2 変量データの共分散・相関係数・回帰分析

樋口さぶろお

龍谷大学理工学部数理情報学科

確率統計☆演習 I L04(2016-10-13 Thu)

最終更新: Time-stamp: "2016-10-13 Thu 07:22 JST hig"

今日の目標

- 高校 数学 I 塚田確率統計 1.7 塚田確率統計 1.8 (塚田確率統計 1.7.1 は省略)
- 2 変数の量的データから, 共分散と相関係数と回帰直線が求められる
- Excel の統計ツールが使える

http://hig3.net

L03-Q1

Quiz 解答:範囲

範囲は $Q_4 - Q_0 = 25 - 14 = 11$, 四分位範囲は $Q_3 - Q_1 = 18 - 14.5 = 3.5$, 四分位偏差は $\frac{1}{2}(Q_3 - Q_1) = 1.75$.

L03-Q2 Quiz 解答:平均値・分散・標準偏差 平均値 $= 90$kg, 分散 $= 4\text{kg}^2$, 標準偏差 $= 2$kg.

L03-Q3 Quiz 解答:度数分布表から分散 平均値 $= 62$(cm), 分散 $= ((50-62)^2 \times 10 + (60-62)^2 \times 20 + (70-62)^2 \times 20)/50 = 112(\text{cm}^2)$.

L03-Q5

Quiz 解答:平均値・分散・標準偏差の換算

1.6m, 0.0025m^2, 0.05m.

ここまで来たよ

1. データのばらつきを表す値

2. 2 変量データの共分散・相関係数・回帰分析
   - 2 変量データとクロス集計表・散布図
   - 2 変量データの相関
   - Excel で統計

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 勝利数
- y (打った) シュート数
- z 失点

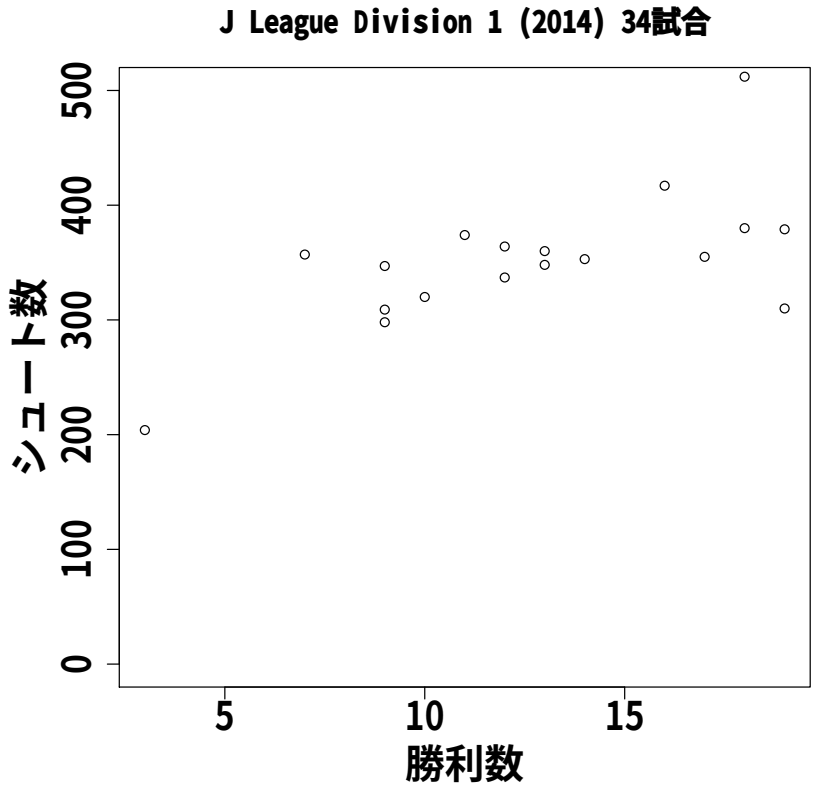
J リーグ Div1. 2014 年の 34 試合. データの個数 $n = 18$(チーム).

