Inference About Two Populations

It is often necessary to draw conclusion on the difference between two populations by their data samples.

In the following tutorials, we discuss how to estimate the difference of means and proportions between two normally distributed data populations.

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.

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 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

Estimate the difference between the means of matched samples using your textbook formula.

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. Using the unpaired t-test, we can obtain an interval estimate of the difference between two population means.

Example

In the data frame column mpg of the data set mtcars, 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, 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.

Example

We used exploratory techniques to identify 92 stars from the Hipparcos data set that are associated with the Hyades. We did this based on the values of right ascension, declination, principal motion of right ascension, and principal motion of declination. We then excluded one additional star with a large error of parallax measurement:

```
#> hip <- read.table("http://astrostatistics.psu.edu/datasets/HIP_star.dat",
#  header=T, fill=T)
> hip <- read.table("HIP_star.dat", header=T, fill=T)
> attach(hip)
> filter1 <-  (RA>50 & RA<100 & DE>0 & DE<25)
> filter2 <- (pmRA>90 & pmRA<130 & pmDE>-60 & pmDE< -10)
> filter  <-  filter1 & filter2 & (e_Plx<5)
> sum(filter)
```

In this section of the tutorial, we will compare these Hyades stars with the remaining stars in the Hipparcos dataset on the basis of the color (B minus V) variable. That is, we are comparing the groups in the boxplot below:

```
> color <- B.V
>  boxplot(color~filter,notch=T)
```

For ease of notation, we define vectors H and nH (for "Hyades" and "not Hyades") that contain the data values for the two groups.

```
> H <- color[filter]
> nH <- color[!filter & !is.na(color)]
```

In the definition of nH above, we needed to exclude the NA values.

A two-sample t-test may now be performed with a single line:

```
> t.test(H,nH)
```

Because it is instructive and quite easy, we may obtain the same results without resorting to the t.test function. First, we calculate the variances of the sample means for each group:

```
> v1 <- var(H)/92
> v2 <- var(nH)/2586
> c(var(H),var(nH))
```

The t statistic is based on the standardized difference between the two sample means. Because the two samples are assumed independent, the variance of this difference equals the sum of the individual variances (i.e., v1+v2). Nearly always in a two-sample t-test, we wish to test the null hypothesis that the true difference in means equals zero. Thus, standardizing the difference in means involves subtracting zero and then dividing by the square root of the variance:

```
> tstat <- (mean(H)-mean(nH))/sqrt(v1+v2)
> tstat
```

To test the null hypothesis, this t statistic is compared to a t distribution. In a Welch test, we assume that the variances of the two populations are not necessarily equal, and the degrees of freedom of the t distribution are computed using the so-called Satterthwaite approximation:

```
> (v1 + v2)^2 / (v1^2/91 + v2^2/2585)
```

The two-sided p-value may now be determined by using the cumulative distribution function of the t distribution, which is given by the pt function.

```
> 2*pt(tstat,97.534)
```

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.

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 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%.