第8回
データの集計と可視化1

pandasを用いた単純集計、クロス集計

目次

事例:アンケートデータの集計

- pandasモジュールを使った、簡単なアンケートデータ(質的変数、量的変数)の集計について試します。
- データファイルをダウンロードする
  - Python講座のサイト
    http://pana4405.u-shizuoka-ken.ac.jp/pyk2
    から、2つのサンプルファイル「python.txt」「python.xlsx」をダウンロードします。前者はテキスト形式、後者はExcel形式のファイルです。
  - 以下、サンプルファイルがユーザのダウンロードフォルダに保存されていると仮定して演習を進めます。
  - ダウンロードした「python.txt」「python.xlsx」を開いて内容を確認します。

	A	B	C	D	E
1	性別	年代	職業	満足度	消費金額
2	女性	20代	学生	満足	¥7,400
3	女性	20代	学生	とても満足	¥4,000
4	男性	60代	無職	ふつう	¥9,300
5	女性	50代	会社員	やや満足	¥5,900
6	女性	20代	会社員	ふつう	¥9,400
7	女性	20代	パート・ア	やや満足	¥6,900
8	女性	40代	会社員	とても満足	¥6,500

質的変数

量的変数

python.xlsxの例

データの取り込み(テキストファイル)

- データをテキストファイル(csv, tsv)から取り込む方法について試します。
- 手順
  - Spyderを起動し、[新規]-[新規ファイル]を作成します。
  - pandasによってテキストファイルを取り込むには、read_csvメソッドを使います。その際、区切り文字を引数sepで指定します。
    - csv(カンマ区切り)ならば","
    - tsv(タブ区切り)ならば"¥t"(または、"/t")
  - ファイル名にURLを記述すれば、ネット上のファイルも取り込めます。
  - 取り込んだデータは、DataFrame型という行列形式のデータ型として管理されます。また、1行目は自動的にラベルとして認識されます。

DataFrame型

pandasのインポート

```
import pandas as pd
import matplotlib.pyplot as plt
plt.rcParams["font.family"] = "IPAexGothic"

data = pd.read_csv("C:/Users/user/Downloads/python.txt", sep="\t")
print(data)
```

userというユーザのDownloadsフォルダに保存されていると仮定(userは環境に合わせて書き換えてください)

データの取り込み（Excelファイル）

- pandasでは、テキストファイルだけでなくExcelファイルを直接読み込むことができます。
  - ExcelFileメソッドでExcelファイルを開く。
  - 開いたファイルのparseメソッドでシート名を指定して取り込む。

```
import pandas as pd
import matplotlib.pyplot as plt
plt.rcParams["font.family"] = "IPAexGothic"

#ダウンロード先の絶対パスを指定（フォルダの区切りはスラッシュで指定）
file = pd.ExcelFile("C:/Users/user/Downloads/python.xlsx")
#または、下記のようにインターネットから直接読み込むことも可能
#file = pd.ExcelFile("http://pana4405.u-shizuoka-ken.ac.jp/wp-content/
uploads/2018/10/python.xlsx")
data = file.parse("アンケート")
print(data)
```

userというユーザのDownloadsフォルダに保存されていると仮定（userは環境に合わせて書き換えてください）

「アンケート」シートの取り込み

取り込んだデータには自動的に連番のindexが割り当てられます。

	性別	年代	職業	満足度	消費金額
0	女性	20代	学生	満足	7400
1	女性	20代	学生	とても満足	4000
2	男性	60代	無職	ふつう	9300
3	女性	50代	会社員	やや満足	5900

特定の行や列を取り出す

- pandasのDataFrameでは特定の行や列を取り出すことができます。

0行目
1行目

	性別	年代	職業	満足度	消費金額
0	女性	20代	学生	満足	7400
1	女性	20代	学生	とても満足	4000

列名

  - 特定の列を取り出す
    - データフレーム.列名
  - インデックス名やラベル名を指定して取り出す
    - データフレーム.loc[行インデックス名, 列ラベル名]
  - 番号を指定して取り出す
    - データフレーム.iloc[行番号, 列番号]

取り出した特定の行や列は、Series型となります

```
#性別の列を取り出し
print(data.性別)
#0行目の性別を取り出し
print(data.loc[0, "性別"])
#0行目の性別と満足度を取り出し（リストで指定）
print(data.loc[0, ["性別","満足度"]])
#0行0列を取り出し
print(data.iloc[0, 0])
#0行の0列から3列までを取り出し（スライスで指定）
print(data.iloc[0, 0:4])
#0行の1列と3列を取り出し
print(data.iloc[0, [1, 3]])
#性別が男性の行だけ取り出し
print(data[data.性別 == "男性"])
```

