MINITAB

PHÂN TÍCH THỐNG KÊ BẰNG PHẦN MỀM MINITAB

MỤC LỤC

CHƯƠNG 1

GIỚI THIỆU VỀ MINITAB

Dữ liệu mà không có khái niệm thì mơ hồ. Khái niệm không có dữ liệu thì vô nghĩa.

Immanuel Kant

Chương 1 giới thiệu tổng quan về Minitab và các thao tác sử dụng Minitab for Windows. Các bạn sẽ làm quen với màn hình và các thành phần của Minitab cũng như các thủ thuật sử dụng, cách nhập liệu, lưu trữ, in ấn và trình bày kết quả.

NỘI DUNG

1.1 SỬ DỤNG MINITAB

Minitab[1] là một phần mềm máy tính giúp hiểu biết thêm về thống kê và tiết kiệm thời gian tính toán. Phần mềm này ban đầu được thiết kế để phục vụ việc giảng dạy môn thống kê, sau đó đã được phát triển thành công cụ phân tích và trình bày dữ liệu rất hữu hiệu.

Để có thể sử dụng Minitab trong môi trường máy tính cá nhân (PC) người sử dụng cần làm quen với hệ điều hành (Operating System), cấu trúc thư mục, ổ dĩa cứng và ổ dĩa mềm. Giao diện Minitab cho phép gõ các câu lệnh trong cửa sổ thao tác (Session Window) và thực thi chương trình bằng cách chọn lệnh từ thanh Menu và điền đầy đủ yêu cầu vào các hộp hội thoại. Các câu lệnh và kết quả sẽ được thể hiện ở cửa sổ thao tác. Ngoài ra bạn có thể chép (Copy), sửa chữa (Edit), và thực thi các lệnh trước.

Sử dụng "Minitab for Windows" người sử dụng cần có một số kiến thức về Windows như di chuyển, thay đổi kích cỡ, dùng thanh menu và các hộp hội thoại. Bạn cần nên làm quen với các thao tác chuột như phóng to, thu nhỏ, nhấp chuột một lần, hai lần, kéo rê chuột. Bạn cũng có thể dùng các tổ hợp phím để thực hiện các thao tác tương tự như nhấp chuột, tuy nhiên các thao tác này đòi hỏi người sử dụng phải nhớ nhiều.

Bạn cần nhớ một số biểu tượng thông thường khi sử dụng các ứng dụng trên môi trường Windows như Maximize  Minimize  Restore  và Close  , và các chức năng thông dụng trên menu.

Để vào Minitab for Windows, bạn nhấp chuột 2 lần vào dấu Minitab trên desktop hoặc **Start > Program > MINITAB 13 for Windows > MINITAB**. Màn hình chính của Minitab thể hiện trên Hình 1.1 gồm các phần: *Menu Bar, Standard Toolbar, Project Manager Toolbar, Title Session Window, Data Window, Status Bar*.

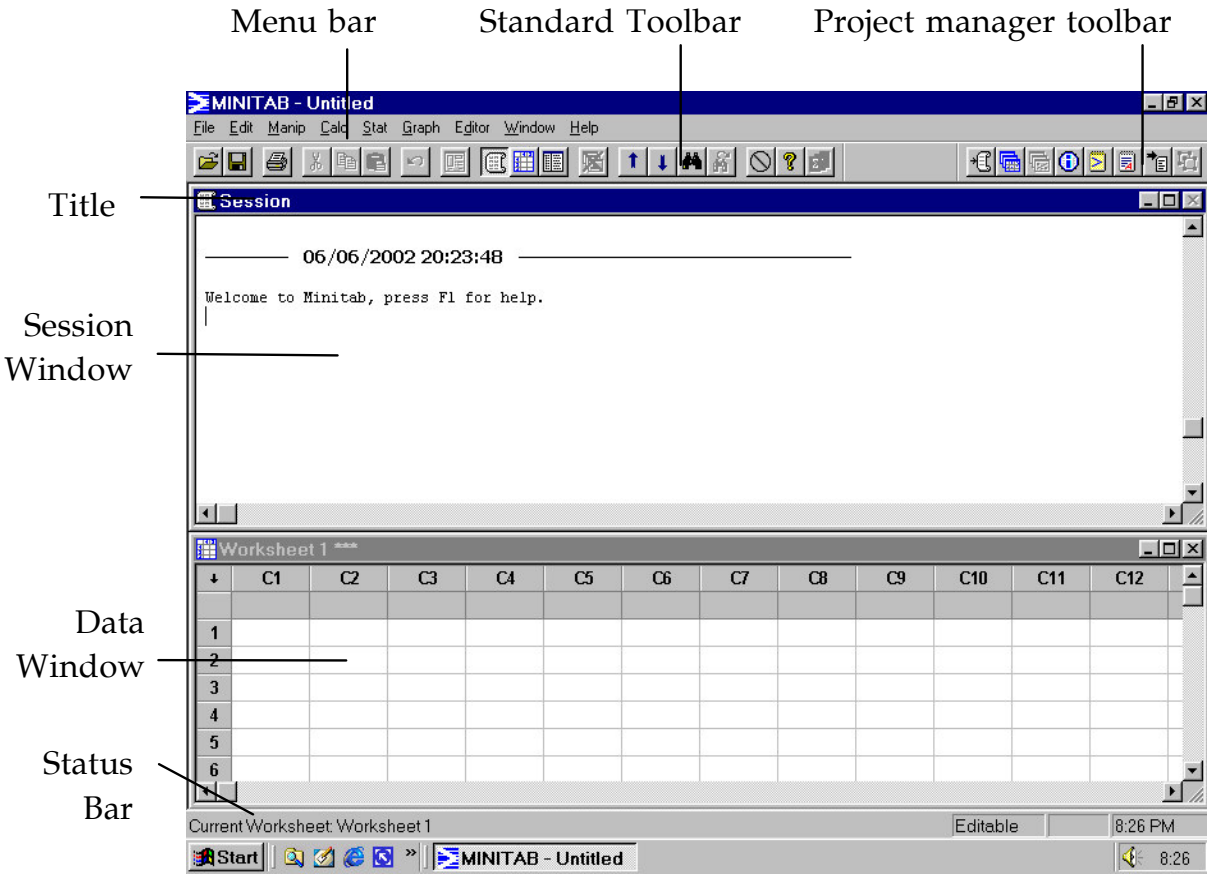
[1] Minitab là thương hiệu của Minitab Inc.

Session Window giúp tạo chương trình bằng cách gõ các câu lệnh. Thông thường các lệnh và kết quả sẽ được thể hiện trên cửa sổ này khi thực thi.

Data Window cho phép bạn nhập dữ liệu vào bảng tính bằng cách gõ hoặc xuất nhập các file dữ liệu từ bên ngoài.

Đồ thị khi thực thi các lệnh của Minitab có yêu cầu vẽ đồ thị, Minitab tự động mở một cửa sổ chứa đồ thị yêu cầu. Lưu ý, Minitab chỉ cho hiển thị tối đa 15 đồ thị trên cùng một project.

Hình 1.1 Màn hình chính của Minitab for Windows

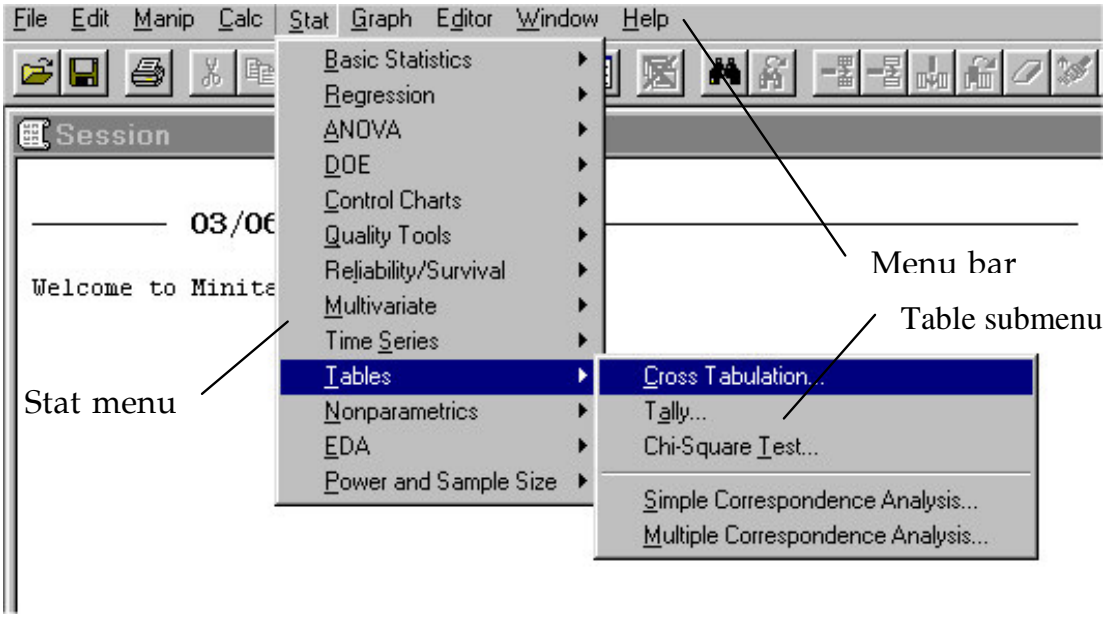


Menu bar giúp mở các menu và chọn các câu lệnh, nhấp chuột vào các mục trên thanh menu sau đó nhấp chuột vào các mục phụ thuộc để thực hiện lệnh hoặc mở hộp hội thoại. Khi các chức năng này không sử dụng được các hạng mục sẽ mờ đi. Để thực hiện lệnh như trên, tập tài liệu này sẽ ghi vắn tắt **Stat > Tables > Cross Tabulation** có nghĩa là

chọn **Stat** trên thanh menu, chỉ vào **Tables** và chọn **Cross Tabulation** trên Submenu của **Tables**. Hình 1.2 thể hiện Menu và Submenu của Minitab.

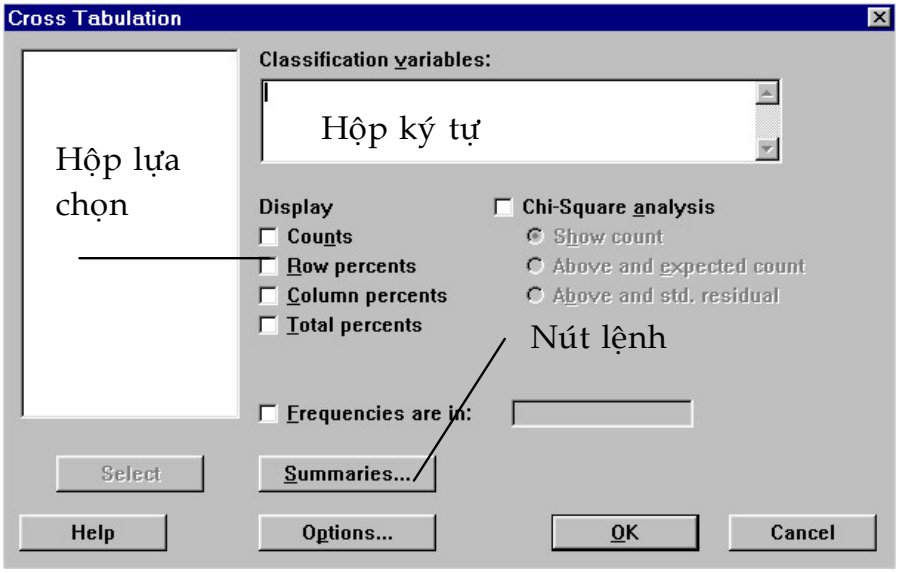
Menu **Edit**, **Manip, Editor, Stat,** và **Graph** hỗ rợ những lựa chọn để sử lý dữ liệu trong Minitab. Menu **Edit** của Minitab cũng giống như hầu hết các ứng dụng khác trên Windows gồm các chức năng như chép (copy), dán (Paste), xóa (Delete) và chọn lựa (Select). Menu **Manip** chứa các lệnh xứ lý trong Minitab. Menu **Editor** chứa các chức năng chuyển dịch và định dạng bảng tính đặc thù của Minitab. Menu **Stat** cung cấp một số công cụ để phân tích thống kê. Menu **Graph** hỗ trợ các công cụ để vẽ các loại đồ thị khác nhau.

Hình 1.2 Menu và submenu



Dialog box: Trong một số trường hợp vận hành các lệnh của Minitab, hộp hội thoại sẽ hiện ra để bạn lựa chọn các chức năng cần thực hiện. Một hộp hội thoại thông thường có những chức năng để người dùng lựa chọn cho những lệnh kế tiếp, giải thích biến, hoặc gán các lựa chọn khác trong từng nút lệnh. Ví dụ, sau khi thực hiện lệnh **Stat > Tables > Cross Tabulation** thì màn hình sẽ xuất hiện dialog box như Hình 1.3.

Hình 1.3 Dialog box của Minitab



Phần **Display** ở hình trên cho phép người dùng chọn những thông số nào sẽ được thể hiện trong kết quả của thao tác. Các nút lệnh cho phép người dùng bổ sung một số các lựa chọn đặc thù của loại dữ liệu vào kết quả lệnh.

Nếu như bạn chọn: **Window>project manager,** màn hình sẽ thể hiện như Hình 1.4:

Project Manager là một phần mới trong Minitab phiên bản 13. Mục này giúp bạn tổ chức và quản lý công việc tốt hơn. Những danh mục dưới đây có các lệnh, dữ liệu, kết quả, đồ thị, và các tài liệu liên quan khác.

Session: Dùng danh mục này để sao chép, xoá hoặc in kết quả và các đồ thị từ Session Window. Bạn có thể bổ sung nội dung của Session Window vào *ReportPad*.

History: Danh mục này chứa tất cả các lệnh được dùng trong suốt phiên làm việc. Dùng danh mục này để lặp lại các thứ tự lệnh hoặc tạo các Macro.

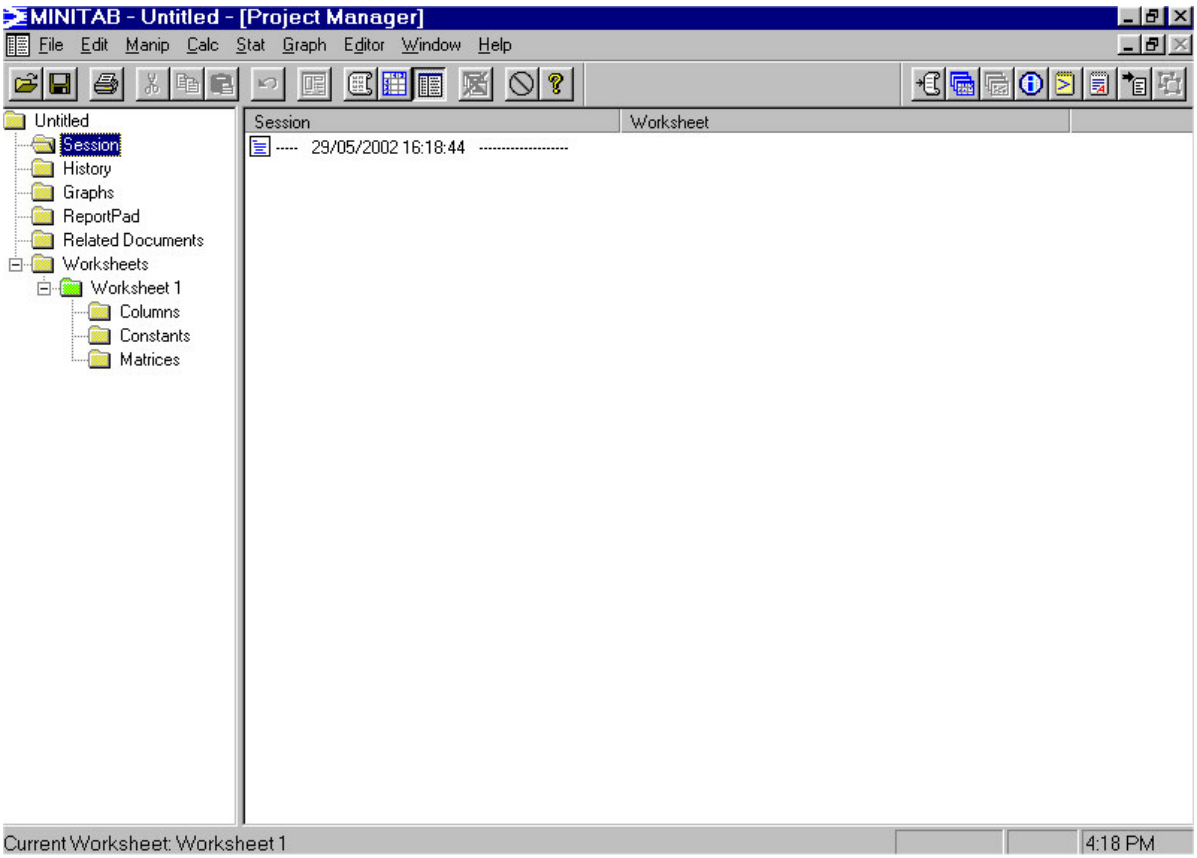
Chart: Dùng danh mục này để quản lý tất cả các đồ thị. Bạn có thể xóa, sắp đặt hoặc bổ sung đồ thị vào *ReportPad*.

ReportPad: Đây là một bộ phận xử lý văn bản mới của Minitab để tạo các báo cáo. Nhấn-chuột-phải vào kết quả hay đồ thị để thêm vào ReportPad. Bạn có thể Sửa chữa và in báo cáo từ ReportPad, hoặc sao chép qua một phần mềm xử lý văn bản khác như Microsoft Word.

Related Document: Danh mục này cho danh sách các file dữ liệu, tài liệu, hoặc các nối kết Internet (URL) không phải của Minitab để tham chiếu hoặc các mục đích sử dụng khác.

Worksheet: Dùng danh mục này cho các thông tin dạng Worksheet bao gồm cột, các biến cố định, số cột, các giá trị thiếu, và ma trận.

Hình 1.4 Màn hình Project manager



Project Manager Toolbar hỗ trợ các shortcut đến những chỉ mục đã nói ở trên.

Standard Toolbar thể hiện những nút các chức năng thường sử dụng, các nút này có thể thay đổi tuỳ vào chức năng sử dụng. Rê chuột đến các nút này để thấy tên của từng chức năng như hình dưới đây.

Status bar ở đáy màn hình thể hiện các câu giải thích cho các mục trên menu hoặc các các chức năng bạn đang sử dụng.

GHI CHÚ

Hướng dẫn này dùng cho Phiên Bản 13 của Minitab vì vậy nếu các bạn dùng các phiên bản trước có thể không có các chức năng được đề cập ở đây.

Các lệnh của Minitab có sẵn trong Menu hoặc thông qua ngôn ngữ trong phần Session Window. Có thể dùng cả hai cách để ghi câu lệnh. Ở đây chúng ta chỉ tập trung vào sử dụng các lệnh trên thanh Menu.

Để dùng các lệnh trên thanh lệnh (Menu Commands), nhấp chuột vào Menu, thực hiện lệnh, mở các mục phụ, mở các hộp đối thoại. Nếu như mục này mờ, có nghĩa là chức năng này không sử dụng được. Ví dụ, để mở cửa sổ dữ liệu, nhấp chuột vào mục **Window**, nhấp vào **Worksheet** (**Window > Worksheet**).

1.2 MINITAB WINDOWS

Minitab sử dụng năm loại cửa sổ (windows) được mô tả trong bảng 1.1 dưới đây. Với Minitab, bạn sẽ làm việc với từng **Project**, trong Project sẽ gồm một **Session Window**, các **Worksheet** và các cửa sổ **Graphics**, cửa sổ **Info**, và cửa sổ **History**.

Bảng 1.1 Các cửa sổ Minitab

Cửa sổ	*Mô tả*
Data	Trình bày dữ liệu bảng tính dưới dạng hàng và cột, cho phép bạn nhập liệu và làm việc với những dữ liệu này. Mội bảng tính (worksheet) có một cửa sổ dữ liệu (data window) riêng.
Session	Trình bày kết quả phân tích dạng chữ. Chỉ có một Session Window. Bạn có thể nhập lệnh vào trong Session Window.
Graph	Trình bày các đồ thị từ lệnh bạn yêu cầu, mỗi đồ thị sẽ hiện diện trên một cửa sổ. Trong một project, chỉ có thể hiện được một lúc 15 đồ thị.
Info	Trình bày tóm tắt về mỗi bảng tính, chỉ có một cửa sổ info.
History	Trình bày các câu lệnh đã thực hiện, không trình bày kết qủa. Chỉ có một cửa sổ History.

Minitab vận hành như một hộp đen, nhận dữ liệu đầu vào (input), phân tích xử lý theo yêu cầu của người sử dụng sau đó hiển thị kết quả (output) thông qua các cửa sổ đẽ đề cập ở trên. Dữ liệu đầu vào có thể nhận được thông qua nhập liệu trực tiếp vào các cửa sổ bảng tính (Worksheets), hoặc lấy file dữ liệu của Minitab hoặc từ các ứng dụng bảng tính khác như Excel hay Lotus. Ngoài ra Minitab cũng nhận nhập liệu từ Session Window thông qua các ngôn ngữ lệnh của Minitab. Kết quả phân tích của Minitab sẽ là các kết quả bằng số, chữ và hình ảnh đồ thị thông qua các cửa sổ của Minitab.

1.3 BẢNG TÍNH MINITAB (MINITAB WORKSHEET)

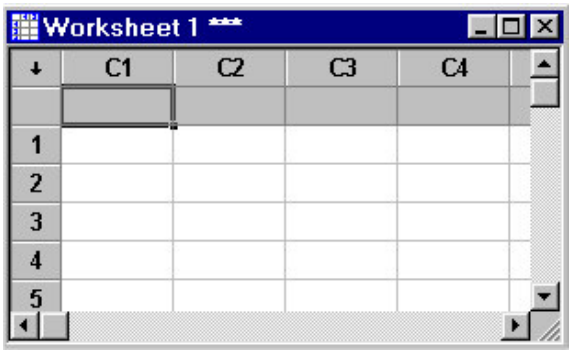
1.3.1 Dữ liệu Minitab

Minitab làm việc trên môi trường bảng tính gồm các hàng (Row) và cột (Column). Thông thường các cột dành cho các biến. Các giá trị quan sát của mỗi biến hoặc điểm dữ liệu sẽ được điền vào các mỗi ô (cell) trên

hàng. Các cột được thể hiện bằng C1, C2, C3… và các hàng trong từng cột sẽ được đánh số 1, 2, 3, hàng đầu tiên dành riêng cho tên của cột.

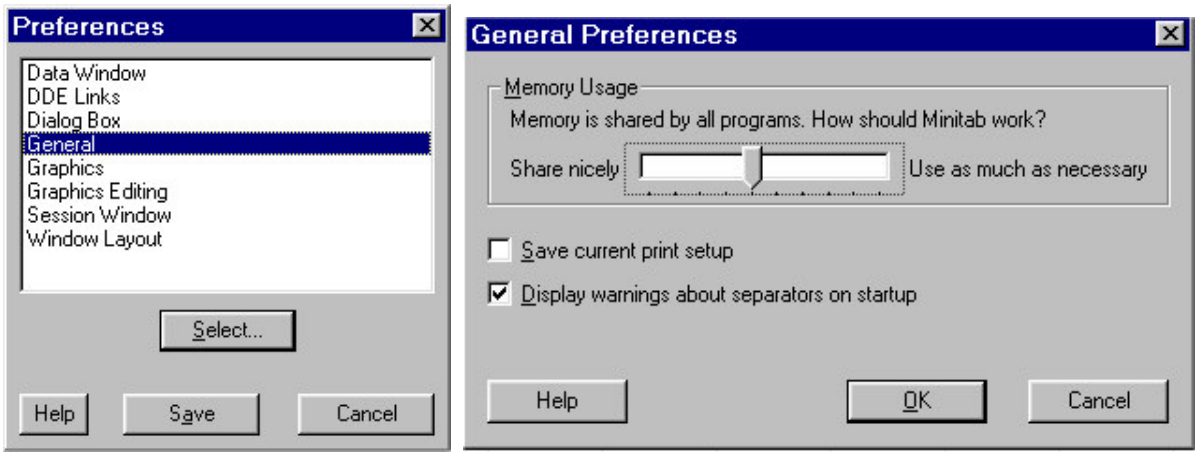
Minitab lưu dữ liệu trong bảng tính ở ba dạng: Cột (Column), các hằng số đơn (Constants) và ma trận (Matrices). Cột được ký hiệu là C1, C2, C3… Hằng số được ký hiệu là K1, K2, K3…, bất cứ một phép tính nào cho ra kết quả là một số đơn sẽ được lưu dưới dạng hằng số. Ma trận ký hiệu là M1, M2, M3,… gồm các giá trị trong các mảng gồm nhiều ô chứa các số.

Hình 1.5 cửa sổ bảng tính (worksheet window)



Trong phiên bản Minitab 13, mỗi bảng tính cho phép dùng đến 4000 cột. Số lượng hàng và ô phụ thuộc vào phân bổ bộ nhớ. Bạn có thể tự kiểm soát được công việc này bằng cách chọn **Edit > Preference > General.**

Hình 1.6 Minitab preferences



Di chuyển thanh trượt cho phần Memory usage qua hai cực điểm:
Share nicely: Minitab sẽ dùng bộ nhớ cùng với các chương trình khác đang chạy.

Use as much as necessary: Minitab sẽ dùng bộ nhớ khi cần thiết, không chia sẻ bộ nhớ với các chương trình khác đang chạy.

Thông thường Minitab sử dụng dữ liệu thông qua Data window trong từng ô dữ liệu. Mỗi cột dữ liệu có **Column header,** thông thường đây là tên biến của dữ liệu. Mỗi hàng dữ liệu có **Row header**, chỉ số thứ tự của giá trị biến quan sát.

1.3.2 Chọn biến

Chọn lệnh từ menu thường phải mở các **hộp thoại,** các hộp thoại này cho phép bạn chọn các biến và các chức năng cần thực hiện. Để chọn các biến, bạn cần phải thấy con trỏ trong hộp bạn muốn nhập biến vào. Bạn có thể làm nổi bật các biến trong danh sách các biến trong hộp và nhấn **Select**. Các biến sẽ xuất hiện trong hộp cùng với con trỏ.

Bạn có thể đặt tên mới cho cột trong hộp thoại. Ví dụ, khi dùng chức năng tính toán, bạn có thể đặt tên cho cột và hằng số để lưu kết quả. Minitab sẽ tự động dùng các cột và biến còn lại.

1.3.3 Dữ liệu thiếu (missing data)

Nhiều tập dữ liệu thiếu một số giá trị và quan sát, trong trường hợp này, khi nhập liệu bạn gõ dấu sao (*) vào nơi có giá trị thiếu. Các lệnh của Minitab sẽ tự động tính đến giá trị (*) này khi phân tích. Nếu như bạn thực hiện một phép tính mà Minitab không thể thực hiện được như là lấy căn của một số âm, Minitab sẽ tự động cho ra giá trị (*).

1.3.4 Loại dữ liệu

Bạn có thể dùng ba loại dữ liệu trong bảng tính Minitab: *số, chữ và ngày tháng*. **Dữ liệu số** gồm các ký tự số và dấu * (giá trị thiếu). Nếu bạn nhập liệu dưới dạng giá trị lũy thừa, dữ liệu số cũng gồm ký tự E như 3.2E12 sẽ là $3.2x10^{12}$. Nếu như ô dữ liệu có chứa các ký tự khác với số hoặc giá trị thiếu, Minitab sẽ tự động hiểu giá trị này là dạng

dữ liệu chữ. **Dữ liệu chữ** bao gồm các ký tự của bàn phím và thường được dùng để để chỉ các biến phân loại. Ví dụ như cột chứa biến giới tính của người tham gia trong nghiên cứu, bạn có thể dùng ký tự "M" và "F" để chỉ hai giới. Biến chữ có thể dài đến 80 ký tự. **Dữ liệu ngày tháng** lưu các giá trị ngày tháng, giờ hoặc cả hai. Minitab lưu dữ liệu này ở dạng số nhưng bạn có thể hiệu chỉnh theo dạng bạn muốn. Minitab mặc định hiểu loại dữ liệu đầu tiên nhập vào sẽ là loại dữ liệu của biến phân tích. Để nhận dạng, dữ liệu chữ thể hiện bên phải của ô dữ liệu, dữ liệu số sẽ thể hiện bên trái của ô dữ liệu. Khi cột biến nhận dữ liệu số, tiêu đề biến mặc định nhận tên C1, C2, C3,.. khi cột biến nhận dữ liệu chữ , tiêu đề biến tự động đổi tên thành "C1-T"

Một số quy luật khi sử dụng dữ liệu ký tự:

- Dữ liệu chữ có thể chứa đến 80 ký tự vào gồm ký tự, số, dấu câu, ký hiệu, hoặc khoảng trống.
- Các con số trong phần chữ được xem như là dữ liệu chữ, và không thể sử dụng để tính toán.
- Dữ liệu chữ và số không thể được dùng trong cùng một cột
- Có thể lưu dữ liệu chữ dưới dạng hằng số
- Dữ liệu chữ có thể được chuyển thành dữ liệu số.

Minitab nhận dạng dữ liệu của biến thông qua giá trị đầu tiên nhập vào, nếu như bạn nhập nhầm biến chữ thành số hay ngược lại, Minitab sẽ không cho nhập tiếp tục những giá trị không cùng loại dữ liệu. Để xử lý, bạn có thể đổi loại dữ liệu của cột bằng cách dùng lệnh **Manip>Change Data Type>**. Lệnh này cho phép bạn có thể chuyển dữ liệu chữ thành số, số thành chữ, chữ thành chữ, số thành số, ngày thành số hoặc thành chữ hoặc ngược lại. Ví dụ, nếu bạn muốn đổi dữ liệu số từ cột C1 thành dữ liệu số cũng trên cột C1 bạn dùng lệnh **Manip>Change Data Type>Numeric to text>** cửa sổ mới yêu cầu bạn nhập dữ liệu nguồn và gốc. Khi sử dụng lệnh này, bạn cần lưu ý đặc điểm của dữ liệu nguồn và dữ liệu đích. Nếu bạn chọn sai nguồn dữ liệu, Minitab sẽ tự động bào lỗi. (xem hình 1.7)

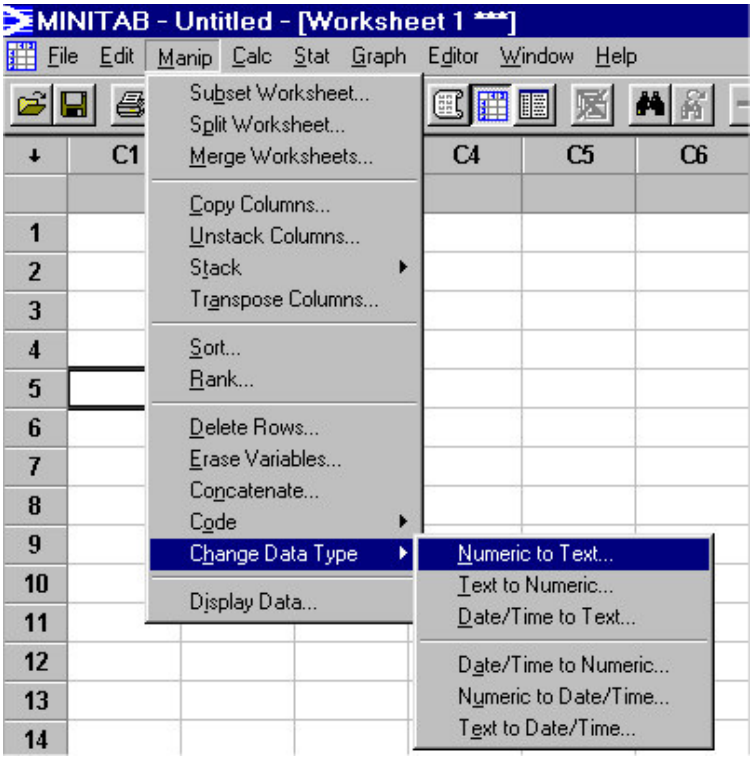
1.3.5 Nhập liệu

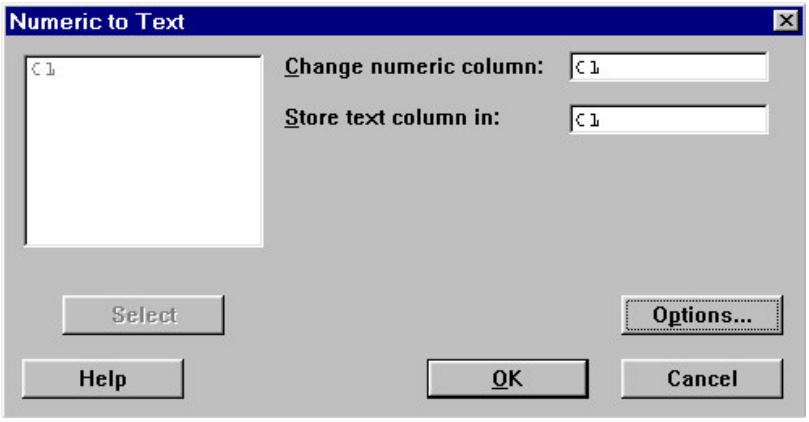
Minitab hỗ trợ nhiều cách để nhập liệu vào bảng tính. Bạn có thể nhập trực tiếp dữ liệu vào Session Window, mở một file mới, dán dữ liệu từ bộ nhớ hoặc phát dữ liệu. Để nhập liệu, chỉ cần hiển thị ô bằng cách dùng mũi tên, nhấp chuột và nhập dữ liệu.

Nhập liệu trực tiếp: Minitab nhận dữ liệu dạng bảng từ số liệu gõ trực tiếp trong Worksheet, hoặc nhập liệu từ Session windows.

Nhập liệu từ file có sẵn: Minitab có thể truy suất file dữ liệu có sẵn của chính Minitab như các file có phần mở rộng dạng **.MTW** hay **.MTP,** hoặc mở file dữ liệu của ứng dụng khác như Excel bằng lệnh File>open worksheet> và chọn file dữ liệu muốn mở.

Hình 1.7 Lệnh đổi loại dữ liệu





Lưu ý, dữ liệu của Minitab cũng được lưu trong file project. Nếu file dữ liệu là dạng project, lệnh để truy xuất file này sẽ là File>Open Project> và chọn file muốn mở. Nếu như bạn mở Project bằng Worksheet hay ngược lại, dữ liệu sẽ không xuất hiện.

Ví dụ 1:
Cho Bảng 1.2 thể hiện hoanh thu của hai loại mặt hàng trong 12 tháng của năm 2001. Nhập tập dữ liệu này vào Minitab.

Bảng 1.2 Doanh thu của hai loại sản phẩm trong 12 tháng của năm 2001 (triệu đồng).

Tháng	Sản Phẩm 1	Sản Phẩm 2
1	103	80
2	115	91
3	123	70
4	120	76
5	129	81
6	130	86

Tháng	Sản Phẩm 1	Sản Phẩm 2
7	132	75
8	136	70
9	140	80
10	142	86
11	146	90
12	150	73

Lời giải: Tạo ba cột, cột 1 đặt tên là "Thang", cột 2 "SP1", cột 3 "SP2" và nhập các giá trị trong bảng, kết quả worksheet sẽ thể hiện như hình 1.8. lưu file dữ liệu này thành doanhthu.mtp theo thao tác **File > Save Current Worksheet As,** Chọn thư mục, Đặt tên file dữ liệu: doanhthu, Chọn **Minitab Portable** trong **Save As Type**, **OK.**

Hình 1.8 Worksheet doanh thu của hai loại sản phẩm

Worksheet 1 ***

↓	C1	C2	C3	C4
	Thang	SP1	SP2	
1	1	103	80	
2	2	115	91	
3	3	123	70	
4	4	120	76	
5	5	129	81	
6	6	130	86	
7	7	132	75	
8	8	136	70	
9	9	140	80	
10	10	142	86	
11	11	146	90	
12	12	150	73	
13				

1.3.6 Sửa chữa nội dung

Khi nhập liệu, bạn cần phải đọc kỹ lại những giá trị và tên biến, nếu như bạn ghi sai, cần phải sửa chữa nội dung của từng ô dữ liệu bằng cách xóa, ghi lại, chèn thêm bằng các thao tác chuột hoặc bàn phím trước khi thực hiện phân tích. Bạn có thể sử dụng lệnh Undo của Minitab để khôi phục lại việc làm trước bằng cách dùng menu hoặc click vào icon  trên thanh lệnh.

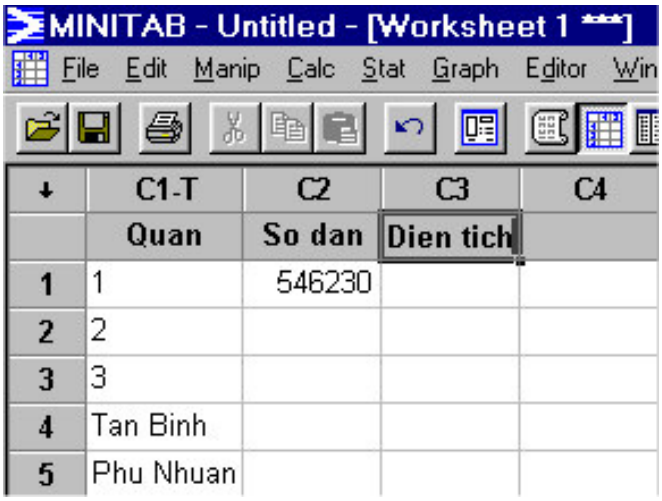
1.3.7 Đặt tên cột

Bạn có thể tham chiếu cột bằng tiêu đề của cột (C1), hoặc bạn có thể đặt tên riêng và tham chiếu bằng tên vì có thể nhớ tên của dữ liệu dễ hơn số cột. Ngoài ra, nếu như bạn dùng các lệnh thao tác trong Session Window, việc tham chiếu này sẽ thuận lợi hơn. Khi sử dụng các phần mềm xử lý dữ liệu, việc sử dụng tiếng Việt sẽ là một hạn chế, một số chương trình không sử dụng font hỗ trợ. Vì vậy, đối với các dữ liệu, để thuận tiện cho việc tham chiếu và xử lý số liệu, các bạn có thể dùng tiếng Anh, ký hiệu hoặc tiếng Việt không dấu.

Lưu ý, khi đặt tên cột hoặc tên biến nên đặt ngắn gọn, thể hiện được đặc tính của biến và tránh lầm lẫn với các biến khác. Nếu biến có quá

nhiều ký tự sẽ gây khó khăn khi thực hiện các phép tính và tham chiếu.

Hình 1.9 Bảng nhập liệu và đặt tên biến.



↓	C1-T	C2	C3	C4
	Quan	So dan	Dien tich	
1	1	546230		
2	2			
3	3			
4	Tan Binh			
5	Phu Nhuan			

1.3.8 Đặt tên bảng tính

Trong một project có thể có nhiều worksheet, bạn có thể đặt tên cho worksheet để dễ kiểm soát và tham chiếu. Nếu bạn không đặt tên, Minitab mặc định đặt theo thứ tự là Worksheet 1, hoặc Worksheet 2. Để đặt tên bạn chọn **Window>manage worksheets** Chọn **Worksheet 1** trong hộp Worksheet, chọn nút **Rename** và gõ vào tên bạn mong muốn. >**OK**, >**Done.** Sau khi đã đặt tên cho worksheet, tên này sẽ xuất hiện trên thanh tiêu đề.

1.4 THỰC HIỆN TÍNH TOÁN & TRUY SUẤT KẾT QUẢ THỐNG KÊ TỪ MINITAB

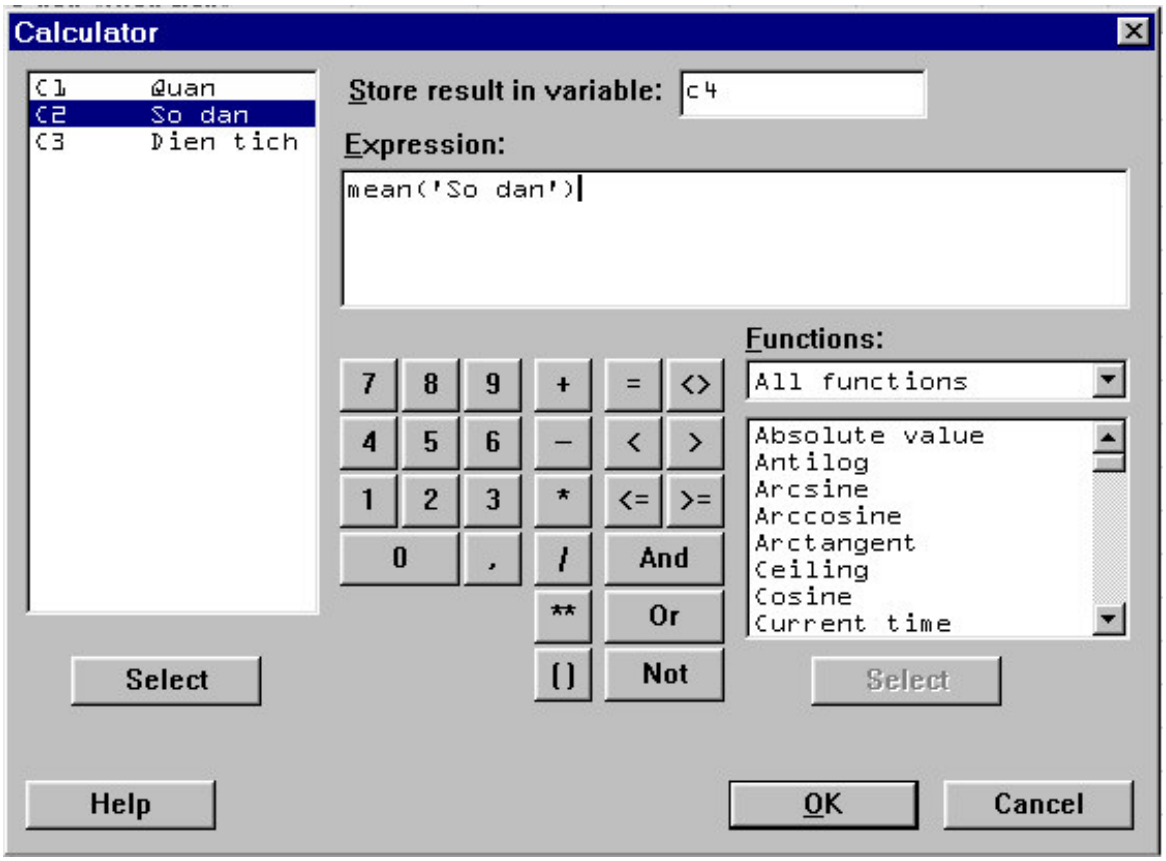
1.4.1 Các phép tính đơn giản

Bạn có thể nhập hoặc phát dữ liệu vào các ô bằng lệnh của Minitab. Các ô dữ liệu của Minitab không thể chứa các công thức giống như các ứng dụng bảng tính MS Excel hay Motus 1-2-3. Để thực hiện các phép tính trong Minitab, các bạn phải dùng Menu Calculator, chính vì thế các dữ liệu kết quả sẽ là dữ liệu tĩnh.

Có thể thực hiện các phép tính đơn giản bằng lệnh **Manip>calculator.** Màn hình lệnh này thể hiện trên Hình 1.10. lệnh yêu cầu nhập ô kết

quả **Store result in Variable.** Kết qủa này sẽ được lưu vào ột cột trên Data window của Minitab. Công thức tính được nhập vào phần **Expression**, bạn có thể dùng các hàm có sẵn của Minitab có sẵn trong phần **Functions** hoặc các chức năng hiện trên lựa chọn này để tính toán.

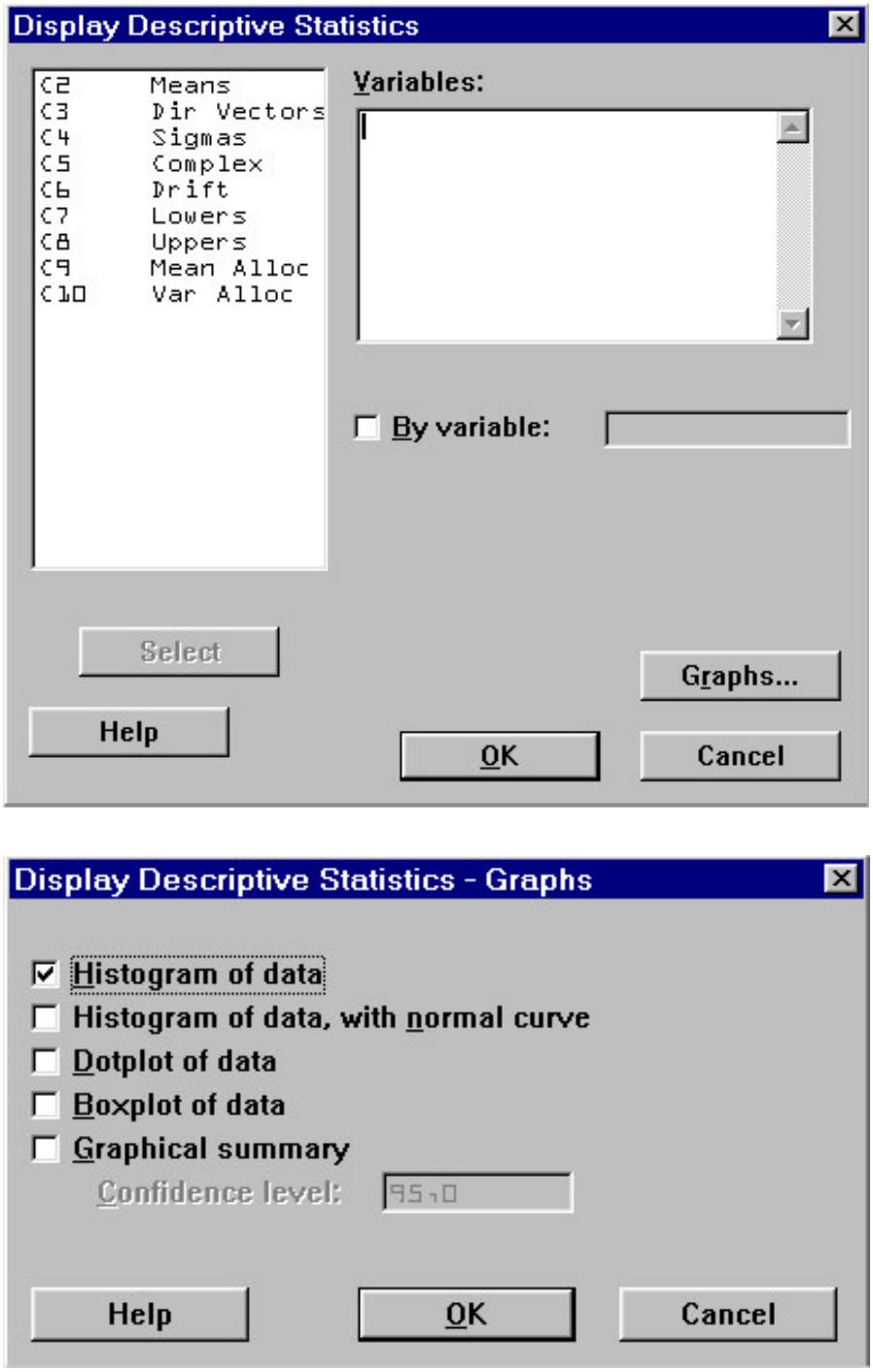
Hình 1.10 Lựa chọn tính toán của Minitab



1.4.2 Phát kết quả thống kê vào session window

Có thể dùng lệnh **Stat>Basic statistics> Display Descriptive Statistics** để phát kết qủa thống kê của tập dữ liệu, kết quả này sẽ thể hiện trong phần Session Window. Bạn có thể chọn một biến hoặc nhiều biến để trình bày các thông số thống kê của biến. Đồng thời cũng có thể mô tả kết quả bằng đồ thị trong lựa chọn **Graphs…** Lựa chọn này cho phép bạn mô tả bằng nhiều loại đồ thị khác nhau.

