



STATISTICS

Essential



By Hapro



1

통계

통계학은 데이터에서 의미를 찾아내는 방법을 다루는 학문

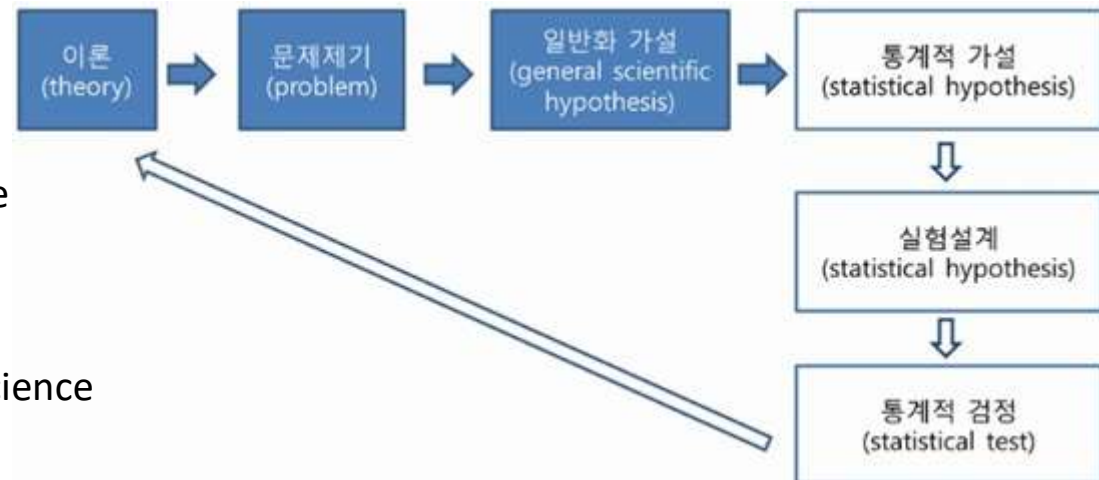
통계학

- 목적

- 이론의 검정
- 데이터로부터 의미 추출
- 사회학의 과학화
- 확률적 과학화

- 분야

- Data Information Science
- Data Science
- Decision-making Science
- Statistical Information Science
- Statistical Science
- Informative Statistical Science
- Information Management Science



용어

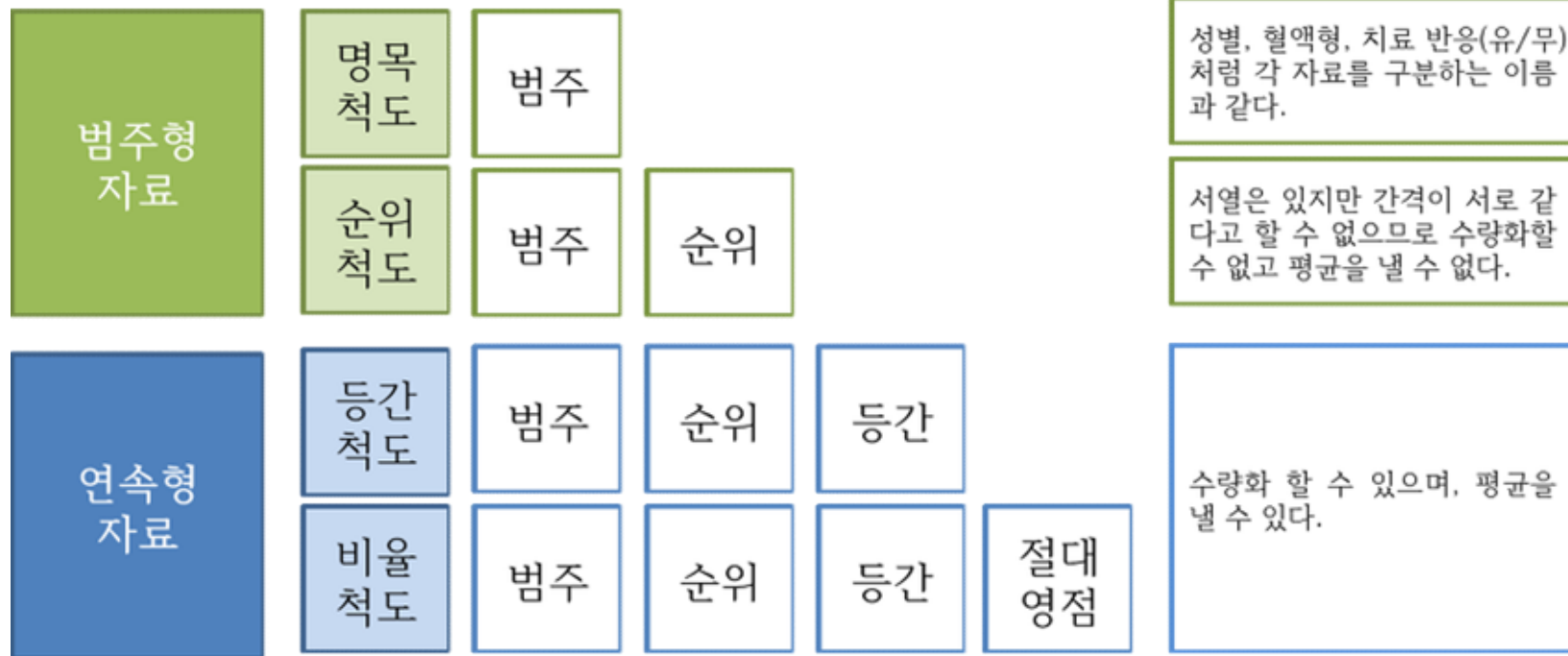
- 양적 자료
 - 숫자로 얻을 수 있는 또는 표현될 수 있는 데이터: 우위 있음
 - 키, 몸무게
- 질적 자료
 - 숫자로 표현할 수 없는 범주 데이터: 우위 없음
 - 주민번호, 성별, 지역
- 개체 요인 변수
 - 개체(Item): 관찰 대상- 학생, 국민, 동물
 - 요인(Factor): 개체의 특성중 연구자가 관심있어 하는 부분- 키, 직무, 성별
 - 변수(Variable): 요인을 구성하고 있는 관측된 요소- 사무원/영업사원, 남/녀

