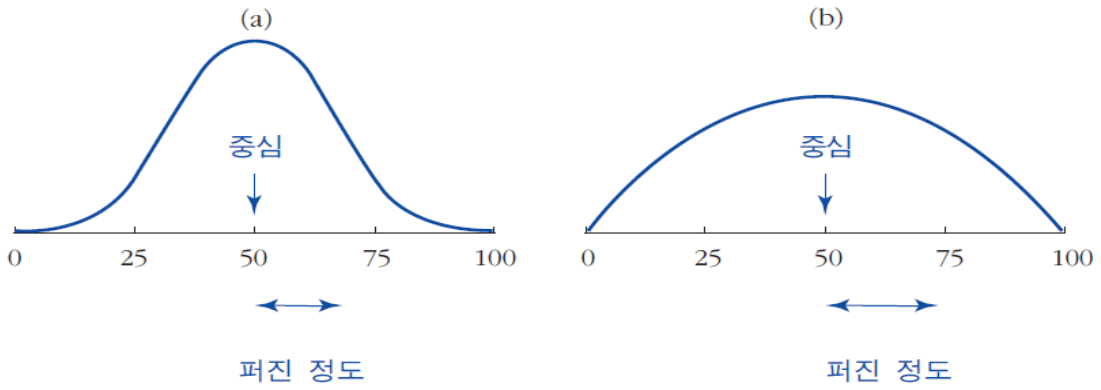
제 3 장 평균과 표준편차

1. 자료의 중심과 퍼진 정도
2. 평균, 중앙값, 최빈치
3. 제곱근-평균-제곱(RMS)
4. 표준편차
5. 자유도
6. 측정오차, 편의, 이탈값

1. 자료의 중심과 퍼진 정도

중심과 퍼진 정도

그림 3-1 히스토그램의 중심과 퍼진 정도



히스토그램에서 자료를 요약할 때 중심(평균, 중앙값)과 중심 주위로 퍼진 정도(표준편차, 사분위수 범위)를 주로 사용

2. 평균, 중앙값, 최빈치

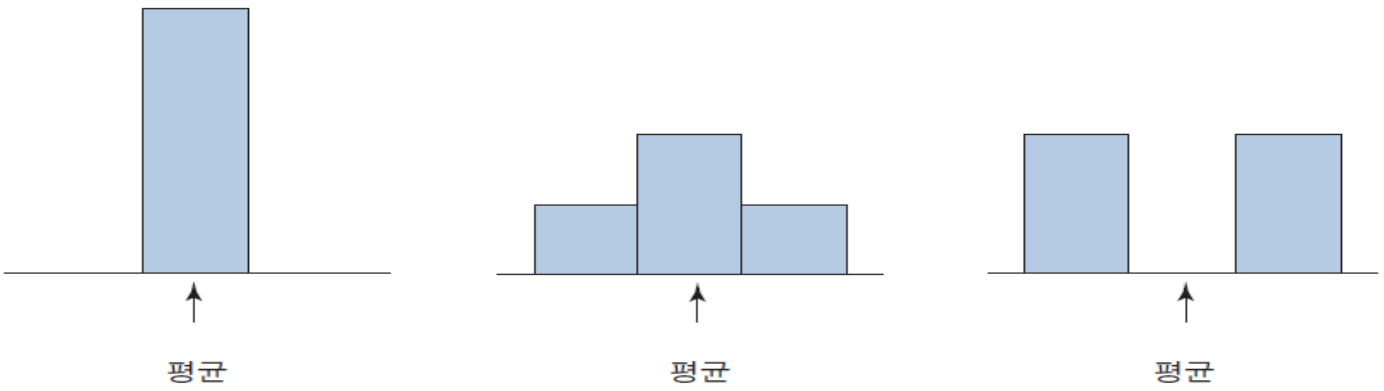
평균

평균(mean)은 관측치의 총합을 관측치의 개수로 나누어 구한다.

