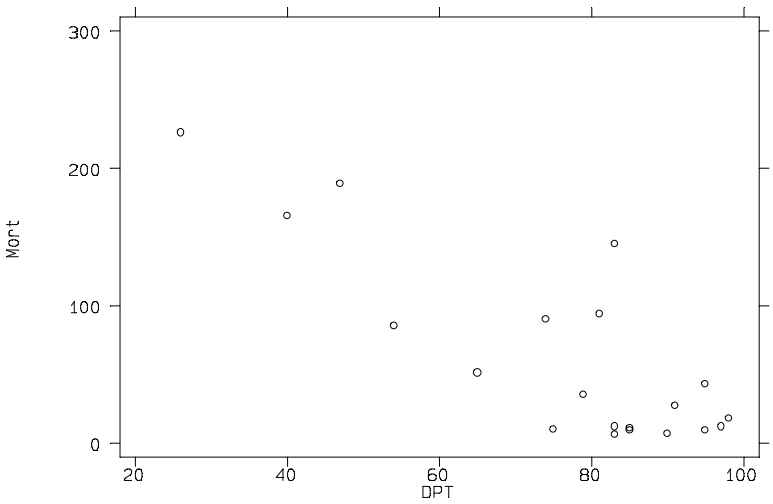
相関（Correlation）

Correlation

2 x 2 table では Yes / No の形で表せる数値を用いてきました。連続的数値を評価するときは correlation analysis を用います。２つの連続変数がお互い関連しているかどうか、直線的関係になるかどうかを検討します。例えば 20 国の 5 歳以下の死亡率（出生 1000 当り）と DPT ワクチン接種率を比較しています。



DPT ワクチン接種率が上がる程、5 歳以下の死亡率は下がりました。しかし DPT ワクチンを行なったこと自体が死亡率を直接下げたわけではないでしょう（もちろん一部はそうでしょうが）。つまりどんなに相関関係が強くでも、それ自体原因―結果の関係を示すわけではないのです。相関関係を評価する際、この点に注意する必要があります。

Pearson's Correlation Coefficient（相関係数）

$$r = \Sigma(x_i - \bar{x})(y_i - \bar{y}) / \sqrt{[\Sigma(x_i - \bar{x})^2][\Sigma(y_i - \bar{y})^2]}$$

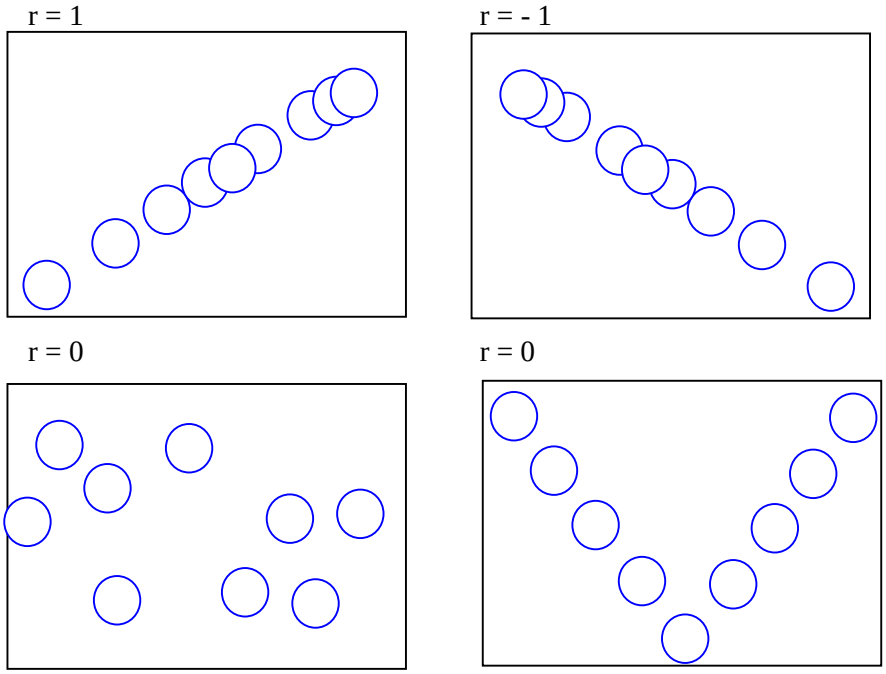
$$-1 \leqq r \leqq 1$$

$$H_0: \rho = 0$$

$$SE = \sqrt{(1 - r^2)/(n-2)}$$

$$t = (r - 0) / SE$$

下のグラフのように r が 1 か−1 であれば完全な相関関係にあります。そして r が 0 の時は相関関係がない状態です。それではその中間、例えば r = 0.5 といった場合相関関係があるといえるのでしょうか？ t test を行なったときと同様に上の公式に従って p の大きさを求め、相関関係がない(r = 0)という仮設を否定することによって、相関関係は存在するという話の持っていき方をします。



