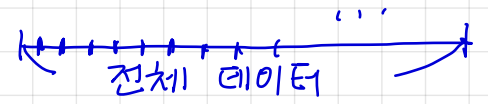
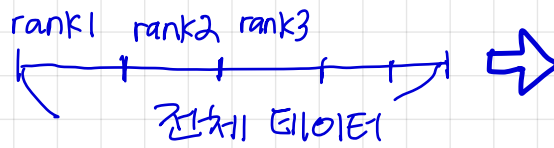
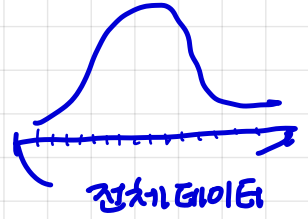


<도수분포표>



<히스토그램>

그러면 거의 대부분
종 모양이 됨.



도수: 각 자급에 속하는 자료의 갯수

계급값: 각 계급을 대표하는 값 (주로 median)

<스터지스 공식>

$$g = 1 + 3.3 \log n$$

of data
g; 얼마나 많은 그룹으로 나눌까

of groups

ex. # of data = 180 → $g = 8.44 \times \approx 9$

— highest value (2738)



range(2579)

— lowest value (159)

class 9

⋮

class 1

$$2579 / 9 = 286.55 \times \approx 287$$

group interval

2531	553	215	1248	159	446	714	486	902
1202	768	472	830	325	460	946	545	1627
1395	347	324	965	235	592	591	553	1703
1365	541	463	817	1169	473	836	510	1202
1250	597	423	893	338	324	736	467	1115
1027	621	249	910	1996	333	796	307	1085
1015	590	352	842	1510	196	836	467	978
983	681	255	814	1576	909	668	440	1028
1002	508	238	785	1539	495	689	369	722
903	503	192	764	1315	1845	784	666	841
891	509	483	642	1637	1658	617	366	943
823	589	2738	612	1054	1243	752	507	703
803	635	1376	596	867	1183	498	311	637
951	471	1419	836	930	1149	969	251	782
695	526	1508	572	788	1235	784	278	596
598	547	1340	617	790	1029	451	269	634
659	334	1496	542	943	983	539	194	653
749	454	1429	523	613	929	580	217	795
757	500	1410	456	652	1147	539	210	555
730	334	1081	315	513	1136	529	877	274

Highest Value	2738
Lowest Number	159
Range	2579
g=	8,443195252 9
Class interval	286,5555556 287

g	Tally	Frequency
159 - 445		31
446 - 732		65
733 - 1019		45
1020 - 1306		18
1307 - 1593		13
1594 - 1880		5
1881 - 2167		1
2168 - 2454		0
2455 - 2741		2

NOTE_ 백분위수: 명확한 정의

데이터의 개수가 짝수라면 (n 이 짝수) 앞선 정의에 따른 경우, 백분위수 값을 정하기 애매하다. 사실 아래 식을 만족하는 순서통계량 $x_{(j)}$ 와 $x_{(j+1)}$ 사이의 어떤 값도 택할 수 있다.

$$100 \cdot \frac{j}{n} \leq P < 100 \cdot \frac{j+1}{n}$$

fixed value
25% 25 30.5%

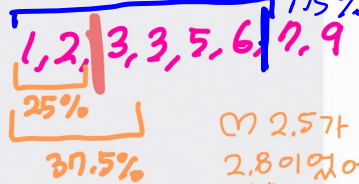
보통은, 백분위수는 아래 수식과 같은 가중평균이다.

$$\text{백분위수}(P) = (1-w)x_{(j)} + wx_{(j+1)}$$

어쨌든 x_j 와 x_{j+1} 사이의 어떤 값이든!

가중치 w 는 0과 1 사이의 값이다. 통계 소프트웨어마다 이 w 를 선택하는 방법이 조금씩 다르다. R의 quantile 함수의 경우, 분위수를 계산하는 9가지 다른 방법들을 제공한다. 데이터 개수가 너무 작지만 일단, 백분위수를 계산할 때 정확도를 걱정할 필요는 없다.

