WEEK 05

통계학의 이해

OPEN CYBER UNIVERSITY OF KOREA

학습목표

1. 통계학의 정의에 대해 학습한다.
2. 통계학이 우리 실생활에서 어떻게 사용되는지 알아본다.
3. 통계학의 기초 원리에 대해 학습한다.

학습목차

1. 당신이 통계학을 사랑해야 하는 이유
2. 지금까지 본 최고의 통계
3. 생활 속의 통계학

오늘의 수업 내용

◆ 통계

- ❖ 앨런 스미스, Why you should love statistics<당신이 통계학을 사랑해야 하는 이유>
  - 우리가 살아가는 세상에 대한 데이터를 다루는 과학, 통계학
  - 사람은 사회적 동물로 집단에 속해야 하고, 이런 상태 집단에 대한 분석은 통계학을 통해 이루어진다.
- ❖ 한스 로슬링, The best stats you've ever seen<지금까지 본 최고의 통계>
  - 자신의 분야에서 우수한 사람이라도 편견 때문에 세상을 바로 볼 수 없으며, 이를 위해 데이터를 통계 처리하여 효과적으로 보여줘야 한다.

1. 당신이 통계학을 사랑해야 하는 이유

1) 통계, 그리고 통계학이란 무엇인가?

2) 오늘의 연사 소개

3) TED 강의 시청 및 강의 내용 정리/요약

1) 통계, 그리고 통계학이란 무엇인가?

◆ 백과 사전에서 찾아본 통계의 정의 (1/2)

- 특히 사회집단 또는 자연집단의 상황을 숫자로 나타낸 것이다. 예를 들어 서울 인구의 생계비, 한국 쌀 생산량의 추이, 추출검사한 제품 중의 불량품의 개수 등이 그것이다. 통계는 집단에 관한 것으로서, 어떤 사람의 재산이라든가 한라산의 높이 등, 어떤 개체에 관한 수적 기술은 아무리 구체적이더라도 통계는 아니다. 통계는 사회의 발전과 함께 발달해 왔는데, 오늘날의 사회생활과 과학은 통계 없이는 존재할 수 없다.
- 집단현상을 통계로 나타낼 때, 그 집단을 구성하는 각 개체를 통계단위 또는 단위라고 한다. 이 단위는 공통의 성질을 가지고 있는데, 이 공통의 성질을 표지(標識)라고 한다. 이를테면 한국의 인구를 구성하는 단위는 일정한 날짜와 시간에 한국에 살고 있는 사람이며, 이 조건이 표지가 된다. 표지에는 남녀, 산업·직업 등 질적인 것과, 연령·소득금액 등 양적인 것이 있다. 질적인 표지의 통계를 속성통계, 양적인 표지의 통계를 변수통계라고 한다. 또, 집단의 성질에 따라 자연현상에 관한 자연통계와, 사회현상에 관한 사회통계로 나누어지는데, 자연통계는 기후·생물통계 등으로, 사회통계는 경제통계·경영통계 등으로 세분할 수 있다.

[출 처] 두산백과(http://www.doopedia.co.kr)

1) 통계, 그리고 통계학이란 무엇인가?

◆ 백과 사전에서 찾아본 통계의 정의 (2/2)

- 또한, 국세조사(國勢調査)와 같이, 집단의 한 시점에 관한 것인 정태통계(靜態統計)와, 1년간의 출생수·사망수·공업생산 등과 같이 어떤 기간에 관한 동태통계(動態統計)로도 나누어진다. 이 밖에 집단의 전체에 걸치는 전수통계(全數統計)와, 일부분을 관찰한 부분통계로 나누는 수도 있는데, 전수통계는 비교적 소박한 기술적 수리(記述的數理)처리에 따른 방법으로 기술통계라고 불리며, 부분통계는 부분에서 전체에로의 추측기법(推測技法)을 포함하기 때문에 추측통계라고 한다.
- 통계를 이용하는 데는, 작성자·작성시기·작성방법·대상(단위표지)·대상의 존재장소 등에 관한 깊은 인식을 필요로 한다. 이 같은 모든 통계는 현실의 일정한 사회관계를 바탕으로, 조사자와 피조사자 사이에서 질문·응답이 행해지는 통계조사(統計調査)라는 특수한 과정을 거쳐 이루어지는데, 거기에는 상호협조와 이해에 따르는 대항관계가 작용한다. 또한 통계는 그 필요성과 작성능력이라는 점으로 보아, 그 대부분이 정부나 지방자치단체 등에 의한 관청 통계로 작성된다는 특성을 지닌다.

