مقدمة مبسطة في لغة R مع تطبيقات إحصائية

Introduction to R Language with Statistical Applications

سبأ محمد علوان

T- tests

The **t.test( )** function produces a variety of t-tests. Unlike most statistical packages, the default assumes unequal variance and applies the Welsh df modification.

```
# independent 2-group t-test
t.test(y1,y2) # where y1 and y2 are numeric
```

the same with
t.test(x,y, paired=FALSE)

```
# paired t-test
t.test(y1,y2,paired=TRUE) # where y1 & y2 are numeric
```

```
# one sample t-test
t.test(y,mu=3) # Ho: mu=3
```

Example

```
> x<-c(1,2,3,4,5,6,7,8,9)
> y<-c(3,4,5,6,7,8,9,11,14)
> t.test(x,y)

                Welch Two Sample t-test

data:  x and y
t = -1.649, df = 15.118, p-value = 0.1198
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.6019780  0.7130891
sample estimates:
mean of x mean of y
 5.000000  7.444444
```

```
> t.test(x,y,paired=FALSE)

        Welch Two Sample t-test

data:  x and y
t = -1.649, df = 15.118, p-value = 0.1198
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.6019780  0.7130891
sample estimates:
mean of x mean of y
 5.000000  7.444444
```

Example

```
> t.test(x,y)

        Welch Two Sample t-test

data:  x and y
t = -1.649, df = 15.118, p-value = 0.1198
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.6019780  0.7130891
sample estimates:
mean of x mean of y
 5.000000  7.444444

> t.test(x,mu=6)

        One Sample t-test

data:  x
t = -1.0954, df = 8, p-value = 0.3052
alternative hypothesis: true mean is not equal to 6
95 percent confidence interval:
 2.894916 7.105084
sample estimates:
mean of x
        5
```

تحليل التباين في اتجاه واحد :
One Way ANOVA

يعد تحليل التباين ANOVA, Analysis of Variance واحدا من أكثر الأدوات شيوعا بالاستخدام في جعبة العاملين بالتحليل الإحصائي، لذا سنقوم في هذه المقالة بتغطية موضوع تنفيذ تحليل التباين باستخدام لغة R وتفسير النتائج التي سنحصل عليها نتيجة تطبيق مثل هكذا تقنية وذلك من وجهة نظر تطبيقية بحتة تبتعد عن التجريد الرياضي وتركز على النواحي العملانية.

تعتبر تقنية تحليل التباين تعميما لما سبق وأن تعاملنا معه في اختبار t والذي ينحصر في تحديد معنوية الفروقات فيما بين مجموعتين اثنتين فقط من البيانات، في حين يستطيع تحليل التباين التعامل مع بيانات مقسمة إلى أكثر من مجموعتين بناء على معامل factor أو أكثر.

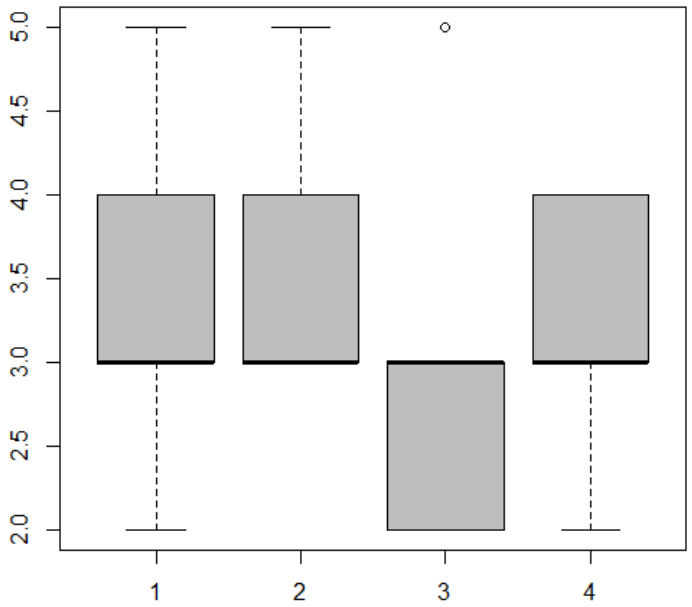
تحليل التباين في اتجاه واحد :
One Way ANOVA

