Chương Trình Giảng Dạy Kinh tế Fulbright

Học kỳ Thu năm 2011

Các Phương Pháp Phân Tích Định Lượng

Gợi ý lời giải Bài tập 1

THỐNG KÊ MÔ TẢ

Ngày Phát: Thứ Hai, 03/10/2010
Ngày Nộp: **8:20 sáng, Thứ Hai, 10/10/2010**
Bản in nộp tại Hộp nộp bài tập trong phòng Lab
Bản điện tử gởi đến thầy Nguyễn Khánh Duy, và cô Nguyễn Thị Phương Thảo theo địa chỉ duynk@fetp.vnn.vn , m3.thaoNTP@fetp.edu.vn

---

Bài 1: (25 điểm)

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

Nhóm 1: 510, 570, 600, 640, 680

Nhóm 2: 590, 595, 600, 610, 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

Số trung bình

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

$$\bar{X}_{nhóm1} = \frac{\sum_{i=1}^{5} X_{Nhóm1_i}}{5} = \frac{510+570+600+640+680}{5} = 600 \text{ ngàn đồng}$$

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

$$\bar{X}_{nhóm2} = \frac{\sum_{i=1}^{5} X_{Nhóm2_i}}{5} = \frac{590+595+600+610+620}{5} = 603 \text{ ngàn đồng}$$

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 Range=Max-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 680 ngàn đồng (Max=680 ngàn đồng), khoảng biến thiên của thu nhập là 170 ngàn đồng (bằng 680-510).

Với Nhóm2, giá trị nhỏ nhất của thu nhập là 590 ngàn đồng (Min=590 ngàn đồng), giá trị lớn nhất của thu nhập là 620 ngàn đồng (Max=620 ngàn đồng), khoảng biến thiên của thu nhập là 30 ngàn đồng (bằng 620-590)

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

Nhóm 1:

$$S^2_{Nhóm1} = \frac{(x_i - \bar{x})^2}{n-1} = \frac{(510-600)^2 + (570-600)^2 + \ldots + (680-600)^2}{5-1} = 4250$$

ngàn đồng^2

$$S_{Nhóm1} = \sqrt{S^2_{Nhóm1}} = \sqrt{4250} = 65.19 \text{ ngàn đồng}$$

$$S^2_{Nhóm2} = \frac{(x_i - \bar{x})^2}{n-1} = \frac{(590-603)^2 + (595-603)^2 + \ldots + (620-603)^2}{5-1} = 145.00$$

ngàn đồng^2

$$S_{Nhóm2} = \sqrt{S^2_{Nhóm2}} = \sqrt{145.00} = 12.04 \text{ ngàn đồng}$$

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, hoặc công cụ **Tools\Data Analysis\Descriptive Statistics** của Excel để tính toán các chỉ tiêu cho từng nhóm.Kết quả như sau:

Bảng 1.1

Nhóm 1		*Nhóm 2*	
Mean	600.00	Mean	603.00
Standard Error	29.15	Standard Error	5.39
Median	600.00	Median	600.00
Mode	#N/A	Mode	#N/A
Standard Deviation	65.19	Standard Deviation	12.04
Sample Variance	4250.00	Sample Variance	145.00
Kurtosis	-0.39	Kurtosis	-0.95
Skewness	-0.27	Skewness	0.60
Range	170.00	Range	30.00
Minimum	510.00	Minimum	590.00
Maximum	680.00	Maximum	620.00
Sum	3000.00	Sum	3015.00
Count	5.00	Count	5.00
CV	10.87		2.00

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

Trong mẫu, thu nhập trung bình của nhóm 2 cao hơn hơn nhóm 1, Thật vậy trung bình thu nhập hàng tuần của nhóm 1 và nhóm 2 lần lượt là 600 ngàn đồng, và 603 ngàn đồng.

