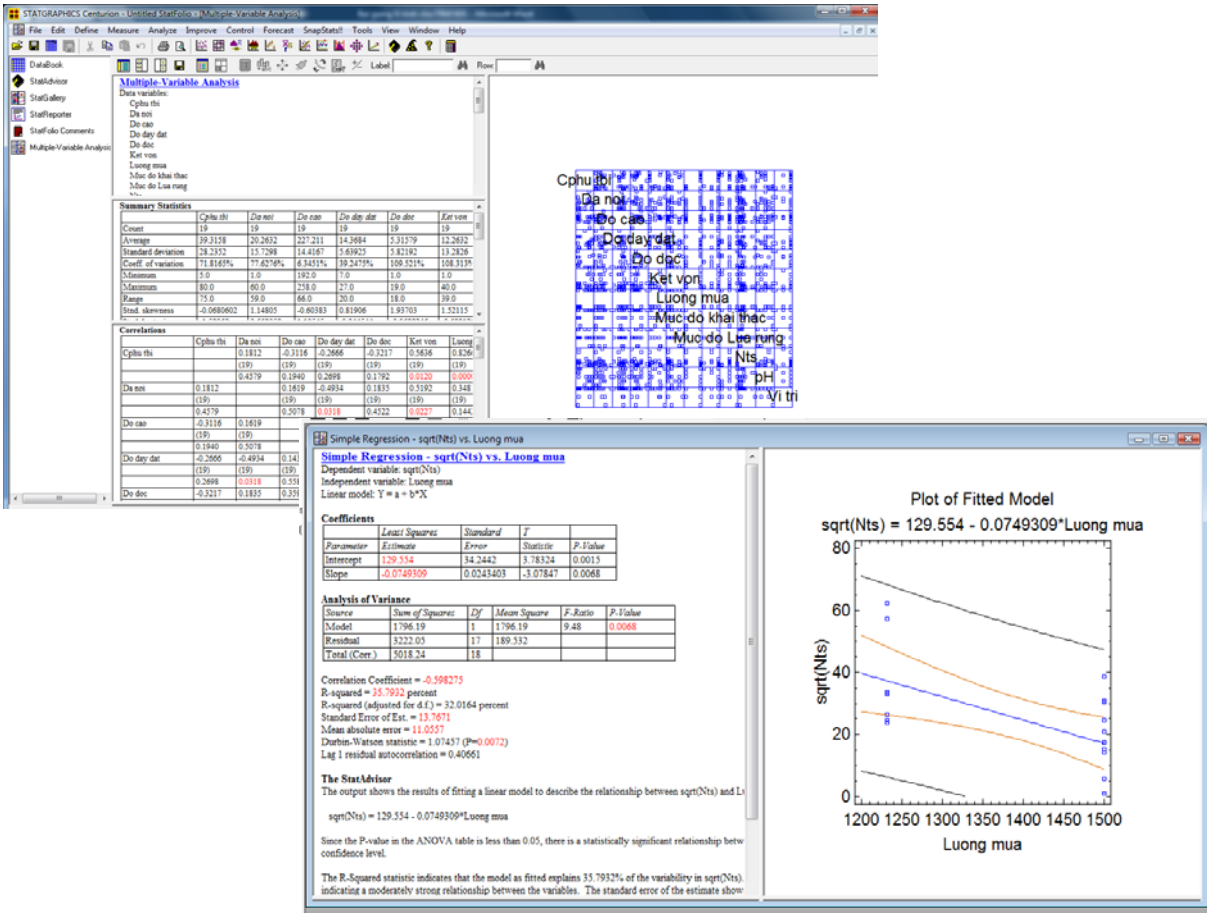
TRƯỜNG ĐẠI HỌC TÂY NGUYÊN
KHOA NÔNG LÂM NGHIỆP

PGS.TS. BẢO HUY

TIN HỌC THỐNG KÊ TRONG QUẢN LÝ TÀI NGUYÊN THIÊN NHIÊN

Xử lý thống kê bằng phần mềm Statgraphics Centurion XV và MS. Excel 2007



Tháng 5 năm 2009

Mục lục

1. TỔNG QUÁT VỀ CHỨC NĂNG XỬ LÝ THỐNG KÊ CỦA MS.EXCEL 2007 VÀ STATGRAPHICS CENTURION XV 7

1.1. Tổng quát về phần xử lý thống kê trong MS. Excel 7
1.2. Tổng quát về phần mềm xử lý thống kê Statgraphics Centurion 8

2. THỐNG KÊ MÔ TẢ 10

3. SẮP XẾP VÀ VẼ BIỂU ĐỒ PHÂN BỐ TẦN SỐ XUẤT HIỆN THEO CẤP, CỠ, HẠNG 12

4. SO SÁNH 1 – 2 MẪU QUAN SÁT BẰNG TIÊU CHUẨN T 14

4.1. So sánh một mẫu với một giá trị cho trước – Kiểm tra T một mẫu 14
4.2. So sánh sự sai khác giữa trung bình 2 mẫu – Kiểm tra T 2 mẫu 16

5. PHÂN TÍCH PHƯƠNG SAI 19

5.1. Phân tích phương sai 1 nhân tố với các thí nghiệm ngẫu nhiên hoàn toàn 19
5.2. Phân tích phương sai 2 nhân tố 22
5.1.1. Phân tích phương sai 2 nhân tố với 1 lần lặp lại: (Bố trí thí nghiệm theo khối ngẫu nhiên đầy đủ (Randomized Complete Blocks) (RCB): 22
5.1.2. Phân tích phương sai 2 nhân tố m lần lặp 28

6. PHÂN TÍCH TƯƠNG QUAN - HỒI QUY 32

6.1. Hồi quy tuyến tính 1 lớp 32
6.2. Dạng phi tuyến đưa về tuyến tính 1 lớp 34
6.2.1. Lập mô hình hàm mũ trong Excel: 34
6.2.2. Lập mô hình hàm mũ một lớp trong Statgraphics: 36
6.3. Ước lượng các dạng hồi quy một lớp tuyến tính hoặc phi tuyến tính trên đồ thị 40
6.4. Hồi quy tuyến tính nhiều lớp 45
6.5. Hồi quy phi tuyến tính nhiều lớp, tổ hợp biến 47

7. MÔ HÌNH HOÁ QUY LUẬT PHÂN BỐ 57

7.1. Mô hình hoá phân bố giảm theo hàm Meyer 57
7.2. Mô phỏng phân bố thực nghiệm theo phân bố khoảng cách-hình học: 60
7.3. Mô phỏng phân bố thực nghiệm theo phân bố Weibull: 62

LỜI NÓI ĐẦU

Trong quản lý tài nguyên thiên nhiên, ứng dụng công nghệ tin học đóng vai trò quan trọng trong phân tích, quản lý cơ sở dữ liệu; trong đó ứng dụng tin học trong xử lý thống kê được áp dụng ngày càng rộng rải. Thông qua xử lý thống kê trên các phần mềm, giúp chúng ta hệ thống hóa cơ sở dữ liệu, đánh giá các thí nghiệm, phân tích các mối quan hệ phức tạp trong tự nhiên và với các nhân tố xã hội để tìm ra quy luật nhằm quản lý bền vững. Xử lý thống kê thông qua công nghệ tin học ngày nay đã phát triển một bước dài, nó giúp cho con người rút ngắn được thời gian tính toán, xử lý được một lượng lớn thông tin và có được những hiểu biết một cách khách quan các quy luật tự nhiên và xã hội. Do đó thành tựu của công nghệ xử lý thống kê tin học cần được ứng dụng một cách rộng rải hơn trong quan lý tài nguyên thiên nhiên.

Có rất nhiều phần mềm ứng dụng để xử lý thống kê như SPSS, Statgraphics, Excel....
Microsoft Excel được mọi người biết đến khi nói đến công cụ bảng tính, tính tóan..., nhưng những chức năng chuyên sâu về ứng dụng thống kê trong sinh học, quản lý tài nguyên thiên nhiên, môi trường lại ít được đề cập đến. Trong khi đó chức năng xử lý thống kê của phần mềm Excel là hết sức phong phú và mạnh để ứng dụng trong các thí nghiệm, phân tích, đánh giá các kết quả nghiên cứu, điều tra khảo sát về lâm nghiệp, quản lý tài nguyên thiên nhiên. Trong đó bao gồm các xử lý thống kế phổ biến như: Phân tích các đặc trưng mẫu, so sánh các mẫu thí nghiệm, phân tích phương sai, tương quan hồi quy, dự báo….. do đó phần mềm Excel được chọn lựa để giới thiệu.
Các phần mềm thống kê chuyên dụng và phổ biến trên thế giới là Statgrahics, SPSS, …. Đây là các phần mềm thống kê được ứng dụng rộng trong hầu hết các lĩnh vực nghiên cứu, phân tích dữ liệu của nhiều ngành khác nhau về xã hội, tự nhiên. Ứng dụng mạnh của các phần mềm này là phân tích các mô hình hồi quy đa biến dạng tuyến tính hay phi tuyến tính với các cách phân tích đa dạng như hồi quy lọc, hồi quy từng bước, tổ hợp biến, mã hóa tự động các biến định tính, ….. Do đó phần mềm Statgraphics Centurion XV cũng được giới thiệu để người đọc có thể tiếp cận với công cụ phân tích thống kê này.

Tài liệu này sẽ không đi sâu vào lý thuyết xác suất thống kê, mà thiên về hướng ứng dụng đơn giản, dễ hiểu, kèm theo các ví dụ để người đọc có thể thực hành các chức năng xử lý, phân tích dữ liệu bằng Excel, Statgraphics Centurion XV một cách nhanh chóng, thuận tiện trong hoạt động quản lý và nghiên cứu lâm nghiệp, quản lý tài nguyên thiên nhiên, môi trường.

1. TỔNG QUÁT VỀ CHỨC NĂNG XỬ LÝ THỐNG KÊ CỦA MS.EXCEL 2007 VÀ STATGRAPHICS CENTURION XV

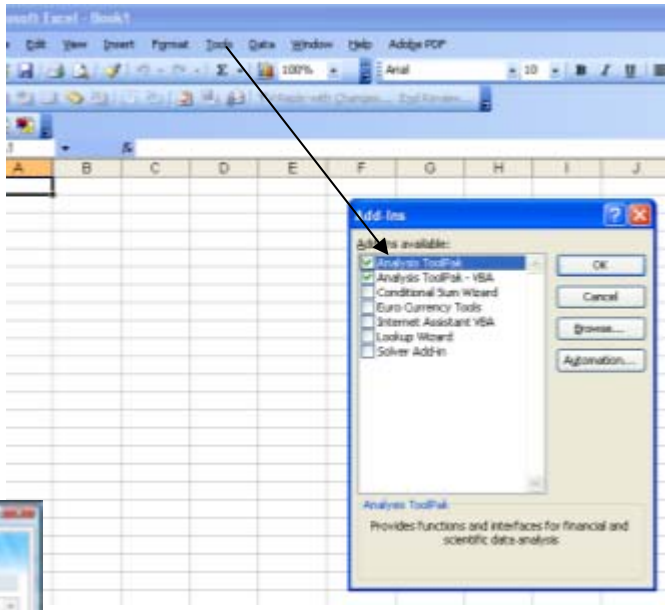
1.1. Tổng quát về phần xử lý thống kê trong MS. Excel

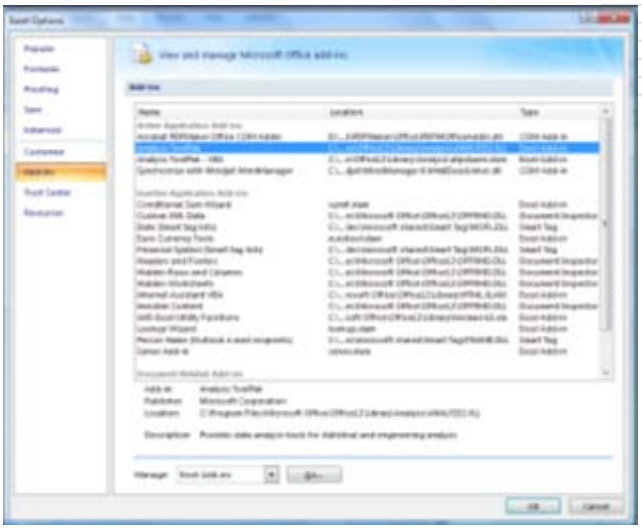
Excel thiết kế sẵn một số chương trình để xử lý số liệu và phân tích thống kê cơ bản ứng dụng trong nhiều lĩnh vực:

- Chức năng xử lý số liệu, tạo bảng tổng hợp dữ liệu: Sắp xếp, tính toán nhanh các bảng tổng hợp từ số liệu thô,...
- Chức năng của các hàm: Cung cấp hàng loạt các hàm về kỹ thuật, thống kê, kinh tế tài chính, hàm tra các chỉ tiêu thống kê như t, F, χ^2
- Chức năng Data Analysis: Dùng để phân tích thống kê như phân tích các đặc trưng mẫu, tiêu chuẩn t để so sánh sự sai khác, phân tích phương sai, ước lượng các tương quan hồi quy
- Phân tích mô hình tương quan hoặc hồi quy để dự báo các thay đổi theo thời gian ngay trên đề thị.

*Lưu ý: Về việc cài đặt chương trình phân tích dữ liệu **(Data Analysis)** trong Excel:*

- *Khi cài đặt phần mềm Excel phải thực hiện trong chế độ chọn lựa cài đặt, sau đó phải chọn mục: Add-Ins và Analysis Toolpak.*
- *Khi chạy Excel lần đầu cần mở chế độ phân tích dữ liệu bằng cách: Menu Tools/Add-Ins và chọn Analysis Toolpak-OK. (Đối với MS. Office 2003)*





Đối với MS. Office 2007, tiến hành mở chế độ phân tích thống kê như sau: Kích vào Microsoft Office Button sau đó chọn excel options, kích vào Add-ins, và chọn Analysis ToolPak trong hộp thoại - OK.

Như vậy trong thực tế quản lý dữ liệu nông lâm nghiệp nói riêng, việc khai thác hết tiềm năng ứng dụng của Excel cũng mang lại hiệu quả tốt mà không nhất thiết phải tìm kiếm thêm một phần mềm chuyên dụng nào khác. Vấn đề đặt ra là xác định chiến lược ứng dụng và khai thác đúng và sâu các công cụ chức năng sẵn có ở một phần mềm phổ biến ở bất kỳ một vi tính cá nhân nào.

Một số hàm thông dụng trong thống kê:

- Tính tổng: =Sum(dãy đs).
- Tổng bình phương: =Sumq(dãy đs).
- Trung bình: =Average(dãy đs).
- Lấy giá trị tuyệt đối: =Abs(đs).
- Trị lớn nhất, nhỏ nhất: =Max(dãy đs), Min(dãy đs).
- Các hàm lượng giác: =Cos(đs), =Sin(đs), =tan(đs).
- Hàm mũ, log: =Exp(đs), =Ln(đs), =Log(đs).
- Căn bậc 2: =Sqrt(đs)..
- Sai tiêu chuNn mẫu chưa hiệu đính: =Stdevp(dãy đs); đã hiệu đính =Stdev(dãy đs).
- Phương sai mẫu chưa hiệu đính: =Varp(dãy đs); đã hiệu đính =Var(dãy đs).
- Giai thừa: =Fact(n).
- Số Pi: =Pi().

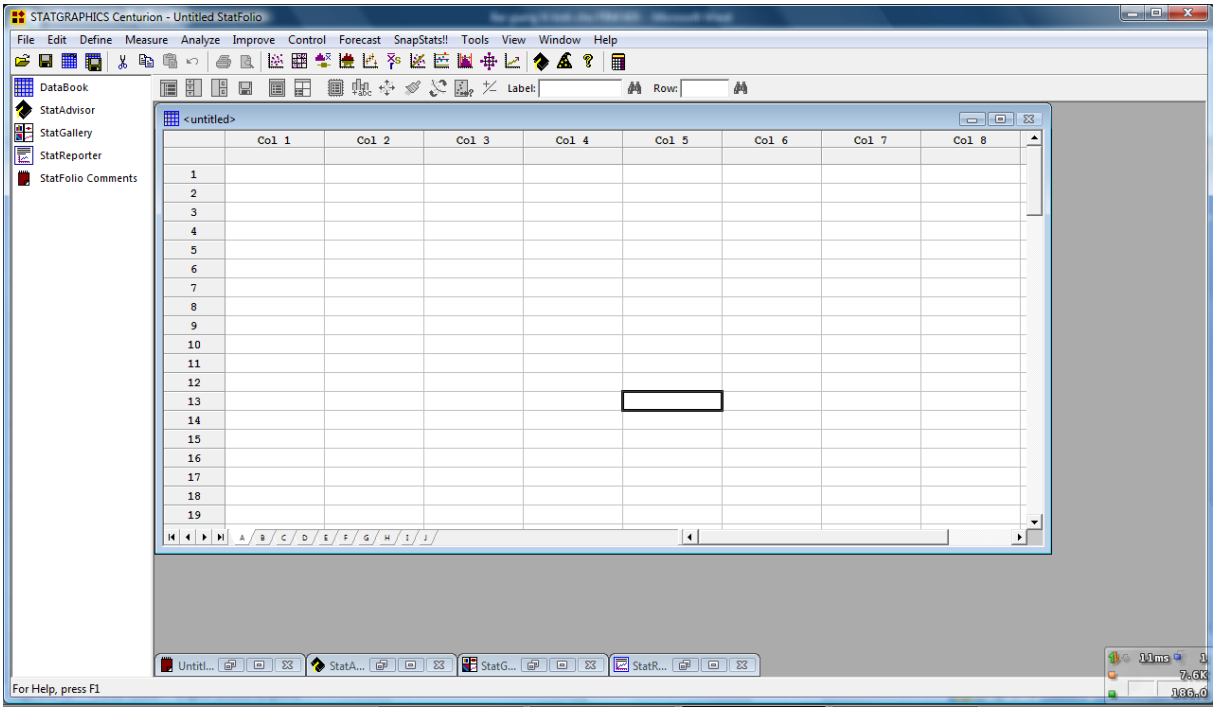
Tra các giá trị T, F, χ2:

- Chọn 1 ô lấy giá trị tra.
- Kích nút fx trên thanh công cụ chuNn. Trong hộp thoại Function Category, chọn Statistical.
- Trong mục Function name, chọn 1 trong các hàm:
  Hàm Tinv: để tra T.
  Hàm Chiinv: để tra χ^2.
  Hàm Finv: để tra F.
  Bấm Next.
- Trong hộp thoại tiếp theo: Function Wizard chọn:
  - Probability (fx): Gõ vào mức ý nghĩa α=0.05 ; 0.01 hay 0.001.
  - Degrees Freedom (fx): Gõ vào bậc tự do. Đối với tiêu chuNn F cần đua vào 2 độ tự do.
  - Finish.

1.2. Tổng quát về phần mềm xử lý thống kê Statgraphics Centurion

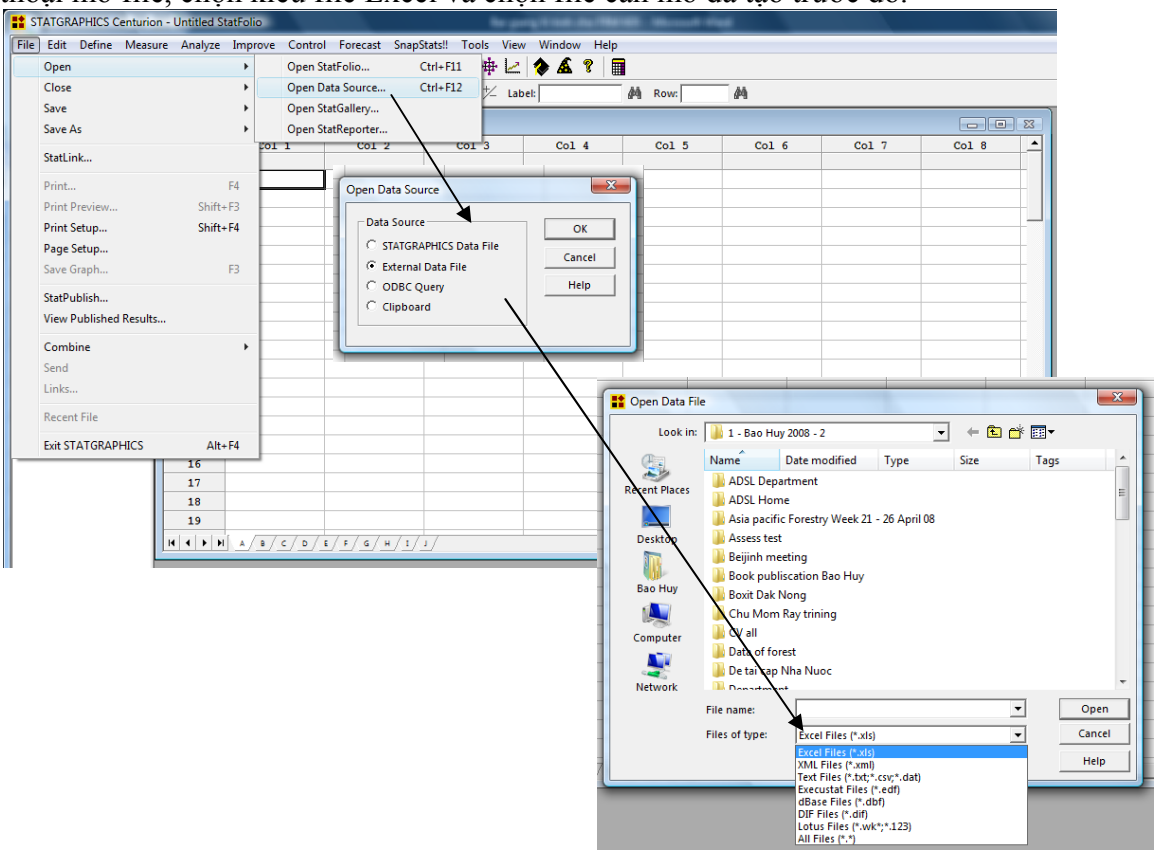
Đây là một phần mềm chuyên dụng trong xử lý thống kê, bao gồm các chức năng:

- Tạo lập cơ sở dữ liệu dưới dạng bảng tính
- Tính toán các đặc trưng mẫu, vẽ sơ đồ, đồ thị quan hệ
- So sánh hai hay nhiều mẫu bằng các tiêu chuNn thống kê t, U, F và nhiều tiêu chuNn phi tham số khác.
- Phân tích phương sai ANOVA.
- Kiểm tra tính chuNn của dữ liệu và đổi biến số.
- Thiết lập các mô hình hồi quy tuyến tính hay phi tuyến tính từ một cho đến nhiều lớp, tổ hợp biến. Với cách xử lý đa dạng để chọn lựa được các biến ảnh hưởng đến một hậu qủa (biến phụ thuộc).



Giao tiếp trong Statgraphics Centurion, số liệu đầu vào có thể được nhập trực tiếp trong file bảng tính và cơ sở dữ liệu; song với các làm này đôi khi không thuận tiện trong các bước xử lý số liệu thô như đổi biến số, tính các biến trung gian, mã hóa biến số. Do đó thông thường nên tạo lập cơ sở dữ liệu trong bảng tính Excel để có thể sử dụng những chức năng bảng tính mạnh của nó trong xử lý dữ liệu thô, tạo lập cơ sở dữ liệu; sau đó sẽ nhập vào Statgraphics Centurion để tính toán, thiết lập mô hình, ... Cơ sở dữ liệu lập trong Excel cần lưu dưới dạng phiên bản của Excel 97 – 2003, vì nó chưa nhận được file Excel ở version 2007.

Sau khi nhập dữ liệu trong Excel 97-2003, đóng file của Excel và mở nó trong Statgrahics Centurion như sau: File/Open/Open Data Source; chọn External Data File – OK. Trong hộp thoại mở file, chọn kiểu file Excel và chọn file cần mở đã tạo trước đó.



2. THỐNG KÊ MÔ TẢ

Để có hiểu biết rõ ràng về một đối tượng quan sát như sinh trưởng cây rừng của một lô rừng, sự đa dạng loài của của lô rừng, biến động mật độ tái sinh, tỷ lệ sống của trồng rừng, ... cần áp dụng thống kê mô tả, bao gồm tiến hành thu thập dữ liệu của mẫu đó và từ đó tính toán đặc trưng của mẫu để ước lượng các chỉ tiêu thống kê cơ bản của tổng thể đó. Đây là các thông tin cơ bản về một đối tượng quan sát, theo một chỉ tiêu, nhân tố quan tâm.

Các đặc trưng mẫu được mô tả bao gồm tính các chỉ tiêu cơ bản: Số trung bình, phương sai, sai tiêu chuNn, độ lệch, độ nhọn của dãy số liệu quan sát được và phạm vi biến động theo một độ tin cậy cho trước.

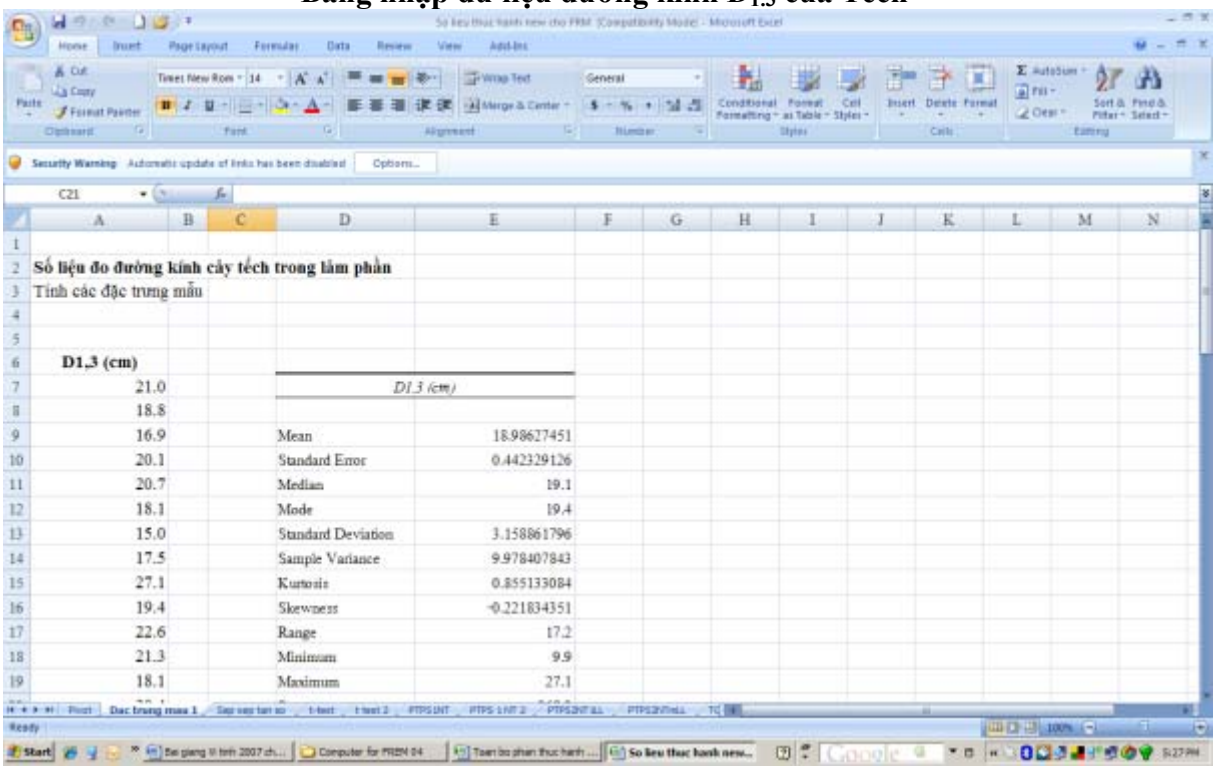
Ví dụ: Khảo sát các đặc trưng cơ bản về sinh trưởng của rừng trồng tếch.

