STAT22200 Spring 2014 Chapter 2

Yibi Huang

April 17, 2014

Chapter 2 Randomization and Design

- Two-Sample Test
- Randomization Test (aka. Permutation Test)
- Matched-Paired Design

Example: Rat's Diet Experiment

- Objective: to investigate the effect of high protein diet on weight gain.
- 8 rats available, randomly choose 4 to be fed with beef, the remaining 4 fed with cereal.
- Response: weight gain (in grams) over a period of time.
- Results:

Protein source	*Weight gain*				*Mean*	*SD*
Cereal	111	56	86	92	86.25	22.81
Beef	104	118	117	111	112.50	6.54

- Questions: Does beef diet yield higher weight gain?

Model for the Simplest Experiment

The simplest experiment compares two **treatments** (sometimes called **treatment** and **control**.)

$$\begin{aligned} \text{Treatment 1}: &\quad y_{11}, y_{12}, \ldots, y_{1n_1} \\ \text{Treatment 2}: &\quad y_{21}, y_{22}, \ldots\ldots, y_{2n_2} \end{aligned}$$

jth obs. from treatment i		treatment effect		error (or noise)	
$\downarrow$		$\downarrow$		$\downarrow$	$i = 1, 2$
y_{ij}	$=$	μ_i	$+$	ε_{ij}	$j = 1, 2, \ldots, n_i$

- Usually we assume that the error terms ε_{ij}, $i = 1, 2$, $j = 1, 2, \ldots, n_i$, are independent and $\varepsilon_{ij} \sim N(0, \sigma_i^2)$.

Two-Sample t-Test

For an experiment with two treatments

$$y_{ij} = \mu_1 + \varepsilon_{ij}, \qquad \varepsilon_{ij}\text{'s are i.i.d.} \sim N(0, \sigma_i^2)$$
$$\text{for } i = 1, 2,\ j = 1, \ldots, n_i$$

1. H_0: $\mu_1 = \mu_2$ v.s. H_a: $\mu_1 \neq \mu_2$
2. Assuming $\sigma_1^2 = \sigma_2^2$, the test t-statistic is

$$t = \frac{\overline{y}_{1\bullet} - \overline{y}_{2\bullet}}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2} \text{ under } H_0,$$

where

$$s_p^2 = \frac{\sum_{j=1}^{n_1}(y_{1j} - \overline{y}_{1\bullet})^2 + \sum_{k=1}^{n_2}(y_{2j} - \overline{y}_{2\bullet})^2}{n_1 + n_2 - 2} = MSE,$$

called the "pooled sample variance", is an estimate of the common variance $\sigma^2 = \sigma_1^2 = \sigma_2^2$.

Example: Rat's Diet Experiment

Protein source	*Weight gain*				*Mean*	*SD*
Cereal	111	56	86	92	$\overline{y}_{1\bullet} = 86.25$	$s_1 = 22.81$
Beef	104	118	117	111	$\overline{y}_{2\bullet} = 112.50$	$s_2 = 6.54$

If assuming $\sigma_1^2 = \sigma_2^2$, the pooled sample variance is

$$s_p^2 = \frac{\sum_{j=1}^{n_1}(y_{1j} - \overline{y}_{1\bullet})^2 + \sum_{k=1}^{n_2}(y_{2j} - \overline{y}_{2\bullet})^2}{n_1 + n_2 - 2}$$

$$= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{3(22.81)^2 + 3(6.45)^2}{4 + 4 - 2} \approx 280.96$$

and the t-statistic is

$$t = \frac{\overline{y}_{1\bullet} - \overline{y}_{2\bullet}}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{86.25 - 112.50}{\sqrt{280.96(\frac{1}{4} + \frac{1}{4})}} \approx -2.21 \sim t_{4+4-2} = t_6$$

The two sided P-value is $2P(t_6 > 2.21) = 0.0687$.

