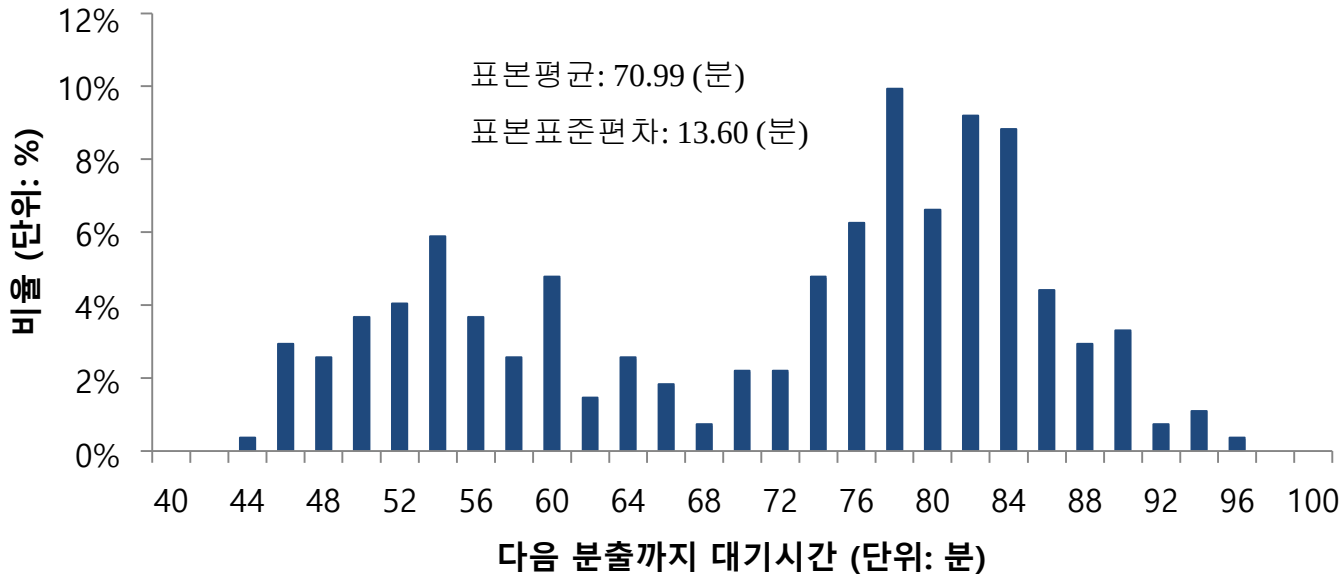
4. Old Faithful: One sample analysis (집단 구분 무시)

다음 분출까지의 대기시간 분석 1

- 미국 Yellowstone 국립공원 내 간헐천 (Geyser)의 분출 대기시간 (y) 분포
- 분출 대기시간의 히스토그램 : 70분 기준, 두 개 봉우리 갖는 쌍봉 분포
- 쌍봉분포라는 사실 무시하고 단일의 정규분포로 잘못 근사하면 대기시간의 95% 예측구간은 70.99 ± 1.96 ×13.60=(44.33, 97.65). 무용지물의 구간임!



