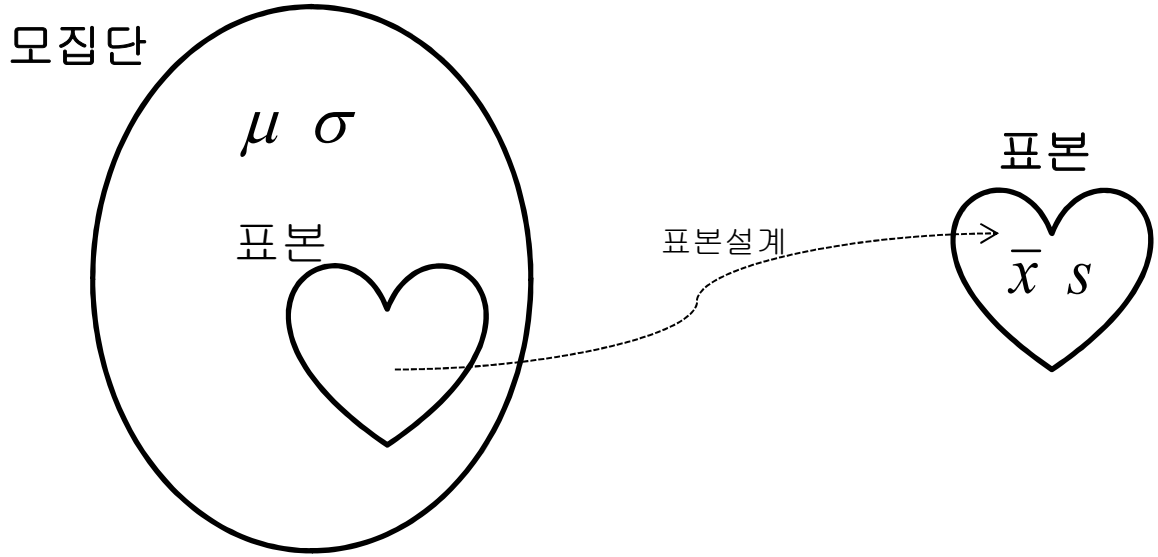

이화여자대학교
EWHA WOMANS UNIVERSITY
확률 및 통계학
기술통계
송수민
soominsong@ewha.ac.kr

# 통계 용어

- 모집단(population)
  : 통계적 질문과 관련하여 관심 있는 개체들 전체의 집합
- 표본(sample) : 모집단에서 선택된 일부 개체의 집합
- 변수(variable) : 개체의 특성을 나타내는 수치적 항목
- 모수(parameter) : 모집단의 수치적 특성
- 통계량(statistics) : 표본의 수치적 특성



# 통계 용어 예

서울 시장 선거에서 **A** 후보의 지지율을 알고 싶다.
**500**명의 유권자를 랜덤하게 뽑아 **A** 후보에 대한 지지여부를 물어보았다.

- 모집단
- 표본
- 관심 변수
- 관심 모수
- 관심 통계량

서울시 중학생들의 몸무게의 평균을  알고싶다.
**1000**명의 서울시 중학생을 랜덤하게 뽑아 몸무게를 측정해 보았다.

