04

데이터분석방법론(1)

One-sample and Two-sample test

통계·데이터과학과 장영재 교수

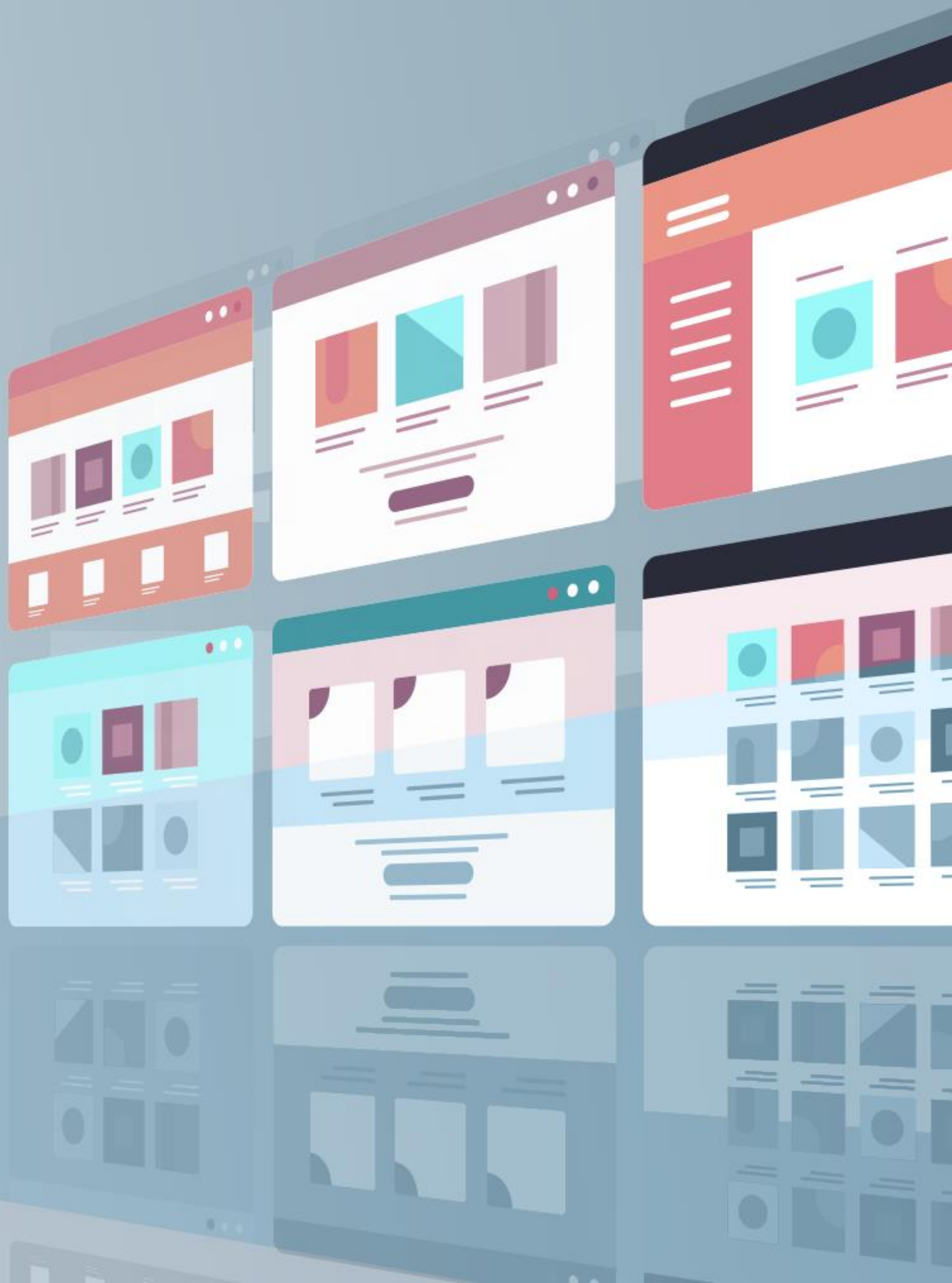
한국방송통신대학교 대학원

학습목차

1. One-sample test
2. Two-sample test (Independent samples)
3. Two-sample test (Paired samples)

01

One-sample test



1. One-sample t test

- $(x_1, \cdots, x_n) \sim N(\mu, \sigma^2)$

  $H_0 : \mu = \mu_0 \quad vs \quad H_1 : \mu \neq \mu_0$

- **검정통계량** $t_0 = \dfrac{\overline{x} - \mu_0}{s / \sqrt{n}} \sim t(n-1)$

ex) example concerning daily energy intake in kJ for 11 women (Altman, 1991, p. 183). To investigate whether the women's energy intake deviates systematically from a recommended value of 7725 kJ.

```
> daily.intake <- c(5260,5470,5640,6180,6390,6515, 6805,7515,7515,8230,8770)
> mean(daily.intake)
[1] 6753.636
> sd(daily.intake)
[1] 1142.123
> quantile(daily.intake)
  0%  25%  50%  75% 100%
5260 5910 6515 7515 8770
```

