

データサイエンス 模擬授業

五十嵐康彦(筑波大学)

全体の概要

- 授業形態(150分/週 x 10週)
 - 講義、演習、ビデオ講義
 - 小テスト、レポート

講義内容・キーワード

週	講義内容	キーワード
1	データサイエンス概論	データサイエンスの概要 データサイエンス導入ビデオ講義
2	データの種類と収集1	データの種類、研究倫理、データの収集、エクセル基本操作
3	データの種類と収集2	データサイエンスの基本プロセス、データの前処理 データの再利用性、エクセル応用操作(統計)
4	データの種類と収集3	データ管理の意義と目的、データ収集項目の設定 情報構造と表現の分離、エクセル応用操作(データ操作)
5	高度なデータ管理 達成度評価(中間)	ビッグデータビデオ講義
6	データの可視化と質的データの解析	可視化の意義と目的、視覚的表現の選び方、データ可視化演習 1つの質的データ(度数分布表)、2つの質的データ(クロス集計表)
7	1つの量的データの解析	1つの量的データの理解、ヒストグラム、記述統計、平均値の違い 統計検定
8	2つの量的データや時系列データの解析	2つの量的データの理解, 相関と相関係数, 時系列データ, (回帰直線)
9	実際のデータの分析	様々なデータの分析、ネットワークデータ、相関と因果
10	高度なデータ分析 達成度評価(期末)	人工知能ビデオ講義

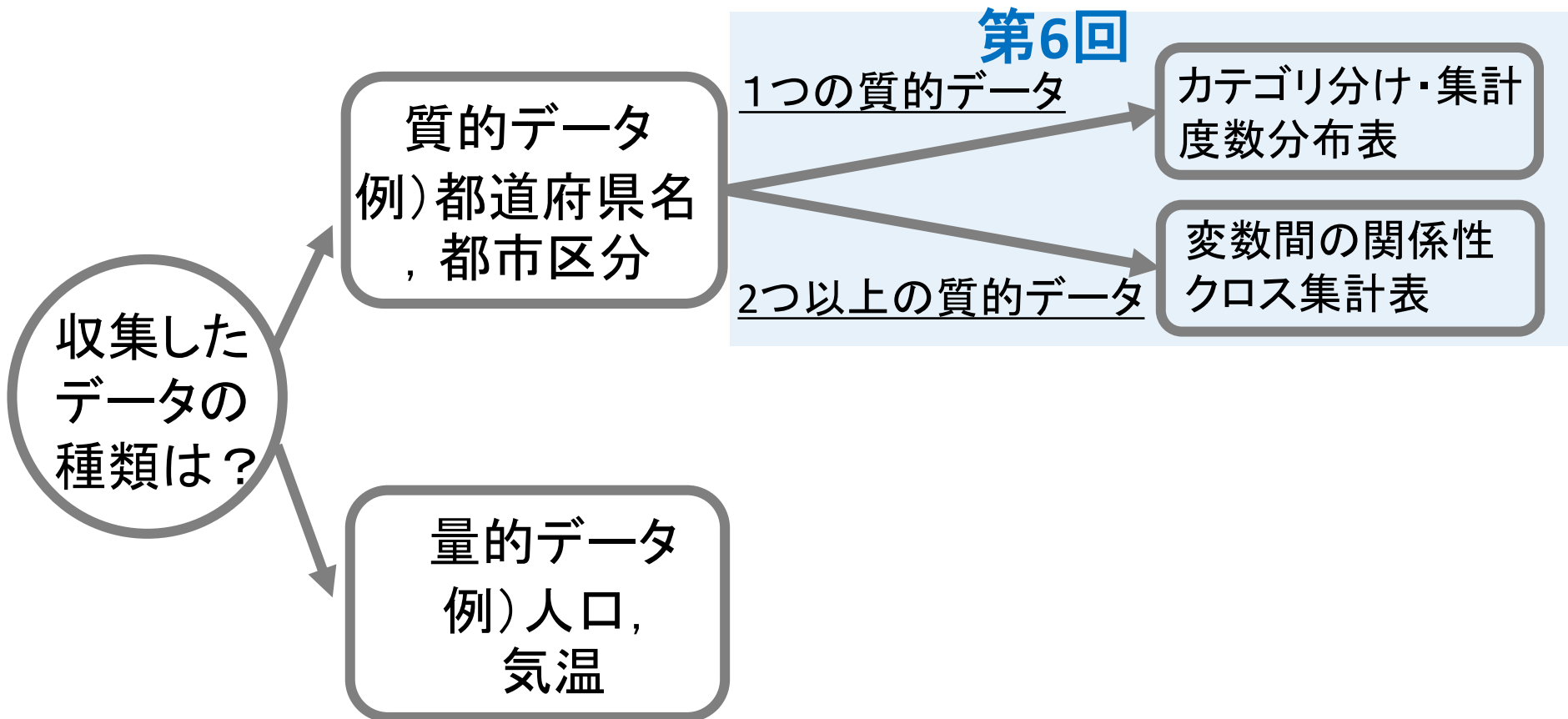
第6回 データの可視化と 質的データの解析 (講義)

