生物統計学
2011 年 1 月 11 日

第１２回　相関分析

Ａ．２つの変数間の関係を調べる

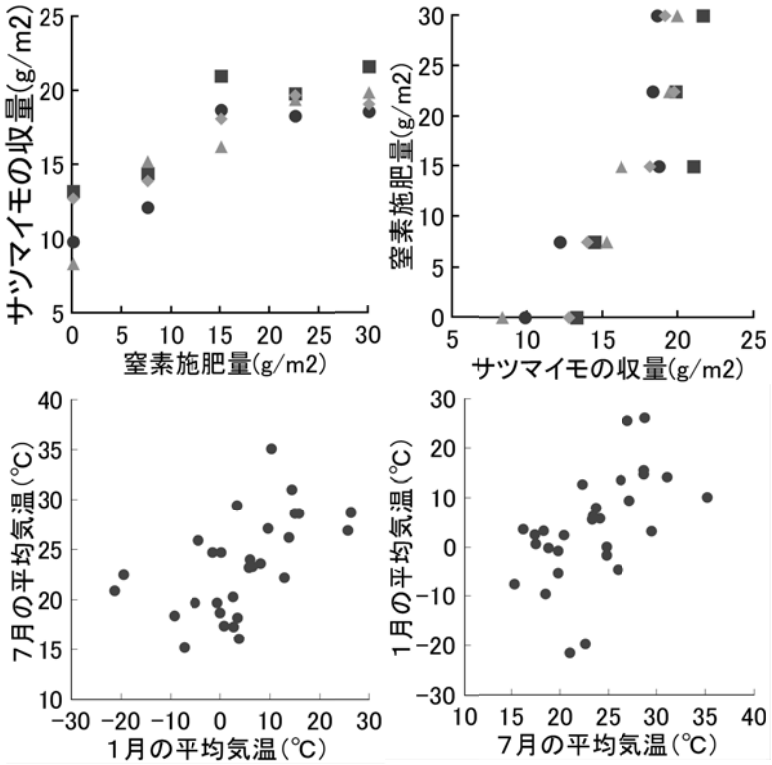
１．散布図を書く

例１　水稲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？

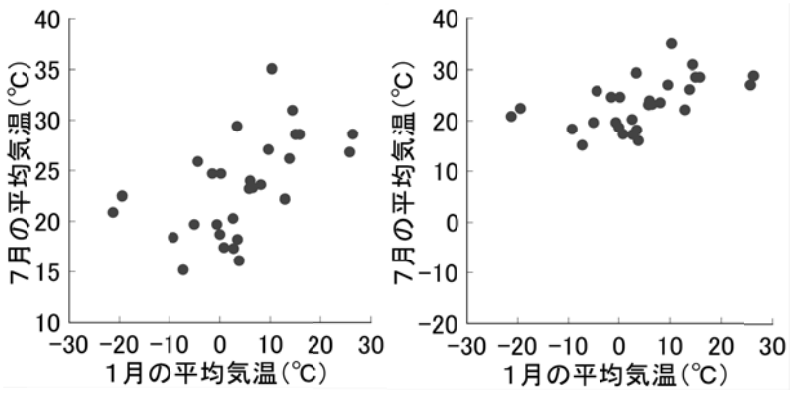
例２　トマトの糖度は施肥量によってどのように変化するかを知りたい．例えば，窒素施肥量を増加させると糖度はどうなるか？

