

Statistics 514: Design of Experiments

Topic 4

Topic Overview

This topic will cover

- Fundamentals/Model of Experimental Design
- Introduction to Randomization
- Permutation Test
- Blocking

Experimental Design

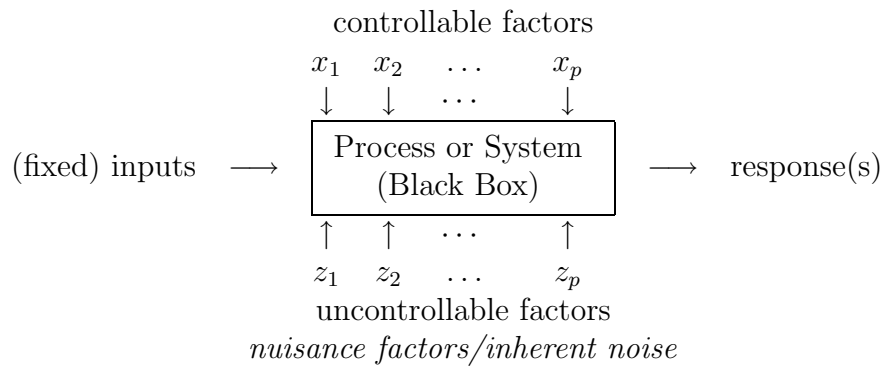
Treatments, units, and assignment methods specify an **experimental design**.

- Underlying this class is the belief that “experiments” are “different.” (Different from what? Different how?)
- XD uses careful problem-solving that require technical *assumptions* having to do with the nature of the data product.
- Since statistics is usually about analysis, understanding of design relies heavily on having good data analysis techniques. (The nature of the data analysis technique will dictate the questions you can ask.)
- Very often, assumed models seek only to *show* effects, not measure them. As such, there are many different models that could suffice. (Thus, the experiment doesn’t help in distinguishing between them.)

Desirable Criteria for Experimental Design

- The design points should exert equal influence on the determination of the regression coefficients and effect estimates.
- The design should be able to detect the need for nonlinear terms.
- The design should be robust to model misspecification, since all models are wrong.
- Designs in the early stage of the use of a sequential set of designs should be constructed with an eye toward providing appropriate information for the follow-up experiments.

Assumed Mechanism



Why Statistical Experimental Design?

Because $X_1 = X_2$ does not imply $Y_1 = Y_2$

May be:

1. $Y = f(X) + e$, random e with mean 0, OR,
2. $Y = f(X, Z)$, Z records "other variables".

Puzzler: which model is more general/more useful?

Strategies of Experimentation

- "Best-guess" experiments
 - Used a lot
 - More successful than you might suspect, but there are disadvantages ...
- One-factor-at-a-time experiments
 - Sometimes associated with the "scientific" or "engineering" method
 - Devastated by interaction, also very inefficient
- Statistically design experiments
 - Based on Fisher's factorial concept (test **all possible combinations**)

A good design must

- avoid systematic error,
- be precise,
- allow estimation of error,
- have broad validity.

Statistical expertise can help by fixing up some common mistakes, chiefly **confounding** (more later).

Failed Experiment

“Did not answer question.”

Not “Proved answer we didn’t want.”

“To call in the statistician after the experiment is done may be no more than asking him to perform a postmortem examination: he may be able to say what the experiment died of.”

- R. A. Fisher

Caveat

Even if you see a statistician, your experiment might still die.

Common things that go wrong

- not a large enough sample
- drop-outs
- ethical challenges
- political/social/legal resistance
- not enough money

Ultimately, randomized experiments are intrusive, thus setting up an artificial situation that has high internal validity but perhaps reduced external validity.

What can go right

- “It is not unusual for a well-designed experiment to analyze itself.”
- “You can see a lot by just looking.” – Yogi Berra

Terminology

Measurement Unit – actual object on which the response is measured. (Example: leaf of a plant)

Experimental Unit – batch of things (material, animal, person, machine) to which a treatment is applied. (Example: plot of land)

Usually, it is better to have more experimental units and fewer measurement units.

Criterion: experimental units should be independent.

Replication – Each treatment is applied to experimental units that are representative of the population of units to which the conclusions of the experiment will apply.

Repetition – Like replication, except that measurement is done on the *same* experimental unit.

Blinding – Evaluators of a response do not know which treatment was given to which unit.

Double-blinding – Both evaluators of the response and the experimental units do not know the assignment of treatment to units.

Control – Treatment is a “standard” treatment that is used as a baseline or base of comparison for other treatments.