척도

- **척도 (Scale)**
 - 어떠한 대상의 특성을 단위를 사용하여 정량화(수로 표현)
 - **대상 특성의 단위**
- **명목 척도 nominal scale**
 - 이름뿐인 척도 성별, 직업, 거주지
 - 숫자로 표현되지만 숫자가 수량의 의미는 없음 남-여?
 - 평균 처리 불가
- **순위 척도 ordinal scale**
 - 숫자가 순위를 나타냄: 절대적 차이 아님 성균관대-한양대
 - 우수학생 순위, 복지 순위
 - 평균 처리 불가
- **등간 척도 interval scale**
 - 관측 대상이 지닌 속성의 차이를 양적인 차이로 측정하기 위하여 척도간 간격을 균일하게 분할하여 측정하는 척도
 - 키, 몸무게, 온도, 리커트 척도
 - 절대 0점은 없고 상대적 0점
- **비율 척도 ratio scale**
 - 절대 영점이 있는 등간 척도: 음의 척도는 존재하지 않음
 - -1km, -2kg

자료

- 자료의 종류



그림으로 이해하는 닥터배의 술술 보건의학통계, 배정민, 한나래

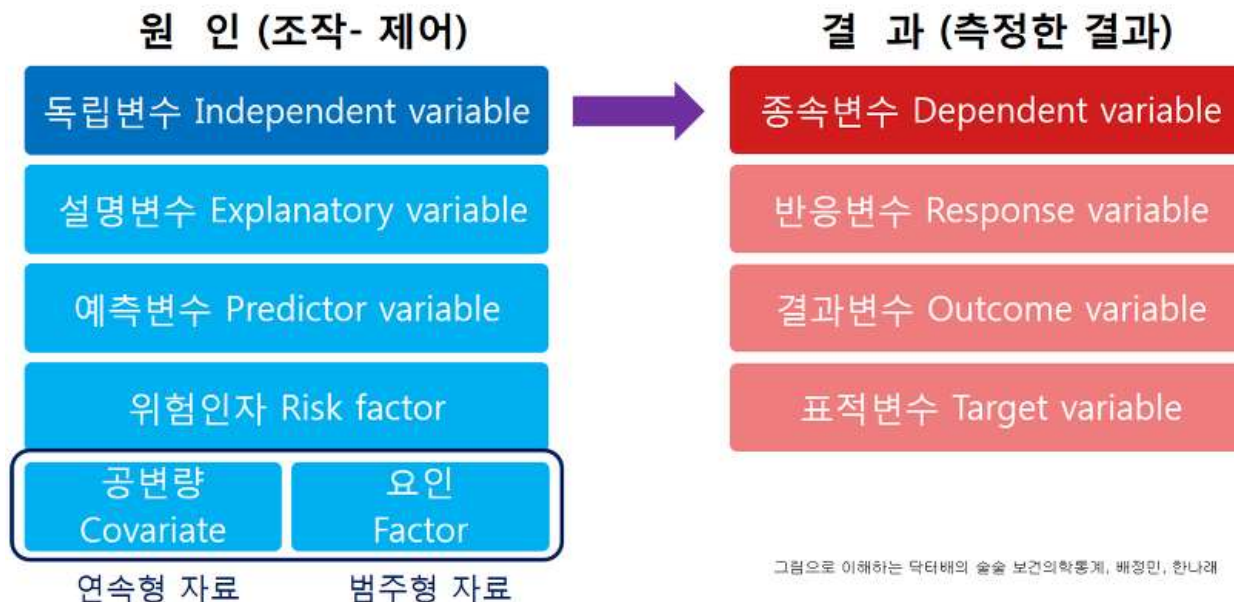
독립변수 vs 종속변수

- 변수
 - 요인을 구성하고 있는 관측된 요소
- 독립변수 Independent variable
 - 연구자가 의도적으로 변화시키는 변수
 - 다른 변수에 영향을 받지 않는다는 뜻
 - 독립변수는 입력값이나 원인
 - 종속변수는 결과물이나 효과
- 파생어
 - 예측변수 predictor variable
 - 회귀자 혹은 회귀변수 regressor
 - 통제변수 controlled variable
 - 조작변수 manipulated variable
 - 노출변수 exposure variable
 - 리스크 팩터 risk factor
 - 반응 변수 (Response variable)
 - 결과 변수 (Outcome variable)

