

Two Sample t Test

Nathaniel E. Helwig

University of Minnesota

1 Paired Samples t Test

Suppose that we have a random sample of bivariate data $(x_i, y_i) \stackrel{\text{iid}}{\sim} F$ from some continuous distribution F . For example, the paired observations (x_i, y_i) could be the same variable measured from the i -th subject before and after some intervention, or the (x_i, y_i) observations could be paired in some other fashion (e.g., the same variable recorded from the i -th twin pair). Suppose that we want to test a null hypothesis about the difference between the expected values for the paired data, i.e.,

- $H_0 : \mu_x - \mu_y = \mu_0$ versus $H_1 : \mu_x - \mu_y \neq \mu_0$ (exact H_0 with two-sided H_1)
- $H_0 : \mu_x - \mu_y \geq \mu_0$ versus $H_1 : \mu_x - \mu_y < \mu_0$ (inexact H_0 with less than H_1)
- $H_0 : \mu_x - \mu_y \leq \mu_0$ versus $H_1 : \mu_x - \mu_y > \mu_0$ (inexact H_0 with greater than H_1)

where μ_0 is the null hypothesized difference between the means (typically $\mu_0 = 0$).

Defining the difference score such as $z_i = x_i - y_i$ for $i = 1, \dots, n$, we could rewrite the null hypotheses as a one sample t test (Student, 1908) on the difference score:

- $H_0 : \mu_z = \mu_0$ versus $H_1 : \mu_z \neq \mu_0$ (exact H_0 with two-sided H_1)
- $H_0 : \mu_z \geq \mu_0$ versus $H_1 : \mu_z < \mu_0$ (inexact H_0 with less than H_1)
- $H_0 : \mu_z \leq \mu_0$ versus $H_1 : \mu_z > \mu_0$ (inexact H_0 with greater than H_1)

where $\mu_z = E(z_i)$ is the expected value of the difference scores. Note that these null hypotheses are equivalent to the null hypotheses that we formed for the one-sample t test. Thus, the paired samples t test simply involves conducting a one sample t test on the difference score.

Example 1. Suppose that $n = 9$ psychiatric patients were treated with a tranquilizer drug, which was meant to reduce their suicidal tendencies. Let X and Y denote the suicidal tendencies of the patients (as measured by the Hamilton Depression Scale) before and after the tranquilizer treatment. Consequently, we will test the null hypothesis $H_0 : \mu_z \leq 0$ versus the alternative hypothesis $H_1 : \mu_z > 0$, where $Z = X - Y$ is the difference score (before minus after). If we reject H_0 , this would be evidence of an effective treatment.

```
> pre <- c(1.83, 0.50, 1.62, 2.48, 1.68, 1.88, 1.55, 3.06, 1.30)
> post <- c(0.878, 0.647, 0.598, 2.050, 1.060, 1.290, 1.060, 3.140, 1.290)
> t.test(pre, post, alternative = "greater", paired = TRUE)
```

Paired t-test

```
data: pre and post
t = 3.0354, df = 8, p-value = 0.008088
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.1673028      Inf
sample estimates:
mean of the differences
      0.4318889
```

We could conduct a nonparametric version of this test using the `np.loc.test` function:

```
> library(np.test)
> np.loc.test(pre, post, alternative = "greater", paired = TRUE)
```

Nonparametric Location Test (Paired t-test)

```
alternative hypothesis: true difference of means is greater than 0
t = 3.0354, p-value = 0.0137
sample estimate:
mean of the differences
      0.4318889
```

Example 2. Suppose that we have collected psychological test scores from dizygotic twins, and want to test the null hypothesis that there is no difference between the first born twin's scores (X) and the second born twin's scores (Y). In other words, we will test the exact null hypothesis $H_0 : \mu_z = 0$ versus the two-sided alternative hypothesis $H_1 : \mu_z \neq 0$, where $Z = X - Y$ is the difference score (twin 1 minus twin 2).

```
> x <- c(277, 169, 157, 139, 108, 213, 232, 229, 114, 232, 161, 149, 128)
> y <- c(256, 118, 137, 144, 146, 221, 184, 188, 97, 231, 114, 187, 230)
> t.test(x, y, paired = TRUE)
```

Paired t-test

```
data: x and y
t = 0.34787, df = 12, p-value = 0.734
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -22.26777 30.72931
sample estimates:
mean of the differences
      4.230769
```

We could conduct a nonparametric version of this test using the `np.loc.test` function:

```
> np.loc.test(x, y, paired = TRUE)
```

