分散分析・2 次元正規分布

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 II L10(2016-06-30 Thu)

最終更新: Time-stamp: "2016-06-30 Thu 13:55 JST hig"

今日の目標

- 分散分析表の F 検定ができる
- 2 次元正規分布の確率密度関数から母平均値と共分散行列が求められる. その逆.

http://hig3.net

L09-Q1

Quiz 解答:F 検定

1. 有意水準 $\alpha = 0.05$ で,
2. 母分散の比の両側 F 検定を行う
3. 帰無仮説 H_0 を, 「…ドーナツの重さの母分散は等しい: $\sigma_1^2/\sigma_2^2 = 1$ 」とする. すなわち, 対立仮説 H_1 を, $\sigma_1^2/\sigma_2^2 \neq 1$ とする.
4. 標本サイズを n_1, n_2, 不偏標本分散を S_1^2, S_2^2 とすると, 量 $F = \frac{S_1^2}{S_2^2}$ は, 帰無仮説のもとで自由度 $(n_1 - 1, n_2 - 1)$ の F 分布に従う. この量を検定統計量として用いる.
5. この標本に対して $F = \frac{28}{4} = 7$ である.
6. F 分布表より, $F_{\alpha/2}(10-1, 5-1) = 8,905 > 7 = F$. また, $F_{1-\alpha/2}(10-1, 5-1) < 7$. よって帰無仮説は棄却できない. 母分散が異なるとは結論できない.

L09-Q2
Quiz 解答:片側 F 検定

1. 有意水準 $\alpha = 0.05$ で,
2. 母分散の比の片側 F 検定を行う
3. 帰無仮説 H_0 を, 「…ドーナツの重さの母分散は等しい: $\sigma_1^2/\sigma_2^2 = 1$ 」とする. すなわち, 対立仮説 H_1 を, $\sigma_1^2/\sigma_2^2 > 1$ とする.
4. 標本サイズを n_1, n_2, 不偏標本分散を S_1^2, S_2^2 とすると, 量 $F = \frac{S_1^2}{S_2^2}$ は, 帰無仮説のもとで自由度 $(n_1 - 1, n_2 - 1)$ の F 分布に従う. この量を検定統計量として用いる.
5. この標本に対して $F = \frac{28}{4} = 7$ である.
6. F 分布表より, $F_\alpha(10 - 1, 5 - 1) = 5.999 < 7 = F$. よって帰無仮説は棄却される. 支店 1 の母分散が大きいと結論する.

ここまで来たよ

3. F 分布・正規分布の 2 標本の母分散の F 検定・分散分析
   - 分散分析

4. 2 次元正規分布
   - 2 変量の連続型確率変数
   - 2 次元正規分布

量的データがカテゴリ変数に依存するか

例

問「ドーナツの重さの母平均値は支店に依存しない」か?

i	支店	データ	個数	標本平均値	不偏標本分散
1	瀬田	79,80,80,81	4	80	$\frac{1}{4-1}[(79-80)^2+\cdots]$
2	石山	78,86,81,83	4	82	
3	草津	81,81,80,82	4	81	
計			12	81	

仮定 各支店のデータは, 正規分布 $\mathrm{N}(\mu_i, \sigma^2)$ にしたがう. (支店番号 $i = 1, 2, 3$).

図解すると? 箱ひげ図や, 信頼区間の図を描いて様子を把握しよう.

分散分析の用語と記号

問「級内平均値は「水準」(=「群」or「級」) に依存しない」か?

水準	データ	個数	級内平均	残差平方和
A_1	$y_{11}, y_{12}, \ldots, y_{1r}$	r	$\overline{y}_{1\bullet}$	$\sum_j (y_{1j} - \overline{y}_{1\bullet})^2$
A_2	$y_{21}, y_{22}, \ldots, y_{2r}$	r	$\overline{y}_{2\bullet}$	$\sum_j (y_{2j} - \overline{y}_{2\bullet})^2$
$\vdots$				
A_ℓ	$y_{\ell 1}, y_{\ell 2}, \ldots, y_{\ell r}$	r	$\overline{y}_{\ell\bullet}$	$\sum_j (y_{\ell j} - \overline{y}_{\ell\bullet})^2$
計		$r\ell$	$\overline{y}_{\bullet\bullet}$	

- $\bullet$ はその添字で平均したという意味.