4. Old Faithful: Two sample analysis (집단 양분)

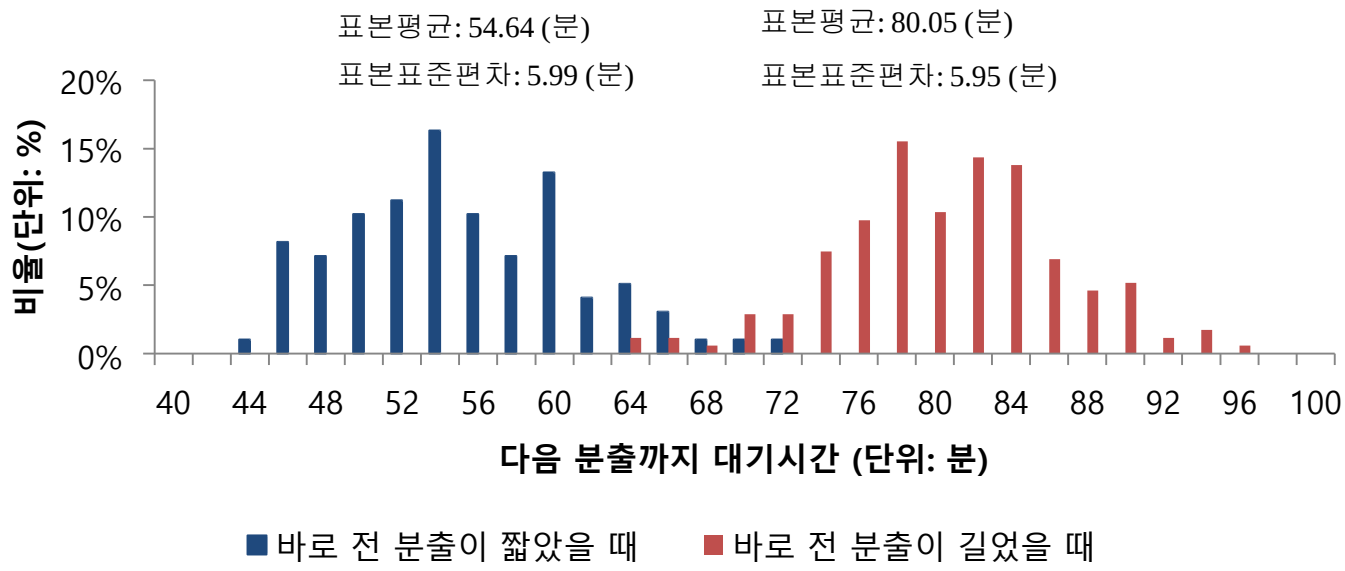
다음 분출까지의 대기시간 분석 2

- 직전의 분출지속기간(x)이 길고 짧았는지에 따라 대기시간 (y) 자료를 양분
  - 직전 분출이 짧았을 때(x<3.2) 개별 y값의 95% 예측구간

    54.64 ± 1.96 ×5.99=(42.90, 66.38)

  - 직전 분출이 길었을 때(x>3.2) 개별 y값의 95% 예측구간

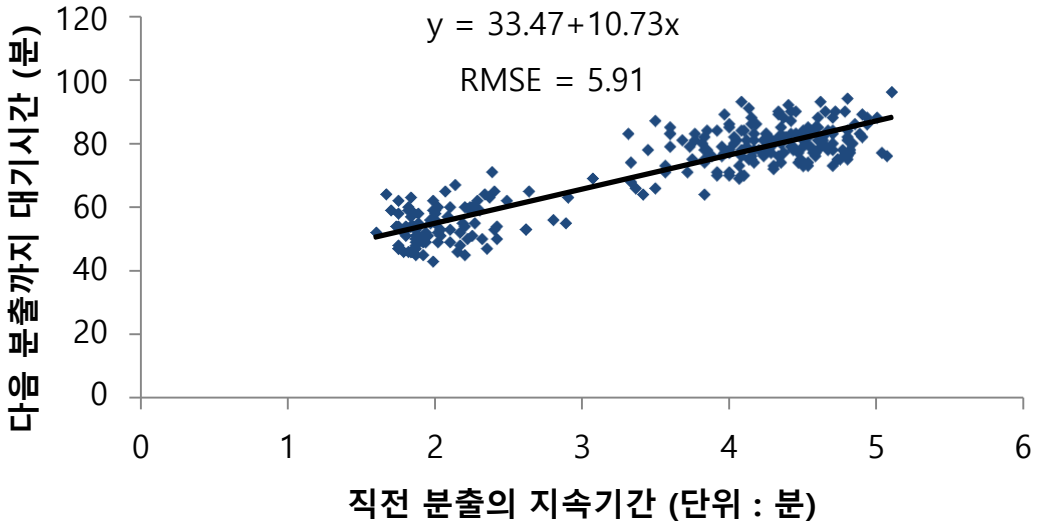
    80.05 ± 1.96 ×5.95=(68.39, 91.71)



