제 13 장 정규분포곡선과 확률히스토그램

1. 이항분포와 정규분포
2. 합의 경험적 히스토그램과 확률 히스토그램
3. 경험적 히스토그램과 확률히스토그램
4. 확률히스토그램과 정규분포곡선
5. 정규분포곡선으로의 근사
6. 정규분포곡선의 사용 범위
7. 중심극한정리와 히스토그램의 수렴
8. 부트스트래핑
9. 컴퓨터 시뮬레이션

1. 이항분포와 정규분포

동전 던지기: 이항분포와 정규분포

동전 던지기를 반복할 때 앞면이 나오는 비율은 점차 50%에 가까워진다.

동전을 다섯 번 던지는 경우, 앞면이 나오는 횟수~B(5, ½)의 이항분포 따름

앞면이 나오는 회수	경우의 수
0	1
1	5
2	10
3	10
4	5
5	1

동전을 백 번 던지면?

1. 이항분포와 정규분포

동전을 백 번 던져 앞면이 정확히 50번 나올 확률

1. 전체 경우의 수

$$2^{100} \approx 1.27 \times 10^{30}$$

2. 앞면이 정확히 50번 나오는 경우의 수

$$\frac{100!}{50! \times 50!} = \frac{100 \times 99 \times \cdots \times 51}{50 \times 49 \times \cdots \times 1} \approx 1.01 \times 10^{29}$$

3. 앞면이 정확히 50번 나올 확률

$$\frac{\text{앞면이 50번 나오는 경우의 수}}{\text{전체 경우의 수}} \approx \frac{1.01 \times 10^{29}}{1.27 \times 10^{30}} \approx 0.08 = 8\%$$

1. 이항분포와 정규분포

동전을 백 번 던져 앞면이 정확히 50번 나올 확률

위와 같은 방식으로 동전을 100번 던져 앞면이 정확히 50번 나올 확률을 계산하기는 불편하다.

대안은? 동전을 많이 던지는 경우 전체 시행횟수 가운데 앞면이 나오는 횟수 또는 앞면이 나오는 비율은 그 분포가 정규분포에 의해 잘 근사된다는 사실을 이용한다.

정규분포곡선을 이용하여 이 문제에 답하면 답은 7.96%가 나온다. 앞에서 이항분포를 이용해 구한 정답 8%와 거의 같다.

2. 합의 경험적 히스토그램과 확률히스토그램

합의 경험적 히스토그램과 확률히스토그램

합의 경험적 히스토그램

여러 차례 반복적으로 관측한 합의 자료를 구간별로 분류하고 구간별 도수를 계산한 뒤 도수를 밀도단위로 바꾸어 밀도단위 히스토그램으로 나타낸 것

합의 확률히스토그램

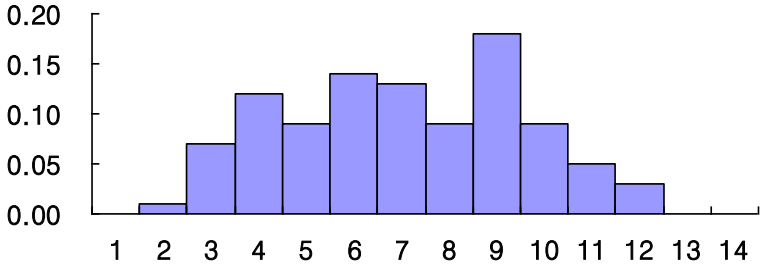
상자의 내용물 및 추출횟수로부터 합이 각각의 값으로 실현될 확률을 계산하여 이를 그래프로 나타낸 것

3. 경험적 히스토그램과 확률히스토그램

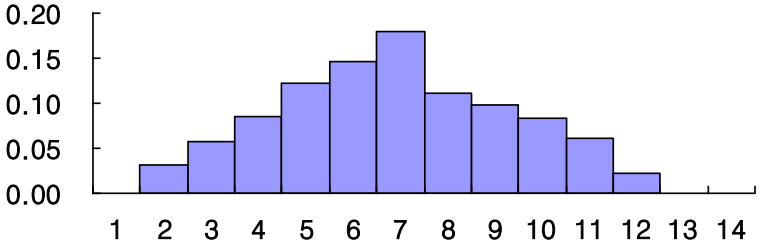
경험적 히스토그램과 확률히스토그램

경험적 히스토그램: 두 개(n=2)의 주사위를 던져 그 합을 기록하는 시행을 백 번, 천 번, 만 번 반복(R=100 → 1,000 → 10,000)

