広島大学

データサイエンス基礎

Fundamental Data Science

第3回　データの種類とデータの要約

広島大学　数理統計学グループ

今日の内容

- **データ，変数の種類**
- **データの要約（記述統計の方法）**

データ
変数の種類

統計データ

表: 家計調査における世帯データ (～費の単位は千円, 1か月の費用)

世帯＼項目	食費	教育費	交際費	世帯人数	勤労者世帯
1	75	40	66	4	1
2	70	80	91	1	0
⋮	⋮	⋮	⋮	⋮	⋮
100	30	51	65	3	1

※ 勤労者世帯: 1が勤労者世帯, 0が勤労者以外の世帯を表す

Q

個体数がいくらで
次元数がいくらの
データ？

- 調査を行う個々の対象 (各世帯のこと) を**個体**といい, 個体の総数を**個体数**という
- 調査して得られた数値を**観測値** (または単に**データ**) という
- 調査項目を**変数**といい, 変数の総数を**次元数**という

変数の種類

量的変数

数値を表すもので, 多い or 少ないがわかるもの

例 身長, 体重, 年収など

質的変数　区別を表すもの

例 車の所持, 学歴, 国籍など

Q 5つの変数それぞれは、**量的変数**と**質的変数**のどちらに分類される？

世帯＼項目	食費	教育費	交際費	世帯人数	勤労者世帯
1	75	40	66	4	1
2	70	80	91	1	0
⋮	⋮	⋮	⋮	⋮	⋮
100	30	51	65	3	1

例: 住宅価格と環境条件データ (高橋等, 2000)

Q どれが**量的変数**? **質的変数**?

ID	Price	Period	Area	Size	JR	St.time	Age	Distance
54	3730	0	148	103	0	9	0	13
55	1480	1	110	67	0	3	23	8
64	3590	1	105	100	1	8	8	8

※ ID number;
Price (単位は万円)　[一戸建て住宅の価格];
Period (いつ調査したものか. 1998年なら0, 1999年なら1);
Area (m^2) [土地面積];
Size (m^2) [床面積];
JR (最寄駅が JR なら1, 他なら0) [最寄駅の種類];
St.time (minute) [最寄の駅, バス停からの徒歩時間];
Age (year) [築後年数];
Distance (km) [広島市中心地からの距離]

住宅データ
引用元: 高橋ら , 広島女学院大学生活科学部紀要, 7: 57-65, 2000.

量的変数の分類

量的変数の分類

四則演算 +,−,×,÷ の意味に注目!!

比率変数

足し算 (引き算) に意味がある
掛け算 (割り算) も意味がある

例 身長, 体重, 年収

200cmは100cmの2倍という
⇒ 掛け算 (割り算) に意味がある

間隔変数

足し算 (引き算) に意味がある
掛け算 (割り算) に意味がない

例 気温 (℃), 時刻

-6℃は-2℃の3倍といわない
⇒ 掛け算 (割り算) に意味がない

5時は1時の5倍といわない
⇒掛け算 (割り算) に意味がない

見分け方
「0」の値に意味があるかないか

比率変数 「0」のとき本当に「何もない」状態
間隔変数 「0」のとき本当に「何もない」状態ではない

質的変数の分類

質的変数の分類　順序の有無に注目!!

順序変数

四則演算に意味がない

大小, 前後 (順序) に意味がある

例　アンケート結果
(1:とても悪い, 2:やや悪い,
3:ややよい, 4:とてもよい)

「とても悪い」＋「やや悪い」
＝「やや良い」とはならない

名義変数

四則演算, 順序に意味がない

例　「1:A型, 2:B型, 3:O型, 4:AB型」, 電話番号

分類の必要性

変数の種類によって
データ分析手法の有効性が異なる

質的変数の平均値は意味ない

～変数では～分析手法を使おう

4つの変数の
違いについて
考えてみましょう

変数の種類によってデータ分析手法の有効性が異なる

比率変数 ＞ **間隔変数** ＞ **順序変数** ＞ **名義変数**

データサイエンス基礎，データの種類とデータの要約：P.9

データの要約

データの要約

ちょっと住宅の価格を調べて
その特徴をまとめておいて

全部で158戸の住宅価格を調べました!!

158戸の住宅価格データ (単位は万円) (高橋等, 2000)

3150	3150	3500	6500	3800	2890	3170	3390	3650	3500	3880	3950	4400	4600	5980	2450	3480	3880	2980	800
2000	2670	2980	2980	3980	4400	4570	3480	3800	2450	4400	2300	2880	2980	3300	3380	3660	3180	3750	4180
1380	2690	1680	2580	2780	2890	3080	2980	3050	2980	2780	3880	3680	3730	1480	2000	2050	2170	2180	2290
2300	2380	1380	3590	2630	2680	2880	3080	3200	3550	3790	5300	3190	2780	4100	5580	1280	2480	2580	2880
2900	3180	3280	3280	3280	3300	3680	3850	4290	7800	1500	2580	3180	3190	3100	2980	3280	3070	3190	3680
2980	4480	3730	3980	3440	3460	3700	3440	3390	4380	1490	3300	3180	2480	3980	1980	2880	4000	4580	4580
3500	3550	3400	3550	2000	2700	2480	1620	2580	2300	850	2290	3180	2050	2580	3780	5780	2680	4380	5977
2280	3090	4400	1950	1980	3200	3380	4280	5980	4180	4350	3730	3980	4000	4650	4680	4680	1580		

部下

でも住宅が多くて特徴はよくわかりません

とりあえず**グラフ**にまとめて

不動産会社の上司

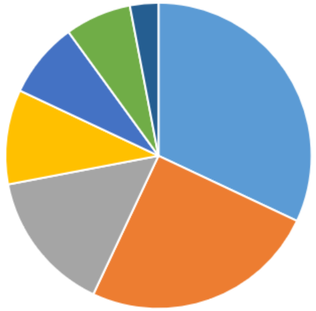
グラフ

グラフ

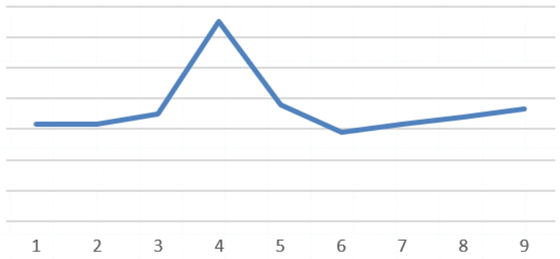
目で見て直感的にデータの特徴が
把握できるようにしたもの

ヒストグラムについて
話していきます

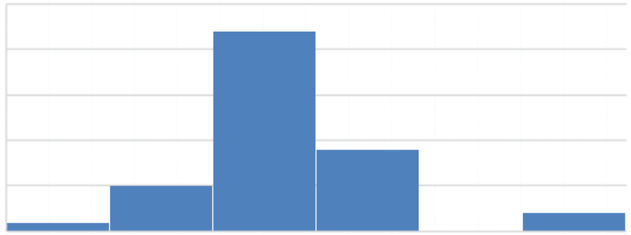
円グラフ



折れ線グラフ



ヒストグラム



データの特徴を知るにはまずはグラフ!!