級内平均値 $\overline{y}_{i\bullet} = \frac{1}{r}\sum_{j=1}^{r} y_{ij}$.

全平均値 $\overline{y}_{\bullet\bullet} = \frac{1}{r\ell}\sum_{i=1}^{\ell}\sum_{j=1}^{r} y_{ij}$.

$Y_{ij} \sim \mathrm{N}(\mu + a_i, \sigma^2)$, 独立. $\sum_i a_i = 0$.

別の書き方: $Y_{ij} = \mu + a_i + E_{ij}$, $\quad E_{ij} \sim \mathrm{N}(0, \sigma^2)$ 独立

問「$a_1 = a_2 = \cdots = a_\ell = 0$」 か?

L10-Q3

Example (分散分析表で使う記号の意味)

上の例で, 次は何に相当する?

r

ℓ

y_{12}

$\overline{y}_{1\bullet}$

$\overline{y}_{\bullet\bullet}$

$\sum_j (y_{1j} - \overline{y}_{1\bullet})^2$

分散分析を使うとき

量的変数 (ドーナツの重さ) の, カテゴリ変数 (支店) への依存性を考えるとき

$\leftrightarrow$ 2 水準の時は 2 標本 t 検定と同じ結果になる

$\leftrightarrow$ 回帰分析 (相関係数…), 2 元分割表の独立性の検定 n

分散 (=ばらつき) の比較に言い換え

横 (級) の中でのばらつきと, 縦 (級の間で) のばらつきは同じ」か?
$a_i \neq 0$ なら縦のばらつきが大きくなるはず.
縦のばらつきの合計 a_i の効果=級間平方和

$$S_{\mathrm{A}} = \sum_{i=1}^{\ell}\sum_{j=1}^{r}(\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet})^2 = r \times \sum_{i=1}^{\ell}(\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet})^2 \sim \chi^2(\ell - 1)$$

横のばらつきの合計 E_{ij} の効果=残差平方和

$$S_{\mathrm{E}} = \sum_{i=1}^{\ell}\sum_{j=1}^{r}(y_{ij} - \overline{y}_{i\bullet})^2 \sim \chi^2((r\ell - 1) - (\ell - 1))$$

すべてのばらつきの合計=全平方和

$$S_{\mathrm{T}} = \sum_{i=1}^{\ell}\sum_{j=1}^{r}(y_{ij} - \overline{y}_{\bullet\bullet})^2 \sim \chi^2(r\ell - 1)$$

実は $S_{\mathrm{A}} + S_{\mathrm{E}} = S_{\mathrm{T}}$. 自由度のカウント $(\ell - 1) + (r\ell - \ell) = r\ell - 1$.

分散分析 (ANOVA) or 分散分析の F 検定 の設計方針.
帰無仮説 $a_i = 0$ のもとで,
S_{A} は自由度 $\phi_{\mathrm{A}} = \ell - 1$ のカイ二乗分布 (*),
S_{E} は自由度 $\phi_{\mathrm{E}} = r\ell - \ell$ のカイ二乗分布にしたがう
よって, $F = \frac{V_A}{V_E} = \frac{S_{\mathrm{A}}/(\ell-1)}{S_{\mathrm{E}}/(r\ell-\ell)}$ は自由度 $(\ell - 1, r\ell - \ell)$ の F 分布にしたがう (**).
もし $a_i \neq 0$ なら, S_{A} は (*) よりも大きい値をとりがち. したがって比 F は (**) よりも大きい値をとりがち. F があまりに大きくて, 片側 F 検定の棄却域に入ったら, 帰無仮説を棄却して $a_i \neq 0$ と結論する.

