빅데이터 분석
-통계분석3-

강의자료 출처:

한밭대학교 컴퓨터공학
임경태

추측 통계에 대한 이해

- **모집단** 통계: 전교 학생(1000명) 의 영어 점수
- **표본** 통계: 표본 집단(100명) 의 영어 점수

- 일부 데이터를 수집해 전체 데이터를 추측할 수 있을까?
  - 일부 데이터로 구한 대표값 평균, 중간 값, 편차, 분산, 표준편차는 어떤 의미가 있나?

- 일부 관측 데이터와 전체 모집단의 대표값들은 어떤 상관관계가 있을까?

- 우리학교 100명의 수능 영어점수 분포를 이용해 전체 수능 점수의 분포와 백분율을 예측할 수 있을까? **(예측) 실제 얼마나 정확할까? (평가)**

CHAPTER

04

추측 통계의 기본

CONTENTS

어느 고등학교에서 전교생 400명이 수학 시험을 동일하게 치렀습니다. 3학년인 A 학생은 이 시험에서 80점을 받았지만, 학교에서 전교생의 평균 점수를 알려주지 않았기 때문에 A 학생은 자신이 전교생 중 어느 정도의 수준인지 알지 못합니다. 자신의 성적이 좋은지 나쁜지가 궁금한 A 학생은 스스로 전교생의 평균 점수를 구해보려 했지만, 400명 전원의 시험 결과를 수소문하는 것은 무리입니다. 그래서 A 학생은 학교 안에서 우연히 만난 20명에게 시험 점수를 물어보고, 그 결과로부터 전교생의 평균 점수를 추측하기로 했습니다.

20명의 시험 점수 평균은 70.4점이었습니다. A 학생은 전교생의 평균도 그 정도일 것으로 생각하고, 자신의 점수가 평균보다 위에 있다는 것에 만족했습니다.

추측 통계: 일부 데이터로부터 전체의 통계적 설질을 추측

In [1] :

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

%precision 3
%matplotlib inline
```

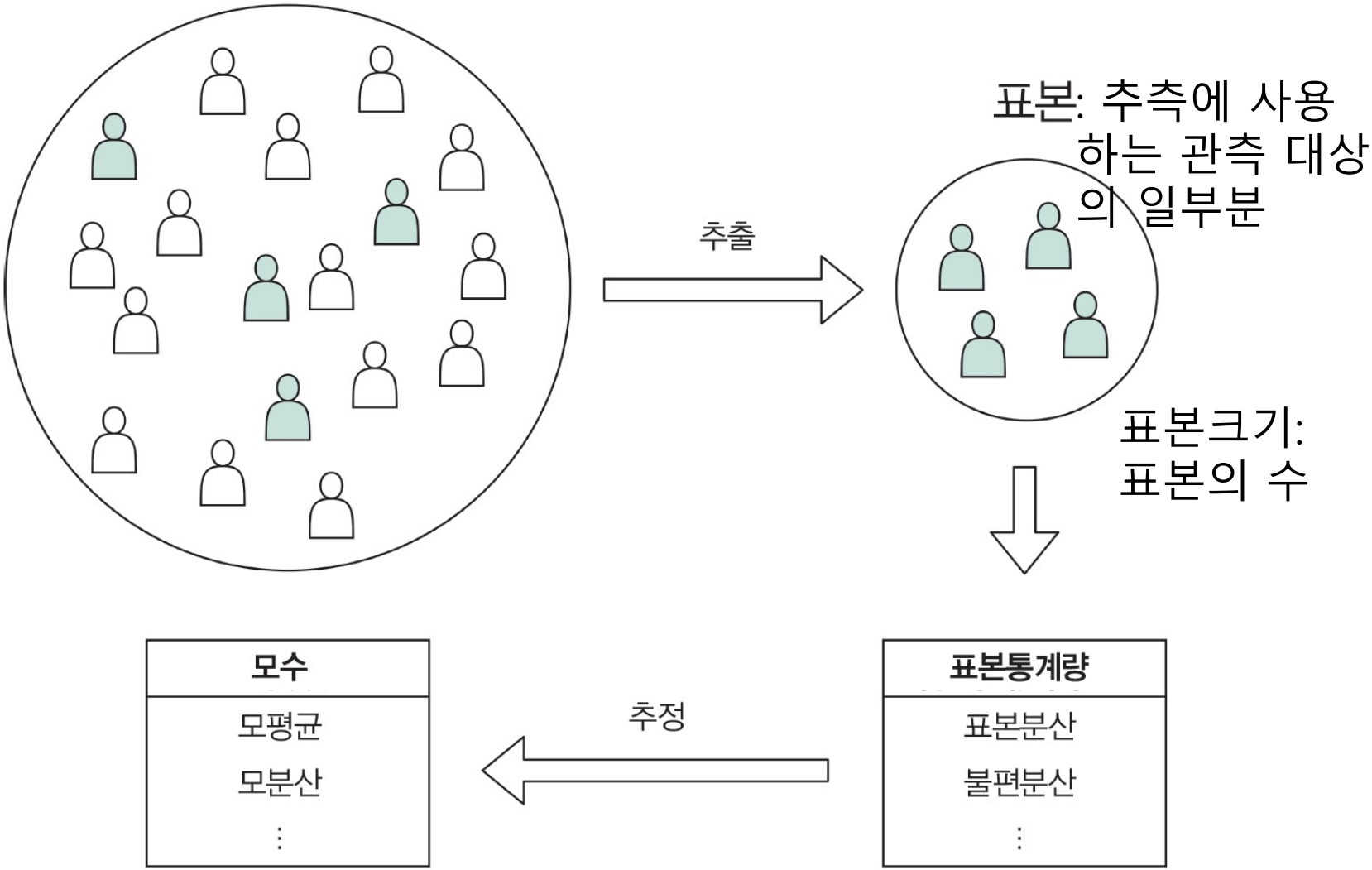
In [2] :