行指定の0を「:」とすると、全行を取り出せます。

質的変数の単純集計(1)

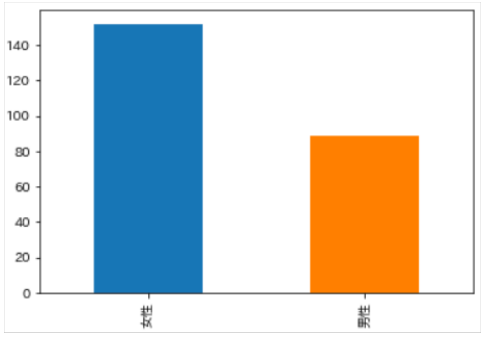
- value_countsメソッドにより、質的変数の項目ごとに単純集計を実行できます。
  - 構文　データフレーム.value_counts()
- 集計により新たなDataFrame（Series）が生成され、そのDataFrame（Series）を元に縦棒グラフを描画します。
  - 構文　データフレーム.plot.bar()

集計結果は新たなデータフレームになる

```
sei = data.性別.value_counts()
print(sei)
sei.plot.bar()
```

```
女性    152
男性     89
```

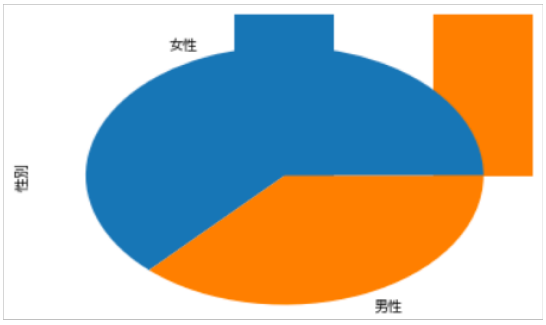
pandasでは内部的にmatplotlibを利用してグラフが描画できる



質的変数の単純集計(2)

- 棒グラフbar以外にも、円グラフpieなど様々な種類のグラフを描くことができます。
  - bar（縦棒グラフ）、barh（横棒グラフ）、pie（円グラフ）、scatter（散布図）、line（折れ線グラフ）など
- 複数のグラフを一度に描きたい場合は、単純にplotメソッドを並べただけでは、グラフが重なって描かれてしまいます。

```
sei.plot.bar()
sei.plot.pie()
```



質的変数の単純集計(3)

- 複数のグラフを並べたい場合は、あらかじめmatplotlibのpyplotで、subplotsメソッドを使いグラフをいくつ並べるか設定しておき、軸（axes）を指定してグラフを描きます。

```
sei = data.性別.value_counts()
print(sei)
fig, axes = plt.subplots(nrows=1, ncols=2, figsize=(10, 5))
sei.plot.bar(ax=axes[0])
sei.plot.pie(ax=axes[1], autopct="%.1f%%")
```

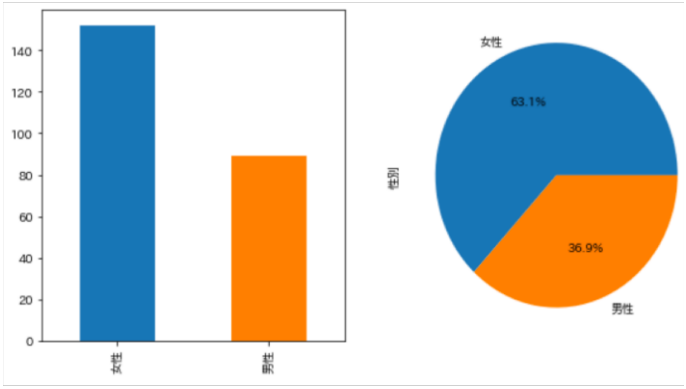
グラフの行数

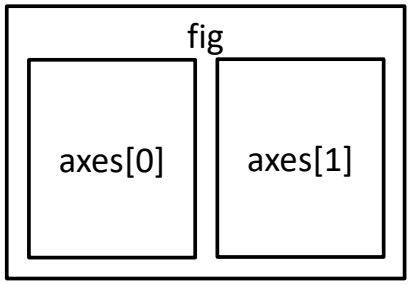
グラフの列数

グラフ全体のサイズ（インチ）

軸の指定

円グラフに割合を表示（小数点以下1位まで%つきで表示）





質的変数の単純集計(4)

