2 変量データの分布

樋口さぶろお

龍谷大学理工学部数理情報学科

使える統計! L05(2013-10-30 Wed)

今日の目標

1. 個々のデータの偏差値が求められる
2. 2 変量データから共分散, (ピアソンの積率) 相関係数が計算できる
3. 相関係数, 散布図から 2 変量の関係を説明できる

http://hig3.net

黒板でやった Quiz の解答は省略します.

L04-S2

Quiz 解答:変動係数

- 1 番目のデータで, $b_1 = 800$ と思う. $X_1 = 5, -20, 5, \ldots$.
- X_1 の平均値は $= \frac{1}{5}[5 + (-20) + 5 + (-5) + (-5)] = -4$. よって, $X_1 + b_1$ の平均値は $800 - 4 = 496$.
- X_1 の分散 $= \frac{1}{5}[(5 - (-4))^2 + ((-20) - (-4))^2 + (5 - (-4))^2 + ((-5) - (-4))^2 + ((-5) - (-4))^2] = 81$. よって, $X_1 + b_1$ の分散は 81.
- $X_1 + b_1$ の標準偏差 $= \sqrt{81} = 9$.
- $X_1 + b_1$ の変動係数 $= 9/796 = 0.011$.
- 2 番目のデータで, $b_2 = 90$ と思う. $X_2 = -3, +3, \ldots$.
- X_2 の平均値 $= \frac{1}{5}[(-3) + 3 + (-1) + 1 + 0] = 0$. $X_2 + b_2$ の平均値は $90 + 0 = 90$.

- X_2 の分散
  $= \frac{1}{5}[((-3)-0)^2+(3-0)^2+((-1)-0)^2+(1-0)^2+(0-0)^2] = 4.$
  $X_2 + b_2$ の分散は 4.
- $X_2 + b_2$ の標準偏差 $= \sqrt{4} = 2.$
- $X_2 + b_2$ の変動係数 $= 2/90 = 0.022.$
- よって 2 番目のデータ $X_2 + b_2$ のほうがばらつきが大きい.