共通科目「情報」データサイエンス



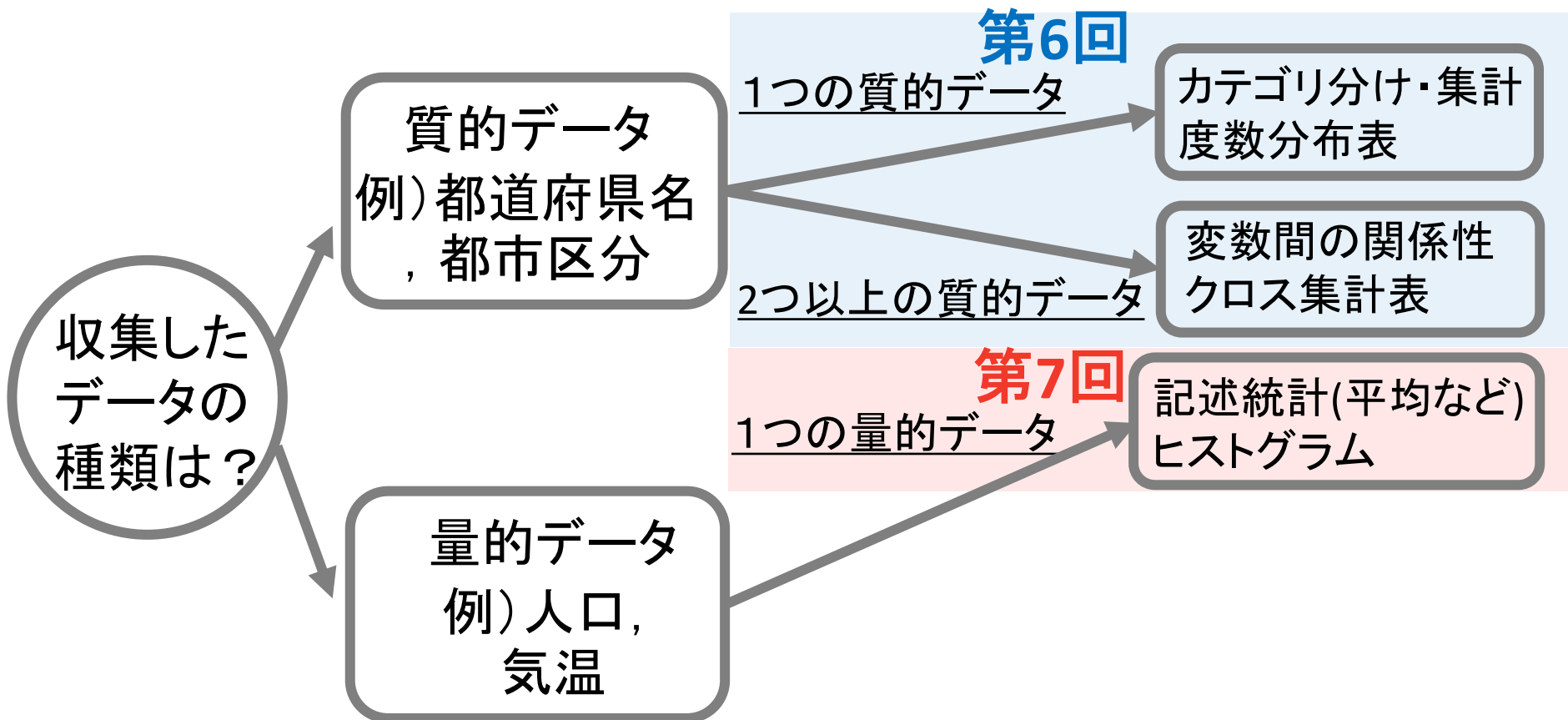
第6回から第8回： 収集したデータをどう扱うか？

収集したデータの種類、数によって
適切な取り扱い方や可視化が変わる。



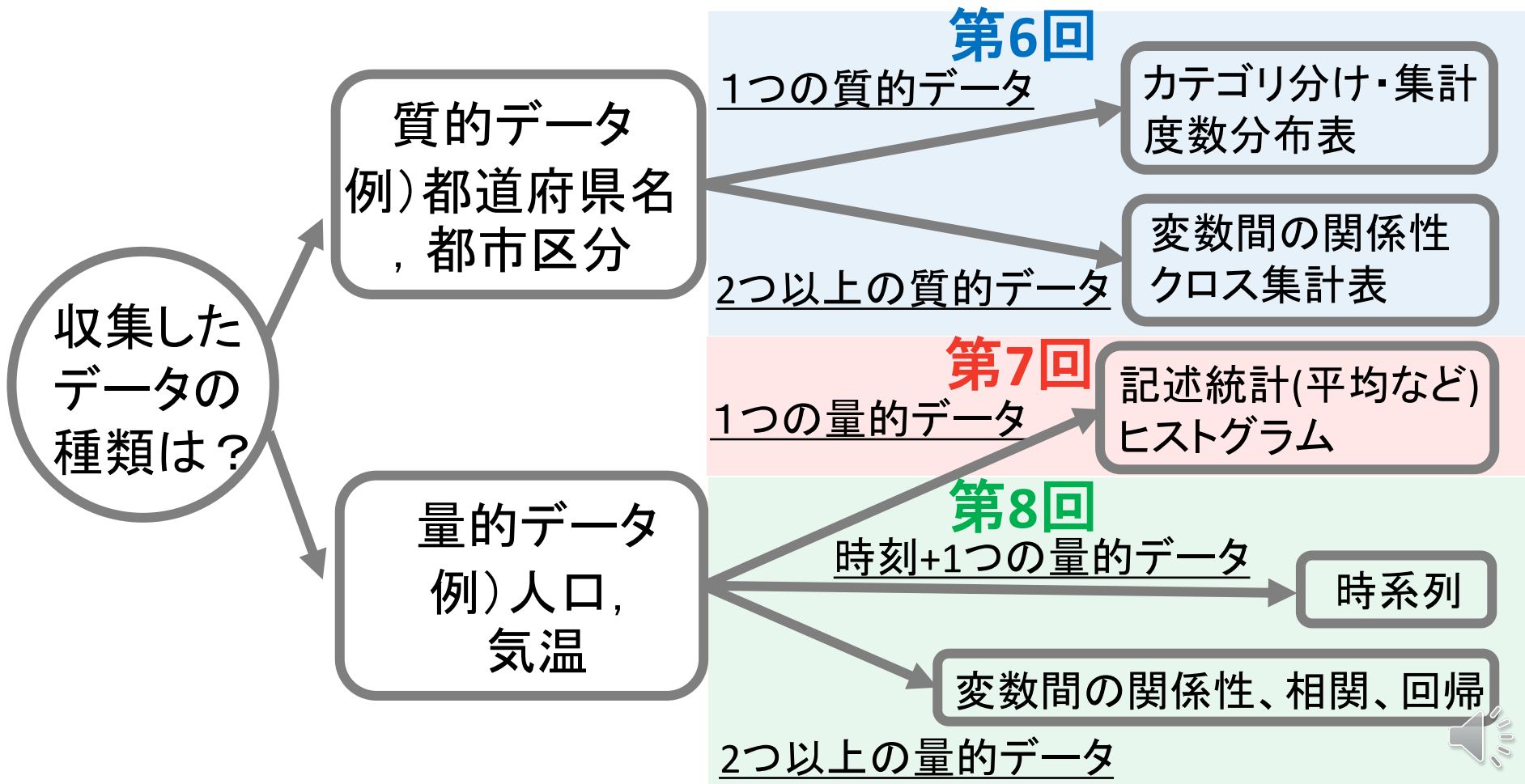
第6回から第8回： 収集したデータをどう扱うか？

収集したデータの種類、数によって
適切な取り扱い方や可視化が変わる。



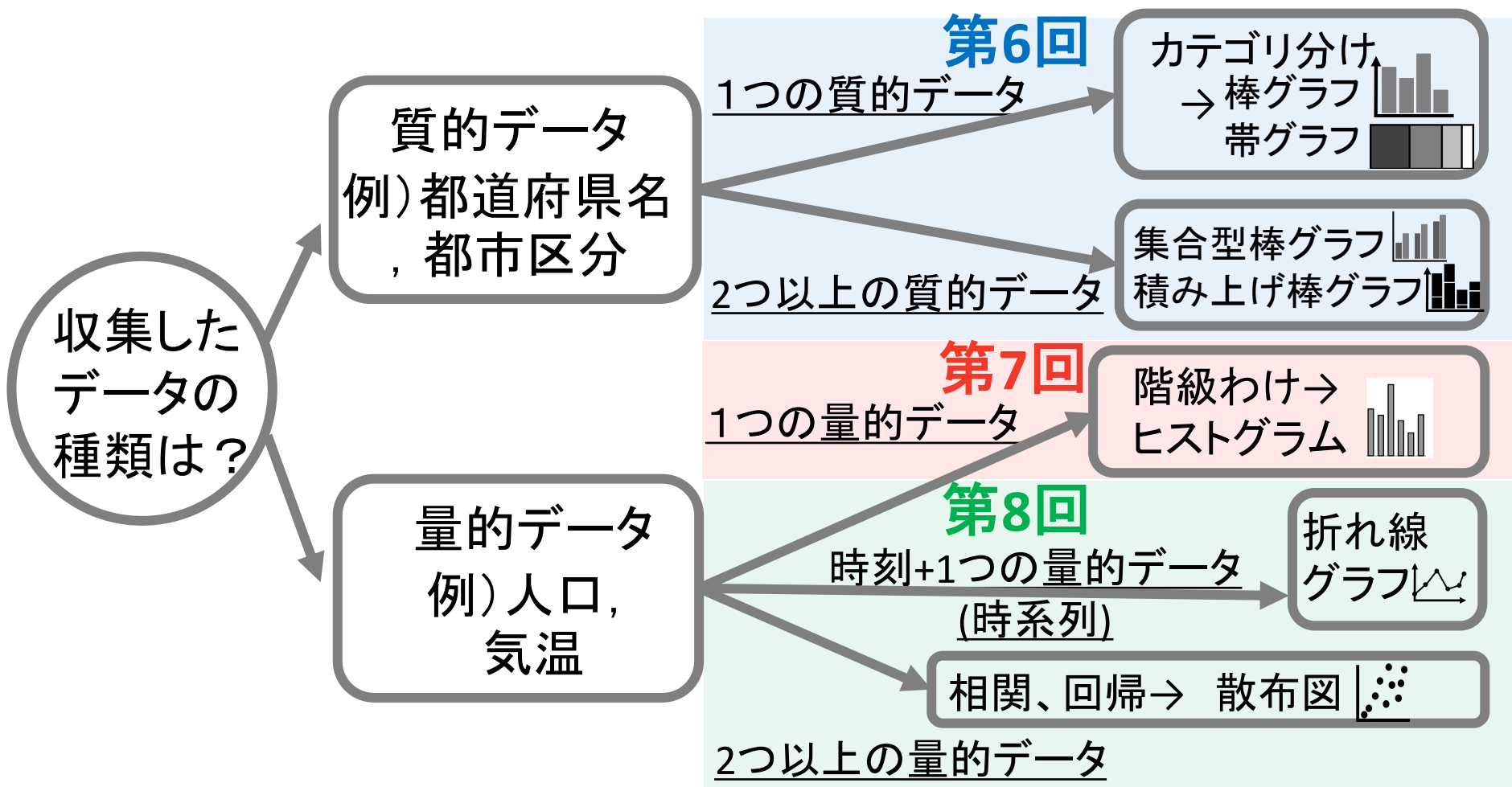
第6回から第8回： 収集したデータをどう扱うか？

収集したデータの種類、数によって
適切な取り扱い方や可視化が変わる。



収集したデータをどう可視化するか？

収集したデータの種類、数によって
適切な可視化も変わる。



今回の目標

データの可視化

- 可視化の意義と可視化処理について知る
- 視覚属性と視覚的表現(グラフ)の選び方を理解する
- 適切／不適切なグラフについて理解する
- Excelで基本的なグラフを描けるようになる

質的データの解析

- カテゴリ分け、度数分布表を理解する
- 質的データの統計量について理解する
- クロス集計表とオッズ比について理解し、活用できるようになる
- 集合型棒グラフについて理解する

人間の視覚能力を利用して、膨大なデータの活用を促進する

