

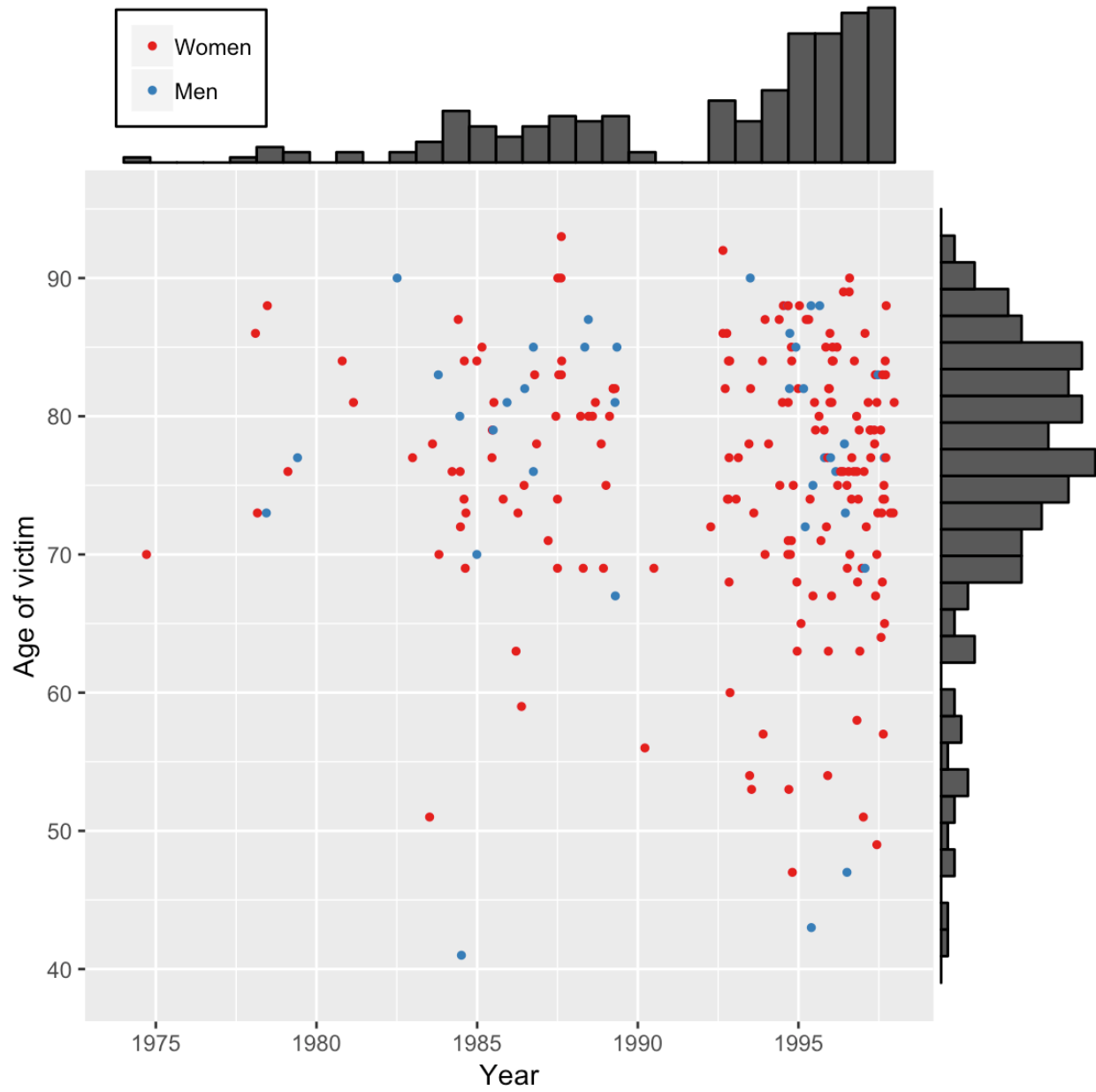
# 데이터 분석(Python)

기초 통계



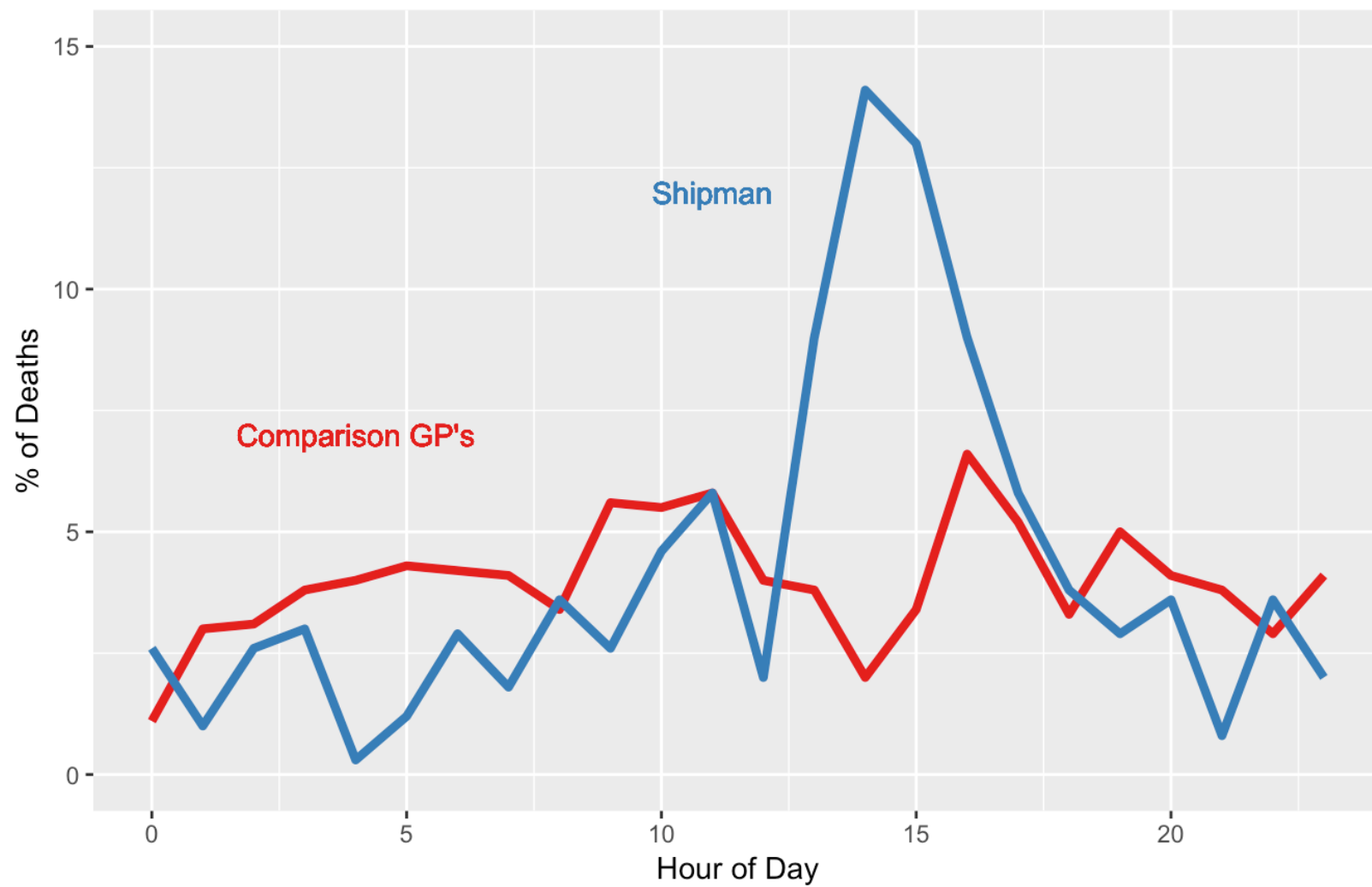
영국	해럴드 시프먼	215
콜롬비아	루이스 가라비토	138
콜롬비아	페드로 로페스	110
독일	닐스 회겔	106
일본	이시카와 미유키	103
파키스탄	자베드 이크발	100
러시아	미하일 포프코프	83
콜롬비아	다니엘 카마르고	72
브라질	페드루 호드리게스 필류	71
인도	캄파티마르 샹카리아	70
중국	양신하이	67
러시아	안드레이 치카틸로	53
우크라이나	아나톨리 오노프리옌코	52
독일	브루노 뢰트케	51
미국	새뮤얼 리틀	50



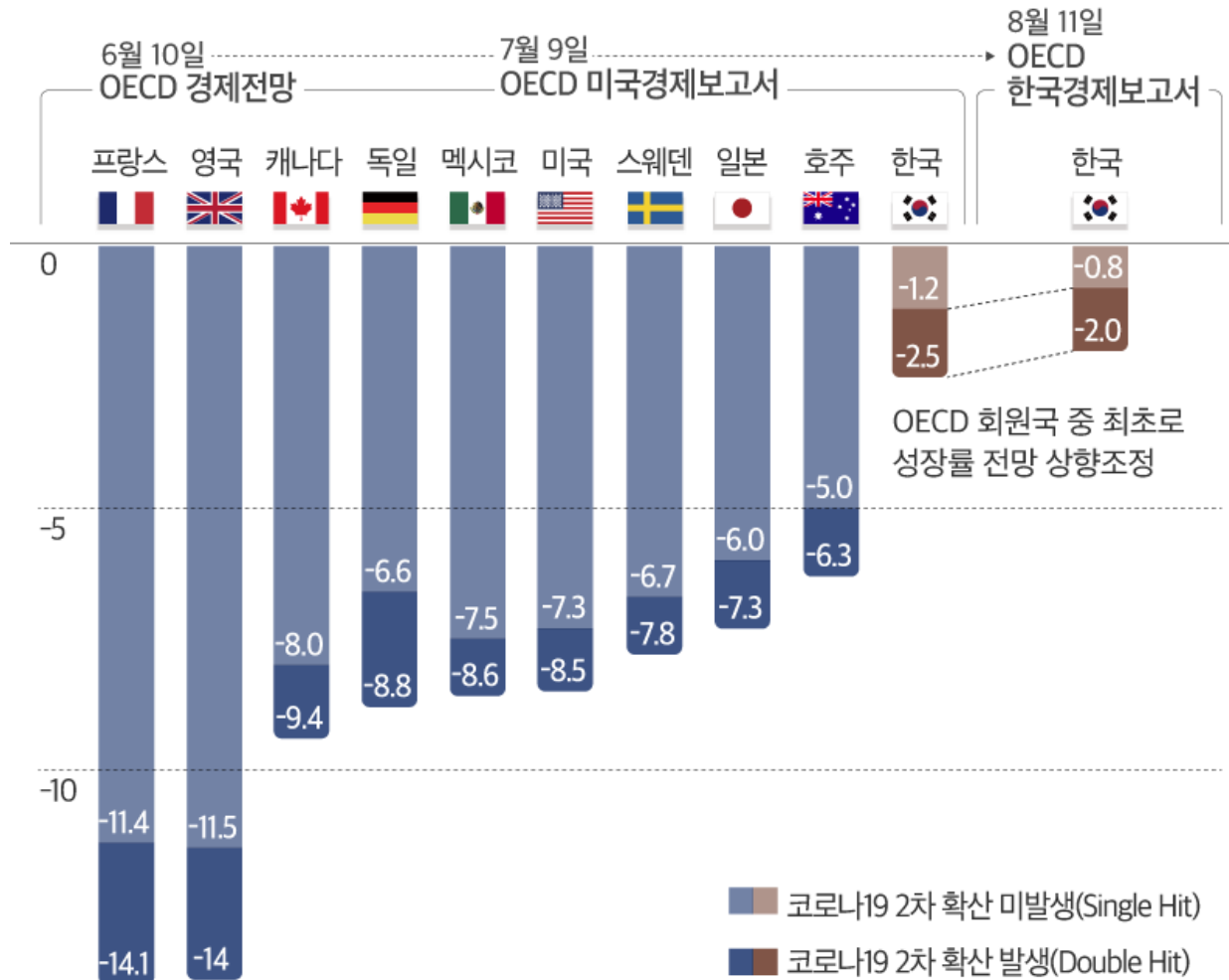


## Deaths by Hour of Day

From Shipman dataset



## OECD 회원국 가운데 주요 10개국 경제성장률 전망 단위: %



❖ 내년 성장률은 37개국 중 34위



# Data Literacy

---



## ❖ 미국 임금 76% 오를때 한국은 154%나 뛰었다

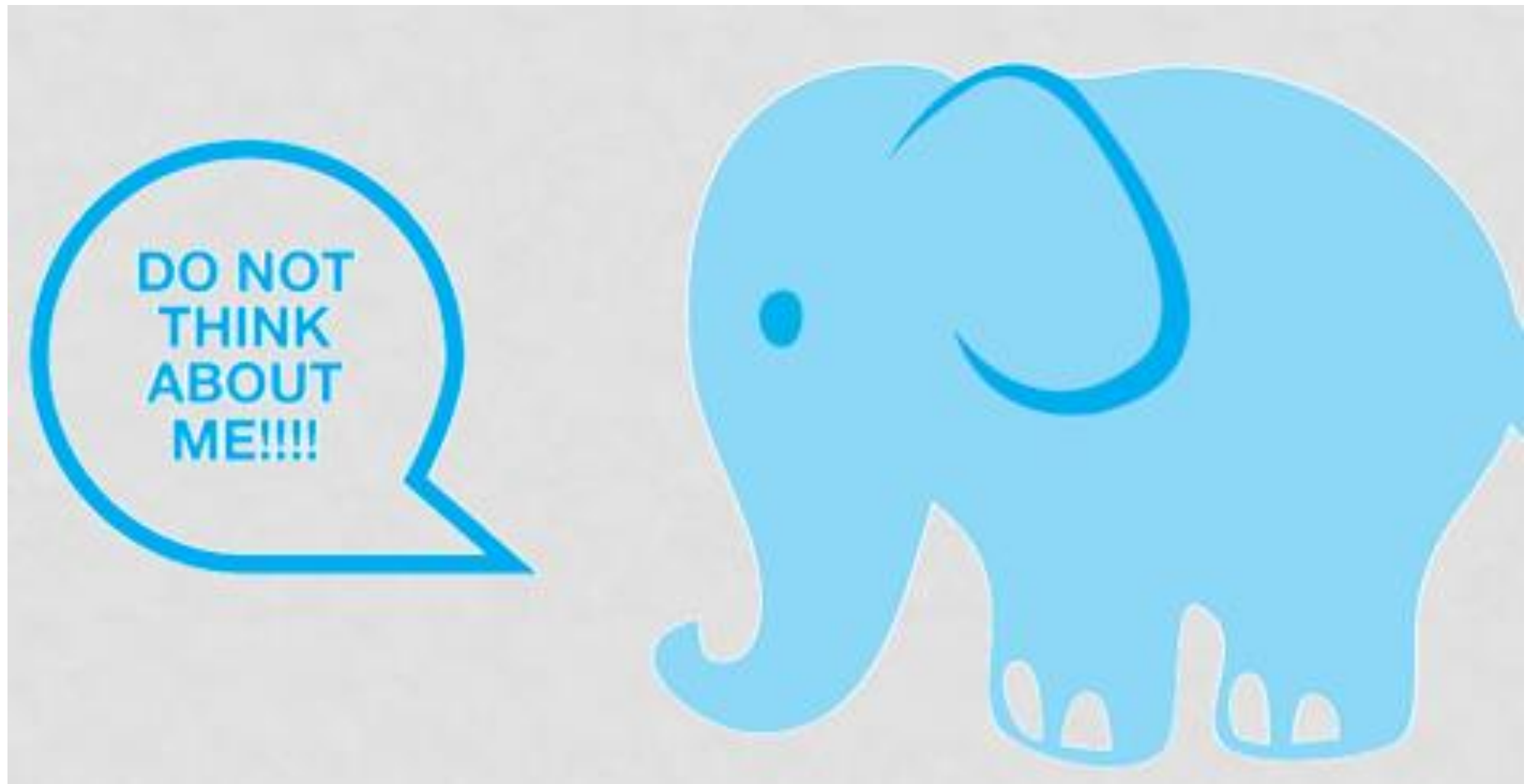
- 지난 20년간 한국 민간부문의 시간당 임금 상승률이 154%를 기록해 미국(76%), 영국(87%) 등 선진국 상승률의 2배에 달하는 것으로 분석됐다. 반면 한국의 생산성은 미국의 절반 수준에 불과한 것으로 나타났다

## ❖ 20년간 2.5배 뛴 임금...구두 단가도 이탈리아보다 높다

- 경제협력개발기구(OECD)에서 집계하는 국가별 임금 통계, 평균 연봉, 노동생산성 등을 분석한 결과, 한국은 지난 20년간(1997~2017년) 시간당 임금이 2.5배(154.1%) 이상 급등한 것으로 나타났다. 미국(76.3%)의 2배, 독일(54.9%)의 3배 수준이다. 반면 한국의 시간당 노동생산성(2017년)은 미국의 54%, 독일의 57% 수준으로 절반을 간신히 웃돌았다

국가	노동생산성(달러)		노동생산성 증가율	임금 증가율	차이(임금증가 율-생산성증가 율)
	1997년	2017년			
한국	15.6	34.3	119.9	154.1	34.2
일본	31.7	41.8	31.9	-9.3	<b>-41.2</b>
미국	48.6	64.1	31.9	76.3	44.4
독일	48.6	60.5	24.5	54.9	30.4
프랑스	48.5	59.8	23.3	66.2	42.9
영국	38.9	52.5	35	87.1	52.1





Framing Effect |

가공육(햄, 베이컨 등)은 담배와 석면이 속해 있는 '1군 발암물질'에 해당되며, 매일 50g의 가공육을 먹으면 장암 발병율이 18% 높아질 수 있다.

- 국제보건기구(WHO)의 국제암연구소(IARC), 2015년 11월

*Relative Risk(상대위험도)* : 하루에 베이컨 두 개가 들어 있는 샌드위치를 먹는 집단이 그러지 않는 집단에 비해 장암에 걸릴 위험이 18% 증가한다.

*Absolute Risk(절대위험도)* : 각 집단에서 장암에 걸릴 것이 라고 예상되는 비율

- 만일 평균적으로 100 명중 6명이 장암에 걸린다고 하면, 100명이 매일 베이컨 두개가 든 샌드위치를 먹는다면 7명이 장암에 걸릴 확률이 있다.

*Expected Frequency(기대 빈도)* : 100명 혹은 1000명에 대해 기대되는 숫자

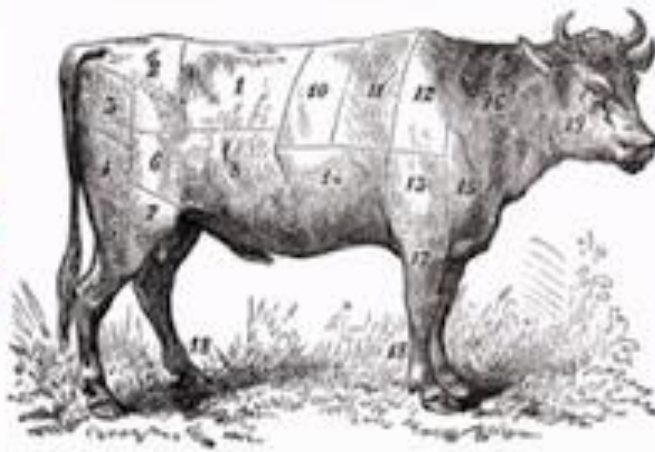
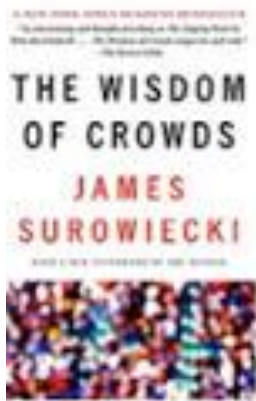
- 장암발생 기대 빈도 :  $6 / 100$ ,  $1 / 16$
- 매일 샌드위치를 먹은 사람의 장암발생 기대 빈도 :  $7 / 100$ ,  $1 / 14$

	베이컨 미섭취	베이컨 섭취
발생률	6%	7%
기대빈도	6/100, 1/16	7/100, 1/14
승산(odds)	6 / 94	7 / 93

❖ *odds(승산)* : 사건이 일어날 가능성 대 사건이 일어나지 않을 가능성

절대위험도 차이	1%
상대위험도	18%
승산비(odds rate)	$(7/93) / (6/94) = 1.18$

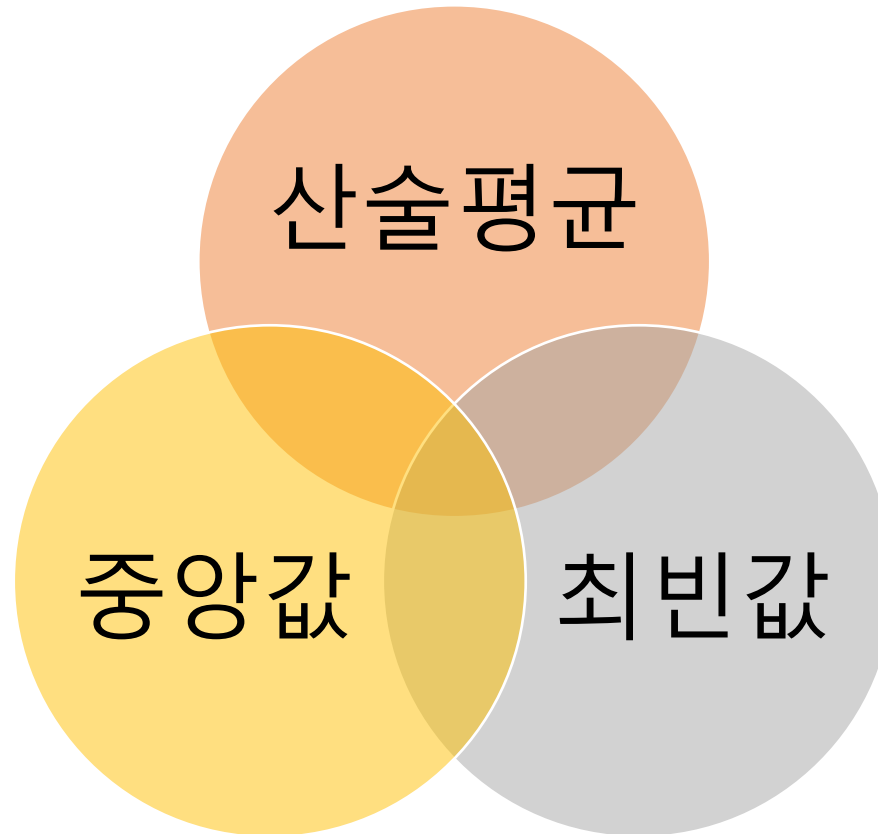
# The Wisdom of Crowds



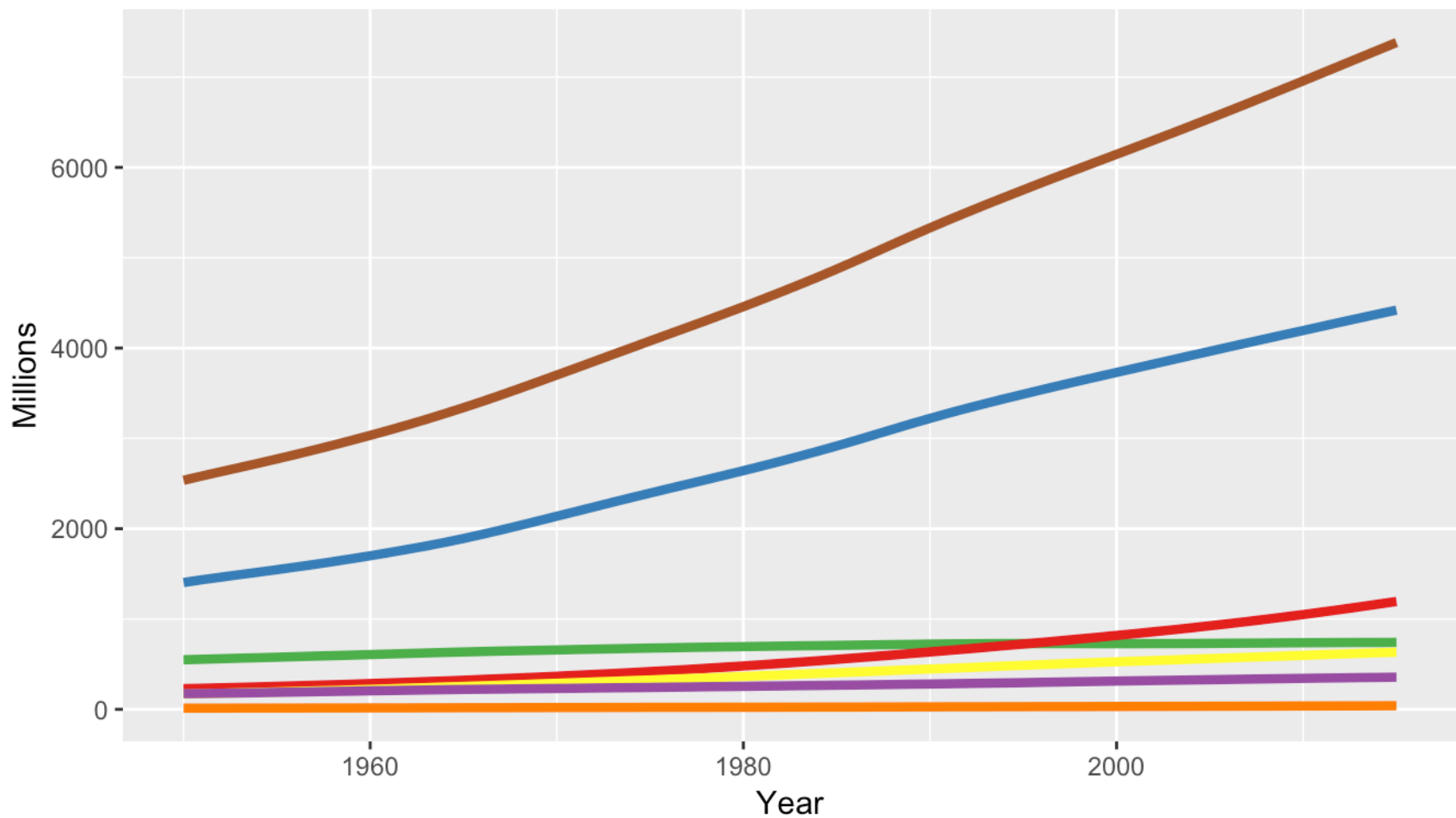
*average of 800 guesses = 1,197*  
*actual weight of the ox = 1,198*

96

## ❖ Average(평균)



[https://www.youtube.com/watch?v=Pp\\_Pd6GZLOE&feature=youtu.be](https://www.youtube.com/watch?v=Pp_Pd6GZLOE&feature=youtu.be)



colour



Africa

Asia

Europe



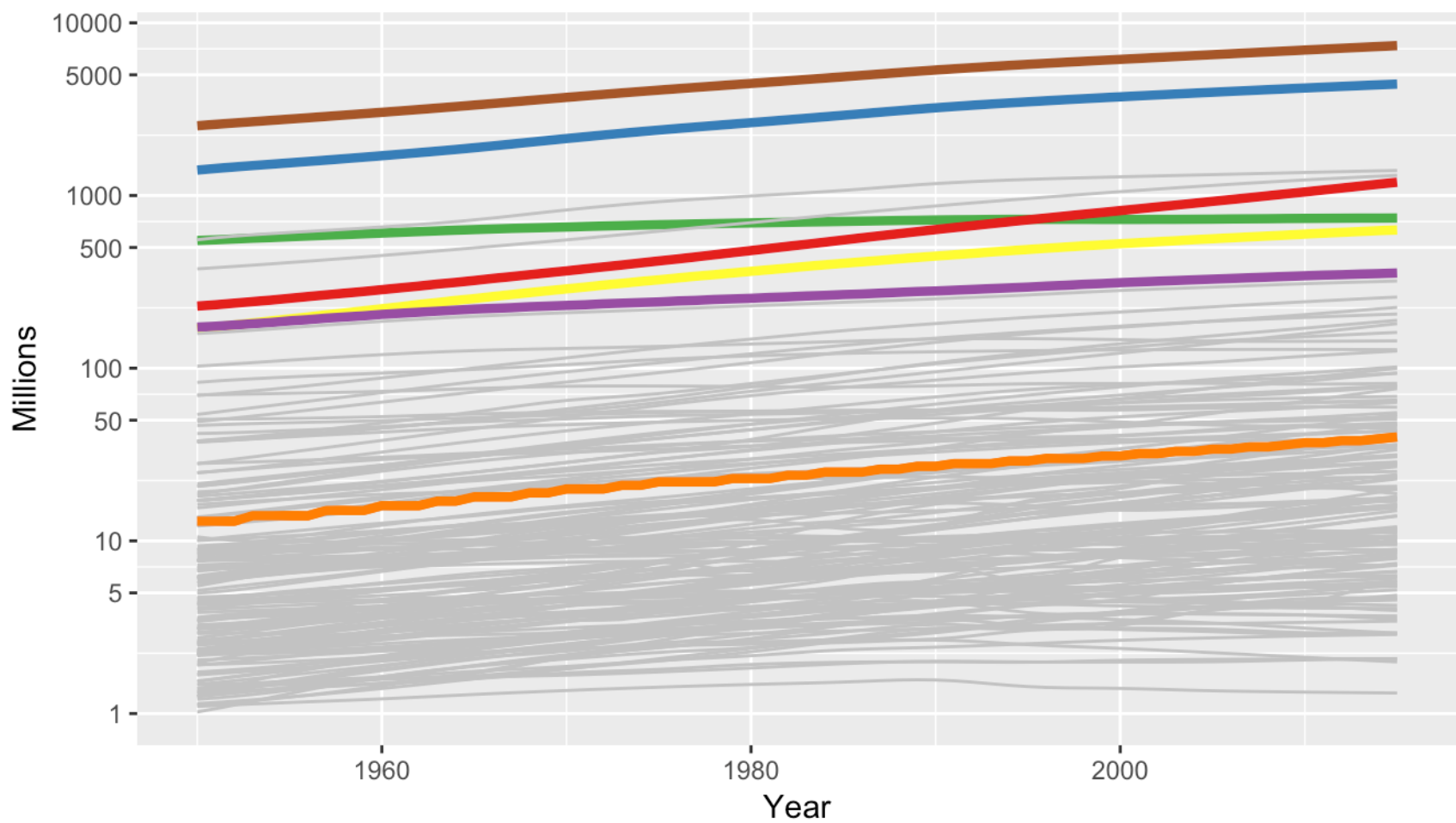
N America

Oceania



S America

World



colour

Africa

Europe

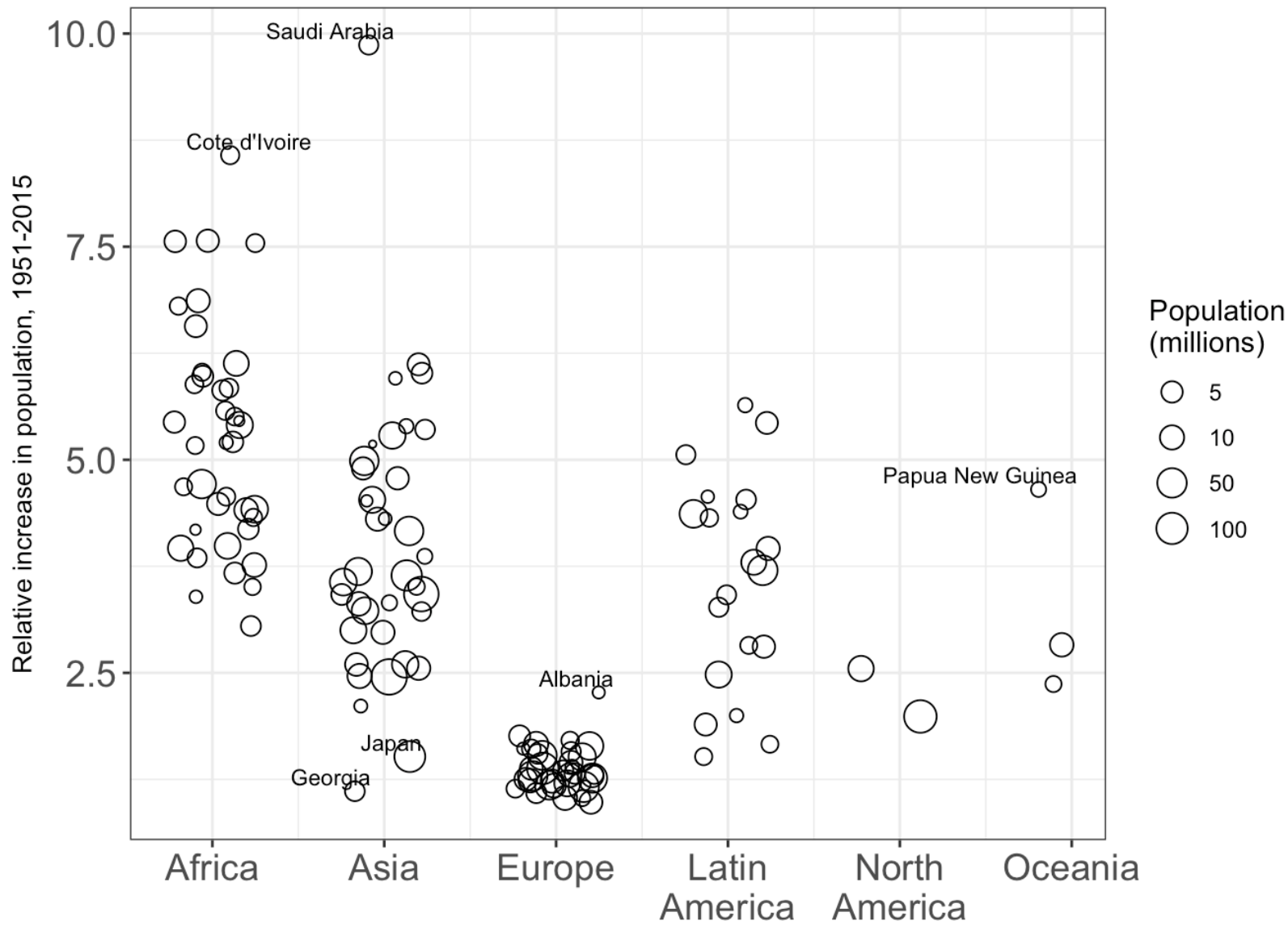
Oceania

World

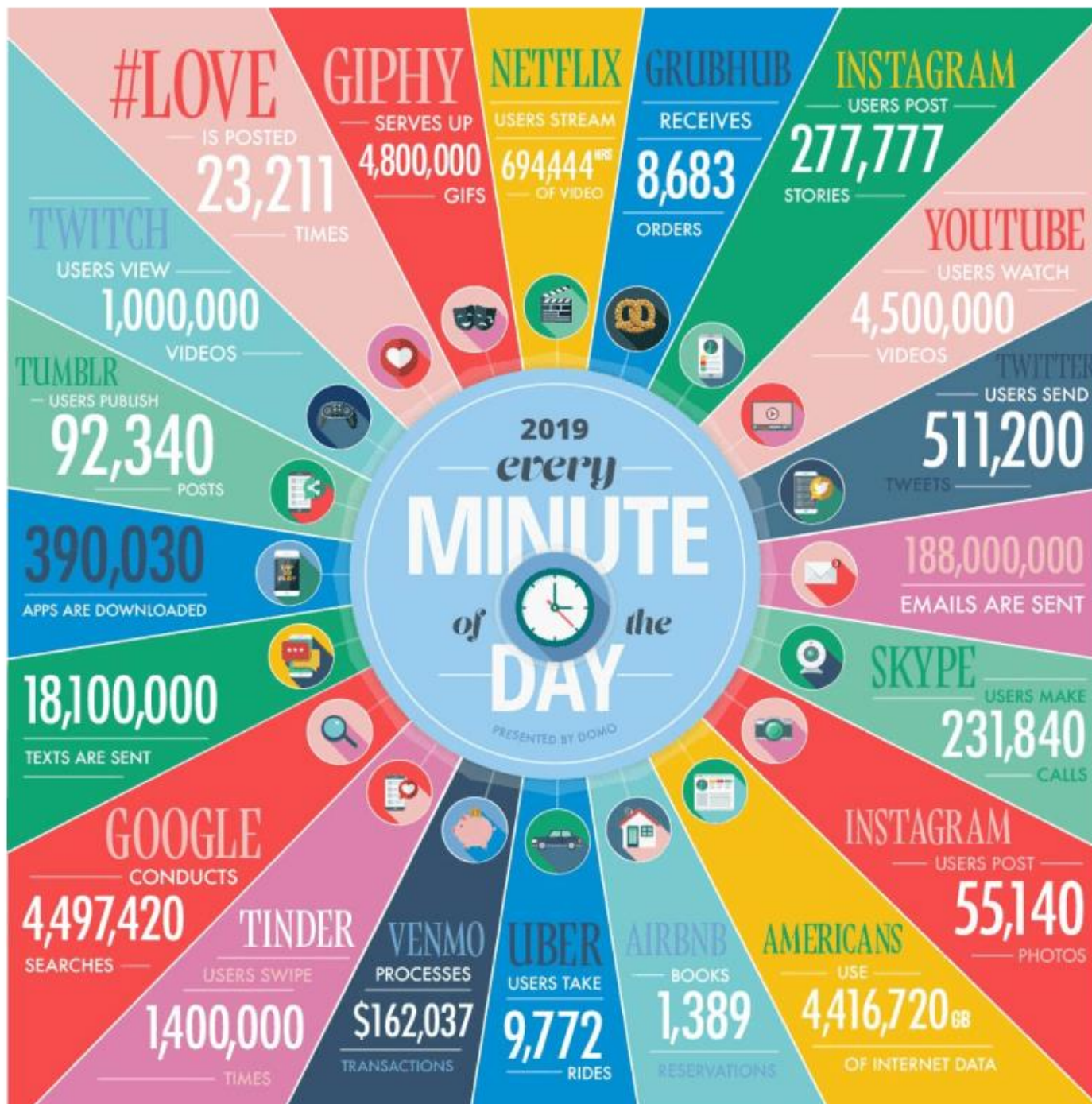
Asia

N America

S America





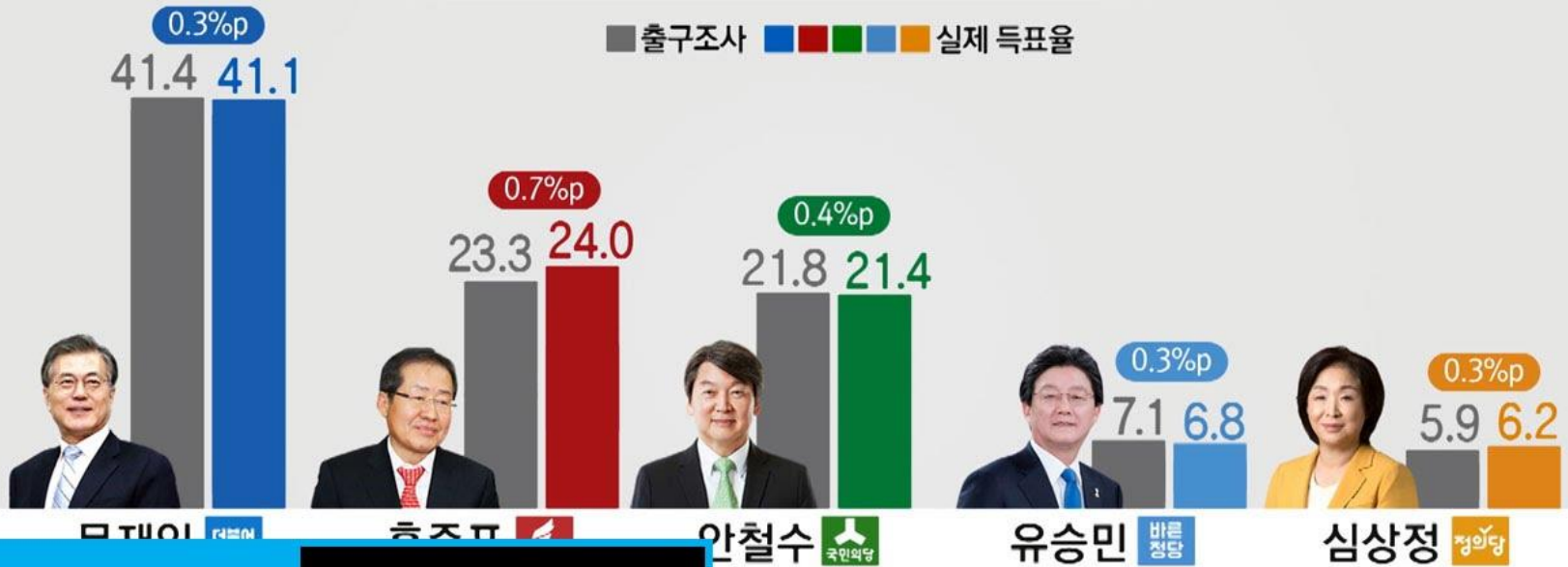


[https://www.ted.com/talks/hans\\_rosling\\_global\\_population\\_growth\\_box\\_by\\_box#t-6756](https://www.ted.com/talks/hans_rosling_global_population_growth_box_by_box#t-6756)



# 19대 대선 후보별 출구조사-실제 득표율 비교 %

■ 출구조사 ■ ■ ■ ■ ■ 실제 득표율



YTN NEWS



## 방송 3사 민주·시민 vs 통합·한국 출구조사 결과



## 2020 우리의 선택 LIVE는 JTBC



## 21대 총선 | JTBC 예측조사

민주·시민 143~175석  
통합·한국 101~134석 예측

4·16 총선

## 21대 총선 정당별 의석수 현황



자료/ 중앙선거관리위원회

YONHAP NEWS

장성구 김토일 기자 / 20200416 / 페이스북 tune.y.kr/LeYN1, 트위터 @yonhap\_graphics

- 기술 통계 : 관측을 통해 얻은 데이터에서 그 데이터의 특징을 뽑아 내기  
위한 기술
- 추리 통계 : 전체를 파악할 수 없을 정도의 큰 대상이나 아직 일어나지 않은,  
미래에 일어날 일에 관해 추측하는 기술
- 통계 : 관측된 데이터의 집합이기 때문에 과거에 일어난 것에 관한 기술
- 확률 : 미래에 일어날 것에 관한 기술

## 중학생 100명의 키(cm)

155	159	152	175	153
172	167	156	144	176
147	149	169	153	152
162	167	157	157	171
175	146	145	155	171
152	173	147	164	167
163	153	151	159	167
161	146	160	177	156
160	140	172	157	170
170	148	177	170	146
174	160	140	166	172
147	154	149	165	179
143	180	165	178	146
167	156	164	146	158
159	171	167	143	147
162	157	148	169	153
175	161	172	158	175
168	145	160	149	153
169	150	163	150	146
149	178	156	162	170

- 데이터(관측치)

- 분포

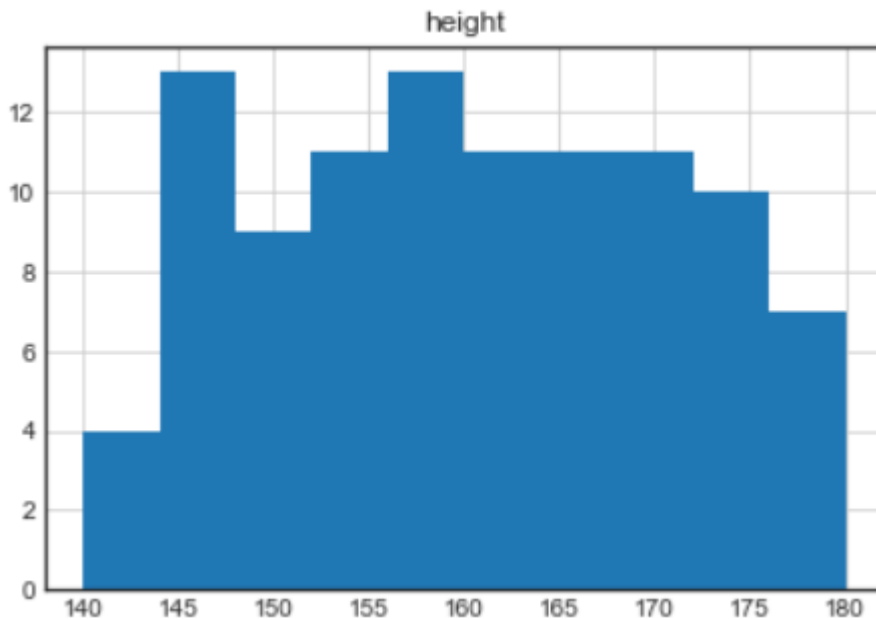
- 통계

- 통계량

## ❖ 도수 분포표

1. 최대값, 최소값
2. 구간 설정(계급)
3. 대표 수치 설정(계급값)
4. 각 계급에 속한 데이터 개수(도수)
5. 각 도수가 전체에서 차지하는 비율(상대도수)
6. 어느 계급까지의 도수의 합(누적도수)

	계급	계급값	도수	상대도수	누적도수
0	140~145	143.0	7	0.07	7
1	145~150	147.0	18	0.18	25
2	150~155	153.0	12	0.12	37
3	155~160	158.0	17	0.17	54
4	160~165	163.0	11	0.11	65
5	165~170	168.0	15	0.15	80
6	170~175	172.0	13	0.13	93
7	175~180	178.0	7	0.07	100



- 비교적 골고루 분포(정규분포 형태)
- 좌우대칭성

## ❖ 평균

- 데이터는 수치적으로 널리 퍼져있지만, 그 널리 퍼져 있는 것 중에 하나의 수를 모든 데이터를 대표하는 수로 뽑은 것
- 데이터들은 평균 주변에 분포
- 많이 나타나는 데이터는 평균값에 주는 영향력이 큼
- 히스토그램이 좌우 대칭이면, 평균값은 대칭이 되는 축에 위치

## ❖ 평균의 종류

- 산술 평균 : 5와 10의 평균은?
- 상승 평균, 기하 평균 : 작년에 50% 성장, 올해 4% 감소, 이때 평균 성장률은?
- 제곱 평균 : 표준편차
- 조화평균 : 비교의 의미



❖ 중간고사와 기말고사의 성적이 각 10, 90 일 때,

- 산술 평균 : 50
- 상승 평균, 기하 평균 : 30
- 제곱 평균 : 표준편차 : 64.03
- 조화평균 : 비교의 의미 : 18

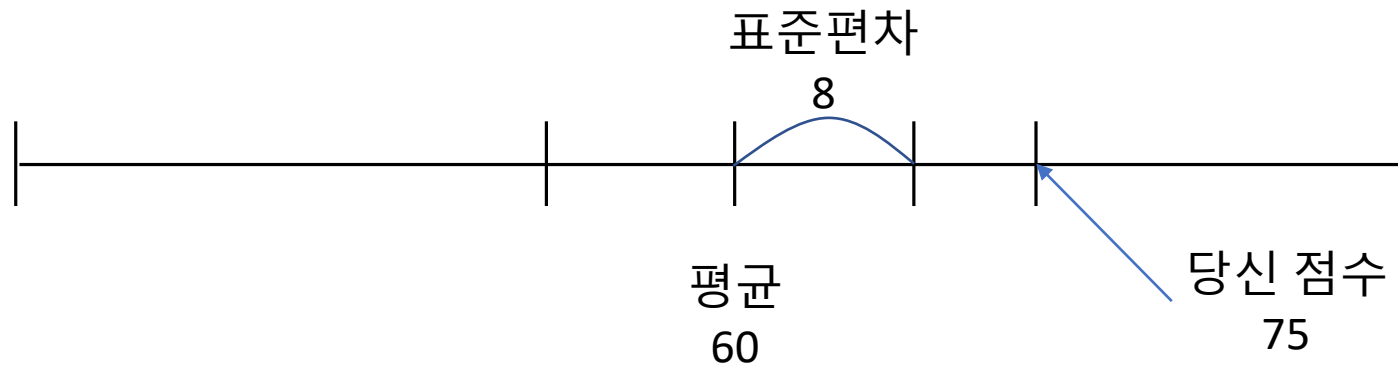
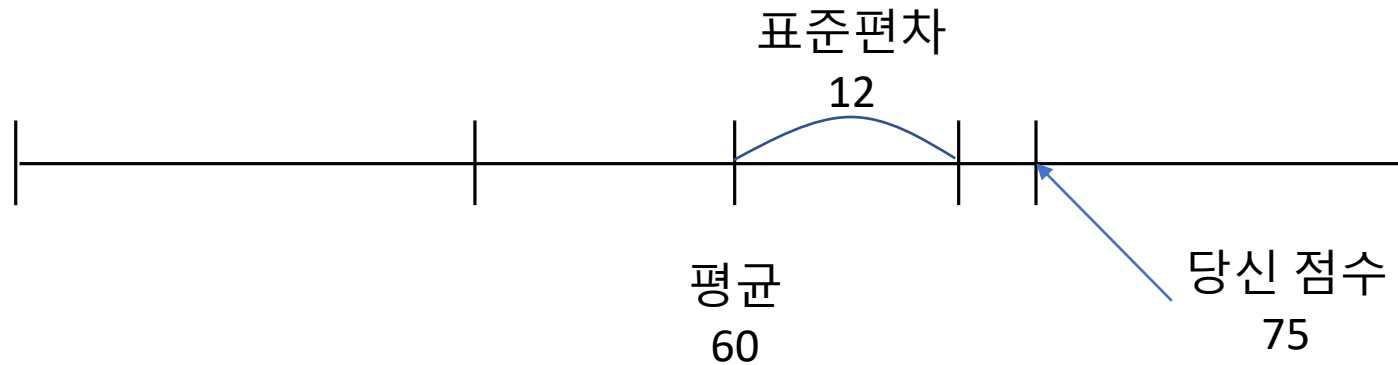
❖ 편차

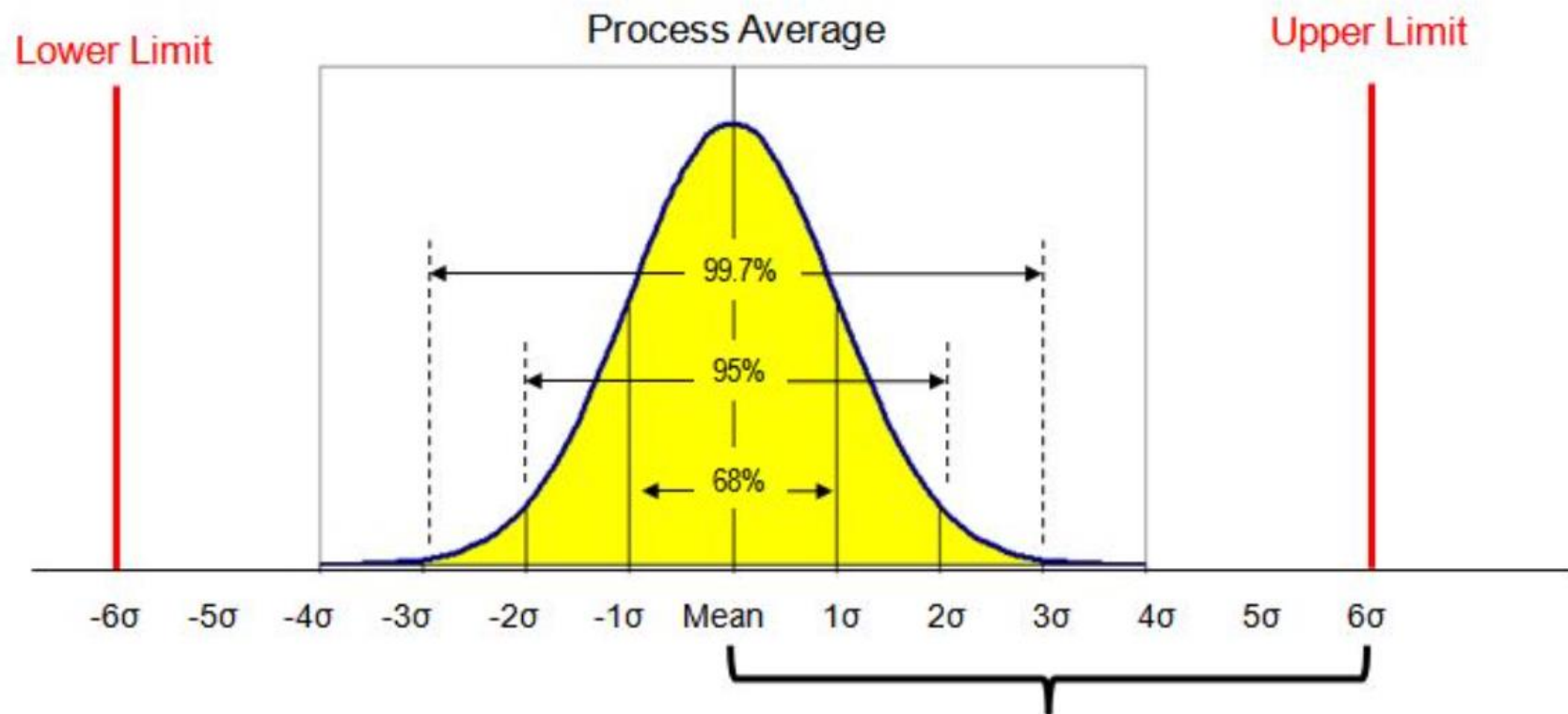
❖ 분산

❖ 표준편차

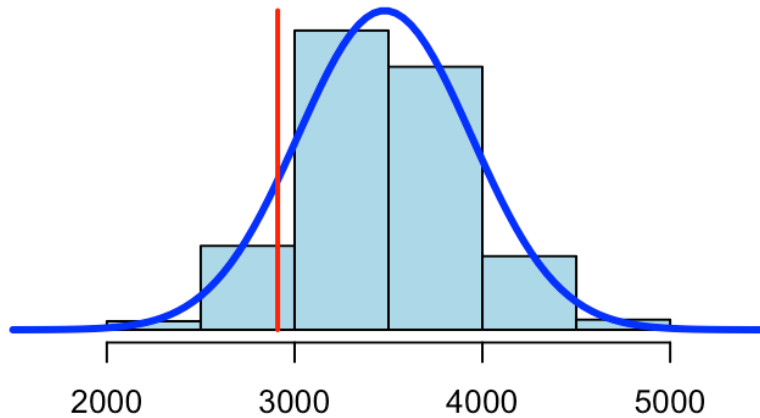
## ❖ 표준 편차

- 데이터 세트 중에 있는 어떤 데이터 하나의 수가 가지는 의미
- 여러 데이터 세트들을 서로 비교해서 나타나는 차이

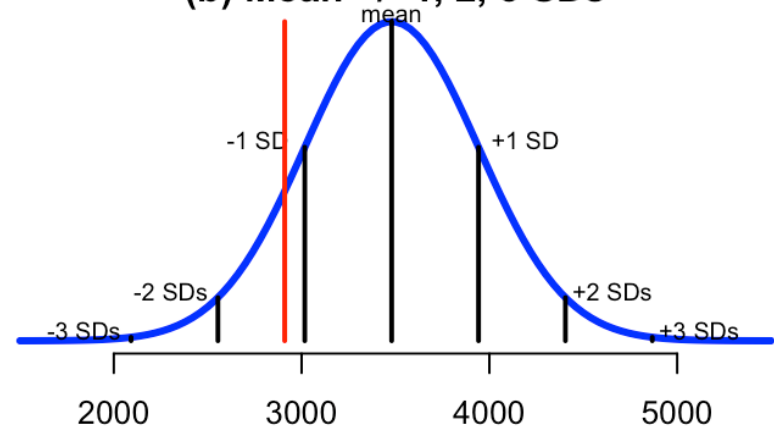




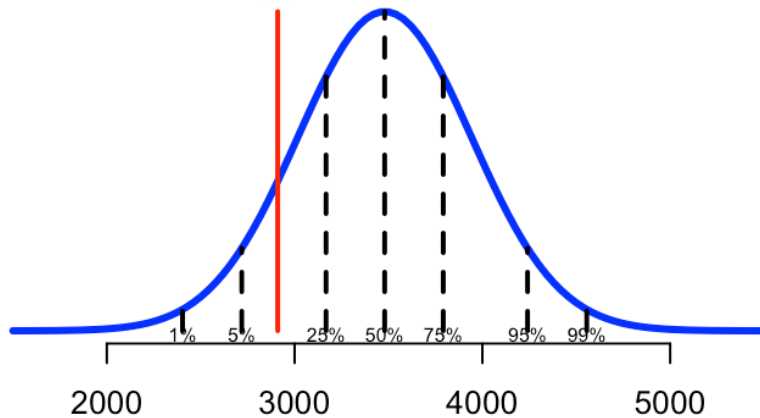
**(a) Distribution of birthweights**



**(b) Mean  $\pm$  1, 2, 3 SDs**

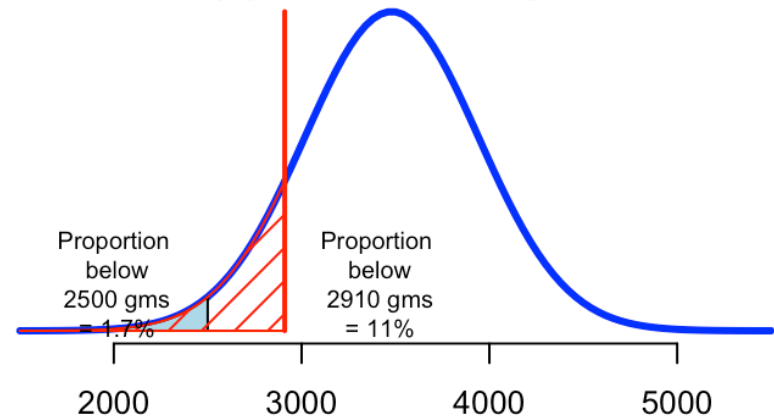


**(c) Percentiles**



Birthweight (gms)

**(d) Low birth weight**

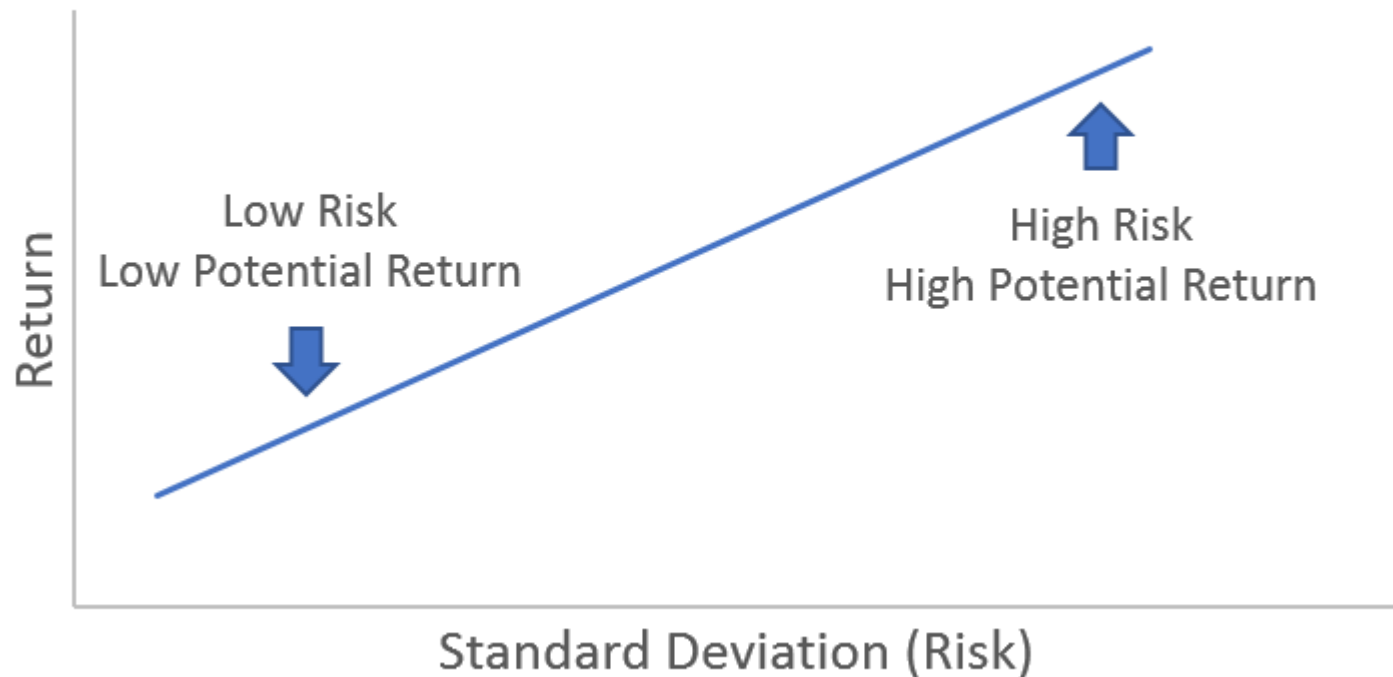


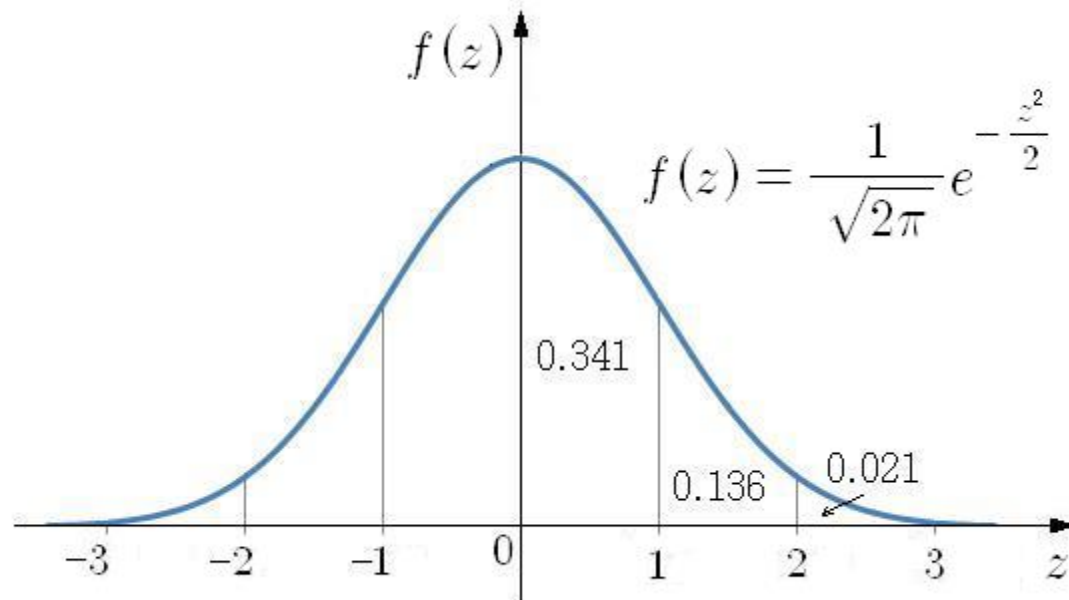
Birthweight (gms)

## ❖ 투자 상품 수익률

	A	B	C	D	E
월평균수익률	2.05	2.46	-1.33	2.04	-0.54
표준편차	5.35	9.11	5.91	5.98	13.7

## Risk-Return Trade-off



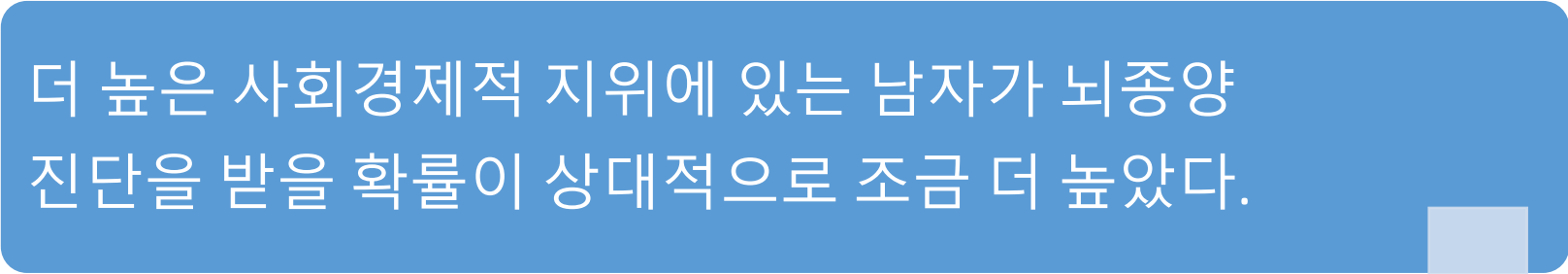


❖ 표준정규분포

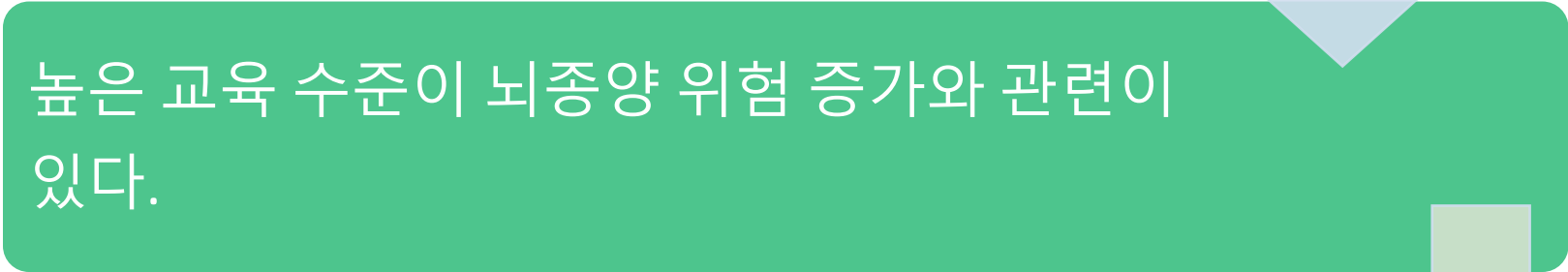
- 평균 : 0, 표준편차 : 1
- -1 ~ 1 의 상대 도수 : 0.6826
- -2 ~ 2 의 상대 도수 : 0.9544

## ascertainment bias(확인 편향)

더 높은 사회경제적 지위에 있는 남자가 뇌종양 진단을 받을 확률이 상대적으로 조금 더 높았다.



높은 교육 수준이 뇌종양 위험 증가와 관련이 있다.



왜 대학에 가면 뇌종양에 걸릴 위험이 커지는가?



## 인과관계(causation) vs. 상관관계(correlation)

- $x$ 가  $y$ 의 원인이다
- $x$ 가 일어날 때마다  $y$ 가 일어난다
- $x$ 가 일어난다면  $y$ 가 일어날 수 밖에 없다
- $x$ 가 일어나게 만들면,  $y$ 가 더 자주 일어나는 경향이 있다



## 제 1상 임상시험

인체를 대상으로 후보 의약품 또는 백신에 대한 임상시험을 처음으로 수행하는 경우 일반적으로 **소수의 건강한 지원자**들에게 약물이 제공됩니다. 그러나 암 등의 말기 질환을 위한 치료제의 경우 해당 질환을 앓고 있는 지원자를 대상으로 임상시험이 수행되기도 합니다.

제 1상 임상시험의 주요 목표는 다음과 같습니다.

- 약물 사용으로 인해 안전상의 중대한 문제가 발현되지는 않는지 확인합니다.
- 목표한 신체 부위로 약물이 미칠 수 있는지, 효능이 전달될 수 있도록 약물이 충분히 지속되는지를 확인합니다.
- 약물이 치료 가치를 제공하거나 질환 또는 상태를 예방해줄 수 있는가에 대한 일차적 근거를 확보합니다.

## 제 2상 임상시험

제 1상 임상시험의 결과가 성공적인 경우, 보다 많은 수의 사람들을 대상으로 하는 임상시험이 수행됩니다. 제 2상 임상시험은 **후보약물을 필요로 하는 환자들을 대상**으로 진행되며 다음과 같은 목표를 가집니다.

- 질환 치료에 대한 유효성
- 질환 예방에 대한 유효성 (지원자가 해당 질환을 앓고 있지 않은 경우)
- 약물의 적정 용량 및 용법

이 단계에서는 약물의 사용이 위약을 제공 받은 환자군을 통해 비교될 수 있습니다. 위약은 후보약물과 동일한 형태를 지녔으나 유효 성분을 지니고 있지 않습니다.

여기에서는 약물의 사용 경과를 판단할 수 있도록 기준이 되는 집단이 설정됩니다. 중요한 점은 환자나 연구자가 각각의 지원자가 어떤 치료제를 제공 받았는지에 대해 알아서는 안된다는 것입니다. 이는 이중맹검 위약대조라고 알려진 연구 방법으로, 시험 결과가 편향되는 것을 방지합니다.

## 제 3상 임상시험

제 2상 임상시험 결과가 긍정적인 경우 다음 단계의 임상시험이 진행됩니다. 이는 **다양한 국가의 수백 또는 수천 명의 지원자**를 대상으로 진행됩니다.

제 3상 임상시험의 주요 목표는 다음과 같습니다.

- 신약 또는 백신 사용 환자 대상 안전성 및 유효성 검사
- 효과적인 용량 및 용법 확인
- 부작용 또는 치료제가 특정 상태의 사람들에서는 사용돼서는 안 되는 이유 확인 (금기사항)
- 약물 또는 백신의 이점에 대한 정보 수집 및 위험 요소 대조
- 기존 치료 요법과의 결과 대조

성공을 위해서는 신약이 기존의 치료제보다 더욱 우수한 치료 효과를 제공할 수 있어야 합니다.

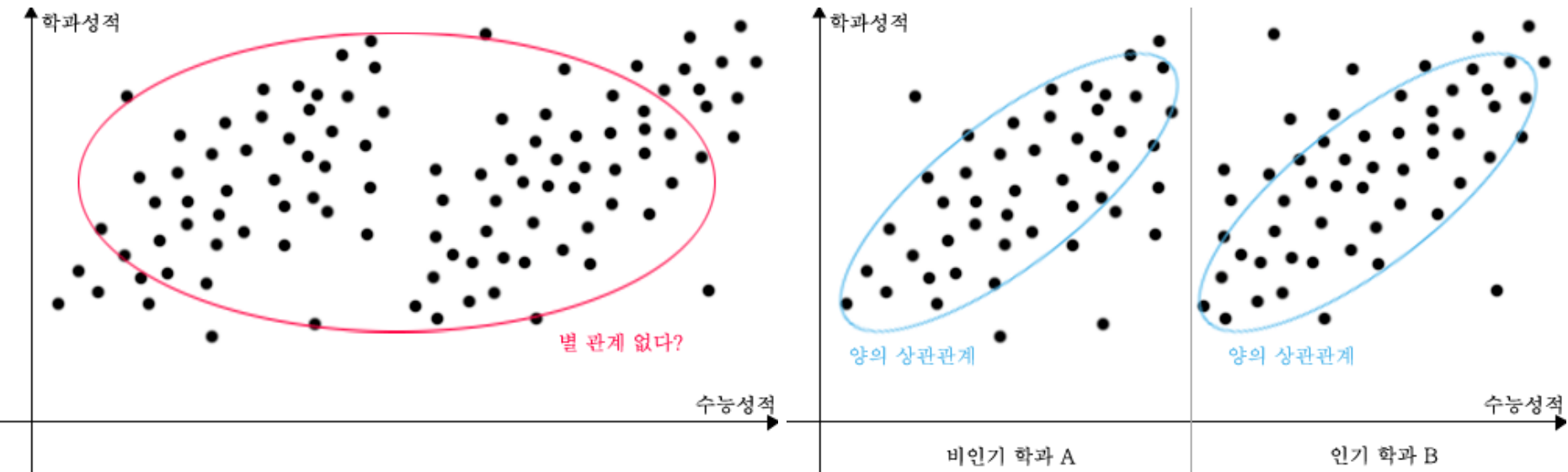
제 3상 임상시험 수행을 위해서는 수 년이 소요됩니다. 새로운 약물이나 백신이 제 3상 임상시험을 통해 긍정적인 결과를 입증할 경우, 국가나 지역 단위로 승인 신청을 할 수 있습니다.

새로운 의약품의 경우 규제기관은 전임상시험 및 임상시험 결과를 바탕으로 이 약물이 어떻게 사용돼야 하는지, 어떤 환자에게 사용돼야 하는지 등을 결정하는데, 이를 의약품의 적응증이라고 합니다.

❖ 아이스크림 판매가 늘어나면 익사 사고가 많이 발생한다.

- 날씨 : confounder(교란변수)
- 조정(adjustment), 층화(stratification) : 교란변수의 수준별 연관성 제고,  
날씨가 같은 날의 아이스크림 판매량과 익사 사고 발생건수의 관계

❖ Simpson's Paradox



- ❖ 인구 1000명당 어린이집 개수가 높은 지역일 수록 출산율이 높다.
- ❖ 탄산음료가 10대 청소년을 폭력적으로 만든다.
- ❖ 카레가 치매 예방에 도움이 된다
- ❖ 왼손잡이가 더 오래 산다.
- ❖ 교황의 평균 수명은 일반인의 평균수명보다 길다.
- ❖ 영유아 예방접종이 자폐증 발생 확률을 높인다.

역인과관계(reverse causation)

잠복변수(lurking variable)

### 안아키 논란 주요 쟁점 3가지

---

 **김효진 한의사**

**엄마가 동의해서** 아이에게 백신을 맞게한 뒤 부작용이 생기면 아무도 **보상해주지 않는다**

**질병관리본부** 

국가 예방접종 때문에 부작용이 생겼다는 **인과 관계가 증명되고, 30만원 이상 비용**이 발생한 경우에는 엄마의 **동의 여부와 관련 없이** 국가에서 해당 진료비 전액과 간병비를 보상한다

---

 **김효진 한의사**

수두는 **어릴 때 앓으면** 가볍게 지나가므로 예방접종을 **안 해도 된다**

**질병관리본부** 

대부분 수두를 **가볍게 앓고 지나갈 수 있다**는 건 맞는 말이지만 **일부에서는 뇌염·폐렴 등 위험한 합병증**이 생길 수 있다

---

 **김효진 한의사**

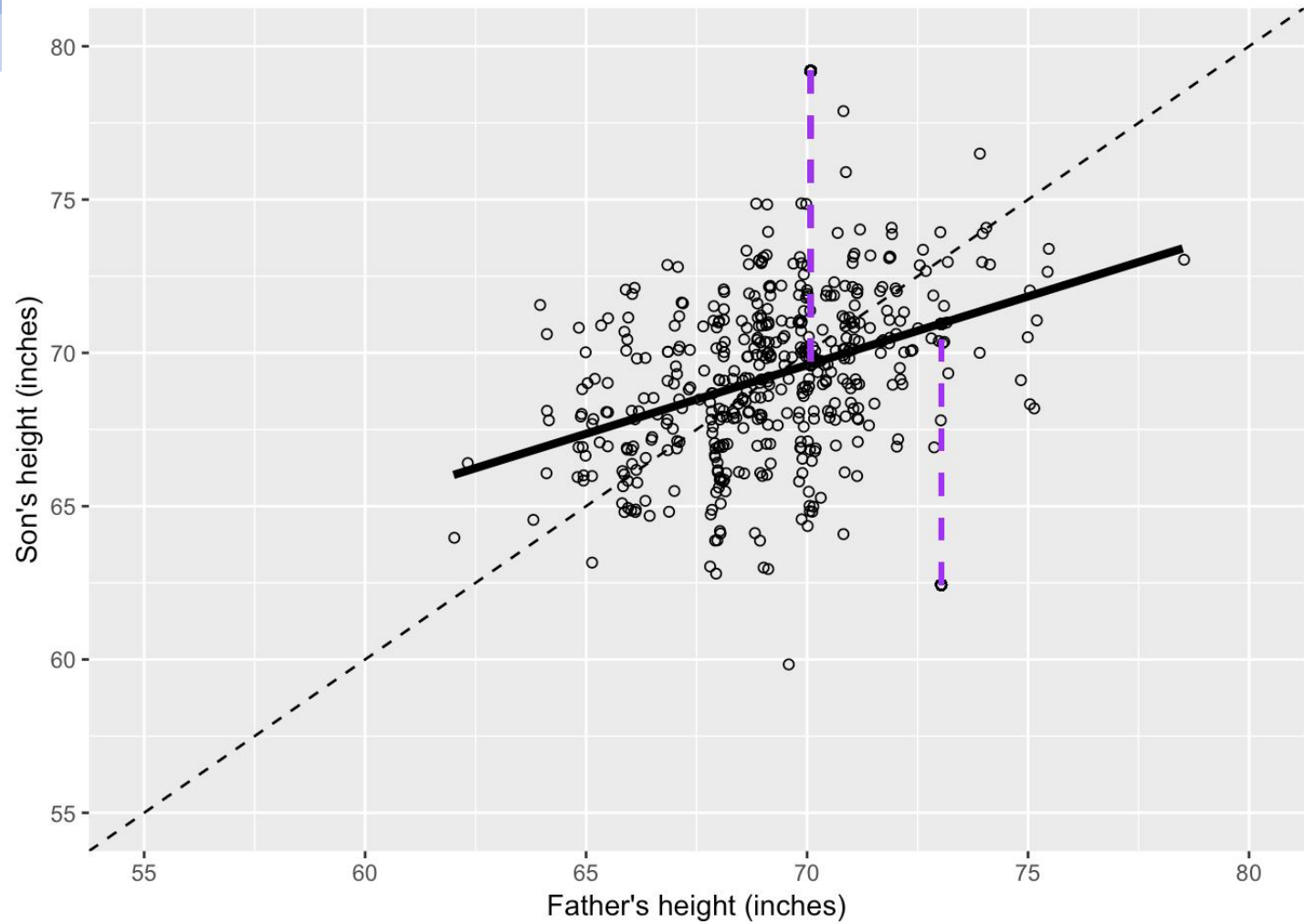
한의사로서 20년을 연구해보니 아토피는 **약을 안써도 되는 증상**이더라

**대한한의사협회** 

안아키의 아토피 치료는 한의학·서양의학을 떠나 **의학적 근거가 부족하다**. 논문을 발표해 동료들과 논의한 뒤 **논리적인 치료 체계**를 잡아야 한다

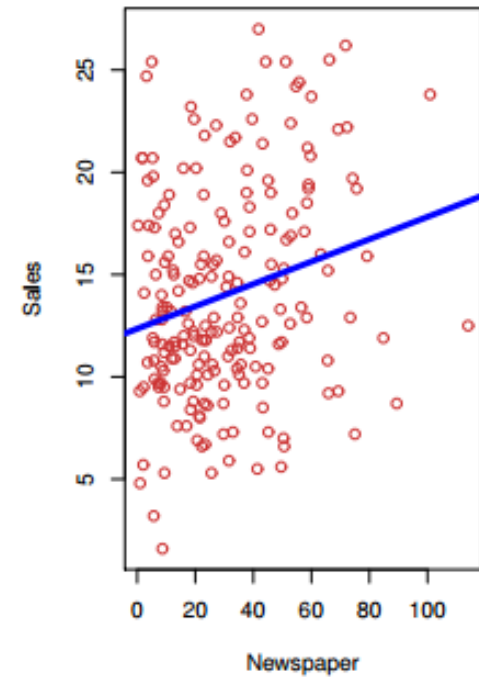
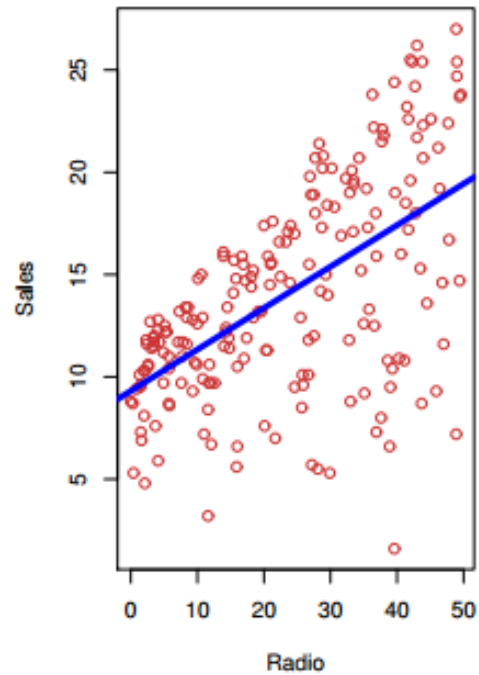
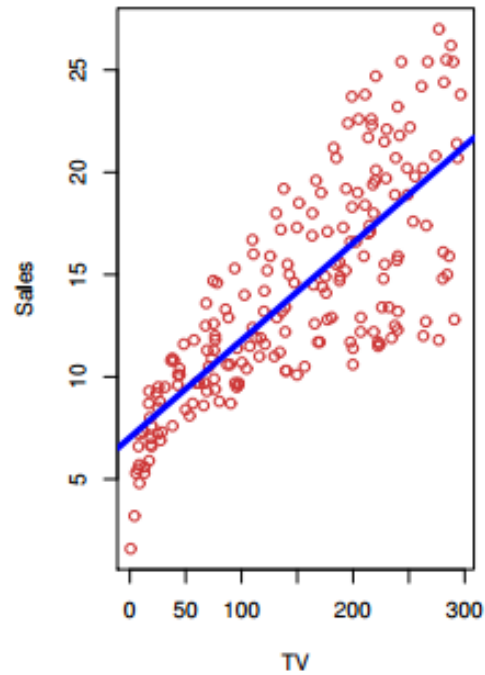
❖ 노출과 결과간의 인과관계 도출 기준(Jeremy Howick)

- 효과가 너무 커서 그럴듯한 혼동요인으로는 설명할 수 없다.
- 적절한 시간적 또는 공간적 근접성이 있다. 원인이 결과에 선행하고 결과는 그럴법한 기간이 지난 후에 발생하며 또한 원인이 결과와 같은 장소에서 발생한다.
- 용량 반응성과 가역성이 있다. 노출이 증가함에 따라 효과가 커진다. 만일 노출이 줄어들때 효과가 감소한다면, 증거는 훨씬 더 강력해 진다.
- 인과 고리를 뒷받침해줄 외적 증거와 함께, 그럴듯한 생물학적/화학적/기계적 매커니즘이 존재한다.
- 그 효과가 이미 알려진 사실과 잘 들어맞는다
- 동일한 효과가 해당 연구를 재현했을 때 발견된다.
- 동일한 효과가 유사 연구에서 발견된다.



$$\text{아들의 키} = 0.33 \times \text{아버지의 키} + 135$$

## ❖ Linear regression



$$\text{Sales} = a + b \cdot \text{TV} + c \cdot \text{Radio} + d \cdot \text{Newspaper} + e$$

## ❖ Linear regression

$$y = \beta_0 + \beta_1 x + \varepsilon$$

intercept    slope    error  
coefficients  
(parameters)

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

prediction value

$$e_i = y_i - \hat{y}_i$$

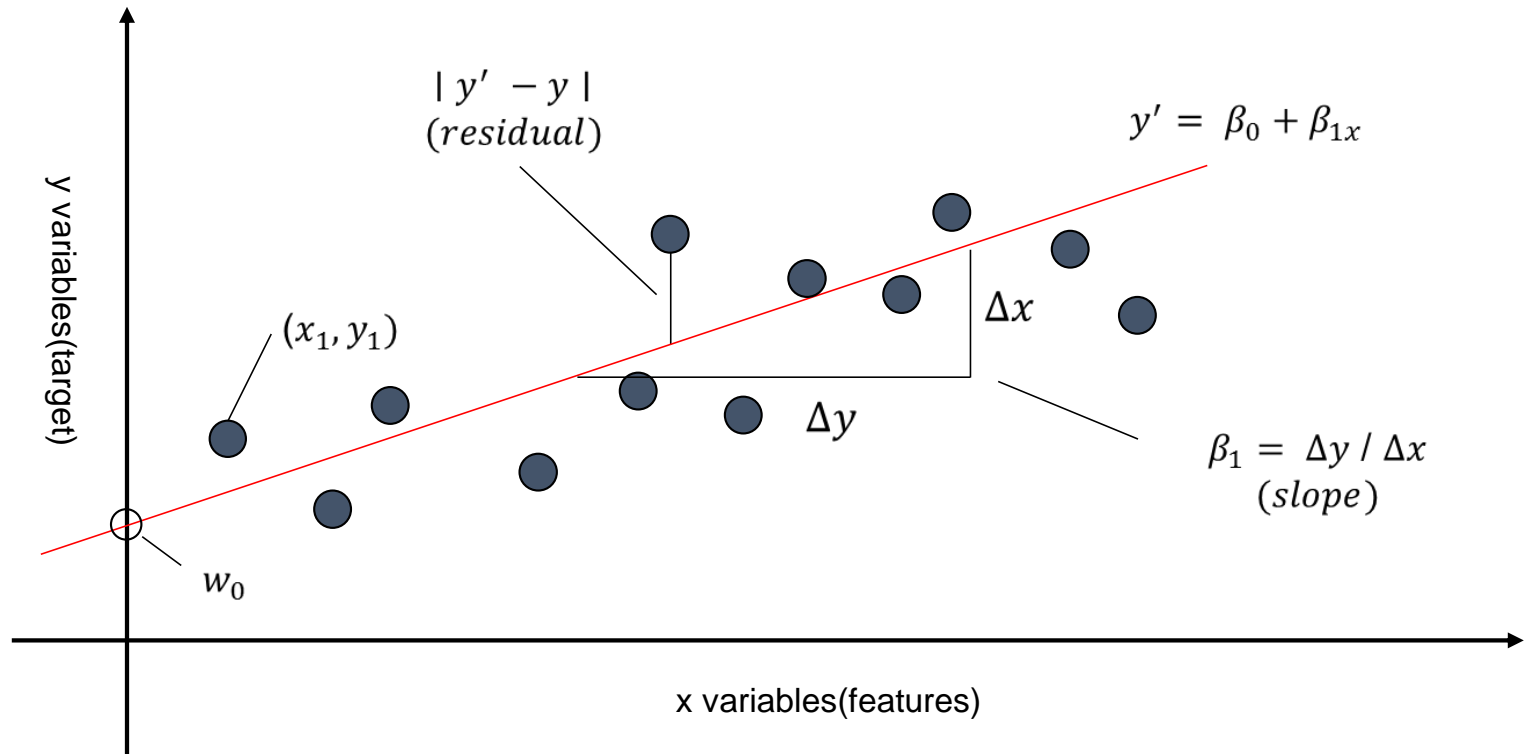
residual

$$RSS = e_1^2 + e_2^2 + \dots + e_n^2$$

residual sum of squares

$$RSS = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \dots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

## ❖ Linear regression





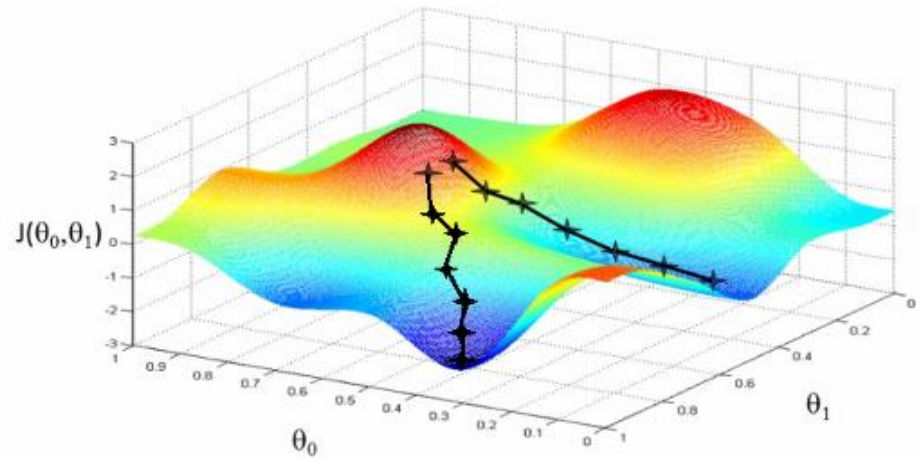
## ❖ Linear regression

❖ OLS(Ordinary least squares) : RSS 최소화 하도록 계수(intercept, coefficient) 추정

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$



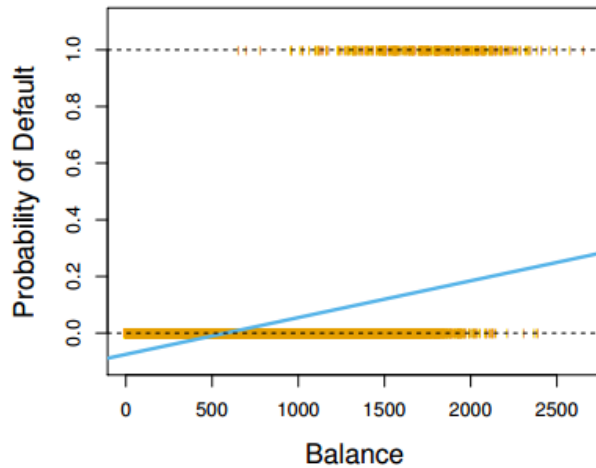
## ❖ Explanatory regression vs. Predictive regression

### ❖ Explanatory regression

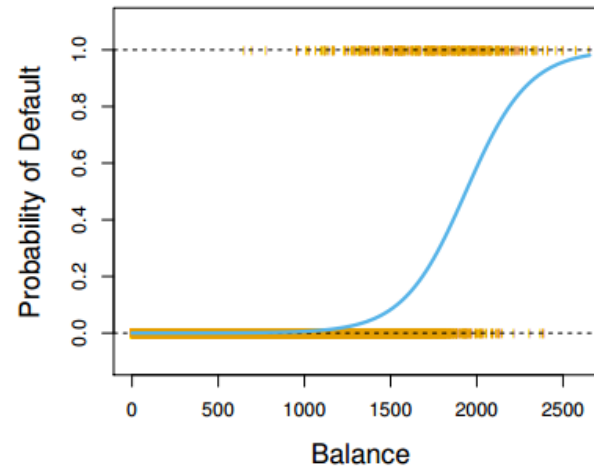
- 타겟 변수와 입력변수들의 관계를 설명
- 타겟 변수(Y)에 입력변수(X)들이 어떠한 영향을 미치는지 살펴보는데 목적
  - ✓ 학습된 회귀모델이 얼마나 많은 정보를 설명하는가?
  - ✓ 각 입력변수들이 통계적으로 출력변수에 유의한 영향을 미치는가?
  - ✓ 이외에 여러 가지 통계 검정을 목적으로 사용
- $\beta_1$  부호의 의미
  - ✓  $\beta_1 > 0$  :  $x_1$  가  $y$  에 긍정적인 영향
  - ✓  $\beta_1 < 0$  :  $x_1$  가  $y$  에 부정적인 영향
- $|\beta_1|$  의 의미
  - ✓  $x_1$  가  $y$  의 값에 공헌한 정도
  - ✓ 예를 들면,  $x_1, x_2$  의 **scale이 값은 경우** 둘다 양의 값이고  $\beta_1 > \beta_2$  이면,  $x_1$ 이  $x_2$  보다  $y$  값의 증가에 더 큰 영향을 준다

## ❖ Logistic regression(Classification)

❖ Linear regression



❖ Logistic regression



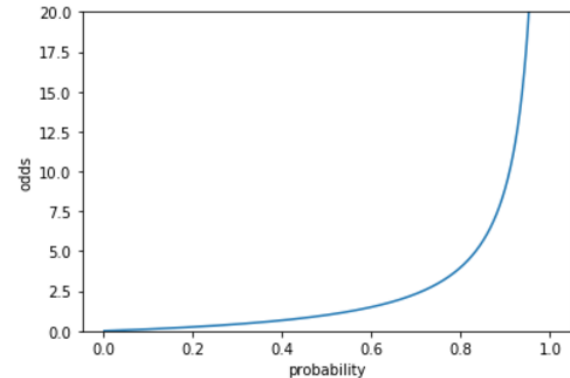
- Linear regression 의 y 값의 범위는  $-\infty \sim \infty$
- 범주형(category, factor, class) 데이터를 y 값으로 사용하면 그 값의 종류가 몇가지로 이루어져 있기 때문에 linear regression 부적합

## ❖ Logistic regression(Classification)

❖ odds : 성공 확률(success probability)이 실패 확률( $1 - \text{성공확률}$ )에 비해 몇 배 더 높은가?

$$\text{odds} = \frac{p}{1 - p}$$

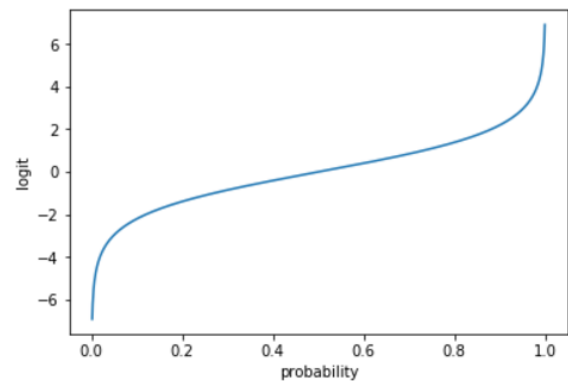
- $0 < \text{odds} < \infty$
- 비대칭



❖ logit : odds에 자연로그(ln)을 취한 값

$$\text{logit} = \ln(\text{odds}) = \ln\left(\frac{p}{1 - p}\right)$$

- $-\infty < \text{logit} < \infty$
- 대칭
- 단순 증가 함수



## ❖ Logistic regression(Classification)

$$y' = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_n x_n + \epsilon$$

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_n x_n$$

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_n x_n)}}$$

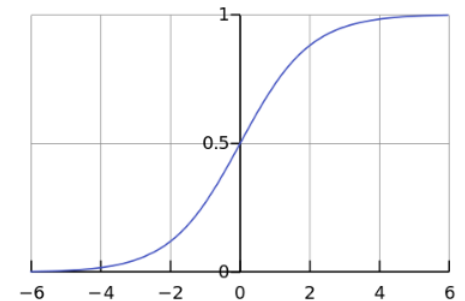
$$p(y == 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_n x_n)}}$$

- Y 값이 두개의 클래스로 이루어진 경우 하나의 데이터가 successive class(y==1)일 확률

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}}$$

$$f(x) = \frac{1}{1 + e^{-x}} \quad (k = 1, x_0 = 0, L = 1)$$

Logistic function



## ❖ Logistic regression(Classification)

❖ Maximum likelihood estimation(MLE) : 최대우도추정

$$p(y = 1|x) = \theta(\beta \cdot X_i)$$

$$Pr(Y_i = y_i | X_i) = p_i^{y_i} (1 - p_i)^{1-y_i} = \theta(y\beta \cdot X_i)$$

$$p(Y|X) = \prod_{i=1}^N Pr(Y_i = y_i | X_i) = \prod_{i=1}^N p_i^{y_i} (1 - p_i)^{1-y_i} \quad \text{최대화}$$

$$\text{Negative Log Likelihood: NLL} = \frac{1}{N} \log(Y|X) = \frac{1}{N} \sum_{i=1}^N \log \frac{1}{Pr(Y_i = y_i | X_i)}$$

최대값을 구하는 문제를 최소값을 구하는 문제로 치환

$$-\frac{1}{N} \sum_{i=1}^N \theta(y_i \beta \cdot X_i) = -\frac{1}{N} \sum_{i=1}^N (y_i \log \frac{1}{\theta(\beta \cdot X_i)} + (1 - y_i) \log \frac{1}{1 - \theta(\beta \cdot X_i)})$$

Y 값이 [0, 1]

## ❖ Logistic regression(Classification)

### ❖ Odds ratio

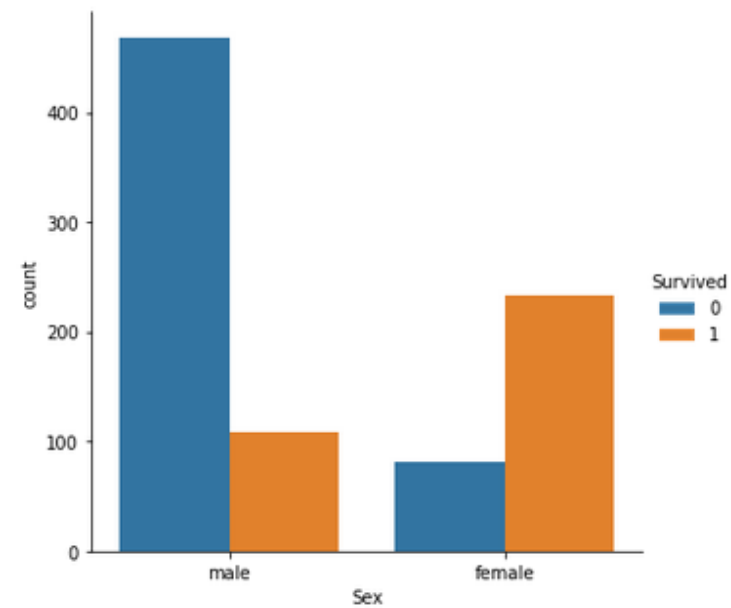
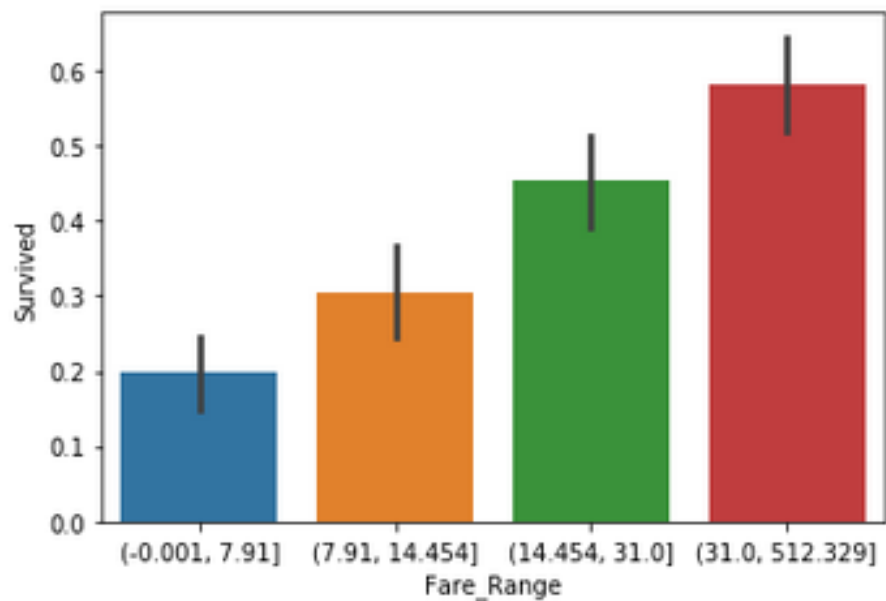
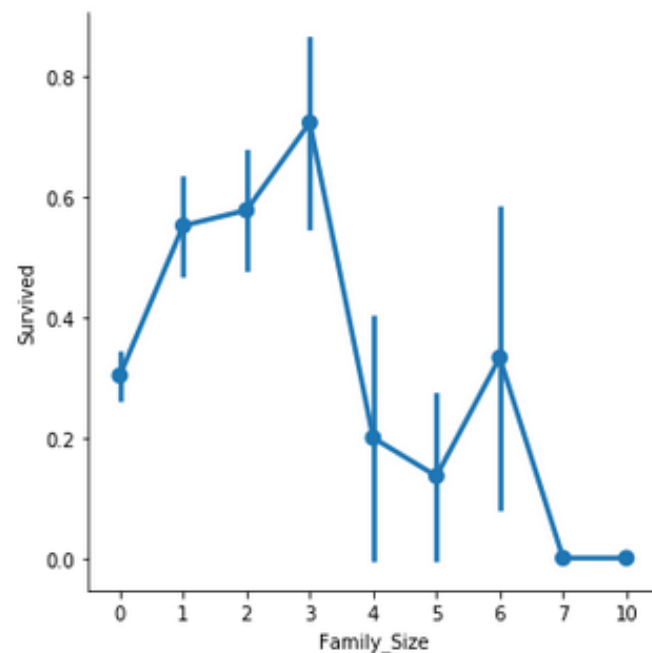
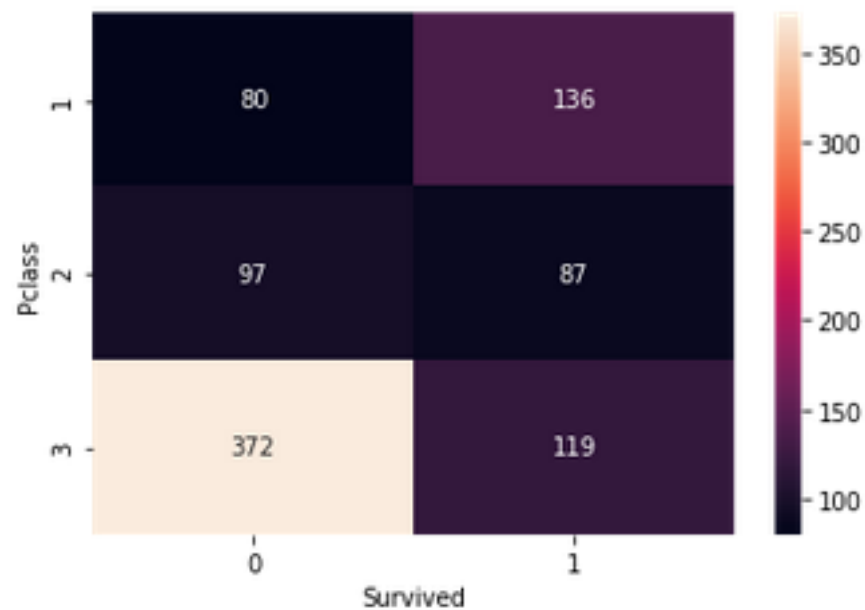
- $x_1$  가 한 단위 증가할 때 odds 의 증가 비율

$$odds\ ratio = \frac{odds(x_1+1 + \dots + x_n)}{odds(x_1 + \dots + x_n)} = \frac{e^{(\beta_0 + \beta_1 (x_1+1) + \beta_2 x_2 + \dots + \beta_n x_n)}}{e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n)}} = e^{\beta_1}$$

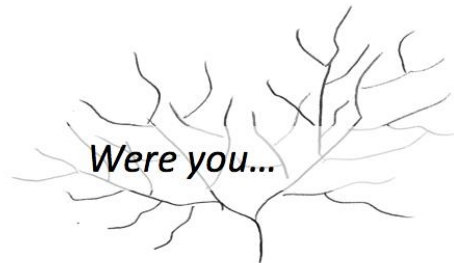
- $x_1$  이 점수,  $y$  가 합격여부라면,  $\beta_1$  이 0.2 일 때, 점수가 1점 올라갈 때 마다 합격에 대한 odds는  $e^{0.2}$  만큼 증가

### ❖ Logistic coefficient ( $\beta_i$ ) : $x_1$ 가 한 단위 증가할 때

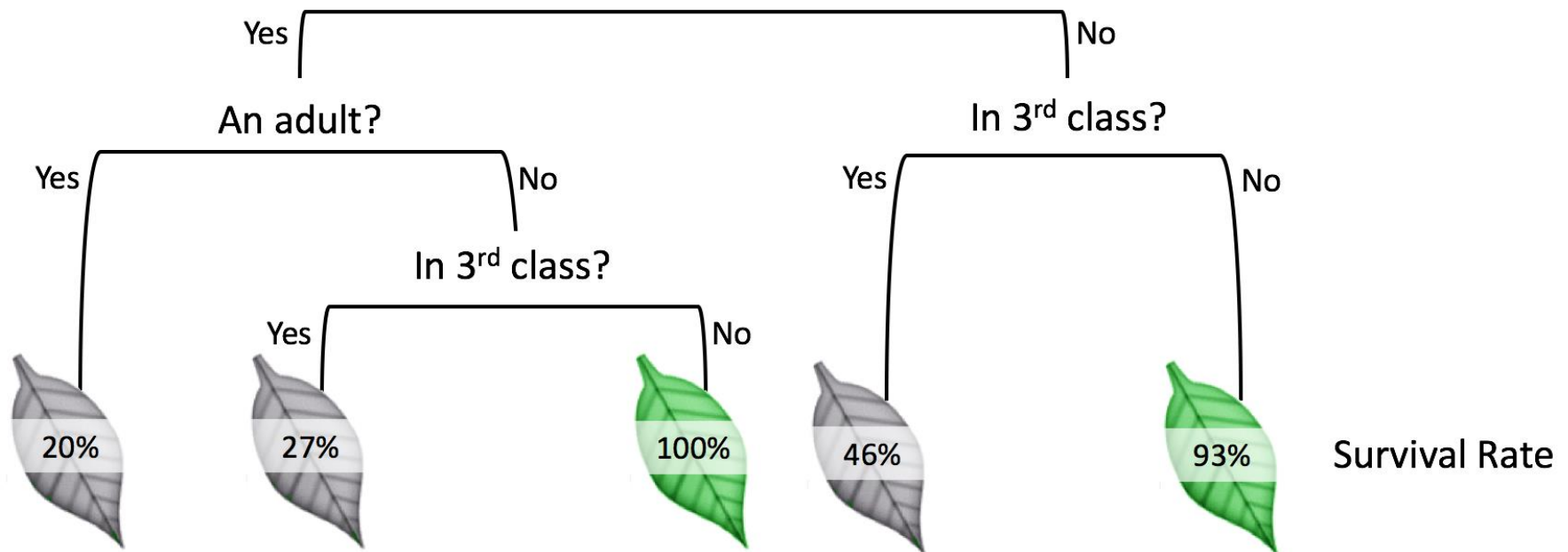
- $\beta_i > 0$  이면,  $e^{\beta_i} > 1$  : odds 증가,  $P(Y=1)$ 이 증가
- $\beta_i = 0$  이면,  $e^{\beta_i} = 1$  : odds 동일,  $P(Y=1)$ 이 동일
- $\beta_i < 0$  이면,  $e^{\beta_i} < 1$  : odds 감소,  $P(Y=1)$ 이 감소



