

# Research Design


28<sup>th</sup> April 2021




# Design of observational studies

- For the description of current status
  - Adequate sampling from the population is essential
  - Increasing the sample size can improve precision, but it costs a lot
  - In the case of rare disease (or other health outcome), long term observation (including past records) is necessary
- Observational studies with hypothesis testing
  - Cross-sectional study
    - Adequate sampling from the population is essential
    - If the target research items are many, large sample may be needed
  - Case-control study
    - In principle, all cases (with eligibility criteria) are recruited at the target institute. The key is how to sample the adequate controls
    - Basically, compare the past exposures between cases and controls
  - Cohort study
    - Follow up the exposed and non-exposed groups, comparing the health outcome measures between those two groups

# Experimental study design

- Any experimental study must be carefully designed.
  - The experimental study design was originated for agricultural study at Rothamstead by R.A. Fisher.
  - In medical and health studies, such kind of design is needed for dose-response relationship analysis of toxicity testing and clinical trial.
  - Of course, the study must pass the ethics committee, which requires the appropriate study design with proper sample size.
- 

# Fisher's three principles

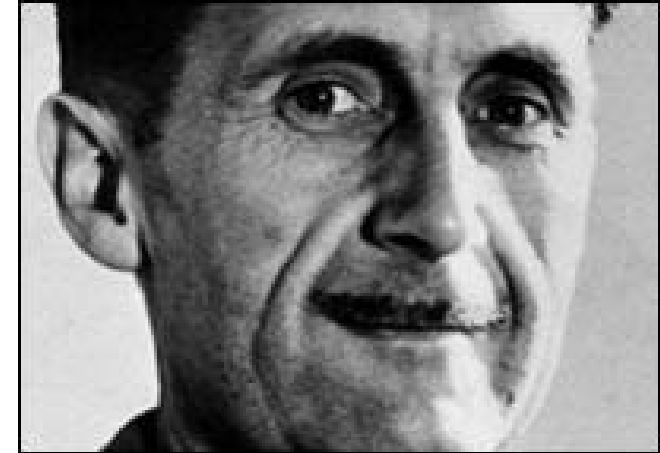
- Replication: at least 2 or more repeated measurements are needed for each treatment
  - Randomization: The order and area allocation of the experiment must be randomly assigned.
  - Local control (blocking): If the experiment is done in a large-scale, the randomization of whole samples is invalid. Instead, making local blocks in several ways, and randomization within the blocks can partly remove bias as inter-block variation.
- 

# A legend of the origin of experimental design


- One fine day at Cambridge, many professors enjoyed afternoon tea. One lady insisted her ability to distinguish the taste of milk tea whether tea first or milk first if she drink it.
  - Discussions and quarrels about her statements
- R.A. Fisher said, “How about experiment?”
- The ability can be tested by making her drink milk tea several times in random order.
- It's necessary to consider the conditions to judge her ability. How many times? Which order?

# (cf.) The order of milk and tea truly affects taste?

- George Orwell: "11 rules for perfect tea making"
  - 10: Add milk to the tea, not vice versa
- At the 100<sup>th</sup> anniversary party of George Orwell's birth, by the Royal Society of Chemistry, Dr. Andrew Stapley stated (2003):
  - It is better to have the chilled milk massed at the bottom of the cup, awaiting the stream of hot tea. This allows the milk to cool the tea, rather than the tea ruinously raise the temperature of the milk.
- Source: BBC News (<http://news.bbc.co.uk/2/hi/uk/3016342.stm>)

