

データリテラシー

第9回 分割表とクロス集計

メディア情報コース
平居 悠（ひらい ゆたか）

到達目標

- データサイエンス・AIが社会でどのように活用されているのかを理解する。
- データを適切に読み解き理解する力をつける。
- データを適切な可視化手法で他者に説明できるようになる。
- 表計算ソフトを使ってデータを集計・加工できるようになる。

前回

第 1 回	社会におけるデータ・AI利活用
第 2 回	表計算ソフトの基本操作
第 3 回	セルの参照と集計
第 4 回	度数分布表とヒストグラム
第 5 回	代表値と散らばりの尺度
第 6 回	データのグラフによる表現とデータ比較
第 7 回	データの分布と標本抽出
第 8 回	推定
第 9 回	分割表とクロス集計
第 1 0 回	2 変量データと相関
第 1 1 回	時系列データ
第 1 2 回	仮説検定
第 1 3 回	回帰分析

前回の目標

表計算ソフトを用いて母平均を推定できるように
なる。

前回学んだこと

1. 母数と統計量
2. 母平均の推定

母数（パラメーター）

母集団分布の特徴を表す
代表値や散らばりの尺度

よく使われる母数

- **母平均**（母集団の平均） μ
- **母分散**（母集団の分散） σ^2
- **母標準偏差**（母集団の標準偏差）

$$\sigma = \sqrt{\sigma^2}$$

標本から推定できるもの

- 標本は母集団の一部でしかないので、標本だけで母集団分布の全てを知ることはできない。
- 標本から推定できるのは母集団の**母数**。

標本平均 $\bar{X}$

母平均 μ の推定に使われる統計量。
母集団から n 個の標本 $X_1, X_2, \dots, X_n$
を無作為抽出したとすると、

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n} = \frac{1}{n} \sum_{i=1}^n X_i$$

不偏分散 s^2

母分散 σ^2 の推定に使われる統計量。

$$s^2 = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2}{n - 1} = \frac{1}{n - 1} \sum_{i=1}^n (X_i - \bar{X})^2$$

標本平均と不偏分散の期待値

標本平均 $\bar{X}$ の期待値は母平均 μ 、
不偏分散 s^2 の期待値は母分散 σ^2
になる。

$$E(\bar{X}) = \mu$$

$$E(s^2) = \sigma^2$$

点推定

単純に標本平均 $\bar{X}$ の値を母平均 μ の推定値とすること。

点推定は**大数の法則**を根拠としている。

大数の法則

標本平均 $\bar{X}$ は n が大きくなるにつれ、母平均 μ に近づく。

区間推定

ある確率 $1 - \alpha$ で μ が含まれる
区間を推定すること。

区間推定

具体的には

$$P(L \leq \mu \leq U) = 1 - \alpha$$

- を満たす区間 $[L, U]$ を求めること。
 $[L, U]$ を $1 - \alpha$ の **信頼区間** という。

区間推定

確率 $1 - \alpha$ は信頼係数といい、好きなように設定して良い。信頼係数には、0.9 (90%), 0.95 (95%), 0.99 (99%) がよく使われる。

