**Population Mean Between Two Matched Samples**

Two data samples are **matched**if they come from repeated observations of the same subject. Here, we assume that the data populations follow the [normal distribution](http://www.r-tutor.com/node/58). Using the **paired t-test**, we can obtain an interval estimate of the difference of the population means.

**Example**

In the built-in data set named **immer**, the barley yield in years 1931 and 1932 of the same field are recorded. The yield data are presented in the [data frame columns](http://www.r-tutor.com/node/15) Y1 and Y2.

> library(MASS)         # load the MASS package   
> head(immer)   
  Loc Var    Y1    Y2   
1  UF   M  81.0  80.7   
2  UF   S 105.4  82.3   
    .....

**Problem**

Assuming that the data in immer follows the normal distribution, find the 95% confidence interval estimate of the difference between the mean barley yields between years 1931 and 1932.

**Solution**

We apply the t.test function to compute the difference in means of the matched samples. As it is a paired test, we set the "paired" argument as TRUE.

> t.test(immer$Y1, immer$Y2, paired=TRUE)   
   
           Paired t-test   
   
data:  immer$Y1 and immer$Y2   
t = 3.324, df = 29, p-value = 0.002413   
alternative hypothesis: true difference in means is not equal to 0   
95 percent confidence interval:   
  6.122 25.705   
sample estimates:   
mean of the differences   
                 15.913

**Answer**

Between years 1931 and 1932 in the data set immer, the 95% confidence interval of the difference in means of the barley yields is the interval between 6.122 and 25.705.

**Exercise**

**Population Mean Between Two Independent Samples**

Two data samples are **independent**if they come from unrelated populations and the samples does not affect each other. Here, we assume that the data populations follow the [normal distribution](http://www.r-tutor.com/node/58). Using the **unpaired t-test**, we can obtain an interval estimate of the difference between two [population means](http://www.r-tutor.com/node/35).

**Example**

In the [data frame column](http://www.r-tutor.com/node/15) **mpg**of the data set [mtcars](http://www.r-tutor.com/node/10), there are gas mileage data of various 1974 U.S. automobiles.

> mtcars$mpg   
 [1] 21.0 21.0 22.8 21.4 18.7 ...

Meanwhile, another data column in mtcars, named **am**, indicates the transmission type of the automobile model (0 = automatic, 1 = manual).

> mtcars$am   
 [1] 1 1 1 0 0 0 0 0 ...

In particular, the gas mileage for manual and automatic transmissions are two independent data populations.

**Problem**

Assuming that the data in mtcars follows the normal distribution, find the 95% confidence interval estimate of the difference between the mean gas mileage of manual and automatic transmissions.

**Solution**

As mentioned in the tutorial [*Data Frame Row Slice*](http://www.r-tutor.com/node/17), the gas mileage for automatic transmission can be listed as follows:

> L = mtcars$am == 0   
> mpg.auto = mtcars[L,]$mpg   
> mpg.auto                    # automatic transmission mileage   
 [1] 21.4 18.7 18.1 14.3 24.4 ...

By applying the negation of L, we can find the gas mileage for manual transmission.

> mpg.manual = mtcars[!L,]$mpg   
> mpg.manual                  # manual transmission mileage   
 [1] 21.0 21.0 22.8 32.4 30.4 ...

We can now apply the t.test function to compute the difference in means of the two sample data.

> t.test(mpg.auto, mpg.manual)   
   
        Welch Two Sample t-test   
   
data:  mpg.auto and mpg.manual   
t = -3.7671, df = 18.332, p-value = 0.001374   
alternative hypothesis: true difference in means is not equal to 0   
95 percent confidence interval:   
 -11.2802  -3.2097   
sample estimates:   
mean of x mean of y   
   17.147    24.392

**Answer**

In mtcars, the mean mileage of automatic transmission is 17.147 mpg and the manual transmission is 24.392 mpg. The 95% confidence interval of the difference in mean gas mileage is between 3.2097 and 11.2802 mpg.

**Alternative Solution**

We can model the response variable mtcars$mpg by the predictor mtcars$am, and then apply the t.test function to estimate the difference of the population means.

> t.test(mpg ~ am, data=mtcars)   
   
        Welch Two Sample t-test   
   
data:  mpg by am   
t = -3.7671, df = 18.332, p-value = 0.001374   
alternative hypothesis: true difference in means is not equal to 0   
95 percent confidence interval:   
 -11.2802  -3.2097   
sample estimates:   
mean in group 0 mean in group 1   
         17.147          24.392

**Note**

Some textbooks truncate down the degree of freedom to an integer, and the result would differ from the t.test.

**Comparison of Two Population Proportions**

A survey conducted in two distinct populations will produce different results. It is often necessary to compare the survey response proportion between the two populations. Here, we assume that the data populations follow the [normal distribution](http://www.r-tutor.com/node/58).

**Example**

In the built-in data set named **quine**, children from an Australian town is classified by ethnic background, gender, age, learning status and the number of days absent from school.

> library(MASS)         # load the MASS package   
> head(quine)   
  Eth Sex Age Lrn Days   
1   A   M  F0  SL    2   
2   A   M  F0  SL   11   
    .....

In effect, the [data frame column](http://www.r-tutor.com/node/15) **Eth**indicates whether the student is Aboriginal or Not ("A" or "N"), and the column **Sex**indicates Male or Female ("M" or "F").

In R, we can tally the student ethnicity against the gender with the table function. As the result shows, within the Aboriginal student population, 38 students are female. Whereas within the Non-Aboriginal student population, 42 are female.

> table(quine$Eth, quine$Sex)   
   
     F  M   
  A 38 31   
  N 42 35

**Problem**

Assuming that the data in quine follows the normal distribution, find the 95% confidence interval estimate of the difference between the female proportion of Aboriginal students and the female proportion of Non-Aboriginal students, each within their own ethnic group.

**Solution**

We apply the prop.test function to compute the difference in female proportions. The Yates’s continuity correction is disabled for pedagogical reasons.

> prop.test(table(quine$Eth, quine$Sex), correct=FALSE)   
   
        2-sample test for equality of proportions   
        without continuity correction   
   
data:  table(quine$Eth, quine$Sex)   
X-squared = 0.0041, df = 1, p-value = 0.949   
alternative hypothesis: two.sided   
95 percent confidence interval:   
 -0.15642  0.16696   
sample estimates:   
 prop 1  prop 2   
0.55072 0.54545

**Answer**

The 95% confidence interval estimate of the difference between the female proportion of Aboriginal students and the female proportion of Non-Aboriginal students is between -15.6% and 16.7%.

**Exercise**

Estimate the difference between two population proportions using your textbook formula.