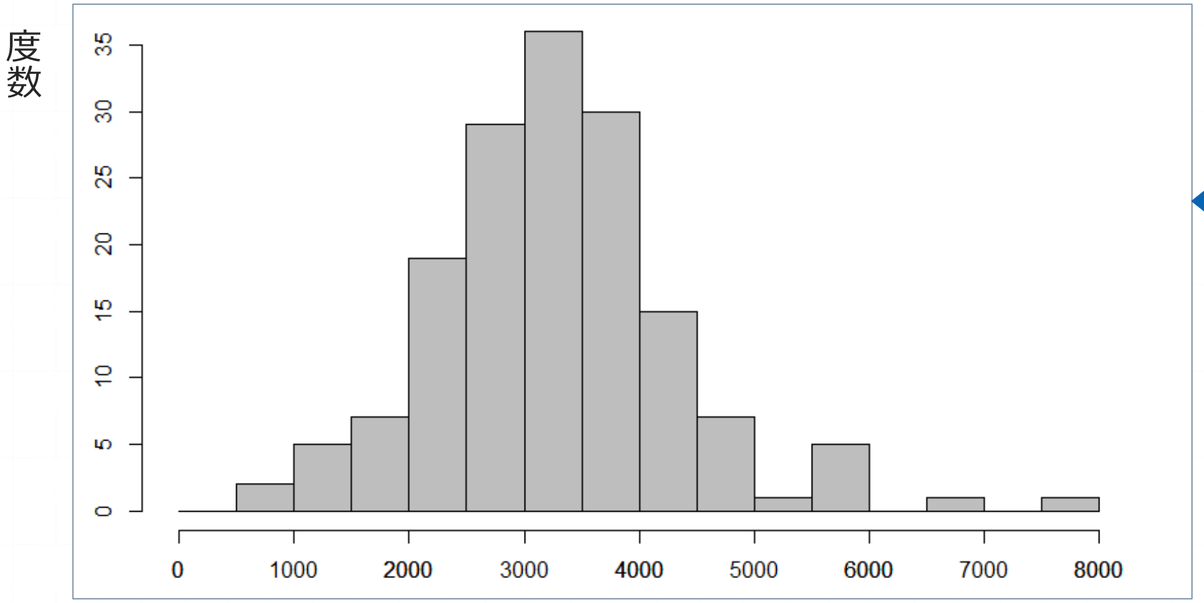
ヒストグラム

ヒストグラム

量的変数の**データの分布**を表すグラフ

質的変数の場合は
通常**棒グラフ**を用いる

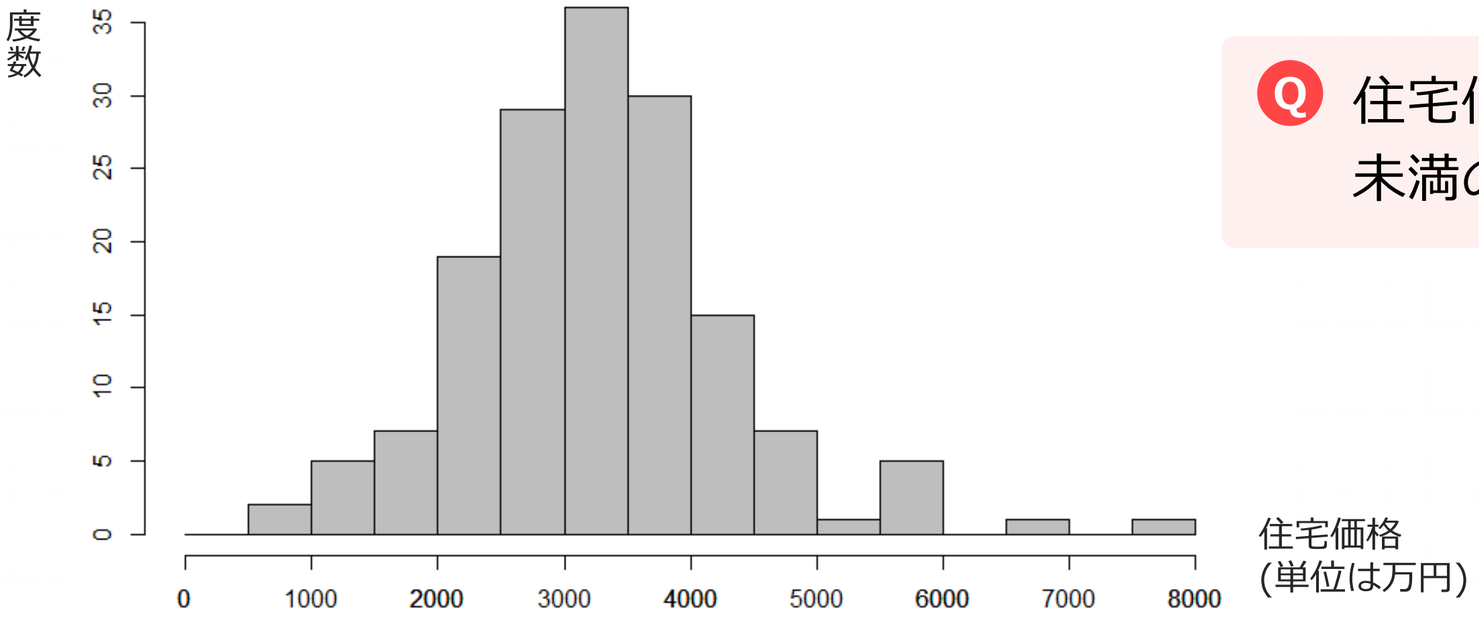
158戸の住宅価格データのヒストグラム



分布: データ全体の形
(集まり具合,
散らばり具合などがわかる)

ヒストグラム

158戸の住宅価格データのヒストグラム



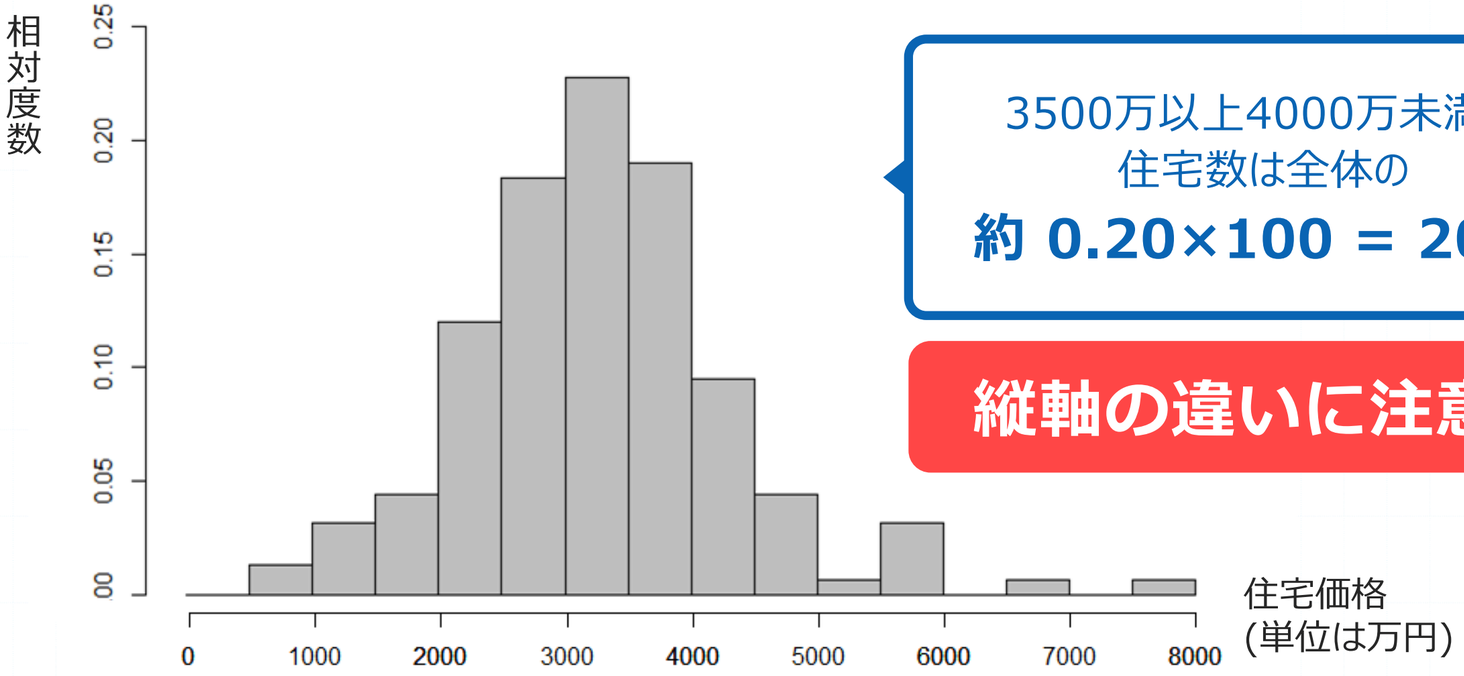
Q 住宅価格が3500万以上4000万未満の住宅は約何戸?