When $\sigma_1^2 \neq \sigma_2^2$

When $\sigma_1^2 \neq \sigma_2^2$, instead of using the pooled sample variance, we use separate estimates for σ_1^2 and σ_2^2

$$t = \frac{\overline{y}_{1\bullet} - \overline{y}_{2\bullet}}{\sqrt{s_1^2/n_1 + s_2^2/n_2}},$$

where s_1^2 and s_2^2 are the sample variances of the 2 groups.

$$s_1^2 = \frac{\sum_{j=1}^{n_1}(y_{1j} - \overline{y}_{1\bullet})^2}{n_1 - 1} \quad \text{and} \quad s_2^2 = \frac{\sum_{j=1}^{n_2}(y_{2j} - \overline{y}_{2\bullet})^2}{n_2 - 1}$$

The t-statistic does NOT have a t-distribution, but it can be approximated by a t-distribution with df $= \nu$ where

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}.$$

This is known as the **Welch-Satterthwaite Method**.

Example: Rat's Diet Experiment

Protein source	*Weight gain*				*Mean*	*SD*
Cereal	111	56	86	92	$\overline{y}_{1\bullet} = 86.25$	$s_1 = 22.81$
Beef	104	118	117	111	$\overline{y}_{2\bullet} = 112.50$	$s_2 = 6.54$

If assuming $\sigma_1^2 \neq \sigma_2^2$, the t-statistic is

$$t = \frac{\overline{y}_{1\bullet} - \overline{y}_{2\bullet}}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} = \frac{86.25 - 112.50}{\sqrt{\frac{22.81^2}{4} + \frac{6.54^2}{4}}} \approx -2.21 \sim t_\nu$$

and the degrees of freedom ν is

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2} = \frac{\left(\frac{22.81^2}{4} + \frac{6.54^2}{4}\right)^2}{\frac{1}{3}\left(\frac{22.81^2}{4}\right)^2 + \frac{1}{3}\left(\frac{6.54^2}{4}\right)^2} = 3.48$$

The two sided P-value is $2P(t_{3.48} > 2.21) = 0.1013$.

Two-Sample t-Tests In R

```
> cereal = c(111, 56, 86, 92)
> beef = c(104, 118, 117, 111)
> t.test(cereal, beef, var.equal = T)
        Two Sample t-test

data:  cereal and beef
t = -2.2147, df = 6, p-value = 0.06869
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -55.251755   2.751755
sample estimates:
mean of x mean of y
    86.25    112.50

> t.test(cereal, beef)
        Welch Two Sample t-test

data:  cereal and beef
t = -2.2147, df = 3.477, p-value = 0.1013
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -61.202402   8.702402
sample estimates:
mean of x mean of y
    86.25    112.50
```

A Closer Look at the Error Term

- The error term ε_{ij} includes all other factors that may affect y_{ij}. For the rat's diet example, it may include

$$\varepsilon_{ij} = \text{age effect} + \text{gender effect} + \text{gene effect} + \cdots + \text{other unknown effects}$$

- The larger the size of errors, the more difficult to discern the treatment effect $\mu_1 - \mu_2$
- How to reduce the size of error?

for factors that	**how to deal with**
known and controllable	1) hold them constant 2) block design (Chap 13-14)
unknown (lurking)	randomization (Chap 2-3)
known but uncontrollable	ANCOVA (Chap 17)

- Randomization is necessary to ensure the allocation is unbiased, i.e., $\mathbb{E}(\varepsilon_{1j}) = \mathbb{E}(\varepsilon_{2j}) = 0$.

Checking Assumptions

The t-test is appropriate when a number of assumptions are true.

1. Randomization was done properly.
   - a trick to check whether randomization is done properly — using some *audit responses*.
     E.g., record the gender and other characteristics of subjects, if randomized properly, the proportion of men and women in the two groups should be close.

2. At least one of the following two assumptions is met
   - the noise ε_{ij} have normal distributions
   - the number of observations is large enough (> 30?)

   However, conducting experiment is often time-consuming and expensive. We may not afford to run an large experiment.

Randomization Test (aka. Permutation Test)

H_0 : the two diets have the same effect on weight gain

H_a : beef diet yields higher weight gain

A reasonable measurement for the effect of beef over cereal is

$$T = \overline{Y}_{beef} - \overline{Y}_{cereal}$$

- If H_0 is true, the weight gain of 8 rats are always

$$\{111, 56, 86, 92, 104, 118, 117, 111\}$$

no matter they are fed with beef or cereal. The variation in weight gain is simply the natural variation in rats. Some rats grow faster, some slower. We just happen to assign more fast-growing rats to the beef group.

