

**Chương trình giảng dạy kinh tế Fulbright
Học kỳ Thu, 2010**

CÁC PHƯƠNG PHÁP PHÂN TÍCH ĐỊNH LƯỢNG

**GỢI Ý LỜI GIẢI BÀI TẬP 1
THỐNG KÊ MÔ TẢ
và phụ lục thực hành thống kê mô tả với phần mềm SPSS**

Bài 1 (25 điểm)

Số liệu thu nhập hàng tuần (đơn vị: ngàn đồng) của hai nhóm công nhân như sau:

Nhóm 1: 510, 580, 600, 640, 670

Nhóm 2: 580, 595, 600, 605, 620

- a. Dựa vào định nghĩa và công thức, hãy tính các giá trị trung bình và trung vị của thu nhập trong mỗi nhóm

Khi chưa rõ số liệu là của mẫu hay của tổng thể, bạn có thể giả định!

Giả sử, chúng ta hiểu rằng đây là dữ liệu của mẫu. Chúng ta sẽ tính trung bình, trung vị của thu nhập cho mỗi nhóm. Nếu bạn hiểu rằng tổng thể là 10 công nhân trong 2 nhóm trên thì cũng không sai, khi đó bạn cần áp dụng các công thức tính các tham số thống kê của tổng thể.

• Số trung bình

Thu nhập trung bình của mỗi công nhân ở nhóm 1 được tính theo công thức

$$\bar{x}_{nhóm1} = \frac{\sum_{i=1}^5 x_{Nhóm1_i}}{5} = \frac{510 + 580 + 600 + 640 + 670}{5} = \frac{3000}{5} = 600 \text{ ngàn đồng}$$

Thu nhập trung bình của mỗi công nhân ở nhóm 2 được tính theo công thức

$$\bar{x}_{nhóm2} = \frac{\sum_{i=1}^5 x_{Nhóm2_i}}{5} = \frac{580 + 595 + 600 + 605 + 620}{5} = \frac{3000}{5} = 600 \text{ ngàn đồng}$$

• Số trung vị

Số lượng số liệu có trong mỗi nhóm là số lẻ, và chúng đã được sắp xếp từ nhỏ nhất đến lớn nhất. Vì vậy, với mỗi nhóm, số trung vị chính là giá trị nằm ở vị trí chính giữa (trong trường hợp này là ở vị trí thứ $(n+1)/2 = (5+1)/2 = 3$)

Trung vị của thu nhập đối với Nhóm 1 là 600 ngàn đồng

Trung vị của thu nhập đối với Nhóm 2 cũng là 600 ngàn đồng

b. Dựa vào định nghĩa và công thức, hãy tính các giá trị Min, Max, Range, phương sai và độ lệch chuẩn của thu nhập trong mỗi nhóm

• **Min, Max, và khoảng biến thiên**

Nhìn vào dãy dữ liệu đã được sắp xếp từ nhỏ đến lớn của mỗi nhóm, chúng ta dễ dàng xác định được Min, Max của mỗi nhóm, và từ đó áp dụng công thức $\text{Range} = \text{Max} - \text{Min}$ để tính ra khoảng biến thiên của mỗi nhóm

Với Nhóm 1, giá trị nhỏ nhất của thu nhập là 510 ngàn đồng (Min=510 ngàn đồng), giá trị lớn nhất của thu nhập là 670 ngàn đồng (Max=670 ngàn đồng), khoảng biến thiên của thu nhập là 160 ngàn đồng (bằng 670-510).

Với Nhóm 2, giá trị nhỏ nhất của thu nhập là 580 ngàn đồng (Min=580 ngàn đồng), giá trị lớn nhất của thu nhập là 620 ngàn đồng (Max=620 ngàn đồng), khoảng biến thiên của thu nhập là 40 ngàn đồng (bằng 620-580)

• **Phương sai và độ lệch chuẩn của mỗi nhóm**

Nhóm 1:

Chúng ta có thể áp dụng công thức, lập ra bảng sau để tính phương sai, độ lệch

i	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	510	-90	8100
2	580	-20	400
3	600	0	0
4	640	40	1600
5	670	70	4900
Tổng	3000		15000

Trung bình 600
Phương sai 3750
Độ lệch 61.24

$$S_{\text{Nhóm1}}^2 = \frac{(x_i - \bar{x})^2}{n - 1} = \frac{(510 - 600)^2 + (580 - 600)^2 + \dots + (670 - 600)^2}{5 - 1} = \frac{15000}{4} = 3750$$

ngàn đồng²

$$S_{\text{Nhóm1}} = \sqrt{S_{\text{Nhóm1}}^2} = \sqrt{3750} = 61.24 \text{ ngàn đồng}$$

Nhóm 2:

i	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	580	-20	400
2	595	-5	25
3	600	0	0
4	605	5	25
5	620	20	400
Tổng	3000		850
Trung bình	600		
Phương sai	212.5		
Độ lệch	14.58		

$$S_{Nhóm 2}^2 = \frac{(x_i - \bar{x})^2}{n - 1} = \frac{(580 - 600)^2 + (595 - 600)^2 + \dots + (620 - 600)^2}{5 - 1} = \frac{850}{4} = 212.5$$

ngàn đồng²

$$S_{Nhóm 2} = \sqrt{S_{Nhóm 2}^2} = \sqrt{212.5} = 14.58 \text{ ngàn đồng}$$

c. Dựa vào các hàm trong Excel, hãy tính các đại lượng thống kê ở câu a và câu b

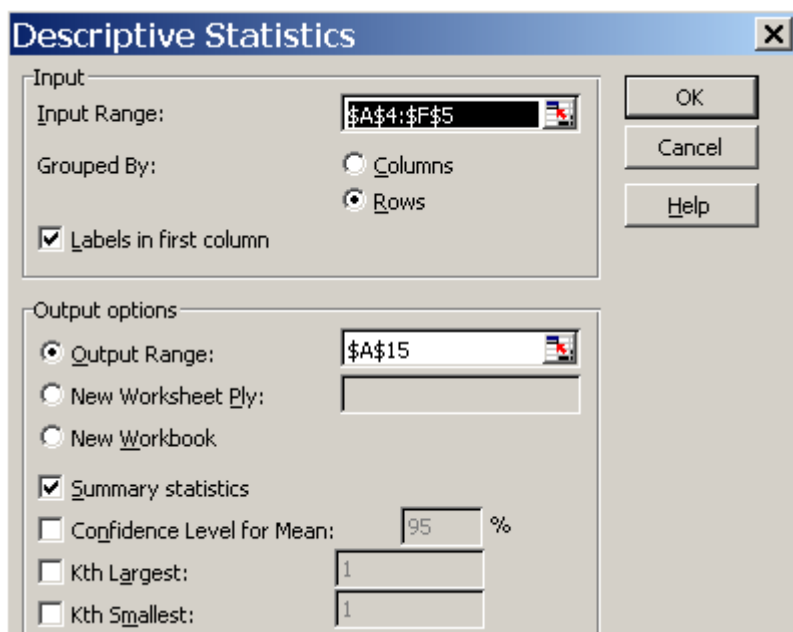
Bạn có thể sử dụng các hàm thống kê (Xem Hình 1.1); hoặc công cụ **Tools\Data Analysis\Descriptive Statistics** của Excel, sau đó khai báo tương tự như Hình 1.2 để tính toán các chỉ tiêu cho từng nhóm

Hình 1.1 Sử dụng hàm số

	A	B	C	D	E	F
1						
2	Bảng 1. Thu nhập của công nhân (ngàn đồng) ở Nhóm 1, và Nhóm 2					
3		Công nhân 1	Công nhân 2	Công nhân 3	Công nhân 4	Công nhân 5
4	Thu nhập của công nhân ở Nhóm 1	510	580	600	640	670
5	Thu nhập của công nhân ở Nhóm 2	580	595	600	605	620
6						
7	Tính một số thống kê mô tả bằng cách sử dụng hàm trong Excel					
8		Nhóm 1	Nhóm 2			
9	Mean	=AVERAGE(B4:F4)	=AVERAGE(B5:F5)			
10	Min	=MIN(B4:F4)	=MIN(B5:F5)			
11	Max	=MAX(B4:F4)	=MAX(B5:F5)			
12	Median	=MEDIAN(B4:F4)	=MEDIAN(B5:F5)			
13	Range	=B11-B10	=C11-C10			
14	Standard Deviation	=STDEV(B4:F4)	=STDEV(B5:F5)			
15	Variance	=VAR(B4:F4)	=VAR(B5:F5)			

Với các hàm như trên, và thao tác như Hình 1.2 bạn sẽ có được kết quả như Hình 1.3

Hình 1.2



Hình 1.3 Kết quả tính toán từ hàm thống kê, và công cụ Data Analysis

	A	B	C	D
6				
7	Tính một số thống kê mô tả bằng cách sử dụng hàm trong Excel			
8		Nhóm 1	Nhóm 2	
9	Mean	600	600	
10	Min	510	580	
11	Max	670	620	
12	Median	600	600	
13	Range	160	40	
14	Standard Deviation	61.24	14.58	
15	Variance	3750	212.5	
16				
17	Kết quả một số thống kê được tính từ công cụ Tools\Data Analysis\Descriptive Statistics			
18		Nhóm 1	Nhóm 2	
19				
20	Mean	600.00		600.00
21	Standard Error	27.39		6.52
22	Median	600.00		600.00
23	Mode	#N/A		#N/A
24	Standard Deviation	61.24		14.58
25	Sample Variance	3750.00		212.50
26	Kurtosis	0.21		0.89
27	Skewness	-0.60		0.00
28	Range	160.00		40.00
29	Minimum	510.00		580.00
30	Maximum	670.00		620.00
31	Sum	3000.00		3000.00
32	Count	5.00		5.00

d. Anh/Chị có nhận xét gì về thu nhập của hai nhóm công nhân này.

Trong mẫu, thu nhập trung bình của hai nhóm công nhân là bằng nhau, và đều bằng 600 ngàn đồng. Tuy nhiên, thu nhập của những công nhân ở nhóm 2 ít biến thiên hơn so với nhóm 1; thật vậy, thu nhập của công nhân ở nhóm 1 dao động từ 510 ngàn đồng đến 670 ngàn đồng với khoảng biến thiên là 160; trong khi đó thu nhập của công nhân ở nhóm 2 có khoảng biến thiên nhỏ hơn, và bằng 40 (dao động từ 580 ngàn đồng đến 620 ngàn đồng).