25 번째 백분위수: 2.5



$$8 \times \frac{3}{4} = 6$$

2.5가 아니라 2.8이었어도 25번째 백분위수 ✓

75 번째 백분위수: 6.5

2. 사분위수 : 자료를 크기 순으로 나열한 후 4등분할 때 경계가 되는 세 위치의 점

→ 사분위수 범위(IQR) : $Q_3 - Q_1$

8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48 50 52

1사분위수, Q_1

2사분위수, Q_2

3사분위수, Q_3

$$IQR = Q_3 - Q_1$$

사분위수 예시

Q_1 : 데이터의 25%는 Q_1 보다 작고, 남은 $(100-25)75\%$ 는 Q_1 보다 크다

→ 상자그림 : 사분위수를 요약한 그림

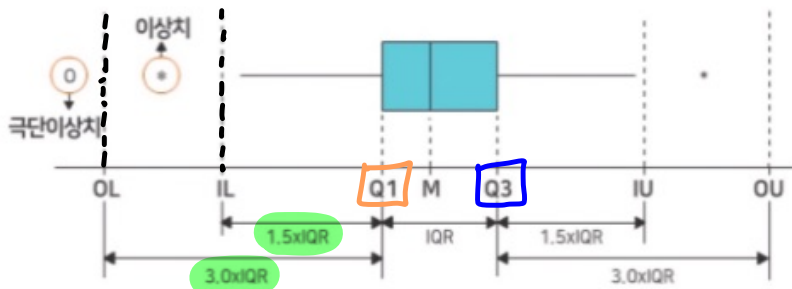
안쪽 윗타리값? $1QR \times 1.5$

