確率と統計　第 3 回

統計データの記述

4 月 27 日

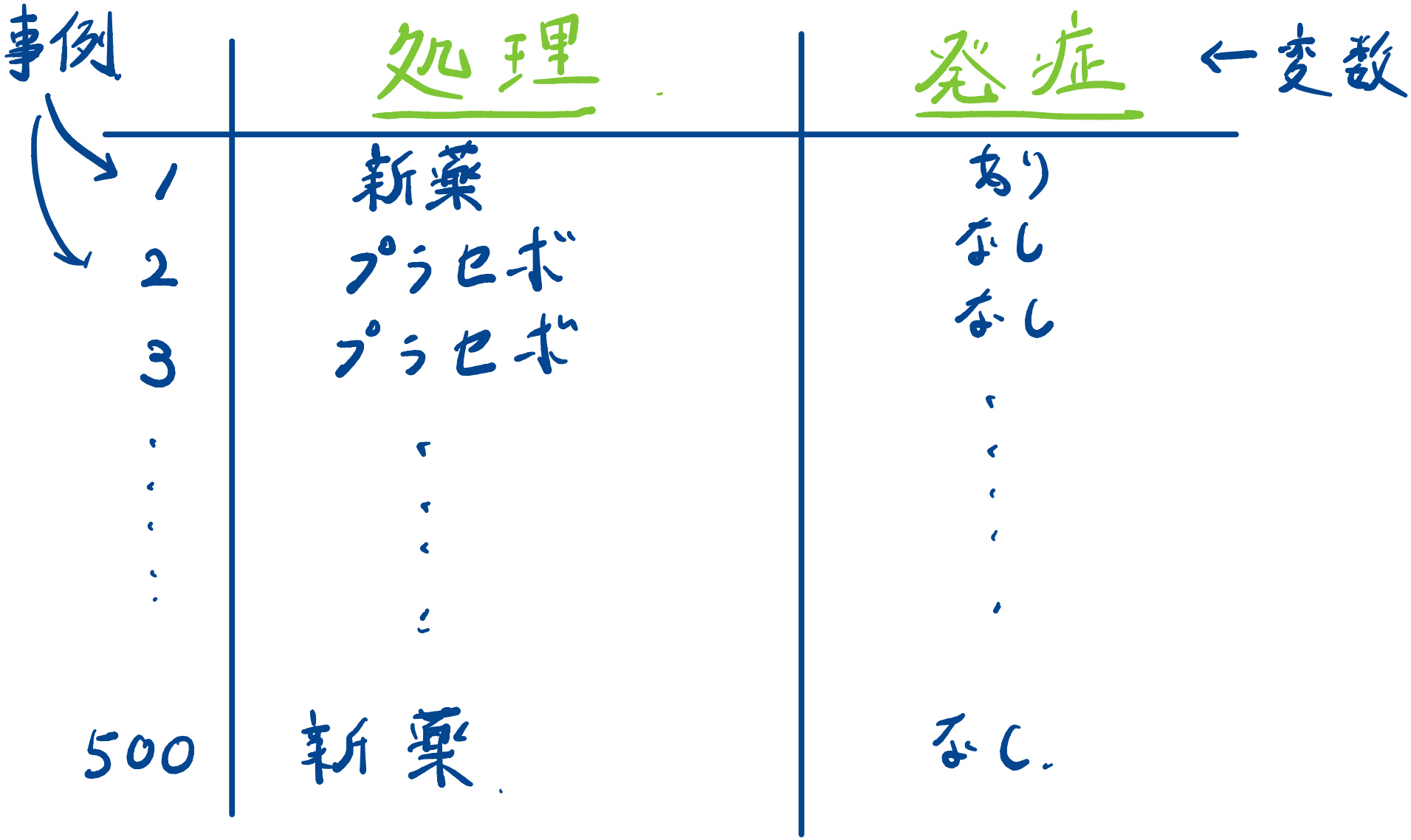
前回の復習

前回の復習

前回の授業では『研究テーマと着目する変数を決めたら, まずは集めたデータをデータ行列の形式に整理する』というデータ分析の鉄則を学んだ.