Bạn cũng có thể sử dụng thêm chỉ tiêu độ lệch, phương sai, hay tính toán thêm hệ số biến thiên ... để nói lên rằng thu nhập của công nhân ở nhóm 2 ổn định hơn so với thu nhập của các công nhân ở nhóm 1. Vì cỡ mẫu của hai nhóm đều bằng 5, thêm vào đó, trung bình của hai nhóm đều như nhau, nên có thể so sánh độ lệch chuẩn của mỗi nhóm. Độ lệch chuẩn của nhóm 2 là 14.58 ngàn đồng nhỏ hơn so với độ lệch chuẩn của nhóm 1 (độ lệch chuẩn của nhóm một là 61.24 ngàn đồng). Điều này càng cho thấy mức độ biến thiên về thu nhập của nhóm 2 ít hơn so với nhóm 1.

Bài 2 (25 điểm)

Tập tin DataRestaurant300.xls ghi nhận dữ liệu về chất lượng bữa ăn và giá bữa ăn tại 300 nhà hàng. Từ bảng dữ liệu này, hãy trả lời những câu hỏi sau:

a. Phân biệt loại biến và loại thang đo cho 2 biến số chất lượng bữa ăn và giá bữa ăn

Hình 2.1

	A	B	C
1	Restaurant	Quality Rating	Meal Price (\$)
2	1	Good	18
3	2	Very Good	22
4	3	Good	28
5	4	Excellent	38
6	5	Very Good	33
7	6	Good	28
8	7	Very Good	19
9	8	Very Good	11
10	9	Very Good	23
11	10	Good	13
12	11	Very Good	33

-Biến Chất lượng bữa ăn (Quality Rating) hiện chỉ có 3 biểu hiện là Good, Very Good, và Excellent thể hiện tính chất của bữa ăn. Vì vậy, có thể được xem là biến định tính. Biến này có thang đo thứ tự (hay thứ bậc)

-Biến giá bữa ăn (Meal Price) là biến định lượng, và có thang đo tỷ lệ (nếu bạn trả lời là thang đo khoảng thì cũng không sai!)

⇒ **Mở rộng** (không tính vào điểm bài tập): Với biến định lượng bạn có thể nhận định thêm xem nó là biến định lượng rời rạc hay liên tục. Và cũng chú ý rằng, định lượng hay định tính cũng chỉ mang tính chất tương đối!

b. Hãy phân nhóm nhà hàng theo chất lượng bữa ăn. Vẽ biểu đồ tần số theo chất lượng bữa ăn

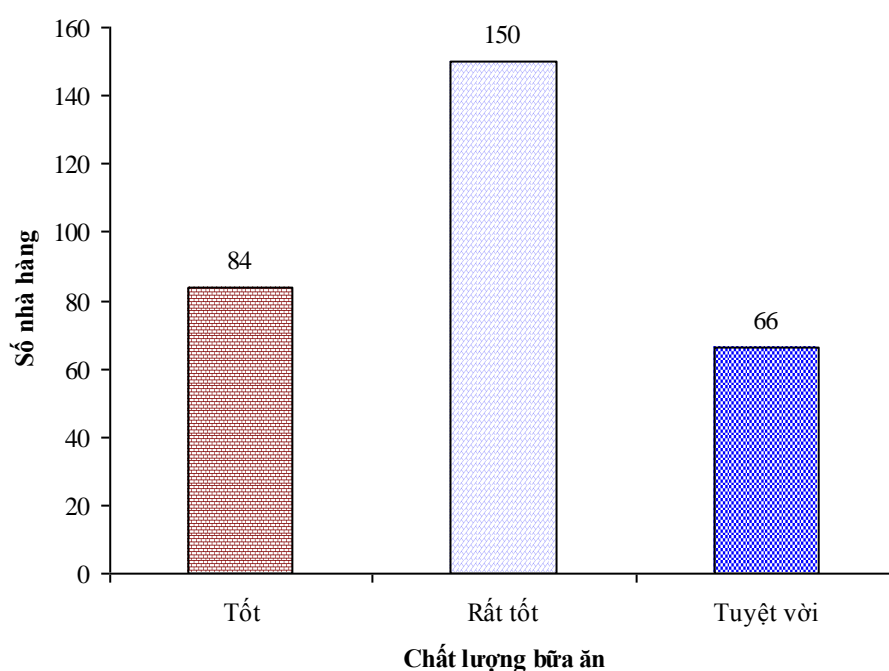
Sử dụng kỹ thuật Data\Pivot Table trong Excel, bạn dễ dàng tạo ra được bảng Kết quả

Hình 2.2

	A	B	C	D
1				
2				
3	Count of Quality Rating			
4	Quality Rating	Total		
5	Excellent	66		
6	Good	84		
7	Very Good	150		
8	Grand Total	300		

Từ dữ liệu đã có, Sử dụng Excel bạn có thể vẽ được biểu đồ tần số như sau

Hình 2.3 Biểu đồ tần số về chất lượng bữa ăn



c. Tìm các giá trị cực đại, cực tiểu, trung bình, trung vị và yếu vị của giá tiền bữa ăn

Với kỹ thuật **Tools\Data Analysis\Descriptive Statistics** bạn có thể dễ dàng tìm được các giá trị cực đại, cực tiểu, trung bình, trung vị, và yếu vị của giá tiền bữa ăn

Hình 2.4

	A	B	C	D	E	F	G	H
1	Restaurant	Quality Rating	Meal Price (\$)					
2	1	Good	18					
3	2	Very Good	22					
4	3	Good	28					
5	4	Excellent	38					
6	5	Very Good	33					
7	6	Good	28					
8	7	Very Good	19					
9	8	Very Good	11					
10	9	Very Good	23					
11	10	Good	13					
12	11	Very Good	33					
13	12	Very Good	44					
14	13	Excellent	42					
15	14	Excellent	34					
16	15	Good	25					
17	16	Good	22					
18	17	Good	26					

Bảng 2.1

<i>Meal Price (\$)</i>	
Mean (Trung bình)	25.90
Standard Error (Sai số chuẩn của giá trị trung bình)	0.54
Median (Trung vị)	25.00
Mode (Yếu vị)	20.00
Standard Deviation (Độ lệch chuẩn)	9.29
Sample Variance (Phương sai)	86.26
Kurtosis (<i>chỉ tiêu Kurtosis đo lường độ nhọn của phân phối; Trong Excel & Eviews, Kurtosis=3 thì phân phối bình thường, Kurtosis>3 thì phân phối nhọn</i>)	-0.58
Skewness (<i>chỉ tiêu Skewness đo lường độ cân xứng của phân phối; Phân phối cân xứng khi Skewness=0, Phân phối lệch phải khi Skewness>0,</i>)	0.31
Range (Khoảng biến thiên)	38.00
Minimum (Giá trị nhỏ nhất)	10.00
Maximum (Giá trị lớn nhất)	48.00
Sum (Tổng)	7769.00
Count (Số quan sát)	300.00

⇨ **Mở rộng** (không tính vào điểm bài tập): Để nhận định về độ cân xứng của phân phối trong tổng thể, một cách thuận tiện, theo kinh nghiệm, nếu Skewness nằm trong đoạn $[-1,1]$ thì có thể nhận định rằng phân phối cân xứng.

d. Có bao nhiêu nhà hàng có chất lượng bữa ăn được xếp loại chất lượng “very good” và giá bữa ăn vào khoảng từ \$30 đến \$39

Bạn có thể sử dụng công cụ lọc dữ liệu, kỹ thuật **Data\Pivot Table** trong Excel; hàm countif; sử dụng các hàm đơn giản if và sum...; hay sử dụng hàm cơ sở dữ liệu DCOUNT để tính ra kết quả này.

Hình 2.5

	A	B	C	D	E	F	G
1	Restaurant	Quality Rating	Meal Price (\$)				
2	1	Good	18		Quality Rating	Meal Price (\$)	Meal Price (\$)
3	2	Very Good	22		Very Good	>=30	<=39
4	3	Good	28				
5	4	Excellent	38				
6	5	Very Good	33		46		
7	6	Good	28				
8	7	Very Good	19				
9	8	Very Good	11				
10	9	Very Good	23				

Với hàm DCOUNT bạn có thể làm như sau: (1) Tạo ra vùng điều kiện như E2:G3, (2) Tại một ô nào đó, ví dụ ô E6, gõ lệnh =DCOUNT(A1:C301,A1,E2:G3), bạn sẽ ra kết quả là 46. Nói cách khác có 46 nhà hàng có chất lượng bữa ăn là “Very Good” và giá nằm trong khoảng từ \$30 đến \$39.

Bài 3 (25 điểm)

Tập tin DataExamScores.xls ghi nhận dữ liệu về điểm thi (tính trên thang điểm 100) của hai trung tâm đào tạo A và B. Mỗi trung tâm có 30 sinh viên được thăm dò. Từ bảng dữ liệu này, hãy trả lời những câu hỏi sau:

a. Hãy tìm trung bình và trung vị của điểm thi ở trung tâm A

Sử dụng công cụ Tools\Data Analysis\Descriptive Statistics của Excel, chúng ta dễ dàng có được bảng kết quả như Bảng 31.

Từ bảng này, ta thấy:

-Trung bình điểm thi của sinh viên ở trung tâm A là 82 điểm

-Trung vị điểm thi của sinh viên ở trung tâm A là 83 điểm

Bảng 3.1

<i>Center A</i>	
Mean	82.00
Standard Error	1.68
Median	83.00
Mode	79.00
Standard Deviation	9.22
Sample Variance	85.03
Kurtosis	-0.12
Skewness	-0.43
Range	37.00
Minimum	60.00
Maximum	97.00
Sum	2460.00
Count	30.00

- b. So sánh các giá trị tính được trong câu a, Anh/Chị có kết luận rằng phân phối của điểm thi ở trung tâm A có bị lệch xiên không. Nếu lệch xiên thì lệch về trái hay phải. Hãy giải thích ngắn gọn câu trả lời của Anh/Chị

Vị trí giữa trung bình, trung vị có thể cho chúng ta biết hình dạng phân phối của biến là lệch trái, lệch phải hay cân xứng. Điểm thi của sinh viên ở trung tâm A có số trung bình nhỏ hơn trung vị nên phân phối điểm thi của sinh viên ở trung tâm A bị lệch xiên về phía trái.