Hình 1.11 Lựa chọn mô tả thống kê



1.5 CÁC ĐỒ THỊ MINITAB

Những phiên bản mới của Minitab đề hỗ trợ cả hai loại đồ thị, ***dạng ký tự*** và dạng ***đồ thị chuyên dùng*** (Professional graphics). Hiện nay đồ thị ký tự đã lạc hậu, ít người sử dụng, kết quả của đồ thị chữ thể hiện trong Session Window. Hầu hết các kết quả đồ thị của Minitab đều

được thể hiện dưới dạng đồ thị chuyên dùng, chất lượng hình ảnh sẽ đẹp và hoàn thiện hơn.

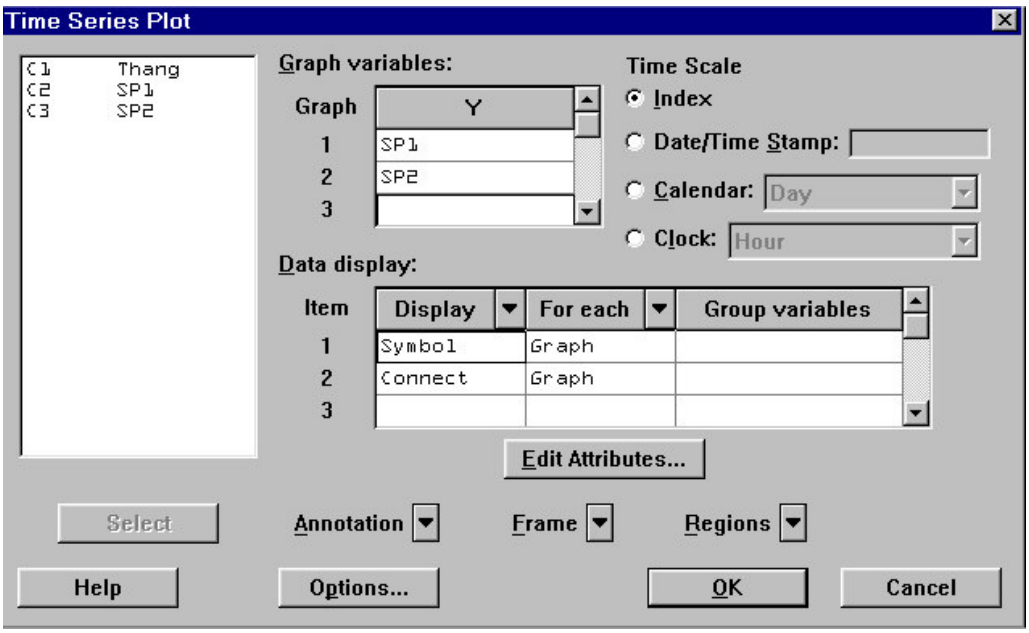
Ví dụ 2:
Bảng 1.2 trong ví dụ 1 thể hiện hoanh thu của hai loại mặt hàng trong 12 tháng của năm 2001. Nhập tập dữ liệu này vào Minitab và so sánh độ dao động của doanh thu hai mặt hàng bằng đồ thị.

Lời giải: Dữ liệu của bài này đã được nhập trong ví dụ 1, mở file dữ liệu bằng lệnh **File>open worksheet** chọn file **doanhthu.mtp**. Ví dụ này là chuỗi dữ liệu thời gian, dữ liệu được thu thập theo từng khoảng.

Vẽ đồ thị:

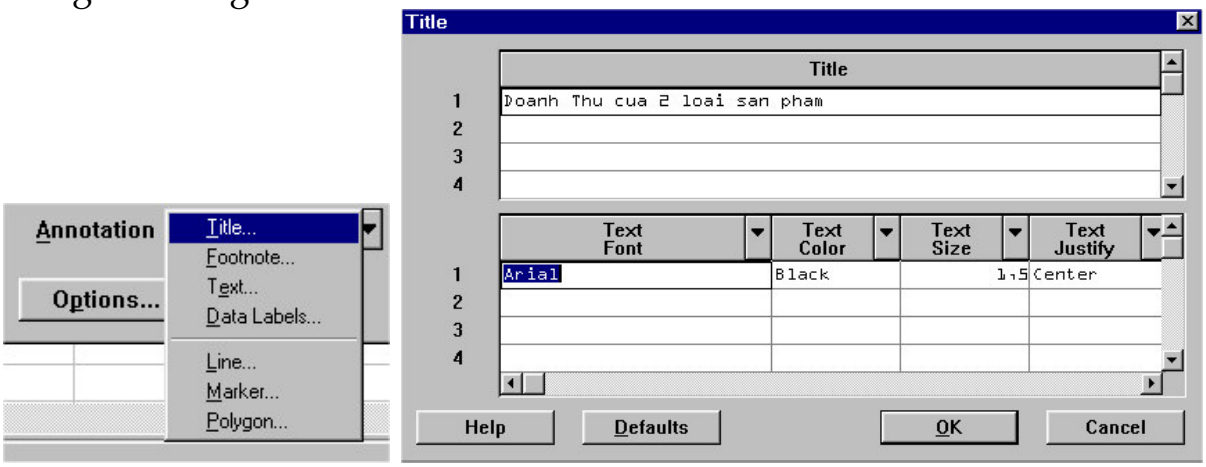
Graph > Time series Plot
Chọn SP1 **Graph** 1Y (xem hình 1.12)
Và SP2 trong **Graph** 2Y;
Trong **Data Display**, chọn **Connect - for each - Graph**;
Chọn Symbol **for each** Graph
Click **Annotation** và **Title**; và đặt tựa đề
Click **Options**
Nhập 1:12 trong **Index**:
Click **Frame**, chọn **Axis**
Trong hàng 1 của **Label**, nhập Thang
Trong hàng 2 của **Label**, nhập Doanh Thu
Click **Frame**, chọn **Multiple Graphs**,
Chọn **Overlay graph on the same page**, và **OK**

Hình 1.12 Các màn hình trong công cụ vẽ đồ thị của Minitab



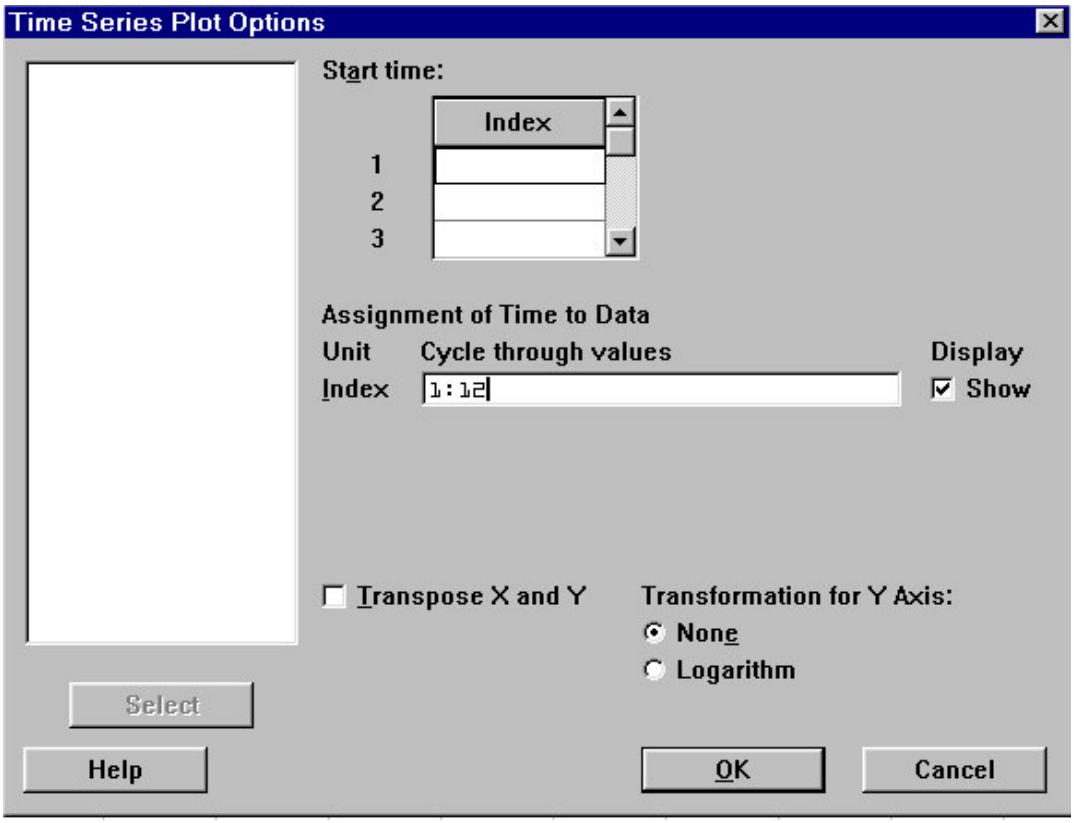
(a) màn hình khi thực hiện lệnh **Graph > Time Serries Plot**

Phần Graph variables cho phép chọn loại chuỗi sữ liệu sẽ vẽ trên biểu đồ, với ví dụ này, vẽ chuỗi dữ liệu doanh thu của hai loại sản phẩm trong 12 tháng.



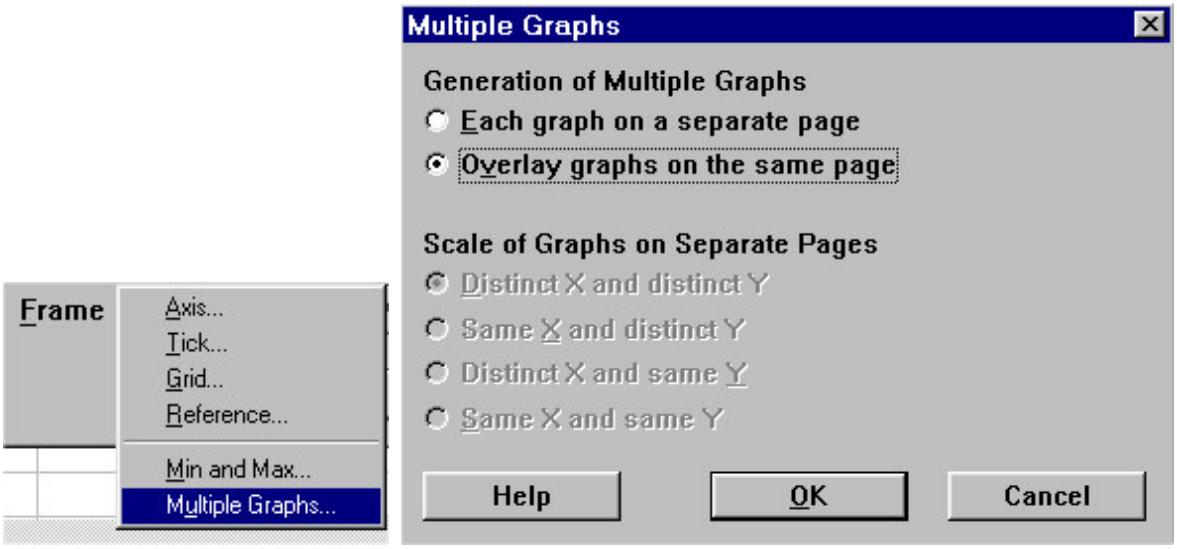
(b) Màn hình khi thực hiện **Graph > Time Serries Plot >Title**

Phần **Annotation** (chú thích) trong **Time Serries Plot** cho phép người dùng lựa chọn các chú thích cho đồ thị muốn vẽ. Phần **Title** trong **Annotation** cho phép bạn đặt tên đồ thị và chọn font chữ cho tựa đề.



(c) Màn hình khi thực hiện **Graph > Time Serries Plot > Options**

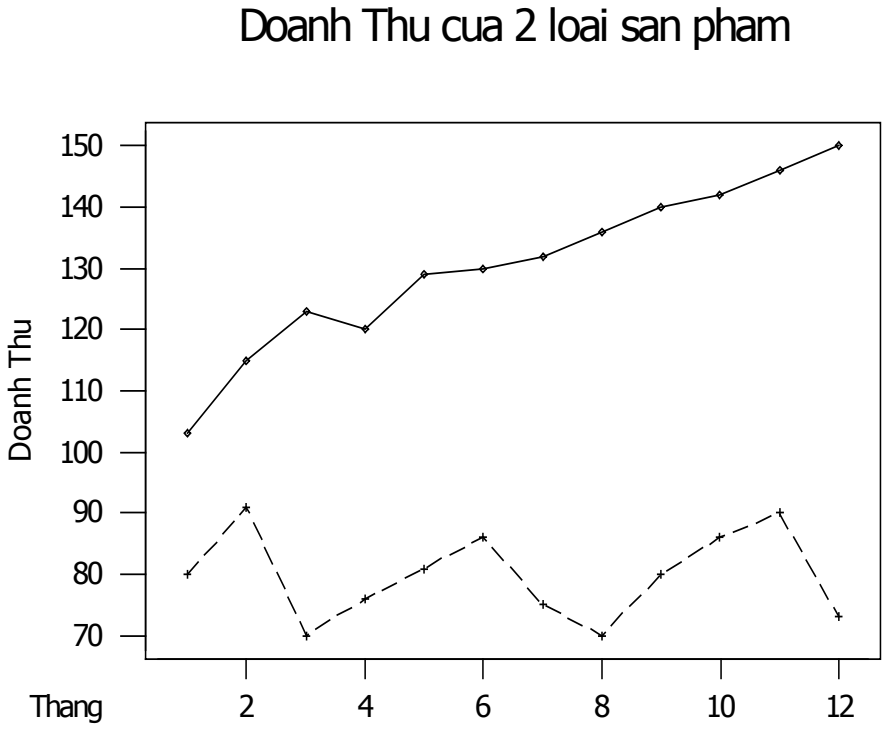
Option trong **Time Serries Plot** cho phép người dùng lựa chọn chỉ số cho đồ thị cũng như đơn vị sẽ dùng trong chuỗi thời gian.



(d) Màn hình khi thực hiện **Graph > Time Serries Plot > Frame> Multiple Graphs**

Phần **Frame** trong **Time Serries Plot** hỗ trợ định dạng đồ thị. Chức năng **Multiple Graphs** giúp vẽ nhiều loại đồ thị với dữ liệu đã có. Lựa

chọn **Overlay Graphs on the same page** giúp trình bày tất cả các chuỗi dữ liệu trên cùng một đồ thị.



(e) Đồ thị doanh thu của hai loại sản phẩm trên thị trường.

Dữ liệu theo thời gian thể hiện sự dao động Doanh thu của hai loại sảm phẩm. Vẽ hai tập dữ liệu này trên cùng một hình sẽ giúp so sánh khả năng sinh lợi cũng như mức độ dao động của doanh thu từng loại sản phẩm. Doanh thu của Sản phẩm 1 liên tục tăng trong 12 tháng, trong khi đó doanh thu Sản phẩm 2 biến động nhiều theo mùa. Người quản lý cần nhận xét được đặc tính của sản phẩm để có thể ra những quyết định thích hợp.

1.6 QUẢN LÝ DỮ LIỆU

1.6.1 File dữ liệu của Minitab

Dữ liệu của Minitab gồm các nhóm chính: Project files, Worksheet files, Session Window files (chữ) và Graphic files. Bảng 1.2 phân loại các loại file dữ liệu của Minitab. Có thể mở (Open), tạo mới (New), lưu (Save) các loại file này trên **Menu File**. Project files sẽ bao gồm tất cả các loại dữ liệu có trong Project, bao gồm các Worksheet, Session window, các

đồ thị đã có trong project. Các dạng file còn lại chỉ lưu kết quả của dạng dữ liệu mặc định của nó.

Bảng 1.2 Các dạng file của Minitab

Đuôi mở rộng	*Loại file*	*Mô tả*
MPJ	Minitab Project	Tất cả các thực thể của Minita liên quan đến phần phân tích thống kê
MTW hay MTP	Bảng tính Minitab	Dạng file chỉ chứa dữ liệu bảng tính. MTW file thường được dùng trong một môi trường máy tính PC, còn MTP có thể dùng cho cả máy Macintosh.
TXT	File Chữ	Bạn có thể lưu nội dung của Session, history và thông tin ở dạng file text, tất cả các ứng dụng khác đểy có thể sử dụng file này.
RFT	Rich Text format	Bạn có thể nội dung của session window dưới dạng RFT nếu như bạn muốn định dạng về sau.
MGF	Minitab Graphics Format	Các đồ thị có thể lưu dưới dạng MGF và có thể sử dụng trên môi trường Minitab

1.6.2 Truy xuất dữ liệu từ các ứng dụng khác

Từ file bên ngoài: Minitab lưu dữ liệu dưới nhiều dạng tùy vào chức năng và yêu cầu. Một số dạng file dữ liệu có thể dùng trong Minitab được thể hiện trong Bảng 1.2. Tuy nhiên bạn có thể nhập dữ liệu từ các file dữ liệu bên ngoài từ Lotus 1-2-3, MS Word, MS Excel hoặc các file dữ liệu Minitab khác bằng câu lệnh **File>open Worksheet** và tìm đúng loại file trong hộp thoại **List Files of Type**, chọn file dữ liệu bạn muốn mở.

Từ bộ nhớ tạm thời (Clipboard)
Bạn có thể sao chép dữ liệu từ các ứng dụng khác vào Window Clipboard và dán vào Minitab. Khi bạn cắt hoặc sao chép, dữ liệu sẽ được đưa vào Clipboard. Phần bộ nhớ này sẽ được Clipboard chia sẻ cho tất các ứng dụng trong Windows. Dùng câu lệnh **Edit>Paste cells**

để chép dữ liệu vào Minitab. Bạn có thể dùng phím tắt để thực hiện các lệnh này.

1.6.3 Xử lý tập dữ liệu

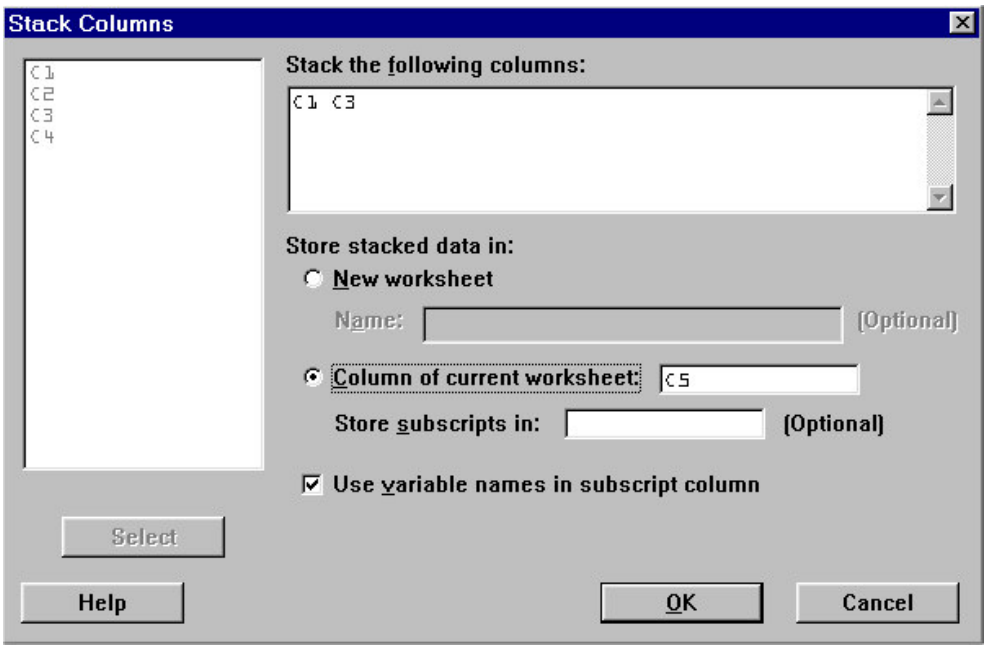
Gộp và tách dữ liệu: (stacking và unstacking data)
Trong Minitab, thông thường một cột chỉ chứa dữ liệu của một biến, với mỗi quan sát trong mỗi hàng hoặc điểm dữ liệu trong từng ô gọi là **Unstacked data.** Dữ liệu có thể ở dạng stacked nếu như các điểm dữ liệu thuộc nhiều nhóm dữ liệu.

Ví dụ 3: Cho tập dữ liệu gồm 4 cột như bảng dưới đây, kết hợp dữ liệu trong C1 và C3 và lưu vào C5

	C1	C2	C3	C4
1	-2	23	-6	45
2	-6	37	-8	62
3	-3	41	-9	65

Lời giải: nhập tập dữ liệu vào Worksheet của Minitab, dùng lệnh Manip >Stack>Stack Culumns để stack dữ liệu theo yêu cầu.

Hình 1.14 Chức năng Stack của Minitab



Kết hợp dữ liệu trong C1 và C3 và lưu vào C5

Manip >Stack>Stack Culumns

Select C1 và C3 trong **Stack the following columns**

Store stacked data in:

Nhấp chuột **Column of current worksheet**: và nhập vào C5 và OK.

Kết hợp dữ liệu trong C2 và C4 vào C6 và lưu tên biến ký hiệu vào C7

Manip >Stack>Stack Culumns

Select C1 và C3 trong **Stack the following columns**

Store staked data in:

Nhấp chuột **Column of current worksheet**: và nhập vào C6;

Nhập C7 vào **Store subscipts in**:

Nhấp **Use variable names in Subscript culumn**. OK

Bảng tính mới sẽ là

	C1	C2	C3	C4	C5	C6	C7
1	-2	23	-6	45	-2	23	C2
2	-6	37	-8	62	-6	37	C2
3	-3	41	-9	65	-3	41	C2
4					-6	45	C4
5					-8	62	C4
6					-9	65	C4

Mã hoá dữ liệu (coding)

Lệnh CODE (mã hóa) biến dữ liệu chữ thành số hoặc ngược lại, lệnh này sẽ tìm kiếm các giá trị và thay đổi giá trị này bằng giá trị mới.

Ví dụ 4: (Xem hình 1.15)

Manip>code>text to Numberic

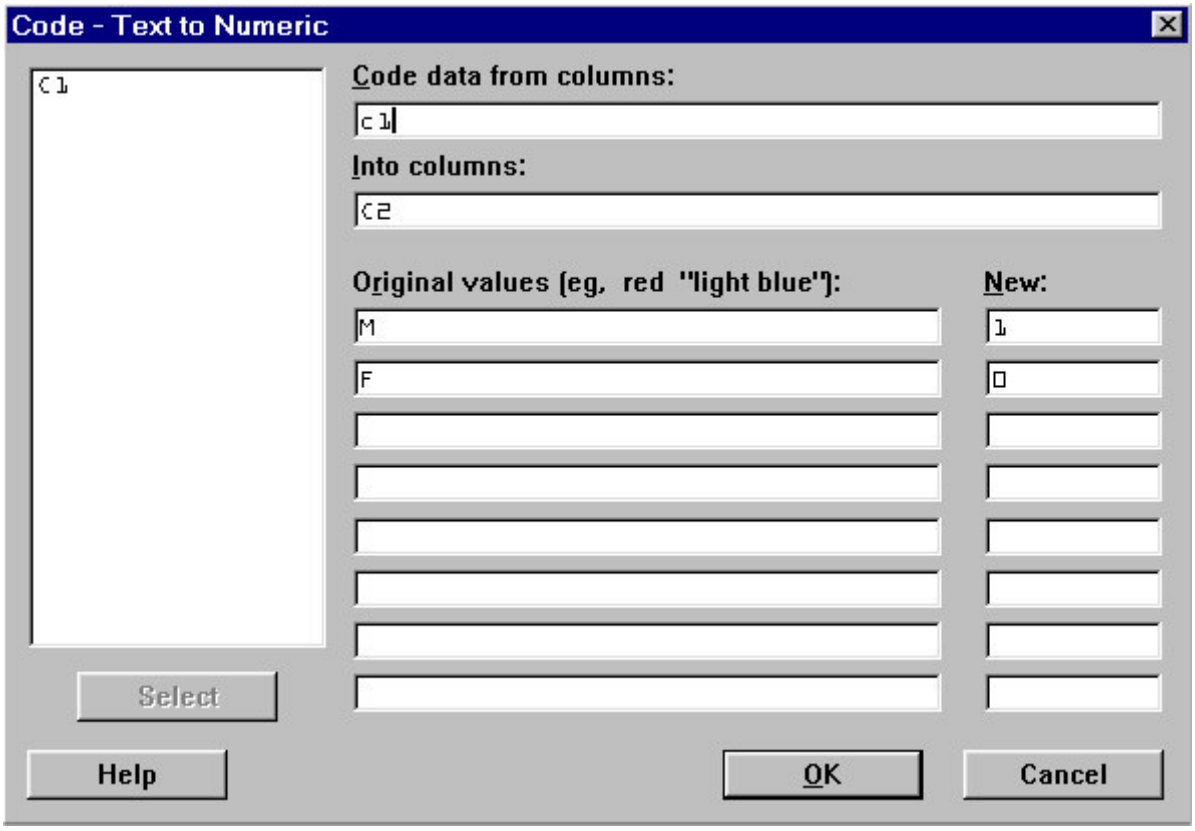
Chọn C1 trong **Code data from Columns**:

Nhập C2 trong **Into columns**:

Nhập M trong **Original Values** và 1 trong **New**

Nhập F trong **Original Values** và 0 trong **New**

Hình 1.15 Mã hoá dữ liệu



Bảng tính cũ

	C1-T
1	M
2	M
3	F
4	M
5	M
6	F

Bảng tính mới

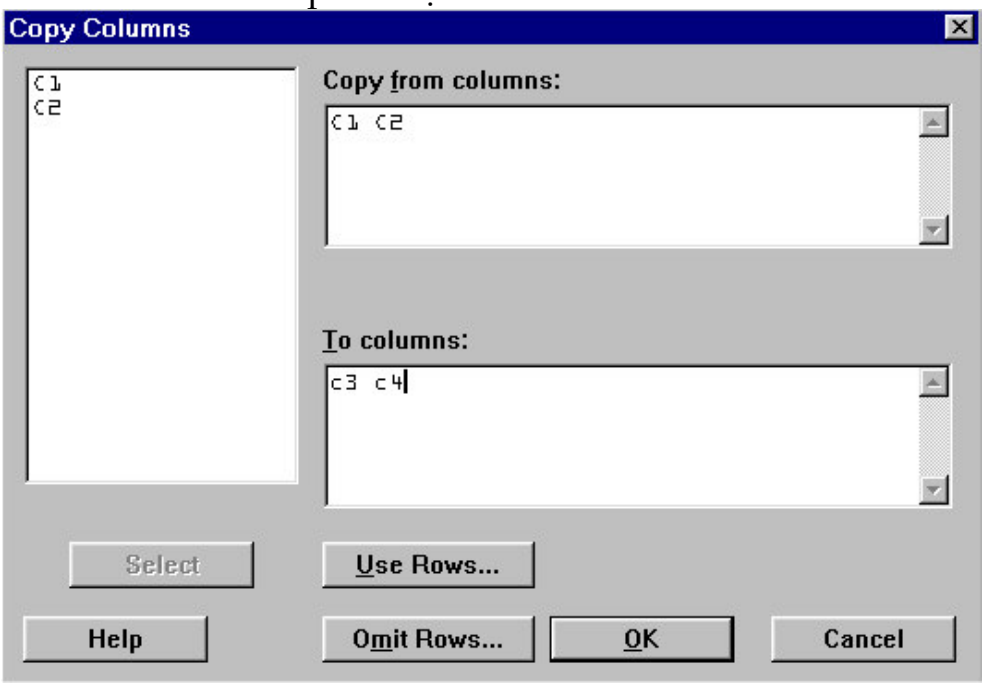
	C1-T	C1
1	M	1
2	M	1
3	F	0
4	M	1
5	M	1
6	F	0

Sao chép dữ liệu (Copy)

Lệnh này cho phép chép dữ liệu từ cột này đến cột khác. Thao tác thực hiện như sau **Manip>Copy Columns.**

Ví dụ 5: giả sử chúng ta có dữ liệu giới tính trong C1 và điểm thi trong C2, và muốn chép dữ liệu của các bạn nam vào C3 và C4.

Hình 1.16 Sao chép dữ liệu



Dữ liệu cũ

	C1-T	C2
1	M	83
2	M	88
3	F	72
4	M	95
5	M	76
6	F	87

Dữ liệu mới

	C1-T	C2	C3	C4
1	M	83	M	83
2	M	88	M	88
3	F	72	F	72
4	M	95	M	95
5	M	76	M	76
6	F	87	F	87

Tìm kiếm giá trị tích lũy

Lệnh này tính và lưu giá trị tổng tích lũy của những hàng cụ thể bằng hàm PARSUM.

Calc > Calculator

Nhập C2 trong **Store result in variable**:

Nhập PARSUM (C1) trong **Expression**, OK

Bảng tính cũ

	C1
1	6
2	9

Bảng tính mới

	C1	C2
1	6	6
2	9	15

3	8
4	3
5	6
6	0
7	1

3	8	23
4	3	26
5	6	32
6	0	32
7	1	33

Sắp xếp dữ liệu (Sort)

Lệnh Manip>Sort của Minitab cho phép bạn sắp xếp dữ liệu theo thứ tự tăng dần hoặc giảm dần theo một tiêu chí nào đó.

Ví dụ 6: sắp xếp bảng dữ liệu cột C1 theo thứ tự tăng dần và lưu trong cột C2.

Manip>copy
Sort column(s) C1
Store sorted culum(s) in C2
Sort column by C1 OK
Kết quả sẽ là

Bảng tính cũ

C1
78
97
84
84
69
42
98
84
78

Bảng tính mới

	C1	C2
1	78	42
2	97	69
3	84	78
4	84	78
5	69	84
6	42	84
7	98	84
8	84	97
9	78	98

Xếp hạng dữ liệu (Rank)

Lệnh RANK xếp hạng giá trị của dữ liệu theo thứ tự tăng dần.
Manip > Rank
Chọn C1 trong **Rank Data in**:
Nhập C2 tring Store ranks in, OK; (C1) trong **Expression**, OK

Bảng tính cũ

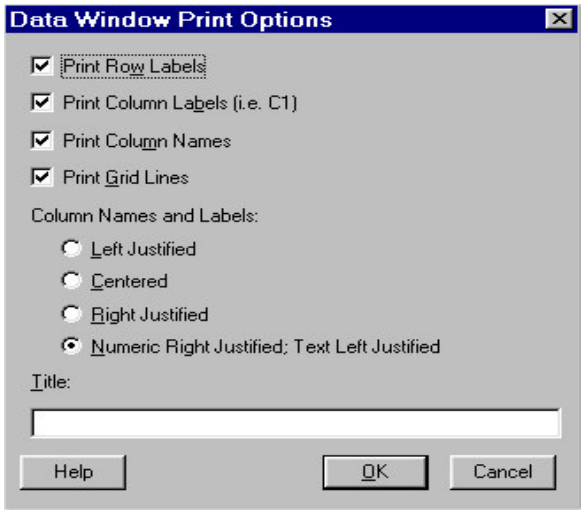
	C1
1	78
2	97
3	84
4	84
5	69
6	42
7	98
8	84
9	78

Bảng tính mới

	C1
1	78
2	97
3	84
4	84
5	69
6	42
7	98
8	84
9	78

1.7 IN CÔNG TÁC

Để in các công việc, dữ liệu, đồ thị, báo cáo, v.v dùng câu lệnh trên thanh **File >Print** Bạn có thể chọn những lựa chọn khác nhau cho chức năng in như trong hình dưới đây. Bạn cũng có thể chọn sao chép dữ liệu hoặc kết quả sang một phần mềm xử lý văn bản khác (như MS Word) để in ra.



Lựa chọn in của Minitab cho phép người dùng chọn lựa những yêu cầu có trong kết quả in như tựa đề của hàng (Print Row Labels), tựa đề của cột (Print Column Labels), tên của Cột (Print Column Names), in các đường ngăn cách dữ liệu (Print Grid lines). Ngoài ra còn có thể lựa chọn vị trí của ký tự và số trong ô dữ liệu và tựa đề của dữ liệu cần in.

BÀI TẬP

1.1 Tạo một Minitab project và đặt tên là Class. Nhập dữ liệu trong bảng dưới đây vào Data window.

MSSV	Điểm 1	Điểm 2
38	74	81
77	93	100
19	52	40
12	78	69
63	80	82

1. Thêm một hàng cho sinh viên số 22 giữa hàng 4 và 5 với số điểm là D1=90 D2=88
2. Xóa dữ liệu của sinh viên số 19 đã bỏ lớp
3. Chép cột MSSV vào C5
4. Lưu dữ liệu

1.2 Dùng tập dữ liệu trong bài 1.1

1. Xóa cột C5 trong đó có chứa dữ liệu sao chép của MSSV
2. Gán dữ liệu của cột C1 dạng text
3. Đặt tên cột C2 với giải thích "diem kiem tra ngày 21/3" (dùng lệnh Editor> column>description) và cột C3 với "diem kiem tra ngày 15/5"
4. Stack hai cột điểm thành 1 cột C8, nhập subscripts D1 và D2 trong C9

1.3 Dùng tập dữ liệu 1.1

1. Xoá C4-C10 trong Class data
2. Sắp xếp dữ liệu C1-C3 theo MSSVvà đặt dữ liệu đã sắp xếp vào C5-C7
3. Xếp hạng sinh viên theo D2 và đặt dữ liệu đã xếp hạng vào C8, đặt tên mới cho cột này là "Hạng"
4. Tạo một cột mới C9 – "DTB" chứa giá trị trung bình của D1 và D2
5. Mã hóa điểm trung bình C9 thành dạng điểm chữ, lưu điểm chữ trong C10 với F=50-59; D=60-69; C=70-79; B=80-89; A=90-100

1.4 Dữ liệu dưới đây là bảng danh sách các giới nghệ sĩ có thu nhập cao nhất được sắp xếp theo mức thu nhập dự báo trong năm 1995 và 1996 (đơn vị là triệu USD)

1. Nhập tập dữ liệu vào máy tính, kiểm tra lại số liệu.
2. Sắp xếp dữ liệu theo thứ tự ABC và theo thu nhập. Lưu dữ liệu qua một cột khác.

	Tên	Thu nhập
1	Winfrey, Oprah	171
2	Spielberg, Steven	150
3	Beatles	130
4	Jackson, Michael	90
5	Rolling Stones	77
6	Eagles	75
7	Schwarzenegger, Arnold	74
8	Copperfield, David	63
9	Carrey, Jim	63
10	Crichton, Michael	59
11	Seinfeld, Jerry	59
12	King, Stephen	56
13	Brooks, Garth	51
14	Webber, Andrew Lloyd	50
15	Hanks, Tom	50
16	Siegfried and Roy	48
17	Cruise, Tom	46
18	Ford, Harrison	44
19	Eastwood, Clint	44
20	R.E.M.	44

	Tên	Thu nhập
21	Stallone, Sylvester	44
22	Grisham, John	43
23	Williams, Robin	42
24	Zemeckis, Robert	42
25	Roseanne	40
26	Douglas, Michael	40
27	Willis, Bruce	36
28	Pavarotti, Luciano	36
29	Kiss	35
30	Schulz, Charles	33
31	Cosby, Bill	33
32	Travolta, John	33
33	Carey, Mariah	32
34	Clancy, Tom	31
35	Costner, Kevin	31
36	Washington, Denzel	30
37	Letterman, David	28
38	Metallica	28
39	Gibson, Mel	28
40	Bullock, Sandra	25

1.5 Bảng dưới đây thể hiện tỷ lệ phần trăm số người sử dụng Internet trên tổng số dân theo báo cáo của Human Development Report Office. Nhập dữ liệu, vẽ đồ thị và nhận xét.

(% của tổng số dân)	1998	2000
United States	26.3	54.3
High-income OECD (excl.US)	6.9	28.2

Latin America and the Caribbean	0.8	3.2
East Asia and the Pacific	0.5	2.3
Eastern Europe and CIS	0.8	3.9
Arab States	0.2	0.6
Sub-Saharan Africa	0.1	0.4
South Asia	0.04	0.4
World	2.4	6.7

CHƯƠNG 2

MÔ TẢ DỮ LIỆU ĐỊNH TÍNH VÀ ĐỊNH LƯỢNG

Thống kê là một môn khoa học – khoa học của thông tin. Công việc phân tích thống kê thông thường tổng hợp dữ liệu, tóm tắt, và trình bày theo một cách có ý nghĩa. Thông tin có thể là *định tính* (Qualitative) hay *định lượng* (Quantitative), việc phân biệt này rất có ý nghĩa trong việc phân tích thống kê. Ví dụ, một người tìm mua hai căn nhà với yêu cầu giá dưới 2 tỷ đồng, diện tích hơn $300m^2$, nhà mặt tiền quay về hướng Đông, có vườn và gara xe. Với thông số này, giá mua là biến định lượng bởi vì nó chứa số lượng thông qua giá trị tiền đồng. Số căn nhà muốn mua cũng là biến định lượng. Hướng của căn hộ là biến định tính vì nó thể hiện tính chất (đông, tây, nam bắc). Có vườn và gara là biến định tính. Biến định lượng được mô tả bằng một con số qua đó có thể thực hiện các phép toán đại số như là lấy giá trị trung bình. Biến định tính (hay biến phân loại) đơn giản chỉ ghi nhận tính chất. Chương này trình bày cách mô tả hai loại biến định tính và định lượng thông qua các công cụ đồ thị và bảng.

NỘI DUNG

2.1 CÁC ĐỒ THỊ MÔ TẢ DỮ LIỆU ĐỊNH TÍNH

Dữ liệu định tính bao gồm dữ liệu dạng *chỉ danh* (Nominal data) và dữ liệu *thứ tự* (Ordinal data). **Dữ liệu chỉ danh** hay dữ liệu phân loại là các đại lượng đo xác định loại đơn vị trong mẫu hoặc tập hợp chính ví dụ như nhãn hiệu, ngành của sinh viên năm nhất, hoặc ý kiến về một vấn đề cụ thể. **Dữ liệu thứ tự** là các các đại lượng thể hiện một số thứ tự hay thứ hạng của đơn vị đo trong mẫu hoặc tập hợp chính. Dữ liệu thứ tự bao gồm tất cả các thông tin về dữ liệu chỉ danh và thứ tự của dữ liệu. Ví dụ, thứ tự đánh giá thứ hạng sinh viên được phân loại như xuất sắc, giỏi, khá, trung bình, yếu kém. Hoặc để phân loại thương tật người ta đánh giá mức độ nghiêm trọng (1) và không nghiêm trọng (5). Bởi vì dữ liệu thư tự chỉ đơn thuẩn sắp xếp đơn vị, vì vậy việc tính toán đại số của các giá trị này thường không có ý nghĩa.

1.2.1 Phân phối tần suất

Lệnh Tally trong Minitab được dùng để tóm tắt dữ liệu định tính hay hay thứ tự. Có nhiều lựa chọn cho kết quả như phân phối tần suất, tần suất phần trăm, tần suất tích lũy và tần suất phần trăm tích lũy.

Câu lệnh: **Stat > tables > Tally**

TALLY C…C	Tổng kết dữ liệu trong bảng
COUNTS	Số quan sát trong mỗi loại
PERCENTS	Tần suất Phần trăm
CUMCOUNTS	tần suất tích lũy
CUMPERCENTS	Phần trăm tích lũy
ALL	Tính tất cả các thông số trên
STORE C…C	Lưu lại kết qủa

Ví dụ 1:
Tập dữ liệu 30 sinh viên và loại xe đang sử dụng. Tìm phân phối tần suất và phần trăm của tập dữ liệu.

Bảng 2.1 Loại xe của 30 sinh viên trong mẫu nghiên cứu.

Honda	Suzuki	Honda	Honda	Suzuki	Honda
Honda	Honda	Sym	Honda	Honda	Honda
Suzuki	yamaha	Honda	suzuki	Yamaha	Honda

Yamaha	Sym	Sym	Yamaha	Sym	Yamaha
Yamaha	Xedap	xedap	Yamaha	Xedap	Yamaha

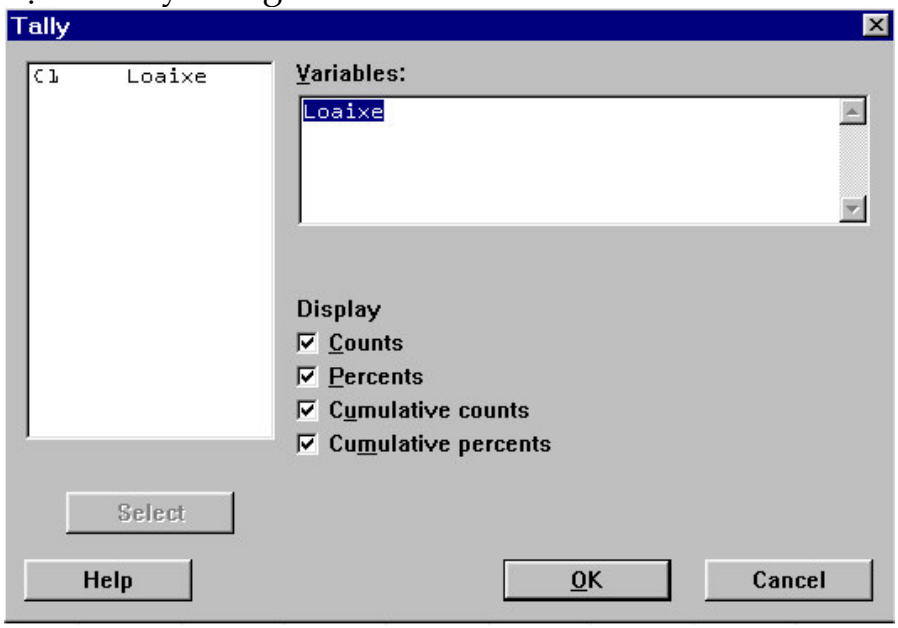
Lời giải: Nhập dữ liệu 30 loại xe vào cột C1 và đặt tên là Loaixe. Dùng lệnh Tally đến tìm các phân phối dữ liệu.

Stat > Tables > Tally

Select loaixe trong **Variables**:

Chọn **Counts, Percents**, **Cumulative counts**, và **Cumulative Percents**.

Hình 2.1 Lệnh Tally trong Minitab



Kết quả

Tally for Discrete Variables: loaixe

Loaixe	Count	CumCnt	Percent	CumPct
Honda	11	11	36,67	36,67
Suzuki	5	16	16,67	53,33
Sym	4	20	13,33	66,67
Xedap	2	22	6,67	73,33
Yamaha	8	30	26,67	100,00
N=	30			

Kết quả thấy rằng phần lớn sinh viên (36,7%) sử dụng xe của Honda, 16% sử dụng xe Suzuki, 6,67% sinh viên sử dụng xe đạp.

2.1.2 Biểu Đồ Thanh

Để trình bày các dữ liệu định tính, người ta dùng nhiều loại đồ thị. Lệnh Chart trong Minitab tạo ra rất nhiều kiểu đồ thị gồm đồ thị thanh, đồ thị đường thẳng, đồ thị diện tích. Biểu đồ thanh (bar chart) được dùng để mô tả dữ liệu định tính khi không cần nhấn mạnh đến phần trăm của từng phân loại (biến).

Có hai cách dựng đồ thị trong Minitab, loại thứ nhất dùng tất cả các nhập liệu của quan sát trong một cột và giá trị của quan sát trong cột còn lại, chiều cao của thanh sẽ là giá trị đếm của quan sát. Loại thứ hai dùng hai cột, một cột chứa quan sát của trục biến y, và trục khác chứa nhóm các biến trục X.

Câu lệnh: **Graph > Chart**

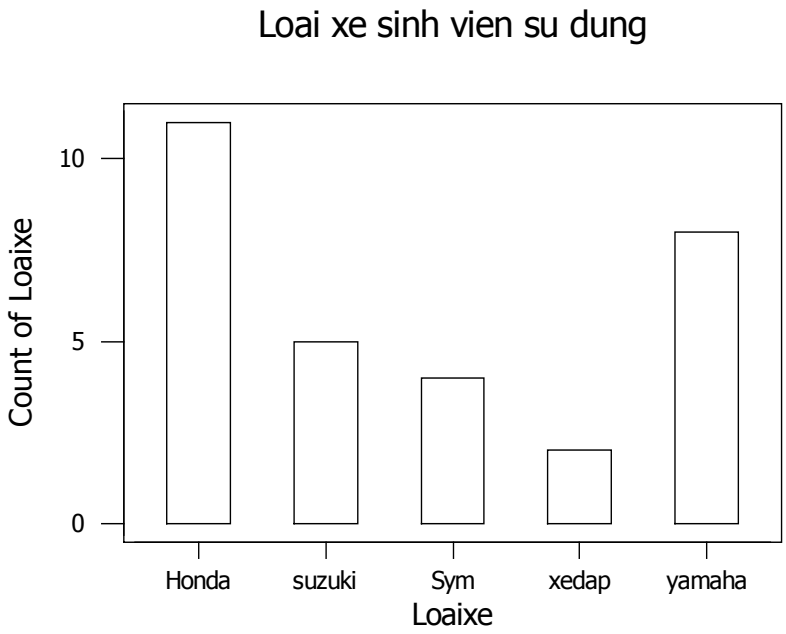
CHART C	Dùng cho dạng đơn biến
CHART C*C	Dùng dạng thứ hai, cột thứ hailà biến nhóm
INCREASING	Tăng thứ tự dựa trên giá trị của biến đầu tiên
DECREASING	Giảm thứ tự dựa trên giá trị của biến đầu tiên
CPERCENT	Thang đo phần trăm
CUMMULATIVE	Thang đo tần suất tích lũy
Hàm	SUM, COUNT, N, NMISS, MEAN, MEDIAN, MINIMUM, MAXIMUM, STDEV, SSQ

Ví dụ 2:
Dùng dữ liệu trong ví dụ 1 để dựng biểu đồ cột

Graph > Chart
Trong Graph hàng 1
Chọn loaixe trong X
Nhấp chuột vào Annotation, chọn Title, nhập loai xe sinh vien su dung, OK, OK, ta có hình 2.2

Chiều cao của biểu đồ thể hiện số lượng xe của từng loại xe của sinh viên. Trong tập dữ liệu này, số lượng xe Honda được sử dụng nhiều nhất.