区間推定

信頼区間は標本から確率的に推定するので、標本の統計量の確率分布（**標本分布**）が必要になる。

区間推定

母集団分布を知らなくても n が大きければ **中心極限定理** によって $\bar{X}$ の標本分布は近似的に正規分布になる。

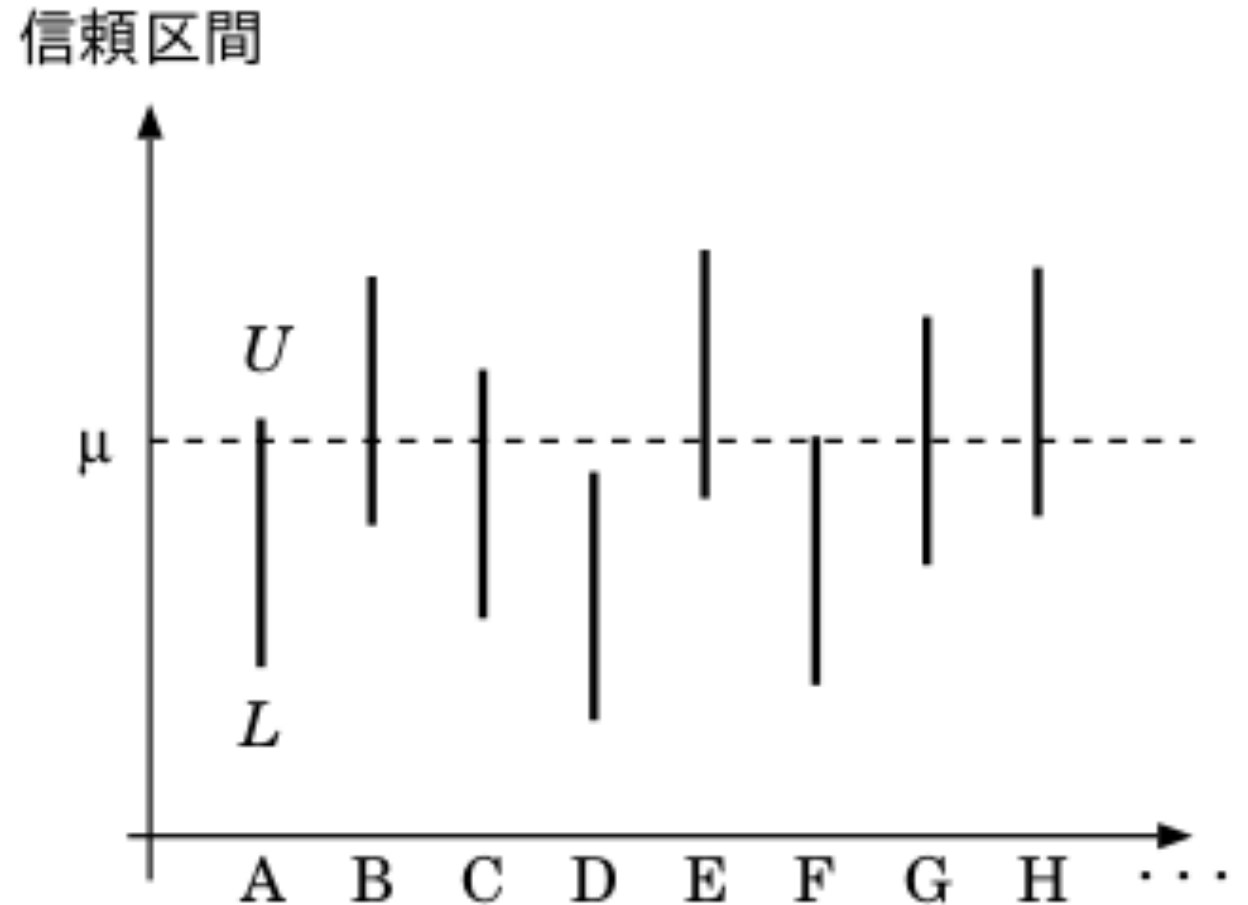
中心極限定理

母平均 μ 、母分散 σ^2 の母集団があるとする。
この母集団がどのような分布でも、 $\bar{X}$ の標本
分布は近似的に $N(\mu, \sigma^2/n)$ になる。

母集団が正規分布の場合は n に関わらず厳密に
いえるが、それ以外の分布の場合は n が大
きいほど近似は良くなる。

信頼係数の意味

信頼係数が大きいほど信頼区間は広く
(右図の縦棒が長くなり)、 μ から外れる可能性は低くなる。



前回の問い

- 母数とは何か？
 - 母集団分布の特徴を表す代表値や散らばりの尺度。
- 母数の推定方法には何があるか？
 - 母数の推定方法には点推定と区間推定がある。

タイピング練習スケジュール

第1回	ホームポジション
第2回	ホームポジション
第3回	ローマ字
第4回	英語初級
第5回	日本国憲法（trr試験）
第6回	ホームポジション
第7回	ローマ字
第8回	英語初級
第9回	日本国憲法（trr試験）
第10回	ホームポジション
第11回	ローマ字
第12回	英語初級
第13回	日本国憲法（trr試験）

