Chapter 19

Tests of Hypotheses for Equality of Two Population Means

In this chapter, we will discuss the following topic:

- How to compare populations means from two Normal distributions with unknown variances using the R-function **t.test()**.

Two-Sample *t* test

If X and Y are two independent simple random samples from two Normal populations, we type:

- `t.test(X,Y)`

  to test whether their populations means are equal. The test incorporated is the Welch's t test for which the degrees of freedom is approximated using the Welch-Satterthwaite equation. The test is valid for samples from two normal populations having possible unequal variances.

- If the two populations in question have the same population variance and the population distributions are exactly Normal, we can type:

  `t.test(X,Y,var.equal=TRUE)`

  Adding the argument **var.equal=TRUE**, specifies that we are using a t test with the pooled variance estimate.

Refer to chapter 18 regarding the possible optional arguments in the t test.

Let $X = (X_1, ..., X_n)$ and $Y = (Y_1, ..., Y_m)$ be two independent simple random samples from two large Normal populations with means μ_x and μ_y, respectively. To test the null hypothesis that their populations means are equal, that is, $H_0 : \mu_x = \mu_y$ or equivalently $H_0 : \mu_x - \mu_y = 0$, we calculate the two-sample T statistics

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}}.$$

The two-sample T statistics has approximately a t distribution with degrees of freedom approximated by the Welch-Satterthwaite equation when the null hypothesis is true. This degrees of freedom is given by

$$v = \frac{\left(\frac{s_x^2}{n} + \frac{s_y^2}{m}\right)^2}{\frac{1}{n-1}\left(\frac{s_x^2}{n}\right)^2 + \frac{1}{m-1}\left(\frac{s_y^2}{m}\right)^2}, \tag{0.1}$$

where s_x and s_y are the sample standard deviations of the samples X and Y, respectively. Notice here that the variances from the two populations are allowed to be different. This test

is called the Welch's t test.
We reject H_0 when the observed value $t = \frac{\bar{x}-\bar{y}}{\sqrt{\frac{s_x^2}{n}+\frac{s_y^2}{m}}}$ of the T-statistics falls within the critical region which is equivalent to having a p-value smaller than α.
The critical region is given by the values of t for which $t > t^*$ for the alternative hypothesis $H_a : \mu_x - \mu_y > 0$, $t < -t^*$ for the alternative hypothesis $H_a : \mu_x - \mu_y < 0$, and finally $|t| > t^*$ for the alternative hypothesis $H_a : \mu_x - \mu_y \neq 0$. Here t^* is the critical value of the approximated t distribution with degrees of freedom approximated by the Welch-Satterthwaite equation given in (0.1).

A level C confidence interval for the difference in population means, $\mu_x - \mu_y$, is given by

$$\bar{x} - \bar{y} \pm t^* \sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}},$$

where t^* is defined such that the area under the density curve between $-t^*$ and t^* is C. Here again t^* is the critical value of the approximated t distribution with degrees of freedom approximated by the Welch-Satterthwaite equation given in (0.1).

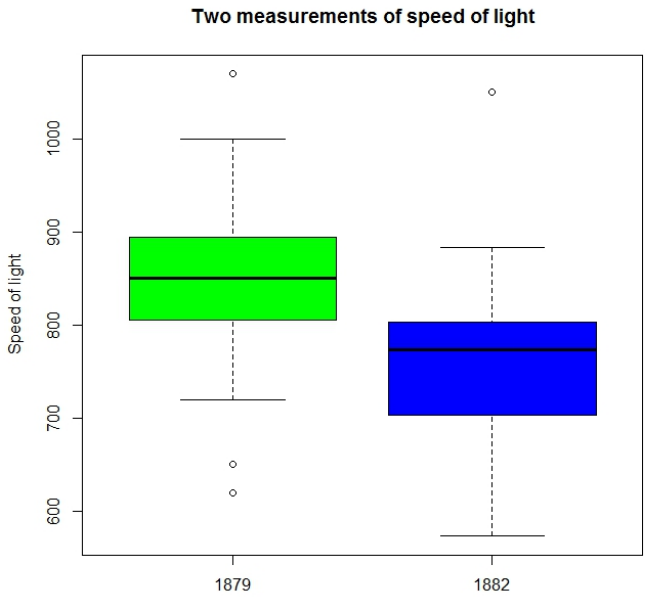
If the two Normal populations have the same variance, we can estimate the variance by using a pooled variance. The pooled variance includes the sample variances from both samples. The resulting T statistics when the null hypothesis is true has an exact t distribution with $n + m - 2$ degrees of freedom.

