環境計測学 12.
実験計画の基本、統計分析の適用（検定と推定、回帰と相関）

１．１変量の統計手順

1) 正規分布に従っているか。
　グラフ（ヒストグラム、正規分位点プロット）の活用、歪度・尖度の計算
　正規分布に該当しなければ、ノンパラメトリック分析へ

2) 平均を求める。

3) 分散を求める。

4) 標準偏差を求める。

5) 標準誤差を求める。　９５％信頼水準の計算

6) 分散の比較
　バラツキの比較　（F 検定）
　F 検定とは、ある母集団が他の母集団より大きなバラツキをもつかどうかを検定する方法。
　分散 $_A$ / 分散 $_B$ の比(但し分散 $_A$ ＞ 分散 $_B$)を　自由度(N_A-1, N_B-1)における F 値と比較する。

7) 平均値の比較
　u 検定と t 検定

　7-A:　u 検定：
　正規分布する母集団の平均と標準偏差が既にわかっている場合、サンプリングした試料の平均値が母集団の平均値とどの程度かけはなれているかを検定する。
　既に管理図の項で説明した方法と同様に、下記の統計変数を計算する。
　母集団の平均　M_0
　試料平均　　M_s
　母集団の標準偏差　σ_0
　試料のサンプリング数　n
　標準誤差　（＝試料平均の標準偏差）＝　$\sigma_0 / \sqrt{n}$
　$M_s - M_0$ が　標準誤差の何倍離れているかによって検定する。
　すなわち、標準正規分布に従い、1.96 倍以上離れる確率は、両側検定で５％以下。3.07 倍以上離れる確率は 0.2%以下となる。

　7-B:　t 検定
　２つのサンプリングされた集団が、同じ母集団に由来するものかどうかを検定する。

　7-B-1) データが対で得られる場合の２群 A と B の比較

　　対になっている値の差（$X_A - X_B$）をもとめる。
　　（$X_A - X_B$）について、分散、標準偏差、標準誤差を求める。
　　平均値の差（$M_A - M_B$）を標準誤差 $\sigma_{A-B} / \sqrt{n}$ で割り　t_{A-B} とする
　　自由度 n-1 で、各確率水準に対応した t 値を調べ、t_{A-B} と比較する。

7-B-2)　データが対でない場合の２つのサンプルの比較

下表のような統計値を計算する

	サンプル群　A	サンプル群　B
サンプル数	N_A	N_B
自由度	N_A － 1	N_B－1
平均	M_A	M_B
偏差平方和	$\Sigma\ (Xi - M_A)^2$	$\Sigma\ (Xi - M_B)^2$
分散	$\Sigma\ (Xi - M_A)^2 / (N_A - 1)$	$\Sigma\ (Xi - M_B)^2 / (N_B - 1)$
A 群と B 群の分散に有意差がない場合		
A 群＋B 群の分散＝分散 $_{AB}$	$[(N_A\text{-}1)\Sigma\ (Xi - M_A)^2 + (N_B\text{-}1)\Sigma\ (Xi - M_B)^2] / (N_A+N_B-2)$	
A 群＋B 群の標準偏差＝ σ_{AB}	$\sigma_{AB} = \sqrt{}$（分散 $_{AB}$）	
A 群＋B 群の標準誤差＝SE_{AB}	$SE_{AB} = \sigma_{AB} \sqrt{}\ (1/N_A+1/N_B)$	
T 値	$t = (M_A - M_B) / SE_{AB}$　自由度＝$N_A + N_B - 2$	

平均値の差の検定に先立って、これらの２群の分散が異なっていないという証明が必要。
すなわち F 検定を先に行う。
分散 $_A$ / 分散 $_B$ の比（但し分散 $_A$ ＞ 分散 $_B$）を　自由度(N_A-1, N_B-1)における F 値と比較する。

以下の t 値を求める。
$t = (M_A - M_B) / SE_{AB}$　自由度＝$N_A + N_B - 2$

自由度 $N_A + N_B - 2$ における各確率水準に対応した t 値を調べ、上記の t 値と比較する。

2. 対応のあるデータと対応のないデータにおける平均値の差の検定の例

1) データ

異なる６本のひもを半分にわけ､片方は A 色に他方は B 色に染色したのちにそれぞれの部分の強度を測定した。それぞれの強度のデータは下表のとおりであった。

ひも	1	2	3	4	5	6	平均	標準偏差
A	14.6	12.1	13.4	14.0	11.5	14.4	13.33	1.27
B	13.8	12.5	11.6	12.0	10.8	13.6	12.38	1.16
A－B	0.8	－0.4	1.8	2.0	0.7	0.8	0.95	0.866

