2007 年度後期　統計データ解析Ｂ　第２回

データを「分布」で見る － 分布

「分布」とは

（測定対象や現象が）分布するとは，ある測定対象や現象から得られる数量が大小ばらばらである，という意味です．例えば，「イチロー選手が 1 試合に打つヒットの数」や「日本男性の身長」は分布します．現実の調査対象についてデータを集めると，そのデータは分布しているほうが自然です[1]．

データが大小ばらばらであるならば，それが「どう」ばらばらかを知ることが，分布を知ることにつながります．すなわち，データの分布を数量的に表現するとは，分布しているデータのどんな値がどのくらい頻繁に現れるか，をとらえることになります．例えば，「イチロー選手が 1 試合に打つヒットの数」で言えば，ヒットの数が 0 本である試合が何試合，1 本である試合が何試合，. . . というように分布を表現することができます．このように，どのくらい頻繁に現れるかを表す量を度数といいます．また，度数を「何試合」と数えるのではなく，全体の試合数に対する割合で「何%」と表すほうが，試合数の違ういろいろな分布を比較するのに便利です．このように%の単位で表した度数を，とくに相対度数といいます．このようにして，度数を使って表現された分布を度数分布といいます．

一方，「日本男性の身長」のようなデータの場合は，身長は「測る」もので，ヒットの数のように「0 本，1 本，. . .」と「数える」ことはできません．そこで，「. . .，160cm 以上 165cm 未満の人が何%，165cm 以上 170cm 未満の人が何%，. . .」のように，数量をある間隔をもつ段階に区切って，各段階に入る数量がどのくらい頻繁に現れるかで分布を表現します．この段階を階級といい，ひとつの階級に入る値の範囲を階級幅といいます．

このとき，「169.4cm 以上 169.5cm 未満の人が何%，169.5cm 以上 169.6cm 未満の人が何%，. . .」などとあまりに細かい話をしても，分布の特徴を把握することはできませんから，適当な間隔の階級を用いる必要があります．

度数分布を作ってみましょう

データから度数分布を作ってみましょう．下の数字は，あるクラス 50 名の試験の得点です．

階級幅の取り方を 10 点として，度数分布表を作って表に書き込んで行きます．「95 点」のデータは 85 点以上 95 点未満の階級に入れます．こういう場合，度数を数えるには，「正」の字を書く，４本の縦棒に 1 本の横棒を重ねる，などの，５ごとにまとめて数える方法がよく用いられます．

35	62	65	23	40	30	70	55	57	65	15	90	67	65	70	45	80
79	46	45	25	50	62	75	78	48	50	60	75	75	60	78	58	78
63	95	20	46	55	56	70	60	79	18	63	67	85	25	40	50	

度数分布表は表 1 のようになります．表の左から 3 列目に階級値というのがあります．これは，各階級の上限下限の中間の値で，その階級に入ったデータ（すなわち試験の得点）は，どれも概略この値であると考えます．

[1] ある党の得票率が 100%であるような選挙は，不自然でしょう．

以上	未満	階級値	度数	相対度数
15	25	20	4	0.08 (8%)
25	35	30	3	0.06 (6%)
35	45	40	3	0.06 (6%)
45	55	50	8	0.16 (16%)
55	65	60	12	0.24 (24%)
65	75	70	8	0.16 (16%)
75	85	80	9	0.18 (16%)
85	95	90	3	0.06 (6%)
x	x	x	計 50	計 1 (100%)

表 1: 度数分布表

ヒストグラム

度数分布を目に見えるようにするために，横軸に階級，縦軸に度数（相対度数）をとり，階級幅を底辺，度数を面積とする柱で各階級の度数を表したグラフを，ヒストグラムといいます．

ヒストグラムは，図 1 のような棒グラフとは違い，図 2 のように柱の間隔を開けずに描きます．

このように，柱の間隔を開けず，また柱の「面積」で度数を表現するのは，階級の区切りかたを自由に変更できるようにするためです．ヒストグラムの横軸は本来連続した値を表しているものであり，柱どうしが分れているのは，連続した値を階級に分割したからです．分割のしかたは自由ですから，ヒストグラムでの階級の区切りかたも自由に変更できるはずです．柱の面積で度数を表現しておけば，柱を分割・結合することで，階級を変更することができます．

