文部科学省 平成24年度採択 大学間連携共同教育推進事業
「データに基づく課題解決型人材育成に資する統計教育質保証」

平成26年度
データ科学特論 Ⅱ

担当責任：狩野 裕＋内田雅之＋西田 豊
日　　程：2014年8月25日～29日
会　　場：大阪大学　基礎工学研究科(本館)B棟B300
連 絡 先：　西田 豊　nishida@sigmath.es.osaka-u.ac.jp

データ科学特論 Ⅱ 講義内容

- 8/25(月) 午前　狩野 裕＋西田 豊（大阪大学）
  - **データ科学特論 Ⅱ の序**
- 8/25(月) 3限～5限　清水裕士（広島大学）
  - **構造方程式モデリングとマルチレベル分析**
- 8/26(火) 3限～5限　盛山和夫（関西学院大学）
  - **社会調査と計量社会学**
- 8/27(水) 3限～5限　岡田謙介（専修大学）
  - **社会科学におけるベイズ統計**
- 8/28(木) 2限～4限　吉田寿夫（関西学院大学）
  - **心理測定と心理統計**
- 8/29(金) 2限～4限　荘島宏二郎（大学入試センター）
  - **テスト理論 (IRT)**

敬称略

単位認定

- 対象
  - 大阪大学大学院生＋特別聴講学生(同志社大＋大阪府立大)
- 出席とクラス内活動 30%
  - 出席確認
- 課題評価 70%
- レポート課題の提出について
  - 課題解答はA4用紙で作成すること
  - 提出締切：2014年9月9日(火)必着
  - 提出方法：
    - CLEによって電子的に提出
      https://cle.koan.osaka-u.ac.jp/
    - レポート提出箱(基礎工学J棟6階J615数理事務室前) [平日 9:00-17:00]
    - 郵送・宅配便
      - 560-8531 豊中市待兼山町1-3
        大阪大学　基礎工学研究科 数理事務室
        「データ科学特論 Ⅱ レポート在中」と朱書のこと
- 講義アンケートにご協力ください

8/27～29は、CLEのバージョンアップ作業のため、CLEを停止いたします。予めご了承願います。
スライド改訂版UL

データ科学特論 Ⅱ の序
---サンプリングについて---

狩野　裕(大阪大学)
西田　豊(大阪大学)

講義内容

- Motivative Examples
- t-検定と標本サイズ
- 事後層別におけるt-検定
- 母集団全体の推測
- Final message
- 付録：サンプリングの基礎

Motivative Exampleの解答

ある母集団から有権者を無作為に抽出し男女に分ける(事後層別)と有権者名簿で性別を分け，男女の集団のそれぞれから一定数無作為抽出する(層別抽出)．

① 事後層別と層別抽出の違いは何か．
- 層別抽出は単純無作為抽出より母平均の推定精度が高い
- 事後層別では各層における標本サイズは確率変数である
- 母集団における層の構成比率が既知or未知？
  - 既知の時，事後層別はかなりよいパフォーマンス
  - 未知の時，事後層別は単純無作為抽出と同等

② 事後層別した後でt-検定してもよいのか．
- 検定してよい

t-検定によって群間比較をするとき，各群の標本サイズは揃えておく方がよいと言われる．

③ それは何故か？
・検出力(検定力)が高いから

④ 標本サイズが揃っていないとき，大きい方の標本をランダムに削除して両群の標本サイズを揃える価値はあるか．
・削除してはいけない．検出力が低下するから

層別の目的

- Stratification
- 層間で比較をしたい
  - 層(strata, stratum)
  - 多群(multiple groups)，多標本(multi-sample)
- 推定の精度を向上させる
  - 層別のコツ
    - 層内のばらつきを小さく，層間の違いを大きくする
  - サンプリング理論と関係

t−検定と標本サイズ

t−検定(独立二標本, 事前層別抽出)

二つの正規母集団からの無作為標本：

$$Y_1^{(1)}, \cdots, Y_{n_1}^{(1)} \sim N(\mu_1, \sigma^2)$$
$$Y_1^{(2)}, \cdots, Y_{n_2}^{(2)} \sim N(\mu_2, \sigma^2)$$

仮説：

$$H_0 : \mu_1 = \mu_2 \quad \text{versus} \quad H_1 : \mu_1 \neq \mu_2$$

標本平均と不偏分散：

$$\bar{Y}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} Y_i^{(1)}, \quad U_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(Y_i^{(1)} - \bar{Y}_1)^2$$
$$\bar{Y}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i^{(2)}, \quad U_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(Y_i^{(2)} - \bar{Y}_2)^2$$