종속변수

- 종속변수

- 연구자가 독립변수의 변화에 따라 어떻게 변하는지 알고 싶어하는 변수
- 종속의 영향을 통해 원리를 알고자 하는 변수
- 독립 변수와 종속 변수 모두
 - 연속형 자료 (예: 몸무게, 키, 성적 등) 가능
 - 범주형 자료(지역, 성별, 학력 등) 가능



그림으로 이해하는 닥터배의 술술 보건의학통계, 배정민, 한나래

기술통계 vs 추리 통계

- **기술 통계 (Descriptive statistics)**
 - 수집한 데이터를 요약 묘사 설명하는 통계 기법
 - 데이터의 집중화 경향 설명 기법 Central tendency
 - 평균 (mean)
 - 중앙값 (median)
 - 최빈값 (mode)
 - 분산 경향 설명 기법 Variation tendency
 - 분산도 (Variation)
 - 표준편차 (standard deviation)
 - 사분위 (quartile)
- **추리 통계 (Inferential statistics)**
 - 수집한 데이터를 바탕으로 추론 예측하는 통계 기법
 - 회귀
 - 확률 통계

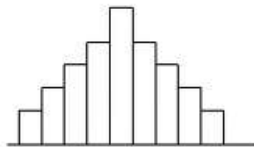
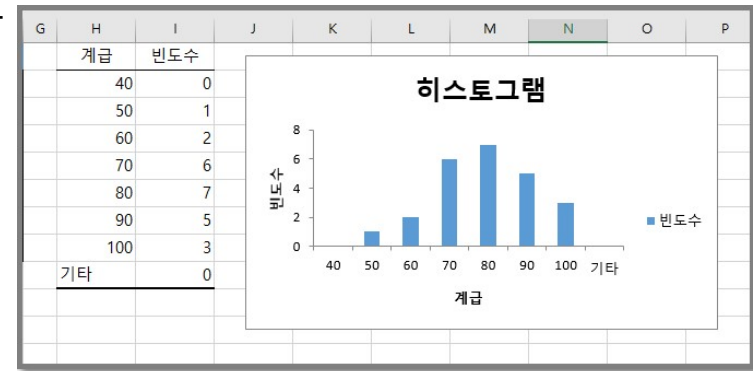
도수분포표 히스토그램

- 도수 분포표 (Frequency table)

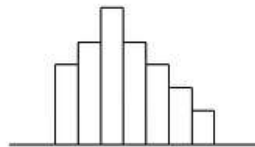
- 특정 구간에 속하는 자료의 개수를 나타내는 표
- 자료의 개수와 최대/최소 나눌 구간의 수 파악
- 구간폭 = (최대-최소) / 구간수
- 구간 경계 값 구하기
- 구간별 자료의 개수(도수)를 적음

- 히스토그램

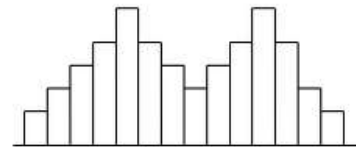
- 도수분포표의 시각화



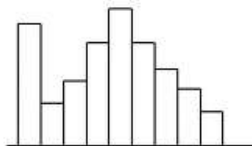
정규분포



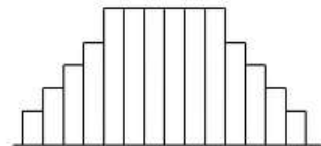
특정한 값보다 작은 값을 모집단(표본)으로부터 제거한 경우



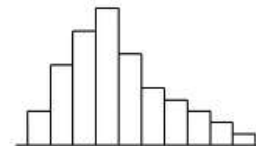
두 모집단이 혼합된 경우



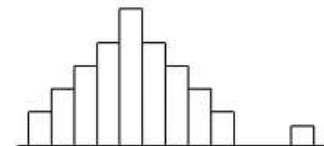
한계값에서 벗어난 값을 모두 한계값으로 대신한 경우



여러개의 모집단이 혼합된 경우



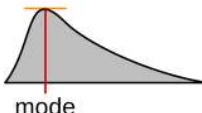
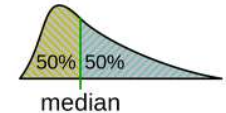
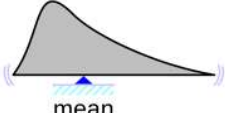
비대칭 분포