4. Old Faithful: Regression analysis (집단 별 분석)

다음 분출까지의 대기시간 분석 3

- 다음 분출까지의 대기시간(y)을 직전 분출의 지속기간(x)에 회귀분석
  - 개별 y값에 대한 95% 예측구간은 33.47+10.73x ± 1.96 ×5.91
    - (43.35, 66.51) for x=2
    - (64.81, 87.97) for x=4



4. Old Faithful: Regression analysis, Real Time Updating

다음 분출까지의 대기시간 분석 4

- 다음 분출 시점을 예측하기 위해 실시간으로 사용할 수 있는 정보는
  - 바로 전 분출의 지속 기간($x = x_0$)
  - 바로 전 분출이 끝난 이후 지금까지 경과한 시간(w : 실시간 업데이트되는 정보)
- 분출 종료 후 막 도착한(w=0) 경우 다음 분출까지 대기시간(y_0) :

  $y_0 \sim N(33.47+10.73x, 5.91^2)$ 로 근사
- 분출 종료 후 일정 시간 경과한(w>0) 경우 다음 분출까지 대기시간 (y_0-w) :

  $y_0 - w \sim N(33.47+10.73x - w, 5.91^2)$의 조건부 분포 (given $y_0 > w$)로 근사.

  즉, 위 정규분포의 truncated normal distribution으로 근사

4. Old Faithful: Regression analysis, Real Time Updating

다음 분출까지의 대기시간 분석 4

- 설명의 편의상 이하 분석에서는 직전 분출이 2분간 지속되었다고 가정 (x=2)
- (i) 직전 분출 종료 후 w분만큼 경과한 경우. 단 $w \le 40$:
  x = 2 로 주어진 경우 y는 평균이 54.93이고 표준편차가 5.91이므로 이제껏 w 만큼 경과했다는 조건이 y의 분포를 사실상 업데이트시키지 못함

  – 총대기시간 y값에 대한 95% 예측구간은 여전히 $33.47+10.73x \pm 1.96 \times 5.91$
    - (43.35, 66.51) for x=2
  – 남은 대기시간인 y-w값에 대한 95% 예측구간은 위 구간에서 w만큼만 차감
    - (43.35-w, 66.51-w) for x=2

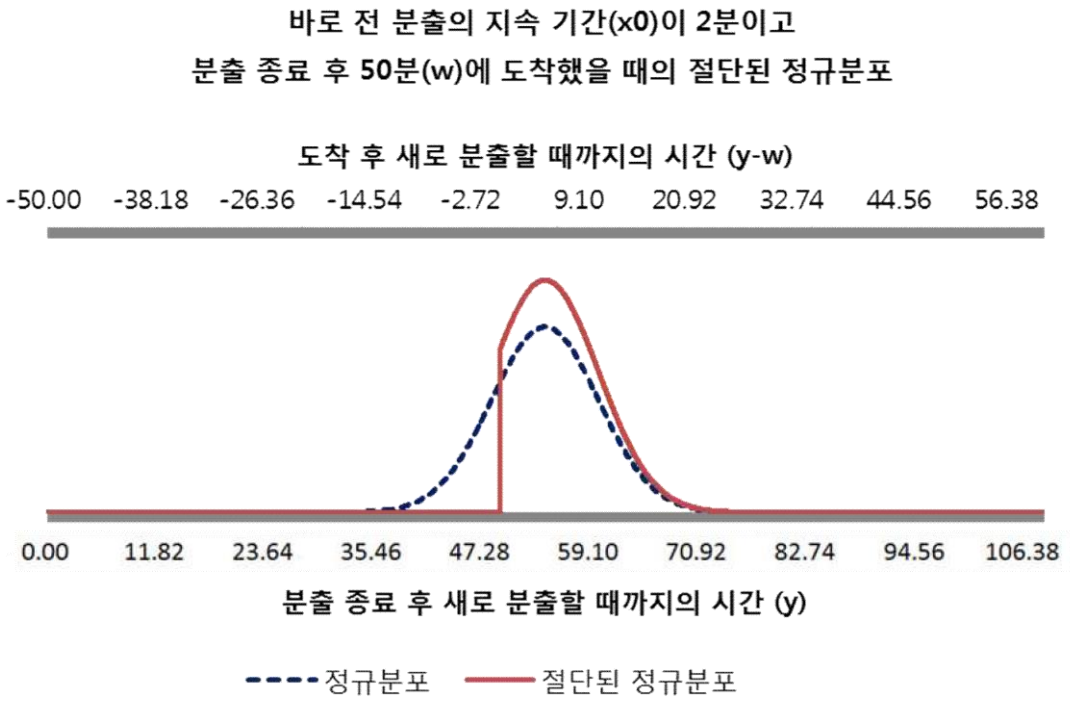
4. Old Faithful: Regression analysis, Real Time Updating

