# Confidence Interval of Difference of Two Indepednent Population Means and Proportions

Cheng Peng

## Contents

## 1 Introduction

This note focuses on converting the hypotheses on the difference between two independent population means and proportions. The building block of these confidence intervals was developed in the previous two modules (two-sample tests).

The same five-step procedure is used in constructing confidence intervals for population means and proportions:

**Step 1**: Identify the confidence level $1 - \alpha$ and find $\alpha/2$ (right tail area) for the critical value from the appropriate table.

**Step 2**: Check conditions to identify appropriate table and find the **critical value**.

**Step 3**: Find the margin of error $E = CV \times$ standard error of $\bar{x}_1 - \bar{x}_2$ or $\hat{p}_1 - \hat{p}_2$. E for the specific confidence interval will be detailed in subsequent sections.

**Step 4**: Write the confidence interval explicitly: $(\bar{x}_1 - \bar{x}_2) \pm E$ for population means and $(\hat{p}_1 - \hat{p}_2) \pm E$.

**Step 5**: Interpret the confidence interval.

In the next few sections, we will briefly discuss the confidence interval of the difference of two population means and proportions under various assumptions.

## 2 Confidence Intervals for Two Population Means

As we already know that constructing a confidence interval requires a certain amount of information contained in the random sample and/or certain assumptions about the population. In the subsequent subsections, we will discuss two types of confidence intervals with different assumptions.

## 2.1 Normal Confidence Intervals

There are two major scenarios where a normal confidence interval is appropriate under different conditions.

### 2.1.1 Scenario 1 - Case of Large Samples

The basic assumption is that both sample sizes are large, say $n_1 > 30$ and $n_2 > 30$ (convention adopted in this class). With this assumption, we know that, by using the Central Limit Theorem (CLT),

$$(\bar{x}_1 - \bar{x}_2) \to N\left(\mu_1 - \mu_2, \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right)$$

This means that the critical value for the given confidence level will be based on the standard normal distribution table, and the standard deviation is given by

$$\text{var}(\bar{x}_1 - \bar{x}_2) \approx \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

Therefore, the margin of error of the confidence interval of $\mu_1 - \mu_2$ is given by

$$E = Z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

Hence, the explicit expression of $100(1 - \alpha)\%$ is given by: $(\bar{x}_1 - \bar{x}_2) \pm E$.

**Example 1**: A study compares the average test scores of students from two different teaching methods (Method A and Method B). Two independent samples were taken from each group.
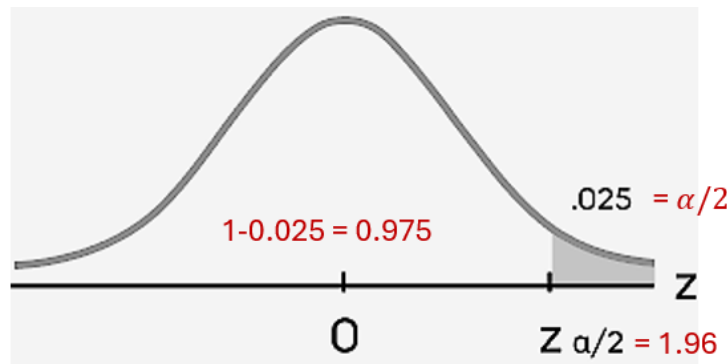
| Sample 1 (Method A) | Sample 2 (Method B) |
|---|---|
| $\bar{X}_1 = 78$ | $\bar{X}_2 = 82$ |
| $s_1 = 10$ | $s_2 = 12$ |
| $n_1 = 50$ | $n_2 = 100$ |

Construct a 95% confidence interval for the difference in population means $(\mu_1 - \mu_2)$.

**Solution** Use the 5-step procedure to construct the confidence interval.

**Step 1**: Since the confidence level 95%, that is, $1 - \alpha = 0.95$ or $\alpha/2 = 0.025$.

**Step 2**: Since both sample sizes are larger than 30, the critical value CV $= Z_{0.025} = 1.96$.

**Step 3**: The margin of error E is

$$E = CV \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 1.96\sqrt{\frac{10^2}{50} + \frac{12^2}{100}} \approx 1.96\sqrt{3.44} \approx 3.64.$$

**Step 4**: 95% confidence interval for $\mu_1 - \mu_2$ is

$$(\bar{x}_1 - \bar{x}_2) \pm E = (78 - 82) \pm 3.64 = (-7.64, -0.36).$$

**Step 4**: We are 95% confident that the true difference in average test scores (Method A – Method B) lies between $-7.64$ and $-0.36$.