タイピング試験 (日本国憲法) 30分間

1. ブラウザを起動し、<https://www.koeki-prj.org/trr/>に繋ぐ。
2. 学籍番号（Cは大文字、省略なし8桁）を入力する。
3. Koeki MAILに届いたパスコードをPasscode: 欄に入力する。

ホームポジション

左手でタイプするキー

右手でタイプするキー



左手の人差指から小指までの
ホームポジション

両手の親指の
ホームポジション

右手の人差指から小指までの
ホームポジション

今回

第1回	社会におけるデータ・AI利活用
第2回	表計算ソフトの基本操作
第3回	セルの参照と集計
第4回	度数分布表とヒストグラム
第5回	代表値と散らばりの尺度
第6回	データのグラフによる表現とデータ比較
第7回	データの分布と標本抽出
第8回	推定
第9回	分割表とクロス集計
第10回	2変量データと相関
第11回	時系列データ
第12回	仮説検定
第13回	回帰分析

今回の目標

多変量データについて分割
表とクロス集計表を作成で
きるようになる

今回学ぶこと

1. 分割表

2. クロス集計表

3. ピボットテーブル

今回の問い

- 分割表とは何か？
- クロス集計表とは何か？
- 表計算ソフトで分割表やクロス集計表を作るには何を使うのか？

今回学ぶこと

1. 分割表

2. クロス集計表

3. ピボットテーブル

多変量データ

ある集団に対して2種類以上の観測を行って得られた多変数のデータ

2変量データ（2次元データ）

2変数のデータ。2変量データを見やすくするには、分割表と散布図の二通りの方法がある。

学年	コース
4	観まち
4	観まち
2	政策
3	情報
3	経営
4	経営
4	政策
2	観まち
4	国際
4	観まち

分割表

2変量データの片方の変数のとりうる値を縦方向（列方向）の項目、もう片方の変数のとりうる値を横方向（行方向）の項目にして、縦と横の項目が交差（クロス）するマスに対応する度数を数えて入れた表

分割表

表側：分割表の一番左の項目のラベルの列

表頭：分割表の一番上の項目のラベルの行

分割表

表側の項目数が s 個、表頭の項目数が t 個の場合、 $s \times t$ 分割表という。

分割表

「合計」の欄は項目数に含めない。列や行の合計は**周辺度数**という。列の周辺度数の合計と行の周辺度数の合計は総数に一致する。

3×6分割表の例

データセット 2 から求めた学年とコース別の学生数の分割表

		表頭						
		経営	政策	福祉	国際	観まち	情報	合計
表側	2年生	40	58	22	22	36	18	196
	3年生	30	42	19	18	22	19	150
	4年生	24	50	23	17	26	14	154
	合計	94	150	64	57	84	51	500

周辺度数

総数

今回の問い

- 分割表とは何か？
 - 2変量データを列と行に入れ、列と行の項目が交差するマスに対応する度数を数えて入れた表
- クロス集計表とは何か？
- 表計算ソフトで分割表やクロス集計表を作るには何を使うのか？

今回学ぶこと

1. 分割表

2. クロス集計表

3. ピボットテーブル

クロス集計表

分割表において度数ではなく、合計、平均、標準偏差などを集計した表。

3×6クロス集計表の例

データセット 2 から求めた学年とコース別の学生一人当たりの授業外学修時間の平均のクロス集計表

	経営	政策	福祉	国際	観まち	情報	平均
2年生	99.7	99.1	95.6	99.4	105.7	106.9	101.7
3年生	95.6	96.5	95.7	94.2	104.3	102.2	98.9
4年生	98.7	94.4	97.1	95.5	102.4	101.9	99.3
平均	98.3	97.1	96.1	96.4	104.2	103.7	100.1