바깥윗타리값? $1QR \times 3.0$

인접값? 양쪽 안쪽 윗타리값에 가장 가까운 값

인접값 선 바깥에 있는 것은 이상치로 판단

Q_3 : 데이터의 75%는 Q_3 보다 작고, 남은 $(100-75)25\%$ 는 Q_3 보다 크다



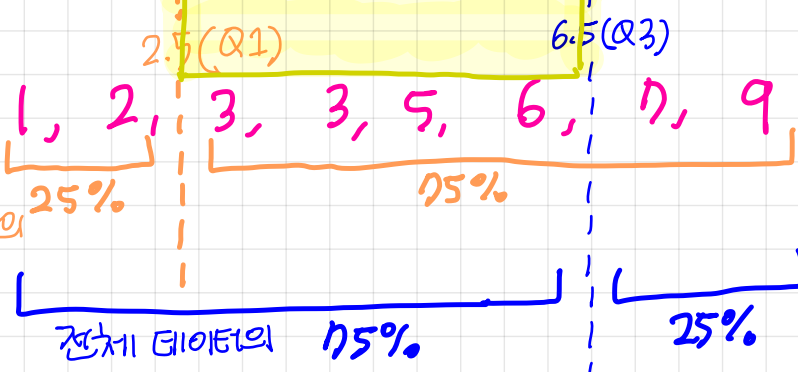
출처: 통계교육원

50%

ex. 데이터가

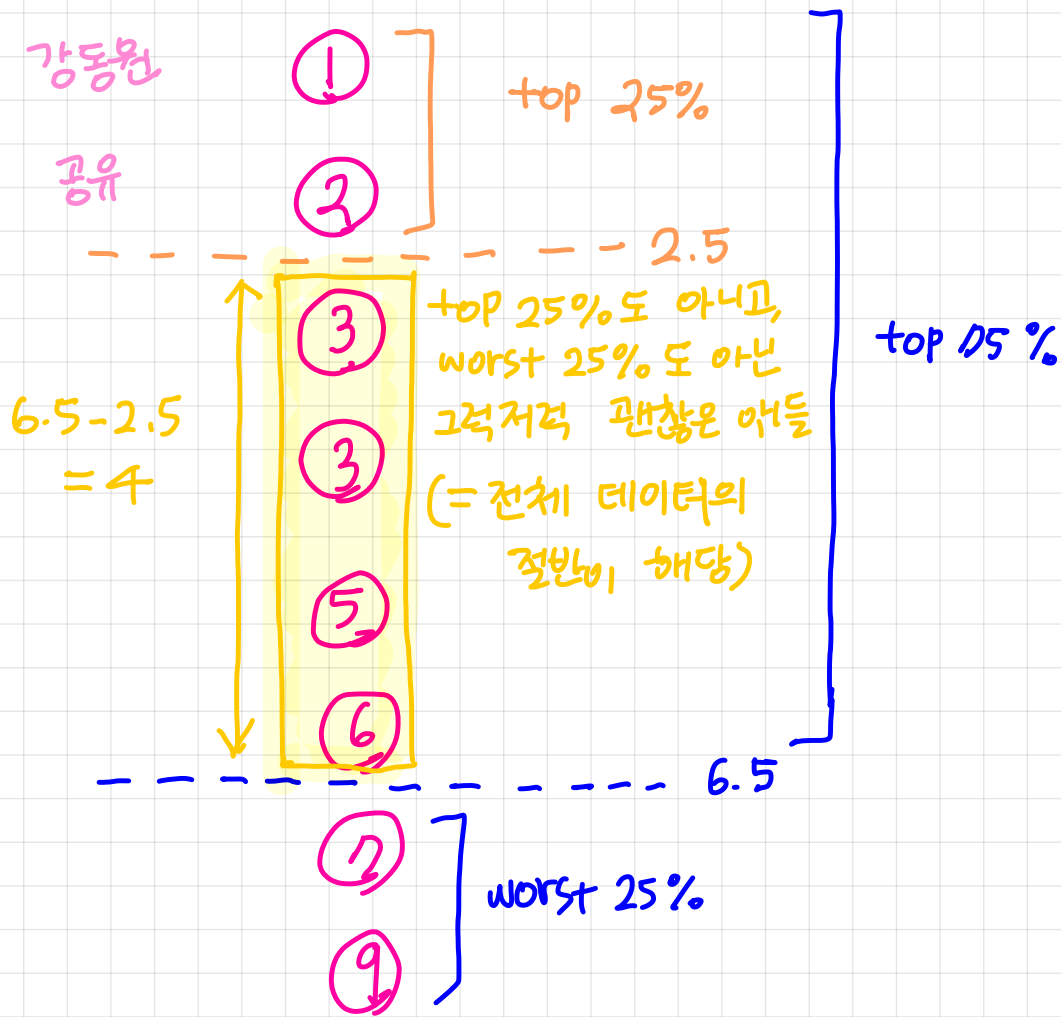
즉 boxplot은 이렇게 그려짐!

(전체 데이터의 절반이 여기에 속함)



라면

제일 좌생긴 순으로 줄세웠다고 하면,





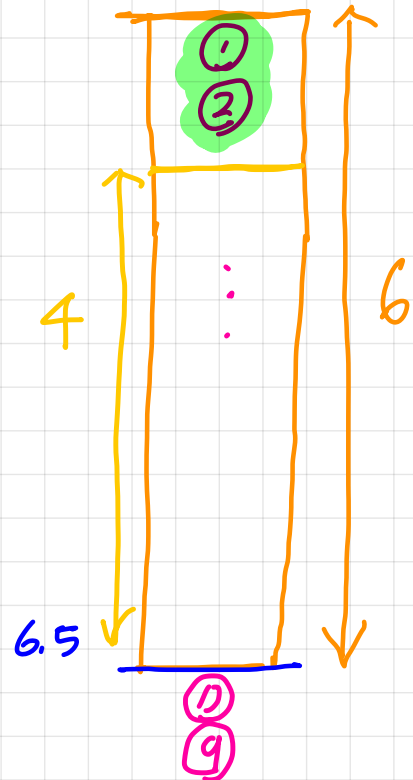
그런데 관측은 애들이
이 폭이 4인 box 안에
다 포함된다면,

⇒ box의 폭을 조금
더 늘리면 "평범"의
기준을 조금 벗어난
데이터들도 웬만하면
담기겠지?