共通分散σ^2の推定 (合併した不偏分散)：

$$U^2 = \frac{(n_1 - 1)U_1^2 + (n_2 - 1)U_2^2}{n_1 + n_2 - 2}$$

t-検定統計量とその分布：

$$T = \frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{U\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

帰無仮説 $H_0 : \mu_1 = \mu_2$ の下で

$$T = \frac{\bar{Y}_1 - \bar{Y}_2}{U\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

棄却域 (有意水準α)：

$$|T| \geq t_{n_1+n_2-2}(\alpha)$$

第一種の過誤：

$$P(|T| \geq t_{n_1+n_2-2}(\alpha) | H_0) = \alpha$$

β が重要

- 標本サイズが群によって異なっていても, 第一種の過誤 α は保たれている
- 第二種の過誤 β, すなわち, 検出力1− β が標本サイズによって異なる

		検定結果	
		H_0	H_1
真の状況	H_0	OK	α
	H_1	β	OK

余談
- 第一種の過誤の意味
  - 過ちという行為・事象
  - 過誤の確率
- 第一種の過誤と有意水準は異なる

βの評価：非心t−分布

$H_1：\mu_1 \neq \mu_2$の下でのt-統計量の分布:

$$T=\frac{\bar{Y}_1-\bar{Y}_2}{U\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}=\frac{\bar{Y}_1-\bar{Y}_2-(\mu_1-\mu_2)+(\mu_1-\mu_2)}{\frac{U}{\sigma}\times\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

$$=\frac{\frac{\bar{Y}_1-\bar{Y}_2-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}+\frac{\mu_1-\mu_2}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}}{U/\sigma}$$

$$=\frac{N(0,1)+\sqrt{\frac{n_1n_2}{n_1+n_2}}\frac{\mu_2-\mu_1}{\sigma}}{\sqrt{\frac{\chi^2_{n_1+n_2-2}}{n_1+n_2-2}}}=\frac{N(0,1)+\delta}{\sqrt{\frac{\chi^2_{n_1+n_2-2}}{n_1+n_2-2}}}$$

d= $\frac{\mu_1-\mu_2}{\sigma}$: （H_0からのズレの大きさ）

δ^2 : （βをコントロール）

各群（層）の標本サイズは揃えておく方がよい

- 検出力が高いから（第二種の過誤が小さいから）
- n_1=50, n_2=50 is better than n_1=30, n_2=70

$n_1+n_2=n$(一定)のとき，

$\sqrt{\frac{n_1n_2}{n_1+n_2}}$は$n_1=n_2=\frac{n}{2}$のとき最大になる．

n_1	n_2	$\sqrt{\frac{n_1n_2}{n_1+n_2}}$	検出力 (d=0.5)
50	50	5.00	0.70
30	70	4.58	0.62
20	80	4.00	0.51
30	149	5.00	0.70
10	200	3.13	0.34
10	10000	3.16	0.35

Rのコード

```
# install.packages("pwr")
library(pwr)

es1=0.5
pwr.t2n.test(d=es1, n1=50, n2=50, sig.level=0.05,
alternative ="two.sided")
pwr.t2n.test(d=es1, n1=30, n2=70, sig.level=0.05,
alternative ="two.sided")
pwr.t2n.test(d=es1, n1=20, n2=80, sig.level=0.05,
alternative ="two.sided")

pwr.t2n.test(d=es1, n1=30, sig.level=0.05, power=0.70,
alternative ="two.sided")
```

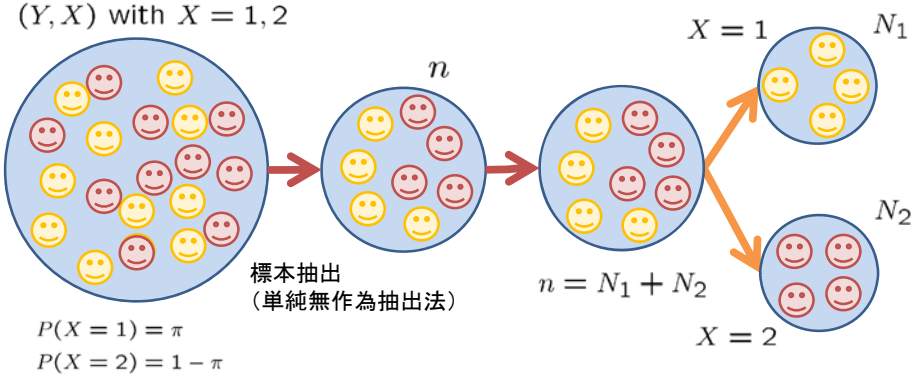

Post-stratification

事後層別における t−検定

22世紀に輝く Harmonious Diversity

事後層別

母集団　　標本　　調査・実験＋層別集計



22世紀に輝く Harmonious Diversity

事後層別後の t−検定は有意水準を保つ

(事前) 層別

棄却域:

$$|T| = \frac{|\bar{Y}_1 - \bar{Y}_2|}{U\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \geq t_{n_1+n_2-2}(\alpha)$$

$n_1 + n_2 = n \quad (n_1, n_2 \geq 2)$

第一種の過誤:

$P(|T| \geq t_{n_1+n_2-2}(\alpha) | H_0) = \alpha$

事後層別

棄却域:

$$|T| = \frac{|\bar{Y}_1 - \bar{Y}_2|}{U\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}} \geq t_{N_1+N_2-2}(\alpha)$$

$N_1 + N_2 = n \quad (N_1, N_2 \geq 2)$

$N_1 \sim B(n, \pi), \quad N_2 \sim B(n, 1-\pi)$

第一種の過誤:

$P(|T| \geq t_{N_1+N_2-2}(\alpha) | H_0) = \alpha$

$$\left. \frac{\frac{1}{N_1}\sum_{i=1}^{N_1} Y_i^{(1)} - \frac{1}{N_2}\sum_{i=1}^{N_2} Y_i^{(2)}}{\sigma\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}} \right|_{\substack{N_1=n_1 \\ N_2=n_2}} \sim N(0,1)$$

22世紀に輝く Harmonious Diversity

検出力

$$1 - \beta = P(|T| \geq t_{N_1+N_2-2}(\alpha) | H_1)]$$

$$= P\left(\left| \frac{N(0,1) + \sqrt{\frac{N_1 N_2}{N_1 + N_2}} \frac{\mu_2 - \mu_1}{\sigma}}{\sqrt{\frac{\chi^2_{N_1+N_2-2}}{N_1 + N_2 - 2}}} \right| \geq t_{N_1+N_2-2}(\alpha) \right)$$

$$= \sum_{\substack{n_1, n_2 \\ n_1+n_2=n}} P\left(\left| \frac{N(0,1) + \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \frac{\mu_2 - \mu_1}{\sigma}}{\sqrt{\frac{\chi^2_{n_1+n_2-2}}{n_1 + n_2 - 2}}} \right| \geq t_{n_1+n_2-2}(\alpha) \middle| \begin{matrix} N_1 = n_1 \\ N_2 = n_2 \end{matrix} \right) \times B(n_1 | n, \pi)$$

$$= \sum_{\substack{n_1, n_2 \\ n_1+n_2=n}} P\left(\left| \frac{N(0,1) + \sqrt{\frac{n_1 n_2}{n}} \frac{\mu_2 - \mu_1}{\sigma}}{\sqrt{\frac{\chi^2_{n-2}}{n-2}}} \right| \geq t_{n-2}(\alpha) \right) \times B(n_1 | n, \pi)$$

層1 : 層2 = 1 : 1

	標本サイズ			検出力
	層1	層2	全標本サイズ	($d = 0.5$)
事前層別	50	50	100	0.697
事後層別	N_1	N_2	100	0.692

層1 : 層2 = 1 : 1

	標本サイズ			検出力
	層1	層2	全標本サイズ	($d = 0.8$)
事前層別	10	10	20	0.395
事後層別	N_1	N_2	20	0.378

層1 : 層2 = 0.3 : 0.7

	標本サイズ			検出力
	層1	層2	全標本サイズ	($d = 0.5$)
事前層別	50	50	100	0.697
事前層別	30	70	100	0.621
事後層別	N_1	N_2	100	0.616

