

第12回 相関分析

A. 2つの変数間の関係を調べる

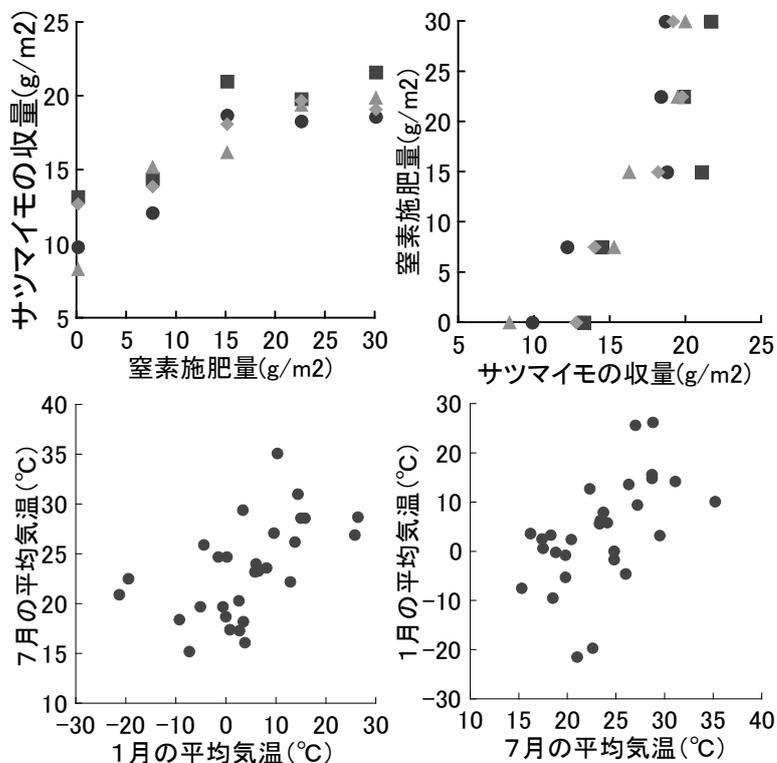
1. 散布図を書く

例1 水稻の収量に関連のある生育指標を知りたい。例えば草丈と収量には関連があるだろうか？

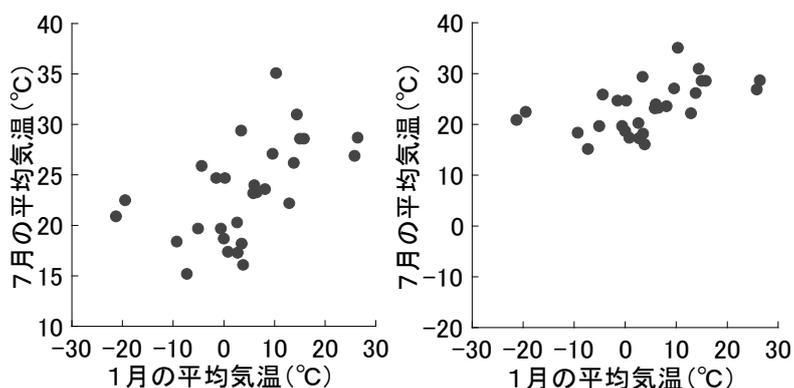
例2 トマトの糖度は施肥量によってどのように変化するかを知りたい。例えば、窒素施肥量を増加させると糖度はどうなるか？

