第二回：1次元のデータ

尚　晋
大学院経済学研究科　助教

2025年4月22日

今日のポイント

1. 度数分布とヒストグラム
2. 代表値
3. 散らばりの尺度

1　度数分布とヒストグラム

1.1　記述統計学に関して

- 正しくしかも効率的に読むためには記述統計学の方法が必要。
- **記述統計学**とは，集団としての特徴を記述するために，観測対象となった各個体について観測し，得られたデータを整理・要約する方法である.
- **観測**とは，広く調査や実験のこと.
- 各「個体」(人，もの)の観測値をまとめたものを**データ(data)**という.

1.2　度数分布

調査や実験によって観測値が得られたとき，最初に度数分布表を作ることから始める.計算するよりも，表や図にする方が全体の分布の状況が明らかになるからである.例えば、ある大学における統計学の試験の受験者数373人の成績を度数分布表にしたのが，表2.1である.

- **度数分布表**は，観測値のとりうる値をいくつかの**階級**に分け，それぞれの階級で観測値がいくつあるか**度数**を数えて，表にしたものである.
- **階級値**とは階級を代表する値のことであって，階級の上限値と下限値の中間値を階級値とするのが普通である.
- 度数/観測値の総数を**相対度数**という.
- **累積度数,累積相対度数**とは，度数を下の階級から順に積み上げたときの度数，相対度数の累積和である.

表2.1 試験得点の度数分布表(某大学の統計学)

階	級	階級値	度数	相対度数	累積度数	累積相対度数
0点以上	10点未満	5	12	0.032	12	0.032
10"	20"	15	10	0.027	22	0.059
20"	30"	25	19	0.051	41	0.110
30"	40"	35	42	0.113	83	0.223
40"	50"	45	72	0.193	155	0.416
50"	60"	55	82	0.220	237	0.635
60"	70"	65	54	0.145	291	0.780
70"	80"	75	38	0.102	329	0.882
80"	90"	85	25	0.067	354	0.949
90"	100点以下	95	19	0.051	373	1.000
合	計		373	1.000		

1.3　ヒストグラム

- 横軸に観測値のとりうる値，各階級に対して階級幅を横幅とし，柱の面積が各階級の(相対)度数と比例するように高さを定める，このようなグラフを**ヒストグラム(柱状グラフ)**という.

図2.1 試験得点のヒストグラム

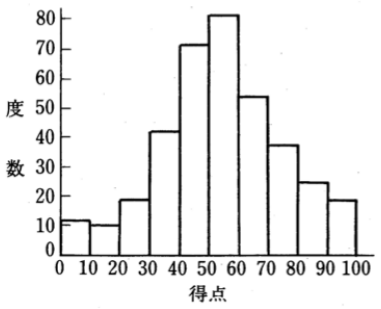


図 2.1　試験得点のヒストグラム
度数分布がこのように整った形となることはむしろめずらしい．階級数，階級幅などをうまくとらねばならない．

試験の成績の分布は，図に示されているように，中央に一つ峰がある山型分布である.しかし，このように左右対称の山型分布にならないものも多くある.そのうち峰が中央から左側に寄っていて，右側に長く裾を引く分布のことを，(感覚とは逆になるが)右に歪んだ分布という．図2.2の従業者規模による事業所数の分布のように，峰が左端に寄り，右に長く尾をひいた分布となる.

補足 1.1 度数分布表やヒストグラムを作成するときに注意すべき点は，階級数の問題と階級幅の問題である.

階級数に関しては，少なすぎても多すぎてもデータの意味するところ(真の分布)が失われる．表 2.1に示した試験の得点の度数分布では10点きざみで10の階級を設定しているが，同じデータに対して，0点以上20点未満というように20点きざみで5階級，および0点以上5点未満というように5点きざみで20階級に分けて，ヒストグラムを作成したものが，それぞれ図2.6, 図2.7である.

きざみが粗すぎるため，きわめてありふれた形となり，真の分布を見出しえない.