22世紀に輝く Harmonious Diversity

検出力

- 標本サイズを揃えた事前層別がベスト
  - 理論的にも証明可能
- 層のサイズが同じ場合，事後層別も良い
- 層のサイズが異なる場合，事後層別は良くない

まとめ

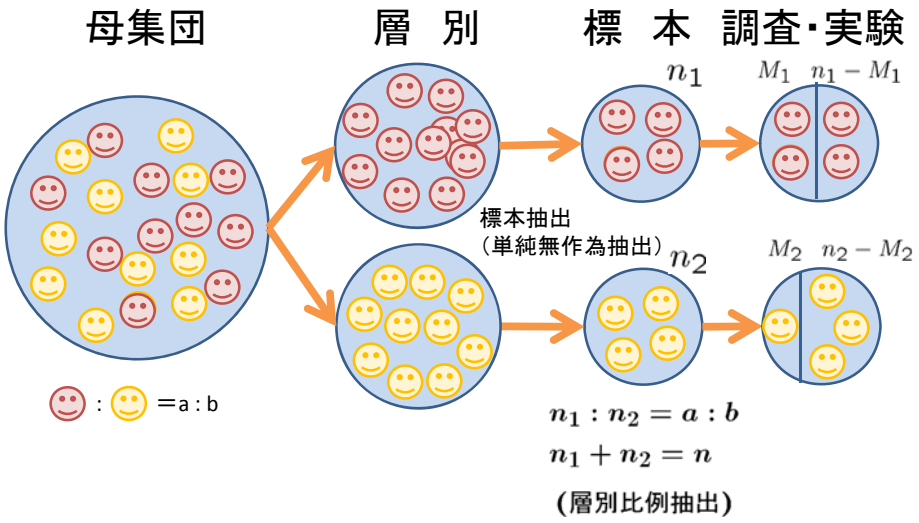
- t-検定によって2つの層(2群)を比較するとき，各層の標本サイズは揃えておく方がよい
  - 検出力が高い(第二種の過誤が小さい)
- 事後層別
  - 標本サイズが確率変数となるが有意水準は保たれる
  - 各層の標本サイズが近い場合，適用可能
  - 各層の標本サイズが相当に異なる場合，検出力の低下は無視できない
  - 極端な場合 n_1=0もありえる
- 可能ならば，標本サイズを揃えた事前層別を行う

Practical issues

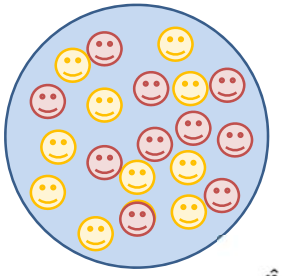
- 層の標本サイズが相当に異なる場合. . .
  - 検出力の低下
    - 非有意を主張したいが故，故意に標本サイズを違えているのではないか，という疑念
      → 標本サイズの設計
  - 標本抽出の適切性に疑問
    - 標本抽出の偏りが疑われる
      - e.g., 標本選択, トランケーション
    - 層のサイズと合っている場合，標本抽出はOK
  - 標本サイズが揃っている場合，多少の不等分散性に対して頑健
    - 永田(1996)
- 事後層別なのに. . .
  - n_1=100, n_2=100?
  - n_1=314, n_2=314?

母集団全体の推測

(事前)層別抽出



（事前）層別抽出



: = a : b

層	層ごと A党支持率	層ごと A党支持者数	母集団全体 A党支持率
	p_1	全数$\times\frac{a}{a+b}\times p_1$	$\frac{ap_1+bp_2}{a+b}$
	p_2	全数$\times\frac{b}{a+b}\times p_2$	

$$\hat{p}_{\text{前層}} = \frac{a\frac{M_1}{n_1}+b\frac{M_2}{n_2}}{a+b} = \frac{n_1\frac{M_1}{n_1}+n_2\frac{M_2}{n_2}}{n_1+n_2} = \frac{M_1+M_2}{n}$$

$$V(\hat{p}_{\text{前層}}) = \frac{1}{n^2}\left(n_1p_1(1-p_1)+n_2p_2(1-p_2)\right)$$

$$= \frac{ap_1(1-p_1)+bp_2(1-p_2)}{n(a+b)}$$

注意
層別比例抽出のとき
$n_1 : n_2 = a : b$

層別比例抽出について

- (size) proportionate stratified sampling
- 母集団のどの個体も抽出される確率は同一
- 推定精度が高い

推定量の分散

$$V\left(\widehat{\frac{ap_1+bp_2}{a+b}}\right) = V\left(\frac{a\frac{M_1}{n_1}+b\frac{M_2}{n_2}}{a+b}\right) = \frac{\frac{a^2p_1(1-p_1)}{n_1}+\frac{b^2p_2(1-p_2)}{n_2}}{(a+b)^2}$$