- c. Sử dụng hàm trong Excel, xác định độ lệch xiên ở câu b. Kết quả ở câu c có phù hợp với câu b hay không

Hình 3.1

	A	B	C	D
1	Center A	Center B		
2	97	64		
3	95	85		
4	89	72	=SKEW(A2:A31)	-0.433
5	79	64		
6	78	74		

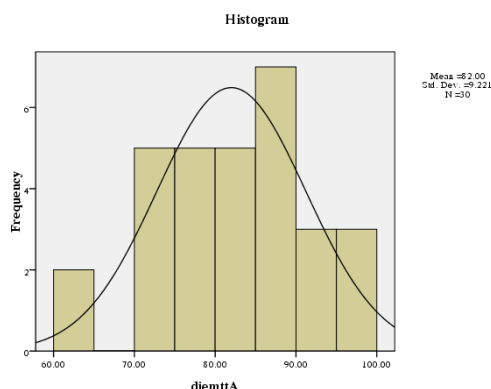
Hệ số Skewness được tính từ hàm Skew trong Excel có kết quả là -0.433. Skew nhỏ hơn 0 nên hệ số này cũng chỉ ra rằng phân phối điểm của sinh viên ở trung tâm A bị lệch trái.

➔ **Mở rộng** (không tính vào điểm bài tập): Chú ý rằng, những nhận định về hình dáng của phân phối ở câu b, câu c là xét cho mẫu được khảo sát. Khi muốn kết luận về hình dạng phân phối xét trên tổng thể, chúng ta cần thực hiện những kiểm định phức tạp hơn. Bạn có thể đọc thêm về kiểm định Jarque-Bera, kiểm định Kolmogorov-Smirnov ...

		dienttA
N		30
Normal Parameters ^a	Mean	82.0000
	Std. Deviation	9.22141
Most Extreme Differences	Absolute	.100
	Positive	.052
	Negative	-.100
Kolmogorov-Smirnov Z		.548
Asymp. Sig. (2-tailed)		.925
a. Test distribution is Normal.		

P-value của thống kê Kolmogorov-Smirnov Z bằng 0.128 (>0.05) nên ở độ tin cậy 95% có thể chấp nhận giả thuyết cho rằng biến điểm thi của trung tâm A có phân phối chuẩn. Nói cách khác, sự lệch xiên của phân phối là không đáng kể.

Nếu dùng SPSS, bạn sẽ dễ dàng vẽ biểu đồ sau để có cái nhìn trực quan về phân phối của biến này



- d. Tìm giá trị chuẩn hóa Z cho giá trị quan sát lớn nhất và nhỏ nhất của điểm thi ở trung tâm B. Các giá trị này có lớn hay nhỏ bất thường không

Bảng 3.2

<i>Center B</i>	
Mean	78.23
Standard Error	1.92
Median	78.00
Mode	78.00
Standard Deviation	10.50
Sample Variance	110.25
Kurtosis	-0.61
Skewness	0.06
Range	42.00
Minimum	57.00
Maximum	99.00
Sum	2347.00
Count	30.00

Trước tiên, bạn tính độ lệch, trung bình, và xác định giá trị nhỏ nhất, giá trị lớn nhất của điểm thi ở trung tâm B.

Điểm thi của trung tâm B có Max=99 điểm , Min=57 điểm , độ lệch chuẩn s=10.5 điểm, \bar{x} =78.23 điểm.

Giá trị chuẩn hoá Z-score của x ở quan sát thứ i được tính bởi công thức $z_i = \frac{x_i - \bar{x}}{s}$

Giá trị lớn nhất có giá trị chuẩn hoá là $z = \frac{99 - 78.23}{10.05} = 1.98$. Giá trị lớn nhất, số 99, có $|z| < 2$ nên là giá trị bình thường, không phải là dị biệt.

Giá trị lớn nhỏ nhất có giá trị chuẩn hoá là $z = \frac{57 - 78.23}{10.05} = -2.02$. Giá trị nhỏ nhất, số 57, có $2 < |z| < 3$ nên có thể là giá trị bất thường.

Bài 4 (25 điểm)

Bảng dữ liệu dưới đây trình bày tên, tốc độ và giá của 10 loại máy in

STT	Tên máy in	Tốc độ	Giá (USD)
1	Minolta-QMS PagePro 1250W	12	199
2	Brother HL-1850	10	499
3	Lexmark E320	12.2	299
4	Minolta-QMS PagePro 1250E	10.3	299
5	HP Laserjet 1200	11.7	399
6	Xerox Phaser 4400/N	17.8	1850
7	Brother HL-2460N	16.1	1000
8	IBM Infoprint 1120n	11.8	1387
9	Lexmark W812	19.8	2089
10	Oki Data B8300n	28.2	2200

- a. Tính các giá trị trung bình, trung vị, yếu vị, phương sai, độ lệch chuẩn của 2 biến X là tốc độ và Y là giá

Sử dụng công cụ Tools\Data Analysis bạn sẽ tính được giá trị trung bình, trung vị, yếu vị, phương sai, độ lệch chuẩn của 2 biến X và Y được tính toán như sau:

Bảng 4.1

		X	Y
Mean	Trung bình	14.99	1022.10
Standard Error	Độ lệch chuẩn	1.80	252.32
Median	Trung vị	12.10	749.50
Mode	Yếu vị	#N/A	299.00
Standard Deviation	Độ lệch chuẩn	5.69	797.91
Sample Variance	Phương sai	32.38	636656.77
Count		10	10

Ghi chú: biến X không xác định được yếu vị do các giá trị của X đều có cùng một tần số

- b. Theo nhận định của Anh/Chị mối quan hệ giữa 2 biến X và Y là đồng biến hay nghịch biến và độ mạnh của mối quan hệ này. Giải thích ngắn gọn về nhận định của Anh/Chị.

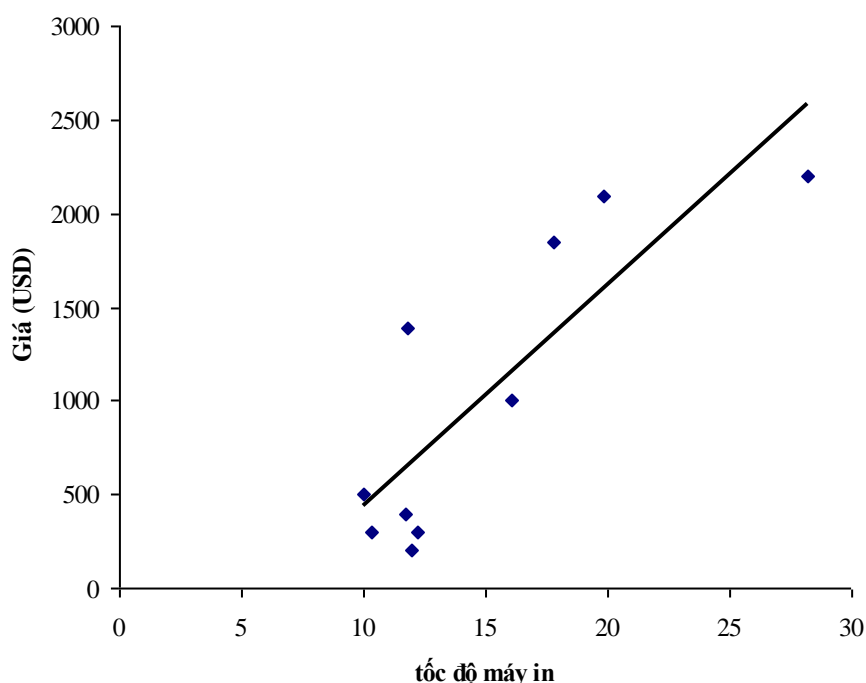
Nhận định này mang tính suy luận của bạn, có thể có nhiều cách giải thích khác nhau. Ví dụ như: X (tốc độ máy in) và Y (giá) thường có mối quan hệ đồng biến, bởi vì người tiêu dùng luôn mong muốn chất lượng máy in tốt (tốc độ in nhanh, ít tốn mực, in đẹp ...), và họ sẵn lòng trả giá cao hơn cho những máy in có chất lượng tốt hơn. Bên cạnh đó, người sản xuất luôn phải cải tiến công nghệ để có thể đáp ứng tốt hơn nhu cầu này của khách hàng, và dĩ nhiên với những điều kiện khác không đổi, những máy in tốt hơn thường tốn chi phí sản xuất cao hơn và vì vậy người sản xuất cũng mong bán những sản phẩm này với giá cao hơn.

Theo ý kiến chủ quan, có thể giữa hai biến X và Y có mối quan hệ chặt, bởi tốc độ in cũng là điều mà nhiều người tiêu dùng quan tâm, và họ sẵn lòng trả giá cao hơn cho những máy in có tốc độ in nhanh.

⇒ **Mở rộng** (Không tính vào điểm bài tập): Nếu bạn tìm được cơ sở lý thuyết, các nghiên cứu thực nghiệm trước đây về vấn đề này, thì những nhận định của bạn càng thuyết phục hơn

- c. Vẽ biểu đồ phân tán (Scatter) thể hiện mối quan hệ giữa 2 biến X và Y và nhận xét về mối quan hệ giữa 2 biến này. Nhận xét này có phù hợp với nhận định của Anh/Chị ở câu b hay không.

Hình 2.1



Đồ thị phân tán ở Hình 2.1 cho thấy giữa X và Y có mối quan hệ đồng biến, những máy in có tốc độ cao thường có giá cao, và những máy in có giá thấp cũng thường là những máy in có tốc độ in thấp.

Thêm đó, trên đồ thị phân tán, các điểm nằm khá gần đường xu thế (được Add Trendline vào đồ thị) nên có thể nói rằng mối quan hệ tuyến tính giữa hai biến X và Y là chặt chẽ (tương quan cao). Những nhận xét này phù hợp với những nhận định ban đầu ở câu b.

- d. Tính các giá trị đồng phương sai và hệ số tương quan giữa 2 biến X và Y này. Có nhận xét gì về mối quan hệ của 2 biến này từ các hệ số tính được và so sánh nó với nhận xét ở câu b và c.