[출 처] 두산백과(http://www.doopedia.co.kr)

1) 통계, 그리고 통계학이란 무엇인가?

◆ 실생활에서 찾아볼 수 있는 통계의 예시

❖ 당신의 평점(GPA)은 3.5이고, 토플 점수는 IBT 기준으로 100점이며, GMAT 점수는 700점이다. 당신은 MBA(Master of Business Administration) 과정에 입학을 원한다. 하버드, MIT, 미시건, 시카고, 스탠포드 등의 경영학 석사 프로그램에서 입학 허가를 받게 될 가능성을 결정할 수 있는가?

◆ 필요한 정보

❖ 대학 요구 점수 최저 기준

❖ 합격자 평균 점수 (연도별)

❖ 합격자 점수의 중간값 (연도별)

❖ 합격자 점수의 분포 (연도별)

1) 통계, 그리고 통계학이란 무엇인가?

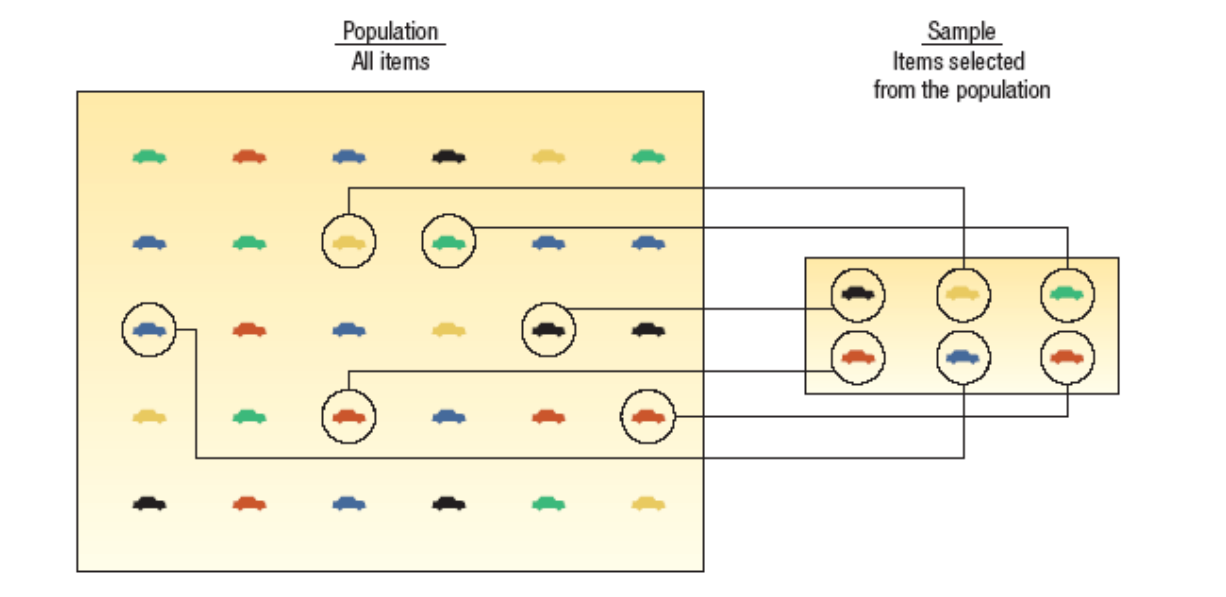
◆ 통계의 기초 지식

❖ 모집단

- 관심의 대상인 모든 개인이나 개체 또는 관심의 대상인 모든 개인이나 개체로부터 얻어진 측정치

❖ 표본

- 관심의 대상인 모집단의 부분 또는 일부



[출 처] Lind, D. A., Marchal, W. G., & Wathen, S. A. (2018). Basic statistics for business & economics. McGraw-Hill

