

데이터와 분포

과학기술연합대학원대학교
한국생명공학연구원 스쿨
시스템생명공학전공

haseong@kribb.re.kr
김하성 선임연구원

강의 개요

- 강사

- 김하성 (한국생명공학연구원 합성생물학전문연구단, 선임연구원)
- haseong@kribb.re.kr (연구동 1143, 042-860-4372)
- 1,2학기 일반강의: 기초통계, R프로그래밍, 생물데이터분석
- 전공강의: 시스템합성생물학

- 강의 목표

- 통계의 필요성
- 데이터의 기초 개념 이해
- 데이터 분석의 기초 사고능력 배양
- 분포의 개념을 이해
- 데이터의 통계적 검증이 갖는 의미를 이해

- 평가

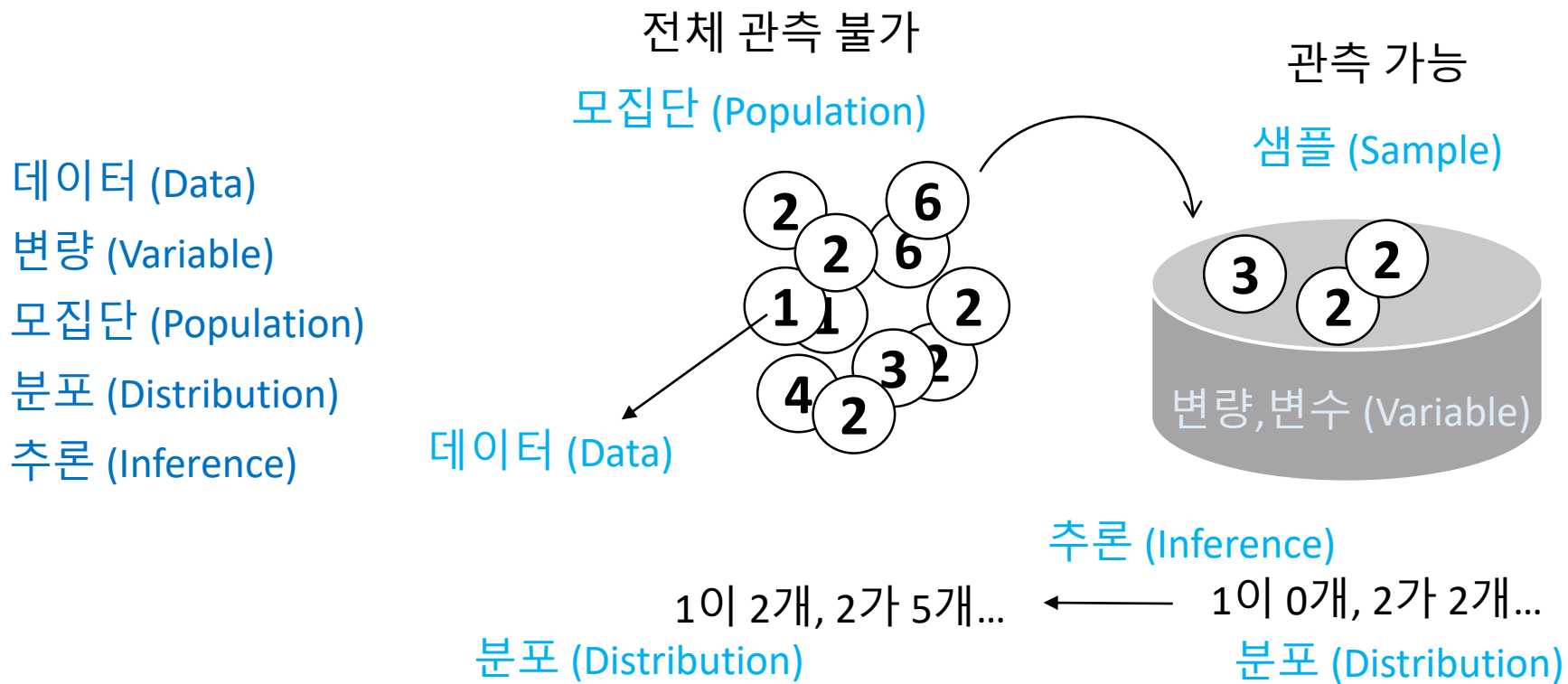
- 출석 50%, 과제 50%

참고자료



강의 목차

- 1일차 : 데이터 개요, 일변량 데이터
- 2일차 : 이변량 데이터
- 3일차 : 다변량 데이터와 모집단
- 4일차 : 분포와 통계적 추론



통계란 무엇인가?

2대 점쟁이 문어 파울



수명 3년, 사망

<월드컵2010>

독일 - 호주: 독일 승→O

독일 - 세르비아: 독일 패→O^[2]

독일 - 가나: 독일 승→O

독일 - 잉글랜드: 독일 승→O

독일 - 아르헨티나: 독일 승→O

독일 - 스페인: 독일 패→O

독일 - 우루과이: 독일 승→O

스페인 - 네덜란드: 스페인 승→O^[3]

데이터를 수집, 정리하여 이로부터
미지의 사실에 대한 신빙성 있는
추론을 수행하는 과정

➡ 데이터 분석을 통한 예측

What is the data?

- A set of values of subjects with respect to qualitative or quantitative variables -wiki-
- 사실을 나타내는 수치

What is the difference with information?

- Processed, organized, structured data in a given context so as to make it useful
- 의미 있는 데이터

지혜 (wisdom) : 패턴화된 지식

지식 (knowledge) : 가치있는 정보

정보 (information) : 의미있는 데이터

데이터 (data) : 단순한 사실의 나열



맥도너 (A.M.McDonough) 정보경제학(1963)

출처: 경기사이버도서관

데이터 in 대한민국

문대통령 "데이터는 4차혁명 '원유'... 1조투자해 산업육성"

"빅데이터 기반을 신기술·신산업 키워야"... 문 대통령, 정부 데이터 활용·개인정보 보호 주문

18.03.17 16:00 | 최종 업데이트 18.03.17 16:00 | 송정익(sungjeik@naver.com)



▶ 문재인 대통령이 17일 서울 용산구 용산문화재단 대강당에서 열린 데이터 유망사업 설명회를 열어 이렇게 말씀하고 있다.

통계청장 경질 논란, 통계참사인가 대응참사인가

송진식 기자 tsujin@kyunghyang.com



문재인 대통령이 5월 29일 경제 관련 부처 장관들이 참석한 가운데 가계소득통합 점검회의를 주재하고 있다. / 청와대 제공

MSIT PR > 보도·해명자료 > 보도자료



과기정통부, 데이터·인공지능(AI) 경제 활성화의 이정표 제시

융합신산업과, 지능정보사회추진단 인공지능정책팀 | 이주식 서기관, 김근영 사무관 연락처 | T : 02-2110-2845, 02-2110-1617

16
19.01

과기정통부, 데이터·인공지능(AI) 경제 활성화의 이정표 제시

- '23년 국내 데이터시장 30조원 규모 성장 -

- 인공지능 유니콘기업 10개 육성 -

- 인공지능 융합 클러스터 조성 및 전문인력 1만명 양성 -

Applications : Policy decision

Prevalence of guns in the movies violence **VS.** violent people



Evidence of cause and effect

https://medium.com/@MrBrown_110/shooting-stars-hollywoods-gun-problem-80c09ef932bb

Price of a hip replacement **VS.** health care quality

Price and transparency



<https://orthoinfo.aaos.org/en/treatment/total-hip-replacement/>

Applications : Industry

23 deadly accidents and 475 fatalities **VS.** one death / 45,000,000 flights (2012)
(One could fly daily for 123,000 years before a fatal plane crash)



demonstrate substantial effect

<https://news.joins.com/article/14644734>

To establish a financial advantage



성격 검사 용도로 수집한 8700여 명의 페이스북 사용자 정보가 지난 미국 대선에서 도널드 트럼프 대통령의 선거캠프에 흘러 들어감



<https://www.facebook.com/notes/facebook-engineering/visualizing-friendships/469716398919>

일변량 (Univariate) data set

변수 (Variable)

- 어떤 현상(사물)을 이해하기 위해서 특징을 선별하고 값을 측정할 때, 특징 = 변수
- 일변량 데이터셋은 임의의 현상을 이해하기 위해 변수 (특징) 하나를 선별해서 측정한 값들의 집합을 말함

$$x_1, x_2, \dots, x_n$$

n measurements

예제

년도: 1964, 1965, 1966, 1967, ...