可視化の意義と可視化処理のモデル

アンスコムの4つ組

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

データの特徴を捉えるためには？

I

x	y
10.0	8.04
8.0	6.95
13.0	7.58
9.0	8.81
11.0	8.33
14.0	9.96
6.0	7.24
4.0	4.26
12.0	10.84
7.0	4.82
5.0	5.68

一般に、データの特徴を捉えるために、
データを統計量(平均など)にまとめる

詳しくは第7、8回参照

左のデータ
の統計量

xの平均: 9.0
xの分散: 11.0
yの平均: 7.5
yの分散: 4.1
相関係数: 0.82

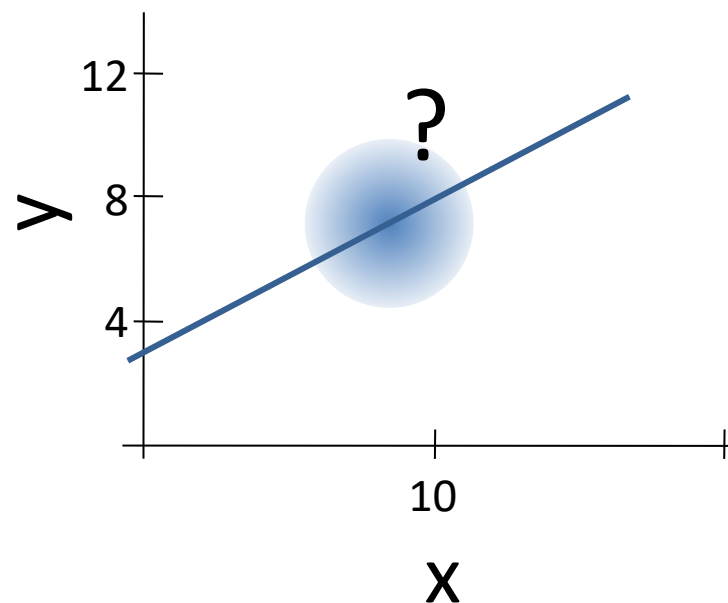
ただし、統計量(平均など)でデータをまとめる
ことで、データの特徴を見落とす場合もある

4組のデータの特徴は？

- 2つの量的データがそれぞれ11個

- x の平均: 9.0
- x の分散: 11.0
- y の平均: 7.5
- y の分散: 4.1
- 相関係数: 0.82

- (2変数間の関係の強さ; 第8回資料参照)



- 平均、分散、相関係数などはどの組でも同じ
- 4組はみな同じなのか？ 何が違うのか？
- 可視化すると、その違いが明らか (Quiz 06-01, 02)