2) 오늘의 연사 소개

◆ 강연자 소개

❖ Alan Smith(앨런 스미스)

- 영국 국가 통계청(UK Office for National Statistics: ONS)에서 근무
- 2015년부터 런던 파이낸셜 타임스로 이직하여 데이터 시각화 업무 담당
- 2010년 왕립통계학회 공식 통계 우수상 수상자
- 2011년 영국 여왕 생일 영예 목록에서 대영제국 훈장 오피스 임명

[출 처] https://www.ted.com/speakers/alan_smith

2) 오늘의 연사 소개

◆ Alan Smith(앨런 스미스) (1/2)

- Alan Smith는 1989에서 1992까지 랭커스트 대학교(University of Lancaster)에서 Geography 2nd Class Honours 학위를 취득하였으며, 1999년부터 2005까지 콜로라도 대학교 볼더(University of Colorado at Boulder)에서 2년간 2학기 석사 수준의 지리 수업을 공부하였다. 또한 샐퍼드 대학교(University of Salford)에서 Geographic Information Science and Cartography Master of Science(MSC)를 받았다.
- 이후 그는 우수한 지도 제작 및 디지털 지도 제작의 능력을 가지고 영국 국가 통계청에서 오랜 경력과 실적을 달성했다. 그는 2001부터 2007까지 6년간 기업지리학과 선임연구관으로 활동하며 출판 매핑 서비스, 기업 GIS 지원 서비스 및 애드호크 연구 프로젝트(예: 데이터 품질, 공간 분석 및 모델링, 웹 매핑, 시각화)를 담당하였고, ON/ORNance Survey relationship을 위한 기업 사무국 기능을 제공하였다.

2) 오늘의 연사 소개

◆ Alan Smith(앨런 스미스) (2/2)

- 그리고 2007부터 2014까지 7년간 2007년에 설립된 ONS 데이터 시각화 센터에서 데이터 시각화 Principal Methodologist로서 데이터 시각화에 대한 기업 표준을 만들고 정부 통계학자 교육 과정을 개발 감독 하는 등 다양한 활동을 하였다. 또한 2007년에서 2015년까지 7년간 Head of Digital Content로 활동하며 ONS 디지털 출판부의 데이터 시각화, 디자인 및 편집 서비스 담당하였다. 하지만 그는 2015년에 런던 파이낸셜 타임즈(Financial Times)로 이직하여 2019까지 데이터 시각화 편집자로 일하였고 현재 데이터 시각화 편집장을 역임하고 있다.
- Alan Smith는 데이터 시각화와 인터랙티브 그래픽(Interactive Graphics) 및 통계에 대한 그의 능력과 공헌으로 다양한 수상 경력을 가졌으며, 2010년 왕립통계학회 공식통계 우수상 수상자로 취임하였으며, 2011년 여왕의 생일 영예 목록에서 대영제국 훈장 오피스(Office)로 임명되기도 하였다.

3) TED 강의 시청 및 강의 내용 정리/요약

◆ TED Talks 시청

- ❖ Alan Smith, Why you should love statistics
- ❖ <당신이 통계학을 사랑해야 하는 이유>

[출 처] https://www.ted.com/talks/alan_smith_why_we_re_so_bad_at_statistics

3) TED 강의 시청 및 강의 내용 정리/요약

◆ 강연 요약

- ❖ 우리는 사람을 두 분류로 나눌 수 있다.
- ❖ 숫자에 익숙하고 계산을 잘 하는 사람, 그렇지 못한 사람.
- ❖ 하지만 통계를 이해하는 데에 뛰어난 계산능력이 필요한 것은 아니다.
- ❖ 통계학의 어원을 보면, 통계학은 우리가 사는 세상의 상태 집단에 대한 데이터를 다루는 과학이다.
- ❖ 또한 사람은 사회적 동물이기에 집단에 속하는 것을 좋아한다.
- ❖ 즉, 통계는 우리에 대한 과학이고, 이것은 우리가 숫자에 매료되어야 하는 이유이다.

