جامعة الملك سعود
قسم الإحصاء وبحوث العمليات

مقدمة للحزم لإحصائية

بإستخدام

R , Python and Excel Data Analysis

تأليف
د. عدنان ماجد عبد الرحمن بري
استاذ الأنظمة العشوائية الحركية

أنواع البيانات و مستويات القياس Data Types and Levels of Measurement

تعريفات:

البيانات Data: هي أي قياسات أو معلومات عن ظاهرة محددة أو متغير محدد. مثل مجموعة ارقام لقياسات أطوال تلاميذ فصل ما. أو الوان عدد من السيارات في صالة عرض ما. أو مستوى تعليم مجموعة من ركاب طائرة ما. او درجات الحرارة في عدد من عواصم العالم. الخ. وتسمى أي مجموعة متجانسة من البيانات بالعينة.

الإحصائة Statistic: هي عدد او صفة مستخرجة من البيانات

نوع البيانات يحدد نوع الإحصائة التي يمكن حسابها . فالإحصائة هي قيمة او صفة مستخرجة (أو محسوبة) من البيانات. مثل متوسط أطوال الطلاب. عدد السيارات ذات اللون الأبيض. عدد الجامعيين من ركاب الطائرة. وسيط درجات الحرارة في العواصم.

من المهم جدا معرفة نوع البيانات لكي نحدد الإحصاآت التي يمكن إستخراجها ونوع التحليل الإحصائي الذي يمكن إجرائه.

تقسم أنواع البيانات إلى أربع أقسام هي:

1- القياس الإسمي Nominal Scale.

2- القياس الترتيبي Ordinal Scale.

3- القياس الفتري Interval Scale.

4- القياس النسبي Ratio Scale.

أولا:

القياس الإسمي: وهو أدنى نوع من القياسات المستخدمة في الإحصاء وهو عبارة عن تصنيف أو وضع البيانات في فئات Categories أو عوامل Factors بدون اي ترتيب Order أو تشكيل Structure.

أمثلة:

1- الوان 10 سيارات في معرض:

أسود ، أسود ، رصاصي ، أحمر ، أزرق ، رصاصي ، أزرق ، أزرق ، أبيض، أبيض.

2- نوع شراب 8 أصدقاء في مقهى:

شاهي ، شاهي ، قهوة ، شاهي ، قهوة ، قهوة ، قهوة ، شاهي.

3- إخوة 5 كان نوع المولود الأول لكل منهم:

ولد ، ولد ، بنت ، ولد ، بنت

أنواع ألإحصاآت والتحليل الإحصائي البسيط للمتغيرات الاسمية:

1- إيجاد المنوال.

2- إيجاد نسبة صفة.

3- التحليل بواسطة الجدولة البينية Crosstabulation مع مربع كاي Chi-square.

ثانيا:

القياس الترتيبى:

ويأتي التالي في مستوى القياسات من حيث قوة القياس عن القياس الإسمي حيث هو عبارة عن قياس إسمي Nominal يوجد به تركيب أو تشكيل Structure أكثر عبارة عن ترتيب Ranking للصفات. هذا الترتيب غير موضوعي أو ظاهري في المسافات بين الصفات فمثلا لو سألك باحث تسويقي عن خدمة تسويقية بحيث تدرجها من (جيدة جدا ، جيدة ، حيادي ، سيئة ، سيئة جدا) فهذا يمثل قياس ترتيبي و لايوجد ترتيب ظاهري بين التدريجات فمثلا جيدة جدا بالنسبة لك أعلى بكثير من جيدة جدا لشخص آخر طرح علية نفس السؤال.

أمثلة:

1- مستوى التعليم ل 10 ركاب في رحلة ما:

ثانوي ، جامعي ، جامعي ، ثانوي ، متوسط ، جامعي ، جامعي ، ثانوي ، ثانوي ، ثانوي.

2- رتب 5 أشخاص في السلك العسكري:

جندي ، جندي ، جندي ، عريف ، ضابط.

أنواع الإحصاآت والتحليل الإحصائي البسيط للمتغيرات الترتيبية:

1- المنوال.

2- الوسيط.

3- التحليل بواسطة الجدولة البينية مع مربع كاي.

ثالثا:

القياس الفترى:

ويأتي التالي في مستوى القياسات من حيث قوة القياس عن القياس الترتيبي حيث هو عبارة عن قياس ترتيبي يوجد به تركيب أو تشكيل Structure أكثر عبارة عن ترتيب Ranking للصفات. وهذا الترتيب موضوعي أو ظاهري في المسافات بين الصفات فالفرق بين صفة والتالية لها ثابتة بين كل الصفات ولها تدريج قياسي Metric معين مثل السنتيمتر و الكيلو الخ

أنواع الإحصاآت والتحليل الإحصائي البسيط للمتغير الفتري:

1- المتوسط و الإنحراف المعياري.

2- الترابط و الإنحدار.

3- تحليل التباين.

رابعاً:

القياس النسبي:

وهو أعلى مستويات القياس وهو قياس فتري حيث يوجد للظاهرة المقاسة نقطة بداية حقيقية وهي الصفر. مثل الطول و الوزن و الدخل الشهري الخ.

لكي نوضح الفرق بين مستويي القياس الفتري و النسبي مثلا درجات الحرارة تقاس على المستوى الفتري فدرجات الحرارة المؤوية لها صفر مئوي ولكنه صفر إختياري حيث القياس الفارنهايتي صفره عند 32 درجة مئوية.

أنواع ألإحصاآت والتحليل الإحصائي البسيط:

نفسه للقياس الفتري.

أنواع الإستجابة و أنواع المتغيرات Types of Response Scales:

1- **متقطع Discerete**

ويشمل جميع القياسات التي تقاس على المستويين الإسمي و الترتيبي.

ففي المقياسي الفتري و النسبي تكون القياسات متقطعة إذا كانت تتشكل من اعداد صحيحة. مثل عدد المتسوقين و عدد الأشجار في طريق الخ.

2- **متصل Continuous**

وهي قياسات تقاس على المستويين الفتري و النسبي حيث يمكن قياس ظاهرة او متغير على هذين المستويين لأي دقة عددية نريد. مثل الطول و الوزن الخ.

مقدمة للغة النمذجة R

تثبيت البرنامج و الحزم Installing R and Packages

مقدمة

يعتبر R بيئة برمجية Programming Environment و لغة نمذجة رياضية Mathematical Modelling Language ويتميز بالتالي:

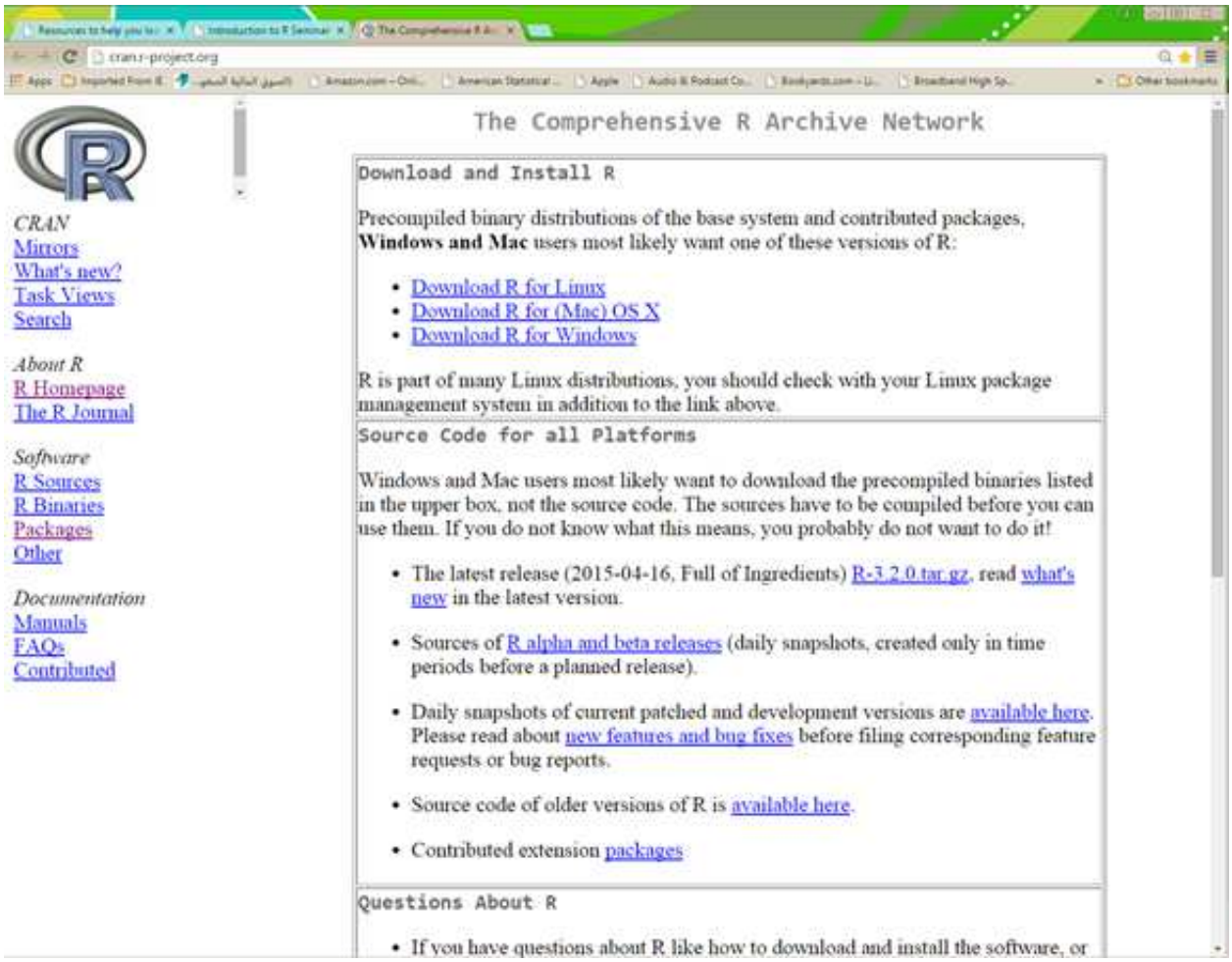
1) إستخدام لغة برمجة بسيطة و عالية التطوير.
2) يسمح بتطويرسريع لأدوات جديدة حسب طلب المستخدم.
3) هذه الأدوات توزع على شكل حزم packages و التي يمكن تحميلها لإضافة خواص جديدة للبرنامج.
4) تتواجد حزم كثيرة للنمذجة في جميع العلوم والتطبيقات العملية والتقنية.
5) من برامج المصدر المفتوح open-source ويتم تطويرة و صيانته من قبل عدد كبير جدا من الأكاديميين و الباحثين.
6) يعمل على أنظمة تشغيل مثل windows و MacOS و Unix وغيرها
7) البرنامج الأساسي Base R ومعظم الحزم packages (وتسمى أيضا مكتبات Libraries) متوفرة للتنزيل من شبكة الأرشيف الشاملة لـ R Comprehensive R Archive Network (CRAN) (8

9) http://cran.r-project.org/

10) البرنامج الأساسي تتوفر فيه أساسيات إدارة البيانات Data Management و التحليل Analysis و أدوات الرسم Graphical Tools

11) ولكن القوة الحقيقية والمرونة القصوى لـ R تأتي من العدد الهائل من الحزم والتي تصل حتى الآن فوق 6400 حزمة لمعظم التطبيقات الإدارية والمالية و الإقتصادية و العلمية والتقنية ألخ.

تنزيل و تثبيت R



أوليات R

إستخدام R كآلة حاسبة

```
> 2 + 5
> sin(2*pi + 10)
> 4*2.4 + (246/3) * 10.5
>
```

تمرين: حاول إجراء مختلف الحسابات. إستكشف مقدرات R كآلة حاسبة.

لإنشاء متغيرات

المتغير a قيمته 5 يعبر عنه كالتالي

```
> a = 5
```

المتغير b قيمته 5 يعبر عنه كالتالي

```
> b <- 5
```

الرموز = أو <- كليهما عمال إسناد لإعطاء قيمة لمعرف identefier ليتم تخزينها و إعادتها عند الحاجة.

لإنشاء متجهات

إنشاء متجه

```
> d = c(3, 4, 5); d
> e = seq(from = 1, to = 3, by = 0.5); e
> f = rep( 7, 6); f
```

بعض الدوال المتجهية المفيدة

```
> length(d)
 > max(d)
 > min(d)
 > mean(d)
```

```
> g <- c(2, 6, 7, 4, 5, 2, 9, 3, 6, 4, 3)
> gg <- sort(g, decreasing = TRUE); gg
> ggg <- sort(g); ggg
```

الكثير من البيانات قد تحوي قيم مفقودة Missing Values والتي قد تسبب بعض الأخطاء في العمليات الحسابية لبعض الدوال. أدرس التالي بعناية

```
> a <- c(1, 2, 3, NA, 6)
```

NA تعني غير متوفرة أو مفقودة Not Available.

```
> is.na(a)
[1] FALSE FALSE FALSE  TRUE FALSE
```

بعض الدوال لا تتقبل القيم المفقودة

```
> mean(a)
[1] NA
> mean(a, na.rm = TRUE)
[1] 3
```

لاحظ حجة الدالة na.rm = TRUE والتي تعطي للدالة mean تعليمات بحذف القيم المفقودة. rm تعني أحذف أو أذل remove

الدالة summary

```
> summary(a)
Min.   1st Qu.  Median   Mean   3rd Qu.   Max.
NA's
```

```
1.00    1.75    2.50       3.00    3.75      6.00
1
```

هذه الدالة تعطي:

(1) القيمة الصغرى Minimum. (2) الربيع الأول 1st Qu أي first quantile.

(3) الوسيط. (4) المتوسط. (5) الربيع الثالث 3rd Qu.

(6) القيمة العظمى Maximum. (7) عدد القيم المفقودة NA's.

إنشاء مصفوفات

```
> mat <- matrix (10:15, nrow = 3, ncol = 2)
> mat
     [,1] [,2]
[1,]   10   13
[2,]   11   14
[3,]   12   15
```

الحجه nrow توجه الدالة لجعل عدد الأسطر 3 والحجة ncol توجهها لجعل عدد الأسطر 2

بعض الدوال المفيدة للمصفوفات

جمع مصفوفتين

```
 > mat + mat
     [,1] [,2]
[1,]   20   26
[2,]   22   28
[3,]   24   30
```

منقول transpose المصفوفة

```
> t(mat)
     [,1] [,2] [,3]
[1,]   10   11   12
[2,]   13   14   15
```

ابعاد مصفوفة

```
> dim(mat)
[1] 3 2
```

ضرب المصفوفات

الضرب عنصر بعنصر

```
> mat * mat
     [,1] [,2]
[1,]  100  169
[2,]  121  196
[3,]  144  225
```

الضرب المتجهي %*%

```
> mat %*% t(mat)
     [,1] [,2] [,3]
[1,]  269  292  315
[2,]  292  317  342
[3,]  315  342  369
```

مجموعات جزئية من المصفوفات

لإسترجاع خلية او عنصر من المصفوفة

```
 > mat[1,2]
[1] 13
```

إسترجاع السطر الثاني

```
> mat[ ,2]
[1] 13 14 15
```

إسترجاع العمود الأول

```
> mat[1, ]
[1] 10 13
```

إسترجاع العناصر 1 و 3 من العمود 2

```
 > mat[c(1,3), 2]
[1] 13 15
```

إنشاء مصفوفات من متجهات

الدالة rbind

```
> d <- seq (1 ,10 ,2)
```

```
> d
[1] 1 3 5 7 9
> mat1 <-rbind (d,d)
> mat1
  [,1] [,2] [,3] [,4] [,5]
d    1    3    5    7    9
d    1    3    5    7    9
>
```

الدالة cbind

```
 > mat2 <-cbind (d,d)
> mat2
     d d
[1,] 1 1
[2,] 3 3
[3,] 5 5
[4,] 7 7
[5,] 9 9
> mat3 <- matrix(c(1, 2), 3, 2, byrow = TRUE)
 > mat3
     [,1] [,2]
[1,]    1    2
[2,]    1    2
[3,]    1    2
```

مجاميع البيات Datasets

لتنزيل بيانات من الإنترنت نستخدم الدالة read.table()

```
> stock.data <-
read.table("http://www.google.com/finance/historica
l?q=NASDAQ:AAPL&output=csv", header=TRUE, sep=",")
>
> head(stock.data)
    ï..Date   Open   High    Low  Close   Volume
1 15-May-15 129.07 129.49 128.21 128.77 38208034
2 14-May-15 127.41 128.95 127.16 128.95 45203456
3 13-May-15 126.15 127.19 125.87 126.01 34694235
4 12-May-15 125.60 126.88 124.82 125.86 48160032
5 11-May-15 127.39 127.56 125.62 126.32 42035757
6  8-May-15 126.68 127.62 126.11 127.62 55550382
```

الخيار header إذا كان TRUE تعني أن السطر الأول من البيانات يحوي أسماء المتغيرات.

الخيار sep يحدد كيفية الفصل بين وحدات البيانات:

sep = “,” يحدد ان البيانات تفصل بينها commas فواصل.

sep = “\t” يحدد ان البيانات بينها تنصيص tab .

sep = “ “ البيانات بينها فراغات space .

ملاحظة: سوف نختار إسم بسيط للبيانات stock.data لسهولة الإستخدام في الأمثلة التالية

```
 s <- stock.data
 > colnames(s)
[1] "ï..Date" "Open"    "High"    "Low"     "Close"
"Volume"
 > s$Open
 [1] 129.07 127.41 126.15 125.60 127.39 126.68 ...
 > s[["High"]]
 [1] 129.49 128.95 127.19 126.88 127.56 127.62  ...
 > x = "Low"
> s[[x]]
[1] 128.21 127.16 125.87 124.82 125.62 126.11 ...
> s[ ,1]
[1] 15-May-15 14-May-15 13-May-15 12-May-15 11-May-
15 8-May-15 ...
> s[2, ]
ï..Date   Open   High    Low  Close   Volume
2 14-May-15 127.41 128.95 127.16 128.95 45203456
> s[3, 3:5]
    High    Low  Close
 3 127.19 125.87 126.01
```

نوع البيانات

```
> class(s)
[1] "data.frame"
```

النوع "إطار بيانات" data.frame

القوائم Lists

وهي متجهات تسمح بتجميع كائنات أو فئات إختيارية مثل

```
> (l = list(a = c(TRUE, FALSE), b = matrix(1:4, 2,
2), "hello"))
$a
[1]  TRUE FALSE
$b
     [,1] [,2]
[1,]    1    3
[2,]    2    4
[[3]]
[1] "hello"
> l$a
[1]  TRUE FALSE
> l[["b"]]
     [,1] [,2]
[1,]    1    3
[2,]    2    4
> l[[3]]
[1] "hello"
> class(l)
[1] "list"
```

الدلة names()

```
> a = c(1, 2, 3, 4)
```

```
> a
[1] 1 2 3 4
> names(a) = c("w", "x", "y", "z")
> a
w x y z
1 2 3 4
> b = matrix(1:4, 2, 2)
> b
     [,1] [,2]
[1,]    1    3
[2,]    2    4
> colnames(b) = c("x", "y")
> b
     x y
[1,] 1 3
[2,] 2 4
> rownames(b) = c("m", "n")
> b
  x y
m 1 3
n 2 4
> b[ , "x"]
m n
1 2
> b["n", ]
x y
2 4
```

تعريف الدوال Defining Functions

تعرف الدوال في R بإستخدام الكلمة المفتاحية keyword و التي هي function كالتالي:

```
> square = function(x) { return(x^2)}
> square(5)
[1] 25
> square(1:5)
[1]  1  4  9 16 25


> cube = function(x) x^3
> cube(2)
[1] 8
> cube(1:5)
[1]   1   8  27  64 125
```

ملاحظة: عامة نضع تعريف الدالة بين { } إذا كان التعريف يمتد على أكثر من سطر واحد.

تعريف قيم إفتراضية Defining Default Values for Arguments

```
> pow = function(x, y = 2) x^y
> pow(2)
[1] 4
```

```
> pow(2, 4)
[1] 16
> pow(y= 4,2)
[1] 16
> pow(y =3, x = 3)
[1] 27
```

تعريف قيم إفتراضية ضمنية بإستخدام (...)

```
> mini.summary = function(...) {
+  n = c(...)
+  return(list(mean = mean(n), median = median(n)))
+  }
> mini.summary(1, 2, 2, 3)
$mean
[1] 2
$median
[1] 2
```

لاحظ أن R تعطي + للدلالة على أن المدخل حتى الآن لم يكتمل تعريفه.

```
> test = function(x, y, ...) {
+  dots = paste(c(...), collapse = " ")
+  print(paste("x=",x," y=",y," ...=",dots,sep =
""))
+  }
> test(1, 2, 3, 4, 5, 6, 7)
```

```
[1] "x=1 y=2 ...=3 4 5 6 7"
> test(1, 2, 3, 4)
[1] "x=1 y=2 ...=3 4"
> test(1, 2, 3)
[1] "x=1 y=2 ...=3"



> test2 = function(x, ...) {
+  l = list(...)
+  n = names(l)
+  for (i in n[n != ""]) {
+  cat(paste(i, "=", l[i]), "\n")
+  }
+  }
> test2(1, 2, z = 3, y = 4)
z = 3
y = 4
> test2(1, 2, mean = 3)
mean = 3
```

التحكم في سير البرنامج و الدورات Flow Control and Loops

if ...else

```
> even.odd = function(x) {
+  if (!is.numeric(x)) {
+  print("neither")  }
+  else if (x%%2 == 0) {
+  print("even") }
+  else {
+  print("odd")
+  }  }
> even.odd(3)
[1] "odd"
> even.odd(4)
[1] "even"
> even.odd("A")
[1] "neither"
```

for

```
> for (x in 1:3) {
+  print(x)
+  }
[1] 1
[1] 2
```

```
[1] 3
> for (x in c("hello", "goodbye")) {
+  print(x)
+  }
[1] "hello"
[1] "goodbye"
> m = matrix(1:4, nrow = 2, ncol = 2)
> for (x in m) print(x)
[1] 1
[1] 2
[1] 3
[1] 4
> d = data.frame(a = c(1, 2), b = "A")
> for (x in d) print(x)
[1] 1 2
[1] A A
Levels: A
> l = list(a = c(1, 2), b = c("A"))
> for (x in d) print(x)
[1] 1 2
[1] A A
Levels: A
```

while

```
> x = 1
```

```
> while (x < 3) {
+  print(x)
+  x = x + 1
+  }
[1] 1
[1] 2
```

apply functions

توجد عائلة من دوال "طبق" apply functions والتي تطبق دالة معطاه على كل عناصر المتجه أو المصفوفة او القائمة الخ. هذه الدوال تكون أسرع بكثير من الدوال المماثلة والتي تستخدم الدورات. ولها التركيب التالي:

```
apply(x, margin, fun, ...)
```

أمثلة:

```
> (x = matrix(1:9, 3, byrow = T))
     [,1] [,2] [,3]
[1,]    1    2    3
[2,]    4    5    6
[3,]    7    8    9
> apply(x,1, mean)
[1] 2 5 8
> apply(x, 2, function(x) sum(x)/length(x))
[1] 4 5 6
> colnames(s)
```

```
[1] "ï..Date" "Open"   "High"   "Low"    "Close"
"Volume"
> sapply(s[ ,3:4], mean)
    High      Low
110.1875 108.2011
> sapply(s[ ,3:4], sd)
    High      Low
13.39210 13.06608
```

إستيراد البيانات من قرص محلي

لمعرفة المجلد الذي تعمل منه R

```
> getwd()
```

إخبار R بالمجلد التي توجد عليه البيانات

```
> setwd("C:/Users/amb/Desktop/R python course")
> getwd()
[1] "C:/Users/amb/Desktop/R python course"
> h2 <- read.table("C:Users/amb/Desktop/R python
course/hsb2.csv")
```

الدالة file.choose تفتح نافذة حوار لإختيار الملف من أي مجلد أو قرص تفاعليا

```
> h2 <- file.choose()
```

إستخدام البيانات الموجودة على R

يأتي مع R وجميع الحزم مجاميع بيانات كثيرة مناسبة لإستخدامها في الأمثلة الكثيرة التي توجد لغرض التوضيح لعمل R أو الحزم.

ولإستخدام بيانات أي حزمة نحملها أولا في R عن طريق الدالة ()library مثل

```
> library ( alr3 )
```

نستطيع مشاهدة اسماء البيانات التي تأتي مع هذه الحزمة

```
> data(package = "alr3")
```

لمشاهدة اسماء كل البيانات الموجودة في الجلسة الحالية نستخدم

```
> data()
```

لتحميل و إستخدام بيانات

```
> data(UN2)
> head(UN2)
             logPPgdp logFertility Purban
Afghanistan  6.614710    2.7655347     22
Albania     10.363040    1.1890338     43
Algeria     10.800900    1.4854268     58
Angola       9.529431    2.8479969     35
Argentina   12.806348    1.2868811     88
Armenia      9.424166    0.2016339     67
```

لإستخدام أسماء المتغيرات عند العمل مع البيانات نستخدم الدالة ()attach

```
> attach(UN2)
> names(UN2)
[1] "logPPgdp"     "logFertility" "Purban"
> ?UN2
> Description
 National health, welfare, and education statistics
for 193 places, mostly UN members, but also other
areas like Hong Kong that are not independent
countries.
```

بعد الإنتهاء من البيانات تخزن نسخة جديدة

```
 > UN2.copy <- UN2
```

تزال من جلسة العمل بالدالة detach()

```
 > detach(UN2)
```

ملخص البيانات

الدالة summary()

```
> summary(UN2)
    logPPgdp       logFertility        Purban
 Min.   : 6.492   Min.   :0.0000   Min.   :  6.00
 1st Qu.: 8.867   1st Qu.:0.9184   1st Qu.: 35.00
 Median :10.920   Median :1.4114   Median : 57.00
 Mean   :10.993   Mean   :1.4687   Mean   : 55.54
 3rd Qu.:12.938   3rd Qu.:2.0909   3rd Qu.: 75.00
 Max.   :15.444   Max.   :3.0000   Max.   :100.00
```

المتوسط

```
> mean(UN2$logPPgdp)
[1] 10.99309
> mean(UN2$logFertility)
[1] 1.468687
> mean(UN2$Purban)
[1] 55.53886
```

الترابط

```
> cor(UN2 [ ,1:2])
              logPPgdp logFertility
logPPgdp      1.000000   -0.677604
logFertility -0.677604    1.000000
```

التفاعل مع R

سوف نستخدم الحزم التالية:

1) Foreign لقرائة البيانات من حزم إحصائية أخرى.
2) xlsx لقرائة دفاتر مكروسوفت إكسل.
3) dplyr لإدارة البيانات.
4) reshape2 لإدارة البيانات أيضا.
5) ggplot2 للرسومات عالية الجودة.
6) GGally لرسم مصفوفات رسوم الإنتشار.

7) vcd لرسم و تحليل البيانات التصنيفية (الإسمية)

لإستخدام الحزم packages في R يجب تثبيتها اولا مستخدمين الدالة:

```
> install.packages
```

والتي تقوم بتنزيل الحزمة من CRAN او أحد مراياتها mirrors ثم تقوم بتثبيتها.

```
> install.packages("foreign")
> install.packages("xlsx")
> install.packages("dplyr")
> install.packages("reshape2")
> install.packages("ggplot2")
> install.packages("GGally")
> install.packages("vcd")
```

ملاحظة: الحزم التي يتم تنزيلها وتثبيتها تظل موجودة دائما في R حتى يتم إزالتها.

عند الحاجة لإستخدام حزمة ما في أحد جلسات R (وتسمى R Session) يجب أن نحمل الحزمة مستخدمين الدوال library أو require

```
> library(foreign)
> library(xlsx)
> library(dplyr)
etc ...
```

لمعرفة الإصدار و الحزم المثبتة نستخدم الدالة **sessionInfo()**

```
> sessionInfo()
R version 3.2.0 (2015-04-16)
Platform: i386-w64-mingw32/i386 (32-bit)
Running under: Windows 8 x64 (build 9200)
locale:
[1] LC_COLLATE=English_United States.1252
[2] LC_CTYPE=English_United States.1252
[3] LC_MONETARY=English_United States.1252
[4] LC_NUMERIC=C
[5] LC_TIME=English_United States.1252
attached base packages:
[1] stats     graphics  grDevices utils
datasets  methods   base
loaded via a namespace (and not attached):
[1] shiny_0.11.1.9004 R6_2.0.1
htmltools_0.2.6   Rcpp_0.11.5
[5] digest_0.6.8      xtable_1.7-4
httpuv_1.3.2      mime_0.3
```

برمجة R

يمكن إدخال أوامر و دوال R بإستخدام خط الأوامر Command Line (CL) او عن طريق نص برمجي script و الذي يمكن إجرائه داخل جلسة بإستخدام الدالة source

الأوامر يمكن الفصل بينها بـ ; او بإدخال خط جديد (newline or enter)

لاحظ أن R حساس لحالة الحرف case sensitive

العلامة # تعنى ان جميع السطر يحوي تعليق او ملاحظة.

ملفات المساعدة help files لدوال R بوضع علامة ? امام الدالة مثل

```
> ?require
```

للبحث في جميع ارشيف R نستخدم ?? مثل

```
> ??logistic
```

تخزن R كل من البيانات و المخرجات من تحليل البيانات (وفي الحقيقة كل شيئ) في كائنات objects

الأشياء تسند و تحفظ في كائنات بإستخدام العمال (operator) -> أو =

يمكن مشاهدة قائمة كل الكائنات في الجلسة الجارية بالدالة ls()

```
> ls()
```

```
[1] "expl_functions" "pkgs"          "r_local"
"r_path"          "r_sessions"
"withMathJaxIP"
>
```

إدخال البيانات Entering Data

تعمل R بشكل سهل مع مجموعات بيانات Datasets مخزنة على شكل ملفات نصية text files والتي غالبا ما تكون مفصولة او منتهية delimited بتصفيف tabs أو فراغات spaces مثل:

```
gender id race ses schtyp prgtype read write
math science socst
0 70 4 1 1 general 57 52 41 47 57
1 121 4 2 1 vocati 68 59 53 63 31
0 86 4 3 1 general 44 33 54 58 31
0 141 4 3 1 vocati 63 44 47 53 56
```

أو قيم مفصولة بالكومة (,) commas separated values (CSV)

```
gender,id,race,ses,schtyp,prgtype,read,write,ma
th,science,socst
0,70,4,1,1,general,57,52,41,47,57
1,121,4,2,1,vocati,68,59,53,63,61
0,86,4,3,1,general,44,33,54,58,31
```

```
0,141,4,3,1,vocati,63,44,47,53,56
```

قرائة بيانات دوال أساس R :

```
read.table

read.csv
```

تقوم بقرائة بيانات مخزنة في ملفات نصية منتهية delimited بأي شيئ (لاحظ حجة الدالة sep = option) في المثال التالي:

```
# comma separated values

> dat.csv <-
read.csv("http://www.ats.ucla.edu/stat/data/hsb
2.csv")

# tab separated values

> dat.tab <-
read.table("http://www.ats.ucla.edu/stat/data/h
sb2.txt",

  header=TRUE, sep = "\t")
```

لاحظ انه بالرغم من أننا نسترجع البيانات في الأمثلة السابقة من الإنترنت إلى أن هذه الدوال تستخدم عادة لقرائة ملفات مخزنة على الجهاز محليا.

كما يمكننا قرائة مجاميع بيانات من حزم إحصائية اخرى مثل SAS و SPSS بإستخدام الدوال الموجودة في الحزمة foreign مثل:

```
> require(foreign)

# SPSS files

> dat.spss <-
read.spss("http://www.ats.ucla.edu/stat/data/hs
b2.sav",

  to.data.frame=TRUE)

# Stata files

> dat.dta <-
read.dta("http://www.ats.ucla.edu/stat/data/hsb
2.dta")
```

قرائة ملفات Excel

الكثير من مجاميع البيانات تخزن غالبا كصفحات نشر إكسل Excel Spreadsheets سوف نستخدم الحزمة xlsx و java لتنزيل مجموعة بيانات من إكسل

```
# these two steps only needed to read excel files

# from the internet

> f <- tempfile("hsb2", fileext=".xls")

>
download.file("http://www.ats.ucla.edu/stat/data/hs
b2.xls", f, mode="wb")

> dat.xls <- read.xlsx(f, sheetIndex=1)
```

مشاهدة البيانات

يمكن مشاهدة البيانات بعدة طرق مثل مشاهدة عدة اسطر من رأس البيانت أو من ذيلها

```
> head(dat.xls)
```

	id	female	race	ses	schtyp	prog	read	write	math	science	socst
1	70	0	4	1	1	1	57	52	41	47	57
2	121	1	4	2	1	3	68	59	53	63	61
3	86	0	4	3	1	1	44	33	54	58	31
4	141	0	4	3	1	3	63	44	47	53	56
5	172	0	4	2	1	2	47	52	57	53	61
6	113	0	4	2	1	2	44	52	51	63	61

```
> tail(dat.xls)
```

	id	female	race	ses	schtyp	prog	read	write	math	science	socst
195	179	1	4	2	2	2	47	65	60	50	56
196	31	1	2	2	2	1	55	59	52	42	56
197	145	1	4	2	1	3	42	46	38	36	46
198	187	1	4	2	2	1	57	41	57	55	52
199	118	1	4	2	1	1	55	62	58	58	61
200	137	1	4	3	1	2	63	65	65	53	61

معرفة أسماء المتغيرات

```
> colnames(dat.xls)
 [1] "id"      "female"  "race"    "ses"
"schtyp"  "prog"    "read"    "write"   "math"
"science"
[11] "socst"
>
```

الدالة view

```
> view(dat.xls)
```

Data: dat.xls

	id	female	race	ses	schtyp	prog	read	write	math	science	socst
1	70	0	4	1	1	1	57	52	41	47	57
2	121	1	4	2	1	3	68	59	53	63	61
3	86	0	4	3	1	1	44	33	54	58	31
4	141	0	4	3	1	3	63	44	47	53	56
5	172	0	4	2	1	2	47	52	57	53	61
6	113	0	4	2	1	2	44	52	51	63	61
7	50	0	3	2	1	1	50	59	42	53	61
8	11	0	1	2	1	2	34	46	45	39	36
9	84	0	4	2	1	1	63	57	54	58	51
10	48	0	3	2	1	2	57	55	52	50	51
11	75	0	4	2	1	3	60	46	51	53	61
12	60	0	4	2	1	2	57	65	51	63	61
13	95	0	4	3	1	2	73	60	71	61	71
14	104	0	4	3	1	2	54	63	57	55	46
15	38	0	3	1	1	2	45	57	50	31	56
16	115	0	4	1	1	1	42	49	43	50	56
17	76	0	4	3	1	2	47	52	51	50	56
18	195	0	4	2	2	1	57	57	60	58	56
19	114	0	4	3	1	2	68	65	62	55	61

اطر البيانات Data Frames

بعد قرائة مجاميع البيانات في R يتم غالبا تخزينها كإطار بيانات data frame والتي لها شكل مصفوفي. ترتب المشاهدات كصفوف و المتغيرات العددية او التصنيفية (وصفية) كأعمدة.

