

2020年度春学期 統計学 第4回

データを「分布」で見る

浅野 晃
関西大学総合情報学部



「※」がついている項目は、
通常の講義ではスライドに書かず、
口頭で話す部分だと思ってください。

度数分布 🤔

※今日から扱うのは、「データの分布」「分布するデータ」というものです。

※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。

初学者には、「分布」という考え方が むずかしいらしい。 🤔

量的データ

※前回の「尺度水準」で、データには、足し算ができる(平均を求めることができる)「量的データ」と、足し算ができない「質的データ」があることを説明しました。

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。

※ところで、しばらく前のWindowsでは「更新をアップロードします」という変なメッセージが出ましたが、これは”update”の直訳で、「更新用データ」という「モノ」も英語ではupdateです。

度数と度数分布

これから考えるのは、
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

それは、
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、
※毎試合必ずヒットを5本打てる選手はいませんから、ヒットの数は分布します。

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)

ヒットの数	[度数] 試合数	[相対度数] 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%

この表しかたで分布を表現したものが
[度数分布]
通常、度数よりも相対度数を用いることが多い

どんな値(ヒットの数)が
何回(何試合)あったか
50試合中の割合で表すと

「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ができない

※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
	...	
この「段階」を [階級] という	160~165	15%
	165~170	20%
この例では [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、

/// という記号や「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

※さらに余談ですが☺

「五円玉」にはどこにも「5」とは書いてないのを知っていましたか？
(財布から出して見てみましょう)

