

統計学I

早稲田大学政治経済学術院

西郷 浩

本日の目標

- 度数分布表からの近似計算
 - ヒストグラムの作成
 - 算術平均、分散の計算
 - 中央値、四分位範囲の計算
 - ローレンツ曲線の描画、ジニ係数の計算
- PC実習

度数分布表による統計の提供(1)

- 公表されている統計(の多く)
 - 原データ x_i : 利用不可
度数分布表: 利用可
 - 例: 総務省統計局『第67回日本統計年鑑 2018』
 - 表21-13 住宅の所有関係、敷地面積別一戸建て住宅数
(借家、2013年)

度数分布表による統計の提供(2)

表1: 敷地面積別一戸建住宅数(借家、2013年)

階級	下限(m^2 以上)	上限(m^2 未満)	度数(1000戸)
1	-	50	216
2	50	75	367
3	75	100	311
4	100	150	375
5	150	200	233
6	200	300	177
7	300	500	90
8	500	-	38
		合計	1,807

資料: 総務省統計局『第67回日本統計年鑑2018』表21-13

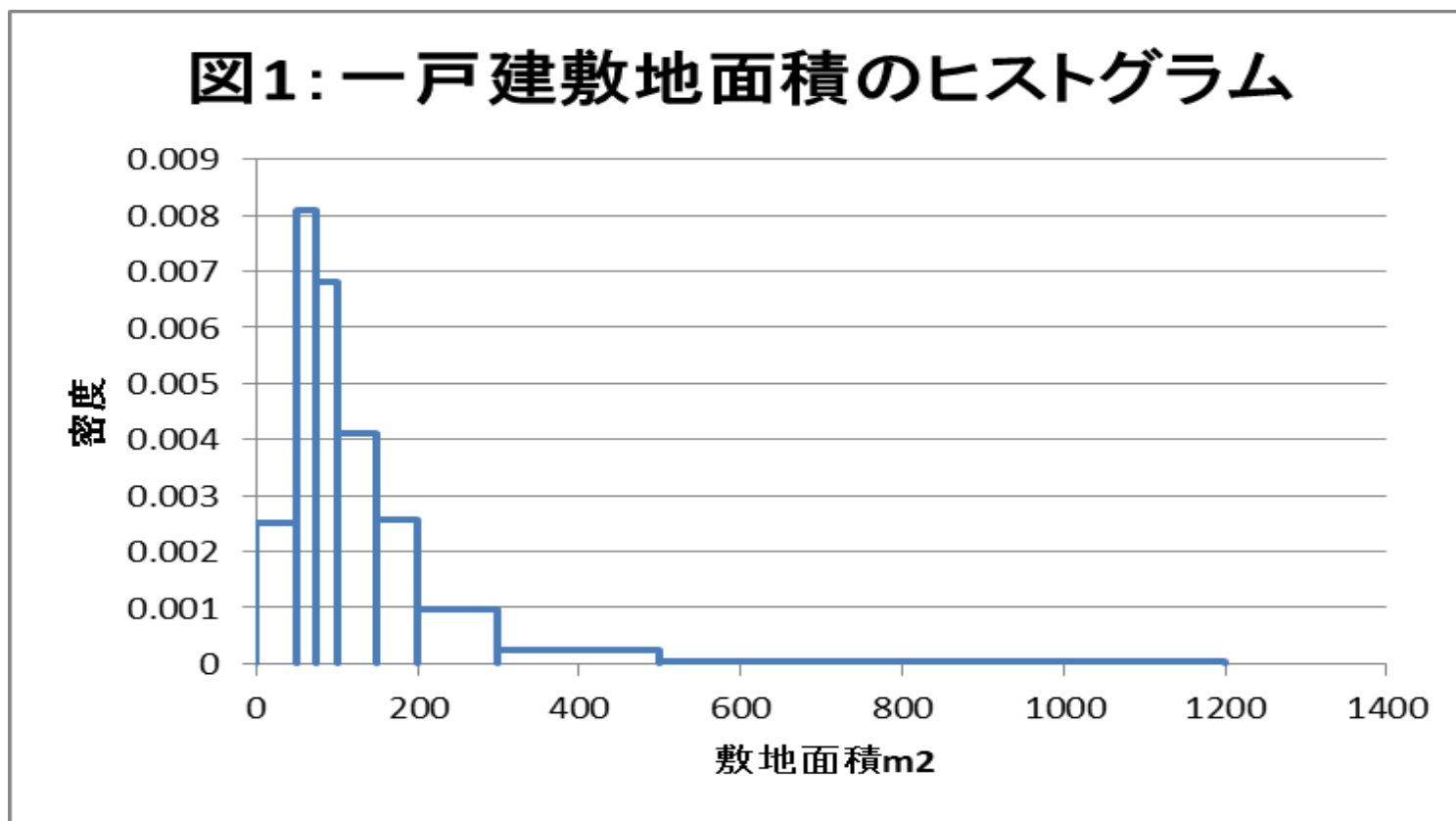
ヒストグラムの作成(1)