Do hai trung bình này khác nhau, nên khi muốn so sánh về độ biến thiên của hai nhóm, bạn không nên so sánh phương sai, hay độ lệch chuẩn; mà nên sử dụng hệ số biến thiên (CV). Hệ số biến thiên của Nhóm 1 là 10.8%, trong khi hệ số biến thiên của nhóm 2 chỉ là 2%. Điều này cho thấy mức biến thiên trong thu nhập của nhóm 1 nhiều hơn nhóm 2. (Trong trường hợp này, khoảng biến thiên cũng cho bạn nhận xét tương tự)

Bài 2: (25 điểm)

Tập tin Employee Data.xls lưu dữ liệu khảo sát 474 người lao động ở nhiều công ty trong ngành Viễn Thông. Từ tập tin dữ liệu này, hãy trả lời những câu hỏi sau:

a. Phân biệt loại biến (định tính/định lượng) và loại thang đo cho các biến trong file dữ liệu (ngoại trừ biến id)

- Các biến định tính: gender (có thang đo định danh), jobcat (có thang đo thứ bậc)
- Các biến định lượng: age, prevexp, educ, salary, salbegin ; các biến này có thang đo tỷ lệ

b. Hãy phân nhóm vị trí công việc theo giới tính.

Dùng công cụ PivotTable cua Excel, bạn có thể tạo ra được kết quả như Hình 2.1

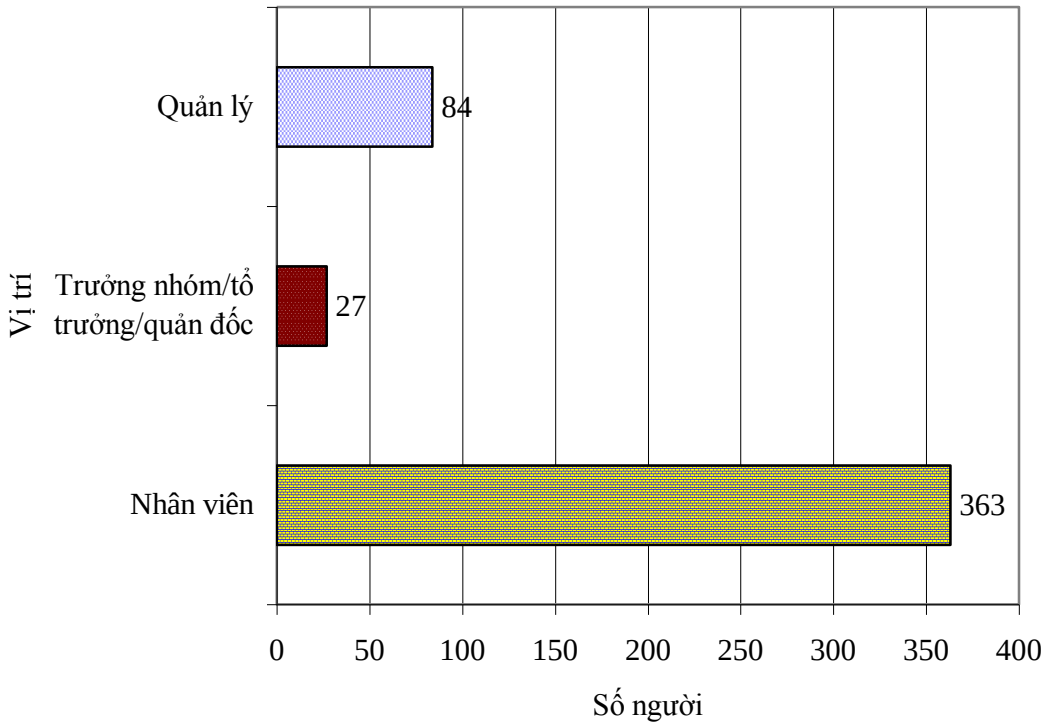
Hình 2.1

	A	B	C	D	E
1					
2					
3	Count of id	gender			
4	jobcat	Nam	Nữ	Tổng	
5	Nhân viên	157	206	363	
6	Giám Sát	27		27	
7	Quản lý	74	10	84	
8	Tổng	258	216	474	
9					
10					

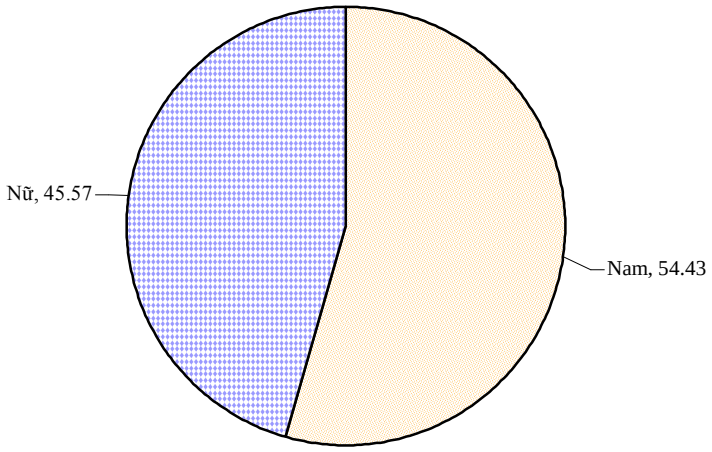
Ghi chú: Giám sát bao gồm các vị trí Trưởng nhóm, tổ trưởng, quản đốc