- Under H_0, the four observations in the treatment group are like 4 random draws without replacement from the total of 8 outcomes

The test-statistic T we observed is

$$T = \frac{104 + 118 + 117 + 111}{4} - \frac{111 + 56 + 86 + 92}{4}$$
$$= 112.5 - 86.25 = 26.25$$

Of the $\binom{8}{4} = \frac{8!}{4!\,(8-4)!} = 70$ possible draws of four from 8 rats $\{104, 118, 117, 111_a, 111_b, 56, 86, 92\}$,

beef diet				cereal diet				test-statistic T	
118	117	111_a	111_b	104	92	86	56	29.75	
118	117	111_a	104	111_b	92	86	56	26.25	
118	117	114	111_b	111_a	92	86	56	26.25	← observed
118	104	111_a	111_b	117	92	86	56	23.25	
104	117	111_a	111_b	118	92	86	56	22.75	
		⋮				⋮		⋮	
104	92	86	56	118	117	111_a	111_b	−29.75	

only 3 of them result in differences at least as extreme as the one we observed (26.25). The **one-sided** P-value is $3/70 \approx 4.3\%$.

Two-Sided Randomization Test

For a two-sided test

H_0 : the two diets have the same effect on weight gain
H_a : the two diets have different effects on weight gain

a reasonable test statistic is $|T| = |\overline{Y}_{beef} - \overline{Y}_{cereal}|$.

By swapping rats in the two groups, we get allocations in the other extreme.

beef diet				cereal diet				test-statistic T	
118	117	111_a	111_b	104	92	86	56	29.75	
118	117	111_a	104	111_b	92	86	56	26.25	
118	117	114	111_b	111_a	92	86	56	26.25	← observed
		⋮				⋮		⋮	
111_a	92	86	56	118	117	114	111_b	-26.25	
111_b	92	86	56	118	117	111_a	104	-26.25	
104	92	86	56	118	117	111_a	111_b	-29.75	

The **two-sided** P-value is thus **twice** the **one-sided** P-value, $2 \times 3/70 = 6/70 \approx 8.6\%$.

Two-Sample Randomization (Permutation) Test

Say we have data $y_{11}, \ldots, y_{1n_1}$ from the treatment group and $y_{21}, \ldots, y_{2n_2}$ from control group.

According to H_0, the treatment makes no difference. So we may mix up the two groups. Any n_1 of the total of $n_1 + n_2$ observations is as likely to be our observations in the treatment group.

Test Procedure:

1. Find the observed difference in means: $d_{observed} = \bar{y}_1 - \bar{y}_2$.
2. For each of the $\binom{n_1+n_2}{n_1}$ allocation of units to the treatment and control group, find the mean differences of the two group

$$d_{new} = \bar{y}_{1,new} - \bar{y}_{2,new}.$$

3. If one-sided, the P-value is the number of allocations having d_{new} at least as great as d_{obs}, over $\binom{n_1+n_2}{n_1}$, if expect to observe greater responses in group 1.
4. If two-sided, the P-value is the number of allocations with $|d_{new}|$ at least as great as $|d_{obs}|$, over $\binom{n_1+n_2}{n_1}$.

Matched-Pairs Designs

- Matched-pairs designs
- t-test for matched-pairs designs
- Randomization test for matched-pairs designs

Example: Coffee & Blood Flow During Exercise

Doctors studying healthy men measured myocardial blood flow (MBF)[1] during bicycle exercise after giving the subjects a placebo or a dose of 200 mg of caffeine that was equivalent to drinking two cups of coffee.

Two possible designs:

- **Completely Randomized Design**: 16 subjects. Randomly choose 8 subjects to be given caffeine, the other 8 placebo
- **Matched Pairs Design**: 8 subjects, each is tested twice. Randomly choose 4 subjects to receive caffeine in the first test and placebo in the second test ; the other 4 receive placebo first and caffeine second. There is a 24-hour gap between the two tests.

Both designs will result in 16 measurements, 8 for caffeine and 8 for placebo. Which design would be more efficient?