```
df = pd.read_csv( ' ../data/ch4_scores400.csv ' )
scores = np.array(df[ ' score ' ])
scores[:10]
```

Out [2] :

```
array([76, 55, 80, 80, 74, 61, 81, 76, 23, 80])
```

모집단: 추측하고 싶은 관측 대상 전체



[그림 4-1] 모집단과 표본

4.1.1 표본추출 방법

- 무작위 추출(임의 추출): 임의로 표본을 추출하는 방법
- 복원추출: 여러 차례 동일한 표본을 선택하는 방법

In [3] :

```
np.random.choice([1, 2, 3], 3)
```

Out [3] :

```
array([1, 2, 2])
```

- 비복원추출: 동일한 표본은 한 번만 선택하는 방법

In [4] :

```
np.random.choice([1, 2, 3], 3, replace=False)
```

Out [4] :

```
array([3, 1, 2])
```

4.1.1 표본추출 방법

- 시드를 0으로 하는 무작위 추출(임의 추출)은 매번 동일한 결과

In [5] :

```
np.random.seed(0)
np.random.choice([1, 2, 3], 3)
```

- 표본크기 20으로 복원추출, 표본 평균 계산

In [6] :

```
np.random.seed(0)
sample = np.random.choice(scores, 20)
sample.mean()
```

Out [6] :

```
70.400
```

- 모평균은 69.530(score.mean())이므로 꽤 괜찮은 추측

4.1.1 표본추출 방법

- 무작위 추측은 실행할 때마다 결과가 달라지므로, 표본평균도 매번 달라짐

In [8] :

```
for i in range(5):
    sample = np.random.choice(scores, 20)
    print(f '{i+1}번째 무작위추출로 얻은 표본평균', sample.mean())
```

Out [8] :

