Experimental Design
Note 1
Introduction

Keunbaik Lee

Sungkyunkwan University

Experiment

- An experiment is a test or a series of tests
- Experiments are used widely in the engineering world
  - Process characterization & optimization
  - Evaluation of material properties
  - Product design & development
  - Component & system tolerance determination
- All experiments are designed experiments, some are poorly designed, some are well-designed

Four Eras in the History of Experimental Design I

- The agricultural origins, 1908 - 1940s
  - W.S. Gossett and the t-test (1908)
  - R. A. Fisher & his co-workers
  - Profound impact on agricultural science
  - Factorial designs, ANOVA
- The first industrial era, 1951 - late 1970s
  - Box & Wilson, response surfaces 안배울거임
  - Applications in the chemical & process industries
- The second industrial era, late 1970s - 1990
  - QC Quality improvement initiatives in many companies
  - Taguchi and robust parameter design, process robustness

Four Eras in the History of Experimental Design II

- The modern era, beginning 1990
  - Popular outside statistics, and an indispensable tool in many scientific/engineering endeavors
  - New challenges:
    - Large and complex experiments, e.g. screening design in pharmaceutical industry, experimental design in biotechnology
    - Computer experiments: efficient ways to model complex systems based on computer simulation

A Systematic Approach to Experimentation

- Choose responses
  - What to measure? How to measure? How good is the measurement system?
- Choose factors and levels
  - Flow chart and cause-and-effect diagram
  - Factor experimental range is crucial for success
- Choose experimental plan
- Conduct the experiment
- Analyze the data
- Conclusion and recommendation
  - iterative procedure
  - confirmation experiments/follow-up experiments

Issues in Experimental Design

- Eliminate bias
  - Use a simultaneous control group
  - Randomization
  - Blinding
- Reduce sampling error
  - Replication
  - Balance
  - Blocking
- Calculate sample size

The Three Principles

- Randomization
  - Running the trials in an experiment in random order
  - Averaging out effects of “lurking” variables
- Replication
  - Sample size (improving precision of effect estimation, estimation of error or background noise)
  - Replication versus repeat measurements? (see pages 12, 13)
- Blocking
  - Dealing with nuisance factors

    성가시지만 관심 밖인 것들

Control Group

experiment group과 직접적인 비교를 위한 그룹

- A control group is a group of subjects left untreated for the treatment of interest but otherwise experiencing the same conditions as the treated subjects
- Example: one group of patients is given an inert placebo

The Placebo Effect

- Patients treated with placebos, including sugar pills, often report improvement
  - Example: up to 40% of patients with chronic back pain report improvement when treated with a placebo
  - Even "sham surgeries" can have a positive effect
- This is why you need a control group

Randomization

- Randomization is the random assignment of treatments to units in an experimental study
- Breaks the association between potential confounding variables and the explanatory variables

Blinding I

- Blinding is the concealment of information from the participants and/or researchers about which subjects are receiving which treatments
  - Single blind: subjects are unaware of treatments
  - Double blind: subjects and researchers are unaware of treatments
- Example: testing heart medication
  Two treatments: drug and placebo
- Single blind: the patients do not know which group they are in, but the doctors do
- Double blind: neither the patients nor the doctors administering the drug know which group the patients are in

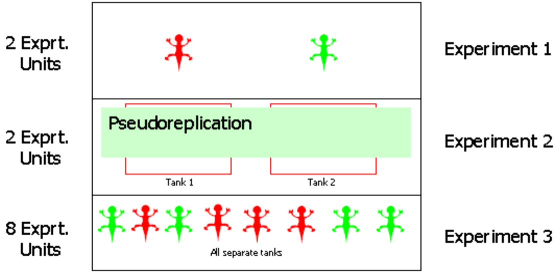
Blinding II

- The key that identifies the subjects and which group they belonged to is kept by a third party and not given to the doctors until the study is over.