- 모집단
- 표본
- 관심 변수
- 관심 모수
- 관심 통계량

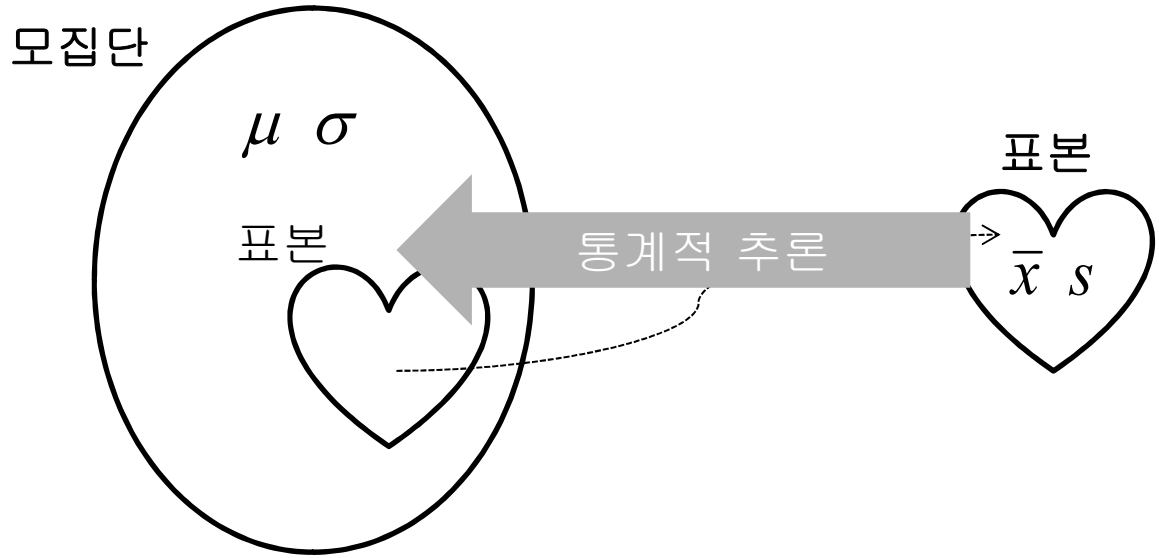
# 통계학의 방법론

- 기술통계

  그래프, 도표, 통계량으로 전체 자료의 특성을 요약하는 기법

- 통계 추론

  - 관심 모집단에서 얻어진 표본자료로 전체 모집단의 특성을 추정하는 기법
  - 기술통계 방법이 선행되어야 함



# 기술통계 활용 예

## 통계청에서 주관하는 인구주택총조사

❍ 인구총조사는 1925년, 주택총조사는 1960년 이후 매 5년마다 실시

- 대한민국 정부 최초의 총조사는 1949년에 실시

-2010년도에 실시한 인구총조사는 제18차, 주택총조사는 제10차에 해당

❍ 조사대상 : 조사기준 시점 현재 대한민국 영토 내에 상주하는 모든 내.외국인과 이들이 살고 있는 거처

❍ 조사항목

| | 전수 (19) | 표본 (31) |
|---|---|---|
| 인구 (28) | ① 성명<br>② 성별<br>③ 나이<br>④ 가구주와의 관계<br>⑤ 교육정도<br>⑥ 혼인상태<br>⑦ 국적<br>⑧ 입국연월 | ① 아들보육 ② 출생지<br>③ 1년 전 거주지 ④ 5년 전 거주지<br>⑤ 출산계획 ⑥ 통근통학 여부<br>⑦ 통근통학 장소 ⑧ 이동교통수단<br>⑨ 통근통학 소요시간 ⑩ 사회활동<br>⑪ 경제활동상태 ⑫ 종사상 지위<br>⑬ 산업 ⑭ 직업<br>⑮ 총 직업 근무연수 ⑯ 교육정도<br>⑰ 혼인 연월 ⑱ 총 출생아 수<br>⑲ 추가계획 자녀 수 ⑳ 고령자생활비 원천 |
| 가구 (13) | ① 가구구분<br>② 사용방수<br>③ 주거시설형태<br>④ 점유형태<br>⑤ 렌틀 및 거주 층<br>⑥ 주인가구 및 타지 주택 소유여부 | ① 거주기관 ② 난방시설<br>③ 수도 및 식수 사용 형태 ④ 정보통신기기 보유 및 이용<br>⑤ 교통수단 보유 및 이용현황 ⑥ 주차장소<br>⑦ 일차로 |

…

** 2010 인구주택총조사 전수집계결과 자료에서 발췌

❍ 일반가구의 96.7%인 16,773천 가구가 지상에 거주하며, 지하(반지하)에는 3.0%, 옥상(옥탑)에는 0.3%가 거주

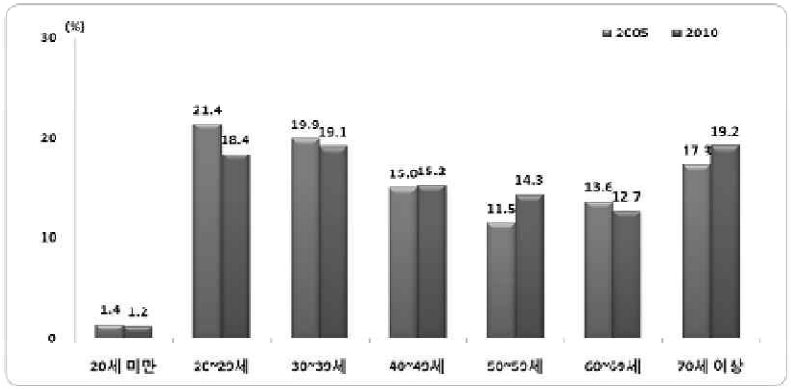
< 표 21 > 거주층 현황(2005, 2010)