1. One-sample t test

```
> t.test(daily.intake,mu=7725)

        One Sample t-test

data:  daily.intake
t = -2.8208, df = 10, p-value = 0.01814
alternative hypothesis: true mean is not equal to 7725
95 percent confidence interval:
 5986.348 7520.925
sample estimates:
mean of x
 6753.636
```

- You can immediately see that $p < 0.05$ and thus that (using the customary 5% level of significance) data deviate significantly from the hypothesis that the mean is 7725.
- Optional arguments : mu (def=0),

  alternative = “grater”(“g”) or “less” (“l”) (def=two-sided),

  conf.level = 0.99 (def=0.95)

2. Wilcoxon signed-rank test

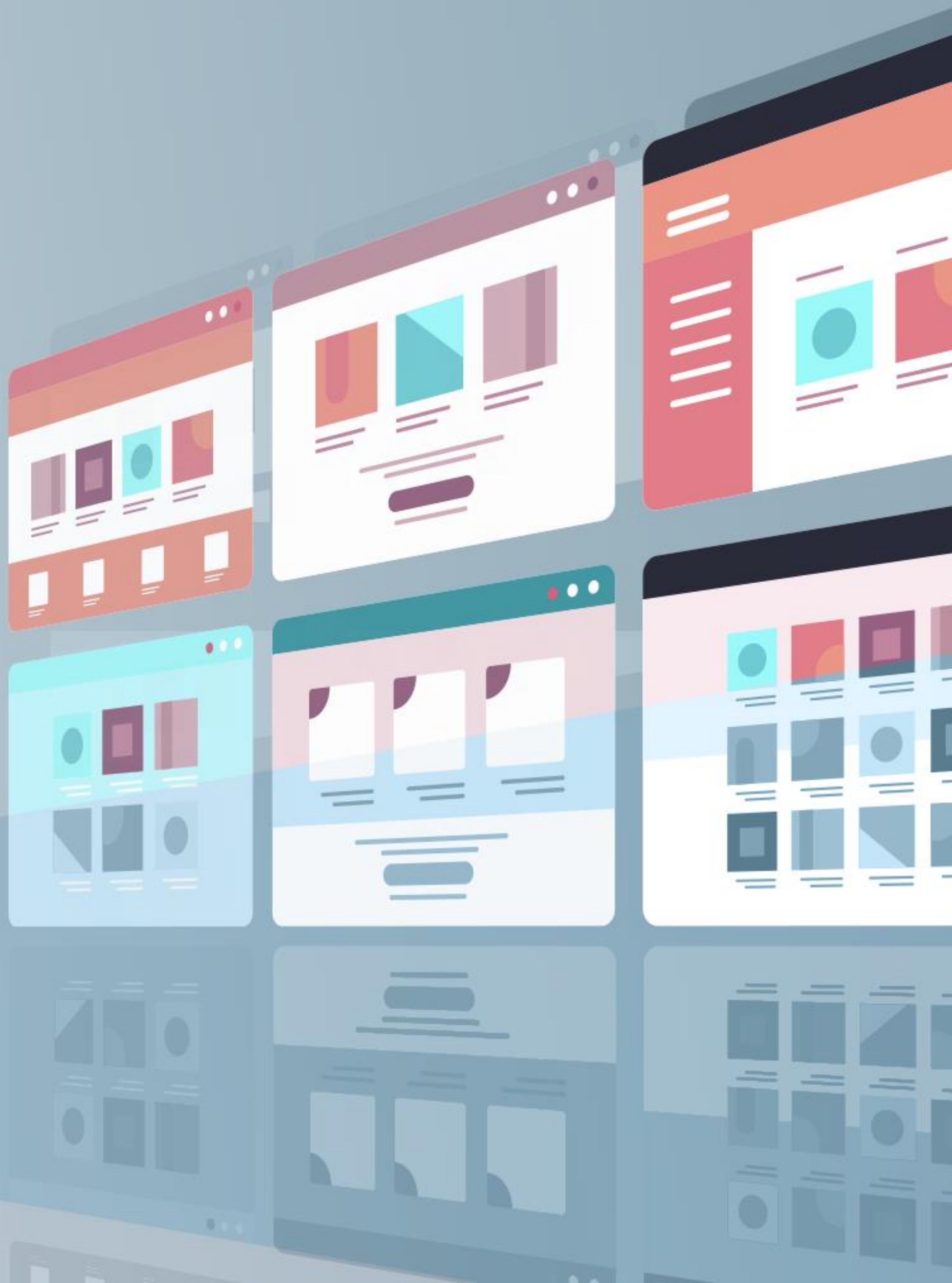
- The t tests are fairly robust against departures from the normal distribution especially in larger samples, but sometimes you wish to avoid making that assumption. To this end, the distribution-free methods are convenient.
- For the one-sample Wilcoxon test, the procedure is to subtract the theoretical μ_0 and rank the differences according to their numerical value, ignoring the sign, and then calculate the sum of the positive or negative ranks.

```
> wilcox.test(daily.intake, mu=7725)

        Wilcoxon signed rank test with continuity correction

data:  daily.intake
V = 8, p-value = 0.0293
alternative hypothesis: true location is not equal to 7725
```

