2-2 データを説明する

東京大学 数理・情報教育研究センター
2020年5月11日

概要

- 本節では、データについて相手に説明するために必要な方法として、グラフによる可視化とデータの比較方法等について学びます。

- グラフにはさまざまな種類があり、それぞれの特徴やどのような時に使用するかについて学び、適切な可視化方法を選択して他者に説明できることを目標とします。

- データの比較方法として、ここでは
  - 条件をそろえた比較
  - 処理の前後での比較
  - A／Bテスト

  の３つについて学びます。

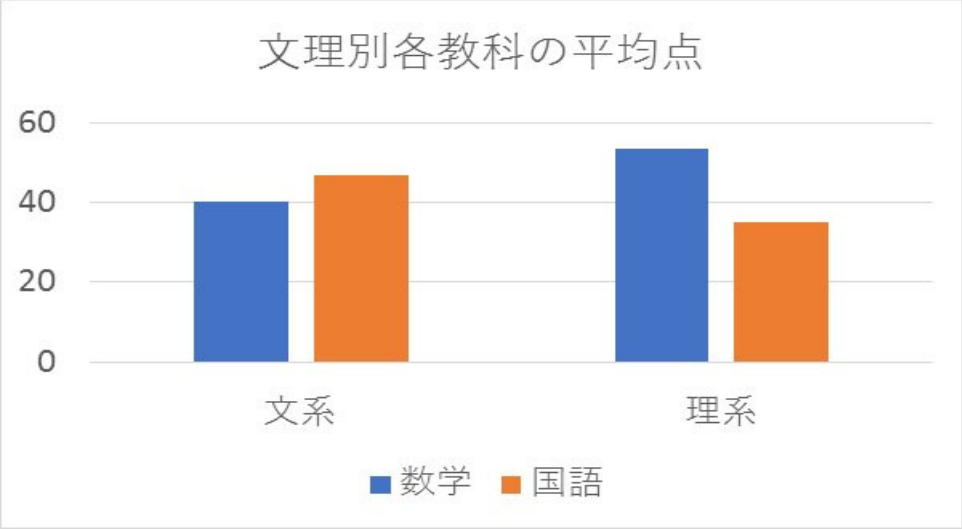
本教材の目次

データの可視化：棒グラフ

- データの可視化の方法として各種グラフを活用すると良いでしょう。
- 以下ではグラフ毎の特徴について説明します。

<u>棒グラフ</u>

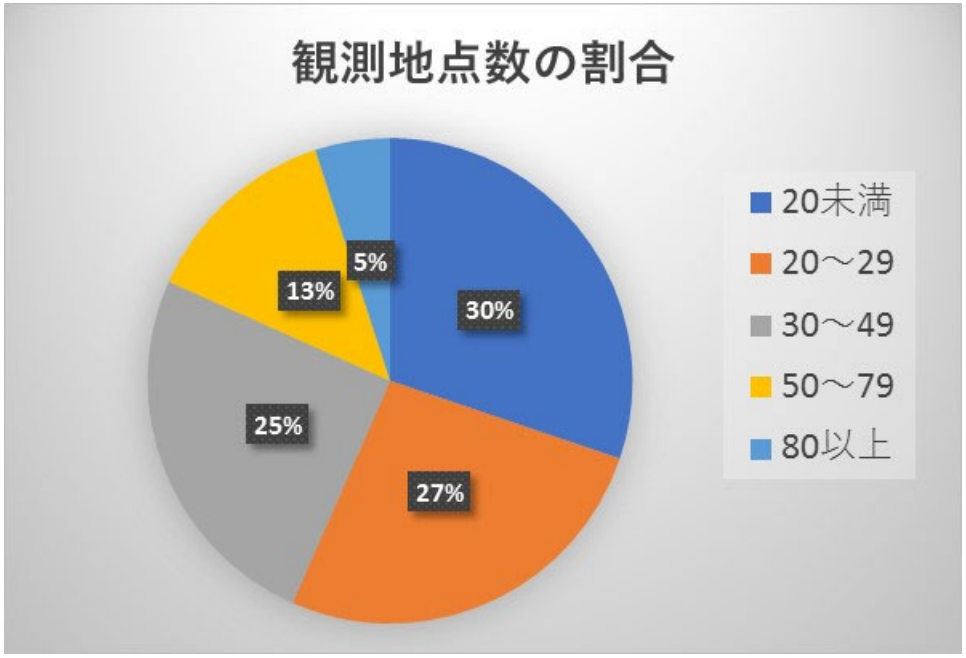
- 長方形の棒を使ってデータを表現し、データの値を比較したい時に使われます。
- 複数項目を表示する時は色を変えて並べます。



