Estimation of $\mu_1 - \mu_2$

Motivation: Two diet programs, one tried on 100 people and the other tried on another 100 people, give the following data:

	Diet 1	Diet 2
Sample size	$n_1 = 100$	$n_2 = 100$
Weight loss	$\bar{x}_1 = 9.3$	$\bar{x}_2 = 3.7$
Sample variance	$s_1^2 = 22.4$	$s_2^2 = 16.3$

What's the difference of the two programs?

1

Let $X_{11}, X_{12}, X_{13}, \ldots, X_{1n_1}$ be an SRS from population 1 with μ_1 *and* σ_1^2.
Let $X_{21}, X_{22}, X_{23}, \ldots, X_{2n_2}$ be an SRS from population 2 with μ_2 *and* σ_2^2.
Then, when n_1 and n_2 are large, we have

$$\bar{X}_1 \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right), \qquad \bar{X}_2 \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right).$$

If the two samples are independent, then

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, \ 1).$$

2

It follows that a large sample $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

if σ_1, σ_2 *are known, or*

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

if σ_1, σ_2 *are unknown.*

3

Example 1.11

Let μ_1 and μ_2 be the mean weight losses in the diet programs. If $1 - \alpha = 0.95$, we find $\alpha = 0.05$ and $z_{0.05/2} = z_{0.025} = 1.96$. Then a 95% confidence interval for $\mu_1 - \mu_2$ is

$$(9.3 - 3.7) \pm 1.96\sqrt{\frac{22.4}{100} + \frac{16.3}{100}} = 5.6 \pm 1.22$$

or (4.38, 6.82).

4

If the sample sizes n_1 and n_2 are small, we need to make further assumptions to estimate $\mu_1 - \mu_2$. Specifically, we assume that the two SRS's are (1) independent, (2) normally distributed, and (3) $\sigma_1^2 = \sigma_2^2$, then we pool the data from the two samples to estimate the common variance by

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

5

It can be shown that under the above assumptions, we have

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2},$$

and a small sample $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2,\ n_1+n_2-2}\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

6

Example 1.12

A new method and a standard method are used to teach 20 slow learners. The scores of a reading test gives

New	Standard
$n_1 = 8$	$n_2 = 12$
$\bar{x}_1 = 76.9$	$\bar{x}_2 = 72.7$
$s_1 = 4.85$	$s_2 = 6.35$

Estimate the mean score difference between the two methods with a 90% C.I.

Solution.

Since $\alpha = 0.1$, $\alpha/2 = 0.05$. From $n_1 = 8$, $n_2 = 12$, we find $t_{\alpha/2,\ n_1+n_2-2} = t_{0.05,18} = 1.734$.

The pooled sample variance is

$$s_p^2 = \frac{(8-1)(4.85)^2 + (12-1)(6.35)^2}{8+12-2} = 33.79.$$

So a 90% C.I. for $\mu_{new} - \mu_{standard}$ is

$$(76.9 - 72.7) \pm 1.734\sqrt{33.79\left(\frac{1}{8}+\frac{1}{12}\right)} = 4.20 \pm 4.60$$

or (–0.40, 8.80).

When we say n_1 and n_2 are small, we mean $n_1 < 30$ or $n_2 < 30$. For small samples, it is not easy to know $\sigma_1^2 = \sigma_2^2$ holds or not, so deciding whether to pool or not is a problem of practical importance. There are formal ways to test whether $\sigma_1^2 = \sigma_2^2$ is true. We, however, take an alternative approach.

Approximate Small Sample Confidence Interval when $\sigma_1^2 \neq \sigma_2^2$:

If $n_1 = n_2 = n$, an approximate $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2,\,2n-2}\sqrt{\frac{s_1^2}{n}+\frac{s_2^2}{n}}$$

$(n + n - 2 = 2n - 2$ is the degrees of freedom)

If $n_1 \neq n_2$, an approximate $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2,\nu}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}},$$

where the degrees of freedom ν is given by

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

Round ν down to the nearest integer.

Example 1.13

A new and an old method are used to store snap beans to study which method keeps ascorbic acid better. This gives

	New method	Old method
Sample size	$n_1 = 8$	$n_2 = 10$
Mean ascorbic acid	449	410
Standard deviation	19	45

Estimate the difference in the mean ascorbic acid levels of the two storage methods with a 90% confidence interval.

Solution. Since $n_1 \neq n_2$, we find ν as

$$\nu = \frac{\left(19^2/8 + 45^2/10\right)^2}{\frac{\left(19^2/8\right)^2}{8-1} + \frac{\left(45^2/10\right)^2}{10-1}} = 12.65 \approx 12.$$

From $t_{\alpha/2,\nu} = t_{0.05,12} = 1.782$, a 90% confidence interval for $\mu_{new} - \mu_{old}$ is

$$(449-410) \pm 1.782\sqrt{\frac{19^2}{8} + \frac{45^2}{10}} = 39 \pm 28.04$$

or (11.0, 67.0).