私は、外国の友人に指摘されて初めて気づきました。「五」は5本の画でできて
いるから「5」の意味だ、と説明しました。

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する

※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。

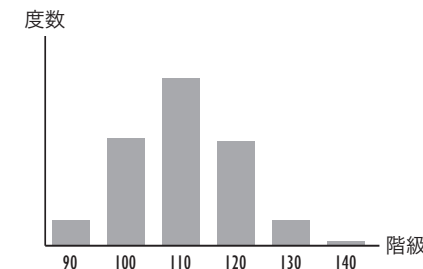
ヒストグラム🤔

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

↑
実は、微妙に違います。

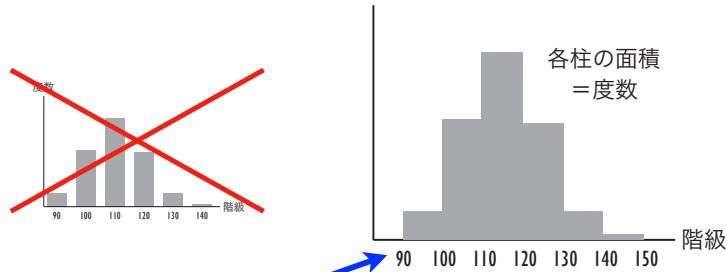


これは
「棒グラフ」

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

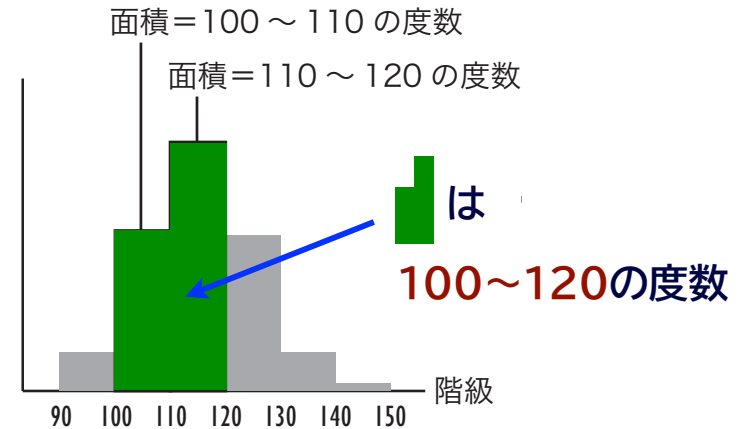


隣同士の階級の間「飛び」はないから、
柱の間には隙間はない

※面積を問題にするんだから、「柱」というより「板」というべきなんです、
なぜか昔から「柱」といわれています。

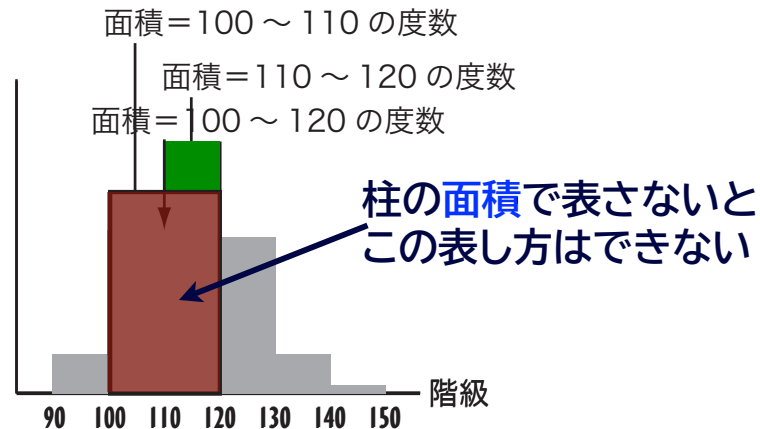
なぜ柱の面積で表すのか？

柱をくっつけたり切り分けたりするため



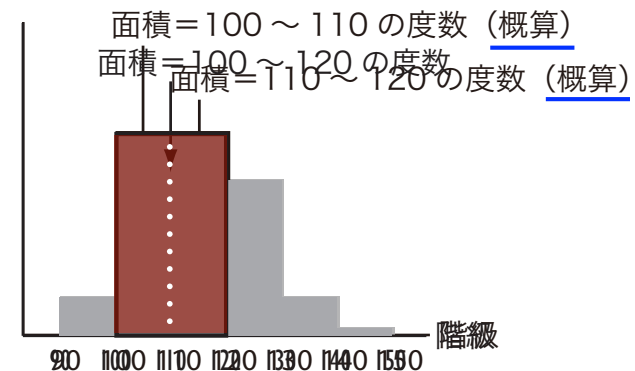
なぜ柱の面積で表すのか？

くっつけた柱を1つにしてもよい



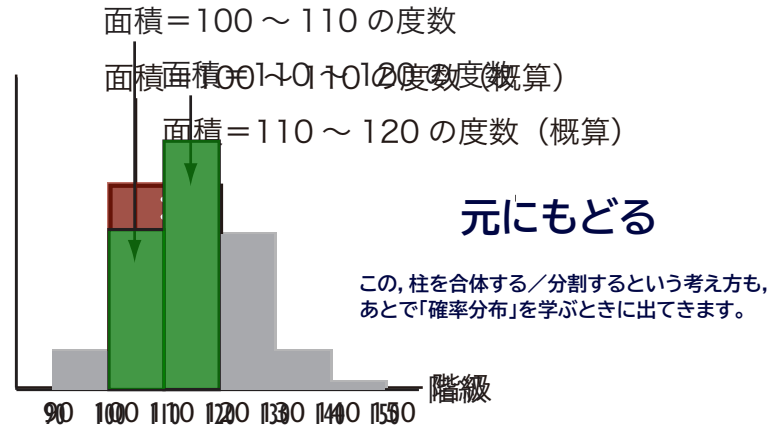
なぜ柱の面積で表すのか？

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



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



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

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

十の位	一の位
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

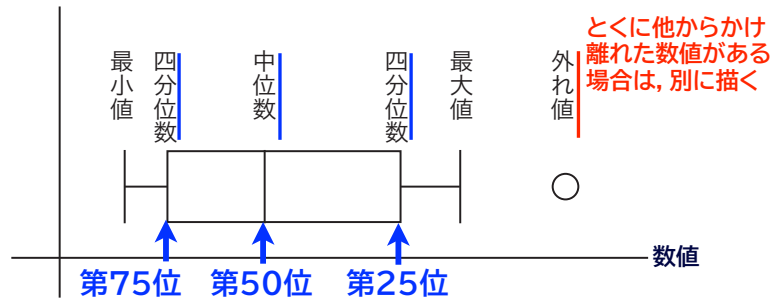
数字の並びによって簡易ヒストグラムになり、数値も見える

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



※中位数のことを中央値、メジアン(median)ともいいます。

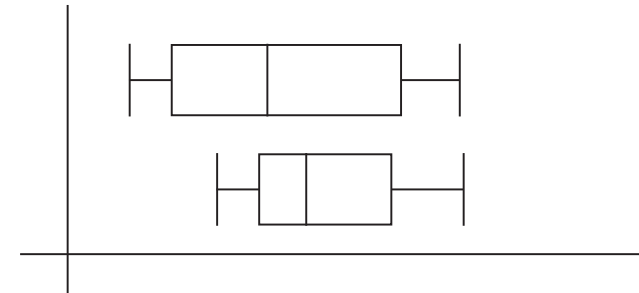
2020年度春学期 統計学

25 / 28

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)



2020年度春学期 統計学

26 / 28

今回も、演習問題は 

今回も、演習問題は

- 今回も、演習問題はテキストの最後に載っています。
- 関大LMSに提出窓口が作ってありますが、**今すぐ回答を提出する必要はありません。**学期末までに回答を提出してください。その内容によって加点します。
- 今回の課題では、**図を描いてもらいます。**パソコンで描いてもいいですし、**手で描いたものを写真に撮って提出してもらってもかまいません。**
- 問題の解答例は、試験の前に公開する予定ですが、**詳細は未定です。**

2020年度春学期 統計学

28 / 28