相関係数・回帰分析の解釈について

相関係数はなぜ標準偏差で割るのか

第 115 回数実研において、旭川南高校の岡崎知之先生のレポート「相関係数はなぜ標準偏差で割るのか」を読み相関関係に興味を持った。更に、相関関係は回帰分析や分散分析に繋がり近年の人工知能に活用されている。医学分野においても高 IF(Impact Factor)である NEJM(New England and Journal Medicine) LANSET,JAMA(Journal of American Medicine Association)等の医学論文の中にも多く統計的手法が見られる。

人間は様々な「現象」に「解釈」をして進歩して来た歴史がある。
現代人にとって、統計分野の知識は欠くことのできない「解釈」の大きな武器になって来ている。

今回は相関係数と回帰分析について考えてみた。

松本睦郎（数実研会員）

Episode 1 回帰分析

大きく回帰分析は次のようなものがある。回帰分析の目的は、「予測」、「判別」「関連性の強弱」等がある。

目的変数 / 説明変数	量的データ（予測）	質的データ（判別）
量的データ	・単回帰分析 ・重回帰分析	・判別分析 ・ロジステック回帰分析
質的データ	・数量化Ⅰ類	・数量化Ⅱ類

Episode 2 相関係数

2 変量 x, y を持つサイズnのデータについて、変量(x, y)の値が$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$のときの基本的な用語の定義について

$$x_i - \bar{x}$$

偏差

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} {x_i}^2 - n(\bar{x})^2$$

偏差平方和

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

偏差積和

$$\sigma_{xy} = \frac{S_{xy}}{n} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$$

共分散

$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

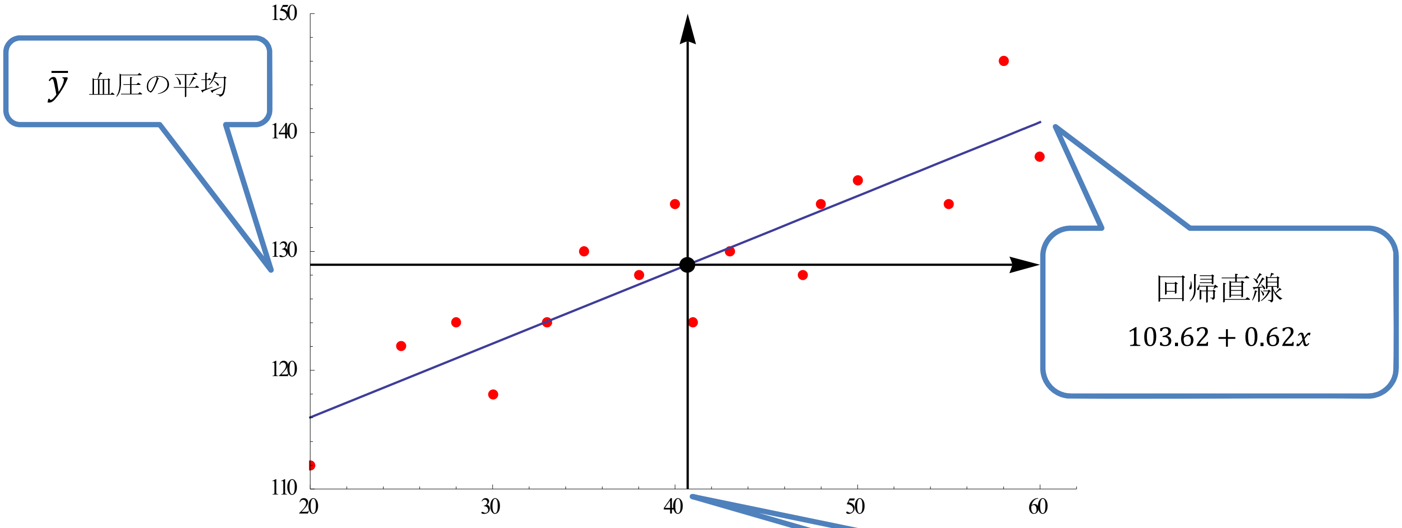
相関係数

$$y = \frac{S_{xy}}{S_{xx}}(x - \bar{x}) + \bar{y}$$

回帰直線

〔例題１〕年齢と血圧の相関関係（架空データ）

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
年齢	48	35	43	28	20	55	40	30	41	50	60	33	38	58	25	47
最高血圧	134	130	130	124	112	134	134	118	124	136	138	124	128	146	122	128



相関係数は 0.887 となる。かなり関連性があると思われる。

$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

旭川南高校　岡崎知之先生が述べた「統計における 2 つの理不尽」

（１）　何故、共分散を標準偏差の積で割るのか。

（２）　何故、$-1 \leq \mathrm{r} \leq 1$ が成立するのか。

について考察してみた。

2 つの変量の偏差を成分とするベクトルで表示する。

$$\vec{x} = (x_1 - \bar{x}, x_2 - \bar{x}, \cdots, x_n - \bar{x}), \vec{y} = (y_1 - \bar{y}, y_2 - \bar{y}, \cdots, y_n - y)$$

とおくと、偏差積和は内積となる。

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \vec{x} \cdot \vec{y}$$

同様にしてそれぞれの偏差平方和もベクトルの大きさとして表示できる。

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = |\vec{x}|^2, S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = |\vec{y}|^2$$

相関係数を 2 種類の変量ベクトルとして表示できる。

$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} = \frac{\vec{x}\cdot\vec{y}}{|\vec{x}||\vec{y}|}$$

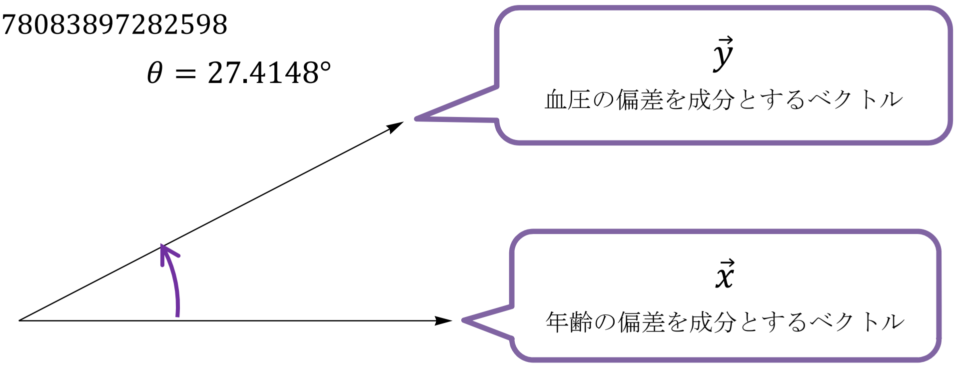
変量ベクトル$\vec{x}$と$\vec{y}$のなす角をθとおくと、$\vec{x}\cdot\vec{y} = |\vec{x}||\vec{y}|cos\theta$