①　散布図の書き方

１）x 軸（横軸）には原因となる変量を，y 軸（縦軸）には結果となる変量をとる．

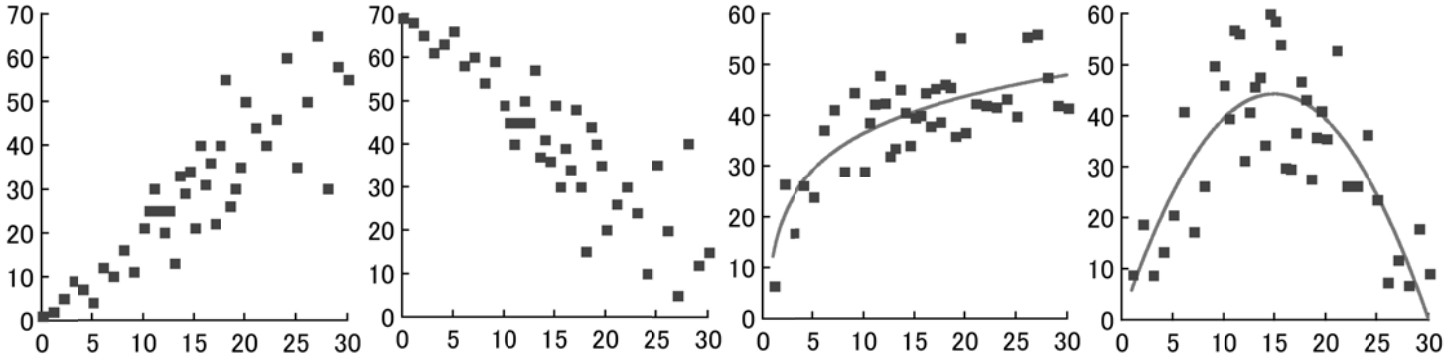


２）できるだけ点が図全体にばらつくように軸の上下限を決める．



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

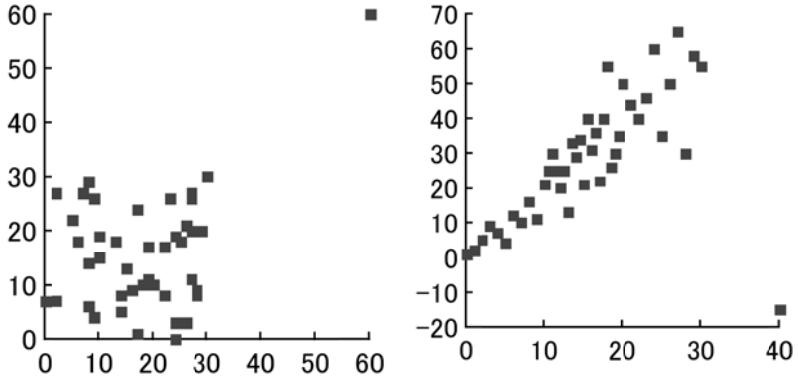
③ 指数・対数・双曲線など

④ その他

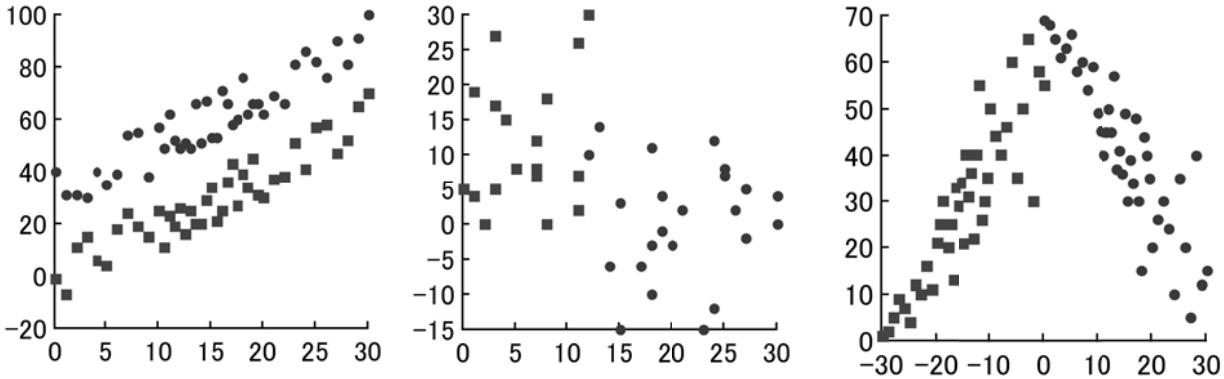
今回の授業では直線的関係についてしか学ばないが，２つの変数間の関係を曲線で表す方がよい場合も多い．