```
1번째 무작위추출로 얻은 표본평균 72.45
2번째 무작위추출로 얻은 표본평균 63.7
3번째 무작위추출로 얻은 표본평균 66.05
4번째 무작위추출로 얻은 표본평균 71.7
5번째 무작위추출로 얻은 표본평균 74.15
```

4.2.1 확률의 기본

- 확률 : 무작위 추출과 같은 불확정성을 수반한 현상을 해석
- 확률 모형 : 무작위 추출 혹은 주사위를 모델링
- 확률변수 : 결과를 알아맞힐 수는 없지만, 취하는 값과 그 값이 나올 확률이 결정되어 있는 것
- 확률분포: 확률모형에 의해 결정된 확률변수가 나타내는 분포
- 시행 : 확률변수의 결과를 관측하는 것
- 실현값 : 시행에 의해 관측되는 값

4.2.1 확률의 기본

- 사건 : 시행 결과로 나타날 수 있는 값(눈이 1, 눈이 홀수)
  - 주사위의 눈은 확률 변수 X
  - 눈이 1이 되는 사건의 확률 $P(X=1)=\frac{1}{6}$
  - 눈이 홀수인 사건의 확률

$$P((X=1)\cup(X=3)\cup(X=5)) = P(X=1)+P(X=3)+P(X=5)$$
$$=\frac{1}{6}+\frac{1}{6}+\frac{1}{6}$$
$$=\frac{1}{2}$$

- 근원사건 : 세부적으로 더 분해될 수 없는 사건(눈이 1)
- 상호배반 : 동시에 일어날 수 없는 사건
  - '눈이 1 또는 2 또는 3'이라는 사건과 '눈이 6'이라는 사건

4.2.2 확률분포

- 확률변수가 어떻게 움직이는지를 나타낸 것
- 공정한 주사위

[표 4-1] 주사위의 확률분포

눈	1	2	3	4	5	6
확률	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

- 불공정한 주사위

[표 4-2] 불공정한 주사위의 확률분포

눈	1	2	3	4	5	6
확률	$\frac{1}{21}$	$\frac{2}{21}$	$\frac{3}{21}$	$\frac{4}{21}$	$\frac{5}{21}$	$\frac{6}{21}$

4.2.2 확률분포

- 파이썬으로 불공정한 주사위의 확률분포를 구하는 실험

In [9] :

```
dice = [1, 2, 3, 4, 5, 6]
prob = [1/21, 2/21, 3/21, 4/21, 5/21, 6/21]
```

In [11] :

```
num_trial = 100
sample = np.random.choice(dice, num_trial, p=prob)
sample
```

Out [11] :