ここまで来たよ

1. 復習:分散の応用
   - 偏差値
2. 2 変量データの分布
   - 2 変量データとは
   - 2 変量データの相関

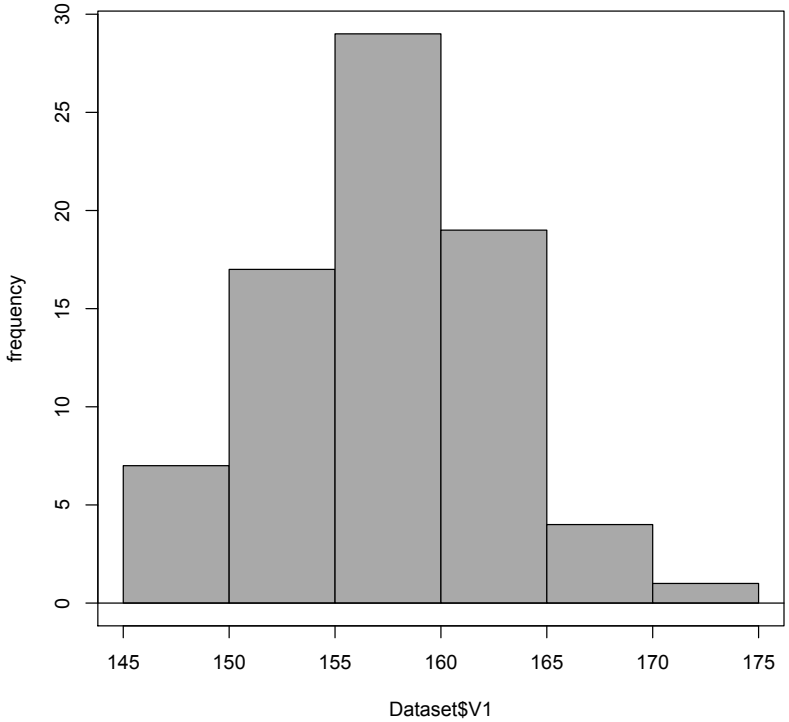
偏差値

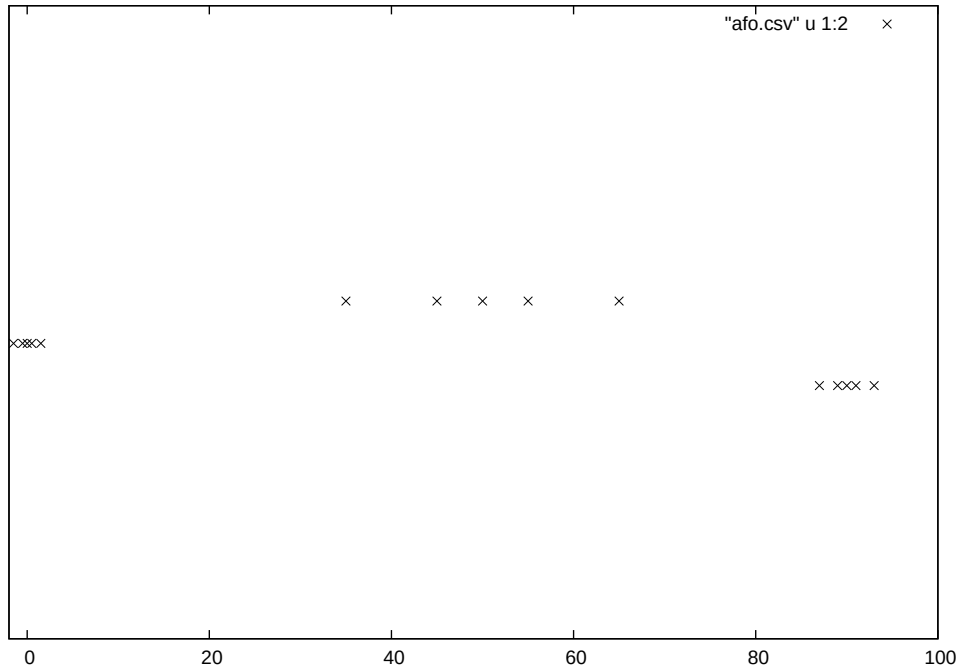
$0 - -100$ の範囲の値をとるデータ (テストの点数や成績?) に使われる.
受験者 1 人 1 人の成績が, 平均値から上, または下に離れている程度を見られる.

偏差値

$$\begin{aligned}(\text{データ 1 個の}) \text{偏差値} &= \text{標準得点} \times 10 + 50 \\ &= \frac{\text{データの値} - \text{平均値}}{\text{標準偏差}} \times 10 + 50\end{aligned}$$

- 異なるテスト, クラスでも比べられる.
- 偏差値の平均値は $\square$
- 偏差値の標準偏差は $\square$
- 偏差値はまあ ‘無次元の数’(1000 点満点と 100 点満点を比較可能)





データ						平均値	標準偏差
X	87	93	89	91	90	90	2
X の標準得点	-1.5	$+1.5$	-0.5	$+0.5$	0	0	1
X の偏差値	35	65	45	55	50	50	10

Q1

Quiz(偏差値)

(学力) 偏差値について, 次のうち正しいのはどれ (とどれ)?

1. 偏差値の最低値は 0 である
2. 偏差値の最高値は 75 である
3. 平均点 (をとった人) の偏差値は 50 である
4. 100 点のテストで満点を取った場合の偏差値は, 他の人の成績しだいである
5. 偏差値 50 の人の順位は上から 1/2 程度である
6. 偏差値 60 の人の順位は上から 15% 程度である.

Q2

Quiz(標準得点と偏差値)

データ 85, 97, 89, 93, 91 で, 85 の標準得点と偏差値を求めよう.

ここまで来たよ

1. 復習:分散の応用
   - 偏差値
2. 2 変量データの分布
   - 2 変量データとは
   - 2 変量データの相関

2 変量データ

これまでやってたのはぜんぶ 1 変量データ.
2 変量データはこんな例. (X, Y) などと書く. X, Y は各チームのデータ.

- X フリーキック回数
- Y 被シュート回数
- Z 失点

データの個数 $N = 18$.

(チーム名)	X	Y
コンサドーレ札幌	389	464
ベガルタ仙台	491	246
⋮	⋮	⋮
計	...	...
平均値	...	...

J League Division 1. 2012-10-06. http://www.j-league.or.jp/data/

クロス集計表と周辺分布

X=フリーキック回数 Y = 被シュート回数

クロス集計表

上の表では…になってる 18 チーム全部のデータから作りました.