周辺度数にあたるところは「合計」ではなく「平均」であることに注意。

3×6クロス集計表の例

	経営	政策	福祉	国際	観まち	情報	平均
2年生	99.7	99.1	95.6	99.4	105.7	106.9	101.7
3年生	95.6	96.5	95.7	94.2	104.3	102.2	98.9
4年生	98.7	94.4	97.1	95.5	102.4	101.9	99.3
平均	98.3	97.1	96.1	96.4	104.2	103.7	100.1

例えば、「経営」コースの「平均」は98.3時間となっているが、これは元データで「経営」コース全体の平均を求めたものになっている。2年生、3年生、4年生のそれぞれの平均の平均 $(99.7+95.6+98.7)/3 = 98.0$ 時間ではない。この違いは各学年の人数が異なることから生じる。

今回の問い

- 分割表とは何か？
 - 2変量データを列と行に入れ、列と行の項目が交差するマスに対応する度数を数えて入れた表
- クロス集計表とは何か？
 - 合計、平均、標準偏差などを集計した表。
- 表計算ソフトで分割表やクロス集計表を作るには何を使うのか？

今回学ぶこと

1. 分割表

2. クロス集計表

3. ピボットテーブル

ピボットテーブル

分割表やクロス集計表を作るための
表計算ソフトの機能。

3変量の生データの例

右のように一番上の行に変数のラベルがあり、その下の列方向に各変数の観測値が並んでいる生データから、分割表やクロス集計表を作成する場合を考える。

	A	B	C
1	変数1	変数2	変数3
2	x_1	y_1	z_1
3	x_2	y_2	z_2
4	x_3	y_3	z_3
⋮	⋮	⋮	⋮

ピボットテーブルの作成

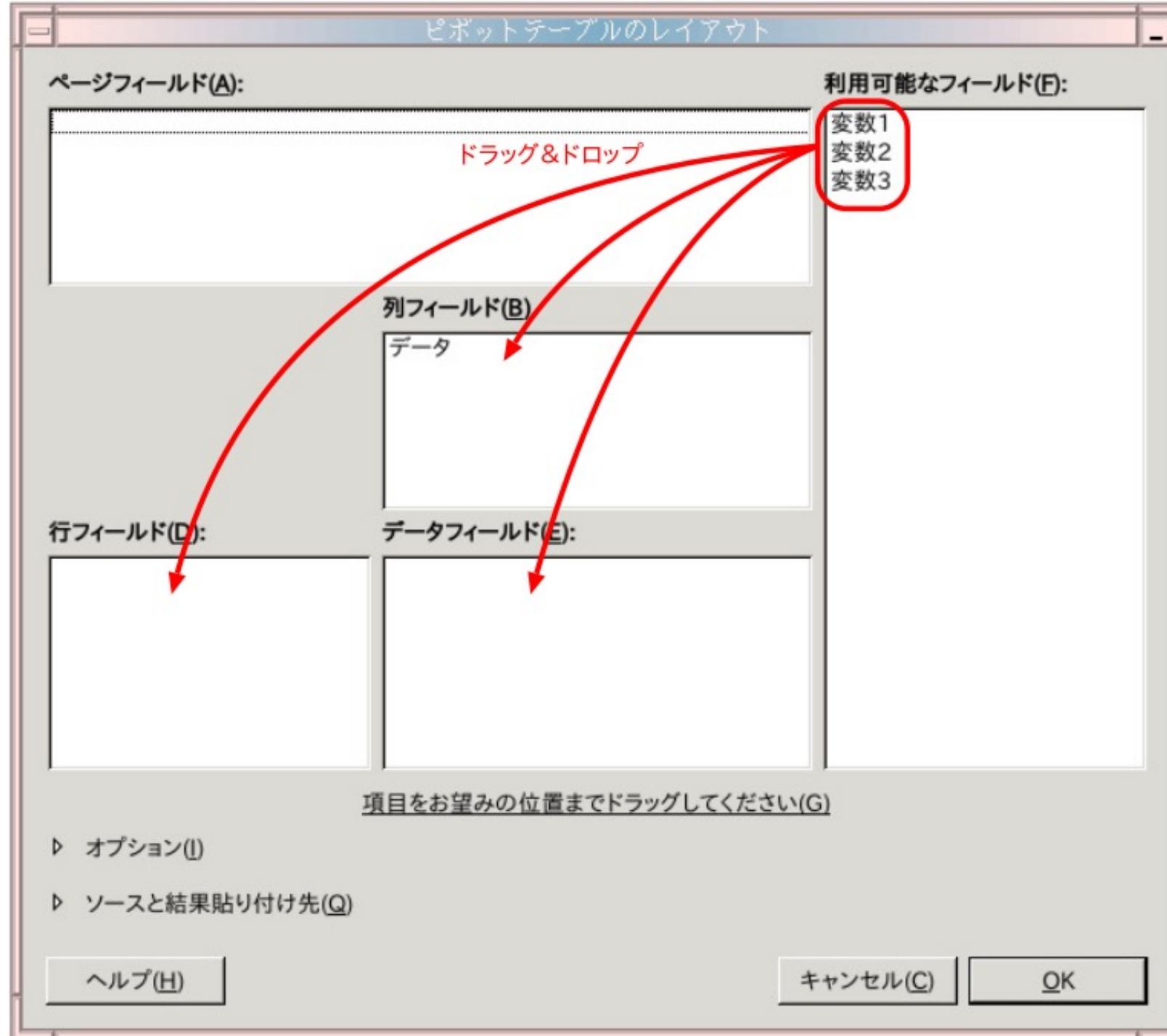
①ピボットテーブルの挿入