Hình 2.2 Đồ thị cột số lượng các loại xe sinh viên sử dụng



2.1.3 Biểu Đồ Tròn (Pie Chart)

Biểu đồ tròn trình bày dữ liệu ở dạng phần trăm trên tổng số. Thang đo có thể là chỉ danh hoặc thứ tự. Biểu đồ tròn trong Minitab lấy dữ liệu ở hai dạng. Loại thứ nhất dùng tất cả các quan sát trong 1 cột, mỗi múi của biểu đồ tương ứng với tần suất xuất hiện của từng quan sát. Loại thứ hai dùng hai cột, một cột chứa tên của quan sát và cột còn lại chứa tân suất của quan sát.

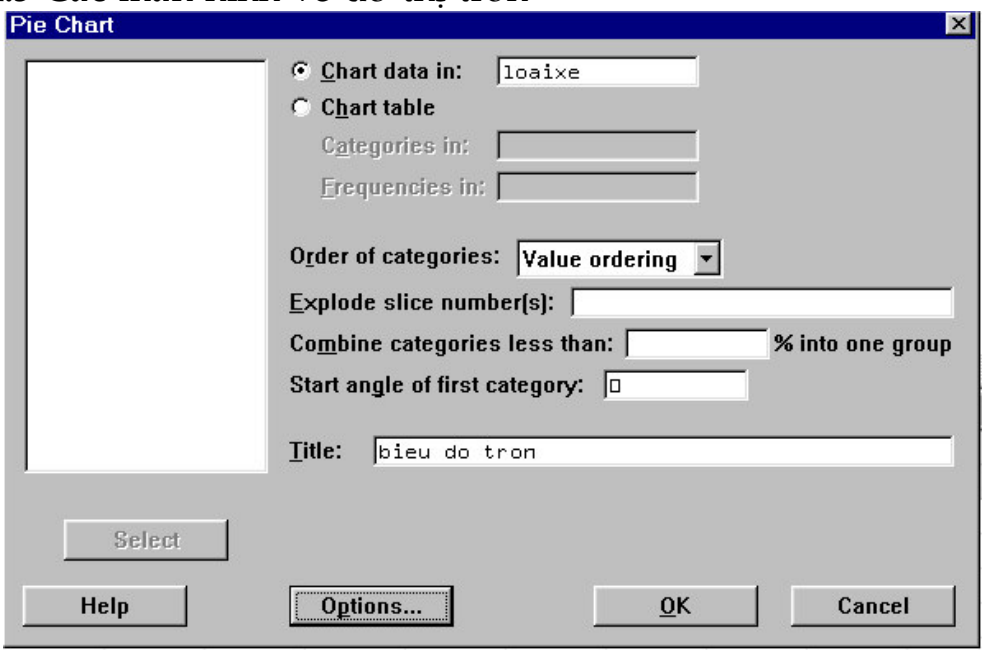
Câu lệnh **Graph > Pie Chart**

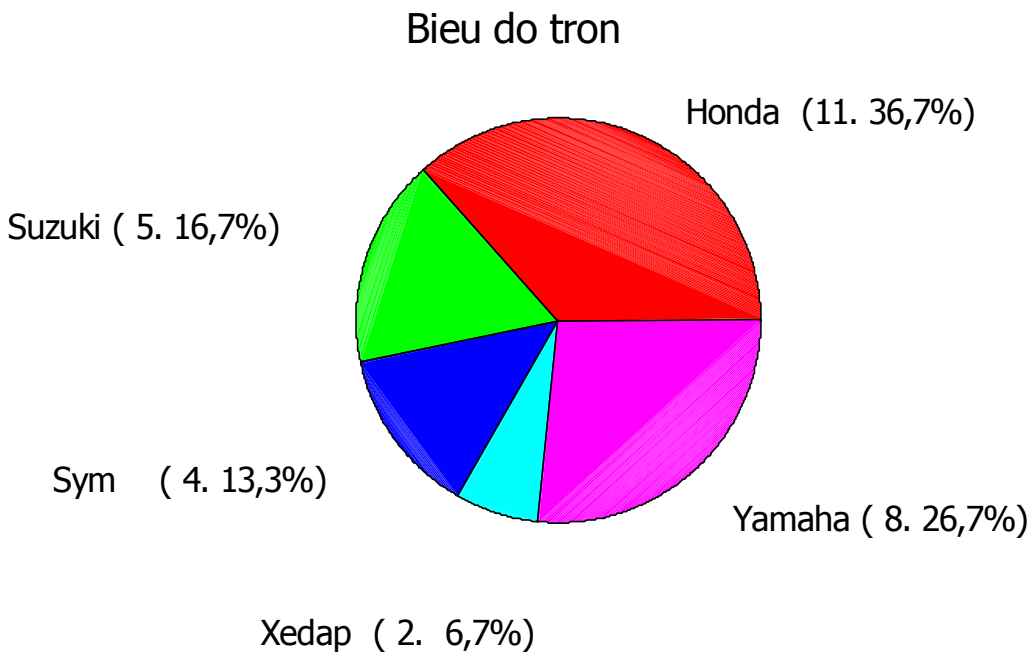
Ví dụ 3:
Vẽ đồ thị tròn từ dữ liệu trong ví dụ 1.

Lời giải: Dùng lệng Graph>Pie Chart

Graph>Pie chart
Chọn **Chart data in** "Loai xe"
Phần **Title** ghi vào "Bieu do tron", OK.

Hình 2.3 Các màn hình vẽ đồ thị tròn





Biểu đồ tròn thể hiện phần trăm số lượng xe sinh viên sử dụng, biểu đồ thể hiện hai dạng dữ liệu, tần suất và phần trăm.

2.1.4 Biểu Đồ Pareto

Biểu đồ Pareto đơn giản chỉ là phân phối tần suất của dữ liệu định tính được sắp đặt theo loại. Trong sản xuất, biểu đồ Pareto là một biểu đồ cột của những đặc tính của quá trình trong sản xuất – thông thường

là các vấn đề hoặc khuyết tật và các giá trị phần trăm của các biến này, tất cả cộng lại bằng 100%. Chú ý rằng, biểu đồ Pareto không xác định những khuyết tật hay vấn đề quan trọng nhất mà là xuất hiện nhiều nhất. Biểu đồ Pareto được sử dụng rất nhiều trong ứng dụng các phương pháp cải tiến chất lượng trong ngành dịch vụ.

Trong Minitab, biểu đồ Pareto nhận 2 loại dữ liệu, loại thứ nhất gồm tất cả các quan sát trong 1 cột. Dạng thứ hai dùng hai cột chứa tên của quan sát và tần suất của quan sát đó.

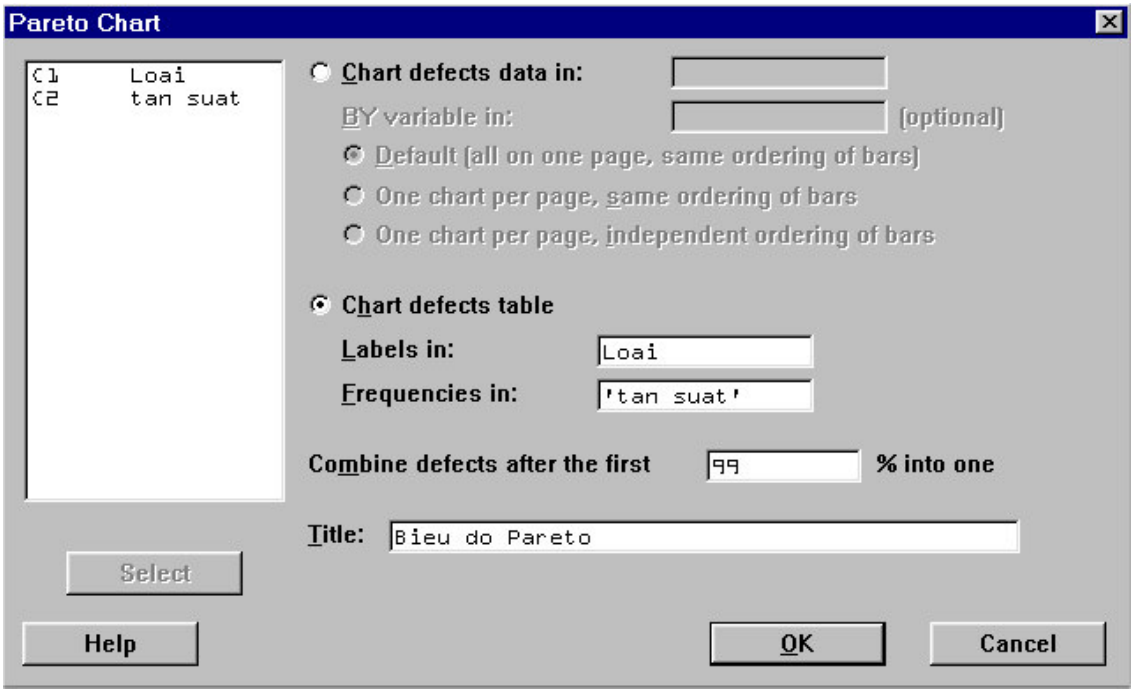
Câu lệnh: **Stat > Quality Tools > Pareto Chart**

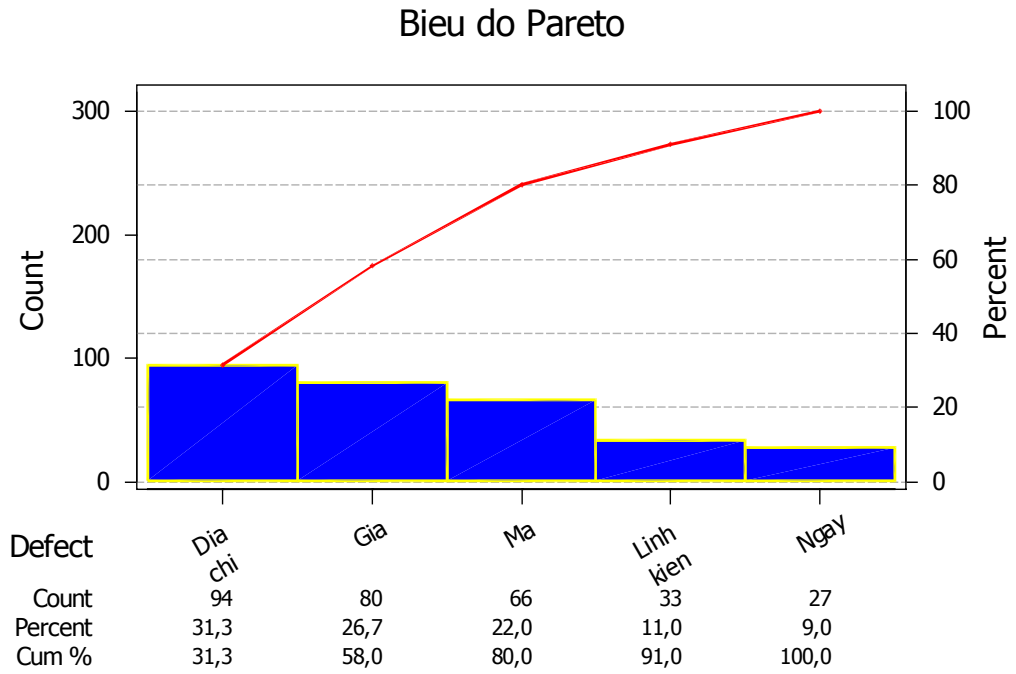
Ví dụ 4:
Một nhóm chất lượng đang điều tra những sai sót trong các đơn đặt hàng của doanh nghiệp nhằm hạn chế việc sửa đổi các đơn đặt hàng bởi vì mỗi lần sửa đổi sẽ tốn kém rất nhiều. Những loại lỗi và số lần xuất hiện thể hiện trong bảng dưới đây. Vẽ biểu đồ Pareto những sai sót trên.

Loại vấn đề	Loai	Số lần xuất hiện
Sai địa chỉ	Dia chi	94
Ghi sai giá	Gia	80
Sai mã nhà cung cấp	Ma	66
Sai linh kiện	Linh kien	33
Sai ngày giao hàng	Ngay	27

Lời giải: dùng lệnh
Stat > Quality Tools > Pareto Chart
Nhấp chuột vào **Chart defects table**
Select "loai" trong **Lables in:**
Select "tan suat" trong **Frequencies in:**
Nhập *99%* trong **Combine defects after the first**
Nhập "Biểu đồ Pareto" vào **Title**, OK

Hình 2.4 vẽ biểu đồ Pareto





Biểu đồ Pareto xếp các thanh dữ liệu từ lớn đến nhỏ và một đường thẳng chỉ phần trăm tích lũy của loại lỗi. Trục tung bên trái chỉ số lỗi, và trục tung bên phải chỉ tỷ lệ phần trăm của từng loại. Bảng ở dưới thể hiện số lượng, phần trăm và phần trăm tích lũy số lỗi. Dữ liệu cho thấy loại lỗi xuất hiện nhiều nhất là ghi sai địa chỉ (31,3%), kế đến là ghi sai giá (26,7%) và ghi sai mã (22%) ba loại lỗi này chiếm 80% số lỗi của hệ thống..

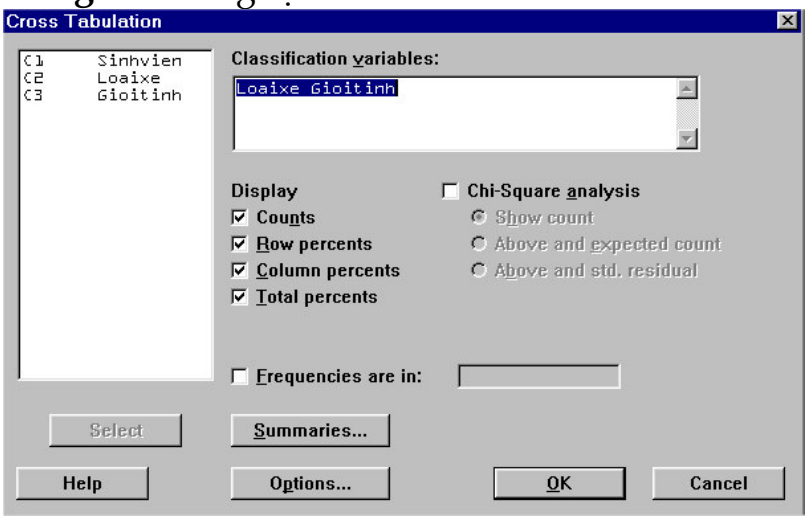
2.2 PHÂN LOẠI DỮ LIỆU ĐỊNH TÍNH

Chúng ta thường thu thập dữ liệu của cùng một biến định tính cho nhiều nhóm khác nhau. Ví dụ như mức độ áp lực của sinh viên nam và sinh viên nữ theo từng mức học vấn khác nhau như thế nào, hoặc trình độ học vấn của nhân viên nam hoặc nữ. Vì vậy, để để thể hiện được sự tương quan và phân loại nhóm của các biến, người ta thực hiện phân tích bảng chéo.

Ví dụ 5:
Cho tập dữ liệu về loại xe và giới tính sinh viên sử dụng những loại xe này, phân tích hành vi sử dụng xe của hai nhóm giới tính bằng phân tích chéo (Cross Tabulation) và dùng đồ thị (Chart).

TT	Loai xe	Gioi tinh	TT	Loai xe	Gioi tinh	TT	Loai xe	Gioi tinh
1	Honda	M	11	Sym	F	21	Sym	F
2	Honda	M	12	Honda	M	22	Yamaha	F
3	Suzuki	F	13	Sym	M	23	Honda	M
4	Yamaha	F	14	Xedap	F	24	Honda	M
5	Yamaha	F	15	Honda	F	25	Suzuki	M
6	Suzuki	M	16	Suzuki	F	26	Honda	F
7	Honda	F	17	Yamaha	M	27	Honda	F
8	Suzuki	F	18	Yamaha	M	28	Yamaha	M
9	Sym	M	19	Honda	M	29	Yamaha	M
10	Xedap	M	20	Yamaha	M	30	Honda	F

Lời giải: Dùng lệnh Stat >tables> Cross Tabulation



Cross Tabulation

Stat >tables> Cross Tabulation

Select loaixe và Gioitinh

Trong **Classification Variables:**

Nhấp chuột **Counts** và **Column pecents**. OK

Kết quả

Tabulated Statistics: Loaixe. Gioitinh

```
 Rows: Loaixe     Columns: Gioitinh

               F         M       All
Honda          5         6        11
            45,45     54,55    100,00
            35,71     37,50     36,67
            16,67     20,00     36,67

Suzuki         3         2         5
            60,00     40,00    100,00
            21,43     12,50     16,67
            10,00      6,67     16,67

Sym            2         2         4
            50,00     50,00    100,00
            14,29     12,50     13,33
             6,67      6,67     13,33

Xedap          1         1         2
            50,00     50,00    100,00
             7,14      6,25      6,67
             3,33      3,33      6,67

 Yamaha        3         5         8
            37,50     62,50    100,00
            21,43     31,25     26,67
            10,00     16,67     26,67

 All           14        16        30
            46,67     53,33    100,00
           100,00    100,00    100,00
            46,67     53,33    100,00

  Cell Contents --
                  Count
                  % of Row
                  % of Col
                  % of Tbl
```

Bảng phân loại xe sinh viên và giới tính. Hàng đầu chỉ số lượng từng xe trong mẫu của từng nhóm nhãn hiệu xe và giới tính, hàng thứ hai

chỉ tỷ lệ phần trăm loại xe trong từng nhóm loại xe, hàng thứ ba thể hiện tỷ lệ phần trăm của từng loại xe trong biến giới tính. Hàng thứ tư thể hiện tỷ lệ phần trăm của những quan sát giới tính và loại xe sử dụng. Với kết quả trên, có 5 sinh viên nữ hay 16,67% sinh viên nữ sử dụng xe Honda trong mẫu.

Chart

Dùng đồ thị cột để so sánh hành vi sử dụng xe của sinh viên.

Graph > Chart

Chọn gioitinh trong **Graph 1 X**

Data display: chọn **Bar** trong **Display**

Nhấp chuột **For each** chọn **Group**;

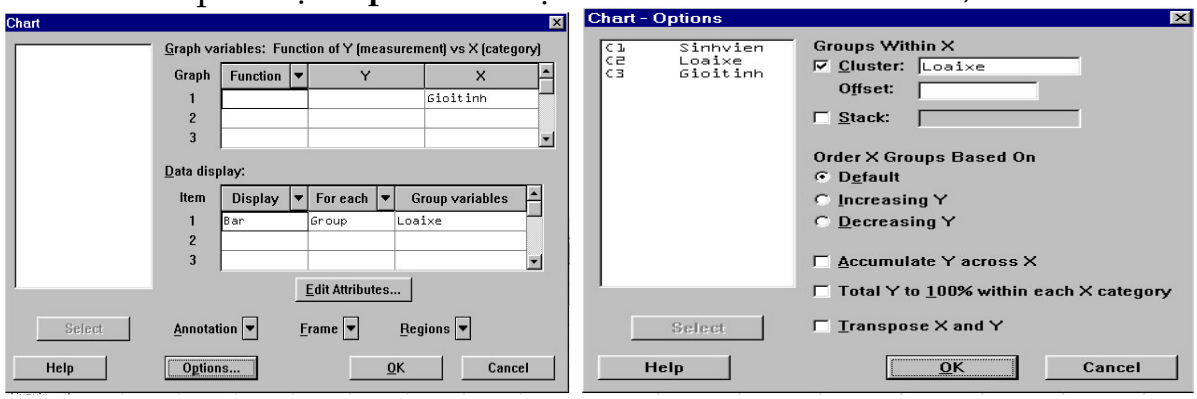
Select loaixe trong **Group variables**;

Nhấp chuột **Annotation**, chọn **Title** và đặt tên

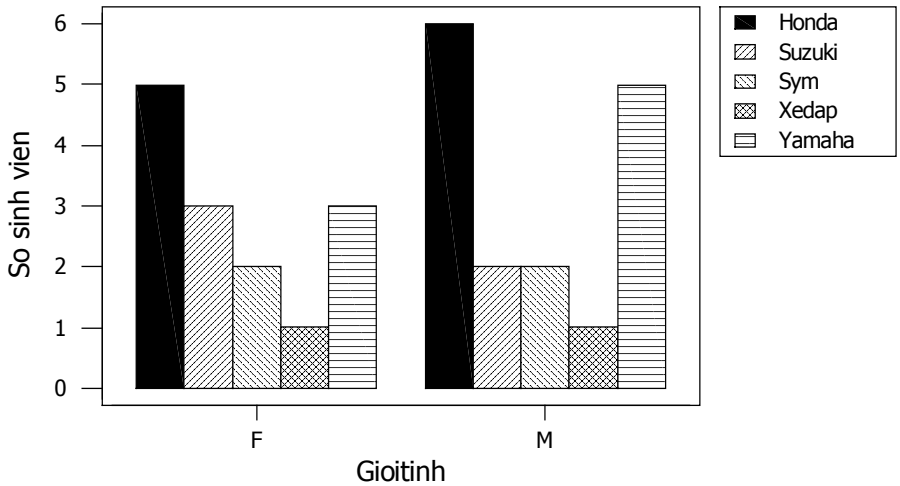
Nhấp chuột **Frame**, nhấp chuột **Axis**.

Nhập So Sinh Vien trong **Label 2.**

Nhấp chuột **Options:** chọn **Cluster: Select** loai xe, **OK**

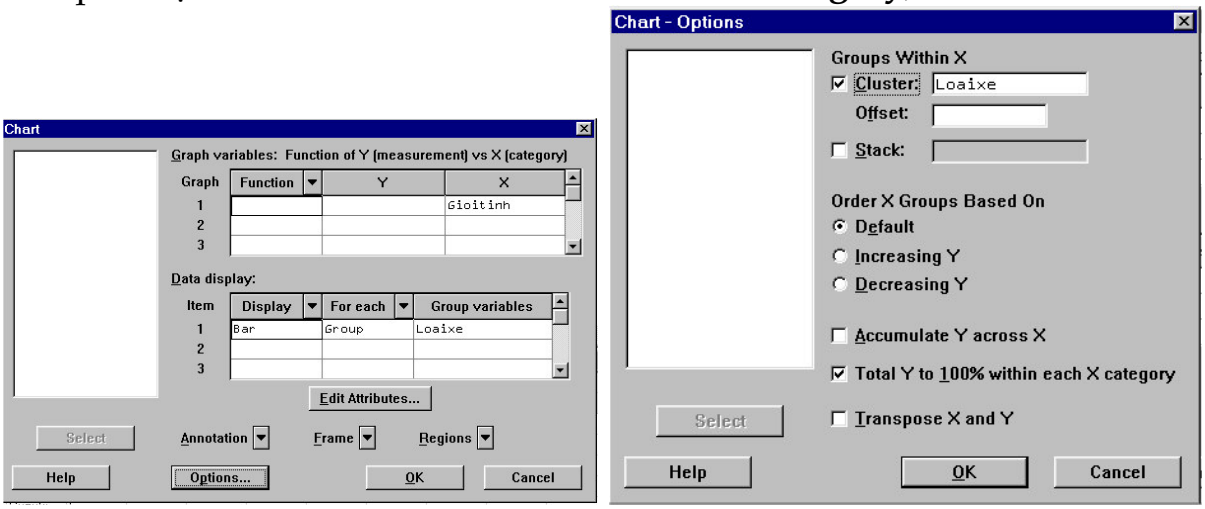


Loai xe sinh vien su dung

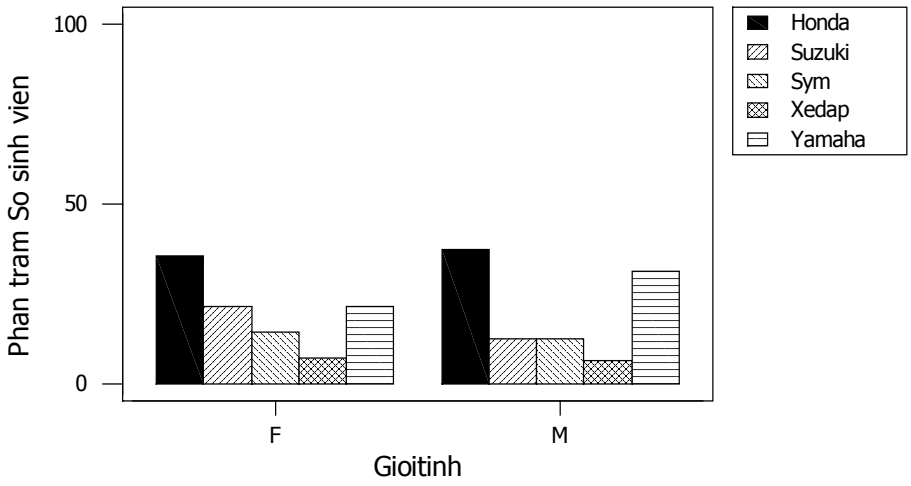


Dạng đồ thị cột thứ hai thể hiện phần trăm số sinh viên sở hữu từng loại xe tương ứng với giới tính. Sự khác biệt chính là trục tung là phần trăm sinh viên sử dụng.

Graph > Chart
Chọn gioitinh trong **Graph 1 X**
Data display: chọn **Bar** trong **Display**
Nhấp chuột **For each** chọn **Group**;
Select loaixe trong **Group variables**;
Nhấp chuột **Annotation**, chọn **Title** và đặt tên
Nhấp chuột **Frame**, nhấp chuột **Axis**.
Nhập Phan tram Sinh Vien trong **Label 2.**
Nhấp chuột **Options:**
Chọn **Cluster: Select** loai xe,
Nhấp chuột **Total Y to 100% within each X category, OK**



Loai xe sinh vien su dung theo gioi tinh



Ví dụ 6:
Một doanh nghiệp muốn tiến hành chương trình đào tạo tại chỗ cho nhân viên của mình, doanh nghiệp tiến hành khảo sát sự quan tâm của nhân viên về chương trình này. Hai câu hỏi trong phần khảo sát là giới tính và trình độ học vấn. Số liệu thu thập 200 nhân viên thể hiện trong bảng dưới đây. Tiến hành so sánh trình độ học vấn của nhân viên nam và nữ. (Dữ liệu trong file **degree.mtp**)

Mã: Giới tính: Nam = 1; Nữ = 2
Bằng cấp: Phổ thông = 1, cử nhân = 2, Thạc ĩ = 3, tiến sĩ = 4

GT	BC	GT	BC	GT	BC	GT	BC	GT	BC	GT	BC	GT	BC	GT	BC
2	2	2	1	2	2	2	1	2	4	1	3	2	1	2	1
1	4	1	3	2	3	1	2	1	2	2	1	1	2	2	2
1	4	2	1	1	2	2	2	2	2	1	2	2	4	2	2
1	2	1	3	1	3	2	1	2	2	1	1	2	1	1	2
1	4	1	4	1	2	1	3	2	2	2	1	2	2	2	2
1	2	1	4	2	2	1	4	2	1	2	2	1	1	2	2
1	2	1	1	1	3	1	2	1	2	2	4	2	3	2	3
2	4	1	2	1	3	1	3	2	1	1	2	1	2	1	3
2	1	1	4	2	1	2	1	2	1	2	2	1	2	1	1
2	4	1	4	2	1	2	1	1	2	1	4	2	1	2	1
2	2	2	1	1	4	1	2	1	2	2	3	2	1	2	2
2	2	1	3	2	1	1	2	1	4	1	2	2	2	1	4
1	3	1	2	1	2	2	1	1	2	2	2	2	2	2	3
1	1	2	2	1	4	2	1	2	3	1	3	2	1	2	3
2	1	2	1	2	2	2	1	2	3	2	1	2	2	2	2
2	2	1	3	1	1	2	3	1	3	1	3	2	2	2	1
2	4	1	2	1	3	2	3	2	1	1	3	2	1	1	2
2	2	1	2	1	4	1	2	1	1	1	2	2	2	2	2
1	3	2	3	2	2	1	4	1	4	1	2	1	3	1	1
2	3	1	4	1	3	2	1	1	2	1	2	1	2	2	2
1	2	2	1	1	2	1	3	2	2	2	1	2	2	1	2
2	2	1	4	2	1	2	1	1	2	1	2	1	3	2	1
2	2	2	2	1	1	2	2	1	2	1	4	1	4	2	1
2	2	2	4	1	2	1	4	2	2	2	1	1	4	2	2
1	2	1	4	2	1	2	2	2	2	2	2	1	4	2	3

Lời giải: Mở file dữ liệu degree.mtp, dữ liệu hiện thời dạng số, cần mã hoá dữ liệu thành biến chữ để dễ nhận dạng. Dùng lệnh Stat>Table> Cross Tabulation> trong Minitab để tìm kết quả

Bước 1: Mã hóa dữ liệu

Manip > Code > Numberic to Text

Select GT trong **Code data from Columns:**

Select GT trong **Into Columns:**

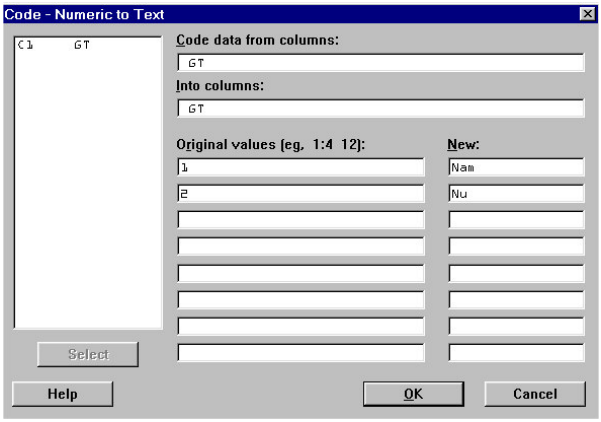
Nhập **Original values** 1, 2 và **New** Nam, Nu. **OK**

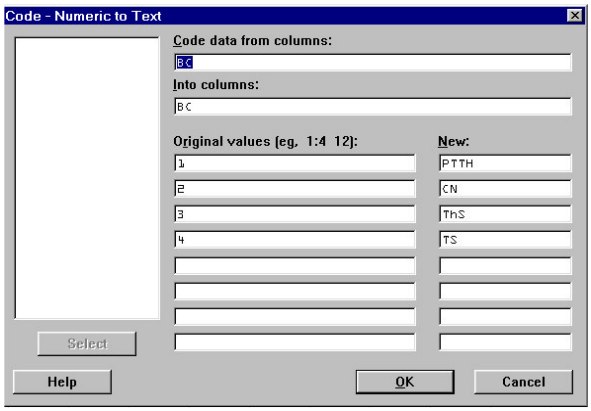
Manip > Code > Numberic to Text

Select BC trong **Code data from Columns:**

Select BC trong **Into Columns:**

Nhập **Original values** 1, 2, 3, 4 và **New** PTTH, CN, ThS, TS. **OK**





Cross Tabulation

Stat >tables> Cross Tabulation

Select GT và BC

Trong **Classification Variables:**

Nhấp chuột **Counts** và **Column pecents**. **OK**

Tabulated Statistics: GT. BC

Rows: GT Columns: BC

	1	2	3	4	All
1	9	41	22	24	96
	18,00	48,81	62,86	77,42	48,00
2	41	43	13	7	104
	82,00	51,19	37,14	22,58	52,00
All	50	84	35	31	200
	100,00	100,00	100,00	100,00	100,00

Cell Contents --
Count
% of Col

Đồ thị

Graph > Chart

Chọn BC trong **Graph 1 X**

Trong **Data Display:**

Nhấp chuột **For each** chọn **Group;**

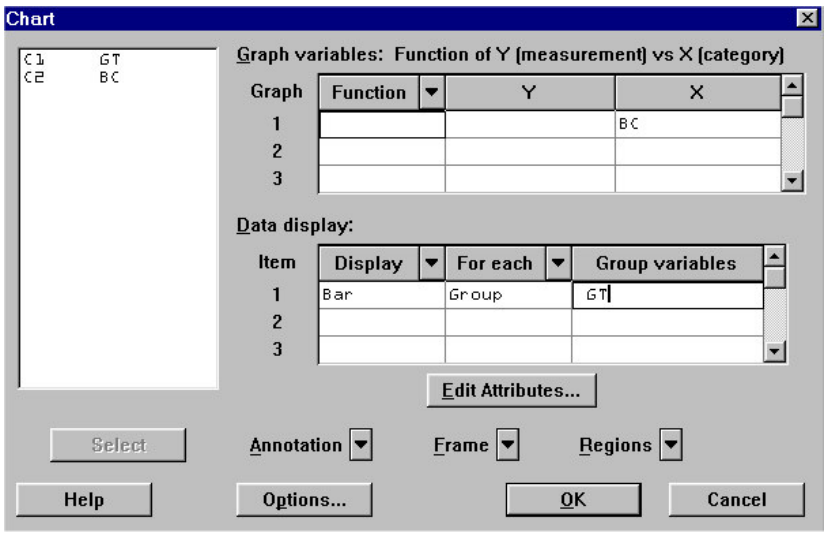
Select GT trong **Group variables;**

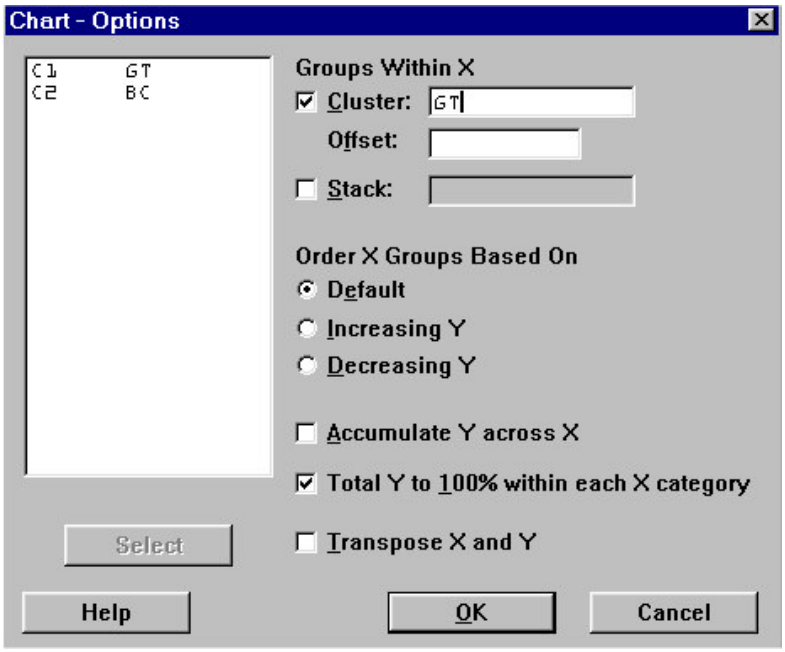
Nhấp chuột **Annotation**, chọn **Title** và đặt tên

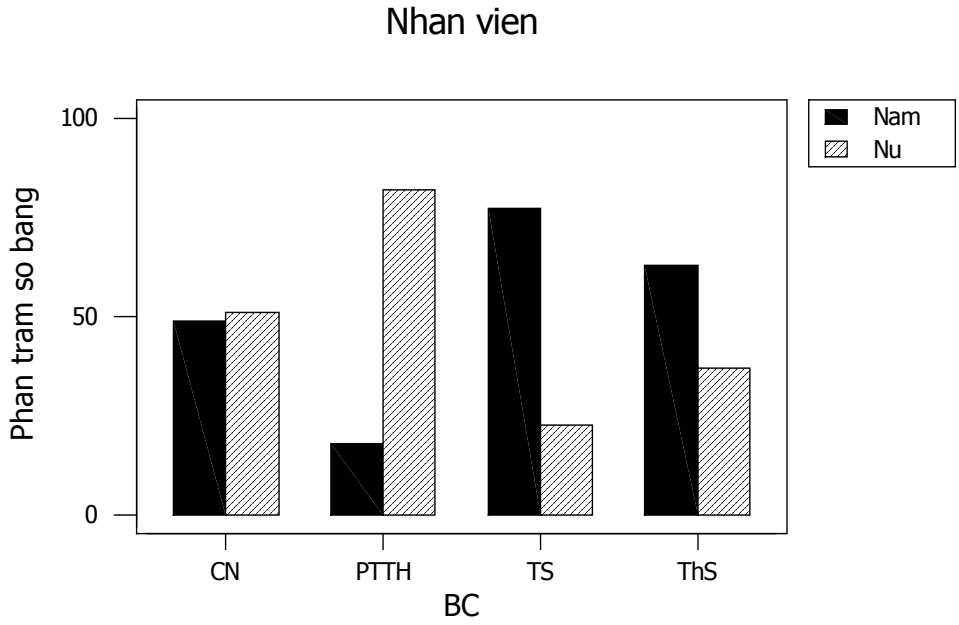
Nhấp chuột **Options**: trong **Groups within X**:

Chọn **Cluster**: **Select** GT;

Nhấp chuột **Total Y to 100% within each X category, OK.**







2.3 ĐỒ THỊ CHO DỮ LIỆU ĐỊNH LƯỢNG

Dữ liệu định lượng là các quan sát được đo theo thang đo bằng số, gồm dữ liệu khoảng (interval data) và dữ liệu tỷ lệ (ratio data). Dữ liệu khoảng là dữ liệu bằng số mà sự chênh lệch giữa hai giá trị có trong dữ liệu là có ý nghĩa, nhưng giá trị 0 không đồng nghĩa với sự vắng mặt của biến đang quan tâm. Ví dụ của dữ liệu khoảng là nhiệt độ cao hàng ngày và các điểm sát hạch được chuẩn hóa như Sát hạch Năng lực Học thuật (SAT). Ví dụ, nhiệt độ được đo theo thang đo bách

phân (^{0}C) và thang Fahrenheit (^{0}F) có điểm 0 khác nhau, và giá trị 0 không đồng nghĩa với sự vắng mặt biến nhiệt độ.

Dữ liệu tỷ lệ, là những đo lường bằng số mà trong đó tỷ lệ giữa 2 giá trị dữ liệu là có nghĩa, và giá trị 0 đồng nghĩa với sự vắng mặt của biến. Số sinh viên vắng mặt trong lớp, tiền lời hàng ngày của một ngân quỹ chung, và độ tuổi của một loại mô sinh học được cấy là các ví dụ. Số sinh viên vắng mặt là 0, tiền lời của ngân quỹ là 0, tuổi mô sinh học bằng 0 đồng nghĩa không có hiện diện biến đang quan tâm.

Các phương pháp đồ thị sẽ tóm lược bằng hình ảnh thông tin thích hợp hàm chứa trong các tập dữ liệu. Phần này mô tả biểu đồ thân – lá (stem – and – leaf displays), lược đồ tần suất (histograms), và biểu đồ điểm (dot plots).

2.3.1 Biểu Đồ Thân Và Lá

Biểu đồ thân – lá có dạng một đồ thị ký tự. Khi làm việc trong chế độ windows (Session window), ta không cần nhập lệnh GSTD trước khi dùng lệnh STEM-AND-LEAF. Dữ liệu gốc gần như nguyên vẹn trong kết quả biểu hiện của đồ thị này. Đồ thị chia mỗi quan sát thành 2 phần: một phần thân và một phần lá. Cột đầu tiên trong biểu đồ cho biết tổng tích lũy của các quan sát bắt đầu từ thân cao nhất xuống thân chứa trung vị, và tổng tích lũy của các quan sát bắt đầu từ thân thấp nhất lên thân chứa trung vị. Số đếm của thân chứa trung vị đặt trong ngoặc đơn. Cột thứ hai chứa phần thân và cột thứ ba chứa phần lá. Minitab dùng 1, 2, và 5 dòng cho mỗi thân tùy thuộc vào số quan sát và vùng rộng của dữ liệu.

Để vẽ biểu đồ thân và lá chúng ta dùng một trong các menu sau:
Graph > Stem-and-Leaf
Graph > Character Graphs > Stem-and-Leaf
Stat > EDA > Stem-and-Leaf

Ví dụ 7: Biểu Đồ Thân-và-Lá
Trong học kỳ mùa xuân năm 2000, một nhóm sinh viên môn thống kê đã lấy mẫu ngẫu nhiên 50 sinh viên đại học để nghiên cứu điểm trung bình tích lũy (GPAs) của họ. Bảng sau cho thấy điểm GPA của 50 sinh

viên đại học trong nghiên cứu này. Hãy vẽ và diễn giải biểu đồ thân lá cho tập dữ liệu.

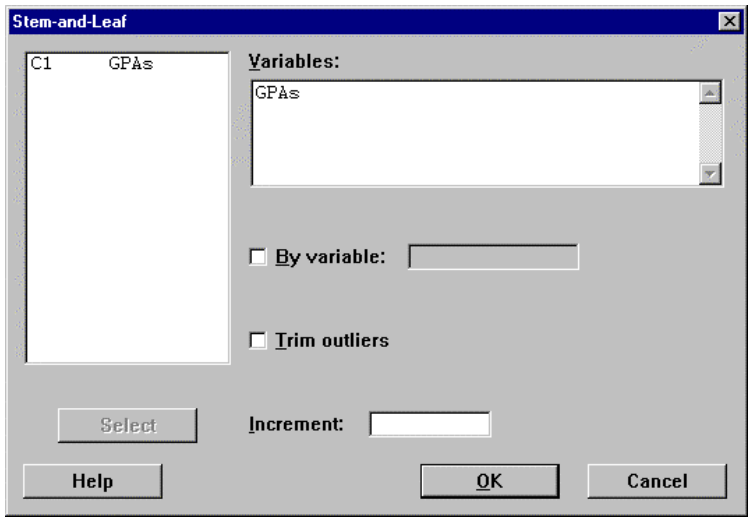
Điểm Trung Bình Tích Lũy (GPA)

2,2	3,4	4,0	3,1	3,3	3,8	3,0	2,8	2,9	2,9
2,8	2,6	3,3	3,2	3,0	2,8	3,5	2,2	2,9	3,6
3,2	3,0	2,5	2,5	2,8	3,4	2,2	2,9	3,7	2,7
2,7	2,7	2,6	2,9	3,0	2,8	2,9	3,0	2,7	2,9
2,9	2,6	3,5	2,5	2,5	2,5	3,5	3,2	2,9	3,4

Lời Giải Điểm trung bình tích lũy được lưu trong file có tên **GPA.mtp**. Dữ liệu trong cột C1 có tên GPAs. Biểu đồ thân lá thể hiện dưới dạng đồ thị ký tự.

Vẽ biểu đồ thân và lá:

Chọn Menu **Graph > Stem-and-Leaf**
Nhấn **Select** chọn biến GPAs vào ô **Variable.**
OK



Stem-and-Leaf Display: GPAs

```
Stem-and-leaf of GPAs      N  = 50
Leaf Unit = 0.10

    3    2 222
    8    2 55555
   15    2 6667777
  (14)   2 88888999999999
   21    3 000001
   15    3 22233
   10    3 444555
    4    3 67
    2    3 8
    1    4 0
```

N = 50 cho ta biết có 50 giá trị trong biểu đồ thân lá. Đơn vị của phần lá (leaf unit) trong biểu đồ là 0,10, nghĩa là đơn vị của phần thân (stem unit) là 1. Giá trị nhỏ nhất, với phần thân là 2 và phần lá là 2, 2,2 GPA, có 3 sinh viên có điểm 2,2. Giá trị lớn nhất là 4,0, và điểm trung vị của phân phối là 2,9.

Cột các số bên tay trái của biểu đồ cho biết tổng tích lũy, bắt đầu từ hai phía: giá trị nhỏ nhất phía trên cùng, và giá trị lớn nhất phía dưới cùng đến phần thân chứa điểm trung vị. Đếm từ giá trị nhỏ nhất, có 3 quan sát trên thân thứ nhất, 8 trên hai thân đầu tiên, 15 trên ba thân đầu tiên, và v.v… Nếu đếm từ giá trị lớn nhất, có 1 quan sát trên thân cuối cùng, 2 trên hai thân cuối cùng, 4 trên ba thân cuối cùng và v.v… Tổng các số trong lớp chứa trung vị là số 14, được để trong dấu ngoặc.

2.3.2 Lược đồ tần suất

Lược đồ tần suất thể hiện dưới dạng đồ thị ký tự và cả dạng đồ họa (professional graph). Đó là đồ thị tần suất, tần suất tương đối hoặc phân phối tần suất phần trăm. Phân phối tần suất của dữ liệu định lượng gồm có theo lớp (theo khoảng), và theo tần suất (hay tổng số đếm) của các giá trị rơi vào mỗi lớp. Tần suất tương đối là tỷ lệ của các giá trị dữ liệu thuộc mỗi lớp so với tổng quan sát; tần suất phần trăm là tần suất tương đối nhưng tính theo phần trăm. Minitab xác định số lượng lớp bởi số lượng giá trị dữ liệu và vùng rộng của dữ liệu. Bạn có thể thay đổi kết quả thể hiện với những lệnh phụ trợ.

Muốn vẽ lược đồ tần suất dạng ký tự dùng Menu
Graph > Character Graphs > Histogram

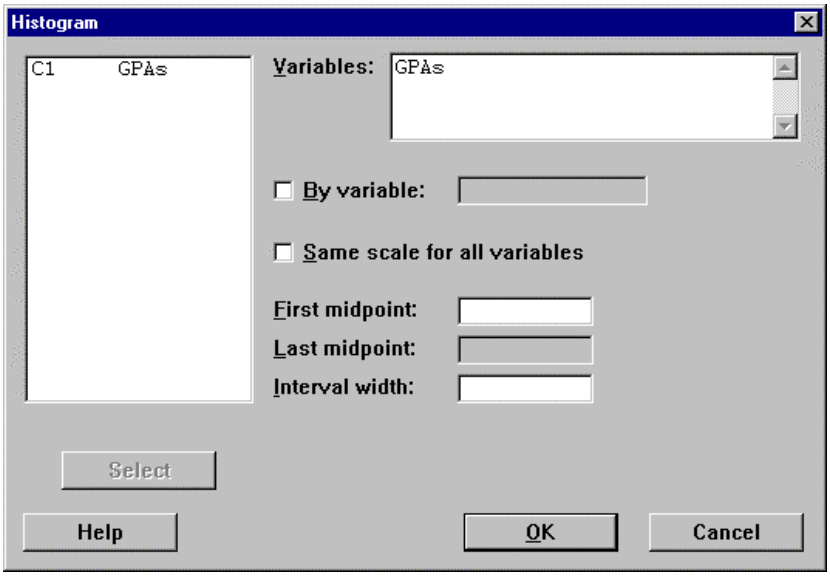
Muốn vẽ lược đồ dạng đồ họa dùng Menu
Graph > Histogram

Ví dụ 8: Lược Đồ Tần Suất
Số liệu giống trong ví dụ tập dữ liệu điểm GPA của 50 sinh viên. Hãy xây dựng và diễn giải lược đồ tần suất.

