IV-3. Statistical test

IV-3-1. Student's t test

Student's t test is used for judgement whether there is a difference between two datasets. Statistically, we discuss whether two sample populations are obtainable from a parent population.

We discuss acceptability of calculated average from sample population (M) as the average value of assuming parental population.

$$M = \mu \pm \alpha\sigma$$

M: average of sample populatio

μ: average of parental population

σ: standard deviation of parental population

α: criterion of difference

Criterion of difference (α) is value expressing degree of difference the of which unit is σ. If we know the value of μ and σ, α is obtainable by following formula

$$\alpha = \frac{M - \mu}{\sigma}$$

When the value is larger than a threshold in standard normal distribution, we can judge the value is in adequate as a candidate of average of parental population. As an example, In a judgement at the 95% confidence limit ($p \leq 0.05$), we judge M is in adequate as candidate of average of parental population, when the value of α is larger than 1.96. This also means that the sample population is not sampled from parental sample. This is logically correct. However, we do not know μ and σ. T distribution is a model including stochastic variability of variance of sample population. T distribution is combination of normal distribution and χ^2 distribution in which variance is a stochastic variable.

$$t = \frac{M - \mu}{\sqrt{s^2}}$$

$$M = \mu \pm ts$$

t: criterion of difference in t distribution

s^2: secondary moment of average of sample population around average of parental population

Secondary moment of estimated average is square of standard error.

$$s = \frac{\sigma_{sample}}{\sqrt{n}}$$

Average of parental population is unknown, though in this discussion $\mu = 0$, because our null hypothesis is "no difference". The average of parental population should be 0. We can express average of ample population by following formula.

$$t = \frac{\mu - M}{\frac{\sigma_{sample}}{\sqrt{n}}}$$

when $\mu = 0$,

$$t = \frac{M}{\frac{\sigma_{sample}}{\sqrt{n}}}$$

This is the model of t test. Actually, we have to consider how to calculate s and n depending on the type of dataset.

We can consider following two types. One is paired dataset. Most simple example is comparison of growth of two species of grass planted in same pots. This called paired t test. The other is comparison of two species of grass planted in different pots. Generally, this is called t test.

Paired t test

In this test each data has its pair data as follow

$A_1 \leftrightarrow B_1$

$A_2 \leftrightarrow B_2$

$\vdots$

$A_i \leftrightarrow B_i$

$\vdots$

$A_n \leftrightarrow B_n$

We can make new dataset (C_i) from this dataset deducting B_i from A_i

$$C_i = A_i - B_i$$

We can obtain mean of C_i

$$M_c = \frac{1}{n}\sum_{i=1}^{n} C_i$$

Observed value of t is as follow

$$t = \frac{M_c}{\frac{\sigma_c}{\sqrt{n}}}$$

Formula 42

Example

i	A	B		C
1	3	1		2
2	5	4		1
3	7	6		1
4	3	1		2
5	9	7		2
n				5
df				4
sum				8
average				1.6
SS				1.2
σ^2				0.3

$$M_c = 1.6$$

$$\sigma_c{}^2 = \frac{1.2}{4} = 0.3$$

$$\sigma_c = \sqrt{0.3}$$

$$n = 5$$

$$\mathrm{t} = \frac{M_c}{\frac{\sigma_c}{\sqrt{n}}} = \frac{1.6}{\frac{\sqrt{0.3}}{\sqrt{5}}} = 6.531973$$

Threshold of t value at $p \leq 0.01$ df = 4 is 4.604

$$\mathrm{t} \geq 4.604$$

We can conclude that the difference is significant at $p \leq 0.01$ and A is larger than B

Unpaired t test

Unpaired t test is logically similar to paired t test, though we need to consider how we can get the standard deviation of difference between sample population A and B.

When we express the standard deviation as s_{A-B}

$$x = \mu \pm t s_{A-B}$$

When we put $x = M_A,\ \mu = M_B$ in the formula

$$|M_A - M_B| = t s_{A-B}$$

$$t = \frac{|M_A - M_B|}{\frac{\sigma_{A-B}}{\sqrt{n_{A-B}}}}$$

n_{A-B}: numbe of data in $A - B$

n_A: numbe of data in A

n_B: numbe of data in B

We can calculate M_A and M_B from sampled dataset, though we do not know σ_{A-B} and n_{A-B}.

