

Example: To determine whether waste discharged by a chemical plant is polluting the local river, the river water was sampled at two locations – one upstream and one downstream from the discharge site. Independent water samples of sizes $n_1 = 10$ and $n_2 = 15$, respectively, were selected from the upstream and downstream locations. The concentration level (ppm) of a suspected chemical pollutant was determined in each water sample, with the following results:

Upstream 24.5, 29.7, 20.4, 28.5, 25.3, 21.8, 20.2, 21.0, 21.9, 22.2

Downstream 32.8, 30.4, 32.3, 26.4, 27.8, 26.9, 29.0, 31.5, 31.2, 26.7, 25.6, 25.1, 32.8, 34.3, 35.4

Is there sufficient evidence that the chemical plant is polluting the river? If yes, explore the level of the pollution by estimating the difference in the average concentrations of the chemical pollutant upstream and downstream of the river.

Write-up

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 - \mu_2 < 0 \text{ where } \mu_1 = \text{mean concentration upstream} \\ \mu_2 = \text{mean concentration downstream}$$

Assumptions check: Boxplots show no outliers in either data set. For the downstream data, statistical tests for normality revealed no evidence that the nearly normal assumption is violated. The data appears reasonably symmetric in the stem plot. However, for the upstream data, statistical tests for normality show some evidence that the data do not come from a normal population (p-value = .0624 for the Kolmogorov-Smirnov test with H_0 : data is a sample from a normal distribution vs. H_a : data is not a sample from a normal distribution.) The stem plot appears right-skewed. The two-sample t procedures are quite robust against non-normality. We will continue the analysis with the two-sample t methods, and compare the results with those from an alternative nonparametric method.

$$\bar{X}_1 = 23.55, s_1 = 3.359, n_1 = 10 \quad (\text{Upstream})$$

$$\bar{X}_2 = 29.88, s_2 = 3.327, n_2 = 15 \quad (\text{Downstream})$$

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t(df) \quad \text{where } df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1} \left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1} \left(\frac{s_2^2}{n_2}\right)^2}$$

$$t_{\text{obs}} = -4.63$$

$$df = 19.3$$

$$p\text{-value} = .0001$$

Reject H_0 in favor of H_1 . There is very strong evidence that the mean concentration of the pollutant downstream of the discharge site is higher than the mean concentration upstream of the discharge site.

A 95% confidence interval for $\mu_2 - \mu_1$ (downstream – upstream) is (3.4739, 9.1861) ppm. We conclude with 95% confidence that the average downstream pollution level is between 3.47 and 9.19 ppm more than that prevailing in the upstream waters.