$$\therefore r = cos\theta$$

〔解釈 1〕『相関係数はベクトルとベクトルのなす角の余弦である。』

〔例題 1〕の相関係数は、$r = 0.8878083897282598$

$$\theta = 27.4148°$$



〔結論〕2 種類のデータの偏差を成分とするベクトルのなす角の余弦が相関係数を表す。

Episode3 相関関係の精度

1. 相関係数の絶対値

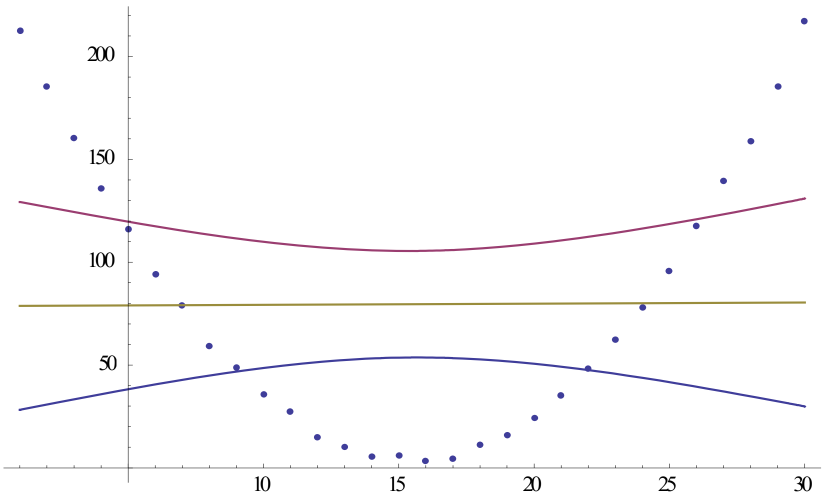
1 に近いほど散布図は直線に接近する。（誤差が少ない）

0 に近いほど散布図は面的に拡散する。（誤差が多い）

$$|cos\theta| \leq 1$$

2. 相関係数は単位を変えても一定。
3. 相関係数は直線の傾きには関係ない。
4. 相関係数は、データの分布がどれだけ**直線に近いか**を測定する指標である。
   変数間に直線的な関係が見られるときに利用できる。回帰直線は必要以上延長して解釈してはいけない。
5. 回帰が直線的でないとき利用できない。

〔例題 2〕二次関数に従った散布図の回帰直線は、相関係数は 0 になる。



6. 相関係数が高いことと、因果関係があることは別の問題である。相関係数が高いだけでは、因果関係には結びつかない。

Episode4 決定係数と相関係数

〔方法Ⅰ〕決定係数（寄与率）について

相関関係の精度を解釈する指標の一つに「決定係数」がある。

$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$
実測値データの偏差平方和

$$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = |\vec{y}|^2$$

予測値は、

$$\hat{y}_i = \frac{S_{xy}}{S_{xx}}(x_i - \bar{x}) + \bar{y}$$

予測値データの偏差平方和

$$S_{\hat{y}} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

予測値と実測値の誤差の平方和

$$S_e = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 = \sum_{i=1}^{n} {e_i}^2$$

決定係数R^2は

$$R^2 = \frac{S_{\hat{y}}}{S_{yy}}$$

と定義される。

〔決定係数（寄与率）と相関係数の関係〕について

回帰直線

$$\hat{y}_i = \frac{S_{xy}}{S_{xx}}(x_i - \bar{x}) + \bar{y}$$

より

$$\hat{y}_i - \bar{y} = \frac{S_{xy}}{S_{xx}}(x_i - \bar{x})$$

$$S_{\hat{y}} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n} \left\{ \frac{S_{xy}}{S_{xx}}(x_i - \bar{x}) \right\}^2 = \frac{{S_{xy}}^2}{{S_{xx}}^2} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{{S_{xy}}^2}{S_{xx}}$$

決定係数（寄与率）の定義より

$$R^2 = \frac{S_{\hat{y}}}{S_{yy}}$$

$$R^2 = \frac{1}{S_{yy}} \times \frac{{S_{xy}}^2}{S_{xx}} = \frac{{S_{xy}}^2}{S_{xx} S_{yy}}$$

回帰係数は、

$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

であったので、

$$R = r^2$$

〔結論〕回帰係数の 2 乗は、決定係数（寄与率）になる。

Episode5 回帰方程式の有効性の検定

〔方法Ⅱ〕分散分析表(ANOVA)について(by Excel)

Excel による ANOVA

「データ」＞「分析ツール」＞「回帰分析」＞目的変数(y)の範囲、説明変数(x)の範囲＞出力先のセルの指定＞OK

$$R^2 = \frac{S_{\hat{y}}}{S_{yy}}$$

決定係数（寄与率）

$0.8 \leq R^2$→よい精度

$0.5 \leq R^2 < 0.8$→まあ良い

回帰統計	
重相関 R	0.887808
重決定 R2	0.788204
補正 R2	0.773075
標準誤差	3.950776
観測数	16

$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

相関係数

説明変数の数

$S_{\hat{y}}$:予測値の変動

分散分析表

	自由度	変動	分散	された分	有意 F
回帰	1	813.2292	813.2292	52.10126	4.44E-06
残差	14	218.5208	15.60863		
合計	15	1031.75			

