データの読み込み

In [1]:
```
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
#  以上3つのライブラリが「3種の神器」
```

In [2]:
```
Data=pd.read_csv("D:2022_数理統計学/StatData/StatData01_2.csv") # 読み出したいファイルのパス
```

In [3]:
```
Data # 読み込んだデータを見る。データには自動的に、0番から番号がつく。
```

Out[3]:

	階級値 (c_j)	度 数 (f_j)
0	22	1
1	24	4
2	26	5
3	28	15
4	30	18
5	32	20
6	34	18
7	36	11
8	38	4
9	40	2
10	42	2

In [4]:
```
# 日本語は避ける方がよい
# ついでに簡単な変数名 (x,f) にしてしまう
Data = Data.rename(columns={'階級値 (c_j )' : 'x'}) # Data= とすることで上書きされる.
Data = Data.rename(columns={'度 数 (f_j )' : 'f'})  # Data= とすることで上書きされる.
Data.head()
```

Out[4]:

	x	f
0	22	1
1	24	4
2	26	5
3	28	15
4	30	18

In [5]:
```
# ファイルを読み込む段階でカラム名を変更しておくことも可能
Data=pd.read_csv("D:2022_数理統計学/StatData/StatData01_2.csv", # 読み出したいファイルのパス
    skiprows=1,          # データファイルの最初の1行を飛ばす
    names=['x', 'f'])    # カラム名を付ける
Data.head()
```

Out[5]:

	x	f
0	22	1
1	24	4
2	26	5
3	28	15
4	30	18

度数分布表で統計量を計算する

In [6]:
```
Data['xf']=Data['x']*Data['f']
Data['x^2f']=Data['x']**2*Data['f']
Data
```

Out[6]:

	x	f	xf	x^2f
0	22	1	22	484
1	24	4	96	2304
2	26	5	130	3380
3	28	15	420	11760
4	30	18	540	16200
5	32	20	640	20480
6	34	18	612	20808
7	36	11	396	14256
8	38	4	152	5776
9	40	2	80	3200
10	42	2	84	3528

In [7]:
```
# 各カラムの総和
Data.sum()
```

Out[7]:
```
x          352
f          100
xf        3172
x^2f    102176
dtype: int64
```

In [8]:
```
size=Data['f'].sum()  # カラム Frequency (f) の総和の計算
sum_xf=Data['xf'].sum()
sum_xxf=Data['x^2f'].sum()
size, sum_xf, sum_xxf
```

Out[8]:
```
(100, 3172, 102176)
```

In [9]:
```
mean=sum_xf/size  # 平均値の計算
mean
```

Out[9]: 31.72

In [10]:
```
var=sum_xxf/size-mean**2   # 分散公式による計算
var
```

Out[10]: 15.601600000000076

In [11]:
```
std=np.sqrt(var)  # 分散の平方根が標準偏差
std
```

Out[11]: 3.949886074306457

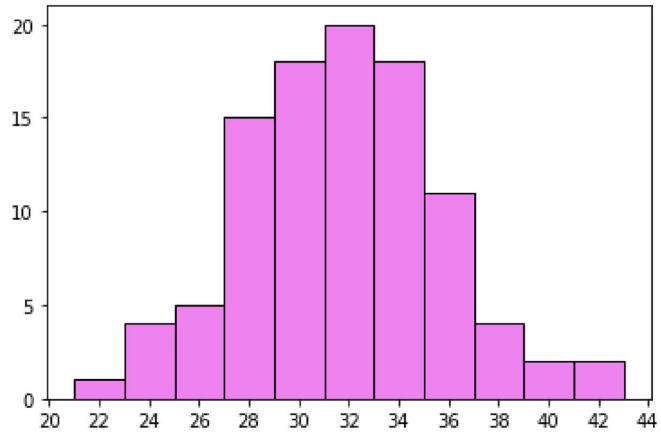
In [12]:
```
# まとめ
StatSummary=pd.DataFrame([
    ['平均値', mean],
    ['分散', np.round(var,2)],
    ['標準偏差', np.round(std,2)],
    ['データ数', size]
])
StatSummary=StatSummary.rename(columns={0:'統計量'})
StatSummary=StatSummary.rename(columns={1:'値'})
StatSummary
```

Out[12]:

	統計量	値
0	平均値	31.72
1	分散	15.60
2	標準偏差	3.95
3	データ数	100.00

ヒストグラム

In [13]:
```
# ヒストグラム
# plt.figure(figsize=(10,5))
plt.bar(Data['x'], Data['f'],      # 棒グラフ（横軸と縦軸の変数を指定）
        width=2,                   # 棒の幅
        color='violet',
        ec='k')
plt.xticks(np.arange(20,46,2))
plt.yticks(np.arange(0,25,5))
plt.show()  # 付帯データを表示せず、グラフのみ表示
```



In [ ]: