

# 全体の流れ

## 1つのデータを読む

- 比較する
- 分解する
- 比率で見る
- ○○あたりで見る
- データの定義
- データの前提条件

## 大量のデータを読む

- 平均値、中央値
- 分布
- ヒストグラム
- 標準偏差
- 傾向
- 関係性

## データを正しく読む

- グラフの注意点
- サンプルの注意点
- データの偏り
- 異常値、欠損値
- 確証バイアス
- フェルミ推定

# 今回のポイント

---

大量データを読むコツは、  
代表値と分布

# 代表値と分布

---

1. データを1つの数値で代表する（代表値）
2. 代表値では分からない「ばらつき」を知る（分布）

## 代表値

---

- 平均値
- 中央値
- 最頻値



## 分布

---

- ヒストグラム
- 標準偏差
- 混合分布
- パレート図

# 大量のデータを読み解く

1. 代表値(1) 平均値

2. 代表値(2) 中央値

3. 代表値(3) 最頻値

4. 分布(1) ヒストグラム

5. 分布(2) 標準偏差

6. 分布(3) 混合分布

7. 分布(4) パレート図

8. まとめ(1) 代表値と分布

9. 傾向(1) 推移

10. 傾向(2) ヒートマップ

11. 関係性(1) 相関分析

12. 関係性(2) 因果関係

13. 関係性(3) 第三因子

14. 関係性(4) 混合グループ

15. 関係性(5) 外れ値

16. まとめ(2) 傾向、関係性

# 今回のポイント

---

代表値

# 代表値と分布

1. データを1つの数値で代表する（代表値）
2. 代表値では分からない「ばらつき」を知る（分布）

## 代表値

---

- 平均値
- 中央値
- 最頻値



## 分布

---

- ヒストグラム
- 標準偏差
- 混合分布
- パレート図

# 代表値

先月の営業チームの成績を知りたい。  
営業マン1人あたりの販売数は  
どれくらいなんだろう？



調べます！

# 代表値

営業マン	販売数
A	3,500
B	2,500
C	2,000
D	1,500
E	1,500



# 代表値



そうですね、Aさんは3,500個、Bさんは  
2,500個、Cさんは・・・

1人ずつ説明されても、分かりづらい・・・  
大体どれくらいなんだ・・・？



# 代表値

---

## 1. 意味

- 集団の中心的傾向  
= たくさんのデータの「真ん中」ってどれくらい？

## 2. ポイント

- 真ん中、といっても色々あります
- **さまざまな視点から「真ん中」を考えることで、**  
その大量データの特徴が見えてくる
  - **平均値、中央値、最頻値・・・**

# 今回のポイント

---

平均値

# 平均値

---

## 1. 計算式

- データの合計 ÷ データの総数
- テストの結果が、30点、20点、10点、40点  
→ 平均点 =  $(30 + 20 + 10 + 40) \div 4人 = 25点$

## 2. ポイント

- とにかくシンプルで分かりやすい！
- データ分析では、まず平均を見ることが多い

# 代表値

営業マン	販売数
A	3,500
B	2,500
C	2,000
D	1,500
E	1,500



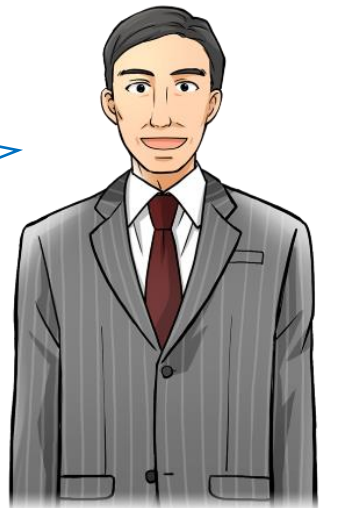
営業マン	販売数
A	3,500
B	2,500
C	2,000
D	1,000
E	1,000
平均値	2,000

# 平均値



先月の販売数は、  
平均すると1人2,000個です！

なるほど、先月より少し改善しているな！



# 今回のポイント

---

平均値

# 大量のデータを読み解く

1. 代表値(1) 平均値
2. 代表値(2) 中央値
3. 代表値(3) 最頻値
4. 分布(1) ヒストグラム
5. 分布(2) 標準偏差
6. 分布(3) 混合分布
7. 分布(4) パレート図
8. まとめ(1) 代表値と分布
9. 傾向(1) 推移
10. 傾向(2) ヒートマップ
11. 関係性(1) 相関分析
12. 関係性(2) 因果関係
13. 関係性(3) 第三因子
14. 関係性(4) 混合グループ
15. 関係性(5) 外れ値
16. まとめ(2) 傾向、関係性



# 今回のポイント

---

中央値

# 中央値

先月の営業チームの成績を知りたい。  
営業マン1人あたりの販売数は  
どれくらいなんだろう？



調べます！

# 中央値

営業マン	販売数
A	9,000
B	2,500
C	2,000
D	1,000
E	1,000
平均値	3,100

# 中央値



平均すると1人3,100個です！

あれ、ずいぶん多いな・・・？



# 中央値

営業マン	販売数
A	9,000
B	2,500
C	2,000
D	1,000
E	1,000
平均値	3,100

→ Aさんの販売数が異常に高い

みんな平均を下回っている

→ 平均値 = 真ん中の数値

# 平均値

---

## 1. 平均値のポイント

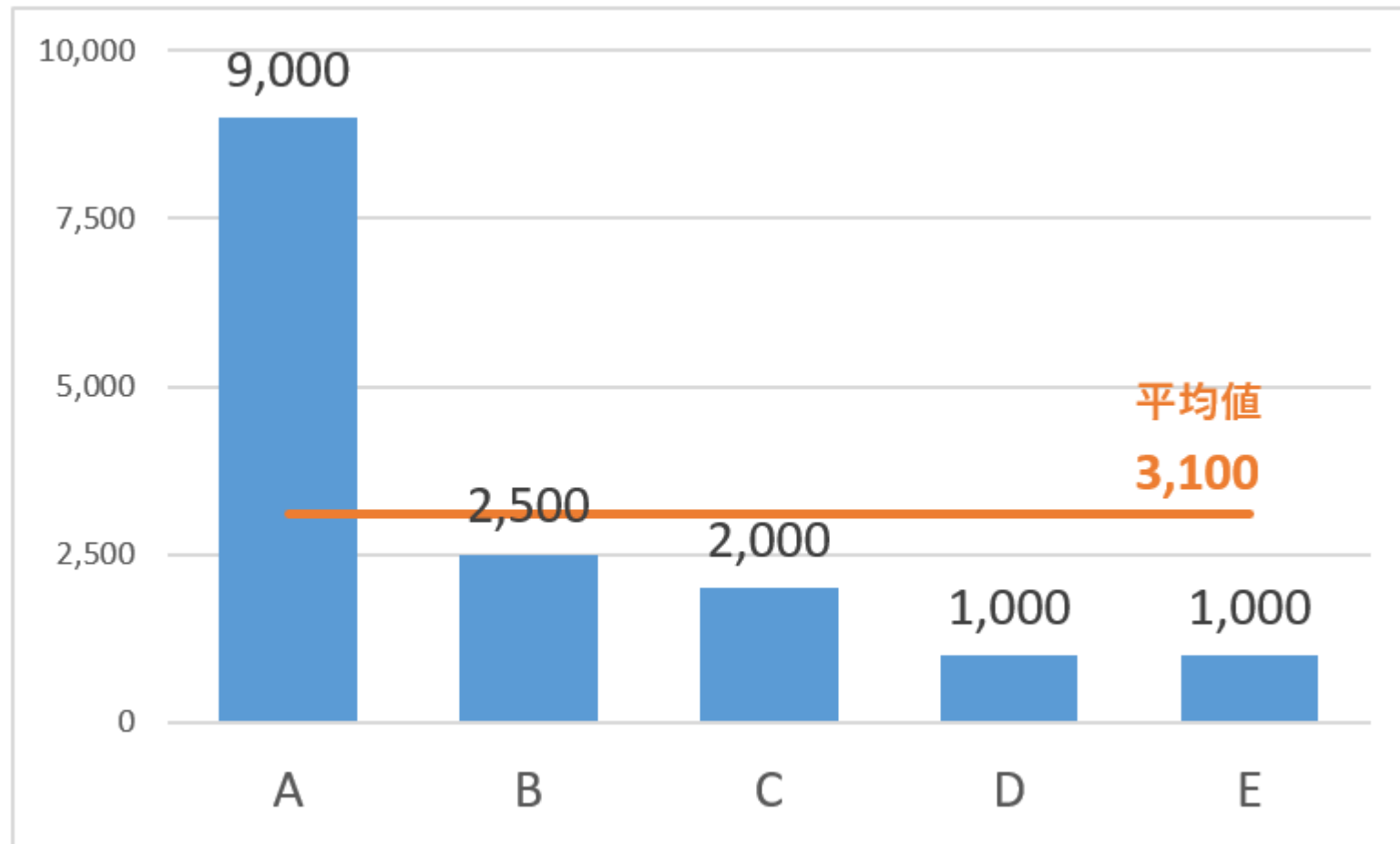
- とにかくシンプルで分かりやすい！
- データ分析では、まず平均を見ることが多い

## 2. ところが

- 1つ異常に大きい数値があると平均値が上がってしまう
  - 1つ異常に小さい数値があると平均値が下がってしまう
- 「真ん中」の数字といえるのか、疑問

# 平均値

平均値 3,100 は、真ん中というにはちょっと高い・・・



# 今回のポイント

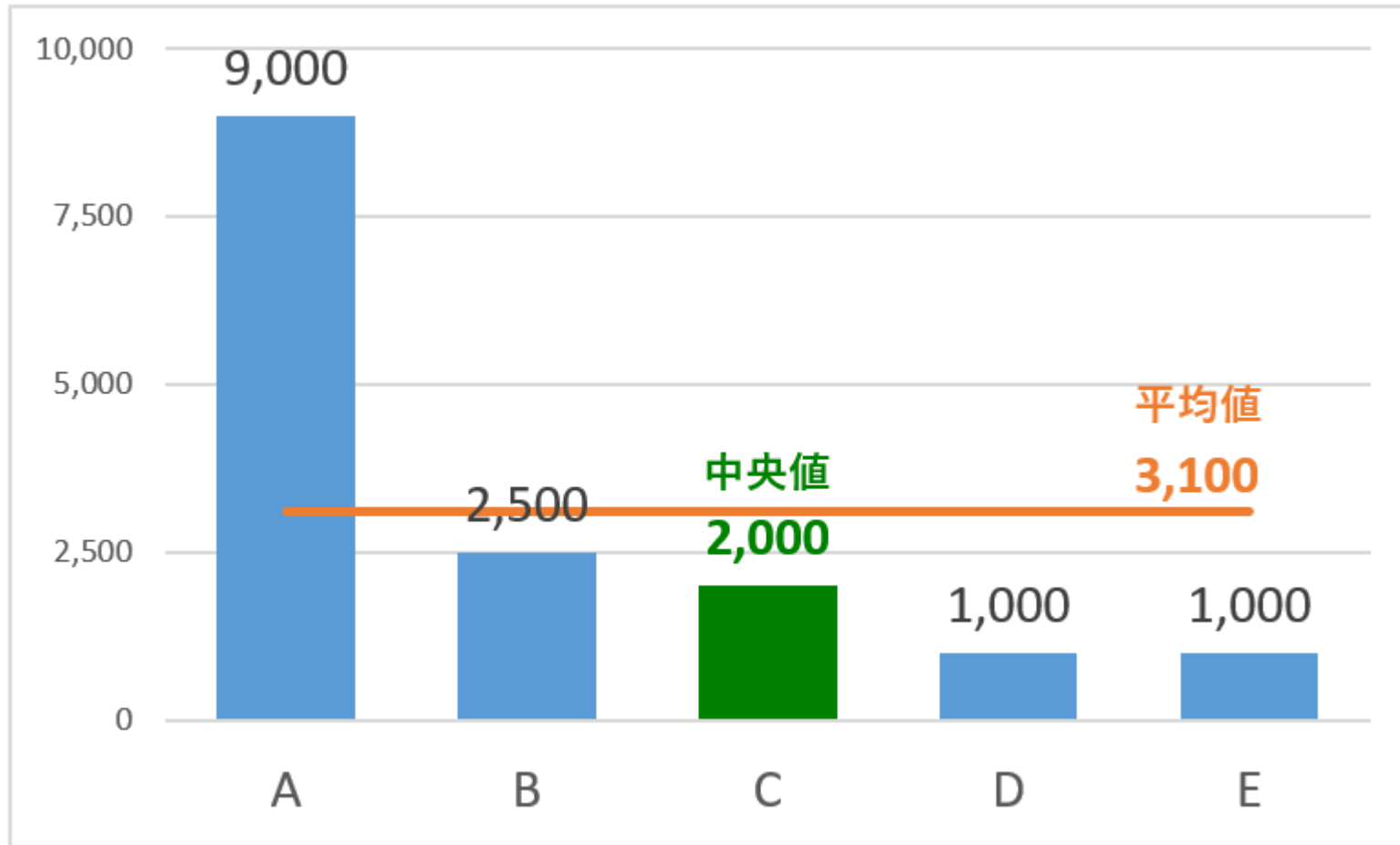
---

中央値



# 中央値

数字を大きい順に並べて、真ん中（3番目）の数値 = 2,000



# 中央値

---

## 1. 計算方法

- 数字を大きい順に並べて、真ん中（3番目）の数値

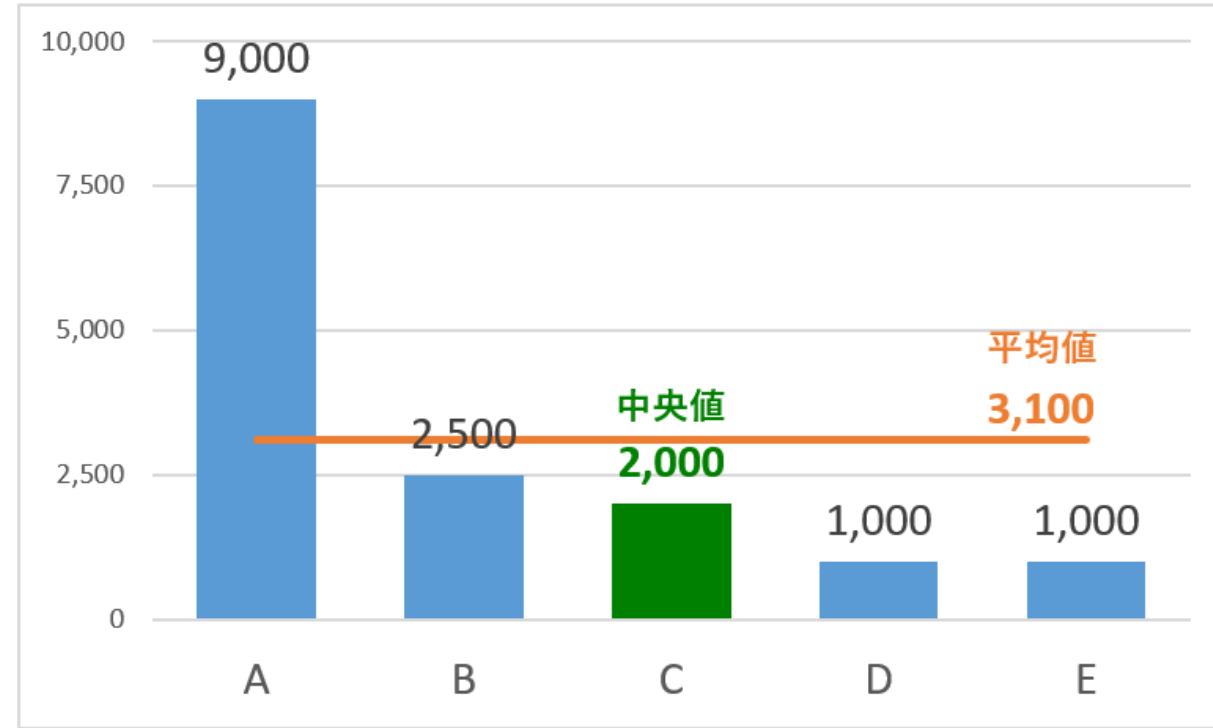
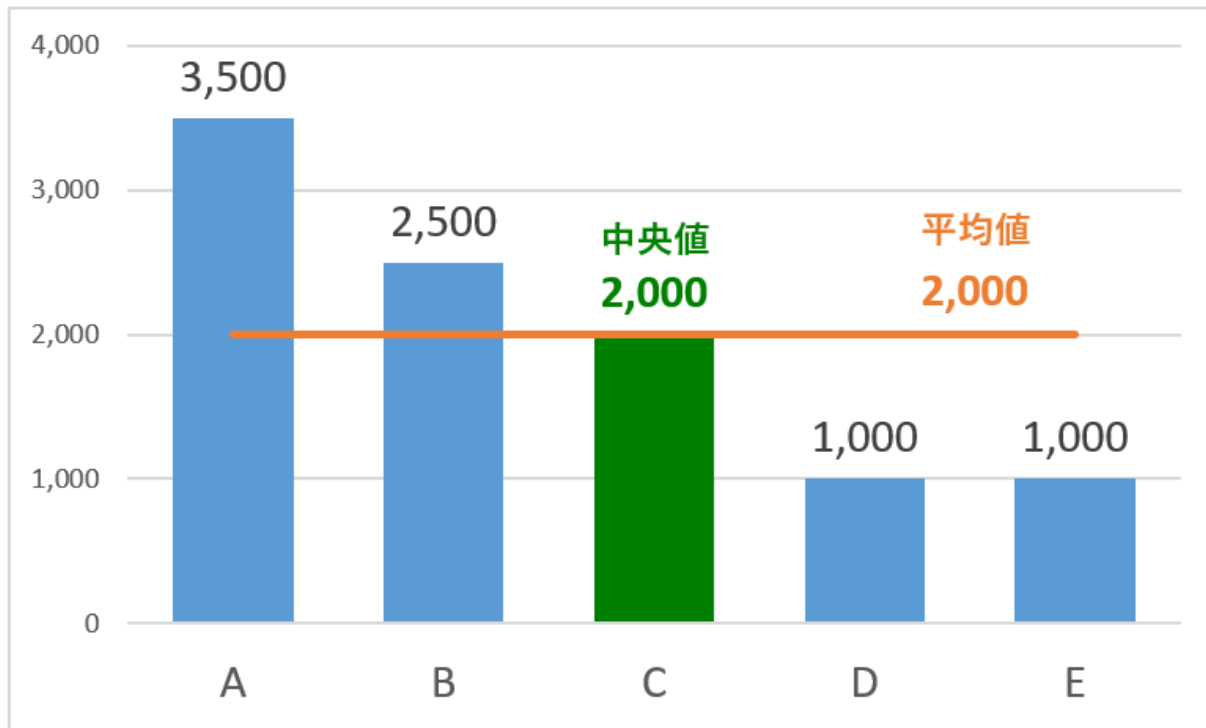
## 2. ポイント

- 1つ異常に大きい数字があっても、あまり影響しない

# 中央値


## ポイント

- 平均値と中央値がズレている場合は、異常値がある可能性あり



# 代表値

平均値と中央値の両方を「代表値」として示す

営業マン	販売数		営業マン	販売数
A	9,000		A	9,000
B	2,500		B	2,500
C	2,000		C	2,000
D	1,000		D	1,000
E	1,000		E	1,000
平均値	3,100		平均値	3,100
			中央値	2,000

# 中央値



平均すると1人3,100個ですが、Aさんの販売数が9,000個だったのが影響しています。  
中央値は2,000個になります

なるほど、Aさんすごいな！  
他の人も先月より増加しているようだな！



# 大量のデータを読み解く

1. 代表値(1) 平均値
2. 代表値(2) 中央値
3. 代表値(3) 最頻値
4. 分布(1) ヒストグラム
5. 分布(2) 標準偏差
6. 分布(3) 混合分布
7. 分布(4) パレート図
8. まとめ(1) 代表値と分布
9. 傾向(1) 推移
10. 傾向(2) ヒートマップ
11. 関係性(1) 相関分析
12. 関係性(2) 因果関係
13. 関係性(3) 第三因子
14. 関係性(4) 混合グループ
15. 関係性(5) 外れ値
16. まとめ(2) 傾向、関係性

# 今回のポイント

---

最頻値

# 最頻値

---

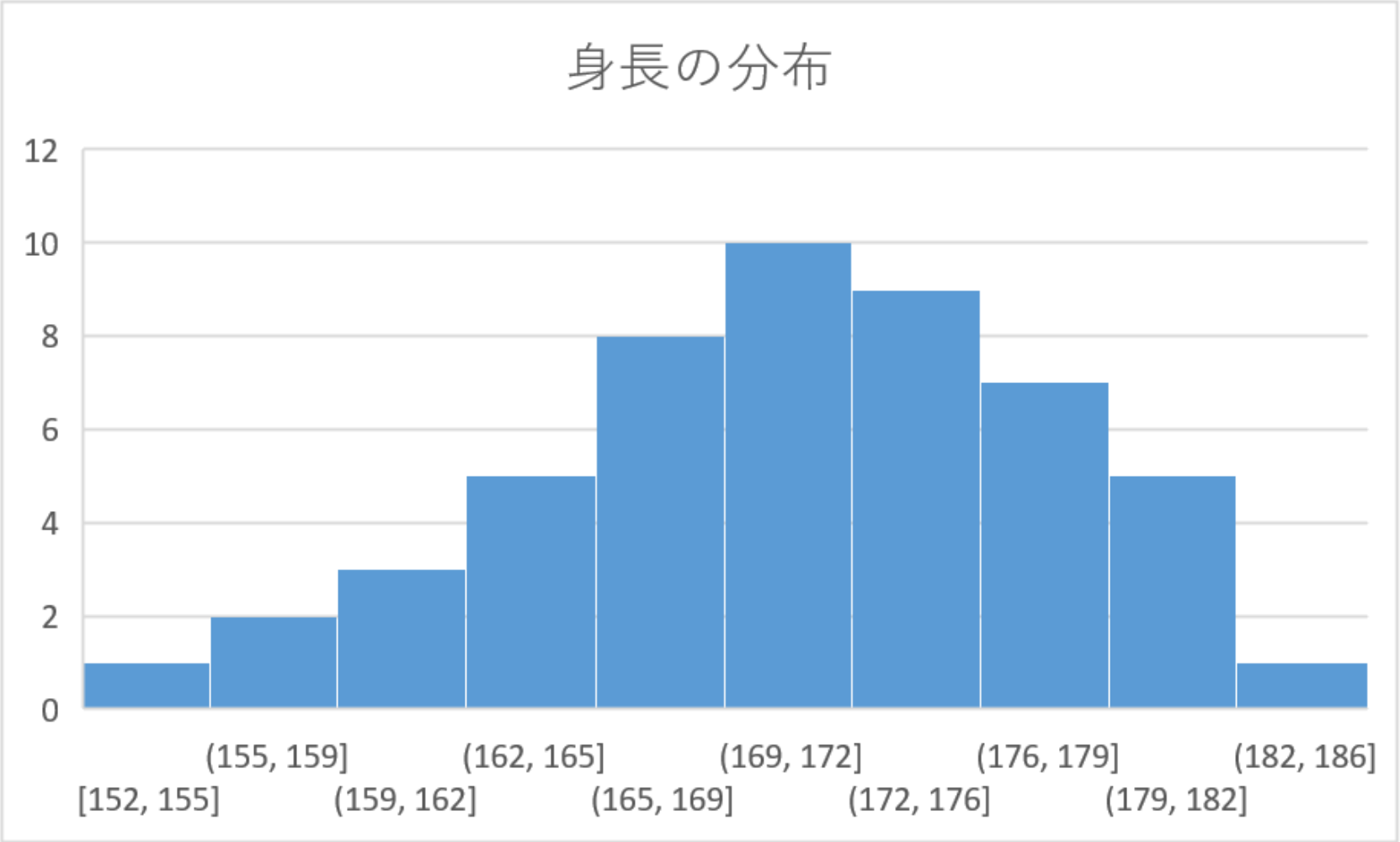
## 1. 最頻値

- もっとも多く出てくる値（頻度）
- ヒストグラムを作ると分かりやすい



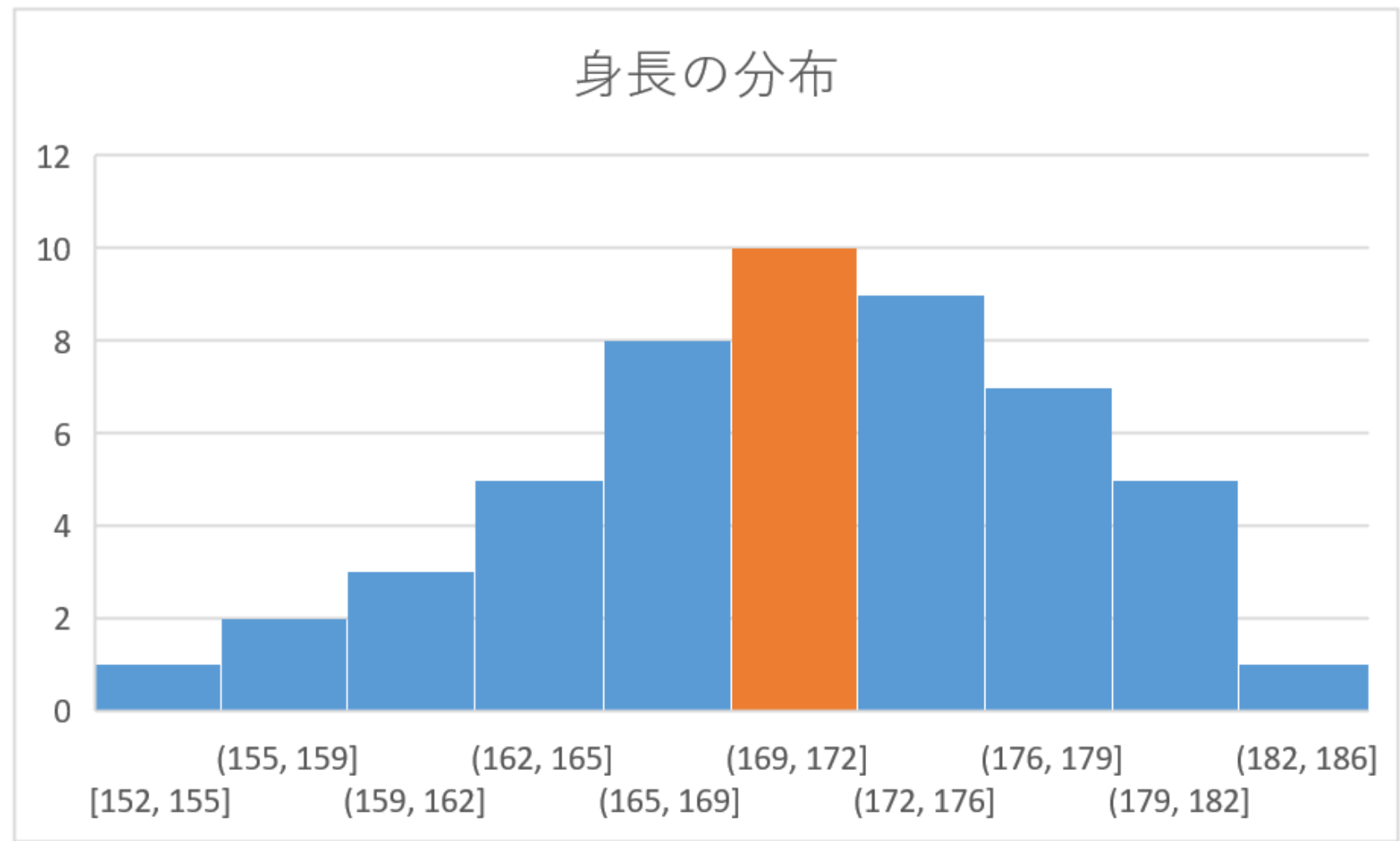
# 最頻値

男性の身長分布（152cm から 186cmまで） 縦軸は人数



# 最頻値

169cm から 172cm の人数が一番多い（最頻値）



# 今回のポイント

---

代表値まとめ

# 代表値まとめ

---

## 1. 今回紹介した代表値

- 平均値、中央値、最頻値

## 2. Excelで計算する場合、関数で簡単に計算できます

- 平均値                      AVERAGE関数
- 中央値                      MEDIAN関数
- 最頻値                      MODE関数

# 大量のデータを読み解く

1. 代表値(1) 平均値
2. 代表値(2) 中央値
3. 代表値(3) 最頻値
4. 分布(1) ヒストグラム
5. 分布(2) 標準偏差
6. 分布(3) 混合分布
7. 分布(4) パレート図
8. まとめ(1) 代表値と分布
9. 傾向(1) 推移
10. 傾向(2) ヒートマップ
11. 関係性(1) 相関分析
12. 関係性(2) 因果関係
13. 関係性(3) 第三因子
14. 関係性(4) 混合グループ
15. 関係性(5) 外れ値
16. まとめ(2) 傾向、関係性

# 今回のポイント

---

分布

# 代表値と分布

---

1. データを1つの数値で代表する（代表値）
2. 代表値では分からない「ばらつき」を知る（分布）

## 代表値

---

- 平均値
- 中央値
- 最頻値



## 分布

---

- ヒストグラム
- 標準偏差
- 混合分布
- パレート図

# 分布

---

## 1. 意味

- いわゆる「数字のばらつき」

## 2. 例えば

- 平均値が50点のテスト結果といっても、
    - 全員が50点
    - 半分が0点、残りの半分が100点
- 平均は同じでも  
意味は異なる

→ 分布（ばらつき）を見る必要がある



# 今回のポイント

---

ヒストグラム

# ヒストグラム

先月の営業チームの状況を知りたい。  
営業マン1人あたり販売数は、  
どれくらい「ばらつき」があるのだろうか？



調べます！

# ヒストグラム

	A	B	C
1			
2		営業マンNo.	販売数
3		0001	456
4		0002	496
5		0003	388
6		0004	434
7		0005	575
8		0006	697
9		0007	573
10		0008	511
11		0009	561
12		0010	468
13		0011	432
14		0012	564
15		0013	638

# ヒストグラム



そうですね・・・

700個近く販売している人もいれば、  
400個くらいの人もありますし・・・

うーん、もう少し分かりやすく  
教えてくれないかな・・・



# ヒストグラム

---

## 1. 分布

- いわゆる「数字のばらつき」

## 2. ポイント

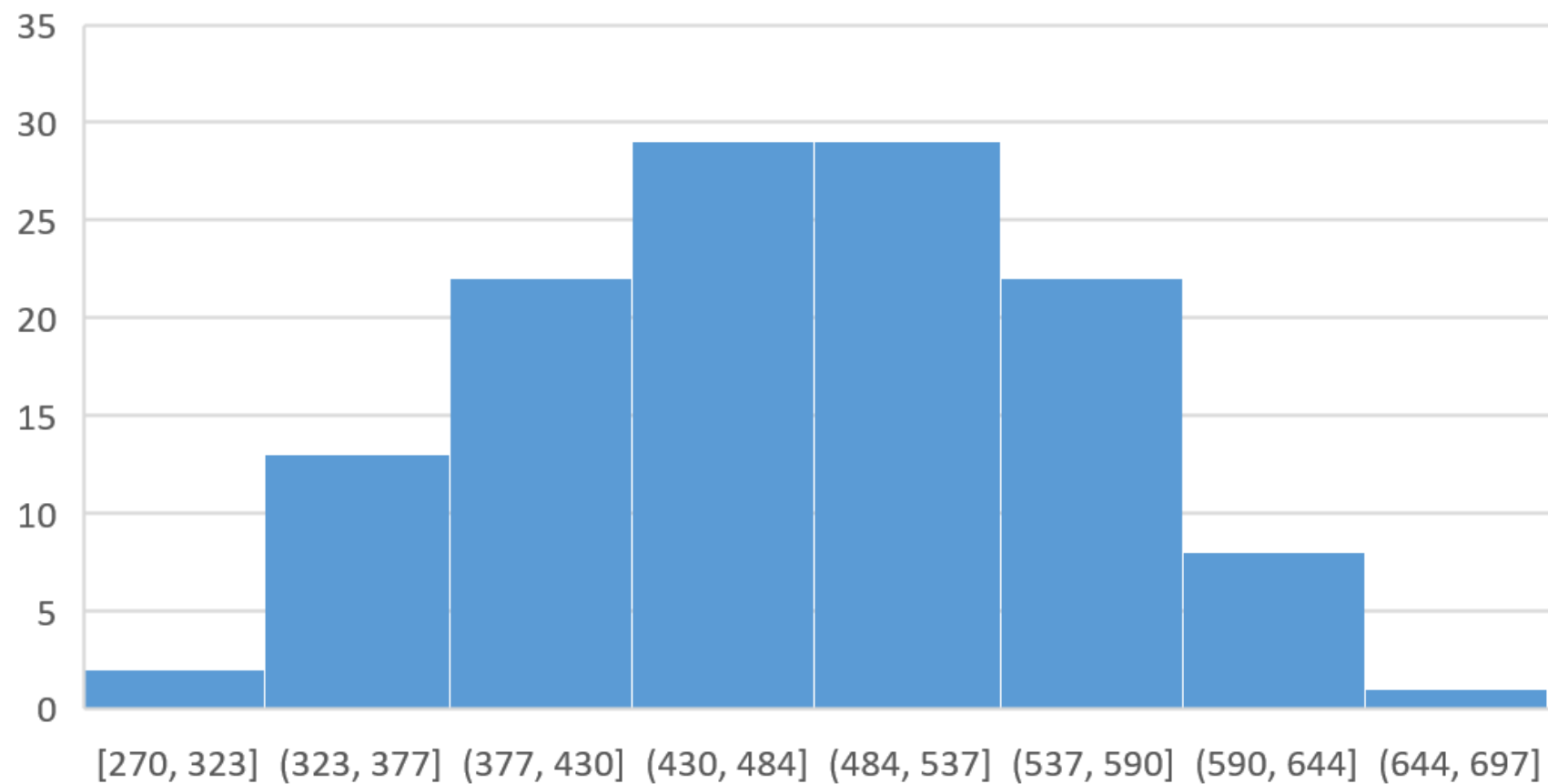
- 分布を説明するのは難しい！
- グラフや表をうまく活用して分かりやすく伝える
- よく使われる「ヒストグラム」を紹介

# ヒストグラム

	A	B	C
1			
2		営業マンNo.	販売数
3		0001	456
4		0002	496
5		0003	388
6		0004	434
7		0005	575
8		0006	697
9		0007	573
10		0008	511
11		0009	561
12		0010	468
13		0011	432
14		0012	564
15		0013	638

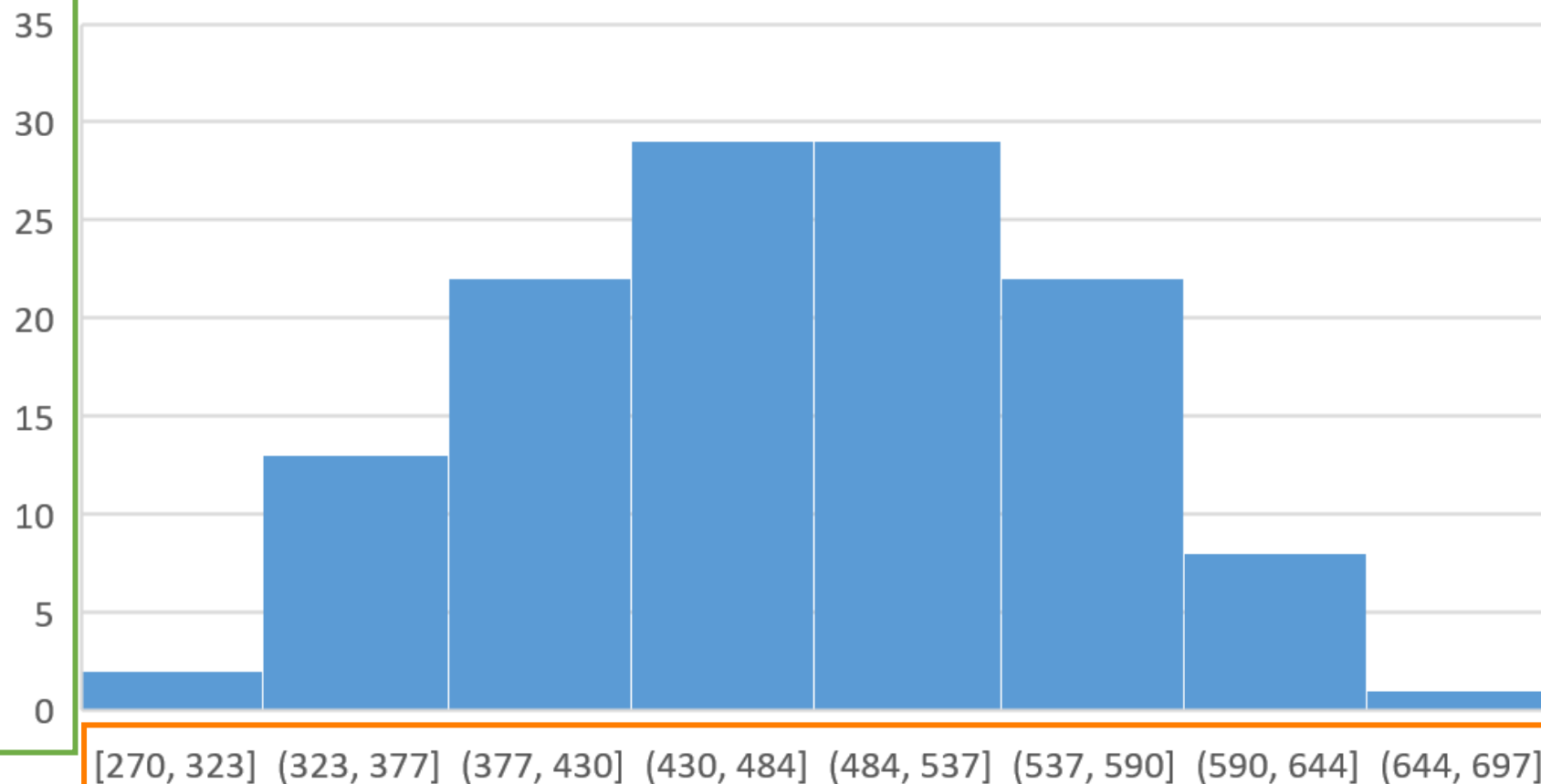
# ヒストグラム

営業マン1人あたり販売数の分布



# ヒストグラム

営業マン1人あたり販売数の分布



## 軸の意味

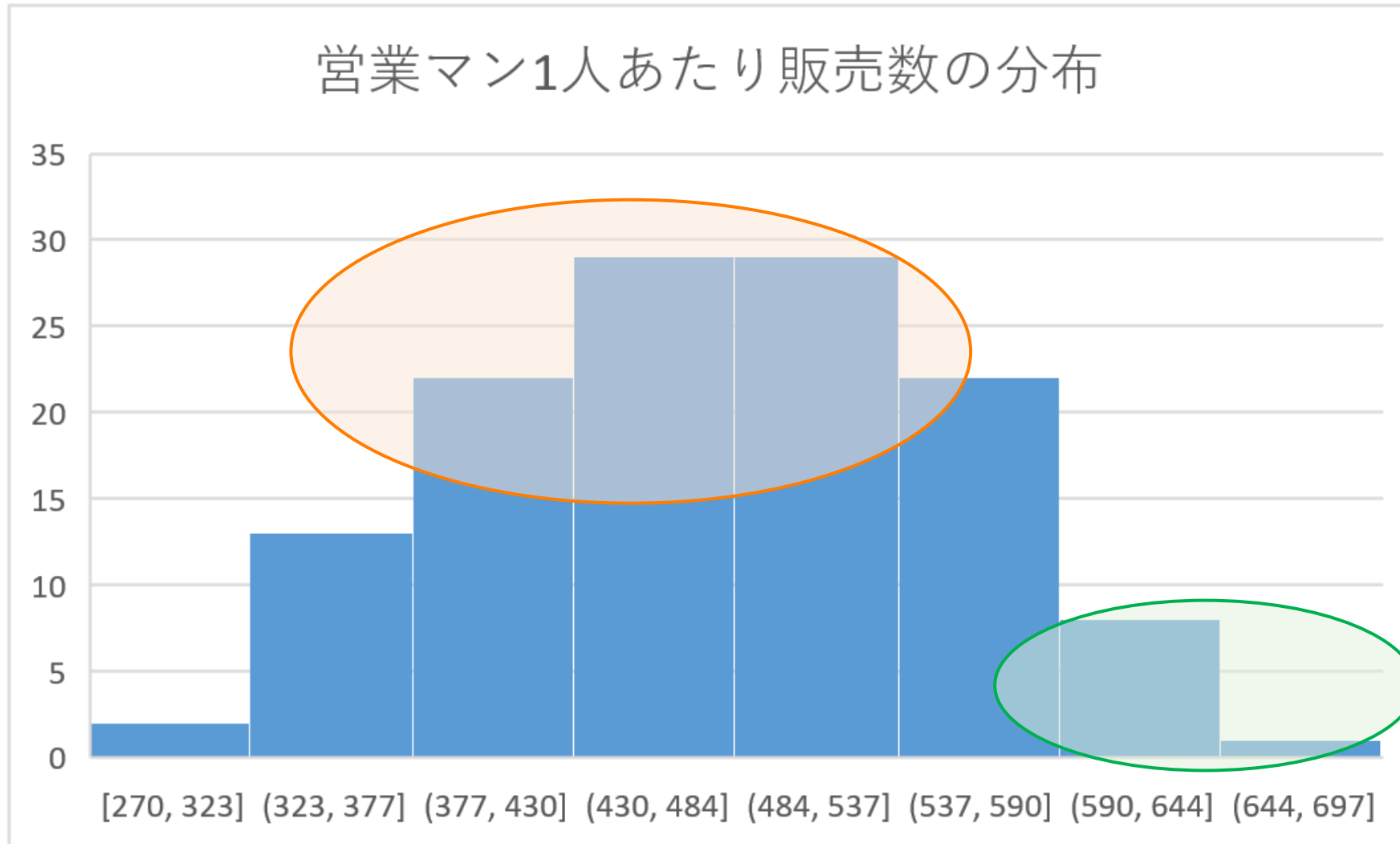
- ・ 横軸：販売数
- ・ 縦軸：人数

## グラフの見方

- ・ 270～323個 2人
- ・ 323～377個 13人
- ・ 377～430個 22人



# ヒストグラム



分布の見方 (1)

だいたい350～550個

くらいの販売数の  
営業マン多いな

分布の見方 (2)

600個以上の営業マン  
は9人だけか。

表彰してあげよう

# ヒストグラム

---

## 1. 分布

- いわゆる「数字のばらつき」

## 2. ポイント

- 分布を説明するのは難しい！
- グラフや表をうまく活用して分かりやすく伝える
- よく使われる「ヒストグラム」を紹介

# 大量のデータを読み解く

1. 代表値(1) 平均値
2. 代表値(2) 中央値
3. 代表値(3) 最頻値
4. 分布(1) ヒストグラム
5. 分布(2) 標準偏差
6. 分布(3) 混合分布
7. 分布(4) パレート図
8. まとめ(1) 代表値と分布
9. 傾向(1) 推移
10. 傾向(2) ヒートマップ
11. 関係性(1) 相関分析
12. 関係性(2) 因果関係
13. 関係性(3) 第三因子
14. 関係性(4) 混合グループ
15. 関係性(5) 外れ値
16. まとめ(2) 傾向、関係性

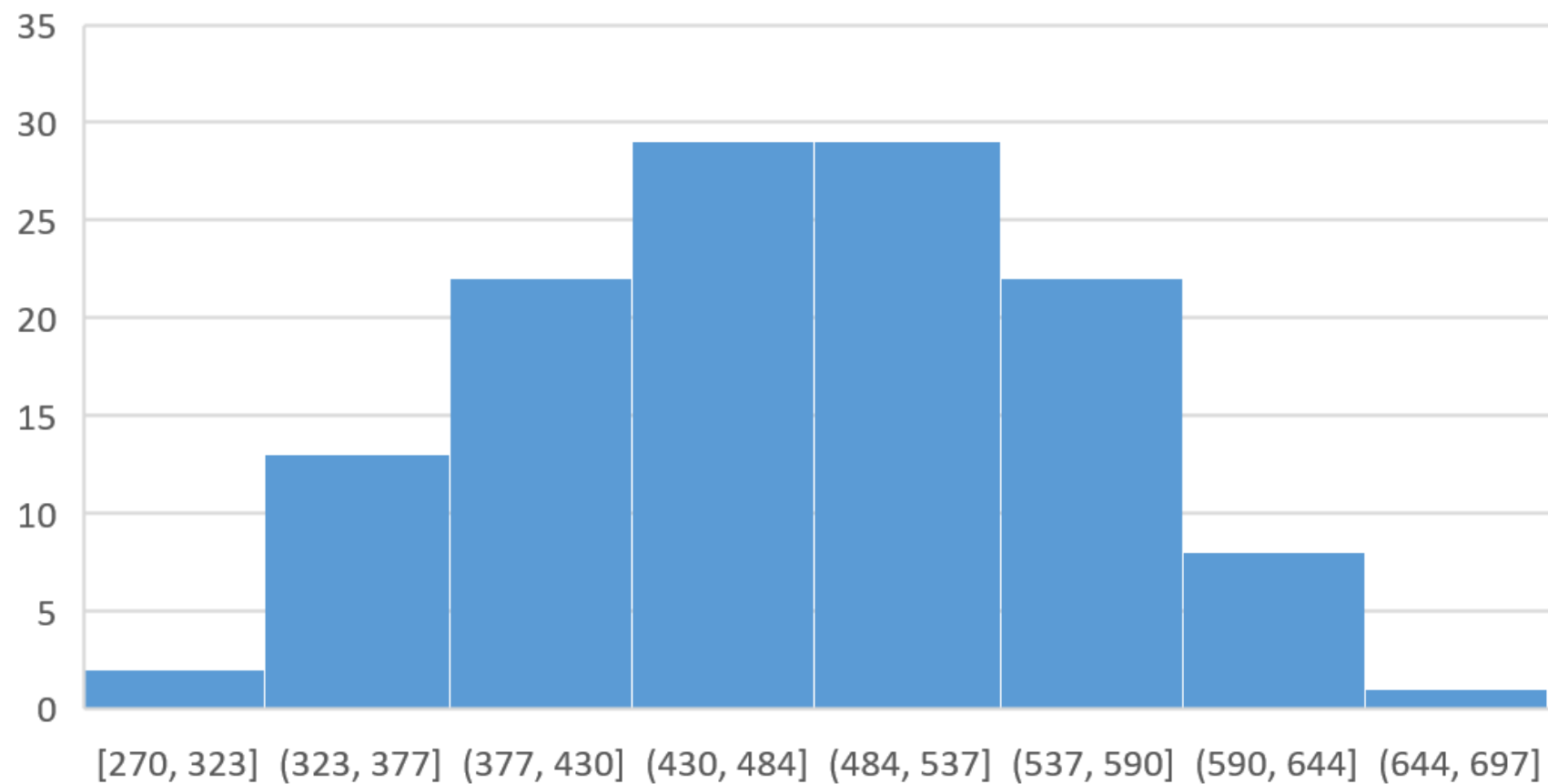
# 今回のポイント

---

標準偏差

# ヒストグラム

営業マン1人あたり販売数の分布



# 今回のポイント

---

「大体どの範囲に収まっている？」

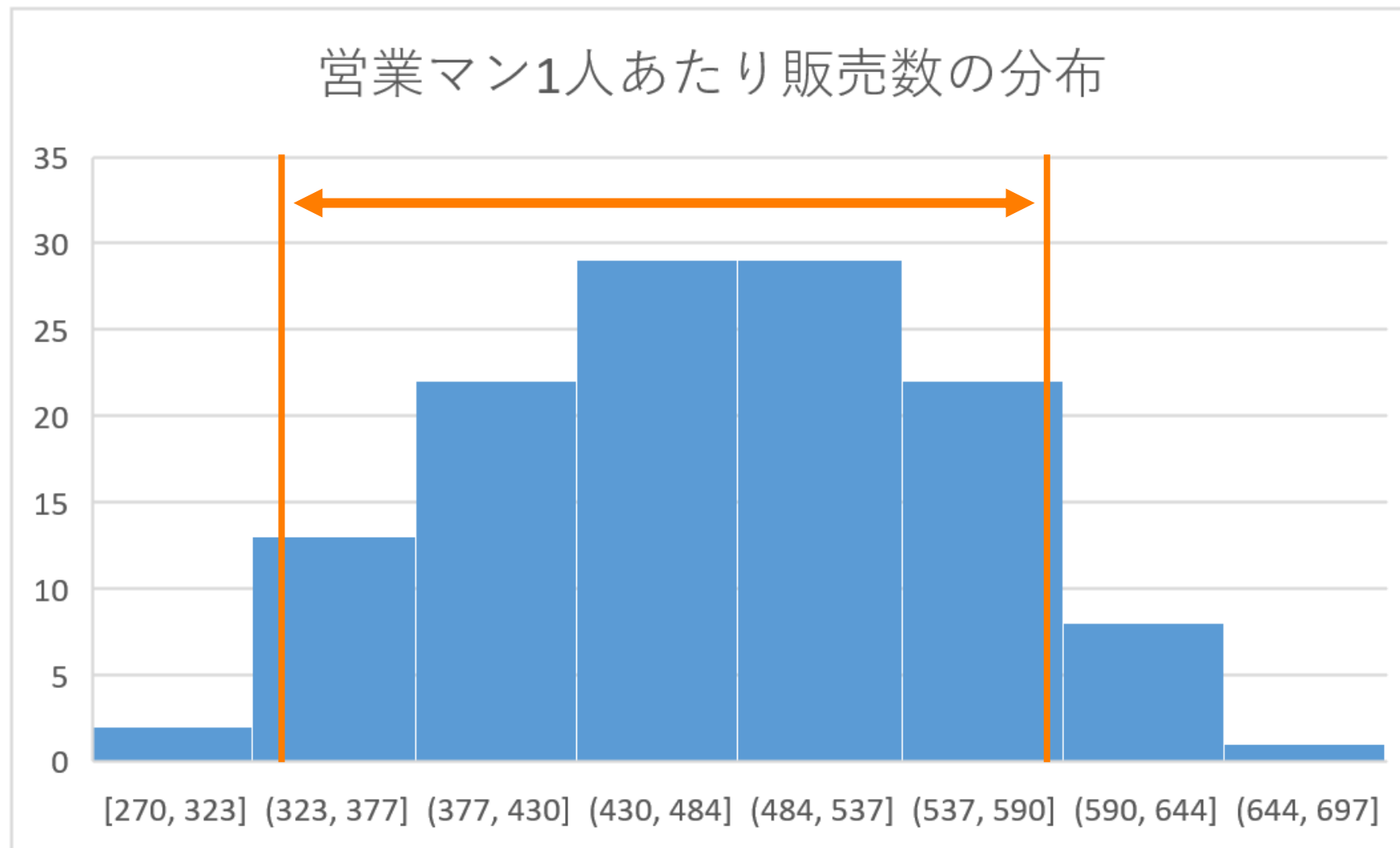
# 標準偏差

先月の営業チームの状況を知りたい。  
営業マン1人あたり販売数は、  
大体どの範囲に収まっているのだろう？



えっ・・・

# 標準偏差



だいたい・・・

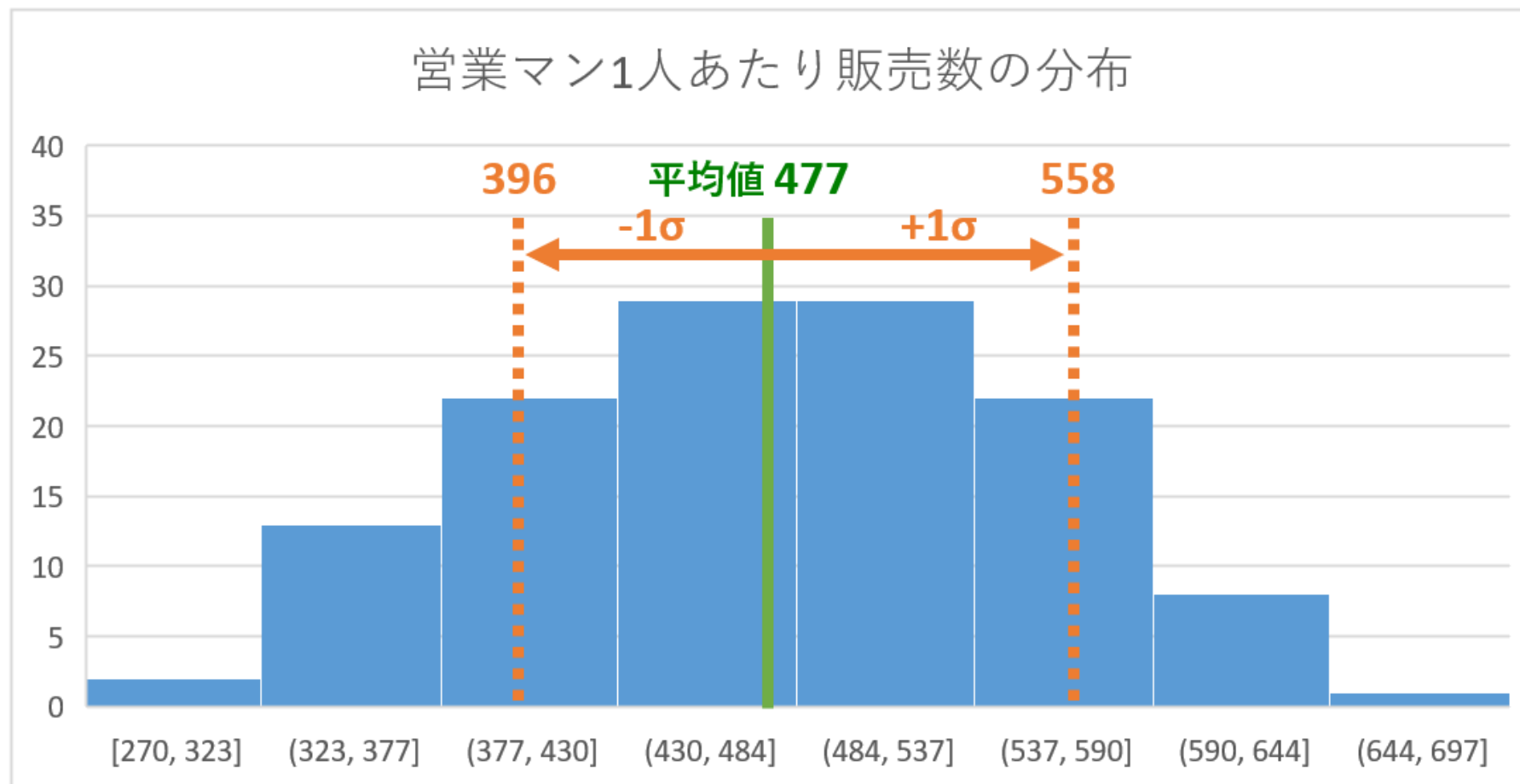
この範囲に収まっている？

→ もっと数字で  
説明したい



# 標準偏差

「396～558の範囲に、約70%が収まっている」



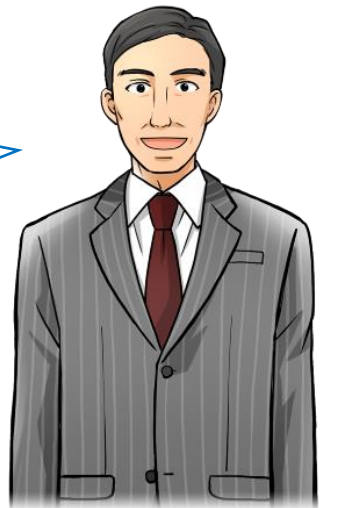
# 標準偏差



約70%の営業マンが、  
約400～550個を販売しています！

なるほど！

営業マンの販売数のイメージがつかめたぞ！



# 今回のポイント

---

標準偏差を使って範囲を計算する

# 標準偏差

---

## 1. 意味

- 数字のばらつきを示す指標の1つ
- $\sigma$ （シグマ）と呼ばれる

## 2. 計算方法

- 手で計算するとかなり面倒
- Excelだと、STDEV.P関数などで簡単に計算できます

# 標準偏差

	A	B	C	D	E	F
1						
2						
3						
4						
5						
6						
7						
8						
9						
10						
11						
12						
13						
14						
15						

営業マンNo.      販売数

0001	456
0002	496
0003	388
0004	434
0005	575
0006	697
0007	573
0008	511
0009	561
0010	468
0011	432
0012	564
0013	638

平均値

477 =AVERAGE(C:C)

標準偏差

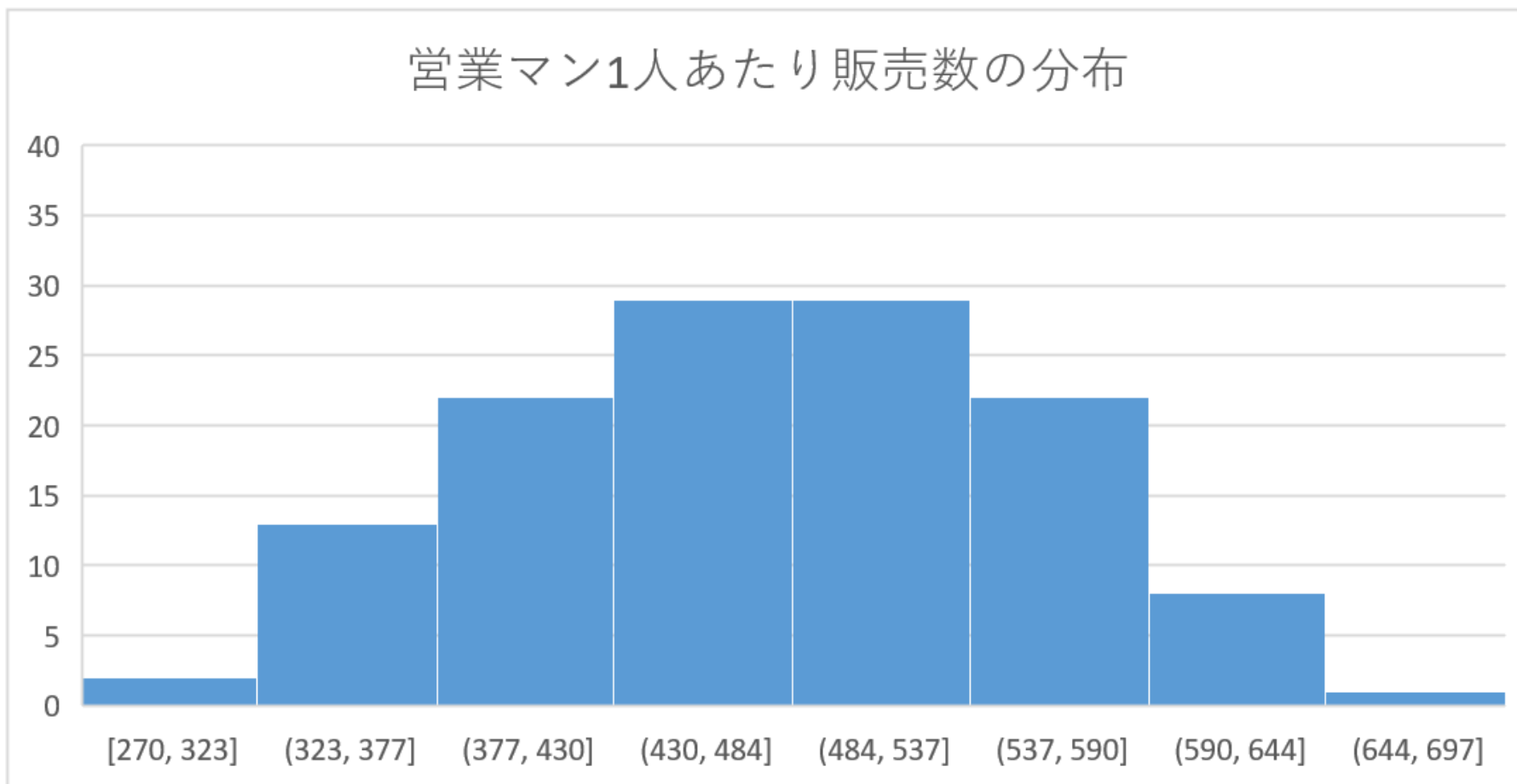
81 =STDEV.P(C:C)

まず、販売数データの  
平均値と標準偏差 (σ)を計算する  
(AVERAGE関数、STDEV.P関数)

# 標準偏差

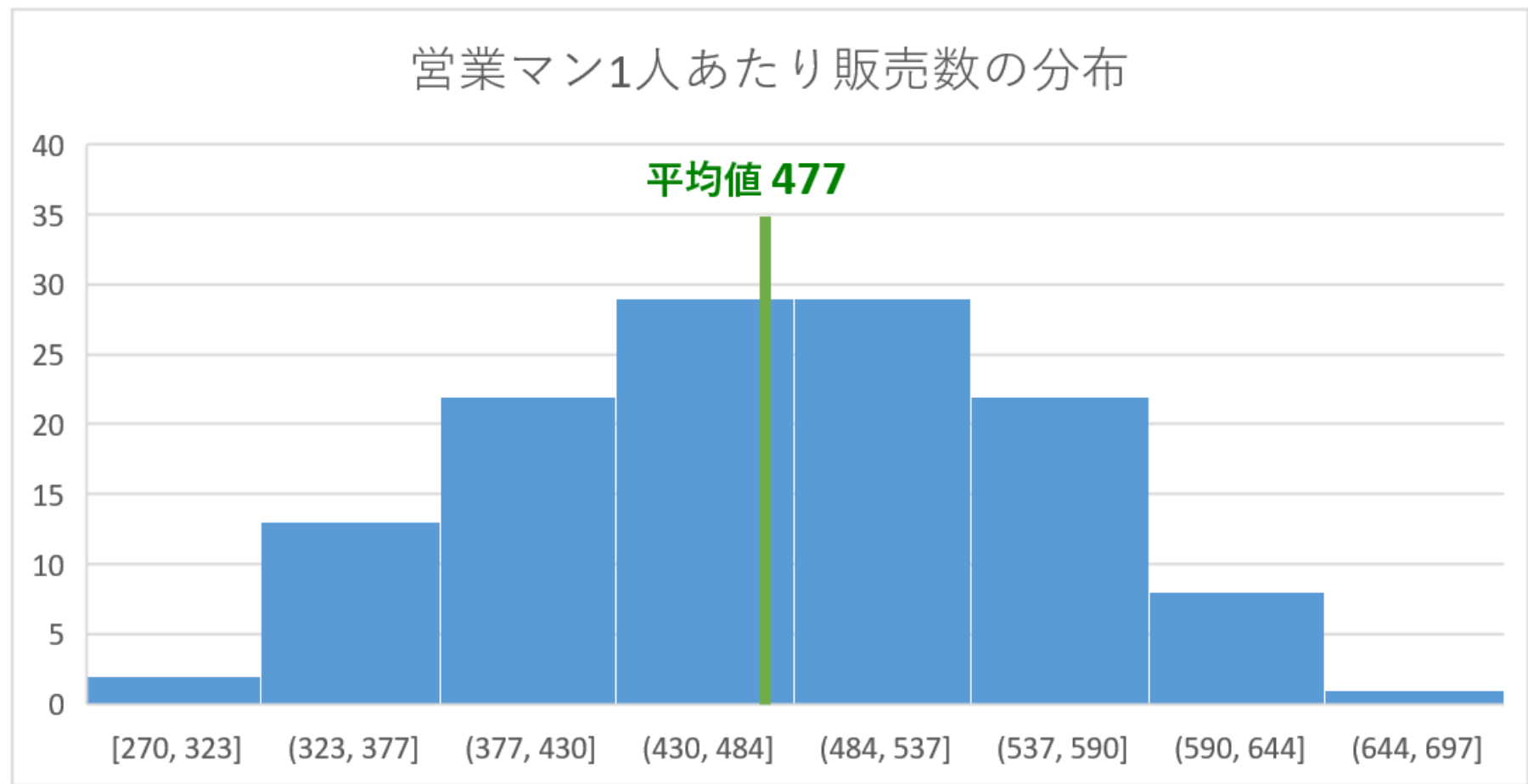
このヒストグラムに、

営業マン1人あたり販売数の分布



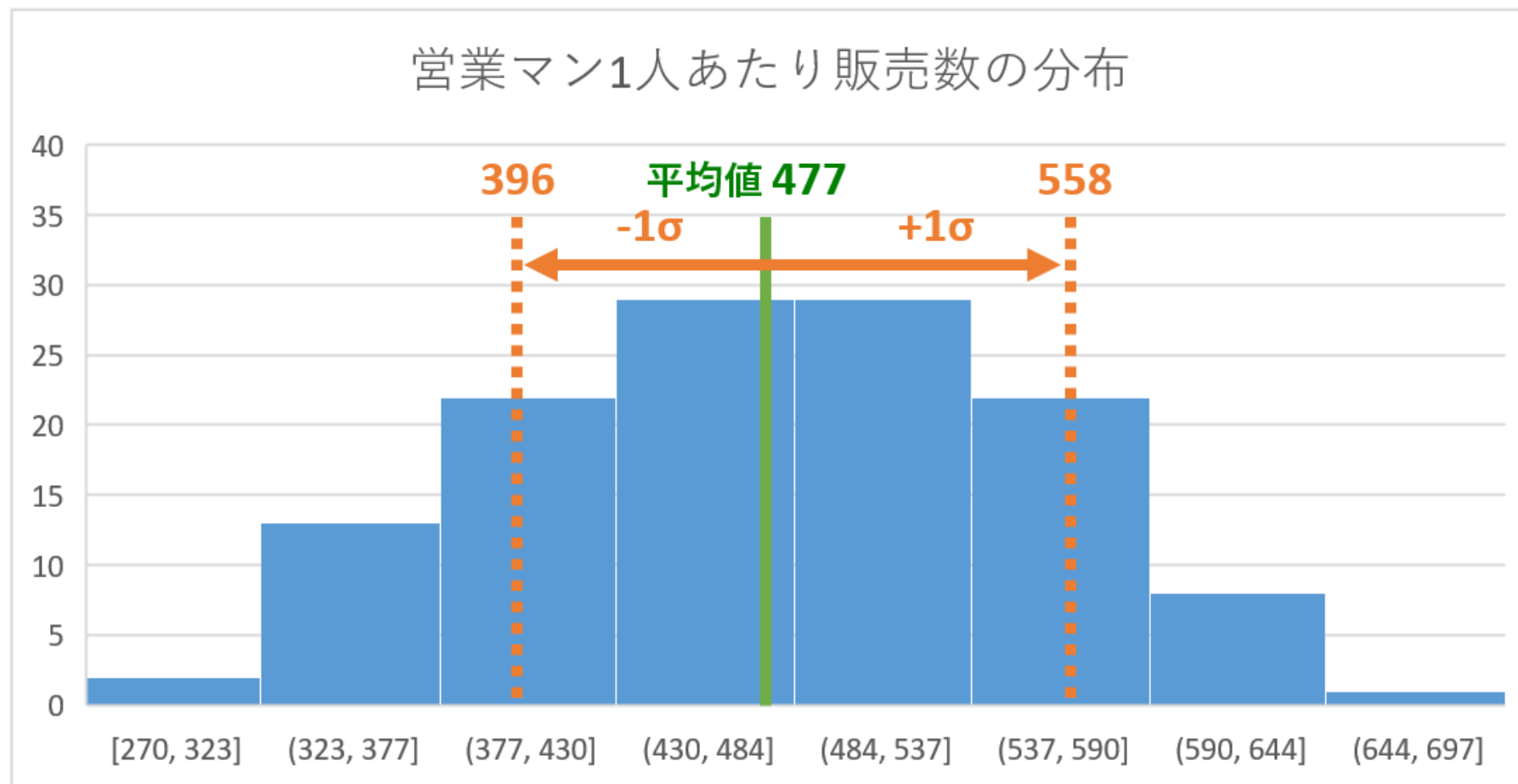
# 標準偏差

平均値（=477）を中心にして、



# 標準偏差

標準偏差（=81）をプラスマイナスして幅をつくる（396～558）





# 標準偏差

---

## 1. 覚えておく と 便利な、範囲イメージ

- 平均値  $\pm 1\sigma$       約68.2%      (だいたいこの範囲)
- 平均値  $\pm 2\sigma$       約95.5%      (ほぼすべての範囲)
- 平均値  $\pm 3\sigma$       約99.7%      (限りなくすべてに近い)

# 標準偏差

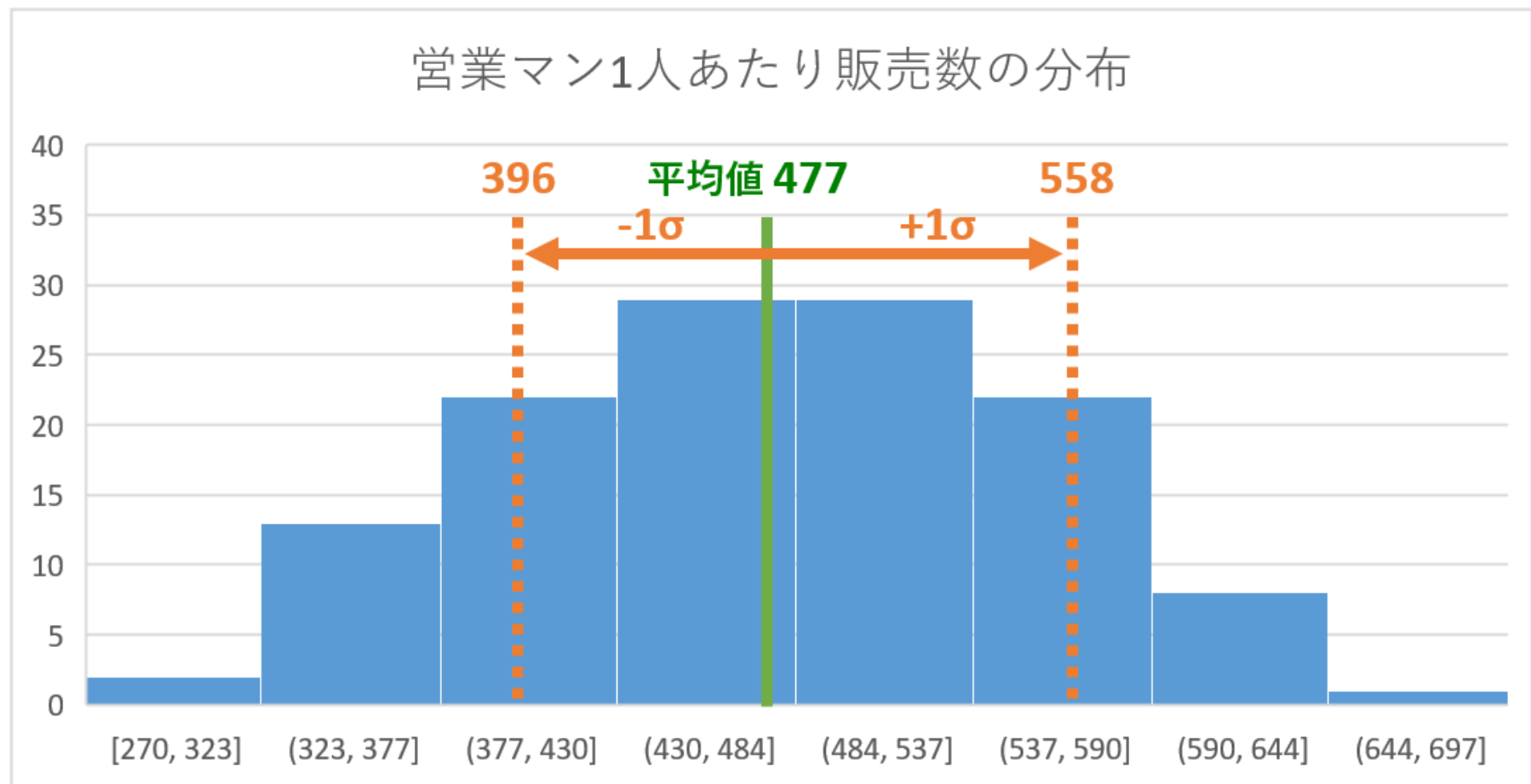
---

## 1. 覚えておくと便利な、範囲イメージ

- 平均値  $\pm 1\sigma$       約68.2%      (だいたいこの範囲)
- 平均値  $\pm 2\sigma$       約95.5%      (ほぼすべての範囲)
- 平均値  $\pm 3\sigma$       約99.7%      (限りなくすべてに近い)

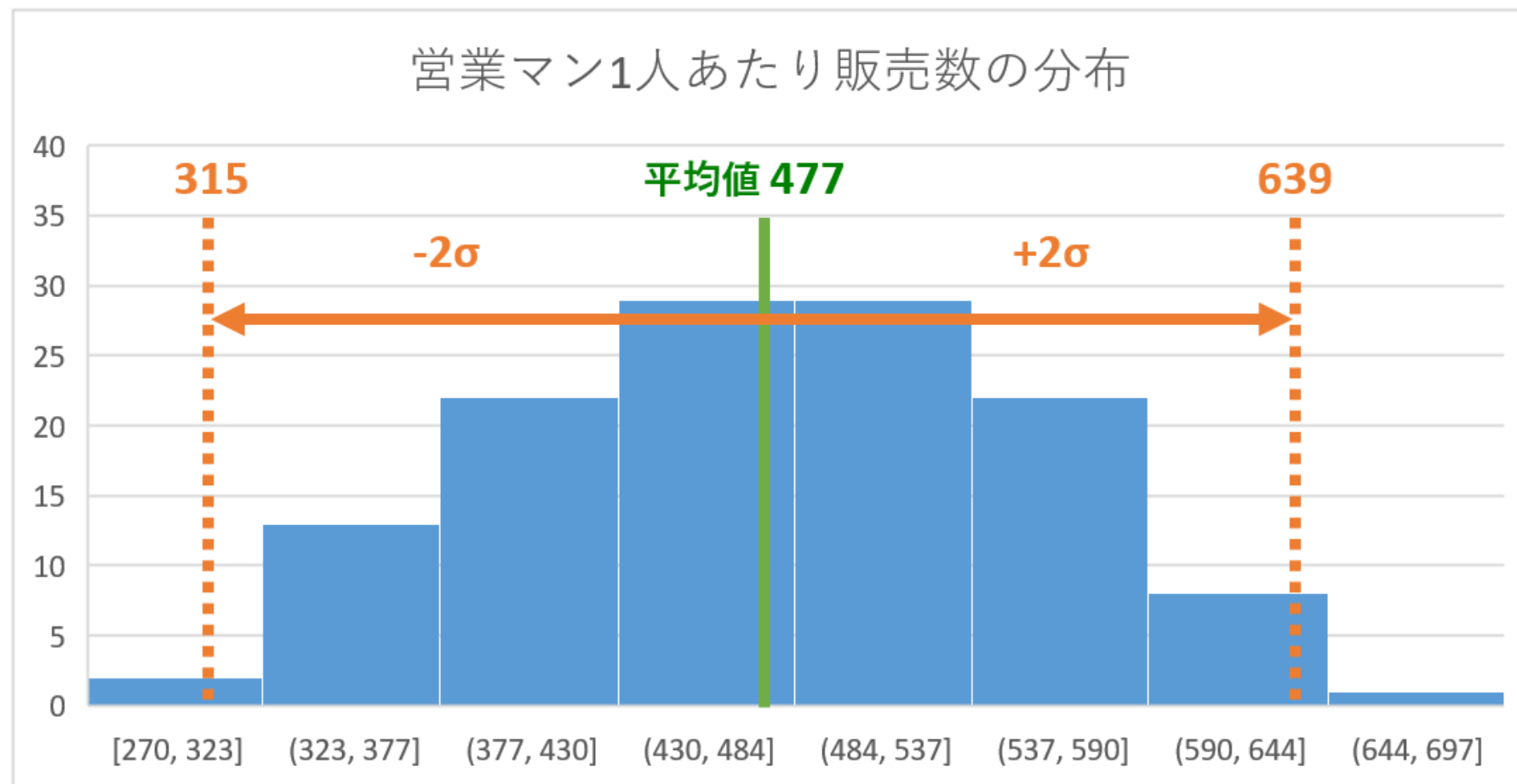
# 標準偏差

1 $\sigma$  : 標準偏差 (=81) をプラスマイナス



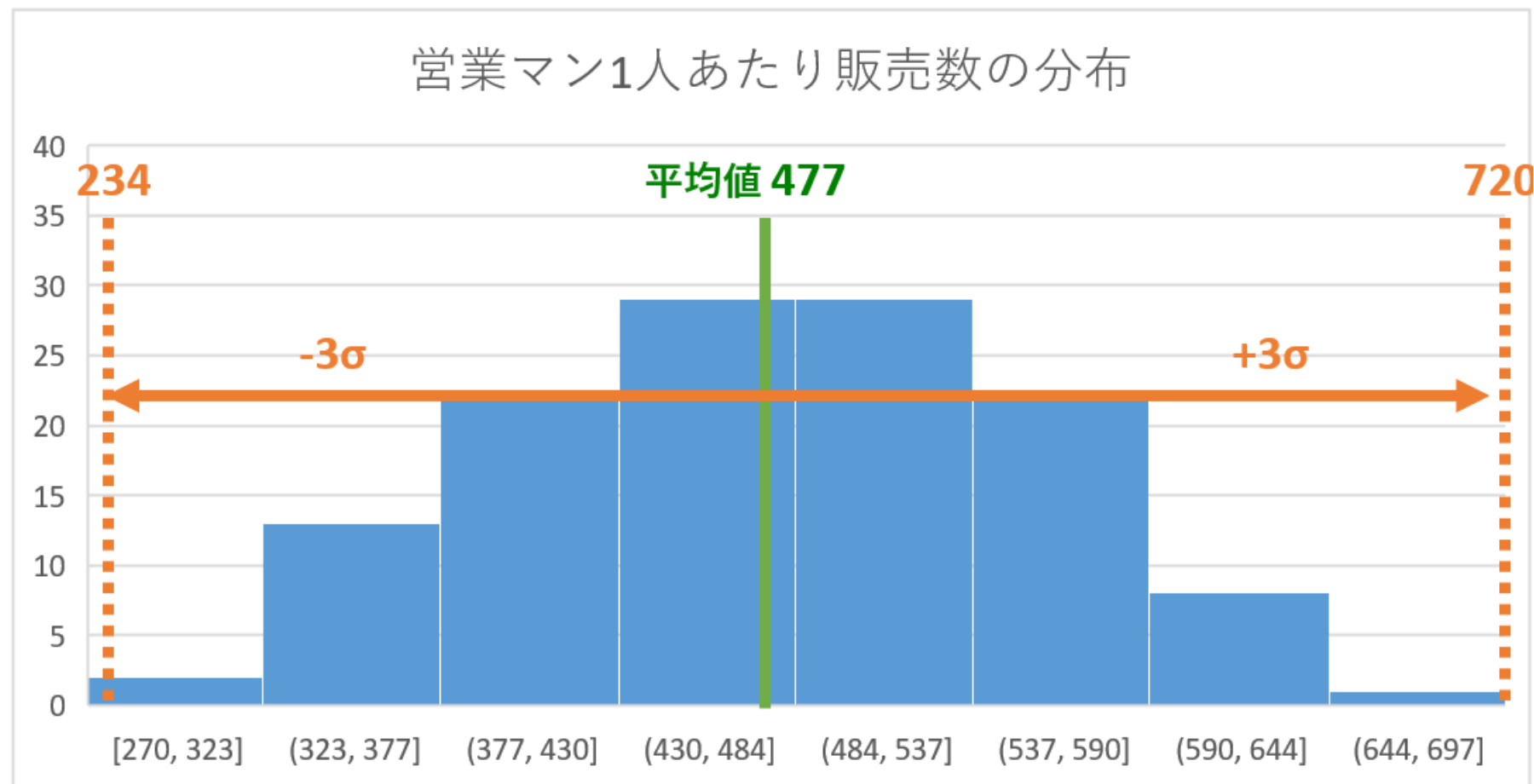
# 標準偏差

$2\sigma$  : 標準偏差 ( $=81$ )  $\times 2 = 162$ をプラスマイナス



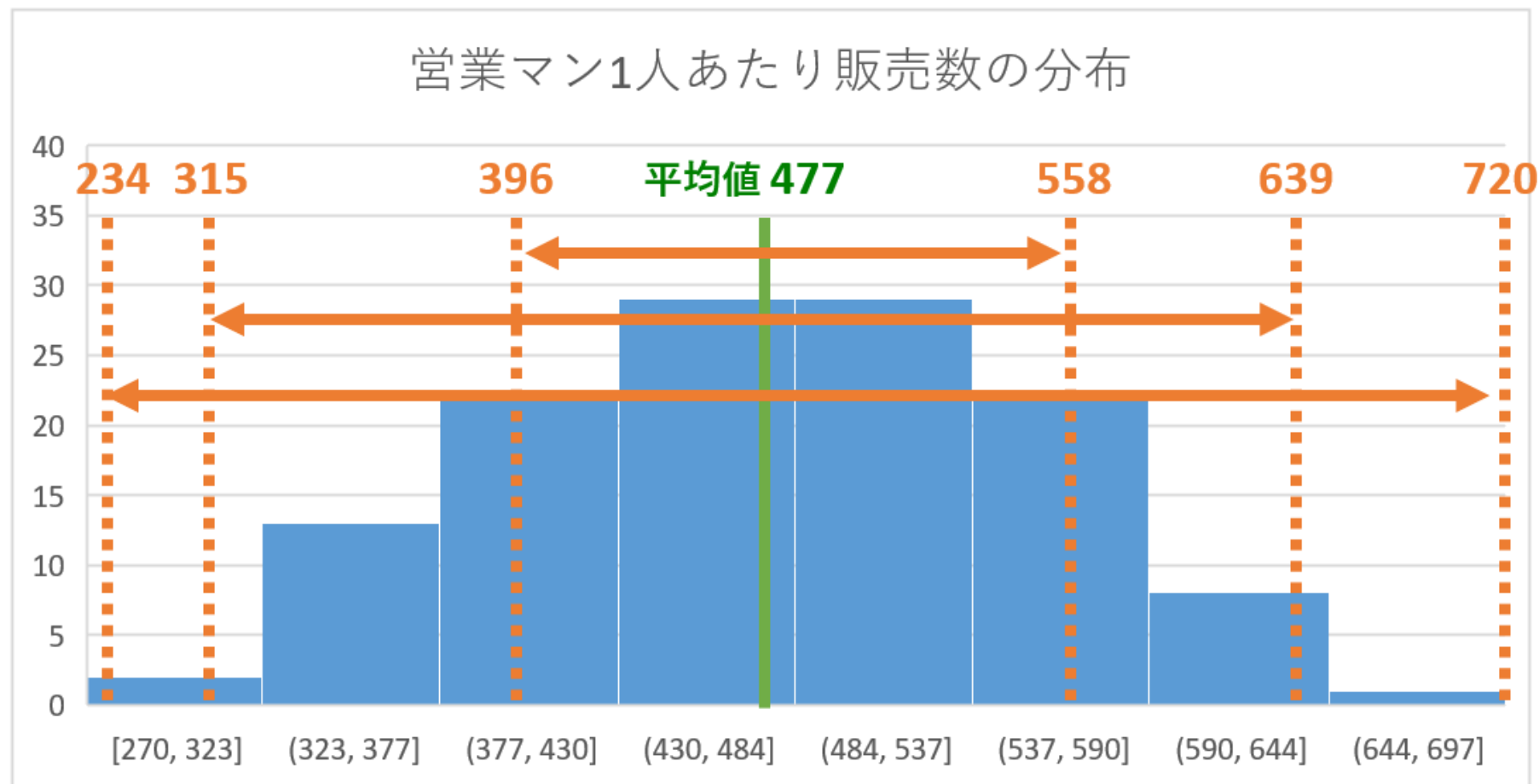
# 標準偏差

**3σ : 標準偏差 (=81) × 3 = 243をプラスマイナス**



# 標準偏差

$1\sigma \sim 3\sigma$  : かなり幅にちがいがあることがわかる



# 標準偏差

---

## 1. 覚えておくと便利な、範囲イメージ

- 平均値  $\pm 1\sigma$       約68.2%      (だいたいこの範囲)
- 平均値  $\pm 2\sigma$       約95.5%      (ほぼすべての範囲)
- 平均値  $\pm 3\sigma$       約99.7%      (限りなくすべてに近い)

## 2. 注意点

- 上記の数値%は、きれいな分布（正規分布）の場合
- 現実には、範囲に収まらない場合もあるので、あくまで目安

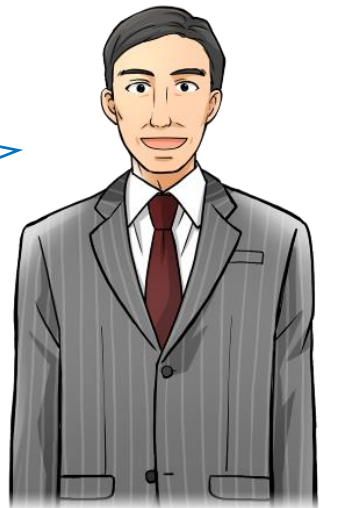
# 先ほどの例



約70%の営業マンが、  
約400～550個を販売しています！

なるほど！

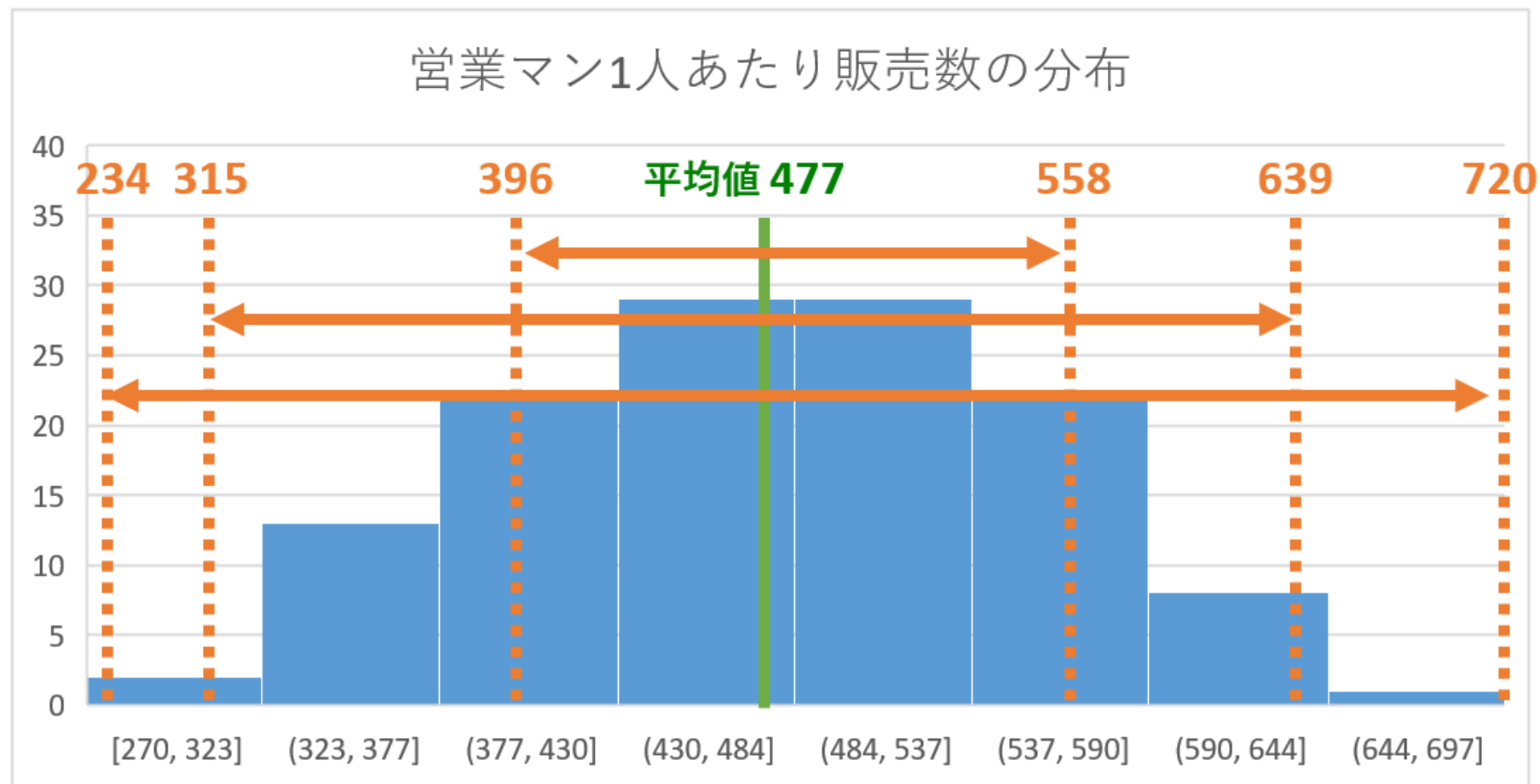
営業マンの販売イメージがつかめたぞ！





# 標準偏差

1 $\sigma$ が約70%、3 $\sigma$ だと約99.5%

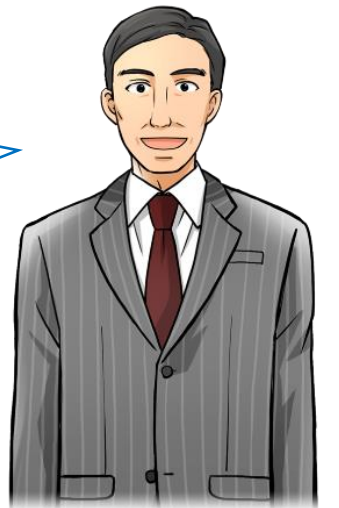


# このように言うこともできる



約99.5%の営業マンが、  
約234～720個を販売してます！

ちょっと幅が広すぎるかな・・・？  
もう少し絞ってもらったほうが分かりやすい



# 標準偏差

---

## 1. ポイント

- 範囲を広げすぎると、逆に意味が分かりにくくなる

## 2. 例

- 「日本人の99.5%の身長は、1 ～ 2メートルの間」  
→ 幅が広すぎてイメージできない
- 「日本人の70%の身長は、1.65 ～ 1.75メートルの間」  
→ 標準の数値をイメージしやすい

# 今回のポイント

---

分布

# 代表値と分布

1. データを1つの数値で代表する（代表値）
2. 代表値では分からない「ばらつき」を知る（分布）

## 代表値

---

- 平均値
- 中央値
- 最頻値



## 分布

---

- ヒストグラム
- 標準偏差
- 混合分布
- パレート図

# 分布

---

## 1. 意味

- いわゆる「数字のばらつき」

## 2. 例えば

- 平均値が50点のテスト結果といっても、

- 全員が50点

- 半分が0点、残りの半分が100点

平均は同じでも

意味は異なる

→ 分布（ばらつき）を見る必要がある

# 分布

---

## 1. 意味

- いわゆる「数字のばらつき」

## 2. 例えば

- 平均値が50点のテスト結果といっても、

- 全員が50点

- 半分が0点、残りの半分が100点

平均は同じでも

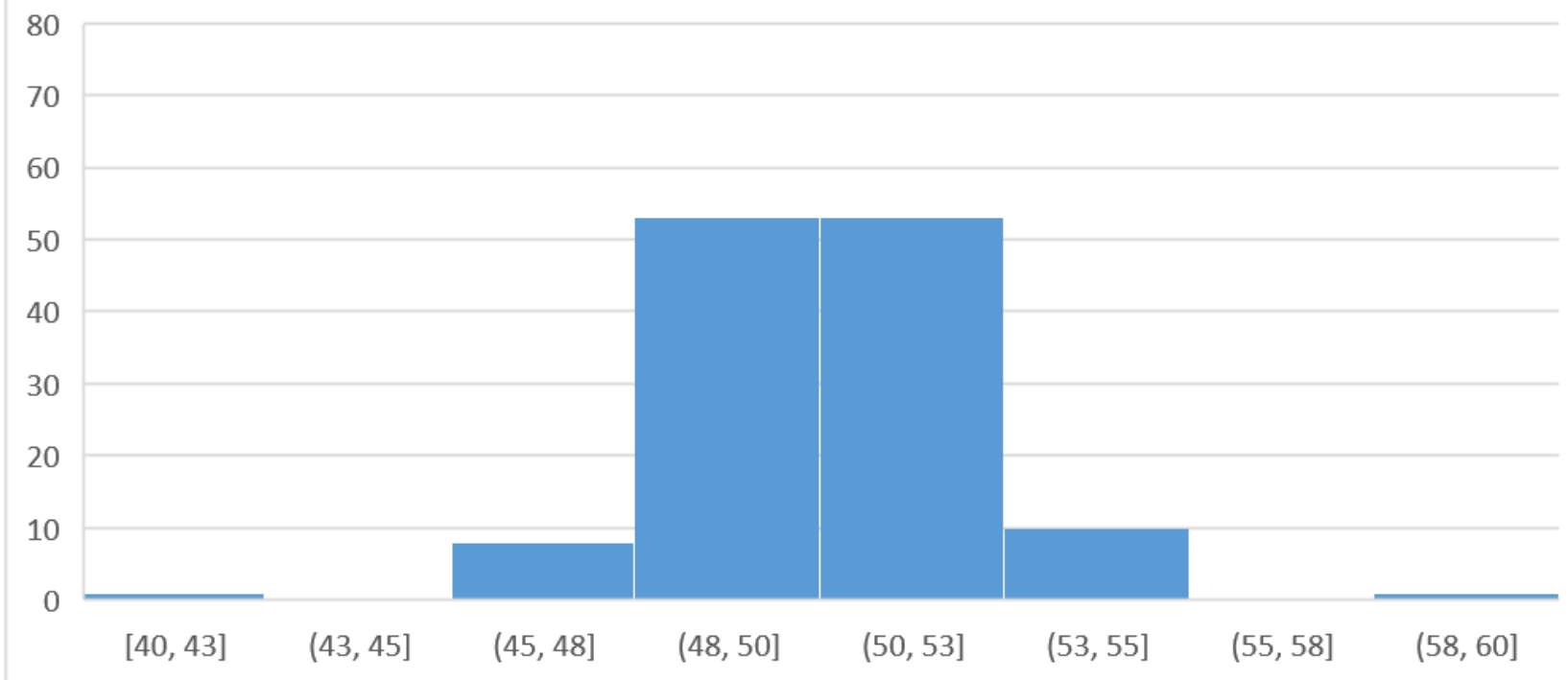
意味は異なる

→ 分布（ばらつき）を見る必要がある

→ 標準偏差を使って分布を計算する

# 標準偏差

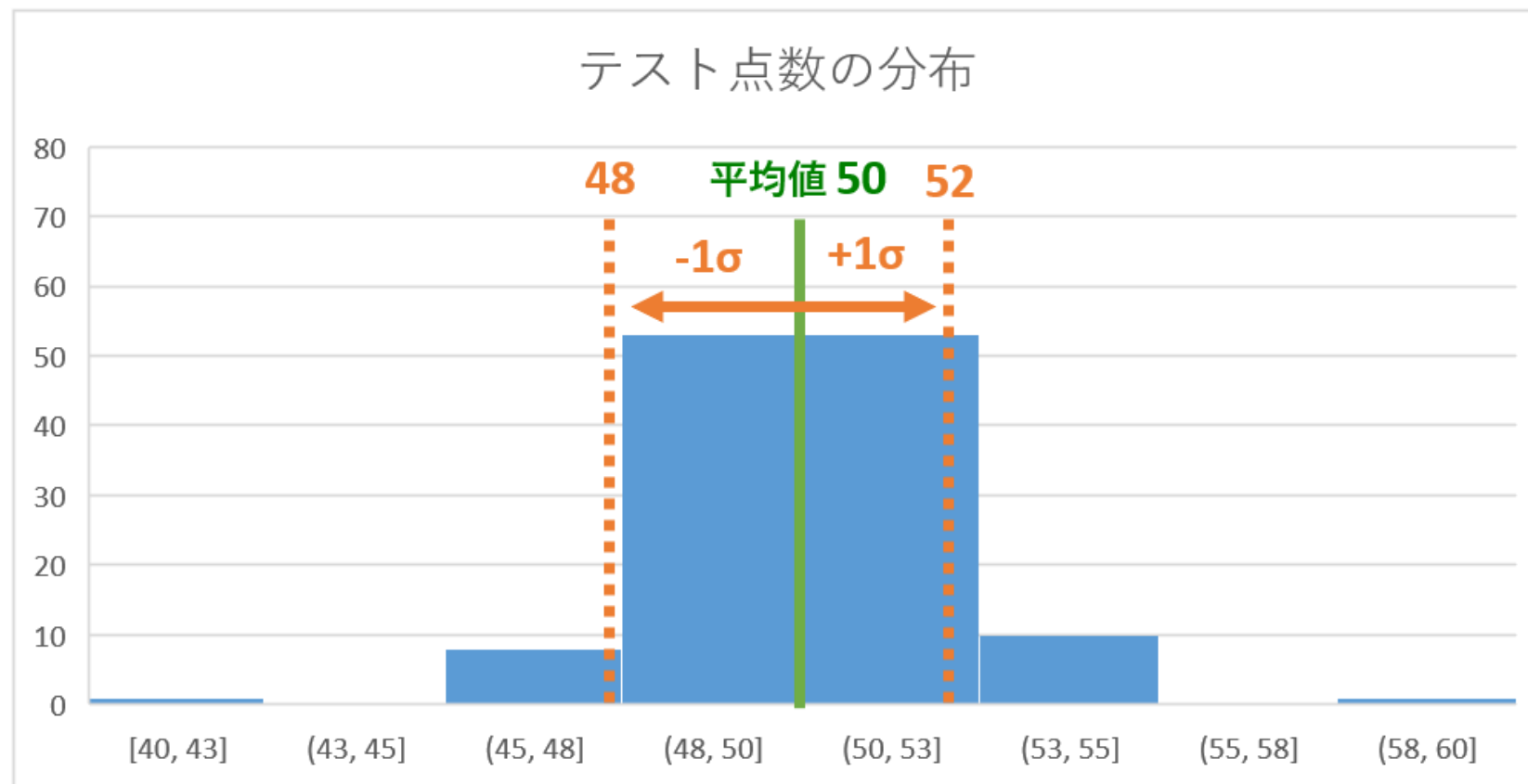
テスト点数の分布





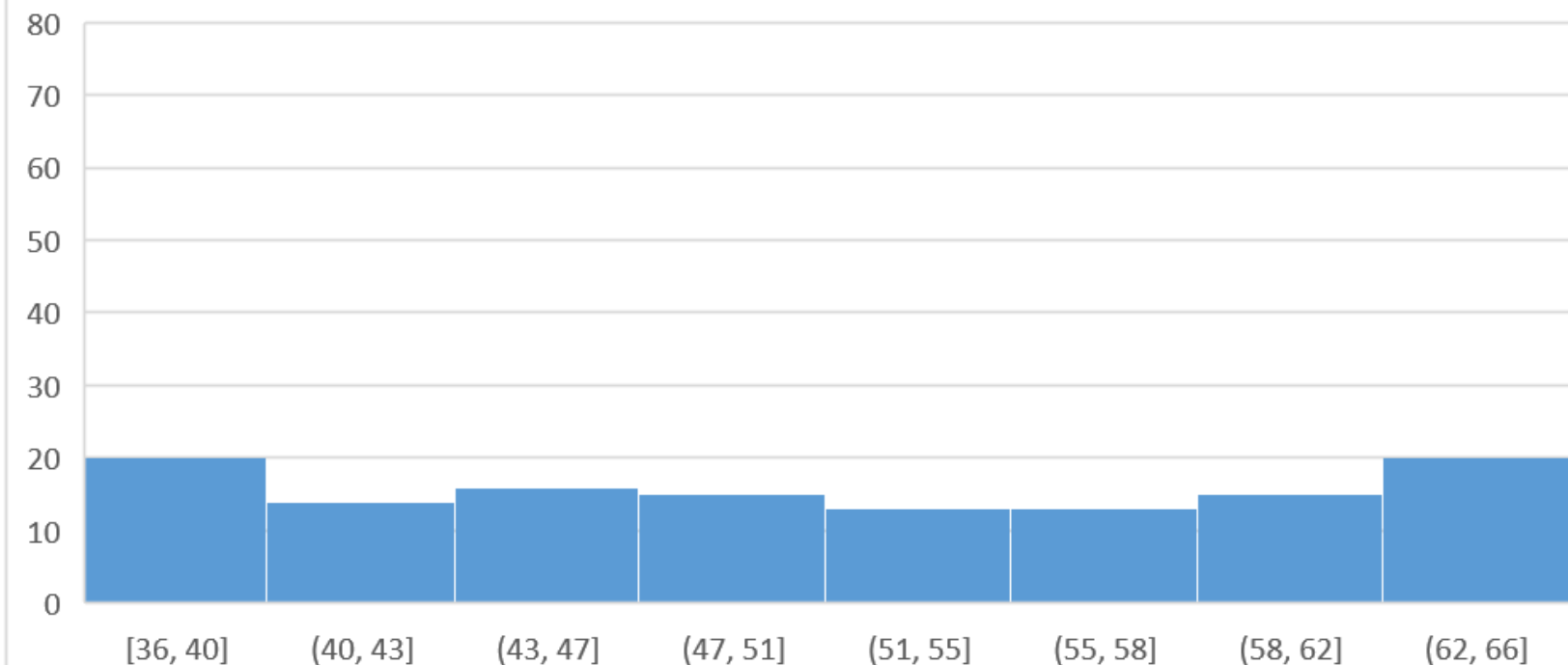
# 標準偏差

1 $\sigma$ で48点～52点に収まっている（＝密集している）



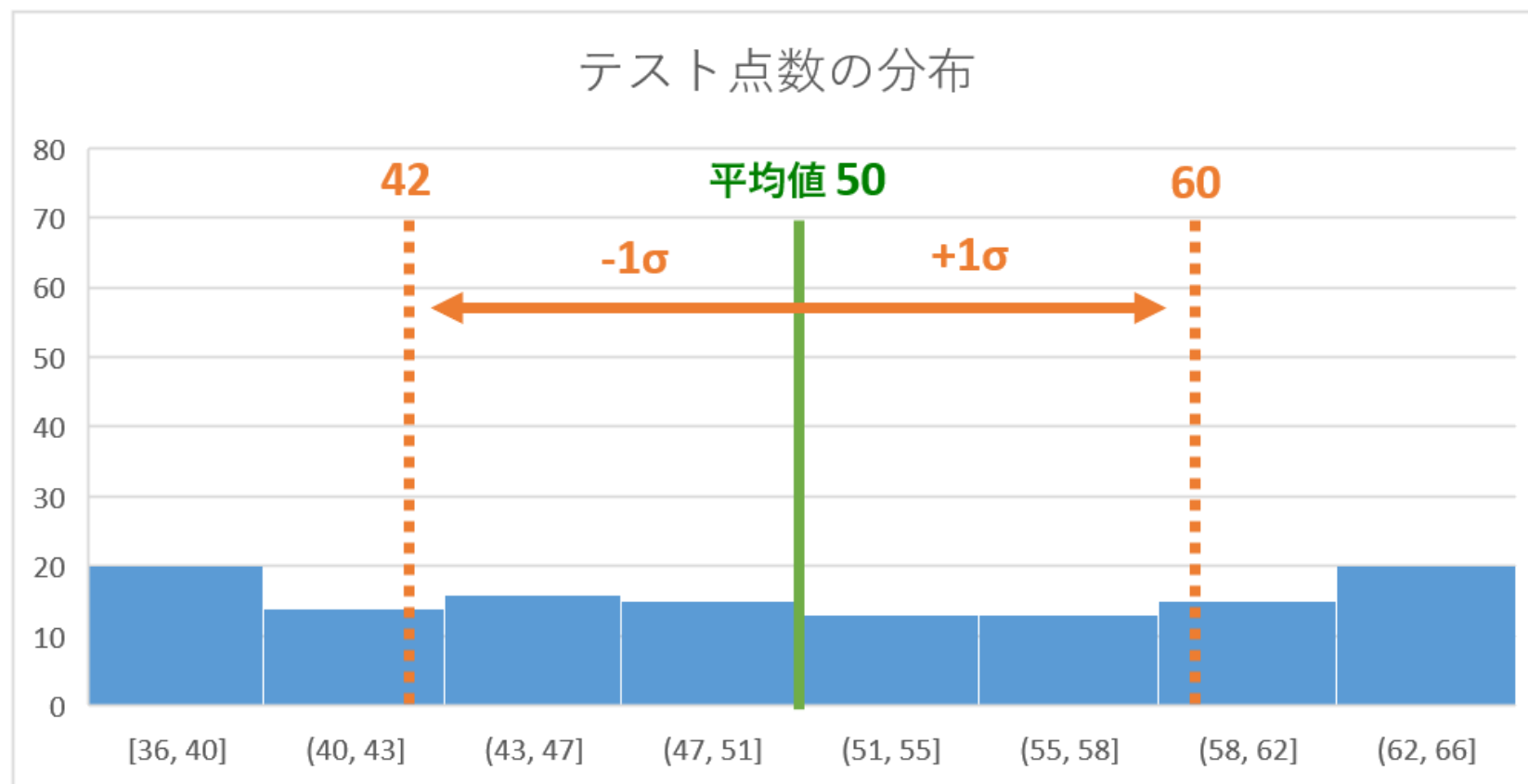
# 標準偏差

テスト点数の分布



# 標準偏差

1 $\sigma$ で42点～60点に収まっている（＝幅広く分散している）



# 分布

## 1. 意味

- いわゆる「数字のばらつき」

## 2. 例えば

- 平均値が50点のテスト結果といっても、

- 全員が50点

- 半分が0点、残りの半分が100点

平均は同じでも

意味は異なる

→ 分布（ばらつき）を見る必要がある

→ 標準偏差を使って分布を計算する

# 今回のポイント

---

標準偏差

# 今回のポイント

---

「大体どの範囲に収まっている？」

# 大量のデータを読み解く

1. 代表値(1) 平均値
2. 代表値(2) 中央値
3. 代表値(3) 最頻値
4. 分布(1) ヒストグラム
5. 分布(2) 標準偏差
6. 分布(3) 混合分布
7. 分布(4) パレート図
8. まとめ(1) 代表値と分布
9. 傾向(1) 推移
10. 傾向(2) ヒートマップ
11. 関係性(1) 相関分析
12. 関係性(2) 因果関係
13. 関係性(3) 第三因子
14. 関係性(4) 混合グループ
15. 関係性(5) 外れ値
16. まとめ(2) 傾向、関係性

