

## Review:

Let  $X_1, X_2, \dots, X_n$  denote  $n$  independent random variables sampled from some distribution (might not be normal!) with mean ( $\mu$ ) and standard deviation ( $\sigma$ ).

Then

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \approx Z \sim N(0, 1)$$

In other words,  $\bar{X}$  is approximately Normally distributed with mean  $\mu$  and variance  $\frac{\sigma^2}{n}$ . The approximation gets better as  $n$  increases.

Let  $X_1, X_2, \dots, X_n$  denote  $n$  independent random variables sampled from some distribution (might not be normal!) with mean ( $\mu$ ) and unknown standard deviation (estimated by  $s$ ).

Then

$$\frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \approx T \sim t_{n-1}$$

Suppose we collect 8 pairs of twins. The first twin in the pair is healthy; the second is not. For each twin, we measure grey matter density. Processed data from the 8 pairs is shown below (units not given).

Pair	Twin 1	Twin 2
1	134	128
2	139	136
3	176	163
4	152	153
5	166	157
6	161	154
7	129	125
8	122	124

Is grey matter density in the populations significantly different?

## General Approach to Hypothesis Testing

1. Define the research question and formulate the appropriate null and alternative hypotheses.
2. Decide on your type I error rate ( $\alpha$ ).
3. Assume that the null is true, construct a test statistic, and identify what is known about the distribution of the test statistic under this assumption.
4. Collect a sample and evaluate the test statistic.
5. Calculate the p-value. Note that calculation of this value depends on the form of the alternative hypothesis.
6. If the p-value is less than  $\alpha$ , reject the null. Otherwise, do not reject the null.

Note: We have not yet considered type II error rates or size of the sample. More on this later.

Consider the population differences,  $d$ ,

where  $\bar{d}$  is the sample mean of the differences,  $\delta$  is the population mean of the differences, and  $s_d$  is the sample standard deviation of the differences. The quantity

$$t = \frac{\bar{d} - \delta}{s_d / \sqrt{n}},$$

is approximately  $t$  distributed with  $n - 1$  degrees of freedom.

Pair	Twin 1	Twin 2	difference
1	134	128	6
2	139	136	3
3	176	163	13
4	152	153	-1
5	166	157	9
6	161	154	7
7	129	125	4
8	122	124	-2

In this sample,  $\bar{d} = 4.875$  and  $s_d = 4.998$ .

Test  $H_0 : \delta = 0$  versus  $H_A : \delta \neq 0$ , at significance level  $\alpha = 0.05$ .

Compute the test statistic

$$\frac{\bar{d} - 0}{s_d/\sqrt{n}} = \frac{4.875}{4.998/\sqrt{8}} = 2.76$$

For 7 degrees of freedom  $t_{0.975} = 2.365$ , so the null hypothesis is rejected and we conclude that there is a non-zero difference in grey matter density.

### Paired Data

Paired data is where two observations are taken from the same individual or from two individuals who are very similar. For example,

- one observation from each eye of an individual,
- a “before” and “after” observation from an individual,
- the response to two different treatments on the same individual,
- a response for both a “case” and a “control” from similar individuals, and
- responses from twins.

In many cases, the difference in response between the two treatments or states can be computed for each pair (such as “twin 1” – “twin 2”).

Suppose that we will sample  $n$  pairs  $(X, Y)$ . From each pair we compute the difference  $D = X - Y$ . Let  $\delta$  denote the true difference in population means ( $\delta = \mu_1 - \mu_2$ ), where  $\mu_1$  is population mean of  $X$  and  $\mu_2$  is population mean of  $Y$ .

With paired data, we are usually interested in testing the hypotheses:

#### Two-sided

$$H_0 : \delta = 0 \text{ versus } H_A : \delta \neq 0$$

#### One-sided

$$H_0 : \delta \leq 0 \text{ versus } H_A : \delta > 0$$

or

$$H_0 : \delta \geq 0 \text{ versus } H_A : \delta < 0$$

A two-sided  $100(1 - \alpha)\%$  confidence interval for the true mean difference ( $\delta = \mu_1 - \mu_2$ ) between two paired samples of size  $n$  from a Normal distribution (when the standard deviation of the differences  $\sigma_d$  is not known) is given by

$$\left( \bar{d} - t_{(n-1), 1-\alpha/2} \frac{s_d}{\sqrt{n}}, \bar{d} + t_{(n-1), 1-\alpha/2} \frac{s_d}{\sqrt{n}} \right)$$

For the last example, the 95% confidence interval is:

$$(0.696, 9.05)$$

HW: What if the data were such that  $\bar{d} = -4.125$  and  $s_d = 5.718$ . Would you reject the null of no difference in grey matter between the populations? Construct a 95% confidence interval for this case. Does it contain zero? What is the relationship between hypothesis testing and confidence intervals?

## Comparison of Two Population Means

Samples can be paired or independent.

Paired Samples: As we just saw, the hypothesis test proceeds just as in the one sample case. The data,  $d_1, d_2, \dots, d_n$ , are defined to be the difference between the values in the first sample and the corresponding value in the second sample ( $d_i = x_i - y_i, i = 1, 2, \dots, n$ ). Just as before, construction of the test statistic depends on if the variance is known or not known.