양의수: 1,2,3,5,8,9,...

달리기 기록: 17, 16, 23, ...

이름: 루크, 한, 오비완, 추바카, 레아, ...

연도별 야구선수 평균 연봉: 0.57, 0.89, 1.08, 1.12, 1.18, 1.07, 1.38



변수



측정값

Four types of variables (Levels of measurement)

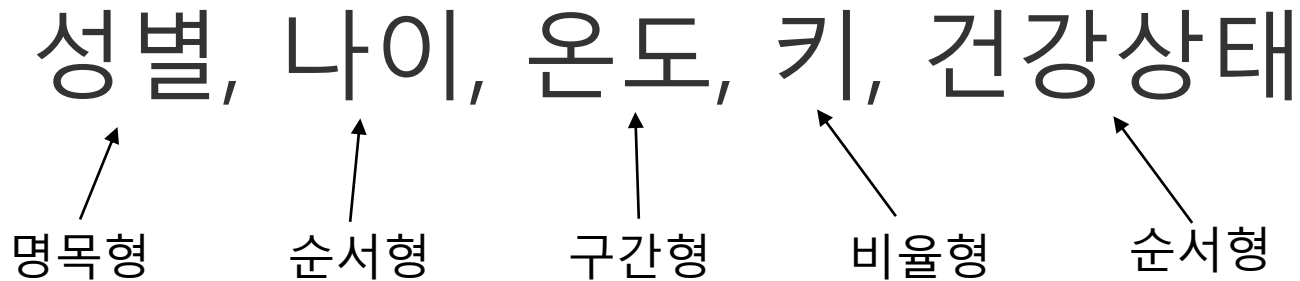
- Nominal (명목형) – 사람 이름
 - Ordinal (순서형) – 달리기 도착 순서
 - Interval (구간형) – 선수1, 선수2 종점통과 시간
 - Ratio (비율형) – 출발시간 기준 종점 통과 시간
- } 범주형 (categorical)
- } 수치형 (numerical)

이름	등수	도착	걸린시간
둘리	1	13:12	1:12
희동	5	14:30	2:30
길동	2	13:30	1:30
철수	4	14:00	2:00
영희	3	13:50	1:50

정보의 수준?

데이터 타입 분류

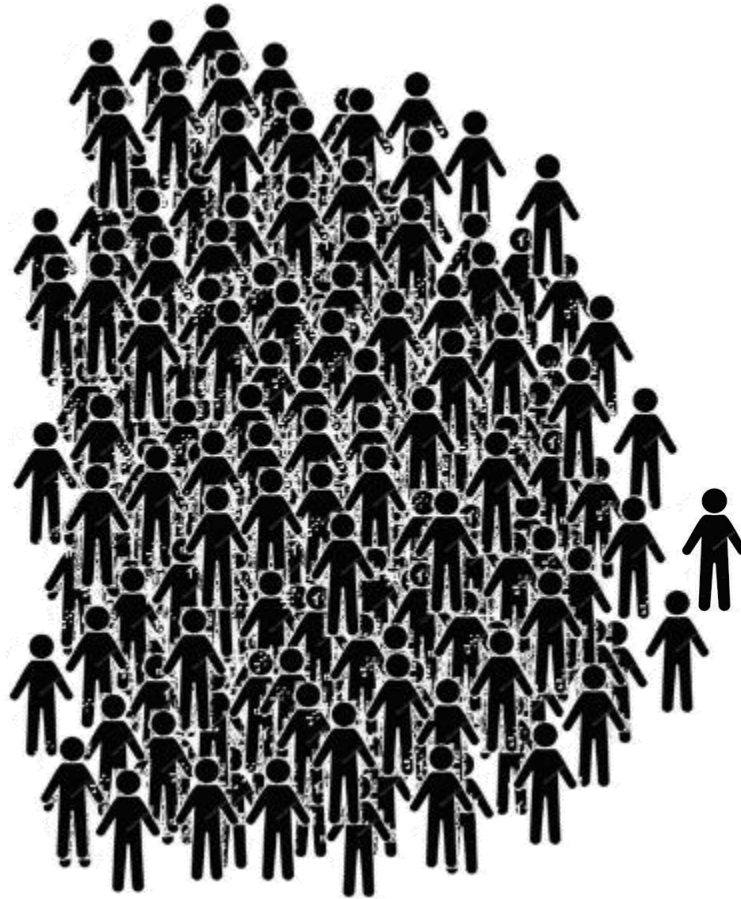
명목형, 순서형, 구간형, 비율형



* 구간형 - 측정대상이 갖고 있는 속성의 양적인 정도에 따라 등간격으로 수치 부여. 해당속성이 전혀 없는 상태인 절대적 원점(absolute zero)이 존재하지 않음. 예로 온도는 0도를 가지고 있지만, 실제로 '0'라는 양을 의미하거나 온도가 없는 것이 아님.

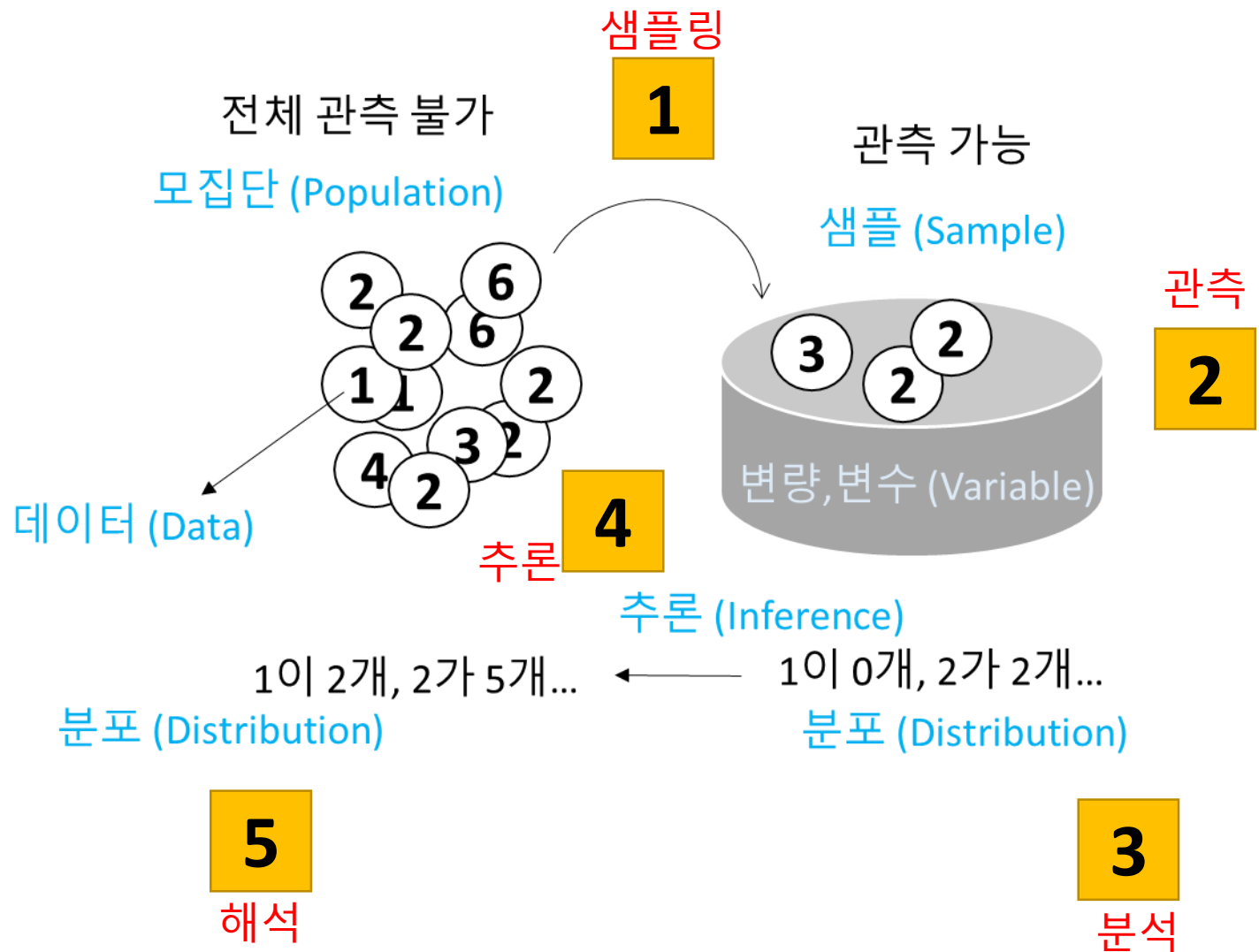
정보의 수준, 보다 정밀한 분석 방법 적용 가능
명목 < 순서 < 구간 < 비율