Hình 4.2

	A	B	C	D	E	F	G	H	I	J
1	i	X	Y	(Xi-X̄)	(Yi-Ȳ)	(Xi-X̄)(Yi-Ȳ)		Sử dụng hàm để tính hiệp phương sai tổng thể, và hs tương quan		
2	1	12.0	199.0	-3.0	-823.1	2461.1		Hiệp phương sai tổng thể	=COVAR(C2:C11,B2:B11)	3435.98
3	2	10.0	499.0	-4.99	-523	2610.3		Hệ số tương quan	=CORREL(C2:C11,B2:B11)	0.84
4	3	12.2	299.0	-2.79	-723	2017.4				
5	4	10.3	299.0	-4.69	-723	3391.3		Bảng hệ số hiệp phương sai		
6	5	11.7	399.0	-3.29	-623	2050.0			X̄	Ȳ
7	6	17.8	1850.0	2.81	827.9	2326.4	X		29.14	
8	7	16.1	1000.0	1.11	-22.1	-24.5	Y		3435.98	572991.09
9	8	11.8	1387.0	-3.19	364.9	-1164.0				
10	9	19.8	2089.0	4.81	1067	5131.8		Bảng hệ số tương quan		
11	10	28.2	2200.0	13.21	1178	15560.1			X̄	Ȳ
12	Tổng	149.9	10221.0			34359.8	X		1	
13	Trung bình	15.0	1022.1				Y		0.84	1
14										
15	Hiệp phương sai=	34359.8/9		=		3817.76	Hiệp phương sai mẫu	=345.98x10/9=		3817.76

Có thể tính hệ số hiệp phương sai, hệ số tương quan bằng hàm hay bằng công cụ Tools\Data Analysis của Excel. (Xem Hình 4.2)

Hệ số hiệp phương sai (đồng phương sai) giữa X và Y bằng **3817.76** (lớn hơn 0) cho thấy giữa X và Y có mối quan hệ đồng biến. Tuy nhiên, hệ số hiệp phương sai chưa cho biết được độ mạnh của mối quan hệ giữa hai biến này. Hệ số tương quan giữa X và Y, $r_{X,Y}=0.84$ (lớn hơn 0) nên X và Y có mối liên hệ tuyến tính. Thêm vào đó $|r_{X,Y}|$ gần về phía 1 nên mối quan hệ giữa hai biến này là chặt chẽ, hay mạnh. Như vậy, Các số liệu khẳng định mạnh hơn những nhận định ở câu b và c.

↔ **Mở rộng** (không tính vào điểm bài tập): Bạn có thể áp dụng một quy tắc kinh nghiệm nào đó để chỉ ra mối quan hệ này là rất chặt (rất mạnh), hay chặt (mạnh)... Chú ý rằng có nhiều quy tắc kinh nghiệm về việc này, tùy thuộc vào kiểu dữ liệu (chuỗi thời gian, chéo, hay bảng), hay lĩnh vực khoa học (khoa học xã hội, khoa học tự nhiên, kỹ thuật công nghệ, y học ...), hay đối tượng khảo sát (hộ, doanh nghiệp, quốc gia...). Khi sử dụng quy tắc nào, bạn có thể trích dẫn nguồn quy tắc ấy. Sau đây là một ví dụ.

Trị tuyệt đối của hệ số tương quan Pearson	Mức độ tương quan
≤ 0.2	Rất yếu, tương quan không đáng kể
0.2 đến 0.4	Yếu, tương quan thấp
0.4 đến 0.7	Vừa phải
0.7 đến 0.9	Mạnh, tương quan cao
> 0.9	Rất mạnh

Nguồn: http://www.bized.co.uk/timeweb/crunching/crunch_relate_expl.htm, truy cập ngày 5/10/2010

Phụ lục. Bài đọc thêm về thao tác trên SPSS để tóm tắt và trình bày dữ liệu

Trong môn học Các phương pháp định lượng, về mặt kỹ năng sử dụng phần mềm trong thống kê, kinh tế lượng, nhóm giảng viên kỳ vọng bạn có thể sử dụng được Excel, Eviews, SPSS. Phần mềm Excel, Eviews sẽ được hướng dẫn nhiều trên lớp học; Bài đọc thêm này sẽ hỗ trợ bạn sử dụng SPSS trong việc làm những bài tập đơn giản liên quan đến chủ đề mô tả các tập dữ liệu; và các thao tác, kết quả tính toán cũng sẽ tương tự như vậy cho những dữ liệu lớn hơn mà bạn sẽ gặp sau này.

• **Khai báo biến**

Để tính các thống kê mô tả trong **Bài tập 1, Bài 1**, bạn cũng có thể tính toán bằng SPSS.

Trước tiên, bạn hãy khởi động SPSS, Tại cửa sổ Quản lý dữ liệu **Data Editor**, Chọn Thẻ **Variable View**, sau đó **khai báo các thuộc tính của biến** bằng cách gõ tên biến vào cột **Name** (tên biến cần viết liền, không có khoảng trắng, không có ký tự đặc biệt, bắt đầu bằng chữ, và mỗi biến có một tên khác nhau ...), gõ nhãn biến vào cột **Label**. Nếu cần thận hơn, bạn hãy chọn Values để quy ước nhãn các giá trị của biến định tính (ví dụ biến Nhóm, với 1 là Nhóm 1, 2 là Nhóm 2), và khai báo đúng thang đo của biến ở cột **Measure** (trong cột này, mặc định là Scale cho những biến định lượng: những biến có thang đo khoảng, hoặc tỷ lệ; Nominal cho những biến có thang đo định danh, và Ordinal cho những biến có thang đo thứ tự)

Hình 1.1

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	Thunhap	Numeric	8	2	Thu nhap cua cong nhan ...	None	None	8	≡ Ri...	Scale
2	Nhom	Numeric	8	0	Nhom	1, Nh...	None	8	≡ Ri...	Nominal
3										Scale
4										Ordinal
5										Nominal
6										
7										
8										

Cửa sổ Quản lý dữ liệu

Kiểu dữ liệu dạng số (mặc định)

Gõ nhãn của biến vào đây, nhãn có thể có khoảng trắng, có thể gõ dấu tiếng Việt...

Gõ tên biến vào đây

Bấm vào đây để quy ước nhãn cho các giá trị của biến định tính. Ví dụ: 1 tương ứng với Nhóm 1, 2 tương ứng với Nhóm 2

Data View dùng để xem dữ liệu, nhập liệu, chỉnh sửa dữ liệu ...

Variable View này để khai báo các thuộc tính của biến

Measure: Khai báo thang đo cho biến

- **Nhập dữ liệu**

Hình 1.2

The screenshot shows the SPSS Data Editor window with a data table. The table has two columns: 'Thunhap' and 'Nhom'. The data is as follows:

	Thunhap	Nhom
1	510.00	1
2	580.00	1
3	600.00	1
4	640.00	1
5	670.00	1
6	580.00	2
7	595.00	2
8	600.00	2
9	605.00	2
10	620.00	2
11		

Annotations in the image:

- 1. Chọn **Data View** để chuyển sang khung nhập liệu ...
- 2. Nhập dữ liệu vào các cột tương ứng (hoặc copy từ Excel sang)

Sau khi khai báo tên biến, nhãn biến... trong **Variable View**, Bạn hãy chọn **Data View** để nhập dữ liệu.

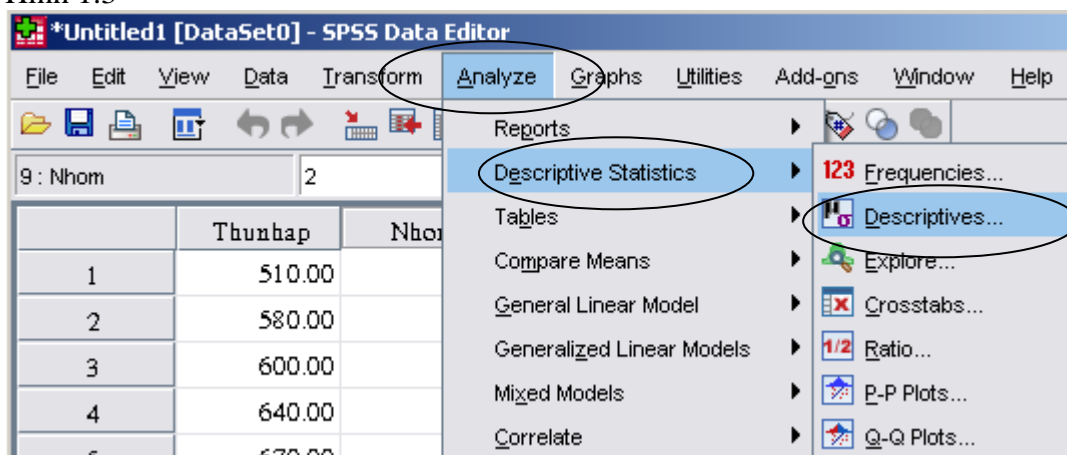
- **Tính các thống kê mô tả cho biến định lượng**

Ví dụ, bạn muốn tính các thống kê mô tả cho biến thu nhập của 10 công nhân, có nhiều cách, cách đơn giản nhất như sau:

Bước 1, Từ thanh Menu của SPSS, chọn **Analyze**

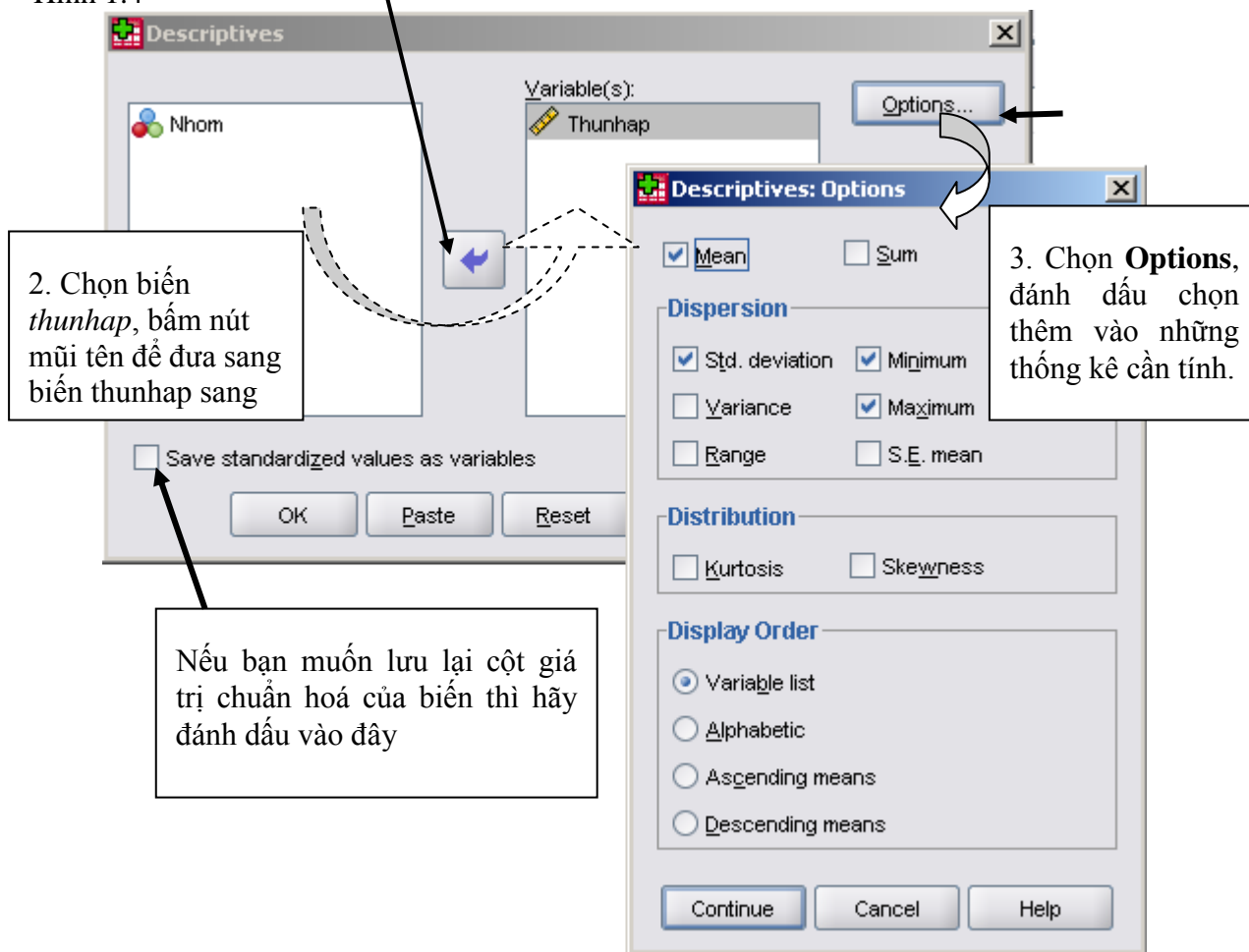
Descriptive Statistics
Descriptives

Hình 1.3



Sau đó, Hộp thoại **Descriptives** xuất hiện, Bước 2, chọn biến định lượng *thunhap* → Bấm vào nút mũi tên ở giữa hộp thoại để đưa biến *thunhap* sang khung **Variable(s)**; Bước 3, chọn **Options** để lựa chọn thêm một số thống kê khác cần tính (mặc định là Mean, Std.deviation, Minimum, Maximum); Cuối cùng, chọn **Continue**, và **OK**

Hình 1.4



Kết quả sẽ như sau:

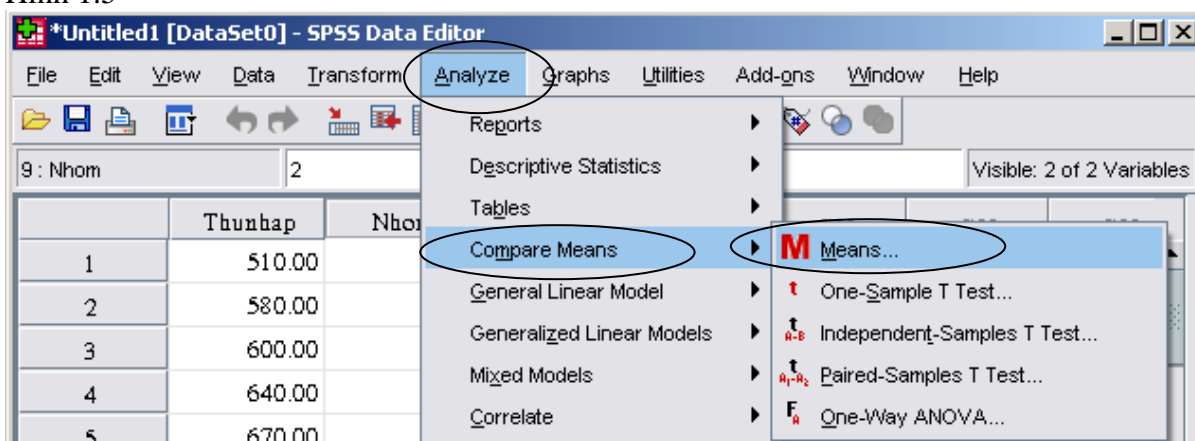
Bảng 1.1

Descriptive Statistics			
		Thunhap Thu nhap cua cong nhan (nghin dong)	Valid N (listwise)
N	Statistic	10	10
Range	Statistic	160.00	
Minimum	Statistic	510.00	
Maximum	Statistic	670.00	
Mean	Statistic	600.0000	
Std. Deviation	Statistic	41.96559	
Variance	Statistic	1761.111	
Skewness	Statistic	-.620	
	Std. Error	.687	
Kurtosis	Statistic	2.181	
	Std. Error	1.334	

- **Tính các thống kê mô tả của biến định lượng phân theo một biến định tính**
 Có nhiều cách trong SPSS để thực hiện điều này, nhưng cách sau đây là đơn giản nhất:

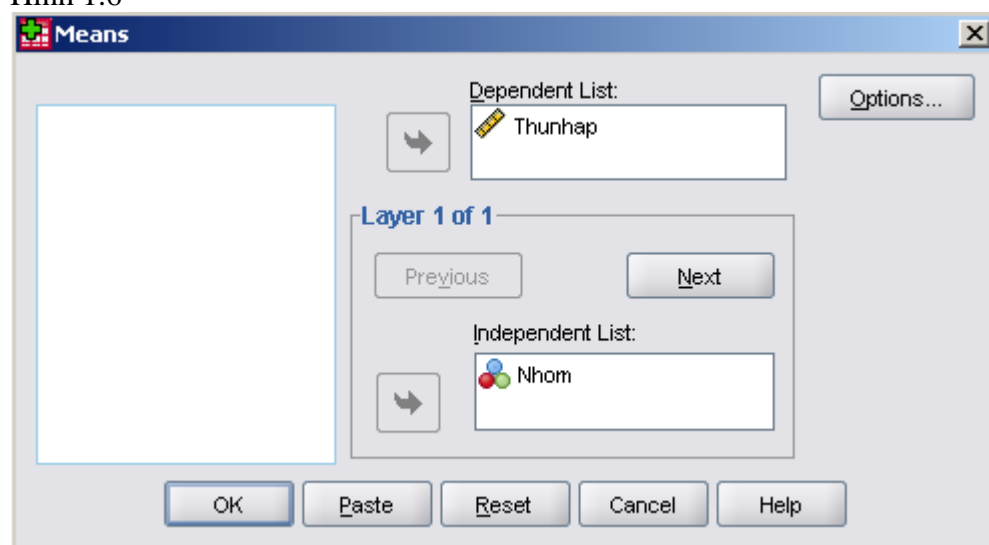
Bước 1. Chọn **Analyze\Compare Means\Means**

Hình 1.5



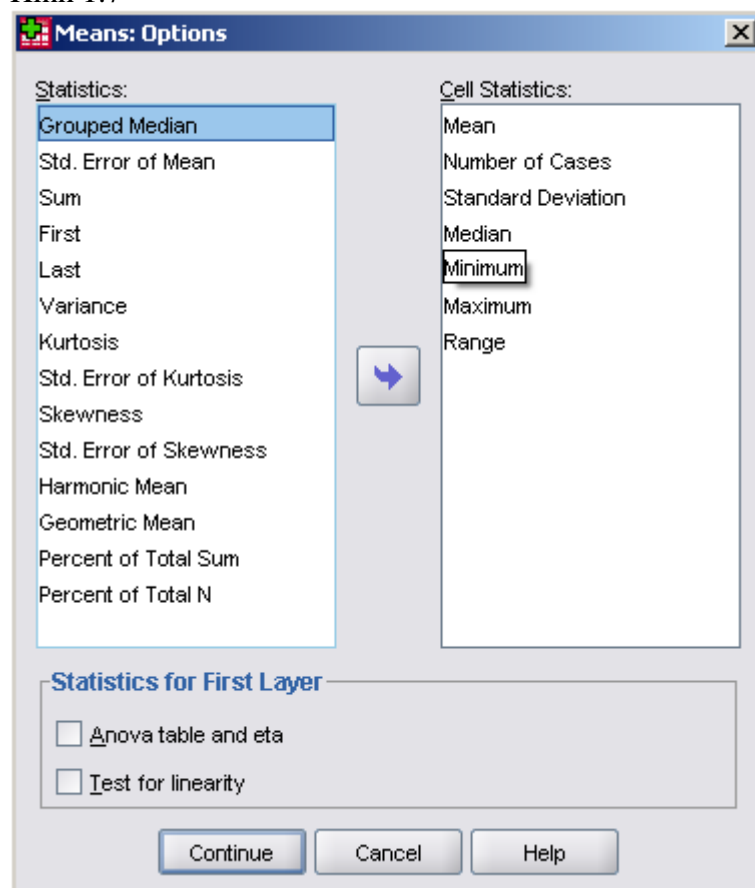
Bước 2. Đưa biến định lượng (biến thunhap) vào ô **Dependent List** & đưa biến định tính (biến nhóm) vào ô **Independent List**

Hình 1.6



Bước 3. Trong Hộp thoại **Means**, Chọn nút **Options** và chọn những thống kê cần tính (Xem Hình 1.7)

Hình 1.7



Cuối cùng, chọn **Continue** (ở Hộp thoại Means: Options), và **OK** (Ở hộp thoại Means), Kết quả như sau

Bảng 1.2

Report

Thunhap Thu nhap cua cong nhan (nghin dong)

	Mean	N	Std. Deviation	Median	Minimum	Maximum	Range
Nhom 1	600.0000	5	61.23724	600.0000	510.00	670.00	160.00
Nhom 2	600.0000	5	14.57738	600.0000	580.00	620.00	40.00
Total	600.0000	10	41.96559	600.0000	510.00	670.00	160.00