다음 분출까지의 대기시간 분석 4

- (ii) 직전 분출 종료 후 50분만큼 경과한 경우

  x = 2 로 주어진 경우 y는 평균이 54.93이고 표준편차가 5.91이므로 이제껏

  50분 만큼 경과했다는 조건은 y의 분포를 의미있게 truncate시켜 업데이트함



총 대기시간: (50 , 65.28)

남은 대기시간: (0 , 15.28)

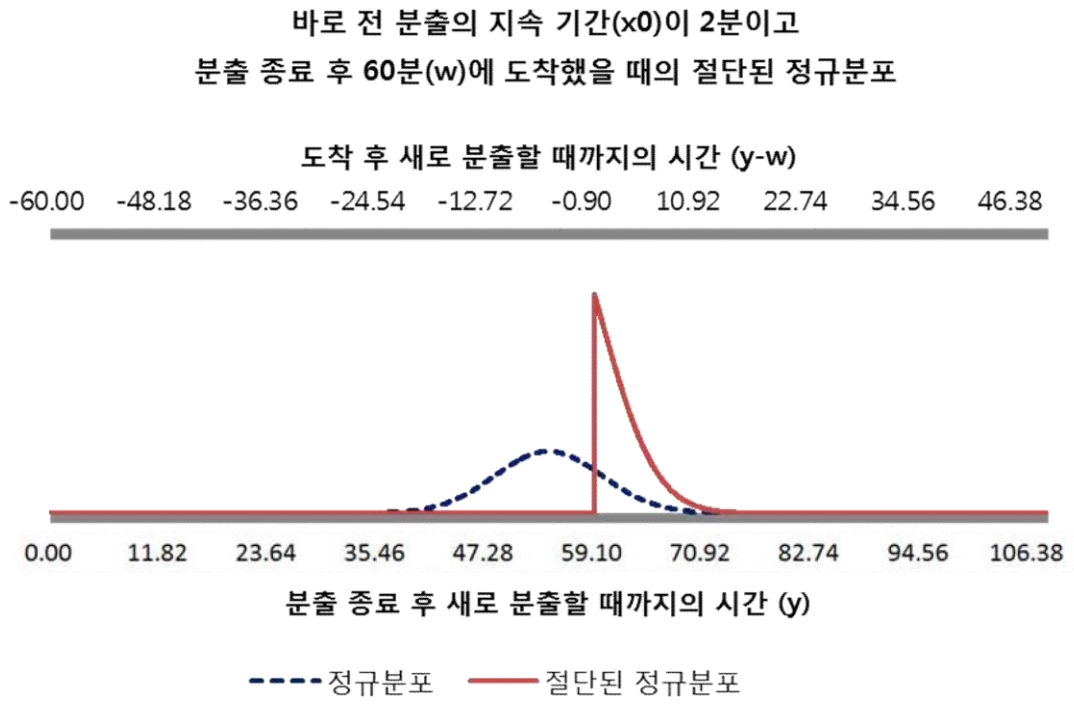
4. Old Faithful: Regression analysis, Real Time Updating