Số liệu đo $D_{1,3}$ rừng trồng Tếch 14 tuổi trong ô tiêu chuNn 500m^2.

Các đặc trưng mẫu có thể tính đồng thời trong Excel theo các bước:

- Nhập số liệu theo cột hoặc hàng.
- Menu Tools/Data Analyisis/Descriptive Statistics/OK (Hoặc Data/Data Analysis trong MS. Ofice 2007). Có hộp thoại, trong đó cần xác định:
  - Input range: Khai báo khối dữ liệu.
  - Grouped by: Chọn dữ liệu nhập theo cột (Columns) hoặc hàng (Rows).
  - Label in first row: Nếu đưa vào cả hàng tiêu đề thì đánh dấu.
  - Output range: Đánh vào địa chỉ ô trên trái nơi đưa ra kết quả.
  - Summary Statistics: Thông tin tóm lược các đặc trưng thống kê (đánh dấu).
  - Confidence Level for Mean: Chọn độ tịnh cậy: 90% hoặc 95% hoặc 99% tùy theo yêu cầu đánh giá, phân tích ướng lượng.
  - Kích nút OK

Bảng nhập dữ liệu đường kính $D_{1.3}$ của Tếch

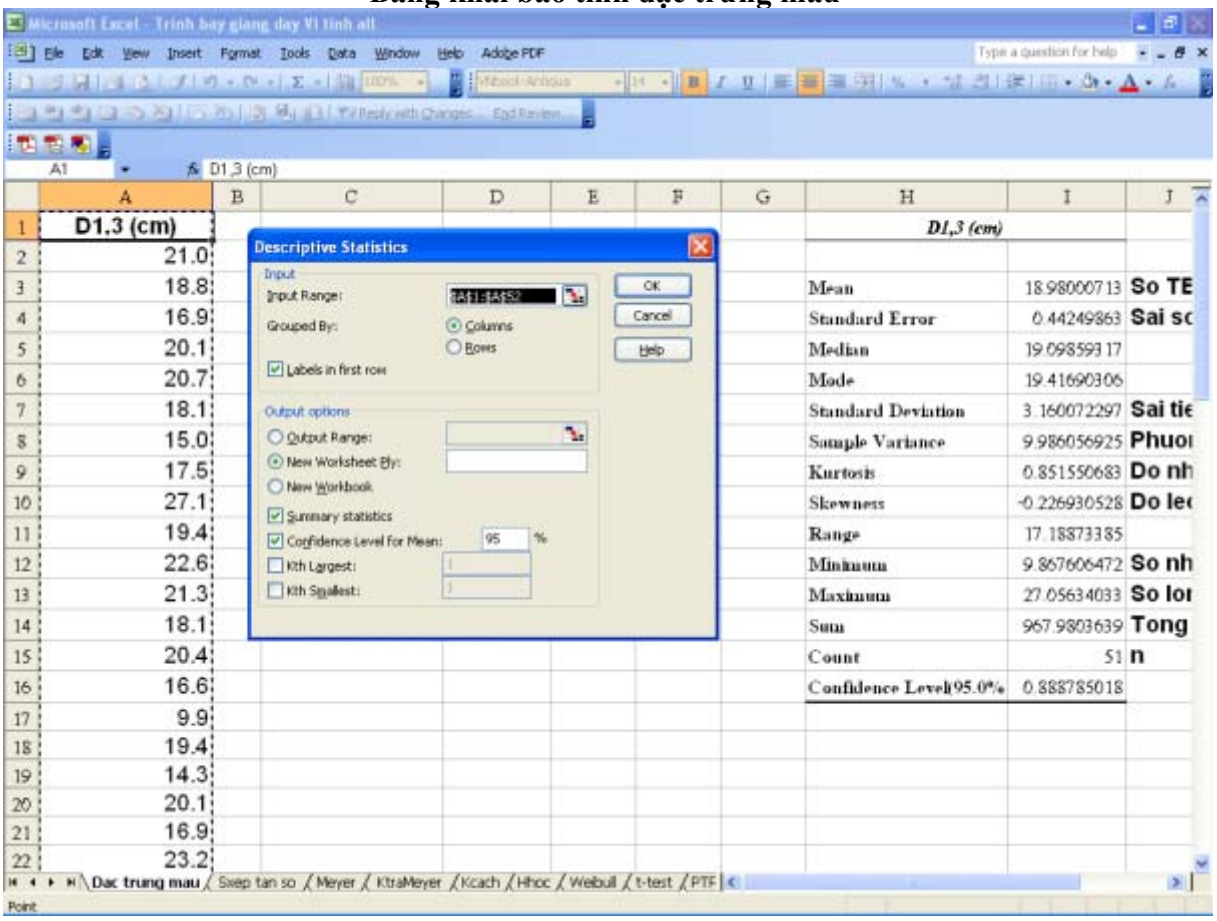


Số liệu đo đường kính cây tếch trong lâm phần
Tính các đặc trưng mẫu

D1,3 (cm)
21.0
18.8
16.9
20.1
20.7
18.1
15.0
17.5
27.1
19.4
22.6
21.3
18.1

D1,3 (cm)	
Mean	18.98627451
Standard Error	0.442329126
Median	19.1
Mode	19.4
Standard Deviation	3.158861796
Sample Variance	9.978407843
Kurtosis	0.855133084
Skewness	-0.221834351
Range	17.2
Minimum	9.9
Maximum	27.1

Bảng khai báo tính đặc trưng mẫu



Kết quả tính đặc trưng mẫu

D1,3 (cm)	
Mean	18,98
Standard Error	0,442
Median	19,1
Mode	19,42
Standard Deviation	3,16
Sample Variance	9,986
Kurtosis	0,852
Skewness	-0,227
Range	17,19
Minimum	9,868
Maximum	27,06
Sum	968
Count	51
Confidence Level (95,0%)	0,889

Giải thích kết quả:

- Mean (Xbq): Số trung bình.
- Standard Error: Sai số của số trung bình mẫu.
- Median: Trung vị mẫu.
- Mode: Trị số ứng với tần số phân bố tập trung nhất.
- Standard deviation (S): Sai tiêu chuNn mẫu.
- Sample variance: Phương sai mẫu.
- Kurtosis (Ku): Độ nhọn của phân bố.
- Skewness (Sk): Độ lệch của phân bố.
- Minimum: Trị số quan sát bé nhất.
- Maximum: Trị số quan sát lớn nhất.
- Sum: Tổng các trị số quan sát.
- Count: Dung lượng mẫu.
- Cofidence level (95%): Sai số tuyệt đối của ước lượng với độ tin cậy 95%.

Với kết quả phân tích đặc trưng mẫu, rút ra được các chỉ số thông kê quan trọng sau:

- Giá trị trung bình và các biến động như sai tiêu chuNn, phương sai, max, min
- Mẫu quan sát đã chuNn hay chưa thông qua Ku và Sk. Mẫu tiệm cận chuNn thì mới bảo đảm số liệu quan sát đủ và các giá trị ước lượng là tin cậy theo độ tin cậy cho trước; nếu không thì giá trị này sẽ sai lệch trong thực tế. Với một mẫu quan sát đạt phân bố chuNn khi Ku và Sk xấp xỉ bằng 0.
    - Kurtosis: Độ nhọn của phân bố
      Ku = 0 phân bố thực nghiệm tiệm cận chuNn.
      Ku > 0 đường cong có dạng bẹt hơn so với phân bố chuNn.
      Ku < 0 đường cong có đỉnh nhọn hơn so với phân bố chuNn.
      Ví dụ Ku = Kurt(A2:A52) = 0.852. Đỉnh đường cong thấp hơn so với phân bố chuNn.
    - Skewness: Độ lệch của phân bố.
      S_k = 0 phân bố đối xứng.
      S_k > 0 đỉnh đường cong lệch trái so với số trung bình.
      S_k < 0 đỉnh đường cong lệch phải so với số trung bình.
      Ví dụ trên S_k = Skew(A2:A52) = -0.227. Đường cong hơi lệch phải.
    - Minimum: Trị số quan sát bé nhất.

Nếu mẫu phân bố chưa chuNn thì cần bổ sung mẫu theo công thức mẫu cần thiết nct:

$$nct \geq t^2.V\%^2/\Delta\%^2$$

Trong đó V% là hệ số biến động: $V\% = \frac{S}{Xbq}100$ và Δ% là sai số tương đối cho trước.

- Ước lượng phạm vi biến động của giá trị trung bình, trong ví dụ trên với độ tin cậy 95% thì đường kính trung bình của khu rừng tếch 14 tuổi biến động trong phạm vi: 18.98 ± 0.89 cm

Hay P(Xbq - Cofidence level (95%) ≤μ≤ Xbq + Cofidence level (95%) = 0.95

3. SẮP XẾP VÀ VẼ BIỂU ĐỒ PHÂN BỐ TẦN SỐ XUẤT HIỆN THEO CẤP, CỠ, HẠNG

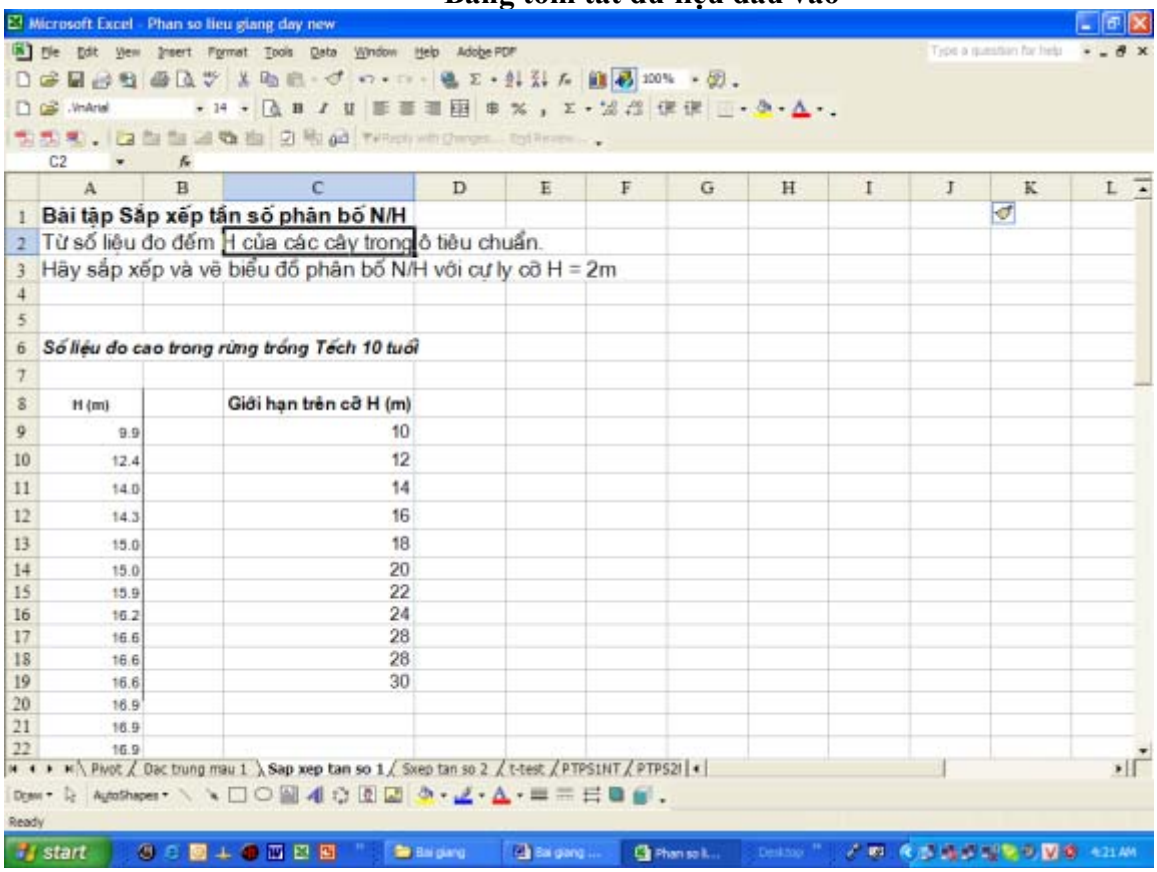
Đây là chức năng sắp xếp bảng phân bố tần số theo một nhân tố theo từng cấp, hạng, ... và vẽ đồ thị phân bố.

Trong nghiên cứu xã hội, người ta cần nghiên cứu tần số phân bố số người theo cấp tuổi để biết sự phân bổ con người theo các thế hệ để có chiến lược quản lý nguồn nhân lực.

Trong quản lý tài nguyên thiên nhiên, thường cần nghiên cứu sự phân bố số lượng cá thể loài theo cấp tuổi, cấp kích thước để biết được quy luật biến đổi cá thể theo thế hệ, theo kích thước, chất lượng, ... là cơ sở quản lý, bảo tồn và định hướng khai thác sử dụng bền vững.
Trong lâm nghiệp thường cần sắp xếp phân bố số cây theo cỡ kính (N/D), số cây theo cỡ chiều cao (N/H), số cây theo cấp thể tích (N/V), số cây theo loài cây theo các tầng rừng, thế hệ để tổ chức quản lý điều chế rừng.
Ví dụ cũng từ số liệu quan sát rừng trồng Tếch 10 tuổi, tiến hành sắp xếp phân bố thực nghiệm N/H và vẽ biểu đồ (cấp H là 2m):

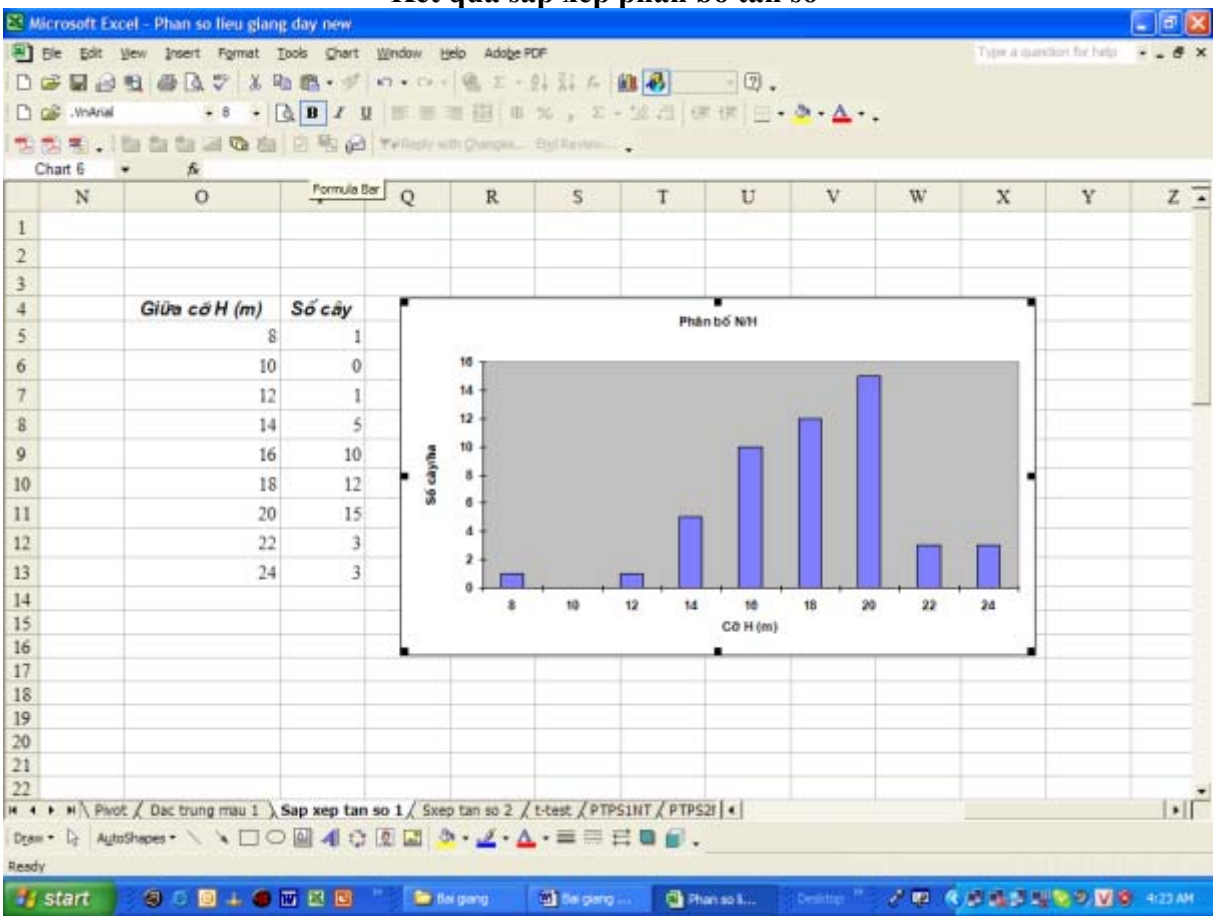
- Nạp số liệu chiều cao vào bảng tính theo cột.
- Lập một cột giới hạn trên cỡ kính. Vd: cỡ 2m.

Bảng tóm tắt dữ liệu đầu vào



- Menu Tools/Data Analysis/Histogram/OK (Data/Data Analysis trong MS Office 20907). Xuất hiện hộp thoại, xác định:
  + Input range: Khai báo khối dữ liệu
  + Bin range: Khai báo khối chứa cự ly tổ.
  + Output range: Khai địa chỉ ô trên trái nơi đưa ra kết quả.
  + Cumulative percentage: Tính phần trăm tần số tích lũy.(Đánh dấu).
  + Chart output: Vẽ biểu đồ. (Đánh dấu chọn).
  + OK.

Kết qủa sắp xếp phân bố tần số



Kết quả sắp xếp tần số cho được một dãy dữ liệu theo cấp và biểu đồ phân bố. Nó phản ảnh cụ thể hơn đặc trưng mẫu và cho thấy hình ảnh của kiểu dạng phân bố theo cấp, thế hệ; từ đó giúp cho việc phân tích quần thể và đưa ra quyết định quản lý, sử dụng bền vững. Ví dụ trong biểu đồ trên, số cây phân hóa khá mạnh theo cấp chiều cao, một số cây sinh trưởng kém ở cấp chiều cao nhỏ 8 – 12m, một số cây vượt tán có cấp H trên 22m; giải pháp đề nghị ở đây là tỉa thưa loại bỏ bớt cây sinh trưởng kém có H < 12m và có thể tỉa thưa một số cây lớn với H>22m để lợi dụng trung gian, lúc này cá thể sẽ có kích thước tập trung trong phạm vi 14 – 22m và có đủ không gian dinh dưỡng để phát triển.

4. SO SÁNH 1 – 2 MẪU QUAN SÁT BẰNG TIÊU CHUẨN T

Kiểm tra mẫu bằng tiêu chuNn t dựa vào giả thiết phân phối chuNn của mẫu quan sát. Có hai loại kiểm tra t: kiểm tra t một mẫu (one-sample t-test), và t cho hai mẫu (two-sample t-test). Kiểm tra t một mẫu để đánh giá số trung bình của một mẫu có phải thật sự bằng một gía trị nào đó hay không?. Kiểm tra t hai mẫu thì để so sánh hai mẫu có cùng một luật phân phối, hay cụ thể hơn là hai mẫu có thật sự có cùng trị số trung bình hay không? Hay nói khác đi có sự sai khác giữa hai mẫu quan sát hay không?

4.1. So sánh một mẫu với một giá trị cho trước – Kiểm tra T một mẫu

Trong mô tả quan sát một mẫu, người ta có thể có yêu cầu đánh giá giá trị trung bình của mẫu với một giá trị cho trước, ví dụ từ đo đếm chiều cao của cây tái sinh trong rừng khộp, so sánh với một giá trị cho trước về chiều cao mong đợi để cây rừng vượt qua được lửa rừng, xem thật sự chiều cao tái sinh của lô rừng đó đã đạt yêu cầu hay chưa?

Để giải quyết vấn đề này, sử dụng kiểm định t một mẫu. Theo lí thuyết thống kê công thức t kiểm tra một mẫu với một giá trị cho trước:

$$t = \frac{Xbq - \mu}{\frac{S}{\sqrt{n}}}$$

Trong đó, Xbq là giá trị trung bình của mẫu, μ là trung bình theo giả thuyết, S là sai tiêu chuẩn và n là số lượng mẫu quan sát.

- Nếu giá trị tuyệt |t| tính cao hơn giá trị t lí thuyết ở mức sai có ý nghĩa, thường là 5% thì có thể kết luận có sự khác biệt có ý nghĩa thống kê giữa trung bình mẫu với giá trị cho trước đó. Và trong trường hợp này nếu t tính <0 thì có nghĩa trung bình của mẫu nhỏ thua có ý nghĩa so với trung bình lý thuyết, ngược lại nếu t tính > 0 thì trung bình của mẫu lớn hơn có ý nghĩa so với trung bình lý thuyết
- Nếu |t| tính ≤ t(0.05, df) thì có thể kết luận ở mức sai 5% trung bình mẫu quan sát xấp xỉ với trung bình lý thuyết.

Trong đó t lý thuyết được tính theo hàm =tinv(0.05, df), với độ tự do df = n-1.

Số liệu đo cao cây tái sinh rừng khộp trong Excel

Stt	Chiều cao cây tái sinh (m)
1	1.5
2	1.3
3	0.8
4	1.9
5	1.7
6	2.2
7	2.5
8	1.0
9	0.7
10	1.9
11	1.8
.....	
58	1.6
59	2.0
60	1.9
61	1.7

Để tính được giá trị t, cần tính toán đặc trưng mẫu để có các giá trị thông kê về Xbq, S.

Kết quả tính đặc trưng mẫu tái sinh rừng khộp

Chiều cao cây tái sinh (m)

Mean	1.64
Standard Error	0.06318
Median	1.7
Mode	1.9
Standard Deviation	0.49347
Sample Variance	0.24351
Kurtosis	-0.4499
Skewness	-0.4627
Range	1.8
Minimum	0.7
Maximum	2.5
Sum	100.3
Count	61
Confidence Level(95.0%)	0.12638

Từ đó tính giá trị thống kê t: So sánh trung bình chiều cao tái sinh với giá trị lý thuyết μ =2m

$$t = \frac{1.64 - 2}{\frac{0.493}{\sqrt{61}}} = -5.63$$

Và t lý thuyết: t (0.05, df = n-1) = tinv(0.05, 60) = 2.00
Kết quả cho thấy |t| = 5.63 > t(0.05, 60). Kết luận: Có sự sai khác có ý nghĩa giữa trung bình chiều cao cây tái sinh rừng khộp với giá trị trung bình lý thuyết mong đợi là 2m. Và t < 0 do đó có nghĩa là chiều cao trung bình cây tái sinh nhỏ thua có ý nghĩa khi so với chiều cao mong đợi là 2m; hay nói khác nếu với yêu cầu cao trên 2m thì mới thoát được ảnh hưởng của lửa rừng, thì lô rừng này cây tái sinh chưa đạt được.

4.2. So sánh sự sai khác giữa trung bình 2 mẫu – Kiểm tra T 2 mẫu

Trong các thí nghiệm thường người ta cần so sánh kết quả của 2 công thức, ví dụ: Bón phân khác nhau, độ tàn che khác nhau, sinh trưởng của cây có xuất xứ khác nhau, nơi bị tác động ảnh hưởng và nơi không, sinh trưởng cây rừng nơi cháy và không cháy.....Việc kiểm tra tiến hành theo 2 mẫu trên cơ sở so sánh 2 số trung bình bằng các tiêu chuẩn t.

Công thức tính giá trị kiểm tra t:

$$\mathbf{t} = \frac{X1 - X2}{\sqrt{\frac{(n1-1)S1^2 + (n2-1)S2^2}{n1+n2-2}\left(\frac{1}{n1} + \frac{1}{n2}\right)}}$$

Với: X_1, X_2: Trung bình của mẫu 1 và 2.
S_1^2, S_2^2: Phương sai mẫu 1 và 2.
n_1, n_2: dung lượng 2 mẫu 1 và 2.

Nếu t tính lớn hơn t bảng với α=0.05 và độ tự do K=n_1+n_2-2 thì bác bỏ giả thuyết Ho, có nghĩa trung bình 2 mẫu sai khác có ý nghĩa, và người ta sẽ chọn mẫu có trung bình cao.

Trước khi sử dụng tiêu chuẩn t, cần kiểm tra 2 điều kiện:

- Hai mẫu có phân bố chuNn.
- Phương sai của hai mẫu có bằng nhau hay không

 Hai mẫu có phân bố chuẩn: Trong thực tế nghiên cứu sinh học, trường hợp dung lượng mỗi mẫu >30 thì có thể xem là tiệm cận chuNn.

 Kiểm tra sự bằng nhau của 2 phương sai của 2 mẫu bằng tiêu chuẩn F.

Trước khi chọn lựa tiêu chuNn t để so sánh trung bình 2 mẫu, cần kiểm tra sự sai khác phương sai của chúng bằng tiêu chuNn F.

Ví dụ: Kiểm tra sinh trưởng chiều cao H của 2 phương pháp trồng thông 3 lá Pinus. kesiya bằng cây con và rễ trần tại trạm thực nghiệm Lang Hanh-Lâm Đồng: Mỗi công thức được rút mẫu theo ô tiêu chuNn 1000m^2, đo đếm chiều cao:

- Dung lượng quan sát mỗi mẫu >90cây, nên chấp nhận giả thuyết phân bố N-H của từng mẫu tiệm cận chuNn; hoặc có thể kiểm tra thêm qua Sk và Ku mỗi mẫu.
- Kiểm tra bằng nhau của 2 phương sai bằng tiêu chuNn F:

Bảng tóm tắt số liệu sinh trưởng H của hai mẫu

	A	B
1	**H (cây con)**	**H (rễ trần)**
2	13,6	13
3	14	13,5
	13,8	12
	13	13,5
	11	15
	12	14
93	12,5	10
94		9

Tính F: Một trong 2 cách:

C1: Kích nút fx, có hộp thoại: Chọn: Statistical (trong Function Category) và Ftest-Next (trong Function name): Xuất hiện hội thoại tiếp theo:

Array 1: Đưa vào dãy 1: A2:A93
Array 2: Đưa vào dãy 2: B2:B94
Finish.

C2: Đưa đến ô kết quả: =Ftest(A2:A93,B2:b94) Enter.