```
array([4, 6, 4, 5, 5, 6, 6, 3, 5, 6, 5, 6, 6, 2, 3, 1, 6, 5, 6, 3,
       4, 5, 3, 4, 3, 5, 5, 4, 4, 6, 4, 6, 5, 6, 5, 4, 6, 2, 6, 4,
       5, 3, 4, 6, 5, 5, 5, 3, 4, 5, 4, 4, 6, 4, 4, 6, 6, 2, 2, 4,
       5, 1, 6, 4, 3, 2, 2, 6, 3, 5, 4, 2, 4, 4, 6, 6, 1, 5, 3, 6,
       6, 4, 2, 1, 6, 4, 4, 2, 4, 1, 3, 6, 6, 6, 4, 5, 4, 3, 3, 4])
```

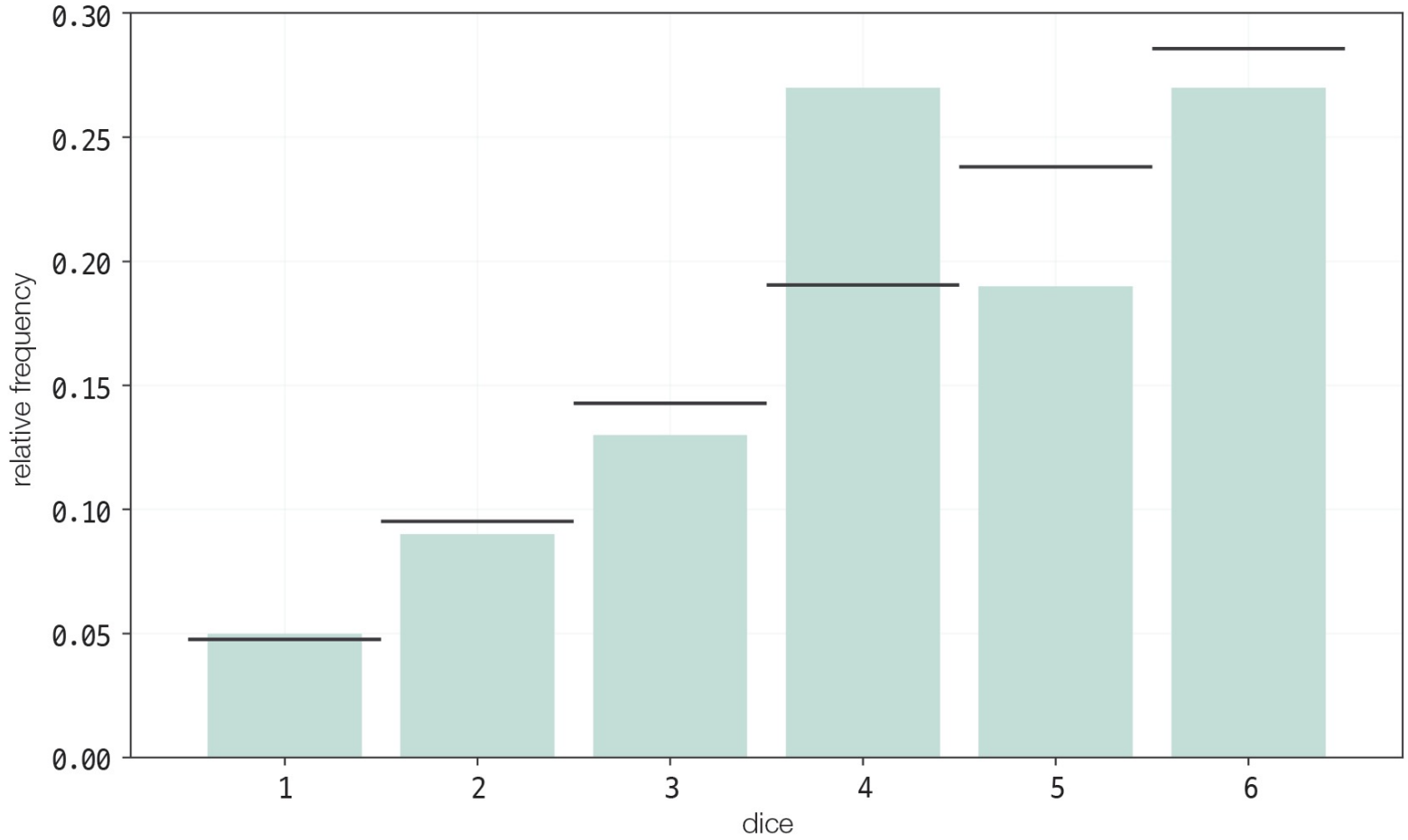
4.2.2 확률분포

- 도수분포표와 히스토그램

dice	frequency	relative frequency
1	5	0.05
2	9	0.09
3	13	0.13
4	27	0.27
5	19	0.19
6	27	0.27

4.2.2 확률분포