Q. 대한민국 5살 아이들의 키는?

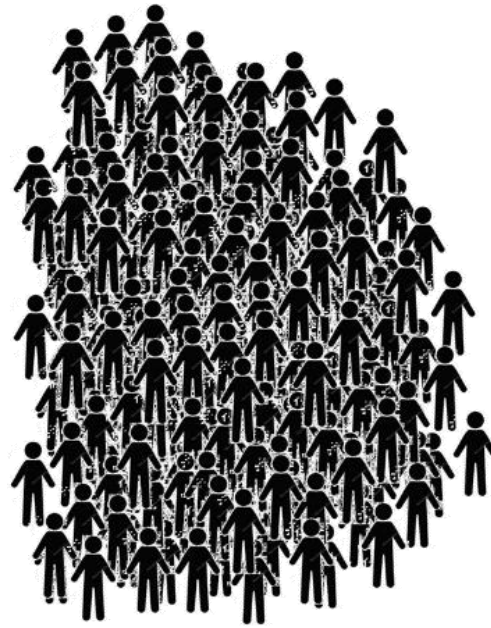


Variable?
Data?

대답은?



Sampling



Boys : Girls

Random sampling

Population의 분포를 가장 잘 나타내는 sample



height

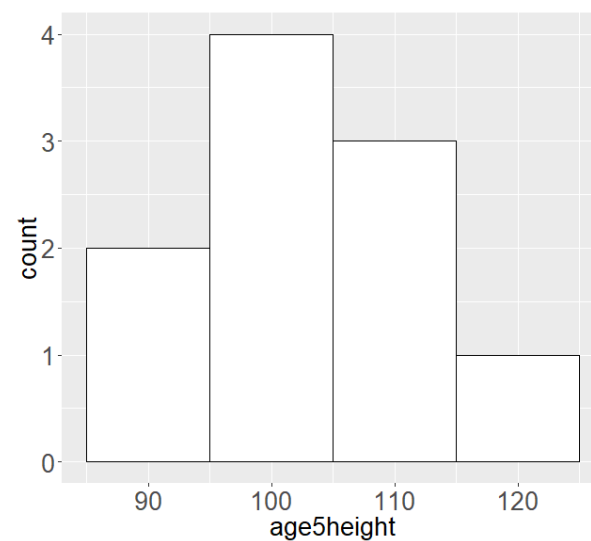
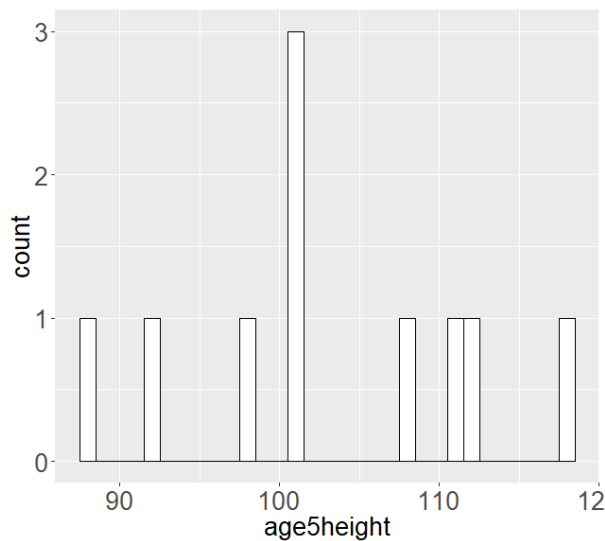
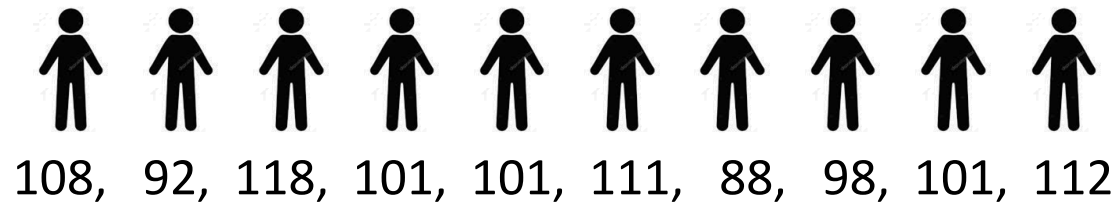
108, 92, 118, 101, 101, 111, 88, 98, 101, 112

gender

F, F, M, F, F, F, M, M, M, F

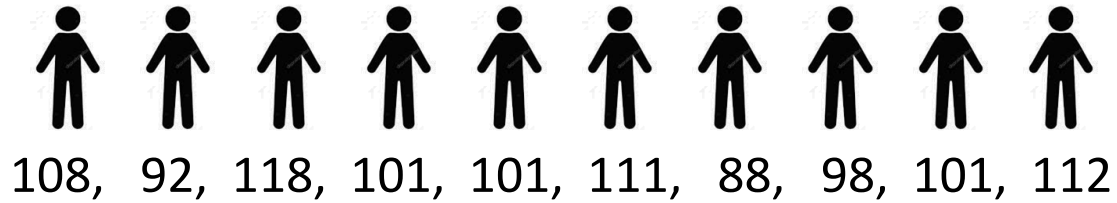
Viewing the data

- Histogram – Break up an interval (구간 나누기), for each subinterval the number of data points are counted



how to compare with population distribution?