Since the **interval does not include 0**, we conclude that there is a statistically significant difference between the two methods at the 95% confidence level.

### 2.1.2   Scenario 2 - Case of Normal Populations with Known Variances

In this scenario, we assume that both populations are normally distributed with respective variances denoted by $\sigma_1^2$ and $\sigma_2^2$. Under these assumptions, there are no restrictions on the sample sizes. To be more specific

$$(\bar{x}_1 - \bar{x}_2) \to N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

Note that this result is not derived from the CLT. Rather, it represents an exact distribution obtained from the properties of normal random variables. Using this sampling distribution, we can define the margin of error in the following way.

$$E = Z_{\alpha/2} \times \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

The steps for constructing a confidence interval for the difference between two population means are the same as those used in Scenario 1.

**Example 2**: A drug company wants to compare the mean blood pressure reduction (in mmHg) between its new medication (Drug A) and the current standard (Drug B). Assume that blood pressure reductions for both drugs follow normal distributions.

Population variances are known from historical data:

- Drug A: $\sigma_1^2 = 25$ mmHg$^2$
- Drug B: $\sigma_2^2 = 16$ mmHg$^2$

The sample information of the two samples is summarized in the following table.

| Drug A (new) | Drug B |
|---|---|
| $\bar{X}_1 = 12$ | $\bar{X}_2 = 10$ |
| $n_1 = 50$ | $n_2 = 25$ |

Construct a 95% confidence interval for the difference of the two normal population means.

**Solution**: We still follow the 5-step procedure.

**Step 1**: The confidence level $1 - \alpha = 0.95$ which gives $\alpha/2 = 0.025$.

**Step 2**: The critical value is the same: $Z_{0.025} = 1.96$.

**Step 3**: The margin of error is given by

$$E = Z_{0.025}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = 1.96\sqrt{\frac{25}{50} + \frac{16}{25}} = 2.09.$$

**Step 4** The explicit confidence interval is given by

$$(\bar{x}_1 - \bar{x}_2) \pm E = (12 - 10) \pm 2.09 = (-0.09, 4.09)$$

**Step 5**: We are 95% confident that the true mean difference in blood pressure reduction (Drug A - Drug B) lies between -0.09 mmHg and 4.09 mmHg. Since the interval includes 0, Drug A is not statistically significantly more effective than Drug B at 95% confidence.

## 2.2 t-Confidence Intervals

We still assume normality of the two populations with **unknown but equal** variances. If both sample sizes are large, we can use the central limit theorem and construct a normal confidence interval. We only discuss the case where at least one of the sample sizes is small.

Since the two unknown variances are equal, we need to combine the two samples and estimate the common variance (pooled variance) as we did in the two-sample test

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

Under the above assumption, the sampling distribution $(\bar{x}_1 - \bar{x}_2)$ is given

$$\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}} \rightarrow t_{n_1+n_2-2}$$

The critical value to be used in the confidence interval is from the t-table. Therefore, the margin of error is given by

$$E = t_{\alpha/2,n_1+n_2-2} \times \sqrt{\frac{s_{\text{pool}}^2}{n_1} + \frac{s_{\text{pool}}^2}{n_2}} = t_{\alpha/2,n_1+n_2-2} \times s_{\text{pool}}\sqrt{1/n_1 + 1/n_2}.$$

With the above building blocks, we can construct a confidence interval of the difference of two normal population means as usual.