- The test statistic V is the sum of the positive ranks. In the example, the p-value is computed from the normal approximation because of the tie at 7515.

02

Two-sample test (Independent samples)



1. Two-sample test

- The two-sample t test is used to test the hypothesis that two smaples may be assumed to come from distributions with the same mean ($\delta_0 = 0$).
- Hypotheses for comparing two population means (μ_1, μ_2).

$$
\begin{aligned}
&① \ H_0: \mu_1 - \mu_2 = \delta_0 \\
&\quad\ H_1: \mu_1 - \mu_2 > \delta_0 \\
&② \ H_0: \mu_1 - \mu_2 = \delta_0 \\
&\quad\ H_1: \mu_1 - \mu_2 < \delta_0 \\
&③ \ H_0: \mu_1 - \mu_2 = \delta_0 \\
&\quad\ H_1: \mu_1 - \mu_2 \neq \delta_0
\end{aligned}
$$

- If the sample size is sufficiently large (usually greater than 30), the following test statistic is used regardless of the population distribution (standard normal distribution).

$$
Z = \frac{(\overline{X}_1 - \overline{X}_2) - \delta_0}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}
$$

1. Two-sample *t* test

- If the sample size is small, under the assumption of normal distribution and equal variance we can use the following test statistics.

① $H_0: \mu_1 - \mu_2 = \delta_0$ $H_1: \mu_1 - \mu_2 > \delta_0$	$\dfrac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{\sqrt{\dfrac{S_p^2}{n_1} + \dfrac{S_p^2}{n_2}}} > t_{n_1+n_2-2,\,\alpha}$ ⇨ reject H_0
② $H_0: \mu_1 - \mu_2 = \delta_0$ $H_1: \mu_1 - \mu_2 < \delta_0$	$\dfrac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{\sqrt{\dfrac{S_p^2}{n_1} + \dfrac{S_p^2}{n_2}}} < -t_{n_1+n_2-2,\,\alpha}$ ⇨ reject H_0
③ $H_0: \mu_1 - \mu_2 = \delta_0$ $H_1: \mu_1 - \mu_2 \neq \delta_0$	$\left\lvert \dfrac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{\sqrt{\dfrac{S_p^2}{n_1} + \dfrac{S_p^2}{n_2}}} \right\rvert > t_{n_1+n_2-2,\,\alpha/2}$ ⇨ reject H_0

$$\boxed{S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

1. Two-sample *t* test

- If variances are not equal, we need to modify the degrees of freedom of the t distribution to φ (Satterthwaite approximation)

$$T = \frac{\bar{X}_1 - \bar{X}_2 - \delta_0}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$$

$$\phi = \frac{\left[\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}\right]^2}{\dfrac{\left(\dfrac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \dfrac{\left(\dfrac{S_2^2}{n_2}\right)^2}{n_2 - 1}}$$

1. Two-sample *t* test

- The two-sample t test is used to test the hypothesis that two samples may be assumed to come from distributions with the same mean.

```
> t.test(expend~stature)
        Welch Two Sample t-test
data:  expend by stature
t = -3.8555, df = 15.919, p-value = 0.001411
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.459167 -1.004081
sample estimates:
 mean in group lean mean in group obese
           8.066154           10.297778
```

- The confidence interval is for the difference in means and does not contain 0, which is in accordance with the p-value indicating a significant difference at the 5% level.
- It is Welch's variant of the t test. This is the test where you do not assume that the variance is the same in the two groups, which (among other things) results in the fractional degrees of freedom.

2. Two-sample *t* test

- To get the usual (textbook) t test, you must specify that you are willing to assume that the variances are the same. This is done via the optional argument var.equal=T;

```
> t.test(expend~stature, var.equal=T)

        Two Sample t-test

data:  expend by stature
t = -3.9456, df = 20, p-value = 0.000799
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.411451 -1.051796
sample estimates:
 mean in group lean mean in group obese
           8.066154           10.297778
```

- Notice that the degrees of freedom now has become a whole number, namely 13 + 9 − 2 = 20. The p-value has dropped slightly (from 0.14% to 0.08%) and the confidence interval is a little narrower, but overall the changes are slight.