Summary of data



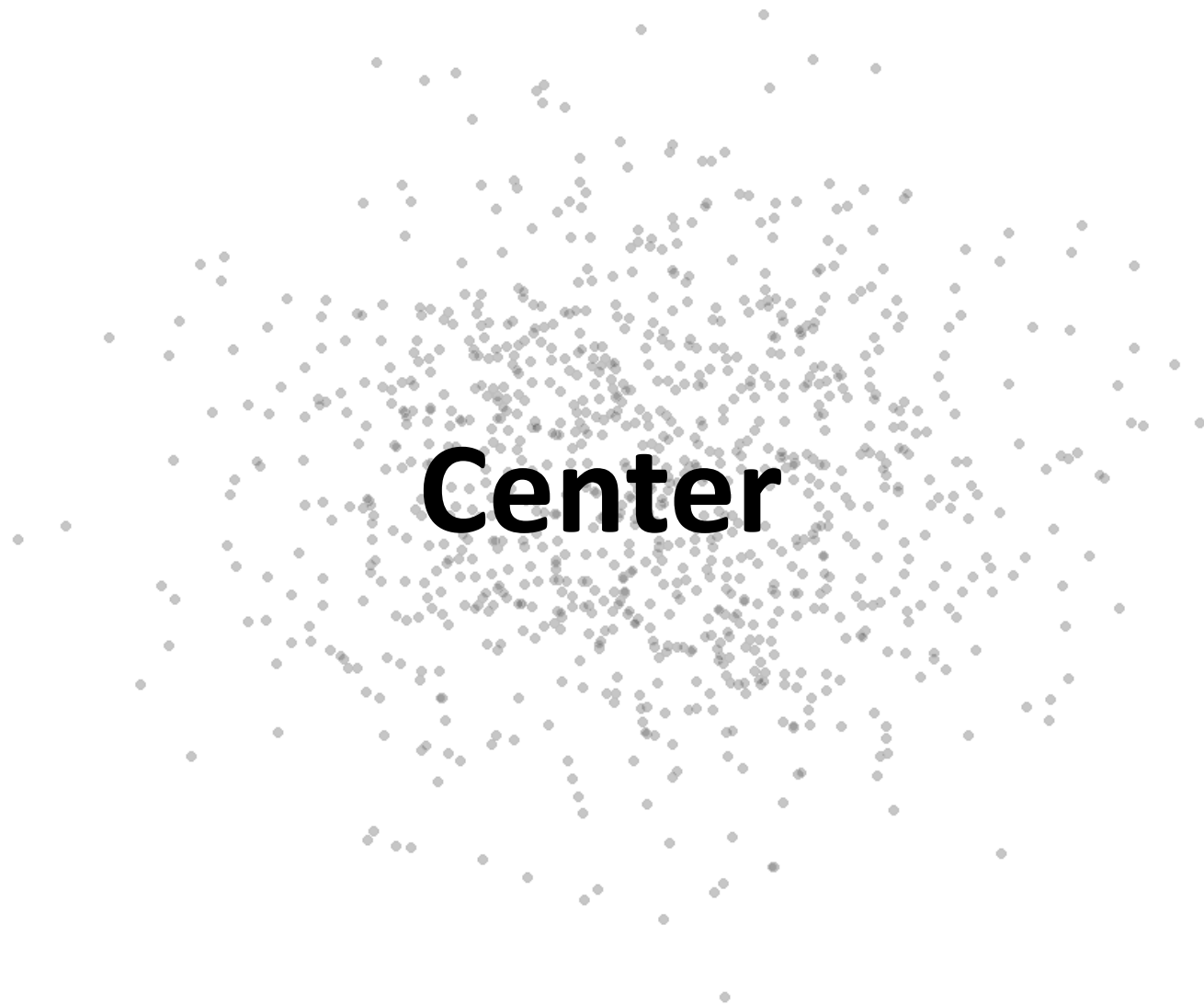
(Sample) Mean: 103



(Population) Mean: 103

Summaries (of numerical data)

- **Center** – commonly known as “**average**” or “**mean**” but not the only one. **median, mode**, etc
- **Spread – Variability** of a data set. No variability – mean is everything vs. Large variability – mean informs much less. confidence of interpretation from knowing center.
Distance from center.
- **Shape** – Degree of interpretation from knowing center and spread. eg) bell shape – two sides are equally likely, large values are rather unlikely and values tend to cluster near the center.



Center

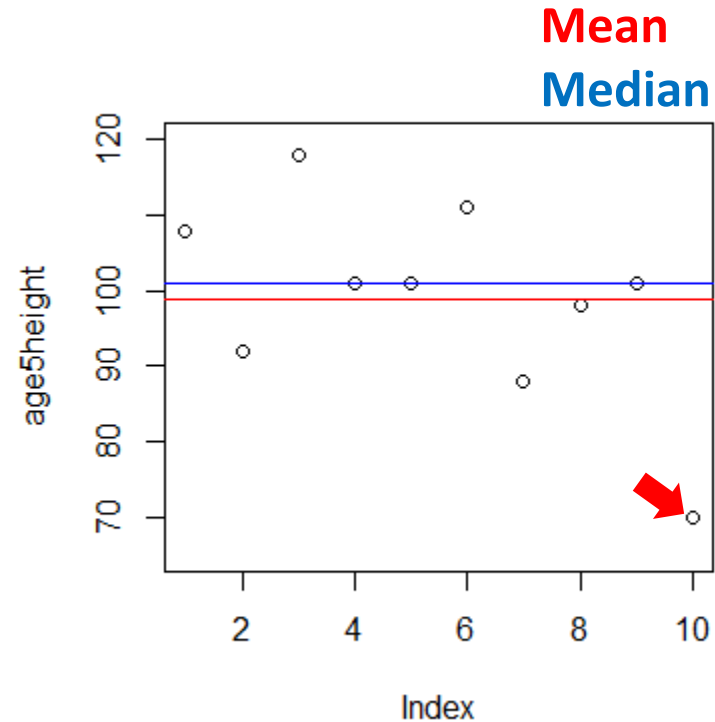
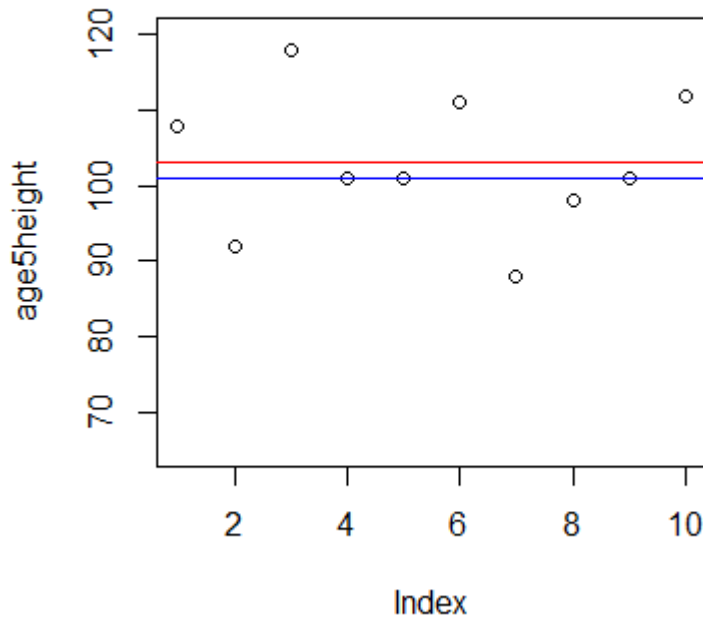
Mean

$$\text{sample mean} = \bar{x} = \frac{1}{n}(x_1 + x_2 + \dots + x_n) = \frac{1}{n} \sum_{i=1}^n x_i$$

