HW07 (제출기한-5월24일자정)

Datascience 모듈의 Table 연습을 위한 문제입니다.

- 반드시 수업시간 배운 내용만을 사용하여 코드를 작성합니다.
- 한 셀에 여러줄의 코드를 사용해도 무방합니다.

제출시 다음 사항에 유의하기 바랍니다.

- 텍스트 셀에 설명되어 있는 각 문항을 잘 읽은 뒤, '답안을 작성하시오'라고 적힌 코드 셀에 적절한 코드를 작성합니다.
- '지우지 마시오'라고 적힌 코드 셀은 절대로 지우면 안 됩니다.
- 작성 후 파일명에 `학번`을 자신의 학번으로 고친 후 ULMS 해당과제 제출란에 업로드합니다.

```
from datascience import *
import matplotlib
%matplotlib inline
import matplotlib.pyplot as plots
plots.style.use('fivethirtyeight')
import numpy as np
```

```
scores = Table.read_table('https://raw.githubusercontent.com/data-8/materials-sp18/master/lec/sc
scores
```

문제1

Section별 median을 구하여 `section_median` Table 변수에 저장하시오.

```
### 답안을 작성하시오.
section_median = scores.group('Section',np.median)

```

```
### 지우지 마시오.
section_median
```

문제2

4번 Section의 학생수를 `no_student` 변수에 저장하시오.

```
### 답안을 작성하시오.
no_student = scores.group('Section').where('Section',4).column('count')[0]

```

```
### 지우지 마시오.
no_student
```

```
30
```

문제3

전체 학생중 4번 Section 학생 수 만큼 랜덤하게 sampling 후 Midterm 점수의 median을 계산하는 시행을 10000번 반복하고 그 분포를 이후에 알아보고자 한다. 이를 위해 table이름, table내 label 이름, sample의 수, repetition횟수를 인자로 받아 시뮬레이션을 하여 얻은 median들의 결과 array를 리턴하는 함수 `sample_median`를 작성하시오. (permuted_sample_average_difference 함수 내용을 참고할 것)

(단, `replacement`는 False로 설정한다.)

```
def sample_median(table, label, no_sample, repetitions):
### 답안을 작성하시오.

  tmp=np.empty(10000)
  for i in np.arange(10000):
    tmp[i]=np.median(scores.sample(no_student, with_replacement=False).column('Midterm'))
  return tmp
```

```
### 지우지 마시오.
results = sample_median(scores, 'Midterm', no_student, 10000)
results
```

```
array([17. , 18. , 19. , ..., 15.5, 16.5, 17.5])
```

문제4

위 `results` 배열에 대해 historgram을 그려보시오.

또한, 관측값(문제 1에서 조사한 4번 section의 median)을 Ch12의 예제처럼 붉은색 원(크기는 100)으로 표시하시오.

```
### 답안을 작성하시오.
Table().with_column('Sample Medain', results).hist()
plots.scatter('section_median', 0, color='red', s=100);

```

문제5

위에서 구한 `results`에 대하여 관측값(4번 Section의 median)의 `p-value`를 계산하시오.

```
### 답안을 작성하시오.
np.sum(results <=section_median.where('Section',4).column("Midterm median")[0])/results.size

```

문제6

`scores` Table에서 Section이 4 또는 5인 경우만을 선택하여 `scores_two` Table 변수에 저장하시오.

```
### 답안을 작성하시오.
scores_two=scores.where('Section',are.between(4, 6))

```

```
### 지우지 마시오.
scores_two
```

문제7

5번 Section의 median에서 4번 Section의 median을 뺀 값을 `observed` 변수에 저장하시오.

```
### 답안을 작성하시오.

observed=scores_two.group('Section',np.median).column("Midterm median")[1]-scores_two.group('Sec


```

```
### 지우지 마시오.
observed
```

문제8

위 `scores_two` Table에 대하여 shuffling 한 후 두 그룹 (5번과 4번 Section) 사이의 median 차이를 구하는 시행을 10000번 반복하여 그 결과를 array에 저장하려고 한다. 이를 위해 Ch12 에서 배운 `permuted_sample_average_difference`을 적절하게 변경하여 `permuted_sample_median_difference`를 작성하시오.

```
def permuted_sample_median_difference(table, label, group_label, repetitions):
    ### 답안을 작성하시오.
  tmp=np.empty(10000)
  for i in np.arange(10000):
    shuffling_scores_two=scores_two.with_column('test',scores_two.sample(with_replacement = Fals
    tmp[i]=np.median(shuffling_scores_two.where('Section',5).column('test'))-np.median(shuffling


```

```
    return tmp
```

```
### 지우지 마시오.
shuffle_results = permuted_sample_median_difference(scores_two, 'Midterm', 'Section', 10000)
shuffle_results

```

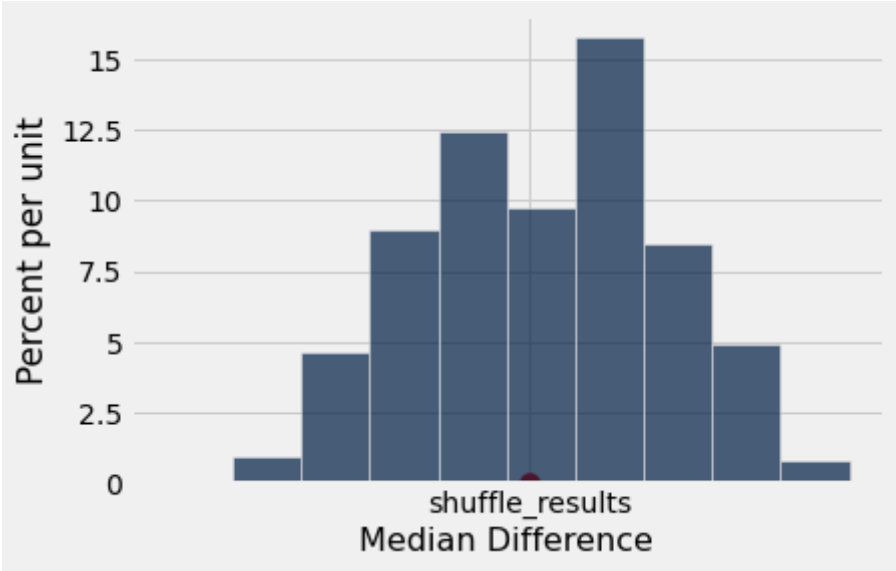
문제9

위 `shuffled_results` 배열에 대해 historgram을 그려보시오.

또한, 관측값(즉 문제7에서 구한 값)을 Ch12의 예제처럼 붉은색 원(크기는 100)으로 표시하시오.

```
### 답안을 작성하시오.
Table().with_column("Median Difference",shuffle_results).hist()
plots.scatter('shuffle_results', 0, color='red', s=100)
```

```
<matplotlib.collections.PathCollection at 0x7fe34fe37ed0>
```



문제10

위에서 구한 `shuffle_results`에 대하여 관측값(`observed` 변수값)의 `p-value`를 계산하시오.

```
### 답안을 작성하시오.
np.sum(shuffle_results >=observed)/shuffle_results.size
```