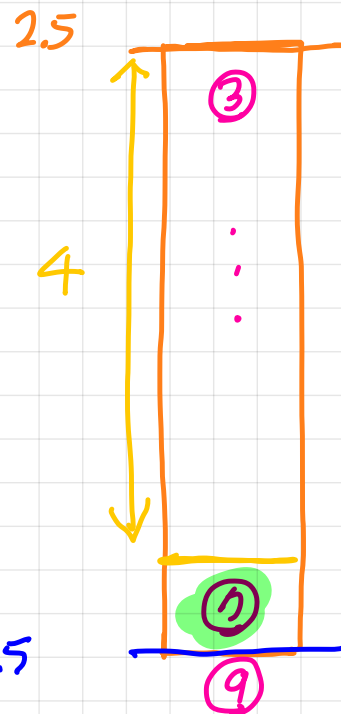
$$4 \times \frac{3}{2} = 6$$

원래 폭보다
50% 증가

now $6.5 - 6 = 0.5$



<늘리는 방법 1 - Q3 값을
fix!>



now $2.5 + 6 = 8.5$

<늘리는 방법 2 - Q1 값을 fix!>

box의 폭을 늘렸다 ⇒ "평범"의 기준을 조금 낮췄다

방법1 box, 방법2 box 中 하나라도 포함되면 "평범" 모임에 쳐주기로!

반대로 여기에도 포함 안 되는 애들 = 이상한 애들

"이상치"

→ 이번엔 평범 box의 크기를 "3배나" 늘리고 동일 과정 반복

마찬가지로 이 느슨해진 기준에도 포함 안 되는 애들 = 완전 이상한 애들

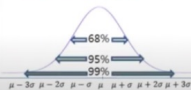
"극단 이상치"

