データサイエンス特論
Data Science

クロス集計とヒートマップ

多次元のデータ（multidimensional data）を直接見ても、人間がその特長を捉えることは困難である。これを何とかしたい。

多次元は，数学用語のn次元（n個の変数）に基づいている。多変量解析（multivariate analysis）もn次元の変量（変数）を扱う統計的手法であり，この一部と考えられても差し支えない

1. はじめに
2. クロス集計
3. 相関行列
4. ヒートマップ

(C) 創造技術専攻　橋本洋志／大久保友幸
{hashimoto, ohkubo-tomoyuki}@aiit.ac.jp
http://hhlab.org/

クロス集計（cross tabulation）

- 分割表（contingency table）とも呼ばれる。
- カウンタブル（countable）な集合数（四則演算可）に対して行われる
  ⇔順序数（ordinal number）
- クロス集計が何を行っているかの説明は，各自の学習に譲る。ここでは，ある例題を通して，Pythonプログラムの説明を行う。

- 他の事例
  - タイタニック号
  - 東京の中小企業の現状
    - 検索サイト　キーワード「東京の中小企業の現状」
    - http://www.sangyo-rodo.metro.tokyo.jp/toukei/chushou/genjyou/
    - 例えば，平成28年度「サービス産業編」この中に 「クロス集計」が多数ある

クロス集計

- **例：第1子（first）と第2子（second）の職業を見る。職業はスポーツ選手か否かである。次の結果を見てどう考えるか？**

	sport
first	73 人
second	25 人

- この結果から，第1子の方がスポーツ選手に向いている，と言えるだろうか？
- この誤りは
  - スポーツ選手以外の職業に就いた人数が不明のため，母数に対する割合がわからない
  - しかも，第1子だけがいて第2子がいない家庭はあるが，その逆は無い。従って，もともと第1子がいる，というサンプル数が多くなったと考えるのが自然である。
- この結果の分析を行う方法の一つとして
  - クロス集計を行った後に，第1子の結果と第2子の結果が独立であるか否かの独立性の検定を行う方法がある。独立性の検定にはピアソンのカイ二乗検定をここでは採用する。

クロス集計

例:このデータは人工的に作成したものである。 Ex_CrossTab

- アンケート項目は，第1子か第2子か
- その人の職業はスポーツ選手(sport)か否(other)か
- 1000人にアンケートを行った結果，第1子が800人，第2子が200人であった。
- そのデータの一部を示す。

```
df = pd.DataFrame({ 'id' : range(n_sum),
                    'child': child,
                    'job' : job },
                  columns=['id', 'child', 'job']) # columnsが無いと，アルファベット順に並ぶ
df.head()
```

	id	child	job
0	0	first	other
1	1	first	other
2	2	first	other
3	3	first	other
4	4	first	other

クロス集計

```
cross = pd.crosstab(df.child, df.job)
cross
```

job	other	sport
child		
first	727	73
second	175	25

クロス集計

クロス集計のカイ二乗検定

- 右の表がクロス集計
- クロス集計は何を表す？

	other	sport	total
first	727	73	800
second	175	25	200
total	902	98	1000

帰無仮説

- 各事象は独立である（仮説）
  - 独立の意味は？
- これに基づくと
  - 第1子，第2子がスポーツ選手，それ以外の職業に就く確率は独立である。
  - すなわち，生まれた順序と職業は関係しない

検定の計算手順

- 他の職業を選ぶ確率が(902/1000)，スポーツを選ぶ確率が(98/1000)を考える。これは，第1子，第2子の要因を取り除いた考え方である。
- これらの確率を 第1子の人数（800）にそれぞれ乗じると，第1子が独立に職業を選ぶ確率となる。これがexpectedである。第2子も同様である。
- 下記のカイ二乗値の式を計算する。この例では，4つの項の総和がカイ二乗値となる。
- カイ二乗分布の式に類似している。

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}},$$

次のページ以降でこの計算を説明する

クロス集計；カイ二乗検定

Ex_CrossTab

例:

- カイ二乗検定は次の関数を用いる
- x2: カイ二乗値，　p:p値，dof：自由度（Degree Of Freedom），
- expected: 帰無仮説から導かれる頻度の期待値

```
1  chi2, p, dof, expected = sp.stats.chi2_contingency(cross, correction=False) # correction=Falseは通常の集計
```

```
1  print(chi2)
2  print(p)
3  print(dof)
4  print(expected)
```

```
2.06174487533
0.151037134524
1
[[ 721.6   78.4]
 [ 180.4   19.6]]
```

有意水準を5%とすれば，p値0.151（>0.05）は棄却できない。
すなわち，第1子と第2子でスポーツ選手に就く確率は独立である，という帰無仮説は棄却できない。
この”棄却できない”とは何を意味するかは、統計分析で明らかになる。

相関行列

相関とは

- 式で見ると

データ$x = \{x_i\}, y = \{y_i\}$の相加平均を$\bar{x}$、$\bar{y}$

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\left(\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)\left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right)\right)^{1/2}}$$

- rの性質

  −1 ≦ r ≦ 1

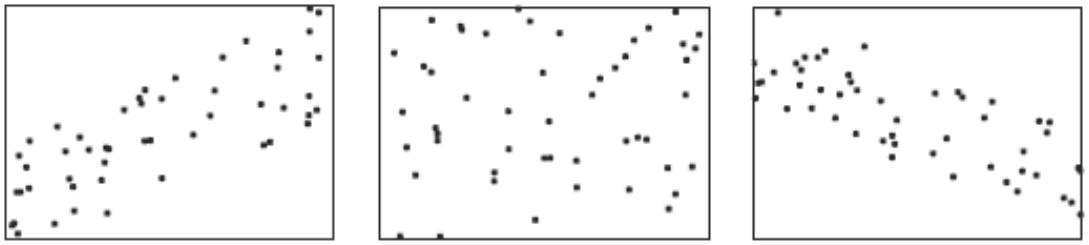
  rが　1に近いほど、「正の相関が高い」

  rが−1に近いほど「負の相関が高い」

  xとyが同じ（自分自身で相関を見る）ならば、r ＝1となる。

と言われる。注意：rが0.6, 0.8 に従い、少し高い、十分高いとは一般には言えない。

- イメージで見ると（例えばの話）
  - 親の身長とその子供の身長は、正の相関が高い（親の身長：x,　子供の身長:y）
  - 体重と100m走のタイムは、負の相関が高い（体重：x, タイム：y）



(a) 正の相関 $\rho_{xy} = 0.633$　(b) 無相関 $\rho_{xy} = 0.062$　(c) 負の相関 $\rho_{xy} = -0.785$

散布図と相関係数

注意：相関係数の絶対値が大きくなるほど，相関が強くなるというのは一般的な解釈であるが，経験上，N < 20 程度のとき，相関係数 = 0.7 程度でも実際にはかなり弱い相関であることがある。したがって，相関係数の数字だけで判断するのではなく，散布図を描いて確かめることも必要である。

相関行列

相関行列とは

確率変数に対する相関行列

n 個の確率変数 $X_1, X_2, \cdots, X_n$ に対して，
ii 成分が 1，ij 成分が ρ_{ij}（X_i と X_j の相関係数）
であるような $n \times n$ 行列 C を相関行列と言います。定義より，相関行列は非対角成分が-1 以上 1 以下であるような対称行列です。

行列の各要素が相関係数
対角は、自分自身ゆえ1となる

注：「$i \neq j$」という文言は省略します。

データに対する標本相関行列

同様に，n 次元のデータに対しても（標本）相関行列が定義されます（対角成分には 1，非対角成分には標本相関係数が並ぶ）。

二組の対応するデータ (X, Y) に対して，相関係数 ρ を以下で定義する：

$$\rho = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

共分散 $\mathrm{Cov}(X,Y)$ は二組の対応するデータの間の関係を表す数値である。
データを$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ とおくとき，
$\mathrm{Cov}(X,Y) = E[(X - \mu_X)(Y - \mu_Y)]$

□ 引用：高校数学の美しい物語 https://mathtrain.jp/correlationmatrix

ヒートマップ（Heat map）

- ここでは，相関行列の表現に用いる
- データ行列（左記の相関行列も含まれる）の各要素の値に応じた色で表現する，可視化グラフの一種である。
- 視覚に訴える図解の資料として用いられる。
- 色の与え方で，恣意的にあるデータを強調されることがあるので，見るときには注意が必要である。

例：天候と家計支出の関係

家計調査

- e-Stat https://www.e-stat.go.jp/ の「データベースから探す」を選択
- 一覧「家計調査」を選択
- 「二人以上の世帯」の「月次」を選択
- 一覧「提供周期で絞込み」から「月次」を選択
- 品目分類「010」「品目分類（平成27年改定）（総数：金額）」の「DB」をクリック
- 表章項目を「金額」，世帯区分「二人以上の世帯のうち勤労者世帯（2000年～）」を選択，地域区分「東京都区部」を選択して，「更新」をクリックしてから「ダウンロード」をクリック
- 現れたダウンロードのデータ形式を設定するウィンドウ，特に何も指定することなく「ダウンロード」をクリック，ダウンロードまで数分かかる。
- ダウンロードされたcsvファイル（“FEH_00200561_xxxxxxxxxxxx.csv”）を開いて，項目が多数ある

この操作は2018年度時点、現在は幾つか変更あり

気象データ

- http://www.data.jma.go.jp/gmd/risk/obsdl/index.php から 地点；「東京」，
- 項目： 日別値，日最高気温，降水量の日合計，日平均相対湿度，期間を選び，
- 「CSVファイルをダウンロード」をクリック
- weather_data.csv

2018年度時点のデータを用いる
フォルダValuable_Kit内にある

2018年度時点のデータを用いる
フォルダValuable_Kit内にある

データの作成

- 気象データ：日次となっているので月次とする。このため，月毎に最高気温と平均湿度は－＞それぞれの平均値，雨量は合計とした。
- 家計支出項目で選んだ項目：アイスクリーム，外食，男性用コート，女性用コート，化粧水

url = 'https://sites.google.com/site/datasciencehiro/datasets/weather_items.csv'