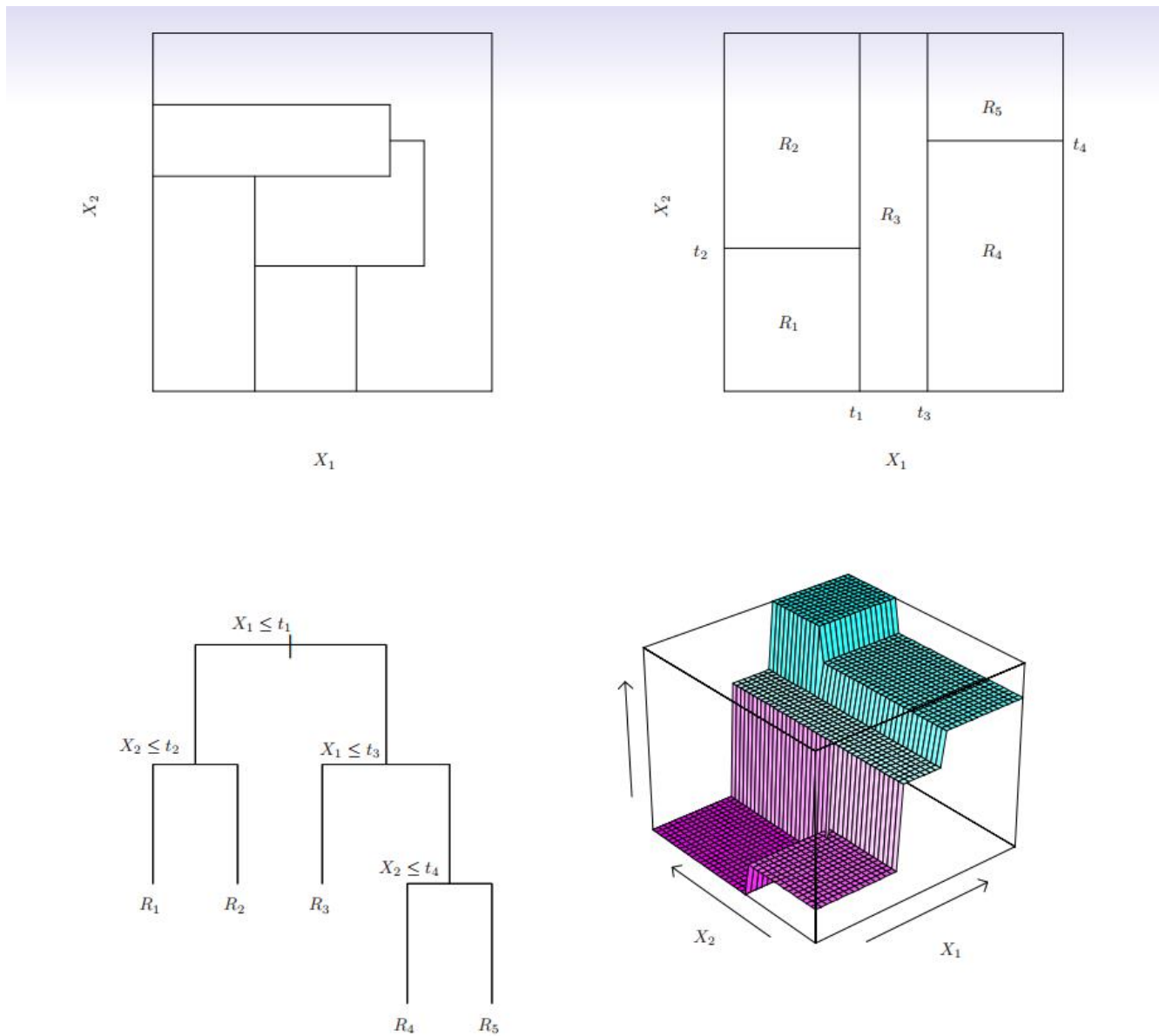




Male?



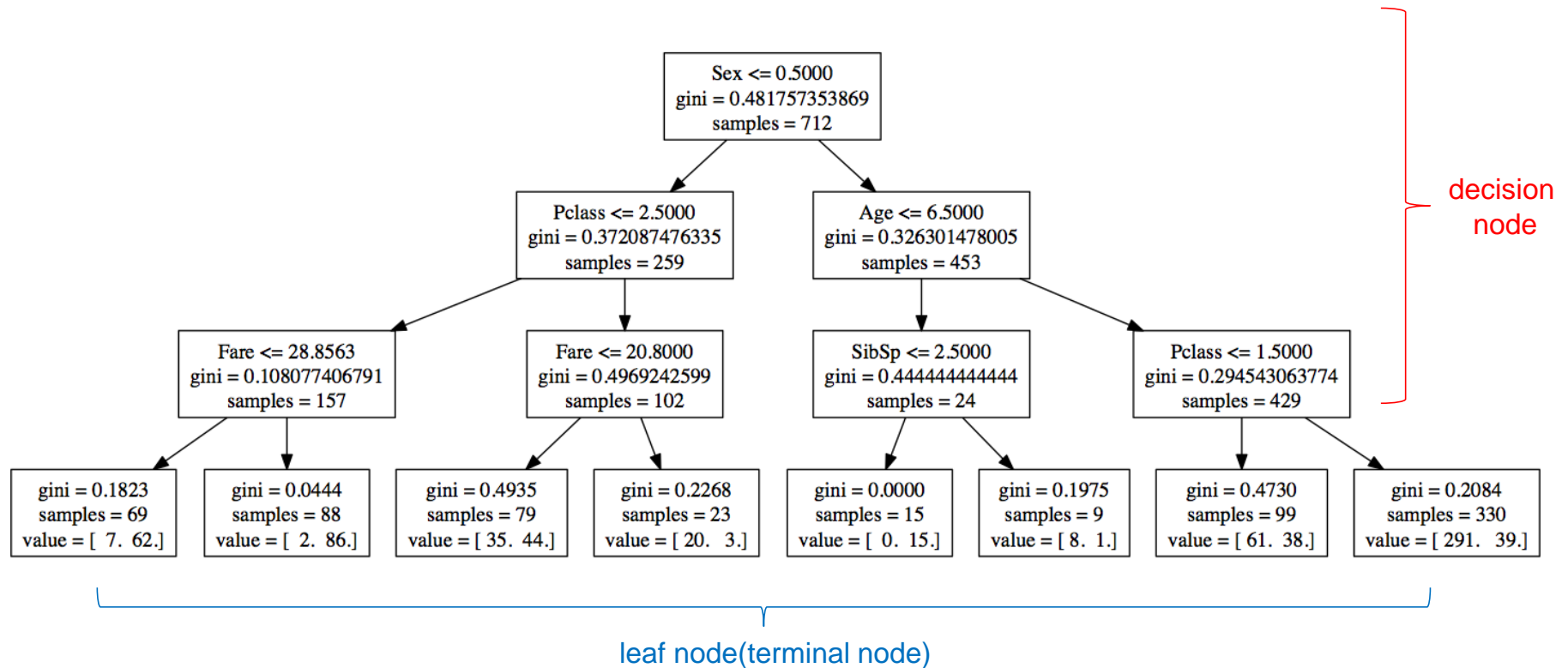
## ❖ Tree-based model



## ❖ Tree-based model

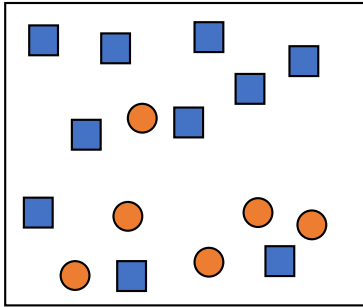
- ❖ 개별 변수의 영역을 반복적으로 분할함으로써 전체 영역에서의 규칙을 생성하는 지도학습 기법
- ❖ If-else-then 형식으로 표현되는 rules을 생성함으로써, 결과에 대한 예측과 함께 그 이유를 설명할 수 있다.
- ❖ Classification, regression 모두 가능
- ❖ 입력 변수들 중 데이터를 가장 잘 분류하는 변수를 택하여 분기가 계속 일어 남
- ❖ 모델에 대한 해석력이 좋다
  - 분기되는 변수는 중요한 변수
  - 분류 규칙 추출 가능
- ❖ 변수 선택이 모델 자체에서 이루어짐
- ❖ Continuous, discrete, category 변수 모두 사용 가능

## ❖ Tree-based model



## ❖ Tree-based model

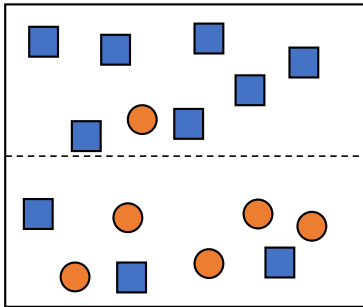
- Gini index



$$I(A) = 1 - \sum_{k=1}^n p_k^2$$

$$I(A) = 1 - \left(\frac{6}{16}\right)^2 - \left(\frac{10}{16}\right)^2 = 0.47$$

- $I(A) = 0$  : 모든 데이터가 같은 클래스
- $I(A) = \text{최대값}$  : 모든 클래스의 데이터가 동일한 숫자



$$I(A) = \sum_{i=1}^d \left( R_i \left( 1 - \sum_{k=1}^n p_k^2 \right) \right)$$

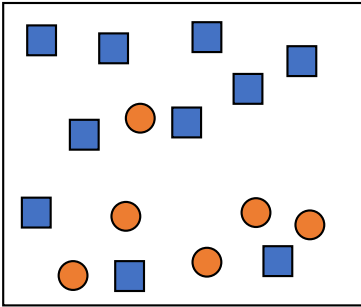
$R_i$  : 클래스의 비율

$$\begin{aligned} I(A) &= 0.5 * \left( 1 - \left(\frac{7}{8}\right)^2 - \left(\frac{1}{8}\right)^2 \right) \\ &\quad + 0.5 * \left( 1 - \left(\frac{3}{8}\right)^2 - \left(\frac{5}{8}\right)^2 \right) \\ &= 0.34 \end{aligned}$$

- information gain : 0.47 - 0.34*

## ❖ Tree-based model

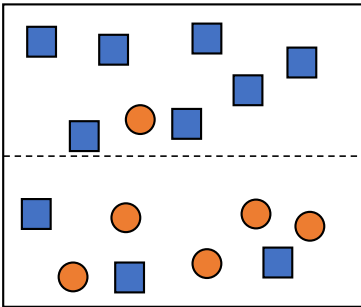
- Entropy



$$I(A) = - \sum_{k=1}^n p_k \log_2(p_k)$$

$$I(A) = 0.95$$

- $0 \leq I(A) \leq \log(n)$



$$I(A) = \sum_{i=1}^d \left( R_i \left( - \sum_{k=1}^n p_k \log_2(p_k) \right) \right)$$

$$I(A) = 0.75$$

$R_i$  : 클래스의 비율

- *information gain : 0.95 - 0.75*

## ❖ Tree-based model

### ❖ Basic algorithm

- i. 전체 데이터를 포함하는 root node 생성
  - ii. 만일 샘플들이 모두 같은 클래스이면 node는 leaf가 되고 해당 클래스로 label 부여
  - iii. 그렇지 않으면 **information gain**이 높은 속성 선택
  - iv. 선택된 속성으로 branch를 만들고 하위 node 생성
  - v. 각 노드에 대하여 ii 부터 반복
- 정지 조건 : 해당 node에 속하는 데이터들이 모두 같은 클래스를 가지거나 상위 node에서 모든 속성을 사용

### ❖ Information gain : 특정 속성을 기준으로 데이터를 구분할 때 감소되는 entropy의 양

$$Gain(S,A)=Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \quad ( S_v = \{s \in S | A(s)=v\} )$$

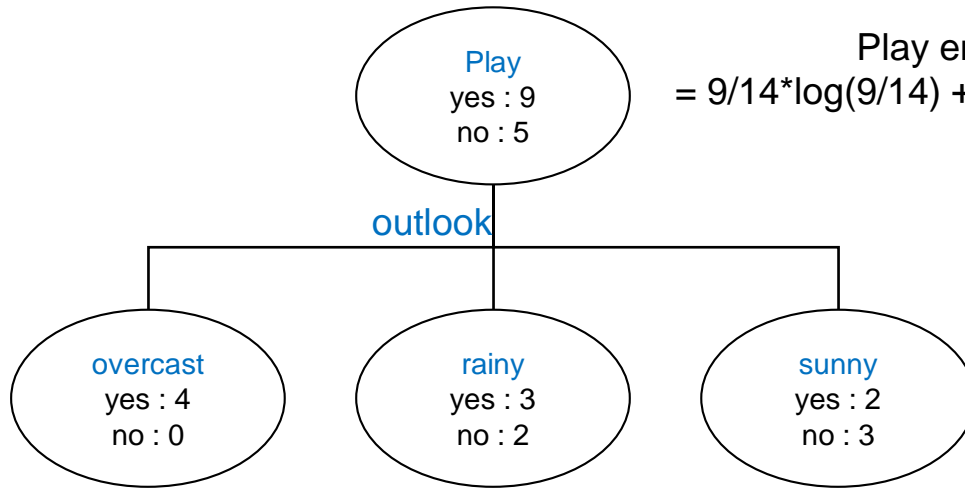
### ❖ Entropy(무질서도) $H(p) = - \sum_{x \in X} p(x) \log p(x)$

## ❖ Tree-based model

Outlook	Temperature Numeric	Temperature Nominal	Humidity Numeric	Humidity Nominal	Windy	Play
overcast	83	hot	86	high	FALSE	yes
overcast	64	cool	65	normal	TRUE	yes
overcast	72	mild	90	high	TRUE	yes
overcast	81	hot	75	normal	FALSE	yes
rainy	70	mild	96	high	FALSE	yes
rainy	68	cool	80	normal	FALSE	yes
rainy	65	cool	70	normal	TRUE	no
rainy	75	mild	80	normal	FALSE	yes
rainy	71	mild	91	high	TRUE	no
sunny	85	hot	85	high	FALSE	no
sunny	80	hot	90	high	TRUE	no
sunny	72	mild	95	high	FALSE	no
sunny	69	cool	70	normal	FALSE	yes
sunny	75	mild	70	normal	TRUE	yes



## ❖ Tree-based model



$$\text{Play entropy}(9,5) = 9/14 * \log(9/14) + 5/14 * \log(5/14) = 0.94$$

$$\begin{aligned} \text{Information gain(outlook)} \\ &= 0.94 - (4/14 * 0 + 5/14 * 0.97 + 5/14 * 0.97) \\ &= 0.23 \end{aligned}$$

$$\begin{aligned} \text{overcast entropy}(4,0) \\ &= 4/4 * \log(4/4) + 0/4 * \log(0/4) = 0 \end{aligned}$$

$$\begin{aligned} \text{sunny entropy}(4,0) \\ &= 2/5 * \log(2/5) + 3/5 * \log(3/5) = 0.97 \end{aligned}$$

$$\begin{aligned} \text{rainy entropy}(4,0) \\ &= 3/5 * \log(3/5) + 2/5 * \log(2/5) = 0.97 \end{aligned}$$

## ❖ Tree-based model

### ❖ 장점

- 사용하기 쉽고 이해하기 쉽다 / 모델에 대한 이해도가 높다
- 변수 선택과 축소가 알고리즘 내에서 이루어 진다
- 통계 모델의 가정이 필요 없다
- 결측치를 정밀하게 다루지 않아도 사용 가능하다

### ❖ 단점

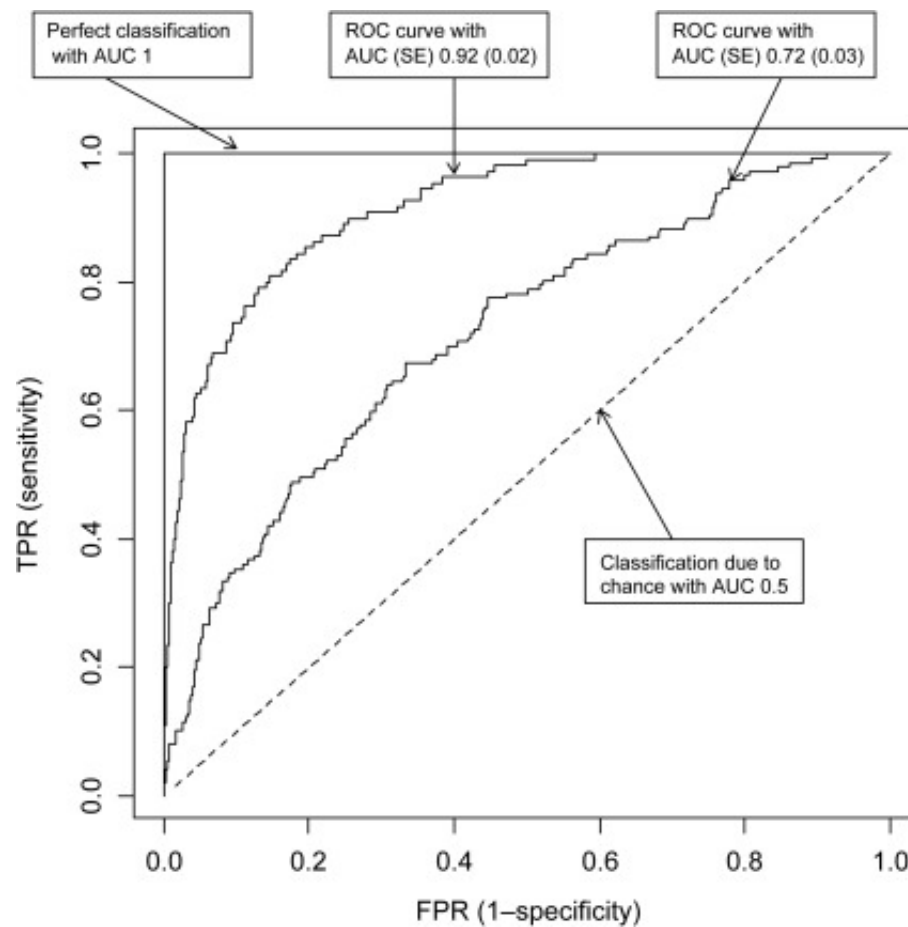
- 많은 데이터들은 수평 또는 수직으로 나누기 어렵다
- 변수 사이의 상호작용을 알아내기 어렵다

## ❖ Confusion matrix

Confusion Matrix		Real			
		Positive	Negative		
Predict	Positive	a	b	Positive Predictive Value (precision)	$a/(a+b)$
	Negative	c	d	Negative Predictive Value	$d/(c+d)$
		Sensitivity (recall)	Specificity	Accuracy $(a+d) / (a+b+c+d)$	
		$a/(a+c)$	$d/(b+d)$		

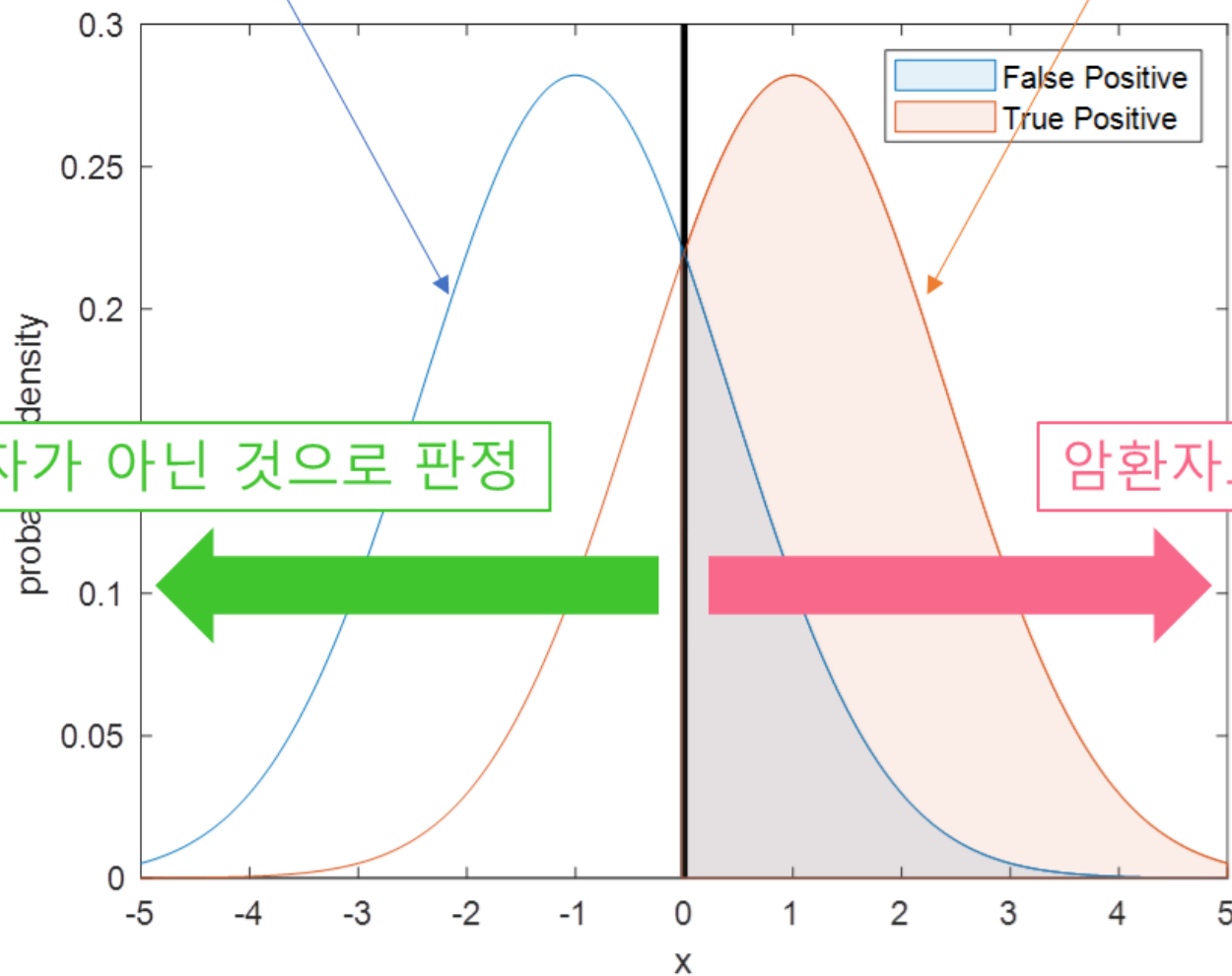
	Training Dataset			Test Dataset		
	사망 예측	생존 예측	합계	사망 예측	생존 예측	합계
실제 사망	475	93	568	228	45	273
실제 생존	71	258	329	35	104	139
합계	546	351	897	263	149	412

	Training Dataset	Test Dataset
Accuracy(정확도)	$(475+258) / 897 = 0.82$	$(228+104)/412 = 0.81$
Sensitivity(민감도)	$258/329 = 0.78$	$104/139 = 0.75$
Specificity(특이도)	$475/568 = 0.84$	$228/273 = 0.84$