الخلايا أو السطور أو الأعمدة المنفردة في إطار بيانات يمكن الوصول الية بعدة طرق من التأشير indexing ومن أشهرها إستخداما

```
object[row,column]
```

مثل خلية واحدة:

```
> dat.xls[2,3]
[1] 4
>
```

بإغفال قيمة السطر يعطي كل الأسطر فمثلا أسطر العمود 3

```
> dat.xls[,3]
  [1] 4 4 4 4 4 4 3 1 4 3 4 4 4 4 3 4 4 4 4 4 4 4 3
1 1 3 4 4 4 2 4 4 4 4 4 4 4 4 1 4 4 4 4 3 4 4 3 4 4
1 2
 [52] 4 1 4 4 1 4 1 4 1 4 4 4 4 4 4 4 4 4 1 4 4 4 4
4 1 4 4 4 1 4 4 4 1 4 4 4 4 4 4 2 4 4 1 4 4 4 4 1 4
4 4
```

```
[103] 3 4 4 4 4 4 3 4 4 1 4 4 1 4 4 4 4 3 1 4 4 4 3
4 4 2 4 3 4 2 4 4 4 4 4 3 1 3 1 4 4 1 4 4 4 4 1 3 3
4 4

[154] 1 4 4 4 4 4 3 4 4 4 4 4 4 4 4 4 4 4 1 3 2 3 4
4 4 4 4 4 4 4 4 2 2 4 2 4 3 4 4 4 2 4 2 4 4 4 4

>
```

إغفال العمود يعطي كل الأعمدة فمثلا للحالة 2

```
> dat.xls[2,]

   id female race ses schtyp prog read write math
science socst

2 121      1    4   2      1    3   68    59   53
63    61

>
```

لمجال range مثلا السطور 2 و 3 و الأعمدة 2 و 3

```
> dat.xls[2:3,2:3]

  female race

2      1    4

3      0    4

>
```

كما يمكننا الوصول للمتغيرات مباشرة بإستخدام أسمائهم إما بإستخدام

```
object[ , "variable"]
```

أو

```
object$variable
```

فمثلا للحصول على الأسطر الـ 10 من المتغير female نستخدم

```
> dat.xls[1:10,"female"]
 [1] 0 1 0 0 0 0 0 0 0 0
> > dat.xls$female[1:10]
 [1] 0 1 0 0 0 0 0 0 0 0
>
```

الدالة c

الدالة c تستخدم بشكل واسع و كثير لوضع قيم ذات نوع واحد في شكل متجه فمثلا يمكن دمج قيم غير متتابعة من الاسطر و الاعمدة من إطار بيانات مثل

```
> dat.xls[c(1,3,5),1]
[1]  70  86 172
> dat.xls[1,c("female", "prog", "socst")]
  female prog socst
1      0    1    57
>
```

أسماء المتغيرات

إذا لم يكن هناك أسماء متغيرات أو نريد تغيير الأسماء نستخدم الدالة colnames مثل

```
> colnames(dat.xls) <- c("ID", "Sex", "Ethnicity",
"SES", "SchoolType",
+   "Program", "Reading", "Writing", "Math",
"Science", "SocialStudies")
> dat.xls[1,]
  ID Sex Ethnicity SES SchoolType Program Reading
Writing Math Science SocialStudies
1 70   0         4   1           1       1      57
52   41      47            57
>
```

لتغيير إسم متغير واحد

```
> colnames(dat.xls)[1] <- "ID2"
>
```

حفظ البيانات

يمكننا حفظ البيانات في عديد من الصيغ منها النصي text و إكسل xlsx وفي شكل مناسب للحزم الإحصائية مثل SAS و SPSS و STATA الخ بإستخدام الدالة write.xlsx و write.dta من حزم foreign و xlsx

```
> write.csv(dat.csv, file =
path/to/save/filename.csv")

> write.table(dat.csv, file =
"path/to/save/filename.txt", sep = "\t", na=".")

> write.dta(dat.csv, file =
"path/to/save/filename.dta")

> write.xlsx(dat.csv, file =
"path/to/save/filename.xlsx", sheetName="hsb2")
```

لحفظ البيانات في شكل خاص بـ R (ملفات غير نصية والتي تخزن عدة أشكال من البيانات

```
> save(dat.csv, dat.dta, dat.spss, dat.txt, file =
"path/to/save/filename.RData")
```

إستكشاف البيانات Exploring Data

سوف نقوم الآن بقرائة بعض البيانات وتخزينها في كائن object نسمه d (سوف نستخدم أسماء قصيرة لسهولة كتابتها)

سوف نستكشف ونتعرف على بيانات تحوي المتغيرات (عدد المدرسة و إختبار و معلومات سكنية Demography لعدد 200 طالب و طالبة)

ملاحظة: هذه البيانات من الولايات المتحدة ونتحصل عليها من جامعة كاليفورنيا.

```
 d <-
read.csv("http://www.ats.ucla.edu/stat/data/hsb2.cs
v")
```

نستخدم الدالة dim والتي تعطينا عدد المشاهدات (الأسطر) و عدد المتغيرات (الأعمدة)

نستخدم str والتي تعطينا تركيب structure البيانات d مشتملة نوع class (type) كل المتغيرات

```
> dim(d)
[1] 200  11
> str(d)
'data.frame':   200 obs. of  11 variables:
 $ id     : int  70 121 86 141 172 113 50 11 84 48
...
 $ female : int  0 1 0 0 0 0 0 0 0 0 ...
 $ race   : int  4 4 4 4 4 4 3 1 4 3 ...
 $ ses    : int  1 2 3 3 2 2 2 2 2 2 ...
 $ schtyp : int  1 1 1 1 1 1 1 1 1 1 ...
 $ prog   : int  1 3 1 3 2 2 1 2 1 2 ...
 $ read   : int  57 68 44 63 47 44 50 34 63 57 ...
 $ write  : int  52 59 33 44 52 52 59 46 57 55 ...
 $ math   : int  41 53 54 47 57 51 42 45 54 52 ...
```

```
 $ science: int  47 63 58 53 53 63 53 39 58 50 ...
 $ socst  : int  57 61 31 56 61 61 61 36 51 51 ...
>
```

الدالة summary وهي دالة عامة generic function تلخص عدة انواع من كائنات R بما فيها مجاميع البيانات

```
> summary(d)
       id              female          race      
 Min.   :  1.00   Min.   :0.000   Min.   :1.00  
 1st Qu.: 50.75   1st Qu.:0.000   1st Qu.:3.00  
 Median :100.50   Median :1.000   Median :4.00  
 Mean   :100.50   Mean   :0.545   Mean   :3.43  
 3rd Qu.:150.25   3rd Qu.:1.000   3rd Qu.:4.00  
 Max.   :200.00   Max.   :1.000   Max.   :4.00  

      ses           schtyp          prog      
 Min.   :1.000   Min.   :1.00   Min.   :1.000  
 1st Qu.:2.000   1st Qu.:1.00   1st Qu.:2.000  
 Median :2.000   Median :1.00   Median :2.000  
 Mean   :2.055   Mean   :1.16   Mean   :2.025  
 3rd Qu.:3.000   3rd Qu.:1.00   3rd Qu.:2.250  
 Max.   :3.000   Max.   :2.00   Max.   :3.000  

      read           write           math      
 Min.   :28.00   Min.   :31.00   Min.   :33.00  
 1st Qu.:44.00   1st Qu.:45.75   1st Qu.:45.00  
 Median :50.00   Median :54.00   Median :52.00  
 Mean   :52.23   Mean   :52.77   Mean   :52.65  
 3rd Qu.:60.00   3rd Qu.:60.00   3rd Qu.:59.00  
 Max.   :76.00   Max.   :67.00   Max.   :75.00  
```

التلخيص الشرطي Conditional Summaries

تسمح R بدوال متداخلة nested functions حيث نتيجة دالة تمرر مباشرة كحجة argument لدالة اخرى. مثل:

```
> summary(subset(d, read >= 60))
```

هنا الدالة subset تسترجع مجموعة بيانات حيث المتغير read >= 60 (كل الطلاب الذين تحصلوا عل درجة 60 أو أكثر في القرائة) ثم تقوم الدالة summary بتلخيص هذه المجموعة الجزئية.

ملاحظة: يمكن وضع المجمعة الجزئية في كائن جديد مثل

```
> dd <- subset( d , read >= 60)
> summary(dd)
```

يمكننا فصل البيانات بطرق اخرى مثلا بوضعها في مجاميع فمثلا لننظر لمتوسطات المتغيرا الخمس (5) الأخيرة لكل نوع من المتغير prog (المتغير program).

هنا سوف نسأل الدالة by لكي تطبق الدالة colMeans للمتغيرات الموجودة في الأعمدة من 7 إلى 11 ومن ثم تحسب متوسطات المتغيرات المتجمعة بواسطة المتغير prog

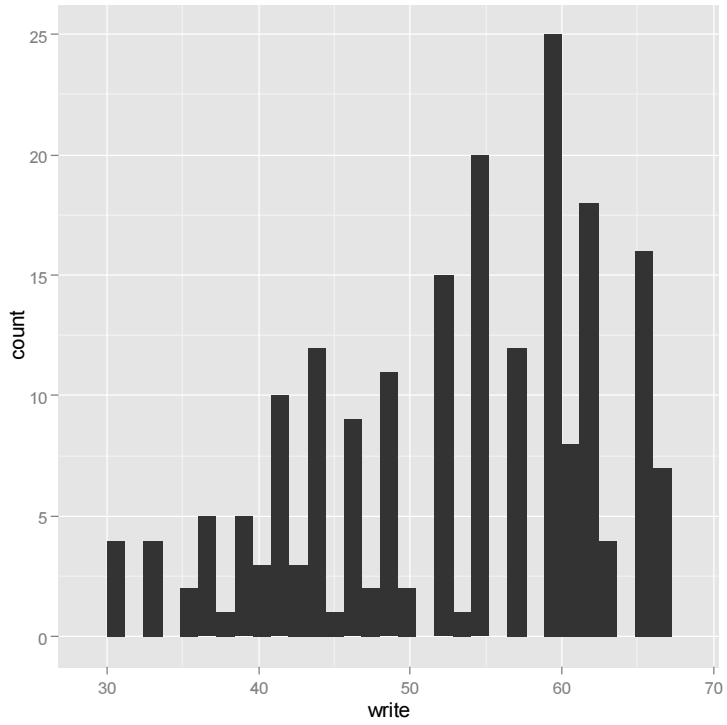
```
> by(d [ , 7:11], d$prog, colMeans)
```

```
d$prog: 1
    read    write     math  science    socst
49.75556 51.33333 50.02222 52.44444 50.60000
------------------------------------------------
d$prog: 2
    read    write     math  science    socst
56.16190 56.25714 56.73333 53.80000 56.69524
------------------------------------------------
d$prog: 3
   read   write    math science   socst
  46.20   46.76   46.42   47.22   45.02
>
```

المدرج التكراري Histogram

غالبا من السهل فحص توزيع المتغير عن طريق الرسم و المدرج التكراري عبارة عن مقدر اولي لهذا التوزيع فمثلا

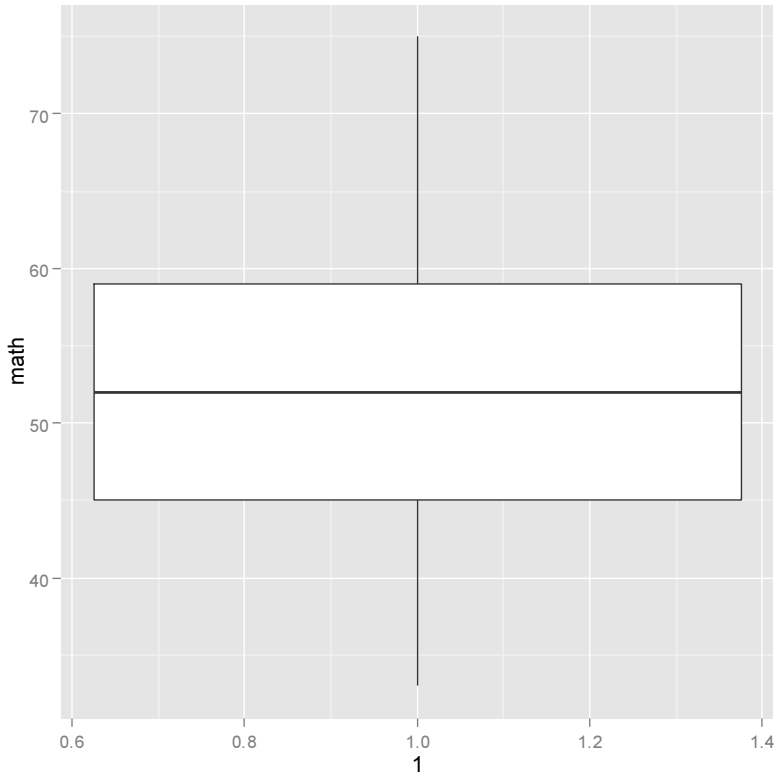
```
> library(ggplot2)
> ggplot(d , aes(x = write)) + geom_histogram()
```



رسومات الصناديق Boxplots

رسومات الصناديق تبين بشكل رسوم الوسيط والربيعين الأدنى و الأعلى و المدى

```
> ggplot(d, aes(x = 1, y = math)) + geom_boxplot()
```



البيانات الوصفية Categorical Data

نستطيع النظر إلى توزيعات المتغيرات الإسمية بإستخدام جداول التكرارات frequency tables

```
> xtabs( ~ female, data = d)

female

  0   1

 91 109
> xtabs( ~ race, data = d)

race
```

```
  1   2   3   4
 24  11  20 145
> xtabs( ~ prog, data = d)
prog
  1   2   3
 45 105  50
```

الجداول الثنائية Two-way Cross tabs

```
> xtabs( ~ ses + schtyp, data = d)
   schtyp
ses  1  2
  1 45  2
  2 76 19
  3 47 11
>
```

الجداول الثلاثية Three-way Cross tabs

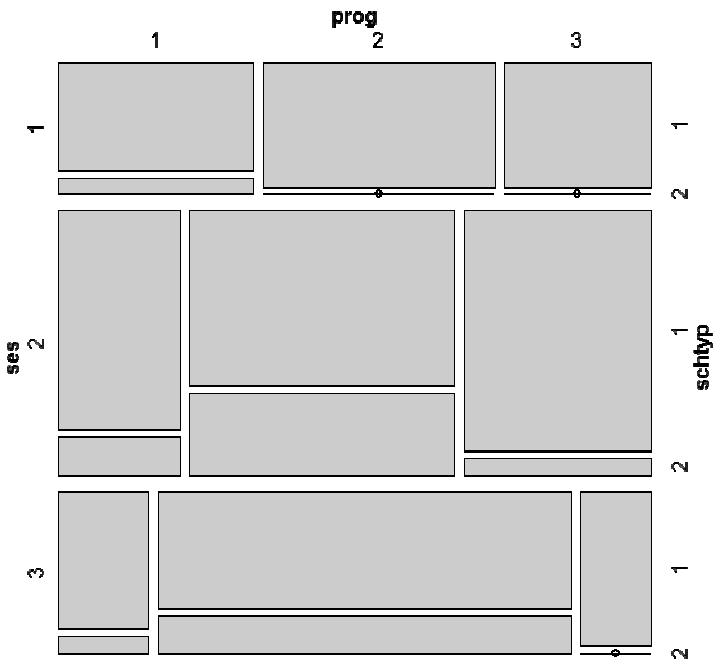
```
> (tab3 <- xtabs( ~ ses + prog + schtyp, data = d))
, , schtyp = 1
```

```
   prog
ses  1  2  3
 1  14 19 12
 2  17 30 29
 3   8 32  7
, , schtyp = 2
   prog
ses  1  2  3
  1  2  0  0
  2  3 14  2
  3  1 10  0
```

مشاهدة البيانات الوصفية Visualizing Cat Data

يمكننا تمثيل كل خلية في جدول تكراري بمساحة على مستطيل بإستخدام الدالة mosaic من الحزمة vcd

```
> library(vcd)
Loading required package: grid
> mosaic(tab3)
```


prog
1
2
3
ses
1
2
3
schtyp
1
2

الترابط Correlations

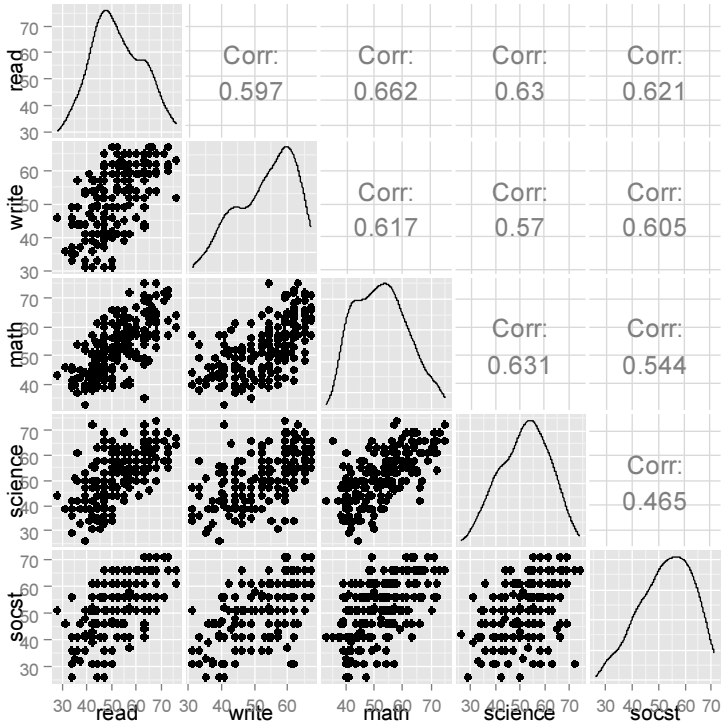
العلاقة بين متغيرين تقاس بإحصائة تسمى الترابط. حزمة R تعطي طرق كثيرة لقياس العلاقة بين متغيرين منها مصفوفات الترابط Correlation Matrices و التي تعطي ملخص عن هذه العلاقات الثنائية. سوف نستخدم الدالة cor مع حجج إفتراضية default arguments في حالة عدم وجود قيم مفقودة missing values و إلا نستخدم الحجة use كالتالي: نوجد الترابطات الثنائية بين المتغيرات في الأعمدة 7 وحتى 11

```
> cor(d[ , 7:11])
```

	read	write	math	science	socst
read	1.0000000	0.5967765	0.6622801	0.6301579	0.6214843
write	0.5967765	1.0000000	0.6174493	0.5704416	0.6047932
math	0.6622801	0.6174493	1.0000000	0.6307332	0.5444803
science	0.6301579	0.5704416	0.6307332	1.0000000	0.4651060
socst	0.6214843	0.6047932	0.5444803	0.4651060	1.0000000

او

```
> library(GGally)
> ggpairs(d[ , 7:11])
```



تعديل البيانات Modifying Data

سوف نستعرض الآن تعديل و إعادة ترتيب للبيانات. سوف نقرأ مجموعة بيانات جديدة و نخزنها في الكائن أو الشيئ d

```
> d <-
read.csv("http://www.ats.ucla.edu/stat/data/hsb2.csv")
```

نستطيع أن نرتب البيانات مستخدمين الدالة arrange من مكتبة dplyr .

سوف نرتب البيانات حسب المتغير female ثم بالمتغير math ثم ننظر إلى مجموعة البيانات الناتجة.

```
> library(dplyr)

>

> d <- arrange(d, female, math)

> head(d)

   id female race ses schtyp prog read write math science
socst

1 167      0    4   2      1    1   63    49   35      66    41

2 128      0    4   3      1    2   39    33   38      47    41

3  49      0    3   3      1    3   50    40   39      49    47

4  22      0    1   2      1    3   42    39   39      56    46

5 134      0    4   1      1    1   44    44   39      34    46

6 117      0    4   3      1    3   34    49   39      42    56
```

تحليل البيانات Analyzing Data

أولا: تحليل البيانات الوصفية Analyzing Cat Data

يستخدم إختبار مربع كاي chisq.test كأحد الطرق لتحليل البيانات الوصفية كالتالي:

إختبار مربع كاي chisq.test

اولا ننشئ جدول تكراري

```
> (tab <- xtabs(~ ses + schtyp, data = d))

> chisq.test(tab)

        Pearson's Chi-squared test

data:  tab

X-squared = 6.3342, df = 2, p-value = 0.04213
```

إختبارات تي t-tests

يستخدم إختبار تي t.test لمقارنة متوسطين. هنا سوف نستعرض إختبار تي لعينة واحدة one sample t-test لمقارنة متوسط write مع القيمة 50 و إختبار تي لزوج من العينات paired samples t-test لمقارنة متوسطات المتغيرين write و read

```
> t.test(d$write, mu = 50)

        One Sample t-test

data:  d$write
```

```
t = 4.1403, df = 199, p-value = 5.121e-05

alternative hypothesis: true mean is not equal to
50

95 percent confidence interval:

 51.45332 54.09668

sample estimates:

mean of x

   52.775
> with(d, t.test(write, read, paired = TRUE))

        Paired t-test

data:  write and read

t = 0.86731, df = 199, p-value = 0.3868

alternative hypothesis: true difference in means is
not equal to 0

95 percent confidence interval:

 -0.6941424  1.7841424

sample estimates:

mean of the differences

                 0.545
```

أيضا يمكننا مقارنة متوسطات المتغير write بين الذكور و الإناث بإستخدام إختبار تي للعينات المستقلة independent sample t-test مفترضين تساوي التباين equal variance أو عدم تساوي التباين

```
> t.test(write ~ female, data = d, var.equal=TRUE)

        Two Sample t-test

data:  write by female

t = -3.7341, df = 198, p-value = 0.0002463

alternative hypothesis: true difference in means is
not equal to 0

95 percent confidence interval:

 -7.441835 -2.298059

sample estimates:

mean in group 0 mean in group 1

       50.12088        54.99083

> t.test(write ~ female, data = d)

        Welch Two Sample t-test

data:  write by female

t = -3.6564, df = 169.71, p-value = 0.0003409

alternative hypothesis: true difference in means is
not equal to 0

95 percent confidence interval:
```

```
 -7.499159 -2.240734
sample estimates:
mean in group 0 mean in group 1
       50.12088        54.99083
```

تحليل التباين و الإنحدار ANOVA and Regression

لتحليل التباين و الإنحدار نستخدم الدالة lm

```
> m <- lm(write ~ prog * female, data = d)
> anova(m)
Analysis of Variance Table
Response: write
             Df  Sum Sq Mean Sq F value    Pr(>F)
prog          1   586.4  586.42  7.2039 0.0078971
**
female        1  1182.9 1182.90 14.5312 0.0001845
***
prog:female   1   154.3  154.32  1.8958 0.1701223
Residuals   196 15955.2   81.40
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05
‘.’ 0.1 ‘ ’ 1
> summary(m)

Call:

lm(formula = write ~ prog * female, data = d)

Residuals:

    Min      1Q  Median      3Q     Max

-23.079  -6.835   1.794   6.973  16.794

Coefficients:

            Estimate Std. Error t value Pr(>|t|)

(Intercept)  57.9520     2.9096  19.917   <2e-16
***

prog         -3.8730     1.3609  -2.846   0.0049 **

female       -0.2943     3.9730  -0.074   0.9410

prog:female   2.5577     1.8576   1.377   0.1701

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05
‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.022 on 196 degrees of
freedom

Multiple R-squared:  0.1076,    Adjusted R-squared:
0.09393
```

```
F-statistic: 7.877 on 3 and 196 DF,  p-value:
5.469e-05
```

الإنحدار

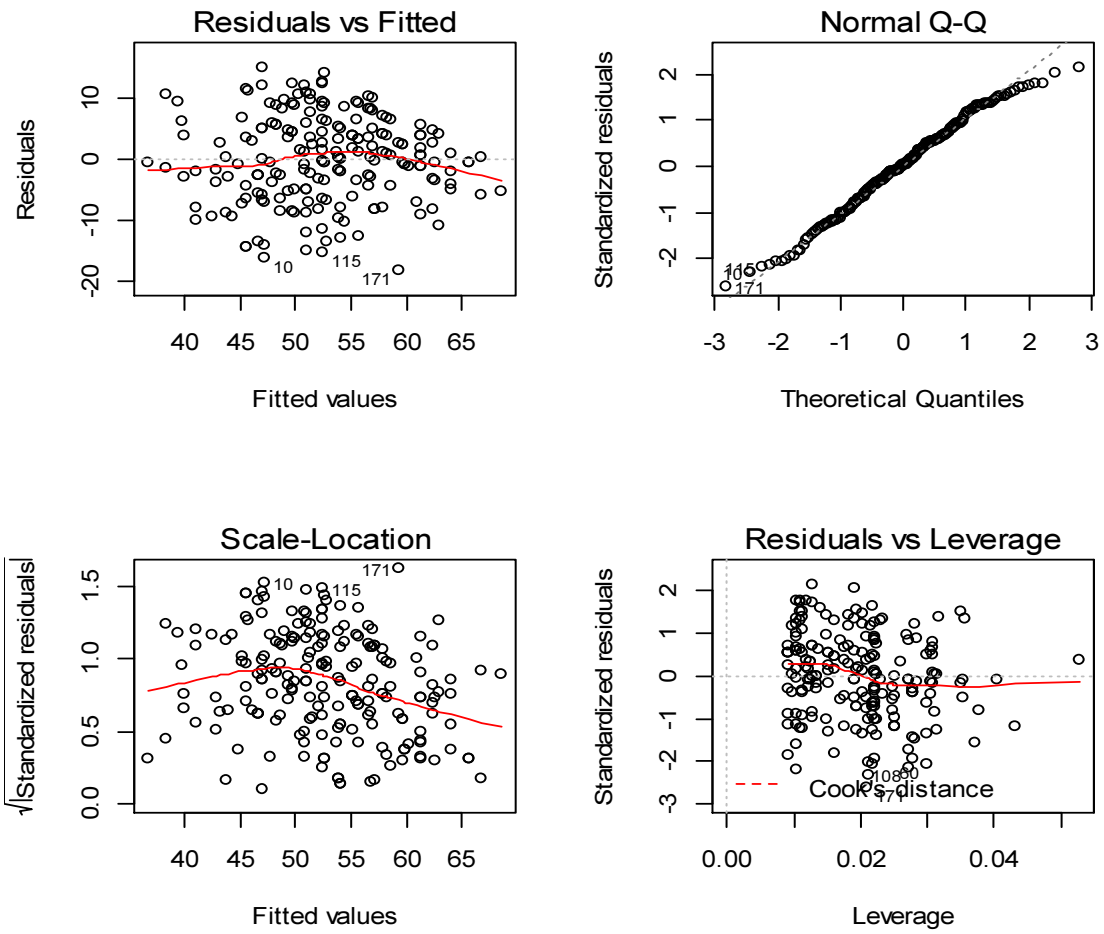
```
> m2 <- lm(write ~ prog + female + read, data = d)
> m2

Call:
lm(formula = write ~ prog + female + read, data =
d)

Coefficients:
(Intercept)         prog       female         read
    23.7154      -1.3933       5.4808       0.5532
> summary(m2)

Call:
lm(formula = write ~ prog + female + read, data =
d)

Residuals:
     Min       1Q   Median       3Q      Max
-18.3357  -4.9757   0.1172   4.9652  15.0705

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) 23.71544    3.26268   7.269 8.43e-12
***
prog        -1.39331    0.73422  -1.898   0.0592 .
female       5.48080    1.00764   5.439 1.58e-07
***
read         0.55321    0.04951  11.173  < 2e-16
***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05
‘.’ 0.1 ‘ ’ 1
Residual standard error: 7.086 on 196 degrees of
freedom
Multiple R-squared:  0.4495,    Adjusted R-squared:
0.4411
F-statistic: 53.35 on 3 and 196 DF,  p-value: <
2.2e-16
```

تشخيص الإنحدار Regression Diagnostics

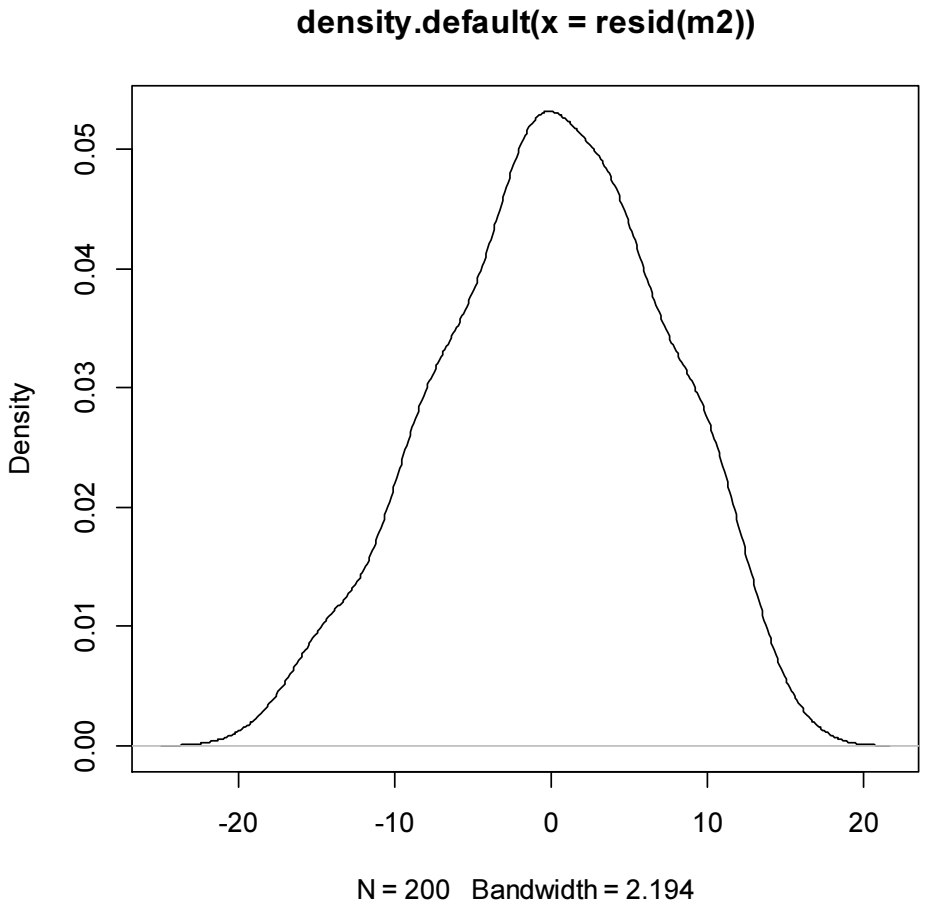
```
> par(mfrow = c(2, 2))
```

```
 > plot(m2)
```



نفحص طبيعية البواقي

```
> plot(density(resid(m2)))
```



(1) http://www.ats.ucla.edu/stat/r/seminars/intro.htm

(2) http://www.stat.ucla.edu/~vlew/stat130a/datasets/

(3) http://www.statmethods.net/

مقدمة للتحليل الإحصائي بواسطة Python

يعتبر Python من لغات البرمجة العالية ذات الأغراض العامة General-purpose High-level Programming Language ويتميز بوضوح جيد وقرائة عالية ومقدرة كبيرة على التعبير عن الخوارزمات بأسطر بسيطة من الترميز بعكس اللغات الاخرى مثل C++ أو Java ويحوي مكتبة شاملة من البرامج.

مترجمة Python (Python Interpreters) متواجدة لجميع أنظمة التشغيل مثل Windows و MacSystems و Unix الخ.

Python من البرامج المفتوحة Open-source وهو مجاني لجميع المستخدمين.

يقوم بتطوير وصيانة Python فريق من المطورين الأكادمين و الباحثين من جميع أنحاء جامعات العالم.

تحميل و تثبيت Python

الموقع الرسمي لـ Python هو

https://www.python.org/

وتوجد واجهات إستخدام عدة لـ Python منها Anaconda وهو برنامج مجاني مفتوح لمعالجة البيانات الكبيرة Larg-scale Data processing و التحليل التوقعي Predictive Analysis و الحسابات العلمية Scientific Computing .

سوف نستخدم Anaconda في محاضراتنا ويمكن تحميل البرنامج من

https://repo.continuum.io/archive/

وآخر إصدار هو:

لنظام win32

https://repo.continuum.io/archive/Anaconda3-2.2.0-Windows-x86.exe

لنظام win64

https://repo.continuum.io/archive/Anaconda3-2.2.0-Windows-x86_64.exe

يمكنك أيضا تحميل البرنامج الأساسي من:

https://www.python.org/downloads/

ومن ثم إضافة المكتبات Modules الاخرى حسب الطلب.

مكتبات Python العلمية

سوف نستخدم المكتبات التالية:

NumPy وهو مكتبة تضاف لإعطاء دعم كبير وسريع لمصفوفات ومتجهات متعددة الأبعاد.

SciPy وهو مكتبة للبرامج العلمية و العددية.

Matplotlib وهو مكتبة للرسومات و الدوال الرياضية.

Pandas وهو مكتبة لتحليل البيانات.