例: ある薬についての対照実験

例: ある薬についての対照実験



例: 消費者金融の顧客データ

例: 消費者金融の顧客データ

	貸付金額	金利	返済期間	グレード	持家
1	8000	5	36	B	あり
2	100	3	12	A	なし
3	1000	5	3	C	なし
⋮	⋮	⋮	⋮	⋮	
50	200	10	24	C	なし

数値変数

カテゴリカル変数

今回学ぶこと

今回学ぶこと

今回の授業では収集したデータの特徴を表す基本的な図や特徴量を学ぶ.

（数）

今回学ぶこと

今回の授業では収集したデータの特徴を表す基本的な図や特徴量を学ぶ.

基本的な図には, 散布図, ドット・プロット, 度数分布表, ヒストグラム, 箱ひげ図などがあり,

今回学ぶこと

今回の授業では収集したデータの特徴を表す基本的な図や特徴量を学ぶ.

基本的な図には, 散布図, ドット・プロット, 度数分布表, ヒストグラム, 箱ひげ図などがあり, 特徴量には平均, 分散, 標準偏差や, 中央値などがある.

今回学ぶこと

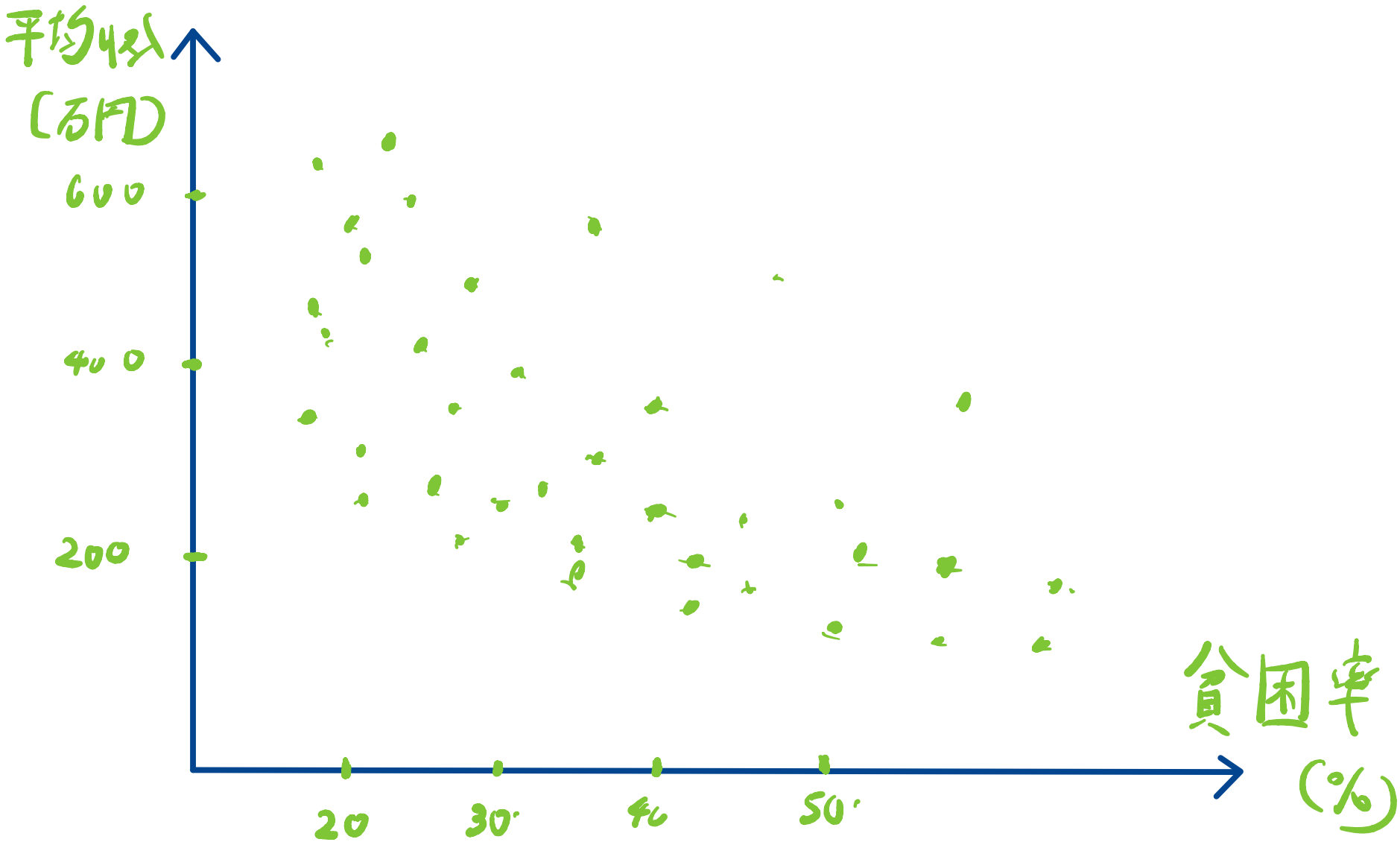
今回の授業では収集したデータの特徴を表す基本的な図や特徴量を学ぶ.