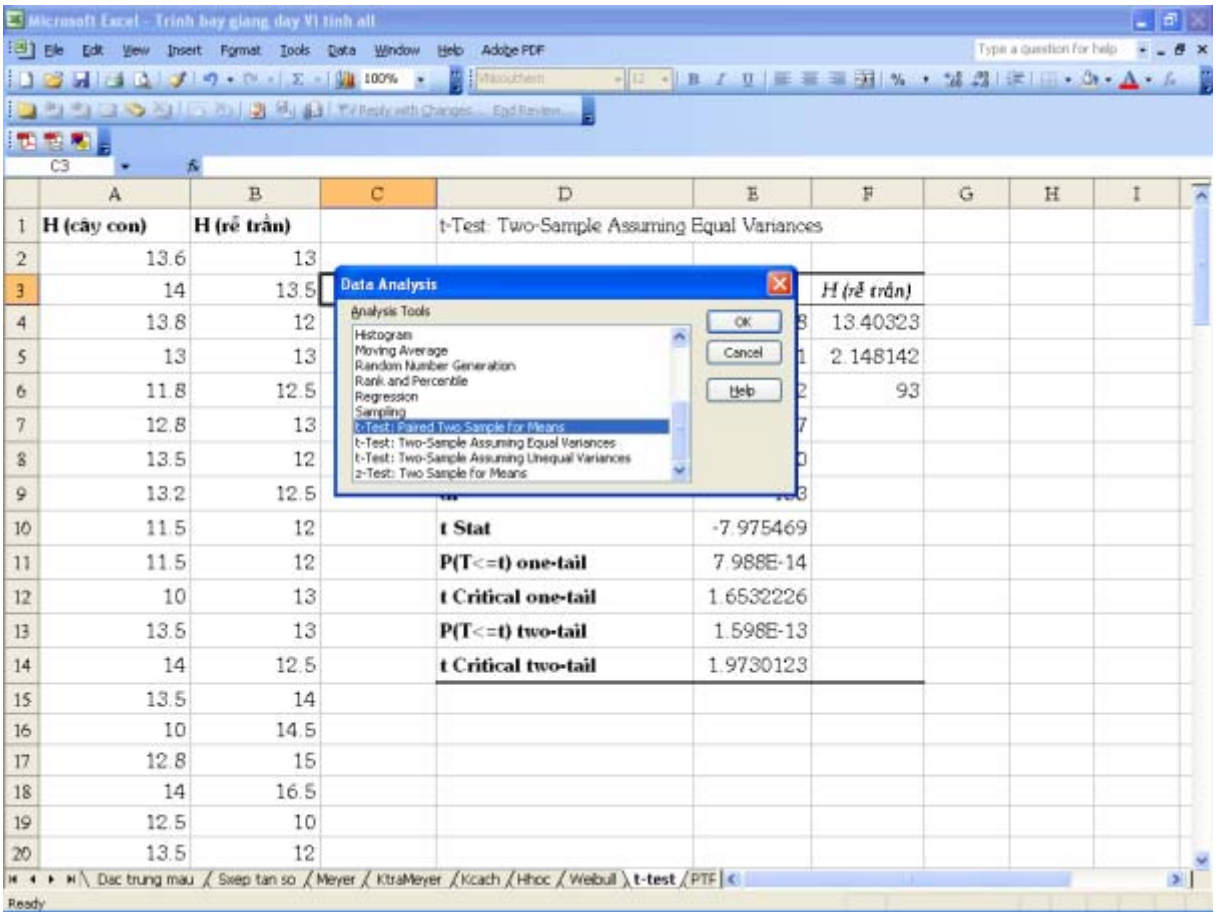
Nếu giá trị xác suất P > 0.05, kết luận hai phương sai bằng nhau, nếu ngược lại thì bác bỏ.

Kết quả ví dụ trên có P=0.40>0.05, kết luận phương sai hai mẫu bằng nhau (chưa có sai dị rõ).

- ***Dùng tiêu chuẩn t để kiểm tra giả thuyết Ho theo trình tự:***

Trong menu Tools/Data Analysis: Chọn trong hộp thoại một trong hai trường hợp tuỳ theo phương sai hai mẫu có bằng nhau hay không qua kiểm tra bằng F ở bước trước

- *t-Test: Two sample assuming equal variance (Trường hợp phương sai bằng nhau).*
- *t-Test: Two sample assuming unequal variance (Trường hợp phương sai không bằng nhau).*



Trong Hộp thoại: Xác định:

- Variable 1 range: Khối dữ liệu mẫu 1 (A1:A93)
- Variable 2 range: Khối dữ liệu mẫu 2 (B1:B94)

  Nên đưa cả tiêu đề.
- Hypothesized mean diference: Đưa vào 0 (Có nghĩa giả thuyết Ho=0).
- Label: Nếu có đưa hàng tiêu đề vào thì cần đánh dấu vài label
- Output range: Đưa địa chỉ ô trên trái nơi xuất kết quả.
- OK.

Nếu: P(T<=t) two tail (hai chiều) <0.05, bác bỏ Ho, có nghĩa 2 mẫu sai dị rõ, ngược lại thì trung bình hai mẫu chưa có sai khác.

Hoặc |t Stat| > t Critical two tail (t hai chiều), bác bỏ Ho, hai mẫu sai dị rõ, ngược lại thì sai khác là ngẫu nhiên.

t-Test: Two-Sample Assuming Equal Variances

	H (cây con)	*H (rễ trần)*
Mean	11,60434783	13,40322581
Variance	2,559761108	2,148141655
Observations	92	93
Pooled Variance	2,352826738	
Hypothesized Mean Difference	0	
df	183	
t Stat	-7,975469453	
P(T<=t) one-tail	7,98781E-14	
t Critical one-tail	1,653222625	
P(T<=t) two-tail	1,59756E-13	
t Critical two-tail	1,973012331	

Từ kết quả trên cho thấy sinh trưởng của P.kesiya trồng bằng 2 phương pháp khác nhau sai dị rõ. Chiều cao bình quân cây trồng bằng rễ trần hơn hẳn trồng bằng cây con, do vậy phương pháp trồng thông 3 lá bằng rễ trần cần được ứng dụng.

5. PHÂN TÍCH PHƯƠNG SAI

Phân tích phương sai là một trong những phương pháp phân tích thống kê quan trọng, đặc biệt là trong các thí nghiệm giống, thí nghiệm các nhân tố tác động đến hiệu quả, chất lượng của cây trồng, vật nuôi, gieo uơm, kiểm nghiệm xuất xứ cây trồng. Chủ yếu đánh giá ảnh hưởng của các công thức, nhân tố đến kết quả thí nghiệm, làm cơ sở cho việc lựa chọn công thức, phương pháp tối ưu trong nông lâm nghiệp.

Điều kiện để phân tích phương sai là:

- Các giá trị quan sát trong từng ô thí nghiệm có phân bố chuNn:
  Nếu dung lượng quan sát đủ lớn (n>30) thì chấp nhận giả thuyết phân bố chuNn.
- Các phương sai của từng nhân tố bằng nhau: Kiểm tra bằng tiêu chuNn Cochran (nếu số lần lặp lại bằng nhau), bằng tiêu chuNn Bartlett (nếu số lần lặp của các công thức không bằng nhau).

5.1. Phân tích phương sai 1 nhân tố với các thí nghiệm ngẫu nhiên hoàn toàn

Phân tích này có một nhân tố như xuất xứ cây trồng, mật độ trồng khác nhau, chế độ chăm sóc khác nhau, ...Trong nhân tố đó được chia thành a công thức, mỗi công thức được lập lại m lần, số lần lặp của mỗi công thức có thể bằng hoặc không bằng nhau.
Trong trường hợp này có thể sử dụng chương trình phân tích phương sai một nhân tố để kiểm tra ảnh hưởng của các công thức đến kết quả thí nghiệm.

Ví dụ: Đánh giá kết quả khảo nghiệm xuất xứ Pinus caribeae tại Lang Hanh-Lâm Đồng.
Theo dự kiến sẽ có 10 xuất xứ P.caribeae được trồng khảo nghiệm tại trạm thực nghiệm Lang Hanh năm 1991. Việc bố trí thí nghiệm ban đầu đã dự kiến tiến hành theo khối ngẫu nhiên

đầy đủ RCB (Randomized Complete Blocks), bao gồm 10 công thức chỉ thị 10 xuất xứ và được lặp lại ở 4 khối.
Nhưng trong quá trình triển khai trồng thực nghiệm, chỉ còn lại 7 xuất xứ và chỉ có 5 xuất xứ lặp lại đủ 4 lần, còn 2 xuất xứ chỉ được lặp lại 2 lần.

7 xuất xứ P.caribeae được trồng thực tế, được đánh số và lặp lại như sau:

1: Xuất xứ	P.alamicamba (NIC)	lặp lại	4 lần.
2:	P.poptun (Guat)	“	4 “
3:	P.guanaja (Nonduras)	“	4 “
4:	P.linures (Nonduras)	“	4 “
5:	P.R482 (Australia)	“	2 “
6:	P.T473 (Australia)	“	4 “
8:	P.little asaco (Bahamas)		2 “

- Mỗi xuất xứ ứng với 1 lần lặp được trồng 25 cây, với cự ly 3x2m, tổng diện tích bố trí thí nghiệm là 1ha.
- Các điều kiện đất đai, vi khí hậu, đia hình, chăm sóc...đều được đồng nhất, nhân tố thay đổi để khảo sát chỉ còn lại là các xuất xứ khác nhau.
- Tại thời điểm điều tra (1996), cây trồng trong các ô thí nghiệm có tuổi là 5. Tiến hành đo đếm toàn diện các chỉ tiêu $D_{1,3}$, H, D_t, phNm chất, tỉa cành, hình thân. Sử dụng 2 chỉ tiêu $D_{1,3}$ và H để đánh giá sinh trưởng của các xuất xứ thử nghiệm.

Dùng phân tích phương sai để đánh giá sự sai khác về sinh trưởng ở các xuất xứ
Trước hết đã kiểm tra 2 điều kiện để phân tích phương sai:

- Điều kiện phân bố chuNn: Các giá trị quan sát ở từng ô thí nghiệm qua kiểm có dạng tiệm cận chuNn nên chấp nhận giả thuyết phân bố chuNn.
- Phương sai bằng nhau: Do dung lượng mẫu ở các xuất xứ không bằng nhau nên dùng tiêu chuNn Bartlett để kiểm tra, kết quả tính được:

  $X^2 = 3,73 < X^2 (0,05 ; 6) = 12,59$

  Do đó chấp nhận giả thuyết bằng nhau của các phương sai mẫu.

  Như vậy 2 điều kiện trên là thỏa mãn để tiến hành phân tích phương sai.

Dùng phân tích phương sai 1 nhân tố để kiểm tra. Trong đó nhân tố là Xuất xứ với 7 công thức:

Giá trị $D_{1,3}$ (cm) bình quân ứng với từng ô thí nghiệm của các Xuất xứ theo khối (lần lặp lại)

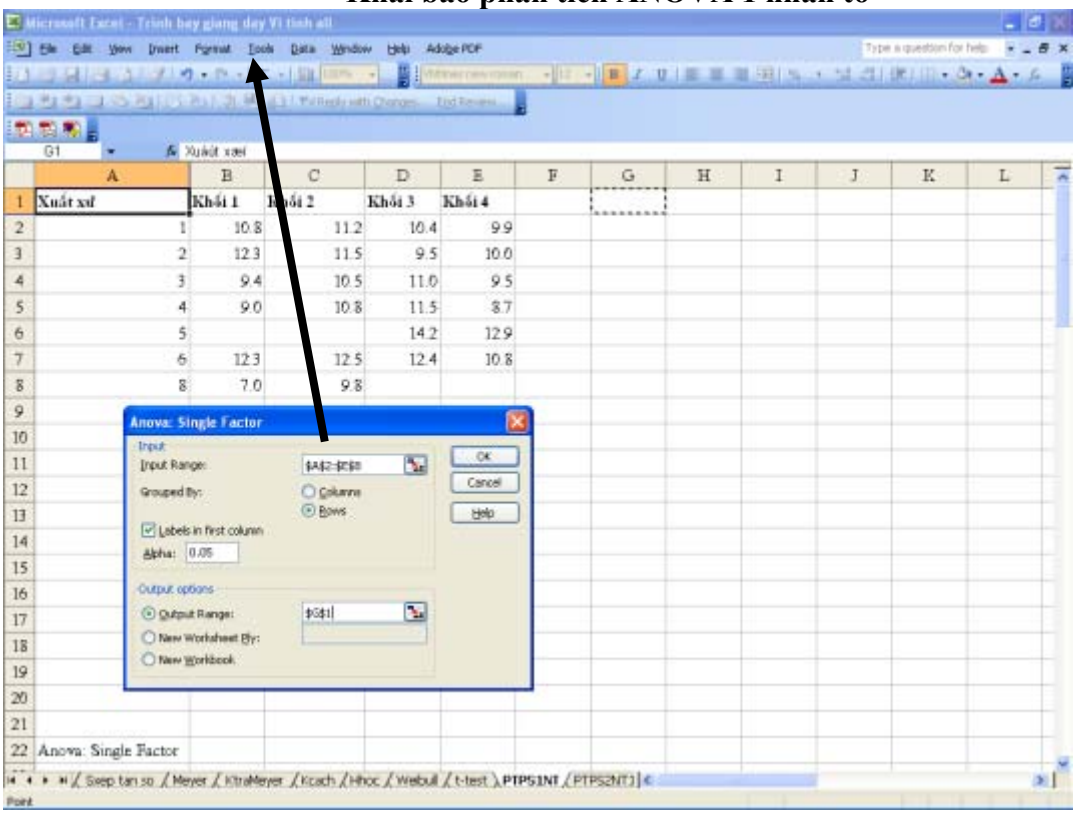
	A	B	C	D	E
1	**Xuất xứ**	**Khối 1**	**Khối 2**	**Khối 3**	**Khối 4**
2	1	10.8	11.2	10.4	9.9
3	2	12.3	11.5	9.5	10.0
4	3	9.4	10.5	11.0	9.5
5	4	9.0	10.8	11.5	8.7
6	5			14.2	12.9
7	6	12.3	12.5	12.4	10.8
8	8	7.0	9.8		

Phân tích phương sai 1 nhân tố:

Vào menu Tools/Data Analysis/Anova (Hoặc Data/Data Analysis/Anova trong MS. Ofice 2007): Chọn ANOVA Single Factor có được Hộp thoại:

- Input range: Nhập địa chỉ khối dữ liệu. Vd: A2:E8. (Có cột đầu chứa số hiệu công thức, nhưng bỏ hàng đầu).
- Grouped by: Chọn Columns hoặc Rows.
- Đánh dấu vào Label in first colum (row).
- Output range: Đưa địa chỉ ô trên trái nơi xuất kết quả.
- Kích OK.

Khai báo phân tích ANOVA 1 nhân tố



Kết quả phân tích phương sai 1 nhân tố

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
1	4	42.3	10.6	0.299523
2	4	43.2	10.8	1.703825
3	4	40.3	10.1	0.616404
4	4	40.0	10.0	1.780196
5	2	27.1	13.5	0.797116
6	4	48.1	12.0	0.673895
8	2	16.7	8.4	3.903367

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	37.53507	6	6.255846	5.338286	0.002925	2.698656
Within Groups	19.92201	17	1.171883			
Total	57.45708	23				

Từ bảng ANOVA nhận được: Đối với các xuất xứ khác nhau: F = 5,33 > $F_{(0,05)}$ = 2,69. Kết luận: Các xuất xứ khác nhau có sự sai khác về sinh trưởng đường kính. Nếu ngược lại thì kết luận rằng giữa các xuất xứ chưa có sự sai khác về sinh trưởng.

Trên cơ sở đó chọn hai xuất xứ có trung bình cao nhất và thứ hai để so sánh bằng tiêu chuNn t. Kết quả cho thấy khoogn có sai khác.

Như vậy, xét theo chỉ tiêu đường kính, xuất xứ tối ưu trong 7 xuất xứ khảo nghiệm là 5 và 6, hai xuất xứ này có chỉ tiêu D lớn nhất, chưa có sai dị với nhau và có sai khác rõ rệt với các xuất xứ còn lại. Đó là 2 xuất xứ: P.R482 (Australia) và P.T473 (Australia).

5.2. Phân tích phương sai 2 nhân tố

Trong các thí nghiệm người ta thường so sánh và phân tích tác động đồng thời 2 nhân tố (ví dụ như đất và xuất xứ) lên kết quả thí nghiệm như: năng suất, sinh khối... Phân tích phương sai lúc này chia 2 trường hợp: *Hai nhân tố với một lần lặp và Hai nhân tố với nhiều lần lặp lại.*

5.1.1. Phân tích phương sai 2 nhân tố với 1 lần lặp lại: (Bố trí thí nghiệm theo khối ngẫu nhiên đầy đủ (Randomized Complete Blocks) (RCB):

Kiểu bố trí thí nghiệm RCB thường được sử dụng, nhân tố A chia làm a cấp và được lặp lại ở b khối (nhân tố B).

Ví dụ: Đánh giá kết quả khảo nghiệm 16 xuất xứ Pinus kesiya tại Lang Hanh-Lâm Đồng: 16 xuất xứ P.kesiya đã được trồng khảo nghiệm tại trạm thực nghiệm Lang Hanh năm 1991. Việc bố trí thí nghiệm đã được tiến hành theo khối ngẫu nhiên đầy đủ RCB (Randomized Complete Blocks), bao gồm 16 công thức chỉ thị 16 xuất xứ và được lặp lại ở 4 cấp đất (khối)

16 xuất xứ P.kesiya được đánh số như sau:

1: Xuất xứ Bengliet.
2: Faplac.
3: Xuân Thọ.
4: Thác Prenn.
5: Lang Hanh.
6: Nong Kiating.
7: Doisupthep.
8: Doiinthranon.
9: Phu Kradung.
10: Nam nouv.
11: Cotomines.
12: Simao.
13: Watchan.
14: Zo khua.
15: Aung ban.
16: Jingdury.

- Mỗi công thức ứng với 1 lần lặp được trồng 25 cây, với cự ly 3x2m, tổng diện tích bố trí thí nghiệm là 1,5ha.
- Các khí hậu, địa hình, chăm sóc...đều được đồng nhất, nhân tố thay đổi để khảo sát chỉ còn lại là các xuất xứ và cấp đất khác nhau.
- Tại thời điểm điều tra (1996), cây trồng trong các ô thí nghiệm có tuổi là 5. Tiến hành đo đếm toàn diện các chỉ tiêu $D_{1,3}$, H, D_t, phNm chất, tỉa cành, hình thân. Sử dụng 2 chỉ tiêu $D_{1,3}$ và H để đánh giá sinh trưởng của các xuất xứ thử nghiệm.

Dùng phân tích phương sai để đánh giá sự sai khác về sinh trưởng, cụ thể cho từng chỉ tiêu sinh trưởng như sau:

Trước hết đã kiểm tra 2 điều kiện để phân tích phương sai:

- Điều kiện phân bố chuNn: Các giá trị quan sát ở từng ô thí nghiệm qua kiểm tra bảo đảm các mẫu tiệm cận chuNn nên chấp nhận giả thuyết phân bố chuNn.
- Phương sai bằng nhau: Dùng tiêu chuNn Cochran, kết quả tính được:

  Gmax = 0,11 < Gmax (0,05 ; 16 ; 3) = 0,28

  Do đó chấp nhận giả thuyết bằng nhau của các phương sai mẫu.

Như vậy 2 điều kiện trên là thỏa mãn để tiến hành phân tích phương sai.

Dùng phân tích phương sai 2 nhân tố 1 lần lặp để kiểm tra:

Với nhân tố thứ nhất là 16 xuất xứ, nhân tố thứ 2 là cấp đất với 4 cấp. Ứng với 1 tổ hợp Xuất xứ - Cấp đất chỉ có 1 ô thí nghiệm (lặp lại 1 lần).

Bảng dữ liệu phân tích phương sai 2 nhân tố 1 lần lặp

Giá trị $D_{1,3}$ (cm) bình quân ứng với từng ô thí nghiệm

	A	B	C	D	E
1	**Xuất xứ**	**Cấp đất 1**	**Cấp đất 2**	**Cấp đất 3**	**Cấp đất 4**
2	1	11.4	11.3	10.8	13.3
3	2	11.4	11.6	10.9	10.9
4	3	11.7	12.6	11.7	12.6
5	4	13.7	12.1	11.6	11.7
6	5	14.1	13.6	13.7	13.7
7	6	13.5	11.4	12.2	11.3
8	7	13.8	12.3	12.6	11.4
9	8	14.1	13.3	15.2	13.0
10	9	13.8	11.8	11.9	12.1
11	10	11.3	11.8	12.1	11.8
12	11	12.6	12.6	13.3	10.9
13	12	11.3	12.4	10.5	12.0
14	13	12.7	13.4	12.1	10.7
15	14	10.1	9.5	9.8	8.0
16	15	10.5	9.4	9.1	10.9
17	16	10.2	11.0	10.8	11.9

Phân tích phương sai 2 nhân tố 1 lần lặp:

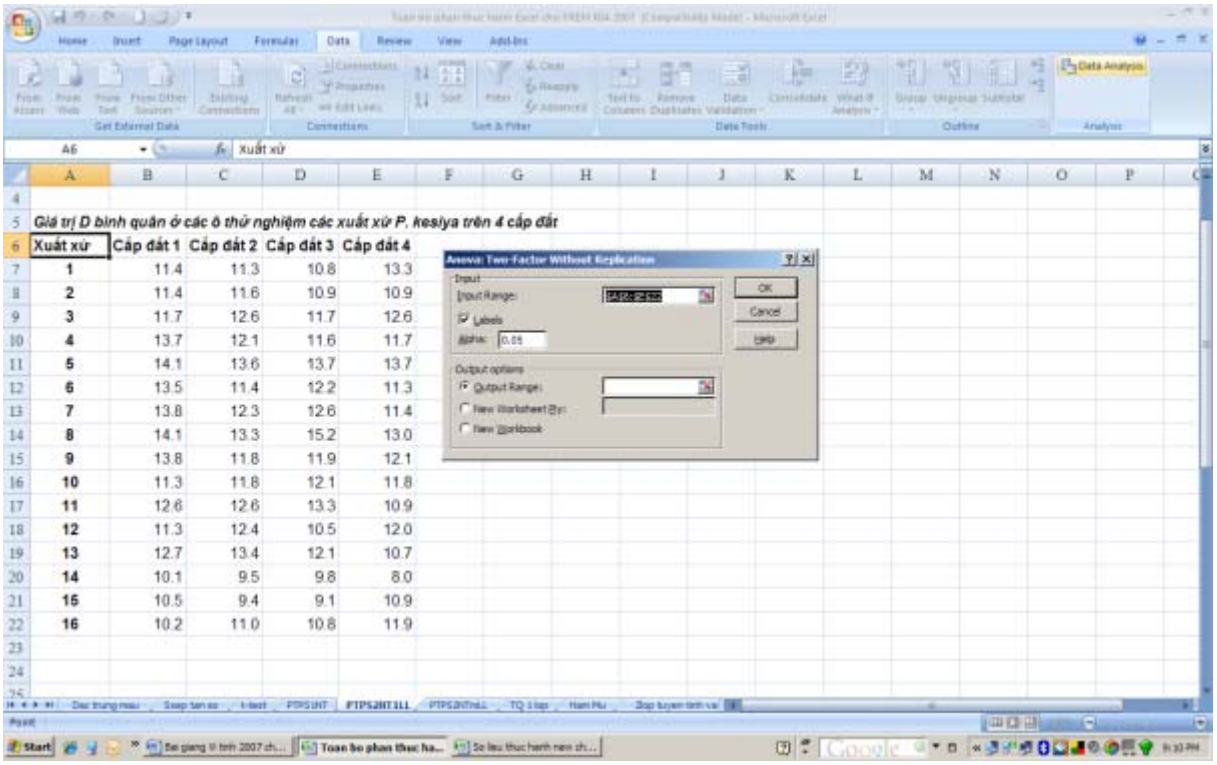
- Tools/Data Analysis/Anova: Two Factor Without Replication - OK.
- Hộp thoại:

  Input range: Địa chỉ khối dữ liệu (Nên quét cả hàng, cột đầu làm nhãn). Vd: A1:E17

  Đánh dấu vào Labels.

  Output range: Địa chỉ ô trên trái nơi xuất kết quả

  OK



Kết quả phân tích phương sai 2 nhân tố 1 lần lặp lại

Anova: Two-Factor Without Replication

SUMMARY	*Count*	*Sum*	*Average*	*Variance*
1	4	46.9	11.7	1.253512
2	4	44.8	11.2	0.156318
3	4	48.6	12.2	0.268337
4	4	49.1	12.3	0.933224
5	4	55.1	13.8	0.049285
6	4	48.5	12.1	1.064903
7	4	50.0	12.5	0.975826
8	4	55.7	13.9	0.926688
9	4	49.7	12.4	0.817143
10	4	47.0	11.7	0.107475
11	4	49.3	12.3	1.054463
12	4	46.1	11.5	0.664541
13	4	48.9	12.2	1.255351
14	4	37.4	9.3	0.85117
15	4	39.9	10.0	0.763403
16	4	43.9	11.0	0.514494

Cấp đất 1	16	196.1	12.3	2.077919
Cấp đất 2	16	190.2	11.9	1.470334
Cấp đất 3	16	188.3	11.8	2.263297
Cấp đất 4	16	186.3	11.6	1.767392

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	82.11826	15	5.474551	7.804468	3.58E-08	1.894875
Columns	3.402532	3	1.134177	1.616873	0.198718	2.811547
Error	31.56586	45	0.701464			
Total	117.0867	63				

Từ bảng ANOVA nhận được:

- Đối với các xuất xứ khác nhau (Hàng - Rows): $F = 7,80 > F_{(0,05)} = 1,89$. Kết luận: Các xuất xứ khác nhau có sự sai khác về sinh trưởng đường kính.
- Đối với các cấp đất (Cột – Collumns): $F = 1,62 < F_{(0,05)} = 2,81$. Kết luận: Các cấp đất khác nhau chưa có ảnh hưởng đến sinh trưởng.

Như vậy 16 xuất xứ khi trồng ở Lang Hanh đã có sinh trưởng khác nhau, do việc cấp đất không ảnh hưởng rệt, nên để đánh giá chính xác hơn chỉ cần phân tích phương sai 1 nhân tố (xuất xứ):

Phân tích phương sai 1 nhân tố

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
1	4	46.9	11.7	1.253512
2	4	44.8	11.2	0.156318
3	4	48.6	12.2	0.268337
4	4	49.1	12.3	0.933224
5	4	55.1	13.8	0.049285
6	4	48.5	12.1	1.064903
7	4	50.0	12.5	0.975826
8	4	55.7	13.9	0.926688
9	4	49.7	12.4	0.817143
10	4	47.0	11.7	0.107475
11	4	49.3	12.3	1.054463
12	4	46.1	11.5	0.664541
13	4	48.9	12.2	1.255351
14	4	37.4	9.3	0.85117
15	4	39.9	10.0	0.763403
16	4	43.9	11.0	0.514494

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	82.11826	15	5.474551	7.514741	3.59E-08	1.880174
Within Groups	34.9684	48	0.728508			
Total	117.0867	63				

Kết quả từ bảng ANOVA cho thấy F = 7,51 > $F_{(0,05)}$ = 1,88. Kết luận: Sinh trưởng đường kính của 16 xuất xú là khác nhau khi trồng ở Lang Hanh.

Sinh trưởng bình quân đường kính các xuất xứ theo thứ tự từ cao đến thấp ở bảng sau:

Thứ tự sinh trưởng đường kính từ tốt đến xấu

Xuất xứ	D1,3 tb(cm)
8	13.9
5	13.8
7	12.5
9	12.4
11	12.3
4	12.3
13	12.2
3	12.2
6	12.1
10	11.7
1	11.7
12	11.5
2	11.2
16	11.0
15	10.0
14	9.3

Xuất xứ 8 có giá trị trung bình cao nhất, sau đó dùng tiêu chuNn t để so sánh sinh trưởng đường kính lớn nhất của xuất xứ 8 với các xuất xứ có đường kính lần lượt nhỏ hơn. Kết quả cho thấy xuất xứ 8 không có sai dị với xuất xứ có trung bình thứ hai là xuất xứ 5.