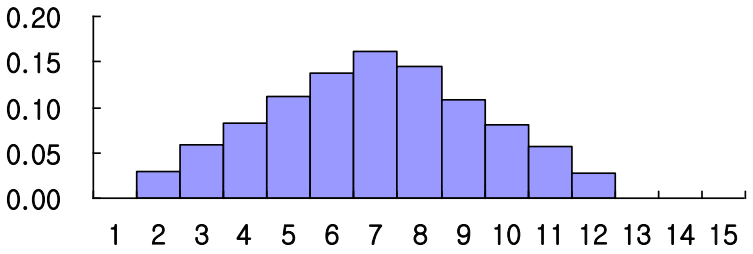
(a) 시행을 100번 반복(R=100)



(b) 시행을 1,000번 반복(R=1,000)



(c) 시행을 10,000번 반복(R=10,000)

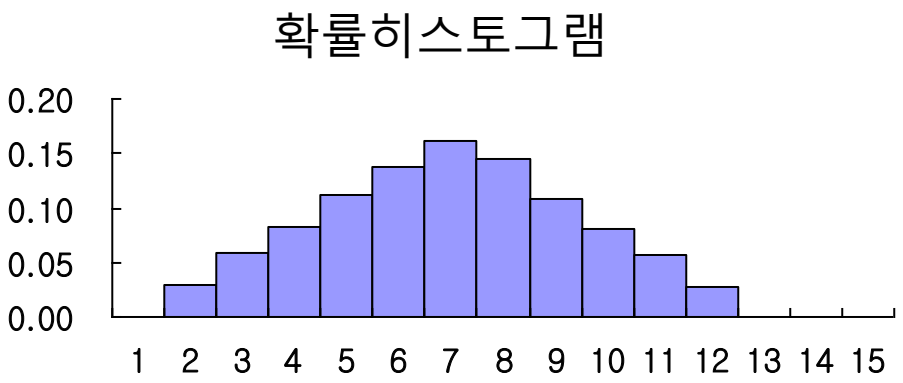


3. 경험적 히스토그램과 확률히스토그램

합의 확률히스토그램

두 개의 주사위를 던졌을 때 나타나는 모든 경우의 수를 고려하여 머리를 써서 계산해 낼 수 있다.

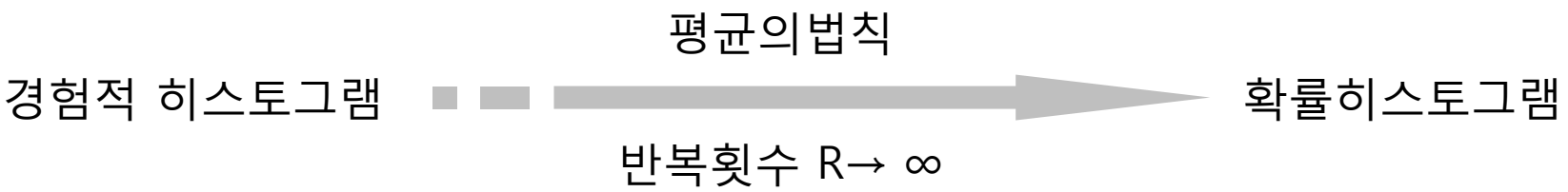
두 개의 주사위를 던져 그 합을 기록하는 시행을 무한히 반복하여 합에 대해 얻어낸 수많은 경험적 자료를 가지고 그린 합의 히스토그램