① 散布図の書き方

1) x軸（横軸）には原因となる変数を、y軸（縦軸）には結果となる変数をとる。

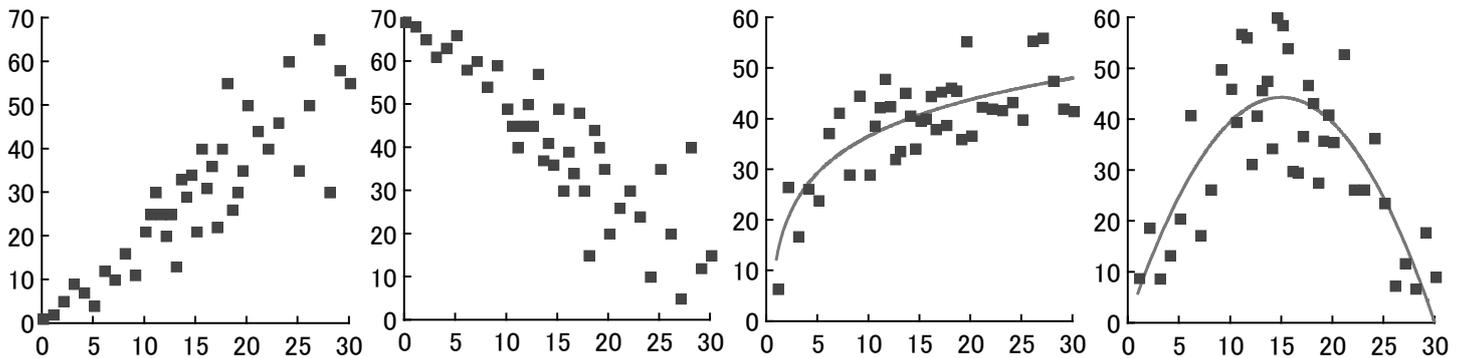


2) できるだけ点が図全体にばらつくように軸の上下限を決める。



② 散布図を書く意義

1) 視覚的にどんな関係かを考えることができる



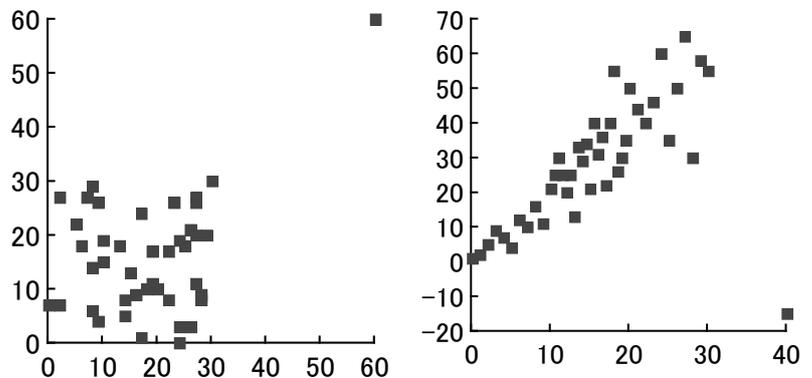
2つの変数間の関係をどう表現するか

- ① 直線的関係
- ② 2次関数
- ③ 指数・対数・双曲線など
- ④ その他

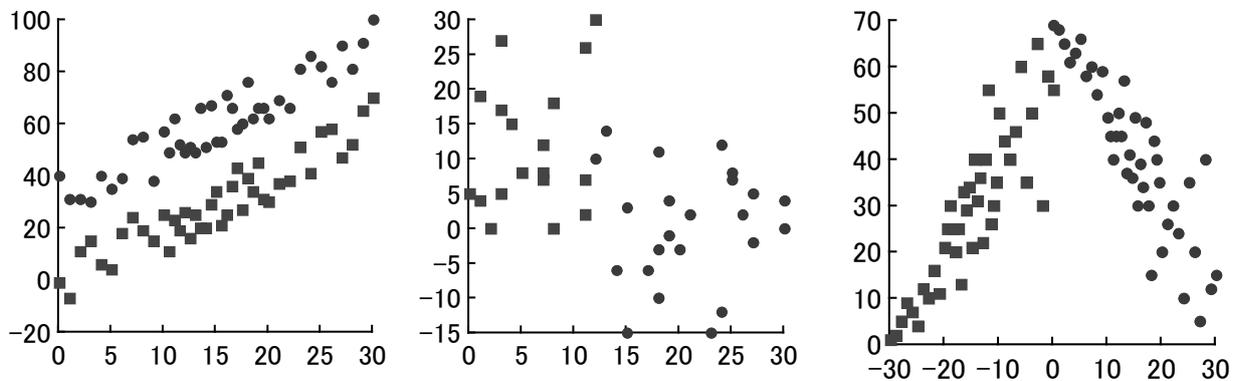
今回の授業では直線的関係についてしか学ばないが、2つの変数間の関係を曲線で表す方がよい場合も多い。

2) 異常値などを発見できる

異常値を除去すると、2つの変数間に相関関係が変わることがある。



3) 異なるグループに分けられることがある



コンピューターを使えば、与えられたデータにいかような関係でも簡単に当てはめられることができる。しかし、その当てはめがいつでも正しいとは限らない。必ず図に書いてから解析を始めるように心がけなければならない。

## B. 相関と回帰

### 1. 相関と回帰の違い

2つの変量 ( $x$ ,  $y$ ) の関係について,  $x$ ,  $y$  ともに正規分布にしたがってばらつく量であるときには両者の直線的な関係を**相関**という. 一方,  $x$ については指定できる変数(独立変数という)であり,  $y$ が指定された $x$ に対して, あるばらつきを含んで決まる場合, **回帰**という.