# 今回のポイント

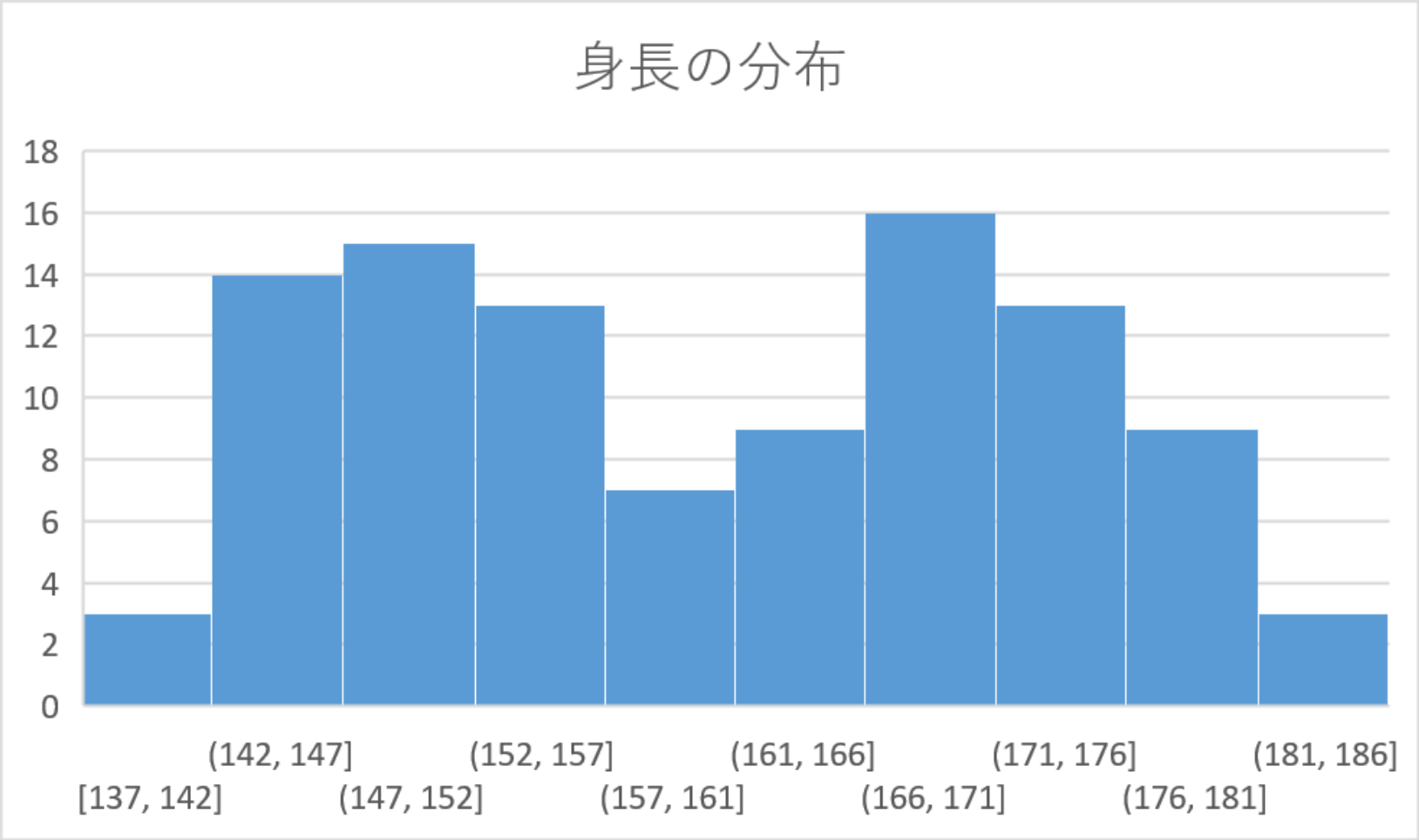
---

混合分布



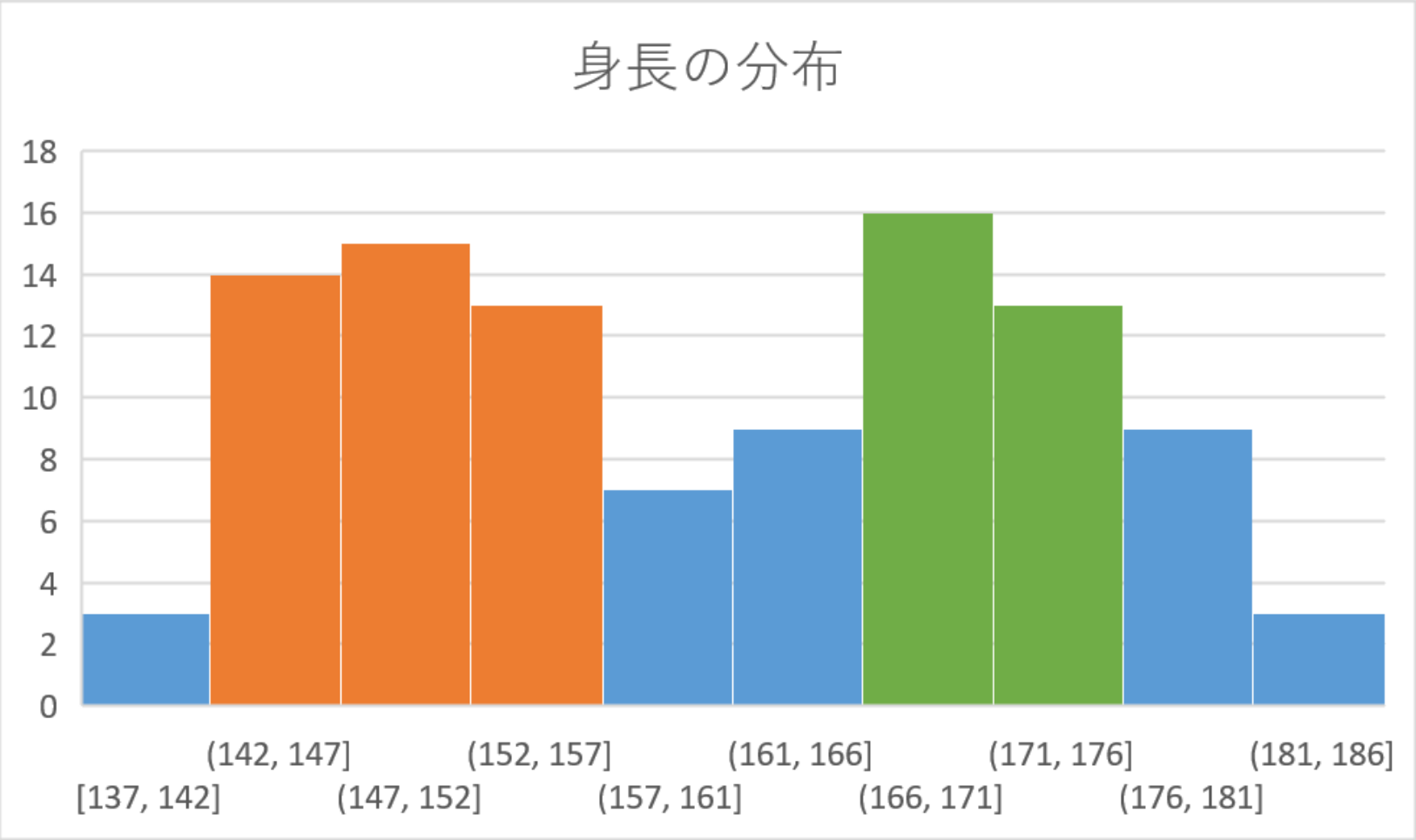
# 混合分布

身長の分布（137cm ～ 186cm）きれいな山になっていない

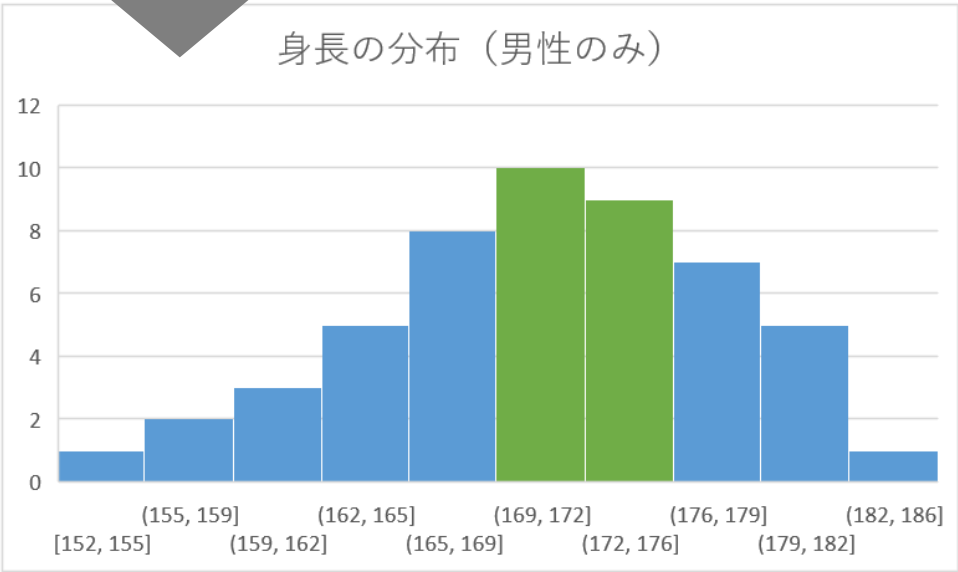
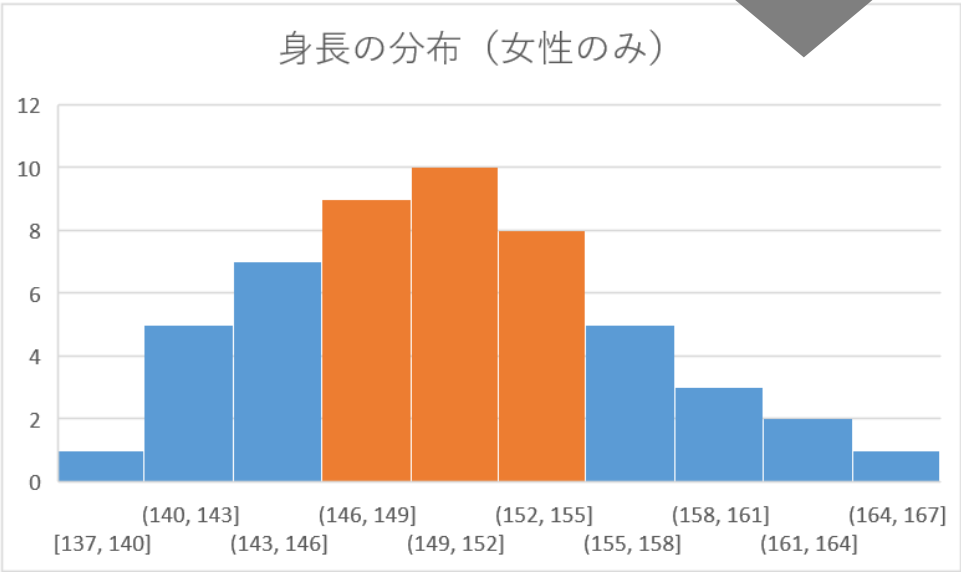
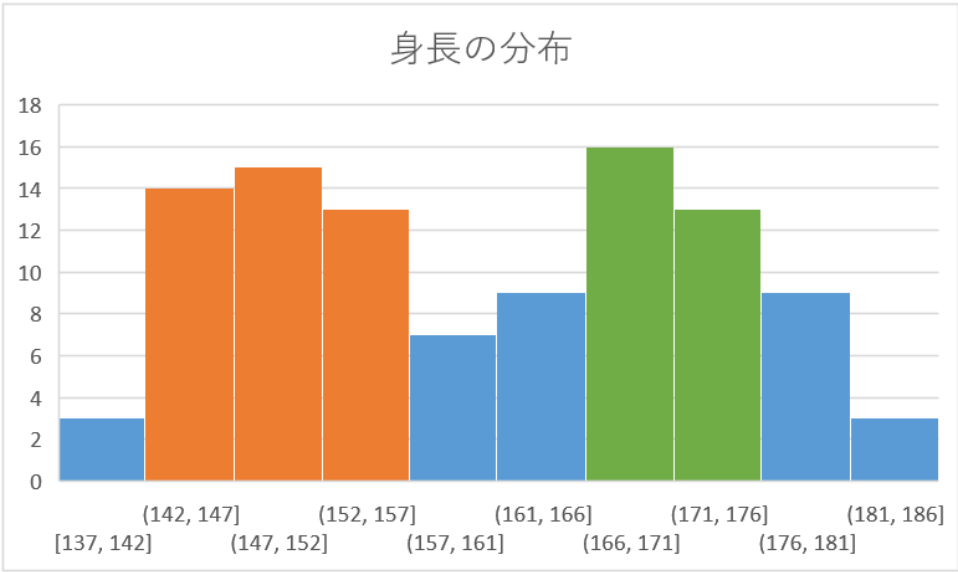


# 混合分布

よく見ると、山が2つある（なぜか？）



# 女性、男性の分布が混在していた



# 混合分布

---

## 1. 意味

- 異なるデータが混ざっていると、  
ヒストグラムが見にくい、標準偏差がゆがむ

## 2. ポイント

- できるだけ近しいデータを集める
- タイプのちがうデータは、分けて計算する  
(女性データと男性データ)

# 大量のデータを読み解く

1. 代表値(1) 平均値
2. 代表値(2) 中央値
3. 代表値(3) 最頻値
4. 分布(1) ヒストグラム
5. 分布(2) 標準偏差
6. 分布(3) 混合分布
7. 分布(4) パレート図
8. まとめ(1) 代表値と分布
9. 傾向(1) 推移
10. 傾向(2) ヒートマップ
11. 関係性(1) 相関分析
12. 関係性(2) 因果関係
13. 関係性(3) 第三因子
14. 関係性(4) 混合グループ
15. 関係性(5) 外れ値
16. まとめ(2) 傾向、関係性

# 今回のポイント

---

パレート図

# 今回のポイント

---

「どの顧客までをターゲットにすれば十分か」

# パレート図

今度の新しいテレビCMだが、  
どの世代までターゲットにすれば十分か、  
調べてくれないか？



調べます！



# パレート図

年齢が高いほうが販売数が多い傾向がある

世代別販売数		
	販売数	シェア
70代以上	9,000	38%
60代	6,000	25%
50代	4,000	17%
40代	2,500	11%
30代	1,000	4%
20代	700	3%
10代以下	500	2%
合計	23,700	100%

# プレート図



そうですね・・・

年齢が高いほうが販売数が多いので、  
ターゲットは60代以上くらいかな・・・

うーん、もう少しデータで提案してもらえると  
うれしいのだが・・・



# 今回のポイント

---

「どの顧客までをターゲットにすれば十分か」

→パレート図を使う

# パレート図

年齢が高いほうが販売数が多い傾向がある

世代別販売数		
	販売数	シェア
70代以上	9,000	38%
60代	6,000	25%
50代	4,000	17%
40代	2,500	11%
30代	1,000	4%
20代	700	3%
10代以下	500	2%
合計	23,700	100%

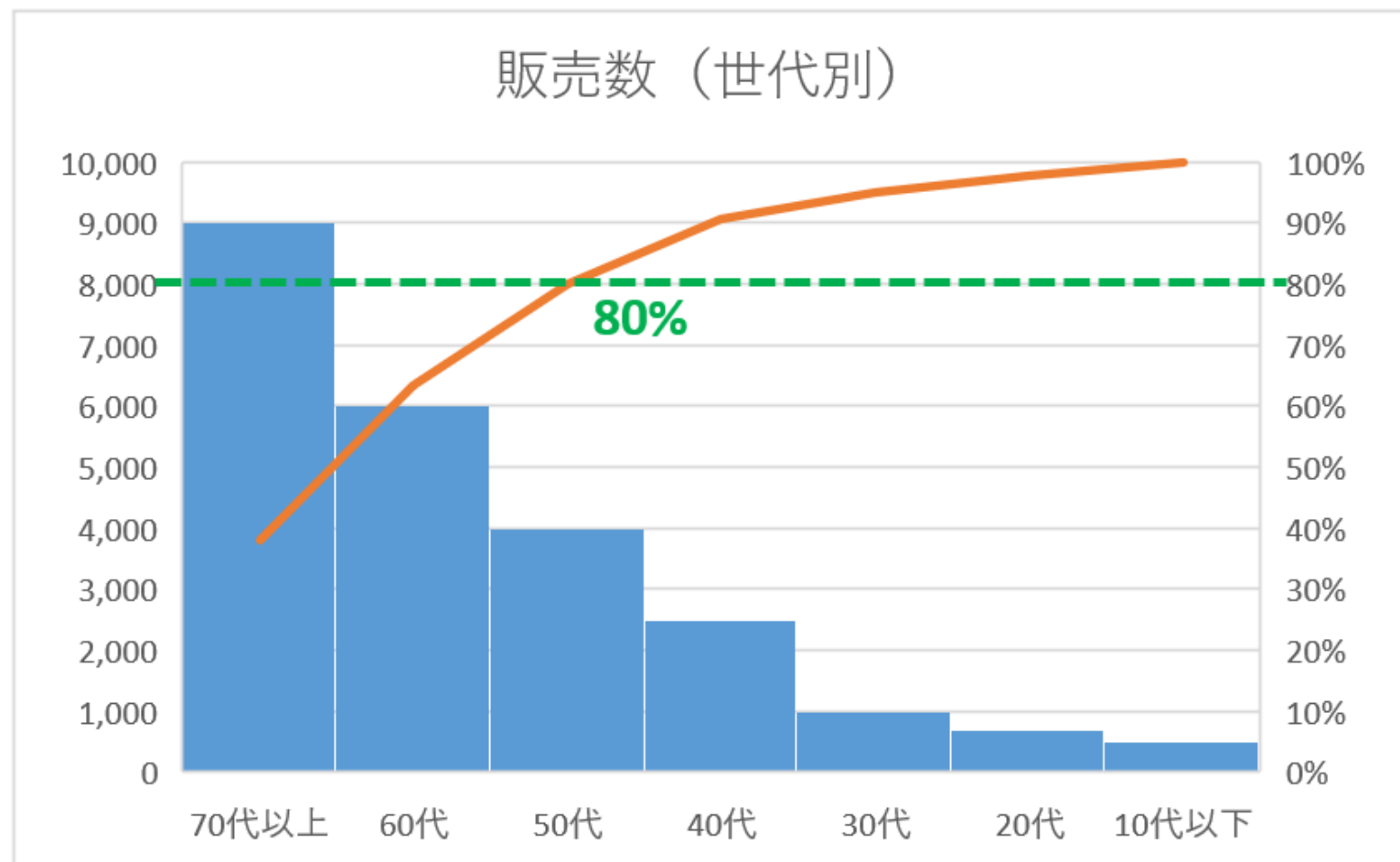
# パレート図

年齢の上から順番にシェアを足していくと、50代までで80%を占める

世代別販売数			
	販売数	シェア	累計シェア
70代以上	9,000	38%	38%
60代	6,000	25%	63%
50代	4,000	17%	80%
40代	2,500	11%	91%
30代	1,000	4%	95%
20代	700	3%	98%
10代以下	500	2%	100%
合計	23,700	100%	

# パレート図

グラフ化したものをパレート図と呼ぶ



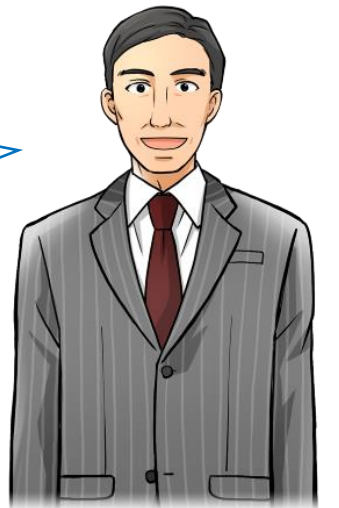
# パレート図



50代～70代以上をターゲットにすれば  
顧客の80%をカバーできます！

なるほど！

80%なら十分カバーしているな！



# パレート図

---

## 1. パレートの法則

- 全体の数値の大部分は、一部の要素が生み出している
- 売上の80%は、20%の優良顧客から生まれている
  - 80：20の法則
- マーケティングの優先順位を考えるときに使われる

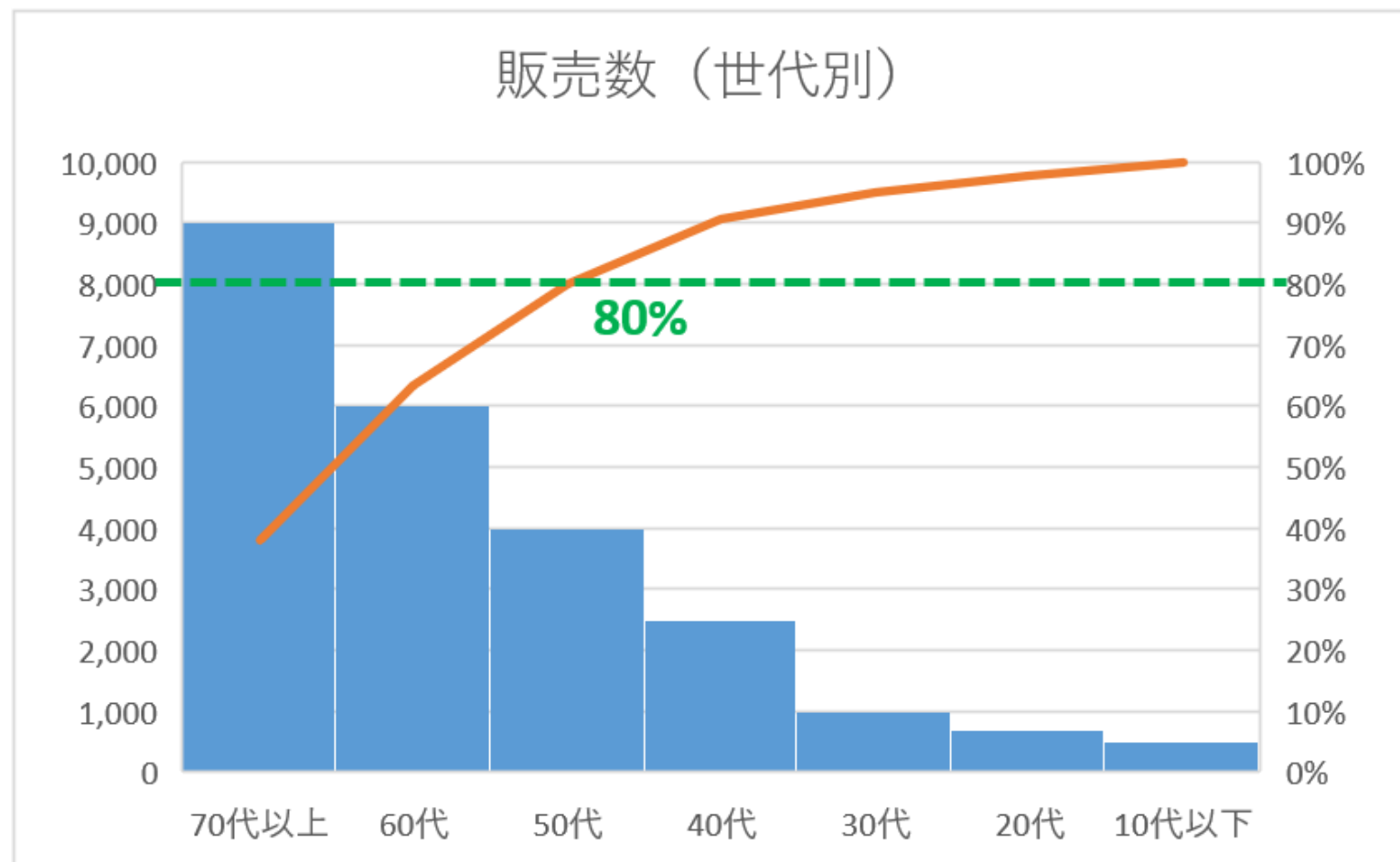
## 2. パレート図

- 「ここまでカバーすれば全体のXX%を占める」  
ことが分かりやすい



# パレート図

グラフ化したものをパレート図と呼ぶ



# 今回のポイント

---

「どの顧客までをターゲットにすれば十分か」

→パレート図を使う

# 大量のデータを読み解く

1. 代表値(1) 平均値
2. 代表値(2) 中央値
3. 代表値(3) 最頻値
4. 分布(1) ヒストグラム
5. 分布(2) 標準偏差
6. 分布(3) 混合分布
7. 分布(4) パレート図
8. まとめ(1) 代表値と分布

9. 傾向(1) 推移
10. 傾向(2) ヒートマップ
11. 関係性(1) 相関分析
12. 関係性(2) 因果関係
13. 関係性(3) 第三因子
14. 関係性(4) 混合グループ
15. 関係性(5) 外れ値
16. まとめ(2) 傾向、関係性

# 今回のポイント

---

傾向を理解する

# 傾向を理解する

---

## 1. 意味

- あるデータを見て「販売数が増え続けている」
  - 時間が経過するごとの数値の変化（＝推移、時系列）
- 「東京都の販売数は増えているが、それ以外はあまり増えていないな」
  - 地域ごとに分解した、推移データ

## 2. ポイント

- 大量データを見ても傾向は理解しにくい  
→ グラフにしたり、色をつけることで理解しやすくする

# 今回のポイント

---

推移

# 推移

1月から10月までの、商品ごとの  
販売数の傾向を教えてくださいか？



調べます！

# 推移

## 1. 商品ごとの1月～10月の販売数の推移

- なにが分かるか？

販売数										
	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月
商品A	1,200	1,315	1,549	1,684	1,814	2,012	2,260	2,369	2,536	2,784
商品B	1,200	960	768	614	492	737	1,106	1,659	2,488	3,732
商品C	1,200	1,320	1,452	1,597	1,757	1,933	1,025	2,338	2,572	2,830



# 推移

## 1. 商品ごとの1月～10月の販売数の推移

- なにが分かるか？
- **たとえば1月と10月を比較すると、増加している**

販売数	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月
商品A	1,200	1,315	1,549	1,684	1,814	2,012	2,260	2,369	2,536	2,784
商品B	1,200	960	768	614	492	737	1,106	1,659	2,488	3,732
商品C	1,200	1,320	1,452	1,597	1,757	1,933	1,025	2,338	2,572	2,830

# 推移



1月と10月を比較すると、  
どの商品の販売数も増えています！

うーん、それは傾向とはいえないな・・・  
どう増えているのか教えてほしい



# 推移

---

## 1. 意味

- あるデータを見て「販売数が増え続けている」
  - 時間が経過するごとの数値の変化（＝推移、時系列）

## 2. ポイント

- 推移、時系列のデータを見るときは、  
表データだけを見てもイメージしにくい  
→ グラフにすると分かりやすい

# 今回のポイント

---

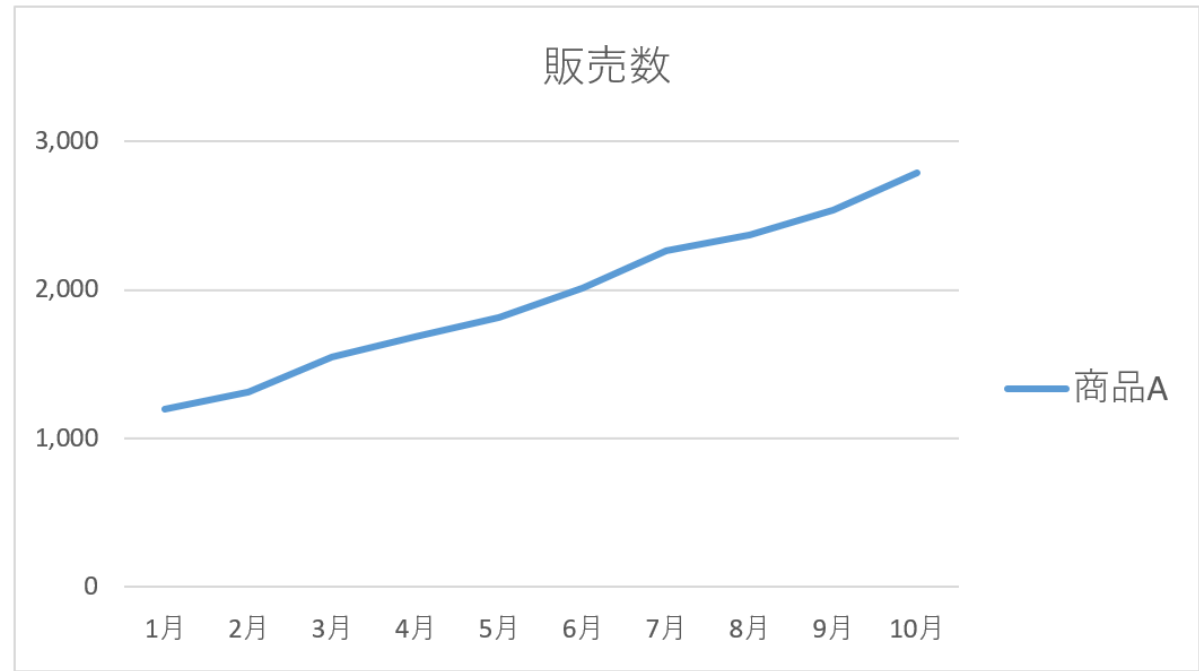
推移の傾向を見るときは、  
グラフにする

# 推移

販売数										
	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月
商品A	1,200	1,315	1,549	1,684	1,814	2,012	2,260	2,369	2,536	2,784
商品B	1,200	960	768	614	492	737	1,106	1,659	2,488	3,732
商品C	1,200	1,320	1,452	1,597	1,757	1,933	1,025	2,338	2,572	2,830

# 推移

販売数										
	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月
商品A	1,200	1,315	1,549	1,684	1,814	2,012	2,260	2,369	2,536	2,784
商品B	1,200	960	768	614	492	737	1,106	1,659	2,488	3,732
商品C	1,200	1,320	1,452	1,597	1,757	1,933	1,025	2,338	2,572	2,830



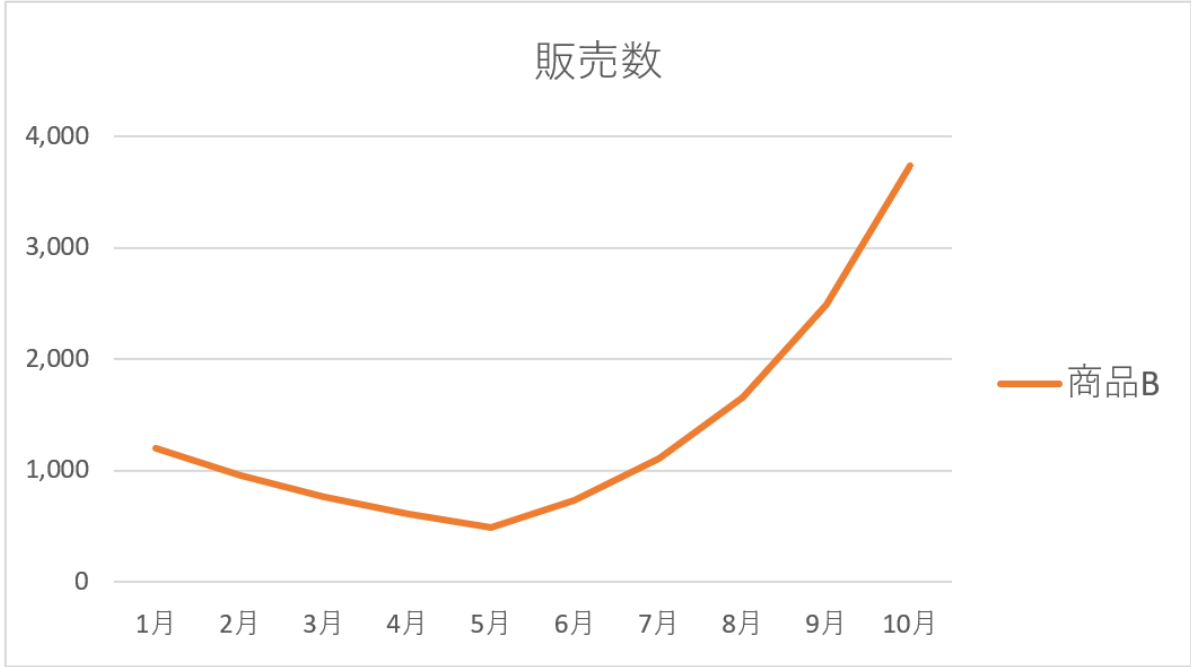
順調に増加している  
ことがわかる

# 推移

販売数										
	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月
商品A	1,200	1,315	1,549	1,684	1,814	2,012	2,260	2,369	2,536	2,784
商品B	1,200	960	768	614	492	737	1,106	1,659	2,488	3,732
商品C	1,200	1,320	1,452	1,597	1,757	1,933	1,025	2,338	2,572	2,830

# 推移

販売数										
	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月
商品A	1,200	1,315	1,549	1,684	1,814	2,012	2,260	2,369	2,536	2,784
商品B	1,200	960	768	614	492	737	1,106	1,659	2,488	3,732
商品C	1,200	1,320	1,452	1,597	1,757	1,933	1,025	2,338	2,572	2,830