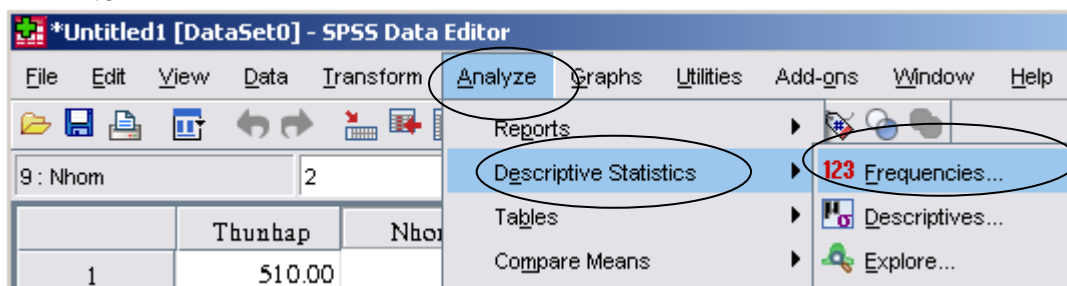
→ **Cách khác:** Bạn có thể tìm hiểu thêm về thủ tục **Analyze\Descriptive Statistics\Explore** hoặc **Analyze\Tables\Custom Tables...**

- **Tạo bảng tần số**

Nhiều khi, bạn muốn tạo bảng tần số cho một biến định tính nào đó, ví dụ như biến *Nhom* chẳng hạn. Thao tác như sau:

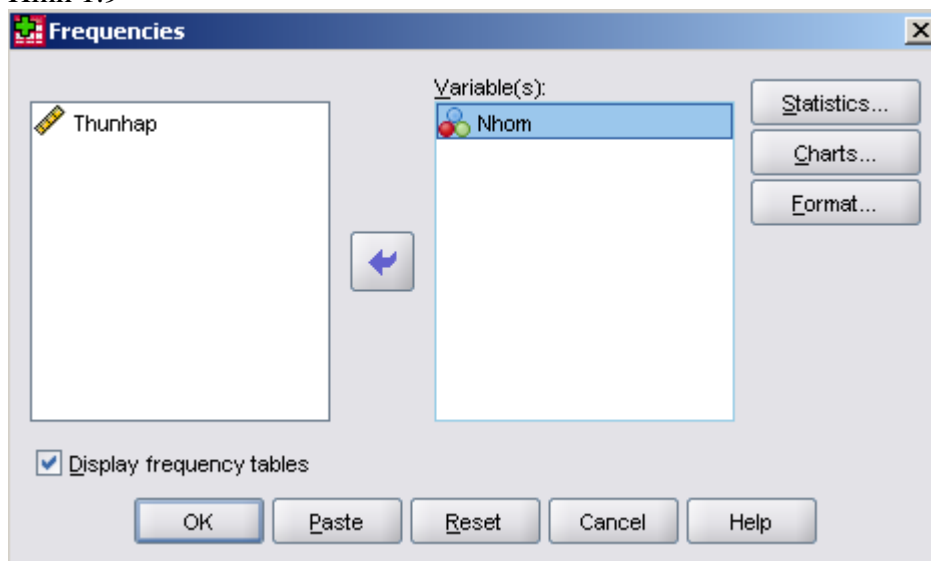
Bước 1. Từ thanh Menu, Chọn **Analyze\Descriptive Statistics\Frequencies**

Hình 1.8



Bước 2. Đưa biến định tính vào khung Variable(s), sau đó chọn OK

Hình 1.9



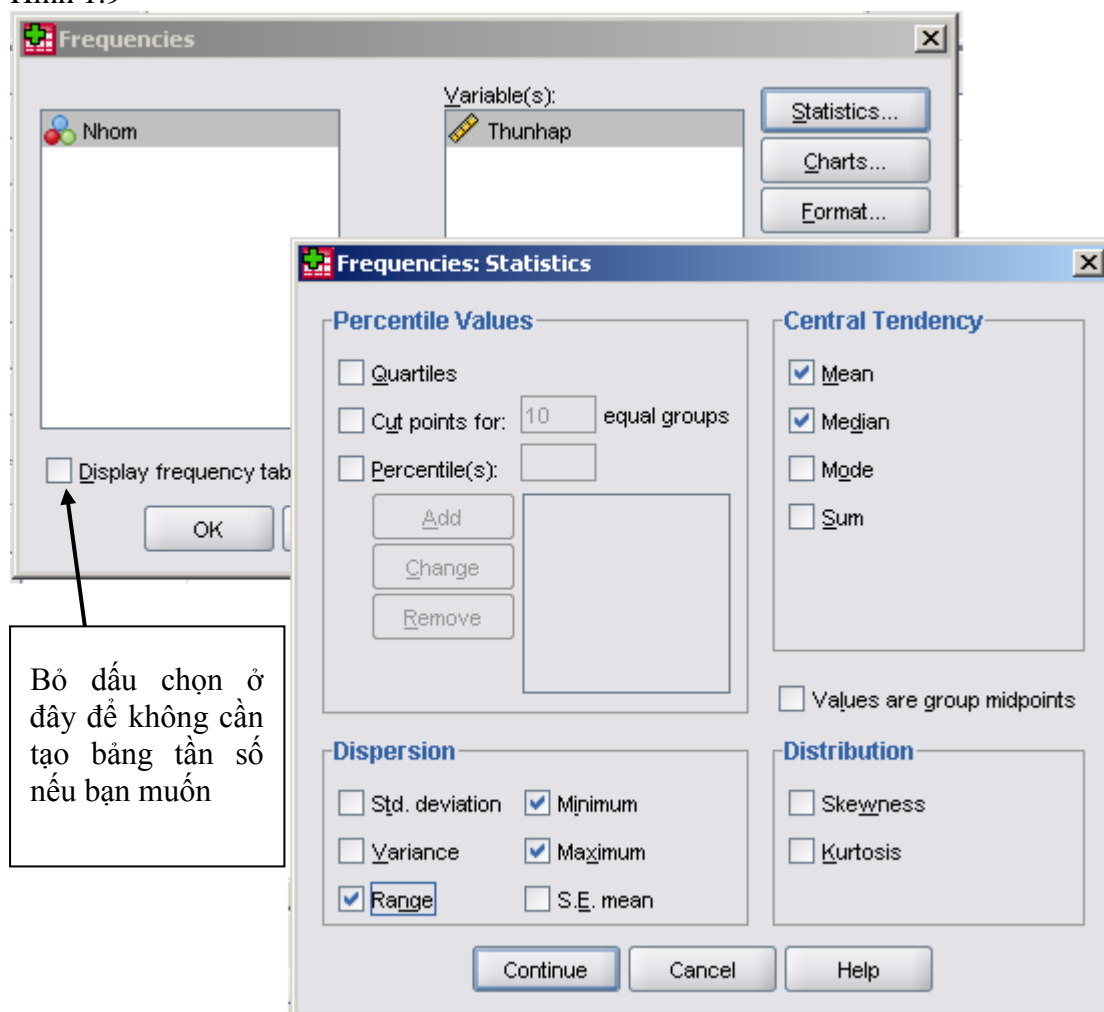
Kết quả tính toán như bảng sau
 Bảng 1.3

		Nhóm			
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Nhóm 1	5	50.0	50.0	50.0
	Nhóm 2	5	50.0	50.0	100.0
	Total	10	100.0	100.0	

↔ Ghi chú:

Trong hộp thoại **Frequencies**, bạn có thể tính các thống kê mô tả cho biến định lượng bằng cách đưa biến định lượng này vào khung **Variable(s)**, chọn **Statistics** và đánh dấu chọn vào các thống kê cần tính

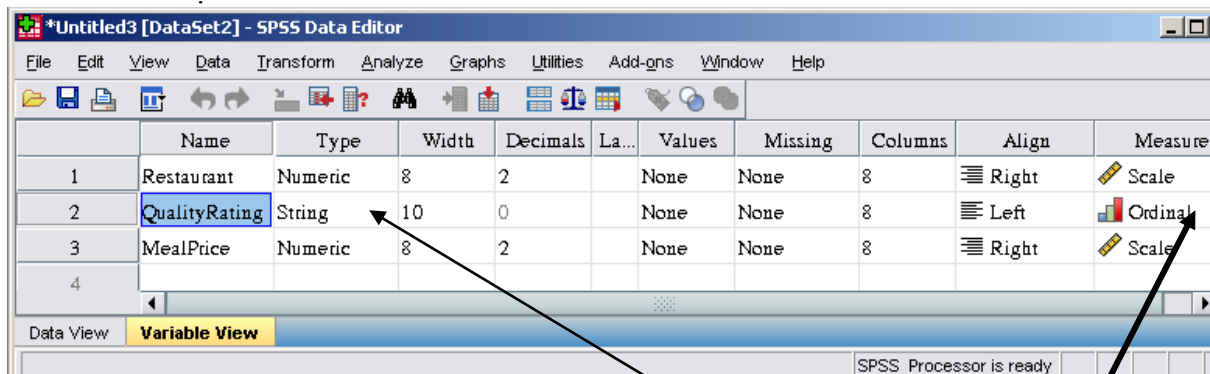
Hình 1.9



Bài 2 – Bài tập 1 với SPSS sẽ dùng đến những thao tác dưới đây

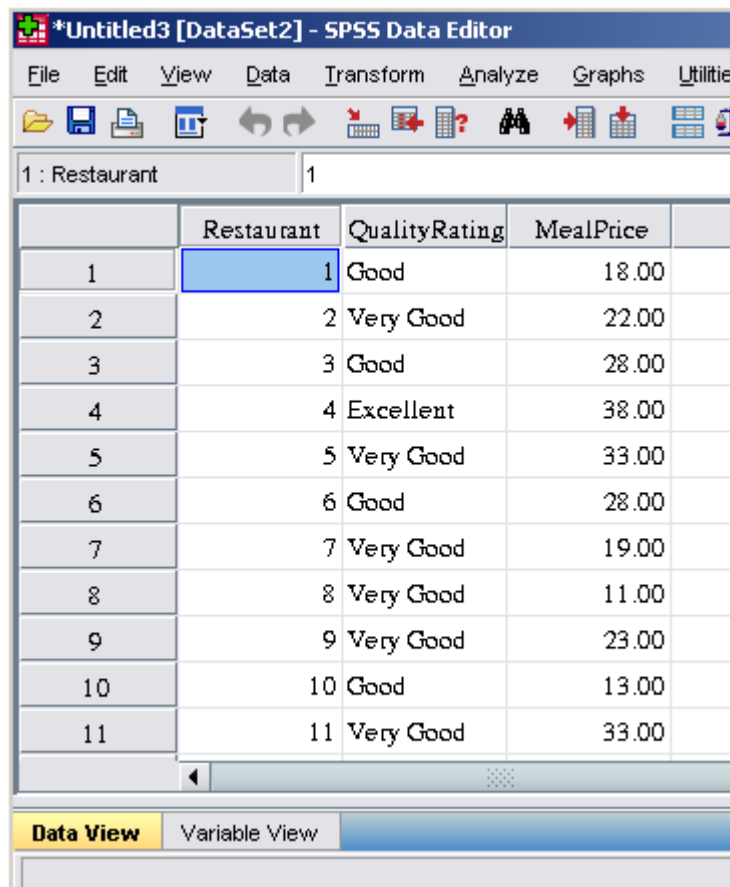
- **Tạo bảng tần số, kết hợp với vẽ đồ thị tần suất**