<변동계수>

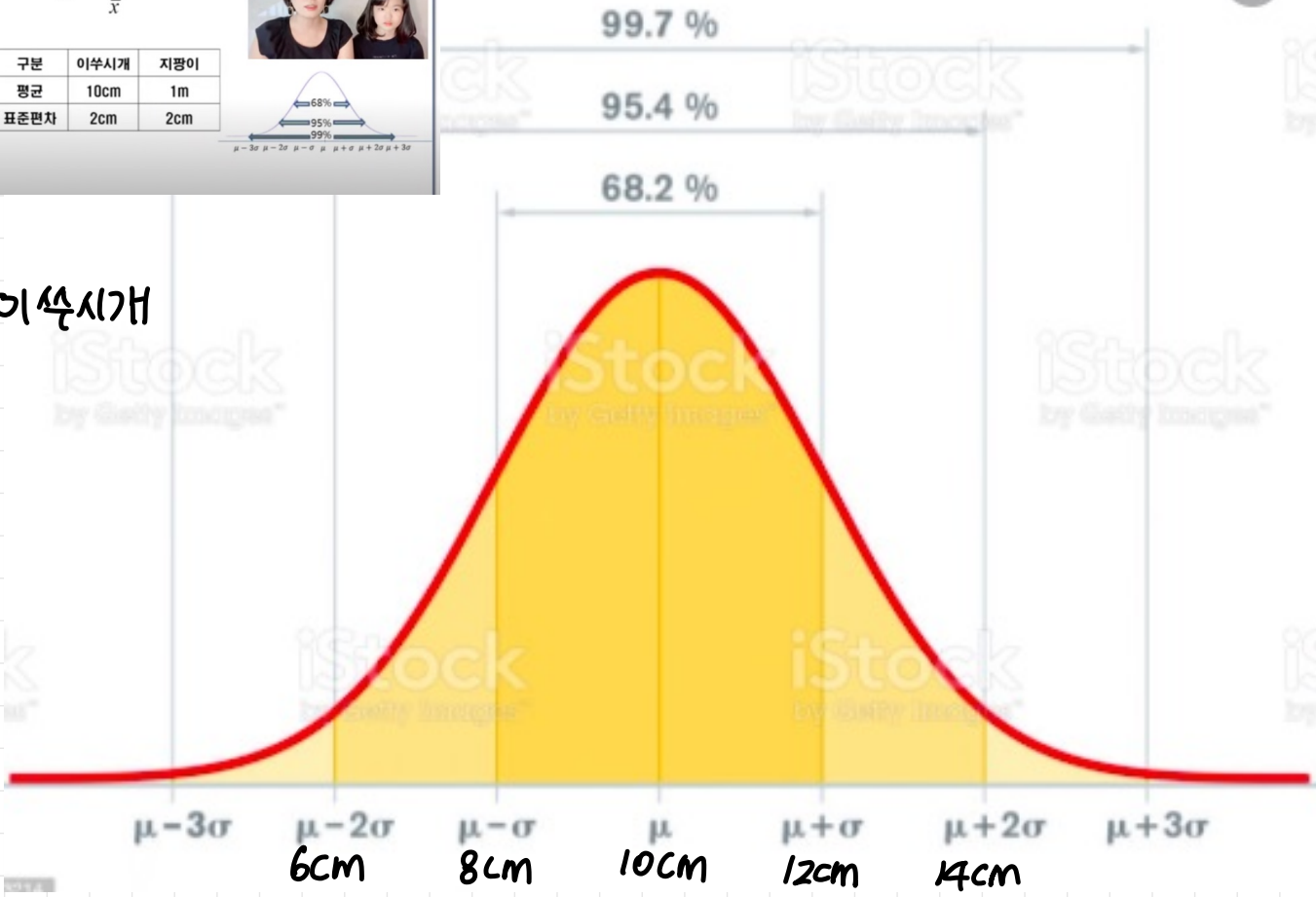
표준편차를 산술평균으로 나눈 것

$$CV = \frac{s}{\bar{x}}$$

구분	이썬시게	지팡이
평균	10cm	1m
표준편차	2cm	2cm



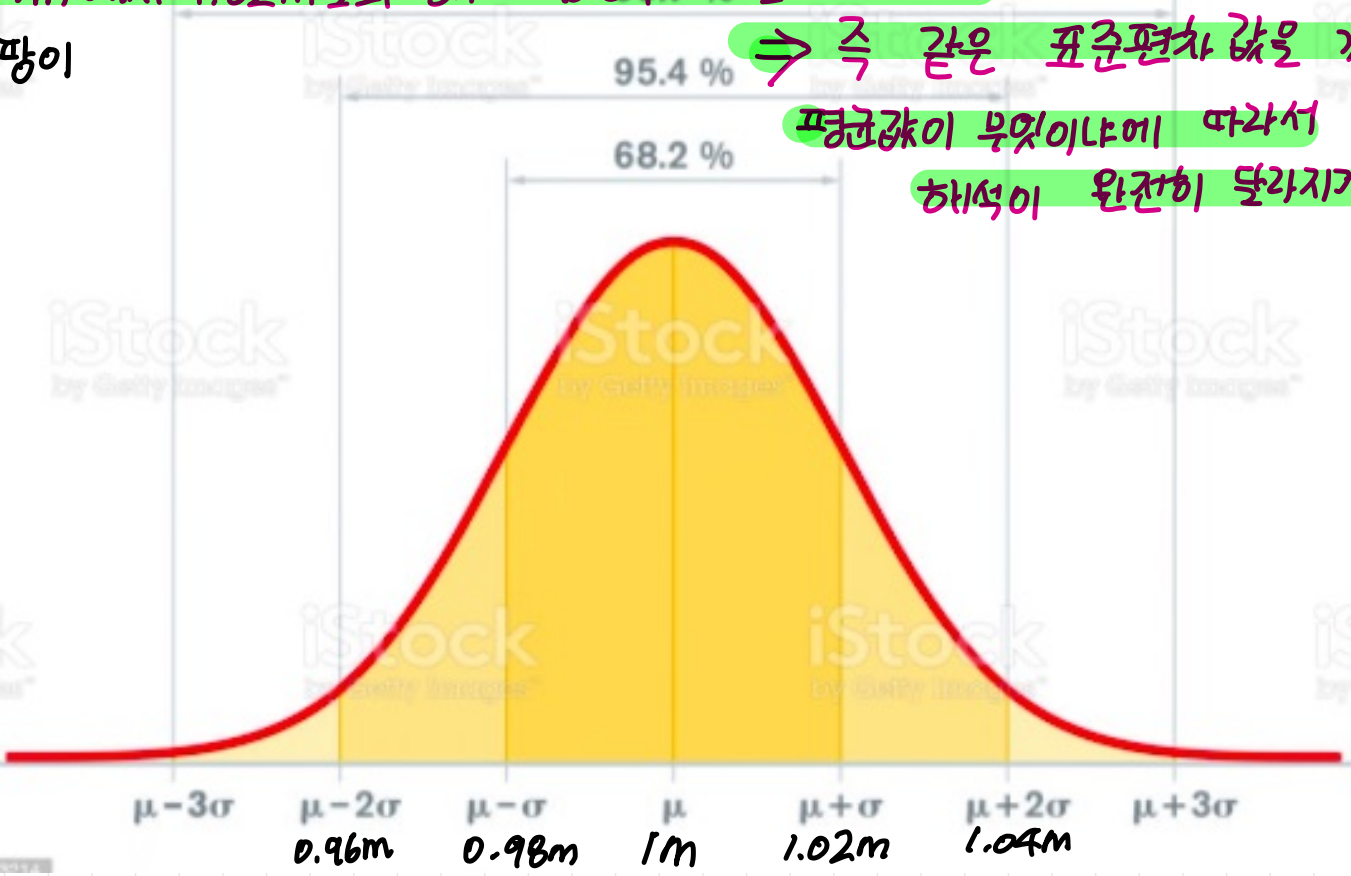
이썬시게



3의 증가 10cm에서 14cm 증가는 엄청나게 큰 변화 (정상품 → 불량품)
0.1m에서 1.02m 3의 증가가 엄청나다고 볼 수 있을까?

지팡이

⇒ 즉 같은 표준편차 값을 가져도
평균값이 무엇이냐에 따라서
해석이 완전히 달라지게 됨



⇒ 평균이 다를 때도 해석할 수 있는 "상대적" 값을 구하자!

$$\text{변동계수 (CV)} = \frac{S}{\bar{X}}$$

↗ 표준편차

이썩시개의 $CV = \frac{2}{10} = 0.2$

지팡이의 $CV = \frac{2}{100} = 0.02$

⇒ 이썩시개가 지팡이보다 "상대적" 변동폭이 10배 더 크다!

<표준점수>

표준점수분포 : 정규분포를 표준화한 것 (평균은 0, 표준편차는 1)