3. 경험적 히스토그램과 확률히스토그램

경험적 히스토그램과 확률히스토그램

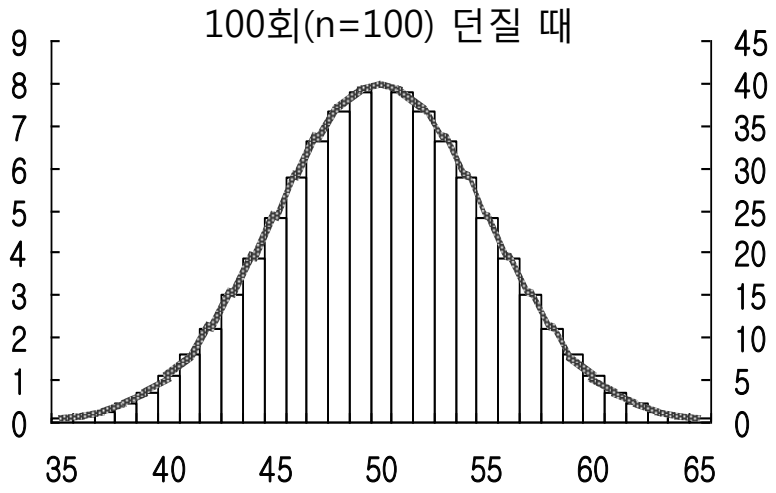
두 개 주사위를 던지는 실험을 반복하여 합에 대해 얻어낸 경험적 자료를 가지고 합의 히스토그램을 그리면 반복하는 횟수가 증가함에 따라 경험적 히스토그램은 확률 히스토그램으로 수렴하게 된다.

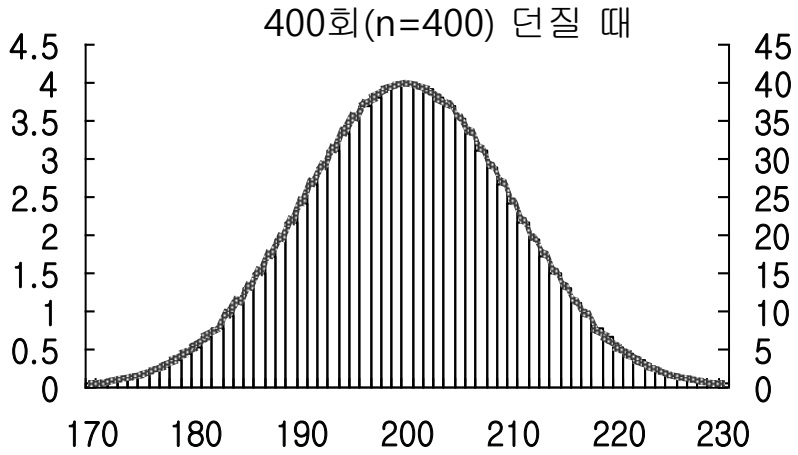


4. 확률히스토그램과 정규분포곡선

확률히스토그램

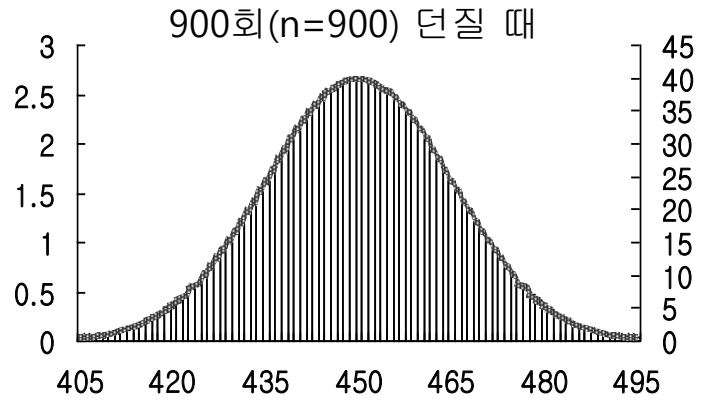
동전을 던질 때 앞면이 나오는 횟수의 확률히스토그램





4. 확률히스토그램과 정규분포곡선

확률히스토그램의 수렴



무한히 많이 던지면?

(n→ ∞의 경우)

중심극한정리(CLT: Central Limit Theorem)

- 관측치수(n)가 증가함에 따라 합이나 평균은 그 확률히스토그램이 정규분포곡선으로 수렴해 간다. 이를 **중심극한정리**라고 한다.

