レシピ57 集約関数の*argと**kwargsをカスタマイズ（P157）

- **自作の集約関数の場合**、pandasは**集約カラムをSeriesとして**1つずつ暗黙的にその関数に渡す
- その際、そのSeriesの他に**他の引数を関数に渡したいことがある**
- そのためには、Pythonで関数に任意の個数の引数を渡す機能を理解する必要がある

In [74]:
```
college = pd.read_csv('college.csv')
grouped = college.groupby(['STABBR', 'RELAFFIL'])

type(grouped)
```

Out[74]:
```
pandas.core.groupby.generic.DataFrameGroupBy
```

- groupbyオブジェクトのaggメソッドのシグネチャをinspectモジュールを使って調べる
- 引数 *args、**kwargsによって、カスタム集約関数に任意個数の非キーワード引数を渡すことが出来る

In [75]:
```
import inspect
inspect.signature(grouped.agg)
```

Out[75]:
```
<Signature (func=None, *args, engine=None, engine_kwargs=None, **kwargs)>
```

- このレシピでは、collegeデータセットで、学部学生数が2つの値の間で、州立と宗教系立かどうかによるグループ分けでの大学の割合を求める

（1）学部学生数が1000～3000である大学のパーセントを返す関数を定義する

`between`、`mean`の使い方

In [76]:
```
s = pd.Series([2, 0, 4, 8, np.nan])
print(s.between(0, 4))
print('-'*20)
print(s.between(0, 4).mean())
# 5つのうち、0≦x≦4に該当する値は、0,2,4の3つ。よって3/5=0.6（mean：平均？割合？）
```

```
0     True
1     True
2     True
3    False
4    False
dtype: bool
--------------------
0.6
```

In [77]:
```
# カスタム関数を作成
def pct_between_1_3k(s):
    return s.between(1000, 3000).mean()
```

（2）このパーセントを州立、宗教系立かどうかのグループ分けで使う

- STABBR:州
- RELAFFIL：宗教系 -> 1(true) or 0(false)
- UGDS:学部学生数

In [78]:
```
college.groupby(['STABBR', 'RELAFFIL'])['UGDS']\
        .agg(pct_between_1_3k).head(9)

# グループ分け：州、宗教系
# 集約カラム：学部学生数
# 集約関数：pct_between_1_3k（1000～3000名である大学のパーセントを返す関数）

# つまり、グループ「AK-0」において、大学は7つあり、1000~3000以上の大学は1つだけ
# つまり、下の1行でいうと、1/7の0.142857が表示されている
```

Out[78]:
```
STABBR  RELAFFIL
AK      0           0.142857
        1           0.000000
AL      0           0.236111
        1           0.333333
AR      0           0.279412
        1           0.111111
AS      0           1.000000
AZ      0           0.096774
        1           0.000000
Name: UGDS, dtype: float64
```

（3）ユーザーが人数の上限下限を設定可能な関数を作成する

In [79]:
```
def pct_between(s, low, high):
    return s.between(low, high).mean()
```

（4）（3）を利用する

In [80]:
```
college.groupby(['STABBR', 'RELAFFIL'])['UGDS']\
        .agg(pct_between, 1000, 10000).head(9)
```

Out[80]:
```
STABBR  RELAFFIL
AK      0           0.428571
        1           0.000000
AL      0           0.458333
        1           0.375000
AR      0           0.397059
        1           0.166667
AS      0           1.000000
AZ      0           0.233871
        1           0.111111
Name: UGDS, dtype: float64
```

- 明示的に、以下のようなやり方もある

In [81]:
```
college.groupby(['STABBR', 'RELAFFIL'])['UGDS']\
        .agg(pct_between, high=10000, low=1000).head()
```

Out[81]:
```
STABBR  RELAFFIL
AK      0           0.428571
        1           0.000000
AL      0           0.458333
        1           0.375000
AR      0           0.397059
Name: UGDS, dtype: float64
```