Nonparametric Location Test (Paired t-test)

```
alternative hypothesis: true difference of means is not equal to 0
t = 0.3479, p-value = 0.7415
sample estimate:
mean of the differences
      4.230769
```

2 Independent Samples t Test

Now suppose that we have observed $x_i \stackrel{\text{iid}}{\sim} N(\mu_x, \sigma_x^2)$ and $y_i \stackrel{\text{iid}}{\sim} N(\mu_y, \sigma_y^2)$, where the X and Y observations are assumed to be independent of one another. Furthermore, suppose that we want to test a null hypothesis about the difference between the expected values for the two independent samples, i.e.,

- $H_0 : \mu_x - \mu_y = \mu_0$ versus $H_1 : \mu_x - \mu_y \neq \mu_0$ (exact H_0 with two-sided H_1)
- $H_0 : \mu_x - \mu_y \geq \mu_0$ versus $H_1 : \mu_x - \mu_y < \mu_0$ (inexact H_0 with less than H_1)
- $H_0 : \mu_x - \mu_y \leq \mu_0$ versus $H_1 : \mu_x - \mu_y > \mu_0$ (inexact H_0 with greater than H_1)

where μ_0 is the null hypothesized difference between the means (typically $\mu_0 = 0$).

If the variances of the two populations are assumed to be equal, i.e., if $\sigma_x^2 = \sigma_y^2$, then the test statistic can be defined as

$$T_0 = \frac{\bar{x} - \bar{y} - \mu_0}{s_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}}$$

where $\bar{x} = \frac{1}{n_x} \sum_{i=1}^{n_x} x_i$ and $\bar{y} = \frac{1}{n_y} \sum_{i=1}^{n_y} y_i$ are the sample means for each group, n_x and n_y denote the sample sizes for each group, and

$$s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}$$

is the *pooled variance estimate* with $s_x^2 = \frac{1}{n_x - 1} \sum_{i=1}^{n_x} (x_i - \bar{x})^2$ and $s_y^2 = \frac{1}{n_y - 1} \sum_{i=1}^{n_y} (y_i - \bar{y})^2$ denoting the variance for each group. Assuming that H_0 is true, the test statistic T_0 follows a Student's t distribution with $n_x + n_y - 2$ degrees of freedom, i.e., $T_0 \sim t_{n_x + n_y - 2}$ under H_0 .

Assuming that $\sigma_x^2 = \sigma_y^2$ is a rather strict assumption, which may not be very reasonable in any real data application. Note that if equal variances are incorrectly assumed, i.e., if we use the previous T_0 test statistic when $\sigma_x^2 \neq \sigma_y^2$, then the resulting hypothesis test will not be valid. In particular, when the variances are unequal, the previous test statistic's sampling distribution will not be well approximated by a $t_{n_x + n_y - 2}$ distribution, so the significance test will not have the desired significance level. Whether or not the type I error rate is too small versus too large will depend on the direction of the alternative hypothesis, the sample sizes n_x and n_y , and the ratio of the true variance σ_x^2 / σ_y^2 (see Helwig, 2019).

Instead of assuming that the variances of the two groups are equal, we could use

$$T_0 = \frac{\bar{x} - \bar{y} - \mu_0}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}}$$

which is the version of the test statistic that was proposed by Welch (1938, 1947). Assuming that the null hypothesis is true, the sampling distribution of Welch's t test statistic can be well approximated by Student's t distribution with the degrees of freedom parameter

$$\nu = \frac{\left(\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}\right)^2}{\frac{s_x^4}{n_x^2(n_x-1)} + \frac{s_y^4}{n_y^2(n_y-1)}}$$

which is known as the Welch-Satterthwaite approximation to the degrees of freedom (Satterthwaite, 1946).

Example 3. Suppose that we are interested in assessing the effectiveness of a Social Skills Training (SST) program for alcoholics that are in a rehabilitation program. Assume that $n_x = 12$ patients (the control group) participated in the normal treatment program, and $n_y = 11$ patients (the test group) participated in the SST supplement in addition to the normal treatment program. The goal is to test the null hypothesis $H_0 : \mu_x - \mu_y \leq 0$ versus the alternative hypothesis $H_1 : \mu_x - \mu_y > 0$, where μ_x and μ_y denote the average amount of alcohol consumed within the first year after the program.

```
> x <- c(1042, 1617, 1180, 973, 1552, 1251, 1151, 1511, 728, 1079, 951, 1319)
> y <- c(874, 389, 612, 798, 1152, 893, 541, 741, 1064, 862, 213)
> t.test(x, y, alternative = "greater")
```