Independent Samples: Hypothesis test proceeds similarly to the one sample case. Construction of the test statistic depends not only on whether the population variance is known, but also on if the population variances are equal.

### Review Notation

Let  $X_1, X_2, \dots, X_{n_1}$  denote random variables from a distribution with mean  $\mu_1$  and standard deviation  $\sigma_1$ ;  $Y_1, Y_2, \dots, Y_{n_2}$  denote random variables from a distribution with mean  $\mu_2$  and standard deviation  $\sigma_2$ .

Let  $x_1, x_2, \dots, x_{n_1}$  denotes a sample from the distribution with mean  $\mu_1$  and standard deviation  $\sigma_1$ ;  $y_1, y_2, \dots, y_{n_2}$  denote a sample from the distribution with mean  $\mu_2$  and standard deviation  $\sigma_2$ .

Question of Interest: Given the sample data, test  $H_0 : \mu_1 = \mu_2$  against some alternative.

### Recall and Note

Recall that if  $X_1, X_2, \dots, X_{n_1}$  denote i.i.d. random variables with mean  $\mu_1$  and standard deviation  $\sigma_1$ ,

$$\frac{\bar{X} - \mu_1}{\frac{\sigma_1}{\sqrt{n_1}}} \approx Z \sim N(0, 1)$$

Note that if  $X_1, X_2, \dots, X_{n_1}$  denote i.i.d. random variables with mean  $\mu_1$  and standard deviation  $\sigma_1$ , and  $Y_1, Y_2, \dots, Y_{n_2}$  denote i.i.d. random variables with mean  $\mu_2$  and standard deviation  $\sigma_2$ ,

$$(\bar{X} - \bar{Y}) \approx N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

and so

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \approx Z \sim N(0, 1)$$

Let  $x_1, x_2, \dots, x_{n_1}$  denotes a sample from a distribution with mean  $\mu_1$  and standard deviation  $\sigma_1$ ;  $y_1, y_2, \dots, y_{n_2}$  denote a sample from a distribution with mean  $\mu_2$  and standard deviation  $\sigma_2$ .

Question of Interest: Given the sample data, test  $H_0 : \mu_1 = \mu_2$  against some alternative.

## Hypothesis Testing for a Difference Between Two Population Means

Two-sided :

$$H_0 : \mu_1 - \mu_2 = (\mu_1 - \mu_2)_0 \text{ versus } H_A : \mu_1 - \mu_2 \neq (\mu_1 - \mu_2)_0$$

One-sided:

$$H_0 : \mu_1 - \mu_2 \leq (\mu_1 - \mu_2)_0 \text{ versus } H_A : \mu_1 - \mu_2 > (\mu_1 - \mu_2)_0$$

or

$$H_0 : \mu_1 - \mu_2 \geq (\mu_1 - \mu_2)_0 \text{ versus } H_A : \mu_1 - \mu_2 < (\mu_1 - \mu_2)_0$$

### Test Statistics

#### Populations with Known Variances

Let  $X_1, X_2, \dots, X_{n_1}$  denote i.i.d. random variables with mean  $\mu_1$  and standard deviation  $\sigma_1$ , and  $Y_1, Y_2, \dots, Y_{n_2}$  denote i.i.d. random variables with mean  $\mu_2$  and standard deviation  $\sigma_2$ . Then,

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \approx Z \sim N(0, 1)$$

To test

$$H_0 : \mu_1 - \mu_2 = 0$$

against some alternative, use the test statistic

$$Z = \frac{(\bar{X} - \bar{Y}) - (0)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Under the null hypothesis that  $\mu_1 = \mu_2$ ,  $Z$  has (approximately) a standard normal distribution. Evaluate the test statistic and compute the p-value using the table for a standard normal distribution.

#### Populations with Unknown but Equal Variances

Let  $X_1, X_2, \dots, X_{n_1}$  denote i.i.d. random variables with mean  $\mu_1$  and standard deviation  $\sigma_1$ , and  $Y_1, Y_2, \dots, Y_{n_2}$  denote i.i.d. random variables with mean  $\mu_2$  and standard deviation  $\sigma_2$ . Assume that we know  $\sigma_1 = \sigma_2 = \sigma$ , but we do not know the value of  $\sigma$ .

To test

$$H_0 : \mu_1 - \mu_2 = 0$$

against some alternative, use the test statistic

$$T = \frac{(\bar{X} - \bar{Y}) - (0)}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$$

Under the null hypothesis that  $\mu_1 = \mu_2$ ,  $T$  has a t-distribution (approximately) with  $n_1 + n_2 - 2$  degrees of freedom. Evaluate the test statistic and compute the p-value using the table for a t-distribution with  $n_1 + n_2 - 2$  degrees of freedom.

Estimation of the common variance.

The pooled estimate of the variance,  $s_p^2$ , combines information from both of the samples to produce a more reliable estimate of  $\sigma^2$ .

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

where

$$s_1^2 = \frac{\sum_{i=1}^{n_1} (x_i - \bar{x})^2}{n_1 - 1}$$

and

$$s_2^2 = \frac{\sum_{i=1}^{n_2} (y_i - \bar{y})^2}{n_2 - 1}$$