

Chebyshev's 정리

Chebyshev's theorem

For *any* set of data (either population or sample) and for any constant k greater than 1, the proportion of the data that must lie within k standard deviations on either side of the mean is *at least*

$$1 - \frac{1}{k^2}$$

Chebyshev's 정리

Results of Chebyshev's theorem

For *any* set of data:

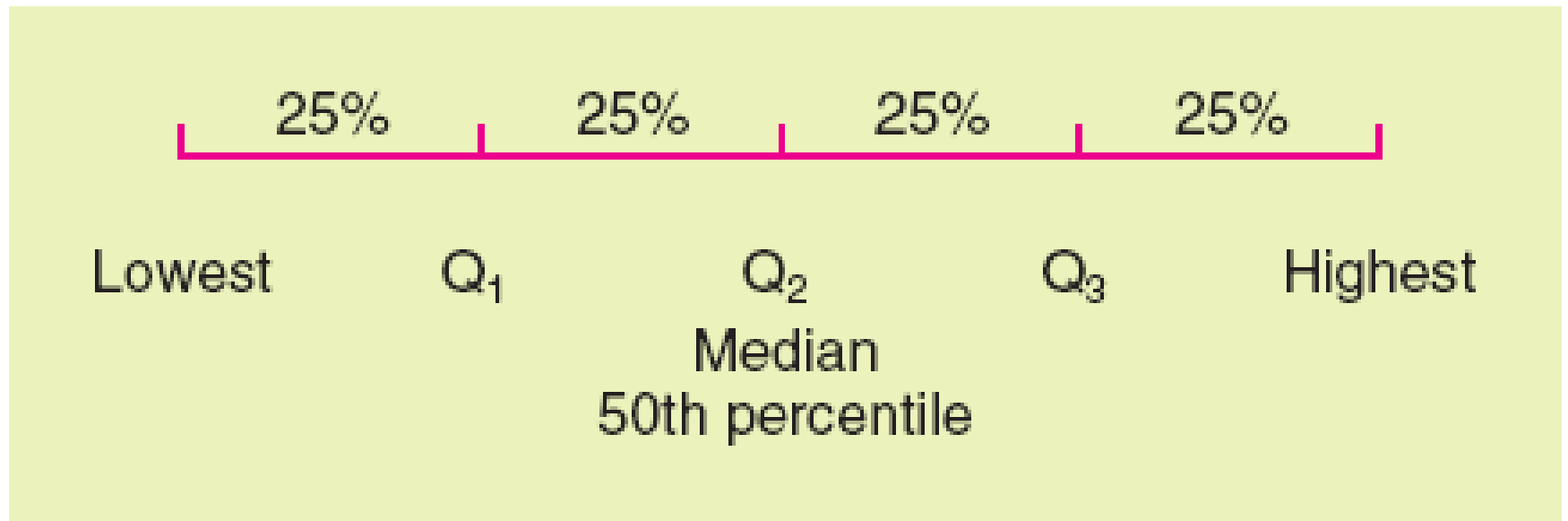
- at *least* 75% of the data fall in the interval from $\mu - 2\sigma$ to $\mu + 2\sigma$.
- at *least* 88.9% of the data fall in the interval from $\mu - 3\sigma$ to $\mu + 3\sigma$.
- at *least* 93.8% of the data fall in the interval from $\mu - 4\sigma$ to $\mu + 4\sigma$.

백분위수와 사분위수

- For whole numbers P , $1 \leq P \leq 99$, the P^{th} percentile of a distribution is a value such that $P\%$ of the data fall below it, and $(100-P)\%$ of the data fall at or above it.
- $Q_1 = 25^{th}$ Percentile
- $Q_2 = 50^{th}$ Percentile = The Median
- $Q_3 = 75^{th}$ Percentile

사분위수와 사분위수 범위(IQR)

Quartiles



$$\text{Interquartile range} = Q_3 - Q_1$$

〈표 3-7〉 누가백분율도표를 그리기 위한 기초통계학 200명의
학기말 시험점수 묶음누가도수표

급 간	급간 상한값	도 수(f)	백분 율(%)	누가백분 율(%)
30.5~37.5	37.5	6	3	3
37.5~44.5	44.5	14	7	10
44.5~51.5	51.5	32	16	26
51.5~58.5	58.5	40	20	46
58.5~65.5	65.5	22	11	57
65.5~72.5	72.5	36	18	75
72.5~79.5	79.5	28	14	89
79.5~86.5	86.5	16	8	97
86.5~93.5	93.5	4	2	99
93.5~100.5	100.5	2	1	100
		200	100	

다섯 숫자 요약

- A listing of the following statistics:
 - Minimum, Q_1 , Median, Q_3 , Maximum
- Box-and-Whisker plot – represents the five-number summary graphically.