(단위 : 천 가구, %)

| 구 분 | 2005년 | | | | 2010년 | | | |
|---|---|---|---|---|---|---|---|---|
| | | 지상 | 지하 (반지하) | 옥상 (옥탑) | | 지상 | 지하 (반지하) | 옥상 (옥탑) |
| 전 국 | 15,887 (100.0) | 15,249 (96.0) | 587 (3.7) | 51 (0.3) | 17,339 (100.0) | 16,773 (96.7) | 518 (3.0) | 49 (0.3) |
| 읍 부 | 1,319 (100.0) | 1,315 (99.7) | 3 (0.2) | 1 (0.1) | 1,468 (100.0) | 1,464 (99.8) | 3 (0.2) | 0 (0.0) |
| 면 부 | 1,823 (100.0) | 1,820 (99.8) | 3 (0.2) | 0 (0.0) | 1,821 (100.0) | 1,818 (99.8) | 3 (0.2) | 0 (0.0) |
| 동 부 | 12,745 (100.0) | 12,114 (95.1) | 581 (4.6) | 50 (0.4) | 14,031 (100.0) | 13,471 (96.0) | 512 (3.3) | 48 (0.3) |

**1인 가구의 19.2%는 70세 이상 고령자**

< 그림 13 > 연령별 1인 가구 비율(2005, 2010)



# 통계추론의 예

## 10·26 서울시장 보궐선거 여론조사 빗나가고 출구조사 근접

2011년 11월 02일 (수) 15:07:05 원성윤 기자 sucool@journalist.kr

10·26 서울시장 보궐선거에서 방송사의 공동 출구조사가 근사치에 다가갔다. 반면 여야 후보의 박빙 승부를 예측했던 여론조사는 빗나갔다.

서울시장 투표가 끝난 직후인 지난달 26일 오후 8시에 발표된 방송 3사(KBS, MBC, SBS)의 출구조사 결과는 무소속 박원순 후보가 54.4%, 한나라당 나경원 후보가 45.2%를 기록했다. 최종 결과는 박원순 후보 53.4%, 나경원 후보 46.2%로 격차는 7.2%였다. 1.0% 오차를 감안해도 실제 결과와 크게 다르지 않았다.

반면 투표 1주일 전까지 여야 후보의 지지도가 엎치락뒤치락했던 여론조사는 실제결과와 다소 차이가 있었다.

주요 여론조사를 비교해 보면 휴대전화 합산 방식을 도입한 곳에서는 박원순 후보의 우세 결과가 나왔고, 유선전화만 합산한 곳에서는 나경원 후보의 우세하다는 결과가 나왔다.

유·무선 합산방식을 도입한 방송3사(16~17일) 조사에서는 박원순 40.5%·나경원 38.2%, YTN(17~19일) 조사에서는 박원순 44.3%·나경원 39.3%, 조선일보(19일) 조사에서는 박원순 43.5%·나경원 41.4%로 나와 모두 박원순 후보의 우세를 점쳤다.

반면 유선만으로 조사한 동아일보(16~17일) 조사에서는 나경원 42.4%·박원순 41.1%, 국민일보(18일) 조사에서는 나경원 42.2%·박원순 39.3%로 집계됐다. 문화일보(19일) 조사에서는 나경원 47.7%·박원순 37.6%로 10% 이상 차이로 나 후보의 우세를 점쳐 실제와 크게 동떨어진 결과를 나타냈다.

홍형식 한길리서치 소장은 "향후 집전화와 휴대전화 합산 비율을 조정하는 작업을 거치면 근사치에 더욱 가깝게 다가갈 것"이라고 답했다.