相関では両変数間の関連の度合いを相関係数で評価することを主な目的とする. 回帰では相関係数で評価することもできるが, 主たる目的は両変数間の数的関係を回帰直線で表し, ある $x$ が指定されたときに $y$ がいくつになるかを求めることである.

### 2. 相関と回帰の例

兄弟の身長について考える. 兄の身長と弟の身長それぞればらつきのある変数であり, 兄の身長を指定しても, そのことで弟の身長が決まるとは考えられない. したがって, 兄弟の身長は相関である. しかし, 父と子の身長を考えると, 遺伝的な要因から父の身長は子の身長に影響を及ぼしているであろう. 父の身長を指定するとあるばらつきを持って, 子の身長が決まると考えられる. 父と子の身長は回帰分析できる. 父と子の身長はともに正規分布するので相関分析もできる. 次に食事で得た蛋白質の量と身長を考えよう. 蛋白質の量を決めればあるばらつきを持って身長が決まるから, 回帰分析できる. この場合は蛋白質の量は指定でき, 正規分布しないので, 相関分析は不適當である.

次の例は相関か回帰か?

最高気温と最低気温

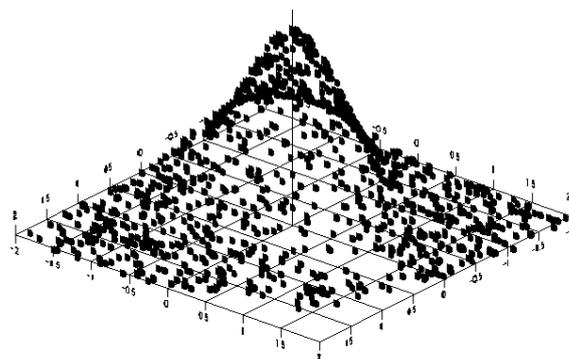
身長と体重

飼料中の脂肪含量と牛の乳脂肪率

テレビを見る時間と血圧

テレビを見る時間とエンゲル係数

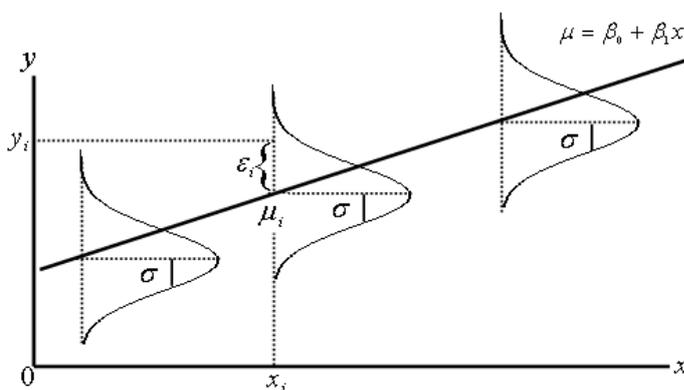
塩分摂取量と血圧



### 3. 相関と回帰のモデル

相関は2変量がそれぞれ正規分布する右上の図のようなデータをモデルとしている.

それに対して回帰では右の図のように $x$ は指定できる変数なので, 誤差は持たない. 一方,  $y$ は指定された $x$ について正規分布し, ある一定の誤差を持つ.  $y$ の誤差は $x$ の値によって変化しない.