암에 걸리지 않은 사람들

이미 암에 걸린 사람들

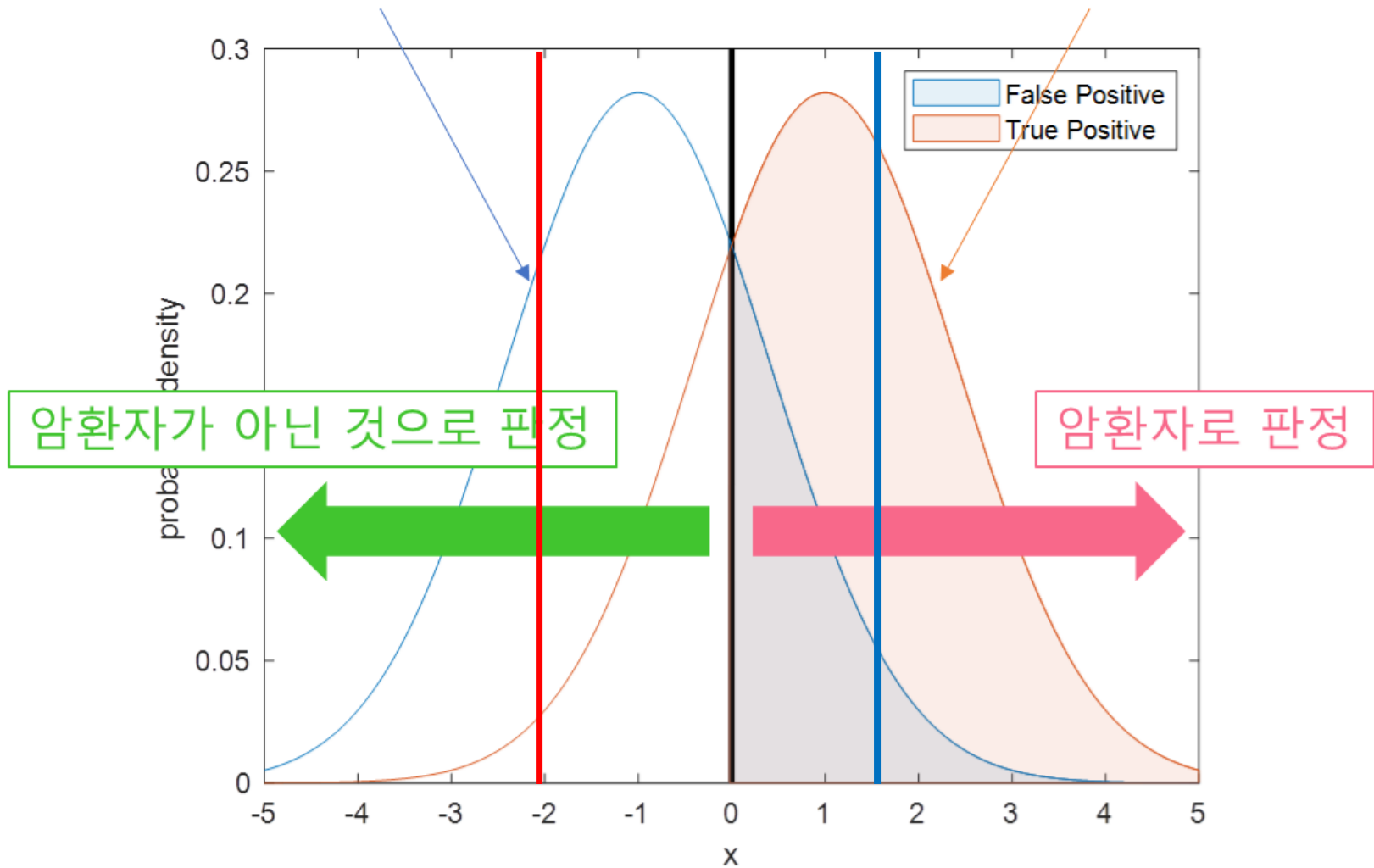


암환자가 아닌 것으로 판정

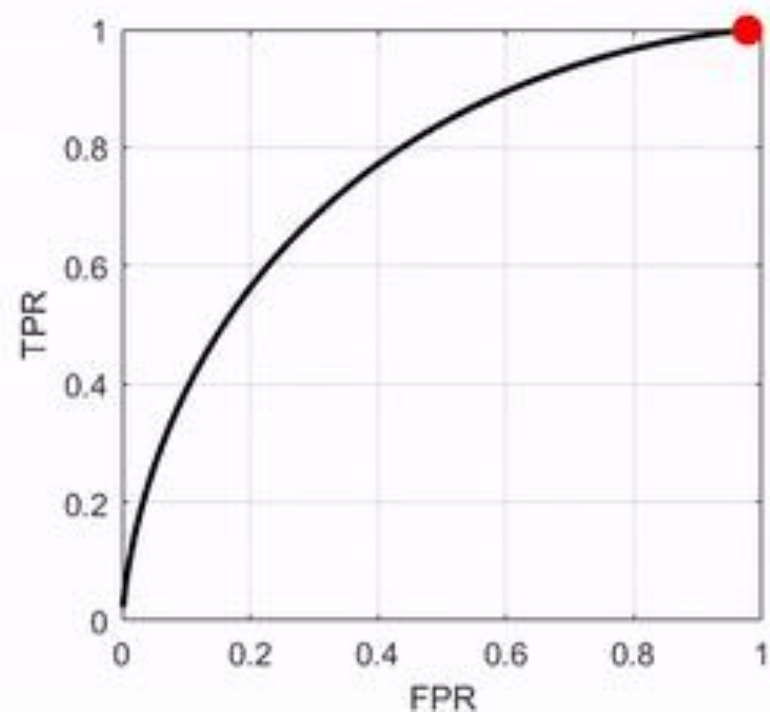
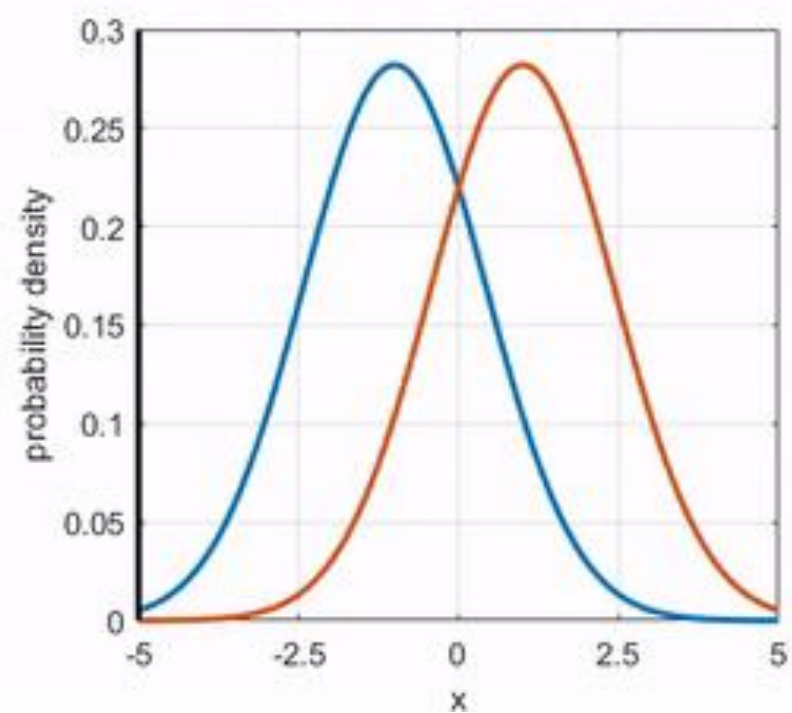
암환자로 판정

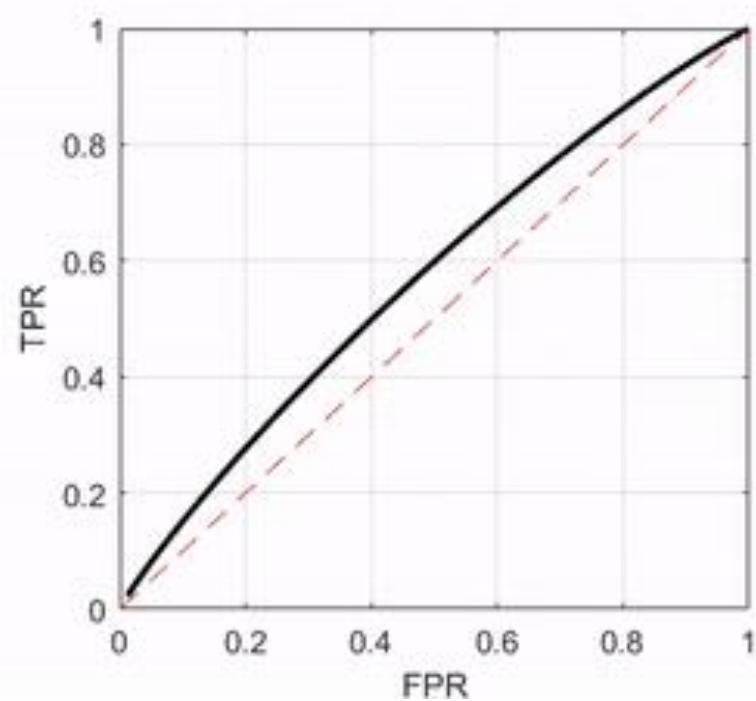
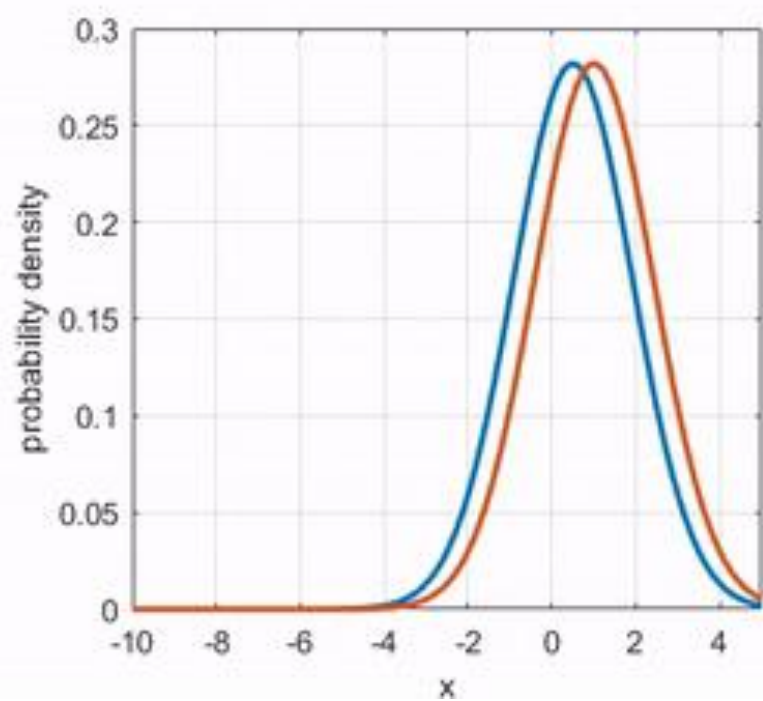
암에 걸리지 않은 사람들

이미 암에 걸린 사람들



판정기준 이동

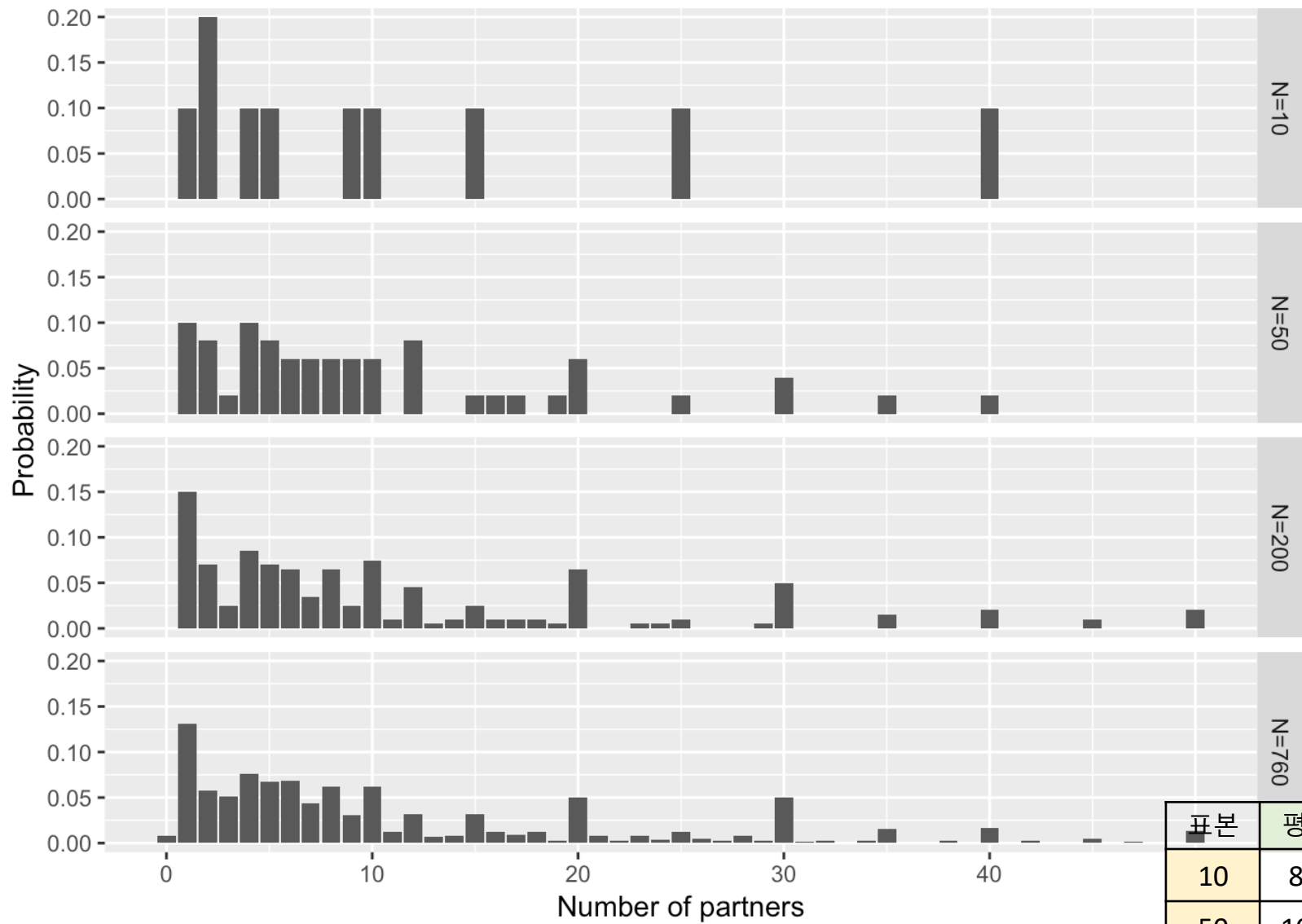




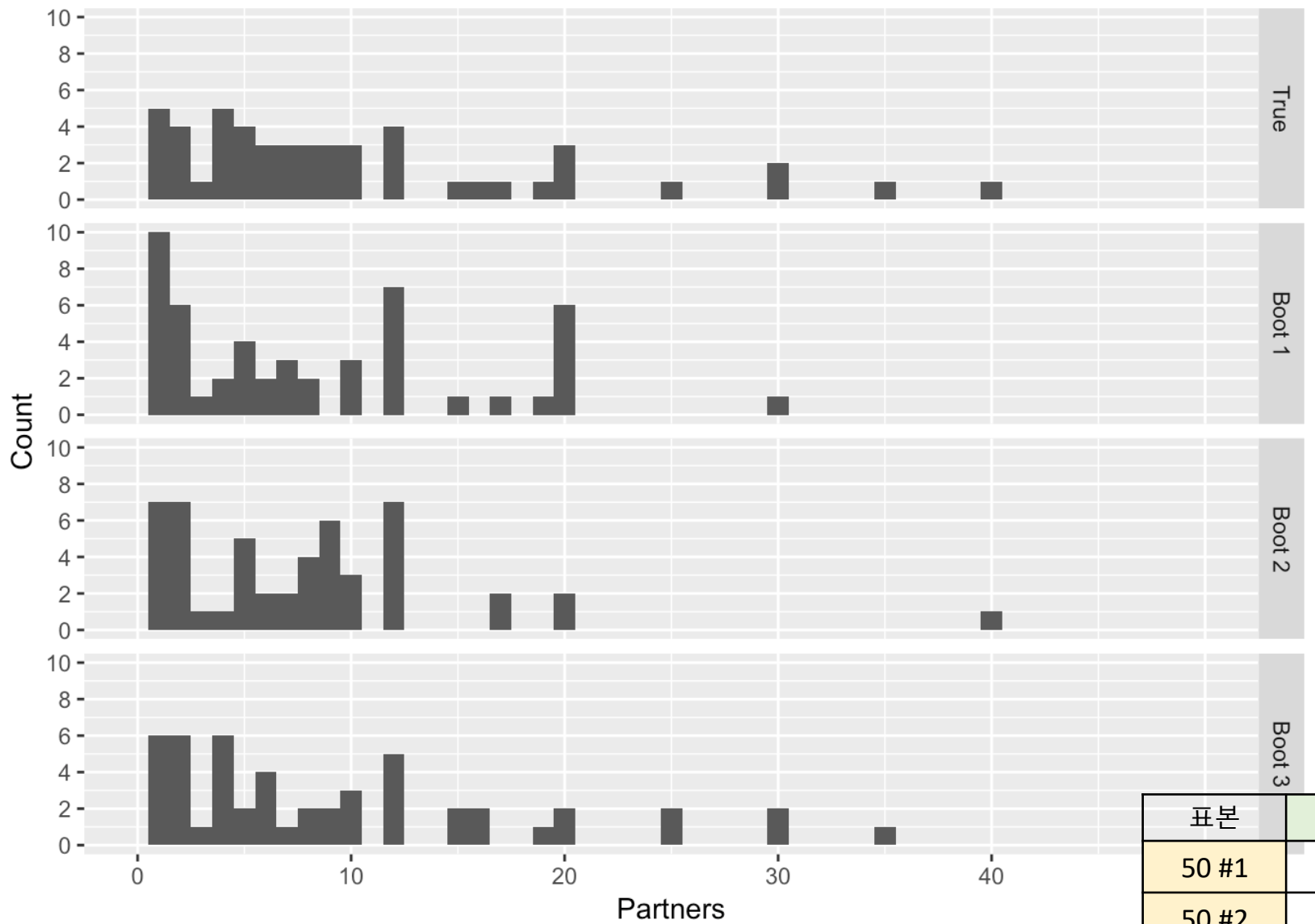


# 강원도철원군

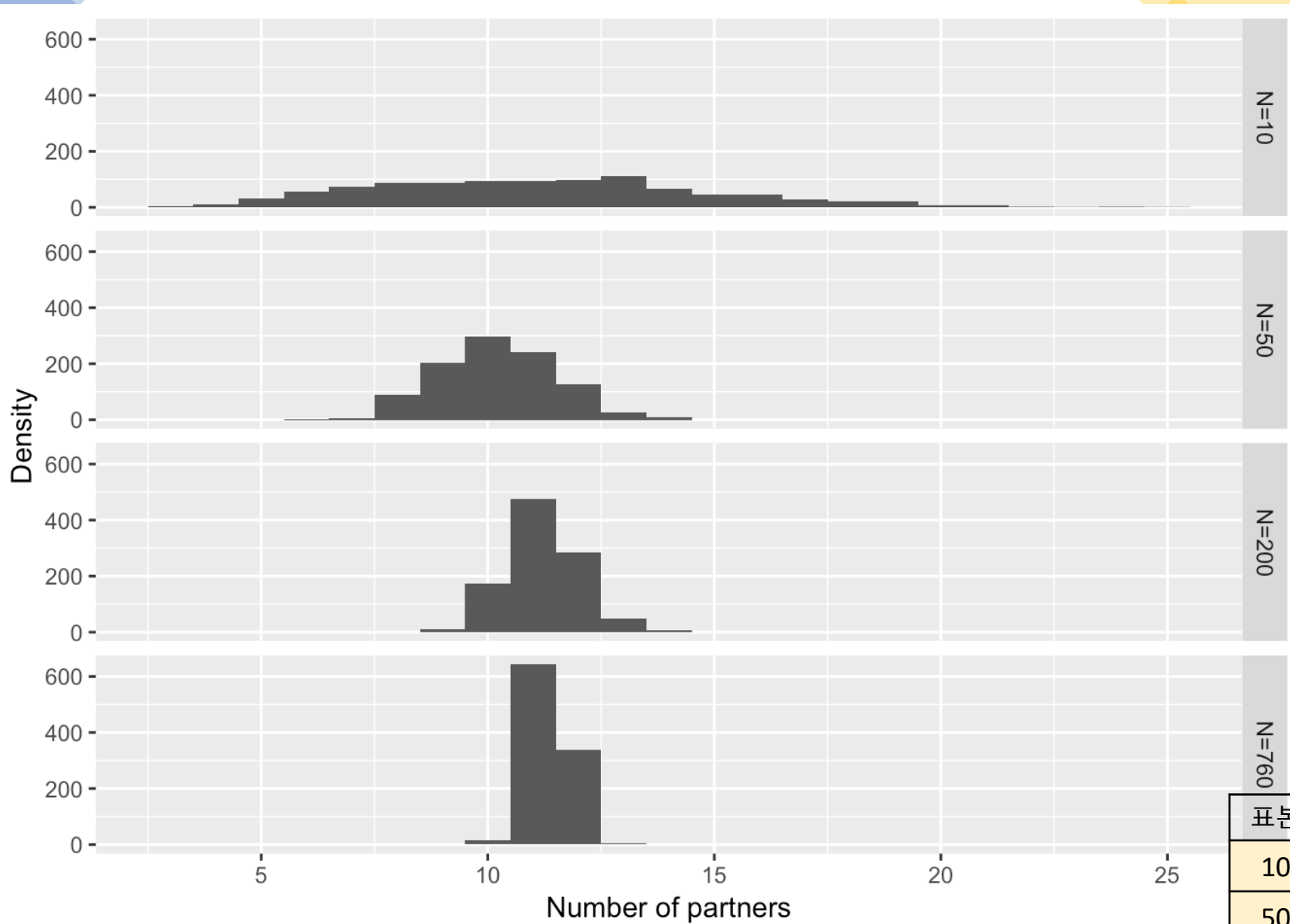
오늘		내일	주간	
날짜	날씨 오전 오후	최고 (℃)	최저 (℃)	강수확률 (%)
8/4(화)	 	27	23	80
8/5(수)	 	26	23	80
8/6(목)	 	26	24	80
8/7(금)	 	28	22	80
8/8(토)	 	27	22	80
8/9(일)	 	27	23	80
8/10(월)	 	27	23	80
8/11(화)	 	28	23	80
8/12(수)	 	29	23	80
8/13(목)	 	29	23	80
8/14(금)		31	25	60
8/15(토)		29	25	60
8/16(일)		30	24	60
8/17(월)		30	24	60



표본	평균	중앙값
10	8.3	9
50	10.5	7.5
200	12.2	8
760	11.4	7



표본	평균
50 #1	10.5
50 #2	8.4
50 #3	9.7
50 #4	9.8



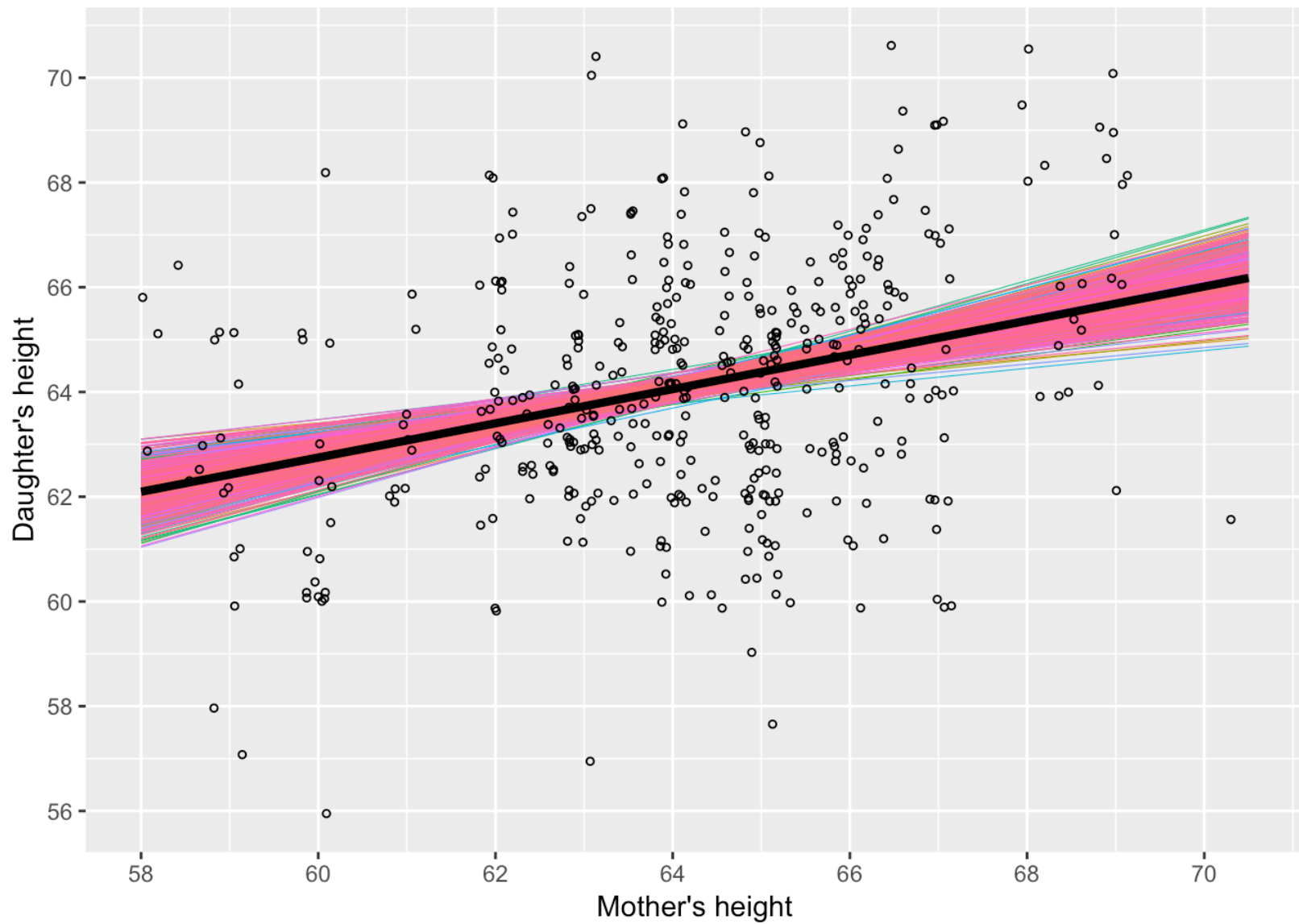
❖ Bootstrap : 1, 000번

표본	평균
10	8.3
50	10.5
200	12.2
760	11.4

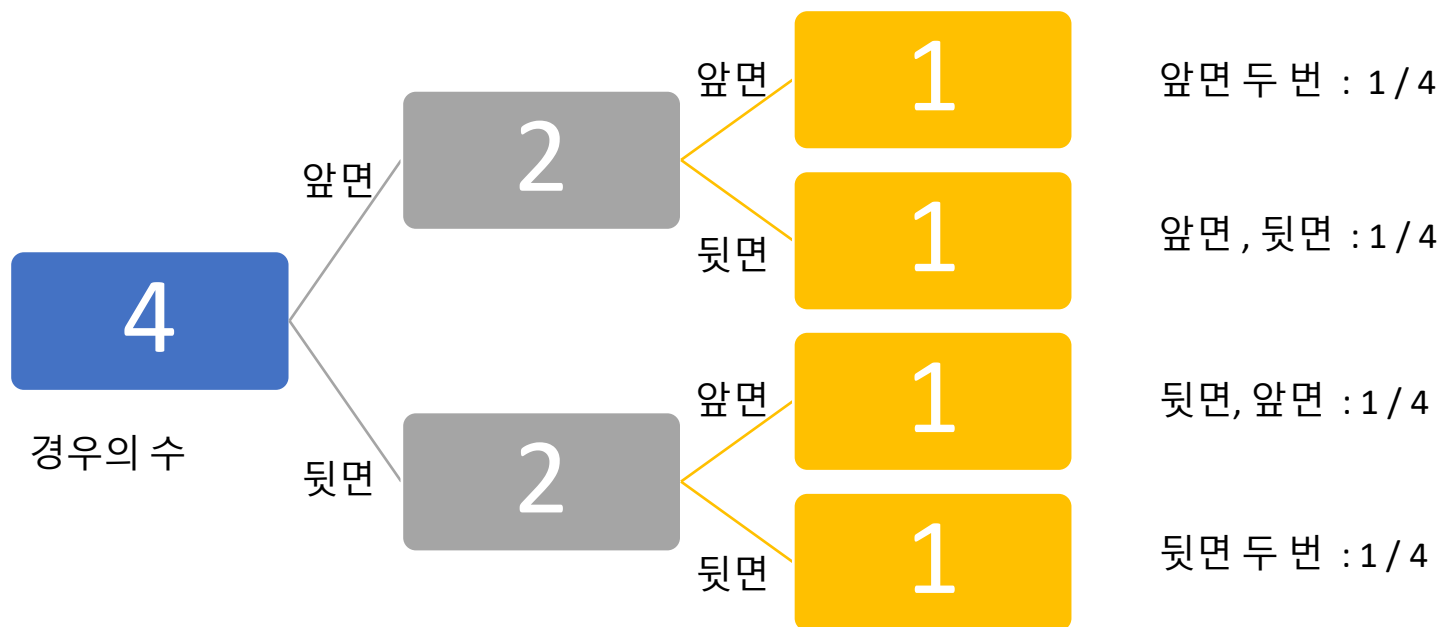
- ❖ 표본의 평균들은 거의 대칭적인 분포를 이룬다.(중심극한정리)
- ❖ 표본의 크기가 증가함에 따라 평균의 분포가 점점 더 좁아진다.