- 度数分布表からのヒストグラムの作成
 - 注意点
 - 異なる階級幅が混在している。
 - 原則：柱の面積が度数に比例するように描く。
 - ヒストグラムの縦軸を密度にすると安全である。
 - $\text{密度} = \text{相対度数} \div \text{階級幅}$
 - 開放間隔の階級の存在
 - 開放間隔：下限または上限が明示されていないこと。
 - 自分で想定してヒストグラムを作成する。

ヒストグラムの作成(2)

- 開放間隔の階級についての想定
 - » 第1階級の下限: 0m^2
 - » 第8階級の上限: 1200m^2
- 大胆な仮定を置くことになるので、明示する必要あり。
- 算術平均等の計算でも、第1階級の下限は 0m^2 、第8階級の上限: 1200m^2 とする。

ヒストグラムの作成(3)



資料: 総務省統計局『第67回日本統計年鑑2018』表21-13

算術平均の近似計算(1)

- 階級内一定値の仮定
 - ある階級に属する個体は、すべて階級値に等しい値を持つ。
 - 例：第2階級について
 - 階級値： $(50+75)/2 = 62.5$
 - 度数：367(千戸)
 - 第2階級に属する個体については、すべて $x_i = 62.5$ と考える。
 - その仮定のもとで：
 - » 第2階級に属する個体の x_i の合計：
 $367(\text{千戸}) \times 62.5 = 22937.5 (\times 1000)$

算術平均の近似計算(2)

仮定のもとで：

$$\begin{aligned}\bar{x} &\approx \frac{1}{1807 \times 1000} \{ 216 \times 1000 \times 25.0 + 367 \times 1000 \times 62.5 \\ &\quad + 311 \times 1000 \times 87.5 + 375 \times 1000 \times 125.0 \\ &\quad + 233 \times 1000 \times 175.0 + 177 \times 1000 \times 250.0 \\ &\quad + 90 \times 1000 \times 400 + 38 \times 1000 \times 850 \} \\ &\approx 141.5\end{aligned}$$

算術平均の近似計算(3)

階級値の加重平均：

$$\begin{aligned}\bar{x} &\approx \frac{216}{1807} \times 25.0 + \frac{367}{1807} \times 62.5 + \frac{311}{1807} \times 87.5 \\ &\quad + \frac{375}{1807} \times 125.0 + \frac{233}{1807} \times 175.0 + \frac{177}{1807} \times 250.0 \\ &\quad + \frac{90}{1807} \times 400 + \frac{38}{1807} \times 850 \\ &\approx 141.5\end{aligned}$$

分散・標準偏差の近似計算(1)

- 階級内一定値の仮定
 - すべての個体の x の値が定まる。
 - 同じ仮定のもとで算術平均を近似する。
 - 計算済み。
 - 分散の公式どおりに計算して近似する。
 - 平均からの偏差の2乗の平均値。
 - 標準偏差＝分散の平方根

分散・標準偏差の近似計算(2)

$$\begin{aligned}s^2 &\approx \frac{1}{1807} \left\{ 216 \times (25.0 - 141.5)^2 + 367 \times (62.5 - 141.5)^2 \right. \\ &\quad + 311 \times (87.5 - 141.5)^2 + 375 \times (125.0 - 141.5)^2 \\ &\quad + 233 \times (175.0 - 141.5)^2 + 177 \times (250.0 - 141.5)^2 \\ &\quad \left. + 90 \times (400.0 - 141.5)^2 + 38 \times (850.0 - 141.5)^2 \right\} \\ &\approx 18630\end{aligned}$$

$$s \approx 136$$

中央値の近似計算(1)