基本的な図には, 散布図, ドット・プロット, 度数分布表, ヒストグラム, 箱ひげ図などがあり, 特徴量には平均, 分散, 標準偏差や, 中央値などがある.

データ行列の型に集めたデータを分析するために, まずは図を描いてから特徴量を求めるとよい.

散布図: 街の平均収入と貧困率

散布図: 街の平均収入と貧困率



この散布図から読み取れることは?

この散布図から読み取れることは?

(a) 2 つの変数には正の相関があるか, 負の相関があるか, あるいは独立か?

この散布図から読み取れることは?

(a) 2 つの変数には正の相関があるか, 負の相関があるか, あるいは独立か?

(b) この相関関係は線形 (直線的) か, 非線形 (曲線的) か?

この散布図から読み取れることは?

(a) 2 つの変数には正の相関があるか, 負の相関があるか, あるいは独立か?

(b) この相関関係は線形 (直線的) か, 非線形 (曲線的) か?

(c) なにかこの散布図から読み取れるか?

ドット・プロットと平均

ドット・プロットと平均

1 から 10 までの数字が描かれた球が 1000 個入ったくじ引きから 10 個の球を選ぶ.

ドット・プロットと平均

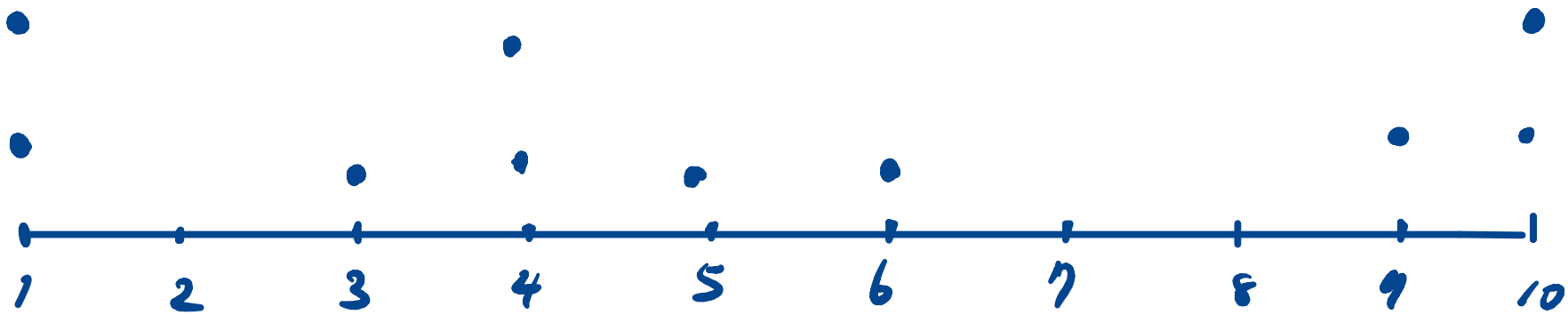
1 から 10 までの数字が描かれた球が 1000 個入ったくじ引きから 10 個の球を選ぶ. (この時, 母集団がくじ引きの球全体で標本が選ばれた 10 個の球.)