5. 정규분포곡선으로의 근사

정규분포곡선으로의 근사

동전을 100번(n=100) 던진다. 다음을 계산하라.

a) 앞면이 정확히 50번 나올 확률

b) 앞면이 45이상 55이하로 나올 확률

c) 앞면이 45초과 55미만으로 나올 확률

5. 정규분포곡선으로의 근사

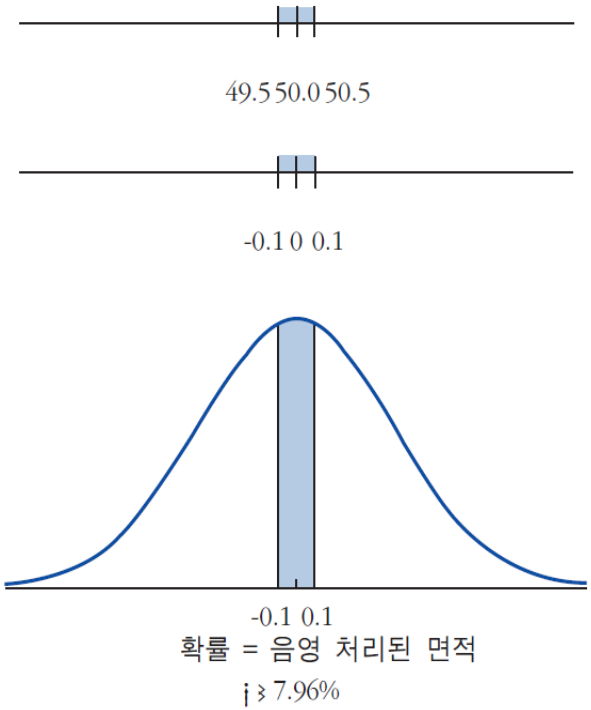
정규분포곡선으로의 근사

a) 앞면이 정확히 50번 나올 확률

- 앞면 횟수는 기대값이 50, 표준오차가 5
- 앞면이 50번 나오는 사건은 이를 연속인 확률변수로 표현할 때 [49.5, 50.5)의 구간에 해당됨. 표준화하면 [-0.01, 0.01) 구간임