一度減少したが、  
そのあと大きく改善  
したことがわかる

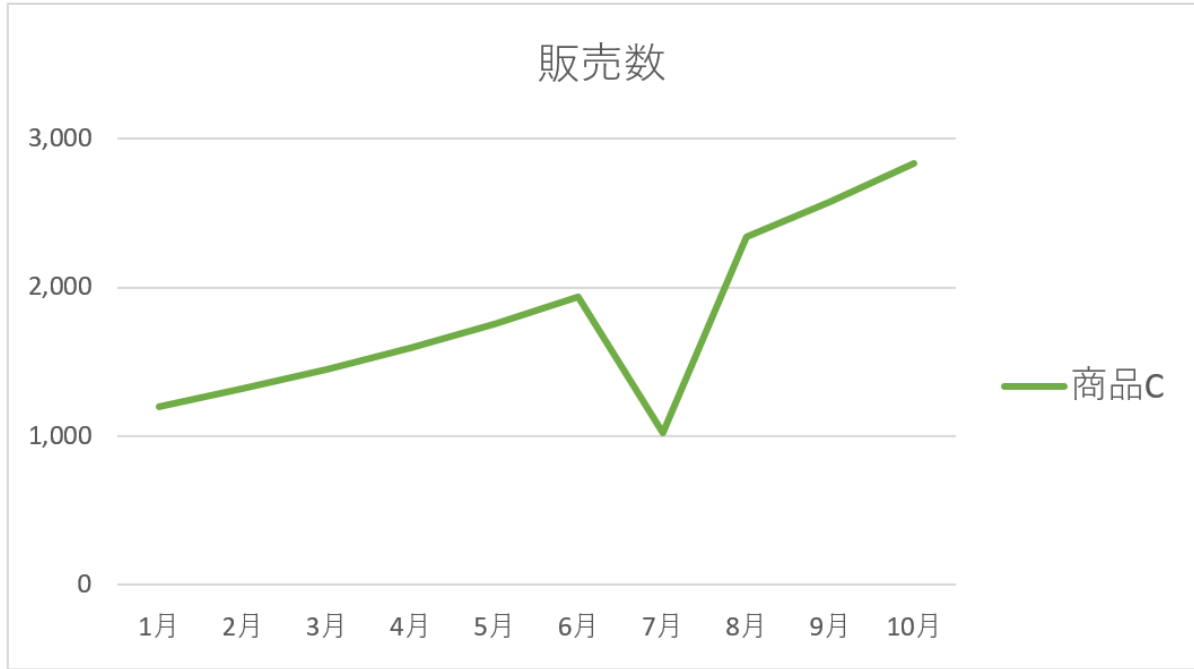


# 推移

販売数										
	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月
商品A	1,200	1,315	1,549	1,684	1,814	2,012	2,260	2,369	2,536	2,784
商品B	1,200	960	768	614	492	737	1,106	1,659	2,488	3,732
商品C	1,200	1,320	1,452	1,597	1,757	1,933	1,025	2,338	2,572	2,830

# 推移

販売数										
	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月
商品A	1,200	1,315	1,549	1,684	1,814	2,012	2,260	2,369	2,536	2,784
商品B	1,200	960	768	614	492	737	1,106	1,659	2,488	3,732
商品C	1,200	1,320	1,452	1,597	1,757	1,933	1,025	2,338	2,572	2,830



7月だけ大きく減少。  
なにか特別なことが起きた？  
(確認が必要)

# 推移

---

## 1. 意味

- あるデータを見て「販売数が増え続けている」
  - 時間が経過するごとの数値の変化（＝推移、時系列）

## 2. ポイント

- 推移、時系列のデータを見る
  - グラフにすると分かりやすい

# 大量のデータを読み解く

1. 代表値(1) 平均値
2. 代表値(2) 中央値
3. 代表値(3) 最頻値
4. 分布(1) ヒストグラム
5. 分布(2) 標準偏差
6. 分布(3) 混合分布
7. 分布(4) パレート図
8. まとめ(1) 代表値と分布
9. 傾向(1) 推移
10. 傾向(2) ヒートマップ
11. 関係性(1) 相関分析
12. 関係性(2) 因果関係
13. 関係性(3) 第三因子
14. 関係性(4) 混合グループ
15. 関係性(5) 外れ値
16. まとめ(2) 傾向、関係性

# 今回のポイント

---

傾向を理解する

# 傾向を理解する

---

## 1. 意味

- あるデータを見て「販売数が増え続けている」
  - 時間が経過するごとの数値の変化（＝推移、時系列）
- 「東京都の販売数は増えているが、それ以外はあまり増えていないな」
  - 地域ごとに分解した、推移データ

## 2. ポイント

- 大量データを見ても傾向は理解しにくい  
→ グラフにしたり、色をつけることで理解しやすくする

# 傾向を理解する

---

## 1. 意味

- あるデータを見て「販売数が増え続けている」
  - 時間が経過するごとの数値の変化（＝推移、時系列）
- 「東京都の販売数は増えているが、それ以外はあまり増えていないな」
  - 地域ごとに分解した、推移データ

## 2. ポイント

- 大量データを見ても傾向は理解しにくい  
→ グラフにしたり、色をつけることで理解しやすくする

# 今回のポイント

---

数字に色を付けると、傾向を理解できる  
(ヒートマップ)



# ヒートマップ

---

## 1. 意味

- 数字に色を付けて傾向を読み取る
- 例
  - 販売数が250より大きいところは緑色
  - 販売数が120より小さいところは赤色

# ヒートマップ

地域ごとに分解した推移データを見ても、傾向が分かりにくい

販売数							
	1月	2月	3月	4月	5月	6月	7月
山形県	235	271	280	192	131	136	157
福島県	239	100	95	108	140	104	262
茨城県	129	194	296	107	261	150	117
栃木県	239	260	213	100	216	128	214
群馬県	177	206	175	111	204	197	223
埼玉県	249	256	365	251	264	380	332
千葉県	298	365	381	247	255	250	325
東京都	522	346	450	538	383	462	486
神奈川県	254	240	263	250	267	247	359

# ヒートマップ

250より大きい数字 (緑) → 関東は販売数が多い傾向

120より小さい数字 (赤) → 4月は販売数が落ち込んだ (原因は?)

販売数

	1月	2月	3月	4月	5月	6月	7月
山形県	235	271	280	192	131	136	157
福島県	239	100	95	108	140	104	262
茨城県	129	194	296	107	261	150	117
栃木県	239	260	213	100	216	128	214
群馬県	177	206	175	111	204	197	223
埼玉県	249	256	365	251	264	380	332
千葉県	298	365	381	247	255	250	325
東京都	522	346	450	538	383	462	486
神奈川県	254	240	263	250	267	247	359

# ヒートマップ

---

## 1. ポイント

- データを分析するときに、代表値を計算、分布を計算、グラフにしてみる・・・などいろいろな工夫をしてきました
  - データを表で見ると理解しにくいから
- 実はそこまでしなくても、表に色をつけるだけで理解できる場合もある

## 2. Excelテクニック

- 「条件付き書式」で簡単にセルの色を変えることができます

# 大量のデータを読み解く

1. 代表値(1) 平均値
2. 代表値(2) 中央値
3. 代表値(3) 最頻値
4. 分布(1) ヒストグラム
5. 分布(2) 標準偏差
6. 分布(3) 混合分布
7. 分布(4) パレート図
8. まとめ(1) 代表値と分布
9. 傾向(1) 推移
10. 傾向(2) ヒートマップ
11. 関係性(1) 相関分析
12. 関係性(2) 因果関係
13. 関係性(3) 第三因子
14. 関係性(4) 混合グループ
15. 関係性(5) 外れ値
16. まとめ(2) 傾向、関係性

# 今回のポイント

---

関係性

# 関係性

今年から始めた広告宣伝（テレビCM）は  
売上につながっているのだろうか？



調べます！

# 関係性

## 1. データ

- 広告宣伝費と売上
- 「広告宣伝費と売上に関係はありそうか？」

広告宣伝費（横軸） と売上（縦軸）

	4月	5月	6月	7月	8月
広告宣伝費	100	200	300	400	500
売上	400	300	550	800	850



# 関係性

## 1. データ

- 広告宣伝費と売上
- 「広告宣伝費と売上に関係はありそうか？」

広告宣伝費（横軸） と 売上（縦軸）

	4月	5月	6月	7月	8月
広告宣伝費	100	200	300	400	500
売上	400	300	550	800	850

# 関係性



4月から8月にかけて、  
広告宣伝費を増やすと売上も増えたので、  
関係あると思います！

うーん、どれくらい関係が強いのか、  
数字で教えてもらえると分かりやすいのだが



# 今回のポイント

---

相関分析

# 相関分析

---

## 1. 目的

- データ分析では、数字の関係性（相関関係）を知ることは重要
- せっかく広告宣伝に投資をしても、  
売上につながらなければ意味がない

## 2. 相関分析

- 2つの数字の推移を見て、その2つの数字に関係性がありそうかを分析する

# 相関分析

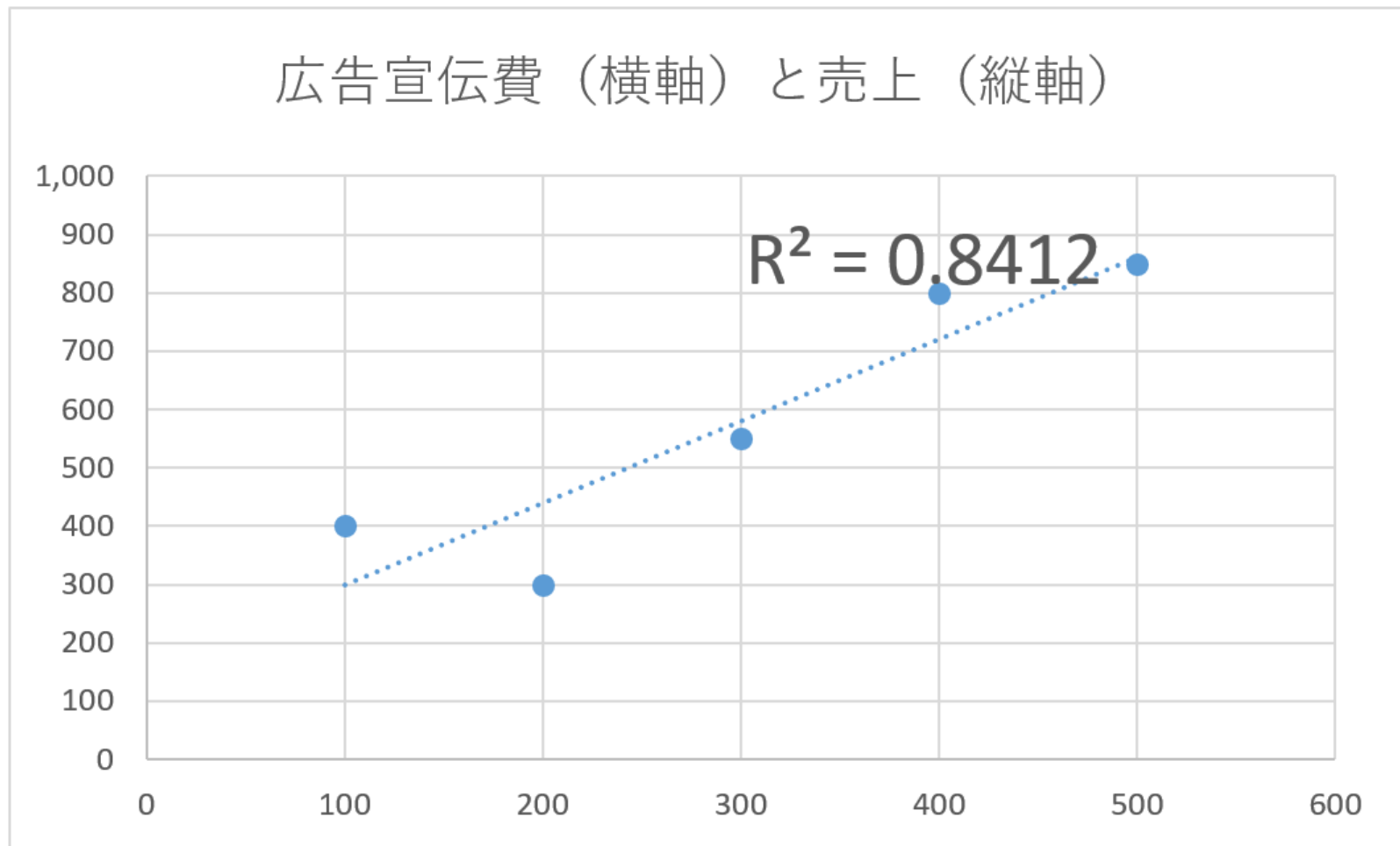
## 1. データ

- 広告宣伝費と売上
- 「広告宣伝費と売上に関係はありそうか？」

広告宣伝費（横軸）と売上（縦軸）					
	4月	5月	6月	7月	8月
広告宣伝費	100	200	300	400	500
売上	400	300	550	800	850

# 相関分析

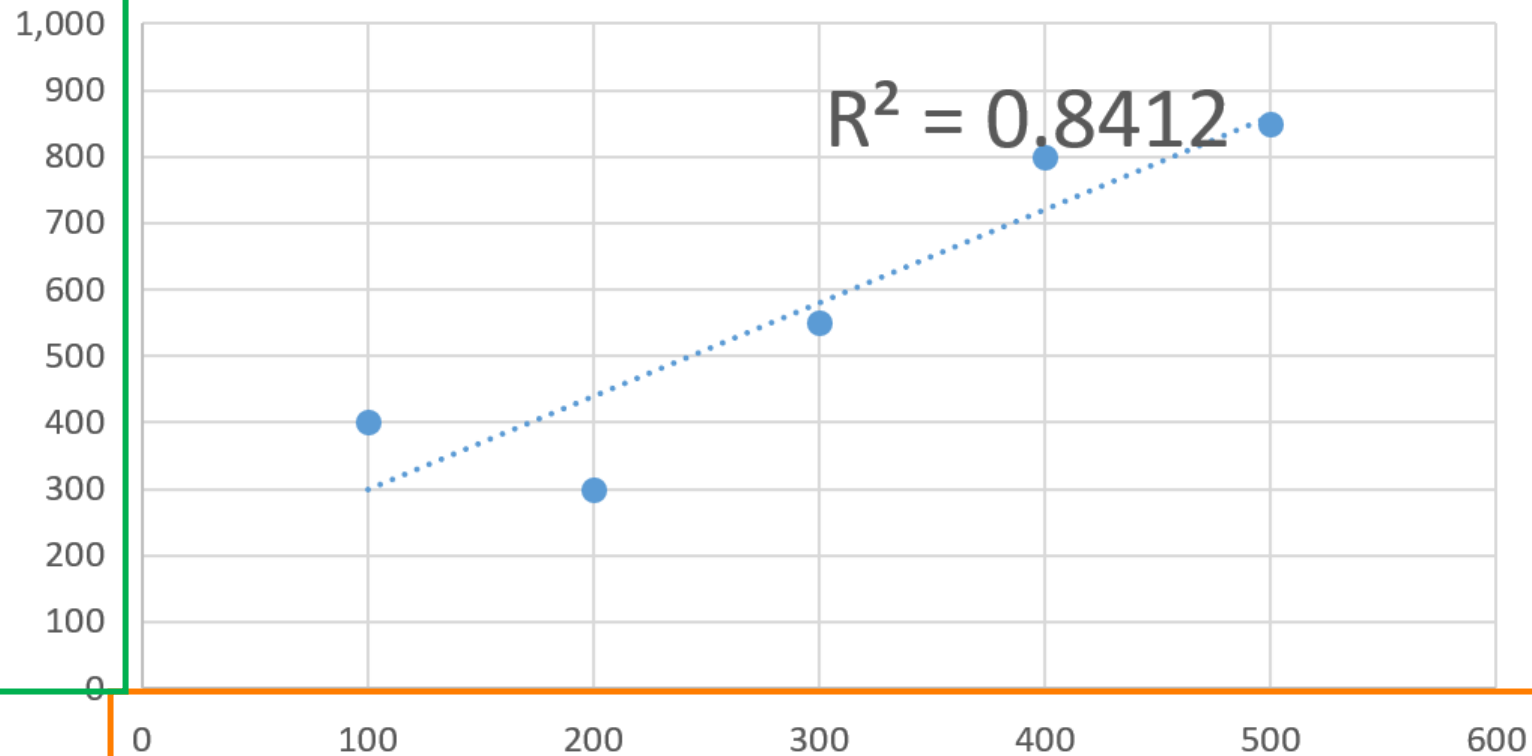
## エクセル「散布図」



# 相関分析

## エクセル「散布図」

広告宣伝費（横軸）と売上（縦軸）



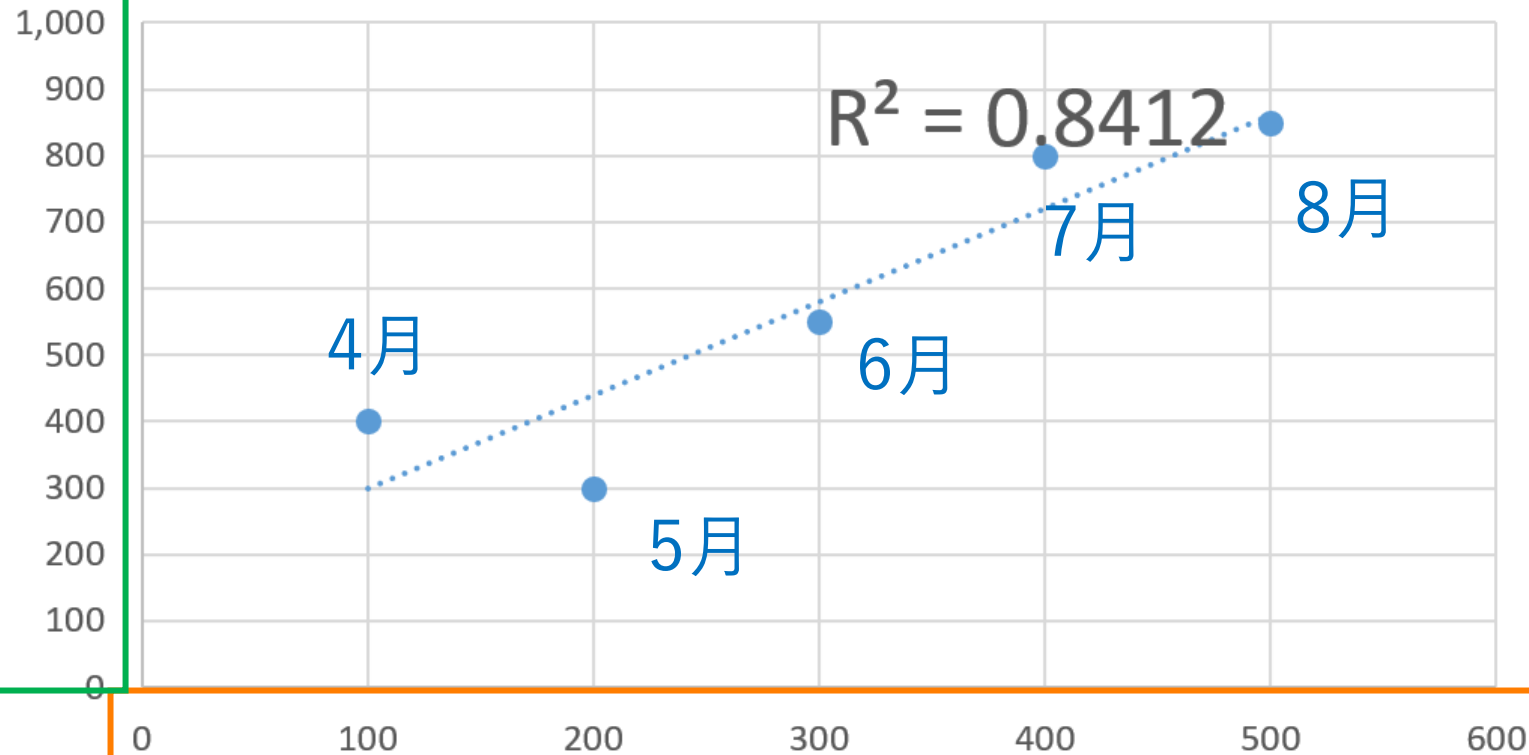
### 軸の意味

- ・ 横軸：広告宣伝費
- ・ 縦軸：売上

# 相関分析

## エクセル「散布図」

広告宣伝費（横軸）と売上（縦軸）



軸の意味

- ・ 横軸：広告宣伝費
- ・ 縦軸：売上

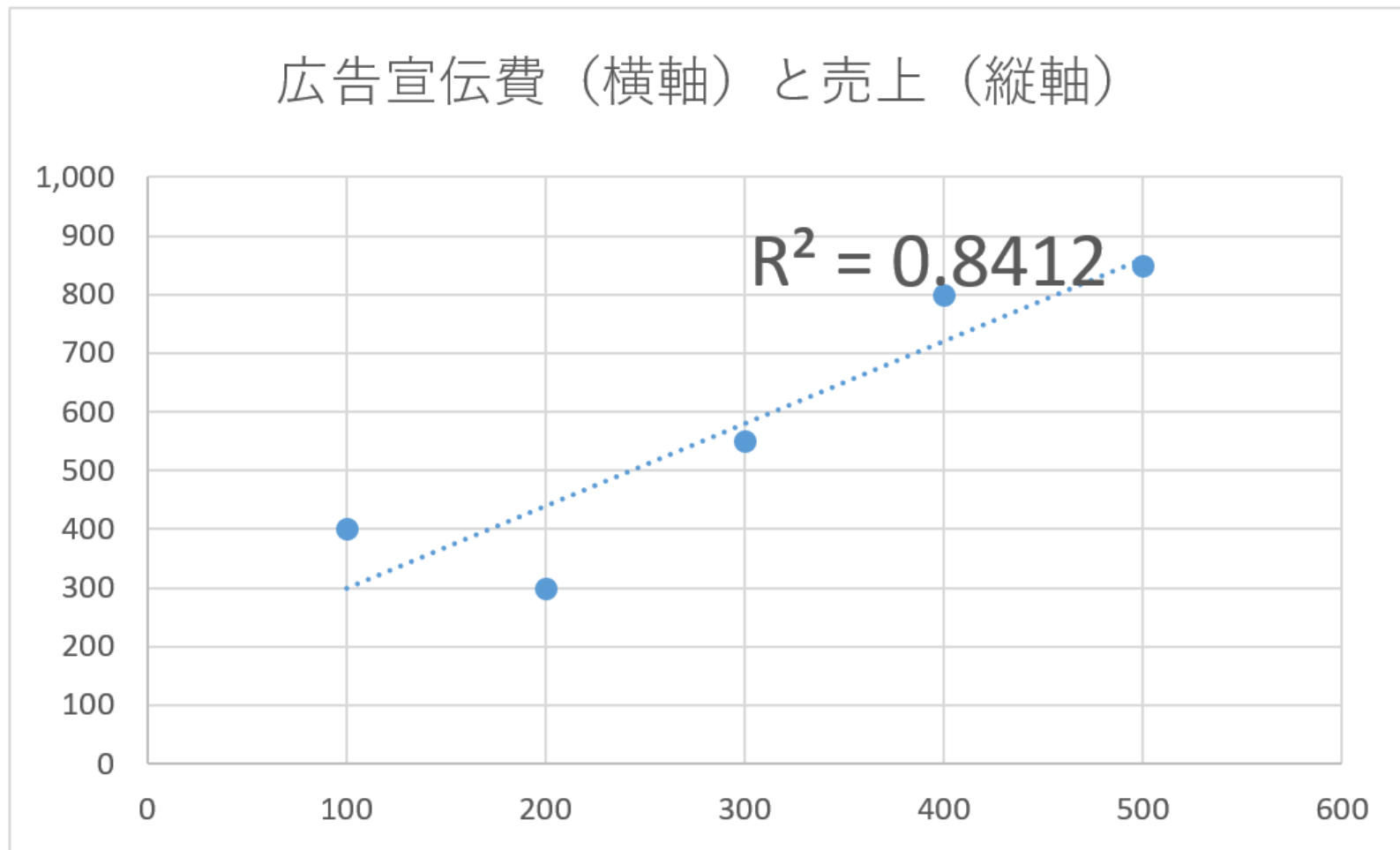
点の意味

- ・ 4月～8月それぞれ
- ・ 点が5個



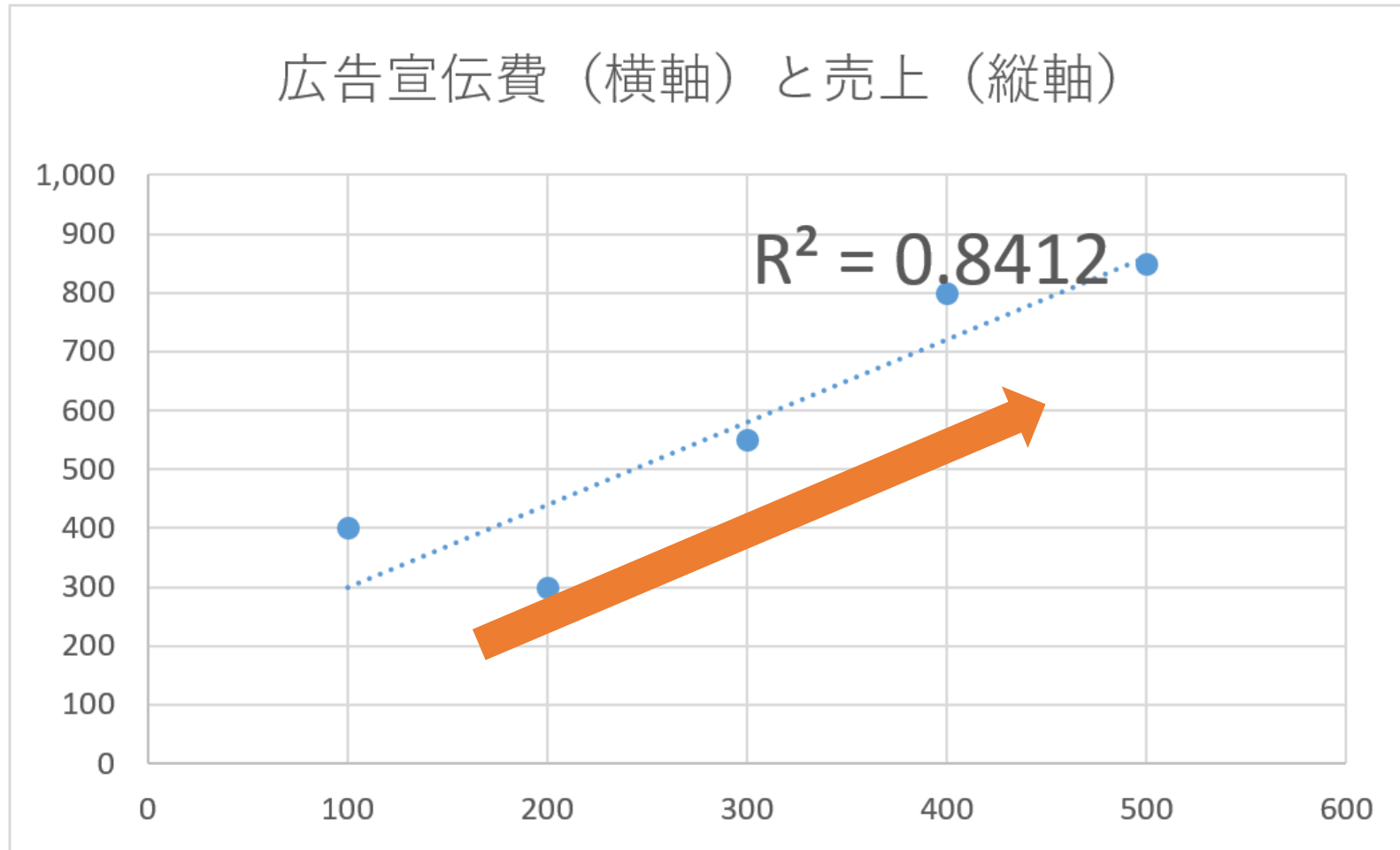
# 相関分析

2つの数字に関係はあるのか？



# 相関分析

広告宣伝費が多いほど、売上も多い（関係性がありそうだ）



# 今回のポイント

---

相関分析のいいところは、  
関係性の強さが数字で分かる

# 関係性



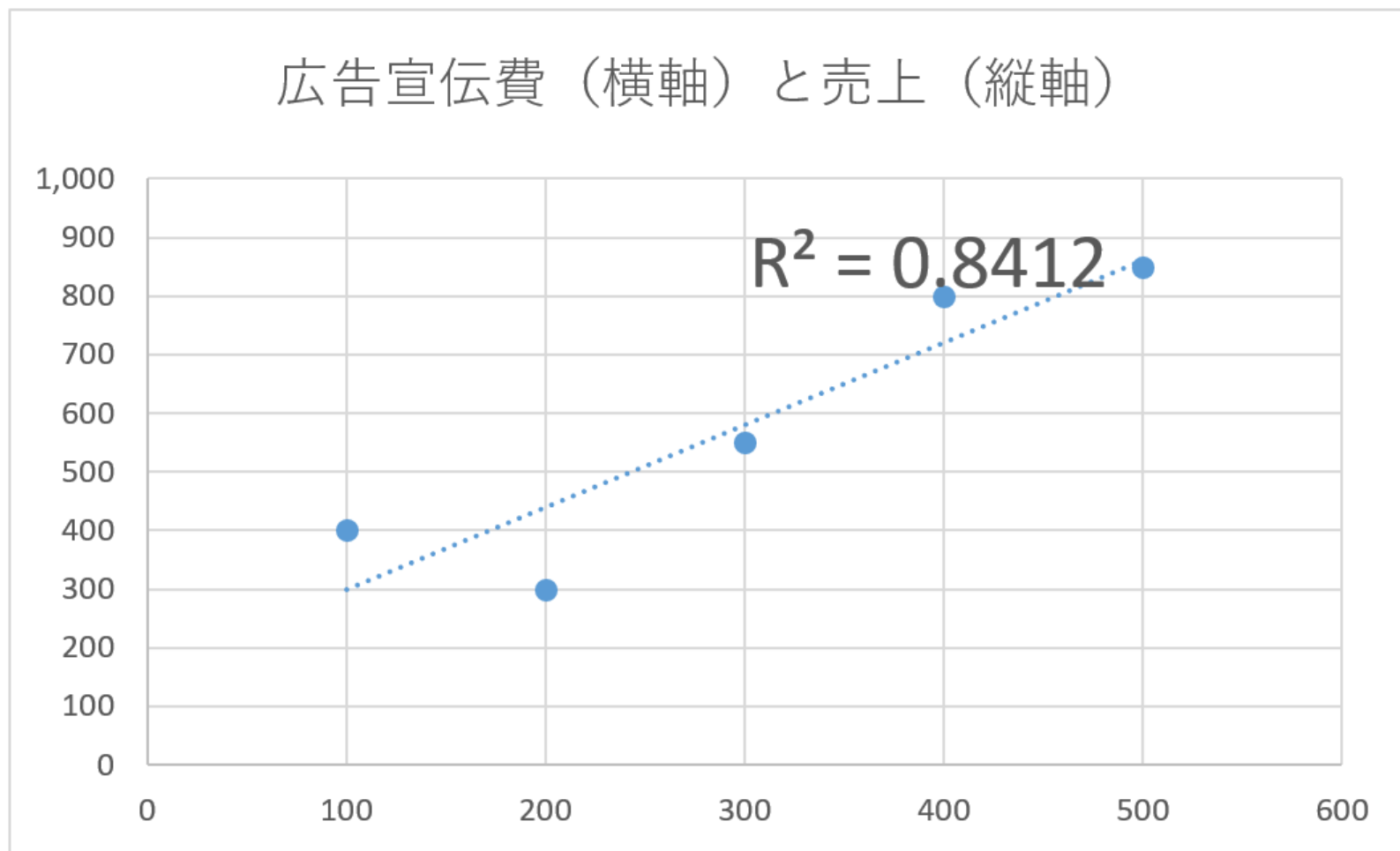
4月から8月にかけて、  
広告宣伝費を増やすと売上も増えたので、  
関係あると思います！

うーん、どれくらい関係が強いのか、  
数字で教えてもらえると分かりやすいのだが



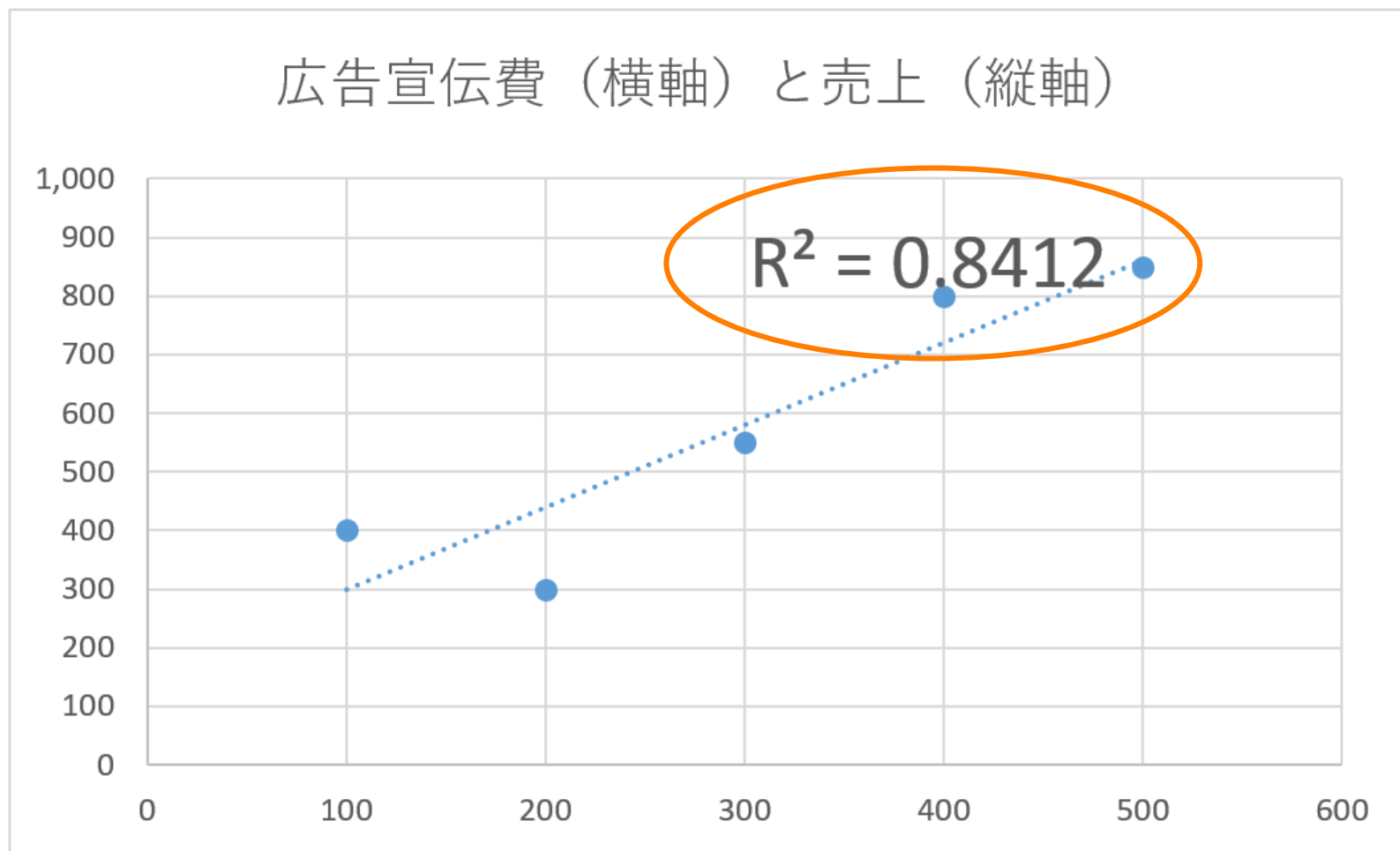
# 相関分析

## エクセル「散布図」



# 相関分析

## エクセル「散布図」



$R^2$

- 関係性の強さを示す数値

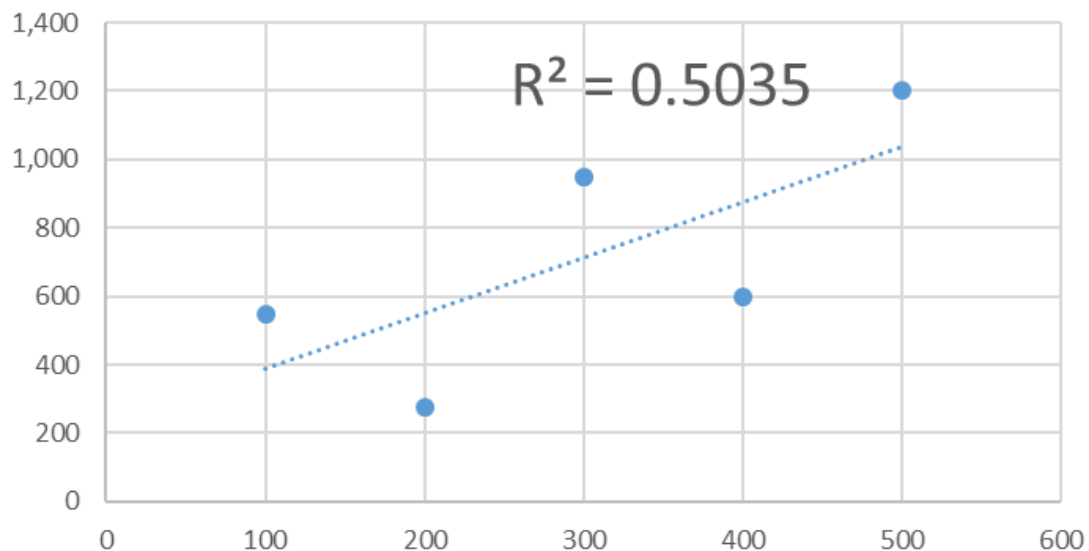
目安

- 0.7 ~ 0.8を超えると強い関係があることが多い
- あくまで目安  
(正解はない)

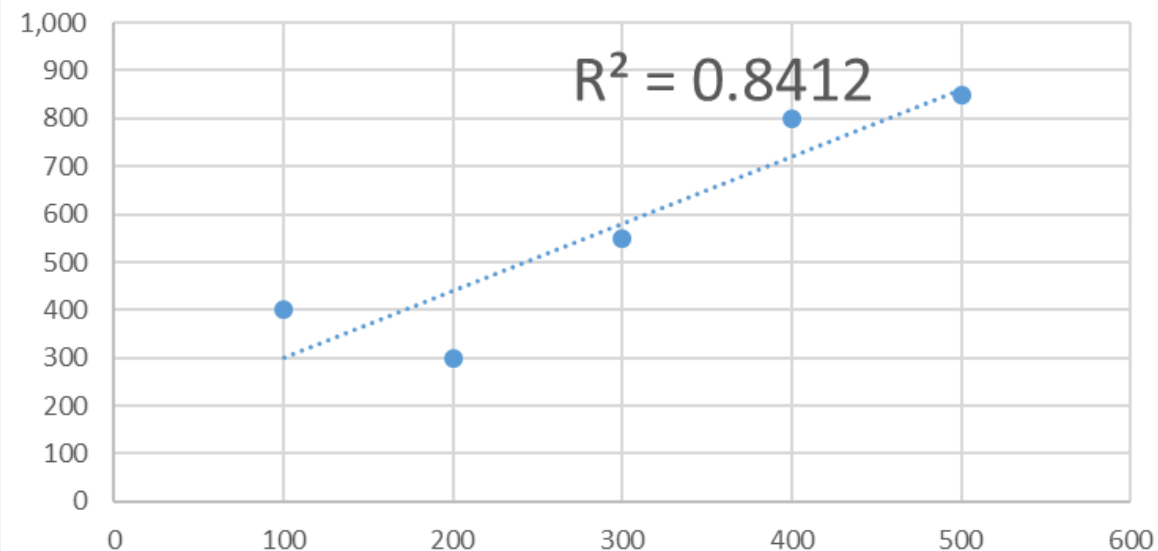
# 相関分析

左右を比べてみると、右図のほうが関係性の強さが分かる

広告宣伝費（横） × 売上（縦）



広告宣伝費（横） × 売上（縦）



# 今回のポイント

---

正の相関、負の相関

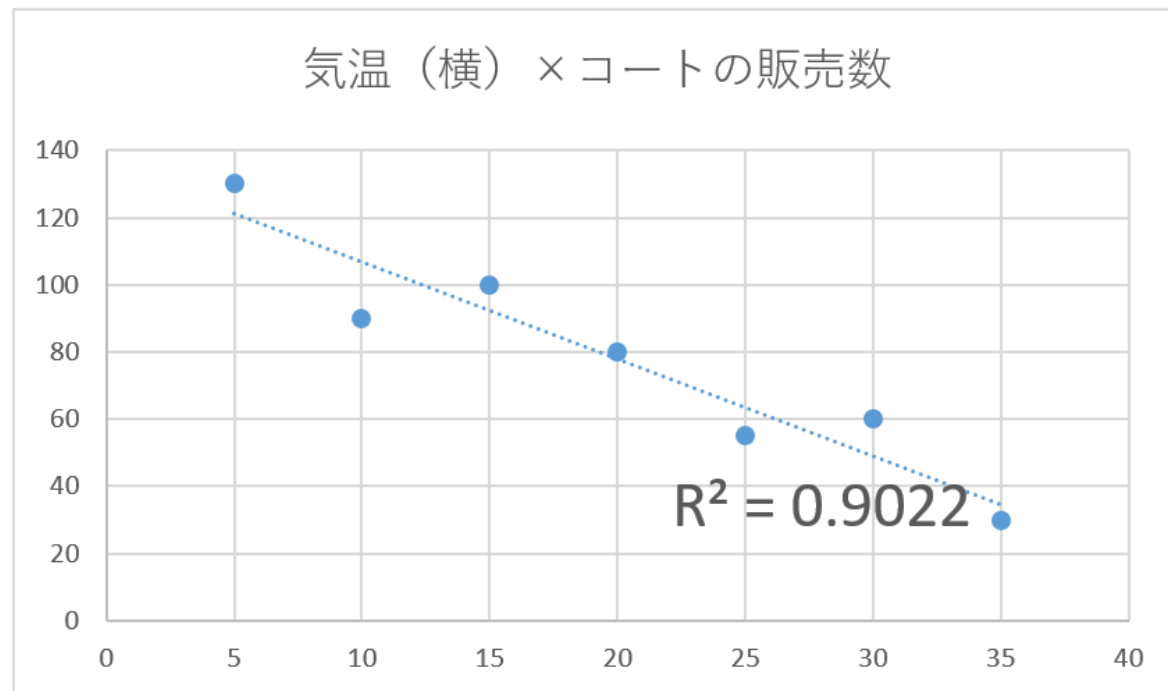
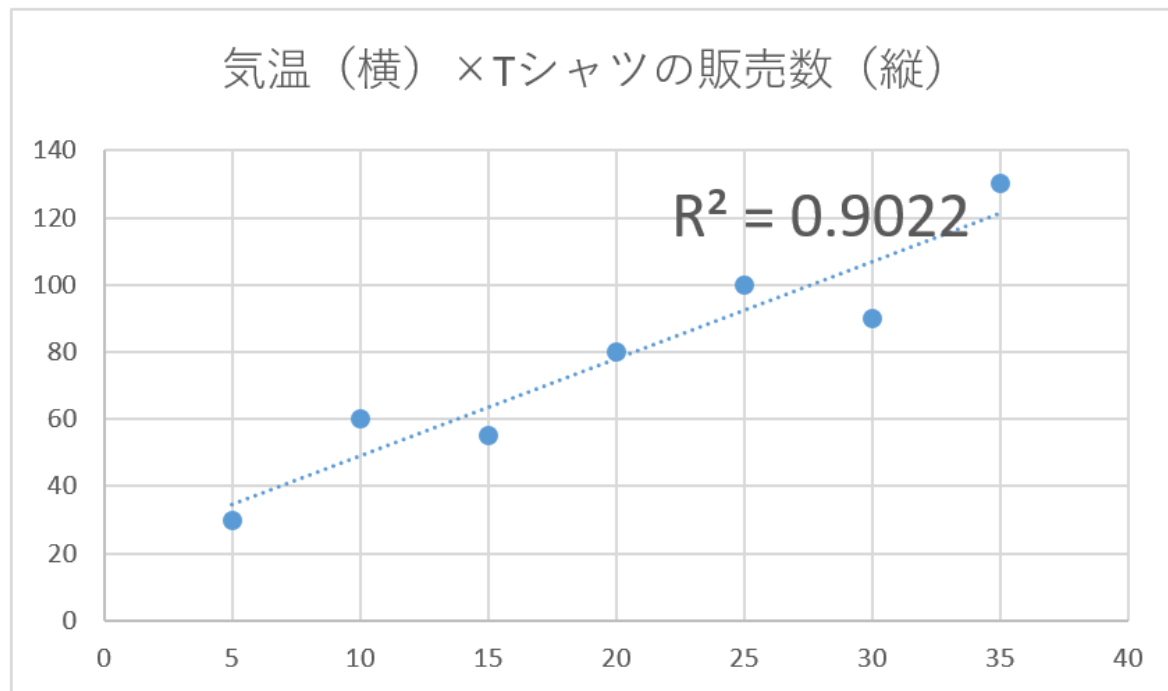


# 相関分析

## 1. 相関の種類

- 気温が上がれば、Tシャツが売れる → 正の相関
- 気温が上がれば、コートが売れない → 負の相関

→ どちらも  $R^2=0.90$  なので注意！ 相関の向きもチェック！



# 今回のポイント

---

補足

# 相関分析

---

## 1. 散布図

- グラフ → 「散布図」

## 2. 相関分析

- 「近似曲線の追加」
- 「グラフにR-2を表示する」

## 3. 相関分析

- 回帰分析、という言葉が使われることもあります

# 大量のデータを読み解く

1. 代表値(1) 平均値
2. 代表値(2) 中央値
3. 代表値(3) 最頻値
4. 分布(1) ヒストグラム
5. 分布(2) 標準偏差
6. 分布(3) 混合分布
7. 分布(4) パレート図
8. まとめ(1) 代表値と分布
9. 傾向(1) 推移
10. 傾向(2) ヒートマップ
11. 関係性(1) 相関分析
12. 関係性(2) 因果関係
13. 関係性(3) 第三因子
14. 関係性(4) 混合グループ
15. 関係性(5) 外れ値
16. まとめ(2) 傾向、関係性

# 今回のポイント

---

因果関係

# 因果関係

---

## 1. 相関関係

- 気温とTシャツの販売数には関係がある
  - 気温が上がるから、Tシャツが売れる
  - × Tシャツが売れるから、気温が上がる
- 相関分析は、どちらが原因で、どちらが結果か分かりにくい場合も

## 2. 因果関係

- 原因と結果
- 「あれ、それ逆じゃない？」と疑問を思うことが大事

# 因果関係

---

## 1. どちらが原因と思われるか？

- 部屋の広さと、不動産の価格
  - 部屋が広いから、不動産の価格も高い
- 売上と、広告宣伝費
  - 広告宣伝するから、売上も増える
- 早起きする人は、年収が高い
  - 早起きして仕事するから年収が高い？
  - 年収が高い高齢者は、朝起きるのが早い？
  - 早起きして年収の高い人に聞いて確認してみたい

# 今回のポイント

---

因果関係



# 大量のデータを読み解く

1. 代表値(1) 平均値
2. 代表値(2) 中央値
3. 代表値(3) 最頻値
4. 分布(1) ヒストグラム
5. 分布(2) 標準偏差
6. 分布(3) 混合分布
7. 分布(4) パレート図
8. まとめ(1) 代表値と分布
9. 傾向(1) 推移
10. 傾向(2) ヒートマップ
11. 関係性(1) 相関分析
12. 関係性(2) 因果関係
13. 関係性(3) 第三因子
14. 関係性(4) 混合グループ
15. 関係性(5) 外れ値
16. まとめ(2) 傾向、関係性

# 今回のポイント

---

第三因子

# 第三因子



商品ごとの販売データを見たところ、  
アイスの販売数とビールの販売数の関係が強い  
ので、アイス卖了らビールも売れます！

アイス卖了ら、ビールも売れる・・・？  
それってつまり・・・？



# 第三因子

---

## 1. データ

- アイスの販売数と、ビールの販売数には関係がある
- がんばってアイス卖了ば、ビールの販売数も増えるはず

アイスの  
販売数



ビールの  
販売数

# 関係性



商品ごとの販売データを見たところ、  
アイスの販売数とビールの販売数の関係が強い  
ので、アイス卖了ばビールも売れます！

アイス卖了ば、ビールも売れる・・・？  
それってつまり・・・？



# 関係性



商品ごとの販売データを見たところ、  
アイスの販売数とビールの販売数の関係が強い  
ので、アイス卖了らビールも売れます！

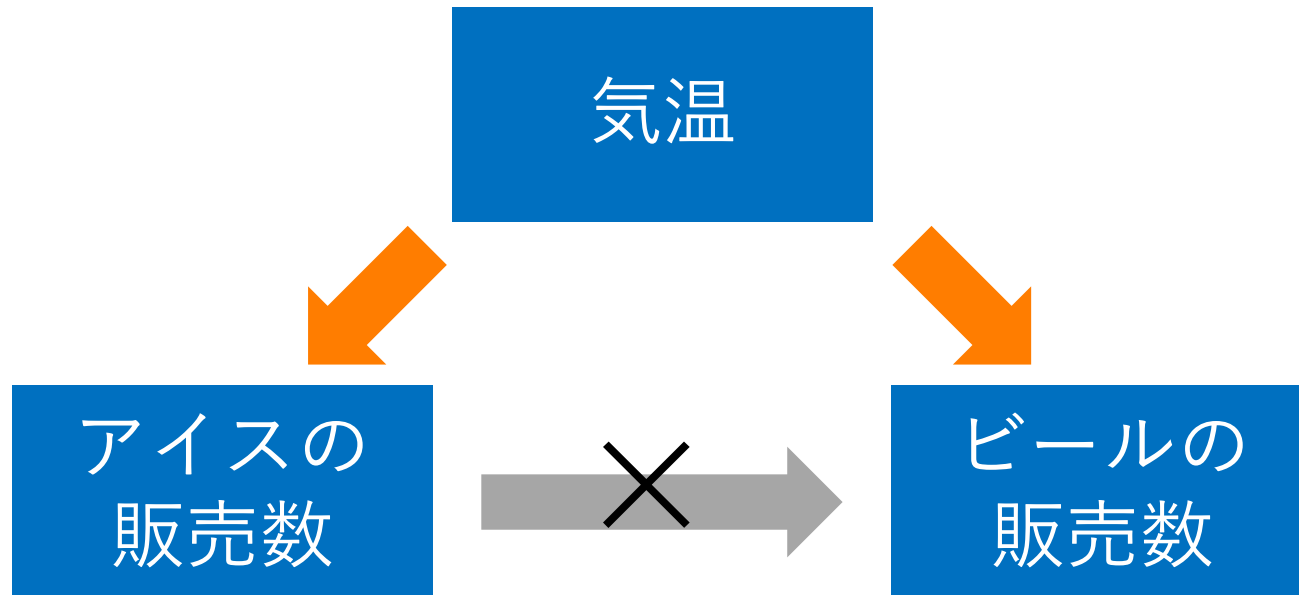
アイス卖了ら、ビールも売れる・・・？  
それってつまり暑いから両方売れただけだね？



# 第三因子

## 1. データ

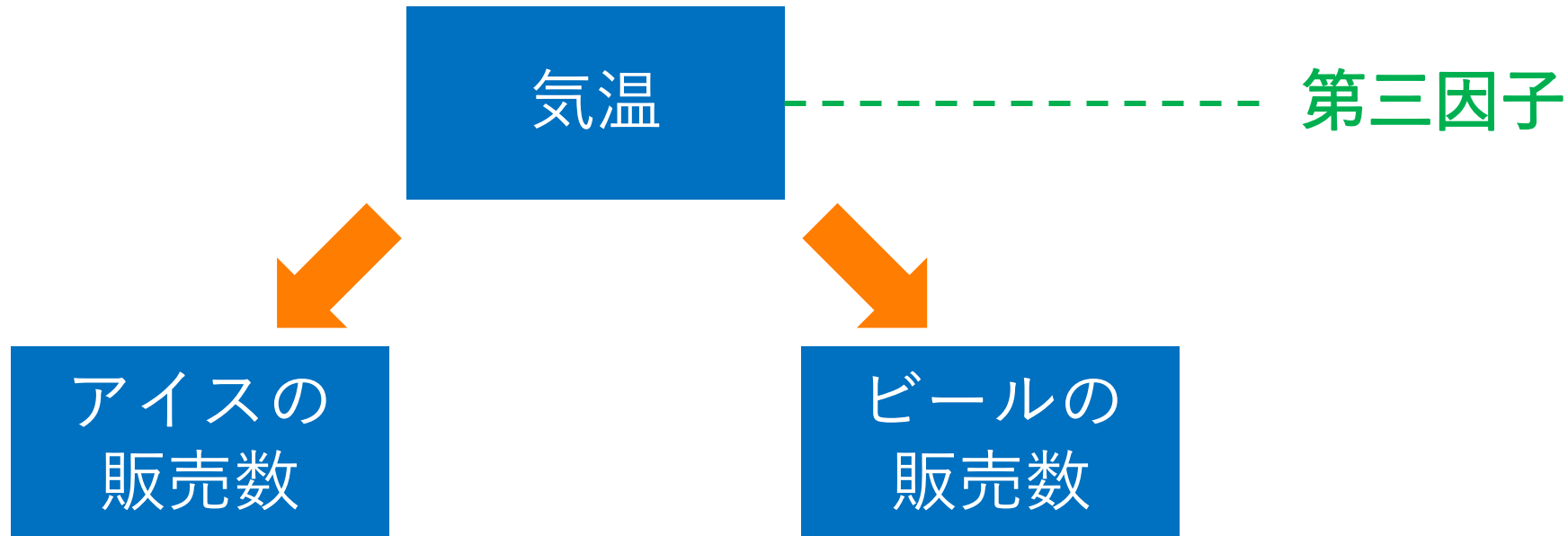
- アイスも、ビールも、気温が高いから売れただけ
- **アイスとビールには直接関係はない**



# 第三因子

## 1. データ

- 隠れていた因子（原因）を第三因子と呼ぶ





# 第三因子

---

## 1. 意味

- ある2つのデータに関係がある場合、  
実は隠れている因子（原因）があるのでは？

## 2. 例

- ある有名なタレントが、  
(1)カメラと(2)化粧品を紹介して、  
2つの商品の売上が大きく増えた  
→ (1)カメラと(2)化粧品、に因果関係がある、わけではない

# 今回のポイント

---

関係性を分析するときには、  
隠れた第三因子がないか考える

# 大量のデータを読み解く

1. 代表値(1) 平均値
2. 代表値(2) 中央値
3. 代表値(3) 最頻値
4. 分布(1) ヒストグラム
5. 分布(2) 標準偏差
6. 分布(3) 混合分布
7. 分布(4) パレート図
8. まとめ(1) 代表値と分布
9. 傾向(1) 推移
10. 傾向(2) ヒートマップ
11. 関係性(1) 相関分析
12. 関係性(2) 因果関係
13. 関係性(3) 第三因子
14. 関係性(4) 混合グループ
15. 関係性(5) 外れ値
16. まとめ(2) 傾向、関係性