$X_1, X_2, \ldots, X_n$ 등 n개의 관측지가 주어져 있을 때 표본 평균은

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

평균이 중요하지만 전부는 아님. 아래 세 그림은 평균이 같지만 퍼진 정도가 다름



2. 평균, 중앙값, 최빈치

중앙값, 최빈치

중앙값(median)

- 절반 이상의 숫자들이 이 값보다 크거나 같고 동시에 절반 이상의 숫자들이 이 값보다 작거나 같은 수
- 히스토그램은 중앙값에서 그 면적이 양분됨
- 중앙값은 n 이 홀수이면 (n+1)/2번째로 크거나 작은 숫자임
- 중앙값은 n이 짝수이면 n/2번째 숫자와 (n+1)/2번째 숫자의 평균으로 정의
- median voter theorem: 중앙값이 LAD의 해로 얻어진다는 것과 수학적으로 같은 내용임. 선호의 비대칭분포 이용하여 후보자의 location choice 문제 설명.

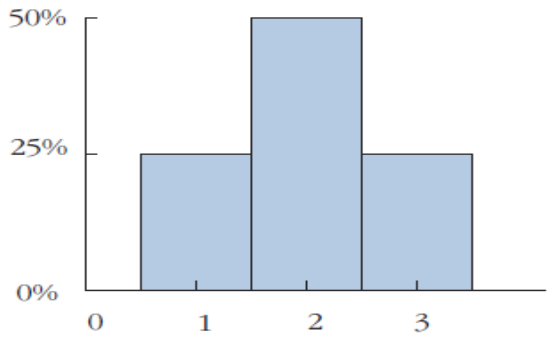
최빈치(mode)