1. 分割表で数えたいデータの範囲を**変数のラベルを含めて**選択する。
2. メニューバーの[挿入]→[ピボットテーブル]を選択すると「ソースの選択」のダイアログが開く。
3. [現在の選択範囲]を選択して[OK]ボタンをクリックすると「ピボットテーブルのレイアウト」のダイアログが開く。

ピボットテーブルの作成

②ピボットテーブルのレイアウト設定

1. 「利用可能なフィールド」から分割表の表側の項目にしたい変数を「行フィールド」にドラッグ&ドロップする。
2. 「利用可能なフィールド」から分割表の表頭の項目にしたい変数を「列フィールド」にドラッグ&ドロップする。
3. 「利用可能なフィールド」から分割表の項目が交差するマスで数えたい変数を「データフィールド」にドラッグ&ドロップする。



ピボットテーブルの作成

②ピボットテーブルのレイアウト設定

4. 「データフィールド」にドラッグ&ドロップした変数は自動的に「合計」になるので、その変数名をダブルクリックする。
5. 「データフィールド」のダイアログが開いたら、計算方式を選択して[OK]ボタンをクリックする。
6. 「ピボットテーブルのレイアウト」の[OK]ボタンをクリックすると新しいシートにピボットテーブルが挿入される。

ピボットテーブルの作成

- 「列フィールド」「行フィールド」「データフィールド」「ページフィールド」に入れた変数は「利用可能なフィールド」にドラッグ&ドロップで戻すことで削除できる。
- ピボットテーブルの表頭と表側の項目は並べ替え（ソート）ができる。
- 「ページフィールド」に変数を入れてフィルタリングを行うことができる。

ピボットテーブルの編集

一度作ったピボットテーブルは編集して項目を変更することができる。

ピボットテーブル上で右クリックしてコンテキストメニューから[プロパティ]を選択すると「ピボットテーブルのレイアウト」のダイアログが開くので、ここで項目を変更する。

