統計分析法　第一回レポート

実行環境

python3.9.7

- pipenv

(これ以降のモジュールは `pipenv update` でダウンロードできます)

- requests
- pandas
- matplotlib
- cogapp
- scipy

準備:データを取得する

演習データは、 `requests.get()` メソッドによって取得した。

そのときに、URLの末端の `n=1000` の部分を変更することで取得する標本数を変更できるため、関数の引数で取得する標本数を変更できるようにした。

また、そのときに返されるデータには、値がダブルクォートで囲まれているため、 `replace()` メソッドでダブルフォートを削除し、加工がしやすいように、csvファイルにまとめた。

取得したデータ

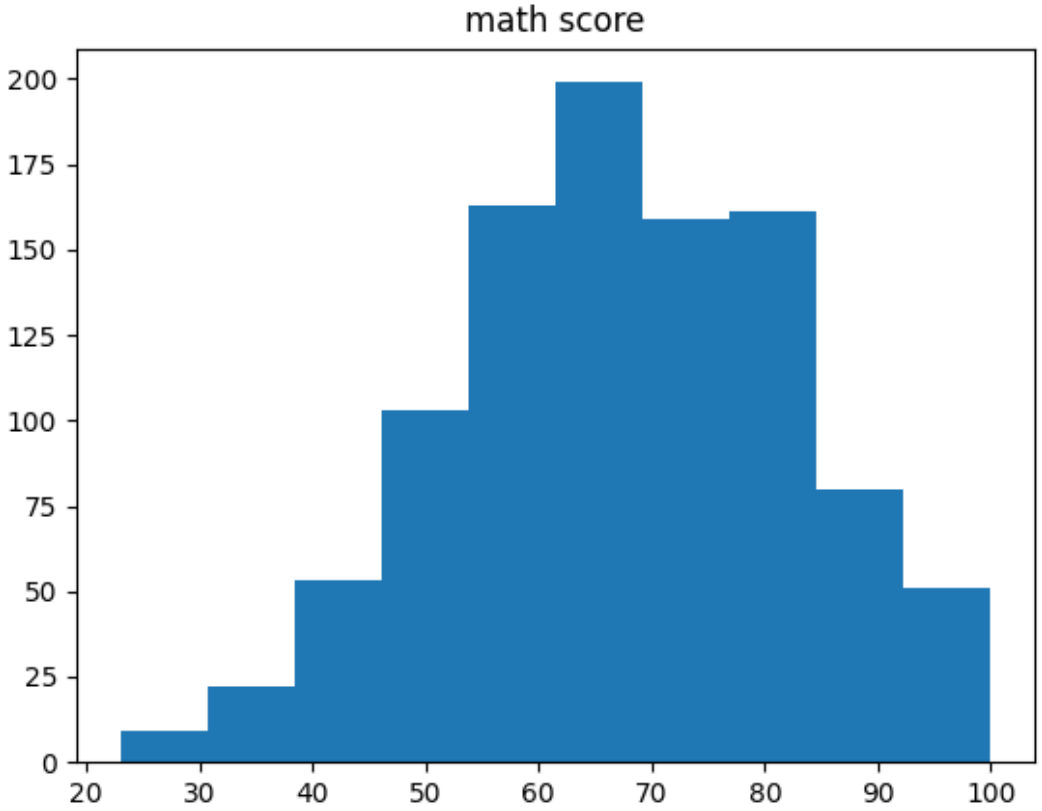
1 量的データの基本統計量

task 1-1:母集団の分布を確認する

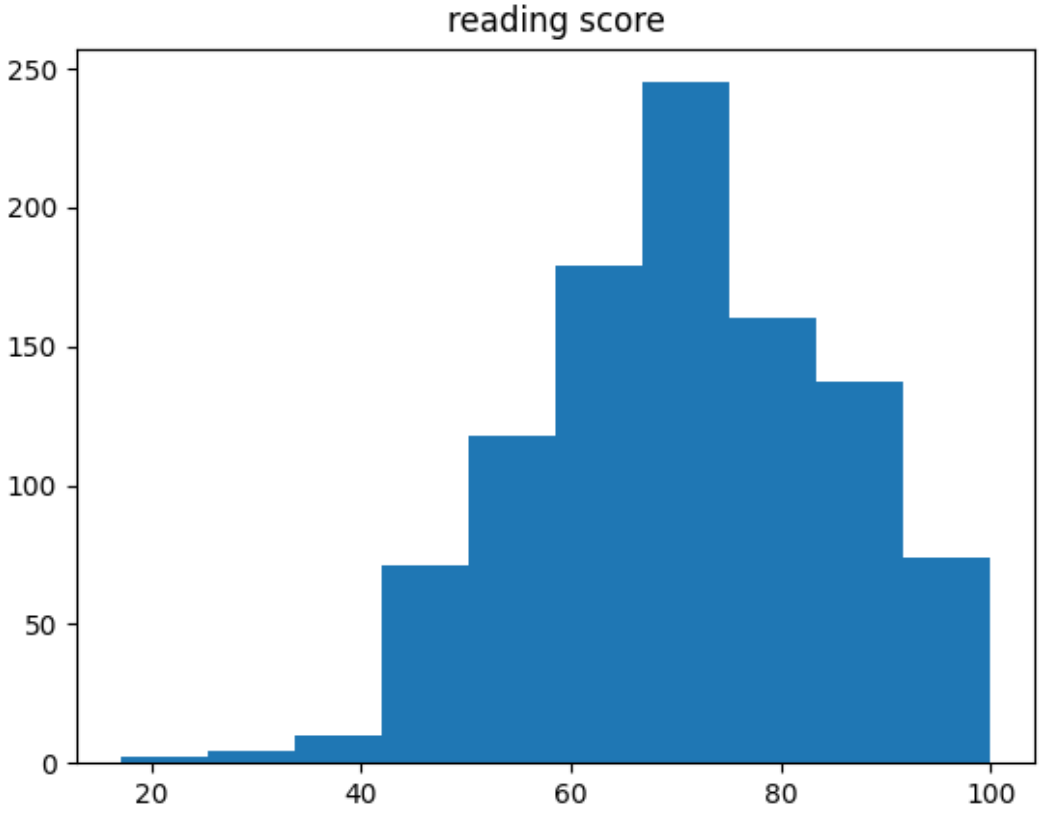
方針: データを取得したファイルから、量的データを取り出し、それぞれの値について、 `matplotlib.pyplot` の `hist` でヒストグラムを作成する。

実行結果

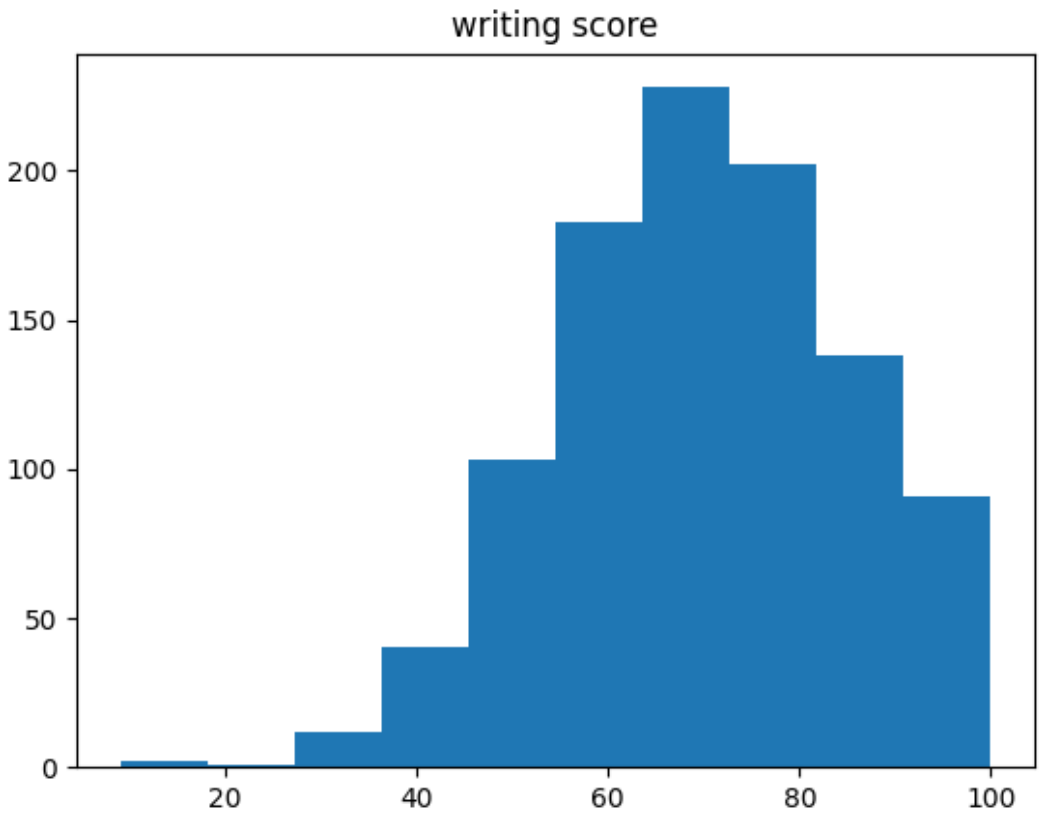
数学のスコア:



読みのスコア　:



書きのスコア:



task 1-2:母集団の基本統計量を調べる

`pandas.mean()` メソッドで平均を、 `pandas.var(ddof=0)` メソッドで分散を、 `pandas.std(ddof=0)` メソッドで標準偏差を、 `pandas.corr()` メソッド相関係数をそれぞれ計算した。

実行結果

```
$ pipenv run python3 task1-2.py
mean:
math score       67.477
reading score    70.706
writing score    69.665
dtype: float64
var:
math score       228.719471
reading score    206.379564
writing score    228.602775
dtype: float64
std:
math score       15.123474
reading score    14.365917
writing score    15.119616
dtype: float64
corr:
```

```
               math score  reading score  writing score
math score       1.000000       0.813027       0.802820
reading score    0.813027       1.000000       0.948475
writing score    0.802820       0.948475       1.000000
```

task 1-3:母平均の推定

標本データの抽出は `pandas.sample()` メソッドによって行った。

点推定

中心極限定理によって推測し、(サンプル数を取得する回数を1とするとそのまま点推定になる)標本平均の平均、標本平均の分散はそれぞれ `pandas.mean()` 、 `pandas.var(ddof=0)` メソッドによって取得した。

区間推定

`scipy.stats.sem()` メソッドによって標準誤差を算出し、、 `scipy.stats.t.interval(alpha,df,loc,scale)` によって、信頼度が `alpha` 、自由度が `df` ,平均が `loc` ,標準誤差が `scale` のときの区間の上限と下限を求めた。

今回は、1回で取得する標本数を20、区間推定で標本を取得する数を1回、信頼度を95%にして推定を行った。

推定に使用したファイル: 点推定 区間推定

実行結果

```
$ pipenv run python3 task1-3.py 20 1 0.95
samples:20, sets:1
CLT:
     math score  reading score  writing score
ave        69.5           73.6          73.25
var         0.0            0.0           0.00

samples:20, confience_level:0.95
interval:
      math score  reading score  writing score
down   53.207854      60.599074       58.98058
up     67.292146      73.700926       72.21942
```

2 仮説検定

今回使うデータについて

今回は、準備:データを取得するのデータを取得する数を引数で指定できることを活かして、 1 量的データの基本統計量とは別の50個のデータを標本として使用した。

検定に使用した標本

task 2-1:質的データの検定

科目平均の差に意味があるかの検定

帰無仮説は `各科目の平均の差が0` である。また、それに対する対立仮説は `各科目の平均の差が0ではない` である。

今回は科目の得点のデータが50個ずつあることからデータの対応がある。そのため有意確率については、 `ave + t x stderr = 0` となるようなt値が自由度49での両側検定での値とする。