#### ❖ 중심극한정리

- 표본의 크기가 증가함에 따라 원래 데이터 분포의 모양이 어떠하든 거의 상관없이 표본 평균들의 분포가 정규분포의 형태로 다가가는 경향



❖ 동전을 두 번 던졌을 때, 두 번 다 앞면이 나올 확률은?



- 한사건의 확률은  $0 \sim 1$ : 앞면도 뒷면도 안나올 확률  $0$ , 4가지 중 하나가 나올 확률  $1$
- 어떤 사건이 일어날 확률은  $1 - \text{그 사건이 일어나지 않을 확률}$ , 모두 앞면이 나올 확률은  $1 - (\frac{1}{4} + \frac{1}{4} + \frac{1}{4}) = \frac{1}{4}$
- 배반사건의 전체 확률은 각 확률의 합, 적어도 한번은 앞면이 나올 확률은  $\frac{1}{4} + \frac{1}{4} + \frac{1}{4} = \frac{3}{4}$
- 독립사건들이 일어날 전체 확률은 각 확률의 곱, 앞면이 두 번 나올 확률은  $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$

❖ 게임 1 : 최대 4번까지 주사위를 한 개 던지는 데, 6이 나오면 이긴다.

❖ 게임 2 : 최대 24번까지 주사위를 두개 던지는 데 , 둘 다 6이 나오면 이긴다.



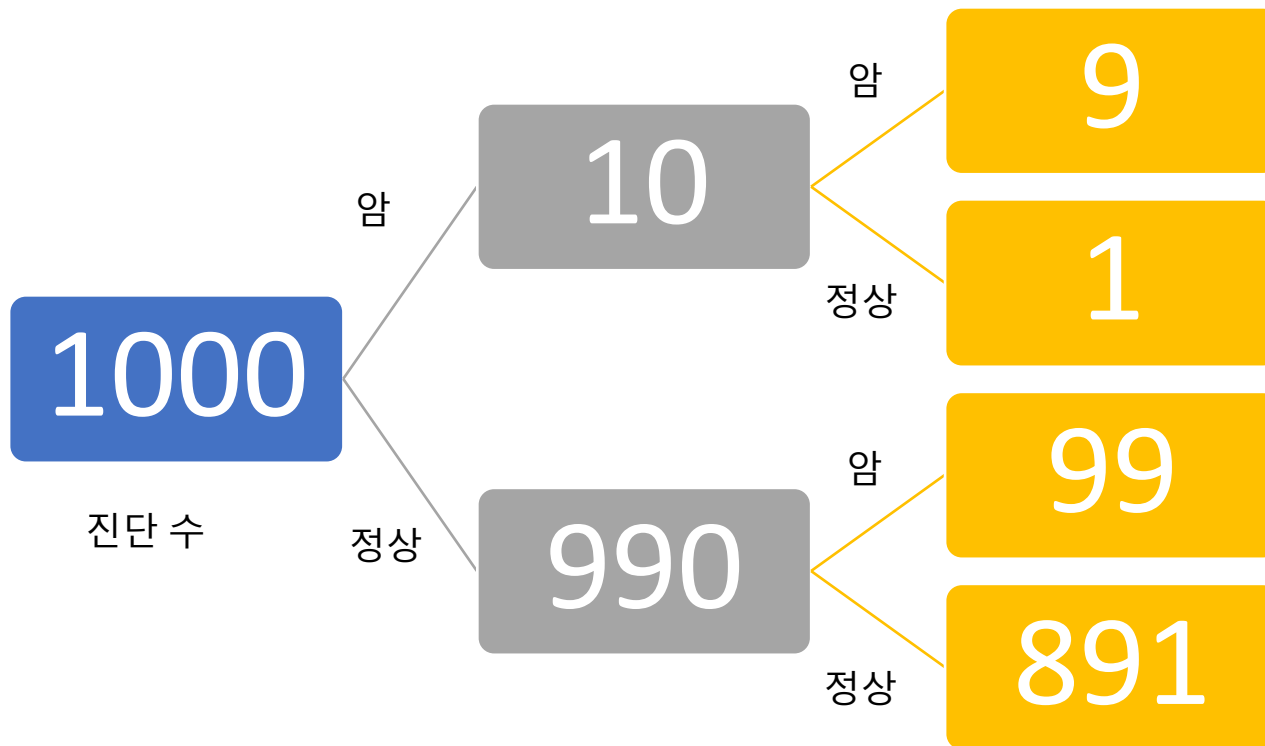
❖ 게임 1 : 최대 4번까지 주사위를 한 개 던지는 데, 6이 나오면 이긴다.

- $1/6 + 5/6 \times 1/6 + 5/6 \times 5/6 \times 1/6 + 5/6 \times 5/6 \times 5/6 \times 1/6$  (이길 확률)
- $1 - 5/6 \times 5/6 \times 5/6 \times 5/6$  (질 확률)

❖ 게임 2 : 최대 24번까지 주사위를 두개 던지는 데 , 둘 다 6이 나오면 이긴다.

- $1 - 35/36^{24}$  (질 확률)

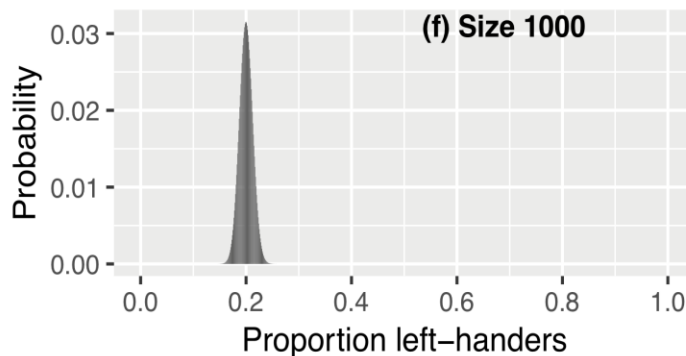
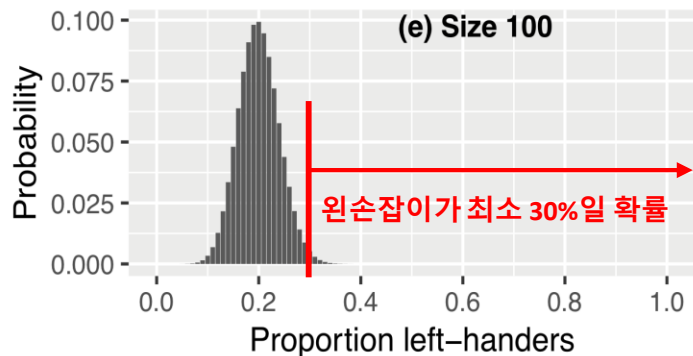
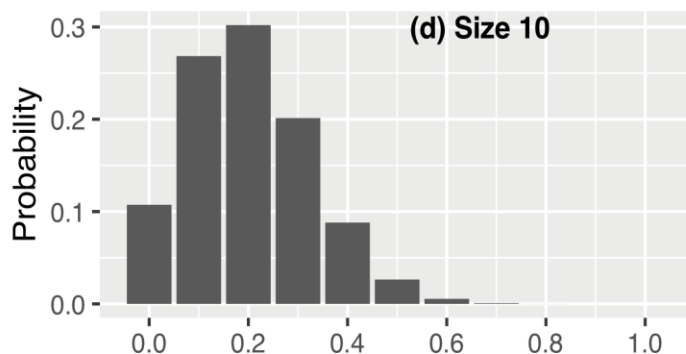
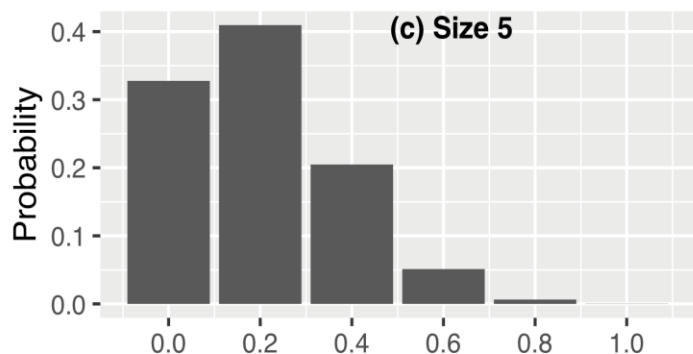
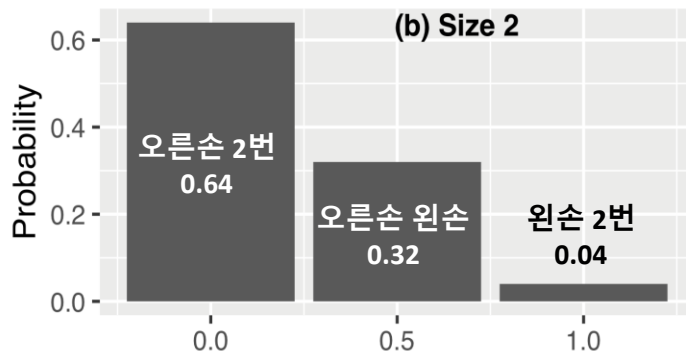
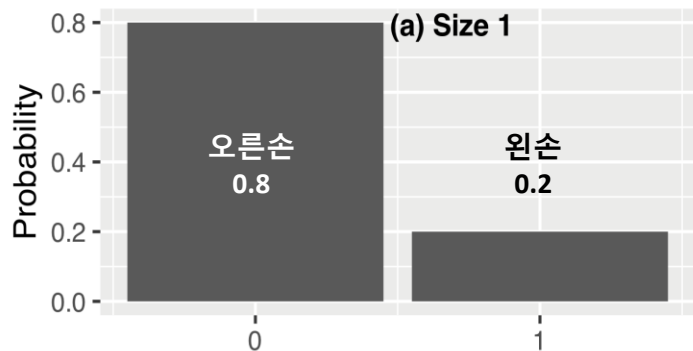
❖ 90%의 정확성을 가진 암 진단 모형이 있다. 검사를 받은 사람의 1%가 실제 암이 있다고 했을 경우, 한 사람의 진단이 양성일 나올 확률은?



암으로 진단될 확률 :  $108 / 1000 = 11 \%$

실제 암일 확률 :  $9 / 108 = 8\%$

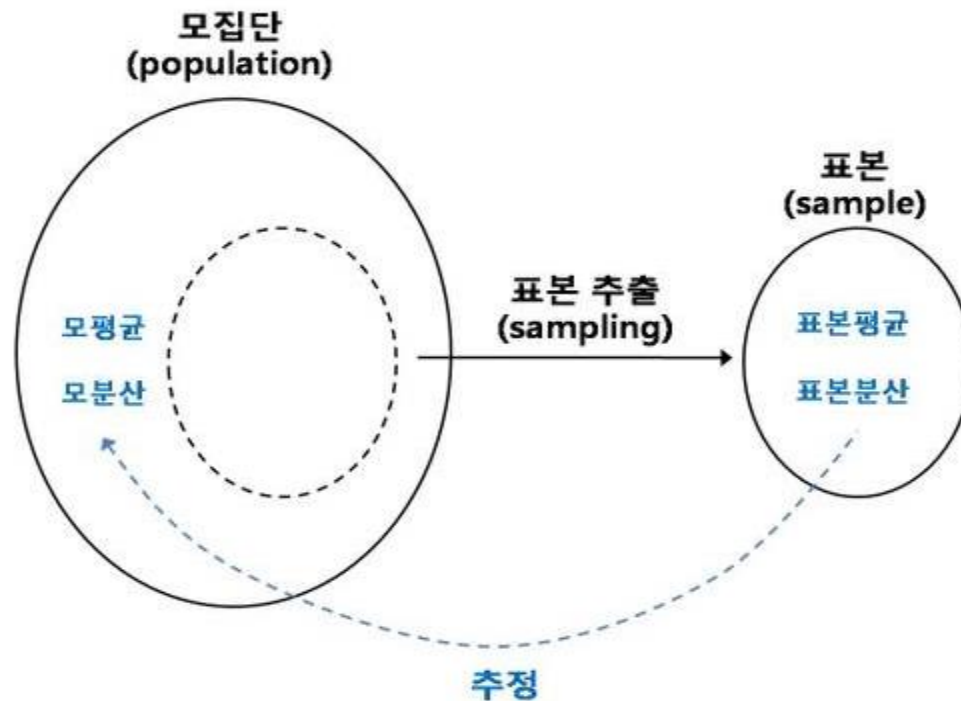
❖ 왼손잡이 20%, 오른손 잡이 80%가 포함된 모집단에서 표본 추출할 때의 확률 분포



- 확률분포 평균 (기댓값) = 0.2
- 부트스트랩의 특성과 동일

이항분포(binomial distribution)

- ❖ 모집단 : 나타날 수 있는 모든 경우의 수
- ❖ 유한모집단 : 나타날 수 있는 경우의 수가 정해진 모집단, 선거 유권자 등
- ❖ 무한모집단 : 무한하게 나타날 수 있는 모집단, 사람 키 등
- ❖ 통계적추정 : 모집단 중에서 나오는 몇 가지 데이터를 가지고 모집단 전체에 대해 어떤 추측을 하는 일(부분으로 전체를 추론)

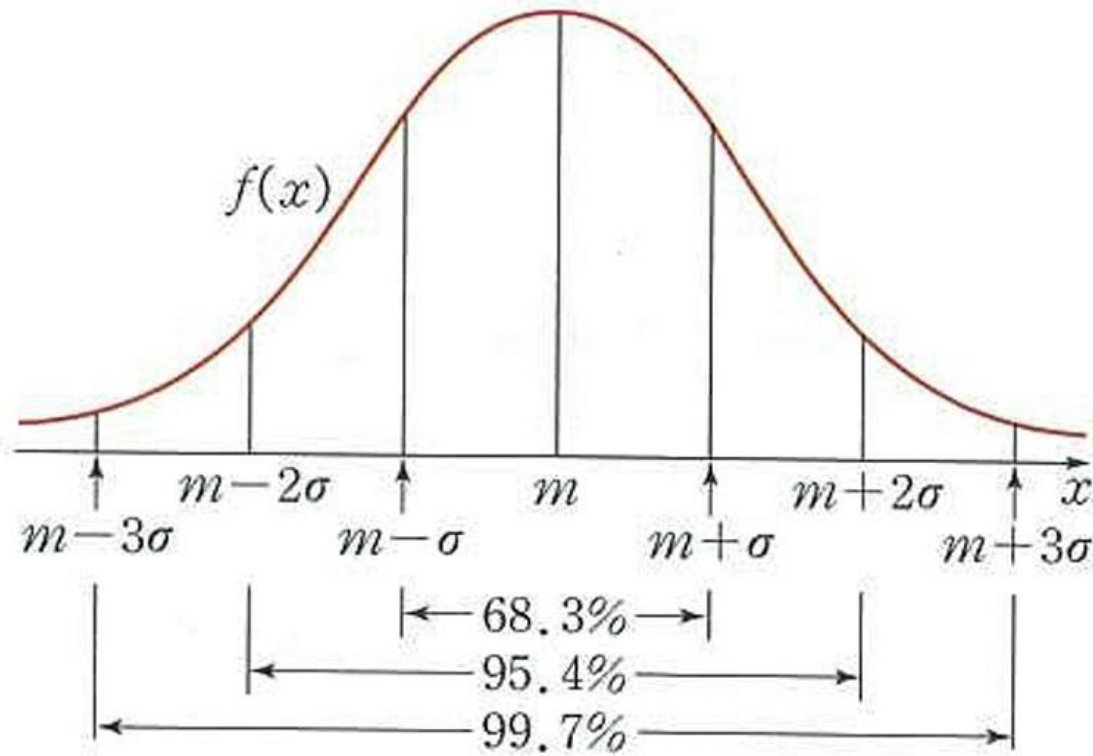


- ❖ 모집단에서 충분할 정도로 많이 반복해서 데이터 관측을 실행해 히스토그램을 작성하면, 이 히스토그램은 거의 모집단과 일치한다.

❖ 표본을 통해 모집단 설명(통계적 추론)

1. 임의의 모집단 모수에 대하여 95% 확률로 관측 통계량이 그 안에 놓여 있기를 기대하는 구간을 확률을 이용해 판단한다.
2. 통계량을 관측한다.
3. 통계량이 95% 예측 구간 안에 놓일 수 있는 모수의 범위를 산출한다. 이 범위를 95% 신뢰구간(confidence interval)이라고 부른다.
4. 반복 적용할 때 그런 구간들의 95%가 참값을 포함해야 하기 때문에, 결과로 나오는 이 신뢰구간을 '95%'라고 한다.

\* 95% 신뢰구간은 이 특정한 구간이 참값을 포함할 확률이 95%라는 뜻이 아님.



❖ 정규 분포 = ( 표준편차 x 정규분포 데이터 ) + 평균

- 95% 데이터 : (평균 - 1.96 x 표준편차) <= 데이터 <= ( 평균 + 1.96 x 표준편차)  
 $-1.96 <= (\text{데이터} - \text{평균}) / \text{표준편차} <= 1.96$

❖ 동전 16개를 던질 때, 앞면이 나오는 개수의 평균은 8, 표준편차는 2라고 할 때, 여러분이 동전을 던질 경우 앞면이 나올 거라고 예측 할 수 있는 개수는?

❖ 동전 N개를 던질 때, 앞면이 10개가 나올 수 있는 N의 범위는?

$$\text{평균} : \mu = \frac{N}{2}, \text{표준편차} : \sigma = \frac{\sqrt{N}}{2}$$

$$(\text{평균} - 1.96 \times \text{표준편차}) \leq \text{데이터} \leq (\text{평균} + 1.96 \times \text{표준편차})$$

$$\mu - 1.96\sigma \leq 10 \leq \mu + 1.96\sigma$$

$$-1.96 \leq \frac{10 - \mu}{\sigma} \leq 1.96$$

$$-1.96 \leq z \leq 1.96$$

❖ 가설 검정

모수 N에 대하여 N개의 동전을 던져서 앞면이 나오는 개수의 데이터를 모집단으로 하면, 이것은

정규분포를 따르며 평균값은  $\mu = \frac{N}{2}$ , 표준편차  $\sigma = \frac{\sqrt{N}}{2}$  이다. 이때, z를  $\frac{10 - \mu}{\sigma}$ 로 계산하여 부등식

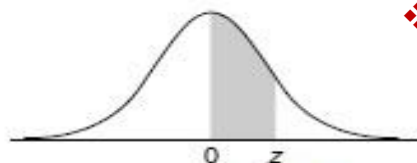
$-1.96 \leq z \leq 1.96$ 이 성립하는 N은 기각하지 않는다(채택한다). 그러나 성립하지 않는 N은 기각한다.

앞면이 나온 개수가 10개로 관측될 때, 모두 N이 95%의 확률로  $13 \leq N \leq 30$ 의 범위에 들어간다(?)

Table AIV.2 Standard Norms Table

Area between 0 and  $z$ 

❖ 표준정규분포표

 $P(0 < Z < 1.55)$ 

$z$	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998

❖ 신뢰구간

구간 추정을 계속 반복하면,  
관측값에 대응하는 여러 구간을  
구할 수 있지만, 그 100번중  
95번은  $N$ 이 구해지는 구간에  
들어간다.



## Muriel Bristol & Ronald Fisher



## ❖ 실험 계획법

- 우유를 먼저 넣은 밀크 티 4잔, 홍차를 먼저 넣은 밀크 티 4잔을 랜덤 추출하여 제공

차 선택	경우의 수	확률
모두 맞힌 경우	$1 \times 1 = 1$	$1 / 70 = 1.4\%$
분리된 4잔에서 1잔만 틀린 경우	$4 \times 4 = 16$	$16 / 70 = 22.8\%$
분리된 4잔에서 2잔이 틀린 경우	$6 \times 6 = 36$	$36 / 70 = 51.4\%$
분리된 4잔에서 3잔이 틀린 경우	$4 \times 4 = 16$	$16 / 70 = 22.8\%$
모두 틀린 경우	$1 \times 1 = 1$	$1 / 70 = 1.4\%$
전체	70	

## ❖ 가설 검정(Hypothesis test)

- 이 클래스의 평균키는 175cm가 아니다.

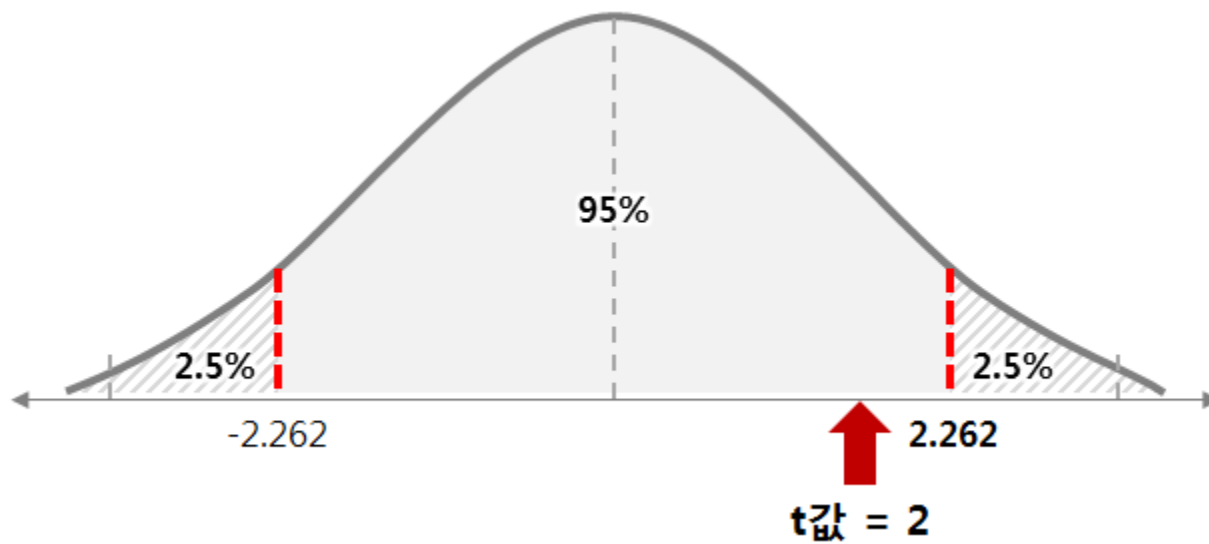
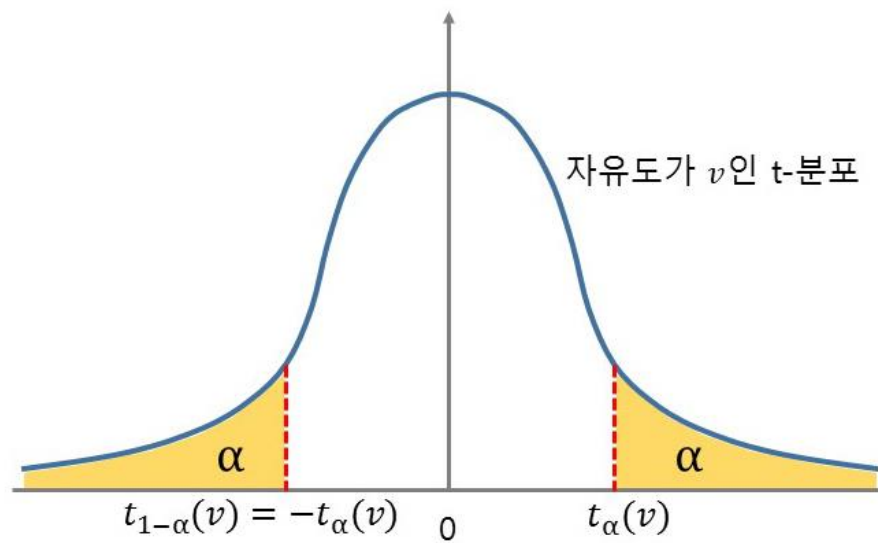
이 클래스의 평균키는 175cm 이다.	귀무가설(Null Hypothesis) 수립
10명의 샘플링 했더니 평균 185cm, 표준편차 5cm 이다. 검정 통계량 = (표본평균 - 모평균) / 표본편차 $t = (185 - 175) / 5 = 2$	샘플링 및 검정 통계량 산출
t 분포 산출(확인)	검정 통계량의 분포 생성(확인)
유의 수준 : 95%	유의수준 설정
p-value : 0.025	p-value(귀무 가설이 참일 경우 극단적인 통계량을 관측할 확률) 계산
귀무가설 채택 이 클래스의 평균키는 175이다.	통계적 유의성 확인(귀무가설 기각/채택)