データの更新

ピボットテーブルの元のデータの値を更新してもピボットテーブルに自動的に反映されないため、手動で更新させる必要がある。

ピボットテーブルで右クリックしてコンテキストメニューから[更新]を選択するとピボットテーブルが更新される。

演習：分割表の作成

dataset2.odsを用いて、学年とコース別の学生数の分割表を作成する。

演習：分割表の作成

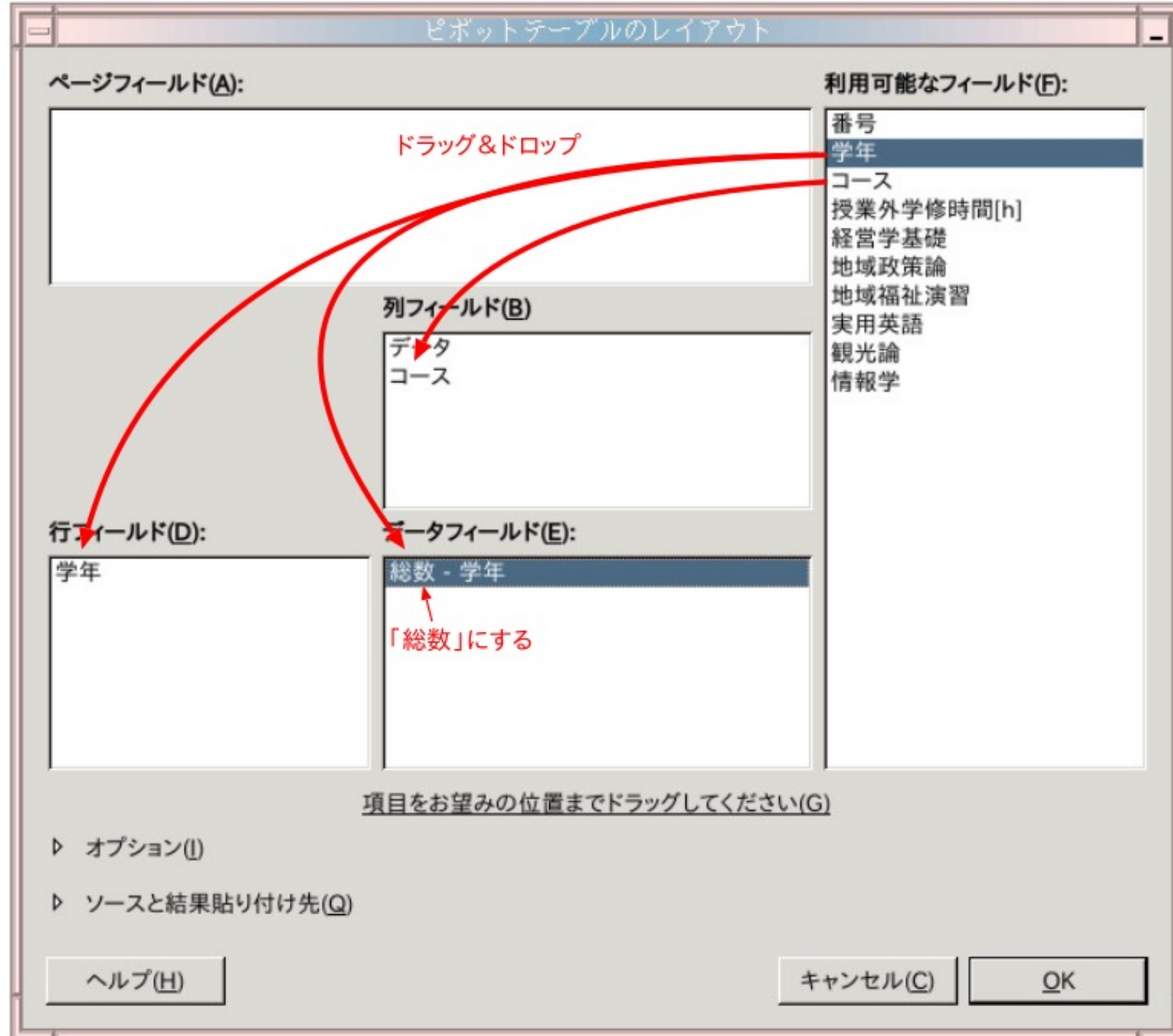
①ピボットテーブルの挿入

1. 列Aから列Jまでを選択する。
2. メニューバーの[挿入]→[ピボットテーブル]を選択し、[OK]ボタンをクリックすると「ピボットテーブルのレイアウト」のダイアログが開く。

演習：分割表の作成

②ピボットテーブルのレイアウト設定

1. 「利用可能なフィールド」から[学年]を「行フィールド」にドラッグ&ドロップする。
2. 「利用可能なフィールド」から[コース]を「列フィールド」にドラッグ&ドロップする。
3. 「利用可能なフィールド」から[学年]を「データフィールド」にドラッグ&ドロップする（この場合は[コース]でもよい）。