また階級幅を小さくとり，階級数を増やすと，分布が階級のとり方に敏感になる.(図2.8と図2.7)

階級をどのようにとるかを決める統一的ルールはない.ただし，階級数に関してはス

図2.2 従業者規模別事業所件数(全国・1986年)

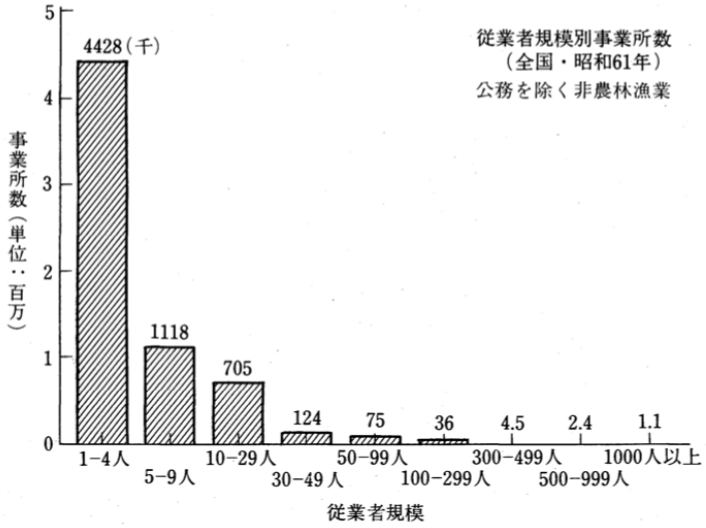


図 2.2　従業者規模別事業所件数(全国・1986 年)
峰が左端に寄り，右に長く尾をひいた分布(右に歪んだ分布)の例である．この図のように，階級幅が各階級で著しく異なる場合には，柱を分離して描く．
(出典：総務庁統計局「事業所統計調査報告」)

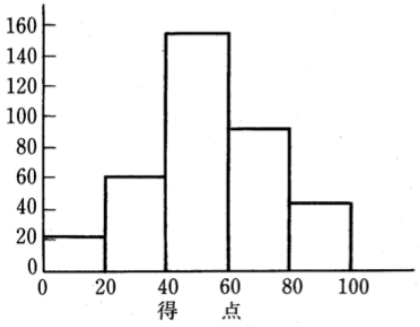


図 2.6　試験得点のヒストグラム
(階級数が少ない場合)

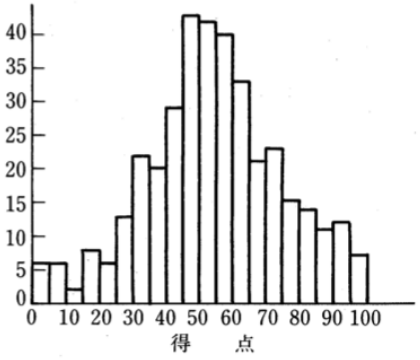


図 2.7　試験得点のヒストグラム
(階級数が多い場合：その 1)

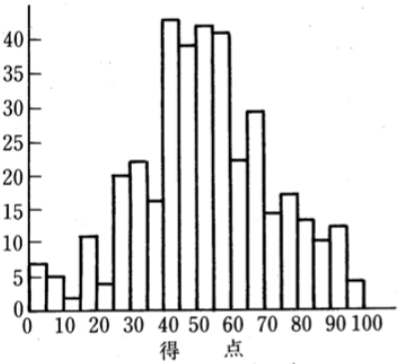


図 2.8　試験得点のヒストグラム
(階級数が多い場合：その 2)

図 2.1 を含むこれら四つは，全て同一のデータからである．度数分布の多様性とともに，'難しさ' もわかるであろう．

タージェスの公式が参考になる.観測値の数をnとし，階級数Kとした時の式は：

$k \fallingdotseq 1 + log_2 n = 1 + (log_{10} n)/(log_{10} 2)$
試験の得点の分布にこの式を適用すると，k=9.543...となり，k=10ととればよい.
階級幅に厳密な決まりはないが、通常は等しい幅が望ましい。ただし、分布の端で度数が極端に少ない場合は、階級幅を広げることがある。

度数分布表からはヒストグラムの他に，累積度数や累積相対度数をもとにしたグラフをつくることもできる.試験の得点の分布に対して累積相対度数のグラフを作成したものが図2.11である。