仮設検定を利用した

p 値

帰無仮説の下で標本の状態が起こる確率

S_e:残差の変動

残差分散=変動/自由度

観測された分散比の値

検定統計量

この回帰分析の結果が採用されるべきかどうかを判定する方法を〔回帰分析の仮設検定〕という。

帰無仮説:H_0　「資料にはこの回帰分析が当てはまらない。」

対立仮設:H_1　「資料にはこの回帰分析が当てはまる。」

「観測された分散比」は先人達の研究によって F分布に従っていることがわかっている。

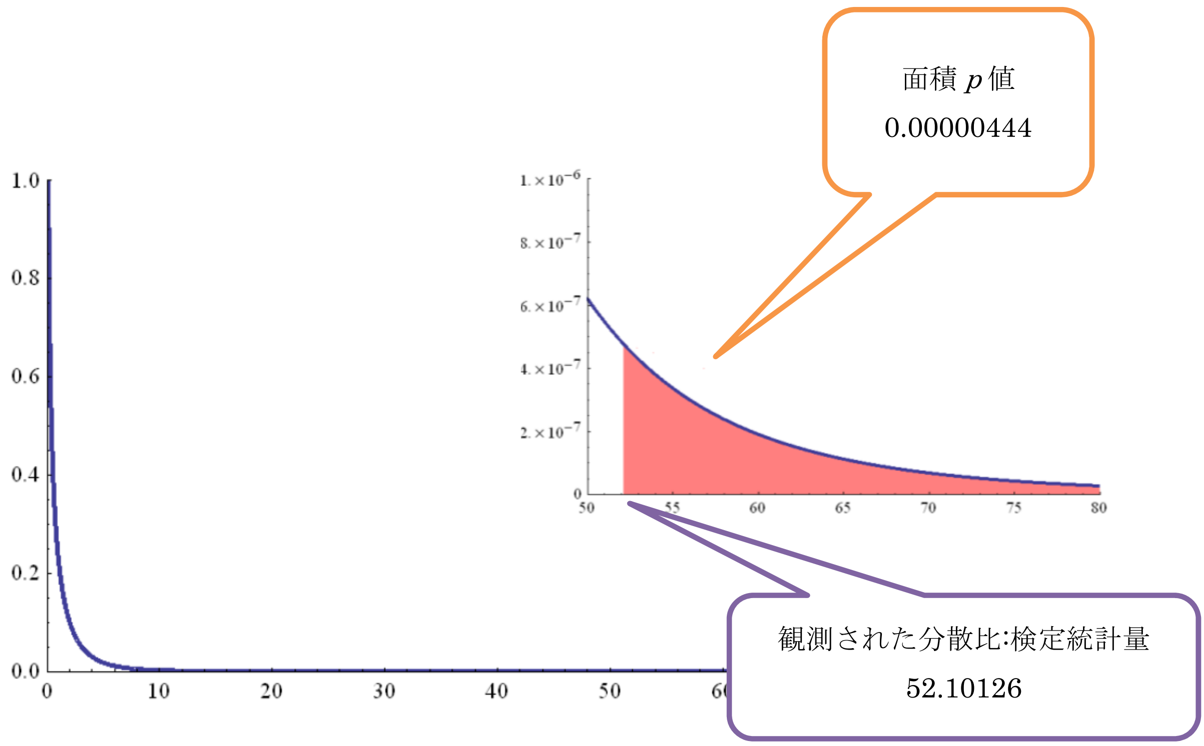
帰無仮説の下で〔例題 1〕の標本は、自由度（1,14）の F分布に従う。

〔注釈〕F 分布とは、

分散が等しい 2 つの正規分布に従う母集団からランダムに抽出した大きさ(n_1, n_2)の 2 組の標本から求めた標本不偏分散S_1^2, S_2^2 とおくと

$$F = \frac{{S_1}^2}{{S_2}^2}$$

は自由度対$(n_1 - 1, n_2 - 1)$の F 分布に従う。



$p > 0.05$ ⇒ H_0を採択 ⇒ この回帰直線は却下

$p < 0.05$ ⇒ H_0を棄却 ⇒ この回帰直線は有効

〔結論〕例題 1 の回帰直線は却下されず、有効と判断される。

Episode6 飲酒と肺癌との相関関係

〔例題 2〕22 人の患者さんの肺がんと飲酒量の相関関係を調べた。横軸（x 軸）に 1 週間当たりの平均飲酒量、縦軸（y 軸）に肺がんマーカー値をとった相関をグラフにした。

（架空のデータ）

Smoke=0：非喫煙者，Smoke=1：喫煙者

)ut[251]//TableForm=

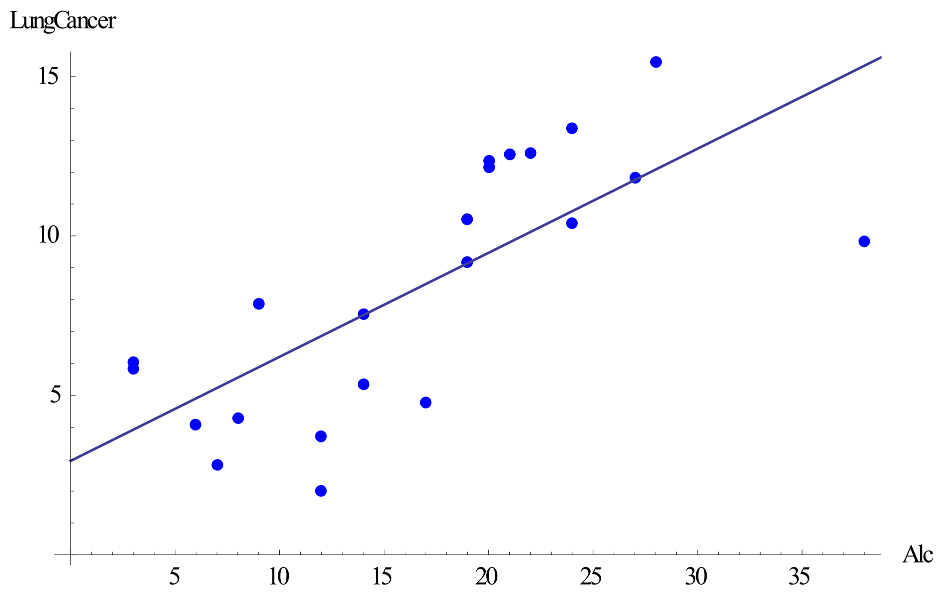
	Alc	LungCan	Smoke
1	3	6.02479	0
2	7	2.81784	0
3	8	4.28487	0
4	12	3.72965	0
5	12	1.99068	0
6	3	5.82547	0
7	14	5.34664	0
8	9	7.88686	0
9	6	4.06798	0
10	14	7.55388	0
11	17	4.76667	0
12	27	11.826	1
13	24	13.3582	1
14	20	12.3714	1
15	21	12.5463	1
16	19	10.5403	1
17	24	10.403	1
18	38	9.84527	1
19	22	12.583	1
20	20	12.1381	1
21	19	9.16425	1
22	28	15.4553	1

相関係数： 0.7323164846959355

回帰直線：2.9477279308863658 + 0.32609365809400537x

ANOVA Table(by Mathematica) 〔補正なし〕喫煙者と非喫煙者を区別しないで回帰分析を実施

""	"DF"	"SS"	"MS"	"F- Statistic"	"P- Value"
x	1	175.53834140813007	175.53834140813007	23.130166089864293	0.00010666645107337822
"Error"	20	151.7830358209589	7.589151791047946		
"Total"	21	327.32137722908897			