- その他、様々なグラフの調整が可能です。

タイトルや配色の指定（r, g, b, c, m, y, k, wや、#ff0000など16進数のカラーコード）が可能

文字サイズ

```
plt.rcParams["font.size"] = 15
fig, axes = plt.subplots(nrows=1, ncols=2, figsize=(10, 5))
fig.tight_layout()
sei.plot.bar(ax=axes[0], title="性別", color=["r", "b"])
sei.plot.pie(ax=axes[1], autopct="%.1f%%")
plt.sca(axes[0])
plt.xticks(rotation=0, color="red")
plt.ylabel("人")
plt.legend()
plt.show()
```

現在の設定対象をaxes[0]に指定（グラフが1つだけの場合は不要）

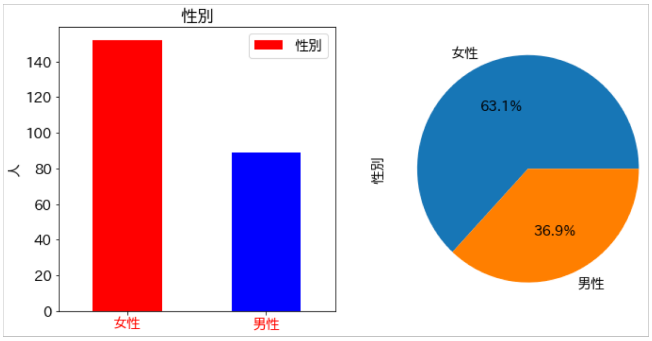
X軸の文字の調整

Y軸の軸ラベルの指定

凡例の表示

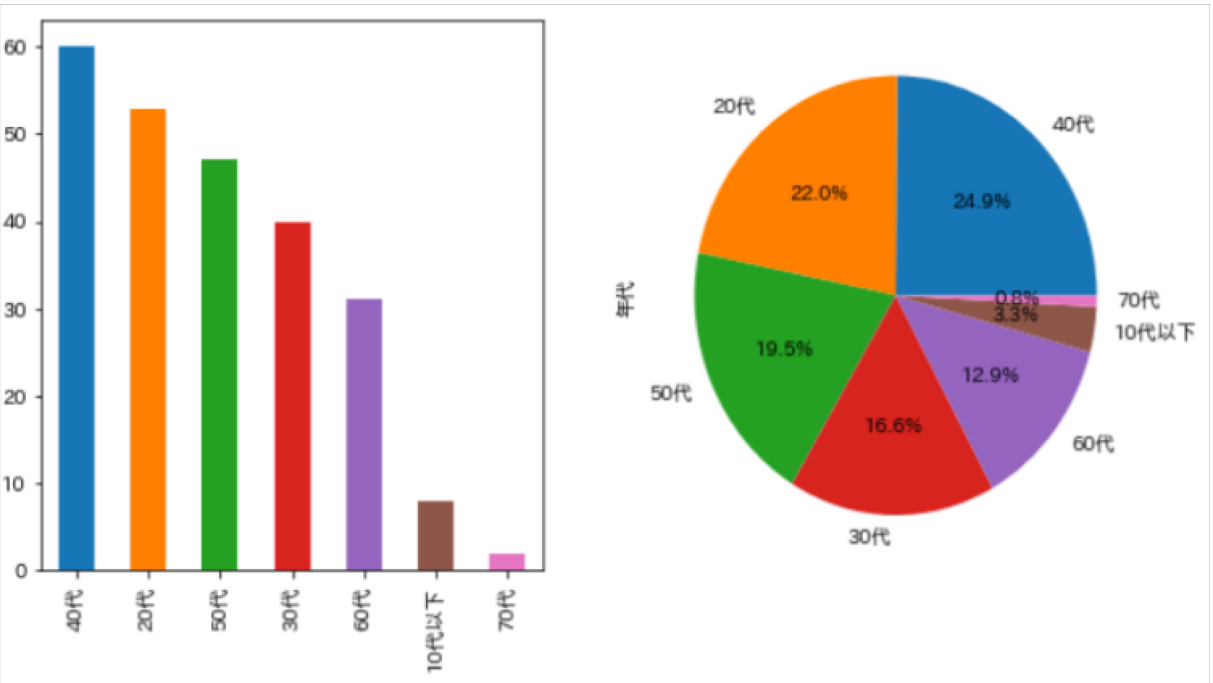
設定の反映

グラフ間の余白を詰める



課題8

- 年代、職業、満足度についてもコードを修正してグラフを描画してください。



量的変数の単純集計

- 量的変数については、和（sum）、平均値（mean）、最小値（min）、最大値（max）などの集計が可能です。
- 基本統計量の一括計算がdescribeメソッドで可能です。
- 量的変数に対してヒストグラムの描画がplot.histメソッドで可能です。

