

Comparison of Two Population Means

Samples can be paired or independent.

Paired Samples: 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. In here, we often do not assume that we know the variance.

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

With paired data, we could consider the differences. Assume that the differences are normally* distributed. If the standard deviation of the differences is known (σ_d), consider

$$Z = \frac{\bar{D} - \delta}{\sigma_d/\sqrt{n}} \quad (1)$$

If the standard deviation of the differences is not known, consider

$$T = \frac{\bar{D} - \delta}{s_d/\sqrt{n}} \quad (2)$$

where \bar{D} is a random variable representing the mean of the differences $X_1 - X_2$, (\bar{d} is the sample mean of the differences) δ is the population mean of the differences, and s_d is the sample standard deviation of the differences. We know that (1) is standard normal and (2) has a t distribution with $n - 1$ degrees of freedom.

**Note that this holds approximately (by CLT) if variables are not Normally distributed (but n_1 is reasonably large). I will use * throughout this lecture to remind you that formulas hold exactly if Normality holds and approximately (by CLT for n reasonably large) if Normality does not hold.

Example (2) from last time: 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 (gmd).

Is grey matter density in the populations significantly different ?

Processed data from the 8 pairs is shown below (units not given).

Consider the population differences, $D = X_1 - X_2$,

Pair i	Twin 1 (x_{1i})	Twin 2 (x_{2i})	difference (d_i)
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

$D =$ gmd of first healthy twin - gmd of unhealthy twin

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

Decide on the type I error rate: $\alpha = 0.05$

If the null is true, we know that the distribution of

$$\frac{\bar{D} - 0}{s/\sqrt{n}}$$

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

In a sample of 8 differences, $\bar{d} = 4.875$ minutes and $s = 4.998$.

Therefore,

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

Calculate the p-value:

$$P(T_7 \leq -2.76) + P(T_7 \geq 2.76) = 0.014 + 0.014 = 0.028$$

Here, the p-value is 0.034.

The 95% confidence interval is:

$$(0.696, 9.05)$$

Recall and Note

Recall that if X_1, X_2, \dots, X_{n_1} denote Normally* distributed random variables mean μ_1 and standard deviation σ_1 ,

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

Note that if X_1, X_2, \dots, X_{n_1} denote Normally* distributed random variables with mean μ_1 and standard deviation σ_1 , and Y_1, Y_2, \dots, Y_{n_2} denote normally* distributed random variables with mean μ_2 and standard deviation σ_2 ,

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

and so

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

Populations with Unknown but Equal Variances

Let X_1, X_2, \dots, X_{n_1} denote normally* distributed random variables with mean μ_1 and standard deviation σ_1 , and Y_1, Y_2, \dots, Y_{n_2} denote normally* distributed random variables with mean μ_2 and standard deviation σ_2 . Assume that we know $\sigma_1 = \sigma_2 = \sigma$, but we do not know the value of σ .

To test

$$H_0 : \mu_1 - \mu_2 = (\mu_1 - \mu_2)_0$$

against some alternative, use the test statistic

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)_0}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$$

Under the null hypothesis that $\mu_1 = \mu_2$, T has a t-distribution 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.

Populations with Known Variances

Note this is not usually the case !! I show it here so we can figure out where the test statistic is coming from when the variance is NOT assumed to be known.

Let X_1, X_2, \dots, X_{n_1} denote normally* distributed random variables with mean μ_1 and standard deviation σ_1 , and Y_1, Y_2, \dots, Y_{n_2} denote normally* distributed random variables with mean μ_2 and standard deviation σ_2 . Then,

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

To test

$$H_0 : \mu_1 - \mu_2 = (\mu_1 - \mu_2)_0$$

against some alternative, use the test statistic

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

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

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 σ^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

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

Populations with Unknown and Possibly Unequal Variances

Let X_1, X_2, \dots, X_{n_1} denote normally* distributed random variables with mean μ_1 and standard deviation σ_1 , and Y_1, Y_2, \dots, Y_{n_2} denote normally* distributed random variables with mean μ_2 and standard deviation σ_2 . Assume that we cannot assume that $\sigma_1 = \sigma_2$ and assume that we do not know the values of σ_1 or σ_2 .

