現代日本論演習／比較現代日本論研究演習 I「統計分析の基礎」

第 3 講　統計分析の基礎

田中重人 (東北大学文学部准教授)

[テーマ]　累積度数の利用と統計分析の基礎

1　前回課題について

- ラベルの利用：「データビュー」の「変数ビュー」タブで、変数の値に「ラベル」をつける →分析結果出力に表示される
- 40 代はなぜ多いのか →1995 年の人口ピラミッド <http://www.ipss.go.jp/site-ad/TopPageData/1995.png>
- 70 代はなぜ少ないのか

「再」マークがついている人は再提出 (来週水曜正午まで)

2　度数分布表の読みかた

- 度数
- 相対度数 （%）
- 累積度数・累積相対度数
- 欠損値のあつかい

(教科書 p. 27–31)

3　今回の課題

年齢（カテゴリ統合していない元の変数）の度数分布から、中央値と四分位を求めよ (提出は不要)。
参考資料：

- 教科書 p. 43
- 総務省統計局「なるほど統計学園高等部：データの特性を見よう」<http://www.stat.go.jp/koukou/howto/process/proc4_3_1.htm>
- 船津好明「統計計算の方法」<http://www.wwq.jp/stacal.htm>

また、任意のパーセンタイル (percentile) を求める方法を考えること。

4　発展問題 (余裕のある人のみ)

次の情報を参考にして、カテゴリ統合した後の年齢の度数分布表から中央値を求める方法を考える

- Yahoo! 知恵袋 <http://detail.chiebukuro.yahoo.co.jp/qa/question_detail/q1214126522> の回答
- 青木繁伸「中央値 (M_e)」<http://aoki2.si.gunma-u.ac.jp/lecture/Univariate/median.html>「One more step!」以降
- 森・吉田 (1990, p.15)

5　データ収集から分析まで

(1) データの収集 (実験／観察)
(2) 分析可能な形に加工
(3) データ・セット作成
(4) クリーニング
(5) データの特徴を少数の数値に要約 ＝ 記述統計
(6) 誤差の評価 (この手続きの一部が推測統計)

(教科書 p. 1–6)

6　標本抽出

標本抽出の 4 段階モデル

- 理論母集団 (universe) = 興味の対象となる人や事物の全体
- 調査母集団 (population) = 調査の対象とする具体的な範囲
- 計画標本 (designed sample) = 母集団から抽出した対象者のこと
- 有効標本 (valid sample / case) = 調査の結果あつまった有効なデータ

「無作為抽出」(random sampling) とは：

- 母集団から計画標本を選ぶ際に、母集団にふくまれるすべての個体の抽出確率が等しくなるように抽出する
- この結果として、**「確率標本」** (probability sample) がえられる

統計的な推測のための理屈は、確率標本を前提として組み立てられている。母集団の人口がわかっていて、全個体を網羅した台帳がないと、無作為抽出はできない。実際にはそういうことはないので、いろいろ工夫して無作為抽出に近づける。

「層化 2 段無作為抽出」はその方法のひとつ：

- まず「地点」を抽出 (第 1 次抽出)
- その際、地域・都市規模等で地点抽出数を割り当てておく (層化)
- その地点の台帳から個人を抽出 (第 2 次抽出)

7　宿題

(1) 教科書 pp. 7–16 を元に、「データの種類」の分類についてまとめよ
(2) SSM 調査の質問項目のうち、比率尺度に当たるものはどれか
(3) 「中央値」「四分位」などに意味があるのはどの種類のデータか
(4) 「収入」や「学歴」を比率尺度として分析するにはどのようにすればよいか

ISTU で 5/11(水) 正午までに提出。