Welch Two Sample t-test

```
data: x and y
t = 3.9747, df = 20.599, p-value = 0.0003559
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 258.5566      Inf
```

```
sample estimates:
mean of x mean of y
1196.1667  739.9091
```

We could conduct a nonparametric version of this test using the `np.loc.test` function:

```
> set.seed(0)
> np.loc.test(x, y, alternative = "greater")
```

```
Nonparametric Location Test (Welch Two Sample t-test)
alternative hypothesis: true difference of means is greater than 0
t = 3.9747,  p-value = 4e-04
sample estimate:
difference of the means
456.2576
```

3 Confidence Interval for Mean Differences

To form a $100(1 - \alpha)\%$ confidence interval for the mean difference $\delta = \mu_x - \mu_y$, we will use the Welch version of the t test. Note that

$$1 - \alpha = P \left(t_{\alpha/2} < \frac{\bar{x} - \bar{y} - \delta}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}} < t_{1-\alpha/2} \right)$$

where t_α denotes the α -th quantile of the t_ν distribution. Rearranging the terms inside the probability statement reveals that

$$\begin{aligned} 1 - \alpha &= P \left(t_{\alpha/2} \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}} < \bar{x} - \bar{y} - \delta < t_{1-\alpha/2} \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}} \right) \\ &= P \left(t_{\alpha/2} \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}} - (\bar{x} - \bar{y}) < -\delta < t_{1-\alpha/2} \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}} - (\bar{x} - \bar{y}) \right) \\ &= P \left(\bar{x} - \bar{y} - t_{\alpha/2} \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}} > \delta > \bar{x} - \bar{y} - t_{1-\alpha/2} \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}} \right) \end{aligned}$$

which implies that the $100(1 - \alpha)\%$ confidence interval is given by

$$\left[\bar{x} - \bar{y} - t_{1-\alpha/2} \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}, \quad \bar{x} - \bar{y} - t_{\alpha/2} \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}} \right]$$

and given that $-t_{\alpha/2} = t_{1-\alpha/2}$ the confidence interval can be written as

$$\bar{x} - \bar{y} \pm t_{1-\alpha/2} \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}$$

Note that the above confidence interval makes sense to use if the alternative hypothesis is two-sided. For directional (one-sided) tests, it would make more sense to use a confidence bound, which places all of the uncertainty in the direction relevant to the alternative:

- For $H_1 : \mu < \mu_0$, use an upper confidence bound: $[-\infty, \bar{x} - \bar{y} + t_{1-\alpha} \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}]$
- For $H_1 : \mu > \mu_0$, use a lower confidence bound: $[\bar{x} - \bar{y} - t_{1-\alpha} \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}, \infty]$

Example 4. For the previous example, the `t.test` function automatically outputs the one-sided lower confidence bound. With 95% confidence, we conclude that (on average) the control group drinks at least 258.5566 more centiliters of alcohol than the SST group during the first year after the treatment program. We wanted to compute a nonparametric 95% confidence interval, would could use the `np.boot` function:

```
> # setup data
> z <- c(x, y)
> g <- factor(rep(c("ctrl", "sst"), c(length(x), length(y))))
> data <- data.frame(alcohol = z, group = g)

> # define statistic function
> statfun <- function(x, data){
+   means <- with(data[x,], tapply(alcohol, group, mean))
+   means[1] - means[2]
+ }
```

```
> # bootstrap data
> set.seed(0)
> np.boot(1:length(z), statfun, data)
```

```
Nonparametric Bootstrap of Univariate Statistic
using R = 9999 bootstrap replicates
```

```
t0: 456.2576
SE: 113.4205
Bias: 0.9594
```

```
BCa Confidence Intervals:
```

```
      lower      upper
90% 276.3364 648.3219
95% 243.5856 685.4309
99% 175.8862 756.5396
```

References

- Helwig, N. E. (2019). Statistical nonparametric mapping: Multivariate permutation tests for location, correlation, and regression problems in neuroimaging. *WIREs Computational Statistics* 2, e1457.
- Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin* 2(6), 110–114.
- Student (1908). The probable error of a mean. *Biometrika* 6(1), 1–25.
- Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika* 39(3/4), 350–362.
- Welch, B. L. (1947). The generalization of ‘Student’s’ problem when several different population variances are involved. *Biometrika* 34(1–2), 28–35.