p 値＜0.05 より帰無仮説H_0を棄却

この回帰直線は有効である。有意差がある。

〔質問 1〕この現象から「肺がんと飲酒量は相関がある。」と結論して良いか？
散布図と ANOVA から、どのような解釈ができるか？

〔解釈例 1〕

① 飲酒量（Alc）の p 値は 0.0001 なので、帰無仮説は棄却される。つまり、この回帰分析は有効である。

② 回帰直線の傾きは、0.32609365809400537なので、1 週間の飲酒量が 1 杯増えると肺癌マーカー値が 0.326 の上昇が期待される。

③ 「飲酒量」と「肺癌マーカー値」は相関がある。

〔解釈例 2〕

① 患者背景を全く無視している。説明変数が 1 つだけで結論をだすのは危険である。

② 情報不足である。〔解釈 1〕だけでは、エビデンスが示されたとは言えない。

Episode7 バイアスの出現

〔解釈 2〕が正しい判断と言える。補う方法として、層別解析と多変量回帰分析がある。

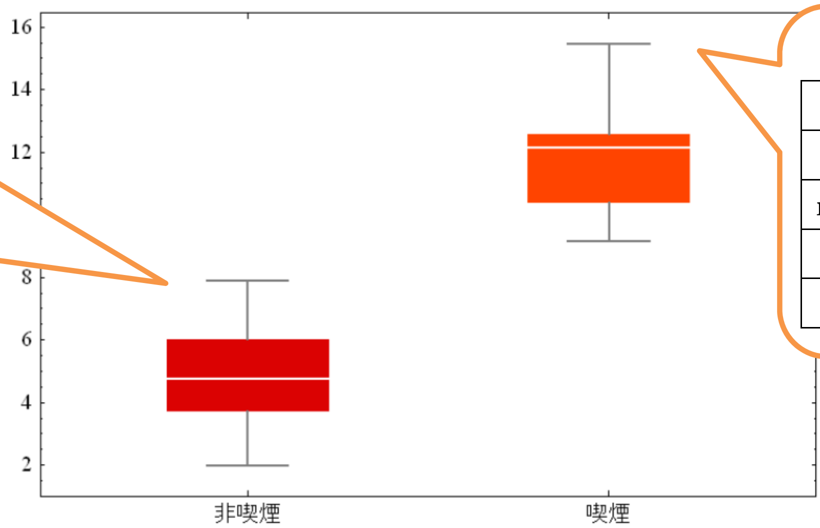
〔層別解析〕

「喫煙グループ」と「非喫煙グループ」のサブグループ別に肺癌マーカー値を求めた

既に『喫煙と肺癌の因果関係がある。』というエビデンスが存在する。

非喫煙グループ

Max	7.88686
75%	6.02479
Median	4.76667
25%	3.72965
min	1.99068



喫煙グループ

max	15.4553
75%	12.583
median	12.1381
25%	10.403
min	9.16425

〔多変数回帰分析〕

回帰モデルに「y:喫煙」（smoke）を説明変数として加える。

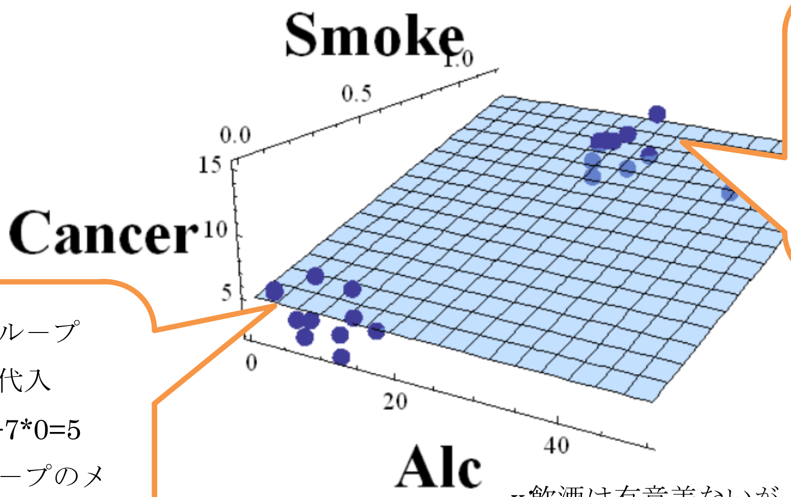
ANOVA Table(by Mathematica)　〔補正あり〕喫煙者と非喫煙者を区別

""	"Estimate"	"Standard Error"	"t- Statistic"	"P- Value"
1	5.000705166796523	0.9494400931834008	5.267004419446347	0.00004394306215778483
x	−0.0067851031882070634	0.080434866823583	−0.08435524861486703	0.933656471876612
y	7.000091472777136	1.3934989472259116	5.023392006655239	0.00007545572102971896

回帰平面式

$$5.000705166796523 - 0.0067851031882070634x + 7.000091472777136y$$

x=Alc，y=Smoke（0;非喫煙，1:喫煙），z=Cancer（肺癌マーカ値）

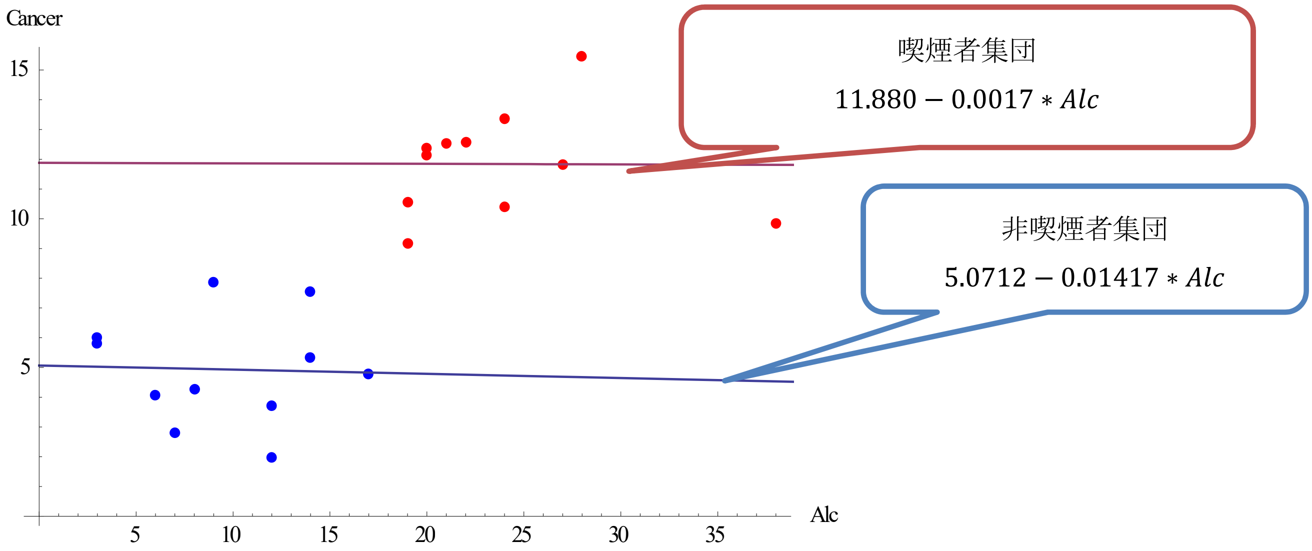


喫煙グループ
y=1 を代入
5+0*Alc+7*1=12
喫煙グループのメジアンと近い

非喫煙グループ
y=0 を代入
5+0*Alc+7*0=5
非喫煙グループのメジアンと近い

x:飲酒は有意差ないが、y:喫煙は有意差が見られる・

〔各層別の単回帰分析〕



ANOVA Table(by Mathematica)

非喫煙者集団〔補正なし〕x=Alc

""	"Estimate"	"Standard Error"	"t- Statistic"	"P- Value"
1	5.071287701947499	1.3845698437149352	3.66271714277753	0.0052134679289054815
x	−0.014179464013547749	0.13163334630006754	−0.10771939187221341	0.9165814492717813

喫煙者集団〔補正なし〕x=Alc

""	"Estimate"	"Standard Error"	"t- Statistic"	"P- Value"
1	11.880982127880182	2.578500337184984	4.607710131560797	0.0012764485555184688
x	−0.0017547229262673853	0.10561948916455237	−0.016613628224745276	0.9871073235580949