このデータを対応のあるデータとして扱う場合と、対応のないデータとして扱う場合を比較する。

2）対応のあるデータとしての有意差検定

B群平均	12.3833	t値	-2.68522
A群平均	13.3333	自由度	5

平均の差	-0.95	N	6
差の標準偏差	0.866		
差の標準誤差	0.354	p値 (Prob>|t|)	0.0435 危険率<5%
上側95%信頼限界	-0.0406	p値(Prob>t)	0.9782
下側95%信頼限界	-1.8594	p値(Prob<t)	0.0218
95%信頼範囲(df=5) ±0.354×2.57	±0.9098		

従って、A 群と B 群の強度は有意に異なる。

3）対応のないデータとしての有意差検定

3-A） t検定

A群の分散	1.27			
B群の分散	1.16			
A群とB群の分散比	1.09	自由度(5,5)のF検定 F(5,5,5%)=5.05>1.09	有意差なし	A、B群の分散が等しいと仮定
差の推定値 M_A-M_B	0.950			
A群＋B群の平方和	14.84			
A群＋B群の分散	1.484			
A群＋B群の標準偏差σ_{A+B}	1.218	T=(M_A-M_B)／σ_{A+B}	1.351	
A群＋B群の標準誤差	0.7033	自由度	10	
下側95%	-0.617	p値(Prob>|t|)	0.2066 21%の危険率	A、B群の強度に有意差なし
上側95%	2.517	自由度10における両側5%のT値	2.23	2.23＞1.351

結論： A, B 群のひもの強度に有意差なし

3-B）F検定

要因	自由度	平方和	平均平方	F値	p値(Prob>F)
色による	1	2.707	2.707	1.8243	0.2066
誤差	10	14.84	1.484		
全体(修正済み)	11	17.55			

結論： A, B 群のひもの強度に有意差なし

4）結論

A群とB群に対応があればA群とB群の差は有意であったが、A群とB群の間に対応関係がない場合は、A群とB群の差は有意でなかった。

3．2 変量の統計手順

1) 散布図を調べる
2) 相関係数を求める
相関係数＝（ｘとｙの共分散）／｛(ｘの標準偏差）×（ｙの標準偏差)｝
ただし
（ｘとｙの共分散）＝Σ｛(ｘ$_i$－ｘの平均値）（ｙ$_i$－ｙの平均値)｝／（ｎ－１）
相関係数だけからでは、相関の有意性はわからない。
ｎ－2 を自由度とした相関係数の検定のための数表があるので使用する。
また、相関係数の二乗を決定係数という。

3) 単回帰式を求める

3-1）ｘが説明変数で、ｙが目的変数の場合 （ただしｘは誤差要因を含まない）
回帰式 ｙ＝ｂｘ＋ａ を求めることである。
この際、回帰直線の周りのｙのバラツキ（分散）が最も小さくなるようにａとｂの値を計算する。
a, b の値はそれぞれ
ｂ＝ （ｘとｙの共分散）／ （ｘの分散）
ａ＝ ｙの平均値 － ｂ ×（ｘの平均値）
で求められる。
実際には、最近はコンピュータソフトで計算するため、手計算で求めることは少なくなった。

3-2）ｙが説明変数でｘが目的変数の場合の回帰式、（ただしｙは誤差要因を含まない） ｙ＝ｂ’ｘ＋ａ’
ｂ’＝ （ｘとｙの共分散）／ (ｙの分散)
ａ’＝ ｙの平均値 － ｂ’ ×（ｘの平均値） となる。
また、 ｂ×ｂ’＝ （相関係数）^{2} ＝ｒ^2 となる。

3-3）ｘ、ｙともに誤差要因を含む場合の回帰式 ｙ＝ｂｘ＋ａ は、
各プロットから回帰直線に垂直におろした垂線の長さの二乗の和が最小になるように求められる。
この場合、
ｂ＝ ｛(ｙの分散）／ （ｘの分散)｝$^{0.5}$
＝（ｙの標準偏差）／ （ｘの標準偏差）
ａ＝ ｙの平均値 － ｂ ×（ｘの平均値）
で求められる。

