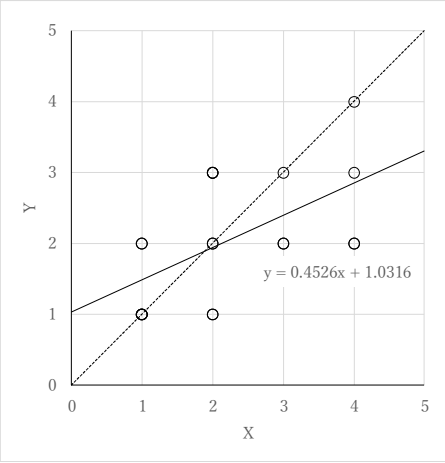
#003★☆☆☆ 2020.3.11 / 2023.7.21

グラフでプロットをずらして描く～EXCEL 乱数の活用とジッタリング。さらに相関係数について。

薬物治療による有効性の評価や、がんなどの疾患による痛みの評価といった整数値で評価を行う場合がある。いわゆるカテゴリ変数とか不連続変数、離散的変数などと呼ばれる数値であるが、こういったデータをグラフ化する場合に、異なる患者から得たデータが同じ値を示した場合にはグラフにすると同じ点に重なってしまって識別できなくなることがある。ここでは EXCEL の乱数発生機能を用いて「少しずらしてプロットする」方法を示す。

下表（左から 3 行）のデータを EXCEL で入力し、図のような個々のデータのプロットを作成してみる。この状態では一部のデータが重なっているため実際に 20 個のプロットすべては見えない。図で直線は「とりあえず」データを線形回帰（直線回帰）した結果でその数式も図中に示してあり、点線は X と Y との 1:1 の線を示す。

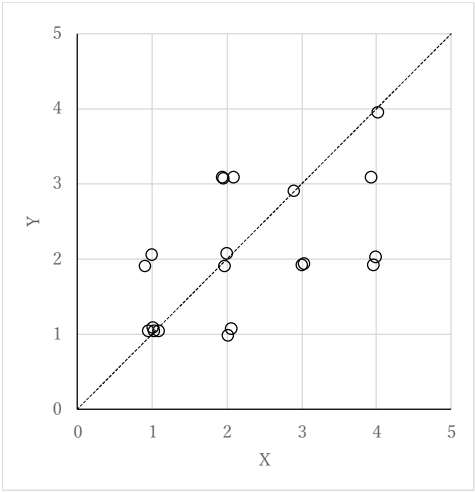
				Jitter				
	Original Data			0.2				
	X	Y		Uniform Random			Jitterd Data	
1	4	4		0.079229	0.016959		4.079229	4.016959
2	3	2		-0.08483	-0.017614		2.915174	1.982386
3	3	2		-0.00115	0.073939		2.998847	2.073939
4	1	1		0.025136	-0.022558		1.025136	0.977442
5	1	2		0.089617	-0.036706		1.089617	1.963294
6	2	1		-0.00114	-0.087068		1.998864	0.912932
7	2	3		0.035749	0.052195		2.035749	3.052195
8	2	1		0.045764	-0.069443		2.045764	0.930557
9	1	1		0.04409	-0.063238		1.04409	0.936762
10	4	3		0.06775	0.073301		4.06775	3.073301
11	2	3		0.022811	0.0165		2.022811	3.0165
12	4	2		-0.01643	0.098453		3.983573	2.098453
13	1	1		0.048911	0.059142		1.048911	1.059142
14	3	3		0.041543	0.034827		3.041543	3.034827
15	2	3		-0.043	-0.033586		1.957	2.966414
16	2	2		-0.09092	-0.014736		1.909082	1.985264
17	2	2		-0.08149	-0.097857		1.918513	1.902143
18	1	2		-0.0321	0.010884		0.967895	2.010884
19	4	2		0.064915	0.010335		4.064915	2.010335
20	1	1		-0.07461	-0.039962		0.925395	0.960038