ドット・プロットと平均

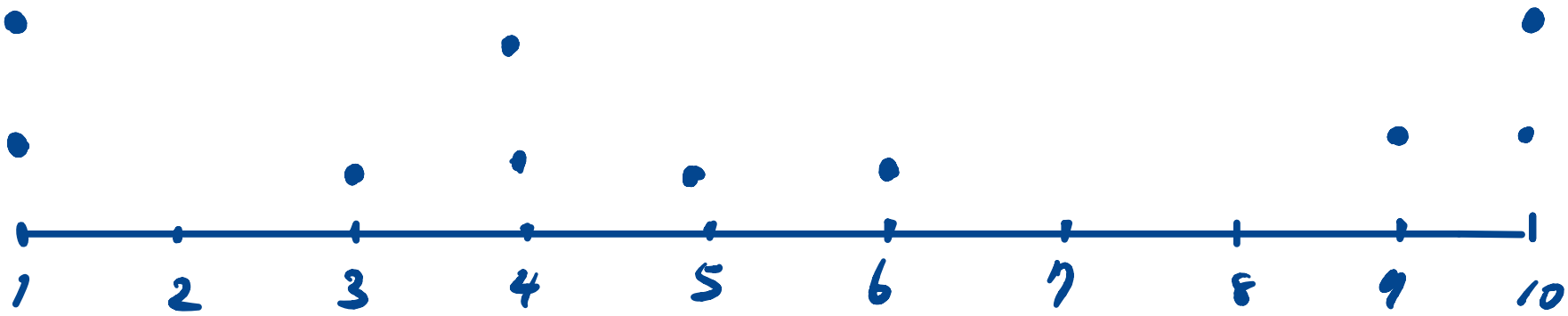
1 から 10 までの数字が描かれた球が 1000 個入ったくじ引きから 10 個の球を選ぶ. (この時, 母集団がくじ引きの球全体で標本が選ばれた 10 個の球.)
標本のドットプロットが以下のようになった.

ドット・プロットの例

ドット・プロットの例



ドット・プロットの例



標本平均は?

標本平均から母平均を推定する

標本平均から母平均を推定する

標本 $x_1, x_2, \ldots, x_n$ の標本平均

標本平均から母平均を推定する

標本 $x_1, x_2, \ldots, x_n$ の標本平均

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} \tag{1}$$

標本平均から母平均を推定する

標本 $x_1, x_2, \ldots, x_n$ の標本平均

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} \tag{1}$$

から母集団 X の母平均 μ(ミュー) の大体の値を予想することができる.

標本平均から母平均を推定する

標本 $x_1, x_2, \ldots, x_n$ の標本平均

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} \tag{1}$$

から母集団 X の母平均 μ(ミュー) の大体の値を予想することができる.

このような足し算による平均はもっとも基本的なデータの特徴量だけれども, 以下のような場合には注意が必要.

例: ある国の一人当たりの平均所得は?

例: ある国の一人当たりの平均所得は?

	州名	人口(万人)	平均所得(万円)
1	A	200	200
2	B	10	300
3	C	100	200
4	D	10	500
5	E	50	100

ヒストグラム

ヒストグラム

データの分布を表現する最も基本的な図がヒストグラム.