**Example 3**: A tech company wants to compare the mean battery life (in hours) of two smartphone models (Model X and Model Y) produced under the same conditions. Assume battery life for both models is normally distributed, and population variances are unknown but assumed equal (e.g., same battery technology). To see whether the lifetimes of the two batteries, we construct a 95% confidence interval of the difference of the lifetimes based on the following two independent samples.

| Model X | Model Y |
|---|---|
| $\bar{X}_1 = 15.2$ | $\bar{X}_2 = 14.1$ |

|          | Model X      | Model Y      |
|----------|--------------|--------------|
|          | $s_1 = 1.8$  | $s_2 = 1.5$  |
|          | $n_1 = 25$   | $n_2 = 20$   |

**Solution**: We construct the confidence interval based on the same 5-step procedure.

**Step 1**: The confidence level is $1 - \alpha = 0.95$ which gives $\alpha/2 = 0.025$.

**Step 2**: The critical value based on the t-table

$$CV = t_{0.025,25+20-2} = t_{0.025,43} = 2.021.$$

## T-Table

| df | 0.50 | 0.25 | 0.20 | 0.15 | 0.10 | 0.05 | 0.025 | 0.01 | 0.005 | 0.001 | 0.0005 |
|------|-------|-------|-------|-------|-------|-------|-------|--------|--------|---------|---------|
| 1 | 0.000 | 1.000 | 1.376 | 1.963 | 3.078 | 6.314 | 12.71 | 31.82 | 63.66 | 318.31 | 636.62 |
| 2 | 0.000 | 0.816 | 1.061 | 1.386 | 1.886 | 2.920 | 4.303 | 6.965 | 9.925 | 22.327 | 31.599 |
| 3 | 0.000 | 0.765 | 0.978 | 1.250 | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 | 10.215 | 12.924 |
| 4 | 0.000 | 0.741 | 0.941 | 1.190 | 1.533 | 2.132 | 2.776 | 3.747 | 4.604 | 7.173 | 8.610 |
| 5 | 0.000 | 0.727 | 0.920 | 1.156 | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 | 5.893 | 6.869 |
| 6 | 0.000 | 0.718 | 0.906 | 1.134 | 1.440 | 1.943 | 2.447 | 3.143 | 3.707 | 5.208 | 5.959 |
| 7 | 0.000 | 0.711 | 0.896 | 1.119 | 1.415 | 1.895 | 2.365 | 2.998 | 3.499 | 4.785 | 5.408 |
| 8 | 0.000 | 0.706 | 0.889 | 1.108 | 1.397 | 1.860 | 2.306 | 2.896 | 3.355 | 4.501 | 5.041 |
| 9 | 0.000 | 0.703 | 0.883 | 1.100 | 1.383 | 1.833 | 2.262 | 2.821 | 3.250 | 4.297 | 4.781 |
| 10 | 0.000 | 0.700 | 0.879 | 1.093 | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 | 4.144 | 4.587 |
| 11 | 0.000 | 0.697 | 0.876 | 1.088 | 1.363 | 1.796 | 2.201 | 2.718 | 3.106 | 4.025 | 4.437 |