4組のデータの特徴は？

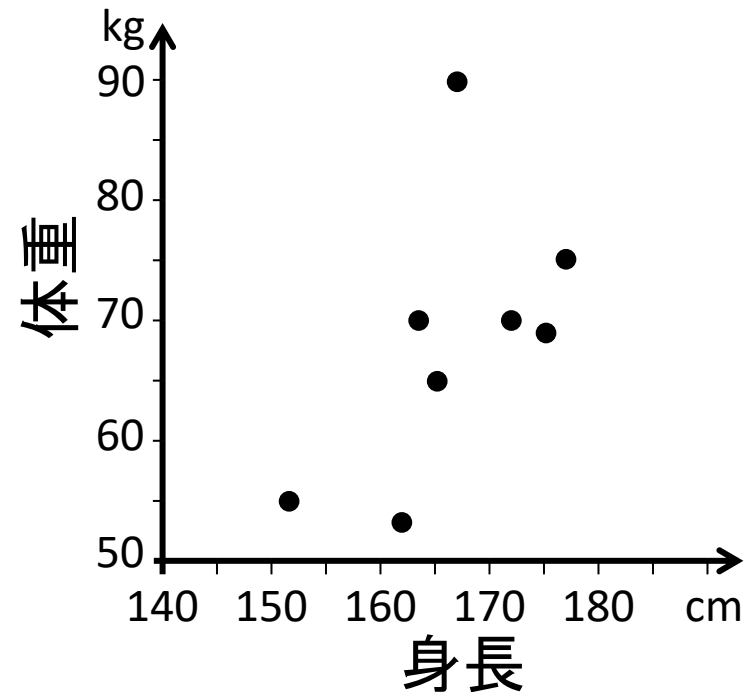
散布図(Scatter plot)での可視化

2つの量的データのペア(具体的には、身長と体重)を散布図で可視化

散布図

対象データ 2つの量的データ

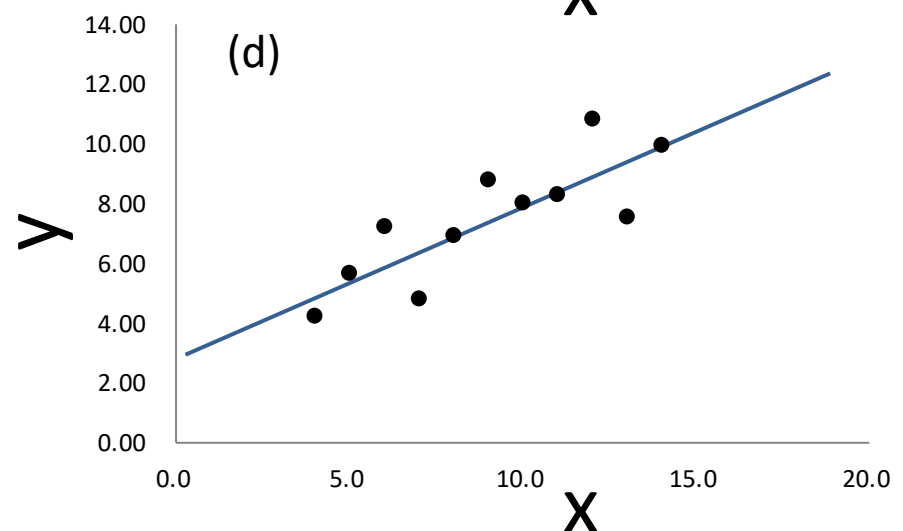
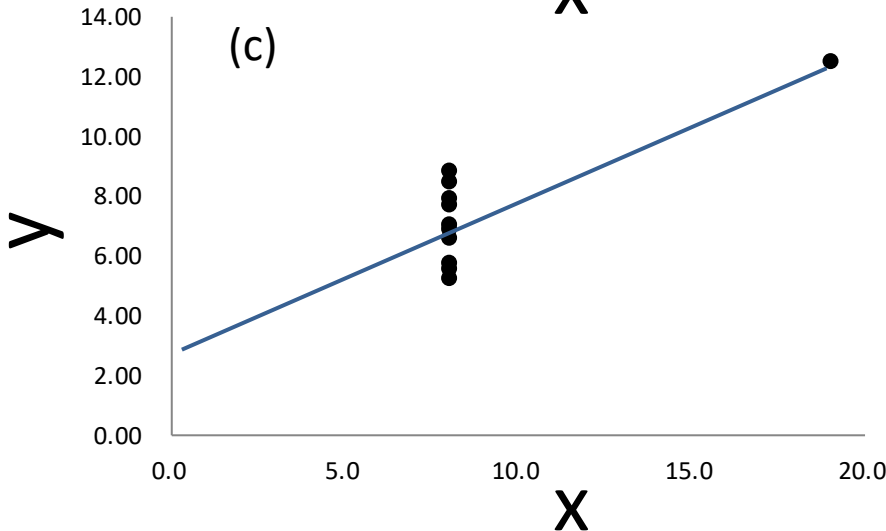
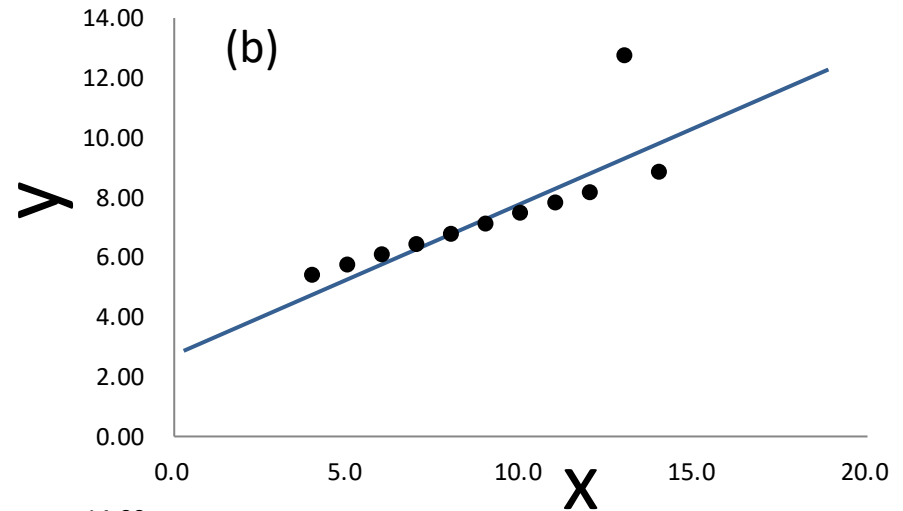
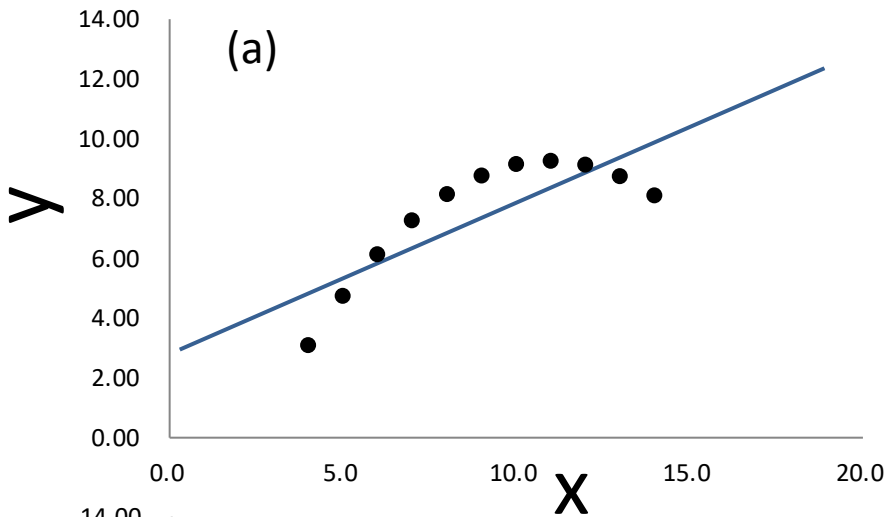
- レコード ⇒ 点
 - 1番目の量的データ ⇒ 横方向の位置
 - 2番目の量的データ ⇒ 縦方向の位置
-
- 直交座標系、座標軸を描く



アンスコムズのデータを散布図で可視化するとどういった特徴になるか？

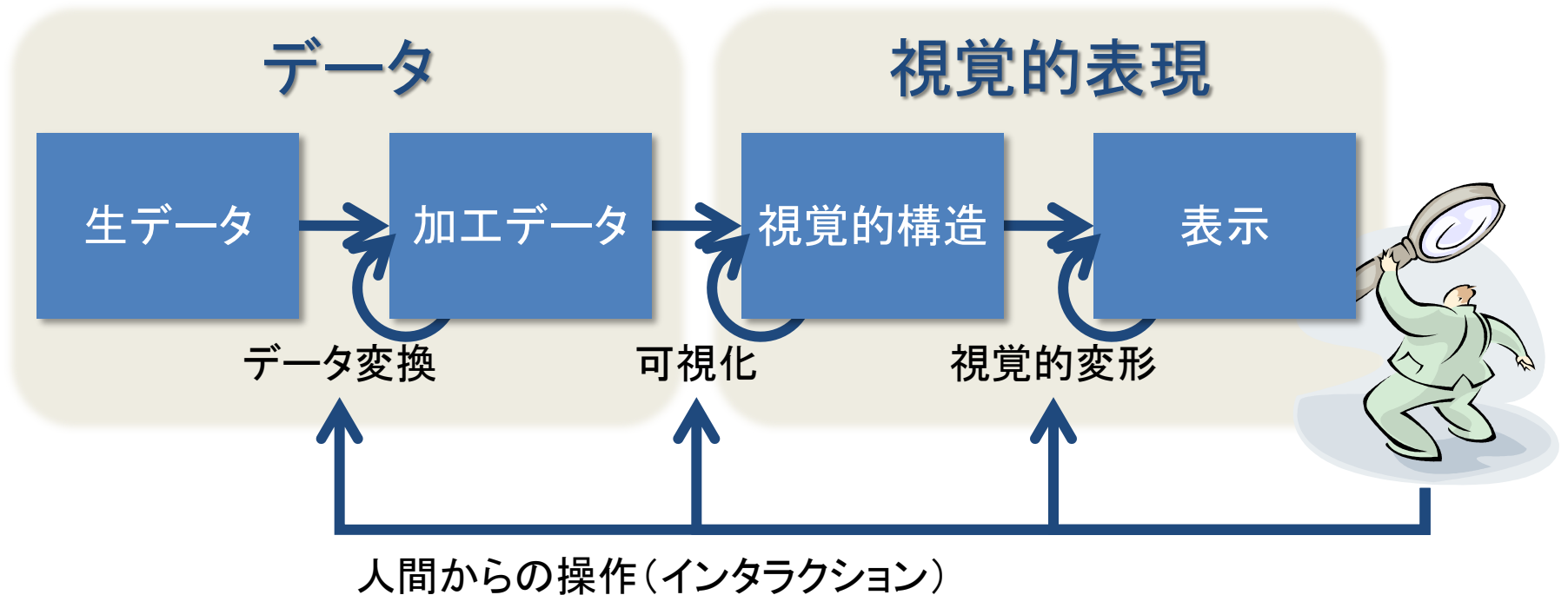
Quiz 06-01と06-02 グラフ(散布図)による可視化

アンスコムズの4つ組(I, II, III, IV)のうち、I(Quiz 6-1)とIII(Quiz 6-2)のデータを散布図で表現した番号をそれぞれ答えなさい。
(manabaには参考のため、データのexcelファイルを添付した)



可視化処理のモデル

- 可視化における処理を一般化したもの



(参考) Figure 1.23 in Stuart K. Card et al., Readings in Information Visualization: Using Vision to Think, Morgan Kaufmann, 1999.

全体の概要

- 授業形態(150分/週 x 10週)
 - 講義、**演習**、ビデオ講義
 - 小テスト、レポート
- 全学必修であるための工夫
 - 講義資料および演習内容については、下記の3種類を用意
 - 簡易版: 数理に馴染みのない文系クラスを想定
 - 標準版: 数理に馴染みのある文系クラスを想定
 - 発展版: 理系クラスを想定
 - 自分たちのデータを自分で解析

演習

- 概要

- エクセルを使用して、各講義内容の演習に取り組む

- エクセルの基本操作、データの前処理
- 可視化、データの入手、データの分析
- など

E03-01: 総合演習A

- シート「E03-01」は各都道府県の人口および面積のデータである

	A	B	C	D
1	都道府県	人口 (単位: 人)	面積 (単位: km ²)	人口密度 (単位: 人/km ²)
2	東京	13,843,403,000	2193.96	
3	神奈川	9,179,835	2416.16	
4	大阪	8,874,566	1905.29	
5	愛知	7,539,185	5172.96	
6	埼玉		3797.75	
7	青森	1,308,265	9645.64	
8	神奈川	9,179,835	2416.16	
9	広島	2,819,962	8479.61	
0	茨城	2,882,943	6097.33	
1	千葉	6,268,585	5157.61	
2	京都	2,591,779	4612.20	
3	熊本	1,756,442	7,409.50km ²	

- 各都道府県の人口密度を集計したい
- 次の手順でデータを処理しなさい
欠損値や重複を含む行を削除しなさい
 1. D列で都道府県ごとの「人口密度」(=「人口」/「面積」)を計算しなさい
 2. の結果、正しく計算できない都道府県が存在する、正しく計算できるように表記揺れを修正しなさい
 3. 日本の全人口は約1.2億人ということをふまえると、記入ミスと思われる大きく外れた値が一つある、該当する値が含まれる行を削除しなさい
 4. D列の書式を変更し小数点以下の桁数表示を2桁に揃えなさい

29

データ前処理

演習課題 E08-03-01

- 課題1
 - 茨城県のデータについて2001年の人口を基準とした各年の人口の伸び率(指数)を求めよ
- 課題2
 - 求めた伸び率を折れ線グラフにプロットせよ
- 課題3
 - 描画した移動平均の折れ線グラフについて考察せよ(穴埋め問題)

64

データの分析

第6回 データの可視化と 質的データの解析 (演習)

共通科目「情報」データサイエンス



(軽量版)

第6回データの可視化と 質的データの解析の演習

以下の課題について、サンプルデータとアンケート集計のデータを用いて、それぞれ表やグラフで可視化および質的データの解析の演習を行う

E06-01:棒グラフ、帯グラフの作成

E06-02:度数分布表の作成

E06-03:アンケート結果の度数分布表と
そのグラフ化(帯グラフ)

E06-05:クロス集計表および、オッズ比の作成

演習 : E06-01

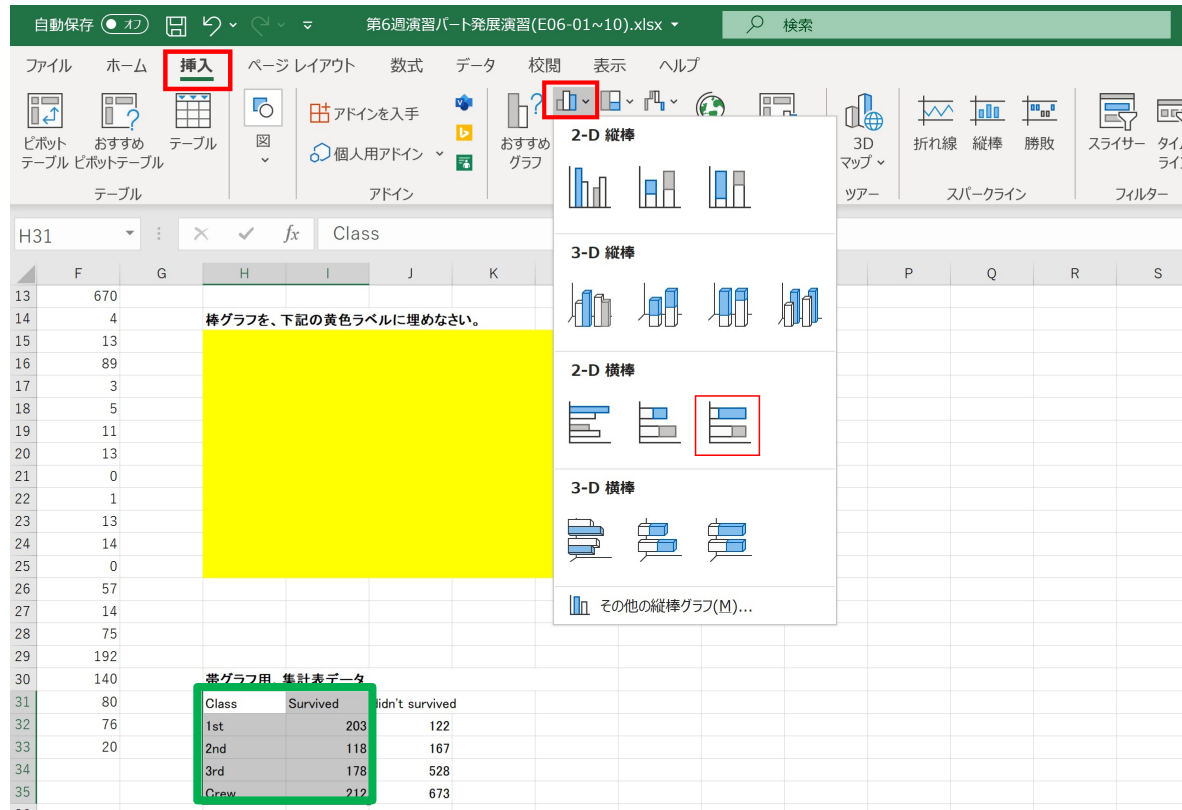
棒グラフと帯グラフ の作成

- タイタニックのデータを用いて、(1) タイタニック号の乗員・乗客数(クラス別)についての棒グラフを作成して下さい。また、(2) タイタニック号の生存者数(クラス別)の帯グラフを作成して下さい。
- グラフのデータは、E06_演習.xlsxの、E06-01シートにある。同じシートにある、黄色く塗りつぶされた範囲にグラフや集計表を回答して下さい。
- 下記のスライドから、作成方法について述べる。



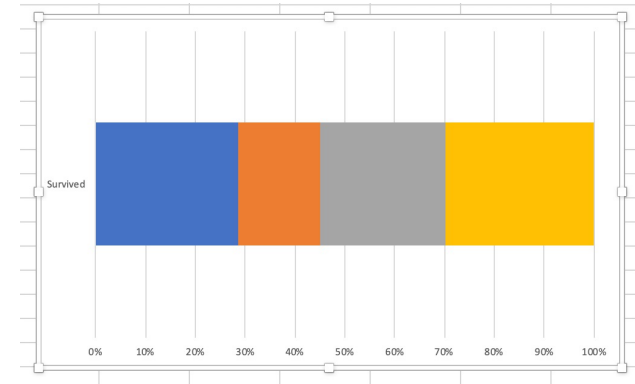
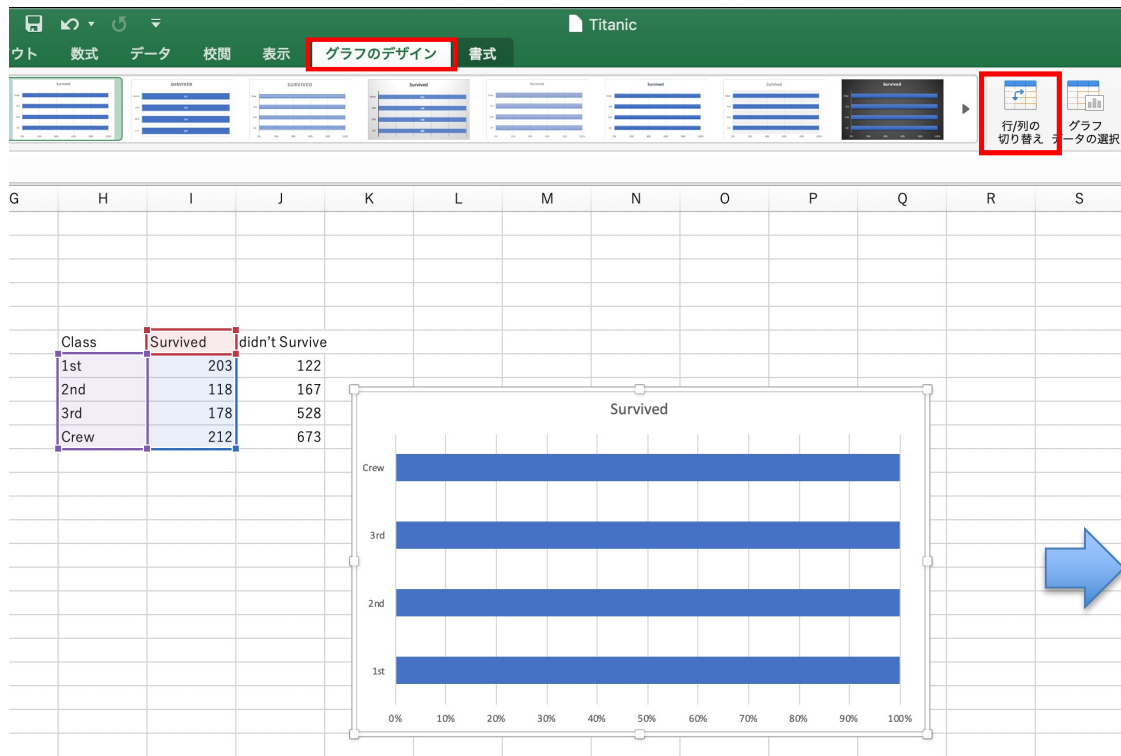
帯グラフを描く

データの選択と図の挿入



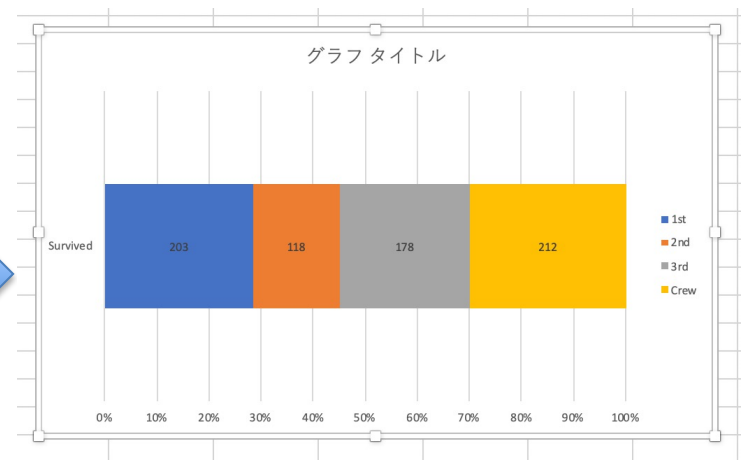
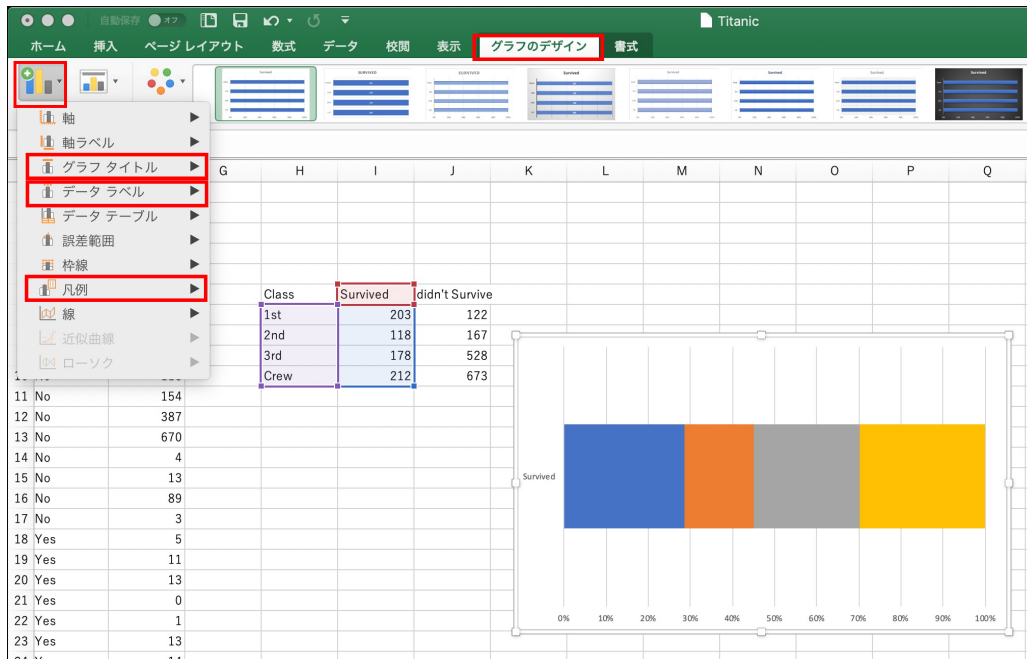
1. 作成した集計表のうち、緑で囲った範囲を選択する
2. 「挿入」タブのグラフ一覧から図のグラフを選択する
→ 「100%積み上げ横棒」と表記が出て、グラフがプロットされる

表示形式の変更



- グラフが一つの帯になるよう、データの行/列を変更する
 1. グラフを選択する
 2. 「グラフのデザイン」タブの「行/列の切り替え」を選択する

各要素の追加



- グラフに「グラフタイトル」、「データラベル」、「凡例」を追加
 1. グラフを選択する
 2. 「グラフのデザイン」タブの「グラフ要素を追加」から「グラフタイトル」、「データラベル」、「凡例」を選択する