図 3 のように，例えば「となりあう 2 つの階級の度数の合計」は，となりあう 2 つの柱の面積の合計となります．同様に，「100〜120 の階級の度数が 10」ということを，「100〜110 の階級の度数が 5，110〜120 の階級の度数が 5」と分割して考えることもできます．

また，階級の幅が途中で違っていると高さと度数は一致せず，同じ度数でも階級の幅が 2 倍ならば高さは半分になります．このように階級の幅が途中で違っている度数分布は，階級幅を一定にすると度数が極端に違ってしまう場合，同じ階級幅でも階級値によって意味が大きく違う場合に用いられます[2]．

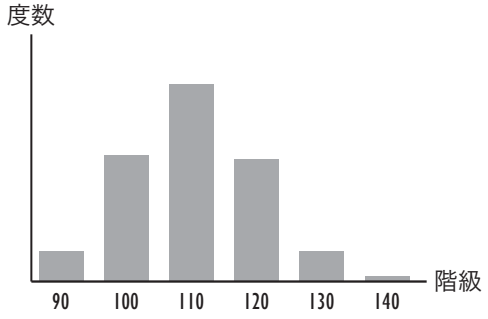

図 1: ヒストグラムはこんなふうには描かない

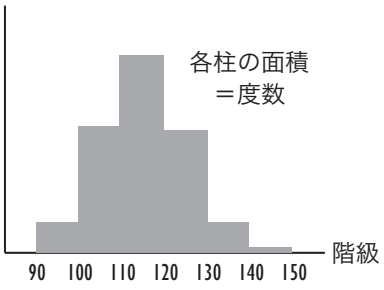

図 2: ヒストグラムはこう描く

[2]年収 300 万円と 400 万円は意味がかなり違いますが，年収 1 億円と 1 億 100 万円はあまり差がないでしょう．

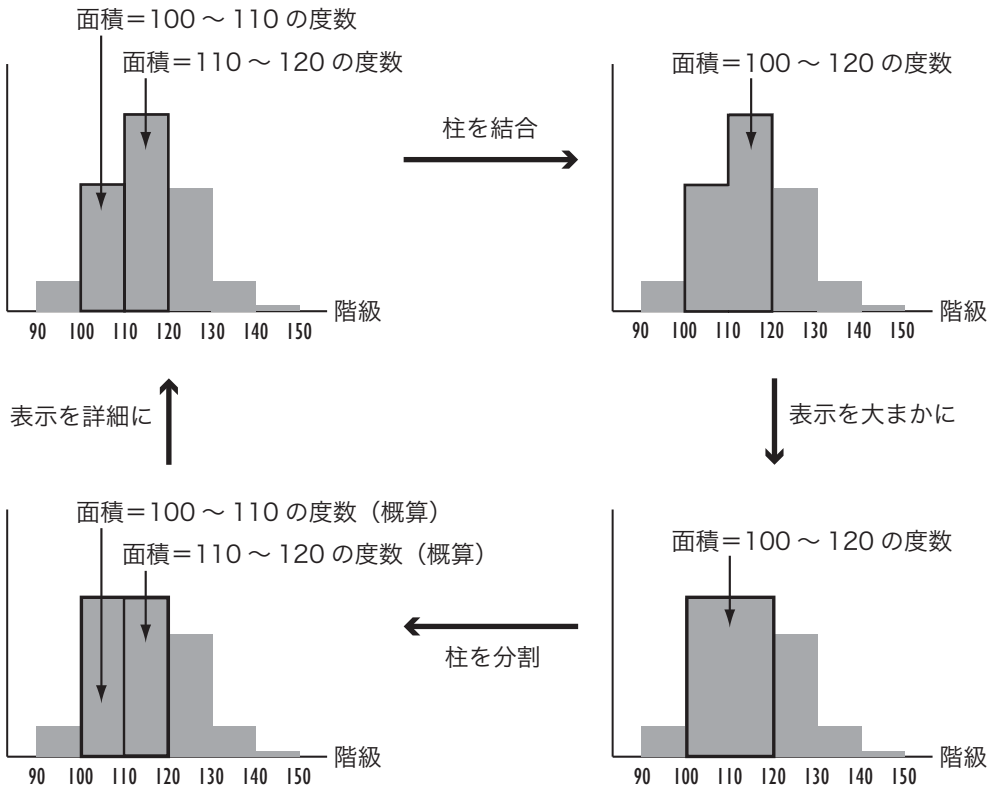

図 3: 柱の分割と結合

ボックスプロット

ヒストグラムをさらに簡略化して表現したのがボックスプロット（箱ひげ図）です．これは図 4 のように，最小値，第 1 （下側）四分位数，中位数（中央値，メディアン），第 3 （上側）四分位数，最大値だけをグラフの中に表示したものです．分布の形を簡単な図で概略つかむことができます．ここで，中位数とは，データを小さいほうから並べたときに順位が 50%（データが１００個のとき 50 位）であるもの，第 1 （第 3 ）四分位数はそれぞれ 25%，75% になるものをさします．