다음 분출까지의 대기시간 분석 4

- (ii) 직전 분출 종료 후 60분만큼 경과한 경우

  x = 2 로 주어진 경우 y는 평균이 54.93이고 표준편차가 5.91이므로 이제껏

  50분 만큼 경과했다는 조건은 y의 분포를 의미있게 truncate시켜 업데이트함



총 대기시간: (60 , 68.73)

남은 대기시간: (0 , 8.73)

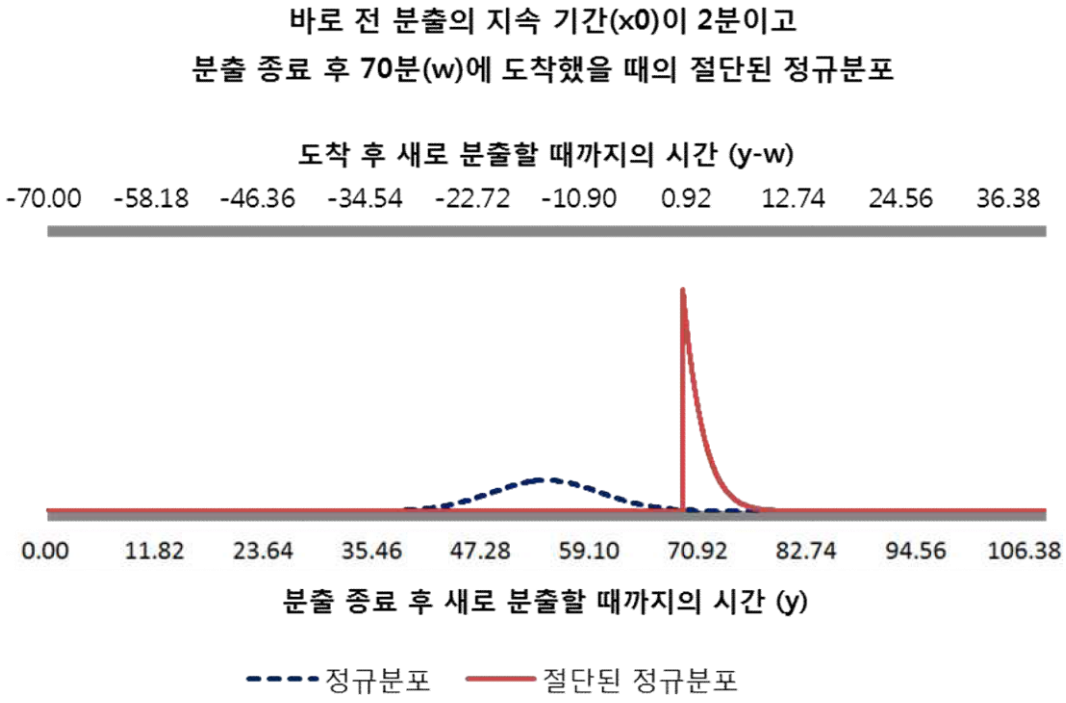
4. Old Faithful: Regression analysis, Real Time Updating

다음 분출까지의 대기시간 분석 4

- (ii) 직전 분출 종료 후 70분만큼 경과한 경우

  x = 2 로 주어진 경우 y는 평균이 54.93이고 표준편차가 5.91이므로 이제껏

  50분 만큼 경과했다는 조건은 y의 분포를 의미있게 truncate시켜 업데이트함



총 대기시간: (70 , 75.38)

남은 대기시간: (0 , 5.38)

