# 데이터로 배우는 통계학

자연과학대학 통계학과
**장원철** 교수

# 자료의 유형과 요약

## 1. 이진 데이터란?

# 수술을 많이 한 병원이 치료를 잘할까?

## 1984년부터 1995년 사이에 브리스틀 병원에서 심장 수술을 한 아이들에게 무슨 일이 생긴 걸까?

- 1995년 1월 브리스틀 왕립병원에서 심장 수술을 받던 16개월 된 아기 조슈아가 수술 도중 사망함
- 1990년 초반 이래 브리스틀 병원에서 수술 생존율이 낮다는 이야기가 지속적으로 나돌고 있었음
- 조슈아의 부모와 아이를 잃은 다른 부모의 항의로 영국의 종합의료위원회 (General Medical Council) 의 조사 결과 1998년 병원장과 2명의 외과 의사가 유죄판결을 받음
- 이후 통계학자들이 1984년부터 1995년 사이 브리스틀 병원과 다른 병원의 생존율을 비교하는 조사를 함

# 수술을 많이 한 병원이 치료를 잘할까?

## 1984년부터 1995년 사이에 브리스틀 병원에서 심장 수술을 한 아이들에게 무슨 일이 생긴 걸까?

- **문제(Problem): 브리스틀 병원의 어린이 심장 수술 사망률은 다른 병원과 비교해서 현저히 낮았는가?**
  - 어린이란? – 16세 미만
  - 심장 수술이란? – 인공심폐기를 이용하는 개복수술
  - 사망의 기준은 무엇인가? – 수술 이후 30일 이내 사망한 경우
- **계획(Plan): 심장 수술 데이터베이스를 이용하여 자료를 수집 후 다른 병원들의 사망률과 비교**

# 수술을 많이 한 병원이 치료를 잘할까?

## 1984년부터 1995년 사이에 브리스틀 병원에서 심장 수술을 한 아이들에게 무슨 일이 생긴 걸까?

- 자료(Data): HES(Hospital Episode Statistics)와 CSR(Cardi-ac Surgical Registry)를 사용함 하지만 두 자료는 매우 상이하다는 것이 밝혀짐. 예를 들면 1991년부터 1995년 사이 개복수술 건수와 사망 건수로 모두 일치하지 않음
- 분석(Analysis): HES와 CSR, 그 외 다른 추가자료를 이용하여 사망위험률 예측
- 결론(Conclusion): 일반적인 사망위험률 예측 기준으로 브리스틀 병원은 그 기간 동안 32명의 사망자가 나왔을 것을 생각됨(실제 사망자는 62명)
- 이 사건을 계기로 영국에서 의료자료의 일반인에 대한 공개와 사용 체계가 확립됨

# 비율로 표시하는 데이터란?

- 특정 사건의 발생 유무와 같이 두 가지 값으로 이루어진 데이터를 이진 데이터(binary data)라고 한다.
- 이러한 이진 데이터들의 평균은 비율로 표시된다.
- 예를 들면 심장 수술 후 사망 여부를 나타내는 데이터는 이진 데이터(1=생존, 0=사망)로 볼 수 있다.
- 심장 수술 생존율은 전체 수술환자 중 생존자의 비율로 위의 이진 데이터의 평균으로 생각할 수 있다.

# 비율로 표시하는 데이터란?

| 병원 | 수술을 받은 어린이 수 | 수술 후 30일 이상 생존한 어린이 수 | 수술 후 30일 이내 사망한 어린이 수 | 생존율 | 사망률 |
|---|---|---|---|---|---|
| London, Harley Street | 418 | 413 | 5 | 96.8 | 1.2 |
| Leicester | 607 | 593 | 14 | 97.7 | 2.3 |
| Newcastle | 668 | 653 | 15 | 97.8 | 2.2 |
| Glasgow | 760 | 733 | 27 | 96.3 | 3.7 |
| Southampton | 829 | 815 | 14 | 98.3 | 1.7 |
| Bristol | 835 | 821 | 14 | 98.3 | 1.7 |
| ... | ... | ... | ... | ... | ... |
| London, Great Ormond Steet | 1,892 | 1,873 | 19 | 99.0 | 1.0 |
| 총합 | 12,933 | 12,670 | 263 | 98.0 | 2.0 |

[ 2012년부터 2015년 영국과 아일랜드 병원에서 어린이 심장수술 결과 ]
(The Art of Statistics, p23)

# 긍정/부정 메시지 프레이밍

- “결론”의 주요 부분 중 하나는 분석 결과를 효과적으로 전달하는 것이다.
- 앞 페이지 표와 같은 데이터의 경우 미국의 경우 수술 사망률을, 영국은 수술 생존율을 제시한다.
- 이와 같이 같은 데이터를 부정적, 혹은 긍정적 의미를 나타내는 결과로 전달하는 것을 부정/긍정 메시지 프레이밍이라고 한다. 5%의 사망률과 95%의 생존율은 똑같은 의미지만 95%의 생존율이 훨씬 긍정적으로 들린다.

# 정보(올바르게) 전달하기

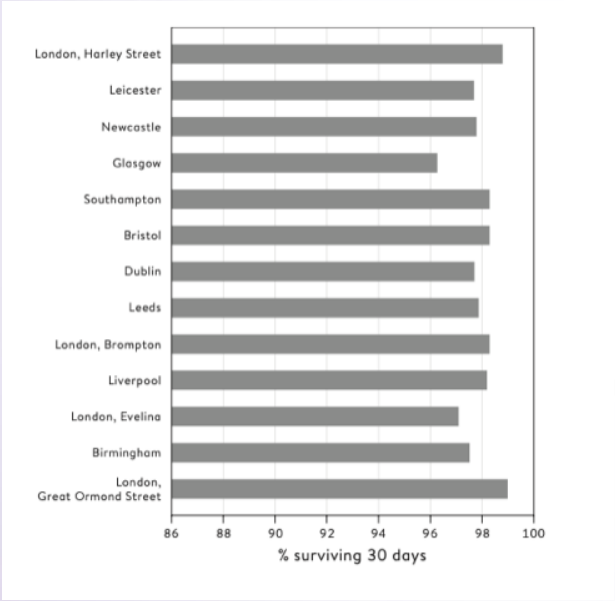
- 2011년 영국 런던의 지하철역에 “런던에 사는 청소년의 99%가 심각한 폭력을 저지르지 않는다”는 광고가 등장한 적이 있었다.
- 부정 메시지 프레이밍으로 바꾼다면 1%의 청소년이 아주 심각한 폭력을 저지른다고 말할 수 있다.
- 런던의 인구가 약 9백만 명이라는 걸 고려하고 이 중 15~25세 인구가 약 100만 명이라는 걸 고려한다면 아주 폭력적인 청소년이 만 명정도 된다는 결론에 도달할 수 있다!
- 정보를 정확히 전달하기 위해서는 결론을 긍정/부정 메시지 프레이밍 모두를 사용해서 표현하고  절대적인 숫자와 상대적인 요약을 모두 제공하는 것이 중요하다.

# 이진 데이터의 비교

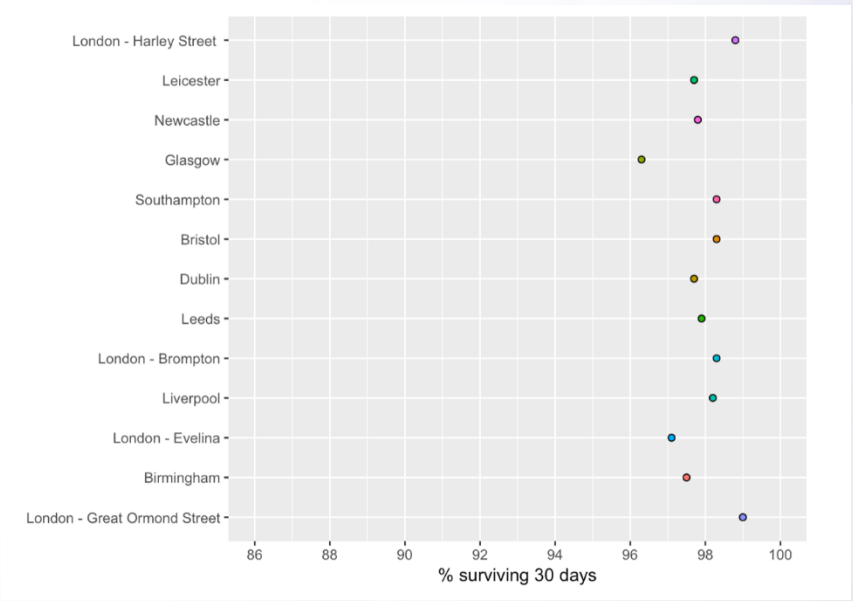
- 출처 #1의 자료를 병원들 간의 생존율 비교를 위해서  막대그래프로 나타낸다고 하자.
- 이 경우 가로축의 시작이 어디인지가 중요하다. 0%로 시작할 경우 병원들 간 생존율은 거의 차이가 없어 보일 것이지만 95%에서 시작한다면 병원들 간의 생존율 차이가 과대포장될 가능성이 높다.
- “How charts lie”라는 책의 저자 알베르토 카이로는 이런 경우 “논리적이고 의미 있는 기준선”을 강조한다.

# 막대그래프와 점 그림을 이용한 비율의 비교

## 각 병원의 30일 생존율 비교




[ 13개 병원의 30일 생존율을 나타낸 막대그래프 ]
(The Art of Statistics. P26)




[ 13개 병원의 30일 생존율을 나타낸 점 그림(Dot plot) ]

# 오늘 강의 요점

- 이진 데이터
- 긍정/부정 메시지 프레이밍
- 막대그래프 vs 점 그림

## 출처

#1~2 D. Spiegelhater, (2019), The Art of Statistics, Penguin Random House

#3 https://bit.ly/3lgfONi

# 자료의 유형과 요약

## 2. 범주형 자료의 소개

# 범주형 자료란?

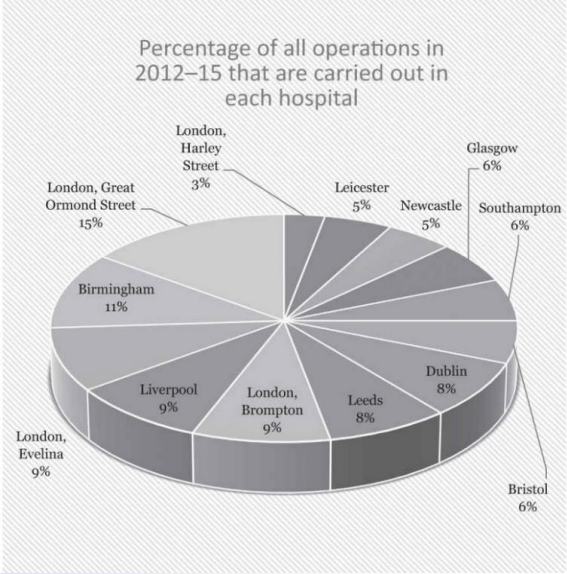
- 변수(variable)란 주어진 상황에 따라 다른 값을 가지는 측정치라고 정의할 수 있다.
- 예를 들면 동전을 던져서 앞면이 나온 상황이라면 1의 값을 가진다고 하고 뒷면이 나오는 경우 0의 값을 가지는 변수를 생각할 수 있다.

# 범주형 자료란?

- **범주형 변수(categorical variable)는 2개 이상의 범주(category)를 값으로 가지는 변수를 의미하며 다음과 같은 경우를 범주형 변수로 정의할 수 있다.**
  - 순서가 없는 범주: 국적, 성별
  - 순서가 있는 범주: 군인계급(이병<일병 <상병 <병장)
  - 일련의 그룹으로 묶인 숫자들: BMI 기준 비만 측도

    (정상의 경우 18.5 < BMI < 22.9)

# 파이차트는 절대 사용하지 마세요!




[ 2012년부터 2015년까지 심장수술을 받은 어린이 환자의 병원별 비율 ]
(The Art of Statistics, P29)

- 이 그림은 2012년부터 2015년까지 심장 수술을 받은 어린이 환자의 병원별 비율을 나타낸 3D 파이차트이다.
- 사람의 눈은 면적의 차이를 식별하기 쉽지 않기 때문에 파이차트의 경우 실제 표시된 숫자를 보고 그 영역의 비율을 인지하는 것이며 어느 영역이 더 크게 보이는지는 시각적인 효과만으로 분별하는 것은 거의 불가능하다!

기타 출처 #1

# 파이차트 대신 막대그래프를 사용하라


[ 심장수술을 받은 어린이의 병원별 비율을 표시한 막대그래프 ]
(The Art of Statistics, P30)

- 이 그림은 똑같은 자료를 막대그래프로 표시한 것이다. 각 병원별 비율의 차이가 뚜렷하게 관측된다.