Median splits the data in half = 0.5 quantile

***pth quantile:** 100p percent of the data is less than the value, 100(1-p) is more

Center

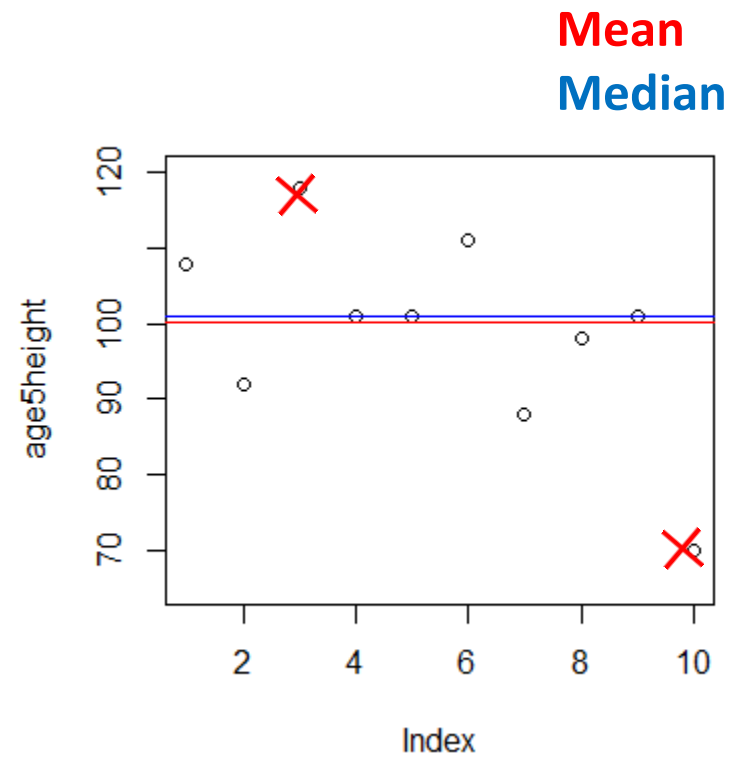
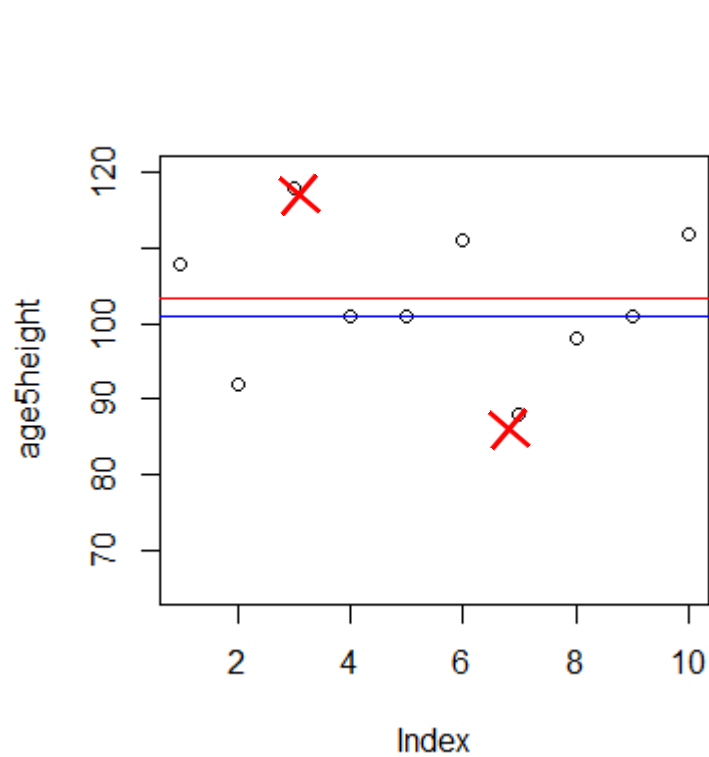


Which one is better?

The median suffers from poor marketing

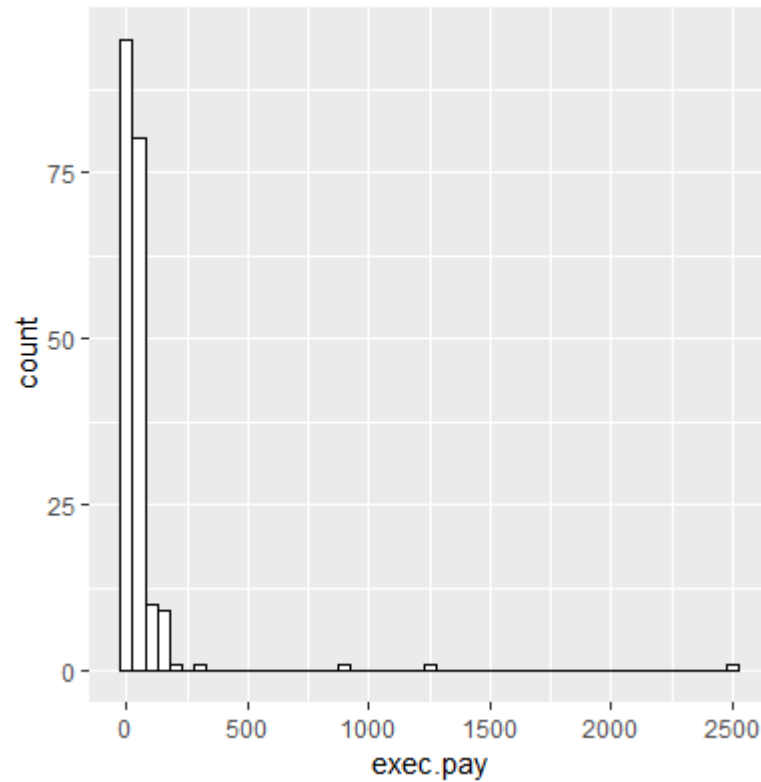
<https://creativemaths.net/blog/median/>

Trimmed mean



Skewed data

Income of CEOs in some American companies



중앙값 (Median)

136 74 8 38 46 43

- Sort the data from smallest to largest
- $n = 2k + 1$ data points, median is the $k+1$ st number
- $n = 2k$, median is $(k\text{th} + k+1\text{th numbers}) / 2$
- 0.5 quantile

분위수 (Quantile)

- Generalized concept of the median
- p th quantile: 이 값보다 작은 데이터의 비율이 $100 \cdot p$ 퍼센트, 큰 데이터의 비율은 $100 \cdot (1 - p)$ 퍼센트
- Ex)
 - $p=0.5$, 밑으로 50% 위로 50% 데이터를 갖는 위치의 값: median
 - $p=0.25$, 25% 밑으로 25% 위로 75% 데이터를 갖는 위치의 값
- 사분위수
 - 제1사분위수 (1Q, $p=0.25$) = 25% 백분위수
 - 제2사분위수 (2Q, $p=0.5$) = 50% 백분위수 = median
 - 제3사분위수 (3Q, $p=0.75$) = 75% 백분위수

예제) 분위수 구하기

9 4 8 11 17 16

1. sorting


4 8 9 11 16 17

$$2. f_i = \frac{i - 1}{n - 1}$$

0 0.2 0.4 0.6 0.8 1

3. Q3 (75%) ?

0.6 —————→ 0.8
11 —————→ 16



보간 (interpolation)

$$0.15/0.2 = x/5$$

$$x = 3.75$$

$$Q3 = 14.75$$

Spread



Spread

- **Range:** the distance between the smallest and largest values
- **Variance**

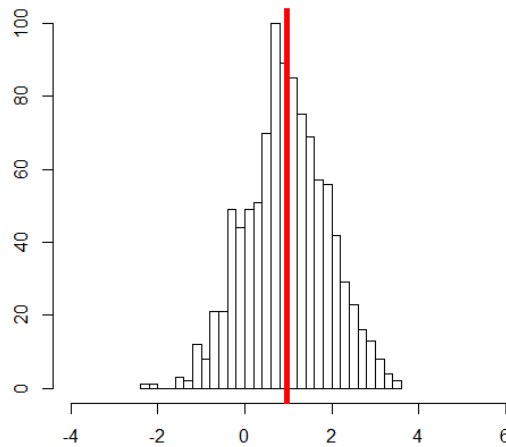
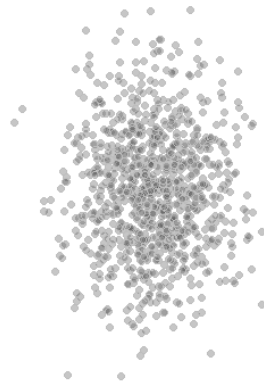
sample variance