1 元配置の分散分析表

変動要因	平方和	自由度	平均平方	F
級間	S_{A}	$\phi_A = \ell - 1$	$V_A = S_A/\phi_A$	V_A/V_E
残差	S_{E}	$\phi_E = (r\ell - 1) - (\ell - 1)$	$V_E = S_A/\phi_A$	
全	S_{T}	$\phi_T = r\ell - 1$		

Example (分散分析)

上の場合に対して分散分析表を作り, 有意水準 $\alpha = 0.05$ で F 検定しよう.

L10-Q4

Quiz(分散分析)

次のデータに対して, 1 元配置の分散分析表を作ろう. 有意水準 $\alpha = 0.05$ で F 検定しよう.

水準					
A_1	11	9	12	9	9
A_2	10	17	18	20	10
A_3	25	23	21	22	24

F 分布表

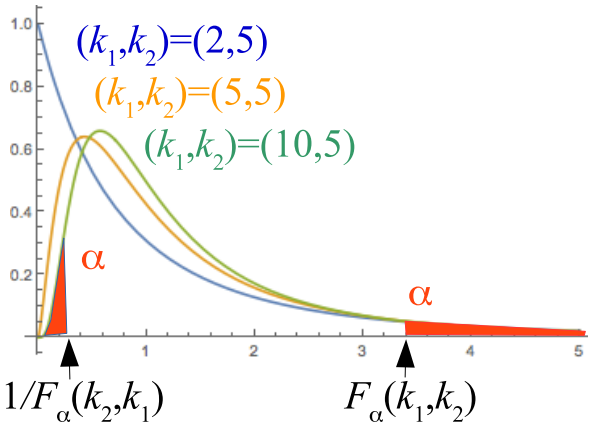
自由度 k_1, k_2 の F 分布にしたがう F に対して, $\alpha = P(F > F_\alpha(k_1, k_2))$ となる $F_\alpha(k_1, k_2)$ の値の表. $F = \frac{Y_{k_1}/k_1}{Y_{k_2}/k_2}$, $Y_k \sim \chi^2(k)$.

$\alpha = 0.05$

$k_2 \backslash k_1$	1	2	3	4	5	6	7	8	9	10	$+\infty$
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.50
3	10.13	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812	8.786	8.526
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	5.964	5.628
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772	4.735	4.365
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099	4.060	3.669
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677	3.637	3.230
8	5.318	4.459	4.066	3.838	3.687	3.581	3.500	3.438	3.388	3.347	2.928
9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230	3.179	3.137	2.707
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020	2.978	2.538
11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948	2.896	2.854	2.404
12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796	2.753	2.296
∞	3841	2.996	2.605	2.372	2.214	2.099	2.010	1.938	1.880	1.831	1.000

$\alpha = 0.025$

$k_2 \backslash k_1$	1	2	3	4	5	6	7	8	9	10	$+\infty$
1	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.7	963.3	968.6	1018
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.50
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	13.90
4	12.22	10.65	9.979	9.605	9.364	9.197	9.074	8.980	8.905	8.844	8.257
5	10.01	8.434	7.764	7.388	7.146	6.978	6.853	6.757	6.681	6.619	6.015
6	8.813	7.260	6.599	6.227	5.988	5.820	5.695	5.600	5.523	5.461	4.849
7	8.073	6.542	5.890	5.523	5.285	5.119	4.995	4.899	4.823	4.761	4.142
8	7.571	6.059	5.416	5.053	4.817	4.652	4.529	4.433	4.357	4.295	3.670
9	7.209	5.715	5.078	4.718	4.484	4.320	4.197	4.102	4.026	3.964	3.333
10	6.937	5.456	4.826	4.468	4.236	4.072	3.950	3.855	3.779	3.717	3.080
11	6.724	5.256	4.630	4.275	4.044	3.881	3.759	3.664	3.588	3.526	2.883
12	6.554	5.096	4.474	4.121	3.891	3.728	3.607	3.512	3.436	3.374	2.725
$+\infty$	5.024	3.689	3.116	2.786	2.567	2.408	2.288	2.192	2.114	2.048	1.000



ここまで来たよ

3. F 分布・正規分布の 2 標本の母分散の F 検定・分散分析
   - 分散分析

4. 2 次元正規分布
   - 2 変量の連続型確率変数
   - 2 次元正規分布

復習:2 変量の離散的確率変数の同時分布

同時分布

確率統計☆演習 I(2016)L01

$P(X = x, Y = y) = f_{\mathrm{XY}}(x, y)$

表で書いたほうが見やすい.