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．

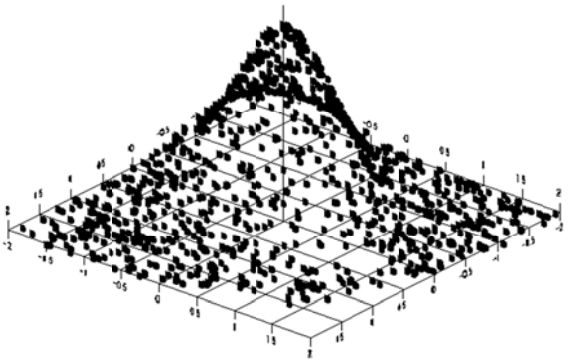
相関では両変数間の関連の度合いを相関係数で評価することを主な目的とする．回帰では相関係数で評価することもできるが，主たる目的は両変数間の数的関係を回帰直線で表し，ある x が指定されたときに y がいくつになるかを求めることである．

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の関係を考えよう．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，相関分析は不適当である．

次の例は相関か回帰か？

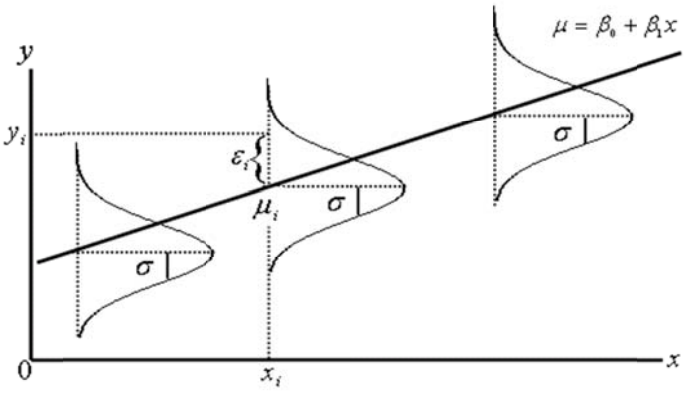
最高気温と最低気温
身長と体重
飼料中の脂肪含量と牛の乳脂肪率
テレビを見る時間と血圧
テレビを見る時間とエンゲル係数
塩分摂取量と血圧



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 は指定できる変量なので，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