Example

D	E	F	G
2	3	3	2
3	4	2	4
4	5	3	3
5	3	2	4
3	3	5	3

اختبر ما إذا كان هناك فروق معنوية بين متوسطات المجموعات الاربع في الاعمدة المقابلة

```
> bloxplot(x ~ Y, col="gray")
```



واضح من الشكل المعروض أعلاه وجود فروق ظاهرية لمقدار الصفة X بين المستويات الاربعة المختلفة، لكننا بحاجة إلى إثبات ذلك بدليل عددي

اختبر ما إذا كان هناك فروق معنوية بين متوسطات المجموعات الاربع في الاعمدة المقابلة

D	E	F	G
2	3	3	2
3	4	2	4
4	5	3	3
5	3	2	4
3	3	5	3

النتائج باستخدام Excel

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Column 1	5	17	3.4	1.3
Column 2	5	18	3.6	0.8
Column 3	5	15	3	1.5
Column 4	5	16	3.2	0.7

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	1	3	0.333333	0.310078	0.817795	3.238872
Within Groups	17.2	16	1.075			
Total	18.2	19				

```
> x<-c(2,3,4,5,3,3,4,5,3,3,3,2,3,2,5,2,4,3,4,3)
> Y<-c("1", "1", "1", "1", "1", "2","2", "2", "2","2", "3","3","3","3","3", "4","4", "4", "4","4")
> model<- aov(x ~ Y)
> summary(model)
            Df Sum Sq Mean Sq F value Pr(>F)
Y            3    1.0  0.3333    0.31  0.818
Residuals   16   17.2  1.0750
```

النتائج باستخدام R

```
> Y<-c("A","A","A","A","A", "B","B","B","B","B","C","C","C","C","C","D","D","D","D","D")
> model<- aov(x ~ Y)
> summary(model)
            Df Sum Sq Mean Sq F value Pr(>F)
Y            3    1.0  0.3333    0.31  0.818
Residuals   16   17.2  1.0750
```

لن تختلف النتائج لو عرفنا مستويات Y بحروف فالمتغير Nominal كما نعلم

المقارنات المتعددة البعدية Multiple Comparison.

```
mc<- TukeyHSD(model, "Y", ordered=TRUE)
```

المقارنات البعدية في المثال السابق

```
> mc<- TukeyHSD(model, "Y", ordered=TRUE)
> mc
  Tukey multiple comparisons of means
    95% family-wise confidence level
    factor levels have been ordered

Fit: aov(formula = x ~ Y)

$Y
    diff       lwr      upr     p adj
D-C  0.2 -1.676096 2.076096 0.9897700
A-C  0.4 -1.476096 2.276096 0.9274133
B-C  0.6 -1.276096 2.476096 0.7972612
A-D  0.2 -1.676096 2.076096 0.9897700
B-D  0.4 -1.476096 2.276096 0.9274133
B-A  0.2 -1.676096 2.076096 0.9897700
```

كتابة الدوال ..... Writing functions

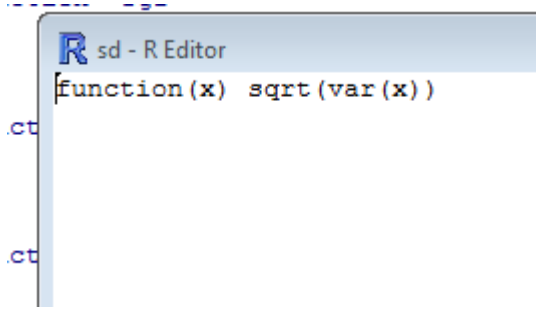
يمكن أنشاء دالة خاصة بالباحث أو المبرمج .. خاصة لتلك التي يستخدما باستمرار و ربما تتضمن الكثير من الاجراءات المرتبة