기타 출처 #2

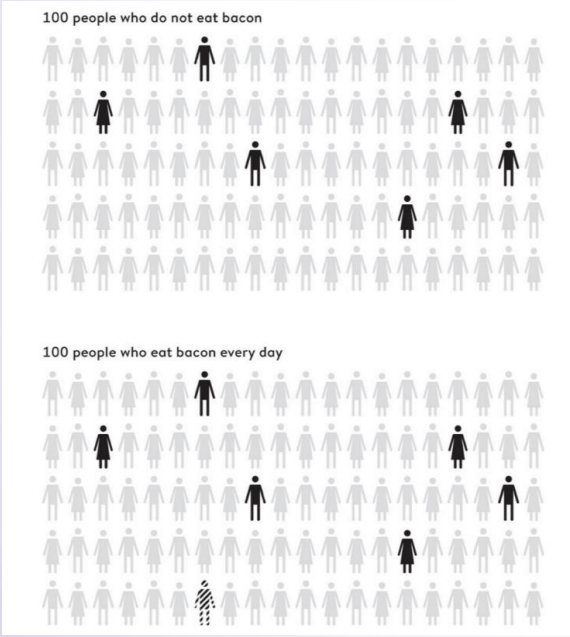
# 베이컨 샌드위치는 대장암 발병률을 얼마나 높이는가?

## 상대위험도

- 2015년 11월 세계보건기구(WHO) 국제 암연구소(IARC)에서 가공육이 담배와 석면이 속하는 1군 발암물질에 속한다고 발표했다.
- 연구소의 보고서에서는 매일 50g의 가공육을 먹으면 대장암 발병률이 18% 높아질 수 있다고 밝혔다.
- 여기서 18%가 의미하는 것은 상대위험도(relative risk)를 의미한다.
- 하루에 베이컨 2개가 들어가는 샌드위치를 먹는 집단은 그렇게 하지 않는 집단에 비해 대장암에 걸릴 위험이 18% 높다는 것을 말한다.

# 베이컨 샌드위치는 대장암 발병률을 얼마나 높이는가?

## 절대위험도



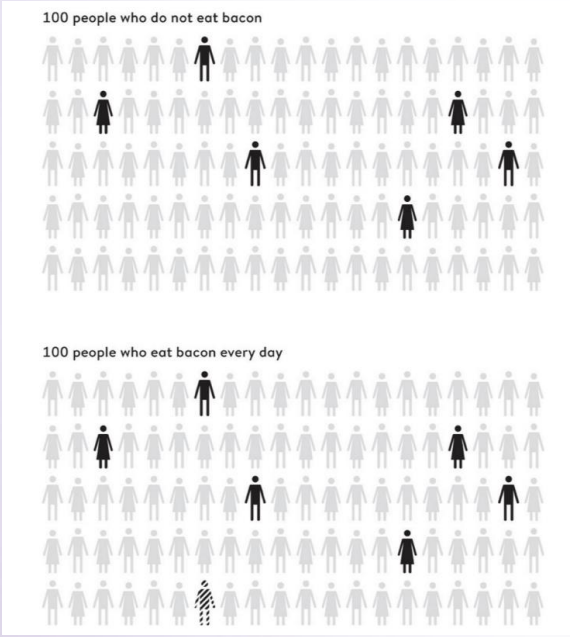

[ 베이컨 샌드위치를 매일 먹는 경우 절대위험도를 설명한 그림 ]
(The Art of Statistics, P33)

- 상대위험도와 대비되는 개념으로 절대 위험도(absolute risk)를 들 수 있다.
- 절대 위험도는 각 집단에서 위험에 처해지는 비율을 의미한다.

기타 출처 #3

# 베이컨 샌드위치는 대장암 발병률을 얼마나 높이는가?

## 절대위험도




[ 베이컨 샌드위치를 매일 먹는 경우 절대위험도를 설명한 그림 ]
(The Art of Statistics, P33)

- 영국에서 대장암의 유병율은 6%이다. 즉 백 명 중 6명 정도 대장암에 걸린다.
- 위의 백 명이 평생동안 매일 베이컨 샌드위치를 먹는다면 대장암에 걸리는 사람은 한 명이 더 추가되어서 7명이 대장암에 걸리게 되는 것이다.

기타 출처 #3

# 베이컨 샌드위치는 대장암 발병률을 얼마나 높이는가?

## 절대위험도와 상대위험도

- 두 집단을 비교하는 경우 사용할 수 있는 측도로 절대위험도의 차이를 고려할 수 있다.
- 상대위험도의 경우 개개인에 대한 비교를 하는 경우 유용한 측도로 매일 베이컨을 먹는 사람과 그렇지 않은 사람의 (대장암의)절대위험도의 비율로 두 사람을 비교한다.

# 베이컨 샌드위치는 대장암 발병률을 얼마나 높이는가?

## 절대위험도와 상대위험도

- 상대위험도 = (위험요인이 있는 집단의 절대위험도)/(control group의 절대위험도)로 정의된다.
- 절대위험도의 차이는 (위험요인이 있는 집단의 절대위험도)-(control group의 절대위험도)로 정의된다.
- 상대위험도의 값이 높다고 하더라도 절대위험도 자체가 작을 경우 실제 위험 자체는 크게 문제가 되지 않을 수 있다.

# 베이컨 샌드위치는 대장암 발병률을 얼마나 높이는가?

## 기대빈도, 오즈와 오즈비

- 기대도수는 주어진 집단에서 특정 사건(예를 들면 대장암에 걸릴 경우)에 일어나는 개수의 예측값을 의미한다.
- 오즈는 도박에서 많이 사용된다. 예를 들면 월드컵에서 한국의 우승 확률이 1/32라고 하면 오즈는 (1/32)/(1-1/32)=1/31로 정의된다. 쉽게 얘기하자면 32개의 경우의 수 중 31개는 한국이 우승하지 못하는 경우이고 1개가 한국이 우승하는 경우라고 생각해서 (우승의 경우의 수)/(우승하지 못하는 경우의 수)로 정의할 수 있다.

# 베이컨 샌드위치는 대장암 발병률을 얼마나 높이는가?

## 기대빈도, 오즈와 오즈비

- 오즈비는 (위험요인이 있는 집단의 오즈)/(control grop의 오즈)로 정의된다.
- 만약 절대위험도가 굉장히 작은 경우 오즈비와 상대위험도의 값은 비슷해진다.