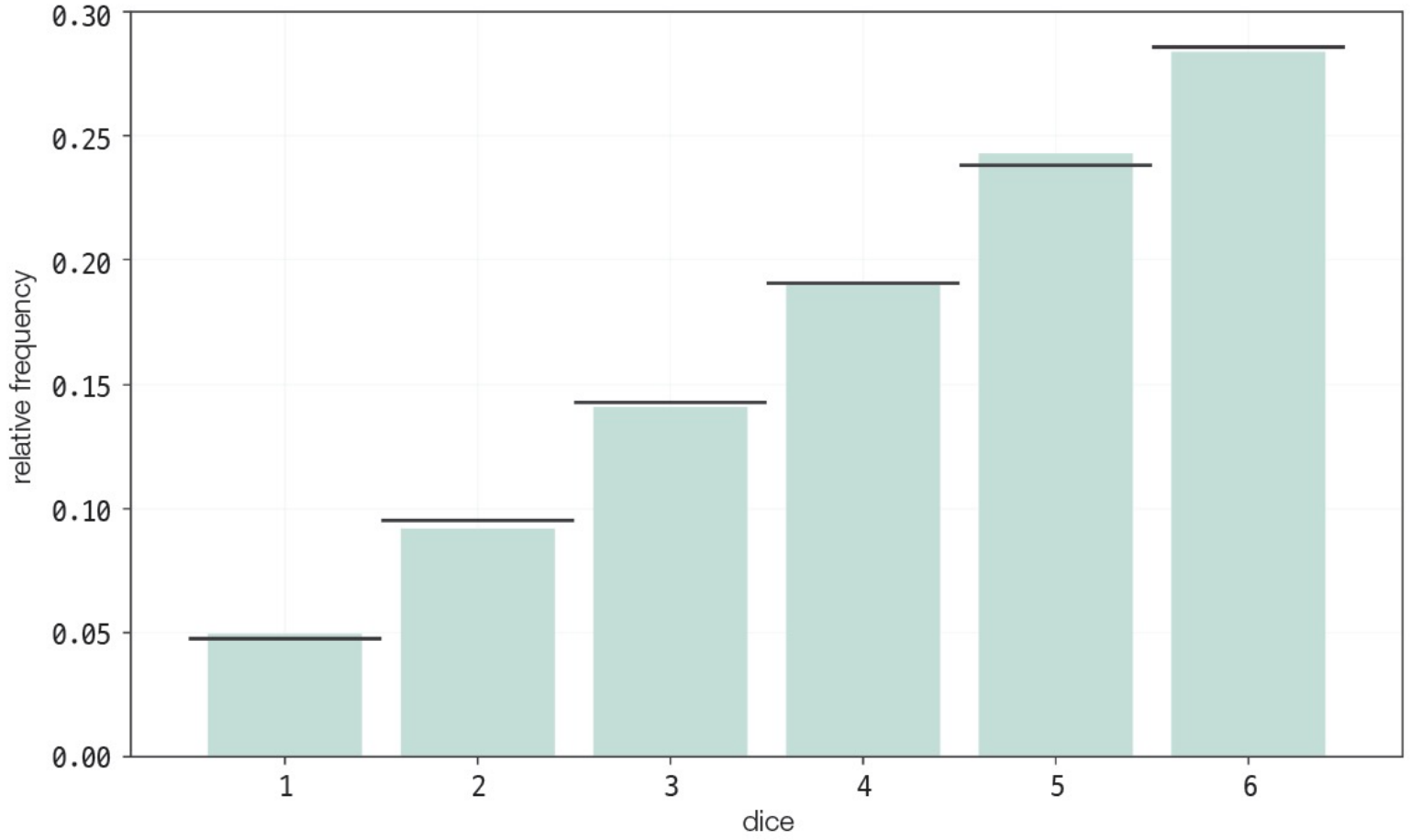
- 도수분포표와 히스토그램



[그림 4-2] 100번 시행했을 때 주사위 눈에 대한 히스토그램

4.2.2 확률분포

- 10000번 시행했을 때의 히스토그램은 실제의 확률분포에 가까워짐



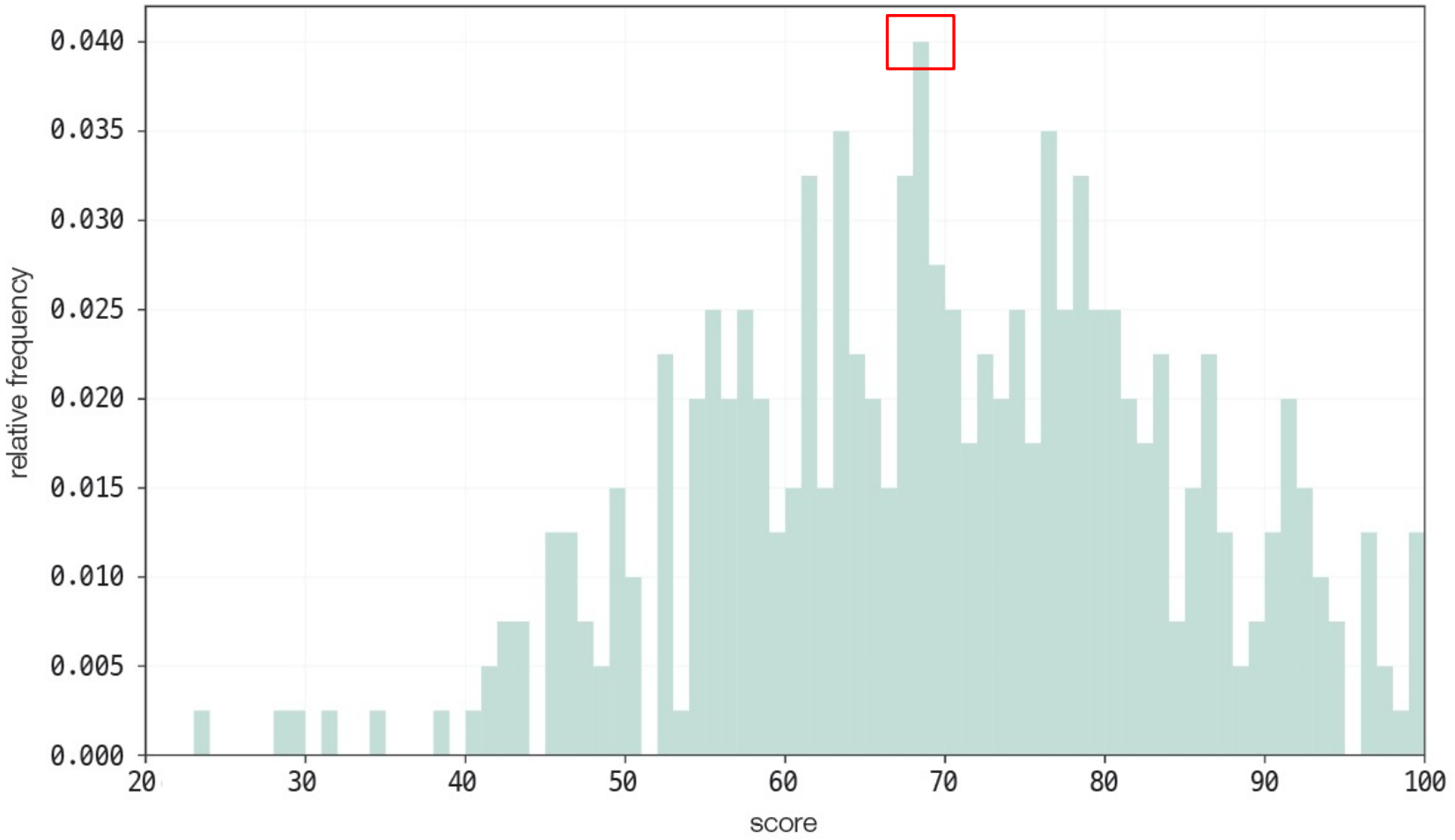
[그림 4-3] 10000번 시행했을 때 주사위 눈의 히스토그램 SAMPLE CODE

계급폭을 1점으로 하는 히스토그램

In [15] :

