데이터 요약

기술 통계

- 통계학
  - 불확실하고 잘 알려지지 않은 사실과 대상에 관련된 자료를 수집 및 요약정리
  - 이를 바탕으로 해석 및 분석하는 데 필요한 이론과 방법을 과학적으로 제시하는 학문
- 기술 통계학
  - 수집된 자료를 정리하여 그림이나 표로 요약하거나 자료의 수치를 요약한 대푯값(평균, 분산, 상관계수 등)과 데이터 분포의 형태와 변동의 크기를 구하는 방법을 다루는 것
- 기술 통계는 표본 자체의 속성이나 특징을 파악하는 데 중점을 두는 데이터 분석 통계
- 자료를 요약하고 조직화, 단순화하는데 목적이 있음
- 데이터 분석시 전수 조사가 불가하기 때문에 표본을 추출하여 표본을 설명해주는 데이터의 최솟값, 최댓값, 중위수 등의 통계량이 기술 통계량
- 모집단의 특성을 유추하는데 사용할 수 있음
- 명목 척도를 대상으로 하는 빈도 분석과 비율 척도 분석이 대표적인 기술 통계 분석 방법

추론통계

- 표본에서 얻은 통계치를 바탕으로 오차를 고려하면서 모수를 확률적으로 추정하는 통계 기법
- 표본에서 얻은 통계치를 가지고 모집단의 특성을 추정하는 데 초점을 두고 가설을 검증하거나 확률적인 가능성을 파악
- 이를 통해 향후 발생할 수 있는 사건을 예측할 수 있음

기술통계와 추론통계의 관계

- 표본의 특성을 분석해 객관적인 데이터로 정리 및 분석하는 것을 기술 통계라고 함
- 표본에서 얻은 기술 통계로 모집단의 특성과 정보를 추측 및 추리하는 것을 추론통계
- 기술 통계는 실재하는 데이터를 기반으로 결과를 얻는 것
- 추론 통계는 밝혀지지 않은 데이터를 기존 표본을 기반으로 추리 및 추측하여 얻은 것

자료의 변화 척도

평균

- 데이터는 수치적으로 널리 펴져있지만, 그 널리 퍼져 있는 것 중에 하나의 수를 모든 데이터를 대표하는 수로 뽑은 것
- 데이터들은 평균값 주변에 분포되어 있다.
- 평균은 흩어져 있는 데이터의 무게중심

- 평균의 함정

```
16일 한국경제연구원이 지난해 임금근로자 1544만명 자료를 분석한 결과, 우리나라 근로자의 평균 연봉은 3387만원, 전체 근로자의 연봉 분포에서 연봉 기준 중간순위에 위치한 근로자의 연봉은 2623만원, 상위 10% 커트라인에 위치한 근로자의 연봉은 6607만원으로 조사됐다./그래픽=뉴시스
- http://news.chosun.com/site/data/html_dir/2017/08/16/2017081601464.html
```

중앙값

- 중앙값은 중위수 또는 Median이라고 한다. 변량의 값을 크기 순으로 나열할 때 꼭 중앙에 오는 수치를 의미한다.
- 1, 2, 3, 3, 5의 중앙값은 3이고, 1,2,3,5 처럼 나열한 숫자가 짝수일 때는 중앙값의 한 가운데로 정한다.
- 평균값은 이상한 수치에 강한 영향을 받는다. 반면 중앙값은 그다지 영향을 받지 않는다.
- 데이터의 개수가 n이 홀수라면, (n+1)/2 번째 데이터가 중앙값
- 데이터의 개수가 n이 짝수라면, n/2번째 데이터와 n/2+1번째 데이터의 평균이 중앙값

최빈값

- 가장 빈도가 많은 데이터 값을 나타낸다.
- 중앙값과 마찬가지로 이상값의 영향을 받지 않는다.

편차

- 개별 관측치에서 평균을 차감한 수

분산

- 평균으로부터 관측치들이 평균적으로 얼마나 떨어져 있는지 요약하는 값
- 편차의 제곱이기 때문에 실제 측정치보다 매우 큰 숫자로 표현

불편분산

- 표본분산은 모집단의 분산, 즉 모분산에 비해 분산을 과소평가해 버리는 경향이 있음
- 이점을 보정하기 위한 것이 불편분산임
- 표본에서 추정한 평균값은 모집단의 평균값과 조금 차이가 있는 게 정상입니다.
- 모평균과 차이가 있는 표본평균을 사용해서 분산을 계산하게 됩니다. 때문에 정확하게 추정하기 힘듬
- 분모가 N-1이 되어 조금 작아지기 때문에 표본분산보다 살짝 더 큰 값을 가지게 됩니다.
- 불편분산 공식