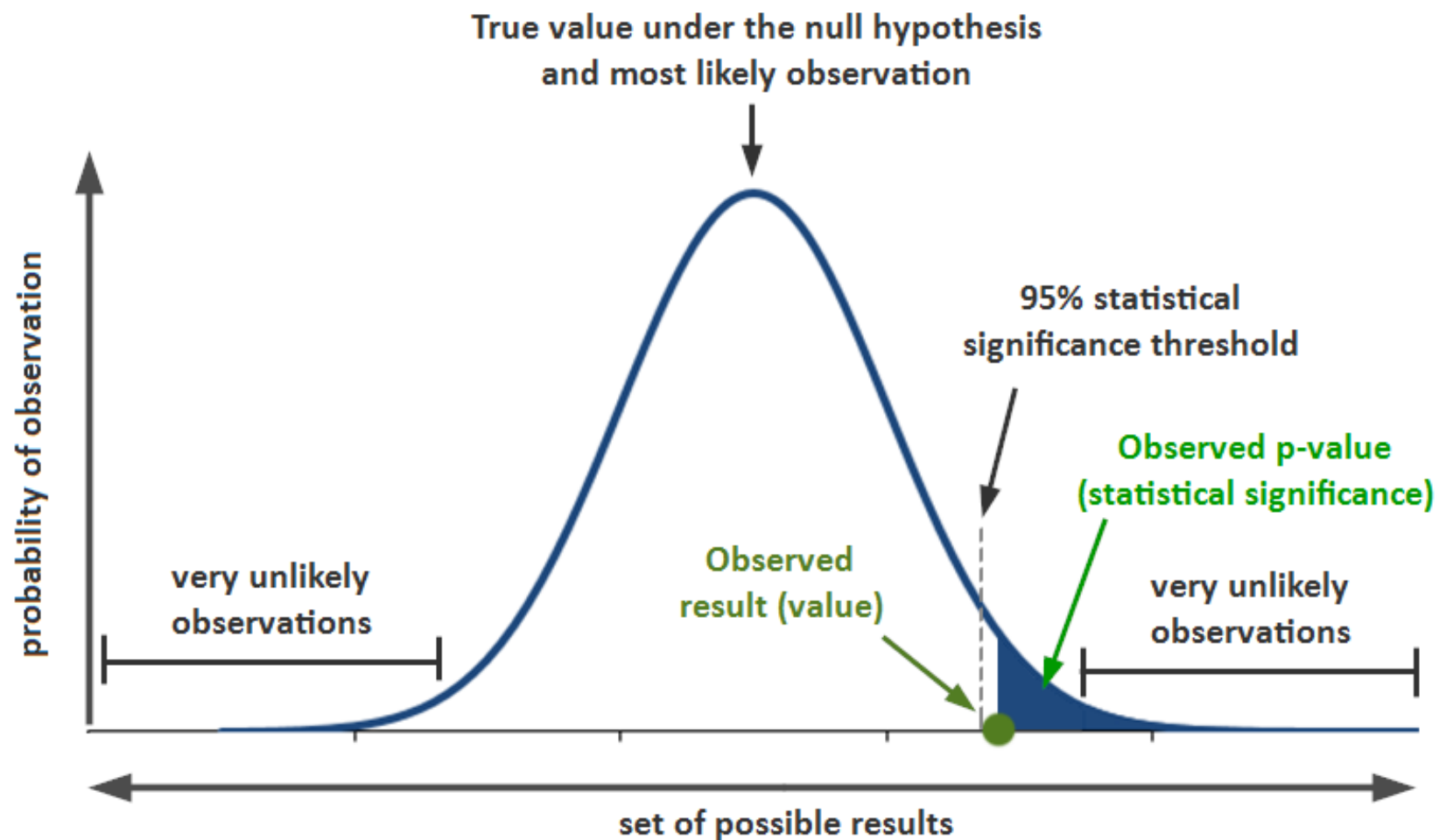
확률

자유도

$\alpha$ df	0.4	0.25	0.1	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	127.32	318.31	636.62
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.599
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.260	0.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.259	0.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.258	0.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.257	0.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.257	0.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.256	0.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.256	0.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.768
24	0.256	0.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.256	0.684	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	0.256	0.684	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	0.256	0.684	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	0.256	0.683	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	0.256	0.683	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	0.255	0.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
60	0.254	0.679	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
120	0.254	0.677	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
$\infty$	0.253	0.674	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

t value





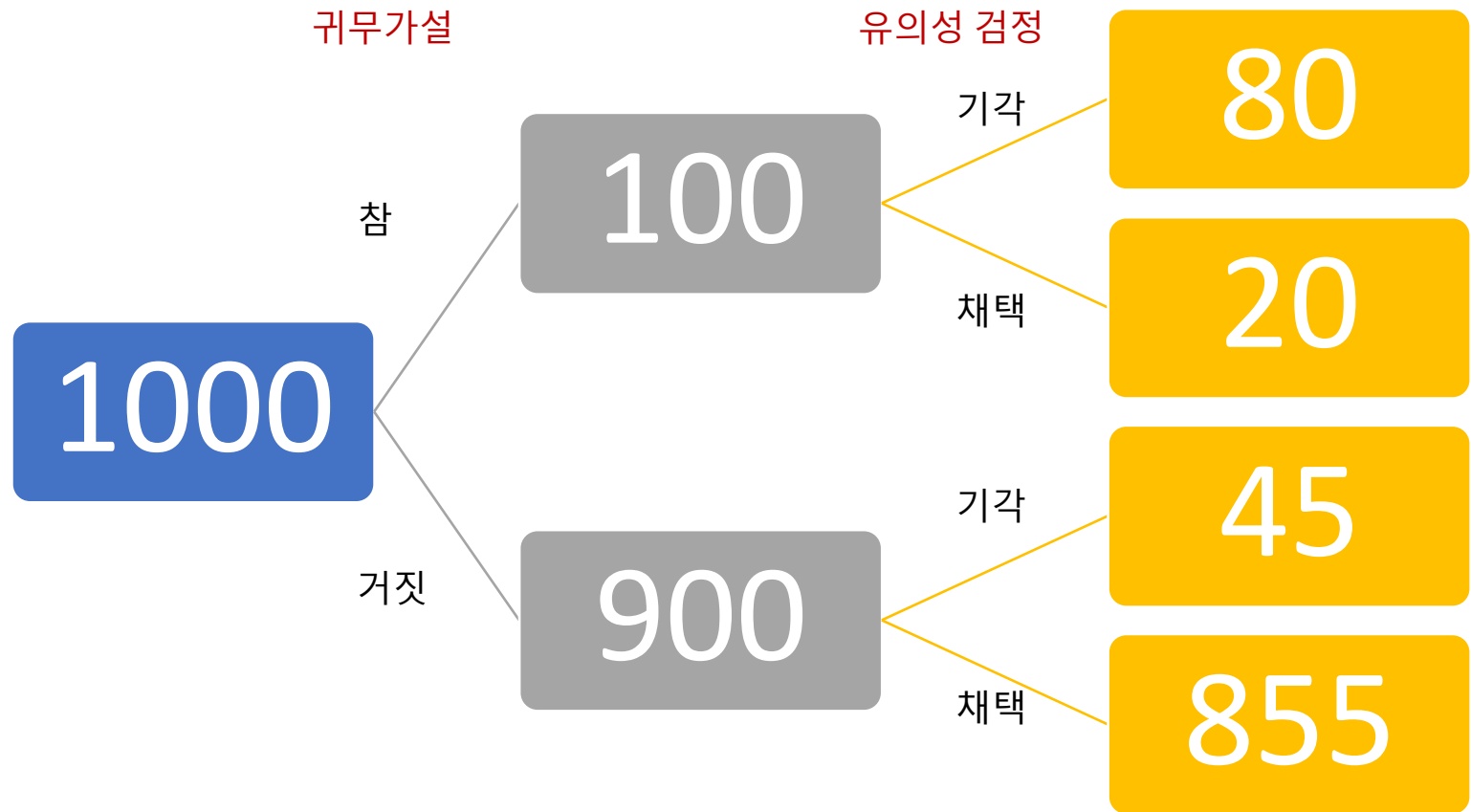
## ❖ 자식의 키와 부모의 키와의 상관 관계

	coef	std err	t	P> t
아버지의 키	0.41175	0.04668	8.820	<2e-16 ***
어머니의 키	0.33355	0.04600	7.7252	1.17e-12 ***
const	69.22882	0.10664	649.168	<2e-16 ***

$$\text{자녀의 키} = 0.41175 \times \text{아버지의 키} + 0.33355 \times \text{어머니의 키} + 69.22882$$

- 귀무 가설 : 계수(coef)들이 0이다(아버지의 키와 어머니의 키는 자녀의 키에 영향을 주지 않는다).
- 절편(const) : 자녀의 평균 키
- 계수(coef) : 부모의 키가 1 변할 때마다 기대되는 아들키의 변화
- t-statistics(t 통계량) : 독립변수와 설명변수간 연관성이 통계적으로 유의미한지 나타내는 지표( coef / std err), 추정 값이 0에서 얼마나 멀리 떨어져 있는지를 표준오차의 몇 배만큼 떨어져 있는지 로 측정한 것

## ❖ P-value의 위험성



125 개의 검증중에 45개는 False-Positive(36%)

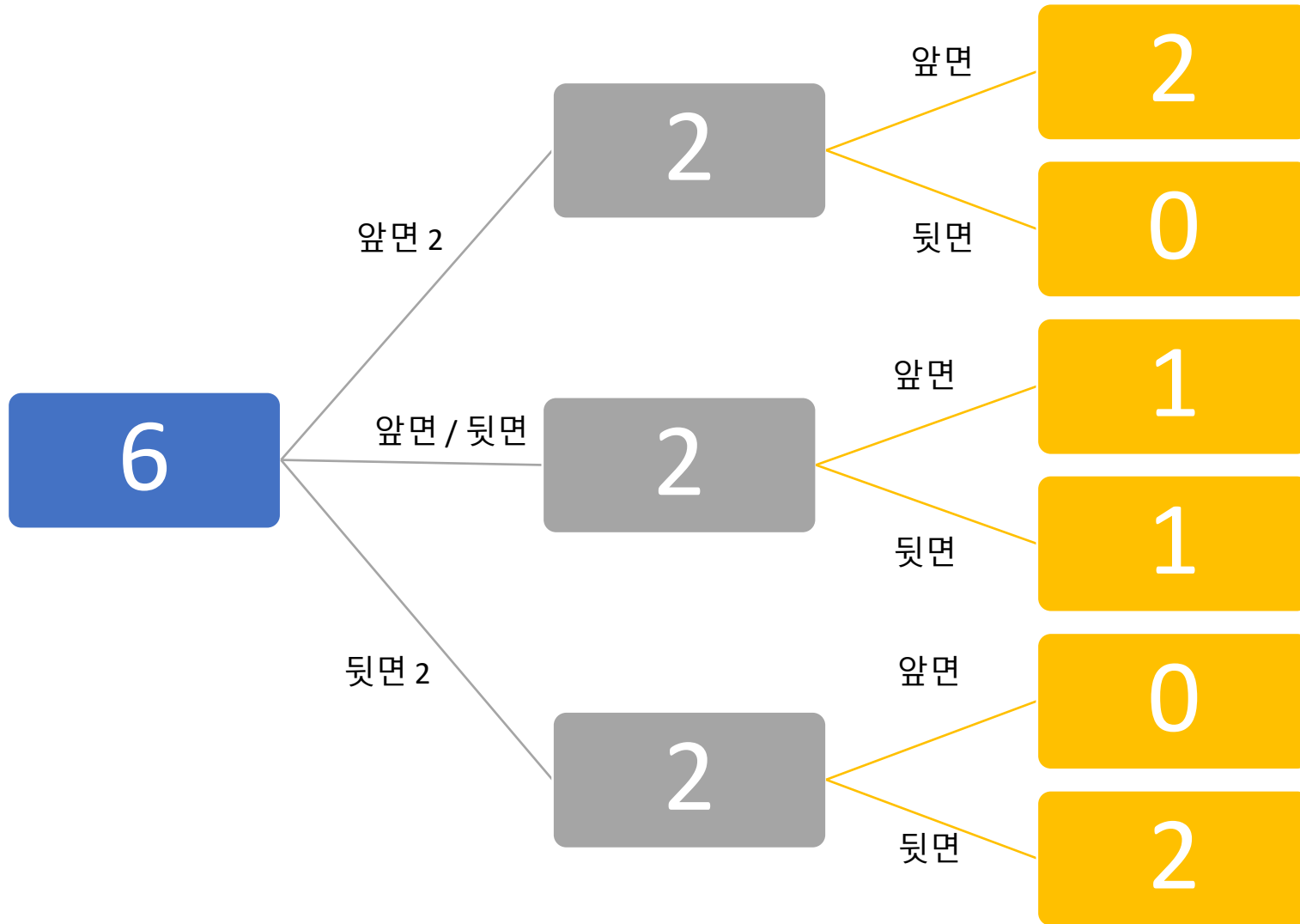


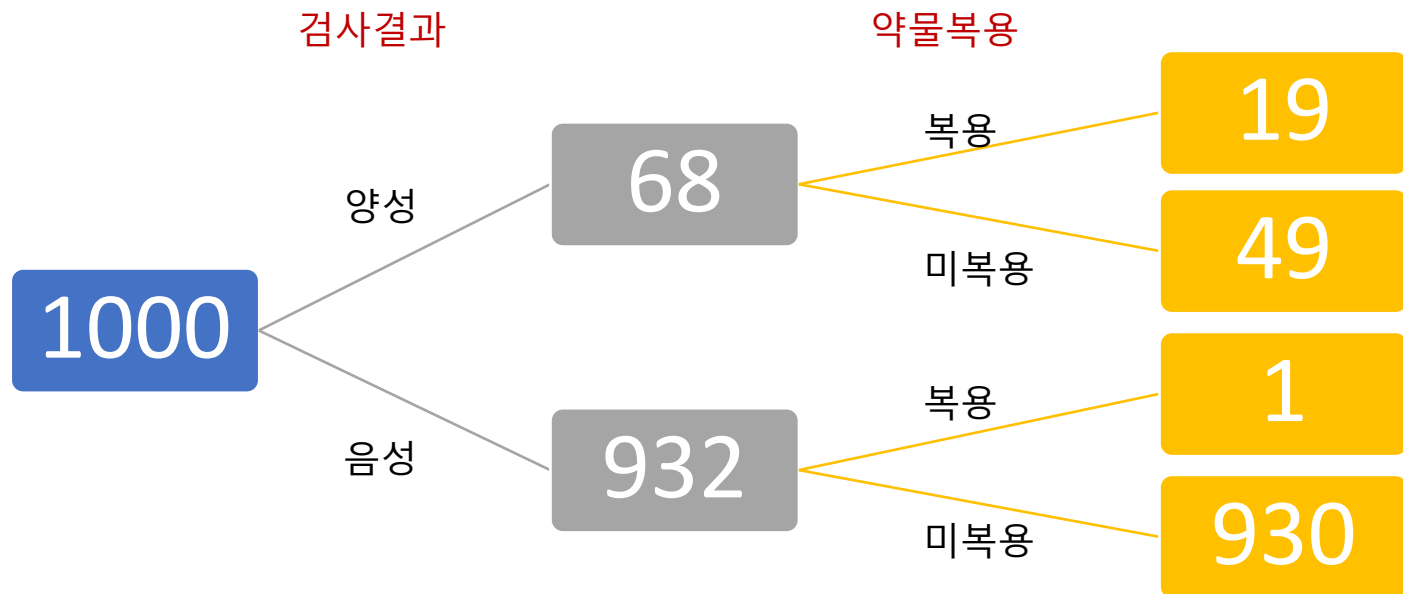
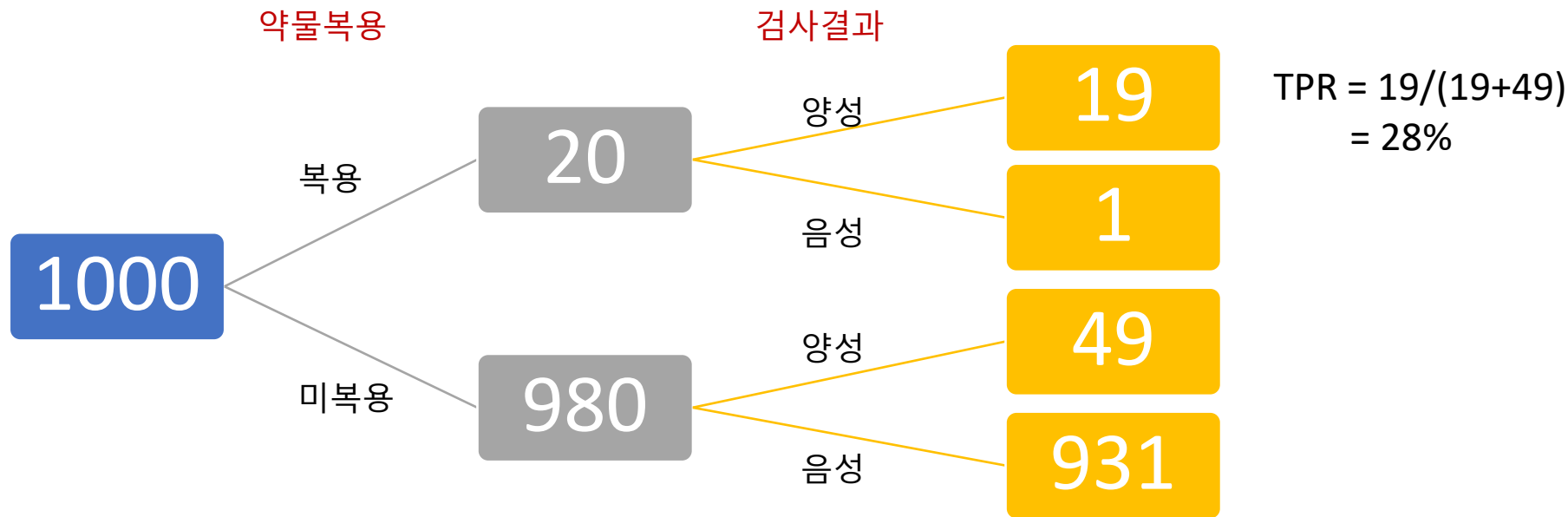
주머니에 동전 세개가 있다. 하나는 앞면만 돌리고, 하나는 앞면과 뒷면이 각각 하나씩 있고, 하나는 뒷면만 돌이다. 당신이 임의로 동전을 하나 골라서 그것을 던졌는데 앞면이 나왔다면, 그 동전의 다른 면이 앞면일 확률은 얼마인가?

스포츠 경기에서 실시되는 금지약물 복용 검사에 정확도가 95%라고 가정하자. 즉 약물 복용자의 95%와 비복용자의 95%를 정확하게 분류한다고 하자. 선수 50명당 1명꼴로 금지약물을 복용하고 있는데 한 선수의 검사 결과가 양성이라면, 그가 정말로 금지약물을 복용하고 있을 확률은 얼마인가?

선택된 동전

던지기 결과





❖ A 선수가 도핑 테스트에서 양성 판정을 받았다. KBO에서 이것을 근거로 징계를 하는 것이 타당한가?

- A 선수의 유죄 증거의 확률 : 95%
- A 선수의 무죄 증거의 확률 : 5%
- A 선수의 가능도비(Likelihood Ratio) :  $0.95 / 0.05 = 19$
- 약물 비복용 선수보다 약물 복용 선수에서 양성 검사 결과가 19배 더 많이 나온다
- 어떤 가설에 대한 승산  $\times$  가능도비 = 가설에 대한 최종 승산
- $1 / 49$ (금지약물 복용에 대한 승산)  $\times 19 = 19 / 49$

## □ Naïve Bayesian theorem

- ❖ 가정 : 주머니에 빨간공 60%, 파란공 40%, 빨간공의 20%는 깨졌고, 파란공의 30%는 깨졌다. 임의로 꺼낸 공이 깨졌을때 이 공이 빨간색일 확률은?
  - $P(\text{깨진공})$  : prior probability, 사전확률
  - $P(\text{파란색}/\text{깨진공})$  : posterior, 우도확률
  - $P(\text{깨진공}/\text{파란색})$  : likelihood, 사후확률
- ❖ 깨진공이 파란색인지 빨간색인지 알아내는 방법
  - Maximum A posterior : 깨진 공이 빨간색일 확률과 깨진 공이 파란색일 확률을 비교해서 더 확률 높은 쪽을 선택
  - Maximum Likelihood : 파란색이 깨진공일 확률과 빨간색이 깨진공일 확률을 계산해서 더 확률 높은 쪽을 선택
- ❖ 가정 : 빨간공 60개 파란공 40개, 빨간공 12개 깨짐, 파란공 12개 깨졌다. 임의로 꺼낸 공이 깨졌을때 이 공이 빨간색일 확률은?

## □ Naïve Bayesian theorem

조건부 확률  $P(B|A) = \frac{P(A \cap B)}{P(A)}$   $P(A \cap B) = P(A)P(B|A) = P(B)P(A|B)$

Bayesian theorem  $p(Y = k|X = x) = \frac{P(X = x|Y = k)P(Y = k)}{p(X = x)}$

- ❖  $P(\text{빨간색}/\text{깨진공}) = P(\text{깨진공}/\text{빨간색}) * P(\text{빨간색}) / P(\text{깨진공})$   
 $= P(\text{깨진공}/\text{빨간색}) * P(\text{빨간색}) / (P(\text{깨진공}/\text{빨간색}) * P(\text{빨간색}) + P(\text{깨진공}/\text{파란색}) * P(\text{파란색}))$   
 $= 0.2 * 0.6 / (0.2 * 0.6 + 0.3 * 0.4)$
- ❖ 확장 : 주머니에 빨간공 35%, 파란공 55%, 흰공 : 10%, 빨간공의 15%는 깨졌고, 파란공의 20%는 깨졌고, 흰공의 35%는 깨짐 임의로 꺼낸 공이 깨졌을때 이 공이 빨간색일 확률은?
- ❖  $P(\text{빨간색}/\text{깨진공}) = P(\text{깨진공}/\text{빨간색}) * P(\text{빨간색}) / P(\text{깨진공})$   
 $= P(\text{깨진공}/\text{빨간}) * P(\text{빨간색}) / (P(\text{깨진공}/\text{빨간색}) * P(\text{빨간색}) + P(\text{깨진공}/\text{파란색}) * P(\text{파란색}) + P(\text{깨진공}/\text{흰색}) * P(\text{흰색}))$   
 $= 0.15 * 0.35 / (0.15 * 0.35 + 0.2 * 0.55 + 0.35 * 0.1)$

## □ Naïve Bayesian theorem

no	words	class
1	fun, couple, love, love	comedy
2	fast, furious, shoot	action
3	couple, fly, fast, fun, fun	comedy
4	furious, shoot, shoot, fun	action
5	fly, fast, shoot, love	action
6	fast, furious, fun	?



no	class	fun	couple	love	fast	furious	shoot	fly
1	comedy	1	1	2	0	0	0	0
2	action	0	0	0	1	1	1	0
3	comedy	2	1	0	1	0	0	1
4	action	1	0	0	0	1	2	0
5	action		0	1	1	0	1	1
6	?	1	0	0	1	1	0	0

$$P(C_1, | x_1, x_2, \dots, x_n) = \frac{P(x_1, x_2, \dots, x_n | C_1)P(C_1)}{P(x_1, x_2, \dots, x_n)}$$

$$P(C_2, | x_1, x_2, \dots, x_n) = \frac{P(x_1, x_2, \dots, x_n | C_2)P(C_2)}{P(x_1, x_2, \dots, x_n)}$$

$$P(x_1, x_2, \dots, x_n | C_1) = P(x_1 | C_1) * P(x_2 | C_1) * \dots * P(x_n | C_1)$$

✓ Feature 들은 서로 독립이라고 가정

## ❑ Naïve Bayesian theorem

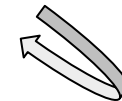
no	words	class
1	fun, couple, love, love	comedy
2	fast, furious, shoot	action
3	couple, fly, fast, fun, fun	comedy
4	furious, shoot, shoot, fun	action
5	fly, fast, shoot, love	action
6	fast, furious, fun	?

$$\begin{aligned}P(\text{comedy}/\text{fast}, \text{furious}, \text{fun}) &= P(\text{fast}/\text{comedy}) * P(\text{furious}/\text{comedy}) \\&\quad * P(\text{fun}/\text{comedy}) * P(\text{comedy}) \\&= 1/9 * 0/9 * 3/9 * 2/5 = 0\end{aligned}$$

$$\begin{aligned}P(\text{action}/\text{fast}, \text{furious}, \text{fun}) &= P(\text{fast}/\text{action}) * P(\text{furious}/\text{action}) \\&\quad * P(\text{fun}/\text{action}) * P(\text{action}) \\&= 2/11 * 2/11 * 1/11 * 3/5 = 0.0018\end{aligned}$$

$$\begin{aligned}P(\text{comedy}/\text{fast}, \text{furious}, \text{fun}) &= P(\text{fast}/\text{comedy}) * P(\text{furious}/\text{comedy}) \\&\quad * P(\text{fun}/\text{comedy}) * P(\text{comedy}) \\&= (1+1)/(9+7) * (0+1)/(9+7) * (3+1)/(9+7) * 2/5 = 0.00078\end{aligned}$$

$$\begin{aligned}P(\text{action}/\text{fast}, \text{furious}, \text{fun}) &= P(\text{fast}/\text{action}) * P(\text{furious}/\text{action}) \\&\quad * P(\text{fun}/\text{action}) * P(\text{action}) \\&= (2+1)/(11+7) * (2+1)/(11+7) * (1+1)/(11+7) * 3/5 = 0.0018\end{aligned}$$



Laplace smoothing



## ❑ Naïve Bayesian theorem

### ❖ 장점

- noise에 민감하지 않음
- 무관한 입력 변수들에 사용하기 적합
- Missing value를 무시할 수 있음
- DTM(document term matrix)와 같은 sparse data에 대해서도 좋은 성능을 보임

### ❖ 단점

- 변수들 간의 상관관계가 있는 데이터
- 큰 biased result
- 충분히 많은 자료 필요