DPT ワクチンの例で r を計算すると – 0.829 になります。t は –6.29 であり p<0.001 で相関関係を認めると結論できます。Pearson's coefficient の場合、極端なデータの存在により結果が大きく左右されてしまいます。また均等に分布していればよいのですが、偏りが存在しても問題です。

STATA で計算してみましょう。

```
. correlate
(obs=20)
             |          DPT          mor
---------+------------------
         DPT |      1.0000
         mor |    -0.8291      1.0000
```

同じ coefficient が得られました。

もしも factor がいくつかある場合でもそれぞれの組み合わせで STATA は各 Correlation を計算してくれます。日本における無菌性髄膜炎の各月の症例数です。それぞれの月の症例数の間に相関があるかどうかみてみましょう。

```
. list
```

	Jan	Apr	Jul	Oct	Total
1.	31	49	148	52	280
2.	83	224	503	394	1204
3.	53	47	157	45	302
4.	41	42	158	42	283
5.	35	50	124	55	264
6.	73	193	61	41	382
7.	35	62	38	47	176
8.	105	110	1427	898	2540
9.	886	1024	1370	1927	5207
10.	1127	1582	2500	2954	8163
11.	77	35	128	161	401
12.	74	85	101	70	330
13.	69	70	95	47	281
14.	40	63	85	49	237

```
. correlate
(obs=14)
             |          Jan          Apr          Jul          Oct        Total
---------+---------------------------------------------
         Jan |      1.0000
         Apr |      0.9892      1.0000
         Jul |      0.8753      0.8821      1.0000
         Oct |      0.9698      0.9689      0.9654      1.0000
       Total |      0.9686      0.9710      0.9669      0.9995      1.0000
```

それぞれ高い相関がみられました。もしも missing data が存在する場合はどうしますか？

. list

	Jan	Apr	Jul	Oct	Total
1.	31	49	148	52	280
2.	83	224	503	394	1204
3.	53	47	157	45	302
4.	41	.	158	42	283
5.	35	.	124	55	264
6.	73	.	61	41	382
7.	35	.	38	47	176
8.	105	.	1427	898	2540
9.	886	1024	1370	1927	5207
10.	1127	1582	2500	2954	8163
11.	77	35	.	161	401
12.	74	85	.	70	330
13.	69	70	.	47	281
14.	40	63	.	49	237

. pwcorr Jan Apr Jul Oct Total

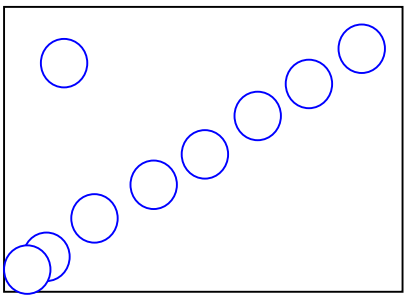
	Jan	Apr	Jul	Oct	Total
Jan	1.0000				
Apr	0.9902	1.0000			
Jul	0.8691	0.9922	1.0000		
Oct	0.9698	0.9986	0.9629	1.0000	
Total	0.9686	0.9994	0.9638	0.9995	1.0000

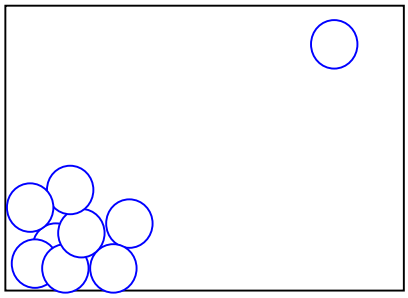
Missing data を用いずに計算します。上表で Jan Oct は missing data を持ちませんが、missing data に相当する部分は計算に用いませんので、Jan vs. Oct はせっかく全部データが揃っているのにデータの一部を捨てていることになってしまいます。しかし上の pairwise correlation (pwcorr) coefficients を用いれば、その点は解決します。Jan vs. Oct は全てのデータを使っているのが下の表を比べても判ります。

. correlate Jan Oct
(obs=14)

