

第2回 1 変量データの整理 (2)

村澤 康友

2022年9月30日

今日のポイント

1. 変量の尺度（名義／順序／間隔／比）により，適切なデータ整理の方法は異なる．
2. 度数分布（表／ヒストグラム）はデータ整理の基本．
3. 記述統計量（位置／散らばり）でデータの特徴をみる．

目次

1	変量の尺度 (p. 27)	1
2	度数分布 (p. 18)	1
2.1	度数 (p. 18)	1
2.2	累積度数 (p. 19)	2
3	記述統計量 (p. 28)	2
3.1	総和記号	2
3.2	位置 (p. 28)	2
3.3	散らばり (p. 35)	2
4	今日のキーワード	3
5	次回までの準備	3
1	変量の尺度 (p. 27)	

変量の尺度により，適切なデータ整理の方法は異なる．

定義 1. 順序がない分類を**名義尺度**という．

注 1. 「最大値」「最小値」「平均」は無意味．

例 1. 婚姻状態（未婚・既婚・離別・死別）．

定義 2. 順序がある分類を**順序尺度**という．

注 2. 「平均」は無意味．

例 2. 最終学歴（中卒・高卒・大卒）．

定義 3. 間隔のみが意味をもつ量を**間隔尺度**という．

例 3. 摂氏・華氏，時刻．

定義 4. 比率が意味をもつ量を**比尺度**という．

注 3. 一般に正の値しかとらない．

例 4. 身長，体重，時間，絶対（熱力学）温度．

2 度数分布 (p. 18)

2.1 度数 (p. 18)

まず最初に観測値の範囲をいくつかの**階級**に分割する．

定義 5. ある階級に含まれる観測値の数を，その階級の**度数**という．

定義 6. (度数) / (観測値の総数) を**相対度数**という．

定義 7. 横軸に値をとり，各階級の（相対）度数を柱の面積で表したグラフを**ヒストグラム（柱状グラフ）**という．

注 4. 柱の高さで表す棒グラフとは異なる．階級分けしない離散変量は棒グラフでよい．

注 5. ヒストグラムの印象は階級の取り方により異なる．粗すぎても細かすぎてもダメ．

例 5. 某大学 1 年生の入試成績（英語）の度数分布（表 1）とヒストグラム（図 1）.

2.2 累積度数 (p. 19)

定義 8. ある階級以下の度数の和を, その階級までの**累積度数**という.

注 6. 名義尺度なら無意味.

定義 9. (累積度数) / (観測値の総数) を**累積相対度数**という.

定義 10. 累積 (相対) 度数の折れ線グラフを**累積 (相対) 度数グラフ**という.

注 7. 階級が細かいほど滑らかなグラフとなる.

例 6. 某大学 1 年生の入試成績（英語）の累積度数分布（表 2）と累積度数グラフ（図 2）.

定義 11. 横軸に累積相対度数, 縦軸に (その階級以下の観測値の総和) / (全観測値の総和) をとった折れ線グラフを**ローレンツ曲線**という.

注 8. 全観測値が等しければ 45 度線に一致. 下に行くほど「不平等」な分布.

例 7. 某大学 1 年生の入試成績（英語）のローレンツ曲線（図 3）.

3 記述統計量 (p. 28)

3.1 総和記号

定義 12.

$$\sum_{i=1}^n x_i := x_1 + \cdots + x_n$$

練習 1. 以下の公式を示しなさい.

1. $\sum_{i=1}^n 1 = n$
2. $\sum_{i=1}^n ax_i = a \sum_{i=1}^n x_i$
3. $\sum_{i=1}^n (x_i + y_i) = \sum_{i=1}^n x_i + \sum_{i=1}^n y_i$

3.2 位置 (p. 28)

定義 13. (観測値の総和) / (観測値の総数) を**(算術) 平均**という.

注 9. 質的変量なら無意味.

注 10. 観測値を (x_1, \dots, x_n) とすると (とりあえず母集団と標本は区別しない)

$$\mu := \frac{x_1 + \cdots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i$$

定義 14. 観測値を小さい方から順に並べたときの中央の値を**中位数**という.

注 11. データの総数が偶数で中央の値が存在しない場合は両隣の間をとる.

注 12. 順序尺度でも意味をもつ.

注 13. 対称な分布なら平均 = 中位数.

定義 15. 観測値を小さい方から順に並べたときの αn 番目の値を α **分位点**という.

注 14. αn 番目の値が存在しない場合は両隣の間をとる.

注 15. 中位数は 0.5 分位点.

注 16. 特に四分位点, 五分位点, 十分位点, 百分位 (パーセント) 点がよく用いられる.