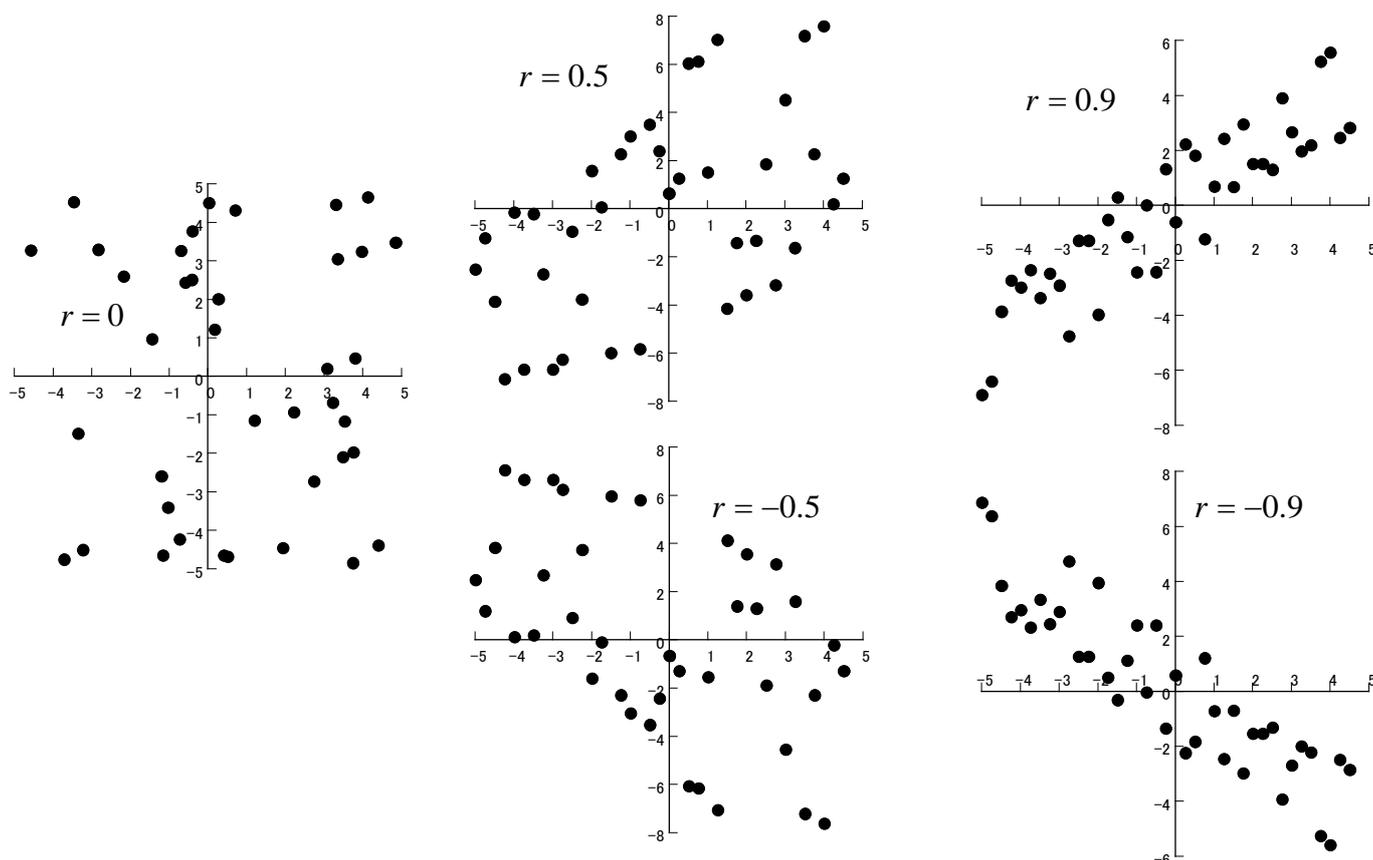
C. 相関

1. 相関係数  $r$

2つの変数間の直線的な関係（相関関係）は相関係数  $r$  によって定量的に示すことができる。

相関係数には以下の性質がある

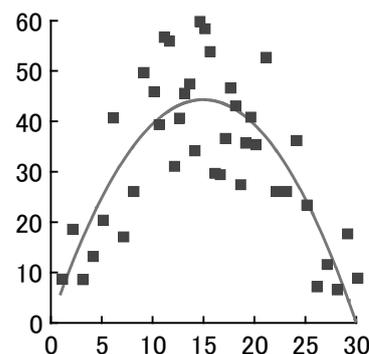
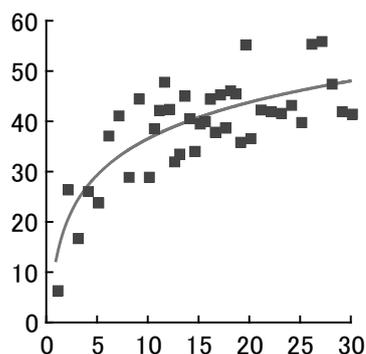
- ①  $-1 \leq r \leq 1$  である。
- ②  $r$  が 1 に近いほど正の相関が強く、 $-1$  に近いほど負の相関が強い。
- ③  $r$  が 0 に近いときは、両変数間には相関がない（無相関）。