棒の長さ = 各区間に含まれるデータの個数 (**度数**という)

例 0以上500未満, 500以上1000未満,…

ヒストグラム (相対度数)

158戸の住宅価格データのヒストグラム



3500万以上4000万未満の
住宅数は全体の
約 0.20×100 = 20%

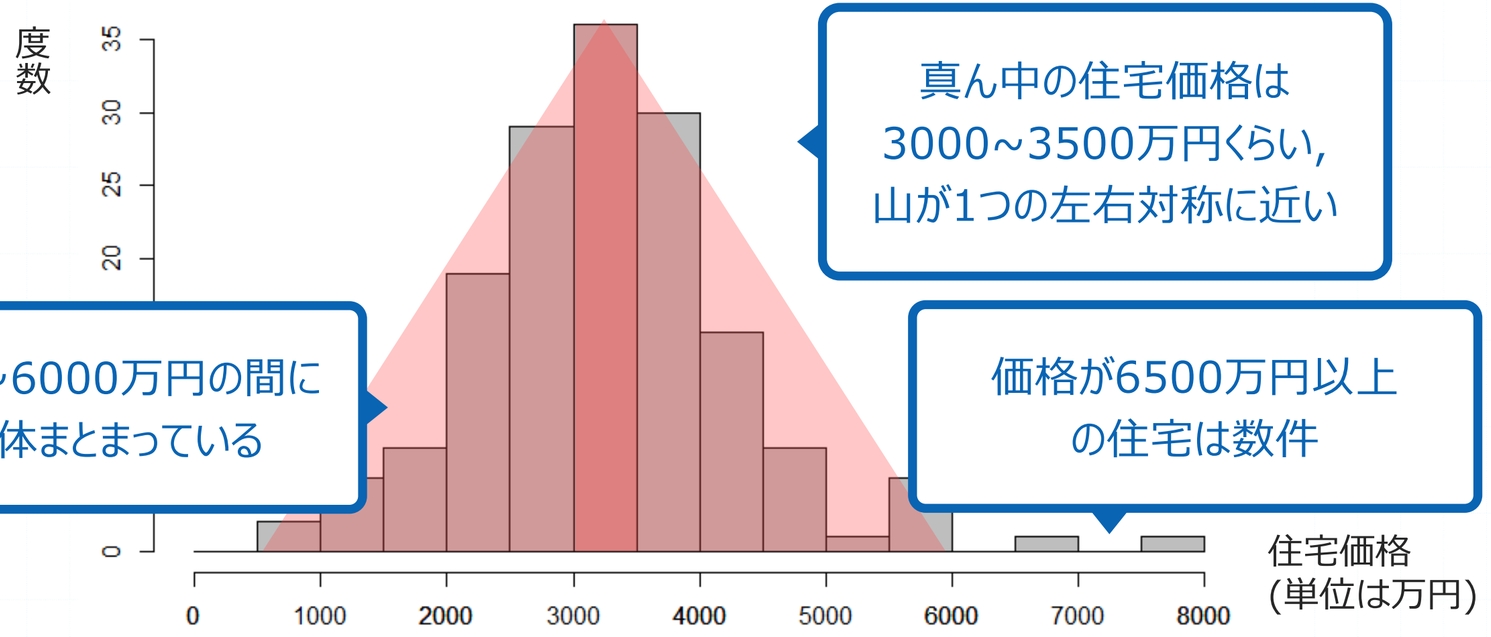
縦軸の違いに注意!!

さっきのと違う場所は…

- 棒の長さ ＝ 度数÷個体数 (**相対度数**という)