To test

$$H_0 : \mu_1 - \mu_2 = (\mu_1 - \mu_2)_0$$

against some alternative, use the test statistic

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

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

Here,

$$v = \frac{[(s_1^2/n_1) + (s_2^2/n_2)]^2}{\left[\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}\right]}$$

v is rounded down to the nearest integer.

11

An approximate two-sided $100(1 - \alpha)\%$ confidence interval for the true mean difference ($\delta = \mu_1 - \mu_2$) between two independent samples of size n_1 and n_2 , respectively, (again n_1 and n_2 large) with unknown and possibly unequal variances:

$$\left((\bar{X} - \bar{Y}) - t_{v, 1-\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, (\bar{X} - \bar{Y}) + t_{v, 1-\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right)$$

Confidence intervals for the true mean difference ($\delta = \mu_1 - \mu_2$) when the data are independent samples.

An approximate two-sided $100(1 - \alpha)\%$ confidence interval for the true mean difference ($\delta = \mu_1 - \mu_2$) between two independent samples of size n_1 and n_2 , respectively, with unknown, but equal, standard deviations $\sigma_1 = \sigma_2 = \sigma$:

$$\left((\bar{X} - \bar{Y}) - t_{(n_1+n_2-2), 1-\alpha/2} \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}, (\bar{X} - \bar{Y}) + t_{(n_1+n_2-2), 1-\alpha/2} \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} \right)$$

12

Recall the example where grey matter density (gmd) was measured in healthy and unhealthy twins. Pairing was reasonable in that example.

Suppose instead I told you that the measurements came from 8 randomly sampled female freshman (left column) and 8 randomly sampled female seniors (right column). We still want to know if the average gmd's are different between the groups, but there is no natural pairing to the data.

Index i	Freshman i	Senior i	difference i
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

We can assume that the variability for the two populations is equal, but we don't know what it is.

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

We 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.

Index i	Freshman i	Senior i	difference i
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

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

Here, $\bar{x} - \bar{y} = 4.875$, $s_1^2 = 371.98$, $s_2^2 = 253.43$, and $s_p^2 = 312.71$ ($s_p = 17.7$).

This gives

$$\frac{4.875}{\sqrt{312.71/8 + 312.71/8}} = 0.55$$

For $(n_1 + n_2 - 2) = 14$ degrees of freedom $t_{0.025} = 2.145$, so the null hypothesis is not rejected.

The 95% confidence interval is:

$$(-14.091, 23.841)$$

Why did we get different answers in the paired t-test and the unpaired t-test ?

1. In this particular (unpaired) example, there could be many factors causing differences between the independently sampled groups.
2. In the example with paired data, twin pairs were being studied. The pairs are very similar in a number of ways (genetically similar, environment might have been similar).
3. With unpaired data, many (unmeasured) factors could contribute to differences between the groups.
4. By pairing, we are controlling for many of these factors.

Hypothesis testing on Proportions

Recall that if X has a *Bernoulli distribution*, we can think of X as a success (with probability p) or failure (with probability $1 - p$).

If X is Bernoulli with success probability p , then

$$f_X(x) = \begin{cases} p^x(1-p)^{1-x} & \text{for } x = 0, 1 \\ 0 & \text{otherwise} \end{cases}$$

$$E[X] = p \text{ and } \text{var}[X] = p(1-p)$$

If X follows a *binomial distribution* with n trials and success probability p , we can think of X as the number of successes in n Bernoulli trials.

If X is Binomial with population parameters n and p ($n \geq 1$) and ($0 \leq p \leq 1$), then

$$f_X(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & \text{for } x = 0, 1, 2, \dots, n \\ 0 & \text{otherwise} \end{cases}$$

$$E[X] = np \text{ and } \text{var}[X] = np(1-p)$$

We are often interested in the proportion of times that an event occurs, rather than the number of times. We are interested in knowing the true value of p . How should we estimate p ?

Suppose we are interested in a group of patients diagnosed with lung cancer. We would like to know what proportion of these patients survive 5 years after diagnosis. We can draw a random sample of size n patients diagnosed with lung cancer and count how many are alive (out of n) after 5 years. Think of each patient as a Bernoulli trial with p , the probability of being alive after 5 years.