# 今回のポイント

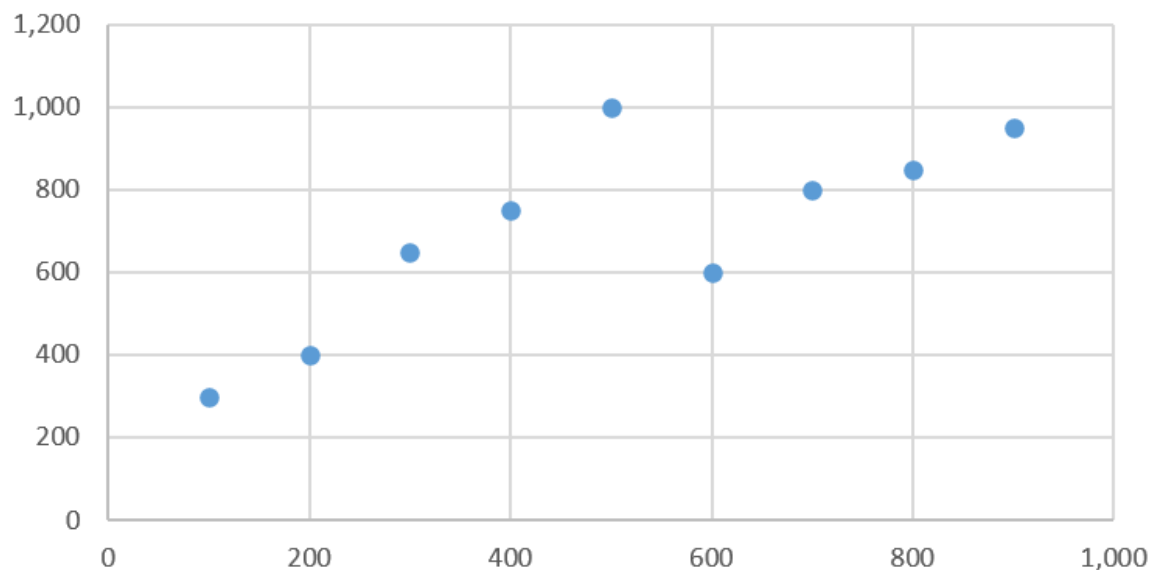
---

混合グループ

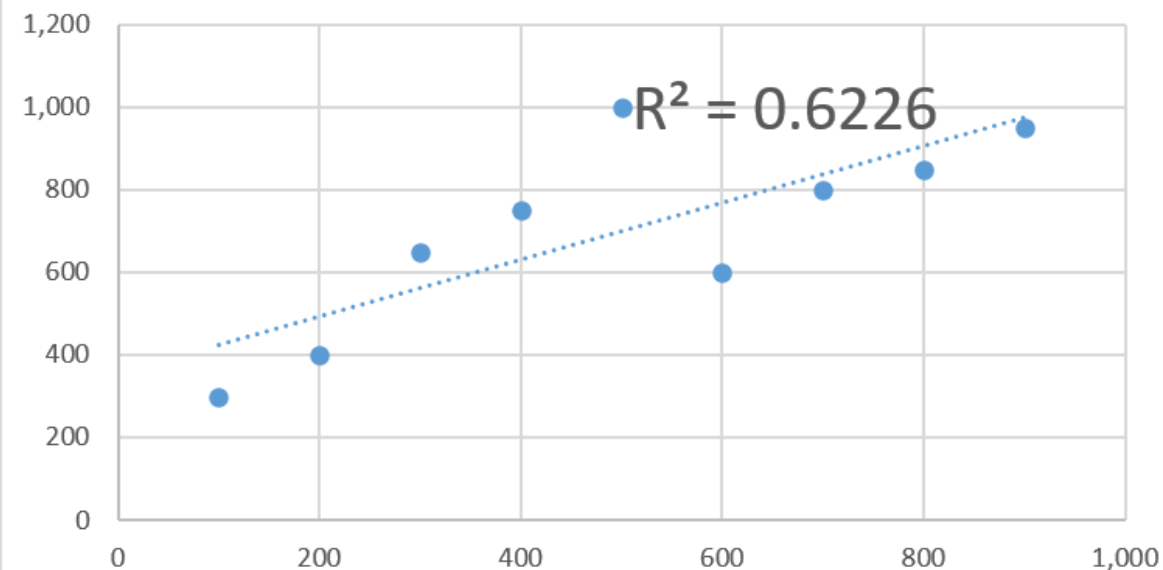
# 混合グループ

## 1. 広告宣伝費と売上、それほど強い関係はなさそう？

広告宣伝費（横） × 売上（縦）



広告宣伝費（横） × 売上（縦）

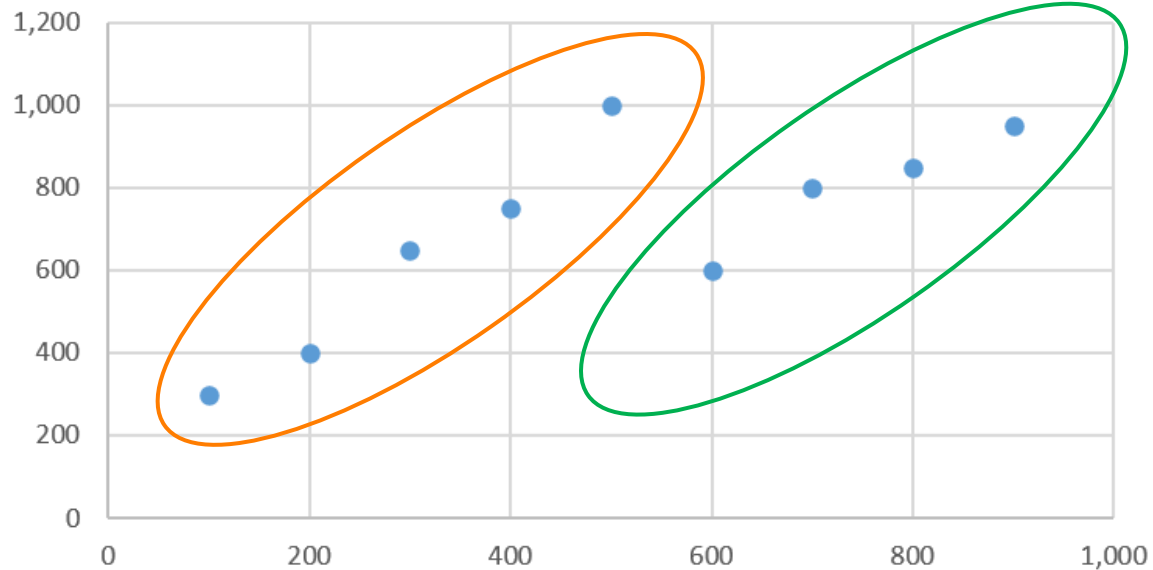


# 混合グループ

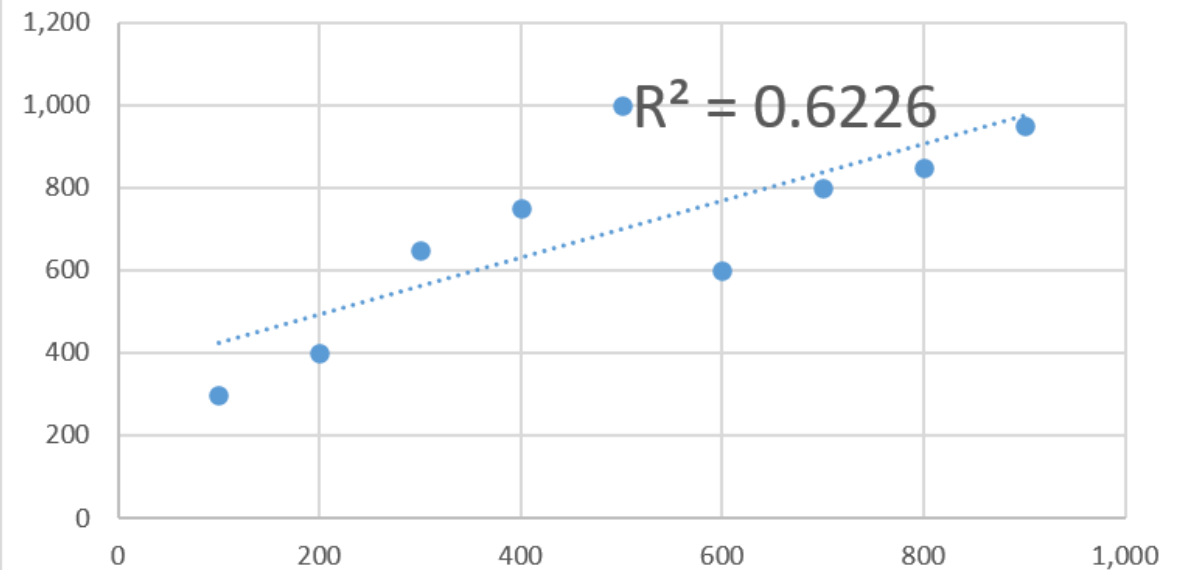
1. 広告宣伝費と売上、それほど強い関係はなさそう？

→ 実は2つの別のグループが混ざっているから関係が低く見える？

広告宣伝費（横） × 売上（縦）



広告宣伝費（横） × 売上（縦）

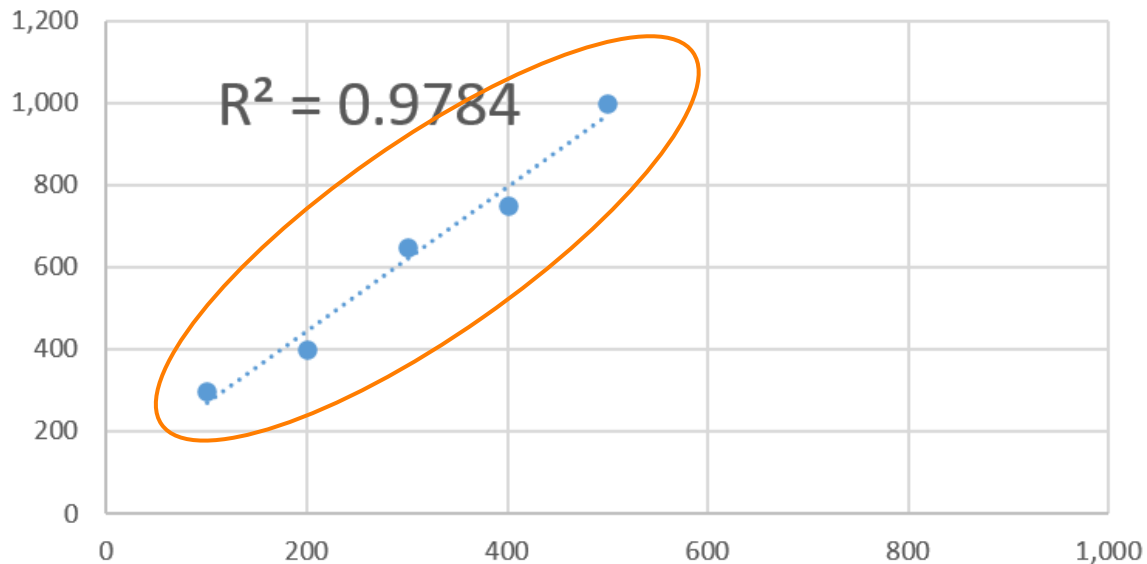


# 混合グループ

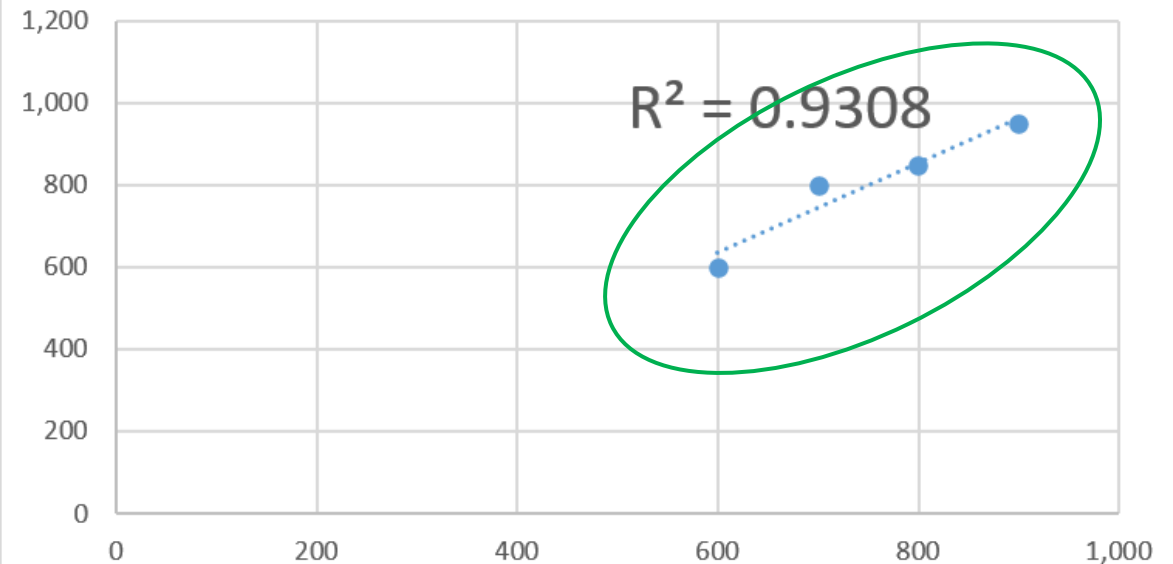
1. 広告宣伝費と売上、それほど強い関係はなさそう？

→ データを分けてみると、それぞれ関係が強いことが分かる

広告宣伝費（横） × 売上（縦）



広告宣伝費（横） × 売上（縦）



# 混合グループ

---

## 1. 意味

- ある2つのデータが混ざっていると、  
関係性の強さが分かりにくい場合がある
  - 商品Aと商品Bの販売データ
  - 大都市と地方の販売データ

## 2. ポイント

- できるだけ近しいデータを集める
- タイプのちがうデータは、分けて計算する



# 大量のデータを読み解く

1. 代表値(1) 平均値
2. 代表値(2) 中央値
3. 代表値(3) 最頻値
4. 分布(1) ヒストグラム
5. 分布(2) 標準偏差
6. 分布(3) 混合分布
7. 分布(4) パレート図
8. まとめ(1) 代表値と分布
9. 傾向(1) 推移
10. 傾向(2) ヒートマップ
11. 関係性(1) 相関分析
12. 関係性(2) 因果関係
13. 関係性(3) 第三因子
14. 関係性(4) 混合グループ
15. 関係性(5) 外れ値
16. まとめ(2) 傾向、関係性

# 今回のポイント

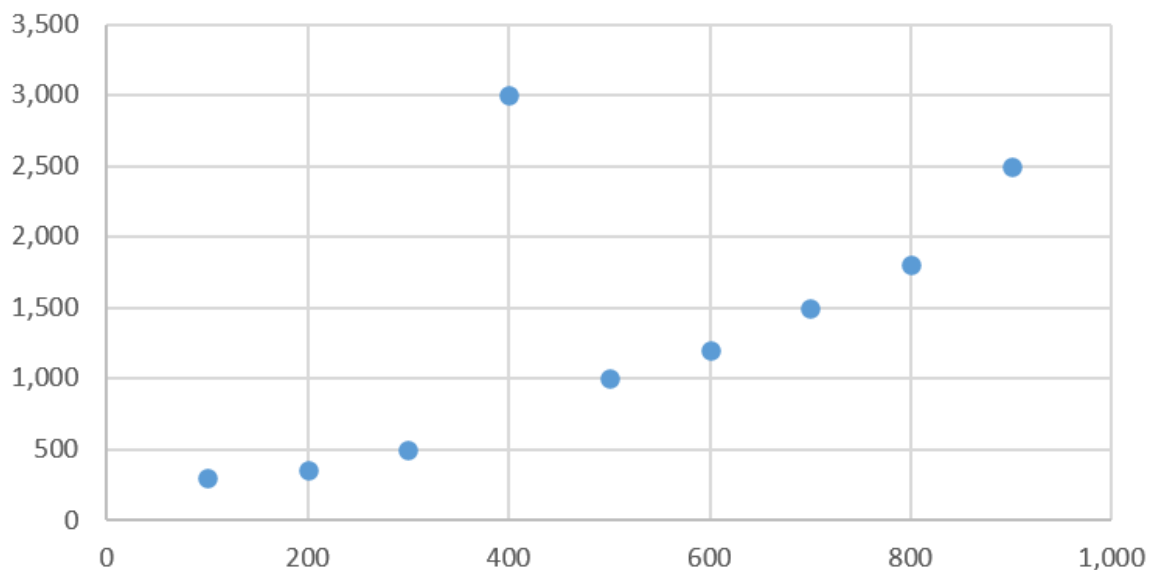
---

外れ値

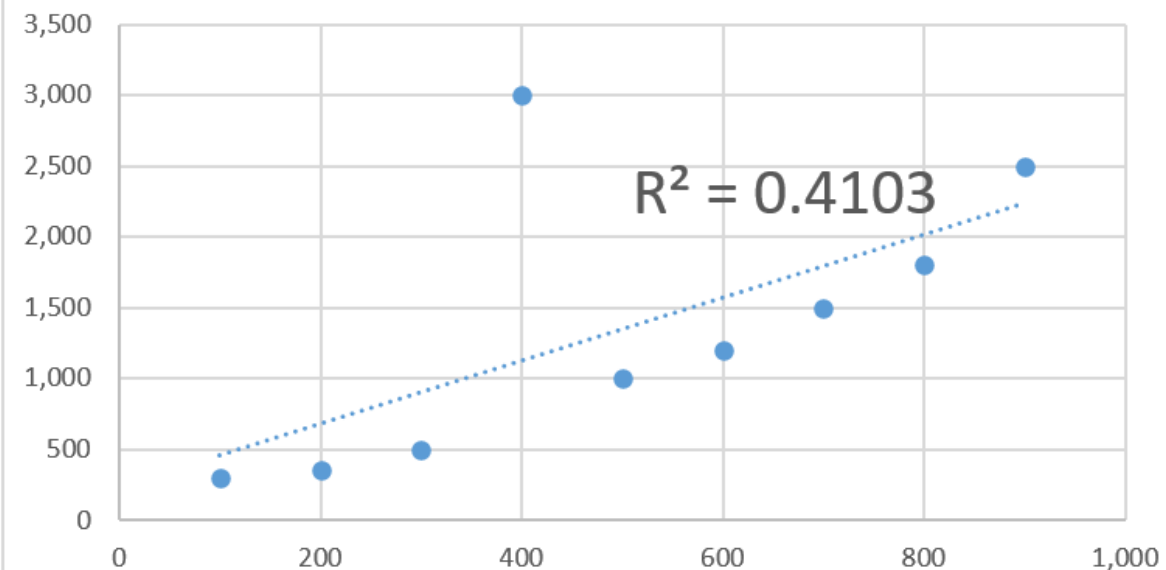
# 外れ値

## 1. 広告宣伝費と売上、それほど強い関係はなさそう？

広告宣伝費（横） × 売上（縦）



広告宣伝費（横） × 売上（縦）

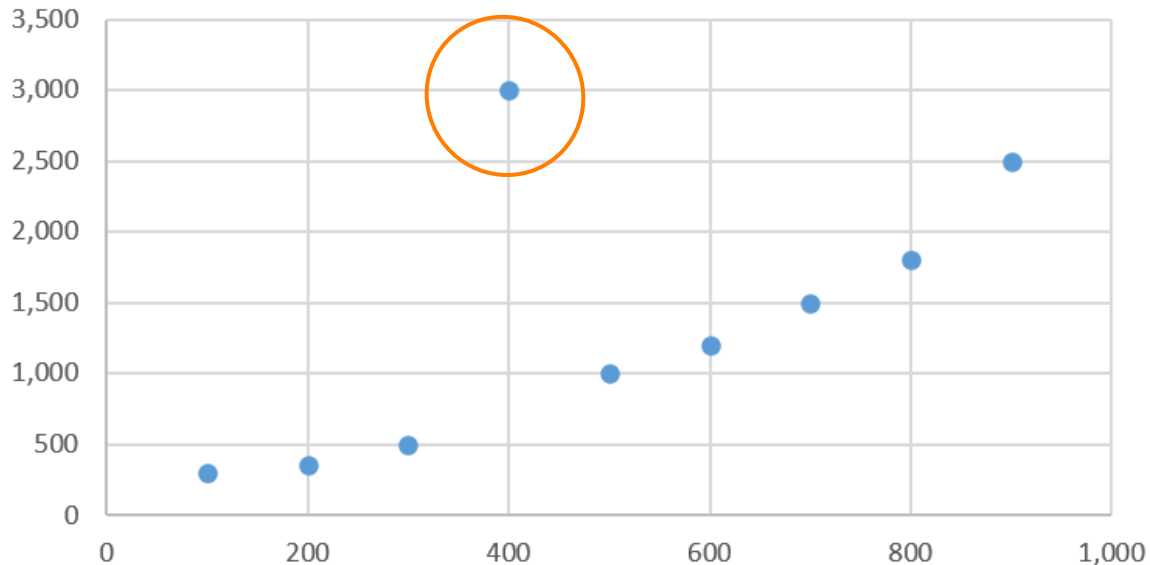


# 外れ値

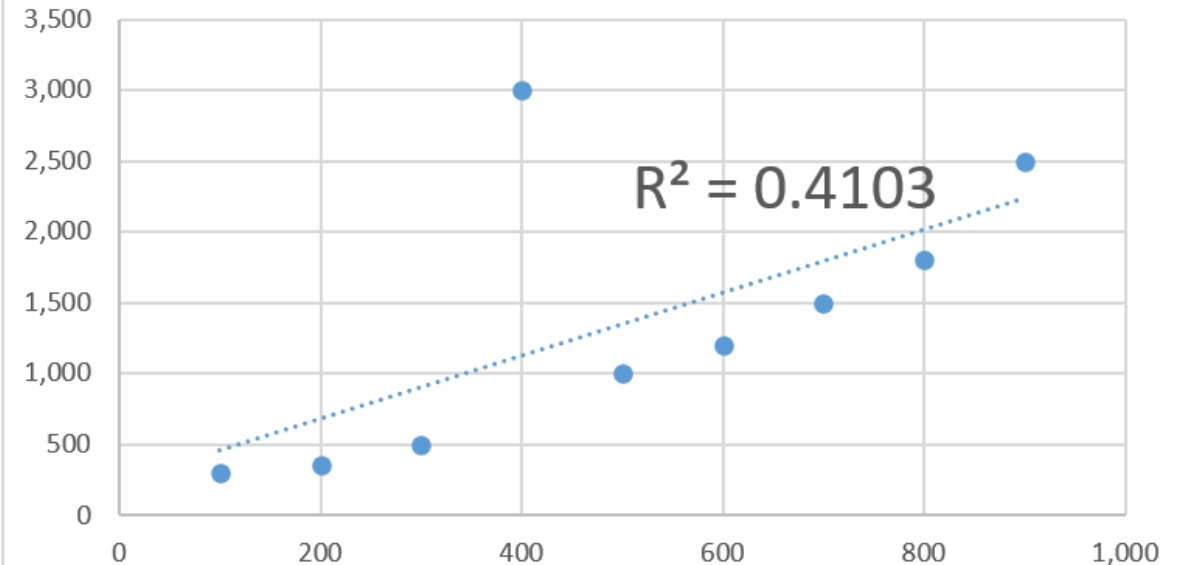
1. 広告宣伝費と売上、それほど強い関係はなさそう？

→ 大きく外れている値があることが分かる

広告宣伝費（横） × 売上（縦）



広告宣伝費（横） × 売上（縦）

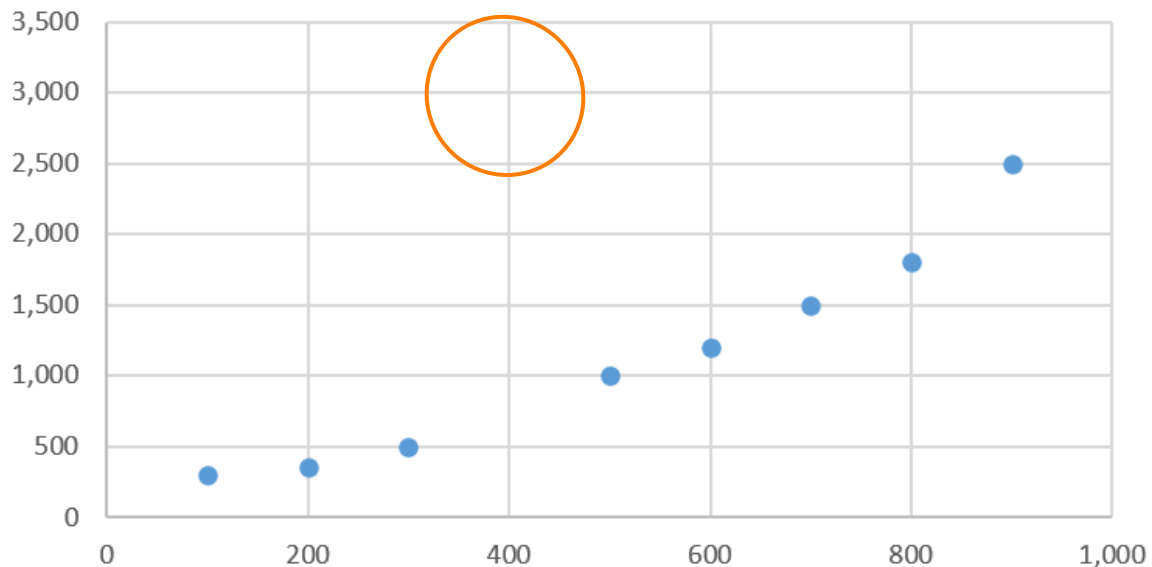


# 外れ値

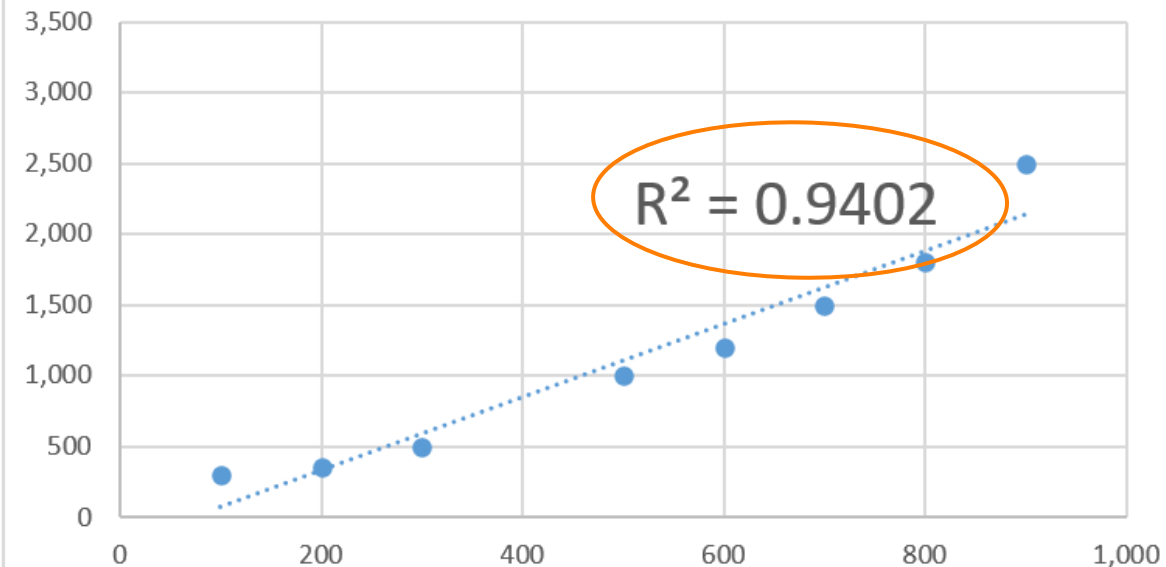
1. 広告宣伝費と売上、それほど強い関係はなさそう？

→ この値を削除すると、関係が強いことが分かる（外れ値）

広告宣伝費（横） × 売上（縦）



広告宣伝費（横） × 売上（縦）



# 外れ値

---

## 1. 意味

- 大きく外れた値があったときに、それを外すことで、関係性が見えてくることがある

## 2. 季節要因

- 毎年12月になると、クリスマス商戦で販売数が増える

## 3. 一時要因

- この年は災害があったため、売上が大きく減った

# 今回のポイント

---

外れ値