# 변수 측정 척도 (1) - 범주형

- 명목척도(**Normal Scale**)

  - 분류척도
  - 변수의 값에 단순히 명목상의 의미로 이름, 숫자, 기호 등을 부여
  - 변수의 수준들 사이에 특별한 대소 관계가 없으며, 수준을 표시하는 기호와 숫자를 임의로 바꾸어도 정보의 손실이 없음
  - 예시) 성별 : 여자 = 0, 남자 = 1
    맨유 선수 : 박지성=13, 웨인 루니=10, 하비에르 에르난데스=14, …

- 순서척도(**Ordinal Scale**)

  - 서열척도
  - 각각의 수준들 사이에 명확한 순서가 있음
  - 예시) 언어 구사 능력 : 상, 중, 하

  서비스 만족도 : 매우 불만족 = 1, 불만족 = 2, 보통 = 3, 만족 = 4, 매우 만족 = 5

# 변수 측정 척도 (2) - 연속형

- 구간척도**(Interval Scale)**

  - 등간척도
  - 변수 값의 차이를 비교 측정할 수 있음 (+ , -)
  - 0이 의미를 갖지 않음
  - 예시) 섭씨 온도, 화씨 온도, 주가 지수, IQ점수

- 비척도**(Ratio Scale)**

  - 변수 값 간의 비가 의미를 가짐(+ , - , * , / )
  - 0이 '전혀 없음' 의 의미를 가짐
  - 예시) 절대 온도, 길이, 무게, 부피, 가격, 내가 소유한 자동차의 수

✓ 자료의 측정 척도에 따라 분석 방법이 달라짐

# 변수 측정 척도 (3)

다음 변수들의 측정 척도는 무엇인가

1. 대통령 선거 출구조사에서 응답 대상자의 성별
2. 화씨온도로 측정된 S병원 환자들의 체온
3. 태권도 선수의 체급
4. 소비자 물가지수
5. 1월 4일 내린 비의 강우량
6. O 농장에서 생산되는 꽃의 색
7. N 중학교 학생들의 몸무게
8. 우리동네 G 중국집의 자장면 사이즈 (보통<곱빼기<특)

# 도수분포표와 도수그래프 (1)

- **범주형 자료의 도수분포표** (예 : 교재 5페이지 표 1.1)
  - 하나의 범주에 해당하는 관측의 수를 도수 또는 빈도라고 함

**우리집 식구들에게 가장 선호하는 남자 배우를 물어보았다.**

데이터

| 식구 | 선호 남자배우 |
|---|---|
| 할머니 | 원빈 |
| 할아버지 | 장동건 |
| 엄마 | 현빈 |
| 아빠 | 현빈 |
| 나 | 소지섭 |
| 언니 | 장근석 |
| 오빠 | 소지섭 |
| 남동생 | 유아인 |
| 여동생 | 장근석 |
| 이모1 | 소지섭 |
| 이모2 | 소지섭 |
| 외삼촌 | 원빈 |
| 고모 | 현빈 |
| 삼촌1 | 장동건 |
| 삼촌2 | 원빈 |

도수분포표

| 선호 남자배우 | 도수 |
|---|---|
| | |
| | |
| | |
| | |
| | |
| | |
| 총합계 | 15 |

도수그래프

# 도수분포표와 도수그래프 (2)

- **연속형 자료의 도수분포표** (예 : 교재 6페이지 표 1.3)

  - 연속형 자료 도수분포표 작성 방법

    1. 전체 자료를 크기 순으로 나열

    2. 관측값의 범위를 적당한 숫자의 계급구간으로 나눔(보통 5-20개 정도)

  - 일반적으로 모든 자료가 하나의 구간에 할당되도록 하기 위해 마지막 유효숫자 아래 소수점을 기준으로 5가 되도록 구간의 끝점을 지정
  - 단점 : 자료를 그룹화하면서 정보의 손실이 생김

    장점 : 전체 자료의 특성을 한눈에 쉽게 파악할 수 있음
  - 연속형 자료의 도수그래프를 히스토그램이라 함

# 도수분포표와 도수그래프 (3)


영화 『트와일라잇』 관람자 **30**명을 대상으로 평점을 조사하였다.