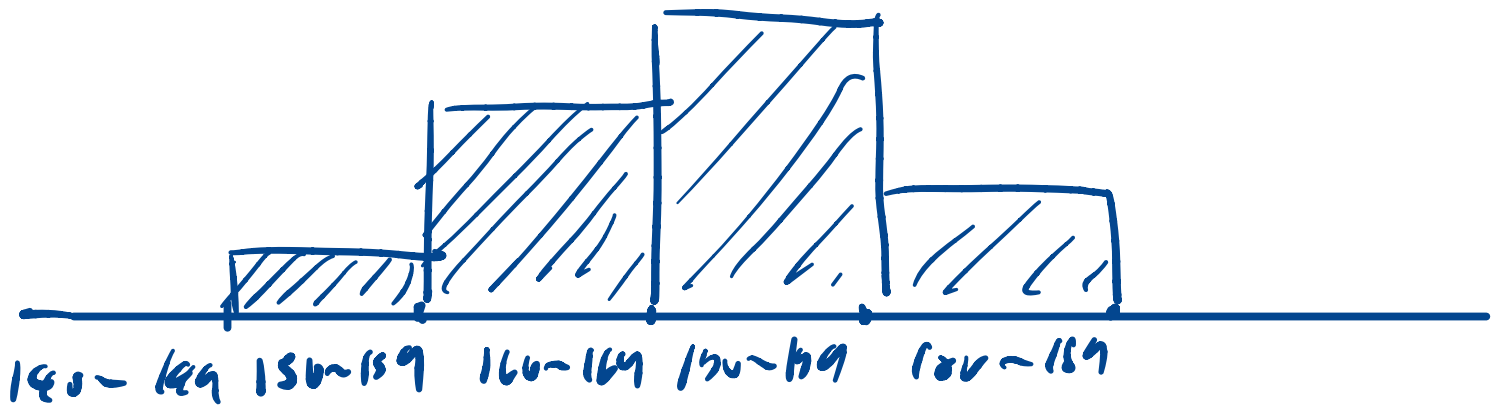
ヒストグラム

データの分布を表現する最も基本的な図がヒストグラム.
データの数が多い場合には, ドット・プロットではなく度数分布表とヒストグラムにデータをまとめる.