$$\frac{\sigma_{A-B}}{\sqrt{n_{A-B}}} = s$$

Variance including A and B is ${\sigma_{\mathrm{A-B}}}^2$. We have to estimate ${\sigma_{\mathrm{A-B}}}^2$ from ${\sigma_A}^2 \, and \, {\sigma_B}^2$. Our null hypothesis is that sample datasets A and B are taken from the same parent population. So, it should be

$${\sigma_A}^2 = {\sigma_B}^2$$

${\sigma_{A-B}}^2$ is average of ${\sigma_A}^2$ and ${\sigma_B}^2$

$${\sigma_{A-B}}^2 = {\sigma_A}^2 = {\sigma_B}^2$$

Expected value of secondary moment of A around average of parent population is as follow

$$\mathrm{E}\left({M_A}^2\right) = \frac{1}{n_A}{\sigma_A}^2$$

$$\mathrm{E}\left({M_A}^2\right) = \frac{1}{n_A}{\sigma_A}^2 = \frac{1}{n_A}{\sigma_{A-B}}^2$$

Similarly,

$$\mathrm{E}\left({M_B}^2\right) = \frac{1}{n_B}{\sigma_B}^2 = \frac{1}{n_B}{\sigma_{A-B}}^2$$

$$\mathrm{s}^2 = \mathrm{E}\left({M_{A-B}}^2\right) = \mathrm{E}\left({M_{A+B}}^2\right) = \mathrm{E}\left({M_A}^2\right) + \mathrm{E}\left({M_B}^2\right)$$

$$\mathrm{s}^2 = \frac{1}{n_A}{\sigma_{A-B}}^2 + \frac{1}{n_B}{\sigma_{A-B}}^2 = \left(\frac{1}{n_A} + \frac{1}{n_B}\right){\sigma_{A-B}}^2$$

$$\mathrm{s} = \sigma_{A-B}\sqrt{\frac{1}{n_A} + \frac{1}{n_A}}$$

Here, $\frac{\sigma_{A-B}}{\sqrt{n_{A-B}}} = s$,

$$\frac{\sigma_{A-B}}{\sqrt{n_{A-B}}} = \sigma_{A-B}\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}$$

$$n_{A-B} = \frac{1}{\frac{1}{n_A} + \frac{1}{n_B}} = \frac{n_A n_B}{n_A + n_B}$$

$$n_{A-B} = \frac{n_A n_B}{n_A + n_B}$$

$$\mathrm{t} = \frac{M_A - M_B}{\sigma_{A-B}\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}}$$

Formula 43

As explanation in IV-2, variance of A-B is similar to that of A+B, it can be obtained as weighted mean by each degree of freedom.

$$\sigma_{A-B}{}^2 = \sigma_{A+B}{}^2 = \frac{(n_A - 1)\sigma_A{}^2 + (n_B - 1)\sigma_B{}^2}{n_A + n_B - 2}$$

$$\sigma_{A-B}{}^2 = \sigma_{A+B}{}^2 = \frac{SS_A + SS_B}{n_A + n_B - 2}$$

Example

	A	B
	1	1
	5	5
	6	6
		8
n	3	4
df	2	3
sum	12	20
average	4	5
SS	14	26
σ^2	7	8.6667

A： $n_A = 3,\ \mathrm{df}_A = 2,\ M_A = 4, SS_A = 14$

B： $n_B = 4, \mathrm{df}_B = 3,\ M_B = 5, SS_B = 26$

A-B： $n_{A-B} = \frac{n_A n_B}{n_A + n_B} = \frac{3\times4}{3+4} = 1.714286$

$$\mathrm{Df}_{A-B} = \mathrm{df}_A + \mathrm{df}_B = 2 + 3 = 5$$

$$|M_A - M_B| = 5 - 4 = 1$$

$$SS_{A-B} = SS_A + SS_B = 14 + 26 = 40$$

$$\sigma_{A-B}{}^2 = \frac{SS_{A-B}}{\mathrm{df}_{A-B}} = \frac{40}{5} = 8$$

$$\sigma_{A-B} = \sqrt{8}$$

$$t = \frac{M_A - M_B}{\frac{\sigma_{A-B}}{\sqrt{n_{A-B}}}} = \frac{1}{\frac{\sqrt{8}}{\sqrt{\frac{12}{7}}}} = 0.46291$$

Threshold of t value ($p \leq 0.05$) is 2.571

$$t < 2.571$$

We cannot deny the null hypothesis that $M_A - M_B = 0$, ($M_A = M_B$).

From this, we cannot conclude that averages of A and B is different, and we cannot say that sample populations A and B are taken from the different parent population.