마케터를 위한 기초통계

- 디지털 마케팅 SCHOOL 5기 (17. 3. 27)

박재근

- 데이터 사이언스 SCHOOL 코스매니저
  - 스쿨 매니징
  - 기초통계 스터디
  - 데이터 분석 프로젝트 매니징
- 1기 수료생
- 교통/물류학 전공

수업의 개요

Chapter 1. 통계를 왜 배워야 하는가 ?

- 통계의 목적
- 문제의 모형화

Chapter 2. 통계의 언어 배우기

- 모집단과 표본
- 모수와 통계량
- 기술통계와 추론통계
- 평균과 분산, 표준편차
- 확률론
  - 조건부확률과 베이즈정리
  - 확률변수와 확률분포

Chapter 3. 분석의 기본 단계 (과정) 알기

- 표본과 표집분포
  - 중심극한정리
- 통계적 추정
  - 신뢰도와 신뢰구간
- 통계적 가설검정
  - 유의수준과 P-value

Chapter 4. 분석 방법론 (모형) 배우기

- ANOVA 분석
- 상관분석
- 회귀분석

Chapter 1. 통계를 왜 배워야 하는가 ?

- 통계의 목적
- 문제의 모형화

1.1. 통계의 목적은 무엇인가 ?

Prediction

- 불확실성을 줄이고, 최대한 정답을 맞춰가는 것 (예측하는 것)

불확실성을 줄이자. 정답에 근접할 가능성을 높이자 !!! (★)

- 우리가 문제의 정답을 맞추기 위해 풀어나가는 사고방식 그대로 통계분석도 똑같은 원리로 작동한다.

1.2. 문제의 모형화

- 현실의 문제를 데이터 분석 문제로 변환
  - ***수학적 기호로 Modeling (★)***
  - ***+ 수많은 가정들이 포함됨 (★)***

현실의 모형화 : $y = f(X)$

⬇

예측 모델 : $\hat{y} = \hat{f}(X) = \text{argmax}\ P(y|X, D, M)$

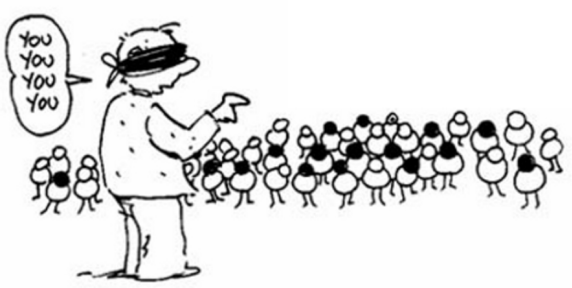
표시 한 내용이 무엇이었나요?

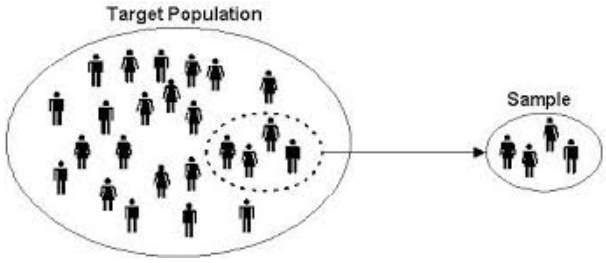
- ***불확실성을 줄이자. 정답에 근접할 가능성을 높이자 !!! (★) ➜ 확률***
- ***수학적 기호로 Modeling (★) ➜ 확률분포, 확률모형***
- ***+ 수많은 가정들이 포함됨 (★) ➜ 정규성***

Chapter 2. 통계의 언어 배우기

- 모집단과 표본
- 모수와 통계량
- 기술통계와 추론통계
- 평균과 분산, 표준편차
- 확률론
  - 조건부확률과 베이즈정리
  - 확률변수와 확률분포