Như vậy, xét theo chỉ tiêu đường kính, xuất xứ tối ưu trong 16 xuất xứ khảo nghiệm là 8 và 5, hai xuất xứ này có chỉ tiêu D lớn nhất, chưa có sai dị với nhau và có sai khác rõ rệt với các xuất xứ còn lại. Đó là 2 xuất xứ: Doiinthranon và Lang Hanh.

5.1.2. Phân tích phương sai 2 nhân tố m lần lặp

Trường hợp này mỗi tổ hợp nhân tố A và B được lặp lại m lần một cách ngẫu nhiên. Lúc này ngoài việc đánh giá ảnh hưởng của từng nhân tố A, B còn phải tính ảnh hưởng qua lại của chúng đến kết quả thí nghiệm.

Ví dụ: Nghiên cứu ảnh hưởng của hai nhân tố thí nghiệm là mật độ và bón phân đến năng suất của bông.

- Nhân tố A: Mật độ chia làm 3 cấp.
- Nhân tố B: Phân bón được chia làm 4 mức
- Mỗi tổ hợp được thí nghiệm lập lại ngẫu nhiên 4 lần.

Bảng số liệu sản lượng bông theo tổ hợp 2 nhân tố và lặp lại 4 lần ở một tổ hợp
(Đ/v: Tạ/ha)

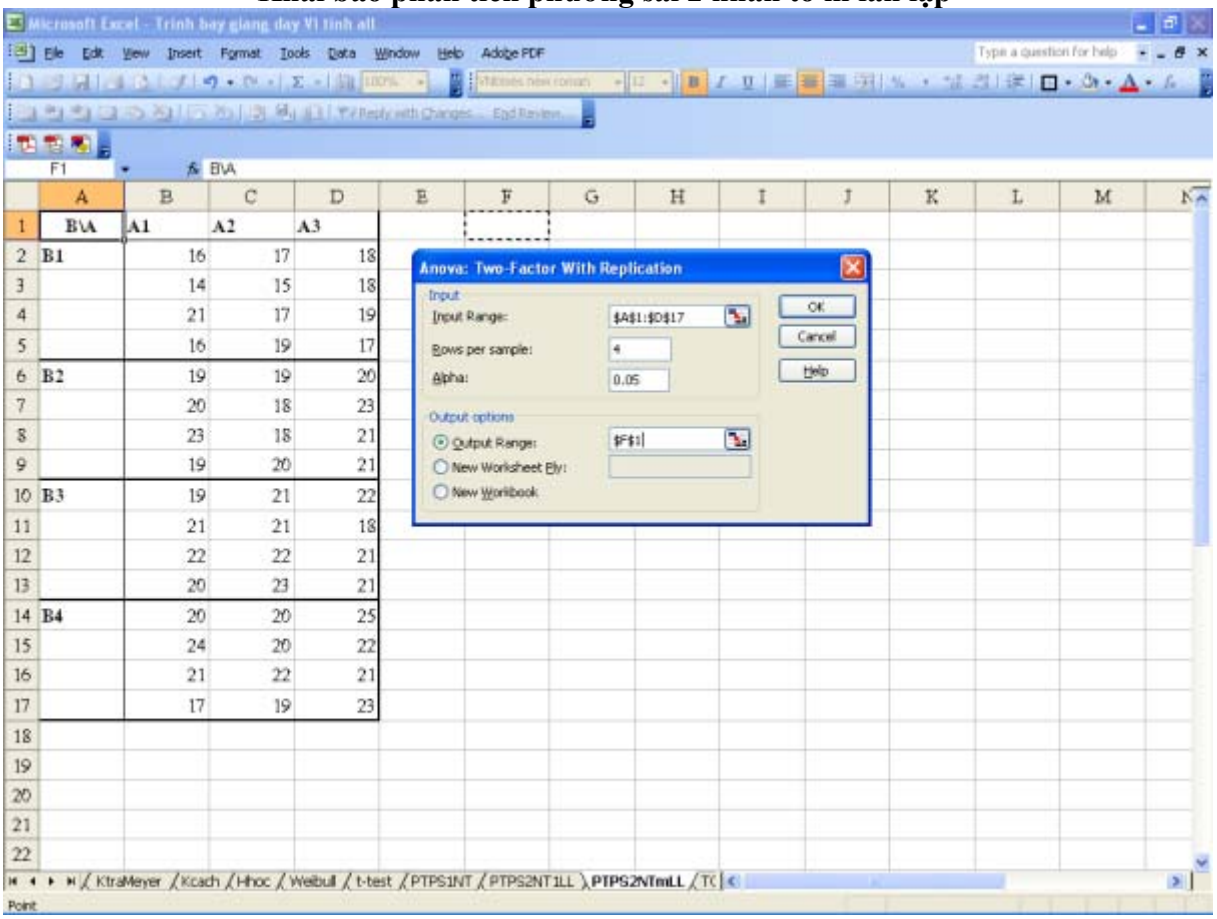
	A	B	C	D
1	**B\A**	A1	A2	A3
2	B1	16	17	18
3		14	15	18
4		21	17	19
5		16	19	17
6	B2	19	19	20
7		20	18	23
8		23	18	21
9		19	20	21
10	B3	19	21	22
11		21	21	18
12		22	22	21
13		20	23	21
14	B4	20	20	25
15		24	20	22
16		21	22	21
17		17	19	23

Phân tích phương sai 2 nhân tố m lần lặp:

- Tools/Data Analysis/Anova: Two Factor With Replication- OK.
- Hộp thoại: Xác định:

  Input range: Nhập khối dữ liệu kể cả hàng cột tiêu đề. Vd: A1:D17.
  Rows per sample: Nhập số lần lặp. Vd: 4.
  Output range: Nhập địa chỉ ô trên trái nơi xuất kết quả.
  OK.

Khai báo phân tích phương sai 2 nhân tố m lần lặp



Kết quả phân tích phương sai 2 nhân tố m lần lặp

Anova: Two-Factor With Replication

SUMMARY	1	2	3	Total
1				
Count	4	4	4	12
Sum	67	68	72	207
Average	16,75	17	18	17,25
Variance	8,916667	2,666667	0,666667	3,659091
2				
Count	4	4	4	12
Sum	81	75	85	241
Average	20,25	18,75	21,25	20,08333
Variance	3,583333	0,916667	1,583333	2,810606
3				
Count	4	4	4	12
Sum	82	87	82	251

Average	20,5	21,75	20,5	20,91667
Variance	1,666667	0,916667	3	1,901515
4				
Count	4	4	4	12
Sum	82	81	91	254
Average	20,5	20,25	22,75	21,16667
Variance	8,333333	1,583333	2,916667	4,878788
Total				
Count	16	16	16	
Sum	312	311	330	
Average	19,5	19,4375	20,625	
Variance	7,2	4,529167	4,783333	

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Sample	116,2292	3	38,74306	12,65079	8,45E-06	2,866265
Columns	14,29167	2	7,145833	2,333333	0,111468	3,259444
Interaction	21,20833	6	3,534722	1,154195	0,352014	2,363748
Within	110,25	36	3,0625			
Total	261,9792	47				

- Bảng Summary: Cho kết quả tính toán từng tổ hợp nhân tố A/B và chung cho từng nhân tố B, nhân tố A, gồm các chỉ tiêu: Dung lượng (Count), Tổng (Sum), Trung bình (Average), Phương sai (Variance).
- Bảng ANOVA:

Cột đầu tiên là các nguồn biến động:

- Sample: Biến động do nhân tố B tạo nên (do được xếp theo hàng).
- Columns: Biến động do nhân tố A tạo nên (do được xếp theo cột).
- Interaction: Tác động qua lại.
- Within: Biến động ngẫu nhiên.
- Total: Biến động chinh của n giá trị quan sát.

Từ kết quả này cho thấy:

$F_B = 12.65 > F_{0.05} = 2.87$. Kl: Phân bón có tác động rõ rệt đến năng suất bông.

$F_A = 2.33 < F_{0.05} = 3.26$. Kl: Mật độ ảnh hưởng không rõ đến năng suất bông.

$F_{AB} = 1.15 < F_{0.05} = 3.36$. Kl: Đồng thời thay đổi mật độ và phân bón ảnh hưởng không rõ đến năng suất.

Lúc này chỉ còn việc lựa chọn công thức bón phân tối ưu. Qua số trung bình năng suất theo từng công thức bón phân cho thấy công thức 4 có năng suất cao nhất là 21.16 tạ/ha. Có thể dùng tiêu chuNn t để kiểm tra lại xem công thức 4 có sai khác với công thức nào còn lại để lựa chọn công thức có hiệu quả nhất.

6. PHÂN TÍCH TƯƠNG QUAN - HỒI QUY

Trong thực tế người ta cần lập các mô hình tương quan hồi quy vì các mục đích:

- Để ước lượng một nhân tố khó đo đếm (gọi là biến phụ thuộc y) thông qua một hay nhiều biến dễ quan sát, đo đếm (gọi là biến độc lập x) và tất nhiên là phải có mối liên hệ giữa y và x. Từ đây có thể lập các biểu điều tra phục vụ cho việc giảm nhẹ các quan sát đo đếm một số nhân tố phức tạp
- Để dự báo một nhân tố trong tương lai (gọi là biến dự báo y) với một số biến độc lập, đầu vào (gọi là là biến độc lập x)
- Để nghiên cứu tác động, ảnh hưởng của một hoặc nhiều nhân tố đến một yếu tố cần quan tâm như sinh trưởng, sản lượng, chất lượng rừng, xói mòn đất, dòng chảy lưu vực. Trên cơ sở đó có giải pháp kỹ thuật thích hợp hoặc các biện pháp quản lý quy hoạch cấp vĩ mô.

Mục đích là sử dụng chương trình Excel hoặc Statgraphics để thiết lập các mô hình tương quan/hồi quy tuyến tính từ một cho đến nhiều biến số độc lập. Trong chương trình này, các tham số được ước lượng bằng phương pháp bình phương tối thiểu. Riêng các dạng phi tuyến khi ứng dụng chương trình này cần đổi biến số để quy về dạng tuyến tính.

6.1. Hồi quy tuyến tính 1 lớp

Hồi quy tuyến tính một lớp có nghĩa là có một biến số độc lập x được nghiên cứu ảnh hưởng đến biến phụ thuộc y, dạng quan hệ được xác định là đường thẳng. Có nghĩa là khi x tăng hoặc giảm thì y cũng tăng hoặc giảm đều theo dạng được thẳng. Dạng phương trình tổng quát: $Y = A + B.X$.

Vd: Lập mô hình tương quan giữa chiều cao dưới cành (Hdc) với chiều cao cả cây (H) rừng Tếch dạng đường thẳng: $Hdc = A + B.H$. Vì Hdc là chỉ tiêu khó đo đếm hơn H, nên dùng quan hệ này để xác định Hdc thông qua H.

- **Nhập số liệu theo bảng:**

Các cặp số liệu Hdc - H

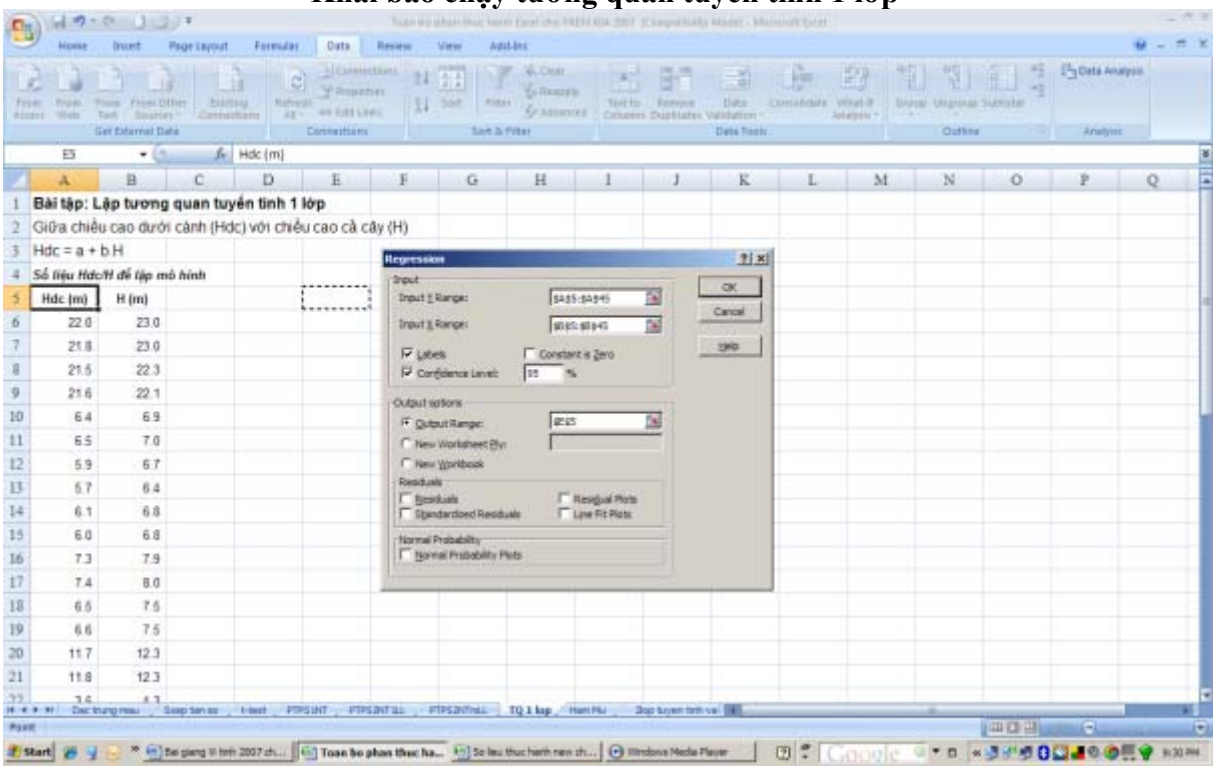
	A	B
1	**Hdc(m)**	**H(m)**
2	22,0	23,0
3	21,8	23,0
4	21,5	22,3
.....	.....	.....
40	9,7	10,9
41	9,8	11,1

- **Ước lượng tương quan hồi quy đường thẳng:**
  - Tools/Data Analysis/Regression (Data/Data Analysis/Regression trong MS. Office 2009). OK.
  - Hộp thoại:

  Input Y range: Nhập địa chỉ cột biến Y (Có thể nhập cả nhãn). Vd: A1:A41.
  Input X range: Nhập địa chỉ cột biến X (Có thể nhập cả nhãn). Vd: B1:B41.
  Label: Đánh dấu nếu đã nhập cả hàng đầu làm nhãn.

Output range: Nhập địa chỉ ô trên trái nơi xuất kết quả.
OK.

Khai báo chạy tương quan tuyến tính 1 lớp



Kết quả ước lượng hồi quy tuyến tính 1 lớp

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,998189546
R Square	0,99638237
Adjusted R Square	0,996287169
Standard Error	0,318271114
Observations	40

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1060,180842	1060,181	10466,12	5,24804E-48
Residual	38	3,84926708	0,101297		
Total	39	1064,030109			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-0,715306008	0,127254043	-5,62109	1,88E-06	-0,972918358	-0,457693658
Hgo(m)	0,994341123	0,009719471	102,304	5,25E-48	0,974665081	1,014017165

Phương trình tương quan:

Hdc = - 0.715 + 0.994.H

Với N = 40 R = 0.998 Fr = 10466.12 với α<0.0000 nên R tồn tại (khác 0)

Từ phương trình hồi quy, có thể xác định Hdc gián tiếp qua H.

6.2. Dạng phi tuyến đưa về tuyến tính 1 lớp

Trong thực tế biến y có thể không có dạng quan hệ đường thẳng với x, do đó cần sử dụng mô hình phi tuyến. Trường hợp các hàm phi tuyến, để ước lượng cần biến đổi thành dạng tuyến tính để ước lượng trong các phần mềm Excel, Statgraphics hoặc ngay trên đồ thị của Excel.

Một số hàm phi tuyến phổ biến như:

$y = a.x^b$ tuyến tính hóa: ln(y) = ln(a) + b.ln(x)

$y = a.e^{bx}$ tuyến tính hóa: ln(y) = ln(a) + b.x

6.2.1. Lập mô hình hàm mũ trong Excel:

Ví dụ: Lập mô hình tương quan H/D rừng trồng Tếch dạng hàm mũ:

$H = a.D^b$

- **Tuyến tính hóa: Logarit neper 2 vế:**

Ln(H) = Ln(a) + b.Ln(D)

Đặt Y = Ln(H) X = Ln(D) A = Ln(a) B = b.

Vậy Y = A + B.X

- **Nhập số liệu và đổi biến số:**
  - Cột A: Số liệu D.
  - Cột B: Số liệu H.
  - Cột C: Ln(D). Tại ô C2: =Ln(A2), copy cho cả cột.
  - Cột D: Ln(H). Tại ô D2: =Ln(B2), copy cho cả cột.

Số liệu H/D và đổi biến số

	A	B	C	D
1	**D(cm)**	**H(m)**	**Ln(D)**	**Ln(H)**
2	31,3	22,0	3,443863	3,091042
3	32,0	21,8	3,466237	3,08191
...	...	....	.....	.....
....	....	....	.....	.....
40	12,6	9,7	2,536373	2,270804
41	13,9	9,8	2,629481	2,277972

Ước lượng tương quan hồi quy đường thẳng trong Excel:

- Tools/Data Analysis/Regression. OK.
- Hộp thoại:

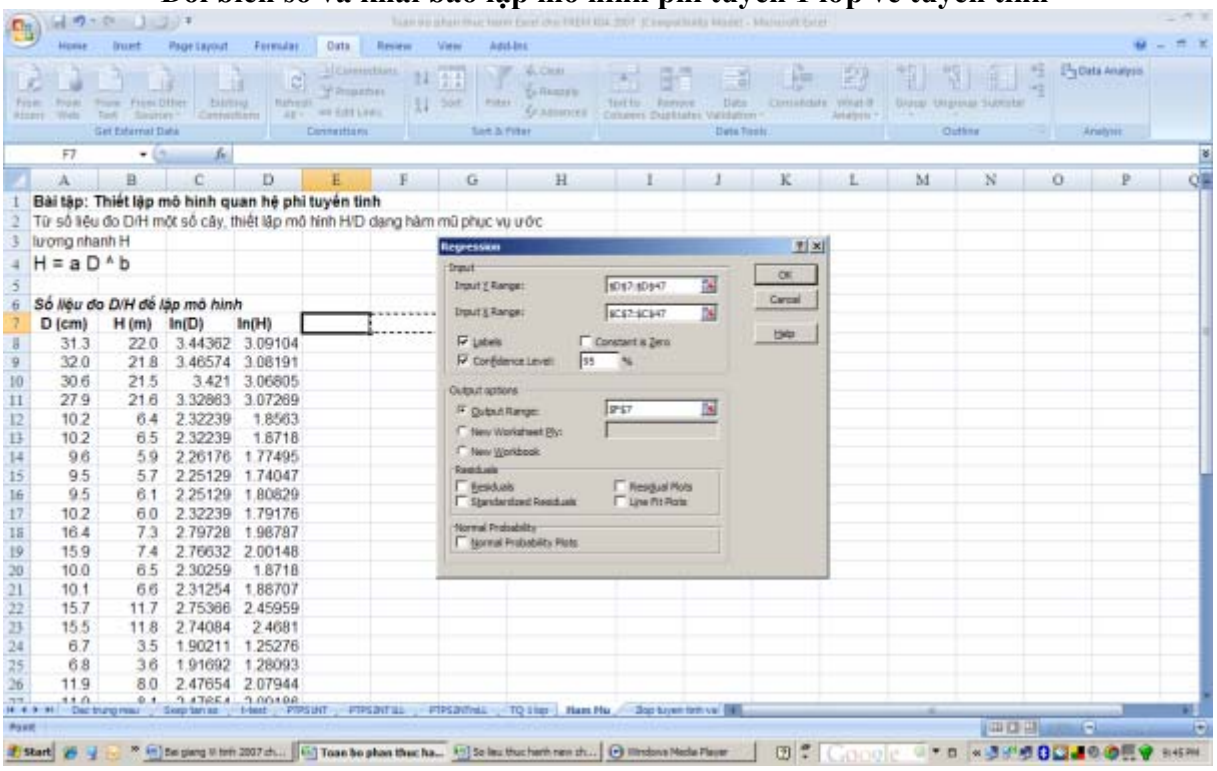
Input Y range: Nhập địa chỉ cột biến Y (Có thể nhập cả nhãn). Vd: D1:D41.

Input X range: Nhập địa chỉ cột biến X (Có thể nhập cả nhãn). Vd: C1:C41.

Label: Đánh dấu nếu đã nhập cả hàng đầu làm nhãn.

Output range: Nhập địa chỉ ô trên trái nơi xuất kết quả.
Kích OK.

Đổi biến số và khai báo lập mô hình phi tuyến 1 lớp về tuyến tính



D (cm)	H (m)	ln(D)	ln(H)
31.3	22.0	3.44362	3.09104
32.0	21.8	3.46574	3.08191
30.6	21.5	3.421	3.06805
27.9	21.6	3.32863	3.07269
10.2	6.4	2.32239	1.8563
10.2	6.5	2.32239	1.8718
9.6	5.9	2.26176	1.77495
9.5	5.7	2.25129	1.74047
9.5	6.1	2.25129	1.80829
10.2	6.0	2.32239	1.79176
16.4	7.3	2.79728	1.98787
15.9	7.4	2.76632	2.00148
10.0	6.5	2.30259	1.8718
10.1	6.6	2.31254	1.88707
15.7	11.7	2.75386	2.45959
15.5	11.8	2.74084	2.4681
6.7	3.5	1.90211	1.25276
6.8	3.6	1.91692	1.28093
11.9	8.0	2.47654	2.07944

Kết quả ước lượng hồi quy tuyến tính

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,940772849
R Square	0,885053553
Adjusted R Square	0,882028647
Standard Error	0,166400069
Observations	40

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	8,101484412	8,101484	292,5887	1,92186E-19
Residual	38	1,052181354	0,027689		
Total	39	9,153665766			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-0,78748559	0,182988537	-4,30347	0,000114	-1,157926531	-0,417044653
Ln(D)	1,153364313	0,067427602	17,10523	1,92E-19	1,016864265	1,289864361

Phương trình tương quan:

Ln(H) = -0.787 + 1.153Ln(D)

Với N = 40 R = 0.941 Fr = 292.59 với α<0.0000, nên R tồn tại

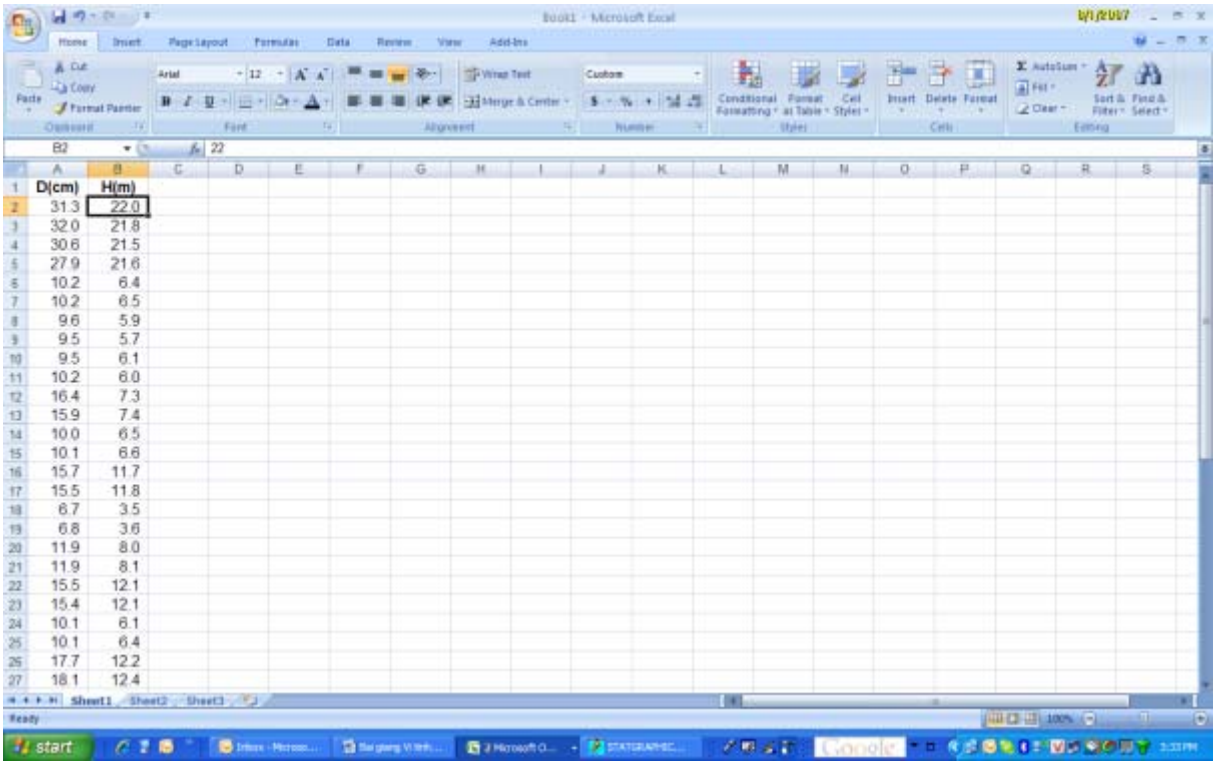
Đưa về dạng nguyên thủy: Tính a = exp(A) = exp(-0.787) = 0.455

Vậy: $\mathbf{H = 0.455.D^{1.153}}$

6.2.2. Lập mô hình hàm mũ một lớp trong Statgraphics:

Trong Statgraphics, việc ước lượng mô hình phi tuyến tính đơn giản hơn vì không cần tạo thêm các cột đổi biến số, biến số được đổi trực tiếp trong hộp thoại khi thiết lập mô hình.