Vẽ lược đồ tần suất:

Vẽ lược đồ tần suất dạng ký tự

Chọn Menu **Graph > Character Graphs > Histogram**
Nhấn **Select** để chọn biến GPAs vào hộp **Variables**. Nhấn **OK**



Histogram

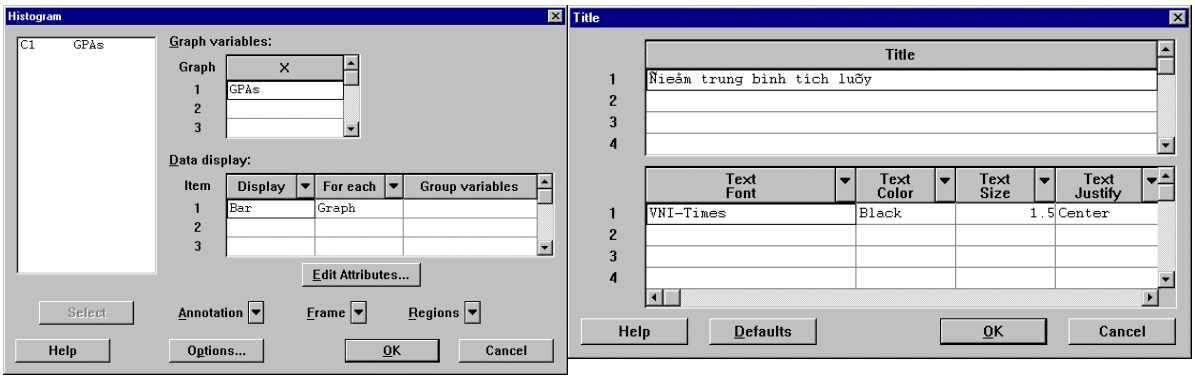
```
Histogram of GPAs   N = 50

Midpoint         Count
     2.2             3  ***
     2.4             0
     2.6             8  ********
     2.8             9  *********
     3.0            14  **************
     3.2             4  ****
     3.4             5  *****
     3.6             4  ****
     3.8             2  **
     4.0             1  *
```

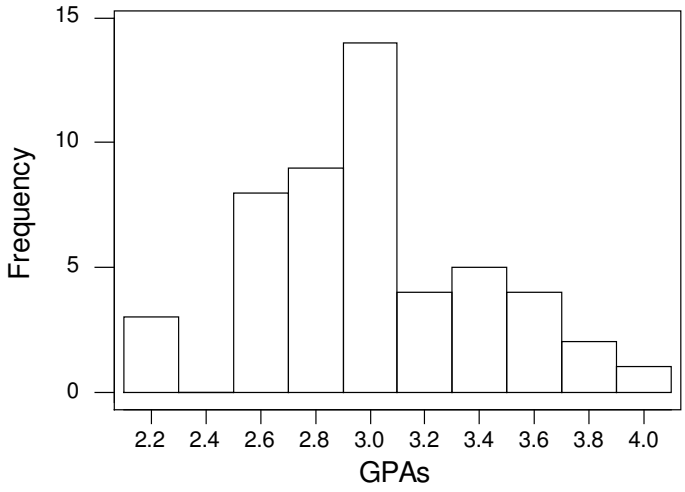
Lệnh HISTOGRAM (cho đồ thị dạng ký tự) cho thấy phân phối của dữ liệu và lược đồ pseudo (pseudo-histogram) với các dấu hoa thị (*) bên phải của các số đếm. Trung điểm có giá trị gần bằng 3, phân phối hơi bị lệch về bên phải.

Vẽ lược đồ tần suất dạng đồ họa

Chọn Menu **Graph > Histogram**
Nhấn **Select** chọn GPAs vào bảng **Graph 1 X;**
Nhấn **Annotation**, chọn **Title**; nhập tên đồ thị "Điểm trung bình tích lũy". Chọn **Text Font** là Vni-times (hoặc bất kỳ font khác mà bạn muốn). Nhấn **OK**.



Điểm trung bình tích lũy



Lược đồ sẽ nhóm điểm GPA theo các lớp có độ rộng là 0,2. Lớp thứ nhất có tâm điểm là 2,2 và có giá trị từ 2,1 đến nhỏ hơn 2,3. Lớp thứ 2 có tâm điểm 2,4 và có giá trị từ 2,3 đến nhỏ hơn 2,5, v.v… Lớp thứ 2 bị trống là do không có sinh viên nào có điểm GPA trong khoảng 2,3 đến nhỏ hơn 2,5. Phân phối dường như tập trung gần khoảng giá trị 3,0 và bị lệch về phía có điểm GPA cao.

2.3.3 Biểu đồ điểm

Biểu đồ điểm thể hiện dưới dạng đồ thị ký tự hay dạng đồ họa bằng cách thực thi một macro. Nếu bạn đang làm việc trong phần Session Window, bạn không cần nhập lệnh GSTD trước khi dùng lệnh DOTPLOT. Đồ thị dùng một trục ngang và các nhóm dữ liệu ít nhất có thể. Các quan sát được trình bày dưới dạng các điểm trên trục nằm ngang.

Để vẽ biểu đồ điểm chúng ta dùng một trong các menu sau:
Graph > Dotplot
Graph > Character Graphs > Dotplot

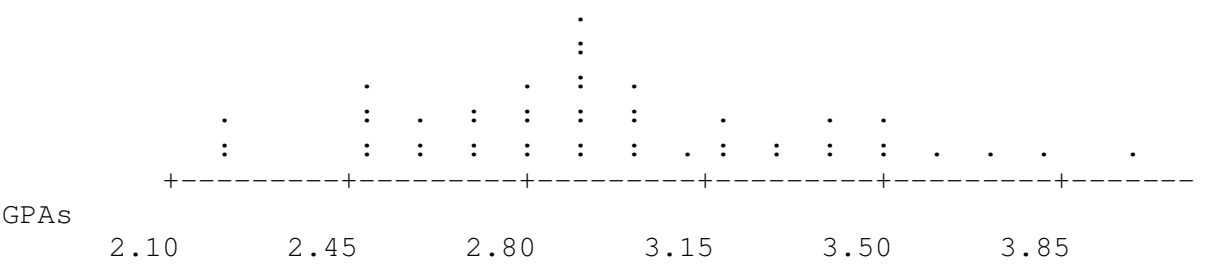
Ví dụ 9: Biểu Đồ Điểm
Dùng dữ liệu điểm GPA của 50 sinh viên trong ví dụ trước, hãy xây dựng và diễn giải biểu đồ điểm.

Vẽ biểu đồ điểm:

Biểu đồ điểm dạng ký tự

Chọn Menu **Graph > Character Graphs > Dotplot**
Nhấn **Select** chọn GPAs vào hộp **Variables**. Nhấn **OK**

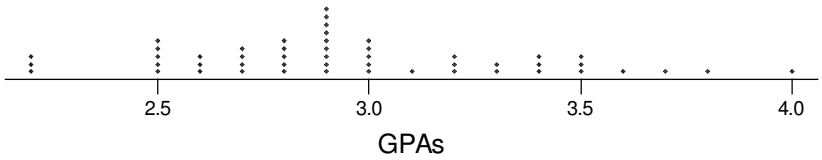
Dotplot: GPAs



Biểu đồ điểm dạng đồ họa

Chọn Menu **Graph > Dotplot**
Nhấn **Select** chọn GPAs vào hộp **Variables**; Nhập tựa đề 'Điểm trung bình tích lũy' vào ô Title. Nhấn **OK**

Điểm trung bình tích lũy



Biểu đồ điểm dạng đồ họa cho thông tin đồ thị về dạng của dữ liệu giống như biểu đồ điểm dạng ký tự. Biểu đồ điểm cho thấy phân phối tập trung gần giá trị 3,0 và bị lệch về phía điểm trung bình tích lũy cao.

2.4 CÁC ĐẠI LƯỢNG THỐNG KÊ CƠ BẢN (MÔ TẢ DỮ LIỆU BẰNG SỐ)

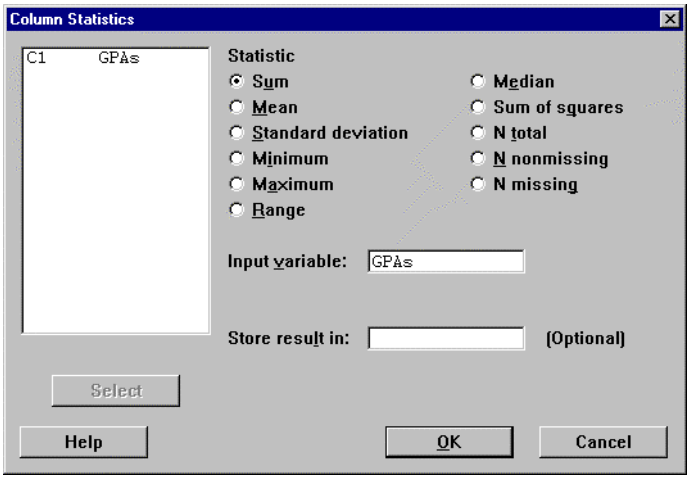
Các đại lượng mô tả bằng số sẽ đặc tả hoặc cho biết thông tin về một tập dữ liệu. Điểm trung bình (mean) và trung vị (median) là các đại lượng đo lường sự tập trung của dữ liệu. Các đại lượng đo lường sự phân tán bao gồm độ lệch chuẩn (standard deviation) và khoảng (range), và các điểm định vị là điểm z (z-score) và điểm phần tư (quartiles). Các đại lượng riêng lẻ có thể được tính cho dữ liệu lưu theo cột và hoặc lưu thep hàng trong bảng Minitab.

2.4.1 Các lệnh dùng cho dữ liệu lưu trữ trong các cột

Mỗi lệnh (option trong hộp thoại) sẽ tính toán và cho ra một con số. Bạn có thể lưu kết quả dưới dạng một hằng số lưu vào bộ nhớ trong Minitab dưới tên một ký tự nào đó.

Để tính toán dữ liệu lưu trong cột, dùng menu
Calc > Column Statistics

Nhấn **Select** để chọn biến muốn tính các giá trị thống kê vào ô **Input variable**.



Các tùy chọn thống kê:

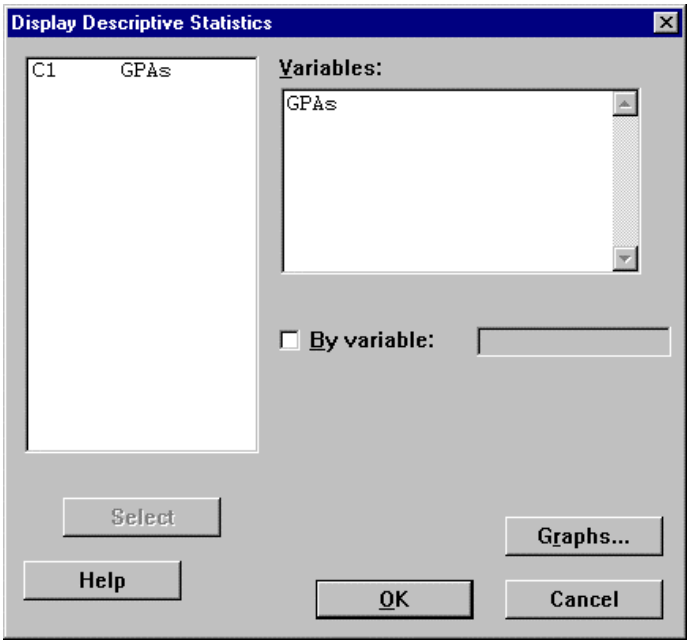
Sum	Tổng các giá trị trong cột (của mẫu)
Mean	Trung bình của các giá trị trong cột
Standard deviation	Độ lệch chuẩn của các giá trị trong cột
Minimum	Giá trị nhỏ nhất của các giá trị trong cột
Maximum	Giá trị lớn nhất của các giá trị trong cột
Range	Khoảng của mẫu, bằng giá trị lớn nhất trừ giá trị nhỏ nhất
Median	Giá trị trung tâm (số trung vị) của các giá trị trong cột
Sum of square	Tổng bình phương các giá trị trong cột
N total	Tổng số quan sát
N nonmissing	Số quan sát có giá trị không bị khuyết
N missing	Số quan sát có giá trị bị khuyết

Nếu bạn nhập vào **Store result in** một giá trị đại số (ví dụ k1, k2, a1,..) thì Minitab sẽ lưu giá trị vừa tính toán vào vào phần **Constant** trong cửa sổ **Project Manager**, sau đó bạn có thể dùng giá trị đại số này để tính toán khi cần thiết về sau.

2.4.2 Các đại lượng thống kê cơ bản

Phần này sẽ tạo một bảng tóm lược nhiều đại lượng đo lường thống kê Chọn Menu **Stat > Basic Statistics > Display Descriptive Statistics**

Nhấn **Select** để chọn biến cần mô tả thống kê, bạn có thể mô tả thống kê biến này theo nhóm được phân loại theo biến điều khiển (chọn trong **By variable**), hoặc bạn muốn tạo thêm các đồ thị thống kê thì chọn thêm tùy chọn **Graph…**



Sau đó bạn nhấn **OK** sẽ tạo một bảng tóm lược các đo lường thống kê gồm:

N	Số quan sát không có giá trị khuyết
N*	Số quan sát có giá trị khuyết
MEAN	Giá trị trung bình của các giá trị trong cột
MEDIAN	Số trung vị
TRMEAN	Giá trị trung bình gọt tỉa (trimmed mean) là giá trị trung bình của dữ liệu sau khi loại bỏ 5% số quan sát có giá trị nhỏ nhất và 5% số quan sát có giá trị lớn nhất
STDEV	Độ lệch chuẩn của mẫu dữ liệu
SEMEAN	Sai số chuẩn của trung bình (standard error of the mean) có giá trị bằng STDEV/√n
MIN	Giá trị nhỏ nhất của các giá trị trong cột
MAX	Giá trị lớn nhất của các giá trị trong cột
Q1	25% số quan sát có giá trị nhỏ hơn Q1
Q3	75% số quan sát có giá trị nhỏ hơn Q3

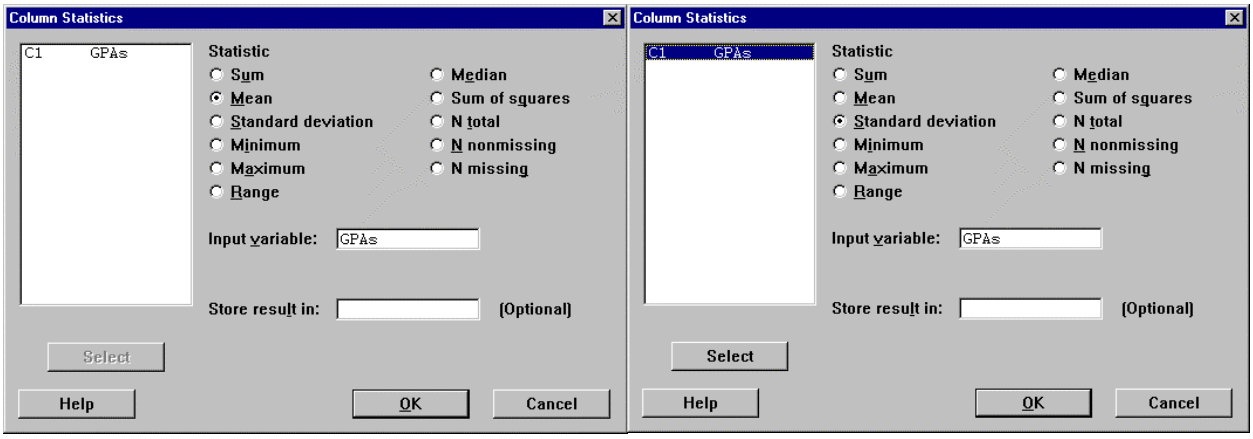
Ví dụ 10: Thống kê mô tả bằng số

Với nghiên cứu về điểm trung bình tích lũy (GPAs) mô tả trong ví dụ trước. Hãy tính trung bình và độ lệch chuẩn. Tính toán và giải thích các thống kê mô tả.

Lời giải: Dữ liệu được lấy trong file GPA ở ví dụ 7 (GPA.mtp). Giá trị trung bình và độ lệch chuẩn được tính riêng biệt theo 2 lệnh sau:

Menu **Calc > Column Statistics**
Select GPAs vào ô **Input variable,**
Chọn **Mean**, nhấn **OK**

Menu **Calc > Column Statistics**
Select GPAs vào ô **Input variable,**
Chọn **Standard deviation**, nhấn **OK**



Bạn sẽ có kết quả sau:

Mean of GPAs

```
Mean of GPAs = 2.9580
```

Standard Deviation of GPAs

```
Standard deviation of GPAs = 0.41012
```

Giá trị trung bình hay trung bình số học của điểm trung bình tích lũy của 50 sinh viên là khoảng 2,96 và độ lệch chuẩn là 0,410.

Tạo bảng các đo lường thống kê:
Menu **Stat ▸ Basic Statistics ▸ Display Descriptive Statistics**
Select GPAs vào ô **Variables. OK** sẽ có kết quả sau

Descriptive Statistics: GPAs

Variable	N	Mean	Median	TrMean	StDev	SE Mean
GPAs	50	2.9580	2.9000	2.9500	0.4101	0.0580

Variable	Minimum	Maximum	Q1	Q3
GPAs	2.2000	4.0000	2.7000	3.2250

N	Có 50 quan sát trong bộ dữ liệu
Mean	Giá trị trung bình của điểm trung bình tích lũy (GPA) gần bằng 2,96
Median	Số trung vị của GPA là 2,9. Giá trị trung bình lớn hơn số trung vị chỉ ra rằng phân phối bị nghiêng về bên phải
TrMean	Đây là trung bình gọt tỉa 90%. Đó là trị trung bình của các giá trị dữ liệu sau khi loại bỏ 5% số quan sát có giá trị nhỏ nhất (lấy số nguyên gần đúng) và 5% số quan sát có giá trị lớn nhất. Vì 5% của 50 là 2,5 gần bằng 3, 3 giá trị GPA nhỏ nhất và 3 giá trị GPA lớn nhất bị loại, và giá trị trung bình của 44 điểm trung bình tích lũy còn lại sẽ được tính toán. Trung bình gọt tỉa là 2,95 hơi nhỏ hơn giá trị trung bình chưa gọt tỉa.
StDev	Độ lệch chuẩn của mẫu là 0,41
SE Mean	Sai số chuẩn của trung bình bằng độ lệch chuẩn (0,410) chia cho căn bậc hai của N (50). Chúng ta dùng đo lường này để rút ra các suy diễn thống kê.
Min	Giá trị GPA nhỏ nhất là 2,2
Max	Giá trị GPA lớn nhất là 4,0
Q1	Khoảng 25% số sinh viên có điểm GPA nhỏ hơn 2,7
Q3	Khoảng 75% số sinh viên có điểm GPA nhỏ hơn 3,2

2.5 DIỄN DỊCH ĐỘ LỆCH CHUẨN

Độ lệch chuẩn là giá trị đo lường độ lệch trung bình của các giá trị dữ liệu so với giá trị trung bình. Có hai quy tắc diễn dịch độ lệch chuẩn: Quy tắc thực nghiệm và Quy tắc Tchebysheff. Cả hai quy tắc đều xem xét tỷ lệ các giá trị dữ liệu rơi vào 1 khoảng nào đó của giá trị trung bình, khoảng này đo bằng độ lệch chuẩn.

Quy tắc Thực nghiệm áp dụng cho dữ liệu có phân phối đối xứng hay có dạng hình chuông:
1. Khoảng 68% các giá trị đo lường rơi vào khoảng ($\bar{x}$ -s, $\bar{x}$ +s)
2. Khoảng 95% các giá trị đo lường rơi vào khoảng ($\bar{x}$ -2s, $\bar{x}$ +3s)
3. Gần 100% các giá trị đo lường rơi vào khoảng ($\bar{x}$ -3s, $\bar{x}$ +3s)
Ghi chú: $\bar{x}$ là giá trị trung bình, s là độ lệch chuẩn của mẫu

Quy tắc Chebyshev áp dụng cho bất kỳ bộ dữ liệu nào bất chấp hình dạng phân phối. Nói chung, quy tắc phát biểu rằng ít nhất ($1-1/k^2$) các giá trị đo lường rơi vào khoảng ($\bar{x} \pm$ ks). Ví dụ, ít nhất 8/9 hay 89% các giá trị đo lường rơi vào khoảng ($\bar{x} \pm$ 3s).

Ví dụ 11: Quy tắc thực nghiệm
Dựa trên nghiên cứu điểm trung bình tích lũy (GPAs) mô tả trong ví dụ trước. Tìm phần trăm các quan sát rơi vào khoảng $\bar{x} \pm$ s, $\bar{x} \pm$ 2s, và $\bar{x} \pm$ 3s. So sánh những giá trị phần trăm này với Quy tắc Thực nghiệm

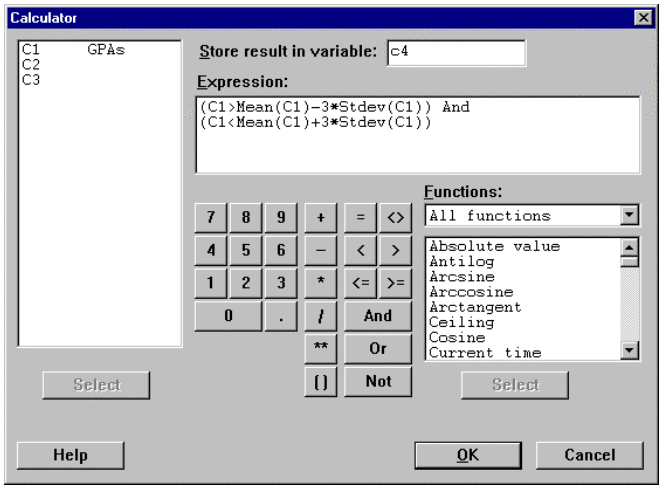
Lời giải: Điểm trung bình tích lũy được lưu trong file GPA.mtp. Và lưu trong cột C1

Ta sẽ dùng menu **Calc > Calculator.** Để tính giá trị trung bình $\bar{x}$ ta dùng hàm Mean, giá trị độ lệch chuẩn s sẽ được tính thông qua hàm Stdev. Sau đó chúng ta dùng so sánh nhỏ hơn (<) và lớn hơn (>), và tính năng logic AND để lọc các giá trị. Nếu một quan sát rơi vào khoảng ($\bar{x} \pm$ s), kết quả sẽ cho là 1, nếu không kết quả sẽ là 0. Lưu kết quả và cột C2.

Như vậy cột C2 sẽ cho biết các quan sát tương ứng trong cột C1 có nằm trong khoảng $\bar{x} \pm$ s hay không? Nếu có là giá trị 1, nếu không sẽ là giá trị 0. Tương tự cột C3 và C4 sẽ cho biết các quan sát tương ứng trong cột C1 nằm trong khoảng $\bar{x} \pm$ 2s và $\bar{x} \pm$ 3s.

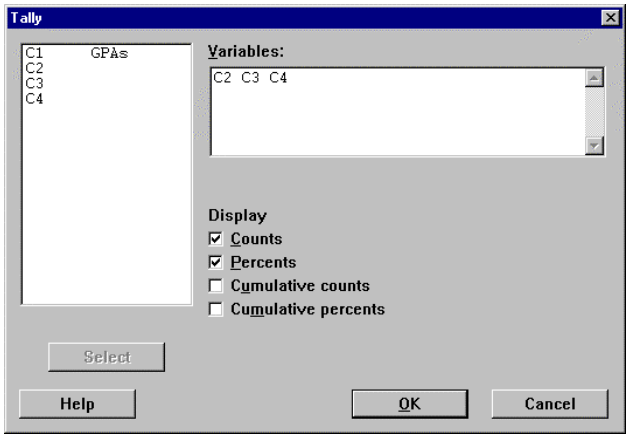
Sau đó chúng ta dùng chức năng TALLY để đếm các quan sát và tính phần trăm

Calc > Calculator
Nhập C2 vào **Store results in variable:**
Nhập (C1>Mean(C1)-Stdev(C1)) And (C1<Mean(C1)+Stdev(C1)) vào **Expression. OK**
Calc > Calculator
Nhập C3 vào **Store results in variable:**
Nhập (C1>Mean(C1)-2*Stdev(C1)) And (C1<Mean(C1)+2*Stdev(C1)) vào **Expression. OK**
Calc > Calculator
Nhập C4 vào **Store results in variable:**
Nhập (C1>Mean(C1)-3*Stdev(C1)) And (C1<Mean(C1)+3*Stdev(C1)) vào **Expression. OK**



Sau đó để tóm lược tần suất và phần trăm trong mỗi khoảng, chúng ta dùng chức năng TALLY.
Vào Menu **Stat > Tables > Tally**

Chọn 3 biến C2 C3 C4 vào ô **Variables:**
Chọn **Counts** và **Percents**. **OK**



Ta sẽ có kết quả sau

Tally for Discrete Variables: C2, C3, C4

C2	Count	Percent	C3	Count	Percent	C4	Count	Percent
0	18	36.00	0	2	4.00	1	50	100.00
1	32	64.00	1	48	96.00	N=	50	
N=	50		N=	50				

Có 32 giá trị GPA hay 64% quan sát trong khoảng $\bar{x} \pm$ s; 48 giá trị GPA hay 96% trong khoảng $\bar{x} \pm 2$s; và 50 hay 100% trong khoảng $\bar{x} \pm 3$s. Các giá trị phần trăm này gần với kết quả cho bởi Quy tắc Thực nghiệm. Các giá trị này gần bằng kết quả thực nghiệm là do phân phối này không đối xứng như chúng ta đã khảo sát trong các ví dụ trước.

2.6 CÁC ĐO LƯỜNG ĐỊNH VỊ TƯƠNG ĐỐI

Các đo lường mô tả mối quan hệ của một giá trị đo lường với toàn bộ giá trị còn lại của dữ liệu gọi là điểm định vị tương đối. Các ví dụ của đo lường này là **điểm chuẩn** (standard score) hoặc **điểm z** (z-score) và các điểm xếp hạng phần trăm (percentile rankings). Đo lường thường dùng nhất là z-score. Nó có thể được dùng để nhận diện các phần tử bất thường (outliers), là những quan sát quá lớn hay quá nhỏ so với những quan sát khác trong một tập dữ liệu. Những điểm bất thường có thể là những giá trị đo lường không đúng hoặc có thể đến từ một tổng thể khác với tập dữ liệu.

2.6.1 Điểm chuẩn hay z-scores

Điểm z chuẩn đại diện cho khoảng cách, tính theo độ lệch chuẩn, là một giá trị đo lường thể hiện sự khác biệt tương đối của các phần tử so với giá trị trung bình của tập dữ liệu. Về công thức, điểm z của giá trị đo lường x là chênh lệch giữa x và giá trị trung bình chia cho độ lệch chuẩn:

$$z = (x - \bar{x})/s.$$

Trong đó $\bar{x}$ là giá trị trung bình và s là độ lệch chuẩn của tập dữ liệu. Các điểm z sẽ có giá trị trung bình là 0 và độ lệch chuẩn là 1.

Để tính điểm z chọn Menu:

Calc > Standardize

2.6.2 Điểm định vị phần trăm

Đối với các đại lượng phần trăm, Minitab cung cấp số trung vị và các số định vị phần tư, tất cả gọi là số tứ phân. Số trung vị là phần trăm thứ 50 hay số tứ phân chính giữa, số tứ phân bên dưới Q1 là số phần trăm thứ 25 (là giá trị vượt qua 25% số lượng giá trị của dữ liệu), và số tứ phân bên trên Q3 là số phần trăm thứ 75. Vùng tứ phân (interquartile range – IQR) là khoảng cách giữa số tứ phân bên trên và bên dưới (= Q3-Q1).

2.6.3 Biểu đồ hộp

Biểu đồ hộp thể hiện dưới dạng đồ thị ký tự lẫn đồ họa. Đồ thị cho thấy những tính chất chính của một tập dữ liệu. Một biểu đồ hộp gồm: thân hộp, đường ria (wishkers) và các điểm bất thường.

Phần hộp đại diện cho một nửa chính giữa của mỗi tập dữ liệu. Các đầu mút của hộp gần bằng với các số phần tư Q1 và Q3, và số trung vị được đánh dấu bởi dấu '+' hoặc một đường thẳng liền nét. Những biểu tượng đặc biệt bên các cạnh của hộp để chỉ phạm vi của dữ liệu và vị trí của các giá trị bất thường.

Vùng tứ phân IQR là khoảng cách giữa số phần tư bên trên và bên dưới (= Q3-Q1). Những phần tử bất thường được phát hiện dựa vào các giới hạn sau:

Giới hạn trong (inner fences) được xác định bởi hai điểm:

Điểm bên dưới = Q1 – 1.5(Q3-Q1)

Điểm bên trên = Q3 + 1.5(Q3-Q1)

Giới hạn ngoài (outer fences) được xác định bởi hai điểm:

Điểm bên dưới = Q1 – 3(Q3-Q1)

Điểm bên trên = Q3 + 3(Q3-Q1)

Đường thẳng hay đường ria nối hai bên cạnh của thân hộp đến 2 giá trị đầu mút của biểu đồ sẽ nằm trong phần giới hạn trong. Giá trị nằm giữa giới hạn trong và giới hạn ngoài được đánh dấu bởi '*' là điểm có thể bất thường (possible outlier). Giá trị nằm ngoài giới hạn ngoài được đánh dấu '0' và là điểm chắc hẳn bất thường (probable outlier).

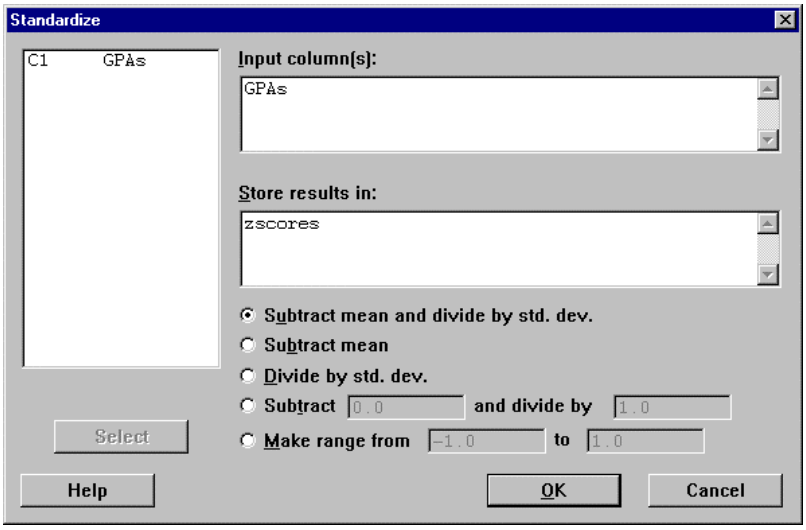
Đối với những phân phối đối xứng, số trung vị nằm tại trung tâm của hộp và các khoảng cách từ cạnh của thân hộp đến giá trị nhỏ nhất và giá trị lớn nhất xấp xỉ bằng nhau. Phân phối không đối xứng có xu hướng có đường ria dài hơn và khoảng cách từ cạnh của thân hộp đến trung vị sẽ lớn hơn về phía bị lệch của phân phối.

Để vẽ biểu đồ hộp, chọn Menu
Graph > Boxplot
Graph > Character Graphs > Boxplot
Graph > EDA > Boxplot

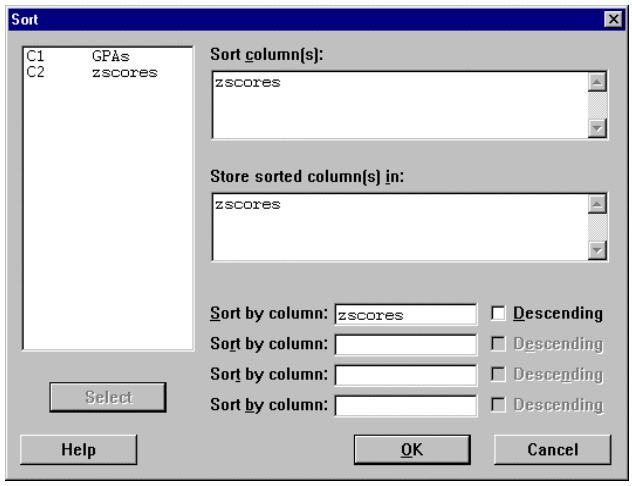
Ví dụ 12: Tính điểm z
Dựa vào nghiên cứu điểm trung bình tích lũy (GPAs) mô tả trong các ví dụ trước, hãy nhận xét xem có điểm bất thường nào hay không trong tập dữ liệu này? Hãy tính và diễn giải điểm z. Hãy xây dựng và diễn giải một biểu đồ hộp.

Lời giải: Ta lấy file GPA.mtp. Dữ liệu trong một cột có tên GPAs. Ta tính điểm z và xếp theo thứ tự để diễn dịch.

Chọn Menu **Calc > Standardize**
Select GPAs vào ô **Input column(s):**
Nhập "zscores' vào ô **Store results in. OK**



Để sắp xếp các điểm z đã tính toán theo thứ tự tăng dần (ascending), vào Menu **Manip > Sort**
Select zscores vào ô **Sort column(s):**
Select zscores vào ô **Store sorted column(s) in:**
Select zscores vào ô **Sort by column. OK**



Sau đó để trình bày kết quả dữ liệu, vào Menu **Manip > Display Data**
Select zscores vào ô **Display**. **OK**

Bạn sẽ có kết quả sau

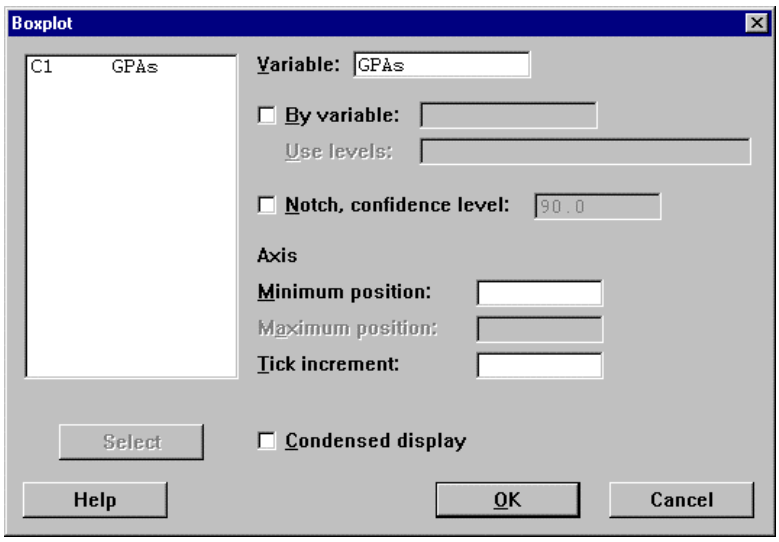
Data Display

```
zscores
  -1.84823  -1.84823  -1.84823  -1.11674  -1.11674  -1.11674  -1.11674
  -1.11674  -0.87291  -0.87291  -0.87291  -0.62908  -0.62908  -0.62908
  -0.62908  -0.38525  -0.38525  -0.38525  -0.38525  -0.38525  -0.14142
  -0.14142  -0.14142  -0.14142  -0.14142  -0.14142  -0.14142  -0.14142
  -0.14142   0.10241   0.10241   0.10241   0.10241   0.10241   0.34624
   0.59007   0.59007   0.59007   0.83390   0.83390   1.07773   1.07773
   1.07773   1.32156   1.32156   1.32156   1.56539   1.80922   2.05305
   2.54071
```

Điểm trung bình nhỏ nhất và lớn nhất có điểm z lần lượt là –1,85 và 2,54. Không có giá trị nào là điểm bất thường. Điểm bất thường là những quan sát có điểm z nhỏ hơn –3 hoặc lớn hơn 3. (Điều này rút ra từ phân phối chuẩn chuẩn hóa có trung bình hay kỳ vọng bằng 0 và độ lệch chuẩn bằng 1)

Chúng ta dùng biểu đồ hộp để vẽ đồ thị của dữ liệu và phát hiện các điểm bất thường. Chúng ta minh họa biểu đồ hộp dạng ký tự và dạng đồ họa.

Vẽ biểu đồ hộp dạng ký tự, dùng Menu **Graph > Character Graphs > Boxplot,** nhấn **Select** chọn GPAs vào ô **Variable**. **OK**



Ta sẽ có kết quả sau

Boxplot

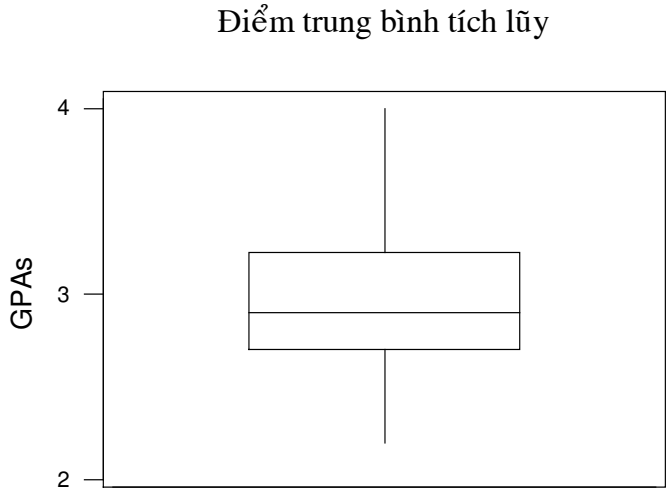
```
                          ---------------
                ----------I      +        I------------------     *
                          ---------------
            +---------+---------+---------+---------+---------+------
GPAs
          2.10      2.45      2.80      3.15      3.50      3.85
```

Biểu đồ hộp dạng ký tự sẽ trình bày dữ liệu theo chiều ngang. Thân hộp trình bày phần chính giữa của các điểm GPAs. Hai mép của hộp có vị trí xấp xỉ bằng các số tứ phân, Q1 và Q3, và trung vị được đánh dấu bởi dấu '+'. Chúng ta đã biết các giá trị này từ thống kê mô tả ví dụ trước. Khoảng 50% số điểm GPAs nằm trong khoảng 2,7 và 3,2; trung vị là 2,9.

Đường thẳng hay đường ria (wishkers) đứt nét chạy từ mép của thân hộp tới hai đầu mút nằm trong giới hạn trong (inner fences). Không có điểm bất thường nào, giá trị được đánh dấu '*' là điểm có thể bất thường. (Giá trị GPA lớn nhất là 4,0 được xác định trong ví dụ về thống kê mô tả)

Phân phối hơi bị bất đối xứng. Nó có đường ria dài và khoảng cách từ trung vị đến mép thân hộp lớn hơn về phía giá trị GPA cao.

Biểu đồ hộp dạng đồ họa được xây dựng bằng cách dùng Menu **Graph > Boxplot**, nhấn **Select** chọn GPAs vào ô **Graph 1Y;**
Chọn **Annotation**, chọn **Title**; nhập vào tiêu đề 'Điểm trung bình tích lũy'. Nhấn **OK**



Biểu đồ hộp cho thấy dữ liệu trình bày theo chiều dọc. Phần nhiều diện tích của thân hộp nằm phía trên trung vị; đường ria phía trên dài hơn đường ria phía dưới chỉ ra rằng phân phối bị lệch về phía có điểm GPAs cao. Biểu đồ hộp không phát hiện ra điểm bất thường nào.

2.6.4 Các lệnh dùng cho dữ liệu lưu trữ theo hàng

Vừa rồi chúng ta tính toán và mô tả các dữ liệu được lưu theo cột trong Minitab. Với dữ liệu được lưu theo hàng, các đại lượng thống kê được tính toán theo hàng cũng bao gồm: tổng, trung bình, độ lệch chuẩn, giá trị nhỏ nhất, giá trị lớn nhất, khoảng, trung vị, tổng bình phương, số lượng quan sát, số lượng quan sát có giá trị khuyết và không có giá trị khuyết. Đặc trưng thống kê được tính cho mỗi hàng trong tập hợp các cột và lưu trữ vào hàng tương ứng của một cột mới.

Để tính các đại lượng thống kê cho các dữ liệu lưu theo hàng ta chọn menu **Calc ▸ Row Statistics**

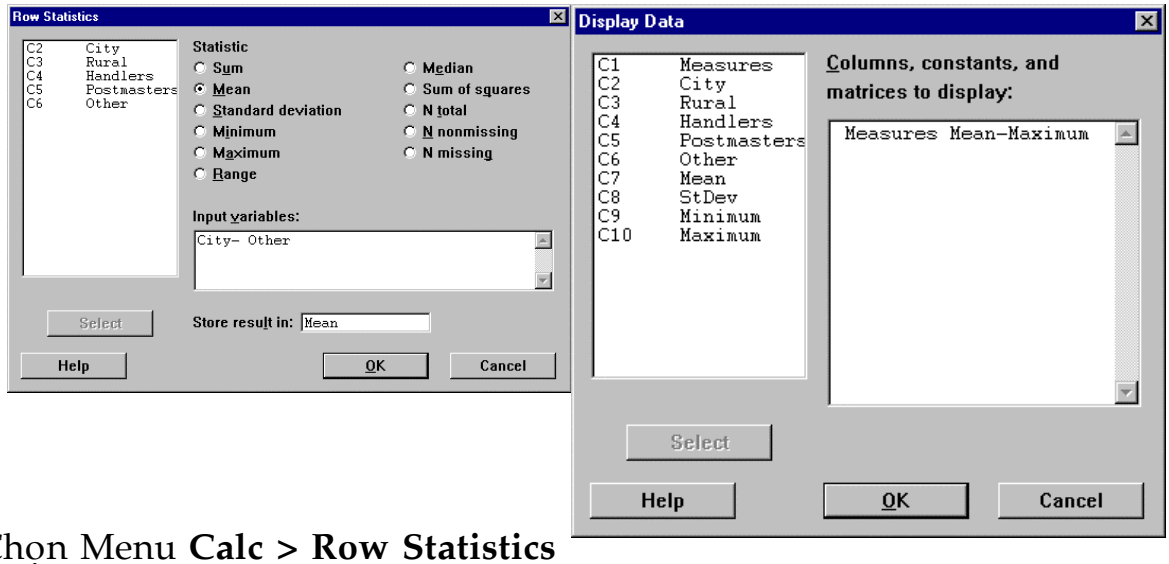
Ví dụ 13: Tính các đại lượng thống kê theo hàng

Vào năm 1998, tổng công ty bưu chính của Mỹ đã lập ra Ủy ban Dịch Vụ Thư tín về An Toàn và Bảo Mật Nơi Công Sở. Trách nhiệm của ủy ban là phát triển một kế hoạch làm cho 38.000 bưu điện và các phương tiện liên quan thành môi trường an toàn nhất cho các nhân viên làm việc. Kế tiếp Ủy ban thực hiện một nghiên cứu tập trung bao gồm một khảo sát toàn diện về tính xung đột ở công sở. Bảng sau cho biết phần trăm số câu trả lời đồng ý về thái độ và đo lường tâm lý của các loại nhân viên khác nhau của một khu vực. Hãy tính giá trị trung bình, độ lệch chuẩn, giá trị min, max của các đo lường này đối với các nhân viên.

	Loại nhân viên				
	City	Rural	Handlers	Postmasters	Other
Đại lượng đo lường	**Đưa thư ở thành phố**	**Đưa thư ở nông thôn**	**Kiểm soát thư**	**Chủ bưu điện**	**Nhà quản lý khác**
Quyền tự trị (Autonomy)	25,2	42,0	34,0	78,5	80,3
Sức ép (Pressure/Burden)	55,0	21,3	37,7	29,6	47,2
Thái độ tiêu cực (Negative Attitude)	30,5	17,9	42,3	8,3	28,7
Sự giận dữ (Anger)	5,3	2,0	4,4	1,3	2,5
Sự thù địch Hostility	14,7	9,3	16,9	7,7	10,5
Sự đối nghịch (Coping)	85,7	91,3	82,4	88,0	86,5
Sự lo lắng và cùng kiệt (Distress and Anxiety)	27,4	16,4	24,6	21,0	32,8
Công kích bằng lời (Verbal Aggressiveness)	30,7	19,5	32,7	13,4	23,3

Lời giải: Dữ liệu khảo sát lưu trong file **Attitudes.mtp.** Tên cột trong bảng Minitab tương tự như đầu đề của bảng trên.

Chọn Menu **Calc > Row Statistics**
Chọn **Mean**;
Nhấn **Select** chọn tất cả các biến từ City - Other vào **Input variables;**
Nhập 'Mean' vào ô **Store results in. OK**



Chọn Menu **Calc > Row Statistics**
Chọn **Standard deviation**; nhấn **Select** chọn tất cả City-Other vào **Input variables;** nhập 'StDev' vào ô **Store results in. OK**

Chọn Menu **Calc > Row Statistics**
Chọn **Minimum**; nhấn **Select** chọn tất cả City-Other vào **Input variables;** nhập 'Minimum' vào ô **Store results in. OK**

Chọn Menu **Calc > Row Statistics**
Chọn **Maximum**; nhấn **Select** chọn tất cả City-Other vào **Input variables; n**hập 'Maximum' vào ô **Store results in. OK**

Chọn Menu **Manip > Display data,** nhấn **Select** chọn tất cả Measures Mean-Maximum vào ô **Display. OK**

Row	Measures	Mean	StDev	Minimum	Maximum
1	Autonomy	52.00	25.7166	25.2	80.3
2	Pressure/Burden	38.16	13.4448	21.3	55.0
3	Negative Attitude	25.54	12.9525	8.3	42.3
4	Anger	3.10	1.6837	1.3	5.3
5	Hostility	11.82	3.8460	7.7	16.9
6	Coping	86.78	3.2538	82.4	91.3
7	Distress and Anxiety	24.44	6.2280	16.4	32.8
8	Verbal Aggressiveness	23.92	7.9632	13.4	32.7

Các lệnh về hàng sẽ lưu trữ các đo lường bằng số vào những cột chỉ định. Ví dụ, đo lường về Anger (sự giận dữ) sắp xếp từ 1,3 đến 5,3 với giá trị trung bình là 3,1 và độ lệch chuẩn là 1,68.

PHỤ LỤC CHƯƠNG 2:

CÁC TÙY CHỌN (OPTIONS) TRONG CÁC LỆNH:

LỆNH **Graph ▸ Stem-and-Leaf:**

Chọn By variable khi bạn muốn vẽ và nhóm các biểu đồ thân lá theo các giá trị của 1 biến nào đó (thường gọi là biến điều khiển).
Chọn Trim outliers khi bạn muốn loại bỏ các phần tử bất thường (outlier) trong dữ liệu
Chọn Increment để hiển thị độ rộng thân. Nếu chọn ô này có giá trị 1; 0.5; 0.2 nhân với đơn vị của phần thân, thì biểu đồ sẽ thể hiện 1, 2 hay 5 dòng cho mỗi thân (xem rõ trong ví dụ 7).

LỆNH **Graph ▸ Character Graphs ▸ Histogram:**

Chọn Same scale for all variables khi bạn muốn tất cả các biến muốn vẽ lược đồ đều có sự canh chỉnh (thang đo) như nhau
Chọn First midpoint để cho biết giá trị của tâm điểm đầu tiên
Chọn Last midpoint để cho biết giá trị của tâm điểm cuối cùng
Chọn Interval width để cho biết độ rộng giữa các tâm điểm

LỆNH Graph ▸ **Histogram:**

Các tùy chọn trong hộp thoại Histogram dialog box:

+ Data display (Kiểu thể hiện dữ liệu)

Display	Kiểu thể hiện dữ liệu
Area	Tô phần diện tích dưới các điểm
Bar	Thanh ngang
Connection lines	Đường thẳng nối liền các điểm với nhau
Lowess (LOcally-Weighted Scatter Plot Smoother)	Đường thẳng nối liền các điểm sau khi làm trơn
Projection lines	Đường thẳng từ các điểm vuông góc với trục x hoặc y

Symbol	Biểu tượng (điểm, hình tròn, hình vuông…) với hình dạng và kích cỡ thay đổi được

For each	Áp dụng kiểu dữ liệu cho
Graph	Cả lược đồ
Group	Từng nhóm phân biệt trong lược đồ

Group variables	Nếu bạn muốn đồ thị thể hiện một lúc nhiều nhóm (ví dụ lược đồ điểm trung bình tích lũy có phân biệt nam và nữ) thì bạn phải chọn thêm 1 biến phân biệt nam nữ vào ô này.

Ghi chú: Với mỗi kiểu thể hiện dữ liệu bạn chọn Edit Attributes *để chỉ định tính chất của mỗi kiểu.*

+ Annotation (Chú thích cho đồ thị)

Annotation	Chú thích cho đồ thị
Title	Nhập tựa đề phía trên cho đồ thị
Footnote	Nhập chú thích phía dưới của đồ thị
Text	Nhập đoạn văn bản bất kỳ trên đồ thị có toạ độ xác định (tọa độ x, y cách nhau 1 khoảng trắng trong phần Point)
Data Labels	Đặt nhãn cho dữ liệu thể hiện trên đồ thị
Line	Vẽ các đường thẳng nối các điểm có tọa độ (x y) xác định
Marker	Đánh dấu 1 biểu tượng (mark) tại một điểm có tọa độ xác định
Polygon	Vẽ một đa giác với các điểm có tọa độ (x y) xác định, để chú thích một vùng nào đó trên đồ thị

Ghi chú: Ngoài cách khai báo trong hộp thoại như trên, bạn có thể sử dụng công cụ trực quan hơn. Khi Minitab đã vẽ xong đồ thị, bạn nhấp đôi chuột vào đồ thị đã vẽ sẵn thì sẽ xuất hiện các hộp công cụ dùng để thực hiện những chú thích này.

+ Frame (Khung dạng của đồ thị)

Frame	Khung dạng của đồ thị
Axis	Định dạng các trục đồ thị: nhãn, kiểu đường trục…
Tick	Định dạng khoảng chia trên trục: nhãn, kiểu thang đo…
Grid	Định dạng các đường kẻ tọa độ (đường ngang và đường dọc)
Reference	Vẽ các đường chú dẫn ngang hoặc dọc trên đồ thị
Min and Max	Thay đổi khoảng đo của các trục bằng cách xác định giá trị nhỏ nhất và lớn nhất cho từng trục đồ thị
Multiple Graph	Thể hiện nhiều đồ thị một lúc cùng với các tùy chọn: - Nhiều đồ thị trên cùng một hệ tọa độ - Các đồ thị riêng biệt nhau và các trục của chúng có thể canh chỉnh khác nhau hay giống nhau.

+ Region (Vùng của đồ thị)

Region	Vùng của đồ thị
Figure	Định dạng vùng mà đồ thị được vẽ
Data	Định dạng vùng mà dữ liệu được vẽ
Legend	Định dạng ô chú thích

+ Option (tùy chọn)

Type of histogram	Kiểu lược đồ
Frequency	Kiểu tần suất Chiều cao của cột là số lượng quan sát rơi trong khoảng biểu diễn
Percent	Kiểu phần trăm Chiều cao của cột là phần trăm so với tổng quan sát của tất cả các khoảng

Density	Kiểu diện tích Tổng diện tích các cột là 1, chiều cao của cột bằng: (số quan sát trong 1 khoảng)/[(tổng số quan sát)*(bề rộng khoảng)]
Cumulative Frequency	Kiểu tần suất tích lũy
Cumulative Percent	Kiểu phần trăm tích lũy
Cumulative Density	Kiểu diện tích tích lũy

Type of Intervals	**Kiểu khoảng chia**
MidPoint	Thể hiện bề rộng mỗi cột theo điểm giữa
CutPoint	Thể hiện bề rộng mỗi cột theo giá trị biên

Definition of Intervals	**Xác định khoảng chia (chọn 1 trong 3 cách)**
Automatic	Minitab tự động chia khoảng
Number of intervals	Xác định số lượng khoảng chia
Midpoint/cutpoint positions	Xác định cụ thể điểm cắt, mỗi điểm cách nhau 1 khoảng trắng.

Transpose X and Y	**Hoán đổi 2 trục tọa độ**

BÀI TẬP

2.1. Sinh viên lớp ngoại ngữ được hỏi về kinh nghiệm sử dụng máy tính. Mức độ kinh nghiệm được gán là không biết (1), biết ít (2), biết nhiều (3). Dùng dữ liệu 32 sinh viên dưới đây để mô tả phân phối kinh nghiệm về máy tính của sinh viên.

1	1	2	3	3	2	3	3
3	2	3	2	3	2	2	3
2	3	1	1	3	1	1	1

2.2. Một cửa hàng sách đã thống kê số lượng sách bán trong mùa khuyến mãi vừa qua và muốn biết phân phối số sách được bán. Sách được bán gồm 3 loại: Tiểu thuyết (TT), truyện tranh thiếu nhi (TN), sách giáo khoa (SGK). Mô tả phân phối loại sách bán được trong đợt khuyến mãi.

TT	TT	TN	SGK	TT	TN
TN	SGK	TT	SGK	TN	TN
TN	TT	TN	TT	TT	TT
TN	SGK	TN	TN	TN	SGK
SGK	TT	TN	TN	SGK	TT

2.3. Một tạp chí muốn khảo sát độc giả về việc họ có đọc quảng cáo hay không, họ tiến hành khảo sát 50 người với hai câu hỏi là tuổi dưới 30 và có đọc quảng cáo hay không. Kết quả thể hiện trong bảng dưới đây.

Stt	Duoi 30	Đọc QC	Stt	Duoi 30	Đọc QC	Stt	Duoi 30	Đọc QC
1	co	co	18	co	co	35	khong	khong
2	co	khong	19	co	khong	36	co	co
3	khong	khong	20	co	khong	37	co	co
4	khong	co	21	khong	co	38	khong	khong
5	co	khong	22	co	khong	39	co	khong
6	khong	co	23	khong	khong	40	co	co
7	co	co	24	co	khong	41	khong	khong

8	khong	co	25	khong	co	42	khong	khong
9	khong	co	26	co	khong	43	co	co
10	co	khong	27	co	khong	44	khong	co
11	khong	co	28	co	khong	45	co	khong
12	khong	khong	29	Khong	co	46	khong	co
13	co	co	30	co	khong	47	co	khong
14	khong	khong	31	co	khong	48	co	khong
15	co	khong	32	Khong	co	49	khong	Co
16	co	co	33	co	khong	50	khong	Co
17	khong	co	34	co	khong			

a. Tính tỷ lệ số người có đọc quảng cáo trong mẫu.
b. Dựng bảng 2 chiều. Liệu tuổi có ảnh hưởng đến việc đọc quảng cáo hay không? Giải thích tại sao.

2.4. Một công ty lớn thuê nhà cho nhân viên tại một thành phố. Bộ phận nhân sự khảo sát tiền thuê hàng tháng các căn hộ trong thành phố. Sau đây là giá thuê của 40 căn hộ loại 3 phòng ngủ được chọn ngẫu nhiên: (File: phoenix.mtp)

Giá thuê ($)

925	900	1175	850	1250
975	935	1325	1095	900
1050	1065	1465	1020	875
865	665	875	925	1075
925	800	875	1150	895
870	895	1035	1275	1100
1475	925	1100	1150	1350
1420	1025	1285	950	1070

a. Hãy mô tả bằng đồ thị và bằng số về dữ liệu giá thuê phòng. Viết một đoạn văn ngắn mô tả dữ liệu
b. Tìm phần trăm các quan sát (các giá thuê) nằm trong khoảng $\bar{x} \pm s$, $\bar{x} \pm 2s$, $\bar{x} \pm 3s$. So sánh các kết quả này với Quy tắc Thực nghiệm.

c. Tính toán và xếp thứ tự các điểm z. Có phần tử bất thường nào không?

2.5. Giám đốc của một công ty quan tâm đến số nhân viên được đào tạo nhưng nghỉ hưu sớm. Nếu quá nhiều nhân viên nghỉ hưu cùng một lúc, việc thay người rất khó khăn. Bảng sau cho biết số năm làm việc của một mẫu ngẫu nhiên của 151 nhân viên. (File: ServiceYears.mtp)

Số năm làm việc

13	16	25	3	27	7	7	2	3	16	7	26	1	8
6	27	6	6	8	23	3	21	4	12	0	9	3	32
27	5	23	9	9	6	9	7	9	13	1	15	20	9
2	1	13	18	10	27	26	4	27	9	10	7	7	13
27	2	7	23	26	16	5	23	6	9	30	4	5	18
4	4	0	10	10	7	2	27	26	3	29	29	7	1
19	19	5	5	10	28	21	20	23	8	3	17	17	26
30	14	17	6	14	20	0	27	22	28	20	0	8	13
19	1	2	18	26	9	3	21	8	17	1	29	21	30
7	6	18	2	10	6	26	9	22	13	7	8	28	44
26	28	16	29	2	9	17	2	8	23	39			

a. Mô tả bằng số và bằng đồ thị dữ liệu về năm làm việc. Bạn có nghĩ giám đốc công ty nên quan tâm về các nhân viên nghỉ hưu không? Viết một tóm tắt ngắn.
b. Tìm phần trăm các quan sát (năm làm việc) nằm trong khoảng $\bar{x} \pm s$, $\bar{x} \pm 2s$, $\bar{x} \pm 3s$. So sánh kết quả này với Quy tắc Thực nghiệm.
c. Hãy tính và sắp xếp các điểm z. Có phần tử nào bất thường hay không?

2.6. Một nha sĩ thực hiện một thí nghiệm để đo lường sự hiệu quả của một cách gây tê răng kiểu mới. Hai mươi lăm bệnh nhân được gây tê theo cách bình thường và hai mươi lăm bệnh nhân khác được gây tê theo cách mới. Mỗi bệnh nhân sẽ đánh giá sự không thoải mái của họ bằng cách cho điểm từ 0 đến 100, điểm

càng cao sự không thoải mái càng cao. Hãy so sánh bằng số và bằng đồ thị hai mẫu này. (File: Dentist.mtp)

Cách bình thường					**Cách mới**				
23	26	44	32	44	62	50	40	35	52
44	34	26	49	67	82	74	87	30	58
44	53	79	52	33	39	51	72	56	50
50	43	49	33	52	85	57	39	48	64
51	6	30	38	22	75	46	48	56	60

2.7. Một nhà quản lý cửa tiệm tra dầu và dầu nhờn nhanh cho xe hơi, ông ta quan tâm đến lượng thời gian cần để phục vụ cho các xe hơi. Hệ thống cũ có một thợ máy thực hiện tất cả mọi việc. Ông ta đề xuất một hệ thống ở đó thợ máy chỉ việc thay dầu. Một thợ thứ hai vừa tra dầu nhờn cho xe vừa kiểm tra mức an toàn. Một mẫu gồm 50 xe hơi dùng hệ thống cũ và 50 xe hơi dùng hệ thống mới, kết quả về thời gian phục vụ theo đơn vị phút. Hãy so sánh hai hệ thống bằng số và đồ thị. Bạn có nghĩ rằng hệ thống mới phục vụ tốt hơn hệ thống cũ? Hãy thảo luận. (File: SystemTimes.mtp)

Thời gian của hệ thống cũ					**Thời gian của hệ thống mới**				
6	11	12	13	13	6	7	8	11	11
14	15	16	17	20	12	13	14	15	16
10	12	12	13	14	6	8	9	11	11
14	15	16	18	20	12	13	14	15	16
10	12	13	13	14	7	8	9	11	12
14	15	17	19	21	12	14	14	15	19
11	12	13	13	14	7	8	9	11	12
15	16	17	19	22	12	14	15	15	19
11	12	13	13	14	7	8	10	11	12
15	16	17	19	28	13	14	15	15	23

2.8. Chỉ số giá tiêu dùng (CPI) đo lường sự thay đổi theo thời gian của giá cả thức ăn và dịch vụ. Bản tóm tắt thống kê của Mỹ năm 1989 đã báo cáo phần trăm thay đổi trong giá tiêu dùng ở nhiều nước từ 1977 đến 1987. (File: CountryCPI.mtp)

Nước	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987
U.S.	7,6	11,3	13,5	10,3	6,1	3,2	4,3	3,5	1,9	3,7
Canada	8,9	9,2	10,2	12,5	10,8	5,8	4,3	4,0	4,2	4,4
Japan	3,8	3,6	8,0	4,9	2,7	1,9	2,2	2,1	0,4	-0,2
Austria	3,6	3,7	6,4	6,8	5,4	3,3	5,6	3,2	1,7	1,4
Belgium	4,5	4,5	6,6	7,6	8,7	7,7	6,3	4,9	1,3	1,6
Denmark	10,0	9,6	12,3	11,7	10,1	6,9	6,3	4,7	3,6	4,0
France	9,1	10,8	13,6	13,4	11,8	9,6	7,4	5,8	2,7	3,1
Italy	12,4	15,7	21,1	18,7	16,3	15,0	10,6	8,6	6,1	4,6
Spain	19,8	15,7	15,5	14,6	14,4	12,2	11,3	8,8	8,8	5,3
U,K,	8,3	13,4	18,0	11,9	8,6	4,6	5,0	6,1	3,4	4,2
W,Germ,	2,7	4,1	5,5	6,3	5,3	3,3	2,4	2,2	-0,2	0,2

Ví dụ, con số ở cột 1978 ở nước Mỹ (U.S) cho thấy rằng giá của hàng hóa tiêu dùng tăng 7,6% từ năm 1977 qua năm 1978.

a. Đối với mỗi năm, hãy tính giá trị trung bình và độ lệch chuẩn của thay đổi giá tiêu dùng. Xây dựng biểu đồ hộp cho mỗi năm. Xu hướng giá tiêu dùng trên khắp thế giới như thế nào?
b. Đối với mỗi nước, hãy tính giá trị trung bình, trung vị, và độ lệch chuẩn của thay đổi giá tiêu dùng. Nhìn chung các nước đều muốn mức tăng thấp và không đổi trong giá tiêu dùng. Độ lệch chuẩn thấp cho thấy mức tăng ổn định. 3 nước nào có độ tăng giá tiêu dùng thấp nhất? 3 nước nào có mức tăng giá tiêu dùng ổn định nhất?

2.9. Một giáo sư hóa học chấm điểm đề án thí nghiệm của mỗi sinh viên. Bảng sau cho thấy điểm thí nghiệm của 13 sinh viên trong 6 lần thực hiện thí nghiệm. (không có file)

Tên sinh viên	1	2	3	4	5	6
Lois	48	50	0	49	44	44
Peter	42	46	47	43	30	0
Tom	48	45	43	38	35	33
Mike	44	47	45	46	40	29
Hugh	44	32	42	50	38	43
Carol	42	48	49	50	48	41
Dean	44	44	49	43	44	41
Roy	48	46	49	47	40	48

Robin	50	38	47	43	35	28
Gina	46	43	48	50	36	32
Bob	42	46	49	50	45	39
Lora	43	41	48	48	44	43
Mark	42	48	49	45	39	37

a. Xác định 3 sinh viên nào có điểm thí nghiệm trung bình cao nhất? Thấp nhất?
b. Thao tác ổn định là một kỹ năng quan trọng khi làm thí nghiệm (thể hiện qua điểm số). Có sinh viên nào thao tác ổn định hơn những sinh viên khác không? Hãy thảo luận.
c. Điểm thí nghiệm có thay đổi qua các lần thực hiện thí nghiệm hay không? Hãy thảo luận.
d. Nhóm tất cả các điểm vào trong 1 cột. Xây dựng một biểu đồ tần suất và một biểu đồ hộp cho dữ liệu đã nhóm. Mô tả phân phối của điểm phòng thí nghiệm.

2.10. Trung tâm bất động sản Minnesota tổng hợp thông tin về nhà bán trong các khu vực của một thành phố. Bảng sau cho biết giá bán của 20 căn nhà chọn ngẫu nhiên từ số nhà bán trong quận A và B của thành phố vào năm 1998. (File: MNHomes.mtp)

Quận A ($)		**Quận B ($)**	
105000	124400	123925	159900
66000	110600	86000	67800
98900	73500	29900	116000
143000	139500	73000	112330
136000	74000	145500	74900
66600	84500	81500	164000
119875	91900	84000	109000
84000	89900	100750	105900
72000	131900	94500	155000
72500	74500	149195	78000

a. So sánh giá bán nhà giữa A và B bằng số và đồ thị. Giá bán nhà ở quận nào cao hơn và biến động hơn?
b. Có phần tử bất thường nào không? Hình dạng phân phối của của giá bán như thế nào?

2.11. Vụ điều tra dân số, Vụ thương mại, và Phòng sức khỏe-nhân sự là các tổ chức trong những tổ chức thu thập dữ liệu về nước Mỹ. Bảng sau cho biết ngân sách năm 1995 gồm thu nhập và chi tiêu quốc gia theo đơn vị tỷ đô-la cho 50 bang và Quận Columbia. (File: StateDept.mtp)

Bang	Thu nhập	Chi tiêu	Bang	Thu nhập	Chi tiêu
Alabama	11,389	10,242	Montana	3,023	2,663
Alaska	7,358	5,423	Nebraska	3,890	3,823
Arizona	10,843	9,783	Nevada	4,500	4,051
Arkansas	6,446	5,915	NewHampshire	3,011	2,970
California	108,222	104,567	NewJersey	29,614	28,923
Colorado	10,028	8,673	NewMexico	6,303	5,599
Connecticut	12,744	12,507	NewYork	78,209	74,280
Delaware	2,876	2,557	NorthCarolina	19,377	19,916
DistColumbia	3,376	3,391	NorthDakota	2,288	2,129
Florida	33,216	30,103	Ohio	38,341	31,685
Georgia	16,585	15,308	Oklahoma	8,679	8,272
Hawaii	5,543	5,606	Oregon	10,826	9,013
Idaho	3,406	2,776	Pennsylvania	37,779	34,359
Illinois	30,351	28,132	RhodeIsland	3,765	4,176
Indiana	14,653	14,136	SouthCarolina	10,637	10,386
Iowa	8,224	7,766	SouthDakota	1,942	1,686
Kansas	6,730	5,742	Tennessee	11,864	11,028
Kentucky	11,011	10,543	Texas	42,019	39,091
Louisiana	13,348	12,893	Utah	5,348	4,833
Maine	3,926	3,889	Vermont	1,953	1,849
Maryland	14,842	13,537	Virginia	16,307	14,721
Massachusetts	21,493	21,557	Washington	19,930	18,003
Michigan	28,760	27,051	WestVirginia	6,047	5,943
Minnesota	16,245	14,295	Wisconsin	18,677	14,621
Mississippi	7,205	6,235	Wyoming	2,181	1,887
Missouri	12,559	10,809			

a. Hãy tính toán các đo lường mô tả cho thu nhập và chi tiêu quốc dân cho 50 bang và Quận Columbia, và cho 50 bang không có Quận Columbia. Hãy so sánh.
b. Xây dựng biểu đồ hộp cho các giá trị ngân sách. Hãy diễn dịch.

c. Hãy tính các điểm z cho thu nhập quốc dân. Xếp các giá trị điểm z từ lớn đến nhỏ. Có phần tử bất thường nào không?

2.12. Kết quả khảo sát tình hình sử dụng ma túy, rượu, thuốc lá trong sinh viên các trường học được báo cáo trong một sách thống kê phát hành bởi Cơ quan công lý. Bảng sau cho biết phần trăm dùng các loại thuốc trong 30 ngày trước thời điểm hỏi của sinh viên. Số người trả lời trong khảo sát từ 1080 đến 1410 cho các năm từ 1984 đến 1991. (File: Drug.mtp)

Drug	**1984**	**1985**	**1986**	**1987**	**1988**	**1989**	**1990**	**1991**
Marihuana	23,0	23,6	22,3	20,3	16,8	16,3	14,0	14,1
Inhalants	0,7	1,0	1,1	0,9	1,3	0,8	1,0	0,9
Hallucinogens	2,6	2,0	3,6	3,4	2,8	3,7	2,5	2,0
Cocaine	7,6	6,9	7,0	5,0	4,7	3,0	2,3	2,3
Heroin	0,0	0,0	0,0	0,1	0,1	0,1	0,0	0,1
Other Opiates	1,4	0,7	0,6	0,8	0,8	0,7	0,5	0,6
Stimulants	5,5	4,2	3,7	2,3	1,8	1,3	1,4	1,0
Sedatives	2,2	1,4	1,3	1,3	1,2	0,4	0,2	0,3
Tranquilizers	1,1	1,4	1,9	1,0	1,1	0,8	0,5	0,6
Alcohol	79,1	80,3	79,7	78,4	77,0	76,2	74,5	74,7
Cigarettes	21,5	22,4	22,4	24,0	22,6	21,1	21,5	23,2

Hãy tính giá trị trung bình, độ lệch chuẩn, giá trị nhỏ nhất và lớn nhất về tình hình sử dụng thuốc trong các năm.

CHƯƠNG 3
PHÂN PHỐI XÁC SUẤT

Chương 3 giới thiệu một số dạng phân phối xác suất, phân phối mẫu cũng như những kết luận thống kê cho trường hợp đơn mẫu và hai mẫu.

3.1 PHÂN PHỐI XÁC SUẤT

Biến ngẫu nhiên là một hàm số có giá trị thực được xác định trên không gian mẫu. Có 2 dạng biến ngẫu nhiên: rời rạc (gián đoạn) và liên tục.

3.1.1 Biến Ngẫu Nhiên Rời Rạc (Discrete Random Variables)

Biến ngẫu nhiên rời rạc là biến ngẫu nhiên mà các giá trị có thể có của nó là hữu hạn hoặc đếm được (ví dụ, giá trị nhận được khi thảy một con xúc sắc).

Kỳ vọng và Phương sai:
Kỳ vọng hay giá trị trung bình số học của một biến ngẫu nhiên rời rạc x, ký hiệu là μ:

$$\mu = \sum_{\forall x} xp(x)$$

Phương sai hay giá trị trung bình trọng số của bình phương khoảng chênh lệch giữa x và μ, ký hiệu là σ^2:

$$\sigma^2 = \sum_{\forall x} (x-\mu)^2 p(x)$$

Độ lệch chuẩn σ bằng căn bậc hai của phương sai σ^2.

Phân phối xác suất của biến ngẫu nhiên rời rạc cho biết tất cả các giá trị có thể có của biến ngẫu nhiên này và xác suất tương ứng của từng giá trị.

Ứng dụng Minitab:

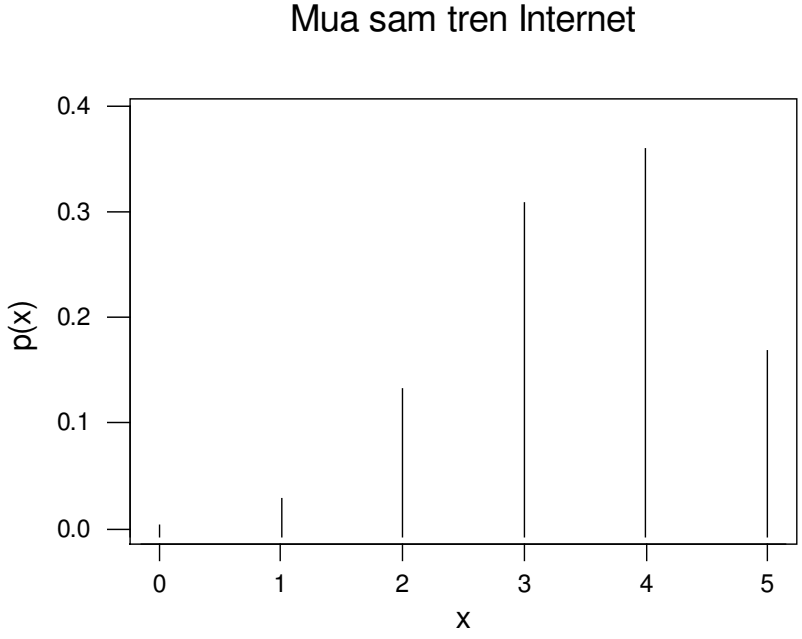
Ví dụ 1: Gọi x là số người thích mua sắm trên Internet trong mẫu gồm 5 khách hàng. Bảng giá trị sau là phân phối xác suất của x. Hãy vẽ phân phối xác suất này, tìm giá trị kỳ vọng, phương sai và độ lệch chuẩn của biến ngẫu nhiên x.

X	p(x)
0	0,003
1	0,028
2	0,132
3	0,309
4	0,360
5	0,168
	1,000

Lời giải: Nhập các giá trị của bảng trên lần lượt vào các cột C1 và C2 có tên là x và p(x). Sau đó dùng lệnh PLOT vẽ phân phối xác suất.

Graph > Plot

Chọn p(x) cho **Graph 1 Y** và x cho **Graph 1 X**;
Nhấp chuột vào **Display**, chọn **Project**;
Nhấp chuột vào **Annotation**, rồi **Title**, nhập tên vào. nhấn **OK.**



Đồ thị này cho thấy phân phối bị lệch về phía các giá trị nhỏ của x.

Để tính giá trị kỳ vọng, phương sai và độ lệch chuẩn của phân phối xác suất, nhập công thức tính μ, σ^2, và σ. (lưu ý: cần sử dụng dấu nháy đơn ' cho việc đặt tên cho cột lưu kết quả vừa tính toán)

Tính giá trị kỳ vọng:
Calc > Calculator
Nhập 'KyVong' vào **Store results in variable**:
Nhập SUM('x'*'p(x)') vào **Expression**. Sau đó chọn **OK**.

Tính giá trị phương sai:
Calc > Calculator
Nhập 'PhuongSai' vào **Store results in variable**:
Nhập SUM(('x'-'KyVong')**2*'p(x)') vào **Expression**. Sau đó chọn **OK**.

Tính giá trị độ lệch chuẩn:
Calc > Calculator
Nhập 'DoLechChuan' vào **Store results in variable**:
Nhập SQRT('PhuongSai') vào **Expression**. Sau đó chọn **OK**.

Sau khi hoàn tất các bước trên, ta có thể xem các giá trị ban đầu của x, p(x) và các kết quả tính toán vừa rồi bằng cách sau:

Manip > Display Data
Chọn tất cả các cột từ x-DoLechChuan ở **Display**. Sau đó chọn **OK**.

Row	x	p(x)	KyVong	PhuongSai	DoLechChuan
1	0	0.003	3.499	1.05400	1.02664
2	1	0.028			
3	2	0.132			
4	3	0.309			
5	4	0.360			
6	5	0.168			

Giá trị kỳ vọng của 5 khách hàng thích mua sắm trên Internet là 3,5 với độ lệch chuẩn là 1,0266

Một số biến ngẫu nhiên rời rạc đặc biệt

Chương này sẽ giới thiệu các dạng biến ngẫu nhiên nhị thức (binomial) và Poisson và các phân phối tương ứng. Biến ngẫu nhiên nhị thức mô tả kết quả của một thực nghiệm có thể là thành công hoặc thất bại. Biến nhẫu nhiên Poisson mô tả số lần xuất hiện của một biến cố trong một khoảng thời gian hay một khoảng xác định nào đó.
Minitab cho phép người sử dụng tính toán xác suất, xác suất tích lũy, và chuyển đổi xác suất tích lũy cho biến ngẫu nhiên x. Tùy chọn **xác suất** (probability) cho biết xác suất xảy ra tại mỗi giá trị của x. Tùy chọn **xác suất tích lũy** (cumulative probability) cho biết xác suất mà biến ngẫu nhiên x nhỏ hơn hay bằng giá trị đang xét. Tùy chọn **chuyển đổi xác xuất tích lũy** (inverse cumulative probability) cho biết giá trị của x_o tương ứng với xác suất p sao cho $P(x \le x_o) = p$. Nếu không có giá trị nào x_o tương ứng với xác suất p chính xác thì Minitab sẽ in ra 2 kết quả tương ứng với 2 xác suất nhỏ hơn và lớn hơn gần p nhất.

Biến Ngẫu Nhiên Nhị Thức (Binomial)

Biến ngẫu nhiên nhị thức xuất hiện khi kết quả của một thực nghiệm rơi vào một trong hai trạng thái. Ví dụ, khi thảy một đồng xu thì kết quả nhận được có thể là mặt sấp hoặc mặt ngữa; một đứa trẻ được sinh ra có thể là trai hoặc gái … Thông thường hai trạng thái này được thể hiện là thành công và thất bại.

Phân phối nhị thức là phân phối xác suất của số lần thành công trong n lần phép thử độc lập. Xác suất thành công, được ký hiệu là π (pi), là không đổi trong mỗi phép thử. Xác suất thất bại là $1 - \pi$. Giá trị kỳ vọng của biến ngẫu nhiên nhị thức là $\mu = n\pi$ và phương sai tương ứng là $\sigma^2 = n\pi(1-\pi)$. Gọi x là biến ngẫu nhiên tuân theo phân bố nhị thức, xác suất của x:

$$P(x) = \frac{n!}{x!(n-x)!}\pi^x(1-\pi)^{n-x}$$

Ứng dụng Minitab:

Ví dụ 2: Một công ty tiến hành điều tra xem khách hàng của công ty có thích mua hàng trên Internet không. Giả sử rằng 70% số khách hàng

của công ty thích mua sắm trên Internet. Công ty chọn ngẫu nhiên 5 khách hàng để tiến hành điều tra và ghi nhận số người thích mua sắm trên Internet trong 5 người này. Gọi x là số khách hàng thích mua sắm trên Internet.

a. Hãy mô tả phân phối xác suất và tìm giá trị kỳ vọng cũng như độ lệch chuẩn tương ứng.
b. Xác suất để không có khách hàng nào thích mua sắm trên Internet là bao nhiêu? Và xác suất để có 4 khách hàng thích mua sắm trên Internet?
c. Xác suất để có nhiều hơn hoặc bằng 3 khách hàng thích mua sắm trên Internet?
d. Tìm số lượng khách hàng x_o tương ứng sao cho xác suất để x có giá trị nhỏ hơn hoặc bằng x_o là 0.5.

Lời giải: Mỗi khách hàng sẽ có câu trả lời là thích hoặc không thích mua sắm trên Internet (chỉ có hai khả năng xảy ra). Cho nên số khách hàng x thích mua sắm trên Internet là biến ngẫu nhiên nhị thức với n = 5, và π = 0,7. Giá trị có thể của x là từ 0 đến 5 khách hàng.

a. Mô tả bảng phân phối xác suất. Các giá trị có thể của x là (0, 1, …, 5) được nhập vào cột C1 có tên là 'x'.

Calc > Probability Distribution > Binomial

Chọn **Probability**

Nhập 5 vào **Number of trials:** và 0.7 vào **Probability of success:**

Chọn x vào **Input column**. Sau đó chọn **OK**.

Probability Density Function

```
Binomial with n = 5 and p = 0.700000

         x     P( X = x )
      0.00         0.0024
      1.00         0.0284
      2.00         0.1323
      3.00         0.3087
      4.00         0.3602
      5.00         0.1681
```

Để ý rằng hàm phân phối này giống với ví dụ 1 trên

Với n = 5 và π = 0,7, ta có:

+ Kỳ vọng: $\mu = n\pi = 3,5$

+ Độ lệch chuẩn: $\sigma = [n\pi(1-\pi)]^{1/2} = 1,0247$

Các giá trị kỳ vọng và độ lệch chuẩn trên gần bằng với giá trị kỳ vọng và độ lệch chuẩn ở ví dụ 1.

b. Từ phân phối xác suất trên, ta thấy rằng xác suất để không có khách hàng nào thích mua sắm trên Internet là 0,0024; và xác suất để có 4 khách hàng thích mua sắm trên Internet là 0,3602.

c. Để tìm được xác suất từ 3 khách hàng trở lên thích mua sắm trên Internet, cần lưu ý rằng Minitab chỉ cho phép tính xác suất tích lũy của biến ngẫu nhiên x nhỏ hơn hay bằng giá trị nào đó đang xét. Để tính P(x≥3), nhận thấy rằng P(x≥3) = 1 – P(x≤2). Như vậy ta phải tính P(x≤2) trước.

Calc > Probability Distribution > Binomial

Chọn **Cumulative probability**

Nhập 5 vào **Number of trials:** và 0.7 vào **Probability of success:**

Nhập 2 vào **Input constant**. Sau đó chọn **OK**.

Cumulative Distribution Function

```
Binomial with n = 5 and p = 0.700000

     x    P( X <= x )
  2.00         0.1631
```

Giá trị 0,1631 là xác suất tích lũy tương ứng với x≤2. Cho nên xác suất để từ 3 khách hàng trở lên thích mua sắm trên Internet là P(x≥3) = 1 – P(x≤2) = 1 – 0,1631 = 0,8369.

d. Để tìm giá trị x_o tương ứng với xác suất để x có giá trị nhỏ hơn hoặc bằng x_o là 0,5, ta sử dụng tùy chọn chuyển đổi xác suất tích lũy: nghĩa là tìm x_0 sao cho $P(x \leq x_o) = 0,5$.

Calc > Probability Distribution > Binomial

Chọn **Inverse cumulative probability**

Nhập 5 vào **Number of trials:** và 0.7 vào **Probability of success:**

Nhập 0.5 vào **Input constant**. Sau đó chọn **OK**.

Inverse Cumulative Distribution Function

```
Binomial with n = 5 and p = 0.700000

          x     P( X <= x )               x     P( X <= x )
          3          0.4718               4          0.8319
```

Minitab cho biết kết quả xác suất tích lũy của x bằng 3 và 4. Trong đó giá trị x=3 có xác suất tích lũy gần bằng 0,5. Như vậy khi số khách hàng thích mua sắm trên Internet bằng 3 thì xác suất tích lũy tương ứng bằng 0,5. ($x_0 = 3$)

Biến Ngẫu Nhiên Poisson

Phân bố xác suất Poisson hữu ích cho trường hợp mô tả số lần xuất hiện của một biến cố trong một khoảng thời gian xác định hoặc trong một đơn vị xác định của đo lường khác. Chẳng hạn như số ca cấp cứu trong một giờ ở bệnh viện cấp cứu Trưng Vương, số máy tính PC bị hỏng trong một tháng, số lỗi chính tả đánh máy trong một trang, số tai nạn lao động ở một phân xưởng sản xuất …

Phân bố Poisson chỉ có một tham số duy nhất là kỳ vọng μ, số lần xuất hiện trung bình của một biến cố trong 1 khoảng xác định. Phương sai của phân bố này σ^2 cũng chính bằng μ. Gọi x là biến ngẫu nhiên tuân theo phân bố Poisson, xác suất của x:

$$p(x) = \frac{e^{-\mu}\mu^{x}}{x!}$$

Ứng dụng Minitab:

Ví dụ 3: Trong một thành phố lớn, tại khu vực giao nhau giữa đường bộ và đường sắt để báo hiệu đoàn xe lửa sắp đến người ta kéo một hồi còi dài. Trung bình một ngày có 80 tiếng còi. Dân cư trong khu vực này đề nghị không kéo còi trong suốt buổi tối. Dữ liệu trong quá khứ cho biết trung bình số ca tai nạn giao thông tại khu vực này nếu không kéo còi là 5,5 ca/năm. Giả sử số tai nạn giao thông/năm ký hiệu là x tuân theo phân bố Poisson. Hãy tìm:

a. Giá trị trung bình và độ lệch chuẩn của x.
b. Phân phối xác suất và phân phối xác suất tích lũy số tai nạn giao thông/năm. Vẽ phân phối xác suất đó.

c. Xác suất để khu vực đó chỉ xảy ra một tai nạn giao thông duy nhất, xác suất để có nhiều nhất là ba tai nạn giao thông khi không kéo còi.

Lời giải:

a. Giá trị trung bình và độ lệch chuẩn của x

$$\sigma^2 = \mu = 5,5$$
$$\sigma = 2,3452$$

b. Phân phối xác suất và phân phối xác suất tích lũy

Để tính được phân phối xác suất, ta cần tạo ra các giá trị của x. Vì giá trị của x không thể là một giá trị âm, ta sẽ tạo ra các giá trị của x từ 0 đến 3 lần độ lệch chuẩn (3σ) so với giá trị trung bình. Nghĩa là nhập giá trị từ 0 đến 15 vào cột C1 và đặt tên cột là 'x'.

Phát giá trị x từ 0 đến 15:

Calc > Make Patterned Data > Simple Set of Numbers

Nhập 'x' vào **Store patterned data in**:

Nhập **From first value:** 0 và **To last value:** 15**.** Sau đó chọn **OK**.

Tính phân bố xác suất của x:

Calc > Probability Distributions > Poisson

Nhấp chuột vào **Probability**:

Nhập 5.5 vào **Mean**:

Chọn x vào **Input column**:

Nhập 'p(x)' vào **Optional storage**. Sau đó chọn **OK**.

Tính phân bố xác suất tích lũy của x:

Calc > Probability Distributions > Poisson

Nhấp chuột vào **Cumulative probability**:

Nhập 5.5 vào **Mean**:, Chọn x vào **Input column**:

Nhập 'cum p(x)' vào **Optional storage**. Sau đó chọn **OK**.

Hiển thị các giá trị:

Manip > Display Data

Chọn x, p(x), và cum p(x) vào **Display**. Sau đó chọn **OK**.

Data Display

Row	x	p(x)	cum p(x)
1	0	0.004087	0.004087
2	1	0.022477	0.026564
3	2	0.061812	0.088376
4	3	0.113323	0.201699
5	4	0.155819	0.357518
6	5	0.171401	0.528919
7	6	0.157117	0.686036
8	7	0.123449	0.809485
9	8	0.084871	0.894357
10	9	0.051866	0.946223
11	10	0.028526	0.974749
12	11	0.014263	0.989012
13	12	0.006537	0.995549
14	13	0.002766	0.998315
15	14	0.001087	0.999401
16	15	0.000398	0.999800

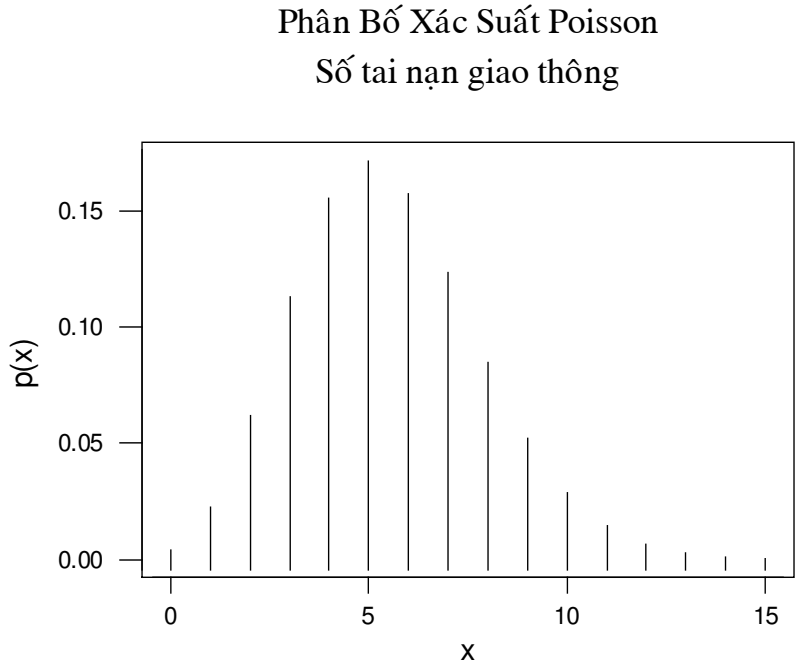
Vẽ phân bố xác suất:

Graph > Plot

Chọn p(x) vào **Graph 1 Y** và x vào **Graph 1 X**

Nhấp chuột vào **Display**, chọn **Project**;

Nhấp chuột vào **Annotation**. Nhấp chuột vào **Title**, rồi đánh tên vào. Sau đó chọn **OK**.



Đồ thị này cho thấy phân bố xác suất hơi lệch về bên phải.

c. Xác suất để chỉ có xảy ra 1 tai nạn giao thông duy nhất: 0,0266
Xác suất để số tai nạn giao thông xảy ra nhiều nhất là 3: 0,2027
(Có thể dùng tùy chọn xác suất tích lũy để tính toán giống như ví dụ 2 trên)

3.1.2 Biến Ngẫu Nhiên Liên Tục (Continuous Random Variables)

Biến ngẫu nhiên liên tục là biến ngẫu nhiên mà các giá trị có thể có của nó là liên tục hoặc không đếm được (ví dụ, tuổi thọ của một chiếc xe hơi).

Biến Ngẫu Nhiên Tuân Theo Phân Bố Chuẩn (Normal Distribution)

Phân bố chuẩn là một trong những phân bố quan trọng nhất trong thống kê. Phân bố chuẩn có dạng hình chuông (đối xứng) và được diễn đạt qua 2 thông số, giá trị kỳ vọng μ và độ lệch chuẩn σ. Các biến phân bố chuẩn có thể là trọng lượng của một đứa bé 1 tuổi, thời gian truy cập Internet, thời gian hoàn tất một thí nghiệm … Hàm mật độ xác suất của biến ngẫu nhiên x tuân theo phân bố chuẩn:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

Biến ngẫu nhiên phân bố chuẩn có thể được chuyển thành biến ngẫu nhiên phân bố chuẩn đơn vị (chuẩn chuẩn hóa), ký hiệu là z, có μ = 0 và σ = 1.

Công thức chuyển đổi: z = (x - μ)/σ; x là giá trị của biến ngẫu nhiên phân bố chuẩn.

Ý nghĩa của giá trị z: khoảng cách của giá trị quan sát x so với kỳ vọng μ tính theo đơn vị độ lệch chuẩn σ.

Ứng dụng Minitab:

Ví dụ 4: Thời gian trung bình làm một bài tập của một sinh viên tuân theo phân bố chuẩn với μ = 50 phút và σ = 5 phút.

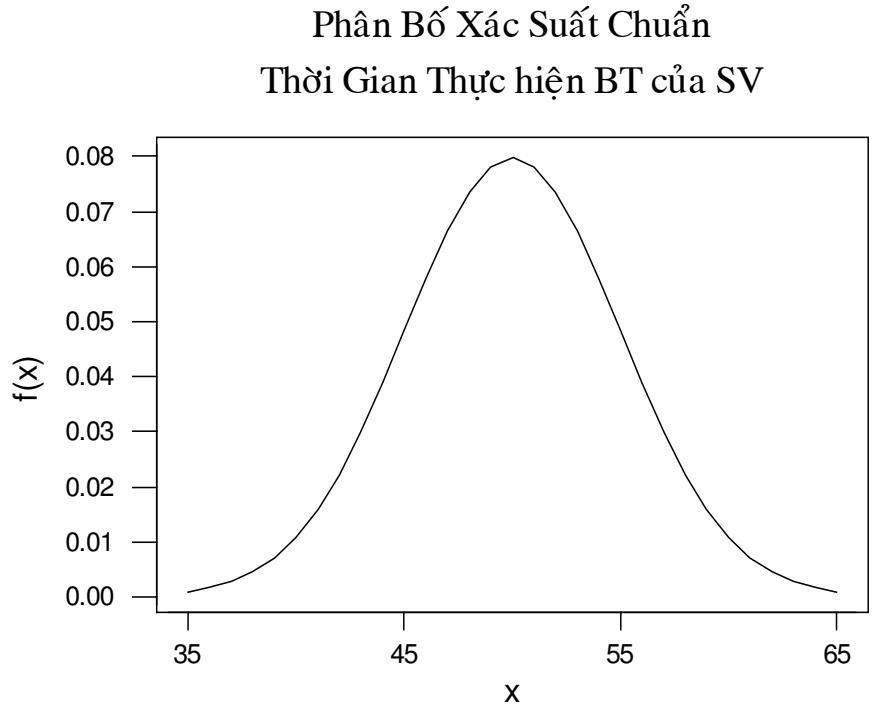
a. Vẽ phân phối chuẩn này.
b. Tìm xác suất để mỗi sinh viên thực hiện một bài tập trong khoảng thời gian từ 40 phút đến 55 phút.
c. Tìm xác suất để mỗi sinh viên thực hiện một bài tập trong khoảng thời gian hơn một giờ.
d. Xác định giá trị x_0 sao cho xác suất để thời gian thực hiện bài tập nhỏ hơn x_0 là 75%.

Lời giải:
a. Đồ thị hàm phân phối
Để vẽ đồ thị hàm phân phối này, ta cần tạo ra các giá trị của x, chỉ cần lấy các giá trị trong khoảng $(\mu \pm 3\sigma)$, nghĩa là (35,65).

Phát giá trị x từ 35 đến 65:
Calc > Make Patterned Data > Simple Set of Numbers
Nhập 'x' vào **Store patterned data in**:
Nhập **From first value:** 35 và **To last value:** 65**.** Sau đó chọn **OK**.