# 베이컨 샌드위치는 대장암 발병률을 얼마나 높이는가?

- 여기서 매일 베이컨을 먹는 사람들의 유병률은 정확히 7.08%이며 상대위험도는 7.08/6 =1.18로 계산된다.

| 위험측도 | 매일 베이컨을 먹지 않는 사람 | 메일 베이컨을 먹는 사람 |
|---|---|---|
| 유병률 (절대위험도) | 6% | 7% |
| 기대도수 | 6/100 | 7/100 |
| | 1/16 | 1/14 |
| 오즈 | 6/94 | 7/93 |

| 비교측도 | |
|---|---|
| 절대위험도 차이 | 1% |
| 상대위험도 | 1.18 |
| 대장암 환자 1명 추가 시 필요한 사람 수 | 100 |
| 오즈비 | (7/93)/(6/94) =1.18 |

[ 매일 베이컨 샌드위치를 먹는 사람과 그렇지 않은 사람들의 대장암 유병률 비교방법 ]
(The Art of Statistics, P35)

기타 출처 #4

# 스타틴은 근육통을 유발하는가?

- 스타틴은 콜레스테롤을 낮추는 약으로 심장마비와 뇌졸중 예방에 도움이 된다고 알려진 약이다. 하지만 일부 의사들은 스타틴의 부작용에 대한 우려를 표시해왔다.

# 스타틴은 근육통을 유발하는가?

- 2013년에 행해진 한 연구에 따르면 스타틴을 복용한 사람 중 87%가 근육통을 호소했고 control group에서 근육통을 호소한 비율은 85%였다.
- 이 경우 절대위험도의 차이는 2%, 상대위험도는 0.87/0.85=1.02이지만 오즈비의 경우

  (0.87/0.13)/(0.85/0.15)=1.18이다.
- 오즈비는 앞의 예제인 베이컨 샌드위치와 같지만 각 그룹의 절대위험도는 전혀 다르다.
- 모 언론매체는 이 연구결과에서 오즈비를 상대위험도로 착각하여 스타틴이 근육통 발병률을 20%까지 증가할 수 있다고 보도했다!

# 오늘의 강의 요점

- **범주형 자료**
- **두 비율의 비교**
  - 절대위험도 vs 오즈
  - 상대위험도 vs 절대위험도 차이 vs 오즈비

## 출처

#1~4 D. Spiegelhater, (2019), The Art of Statistics, Penguin Random House

# 자료의 유형과 요약

## 3. 연속형 자료와 요약

# 집단지성이란?

## 황소 몸무게 맞추기

- 1907년 찰스 다윈의 사촌이자 우생학의 창시자인 프랜시스 골턴이 다음과 같은 일화를 통해 집단지성(wisdom of crowds)를 소개하는 논문을 Nature에 출간하였다.
- 그해 항구도시 플리머스에 가축박람회가 열렸는데 거기서 나온 황소의 무게를 맞추는 게임을 진행하였다.

# 집단지성이란?

## 황소 몸무게 맞추기

- 총 787명의 참가자들이 6펜스를 내고 황소의 몸무게를 적어서 제출하였다. 골턴의 참가자들이 써낸 값 중 중앙값인 547kg을 황소의 몸무게로 추정하였는데 실제 몸무게는 543kg이었다.
- 골턴은 이러한 선택을 Vox Populi(voice of people)이라는 제목을 붙였는데 이러한 현상은 오늘날 집단지성으로 알려져 있다.

# 집단지성의 힘

## 젤리개수 맞추기

[ 젤리의 RO수는? ]
(The Art of Statistics, p41)

- 집단지성의 힘을 알아보기 위해 The Art of Statistics의 저자 데이비드 스피겔헬터는 다음과 같은 실험을 유튜브에서 실시하였다.
- 먼저 그림에 보이는 젤리상자를 유튜브에서 보여준 후에 상자 안의 젤리의 개수를 맞춰보라는 퀴즈를 제시하였다.
- 총 915명이 이 퀴즈에 참가하였다.

기타 출처 #1

# 자료의 유형

- 젤리의 개수는 1, 2, 3 등과 같은 자연수로 표시될 수 있다.
- 우리가 이때까지 배운 자료의 유형은(분석목적의 관점에서) 다음과 같이 정리할 수 있다.

→ 범주형
  - 순서가 있는 범주형
  - 순서가 없는 범주형

→ 수치형
  - 이산형: 값이 정수인 경우(예: 교통사고 횟수)
  - 연속형: 값이 실수인 경우(예: 몸무게, 키)

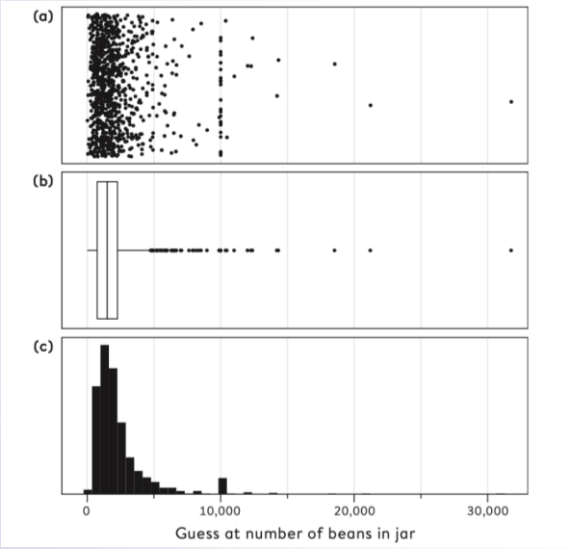
# 자료의 요약

## 통계량과 분포

- 데이터가 굉장히 많은 경우 모든 데이터 값을 일일이 살펴보는 것은 불가능하다.
- 이럴 경우 자료의 특징을 나타내는 몇가지 요약값을 대신 제시할 수 있는데 이러한 요약치를 통계량(statistic)이라고 한다.
- 또한 데이터가 가지는 모든 값의 형태를 통칭해서 분포(distr-ibution)라고 한다.

# 연속형 자료의 시각화

## strip-chart, 상자그림(boxplot), 히스토그램



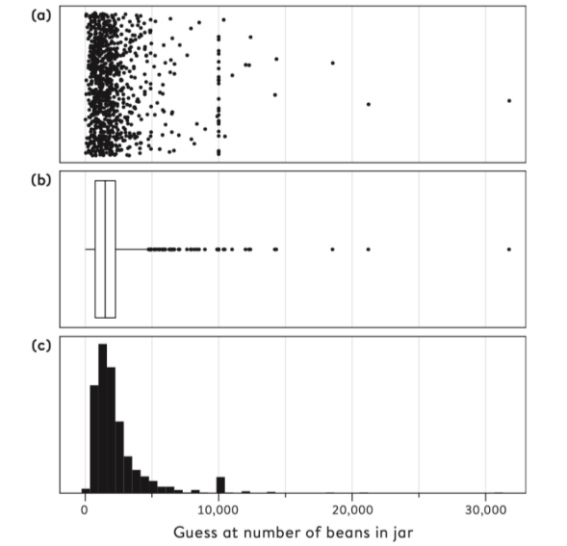

[ 병 속 젤리 개수에 대한 915명의 추정값 요약 ]
(The Art of Statistics, P42)

- 퀴즈 참가자 전원의 젤리 숫자에 관한 추정값을 보는 것은 전반적인 경향을 파악하기 어렵기 때문에 데이터 시각화를 통해서 전체 데이터를 살펴보자.

