統計の手法レポート： 解説と解答例　　　　担当： 渕野 昌

以下は 5 月 13 日の講義の初めに提出してもらった第一回目レポートの解説 + 解答例です．

http://math.cs.kitami-it.ac.jp/fuchino/chubu/statistics-04s-report01.pdf

に置いてあるものと同じものです．また，講義のホームページを，

http://math.cs.kitami-it.ac.jp/fuchino/chubu/statistics-04s.html

に作りました．レポートの解説や講義の補足説明なども色々リンクしたいと思っています．

1. 次の表は，50 人のクラスで行なったあるテスト（100 点満点）の得点である．

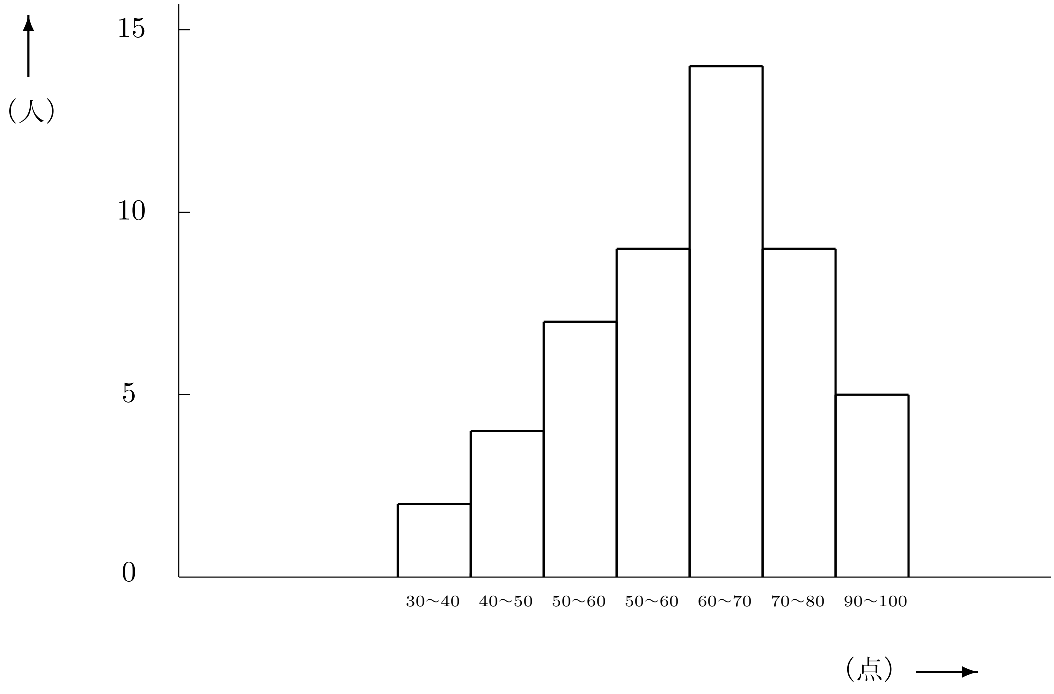
57 71 62 65 62 88 95 66 100 71 73 79 45 32 78 75 86 67 50 56 100 73 79 35 75
50 56 88 75 90 63 82 63 82 63 72 87 49 48 69 45 90 73 82 78 72 83 55 80 54

このデータを級間隔 10 の度数分布表にまとめて，そのヒストグラムを作成してください．この度数分布表での平均値，モード，分散と標準偏差を求めてください．

このデータの度数分布表は，階級を，30 点以上 40 点未満，40 点以上 50 点未満，...，90 点以上 100 以下として次のようになる．

階級	階級値	度数
30〜40	35	2
40〜50	45	4
50〜60	55	7
60〜70	65	9
70〜80	75	14
80〜90	85	9
90〜100	95	5
合計		50

これに対応するヒストグラムは：



度数分布表のデータの平均値は,

$$\bar{x} = \left(\sum_i \text{階級値}_i \times \text{度数}_i \right) \ / \ \text{データの総数}$$