$y \backslash x$	158	160	165
45	3/8	0	1/12
50	1/8	1/3	1/12

$y \backslash x$	x_1	x_2	x_3
y_1	$f_{\mathrm{XY}}(x_1, y_1)$	$f_{\mathrm{XY}}(x_2, y_1)$	$f_{\mathrm{XY}}(x_3, y_1)$
y_2	$f_{\mathrm{XY}}(x_1, y_2)$	$f_{\mathrm{XY}}(x_2, y_2)$	$f_{\mathrm{XY}}(x_3, y_2)$

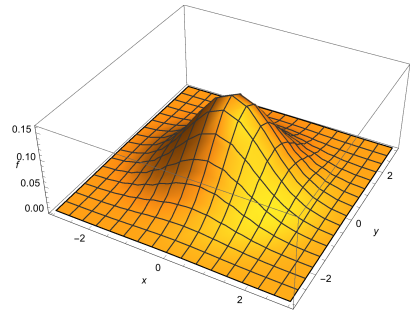
2 変量の離散型確率変数の母期待値

$$\mathrm{E}[\phi(X, Y)] = \sum_{i=1}^{a} \sum_{j=1}^{b} \phi(x_i, y_j) f_{\mathrm{XY}}(x_i, y_j)$$

2 変量の連続型確率変数の同時分布

確率密度関数 (2 変数関数)

$$f_{\mathrm{XY}}(x, y)$$



2 変量の連続型確率変数の母期待値

$$\mathrm{E}[\phi(X, Y)] = \int_{-\infty}^{+\infty} \mathrm{d}x \int_{-\infty}^{+\infty} \mathrm{d}y \; \phi(x, y) f_{\mathrm{XY}}(x, y)$$

2 変量の連続型確率変数の確率 (母比率)

$$\begin{aligned} P(a \leq X < b, c \leq Y < d) =& \mathrm{E}[\mathbf{1}_{[a \leq x < b, c \leq y < d]}(X, Y)] \\ =& \int_a^b \mathrm{d}x \int_c^d \mathrm{d}y \; f_{\mathrm{XY}}(x, y). \quad \text{体積} \end{aligned}$$

