Chapter 9-10

Confidence Intervals and Hypothesis Testing

CI and HT for the difference of Two Means - Examples

Statistics

Mehmet Güray Güler, PhD

Last updated 17.07.2020

Example – 1: Comparing Gas Mileage

- A study was conducted in which two types of engines, A and B, were compared.
- Gas mileage, in km/l, was measured.
  - 50 experiments for engine type A
  - 60 experiments for engine type B.
- The same type of gasoline and road/ weather conditions apply.
- The variables are
  - X_{A} = gas mileage using engine type A, and
  - X_{B} = gas mileage using engine type B.

Example – 1: Comparing Gas Mileage

- The data were summarized as follows:

$$\bar{x}_A = 14.4 \text{ km / l}, \quad \bar{x}_B = 12.2 \text{ km / l}.$$

- Assume that the population standard deviations are known as:
- $\sigma_\mathrm{A} = 5$ km/l and, $\sigma_\mathrm{B} = 4$ km/l,
- A) Test whether $\mu_A = \mu_B$ or not with $\alpha = 0.04$
- B) Find a 96% confidence interval for on $\mu_\mathrm{A} - \mu_\mathrm{B}$.

Example – 1: Comparing Gas Mileage

- **Step1:** The hypothesis are:
  - $H_0: \mu_A - \mu_B = 0$
  - $H_1: \mu_A - \mu_B \neq 0$

- **Step2:** The test statistic for step 2: $z = \dfrac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$

- **Step 3**: R- [$-z_{0.02}$,$z_{0.02}$] = R – [-2.05,2.05]

- **Step 4**: $z_{obs} = \dfrac{14.4-12.2\ -0}{(5^2/50+4^2\ /\ 60)^{0.5}} = 2.51$

- **Step5:** Since z_{obs} = 2.51 >2.05 we reject the hypothesis.

Example – 1: Comparing Gas Mileage

- Find z-value from the standard normal table (or EXCEL) and
- Substitute into the formula:

$$\mu_A - \mu_B \in (\bar{x}_A - \bar{x}_B) \pm z_{\alpha/2}\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}$$

$$\mu_A - \mu_B \in 2.2 \pm 2.05\sqrt{\frac{25}{50} + \frac{16}{60}} = 2.2 \pm 1.795$$

$$\Rightarrow 0.41\,\text{km/l} < \mu_A - \mu_B < 3.99\,\text{km/l}.$$

Example – 2: Comparing Chemical Concentration

- A study was conducted to estimate the difference in the amounts of the chemical orthophosphorous measured (in mg/l) at two different stations on a river.
  - 15 samples from station 1,
  - 12 samples from station 2, all at random locations.
- The variables:
  - X_1 = amount of chemical measured at station 1, and
  - X_2 = amount of chemical measured at station 2.,

Example – 2: Comparing Chemical Concentration

- The data were summarized as follows:
- $\bar{x}_1 = 3.84\, mg/\, l\,, s_1 = 1.65\, mg/l\,, n_1 = 15$
- $\bar{x}_2 = 1.49\, mg/\, l\,, s_2 = 0.92\, mg/l\,, n_2 = 12$

- Assume $\sigma_1^2 = \sigma_2^2$.

- Find a 95% CI for the difference between the true mean content at these two stations.

Example – 2: Comparing Chemical Concentration

- Since we assume that $\sigma_1{}^2 = \sigma_2{}^2 = \sigma^2$, we first estimate this common variance for both stations from the pooled estimate:

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}, \text{hence}$$

$$s_p^2 = \frac{(14)(1.65)^2 + (11)(0.92)^2}{25} = 1.897.$$

- From t-table, with 25 degrees of freedom, we get $t_{0.025} \approx 2.06$

Example – 2: Comparing Chemical Concentration

- Assuming ${\sigma_1}^2 = {\sigma_2}^2 = \sigma^2$, and using the summary statistics
- A 95% CI is calculated as:
  - $\mu_1 - \mu_2 \in (\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ = 2.35 $\pm$1.10 = [1.25, 3.45]
- How do we interpret this interval?

Example – 3: Comparing Chemical Concentration Revisited with Revisited with $\sigma_1^2 \neq \sigma_2^2$

- In Example 2, we assumed that the population variances were equal.
- One may *not be very comfortable* with the assumption of equal variances as the sample variances are not very close.
- Let’s drop the assumption of equal variances and construct an approximate CI again.

Example – 3: Comparing Chemical Concentration Revisited with Revisited with $\sigma_1^2 \neq \sigma_2^2$

- Assuming $\sigma_1{}^2 \neq \sigma_2{}^2$, and using the observed statistics value:
- using the messy formula =>> v= 22.6 ~ 23
- $t_{0.025}$(v=23) = 2.07
- The CI is given by:
- $\mu_1 - \mu_2 \in (\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ = 2.35 $\pm$1.04 = [1.31, 3.39]

Paired Observations

HT and CI for paired observations

Paired Observations

- **Example**: tires of the two brands are assigned at random to the left and right rear wheels of 8 taxis and the following distances, in kilometers, are recorded. Use $\alpha = 0.01$

- A) Test the hypothesis that they have equal lifetime

- B) Find a CI for $\mu_1 - \mu_2$

Taxi	Brand A	Brand B
1	34,400	36,700
2	45,500	46,800
3	36,700	37,700
4	32,000	31,100
5	48,400	47,800
6	32,800	36,400
7	38,100	38,900
8	30,100	31,500

HT for Paired Observations

- Let $\overline{D}$ and S_d are mean and standard deviation of difference of n random pairs of measurement, then

$$\mathrm{T} = \frac{\overline{D} - \mu_D}{S_d / \sqrt{n}}$$

- is a T RV with n-1 dof.

HT for Paired Observations

- **Step1:** The hypothesis are:
  - $H_0: \mu_1 - \mu_2 = \mu_D$
  - $H_1: \mu_1 - \mu_2 \neq \mu_D$
- **Step2:** The test statistic: $\mathrm{T} = \frac{\bar{D} - \mu_D}{S_d / \sqrt{n}}$ with v = n – 1
- **Step 3**: R- [-$t_{\alpha \backslash 2}$, $t_{\alpha \backslash 2}$],
- **Step 4**: Calculate t_{obs} using the formula in step2 using $\bar{d}_{\boldsymbol{obs}}$
- **Step5:** if t_{obs} is in the critical region, we reject the null hypothesis.

HT for Paired Observations

Taxi	Brand A	Brand B	D
1	34,400	36,700	-2,300
2	45,500	46,800	-1,300
3	36,700	37,700	-1,000
4	32,000	31,100	900
5	48,400	47,800	600
6	32,800	36,400	-3,600
7	38,100	38,900	-800
8	30,100	31,500	-1,400

- $\bar{d}$=-1113, $s_d = 1454$
- **Step1:** $H_0: \mu_1 - \mu_2 = 0$ vs $H_1: \mu_1 - \mu_2 \neq 0$
- **Step2:** $\mathrm{T} = \frac{\bar{D}-\mu_D}{S_d/\sqrt{n}}$ with v = 7
- **Step 3**: R - [-3 .499, 3.499],
- **Step 4**: $t_{obs} = \frac{1.113-0}{1454/\sqrt{8}}$ =-2.16
- **Step5:** Do not reject.

CI for Paired Observations

If $\bar{d}$ and s_d are the mean and standard deviation, respectively, of the normally distributed differences of n random pairs of measurements, a $100(1-\alpha)\%$ confidence interval for $\mu_D = \mu_1 - \mu_2$ is

$$\bar{d} - t_{\alpha/2}\frac{s_d}{\sqrt{n}} < \mu_D < \bar{d} + t_{\alpha/2}\frac{s_d}{\sqrt{n}},$$

where $t_{\alpha/2}$ is the t-value with $v = n-1$ degrees of freedom, leaving an area of $\alpha/2$ to the right.

CI for Paired Observations

- Solution:

$n = 8, \bar{d} = -1112.5, s_d = 1454$, with $t_{0.005} = 3.499$ with 7 degrees of freedom. So,

$$-1112.5 \pm (3.499)\frac{1454}{\sqrt{8}} = -1112.5 \pm 1798.7,$$

which yields $-2911.2 < \mu_D < 686.2$.