相関係数  $r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$  ここで積和  $S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$ ,  $x$  の平方和  $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$ ,

$y$  の平方和  $S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2$

相関係数は2変数の間に直線的な関係があるかだけ进行评估する。したがって、2次曲線のような関係があっても相関係数  $r$  はほとんど0かもしれない。



相関係数の計算方法（右のデータについて）

① 関数を使う方法

相関係数	0.543962	=CORREL(C4:C32,D4:D32)
------	----------	------------------------

② 分析ツールを使う方法



	A	B	C	D	E
1					
2		北半球各地点での1月の気温と7月の気温の関係			
3			1月	7月	
4		オスロ	-7.5	15.2	
5		ロンドン	3.6	16.1	
6		パリ	3.3	18.2	
7		リヨン	2.4	20.3	
8		マルセイユ	6.3	23.3	
9		マドリード	5.8	24	
10		ライプチヒ	0.6	17.4	
11		ベルリン	-0.2	18.7	
12		ウィーン	-0.8	19.7	
13		ローマ	7.9	23.6	
14		アテネ	9.4	27.1	
15		イスタンブール	5.6	23.2	
16		モスクワ	-9.5	18.4	
17		ハバロフスク	-21.5	20.9	
18		キエフ	-5.3	19.7	
19		バグダッド	10.1	35.1	
20		テヘラン	3.2	29.4	
21		カブール	-1.7	24.7	
22		ニューデリー	14.2	31	
23		香港	15.6	28.6	
24		台北	14.8	28.6	
25		バンコク	26.2	28.7	
26		シンガポール	25.6	26.9	
27		ハルビン	-19.7	22.5	
28		北京	-4.6	25.9	
29		カサブランカ	12.7	22.2	
30		アレクサンドリア	13.6	26.2	
31		バンクーバー	2.5	17.3	
32		ニューヨーク	0	24.7	

	1月	7月
1月	1	
7月	0.543962	1

練習：右のデータ（授業用データ）について相関係数を計算せよ。

	北半球各地点の各月の気温					
	1月	2月	3月	4月	5月	6月
オスロ	-7.5	-7.8	-2.7	2.7	9	14
ロンドン	3.6	4.1	5.6	7.9	11.1	14
パリ	3.3	4	6.6	9.6	13.3	16
リヨン	2.4	4	7.1	10.3	14.4	17
マルセイユ	6.3	7.3	10	12.9	16.9	20
マドリード	5.8	7	9.8	12.3	16.2	20
ライプチヒ	0.6	1	4.3	6.8	12.3	16
ベルリン	-0.2	0.5	3.8	8.5	13.6	17
ウィーン	-0.8	1.1	4.9	10	14.5	18
ローマ	7.9	8.8	10.5	13.2	17.2	21
アテネ	9.4	10.1	11.6	15.1	20.1	24
イスタンブール	5.6	6.1	7.2	11.5	16.4	20
モスクワ	-9.5	-8.4	-3.3	5.1	12.4	16
ハバロフスク	-21.5	-17.6	-7.7	3.3	11.6	17
キエフ	-5.3	-4.6	0	8.3	14.7	18
バグダッド	10.1	12.6	16.5	22.5	28.2	33
テヘラン	3.2	5.7	10.3	15.8	21.9	26
カブール	-1.7	-0.7	5.9	12.9	17.6	22
ニューデリー	14.2	17.2	22.7	28.9	32.8	33
香港	15.6	15.9	18.4	22.1	26	27
台北	14.8	15.5	17.8	21.3	24.9	26
バンコク	26.2	27.7	29.2	30.3	29.7	29
シンガポール	25.6	26.1	26.6	27	27.3	27

## 2. 相関係数に関する推定と検定

### ① 推定

相関係数  $r$  は集めてきたデータ (標本) から求めたものであるから、統計量である。母集団の相関係数である母相関係数  $\rho$  を  $r$  から区間推定することができる。