有意水準は5%とする。つまり、有意確率が5%より低ければ、帰無仮説は却下され、各科目の平均の差があると言える。

また、有意確率の計算については、調べたい変数を抽出したあと、
`scipy.stats.ttest_rel()` メソッドによって行った。

異なる学生グループの科目平均の差に意味があるかの検定

今回は、性別で学生をグループ分けを行った。

帰無仮説は `両グループの各科目の平均の差が0` である。また、それに対する対立仮説は `両グループの各科目の平均の差が0ではない` である。

今回は科目の平均のデータが、それぞれのグループによって異なるため、データの対応はない。そのため有意確率については、両グループの分散が等しいものと仮定して、 `(両グループの平均の差) + t x (両グループを使った標準誤差) = 0` となるようなt値が自由度:両グループの自由度の和での両側検定の値とする。ただし、両グループの分散が等しいとするには、F検定を行う必要があるが、今回は省略する。

有意水準は5%とする。つまり、有意確率が5%より低ければ、帰無仮説は却下され、各科目の平均の差があると言える。

また、有意確率の計算については、調べたい変数を抽出したあと、
`scipy.stats.ttest_ind()` メソッドによって行った。

実行結果

```
$ pipenv run python3 task2-1.py
```

```
linked hyothesis:

math score vs reading score:
 t-value : -3.7744334476190105
 p-value : 0.0004336324732439187
 -> rejected
math score vs writing score:
 t-value : -2.564351071600227
 p-value : 0.01345290292047778
 -> rejected
reading score vs writing score:
 t-value : 1.3427871915534026
 p-value : 0.1855283477979288
 -> accepted
unlinked hyothesis:

female vs male:

(math score):
 t-value : -2.301306099865485
 p-value : 0.02575930728902603
-> rejected
(reading score):
 t-value : 0.270894089323193
 p-value : 0.787634029643048
-> accepted
(writing score):
 t-value : 0.653492759874751
 p-value : 0.5165570517984135
-> accepted
```

全学生の科目平均の差については、読みと書きのスコアについては帰無仮説が採択されたが、数学と読み、数学と書きのスコアについては帰無仮説が却下されたことより、読みと書きとの間の平均の差については違いがあるとは言えないが、数学と読み、数学と書きとの間の平均の差については、t値がどちらも負であることから数学は読みや書きよりも有意に平均点が高いと言える。

一方で、男女別の科目平均の差については、数学については帰無仮説が棄却されたが、読みと書きについては帰無仮説が採択されたことより、数学についてはt値が負であったことから女子のほうが有意に高いと言えるが、読みと書きについては男女別の平均の差に意味があるとは言えない。

task2-2 : 名義データについて、関係があるかを検定する

帰無仮説は 両変数に関係がない である。また、それに対する対立仮説は 両変数に関係がある である。

今回は、標本数が50のため、期待度数が小さくなる可能性がある。そのため、イエーツの補正を用いる。

有意水準は5%とする。つまり、有意確率が5%より低ければ、帰無仮説は却下され、両変数に関係があると言える。

また、有意確率の計算については、調べたい変数を `pandas.crosstab()` メソッドを用いて、クロス表を作った後、 `scipy.stats.chi2_contingency()` で行った。

実行結果

```
$ pipenv run python3 task2-2.py
race/ethnicity & parental level of education:
chi_vale : 21.63720538720539
prob : 0.36051681812077707
->accepted
race/ethnicity & lunch:
chi_vale : 1.630273321449792
prob : 0.8033410778961504
->accepted
parental level of education & lunch:
chi_vale : 1.69160409802121
prob : 0.889955797846352
->accepted
```

いずれの変数間において、帰無仮説が採択されたことより、どの変数も他の変数と関係があるとは言えない。