모집단과 표본





2.1. 모집단과 표본

- **모집단 (Populataion)**
  - 분석가의 관심 대상이 되는 모든 개체의 집합
  - 규모가 작을 때에는 모든 개체를 조사 가능 (전수조사)
  - 그러나 규모가 클 경우, 다 조사하기 힘듦
  - ex. 대한민국 고등학교 학생들의 키
- **표본 (Sample)**
  - 조사 대상으로 채택된 일부 집합
  - 모집단의 규모가 큰 경우, 일부 표본을 뽑아 이를 분석
  - 모집단의 특성을 파악 (추측)
  - ex. 서울, 부산, 광주, ... 일부 고등학교 학생 5000명의 키

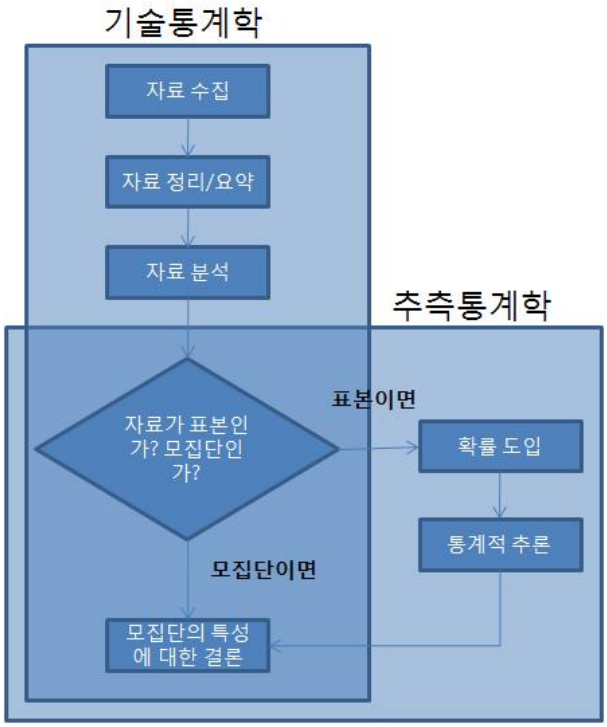
2.2. 모수와 통계량

- **모수 (Parameter)**
  - 모집단의 특성을 수치로 나타낸 것
  - 평균, 분산, 표준편차 : μ, σ^2, σ
- **통계량 (Statistic)**
  - 표본의 특성을 수치로 나타낸 것
  - 평균, 분산, 표준편차 : $\overline{X}, S^2, S$
- 모집단과 표본을 구분하기 위해 기호를 달리하여 사용

2.3. 기술통계와 추론통계

- **기술통계학 (Descriptive Statistics)**
  - 자료를 정리하고 요약하는 자료특성의 계산과 관련된 통계학
  - ex. 우리반 평균 시험점수를 알기 위해 학생들 점수를 모아 평균을 구하는 것
- **추론통계학 (Inferential Statistics)**
  - 모집단에서 뽑은 표본을 분석하여 이를 기초로 모집단의 특성을 추론하는 통계학
  - ex. 우리나라 고등학교 수능모의고사 평균을 알아보기 위해 몇만 명만 표본으로 뽑아 그들의 점수를 바탕으로 전체 평균을 추측하는 것

2.3. 기술통계와 추론통계



여기서 잠깐!

GA 는 기술통계학 vs 추론통계학 ?

2.4. 변수 (1)

양적변수 (quantitative variable, real value)

수치로 나타낼 수 있는 변수

- **이산변수 (discrete)**
  - 정숫값을 취한 수 있는 변수
  - ex. 자녀수, 자동차판매대수, 세션수 등
- **연속변수 (continuous)**
  - 모든 실수값을 취할 수 있는 변수
  - ex. 길이, 무게 등

2.4. 변수 (2)

질적변수 (qualitative variable, categorical value)

수치로 나타낼 수 없는 변수

- **명목변수 (nominal)**
  - ex. 성별, 종교, 출생지, 운동선수 등번호 등
- **서열변수 (ordinal)**
  - 측정대상 간의 순서를 매기기 위해 사용되는 변수
  - ex. 성적 A, B, C 등급