$Y \backslash X$	400 未満	450 未満	500 未満	550 未満	計
200 以上 250 未満		1	2	1	4
300 未満			4	1	5
350 未満	2	2	1	1	6
400 未満		2			2
450 未満					0
500 未満	1				1
計	3	5	7	3	18

周辺分布

Q3

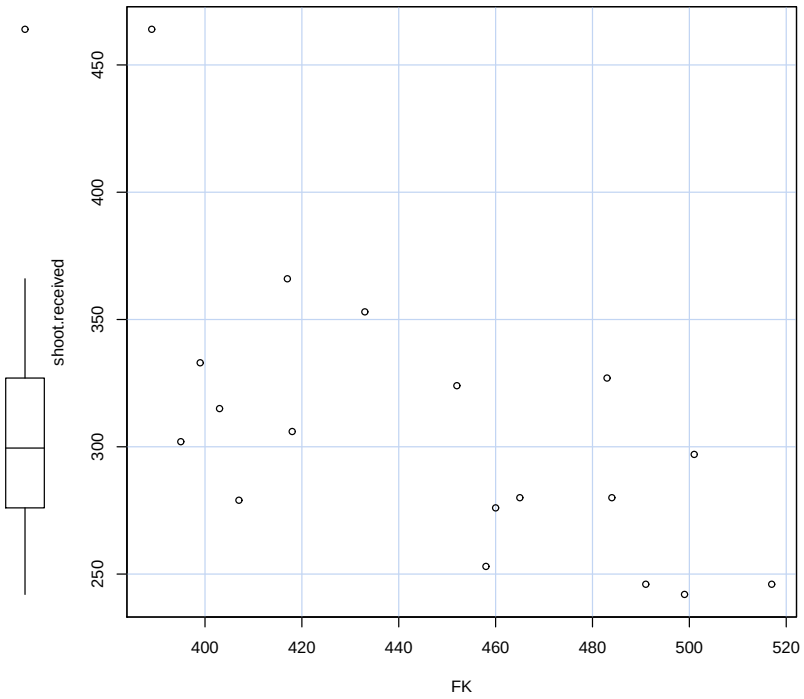
Quiz(クロス集計表)

1. クロス集計表を作ろう. X の階級の幅は 2, Y の階級の幅は 5 で.
2. 散布図を描こう.