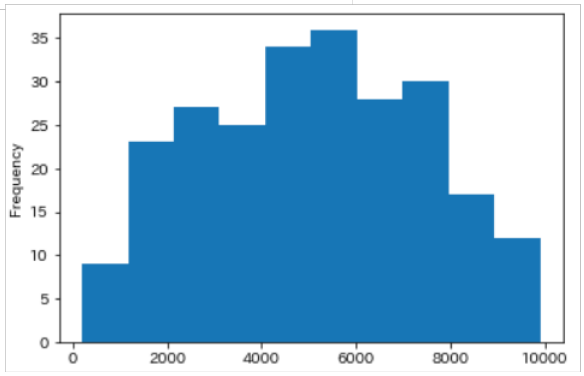
```
print(data.消費金額.sum())
print(data.消費金額.mean())
print(data.消費金額.min())
print(data.消費金額.max())
print(data.消費金額.describe())
data.消費金額.plot.hist(bins=10)
plt.show()
```

和や平均値の計算

基本統計量

binsでヒストグラムの区分数を指定

```
1211530
5027.095435684647
200
9900
count     241.000000
mean     5027.095436
std      2351.373328
min       200.000000
25%      3200.000000
50%      5100.000000
75%      6900.000000
max      9900.000000
```



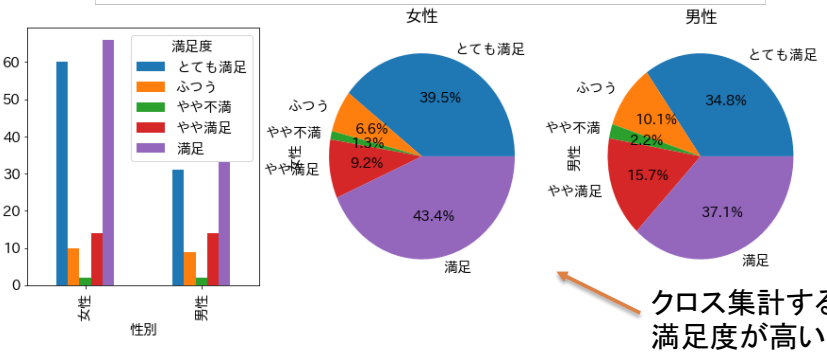
クロス集計(1)

- 2つの項目を行と列に割り当てて集計するクロス集計は、pandasのcrosstabメソッドで簡単に実行できます。

```
cross = pd.crosstab(data.性別, data.満足度)
print(cross)
fig, axes = plt.subplots(nrows=1, ncols=3, figsize=(15, 5))
cross.plot.bar(ax=axes[0])
cross.loc["女性"].plot.pie(ax=axes[1], autopct="%.1f%%", title="女性")
cross.loc["男性"].plot.pie(ax=axes[2], autopct="%.1f%%", title="男性")
plt.show()
```

stacked=Trueを追記すると積み上げ型に

満足度	とても満足	ふつう	やや不満	やや満足	満足
性別					
女性	60	10	2	14	66
男性	31	9	2	14	33



クロス集計することで、女性の方が満足度が高いことがわかる。

クロス集計(2)

- pandasのcrosstabメソッド以外に、DataFrameのpivot_tableメソッドでクロス集計することもできます。下記の例では、消費金額の合計値を集計しています。

集計項目　行　列

```
cross = data.pivot_table(values="消費金額", index="性別", columns="年代",
aggfunc="sum")
print(cross)
```

年代	10代以下	20代	30代	40代	50代	60代	70代
性別							
女性	27300.0	174910.0	125400.0	220300.0	148000.0	54500.0	NaN
男性	14000.0	81600.0	78700.0	72500.0	94100.0	116220.0	4000.0

  - 値が存在しない項目にはNaNが表示されますが、fill_value=0を設定することで、強制的に0で埋めることができます。

```
cross = data.pivot_table(values="消費金額", index="性別", columns="年代",
aggfunc="sum", fill_value=0)
print(cross)
```