The Welch's t test is more safe to use since it does not require any previous knowledge of the variance. In addition, the Welch's t procedure is not so dependent of having the underlying distribution being exactly Normal.

Problem. In 1879 and 1882 Michelson [1] determined the speed of light. He had n = 100 observations in 1879 and n=23 observations in 1882 (he subtracted 299000 from the values and the units are km/sec).
Is there a difference in the mean of the speed of light of the 1879 observations and the 1882 observations at the 1% confidence level? Determine a 99% confidence interval for the difference in the populations means.

```
Vel1:
 850  740  900 1070  930  850  950  980  980  880 1000  980  930  650  760
 810 1000 1000  960  960  960  940  960  940  880  800  850  880  900  840
 830  790  810  880  880  830  800  790  760  800  880  880  880  860  720
 720  620  860  970  950  880  910  850  870  840  840  850  840  840  840
 890  810  810  820  800  770  760  740  750  760  910  920  890  860  880
 720  840  850  850  780  890  840  780  810  760  810  790  810  820  850
 870  870  810  740  810  940  950  800  810  870

Vel2:
883  816  778  796  682  711  611  599 1051  781  578  796  774  820  772
696  573  748  748  797  851  809  723
```

Solution. We first include a boxplot of the two samples:



It can be checked that the first data set is approximately Normal while the second data set is skewed with a possible outlier. However, since we have large sample sizes, we can use the t test. The sample variances looks somewhat different. We apply the two-sided two-sample Welch's t test.

Let X=vel1 and Y=vel2. We test

$$H_0 : \mu_X = \mu_Y \quad \text{against} \quad H_a : \mu_X \neq \mu_Y.$$

We obtain:

```
> data=read.delim("C:/Folder/DifferenceVel.txt")
> Vel1=data$Vel1
> Vel2=data$Vel2
> t.test(Vel1,Vel2,conf.level=0.99)

        Welch Two Sample t-test

data:  Vel1 and Vel2
t = 4.0598, df = 27.754, p-value = 0.0003625
alternative hypothesis: true difference in means is not equal to 0
99 percent confidence interval:
  30.67544 161.68977
sample estimates:
mean of x mean of y
 852.4000  756.2174
```

Since the p-value= $0.0003625 < 0.05$, we reject H_0 at the 1% level. Thus, there is a difference in the means of the speed of light of the 1879 observations and the 1882 observations at the 1% confidence level. The 99% confidence interval for the difference in the means $\mu_x - \mu_y$ is (30.67544, 161.68977). Here is the R-code for the boxplot:

```
> Speed_light=data.frame(Vel1,Vel2)
> colnames(Speed_light)=c("1879","1882")
> boxplot(Speed_light,ylab="Speed of light",col=c("green","blue"),
+ main="Two measurements of speed of light")
```

Explanation. Recall the following from chapter 2:

- We put the two data sets in a data frame and name their columns. Then each boxplot will include the individual name of each of the data sets. We also added two different colors to the two boxplots.

Problem. The table below shows the women labor participation rate in 1968 and 1972 in 19 U.S citites [2]:

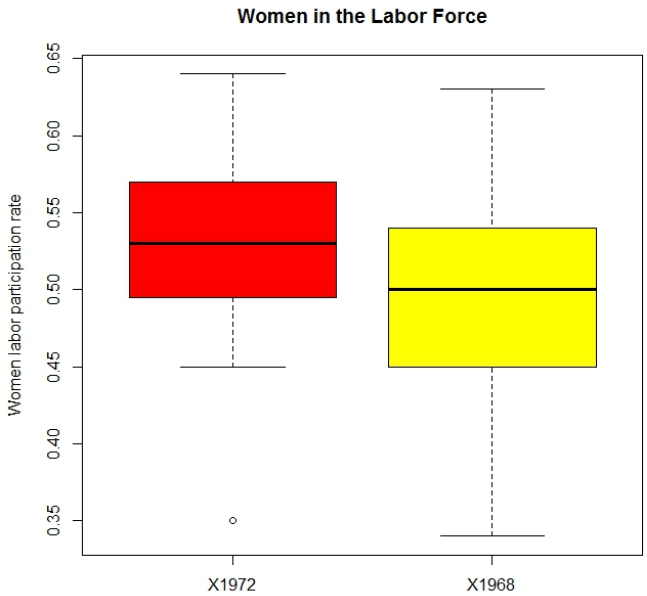
```
            City X1972 X1968
1           N.Y.  0.45  0.42
2           L.A.  0.50  0.50
3        Chicago  0.52  0.52
4   Philadelphia  0.45  0.45
5        Detroit  0.46  0.43
6  San Francisco  0.55  0.55
7         Boston  0.60  0.45
8          Pitt.  0.49  0.34
9      St. Louis  0.35  0.45
10   Connecticut  0.55  0.54
11   Wash., D.C.  0.52  0.42
12         Cinn.  0.53  0.51
13     Baltimore  0.57  0.49
14        Newark  0.53  0.54
15  Minn/St.Paul  0.59  0.50
16       Buffalo  0.64  0.58
17       Houston  0.50  0.49
18     Patterson  0.57  0.56
19        Dallas  0.64  0.63
```

Did the labor force participation rate of women increase from 1968 to 1972 at the 5% significance level?

Solution. We first create a stemplot and boxplot of the data:

```
> table=read.delim("labor_force.txt")
> boxplot(table[,2:3],col=c("red","yellow"),
+ ylab="Women labor participation rate",main="Women in the Labor Force")
```

```
> attach(table)
> stem(X1972)

  The decimal point is 1 digit(s) to the left of the |

  3 | 5
  4 | 5569
  5 | 00223355779
  6 | 044

> stem(X1968)

  The decimal point is 1 digit(s) to the left of the |

  3 | 4
  4 | 22355599
  5 | 001244568
  6 | 3
```



There is no signs of departure from Normality but there is a possible outlier in the 1972 data. However, since there was not much change in the U.S. from 1968 to 1972, we assume that the population variances are equal and perform a pooled t test. We wish to test the null hypothesis that there is no significant difference in labor force participation rate of women from 1968 to 1972 against the alternative hypthesis that the labor force participation rate of

women was higher in 1972 compared to 1968. Let X=X1972 and Y=X1968. We test

$$H_0 : \mu_X = \mu_Y \quad \text{against} \quad H_a : \mu_X > \mu_Y.$$

We have:

```
> t.test(X1972,X1968,alternative=c("greater"),var.equal=TRUE)

        Two Sample t-test

data:  X1972 and X1968
t = 1.4959, df = 36, p-value = 0.0717
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.004333581          Inf
sample estimates:
mean of x mean of y
0.5268421 0.4931579
```

Since the p-value =0.0717, there is no significant difference in labor participation rate for women at the 5% significance in 1968 compared to 1972.

However, since the data was obtained from the same cities for both 1968 and 1972, the data are paired, so it will be better to use a Matched pairs test. Let us see if we have enough evidence against the null hypothesis:

```
> t.test(X1972,X1968,alternative=c("greater"),paired=T)

        Paired t-test

data:  X1972 and X1968
t = 2.4577, df = 18, p-value = 0.01218
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.009917895         Inf
sample estimates:
mean of the differences
             0.03368421
```

Since the p-value is 0.01218, there is a significant difference at the 5% level in labor participation rate for women in 1968 and 1972. This is an example of how a matched pairs test provide a more effective test when its conditions are satisfied.

References

[1] S. M. Stigler, *Do robust estimators work with real data?* Annals of Statistics 5, 1055-1098. (See Table 6.), 1977. The data and the story are also listed on DASL at http://lib.stat.cmu.edu/DASL/

[2] United States Department of Labor Statistics
The data and the story are also listed on DASL at http://lib.stat.cmu.edu/DASL/

00