이상값이 존재한 경우

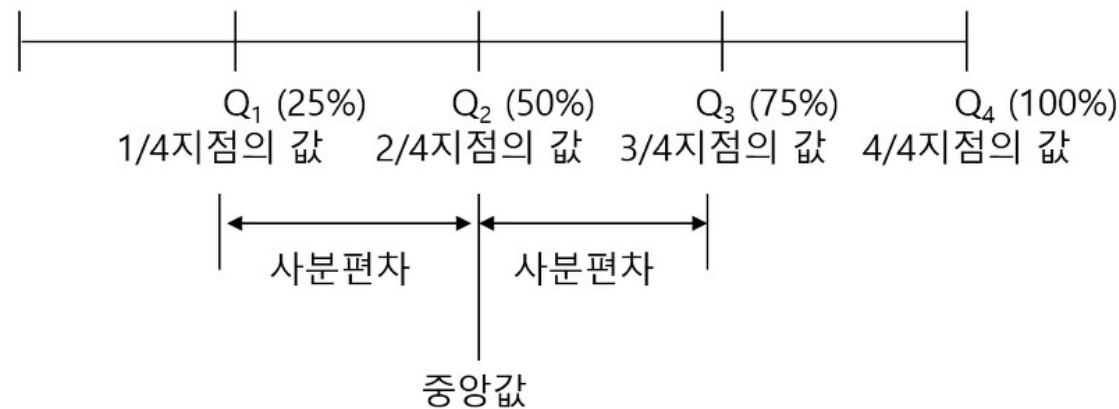
중심화 경향

- **중심화 경향**
 - 수집한 자료 전체를 대표하는 값이 무엇인지 나타내는 통계 (대표값)
- **평균값**
 - 자료를 모두 더해서 전체 자료의 갯수로 나눈 값 (정량적 자료의 대표값)
- **최빈값**
 - 수집한 데이터 중 그 빈도가 가장 많이 나타나는 데이터 (명목 자료의 대표값)
- **중앙값**
 - 자료를 크기 순으로 정렬했을 때, 중앙에 위치하는 값 (순위 자료의 대표값)

	최빈치(mode)	중앙치(median)	산술평균(mean)
의미	<p>• 가장 빈번하게 나타나는 값</p>  <p>mode</p>	<p>• 자료를 크기 순으로 나열했을 때, 중앙에 위치하는 값</p>  <p>median</p>	<p>• 자료를 모두 더해서 자료의 개수로 나눈 값</p>  <p>mean</p>
특징	<p>• 명목자료에서는 최빈치가 대푯값이다.</p>	<p>• 서열자료의 경우 평균을 사용할 수 없으므로 중앙치를 사용한다.</p>	<p>• 일부 극단적인 값들에 크게 영향을 받는다. • 수학적 연산에 의해 계산되므로 수리적인 조작이 용이하다.</p>
예	<p>유행하는 가방 인기 투표</p>	<p>학교 석차 100명 중 50 등</p>	<p>년간 평균 강우량 기말 고사 평균 점수</p>

분산 경향

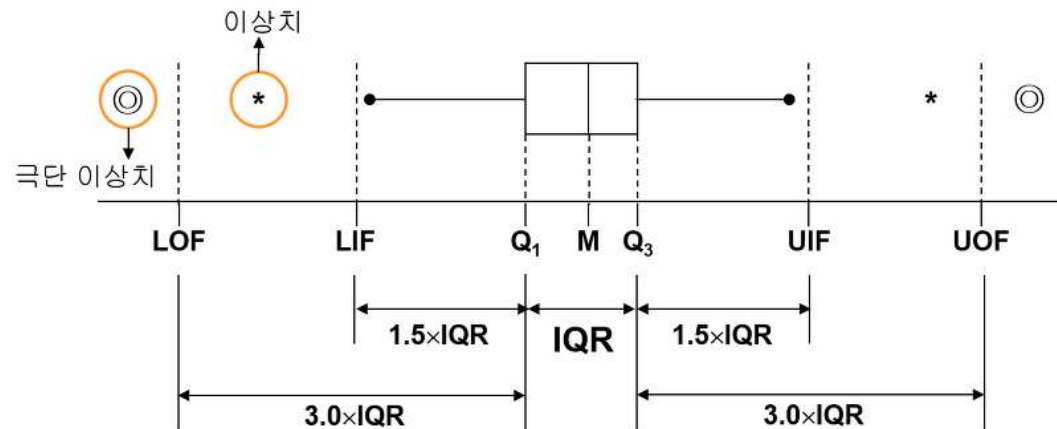
- **분산도**: 데이터가 어떻게 분포되어 있는지를 설명하는 통계치
- **범위**: 자료의 최대값에서 최소값의 차이
- **사분편차**: 자료를 크기순 정렬 후 전 자료 분포의 중앙부에서 전자료의 50%를 포함한 범위의 반



$$\text{Quartile deviation} = \frac{Q_3 - Q_1}{2}$$

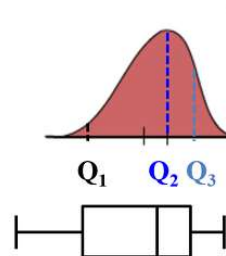
편향 경향

- IQR: Inter Quatile Range $Q_1 \sim Q_3$ 간의 구간
- Outlier : Q_1 바깥 Q_3 바깥 Box whisker (수염) 길이는 보통 IQR의 1.5배
 - 벗어나면 이상치
 - $3IQR$ 이상은 극단치
 - 일반적 표현