- 階級内一様分布の仮定
 - ある階級の中に属する個体が、階級幅全体に一様に分布している。
 - たとえば、第2階級について：
 - 下限: 50、上限: 75、階級幅: 25 ($=75-50$)
 - 度数: 367 (千戸)
 - » 下限($=50$) から、階級幅の $1/5$ まで ($50+25/5 = 55$) の幅に、度数の $1/5$ ($=367/5$ 千戸) が分布している。
 - » 下限から階級幅の $2/5$ まで ($=60$) の幅に、度数の $2/5$ ($=367 \times 2/5$) が分布している、等々。

中央値の近似計算(2)

表2: 敷地面積別一戸建住宅数(借家、2013年、累積相対度数入り)

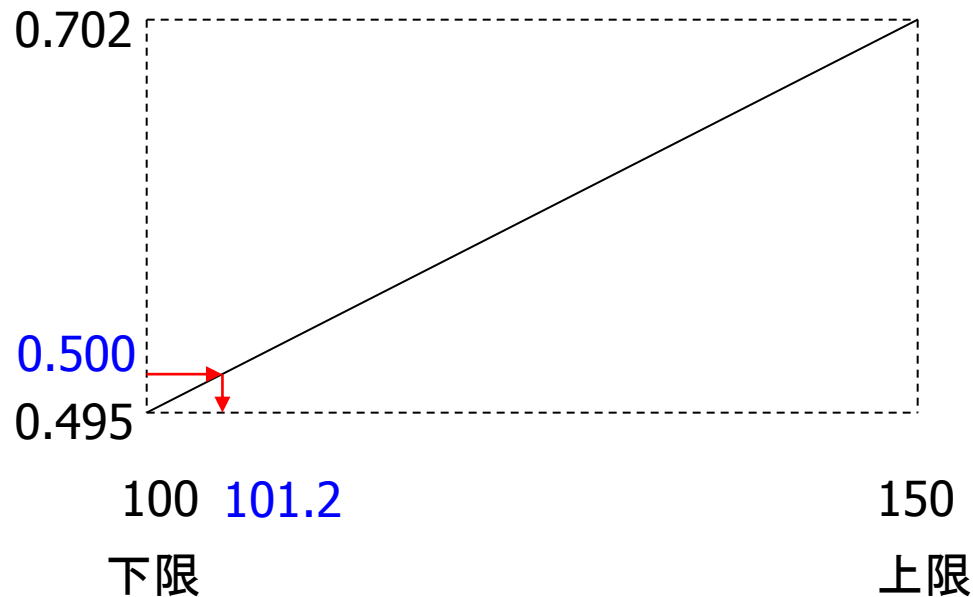
階級	下限(m^2 以上)	上限(m^2 未満)	度数(1000戸)	累積相対度数
1	-	50	216	0.120
2	50	75	367	0.323
3	75	100	311	0.495
4	100	150	375	 0.702
5	150	200	233	0.831
6	200	300	177	0.929
7	300	500	90	0.979
8	500	-	38	1.000
		合計	1,807	

資料: 総務省統計局『第67回日本統計年鑑2018』表21-13

中央値の近似計算(3)

図2: 第4階級における近似的な中央値の位置

累積相対度数 $100 + (150-100) \times (0.5 - 0.495) / (0.702 - 0.495) = 101.2$



四分位範囲の近似計算

- 階級内一様分布の仮定
 - 四分位点も中央値と同じように計算する。
 - 第1四分位点 Q_1 :
 - 表2から、第2階級にふくまれる。
 - » $Q_1 = 50 + (75 - 50) \times (0.25 - 0.120) / (0.323 - 0.120) = 66.0$
 - 第3四分位点 Q_3 :
 - 表2から、第5階級にふくまれる。
 - » $Q_3 = 150 + (200 - 150) \times (0.75 - 0.702) / (0.831 - 0.702) = 168.6$
 - 四分位範囲
 - $Q_3 - Q_1 = 168.6 - 66.0 = 102.6$

ローレンツ曲線の作成 (1)

- 階級内一定値の仮定
 - ある階級に属する個体は、すべて階級値に等しい値を持つ。
 - その階級の中では、個体数の相対的な累積のスピードと、変数の相対的な累積量のスピードとが同じになる。
 - その階級の中では、ローレンツ曲線が直線になる。

ローレンツ曲線の作成 (2)

- ローレンツ曲線の作成手順
 - 横軸: 階級 j の累積相対度数
 - 縦軸: 階級 j の変数の相対的な累積量
 - 計算方法:
 - 度数 $\times$ 階級値で各階級の変数合計を近似する。
 - » 算術平均の近似と同じ計算
 - 近似値を累積する。
 - 累積値を合計で除して相対的な累積量にする。