What is the best estimate of p , the proportion of interest?

Suppose in the context of the lung cancer example above, we would like to test

$$H_0 : p = 0.082 \text{ against } H_A : p \neq 0.082$$

We identify 52 patients that were diagnosed with lung cancer 5 years ago. We confirm that 6 are still alive. Therefore, we calculate $\hat{p} = 0.115$. Conduct the test.

What is the 99% confidence interval for the true proportion?

$$Z = \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}} \text{ is approximately } N(0, 1)$$

(I have capitalized the P in \hat{P} to stress that it is a random variable. This is not standard in this context, and will not be continued).

An approximate $100(1 - \alpha)$ % confidence interval for p :

$$\left(\hat{p} - z_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

If we count the number of people alive and divide by the total number n , this serves as the best estimate of p .

Let

$$\hat{P} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

denote the *sample proportion*. As before, we would like to address questions about p , given \hat{P} .

Recall that if X is Bernoulli with success probability p , then

$$E[X] = p \text{ and } \text{var}[X] = p(1 - p).$$

The confidence interval on the previous page is approximate for two reasons:

1. The distribution of the test statistic is approximately normal.
2. We do not know the standard error, so we use $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. To be consistent with distinctions we have made previously, we should use $t_{1-\alpha/2}$ instead of $z_{1-\alpha/2}$ here (since we are estimating the standard error).

Suppose we would like to test

$$H_0 : p_1 = p_2 \text{ against } H_A : p_1 \neq p_2,$$

We collect a sample of size n_1 from the first population and a sample of size n_2 from the second population. Under the null, the estimate of the standard error of the difference $p_1 - p_2$ takes the form

$$\sqrt{\hat{p}(1 - \hat{p}) \left[\frac{1}{n_1} + \frac{1}{n_2} \right]}$$

$$\text{where } \hat{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$$

The appropriate test statistic is

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}(1 - \hat{p}) \left[\frac{1}{n_1} + \frac{1}{n_2} \right]}}$$

If n_1 and n_2 are large, this statistic is approximately standard normal.

How large is large for tests on proportions ?

Commonly used criterion:

One sample case: $n\hat{p}$ and $n(1 - \hat{p})$ are both greater than 5.

Two sample case: $n_1\hat{p}$, $n_1(1 - \hat{p})$, $n_2\hat{p}$, and $n_2(1 - \hat{p})$ are each greater than 5.

Suppose we want to study the effectiveness of bike helmets in preventing head injury. We collect two random samples: one of size 147 from a population of people that wear helmets and the other of size 646 from a population of people that do not wear helmets. We record that 17 of the 147 suffered a serious head injury and 218 of the 646 suffered a serious head injury. We want to know if the proportion of serious head injuries is the same in the two populations.

$$n_1 = 147, n_2 = 646, \hat{p}_1 = \frac{17}{147} = 0.116, \hat{p}_2 = \frac{218}{646} = 0.337$$

The appropriate test statistic is

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}(1 - \hat{p}) \left[\frac{1}{n_1} + \frac{1}{n_2} \right]}}$$

$$\text{where } \hat{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2} = 0.296$$

The evaluated test statistic is

$$z = \frac{(0.116 - 0.337)}{\sqrt{0.296(1 - 0.296) \left[\frac{1}{147} + \frac{1}{646} \right]}} = -5.3$$

$$\begin{aligned} \text{p-value} &= P(Z \leq -5.3) + P(Z \geq 5.3) \\ &= 5.8 \cdot 10^{(-8)} + 5.8 \cdot 10^{(-8)} \\ &= 1.16 \cdot 10^{(-7)} \end{aligned}$$

The null is rejected at significance level $\alpha = 0.01$.

We could have approached this same question a different way.

Wearing Helmet			
Head Injury	(Y)	(N)	total
+ (Y)	17	218	235
- (N)	NA	NA	NA
total	147	646	793

Considering that we know the totals, we can figure out what missing numbers are:

Wearing Helmet			
Head Injury	(Y)	(N)	total
+ (Y)	17	218	235
- (N)	130	428	558
total	147	646	793

What would we expect this table to look like if the null was true ?