として計算できる（教科書の (1.2)）．ただし，“$\sum_i \cdots$” は，各階級で式 $\cdots$ の値を計算してその和をとる計算を表している．上の度数分布表での値をこれに代入して，平均値

$$(35 \times 2 + 45 \times 4 + 55 \times 7 + 65 \times 9 + 75 \times 14 + 85 \times 9 + 95 \times 5)/50 = 70.2$$

を得る．

上の度数分布表のデータのモードは，80〜90 の階級の度数が最大であることから，その階級値 85 である．

最後に，分散と標準偏差を求める．このために，まず平方和 S を求める．これは，教科書の (1.6) 式を度数分布表のデータ用に書き直した (1.7) の式:

$$S = \sum_i (\text{階級値}_i - \text{平均値})^2 \times \text{度数}_i$$

で計算されるから,

$$\begin{aligned}
&(35 - 70.2)^2 \times 2 + (45 - 70.2)^2 \times 4 + (55 - 70.2)^2 \times 7 \\
&\quad + (65 - 70.2)^2 \times 9 + (75 - 70.2)^2 \times 14 + (85 - 70.2)^2 \times 9 + (95 - 70.2)^2 \times 5 \\
&= 12248
\end{aligned}$$

である．したがって，N を度数の和（つまりデータの総数：ここでは 50）とすると分散と標準偏差はそれぞれ：

$$\frac{S}{N-1} = \frac{12248}{49} = 644.631579\cdots$$

$$\sqrt{\frac{S}{N-1}} = \sqrt{\frac{12248}{49}} = 25.389596\cdots$$

となる．

平方和の計算は (1.16) またはこれを度数分布表用に書き直した (1.17) により計算することもできる．(1.16) や (1.17) は作表ソフトを用いて統計処理をするときに使いやすいものになっている．(1.17) は，平方和 S が

$$S = \sum_i (\text{階級値}_i)^2 \times \text{度数}_i \quad - \quad \frac{(\sum_i \text{階級値}_i \times \text{度数})^2}{\sum_i \text{度数}_i}$$

で計算できることを主張している．したがって,「階級値 × 度数」「(階級値)^2× 度数」の項目を上で作成した度数分布表に付け加えて,

階級	階級値	度数	階級値 × 度数	(階級値)^2×度数
30〜40	35	2	70	2450
40〜50	45	4	180	8100
50〜60	55	7	385	21175
60〜70	65	9	585	38025
70〜80	75	14	1050	78750
80〜90	85	9	765	65025
90〜100	95	5	475	45125
計		50	3510	258650

(1.17) に，ここで得られた値を代入して，

$$S = 258650 - 3510^2/50 = 12248$$

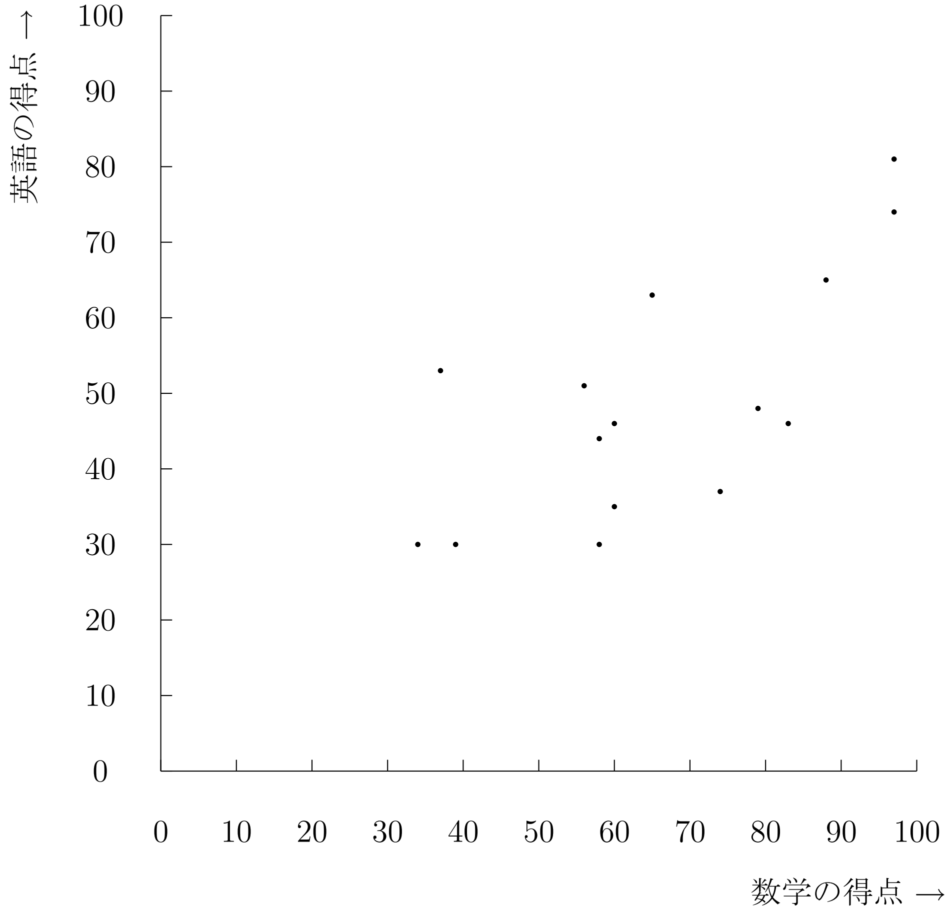
として平方和が計算できる！この値を用いて上と同様に分散と標準偏差を計算すればよい．