「非喫煙集団」と「喫煙集団」の p 値＞0.05 となる。

帰無仮説H_0「資料にはこの回帰分析が当てはまらない。」は棄却されない。

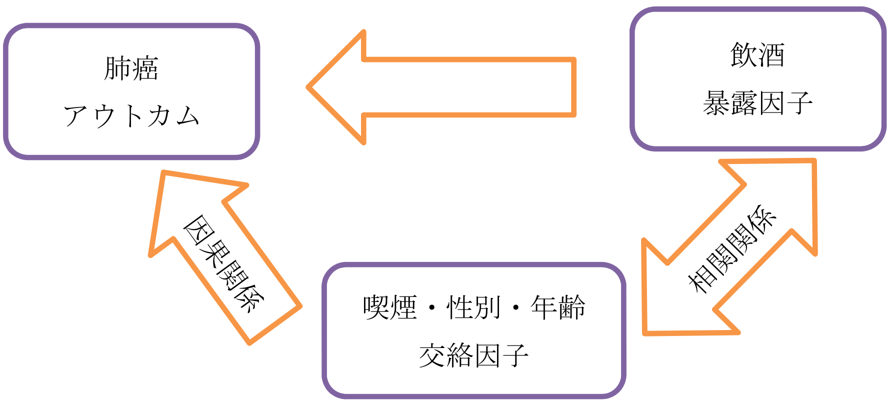
飲酒と肺がんとの相関関係は無い。

〔飲酒が交絡因子 confounding になる〕:患者背景（喫煙）が比較群間で偏り（バイアス）が発生する。

① 飲酒と肺癌には相関がない。

② 肺癌と喫煙には関連がある。

③ 喫煙者は飲酒量が多い。



結論：正確な解釈をするのには、単回帰分析よりも多変数回帰分析の方が良い。

〔交絡の例〕

① コーヒーの飲用と心筋梗塞との関係。

② ブロッコリーの摂取と高脂肪血症との関係。

③ 小学校でのボール投げと漢字テストとの関係。

④ 景気指数とサザエさんの視聴率は。負の相関関係がある。

相関関係が高いことだけで、因果関係があると解釈することは危険である。
交絡因子を調整するには、層別解析（サブグループ解析）と多変数回帰分析が必要となってくる。

//TableForm=

	Alc	LungCan	Smoke
1	1.40961	7.79862	0
2	2.65795	10.6224	0
3	1.72575	8.62198	0
4	3.04908	10.9086	0
5	2.15673	11.0176	0
6	2.66638	10.4908	0
7	1.13295	5.23828	0
8	2.18686	10.6239	0
9	2.22674	11.5406	0
10	2.24292	10.9251	0
11	0.854722	3.52403	0
12	1.31681	12.0382	1
13	2.05027	10.1072	1
14	1.2397	11.3382	1
15	1.65431	7.88453	1
16	3.04035	11.4092	1
17	1.37478	7.81916	1
18	3.0134	10.3375	1
19	1.36117	7.03306	1
20	2.94495	11.6502	1
21	1.80728	8.57033	1
22	3.74808	10.4621	1

Episode8 交互作用因子-インターアクション

〔例題 3〕22 人の患者さんの膵臓癌と飲酒量の相関関係を調べた。横軸（x 軸）に 1 週間当たりの平均飲酒量、縦軸（y 軸）に膵臓癌マーカー値をとった相関をグラフにした。（架空のデータ）

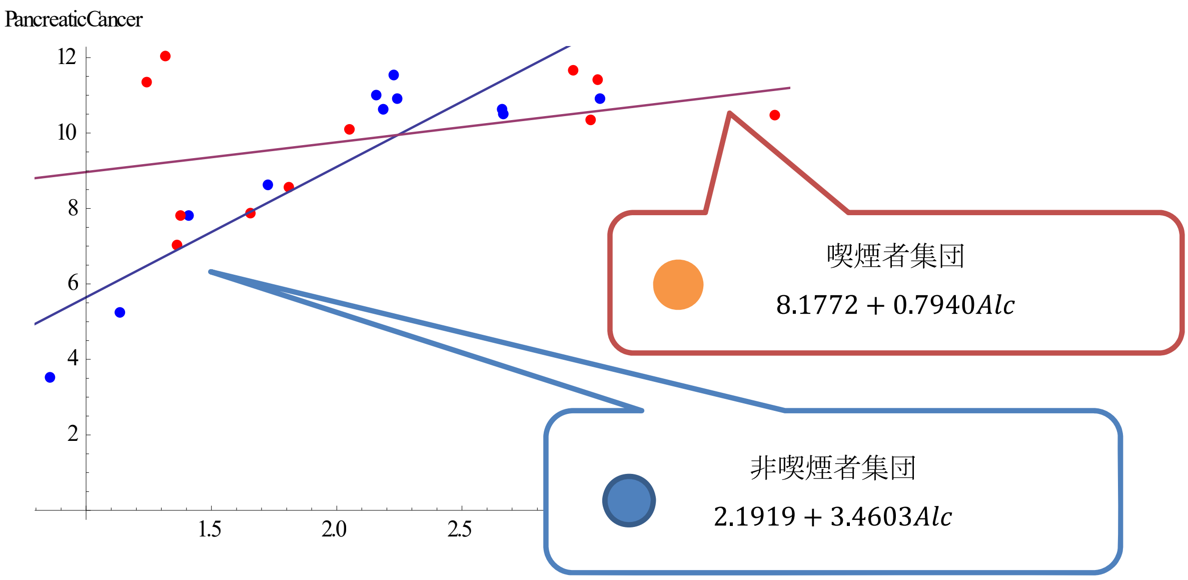
〔各層別の単回帰分析〕

非喫煙喫煙者〔補正なし〕 x=*Alc*

""	"Estimate"	"Standard Error"	"t- Statistic"	"P- Value"
1	2.1919568842276203	1.2870287409945245	1.7031141686345181	0.12274845167815664
x	3.460394027589656	0.6044362767042544	5.72499395049182	0.0002851503765784526

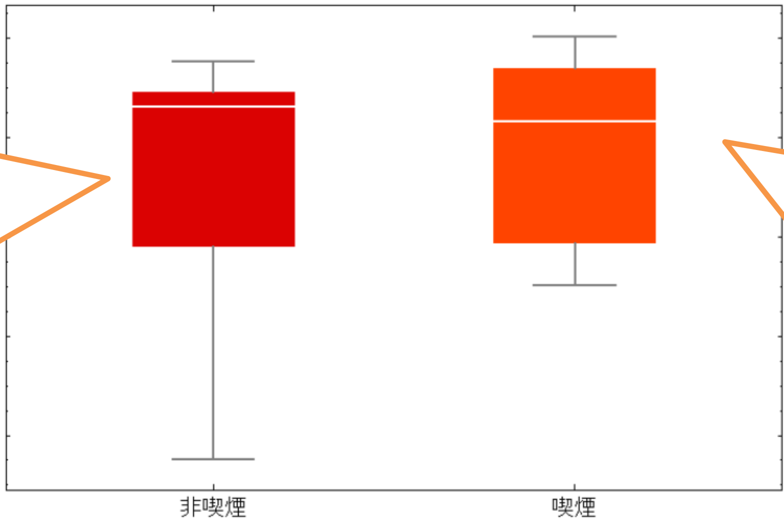
喫煙者〔補正なし〕 x=*Alc*

""	"Estimate"	"Standard Error"	"t- Statistic"	"P- Value"
1	8.177239561866347	1.3958082245717685	5.858426263661765	0.00024121359770875277
x	0.7940189401944431	0.6064804025631857	1.3092243984119813	0.22289527161737732