c. Vẽ biểu đồ tần số hình cột cho biến vị trí công việc (jobcat), biểu đồ hình tròn thể hiện cơ cấu nam/nữ trong mẫu điều tra

Hình 2.2 Số lượng người lao động được khảo sát phân theo vị trí công tác



Hình 2.3 Cơ cấu người lao động được điều tra theo giới tính (ĐVT: %)



d. Tìm các giá trị cực đại, cực tiểu, trung bình, trung vị, yếu vị của tiền lương hiện tại (salary)

Bạn có thể sử dụng các hàm thống kê, hoặc công cụ **Tools\Data Analysis\Descriptive Statistics** của Excel để tính toán các chỉ tiêu cho từng nhóm.Kết quả như sau:

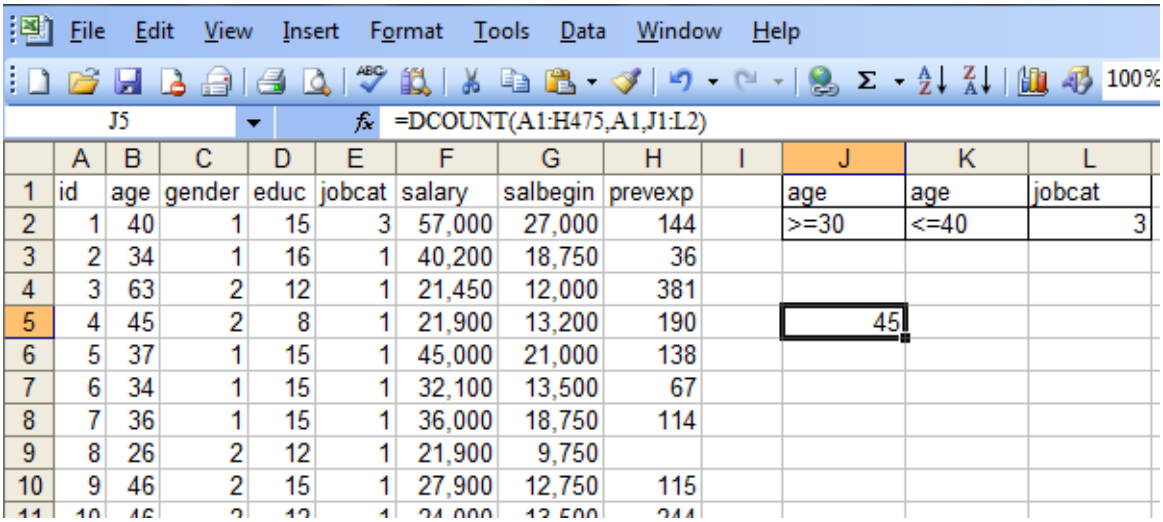
Bảng 2.1 Các thống kê mô tả cho biến salary

	salary
Mean	34,419.57
Standard Error	784.31
Median	28,875.00
Mode	30,750.00
Standard Deviation	17,075.66
Sample Variance	291,578,214.45
Kurtosis	5.38
Skewness	2.12
Range	119,250.00
Minimum	15,750.00
Maximum	135,000.00
Sum	16,314,875.00
Count	474.00

e. Có bao nhiêu người ở vị trí quản lý và ở độ tuổi từ 30 đến 40?