$$\frac{49.5-50}{50}=-0.01, \quad \frac{50.5-50}{50}=0.01$$

- 표준정규분포 곡선을 이용하여 [-0.01, 0.01) 구간의 면적을 구하면 7.96% 얻음. 이는 앞에서 구한 8%와 별 차이가 없음

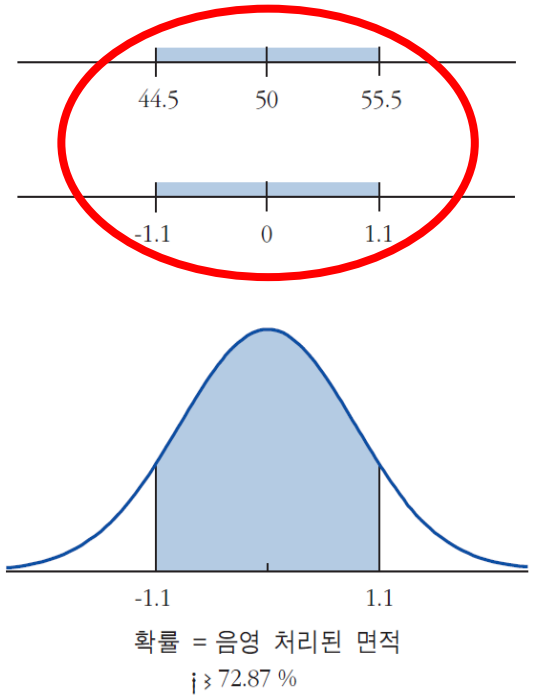


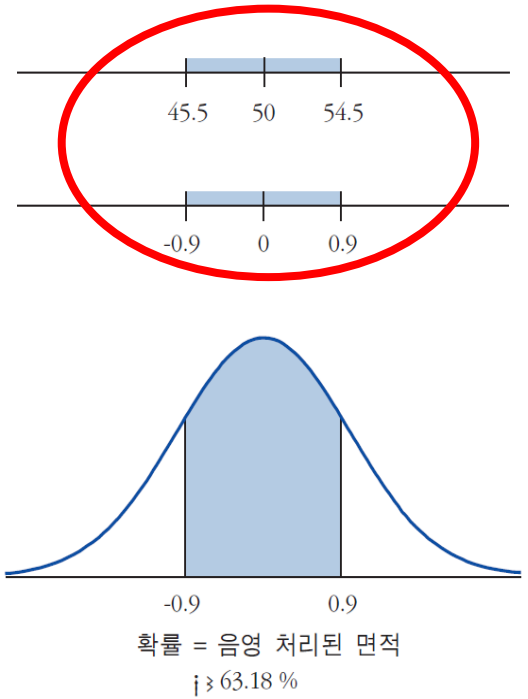
5. 정규분포곡선으로의 근사

정규분포곡선으로의 근사

b) 앞면이 45이상 55이하로 나올 확률

c) 앞면이 45초과 55미만으로 나올 확률





6. 정규분포곡선의 사용 범위

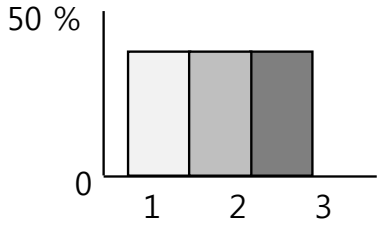
정규분포곡선으로의 근사

동전 던지기와는 달리 내용물의 분포가 대칭이 아닌 상자로부터 숫자를 추출하더라도, 추출횟수만 충분히 크면 추출된 숫자의 합은 상자의 내용물 분포에 관계없이 정규분포곡선으로 잘 근사 된다.

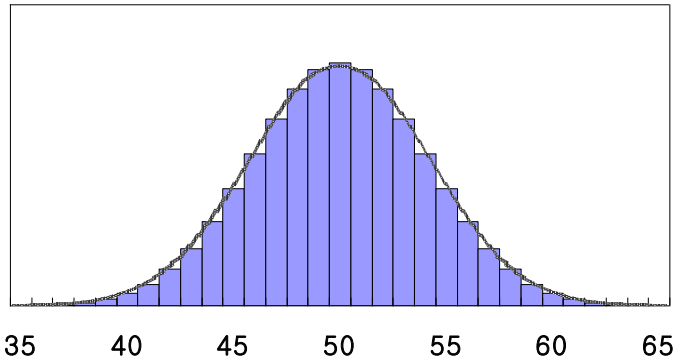
다만, 상자의 내용물 분포에 따라 합의 확률히스토그램이 정규분포곡선으로 닮아가는 속도는 다르다. 상자의 내용물 분포가 정규분포곡선과 비슷하면 비슷할수록 합의 확률히스토그램이 정규분포곡선으로 닮아가는 속도는 빨라진다.

