Data Science

데이터

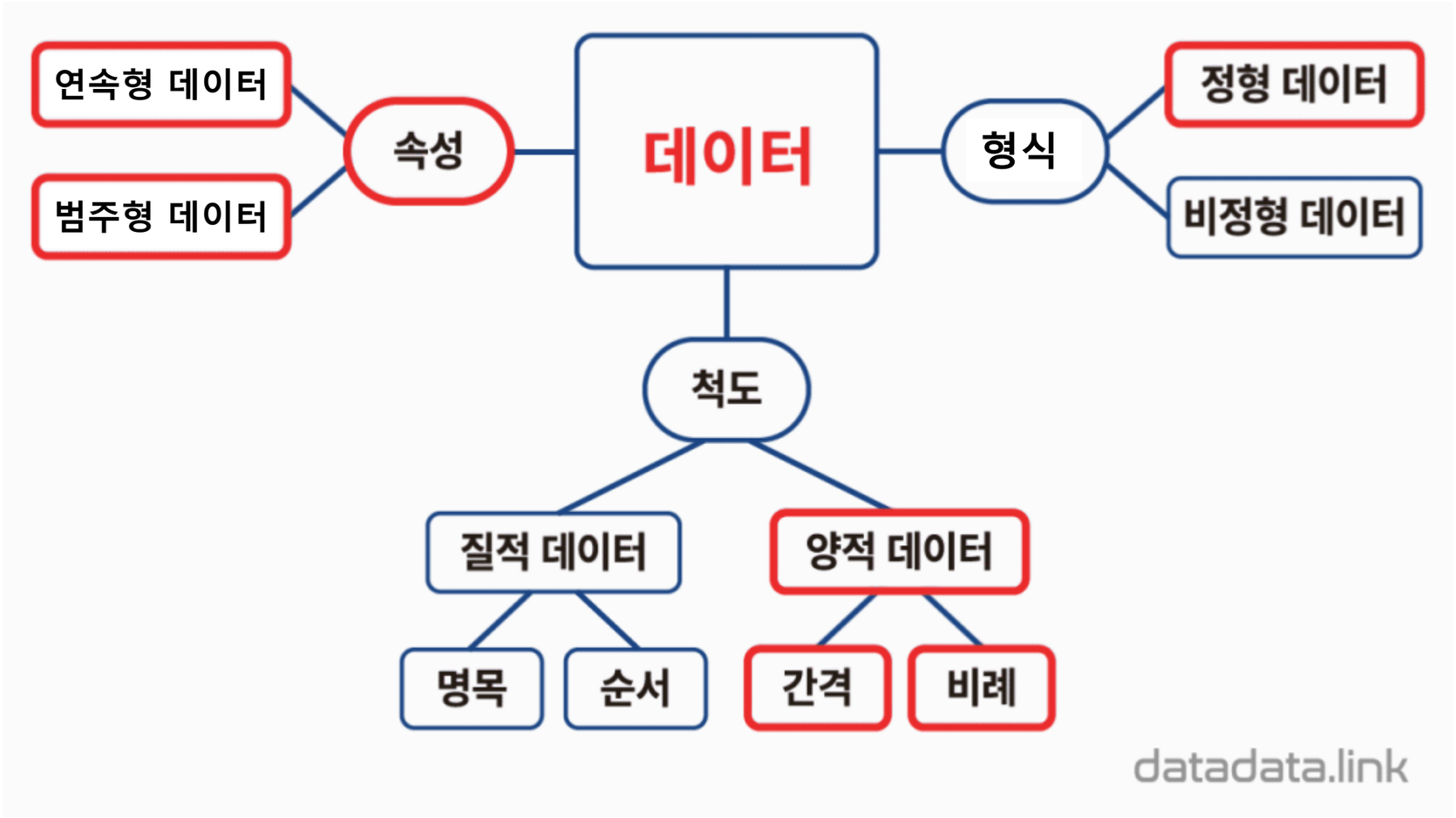
잡초 종류에 따른 쌀 수확량 비교

독립된 두 집단 모평균 비교
독립표본 t검정

학습순서

- 데이터
- 데이터시각화
- 표본통계량

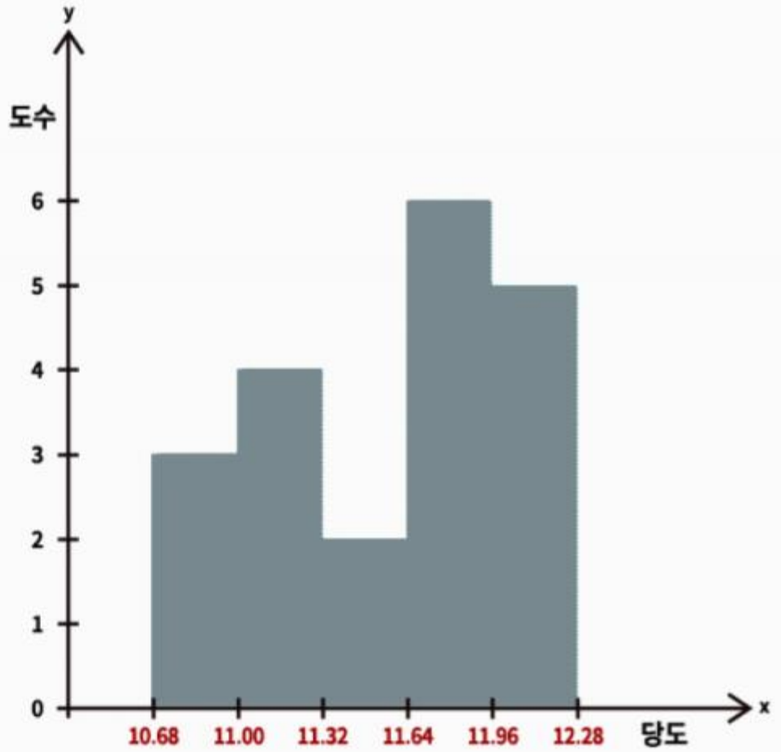
데이터 종류


데이터
속성
연속형 데이터
범주형 데이터
형식
정형 데이터
비정형 데이터
척도
질적 데이터
양적 데이터
명목
순서
간격
비례
datadata.link

히스토그램은 두 변수의 관계를 이어진 면적으로 표현

구간 (이상~미만)	구간중앙값	도수(빈도수)	상대도수
10.68 ~ 11.00	10.84	3	0.15
11.00 ~ 11.32	11.16	4	0.20
11.32 ~ 11.64	11.48	2	0.10
11.64 ~ 11.96	11.80	6	0.30
11.96 ~ 12.28	12.12	5	0.25
합		20	1