[1] MBF was measured by taking positron emission tomography (PET) images after oxygen-15 labeled water was infused in the patients.

Models

Completely Randomized Design:

$$y_{ij} = \underset{\binom{\text{treatment}}{\text{effect}}}{\mu_i} + \underset{(\text{noise})}{\tau_{ij}}$$

Matched Pairs Design

$$y_{ij} = \underset{\binom{\text{treatment}}{\text{effect}}}{\mu_i} + \underset{\binom{\text{subject}}{\text{effect}}}{\alpha_j} + \underset{\binom{\text{other}}{\text{noise}}}{\eta_{ij}}$$

for $i = 1, 2$, $j = 1, 2, \ldots, 8$.

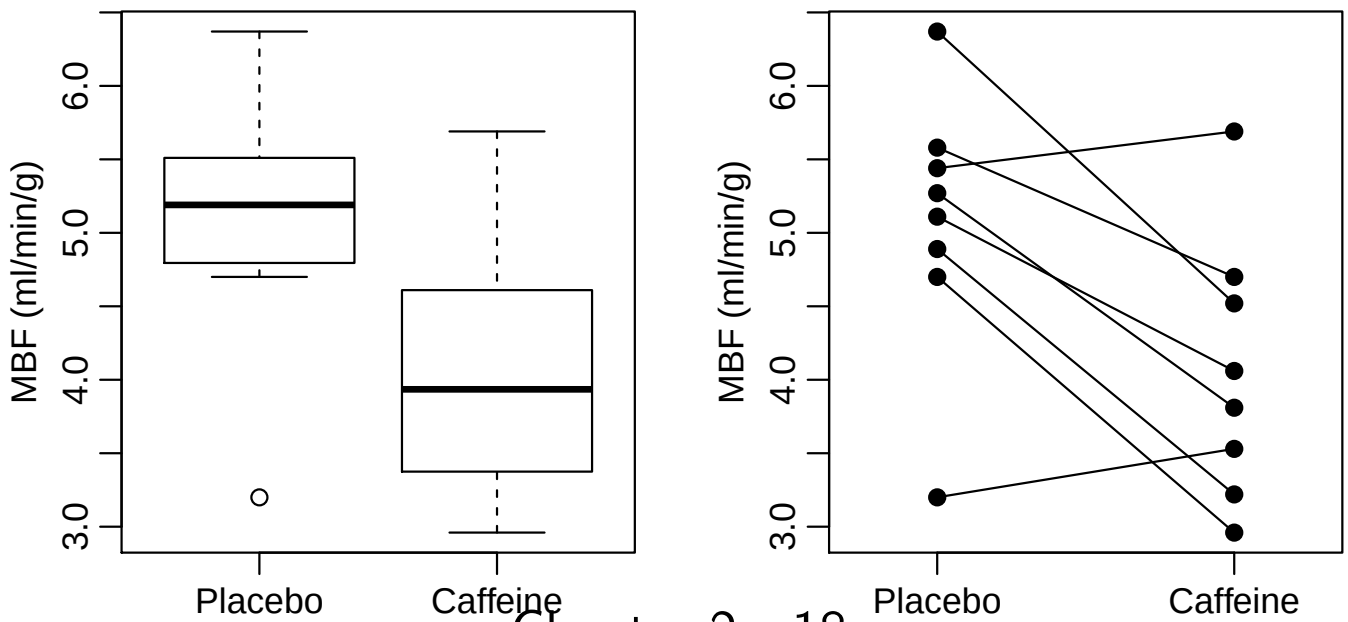
In the matched pairs design, the subject effect can be removed by taking the difference between the two measurements on the same individual.

$$d_j = y_{2j} - y_{1j} = \mu_2 - \mu_1 + \underbrace{\eta_{2j} - \eta_{1j}}_{\varepsilon_j}$$

When subject-to-subject variation is considerable, the matched pairs design can effectively reduce the size of noise.