2. 次の表は 15 人のクラスにおける数学と英語のテストの成績（どちらも 100 点満点）の記録である．

学生番号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
数学	88	83	79	34	65	58	58	39	37	97	60	56	74	97	60
英語	65	46	48	30	63	30	44	30	53	74	35	51	37	81	46

a) このデータの散布図を作成してください．b) このデータでの数学の成績と英語の成績の相関係数を求めてください．この値から数学の成績と英語の成績の間の相関関係についてどのようなことが言えますか？ c) このデータに関する回帰直線の式を算出して，a) で求めた散布図に書き加えてください．

a):


英語の得点 →
100
90
80
70
60
50
40
30
20
10
0
0
10
20
30
40
50
60
70
80
90
100
数学の得点 →

b): 上の散布図は正の相関があるように見える．相関係数を計算してみることにする．変量の組のデータが (x_i, y_i) $i = 1, 2, \ldots, n$ で与えられているとき，これらの変量の間の相関係数は，(1.25) 式により，

$$r = \frac{\displaystyle\sum_{i=1}^{n} x_i \cdot y_i \; - \; \frac{1}{n} \cdot \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{\displaystyle\sum_{i=1}^{n} {x_i}^2 \; - \; \frac{1}{n} \cdot \left(\sum_{i=1}^{n} x_i\right)^2}\;\sqrt{\displaystyle\sum_{i=1}^{n} {y_i}^2 \; - \; \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2}}$$

により求まる．この計算をするために．まず x_i, y_i, ${x_i}^2$, ${y_i}^2$, $x_i y_i$ の項目を持つ表を作る．

学生番号	x_i(= 数学の点)	y_i(= 英語の点)	${x_i}^2$	${y_i}^2$	$x_i y_i$
1	88	65	7744	4225	5720
2	83	46	6889	2116	3818
3	79	48	6241	2304	3792
4	34	30	1156	900	1020
5	65	63	4225	3969	4095
6	58	30	3364	900	1740
7	58	44	3364	1936	2552
8	39	30	1521	900	1170
9	37	53	1369	2809	1961
10	97	74	9409	5476	7178
11	60	35	3600	1225	2100
12	56	51	3136	2601	2856
13	74	37	5476	1369	2738
14	97	81	9409	6561	7857
15	60	46	3600	2116	2760
計	985	733	70503	39407	51357

上の表で求めた値を (1.25) に代入すると，分散は，

$$r = \frac{51357 - \frac{1}{15} \times 985 \times 733}{\sqrt{70503 - \frac{1}{15} \times 985^2} \;\cdot\; \sqrt{39407 - \frac{1}{15} \times 733^2}} = 0.7053\cdots$$