기타 출처 #2

# 연속형 자료의 시각화

## strip-chart, 상자그림(boxplot), 히스토그램



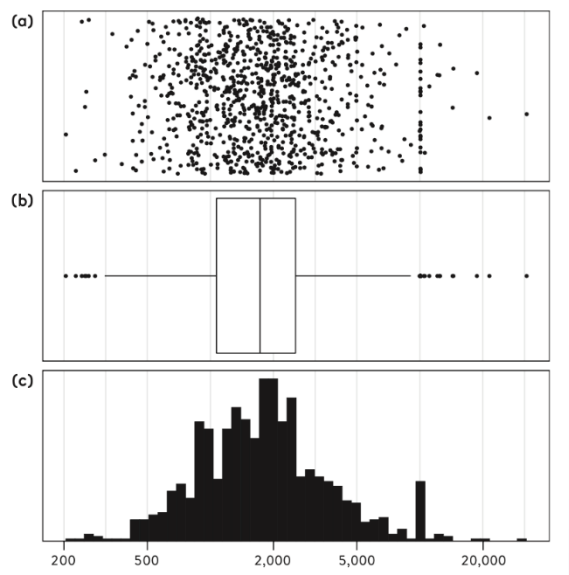

[ 병 속 젤리 개수에 대한 915명의 추정값 요약 ]
(The Art of Statistics, P42)

- 왼쪽 그림은 수치형 자료들의 대표적인 3가지 시각화 방법을 이용하여 젤리숫자 추정값을 제시하였다.
  - strip-chart( 또 는 dot-diagr -am)
  - 상자그림(box plot)
  - 히스토그램

기타 출처 #2

# 연속형 자료의 시각화

## 데이터의 치우침과 로그변환



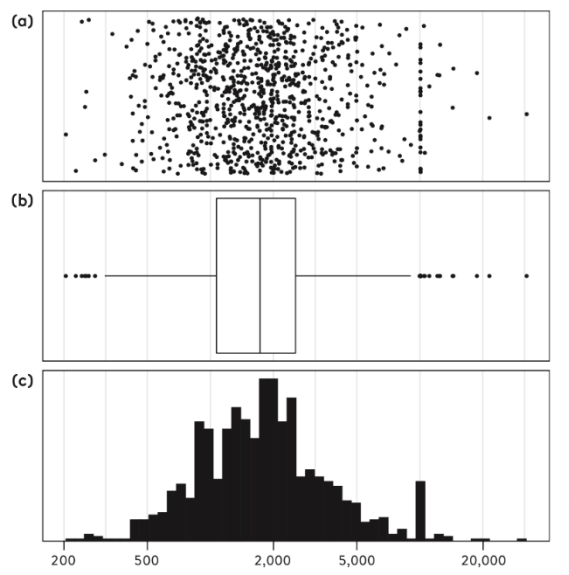

[ 젤리 개수 추정치(로그스케일) ]
(The Art of Statistics, P45)

- 원래 그래프에서 소수의 극단값을 왼쪽에서 관측할 수 있었다.
- 실제로 9,000이상의 값을 임의로 제거한 후 그림을 그렸었다. 만약 이런값을 포함한다면 대부분의 데이터들이 뭉쳐보여서 데이터 시각화가 효율적으로 이루어지지 않을 수 있다.

기타 출처 #3

# 연속형 자료의 시각화

## 데이터의 치우침과 로그변환




[ 젤리 개수 추정치(로그스케일) ]
(The Art of Statistics, P45)

- 이러한 극단값을 포함하면서 영향을 최소화하기 위해 데이터의 값을 로그변환 변환하는 것을 고려할 수 있다.
- 왼쪽 그림은 로그변환후의 시각화 결과를 제시하고 있으며 분포가 대칭적이고 극단값이 관측되지 않음을 알 수 있다.

기타 출처 #3

# 연속형 자료의 요약

## 자료의 위치 대푯값

- **연속형 자료의 중앙을 나타내는 대표적인 통계량(요약값)은 다음 3가지를 들 수 있다.**
  - 평균: 데이터의 총합을 데이터의 개수로 나눈 값
  - 중앙값: 자료를 순서대로 나열했을 때 가운데 값 만약 관측치의 개수가 짝수라면 가운데에 가장 가까운 두개의 값의 평균을 사용
  - 최빈수: 가장 많이 관측되는 값

# 연속형 자료의 요약

## 자료의 위치 대푯값

- 소득과 같이 분포의 모양이 비대칭일 경우 평균은 좋은 대푯값이 아니다. 예를 들어 국회의원의 평균소득의 전체 국회의원의 소득을 대표한다고 얘기하기 힘들다.
- 젤리 개수 맞추기에서 중앙값은 1,775였으며 실제 젤리의 개수는 1,616이었다.

# 연속형 자료의 요약

## 자료의 퍼짐 대푯값

- **데이터의 중앙을 나타내는 대푯값과 더불어 데이터가 얼마나 퍼져 있는지 여부를 제시하는 대푯값은 자료의 요약에 필수적인 요소이다.**
- **대표적인 퍼짐을 나타내는 통계량은 다음과 같다.**
  - 범위 = 최댓값 - 최솟값
  - IQR = $Q_3 - Q_1$, 여기서 $Q_1$과 $Q_3$는 1사분위수(하위 50%데이터의 중앙값)과 3사분위수(상위 50%데이터의 중앙값)을 나타낸다.
  - 표준편차 = 분산의 제곱근, 여기서 분산은 각 데이터가 평균에서 떨어진 거리의 제곱의 평균이다.

# 연속형 자료의 시각화

## 자료의 퍼짐 대푯값

| 통계량 | 젤리 개수의 추정치 |
|---|---|
| 평균 | 2,408 |
| 중앙값 | 1,775 |
| 최빈값 | 10,000 |
| 범위 | 31,337-219=31,118 |
| IQR | 2,599-1,109=1,490 |
| 표준편차 | 2,422 |

[ 젤리 추측값 915개를 요약한 다양한 통계량 ]
(The Art of Statistics, P50)

- 왼쪽 표는 915개의 젤리 개수 추정치에 대한 다양한 통계량을 보여준다.
- 앞에서 본 것처럼 이 데이터에는 몇 개의 극단값이 존재한다. 평균과 표준편차와 같은 통계량들은 이러한 극단값에 영향을 많이 받는다.