非喫煙グループ

Max	11.5406
75%	10.9251
Median	10.6224
25%	7.79862
min	3.52403



喫煙グループ

max	12.0382
75%	11.4092
median	10.3375
25%	7.88453
min	7.03306

喫煙者集団と非喫煙者集団の間には、大きな差が見られない。

喫煙者集団と膵臓癌マーカー値の分散分析表から見ると、p 値が 0.222 である。

帰無仮説H_0「資料にはこの回帰分析が当てはまらない。」は棄却されない。

飲酒と膵臓がんとの相関関係は無い。

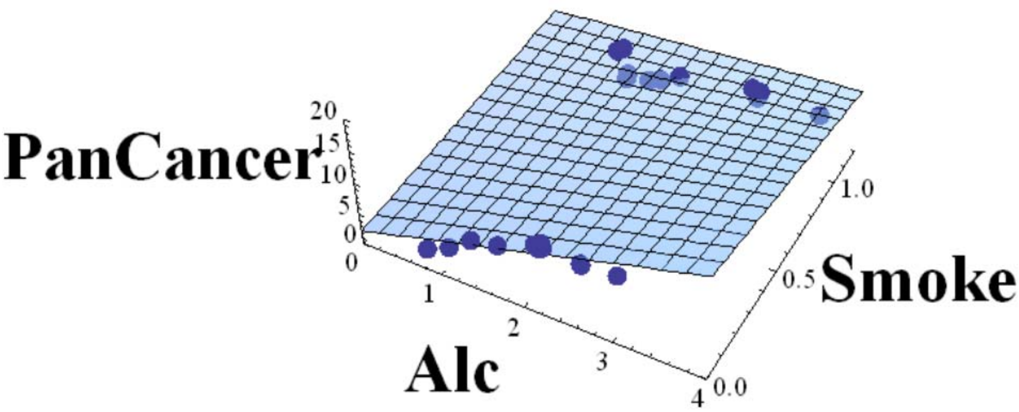
〔交互作用（インターアクション）解析〕

喫煙を表す変数 y と飲酒を表す変数 x に更に新しい変数$x \times y$を加えて回帰分析を行う。

ANOVA Table(by Mathematica)〔補正調整あり〕

""	"Estimate"	"Standard Error"	"t- Statistic"	"P- Value"
1	2.191956884227625	1.4972678305375118	1.463971134303155	0.16044776293373342
x	3.4603940275896545	0.7031723254445963	4.921118056518711	0.00011015764577867712
y	5.985282677638726	1.9459444132578854	3.0757726874727176	0.0065153944257306876
xy	−2.666375087395214	0.8866300844902659	−3.0073140242338434	0.007564114815944046

x=Alc,　y=$Smoke$（0;非喫煙，1:喫煙），z=Cancer（膵臓癌マーカ値）



回帰曲面

$$2.191956884227625 + 3.4603940275896545x + 5.985282677638726y - 2.666375087395214xy$$