※軸は 0 から始めます。

データの可視化：円グラフ

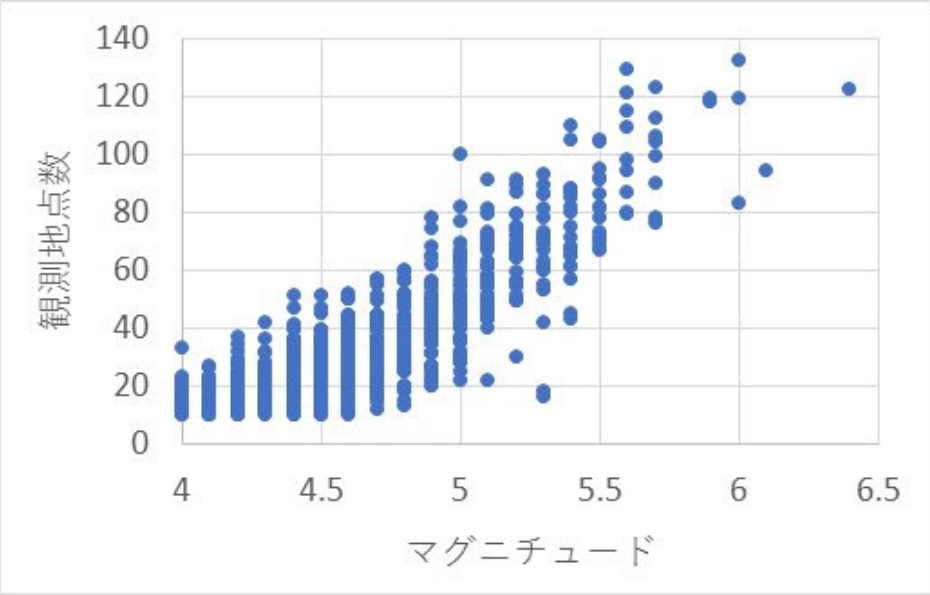
- 円グラフは、データの各数値を円の中の扇で表現し、割合化したもの。数値の比率を見る時に用いられます。



フィジーの地震における観測時点数の割合

データの可視化：散布図

- 前述の散布図は、データの2種類の項目の関係性を見るための可視化方法の一つと言えます。



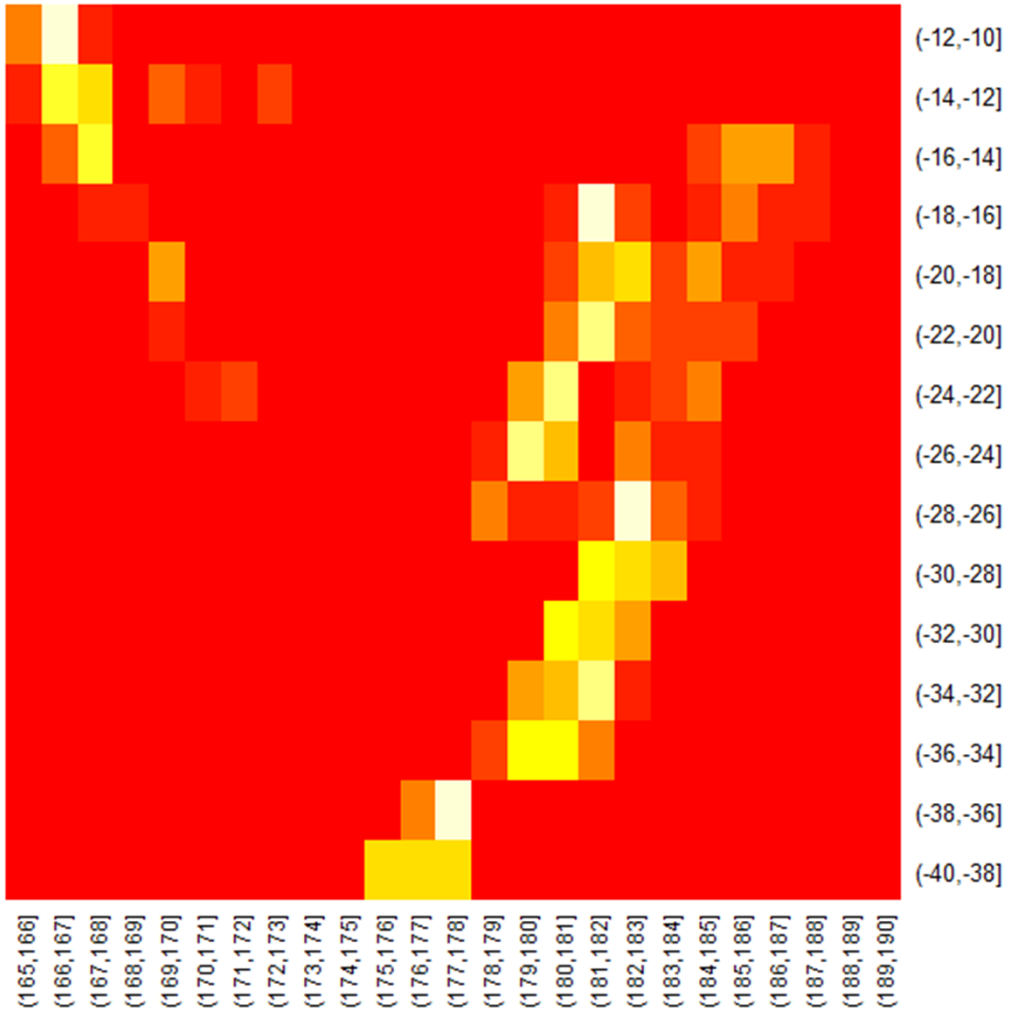
データの可視化：折れ線グラフ

- 折れ線グラフは散布図の一種で、各点を直線で結んだグラフになります。
- 横軸として時間の項目がとられることが多く、データの時間推移を見る時によく用いられます。

データの可視化：ヒートマップ

- 散布図においてデータの密度を色を変えて表現したものをヒートマップといいます。
- 右図はフィジーの地震データを緯度と経度をもとに、地震の発生頻度を色を分けて表現しています。

  ※白に近い色がより地震の頻度が高いです。



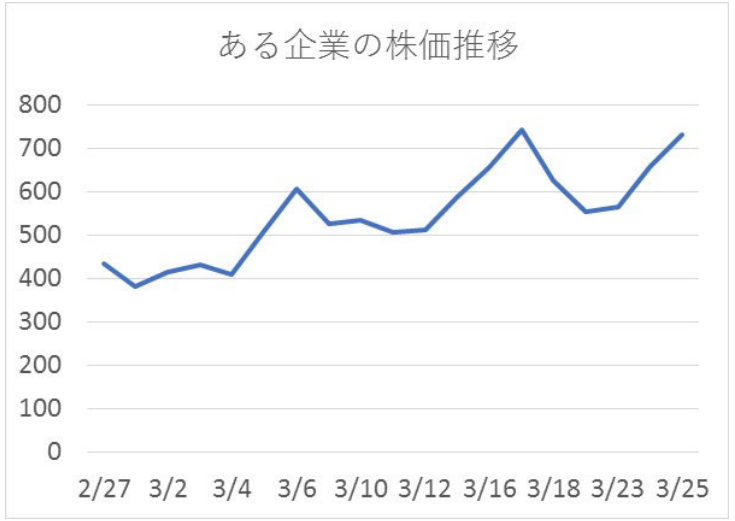
データのグラフ化

- 折れ線グラフ

日付	株価
2/27	435
2/28	382
3/2	416
3/3	433
3/4	409
3/5	508
3/6	606
・	・
・	・

対称の2項目を選んで

スプレッドシートの
機能等を用いてグラフ化



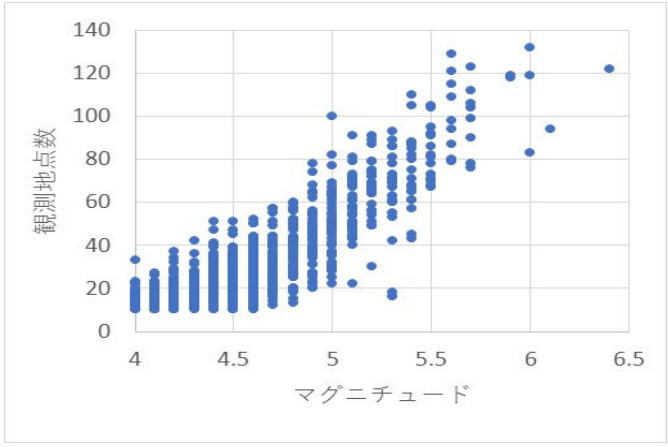
※折れ線グラフは１項目だけで作ることもできます。

- 散布図

マグニチュード	計測地点数
4.8	41
4.2	15
5.4	43
4.1	19
4	11
4	12
4.8	43
・	・
・	・

対称の2項目を選んで

スプレッドシートの
機能等を用いてグラフ化



データのグラフ化

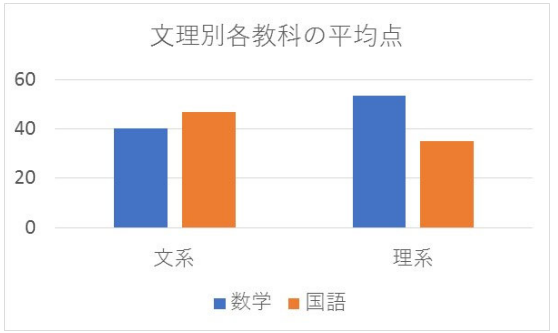
- 棒グラフ、円グラフ
  - 必要に応じてデータの合計・平均等を計算してからグラフにします。

文理	数学	国語
理系	93	46
理系	48	50
文系	41	64
文系	28	31
理系	75	23
文系	68	42
理系	63	24
・	・	・
・	・	・

文系理系別に
平均値を計算

	数学	国語
文系	40.2	47.0
理系	53.5	35.2

グラフ化

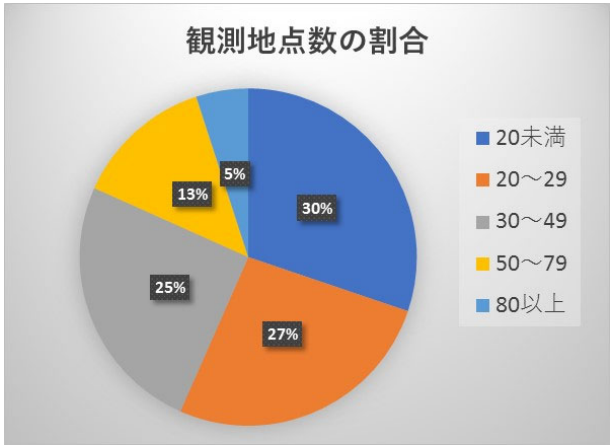


緯度	経度	地震の深さ(km)	マグニチュード	計測地点数
-20	182	562	4.8	41
-21	181	650	4.2	15
-26	184	42	5.4	43
-18	182	626	4.1	19
-20	182	649	4	11
-20	184	195	4	12
-12	166	82	4.8	43
-28	182	194	4.4	15
-29	182	211	4.7	35
-17	180	622	4.3	19
・	・	・	・	・
・	・	・	・	・
・	・	・	・	・

度数分布を
作成

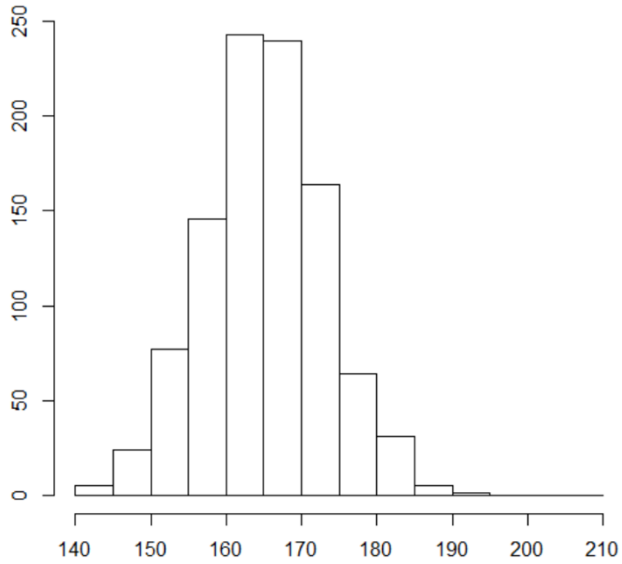
観測地点数	件数
20未満	302
20～29	264
30～49	251
50～79	133
80以上	50

グラフ化

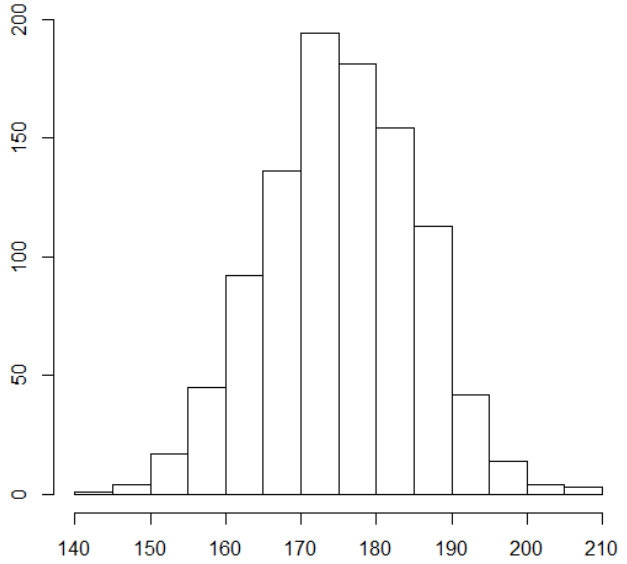


データの比較：条件をそろえた比較

- A国とB国の身長データの比較
  - 地域の差を比較したい場合は他の条件をそろえて比較します。
  - そうでないと年齢や男女のデータ数の違いによる影響が出てしまいます。