ここまで来たよ

3. F 分布・正規分布の 2 標本の母分散の F 検定・分散分析
   - 分散分析

4. 2 次元正規分布
   - 2 変量の連続型確率変数
   - 2 次元正規分布

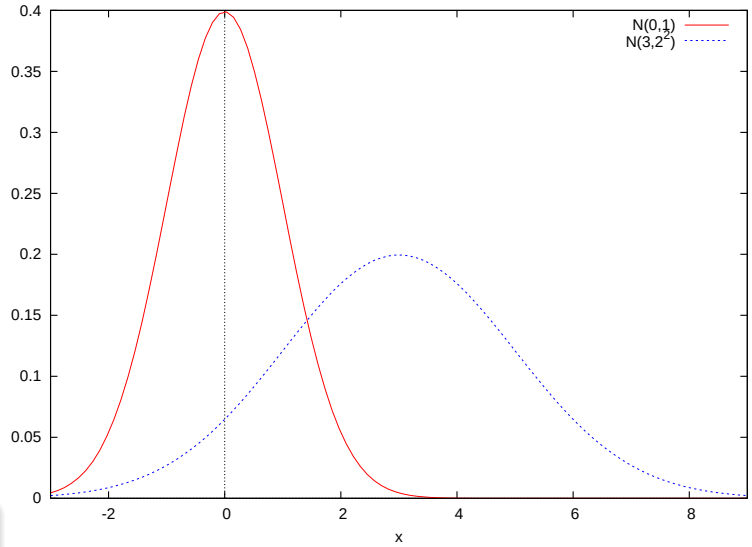
復習:1 変数の正規分布

標準正規分布の確率密度関数

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{z^2}{2}}$$

$X = aZ + b$ を考える. 確率統計☆演習 II(2016)L06

確率密度関数は, z のところに $z = \frac{x-b}{a} = \frac{x-\mu}{\sigma}$ を代入すればいいので,

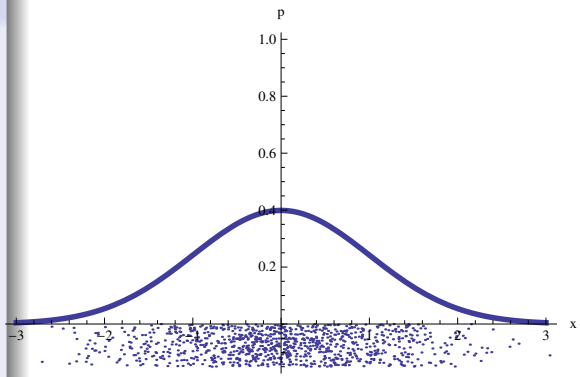


正規分布 $\mathrm{N}(\mu, \sigma^2)$ の確率密度関数

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

パラメタ μ(= 実は $\mathrm{E}[X]$),
σ^2(= 実は $\mathrm{V}[X]$).

確率統計☆演習 I(2015)L08

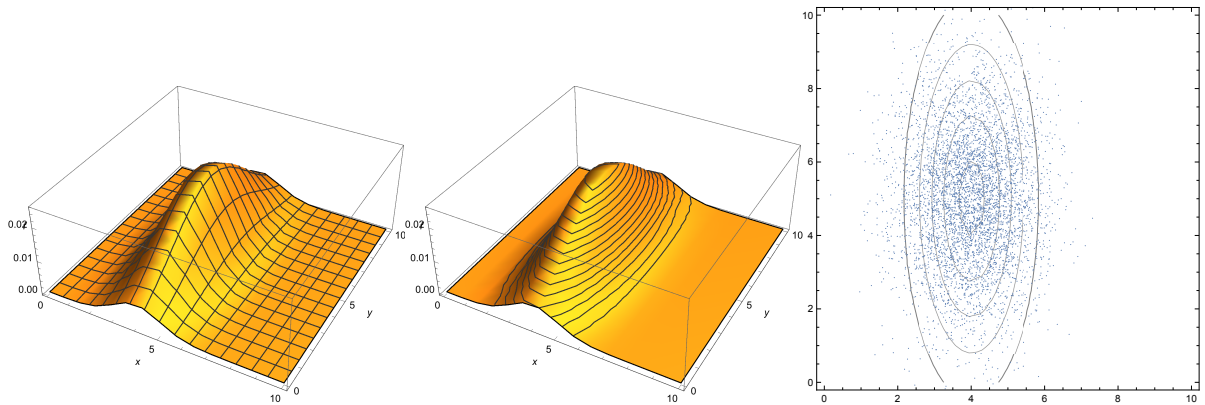


2 次元正規分布 (のうち X, Y が独立な簡単なケース)

$$f_{\mathrm{XY}}(x,y) = \frac{1}{\sqrt{2\pi\sigma_{\mathrm{X}}^2}} \mathrm{e}^{-\frac{(x-\mu_{\mathrm{X}})^2}{2\sigma_{\mathrm{X}}^2}} \times \frac{1}{\sqrt{2\pi\sigma_{\mathrm{Y}}^2}} \mathrm{e}^{-\frac{(y-\mu_{\mathrm{Y}})^2}{2\sigma_{\mathrm{Y}}^2}}.$$

$$\mathrm{E}[X] = \mu_{\mathrm{X}}, \quad \mathrm{E}[Y] = \mu_{\mathrm{Y}},$$
$$\mathrm{V}[X] = \sigma_{\mathrm{X}}^2, \quad \mathrm{V}[Y] = \sigma_{\mathrm{Y}}^2,$$

母共分散 $C_{XY} = \mathrm{Cov}[X,Y] = \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y] = 0 - 0 \cdot 0 = 0.$