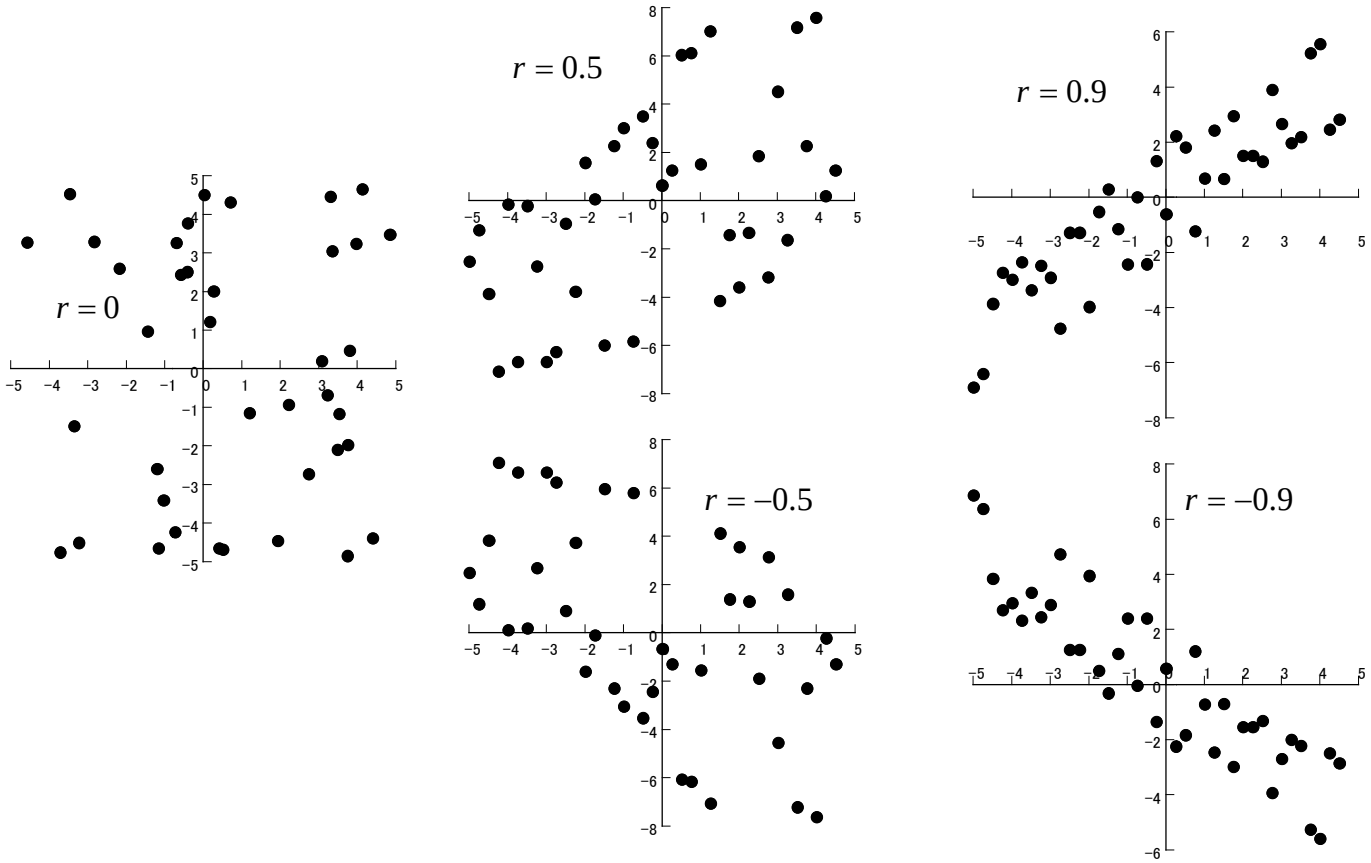
２つの変数間の**直線的な**関係（相関関係）は相関係数 r によって定量的に示すことができる．

相関係数には以下の性質がある

① $-1 \le r \le 1$ 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．

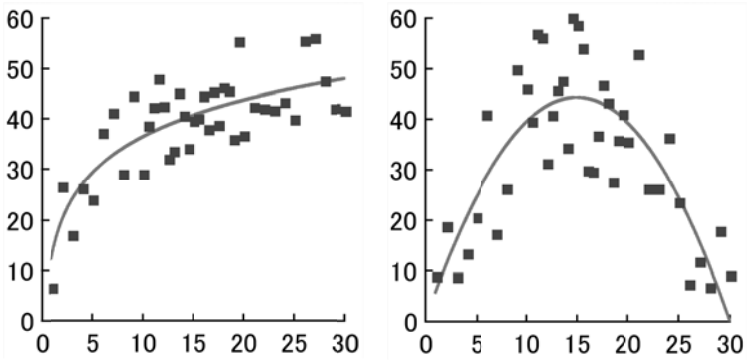
③ r が 0 に近いときは，両変数間には相関がない（無相関）．



相関係数　$r = \dfrac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$　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を評価する．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

	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.

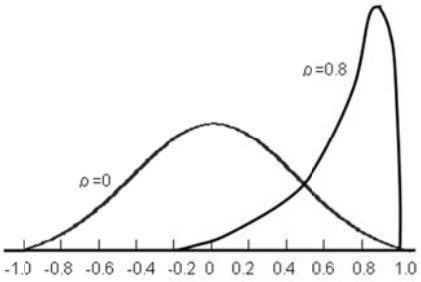
	A	B	C	D	E
1					
2	北半球各地点での1月の気温と7月の気温の関係				
3			1月	7月	
4		オスロ	-7.5	15.2	
5		ロンドン	3.6	16.1	
6		パリ	3.3	18.2	
7		リヨン	2.4	20.3	
8		マルセイユ	6.3	23.3	
9		マドリード	5.8	24	
10		ライプチヒ	0.6	17.4	
11		ベルリン	-0.2	18.7	
12		ウイーン	-0.8	19.7	
13		ローマ	7.9	23.6	
14		アテネ	9.4	27.1	
15		イスタンブ	5.6	23.2	
16		モスクワ	-9.5	18.4	
17		ハバロフス	-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	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	25.6	26.9	
27		ハルピン	-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	13.6	26.2	
31		バンクーバ	2.5	17.3	
32		ニューヨーク	0	24.7	