1. 당신이 통계학을 사랑해야 하는 이유

1교시 수업을 마치겠습니다.

2. 지금까지 본 최고의 통계

1) 통계 자료의 시각화 방법

2) 오늘의 연사 소개

3) TED 강의 시청 및 강의 내용 정리/요약

1) 통계 자료의 시각화

◆ 통계 자료의 시각화

- 통계학은 조금 효과적인 의사 결정을 보조하기 위하여 데이터를 수집하고, 정리하고, 표현하고, 분석하고, 조직하는 과학이다.
- 이를 위해 통계학의 결과를 시각화해서 보여주는 것은 의사결정에 큰 도움이 된다고 할 수 있다.

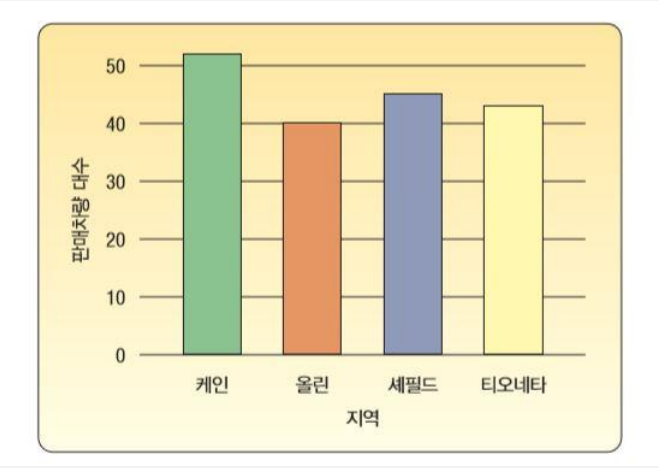
❖ 도수표(Frequency Table)

- 정성적 데이터를 상호배타적(mutually exclusive)이고, 총망라(collectively exhaustive)된 계급으로 분류한 후 각 계급에 존재하는 관측치의 도수를 나타낸 표
- 상호배타적: 데이터가 단 하나의 계급에만 소속된다는 것을 의미
- 총망라: 모든 값이 계급 내에 포함되는 것을 의미

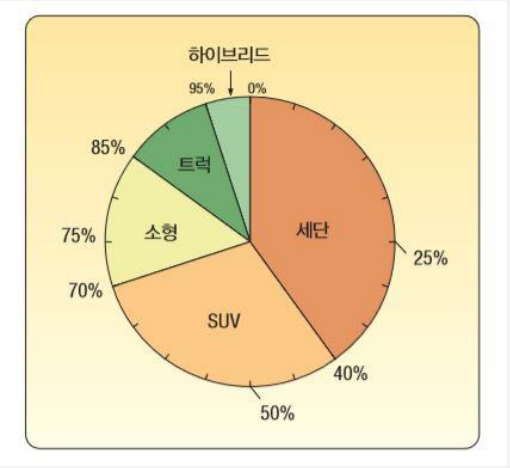
1) 통계 자료의 시각화

◆ 통계 자료의 시각화 (1/3)

- 막대도표: 수평축에는 정성계급을 수직축에는 계급의 도수를 나타내는 도표이다. 계급의 도수는 막대의 높이에 따라 비율적으로 반영된다.
- 파이도표: 각 계급을 총 도수에 대한 비율 또는 백분율로 나타내는 도표
- 각 계급의 관측치의 수에 대한 비교를 하고 싶을 때는 막대 도표를 활용하고, 백분율에 대한 상대적 차이를 비교하고 싶을 때 파이도표를 사용하는 것이 시각화에 도움이 된다.