3-3) の場合の回帰式は、エクセルやエクセル統計などのソフトでは自動的に計算

できないので、手計算する必要がある。
統計の専門ソフトでは可能である。（例えば JMP5.1 SAS Institute Japan）

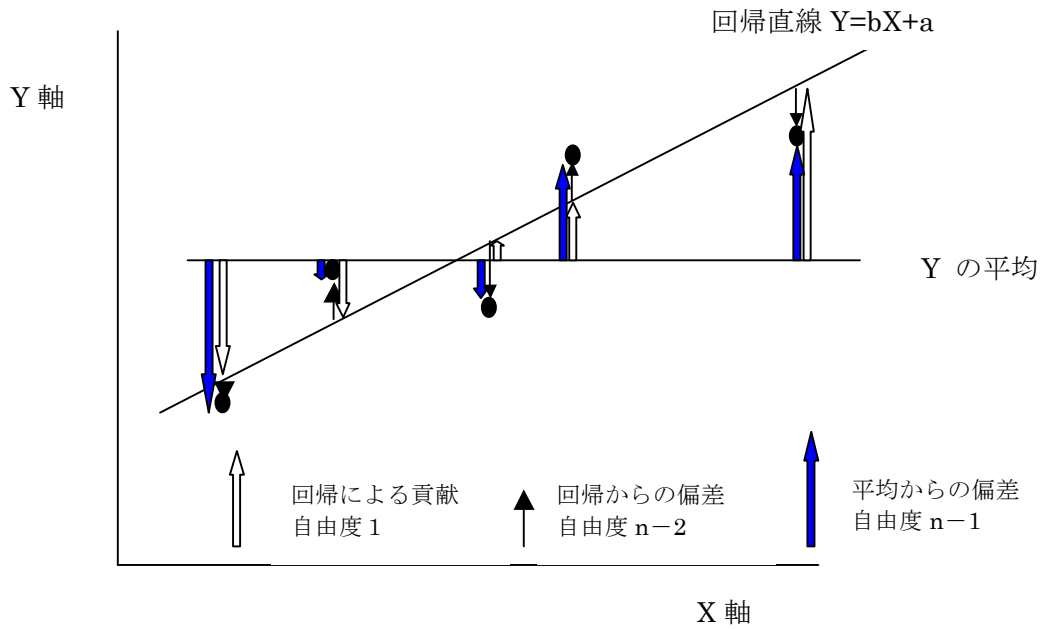
相関係数および決定係数は 3-1), 3-2) , 3-3）の全ての場合に同一である。

4）分散分析表で F 検定
F 検定による回帰式の有意性の検定は以下の変数を計算して行う。

	平方和	自由度	分散（平均平方）	F 値
回帰によるバラツキの減少	[Σ(x 実測値 –Mx) (y 実測値 –My)] ²／Σ(x－Mx)²	1	A=C－B	A/B
回帰からの残差	Σ(y 実測値 –y 予測値)²	n-2	B= [Σ(y 実側値 –y 予測値)²]/(n-2)	
全体のバラツキ	Σ(y 実測値 –y 平均値)²	n-1	C= [Σ(y 実側値 –y 平均値)²]/(n-1)	B,C の関係 (1-B/C)=r²