ヒストグラム

158戸の住宅価格データのヒストグラム



真ん中の住宅価格は
3000~3500万円くらい,
山が1つの左右対称に近い

500~6000万円の間に
大体まとまっている

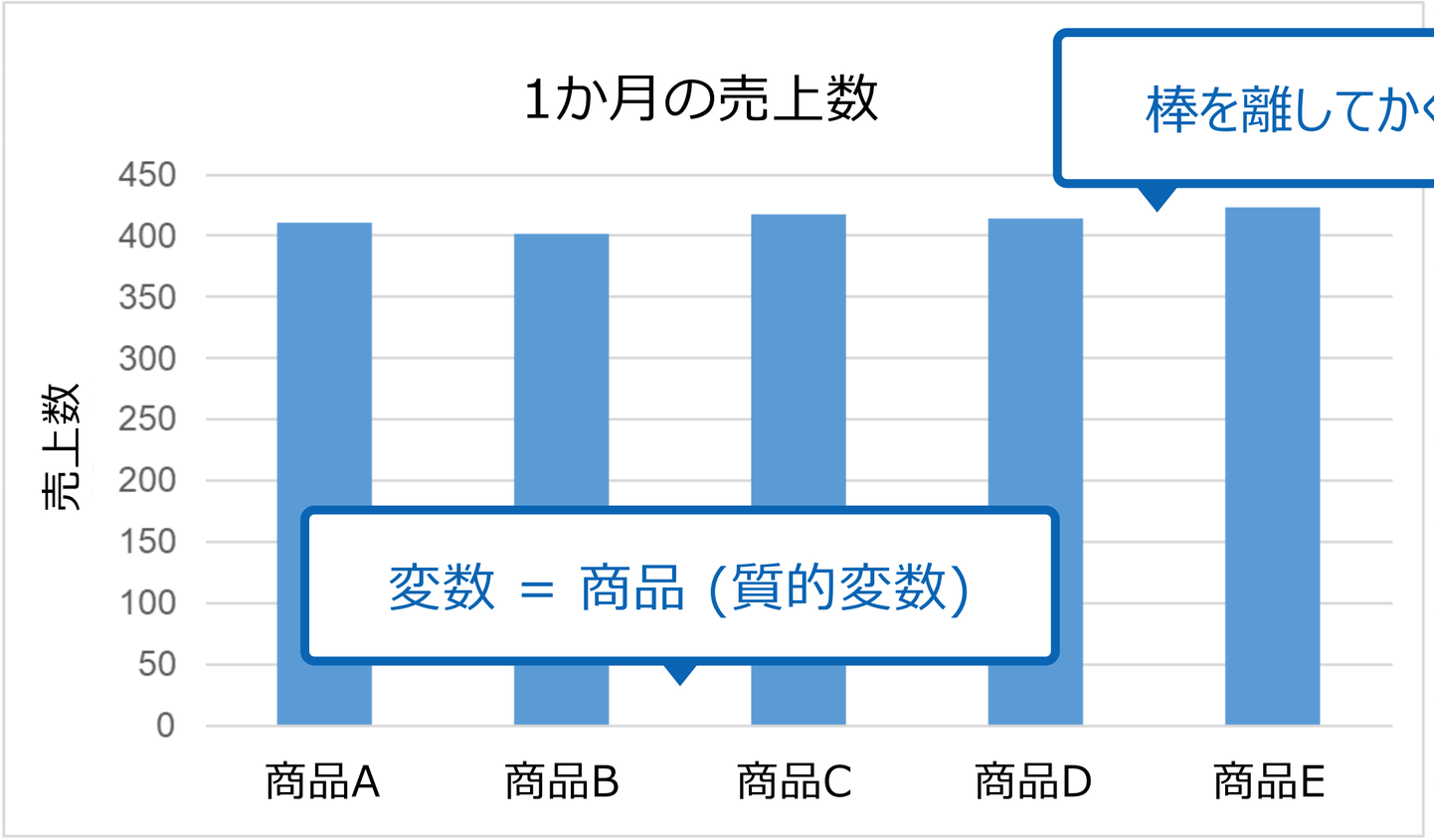
価格が6500万円以上
の住宅は数件

目で見るだけで
いろんなことが
読み取れますね

**データの数値をみなくても
データ全体の特徴が目でみてわかる!!**

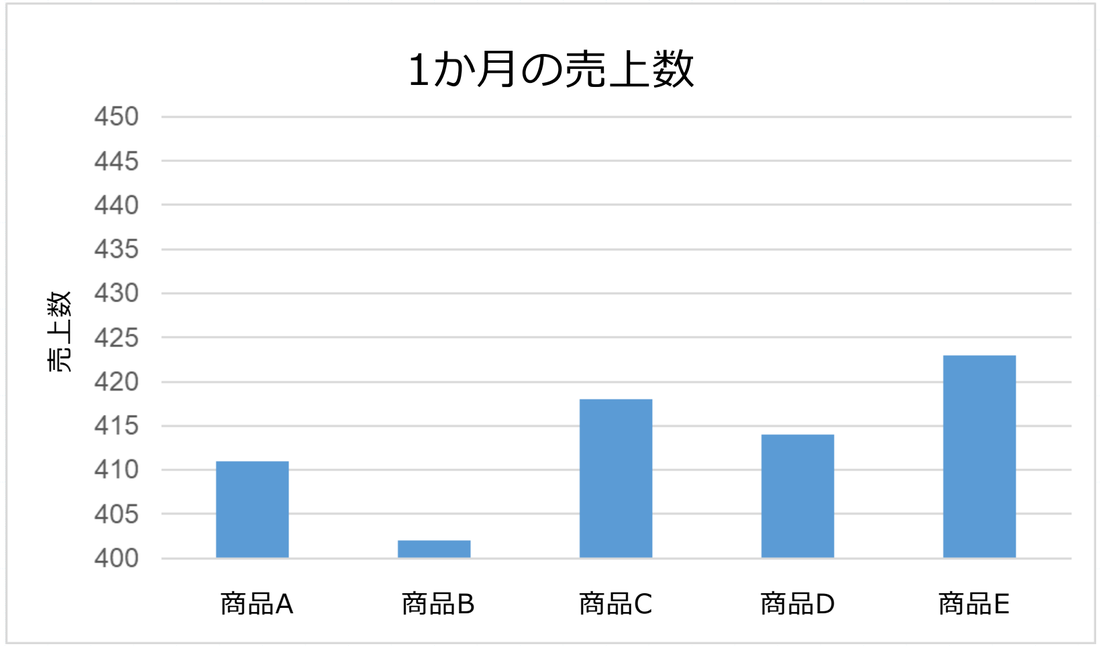
質的変数の場合

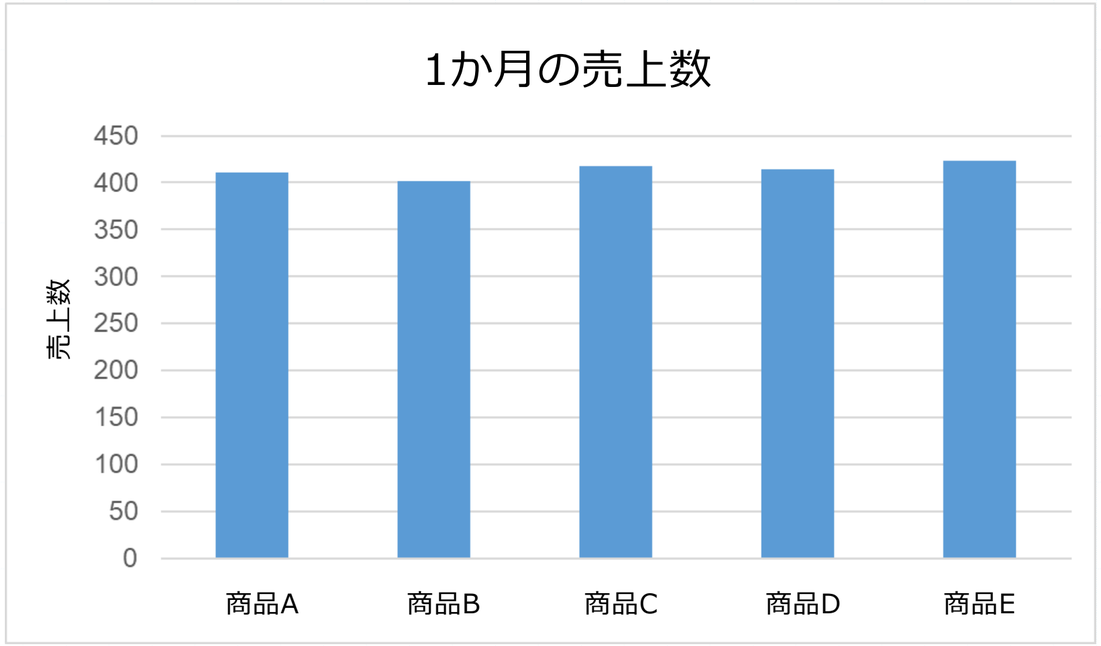
棒グラフ



ある商品の
1ヶ月の売り上げ数を
表したグラフです

グラフの注意

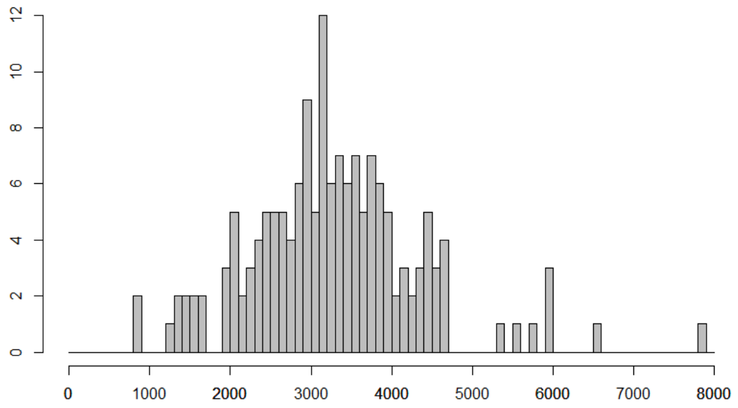




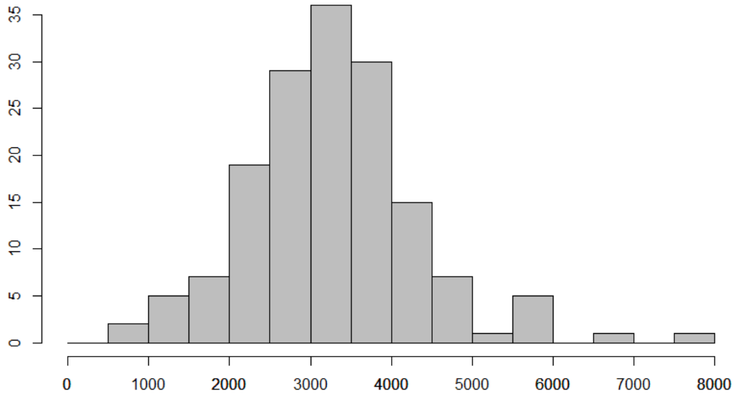
縦軸の違いで見た目も変わる

ヒストグラムの注意

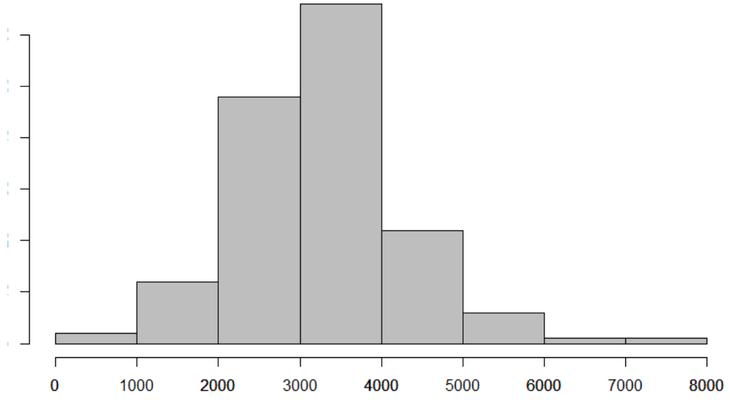
各区間の幅：100刻み



各区間の幅：500刻み



各区間の幅：1000刻み

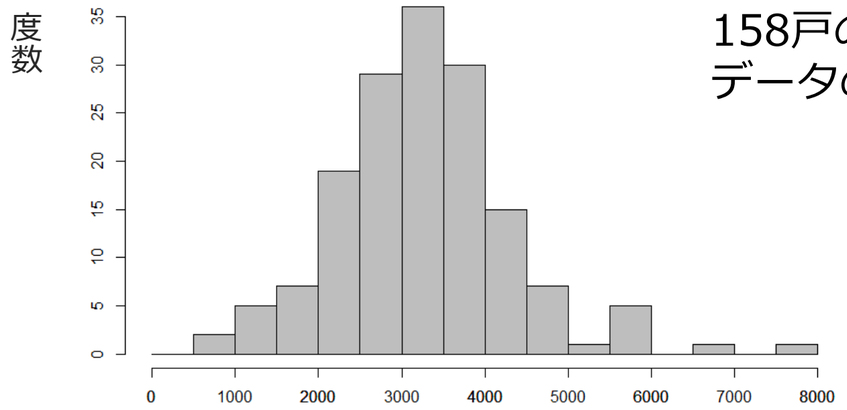


各区間の幅の違いで見た目も変わる

要約統計量

データの要約

ヒストグラムをかくとこんな特徴でした



158戸の住宅価格
データのヒストグラム

部下

真ん中の価格は大体3000~3500万円です

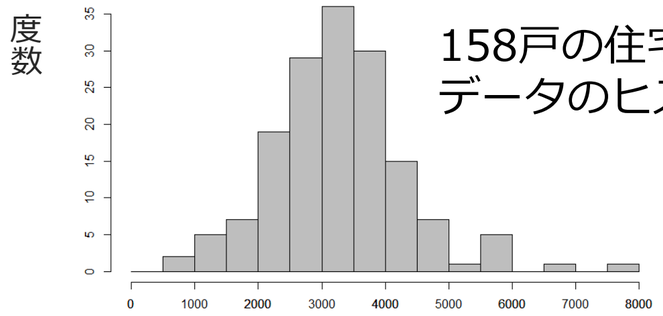
なんとなくわかったけど…
大体の範囲じゃなくてもっとわかりやすく教えて

不動産会社の上司

要約統計量

要約統計量

データの分布の特徴を表した数値. 伝達が簡単で客観性がある



158戸の住宅価格
データのヒストグラム

＋

平均値は〇〇万円
分散は〇〇

基本的な要約統計量の例

分布の位置: 平均値, 中央値, 最頻値など