막대도표



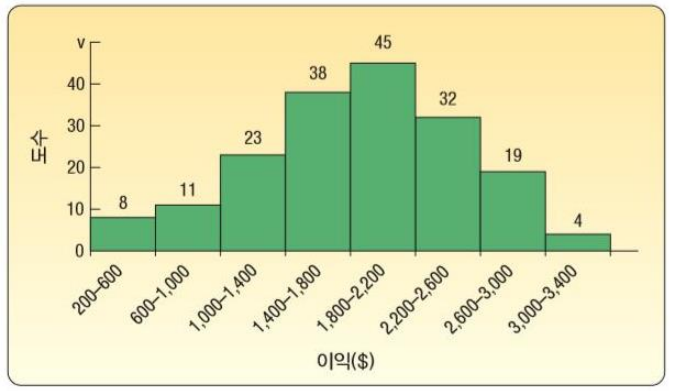
파이도표

[출 처] Lind, D. A., Marchal, W. G., & Wathen, S. A. (2018). Basic statistics for business & economics. McGraw-Hill

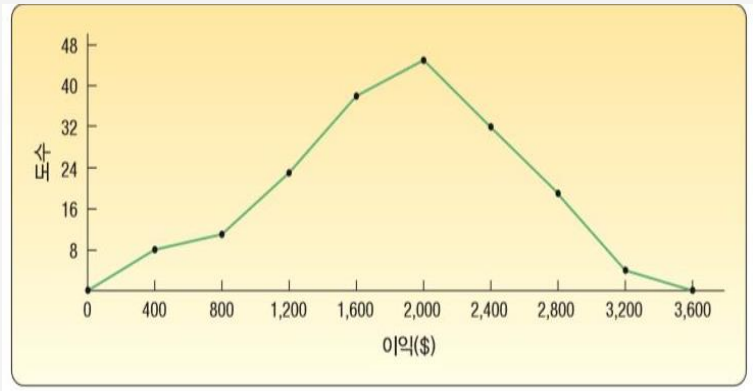
1) 통계 자료의 시각화

◆ 통계 자료의 시각화 (2/3)

- 히스토그램: 계급이 수평축에 표시되고, 계급도수가 수직축에 표시되는 그래프. 계급도수는 막대 의 높이로 표현되고, 각 막대는 데이터의 연속적인 성질을 반영하기 위하여 붙여서 그려진다.
- 도수다각형: 도수 다각형은 히스토그램과 유사하지만 계급 중간점과 계급도수의 교차점에 의하여 형성된 점들을 연결하는 직선으로 구성되어 있다.
- 도수다각형은 두 개 이상의 분포를 비교할 때 사용하기가 용이하다.