定義 16. 度数が最大となる値を**最頻値**という.

注 17. 階級の取り方に依存する.

注 18. 名義尺度でも意味をもつ.

注 19. 対称で単峰な分布なら平均 = 中位数 = 最頻値.

3.3 散らばり (p. 35)

定義 17. 平均からの偏差の 2 乗の平均を**分散**という.

注 20. 式で表すと

$$\sigma^2 := \frac{(x_1 - \mu)^2 + \cdots + (x_n - \mu)^2}{n} = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$$

定義 18. 分散の平方根を**標準偏差**という.

定義 19. (標準偏差) / (平均) を**変動係数**という.

注 21. 変動係数は測定単位の影響を受けない.

表 1 某大学 1 年生の入試成績（英語）の度数分布

階級	度数	相対度数
200～250	2	.00
250～300	11	.03
300～350	15	.04
350～400	30	.07
400～450	63	.15
450～500	95	.22
500～550	92	.22
550～600	67	.16
600～650	33	.08
650～700	19	.04
計	427	1.00

表 2 某大学 1 年生の入試成績（英語）の累積度数分布

階級	累積度数	累積相対度数
200～250	2	.00
250～300	13	.03
300～350	28	.07
350～400	58	.14
400～450	121	.28
450～500	216	.51
500～550	308	.72
550～600	375	.88
600～650	408	.96
650～700	427	1.00

注 22. 平均が正でないと（比尺度でないと）無意味.

定義 20.（ローレンツ曲線と 45 度線の間の面積）
／（45 度線の下での面積）を**ジニ係数**という.

注 23. 45 度線の下での面積は $1/2$.

注 24. 不平等度（格差）を表す.