(チーム名)	x	y	z
ベガルタ仙台	9	347	50
鹿島アントラーズ	18	512	39
⋮	⋮	⋮	⋮
計	...	...	...
平均値	...	...	...

http://www.j-league.or.jp/data/

他にも…(x, y) =(身長 (cm), 体重 (kg)), (人口 (人), 面積 (m^2), (打率, 本塁打数), (カロリー, 糖分含有量)....

散布図=相関図 塚田確率統計 1.7.2



クロス集計表=相関表 塚田確率統計 1.7.2 と周辺分布

x:勝利数, y (打った) シュート数

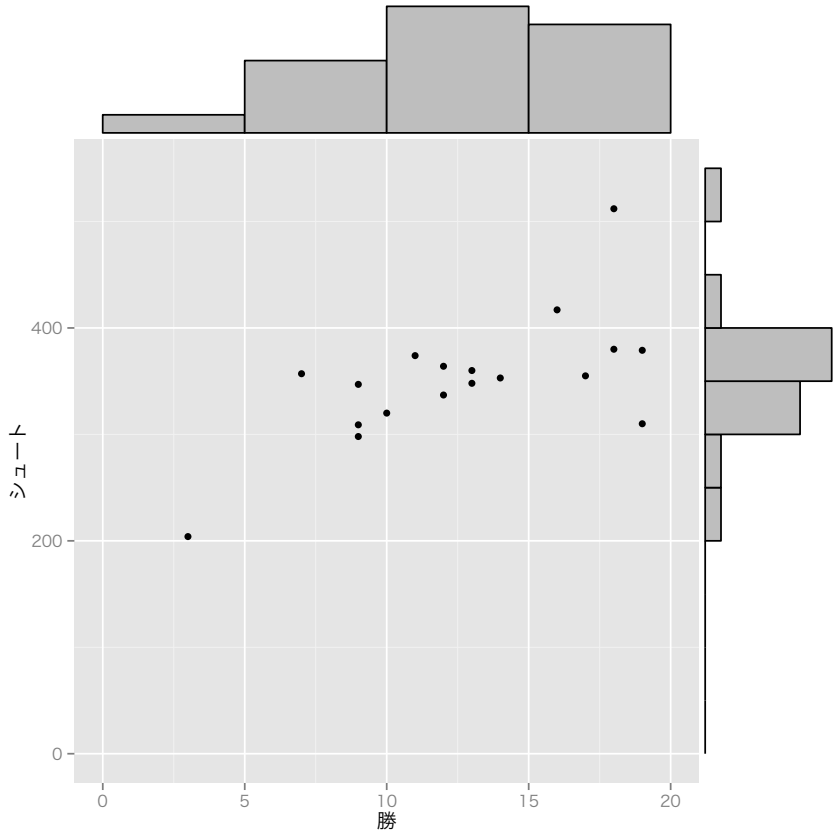
クロス集計表 度数分布表の 2 変数版

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.

$\downarrow y$　$\backslash x$ の階級 $\rightarrow$	0 以上 5 未満	10 未満	15 未満	20 未満	計
200 以上 250 未満	1				1
250 以上 300 未満		1			1
300 以上 350 未満		2	3	1	6
350 以上 400 未満		1	4	3	8
400 以上 450 未満				1	1
450 以上 500 未満				0	0
500 以上 550 未満				1	1
計	1	4	7	6	18

周辺分布とは

周辺分布のヒストグラム



周辺分布のヒストグラムは, 散布図で

[　　] して作れる.

L04-Q1

Quiz(クロス集計表)

1. 散布図を描こう.
2. クロス集計表を作ろう. x の階級は 0 以上 2 未満, …, y の階級は 0 以上 5 未満, … で.

x	y
1	5
3	15
4	14
5	11
7	20

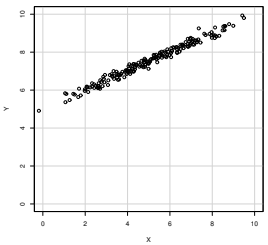
ここまで来たよ

1. データのばらつきを表す値

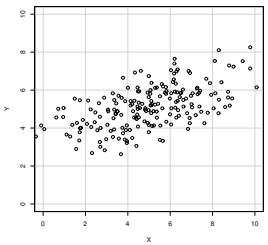
2. 2 変量データの共分散・相関係数・回帰分析
   - 2 変量データとクロス集計表・散布図
   - 2 変量データの相関
   - Excel で統計