NaNを0で埋める

年代	10代以下	20代	30代	40代	50代	60代	70代
性別							
女性	27300	174910	125400	220300	148000	54500	0
男性	14000	81600	78700	72500	94100	116220	4000

クロス集計(3)

- crosstabやpivot_table以外に、同じ値を持つ行をグループ化するgroupbyメソッドによっても項目ごとの集計が可能です。
  - groupbyの場合には、列に並ぶ項目は無く、すべて行方向に並ぶ形になります。

性別と年代でグループ化

```
print(data.groupby(["性別", "年代"]).消費金額.sum())
```

消費金額の和を計算

```
性別   年代
女性   10代以下      27300
       20代        174910
       30代        125400
       40代        220300
       50代        148000
       60代         54500
男性   10代以下      14000
       20代         81600
       30代         78700
       40代         72500
       50代         94100
       60代        116220
       70代          4000
```

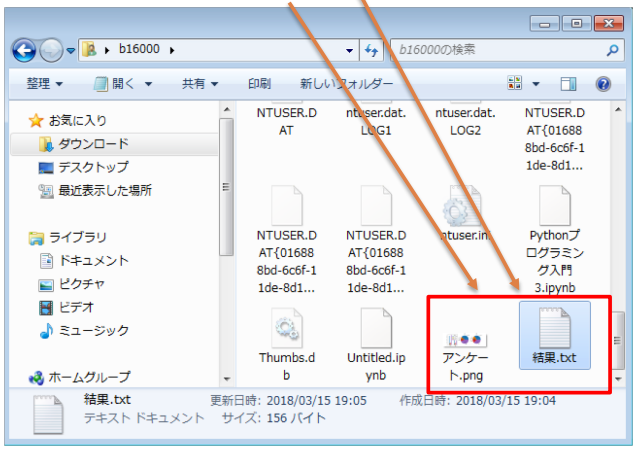
下記のpivot_tableでも同じ結果が得られます。

```
print(data.pivot_table(values="消費金額",
index=["性別", "年代"], aggfunc="sum",
fill_value=0))
```

列ではなく全て行に項目を指定

結果の保存

- matplotlibで描いた図は、plt.savefigメソッドで画像として保存できます。
  - ただし、plt.showメソッドの前に書く必要あります。
  - 例：`plt.savefig("アンケート.png")`
- pandasのDataFrameは、to_csvメソッドでcsvファイルとして保存できます。
  - 例：`cross.to_csv("結果.txt")`



(補足)Seriesの各要素の値の加工(1)

- DataFrameの各行や各列はSeries型となる
- Series型の各要素の値を、あるルールで加工したい場合は？
  - 例：消費金額の各要素に、消費税8%分を加えたい
  - mapメソッド、applyメソッドで各要素に関数を適用すれば実現できる

```
def plus8(x):
    return int(x * 1.08)

print(data.消費金額.map(plus8))
```

各要素にplus8関数が適用される

applyでもOK

0	3000
1	2500
2	2400
3	5900
4	5100
5	6900
6	6500

8%up

0	3240
1	2700
2	2592
3	6372
4	5508
5	7452
6	7020

- ラムダ式を使うと関数を定義しなくても同じことができる(無名関数)
  - 構文・・・ lambda 各要素の値をとる変数 : 変数に対する処理

```
print(data.消費金額.map(lambda x: int(x * 1.08)))
```

- mapやapplyを実行しても元のSeriesは変更されないが、上書きすれば変更できる

```
data.消費金額 = data.消費金額.map(lambda x: int(x * 1.08))
print(data.消費金額)
```


(補足)Seriesの各要素の値の加工(2)

- 要素が文字列の場合は、str.文字列メソッドが使用できる
  - 例：満足度の「ふつう」を「普通」と漢字に置換する

```
print(data.満足度.str.replace("ふつう", "普通"))
```

  - その他
    - str.strip() ・・・空白の削除
    - str.lower()・・・小文字に
    - str.upper()・・・大文字に
    - ・・・

- mapでディクショナリを引数として指定すると、複数種類の置換が可能(質的変数の量的変数への置き換えに便利)
  - 例：男性を0、女性を1に変換

```
print(data.性別.map({"男性": 0, "女性": 1}))
```

0	女性
1	女性
2	男性
3	女性
4	女性

0	1
1	1
2	0
3	1
4	1