◀예제 ▶ 사분위수 구하기

♣ case 1 (맥과이어의 홈런)
⇒ 홀수 개의 자료

9	9	22	32	33	39	39	42	49	52	58	65	70
		↑				↑			↑			
		Q_1				M			Q_2			

♣ case 2 (로저 메리스의 홈런)
⇒ 짝수 개의 자료

8	13	14	16	23	26	28	33	39	61
		↑			↑		↑		
		Q_1			M		Q_2		

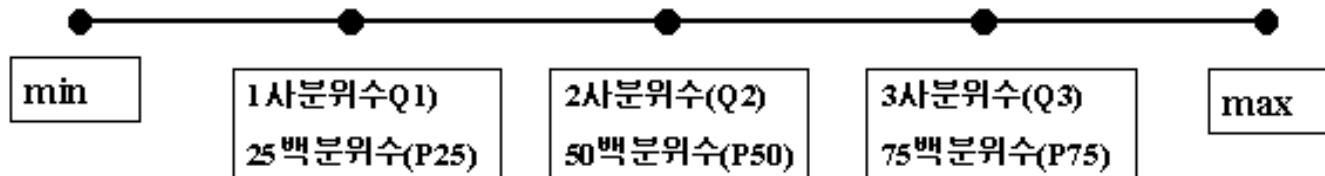
<Note> 다섯 숫자 요약과 상자 그림

⇒ 중앙값과 사분위수를 통해서 알 수 없는 분포의 꼬리에 대한 정보를 얻을 수 있다.

→ 중심과 퍼짐에 대한 정보를 얻기 위해서 다섯 숫자 요약을 사용할 수 있다.

<알아두기> 다섯 숫자 요약

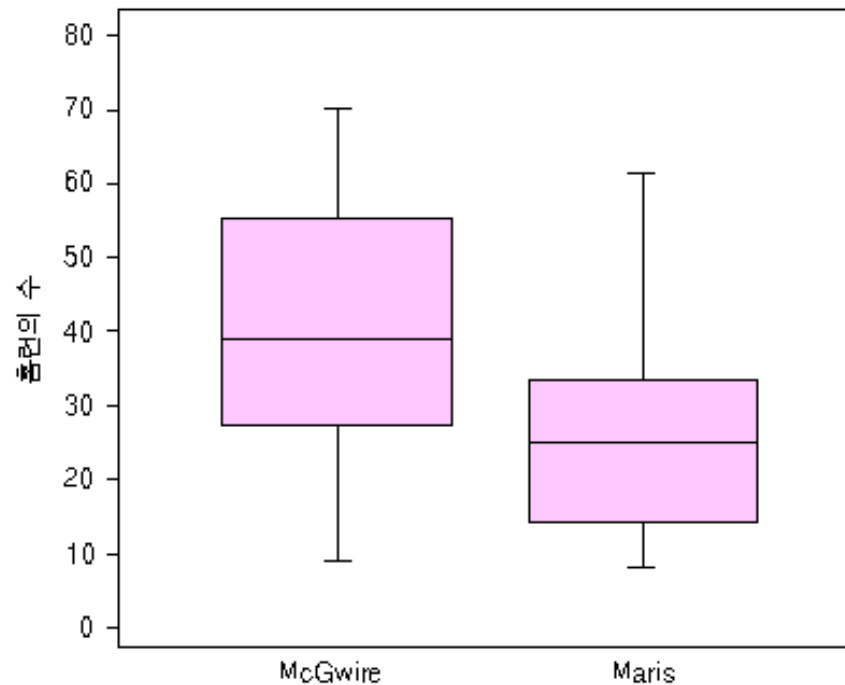
⇒ 분포에 대한 다섯 숫자 요약은 최소값, 제1사분위수, 중앙값, 제3사분위수, 최대값으로 구성된다.