	Jan	Oct
Jan	1.0000	
Oct	0.9698	1.0000

Spearman's Rank Correlation Coefficient





上図のように outlier が存在した場合 Pearson's Correlation Coefficient は強く影響を受けます。左の場合、１つの outliner によって明らかに存在する相関関係が過小評価されるかもしれません。また右のような場合も間違ったデータが１つ存在することにより、強い相関関係ありと間違った結論に達しかねません。そこで１つ１つのデータ（x と y 別々に）に順番をつけます。先の例を用いて Pearson's Correlation Coefficient を計算してみましょう。

	% immunized	Rank (A)	Mortality rate	Rank (B)	d_i (A – B)	d_i^2 $(A – B)^2$
Ethiopia	0.26	1	0.0226	20	-19	361
Bolivia	0.40	2	0.0165	18	-16	256
Senegal	0.47	3	0.0189	19	-16	256
Brazil	0.54	4	0.0085	14	-10	100
Mexico	0.65	5	0.0051	13	-8	64
Turkey	0.74	6	0.0090	15	-9	81
UK	0.75	7	0.0010	5	2	4
USSR	0.79	8	0.0035	11	-3	9
Egypt	0.81	9	0.0094	16	-7	49
Japan	0.83	10	0.0006	1	10	100
Greece	0.83	11	0.0012	7.5	3.5	12.25
India	0.83	12	0.0145	17	-6	36
Canada	0.85	13	0.0009	3.5	10	100
Italy	0.85	14	0.0011	6	7.5	56.25
Finland	0.90	15	0.0007	2	13	169
Yugoslavia	0.91	16	0.0027	10	6	36
France	0.95	17	0.0009	3.5	14	196
China	0.95	18	0.0043	12	5.5	30
USA	0.97	19	0.0012	7.5	11.5	132.25
Poland	0.98	20	0.0018	9	11	121
total						2169

$$r_s = 1 - [6 \sum d_i^2]/n(n^2 - 1)$$

$$= 1 - [6 \times 2169]/20 \times (20^2 - 1) = -0.631$$

$$t_s = r_s \sqrt{(n-2)/(1-r_s^2)}$$

= - 3.45　p < 0.01

よって相関関係がないという仮設を reject して、相関関係が存在すると結論できます。Spearman は nonparametric であり実際の数値は用いずランクを用います。ですから outliers にあまり影響されません。

Stata で計算してみましょう。まずデータを入力して下のようにタイプするとデータが算出されます。

```
. spearman　DPT mor

  Number of obs =        20
Spearman's rho =      -0.6357

Test of Ho: DPT and mor independent
        Pr > |t| =       0.0026
```

Ktau はKendall's rank correlation coefficient を示しています。p<0.05 のためDPT接種率と死亡率は独立している（相関しない）というH_0が否定され、両者の間には相関関係が認められました。

```
. ktau　DPT mor

   Number of obs =        20
Kendall's tau-a =      -0.4368
Kendall's tau-b =      -0.4451
Kendall's score =       -83
      SE of score =      30.698    (corrected for ties)

Test of Ho: DPT and mor independent
         Pr > |z| =       0.0076   (continuity corrected)
```

相関係数は Spearman's rank correlation において高くなる傾向にあります。

```
. spearman　Jan Oct

  Number of obs =        14
Spearman's rho =       0.6751
```

Test of Ho: Jan and Oct independent

Pr > |t| = 0.0081

Apgar score とSpearman rank-correlation

Apgar score は新生児の仮死状態を示すスコアで、1953 年 Apgar により考案されました（Current Researches in Anesthesia and Analgesia, 260-267,1953）。出生後 5 分で測定するのが一般的ですが、1 分後も通常測定します。我々は 1 分後と 5 分後の apgar score の関係について調べようと思います。Apgar score の分布は 9, 10 点に偏るため正規分布にはなりません。よって Pearson (parametric)を使わず Spearman Rank (nonparametric)を使用します。

. list

	var1	var2
1.	10	10
2.	3	6
3.	8	9
4.	9	10
5.	8	9
6.	9	10
7.	8	9
8.	8	9
9.	8	9
10.	8	9
11.	7	9
12.	8	9
13.	6	9
14.	8	10
15.	9	10
16.	9	10
17.	9	10
18.	9	9
19.	8	10
20.	9	9
21.	3	3
22.	9	9
23.	7	10
24.	10	10

```
. spearman    var1 var2

  Number of obs =          24
Spearman's rho =      0.5927

Test of Ho: var1 and var2 independent
          Pr > |t| =      0.0023
```

よって 1 分後と 5 分後の apgar score の間には統計学的に有意な相関関係を認めます。
相関係数は 0.6 です。