لمعرفة تفاصيل أكثر أنظر

http://en.wikipedia.org/wiki/Python_(programming_language)#Implementations

تراكيب البيانات الأساسية في Python

أي عملية على البيانات في Python تعتمد على تركيب تلك البيانات فمثلا لو حاولنا تطبيق دالة على بيانات غير معروف نوعها لتلك الدالة سوف تصدر رسالة خطأ (سوف تصادف بعضها أثناء الإستخدام)

1) Lists (قوائم) متتبعات مرتبة يمكن تغيير عناصرها mutable وتعرف بـ [ ]

2) Tuples (مرتبات) متتابعات مرتبة لايمكن تغيير عناصرها immutable وتعرف بـ ( ) أو بدون أقواس

3) Dictionaries (قاموس) وتعرف تطبيق mapping من مفاتيح keys إلى قيم values وتعرف بـ { }

3) Strings (نصوص) وهي متتابعة من الرموز وتعرف بـ " " أو ' '

أمثلة:

```
>>> s = " a string"
>>> t = ' a string '
>>> a = ['ahmad','red',100,1872]# list of 4
                                  # elements
>>> b = [ ]                  # an empty list
>>> c = 1, 2, 3          # a tuble of 3 elements
>>> d = (1, 2, 3)      # a tuble of 3 elements
>>> e = ( )               # an empty tuble
>>> f = 'ahmad',  # a tuble with on element
                  #(comma is required)
>>> g = ('ahmad',) # a tuble with on element
                   #(comma is required)

>>> h = {'x':1, 'y':2, 'z':2} # a dictionary
                              #  mapping strings to
                              # integers
>>> i = {1:'x', 2:'y', 3:'z'} # a dictionary
                              #  mapping integers
                              # to strings
>>> j = {}            # a empty dictionary
```

التأشير Indexing

يمكن تأشير القوائم و المرتبات و النصوص و أول مؤشر هو 0

```
>>> s = 'a string'
>>> s[0]
'a'
>>> s[3]
't'
>>> t = 99,100,101
>>> t[2]
101
>>> v = [33,44,'ahmad']
>>> v[2]
'ahmad'
>>> v[2][1:3]
'hma'
>>> v = [10,11,12,13,14]
>>> v[-2]
13
>>> v
[10,11,12,13,14]
>>> v[3] = 99
>>> v
[10,11,12,99,14]
```

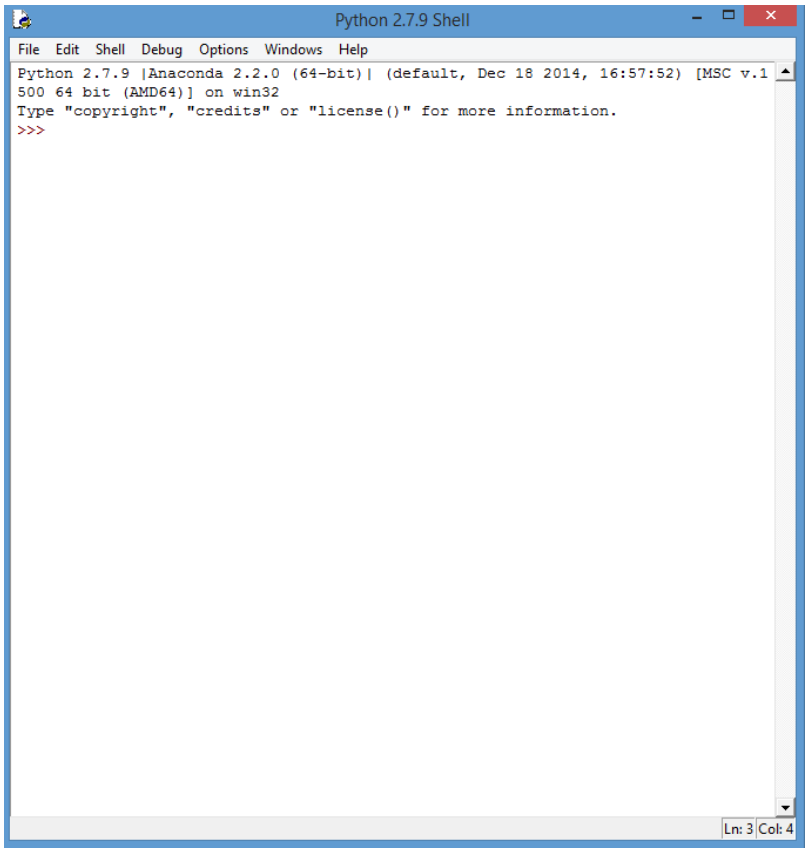
التشريح Slicing

التشريح يعطي متتابعة جزئية subsequence

```
>>> v =[1,2,3,4,5,6,7,8,9,10]
>>> v[3:4]
[4]
>>> v[3:9:2]
[4,6,8]
>>> v[-7:-1]
[4,5,6,7,8,9]
>>> v[3:5] = [99,100]
>>> v
[1,2,3,99,100,6,7,8,9,10]
```

Python كآلة حاسبة

عند تشغيل Python تظهر النافذة



أدخل التالي عند محث الأوامر >>>

```
>>> 2 + 2
4
>>> 50 - 5*6
20
>>> (50 - 5*6)/4
5
```

```
>>> 8 /5
1
>>> 8 / 5.0
1.6
```

إدخال بيات و إجراء بعض العمليات

```
>>> import numpy as np
>>> import scipy as sc
>>> x = np.array([474.688, 506.445, 524.081,
530.672, 530.869, 566.984, 582.311, 582.940,
603.574, 792.358])
>>> x
array([ 474.688,  506.445,  524.081,  530.672,
530.869,  566.984, 582.311,  582.94 ,  603.574,
792.358])
>>> from scipy import stats
>>> sc.stats.describe(x)
DescribeResult(nobs=10L,
minmax=(474.68799999999999, 792.35799999999995),
mean=569.49220000000003,
variance=7689.5458092888857,
skewness=1.7024129743728933,
kurtosis=2.3738129817185847)
>>>
```

قرائة ملفات القيم المفصولة بفاصلة Reading csv files

```
>>> import pandas as pn
>>> from pandas import *
>>> iris = read_csv("C:/Users/ibin/Desktop/iris.csv")
>>> print iris
    sepal_length  sepal_width  petal_length  petal_width  name
0     5.1         3.5          1.4           0.2     setosa
1     4.9         3.0          1.4           0.2     setosa
2     4.7         3.2          1.3           0.2     setosa
...
>>> print iris["name"]
0    setosa
1    setosa
2    setosa
...

>>> print iris["name"][12]
setosa
>>> print iris["sepal_length"][12]
4.8
>>> print iris[iris["sepal_length"]> 5.0]
sepal_length  sepal_width  petal_length petal_width  name
0      5.1          3.5           1.4          0.2      setosa
5      5.4          3.9           1.7          0.4      setosa
10     5.4          3.7           1.5          0.2      setosa
14     5.8          4.0           1.2          0.2      setosa
15     5.7          4.4           1.5          0.4      setosa
```

```
. . .
>>> print iris[iris["sepal_length"]< 5.0]
  sepal_length sepal_width petal_length petal_width  name
1     4.9          3.0          1.4          0.2    setosa
2     4.7          3.2          1.3          0.2    setosa
3     4.6          3.1          1.5          0.2    setosa
6     4.6          3.4          1.4          0.3    setosa
. . .
>>> print iris["sepal_length"].mean()
5.84333333333
>>> import scipy
>>> from scipy import stats
>>> print scipy.stats.sem(iris["sepal_length"])
0.0676113162276

>>> print iris["sepal_length"].std()
0.828066127978
>>> print iris["sepal_length"].var()
0.685693512304
>>> print iris["sepal_length"].median()
5.8
>>> print iris["sepal_length"].mode()
0    5
>>> print iris["sepal_length"].sem()
0.0676113162276
>>> f_val, p_val =
stats.f_oneway(iris.sepal_length, iris.sepal_width)
>>> f_val
1335.7678308241761
```

```
>>> p_val
3.9878381148481482e-112
>>> x = iris.sepal_length
>>> y = iris.sepal_width
>>> z = iris.petal_length
>>> w = iris.petal_width
>>> f_val, p_val = stats.f_oneway(x, y, z, w)
>>> f_val
483.57128302425951
>>> p_val
3.4996987081941695e-159
>>>
```

مثال آخر

```
>>> import pandas
>>> data =
pandas.read_csv('C:/Users/ibin/Desktop/brain_size.c
sv', sep=';', na_values=".")
>>> print data
Unnamed: 0  Gender  FSIQ  VIQ  PIQ  Weight  Height  MRI_Count\t
0        1  Female   133  132  124     118    64.5       816932
1        2    Male   140  150  124     NaN    72.5      1001121
2        3    Male   139  123  150     143    73.3      1038437
...
>>> print data ['Gender']
0     Female
1       Male
...
38      Male
```

```
39      Male
>>> gender_data = data.groupby('Gender')
>>> print gender_data.mean()
Unnamed: 0   FSIQ    VIQ    PIQ     Weight    Height MRI_Count\t
Gender
Female  19.65 111.9 109.45 110.45 137.200000 65.765000    862654.6
Male    21.35 115.0 115.25 111.60 166.444444 71.431579    954855.4
>>> for name, value in gender_data['VIQ']:
    print name, np.mean(value)
Female 109.45
Male 115.25
>>> data[data['Gender'] == 'Female']['VIQ'].mean()
109.45
>>> from pandas.tools import plotting
>>> plotting.scatter_matrix(data[['PIQ', 'VIQ',
'FSIQ']])
>>> from scipy import stats
>>> stats.ttest_1samp(data['VIQ'], 0)
>>> female_viq = data[data['Gender'] ==
'Female']['VIQ']
>>> male_viq = data[data['Gender'] ==
'Male']['VIQ']
>>> stats.ttest_ind(female_viq, male_viq)
>>> stats.ttest_ind(data['FSIQ'], data['PIQ'])
>>> stats.ttest_rel(data['FSIQ'], data['PIQ'])
>>> stats.ttest_1samp(data['FSIQ'] - data['PIQ'],
0)
>>> from scipy import stats
```

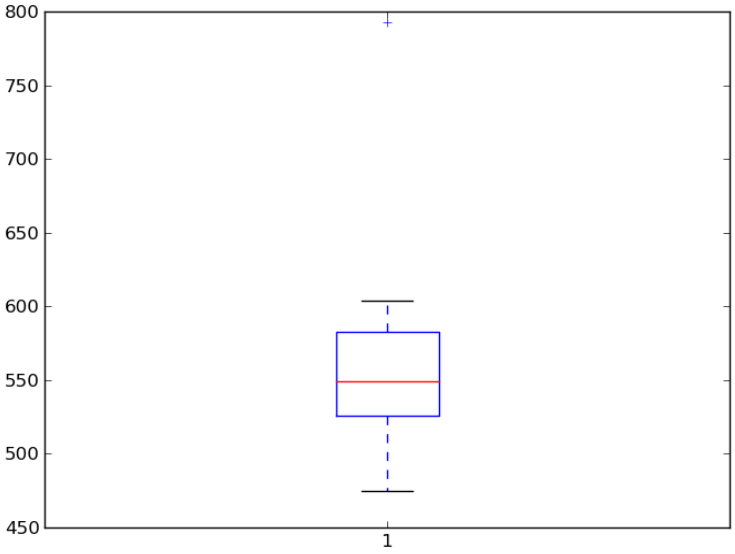
```
>>> import numpy as np
>>> x = [10, 11, 13, 9, 7, 12, 12, 9, 10, 12]
>>> y = [13, 21, 5, 10, 8, 14, 10, 12, 7, 15]
slope, intercept, r_value, p_value, std_err =
stats.linregress(x,y)
# To get coefficient of determination (r_squared)
>>> print "r-squared:", r_value**2
r-squared: 0.15286643777
```

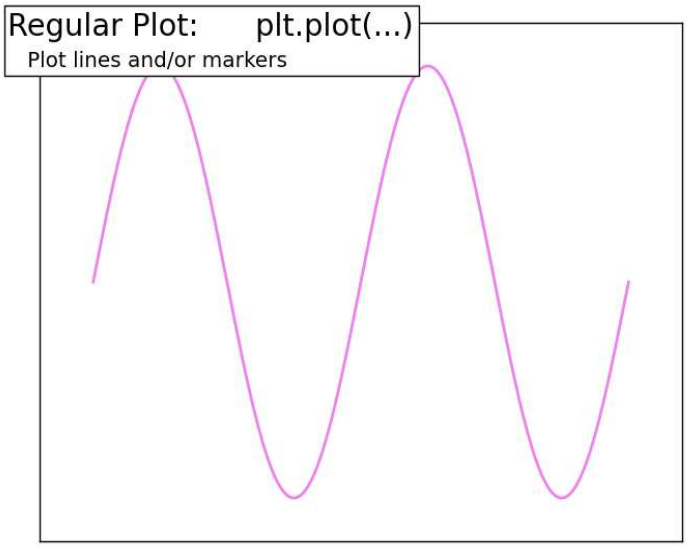
المكتبة numpy:

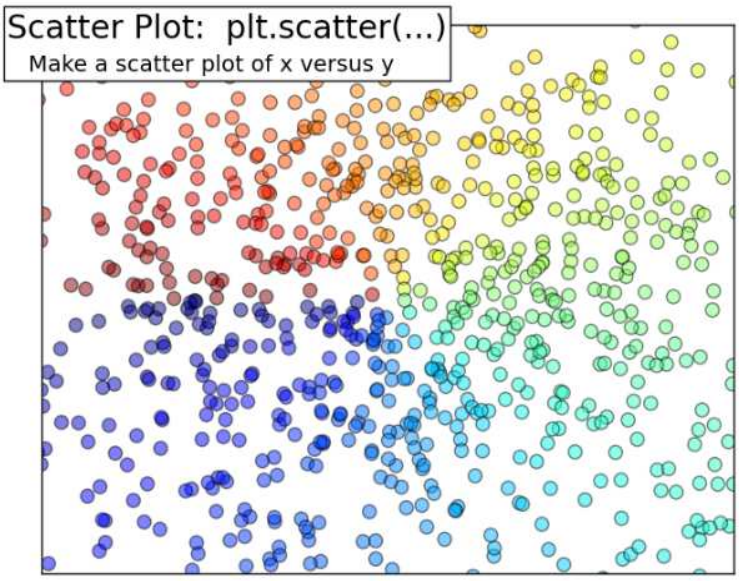
```
>>> import numpy as np
>>> x = np.array([474.688, 506.445, 524.081,
530.672, 530.869, 566.984, 582.311, 582.940,
603.574, 792.358])
>>> x[0]    # First element
474.68799999999999
>>> x[:3]   # Up to the third element
array([ 474.688,  506.445,  524.081])
>>> x[-1]   # Last element
792.35799999999995
>>> np.max(x)
792.35799999999995
>>> np.sum(x[:3])
1505.2139999999999
>>> np.mean(x)
569.49219999999991
>>> np.log(x)
array([ 6.16265775,  6.22741573,  6.26164625,
6.27414413,  6.27451529, 6.34033108,  6.36700467,
6.36808427,  6.40286865,  6.67501331])
```

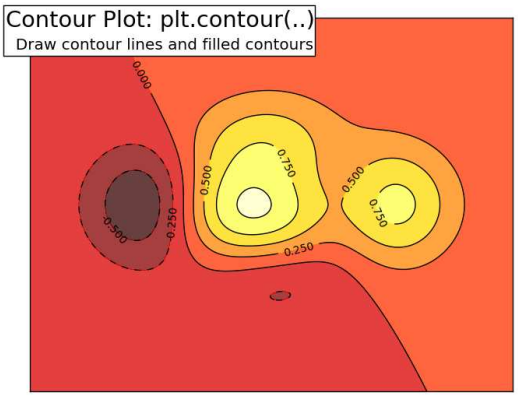
الرسومات البسيطة مع pylab

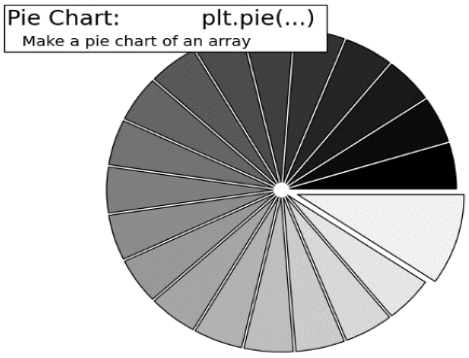
```
>>> from matplotlib import pyplot as plt
>>> plt.boxplot(x)
```

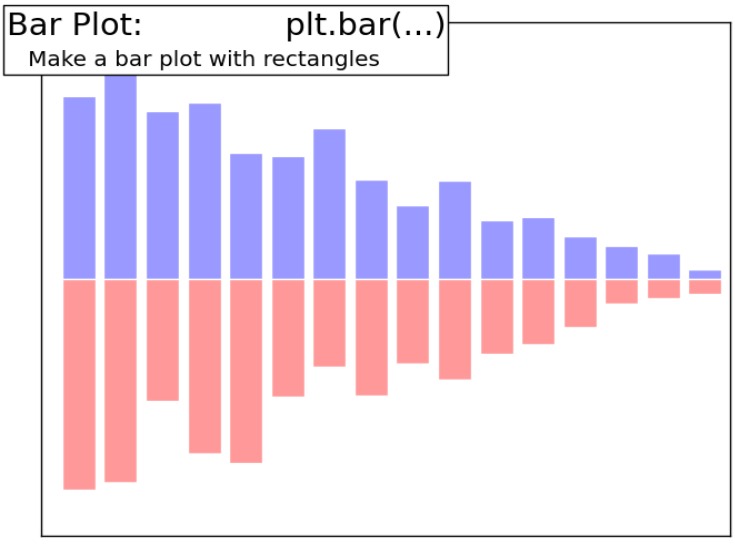





Scatter Plot: plt.scatter(...)
Make a scatter plot of x versus y


Contour Plot: plt.contour(..)
Draw contour lines and filled contours
0.000
0.500
0.750
0.500
0.750
-0.500
0.250
0.250


Pie Chart: plt.pie(...)
Make a pie chart of an array

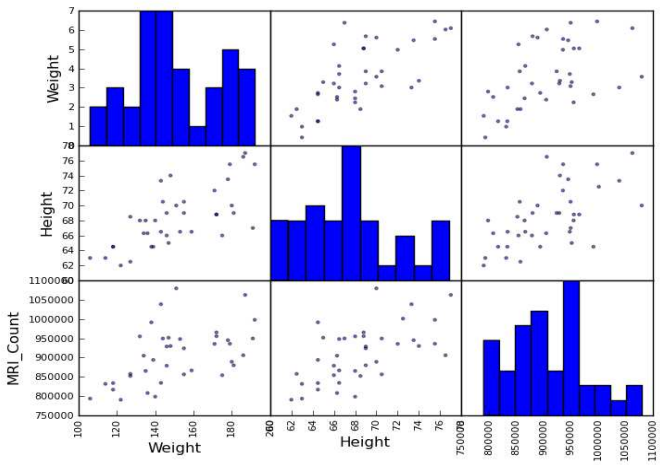

Bar Plot: plt.bar(...)
Make a bar plot with rectangles

```
>>> import pandas
>>> data =
pandas.read_csv('examples/brain_size.csv', sep=';',
na_values=".")
>>> print data
    Unnamed: 0  Gender  FSIQ  VIQ  PIQ  Weight  Height  MRI_Count
0            1  Female   133  132  124     118    64.5     816932
1            2    Male   140  150  124     NaN    72.5    1001121
2            3    Male   139  123  150     143    73.3    1038437
. . .
>>> print data['Gender']
0     Female
1       Male
2       Male
3       Male
4     Female
...
>>> gender_data = data.groupby('Gender')
>>> print gender_data.mean()
Unnamed: 0   FSIQ    VIQ    PIQ     Weight    Height  MRI_Count
Gender
Female  19.65 111.9 109.45 110.45 137.200000 65.765000  862654.6
Male    21.35 115.0 115.25 111.60 166.444444 71.431579  954855.4
>>> for name, value in gender_data['VIQ']:
...     print name, np.mean(value)
Female 109.45
Male 115.25
```

```
>>> data[data['Gender'] == 'Female']['VIQ'].mean()
109.45
```

الرسومات مع pandas

```
>>> from pandas.tools import plotting
>>> plotting.scatter_matrix(data[['Weight', '
Height', 'MRI_Count']])
```



إختبار الفرضيات Hypothesis testing:

```
>>> from scipy import stats
```

إختبار t لعينة واحدة 1-sample t-test

```
>>> stats.ttest_1samp(data['VIQ'], 0)
```

إختبار t لعينتين 2-sample t-test

```
>>> female_viq = data[data['Gender'] ==
'Female']['VIQ']
>>> male_viq = data[data['Gender'] ==
'Male']['VIQ']
>>> stats.ttest_ind(female_viq, male_viq)
```

مثال آخر:

رميت عملة 100 مرة. عدد مرات ظهور وجه العملة 61 مرة. هل العملة متزنة. أي هل عدد ظهور الوجة يساوي تقريبا عدد ظهور الكتابة.

```
>>> import numpy as np
>>> import scipy.stats as st
>>> import scipy.special as sp
>>> n = 100
>>> h = 61
>>> p = .5
>>> xbar = float(h)/n
```

```
>>> xbar
0.61
>>> z = (xbar - p) * np.sqrt(n / (p*(1-p))); z
2.1999999999999997
>>> pval = 2 * (1 - st.norm.cdf(z)); pval
0.027806895026997 18
>>>
```

نلاحظ أن إحتمال الحصول على 61 وجه عند رمي عملة متزنة 100 مرة هو 0.028 وهذا يؤيد رفض الفرضية الصفرية على ان العملة متزنة (نرفض عند 0.05 مستوى معنوية)

إنحدار خطي بسيط

A simple linear regression

نولد بيانات صناعية بالمحاكاة

```
>>> x = np.linspace(-5, 5, 20)
>>> np.random.seed(1)
>>> y = -5 + 3*x + 4 *
np.random.normal(size=x.shape)
>>> data = pandas.DataFrame({'x': x, 'y': y})
>>> from statsmodels.formula.api import ols
>>> model = ols("y ~ x", data).fit()
>>> print(model.summary())
```

إنحدار متعدد

Multiple Regression

```
>>> data = pandas.read_csv('examples/iris.csv')
```

```
>>> model = ols('sepal_width ~ name +
petal_length', data).fit()
>>> print(model.summary())
```

أمثلة

```
>>> from scipy import stats
>>> from scipy.stats import norm
>>> norm.cdf(0)
>>> norm.mean(), norm.std(), norm.var()
>>> from scipy.stats import expon
>>> expon.mean(scale=3.)
>>> from scipy.stats import uniform
>>> uniform.cdf([0, 1, 2, 3, 4, 5], loc = 1, scale
= 4)
>>> np.mean(norm.rvs(5, size=500))
>>> from scipy.stats import gamma
>>> gamma.numargs
>>> gamma.shapes
>>> from scipy.stats import hypergeom
>>> [M, n, N] = [20, 7, 12]
>>> np.random.seed(282629734)
>>> x = stats.t.rvs(10, size=1000)
>>> print x.max(), x.min()  # equivalent to
np.max(x), np.min(x)
>>> print x.mean(), x.var() # equivalent to
np.mean(x), np.var(x)
>>> m, v, s, k = stats.t.stats(10, moments='mvsk')
```

```
>>> n, (smin, smax), sm, sv, ss, sk =
stats.describe(x)
>>> print 'distribution:',
>>> sstr = 'mean = %6.4f, variance = %6.4f, skew =
%6.4f, kurtosis = %6.4f'
>>> print sstr %(m, v, s ,k)
>>> print 'sample:      ',
>>> print sstr %(sm, sv, ss, sk)
>>> print 't-statistic = %6.3f pvalue = %6.4f' %
stats.ttest_1samp(x, m)
>>> tt = (sm-m)/np.sqrt(sv/float(n))  # t-statistic
for mean
>>> pval = stats.t.sf(np.abs(tt), n-1)*2  # two-
sided pvalue = Prob(abs(t)>tt)
>>> print 't-statistic = %6.3f pvalue = %6.4f' %
(tt, pval)
>>> print 'KS-statistic D = %6.3f pvalue = %6.4f' %
stats.kstest(x, 't', (10,))
>>> print 'KS-statistic D = %6.3f pvalue = %6.4f' %
stats.kstest(x,'norm')
>>> d, pval = stats.kstest((x-x.mean())/x.std(),
'norm')
>>> print 'KS-statistic D = %6.3f pvalue = %6.4f' %
(d, pval)
>>> rvs1 = stats.norm.rvs(loc=5, scale=10,
size=500)
```

```
>>> rvs2 = stats.norm.rvs(loc=5, scale=10,
size=500)
>>> stats.ttest_ind(rvs1, rvs2)
>>> stats.ks_2samp(rvs1, rvs2)
```

تعريف دوال

```
import random
def die():
      return random.choice([1,2,3,4,5,6])

die()

def roll(n):
      result=''
      for i in range(n):
          result = result + str(die())
          print(result)

roll(10)



import matplotlib
from matplotlib import pylab
vals = [1,200]
for i in range(1000):
              num1 = random.choice(range(1,100))
              num2 = random.choice(range(1,100))
              vals.append(num1+num2)
```

```
pylab.hist(vals, bins=10)
```

فتح ملف تفاعليا:

```
Python 2.7.9 |Anaconda 2.2.0 (64-bit)| (default, Dec 18
2014, 16:57:52) [MSC v.1500 64 bit (AMD64)] on win32
Type "copyright", "credits" or "license()" for more
information.
>>> import numpy as np
>>> import scipy as sc
>>> import pandas as pd
>>> from pandas import *
>>> import Tkinter,tkFileDialog
>>> ## C:\Users\ibin\Desktop\R python course\hsb2.csv
>>> hsb = tkFileDialog.askopenfile()
>>> hsb2 = read_csv(hsb)
>>> hsb.close()
>>> hsb2
      id  female  race  ses  schtyp  prog  read  write  math  science  socst
0     70       0     4    1       1     1    57     52    41       47     57
1    121       1     4    2       1     3    68     59    53       63     61

198  118       1     4    2       1     1    55     62    58       58     61
199  137       1     4    3       1     2    63     65    65       53     61


>>> hsb2["race"]
0     4
1     4

198    4
199    4
Name: race, Length: 200, dtype: int64
>>> x = hsb2["read"]
>>> y = hsb2["write"]
>>> z = hsb2["math"]
```

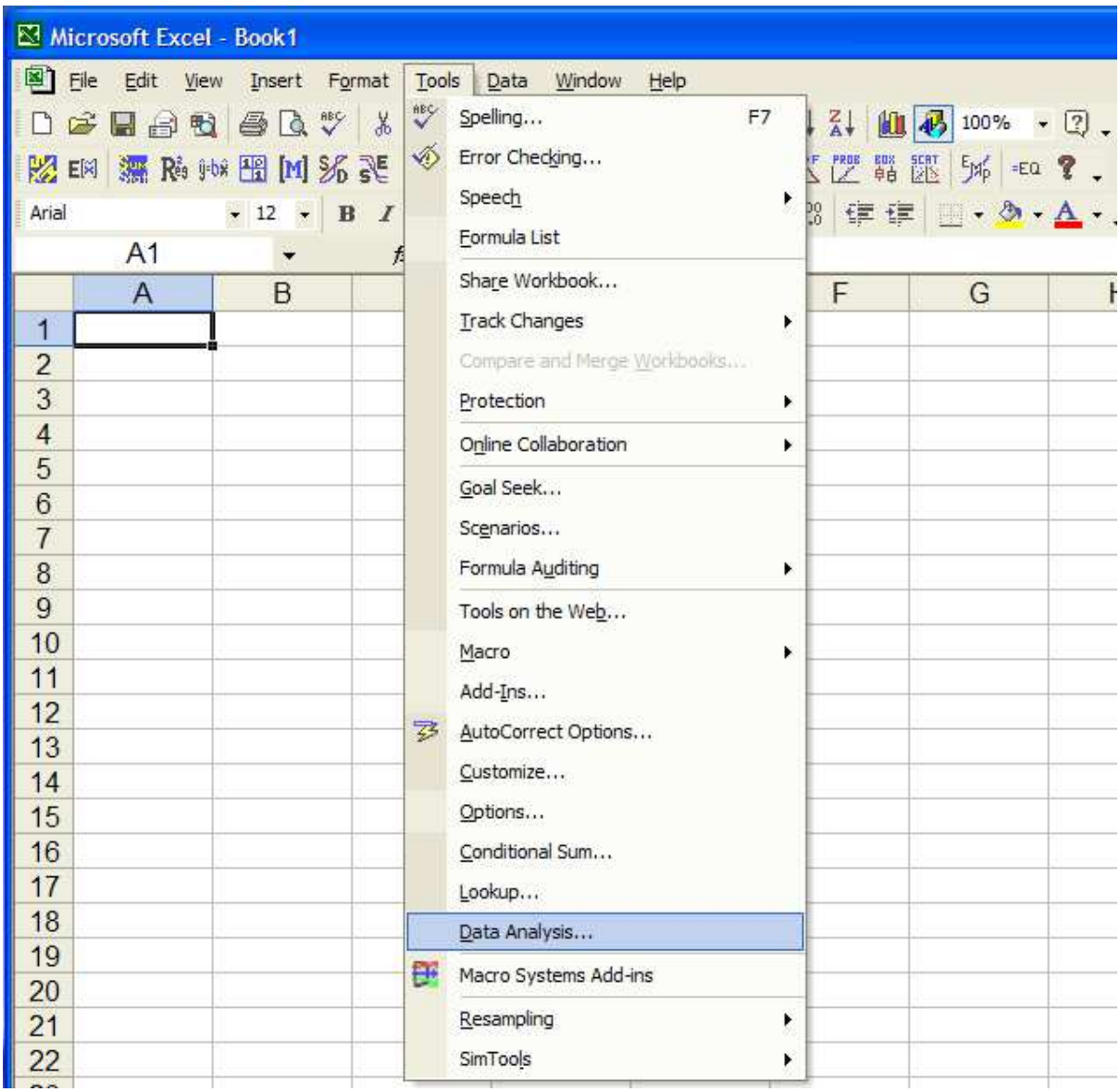
```
>>> w = hsb2['science']
>>> from scipy import stats
>>> stats.ttest_ind(x,y)
(-0.551990645527904, 0.58126457287969857)
>>> stats.ttest_ind(x,z)
(-0.42257874015791508, 0.67283084594388431)
>>> print hsb2.mean()
id          100.500
female        0.545
race          3.430
ses           2.055
schtyp        1.160
prog          2.025
read         52.230
write        52.775
math         52.645
science      51.850
socst        52.405
dtype: float64
>>> hsb2.describe()
                id      female        race         ses      schtyp        prog  \
count  200.000000  200.00000  200.000000  200.000000  200.000000  200.000000
mean   100.500000    0.54500    3.430000    2.055000    1.160000    2.025000
std     57.879185    0.49922    1.039472    0.724291    0.367526    0.690477
min      1.000000    0.00000    1.000000    1.000000    1.000000    1.000000
25%     50.750000    0.00000    3.000000    2.000000    1.000000    2.000000
50%    100.500000    1.00000    4.000000    2.000000    1.000000    2.000000
75%    150.250000    1.00000    4.000000    3.000000    1.000000    2.250000
max    200.000000    1.00000    4.000000    3.000000    2.000000    3.000000

             read       write        math     science       socst
count  200.000000  200.000000  200.000000  200.000000  200.000000
mean    52.230000   52.775000   52.645000   51.850000   52.405000
std     10.252937    9.478586    9.368448    9.900891   10.735793
min     28.000000   31.000000   33.000000   26.000000   26.000000
25%     44.000000   45.750000   45.000000   44.000000   46.000000
```

```
50%     50.000000   54.000000   52.000000   53.000000   52.000000
75%     60.000000   60.000000   59.000000   58.000000   61.000000
max     76.000000   67.000000   75.000000   74.000000   71.000000
>>> help(ols)
>>> result = ols(y=y, x=x)
>>> result
---------Summary of Regression Analysis------------------

Formula: Y ~ <x> + <intercept>

Number of Observations:         200
Number of Degrees of Freedom:   2

R-squared:         0.3561
Adj R-squared:     0.3529

Rmse:              7.6249

F-stat (1, 198):   109.5213, p-value:     0.0000

Degrees of Freedom: model 1, resid 198

--------Summary of Estimated Coefficients---------------
Variable   Coef  Std Err  t-stat p-value CI 2.5%   CI 97.5%
-------------------------------------------------------------
    x      0.5517  0.0527   10.47  0.0000   0.4484  0.6550
intercept 23.9594  2.8057    8.54  0.0000  18.4602 29.4587
------------End of Summary---------------------------------

>>>
```

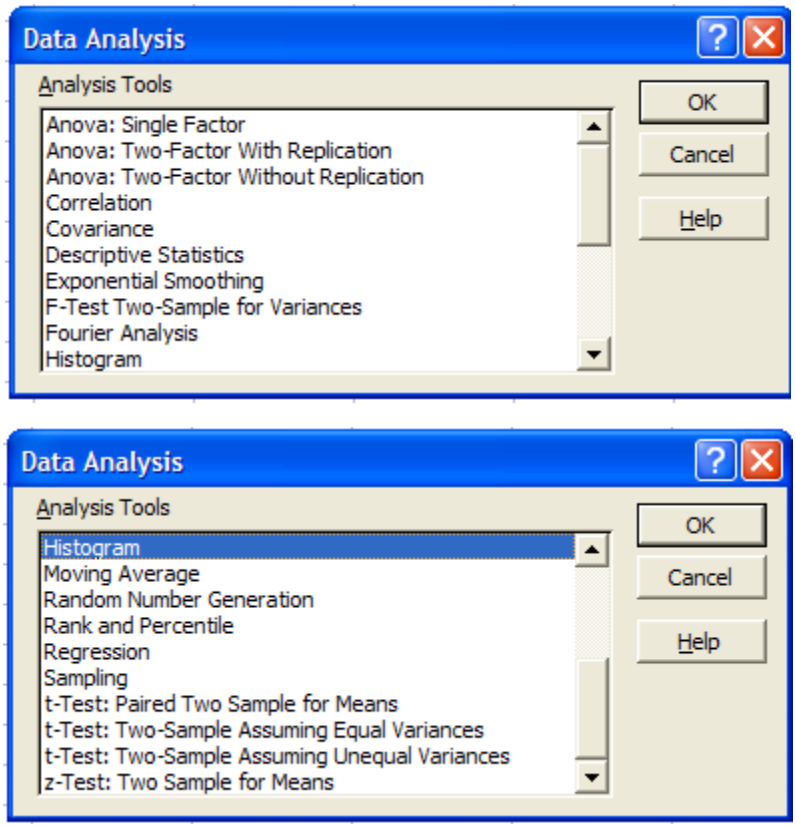
مقدمة لدوال التحليل الإحصائي في Excel:

إستخدام تحليل البيانات المبني في إكسل

يوجد في إكسل إختيار ضمن قائمة الأدوات لتحليل البيانات



والذي يحوي التالي:



1- تحليل التباين لعامل واحد

2- تحليل التباين لعاملين مع تكرار

3- تحليل التباين لعاملين بدون تكرار

4- الترابط

5- التغاير

6- إحصائات وصفية

7- التمهيد الاسي

8- إختبار F للتباين لعينتين

9- تحليل فوريه

10- المدرج التكراري

11- المتوسط المتحرك

12- توليد ارقام عشوائية

13- الرتب والمئينات

14- الإنحدار

15- المعاينة

16- إختبار t للمتوسطات لعينتين متقارنة

17- إختبار t لعينتين على إفتراض تساوي التباين

18- إختبار t لعينتين على إفتراض عدم تساوي التباين

19- إختبار z للمتوسطات لعينتين

وسوف نستعرض بعض هذه الطرق فيما يلي:

تحليل التباين لعامل واحد Anova: Single Factor

اجريت دراسة لمعرفة الفرق بين تأثير ثلاثة طرق لتدريس مبادئ الحساب لطلاب المرحلة الأولى الإبتدائية فاختير 27 تلميذا عشوائيا وتم تخصيص 9 تلاميذ بطريقة عشوائية لكل طريقة من الطرق الثلاثة. تم اختبار جميع التلاميذ بعد فترة معينة وكانت نتائج الإختبارات كالتالي:

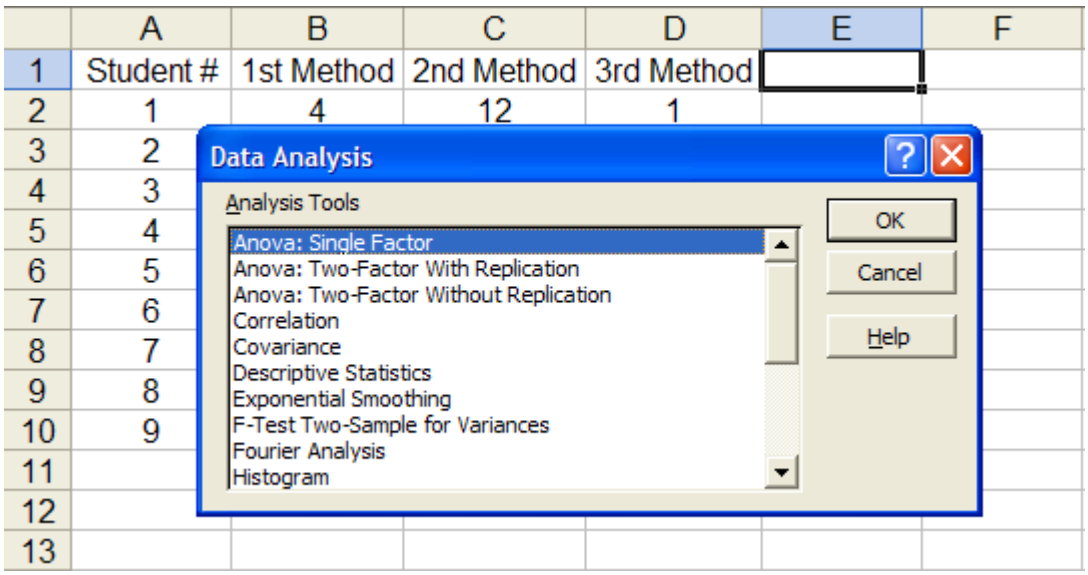
رقم الطالب	1	2	3	4	5	6	7	8	9	المجموع
الطريقة 1	4	5	4	3	6	10	1	8	5	46
الطريقة 2	12	8	10	5	7	9	14	9	4	78
الطريقة 3	1	3	4	6	8	5	3	2	2	34

المطلوب معرفة هل هناك فرق معنوي بين طرق التدريس المختلفة. اختبر عند مستوى معنوية 0.05

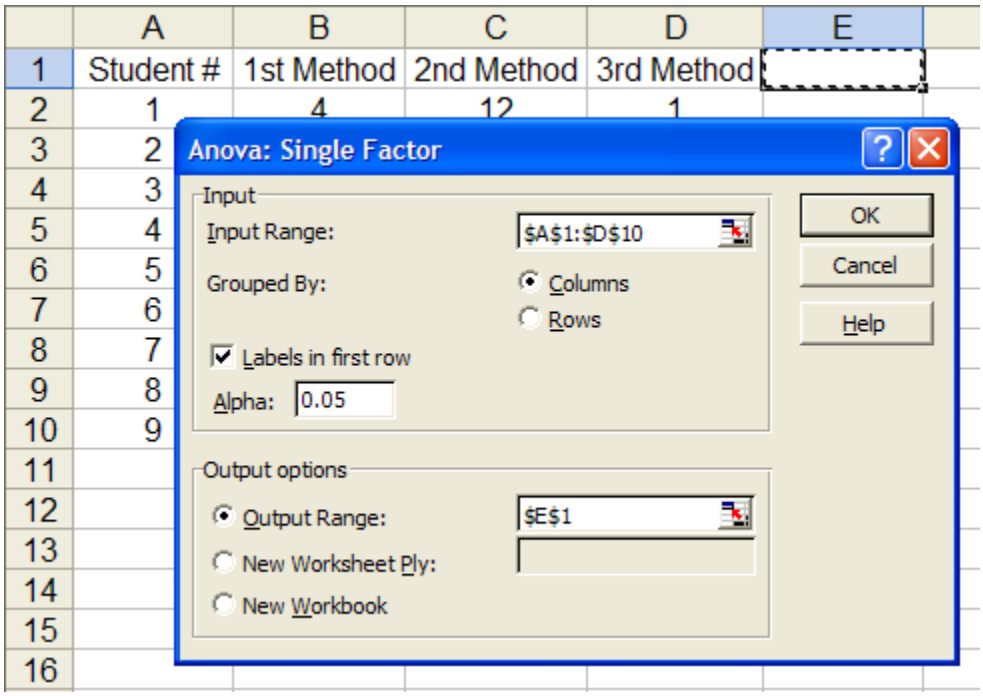
ندخل البيانات في صفحة من إكسل كالتالي:

	A	B	C	D
1	Student #	1st Method	2nd Method	3rd Method
2	1	4	12	1
3	2	5	8	3
4	3	4	10	4
5	4	3	5	6
6	5	6	7	8
7	6	10	9	5
8	7	1	14	3
9	8	8	9	2
10	9	5	4	2

من قائمة الأدوات Tools نختار Data Analysis فتظهر النافذة



نختار تحليل التباين لعامل واحد Anova: Single Factor فتظهر النافذة



ندخل البيانات المطلوبة ثم OK فينتج

E	F	G	H	I	J	K
Anova: Single Factor						
SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
Student #	9	45	5	7.5		
1st Method	9	46	5.111111111	7.111111111		
2nd Method	9	78	8.666666667	10		
3rd Method	9	34	3.777777778	4.944444444		
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	119.8611111	3	39.9537037	5.40726817	0.003997338	2.901117568
Within Groups	236.4444444	32	7.388888889			
Total	356.3055556	35				

تحليل التباين لعاملين مع تكرار Anova: Two-Factor With Replication

قام احد الباحثين بتجربتين على مجموعتين لتحديد درجة الاستيعاب التي تقاس كجزء من 100 فتحصل على النتائج التالية

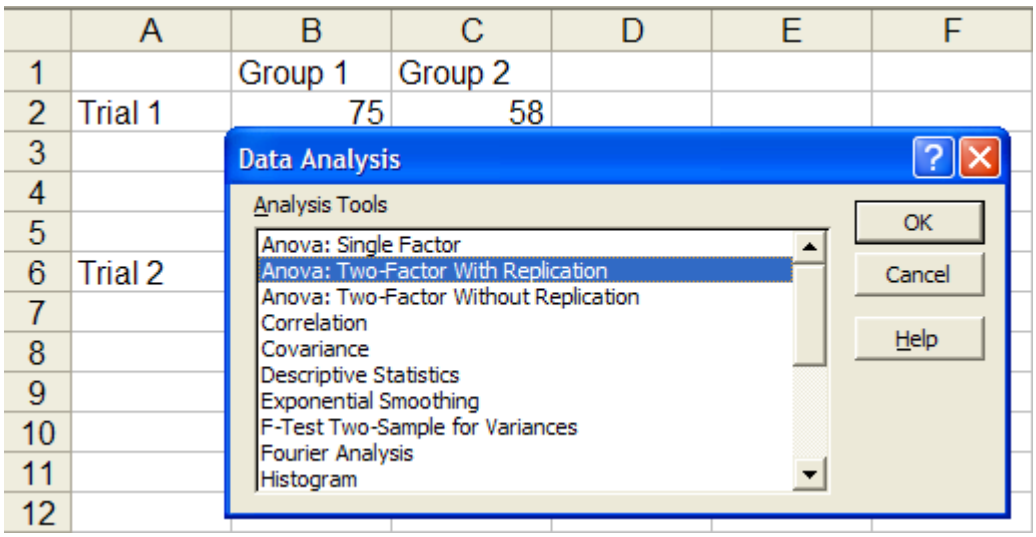
	مجموعة 1	مجموعة 2
تجربة 1	75	58
	68	56
	71	61
	75	60
تجربة 2	66	62
	70	60
	68	59
	68	68

هل هناك فرق بين التجارب وفرق بين المجموعات ؟ اختبر عند مستوى معنوية 0.05

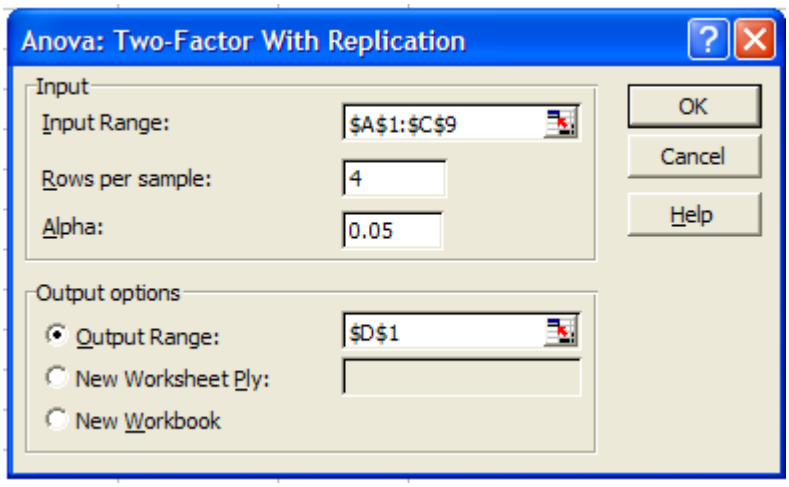
ندخل البيانات في صفحة من إكسل

	A	B	C
1		Group 1	Group 2
2	Trial 1	75	58
3		68	56
4		71	61
5		75	60
6	Trial 2	66	62
7		70	60
8		68	59
9		68	68

كالسابق من قائمة الأدوات وتحت تحليل البيانات نختار تحليل التباين لعاملين مع تكرار



فتظهر النافذة



ندخل البيانات كماهو موضح ثم OK فينتج

D	E	F	G
Anova: Two-Factor With Replication			
SUMMARY	Group 1	Group 2	Total
Trial 1			
Count	4	4	8
Sum	289	235	524
Average	72.25	58.75	65.5
Variance	11.58333333	4.916666667	59.14285714
Trial 2			
Count	4	4	8
Sum	272	249	521
Average	68	62.25	65.125
Variance	2.666666667	16.25	17.55357143
Total			
Count	8	8	
Sum	561	484	
Average	70.125	60.5	
Variance	11.26785714	12.57142857	

ANOVA		
Source of Variation	*SS*	*df*
Sample	0.5625	1
Columns	370.5625	1
Interaction	60.0625	1
Within	106.25	12
Total	537.4375	15

MS	*F*	*P-value*	*F crit*
0.5625	0.063529412	0.805267228	4.747221283
370.5625	41.85176471	3.07191E-05	4.747221283
60.0625	6.783529412	0.023033141	4.747221283
8.854166667			

تحليل التباين لعاملين بدون تكرار Anova: Two-Factor Without Replication

استخدم أحد الباحثين 4 انواع من السماد A,B,C,D لمعالجة 4 قطاعات من الأراضي قطاع 1 وحتى قطاع 4 فتحصل على الإنتاج التالى بالأطنان

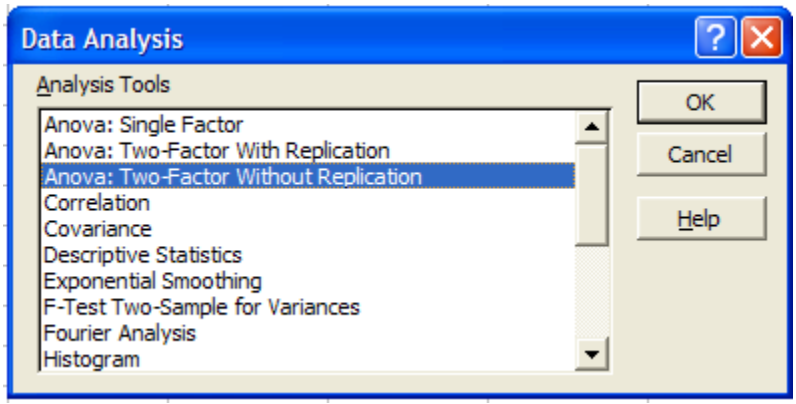
Treatment	Sector 1	Sector 2	Sector 3	Sector 4
A	9.3	9.4	9.6	10
B	9.4	9.3	9.8	9.9
C	9.2	9.4	9.5	9.7
D	9.7	9.6	10	10.2

هل هناك فرق بين المعالجات؟ هل هناك فرق بين القطاعات؟ اختبر عند 0.05

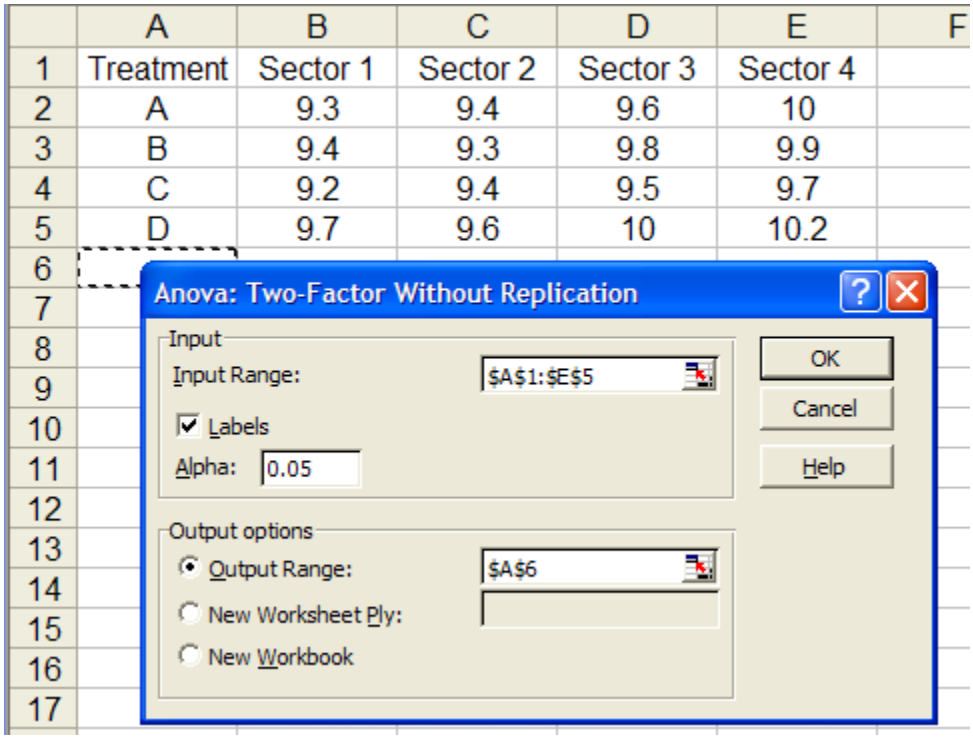
تدخل البيانات كالتالي:

	A	B	C	D	E
1	Treatment	Sector 1	Sector 2	Sector 3	Sector 4
2	A	9.3	9.4	9.6	10
3	B	9.4	9.3	9.8	9.9
4	C	9.2	9.4	9.5	9.7
5	D	9.7	9.6	10	10.2
6					

كالسابق من قائمة الأدوات وتحت تحليل البيانات نختار تحليل التباين لعاملين بدون تكرار



ثم ندخل المطلوب كالتالي



فينتج

Anova: Two-Factor Without Replication				
SUMMARY	*Count*	*Sum*	*Average*	*Variance*
A	4	38.3	9.575	0.095833333
B	4	38.4	9.6	0.086666667
C	4	37.8	9.45	0.043333333
D	4	39.5	9.875	0.075833333
Sector 1	4	37.6	9.4	0.046666667
Sector 2	4	37.7	9.425	0.015833333
Sector 3	4	38.9	9.725	0.049166667
Sector 4	4	39.8	9.95	0.043333333

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	0.385	3	0.128333333	14.4375	0.000871272	3.862538733
Columns	0.825	3	0.275	30.9375	4.52327E-05	3.862538733
Error	0.08	9	0.008888889			
Total	1.29	15				

الترابط Correlation

الجدول التالي يوضح السن X وضغط الدم Y لثمان من الإناث :

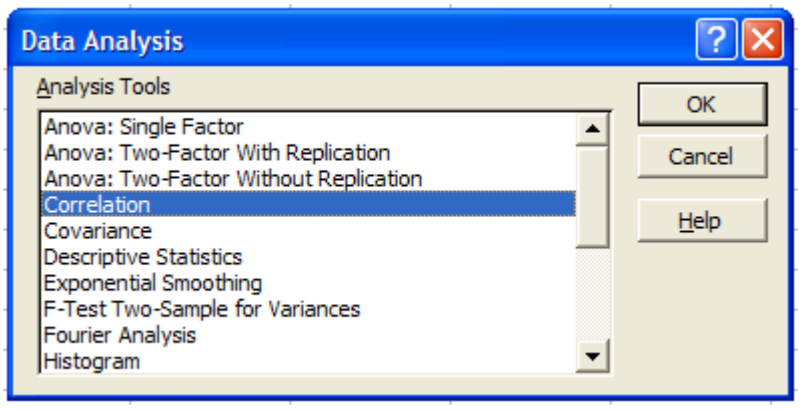
السن X	42	36	63	55	42	60	49	68
ضغط الدم Y	125	118	140	150	140	155	145	152

أوجد معامل الارتباط بين Y و X .

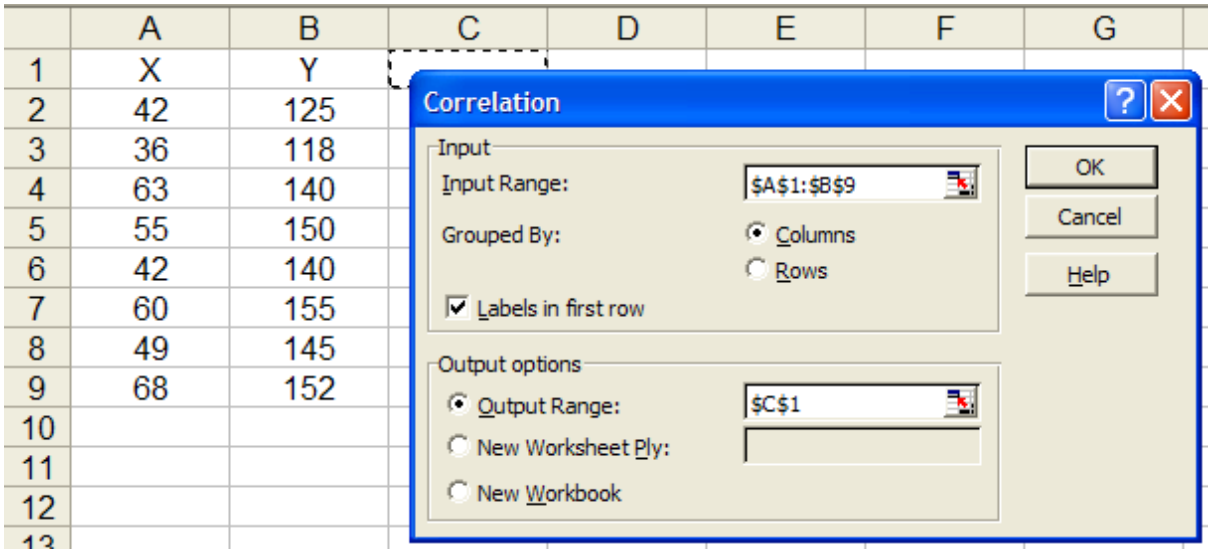
تدخل البيانات في صفحة من إكسل

	A	B
1	X	Y
2	42	125
3	36	118
4	63	140
5	55	150
6	42	140
7	60	155
8	49	145
9	68	152

كالسابق من قائمة الأدوات وتحت تحليل البيانات نختار الترابط



ثم OK فتظهر النافذة



ندخل المطلوب كما في الشكل فينتج

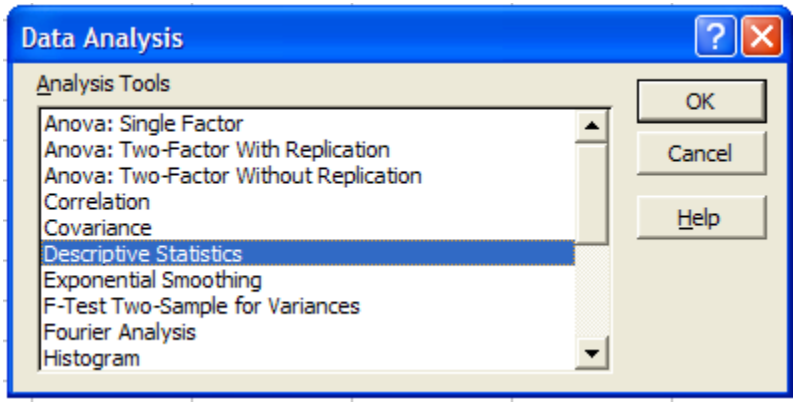
	A	B	C	D	E
1	X	Y		*X*	*Y*
2	42	125	X	1	
3	36	118	Y	0.791832	1
4	63	140			
5	55	150			
6	42	140			
7	60	155			
8	49	145			
9	68	152			

الإحصائات الوصفية Descriptive Statistica

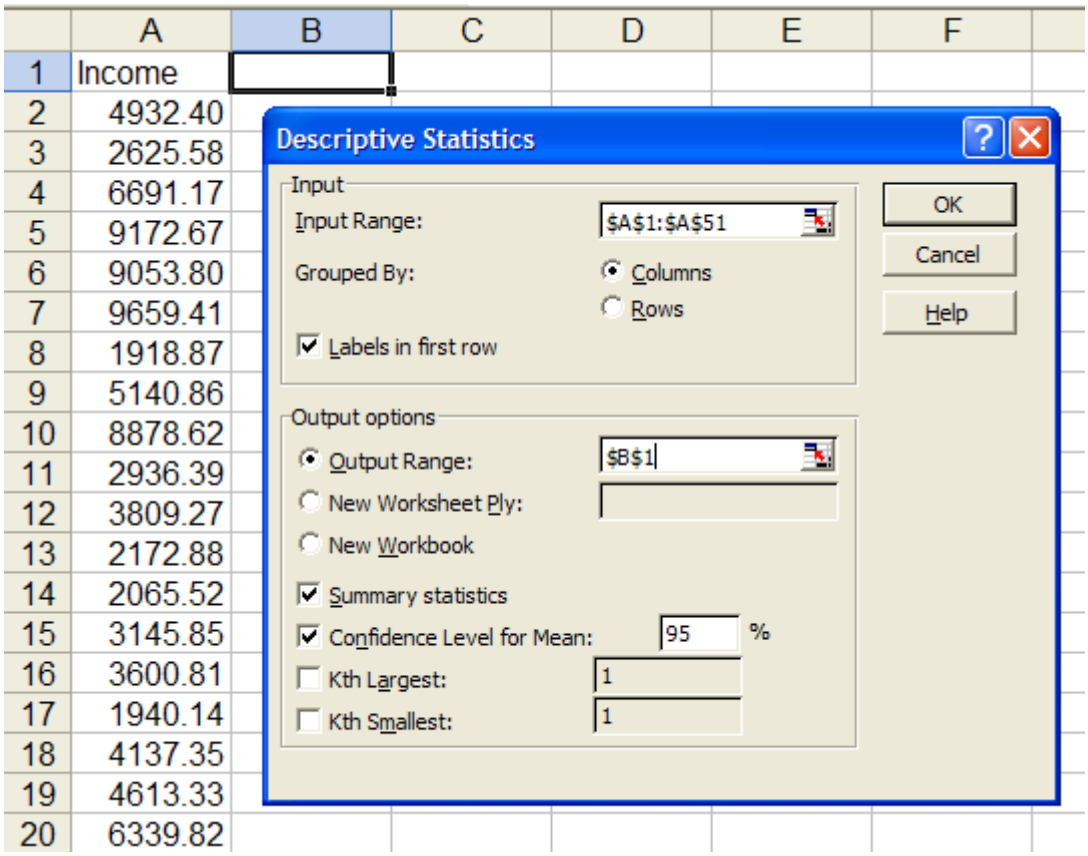
البيانات التالية هي الدخل الشهري بالريال (لأقرب هللة) لعينة من 50 متخرجا من جامعة الملك سعود (للكليات غير الطبية)

4932.40, 2625.58, 6691.17, 9172.67, 9053.80, 9659.41, 1918.87, 5140.86, 8878.62, 2936.39, 3809.27, 2172.88, 2065.52, 3145.85, 3600.81, 1940.14, 4137.35, 4613.33, 6339.82, 4730.45, 4849.07, 4715.93, 9264.51, 5621.34, 5294.52, 4292.01, 9800.80, 8414.65, 9928.18, 3901.36 9603.85, 2238.19, 7581.32, 8495.49, 9774.52, 5623.85, 4261.73, 7951.69, 4682.15, 8160.40, 2409.61, 3427.14, 2325.28, 4738.46, 5793.77, 5991.97, 4862.33, 9884.38, 2133.84, 3691.90

بإختيار إحصائات وصفية من قائمة الإختيارات



تظهر النافذة



هنا أخترنا جميع الإحصائيات الملخصة وكذلك فترة 95% للثقة بالضغط على OK ينتج

	A	B	C
1	Income	*Income*	
2	4932.40		
3	2625.58	Mean	5545.59
4	6691.17	Standard Error	369.81
5	9172.67	Median	4855.70
6	9053.80	Mode	#N/A
7	9659.41	Standard Deviation	2614.93
8	1918.87	Sample Variance	6837855.24
9	5140.86	Kurtosis	-1.16
10	8878.62	Skewness	0.37
11	2936.39	Range	8009.31
12	3809.27	Minimum	1918.87
13	2172.88	Maximum	9928.18
14	2065.52	Sum	277279.42
15	3145.85	Count	50.00
16	3600.81	Confidence Level(95.0%)	743.15
17	1940.14		

التمهيد الاسى Exponential Smoothing

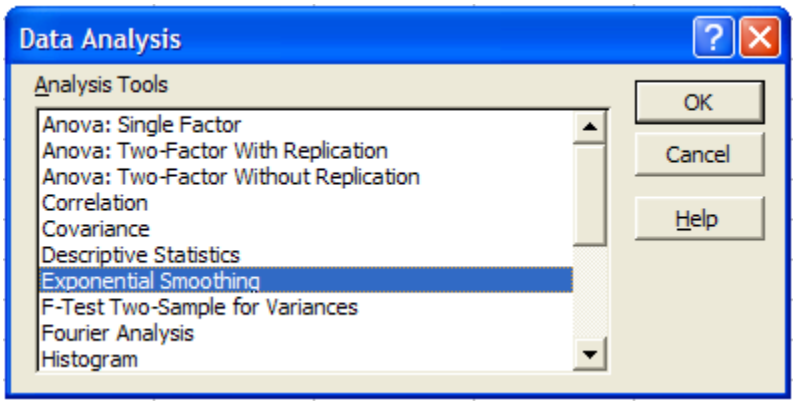
يستخدم التمهيد الاسي للتنبؤ عن القيمة المستقبلية التالية في سلسلة من المشاهدات لمتغير عشوائي معطى. التمهيد الاسي هو احد الطرق المستخدمة في التنبؤ الإحصائي (انظر كتاب: طرق التنبؤ الإحصائي – الجزء الأول- تأليف: د. عدنان ماجد بري)

البيانات التالية لظاهرة عشوائية

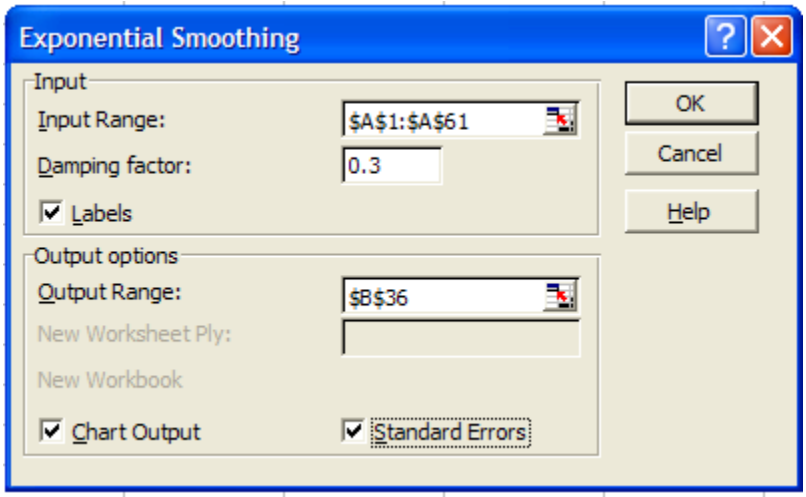
44.2	44.3	44.4	43.4	42.8	44.3
44.4	44.8	44.4	43.1	42.6	42.4
42.2	41.8	40.1	42.0	42.4	43.1
42.4	43.1	43.2	42.8	43.0	42.8
42.5	42.6	42.3	42.9	43.6	44.7
44.5	45.0	44.8	44.9	45.2	45.2
45.0	45.5	46.2	46.8	47.5	48.3

48.3	49.1	48.9	49.4	50.0	50.0
49.6	49.9	49.6	50.7	50.7	50.9
50.5	51.2	50.7	50.3	49.2	48.1

سوف نستخدم التمهيد الاسي (البسيط) علي البيانات

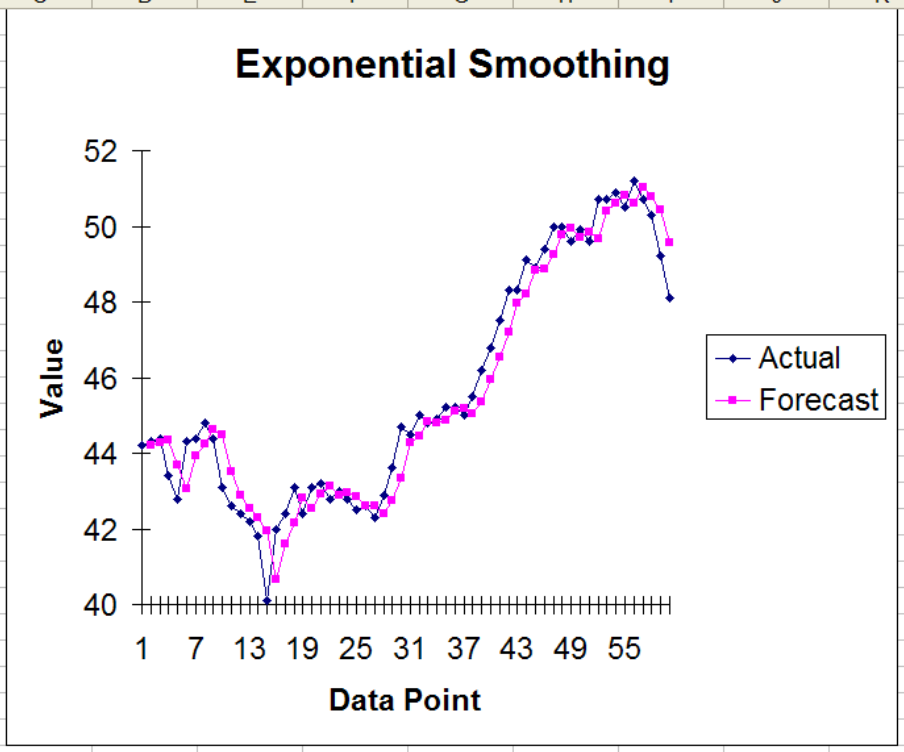


فتظهر نافذة الإدخال



وينتج

	A	B
1	Value	#N/A
2	44.2	44.2
3	44.3	44.27
4	44.4	44.361
5	43.4	43.6883
6	42.8	43.06649
7	44.3	43.92995
8	44.4	44.25898
9	44.8	44.6377
10	44.4	44.47131
11	43.1	43.51139
12	42.6	42.87342
13	42.4	42.54203
14	42.2	42.30261
15	41.8	41.95078
16	40.1	40.65523
17	42	41.59657
18	42.4	42.15897
19	43.1	42.81769
20	42.4	42.52531
21	43.1	42.92759
22	43.2	43.11828
23	42.8	42.89548
24	43	42.96864
25	42.8	42.85059
26	42.5	42.60518
27	42.6	42.60155
28	42.3	42.39047
29	42.9	42.74714

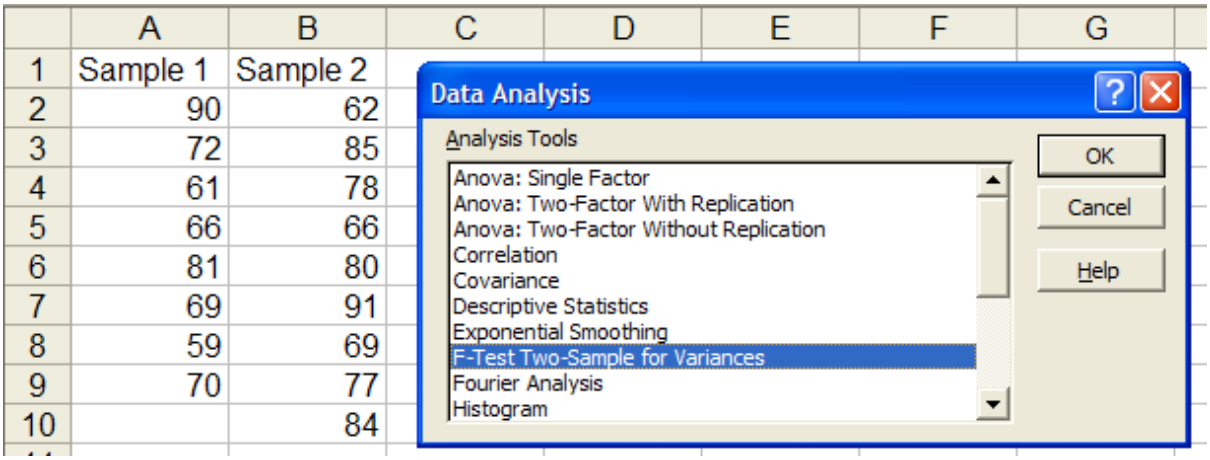


إختبار F للتباين لعينتين:

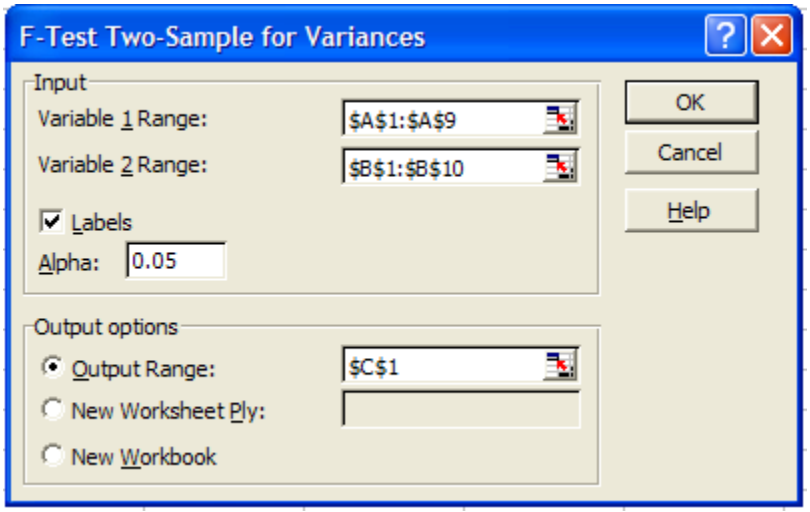
نريد ان نختبر تساوي تباين العينتين التالية:

عينة 1: 70 59 69 81 66 61 72 90

عينة 2: 84 77 69 91 80 66 78 85 61



وندخل البيانات كالتالي



فينتج

C	D	E
F-Test Two-Sample for Variances		
	Sample 1	*Sample 2*
Mean	71	76.88888889
Variance	105.1428571	91.11111111
Observations	8	9
df	7	8
F	1.154006969	
P(F<=f) one-tail	0.418377164	
F Critical one-tail	3.500460366	

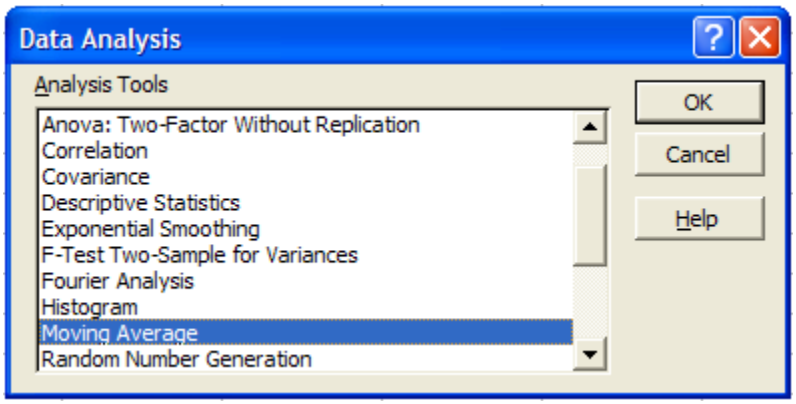
المتوسط المتحرك Moving Average

يستخدم المتوسط المتحرك مثل التمهيد الاسي للتنبؤ عن القيمة المستقبلية التالية في سلسلة من المشاهدات لمتغير عشوائي معطى. المتوسط المتحرك هو احد الطرق المستخدمة في التنبؤ الإحصائي (انظر كتاب: طرق التنبؤ الإحصائي – الجزء الأول - تأليف: د. عدنان ماجد بري)

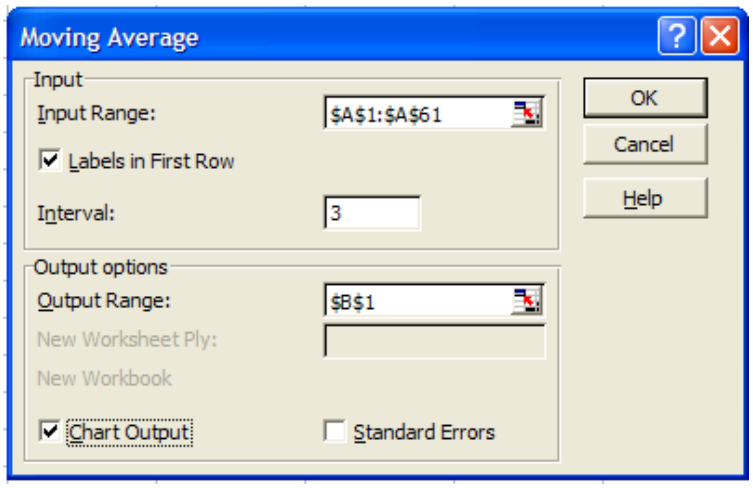
البيانات التالية لظاهرة عشوائية

44.2	44.3	44.4	43.4	42.8	44.3	44.4	44.8	44.4	43.1
42.6	42.4	42.2	41.8	40.1	42.0	42.4	43.1	42.4	43.1
43.2	42.8	43.0	42.8	42.5	42.6	42.3	42.9	43.6	44.7
44.5	45.0	44.8	44.9	45.2	45.2	45.0	45.5	46.2	46.8
47.5	48.3	48.3	49.1	48.9	49.4	50.0	50.0	49.6	49.9
49.6	50.7	50.7	50.9	50.5	51.2	50.7	50.3	49.2	48.1

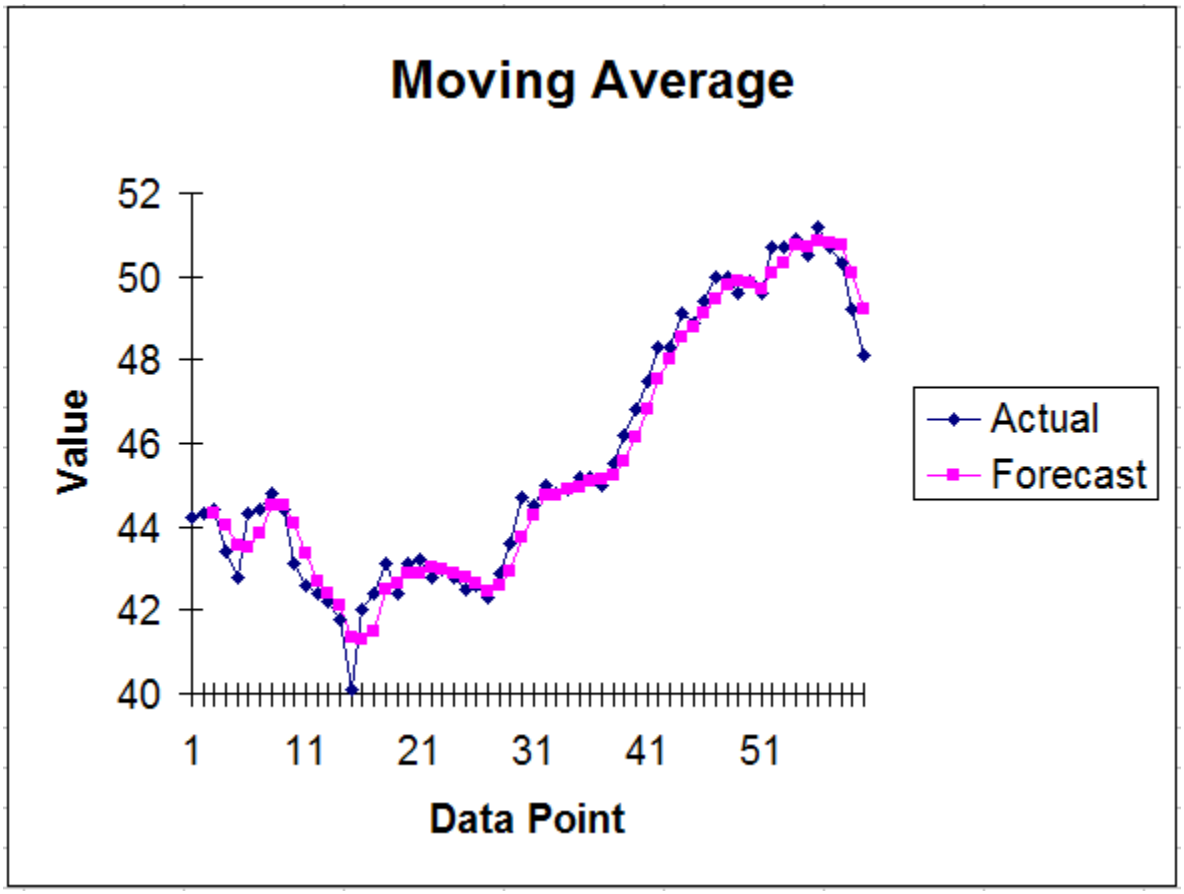
سوف نستخدم المتوسط المتحرك عليها



فتظهر نافذة الإدخال


Moving Average
Input
Input Range:
$A$1:$A$61
Labels in First Row
Interval:
3
OK
Cancel
Help
Output options
Output Range:
$B$1
New Worksheet Ply:
New Workbook
Chart Output
Standard Errors

وينتج

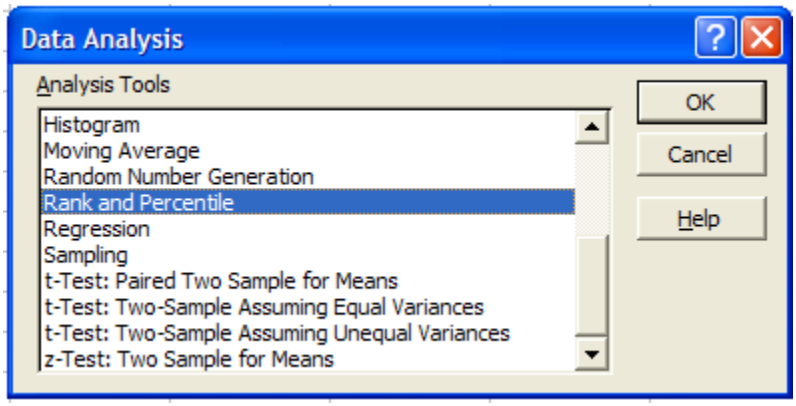

Moving Average
Value
52
50
48
46
44
42
40
1
11
21
31
41
51
Data Point
Actual
Forecast

الرتب والمئينات Rank and Percentile

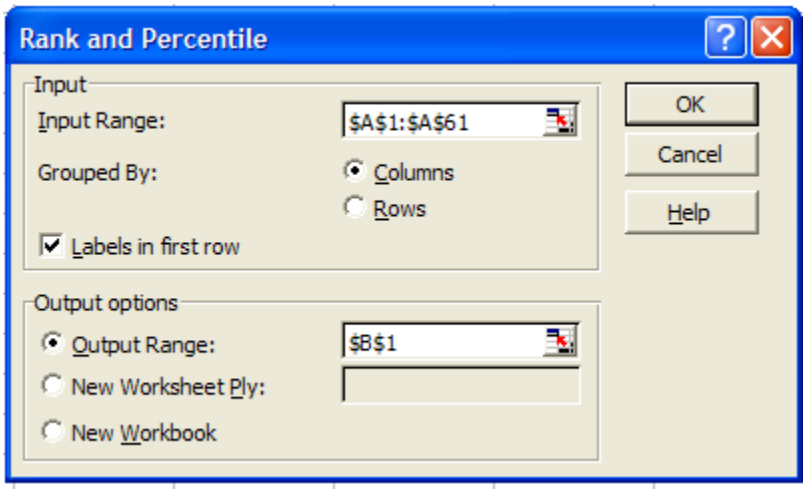
سوف نوجد رتب ومئينات البيانات التالية

44.2	44.3	44.4	43.4	42.8	44.3	44.4	44.8	44.4	43.1	42.6
42.4	42.2	41.8	40.1	42.0	42.4	43.1	42.4	43.1	43.2	42.8
43.0	42.8	42.5	42.6	42.3	42.9	43.6	44.7	44.5	45.0	44.8
44.9	45.2	45.2	45.0	45.5	46.2	46.8	47.5	48.3	48.3	49.1
48.9	49.4	50.0	50.0	49.6	49.9	49.6	50.7	50.7	50.9	50.5
51.2	50.7	50.3	49.2	48.1						

من قائمة الأدوات ومن تحليل البيانات نختار Rank and Percentile كالتالي:



وفي نافذة المدخلات ندخل المطلوب كالتالي:



فينتج

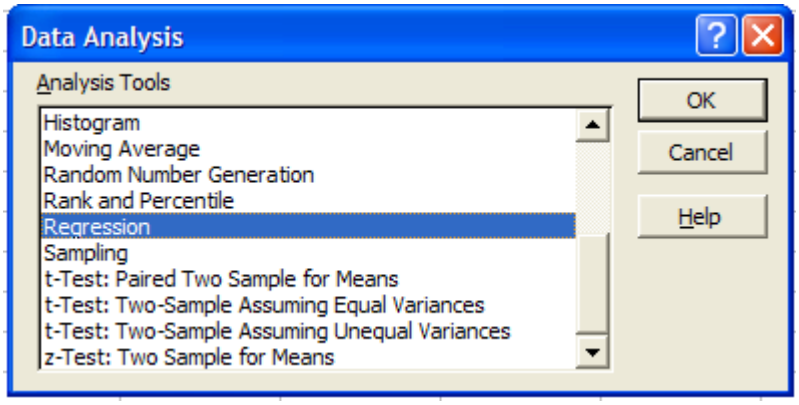
	A	B	C	D	E
1	Value	*Point*	*Value*	*Rank*	*Percent*
2	44.2	56	51.2	1	100.00%
3	44.3	54	50.9	2	98.30%
4	44.4	52	50.7	3	93.20%
5	43.4	53	50.7	3	93.20%
6	42.8	57	50.7	3	93.20%
7	44.3	55	50.5	6	91.50%
8	44.4	58	50.3	7	89.80%
9	44.8	47	50	8	86.40%
10	44.4	48	50	8	86.40%
11	43.1	50	49.9	10	84.70%
12	42.6	49	49.6	11	81.30%
13	42.4	51	49.6	11	81.30%
14	42.2	46	49.4	13	79.60%
15	41.8	59	49.2	14	77.90%

16	40.1	44	49.1	15	76.20%
17	42	45	48.9	16	74.50%
18	42.4	42	48.3	17	71.10%
19	43.1	43	48.3	17	71.10%
20	42.4	60	48.1	19	69.40%
21	43.1	41	47.5	20	67.70%
22	43.2	40	46.8	21	66.10%
23	42.8	39	46.2	22	64.40%
24	43	38	45.5	23	62.70%
25	42.8	35	45.2	24	59.30%
26	42.5	36	45.2	24	59.30%
27	42.6	32	45	26	55.90%
28	42.3	37	45	26	55.90%
29	42.9	34	44.9	28	54.20%
30	43.6	8	44.8	29	50.80%
31	44.7	33	44.8	29	50.80%
32	44.5	30	44.7	31	49.10%

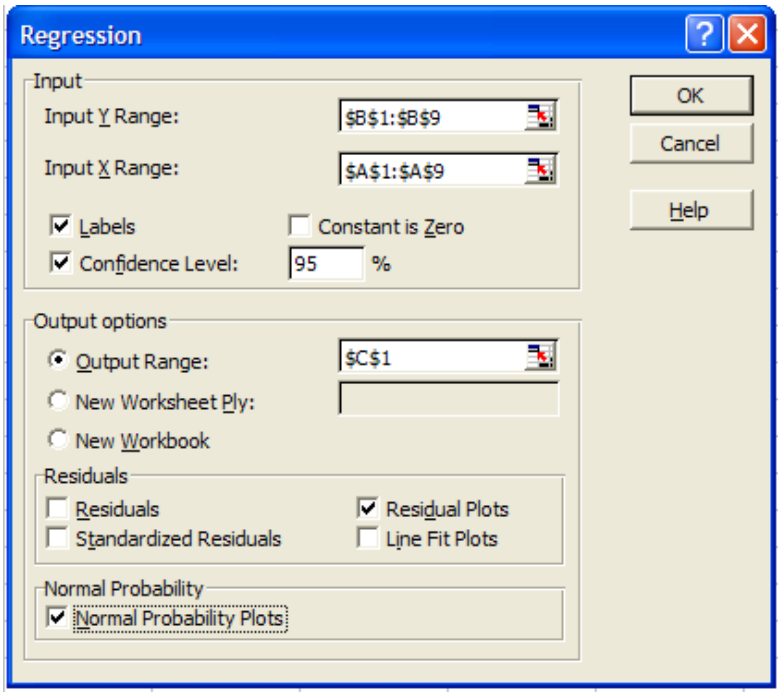
الإنحدار Regression

سوف نستعرض الإنحدار الخطي على البيانات التالية

X	Y
42	125
36	118
63	140
55	150
42	140
60	155
49	145
68	152



فتظهر نافذة المدخلات



والنتائج

SUMMARY OUTPUT					
Regression Statistics					
Multiple R	0.791831817				
R Square	0.626997626				
Adjusted R Square	0.564830564				
Standard Error	8.636706775				
Observations	8				
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	752.3187765	752.3187765	10.0856885	0.019177545
Residual	6	447.5562235	74.59270392		
Total	7	1199.875			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	93.58382293	15.12386375	6.187825046	0.000819999
X	0.906817871	0.285540223	3.175797302	0.019177545

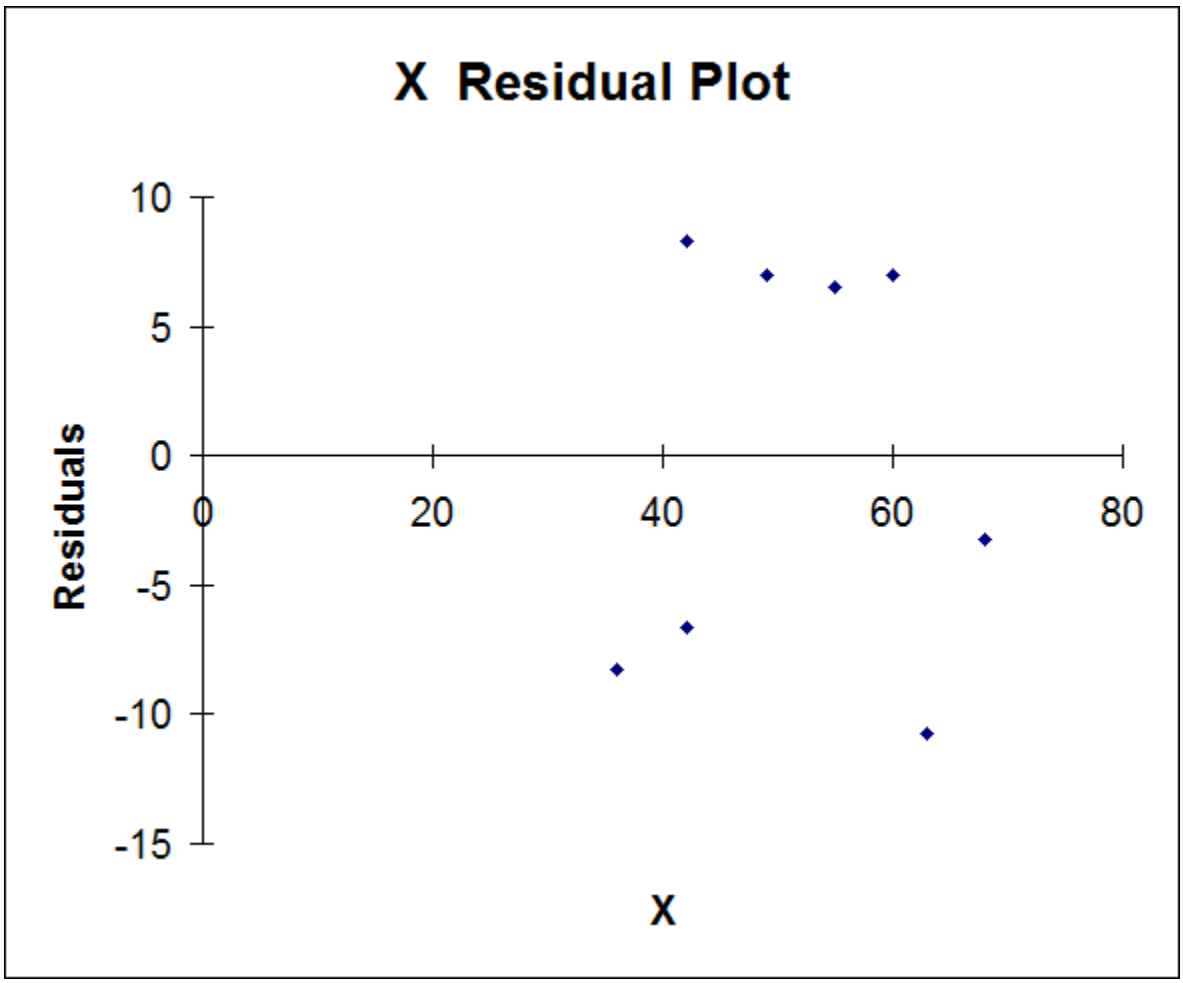
Lower 95%	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
56.57703441	130.5906114	56.57703441	130.5906114
0.208125604	1.605510139	0.208125604	1.605510139

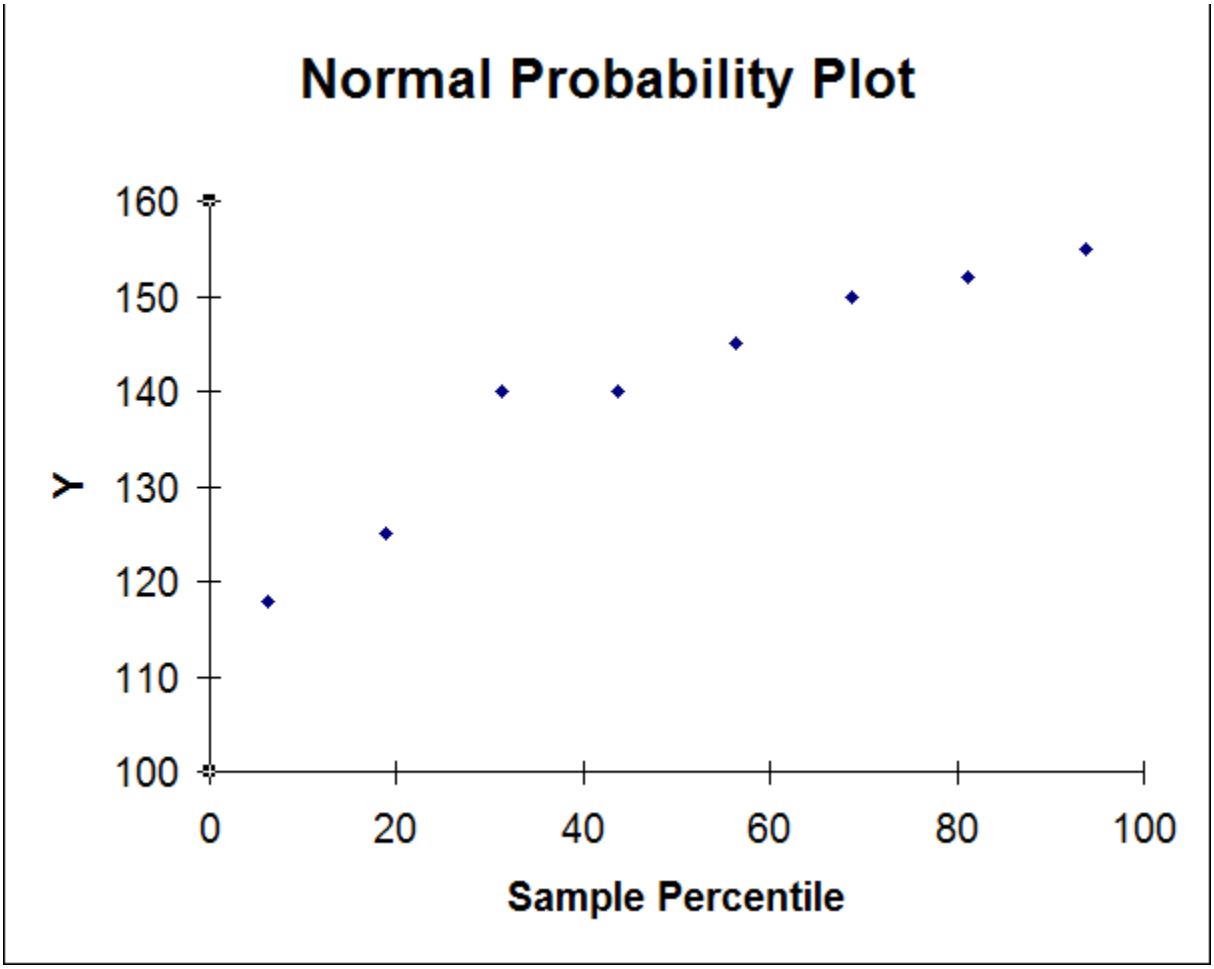
RESIDUAL OUTPUT

Observation	*Predicted Y*	*Residuals*
1	131.6701735	-6.670173521
2	126.2292663	-8.229266293
3	150.7133488	-10.71334882
4	143.4588058	6.541194152
5	131.6701735	8.329826479
6	147.9928952	7.007104796
7	138.0178986	6.98210138
8	155.2474382	-3.247438175

PROBABILITY OUTPUT

Percentile	*Y*
6.25	118
18.75	125
31.25	140
43.75	140
56.25	145
68.75	150
81.25	152
93.75	155


X Residual Plot
10
5
0
-5
-10
-15
Residuals
0
20
40
60
80
X


Normal Probability Plot
160
150
140
130
120
110
100
Y
0
20
40
60
80
100
Sample Percentile

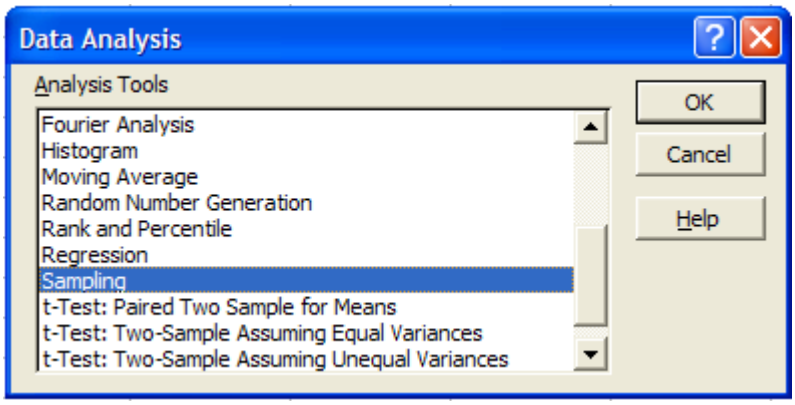
المعاينة Sampling

سوف نقوم بسحب عينة عشوائية من المجتمع $\{0,1\}$ حجمها 60 وحدة كالتالي:

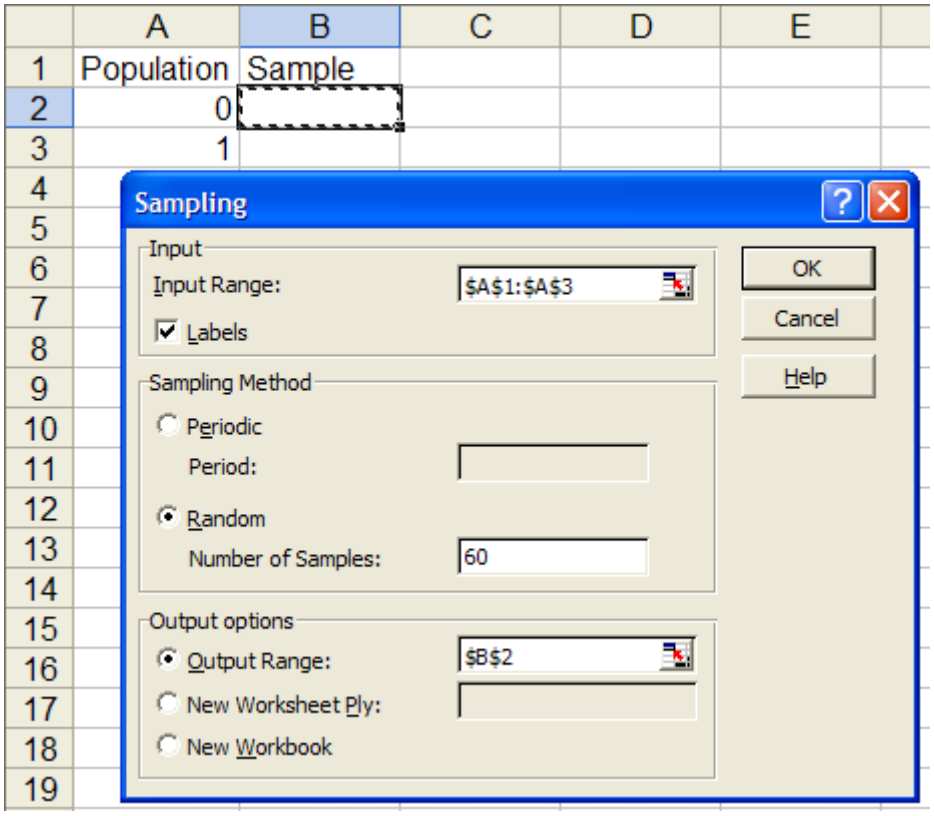
في صفحة من إكسل ادخل التالي

	A	B
1	Population	Sample
2	0	
3	1	

من قائمة الأدوات نختار تحليل البيانات ثم المعاينة



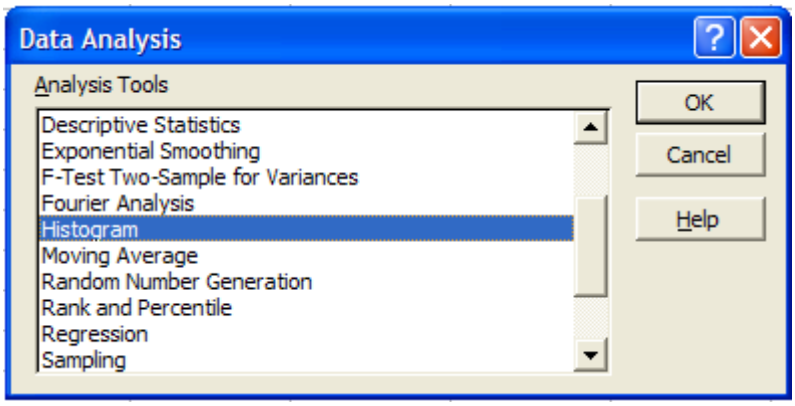
فتظهر نافذة إدخال المعلومات



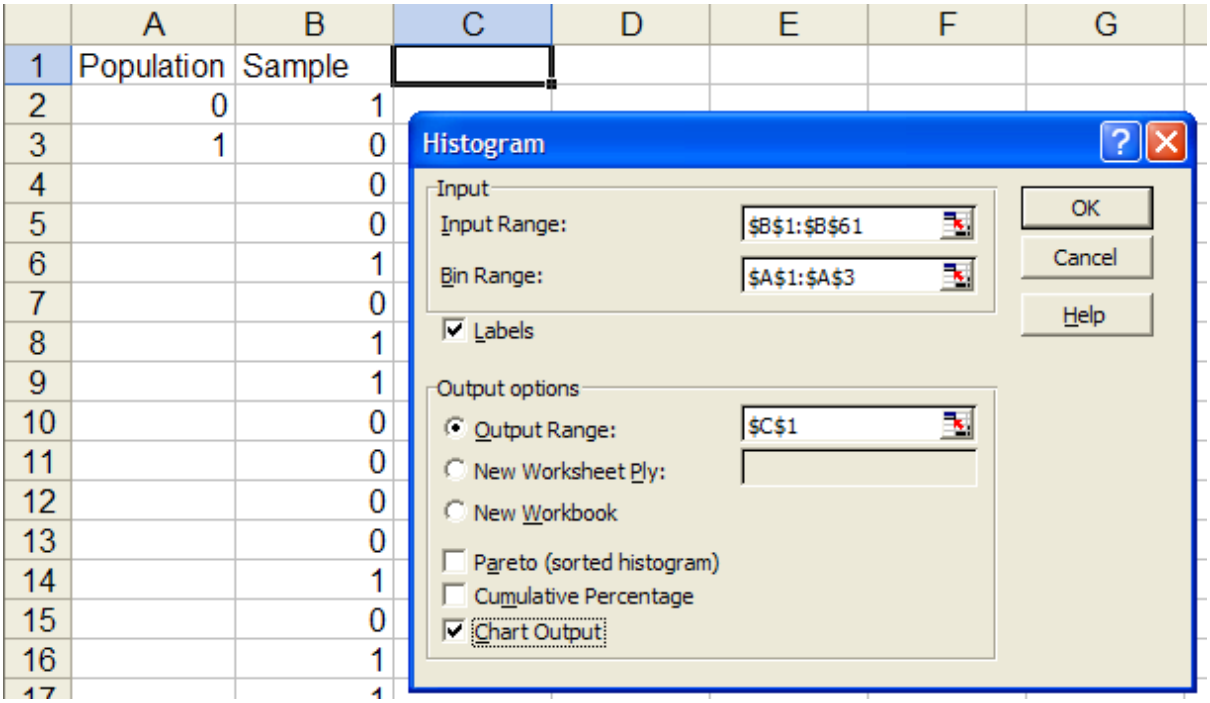
فينتج (جزء من المخرجات)

	A	B	C
1	Population	Sample	
2	0	1	
3	1	0	
4		0	
5		0	
6		1	
7		0	
8		1	
9		1	
10		0	

المعاينة هنا كانت بإحلال، سوف نوجد التوزيع التكراري للعينة بإستخدام أداة المدرج التكراري HISTOGRAM الموجودة ضمن تحليل البيانات



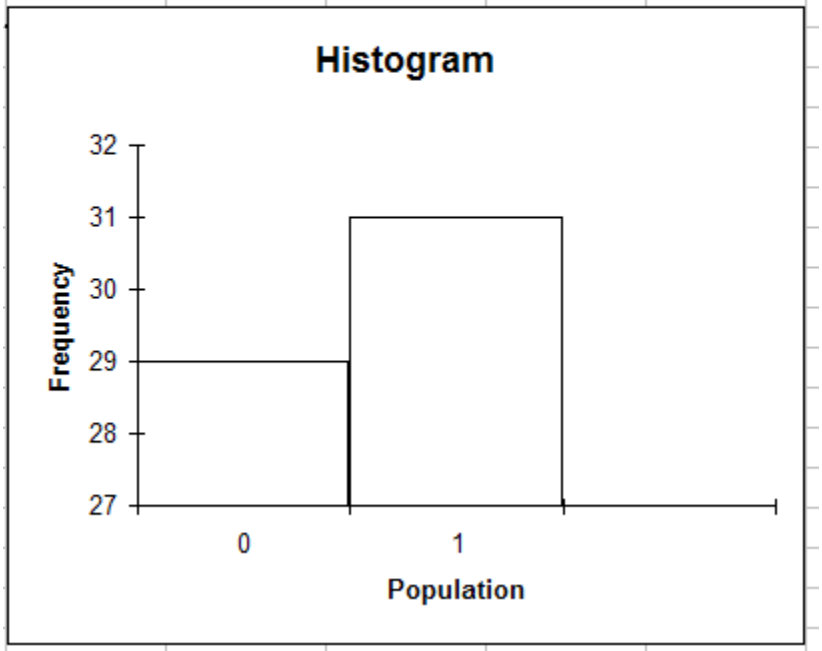
وفي نافذة الإدخال



	A	B	C	D	E	F	G
1	Population	Sample					
2	0	1					
3	1	0					
4		0					
5		0					
6		1					
7		0					
8		1					
9		1					
10		0					
11		0					
12		0					
13		0					
14		1					
15		0					
16		1					