正の相関・負の相関・無相関

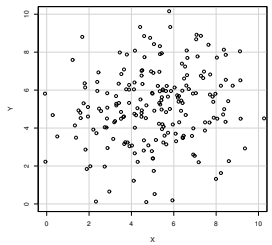
塚田確率統計 1.7.2 | 塚田確率統計 p.40



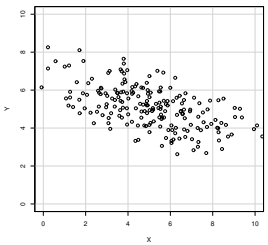
強い正の相関
$r = 0.99$



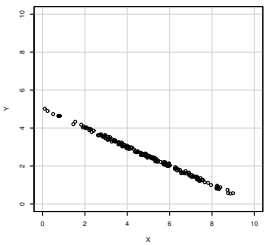
弱い正の相関
$r = 0.55$



無相関
$r = 0$



弱い負の相関
$r = -0.55$



強い負の相関
$r = -0.99$

相関

‘正の相関’: x が大きい $\Leftrightarrow$ y が大きい

‘負の相関’: x が大きい $\Leftrightarrow$ y が小さい

強い/弱い: 傾向がはっきりしている/していない

r: 相関係数 計算方法は以下.

共分散

塚田確率統計 p.44 高校 数学 I 発展

相関の強さを数で表したい

$$x \text{ の平均値} \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$x \text{ の分散 } s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(x_i - \overline{x})$$

$\overline{y}, s_y^2$ も同様.

共分散 (covariance)

塚田確率統計 p.43 下から 2 行目

$$x, y \text{ の共分散 } s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}) \times (y_i - \overline{y})$$

注: x の分散を $s_x^2 = s_{xx}$, y の分散を $s_y^2 = s_{yy}$ と書く自然な記法がある.

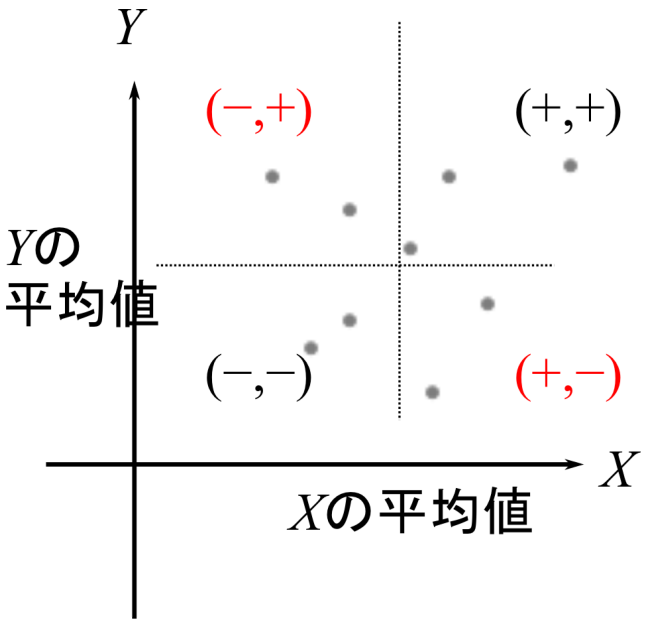
L04-Q2

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 $= \sqrt{\frac{122}{5}} = 4.94$ は使っちゃっていい.

x	y
1	5
3	15
4	14
5	11
7	20

共分散の意味 塚田確率統計 1.8



$(+,-) = (x_i - \bar{x}$の符号$, y_i - \bar{y}$の符号$)$.

共分散が正に/負に大きい $\Leftrightarrow$ 正の/負の相関が強い (?)

なぜなら

しか〜し.

相関係数 塚田確率統計式 (1.9) 高校 数学 I

共分散は

- 次元のある量なので単位を変えると［　　］→ 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，これらの影響を受けずに，相関の強さをそのまま表す．

相関係数 (correlation coefficient) 塚田確率統計式 (1.9)

$$x, y \text{ の相関係数 } r = \frac{s_{xy}}{s_x \times s_y}$$

相関係数の性質

- 相関係数は
- $-1 \leq r \leq +1$
- $r = 0 \Leftrightarrow$ ' 無相関'
- $r = \pm 1 \Leftrightarrow$ 散布図の点が傾き正/負の一直線上 $\Leftrightarrow$ y は x の 1 次関数.