ツ曲線,（算術）平均, 中位数, 分位点, 最頻値, 分散, 標準偏差, 変動係数, ジニ係数

5 次回までの準備

復習 教科書第 2 章, 復習テスト 2

予習 教科書第 3 章

4 今日のキーワード

名義尺度, 順序尺度, 間隔尺度, 比尺度, 度数,
相対度数, ヒストグラム（柱状グラフ）, 累積度数,
累積相対度数, 累積（相対）度数グラフ, ローレン

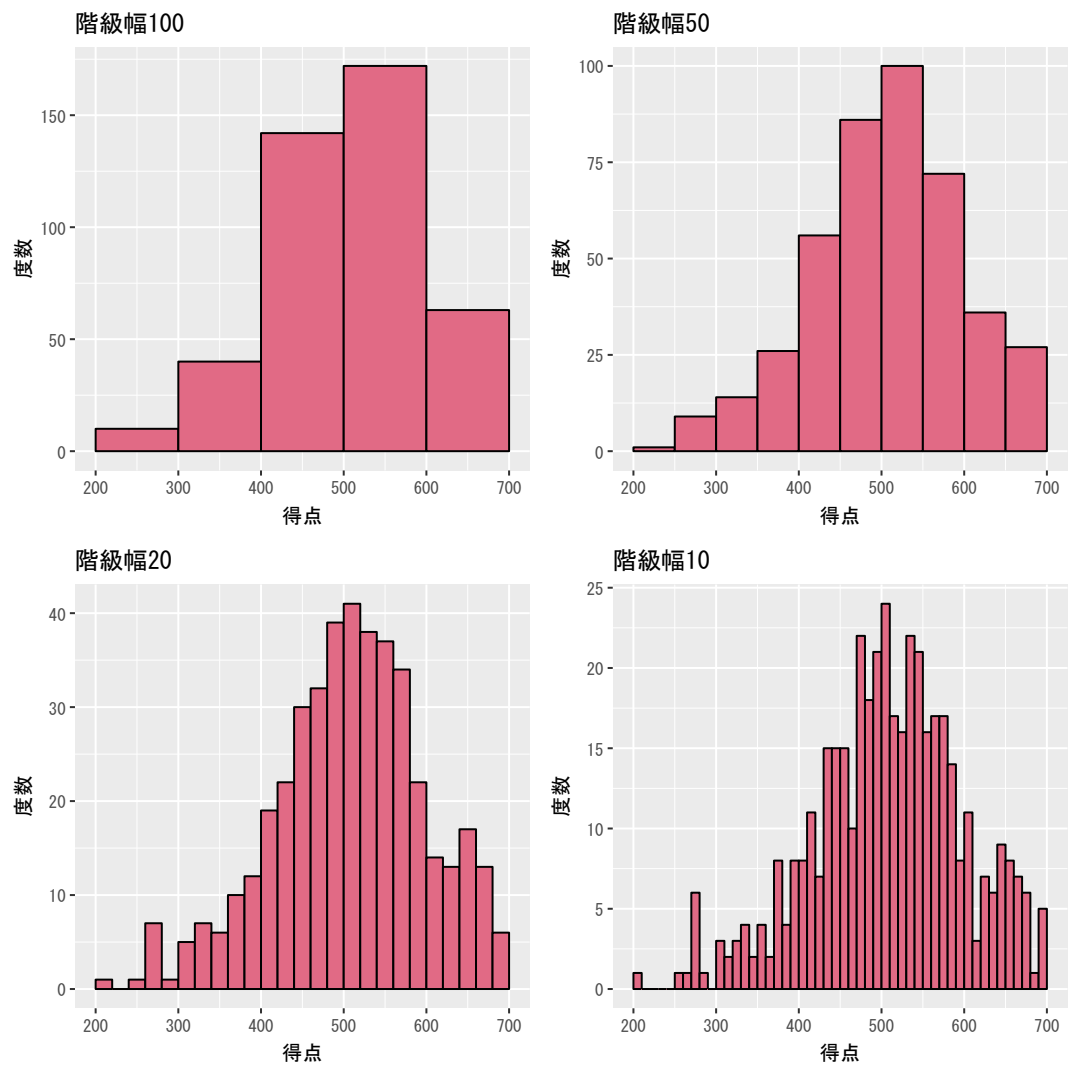


図1 某大学1年生の入試成績（英語）のヒストグラム

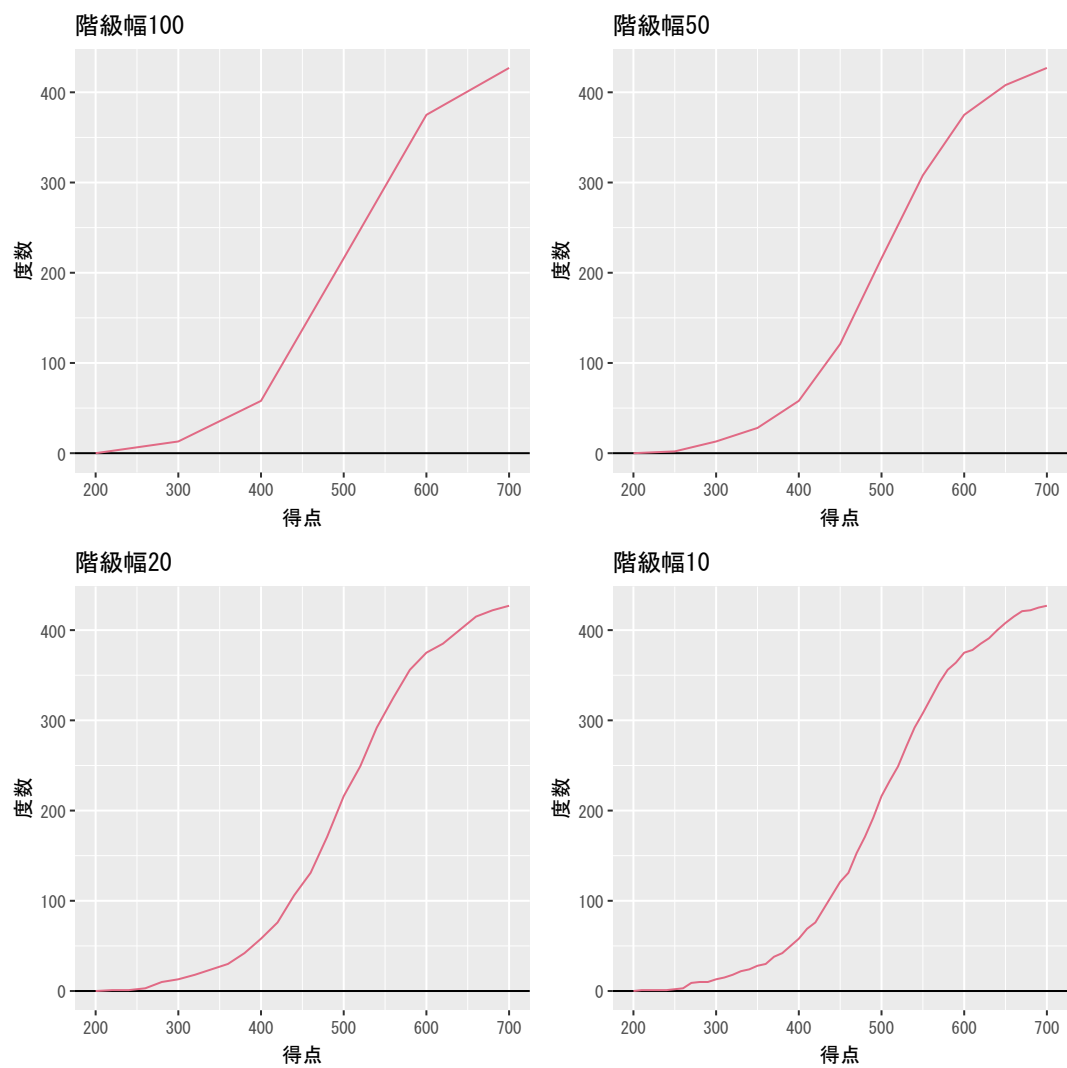


図2 某大学1年生の入試成績（英語）の累積度数グラフ

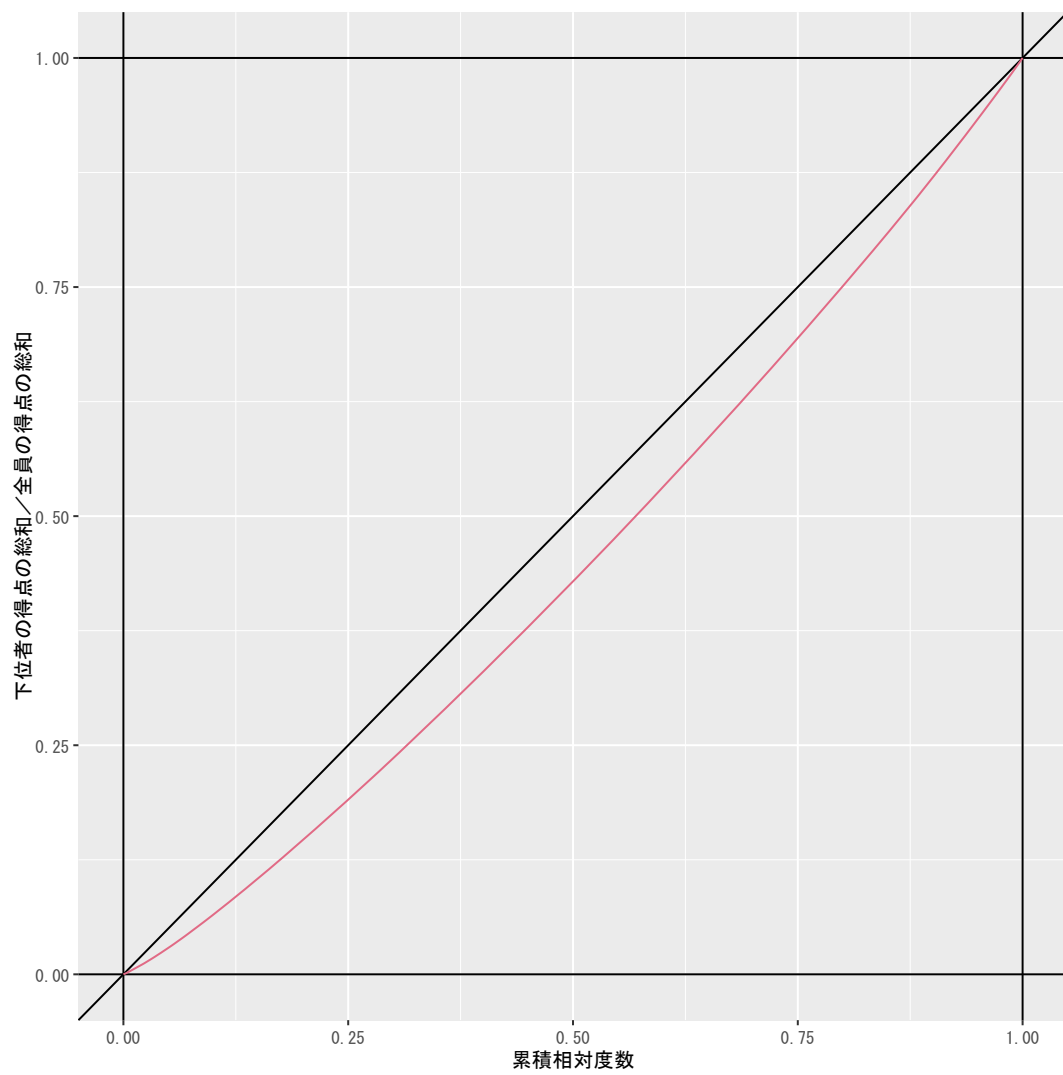


図3 某大学1年生の入試成績（英語）のローレンツ曲線