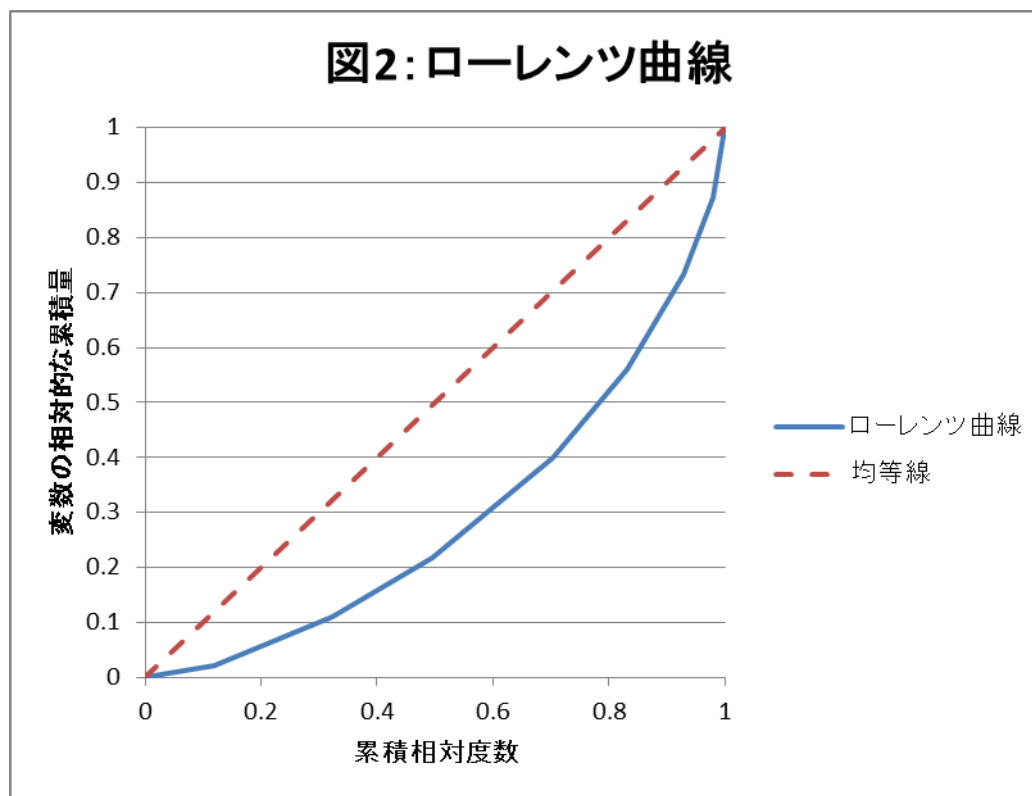
ローレンツ曲線の作成 (3)

表3: 変数の相対的な累積量の計算

階級	下限 (m2以上)	上限 (m2未満)	度数 (1000戸)	累積相対 度数	階級値	変数合計	変数の相 対的な累 積量
				0			0
1	0	50	216	0.120	25	5400	0.021
2	50	75	367	0.323	62.5	22937.5	0.111
3	75	100	311	0.495	87.5	27212.5	0.217
4	100	150	375	0.702	125	46875	0.400
5	150	200	233	0.831	175	40775	0.560
6	200	300	177	0.929	250	44250	0.733
7	300	500	90	0.979	400	36000	0.874
8	500	1200	38	1.000	850	32300	1.000
		合計	1807			255750	

資料: 総務省統計局『第67回日本統計年鑑2018』表21-13

ローレンツ曲線の作成 (4)



資料: 総務省統計局『第67回日本統計年鑑2018』表21-13

ジニ係数の計算 (1)

- ローレンツ曲線から、通常と同じように計算する。
 - 台形の面積の公式を利用して、ローレンツ曲線の下側の面積を計算する。
 - その2倍を1から減じる。
 - 借家の一戸建敷地面積については、
 - $GI = 0.42$

ジニ係数の計算 (2)

- 度数分布表からの近似計算の方が、もとのデータから計算したものよりも、不均等度が低くなりやすい。
 - ローレンツ曲線は均等線に若干近くなりやすい。
 - ジニ係数は0に若干近くなりやすい。
 - 階級内一定値の仮定のため。
 - つまり、階級内の不均等が無視されるため。

PC実習

- 度数分布表・ヒストグラムの作成
- 算術平均・分散の近似計算
- 中央値・四分位範囲の近似計算
- ローレンツ曲線の近似、ジニ係数の近似計算