非喫煙者集団
y=0 を代入した

$8.1772 + 0.7940Alc$

喫煙グループ
y=1 を代入した

$2.1919 + 3.4603Alc$

〔喫煙集団の回帰係数〕－〔非喫煙集団の回帰係数〕＝〔インターアクション項の回帰係数〕

0.7940－3.4603＝－2.6663

交互作用の項の p 値=0.00756

帰無仮説H_0「資料にはこの交互作用の回帰分析が当てはまらない。」は棄却される。

統計的に有意であり、飲酒と膵臓がんとの交互作用（インターアクション）の回帰方程式が認められる。

喫煙者か非喫煙者の違いによって飲酒と膵臓癌との影響が変化する。

〔交互作用因子の例〕

① あるワクチンは大人には有効であるが、子供には有害である。

② 納豆（ビタミンK）はワルファリンの効能を抑制する。納豆を食べないとワルファリンは有効であるが、食べると無効になる。

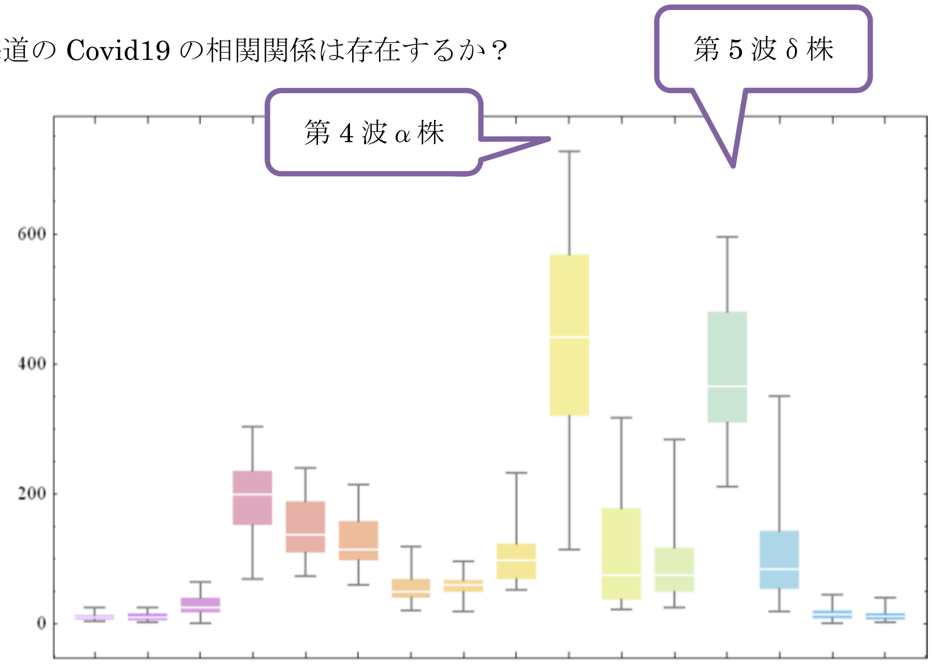
③ アルコールはハルシオン（睡眠導入剤）の効果を増強させる。

④ ある遺伝子を持った人には良い効果があるが、持たない人には効果がないワクチンがある。

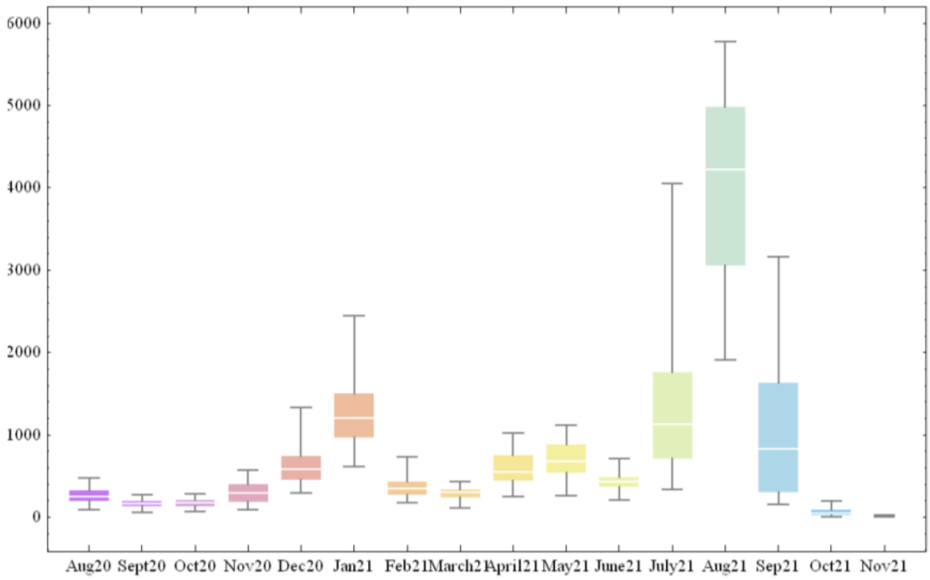
Episode9　Covid19

〔例題〕東京都と北海道の Covid19 の相関関係は存在するか？

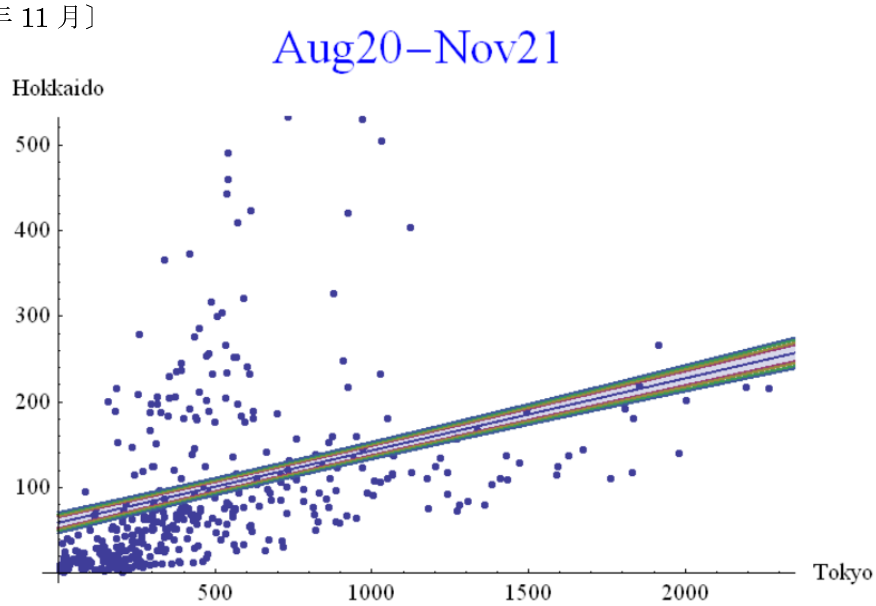
〔北海道→〕



〔東京都→〕



〔2020 年 8 月～2021 年 11 月〕



x=東京都感染者数

ANOVA Table(by Mathematica)〔補正調整なし〕

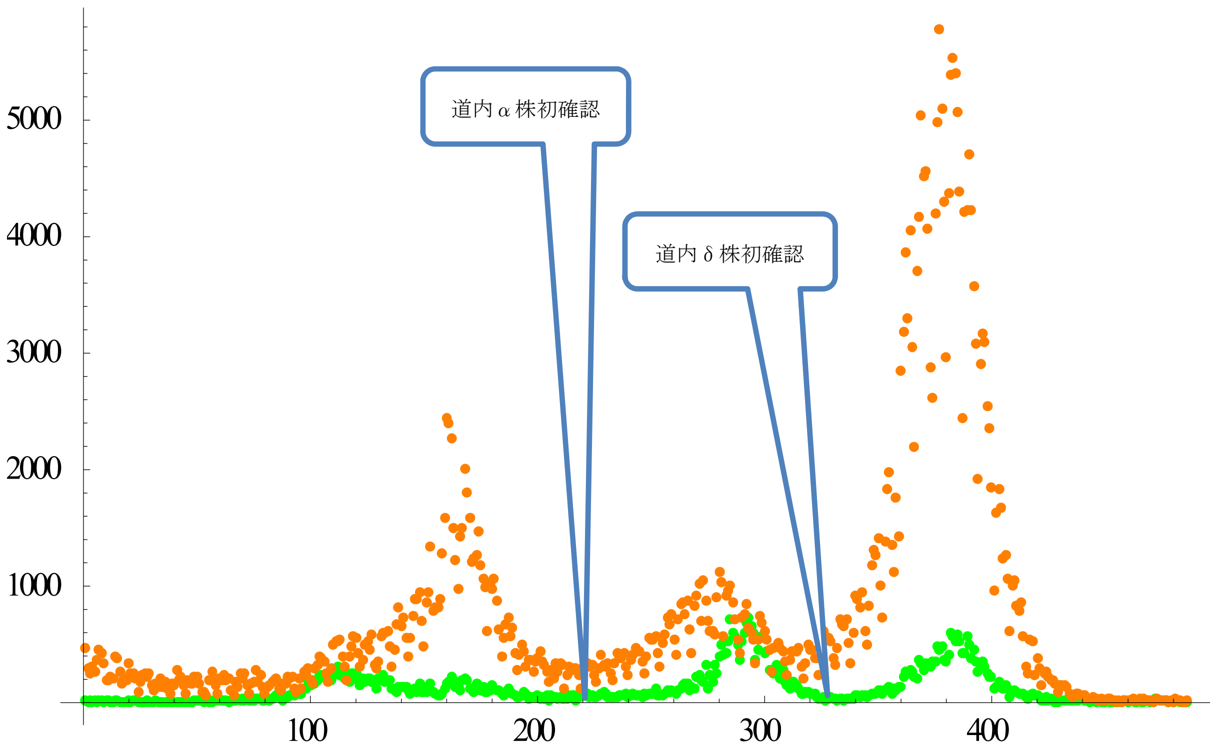
""	"Estimate"	"Standard Error"	"t- Statistic"	"P- Value"
1	58.99226428942755	6.193885432119229	9.524274372837962	$7.965104417190235\times10^{-20}$
x	0.08431062886580394	0.004791999166977777	17.594040801759373	$6.278998556352676\times10^{-54}$