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

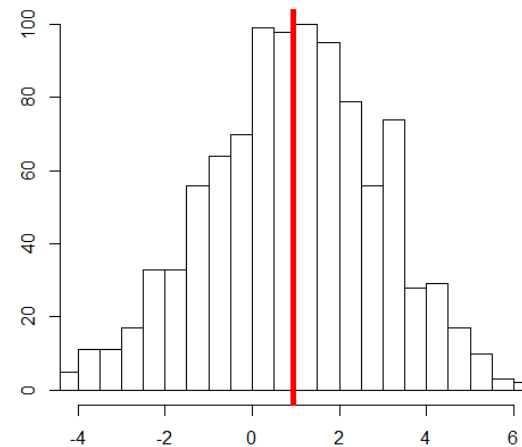
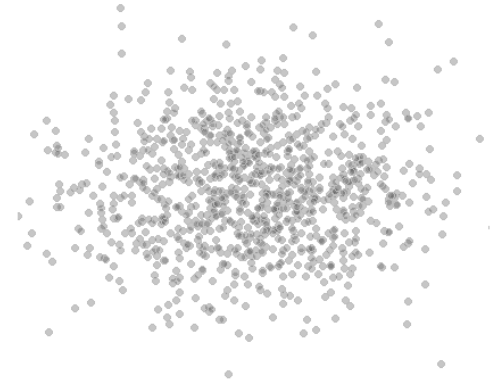
sample standard deviation

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

Spread



sd 1.09



sd 2.05

측정값들이 평균에서 떨어진 정도

Spread

- **Deviation** – How far (Big, small) relative to the center
- **z-score** - How big (small) is the value relative to the others

$$z \text{ score of } x_i = \frac{x_i - \bar{x}}{s}$$

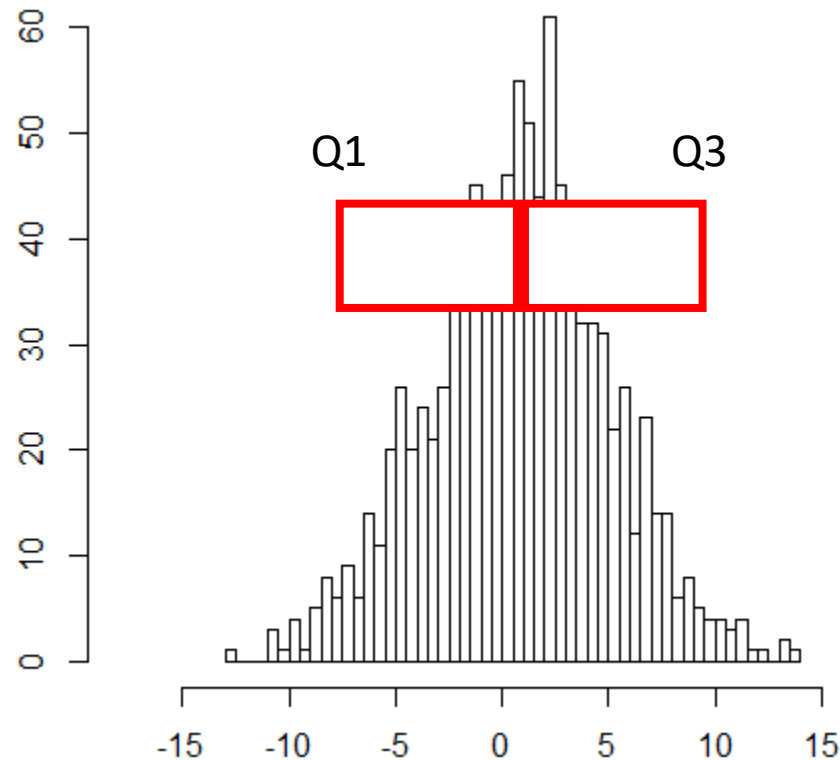
- 해석) z-score= 3: 이 값은 평균에 비해서 3표준편차만큼 크다
- 예) 다음 학생들의 점수에서 z값이 1.28보다 큰 학생이 A이다 라고함. 누가 A를 받을까?

54, 50, 79, 62, 100, 80

다른 스케일 데이터 비교 가능, z-test, normal distribution

Spread

- Range suffers from outliers (one large or small value)
- SD is very sensitive to a single large or small value
- **Interquartile range (IQR)** – middle 50% of the data
- Difference between Q3 and Q1





Shape

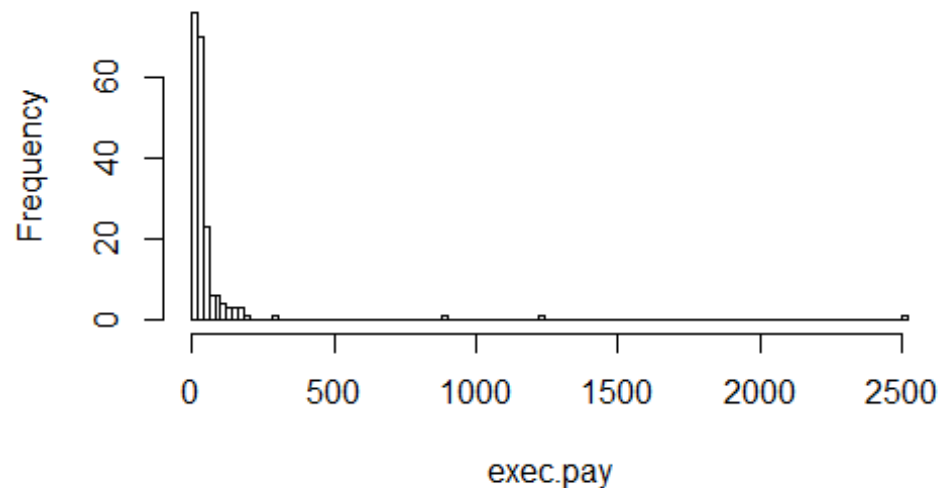
Shape

- Symmetry and skew

$$\text{sample skewness} = \sqrt{n} \frac{\sum (x_i - \bar{x})^3}{(\sum (x_i - \bar{x})^2)^{3/2}} = \frac{1}{n} \sum z_i^3$$

$$z_i = (x_i - \bar{x})/s$$

Histogram of exec.pay



Shape

- Tail – how much data is far from the bulk of the data

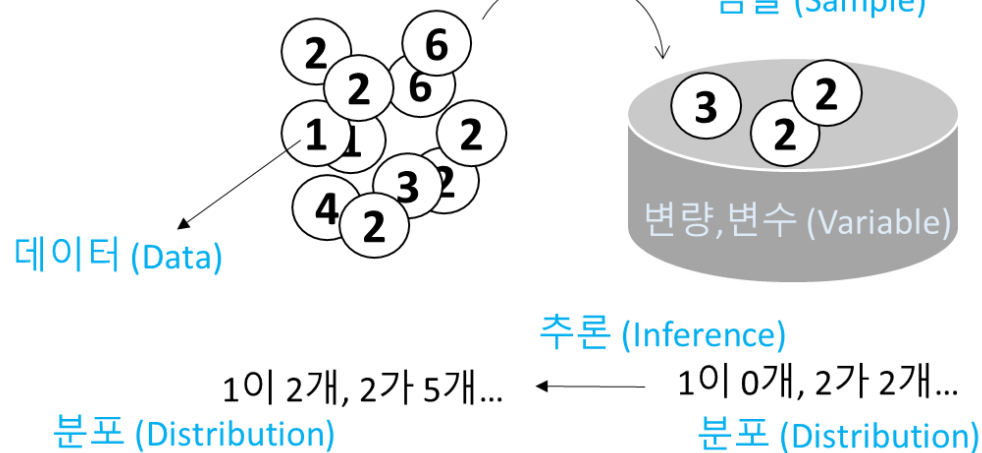
$$\text{sample excess kurtosis} = n \frac{\sum (x_i - \bar{x})^4}{(\sum (x_i - \bar{x})^2)^2} - 3 = \frac{1}{n} \sum z_i^4 - 3$$

$$z_i = (x_i - \bar{x})/s$$



전체 관측 불가
모집단 (Population)