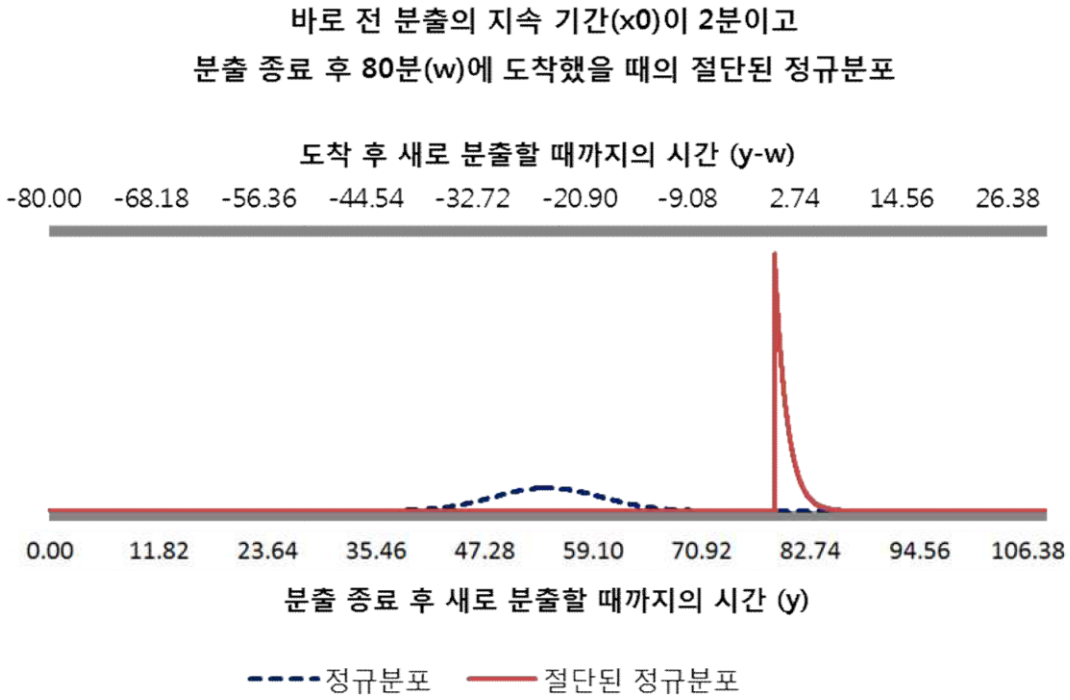
4. Old Faithful: Regression analysis, Real Time Updating

다음 분출까지의 대기시간 분석 4

- (ii) 직전 분출 종료 후 80분만큼 경과한 경우

  x = 2 로 주어진 경우 y는 평균이 54.93이고 표준편차가 5.91이므로 이제껏

  50분 만큼 경과했다는 조건은 y의 분포를 의미있게 truncate시켜 업데이트함



총 대기시간: (80 , 83.72)

남은 대기시간: (0 , 3.72)

제 21 장 χ^2 - 검정

1. χ^2 -검정의 소개
2. χ^2 -검정의 구조
3. 확률적 독립의 검정

1. χ^2- 검정의 소개

카이제곱검정

1. χ^2- 검정의 소개

카이제곱검정

범주(category)별로 관측된 빈도와 기대빈도의 차이를 봄으로써 하나의 확률모형이 전반적으로 자료를 얼마나 잘 설명하는지 검정하는데 사용

종합주가지수가 오를지 내릴지 예측하는 경우
2개의 범주만 존재 → $z-$ 검정
부호 검정

종합주가지수의 수준을 구간별로 예측하는 경우
3개 이상의 범주가 존재 → χ^2- 검정

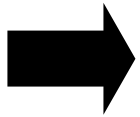
1. χ^2- 검정의 소개

카이제곱검정

도박사는 균형이 잘 잡히지 않은 주사위를 사용했는가?

주사위를 60번 던진 결과

4 3 3 1 2 3 4 6 5 6
2 4 1 3 3 5 3 4 3 4
3 3 4 5 4 5 6 4 5 1
6 4 4 2 3 3 2 4 4 5
6 3 6 2 4 6 4 6 3 2
5 4 6 3 3 3 5 3 1 4



주사위 눈	관측된 도수	기대도수
1	4	10
2	6	10
3	17	10
4	16	10
5	8	10
6	9	10
합	60	60

실제 관측된 도수와 기대도수의 차이가 크다.