2.5. 자료의 정리 - 도수분포표

수집된 자료를 적절한 등급(또는 범주)으로 분류하고 각 등급에 해당되는 빈도수 등을 정리한 표

학생들의 키

(단위 : cm)

144	168	148	129
162	130	153	154
167	135	128	140
134	159	149	145
138	151	146	150

<정리되지 않은 자료>

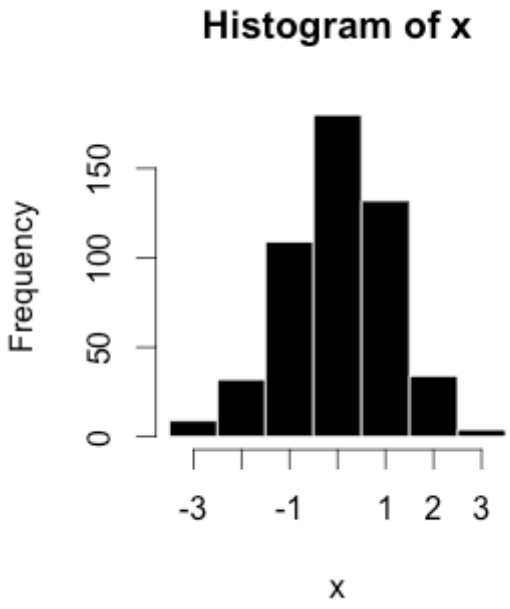
학생들의 키

키(cm)	학생 수(명)
120이상~130미만	2
130 ~ 140	4
140 ~ 150	6
150 ~ 160	5
160 ~ 170	3
합 계	20

<도수분포표>

2.5. 자료의 정리 - 히스토그램

- 도수분포표를 그래프로 나타낸 것
- 보통 히스토그램에서는 가로축이 계급, 세로축이 도수를 뜻함



히스토그램 (예제) - 문제

- 질적 자료의 도수분포표 & 히스토그램 작성하기
  - 아파트 A 에 어떤 직업을 가진 사람들이 얼마나 살고 있는가를 알기 위해서 세대주들의 직업을 조사해 보니 다음과 같았다. 아래 데이터를 이용하여 4개의 직업을 구분한 도수분포표와 히스토그램을 그려라.

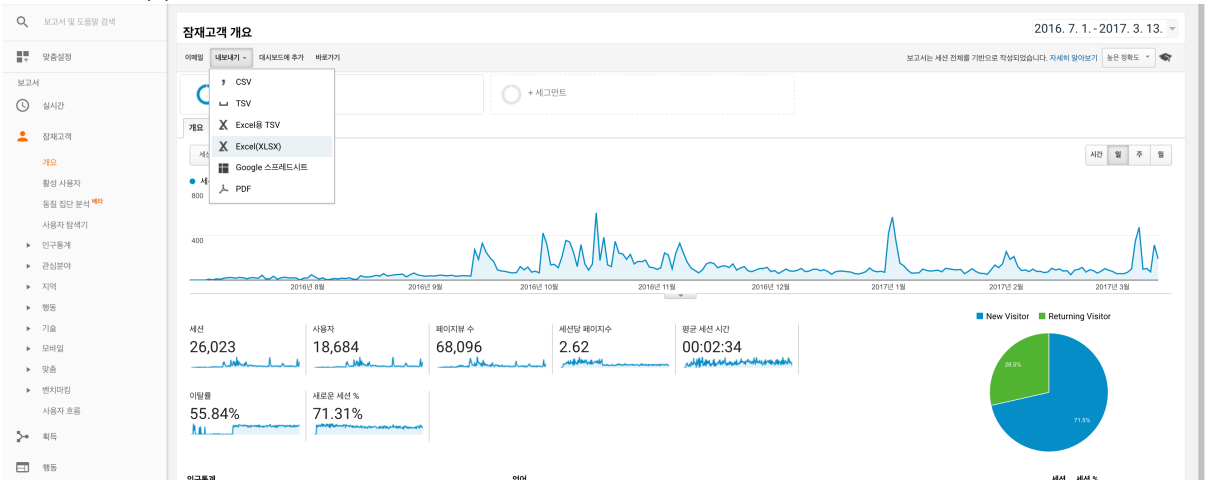
표. 세대주들의 직업							
101호	사무직	201호	전문직	301호	상업	401호	사무직
102호	전문직	202호	전문직	302호	상업	402호	상업
103호	전문직	203호	상업	303호	상업	403호	사무직
104호	전문직	204호	노동	304호	전문직	404호	상업
105호	사무직	205호	상업	305호	상업	405호	노동