散布図の点が傾き正/負の一直線上 $\Rightarrow$ $r = \pm 1$ であることの証明

$y_i = ax_i + b$ とすると.

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}) \cdot ((ax_i + b) - (a\overline{x} + b)) = as_x^2$$

ところで, $s_y = |a|s_x$ なので,

$$r = \frac{as_x^2}{s_x|a|s_x} = \pm 1$$

.

相関係数 $= 0$ にだまされるな 塚田確率統計 p.41

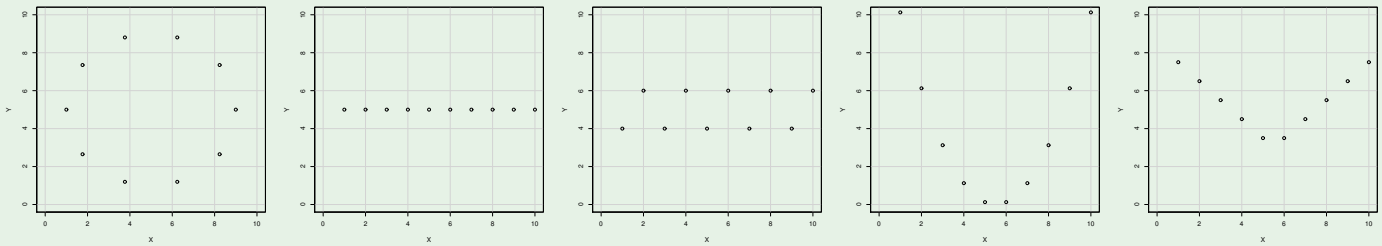
相関係数 $r = 0 \Leftrightarrow x$ と y の間に ‘関係’ がない?

- 相関係数 $r = 0 \Leftrightarrow x$ が増えた
  ら［　　］言えない
- 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

L04-Q3

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?



Anscombe(1973)

にせの因果関係にだまされるな

因果関係=原因と結果
勝利数とシュート回数は正の相関

- 原因:シュートが多い, 結果: 勝利が多い?
- 原因:勝利が多い, 結果:シュートが多い?

(打った) フリーキック回数と被シュート本数は負の相関

- 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相関係数からは原因と結果を区別できない

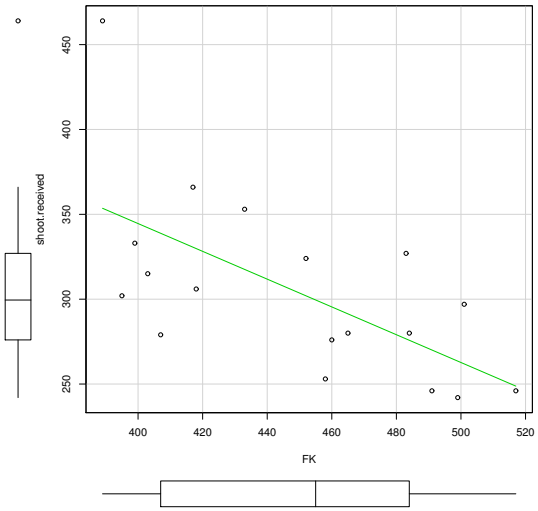
回帰分析 塚田確率統計 1.8

回帰 (regression), 直線回帰=単回帰分析=1 変数回帰分析 物理実験

2 変量データ (x, y) が
相関係数 $r = \pm 1$ に近い $\Leftrightarrow$ 散布図上のデータ点 (x, y) がほぼ直線に載っている

その直線 ( ) の式 $y = ax + b$ を知りたい!

つまり a, 定数項 b を決めたい. (塚田確率統計 p.44 と逆の定義)



y: 目的変数 (従属変数)
x: 説明変数 (独立変数)
何でそんなことしたいの?

- 法則を見つけたい
- x から y を予測したい

回帰直線の決め方

1. 定規をあてて ‘真ん中’ を通るように
2. 最小 2 乗法で.