Đầu tiên nhập dữ liệu trong Excel với hai cột x và y, ví dụ là D và H như sau

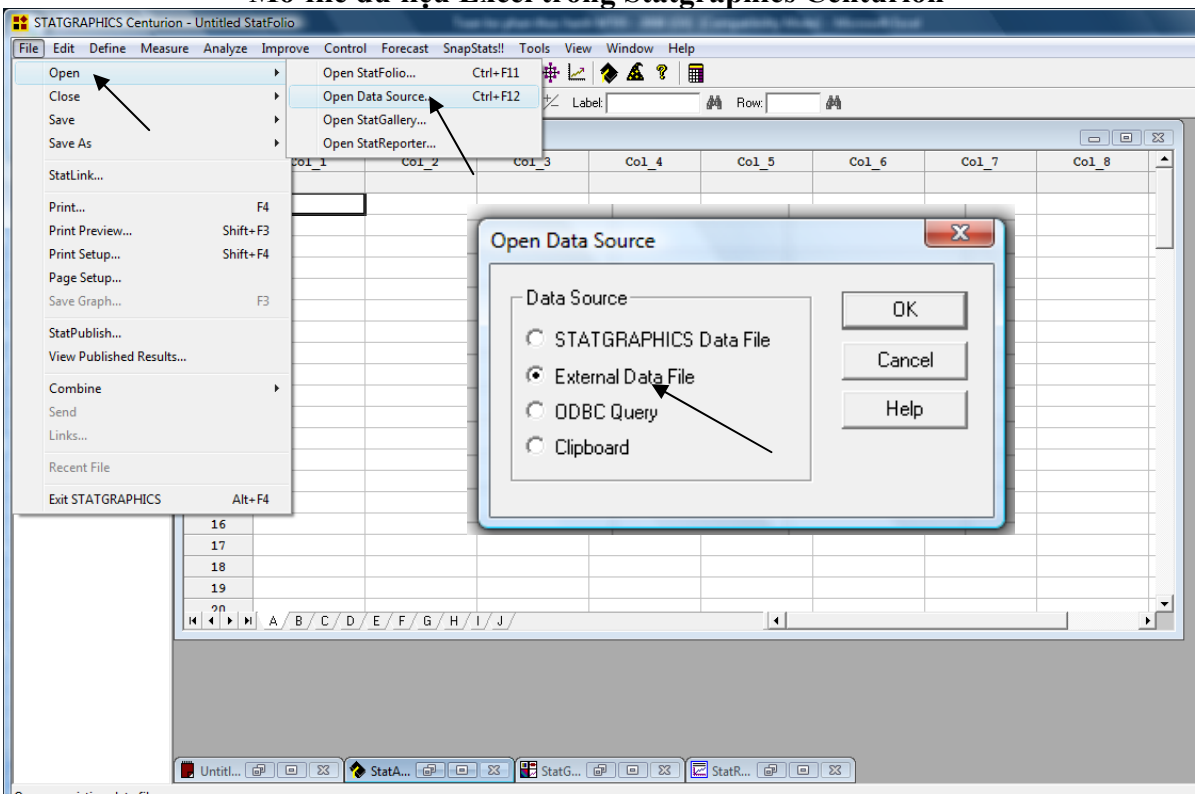


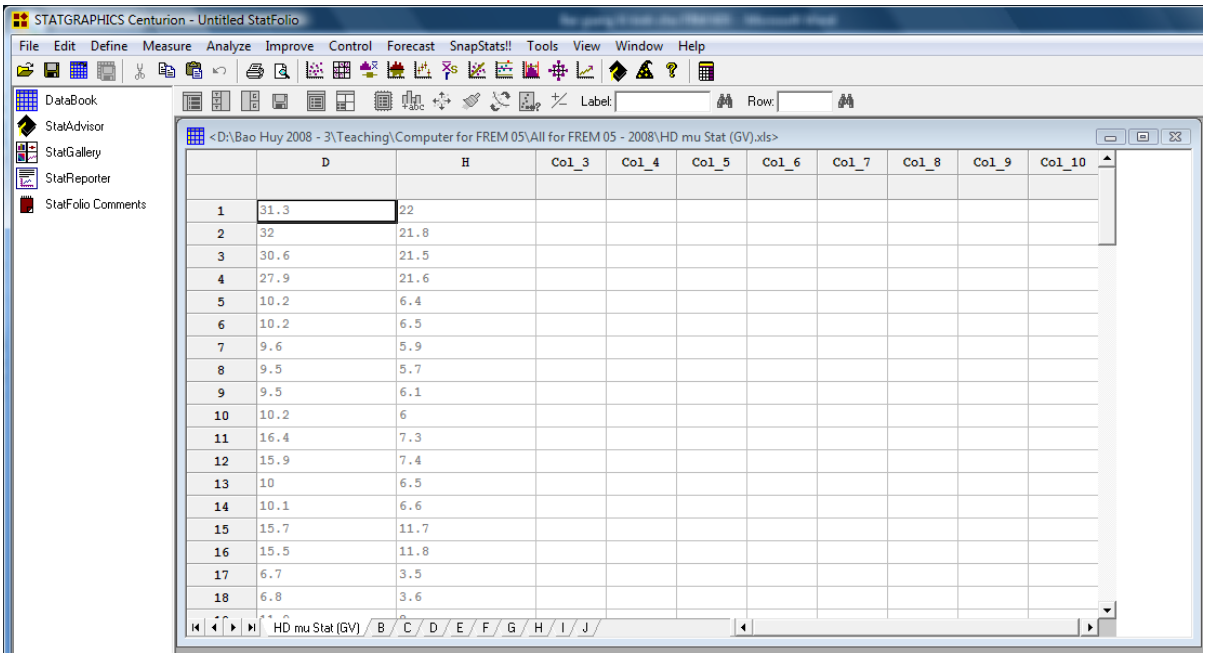
	A	B
1	D(cm)	H(m)
2	31.3	22.0
3	32.0	21.8
4	30.6	21.5
5	27.9	21.6
6	10.2	6.4
7	10.2	6.5
8	9.6	5.9
9	9.5	5.7
10	9.5	6.1
11	10.2	6.0
12	16.4	7.3
13	15.9	7.4
14	10.0	6.5
15	10.1	6.6
16	15.7	11.7
17	15.5	11.8
18	6.7	3.5
19	6.8	3.6
20	11.9	8.0
21	11.9	8.1
22	15.5	12.1
23	15.4	12.1
24	10.1	6.1
25	10.1	6.4
26	17.7	12.2
27	18.1	12.4

File dữ liệu Excel cần được lưu với version của Microsoft Excel 97-2003 về trước, vì Statgraphics chưa nhận được kiểu file MS. Office 2007

Sau đó mở file dữ liệu này trong Statgraphics Centurion: File/Open/Open Data Source/External Data file - OK

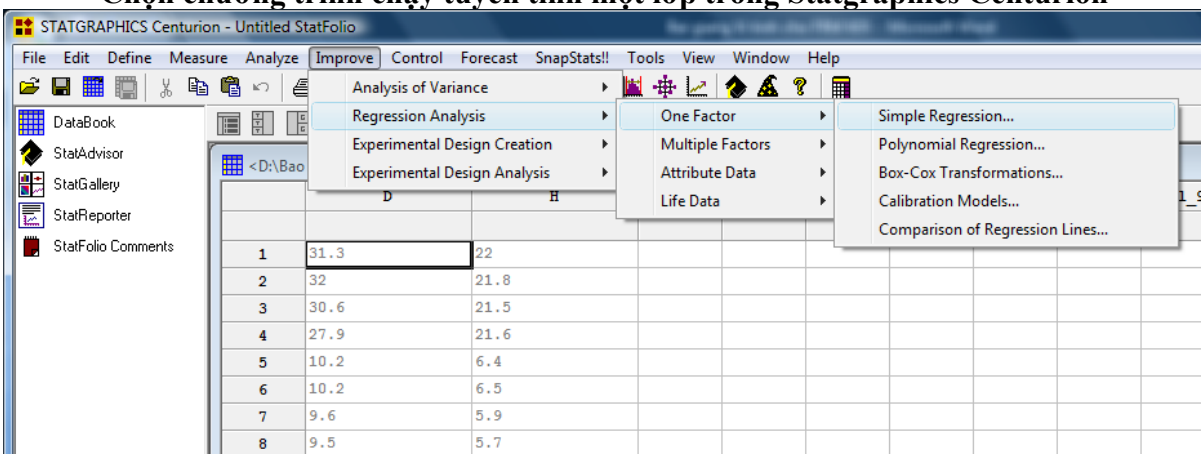
Mở file dữ liệu Excel trong Statgraphics Centurion



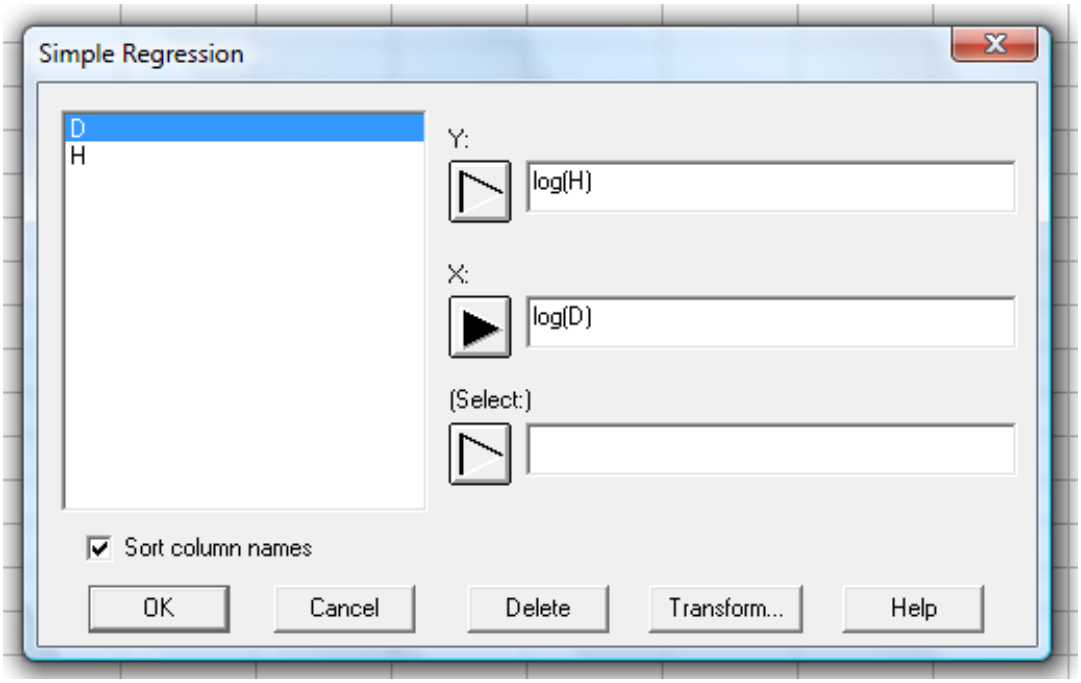


Chạy phần xử lý hàm tương quan một lớp: Improve/Regression Analysis/One Factor/Simple Regression

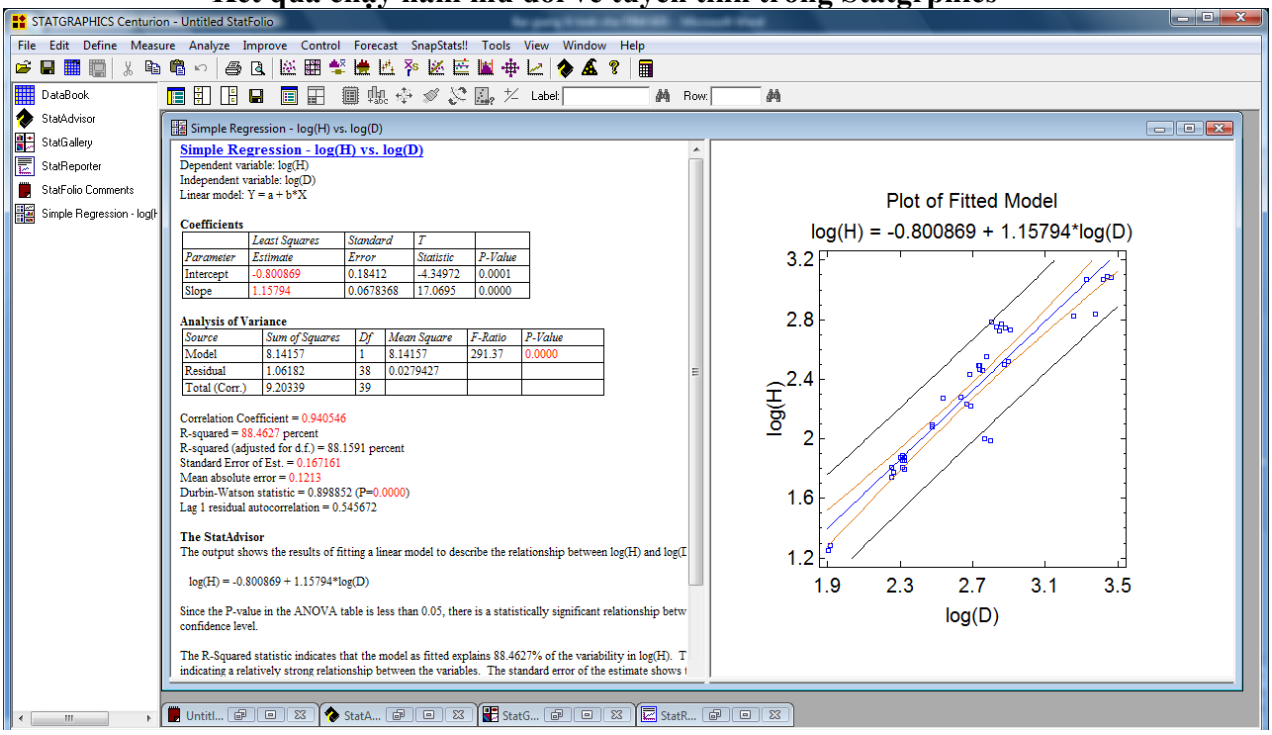
Chọn chương trình chạy tuyến tính một lớp trong Statgraphics Centurion



Trong hộp thoại chọn biến y và x và đổi biến số ngay trong hộp thoại: log(H) và log(D). Kích OK để có kết quả. (Lưu ý ký hiệu log trong Statgraphics là logarit neper)



Kết quả chạy hàm mũ đổi về tuyến tính trong Statgrphics



Simple Regression - log(H) vs. log(D)

Dependent variable: log(H)
Independent variable: log(D)
Linear model: Y = a + b*X

Coefficients

	Least Squares	*Standard*	*T*	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*
Intercept	-0.800869	0.18412	-4.34972	0.0001
Slope	1.15794	0.0678368	17.0695	0.0000

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	8.14157	1	8.14157	291.37	0.0000
Residual	1.06182	38	0.0279427		
Total (Corr.)	9.20339	39			

Correlation Coefficient = 0.940546
R-squared = 88.4627 percent
R-squared (adjusted for d.f.) = 88.1591 percent
Standard Error of Est. = 0.167161
Mean absolute error = 0.1213
Durbin-Watson statistic = 0.898852 (P=0.0000)
Lag 1 residual autocorrelation = 0.545672

The StatAdvisor
The output shows the results of fitting a linear model to describe the relationship between log(H) and log(D). The equation of the fitted model is

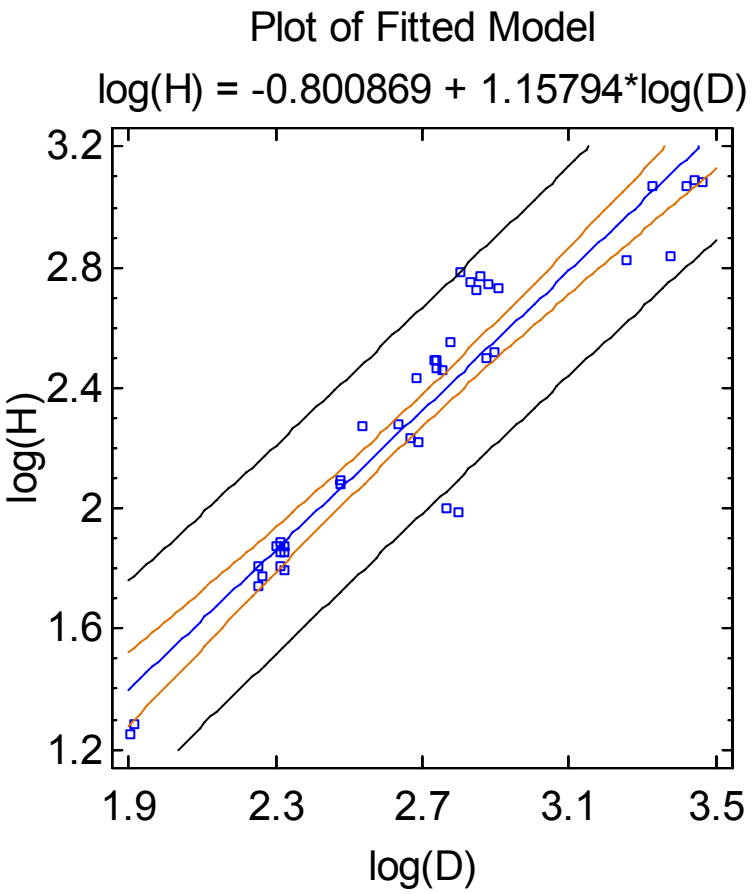
log(H) = -0.800869 + 1.15794*log(D)

Since the P-value in the ANOVA table is less than 0.05, there is a statistically significant relationship between log(H) and log(D) at the 95.0% confidence level.

The R-Squared statistic indicates that the model as fitted explains 88.4627% of the variability in log(H). The correlation coefficient equals 0.940546, indicating a relatively strong relationship between the variables. The standard error of the

estimate shows the standard deviation of the residuals to be 0.167161. This value can be used to construct prediction limits for new observations by selecting the Forecasts option from the text menu.

The mean absolute error (MAE) of 0.1213 is the average value of the residuals. The Durbin-Watson (DW) statistic tests the residuals to determine if there is any significant correlation based on the order in which they occur in your data file. Since the P-value is less than 0.05, there is an indication of possible serial correlation at the 95.0% confidence level. Plot the residuals versus row order to see if there is any pattern that can be seen.



Kết quả cho ra hàm trực tiếp viết dưới dạng tuyến tính đã đổi biến số
Các kết quả kiểm tra hệ số tương quan R và các biến số được hiểu giống như trong Excel

6.3. Ước lượng các dạng hồi quy một lớp tuyến tính hoặc phi tuyến tính trên đồ thị

Trong thực tế trực quan các mối quan hệ, người ta thường dùng đồ thị để biểu diễn, và để dễ dàng trong việc xem xét các sự báo, Excel hỗ trợ chương trình xác định mô hình hồi quy tuyến tính và phi tuyến tính một lớp ngay trên đồ thị. Excel lập sẵn 5 dạng hàm phổ biến trong phần này.
Ví dụ: Lựa chọn mô hình hồi quy H/D cho rừng trồng Tếch ngay trên đồ thị quan hệ

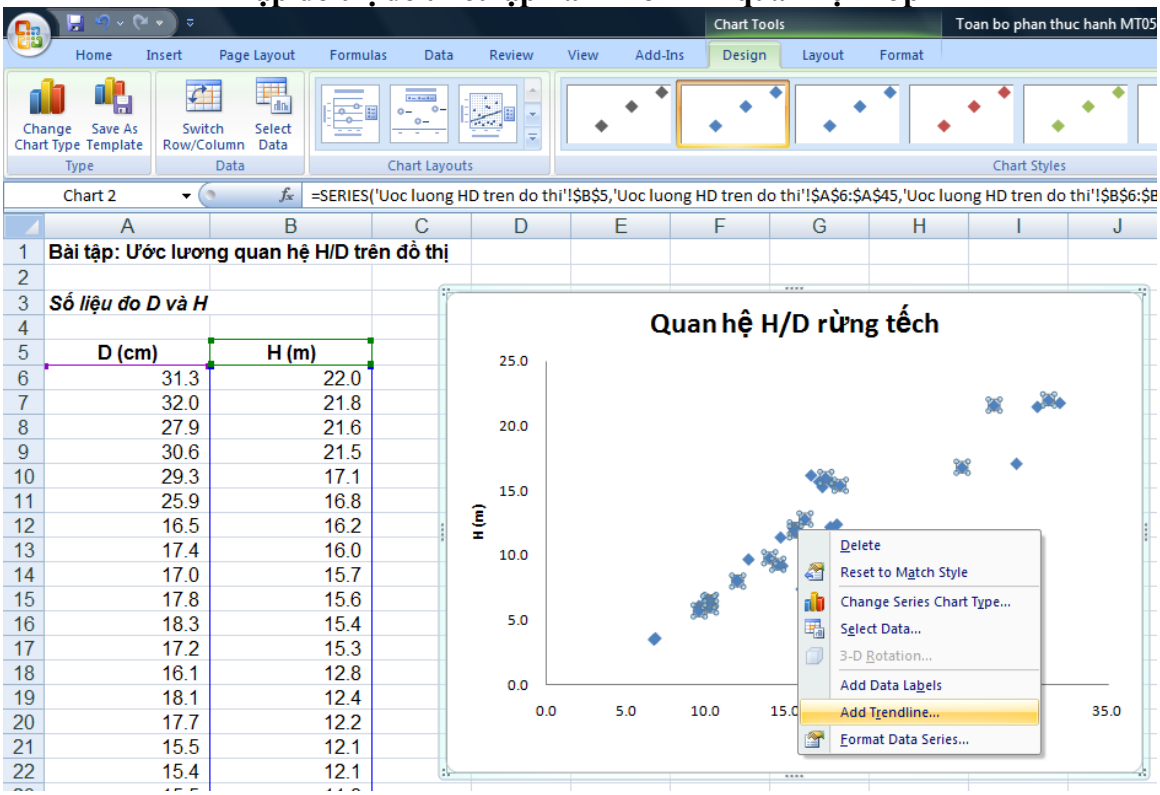
- **Nhập số liệu:**

Số liệu về quan hệ H/D

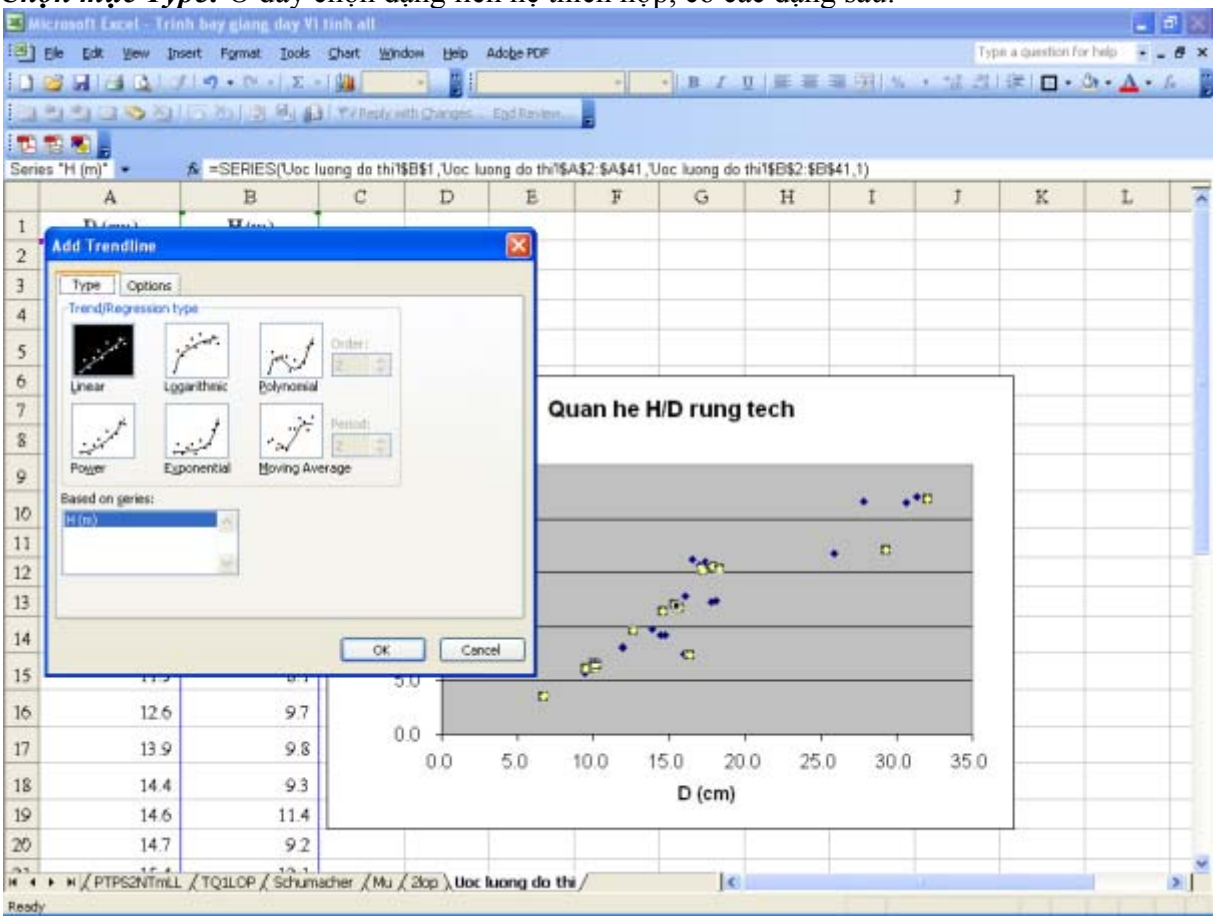
	A	B
1	**D(cm)**	**H(m)**
2	6,7	3,5
3	6,8	3,6
4	9,5	5,7
5	9,5	6,1
...	...	...
40	31,3	22,0
41	32,0	21,8

- **Vẽ đồ thị:** Tiến hành các bước vẽ đồ thị quan hệ H/D. (Nên vẽ dạng đám mây điểm).
- **Tính toán mô hình quan hệ dựa vào đồ thị:**
  - Kích hoạt đồ thị: Kích chuột trái.
  - Chọn đám mây điểm trên đồ thị: Kích chuột phải vào đám mây đểm này.
  - Chọn Add Trendline

Lập đồ thị để thiết lập hàm mô hình quan hệ 1 lớp



Chọn mục Type: Ở đây chọn dạng liên hệ thích hợp, có các dạng sau:



Linear: $y = mx + b$

Logarithmic: $y = c\ln x + b$

Polynomial: $y = b + c_1x + c_2x^2 + ...c_6x^6$

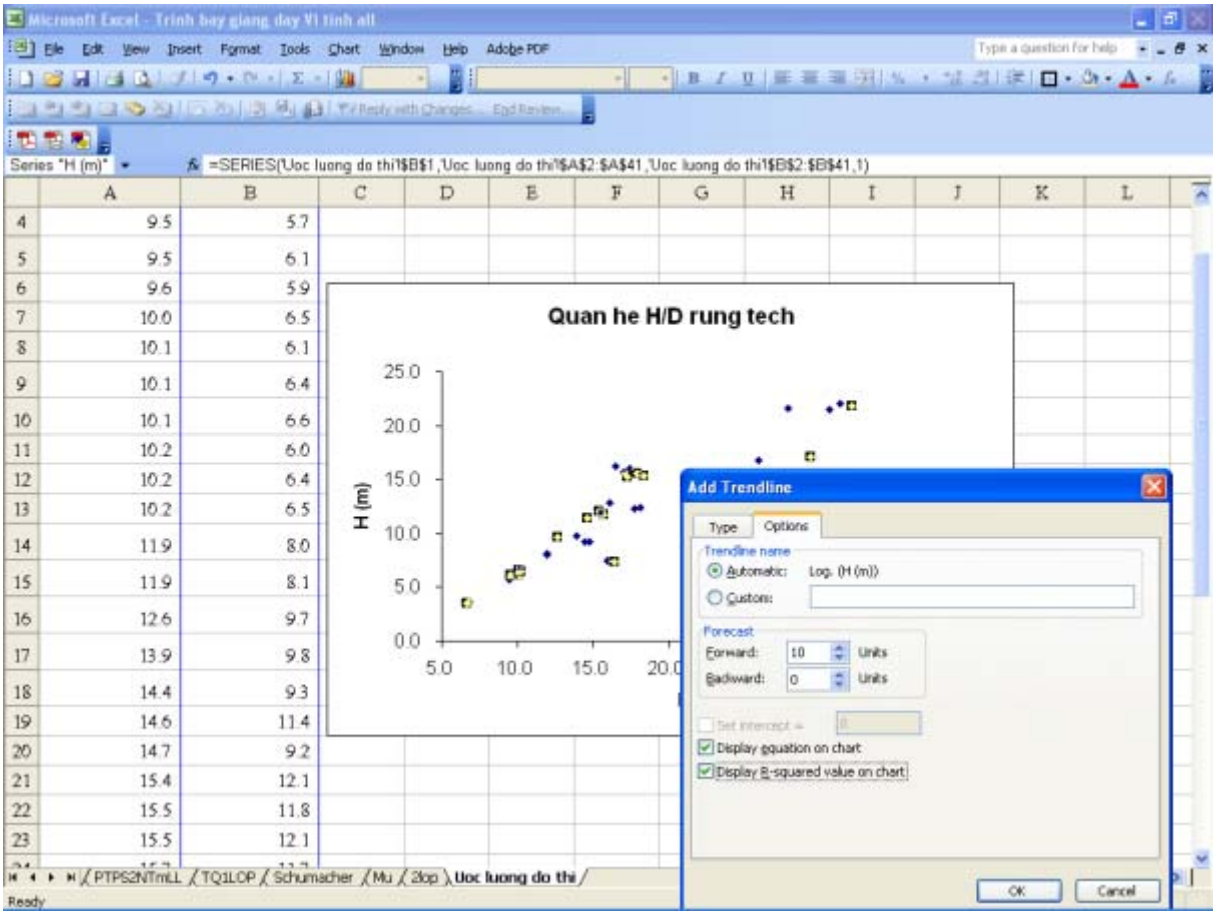
Có thể chọn 1 đến 6 bậc trong ô Order: Xác định số bậc.