条件（年齢・性別）を揃えて比較します。

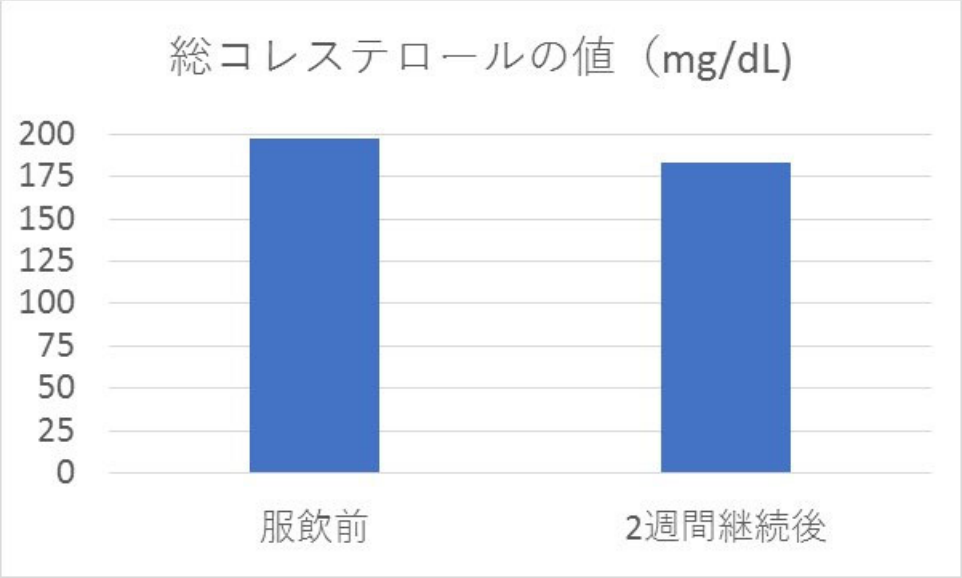


A国の身長データ（20代、男性）
平均：165.1cm
標準偏差：7.9cm

B国の身長データ（20代、男性）
平均：175.2cm
標準偏差：10.1cm

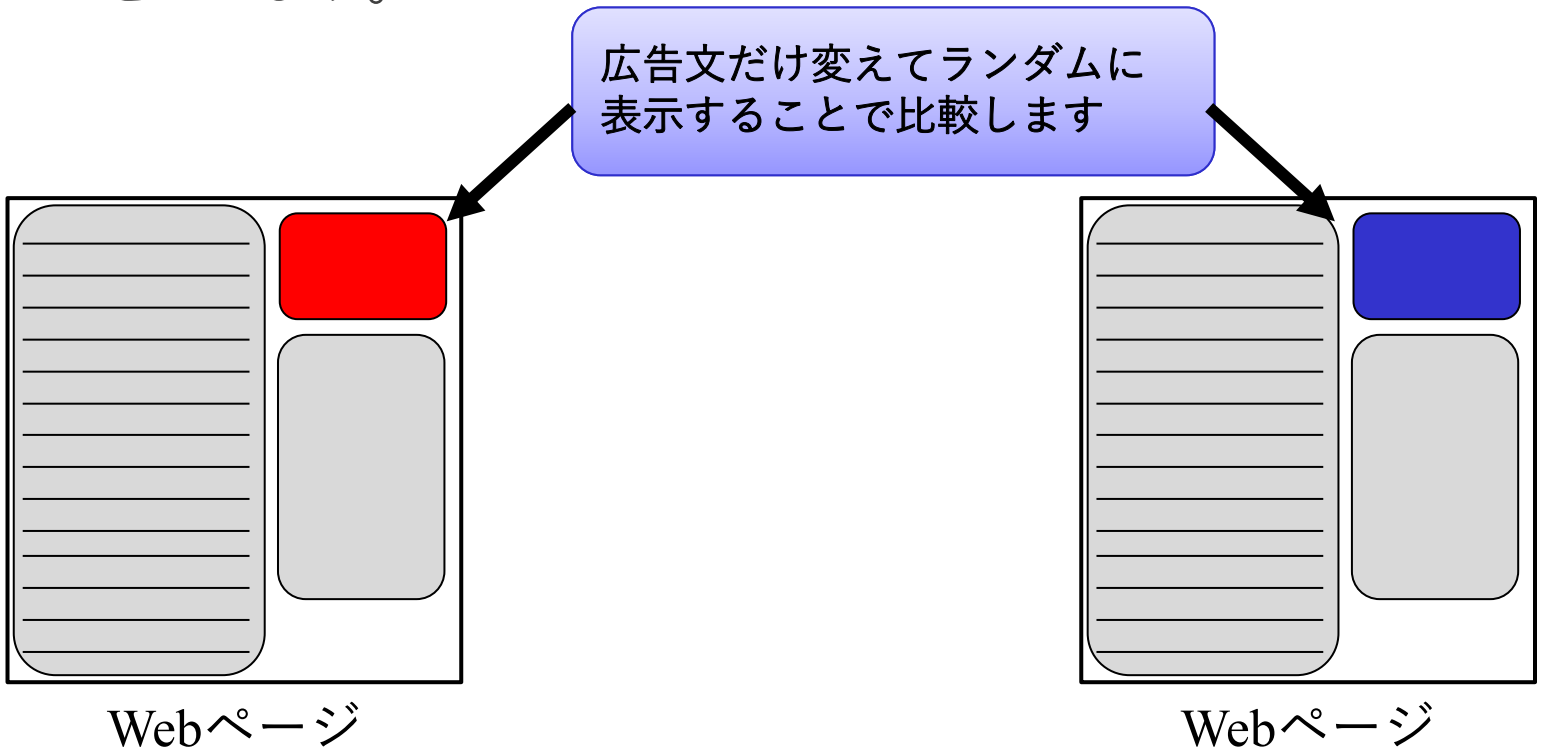
データの比較：処理の前後での比較

- 総コレステロールを下げる飲料の効果の検証
  - 処理の前後のデータを比較することで処理の効果を測ることができます。



データの比較： A/Bテスト

- A/Bテストとは、Web広告等で広告文を変えてランダムに広告を表示することで、よりクリックにつながるような広告文を検証する手法のことをいいます。