```
fig = plt.figure(figsize=(10, 6))
ax = fig.add_subplot(111)
ax.hist(scores, bins=100, range=(0, 100), density=True)
ax.set_xlim(20, 100)
ax.set_ylim(0, 0.042)
ax.set_xlabel( ' score ' )
ax.set_ylabel( ' relative frequency ' )
plt.show()
```

69점을 얻은 학생은 전교생의 0.04(4%)이므로 무작위추출을 수행하면 4%의 확률로 69점이라는 표본 데이터 획득



[그림 4-4] 전교생 시험 점수에 대한 히스토그램

무작위추출은 확률분포를 따르는 확률변수의 시행

In [16] :

```
np.random.choice(scores)
```

Out [16] :

```
89
```

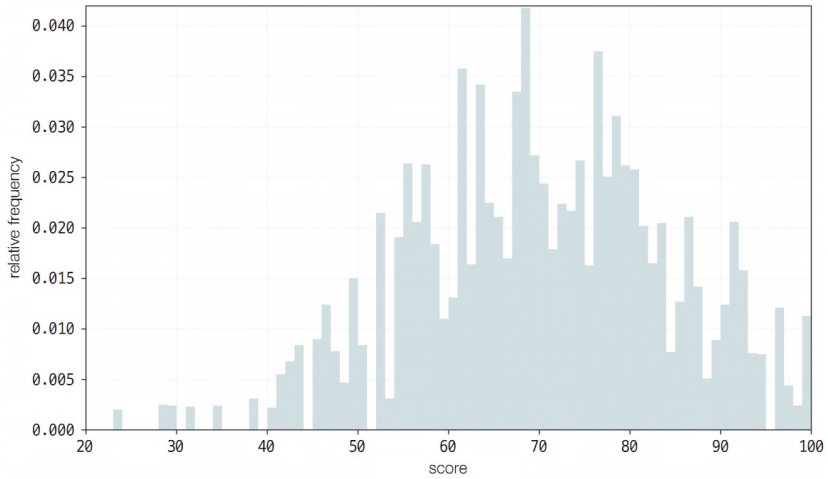
무작위추출로 얻은 표본 데이터가 89점

- 시행 횟수를 늘리면 주사위의 상대도수는 실제의 확률분포에 가까워짐
- 무작위추출에서도 표본의 크기가 커지면, 표본 데이터의 상대도수는 실제의 확률분포에 근사
- 무작위추출로 샘플 사이즈가 10000인 표본 추출

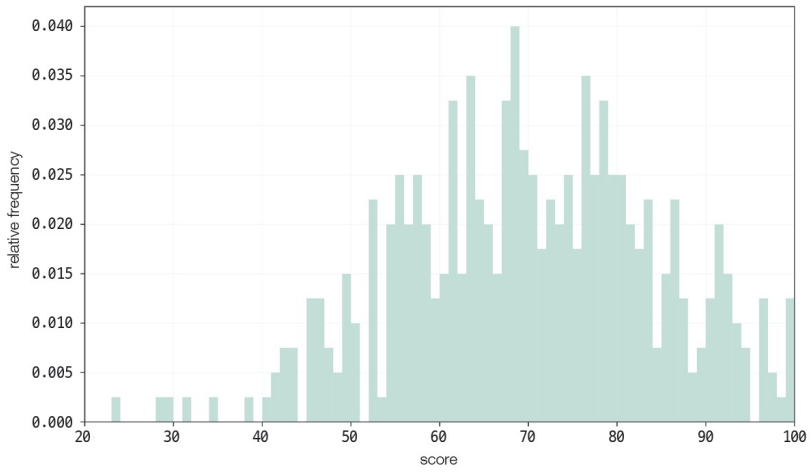
In [17] :

```
sample = np.random.choice(scores, 10000)

