2022 年度秋学期　統計学　第４回

データを「分布」で見る

ここからは，尺度基準でいう質的データと量的データのうち，量的データを扱っていきます。統計学では，大小ばらばらな数値からなる量的データを扱い，そこから何らかの情報を取り出します。「大小ばらばら」な集団を扱うために必要なのが，「分布」という考え方と，分布を表現する方法です。今日は，その方法である「度数分布」と，それを図示するヒストグラムを説明します。

「分布」とは

「分布する」「分布」という言葉

量的データは，数値の集まりです[1]。量的データが**分布する**とは，ある測定対象や現象から得られる量的データが，大小ばらばらの数値で構成されている，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[2]。さらに，大小ばらばらの数値からなるデータを，そのばらつきのようすがわかる形で表したものを「分布」という名詞で表します。

度数，相対度数，度数分布

データが大小ばらばらの数値でできているならば，それが「どう」ばらばらかを知ることが，分布のようす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うちに**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，「ある野球選手が 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。

「数える」のでなく「測る」データの場合は

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。

---

[1] この講義では，「データ」とは数値の集まり（集合）をさすものとし，データに含まれる個々の数値は「数値」とよぶことにします。

[2] 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。

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」という数値は 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	

表 1: 度数分布表

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 (18%)
85	95	90	3	0.06 (6%)
x	x	x	計 50	計 1 (100%)