الطريقة الاولى : مثلا يوجد في لغة R دالة var() التي تحسب التباين لمجموعة من البيانات ، لكن دعنا نوجد دالة تحسب الانحراف المعياري ... و لنقم بذلك الان

```
> sd <- function(x) sqrt(var(x))
> x<-c(9,5,2,3,7)
> sd(x)
[1] 2.863564
> y<-c(4,5,6,7)
> sd(y)
[1] 1.290994
```

لتعديل الدالة ما عليك سوى كتابة fix(sd) الذي يفتح لك الدالة لتقوم بتحريرها أو التعديل عليها

```
> fix(sd)
```



الطريقة الثانية: تعريف الدالة وفقا للشكل التالي

```
fix(several.plots)
```

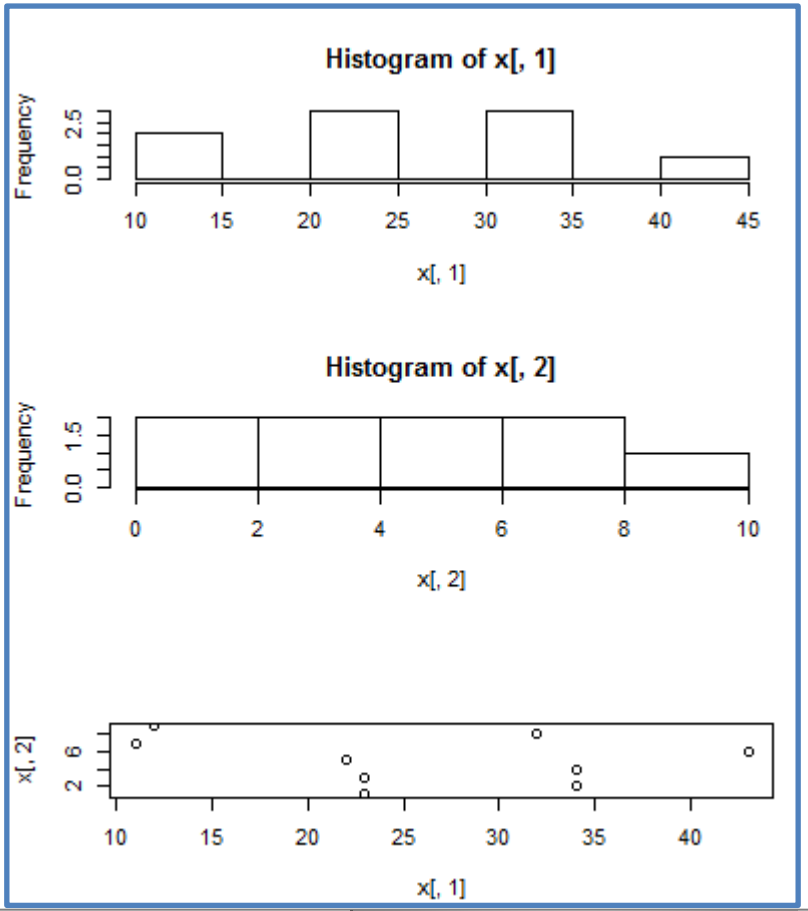
وممكن من خلال هذا الامر تفتح نافذة لعمل تغييرات فيها

several.plots - R Editor

```
function ()
{
}
```

```
function(x){
par(mfrow=c(3,1))
hist(x[,1])
hist(x[,2])
plot(x[,1],x[,2])
par(mfrow=c(1,1))
apply(x,2,summary)
}
```

```
> x<- c(23,34,23,34,22,43,11,32,12)
> y<-c(1,2,3,4,5,6,7,8,9)
> z<-cbind(x,y)
> several.plots(z)
          x y
Min.     11 1
1st Qu.  22 3
Median   23 5
Mean     26 5
3rd Qu.  34 7
Max.     43 9
```