도수분포표

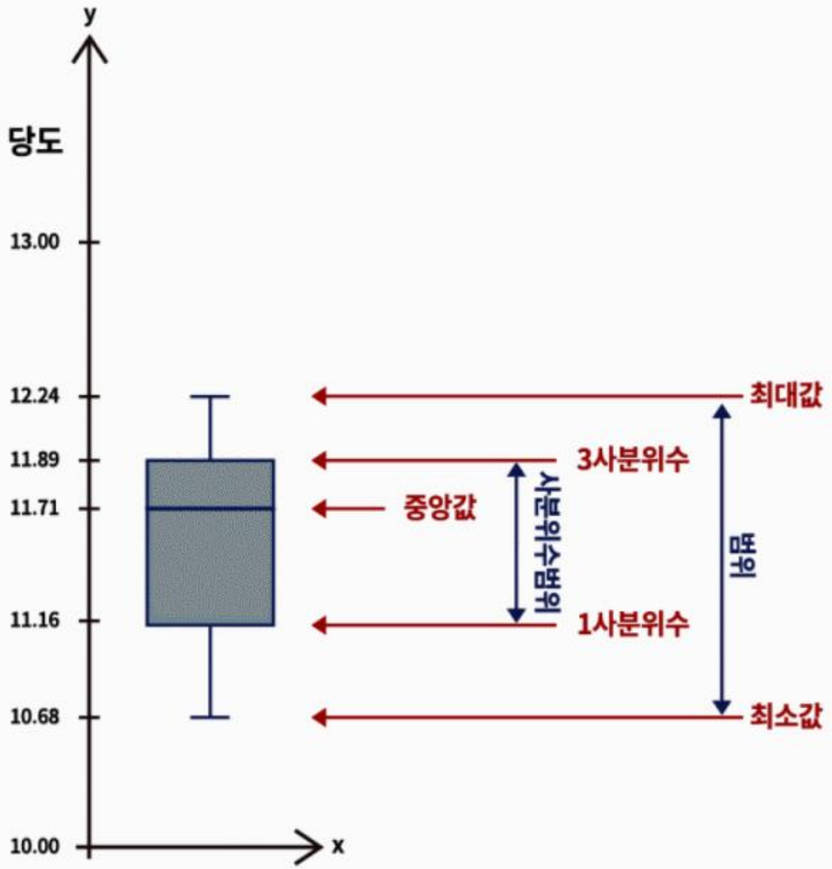


히스토그램

상자그림

당도(내림차순)	사분위	사분위수
12.24	4쿼터	최대값
12.08		
12.03		
12.02		
11.98		
11.89	3쿼터	3사분위수 Q3
11.85		
11.85		
11.80		
11.73		
	11.71	중앙값
11.68	2쿼터	2사분위수 Q2
11.60		
11.41		
11.21		
11.18		
11.16	1쿼터	1사분위수 Q1
11.08		
10.96		
10.75		
10.68		최소값

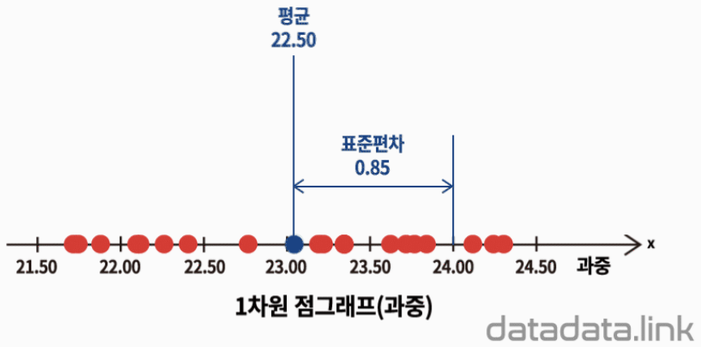
사분위표



상자그림

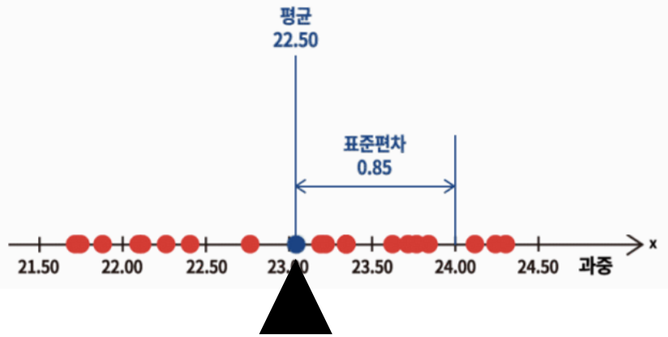
평균은 편차(deviation)의 합이 0이 되게 하는 “편차의 기준”

딸기ID	과중
1	24.21
2	24.28
3	23.88
4	23.85
5	23.73
6	23.17
7	24.14
8	23.63
9	23.37
10	23.37
11	23.37
12	23.19
13	22.78
14	22.25
15	22.13
16	21.86
17	22.10
18	21.72
19	21.69
20	22.42

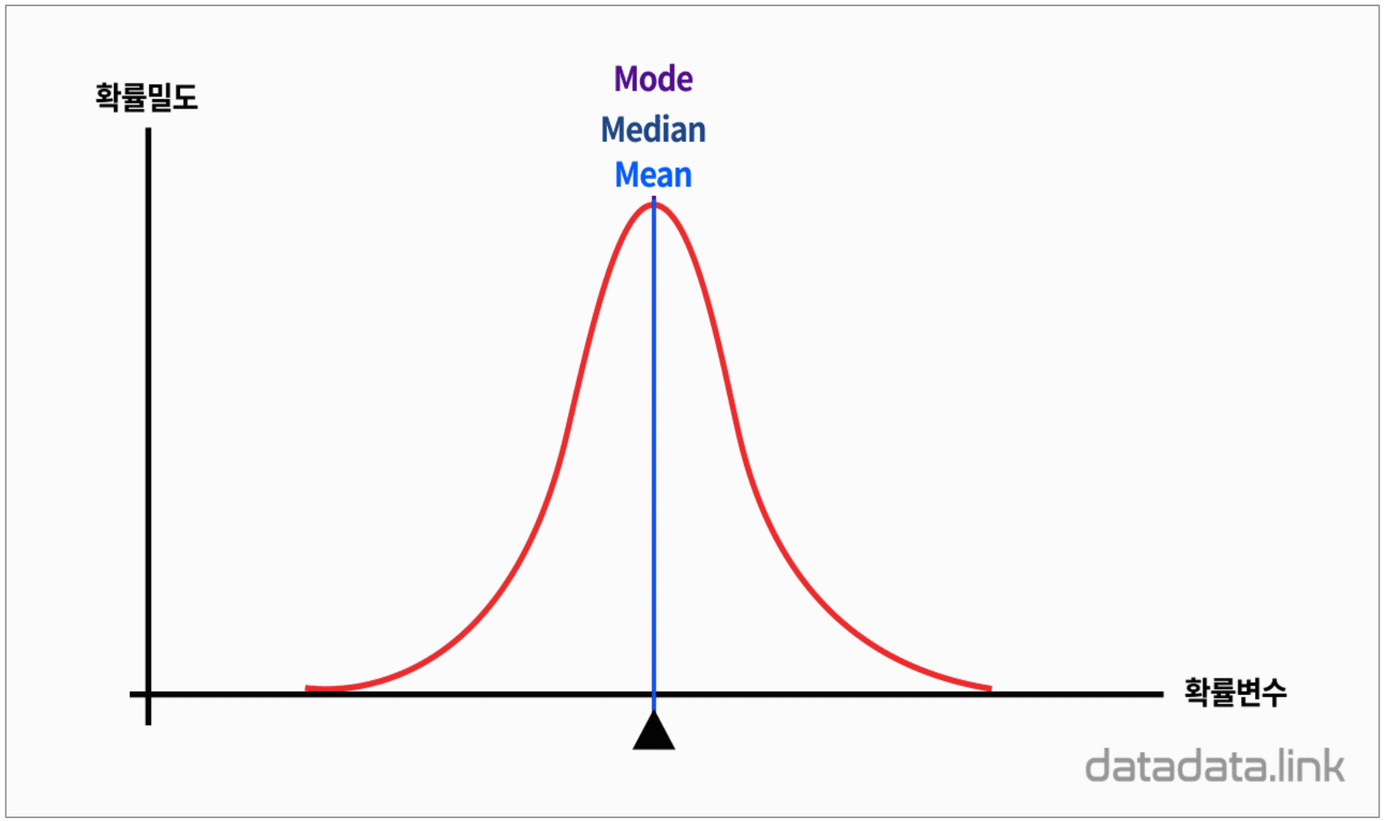


20개의 딸기 과중의 산점도와 평균

평균은 확률질량의 무게중심



평균, 중앙값, 최빈값을 비교하여 확률분포 모양을 추론



- 평균(Mean)은 확률질량의 무게중심
- 중앙값(Median)은 확률질량을 이등분
- 최빈값(Mode)은 확률밀도의 피크를 지정
- 평균, 중앙값, 최빈값이 같은 대표적인 확률분포는 정규분포(Normal distribution)

