[2020 혁신성장 청년인재 집중양성 사업]

프로젝트 기반 데이터 과학자 양성과정

# 빅데이터 분석

- 5주차 -

**#통계학의 이해 #분포 # 검정 # 추정**

# Part 1

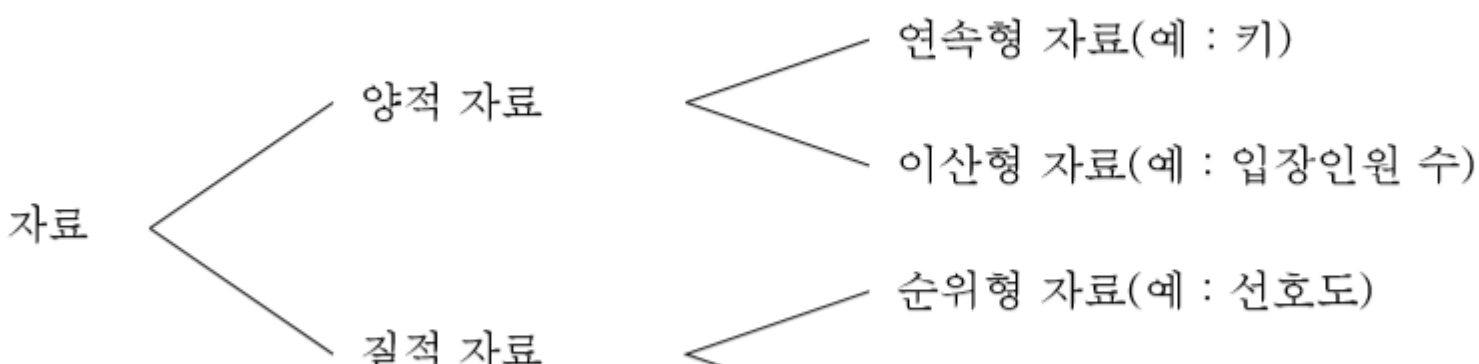
## 통계학의 이해

# 통계학의 이해

## 자료의 이해

- 개체와 변수
- 질적 자료와 양적 자료

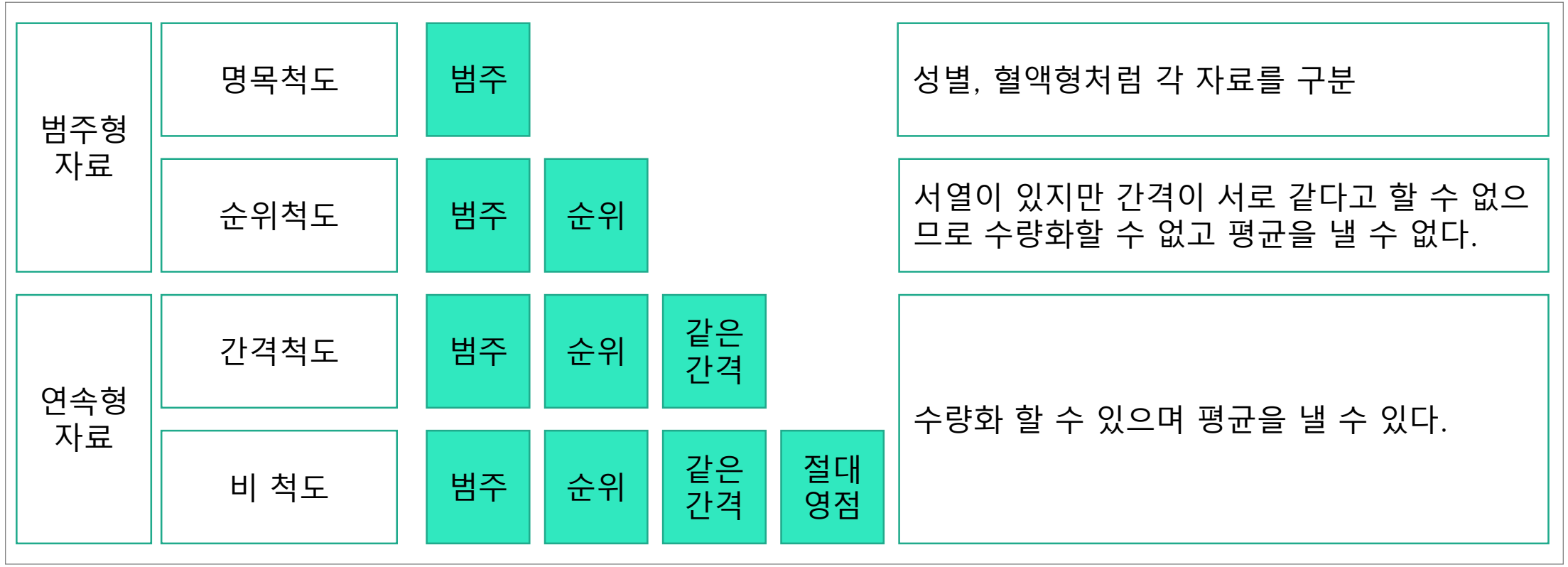
■ 자료의 종류



# 통계학의 이해

## 자료의 분류

- 범주형과 연속형 자료

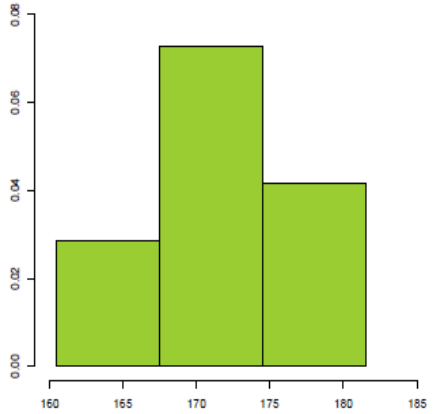


# 통계학의 이해

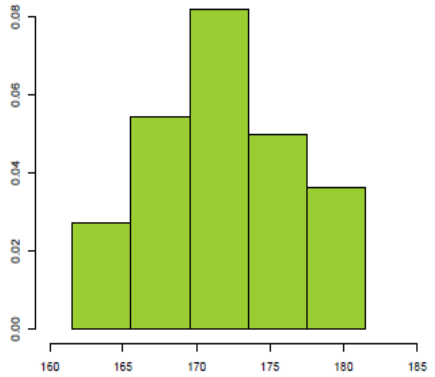
## 자료의 요약

- 범주형과 연속형 자료

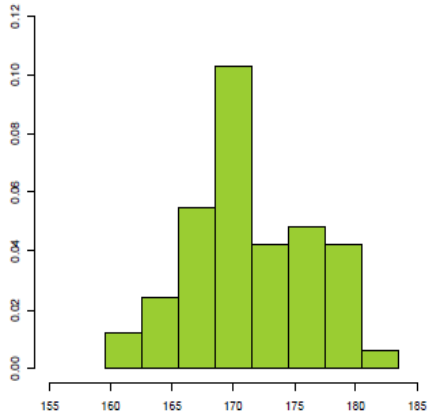
| 계급구간(cm) | 상대도수 | 높이 |
|---|---|---|
| 161.5 이상 165.5 미만 | 0.109 | 0.027 |
| 165.5 이상 169.5 미만 | 0.218 | 0.055 |
| 169.5 이상 173.5 미만 | 0.327 | 0.082 |
| 173.5 이상 177.5 미만 | 0.200 | 0.050 |
| 177.5 이상 181.5 미만 | 0.146 | 0.036 |

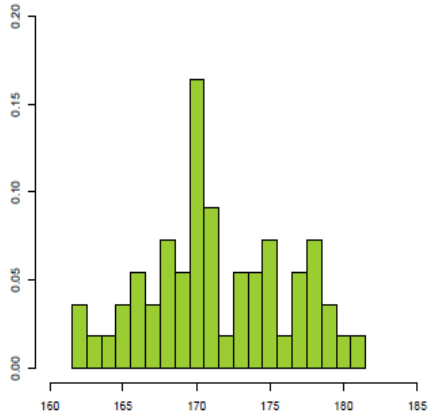



(a) 계급의 수 : 3




(b) 계급의 수 : 5



# 통계학의 이해

**분포**

- 확률이란? 어떠한 결과에 대해 확신하는 정도를 나타낸 수치적 척도

$$P(\mathrm{A}) = \frac{\text{사건 A 에 속하는 결과의 수}}{\text{표본공간에 속하는 결과의 수}}$$

- 확률변수란? 표본공간에 속하는 각각의 변수들에 대해 실수값을 대응하여 나타낸 함수
- 확률분포란? 확률변수가 가지는 값과 그 값을 가질 확률을 정해주는 규칙 또는 관계
- 분산이란? 데이터 펼쳐진 정도를 볼 수 있음.
- 표준편차란? 분산의 양의 제곱근

# 통계학의 이해

## 이항분포

- ✓ 이항분포란? 베르누이 시행을 N번 반복한 경우 성공횟수를 확률변수 X 라 하면, 이 확률변수 X 의 확률분포는 이항분포를 따르게 된다

$$X \sim \mathrm{Bin}(n, p)$$

# 통계학의 이해

## 가설검정

**용어 정의**

귀무가설(null hypothesis, $H_0$) : 기존의 주장

대립가설(alternative hypothesis, $H_1$) : 증명을 필요로 하는 새로운 주장

기각치(critical value) : 귀무가설 $H_0$를 기각하는 기준값

기각역(critical region) : 기각치를 기준으로 귀무가설을 기각할 수 있는 범위

검정통계량(test statistic) : 가설 검정을 위한 모수의 점추정량

유의수준(significance level, $\alpha$) : 귀무가설 $H_0$를 잘못 기각할 확률

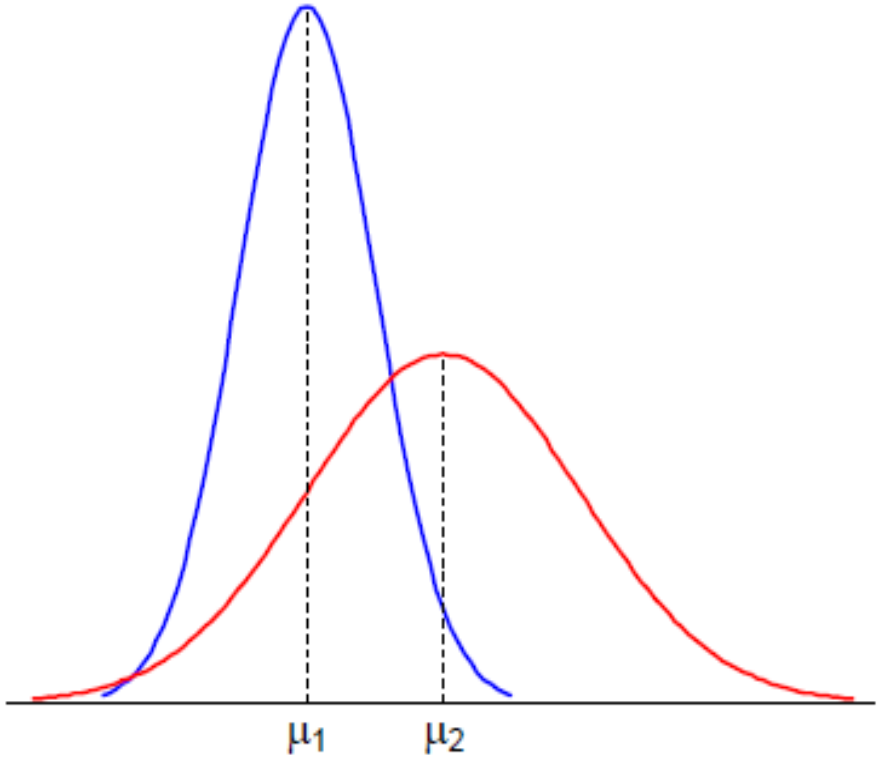
**P-value**

- $H_0$가 참일 때 검정통계량이 실제 표본을 통한 검정통계량의 관측값과 같거나 더 지나친 값(크거나 작은 값)을 취할 확률이다.
- $p$-값$< \alpha \Rightarrow$ 귀무가설을 기각한다.

# 통계학의 이해

**정규분포**

- 정규분포란? 모든 분야에서 가장 중요하게 생각되는 대표적인 연속확률변수
- 표준정규분포란? 정규분포 중에서 평균이 0이고, 분산이 1인
- 정규분포를 말함



# 통계학의 이해

## 중심극한정리

- 모집단의 확률분포가 연속형 이거나  이산형, 혹은 대칭이거나 비대칭에 상관관계 없이 표본의 크기가 충분히 크다면 표본평균의 확률분포가 근사적으로 정규분포를 따르게 된다.

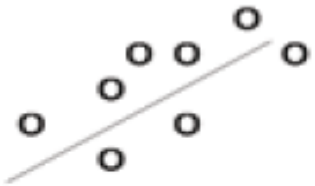
평균이 $\mu$이고 분산이 $\sigma^2$인 모집단으로부터 추출한 크기 $n$의 확률표본의 표본평균 $\overline{X}$는 표본의 크기가 큰 경우 (보통 30 이상), 근사적으로 평균이 $\mu$이고 분산이 $\sigma^2/n$인 정규분포를 따르게 된다.

# 통계학의 이해

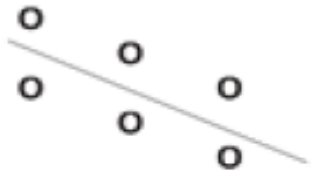
**상관분석**

- 두 변수 사이의 선형관계를 산점도로부터 대략적으로 알 수 있다. 하지만 이를 하나의 수치값인 통계량으로 나타낸 것이 상관계수

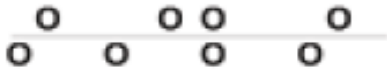
$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$



(1) $0 < r < 1$



(2) $-1 < r < 0$



(3) $r = 0$

# 통계학의 이해

## T test

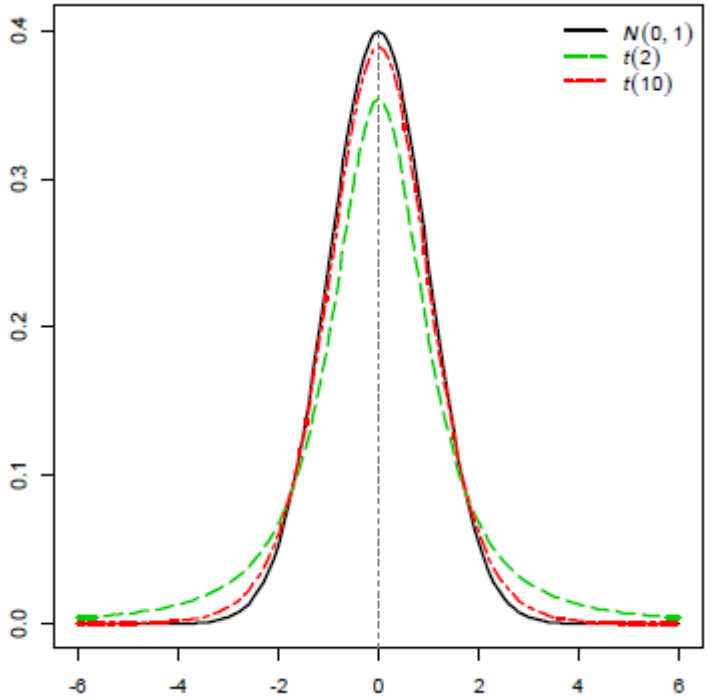
평균이 $\mu$이고 분산이 $\sigma^2$인 정규모집단으로부터 추출된 크기 $n$의 표본을 $X_1, X_2, \cdots, X_n$이라 할 때, 이들에 대한 표본평균과 표본분산을 각각

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

이라 정의하면, 확률변수

$$t = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

는 자유도(degree of freedom, $df$)가 $df = (n-1)$인 $t$-분포를 따른다 한다.

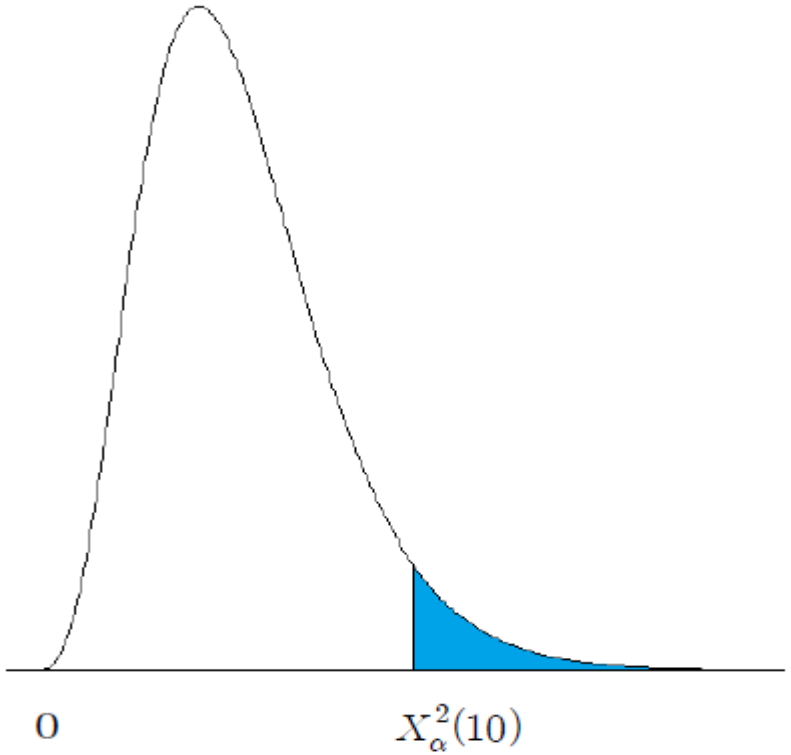


# 통계학의 이해

**카이제곱 검정**

- 귀무가설에 언급되는 내용이 적합한지 알아보는 척도로 피어슨이 처음 제시
- 카이제곱 분포를 가정하는 범주형 데이터에 대한 통계적 검정

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(\text{관찰도수}-\text{기대도수})^2}{\text{기대도수}} = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$




[자유도가 10인 카이제곱 통계량]