Power: $y = cx^b$

Exponential: $y = c.e^{bx}$

Chọn các kiểu hàm khác nhau để có được R^2 lớn nhất.

Chọn mục Option: Xác định:



Forecast: Foward: Xác định độ dài dự đoán tiếp theo.

Backward: Xác định độ dài dự đoán lùi.

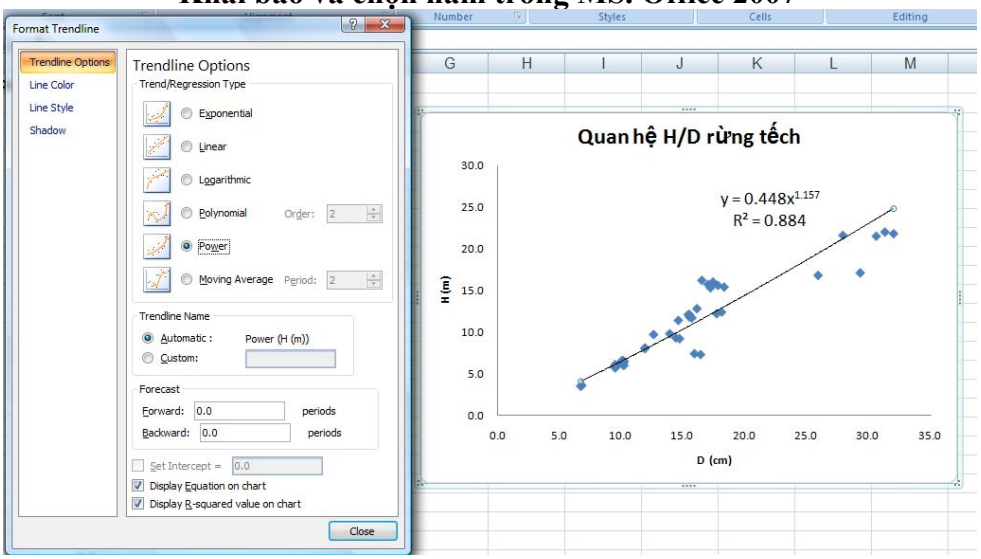
Set intercept (0): Nếu đánh dấu thì tham số b=0 trong các hàm đường thẳng

Display Equation on Chart: Đánh dấu để đưa hàm lên đồ thị.

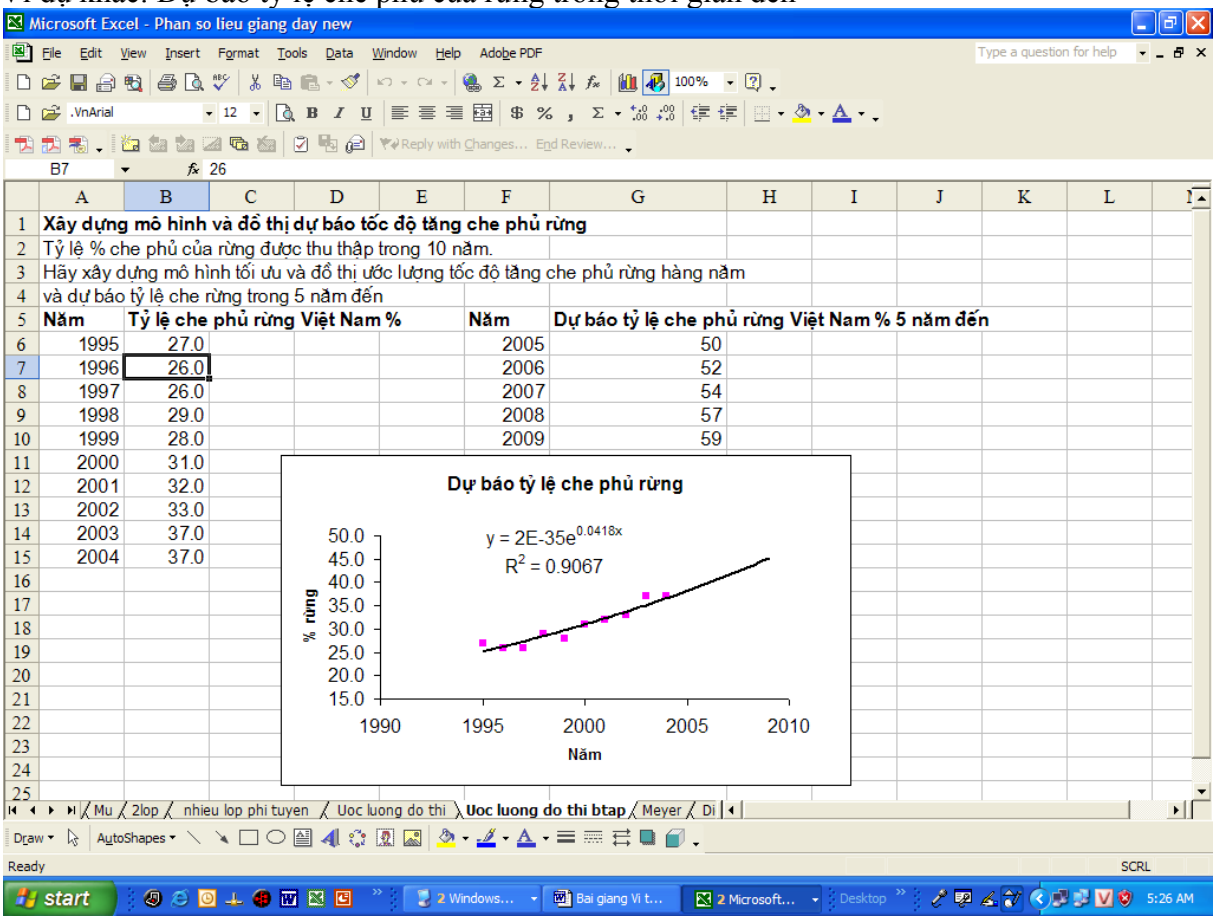
Display R-squared Value on Chart: Đánh dấu nếu muốn tính hệ số tương quan bình phương.

Cuối cùng là OK.

Khai báo và chọn hàm trong MS. Office 2007



Ví dụ khác: Dự báo tỷ lệ che phủ của rừng trong thời gian đến

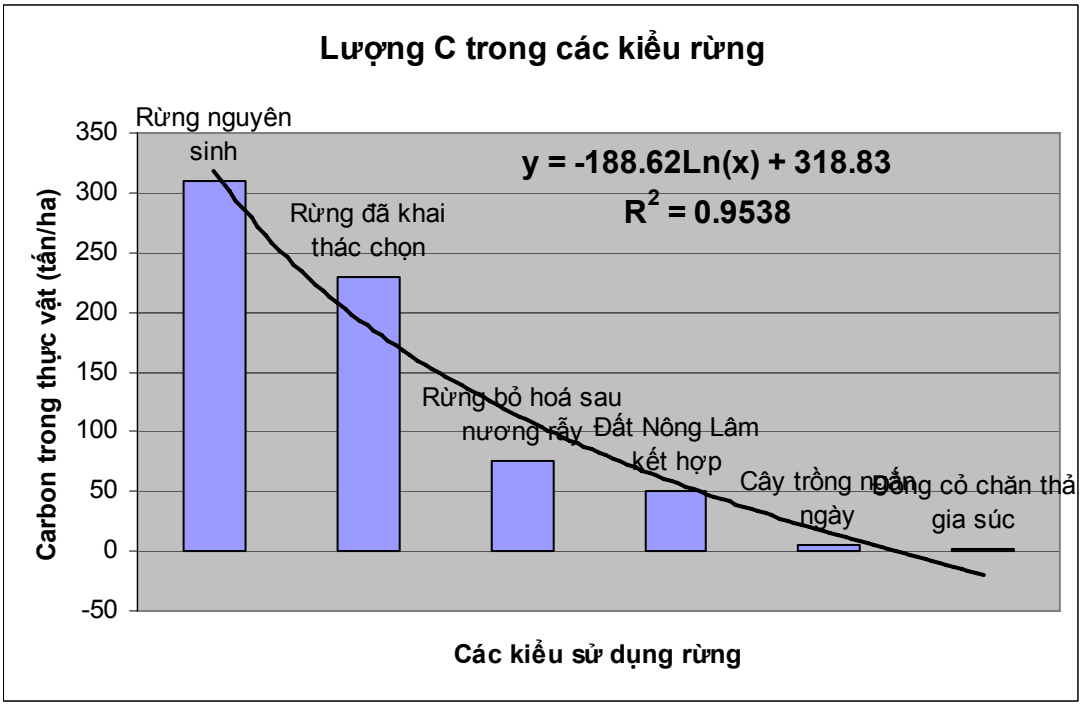


Ví dụ khác: Lượng carbon được lưu trữ trong các kiểu rừng khác nhau được mô phỏng bằng dạng hàm phi tuyến trên đồ thị. Trong đó không cần mã hóa biến số x (kiểu rừng), lúc này sử dụng sơ đồ cột để vẽ và chạy phương trình thích hợp. Lúc này máy tính đã tự động mà hóa các kiểu rừng là 1, 2, 3, 4 .....

Lượng carbon trên và dưới mặt đất ở các kiểu sử dụng đất rừng

Các vùng rừng ở Brazil, Cameroon và Indonesia

Các kiểu rừng	**Lượng carbon (tấn/ha)**	
	Dưới mặt đất	Trong thực vật
Rừng nguyên sinh	48	310
Rừng đã khai thác chọn	48	230
Rừng bỏ hoá sau nương rẫy	48	75
Đất Nông Lâm kết hợp	45	50
Cây trồng ngắn ngày	25	5
Đồng cỏ chăn thả gia súc	20	2



6.4. Hồi quy tuyến tính nhiều lớp

Trong thực tế biến phụ thuộc Y bị chi phối bởi nhiều biến số độc lập Xi. Ví dụ như trữ lượng rừng được đóng góp bởi nhiều nhân tố như mật độ, tiết diện ngang, chiều cao, cấp đất; hoặc biến đổi dòng chảy, mức độ xung yếu của lưu vực bị chi phối bởi nhiều nhân tố như lượng mưa, độ dốc, địa hình, loài đất, che phủ thảm thực vật, ...

Trong trường hợp này để ước lượng biến phụ thuộc Y người ta cần lập mô hình hồi quy nhiều biến số để có thể phản ảnh chính xác giá trị ước lượng, dự báo Y.

Dạng phương trình tổng quát:

$$\mathbf{Y = a_o + b_1X_1 + b_2X_2 + ... + b_nX_n}$$

Ví dụ: Thiết lập mô hình dự đoán trữ lượng rừng (M) Tếch theo 2 biến số mật độ (N) và chiều cao bình quân (H) theo dạng tuyến tính 2 lớp:

$$M = a + b_1 N + b_2 H$$

Đây là dạng tuyến tính 2 lớp $Y = a + b_1X_1 + b_2X_2$

Dùng phương pháp bình phương tối thiểu ước lượng phương trình

- **Nhập số liệu**

Bảng số liệu M/N/H

	A	B	C
1	**N(c/ha)**	**H(m)**	**M(m3/ha)**
2	180	23,0	163,452
3	170	23,0	160,154
4	220	22,3	184,167
...	....	....	.....
...	....	....	.....

	A	B	C
1	**N(c/ha)**	**H(m)**	**M(m3/ha)**
40	570	10,9	43,846
41	570	11,1	53,212

- **Ước lượng tương quan tuyến tính nhiều lớp:**
  - Tools/Data Analysis/Regression.OK. (Data/Data Analysis/Regression trong MS Office 2007)
  - Hộp thoại:

Input Y range: Nhập địa chỉ cột biến Y (Có thể nhập cả nhãn). Vd: C1:C41.

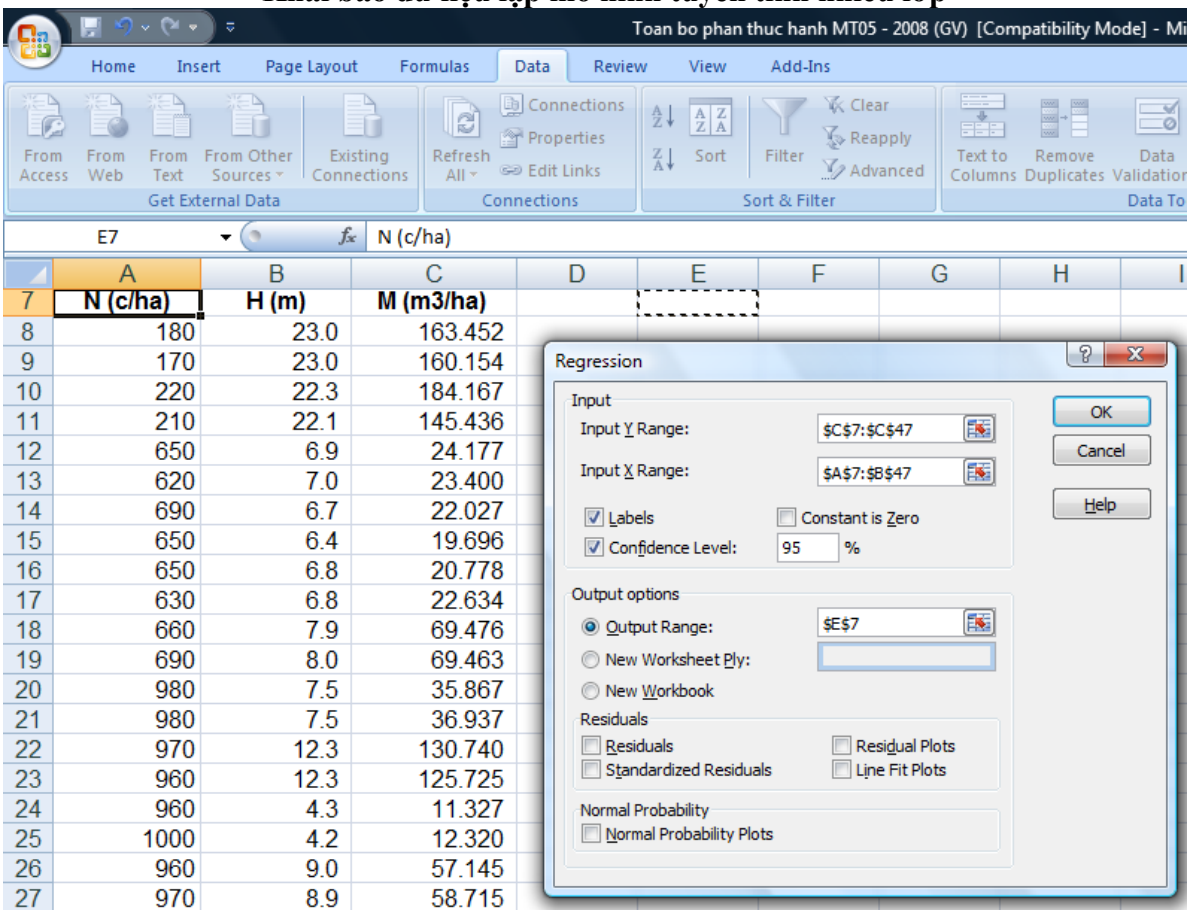
Input X range: Nhập địa chỉ khối các biến X (Có thể nhập cả nhãn). Vd: A1:B41.

Label: Đánh dấu nếu đã nhập cả hàng đầu làm nhãn.

Output range: Nhập địa chỉ ô trên trái nơi xuất kết quả.

OK.

Khai báo dữ liệu lập mô hình tuyến tính nhiều lớp



Kết quả ước lượng mô hình hồi quy tuyến tính 2 lớp

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.9256776
R Square	0.856879
Adjusted R Square	0.8491427
Standard Error	28.140919
Observations	40

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	175426.2	87713.1	110.7613	2.40166E-16
Residual	37	29300.72	791.9113		
Total	39	204726.9			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-154.77144	22.13662	-6.99165	2.91E-08	-199.6244851	-109.918392
N (c/ha)	0.1095484	0.016994	6.446152	1.57E-07	0.075114494	0.143982284
H (m)	14.52156	0.97677	14.86692	3.49E-17	12.54243676	16.50068344

Phương trình tương quan hồi quy:
M = - 154.771 + 0.109 N + 14.521 H
Với N = 40 R = 0.926 Fr = 110.76 với P<0.00
tb_1 = 6.44 tb_2 = 14.86 với P<0.00

Lưu ý quan trọng: Khi phân tích mô hình nhiều lớp, ngoài việc kiểm tra sự tồn tại của hệ số tương quan R bằng tiêu chuẩn F, với R tồn tại khi Significance F (P) < 0.05; đồng thời phải kiểm tra sự tồn tại của các tham số gắn các biến số Xi bằng tiêu chuẩn tstat, tham số tồn tại khi P-value < 0.05. (Thể hiện trong kết quả ở bảng cuối cùng). Nếu một tham số không tồn tại thì có nghĩa: i) Biến số (nhân tố) đó không ảnh hưởng đến Y, lúc này cần loại biến đó khỏ mô hình; hoặc dạng đường thẳng là chưa thích hợp (lúc này phải chuyển sang dạng phi tuyến để xem xét sự ảnh hưởng của nhân tố này)
Trong trường hợp trên hai biến N và H ảnh hưởng rõ ràng đến M ở dạng đường thẳng, với P <0.05 rất nhiều.

6.5. Hồi quy phi tuyến tính nhiều lớp, tổ hợp biến

Trong trường hợp nhiều biến số xi ảnh hưởng đến y không theo dạng tuyến tính mà có dạng quan hệ phi tuyến, trường hợp này cần đổi biến số để trở về dạng tuyến tính, hoặc lập mô hình tổ hợp biến.
Một số dạng phi tuyến nhiều lớp phổ biến và cách quy về tuyến tính hoặc tổ hợp biến:

$y = a.x1^{b1}\ x2^{b2}....xn^{bn}$ tuyến tính hóa: ln(y) = ln(a) + b1.ln(x1) + b2ln(x2) + ... + bnln(xn)

$y = a.e^{b1x1+b2x2+...+bnxn}$ tuyến tính hóa: ln(y) = ln(a) + b1x1 + b2x2 + ...+ bnxn

Hoặc dạng tổ hợp biến và đổi biến số kết hợp:
ln(y) = a + b1.log(x1.x2) + b2 exp(x3/x4) + ...

Trong Statgraphics, việc tính toán mô hình kiểu này rất đơn giản vì không cần tạo thêm các cột đổi biến số, biến số được đổi trực tiếp trong hộp thoại khi thiết lập mô hình.

Các bước tiến hành như sau:

i) Kiểm tra dạng chuẩn của mỗi biến số, nếu chưa chuẩn phải đổi biến số để đưa về chuẩn (log(x), 1/x, sqrt(x), exp(x), ...)

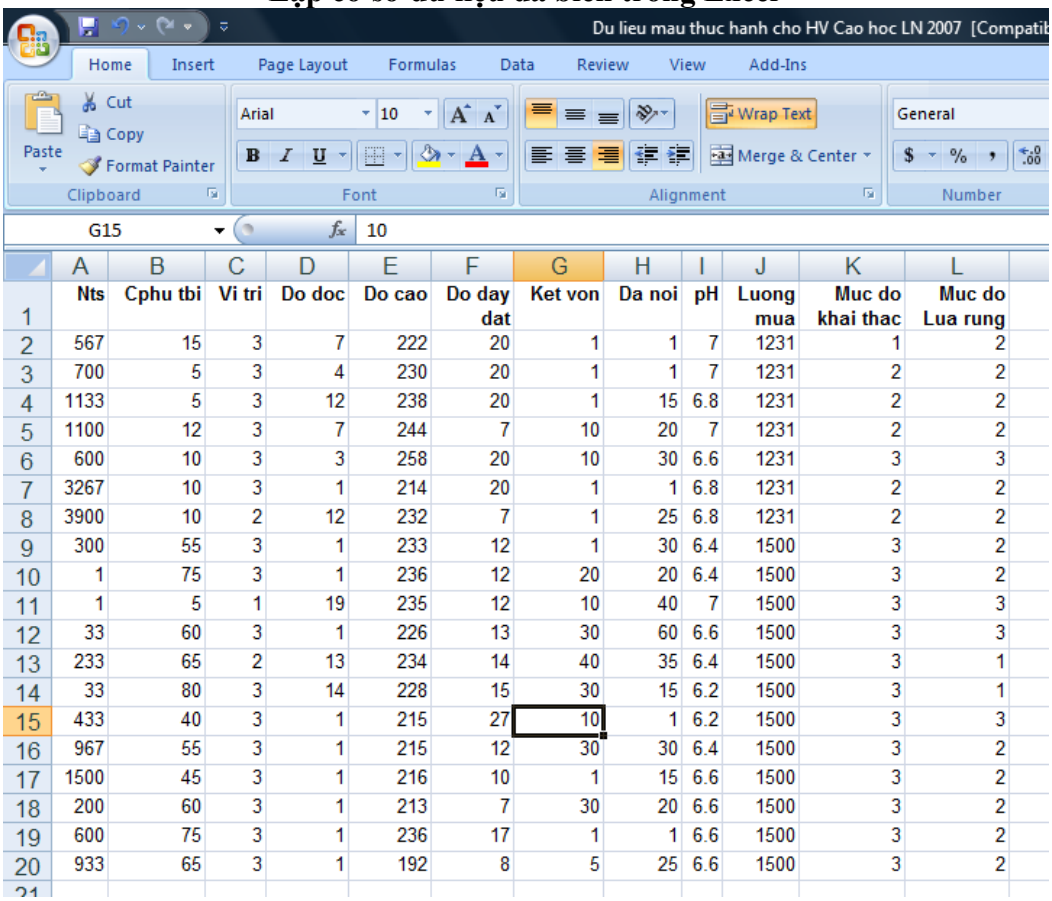
ii) Chọn biến số xi có ảnh hưởng đến y

iii) Chạy mô hình tuyến tính nhiều lớp được đổi biến số, khi cần thiết phải tổ hợp biến nếu các biến xi có quan hệ với nhau

iv) Kiểm tra mô hình: Hệ số xác định R^2 có P <0.05 và các tham số gắn biến số qua kiểm tra theo t phải có P < 0.05. Nếu một biến số chưa bảo đảm P <0.05 thì phải loại khỏi mô hình hoặc đổi biến số, hoặc tổ hợp với biến số khác.