度数分布表とヒストグラムの例

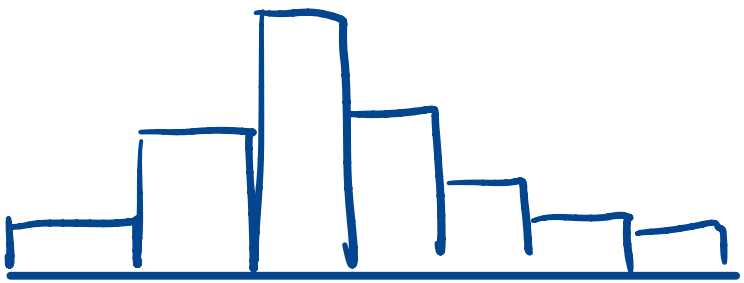
度数分布表とヒストグラムの例

	140~149	150~159	160~169	170~179	180~189	
	0	2	5	7	3	

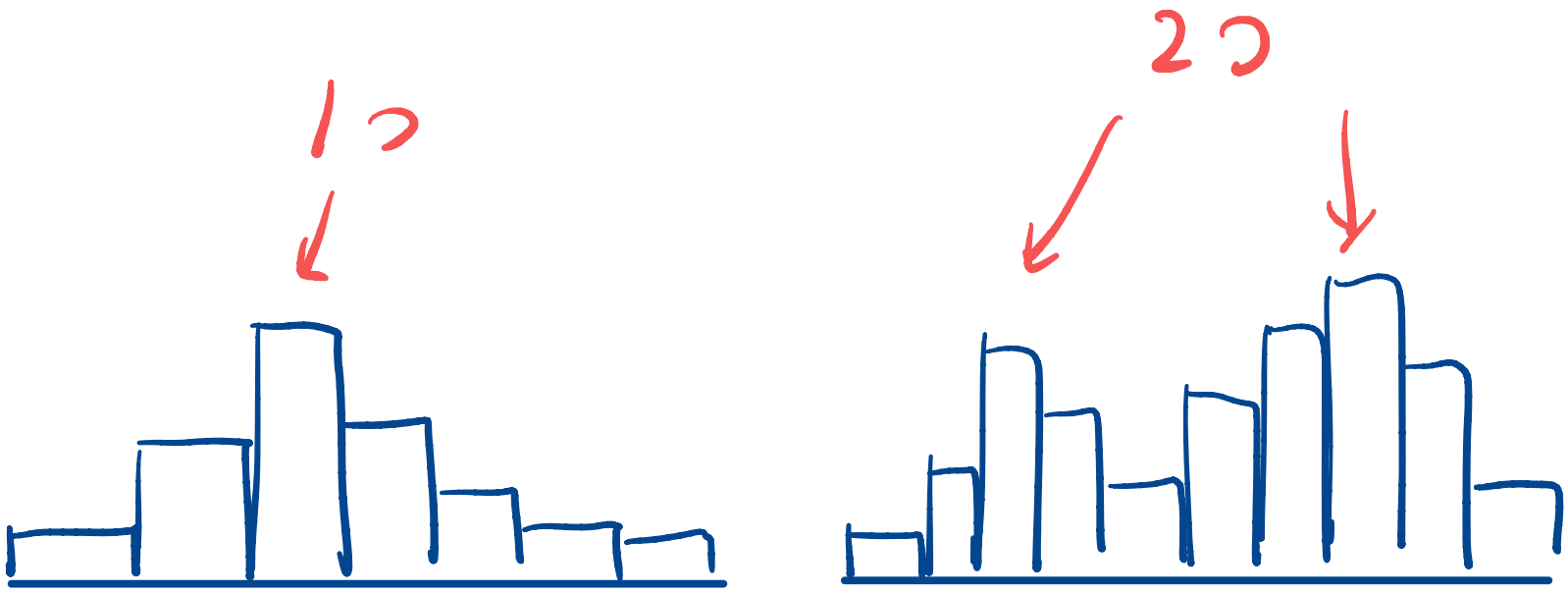


山の数はいくつか?

(ピーク)



山の数はいくつか?



大事な特徴① 山(ピーク)の数.

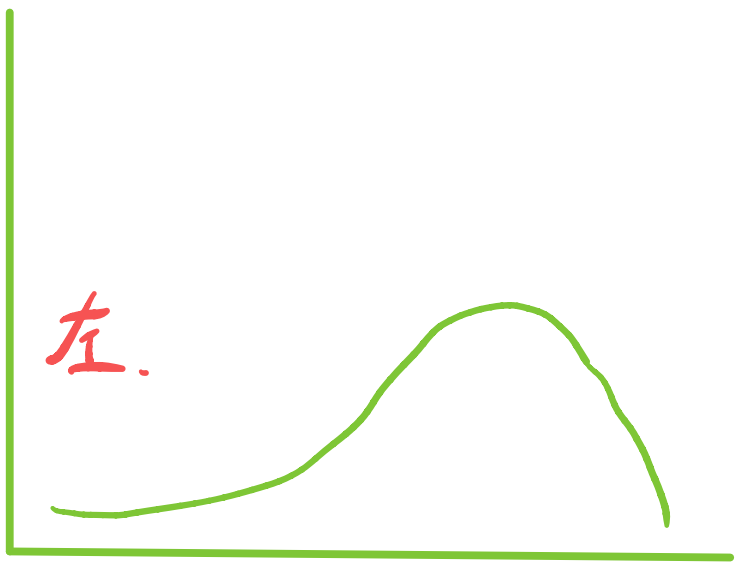
右に歪んでいるか? 左に歪んでいるか? 対称か?