この計算もパソコンソフトで行えるので手計算の必要はない。

3-2）と 3-3）に対応する F 検定も結果は同じになる。



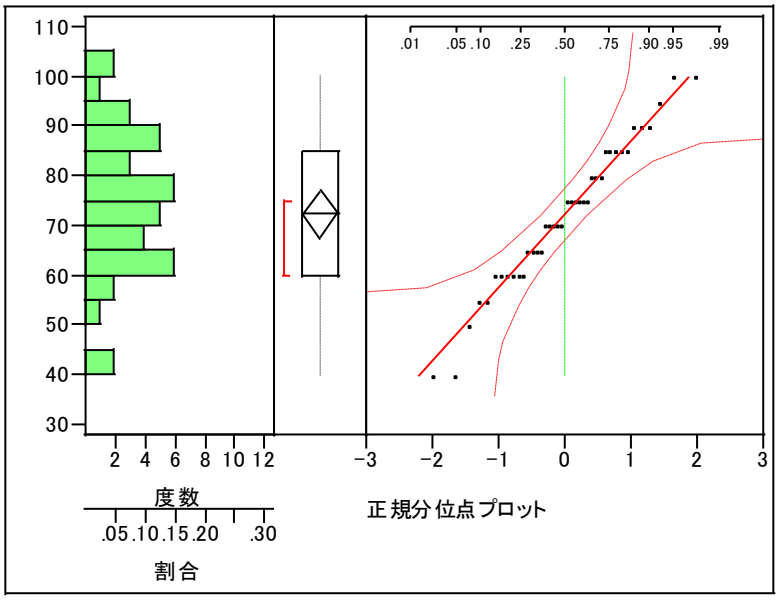
4．２変量の統計分析例　学生の成績分析

1）データ

学籍番号	性別	成績	勉強時間	支出
1	0	55	2	6
2	1	70	7	3
3	0	60	1	6
4	1	90	10	2
5	0	85	6	5
6	1	80	2	4
7	0	75	5	4
8	0	60	3	2
9	0	40	3	10
10	1	85	3	3
11	0	90	7	3
12	0	90	7	3
13	1	65	4	6
14	1	65	5	5
15	1	60	5	2
16	1	95	7	3
17	0	55	3	7
18	0	60	2	5
19	0	75	9	5
20	1	100	9	2
21	1	70	6	3
22	0	100	12	4
23	0	70	3	3
24	0	75	5	2
25	1	85	6	3
26	1	70	4	4
27	0	80	6	3
28	0	60	3	6
29	1	50	3	7
30	1	70	4	5
31	0	80	10	4
32	1	75	7	4
33	1	65	3	5
34	0	75	3	5
35	0	60	1	8
36	1	85	8	3
37	0	85	5	4
38	0	40	2	5
39	1	75	5	3
40	0	65	3	3

2）統計分析　（統計ソフトJMP5による）

成績の分布（正規分布への適合性の検討）



分位点

100.0%	最大値	100.00
75.0%	4分位点	85.00
50.0%	中央値(メディアン)	72.50
25.0%	4分位点	60.00
0.0%	最小値	40.00

モーメント

平均	72.25
標準偏差	14.63
平均の標準誤差	2.31
平均の上側95%信頼限界	76.93
平均の下側95%信頼限界	67.57
N	40
分散	214.0
歪度	−0.153
尖度	−0.225

歪度：　$\Sigma[(x_i-m)/s]^3 \times n/[(n-1)(n-2)]$

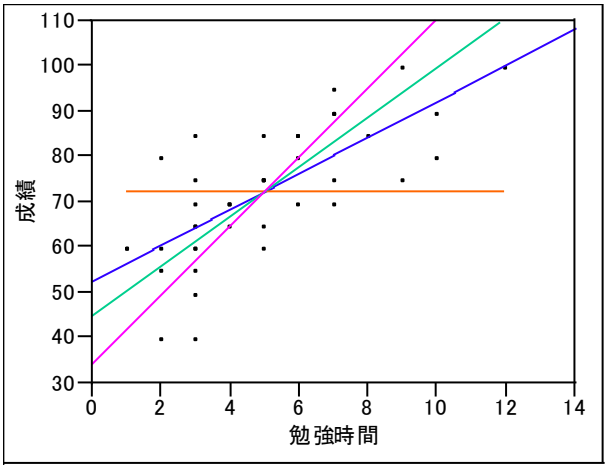
尖度：　$\Sigma[(x_i-m)/s]^4 \times n(n+1)/[(n-1)(n-2)(n-3)] - 3\times(n-1)^2/[(n-2)(n-3)]$

ただし、　mは平均、 s は標準偏差

分布が左右対称であれば歪度は０になる。

また、正規分布の尖度は０になる。

勉強時間と成績の二変量の関係



回帰直線(x→ｙ)のあてはめ

成績 = 52.4 + 3.99 勉強時間

あてはめの要約

R2乗	0.524
自由度調整R2乗	0.512
誤差の標準偏差(RMSE) σ_y	10.22
Yの平均	72.25
オブザベーション(または重みの合計)	40

分散分析

要因	自由度	平方和	平均平方	F値
モデル	1	4378.4	4378.4	41.9
誤差	38	3969.1	104.5	p値(Prob>F)
全体(修正済み)	39	8347.5		<.0001

直交回帰

変数	平均	標準偏差	分散比	相関
勉強時間	4.975	2.66	30.3	0.7242
成績	72.3	14.6		

切片	傾き	下側信頼限界	上側信頼限界	有意水準(α)
44.8	5.51	3.99	7.61	0.050

成績 = 44.8 + 5.51勉強時間

Y→X回帰

変数	平均	標準偏差	分散比	相関
勉強時間	4.98	2.66	0	0.7242
成績	72.3	14.6		

切片	傾き
34.4	7.61

成績 = 34.4 + 7.61勉強時間

5．実験計画法

1）実験計画法の対象となるのは、一つ以上の実験条件を変えたときの効果である。
与えられた一つの実験条件の組を因子という。
ある実験において用いる特定の条件（因子の値）を水準という。
実験の数値的な結果を応答という。
因子の効果とは、因子の水準の変化による応答の変化をいう。