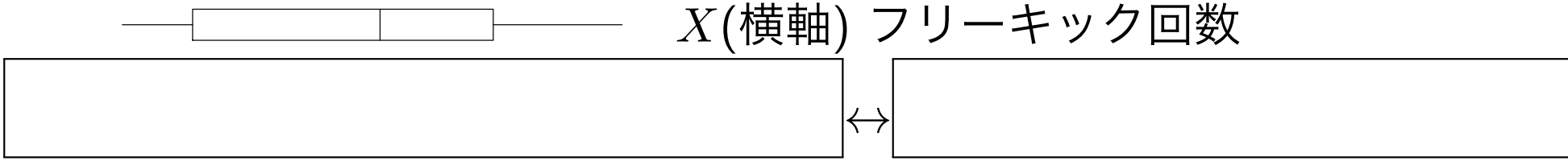
X	Y
1	5
3	11
4	14
5	15
7	20

散布図

Y(横軸) 被シュート回数

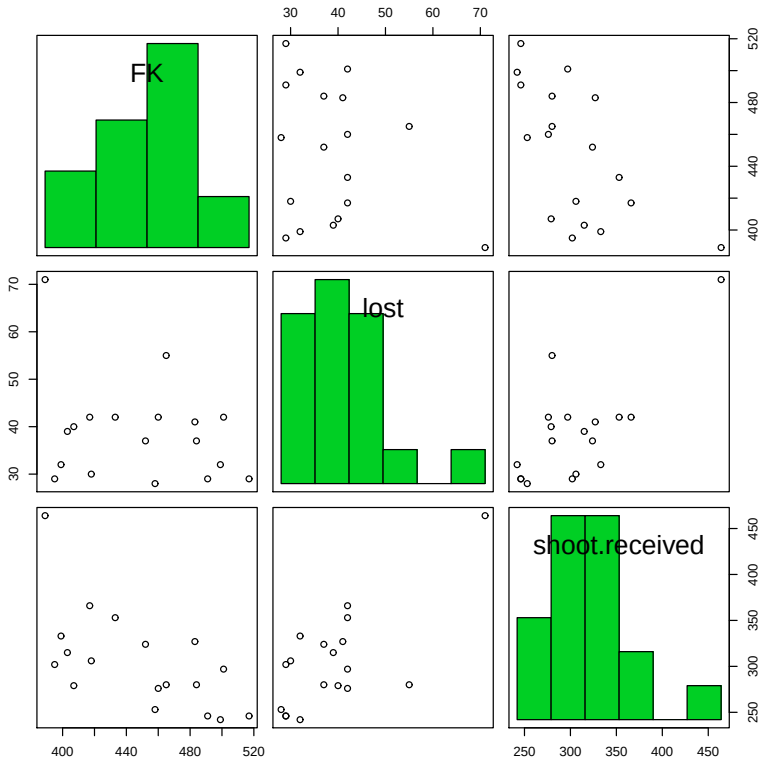


X(横軸) フリーキック回数



散布図と周辺分布

上 (左) から, X:フリーキック回数, Z:失点, Y:被シュート回数

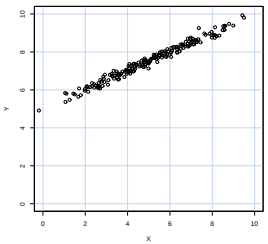


対角線上にあるのは, 周辺分布のヒストグラム

ここまで来たよ

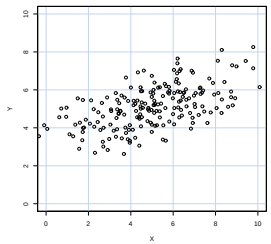
1. 復習:分散の応用
   - 偏差値
2. 2 変量データの分布
   - 2 変量データとは
   - 2 変量データの相関

正の相関・負の相関・無相関



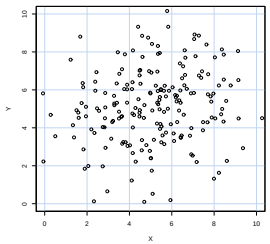
強い正の相関

$r = 0.99$



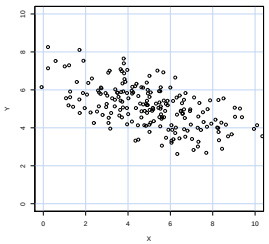
弱い正の相関

$r = 0.55$



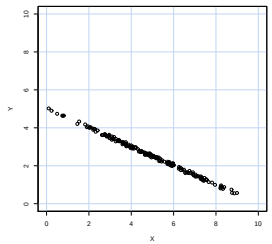
無相関

$r = 0$



弱い負の相関

$r = -0.55$



強い負の相関

$r = -0.99$

‘正’: X が大きい $\Leftrightarrow$ Y が大きい

‘負’: X が大きい $\Leftrightarrow$ Y が小さい

r:

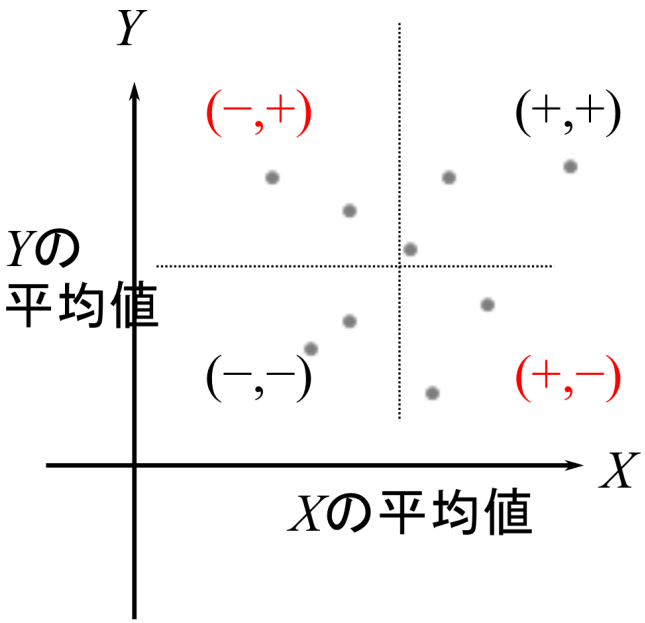
共分散

$$\text{分散} = \frac{1}{\text{データの個数}}[(X\ \text{のデータ}\ 1 - \text{平均値}) \times (X\ \text{のデータ}\ 1 - \text{平均値}) +$$