1~6까지의 숫자가 한 장씩 든 상자로부터 60회 복원추출하는 상황에서의 기대값

1. χ^2- 검정의 소개

카이제곱검정

여러 행 중에서 한두 행이 이상하다고 전체가 이상하다고 결론짓기는 이르다.
각각의 관측된 도수와 기대도수의 차이를 한 번에 모아서 전반적인 차이를 재는 척도가 필요하다.

$$\chi^2 = \sum \frac{(\text{관측도수-기대도수})^2}{\text{기대도수}}$$

χ^2- 통계량이 크다는 것은 관측된 도수와 기대도수가 전반적으로 큰 차이를 보임을 뜻한다.

$$\chi^2 = \frac{(4-10)^2}{10} + \frac{(6-10)^2}{10} + \frac{(17-10)^2}{10} + \frac{(16-10)^2}{10} + \frac{(8-10)^2}{10} + \frac{(9-10)^2}{10} = 14.2$$

1. χ^2- 검정의 소개

카이제곱검정

14.2 라는 값은 지나치게 크기 때문에 모형을 의심한다.
균형이 잘 잡힌 주사위를 60회 던지더라도 운에 따라 큰 카이제곱의 값이 나올 수도 있지만, 그러한 값이 나올 가능성의 크기가 문제다.
균형이 잘 잡힌 주사위를 60회 던져서 χ^2-통계량 값을 구하는 과정을 1,000번 반복하여 1,000개의 χ^2–통계량의 값을 얻는다.

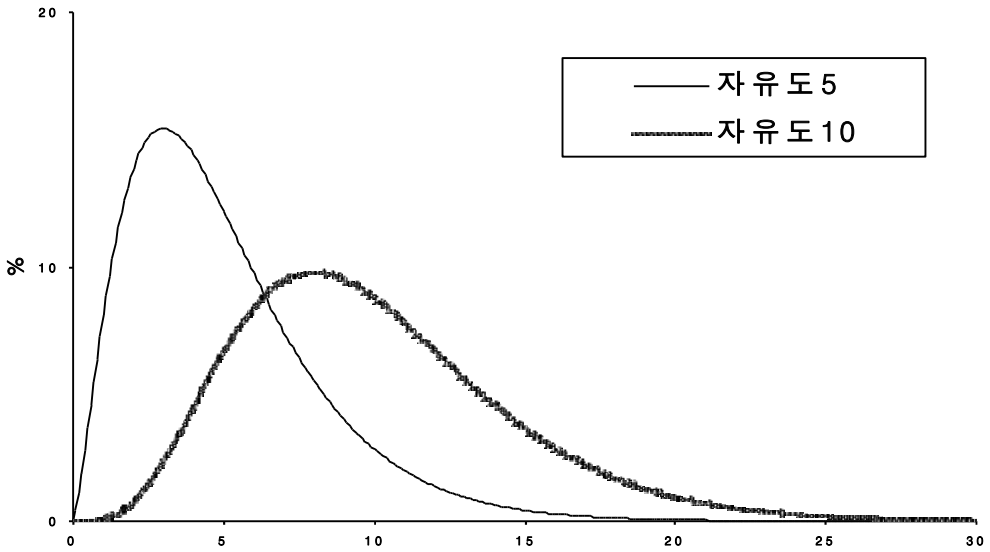
χ^2–통계량 값을 히스토그램(χ^2–분포의 경험적 히스토그램)으로 정리했을 때 14.2 오른쪽의 면적 = 1,000개의 χ^2-통계량 값 중 14.2보다 같거나 큰 것의 비율 = 1.4%. (p-값)

- 관측된 14.2 보다 큰 χ^2-통계량 값은 주사위가 균형 잡혔다는 귀무가설에 반하는 더 강력한 반증임
- 하나의 확률모형이 자체적으로 자신에 대해 이토록 강력한 반증을 만들어낼 확률은 얼마인가? → p-값의 의미

1. χ^2- 검정의 소개

카이제곱분포와 자유도

자유도 5와 자유도 10에 대응하는 χ^2-분포 곡선



이 분포곡선들은
오른쪽으로 늘어진 꼬리를
갖는다.
자유도가 증가함에 따라
곡선은 오른쪽으로
움직이면서 좀더
대칭적으로 된다.