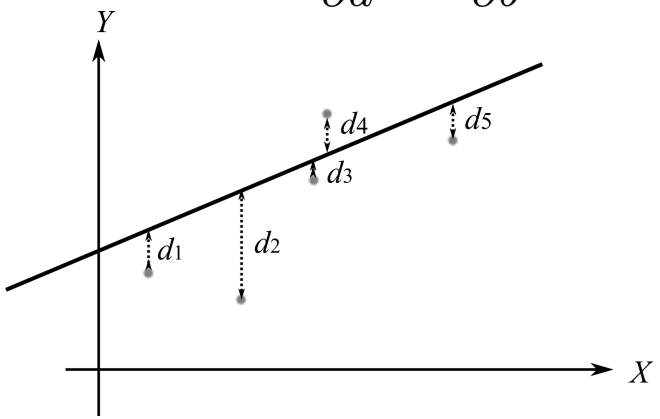
最小 2 乗法

直線からのずれの 2 乗 d^2 の合計

$$f(a,b) = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (y_i - (ax_i + b))^2$$

の最小条件 $\frac{\partial f}{\partial a} = \frac{\partial f}{\partial b} = 0$ で a, b を決める.

微積分 I



物理実験

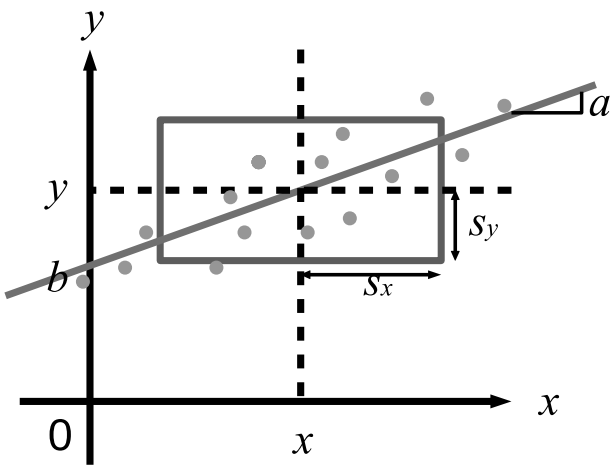
直線回帰の公式

回帰直線 塚田確率統計 p.44

x_i, y_i $(i = 1, \ldots, n)$ の平均値を $\overline{x}, \overline{y}$, 標準偏差を s_x, s_y, 相関係数を r とする. このとき回帰直線は,

$$y = \frac{r \times s_y}{s_x} \times (x - \overline{x}) + \overline{y} = ax + b.$$

傾きは $a = \frac{r \times s_y}{s_x} = \frac{s_{xy}}{s_x^2}$, 切片は $b =$ (点 $(\overline{x}, \overline{y})$ を通るような値)



a: 回帰係数 (x を 1 だけ変えたときの y の変化量)
r^2: 決定係数 (あてはまりのよさ)

回帰直線の傾きのおぼえ方 I

広がり方

散布図上のデータ点の分布は, 横 $2s_x$, 縦 $2s_y \to$ 傾き $\frac{s_y}{s_x}$ くらい?
しか〜し, 傾きには正負があるし, 相関がなかったら傾きを 0 にしたいので, 相関係数 r をかけ算しておく.

単位チェック

(x, y) の単位が (m,kg) だとする.
r は無次元. 単位無し.
左辺 y (kg).
右辺 $r \times \frac{s_y(\text{kg})}{s_x(\text{m})} \times x(\text{m}) + b(\text{kg})$
で, s_x/s_y かけると単位があう.

L04-Q4

Quiz(共分散)

y を応答変数, x を説明変数として, 回帰直線の式を求めよう.

x	y
1	5
3	15
4	14
5	11
7	20

L04-Q5

Quiz(共分散と相関係数)

下のデータを考える.

x	y
1	3
2	7
4	10
5	9
8	16

1. 共分散を求めよう.
2. 相関係数を求めよう.
3. 回帰直線の式を求めよう.

ただし, 平均値 $\overline{x} = 4, \overline{y} = 9$, 分散 $s_x^2 = 6, s_y^2 = 18$ であることを使っていい.