평점 데이터

| | | |
|---|---|---|
| 67 | 56 | 81 |
| 88 | 99 | 85 |
| 92 | 81 | 77 |
| 93 | 81 | 79 |
| 95 | 52 | 80 |
| 75 | 65 | 38 |
| 25 | 66 | 42 |
| 71 | 61 | 13 |
| 86 | 98 | 85 |
| 88 | 5 | 76 |

순위화된 평점 데이터

| | | |
|---|---|---|
| 5 | 67 | 85 |
| 13 | 71 | 85 |
| 25 | 75 | 86 |
| 38 | 76 | 88 |
| 42 | 77 | 88 |
| 52 | 79 | 92 |
| 56 | 80 | 93 |
| 61 | 81 | 95 |
| 65 | 81 | 98 |
| 66 | 81 | 99 |

첫번째 구간의 끝점을 0.5
구간의 폭을 10으로

도수분포표

| 평점 | 도수 |
|---|---|
| | |
| | |
| | |
| | |
| | |
| | |
| | |
| | |
| | |
| | |
| 계 | 30 |

도수그래프

# 도수분포표와 도수그래프 (4)

● 상대도수와 누적도수 (예 : 교재 11페이지 표 1.6)

- 상대도수는 절대도수를 전체 자료의 개수로 나누어 준 것
- 누적도수는 각 구간의 상한값 이하 모든 관측값들의 수
- 누적상대도수는 누적도수를 전체 자료의 개수로 나누어 준 것
- 상대도수분포표는 관측값의 개수가 서로 다른 데이터를 비교할 때 용이 (교재 11페이지 그림 1.5 참고)

『트와일라잇』 평점 도수분포표로 누적도수, 상대도수, 누적상대도수를 표에 작성

| 평점 | 도수 | 누적도수 | 상대도수 | 누적상대도수 | 상대도수(%) | 누적도수(%) |
|---|---|---|---|---|---|---|
| 0.5-10.5 | | | | | | |
| 10.5-20.5 | | | | | | |
| 20.5-30.5 | | | | | | |
| 30.5-40.5 | | | | | | |
| 40.5-50.5 | | | | | | |
| 50.5-60.5 | | | | | | |
| 60.5-70.5 | | | | | | |
| 70.5-80.5 | | | | | | |
| 80.5-90.5 | | | | | | |
| 90.5- | | | | | | |
| 계 | 30 | 30 | 1.000 | 1.000 | 100.0 | 100.0 |

# 위치를 나타내는 통계량 (1)

- 평균**(mean)**
- 중앙값**(median)**
- 최빈값**(mode)**
- 그 외에도 절단평균, 윈저화평균 등의통계량이 존재

# 위치를 나타내는 통계량 (2)

## 『트와일라잇』 평점 데이터

순위화된 평점 데이터

| | | |
|---|---|---|
| 5 | 67 | 85 |
| 13 | 71 | 85 |
| 25 | 75 | 86 |
| 38 | 76 | 88 |
| 42 | 77 | 88 |
| 52 | 79 | 92 |
| 56 | 80 | 93 |
| 61 | 81 | 95 |
| 65 | 81 | 98 |
| 66 | 81 | 99 |

- 평균
- 중앙값
- 최빈값

# 변동을 나타내는 통계량 (1)

- 범위**(range)**
- 분산**(variance)**
- 표준편차**(standard deviation)**
- 사분위범위**(IQR : interquartile range)**

# 변동을 나타내는 통계량 (2)

『트와일라잇』 평점 데이터

| 순위화된 평점 데이터 | | |
|---|---|---|
| 5 | 67 | 85 |
| 13 | 71 | 85 |
| 25 | 75 | 86 |
| 38 | 76 | 88 |
| 42 | 77 | 88 |
| 52 | 79 | 92 |
| 56 | 80 | 93 |
| 61 | 81 | 95 |
| 65 | 81 | 98 |
| 66 | 81 | 99 |

- 범위
- 분산
- 표준편차
- 사분위범위

# 줄기 잎 그림

- 줄기 잎 그림의 특징 (예 : 교재 20페이지 그림 1.8)