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 2.4



	A	B	C	D	E	F	G	H	I	J	K	L
1	id	age	gender	educ	jobcat	salary	salbegin	prevexp		age	age	jobcat
2	1	40	1	15	3	57,000	27,000	144		>=30	<=40	3
3	2	34	1	16	1	40,200	18,750	36				
4	3	63	2	12	1	21,450	12,000	381				
5	4	45	2	8	1	21,900	13,200	190		45		
6	5	37	1	15	1	45,000	21,000	138				
7	6	34	1	15	1	32,100	13,500	67				
8	7	36	1	15	1	36,000	18,750	114				
9	8	26	2	12	1	21,900	9,750					
10	9	46	2	15	1	27,900	12,750	115				

Với hàm DCOUNT bạn có thể làm như sau: (1)Tạo ra vùng điều kiện như J1:L2, (2) Tại một ô nào đó, ví dụ ô J5, gõ lệnh =DCOUNT(A1:H475,A1,J1:L2), bạn sẽ ra kết quả là 45. Nói cách khác có 45 người ở vị trí quản lý trong độ tuổi từ 30 đến 40.

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.

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

Sử dụng công cụ **Tools\Data Analysis\Descriptive Statistics** của Excel, chúng ta dễ dàng có được bảng kết quả như Hình 3.1

Từ bảng này, ta thấy:

-Trung bình điểm thi của sinh viên ở trung tâm B là 78.23 điểm

-Trung vị điểm thi của sinh viên ở trung tâm B là 78.00 điểm

Hình 3.1

	A	B	C	D	E	F	G	H	I
1	**Center A**	**Center B**	ZA		*Center A*		*Center B*		
2	97	64	1.63						
3	95	85	1.41		Mean	82.00	**Mean**	**78.23**	
4	89	72	0.76		Standard Error	1.68	Standard Error	1.92	
5	79	64	-0.33		Median	83.00	**Median**	**78.00**	
6	78	74	-0.43		Mode	79.00	Mode	78.00	
7	87	93	0.54		Standard Deviation	9.22	Standard Deviatio	10.50	
8	83	70	0.11		Sample Variance	85.03	Sample Variance	110.25	
9	94	79	1.30		Kurtosis	-0.12	Kurtosis	-0.61	
10	76	95	-0.65		Skewness	-0.43	Skewness	0.06	
11	79	75	-0.33		Range	37.00	Range	42.00	
12	83	66	0.11		Minimum	60.00	Minimum	57.00	
13	84	83	0.22		Maximum	97.00	Maximum	99.00	
14	76	74	-0.65		Sum	2460.00	Sum	2347.00	
15	82	70	0.00		Count	30.00	Count	30.00	
16	85	82	0.33						
17	85	82	0.33		Mean (B) > Median (B) => Lệch xiên về phía phải				
18	91	75	0.98		SK(B) > 0 => Lệch xiên về phía phải				
19	72	78	-1.08						
20	86	99	0.43						
21	70	57	-1.30						
22	91	91	0.98						
23	82	78	0.00						
24	73	87	-0.98						
25	96	93	1.52						
26	64	89	-1.95						
27	74	79	-0.87						
28	88	84	0.65						
29	88	65	0.65						
30	**60**	**78**	**-2.39**		--> dị biệt				
31	73	66	-0.98						
32									

Q4 Q3 Q2 Q1

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 B 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

Do điểm thi trung bình của sinh viên ở trung tâm B lớn hơn trung vị nên phân phối của điểm thi ở trung tâm B bị lệch xiên về phía bên phải.

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

Giá trị của Skewness=0.06 (>0) nên phân phối điểm thi của trung tâm B bị lệch xiên về phía phải. Kết quả này tương tự như kết quả ở câu b.

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

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 A.

Điểm thi của trung tâm A có Max=97.00 điểm , Min=60.00 điểm , độ lệch chuẩn s=9.22 điểm, $\bar{x}$ =82.00 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 $z = \frac{97 - 82.00}{9.22} = 1.63$. Giá trị lớn nhất, số 97, 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 $z = \frac{60 - 82.00}{9.22} = -2.39$. Giá trị nhỏ nhất, số 60, có 2<|z|<3 nên có thể là giá trị bất thường.