히스토그램



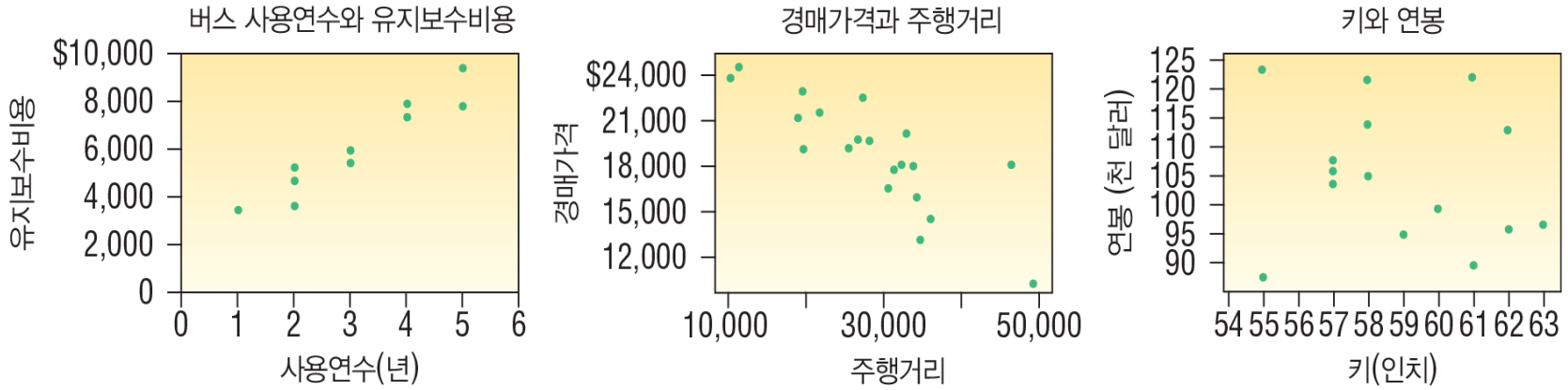
도수다각형

[출 처] Lind, D. A., Marchal, W. G., & Wathen, S. A. (2018). Basic statistics for business & economics. McGraw-Hill

1) 통계 자료의 시각화

◆ 통계 자료의 시각화 (3/3)

- 산점도: 두 개 변수 사이의 관계를 보여주는 도표
- 두 변수 모두 구간 척도는 비율 척도로 측정되어야 함
- 점들이 왼쪽 하단으로부터 오른쪽 상단으로 퍼져 있으면 그 변수들은 양의 관계를 가짐
- 점들이 왼쪽 상단으로부터 오른쪽 하단으로 퍼져 있으면 그 변수들은 음의 관계를 가짐



[출 처] Lind, D. A., Marchal, W. G., & Wathen, S. A. (2018). Basic statistics for business & economics. McGraw-Hill

2) 오늘의 연사 소개

◆ 강연자 소개

❖ Hans Rosling(한스 로슬링)

- 스웨덴의 의사, 학자, 연설가, 국제보건학 교수, 갭마인더 재단 회장
- 2017년 사망, '데이터를 생명으로 가져올 수 있는 사람'

[출 처] ko.wikipedia.org

3) TED 강의 시청 및 강의 내용 정리/요약

◆ TED Talks 시청

- ❖ Hans Rosling, The best stats you've ever seen
- ❖ <지금까지 본 최고의 통계>

[출 처] https://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen

3) TED 강의 시청 및 강의 내용 정리/요약

◆ 강연 요약

- Hans Rosling은 자신의 분야에서 아무리 우수한 성취를 보인 사람들이더라도, 세계에 대해서는 많은 것을 모르고 있고, 그 원인은 편견이라는 사실을 알게 된다.
- 문제를 해결하기 위해 많은 양의 데이터를 간단히 시각화해 보여주는 도구, Gap miner를 만든다.
- 유엔과 국립 통계 대행사, 대학, 그 밖에 다른 비정부 조직들의 자료 등 수많은 자료가 있어도, 대중들은 큰 데이터베이스 안에 있는 자료들은 효과적으로 사용하지 못했기 때문이다.

2. 지금까지 본 최고의 통계

2교시 수업을 마치겠습니다.

3. 생활 속의 통계학

1) 시장에서 사용되는 통계학

2) 평균적인 미국인

3) 위치 척도와 산포 척도

1) 시장에서 사용되는 통계학

- 미국의 자동차 판매업은 연간 수십억 달러 이상의 매출을 올리는 거대 판매회사들의 전쟁터로 10,000여명 이상 고용, 50개 이상의 대리점을 소유하고 있다.
- 이러한 기업들은 통계학을 이용하여 데이터를 요약하고 의사 결정을 지원한다.
- 이를 위해서 기본적으로 그들이 궁금한 것은 차량 구매자의 인구통계학적 정보를 기술하는 것이다.
- 구매자의 연령, 과거 구매 이력, 구매한 차의 종류, 차량이 판매된 대리점 등의 데이터를 활용하게 된다.

[출 처] Lind, D. A., Marchal, W. G., & Wathen, S. A. (2018). Basic statistics for business & economics. McGraw-Hill

2) 평균적인 미국인

- 이름 : 로버트
- 나이 : 31세
- 키 : 69.5인치
- 수면시간은 7.7시간, 몸무게는 172파운드, 신발은 9.5사이즈, 허리둘레는 34인치, 양복사이즈는 40.
- 평균의 미국 남자는 매년 감자칩 4파운드를 먹고 TV 시청은 1,456시간이다.
- 평균 미국 모델인 경우, 키 5피트 11인치, 몸무게는 117파운드
- 평균 미국 여자는 키 5피트 4인치, 몸무게 140파운드
- 1950년대의 우상 마릴린 먼로를 지금의 기준으로 본다면?