فينتج

Population	*Frequency*
0	29
1	31

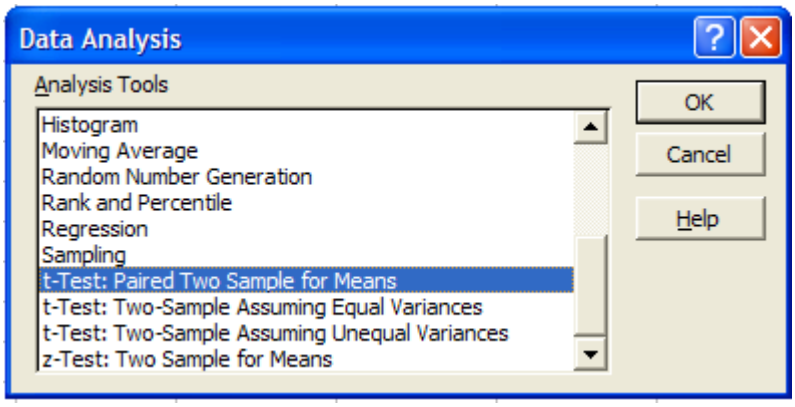


إختبار t للمتوسطات لعينتين متقارنة t-Test: Paired Two Sample for Means

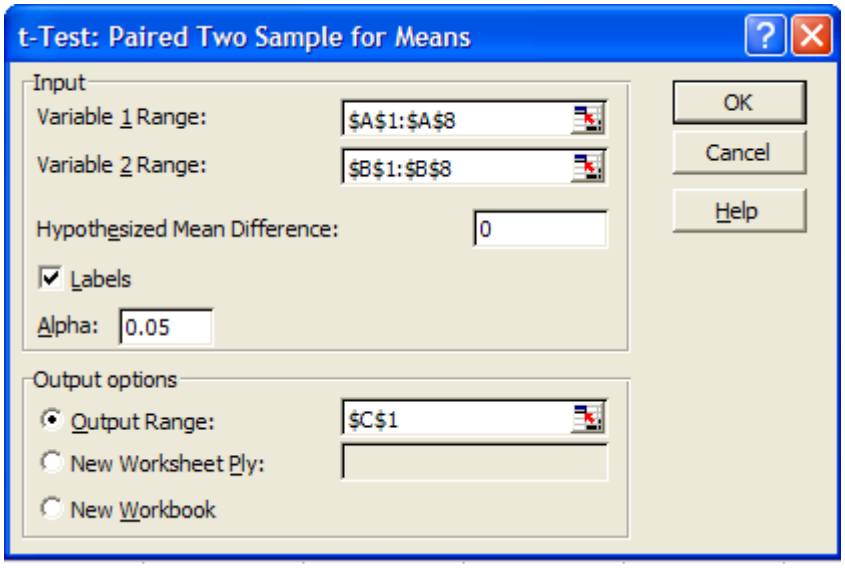
البيانات التالية هي درجات 7 طلاب في مادتين مختلفة

	A	B
1	X	Y
2	62	53
3	82	75
4	77	65
5	57	55
6	62	67
7	90	85
8	82	79

اختبر الفرض القائل انه لايوجد فرق بين متوسطي درجات المادتين عند مستوى معنوية 0.05



وندخل البيانات

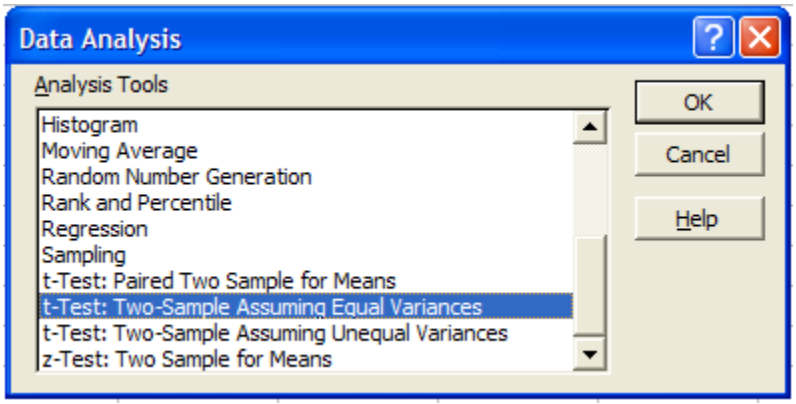


والنتائج

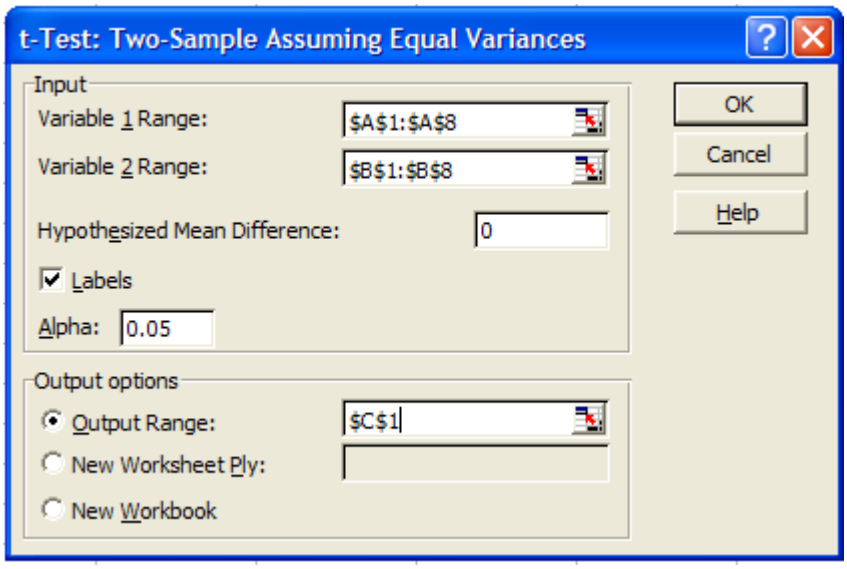
C	D	E
t-Test: Paired Two Sample for Means		
	X	*Y*
Mean	73.14285714	68.42857143
Variance	160.8095238	143.6190476
Observations	7	7
Pearson Correlation	0.902112013	
Hypothesized Mean Difference	0	
df	6	
t Stat	2.268233209	
P(T<=t) one-tail	0.031911178	
t Critical one-tail	1.943180905	
P(T<=t) two-tail	0.063822356	
t Critical two-tail	2.446913641	

إختبار t لعينتين على إفتراض تساوى التباين t-Test: Two-Sample Assuming Equal Variances

سوف نقوم بهذا الإختبار للمثال السابق



تدخل البيانات المطلوبة

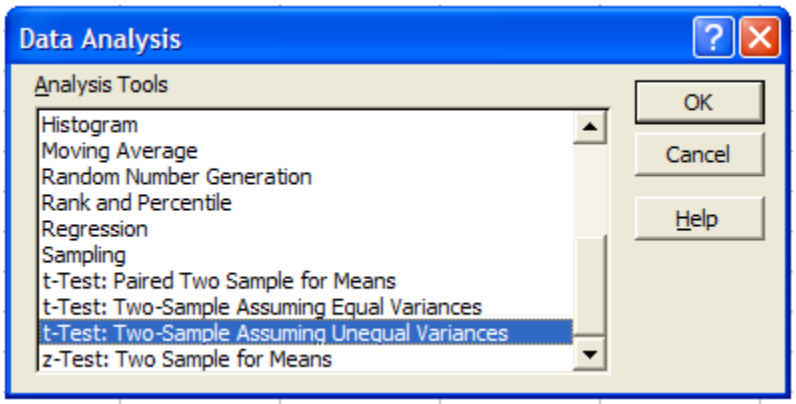


وينتج

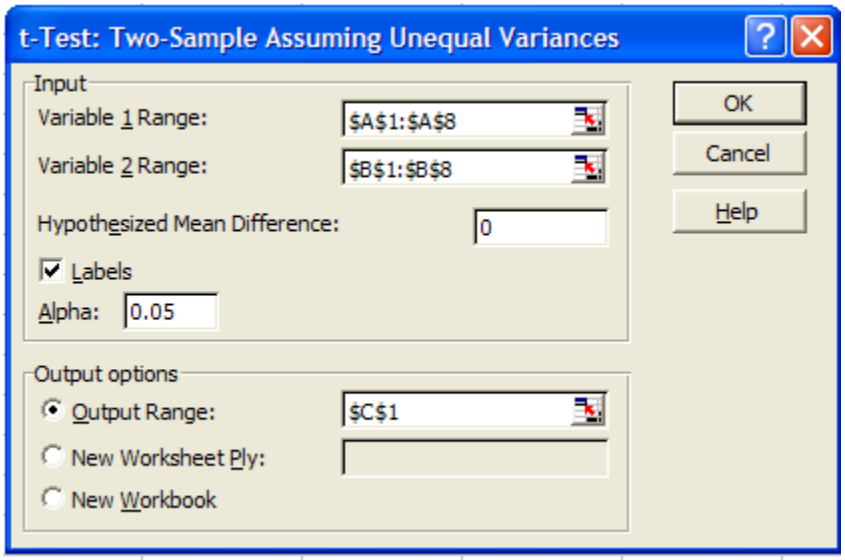
C	D	E
t-Test: Two-Sample Assuming Equal Variances		
	X	*Y*
Mean	73.14285714	68.42857143
Variance	160.8095238	143.6190476
Observations	7	7
Pooled Variance	152.2142857	
Hypothesized Mean Difference	0	
df	12	
t Stat	0.714862005	
P(T<=t) one-tail	0.244184537	
t Critical one-tail	1.782286745	
P(T<=t) two-tail	0.488369074	
t Critical two-tail	2.178812792	

إختبار t لعينتين على إفتراض عدم تساوى التباين t-Test: Two-Sample Assuming Unequal Variances

سوف نقوم بهذا الإختبار للمثال السابق



تدخل البيانات المطلوبة

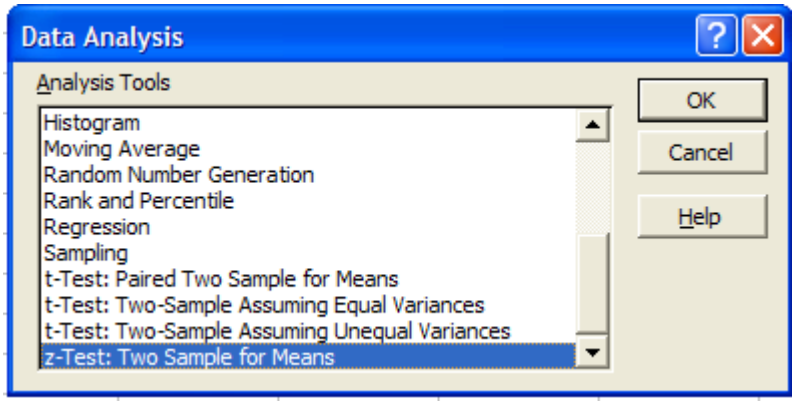


وينتج

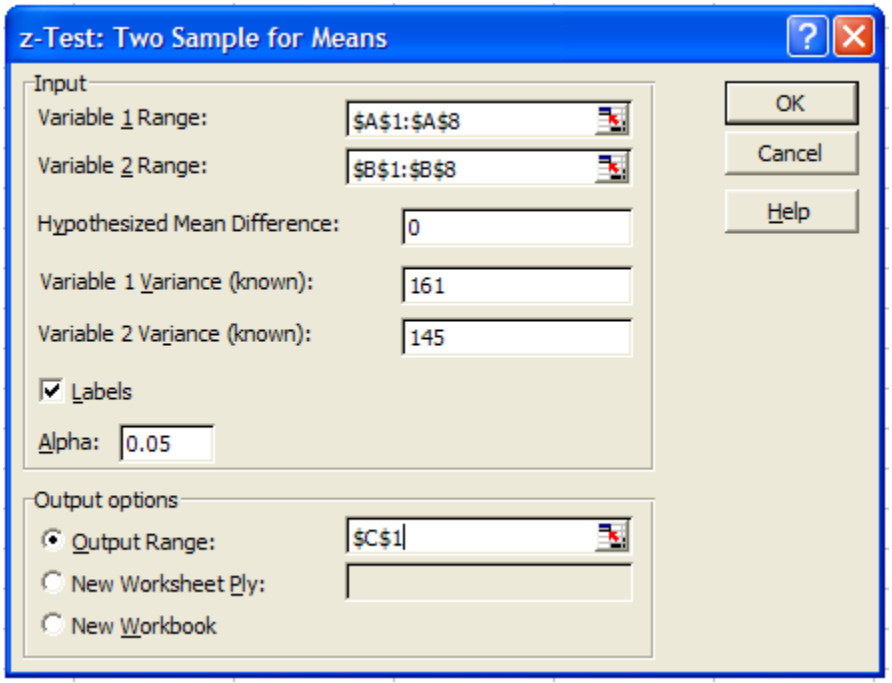
C	D	E
t-Test: Two-Sample Assuming Unequal Variances		
	X	*Y*
Mean	73.14285714	68.42857143
Variance	160.8095238	143.6190476
Observations	7	7
Hypothesized Mean Difference	0	
df	12	
t Stat	0.714862005	
P(T<=t) one-tail	0.244184537	
t Critical one-tail	1.782286745	
P(T<=t) two-tail	0.488369074	
t Critical two-tail	2.178812792	

إختبار z للمتوسطات لعينتين z-Test: Two-Sample for Means

سوف نستعرض هذا الإختبار على المثال السابق



ندخل البيانات المطلوبة



ويكون الناتج

C	D	E
z-Test: Two Sample for Means		
	X	*Y*
Mean	73.14285714	68.42857143
Known Variance	161	145
Observations	7	7
Hypothesized Mean Difference	0	
z	0.713024096	
P(Z<=z) one-tail	0.237915349	
z Critical one-tail	1.644853476	
P(Z<=z) two-tail	0.475830698	
z Critical two-tail	1.959962787	

ويترك للطالب المقارنة بين الإختبارات السابقة.

ملحق:

هذا الملحق يحوي أقل معلومات يجب على طالب الإحصاء الإلمام بها عند تخرجه.

تداريج او مستويات القياس Scales of Measurement

لكي نستعرض ونفسر ونحلل البيانات لابد أن نميز بين تداريج أو مستويات قياس المتغيرات.

يوجد أربع تصانيف أساسية لمستويات قياس المتغيرات

1- **التدريج أو التصنيف الإسمي Nominal Scale** ويصف المتغيرات النوعية أو الفئوية Categorical or Qualitative Variables فمثلا المتغير جنس gender يصنف ذكر او انثى والمتغير جنسية يصنف سعودي او اماراتي او عماني او مصري او جزائري الخ. في بعض الحزم الإحصائية لابد من إعادة ترميز المتغيرات النوعية حتى يمكن تحليلها فمثلا الذكر يرمز له بالرقم 1 والانثى الرقم 2 فقط لاحظ ان ارقام الترميز هذه لا تعنى اي ترتيب.

2- **التدريج او التصنيف الترتيبي Ordinal Scale** ويصف ايضا المتغيرات النوعية أو الفئوية وهذا التدريج يوسع المعلومات عن التدريج الإسمي ويعطيه ترتيب فمثلا المتغير المستوى الدراسي يمكن ان يصنف اول او ثاني او ثالث وهذا يعطي وصفا وترتيب ايضا المتغير مستوى الرضى عن خدمة يمكن ان يوصف غير راضي او راضي بعض الشيئ او راضي الخ.

3- **التدريج او التصنيف الفتري Interval Scale** ويصنف المتغيرات الكمية Quantitative Variables فبالإضافة الى الترتيب بين قيم المتغير هناك معنى لكمية الفرق بين القياسات. التصنيف الفتري لايوجد له نقطة صفر حقيقية فمثلا درجة حرارة 40 مئوية لاتعني ان الحرارة مرتين اعلى من 20 درجة مئوية لأن الصفر في التدريج المئوي إختياري حيث ان درجة الحرارة صفر لاتعنى عدم وجود حرارة.

4- **التدريج او التصنيف النسبي Ratio Scale** ويصنف المتغيرات الكمية ايضا و هو مثل التدريج الفتري إلا انه يمتلك صفر حقيقي فمثلا إذا اخذنا الوزن بوحدات الكيلوجرام فإن الفرق في الوزن بين شخصين وزن احدهم 82 كجم والآخر 69 كجم هو نفسه كالفرق في الوزن بين شخصين وزن احدهم 64 كجم و آخر وزنه 51 كجم أي ان الفرق 13 كجم وله نفس المعنى والتفسير ايضا شخص وزنه 100 كجم يزن ضعف شخص وزنه 50 كجم الوزن له صفر حقيقي.

طرق و انواع التحليل الإحصائي على المتغير يعتمد على تدريجه كالتالي:

1- المتغير الإسمي لايوجد معنى لمتوسطه او وسيطه ولكن يوجد لة منوال ونسبة.

2- المتغير الترتيبي لا يوجد معنى لمتوسطه ولكن يمكن ايجاد وسيطه ومنواله ونسبته.

3- المتغيرات الفتريه والنسبية يمكن ايجاد المتوسط الخ لها.

الربيعات والعشيرات والمئينات Quantiles, Deciles and Percentiles

إذا رتبت المشاهدات تصاعديا فإن المشاهدة التي يكون $n\%$ من المشاهدات أقل منها في القيمة تسمى المئين n ويرمز له P_n. المئينات 25 و 50 و 75 تعرف على أنها الربيع الأول و الربيع الثاني (او الوسيط) والربيع الثالث أي

$$Q_1 = P_{25}$$
$$Q_2 = P_{50} = median$$
$$Q_3 = P_{75}$$

المئينات 10 و 20 و ... و 90 تعرف على انها العشير الأول والعشير الثاني و ... والعشير التاسع أي

$$D_1 = P_{10}$$
$$D_2 = P_{20}$$
$$D_5 = P_{50} = median$$
$$D_9 = P_{90}$$

معامل الاختلاف Coefficient of Variation:

أو التشتت النسبي ويعرف كالتالي :

$$C.V. = \frac{s}{\bar{x}}$$

$$C.V. = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

المتغير المعياري والدرجات المعياري (مقياس التمركز):

إذا كان لدينا المتغير العشوائي X الذي له القيم الممكنة $x_1, x_2, \ldots, x_n$ والتي متوسطها $\bar{x}$ وانحرافها المعياري s فإن المتغير Z والذي له القيم $z_1, z_2, \ldots, z_n$ والتي تعطى بالعلاقة التالية :

$$z_i = \frac{x_i - \bar{x}}{s}, \quad i = 1, 2, \ldots, n$$

حيث z_i تقيس الانحرافات عن المتوسط الحسابي بوحدات من الانحراف المعياري يسمى بالمتغير المعياري ويستخدم للمقارنة بين التوزيعات المختلفة.

مثال

حصل طالب على 82 درجة في مقرر للإحصاء حيث كان متوسط الدرجات هو 75 درجة وانحراف معياري 10 درجات ثم حصل على 89 درجة في مقرر للرياضيات وكان متوسط الدرجات للرياضيات هو 81 درجة وانحراف معياري 16 درجة في أي من المقررين كانت درجة استيعاب هذا الطالب أعلى ؟

الحـــــل

إذا كانت z_1 ترمز للدرجة المعيارية للإحصاء فإن :

$$z_1 = \frac{82 - 75}{10} = 0.7$$

وإذا كانت z_2 ترمز للدرجة المعيارية للرياضيات فإن :

$$z_2 = \frac{89 - 81}{16} = 0.5$$

وهذا يعطي أن إستيعاب الطالب النسبي لمقرر الإحصاء أعلى من الرياضيات .

العزوم Moments :

للمتغير العشوائي X والذي له دالة توزيع $F_X(x)$ معرفة على جميع قيم

$-\infty < x < \infty$ يعرف العزوم الـ r حول الصفر

$$E(X^r) = \int_x x^r \, dF_X(x)$$

والعزم الـ r حول المتوسط

$$E\left((X-\mu)^r\right) = \int_x (x-\mu)^r \, dF_X(x)$$

حيث

$$\mu = E(X) = \int_x x \, dF_X(x)$$

العزوم من عينة $x_1, x_2, \ldots, x_n$

1- العزوم حول المتوسط

يعرف العزم النوني حول المتوسط بالعلاقة

$$m_r = \frac{\sum (x-\bar{x})^r}{n}$$

2- العزوم حول الصفر

$$m'_r = \frac{\sum x^r}{n}$$

مقياس الالتواء (ν)

$$\nu = \frac{m_3}{s^3}$$

$$m_3 = \frac{\sum (x - \bar{x})^3}{n}$$

$$s = \sqrt{\frac{1}{n-1}\sum (x-\bar{x})^2}$$

للتوزيع الطبيعي $\nu = 0$.

التفلطح k بالعلاقة التالية:

$$k = \frac{m_4}{s^4}$$

حيث :

$$m_4 = \frac{\sum (x - \bar{x})^4}{n}$$

للتوزيع الطبيعي $k = 3$.

نظرية النهاية المركزية Central limit Theorem

وتعتبر هذه النظرية أساسية في علم الإحصاء والتي تعطي التوزيع العيني للمتوسط وهي كالتالي :

إذا فرضنا أننا أخذنا عينة حجمها n من مجتمع له متوسط μ وانحراف معياري σ . فإن المتغير العشوائي $\frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$ له الخصائص التالية :

(1) يكون له (تقريباً) التوزيع الطبيعي القياسي إذا كانت $n \geq 30$ مهما كان توزيع المجتمع .

(2) يكون له (تماماً) التوزيع الطبيعي القياسي إذا كان توزيع المجتمع طبيعياً مهما كان حجم العينة .

ملاحظة

أ ـ إذا كانت $n \geq 30$ فإن $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

ب ـ إذا كان المجتمع طبيعيا فإن : $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

مهما كانت n .

تقدير معالم المجتمع واختبار الفروض Estimation and Testing Hypothese

التقدير بنقطة

يحتوي اي مجتمع إحصائي على معالم تكون غير معلومة مثل متوسطه μ أو انحرافه المعياري σ أو نسبة معينة R ..... الخ ويمكن إيجاد تقديرات لهذه المعالم من بيانات مأخوذة من عينة عشوائية من هذا المجتمع الإحصائي وذلك بحساب ما يسمى بالإحصاءات (وهي دوال في المشاهدات) فمثلاً متوسط العينة العشوائية $\bar{X}$ يستخدم كمقدر لمتوسط المجتمع μ وكذلك الانحراف المعياري s يستخدم كمقدر للانحراف المعياري للمجتمع σ وهكذا.... وتسمى هذه التقديرات التقدير بنقطة لأنها قيمة وحيدة محسوبة من العينة.

التقدير بفترة

التقدير بفترة لإحدى معالم المجتمع المجهولة مثل μ أو σ أو R هي عبارة عن إيجاد فترة تحدد بقيمتين تحسب من مشاهدات العينة العشوائية المأخوذة من المجتمع محل الدراسة ، ونتوقع احتواء هذه الفترة على معلمة المجتمع باحتمال معين ($1 - \alpha$) حيث عادة α تأخذ قيماً صغيرة مثل $0.1, 0.05, 0.01$ ويمكن أيضاً إيجاد دقة التقدير بفترة للمعلمة . وكلما كان طول الفترة صغيراً زادت دقة التقدير . لذلك سميت بتقدير فترة الثقة . فإذا كان مثلاً درجة الدقة في الخطأ بين متوسط المجتمع μ ومتوسط العينة العشوائية $\bar{X}$ هو المقدار الموجب ε فإنه يمكن تحديد حد أدنى للاحتمال يكتب كالتالي :

$$P\left(|\mu - \bar{x}| \le \varepsilon\right) \ge 1 - \alpha$$

ويمكن كتابة العلاقة السابقة كالتالي :

$$P\left(\bar{x} - \varepsilon \le \mu \le \bar{x} + \varepsilon\right) \ge 1 - \alpha$$

وهذه العلاقة تمثل فترة ثقة $\left(\bar{x} - \varepsilon\ ,\ \bar{x} + \varepsilon\right)$ للمعلمة المجهولة μ باحتمال لا يقل عن $\left(1 - \alpha\right)$. والمقدار $\left(1 - \alpha\right)$ يسمى درجة الثقة فإذا كانت قيم α هي $0.1, 0.05, 0.01$ فإن درجات الثقة المناظرة هي على الترتيب $99\ \%\ ,\ 95\ \%\ ,\ 90\ \%$.

وسوف نقوم باستعراض كيفية تقدير فترة الثقة لبعض المعالم المجهولة للمجتمع .

1 - تقدير فترة الثقة للمتوسط μ لمجتمع تباينه σ^2 معروف أو غير معروف للعينات الكبيرة .

2 - تقدير فترة الثقة للمتوسط μ لمجتمع ذات توزيع طبيعي تباينه σ^2 معروف أو غير معروف للعينات الصغيرة .

3 - تقدير فترة الثقة للنسبة R .

وقبل التعرض لدراسة كيفية حساب تقدير فترات الثقة للحالات السابقة . سوف نستعرض فيما يلي ما يسمى بالقيمة العظمى في خطأ التقدير ، والتي تساعدنا في إيجاد حجم العينة عند مستوى دقة معين α ، وكذلك في إيجاد فترات الثقة السابق ذكرها .

القيمة العظمى للخطأ في التقدير

لقد سبق دراسة توزيع المعاينة لمتوسط العينة $\bar{X}$ ووجد أن الخطأ المعياري له $\sigma_{\bar{x}}$ يساوي $\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$ وذلك عندما يكون حجم العينة n مأخوذاً من مجتمع صغير أو محدود حجمه N . أو يساوي $\frac{\sigma}{\sqrt{n}}$ في المجتمعات الكبيرة جداً أي عندما يكون حجم العينة n يمثل نسبة صغيرة جداً من المجتمع الذي حجمه N . ولقد سبق أيضاً دراسة نظرية النهاية المركزية . والتي تقول إن توزيع المعاينة للمتوسط يقترب من التوزيع الطبيعي وأنه يمكن أن يؤكد احتمال قدره $(1-\alpha)$ بأن متوسط العينة $\bar{X}$ يختلف عن متوسط المجتمع μ بمقدار يقل عن $z_{\alpha/2}$ من الخطأ المعياري $\sigma_{\bar{x}}$. ويمكن التعبير عما سبق بالعبارة الاحتمالية التالية :

$$P\left(\left|\bar{X}-\mu\right| \le z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)=1-\alpha$$

وذلك في حالة المجتمعات ذات الحجم الكبير أو الانهائية. والمقدار $\left|\bar{X}-\mu\right|$ يسمى خطأ التقدير ويرمز له بالرمز E وذلك عندما يزيد مقدار الخطأ في التقدير إلى نهايته العظمى. فمن المعادلة السابقة نجد أن خطأ التقدير يأخذ القيم التالية:

$$E=z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

أحياناً تسمى القيمة العظمى في خطأ التقدير بالدقة في التقدير.

حجم العينة

عند تحديد مقدار الدقة المطلوب أو القيمة العظمى للخطأ في التقدير عند احتمال معين $(1-\alpha)$ يمكن تحديد حجم العينة n من المعادلة السابقة ويكون كالتالي :

$$n=\left(\frac{z_{\alpha/2}\cdot\sigma}{E}\right)^2$$

تقدير فترة الثقة للمتوسط

في العينات الكبيرة

يمكننا أن نقول إنه باحتمال $(1-\alpha)$ أن :

$$|\bar{x}-\mu|\le z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

أي أن :

$$-z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\le \bar{x}-\mu\le z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

أي أن :

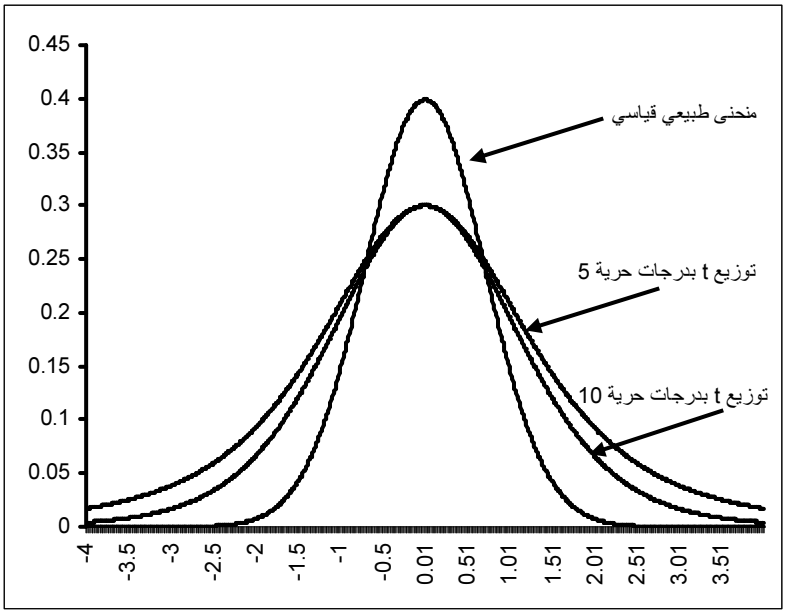
$$\bar{x}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\le \mu\le \bar{x}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

وهذا يعني بأن متوسط المجتمع μ واقع داخل الفترة الممتدة من الحد الأعلى $\bar{x}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ إلى الحد الأدنى $\bar{x}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ ويسمى هذا بتقدير فترة الثقة عند مستوى معنوي أو بدرجة ثقة قدرها $(1-\alpha)$ 100 . فإذا كانت $\alpha=0.05$ فإن درجة الثقة هي 95% . وإذا كانت $\alpha=0.01$ فإن درجة الثقة تكون 90% وهكذا .

في العينات الصغيرة

توزيع المعاينة للمتوسط يقترب من التوزيع الطبيعي عندما يكون الانحراف المعياري σ للمجتمع معلوماً أو غير معلوم للعينات الكبيرة أو معلوماً للعينات الصغيرة من

مجتمع ذا توزيع طبيعي ولكن عندما يكون الانحراف المعياري للمجتمع الطبيعي σ مجهولاً ونستعيض عنه بالانحراف المعياري المقدر من العينة s و توزيع المعاينة يكون في هذه الحالة توزيع t و هو توزيع متماثل ويختلف عن التوزيع الطبيعي . وتوزيع t يشبه التوزيع الطبيعي ويقترب من التوزيع الطبيعي القياسي كلما كبر حجم العينة n أو زادت درجات الحرية ν حيث $\nu = n - 1$ وينطبق على التوزيع الطبيعي عندما يصبح $\nu = 30$. ونوضح شكل توزيع t عند درجات حرية 5 , 10 ومنحنى التوزيع الطبيعي القياسي بالشكل التالي :



شكل يمثل توزيع t مقارنة بالتوزيع الطبيعي

ويمكن كتابة فترة الثقة للعينات الصغيرة باحتمال قدره $(1-\alpha)$ لمتوسط المجتمع μ مثل ما تم بالنسبة للعينات الكبيرة كالتالي :

$$\bar{x} - t_{\left(\frac{\alpha}{2},\nu\right)}\left(\frac{s}{\sqrt{n}}\right) < \mu < \bar{x} + t_{\left(\frac{\alpha}{2},\nu\right)}\left(\frac{s}{\sqrt{n}}\right)$$

حيث استبدلنا $z_{\alpha/2}$ بالقيمة $t_{\alpha/2}$. واستبدلنا الانحراف المعياري σ للمجتمع بالانحراف المعياري للعينة s . و $t_{\alpha/2}$ تعطى من الحزم الإحصائية حيث تعطى قيم t لكل درجة من درجات الحرية $\nu = n$-1 وذلك لبعض قيم α مثل 0.1, 0.05, 0.01,

التقدير بإستخدام الحزم الإحصائية:

R

```
> height = read.table("clipboard",sep=" ",header=TRUE)
> height
    height
1       56
2       70
.........
99      63
100     73
> n = length(height)
> a = mean(height)
> b = sd(height)
> err = qt(0.975, df = n-1)*b/sqrt(n)
> lft = a - err
> rt = a + err
> lft
  height
57.21374
> rt
  height
61.22626
```

```
> t.test(height)

        One Sample t-test

data:  height
t = 58.5694, df = 99, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 57.21374 61.22626
sample estimates:
mean of x
    59.22
```

Excel

المخرجات:

height	
Mean	59.22
Standard Error	1.011108003
Median	59
Mode	56
Standard Deviation	10.11108003
Sample Variance	102.2339394
Kurtosis	-0.701168047
Skewness	-0.126404815
Range	46
Minimum	34
Maximum	80
Sum	5922
Count	100
Largest(75)	52
Smallest(25)	51
Confidence Level(95.0%)	2.006257588

اختبارات الفروض الإحصائية Test of Statistical Hypothesis

الفرع الثاني من الاستدلال الإحصائي هو اختبارات الفروض . يحاول الباحث في كثير من الأحيان اتخاذ قرار بشأن خواص توزيع مجتمع ما ، وذلك بناءً على بيانات عينة عشوائية اختيرت من المجتمع نفسه . فمثلاً يريد طبيب أن يختبر فعالية دواء جديد بالنسبة لعلاج مرض معين . أو يريد باحث في التربية أن يختبر كون منهج معين أكثر فائدة من منهج آخر .... الخ . وهكذا وللوصول إلى هذه القرارات الإحصائية (Statistical Decision) نقوم عادة بوضع فروض عن خواص المجتمع ، متوسطه ، انحرافه المعياري ... الخ . ونختبر هذا الفرض بناء على عينة عشوائية نختارها من المجتمع ، وهذه الفروض ـ التي قد تكون صحيحة ، أو غير صحيحة ـ تسمى بالفروض الإحصائية وتنقسم الفروض إلى قسمين :

i) فروض عن معالم المجتمع (Parametric) .

ii) فروض عن صورة دالة التوزيع (Nonparametric) .

وسوف نكتفي بدراسة الفروض عن بعض معالم المجتمع فقط .

ونبدأ بافتراض إحصائي يسمى **فرض العدم (Null Hypothesis)** (وسمي بهذا لأن الغرض منه هو عدم قبوله أو محوه) ويرمز له بـ H_0 ، ويسمى الافتراض الذي يختلف عن H_0 **بفرض البديل** ويرمز له بـ H_1 **(Alternative)** .