となり，1 にある程度近い値をとるので，正の相関があると仮定してもよいと結論できそうである．

(c): (b) で求めた値を (1.32) 式と (1.33) 式に代入して回帰直線をあらわす式を求める．まず，x_i, $1 \leq i \leq 15$ と y_i, $1 \leq i \leq 15$ の平均 $\bar{x}$, $\bar{y}$ は，それぞれ

$$\bar{x} = 985/15; \quad \bar{y} = 733/15$$

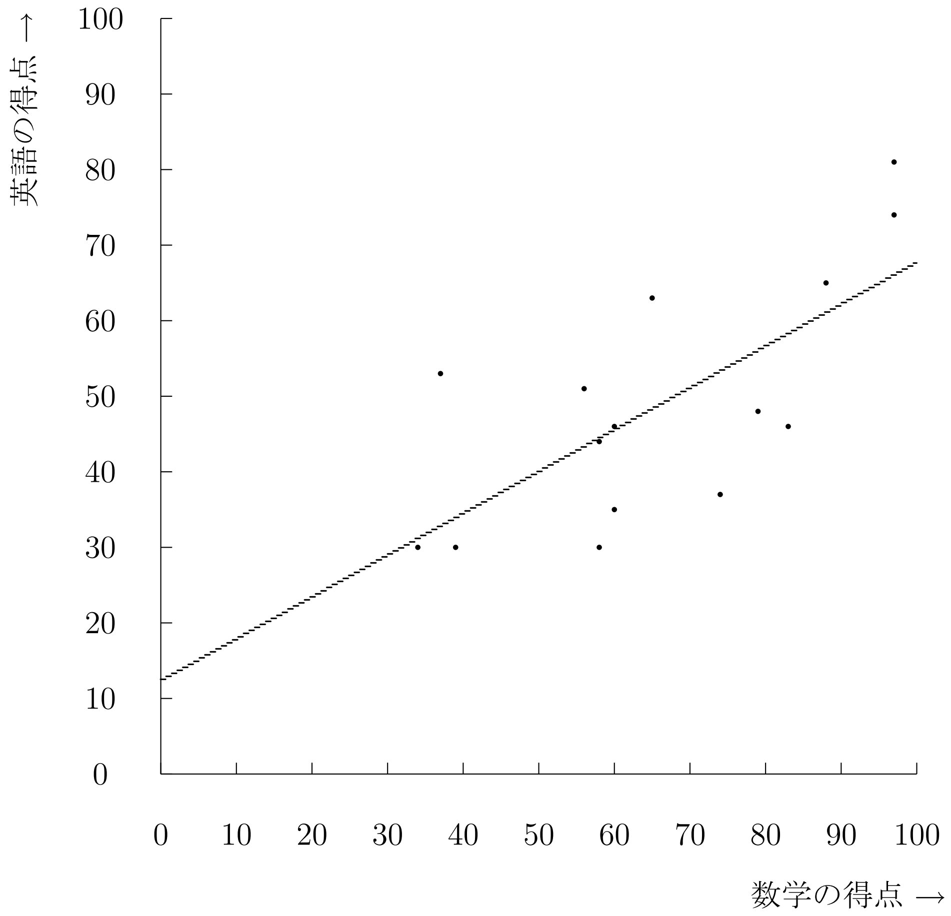
となることに注意する．y の x への回帰直線の傾きは，(1.32) 式により，

$$b \;=\; \frac{n\displaystyle\sum_{i=1}^{n} x_i y_i \; - \; \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\displaystyle\sum_{i=1}^{n} x^2 \; - \; \left(\sum_{i=1}^{n} x_i\right)^2} \;=\; \frac{15 \times 51357 - 985 \times 733}{15 \times 70503 - 985^2} = 4835/8732$$

となる．したがって，この回帰直線の方程式は，

$$y - \frac{733}{15} = \frac{4835}{8732} \left(x - \frac{985}{15} \right)$$

となる．この式の x に 0 と 100 を代入すると，それぞれ $y = 546027/43660 \approx 12.51$, $2963527/43660 \approx 67.88$ を得るから，(a) で求めた散布図に $(0, 12.51)$ と $(100, 67.88)$ をむすぶ直線を書き込むと，



(1.32) の代わりに (1.33) を用いて全く同様の作業を行なうと y の x への回帰直線を得ることができる.（これについてはここでは省略する）.