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	1
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
ウイーン	-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	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	-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	3
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	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	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 ρ を r から区間推定することができる．なお母相関係数 ρ の点推定値は標本から得た相関係数 r と同じである．

その前に母相関係数 ρ 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．

右上の図のように母相関係数 ρ 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±1 に近づくと著しくゆがんだ分布をする．そのため，相関係数の区間推定値は点推定値±〇〇と表現はできない（分散，標準偏差の区間推定と同じである）．

次の例題で母相関係数 ρ の推定を行ってみよう．

例題　右のデータは 20 頭の羊について胴回りと体重を測定したデータである．95%信頼区間をつけて，母相関係数を区間推定せよ．

羊の胴回り(cm)	体重(kg)
125.5	37.2
130	46.3
150.5	71.4
151.5	70.6
132	57.8
152.5	69
125	34.7
141	60.8
131	47
124.5	38.9
146	55.8
123	29.8
125	37.5
148.5	57.4
145.5	59
129.5	44.4
137.5	55.4
146.5	67.2
135	55.6
142	59.8

母相関係数 ρ の推定の手順

（1）　点推定：$\hat{\rho} = r$

（2）　区間推定：信頼率 p%の ρ の信頼区間は生物統計学__授業用データ集のエクセルファイルにデータ（100 個以内）を入力すると，右のように計算できる．

点推定値 $\hat{\rho} = 0.931$

すなわち 95%信頼区間をつけた母相関係数の推定値は $0.831 \le \rho \le 0.973$ となる．

下の測定値に100以内のデータセットを入れると相関係数，信頼率p%のときの相関係数の区間推定，あるいは有意水準p%のときに帰無仮説：母相関係数=0（無相関）とした場合の有意確率p−値を自動的に計算する．ただし測定値に値を誤入力したときはDelキーで削除すること．セルを移動させると式が変わってしまう．

標本番号	測定値A	測定値B			
1	125.5	37.2	相関係数の区間推定		
2	130	46.3			
3	150.5	71.4	相関係数	0.93107998	
4	151.5	70.6			
5	132	57.8			
6	152.5	69	信頼率%	95	%
7	125	34.7			
8	141	60.8	相関係数上限	0.972789749	
9	131	47	相関係数下限	0.830914418	
10	124.5	38.9			
11	146	55.8			
12	123	29.8	相関係数の検定		
13	125	37.5			
14	148.5	57.4	有意確率p−値	2.59012E-09	
15	145.5	59			
16	129.5	44.4			
17	137.5	55.4			
18	146.5	67.2			
19	135	55.6			
20	142	59.8			

練習１　右のデータは 20 個のジャガイモについて重量と芽の数を測定したデータである．95%信頼区間をつけて，母相関係数を区間推定せよ．

重量(g)	芽の数(個)
121.68	8
123.62	10
120.84	9
125.59	11
113.39	6
132.47	10
123.11	11
127.02	13
126.17	12
121.11	6
131.58	10
122.45	8
137	9
117.47	12
155.34	7
129.81	9
132.92	12
142.46	9
136.88	11
138.77	9

練習２　右のデータは 20 匹のミミズについて長さと重さを測定したデータである．99%信頼区間をつけて，母相関係数を区間推定せよ．

ミミズの長さ(cm)	ミミズの重さ(g)
9.7	0.973
7.4	0.421
10.2	0.453
6.9	0.412
9.3	0.453
10.5	1.093
4.2	0.231
5.3	0.621
10.2	0.593
5.3	0.193
9.7	0.942
4.5	0.132
7.8	0.695
6.3	0.823
5.4	0.621
10.3	0.741
7.2	0.632
3.4	0.348
7.2	0.731
5.6	0.554