وكما سبق أن وضحنا بأن اختبار الفروض يعتمد على بيانات العينة وهي تعني بالإجابة على السؤال التالي : هل بيانات العينة متناسقة مع فرض معين ؟ وهل تميل إلى تأكيده أو نفيه ؟

فإذا فرضنا قيمة لمعلم من معالم المجتمع (مثلاً المتوسط) وأخذنا عينة عشوائية من المجتمع فسيكون هناك غالباً فرق راجع إلى مجرد الصدفة والخطأ المتوقع نتيجة العينة ؟ أم أنه فرق حقيقي معنوي (Significant) ، وأن هناك أسباباً أخرى ساعدت على كبر

هذا الفرق ؟ فإذا وجدنا أن الفرق معنوي فإننا نميل إلى أن الفرض H_0 قد يكون غير صحيح (أو على الأقل نرفضه بناء على البيانات المتاحة) . وإذا كان الفرق غير معنوي (Nonsignificant) فليس هناك ما يدعونا لرفض فرض العدم H_0 .

فمثلاً إذا قذفنا قطعة نقود 20 مرة وكانت النتيجة المشاهدة هي 16 للصورة و 4 للكتابة ، فإننا قطعاً نميل إلى رفض H_0 بأن العملة متزنة ، على الرغم من أننا قد نكون مخطئين في اتخاذ هذا القرار الإحصائي . ولكن إن كانت النتيجة مثلاً 11 صورة و 9 كتابة فليس هناك ما يبرر رفض H_0 .

وكمثال آخر لفرض العدم (الفرضية الأولية) . إذا افترض أحد الباحثين بأن أطوال الذكور البالغين في إحدى البلاد هو 165 سم واختار عينة مكونة من 64 ذكراً بالغاً . وحسب متوسط الطول لهم فكان 160 سم ، وكان الانحراف المعياري لهذا المجتمع معروفاً ويساوي 5 سم . فيكون فرض العدم في هذه الحالة $H_0 : \mu = 165$ cm (أو الفرضية الأولية H_0 أن متوسط المجتمع الذي سحبت منه العينة هو 165 سم) . حيث نفترض عدم وجود فروق حقيقية بين متوسط المجتمع والقيمة المفروضة . والفروق المشاهدة إنما تعزى للصدفة ويقابل الفرضية الأولية فرضية أخرى تسمى بالفرضية البديلة ونرمز لها بالرمز H_1 ، في حالة متوسط أطوال الذكور البالغين يمكن أن يأخذ الفرض البديل إحدى الحالات التالية :

$$\mu \neq 165 \quad \text{أو} \quad \mu < 165 \quad \text{أو} \quad \mu > 165 \quad \text{أو} \quad \mu = 170 \quad \ldots \text{الخ}$$

اختبارات المعنوية

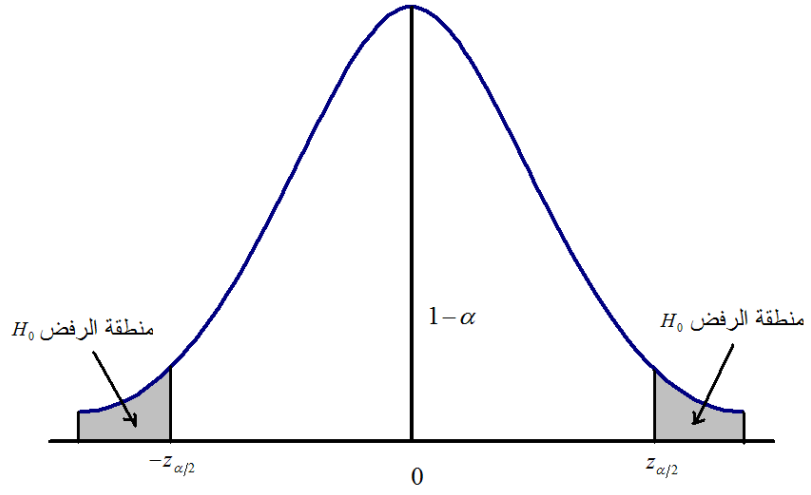
ولاختبار صحة الفرضية الأولية H_0 يجب علينا تكوين إحصاءة وهي دالة من مشاهدات العينة العشوائية ومعالم H_0 مثل $z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$ ، وعادة يكون توزيع الإحصاءة معروفاً ويقسم المجال المقابل لهذه الإحصاءة إلى منطقتين ، المنطقة الأولى يمكن تسميتها منطقة عدم الرفض وهي التي يكون فيها احتمال حدوث قيم الإحصاءة والذي

هو $1-\alpha$ كبيراً عندما تكون الفرضية الأولية H_0 صحيحة . والمنطقة الثانية تسمى بمنطقة الرفض وهي التي يكون احتمال حدوث قيم الإحصاءة والذي هو α صغيراً عندما تكون الفرضية الأولية صحيحة . وفي حالة متوسط أطوال الذكور البالغين نأخذ الإحصائية $z_0 = \frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}}$ حيث سبق معرفة توزيعها بأنه يقترب من التوزيع الطبيعي القياسي $N(\ 0,\ 1\ )$ ويمكن توضيح منطقة عدم الرفض ومنطقة الرفض للفرضية $H_0\ :\ \mu\ =\ 160$ على أساس صحة الفرضية H_0 مع الأخذ في الاعتبار الفرضية البديلة H_1 في الحالات التالية :

$H_0 : \mu = 160$

$H_1 : \mu \neq 160$

في هذه الحالة الفرض البديل H_1 له طرفان كما هو موضح بالمنطقة المظللة في الرسم التالي .



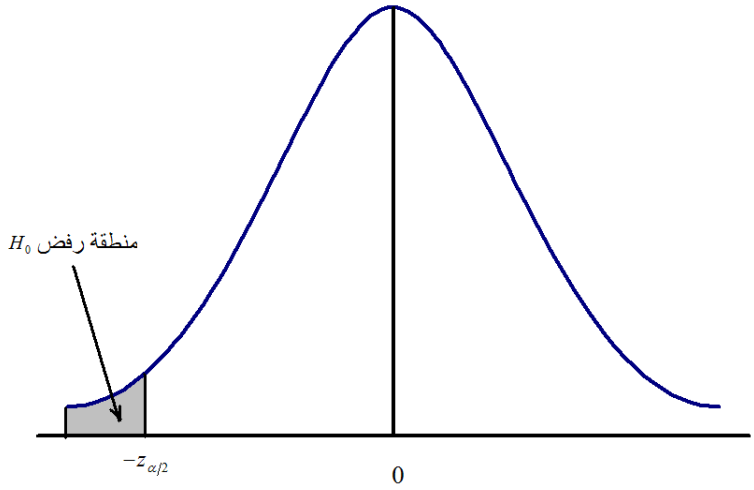
شكل يمثل منطقة الرفض للفرض H_0 من الطرفين

منطقة عدم الرفض H_0 هي $-z_{\alpha/2} \le z_0 \le z_{\alpha/2}$ وهي غير مظللة في شكل السابق .

$H_0 : \mu = 160$

$H_1 : \mu < 160$

الفرض البديل H_1 له طرف واحد أدنى ونوضح ذلك بالشكل التالي :



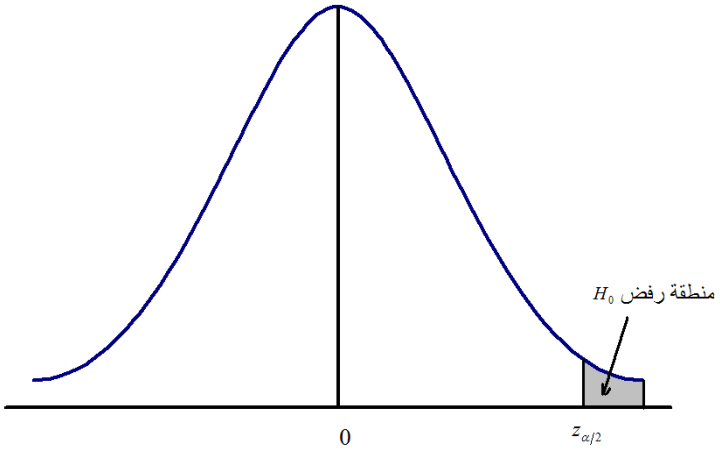
شكل يمثل منطقة الرفض للفرض H_0 من الطرف الأيسر

منطقة عدم رفض H_0 هي $z_0 > z_\alpha$.

$H_0 : \mu = 160$

$H_1 : \mu > 160$

الفرض البديل له طرف واحد من أعلى ونوضح ذلك بالشكل التالي :



شكل يمثل منطقة الرفض للفرض H_0 من الطرف الأيمن

منطقة عدم رفض H_0 هي $z_0 < z_\alpha$.

وعموماً إن كانت الفرضية الأولية $H_0 : \mu = \mu_0$ فإنه يمكن تلخيص مناطق الرفض وعدم الرفض للفرضية الأولية وذلك حسب الفرض البديل H_1 . وذلك لقيم الإحصاءة z_0 المحسوبة من العينة كالتالي :

H_1الفرض البديل	$\mu \neq \mu_0$	$\mu < \mu_0$	$\mu > \mu_0$
منطقة الرفض لـ H_0	$z_0 < -z_{\alpha/2}$ أو $z_0 < z_{\alpha/2}$	$z_0 < -z_\alpha$	$z_0 < z_\alpha$
منطقة عدم الرفض لـH_0	$-z_{\alpha/2} < z_0 < z_{\alpha/2}$	$z_0 < -z_\alpha$	$z_0 < z_\alpha$

الخطأ من النوع الأول α والخطأ من النوع الثاني β

أي قرار إحصائي يترتب عليه نوعان من الخطأ . خطأ من النوع الأول : هو عندما نرفض الفرضية الأولية H_0 وهي صحيحة ، ويكون احتمال هذا الخطأ هو قيمة α ، وهي عادة ما تكون صغيرة ومثال على ذلك لقيم α 0.1, 0.05, 0.01, وأحياناً تسمى α مستوى المعنوية . وعندما لانرفض الفرضية الأولية H_0 وهي خاطئة نكون قد ارتكبنا خطأ من النوع الثاني ، ويكون احتمال هذا الخطأ هو القيمة β . ويمكن تلخيص ذلك في الجدول التالي :

القرار / الفرضية H_0	عدم رفض H_0	رفض H_0
الفرضية H_0 صحيحة	قرار صحيح	خطأ من النوع الأول α
الفرضية H_0 خطأ	خطأ من النوع الثاني β	قرار صحيح

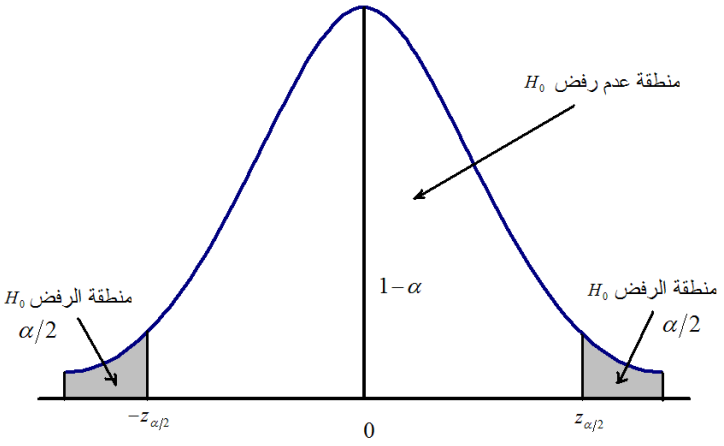
اختبار الفرضيات للمتوسط μ

عندما يكون **الانحراف المعياري σ للمجتمع معلوماً** فإن الإحصائية $\frac{\bar{x}-\mu}{\sigma_{\bar{x}}}$ يكون لها توزيع يقترب من التوزيع الطبيعي القياسي $N(0, 1)$ وبذلك يمكننا من تكوين الفرضية الأولية (فرض العدم) H_0 والفرضية البديلة H_1 كما يلي :

$H_0 : \mu = \mu_0$

$H_1 : \mu \neq \mu_0$

حيث الفرض البديل له طرفان كما هو موضح بالرسم التالي :

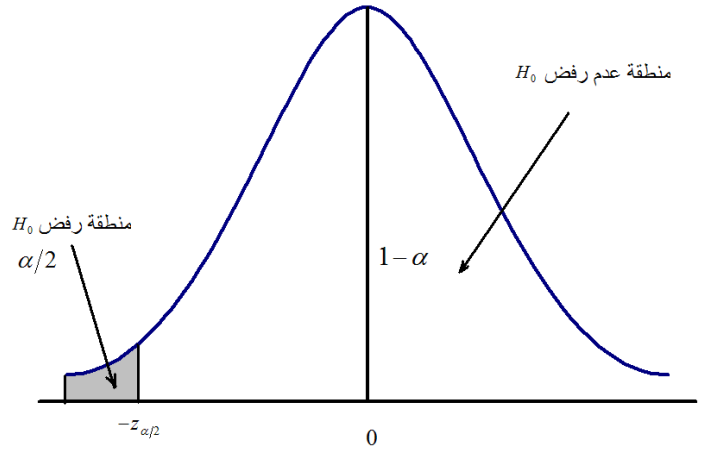


شكل يمثل منطقة الرفض وعدم الرفض للفرض H_0

$H_0 : \mu = \mu_0$

$H_1 : \mu < \mu_0$

حيث الفرض البديل له طرف واحد كما هو موضح بالرسم التالي:



شكل يمثل منطقة الرفض للفرض H_0 للطرف الأدنى

ثم تكون الإحصائية $\left(z_0 = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \right)$ باستخدام البيانات المشاهدة من العينة

ثم نختار مستوى المعنوية α الذي على أساسه نحدد القيم الحرجة $z_{\alpha/2}$, $-z_{\alpha/2}$ عندما تكون الفرضية البديلة H_1 من طرفين ونوضح ذلك بالمثال التالي :

درجات الحرية

إن كان لدينا مجتمع ما ويوجد به عدد من المعالم . ونرغب في تقدير هذه المعالم من عينة تحتوي على n من البيانات المستقلة . فإن درجة الحرية التي يرمز لها بالرمز ν تساوي عدد البيانات المستقلة للعينة مطروحاً منها عدد المعالم للمجتمع التي يجب تقديرها من العينة ويعبر عن ذلك رياضياً كما يلي :

$$\nu = n - k$$

حيث k تساوي عدد المعالم التي يجب تقديرها . فإن كان المطلوب تقدير المتوسط μ بـ $\overline{X}$ من العينة فإن $k = 1$ وبالتالي ($\nu = n - 1$) وإن كان المطلوب تقدير كل من μ و σ فإن $k = 2$ أي أن درجات الحرية في هذه الحالة ($\nu = n - 2$) وهكذا .....

إختبارات t :

إختبار t يستخدم لتثمين الفرق بين متوسطين لمجموعتين مستقلتين من البيانات قد يكون أحد المتوسطين مفترض نظريا . فمثلا يمكن إختبار نتائج فحص مجموعتين من المرضى اعطي لمجموعة دواء حقيقي و للاخرى دواء مزيف اونتائج فحص مجموعة

من المرضى مع نتيجة او معيار سابق. ويستخدم إختبار t على عينة مسحوبة من توزيع طبيعي غير معلوم تباينه وعلى اساس ان العينات مسحوبة من مجتمعين لهم نفس التباين.

<u>إختبار t للعينات المستقلة:</u>

ويستخدم لإختبار الفرضية الصفرية لتساوي متوسطات مجتمعين أي

$$H_0 : \mu_1 = \mu_2$$

عند توفر عينات مستقلة من كل من المجتمعين. كما أن المتغير الذي نقارنه يفترض ان له توزيع طبيعي وله نفس الإنحراف المعياري في كلا المجتمعين.

إحصائة الإختبار هي

$$t = \frac{\overline{x}_1 - \overline{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

حيث $\overline{x}_1$ و $\overline{x}_2$ متوسطات العينات و n_1 و n_2 احجام العينات و s الإنحراف المعياري المجمع والذي يحسب من العلاقة

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

حيث s_1 و s_2 الإنحرافات المعيارية المحسوبة من العينات.

تحت الفرضية الصفرية الإحصائة t لها توزيع t بدرجات حرية $n_1 + n_2 - 2$.

فترة الثقة الناتجة من إختبار عند مستوى معنوية α يعطى بالعلاقة

$$(\overline{x}_1 - \overline{x}_2) \pm t_\alpha s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

حيث t_α هي القيمة الحرجة لإختبار بذيلين بدرجات حرية $n_1 + n_2 - 2$.

<u>إختبار t للعينات المتزاوجة:</u>

الإختبار المتزاوج *Paired t-test* يستخدم لإختبار متوسطات مجتمعين والتي يكون فيها كل فرد من المجتمع الأول متزاوج (مرتبط) بفرد من المجتمع الثاني فمثلا مقارنة أوزان لأطفال اوزانهم فوق العادة مع أخوانهم العاديين أو مقارنة حالة مريض قبل وبعد العلاج.

إذا رمزنا للمتغير الذي نهتم به بالرمز x فيكون x_1 و x_2 رموز للمتغير للقياسات الأولى و الثانية على التوالي ونرمز للزوج i بالرموز x_{1i} و x_{2i} ويكون الفرق $d_i = x_{1i} - x_{2i}$ و الذي نفترض ان له توزيع طبيعي. الفرضية الصفرية هي

$$H_0 : \mu_d = 0$$

و إحصائية الإختبار

$$t = \frac{\bar{d}}{s_d / \sqrt{n}}$$

حيث $\bar{d}$ متوسط الفروق d_i و s_d الإنحراف المعياري للفروق d_i و n عدد الأزواج. تحت الفرضية الصفرية الإحصائة t لها توزيع t بدرجات حرية $n-1$. فترة ثقة $100(1-\alpha)\%$ تعطى بالعلاقة

$$\bar{d} \pm t_\alpha \, s_d / \sqrt{n}$$

حيث t_α هي القيمة الحرجة لإختبار بذيلين بدرجات حرية $n-1$.

<u>مثال:</u>

R

```
> male = c(13.3,19,20,8,18,22,20,31,21,12,16,12,24)
> female = c(22,26,16,12,21.7,23.2,21,28,30,23)
> t.test(male,female)

        Welch Two Sample t-test
```

```
data:  male and female
t = -1.7336, df = 20.539, p-value = 0.09798
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 -9.0538774  0.8277235
sample estimates:
mean of x mean of y
 18.17692  22.29000

>
```

Excel

المخرجات:

m	f
13.3	22
19	26
20	16
8	12
18	21.7
22	23.2
20	21
31	28
21	30
12	23
16	
12	
24	

t-Test: Two-Sample Assuming Unequal Variances

	m	*f*
Mean	18.17692308	22.29
		28.2987777
Variance	36.39025641	8
Observations	13	10
Hypothesized Mean Difference	0	
df	21	
	-	
t Stat	1.733589466	
P(T<=t) one-tail	0.048825134	
t Critical one-tail	1.720742871	
P(T<=t) two-tail	0.097650268	
t Critical two-tail	2.079613837	

إختبارات مربع كاى Chi Square Tests:

إختبارات حسن المطابقة Goodness of Fit Test:

لنفترض اننا اخذنا عينة من 100 طفل تمت ولادتهم في مستشفى الجامعة فوجدنا 66 طفلا من نوع الذكور و 34 طفلا من نوع الإناث و المفترض نظريا ان يكون عدد الذكور وعدد الإناث متساويا فهل هذا الإختلاف عائد للمعاينة و الصدفة أم أن هناك إختلاف بيئي و إجتماعي.

إذا فرضيتنا هي ان نسبة عدد الذكور لعدد الإناث هي 1:1 أي في عينة من 100 طفل نتوقع أن يكون عدد الذكور 50 وعدد الإناث 50 ولكي نختبر هذا نكون الإحصائة

$$\chi_0^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}$$

حيث O_i العدد المشاهد من النوع $i = 1, 2$ (1 ذكر و 2 انثى) و E_i العدد المتوقع.

للمثال السابق

$$\chi_0^2 = \frac{(66-50)^2}{50} + \frac{(34-50)^2}{50} = 5.12 + 5.12$$
$$= 10.24$$

هذه الإحصائة لهذا توزيع كاي تربيع بدرجة حرية 1 تحت الفرضية الصفرية. القيمة الناتجة لها $p\text{-}value = 0.00137$ وهي تدل على رفض الفرضية الصفرية ان نسبة عدد الذكور لعدد الإناث هي 1:1 .

أو في رمي زهرة نرد متزنة 60 مرة ونتج التالي:

Draw	1	2	3	4	5	6
Number Observed	10	10	8	11	12	9
Number Expected	10	10	10	10	10	10

تحت الفرضية الصفرية أن الزهرة متوازنه نحسب الإحصائة

$$\chi_0^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

حيث O_i العدد المشاهد من النوع $i = 1, 2,\ldots,n$ و E_i العدد المتوقع.

$$\chi_0^2 = \frac{(10-10)^2}{10} + \frac{(10-10)^2}{10} + \cdots + \frac{(9-10)^2}{10} = 1$$

ولها درجات حرية 5 = 1 - 6 ونجد قيمة $p\text{-}value = 0.962566$ أي لانرفض الفرضية الصفرية.

كمثال لحسن مطابقة بيانات على توزيع معين لنفترض أن أحد الباحثين تحصل على البيانات التالية والتي يعتقد انها موزعة توزيع بواسون بمتوسط 4.

2	2	3	6	4	3	5	6	2	5	6
4	4	3	4	1	7	2	4	3	0	2
2	5	3	5	0	5	7	4	3	3	8
2	4	5	4	5	5	4	5	4	2	5
3	11	7	5	4	7					

نكون توزيع تكراري للبيانات

value	0	1	2	3	4	5	6	7	8	9	10	11
obs	2	1	8	8	11	11	3	4	1	0	0	1
exp	0.9	3.66	7.3	9.77	9.77	7.8	5.21	2.98	1.49	0.66	0.26	0.1

تحسب القيم المتوقعة تحت الفرضية الصفرية ان البيانات تأتي من توزيع بواسون بمتوسط 4 كالتالي:

من العلاقة

$$E_i = n \times P(x) = n\frac{\lambda^x}{x!}e^{-\lambda}, x = 0,1,2,...$$

حيث n عدد المشاهدات و $P(x) = \frac{\lambda^x}{x!}e^{-\lambda}, x = 0,1,2,...$ دالة الكتلة لتوزيع بواسون

ونحسب القيم المتوقعة:

$$E_0 = 50 \times P(x=0) = 50\frac{4^0}{0!}e^{-4} = 0.915782$$

$$E_1 = 50 \times P(x=1) = 50\frac{4^1}{1!}e^{-4} = 3.663129$$

$$E_2 = 50 \times P(x=2) = 50\frac{4^2}{2!}e^{-4} = 7.326256$$

الخ

نحسب الإحصائة

$$\chi_0^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

$$\chi_0^2 = \frac{(2-0.9)^2}{0.9} + \frac{(1-3.66)^2}{3.66} + \cdots + \frac{(1-0.1)^2}{0.1} = 15.919$$

ولها 11 درجات حرية (عدد الخلايا - 1) القيمة الإحتمالية هي 0.144159 وهذه القيمة تدعوا لعدم رفض الفرضية الصفرية.

مثال على إختبار حسن التطابق:

R

```
> observed <- c(152,39,53,6)
> chisq.test(observed, p=c(9,3,3,1)/16) -> results
> results

        Chi-squared test for given probabilities

data:  observed
X-squared = 8.9724, df = 3, p-value = 0.02966

> results$expected
[1] 140.625  46.875  46.875  15.625
> results$residuals
[1]  0.9592242 -1.1502174  0.8946135 -2.4349538
>
```

Excel

	A	B	C	D	E	F
1	observed	ratios	expected	chi	chi-test =	=CHITEST(A2:A5,C2:C5)
2	152	9	=(B2/$B$7)*$A$7	=(A2-C2)*(A2-C2)/C2		
3	39	3	=(B3/$B$7)*$A$7	=(A3-C3)*(A3-C3)/C3		
4	53	3	=(B4/$B$7)*$A$7	=(A4-C4)*(A4-C4)/C4		
5	6	1	=(B5/$B$7)*$A$7	=(A5-C5)*(A5-C5)/C5		
6	sum			chi-sq	p-value	
7	=SUM(A2:A5)	=SUM(B2:B5)	=SUM(C2:C5)	=SUM(D2:D5)	=CHIDIST(D7,3)	
8						

	A	B	C	D	E	F
1	observed	ratios	expected	chi	chi-test =	0.029659504
2	152	9	140.625	0.920111111		
3	39	3	46.875	1.323		
4	53	3	46.875	0.800333333		
5	6	1	15.625	5.929		
6	sum			chi-sq	p-value	
7	250	16	250	8.972444444	0.029659504	
8						

جداول الرجحان و إختبار التوافق و الإقتران Contingency Tables

(Independence and Association)

ونشرحها بمثال: البيانات التالية اخذت من مجموعة من زوار معرض الكتاب لتفضيلهم للكتب الإلكترونية أو الورقية

	Education Level		
Book Type	High Scool	Bachelors	Master
eBook	23	35	42
Paper Book	45	30	25

الفرضية الصفرية: لا توجد علاقة بين مستوى التعليم ونوع الكتب المفضلة

لحساب القيم المتوقعة نكون الجدول التالي

	Education Level			Row Totals
Book Type	High Scool	Bachelors	Master	
eBook	23 (34.0) (3.559)	35 (32.5) (0.192)	42 (33.5) (2.157)	100
Paper Book	45 (34.0) (3.559)	30 (32.5) (0.192)	25 (33.5) (2.157)	100
Column Totals	68	65	67	200

القيمة المتوقعة للخلية = (مجموع السطر الذي تقع فيه الخلية X مجموع العمود الذي تقع فيه الخلية) \ المجموع الكلي

القيمة المتوقعة لخلية (مستوى التعليم ثانوي و كتاب إلكتروني) تحسب من

100 X 68 / 200 = 34.0

القيمة المتوقعة لخلية (مستوى التعليم جامعي و كتاب إلكتروني) تحسب من

100 X 65 / 200 = 32.5

وهكذا. نحسب إحصائة الإختبار من

$$\chi_0^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

نحسب قيمة مربع كاي للخلية الواحدة من $(O-E)^2/E$ فمثلا للخلية (مستوى التعليم ثانوي و كتاب إلكتروني) نجد O = 23 و E = 34.0 فيكون

كاي تربيع (مستوى التعليم ثانوي و كتاب إلكتروني) =

$$(23-34.0)^2/34.0 = 3.5588$$

وهكذا الخ. مربع كاي الكلي ينتج عن جمع مربعات كاي الجزئية $chi\text{-}sq = 11.816$

تحسب درجات الحرية $\nu = (n-1)\times(m-1)$ حيث m عدد الأعمدة و n عدد السطور ويكون $\nu = (2-1)\times(3-1) = 2$ القيمة الإحتمالية لقيمة $chi\text{-}sq =$ 11.816 بدرجات حرية 2 هي $p\text{-}value = 0.003$ أي ترفض الفرضية الصفرية.

مثال :

R

```
> x = c(50,88,155,379,18,63)
> y = c(43,62,110,300,14,144)
> chisq.test(x,y)

        Pearson's Chi-squared test

data:  x and y
X-squared = 30, df = 25, p-value = 0.2243

Warning message:
In chisq.test(x, y) : Chi-squared approximation may
be incorrect
>
```

مثال آخر في R

```
> rm(list=ls(all=TRUE))
> library(MASS)
> head(survey)
     Sex Wr.Hnd NW.Hnd W.Hnd    Fold Pulse    Clap Exer Smoke Height
1 Female   18.5   18.0 Right  R on L    92    Left Some Never 173.00
2   Male   19.5   20.5  Left  R on L   104    Left None Regul 177.80
3   Male   18.0   13.3 Right  L on R    87 Neither None Occas     NA
4   Male   18.8   18.9 Right  R on L    NA Neither None Never 160.00
5   Male   20.0   20.0 Right Neither    35   Right Some Never 165.00
6 Female   18.0   17.7 Right  L on R    64   Right Some Never 172.72
       M.I    Age
1   Metric 18.250
2 Imperial 17.583
3     <NA> 16.917
4   Metric 20.333
5   Metric 23.667
6 Imperial 21.000
> tbl = table(survey$Smoke, survey$Exer)
> tbl

        Freq None Some
  Heavy    7    1    3
  Never   87   18   84
  Occas   12    3    4
  Regul    9    1    7
> chisq.test(tbl)

        Pearson's Chi-squared test

data:  tbl
```

```
X-squared = 5.4885, df = 6, p-value = 0.4828

Warning message:
In chisq.test(tbl) : Chi-squared approximation may
be incorrect
> ctbl = cbind(tbl[,"Freq"], tbl[,"None"] +
tbl[,"Some"])
> ctbl
      [,1] [,2]
Heavy    7    4
Never   87  102
Occas   12    7
Regul    9    8
> chisq.test(ctbl)

        Pearson's Chi-squared test

data:  ctbl
X-squared = 3.2328, df = 3, p-value = 0.3571

>
```

Excel

	A	B	C	D
1	observed			
2	crm	Arsn	Rp	Vlnc
3	1	50	88	155
4	0	43	62	110
5	col totals	=SUM(B3:B4)	=SUM(C3:C4)	=SUM(D3:D4)
6	expected			
7	crm	Arsn	Rp	Vlnc
8	1	=$H3*B$5/$H$5	=$H3*C$5/$H$5	=$H3*D$5/$H$5
9	0	=$H4*B$5/$H$5	=$H4*C$5/$H$5	=$H4*D$5/$H$5
10		=SUM(B8:B9)	=SUM(C8:C9)	=SUM(D8:D9)
11	chi	=(B3-B8)*(B3-B8)/B8	=(C3-C8)*(C3-C8)/C8	=(D3-D8)*(D3-D8)/D8
12		=(B4-B9)*(B4-B9)/B9	=(C4-C9)*(C4-C9)/C9	=(D4-D9)*(D4-D9)/D9
13				
14	chi-sq	=SUM(B11:G12)		
15	p-value	=CHIDIST(B14,5)		

E	F	G	H
Stlng	Cnng	Frd	row total
379	18	63	=SUM(B3:G3)
300	14	144	=SUM(B4:G4)
=SUM(E3:E4)	=SUM(F3:F4)	=SUM(G3:G4)	=SUM(H3:H4)
Stlng	Cnng	Frd	
=$H3*E$5/$H$5	=$H3*F$5/$H$5	=$H3*G$5/$H$5	=SUM(B8:G8)
=$H4*E$5/$H$5	=$H4*F$5/$H$5	=$H4*G$5/$H$5	=SUM(B9:G9)
=SUM(E8:E9)	=SUM(F8:F9)	=SUM(G8:G9)	=SUM(H8:H9)
=(E3-E8)*(E3-E8)/E8	=(F3-F8)*(F3-F8)/F8	=(G3-G8)*(G3-G8)/G8	
=(E4-E9)*(E4-E9)/E9	=(F4-F9)*(F4-F9)/F9	=(G4-G9)*(G4-G9)/G9	

	A	B	C	D	E	F	G	H
1	observed							
2	crm	Arsn	Rp	Vlnc	Stlng	Cnng	Frd	row total
3	1	50	88	155	379	18	63	753
4	0	43	62	110	300	14	144	673
5	col totals	93	150	265	679	32	207	1426
6	expected							
7	crm	Arsn	Rp	Vlnc	Stlng	Cnng	Frd	
8	1	49.10869565	79.20757	139.9334	358.5463	16.89762	109.3065	753
9	0	43.89130435	70.79243	125.0666	320.4537	15.10238	97.69355	673
10		93	150	265	679	32	207	1426
11	chi	0.016176839	0.976002	1.622222	1.166808	0.071918	19.61721	
12		0.018099791	1.09202	1.815057	1.305507	0.080468	21.94912	
13								
14	chi-sq	49.73060762						
15	p-value	1.57332E-09						

الارتباط والانحدار Correlation and Regression

سوف نتناول دراسة البيانات التي يكون لأفرادها متغيران يتغيران معاً في وقت واحد ، وذلك لمعرفة نوع العلاقة التي تربط بينهما . والأمثلة كثيرة على هذا النوع من البيانات ، مثل دراسة العلاقة بين أوزان وأطوال مجموعة من الطلاب ، أو أعمار ودرجات مجموعة من الطلاب أو أجـور و إنتاج مجموعـة من العمـال ، أو الدخل والإنفاق لمجموعة من الأسر ، أو العلاقة بين صفة الطول للأب والابن ، أو صفة الذكاء للأب والابن وهكذا . ثم إيجاد مقاييس تقيس درجة هذه العلاقة .

سوف نتناول دراسة العلاقة بين المتغيرين (X, Y) ، فإذا كان هناك علاقة بين المتغير X والمتغير Y ، فكيف يمكن التعبير عنها بمعادلة رياضية ومنها يمكن التنبؤ بقيمة أحد المتغيرين إذا علمت قيمة المتغير الآخر .

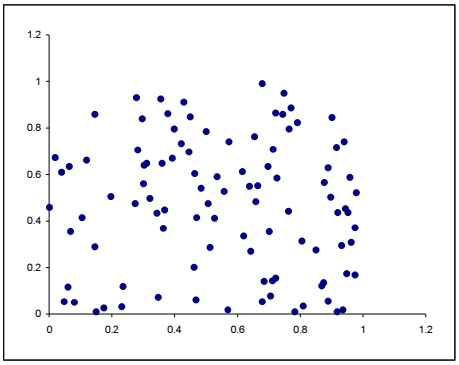
فإذا أردنا دراسة العلاقة بين الطول Y والوزن X لمجموعة عددها n من طلاب جامعة الملك سعود . فإنه يمكن لكل طالب قياس طوله ووزنه ويصبح لدينا أزواج من القراءات التالية :

$$(x_1, y_1), (x_2, y_2), \ldots . , (x_n, y_n)$$

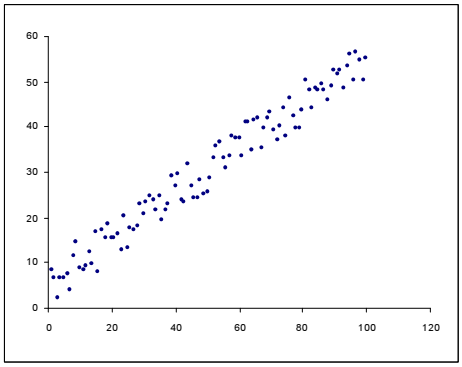
ولتمثيل هذه الأزواج من القراءات بيانياً نرسم محورين متعامدين . المحور الأفقي يمثل الوزن X مثلاً والمحور الرأسي ويمثل الطول Y ونقوم بتمثيل القراءات السابقة بنقاط فنحصل على ما يسمى بشكل الانتشار (*Scatter Diagram*) .