ボックスプロットを描くときに，最大値や最小値が他のデータから飛び離れている場合は，それを別扱いにして表現することもあります．これは，このような 1 つだけ飛び離れた値は，他のデータが分布している理由とは別の理由によって生じている場合があるからです．このような飛び離れた値を外れ値（**outlier**）といいます．外れ値がある場合，最大値・最小値は外れ値を除いたものを表示します．

ボックスプロットの利点は，図 5 のように複数のボックスプロットを並べたパラレルボックスプロットによって比較しながら見られるところです．ただし，ボックスプロットにはデータから抽出した量しか表示されておらず，データそのものは隠れてしまっているので注意が必要です．

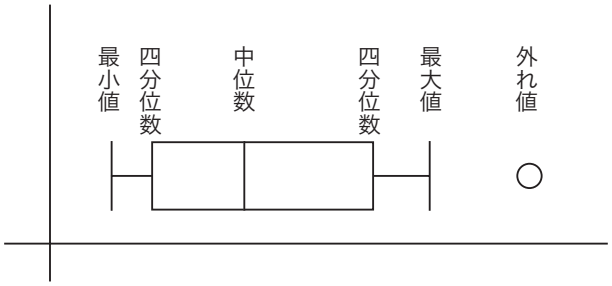

図 4: ボックスプロット

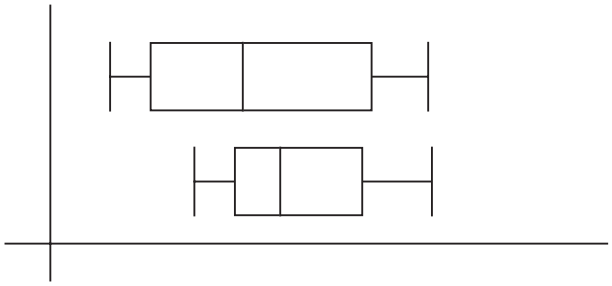

図 5: パラレルボックスプロット

今日の演習

(1) ヒストグラムについて，以下の設問に答えてください．

1. 本文中の度数分布表から，ヒストグラムを描いてください．
2. 「75 点以上 85 点未満」「85 点以上 95 点未満」の 2 つの階級を合わせて「75 点以上 95 点未満」とした場合のヒストグラムを描いてください．