<알아두기> 상자그림

⇒ 상자 그림은 다섯 숫자 요약에 대한 그래프이다.

- ♣ 1단계 : 가운데 있는 상자는 Q1 에서Q3 까지 그린다.
- ♣ 2단계 : 상자 안에 있는 선은 중앙값을 나타낸다.
- ♣ 3단계 : 상자 밖의 선은 가장 작은 관측 값과 가장 큰 관측 값까지 이어진다.



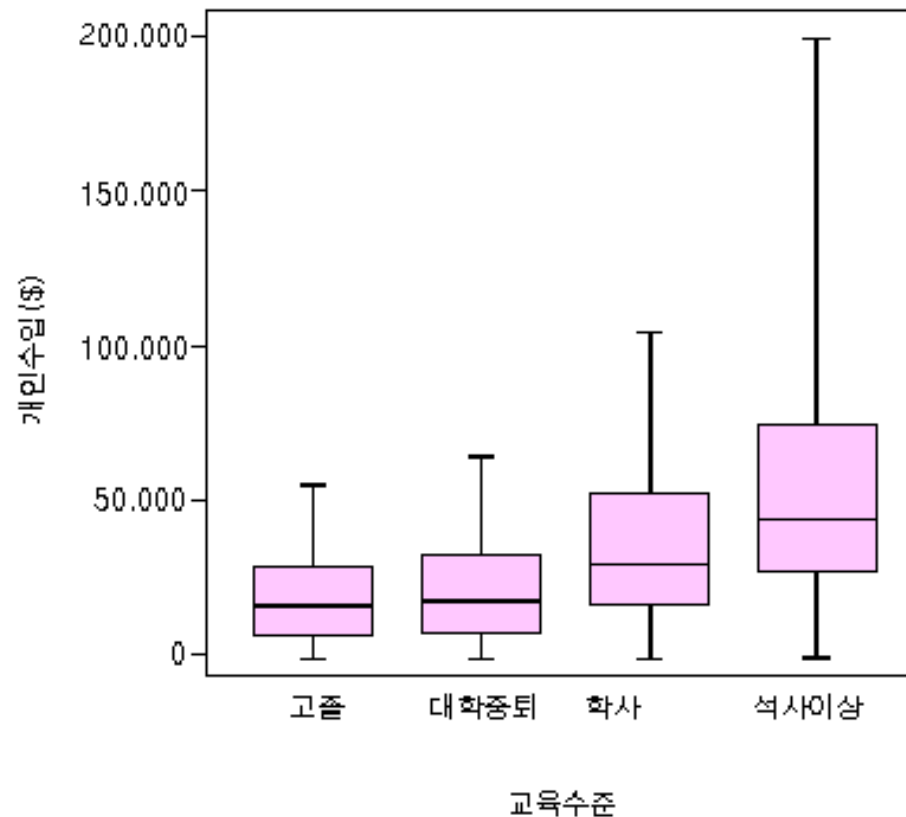
[그림 12.2] McGwire와 Maris의 홈런수를 비교한 상자 그림

<참고> 상자그림에 대하여

- ① 상자 그림은 가로나 세로로 그릴 수 있다.
- ② 그래프에는 숫자 눈금이 포함되어야 한다.
- ③ 상자 그림을 보는 때는 제일 먼저 분포의 중심을 나타내는 중앙값의 위치를 살펴 보아야 한다.
- ④ 제1사분위수와 제3사분위수는 자료의 가운데 절반의 퍼진 정도를 나타내고 있고, 최대, 최소값은 자료 전체의 퍼짐을 나타내고 있다.
- ⑤ 중앙값, 제1사분위수, 제3사분위수를 나타내는 [그림 12.2]의 상자를 통해서 McGwire의 성적이 Maris의 성적보다 더 좋다는 것을 알 수 있다.
- ⑥ 하나 이상의 분포들을 나란히 비교하기 위해서 사용할 수 있다.

◀예제 ▶ 교육과 수입

⇒ 교육 수준에 따라 수입이 얼마나 변화는 지...



[그림 12.3] 교육수준별 수입의 분포를 비교한 상자 그림. 각 그림의 끝은 분포에서 5%와 95%에 있는 관측 값이다.