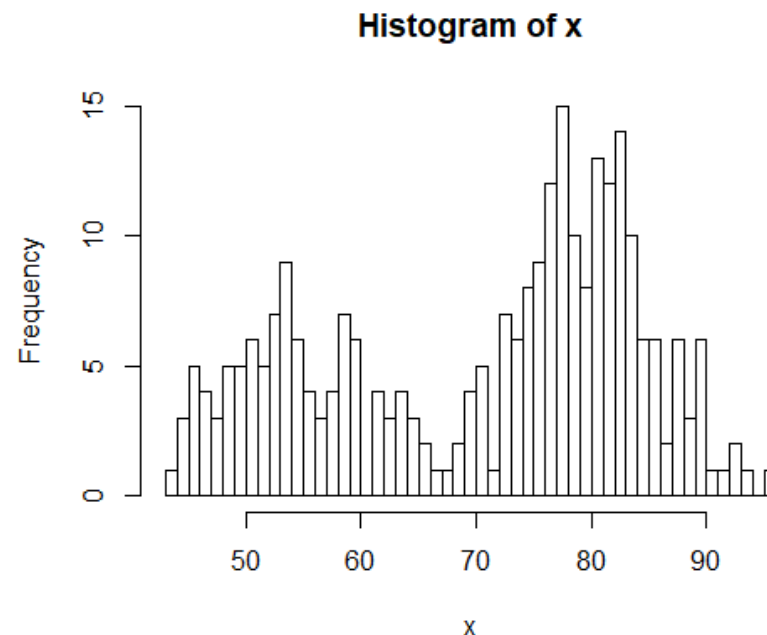
관측 가능
샘플 (Sample)



얼굴 == 데이터
생김새 == 분포
눈코입 == 요약통계량

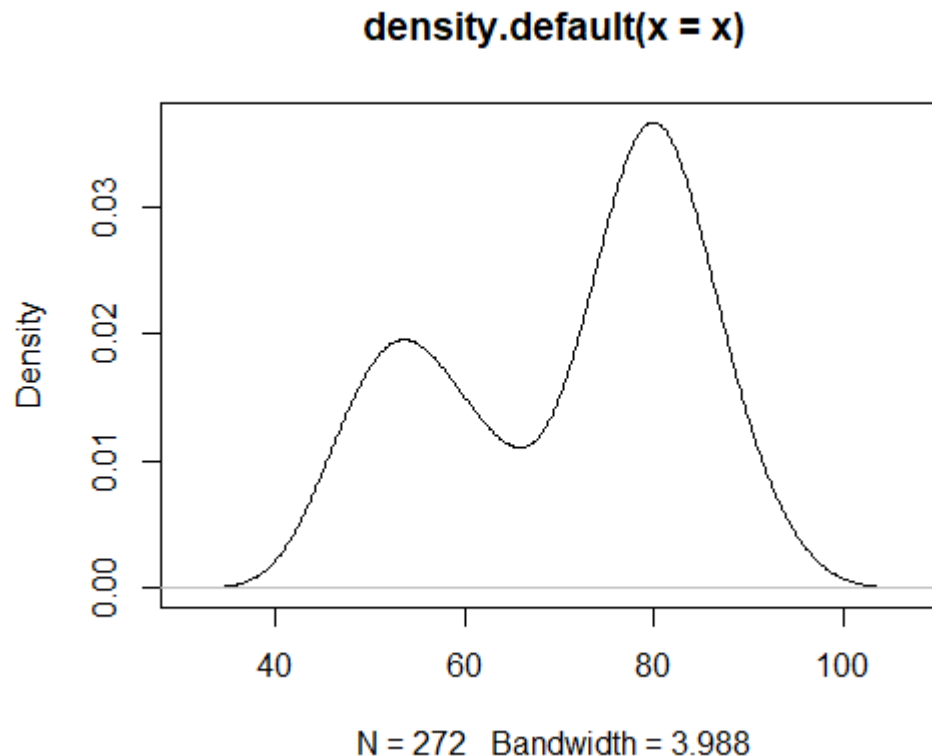
Viewing the shape

- Dot plots – trouble with repeated values, only used for small data sets
- Stem and leaf plot – shows range, median, shape. But only for small data sets. trouble with clustered data. Rounding
- Histogram – Break up an interval, for each subinterval the number of data points are counted



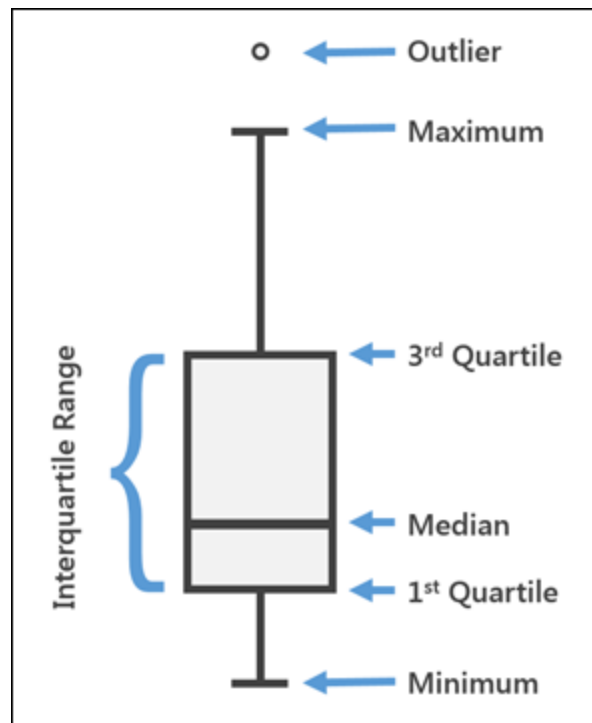
Viewing the shape

- Density plots – We have a sample and a histogram. If we pick a data point from the sample at random what is the chance we pick a value in a given bin?