Data for Matched Pairs Design

	MBF (ml/min/g)	
Subject	Placebo	Caffeine
1	6.37	4.52
2	5.44	5.69
3	5.58	4.70
4	5.27	3.81
5	5.11	4.06
6	4.89	3.22
7	4.70	2.96
8	3.20	3.53
Mean	5.07	4.06
SD	0.91	0.89



Matched-pair data cannot be analyzed as two independent samples since the 2 measurements on the same subject are *dependent*.
Method: take differences and analyze like a **one-sample data**.

	MBF (ml/min/g)		
Subject	Placebo	Caffeine	Difference
j	y_{2j}	y_{1j}	$d_j = y_{2j} - y_{1j}$
1	6.37	4.52	1.85
2	5.44	5.69	−0.25
3	5.58	4.70	0.88
4	5.27	3.81	1.46
5	5.11	4.06	1.05
6	4.89	3.22	1.67
7	4.70	2.96	1.74
8	3.20	3.53	−0.33
Mean	5.07	4.06	1.01
SD	0.91	0.89	0.87

To test H_0: $\mu_1 = \mu_2$, the test statistic is

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} \sim t_{n-1}$$

where

$$s_d = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(d_j - \bar{d})^2}$$

In this example, $\bar{d} = 1.01$, $s_d = 0.87$, $t = \frac{1.01-0}{0.87/\sqrt{8}} \approx 3.28$. The 2-sided P-value is

```
> 2*pt(-3.28,df=7)
[1] 0.01348706
```

Randomization Test for Matched-Pairs

- Under the H_0 that treatment has no effect, if we had changed the order of caffeine and placebo for subject 1 we would have Placebo $-$ Caffeine $= -1.85$ rather than 1.85.
- So for a randomization test, we can take as our null distribution that each difference is positive with probability $1/2$ and negative with probability $1/2$.
- Use the sum as the test statistic:

$$\sum_{j=1}^{8} d_j = d_1 + d_2 + \ldots + d_8$$

The observed value is

$$\begin{aligned} T &= \sum_{j=1}^{8} d_j \\ &= 1.85 - 0.25 + 0.88 + 1.46 + 1.05 + 1.67 + 1.74 - 0.33 \\ &= 8.07 \end{aligned}$$

For a randomization test, look at the test statistics associated with the outcomes under sign changes/permutation within each pair.

There are $2^8 = 256$ sets of randomized $\{d_1, d_2, \ldots, d_8\}$:

Randomization								test-statistic $\sum_j d_j$	
1.85	0.25	0.88	1.46	1.05	1.67	1.74	0.33	9.23	
1.85	-0.25	0.88	1.46	1.05	1.67	1.74	0.33	8.73	
1.85	0.25	0.88	1.46	1.05	1.67	1.74	-0.33	8.57	
1.85	-0.25	0.88	1.46	1.05	1.67	1.74	-0.33	8.07	← observed
		⋮			⋮			⋮	
-1.85	0.25	-0.88	-1.46	-1.05	-1.67	-1.74	0.33	-8.07	
-1.85	-0.25	-0.88	-1.46	-1.05	-1.67	-1.74	0.33	-8.57	
-1.85	0.25	-0.88	-1.46	-1.05	-1.67	-1.74	-0.33	-8.73	
-1.85	-0.25	-0.88	-1.46	-1.05	-1.67	-1.74	-0.33	-9.23	

Foe one-sided test, there are 4 randomizations that will result in a $\sum_j d_j$ that is at least as extreme as the one observed 8.07

$$P\text{-value} = 4/2^8 = 0.015625$$

For two-sided test, there are 8 randomizations, so

$$P\text{-value} = 8/2^8 = 0.03125.$$

Randomization Tests for Completely Randomized Designs

Experimental Unit versus Measurement Unit

Experimental units are the smallest groupings of the experimental material that could have gotten different treatments.
Measurement units are the actual objects on which the response is measured.

- In many cases, the measurement units are just the experimental units
- Sometimes a measurement unit is only *part* of an experimental unit.
- 12 pens of young turkeys are randomly assigned 3 different diets (20 turkeys per pen)
  - A measurement unit is one turkey, and an experimental unit is a whole pen of turkeys.
- A class full of students is assigned a certain pedagogical intervention.
  - Suppose classes of students are assigned to two different pedagogy scheme. A measurement unit is one student, and an experimental unit is a whole class of students