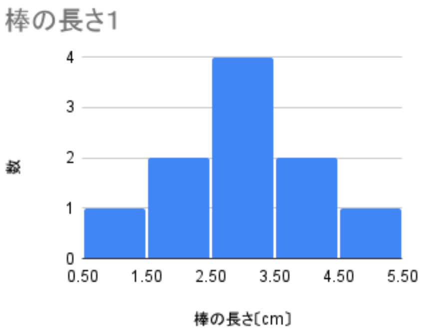
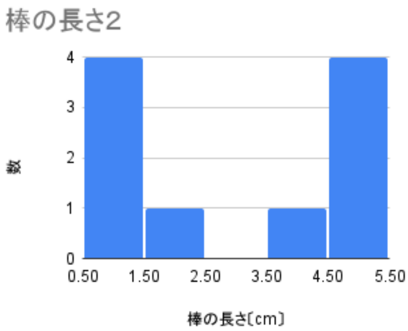
データサイエンス入門　授業プリント No.9

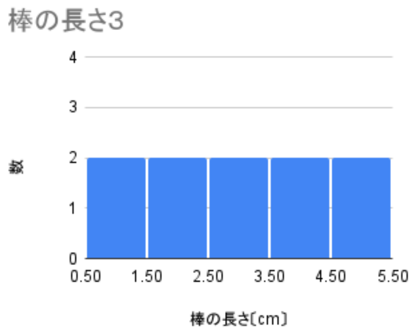
【データの代表】（その２）

平均値以外にデータを代表するような統計量は…

平均値以外の統計量というと，最大値や最小値といったものが思いつくかもしれませんが，データの特徴を表す重要なものとして「ばらつき」があります。例えば，長さ 1cm,2cm,3cm,4cm,5cm の小さな棒 10 本セットが３組あり，平均がすべて 3cm だとします。その３組のヒストグラムが次のようになった場合，「ばらつき」はどうでしょう？







どのくらいバラバラになっているのか，逆を言うとどのくらい均一になっているのかを表す統計量を見てみましょう。

例１．バス停にバスが到着する時刻の「ばらつき」

6 時 30 分着のバスの最近 1 週間の到着時刻が次の表のようになっていた場合を考えて見ましょう。

6 時 30 分着のバスの最近 1 週間の到着時刻の記録（分のみ）

曜日	日	月	火	水	木	金	土	平均値
到着時刻	27	34	29	31	33	34	29	31
平均値との差								

→到着時刻を平均すると 31 分になることがわかります。平均値からの差（ずれ）はすぐに求めることができる。この平均値との差のことを（　　　　　　）という。しかし，この値は平均すると０（ゼロ）になってしまうので，平均的な差（ずれ）を差の（　　　　　　）では表すことができない。