6. 정규분포곡선의 사용 범위

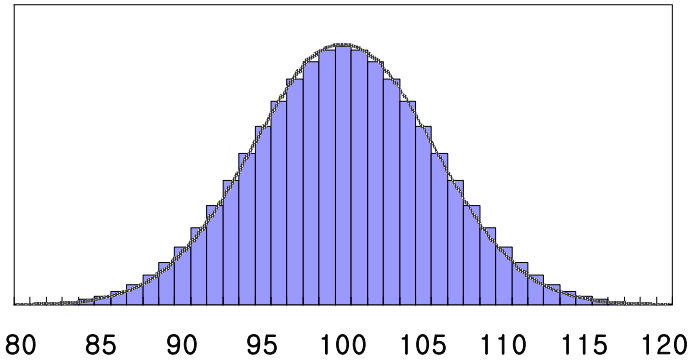
상자의 내용물이 대칭인 경우 합의 확률히스토그램



상자의 확률히스토그램



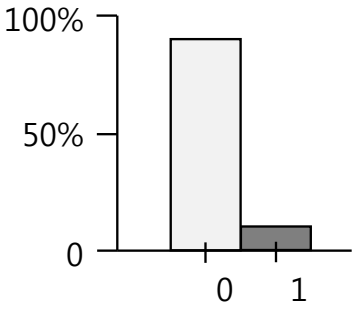
(i) 25회 추출



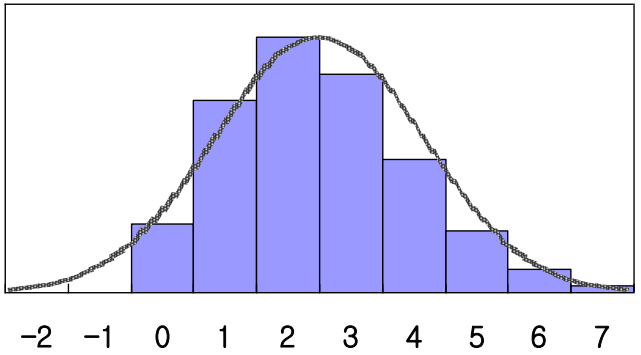
(ii) 50회 추출

6. 정규분포곡선의 사용 범위

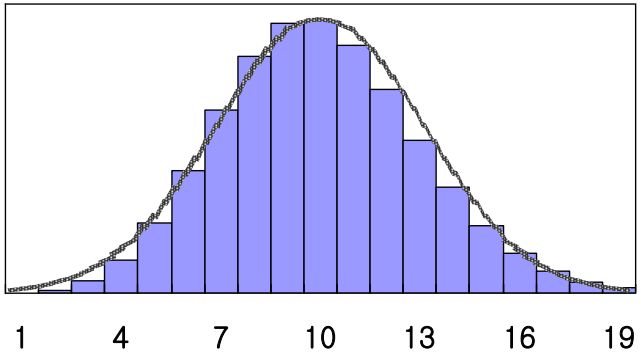
상자의 내용물이 비대칭인 경우 합의 확률히스토그램



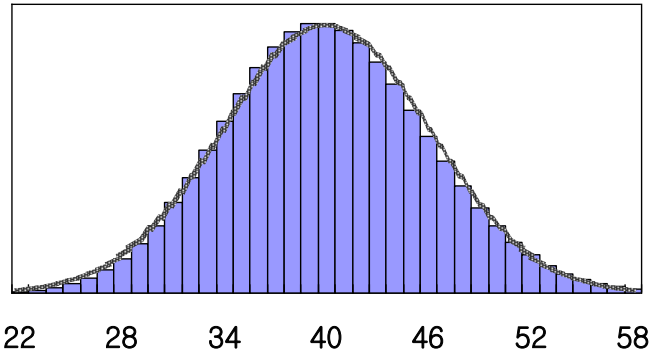
상자의 확률히스토그램



(i) 25회 추출