これを$n_1+n_2=n$の下で最小にするのは次のとき：

$$n_1 : n_2 = a\sqrt{p_1(1-p_1)} : b\sqrt{p_2(1-p_2)} \quad \text{(ネイマン抽出法)}$$

$$\approx a : b \quad \text{(層別比例抽出法)}$$

シュワルツの不等式

$$\left(\frac{A_1^2}{n_1}+\frac{A_2^2}{n_2}\right)(n_1+n_2) \geq (A_1+A_2)^2$$

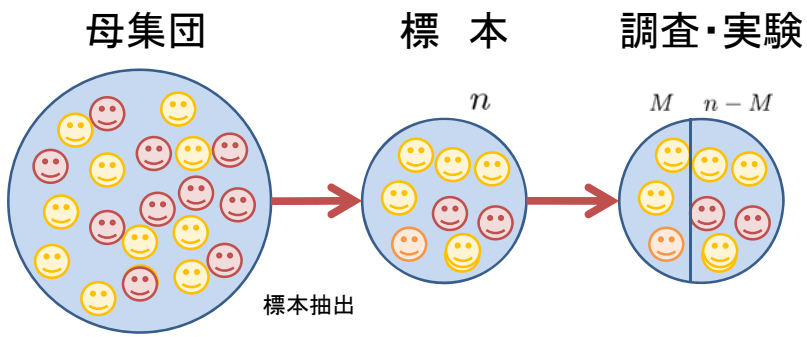
等号成立は

$$n_1 : n_2 = A_1 : A_2$$

のとき.

ネイマン抽出法がbestであることの証明

単純無作為抽出法



A党支持の割合

$$p = \frac{ap_1+bp_2}{a+b}$$

$$\hat{p}_{\text{単純}} = \frac{M}{n}$$

$$V(\hat{p}_{\text{単純}}) = \frac{p(1-p)}{n} \text{ with } p = \frac{ap_1+bp_2}{a+b}$$

単純無作為抽出 vs 層別抽出

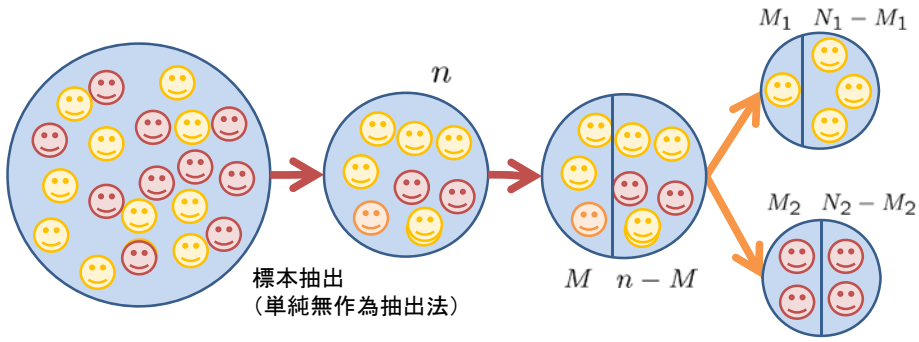
$$V(\hat{p}_{\text{単純}}) - V(\hat{p}_{\text{前層}}) = \frac{\frac{ap_1+bp_2}{a+b}\left(1-\frac{ap_1+bp_2}{a+b}\right)}{n} - \frac{ap_1(1-p_1)+bp_2(1-p_2)}{n(a+b)}$$

$$= \frac{ab(p_1-p_2)^2}{n(a+b)^2} \geq 0$$

- 層別抽出法は単純無作為抽出法より精度が高い
- 層の違いが大きいほど精度の差がひらく
- 層に差がないときは精度は同一
  - 層別抽出する価値は無い

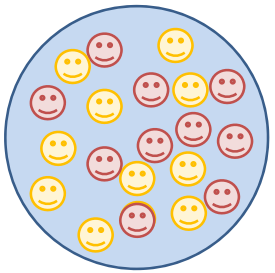
事後層別

母集団　　標　本　　調査・実験　層別集計



$$n = N_1 + N_2$$

事後層別（各層のサイズ比が既知）



：　＝a : b

層	層ごと A党支持率	層ごと A党支持者数	母集団 A党支持率
😊	p_1	全数 $\times\frac{a}{a+b}\times p_1$	$\frac{ap_1+bp_2}{a+b}$
🙂	p_2	全数 $\times\frac{b}{a+b}\times p_2$	