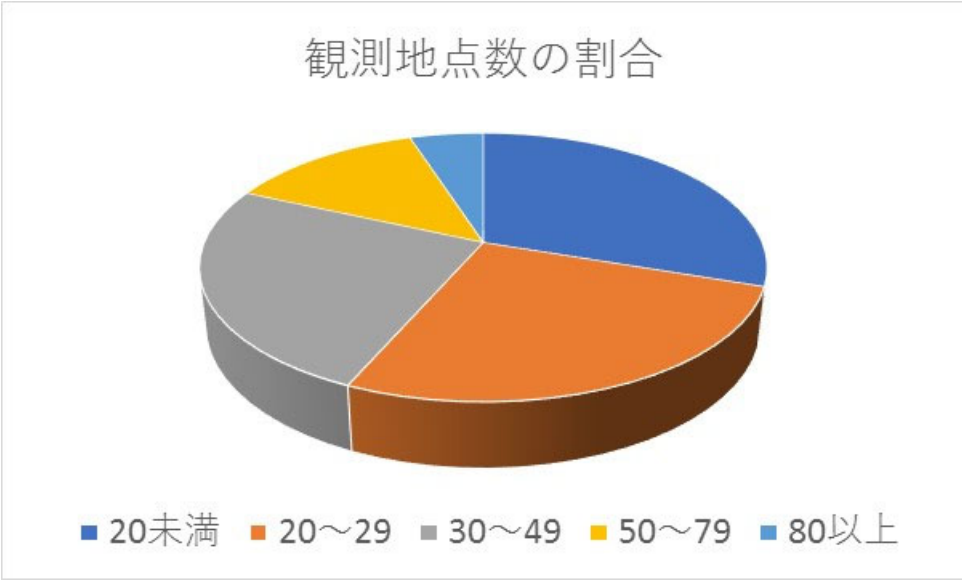


※A／Bテストは、ランダムに比較項目を変更することで他の要因による影響を除去する「ランダム化比較試験」を簡易に実現しています。

不適切なグラフ表現

- グラフに対して必要以上に視覚的な加工を行うことはデータを誤って読み取ってしまう原因になるので注意が必要です。
- 例えば右下の円グラフは、上述の「観測地点数の割合」の円グラフを立体化したもので、観測地点数が「20未満」の割合は「20～29」の割合よりも多いのですが、右下のグラフをみると「20～29」の割合の方が多い印象を受けます。