分布の散らばり (具合) : 分散, 標準偏差, 範囲など

分布の**中心**の位置を表す
要約統計量を**代表値**という

平均値

表. 5人の体重データ

人	体重 (kg)
1	60
2	52
3	44
4	74
5	60

文字で一般的に

表. 個体数 n, 変数 x の1次元データ

個体	変数 x
1	x_1
2	x_2
⋮	⋮
$n-1$	x_{n-1}
n	x_n

$$\text{平均値} = \frac{60+52+44+74+60}{5}$$
$$= 58$$

データをすべて足して個体数で割ったもの

※ 体重データだと,

$n = 5$

x: 体重 (kg)

$x_1 = 60, x_2 = 52, x_3 = 44,$
$x_4 = 74, x_5 = 60$

文字を使って
データを一般化してみましょう

平均値

平均値 (mean)

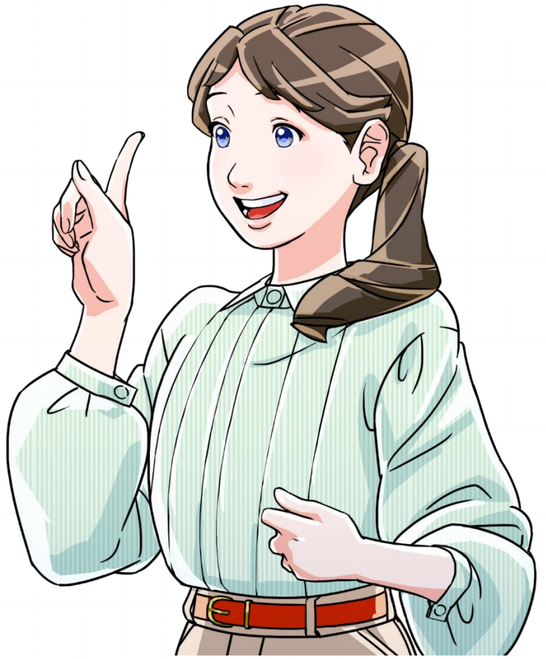
変数 x の n 個のデータ $x_1, \ldots, x_n$ に対する平均値:

$$\text{平均値 } \bar{x} = \frac{x_1 + \cdots + x_n}{n}$$

個体	変数 x
1	x_1
2	x_2
⋮	⋮
$n-1$	x_{n-1}
n	x_n

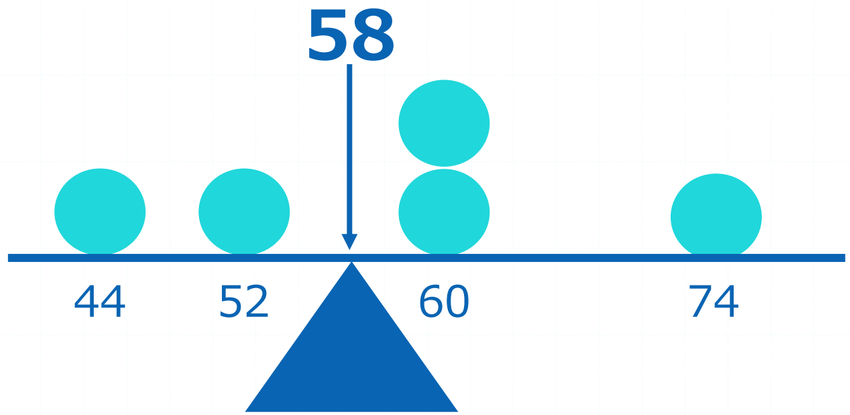
データをすべて足して
個体数で割ったもの
(文字で書いただけ)

計算式も
文字を使って
一般化できそうですね



平均値のコメント

- 分布の**中心** (特に,**重心**) を表す



- **量的変数**に適した代表値

量的変数の例　**身長, 体重, 年収, 年度, 気温 (℃)**

質的変数 (名義変数) の例　**「1:A型, 2:B型, 3:O型, 4:AB型」**

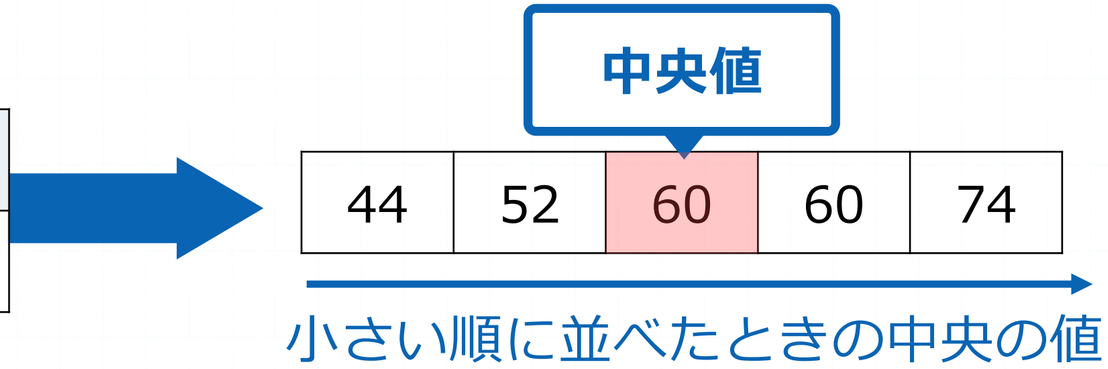
数字自体に意味はない,
ただ区別しているだけ

平均値が2.5となった ⇒ 分布の中心はB型とO型の間??