fig = plt.figure(figsize=(10, 6))
ax = fig.add_subplot(111)
ax.hist(sample, bins=100, range=(0, 100), density=True)
ax.set_xlim(20, 100)
ax.set_ylim(0, 0.042)
ax.set_xlabel( ' score ' )
ax.set_ylabel( ' relative frequency ' )
plt.show()
```



[그림 4-5] 무작위추출로 얻은 표본 데이터의 히스토그램 SAMPLE CODE



[그림 4-4] 전교생 시험 점수에 대한 히스토그램

- 히스토그램이 실제의 점수 분포에 가까운 형태
- 표본 크기가 커지면 실제의 분포에 수렴

표본크기가 20인 표본을 추출하여 표본평균을 계산하는 작업을 10000번 수행

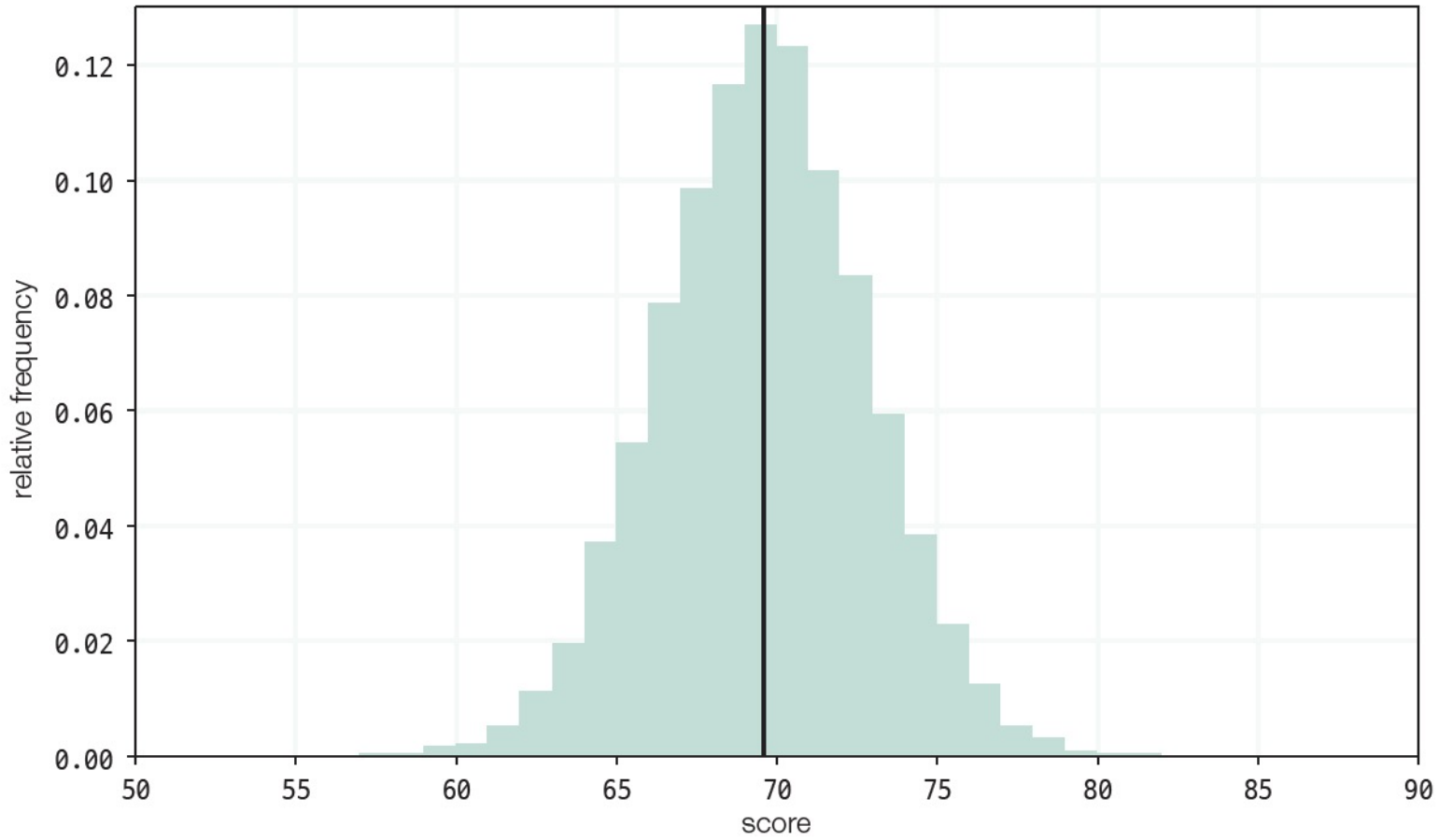
In [18] :

```
sample_means = [np.random.choice(scores, 20).mean()
                for _ in range(10000)]

fig = plt.figure(figsize=(10, 6))
ax = fig.add_subplot(111)
ax.hist(sample_means, bins=100, range=(0, 100), density=True)
# 모평균을 세로선으로 표시
ax.vlines(np.mean(scores), 0, 1, 'gray')
ax.set_xlim(50, 90)
ax.set_ylim(0, 0.13)
ax.set_xlabel('score')
ax.set_ylabel('relative frequency')
plt.show()
```

표본평균은 모평균을 중심으로 분포

=> **무작위추출에 의한 표본평균으로 모평균 추측 가능**



[그림 4-6] 표본평균의 분포 SAMPLE CODE

Q&A