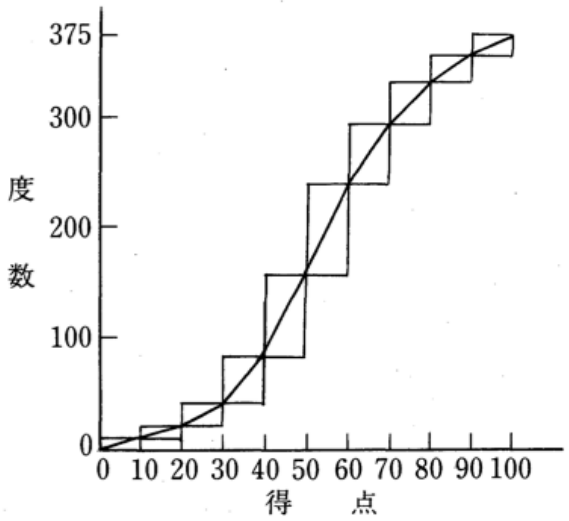


図 2.11　ヒストグラムと累積度数グラフ

累積相対度数のグラフでは、異なる2つのデータ（例：表2.3の事業所数と従業者数）を組み合わせて表すことができる。例えば、横軸に事業所数、縦軸に従業者数の累積相対度数を取り、各点を線で結ぶと図2.13のようなグラフが作成できる。これはローレンツ曲線であり、事業所数の最初の何％に従業者数の何％が含まれるかを示すグラフである。全ての事業所に同じ従業者数がいれば、線は対角線になる。対角線からのずれが大きいほど、規模の不均等が大きい。ローレンツ曲線は、所得や資産の不平等を示すのにも使われる。

補足 1.2 測定の尺度:

- 名義(名目)尺度:ある個体(対象)が他とは異なるか同一か。例：性別‘男”女’。
- 順序尺度:ある個体が他より‘大きい’, 他より‘良い’, 他より (何かについて)‘多い’といえる判断の基準。例：住みやすいを‘非常によい”よい”中程度”悪い”非常に悪い’など.
- 間隔尺度：ある個体は他よりもある単位によってどれだけ多い(少ない)といえる判断の基準.例：温度，時刻など。
- 比尺度：ある個体は他よりある単位によって何倍だけ多い(少ない)といえる判断の基準.例：身長(長さ)，体重(重さ)。

2　記述統計量

2.1　位置の尺度

定義：算術平均（観測値の総和）/（観測値の総数)、即ち観測値の総和を観測値の総数で割ったもの。

表 2.3 従業者規模別事業所数および従業者数
(全国・1986 年・公務を除く非農林漁業)

従業者規模	事業所数 (千件)	累積相対度数	従業者数 (千人)	累積相対度数
1- 4人	4,428	0.682	9,486	0.194
5- 9人	1,118	0.854	7,214	0.341
10- 29人	705	0.963	11,134	0.568
30- 49人	124	0.982	4,648	0.663
50- 99人	75	0.993	5,103	0.767
100-299人	36	0.999	5,734	0.884
300-499人	4.5	0.999	1,706	0.919
500-999人	2.4	1.000	1,651	0.953
1000人以上	1.1	1.000	2,320	1.000
合 計	6,494		48,995	

(出典:総務庁統計局「事業所統計調査報告」)

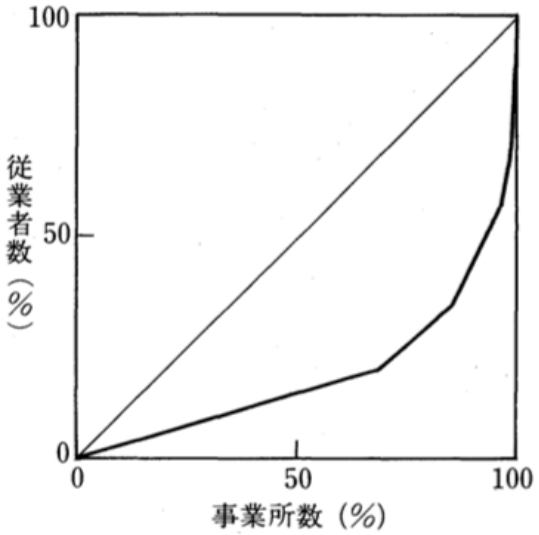


図 2.13 事業所規模のローレンツ曲線
(全国・昭和 61 年)