| 12 | 0.000 | 0.695 | 0.873 | 1.083 | 1.356 | 1.782 | 2.179 | 2.681 | 3.055 | 3.930 | 4.318 |
| 13 | 0.000 | 0.694 | 0.870 | 1.079 | 1.350 | 1.771 | 2.160 | 2.650 | 3.012 | 3.852 | 4.221 |
| 14 | 0.000 | 0.692 | 0.868 | 1.076 | 1.345 | 1.761 | 2.145 | 2.624 | 2.977 | 3.787 | 4.140 |
| 15 | 0.000 | 0.691 | 0.866 | 1.074 | 1.341 | 1.753 | 2.131 | 2.602 | 2.947 | 3.733 | 4.073 |
| 16 | 0.000 | 0.690 | 0.865 | 1.071 | 1.337 | 1.746 | 2.120 | 2.583 | 2.921 | 3.686 | 4.015 |
| 17 | 0.000 | 0.689 | 0.863 | 1.069 | 1.333 | 1.740 | 2.110 | 2.567 | 2.898 | 3.646 | 3.965 |
| 18 | 0.000 | 0.688 | 0.862 | 1.067 | 1.330 | 1.734 | 2.101 | 2.552 | 2.878 | 3.610 | 3.922 |
| 19 | 0.000 | 0.688 | 0.861 | 1.066 | 1.328 | 1.729 | 2.093 | 2.539 | 2.861 | 3.579 | 3.883 |
| 20 | 0.000 | 0.687 | 0.860 | 1.064 | 1.325 | 1.725 | 2.086 | 2.528 | 2.845 | 3.552 | 3.850 |
| 21 | 0.000 | 0.686 | 0.859 | 1.063 | 1.323 | 1.721 | 2.080 | 2.518 | 2.831 | 3.527 | 3.819 |
| 22 | 0.000 | 0.686 | 0.858 | 1.061 | 1.321 | 1.717 | 2.074 | 2.508 | 2.819 | 3.505 | 3.792 |
| 23 | 0.000 | 0.685 | 0.858 | 1.060 | 1.319 | 1.714 | 2.069 | 2.500 | 2.807 | 3.485 | 3.768 |
| 24 | 0.000 | 0.685 | 0.857 | 1.059 | 1.318 | 1.711 | 2.064 | 2.492 | 2.797 | 3.467 | 3.745 |
| 25 | 0.000 | 0.684 | 0.856 | 1.058 | 1.316 | 1.708 | 2.060 | 2.485 | 2.787 | 3.450 | 3.725 |
| 26 | 0.000 | 0.684 | 0.856 | 1.058 | 1.315 | 1.706 | 2.056 | 2.479 | 2.779 | 3.435 | 3.707 |
| 27 | 0.000 | 0.684 | 0.855 | 1.057 | 1.314 | 1.703 | 2.052 | 2.473 | 2.771 | 3.421 | 3.690 |
| 28 | 0.000 | 0.683 | 0.855 | 1.056 | 1.313 | 1.701 | 2.048 | 2.467 | 2.763 | 3.408 | 3.674 |
| 29 | 0.000 | 0.683 | 0.854 | 1.055 | 1.311 | 1.699 | 2.045 | 2.462 | 2.756 | 3.396 | 3.659 |
| 30 | 0.000 | 0.683 | 0.854 | 1.055 | 1.310 | 1.697 | 2.042 | 2.457 | 2.750 | 3.385 | 3.646 |
| 40 | 0.000 | 0.681 | 0.851 | 1.050 | 1.303 | 1.684 | 2.021 | 2.423 | 2.704 | 3.307 | 3.551 |
| 60 | 0.000 | 0.679 | 0.848 | 1.045 | 1.296 | 1.671 | 2.000 | 2.390 | 2.660 | 3.232 | 3.460 |
| 80 | 0.000 | 0.678 | 0.846 | 1.043 | 1.292 | 1.664 | 1.990 | 2.374 | 2.639 | 3.195 | 3.416 |
| 100 | 0.000 | 0.677 | 0.845 | 1.042 | 1.290 | 1.660 | 1.984 | 2.364 | 2.626 | 3.174 | 3.390 |
| 1000 | 0.000 | 0.675 | 0.842 | 1.037 | 1.282 | 1.646 | 1.962 | 2.330 | 2.581 | 3.098 | 3.300 |
| z | 0.000 | 0.674 | 0.842 | 1.036 | 1.282 | 1.645 | 1.960 | 2.326 | 2.576 | 3.090 | 3.291 |