演習：分割表の作成

②ピボットテーブルのレイアウト設定

4. 「データフィールド」の[合計-学年]をダブルクリックする。
5. 「データフィールド」のダイアログが開いたら、計算方式の[総数]を選択して[OK]ボタンをクリックする。
6. 「ピボットテーブルのレイアウト」の[OK]ボタンをクリックすると新しいシートにピボットテーブルが挿入される。
7. 作成したピボットテーブルを空いている場所にコピー＆ペーストして分割表の形式を整える。

演習：平均のクロス集計表の作成

前の演習のピボットテーブルを編集して授業外学修時間の平均のクロス集計表を作成する。

演習：平均のクロス集計表の作成

1. 作成したピボットテーブル上で右クリックしてコンテキストメニューから[プロパティ]を選択すると、「ピボットテーブルのレイアウト」のダイアログが開く。
2. 「データフィールド」の[総数-学年]を「利用可能なフィールド」にドラッグ&ドロップで戻して削除する。
3. 「利用可能なフィールド」から[授業外学修時間[h]]を「データフィールド」にドラッグ&ドロップする。
4. 「データフィールド」の[合計-授業外学修時間[h]]をダブルクリックする。
5. 「データフィールド」のダイアログが開いたら、計算方式の[平均値]を選択して[OK]ボタンをクリックする。
6. 「ピボットテーブルのレイアウト」の[OK]ボタンをクリックする。

ピボットテーブルのレイアウト

ページフィールド(A):

利用可能なフィールド(F):

番号
学年
コース
授業外学修時間[h]
経営学基礎
地域政策論
地域福祉演習
実用英語
観光論
情報学

ドラッグ&ドロップ

列フィールド(B)

データ
コース

行フィールド(D):

学年

データフィールド(E):

平均 - 授業外学修時間[h]
「平均」にする

項目をお望みの位置までドラッグしてください(G)

オプション(I)

ソースと結果貼り付け先(Q)

ヘルプ(H)

キャンセル(C)

OK

作成したピボットテーブルでは、周辺度数に当たるところは「合計」と表示されるが、実際は「合計」ではなく「平均」であることを注意する。作成したピボットテーブルは空いている場所にコピー&ペーストして、分割表の形式を整える。

演習：平均のクロス集計表の作成

「行フィールド」「列フィールド」
「データフィールド」にその他の項目
を追加するとどうなるだろうか。
色々試してみよう。

今回の問い

- 分割表とは何か？
 - 2変量データを列と行に入れ、列と行の項目が交差するマスに対応する度数を数えて入れた表
- クロス集計表とは何か？
 - 合計、平均、標準偏差などを集計した表。
- 表計算ソフトで分割表やクロス集計表を作るには何を使うのか？
 - ピボットテーブルを用いる。

今回学んだこと

1. 分割表

2. クロス集計表

3. ピボットテーブル

今回の目標

多変量データについて分割
表とクロス集計表を作成で
きるようになる

課題

dataset1.odsにおいて、年齢と性別で人数を数えた分割表を作りなさい。表頭を「年齢」、表側を「性別」とする。また、同じデータセットで年齢と性別で身長を求めたクロス集計表を作りなさい。作成したファイルはODF表計算ドキュメント(拡張子.ods)で提出すること。

提出先：yutaka.hirai@koeki-u.ac.jp

件名：データリテラシー第9回課題 [学籍番号]

締め切り：12月11日(金)

次回

第 1 回	社会におけるデータ・AI利活用
第 2 回	表計算ソフトの基本操作
第 3 回	セルの参照と集計
第 4 回	度数分布表とヒストグラム
第 5 回	代表値と散らばりの尺度
第 6 回	データのグラフによる表現とデータ比較
第 7 回	データの分布と標本抽出
第 8 回	推定
第 9 回	分割表とクロス集計
第 1 0 回	2 変量データと相関
第 1 1 回	時系列データ
第 1 2 回	仮説検定
第 1 3 回	回帰分析