ここで使った「公式」

モーメント母関数 $M_X(t) = \mathrm{e}^{\mu t + \frac{1}{2}\sigma^2 t^2}$. から簡単に示せる.
$f(x;\mu,\sigma^2)$: 1 次元正規分布.

$$\int_{-\infty}^{+\infty} x^0 \, f(x;\mu,\sigma^2)\,\mathrm{d}x = \int_{-\infty}^{+\infty} x^0 \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}\,\mathrm{d}x = 1.$$

$$\int_{-\infty}^{+\infty} x^1 \, f(x;\mu,\sigma^2)\,\mathrm{d}x = \int_{-\infty}^{+\infty} x^1 \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}\,\mathrm{d}x = \mu.$$

$$\int_{-\infty}^{+\infty} x^2 \, f(x;\mu,\sigma^2)\,\mathrm{d}x = \int_{-\infty}^{+\infty} x^2 \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}\,\mathrm{d}x = \sigma^2 + \mu^2.$$

$$\int_{-\infty}^{+\infty} (x-\mu)^2 \, f(x;\mu,\sigma^2)\,\mathrm{d}x = \int_{-\infty}^{+\infty} (x-\mu)^2 \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}\,\mathrm{d}x = \sigma^2$$

L10-Q5

Quiz(2 次元正規分布)

次の 2 変数確率密度関数は 2 次元正規分布を定める.

$$f(x,y) = C \cdot \mathrm{e}^{-x^2-4x-2y^2+12y-5}.$$

1. X, Y の母平均値, 母分散, 母共分散を求めよう.
2. $\mathrm{E}[1] = 1$ が満たされるように定数 C を定めよう.

2 次形式の標準化 (線形代数 I)

L10-Q6

Quiz(2 次元正規分布)

次の 2 変数確率密度関数は 2 次元正規分布を定める.

$$f(x,y) = C \cdot \mathrm{e}^{-4x^2 - \frac{1}{6}y^2 + 2y}$$

1. X, Y の母平均値, 母分散, 母共分散を求めよう.
2. $\mathrm{E}[1] = 1$ が満たされるように定数 C を定めよう.

お知らせ

- 予習問題と同じタイミングで,「学期途中のリフレクションレポート」をやりましょう. 100 ピーナッツ以外の 3 ピーナッツ.
- 確率統計☆演習 I と同じセッティングで予習問題をやりましょう.
  `http://hig3.net` → RaMMoodle
  `https://el.math.ryukoku.ac.jp/moodle/` → 確率統計☆演習 II(2016)
- チューター/Math ラウンジ 月火水木昼 1-614

`https://manaba.ryukoku.ac.jp`
マイページの下の方に
manaba 出席カード提出

瀬田龍大生調査プロジェクト

何回かの授業にまたがって, チーム別で, 問題 (RQ=Research Question) をたて, 調査し, 検定して答をだします.

制約

- 指定の検定で答えられるような問題で.
- 母集団=瀬田学舎の龍大生. したがって問題は「瀬田学舎の龍大生の…は…か?」のようになるでしょう.
- 標本=確率統計☆演習 II 参加者. どこかの回で Web で調査します.

今日のタスク

1. クラス別の座席に移動してメンバーを確認する
2. 3 人以上 5 人以下の 1-2 チームに分かれる. 今日欠席していても参加が確実な人はカウントしておく. 1 チームで 1 枚のシートを確保する.
3. manaba の科目のプロジェクトに投稿された問題をチェックする.
4. 問題の第 1 候補, 第 2 候補を選ぶ. manaba にないものを新たに考えてもよい.
    1. 指定の検定で答が得られる
    2. 結果が分かりきってなく「おもしろい」
5. 第 1,2 候補について, 帰無仮説, 対立仮説を書く
6. 第 1,2 候補について, 検定に必要なデータを集めるための質問をそれぞれ 1 または 2 問書く. 多肢選択または実数で回答. 標本=このクラスのメンバー があいまいさなく答えられる質問で.
7. 紙を提出