グラフを見やすくするため実測データに乱数で少しだけ「ずれ」を加えてみる。まずは EXCEL コマンド RAND() を用いて一様乱数を発生させる。一様乱数とは 0 から 1 の範囲でいろいろな数値を同じ確率で（一様に）発生させたときの数値であり、少し工夫して RAND()–0.5 と定義すると –0.5 から +0.5 までの一様乱数を発生できることになり、さらに、その幅を 20%に縮めたい場合には：「ずれ」を加えた後のデータ ＝ もとのデータ ＋ 0.2 ×（RAND() – 0.5）と定義すれば計算できる。
EXCEL に定義した数式を下図に示した（一部のデータ分のみ表示)。

	A	B	C	D	E	F	G	H	I	J
1						Jitter				
2			Original Data			0.2				
3			X	Y		Uniform Random			Jitterd Data	
4		1	4	4		=(RAND()-0.5)*F2	=(RAND()-0.5)*F2		=C4+F4	=D4+G4
5		=B4+1	3	2		=(RAND()-0.5)*F2	=(RAND()-0.5)*F2		=C5+F5	=D5+G5
6		=B5+1	3	2		=(RAND()-0.5)*F2	=(RAND()-0.5)*F2		=C6+F6	=D6+G6
7		=B6+1	1	1		=(RAND()-0.5)*F2	=(RAND()-0.5)*F2		=C7+F7	=D7+G7
8		=B7+1	1	2		=(RAND()-0.5)*F2	=(RAND()-0.5)*F2		=C8+F8	=D8+G8
9		=B8+1	2	1		=(RAND()-0.5)*F2	=(RAND()-0.5)*F2		=C9+F9	=D9+G9
10		=B9+1	2	3		=(RAND()-0.5)*F2	=(RAND()-0.5)*F2		=C10+F10	=D10+G10

4 列目（F 列）から、X に加える乱数、Y に加える乱数、X に乱数を加えた値、Y に乱数を加えた値それぞれを EXCEL 関数式で示しており、F2 セルには 0.2（上記の 2 割）という数値が入力されていてそれを各式から参照しており、$は EXCEL での「絶対参照座標（コピー&ペーストしても参照位置が変化しない）」を意味する。20 個すべてのデータについてこの式を定義すると前ページの EXCEL の中央および右の表が得られるが、乱数は EXCEL でなんらかの操作をするごとに変化するので表の数値はそのたびに変動し同じにはならない。こうして図を少しだけずらすことを「ジッタリング」と呼ぶことがあり、ジッタリングしたデータをプロットすると次図のようにいくつかのデータが重なっていることが視覚的にわかりやすくなる。



他に同じ点にプロットされる場合にグラフを見やすくする方法として「バブルチャート」がある。同一ポイントにあるデータ数を考慮してそのプロットの大きさで表す方法である。

2 変量（X と Y）の相関性を示す指標はいくつか定義が異なるものがあり、もとのデータが連続的か、離散的かによって使い分ける必要がある。大きくはピアソン（Pearson）の相関係数とスピアマン（Spearman）の相関係数とがあり、前者はもとのデータが正規分布している（と仮定できる）場合に用いる方法で、後者は正規分布を仮定できない場合に用いる方法とされている。データが正規分布になっているかどうかを厳密に検定する方法もあるが、今回紹介したデータは整数値なのでスピアマンの相関係数がより適切と考えられる。

今回のデータでそれぞれの相関係数の値を求めてみると、それぞれ 0.571（ピアソン）、0.580（スピアマン）となり、定義が違うので値も異なる。相関係数の有意性評価方法や信頼区間の算出方法も定義されており、詳しくは成書を参照してほしい。

相関性は必ずしも直線だけで説明できるとは限らないので、もとのデータをプロットして相関係数を適用するための前提が正しいかどうかを常に意識することが大切である。前ページの図には EXCEL で簡単に操作できるように EXCEL で直線回帰した結果の式を掲載したが、この直線回帰も厳密にはデータが正規分布している（と仮定できる）場合に利用できるものであり、本来今回のような整数値データには適さないと思わる。