共分散 (covariance)

$$\begin{aligned}
&X, Y\ \text{の共分散}\ C \\
&= \frac{1}{\text{データの個数}\ N} \\
&\times[(X\ \text{のデータ}\ 1 - X\ \text{の平均値}) \times (Y\ \text{のデータ}\ 1 - Y\ \text{の平均値}) \\
&+(X\ \text{のデータ}\ 2 - X\ \text{の平均値}) \times (Y\ \text{のデータ}\ 2 - Y\ \text{の平均値}) \\
&+\cdots(\text{データすべて})\cdots \\
&+(X\ \text{のデータ}\ N - X\ \text{の平均値}) \times (Y\ \text{のデータ}\ N - Y\ \text{の平均値})]
\end{aligned}$$

共分散の意味



共分散が正に/負に大きい ⇔ 正の/負の相関が強い (?)

しかし

相関係数

共分散は

- 次元のある量なので
  比較に不便
  →
- 広い範囲にばらついていたほうが

相関係数は, 相関の強さを直接的に表す.

相関係数 (correlation coefficient)

$$X, Y \text{ の相関係数 } r = \frac{X, Y \text{ の共分散 } C}{X \text{ の標準偏差 } s_X \times Y \text{ の標準偏差 } s_Y}$$

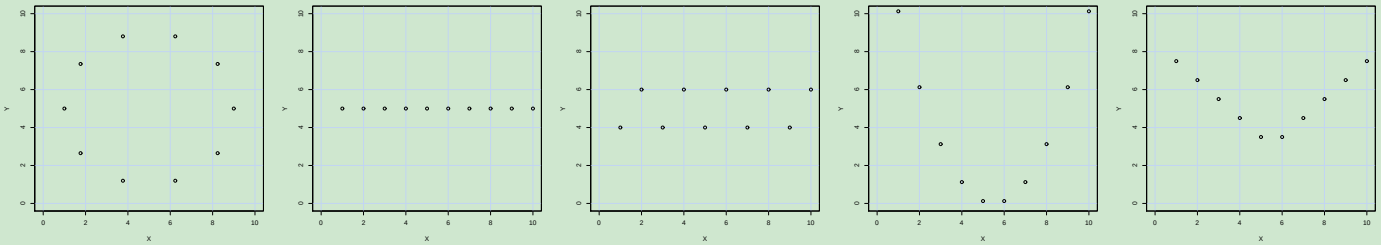
相関係数の性質

- 
- $-1 \leq r \leq +1$
- $r = \pm 1$

  $\Leftrightarrow$ 
- $\Leftrightarrow Y$ が X の 1 次関数で書ける
- $r = 0 \Leftrightarrow$ ' 無相関'

Quiz(相関係数)

次のうち, 相関係数 r がもっとも大きいものはどれ?



相関係数=0 にだまされるな

相関係数=0 $\Leftrightarrow$ X と Y の間に ‘関係’ がない?

- 相関係数 $r = 0 \Leftrightarrow$

- 相関係数 $r = 0$ だから X, Y は無関係な量, というわけではない

Q4

Quiz(共分散)

1. X, Y の共分散を求めよう
2. X, Y の相関係数を求めよう．ただし，Y の標準偏差 $= 11/\sqrt{5} = 4.92$ は使っちゃっていい．

X	Y
1	5
3	11
4	14
5	15
7	20

Q5

Quiz(相関係数)

次のうち, X, Y の相関係数について本当はどれ?

1. X を一斉に -2 倍すると, r は -2 倍になる.
2. X を一斉に -2 倍すると, r は 2 倍になる.
3. X を一斉に -2 倍すると, r は -1 倍になる.
4. X を一斉に -2 倍すると, r は $+1$ 倍になる (かわらない).
5. X を一斉に -2 倍すると, r は $-1/2$ 倍になる.
6. X を一斉に -2 倍すると, r は $1/2$ 倍になる.