Viewing the shape

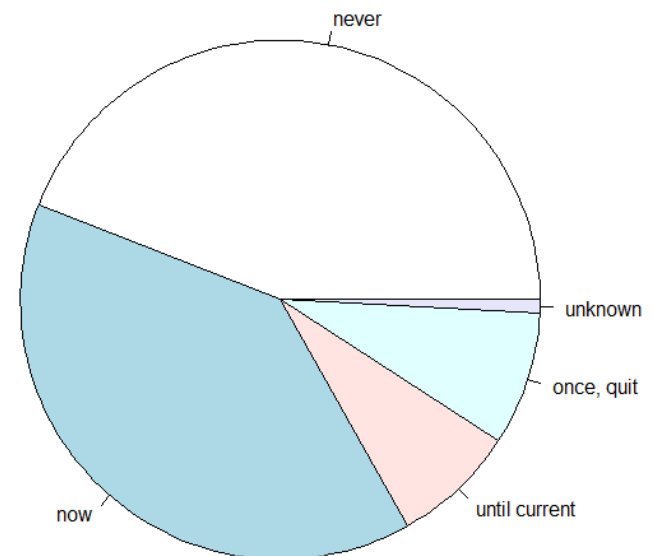
- boxplots – Five-number summary of a univariate data set: min, max, Q1, Q3, and median. These are good summary of even very large data sets. It shows center, spread, shape
- Outliers – $1.5 \times \text{IQR}$



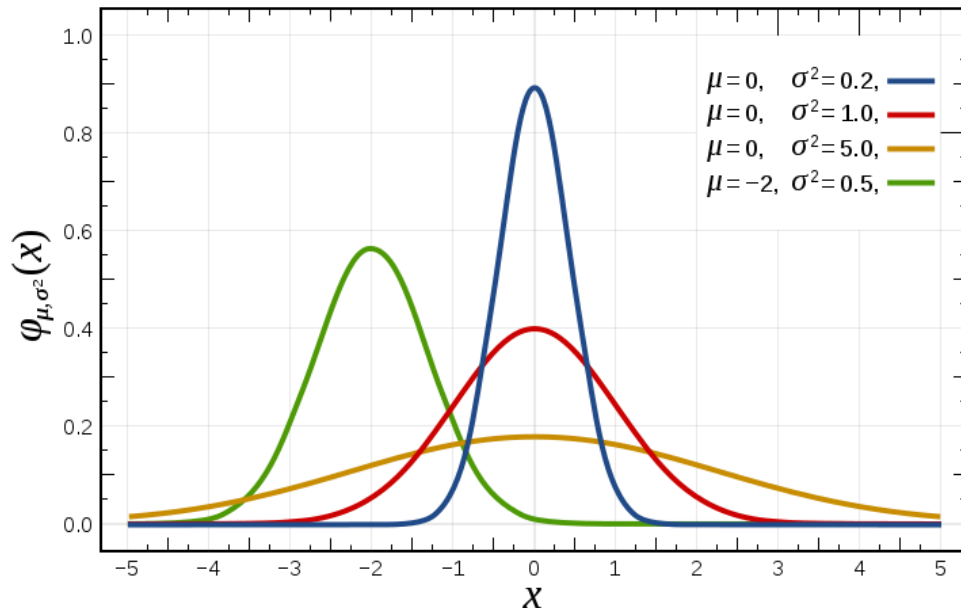
Categorical data

- Tabulating data

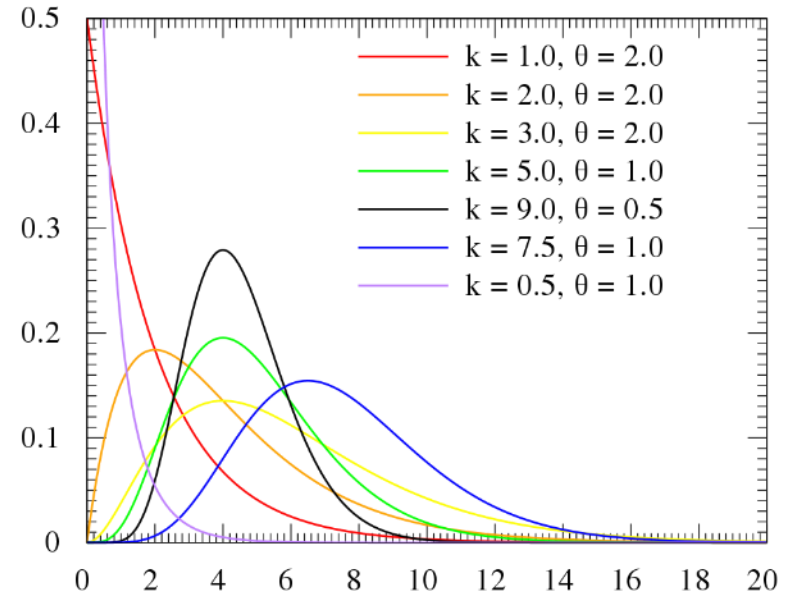
		Patients with bowel cancer (as confirmed on endoscopy)	
		Condition positive	Condition negative
Fecal occult blood screen test outcome	Test outcome positive	True positive (TP) = 20	False positive (FP) = 180
	Test outcome negative	False negative (FN) = 10	True negative (TN) = 1820



착한 데이터



Normal distribution



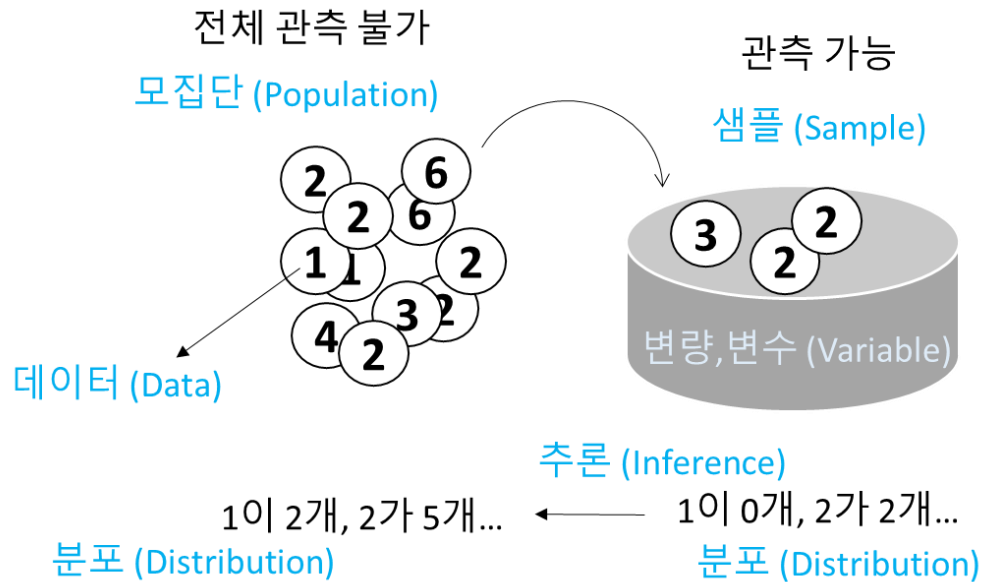
Gamma distribution

(Exponential distribution,
chi-squared distribution)

Summary

통계
데이터, 정보
일변량
요약통계량

- Center: mean, median..
- Spread: variance, range..
- Shape: skewness, ..



Next

Bivariate data

- Two variables at once
- Similarity, Relationship (independence)
- Paired data
- Bivariate categorical data

숙제 #1 (다음시간제출, A4용지 사용, 이름, 학번 명시)

1. 다음 데이터들의 타입을 구분하시오 (명목, 순서, 구간, 비율 중 하나)

직업, 지역, 물가지수, 돈, 성적, 소득, 학년, 혼인 상태, 지지도, 선호도, 몸무게

2. 다음 데이터셋의 mean, median, variance 를 구하시오

11, 20, 9, 95, 34, 7, 14, 39, 12, 29, 21

3. 다음 데이터셋의 histogram을 그리시오 (0부터 150까지 10개 구간)

21, 60, 35, 17, 36, 29, 162, 88, 31, 6, 135, 13, 20, 9, 14, 28, 42, 10, 35, 2, 16

4. 위 예제3의 데이터를 z-score로 변환하고 -1부터 3까지 10개 구간으로 나누어 histogram을 그리시오

5. 위 예제3 데이터의 boxplot을 그리시오 (outlier 무시)