기타 출처 #4

# 연속형 자료의 시각화

## 자료의 퍼짐 대푯값

| 통계량 | 젤리 개수의 추정치 |
|---|---|
| 평균 | 2,408 |
| 중앙값 | 1,775 |
| 최빈값 | 10,000 |
| 범위 | 31,337-219=31,118 |
| IQR | 2,599-1,109=1,490 |
| 표준편차 | 2,422 |

[ 젤리 추측값 915개를 요약한 다양한 통계량 ]
(The Art of Statistics, P50)

- 만약 이 데이터에서 최댓값 31,337을 제외한다면 표준편차는 2,422에서 1,398로 줄어든다.

기타 출처 #4

# 영국인들의 평생 동안 얼마나 많은 사람들과 성관계를 가지는가?

## 영국의 성생활에 관한 설문조사

- 1980년대 에이즈가 심각한 공중보건의 위기문제로 대두되었을 때 사람들은 성생활에 대한 데이터가 전무하다는 것을 깨달았다.
- 사람들은 얼마마다 sex파트너를 바꿀까?
- 얼마나 많은 사람들이 동시에 여러 명과 성관계를 맺을까?
- 사람들은 어떤 종류의 성행위를 할까?

기타 출처 #5

# 영국인들의 평생 동안 얼마나 많은 사람들과 성관계를 가지는가?

## 영국의 성생활에 관한 설문조사

- 이런 정보는 성병확산을 예방하고 공중보건 서비스를 설계하는데 꼭 필요한 정보이다.
- 이에 따라 1990년부터10년마다 영국에서 성생활에 관한 대규모 설문조사 National Sexual Attitudes and Lifestyle Survey가 실시되고 있다.

기타 출처 #5

# 영국인들의 평생 동안 얼마나 많은 사람들과 성관계를 가지는가?

## 자료와 요약을 통한 두 집단의 비교

| 성관계 상대수 | 35-44세 남성 | 35-44세 여성 |
|---|---|---|
| 평균 | 14.3 | 8.5 |
| 중앙값 | 8 | 5 |
| 최빈값 | 1 | 1 |
| 범위 | 500-0=500 | 550-0=550 |
| IQR | 18-4=14 | 10-3=7 |
| 표준편차 | 24.2 | 19.7 |

[ 영국 중년 성인의 성별 sex 파트너 숫자 ]
(The Art of Statistics, P53)

- 왼쪽 표는 700만 파운드의 비용을 들여 2010년에 이루어진 3차 설문조사의 자료 요약을 보여준다.
- 여기서 표준편차가 매우 크고 자료가 양수로만 이루어진다는 걸 감안하며 sex 파트너 숫자의 분포는 오른쪽으로 꼬리가 긴 분포임을 알 수 있다.

기타 출처 #6

# 영국인들의 평생 동안 얼마나 많은 사람들과 성관계를 가지는가?

## 자료와 요약을 통한 두 집단의 비교

| 성관계 상대수 | 35-44세 남성 | 35-44세 여성 |
|---|---|---|
| 평균 | 14.3 | 8.5 |
| 중앙값 | 8 | 5 |
| 최빈값 | 1 | 1 |
| 범위 | 500-0=500 | 550-0=550 |
| IQR | 18-4=14 | 10-3=7 |
| 표준편차 | 24.2 | 19.7 |

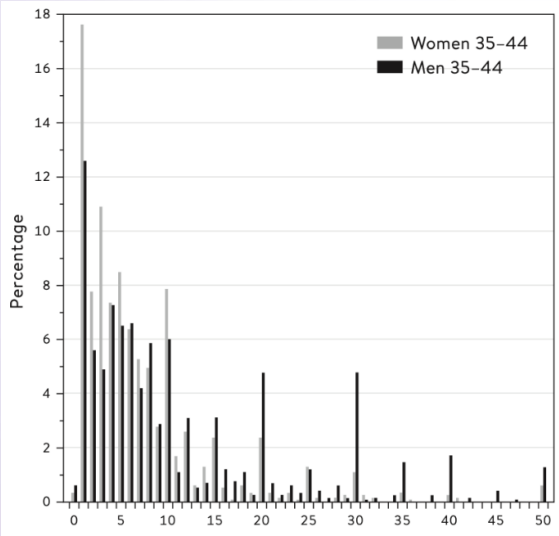
[ 영국 중년 성인의 성별 sex 파트너 숫자 ]
(The Art of Statistics, P53)

- 또 하나의 특징은 여성에 비해서 남성의 평균이 월등히 높다는 점이다. 이러한 현상은 중앙값을 기준으로도 볼 수 있다.
- 하지만 비슷한 연령대의 남녀로 구성된 모집단에서 이성 성관계 상대 수의 평균은 남녀 모두 (거의) 동일해야 한다.

기타 출처 #6

# 영국인들의 평생 동안 얼마나 많은 사람들과 성관계를 가지는가?

## 자료와 요약을 통한 두 집단의 비교



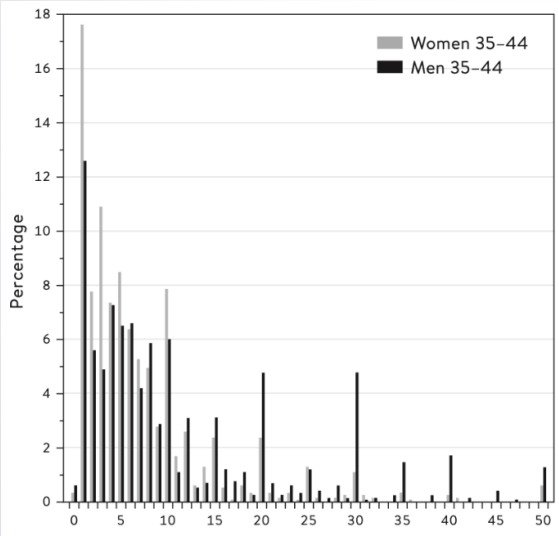

[ 영국 중년 성인의 성별 sex 파트너 숫자 ]
(The Art of Statistics, P55)

- 왼쪽 그림은 실제 데이터를 보여주고 있다. 예상했던대로 오른쪽으로 긴 꼬리를 가지는 분포임을 알 수 있다.
- 실제 데이터에서 최댓값은 남녀 모두 500명 이상이라서 데이터를 잘 볼 수 있도록 50명까지만 데이터를 제시하였다.

기타 출처 #7

# 영국인들의 평생 동안 얼마나 많은 사람들과 성관계를 가지는가?

## 자료와 요약을 통한 두 집단의 비교




[ 영국 중년 성인의 성별 sex 파트너 숫자 ]
(The Art of Statistics, P55)

- 또 하나의 특징은 10명 이상의 경우 남녀 모두 숫자를 5(또는 10)의 배수로 대답하는 경향이 있다는 점이다.

기타 출처 #7

# 오늘의 강의 요점

- **자료의 유형**
  - 범주형
  - 수치형
- **수치형 자료의 시각화**
- **수치형 자료의 요약**
  - 자료의 중앙 대푯값
  - 자료의 퍼짐 대푯값

## 출처

#1~7 D. Spiegelhater, (2019), The Art of Statistics, Penguin Random House

# 자료의 유형과 요약

## Lab 2: 자료의 요약과 시각화

# 평균과 분산

- $n$개의 데이터 $x_1, x_2, \cdots, x_n$에 대해서 평균과 분산은 다음과 같이 계산할 수 있다.

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}, s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}$$

- 은행 융자 50건에 대한 이자율 자료를 가지고 있다고 하자. 평균은 다음과 같이 구할 수 있다.

$$\overline{x} = \frac{10.90\% + 9.92\% + 26.30\% + \cdots + 6.08\%}{50} = 11.57\%$$

# 평균과 분산

- 이자율에 대한 분산은 다음과 같이 계산할 수 있다.

$$s^2 = \frac{(10.90 - 11.57)^2 + (9.92 - 11.57)^2 + \cdots + (6.08 - 11.57)^2}{49} = 25.52$$

- 표준편차는 분산의 제곱근으로 위의 예제의 경우

$s = \sqrt{25.52} = 5.05$이다.

# 분산공식

- 분산공식에서 분모는 $n$이 아니라 $n-1$일 이유는 무엇인가?
- 제곱대신 절대값을 사용하면 괜찮을까?

# 중앙값과 $Q_1$, $Q_3$

- 중앙값은 자료를 크기 순으로 나열했을 경우 중앙에 해당하는 값이다.
- 데이터가 짝수일 경우 중앙에 가장 가까운 두 값의 평균으로 중앙값을 사용한다.
- 이자율 데이터의 경우 데이터의 개수가 50이기 때문에 중앙에(25.5)에 가장 가까운 값인 크기가 작은 순서로 25번째와 26번째인 데이터의 값의 평균을 사용한다. 이 경우에 두 값이 동일해서 중앙값은(9.93%+9.93%)/2=9.93%이다.

# 중앙값과 $Q_1$, $Q_3$

- $Q_1$과 $Q_3$는 데이터를 반으로 나눈 후 하위와 상위 50% 자료의 중앙값이 해당하면 각각 1사분위수(first quartile)와 3사분위수(third quartile)라는 명칭을 사용한다.
- 사분위수와 비슷한 개념으로 백분위수(percentile)를 들 수 있다. 예를 들면 1사분위수는 25 백분위수에 해당한다. 즉 25%의 자료가 25백분위수보다 작다.

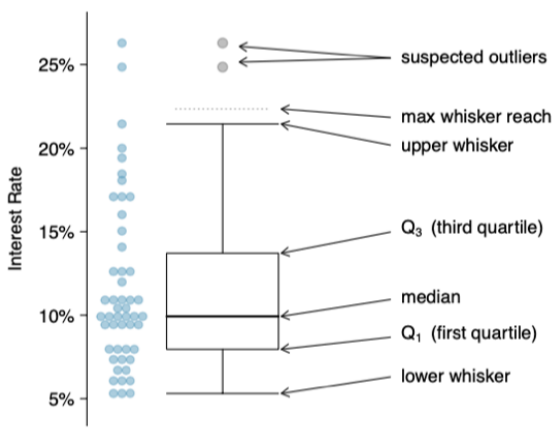
# 그림 상자

- 그림 상자는 5가지 통계량(upper whisker, $Q_3$, 중앙값, $Q_1$, lower whisker)과 이상치를 함께 제시하는 시각화방식이다.
- 그림상자에서 whisker는 사분위수로부터 1.5×IQR 이내에 떨어진 점 혹은 최댓/최솟점을 말한다. 여기서 IQR = $Q_3 - Q_1$ 이다. 먼저 max/min whisker reach를 정의하자.

  ↳ max whisker reach = $Q_1 - 1.5 \times IQR$

  ↳ min whisker reach= $Q_3 + 1.5 \times IQR$

# 그림 상자

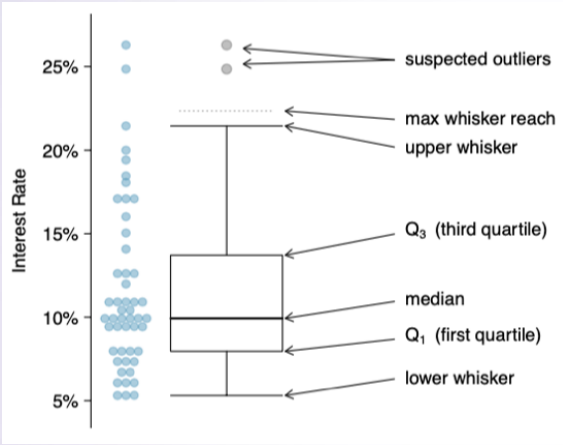



[ 이자율에 대한 그림 상자 ]
(OpenIntro Statistics, P49)

- 왼쪽 그림은 이자율 자료에 관한 그림상자이다.
- 여기서 lower whisker는 최솟값으로, upper whisker는 max whisker reach아래 데이터 중 가장 큰 값으로 정의된다.

사진 출처 #1

# 그림 상자




[ 이자율에 대한 그림 상자 ]
(OpenIntro Statistics, P49)

- max whisker reach보다 더 큰 데이터가 2개 관측되는데 이러한 데이터를 이상점(outlier)이라고 한다.
- 이자율 데이터에서 이상점의 값은 각각 24.85%와 26.30%이다.

사진 출처 #1

# 이상점과 로버스트 통계량(Robust Statistics)

- 이자율 데이터에서 최댓값인 26.3% 자료가 15%로 변경이 된다면 각종 통계량에는 어떤 변화가 생길까?
- 또한 이자율 데이터에서 최댓값인 26.3% 자료가 35%로 변경이 된다면 각종 통계량에는 어떤 변화가 생길까?
- 한 개의 데이터 값의 변화에 크게 좌우되지 않는 통계량을 로버스트(Robust)하다고 한다.