中央値

表. 5人の体重データ

人	1	2	3	4	5
体重(kg)	60	52	44	74	60



中央値 (median)

n 個のデータ $x_1, \ldots, x_n$ の中央値:

n が**奇数**のとき, 小さい順に並べたときの**真ん中の値**

例 データ 5,7,8,11,12 ⇒ 中央値: 8

n が**偶数**のとき, 小さい順に並べたときの**真ん中2つの値の平均値**

例 データ 1,5,7,8,11,12 ⇒ 中央値: $\frac{7+8}{2} = 7.5$

中央値のコメント

- 分布の中心を表す (順番の真ん中)
- **量的変数**, **順序変数**に適した代表値

順番の真ん中だから
大小関係があれば
OK

大小関係に
意味がある変数だと
中央値は適しています

順序変数の例　**アンケート結果**

1:とても悪い, 2:やや悪い,
3:ややよい, 4:とてもよい

➡ 中央値は3.5

データの真ん中は **ややよい, とてもよい** の間

練習問題 (1)

5人の体重データの

平均値: 58 (kg)

中央値: 60 (kg)

表. 5人の体重データ

人	1	2	3	4	5
体重 (kg)	60	52	44	74	60

Q.1 上の表に6人目の観測値61 (kg) が追加されたときの
6人の平均値と中央値を求めよ

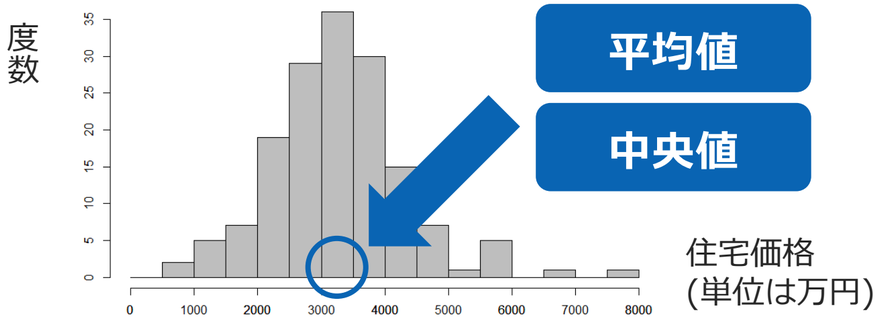
Q.2 上の表に6人目の観測値を誤って1000 (kg) と追加してしまったときの
6人の平均値と中央値を求めよ

Q.3 1と2の結果から平均値と中央値の違いを考察せよ

データの要約

分布の位置を
調べました

158戸の住宅価格データのヒストグラム



部下

平均値は3291万1200円,中央値は3195万円です

ほとんど同じだけど微妙に値が違うね

どっちの値に注目すればいいの??
平均値と中央値の違いはなに??

不動産会社の上司

平均値と中央値の違い

練習問題 (1)

5人の体重データの

平均値: 58 (kg)

中央値: 60 (kg)

表. 5人の体重データ

人	1	2	3	4	5
体重 (kg)	60	52	44	74	60

Q.1 上の表に6人目の観測値61 (kg) が追加されたときの
6人の平均値と中央値を求めよ

Q.2 上の表に6人目の観測値を誤って1000 (kg) と追加してしまったときの
6人の平均値と中央値を求めよ

Q.3 1と2の結果から平均値と中央値の違いを考察せよ

解答例: 練習問題 (1)

A.1 データ 44, 52, 60 ,60, 61, 74

平均値 $\frac{40+52+60+60+61+74}{6} = 58.5$

中央値 $\frac{60+60}{2} = 60$

A.2 データ 44, 52, 60 ,60, 74, 1000

平均値 $\frac{44+52+60+60+74+1000}{6} = 215$

中央値 $\frac{60+60}{2} = 60$

3問目については
次のスライドで
詳しく解説していきます

平均値, 中央値の違い1

5人のデータ

平均値: 58

中央値: 60

1人を追加

61 (kg) 追加

平均値: 58.5

中央値: 60

1000 (kg) 追加

平均値: **215**

中央値: 60

極端な値の影響

極端な値のことを**外れ値**という

平均値: 外れ値の影響を大きく受ける

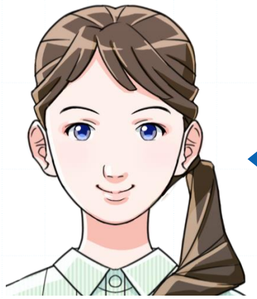
中央値: 外れ値の影響をあまり受けない

外れ値の影響を受けたくないなら中央値

外れ値のコメント

例. 5人の点数データ

人	1	2	3	4	5
点数 (点)	40	40	40	40	100



100点は外れ値だから
無視できる中央値を使う!!
中央値: 40点

100点も重要なデータ
だから平均値を使う!!
平均値: 52点

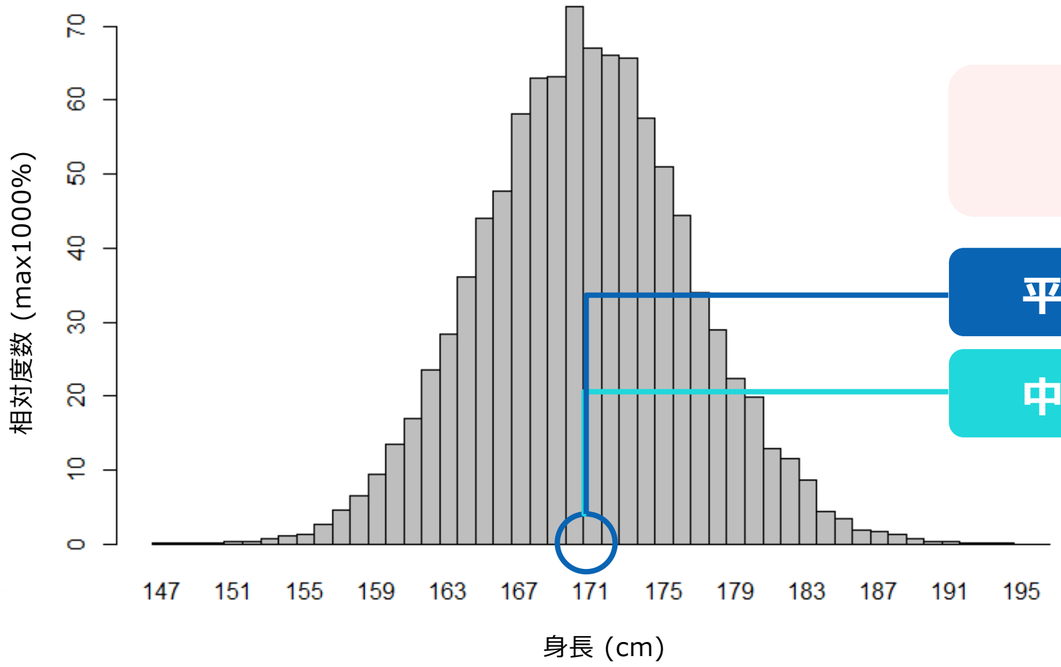
順番の真ん中を知りたい (普通の人の点数)
⇒ **中央値**

数値の真ん中を知りたい
⇒ **平均値**

外れ値があってもどんな真ん中を知りたいかで使い分け

平均値, 中央値の違い2

図. 2019年度17歳男子の身長のヒストグラム



左右対称な分布

平均値 170.6cm

中央値 約171cm

他には
どんな違いがあるんですか?