  - 분포의 모양, 분포의 집중도, 범위 등을 한눈에 알 수 있는 그림
  - 자료의 손실이 없이 전반적인 자료의 추이를 쉽게 알 수 있음
  - 줄기와 잎을 결정하는 방법은 일률적이지 않음

『트와일라잇』 평점 데이터

순위화된 평점 테이터

| | | |
|---|---|---|
| 5 | 67 | 85 |
| 13 | 71 | 85 |
| 25 | 75 | 86 |
| 38 | 76 | 88 |
| 42 | 77 | 88 |
| 52 | 79 | 92 |
| 56 | 80 | 93 |
| 61 | 81 | 95 |
| 65 | 81 | 98 |
| 66 | 81 | 99 |

# 상자그림

- 상자그림의 특징 (예 : 교재 21페이지 그림 1.9)
  - 분포의 모양, 분포의 집중도, 범위 등을 한눈에 알 수 있는 그림
  - 중앙값, 사분위값, 최대값, 최소값을 그림에 표시

『트와일라잇』 평점 데이터

순위화된 평점 데이터

| | | |
|---|---|---|
| 5 | 67 | 85 |
| 13 | 71 | 85 |
| 25 | 75 | 86 |
| 38 | 76 | 88 |
| 42 | 77 | 88 |
| 52 | 79 | 92 |
| 56 | 80 | 93 |
| 61 | 81 | 95 |
| 65 | 81 | 98 |
| 66 | 81 | 99 |

# 기술통계 예제

다음 데이터로 위치, 변동을 나타내는 통계량과 줄기잎그림, 상자그림을 구하라.

확률과 기초통계 중간고사 점수 데이터

| | | | | |
|---|---|---|---|---|
| 35 | 55 | 61 | 62 | 62 |
| 38 | 65 | 57 | 48 | 54 |
| 57 | 56 | 57 | 59 | 59 |
| 47 | 53 | 55 | 56 | 56 |
| 54 | 57 | 52 | 57 | 54 |
| 35 | 48 | 57 | 63 | 57 |
| 55 | 58 | 60 | 52 | 61 |
| 58 | 47 | 54 | 72 | 65 |
| 53 | 56 | 58 | 61 | 51 |
| 54 | 53 | 53 | 45 | 58 |

순위화 →

확률과 기초통계 중간고사 점수 데이터

| | | | | |
|---|---|---|---|---|
| 35 | 52 | 55 | 57 | 60 |
| 35 | 53 | 55 | 57 | 61 |
| 38 | 53 | 55 | 57 | 61 |
| 45 | 53 | 56 | 57 | 61 |
| 47 | 53 | 56 | 58 | 62 |
| 47 | 54 | 56 | 58 | 62 |
| 48 | 54 | 56 | 58 | 63 |
| 48 | 54 | 57 | 58 | 65 |
| 51 | 54 | 57 | 59 | 65 |
| 52 | 54 | 57 | 59 | 72 |

# [별첨] 그룹화된 자료 (1)

● 그룹화된 자료의 평균과 분산

k개의 계급구간이 있는 그룹화 자료에서

$m_i$ = i번째 계급구간의 중간점

$f_i$ = i번째 계급구간의 도수

n = 총관측수 라 할 때,

- 그룹화 자료의 평균 $\bar{x} = \sum_{i=1}^{k} m_i f_i / n$

- 그룹화 자료의 분산

$$s^2 = \sum_{i=1}^{k} (m_i - \bar{x})^2 f_i / (n-1) = (\sum_{i=1}^{k} m_i^{\,2} f_i - n\bar{x}^{2}) / (n-1)$$

# [별첨] 그룹화된 자료 (2)

## 『트와일라잇』 평점 도수분포표 예

| 평점 | $f_i$ |
|---|---|
| 0.5-10.5 | 1 |
| 10.5-20.5 | 1 |
| 20.5-30.5 | 1 |
| 30.5-40.5 | 1 |
| 40.5-50.5 | 1 |
| 50.5-60.5 | 2 |
| 60.5-70.5 | 4 |
| 70.5-80.5 | 6 |
| 80.5-90.5 | 8 |
| 90.5-100.5 | 5 |
| 계 | 30 |

- 평균

- 분산