히스토그램 (예제) - 정답

히스토그램 (실습) - GA데이터로 히스토그램 그리기

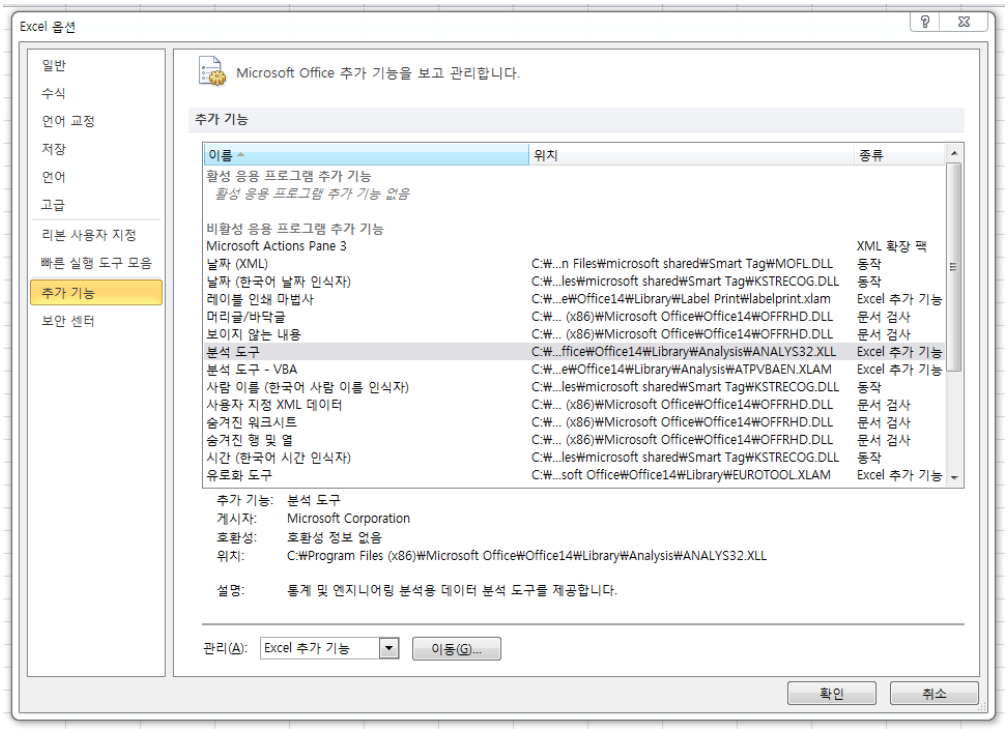
히스토그램 (실습) - 준비단계 (1)

- 데이터 수집
  - GA접속 - 잠재고객 개요 - 내보내기 - Excel(XLSX)
  - "DMS실습(1)-히스토그램, 분포특성.xlsx" 로 함께 실습



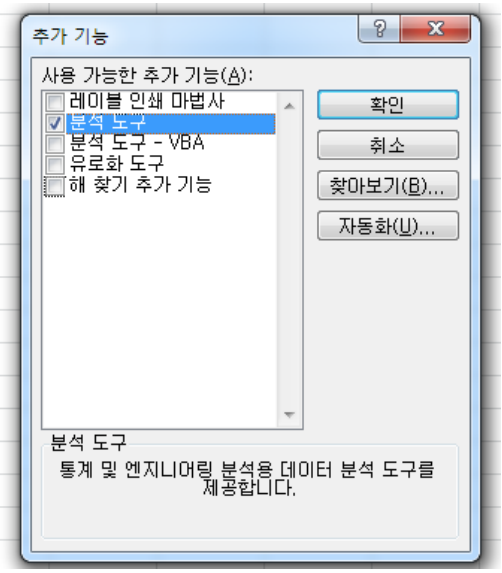
히스토그램 (실습) - 준비단계 (2)

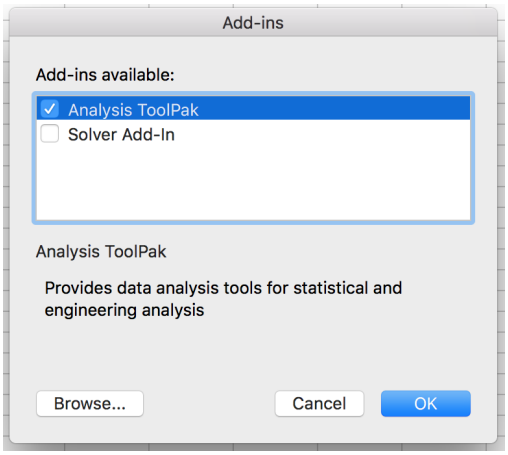
- 엑셀 **분석도구** 추가하기 (Windows OS)
  - 파일 - 옵션 - 추가기능 - 분석도구 클릭
  - 이동 클릭



히스토그램 (실습) - 준비단계 (3)

- 엑셀 **분석도구** 추가하기 (Window OS)
  - 추가기능 상자에서 **분석도구** 클릭 - 확인
- 엑셀 **분석도구** 추가하기 (Mac OS)
  - 도구(Tools) - 추가기능(Add-ins) - 분석도구(Analysis ToolPack)





2.6. 분포의 특성 (1) - 집중화경향

최빈값 (mode)

최빈값은 빈도수가 가장 많이 발생한 관찰값을 말함

- ex) 1, 3, 6, 6, 6, 7, 7, 12, 12, 19 있을때, 최빈값은 6이다.

중앙값 (median)

중앙값은 수치로 된 자료를 크기순서대로 나열할 때, 가장 가운데에 위치하는 관찰값을 말한다.

- ex) 1, 2, 4, 5, 7, 9, 10 있을때, 중앙값은 5이다.

2.6. 분포의 특성 (1) - 집중화경향

산술평균 (arithmetic mean)

우리가 흔히 사용하는 간단한 평균, 그냥 "평균" 이라고도 한다.

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{\sum X_i}{n}$$

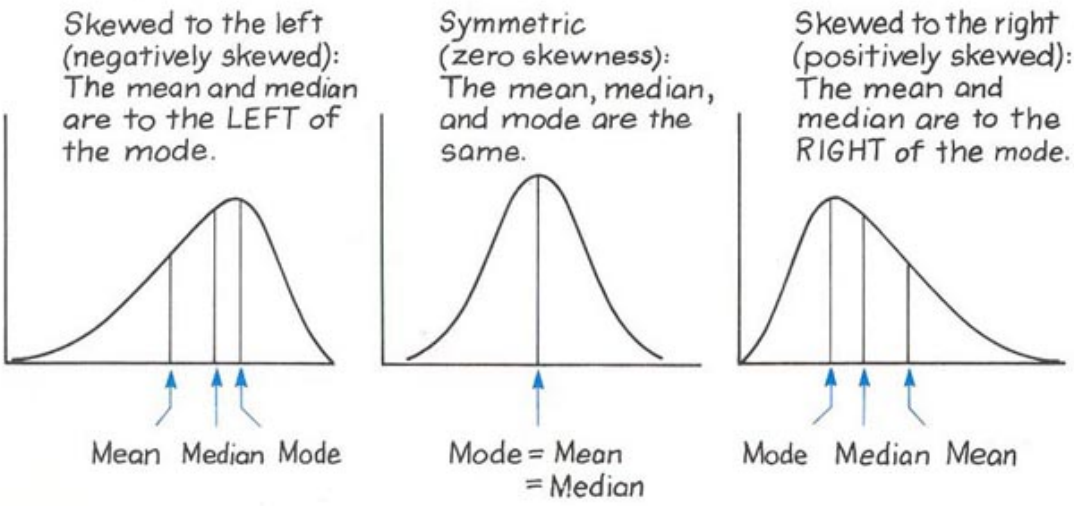
가중평균 (weighted arithmetic mean)