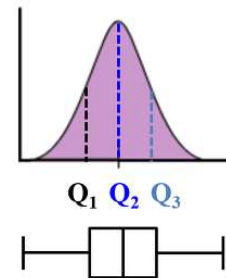


- 비대칭도
 - 왜도 skewed

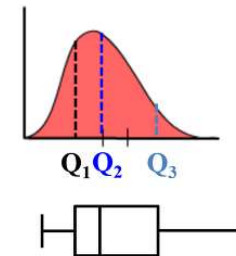
Negatively-Skewed



Symmetrical

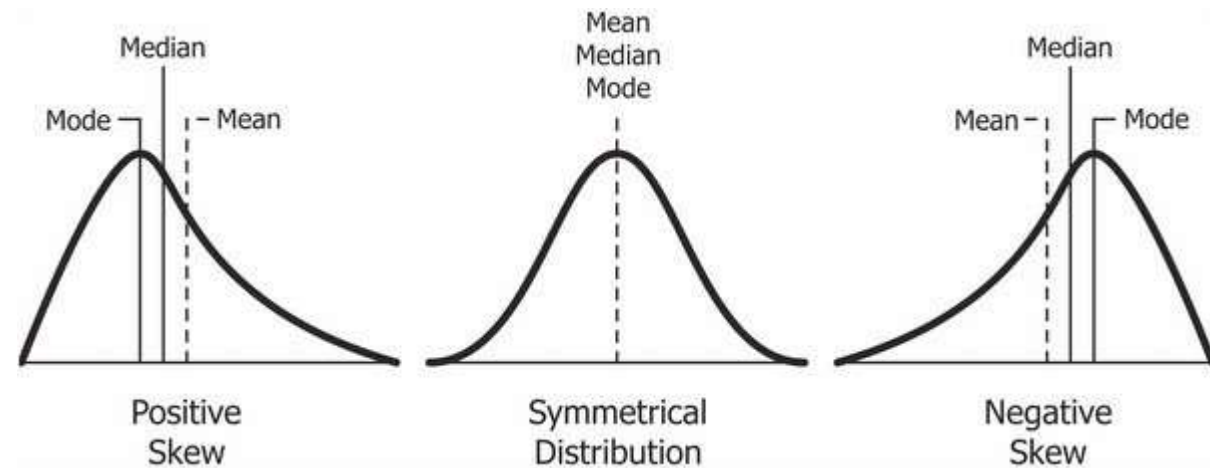


Positively-Skewed



왜도

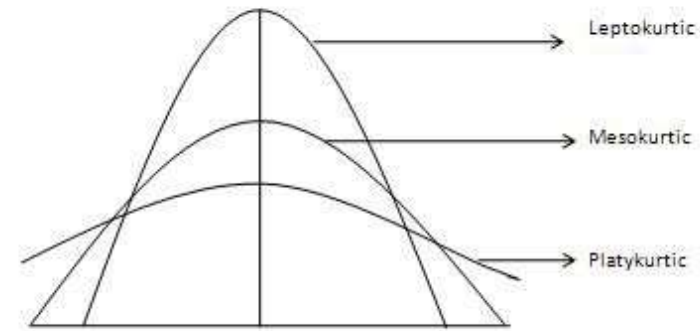
- 왜도 Skewness
 - symmetrical bell curve



- skewness 0 : 대칭
- skewness -0.5 ~ 0,5 : 데이터는 대칭
- Skewness -1~-0.5 이거나 0.5~1 : 적당히 치우침
- Skewness -1보다 작거나 1보다 클 경우 :상당히 치우침

첨도

- 첨도 Kurtosis
 - 첨도는 분포 그래프의 꼬리 부분에 관한 것
 - **Kurtosis가 높으면**
 - 데이터가 두꺼운 꼬리
 - outlier를 가지고 있다는 것을 의미
 - 다수의 outlier는 조사할 필요
 - **Kurtosis가 낮으면**
 - 데이터가 얇은 꼬리
 - outlier를 가지고 있지 않다
 - 이상 결과의 데이터를 다듬을 필요 검토.



- 구분
 - **Mesokurtic** : 이 분포는 정규 분포와 유사
 - 표준정규분포는 첨도 3 정도
 - **Leptokurtic (Kurtosis > 3)** : 분포가 길고, 꼬리가 더 뚱뚱
 - Mesokurtic보다 높고 날카와 특이치(outlier)가 많다는 것을 의미
 - **Platykurtic (Kurtosis < 3)** : 분포는 짧고 꼬리는 정규 분포보다 얇음
 - 피크는 Mesokurtic보다 낮고 넓으며, 이는 데이터가 가벼운 편이나 특이치(outlier)가 적음

분산도

- 편차(Deviation):

- 개별 자료와 전체 자료 평균 간의 차이
- 개별 자료들이 평균자료에서 얼마나 떨어져 있는가?(거리)

$$Deviation = x_i - \bar{X}$$

- 분산(Variance)

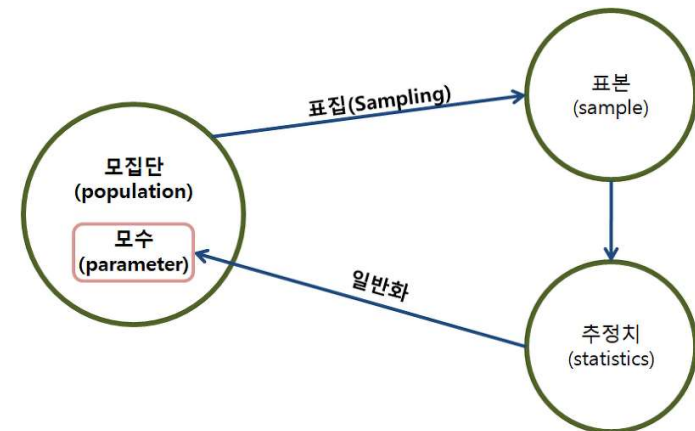
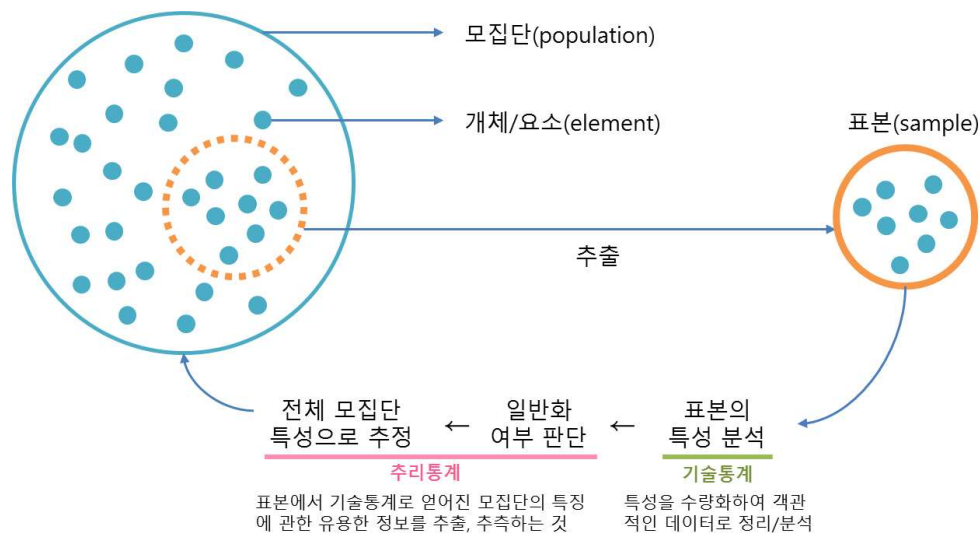
- 편차의 제곱을 모두 더해 평균 낸 값
- 학생 키의 분산은 cm^2 이됨

- 표준편차(Standard deviation):

- 분산에 제곱근을 취한 값 단위를 원래의 자료의 단위에 맞게 전환

모집단 표본

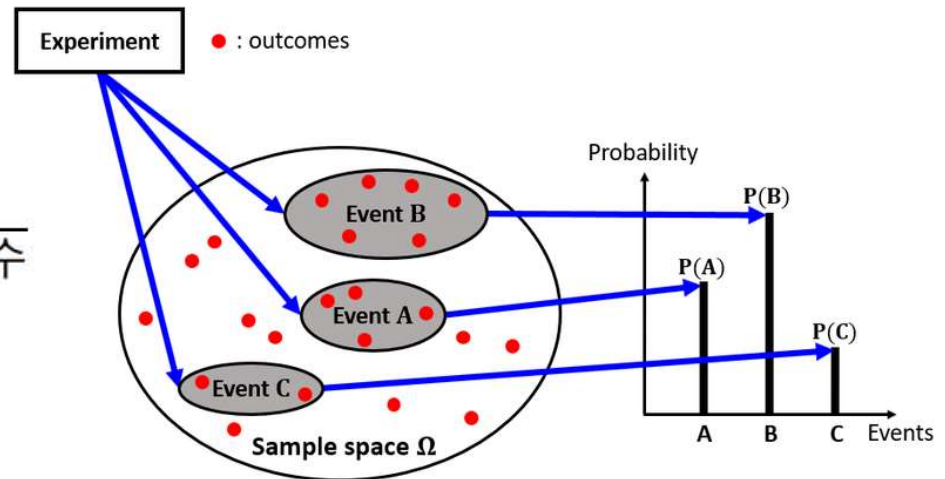
- 모집단 (Population)
 - 연구자가 알고 싶어하는 대상 / 집단 전체
 - 추상적/ 실제 전체를 상대하기 어려움
- 표본 (Sample)
 - 연구자가 측정 또는 관찰한 결과들의 집합
 - 모집단을 위한 실험체



표본 공간

- **표본 공간 (sample space)**
 - 어떤 특정 실험 또는 무작위 실험을 했을 때, 나올 수 있는 가능한 모든 결과들의 집합(the set of all possible outcomes or results of that experiment)
- **사건(event)**
 - 표본공간의 부분집합으로 어떤 조건을 만족하는 특정한 표본 점들의 집합
- **확률(probability)**
 - 동일한 조건 하에서 동일한 실험을 무수히 많이 반복하여 실시할 때, 어떤 특정한 사건이 발생하는 비율

$$P(A) = \frac{A \text{ 사건이 일어나는 경우의 수}}{\text{모든 사건이 일어나는 경우의 수}}$$



확률

- **확률의 주관적 정의**

- 한 개인의 경험이나 지식, 정보, 직관 등을 토대로 각자의 주관적
- 주관적인 믿음의 척도 어떤 일이 발생할 가능성에 대한 믿음의 정도

- **고전적 확률**

- 결과가 발생할 가능성이 같다고 볼 수 있을 때 사건 A의 확률을 다음과 같이 정의
- $P(A) = \frac{\text{사건 A의 경우의 수}}{\text{전체 경우의 수}}$
- 단점: 모든 사건의 발생확률이 같은 경우 별로 없다