Bài 4: (25 điểm)

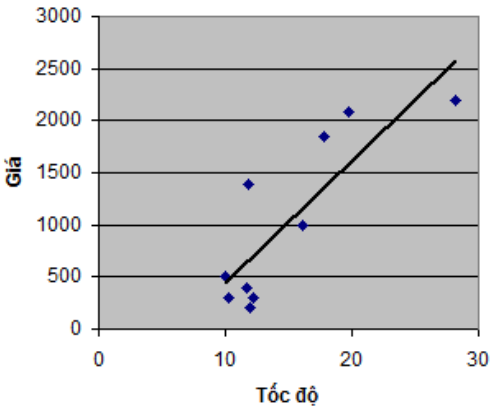
Bảng dữ liệu dưới đây trình bày tên, tốc độ và giá của 10 loại máy in (dữ liệu mẫu)

STT	**Tên máy in**	**Tốc độ**	**Giá** (USD)
1	Minolta-QMS PagePro 1250W	12	200
2	Brother HL-1850	10	500
3	Lexmark E320	12.2	300
4	Minolta-QMS PagePro 1250E	10.3	300
5	HP Laserjet 1200	11.7	400
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

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 Tốc độ (Speed) và Giá (Price) như Hình 4.1:

Hình 4.1

	A	B	C	D	E	F	G	H	I
1	**Name**	**Speed**	**Price**						
2	Minolta-QMS PagePro 1250W	12	200			*Speed*		*Price*	
3	Brother HL-1850	10	500						
4	Lexmark E320	12.2	300			Mean	15.0	Mean	1022.6
5	Minolta-QMS PagePro 1250E	10.3	300			Standard Error	1.8	Standard Error	252.2
6	HP Laserjet 1200	11.7	400			Median	12.1	Median	750.0
7	Xerox Phaser 4400/N	17.8	1850			Mode	#N/A	Mode	300.0
8	Brother HL-2460N	16.1	1000			Standard Deviation	5.7	Standard Deviation	797.4
9	IBM Infoprint 1120n	11.8	1387			Sample Variance	32.4	Sample Variance	635898.0
10	Lexmark W812	19.8	2089			Kurtosis	2.4	Kurtosis	-1.7
11	Oki Data B8300n	28.2	2200			Skewness	1.6	Skewness	0.5
12						Range	18.2	Range	2000.0
13						Minimum	10.0	Minimum	200.0
14						Maximum	28.2	Maximum	2200.0
15						Sum	149.9	Sum	10226.0
16						Count	10.0	Count	10.0
17									
18						r =	**0.841**	'=CORREL(B2:B11,C2:C11)	
19						Population Covar =	3434.106	=COVAR(B2:B11,C2:C11)	
20						Sample Covar =	**3815.673**	=G19*10/9	
21						r =	0.841	=G20/(G8*I8)	
22									
23									
24									
25									



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

Bảng 4.1

		Tốc độ (X)	*Giá cả (Y)*
Mean	Trung bình	15.0	1022.6
Standard Error	Sai số chuẩn	1.8	252.2
Median	Trung vị	12.1	750.0
Mode	Yếu vị	#N/A	300.0
Standard Deviation	Độ lệch chuẩn	5.7	797.4
Sample Variance	Phương sai	32.4	635898.0
Kurtosis	Hệ số Kurtosis	2.4	-1.7
Skewness	Hệ số Skewness	1.6	0.5
Range	Khoảng biến thiên	18.2	2000.0
Minimum	Giá trị nhỏ nhất	10.0	200.0
Maximum	Giá trị lớn nhất	28.2	2200.0
Sum	Tổng	149.9	10226.0
Count	Số quan sát	10.0	10.0

Ghi chú: biến X không xác định được yếu vị do các giá trị của X đều có cùng một tần số là 1.

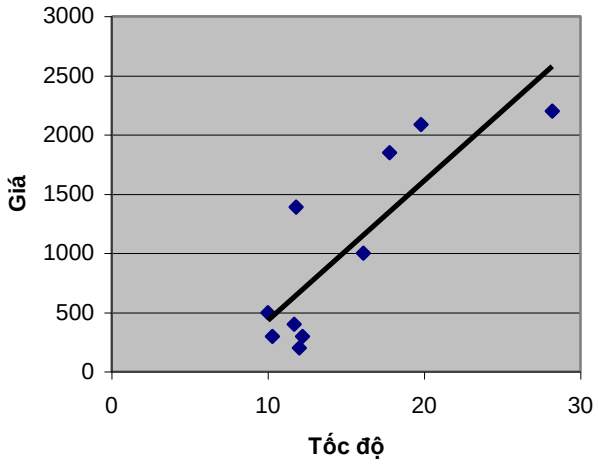
b. Theo nhận định của Anh/Chị mối quan hệ giữa 2 biến X và Y là đồng biến hay nghịch biến và độ mạnh của mối quan hệ này. Giải thích ngắn gọn về nhận định của Anh/Chị.

