2024年度春学期

統計学

第4回
データを「分布」で見る

関西大学総合情報学部
浅野 晃

度数分布🤔

初学者には、「分布」という考え方が
むずかしいらしい。🤔

※ 鳥居泰彦氏の「はじめての統計学」という本に，最初にこう書いてあります。

量的データ

ここからは，量的データを扱う

量的データは，数値の集まり

データという言葉は，ひとつひとつの数字ではなく，
数値の集まり（集合）をさす

データに含まれる個々の数値は（ここでは）「数値」とよぶことにする

分布とは

- データが「分布している」「分布する」
  ＝ データが，大小ばらばらの数値が集まってできている

- 「分布」
  ＝ 大小ばらばらの数値が集まってできたデータ

統計学が相手にするのは[分布]です

※分布しないデータには，統計学の出番はありません。
全員が💯をとるテスト，有権者が全員統一〇〇党に入れる選挙…には，
統計学は必要ありません。

分布は，「コト」？「モノ」？

データが「分布している」「分布する」
この「分布」とは，分布している「コト」

「分布」とは，大小ばらばらの数値が集まってできたデータ
この「分布」とは，分布という「モノ」

※ 日本語は「コト」か「モノ」かを気にしますが，英語で「分布」にあたる”distribution”にはどちらの意味もあって，あまり気にしません。

※ 日本語としてはおかしいわけですが，こういう「翻訳調」なのは，明治時代の日本人が西洋の学問を翻訳したためです。

度数と度数分布

ばらばらなデータがどんなふうにばらばらか　　を表現する方法

どんな値がどのくらい頻繁に現れるか　を調べることで分布を表現する

「ある野球選手が1試合に打つヒットの数の分布」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，… と数えてみる

度数と度数分布

数えた結果（例）

どんな値（ヒットの数）が	ヒットの数	[度数] 何回（何試合）あったか 試合数	[相対度数] 50試合中の割合で表すと 割合
	0本	5試合	10%
	1本	15試合	30%
	2本	15試合	30%
	3本	10試合	20%
	4本	5試合	10%
		計 50試合	100%

[度数分布]

通常，度数よりも相対度数を用いることが多い

「測るデータ」の場合は？

「身長」のデータは，1本，2本と「数える」 のではなく「測る」

数値を「1本，2本」とランクにわけることはできない

※身長170cmの次は，171cm？170.5cm？170.1cm？

さて，
「どんな値がどれくらい頻繁に現れるか」，つまり「度数（相対度数）」は
どうやって表すのか？

「階級と階級値」を用いる

測ったデータを，ある間隔で段階に区切る　段階の間隔は自分で決める

	身長	相対度数
	…	
[階級]	160～165	15%
	165～170	20%
[階級幅]は5cm	170～175	20%
	175～180	10%
	…	

度数分布を作るには

階級を決めて，各階級に入る数値を数える

数値例

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	

※例えば，階級を「15以上25未満」「25以上35未満」…と決めて，
各階級に入る数値がいくつあるかを数える

手作業で数えるときには，

👈こんな記号や　「正」の字を書いて5ずつまとめて，
間違えないようにする

ちょっと余談ですが

統計学の教科書には，数え方の標準は  と書かれています

私は，日本伝統の「正」の字も，優れていると思います。
なぜならば，字画が正方形だから。

←急いで書くと間延びしやすい🌀

正 正 正

←「正」の字の幅が一定なので，
見た目で「正」が何文字あるか　すぐわかる

度数分布表はこうなります

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%)

[階級値]

その階級に入った数値は,
どれも概略この値であると考える

※ 例えば,「15以上25未満」の階級に入っていれば,
17という数値も, 20も22も, みんな「20である」
とみなす

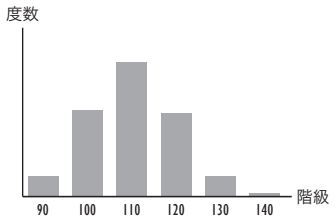
※ あとで「確率分布」が出てくるときに, この考えが重要です。

ヒストグラム🤔

ヒストグラム

度数分布を表すためのグラフ

~~横軸に階級, 縦軸に(相対)度数~~ ← そういいたいところですが, 違います
(次ページ)