- 가장 많이 관찰되는 값
- 히스토그램은 최빈치에서 그 높이가 제일 높음

2. 평균, 중앙값, 최빈치

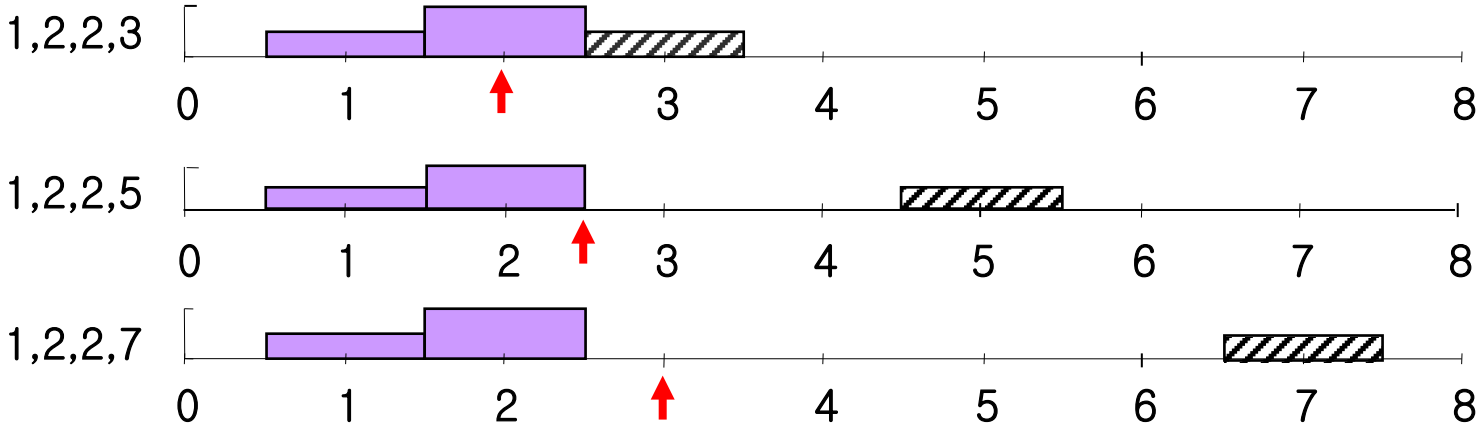
평균과 중앙값의 관계

숫자열 1, 2, 2, 3에 대한 히스토그램



히스토그램이 대칭이면 평균=중앙값

숫자열의 변화에 따른 평균의 변화

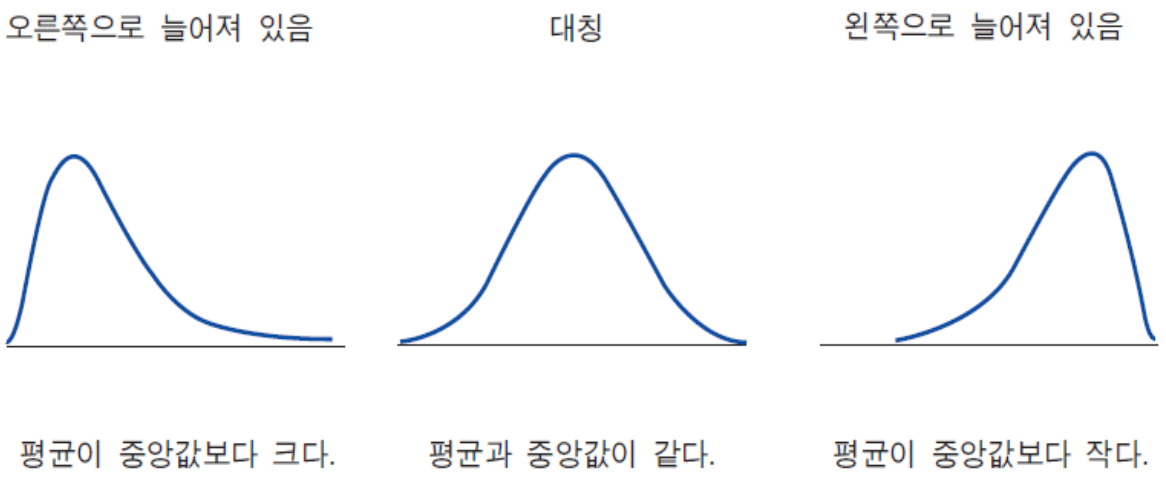


위의 세 히스토그램에서 중앙값은 언제나 2이나 평균은 ↑ 따라서 이동

2. 평균, 중앙값, 최빈치

히스토그램의 세 가지 꼬리 유형

그림 3-7 히스토그램의 세 가지 꼬리 유형



평균은 극단적인 값의 영향을 받음

중앙값은 극단적인 값의 영향을 받지 않음

극단적인 값이 존재하는 경우 평균보다 중앙값이 중심을 더 잘 나타냄

2. 평균, 중앙값, 최빈치

야구 통계: 희생번트의 득실

[표 1] 주자와 아웃카운트 상황별 득점 확률 및 평균 득점

주자	아웃 카운트	해당 이닝 동안 적어도 1점 이상 득점한 비율	이닝 당 평균 득점	관측치수
1루	0	0.396	0.813	1,728
2루	1	0.390	0.671	657
2루	0	0.619	1.194	294
3루	1	0.693	0.980	202
1,2루	0	0.605	1.471	367
2,3루	1	0.730	1.560	176

Source: Lindsey (1963), Tanur et al. (1976)에서 재인용.

2. 평균, 중앙값, 최빈치

야구 통계: 2010년 한국 프로야구 팀별 공격력과 수비력

<표 1: 2010년 한국프로야구 팀별 공격과 수비 성적표>

공격력 순위	공격력				수비력 순위	수비력			
	타율	장타율	출루율	도루		방어율	실점	자책점	실책
롯데	0.288	0.461	0.352	124	SK	3.71	545	494	81
두산	0.281	0.440	0.365	128	삼성	3.94	575	523	92
LG	0.276	0.411	0.349	169	기아	4.39	641	577	86
SK	0.274	0.412	0.355	161	넥센	4.55	652	599	93

* 2010시즌 공격력 수비력 각각 상위 4개 팀만을 대상으로 함. 전체 팀에 대한 순위는 류근관의 웹사이트 ezstat.co.kr=>current lectures=>baseball Academy 내 관련 자료 참조)

*위 표에서 도루, 실점, 자책점, 실책 등은 팀별로 시즌 133 경기에 걸친 전체 값을 나타냄. 한편 장타율과 출루율은 다음 공식에 따라 구해짐 (4사구에는 "hit by pitch"도 포함됨)

※ 장타율 = {(1루타×1) + (2루타×2) + (3루타×3) + (홈런×4)} ÷ 타수

※ 출루율 = (안타+4사구) ÷ (타수+4사구+희생플라이)

2. 평균, 중앙값, 최빈치

야구 통계: 2010년 한국 프로야구 투수 순위

<표 3: 2010년 한국프로야구 투수 개인별 성적표: 상위 6인>

순위*	방어율 순위	이름	팀명	이닝	탈삼진/(4사구-고의4구)	탈삼진	4사구	고의4구	방어율
1	1	류현진	한화	192 2/3	3.596	187	54	2	1.82
2	3	카도쿠라	SK	153 2/3	2.554	143	58	2	3.22
3	6	장원삼	삼성	151	2.396	115	49	1	3.46
4	9	김선우	두산	154 2/3	2.246	128	58	1	4.02
5	12	장원준	롯데	144 1/3	2.093	113	56	2	4.43
6	2	김광현	SK	193 2/3	2.080	183	89	1	2.37