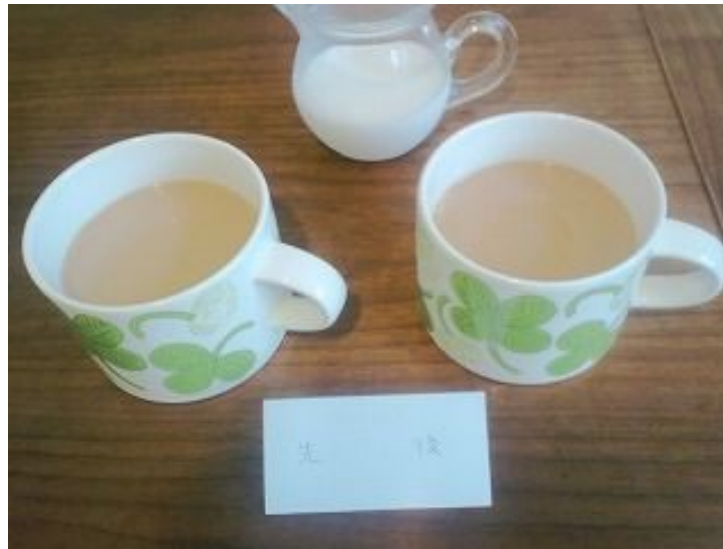


# George Orwell's 11 rules

- 1. Use tea from India or Ceylon (Sri Lanka), not China
  - 2. Use a teapot, preferably ceramic
  - 3. Warm the pot over direct heat
  - 4. Tea should be strong - six spoons of leaves per 1 litre
  - 5. Let the leaves move around the pot - no bags or strainers
  - 6. Take the pot to the boiling kettle
  - 7. Stir or shake the pot
  - 8. Drink out of a tall, mug-shaped tea cup
  - 9. Don't add creamy milk
  - 10. Add milk to the tea, not vice versa
  - 11. No sugar!
- 

# Perfect cuppa (cnt'd)

- A Japanese blogger examined the taste of 130cc tea and 30 cc Takanashi's pasteurized milk about which first makes better taste.
- According to the subjective judge, “milk first” makes better taste.
- [http://blog.livedoor.jp/teatime312/archives/cat\\_123365.html](http://blog.livedoor.jp/teatime312/archives/cat_123365.html)





# How many cups of tea are needed?

- Correct judge of 1/1 may occur at 50%.
- Correct judges of 2/2 may occur at 25%
- Correct judges of 3/3 may occur at 12.5%
- Correct judges of 4/4 may occur at 6.25%
- Correct judges of 5/5 may occur at 3.125%
  - 3.125% is too rare to accidentally occur. Usually the criteria is set at 5%. That is significance level. The null hypothesis “She has no ability to judge” is rejected when the result shows significant probability is less than 5%.
- Testing this hypothesis requires at least 5 cups.

# Famous experimental designs

- One group pretest-posttest design: paired t-test.
- Completely randomized design: t-test/ANOVA for quantitative data, chi-square test for proportion
- Randomized block design: similar to completely randomized design / considering block's effect
- Latin-square: usually ANOVA
- Crossover design: Matched (paired) analysis of variance (within-subject difference will be zero or not, adjusted by the order of treatment)