ここまで来たよ

1. データのばらつきを表す値

2. 2 変量データの共分散・相関係数・回帰分析
   - 2 変量データとクロス集計表・散布図
   - 2 変量データの相関
   - Excel で統計

準備

統計ソフトウェア実習室にインストールされているのは

R 無料. オープンソース. 解説書が多い.
SPSS 伝統ある高級品.
Excel 機能は限られ怪しいところもあるが, 普及率高い. 龍大では Office365 で無料.

今日は Excel を使ってみます.
スタートボタン >Excel 2013
統計分析のための準備
ファイル > オプション > アドイン > Excel のアドイン > 設定 > 分析ツール にチェックを入れて OK する.

表計算ソフトウェア (Excel) による主な分析 高校 数学 I

どこかの段階でデータ範囲を指定, または関数の引数にデータ範囲を指定.

	メニューベース	関数ベース
平均値, 分散, 標準偏差	データ > 分析 > データ分析 > 基本統計量 > 統計情報	平均値 average, 分散 var.p, 標準偏差 stdev.p, 最頻値 mode
四分位数	データ > 分析 > データ分析 > 順位と百分位数	中央値 median, 四分位数 quartile
度数分布表, ヒストグラム	データ > 分析 > データ分析 > ヒストグラム > 入力範囲とデータ区間	frequency + グラフ
散布図	挿入 > グラフ > 散布図	
共分散, 相関係数	データ > 分析 > データ分析 > 共分散, 相関	covar=covariance.p, correl
回帰分析	データ > 分析 > データ分析 > 回帰分析	linest
クロス集計表	挿入 > テーブル > ピボットテーブル	
	結果が $\frac{n}{n-1}$ 倍違うことあり	

メニューベースの分析をするときの注意

- Excel は, 1 種類のデータは列方向 (縦方向) にならんでいるとデフォルトでは想定する. 分析の種類によっては, 列方向, 行方向のどちらに並んでいるかを指定できるものもある.
- 2 変量 (n 変量) の統計量である, 共分散 s_{xy} や相関係数 r_{xy} の出力は

$$\begin{matrix} s_{xx} & s_{yx} \\ s_{xy} & s_{yy} \end{matrix}, \qquad \begin{matrix} r_{xx} & r_{yx} \\ r_{xy} & r_{yy} \end{matrix}$$

のように行列状にになっている. s_{yy} や r_{yy} は, $y = x$ であるときの s_{xy}, r. よく考えると, $s_{yy} = s_y^2, r_{yy} = 1$ であることに気づく. $n \geq 3$ のときは $n \times n$ 行列になる.
- 回帰分析の出力では
  - 重相関 R $=$ 相関係数 r
  - 従決定 R2 $=$ 決定係数 r^2
  - 切片の係数 $=$ 回帰直線の切片 b
  - X 値 1 の係数 $=$ 回帰係数 a
  - $n \geq 3$ の重回帰 $(x_1, x_2, \ldots, x_{n-1}, y)$ というものがあり, そのときは X 値 2,$\cdots$ などとなっていく.
- ここで紹介したメニューベースの分析では, 実はここまで学んだ「データの分散」すなわち var.p でなく, 今後学ぶ「不偏標本分散」 var.s を計算している$\cdots$ 両者の区別は考え方としては超重要だが, Excel で扱いたくなるくらいデータ数が多いときは, 近い値になる.

連絡

欠席届 毎回出席を前提に進めます．やむを得ず欠席して，ピーナッツ的に考慮されたい場合は，専用用紙に事情を説明する書類を貼って，授業前後各 5 分に提出 (事前事後とも可．ファイナルトライアルが締切)．欠席に事前連絡は原則不要．何回欠席してもファイナルトライアル参加資格を失うことはありません．

- 配布資料は 1-503 向かいの引出，`http://hig3.net` で再配布．
- 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．
- 樋口オフィスアワー木 6 金昼 (1-502)，Math ラウンジ月-木昼 (1-614)
- 次回は 塚田確率統計 2.1 | 塚田確率統計 2.2 | 塚田確率統計 3.1 | 塚田確率統計 3.2 .

`https://manaba.ryukoku.ac.jp`