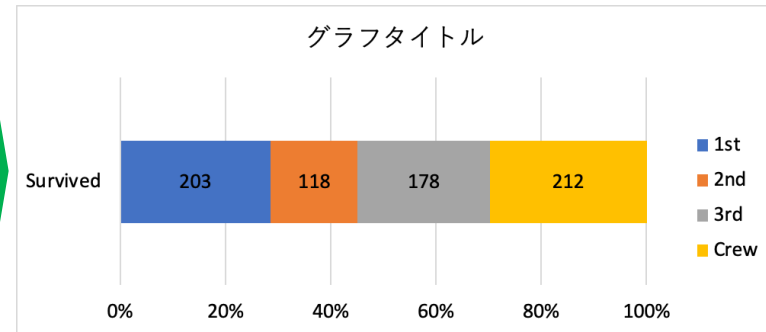
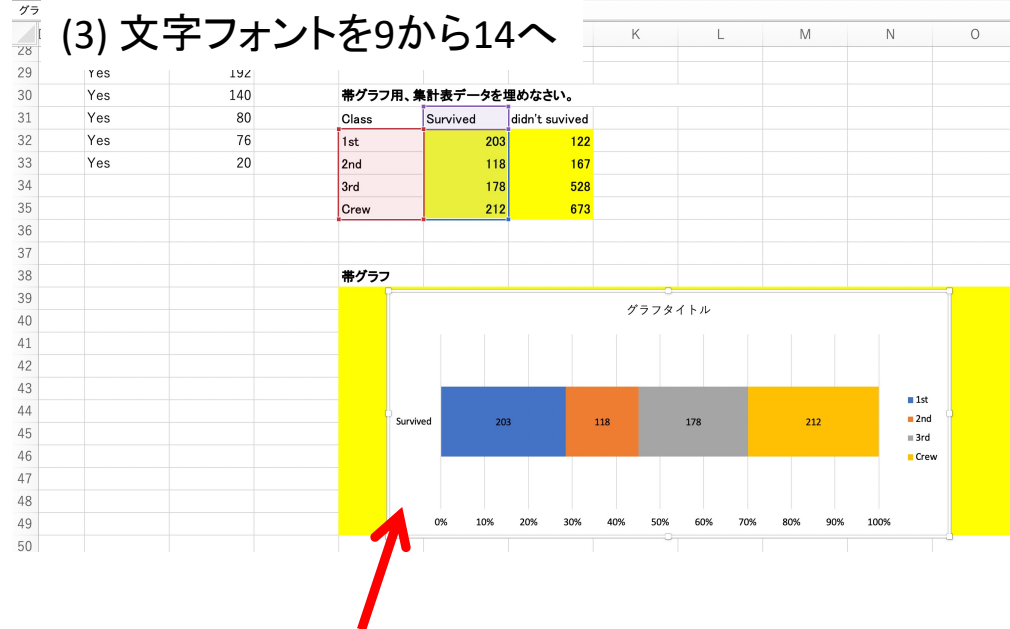
データラベル、グラフタイトル、凡例の文字フォントを変更する

- 次の手順で、データラベルのフォントサイズを、より識別しやすいように変更

(2) “ホーム”タブを選択



(3) 文字フォントを9から14へ



(1) “グラフエリア”全体を選択

データラベルの色変更

ホーム

(2) “ホーム”タブを選択

(3) 文字の色を黒から白へ

グラフ

帯グラフ用、集計表データを埋めなさい。

Class	Survived	didn't survived
1st	203	122
2nd	118	167
3rd	178	528
Crew	212	673

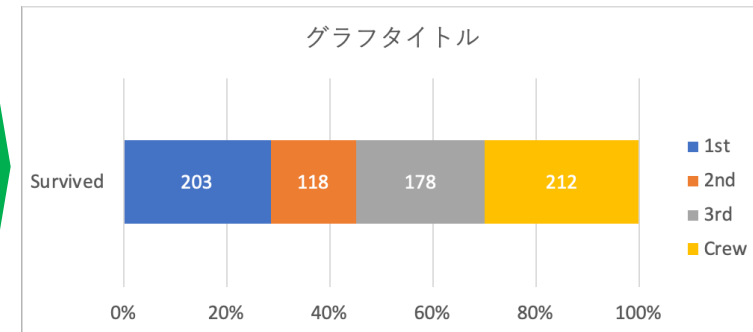
帯グラフ

グラフタイトル

Survived

0% 20% 40% 60% 80% 100%

1st
2nd
3rd
Crew

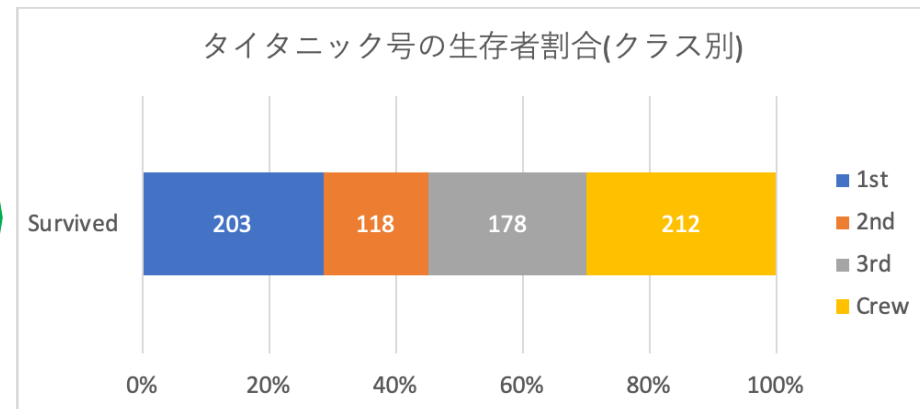
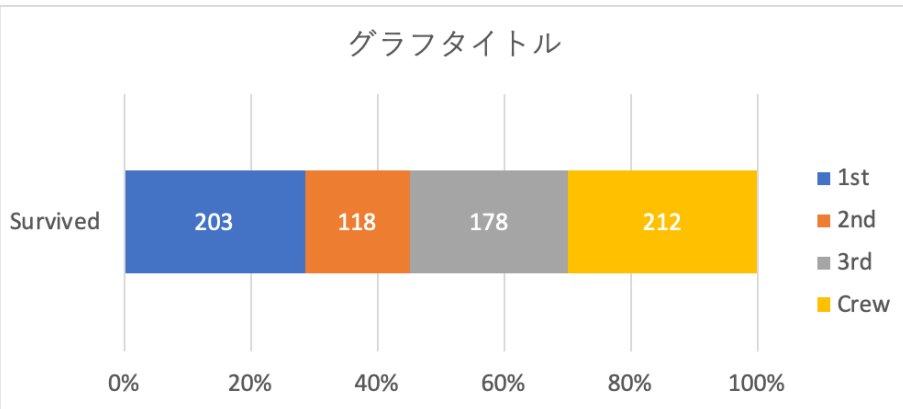


(1)色を変更したいラベルを選択

- データラベルの色やフォントサイズを、より識別しやすいように変更
 - (1) 色を変更したいラベルを選択
 - (2) “ホーム”タブを選択
 - (3) 文字の色を見やすい色(例: 白)に変更 (他に、「書式」タブの「文字の塗りつぶし」からでも変更可能)

タイトルの変更

「タイタニック号の生存者割合(クラス別)」



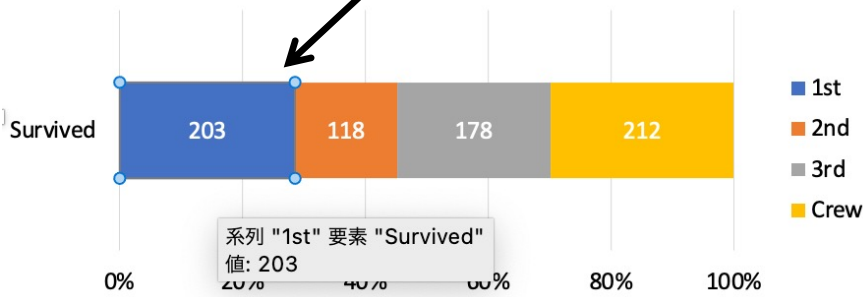
- タイトルを適切なものへ変更
 1. タイトルを選択し、テキストの編集が可能な状態にする
 2. タイトルを「タイタニック号の生存者割合(クラス別)」に変更する

グラフの色の変更

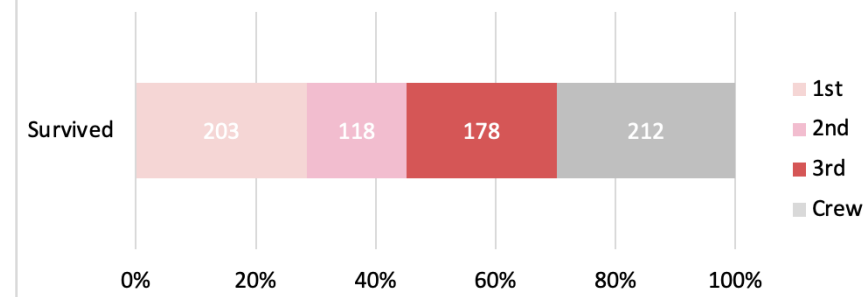
クラス別に順序尺度となっているため、
色分けをして、結果を視覚的にわかりやすくする