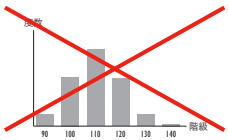


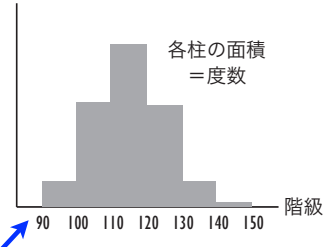
これは「棒グラフ」

ヒストグラムはこれとは違います

ヒストグラムは面積で度数を表す

横軸に階級をとり,
階級幅を底辺とする柱の 面 積 で(相対)度数を表す

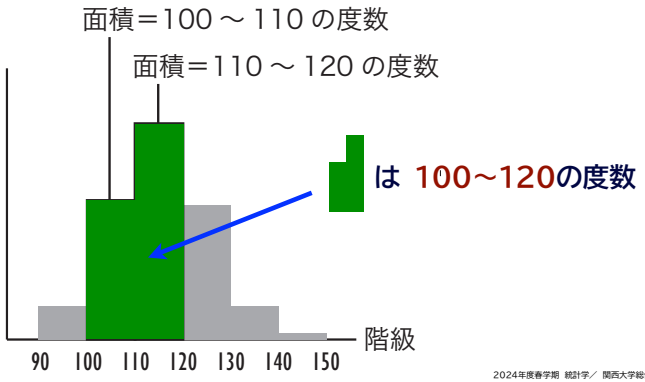




隣どうしの階級の間に「飛び」はないから, 柱の間には隙間はない

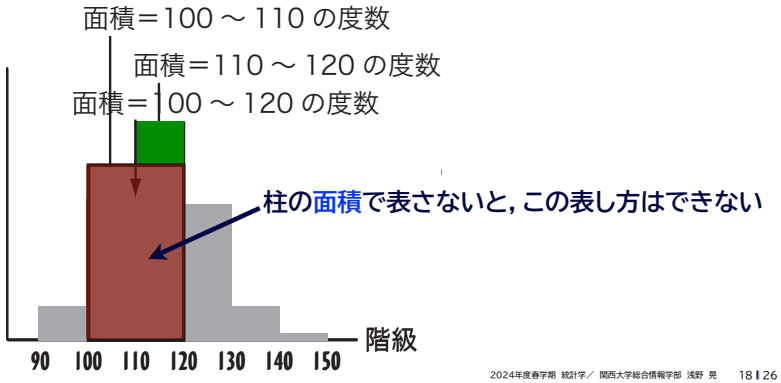
なぜ柱の面積で表すのか？

柱をくっつけたり切り分けたりするため



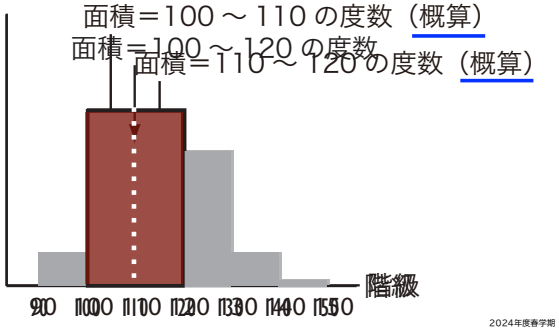
なぜ柱の面積で表すのか？

くっつけた柱を1つにしてもよい



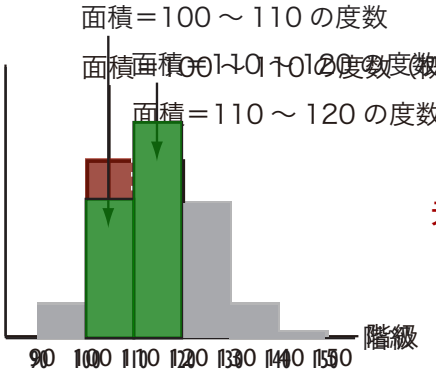
なぜ柱の面積で表すのか？

柱を分割してもよい



なぜ柱の面積で表すのか？

分割した柱を詳細に表すと



元にもどる

※この，柱を合体する／分割するという考え方も，あとで「確率分布」を学ぶときに出てきます。

階級幅が一定でないヒストグラム

階級幅が一定でないヒストグラムは，
実際のデータではごく普通

年収を考えると
500万円と600万円は「かなり異なる」が
1,000万円と1,100万円は「多少異なる」だし
1億円と1億100万円は「そんなに変わらない」なので

「階級幅がつねに100万円」のほうがおかしい

分布を可視化する他の方法🤔

幹葉(かんよう，みきは)表示

度数が目に見えるように，数値を並べる

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	

35という数値は
65という数値は
62という数値は

十の位	一の位	
0		
1		
2		
3	5	十の位が3　その行に「5」を置く
4		
5		
6	2 5	十の位が6　その行に「2」を置く 十の位が6　その行に「5」を置く
7		
8		
9		

幹葉表示

度数が見えるように，数値を並べる

データ全体を並べるとこうなる

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

一の位を昇順に(小さい数字から大きい数字に向かって)並べる

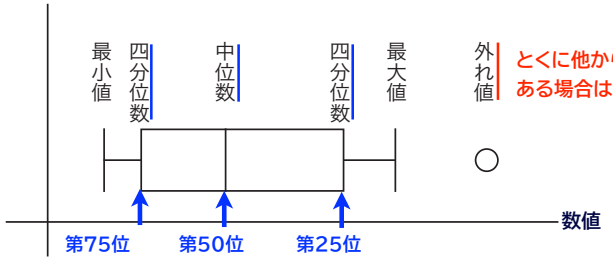
0	
1	5 8
2	0 3 5 5
3	0 5
4	0 0 5 5 6 6 8
5	0 0 0 5 5 6 7 8
6	0 0 0 2 2 3 3 5 5 5 7 7
7	0 0 0 5 5 5 8 8 8 9 9
8	0 5
9	0 5

数字の並びによって簡易ヒストグラムになり，数値も見える

ボックスプロット

ヒストグラムをさらに簡略にしたもの　順位にもとづく指標を用いて描く

100個の数値からなるデータがあるとして，大きさの順に並べ替えた時，順位が



パラレルボックスプロット

ボックスプロットは，複数個並べて比較することができる

（ヒストグラムにはできないこと）