上の2つのデータを加工した後にまとめたもの

例：天候と家計支出の関係

Ex_Heatmap

```
: 1 url = 'https://sites.google.com/site/datasciencehiro/datasets/weather_items.csv'
  2 df = pd.read_csv(url, index_col='date', parse_dates=[0],
  3                        comment='#', encoding='SHIFT-JIS')
  4 df.head()
```

:

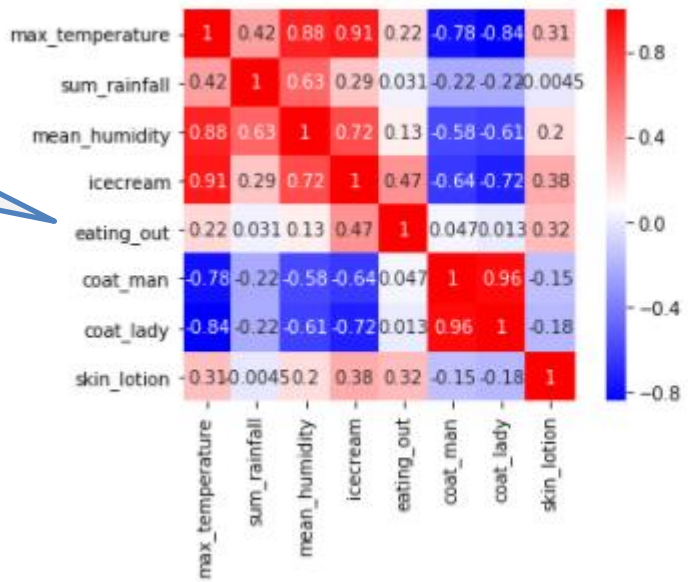
date	season	max_temperature	sum_rainfall	mean_humidity	icecream	eating_out	coat_man	coat_lady	skin_lotion
2016-01-01	winter	10.6	85.0	54.5	464	14703	272	709	291
2016-02-01	winter	12.2	57.0	56.2	397	12428	118	364	347
2016-03-01	spring	14.9	103.0	60.5	493	14506	85	342	399
2016-04-01	spring	20.3	120.0	66.8	617	13361	14	135	352
2016-05-01	spring	25.2	137.5	66.1	890	15311	48	58	364

例：天候と家計支出の関係

```
1 corr = df.corr()
2 corr
```

	max_temperature	sum_rainfall	mean_humidity	icecream	eating_out	coat_man	coat_lady	skin_lotion
max_temperature	1.000000	0.418414	0.875356	0.909101	0.219756	-0.782589	-0.843763	0.308308
sum_rainfall	0.418414	1.000000	0.629814	0.285431	0.031304	-0.224589	-0.221839	-0.004549
mean_humidity	0.875356	0.629814	1.000000	0.724164	0.132064	-0.583589	-0.607045	0.199416
icecream	0.909101	0.285431	0.724164	1.000000	0.469277	-0.638069	-0.723326	0.379101
eating_out	0.219756	0.031304	0.132064	0.469277	1.000000	0.046847	0.013107	0.316940
coat_man	-0.782589	-0.224589	-0.583589	-0.638069	0.046847	1.000000	0.959226	-0.149612
coat_lady	-0.843763	-0.221839	-0.607045	-0.723326	0.013107	0.959226	1.000000	-0.177888
skin_lotion	0.308308	-0.004549	0.199416	0.379101	0.316940	-0.149612	-0.177888	1.000000

```
1 heatmap = sns.heatmap(corr, annot=True, square=True, cmap='bwr')
```



ヒートマップ
この見方に慣れましょう

結果から

- 最高気温，アイスクリームとコート売上に相関が強い
- 外食は天候と相関が無い
- スキンローションは，雨や湿度とほとんど相関が無い

これは正しい？？

例：天候と家計支出の関係

クロス集計を行うために，気象データを削除する
次に，groupby機能を用いて，四季毎にデータをまとめる

```
df_1 = df.drop(['max_temperature', 'sum_rainfall', 'mean_humidity'], axis=1)
df_season = df_1.groupby('season')
```

メソッドsum()を用いて，クロス集計を行う

```
df_season_sum = df_season.sum()
df_season_sum
```

season	icecream	eating_out	coat_man	coat_lady	skin_lotion
autumn	3837	80265	814	2187	2079
spring	4011	86175	257	1081	2124
summer	7240	86859	36	238	2293
winter	2866	83413	1405	3231	2103

結果から何が言える？

備考：クロス集計を行っているからと言って，この例のように，セルに０か１の数をカウントするという頻度を扱っていないので，カイ2乗検定は適用しない。このデータは時系列データゆえ，先に示したように，相関係数からなる相関行列を見る方が良い。

おわりに

- 課題がでていますので、それに取り掛かってください。