# One group pretest-posttest design

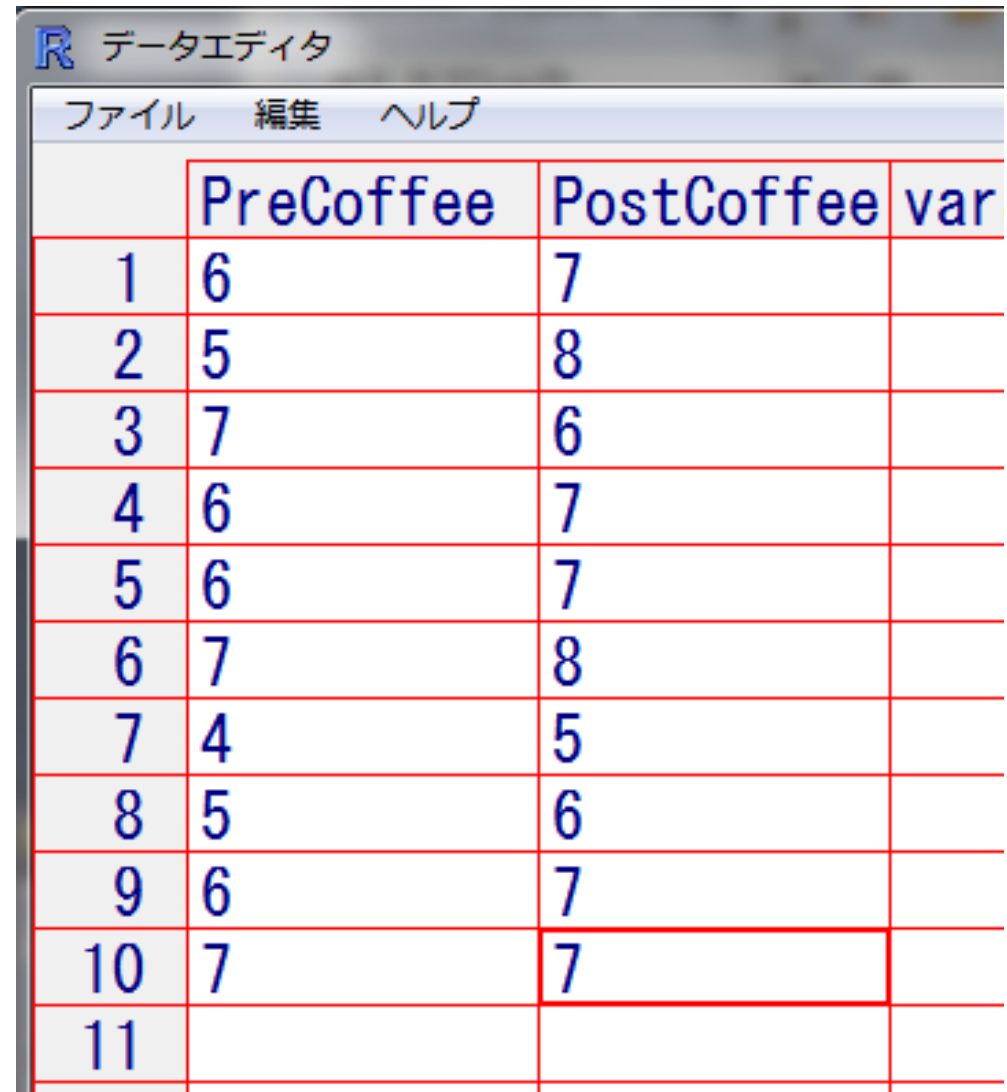
- The design enables a researcher to compute a contrast between means in which the pretest and posttest means are measured with the same precision.
  - Compare serum cortisol levels before and after the surgery in rheumatoid arthritis patients
  - Compare the depression score before and after the sound-therapy in depression patients
  - Compare the simple calculation test score before and after drinking coffee.
- The statistical test is usually paired t-test.
- Null hypothesis: the mean of difference is zero.

# Example of paired t-test

- We can use “survey” dataframe of MASS package in R (EZR), whereas it is the result of cross-sectional study.
- In EZR, select [File] and [Read data included in package], then select [MASS] and [survey].
- The “survey” contains the responses of 237 students at the University of Adelaide to a number of questions (Venables and Ripley, 1999). Variables include the span (distance from tip of thumb to tip of little finger of spread hand) of writing hand as [Wr.Hnd] and that of non-writing hand as [NW.Hnd].
- Select [Statistical Analysis], [continuous variables], then [paired t-test]. Select [NW.Hnd] at left panel and [Wr.Hnd] at right panel and click [OK]. That's all.

# Exercise

- Compare the results of simple calculation test before and after drinking coffee.
- In EZR, at first, making data: select [File], [New data], then enter the data as right screen-capture.
- Conducting paired t-test can be done in similar manner as NW.Hnd-Wr.Hnd
- [t = -2.862, df = 9, p-value = 0.01872] mean significant difference.



