Introduction to R　Part4

Satoru UCHIDA (Visiting Student Researcher, UC Berkeley)
http://realize.jounin.jp/R.html
2010/3/19

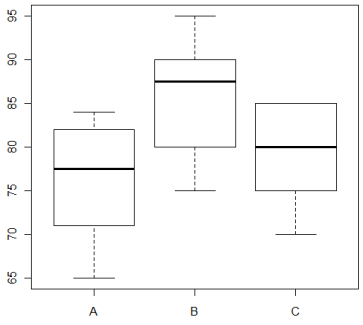
1. Comparing the averages of more than two groups

Remember that we use t-test for comparing two averages. But this cannot be applied to more than two averages.→ANOVA

<Data>　coffee3

CoffeeType	Score
A	80
A	72
A	65
A	71
A	75
A	71

…



Step1 set up H_0 and H_1

H_0: There is no difference among coffee A, coffee B and coffee C.
H_1: There is a difference either between coffee A and coffee B, coffee A and coffee C, and coffee B and coffee C.

Step2 determine the significance level
This time, let's set $p < .05$

Step3 Conduct an analysis
We use **aov** (analysis of variance) function.

```
> coffee3=read.delim("clipboard",header=T)
> boxplot(coffee3$Score~coffee3$CoffeeType)
> aov(coffee3$Score~coffee3$CoffeeType)

Call:
    aov(formula = coffee3$Score ~ coffee3$CoffeeType)
Terms:
                    coffee3$CoffeeType Residuals
Sum of Squares                552.4667 1018.2000
Deg. of Freedom                      2        27

Residual standard error: 6.140937
Estimated effects may be unbalanced
#This is difficult to interpret. So use the following command:
> summary(aov(coffee3$Score~coffee3$CoffeeType))
                          Df  Sum Sq Mean Sq F value    Pr(>F)
coffee3$CoffeeType     2  552.47 276.233   7.325 0.002875 **
Residuals                 27 1018.20  37.711
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Step4 Interpret the result
p-value is less than 0.05. Therefore, we reject the H_0 and there is exist a difference among coffee A, coffee B and coffee C.

Think
We now learn that there is a difference. However, is it between A and B, A and C, B and C or ALL?

```
> TukeyHSD(aov(coffee3$Score~coffee3$CoffeeType))
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = coffee3$Score ~ coffee3$CoffeeType)

$`coffee3$CoffeeType`
      diff        lwr        upr     p adj
B-A 10.2   3.390749 17.0092512 0.0026149
C-A  2.9  -3.909251  9.7092512 0.5489356
C-B -7.3 -14.109251 -0.4907488 0.0337845
```

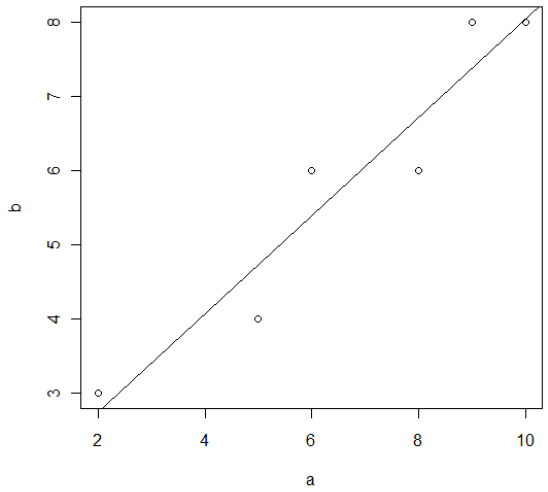
Paired-data
If the data is paired, we use conduct the analysis as follows: (based on the coffee4 data)

```
> TukeyHSD(aov(coffee4$Score~coffee4$CoffeeType+coffee4$No))
```

2. Testing correlation

<data>

```
> a=c(5,6,10,8,9,2)
> b=c(4,6,8,6,8,3)
> plot(b~a)
> abline(lm(b~a))
```



Step1 set up H_0 and H_1

H_0: There is no correlation between a and b.
H_1: There is a correlation between a and b.

Step2 determine the significance level
This time, let's set $p < .05$

Step3 Conduct an analysis
We use **cor.test** (correlation test) function.

```
> cor.test(a,b)

        Pearson's product-moment correlation

data:  a and b
t = 6.3704, df = 4, p-value = 0.003114
```

alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6314654 0.9951235
sample estimates:
      cor
0.9540843

Step4 Interpret the result
p-value is less than 0.05. Therefore, we reject the H_0 and there is a correlation between a and b.

3. Using statistics for your own analysis

(1) We conducted a questionnaire which consists of four answers in England and USA.
How many times have you been to abroad?

a. never
b. 1-5 times
c. 5-10 times
d. More than 10 times

We got the following data:

ans/country	England	USA
a	25	24
b	30	34
c	18	20

Can we say that there is a difference between England and USA? [$p < .05$]
H_0:
H_1:
used function:
p value:
conclusion:

(2) Tests were conducted with respect to aspiration sensitivity among Chinese, Japanese and Korean. The scores are as follows:

Subject	Country	Score
sub1	Chinese	540
sub2	Chinese	450
sub3	Chinese	600

…
…
The higher the score of the test, the more sensitive the subject is to the sound of aspiration. Are there any differences among them? If yes, where is the difference? [$p < .05$]

H_0:
H_1:
used function:
p value:
conclusion:

(3) We conducted two tests as to the pronunciations of vowels and consonants of English learners. The score is judged by a native speaker of English (/100). The results are following:

Subject	Vowel	Consonant
A	80	95
B	54	60
C	75	81
D	60	55

Can we say that the scores of these two have some relationship? [$p < .05$]
H_0:
H_1:
used function:
p value:
conclusion:

(4) A teacher wanted to know how effective the new method of teaching is. So he conducted a pre-word test for his students and also post-test after his lecture. Can we say that his new method is effective enough? [$p < .05$]

Student	Pre	Post
A	7	9
B	4	5
C	11	10
D	12	17

…

H_0:
H_1:
used function:
p value:
conclusion: