Compare two means

We have two SRSs, from two distinct populations.
The samples are independent.
sample sizes n_1, n_2.
sample means $\bar{x}_1, \bar{x}_2$.
sample standard deviation: s_1, s_2.
We want to compare two population means μ_1 and μ_2: Are they equal or not? What is the size of $\mu_1 - \mu_2$?

Section 9.1

Conditions: SRS from each population.
Two samples are independent.
Each sample size is large (> 30) or its population distribution is normal.

Confidence interval for $\mu_1 - \mu_2$:
$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
Here $t_{\alpha/2}$ depends on the confidencel level and the degrees of freedom.
Simple way to compute d.f.= $\min(n_1 - 1, n_2 - 1)$.,i.e., d.f.=the smaller sample size -1.
Still use the t chart to find the t critical value as before.
Significance test: $H_0 : \mu_1 = \mu_2$ or $H_0 : \mu_1 - \mu_2 = 0$.
t statistic=$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$.
Then use the t chart to find the p-value based on t test statistic value

CI Example

9.3. A sample of 15 patients took a new drup designed to reduce blood pressure. Their blood pressures are reduced by 28.3 mmHg on average with a sample standard deviation of 12.0 mmHg. Another sample of 20 patients took a standard drug and their blood pressure are reduced by an average of 17.1 mmHg with a standard deviation of 9.0 mmHg. Find a 95% CI for $\mu_1 - \mu_2$, the difference between two population mean reductions.

Summary of data:
$\bar{x}_1 = 28.3, s_1 = 12, n_1 = 15, \bar{x}_2 = 17.1, s_2 = 9, n_2 = 20.$
d.f. = 15-1=14, $t_{0.025}$= 2.145.
95% CI: $28.3 - 17.1 \pm 2.145\sqrt{\frac{12.0^2}{15} + \frac{9.0^2}{20}}$
$= 11.2 \pm 7.925 = (3.3, 19.1).$

Two sample t test Example

9.2. Two methods to treat wastewater are compared. Treatment 1 is applied to 5 specimens of wastewater and treatment 2 is applied to 7 specimens. The benzene concentrations for each specimen are given below:
Treatment 1: 7.8, 7.6, 5.6, 6.8, 6.4
Treatment 2: 4.1, 6.5, 3.7, 7.7, 7.3, 4.7, 5.9
How we claim the mean concentration for treatment 2, μ_2 is less than that for treatment 1, μ_1? Use $\alpha = 0.05$.

Two sample t test example

Summary of data:
$\bar{x}_1 = 6.84, s_1 = 0.8989, n_1 = 5, \bar{x}_2 = 5.70, s_2 = 1.5706, n_2 = 7$
(computed from sample data)
$H_0 : \mu_1 = \mu_2$
$H_1 : \mu_1 > \mu_2$ (implied in the problem)
t = $\frac{6.84 - 5.70}{\sqrt{\frac{0.8989^2}{5} + \frac{1.5706^2}{7}}} = 1.590.$
d.f.= 4.
p-value = $P(t > 1.590)$ is between 0.05 and 0.10. (one tailed p-value)
We do not reject H_0.

Section 9.2: Compare two population proportions

Conditions:
Two simple random samples and are independent of each other.
Each sample contains at least 10 successes and 10 failures.

A confidence interval for the difference $p_1 - p_2$ between two population proportion is
$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$.

Example

An experiment was designed to compare treating cancer patients with chemotherapy to treatment with a combination of chemotherapy and radiation. Of the 154 patients who received only chemotherapy, 76 survived at least 15 years whereas 98 of the 164 patients who received the hybrid treatment survived at least that long.
Let p_1, p_2 denote the proportion of patients, when treated with the respective treatment, survive at least 15 years.
Get a 99% CI for $p_1 - p_2$.

Solutions

$\hat{p}_1 = 76/154 = 0.494, \hat{p}_2 = 98/164 = 0.598.$
CI is
$0.494 - 0.598 \pm 2.58\sqrt{\frac{0.494*0.506}{154} + \frac{0.598*0.402}{164}}$
$= -0.104 \pm 0.143 = (-0.247, 0.039).$

Testing for $H_0 : p_1 = p_2$ will be given in chapter 10 as it is a special case of the chi square test there.