Hình 2.1 Dữ liệu



Biến QualityRating có thể được nhập theo kiểu ký tự (String), và có thang đo là Ordinal
Biến MealPrice có thang đo là Scale

Hình 2.2



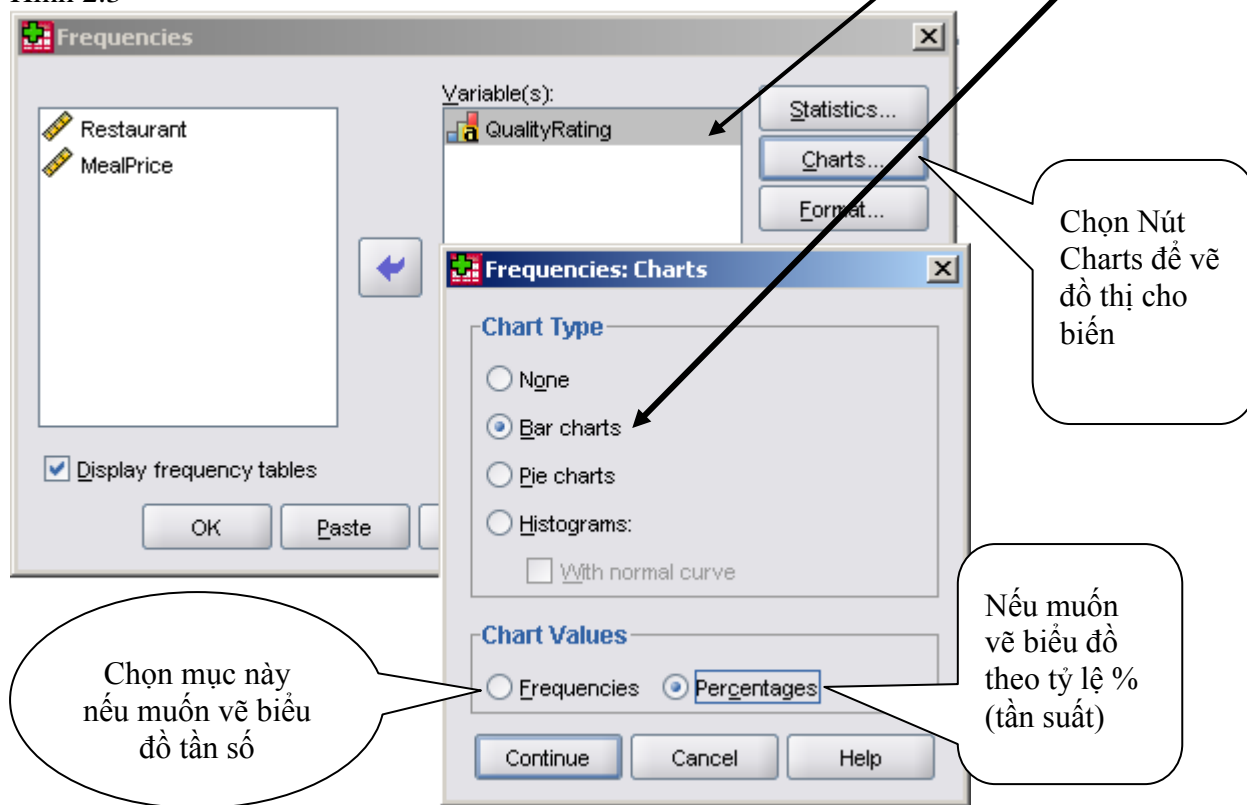
Sau khi khai báo biến như Hình 2.1, nhập dữ liệu như Hình 2.2, bạn hãy thực hiện thao tác tạo bảng tần số bằng cách:

Bước 1. Từ thanh Menu, Chọn **Analyze\Descriptive Statistics\Frequencies**

Bước 2. Trong hộp thoại **Frequencies**, đưa biến **QualityRating** vào ô **Variables**

Bước 3. Để vẽ đồ thị tần số, hãy chọn nút **Charts** và đánh dấu chọn vào loại đồ thị cần thiết như Hình 2.3

Hình 2.3

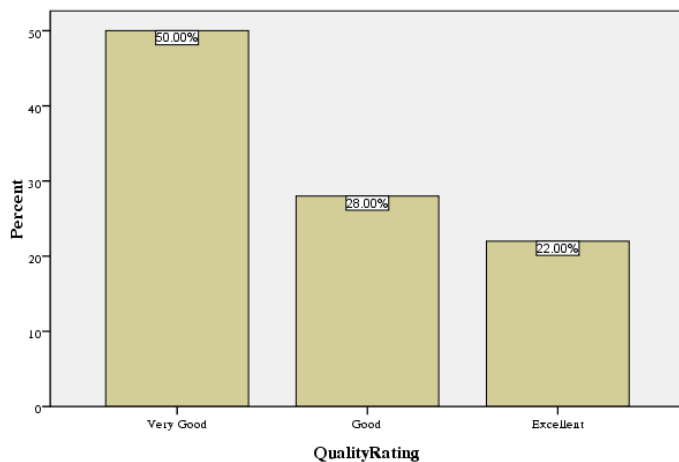


Tại cửa sổ **Viewer** bạn sẽ thấy các kết quả bảng tần số, và đồ thị sau

Bảng 2.1

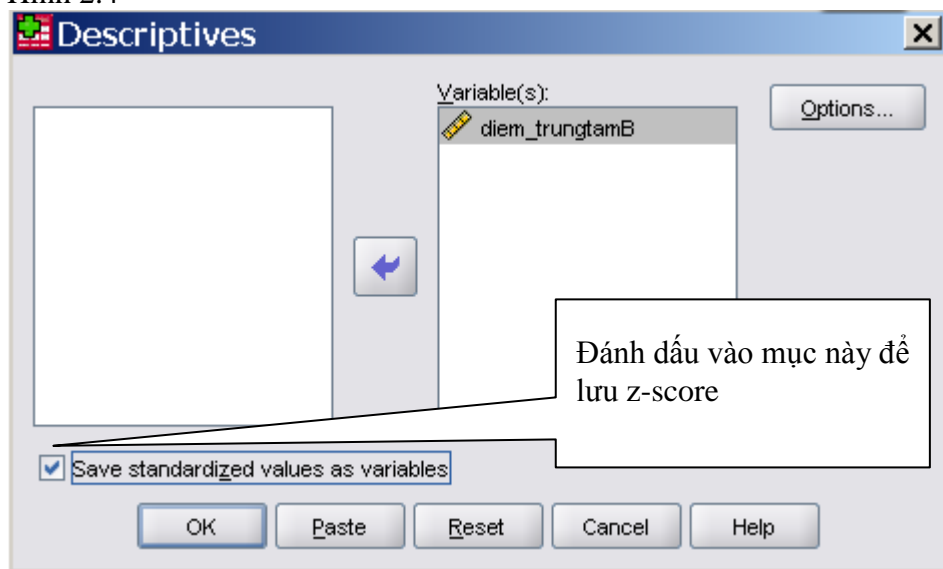
QualityRating					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Very Good	150	50.0	50.0	50.0
	Good	84	28.0	28.0	78.0
	Excellent	66	22.0	22.0	100.0
	Total	300	100.0	100.0	

Hình 2.4



Bài 3 liên quan đến việc phát hiện ra các quan sát bất thường. Có nhiều kỹ thuật khác nhau, tuy nhiên áp dụng z-score (giá trị z) là cách đơn giản nhất. Bạn có thể tính z-score cho một biến bằng cách sau: **Analyze\ Descriptive Statistics\ Descriptives** → Chọn biến cần tính.

Hình 2.4



Hình 2.5

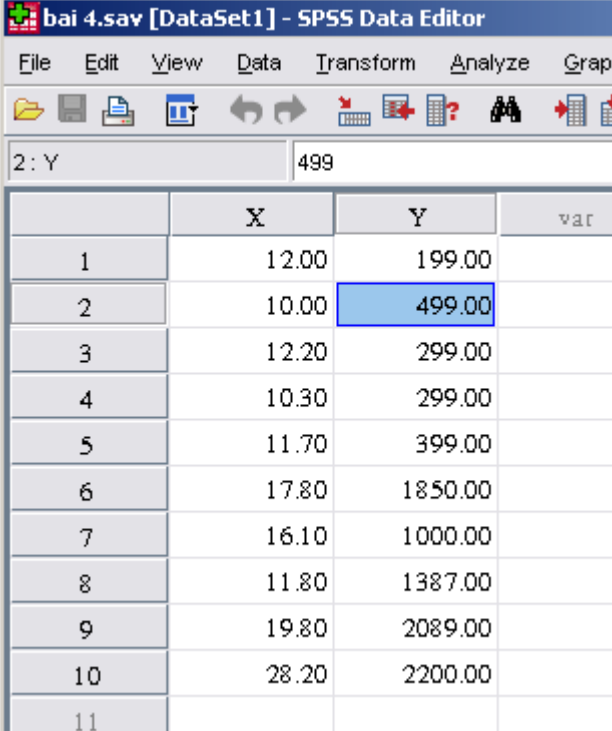
	diem_trungtamB	Zdiem_trungtamB
19	99.00	1.97774
20	57.00	-2.02219
21	91.00	1.21585
22	78.00	-0.02222
23	87.00	0.83491

Máy tính sẽ tạo ra một cột dữ liệu mới lưu giá trị z-score.