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.

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 4.2 Đồ thị phân tán của X và Y



Đồ thị phân tán ở Hình 4.2 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.

Hệ số hiệp phương sai tính từ dữ liệu mẫu cho hai biến X và Y là 3815.673 (Xem công thức tính toán ở Hình 4.1). Hệ số này lớn hơn 0, cho thấy hai biến X và Y có tương quan tuyến tính thuận. Hệ số tương quan giữa X và Y là 0.841 (lớn hơn 0) cho thấy mối liên hệ tuyến tính thuận giữa hai biến này, đồng thời trị tuyệt đối của hệ số tương quan rất gần 1

nên có thể xem rằng mối quan hệ tuyến tính giữa X và Y là chặt. Trong trường hợp này, kết luận tương tự như câu b và c, nhưng có căn cứ rõ ràng hơn (đặc biệt là về độ mạnh của mối tương quan tuyến tính)

🗝**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, tuỳ 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.

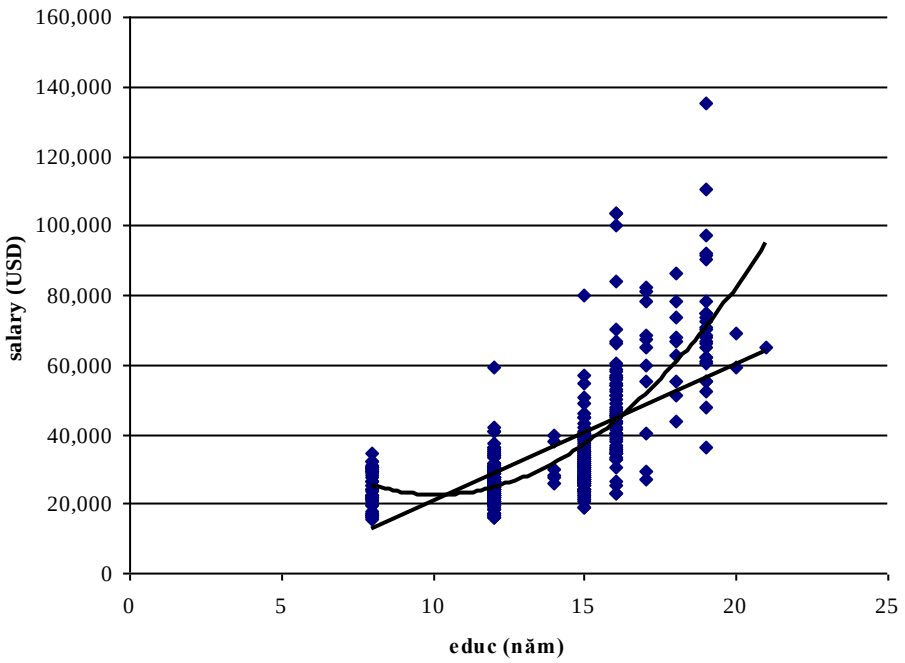
Trị tuyệt đối của hệ số tương quan Pearson	Mức độ tương quan tuyến tính
<=0.2	Rất yếu, tương quan không đáng kể
0.2 đến 0.4	Yếu, tương quan thấp
0.4 đến 0.7	Vừa phải
0.7 đến 0.9	Mạnh, tương quan cao
>0.9	Rất mạnh

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

---

e. Bạn hãy sử dụng tập tin Employee Data.xls, vẽ đồ thị phân tán của 2 biến educ, salary; tính hệ số tương quan giữa 2 biến này và cho biết nhận định của bạn.

Hình 4.3 Đồ thị phân tán giữa educ và salary



Đồ thị trên cho thấy mối quan hệ tuyến tính thuận giữa hai biến. Tuy nhiên, mối quan hệ tuyến tính ở mức vừa phải. Có vẻ như quan hệ dạng đường cong bậc hai giữa hai biến sẽ phù hợp hơn so với dạng đường thẳng tuyến tính.

Sử dụng hàm correl, hay công cụ Tools\Data Analysis của Excel bạn sẽ tìm được hệ số tương quan tuyến tính giữa hai biến này là 0.661. Hệ số này nằm trong khoảng 0.4 đến 0.7 nên mức độ tương quan tuyến tính là vừa phải.