その前に母相関係数  $\rho$  が与えられたときに、標本の相関係数  $r$  はどのように分布するかをみてみよう。

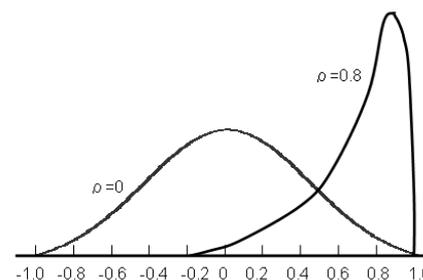


図 相関係数の分布

右上の図のように母相関係数  $\rho$  が 0 であるときには、その母集団から無作為に抽出した標本の相関係数は左右対称に分布する。しかし、母相関係数が  $\pm 1$  に近づくと著しくゆがんだ分布をする。

標本の相関係数  $r$  の分布は

$\rho = 0$  のとき  $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$  は自由度  $n-2$  の  $t$  分布に従う。

$\rho \neq 0$  のとき  $Z = \frac{1}{2} \ln \frac{1+r}{1-r}$  は近似的に正規分布  $N(\frac{1}{2} \ln \frac{1+\rho}{1-\rho}, \frac{1}{n-3})$  に従う。

この関係を利用して、標本の相関係数  $r$  が得られたときに母相関係数を区間推定できる。

ここで  $\ln$  は自然対数 (底を  $e$  とする対数) である。 $e$  は自然対数の底であり、 $e \approx 2.71828\dots$  エクセルの関数を利用して計算する以下の方法では自然対数についての知識は必要ない。

母相関係数  $\rho$  の推定の手順

- (1) 点推定:  $\hat{\rho} = r$
- (2) 区間推定: 信頼率  $p\%$  の  $\rho$  の信頼区間はエクセルでは以下のように計算できる。

$$FISHERINV((FISHER(r) - NORMSINV((100 + p)/200)/SQRT(n-3))) \leq \rho \leq FISHERINV((FISHER(r) - NORMSINV((100 - p)/200)/SQRT(n-3)))$$

例: 相関係数  $r=0.959$  (標本数  $n=6$ )

	B	C	D	E	F	G	H	I	J
88									
89	標本相関係	0.959							
90									
91	標本数n	6							
92									
93	信頼率	95 %							
94									
95	$\rho$ 下限	0.664999	=FISHERINV((FISHER(C89)-NORMSINV((100+C93)/200)/SQRT(C91-3)))						
96									
97	$\rho$ 上限	0.995655	=FISHERINV((FISHER(C89)-NORMSINV((100-C93)/200)/SQRT(C91-3)))						

すなわち 95%信頼区間をつけた母相関係数の推定値は  $0.665 \leq \rho \leq 0.996$  となる。

練習 20個のデータセットから標本相関係数  $r=0.755$  を得た。母相関係数を 95%の信頼区間をつけて、区間推定せよ。

② 検定

母相関係数  $\rho$  に関する検定は、たいていの場合、帰無仮説  $H_0: \rho = 0$ 、対立仮説  $H_1: \rho \neq 0$  とする無相関の検定である（2つの変数間に相関がないという帰無仮説を検定する）。

帰無仮説：両変数間には相関がない。母相関係数  $\rho = 0$

対立仮説：両変数間には相関がある。母相関係数  $\rho \neq 0$

帰無仮説が棄却されたときは両変数間には相関があると結論できる。

帰無仮説が棄却できなかったときは両変数間には相関があるとはいえないと結論できる。

母集団の母相関係数  $\rho = 0$  のときでも、そこから無作為に取り出した標本の相関係数が 0.5 程度のかかなり大きな値となることもよくある。

標本の相関係数  $r$  の分布は

$$\rho = 0 \text{ のとき } t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \text{ は自由度 } n-2 \text{ の } t \text{ 分布に従う。}$$

このことを利用して、 $t$  検定する。

エクセルでは

$$p\text{-値} = TDIST\left(\frac{ABS(r) \times \sqrt{n-2}}{\sqrt{1-r^2}}, n-2, 2\right) = TDIST\left(\frac{ABS(r) \times SQRT(n-2)}{SQRT(1-r^2)}, n-2, 2\right)$$

例：相関係数  $r=0.959$ （標本数  $n=6$ ）を得た。相関の有無を検定せよ。

帰無仮説：両変数間には相関がない。母相関係数  $\rho = 0$ 。母集団に相関がない。

対立仮説：両変数間には相関がある。母相関係数  $\rho \neq 0$ 。母集団に相関がある。

p-値は 0.002487 となるので、1%の有意水準で帰無仮説を棄却でき、相関があると認められる。

	B	C	D	E	F	G	H	I
123								
124	標本相関係数r	0.959						
125								
126	標本数n	6						
127								
128	P値	0.002487	=TDIST((ABS(C124)*SQRT(C126-2))/SQRT(1-C124^2),C126-2,2)					
129								