- このように円グラフを立体化すると手前が大きく見えてしまうので不適切なグラフ表現になってしまいます。



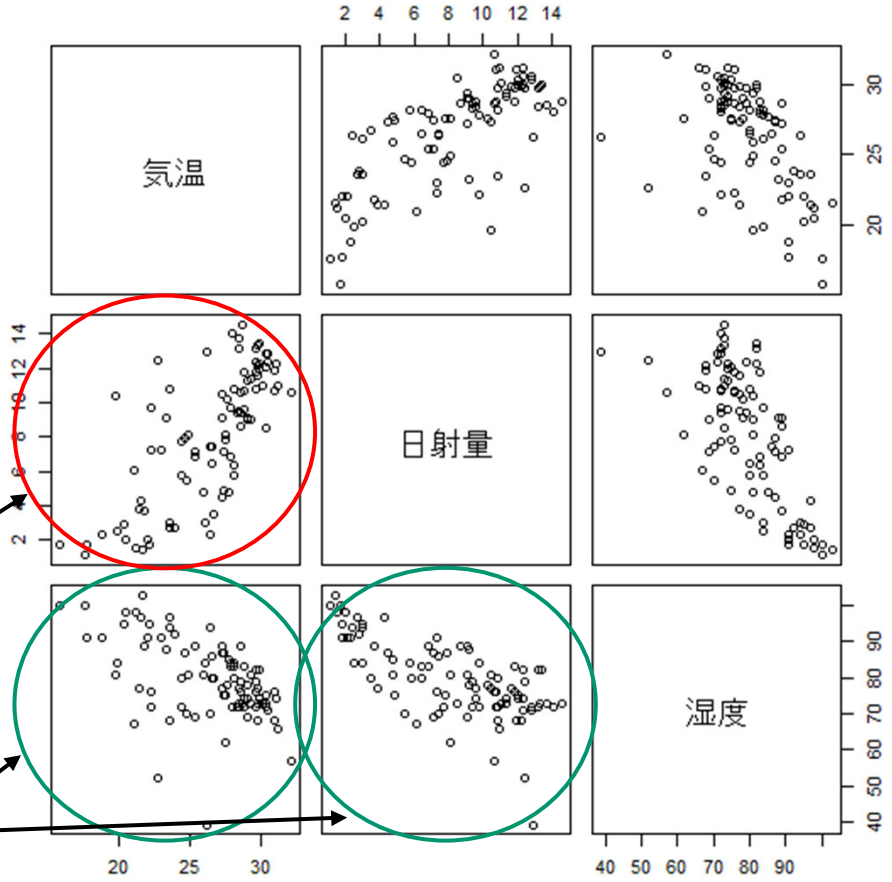
可視化による気づき

- データを可視化することによってデータの新たな傾向などに気づくことも多いです。例えば下記の気候のデータに対して散布図行列を計算することで各項目にどのような関係があるかがわかります。

＜７～9月の気候データ＞

月	日	気温	日射量	湿度
7	1	28	14.08	72
7	2	28.7	14.54	73
7	3	28.4	13.18	72
7	4	27.5	7.82	75
7	5	26.7	3.49	80
7	6	21.6	1.42	103
7	7	24.9	8.11	81
・	・	・	・	・
・	・	・	・	・

散布図による可視化で各項目の関係性が捉えられます。



右肩上がり＝同じ方向に動く

右肩下がり＝逆に動く