The screenshot shows the R Data Editor window titled "データエディタ". The menu bar includes "ファイル", "編集", and "ヘルプ". The data is organized into a table with three columns: "PreCoffee", "PostCoffee", and "var". The rows are numbered 1 through 11. The data values are as follows:

	PreCoffee	PostCoffee	var
1	6	7	
2	5	8	
3	7	6	
4	6	7	
5	6	7	
6	7	8	
7	4	5	
8	5	6	
9	6	7	
10	7	7	
11			

# Pararell group design (=Completely randomized design)

- Very simple. The subjects who signed informed consent are completely randomly (not haphazardly) assigned to one of the several treatment (exposure).
- There are several randomization methods. Fleiss JL (1986) “The design and analysis of clinical experiments” recommends to use “random permutation tables” instead of “random number tables”.
- However, now we can use computer software. If we want to assign 45 subjects into 3 treatments, type `matrix(sample(1:45, 45, replace=FALSE),3)` in R.

```
> matrix(sample(1:45, 45, replace=FALSE), 3) -> x
```

```
> x[1, ]
```

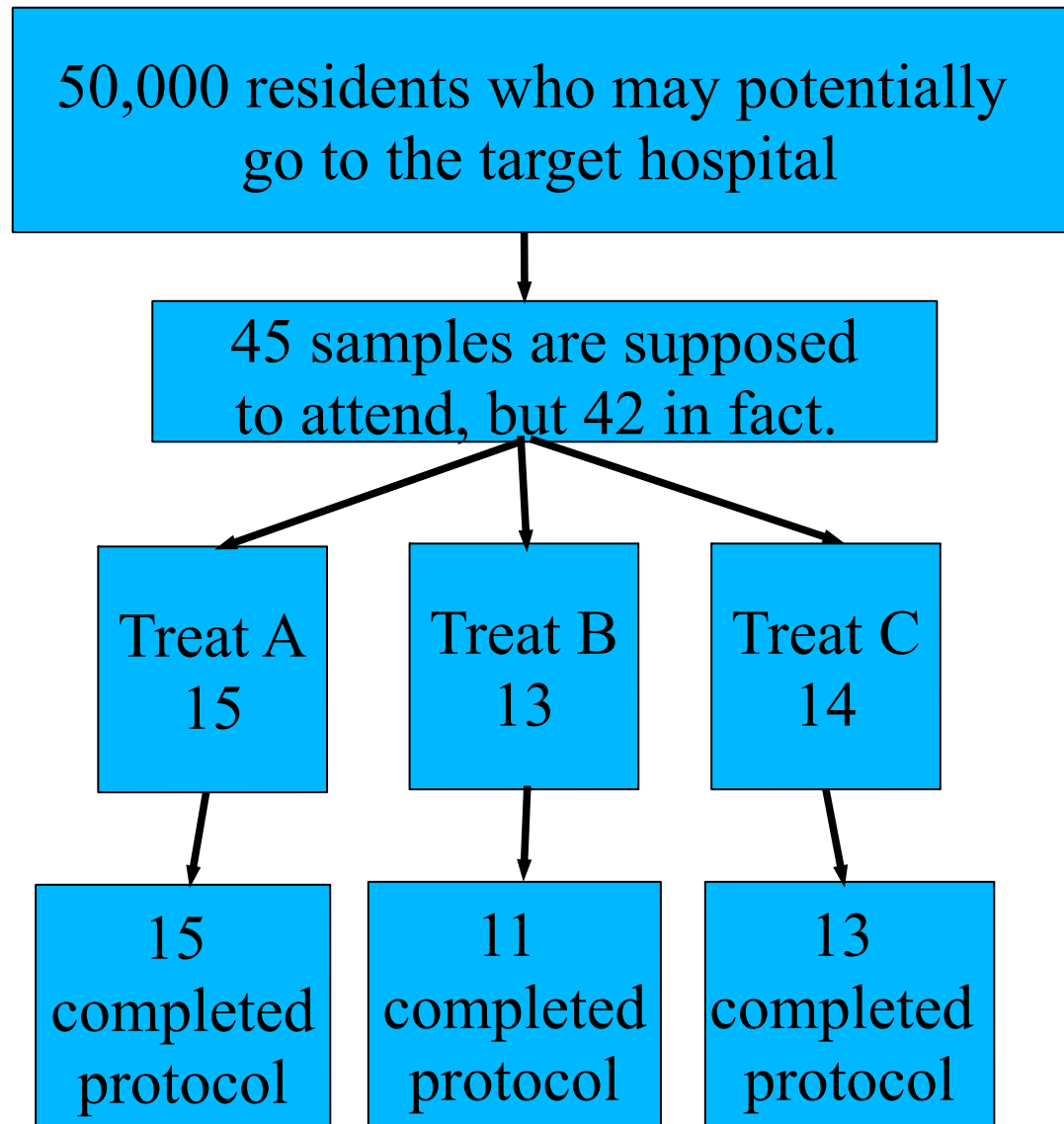
```
[1] 34 13 37 17 3 11 4 23 18 39 42 40 36 8 25
```

```
> t(apply(x, 1, sort))
```