표준점수의 역할

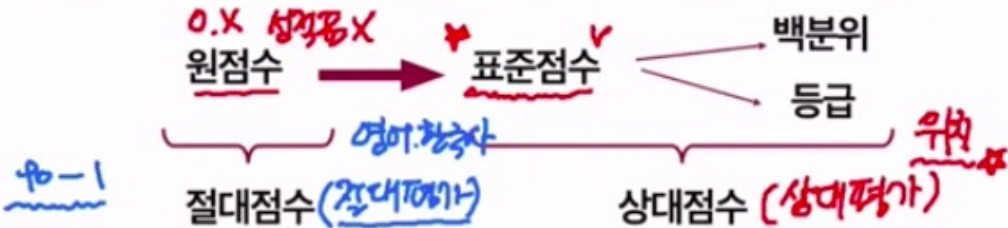
ETOOS 교육평가연구소

✓ 최고점 변화에 따른 수험생의 분포 표현

영역	국어			수학(가)			수학(나)		
학년도	2018	2019	2020	2018	2019	2020	2018	2019	2020
만점	134	150	140	130	133	134	135	139	149
1등급	128	132	131	123	126	128	129	130	135
2등급	123	125	125	120	123	122	126	127	128
3등급	117	117	117	116	117	118	121	119	118
4등급	109	107	107	111	110	110	108	108	106
5등급	98	95	95	102	99	97	91	92	92

성적 지표에 대한 이해

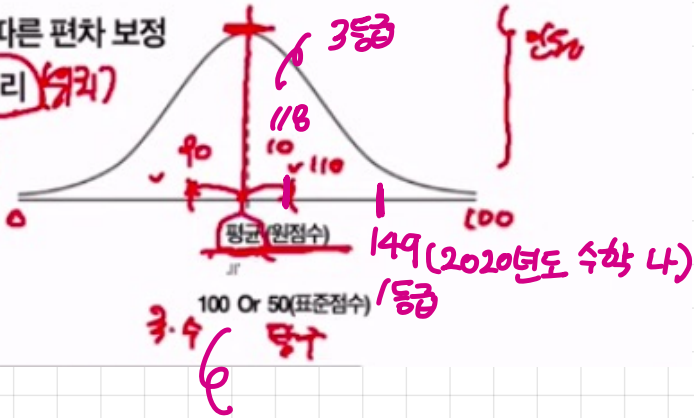
ETOOS 교육평가연구소



상대점수 : 응시 집단의 규모와 질에 따른 편차 보정

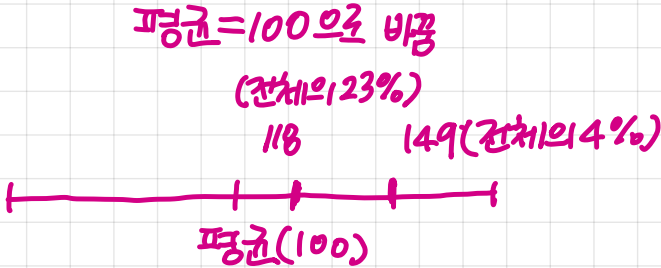
표준점수 : 원점수 평균으로부터의 거리 (거리)

비정규분포 \rightarrow 정규분포 \rightarrow



1등급 : 전체의 4%

3등급 : 전체의 23%



<자유도>

계산의 자유도

∴ 표본평균의 자유도, 표본분산의 자유도 다 해당!

미지수가 많을수록 자유도가 올라가고

(Why? 계산에 다양한 숫자들을 활용할 수 있음)

추정치가 많을수록 자유도가 떨어짐 ex. $\frac{(a-M) + (b-M) + (c-M)}{3} =$ a, b, c가 어떠한 값이라도 M이라는 추정치가 생김으로써 값이 0으로 고정되었음!

* 추정치: 고유한 값이 아니라 다른 값들로부터 추정하여서 탄생한 값

ex. 표본분산의 자유도에서 표본의 크기가 3이면

a, b, c, M 미지수가 4개인 것처럼 보이지만 사실 M은 추정치!

$$\frac{(a-M)^2 + (b-M)^2 + (c-M)^2}{}$$

자유도 (2)

$$\therefore \frac{a+b+c}{3} = M$$

② 샘플 수로 나눈다고 알고 있었는데 더 엄밀하게 말하면 자유도로 나눈 것!

미지수 3개 ∴ 자유도 = 3

추정치 1개 ∴ 자유도 = 2

⇒ 최종 자유도 2