**Step 3**: To find the margin of error, we first estimate the pooled variance.

$$s_{\text{pool}}^2 = \frac{(25-1)1.8^2 + (20-1)1.5^2}{25 + 20 - 2} = 2.80$$

is evaluated in the following

$$E = t_{\alpha/2,n_1+n_2-2} \times s_{\text{pool}}\sqrt{1/n_1 + 1/n_2} = 2.201 \times \sqrt{2.80}\sqrt{1/25 + 1/20} \approx 1.01$$

5

**Step 4**: The confidence interval of $\mu_1 - \mu_2$ is given by

$$(\bar{x}_1 - \bar{x}_2) \pm E = (15.2 - 14.1) \pm 1.01 = (0.09, 2.11).$$

**Step 5**: We are 95% confident that the true mean difference in battery life (Model X - Model Y) is between 0.09 and 2.11 hours. The interval excludes 0. This implies that Model X likely outperforms Model Y statistically.

# 3 Confidence Interval of the Difference of Two Population Proportions

Similar to one-sample proportion confidence interval and testing hypothesis, there are analogous assumptions to ensure a good approximation (using CLT). For two-sample proportion confidence intervals, the following assumptions should be checked before constructing the confidence interval.

- Independent random samples from two populations.
- Large sample sizes (for normal approximation)
    - $n_1 \hat{p}_1 \geq 5$ and $n_1(1 - \hat{p}_1) \geq 5$
    - $n_2 \hat{p}_2 \geq 5$ and $n_2(1 - \hat{p}_2) \geq 5$

With the above assumptions, we have the following sampling distribution.

$$\hat{p}_1 - \hat{p}_2 \to N\left(p_1 - p_2, \sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}}\right).$$

Since the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is approximated by the central limit theorem, the critical value should be based on the standard normal distribution table. The margin of error is given by

$$E = Z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}.$$

The steps for constructing the confidence interval of $p_1 - p_2$ are the same as those for constructing other confidence intervals.

**Example 4**: A pharmaceutical company wants to compare the effectiveness of two new drugs (Drug A and Drug B) for treating high blood pressure. In a clinical trial:

- Drug A was given to 200 patients, and 140 showed improvement.

- Drug B was given to 150 patients, and 90 showed improvement.

The company wants to estimate the difference in the true proportions of improvement between the two drugs with 95% confidence.

**Solution**: Since the sample proportions are not explicitly given, we first calculate these sample proportions in the following

- $\hat{p}_1 = X_1/n_1 = 140/200 = 0.70$
- $\hat{p}_2 = X_2/n_2 = 90/150 = 0.60$

**Caution**: *proportions in formulas above must be expressed in the decimal form*

**Step 1**: The significance level $1 - \alpha = 0.95$ which gives $\alpha/2 = 0.025$.

**Step 2**: The critical value can be found from the normal table $Z_{0.025} = 1.96$ since the following conditions

- $n_1 \hat{p}_1 = 200 \times 0.7 = 140 \geq 5$ and $n_1(1 - \hat{p}_1) = 200 \times 0.3 = 60 \geq 5$

- $n_2\hat{p}_2 = 150 \times 0.6 = 90 \geq 5$ and $n_2(1 - \hat{p}_2) = 150 \times 0.4 = 60 \geq 5$

**Step 3**: The margin of error is given by

$$E = Z_{0.025}\sqrt{\frac{0.7(1 - 0.7)}{200} + \frac{0.6(1 - 0.6)}{150}} \approx 0.1009$$

**Step 4**: The confidence interval for $p_1 - p_2$ is

$$(\hat{p}_1 - \hat{p}_2) \pm E = (0.7 - 0.6) \pm 0.1009 = (-0.009, 0.2009)$$

**Step 5**: We are 95% confident that the true difference in improvement rates (Drug A – Drug B) lies between 0% and 20%. Since the interval includes 0, we cannot conclusively say that Drug A is better than Drug B at the 95% confidence level (though the data suggests a possible advantage).

# 4 Practice Exercises

**Problem 1**. In order to investigate the relationship between mean job tenure in years among workers who have a bachelor's degree or higher and those who do not, random samples of each type of worker were taken, with the following results.

| population | n | $\bar{X}$ | s |
|---|---|---|---|
| Bachelor's degree or higher | 155 | 5.2 | 1.3 |
| No degree | 210 | 5.0 | 1.5 |

Construct the 99% confidence interval for the difference in the population means based on these data.

**Problem 2**. Records of 49 used passenger cars and 36 used pickup trucks (none used commercially) were randomly selected to investigate whether there was any difference in the mean time in years that they were kept by the original owner before being sold. For cars, the mean was 5.3 years with a standard deviation of 2.2 years. For pickup trucks, the mean was 7.1 years with a standard deviation of 3.0 years.

Construct the 95% confidence interval for the difference in the means based on these data.

**Problem 3**. The coaching staff of a professional football team believes that the rushing offense has become increasingly potent in recent years. To investigate this belief, 20 randomly selected games from one year's schedule were compared to 11 randomly selected games from the schedule five years later. The sample information on rushing yards per game (rypg) is summarized below.

| Sample Type | n | $\bar{x}$ | s |
|---|---|---|---|
| rypg previously | 20 | 112 | 24 |
| rypg recently | 11 | 114 | 21 |

Assume that both populations are normally distributed and with equal variance. Construct the 95/ confidence interval for the difference in the population means based on these data.

**Problem 4**. A simple random sample is taken of front-seat occupants involved in car crashes. Of the 2823 occupants not wearing seat belts, 31 were killed. Among the 7765 occupants wearing seat belts, 16 were killed. Construct a 90% confidence interval estimate of the difference between the fatality rates for those not wearing seat belts and those wearing seat belts. What does the result suggest about the effectiveness of seat belts?