[출 처] Lind, D. A., Marchal, W. G., & Wathen, S. A. (2018). Basic statistics for business & economics. McGraw-Hill

3) 위치 척도와 산포 척도

❖ 위치척도: 데이터 분포의 중심을 정확히 찾아내는 것

- 예1) 미국인은 연간 평균 568통의 우편물을 받는다.
- 예2) 미국 가정에는 평균적으로 구성원보다 TV가 더 많다. (2.73TV, 2.55명)
- 예3) 미국 주택은 평균 11.8년 만에 소유자가 바뀐다.

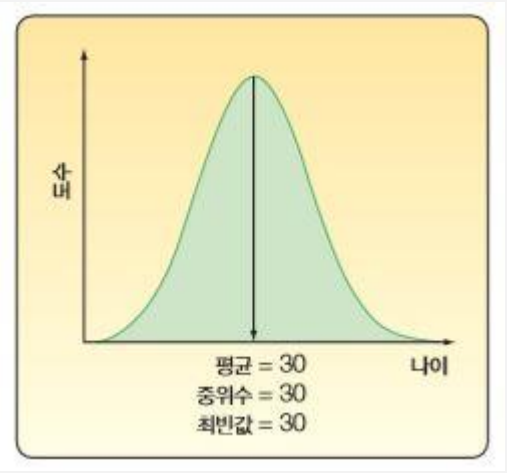
❖ 산포척도: 변동(variation) 또는 흩어짐(spread)을 고려한 정보

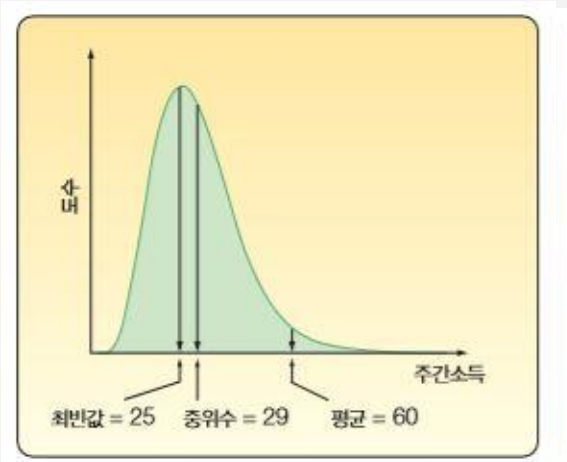
- 예) 인터넷 기업 임원의 연평균 소득이 $80,000 이고, 제약회사 임원의 연평균소득도 이와 같다고 가정하자. 하지만 실제 데이터를 살펴보면, 인터넷 기업 임원의 연봉은 $70,000 - $90,000 인 반면에, 제약회사 임원의 연봉은 $40,000 - $120,000의 범위를 갖는다.

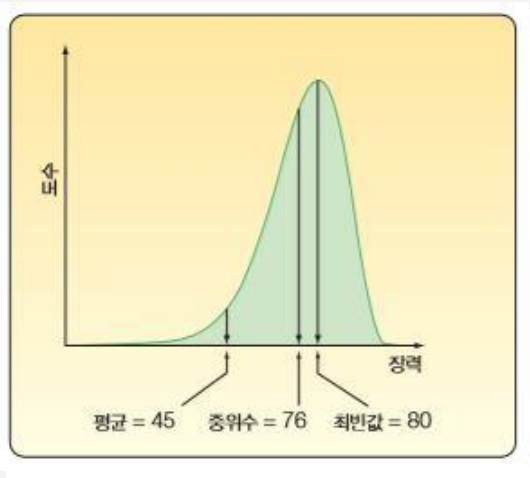
[출 처] Lind, D. A., Marchal, W. G., & Wathen, S. A. (2018). Basic statistics for business & economics. McGraw-Hill

3) 위치 척도와 산포 척도

- 위치척도: 데이터 분포의 중심을 정확히 찾아내는 것
- 위치척도는 데이터 집합의 중심 경향을 기술하는 데 사용되는 값이다.
- 일반적 위치척도
  - 평균 (모평균, 표본평균)
  - 중위수
  - 최빈값
- 산술평균은 가장 널리 사용되는 위치척도이다.