補足 2.1 他の平均:幾何平均，調和平均

- 幾何平均：正数$x_1, x_2, ..., x_n$の幾何平均は$x_G = \sqrt[n]{x_1 \cdot x_2 \cdot ... \cdot x_n}$.例：1983年から88年までの5年間にはそれぞれ21.8%, 30.5%, 53.6%, 50.0%, 12.9%上昇した．この間の年平均上昇率は算術平均33.8%ではなくて，幾何平均を用いなければならない.
- 調和平均:$\frac{1}{x_H} = \frac{1}{n}(\frac{1}{x_1}+...+\frac{1}{x_2})$.例：ある路線バスが行きは時速25km, 帰りは時速15kmで往復したとすると，その平均時速は20kmではない．路線の距離をdとすると，往復距離$2d$を時間で割って平均時速vは調和平均を用いなければならない。

定義：中位数(中央値或いはメディアン)

観測値を小さい方から順に並べたときの中央の値。

対称な分布なら平均＝中位数。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．データの大きさが奇数n=2m+lのときにはm+l番目の観測値X_{m+1}であるが，偶数n=2mの場合は中央が一つに決まらないので，m番目とm+l番目の観測値の平均をとる。

補足 2.2 メディアンの考え方を拡張したものしたものに**分位点**がある。

- 観測値を小さいものの順に並びかえたとき，小さい方から$100p\%(0 \leqq p \leqq l)$の所にある値を100pパーセンタイルまたは分位点という．
- よく用いられる分位点には**四分位点**がある:第1四分位点Q_1は25%分位点，第2四分位点Q_2は50%分位点(メディアン),第3四分位点Q_3は75%分位点

定義：モード(最頻値)

その度数が最大である階級の階級値がモードとなる。

補足 2.3 分布の代表値としては平均，メディアン，モードの三つが有名である．平均，メディアン，モードの関係に関していうと，峰が一つある単峰性の分布で，分布が完全に左右対称の場合この三つは完全に一致する．

2.2 散らばりの尺度

大きさが同じ$n = 10$の以下3つのデータA,B,Cがあったとする．これらの平均，メディアン，モードはいずれも5である．

A:	0	3	3	5	5	5	5	7	7	10
B:	0	1	2	3	5	5	7	8	9	10
C:	3	4	4	5	5	5	5	6	6	7

このA, B, C3つの分布は，なめらかな分布を想定すると図2.17のような関係になっている．前節で説明した代表値は分布の位置を示す指標であって，この例のようなA, B, C3つの分布を区別するには，分布の形状を示す他の指標が必要となることがわかる．

分布の形状を示す指標は多くある．**散らばりの尺度**と呼ばれるものはよく使われる。位置の指標と散らばりの尺度の二つを用いれば，まずは分布のおおまかな形状を記述することができる．

定義：レンジ(範囲)

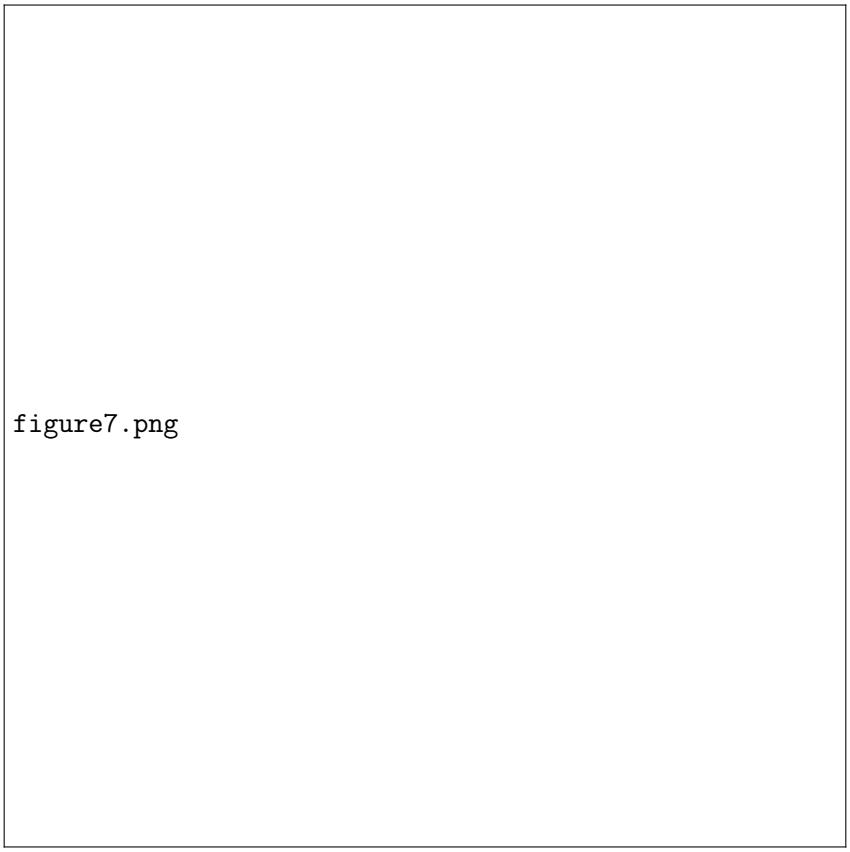
もっとも単純なものはレンジまたは範囲と呼ばれ，分布の存在する範囲を示すもの。最大値と最小値の差と計算する。かなり粗いものである．あまり使わない。