- **통계적 확률**

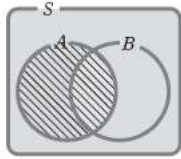
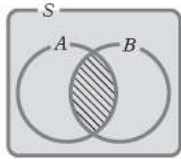
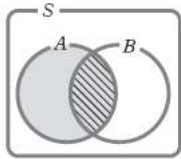
- 어떤 시행을 반복할 때, N번 시행에서 사건 A가 발생한 횟수를 $n(A)$ 라 할 때, 사건 A의 상대도수
- 사건 A의 상대도수 = $n(A)/N$
- $P(A)$ 는 N이 무한대로 갈때 확률
- 단점: N이 무한으로 갈때 값이 수렴 안할 수 있다. 반복적 시행이 불가능한 경우

공리적 확률론

- 러시아 수학자 안드레이 콜모고로프 정의 (A. N. Kolmogorov)
- 표본공간 Ω 의 모든 사건들의 집합 F 위에서 정의된 함수 p 가
- 다음 세 가지를 만족할 때, p 를 확률측도(probability measure)라 한다.
 - 1. 모든 $A \in F$ 에 대해 $0 \leq p(A) \leq 1$
 - 2. $p(\Omega) = 1$
 - 3. 표본공간의 모든 사건이 서로 배반이면,
 - $p(A_1 \cup A_2 \cup \dots \cup A_n) = p(A_1) + p(A_2) + \dots + p(A_n)$
- 정의
 - 모든 사건을 모아놓은 집합을 F
 - 표본공간 Ω
 - 확률측도 p
 - 이 세 쌍으로 구성된 공간을 '확률공간(probability space)'이라고 함
- 해설
 - 첫 번째로 표본공간의 모든 확률은 항상 1이어야 한다.
 - 두 번째로 각 사건의 확률은 항상 0과 1 사이의 값을 갖는다.
 - 세 번째로 각 사건이 서로 배반이면 합사건의 확률은 각 확률을 더한 것과 같다.

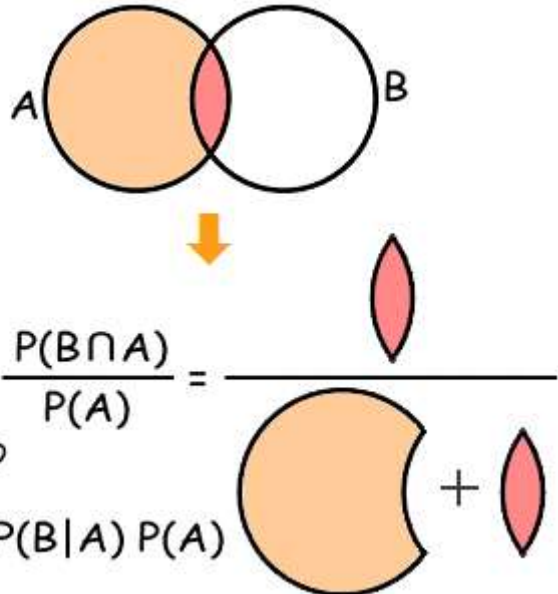
조건부 확률

- 어떤 사건이 일어날 가능성에 대한 믿음 강화-> 사건에 대한 정보가 필요
- 사건에 대한 정보가 많으면 많을수록 우리의 믿음(확률)은 좀 더 확실
- 즉 주어진 정보에 따라 확률이 달라진다!
- 예)
 - 야구의 경우 투수에 따른 타율 변화
 - 경기장에 따른 승률의 변화

$P(A) = \frac{P(A)}{P(S)}$	$P(A \cap B) = \frac{P(A \cap B)}{P(S)}$	$P(B A) = \frac{P(A \cap B)}{P(A)}$
전체에서 A의 비율	전체에서 $A \cap B$ 의 비율	A에서 B의 비율
		

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{\text{Intersection of A and B}}{\text{Area of A}}$$

$$P(B \cap A) = P(B|A) P(A)$$



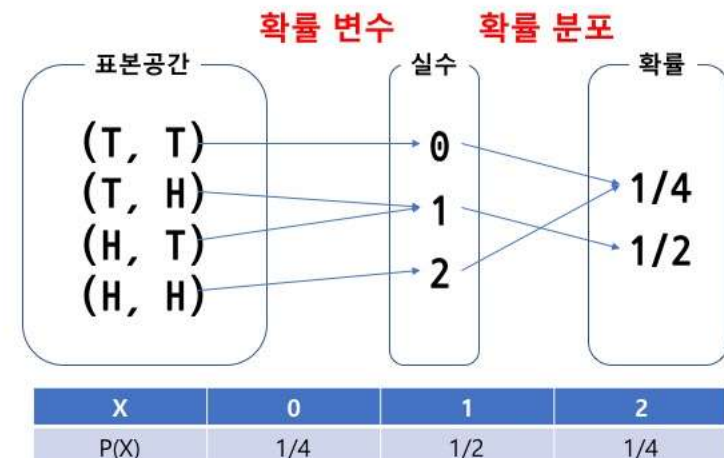
베이즈 정리

- 새로운 정보를 토대로 어떤 사건이 발생했다는 주장에 대한 신뢰도를 갱신해 나가는 방법
- 확률론 패러다임의 전환: 연역적 추론에서 귀납적 추론으로의 변화
- 연역적
 - 빈도주의 기존 통계학(연역적:집단통계를 규정하고 분포 측정 유의성 판단)
- 귀납적
 - 경험에 기반한 선택적인, 혹은 불확실성을 내포하는 수치를 기반
 - 거기에 추가 정보를 바탕으로 사전 확률을 갱신