[출 처] Lind, D. A., Marchal, W. G., & Wathen, S. A. (2018). Basic statistics for business & economics. McGraw-Hill

3) 위치 척도와 산포 척도

- 산포척도: 변동(variation) 또는 흩어짐(spread)을 고려한 정보
- 지금까지 배운 평균, 중위수, 최빈값과 같은 위치척도는 데이터의 중심을 찾는 것에만 그 초점이 맞추어져 있다. 과연 데이터의 중심만 중요한가?
- 분산(Variance): 평균으로부터의 제곱편차에 대한 산술 평균
- 표준편차(Standard Deviation): 평균으로부터의 제곱거리 평균의 제곱근이다.

[출 처] Lind, D. A., Marchal, W. G., & Wathen, S. A. (2018). Basic statistics for business & economics. McGraw-Hill

주차 정리

◆ 통계

❖ 앨런 스미스, Why you should love statistics<당신이 통계학을 사랑해야 하는 이유>

- 우리가 살아가는 세상에 대한 데이터를 다루는 과학, 통계학
- 사람은 사회적 동물로 집단에 속해야 하고, 이런 상태 집단에 대한 분석은 통계학을 통해 이루어진다.

❖ 한스 로슬링, The best stats you've ever seen<지금까지 본 최고의 통계>

- 자신의 분야에서 우수한 사람이라도 편견 때문에 세상을 바로 볼 수 없으며, 이를 위해 데이터를 통계 처리하여 효과적으로 보여줘야 한다.

1. 모집단과 표본의 차이에 대해 설명하라.

정답

모집단

: 관심의 대상인 모든 개인이나 개체 또는 관심의 대상인 모든 개인이나 개체로부터 얻어진 측정치

표본

: 관심의 대상인 모집단의 부분 또는 일부

2. 일반적인 위치척도의 종류를 설명하라.

정답

평균(모평균, 산술평균)

중위수: 최소에서 최대의 순서로 정렬된 값들의 중간점

최빈값: 가장 빈번하게 출현하는 관측치 값

3. 한스 로슬링은 자신의 분야에서 우수한 성취를 보인 사람이라도 세계에 대해서는 많은 것을 모르고 있는데, 이 원인을 편견이라 보았다.

X

정답 : O

맞는 설명(T)이다.

4. 데이터를 시각화하는 방법 중, 히스토그램은 도수다각형과 유사하지만 두 개 이상의 분포를 비교할 때 더 자주 사용된다.

O X

정답 : X

틀린 설명이다.
도수다각형이 두 개 이상의 분포를 비교할 때 사용하기가 용이하다.

차주 예고

◆ 데이터 사이언스

- ❖ 한스 로슬링 “Let my dataset change your mindset(내 데이터셋이 당신의 사고방식을 바꾸게 해주세요.)”
  - 개발 관련 통계를 그래픽 디스플레이로 변화시켜 제시
  - 이분법적 분류에서 벗어난 새로운 태도 필요
- ❖ 케네스 쿠키어 “Big data is better data(빅 데이터가 더 좋은 데이터)”
  - 머신러닝과 빅데이터
  - 빅데이터의 문제점: 예측에 의한 처벌, 일자리의 감소
- ❖ 조엘 셀라니키오 “The big-data revolution in health care(의료 분야의 빅데이터 혁명)”
  - 국제 보건 분야 자료 수집 방식의 혁명적 변화
  - 예방 가능한 질병의 자료를 디지털 형식으로 수집하는 방식

3. 생활 속의 통계학

3교시 수업을 마치겠습니다.