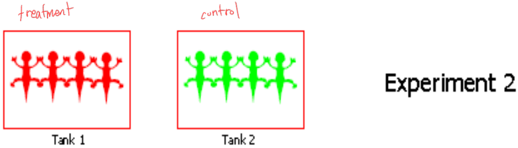
Lecture 1

Replication I

- Experimental unit: the smallest unit to which a treatment is applied
- Observational (Sampling) unit: the unit on which observation is made



Replication II



- Why is pseudoreplication bad?
  - problem with confounding and replication 교락
  - Imagine that something strange happened, by chance, to tank 2 but not to tank 1
  - Example: light burns out
  - All four lizards in tank 2 would be smaller
  - You might then think that the difference was due to the treatment, but its actually just random chance

Replication III

독립적인 반복

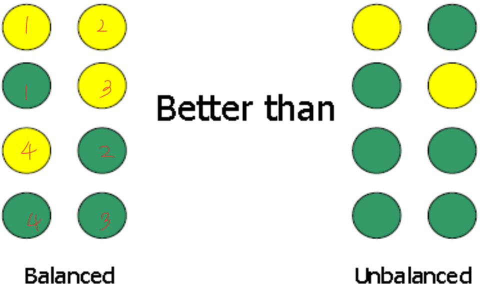
- **Replication** is the repetition of an experimental condition so that the variability associated with the phenomenon can be estimated
  - Imaging you flip a coin and it comes up heads
  - You ask a colleague to look at it and call out the result. Then another colleague is asked to observe the coin and state which side came up. This is a repeated measure.
  - A true replication is accomplished only by re-flipping the coin.
- Why is replication good?
  - Consider the formula for standard error of the mean:

$$SE_{\bar{Y}} = \frac{s}{\sqrt{n}}$$

Larger $n \Rightarrow$ Smaller SE 를 얻을 수 있으니까

Balance I

- In a balanced experimental design, all treatments have equal sample size



Balance II

- In a balanced experimental design, all treatments have equal sample size
- The test will have larger statistical power
- Also makes tests more robust to violating assumptions
- The test statistic is less susceptible to small departures from the assumption of equal variances (homoscedasticity)
- However, for single factor ANOVA, a lack of balance does not usually affect the results (Milliken and Johnson, 1984).

Blocking I

- Blocking is the grouping of experimental units that have similar properties
- Within each block, treatments are randomly assigned to experimental units
- Randomized block design
- Advantages of Blocking
  - Blocking allows you to remove extraneous variation from the data
  - Like replicating the whole experiment multiple times, once in each block
- When need consider blocking, when need consider randomization?

Blocking II

- Nuisance factors are not our interest but they do affect the response
- For the nuisance factors
  - Block what you can control
  - Randomize what you cannot control

Sample size calculation I

- Before carrying out an experiment you must choose a sample size
- Too small: no chance to detect treatment effect
- Too large: too expensive
- Sample size calculation - plan for precision
  Example:
  - Assume that the standard deviation of exam scores for a class is 10. We want to compare scores between two lab sections.
  - How many exams do we need to mark to obtain a confidence limit for the difference in mean exam scores between two sections that has a width (precision) of 5?

  Using confidence interval approach

Sample size calculation II

- Sample size calculation - plan for power
  Example:
  - Assume that the standard deviation of exam scores for a class is 10. We want to compare scores between two lab sections.
  - How many exams do we need to mark to have sufficient power (80%) to detect a mean difference of 10 points between the sections?

FP

Using power $=>$ type II error

1 - Power

power = 1 - Type II error

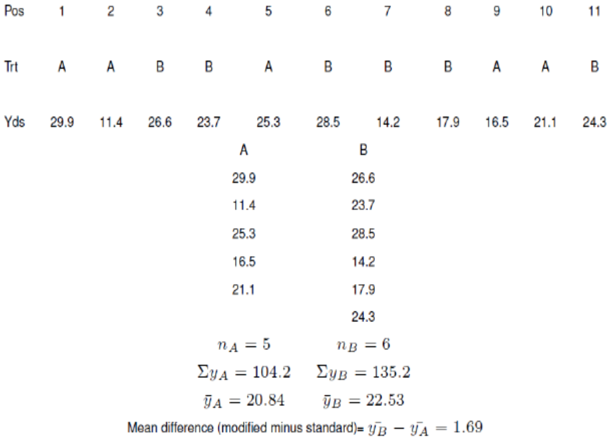
Randomized Experiment: Modified Fertilizer Mixtures for Tomato Plants I