$$\underset{\substack{\text{사후 확률} \\ \text{(posterior)}}}{P(H|E)} = \frac{P(E|H) \overset{\substack{\text{사전 확률} \\ \text{(prior)}}}{P(H)}}{P(E)}$$

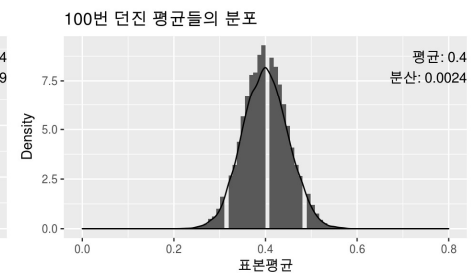
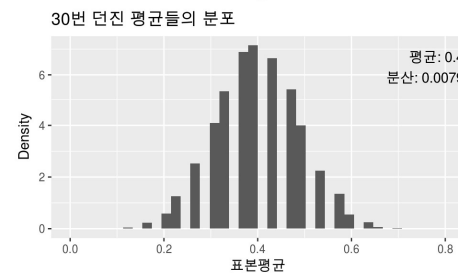
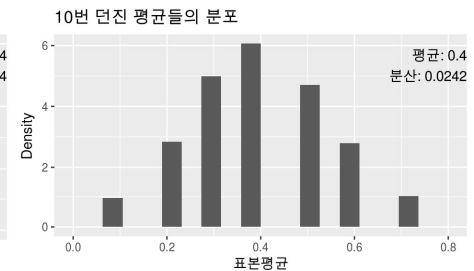
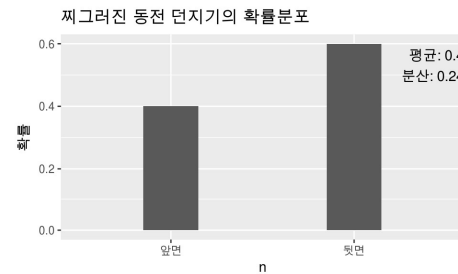
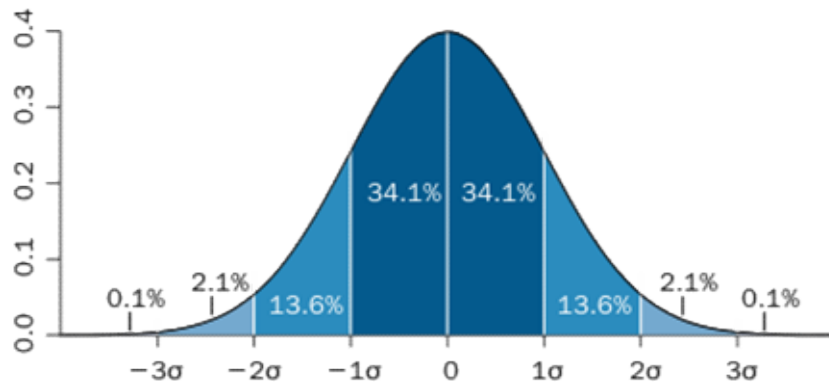
확률 변수

- 변수(Variable)이란 무엇일까요? 변수란 특정 조건에 따라 변하는 값
 - 확률변수(Random variable)는 확률에 따라 변하는 값
- 확률 변수
 - 무작위 실험을 했을 때, 특정 확률로 발생하는 각각의 결과를 수치적 값으로 표현하는 변수
 - 확률 변수(R)가 취하는 모든 실수들의 집합을 상태공간(State space)→확률 분포
- 종류
 - 이산확률 변수(Discrete random variable)
 - 확률 변수 x 가 어느 구간의 모든 실수값을 택하지 않고, 0,1,2 ...와 같은 고립된 값만을 택하는 변수
 - 연속확률 변수(Continuous random variable)
 - 확률 변수 x 가 취하는 값이
 - 연속된 구간으로 나타나는 확률 변수



정규분포

- 집단group
 - 자료 2개이상 모임
 - 모집단 :어떤 정보를 얻고자 하는 전체 대상 또는 전체 집합
 - 표본집단: 모집단으로부터 추출된 모집단의 부분 집합
- 분포(distribution):
 - 집단을 구성하는 개별 수치들이 가지는 자료의 전반적인 특성
 - 평균 중심의 좌우대칭 종모양의 분포를 정규분포라 칭함
 - 대부분 분포는 무한 반복 추출하면 정규분포로 수렴



2

기초 통계

통계학은 데이터에서 의미를 찾아내는 방법을 다루는 학문