Placebo – A null treatment used when the act of applying a treatment – any treatment – has an effect.

Three Sources of Variability

1. Variability due to *conditions of interest* (wanted)
2. Variability in *measurement process* (unwanted)
3. Variability in *experimental material* (unwanted)

Good design lets you estimate amount of variability due to each source.

Three Kinds of Variability

1. **Planned, systematic variability** – the kind we want
2. **Chance-like variability** – the kind we can live with
3. **Unplanned, systematic variability** – the kind that THREATENS DISASTER.

Confounding and Selection Bias

Two influences on a response are *confounded* if the design makes it impossible to isolate the effects of one from the effects of another.

Selection bias occurs in observational studies when the process of selecting groups to be compared confounds the effects of interest with other effects.

How long does it take for a car's brakes to stop it, from say, 50 miles per hour?

Blatant confounding

Compare Mercedes and minivans.
Do 10 Mercedes trials (on wet pavement).
Do 10 minivan trials (on dry pavement).
May see differences, but can't tell why.

Subtle confounding

Compare wet and dry pavement, for minivans.
While one driver does 10 trials on wet pavement,
another does 10 trials on dry pavement.

More subtle confounding

Compare wet and dry pavement, for minivans, on driver.
First do 10 trials on dry pavement,
then do 10 trials on wet pavement.
Could be confounded with run order.

Basic Principles of Experimental Design

- **Intervention** – if factors are not assigned, can't validly predict effect (or even show that there *is* an effect) after intervention.
 - Common Example: Cannot randomly assign people to smoke or not. Thus, there is little strictly valid evidence that smoking is harmful.
- **Randomization** – running trials in an experiment in random order
 - to avoid confounding with hidden factor, confound treatment assignment (or run order) with random variable that is generated to be independent of response
 - protection – averages out unknown, “lurking” factors
 - independence of trials (avoids bias)
 - randomization test \equiv Anova F -test
- **Replication** – decrease uncertainty by averaging out experimental variability
 - improves precision of effect estimation, estimation of error or background noise
- **Blocking** – decrease uncertainty by adjusting for (controlling) specific nuisance factors
 - accounts for variability but does not stem from identifiable agent (since not randomly assigned)

- **Balance/Completeness** – guarantees that there is no ambiguity as to where the effect is coming from.
- **Random factors** – even if we don't see *every* level of a factor, can infer that factor has *some* effect. In this case, get *inference* but no real predictive power.

Example of Randomization

2 groups of tomatoes

Assign varieties A/B of tomatoes to plot, measure yield:

A	A	A	A	A	B	B	B	B	B
---	---	---	---	---	---	---	---	---	---

Maybe the land isn't uniform, try:

A	B	A	B	A	B	A	B	A	B
---	---	---	---	---	---	---	---	---	---

... or a random allocation

A	B	B	B	A	A	B	A	A	B
---	---	---	---	---	---	---	---	---	---

if you're worried about periodic effects.

A strong effect is unlikely to match a random allocation [although there are no guarantees].

Randomization Principle

Whenever possible, any assigning or sampling should be done using a chance device.

- **Typically**, all allocations should be possible.
- **Typically**, all allocations should be equally likely.

How Do We Randomize?

Randomizing run order

Write out treatments in any order; apply random permutation.

Ranking Method

1. Generate $U_i \sim U[0, 1], i = 1, \dots, n$. **no ties**
2. Rank U_i is $\#\{j | U_j \leq U_i\}$.
3. $\pi(i) = \text{rank}(U_i)$.

Sampling method

1. Draw $\pi(1)$ from $\{1, \dots, n\}$.

2. Draw $\pi(2)$ from $\{1, \dots, n\} - \{\pi(1)\}$.
3. Draw $\pi(3)$ from $\{1, \dots, n\} - \{\pi(1), \pi(2)\}$.
- ⋮
4. Draw $\pi(n)$ (only one choice left).

Example

Raw	Trt	U_i	Rank	Run	Trt
1	1	0.1398928	1	1	1
2	1	0.4903066	6	2	2
3	1	0.8459779	9	3	3
4	1	0.8692369	11	4	3
5	2	0.6389887	8	5	2
6	2	0.3783782	4	6	1
7	2	0.4057894	5	7	2
8	2	0.8906754	12	8	3
9	3	0.6366516	7	9	2
10	3	0.3087094	3	10	1
11	3	0.8491306	10	11	3
12	3	0.2690837	2	12	1

Should we randomize?

1. Protects against unforeseen error patterns. More likely to get “genuine” replicates.
2. Allows randomization analysis.
 - Can we do the analysis without randomization?
 - Of course, statistical analysis does not check for randomness, but...
 - the resulting conclusions tend to be overly optimistic.
3. Might cost too much (cheaper alternatives come from sampling theory)

What Can Go Wrong Without Randomization: Patterns in Errors

Obs	1	2	3	4	5	6	7	8	9	10
Des1	A	A	A	A	A	B	B	B	B	B
Des2	A	B	B	B	A	A	B	A	A	B

$$\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2) =$$

$$\text{Des1: } \frac{1}{5}(\mu_1 + e_1 + \mu_1 + e_2 + \dots + \mu_1 + e_5) - \frac{1}{5}(\mu_2 + e_6 + \dots + \mu_2 + e_{10}) - (\mu_1 - \mu_2) =$$

$$e_1 + e_2 + e_3 + e_4 + e_5 - e_6 - e_7 - e_8 - e_9 - e_{10}$$

Des2: $e_1 - e_2 - e_3 - e_4 + e_5 + e_6 - e_7 + e_8 + e_9 - e_{10}$

Trend: If $E(e_i) = C \times (i - 5.5)$ (linear trend),

Trend adds $-25C$ under Des1; Trend adds $3C$ under Des2.

Optimal versus linear trend

A	B	B	A	A	B	B	A	A	B
---	---	---	---	---	---	---	---	---	---

Autocorrelations

$$\text{Cor}(e_i, e_j) = \begin{cases} 1 & i = j \\ \rho & |i - j| = 1 \\ 0 & |i - j| > 1. \end{cases}$$

$$\begin{aligned} \text{Var}((e_1 + e_2 + e_3 + e_4 + e_5 - e_6 - e_7 - e_8 - e_9 - e_{10})/5) &= \frac{2}{5}\sigma^2 + \frac{14}{25}\rho\sigma^2 \\ \text{Var}((e_1 - e_2 - e_3 - e_4 + e_5 + e_6 - e_7 + e_8 + e_9 - e_{10})/5) &= \frac{2}{5}\sigma^2 + \frac{2}{25}\rho\sigma^2 \end{aligned}$$

$\rho \neq 0$ changes $\text{Var}(\bar{Y}_1 - \bar{Y}_2)$.

Estimates of s_i^2 don't capture this.

A random design mitigates.

- Balance is important. Keep the treatment group sizes equal (or approximately so). There are versions of randomization that don't preserve balance.
- Randomization might occur in space/time or some other dimension.
- Randomization sometimes needs to be constrained. (Example: two queues to two different evaluators.)
- Haphazard is not randomized.

Example

4 treatment/16 units

HAPHAZARD

Treatment A is assigned to the first four units we happen to encounter, treatment B to the next four units, and so on.

As each unit is encountered, we assign treatment A, B, C, and D based on whether the "seconds" reading on the clock is between 1 and 15, 16 and 30, 31 and 45, or 46 and 60.

RANDOM

We use 16 identical slips of paper: four marked with A, four with B, and so on to D. We put the slips of paper into a basket and mix them thoroughly. For each unit, we draw a slip of paper from the basket and use the treatment marked on the slip.

Randomization Inference/Permutation Test

Data:

A	A	B	B	test statistic
22	40	10	16	18

List all $\binom{n_1 + n_2}{n_1}$ orderings:

Y:	22	40	10	16	$\bar{Y}_A - \bar{Y}_B$	
P_1	A	A	B	B	$(+22 + 40 - 10 - 16)/2$	$= +18$
P_2	A	B	A	B	$(+22 - 40 + 10 - 16)/2$	$= -12$
P_3	A	B	B	A	$(+22 - 40 - 10 + 16)/2$	$= -6$
P_4	B	A	A	B	$(-22 + 40 + 10 - 16)/2$	$= +6$
P_5	B	A	B	A	$(-22 + 40 - 10 + 16)/2$	$= +12$
P_6	B	B	A	A	$(-22 - 40 + 10 + 16)/2$	$= -18$

Logic

1. A, B identical (null hypothesis true) \Rightarrow distribution of $\bar{Y}_A - \bar{Y}_B$ independent of observation labels (null distribution)
2. Make histogram of permuted $\bar{Y}_A - \bar{Y}_B$ values.
3. If actual value (based on data) is extreme, conclude groups differ. (What's the p -value in the example?)
 - For $n_1 = n_2 = 5$, there are $\binom{10}{5} = 252$ possible orders.
 - Can do $\text{median}_A - \text{median}_B$.
 - Replace Y_{Ai} by $Y_{Ai} - \Delta$, to test $A \equiv B + \Delta$.
 - For large n_i , equivalent to t test.