事後層別による推定量とその分散（ばらつき）：

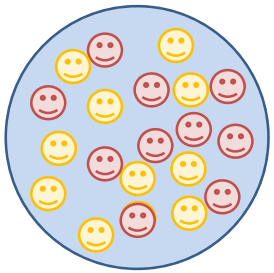
$$\hat{p}_{\text{後層}} = \frac{a\frac{M_1}{N_1}+b\frac{M_2}{N_2}}{a+b} \quad (0 < N_1, N_2 < n)$$

$$V(\hat{p}_{\text{後層}}) = \frac{a^2p_1(1-p_1)E\left(\frac{1}{N_1}\right)+b^2p_2(1-p_2)E\left(\frac{1}{N_2}\right)}{(a+b)^2}$$

単純無作為抽出 → 層サイズ比による調整：

$$\hat{p}_{\text{単純}} = \frac{M}{n} \longrightarrow \frac{a\frac{M_1}{N_1}+b\frac{M_2}{N_2}}{a+b}$$

事後層別（各層のサイズ比が未知）



：　＝a : b

未知！

層	層ごと A党支持率	層ごと A党支持者数	母集団 A党支持率
😊	p_1	全数 $\times\frac{a}{a+b}\times p_1$	$\frac{ap_1+bp_2}{a+b}$
🙂	p_2	全数 $\times\frac{b}{a+b}\times p_2$	

$$\hat{p}_{\text{後層}} = \frac{\hat{a}\frac{M_1}{N_1}+\hat{b}\frac{M_2}{N_2}}{\hat{a}+\hat{b}} \quad (0 < N_1, N_2 < n)$$

$$= \frac{N_1\frac{M_1}{N_1}+N_2\frac{M_2}{N_2}}{N_1+N_2}$$

$$= \frac{M_1+M_2}{N_1+N_2} = \frac{M}{n} = \hat{p}_{\text{単純}}$$

事後層別の価値は無い！

推定精度のまとめ

$$V(\hat{p}_{\text{単純}}) = \frac{\frac{ap_1+bp_2}{a+b}\left(1-\frac{ap_1+bp_2}{a+b}\right)}{n}$$

$$V(\hat{p}_{\text{前層}}) = \frac{ap_1(1-p_1)+bp_2(1-p_2)}{n(a+b)} \quad \text{(層別比例抽出)}$$

$$V(\hat{p}_{\text{後層}}) = \frac{a^2p_1(1-p_1)E\left(\frac{1}{N_1}\right)+b^2p_2(1-p_2)E\left(\frac{1}{N_2}\right)}{(a+b)^2}$$

（各層のサイズが既知）

nが大きいとき，次の関係が成立する：

$$V(\hat{p}_{\text{単純}}) \geq V(\hat{p}_{\text{後層}}) \geq V(\hat{p}_{\text{前層}})$$

大阪大学 OSAKA UNIVERSITY

数値比較

- 母集団
  - 層：性別
  - 男女比率＝1：1
  - 真値
    - 政党Aの母支持率：男性 0.3，女性 0.7
- 標本
  - n=100
- 標準誤差（理論）
  - 0.0500 単純無作為抽出
  - 0.0461 事後層別
  - 0.0458 (事前)層別抽出
- 標準誤差(数値実験，反復回数= 10,000)
  - 0.0462 事後層別

大阪大学 OSAKA UNIVERSITY

まとめ

- 母集団全体の平均の推定
  - 層のサイズ（比）が既知
    - 層別が単純無作為抽出より良い
      - 事前層別＋層別比例抽出
        or 事後層別＋層サイズ比による調整
      - 理由
        - » 抽出の方法の違い
        - » 層サイズ(比)が分かっていること
    - 特性値に関して層間の違いが大きいほど層別は良い
  - 層のサイズ（比）が未知
    - 単純無作為抽出法
    - 層別抽出は適用不可能

大阪大学 OSAKA UNIVERSITY

Summary

- データは，数値（カテゴリー）とその履歴が分かって初めて意味をもつ
  - どのようにして得られた（出てきた）データかということ
  - サンプリング（標本抽出）の方法は典型的なデータの履歴

$$\hat{p}_{\text{単純}} = \frac{M}{n}, \quad \hat{p}_{\text{前層}} = \frac{M_1+M_2}{n}$$