Tính phân bố xác suất của x:
Calc > Probability Distributions > Normal
Nhấp chuột vào **Probability density**:
Nhập 50 vào **Mean** và 5 vào **Standard deviation**
Chọn x vào **Input column**:
Nhập 'f(x)' vào **Optional storage**. Sau đó chọn **OK**.

Vẽ phân bố xác suất:
Graph > Plot
Chọn f(x) vào **Graph 1 Y** và x vào **Graph 1 X**
Nhấp chuột vào **Display**, chọn **Connect**;
Nhấp chuột vào **Annotation**. Nhấp chuột vào **Title**, rồi đánh tên vào. Sau đó chọn **OK**.



b. Để tìm xác suất để mỗi sinh viên thực hiện một bài tập trong khoảng thời gian từ 40 phút đến 55 phút, ta phải sử dụng phân phối xác suất tích lũy:

$$P(40 \le x \le 55) = P(x \le 55) - P(x < 40)$$

Tính P(x ≤55)

Calc > Probability Distributions > Normal

Nhấp chuột vào **Cumulative probability**:

Nhập 50 vào **Mean** và 5 vào **Standard deviation**

Nhấp chuột vào **Input constant**: và nhập 55. Sau đó chọn **OK**.

Cumulative Distribution Function

```
Normal with mean = 50.0000 and standard deviation = 5.00000

         x     P( X <= x )
   55.0000          0.8413
```

Tính P(x < 40)

Calc > Probability Distributions > Normal

Nhấp chuột vào **Cumulative probability**:

Nhập 50 vào **Mean** và 5 vào **Standard deviation**

Nhấp chuột vào **Input constant**: và nhập 40. Sau đó chọn **OK**.

Cumulative Distribution Function

```
Normal with mean = 50.0000 and standard deviation = 5.00000

          x    P( X <= x )
   40.0000         0.0228
```

$P(40 \le x \le 55) = P(x \le 55) - P(x < 40) = 0,8413 - 0,0228 = 0,8185$

c. Để tìm xác suất để mỗi sinh viên thực hiện một bài tập trong khoảng thời gian hơn một giờ, ta tìm $P(x > 60) = 1 - P(x \le 60)$.

Tính $P(x \le 60)$

Calc > Probability Distributions > Normal

Nhấp chuột vào **Cumulative probability**:
Nhập 50 vào **Mean** và 5 vào **Standard deviation**
Nhấp chuột vào **Input constant**: và nhập 60. Sau đó chọn **OK**.

Cumulative Distribution Function

```
Normal with mean = 50.0000 and standard deviation = 5.00000

          x    P( X <= x )
   60.0000         0.9772
```

Xác suất để bất kỳ một sinh viên thực hiện một bài tập với thời gian hơn một giờ là 1 – 0,9772 = 0,0228. Điều này có nghĩa là chỉ có khoảng 2 sinh viên trong số 100 sinh viên cần thời gian hơn một giờ để thực hiện một bài tập.

d. Để tìm thời gian thực hiện một bài tập với xác suất 75%, ta sử dụng tùy chọn phân phối xác suất tích lũy nhằm tìm giá trị x_0 sao cho $P(x \le x_0) = 0,75$

Tính $P(x \le x_0) = 0,75$

Calc > Probability Distributions > Normal

Nhấp chuột vào **Inverse cumulative probability**:
Nhập 50 vào **Mean** và 5 vào **Standard deviation**
Nhấp chuột vào **Input constant**: và nhập 0.75. Sau đó chọn **OK**.

Inverse Cumulative Distribution Function

```
Normal with mean = 50.0000 and standard deviation = 5.00000

P( X <= x )           x
    0.7500       53.3724
```

Có 75% sinh viên có thời gian thực hiện một bài tập nhỏ hơn 53 phút.

Biến Ngẫu Nhiên Tuân Theo Phân Bố Hàm Số Mũ (Exponential Distributon)

Phân bố hàm số mũ dùng để mô tả khoảng thời gian giữa các lần xuất hiện của các biến cố. Ví dụ như khoảng thời gian giữa các lần đến của xe buýt ở các trạm, khoảng thời gian hỏng máy tính PC.

Phân bố hàm số mũ chỉ có một tham số duy nhất là kỳ vọng μ, khoảng thời gian trung bình giữa các lần xuất hiện biến cố. Độ lệch chuẩn của phân bố này, σ, cũng chính bằng μ. Hàm mật độ xác suất của biến ngẫu nhiên x tuân theo phân bố hàm mũ:

$$f(x) = \frac{1}{\mu} e^{-x/\mu}$$

Ứng dụng Minitab trong lý thuyết xếp hàng: Thời gian khách hàng chờ được phục vụ và thời gian phục vụ tuân theo phân phối hàm số mũ.

Ví dụ 5: Giả sử khoảng thời gian giữa những lần khách hàng đến tiệm hớt tóc tuân theo phân phối hàm số mũ với giá trị kỳ vọng là μ = 10 phút.

a. Tìm giá trị kỳ vọng và độ lệch chuẩn của x. Vẽ phân phối xác suất này.
b. Tìm xác suất để khoảng thời gian giữa hai lần khách đến lớn hơn 15 phút.
c. Tìm xác suất để khoảng thời gian giữa hai lần khách đến nằm trong khoảng tin cậy (x±2σ)

Lời giải:

a. Giá trị kỳ vọng và độ lệch chuẩn của x: μ = σ = 10 phút.

Để vẽ đồ thị hàm phân phối này, ta cần tạo ra các giá trị của x, và chỉ cần lấy các giá trị trong khoảng ($\mu \pm 3\sigma$), nghĩa là (-20,40). Vì giá trị của x phải dương nên chỉ nhập từ 0 đến 40.

Phát giá trị x từ 0 đến 40:

Calc > Make Patterned Data > Simple Set of Numbers

Nhập 'x' vào **Store patterned data in**:

Nhập **From first value:** 0 và **To last value:** 40**.** Sau đó chọn **OK**.

Tính phân bố xác suất của x:

Calc > Probability Distributions > Exponential

Nhấp chuột vào **Probability density**:

Nhập 10 vào **Mean**

Chọn x vào **Input column**:

Nhập 'f(x)' vào **Optional storage**. Sau đó chọn **OK**.

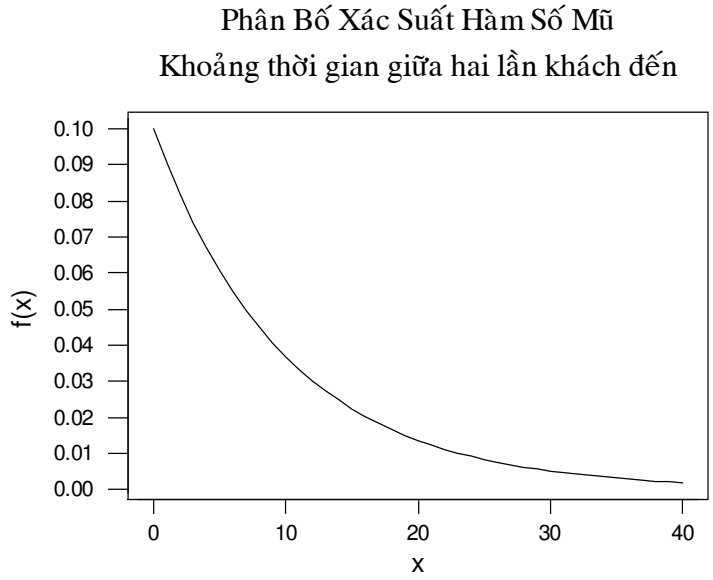
Vẽ phân bố xác suất:

Graph > Plot

Chọn f(x) vào **Graph 1 Y** và x vào **Graph 1 X**

Nhấp chuột vào **Display**, chọn **Connect**;

Nhấp chuột vào **Annotation**. Nhấp chuột vào **Title**, rồi đánh tên vào. Sau đó chọn **OK**.



Đồ thị cho thấy phân bố hàm mũ này lệch nhiều bên phải.

b. Để tìm xác suất để khoảng thời gian giữa hai lần khách đến lớn hơn 15 phút, ta tìm $P(x > 15) = 1 - P(x \leq 15)$.

Tính $P(x \leq 15)$:

Calc > Probability Distributions > Exponential

Nhấp chuột vào **Cumulative probability**:

Nhập 10 vào **Mean**

Nhấp chuột vào **Input constant**: và nhập 15. Sau đó chọn **OK**.

Cumulative Distribution Function

```
Exponential with mean = 10.0000

         x    P( X <= x )
   15.0000         0.7769
```

Xác suất để khoảng thời gian giữa hai lần khách đến lớn hơn 15 phút là: 1 – 0,7769 = 0,2231

c. Để tìm xác suất để khoảng thời gian giữa hai lần khách đến nằm trong khoảng $(x \pm 2\sigma)$, ta tìm $P(\mu - 2\sigma \leq x \leq \mu + 2\sigma) = P(-10 \leq x \leq 30)$. Bởi vì thời gian phải là giá trị dương nên ta chỉ tìm $P(0 \leq x \leq 30)$.

Tính $P(x \leq 30)$:

Calc > Probability Distributions > Exponential

Nhấp chuột vào **Cumulative probability**:

Nhập 10 vào **Mean**

Nhấp chuột vào **Input constant**: và nhập 30. Sau đó chọn **OK**.

Cumulative Distribution Function

```
Exponential with mean = 10.0000

         x    P( X <= x )
   30.0000         0.950
```

Xác suất để thời gian giữa hai lần khách đến nằm trong khoảng $(x \pm 2\sigma)$ là 95%. Kết quả cũng rất gần so với qui tắc thực nghiệm.

3.1.3 Các phân phối xác suất có trong MINITAB

Phân Phối Liên Tục

Phân phối chuẩn Normal	Phân phối Lognormal	Phân phối Gamma
Phân phối Uniform	Phân phối Student t	Phân phối Weibull
Phân phối Cauchy	Phân phối Fisher F	Phân phối Beta
Phân phối Laplace	Phân phối Chi-Square	Phân phối hàm số mũ Exponential
Phân phối Logistic		

Phân Phối Rời Rạc

Phân phối Bernoulli	Phân phối Poisson
Phân phối nhị thức Binomial	Phân phối Integer
Phân phối Hypergeometric	

3.2 PHÂN PHỐI MẪU

Ta thường hay muốn biết thông tin về tập hợp thống kê. Tập hợp thống kê có thể là tất cả những bé gái mười tuổi hiện đang sống ở TP.HCM, ta muốn biết giá trị trung bình huyết áp của tập hợp những bé gái này. Tập hợp thống kê cũng có thể là số bóng đèn được sản xuất tại một nhà máy của xí nghiệp bóng đèn Điện Quang, và ta muốn biết tỷ lệ sản phẩm sản xuất hỏng là bao nhiêu. Trong những trường hợp này ta không có khả năng để nghiên cứu toàn bộ những tập hợp thống kê này (hạn chế về thời gian và tiền bạc). Do đó ta lấy một mẫu nhỏ – chẳng hạn như chỉ cần 200 bé gái hay 80 bóng đèn – và ta dùng thông tin về mẫu này để ước lượng cho những thông tin mà ta muốn biết về tập hợp. Nếu mẫu đại diện được cho tập hợp thống kê thì ước lượng về tập hợp sẽ tương đối chính xác.

Trong ví dụ về những bé gái mười tuổi, ta muốn biết giá trị trung bình huyết áp của tập hợp những bé gái này. Giá trị trung bình này của tập hợp được ký hiệu là μ. Ta có thể dùng giá trị trung bình của mẫu 200 bé gái, $\bar{x}$, để ước lượng cho μ. Ở ví dụ bóng đèn, ta muốn biết tỷ lệ sản xuất bóng đèn hỏng của tập hợp. Tỷ lệ hỏng này của tập hợp được ký hiệu là p. Ta có thể ước lượng p bằng cách sử dụng giá trị tỷ lệ hỏng của mẫu 80 bóng đèn. Tỷ lệ của mẫu thường được ký hiệu là $\hat{p}$.

Nhưng làm thế nào để có được một mẫu đại diện cho tập hợp? Một phương pháp lấy mẫu thường được sử dụng là phương pháp lấy mẫu ngẫu nhiên đơn giản. Một mẫu ngẫu nhiên đơn giản có 2 đặc tính: Mỗi một phần tử của tập hợp thống kê của mẫu phải có cơ hội chọn lấy mẫu bằng nhau và quá trình lấy mẫu phải là ngẫu nhiên. Ta có thể dùng cách phát số ngẫu nhiên của phần mềm Minitab để tiến hành lấy mẫu ngẫu nhiên từ một tập hợp thống kê hữu hạn hay vô hạn.

Một tập hợp thống kê hữu hạn gồm số lượng hữu hạn các phần tử quan sát. Ví dụ nếu ta có thể thu thập được toàn bộ số liệu nồng độ cholesterol của những người mắc bệnh tim và ta muốn có một nghiên cứu sơ bộ về tập hợp này thì ta có thể tiến hành lấu mẫu ngẫu nhiên từ tập hợp hữu hạn những người mắc bệnh tim. Hoặc ta có thể tiến hành lấy mẫu ngẫu nhiên về mức lương khởi điểm từ tập hợp hữu hạn những sinh viên tốt nghiệp năm 2000.

Một tập hợp thống kê vô hạn gồm số lượng vô hạn (không đếm được) các phần tử quan sát. Nếu một quá trình sản xuất liên tục trong cùng một số điều kiện không đổi thì các sản phẩm đầu ra của quá trình lúc bấy giờ được xem như là một tập hợp thống kê vô hạn.

Ứng dụng Minitab:

3.2.1 Lấy Mẫu Ngẫu Nhiên Từ Tập Hợp Thống Kê Hữu Hạn

Ví dụ 6: tập tin Home.mtp chứa dữ liệu gồm giá bán của 200 căn nhà ở Minnesota, Mỹ. Hãy chọn 3 mẫu ngẫu nhiên có kích thước mẫu n = 5 từ tập hợp 200 giá bán nhà này (dùng cột SellingPrice trong file Home.mtp). Tính giá trị kỳ vọng của các mẫu.

Lời giải:

Tạo mẫu ngẫu nhiên 1 với n = 5:

Calc > Random Data > Sample From Columns

Nhập 5 vào **Sample rows from column(s):**

Chọn SellingPrice ở **Columns:**

Nhập ‘Mau1’ vào **Store sample in.** Sau đó chọn **OK**.

Tạo mẫu ngẫu nhiên 2 với n = 5:

Calc > Random Data > Sample From Columns

Nhập 5 vào **Sample rows from column(s):**
Chọn SellingPrice ở **Columns:**
Nhập 'Mau2' vào **Store sample in.** Sau đó chọn **OK**.

Tạo mẫu ngẫu nhiên 3 với n = 5:

Calc > Random Data > Sample From Columns

Nhập 5 vào **Sample rows from column(s):**
Chọn SellingPrice ở **Columns:**
Nhập 'Mau3' vào **Store sample in.** Sau đó chọn **OK**.

Tính giá trị kỳ vọng mẫu 1:

Calc > Calculator

Nhập 'KyVong1' vào **Store results in:**
và Mean(Mau1) vào **Expression**. Sau đó chọn **OK**.

Tính giá trị kỳ vọng mẫu 2:

Calc > Calculator

Nhập 'KyVong2' vào **Store results in:**
và Mean(Mau2) vào **Expression**. Sau đó chọn **OK**.

Tính giá trị kỳ vọng mẫu 3:

Calc > Calculator

Nhập 'KyVong 3' vào **Store results in:**
và Mean(Mau3) vào **Expression**. Sau đó chọn **OK**.

Hiển thị tất cả các kết quả vừa tính toán:

Manip > Display Data

Chọn tất cả các cột từ Mau1 – KyVong3 ở **Display**. Sau đó chọn **OK**.

Data Display

Row	Mau1	Mau2	Mau3	KyVong1	KyVong2	KyVong3
1	54800	80200	90500	91540	88480	92040
2	116300	88700	71600			
3	104700	82400	89900			
4	72200	104200	129200			
5	109700	86900	79000			

Giá trị kỳ vọng của mẫu có thể được sử dụng làm giá trị ước lượng cho giá trị kỳ vọng giá nhà bán của tập hợp 200 căn nhà ở Minnesota, Mỹ. Để ý, mỗi lần phát giá trị nhẫu nhiên cho mẫu đều khác nhau và do đó giá trị kỳ vọng của các mẫu cũng khác nhau.

3.2.2 Lấy Mẫu Ngẫu Nhiên Từ Tập Hợp Thống Kê Vô Hạn

Từ các hàm phân phối xác suất (có 18 dạng phân phối) sẵn có của Minitab, ta có thể phát các giá trị ngẫu nhiên tuân theo phân phối xác định. Phần này chỉ giới thiệu cách lấy mẫu ngẫu nhiên theo phân bố chuẩn.

Ví dụ 7: Giả sử thời gian trung bình một sinh viên thực hiện một bài tập tuân theo phân bố chuẩn với $\mu = 50$ phút và $\sigma = 5$ phút. Hãy tiến hành mô phỏng thực nghiệm này và vẽ đồ thị.

Lời giải: Đầu tiên ta thực hiện 100 lần phát số ngẫu nhiên rồi lần lượt tính các thông số thống kê, vẽ đồ thị. Sau đó thực hiện lại các bước trên với số lần phát ngẫu nhiên là 1000.

Phát 100 giá trị ngẫu nhiên:

Calc > Random Data > Normal

Nhập 100 vào **Generate**;
Nhập 'ThoiGianThucHien' vào **Store in column(s)**:
Nhập **Mean** 50, **Standard deviation** 5. Sau đó chọn **OK**.

Tính các thông số thống kê và vẽ đồ thị cho 100 giá trị ngẫu nhiên:

Stat > Basic Statistics > Display Descriptive Statistics

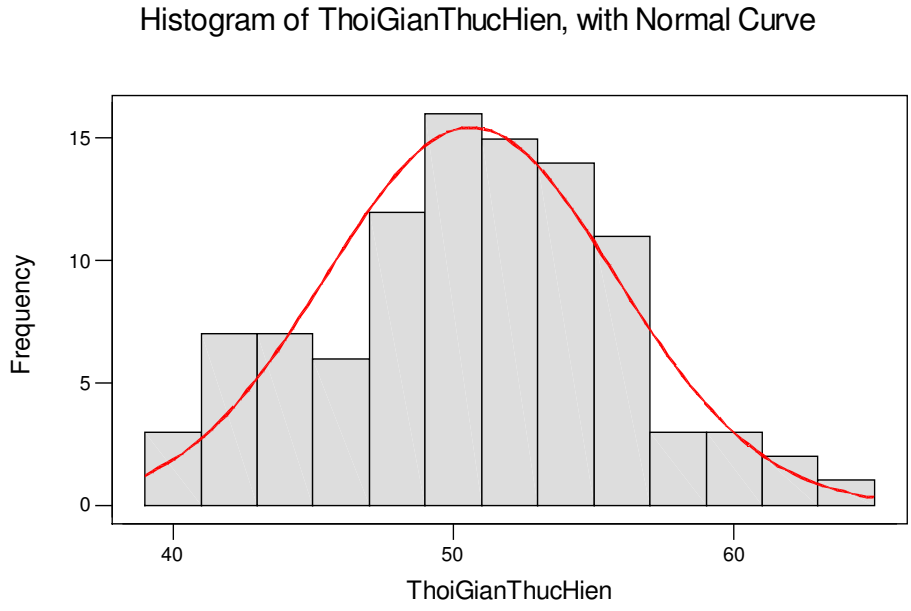
Chọn ThoiGianThucHien vào **Variables:**

Nhấp chuột vào **Graph** và chọn **Histogram of data, with normal curve.** Sau đó chọn **OK**.

Descriptive Statistics: ThoiGianThucHien

Variable	N	Mean	Median	TrMean	StDev	SE Mean
ThoiGian	100	50.630	50.968	50.603	5.167	0.517

Variable	Minimum	Maximum	Q1	Q3
ThoiGian	39.312	63.264	47.675	54.042



Đồ thị Histogram cho thấy phân phối của mẫu 100 giá trị phát ngẫu nhiên xấp xỉ bằng với phân phối chuẩn. Giá trị trung bình mẫu $\bar{x}$ = 50,6 phút gần bằng μ = 50 phút; và độ lệch chuẩn mẫu s = 5,167 gần bằng σ = 5 phút.

Phát 1000 giá trị ngẫu nhiên:

Calc > Random Data > Normal

Nhập 1000 vào **Generate**;

Nhập 'ThoiGianThucHien1' vào **Store in column(s)**:

Nhập **Mean** 50, **Standard deviation** 5. Sau đó chọn **OK**.

Tính các thông số thống kê và vẽ đồ thị cho 1000 giá trị ngẫu nhiên:

Stat > Basic Statistics > Display Descriptive Statistics

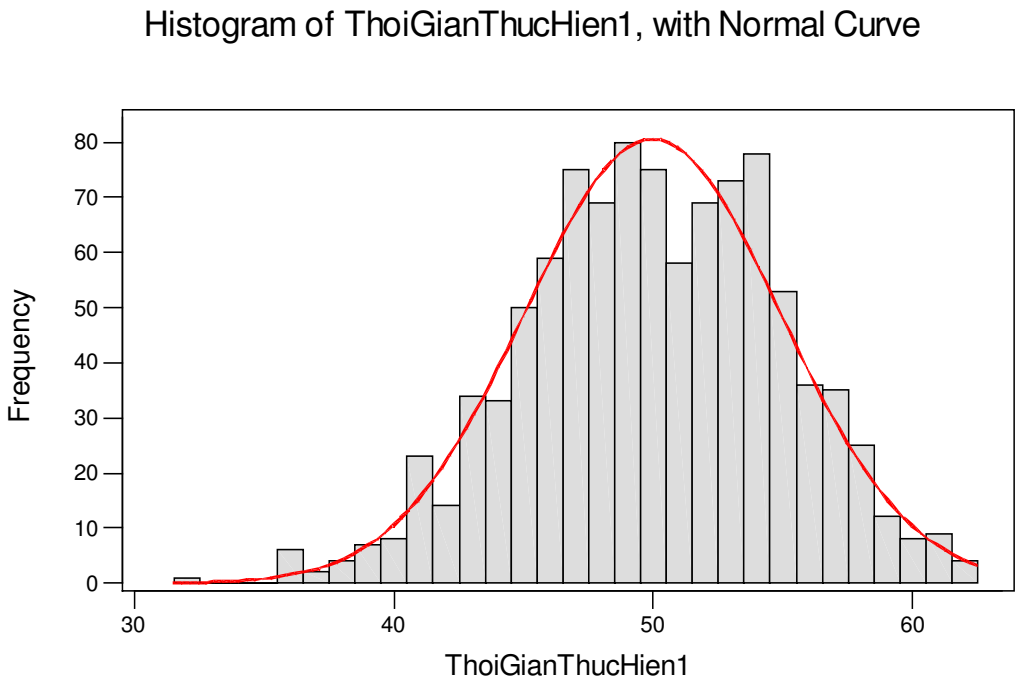
Chọn ThoiGianThucHien1 vào **Variables:**

Nhấp chuột vào **Graph** và chọn **Histogram of data, with normal curve.** *Sau đó chọn* ***OK****.*

Descriptive Statistics: ThoiGianThucHien1

Variable	N	Mean	Median	TrMean	StDev	SE Mean
ThoiGian	1000	49.973	49.901	50.036	4.953	0.157

Variable	Minimum	Maximum	Q1	Q3
ThoiGian	31.747	62.398	46.626	53.659



Đồ thị Histogram cho thấy phân phối của mẫu 1000 giá trị phát ngẫu nhiên tiến gần phân phối chuẩn hơn so với mẫu 100 giá trị ngẫu nhiên. Giá trị trung bình mẫu $\bar{x}$ = 49,97 phút gần như bằng μ = 50 phút; và độ lệch chuẩn mẫu s = 4,953 phút gần như bằng σ = 5 phút.

3.3 KẾT LUẬN THỐNG KÊ TRƯỜNG HỢP ĐƠN MẪU

3.3.1 Ước Lượng Kỳ Vọng Tập Hợp Thống Kê

Một trong những thông số quan trọng nhất của tập hợp thống kê là giá trị kỳ vọng μ. Để ước lượng giá trị kỳ vọng tập hợp μ, ta tiến hành chọn mẫu ngẫu nhiên từ một tập hợp thống kê và tính giá trị kỳ vọng mẫu $\bar{x}$. Giá trị kỳ vọng mẫu $\bar{x}$ chính là ước lượng điểm của giá trị kỳ vọng tập hợp μ. Khi kích thước mẫu càng lớn thì phân bố mẫu $\bar{x}$ càng tập trung gần giá trị kỳ vọng tập hợp μ do đó ước lượng sẽ càng chính

xác. Khi đó giá trị kỳ vọng mẫu $\bar{x}$ bằng với giá trị kỳ vọng tập hợp μ, $\mu_{\bar{x}} = \mu$ và độ lệch chuẩn của kỳ vọng mẫu là $\sigma_{\bar{x}} = \sigma/\sqrt{n}$.

Tuy nhiên một nhược điểm của việc sử dụng ước lượng điểm là nó không cho thấy độ chính xác của ước lượng. Để khắc phục điều này người ta sử dụng ước lượng khoảng.

Khoảng ước lượng của μ với độ tin cậy 100*(1-α)% được định nghĩa như sau:

$$[\bar{x} \pm z_{\alpha/2}\sigma_{\bar{x}}] = [\bar{x} \pm z_{\alpha/2}(\sigma/\sqrt{n})]$$

Trong đó: $\bar{x}$: giá trị kỳ vọng mẫu

σ : độ lệch chuẩn của tập hợp

n : kích thước mẫu

$z_{\alpha/2}$: giá trị z mà ở đó phần diện tích nằm dưới đường cong phân bố chuẩn ở về phía bên phải của nó bằng với α/2.

Trên đây là phần trình bày khoảng ước lượng của kỳ vọng tập hợp thống kê với kích thước mẫu lớn (n ≥ 30) hoặc phân phối của tập hợp xấp xỉ phân phối chuẩn và biết được giá trị σ.

Trong trường hợp kích thước mẫu nhỏ (n < 30) và không biết giá trị σ, khoảng ước lượng của kỳ vọng tập hợp thống kê μ với độ tin cậy 100(1-α)% được định nghĩa như sau:

$$[\bar{x} \pm t_{\alpha/2,n-1}\sigma_{\bar{x}}] = [\bar{x} \pm t_{\alpha/2,n-1}(s/\sqrt{n})]$$

Với $t_{\alpha/2,n-1}$ là giá trị t mà ở đó phần diện tích nằm dưới đường cong phân bố Student với bậc tự do là (n –1) ở về phía bên phải của nó bằng với α/2, s là độ lệch chuẩn của mẫu.

Ứng dụng Minitab:

Tìm Khoảng Ước Lượng Khi Biết Giá Trị σ

Ví dụ 8: Giả sử ta muốn tìm khoảng ước lượng với độ tin cậy 95% cho μ với mẫu có kích thước n = 8 và giả sử biết rằng σ = 0,2. Hãy tính khoảng ước lượng (khoảng tin cậy Confidence Interval – CI).

Lời giải: Nhập vào cột C1 8 giá trị của mẫu, bao gồm: 4,9; 4,7; 5,1; 5,4; 4,7; 5,2; 4,8; và 5,1.

Stat > Basic Statistics > 1-Sample Z

Nhập C1 vào **Variables** và 0.2 vào **Sigma**.
Nhấp chuột vào **Option**, nhập 95 vào **Confidence Level** và chọn not equal ở **Alternative.** Sau đó chọn **OK**.

One-Sample Z: C1

```
The assumed sigma = 0.2

Variable     N      Mean     StDev   SE Mean          95.0% CI
C1           8    4.9875    0.2532    0.0707   ( 4.8489,  5.1261)
```

Với $\bar{x}$ = 4,9875 và σ = 0,2, ta có thể biết rằng μ có giá trị thuộc khoảng (4,8489, 5,1261) với độ tin cậy 95%. Điều này có nghĩa là có 5% trường hợp μ nằm ngoài khoảng ước lượng này.

Tìm Khoảng Ước Lượng Khi Không Biết Giá Trị σ

Ví dụ 9: Một bệnh viện ở Mỹ đã tiến hành lấy mẫu ngẫu nhiên về những người mắc bệnh tim trong suốt mùa hè năm 1999. Hãy tính và diễn giải ý nghĩa của khoảng ước lượng với độ tin cậy là 90% của giá trị kỳ vọng tuổi μ của tập hợp. Sử dụng dữ liệu ở cột Age trong tập tin Heart.mtp.

Lời giải: Vì không biết được độ lệch chuẩn của tập hợp nên ta phải sử dụng lệnh **1-Sample t.** Biểu đồ Histogram cung cấp một bức tranh tổng quát tốt về dữ liệu mẫu.

Stat > Basic Statistics > 1-Sample t

Nhập C1 vào **Variables**

Nhấp chuột vào **Option**, nhập 90 vào **Confidence Level** và chọn not equal ở **Alternative.** Sau đó chọn **OK**.

Nhấp chuột vào **Graph**, chọn **Histogram of data**. Sau đó **OK**.

One-Sample T: Age

```
Variable     N      Mean     StDev    SE Mean          90.0% CI
Age         41     72.78     10.07       1.57   (    70.13,    75.43)
```

Giá trị kỳ vọng của mẫu khoảng 73 tuổi là điểm ước lượng cho giá trị kỳ vọng tuổi của tập hợp. Với 90% độ tin cậy ta có thể kết luận rằng tuổi trung bình của các bệnh nhân mắc bệnh tim trong khoảng (70, 75).

Biểu đồ Histogram cho biết hầu hết các bệnh nhân bệnh tim đều có độ tuổi từ 70 đến 85.

Lưu ý: khi độ tin cậy tăng thì bề rộng khoảng ước lượng cũng tăng.

3.3.2 Kết Luận Về Tỷ Lệ Tập Hợp

Chúng ta đã làm quen với phân phối xác suất nhị thức, được đặc trưng bởi 2 kết xuất: thành công và thất bại. Xác suất để thành công là π và xác suất thất bại là $(1 - \pi)$.

Gọi p, là tỷ lệ số lần thành công trên n phép thử trên một mẫu ngẫu nhiên, là ước lượng điểm. Nếu x là số lần thành công trong n phép thử thì p = x/n. Phân bố mẫu của p sẽ xấp xỉ gần bằng phân phối chuẩn khi kích thước mẫu lớn. (Lưu ý: đối với tỷ lệ tập hợp thì kích thước mẫu được xem là lớn khi và chỉ khi np ≥ 5 và n(1-p) ≥ 5). Giá trị kỳ vọng của phân phối mẫu của p là π và độ lệch chuẩn là $\sqrt{\pi(1-\pi)/n}$. Khoảng ước lượng của π với độ tin cậy 100*(1-α)% được định nghĩa như sau:

$$p \pm z_{\alpha/2}\sqrt{p(1-p)/n}$$

Ứng dụng Minitab:

Ví dụ 10: Xét một mẫu gồm 120 sinh viên trong đó số sinh viên không có máy tính cá nhân PC là x = 21. Hãy xây dựng và diễn giải khoảng tin cậy 95% cho π, tỷ lệ số sinh viên không có máy tính.

Lời giải: Vì np ≥ 5 và n(1-p) ≥ 5 cho nên trong trường hợp này mẫu đủ lớn để phân phối mẫu tuân theo phân phối chuẩn.

Tính khoảng tin cậy 95% cho π:

Stat > Basic Statistic > 1 Proportion

Nhấp chuột vào **Summarized data,** nhập 120 vào **Number of trials:**

Nhập 21 vào **Number of successes:**

Nhấp chuột vào **Option**, nhập 95 vào **Confidence level:**

Nhấp chuột vào **Use test and interval based on normal distribution**. Sau đó chọn **OK.**

Test and CI for One Proportion

```
Test of p = 0.5 vs p not = 0.5

Sample   X      N  Sample p          95.0% CI       Z-Value  P-Value
1       21    120  0.175000  (0.107017, 0.242983)    -7.12    0.000
```

Điểm ước lượng của tỷ lệ tập hợp tất cả những sinh viên không có máy tính cá nhân PC là 0,175. Ta có 95% tin cậy khẳng định rằng tỷ lệ

tập hợp sinh viên không có máy tính cá nhân PC thuộc khoảng (0,11, 0,24).

3.4 KẾT LUẬN THỐNG KÊ TRƯỜNG HỢP HAI MẪU

3.4.1 Ước Lượng Sự Khác Nhau Về Kỳ Vọng Giữa 2 Tập Hợp Thống Kê

Trong phần này, chúng ta nghiên cứu các kết luận thống kê về sự khác nhau giữa kỳ vọng của hai tập hợp thống kê được *lấy mẫu độc lập*. Giả sử rằng μ_1, μ_2 là kỳ vọng của hai tập hợp thống kê đang xét và $\overline{x}_1, \overline{x}_2$ là giá trị trung bình số học của hai tập mẫu tương ứng. Chúng ta có thể cảm nhận được rằng, hiệu số $(\overline{x}_1 - \overline{x}_2)$sẽ có thể được sử dụng để đưa ra những kết luận về sai lệch (μ_1 - μ_2). Do vậy phân bố mẫu của $(\overline{x}_1 - \overline{x}_2)$ có được khi các thí nghiệm lấy mẫu được thực hiện lặp đi lặp lại nhiều lần cần phải được nghiên cứu.

Khoảng tin cậy (1 - α)100% cho (μ_1 - μ_2) được xác định bởi:

$$(\overline{x}_1 - \overline{x}_2) \pm t_{\alpha/2}\sqrt{s_1^2/n_1 + s_2^2/n_2}$$

Trong đó: s_1 và s_2 là độ lệch chuẩn của 2 mẫu
$t_{\alpha/2}$ là giá trị t mà ở đó phần diện tích nằm dưới đường cong phân bố Student với bậc tự do là (n_1 + n_2 –2) ở về phía bên phải của nó bằng với $\alpha/2$

Lưu ý: Bậc tự do trong phân bố Student t rất quan trọng vì nó ảnh hưởng đến giá trị t_α, do đó ảnh hưởng đến khoảng tin cậy.

Trong trường hợp kích thước mẫu nhỏ - n_1 hoặc n_2 hoặc cả 2 giá trị n_1, n_2 nhỏ hơn 30 thì phải giả thiết rằng phương sai của cả 2 tập hợp bằng nhau ($\sigma_1^2 = \sigma_2^2 = \sigma^2$). Điều này thực hiện trong Minitab bằng cách chọn **Assume equal variances** trong **2-Sample t** hoặc tiến hành kiểm định Fisher (kiểm định này không được trình bày trong chương này). Có giả thiết như vậy thì bấy giờ DF của ước lượng mới được xác định đúng theo công thức trên. Trường hợp kích thước mẫu của cả 2 mẫu đều lớn hơn 30 thì giả sử phương sai của cả 2 tập hợp bằng nhau là không cần thiết lắm vì lúc bấy giờ phân bố Student t sẽ gần bằng phân bố chuẩn

Normal z. Hay khi n_1 và n_2 đều lớn hơn 30, giá trị t sẽ không thay đổi lắm nếu ta có hoặc không có chọn **Assume equal variances** trong **2-Sample t**.

Trường hợp hai tập hợp có phương sai bằng hay xấp xỉ bằng nhau thì phương sai s_p^2 của $(\overline{x}_1 - \overline{x}_2)$ được xác định như sau:

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

Ứng dụng Minitab:

Dữ liệu nhập vào hàm **2-Sample t** ở Minitab có thể ở 2 dạng: hoặc nhập vào từ 1 cột dữ liệu gộp (Stacked data column) hoặc nhập vào từ 2 cột dữ liệu.

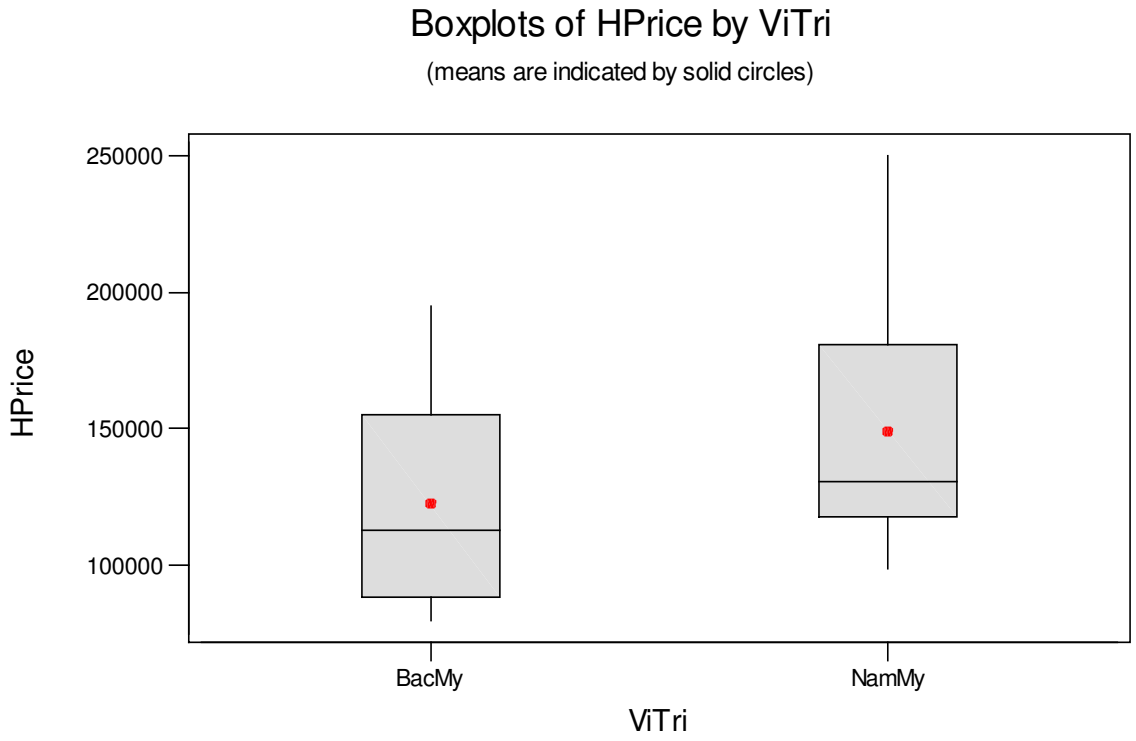
Ví dụ 11: Người ta tiến hành nghiên cứu sự thay đổi giá nhà trong các thành phố thuộc khu vực phía Bắc và Nam của nước Mỹ vào năm 1995 và năm 1996. Dữ liệu được lưu trong tập tin Cityhomes.mtp. Người ta hoài nghi về việc có hay không sự khác biệt về giá nhà trung bình ở 2 khu vực Bắc Mỹ và Nam Mỹ. Hãy ước lượng khoảng tin cậy cho (μ_1 - μ_2) với mức ý nghĩa α = 0,1.

Lời giải: Vì đây là trường hợp dữ liệu gộp, tất cả các giá nhà được nhập vào 1 cột duy nhất. Do đó trước khi tiến hành ước lượng thống kê về sự khác biệt giá nhà ở 2 khu vực Bắc Mỹ và Nam Mỹ ta cần phải nhập vào 1 biến để phân biệt các giá thành 2 khu vực. Trong ví dụ này, ta nhập vào cột ViTri sẽ có một trong hai giá trị BacMy, NamMy tương ứng với vị trí của từng thành phố Mỹ (xem bảng 3.1).

Tính khoảng 90% tin cậy cho (μ_1 - μ_2):
Stat > Basic Statistics > 2-Sample t
Nhấp chuột vào **Sample in one column:**
Chọn HPrice vào **Samples:** và ViTri vào **Subscripts:**
Nhấp chuột vào **Assume equal variances**
Nhấp chuột vào **Options** và nhập giá trị 90% vào **Confidence level:**
Nhấp chuột vào **Graph** và chọn **Boxplots of data**. Sau đó chọn **OK.**

Bảng 3.1

Row	City	ViTri	Row	City	ViTri
1	Albany, NY	BacMy	20	Orange County, CA	NamMy
2	Allentown, PA	BacMy	21	Philadelphia, PA	BacMy
3	Bakersfield, CA	NamMy	22	Portland, OR	NamMy
4	Baltimore	BacMy	23	Raleigh/Durham, NC	BacMy
5	Bergen, NJ	BacMy	24	Riverside, CA	NamMy
6	Boston, MA	BacMy	25	Rochester, NY	BacMy
7	Buffalo, NY	BacMy	26	Sacramento, CA	NamMy
8	Charleston, SC	BacMy	27	San Diego	NamMy
9	Charlotte, NC	BacMy	28	San Francisco, CA	NamMy
10	Fresno, CA	NamMy	29	San Jose, CA	NamMy
11	Greensboro, NC	BacMy	30	Seattle, WA	NamMy
12	Greenville, SC	BacMy	31	Springfield, MA	BacMy
13	Harrisburg, PA	BacMy	32	Stockton, CA	NamMy
14	Hartford, CN	BacMy	33	Syracuse, NY	BacMy
15	Middlesex, NJ	BacMy	34	Tacoma, WA	NamMy
16	Monmouth, NJ	BacMy	35	Vallejo, CA	NamMy
17	New Haven, CN	BacMy	36	Ventura, CA	NamMy
18	New York City	BacMy	37	Washington, D.C.	BacMy
19	Newark, NJ	BacMy			



Two-Sample T-Test and CI: HPrice, ViTri

```
Two-sample T for HPrice

ViTri        N      Mean     StDev   SE Mean
BacMy       23    122498     37196      7756
NamMy       14    149327     46814     12512

Difference = mu (BacMy) - mu (NamMy)
Estimate for difference:  -26829
90% CI for difference: (-50330, -3329)
T-Test of difference = 0 (vs not =): T-Value = -1.93   P-Value =
0.062  DF = 35
Both use Pooled StDev =  41032
```

Đồ thị dạng hộp (boxplot) cho thấy giá nhà ở Bắc Mỹ có khuynh hướng thấp hơn giá nhà ở Nam Mỹ bởi ta thấy giá nhà của mẫu Bắc Mỹ có giá trị trung bình và các phần tư (Q1, Q3) và đồng thời cũng có độ lệch chuẩn thấp hơn so với mẫu Nam Mỹ.

Kết quả thống kê cho thấy: ước lượng điểm cho sự khác biệt về giá nhà trung bình của 2 miền là $149.327 – $122.498 = $26.829. Ta có thể kết luận rằng giá nhà miền Bắc Mỹ thấp hơn Nam Mỹ trung bình dao động trong khoảng $ (3329, 50330) với 90% độ tin cậy. (Lưu ý: vì kích thước mẫu Nam Mỹ nhỏ hơn 30 nên phải chọn Assume equal variances khi gọi lệnh 2-Sample t)

3.4.2 Ước Lượng Sự Khác Nhau Về Kỳ Vọng Giữa 2 Tập Hợp Thống Kê – Trường Hợp Lấy Mẫu So Sánh Từng Cặp

Trong thực tế, khi ước lượng sự khác nhau về kỳ vọng giữa hai tập hợp thống kê, nhiều lúc việc lấy mẫu so sánh từng cặp là cần thiết. Trong những trường hợp như vậy, việc ước lượng dựa trên phân bố Student t như đã trình bày trong phần trên có thể đưa đến kết luận sai lầm. Điều này được minh họa thông qua ví dụ dưới đây.

Ví dụ 12: Giả sử một tổ hợp kinh doanh nhà hàng muốn so sánh doanh thu hàng ngày của hai nhà hàng trong hệ thống của mình. Doanh thu tương ứng trong hai tuần hoạt động ngẫu nhiên được thu thập như sau:

Ngày	Nhà hàng 1 (trăm triệu)	Nhà hàng 2 (trăm triệu)
1	759	678
2	981	933
3	1005	918
4	1449	1302
5	1905	1782
6	2073	1971
7	693	639
8	873	825
9	1074	999
10	1338	1281
11	1932	1827
12	2106	2049

Tương tự như ví dụ 11, ta tiến hành ước lượng dựa theo phân bố Student t như sau:
Dữ liệu cho trong ví dụ này ở 2 cột nên ta phải sử dụng **Sample in different columns** khi chạy lệnh **2-Sample t**.

Tính khoảng 90% tin cậy cho (μ_1 - μ_2):

Stat > Basic Statistics > 2-Sample t

Nhấp chuột vào **Sample in different columns:**
Chọn NhaHang1 vào **First:** và NhaHang2 vào **Second:**
Nhấp chuột vào **Assume equal variances**
Nhấp chuột vào **Options** và nhập giá trị 90% vào **Confidence level:**. Sau đó chọn **OK**.

Two-Sample T-Test and CI: NhaHang1, NhaHang2

```
Two-sample T for NhaHang1 vs NhaHang2

           N      Mean     StDev   SE Mean
NhaHang1  12      1349       530       153
NhaHang2  12      1267       516       149

Difference = mu NhaHang1 - mu NhaHang2
Estimate for difference:  82
90% CI for difference: (-285, 449)
T-Test of difference = 0 (vs not =): T-Value = 0.38  P-Value = 0.705
DF = 22
```

```
Both use Pooled StDev =  523
```

Tính khoảng 99% tin cậy cho (μ_1 - μ_2):

Stat > Basic Statistics > 2-Sample t

Nhấp chuột vào **Sample in different columns:**

Chọn NhaHang1 vào **First:** và NhaHang2 vào **Second:**

Nhấp chuột vào **Assume equal variances**

Nhấp chuột vào **Options** và nhập giá trị 99% vào **Confidence level:**, Sau đó chọn **OK**.

Two-Sample T-Test and CI: NhaHang1, NhaHang2

```
Two-sample T for NhaHang1 vs NhaHang2

            N      Mean     StDev   SE Mean
NhaHang1   12      1349       530       153
NhaHang2   12      1267       516       149

Difference = mu NhaHang1 - mu NhaHang2
Estimate for difference:  82
99% CI for difference: (-520, 684)
T-Test of difference = 0 (vs not =): T-Value = 0.38  P-Value = 0.705
DF = 22
Both use Pooled StDev =  523
```

Thay đổi độ tin cậy từ 90% lên 99% dẫn đến khoảng tin cậy tương ứng thay đổi, bề rộng khoảng tăng từ (-285, 449) đến (-520, 684).

Sự khác nhau về kỳ vọng doanh thu của 2 nhà hàng này cho thấy doanh thu nhà hàng 1 có lúc nhỏ hơn doanh thu nhà hàng 2. Tuy nhiên nếu xem xét kỹ lại các số liệu thống kê về doanh thu ta có thể kết luận ngay là nhà hàng 1 có doanh thu luôn cao hơn nhà hàng 2 (đương nhiên!)?