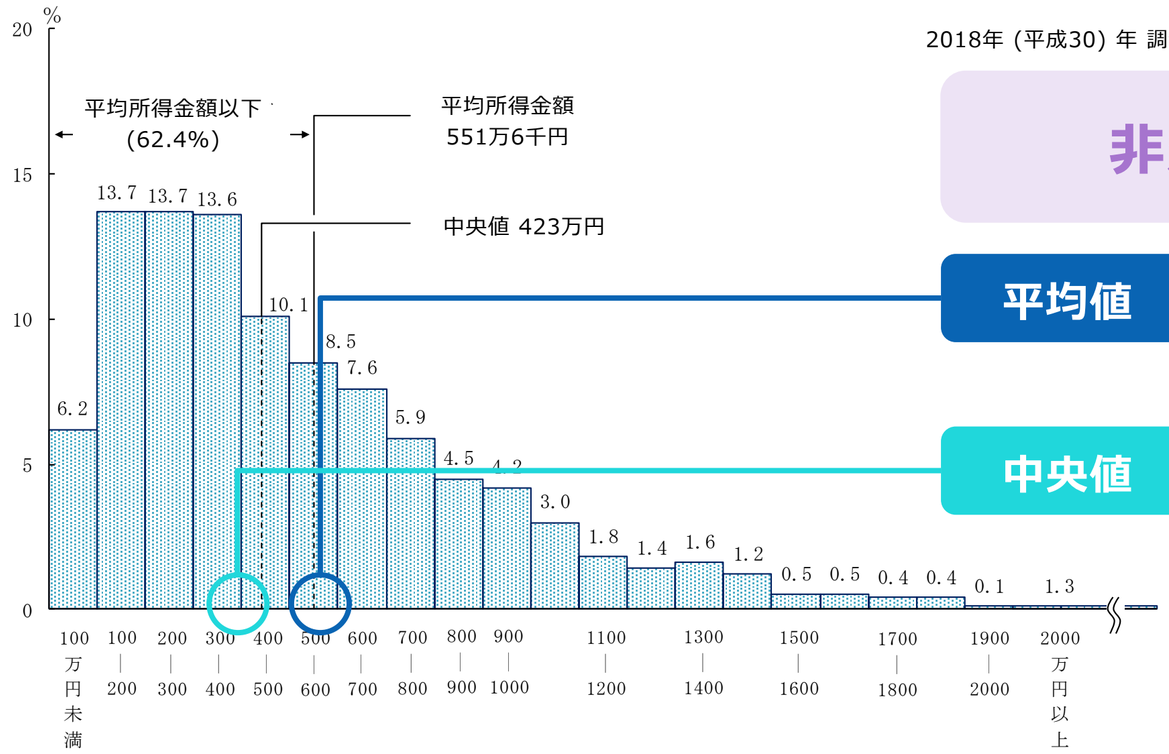
左右対称な分布だと
平均値も中央値もそれほどかわらない

2019年度17歳男子の身長ヒストグラム
引用元: e-Stat_学校保健統計調査_令和元年度_全国表_身長の年齢別分布からヒストグラムを自作 (学校保健統計調査 令和元年度 全国表 | ファイル | 統計データを探す | 政府統計の総合窓口 (e-stat.go.jp))

データサイエンス基礎, データの種類とデータの要約: P.35

平均値, 中央値の違い2

図. 2018年所得金額階級別世帯数のヒストグラム



非対称な分布

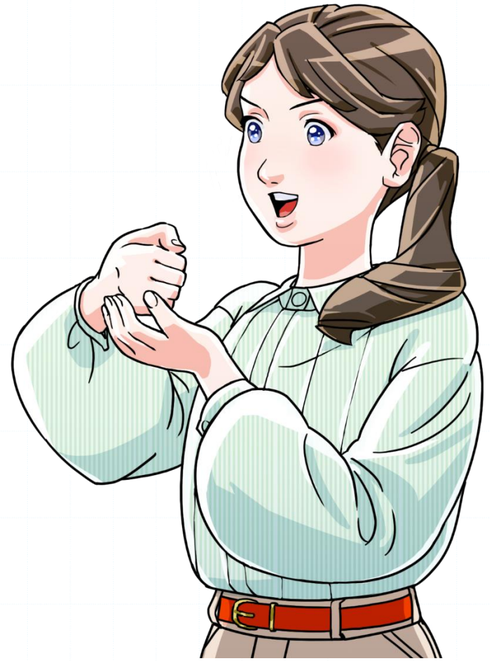
平均値

551万6千円
(下から約60%の位置)

中央値

423万円
(下から50%の位置)

この例のような場合は
中央値を使う方が
いいですね



非対称な分布だと平均値と中央値は大きく異なる

図引用: 厚生労働省 (以下のURLから加工して作成)
https://www.mhlw.go.jp/toukei/saikin/hw/k-tyosa/k-tyosa18/index.html

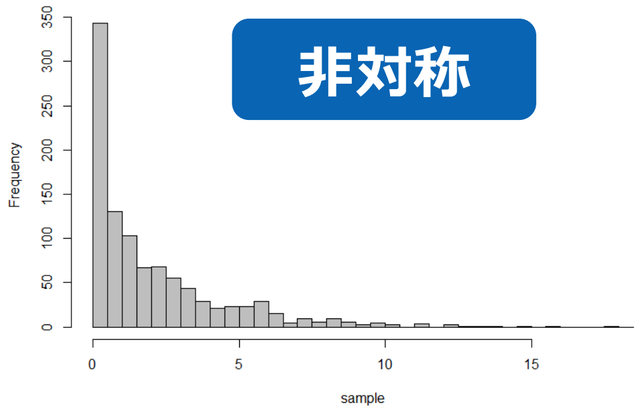
平均値と中央値の使い分け（目安）

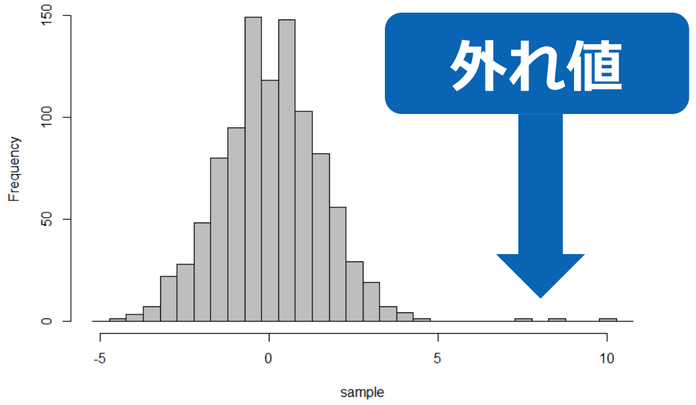
どっちも計算はしよう!!

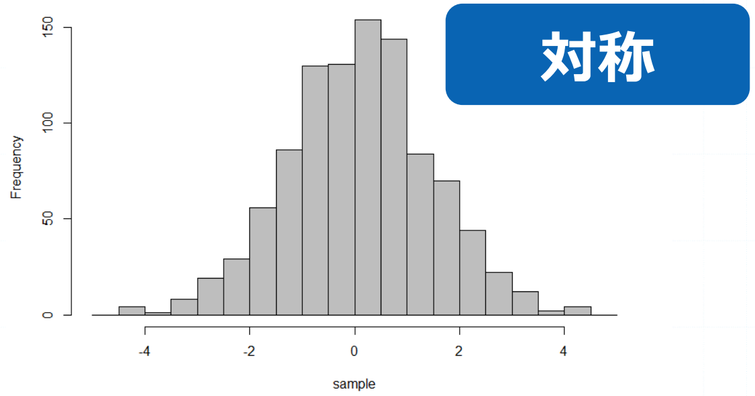
- **順序変数**なら中央値
- **量的変数**ならヒストグラムをみて決める

分布が非対称 or 外れ値がありそうなら**中央値**

それ以外 (対称で外れ値がなさそう) なら**平均値**　　※ しかしどんな中心を知りたいかで使い分けが重要

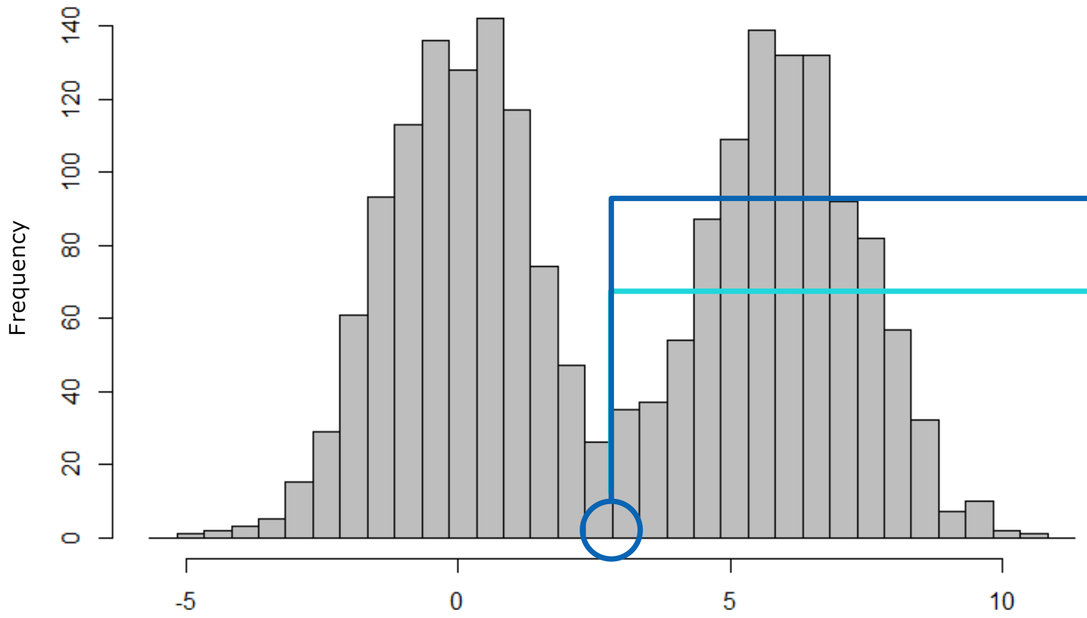






補足1: 平均値, 中央値

山が2つあるヒストグラム



平均値

中央値

分布の中心を表してるとは
言えなさそう…

特徴の異なる集団が混じっている可能性!!