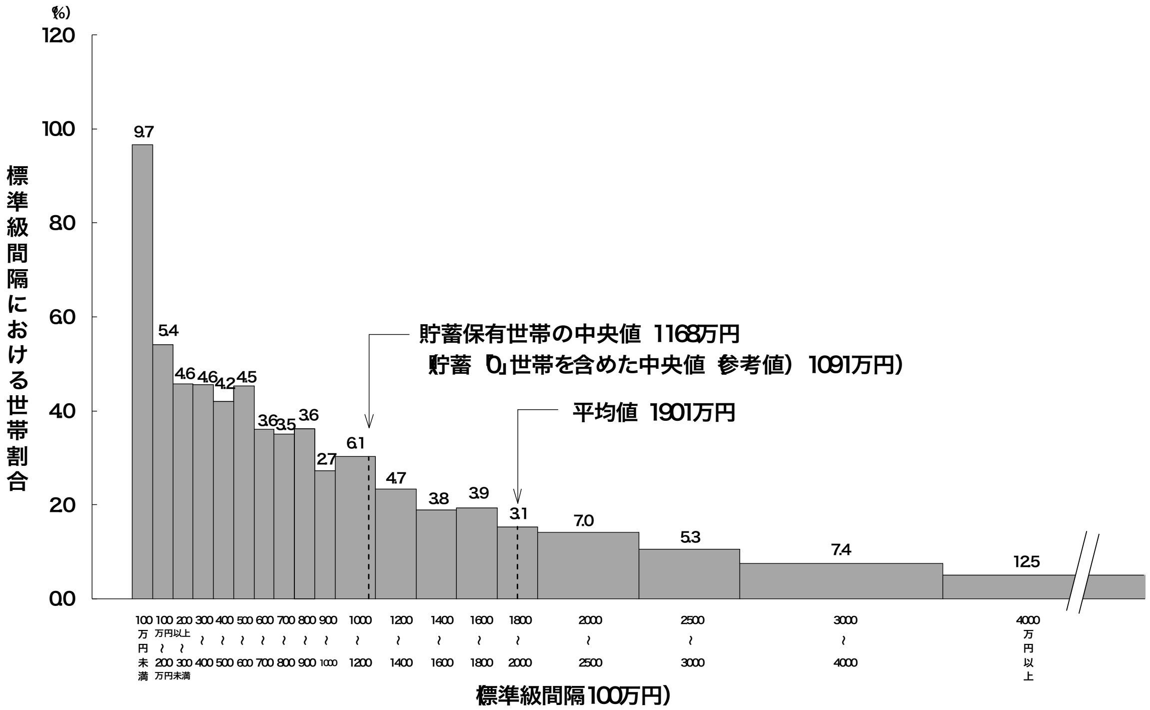
→そこで，差（ずれ）がプラス・マイナスで打ち消し合わないようにする工夫として平均値との差を（　　　　　　）すると，平均値が０にならないようにすることができる。こうして得られた平均値を（　　　　　　）という。例であれば（　　　　）分 ²

→そのままだと単位が異なるので，（　　　　　　）の平方根をとって単位を元に戻すことができる。この値を（　　　　　　　）といい，重要な統計量である。例だと約（　　　）分

（　　　　　　　　） $=\sqrt{(\qquad\qquad)}=$

［練習７］例に挙げた 6 時 30 分着のバスの最近 1 週間の到着時刻の記録の表を，Classroom で配付するスプレッドシートの空欄に計算式を入力して完成させてください。

例２．貯蓄現在高の分布の「ばらつき」



注１）貯蓄保有世帯の中央値とは、貯蓄「０」世帯を除いた世帯を貯蓄現在高の少ない方から順番に並べたときに、ちょうど中央に位置する世帯の貯蓄現在高をいう。
注２）標準級間隔100万円（貯蓄現在高1000万円未満）の各階級の度数は縦軸目盛りと一致するが、貯蓄現在高1000万円以上の各階級の度数は階級の間隔が標準級間隔よりも広いため、縦軸目盛りとは一致しない。

貯蓄現在高階級別世帯分布（2 人以上世帯）　（2022 年家計調査）

例 1 の場合，平均値を元に「ばらつき」を表す（　　　　　　　　）を求めていました。つまり，分布が平均値を代表値にできる（　　　　　　　　）に近い場合に使うものになります。しかし，以前紹介した貯蓄現在高のような偏った分布では，代表値としては平均値よりも（　　　　　　　）の方が適しています。

→例２のような偏った分布の「ばらつき」を表すには，（　　　　　　　　　　　）を使う。これは，データを大きさ順に並べたときの 4 分の 1～4 分の 3 の間の範囲をいう。つまり，この範囲には半数のデータが含まれることになる。

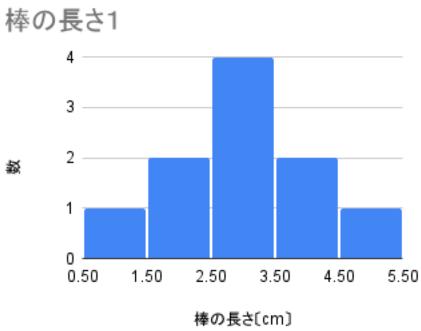
［考えてみよう７］例２の貯蓄現在高の四分位範囲を求めてみましょう。グラフの各階級の上に表示されているのが世帯の割合です。