- 標本抽出の性質を理解し，与えられた状況でベスト（ベター）な方法を選択する

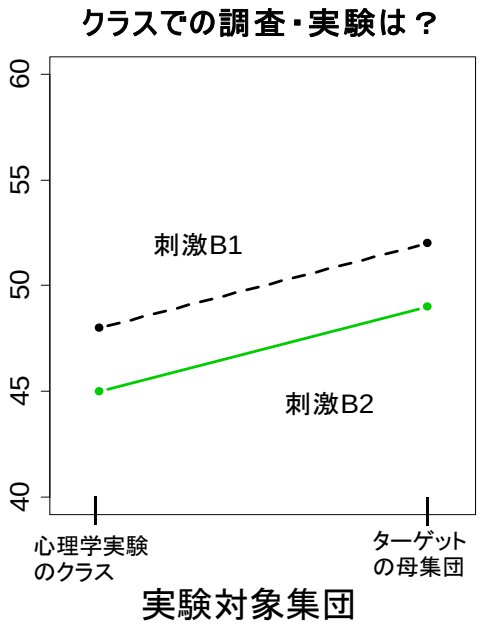
大阪大学 OSAKA UNIVERSITY

Summary 2

- 母集団全体＝層1＋層2
  - 層1:層2 ＝ a : b
- 群比較
  - 標本サイズは揃えておく: $n_1 : n_2 = 1 : 1$
  - 推測の対象: $\mu_1 - \mu_2, \quad p_1 - p_2$
- 母集団全体の推定
  - 層別比例抽出: $n_1 : n_2 = a : b$
  - 推測の対象: $\frac{a\mu_1 + b\mu_2}{a+b}, \quad \frac{ap_1 + bp_2}{a+b}$

Final message

- 社会学系は標本抽出を重視
  - どの構成要素も選ばれる確率は同一
- 心理学系は条件統制を重視
  - 比較を重視
- 理工生物系は個体差をあまり問題にしない
  - 再現性を仮定
- それぞれの分野においてそれなりの理屈はある
  - データの特徴
  - 実行可能性



文献

- 浅井晃(1987). 調査の技術. 日科技連
- 鈴木督久・佐藤寧・棟近雅彦(2012). アンケート調査の計画・分析入門. 日科技連
- 盛山和夫(2004). 社会調査法入門. 有斐閣ブックス
- 永田靖(1996). 統計的方法のしくみ. 日科技連
- 日本統計学会編(2012)「統計検定2級対応　統計学基礎」 東京図書
- 南風原朝和 (2002)「心理統計学の基礎―統合的理解のために」有斐閣
- 森敏昭・吉田寿夫(1990). 心理学のためのデータ解析テクニカルブック. 北大路書房

22世紀に輝く Harmonious Diversity

JMOOC

- 無料で学べる大学講座
  - http://gacco.org/
- ga014: 統計学Ⅰ:データ分析の基礎
  - 2014年11月12日開講
  - 講師
    - 東京大学 竹村彰通
    - 山梨大学 下川敏雄
    - 中央大学 酒折文武
    - 首都大学東京 中山厚穂
    - 総務省統計局

大阪大学 OSAKA UNIVERSITY

22世紀に輝く Harmonious Diversity

大学院等高度副プログラム「データ科学」

データ科学とは何か

データ科学には定まった定義はないが，データ科学をデータが関わる研究を行う学問と考えるならばその守備範囲は広大である．大学は学問の府であるから，データが重要な役割を果たす実証研究に直結する研究のデザインやデータのハンドリングの方法（統計手法）の習得が，データ科学の中でも，肝要である．本副プログラムは，こういった意味でのデータ科学の実践的かつ包括的な教育コースを提供する．実証研究のデータ科学を身に付けた修了生は，実社会でもデータに関わる実務においてそのスキルを十分に活かすことができる．

データ科学の目的

・データ科学の基本的な考え方と統計手法の数理的基礎を理解する
・主専攻の研究分野に直結する統計手法を体系的に学ぶ
・主専攻でない分野におけるデータ科学を知り学際的な視点を養う
・最新の統計手法に関する情報を得る
・データ科学の教育における課題を発見し教育方法の改善に資する

統計数理コース

授業科目名	単位数 選択必修	単位数 選択	開講学期	開講部局
データ科学特論Ⅰ	2		1	基礎工
データ科学特論Ⅱ	2		1	基礎工
統計的推測	2		2	基礎工
多変量解析	2		2	基礎工
時系列解析		2	2	基礎工
確率解析		2	1	基礎工
確率微分方程式		2	2	基礎工
行動統計科学特講Ⅱ		2	1	基礎工
統計・情報数学概論		2	1	基礎工