Nguyên nhân chủ yếu dẫn đến một kết luận thống kê sai lầm như ở trên là do *hai mẫu thử đã không được tiến hành độc lập*. Tuy nhiên việc tiến hành thu thập hai mẫu thử độc lập trong trường hợp này sẽ là không thực tế vì có thể thấy được rằng doanh thu biến động rất nhiều và có quy luật (patterns), do đó giá trị s_p^2 sẽ rất lớn và tiêu chuẩn kiểm định dựa vào phân bố Student sẽ không thể phát hiện được sai lệch kỳ vọng. Nếu tiến hành lấy mẫu độc lập trong trường hợp này thì số mẫu phải khá lớn và tiêu chuẩn kiểm định dựa trên phân bố chuẩn có thể được dùng.

Vấn đề đặt ra ở đây là có cách nào kiểm định được sự sai lệch kỳ vọng nhưng vẫn sử dụng mẫu nhỏ hay không? Để thực hiện điều này, phương pháp ước lượng dựa trên các mẫu được lấy song song (so sánh theo cặp) có thể được áp dụng. Theo phương pháp ước lượng này, tập hợp thống kê biểu diễn cho sai lệch của các cặp giá trị tương ứng sẽ được xem xét. Lưu ý rằng kỳ vọng của tập hợp thống kê này là $\mu_D = (\mu_1 - \mu_2)$ và $\overline{x}_D = (\overline{x}_1 - \overline{x}_2)$ sẽ là một ước lượng không lệch của μ_D. Bài toán như vậy đã được chuyển về trường hợp kiểm định đơn mẫu và việc ước lượng được thực hiện tương tự như đã trình bày ở phần trên.

Lúc bấy giờ ví dụ 12 được giải quyết theo cách giải ở ví dụ 9 bằng cách tạo thêm 1 cột dữ liệu mới C3 với tên là *Sailech*. *Sailech* chính là độ chênh lệch giữa hai doanh thu của hai nhà hàng, với giá trị trong cột C3 bằng giá trị cột C1 trừ đi giá trị cột C2. Dựa vào các giá trị sai lệch từ hai mẫu thử ta sẽ tính được x_D và s_D.

Ví dụ 13: Giải lại ví dụ 12 trên theo cách vừa trình bày. Kết quả như sau:

Lời giải:

Tính sai lệch giữa 2 mẫu:

Calc > Calculator

Nhập Sailech vào **Store result in variable:**

Nhập (C1 – C2) vào **Expression**. Sau đó chọn **OK.**

Stat > Basic Statistics > 1-Sample t

Nhập C3 vào **Variables**

Nhấp chuột vào **Option**, nhập 95% vào **Confidence Level** và chọn not equal ở **Alternative**. Sau đó chọn **OK**.

(Lưu ý: vì không biết được độ lệch chuẩn của tập hợp và kích thước mẫu nhỏ nên ta phải sử dụng lệnh **1-Sample t**)

One-Sample T: Sailech

```
Variable      N      Mean     StDev   SE Mean          95.0% CI
Sailech      12     82.00     31.99      9.23  (   61.68,  102.32)
```

Từ khoảng ước lượng, thấy rằng nói chung doanh thu nhà hàng 1 lớn hơn doanh thu nhà hàng 2

Ngoài cách giải quyết bài toán như trên, ta có thể sử dụng hàm **Paired t** trong Minitab để tính ước lượng sự khác nhau về kỳ vọng giữa hai tập hợp thống kê hoặc dùng đồ thị để vẽ từng cặp giá trị quan sát trên cùng một đồ thị (tuy nhiên cách dùng đồ thị này chỉ phù hợp khi tất cả các cặp giá trị đều có sai lệch dương hay nói cách khác mỗi giá trị quan sát ở mẫu này lớn hơn giá trị quan sát tương ứng ở mẫu kia).

Ví dụ 14: Giải ví dụ 12 trên theo cách lấy mẫu so sánh từng cặp (sử dụng hàm **Paired t** trong Minitab) với độ tin cậy 95%.

Lời giải:

Stat > Basic Statistics > Paired t

Chọn C1 vào **First Sample:** và C2 vào **Second Sample:**

Nhấp chuột vào **Option**, sau đó nhập 95% vào **Confidence level:** và chọn not equal tại Alternative:

Paired T-Test and CI: C1, C2

```
Paired T for C1 - C2

                 N      Mean     StDev   SE Mean
NhaHang1        12      1349       530       153
NhaHang2        12      1267       516       149
Difference      12     82.00     31.99      9.23

95% CI for mean difference: (61.68, 102.32)
T-Test of mean difference = 0 (vs not = 0): T-Value = 8.88  P-Value
= 0.000
```

Rõ ràng là kết quả của doanh thu nhà hàng 1 lớn hơn doanh thu nhà hàng 2.

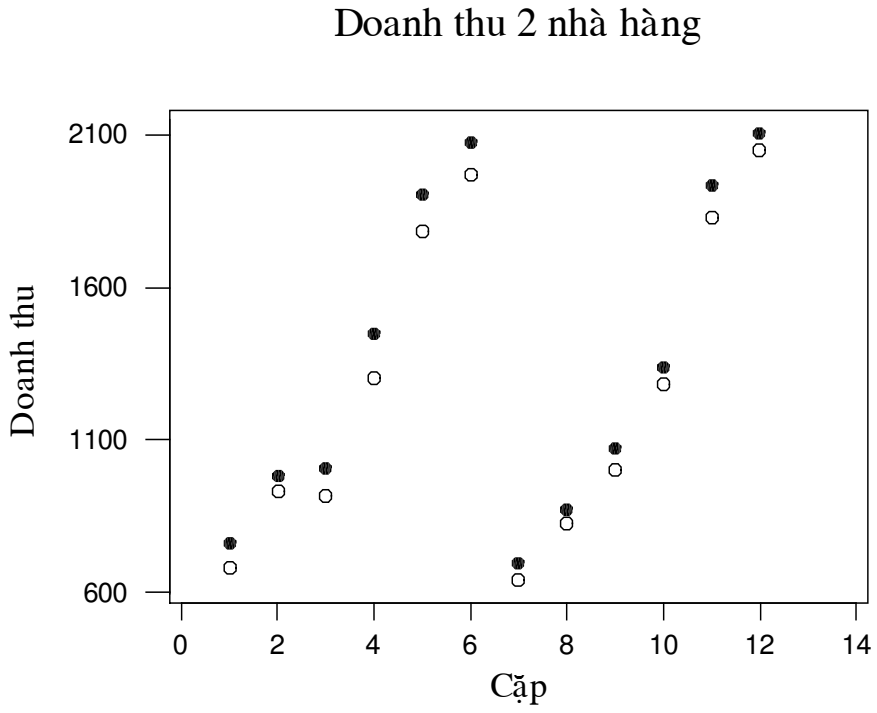
Kết quả của cách sử dụng hàm **Paired t** cho ta kết quả giống như cách giải ở ví dụ 13.

Ví dụ 15: Giải ví dụ 12 trên theo cách vẽ đồ thị.

Lời giải: Chèn 1 cột mới vào trước cột C1 (Doanh thu 1) với tên là Cap. Lần lượt đánh số thứ tự từ trên xuống cho các cặp giá trị tương ứng.

Graph > Plot

Chọn NhaHang1 vào **Graph 1Y** và Cap vào **Graph 1X**
Chọn NhaHang2 vào **Graph 2Y** và Cap vào **Graph 2X**
Nhấp chuột vào **Edit Attributes**, rồi nhấp chuột vào **Type**, chọn Solid Circle cho **Graph 1** và Circle cho **Graph 2**
Nhấp chuột vào **Annotation** và **Title**, nhập tiêu đề Doanh thu 2 nha hang vào.
Nhấp chuột vào **Frame** và **Multiple Graphs**, chọn **Overlay graphs on same page**
Nhấp chuột vào **Frame** và **Axis**, nhập Doanh thu vào Label 2. Sau đó chọn **OK**.



Đồ thị cho thấy doanh thu nhà hàng 2, được biểu diễn bằng vòng tròn rỗng, tại mỗi cặp giá trị quan sát luôn thấp hơn so với nhà hàng 1. (Trong trường hợp nhà hàng 2 có 5 ngày có doanh thu/ngày cao hơn nhà hàng 1 và 7 ngày có doanh thu/ngày thấp hơn nhà hàng 1 thì phương pháp đồ thị này có sử dụng được không?)

3.4.3 Ước Lượng Sự Khác Nhau Về Kỳ Vọng Giữa 2 Tập Hợp Thống Kê – Trường Hợp Lấy Mẫu Nhị Thức Độc Lập

Giả sử tỷ lệ thành công của hai tập hợp thống kê lần lượt là π_1 và π_2. Ước lượng điểm cho sự khác biệt (π_1 - π_2) của 2 tỷ lệ thành công của 2 tập hợp này là sự khác biệt ($p_1 - p_2$) từ 2 tỷ lệ của 2 mẫu nhị thức độc lập tương ứng với 2 tập hợp. Trong đó $p_1 = x_1/n_1$, và $p_2 = x_2/n_2$ với x_1, x_2 lần lượt là số lần thành công trên 2 mẫu nhị thức độc lập có kích thước là n_1 và n_2. Nếu kích thước mẫu đủ lớn thì phân bố mẫu của ($p_1 - p_2$) xấp xỉ phân bố chuẩn.

Với độ tin cậy 100(1 - α)%, khoảng ước lượng của (π_1 - π_2) được xác định bằng:

$$(p_1 - p_2) \pm z_{\alpha/2}\sqrt{(\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2})}$$

Với $q_1 = 1 - p_1$ và $q_2 = 1 - p_2$

Ứng dụng Minitab:

Dữ liệu nhập vào hàm **2 Proportions** trong Minitab có thể ở 3 dạng: hoặc nhập vào từ 1 cột dữ liệu gộp (Stacked data column) hoặc nhập vào từ 2 cột dữ liệu hoặc nhập các giá trị n_1, p_1, và n_2, p_2.

Ví dụ 16: Hội phụ nữ TP.HCM tiến hành khảo sát so sánh tỷ lệ số giảng viên nữ trong các trường ĐH khu vực TP.HCM năm 2001. Tiến hành lấy mẫu ngẫu nhiên ở 2 trường ĐH Bách Khoa TPHCM và Sư Phạm TPHCM. Trong mẫu ngẫu nhiên của trường ĐH Sư Phạm TPHCM có 36/80 giảng viên là nữ, trong khi đó tỷ lệ này ở mẫu trường ĐH Bách Khoa TPHCM là 15/70. Dựa vào hai mẫu ngẫu nhiên này ta có thể kết luận với độ tin cậy 90% rằng có sư khác biệt về tỷ lệ giảng viên nữ ở 2 trường ĐH này không?

Lời giải:

Tính khoảng ước lượng của (π_1 - π_2):

Stat > Basic Statistics > 2 Proportions

Nhấp chuột vào **Summarized data:**

Ở **First sample:** nhập 80 Trials: và 36 Successes:

Ở **Second sample:** nhập 70 Trials: và 15 Successes:

Nhấp chuột vào **Options** và nhập 90% vào **Confidence level:**

Sau đó chọn **OK**.

Test and CI for Two Proportions

```
Sample      X      N  Sample p
1          36     80  0.450000
2          15     70  0.214286

Estimate for p(1) - p(2):  0.235714
90% CI for p(1) - p(2):  (0.113740, 0.357689)
Test for p(1) - p(2) = 0 (vs not = 0):  Z = 3.18  P-Value = 0.001
```

Khoảng tin cậy của (π_1 - π_2) là 0,11÷0,36. Với độ tin cậy 90%, ta có thể kết luận rằng trường ĐH Sư Phạm TPHCM có tỷ lệ giảng viên nữ nhiều hơn từ 11% đến 36% so với trường ĐH Bách Khoa TPHCM.

BÀI TẬP

3.1. Gọi x là số khuyết tật của một cái áo do nhà may A sản xuất. Giả sử biến ngẫu nhiên x có phân phối như ở bảng dưới. Dùng Minitab để trình bày phân phối xác suất dạng đồ thị. Tính giá trị trung bình (mean), độ lệch chuẩn (standard deviation) của phân phối này.

Số khuyết tật X	Xác suất P(x)
0	0,85
1	0,10
2	0,05

3.2. Cửa hàng sách New Edition thường cần phải đặt các loại sách mới nhất. Người quản lý dùng phân phối xác suất số lượng sách cần trong tuần như sau:

Nhu cầu sách (cuốn/tuần) X	Xác suất P(x)
5	0,1
6	0,3
7	0,3
8	0,2
9	0,1

a. Dùng Minitab để trình bày phân phối xác suất dưới dạng đồ thị.
b. Tính giá trị trung bình (mean), phương sai (variance), và độ lệch chuẩn (standard deviation) của phân phối này.

3.3. Giả sử x là biến ngẫu nhiên nhị thức với n=10.
a. Với π = 0,01; 0,1; 0,5; 0,8; và 0,9; ứng với mỗi giá trị tìm phân phối xác suất nhị phân và đồ thị của mỗi phân phối.
b. Hình dạng của đồ thị sẽ như thế nào khi π tăng?
c. Với giá trị nào của π, phân phối xác suất sẽ đối xứng và gần đối xứng?

3.4. Giả sử x là biến ngẫu nhiên nhị thức với $\pi = 0,02$.
a. Tìm phân phối xác suất và vẽ đồ thị của mỗi phân phối với ứng vớ các giá trị n = 2, 5, 10, 20 và 40.
b. Khi n tăng, hình dạng của đồ thị sẽ thay đổi như thế nào?

3.5. Giả sử x là biến ngẫu nhiên nhị thức với $\pi = 0,02$. Với n = 40, tính xác suất sao cho giá trị x nằm trong khoảng một độ lệch chuẩn ($\mu \pm \sigma$), hai độ lệch chuẩn ($\mu \pm 2\sigma$). So sánh những xác suất này với Quy Tắc Kinh Nghiệm.

3.6. Một doanh nghiệp lấy mẫu 200 linh kiện trong chuyến hàng. Họ chỉ chấp nhận chuyến hàng nếu có ít hơn hoặc bằng 10 phế phẩm trong mẫu, và từ chối nếu số phế phẩm lớn hơn 10.
a. Tìm xác suất chấp nhận nếu chuyến hàng có 3% phế phẩm (π=0,03).
b. Tìm xác suất từ chối nếu chuyến hàng có 10% phế phẩm.

3.7. Một công ty quảng cáo được thuê giới thiệu một sản phẩm mới. Công ty này cho rằng sau chiến dịch quảng cáo, có 40% khách hàng sẽ quen với sản phẩm. Gọi x là số khách hàng quen với sản phẩm trong mẫu ngẫu nhiên 25 khách hàng.
a. Tính giá trị trung bình và độ lệch chuẩn của x, vẽ biểu đồ phân bố xác suất.
b. Xác suất có hơn một nửa số khách hàng ($\geq$13) trong mẫu quen với sản phẩm là bao nhiêu.

3.8. Gọi x là số người hơn 65 tuổi mà vẫn làm việc trong mẫu ngẫu nhiên 50 người được chọn từ tất cả những người tuổi trên 65. Theo những báo cáo đáng tin cậy mới đây, khoảng 30% số người trên 65 ở trên thế giới vẫn còn làm việc.
a. Tính giá trị trung bình và độ lệch chuẩn của x, vẽ biểu đồ phân bố xác suất.
b. Xác định các khoảng giá trị $\mu \pm \sigma$, $\mu \pm 2\sigma$, và $\mu \pm 3\sigma$, Tìm xác suất để x rơi vào trong từng khoảng trên. So sánh xác suất tìm được với Quy Tắc Kinh Nghiệm.

3.9. Hiệp hội Internet mới đây báo cáo rằng 82,3% số trang chủ (homepage) sử dụng tiếng Anh. Giả sử chọn ngẫu nhiên 100 trang chủ, gọi x là số trang chủ dùng tiếng Anh.

a. Có bao nhiêu trang chủ dùng ngôn ngữ không phải là tiếng Anh?

b. Tìm xác suất để có thể thấy hơn 10 trang không phải là tiếng Anh, xác suất không thấy một trang web không phải là tiếng Anh là bao nhiêu.

3.10. Nhân viên kiểm tra chất lượng tại một dây chuyền sản xuất xe nhận thấy rằng số khuyết tật sơn không nhìn thấy rõ của một chiếc xe có phân phối Poisson với giá trị trung bình là 2 khuyết tật trên một chiếc xe. Gọi x là số khuyết tật sơn trên 1 chiếc xe

a. Tình giá trị trung bình và độ lệch chuẩn của x

b. Vẽ đồ thị phân bố xác suất.

c. Tìm xác suất không tìm thấy khuyết tật sơn nào? Xác suất thấy số khuyết tật lớn hơn hay bằng 5?

3.11. Giả sử số khuyết tật trên 1 cái áo sơ mi của một nhà may tuân theo phân phối Poisson với giá trị trung bình $\mu=0,2$. Gọi x là số khuyết tật trên 1 cái áo

a. Tìm giá trị trung bình và độ lệch chuẩn của x. Vẽ đồ thị phân bố xác suất.

b. Tìm xác suất để một chiếc áo không có khuyết tật.

c. Nếu chiếc áo có nhiều hơn hoặc bằng 1 khuyết tật, nhà may xếp chúng vào loại hai. Xác định tỷ lệ tất cả các áo được xem là loại 2.

3.12. Số khách hàng đến ngân hàng có phân phối xác suất Poisson với giá trị trung bình là một khách hàng đến trong một phút.

a. Dùng Minitab để trình bày phân phối xác suất bằng bảng và đồ thị. Tìm phương sai và độ lệch chuẩn.

b. Tìm xác suất có hơn hai người đến quầy ngân hàng trong một phút.

c. Ngân hàng muốn có đủ khả năng phục vụ để xác suất phục vụ tất cả các khàng đến trong 1 phút dưới 0,99%. Ngân hàng có thể phục vụ bao nhiêu khách hàng?

3.13. Thống kê từ Cơ Quan Thuế Vụ Hoa Kỳ (IRS) cho thấy rằng cơ hội kiểm toán khai thuế đối với những người có thu nhập dưới $50000 vào khoảng 0,009. Gọi x là số lần kiểm toán khai thuế trong một mẫu ngẫu nhiên của 100 người nộp thuế.

a. Vẽ đồ thị phân bố xác suất nhị thức của x
b. Tìm xác suất không có người nào bị kiểm toán, giả thiết rằng thông báo của IRS là chính xác.

3.14. Giả sử số lỗi bảng tính trong một giờ của người kế toán viên tuân theo phân bố Poisson với trị trung bình $\mu=1,2$. Gọi x là số lỗi bảng tính trong 1 giờ.

a. Tìm giá trị trung bình và độ lệch chuẩn của x. Vẽ đồ thị phân bố xác suất.
b. Tìm xác suất không có một lỗi nào xuất hiện trong một giờ.
c. Tìm xác suất mà người kế toán viên mắc phải 1 đến 4 lỗi trong một giờ.

3.15. Tính và vẽ hàm mật độ xác suất cho một biến ngẫu nhiên phân phối chuẩn với $\mu=100$ và $\sigma=10$. Dùng các giá trị x trong khoảng 3σ trên và dưới giá trị trung bình.

3.16. Công ty Sữa Diary dùng máy để rót sữa vào các bình sữa 32 ounces. Trọng lượng tịnh của sữa rót vào bình có phân phối chuẩn với đệ lệch chuẩn là 0,5 ounces. Thao tác điều chỉnh máy sẽ khống chế được lượng sữa trung bình rót vào bình.

a. Giả sử máy được điều chỉnh để rót vào bình lượng sữa có giá trị trung bình 32,2 ounces. Tỷ lệ số bình được rót ít hơn 32 ounces là bao nhiêu?
b. Công ty phải đáp ứng một số tiêu chuẩn và giả sử rằng Dairy cần phải tuân thủ quy định không quá 1% số chai có ít hơn 32 ounces. Cần phải điều chỉnh máy khống chế lượng sữa rót ít nhất là bao nhiêu để đáp ứng yêu cầu?

3.17. Nhân viên kiểm tra chất lượng tin rằng máy dùng để rót vào các chai nước khoáng được cài đặt rót bình quân 12 ounces. Giả thiết

rằng lượng nước rót vào mỗi bình này (gọi là x) có phân phối chuẩn với độ lệch chuẩn là 0,5 ounces.

a. Vẽ đồ thị phân bố xác suất của x.
b. Giả sử các bình có lượng nước ít hơn 11,9 ounces sẽ phải bơm lại. Tính tỷ lệ các bình cần phải bơm lại.

3.18. Giả sử thời gian khởi động máy tính (gọi là x) tuân theo phân phối chuẩn với giá trị trung bình là 70 giây, độ lệch chuẩn là 15 giây.

a. Vẽ đồ thị phân phối xác suất của x.
b. Tìm xác suất để thời gian khởi động nằm trong khoảng 2 độ lệch chuẩn so với trung bình.
c. Xác suất để thời gian khởi động ít hơn 1 phút là bao nhiêu?

3.19. Dữ liệu sau đây ghi nhận thời gian đến (tính theo phút) giữa mỗi người khách tại Thảo Cầm Viên Sài Gòn. (File: ZooArrival.mtp)

0,2	8,0	9,0	0,4	13,2	7,2	0,7	1,2	0,4
10,5	1,4	6,4	0,3	2,4	6,6	18,4	0,5	7,1
5,7	0,1	14,9	0,3	4,4	0,9	13,5	7,2	3,8
5,3	0,6	10,8	1,2	0,5	2,2	18,2	15,5	1,3
4,3	1,9	0,5	3,2	5,8	2,1	8,3	3,5	3,2
5,5	0,5	0,4	1,9	1,9	1,0	1,2	2,2	5,0
0,3	8,1	3,1	11,4	0,2	0,7	0,5	0,7	1,0
0,3	1,1							

a. Vẽ đồ thị dữ liệu. Dữ liệu có tuân theo phân phối hàm mũ hay không?
b. Nếu dữ liệu tuân theo phân phối hàm mũ, ước lượng μ và σ.

3.20. Giả sử khoảng thời gian một người phải xếp hàng để được phục vụ tại một cửa hàng thức ăn nhanh (fast food) tuân theo phân phối chuẩn với $\mu = 1,5$ phút.

a. Vẽ đồ thị phân phối hàm mũ. Tìm giá trị trung bình và độ lệch chuẩn.
b. Tìm xác suất khách hàng phải đợi trên 3 phút trước khi được phục vụ.

c. Tìm khoảng thời gian x sao cho 80% tất cả khách hàng đợi lâu hơn x.

3.21. Mô phỏng 50 mẫu với cỡ mẫu n = 15 từ một phân phối chuẩn với $\mu = 300$ và $\sigma = 50$. Hãy xác định khoảng tin cậy 90% cho 50 mẫu này. Có bao nhiêu khoảng tin cậy chứa giá trị $\mu = 100$? Bạn kỳ vọng bao nhiêu khoảng chứa giá trị $\mu = 300$?

3.22. Một cửa hàng lớn có một máy đóng gói để cân và đóng gói gà tây. Bộ phận điều chỉnh máy cho phép người điều khiển đóng gói với nhiều trọng lượng khác nhau. Giả sử cửa hàng điều chỉnh để máy đóng gói 1kg/bao. Tất cả các bao đều không chứa chính xác 1kg bởi vì sự biến động ngẫu nhiên của quá trình cân, hoặc là do điều chỉnh không chính xác. Để kiểm tra máy điều chỉnh chính xác hay không, người ta thu thập ngẫu nhiên mẫu gồm 40 bao như sau (file: Weights.mtp):

Trọng lượng bao

0,97	1,05	1,09	1,05	1,05	1,05	0,94	0,98	0,96	1,05
1,08	1,12	1,00	1,00	1,07	1,04	1,03	0,96	1,03	1,05
0,98	1,02	1,00	1,06	1,03	1,02	1,05	1,07	0,98	1,07
1,05	1,09	1,09	1,01	0,98	1,00	1,01	1,06	1,10	1,09

Hãy xác định khoảng tin cậy 95% để dự báo trọng lượng trung bình của toàn bộ các bao. Tóm tắt dữ liệu bằng đồ thị, diễn giải khoảng tin cậy.

3.23. Từ tập dữ liệu dưới đây (file: Homes.mtp), biến D là số ngày một căn nhà còn xuất hiện trên thị trường trước khi được bán. Hãy tóm tắt bằng số và đồ thị của phân phối số ngày tất cả các căn nhà trong tập hợp. (A: khu vực, D: số ngày)

A	D	A	D	A	D	A	D	A	D	A	D	A	D	A	D
1	606	1	16	1	79	1	107	1	74	2	125	3	51	3	22
1	93	1	13	1	343	1	119	1	169	2	135	3	408	3	94
1	22	1	136	1	92	1	10	1	14	2	142	3	61	3	219
1	24	1	16	1	45	1	89	1	85	2	50	3	37	3	105

1	150	1	5	1	91	1	10	1	29	2	88	3	77	3	323
1	239	1	66	1	18	1	47	2	180	2	11	3	84	3	14
1	42	1	30	1	2	1	354	2	6	2	41	3	135	3	233
1	281	1	140	1	256	1	127	2	240	2	45	3	139	3	135
1	28	1	14	3	127	1	43	2	141	2	35	3	57	3	25
1	161	1	44	1	15	1	9	2	189	2	282	3	134	3	142
1	17	1	13	1	74	1	11	2	133	2	98	3	343	3	157
1	2	1	17	1	4	1	143	2	39	2	17	3	59	3	160
1	407	1	10	1	143	1	55	2	76	2	44	3	86	3	20
1	7	1	147	1	1	1	38	2	32	2	39	3	73	3	69
1	26	1	166	1	40	1	4	2	11	2	232	3	44	3	24
1	62	1	90	1	73	1	454	2	93	2	201	3	82	3	84
1	14	1	100	1	41	1	19	2	43	2	1	3	118	3	218
1	363	1	32	1	26	1	*	2	138	2	111	3	441	3	261
1	76	1	164	1	21	1	115	2	124	2	2	3	454	3	175
1	153	1	16	1	48	1	47	2	50	2	79	3	12	3	25
1	95	1	419	1	32	1	4	2	294	2	148	3	501	3	102
1	*	1	178	1	45	1	130	2	234	2	57	3	101	3	220
1	114	1	33	1	48	1	1	2	24	2	50	3	108	3	66
1	85	1	92	1	140	1	46	2	80	2	63	3	29	3	30
1	67	1	72	1	23	1	82	2	74	2	39	3	18	3	5

3.24. Một công ty dịch vụ quan tâm đến thời gian khách hàng phải đợi trước khi nhận được hỗ trợ theo yêu cầu. Công ty lấy ngẫu nhiên 40 khách hàng và ghi nhận thời gian đợi theo phút với kết quả dưới đây (file: WaitingTime.mtp). Xác định khoảng tin cậy 95% để ước lượng thời gian trung bình khách hàng phải đợi. Tóm tắt dữ liệu bằng đồ thị, đồ thị và diễn dịch khoảng tin cậy.

9	16	19	21	11	16	19	22	12	17
19	22	13	17	20	22	13	17	20	24
14	17	20	24	14	18	21	25	15	18
21	27	15	18	21	29	16	19	21	38

3.25. Nghiên cứu mới đây của Khoa Toán cho thấy rằng tỷ lệ bỏ môn học Đại số tại trường đại học năm vừa rồi là 40%. Khoa đang thử một kế hoạch mới để chỉ định sinh viên vào các lớp toán nhằm giảm tỷ lệ bỏ lớp trong năm học này. Một mẫu ngẫu nhiên

gồm 100 sinh viên học môn đại số trong năm nay được chọn và ghi nhận tỷ lệ bỏ lớp.

a. Vẽ đồ thị phân bố mẫu của p, giả thiết rằng tỷ lệ bỏ lớp năm này không thay đổi so với năm trước.
b. Trong mẫu 100 sinh viên, tỷ lệ bỏ lớp là p = 0,35. Liệu mẫu này có cho đủ bằng chứng rằng tỷ lệ bỏ lớp môn Đại số này đã giảm so với năm trước? Dùng mức ý nghĩa $\alpha = 0,1$.
c. Bạn có đề nghị khoa toán tiếp tục dùng phương pháp này để giảm tỷ lệ bỏ lớp, hay tìm một giải pháp khác để giải quyết vấn đề này?

3.26. Giám đốc của một công ty lớn quan tâm đến số nhân viên được đào tạo tốt mà có thể nghỉ hưu sớm. Nếu quá nhiều nhân viên nghỉ hưu cùng một lúc, việc thay người rất khó khăn. Bảng sau cho biết số năm làm việc của một mẫu ngẫu nhiên của 151 nhân viên (file: ServiceYears.mtp).

Số năm làm việc

13	16	25	3	27	7	7	2	3	16	7	26	1	8
6	27	6	6	8	23	3	21	4	12	0	9	3	32
27	5	23	9	9	6	9	7	9	13	1	15	20	9
2	1	13	18	10	27	26	4	27	9	10	7	7	13
27	2	7	23	26	16	5	23	6	9	30	4	5	18
4	4	0	10	10	7	2	27	26	3	29	29	7	1
19	19	5	5	10	28	21	20	23	8	3	17	17	26
30	14	17	6	14	20	0	27	22	28	20	0	8	13
19	1	2	18	26	9	3	21	8	17	1	29	21	30
7	6	18	2	10	6	26	9	22	13	7	8	28	44
26	28	16	29	2	9	17	2	8	23	39			

a. Bình luận kết quả hình dạng phân bố của các năm làm việc ở bảng trên.
b. Công ty có chính sách nghĩ hưu là sau 25 làm việc ở công ty. Hãy ước lượng tỷ lệ nhân viên nghĩ hưu ở công ty bằng cách sử dụng khoảng tin cậy 90%.

3.27. Trong suốt mùa hè năm 1999, trạm y tế đã tiến hành đo lường chất lượng chăm sóc sức khỏe dân địa phương. Chất lượng được phân làm 3 loại: Tốt: thời gian chữa trị dưới 15 phút so với lịch; Chấp nhận được: từ 15-30 phút; xấu: trên 30 phút. Một mẫu ngẫu nhiên gồm 200 ca, trong đó 128 ca tốt, 46 ca chấp nhận được và 26 ca xấu. Dùng 90% khoảng tin cậy để dự báo tỉ lệ của tất cả các ca cho chất lượng tốt.

3.28. Nhiều trường ĐH trên thế giới đã thực hiện nhiều chương trình nhằm duy trì số lượng sinh viên trong trường. Một trong những nguyên nhân khiến cho sinh viên rời trường (chọn trường khác để học) là cảm giác cô đơn. Một cuộc khảo sát được tiến hành tại một trường ĐH lớn vào thời gian trước và sau khi thực hiện chương trình khắc phục nguyên nhân cô đơn này. Kết quả dưới đây là điểm đánh giá sự cô đơn trong trường của 2 mẫu có kích thước lần lượt là 40, và 30 (tương ứng với việc trước và sau thực hiện chương trình). Mức điểm càng cao thì sự cô đơn càng lớn. (file: LonelinessScores.mtp)

Trước chương trình

69	39	64	49	37	17	39	54	61	48
32	16	38	63	19	33	49	21	49	47
52	83	58	57	36	24	40	49	42	59
41	40	19	52	36	44	48	30	54	25

Sau chương trình

26	61	16	31	63	57	26	53	39	45
52	32	31	5	41	28	23	38	34	40
11	45	24	36	55	52	22	26	2	43

a. Tính khoảng tin cậy 95% cho điểm trung bình ở lần trước và sau nghiên cứu của trường ĐH này.
b. Hãy so sánh các điểm của 2 mẫu bằng phương pháp đồ thị.
c. Tính khoảng tin cậy 95% cho sự sai lệch giữa điểm trước và sau khi thực hiện chương trình. Giải thích kết quả.

3.29. Người ta tiến hành nghiên cứu điểm trung bình tích lũy (GPA) môn thống kê của các sinh viên. Người ta chia thành 2 nhóm sinh viên: học ngành kinh doanh và không học ngành kinh doanh. Điểm GPA của 2 nhóm sinh viên được ghi nhận lại như sau (file: MajorGPAs.mtp):

Học ngành kinh doanh		Không học ngành kinh doanh	
3,440	2,650	2,790	3,100
3,941	3,360	3,400	2,890
2,730	3,680	4,000	2,560
3,625	2,760	2,666	3,000
3,330	3,100	3,300	3,900
3,460	3,890	3,750	2,680
2,700	2,500	2,000	2,900
2,690	3,010	3,428	3,650
3,450	3,200	2,750	2,850
3,125	2,800	3,800	2,768

a. Dùng đồ thị tóm tắt điểm GPA cho 2 nhóm sinh viên này.
b. Tính khoảng tin cậy 99% cho sự khác biệt về kỳ vọng của GPA của 2 nhóm, Giải thích kết quả.

3.30. Một nhà quản lý cửa tiệm tra dầu và dầu nhờn nhanh cho xe hơi, ông ta quan tâm đến lượng thời gian cần để phục vụ cho các xe hơi. Hệ thống cũ có một thợ máy thực hiện tất cả mọi việc. Ông ta đề xuất một hệ thống ở đó thợ máy chỉ việc thay dầu, một thợ thứ hai vừa tra dầu nhờn cho xe vừa kiểm tra mức an toàn. Một mẫu gồm 50 xe hơi dùng hệ thống cũ và 50 xe hơi dùng hệ thống mới được chọn, kết quả về thời gian phục vụ theo đơn vị phút cho như sau (file: SystemTimes.mtp)

Thời gian của hệ thống cũ					**Thời gian của hệ thống mới**				
6	11	12	13	13	6	7	8	11	11
14	15	16	17	20	12	13	14	15	16
10	12	12	13	14	6	8	9	11	11
14	15	16	18	20	12	13	14	15	16
10	12	13	13	14	7	8	9	11	12

14	15	17	19	21	12	14	14	15	19
11	12	13	13	14	7	8	9	11	12
15	16	17	19	22	12	14	15	15	19
11	12	13	13	14	7	8	10	11	12
15	16	17	19	28	13	14	15	15	23

a. Tính khoảng tin cậy 95% cho thời gian phục vụ của hệ thống cũ và mới.
b. Hãy so sánh xem thời gian phục vụ trung bình của hệ thống mới có nhanh hơn hệ thống cũ không? Sử dụng $\alpha = 0,02$.

3.31. Một nha sĩ thực hiện một thí nghiệm để đo lường sự hiệu quả của một cách gây tê răng kiểu mới. Hai mươi lăm bệnh nhân được gây tê theo cách bình thường và hai mươi lăm bệnh nhân khác được gây tê theo cách mới. Mỗi bệnh nhân sẽ đánh giá sự không thoải mái của họ bằng cách cho điểm từ 0 đến 100, điểm càng cao sự không thoải mái càng cao (file: Dentist.mtp).

Cách bình thường					**Cách mới**				
23	26	44	32	44	62	50	40	35	52
44	34	26	49	67	82	74	87	30	58
44	53	79	52	33	39	51	72	56	50
50	43	49	33	52	85	57	39	48	64
51	6	30	38	22	75	46	48	56	60

a. Hãy so sánh hai mẫu này bằng đồ thị.
b. Tính khoảng tin cậy 98% cho từng cách và cho sự khác biệt giá trị trung bình giữa 2 cách. Diễn giải kết quả cụ thể.

3.32. Trung tâm bất động sản Minnesota tổng hợp thông tin về nhà bán trong các khu vực của một thành phố. Bảng sau cho biết giá bán của 20 căn nhà chọn ngẫu nhiên từ số nhà bán trong quận A và B của thành phố vào năm 1998.

Quận A ($)		Quận B ($)	
105000	124400	123925	159900
66000	110600	86000	67800
98900	73500	29900	116000
143000	139500	73000	112330
136000	74000	145500	74900
66600	84500	81500	164000
119875	91900	84000	109000
84000	89900	100750	105900
72000	131900	94500	155000
72500	74500	149195	78000

a. Dùng đồ thị so sánh 2 mẫu nói trên.
b. Tính khoảng tin cậy 98% cho từng mẫu.
c. Tính khoảng tin cậy 95% cho sự khác biệt giá trị trung bình của 2 mẫu.
d. Tính khoảng tin cậy 98% cho sự khác biệt giá trị trung bình của 2 mẫu. So sánh kết quả với câu c.

3.33. Trường tiểu học địa phương đang thực hiện so sánh hai phương pháp dạy kỹ năng đọc cho học sinh lớp một. Học sinh lớp một sẽ được bắt cặp theo chỉ số IQ, dân tộc và các yếu tố khác có ảnh hưởng đến khả năng đọc. Lấy ngẫu nhiên một học sinh trong mỗi cặp và dạy theo phương pháp A; Học sinh kia học phương pháp B. Sau khi kết thúc lớp học, kết quả kiểm tra được ghi nhận theo bảng dưới đây (file: Teaching Methods.mtp).

	Điểm của học sinh				
Cặp	**Phương pháp A**	**Phương pháp B**	**Cặp**	**Phương pháp A**	**Phương pháp B**
1	63	80	11	69	43
2	82	86	12	63	60
3	99	56	13	84	74
4	58	54	14	57	91
5	92	73	15	52	83
6	67	68	16	91	81
7	80	94	17	97	97

8	88	67	18	50	87
9	85	79	19	82	90
10	76	71	20	59	88

a. So sánh bằng đồ thị hai phương pháp dạy.
b. Xác định và diễn dịch khoảng tin cậy 95% cho sự khác biệt giữa giá trị trung bình điểm kiểm tra. Mô tả bằng đồ thị sự khác biệt này.

3.34. Nhiều người hút thuốc trong một thời gian dài khó có thể bỏ thuốc. Lấy ngẫu nhiên một nhóm 30 người nghiện hút thuốc và muốn bỏ thuốc và tách thành hai nhóm, mỗi nhóm 15 người. Một nhóm được đưa vào chương trình cai thuốc mới và nhóm còn lại đưa vào chương trình cũ. Một tháng sau khi hoàn tất chương trình, dữ liệu ghi nhận số thuốc lá được hút trong ngày như sau

CT cũ	14	0	0	20	15	12	0	0	5	3	0	15	2	0	12
CT mới	3	20	0	27	10	0	1	8	0	19	24	14	8	22	14

a. So sánh dữ liệu bằng số và bằng đồ thị hai nhóm.
b. Có bằng chứng nào trong mẫu cho rằng chương trình mới hiệu quả hơn chương trình cũ hay không? Chọn thủ tục kiểm định thích hợp, dùng α=0,05 (xem thêm lý thuyết kiểm định giả thuyết thống kê).

3.35. Trong khoá học mùa Thu năm 2001, một nhóm sinh viên thống kê đã so sánh điểm trung bình của sinh viên sống trong ký túc xá và sinh viên sống bên ngoài. Nhóm này quan tâm liệu có một sự khác biệt nào giữa phân phối điểm trung bình của hai nhóm sinh viên. Mẫu ngẫu nhiên thể hiện bảng dưới đây. (file: CampusGPA.mtp)

Trong KTX				**Bên ngoài**				
2,2	3,4	4,0	3,1	2,8	3,4	2,2	2,9	2,9
2,9	2,9	2,8	2,6	2,6	2,9	3,0	2,8	3,4
3,5	2,2	2,9	3,6	2,9	2,6	3,5	2,5	

3,3	3,8	3,0	2,8	3,7	2,7	2,7	2,7
3,3	3,2	3,0	2,8	2,9	3,0	2,7	2,9
3,2	3,0	2,5	2,5	2,5	2,5	3,5	3,2

a. So sánh bằng số và đồ thị hai mẫu.
b. Xây dựng và diễn dịch khoảng tin cậy 90% của sự khác biệt về điểm trung bình của hai mẫu.
c. Có đủ bằng chứng trong hai mẫu cho thấy rằng điểm trung bình của hai nhóm khác nhau? Dùng $\alpha = 0,10$

3.36. Một nghiên cứu về các cặp vợ chồng đều đang làm việc bên ngoài và cho thấy rằng tất cả các bà vợ đều dành nhiều thời gian cho việc nhà. Người ta lấy ngẫu nhiên 15 cặp vợ chồng và hỏi tỷ lệ về công việc nhà họ làm. Dữ liệu thể hiện trong bảng dưới đây (file: Housework.mtp)

Stt	Chồng	Vợ	Stt	Chồng	Vợ
1	51	64	9	47	62
2	46	57	10	62	61
3	54	65	11	47	65
4	50	69	12	39	57
5	50	59	13	62	57
6	51	63	14	44	55
7	57	57	15	57	60
8	49	65			

a. Ước lượng sự khác biệt trong giá trị trung bình về tỷ lệ công việc nhà do vợ và chồng đảm trách, dùng khoảng tin cậy 95%.
b. Có đủ bằng chứng trong mẫu cho thấy rằng các bà vợ có tỷ lệ làm công việc nội trợ nhiều hơn? Dùng $\alpha = 0,10$.

CHƯƠNG 4
KIỂM SOÁT QUÁ TRÌNH

4.1 GIỚI THIỆU KHÁI QUÁT

Quá trình là trình tự các bước để tạo ra một kết quả. Ví dụ một nhân viên tiếp tân trả lời điện thoại và chuyển cuộc gọi đến người liên quan. Một phòng thí nghiệm đo lượng cholesterol trong một mẫu máu. Một máy đóng chai nước soda loại 12 ounce (1 ounce ≈ 28g). Trong mỗi trường hợp các công việc cụ thể được lặp đi lặp lại theo thời gian. Theo cách lý tưởng thì những công việc này phải được thực hiện "hoàn hảo". Nhân viên tiếp tân trả lời mỗi cuộc gọi ngay tiếng chuông đầu tiên và chuyển cuộc gọi đến đúng người. Phòng thí nghiệm thì xác định chính xác lượng cholesterol trong mỗi mẫu máu. Máy đóng chai thì phải đong chính xác 12 ounces soda vào mỗi chai. Nhưng thực tế thì sẽ không thể như vậy, quá trình có thay đổi.

Có hai kiểu cơ bản của sự thay đổi quá trình. Thay đổi do nguyên nhân thông thường (common-cause variation) – phát sinh do khả năng hay tính biến thiên cố hữu trong hệ thống. Cách duy nhất để giảm sự thay đổi này là phải thay đổi chính quá trình. Ví dụ, trong trường hợp nhân viên tiếp tân, chúng ta phải cài đặt một hệ thống điện thoại mới, vẽ một sơ đồ để nhân viên tiếp tân có thể xác định nhanh chóng và chính xác người hay bộ phận để chuyển cuộc gọi điện thoại đến. Hoặc chúng ta phải giảm sự sao lãng và tiếng ồn trong văn phòng để nhân viên tiếp tân có thể làm việc hiệu quả hơn.

Loại thay đổi thứ hai, thay đổi do nguyên nhân đặc biệt (special-cause variation), phát sinh do những nguyên nhân đặc biệt. Trong ví dụ nhân viên tiếp tân, có thể một tháng nào đó công ty đặt thêm chương trình quảng cáo, làm cho số cuộc gọi điện thoại nhiều hơn số lượng mà nhân viên tiếp tân có khả năng kiểm soát, dẫn đến thời gian để đáp ứng cuộc gọi tăng lên. Một quá trình được xem là trong vòng kiểm soát nếu không có nguyên nhân bất thường nào tác động lên, mà chỉ có nguyên nhân thông thường mà thôi.

Chương này sẽ xem xét vài thủ tục kiểm soát thống kê để đo lường và điều chỉnh đặc tính của sản phẩm và dịch vụ. Đồ thị kiểm soát được dùng để nghiên cứu quá trình theo thời gian. Có hai loại cơ bản: các đồ thị dành cho dữ liệu đo lường được và các đồ thị dành cho dữ liệu có tính thuộc tính (dữ liệu đếm). Nếu chúng ta ghi nhận thời gian từ lúc một cuộc điện thoại gọi đến công ty đến khi người gọi tiếp xúc được người muốn gặp, ta có dữ liệu đo lường. Nếu mỗi ngày chúng ta ghi nhận tổng số các cuộc gọi và tổng số cuộc gọi bị chỉ dẫn sai, ta có dữ liệu có tính thuộc tính (ví dụ trường hợp này là cuộc gọi có thuộc tính là bị chỉ dẫn sai)

Một đồ thị kiểm soát sẽ giám sát liên tục một đặc tính định lượng hay định tính của một quá trình. Chúng ta minh họa các đồ thị kiểm soát thống kê cho những quan sát riêng lẻ, trung bình của quá trình, sự phân tán, phần tỷ lệ và số khuyết tật mỗi sản phẩm.

4.2 CÁC ĐẶC TÍNH CHUNG CỦA ĐỒ THỊ KIỂM SOÁT

Đồ thị kiểm soát (Control charts) là sơ đồ của các quan sát hay thống kê của mẫu được thu thập định kỳ từ một quá trình liên tục. Các ví dụ của thống kê mẫu bao gồm trung bình và sự phân tán của một quá trình hay tỷ lệ của các sản phẩm khuyết tật được sản xuất bởi một quá trình.

Để xây dựng một đồ thị kiểm soát, kết quả đầu ra của quá trình sẽ được lấy mẫu theo thời điểm định kỳ, sau đó các đặc trưng thống kê được tính toán và được vẽ theo thời gian. Đường trung tâm của đồ thị kiểm soát là trung bình của quá trình. Giới hạn kiểm soát bên trên và bên dưới (Upper and Lower Control Limit – UCL, LCL) nhìn chung sẽ bằng trung bình của quá trình cộng và trừ 3 lần độ lệch chuẩn. Nếu chưa biết trước, trung bình và độ lệch chuẩn được ước lượng từ dữ liệu mẫu.

Sự thay đổi giá trị của đặc trưng thống kê của mẫu theo thời gian sẽ cho biết thông tin hữu ích về quá trình chính. Quá trình có thể không kiểm soát được nếu một thống kê mẫu rơi ra ngoài giới hạn kiểm soát, hoặc nếu hình dáng thay đổi dữ liệu cho thấy đó là một quá trình không ngẫu nhiên (xem 8 loại kiểm tra). Các đồ thị kiểm soát sẽ cho

thấy liệu rằng có nên thực hiện hành động điều chỉnh để thay đổi quá trình hay không.

4.2.1 Các Tùy Chọn Lệnh Trong Minitab

Phần này mô tả những chọn lựa thông thường cho hầu hết các đồ thị kiểm soát của Minitab. Các đồ thị này bao gồm:

I chart Đồ thị kiểm soát các quan sát riêng lẻ (individual observations)

$\bar{x}$ - chart Đồ thị kiểm soát giá trị trung bình

S- chart Đồ thị kiểm soát độ lệch chuẩn trong quá trình (standard deviation)

R- chart Đồ thị kiểm soát khoảng (ranges)

C- chart Đồ thị kiểm soát số khuyết tật

P- chart Đồ thị kiểm soát tỷ lệ khuyết tật

Ghi chú: C-chart và P-chart là đồ thị dùng cho dữ liệu có tính thuộc tính

4.2.2 Kiểm Tra Thống Kê

Minitab có 8 kiểm tra thống kê để giám sát một quá trình. Mỗi loại đồ thị kiểm soát đều áp dụng một vài hoặc tất cả các kiểm tra này.