② 検定

母相関係数 ρ 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 程度のかなり大きな値となることもよくある．

例題　右のデータは 20 頭の羊について胴回りと体重を測定したデータである．相関の有無を検定せよ．

羊の胴回り(cm)	体重(kg)
125.5	37.2
130	46.3
150.5	71.4
151.5	70.6
132	57.8
152.5	69
125	34.7
141	60.8
131	47
124.5	38.9
146	55.8
123	29.8
125	37.5
148.5	57.4
145.5	59
129.5	44.4
137.5	55.4
146.5	67.2
135	55.6
142	59.8

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．

相関の有無に関する検定は生物統計学__授業用データ集のエクセルファイルにデータ（100 個以内）を入力すると，以下のように計算できる．

標本番号	測定値A	測定値B			
1	125.5	37.2	相関係数の区間推定		
2	130	46.3			
3	150.5	71.4	相関係数	0.93107998	
4	151.5	70.6			
5	132	57.8			
6	152.5	69	信頼率%	95	%
7	125	34.7			
8	141	60.8	相関係数上限	0.972789749	
9	131	47	相関係数下限	0.830914418	
10	124.5	38.9			
11	146	55.8			
12	123	29.8	相関係数の検定		
13	125	37.5			
14	148.5	57.4	有意確率p-値	2.59012E-09	
15	145.5	59			
16	129.5	44.4			
17	137.5	55.4			
18	146.5	67.2			
19	135	55.6			
20	142	59.8			

p-値は 2.59×10^{-9} となるので，0.1%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．

練習１　右のデータは 20 個のジャガイモについて重量と芽の数を測定したデータである．相関の有無を検定せよ．

重量(g)	芽の数(個)
121.68	8
123.62	10
120.84	9
125.59	11
113.39	6
132.47	10
123.11	11
127.02	13
126.17	12
121.11	6
131.58	10
122.45	8
137	9
117.47	12
155.34	7
129.81	9
132.92	12
142.46	9
136.88	11
138.77	9

帰無仮説：

対立仮説：

p-値は（　　）である．したがって，帰無仮説は棄却（　される　・　されない　）ので，相関は（　ある　・　ない　・　あるとはいえない　・　ないとはいえない　）．

練習２　右のデータは 20 匹のミミズについて長さと重さを測定したデータである．相関の有無を検定せよ．

ミミズの長さ(cm)	ミミズの重さ(g)
9.7	0.973
7.4	0.421
10.2	0.453
6.9	0.412
9.3	0.453
10.5	1.093
4.2	0.231
5.3	0.621
10.2	0.593
5.3	0.193
9.7	0.942
4.5	0.132
7.8	0.695
6.3	0.823
5.4	0.621
10.3	0.741
7.2	0.632
3.4	0.348
7.2	0.731
5.6	0.554

帰無仮説：

対立仮説：

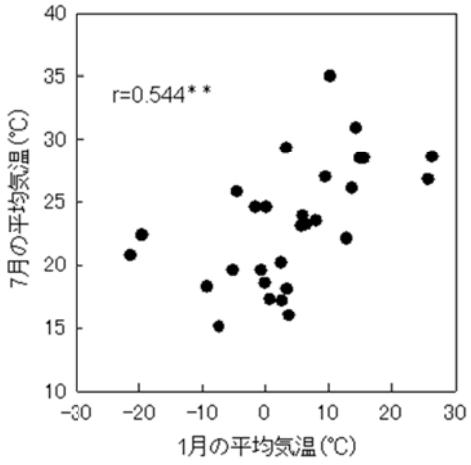
p-値は（　　）である．したがって，帰無仮説は棄却（　される　・　されない　）ので，相関は（　ある　・　ない　・　あるとはいえない　・　ないとはいえない　）．

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，すなわち相関の有無を検定しており，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てくることがよくある．

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う結果が得られるか？

その結果から，どういう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．