같은 모집단에서 표본을 서로 다른 개수로 뽑은 경우 (가중치가 존재하는 경우) 평균값을 구할때 사용

$$\bar{X} = \frac{n_1\bar{X_1} + n_2\bar{X_2} + \cdots + n_k\bar{X_k}}{n_1 + n_2 + \cdots + n_k} = \frac{\sum n_i\bar{X_i}}{n_i}$$

평균과 최빈값, 중앙값 위치 비교

- 대칭분포
  - 평균 = 중앙값 = 최빈값 모두 같다
- 왼쪽꼬리분포 (skewed to the left)
  - 평균은 중앙값보다 작다
- 오른쪽꼬리분포 (skewed to the right)
  - 평균은 중앙값보다 크다



2.6. 분포의 특성 (2) - 흩어진 정도

분산 (variance)

자료가 평균으로부터 얼마나 떨어져 분포하는지를 가늠하는 숫자
분산이란 각각의 관찰값에 대한 평균과의 편차를 제곱하여 그 평균을 구한 것

- 모집단의 분산 (σ^2)
- 표본의 분산 (S^2)
  - n 대신 $(n-1)$을 나누는 이유는, $(n-1)$을 나누어줌으로써 모집단의 σ를 추정하는데 더 적절한 표준편차를 구하기 위함이다.

$$\sigma^2 = \frac{\sum_1^N (x_i-\mu)^2}{N} = \frac{\sum_1^N x_i^2 - \mu^2}{N}; \qquad S^2 = \frac{\sum_1^n (x_i-\overline{X})^2}{n-1} = \frac{\sum_1^n x_i^2 - n\overline{X^2}}{n-1}$$

여기에서 N: 모집단의 크기
n: 표본의 크기

2.6. 분포의 특성 (2) - 흩어진 정도

표준편차 (standard deviation)

분산의 양의 제곱근

- 모집단의 표준편차 ($\sigma = \sqrt{\sigma^2}$)
- 표본의 표준편차 ($S = \sqrt{S^2}$)

분포의 특성 (실습)

- 전체 세션의 평균과 분산, 표준편차 등을 구해보자 (기술통계)
- "DMS실습(1)-히스토그램, 분포특성.xlsx" - "Descriptive Sheet"

2.7. 확률론

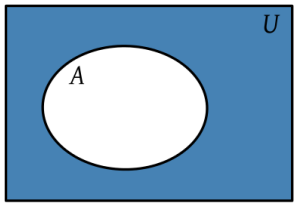
- 확률(probability)이란 어떤 상황이 발생할 가능성

2.7.1. 집합이론

- 확률이론을 쉽게 설명하기 위해서는 집합이론의 용어와 부호 사용하는 것이 편리
- 집합 (set) 이란 개체 또는 원소 (element)의 모임이라 정의
- 원소는 { ... } 속에 넣는 것이 관례
  - ex. A = { 남자, 여자 }, B = { 10대, 20대, 30대, ... }