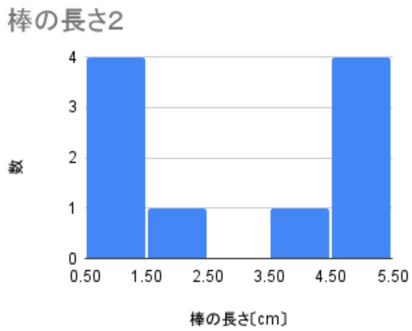
四分位範囲は，（　　　　　）万円～（　　　　　）万円の範囲

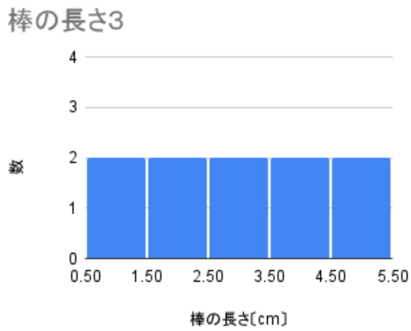
【データの代表】（その２）

平均値以外にデータを代表するような統計量は…

平均値以外の統計量というと，最大値や最小値といったものが思いつくかもしれませんが，データの特徴を表す重要なものとして「ばらつき」があります。例えば，長さ 1cm,2cm,3cm,4cm,5cm の小さな棒 10 本セットが３組あり，平均がすべて 3cm だとします。その３組のヒストグラムが次のようになった場合，「ばらつき」はどうでしょう？







どのくらいバラバラになっているのか，逆を言うとどのくらい均一になっているのかを表す統計量を見てみましょう。

例１．バス停にバスが到着する時刻の「ばらつき」

6 時 30 分着のバスの最近 1 週間の到着時刻が次の表のようになっていた場合を考えて見ましょう。　　6 時 30 分着のバスの最近 1 週間の到着時刻の記録（分のみ）

	曜日	日	月	火	水	木	金	土	平均値	
	到着時刻	27	34	29	31	33	34	29	31	
（偏差）→	平均値との差	-4	+3	-2	0	+2	+3	-2	0	
	(平均値との差)^2	16	9	4	0	4	9	4	$\boxed{6.6}$	←（分散）

→到着時刻を平均すると 31 分になることがわかります。平均値からの差（ずれ）はすぐに求めることができる。この平均値との差のことを（　偏差　）という。しかし，この値は平均すると０（ゼロ）になってしまうので，平均的な差（ずれ）を差の（　平均値　）では表すことができない。