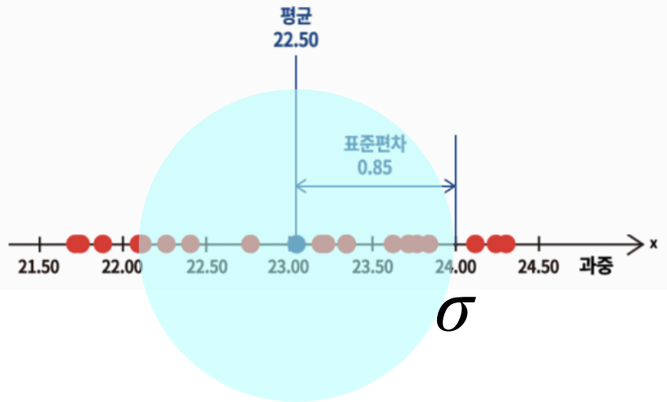
분산은 “편차제곱의 평균” : 편차제곱합을 자유도로 나누어 구함

당도	
평균	
분산	
표준편차	

딸기 번호	당도	당도편차	당도편차제곱
1	11.98	0.42	0.1772
2	12.08	0.52	0.2714
3	12.03	0.47	0.2218
4	11.89	0.33	0.1096
5	12.24	0.68	0.4638
6	11.60	0.04	0.0017
7	12.02	0.46	0.2125
8	11.85	0.29	0.0847
9	11.85	0.29	0.0847
10	11.80	0.24	0.0581
11	11.73	0.17	0.0292
12	11.68	0.12	0.0146
13	11.41	-0.15	0.0222
14	10.96	-0.60	0.3588
15	11.18	-0.38	0.1436
16	11.08	-0.48	0.2294
17	11.16	-0.40	0.1592
18	10.75	-0.81	0.6545
19	10.68	-0.88	0.7726
20	11.21	-0.35	0.1218

표준편차(σ)는 분산(σ^2)의 양의 제곱근

$$\sigma^2=\pi\sigma^2/\pi$$



표본통계량은 표본의 속성을 기술(記述)한 것

딸기 ID	당도	당도편차	당도편차제곱
1	11.98	0.42	0.18
2	12.08	0.52	0.27
3	12.03	0.47	0.22
4	11.89	0.33	0.11
5	12.24	0.68	0.46
6	11.60	0.04	0.00
7	12.02	0.46	0.21
8	11.85	0.29	0.08
9	11.85	0.29	0.08
10	11.80	0.24	0.06
11	11.73	0.17	0.03
12	11.68	0.12	0.01
13	11.41	-0.15	0.02
14	10.96	-0.60	0.36
15	11.18	-0.38	0.14
16	11.08	-0.48	0.23
17	11.16	-0.40	0.16
18	10.75	-0.81	0.65
19	10.68	-0.88	0.77
20	11.21	-0.35	0.12
합계	231.18	0.00	4.19
표본크기	20	20	20
자유도	20	19	19
합계/자유도	11.56	0.00	0.22

표본통계량	
당도 표본평균	11.56
당도 표본분산	0.22

datadata.link

- 표본평균은 표본데이터를 대표하는 대표값 중 하나
- 표본평균은 편차(deviation)의 합을 0이 되게 하는 기준값
- 표본평균은 편차제곱의 합을 가장 작게 하는 기준값

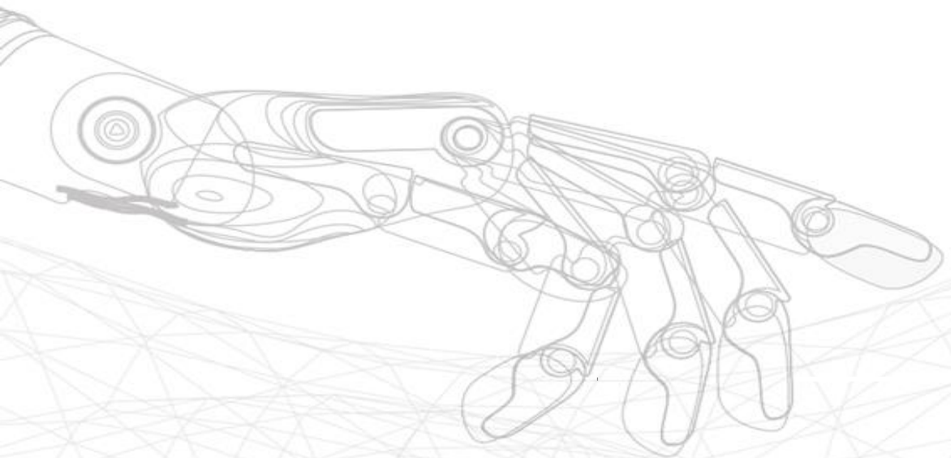
$$\bar{X} = \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right) = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

- 표본분산은 표본평균을 기준으로 하는 편차제곱들의 대표값

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

- 표본표준편차는 표본분산의 양의 제곱근, 데이터와 단위동일

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



감사합니다

www.datadata.link