13

A simpler working rule is: if $0.5 \le s_1/s_2 \le 2$, it is reasonable to assume $\sigma_1^2 = \sigma_2^2$. If $s_1/s_2 <$ 0.5 or $s_1/s_2 > 2$, the following conservative small sample 100(1– α)% confidence interval for $\mu_1 - \mu_2$ is recommended:

$$\left(\bar{x}_1 - \bar{x}_2\right) \pm t_{\alpha/2,\, n^*}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},$$

where n* is the smaller of $n_1 - 1$ and $n_2 - 1$.

14

Example 1.13 continued

Since $s_1/s_2 = 19/45 = 0.42$, we don't pool. As the smaller of 8 – 1 and 10 – 1 is 7, we have n* = 7 and $t_{\alpha/2,\, n^*} = t_{0.05,\, 7} = 1.895$. A 90% confidence interval for $\mu_{new} - \mu_{old}$ is

$$(449-410) \pm 1.895\sqrt{\frac{19^2}{8} + \frac{45^2}{10}} = 39 \pm 29.82$$

or (9.2, 68.8). (vs (11.0, 67.0))

15

The above discussions (large or small samples) are based on the assumption that the two SRS's are independent. This is, however, not always the case. For example, a training program is designed to help reduce the cholesterol levels of some patients. Eleven patients participated in the training program giving the following data:

16

Cholesterol Levels			
Subject	Before T	After T	Difference
1	182	198	-16
2	232	210	22
3	191	194	-3
4	200	220	-20
5	148	138	10
6	249	220	29
7	276	219	57
8	213	161	52
9	241	210	31
10	480	313	167
11	262	226	36

17

In this case, each patient's cholesterol level is measured before (X_1) and after training (X_2). Therefore, the two samples are matched and are NOT independent. We note that the magnitudes of the values within a sample are of little interest (Why?), but the differences, $X_1 - X_2$, for the 11 patients are indicating whether the training program is effective.

18

In general, let $D_1, D_2, \ldots, D_n$ be the n differences obtained from subtracting sample 2 from sample 1 for two matched samples. Let $\bar{x}_D$ and s_D be the sample mean and sample standard deviation for the D_i's. Then a $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is

$$\left(\bar{x}_D - t_{\alpha/2,n-1}\frac{s_D}{\sqrt{n}}, \quad \bar{x}_D + t_{\alpha/2,n-1}\frac{s_D}{\sqrt{n}}\right).$$

19

Example 1.14

For the cholesterol data, we have $\bar{x}_D = 33.18$ and $s_D = 51.07$. If a 90 confidence interval is desired, we find

$$t_{\alpha/2,\ n-1} = t_{0.1/2,\ 11-1} = t_{0.05,\ 10} = 1.812,$$

and

$$\left(33.18 - 1.812\frac{51.07}{\sqrt{11}}, \quad 33.18 + 1.812\frac{51.07}{\sqrt{11}}\right)$$

or (5.28, 61.08).

20

The fact that the lower end point of the above 90% confidence interval is quite a bit above zero indicates that the training program is successful (at 90% level).

21

Estimation of $p_1 - p_2$

Motivation: A survey of smoking habit results in the following data:

	1985	1990
Sample size	$n_1 = 1500$	$n_2 = 2000$
# of smokers	$x_1 = 576$	$x_2 = 652$

What is the difference between the population proportions of smokers in 1985 and 1990?

22

Let $\bar{x}_1 = \hat{p}_1, \quad \bar{x}_2 = \hat{p}_2,$

$$s_1^2 = \hat{p}_1(1-\hat{p}_1), \quad s_2^2 = \hat{p}_2(1-\hat{p}_2).$$

Then the results for $\mu_1 - \mu_2$ carry over here for $p_1 - p_2$, that is, a $100(1-\alpha)\%$ confidence interval for $p_1 - p_2$ is

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

23

The above confidence interval is approximately valid if

$$n_1\hat{p}_1 \geq 15, \quad n_1\hat{q}_1 \geq 15,$$
$$n_2\hat{p}_2 \geq 15, \quad n_2\hat{q}_2 \geq 15.$$

24

Example 1.15

For the smoking survey data, we have

$$\hat{p}_1 = \frac{576}{1500} = 0.384, \quad \hat{p}_2 = \frac{652}{2000} = 0.326.$$

If a 90% confidence interval is desired, $\alpha = 0.10$, $z_{\alpha/2} = z_{0.1/2} = z_{0.05} = 1.645$, and

$$(0.384 - 0.326) \pm 1.645\sqrt{\frac{0.384(1-0.384)}{1500} + \frac{0.326(1-0.326)}{2000}}$$

or 0.058 ± 0.0269, *or* $(0.031, \quad 0.085)$.

25

Practice Problems

9.1, 9.4, 9.5, 9.7, 9.8, 9.9 (a, c), 9.17 (use C.I.), 9.19, 9.21

9.30, 9.31, 9.34 (d), 9.39, 9.42 (c: -0.56 ± 1.44)

9.50, 9.51, 9.52, 9.53, 9.56 (d), 9.59, 9.64 (0.0074 ± 0.03153)

26