*첫 번째 열의 순위는 2010 정규시즌 규정 이닝인 133이닝을 채운 전체 15명의 투수 가운데 "탈삼진/(4사구-고의4구)" 기준 상위 6인의 순위임. 전체 15명 투수의 순위는 류근관의 웹사이트 ezstat.co.kr=>current lectures=>baseball Academy 내 관련 자료 참조)

2. 평균, 중앙값, 최빈치

야구 통계: 2010년 한국 프로야구 타자 순위

<표 4: 2010년 한국프로야구 타자 개인별 성적표: 상위 10인>

기여점수 순위	타율 순위	이름	소속팀	기여점수	타수	안타	1루타	2루타	3루타	홈런	4사구
1	1	이대호	롯데	1.121	3.764	1.370	0.921	0.102	0	0.346	0.559
2	2	홍성흔	롯데	1.005	3.883	1.360	0.865	0.252	0.00901	0.234	0.550
3	7	김현수	두산	0.799	3.583	1.136	0.735	0.220	0	0.182	0.636
4	21	김동주	두산	0.733	3.518	1.036	0.655	0.200	0	0.182	0.700
5	6	조인성	LG	0.723	3.436	1.090	0.692	0.180	0.00752	0.211	0.331
6	5	최준석	두산	0.722	3.339	1.071	0.685	0.205	0.00787	0.173	0.433
7	29	최형우	삼성	0.714	3.471	0.967	0.521	0.240	0.00826	0.198	0.653
8	14	강민호	롯데	0.699	3.504	1.068	0.701	0.162	0.00855	0.197	0.427
9	19	최정	SK	0.698	3.252	0.976	0.553	0.252	0.00813	0.163	0.561
10	11	박정권	SK	0.688	3.476	1.065	0.734	0.153	0.0323	0.145	0.548
평균				**0.707**	**3.503**	**1.072**	**0.707**	**0.192**	**0.007**	**0.167**	**0.526**

*위 표의 모든 데이터는 개인별로 2010시즌 경기당 평균 수치이고, 개인 순위는 2010시즌 규정타석을 채운 총 45명의 타자 가운데 상위 10명의 순위임. 규정 타석을 전부 채운 타자 상위 20명의 순위는 류근관의 웹사이트 ezstat.co.kr=>current lectures=>baseball Academy 내 관련 자료 참조). 참고로 2010시즌 규정타석은 133(경기수)×3.1=412.3 타석임. 마지막 행의 '평균'은 기여점수 순위 1~20위 타자들의 평균값임.

**기여점수와 총잔루수는 다음 공식에 따라 계산된다.

※ 기여점수 = (안타+4사구) × 총잔루수 ÷ (타수+4사구)

※ 총잔루수 = 1루타×1 + 2루타×2 + 3루타×3 + 홈런×4

3. 제곱근-평균-제곱(RMS)

제곱근-평균-제곱 (Root Mean Square)

계산은 표현의 역순(제곱 후 평균, 최종적으로 제곱근)

(1) 제곱 (S) : 모든 수를 제곱하여 부호를 없앤다.

(2) 평균 (M) : 제곱된 값들의 평균을 구한다.

(3) 제곱근 (R) : 제곱-평균된 값에 제곱근을 취한다.

$$RMS = \sqrt{\text{숫자들의제곱의평균}}$$

예: 0, 1, -3, 3, -1의 RMS

$$\sqrt{\frac{0^2 + 1^2 + (-3)^2 + 3^2 + (-1)^2}{5}} = \sqrt{4} = 2$$

4. 표준편차

표준편차의 계산

표준편차(SD)는 "평균으로부터의 편차들"의 RMS와 "대략" 비슷

표본 분산 및 표본 표준편차는

$$S^2 = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2}{n-1}$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, \quad s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