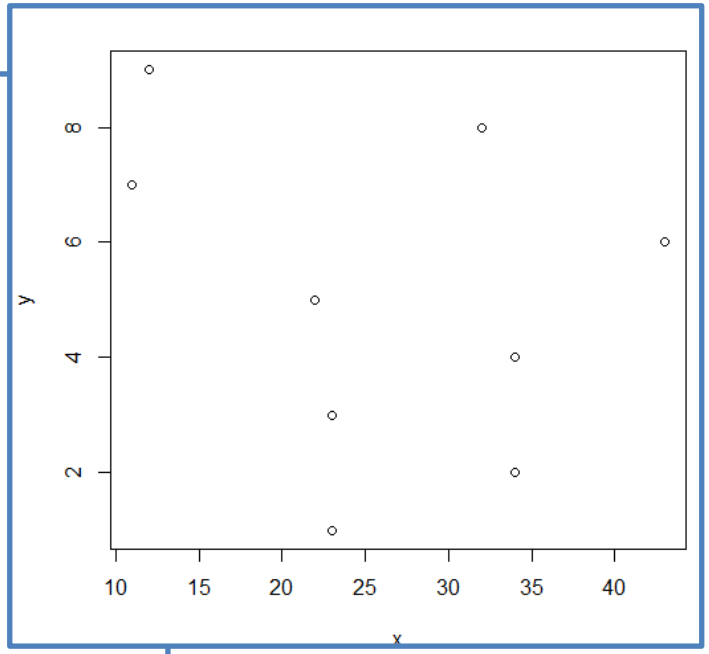


مثال آخر : دعنا نكون دالة تقوم برسم شكل الانتشار لمتغيرين ، ثم تقوم بإيجاد معامل الارتباط بينهم و من ثم تحديد مدى قوة العلاقة الخطية بينهم و إيجاد نموذج الانحدار الخطي المناسب

```
function (z)
{
z<- cbind(x,y)
cor(x, y)
plot(x, y)
cor.test(x,y)
fit<-lm(x ~y)
summary (fit)
}
```

تكون النتائج كما يلي:

شكل الانتشار




```
> fix(corr.plot)
> x<- c(23,34,23,34,22,43,11,32,12)
> y<-c(1,2,3,4,5,6,7,8,9)
> corr.plot(z)

Call:
lm(formula = x ~ y)

Residuals:
    Min      1Q  Median      3Q     Max
-12.833  -7.333  -4.000   6.917  18.083

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   31.417      7.947   3.953  0.00551 **
y             -1.083      1.412  -0.767  0.46808
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.94 on 7 degrees of freedom
Multiple R-squared:  0.07755,   Adjusted R-squared:  -0.05423
F-statistic: 0.5885 on 1 and 7 DF,  p-value: 0.4681
```

Exercises

1. Write a function that takes as its argument two vectors, `x` and `y`, produces a scatterplot, and calculates the correlation coefficient (using `cor(x,y)`).

2. Write a function that takes a vector $(x_1, \ldots, x_n)$ and calculates both $\sum x_i$ and $\sum x_i^2$. (Remember the use of the function `sum`).

مثال : حساب بعض التكاملات :

$$\int_0^{\infty} r\, e^{-r*x} \quad dx$$

```
> r<- 0.5
> f<- function (x) {r*exp(-r *x)}
> integrate(f, lower=0, upper= 5)#P(x<5):x~exp(r)
0.917915 with absolute error < 1e-14
> integrate(f, lower=0, upper= Inf)
1 with absolute error < 3.4e-05
```

بعض المراجع :

https://academy.hsoub.com/tags/%D9%85%D8%AF%D8%AE%D9%84%20%D8%A5%D9%84%D9%89%20%D9%84%D8%BA%D8%A9%20r/

https://www.programiz.com/r-programming/for-loop

https://www.tutorialspoint.com/r/

https://www.youtube.com/