An experiment was conducted by an amateur gardener whose object was to discover whether a change in the fertilizer mixture applied to his tomato plants would result in an improved yield. He had 11 plants set out in a single row; 5 were given the standard fertilizer mixture A, and the remaining 6 were fed a supposedly improved mixture B. The A's and B's were randomly applied to the positions in the row to give the design shown in next slide. The gardener arrived at this random arrangement by taking 11 playing cards, 5 red corresponding to fertilizer A and 6 black corresponding to fertilizer B. The cards were thoroughly shuffled and dealt to give

Randomized Experiment: Modified Fertilizer Mixtures for Tomato Plants II

the sequence shown in the design. The first card was red, the second was red, the third was black, and so forth.

Randomized Experiment: Modified Fertilizer Mixtures for Tomato Plants III



Pos	1	2	3	4	5	6	7	8	9	10	11
Trt	A	A	B	B	A	B	B	B	A	A	B
Yds	29.9	11.4	26.6	23.7	25.3	28.5	14.2	17.9	16.5	21.1	24.3

A	B
29.9	26.6
11.4	23.7
25.3	28.5
16.5	14.2
21.1	17.9
	24.3
$n_A = 5$	$n_B = 6$
$\Sigma y_A = 104.2$	$\Sigma y_B = 135.2$
$\bar{y}_A = 20.84$	$\bar{y}_B = 22.53$

Mean difference (modified minus standard)= $\bar{y_B} - \bar{y_A} = 1.69$

H_0 : the modified fertilizer does not improve the (mean) yield
H_a : the modified fertilizer improves the (mean) yield
Under the null hypothesis, A and B are mere labels and should not affect the yield. For example, the first plant would yield 29.9 pounds of tomatoes no matter it had been labeled as A or B (or fed A or B).
There are $\frac{11!}{5!6!} = 462$ ways of allocating 5 A's and 6 B's to the 11 plants, any one of which could equally be chosen. The used design is just one of 462 equally likely possibilities (why?)

For example:

Pos	1	2	3	4	5	6	7	8	9	10	11
Yds	29.9	11.4	26.6	23.7	25.3	28.5	14.2	17.9	16.5	21.1	24.3
LL1	A	A	A	A	A	B	B	B	B	B	B
LL2	A	A	A	A	B	A	B	B	B	B	B
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

LL1, LL2, etc are equally likely.

LL1: mean difference between B and A is -2.96

LL2: mean difference between B and A is -4.14

⋮

Under the null hypothesis, these differences are equally likely.

An Aside: Hypothesis Testing: Criminal Trial Analogy I

- Two Hypotheses:
  H_0 : Defendant not guilty (innocent assumption)
  H_a : Defendant guilty
  Note: H_0 represents the status quo. H_a is the conclusion that the persecution (researcher) tries to make.
- Collecting evidence:
  - In trial, finger prints, blood spots, hair samples, carpet fibers, shoe prints, ransom notes, etc.
  - In testing, survey, experiment, data.
- Fundamental Assumption
  - In trial, defendant is innocent until proven guilty, i.e., H_0 is assumed to be true.
  - In testing, similarly, we always assume H_0 is true.

An Aside: Hypothesis Testing: Criminal Trial Analogy II

- Summarizing Evidence:
  - In trial: Cross examination, argument, jury deliberation.
  - In testing: test statistic, its sampling distribution (under H_0), and observed test statistic.
- Decision Rule:
  - In trial: Reject H_0, if beyond a reasonable doubt (under the innocent assumption).
  - In testing: Reject H_0, if the observed test statistic is extreme enough:
    more extreme than a critical value or its P-vlaue is less than a theshold.

An Aside: Hypothesis Testing: Criminal Trial Analogy III

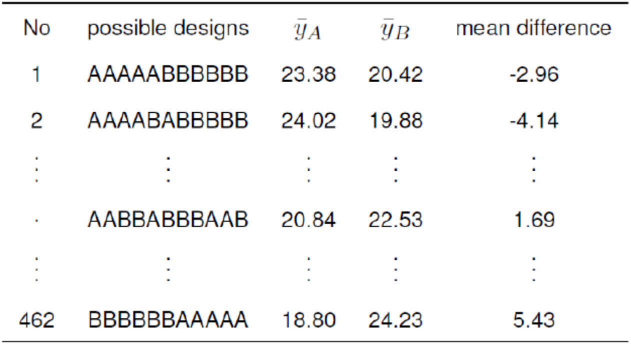
- An Important Point
  Neither decision entails proving H_0 or H_a. We merely state there is enough evidence to behave one way or the other. This is true in both trial and testing. No matter what decision we make, there is always a chance we made an error.

An Aside: Hypothesis Testing: Criminal Trial Analogy IV

Significance of Observed Difference

A summary of possible allocations and their corresponding mean differences:

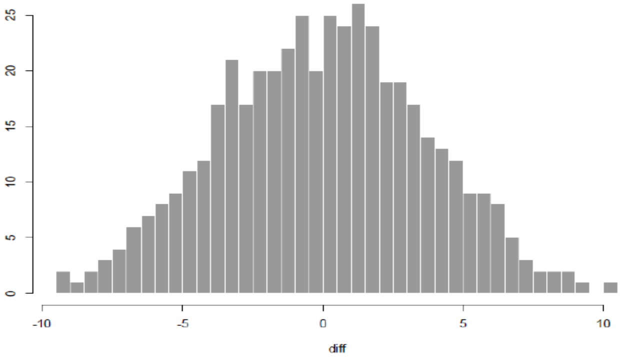
An Aside: Hypothesis Testing: Criminal Trial Analogy V



No	possible designs	$\bar{y}_A$	$\bar{y}_B$	mean difference
1	AAAAABBBBBB	23.38	20.42	-2.96
2	AAAABABBBBB	24.02	19.88	-4.14
⋮	⋮	⋮	⋮	⋮
·	AABBABBBAAB	20.84	22.53	1.69
⋮	⋮	⋮	⋮	⋮
462	BBBBBBAAAAA	18.80	24.23	5.43

An Aside: Hypothesis Testing: Criminal Trial Analogy VI

Randomization Distribution (Histogram) of the Mean Differences



An Aside: Hypothesis Testing: Criminal Trial Analogy VII

$H_0 : \mu_A = \mu_B$ vs $H_a : \mu_B > \mu_A$ $(\alpha = 5\%)$

- Randomization Test: nonparametric approach
  Observed Diff$=$ 1.69 from data
  P-value$=$ $P(\text{Diff} \geq 1.69|\textit{randomization}) = \frac{155}{462} = .335$ under H_0
  Because p-value$\geq \alpha$, do not reject H_0.

An Aside: Hypothesis Testing: Criminal Trial Analogy VIII

- Two sample $t-$test: parametric approach

$$
\begin{aligned}
s_A^2 &= 52.50, \quad s_B^2 = 29.51 \\
s_{pool}^2 &= \frac{(n_A-1)s_A^2 + (n_B-1)s_B^2}{n_A+n_B-2} = 39.73 \\
t_0 &= \frac{\bar{y}_B - \bar{y}_A}{s_{pool}\sqrt{1/n_A + 1/n_B}} = .44 \\
P - \text{value} &= P(t > t_0 | t_{(n_A+n_B-2)}) = P(t > .44 | t_{(9)}) = .34
\end{aligned}
$$

Because p-value$\geq \alpha$, do not reject H_0.

Designs

- Randomized block design
- Factorial design
- Fractional factorial design
- Latin square design
- Response surface design
- Split-plot design
- Nested design
- . . .