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。

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。

棒グラフと何が違うのか

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[3]。

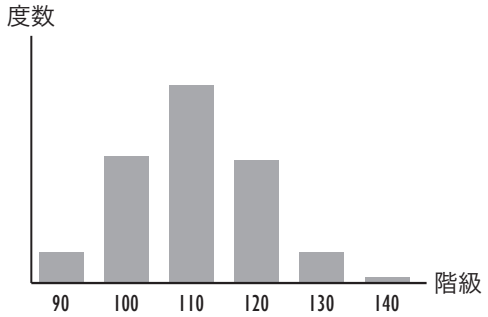

図 1: ヒストグラムはこんなふうには描かない

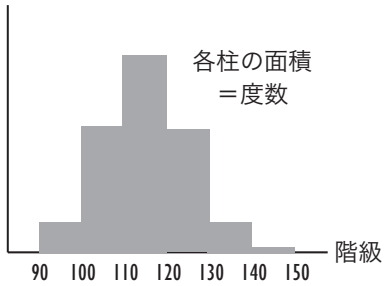

図 2: ヒストグラムはこう描く

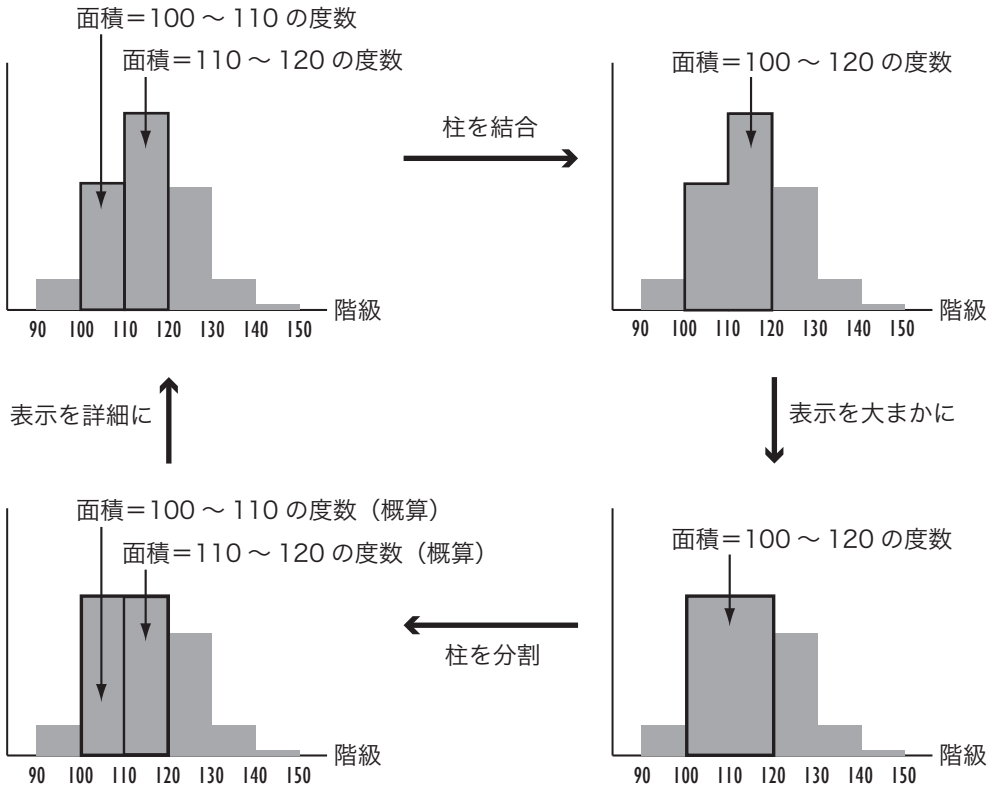

図 3: 柱の分割と結合

幹葉表示

先に，50 名の試験の得点から度数分布を作りました。このとき，「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」と

[3]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。

表 2: 幹葉表示を作ってみましょう

十の位	一の位
0	
1	
2	
3	5
4	
5	
6	2 5
7	
8	
9	

表 3: 幹葉表示（整理前）

0	
1	5 8
2	3 5 0 5
3	5 0
4	0 5 6 5 8 6 0
5	5 7 0 0 8 5 6 0
6	2 5 5 7 5 2 0 0 3 0 3 7
7	0 0 9 5 8 5 5 8 8 0 9
8	0 5
9	0 5

表 4: 幹葉表示

0	
1	5 8
2	0 3 5 5
3	0 5
4	0 0 5 5 6 6 8
5	0 0 0 5 5 6 7 8
6	0 0 0 2 2 3 3 5 5 5 7 7
7	0 0 0 5 5 5 8 8 8 9 9
8	0 5
9	0 5

書きました。

このような方法の他に，数値を書き並べる際に，数値の並べ方を工夫して，ヒストグラムのように直観的に頻度が見えるように表す方法があります。これは，数値のある一つの位に注目して，その位での数字によって 0 から 9 の 10 行に分けて，各行に，その行に入る数値を，一つ下の位の値を横に並べて書いていくものです。これを**幹葉表示**といいます。

試験の得点のデータを使って，幹葉表示を作ってみましょう。表 2 は，試験の得点の数値を十の位によって 10 行に分けたものです。最初の数値は「35」ですので，十の位が「3」の行に，「5」と書き込みます。次の 2 つは「62」と「65」ですから，十の位が「6」の行に「2」「5」と書きます。他の数値も，表 2 に書き込んでみてください。

書き込んだ結果は，表 3 となります。さらに，一の位の数字を大きさ順に並べ替えたのが表 4 となります。データをこのように幹葉表示にすると，簡便なヒストグラムのように見ることができます。

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

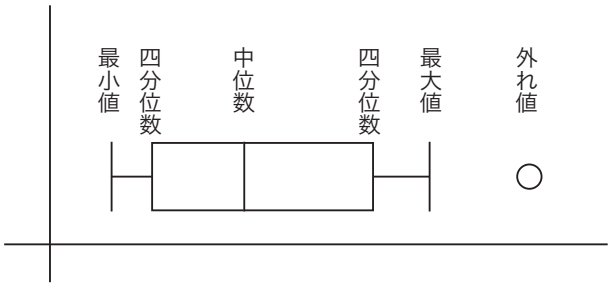

図 4: ボックスプロット

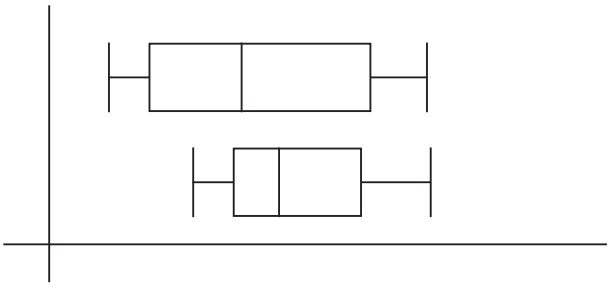

図 5: パラレルボックスプロット

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，データを小さい数値から順に並べて表したときに順位が 50% （データが 100 個の数値からなるとき 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。

「外れ値」

ボックスプロットを描くときに，最大値や最小値が他の数値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，他の数値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。

演習問題

以下の設問に答えてください。

1. 表 1 の度数分布表から，ヒストグラムを描いてください。
2. 「45 点以上 55 点未満」「55 点以上 65 点未満」の 2 つの階級を合わせて「45 点以上 65 点未満」とした場合のヒストグラムを描いてください。
3. 表 4 の幹葉表示をもとに，ボックスプロットを描いてください。