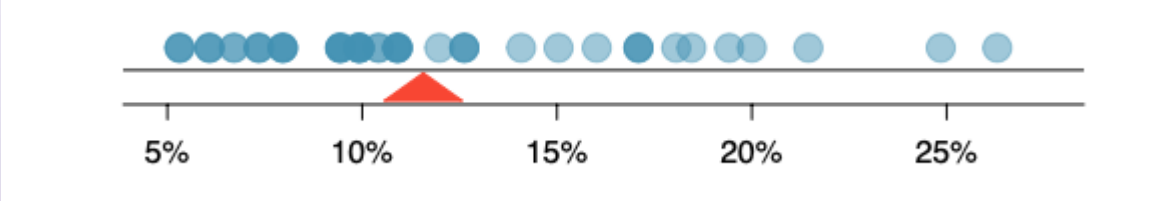
| | Robust 통계량 | | Robust 하지 않은 통계량 | |
|---|---|---|---|---|
| | 중앙값 | IQR | 평균 | 표준편차 |
| 원 자료 | 9.93% | 5.76% | 11.57% | 5.05% |
| 26.3%가 15%로 바뀐 경우 | 9.93% | 5.76% | 11.34% | 4.61% |
| 26.3%가 35%로 바뀐 경우 | 9.93% | 5.76% | 11.74% | 5.68% |

[ 로버스트 통계량 ]
(OpenIntro Statistics, P51)

기타 출처 #2

# Strip chart(dot plot)

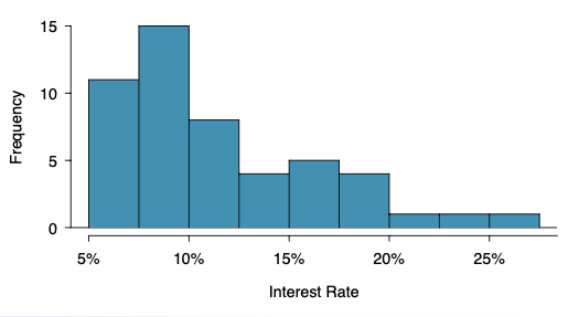
- Strip chart는 점 그림(dot plot)으로도 많이 알려져 있다.
- 아래 그림에서 볼 수 있는 것처럼 자료들이 겹치는 부분은 진한 색으로 표시된다. 이럴 경우 시각적 효과를 살리기 위해 자료를 흐트려서(jittering) 겹치지 않게 보이게 할 수 있다. 젤리 개수 추정치의 strip chart의 경우가 그렇게 표시된 예이다.
- 아래그림에서 빨간 삼각형이 가리키는 부분은 평균에 해당하는 지점이다.




[ 이자율에 대한 점 상자 ]
(OpenIntro Statistics, P42)

사진 출처 #3

# 히스토그램

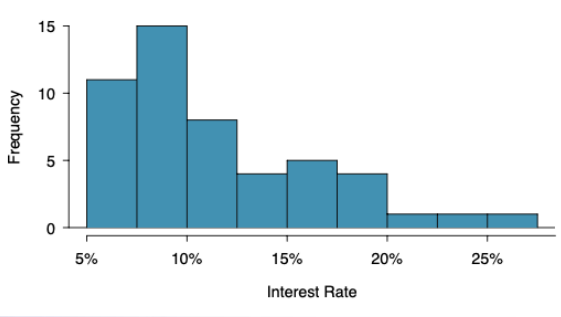



[ 이자율에 대한 히스토그램 ]
(OpenIntro Statistics, P45)

- Strip chart는 모든 데이터의 값은 정확히 보여주지만 데이터가 많은 경우 사용하기 어려울 수 있다.
- 이런 단점을 보완하기 위해 히스토그램이 사용될 수 있다. 즉 데이터를 몇 개의 구간으로 나눈 후에 각 구간에 포함된 데이터 개수를 그림으로 나타내는 것이다.

사진 출처 #4

# 히스토그램



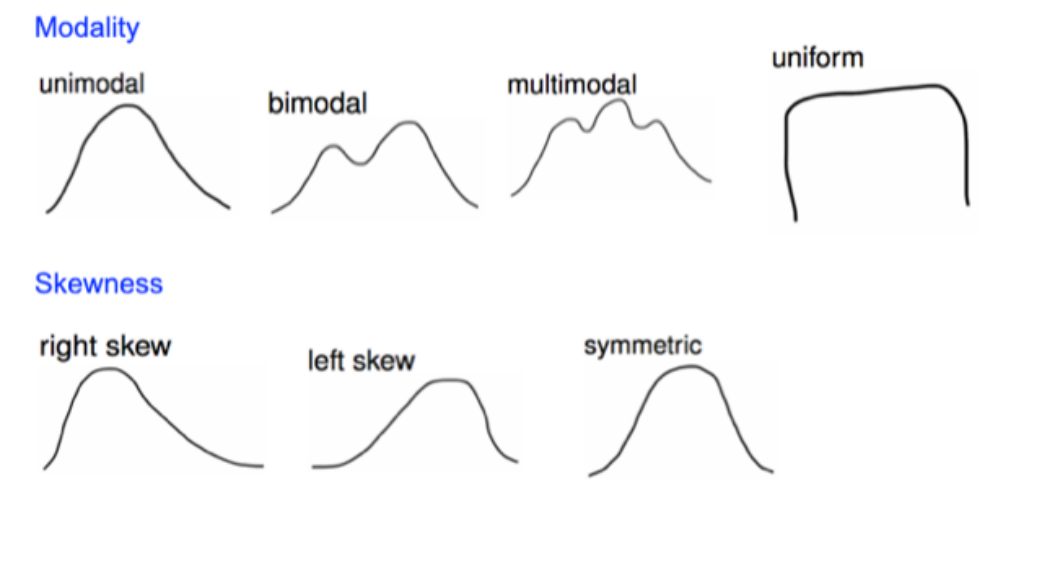

[ 이자율에 대한 히스토그램 ]
(OpenIntro Statistics, P45)

- 왼쪽 그림은 이자율데이터를 총 9개의 구간으로 나눈 후 히스토그램을 그린 것이다. 여기서 구간의 길이는 2.5%이며 5%부터 구간을 시작하였다.

사진 출처 #4

# 분포의 형태

## Modality와 Skewness




[ 분포의 형태 ]
(Openintro Statistics Lecture Slide Chapter 2, p51)

사진 출처 #5

## 출처

#1~4 Diez, D.M., Barr, C.D. and Çetinkaya-Rundel, M, (2019), OpenIntro Statistics, 4th edition, OpenIntro, Inc.

#5 Diez, D.M., Barr, C.D. and Çetinkaya-Rundel, M, (2019), Openintro Statistics Lecture Slide Chapter 2, https://bit.ly/3jCr7