Bài 4 – Bài tập 1 sẽ cần đến thao tác vẽ đồ thị phân tán, tính hệ số tương quan

- **Vẽ đồ thị phân tán**

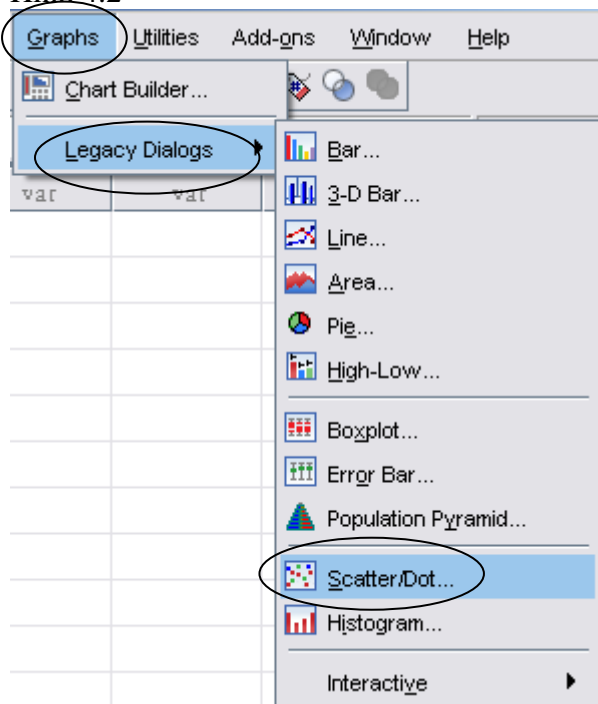
Hình 4.1



	X	Y	var
1	12.00	199.00	
2	10.00	499.00	
3	12.20	299.00	
4	10.30	299.00	
5	11.70	399.00	
6	17.80	1850.00	
7	16.10	1000.00	
8	11.80	1387.00	
9	19.80	2089.00	
10	28.20	2200.00	
11			

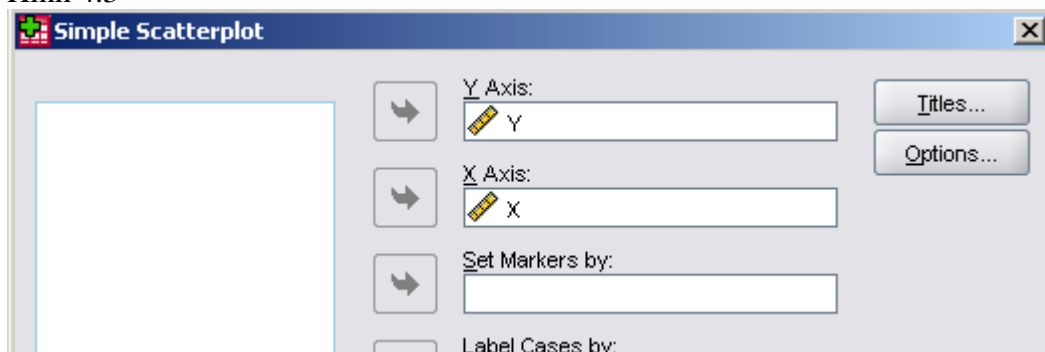
Dữ liệu sau khi được nhập vào SPSS như Hình 4.1, bạn có thể vẽ đồ thị phân tán (Hình 4.4) bằng cách: Bước 1. chọn **Graphs\Legacy Dialogs\Scatter/Dot**

Hình 4.2

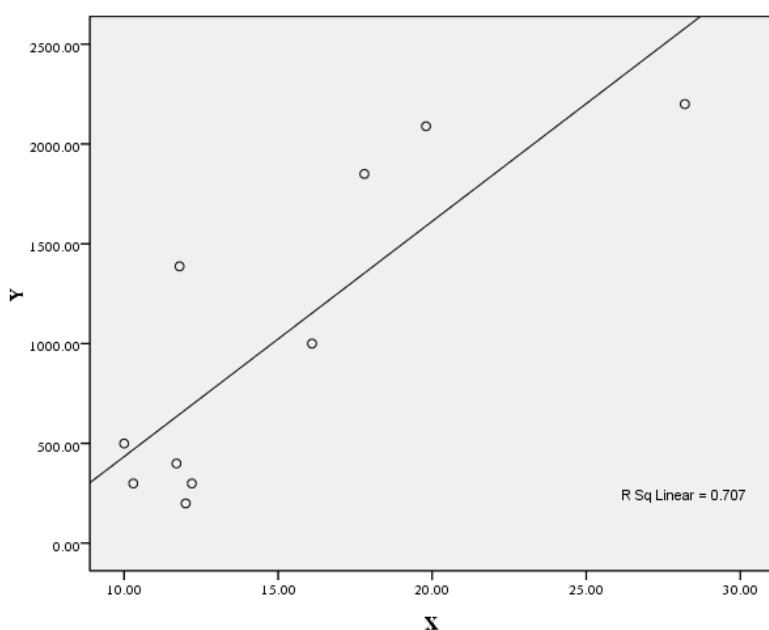


Bước 2. Đưa biến vào các trục thích hợp, sau đó bấm **OK**

Hình 4.3

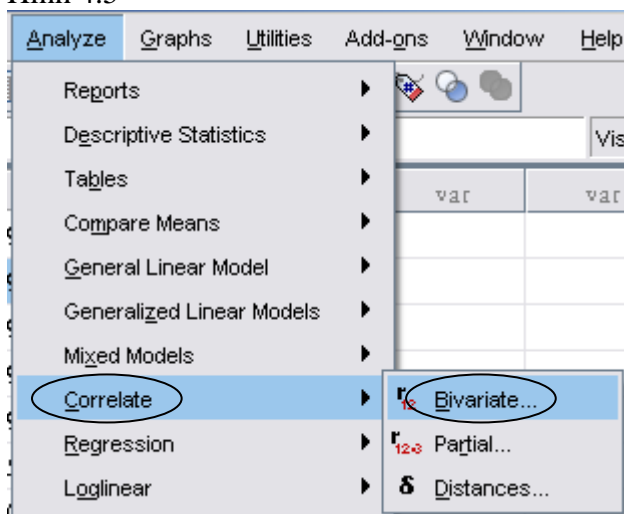


Hình 4.4



- **Tính hệ số tương quan**

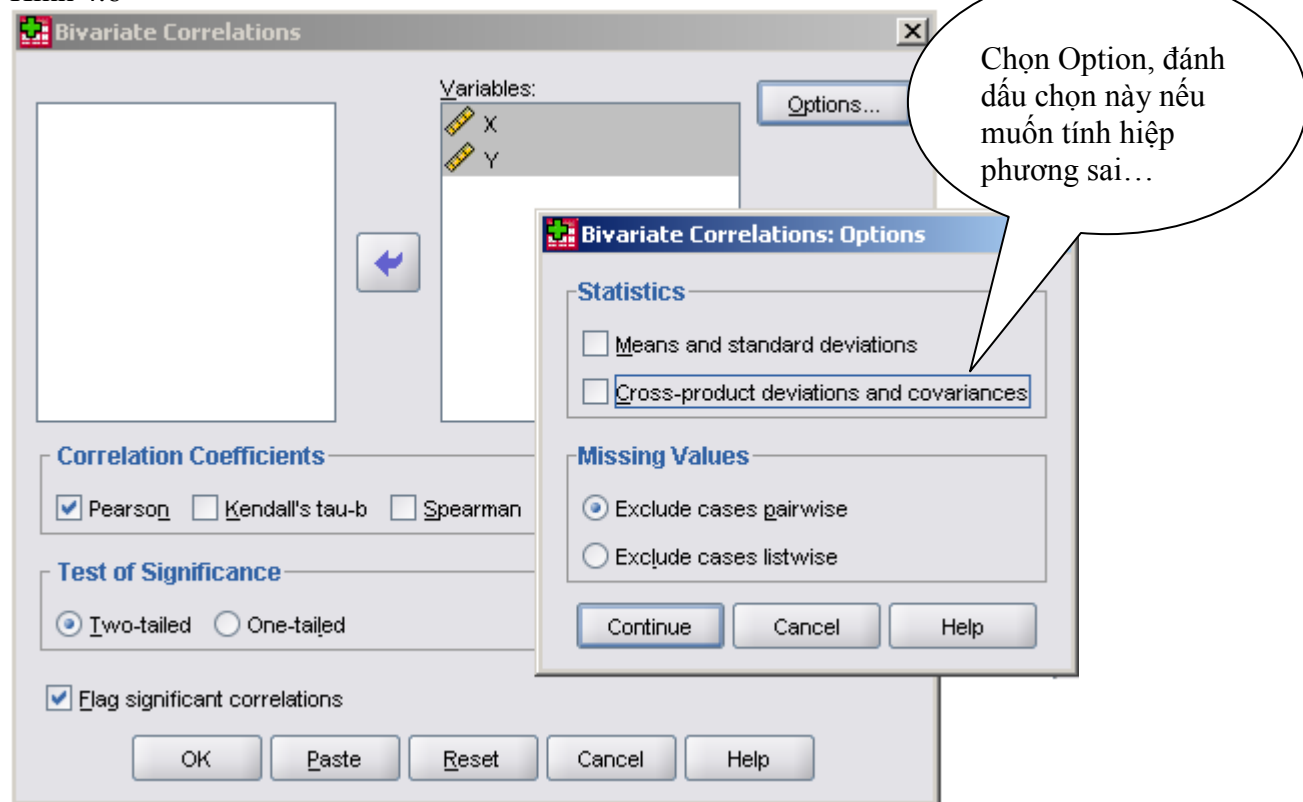
Hình 4.5



Bước 1. Analyze\Correlate\Bivariate

Bước 2. Đưa các biến cần tính hệ số tương quan vào Variables

Hình 4.6



Kết quả sẽ như sau

Bảng 4.1

Correlations

		X	Y
X	Pearson Correlation	1	.841**
	Sig. (2-tailed)		.002
	N	10	10
Y	Pearson Correlation	.841**	1
	Sig. (2-tailed)	.002	
	N	10	10

** . Correlation is significant at the 0.01 level (2-tailed).

Correlations

		X	Y
X	Pearson Correlation	1	.841**
	Sig. (2-tailed)		.002
	Sum of Squares and Cross-products	291.389	34359.810
	Covariance	32.377	3817.757
	N	10	10
Y	Pearson Correlation	.841**	1
	Sig. (2-tailed)	.002	
	Sum of Squares and Cross-products	34359.810	5729910.900
	Covariance	3817.757	636656.767
	N	10	10

** . Correlation is significant at the 0.01 level (2-tailed).