2. Comparison of Variances

- Two population variances σ_1^2, σ_2^2, sample sizes n_1, n_2, and sample variances S_1^2, S_2^2 respectively
- The following test statistic follows F distribution with degrees of reedom $(n_1 - 1, n_2 - 1)$ under the null hypothesis $\sigma_1^2 = \sigma_2^2$

$$F = \left(\frac{S_1^2}{\sigma_1^2}\right) \Big/ \left(\frac{S_2^2}{\sigma_2^2}\right)$$

$H_0 : \sigma_1^2 = \sigma_2^2$ $H_1 : \sigma_1^2 \neq \sigma_2^2$	$\frac{S_1^2}{S_2^2} > F_{n_1-1,\ n_2-1,\ \alpha/2}$ or $\frac{S_1^2}{S_2^2} < F_{n_1-1,\ n_2-1,\ 1-\alpha/2}$ $\Rightarrow$ reject H_0

2. Comparison of Variances

- Testing the assumption that the variances are the same

```
> var.test(expend~stature)

        F test to compare two variances

data:  expend by stature
F = 0.7844, num df = 12, denom df = 8, p-value = 0.6797
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1867876 2.7547991
sample estimates:
ratio of variances
          0.784446
```

- The test is not significant, so there is no evidence against the assumption that the variances are identical. However, the confidence interval is very wide. For small data sets such as this one, the assumption of constant variances largely a matter of belief.

3. Two-sample Wilcoxon test

- You might prefer a nonparametric test if you doubt the normal distribution assumptions of the t test. The two-sample Wilcoxon test is based on replacing the data by their rank (without regard to grouping) and calculating the sum of the ranks in one group, thus reducing the problem to one of sampling n1 values without replacement from the numbers 1 to n1+n2.

```
> wilcox.test(expend~stature)

        Wilcoxon rank sum test with continuity correction

data:  expend by stature
W = 12, p-value = 0.002122
alternative hypothesis: true location shift is not equal to 0
```

- The test statisticW is the sum of ranks in the first group minus its theoretical minimum (i.e., it is zero if all the smallest values fall in the first group).

03

Two-sample test (Paired samples)

1. The paired *t* test

- Paired tests are used when there are two measurements on the same experimental unit.

Sample from pop1 (X_{i1})	Sample from pop2 (X_{i2})	$D_i = X_{i1} - X_{i2}$
X_{11}	X_{12}	$D_1 = X_{11} - X_{12}$
X_{21}	X_{22}	$D_2 = X_{21} - X_{22}$
$\vdots$	$\vdots$	$\vdots$
X_{n1}	X_{n2}	$D_n = X_{n1} - X_{n2}$

$\overline{D} = \sum D_i / n$

$s_D^2 = \sum (D_i - \overline{D})^2 / (n-1)$

1. The paired *t* test

- Paired tests are used when there are two measurements on the same experimental unit.

① $H_0: \mu_1 - \mu_2 = D_0$ $H_1: \mu_1 - \mu_2 > D_0$	$\dfrac{\bar{D} - D_0}{\dfrac{S_D}{\sqrt{n}}} > t_{n-1,\,\alpha}$ ⇨ reject H_0
② $H_0: \mu_1 - \mu_2 = D_0$ $H_1: \mu_1 - \mu_2 < D_0$	$\dfrac{\bar{D} - D_0}{\dfrac{S_D}{\sqrt{n}}} < -t_{n-1,\,\alpha}$ ⇨ reject H_0
③ $H_0: \mu_1 - \mu_2 = D_0$ $H_1: \mu_1 - \mu_2 \neq D_0$	$\left\lvert \dfrac{\bar{D} - D_0}{\dfrac{S_D}{\sqrt{n}}} \right\rvert > t_{n-1,\,\alpha/2}$ ⇨ reject H_0

1. The paired *t* test

- Paired tests are used when there are two measurements on the same experimental unit.

```
> intake
   pre post
1  5260 3910
2  5470 4220
3   5640 3885
   --- ---
> attach(intake)
> t.test(pre, post, paired=T)
       Paired t-test
data:  pre and post
t = 11.9414, df = 10, p-value = 3.059e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1074.072 1566.838
sample estimates:
mean of the differences
             1320.455
```

2. The matched-pairs Wilcoxon test

- The paired Wilcoxon test is the same as a one-sample Wilcoxon signedrank test on the differences.

```
> wilcox.test(pre, post, paired=T)

        Wilcoxon signed rank test with continuity correction

data:  pre and post
V = 66, p-value = 0.00384
alternative hypothesis: true location shift is not equal to 0
```

다음시간 안내

05

Analysis of Variance