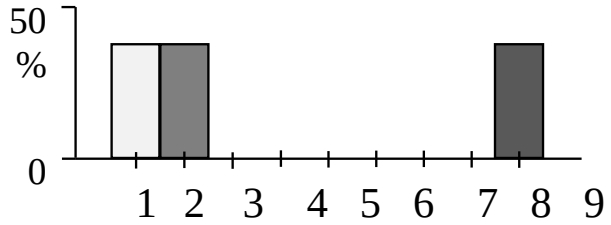
(ii) 100회 추출



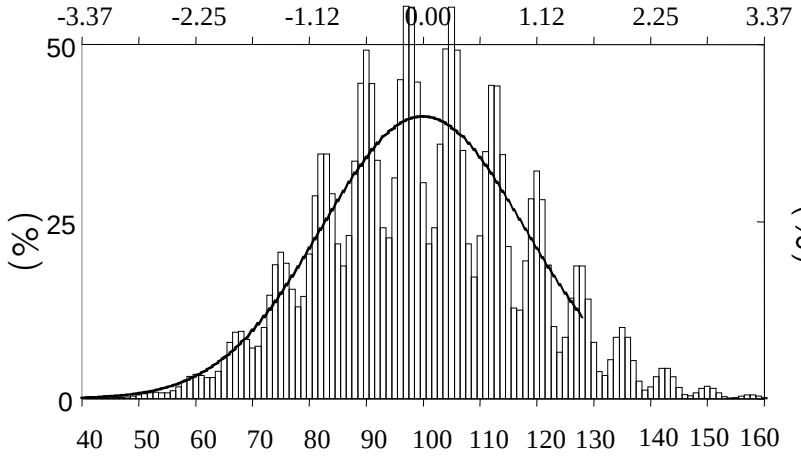
(iii) 400회 추출

6. 정규분포곡선의 사용 범위

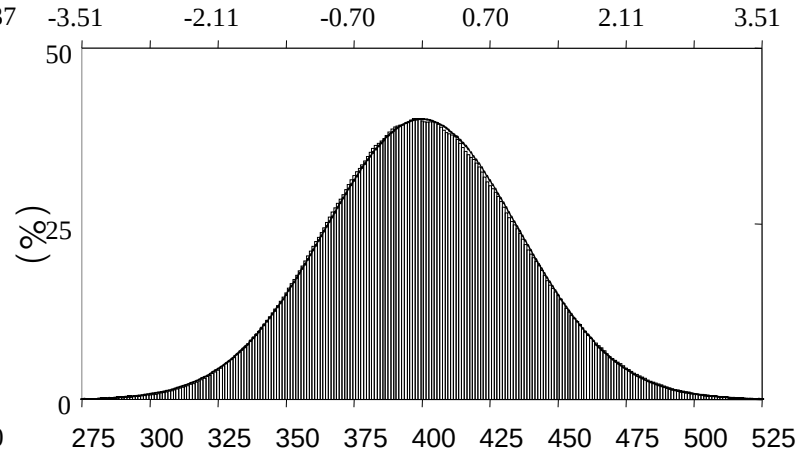
상자의 내용물이 비대칭인 경우 합의 확률히스토그램



상자의 확률히스토그그램



(i) 25회 추출



(ii) 100회 추출

6. 정규분포곡선의 사용 범위

곱의 히스토그램

중심극한정리는 곱에 대해서는 성립하지 않는다.

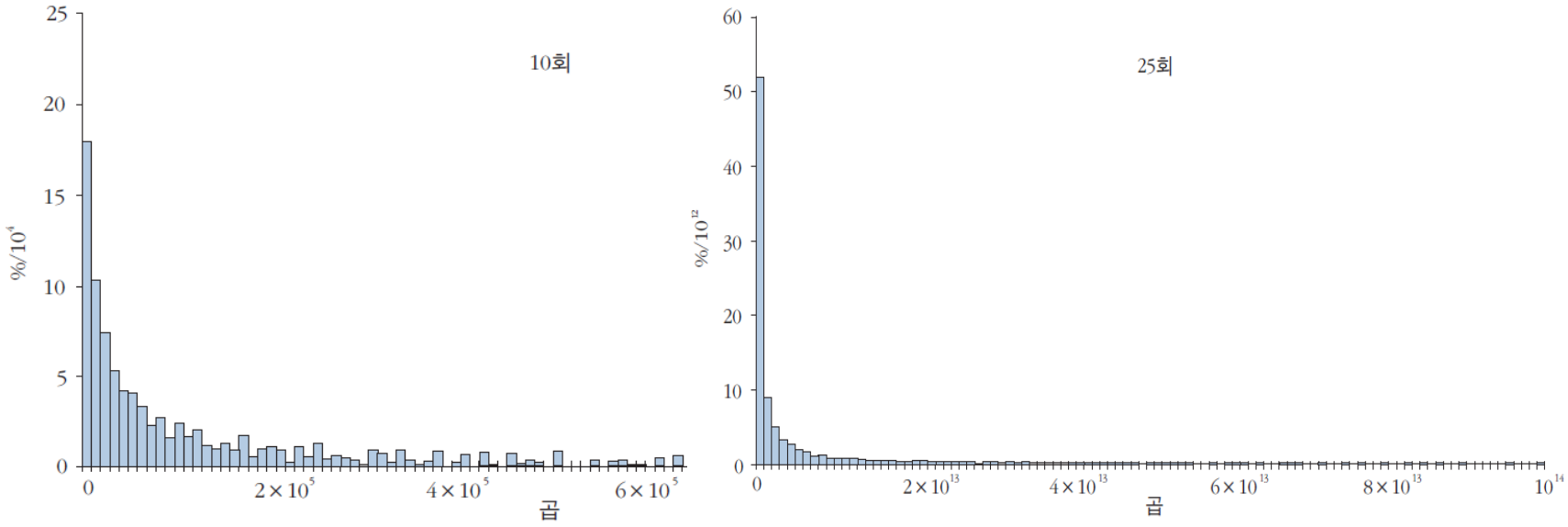
- 곱의 히스토그램은 정규분포곡선과 다르다.
- 곱의 분포는 합의 분포와 모양이 아주 다르다.

