter
5	high	low	low	John
6	high	low	high	John
7	high	high	low	John
8	high	high	high	John

aliased

	A	B	C	AB
2	low−	low−	high+	high+
3	low−	high+	low−	low−
5	high+	low−	low−	low−
8	high+	high+	high+	high+

= A×B

Q: what if operators have an effect on response?

- Q: Is aliasing/confounding always a bad thing?
  - pros & cons

Randomization

But we don't know

randomize run order

Q: what if operators have an effect on response?

confounded

slightly confounded

	A	B	C	operator	unknown Z
1	low	low	low	Peter	EU_{11}
2	low	low	high	Peter	EU_{12}
3	low	high	low	Peter	EU_{13}
4	low	high	high	Peter	EU_{14}
5	high	low	low	John	EU_{21}
6	high	low	high	John	EU_{22}
7	high	high	low	John	EU_{23}
8	high	high	high	John	EU_{24}

	A	B	C	operator
5	high	low	low	Peter
2	low	low	high	Peter
8	high	high	high	Peter
4	low	high	high	Peter
3	low	high	low	John
1	low	low	low	John
6	high	high	low	John
7	high	high	high	John

run order

EUs are heterogeneous, but unknown to the exp'ters. Exp'ters still assume EUs are homogeneous.

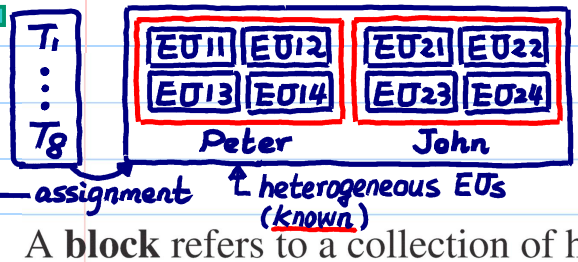
Randomization provides protection against *extraneous factors* that are unknown to the experimenter, but may impact the response

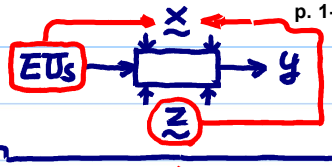
like firewall, immune system

- what should be randomized?
  - ➢ allocation of exp'tal materials to treatments; the order of applying treatments; the order of measuring responses; ...

Blocking

for known Z





fitted model: $Y = X\beta + Z\gamma + \varepsilon$ cf. $Y = X\beta + \varepsilon'$ (LNp.1-19)

① $Var(\varepsilon') \geq Var(\varepsilon)$
② $\hat{\beta}$ not biased by $Z\gamma$

A **block** refers to a collection of homogeneous units. Effective blocking : larger between-block variations than within-block variations.
(Examples: hours, batches, lots, street blocks, pairs of twins.)

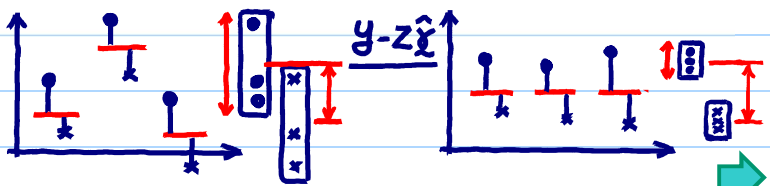
can try to achieve $X \perp Z$ by design ⇒ orthogonality

- Run and compare treatments within the same blocks. (Use randomization within blocks.) It can eliminate block-block variation and reduce variability of treatment effects estimates.

∵ $Z\gamma$ is put in the systematic part of the fitted model

- **Block what you can and randomize what you cannot.**

- Discuss **typing experiment** to demonstrate possible elaboration of the blocking idea. See LNp.1-24



Blocking

homogeneous EUs or same z value

a known z

Q: If operator effect is identified as significant before exp't, what can we do?

	A	B	C	operator	
1	low	low	low	Peter	EU11
2	low	low	high	John	EU21
3	low	high	low	John	EU22
4	low	high	high	Peter	EU12
5	high	low	low	John	EU23
6	high	low	high	Peter	EU13
7	high	high	low	Peter	EU14
8	high	high	high	John	EU24

Orthogonality

Known $\underset{\sim}{z}$ or systematic structure in EUs

block factor : factors that are controllable and may influence the response but in which we are ***not directly interested***. (cf. treatment factors) — $\underset{\sim}{x}$

& usually no interaction with $\underset{\sim}{x}$

Examples of blocking factors:

lot-to-lot, brand-to-brand, operator-to-operator, day-to-day, ...

block what you can & randomize what you cannot

Illustration: Typing Experiment

- To compare two keyboards A and B in terms of typing efficiency. Six manuscripts 1-6 are given to the same typist.
- Several designs (i.e., orders of test sequence) are considered:

1. ★ extra variable (factor or unknown): tying order
   2 levels: I, II, → block factor.

   1. A,B (I II), 2. A,B (I II), 3. A,B (I II), 4. A,B (I II), 5. A,B (I II), 6. A,B (I II)

   (A always followed by B, why bad ?)

2. Randomizing the order leads to a new sequence like this

   1. A,B (I), 2. B,A (II), 3. A,B (I), 4. B,A (II), 5. A,B (I), 6. A,B (I)

   (an improvement, but there are four with A,B and two with B,A. Why is this not desirable? Impact of *learning effect*.)

3. *Balanced randomization*: To mitigate the learning effect, randomly choose three with A,B and three with B,A. (Produce one such plan on your own).

4. Other improved plans?

- factors
  - ★ keyboard: 2 levels: A, B; treatment factor
  - ★ manuscript: 6 levels: 1~6; block factor
- EUs: a manuscript type once; $6 \times 2 = 12$ EUs

treated as unknown

treated as block factor

orthogonality

only one typist. What if several typists included? → reprsentative of typist population

1. $\underset{\sim}{X}$ known $\underset{\sim}{z}$ unknown $\underset{\sim}{z}$

KB	manu.	order
A	1	I
B	1	II
A	2	I
B	2	II
A	3	I
B	3	II
A	4	I
B	4	II
A	5	I
B	5	II
A	6	I
B	6	II

confounded

2. $\underset{\sim}{X}$ known $\underset{\sim}{z}$ unknown $\underset{\sim}{z}$

KB	manu.	order
A	1	I
B	1	II
B	2	I
A	2	II
A	3	I
B	3	II
B	4	I
A	4	II
A	5	I
B	5	II
A	6	I
B	6	II

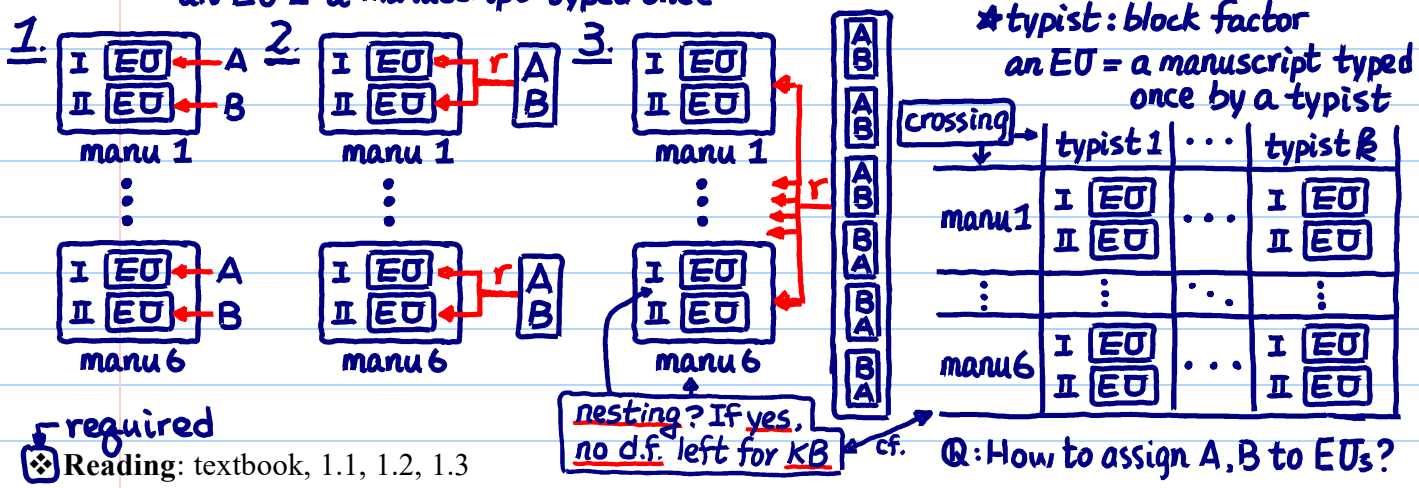
slightly confounded

3. $\underset{\sim}{X}$ known $\underset{\sim}{z}$ known $\underset{\sim}{z}$

KB	manu.	order
A	1	I
B	1	II
B	2	I
A	2	II
B	3	I
A	3	II
A	4	I
B	4	II
B	5	I
A	5	II
A	6	I
B	6	II

randomly chosen

an EU = a manuscript typed once



*typist: block factor

an EU = a manuscript typed once by a typist

crossing	typist 1	...	typist k
manu 1	I EU / II EU	...	I EU / II EU
⋮	⋮	⋱	⋮
manu 6	I EU / II EU	...	I EU / II EU

Q: How to assign A, B to EUs?

Reading: textbook, 1.1, 1.2, 1.3