にせの因果関係にだまされるな

- 原因:被シュートが多い, その結果: 失点が多い?
- 原因:失点が多い, 結果: 被シュートが多い?
- 原因:フリーキックが多い, 結果:被シュートが少ない?
- 原因:被シュートが少ない, 結果:フリーキックが多い?
- 原因:???, 結果:被シュートが少ない, かつ, フリーキックが多い?

- 相関が強くても
- 因果関係があっても

連絡

- 2013-11-06 水 臨時教室変更 3-B105 計算機実習室 (この建物の地下, 前回と同じ)
- 2013-11-13 水 教室, 授業形態など変更あるかも
- 2013-11-20 水 プチテスト 公式外部記憶ペーパーのみ持込可 出題計画など詳細は来週以降に
- いつか 台風の分の補講
- 学期半ば授業アンケートにご協力ありがとうございました. 随時追加の意見・感想を送れます.
- 加減乗除と平方根 (ルート) の使える電卓持ってきてね. 関数電卓でなくてもいいです. 携帯電話の機能・アプリでもかまいません.

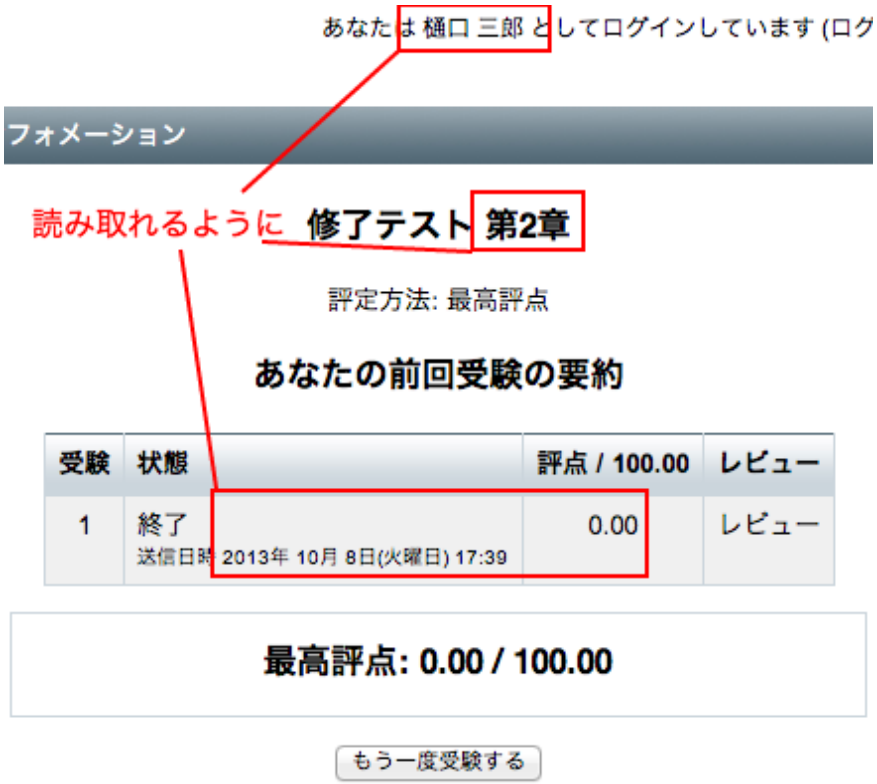
新たなる課題

- 各追加 2 ピーナッツ=計 4 ピーナッツになる 新たな 課題.

  提出: 2013-11-06 水 の授業 or 2013-11-20 水 のテスト前

  1. 龍谷大学 e ラーニングシステム
     `https://moodle.media.ryukoku.ac.jp/` → リメディアルコース統計学 → 第 3 章修了テスト
  2. 龍谷大学 e ラーニングシステム
     `https://moodle.media.ryukoku.ac.jp/` → リメディアルコース統計学 → 第 5 章修了テスト

  このサイトには, `http://hig3.net` → 龍大 Moodle, や Info Seta → e ラーニングサイト → 新 e ラーニングシステム でも到達できます.
  すべてを送信して終了する → レビューを終了する の後に出る, 「あなたの前回受験の要約」ページ (下) を印刷して, 紙で提出. (スクリーンショットを課題にアップロードしてもいい)
- 今週は授業内で紙を 1 枚提出 (+修了テストも提出できます)



クリッカー学籍番号送信の方法

- t012345 → 1012345 OK
- c012345 → 4012345 OK
- w012345 → 7012345 OK