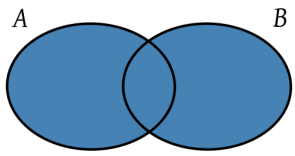
1. 여집합

- A^C = { 전체집합 중에서 집합 A 에 포함되지 않는 원소들 }



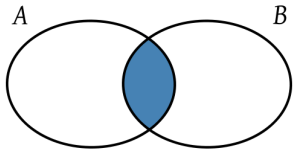
2. 합집합

- $A \cup B$ = { 집합 A 또는 집합 B에 속하는 원소 }



3. 교집합

- $A \cap B$ = { 집합 A와 B의 공통 원소 }



합집합의 계산

- $A \cup B = A + B - A \cap B$
  - if) 집합 A와 B가 서로 배타적(mutually exclusive)일 때 ($A \cap B = \varnothing$)
  - $A \cup B = A + B$

집합이론 (예제) - 문제

- 100명의 학생이 시험을 본 결과, 영어시험에서 60점 이상 받은 학생은 40명, 국어시험에서 60점 이상 받은 학생이 50명이었다. 영어시험과 국어시험 모두 60점 이상 받은 학생이 15명이라면, 최소한 한 과목에서 60점 이상을 받은 학생은 몇명인가?

집합이론 (예제) - 정답

2.7.2. 확률법칙 (1)

덧셈법칙

- 집합이론에서 합집합의 개념
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
  - if) A사건과 B사건이 서로 배타적일 때 (서로 독립일 때)
  - $P(A \cup B) = P(A) + P(B)$

2.7.2. 확률법칙 (2)

조건부확률

- 사건 B가 발생했다는 조건하에서 사건 A가 발생할 확률

$$P(A|B) = \frac{p(A \cap B)}{p(B)}$$

- *조건부확률은 표본공간이 전체사건이 되는 것이 아니라, 새로운 조건이 부여되어 관심대상이 **새로운 표본공간**이 되는 경우에 쓰이는 개념이다.*

2.7.2. 확률법칙 (3)

곱셈법칙

- 집합이론에서 교집합의 개념
- $P(A \cap B) = P(B) \bullet P(A|B) = P(A) \bullet P(B|A)$
  - 사건 A와 B가 동시에 일어날 확률은
  - 사건 A가 일어날 확률과 사건 A가 일어난 다음 사건 B가 일어날 확률을 곱한 것이란 의미

2.7.3. 베이즈정리 (Bayes' theorem)

- 사전에 알고 있는 정보에 기준을 두고, 어떤 사건이 일어나게 될 확률을 계산하는 이론

$$P(A|B) = \frac{P(B|A)P(A)}{p(B)}$$

조건부확률 (예제) - 문제

- 패스트캠퍼스 스쿨 수강생 200명 중에서 여자 수강생은 40명이다. 30명이 디지털 마케팅 수강생이며, 이 중 여자 수강생은 10명이다. 다음 조건부 확률을 구해보자.

1. 여자 수강생 중에서 디지털 마케팅 수강생의 구성은 어떻게 되는가?
2. 디지털 마케팅 수강생 중에서 여자 수강생의 구성은 어떻게 되는가?

조건부확률 (예제) - 정답

조건부확률 (실습) - 인구통계 GA 데이터로 조건부확률 구하기

조건부확률 (실습) - 1. 목표 설정하기

- 성별에 따라 재방문 할 확률은 어떻게 될까?
- 성별에 따라 이탈률은 달라질까?
- 성별에 따른 모바일 접속자 비율은 어떻게 될까?

조건부확률 (실습) - 2. 데이터 수집

- GA접속 - 잠재고객 - 인구통계 개요
- 세그먼트 추가 (재사용자, 이탈한 세션수, 모바일 트래픽)
- 내보내기 - Excel(XLSX)
- "DMS실습(2)-조건부확률.xlsx" 로 함께 실습

조건부확률 (실습) - 3. 데이터 탐색 및 확률 구하기

- 수업 후 실행한 엑셀 파일 나눠 드릴게요!

수업끝

- 디지털 마케팅 SCHOOL 5기 (17. 3. 27)