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

- 모분산

$$\sigma^2 = \frac{\sum(x_i - \mu)^2}{n}$$

표준편차

- 분산에 제곱근을 적용해 구한 값
- 분산처럼 변화의 폭을 쉽게 파악할 수 있음
- 실제 관측치의 단위와 동일한 단위로 변화를 파악할 수 있음

통계량 예시

- 불규칙한 통계량을 아는 것이 중요
- 평균값이라는 것은 데이터의 분포 중에서 하나의 수를 꺼낸 것에 불과하며
- 데이터가 그 주변에 어느 정도 퍼져 있는지, 또는 흩어져 있는지는 알수가 없다.
- 버스 운행의 예시 (버스의 시간표는 30분)

```
- A 버스는 2분 늦거나, 2분 빠르게 도착
- B 버스는 10분 늦거나, 10분 빠르게 도착
- A, B 모두 평균은 동일
```

- 버스의 예처럼 평균값보다는 불규칙한 상태의 통계량을 파악할 필요가 있다.
- 평균 : 31분

32	**27**	**29**	**34**	**33**
+1	-4	-2	+3	+2

- 버스 도착시간으로 분산을 이해
  - 각 데이터가 평균값으로부터 어느 정도 큰가, 혹은 작은가를 나타내고 있다. 이 수치를 통계학에선 편차(Deviation)이라고 한다.
  - 위의 5개의 편차를 축약하고, 하나의 수로 대표할 수 있는 평균을 구해보면 0이 된다.
  - 표준편차(Standard Deviation)은 평균을 구하고 싶은 수치들을 각각 제곱하고 모두 합하여 총 개수로 나눈 뒤에 루트를 하는 방법을 취한다. 이 때 제곱한 숫자들의 합을 분산(Variance)라고 한다.
- 버스 시간의 표준편차

```
편차의 합 / 5 의 루트 값은 약 2.6이 나온다. 곧 표준편차가 2.6분인데 이는 평균적으로 시간표보다 1분 늦게 도착하지만(버스 시간표는 30분 도착), 실제 도착 시간은 정해진 시간보다 전후로 대략 2.6분 정도 다를 수 있다고 생각해도 좋다.
```

평균값이 이 데이터의 분포를 대표하는 수치지만, 표준편차는 그 대표값을 기점으로 해서 데이터가 대략 어느 정도 멀리까지 위치해 있는지를 나타내는 통계량이다.

- 표준편차의 의미

평균값은 분포하고 있는 데이터 중에서 대표적인 수로 꺼낸 것이다. 그래서 데이터는 평균값을 기점으로 해서 그 앞 뒤에 널리 퍼져 있다고 생각해도 좋다. 그러나 어느 정도 퍼져 있거나 흩어져 있는 지는 평균값으로 알 수가 없다. 퍼져 있거나 흩어져 있는 정도를 평가하는 것이 표준편차다. 표준편차는 데이터들의 평균값에서 떨어져 있는 것을 평균화하는 것이다. 이 때 멀리 떨어져 있든지 가까운 곳에 있든지, 모두 양수로 평가하여 상쇄하지 않도록 해서 평균을 구한다.

- 누구의 점수가 더 좋은 점수일까?
  - 평균이 60점인 시험에서 당신은 75점을 받았다. 이 때 표준편차가 12점일 때와 표준편차가 8점일때 중 어느 경우가 기분이 더 좋을까? (물론 성적이 낮아서 더 기분 좋다고 생각하는 사람이 없다고 가정하에...)
  - 표준편차가 12점인 경우 보다 편차가 8점일때 내가 받은 점수 75점이 더 좋은 성적일 것이다.
  - 한 데이터 세트 중에 있는 어떤 하나의 데이터가 가진 특수성은 평균에서 떨어진 정도인 편차만으로 설명할 수 없고, 표준편차를 기준으로 가정해야만 알 수 있다.
- 데이터 특수성의 평가기준
  - 데이터 세트 중에 있는 어느 한 데이터의 편차가 표준편차로 계산해서 ± 1배 전후라면 이 것은 '평범한 데이터'
  - ±2배로 멀리 있다면 이 데이터는 특수한 데이터라고 할 수 있다.
  - 데이터가 정규분포라고 가정하면 평균값에서 표준편차 ± 1배의 범위 내에 약 70%의 데이터가 들어간다고 생각하면 된다.
  - 표준편차 ± 2배보다 멀리 떨어진 데이터는 좌우 양쪽 합쳐서 5%밖에 없다고 생각하면 대략 맞다
  - 당신의 데이터가 평균값보다 큰 쪽으로 표준편차 2배 이상 떨어져 있다면, 그것은 전체의 2.5% 범위 내에 드는 데이터라는 것을 의미하기 때문에 상당히 특수한 경우에 있다고 해도 좋을 것이다.

표본 추출

표본 개념

모수

- 관심을 갖고 있는 모집단 관측치의 대푯값
- 대표적인 모수는 모비율, 모평균, 모총계등 있음

통계량

- 표본(Sample)을 조사하여 얻은 데이터를 가지고 모수를 추정하기 위해 만든 공식

- 표본을 뽑을 때마다 통계량이 달라지는 것을 표본 추출 변동이라고 함