예: 20, 10, 15, 15 에 대한 표준편차

- 평균은 (20+10+15+15)/4 = 15이고, 평균으로부터의 편차들은 5, -5, 0, 0 이므로, 표준편차는

$$\sqrt{\frac{5^2 + (-5)^2 + 0^2 + 0^2}{4-1}} = \sqrt{16.7} \approx 4.1$$

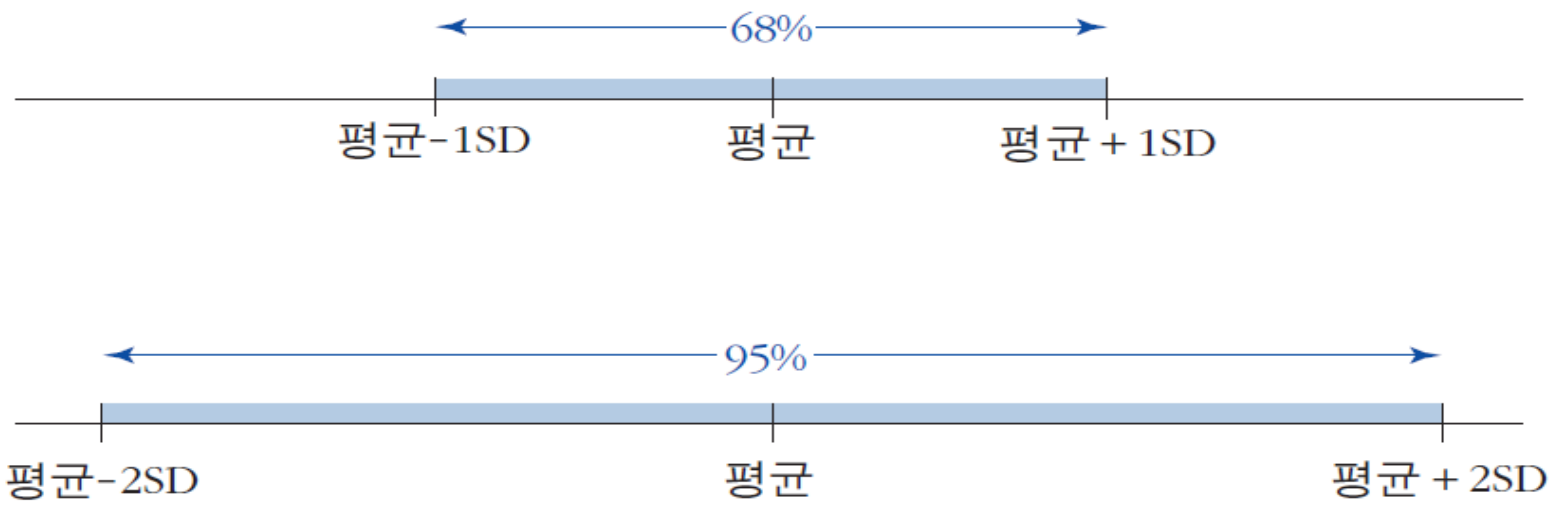
4. 표준편차

표준편차의 의미

표준편차는 관측치들이 평균으로부터 얼마나 떨어져 있는지 알려줌

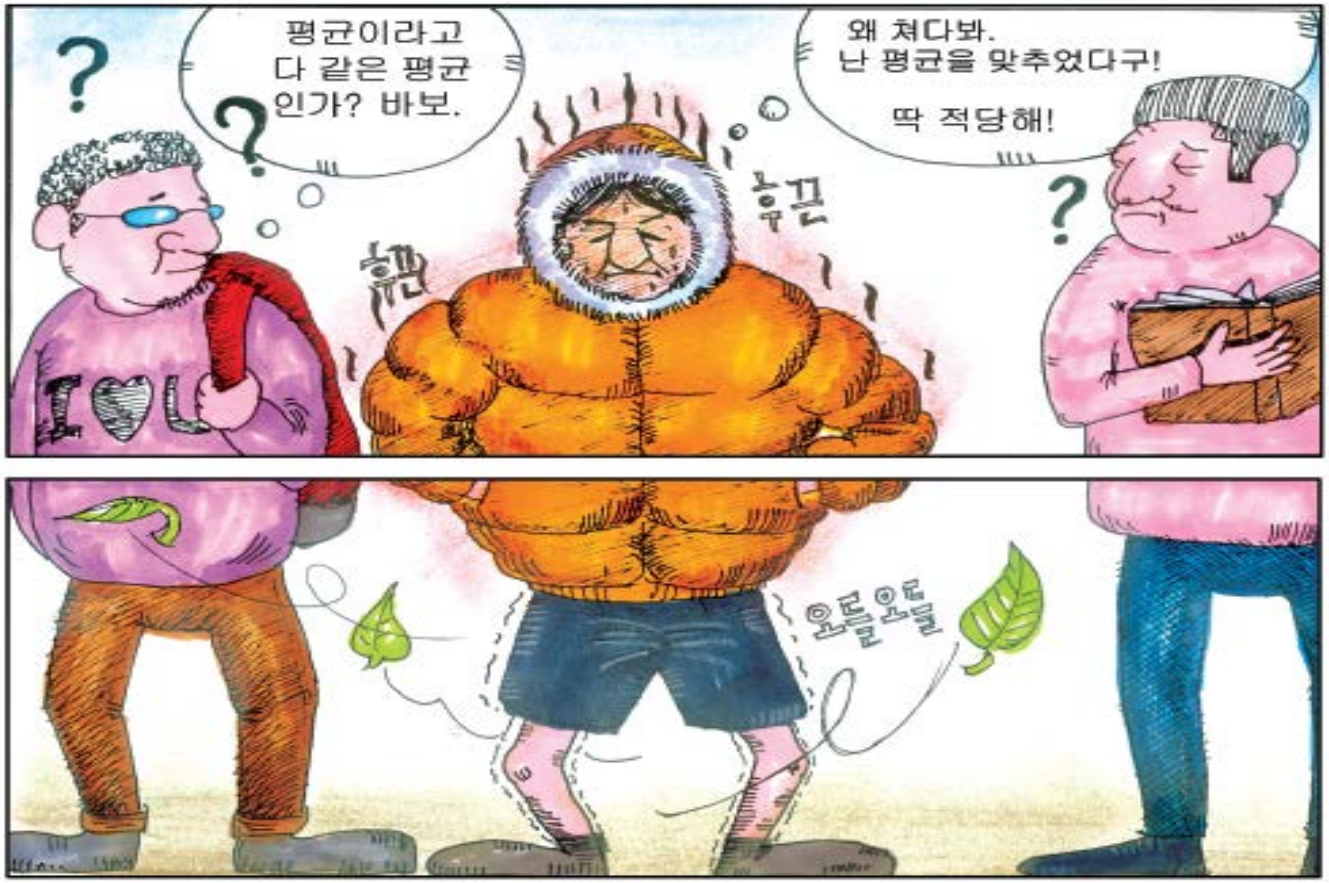
68-95법칙

- 관측치들의 약 68% 정도가 평균으로부터 1 표준편차 이내로 떨어져 있다.
- 관측치들의 약 95% 정도가 평균으로부터 2 표준편차 이내로 떨어져 있다.



4. 표준편차

평균뿐만 아니라 표준편차도 중요



평균만으론 곤란!

5. 자 유 도

자유도의 정의

자유도는 합쳐진 값들 중에서 실질적으로 독립인 값들의 개수

표준편차 계산하는 경우의 자유도는 "자료의 개수 - 1"

표준편차 계산의 대상이 되는 편차들의 합은 0이 됨. 편차들의 합이 0이 된다는 하나의 제약조건이 자유도를 1만큼 감소시킨 것임