まず、変えたいエリアをクリックして選択

タイタニック号の生存者割合(クラス別)

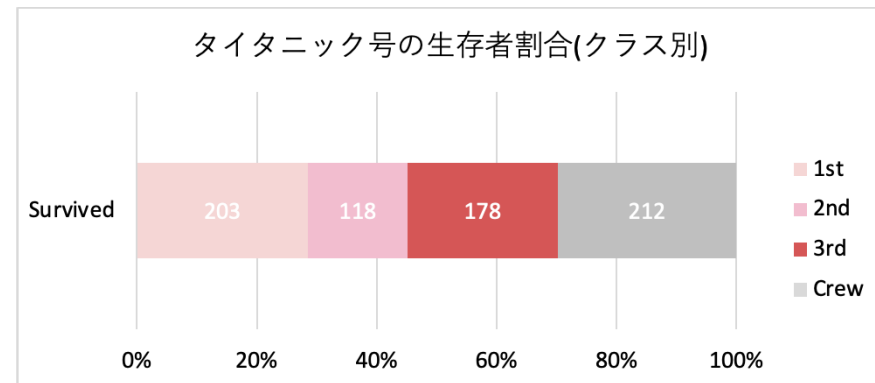
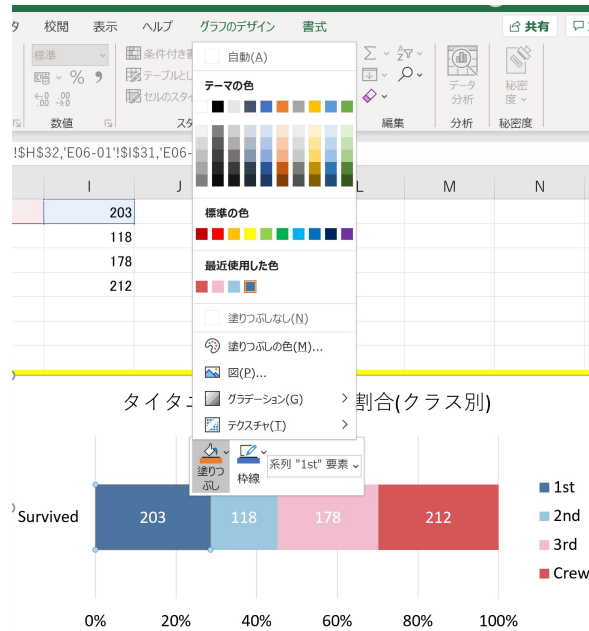


タイタニック号の生存者割合(クラス別)



グラフの色の変更

クラス別に順序尺度となっているため、
色分けをして、結果を視覚的にわかりやすくする



右クリックし、“塗りつぶし”の色を変更
⇒各クラスの色を同様に変更する

* 資料と同一の色設定にしなくてもよい

アンケート結果の度数分布表作成と そのグラフ化(帯グラフ)

事前に行うアンケート(一部)

DS 全学アンケート項目案
2020.6.13.

<個人属性に関する質問>

1. あなたの所属学群を以下より選択してください

- (ア)人文・文化学群
- (イ)社会・国際学群
- (ウ)人間学群
- (エ)生命環境学群
- (オ)理工学群
- (カ)情報学群
- (キ)医学群
- (ク)体育専門学群
- (ケ)芸術専門学群
- (コ)総合学域群

2. 2015 年に国際連合にて 193 ヶ国の首脳により署名された持続可能な開発目標 (SDG)、またはグローバル目標はご存知ですか。

- (ア)はい
- (イ)いいえ

3. SDG の 17 個のグローバル目標のうち、どのグローバル目標が、あなたとあなたの家族にとって目前に迫っている課題ですか？（6つ選択してください。）



(ア) 貧困をなくそう

- ・あらゆる場所で、あらゆる形態の貧困に終止符を打つ



(イ) 飢餓をゼロに

- ・飢餓に終止符を打ち、食料の安定確保と栄養状態の改善を達成するとともに、持続可能な農業を推進する



(ウ) すべての人に健康と福祉を

- ・あらゆる年齢のすべての人の健康的な生活を確保し、福祉を推進する



(エ) 質の高い教育をみんなに

- ・すべての人に包摂的かつ公平で質の高い教育を提供し、生涯学習の機会を促進する



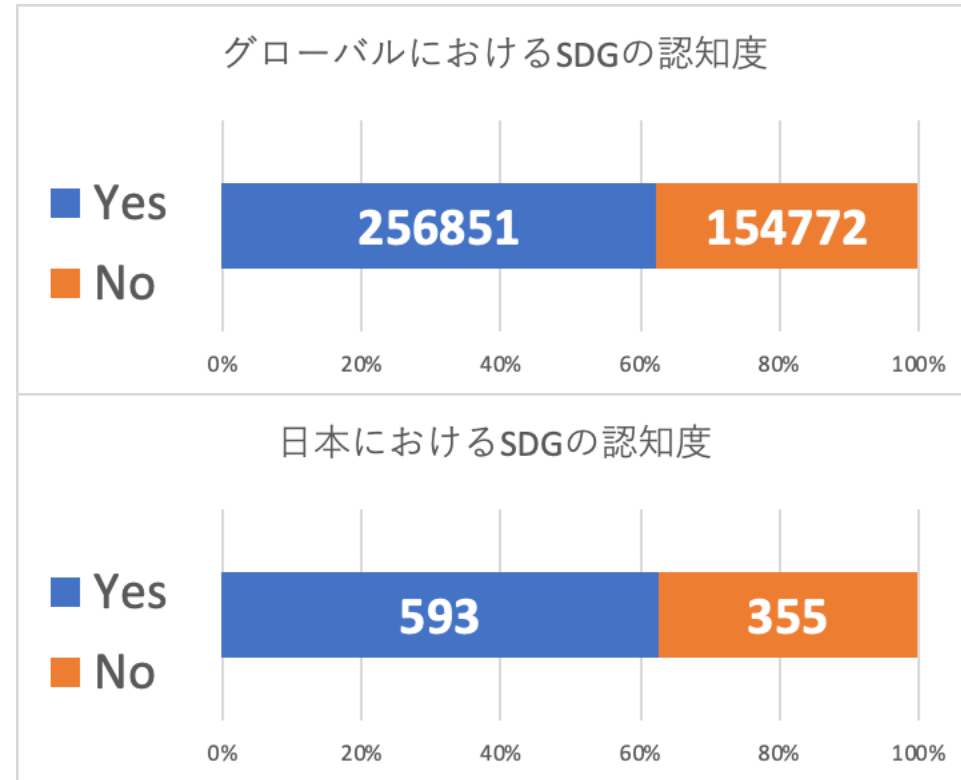
(オ) ジェンダー平等を實現しよう

- ・ジェンダーの平等を達成し、すべての女性と女児のエンパワーメントを図る

事前に行ったアンケートをもとに、
度数分布表と帯グラフを作成する

演習：E06-03: SDGsの認知度のアンケート結果について、次の問いに答えなさい

- 下図に示す、SDGsの認知度についての令和元年のアンケート結果から、日本とグローバルにおけるSDGsの認知度について比較しなさい。
- 事前に収集され、配布されたアンケートデータ(Q2)を用いて、SDGsの認知度についての度数分布表および帯グラフを作成しなさい。
- (b)で作成したグラフを用いて、筑波大学における大学1年生と、令和元年における全世界のSDGsの認知度についての比較しなさい。



(令和元年(2019年) 6月19日現在)

MY WOLRD 2030より

https://www.unic.or.jp/activities/economic_social_development/sustainable_development/2030agenda/myworld2030/

度数分布表とグラフの作成 における注意点

* 調査対象単位での無回答の場合には、集計から完全に除外し、有効票だけを集計せよ。