Đầu tiên lập cơ sở dữ liệu trong Excel, bao gồm các trường (cột) biến y và xi, ví dụ nghiên cứu để phát hiện các nhân tố sinh thái nhân tác đa biến ảnh hưởng đến tái sinh rừng; biến y là mật độ tái sinh (Ntx), biến xi bao gồm nhiều nhân tố sinh thái và nhân tác

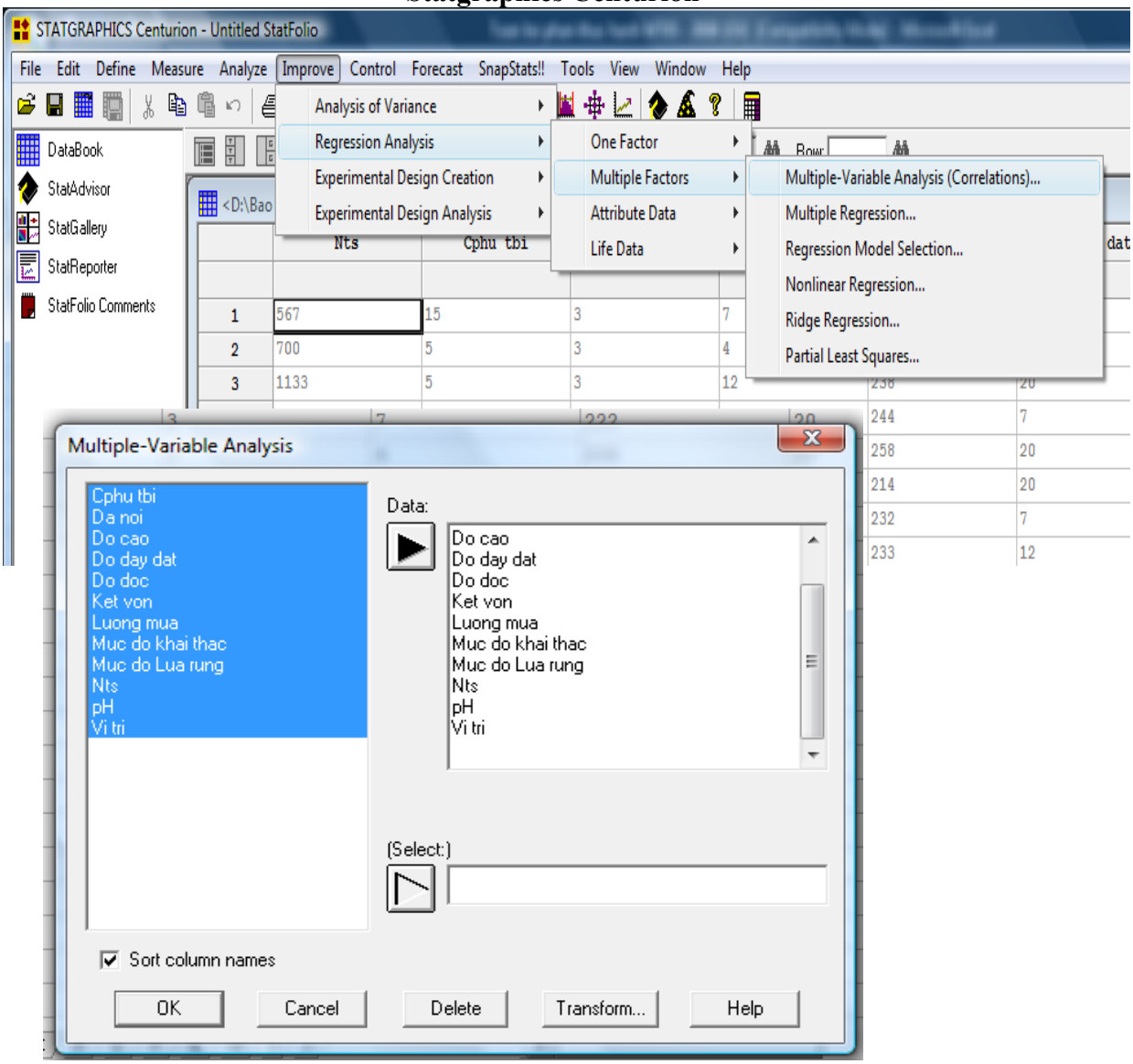
Lập cơ sở dữ liệu đa biến trong Excel



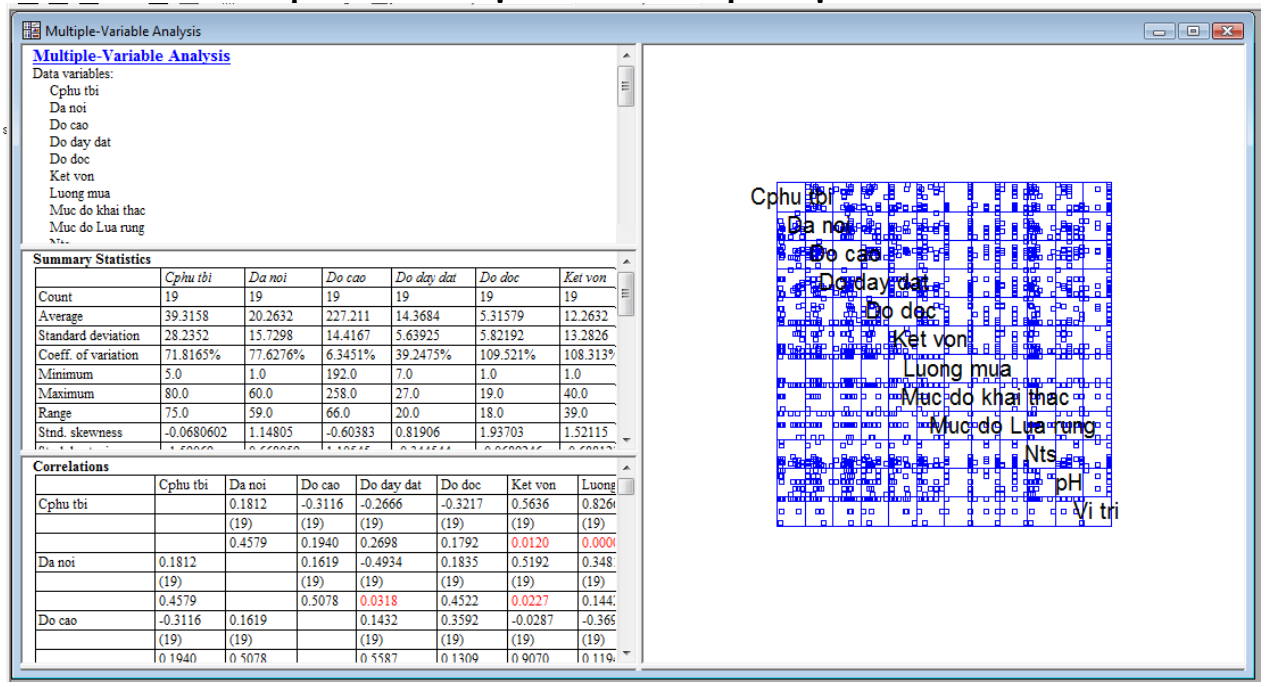
	A	B	C	D	E	F	G	H	I	J	K	L
1	Nts	Cphu tbi	Vi tri	Do doc	Do cao	Do day dat	Ket von	Da noi	pH	Luong mua	Muc do khai thac	Muc do Lua rung
2	567	15	3	7	222	20	1	1	7	1231	1	2
3	700	5	3	4	230	20	1	1	7	1231	2	2
4	1133	5	3	12	238	20	1	15	6.8	1231	2	2
5	1100	12	3	7	244	7	10	20	7	1231	2	2
6	600	10	3	3	258	20	10	30	6.6	1231	3	3
7	3267	10	3	1	214	20	1	1	6.8	1231	2	2
8	3900	10	2	12	232	7	1	25	6.8	1231	2	2
9	300	55	3	1	233	12	1	30	6.4	1500	3	2
10	1	75	3	1	236	12	20	20	6.4	1500	3	2
11	1	5	1	19	235	12	10	40	7	1500	3	3
12	33	60	3	1	226	13	30	60	6.6	1500	3	3
13	233	65	2	13	234	14	40	35	6.4	1500	3	1
14	33	80	3	14	228	15	30	15	6.2	1500	3	1
15	433	40	3	1	215	27	10	1	6.2	1500	3	3
16	967	55	3	1	215	12	30	30	6.4	1500	3	2
17	1500	45	3	1	216	10	1	15	6.6	1500	3	2
18	200	60	3	1	213	7	30	20	6.6	1500	3	2
19	600	75	3	1	236	17	1	1	6.6	1500	3	2
20	933	65	3	1	192	8	5	25	6.6	1500	3	2

Kiểm tra dạng chuNn của các biến số trong Statgrahics và định hướng đổi biến số: Improve/Regression Analysis/Multiple Factors/Multiple Variable Analysis. Sau đó đưa tất cả biến y và xi vào hộp thoại data.

Chọn chương trình kiểm tra luật chuẩn và định hướng đổi biến số để chuẩn hóa trong Statgraphics Centurion



Kết quả kiểm tra luật chuẩn và mối quan hệ các biến số



- **Kết quả kiểm tra phân bố chuẩn của các biến số:**

- **Summary Statistics**

	Cphu tbi	*Da noi*	*Do cao*	*Do day dat*	*Do doc*	*Ket von*	*Luong mua*	*Muc do khai thac*
Count	19	19	19	19	19	19	19	19
Average	39.3158	20.2632	227.211	14.3684	5.31579	12.2632	1400.89	2.63158
Standard deviation	28.2352	15.7298	14.4167	5.63925	5.82192	13.2826	133.315	0.597265
Coeff. of variation	71.8165%	77.6276%	6.3451%	39.2475%	109.521%	108.313%	9.51641%	22.6961%
Minimum	5.0	1.0	192.0	7.0	1.0	1.0	1231.0	1.0
Maximum	80.0	60.0	258.0	27.0	19.0	40.0	1500.0	3.0
Range	75.0	59.0	66.0	20.0	18.0	39.0	269.0	2.0
Stnd. skewness	-0.0680602	1.14805	-0.60383	0.81906	1.93703	1.52115	-1.05608	-2.56858
Stnd. kurtosis	-1.59069	0.668059	1.10545	-0.344544	-0.0689246	-0.688123	-1.65147	1.22788

-

	Muc do Lua rung	*Nts*	*pH*	*Vi tri*
Count	19	19	19	19
Average	2.10526	868.474	6.63158	2.78947
Standard deviation	0.567131	1054.29	0.260454	0.535303
Coeff. of variation	26.9387%	121.395%	3.92748%	19.1901%
Minimum	1.0	1.0	6.2	1.0
Maximum	3.0	3900.0	7.0	3.0
Range	2.0	3899.0	0.8	2.0
Stnd. skewness	0.0906087	3.63749	0.0232827	-4.72906
Stnd. kurtosis	0.52516	3.5476	-0.823423	6.1244

The StatAdvisor

This table shows summary statistics for each of the selected data variables. It includes measures of central tendency, measures of variability, and measures of shape. Of particular interest here are the standardized skewness and standardized kurtosis, which can be used to determine whether the sample comes from a normal distribution. Values of these statistics outside the range of -2 to +2 indicate significant departures from normality, which would tend to invalidate many of the statistical procedures normally applied to this data. In this case, the following variables show standardized skewness values outside the expected range:

Muc do khai thac
Nts
Vi tri

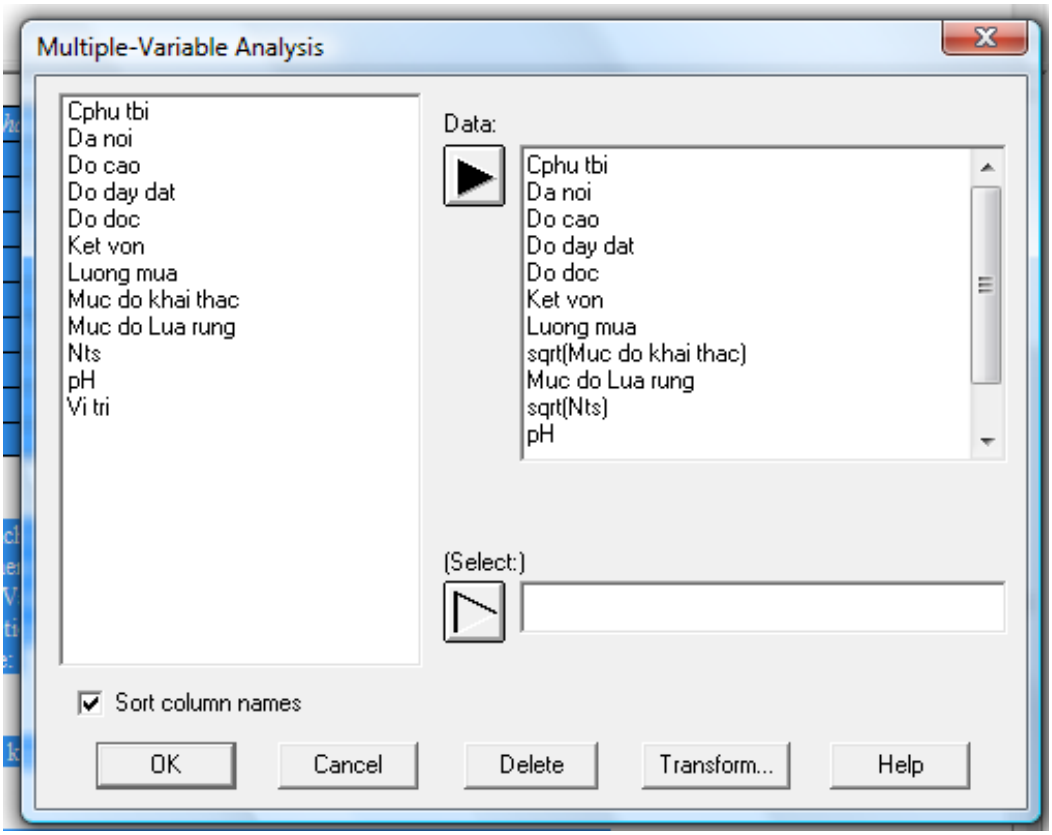
The following variables show standardized kurtosis values outside the expected range:

Nts
Vi tri

To make the variables more normal, you might try a transformation such as LOG(Y), SQRT(Y), or 1/Y.

Kết quả cho thấy có 3 biến số có ***Standardized Sk hoặc Ku*** không bảo đảm có phân bố chuNn là: ***Nts, Muc do khai thac và Vi tri.*** Và 3 biến này cần đổi biến số ở các dạng LOG(Y), SQRT(Y), or 1/Y để chuNn hóa.

Đổi biến số để chuẩn hóa



Summary Statistics

	Cphu tbi	*Da noi*	*Do cao*	*Do day dat*	*Do doc*	*Ket von*	*Luong mua*
Count	19	19	19	19	19	19	19
Average	39.3158	20.2632	227.211	14.3684	5.31579	12.2632	1400.89
Standard deviation	28.2352	15.7298	14.4167	5.63925	5.82192	13.2826	133.315
Coeff. of variation	71.8165%	77.6276%	6.3451%	39.2475%	109.521%	108.313%	9.51641%
Minimum	5.0	1.0	192.0	7.0	1.0	1.0	1231.0
Maximum	80.0	60.0	258.0	27.0	19.0	40.0	1500.0
Range	75.0	59.0	66.0	20.0	18.0	39.0	269.0
Stnd. skewness	-0.0680602	1.14805	-0.60383	0.81906	1.93703	1.52115	-1.05608
Stnd. kurtosis	-1.59069	0.668059	1.10545	-0.344544	-0.0689246	-0.688123	-1.65147

	sqrt(Muc do khai thac)	*Muc do Lua rung*	*sqrt(Nts)*	*pH*	*log(Vi tri)*
Count	19	19	19	19	19
Average	1.60988	2.10526	24.5836	6.63158	0.99811
Standard deviation	0.205131	0.567131	16.697	0.260454	0.273236
Coeff. of variation	12.742%	26.9387%	67.9193%	3.92748%	27.3753%
Minimum	1.0	1.0	1.0	6.2	0.0
Maximum	1.73205	3.0	62.45	7.0	1.09861
Range	0.732051	2.0	61.45	0.8	1.09861
Stnd. skewness	-3.07989	0.0906087	1.22414	0.0232827	-5.60515
Stnd. kurtosis	2.6152	0.52516	0.490076	-0.823423	9.35136

The StatAdvisor
This table shows summary statistics for each of the selected data variables. It includes measures of central tendency, measures of variability, and measures of shape. Of particular interest here are the standardized skewness and standardized kurtosis, which can be used to determine whether the sample comes from a normal distribution. Values of these statistics outside the range of -2 to +2 indicate significant departures from normality, which would tend to invalidate many of the statistical procedures normally applied to this data. In this case, the following variables show standardized skewness values outside the expected range:
sqrt(Muc do khai thac)
log(Vi tri)
The following variables show standardized kurtosis values outside the expected range:
sqrt(Muc do khai thac)
log(Vi tri)
To make the variables more normal, you might try a transformation such as LOG(Y), SQRT(Y), or 1/Y.

Ví dụ sau khi thử đổi biến số thì biến sqrt(Nts) bảo đảm luật chuNn, trong khi đó thì 2 biến Muc do khai thac và Vi tri vẫn chưa thỏa mãn; nếu tiếp tục đổi biến số mà cũng không bảo đảm thì có 2 phương án: i) Đổi biến số theo kiểu khác; ii) Thu thập thêm dữ liệu để bảo đảm chuNn;

Kết quả phân tích này cũng chỉ ra được các biến số có quan hệ với nhau và ảnh hưởng đến y (Nts)

Correlations

	Cphu tbi	Da noi	Do cao	Do day dat	Do doc	Ket von	Luong mua	sqrt(Muc do khai thac)
Cphu tbi		0.1812	-0.3116	-0.2666	-0.3217	0.5636	0.8266	0.6420
		(19)	(19)	(19)	(19)	(19)	(19)	(19)
		0.4579	0.1940	0.2698	0.1792	0.0120	0.0000	0.0030
Da noi	0.1812		0.1619	-0.4934	0.1835	0.5192	0.3481	0.4579
	(19)		(19)	(19)	(19)	(19)	(19)	(19)
	0.4579		0.5078	0.0318	0.4522	0.0227	0.1442	0.0486
Do cao	-0.3116	0.1619		0.1432	0.3592	-0.0287	-0.3695	-0.0594
	(19)	(19)		(19)	(19)	(19)	(19)	(19)
	0.1940	0.5078		0.5587	0.1309	0.9070	0.1194	0.8092
Do day dat	-0.2666	-0.4934	0.1432		-0.0680	-0.2313	-0.2668	-0.2309
	(19)	(19)	(19)		(19)	(19)	(19)	(19)
	0.2698	0.0318	0.5587		0.7820	0.3407	0.2695	0.3415
Do doc	-0.3217	0.1835	0.3592	-0.0680		0.1117	-0.1692	-0.1966
	(19)	(19)	(19)	(19)		(19)	(19)	(19)
	0.1792	0.4522	0.1309	0.7820		0.6490	0.4885	0.4197
Ket von	0.5636	0.5192	-0.0287	-0.2313	0.1117		0.5135	0.4748
	(19)	(19)	(19)	(19)	(19)		(19)	(19)
	0.0120	0.0227	0.9070	0.3407	0.6490		0.0245	0.0400
Luong mua	0.8266	0.3481	-0.3695	-0.2668	-0.1692	0.5135		0.8012
	(19)	(19)	(19)	(19)	(19)	(19)		(19)
	0.0000	0.1442	0.1194	0.2695	0.4885	0.0245		0.0000
sqrt(Muc do khai thac)	0.6420	0.4579	-0.0594	-0.2309	-0.1966	0.4748	0.8012	
	(19)	(19)	(19)	(19)	(19)	(19)	(19)	
	0.0030	0.0486	0.8092	0.3415	0.4197	0.0400	0.0000	
Muc do Lua rung	-0.3769	0.2521	0.1194	0.2478	-0.2294	-0.2546	-0.0520	0.1167
	(19)	(19)	(19)	(19)	(19)	(19)	(19)	(19)
	0.1117	0.2979	0.6262	0.3064	0.3449	0.2928	0.8325	0.6343
sqrt(Nts)	-0.4810	-0.3686	-0.1715	-0.0247	-0.1215	-0.5421	-0.5983	-0.4547
	(19)	(19)	(19)	(19)	(19)	(19)	(19)	(19)
	0.0371	0.1204	0.4826	0.9199	0.6203	0.0165	0.0068	0.0505
pH	-0.7690	-0.1160	0.1786	-0.0916	0.2715	-0.5164	-0.6796	-0.6910
	(19)	(19)	(19)	(19)	(19)	(19)	(19)	(19)
	0.0001	0.6361	0.4643	0.7093	0.2608	0.0236	0.0014	0.0011
log(Vi tri)	0.2821	-0.3823	-0.1869	0.2069	-0.7285	-0.0642	-0.1223	-0.1035
	(19)	(19)	(19)	(19)	(19)	(19)	(19)	(19)
	0.2420	0.1062	0.4436	0.3953	0.0004	0.7940	0.6180	0.6733

	Muc do Lua rung	sqrt(Nts)	pH	log(Vi tri)
Cphu tbi	-0.3769	-0.4810	-0.7690	0.2821
	(19)	(19)	(19)	(19)
	0.1117	0.0371	0.0001	0.2420
Da noi	0.2521	-0.3686	-0.1160	-0.3823
	(19)	(19)	(19)	(19)
	0.2979	0.1204	0.6361	0.1062
Do cao	0.1194	-0.1715	0.1786	-0.1869
	(19)	(19)	(19)	(19)
	0.6262	0.4826	0.4643	0.4436
Do day dat	0.2478	-0.0247	-0.0916	0.2069
	(19)	(19)	(19)	(19)
	0.3064	0.9199	0.7093	0.3953
Do doc	-0.2294	-0.1215	0.2715	-0.7285
	(19)	(19)	(19)	(19)
	0.3449	0.6203	0.2608	0.0004
Ket von	-0.2546	-0.5421	-0.5164	-0.0642
	(19)	(19)	(19)	(19)
	0.2928	0.0165	0.0236	0.7940
Luong mua	-0.0520	-0.5983	-0.6796	-0.1223
	(19)	(19)	(19)	(19)
	0.8325	0.0068	0.0014	0.6180
sqrt(Muc do khai thac)	0.1167	-0.4547	-0.6910	-0.1035
	(19)	(19)	(19)	(19)
	0.6343	0.0505	0.0011	0.6733
Muc do Lua rung		-0.1064	0.2019	-0.1764
		(19)	(19)	(19)
		0.6648	0.4071	0.4699
sqrt(Nts)	-0.1064		0.3337	0.1746
	(19)		(19)	(19)
	0.6648		0.1627	0.4748
pH	0.2019	0.3337		-0.2960
	(19)	(19)		(19)
	0.4071	0.1627		0.2186
log(Vi tri)	-0.1764	0.1746	-0.2960	
	(19)	(19)	(19)	
	0.4699	0.4748	0.2186	

Correlation
(Sample Size)
P-Value

The StatAdvisor

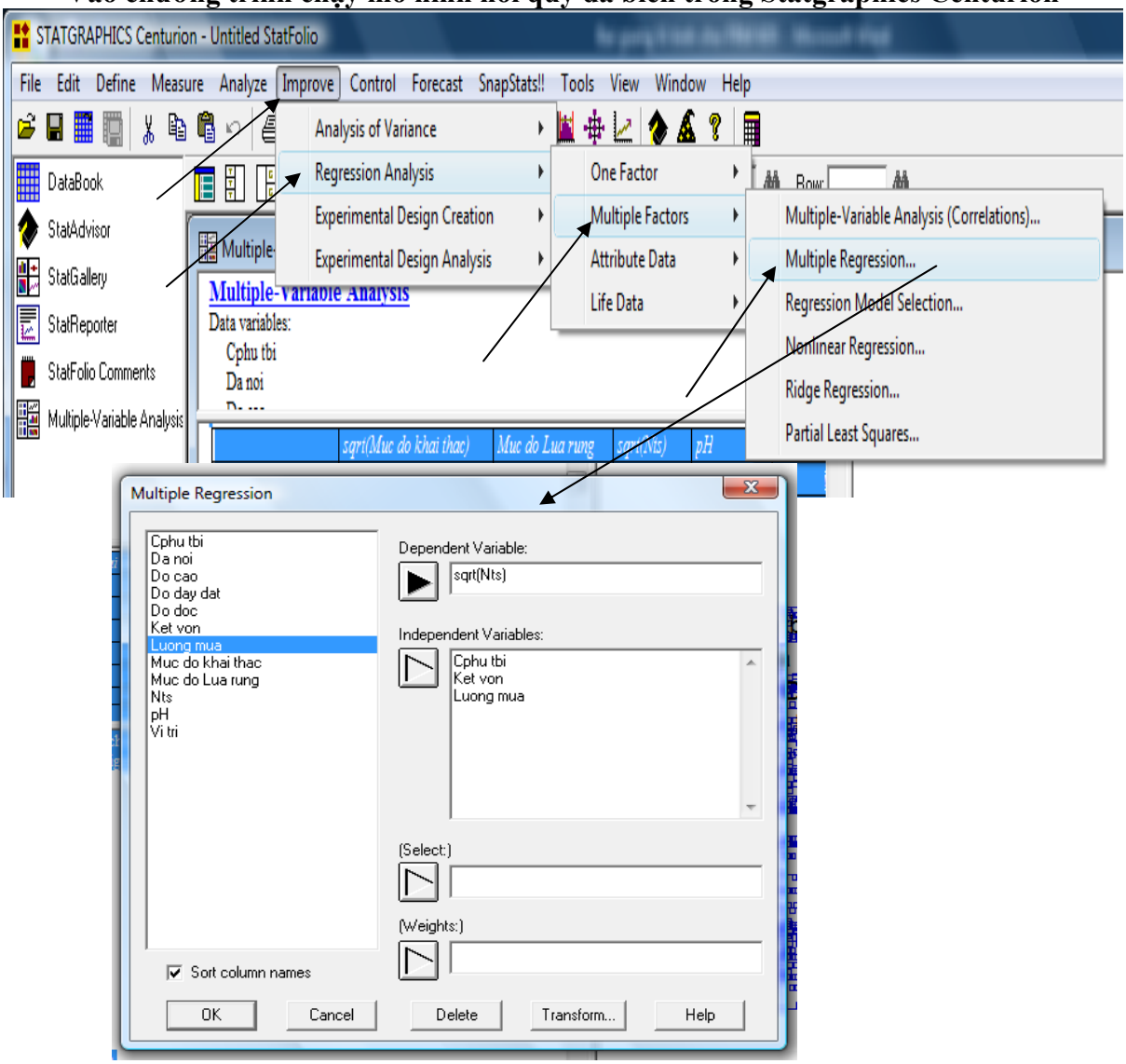
This table shows Pearson product moment correlations between each pair of variables. These correlation coefficients range between -1 and +1 and measure the strength of the linear relationship between the variables. Also shown in parentheses is the number of pairs of data values used to compute each coefficient. The third number in each location of the table is a P-value which tests the statistical significance of the estimated correlations. P-values below 0.05 indicate statistically significant non-zero correlations at the 95.0% confidence level. The following pairs of variables have P-values below 0.05:

Cphu tbi and Ket von
Cphu tbi and Luong mua
Cphu tbi and sqrt(Muc do khai thac)
Cphu tbi and sqrt(Nts)
Cphu tbi and pH
Da noi and Do day dat
Da noi and Ket von
Da noi and sqrt(Muc do khai thac)
Do doc and log(Vi tri)
Ket von and Luong mua
Ket von and sqrt(Muc do khai thac)
Ket von and sqrt(Nts)
Ket von and pH
Luong mua and sqrt(Muc do khai thac)

Luong mua and sqrt(Nts)
Luong mua and pH
sqrt(Muc do khai thac) and pH

Từ kết quả này cho thấy Nts bị chi phối bởi 3 nhân tố chính là: Cphu tbi, Kvon, Luong mua. Từ đây thiết lập mô hình quan hệ Nts với 3 biến này để lượng hóa sự ảnh hưởng: Improve/Regression Analysis/Multiple Factors/Mutiple Regression – Sau đó chọn các biến y, xi vào trong hộp thoại. Lưu ý đổi biến số để chuNn hóa như đã xác định ở bước trên.

Vào chương trình chạy mô hình hồi quy đa biến trong Statgraphics Centurion



Multiple Regression - sqrt(Nts)

Dependent variable: sqrt(Nts)
Independent variables:
Cphu tbi
Ket von
Luong mua

		Standard	*T*	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*
CONSTANT	127.22	53.9381	2.35863	0.0323
Cphu tbi	0.118008	0.21119	0.558777	**0.5846**
Ket von	-0.4484	0.29441	-1.52305	**0.1485**
Luong mua	-0.0726513	0.0430591	-1.68725	**0.1122**

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	2230.26	3	743.419	4.00	0.0281
Residual	2787.98	15	185.866		
Total (Corr.)	5018.24	18			

R-squared = 44.443 percent
R-squared (adjusted for d.f.) = 33.3316 percent
Standard Error of Est. = 13.6333
Mean absolute error = 10.1868
Durbin-Watson statistic = 1.17117 (P=0.0106)
Lag 1 residual autocorrelation = 0.363982

The StatAdvisor
The output shows the results of fitting a multiple linear regression model to describe the relationship between sqrt(Nts) and 3 independent variables. The equation of the fitted model is

sqrt(Nts) = 127.22 + 0.118008*Cphu tbi - 0.4484*Ket von - 0.0726513*Luong mua

Since the P-value in the ANOVA table is less than 0.05, there is a statistically significant relationship between the variables at the 95.0% confidence level.

The R-Squared statistic indicates that the model as fitted explains 44.443% of the variability in sqrt(Nts). The adjusted R-squared statistic, which is more suitable for comparing models with different numbers of independent variables, is 33.3316%. The standard error of the estimate shows the standard deviation of the residuals to be 13.6333. This value can be used to construct prediction limits for new observations by selecting the Reports option from the text menu. The mean absolute error (MAE) of 10.1868 is the average value of the residuals. The Durbin-Watson (DW) statistic tests the residuals to determine if there is any significant correlation based on the order in which they occur in your data file. Since the P-value is less than 0.05, there is an indication of possible serial correlation at the 95.0% confidence level. Plot the residuals versus row order to see if there is any pattern that can be seen.

In determining whether the model can be simplified, notice that the highest P-value on the independent variables is 0.5846, belonging to Cphu tbi. Since the P-value is greater or equal to 0.05, that term is not statistically significant at the 95.0% or higher confidence level. Consequently, you should consider removing Cphu tbi from the model.

Kết quả cho thấy cả 3 biến số đều có Pvalue>0.05; do đó chưa tham gia được vào mô hình; lúc này cần đổi biến số (log, exp, sqrt, 1/xi, ...) hoặc tổ hợp biến để bảo đảm sự tồn tại của biến số đó. Nếu một biến nào chưa tìm được cách đổi biến số thích hợp hoặc tổ hợp biến thì cần loại khỏi mô hình, tuy nhiên thực tế biến này có ảnh hưởng đến y, nhưng chưa được phát hiện dạng biến số thích hợp.