H_0	Value of Test Statistic	H_1	Critical Region
$\mu = \mu_0$	$z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$; σ known	$\mu < \mu_0$ $\mu > \mu_0$ $\mu \neq \mu_0$	$z < -z_\alpha$ $z > z_\alpha$ $z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$
$\mu = \mu_0$	$t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$; $v = n - 1$, σ unknown	$\mu < \mu_0$ $\mu > \mu_0$ $\mu \neq \mu_0$	$t < -t_\alpha$ $t > t_\alpha$ $t < -t_{\alpha/2}$ or $t > t_{\alpha/2}$
$\mu_1 - \mu_2 = d_0$	$z = \dfrac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$; σ_1 and σ_2 known	$\mu_1 - \mu_2 < d_0$ $\mu_1 - \mu_2 > d_0$ $\mu_1 - \mu_2 \neq d_0$	$z < -z_\alpha$ $z > z_\alpha$ $z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$
$\mu_1 - \mu_2 = d_0$	$t = \dfrac{(\bar{x}_1 - \bar{x}_2) - d_0}{s_p\sqrt{1/n_1 + 1/n_2}}$; $v = n_1 + n_2 - 2$, $\sigma_1 = \sigma_2$ but unknown, $s_p^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$	$\mu_1 - \mu_2 < d_0$ $\mu_1 - \mu_2 > d_0$ $\mu_1 - \mu_2 \neq d_0$	$t < -t_\alpha$ $t > t_\alpha$ $t < -t_{\alpha/2}$ or $t > t_{\alpha/2}$
$\mu_1 - \mu_2 = d_0$	$t' = \dfrac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$; $v = \dfrac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$, $\sigma_1 \neq \sigma_2$ and unknown	$\mu_1 - \mu_2 < d_0$ $\mu_1 - \mu_2 > d_0$ $\mu_1 - \mu_2 \neq d_0$	$t' < -t_\alpha$ $t' > t_\alpha$ $t' < -t_{\alpha/2}$ or $t' > t_{\alpha/2}$
$\mu_D = d_0$ paired observations	$t = \dfrac{\overline{d} - d_0}{s_d/\sqrt{n}}$; $v = n - 1$	$\mu_D < d_0$ $\mu_D > d_0$ $\mu_D \neq d_0$	$t < -t_\alpha$ $t > t_\alpha$ $t < -t_{\alpha/2}$ or $t > t_{\alpha/2}$

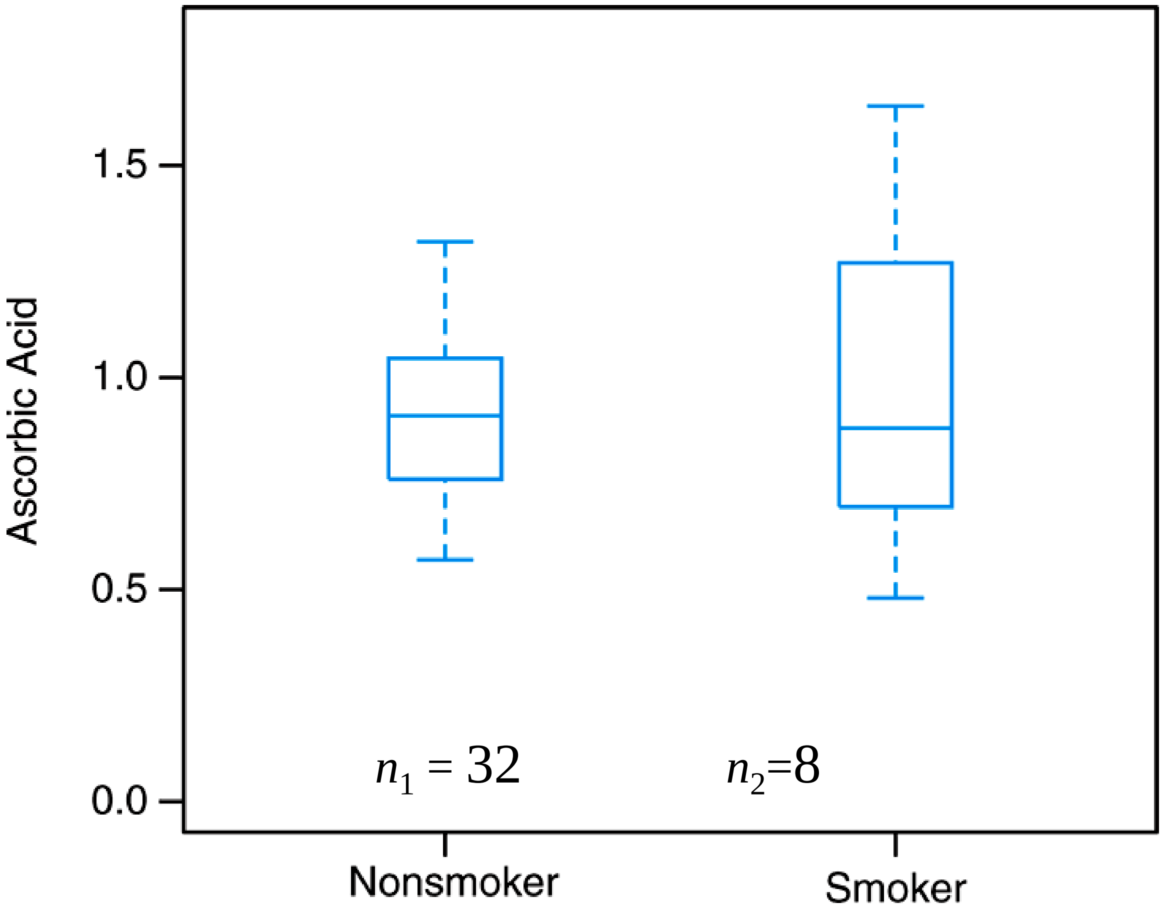
Graphical Methods for Comparing Means

Graphical Methods for Comparing Means

Box plots can be used for visual comparison of data from several samples to compare the corresponding population means and variances.

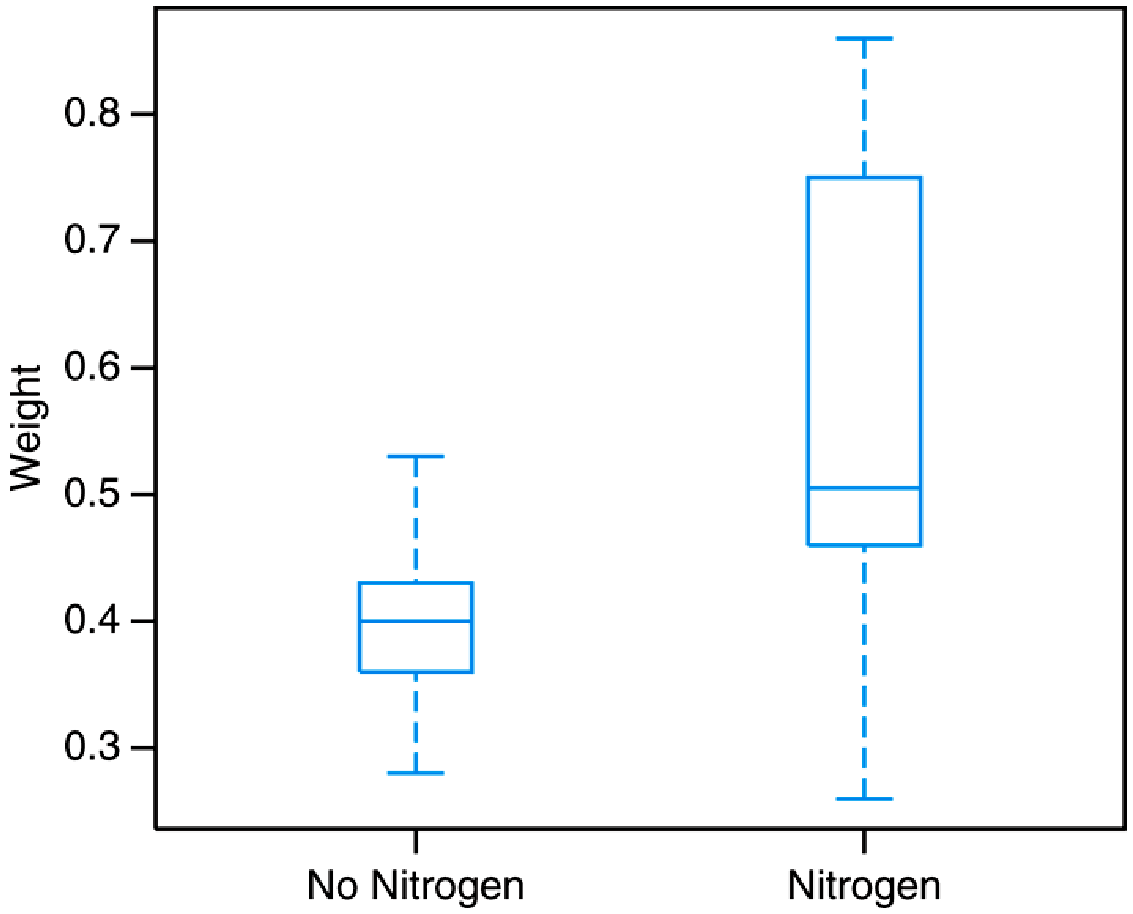
How do you interpret Figure 1? See Exercise 10.40

Figure 1. Box plots of plasma ascorbic acid in smokers and nonsmokers



Graphical Methods for Comparing Means

Figure 2. Box plots of seedling data



Interpretation?
See Exercise 10.40