定義：四分位偏差

データの第3四分位点Q_3と第1四分位点Q_1の隔たりの半分として定義され，真中の半分のデータが散らばっている範囲の平均を表す．四分位偏差が大きいほど散らばった分布となる．

補足 2.4 レンジも四分位偏差も，与えられた観測値の散らばり具合を表現するのに，たかだか2個ないし4個の観測値を用いるだけであり，すべての観測値を用いていない．平均偏差，標準偏差はすべての観測値のもつ情報を利用した散らばりの尺度，いずれも各観測値x_iと平均$\bar{x}$との隔たり（偏差という）をもとに計算される．

定義：偏差



各観測値x_iと平均$\bar{x}$との隔たり。

定義：平均偏差
各観測値が平均からどれだけ離れているかについての平均を求めたもの。平均偏差の計算式： $d = \frac{1}{n}\{|x_1 - \bar{x}| + |x_2 - \bar{x}| + ... + |x_n - \bar{x}|\}$

定義：分散
平均からの偏差の2乗の平均. 分散はS^2という記号で表され，分散の計算式：

$$S^2 = \frac{1}{n}\{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + ... + (x_n - \bar{x})^2\} = \frac{1}{n}\Sigma_{i=1}^{n}(x_i - \bar{x})^2$$

定義：標準偏差
分散の平方根.

補足 2.5 平均偏差も標準偏差も，ともに分布の散らばりの程度を示す指標である．しかし現実には標準偏差が用いられる場合がほとんどであって，平均偏差が用いられることはほとんどない．それは，分散や標準偏差の方が理論的に扱いやすくすぐれた性質をもっているからである．

定義：変動係数
（標準偏差）/（平均）.変動係数はよくC.V.という記号で表され、計算式は：$C.V. = \frac{S_x}{\bar{x}}$

補足 2.6 変動係数の例：

- 1965年には1人あたり県民所得の平均は$\bar{x} = 26.6$万円，標準偏差は$S = 7.5$万円であ

ったのが，1975年には平均$\bar{x} = 117.5$万円，標準偏差$S = 23.8$万円になっている．地域間の所得格差は大きくなっているのであろうか．

- 単純に標準偏差を比較すれば約3倍になっており，大きくなっている．しかしその間に平均も約4.5倍に増えている．この例のように直接の比較が困難な場合に，平均えを考慮した上で散らばり具合を相対的に比較するのに便利な指標である．この例では1965年の変動係数は $7.5/26.6 = 0.28(28\%)$,1975年は$23.8/117.5 = 0.20(20\%)$であり，安定している中にも相対的な地域間所得格差はむしろ小さくなっている．

定義：標準得点(標準化)

$z_i = \frac{x_i - \bar{x}}{S_x}$

平均を差し引き，標準偏差で割って，位置，尺度の調整をした結果，このzをデータxの標準化とか，標準得点といぅ．

今日のキーワード

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，ローレンツ曲線，（算術）平均，中位数(メディアン)，分位点，最頻値(モード)，偏差，分散，標準偏差，変動係数