相関係数は、0.6245660488175496となり、北海道と東京都のコロナ感染者数の相関関係は、まあああると判断して良い。

回帰直線の方程式は、$0.08431062886580394x + 58.99226428942755$

95%,90%,85%信頼区間を求めてグラフにした。（↑）

p 値<0.05 となり統計的にこの回帰係数は有効であると言える。



〔質問 2〕東京都と北海道のコロナウイルス感染者数についての解釈はどうする？

	相関係数	p 値	東京都	北海道	千歳空港旅客数
2020 年 8 月	0.478806	0.00643	8125	353	726575
2020 年 9 月	0.115924	0.541846	4918	326	799662
2020 年 10 月	0.174193	0.348654	5361	971	969219
2020 年 11 月	0.786895	0.000000251041	9869	5729	814706
2020 年 12 月	−0.214647	0.246224	19257	4593	522527
2021 年 1 月	0.734031	0.00000390421	38895	3856	386701
2021 年 2 月	0.739515	0.0000069185	10997	1603	361461
2021 年 3 月	0.427857	0.016348	9310	1788	685574
2021 年 4 月	0.8151	4.14917*10^-8	18090	3169	564921
2021 年 5 月	0.16113	0.386519	21829	13784	452415
2021 年 6 月	0.0105107	0.956039	12979	3367	434171
2021 年 7 月	0.961629	7.83013*10^-18	44061	3084	769988
2021 年 8 月	0.659442	0.0000546188	125603	12377	853086
2021 年 9 月	0.950825	1.97878*10^-21	31842	3432	646163
2021 年 10 月	0.346368	0.0562868	2156	511	900902
2021 年 11 月	0.333201	0.0719742	562	400	

〔月別感染者数と千歳空港利用者数の推移〕

変異株の α 株や δ 株による感染者が増加した 7 月と 9 月は、相関係数が 1 に接近している。
月別によって相関係数とその p 値は変化していることが分かる。
約 1 年間の東京都と北海道の月別感染者数の相関係数は、0.6539986319204625　となり相関があると判断して良い。

ANOVA Table(by Mathematica)〔補正調整なし〕

""	"Estimate"	"Standard Error"	"t- Statistic"	"P- Value"
1	1735.9939332493286	995.0366311194225	1.7446532911017802	0.10294740732803236
x	0.08676858593834542	0.026824284287317	3.2347027420735754	0.005992312047841879

p 値<0.05 となり統計的にこの回帰係数は有効であると言える。

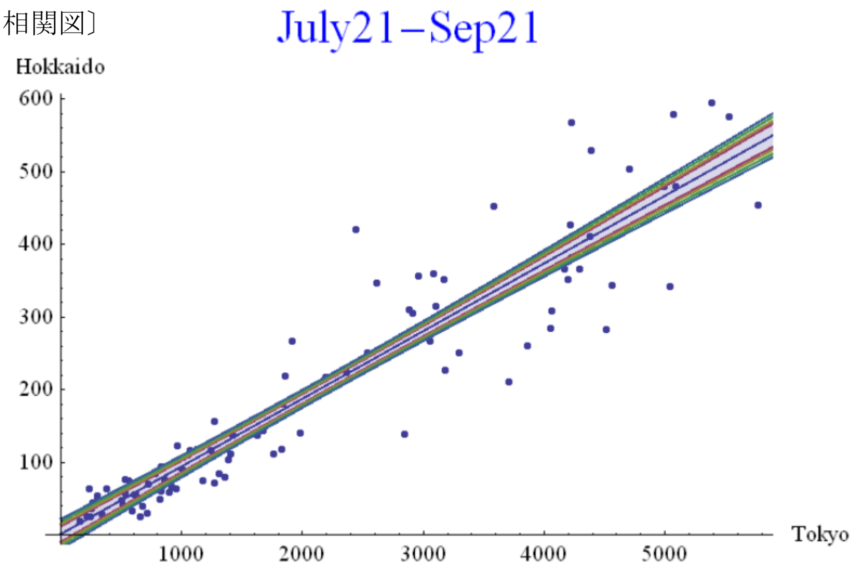
さらに千歳空港旅客数 y を変数に加え、更に変数 $x \times y$ も加えて交互作用（インターアクション）解析を行った。

ANOVA Table(by Mathematica)〔補正調整あり〕

""	"Estimate"	"Standard Error"	"t- Statistic"	"P- Value"
1	6025.246034476577	6056.54586948059	0.9948320650617481	0.34119836912762713
x	0.04104886370934555	0.25625233995340735	0.16018922487423606	0.8756357785160478
y	−0.0060876509115858155	0.00769201796585548	−0.7914244270630442	0.44542666196669556
xy	$5.588173422785678 \times 10^{-8}$	$3.067068451166582 \times 10^{-7}$	0.18219917526327709	0.8587407189193135

p 値>0.05 となり統計的にこの回帰係数は無効であると言える。
つまり、北海道の感染者数は東京都と千歳空港利用者数を加えると直線的な相関関係が認められないことになる。
1 年間の千歳空港旅客数の増減が北海道感染者数にはほとんど相関なしと解釈できる。

〔第 5 波デルタ株の相関図〕



ANOVA Table(by Mathematica)〔補正調整なし〕

""	"Estimate"	"Standard Error"	"t- Statistic"	"P- Value"
1	1.514683778214857	10.281407008226221	0.14732261615583828	0.8832069145020391
x	0.09306744758173072	0.0037544768947810445	24.788392681574425	$5.493639607084101 \times 10^{-42}$

回帰方程式は

$$1.514683778214857 + 0.09306744758173072x$$

東京オリパラがあった 7 月から 9 月にかけての東京都と北海道の感染者数の相関図である。

7 月と 9 月は、回帰係数が 1 に近く、p 値<0.05 となり統計的にこの回帰係数は有効であると言える。

千歳空港旅客数は因果関係が認められない。

〔推察〕北海道のコロナ感染者数は、変異株の発見から増加している。道内で 3 月 10 日に α 株が報告されて 2.5 カ月後の 5 月中旬にピークを迎えた。道内で 6 月 29 日にデルタ株の感染が報告されて、やはり 2.5 か月後の 8 月にピークを迎えた。国内では 11 月 30 日にオミクロン株が報告された。過去の例から、約 3 か月後 2022 年 2 月に道内に上陸し、2.5 カ月の 4 月にはピークを迎える可能性がある。

広い検査と感染者の隔離を徹底するのが重要である。

Episode10 最後に

「データサイエンス」とは社会に存在する大量のデータから因子間の関係を予測する分野である。

「因果推論」では因果関係を推定・検定する分野である。

人類が社会生活を営む様になってから、歴史を振り返るとその時代によって様々な解釈をして来た。

現代社会は、様々なデータをコンピューターを駆使して数学的に解釈してする世界である

医薬品の有効性を求める。傷病が陽性か陰性を判定する。危険か安全かを判定する。

数学は、いろいろな分野に活用されている。

高校数学での統計分野の指導手順が教条的なのが気になる。

〔参考図書〕

・「基礎医学統計学」南江堂

・「みんなの医療統計／多変量解析編」講談社

・「医学論文の難解な統計手法が手に取るようにわかる本」金原出版

・「意味がわかる多変量解析」ペレ出版