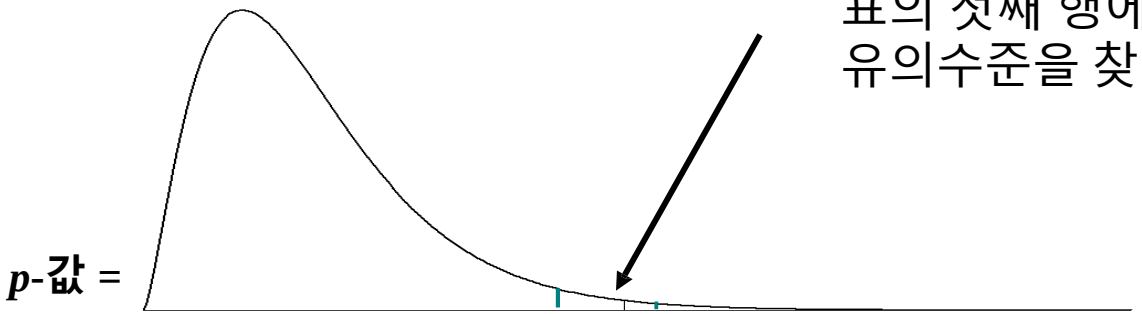
모형이 구체적으로 설정되어 모수 추정이 전혀 불필요한 경우:

자유도 = χ^2-통계량 계산에 사용된 항의 개수 - 1 → 자유도
= 6-1 = 5

1. χ^2- 검정의 소개

카이제곱분포의 임계치

자유도 5의 χ^2-분포곡선



표의 첫째 행에서 미리 설정한 유의수준을 찾는다.

면적과 자유도가 만나는 위치에 놓인 수치를 읽는다.

→ 14.2 오른쪽의 면적은 5%와 1% 사이의 값이다.

χ^2-분포표의 일부

자유도	50%	10%	5%	1%
3	2.37	6.25	7.82	11.34
4	3.36	7.78	9.49	13.28
5	4.35	9.24	11.07	15.09
6	5.35	10.65	12.59	16.80
7	6.35	12.02	14.07	18.48

1. χ^2- 검정의 소개

z-검정, t-검정, 카이제곱검정

1부터 6까지 숫자가 적힌 카드가 들어 있는 상자에서 복원추출

상자의 평균이 3.5라는 가설을 검정 ⟹ z – 검정, t - 검정

각 카드가 나올 확률이 1/6 씩이라는 가설을 검정 ⟹ χ^2 – 검정

χ^2-검정은 상자의 내용 구성이 알려져 있을 때 관측된 자료를 이 상자로부터 무작위 추출한 결과로 볼 수 있는지 알려준다.

z-검정 또는 t-검정은 상자의 평균만 주어져 있을 때 관측된 자료를 이 상자로부터 무작위 추출한 결과로 볼 수 있는지 알려준다.

2. χ^2 - 검정의 구조

카이제곱검정의 구조

기본 자료	확률모형	도수 분포표
자료의 크기는 일반적으로 n 으로 나타낸다. 예) n=60	상자모형 예) 주사위 모형: 1~6 이 한 장씩 들어 있는 상자 내용구성이 알려진 상자로부터 무작위로 복원추출	각각의 값에 대해 관측된 도수를 기록 이를 도수분포표의 형태로 정리

2. χ^2 - 검정의 구조

카이제곱검정의 구조

χ^2-통계량	자유도	관측된 유의수준(p-값)
$\sum \frac{(\text{관측도수}-\text{기대도수})^2}{\text{기대도수}}$ 예) 14.2	모수의 추정이 불필요한 경우 자유도는 χ^2-통계량 계산 시 사용한 항의 개수-1 예) 6-1=5	해당 자유도를 가진 χ^2-분포곡선 아래의 면적 중 관측된 χ^2-통계치 오른쪽의 면적이 p-값 예) p-값=1.4%