機械学習コース

授業科目名	単位数 選択必修	単位数 選択	開講学期	開講部局
データ科学特論Ⅰ	2		1	基礎工
データ科学特論Ⅱ	2		1	基礎工
データマイニング工学	2		2	工
統計解析	2		1	基礎工
リスク評価論		2	1	工
統計モデリング		2	1	基礎工
データ解析		2	2	基礎工
数理特論 Ⅱ		2	1	基礎工

大阪大学 OSAKA UNIVERSITY

22世紀に輝く Harmonious Diversity

人文社会統計学コース

授業科目名	単位数 選択必修	単位数 選択	開講学期	開講部局
データ科学特論Ⅰ	2		1	基礎工
データ科学特論Ⅱ	2		1	基礎工
行動統計科学特講 Ⅰ	2		2	人間科
経験社会学特講	2		2	人間科
行動統計科学特講 Ⅱ		2	1	人間科
計量社会学特講		2	1	人間科
教育動態学特講		2	2	人間科
多変量解析		2	2	基礎工
標本調査		2	2	経済

保健医療統計学コース

授業科目名	単位数 選択必修	単位数 選択	開講学期	開講部局
データ科学特論Ⅰ	2		1	基礎工
データ科学特論Ⅱ	2		1	基礎工
保健情報論	2		1	医学系
医学統計学基礎	2		2	医学系
医学統計学応用		2	1	医学系
臨床試験デザイン基礎		2	2	医学系
観察研究の統計的方法		2	1	医学系
リスク評価論		2	1	工
行動統計科学特講 Ⅰ		2	2	人間科
行動統計科学特講 Ⅱ		2	1	人間科

経済経営統計学コース

授業科目名	単位数 選択必修	単位数 選択	開講学期	開講部局
データ科学特論Ⅰ	2		1	基礎工
データ科学特論Ⅱ	2		1	基礎工
エコノメトリックス Ⅰ	2		1	経済
行動統計科学特講 Ⅰ	2		2	人間科
統計解析		2	1	経済
エコノメトリックス Ⅱ		2	2	経済
マーケティング・サイエンス		2	2	経済
標本調査		2	2	経済
多変量解析		2	2	基礎工
データ解析		2	2	基礎工

大阪大学 OSAKA UNIVERSITY

データ科学 II の序 (狩野＋西田) の課題 (H26/8/25 改訂版)
1 または 2 のいずれかを選択し解答すること.

1 ある（無限）集団 G において，男女比が $\pi : 1-\pi$ であり，男性のビール消費量 Y_1 は $N(\mu_1, \sigma^2)$，女性のビール消費量 Y_2 は $N(\mu_2, \sigma^2)$ に従う．ここで π $(0 < \pi < 1)$ は既知である．このとき，この集団から任意に一人抽出したときのビール消費量 Y の分布は

$$Y \sim \pi N(\mu_1, \sigma^2) + (1-\pi) N(\mu_2, \sigma^2)$$

である．集団 G から単純無作為抽出法によって大きさ n (≥ 4) の標本を抽出する，もしくは，(事前) 層別抽出法によって男女の層からそれぞれ大きさ n_1, n_2 の標本を抽出する．ただし，$n = n_1 + n_2$ である．以下の設問に答えよ.

(1) $E[Y]$ を求めよ.

(2) シュワルツの不等式 (スライド 27) を証明せよ.

(3) 下記 3 つの母数（の関数）を（事前）層別抽出法によってデータを採取し推定する．各母数に対して最適な標本サイズを設計せよ.

(a) $\pi\mu_1 + (1-\pi)\mu_2$

(b) $\mu_1 - \mu_2$

(c) $\pi\mu_1 - (1-\pi)\mu_2$

(4) (a) の母数を推定する際，層別抽出法と単純無作為抽出法を比較せよ.

(5) この課題を解くにあたって考えたことを記せ.

補足．上述の母数はそれぞれ次のような意味をもつ.

(a) 集団 G における総消費量を構成員一人あたりに変換した量

(b) 集団 G における一人あたりの消費量の男女差

(c) 集団 G における男性の総消費量と女性の総消費量の差を構成員一人あたりに変換した量

2 （事前）層別抽出法 (+層別比例抽出）が単純無作為抽出法に勝るという主張について以下の設問に答えよ.

(1) 講義では推定量の分散の比較に基づいて上述の主張を行った．このことを復習し纏めよ.

(2) 上述の主張を数式を使わずに直感的に説明せよ.

以上