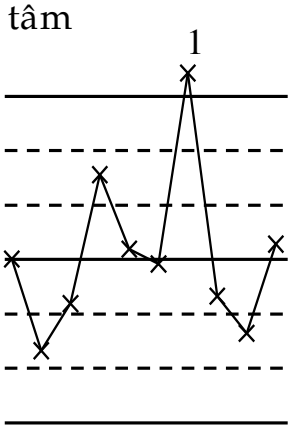
4.2.3 Các Kiểm Tra Dùng Trong Đồ Thị Kiểm Soát

Hình 4.1 trình bày 8 loại kiểm tra có thể dùng để phát hiện dạng thay đổi đặc biệt và bất thường của quá trình. Giá trị bằng số được cho trong mỗi kiểm tra là mặc nhiên; giá trị này có thể thay đổi trong một khoảng nào đó tùy vào sự khai báo và điều chỉnh.

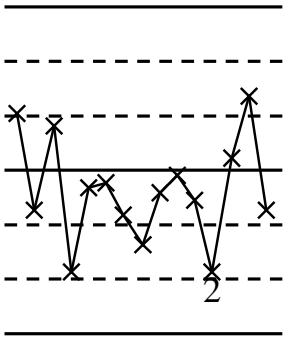
Một điều chúng ta cần lưu ý là thực tế quá trình ổn định (nằm trong sự kiểm soát) không có nghĩa quá trình được chấp nhận sử dụng. Ví dụ quá trình đong soda vào chai, lượng đong trung bình của quá trình là 11 ounces mà không phải 12 ounces theo yêu cầu chất lượng, và giả sử tất cả các điểm kiểm tra nằm trong giới hạn và không phát hiện ra dạng đặc biệt nào. Về mặt thống kê thì quá trình nằm trong sự kiểm soát. Về mặt thực tế hoạt động của công ty thì những chai soda sản xuất ra không đạt yêu cầu về trọng lượng và dĩ nhiên không được thị trường chấp nhận.

Hình 4.1 Các loại kiểm tra

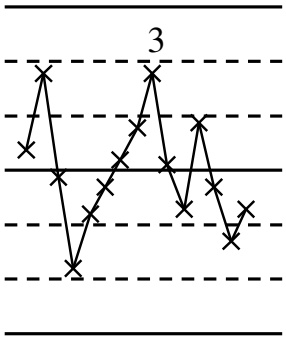
Kiểm tra 1
1 điểm (phần tử) nằm ngoài phạm vi 3σ từ đường trung tâm



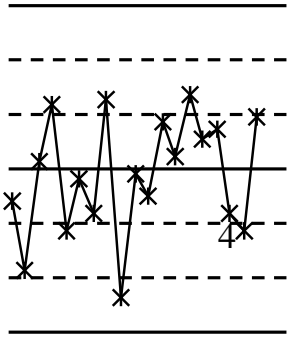
Kiểm tra 2
9 điểm liên tiếp nằm về một phía của đường trung tâm



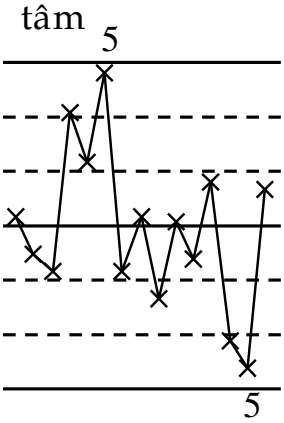
Kiểm tra 3
6 điểm tăng hoặc giảm liên tục



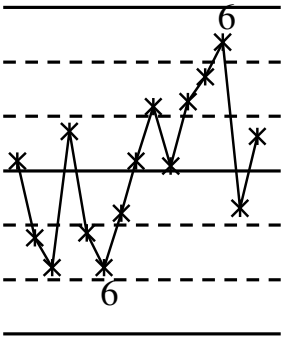
Kiểm tra 4
14 điểm liên tiếp lên và xuống thay phiên nhau



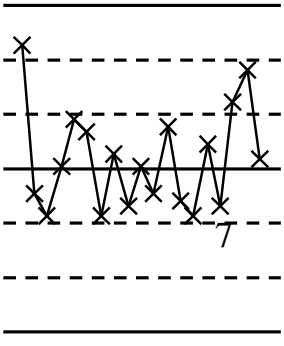
Kiểm tra 5
2 trong 3 điểm liên tiếp nằm ngoài phạm vi 2σ và về cùng một phía của đường trung tâm



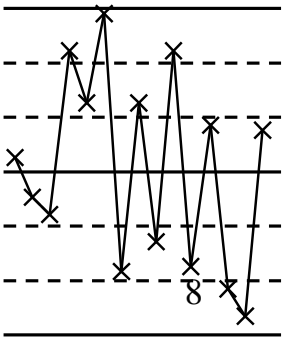
Kiểm tra 6
4 trong 5 điểm liên tiếp nằm ngoài phạm vi 1σ và về cùng một phía của đường trung tâm



Kiểm tra 7
15 điểm liên tiếp nhau nằm trong khoảng 1σ của đường trung tâm



Kiểm tra 8
8 điểm liên tiếp nằm ngoài phạm vi 1σ của đường trung tâm



Ghi chú: đường liền nét ở giữa là đường trung tâm. Các đường đứt nét cách đường trung tâm 1σ và 2σ. Đường liền nét phía trên và phía dưới cách 3σ so với đường trung tâm.

4.3 ĐỒ THỊ DÀNH CHO CÁC QUAN SÁT RIÊNG LẺ: I CHART

I Chart giám sát một đặc trưng định lượng nào đó của sản phẩm, ví dụ trọng lượng, chiều dài, hay đường kính. Đồ thị này vẽ các đo lường riêng lẻ được ghi nhận theo các thời điểm định kỳ.

Đồ Thị Kiểm Soát Dành Cho Các Quan Sát Riêng Lẻ

Đường trung tâm của đồ thị là trung bình của quá trình μ. Giới hạn kiểm soát trên và dưới được xác định bởi

$$UCL = \mu + 3\sigma \qquad LCL = \mu - 3\sigma$$

Trong đó σ là độ lệch chuẩn của tổng thể. Nếu chưa biết, thì μ và σ được ước lượng từ dữ liệu. Một đo lường có thể rơi ra ngoài phạm vi kiểm soát này nếu có một sự cố thật đặc biệt xảy ra hoặc nếu quá trình đi ra ngoài sự kiểm soát.

Để vẽ đồ thị này dùng Menu Stat > Control Charts > Individuals

Ví dụ 1: Đồ thị kiểm soát quan sát riêng lẻ

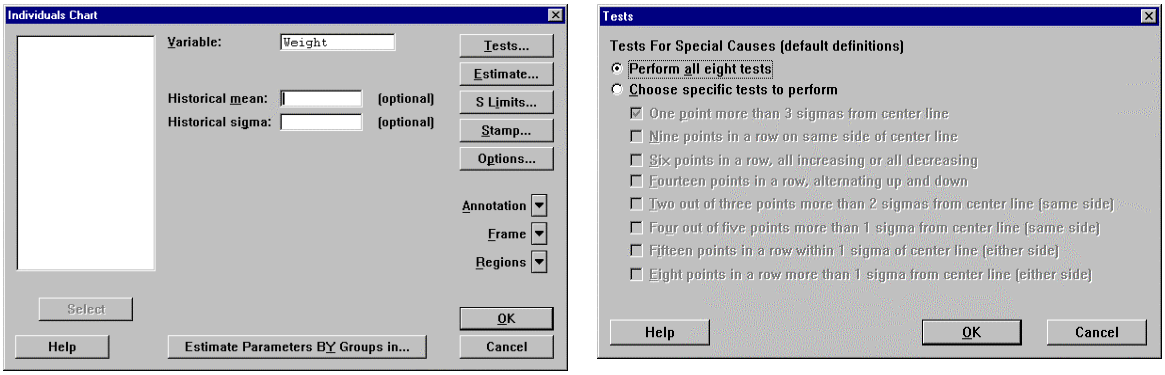
Một cửa hàng tạp hóa lớn có một máy cân và gói thịt. Một bộ phận điều chỉnh trên máy cho phép người vận hành đổ đầy gói với các trọng lượng khác nhau. Giả sử rằng người vận hành điều chỉnh máy để đổ đầy gói với trọng lượng 1 pound (khoảng 450 gam). Để giám sát quá trình đổ đầy, cứ sau 5 phút người vận hành chọn ngẫu nhiên một gói và đem cân. Trọng lượng của 20 gói được chọn cho trong bảng sau. Hãy xây dựng một đồ thị kiểm soát. Xác định xem quá trình có trong sự kiểm soát hay không.

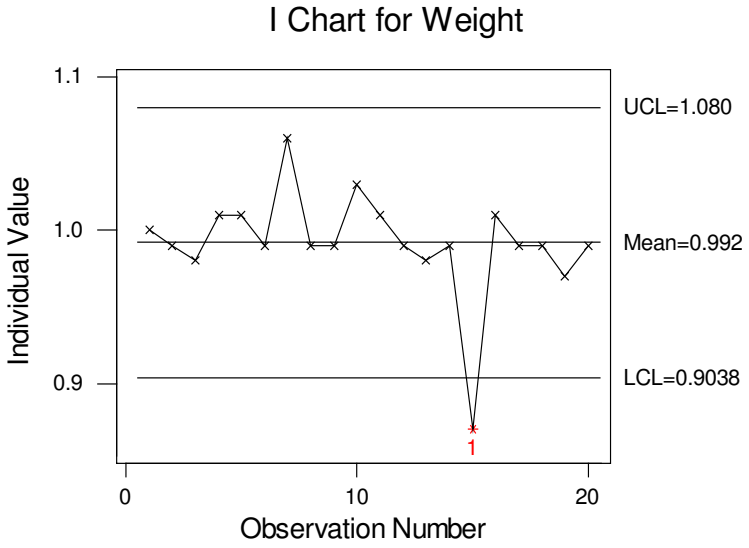
Gói	Trọng lượng	Gói	Trọng lượng
1	1.00	11	1.01
2	0.99	12	0.99
3	0.98	13	0.98
4	1.01	14	0.99
5	1.01	15	0.87
6	0.99	16	1.01
7	1.06	17	0.99

8	0.99	18	0.99
9	0.99	19	0.97
10	1.03	20	0.99

Lời giải: Dữ liệu được cho trong file Weight-1.mtp

Vào menu **Stat > Control Charts > Individuals**
Chọn biến Weight vào ô **Variable:**
Nhấn **Tests**, Chọn **Perform all eight tests**. Nhấn **OK** và **OK**.






```
TEST 1. One point more than 3.00 sigmas from center line.
Test Failed at points: 15
```

Đồ thị cho thấy trọng lượng của gói thứ 15 giảm xuống quá 3 độ lệch chuẩn của trung bình. Quá trình không nằm trong sự kiểm soát. Đồ thị kiểm soát này không nên dùng để giám sát kết quả đầu ra của quá trình trong tương lai.

4.4 ĐỒ THỊ KIỂM SOÁT GIÁ TRỊ TRUNG BÌNH: $\bar{x}$ - CHART

$\bar{x}$ - chart giám sát một đặc trưng định lượng của một quá trình, ví dụ như trọng lượng, độ dài hay đường kính. Đồ thị này cho thấy trung bình của mẫu ngẫu nhiên được chọn theo từng thời điểm định kỳ. Đường trung bình của $\bar{x}$ - chart là trung bình của quá trình μ. Giới hạn trên và dưới được xác định bởi

$$UCL = \mu + 3\sigma/\sqrt{n} \qquad LCL = \mu - 3\sigma/\sqrt{n}$$

Trong đó σ là độ lệch chuẩn của tổng thể và n là kích thước mẫu. Nếu chưa biết, thì μ và σ được ước lượng từ dữ liệu. Trung bình của mẫu có thể rơi ra ngoài các giới hạn này nếu có một sự cố đặc biệt xảy ra hoặc nếu quá trình đi ra ngoài sự kiểm soát. Sự phân tán của quá trình được giả thiết là ổn định (độ lệch chuẩn của quá trình nằm trong sự kiểm soát).

Đồ Thị Kiểm Soát Để Giám Sát Trung Bình Của Quá Trình

Lệnh này sẽ vẽ các trung bình của mẫu. Khi dùng lệnh chú ý cách khai báo cho trường hợp những phân nhóm nằm theo các cột, và những phân nhóm nằm theo các hàng. Nếu phân nhóm nằm theo cột, thì khi nhập dữ liệu cột phải kèm theo 1 hằng số hoặc cột định rõ kích thước mẫu. Nếu μ và σ chưa biết, thì chúng được ước lượng từ dữ liệu.

Để vẽ đồ thị này dùng menu **Stat > Control Charts > Xbar**

Ví dụ 2: Đồ thị kiểm soát: Trung bình của Quá trình

Dựa vào ví dụ 1. Một tiệm tạp hóa lớn có một máy để cân và gói thịt. Một bộ phận điều chỉnh trên máy cho phép người vận hành đổ đầy các bao với trọng lượng khác nhau. Giả sử người vận hành điều chỉnh máy để đổ các bao với trọng lượng 1 pound (≈ 450 gram). Để giám sát quá trình đổ, cứ mỗi 15 phút người vận hành chọn ngẫu nhiên một mẫu gồm 5 bao. Trọng lượng của 20 mẫu được cho trong bảng dưới đây. Hãy xây dựng và diễn dịch đồ thị $\bar{x}$ - chart, dùng dữ liệu để ước lượng μ và σ. Hãy dùng tất cả kiểu kiểm tra cho những nguyên nhân đặc biệt của sự thay đổi.

Mẫu	Trọng lượng của các bao (pound)				
1	1.01	1.03	1.00	1.06	1.02
2	1.05	1.05	1.05	1.04	1.08
3	1.09	1.03	0.97	1.04	0.96
4	1.01	1.07	1.07	1.07	1.11
5	1.03	0.99	0.92	1.00	1.04
6	1.07	1.03	1.00	1.04	0.99
7	0.99	1.00	1.09	0.97	1.03
8	0.87	1.00	1.03	1.05	1.07
9	0.99	0.93	1.01	0.95	0.96
10	1.02	1.02	1.00	1.02	0.93
11	0.94	1.09	1.06	1.05	1.01
12	0.98	0.98	1.05	1.00	1.07
13	1.05	1.00	1.03	1.14	1.09
14	0.95	0.97	1.06	1.00	1.08
15	1.06	0.99	1.00	0.98	1.01
16	1.06	1.03	1.03	1.01	1.09
17	1.09	1.09	1.17	1.05	0.89
18	1.02	1.01	1.04	0.98	1.08
19	1.10	0.99	1.05	0.96	1.10
20	1.08	0.91	0.94	0.97	0.98

Lời giải: Dữ liệu được lưu trong file Weight-5.mtp; trọng lượng nằm trong cột C1 đến C5. Mỗi hàng của bảng tính Minitab cho biết 5 quan sát trong một mẫu.

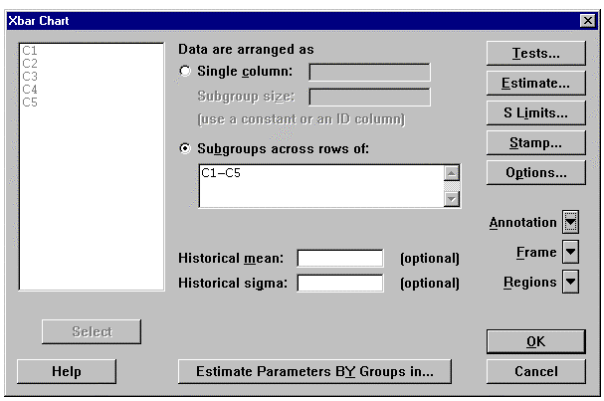
Dùng menu **Stat > Control Charts > Xbar**

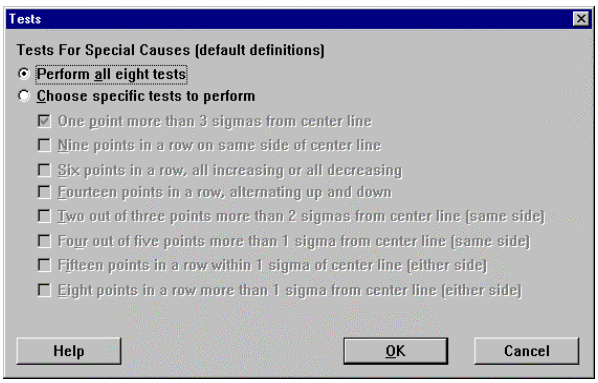
Chọn **Subgroups across rows of:**

Nhấn **Select** chọn các biến từ **C1-C5**

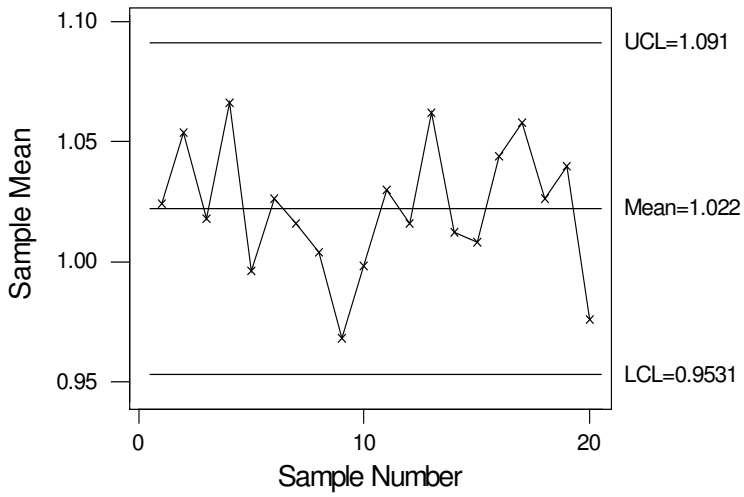
Nhấn **Tests**, chọn **Perform all eight tests**; nhấn **OK**.

Nhấn **Annotation**, chọn **Title**, đánh tiêu đề đồ thị 'Đồ thị kiểm soát trọng lượng bao', format font VNI. **OK**





Đồ thị kiểm soát trọng lượng bao



Chúng ta thấy không có nguyên nhân đặc biệt nào của sự thay đổi trong đồ thị $\bar{x}$ - chart. Quá trình đổ không nằm ngoài sự kiểm soát. Các giới hạn kiểm soát có thể được dùng để giám sát đầu ra tiếp theo của quá trình.

4.5 ĐỒ THỊ KIỂM SOÁT SỰ PHÂN TÁN CỦA QUÁ TRÌNH

Đồ thị kiểm soát sự phân tán của quá trình thường vẽ trước đồ thị trung bình $\bar{x}$ - chart. Đồ thị S-chart giám sát sự phân tán, được đo bằng độ lệch chuẩn. Sự phân tán của quá trình sản xuất nằm trong sự kiểm soát nếu tất cả độ lệch chuẩn của mẫu rơi vào phạm vi giới hạn kiểm soát và không có nguyên nhân đặc biệt nào của sự biến đổi được phát hiện.

Đồ Thị Kiểm Soát Độ Lệch Chuẩn Của Quá Trình

Lệnh này sẽ vẽ một đồ thị kiểm soát của các độ lệch chuẩn. Khi dùng lệnh chú ý cách khai báo với trường hợp các phân nhóm được lưu trong cột hay là hàng. Nếu các phân nhóm được lưu theo cột thì khi nhập dữ liệu cột phải kèm theo 1 hằng số hoặc cột định rõ kích thước mẫu.

Giá trị ước lượng mặc nhiên của sigma là trung bình của các độ lệch chuẩn các phân nhóm. Bạn có thể nhập một giá trị SIGMA quá khứ.

Để vẽ đồ thị S-chart dùng menu **Stat > Control Charts > S**

Ngoài đồ thị S, đồ thị R-chart cũng giám sát sự phân tán của quá trình, được đo bằng khoảng (range) của các mẫu. Khoảng là sự chênh lệch giữa giá trị lớn nhất và nhỏ nhất trong mỗi mẫu. Sự phân tán của quá trình nằm trong sự kiểm soát khi tất cả các khoảng của mẫu nằm trong giới hạn kiểm soát và không có nguyên nhân đặc biệt của sự biến đổi được phát hiện. Đường trung tâm của R-chart đặt tại giá trị trung bình của các khoảng μ_R, và giới hạn kiểm soát trên và dưới đặt tại giá trị $\mu_R \pm 3\sigma_R$. Nếu chưa biết thì μ_R và σ_R được ước lượng từ dữ liệu mẫu.

Đồ Thị Kiểm Soát Khoảng Của Quá Trình

Lệnh này sẽ vẽ một đồ thị kiểm soát các khoảng của mẫu. Lưu ý các cách khai báo lệnh cho trường hợp dữ liệu được xếp theo cột và theo hàng.

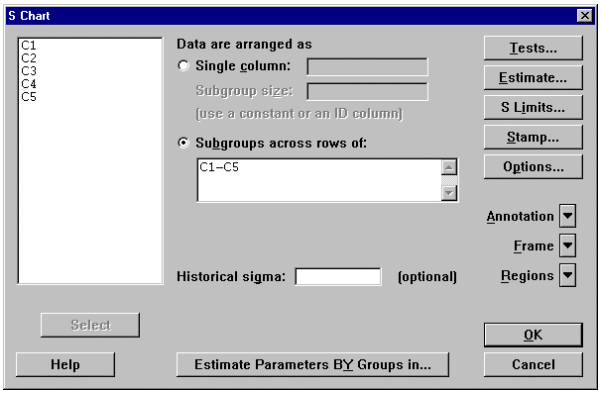
Để vẽ đồ thị R-chart dùng menu **Stat > Control Charts > R**

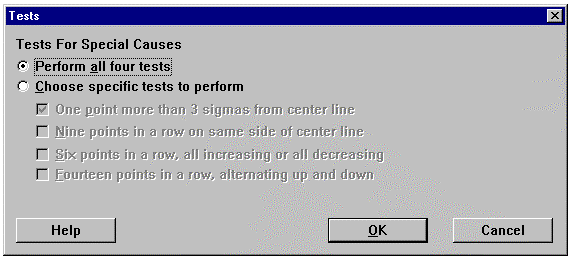
Ví dụ 3: Đồ thị kiểm soát sự phân tán của quá trình
Dựa vào ví dụ của quá trình đổ đầy và cân bao thịt trong ví dụ 2 trước. Một tiệm tạp hóa lớn có một máy để cân và gói thịt. Cứ mỗi 15 phút người vận hành chọn ngẫu nhiên 5 bao để cân. Dữ liệu cho trong ví dụ 2. Hãy xây dựng và diễn dịch một đồ thị S và một đồ thị R.

Lời giải: Lấy dữ liệu lưu trong file Weight-5.mtp. Các quan sát của mẫu được xếp theo hàng của các cột từ C1 đến C5.

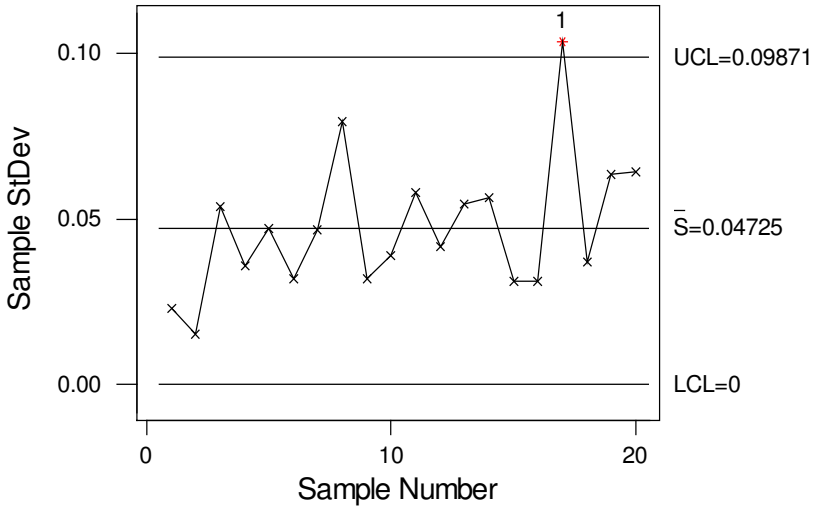
Kiểm soát độ lệch chuẩn:

Vào menu **Stat > Control Charts > S**
Chọn **Subgroups across rows of:**
Chọn các biến từ C1-C5;
Nhấn **Tests**, chọn **Perform all four tests; OK.**
Nhấn **Annotation**, chọn **Title**, nhập tên đồ thị 'Độ lệch chuẩn của trọng lượng', format font VNI. **OK.**





Độ lệch chuẩn của trọng lượng

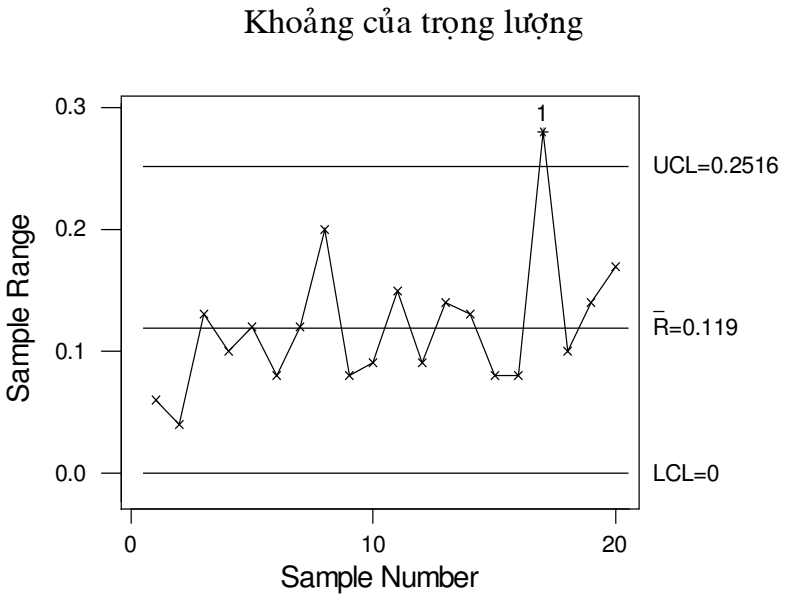



```
TEST 1. One point more than 3.00 sigmas from center line.
Test Failed at points: 17
```

Đồ thị dạng S cho thấy sự phân tán của quá trình nằm ngoài sự kiểm soát. Một kiểm tra bị vi phạm ở mẫu số 17 do có độ lệch chuẩn lớn hơn một cách bất thường.

Kiểm soát khoảng:

Vào menu **Stat > Control Charts > R**
Chọn **Subgroups across rows of:**
Chọn các biến từ C1-C5;
Nhấn **Tests**, chọn **Perform all four tests; OK.**
Nhấn **Annotation**, chọn **Title**, nhập tên đồ thị 'Khoảng của các trọng lượng', format font VNI. **OK.**




```
TEST 1. One point more than 3.00 sigmas from center line.
Test Failed at points: 17
```

Đồ thị R cũng cho thấy rằng sự phân tán quá trình nằm ngoài sự kiểm soát. Kiểm tra này cũng thất bại ngay tại mẫu số 17 giống như đồ thị S. Người vận hành máy nên xác định nguyên nhân gây ra sự chênh lệch trọng lượng trong mẫu lớn một cách bất thường.

Lưu ý: Đồ thị trung bình (Xbar) và đồ thị độ lệch chuẩn (S) hoặc khoảng có thể vẽ cùng một lúc bằng cách dùng menu Stat > Control Charts > Xbar-S hoặc Stat > Control Charts > Xbar-R.

5.5 ĐỒ THỊ KIỂM SOÁT PHẦN TỶ LỆ: P-CHART

Đồ thị P giám sát một tính chất định tính của một quá trình, ví dụ như tỷ lệ của sản phẩm không phù hợp (nonconforming) hay khuyết tật (defective). Đồ thị chỉ ra tỷ lệ của khuyết tật được quan sát trong các mẫu thu thập theo thời gian.

Đường trung tâm là p, tỷ lệ khuyết tật của quá trình. Giới hạn kiểm soát trên và dưới là

$$\text{UCL} = \quad p + 3\sqrt{p(1-p)/n} \qquad\qquad \text{LCL} = \quad p - 3\sqrt{p(1-p)/n}$$

Trong đó n là kích thước mẫu. Vì $0 \le p \le 1$, UCL sẽ lấy bằng 1 nếu giá trị tính toán của UCL lớn hơn 1, LCL sẽ cho bằng 0 nếu giá trị LCL tính toán nhỏ hơn 0. Nếu p chưa biết thì sẽ ước lượng từ dữ liệu mẫu.

Đồ Thị Kiểm Soát Tỷ Lệ

Lệnh này sẽ vẽ đồ thị P về tỷ lệ sản phẩm không phù hợp. Cột của dữ liệu chứa số sản phẩm khuyết tật hay không phù hợp trong các mẫu. Một hằng số hay cột định rõ kích thước mẫu. Nếu E là một hằng số, mọi mẫu đều có kích thước là E. Nếu E là một cột, các số trong cột xác định các kích thước mẫu.

Để vẽ đồ thị vào menu **Stat > Control Charts > P**

Ví dụ 4: Đồ thị P: Tỷ lệ của Quá trình
Bộ phận quản lý hàng không liên bang liên tục giám sát các chuyến bay của các máy bay, mối quan tâm đặc biệt là tỷ lệ các chuyến bay không đúng giờ. Trong quá khứ, khoảng 30% các chuyến bay không đến đúng lịch trình trong vòng 15 phút và được xếp loại là bị trễ giờ. Giả sử một hãng hàng không lớn có tỷ lệ trễ chuyến 30% phát động một chương trình giảm chuyến bay trễ. Để giám sát sự thành công của chương trình, hãng hàng không chọn ngẫu nhiên 25 chuyến bay mỗi ngày trong 28 ngày và ghi nhận số chuyến trễ sau đây. Hãy xây dựng và diễn dịch đồ thị P về tỷ lệ của các chuyến bay trễ.

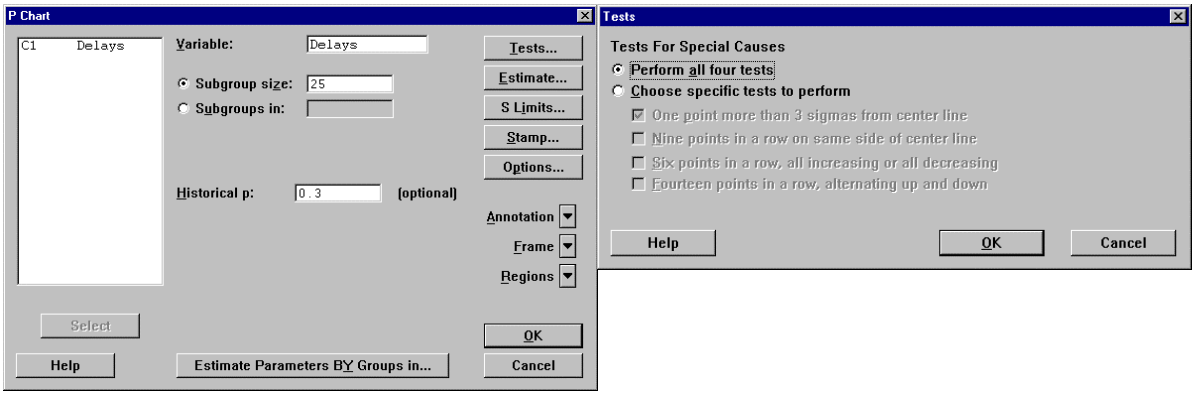
Ngày	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Số trễ	12	6	8	6	13	8	10	6	9	7	3	5	11	7

Ngày	15	16	17	18	19	20	21	22	23	24	25	26	27	28
Số trễ	3	4	7	7	9	6	5	7	2	7	3	2	4	5

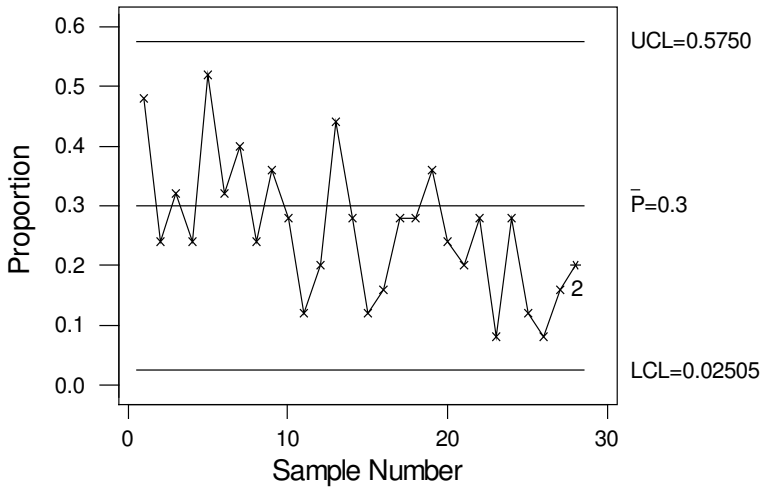
Lời giải: Dữ liệu lưu trong file Airline.mtp. Số chuyến trễ lưu trong cột Delays. Kích thước mẫu là n = 25 cho mỗi ngày và tỷ lệ quá khứ p = 0,3 được chỉ rõ trong hộp thoại.

Vào menu **Stat > Control Charts > P**
Chọn biến Delays vào ô **Variable:**
Nhập giá trị 25 vào **Subgroup size:**

Và giá trị 0.3 vào **Historical p:**
Nhấn **Tests**, chọn **Perform all four tests**;
Nhấn **Annotation**, chọn **Title**, nhập tiêu đề 'Tỷ lệ chuyến bay trễ hàng ngày'. Nhấn **OK**.



Tỷ lệ chuyến bay trễ hàng ngày




```
TEST 2. 9 points in a row on same side of center line.
Test Failed at points: 28
```

Đồ thị P cho thấy chương trình của hãng hàng không đem lại kết quả là số chuyến bay trễ có khuynh hướng giảm theo thời gian. Tuy nhiên, khuynh hướng này không tiếp tục vào ngày cuối cùng (thứ 28), vì khi thể hiện tỷ lệ chuyến trễ giảm dựa theo tỷ lệ quá khứ 30% thì kiểm tra số 2 bị vi phạm.

4.6 ĐỒ THỊ KIỂM SOÁT SỐ LƯỢNG KHUYẾT TẬT: C-CHART

Đồ thị C giám sát số khuyết tật trên mỗi sản phẩm. Số khuyết tật c chứa trong mỗi sản phẩm được giả thiết có phân phối xác suất Poisson, với giá trị trung bình μ. Đường trung tâm đặt tại giá trị cho biết số trung bình khuyết tật trên mỗi sản phẩm, và giới hạn kiểm soát trên và dưới đặt tại giá trị 3σ trên và dưới đường trung tâm.

Đồ Thị Kiểm Soát Số Lượng Khuyết Tật

Lệnh này vẽ đồ thị kiểm soát số lượng khuyết tật trong mỗi mẫu. Các mẫu khuyết tật được giả thiết có phân phối Poisson với trị trung bình là μ.

Để vẽ đồ thị vào menu **Stat > Control Charts > C**

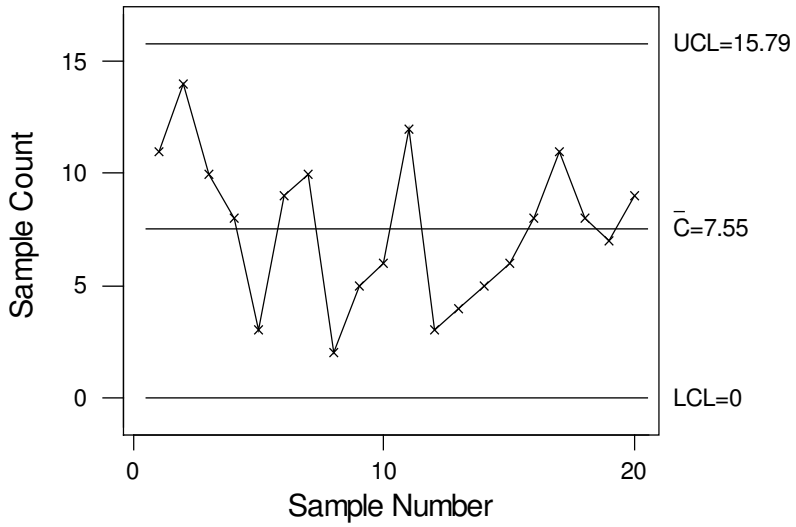
Ví dụ 5: Đồ thị C: Số khuyết tật
Mỗi giờ, một nhà máy sản xuất xe hơi chọn chiếc xe hoàn tất kế tiếp trong một dây chuyền sản xuất, và đếm số lỗi sơn trên bề mặt xe. Kết quả sau một thời đoạn 20 giờ cho như sau. Hãy xây dựng một đồ thị kiểm soát để giám sát chất lượng của bề mặt sơn.

Giờ	1	2	3	4	5	6	7	8	9	10
Số khuyết tật	11	14	10	8	3	9	10	2	5	6
Giờ	11	12	13	14	15	16	17	18	19	20
Số khuyết tật	12	3	4	5	6	8	11	8	7	9

Lời giải: Dữ liệu được lưu trong file Auto.mtp. Số khuyết tật lưu trong cột có tên Defects. Đồ thị C sẽ giám sát số khuyết tật cho mỗi xe hơi

Vào menu **Stat > Control Charts > C**
Chọn Defects vào ô **Variable:**
Nhấn **Tests**, chọn cả 4 kiểu kiểm tra;
Nhấn **Annotation**, chọn **Title**, nhập tiêu đề, nhấn **OK**.

Số khuyết tật sơn trên xe hơi



Tất cả 20 quan sát đều nằm trong giới hạn kiểm soát. Các kiểu kiểm tra thống kê để giám sát chất lượng bề mặt sơn đã không phát hiện bất kỳ dạng biến đổi đặc biệt nào. Quá trình cho thấy nằm trong sự kiểm soát.

BÀI TẬP

4.1. Tại một nhà máy đóng chai cola, mỗi ngày một kỹ thuật viên kiểm soát chất lượng lấy mẫu ngẫu nhiên một chai đã hoàn tất từ dây chuyền sản xuất và ghi nhận trọng lượng của chai, trọng lượng đo bằng ounces (1 ounce ≈ 28 g). Việc lấy mẫu được thực hiện trong 20 ngày. Dữ liệu cho trong bảng sau. Hãy xây dựng một đồ thị kiểm soát cho loại quan sát riêng lẻ, dùng cả 8 loại kiểm tra. Quá trình có nằm trong sự kiểm soát hay không? (file: Cola.mtp)

Day	Weight	Day	Weight
1	5.6	11	6.2
2	5.7	12	5.9
3	6.1	13	5.2
4	6.3	14	6.0
5	5.2	15	6.3
6	6.0	16	5.8
7	5.8	17	6.1
8	5.8	18	6.2
9	6.4	19	5.3
10	6.0	20	6.0

4.2. Một máy đóng gói một loại ngũ cốc khô ăn nhanh, máy này được lắp đặt để đổ đầy một hộp 300 inch khối (1 inch ≈ 2.5 cm) với 20 ounces ngũ cốc (1 ounce ≈ 28 g). Nhà sản xuất muốn giám sát trọng lượng và thể tích của quá trình đóng gói. Bảng sau cho thấy dữ liệu của 20 mẫu, mỗi mẫu lấy một lúc 3 hộp. (File: Cereal.mtp)

	Trọng lượng			Thể tích		
Sample	**Wt.1**	**Wt.2**	**Wt.3**	**Vol.1**	**Vol.2**	**Vol.3**
1	20.0	19.9	19.9	284	280	275
2	20.1	19.9	20.0	279	279	282
3	20.0	19.8	19.9	281	279	277
4	19.9	19.9	19.9	280	279	276
5	20.1	20.4	20.0	285	287	276
6	20.3	20.2	20.0	285	284	284
7	20.1	20.7	19.9	282	294	278
8	20.0	19.9	20.2	278	279	286

9	20.2	20.4	20.0	281	287	280
10	20.2	20.2	20.3	280	283	286
11	20.0	19.8	19.8	278	277	275
12	20.0	20.0	20.1	283	276	281
13	20.0	19.5	20.0	275	272	284
14	19.9	20.1	19.9	279	283	274
15	19.9	20.1	20.1	275	281	282
16	19.8	19.9	20.1	281	277	280
17	20.1	20.1	20.3	283	287	283
18	19.9	20.1	20.1	276	283	280
19	20.0	20.2	20.2	277	277	281
20	20.2	20.1	20.1	282	279	282

a. Hãy xây dựng và diễn dịch các đồ thị S và các đồ thị R cho trọng lượng và thể tích của ngũ cốc trong các gói.

b. Đồ thị x trung bình được vẽ chỉ khi sự phân tán của quá trình nằm trong sự kiểm soát. Trong trường hợp này có nên vẽ đồ thị x trung bình cho trọng lượng và thể tích của ngũ cốc trong gói hay không? Nếu có, hãy vẽ các đồ thị. Khi vẽ hãy xét tất cả 8 loại kiểm tra cho dạng thay đổi của quá trình.

c. Hãy vẽ đồ thị trọng lượng đối với thể tích (hay ngược lại). Tính hệ số tương quan. Có mối quan hệ giữa 2 biến này hay không? (Sinh viên trả lời câu này tham khảo thêm kiến thức *Phân Tích Hồi Qui*)

4.3. Dữ liệu sau được thu thập để giám sát lượng của một loại nước cola uống kiêng đóng trong chai loại 16 ounces. Trong 24 tiếng đồng hồ, cứ mỗi tiếng lấy liên tục 4 chai. (File: DietCola.mtp)

Giờ	**Lượng Cola đo được**				**Giờ**	**Lượng Cola đo được**			
1	16.01	16.03	15.98	16.00	13	15.96	16.00	16.01	16.00
2	16.03	16.02	15.97	15.99	14	15.98	16.01	16.02	15.99
3	15.98	16.00	16.03	16.04	15	15.99	16.03	16.00	15.98
4	16.00	16.03	16.02	15.98	16	16.02	16.02	16.01	15.97
5	15.97	15.99	16.03	16.01	17	16.01	16.05	15.99	15.99
6	16.01	16.03	16.04	15.97	18	15.98	16.03	16.04	15.98
7	16.04	16.05	15.97	15.96	19	15.97	15.96	15.99	15.99
8	16.02	16.05	16.03	15.97	20	16.03	16.01	16.04	15.96
9	15.97	15.99	16.02	16.03	21	15.99	16.03	15.97	16.05
10	16.00	16.01	15.95	16.04	22	15.98	15.95	16.07	16.01

11	15.95	16.04	16.07	15.93	23	15.99	16.06	15.95	16.03
12	15.98	16.07	15.94	16.08	24	16.00	16.01	16.08	15.94

a. Hãy xây dựng và diễn dịch đồ thị S, R, x trung bình, sử dụng các thông số của quá trình là $\mu = 16$ ounce và $\sigma = 0,05$

b. Hãy xây dựng và diễn dịch đồ thị S, R, x trung bình, sử dụng dữ liệu để ước lượng μ và σ. So sánh các đồ thị này với các đồ thị trong phần a.

4.4. Một công ty điện tử sản xuất màn hình máy tính. Mỗi ngày, 50 màn hình máy tính được chọn ngẫu nhiên để xem số màn hình bị khuyết tật, việc này được thực hiện trong 21 ngày. Bảng sau cho biết số màn hình bị lỗi mỗi ngày. Hãy xây dựng và diễn dịch đồ thị P cho quá trình sản xuất. (bài này không có file, sinh viên tự nhập dữ liệu)

Ngày	1	2	3	4	5	6	7	8	9	10	11
Số màn hình khuyết tật	3	7	4	2	1	4	4	6	1	5	7

Ngày	12	13	14	15	16	17	18	19	20	21
Số màn hình khuyết tật	15	7	4	3	3	8	7	2	5	4

4.5. Lỗi in ấn được phát hiện trong nhiều cuốn sách. Bảng sau cho biết tổng số lỗi in ấn trong 10 trang sách được chọn ngẫu nhiên được thực hiện từ 25 cuốn sách mới xuất bản bởi một nhà xuất bản. Hãy xây dựng và diễn dịch đồ thị C đối với số lỗi trong 10 trang của một cuốn sách. Các giới hạn kiểm soát thu được có thể được dùng cho dữ liệu tương lai hay không? Hãy giải thích. (bài này không có file, sinh viên tự nhập dữ liệu)

Cuốn sách	1	2	3	4	5	6	7	8	9	10	11	12	13
Số lỗi in	7	6	6	7	4	7	8	12	9	9	8	5	5

Cuốn sách	14	15	16	17	18	19	20	21	22	23	24	25
Số lỗi in	9	8	15	6	4	13	7	8	15	6	6	10

4.6. Dựa vào bài số 5. Bảng sau cho biết lỗi in ấn của mỗi cuốn đối với 25 cuốn sách được chọn tiếp theo. Hãy thêm những dữ liệu này vào đồ thị C đã vẽ trong bài 5. Hãy sử dụng dữ liệu từ 25 cuốn sách đầu tiên để xác định các giới hạn kiểm soát. Quá trình CÓ CÒN trong sự kiểm soát hay không?

Cuốn sách	26	27	28	29	30	31	32	33	34	35	36	37	38
Số lỗi in	7	13	4	5	9	3	4	6	7	14	18	11	11

Cuốn sách	39	40	41	42	43	44	45	46	47	48	49	50
Số lỗi in	11	8	10	8	7	16	13	12	9	11	11	8

4.7. Một nhà sản xuất các ổ đĩa cứng liên tục theo dõi các ổ đĩa để tìm vết rạn nứt trên bề mặt lưu trữ. Các vết rạn nứt sẽ làm giảm không gian lưu trữ của đĩa. Mỗi giờ một đĩa cứng sẽ được chọn và xác định số rạn nứt. Bảng sau cho biết kết quả của 50 mẫu. (File: DiskDrives.mtp). Hãy xây dựng và diễn dịch đồ thị C đối với số vết nứt của mỗi ổ đĩa

Mẫu	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Số vết	8	11	7	11	10	10	8	9	6	14	11	8	11	9	18

Mẫu	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Số vết	13	7	14	8	9	17	11	12	12	16	10	20	13	12	22

Mẫu	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
Số vết	14	8	12	6	11	6	10	9	8	10	13	11	9	14	15

Mẫu	46	47	48	49	50
Số vết	11	12	11	6	13