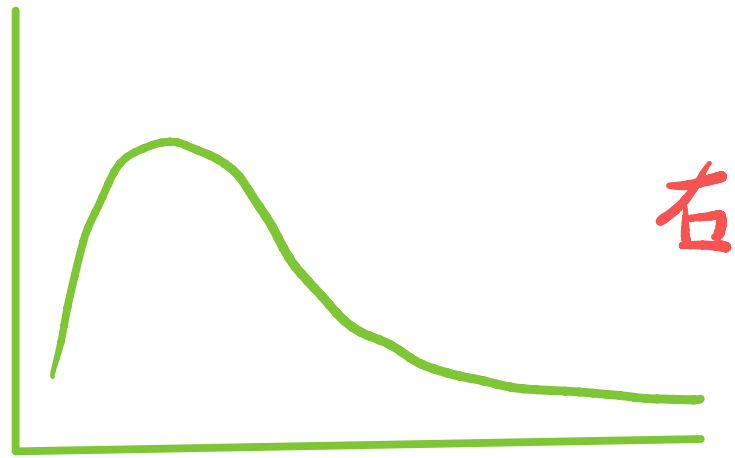
右に歪んでいるか? 左に歪んでいるか? 対称か?

ヒストグラムが右に広がっているとき右に歪んでいるといい, 左に広がっているとき左に歪んでいるという.

右に歪んでいるか? 左に歪んでいるか? 対称か?

ヒストグラムが右に広がっているとき右に歪んでいるといい, 左に広がっているとき左に歪んでいるという.





大事な特徴②
グラフの歪み

分散と標準偏差

分散と標準偏差

データのばらつきを表す量に分散と標準偏差がある.

分散と標準偏差

データのばらつきを表す量に分散と標準偏差がある.標本 $\{x_1, x_2, \ldots, x_n\}$ の標本分散と標準偏差をそれぞれ s^2, s と表す.

分散と標準偏差

データのばらつきを表す量に分散と標準偏差がある.標本 $\{x_1, x_2, \ldots, x_n\}$ の標本分散と標準偏差をそれぞれ s^2, s と表す.また母集団 X の母分散と標準偏差をそれぞれ σ^2, σ と表す.

分散と標準偏差

データのばらつきを表す量に分散と標準偏差がある.標本 $\{x_1, x_2, \ldots, x_n\}$ の標本分散と標準偏差をそれぞれ s^2, s と表す.また母集団 X の母分散と標準偏差をそれぞれ σ^2, σ と表す.$\bar{x}$ を標本平均として,

$$s^2 = \frac{1}{n-1}\{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2\},$$

$$s = \sqrt{s^2}$$

と定義する.

全体の約 70%が (平均) $\pm$ (標準偏差) の範囲に入ることが多く,

全体の約 70%が (平均) $\pm$ (標準偏差) の範囲に入ることが多く,
全体の約 95%が (平均) $\pm 2 \times$ (標準偏差) の範囲に入ることが多い.

全体の約 70%が (平均) $\pm$ (標準偏差) の範囲に入ることが多く,
全体の約 95%が (平均) $\pm 2 \times$ (標準偏差) の範囲に入ることが多い.

このように平均と標準偏差でデータの大体の分布がわかる.

しかし...

平均と標準偏差の弱点

平均と標準偏差の弱点

全て同じ平均と標準偏差をもつデータのヒストグラム.

平均と標準偏差の弱点

全て同じ平均と標準偏差をもつデータのヒストグラム. $\bar{x}$ s

平均と標準偏差の弱点

全て同じ平均と標準偏差をもつデータのヒストグラム. $\bar{x}$ s

左右への歪みや山の数は平均と標準偏差には反映されない. ゆえにヒストグラムを描いたり, 平均値と中央値を比較したりする必要がある.