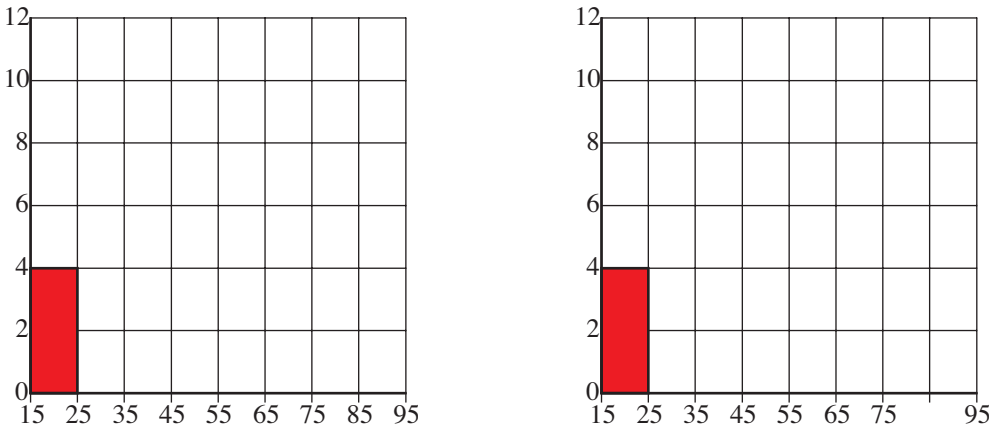

図 6: ヒストグラム．左・問 (1)1，右・問 (1)2

(2) 視覚的表現は直観的印象でデータを表現するため，説得力の高い表現を行うことができます．しかし，直観的印象はグラフの描き方によってかなり変化するため，グラフを読むほうはデータについての錯覚を起こさないように注意する必要があります．各データの大きさを棒の長さで表して比較する「棒グラフ」は，小学生の時から知っているおなじみのグラフですが，慣れているだけに，よく注意して見ないとだまされるおそれがあります．次の設問について答えてください．

1. 図 7 は，いずれも同じデータをグラフにしたものです．差が際立って見えるのはどれでしょうか．
2. 棒グラフの棒を，いろいろな形で表すと，親しみが持てるグラフにはなりますが，誤解を生みやすくもなります．図 8 の例は，棒グラフの棒を缶の形にしたものですが，このような描き方は正当でしょうか？

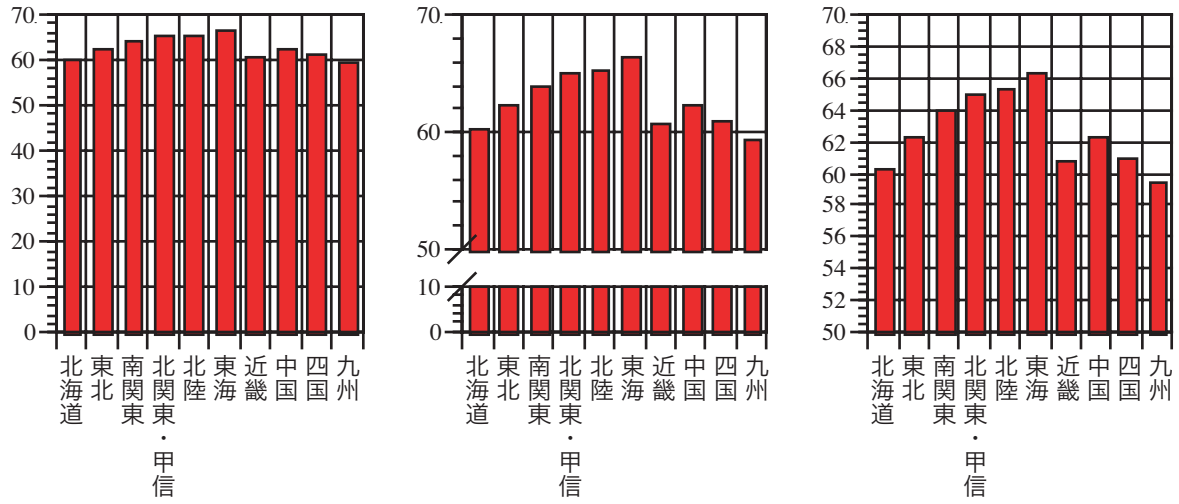

図 7: 棒グラフの例（平成 9 年就業構造基本調査より）

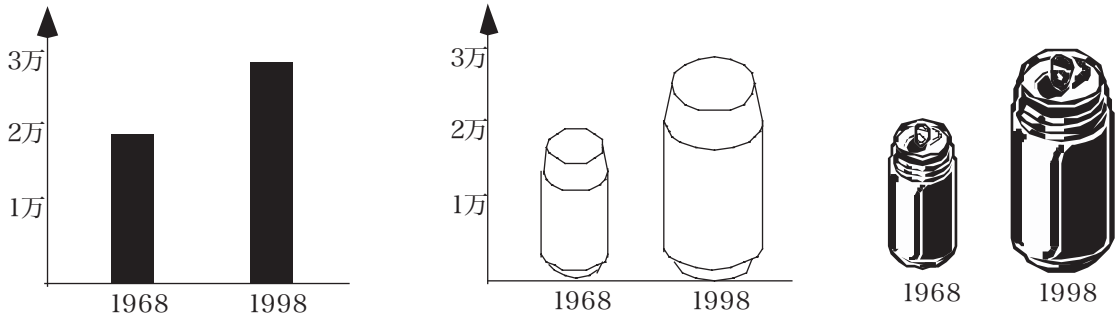

図 8: 怪しいグラフ（架空のデータ）