	[, 1]	[, 2]	[, 3]	[, 4]	[, 5]	[, 6]	[, 7]	[, 8]	[, 9]	[, 10]	[, 11]	[, 12]	[, 13]	[, 14]	[, 15]
[1, ]	3	4	8	11	13	17	18	23	25	34	36	37	39	40	42
[2, ]	1	6	7	10	14	15	16	19	26	29	31	33	41	43	44
[3, ]	2	5	9	12	20	21	22	24	27	28	30	32	35	38	45

# How to describe?

- Dropouts sometimes occur.
- The design is usually shown as diagram (right).
- In this diagram, the quantitative data of the subjects can be compared by one-way ANOVA , proportion by chi-square test.
- Unbalanced sample size may reduce the statistical power.



# Randomized block(s) design (乱塊法)

- Due to incompleteness of study, completely randomized design may lead to unbalanced sample sizes among groups.
- If the sample size for each treatment is 15 and the kind of treatment is 3, randomized block design randomly selects one of 6 possible blocks ( $\{A, B, C\}$ ,  $\{A, C, B\}$ ,  $\{B, C, A\}$ ,  $\{B, A, C\}$ ,  $\{C, A, B\}$ ,  $\{C, B, A\}$ ) 15 times. By doing so, if the study may suspend in the middle, the sample size difference is at most 1 among groups. Description and analysis can be similar to complete randomization, but the analysis considering blocks' effect is also possible.
- Another method to keep size balance is “Minimization design”. It minimizes the sample size difference at each time of sampling.



# Factorial design

- Example of 2x2 factorial design
- McMaster et al. (1985): a randomized trial to evaluate breast self-examination teaching materials as leaflets or tape/slides.
- The treatment groups were designed as four parallel groups as
  1. No leaflets nor tape/slides given (control)
  2. Leaflets displayed
  3. Tape/slides program
  4. Both given
- The effect of teaching can be evaluated using ANOVA: two kinds of materials can be evaluated.