例 男女が混じった身長のデータ
ストレート, 変化球が混じった球速のデータ

➡ 各集団ごとに分けて分布の特徴をみる

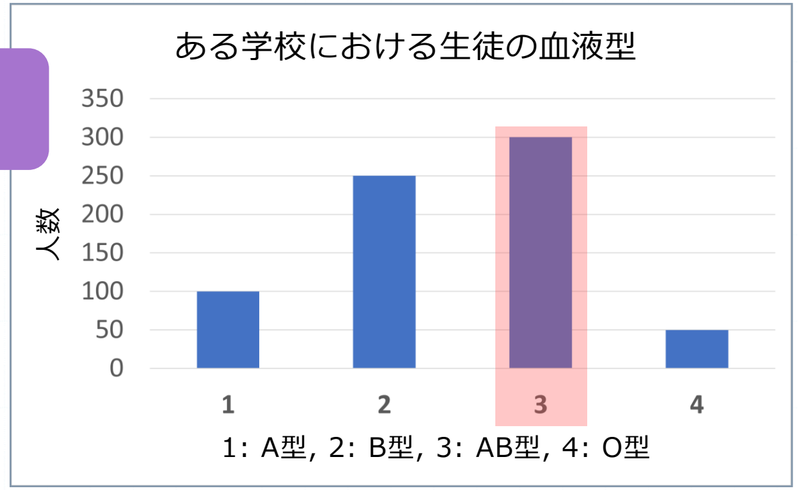
補足2: 最頻値

最頻値 (mode)

最も頻繁に出現する値のことを**最頻値**という

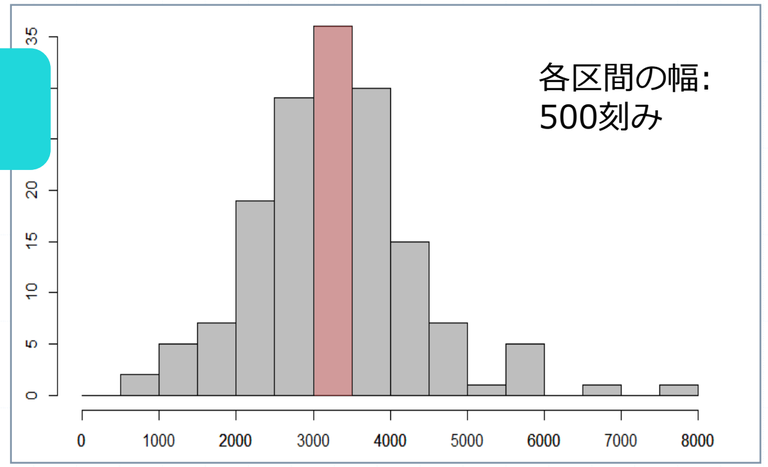
※ **量的変数**の場合，　最頻値 ＝ ヒストグラムの最も高い棒の**底辺の中点**，
または**その中に入るデータの平均値**

質的変数



最頻値: 3

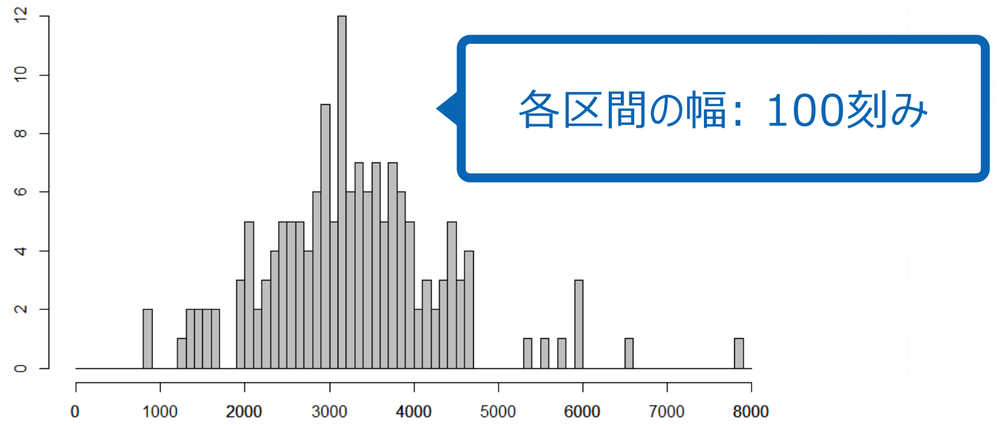
量的変数



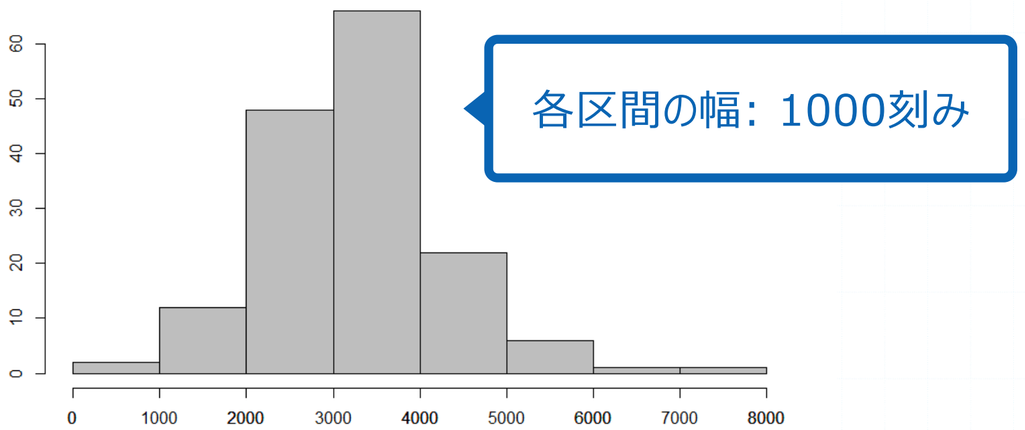
最頻値: $\frac{3000+3500}{2}$ ＝ 3250 (底辺の中点の場合)

補足2: 最頻値のコメント

- (頻度の意味で) 分布の**中心**を表す
- **量的変数**, **質的変数**のどちらにも適した代表値
- **量的変数**の場合, ヒストグラムの区間幅の設定の仕方で
  最頻値は変わることに注意



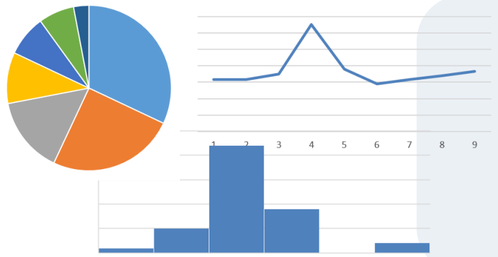
最頻値: $\frac{3100+3200}{2}$ = **3050** (底辺の中点の場合)



最頻値: $\frac{3000+4000}{2}$ = **3500** (底辺の中点の場合)

今日のまとめ

- **データの特徴を調べるには…**



グラフで視覚的に!!

＋

要約統計量で数値的に!!
(平均値や分散など)

データを要約

- **ヒストグラム (分布の形)**
- **要約統計量: 平均値, 中央値, 最頻値** (代表値: 分布の**中心**の位置)

変数の種類などで使い分け

グラフと要約統計量はどちらも重要!!