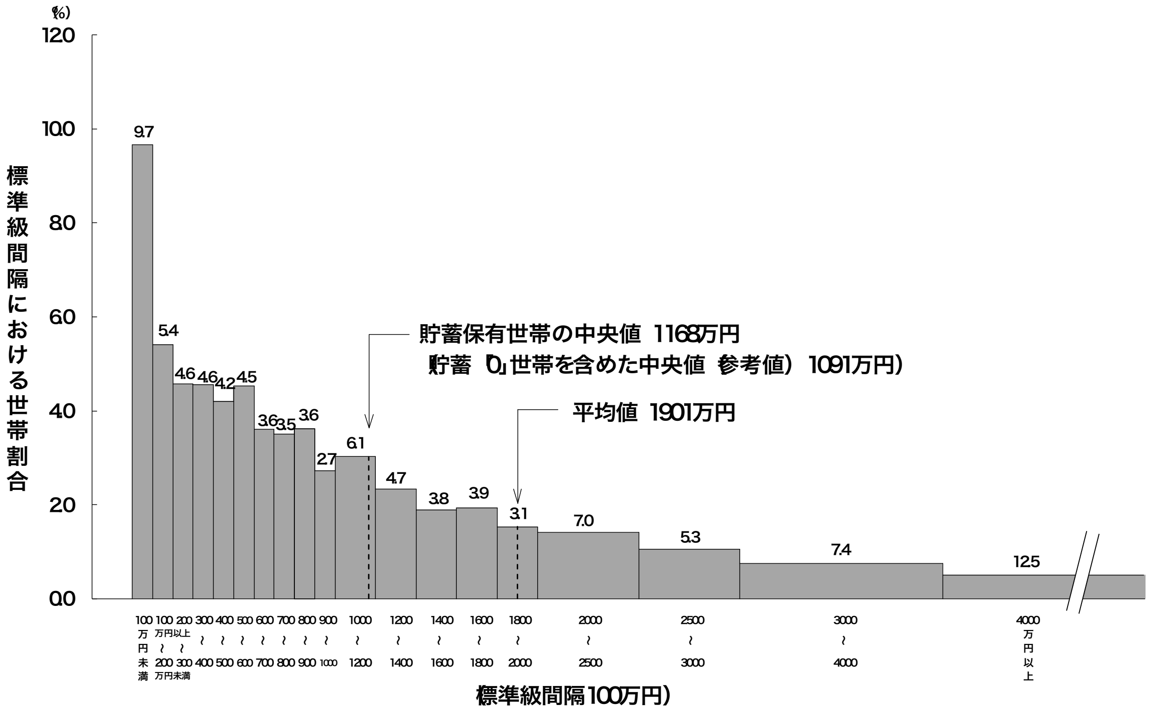
→そこで，差（ずれ）がプラス・マイナスで打ち消し合わないようにする工夫として平均値との差を（　２乗　）すると，平均値が０にならないようにすることができる。こうして得られた平均値を（　分散　）という。例であれば（　6.6　）分^2

→そのままだと単位が異なるので，（　分散　）の平方根をとって単位を元に戻すことができる。この値を（　標準偏差　）といい，重要な統計量である。例だと約（　2.6　）分

$$（\quad 標準偏差 \quad）=\sqrt{(\quad 分散 \quad)}=\sqrt{\frac{\{(個々のデータ)-(平均値)\}^2 の合計}{(データの個数)}}$$

［練習７］例に挙げた 6 時 30 分着のバスの最近 1 週間の到着時刻の記録の表を，Classroom で配付するスプレッドシートの空欄に計算式を入力して完成させてください。

例２．貯蓄現在高の分布の「ばらつき」



注１）貯蓄保有世帯の中央値とは、貯蓄「０」世帯を除いた世帯を貯蓄現在高の少ない方から順番に並べたときに、ちょうど中央に位置する世帯の貯蓄現在高をいう。
注２）標準級間隔100万円（貯蓄現在高1000万円未満）の各階級の度数は縦軸目盛りと一致するが、貯蓄現在高1000万円以上の各階級の度数は階級の間隔が標準級間隔よりも広いため、縦軸目盛りとは一致しない。

貯蓄現在高階級別世帯分布（2 人以上世帯）　　（2022 年家計調査）

例 1 の場合，平均値を元に「ばらつき」を表す（　　標準偏差　　）を求めていました。つまり，分布が平均値を代表値にできる（　　正規分布　　）に近い場合に使うものになります。しかし，以前紹介した貯蓄現在高のような偏った分布では，代表値としては平均値よりも（　　中央値　　）の方が適しています。

→例２のような偏った分布の「ばらつき」を表すには，（　　四分位範囲　　）を使う。これは，データを大きさ順に並べたときの 4 分の 1～4 分の 3 の間の範囲をいう。つまり，この範囲には半数のデータが含まれることになる。

［考えてみよう７］例２の貯蓄現在高の四分位範囲を求めてみましょう。グラフの各階級の上に表示されているのが世帯の割合です。

四分位範囲は，（　４００　）万円～（　２５００　）万円の範囲