Kết quả thử nghiệm đổi biến số, tổ hợp biến, loại biến số

Multiple Regression - sqrt(Nts)

Dependent variable: sqrt(Nts)
Independent variables:
log(Luong mua*Ket von)

		Standard	*T*	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*
CONSTANT	83.901	18.0012	4.66085	0.0002
log(Luong mua*Ket von)	**-6.68159**	**1.99815**	**-3.34389**	**0.0038**

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	1991.09	1	1991.09	11.18	0.0038
Residual	3027.15	17	178.068		
Total (Corr.)	5018.24	18			

R-squared = 39.677 percent
R-squared (adjusted for d.f.) = 36.1286 percent
Standard Error of Est. = 13.3442
Mean absolute error = 10.4431
Durbin-Watson statistic = 1.34835 (P=0.0522)
Lag 1 residual autocorrelation = 0.293351

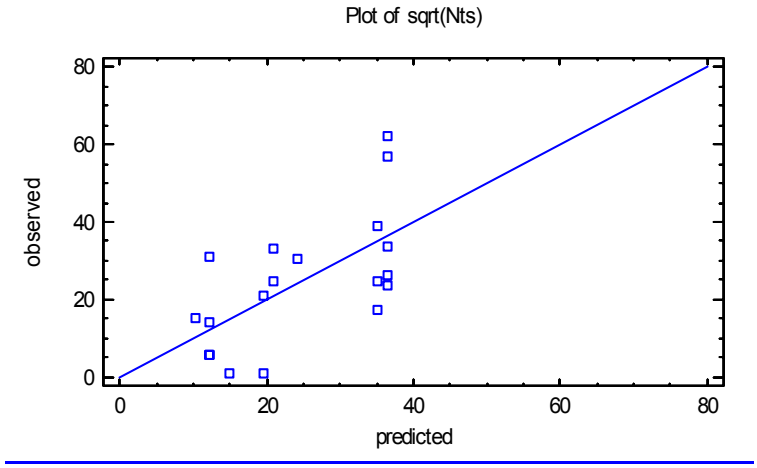
The StatAdvisor
The output shows the results of fitting a multiple linear regression model to describe the relationship between sqrt(Nts) and 1 independent variables. The equation of the fitted model is

sqrt(Nts) = 83.901 - 6.68159*log(Luong mua*Ket von)

Since the P-value in the ANOVA table is less than 0.05, there is a statistically significant relationship between the variables at the 95.0% confidence level.

The R-Squared statistic indicates that the model as fitted explains 39.677% of the variability in sqrt(Nts). The adjusted R-squared statistic, which is more suitable for comparing models with different numbers of independent variables, is 36.1286%. The standard error of the estimate shows the standard deviation of the residuals to be 13.3442. This value can be used to construct prediction limits for new observations by selecting the Reports option from the text menu. The mean absolute error (MAE) of 10.4431 is the average value of the residuals. The Durbin-Watson (DW) statistic tests the residuals to determine if there is any significant correlation based on the order in which they occur in your data file. Since the P-value is greater than 0.05, there is no indication of serial autocorrelation in the residuals at the 95.0% confidence level.

In determining whether the model can be simplified, notice that the highest P-value on the independent variables is 0.0038, belonging to log(Luong mua*Ket von). Since the P-value is less than 0.05, that term is statistically significant at the 95.0% confidence level. Consequently, you probably don't want to remove any variables from the model.



Kết quả thiết lập được mô hình:

sqrt(Nts) = 83.901 - 6.68159*log(Luong mua*Ket von)

Với R-squared = 39.677 percent; Pvalue <0.05

Các tham số đều tồn tại với Pvalue = 0.0038 < 0.05

Từ mô hình này cho thấy có hai nhân tố là lượng mưa và % kết von ảnh hưởng rõ rệt đến tái sinh ở khu vực nghiên cứu. Lượng mưa và kết von gia tăng làm giảm số cây tái sinh; đây là cơ sở quy hoạch cảnh quan và áp dụng biện pháp lâm sinh để xúc tiến tác sinh.

7. MÔ HÌNH HOÁ QUY LUẬT PHÂN BỐ

Trong nghiên cứu các lâm phần, người ta thường khái quát quy luật phân bố số cây theo cỡ kính, chiều cao để làm cơ sở cho việc điều tra rừng và xác định các giải pháp lâm sinh thích hợp để dẫn dắt rừng. Hoặc nghiên cứu phân bố số cá thể theo tuổi, thế hệ; phân bố số loài theo tầng thứ, phân bố vi sinh vật đất theo các lớp đất, ... để hiểu rõ quy luật sinh học, sinh thái học làm cơ sở quản lý tài nguyên thiên nhiên bền vững.

7.1. Mô hình hoá phân bố giảm theo hàm Meyer

Hàm Mayer có dạng: $y = \alpha.e^{-\beta.x}$. Kiểu dạng này thích hợp cho mô tả mô phỏng phân bố số cây theo cỡ kính (N/D) rừng chặt chọn có dạng giảm, hoặc mô phỏng sự giảm của số loài theo tầng, theo cỡ kính, .,,,,,

Trong Excel có chương trình lập sẵn tính quan hệ Mayer ngay trên đồ thị.

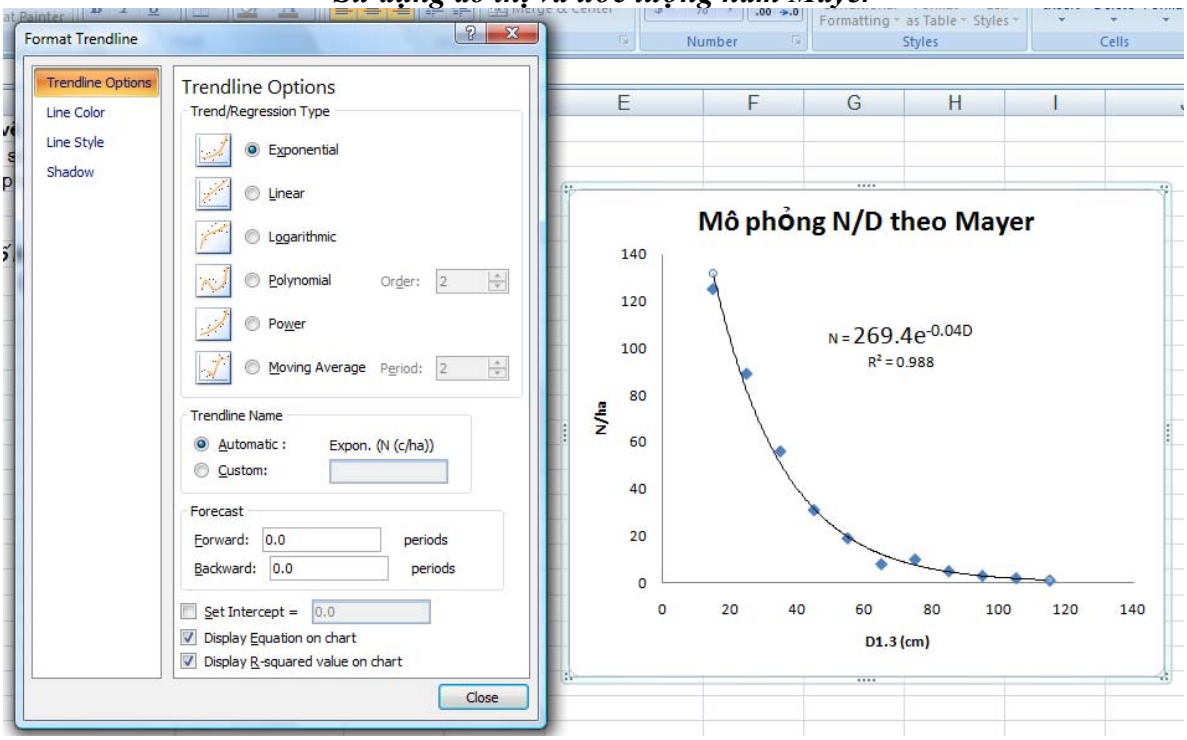
Ví dụ mô phỏng phân bố N/D theo dạng Mayer: $N = \alpha.e^{-\beta.D}$

Nhập số liệu: Cột A là giá trị giữa cỡ kính (D) ; Cột B là tần số thực nghiệm (N).

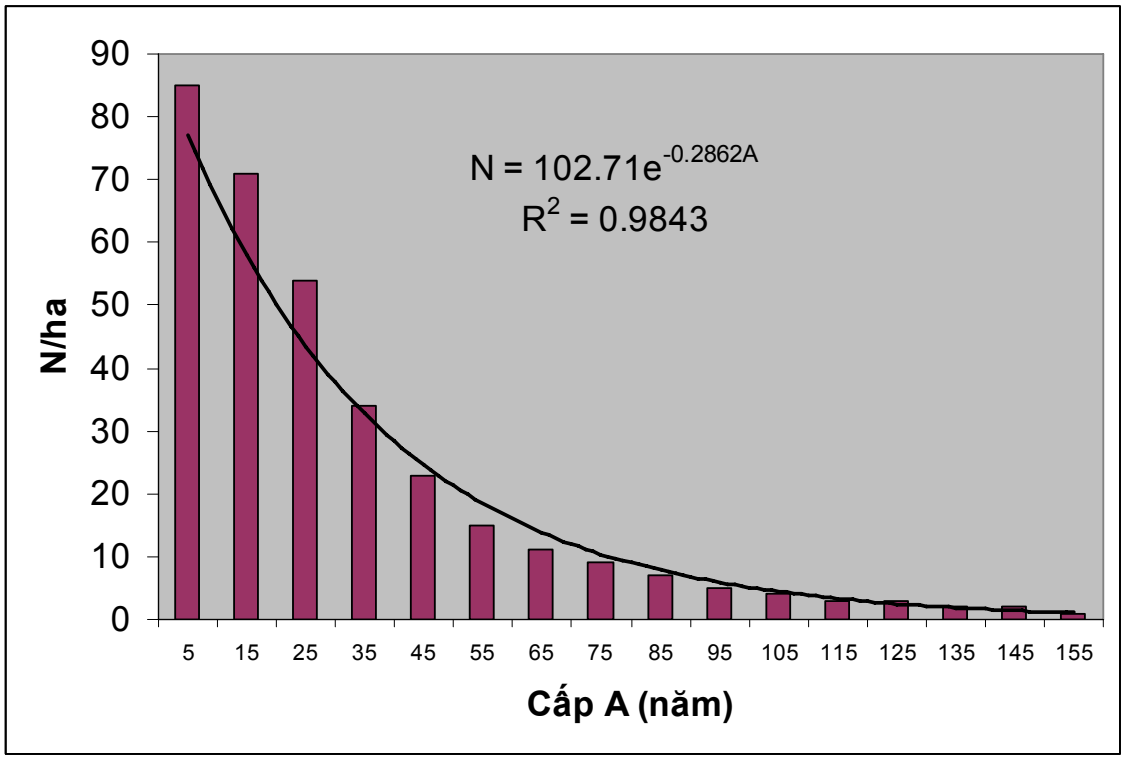
Bảng dữ liệu tần số phân bố N/D

	A	B
1	D1,3 (cm)	N (c/ha)
2	15	125
3	25	89
4	35	56
5	45	31
6	55	19
7	65	8
8	75	10
9	85	5
10	95	3
11	105	2
12	115	1

Sử dụng đồ thị và ước lượng hàm Mayer

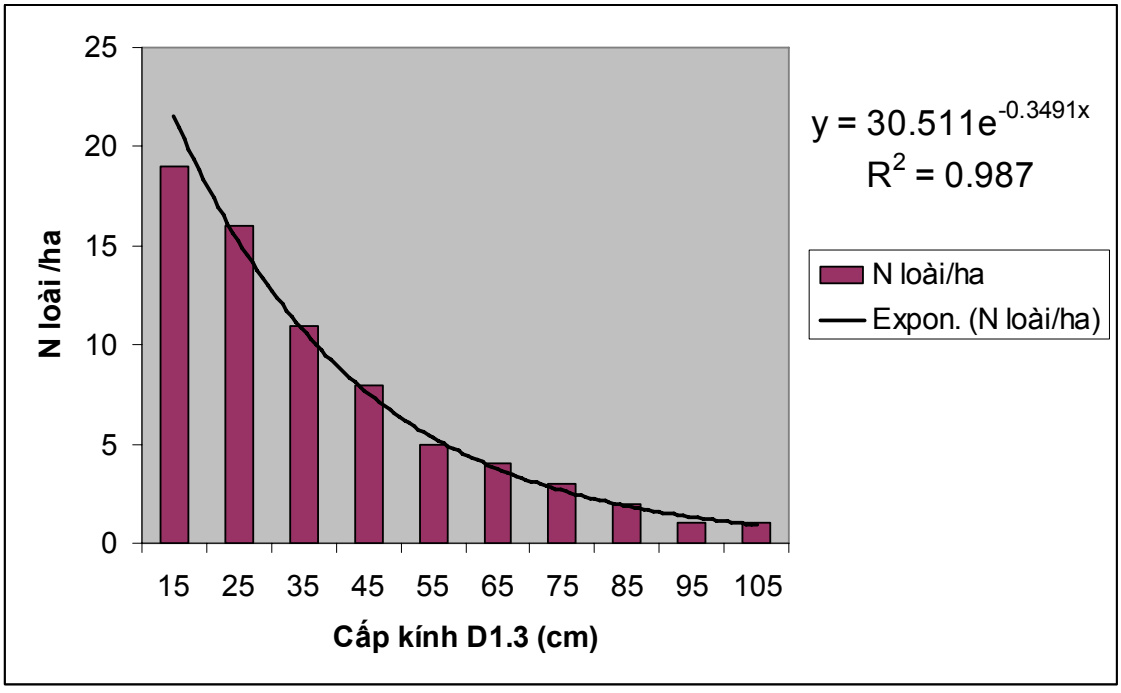


Phân bố Mayer còn có thể sử dụng để xem xét phân bố số lượng cá thể của một loài theo các giai đoạn tuổi. Kiểu dạng cấu trúc số cây theo tuổi (N/A) rừng nhiệt đới nhìn chung có dạng giảm, tuổi càng cao thì số cá thể càng ít, bảo đảm cho sự kế tục các thế hệ cây rừng và ổn định quần thể thực vật rừng theo thời gian. Với đặc trưng cấu trúc dạng giảm theo thế hệ, tuổi như vậy nên phương thức khai thác chính của rừng tự nhiên là chặt chọn theo cấp kính. Khai thác lớp cây thành thục và nuôi dưỡng rừng trong một luân kỳ để rừng phục hồi trạng thái ban đầu và tiếp tục khai thác lần 2. Việc xác định được cấu trúc N/A của lâm phần và N/A theo từng loài/nhóm loài chính sẽ rất thuận tiện cho việc xác định kỹ thuật lâm sinh như tuổi, đường kính khai thác, luân kỳ,.... Tuy nhiên trong thực tế việc xác định A là rất khó khăn, do đó thông thường được thay bằng đường kính, và kiểu cấu trúc phổ biến được nghiên cứu là số cây theo cỡ kính N/D để phục vụ cho điều tra, xác định chỉ tiêu kỹ thuật nuôi dưỡng, khai thác rừng. Mô hình hoá cấu trúc N/A thường được biểu diễn tốt bằng hàm Mayer với hệ số tương quan R^2 rất cao.



Ví dụ mô hình cấu trúc N/A rừng hỗn loài khác tuổi theo hàm Mayer

Rừng mưa nhiệt đới có khu hệ thực vật đa dạng với thành loài phong phú, phân bố ở nhiều thế hệ, cấp tuổi khác nhau. Trên 01 ha rừng có thể phát hiện trên 60 loài cây thân gỗ, ngoài ra rừng mưa rất phong phú các loài dây leo, song mây, rêu, dương xỉ, phong lan. Các loài cây nói chung là ưa sáng, cố gắng vươn lên cạnh tranh ánh sáng, tuy vậy cũng có loài chịu được ở tầng dưới và hình thành sự phân bố loài theo tầng, theo cấp tuổi, cấp kính khá rõ rệt. Trong thực tế việc xác định tuổi cây rừng là khó khăn, do đó thường nghiên cứu cấu trúc số loài theo cấp kính (Nloài/D)



Cấu trúc số loài theo cấp kính rừng nửa rụng lá ưu hợp bằng lăng – căm xe ở Dăk Lăk

Cấu trúc Nloài/D của kiểu rừng nửa rụng ưu hợp bằng lăng –căm xe ở Dăk Lak có kiểu dạng phân bố là dạng giảm liên tục, có nghĩa khi lên tầng cao, cấp kính lớn, số loài chiếm tỷ lệ thấp, đây là các loài ưu thế sinh thái. Với kiểu rừng này, số loài trên ha là 70 loài thân gỗ, và với cỡ kính thành thục từ 55cm trở lên thì số loài còn khoảng 5 loài. Kiểu dạng cấu trúc này cũng có thể mô phỏng tốt bằng dạng hàm Mayer.

7.2. Mô phỏng phân bố thực nghiệm theo phân bố khoảng cách-hình học:

i) <u>Dạng phân bố khoảng cách:</u>

$$P(x) = \begin{cases} \Upsilon & x=0 \\ (1-\alpha).(1-\Upsilon).\alpha^{x-1} & x>=1 \end{cases}$$

Với x là mã số các cỡ kính từ nhỏ đến lớn 0,1,2,3....

Khi: $\Upsilon < (1-\Upsilon)(1-\alpha)$ Phân bố có đỉnh tại x=1.

$\Upsilon = 1 - \alpha$ Phân bố giảm có thể thay thế bằng phân bố hình học.

$\Upsilon > (1-\Upsilon)(1-\alpha)$ Phân bố giảm.

Ước lượng 2 tham số bằng phương pháp cực đại hợp lý:

$$\Upsilon = N_0/N$$

$$\alpha = 1 - \frac{\sum_{i=1}^{r} Ni}{\sum_{i=1}^{r} Ni.xi}$$

Trình tự tính trong Excel: *Vd: Mô phỏng phân bố N/D có dạng 1 đỉnh:*

* Cột A: Mã số x
* Cột B: Giá trị giữa cỡ D.
* Cột C: Số cây theo cỡ kính. Tổng tại ô C13=sum(c2:c12)
* Cột D: Ni.xi. Tại ô D2:=A2*C2; copy cho các ô dưới. Tổng tại ô D13

* Tính 2 tham số:
  Υ = C2/Sum(c2:c12)
  α = 1- Sum(c3:c12)/sum(d2:d12)
* Cột E:Xác suất từng cỡ kính P(xi): Ô E2: Pxo=Υ; ô E3: Px1 = (1-Υ)(1-α)α^(a3-1); copy cho các ô dưới.
* Cột F: Tần số lý thuyết: Nlti: Ô F2: =$C$13*E2; copy cho các ô dưới
* Cột G: Tính χ^2 từng cỡ và tổng. Ô G2: = (f2-c2)^2/f2, copy cho các ô dưới, cộng tổng.
* Ô G14: Tra χ^2 bảng (α=0,05 ; K = 8-2-1=5): =Chiinv(0.05,5)

Kết quả χ^2 tính < χ^2 bảng . Kl: Phân bố Khoảng cách mô phỏng tốt phân bố thực nghiệm N/D.

Kết quả mô phỏng phân bố N/D theo phân bố khoảng cách

	A	B	C	D	E	F	G
1	**x**	**Cỡ D1,3 (cm)**	**N (c/ha)**	**Nixi**	**Px**	**Nlt (c/ha)**	**X2**
2	0	15	70	0	0,212121	70	0,00
3	1	25	125	125	0,345444	114	1,06
4	2	35	56	112	0,193985	64	1,00
5	3	45	31	93	0,108932	36	0,68
6	4	55	19	76	0,061171	20	0,07
7	5	65	8	40	0,034351	11	0,98
8	6	75	10	60	0,01929	6	2,08
9	7	85	5	35	0,010832	4	1,82
10	8	95	3	24	0,006083	2	
11	9	105	2	18	0,003416	1	
12	10	115	1	10	0,001918	1	
13	**Tổng**		**330**	**593**	**0,997543**	**329**	**7,70**
14			Gamma=	0,212121		X2 bảng=	11,07
15			Alpha=	0,561551		K=8-2-1=5	

ii) Phân bố hình học:

$P(x) = \alpha^x . (1-\alpha)$ x=0,1,2,3...r

Ước lượng α bằng phương pháp cực đại hợp lý:

$$\alpha = \frac{x}{x+1}$$

$$\mathbf{x} = \frac{1}{N}\sum_{i=1}^{r} Ni.xi$$

Phân bố hình học dùng mô tả các phân bố thực nghiệm dạng giảm

Trình tự tính trong Excel: *Vd: Mô phỏng phân bố N/D có dạng giảm:*

* Cột A: Mã số x
* Cột B: Giá trị giữa cỡ D.
* Cột C: Số cây theo cỡ kính. Tổng tại ô C13=sum(c2:c12)
* Cột D: Ni.xi. Tại ô D2:=A2*C2; copy cho các ô dưới. Tổng tại ô D13
* Tính tham số α:
  x = D13/c13
  α = x/(x+1)
* Cột E:Xác suất từng cỡ kính P(xi): Ô E2: Pxo = (1-α)α^a3; copy cho các ô dưới.
* Cột F: Tần số lý thuyết: Nlti: Ô F2: =$C$13*E2; copy cho các ô dưới
* Cột G: Tính χ^2 từng cỡ và tổng. Ô G2: = (f2-c2)^2/f2, copy cho các ô dưới, cộng tổng.
* Ô G14: Tra χ^2 bảng (α=0,05 ; K = 8-1-1=6): =Chiinv(0.05,6)

Kết quả χ^2 tính < χ^2 bảng . Kl: Phân bố hình học mô phỏng tốt phân bố thực nghiệm N/D.

Kết quả mô phỏng phân bố N/D theo phân bố hình học

	A	B	C	D	E	F	G
1	**x**	**Cỡ D1,3 (cm)**	**N (c/ha)**	**Nixi**	**Px**	**Nlt (c/ha)**	**X2**
2	0	15	125	0	0,38521	134	0,66
3	1	25	89	89	0,236823	83	0,49
4	2	35	56	112	0,145597	51	0,53
5	3	45	31	93	0,089511	31	0,00
6	4	55	19	76	0,055031	19	0,00
7	5	65	8	40	0,033832	12	1,23
8	6	75	10	60	0,0208	7	1,03
9	7	85	5	35	0,012788	4	0,12
10	8	95	3	24	0,007862	3	
11	9	105	2	18	0,004833	2	
12	10	115	1	10	0,002971	1	
13	**Tổng**		**349**	**557**	**0,995258**	**347**	**4,06**
			xbq=	1,595989		X2 bảng=	12,59
			Alpha=	0,61479		K=8-1-1=6	

7.3. Mô phỏng phân bố thực nghiệm theo phân bố Weibull:

Phân bố Weibull là phân bố xác suất của biến ngẫu nhiên liên tục với miền giá trị $x \in (0,+\infty)$.

Hàm mật độ:

$f(x) = \alpha.\lambda(x - x_{min})^{\alpha-1}.\exp(-\lambda(x - x_{min})^{\alpha}$

Hàm phân bố:

$F(x) = 1 - \exp(-\lambda(x - x_{min})^{\alpha}$

Với x_{min}: trị số quan sát nhỏ nhất.

x: các giá trị quan sát, nếu xếp theo tổ thì x là giá trị giữa mỗi tổ.

Khi:

$\alpha <= 1$: Phân bố giảm.

$1<\alpha<3$:Phân bố lệch trái

$\alpha=3$: Phân bố đối xứng.

$\alpha>3$: Phân bố lệch phải.

**** Ước lượng 2 tham số α và λ:***

Tham số α thường được thăm dò trong một khoảng thích hợp dựa trên các đặc trưng mẫu, cho chạy α để tính λ. Sau đó kiểm tra sự phù hợp của phân bố lý thuyết bằng tiêu chuẩn χ^2, chọn cặp tham số có χ^2 bé nhất và nhỏ thua χ^2 bảng.

Tham số λ được ước lượng bằng phương pháp cực đại hợp lý:

r

$$\lambda = N/\sum_{i=1} N_i.(x_i - x_{min})^{\alpha}$$

N: Tổng dung lượng quan sát.
N_i: Tần số tổ i.

*** Tính xác xuất cho từng tổ:**

+ Tổ 1: $P(x_1)=F(x_1) = 1 - \exp(-\lambda(x_1 + A - x_{min})^{\alpha})$
+ Tổ 2: $P(x_2)=F(x_2) - F(x_1) = \exp(-\lambda(x_1 + A - x_{min})^{\alpha}) - \exp(-\lambda(x_2 + A - x_{min})^{\alpha})$
+ Tổ 3: $P(x_3)=F(x_3) - F(x_2) = \exp(-\lambda(x_2 + A - x_{min})^{\alpha}) - \exp(-\lambda(x_3 + A - x_{min})^{\alpha})$

.....

+ Tổ r: $P(x_r)=F(x_r) - F(x_{r-1}) = \exp(-\lambda(x_{r-1} + A - x_{min})^{\alpha}) - \exp(-\lambda(x_r + A - x_{min})^{\alpha})$

Với A: giá trị 1/2 cự ly tổ.

*** Tần số lý thuyết Nlt cho từng tổ:**

$Nlt_i = N.P(x_i)$.

*** Kiểm tra sự phù hợp bằng tiêu chuẩn χ^2.**

Kết quả mô phỏng phân bố N/D theo hàm Weibull

	A	B	C	D	E	F	G	H
1	**Cỡ D1,3 (cm)**	**N (c/ha)**	**Alpha**	**N(x-xmin)^alpha**	**Lamda**	**P(x)**	**Nlt (c/ha)**	**X2**
2	15	125	1	625,0	0,047710	0,379420	132	0,42
3	25	89		1335,0		0,235460	82	0,57
4	35	56		1400,0		0,146121	51	0,49
5	45	31		1085,0		0,090680	32	0,01
6	55	19		855,0		0,056274	20	0,02
7	65	8		440,0		0,034922	12	1,44
8	75	10		650,0		0,021672	8	0,78
9	85	5		375,0		0,013449	5	0,02
10	95	3		255,0		0,008346	3	0,00
11	105	2		190,0		0,005179	2	
12	115	1		105,0		0,003214	1	
13	**Tổng**	**349**		**7315,0**		**1,0**	**347**	**3,76**
14							X2 bảng=	14,07
15							K=9-1-1=7	

* Cột A: Giá trị giữa cỡ kính 15, 25,....115 với cự ly cỡ 10 cm.
* Cột B: Số cây từng cỡ Ni. Ô B13: tổng N= Sum(b2:b12)
* Ô C2: Đưa tham số α thăm dò.
* Cột D: Giá trị: $Ni(x_i - 10)^{\alpha}$. Với x_{min}=10. Tính tại ô d2: =B2*(A2-10)^$C$2, sau đó copy cho các ô dưới. Ô D13 tính tổng =Sum(d2:d12).
* Ô E2: Tính tham số λ: = B13/Sum(d2:d12).
* Cột F: Tính xác suất P(x) từng tổ: Tính theo công thức địa chỉ ô.
* Cột G: Nlt từng tổ: Ô G2: =$B$13*F2, sau đó copy xuống và tính tổng.
* Cột H: Tính χ^2 từng tổ và tổng χ^2=3.76

* Ô H14: Tra $\chi^2(0.05,7)$ =Chiinv(0.05,7)=14.07
* KL: Phân bố Weibull mô phỏng tốt phân bố thực nghiệm.

Chú ý: Để chọn được α tối ưu, lần lượt thay giá trị ở ô C2, bảng tính sẽ tự động tính lại, sau đó chọn một α với χ^2 bé nhất.