6. 정규분포곡선의 사용 범위

곱의 히스토그램

곱의 확률히스토그램은 표본크기가 커져도 정규분포곡선으로 수렴하지 않는다

그림 13-9 주사위를 10회, 25회 던질 때 곱의 확률히스토그램



7. 중심극한정리와 히스토그램의 수렴

중심극한정리

상자로부터 무작위로 복원추출할 때 추출횟수가 증가함에 따라 합 또는 평균의 확률히스토그램은 정규분포곡선으로 수렴해간다. 이를 중심극한정리라고 한다.

- 중심극한정리는 일반적으로 상자의 내용물에 관계없이 성립한다.
- 다만, 확률히스토그램을 정규분포곡선으로 근사시킬 때 근사에 필요한 최소한의 추출횟수는 달라진다.
- 상자의 내용물 분포가 정규분포곡선과 비슷하면 비슷할수록 추출횟수가 적어도 정규분포로 근사가 잘 되나 그렇지 않으면 추출횟수가 많아야 한다.

7. 중심극한정리와 히스토그램의 수렴

기대값과 표준오차

확률히스토그램이 정규분포곡선을 따르면 그 전체 모양은 기대값과 표준오차만으로 완벽하게 묘사된다.

합의 기대값과 표준오차는 다음으로부터 계산한다.

(i) 상자의 평균 (ii) 상자의 표준편차 (iii) 추출횟수

기대값은 확률히스토그램의 중심을 나타내고 표준오차는 퍼진 정도를 나타낸다.

7. 중심극한정리와 히스토그램의 수렴

히스토그램의 수렴

표 13-2 히스토그램의 수렴(convergence of histogram) : 개념이 다른 두 종류의 수렴

	합의 경험적 히스토그램이 합의 확률히스토그램으로 수렴	합의 확률히스토그램이 정규분포곡선으로 수렴
책의 해당 부분	제13장 2절과 3절	제13장 4절-6절
수렴현상의 예시	[그림 13-1]	[그림 13-3], [그림 13-5] [그림 13-6], [그림 13-8]

7. 중심극한정리와 히스토그램의 수렴

히스토그램의 수렴: 평균의 법칙과 중심극한정리

수렴의 조건	반복시행의 횟수가 무한히 증가(한 번의 시행시 추출 횟수는 고정한 채, 전체 시행을 무한히 반복함)	추출 횟수가 무한히 증가(확률히스토그램은 개념상 전체 시행의 횟수가 무한대인 경우에 해당함. 즉, 전체 시행의 횟수는 언제나 무한대인 상태에서 매 번의 시행시 추출 횟수를 무한히 증가시킴)
관련된 법칙	평균의 법칙(대수의 법칙)	중심극한정리

8. 부트스트래핑

부트스트래핑

표본 크기도 크지 않고 마땅히 이론의 도움도 기대할 수 없는 상황에서 표본합이나 평균의 확률히스토그램을 주어진 자료만을 가지고 근사시키는 방법

- 추출횟수가 적을 때 표본합 내지 표본평균의 확률히스토그램을 무작정 정규분포로 근사시키는 것은 위험

- 이러한 경우 표본합 내지 표본평균의 확률히스토그램, 즉 표본분포(sampling distribution)를 근사시키기 위한 현실적 대안 중 하나가 부트스트래핑

8. 부트스트래핑

표본평균의 확률히스토그램을 부트스트래핑으로 근사시키는 절차

주어진 원래의 표본으로부터 복원추출(sampling with replacement)로 같은 크기의 "부트스트랩 표본" 추출

부트스트랩 표본으로부터 새로운 표본평균 계산

위의 절차를 총 1,000번 반복

이와 같이 얻은 1,000개의 표본평균을 히스토그램으로 나타내면 이것이 바로 표본평균의 확률히스토그램에 대한 부트스트랩 근사임

9. 컴퓨터 시뮬레이션

컴퓨터 시뮬레이션의 응용

컴퓨터로 생성한 난수를 이용하여 π의 값을 추정하라.

- 가로 길이와 세로 길이가 각각 1인 정사각형의 면적에서 원점 기준 반지름이 1인 원의 1상한에 해당되는 ¼원의 면적이 차지하는 비율은 바로 π/4
- [0, 1] 의 일양분포로부터 서로 독립인 x와 y 생성하여, y< $\sqrt{1-x^2}$ 인지 판단
- 위의 조건을 만족하는 (x, y) 쌍의 경험적 비율은 대수의 법칙에 의해 π/4로 수렴