Example

Paired t-test/Randomization Paired Test

In a study of egg cell maturation, the eggs from each of four female frogs were divided into two batches, and one batch was exposed to progesterone. After two minutes, the cAMP content was measured. It is believed that cAMP is a substance that can mediate cellular response to hormones.

FROG	cAMP Content		
	Control	Progesterone	Diff
1	6	4	2
2	4	5	-1
3	5	2	3
4	4	2	2

- **t-test:** $d = \{2, -1, 3, 2\} \rightarrow \bar{d} = 1.5$ and $s_{\bar{d}} = 0.866$. The test statistic is 1.732. Using Table II and 3 degrees of freedom, the p -value is between 0.05 and 0.10 (one-sided), 0.10 and 0.20 (two-sided). The actual two-sided p -value is close to 0.18.
- **randomization:** The result of each pair does not depend on the allocation of treatments. Thus, there are $2^4 = 16$ possible outcomes. The observed outcome is $2 - 1 + 3 + 2 = 6$.

$ \sum d $	# of occurrences
8	2
6	2
4	4
2	6
0	2

From the table, there are four of sixteen outcomes as “unlikely” or more, simply due to chance. Thus, the p -value is 0.25.

Discussion

- We will be discussing the randomization/permutation components of future designs (≥ 2 levels, ≥ 1 factors, mixed models, nested models, etc.).
- Understanding where “non-parametric” methods come from is becoming more important. (In the mechanics of inference, different methods affect how the reference distribution is generated or how big the confidence interval is.)
- It’s the **independence** of the randomization distribution from the experimental mechanism that makes the randomization hypothesis valid.

Poser

What do we do if we get

Obs	1	2	3	4	5	6	7	8	9	10
Des1	A	A	A	A	A	B	B	B	B	B

or

Obs	1	2	3	4	5	6	7	8	9	10
Des2	A	B	A	B	A	B	A	B	A	B

by chance?

This American Life: “It’s part of the game.”

http://www.thislife.org/Radio_Episode.aspx?sched=887

or Search on “Meet the Pros” at
<http://www.thislife.org/>

Story times: 20:45 - 47:00

Example of Problems with Randomization

In a trial on newborn infants with respiratory failure, the new treatment T was highly invasive: extracorporeal membrane oxygenation (EMCO), while the control treatment C was conventional medical management. A randomized trial was set up which saw a binary response: success or failure of the treatment.

Adaptive Urn Scheme

- At each subsequent trial, a treatment (T or C) was chosen as a ball from an urn.
- The initial trial has two balls marked T and C in the urn.
- Each time a success is observed, a ball marked by the successful treatment is added to the urn.

First Trial

Trial	Treatment	Outcome	Trial	Treatment	Outcome
1	T	survival	7	T	survival
2	C	death	8	T	survival
3	T	survival	9	T	survival
4	T	survival	10	T	survival
5	T	survival	11	T	survival
6	T	survival	12	T	survival

Analysis of the efficacy of T over C was considered to be inconclusive.

Second Trial (Boston, 1986)

- Patients were randomized equally to T and C in blocks of size 4.
- Stopping Rule: Four deaths (cumulative) on one of either T or C .

Results

- T : 9 units with no failures
- C : 10 units with 4 failures

Conclusions

Substantial but not overwhelming evidence in favor of EMCO (T).

Moral

One should be *very* careful when trying to do better than chance; sometimes there is no way to avoid being unlucky.

Upshot

Statistics lets us *quantify* our chances of being unlucky.

Blocking

A **nuisance factor** is any possible source of variability other than the conditions you want to compare; that is, anything other than effects of interest that might affect the response.

- *Randomizing* turns any bias resulting from a nuisance influence into chance error. However, this increases the size of the chance error, making it harder to detect and measure the effects of interest.
- *Blocking* turns a nuisance factor into a factor of the design.

Goal

within-block variation \ll between-block variation

The more similar the units in a block, the more effective blocking will be.

Example

Goal: Study the effect of vitamin B6 on premenstrual syndrome.

Units: Human volunteers, sorted into pairs. (One got B6; the other got a placebo.)

Grouping: Severity of symptoms (as evaluated by a questionnaire).

Another nuisance influence: stress at different times of year

“For many students, the beginning of a semester tends to be less stressful than the end, when there are exams to take and papers to write.”

“For many people, major holidays are often stressful.”

As a result...

December and January should be treated as different blocks of time in the study.

Maxim

Block what you can, and randomize what you cannot.