أشكال الانتشار

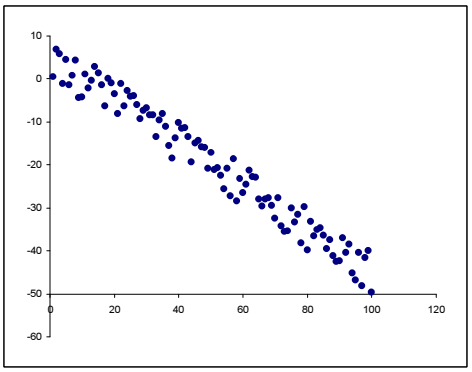
وأشكال الانتشار تأخذ صوراً مختلفة وذلك حسب طبيعة العلاقة بين المتغيرين (X, Y) تحت الدراسة . وفيما يلي نعرض بعض أشكال الانتشار (1) ، (2) ، (3) ، (4) التالية .



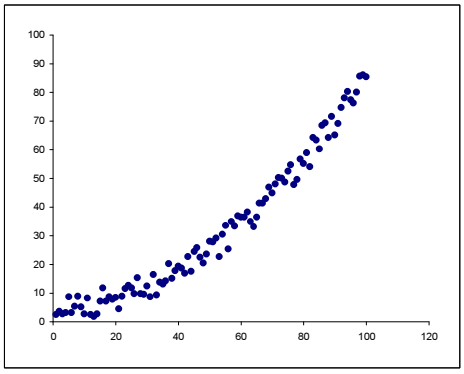
شكل(1)



شكل(2)



شكل(3)



شكل(4)

ونلاحظ من أشكال الانتشار السابقة ما يلي :

شكل الانتشار (1)

تكون فيه النقاط منتشرة بدون ترابط حول اتجاه محدد مما يدل على أنه لا توجد علاقة بين المتغيرين (X, Y) .

شكل الانتشار (2)

تكون فيه النقاط منتشرة حول خط مستقيم تزيد فيه قيم Y مع زيادة قيم X ، ونستنتج منه وجود علاقة خطية طريدة بين المتغيرين (X, Y) .

شكل الانتشار (3)

تكون فيه النقاط منتشرة حول خط مستقيم وفيه تنقص قيم Y مع زيادة قيم X ونستنتج منه وجود علاقة خطية عكسية بين المتغيرين (X, Y) .

شكل الانتشار (4)

تكون فيه النقاط منتشرة حول منحنى نستنتج منه وجود علاقة غير خطية بين المتغيرين (X, Y) .

وسوف نكتفي بإيجاد مقاييس تقيس قوة الارتباط بين المتغيرين (X, Y) في الحالة الخطية فقط وسندرس منها معامل الارتباط الخطي لبيرسون $(Person)$ ، وكذلك دراسة معادلة خط الانحدار للمتغير Y على المتغير X أو العكس للمتغير X على Y .

معامل الإرتباط الخطي لبيسرون Coefficient of Linear Correlation

ويستخدم معامل الارتباط الخطي لبيرسون لقياس التغير الذي يطرأ على المتغير Y عندما تتغير قيم X أو العكس . ويستخدم عادة في حالة البيانات الكمية .

معامل ارتباط بيرسون

إذا كان لدينا ازدواج المشاهدات التالية :

$$(x_1, y_1), (x_2, y_2), \dots , (x_n, y_n)$$

فإن معامل الارتباط r لبيرسون يعطي بمتوسط مجموعة حاصل ضرب القيم المعيارية للمتغيرين X', Y' ويكون كالتالي :

$$r = \frac{1}{n}\sum x'y'$$

حيث :

$$y' = \frac{y - \bar{y}}{\sigma_y}, \qquad x' = \frac{x - \bar{x}}{\sigma_x}$$

أي أن العلاقة السابقة يمكن كتابتها كالتالي :

$$r = \frac{1}{n}\frac{\sum\left(x - \bar{x}\right)\left(y - \bar{y}\right)}{\sigma_x \sigma_y}$$

ويفضل في حالة البيانات المأخوذة من عينة وضع العلاقة السابقة في الصورة التالية :

$$r = \frac{\sum\left(x - \bar{x}\right)\left(y - \bar{y}\right)}{\left(n-1\right)s_x s_y}$$

ويكون لمعامل الارتباط r الخصائص التالية .

خصائص معامل الارتباط الخطي لبيرسون

1) قيمته تساوي الصفرعندما تكون الظاهرتان مستقلتين تماماً ولا يوجد ارتباط بينهما ألبتة.

2) قيمته مقدار موجب عندما يكون الارتباط بين المتغيرين طردياً . ويكون قوياً عندما يكون المقدار الموجب قريباً من الواحد الصحيح وضعيفاً عندما يكون المقدار الموجب قريباً من الصفر .

3) قيمته مقدار سالب عندما يكون الارتباط بين المتغيرين عكسياً ، ويكون قوياً عندما تقترب قيمته من المقدار 1- وضعيفاً عندما تقترب من الصفر .

ملاحظة

يجب ملاحظة عدم ربط الارتباط بين متغيرين بالسببية أي أن التغير في أحد المتغيرين ليس بالضرورة يؤدي إلى التغير في المتغير الثاني (يتسبب فيه) فمثلاً إذا وجدنا ارتباطاً قوياً بين سلسلة من بيانات خاصة بأعداد مرضى السكري في سنوات متتالية وسلسلة بيانات الزيادة في دخل الوظائف العسكرية فهذا لايدل على أن هناك سبب بين الظاهرتين. أو إذا كان يوجد ارتباط عكسي قوي بين سلسلة مبيعات البطاطين في المملكة العربية السعودية في شهور سنة ما مع درجات الحرارة في انجلترا في نفس شهور السنة فلا يمكن تفسير ذلك بأن انخفاض درجة الحرارة في انجلترا يتسبب في زيادة مبيعات البطاطين في السعودية.

خـط الانحـدار Regression Line

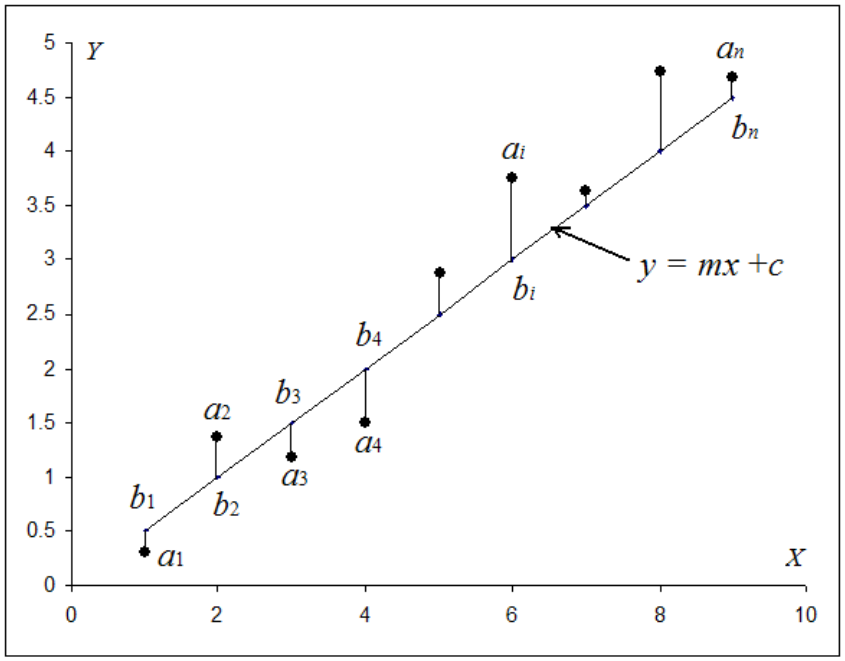
سبق لنا أن بيّنا أنه إذا كان لدينا متغيران (X , Y) أحدهما متغير مستقل X مثلاً والآخر متغير تابع وليكن Y مثلاً لظاهرتين محل الدراسة مثل العلاقة بين الطول X والوزن Y . أو العلاقة بين الإنفاق Y والدخل X . فإنه يمكن عمل أشكال الانتشار لهذه المتغيرات . وقد تم دراسة إيجاد معامل الارتباط بعدة طرق لقياس قوة الارتباط بين المتغيرين (X , Y) وذلك في الحالة التي يكون فيها شكل الانتشار في صورة خطية . وفيما يلي نبحث عن إيجاد معادلة رياضية تمثل أحسن توفيق لخط مستقيم يعبر عن البيانات في شكلها الخطي . وهذه العلاقة تسمى بمعادلة خط الانحدار . فإذا كان X متغيراً مستقلاً ، و Y متغيراً تابعاً ، فإن المعادلة التي نحصل عليها تسمى بمعادلة خط انحدار Y على X وهي على الصورة التالية :

$$y = mx + c + error$$

وإذا كان المتغير Y متغيراً مستقلاً والمتغير X متغيراً تابعاً فإن معادلة الخط المستقيم تسمى بمعادلة خط انحدار X على Y وتعطى بالعلاقة التالية :

$$x = m'y + c' + error$$

والغرض من إيجاد معادلة خط الانحدار هو التنبؤ بقيمة المتغير التابع لقيمة محددة من قيم المتغير المستقل . ويمكن الحصول على رسم خط الانحدار للبيانات من شكل الانتشار وذلك بالتمهيد باليد ، ولكن هذا التمهيد باليد يختلف من شخص إلى آخر ولذلك دعت الحاجة إلى إيجاد خط الانحدار بطريقة لا تعتمد على الرسم وإنما تعتمد على استخدام البيانات المعطاة للحصول على قيم الثوابت c , m في حالة خط انحدار Y على X وللحصول على m' , c' في حالة خط الانحدار X على Y وفي حالة انحدار Y على X يمكن إيجاد قيمة الثابتين c , m بطريقة المربعات الصغرى كما يلي :



شكل يبين انحرافات النقط التي تمثل المشاهدات عن خط الانحدار

نفرض أن لدينا مجموعة أزواج المشاهدات (x_1 , y_1) , ... , (x_n , y_n) ويرسم شكل الانتشار لهذه الأزواج نحصل على النقاط $a_1 , \ldots , a_n$ نفرض أن خط الانحدار $y = mx + c$ كما هو موضح بالرسم . وبرسم خطوط موازية لمحور Y مارة بالنقاط $a_1 , \ldots , a_n$ فإنها تقطع خط الانحدار في النقاط $b_1, \ldots , b_n$. نرمز للأطوال $a_1b_1 , \ldots , a_nb_n$ بالقيم $d_1, \ldots , d_n$ وهذه الأطوال عبارة عن انحرافات المشاهدات عن خط الانحدار والتي سميناها *error* في المعادلات السابقة . وقد تكون هذه الانحرافات سالبة أو موجبة حسب وضع النقط $a_1, \ldots , a_n$ بالنسبة لخط الانحدار .

لكي نحصل على أجود خط لابد أن يكون مجموع مربعات هذه الانحرافات او الأخطاء أقل ما يمكن أي أن :

$$d = \sum_{i=1}^{n} \left(mx_i + c - y_i \right)^2$$

أقل ما يمكن .

بتفاضل d بالنسبة إلى m , c ومساواة الناتج بالصفر نحصل على :

$$\sum y = m \sum x + x + nc$$
$$\sum xy = m \sum x^2 + c \sum x$$

وبحل هاتين المعادلتين نحصل على :

$$m = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2}$$

$$c = \frac{\sum y}{n} - m\frac{\sum x}{n}$$

وبالطريقة نفسها يمكن الحصول على m', c' :

$$m' = \frac{n\sum xy - \sum x \sum y}{n\sum y^2 - \left(\sum y\right)^2}$$

$$c' = \frac{\sum x}{n} - m' \ \frac{\sum y}{n}$$

سوف نطبق نموذج الإنحدار

$$\mathbf{Y} = \boldsymbol{\beta}\mathbf{X} + \boldsymbol{\varepsilon}$$

على البيانات

$$\mathbf{Y} = \begin{pmatrix} 136 \\ 144 \\ 145 \\ 169 \\ 176 \\ 195 \\ 211 \\ 224 \\ 231 \\ 256 \\ 281 \\ 312 \\ 347 \\ 377 \\ 423 \\ 477 \\ 553 \\ 613 \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 & 91 & 11 \\ 1 & 105 & 13 \\ 1 & 109 & 17 \\ 1 & 130 & 19 \\ 1 & 146 & 23 \\ 1 & 155 & 29 \\ 1 & 160 & 36 \\ 1 & 180 & 41 \\ 1 & 200 & 59 \\ 1 & 215 & 82 \\ 1 & 240 & 100 \\ 1 & 275 & 110 \\ 1 & 320 & 134 \\ 1 & 360 & 139 \\ 1 & 410 & 138 \\ 1 & 460 & 182 \\ 1 & 510 & 220 \\ 1 & 575 & 271 \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{18} \end{pmatrix}$$

بإستخدام R:

أولا: بالطريقة المطولة:

تقدر $\boldsymbol{\beta}$ من العلاقة (أنظر الورقة النظرية)

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}$$

أدخل التالي في R

```
X = matrix(c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
91,105,109,130,146,155,160,180,200,215,240,275,320,360,410,
460,510,575,11,13,17,19,23,29,36,41,59,82,100,110,134,139,
138,182,220,271),byrow=F,18,3)

Y = matrix(c(136,144,145,169,176,195,211,224,231,256,281,
312,347,377,423,477,553,613),byrow=F,18,1)

XX = solve(t(X)%*%X)
XY = t(X)%*%Y
beta = XX%*%XY
options(digits=4)
beta
        [,1]
[1,] 60.4599
[2,]  0.7563
[3,]  0.4135
```

أي

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} 60.4599 \\ 0.7563 \\ 0.4135 \end{pmatrix}$$

والنموذج المقدر

$$\hat{y}_i = 60.4599 + 0.7563x_{i1} + 0.4135x_{i2}\ , i = 1,2,...,18$$

تقدير الخطأ المعياري للبواقي:

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} \hat{\varepsilon}_i}{n-k}}$$

حيث $n = 18$ (حجم العينة) و $k = 3$ (عدد المعالم المقدرة) و

$$\hat{\varepsilon}_i = y_i - \hat{y}_i, i = 1,2,...,18$$

هي البواقي أو مقدرات للخطأ.

نحسب $\hat{y}_i, i = 1,2,...,18$ من

```
Ye = beta[1]*X[,1]+beta[2]*X[,2]+beta[3]*X[,3]
erro = Y-Ye
```

مجموع مربعات الأخطاء $\sum_{i=1}^{18} \hat{\varepsilon}_i^2$ يحسب من

```
sum(erro^2)
[1] 1084
```

ويكون الخطأ المعياري للبواقي

```
n = length(Y)
k = 3
sigma2 = sum(erro^2)/(n-k)
sigma = sqrt(sigma2)
sigma
[1] 8.503
```

إذا الخطأ المعياري للبواقي $\hat{\sigma} = 8.503$ بدرجات حرية 15

الخطأ المعياري لمقدرات المعالم:

الاخطاء المعيارية هي عناصر القطر لمصفوفة التباين - التغاير

$$V\left(\hat{\boldsymbol{\beta}}\right) = \hat{\sigma}^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

```
var.beta = sigma2*XX
var.beta
        [,1]      [,2]     [,3]
[1,] 67.6512 -0.665858  1.19755
[2,] -0.6659  0.007659 -0.01451
[3,]  1.1976 -0.014508  0.02819
```

قطر المصفوفة

```
diag(var.beta)
[1] 67.651189  0.007659  0.028188
```

والأخطاء المعيارية للمعالم

```
sqrt(diag(var.beta))
[1] 8.22503 0.08752 0.16789
```

ونضع النموذج على الشكل

$$\hat{y}_i = \underset{(8.225)}{60.4599} + \underset{(0.088)}{0.7563}\, x_{i1} + \underset{(0.168)}{0.4135}\, x_{i2} \;, i = 1,2,...,18$$

معامل التحديد:

معامل التحديد R^2 غير المعدل يعطى بالعلاقة

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}{\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2}$$

ويحسب

```
R2 <- 1-sum(erro^2)/sum((Y-mean(Y))^2)
R2
[1] 0.9969
```

لحساب معامل التحديد R_a^2 المعدل

$$R_a^2 = 1-\left(1-R^2\right)\frac{n-1}{n-k}$$

```
R22 <- 1-(1-R2)*(18-1)/(18-3)
R22
[1] 0.9965
```

إختبار فرضيات:

سوف نختبر هل هناك علاقة بين المتغيرات المستقلة والمتغير التابع كالآتي:

$$H_0: \quad \beta_1 = 0$$
$$\beta_2 = 0$$

وبوضعها في شكل مصفوفة

$$H_0 : \mathbf{RB} = \mathbf{r}$$

حيث:

$$\mathbf{R} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \mathbf{B} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}, \mathbf{r} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

ونحسب النسبة F

$$F = \frac{\left(\mathbf{R}\hat{\beta} - \mathbf{r}\right)' \left|\mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}\right|^{-1} \left(\mathbf{R}\hat{\beta} - \mathbf{r}\right)/q}{\frac{\sum_{i=1}^{n}\hat{\varepsilon}_i^2}{n-k}}$$

حيث q هو عدد عدد القيود على الفرضية الصفرية ويساوي هنا 2

```
R = matrix(c(0,1,0,0,0,1),2,3,byrow=T)
r = matrix(c(0,0),2,byrow=T)
q = 2
(t(R%*%beta-r)%*%solve(R%*%XX%*%t(R))%*%(R%*%beta-
r)/q)/(sum(erro^2)/(n-k))
```

```
     [,1]
[1,] 2450
qf(0.95,2,15)
[1] 3.682
```

وحيث أن $2450 \in [3.682, \infty)$ فإننا نرفض الفرضية الصفرية.

ثانيا بالطريقة المختصرة:

بإستخدام lm:

```
summary(lm(Y~X[,2]+X[,3]))

Call:
lm(formula = Y ~ X[, 2] + X[, 3])

Residuals:
   Min     1Q Median     3Q    Max
-13.22  -4.85  -1.60   4.58  15.83

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  60.4599     8.2250    7.35  2.4e-06 ***
X[, 2]        0.7563     0.0875    8.64  3.3e-07 ***
X[, 3]        0.4135     0.1679    2.46    0.026 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
’ 1

Residual standard error: 8.5 on 15 degrees of freedom
Multiple R-squared: 0.997,      Adjusted R-squared: 0.997
F-statistic: 2.45e+03 on 2 and 15 DF,  p-value: <2e-16
```

Excel

المخرجات:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.998473
R Square	0.996948331
Adjusted R Square	0.996541442
Standard Error	8.502626793
Observations	18

ANOVA

	df	*SS*	*MS*
Regression	2	354268.6912	177134.3456
Residual	15	1084.419936	72.29466238
Total	17	355353.1111	

	Coefficients	*Standard Error*	*t Stat*
Intercept	60.45994709	8.225034304	7.350722788
x1	0.75634722	0.087517909	8.642199411
x2	0.413493537	0.167893881	2.462826727

F	*Significance F*
2450.171835	1.36154E-19

P-value	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
2.40303E-06	42.92870154	77.99119264	42.92870154	77.99119264
3.28542E-07	0.569807214	0.942887226	0.569807214	0.942887226
0.026368095	0.055636203	0.771350871	0.055636203	0.771350871

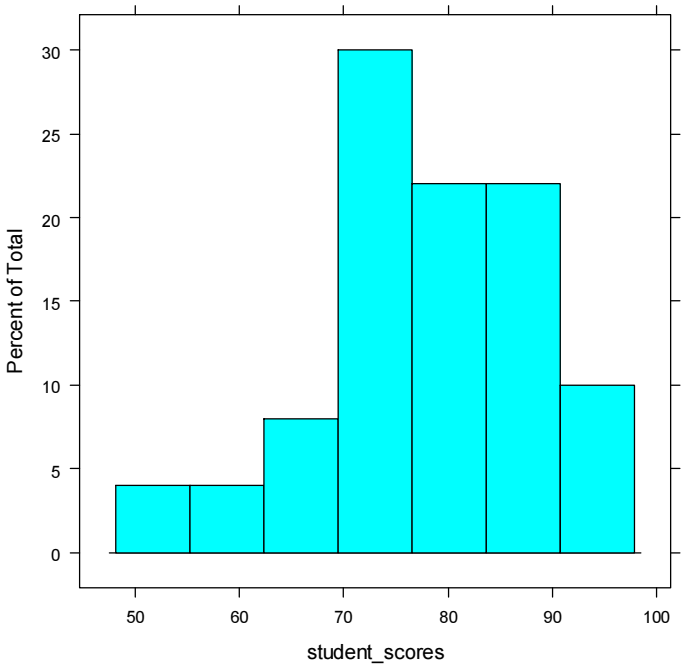
أمثلة متنوعة:

البيانات التالية تمثل درجات 50 طالباً في إحدى المواد :

51	95	70	74	73	90	71	74	90	67
91	72	83	89	50	80	72	84	85	69
62	82	87	76	91	76	87	75	78	79
71	96	81	88	64	82	73	57	86	70
80	81	75	85	74	90	83	66	77	91

```
> student_scores =
c(67,90,74,71,90,73,74,70,95,51,69,85,84,72,80,50,89,83,72,
91,79,78,75,87,76,91,76,87,82,62,70,86,57,73,82,64,88,81,96
,71,91,77,66,83,90,74,85,75,81,80)
> student_scores
[1] 67 90 74 71 90 73 74 70 95 51 69 85 84 72 80 50 89 83
72 91 79 78 75 87 76
[26] 91 76 87 82 62 70 86 57 73 82 64 88 81 96 71 91 77 66
83 90 74 85 75 81 80
>
> mean(student_scores)
[1] 77.86
> sd(student_scores)
[1] 10.51337

> library(mosaic)

> tally(student_scores)
```

```
> histogram(student_scores)

> t.test(student_scores, alternative = c("two.sided"),mu =
75,conf.level = 0.95)

        One Sample t-test

data:  student_scores
t = 1.9236, df = 49, p-value = 0.06023
alternative hypothesis: true mean is not equal to 75
95 percent confidence interval:
 74.87213 80.84787
sample estimates:
mean of x
    77.86

>
```



البيانات التالية هي درجات طلاب 328 إحص في إختبارين

Exam Data

Student	Exam 1	Exam 2
1	93	98
2	88	74
3	89	67
4	88	92
5	67	83
6	89	90
7	83	74
8	94	97
9	89	96
10	55	81
11	88	83
12	91	94
13	85	89
14	70	78
15	90	96
16	90	93
17	94	81
18	67	81
19	87	93
20	91	83

أوجد كل المعلوات الإحصائية الممكنة بإستخدام R و Python و Excel Data Analysis Tools (EDAT)

ملخص:

الحزم الإحصائية R و Python و EXCEL:

سوف نستعرض بعض الحزم الإحصائية لحساب بعض الإحصائيات البسيطة.

اولا: مقاييس النزعة المركزية و الإختلاف:

سوف نستخدم البيانات التالية:

```
x = 13.3,19,20,8,18,22,20,31,21,12,16,12,24
y = 22,26,16,16,16,21.7,23.2,21,28,30,23
```

المتوسط

1- حزمة R

```
> x = c(13.3,19,20,8,18,22,20,31,21,12,16,12,24)
> mean(x)
```

2- حزمة Python

```
>>> import numpy as np
>>> x = (13.3,19,20,8,18,22,20,31,21,12,16,12,24)
>>> np.mean(x)
```

3- صفحة النشر Excel

```
 x
 13.3
```

```
   19
   20
    8
   18
   22
   20
   31
   21
   12
   16
   12
   24
=AVERAGE(A2:A14)
```

الوسيط:

R

```
> x = c(13.3,19,20,8,18,22,20,31,21,12,16,12,24)
> median(x)
```

Python

```
>>> x = (13.3,19,20,8,18,22,20,31,21,12,16,12,24)
>>> np.median(x)
```

EXCEL

```
=MEDIAN(A2:A14)
```

X

	MEDIN	
13.3	=	19
19		
20		
8		
18		
22		
20		
31		
21		
12		
16		
12		
24		

الدالة SUMMARY في R:

```
> x = c(13.3,19,20,8,18,22,20,31,21,12,6,12,24)
> summary(x)
```

الدالة DESCRIBE في Python.pandas:

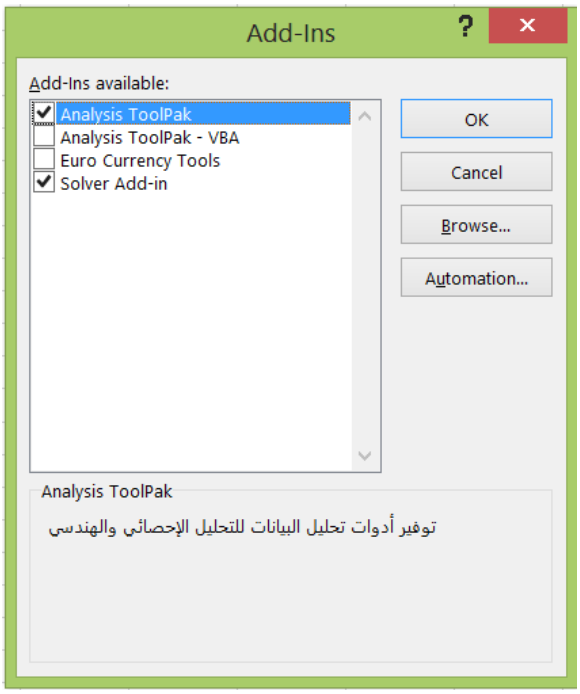
```
>>> import pandas as pd
>>> x = pd.DataFrame(x)
>>> x.describe()
```

أدوات تحليل البيانات DATA ANALYSIS TOOLS في EXCEL وتختصر DAT:

نفعل DAT من

Options -> Add-Ins -> Manage [Excel Add-ins] Go

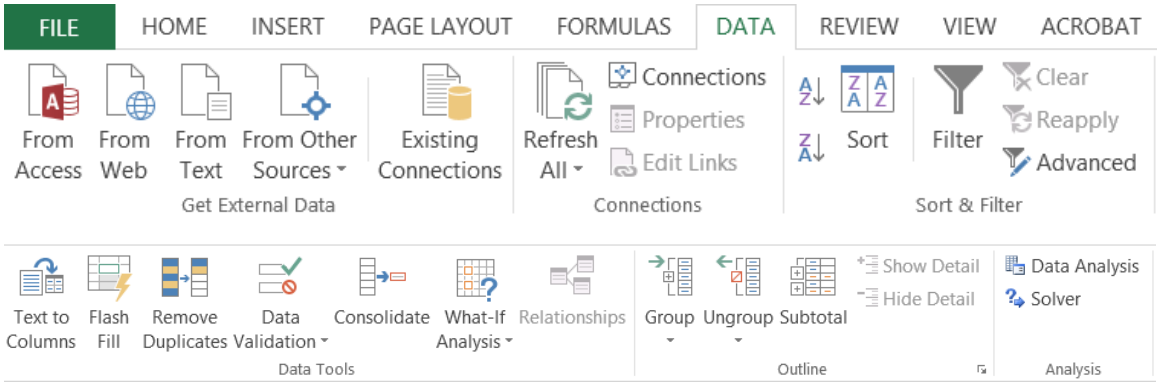
ثم أختار Analysis ToolPak

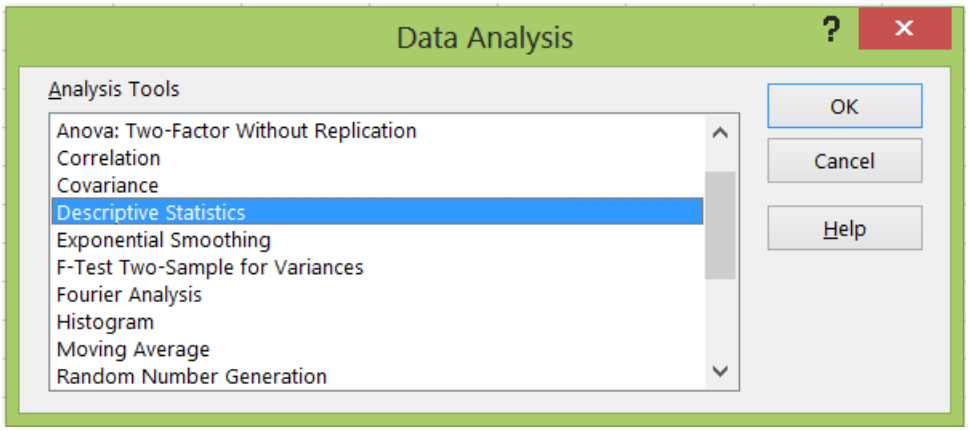


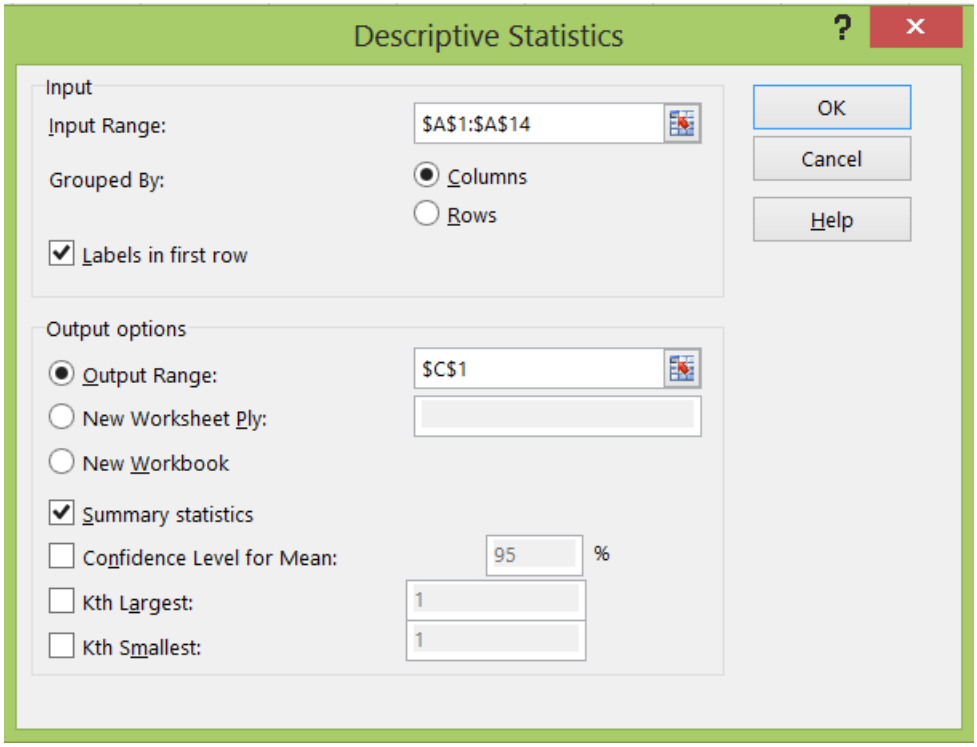
ثم

OK

من DATA أختار Data Analysis




Data Analysis
Analysis Tools
Anova: Two-Factor Without Replication
Correlation
Covariance
Descriptive Statistics
Exponential Smoothing
F-Test Two-Sample for Variances
Fourier Analysis
Histogram
Moving Average
Random Number Generation
OK
Cancel
Help


Descriptive Statistics
Input
Input Range:
$A$1:$A$14
Grouped By:
Columns
Rows
Labels in first row
Output options
Output Range:
$C$1
New Worksheet Ply:
New Workbook
Summary statistics
Confidence Level for Mean:
95
%
Kth Largest:
1
Kth Smallest:
1
OK
Cancel
Help

	A	B	C	D
1	X		*X*	
2	13.3			
3	19		Mean	18.17692308
4	20		Standard Error	1.67309608
5	8		Median	19
6	18		Mode	20
7	22		Standard Deviation	6.032433705
8	20		Sample Variance	36.39025641
9	31		Kurtosis	0.519983741
10	21		Skewness	0.334378727
11	12		Range	23
12	16		Minimum	8
13	12		Maximum	31
14	24		Sum	236.3
15			Count	13

الإحصائآت الوصفية:

Summary Statistics

1- R:

```
max(x), min(x), sum(x), mean(x), median(x),
range(x),var(x), sort(x), rank(x), order(x),
quatile(x), cumsum(x),cumprod(x), etc ...
```

Example:

```
> x = c(13.3,19,20,8,18,22,20,31,21,12,16,12,24)
> max(x)
> min(x)
> sum(x)
> mean(x)
> median(x)
> range(x)
> var(x)
> sort(x)
> rank(x)
> order(x)
> quatile(x)
> cumsum(x)
> cumprod(x)
```

2- Python

```
>>> x.max()
>>> x.min()
>>> x.sum()
>>> x.mean()
>>> x.median()
>>> x.var()
>>> x.std()
>>> x.skew()
>>> x.kurt()
>>> x.quantile()
```

3- Excel

```
=AVERAGE(A2:A14)
=COUNT(A2:A14)
=KURT(A2:A14)
=LARGE(A2:A14,2)
=MAX(A2:A14)
=MEDIAN(A2:A14)
=MIN(A2:A14)
=MODE.MULT(A2:A14)
=MODE.SNGL(A2:A14)
=PERCENTILE.EXC(A2:A14,0.5)
=QUARTILE.EXC(A2:A14,1)
```

```
=SKEW(A2:A14)
=SMALL(A2:A14,3)
=STDEV.P(A2:A14)
=VAR.P(A2:A14)
```

تحليل البيانات الفئوية Categorical Data Analysis

كما ذكرنا سابقا البيانات الفئوية تقاس على المستويين الإسمي و الترتيبي و أحد طرق تحليلها يكون عن طريق جداول الإحتمالات Contingency Tables والتي تقيس ما يسمى معامل الإقتران و معامل التوافق والتي يطلق عليها Measures of Association وسوف نستعرضها في الأمثلة التالية:

R

```
> auto = read.csv("http://www-
bcf.usc.edu/~gareth/ISL/Auto.csv")
> head(auto)
> y = auto$weight
> z = auto$origin
> table(z,y)
> table(y,z)
> b =auto$cylinders
> table(b)
> table(b,z)
> table(b,y)
>
```

Python

```
>>> auto = pd.read_csv("http://www-
bcf.usc.edu/~gareth/ISL/Auto.csv")
>>> auto.head()
>>> y = auto.weight
>>> z = auto.origin
>>> pd.crosstab(y,z)
>>> pd.crosstab(z,y)
>>> b = auto.cylinders
>>> pd.crosstab(b,z)
```

Excel

تستخدم في إكسل جداول المحور او الركيزة Pivot Tables لهذا الغرض. لشرح وافر عن جداول الركيزة انظر كتاب " مبادئ الإحصاء و الإحتمالات مع حل الأمثلة بإستخدام ميكروسوفت إكسل" تأليف "عدنان ماجد بري" و "محمود محمد هندي" صفحة 69.