5. 자 유 도

자유도의 정의

예: 극단적으로 자료의 개수가 하나인 경우

- 편차는 단 하나뿐이고 그 값은 0임
- 0에 대해 제곱의 평균, 즉 MS(mean square)를 구할 때 자유도 고려치 않으면 0/1=0이고 자유도를 고려하면 0/(1-1)=0/0으로 부정형(indefinite form)이 됨
- 단 하나의 자료만을 가지고는 퍼진 정도를 알 수 없음. 이 때 퍼진 정도는 0이 아니라 '알 수 없다(부정형)'가 정답임. 즉, 자유도를 고려해야 함

6. 측정오차, 편의, 이탈값

정의

측정오차(measurement error)

- 관측치와 실제 값의 차이
- 측정오차가 존재하면, (관측치)=(실제 값)+(측정오차)
- 측정오차의 대략적인 크기는 관측치들의 표준편차(SD)를 통해 알 수 있음
- 표준편차(SD)의 크기는 한 번의 관측에서 측정오차가 어느 정도 될지 알려 줌

편의(bias)

- 방향성을 갖는 하나의 체계적인 오차
- 측정오차와 함께 편의가 있으면, (관측치)=(실제 값)+(편의)+(측정오차)

이탈값(outlier)

- 극단적인 관측치