Check your understanding exercise

A new postsurgical treatment was compared with a standard treatment. Seven subjects received the new treatment while seven others (controls) ceceived the standard treatment. THeir recovery time, in days, are given below:
Treatment: 12,13,15,19,20,21,24
Control: 18,23,24,30,32,35,39
Can you conclude the mean recovery time for the new treatment, μ_1, is shorter than that for the standard treatment, μ_2? Use $\alpha = 0.05$.
Also obtain a 95% CI for $\mu_1 - \mu_2$.

Solutions

summary of data: mean and sd for treatment: $\bar{x}_1 = 17.71, s_1 = 4.46$.
mean and sd for control: $\bar{x}_2 = 28.71, s_2 = 7.39$.
$n_1 = n_2 = 7$.
95% CI for $\mu_1 - \mu_2$:
$17.71 - 28.71 \pm 2.447\sqrt{\frac{4.46^2}{7} + \frac{7.39^2}{7}} = -11 \pm 7.98 = (-18.98, -3.02)$.

$H_0 : \mu_1 = \mu_2$
$H_1 : \mu_1 < \mu_2$
$t = \frac{17.71-28.01}{\sqrt{4.46^2/7+7.39^2/7}} = -3.37$.
P-value $= P(t < -3.37)$ between 0.005 and 0.01.
Reject H_0. We conclude the mean recovery time is shorter for the new treatment.

Check your understanding exercise

A sample survey asked random sample of teens with online profiles if they included false information in their profiles. Of 170 younger teens (ages 12 to 14), 117 said yes. Of 317 older teens (ages 15 to 17), 152 said yes. Give a 95% CI for the difference between the proportion of younger and older teens who include false information in their online profiles.

$\hat{p}_1 = 117/170 = 0.688, \hat{p}_2 = 152/317 = 0.479.$
95% CI is
$0.688 - 0.479 \pm 1.96\sqrt{\frac{0.688*0.312}{170} + \frac{0.479*0.521}{317}} = (0.120, 0.298).$

For matched pair design, we take the difference between each pair and use one sample method.
Testing $H_0 : \mu_1 = \mu_2$ then becomes testing $\mu_d = 0$ where $\mu_d = \mu_1 - \mu_2$.
This is called **matched pairs t test**.

Each of six subjects were given both diets (diet 1: high fat, diet 2: high carbodydrates) in a random order. Each diet lasted 3 days. At the end of the 3 day period, the subject was put on the treadmill and the time to exhaustion, in seconds, was measured. The data are given below.

Subject	Diet1	Diet2	difference(d)
1	91	122	31
2	48	53	5
3	71	110	39
4	45	71	26
5	61	91	30
6	61	122	61
mean:	62.83	94.83	32

(standard deviation of d is 18.221)

Get a 95% confidence interval for the mean difference in treadmill time between the two diets. Also perform a t test to examine if the two diets produce the same mean treadmill time.

$\bar{d} \pm t\frac{s_d}{\sqrt{n}} = 32 \pm 2.571 * 18.221/\sqrt{6} = 32 \pm 19.12 = (12.88, 51.12)$.
$H_0 : \mu_d = 0$ or $\mu_1 = \mu_2$
$H_1 : \mu_d \neq 0$ or $\mu_1 \neq \mu_2$
$t = \frac{32-0}{\frac{18.221}{\sqrt{6}}} = 4.30$
P-value is between 0.002 and 0.01. Reject H_0. There is sufficient evidence the two diets do not produce the same mean treadmill time.

One more exercise with proportions

A poll taken by GSS asked whether people are satisfied with their financial situation. A total of 478 out of 2038 people said they were. The same question was asked two years later, and 537 out of 1967 people said they were. Get a 95% confidence interval for the increase in the proportion of people who were satisfied with their financial condition.

solution

$\hat{p}_1 = \frac{478}{2038} = 0.235, \hat{p}_2 = \frac{537}{1967} = 0.273.$

$0.273 - 0.235 \pm 1.96 * \sqrt{\frac{0.235*0.765}{2038} + \frac{0.273*0.727}{1967}} = 0.038 \pm 0.027 = (0.011, 0.065).$