2）ブロック化の原則

実験単位を全体のバラツキよりも小さいバラツキをとると思われる単位に分けること。
このような単位をブロックと呼ぶ。
因子水準の比較はブロック内で行なう。
ブロック内での応答の変動は、ブロック間での応答の変動とは独立である。
ブロック内ではランダムに実験を行なう。

3）乱塊法　（ランダムブロック法）

実験をいくつかのブロックに分け、各ブロック内で全ての実験条件で実験を行なう。各実験は各ブロック内でランダムに割り付ける。実験結果の解析は、要因をブロックおよび実験条件とした２元配置の分散分析となる。ブロックと実験条件の間に交互作用がないと仮定すれば、残渣平方和を誤差分散の推定に用いることができる。実験を繰り返せば、誤差の推定値を独立に求められる。

4）ラテン方格

乱塊法において、さらに実験条件を２通り以上に分割する場合、２通りの分割が可能な場合はラテン方格による計画を用いる。この計画は方格の形をとる。一組の条件（因子）を方格の行に割り付け、他の組の条件を列に割り付ける。各行および各列に各処理がただ一度だけ起こるようにする。

5）乱塊法およびラテン方格による実験計画の例

土壌の種類	肥料の種類	施肥量	繰り返し数	応答
A	有機（α）	無施肥（N）	３（1）	作物の収量
B	無機（β）	少　（L）	（2）	
C		中　（M）	（3）	
		多　（H）		

実験区は　３×２×４×３＝５４　必要

6）土壌 A 地点における実験

α-N-1	β-N-1	α-L-1	β-L-1	α-M-1	β-M-1
β-H-1	α-H-1	β-N-2	α-N-2	β-L-2	α-L-2
α-M-2	β-M-2	α-H-2	β-H-2	α-N-3	β-N-3
β-L-3	α-L-3	β-M-3	α-M-3	β-H-3	α-H-3

他の土壌に属する B・C 地点内でもほぼ同様に配置するが、N,L,M,H の順番を変える。

7）３元配置の分散分析　（エクセル統計 2000 による）

土壌、肥料の種類、施肥量の違いによる作物収量の応答

	土壌A		土壌B		土壌C	
	有機	無機	有機	無機	有機	無機
無施用	30	30	35	35	40	40
	35	35	37	37	39	39
	28	28	39	39	38	38
少量	35	36	40	41	43	44
	37	38	42	44	42	42
	38	39	43	40	44	43
中量	42	43	48	49	50	51
	40	46	45	44	48	52
	38	44	46	45	52	54
多量	45	50	52	50	54	55
	46	48	54	48	52	50
	42	45	51	47	55	54

平均値の差の検定:最小有意差法　　**:1%有意 *:5%有意

因　子	水 準 1	水 準 2	平均値1	平均値2	差	P 値	判定
因子A	土壌A	土壌B	39.1	43.8	−4.7	0.000	**
		土壌C	39.1	46.6	−7.5	0.000	**
	土壌B	土壌C	43.8	46.6	−2.8	0.000	**
因子B	有機	無機	42.9	43.4	−0.5	0.321	
因子C	無施用	少量	35.7	40.6	−4.9	0.000	**
		中量	35.7	46.5	−10.8	0.000	**
		多量	35.7	49.9	−14.2	0.000	**
	少量	中量	40.6	46.5	−5.9	0.000	**
		多量	40.6	49.9	−9.3	0.000	**
	中量	多量	46.5	49.9	−3.4	0.000	**

分散分析表　　**:1%有意 *:5%有意

要　因	偏差平方和	自 由 度	平均平方	F 値	P 値	判定
因子A	696.6	2	348.3	77.83	0.000	**
因子B	4.5	1	4.5	1.01	0.321	
因子C	2143.4	3	714.5	159.66	0.000	**
A×C	30.6	6	5.1	1.14	0.353	
B×C	16.9	3	5.6	1.26	0.297	
A×B×C	26.1	6	4.4	0.97	0.453	
誤差	223.8	50	4.5			
全体	3142.0	71				