練習：以下の相関係数について，相関の有無を検定せよ．

①  $r = 0.667$ ,  $n = 12$

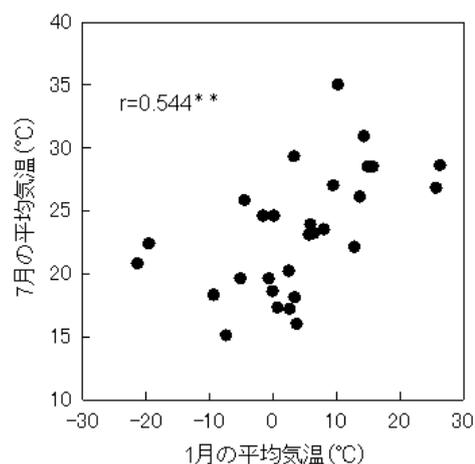
②  $r = -0.403$ ,  $n = 40$

相関係数  $r$  の検定の結果，相関が有意であることがわかったら，相関自体の強さは相関係数の絶対値で判断する．おおむね次のように考える．

- 1.000~-0.600 高い負の相関
- 0.599~-0.400 中位の負の相関
- 0.399~-0.200 低い負の相関
- 0.199~+0.199 無相関
- +0.200~+0.399 低い正の相関
- +0.400~+0.599 中位の正の相関
- +0.600~+1.000 高い正の相関

したがって，相関係数が 1%あるいはそれより小さい有意水準で有意であったとしても，相関係数自体の値が 0 に近ければ，2つの変数間の相関はあまり大きいとはいえない．標本数が多くなると，相関係数がかなり 0 に近くても有意にはなるので，この点に注意すること．

論文などで相関係数に\*や\*\*が付いていることをよく見るが，母相関係数が 0 でないことを検定しており，ふつう\*は 5%の有意水準で相関があるとき，\*\*は 1%の有意水準で相関があることを示している．



### 3. みかけの（偽の）相関関係

相関係数が高いからといって，両者の間に因果関係などが必ずあるとは限らない．例えば，年齢を問わずに調査したら，血圧と垂直飛びに負の相関関係があるかもしれない．しかし，加齢とともに血圧は上がり，運動能力は落ちるから，この関係は見かけのものでしかない．あるいはテレビの普及率と米の消費量を 1960 年代について調べたら，負の相関があるだろう．一般に時間の絡むデータでは見かけの相関関係の出ることがよくある．

① 時系列データ

1955年から1970年におけるテレビの販売数と自動車事故の数  
1930年から1970年におけるタバコの消費本数と平均寿命

以上のことを調べるとどういふ結果が得られるか？  
その結果から、どういふ誤った結論が引き出せるか？

② 年齢などに関わるデータ

血圧と原宿あるいは巣鴨で遊ぶ時間

③ その他

小学1～6年生までの身長と体重の相関関係は同年代だけの相関係数よりもかなり大きくなる。

4. 相関分析の手順

- ① 2つの変量間の相関係数  $r$  を計算する
- ②  $\rho=0$  という帰無仮説を検定し、相関関係が有意であるかを調べる
- ③ 有意であれば、相関の強さを相関係数の大きさから評価する。相関があっても、それは2つの変量間に必ずしも何らかの関係があることを証明するわけではない。

注意点：2つの変量間に実際にどんな結びつきがあるのかを相関分析の後、考える。

D. 宿題

1. 自分の調べた2つの変量の間関係について、両方のデータともに以下の①～③までを行え。

- ① 散布図を書く上での注意にしたがって、散布図を書き直せ。
- ② 95%信頼区間をつけて母相関係数を区間推定せよ。
- ③ 相関の有無（帰無仮説  $\rho=0$ ）を検定せよ。

2. 自分の調べた相関についてのデータに相関があるときはどういふ関係があるのかを考えよ。

すなわちみかけの相関ではないのか、直接的な因果関係があるのか、あるいは何か別の要因が間に入っていないか？相関がないときは（始めに相関があると予想したにもかかわらず）なぜ相関がなかったのかを考えよ。