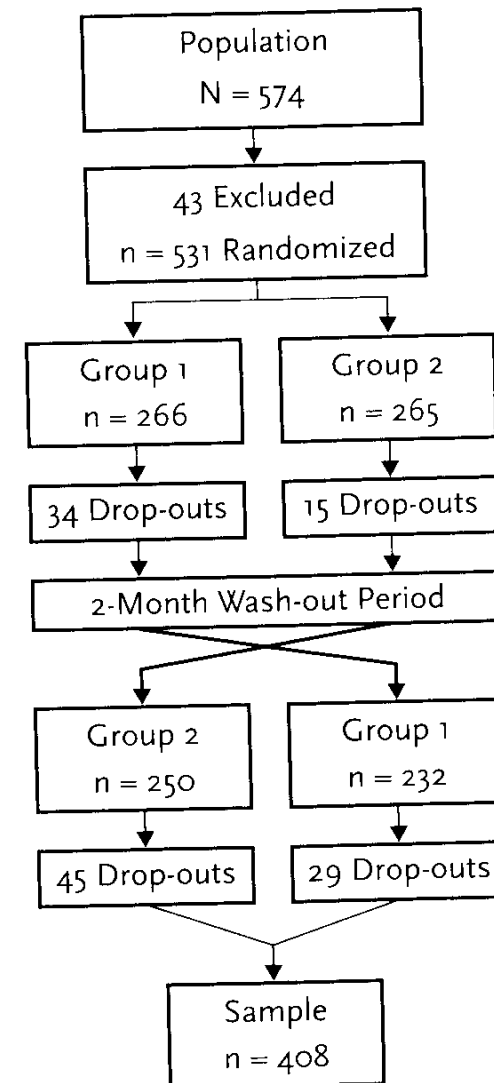
# Latin-square design (ラテン方格法)

- When the experiment have one treatment (A) with  $p \geq 2$  levels, 2 nuisance variables (B, C) each with  $p$  levels, this design is useful. The name is originated from ancient puzzle.
- Assume  $p$  is 3. Latin square is shown as right. Group 1 gets the combination of treatment  $a_1b_1c_1$  for  $n_1$  individuals. Following  $n_2$  individuals get treatment  $a_1b_2c_3$  as group 2.
- By doing so, the effects of B and C on the outcome measure can be controlled in ANOVA.

	c1	c2	c3
b1	a1	a2	a3
b2	a2	a3	a1
b3	a3	a1	a2

# Crossover design (クロスオーバー法)

- Subjects will be serially treated by 2 kinds of intervention with proper interval (wash-out period to avoid carry-over) in different order.
- (Example) Hilman BC et al. "Intracutaneous Immune Serum Globulin Therapy in Allergic Children.", JAMA. 1969; 207(5): 902-906.



# Types of outcome measure

- **Superiority trials:** The effect of new treatment is significantly better than control or not
  - Test the null-hypothesis of "no difference". If p-value is less than the significance level (usually 0.01 or 0.05), reject the null-hypothesis
- **Equivalence trials:** The effect of new treatment is similar to control or not
  - Don't use the hypothesis test. Using enough size of sample, confidence interval (CI) are used. If the CI is included within the equivalence margin (regarded as clinically equivalent), the effects by the two treatments are judged as similar.
- **Non-inferiority trials:** Special case of equivalence trials.
  - Using only half of the confidence interval, sample size will be saved.
- In R, ThreeArmedTrials package is useful for non-inferiority trials with three groups of experimental, reference and placebo.

<https://cran.r-project.org/web/packages/ThreeArmedTrials/vignettes/ThreeArmedTrials.html>

# Effect size

- Statistical revolution in psychology (triggered by Cohen 1994)
  - APA Publication Manual 6<sup>th</sup> Ed (2009) wrote that the confidence intervals should be always shown with brackets and effect size should always follow p-value as minimum requirement in APA journals.
  - In addition to psychology, effect size is used in education studies, social research and a part of epidemiological studies
- Definition "Statistical indicator to show the size of the effect"
  - Showing the extent of incorrectness of null hypothesis
  - Statistical estimates is the products of sample size function and effect size function. Thus, effect size can be interpreted as the remains of statistical estimates after subtracting the part depending on sample size
- d-family effect size (to show the size of intergroup difference, `cohen.d()` in `effsize` package)
  - Cohen's d: the difference of sample means divided by pooled SD
  - Hedges' g: the difference of sample means divided by unbiased SD which is estimated from common population variance
  - See, <http://minato.sip21c.org/ebhc/dfamefs.R>
- r-family effect size (to show the size of association. Pearson's r and so on)