

1주차

<https://www.coursera.org/learn/basic-statistics>

- Week 2. Correlation and Regression
- Week 3. Probability
- Week 4. Probability Distributions

2주차

- Week 5. Sampling Distributions
- Week 6. Confidence Intervals
- Week 7. Significance Tests

<https://www.coursera.org/learn/statistical-inference#syllabus>

- Week 4: Power, Bootstrapping, & Permutation Tests

3주차

<https://www.edx.org/course/linear-algebra-foundations-frontiers-utaustinx-ut-5-05x>

<https://courses.edx.org/courses/course-v1:UTAustinX+UT.5.05x+2T2017/course/>

- Week1. Vectors in Linear Algebra
- Week2. Linear Transformations and Matrices
- Week4. Matrix-Vector Operations
- Week5. Matrix-Matrix Multiplication

(선형대수학을 이미 들으신 분들은 알아서 패스)

- Week6. Gaussian Elimination
- Week7. More Gaussian Elimination and Matrix Inversion

4주차

- Week9. Vector Spaces.
- Week10. Vector Spaces, Orthogonality, and Linear Least-Squares
- Week11. Orthogonal Projection, Low Rank Approximation, and Orthogonal Bases

5주차

<https://www.coursera.org/learn/machine-learning>

→ 구매함! 프로그래밍 과제도 같이 다 할 것! 안되면 되게 하라! → octave는 못할 듯하다!! 개념만!!

- Week1. Linear Regression with One Variable
- Week2. Linear Regression with Multiple Variables

6주차

- Week3. Logistic Regression
- Week8. Unsupervised Learning
- Week8. Dimensionality Reduction

7주차

- Week9. Anomaly Detection
- Week9. Recommender Systems

1~8주차 : Udemy : <https://www.udemy.com/complete-python-bootcamp/>

Complete python bootcamp

Machine Learning A-Z™ → 머신러닝은 아직은 필요가 없을 듯하다!!

반드시 강의를 complete하자!

<EdX LAFF> 8장도 언젠가는...

2월 11일(일요일) : 10장 끝내기 (완료!)

2월 18일까지(일요일) : 11장 끝내기 (완료!)

3월 4일까지(일요일) : 12장 끝내기 (못함)

<Complete python bootcamp>

2월 18일까지(일요일) : 섹션 6.

2월 25일까지(일요일) : 섹션 7. milestone project 1.

3월 4일까지(일) : 섹션 8/18

3월 11일까지(일) : 섹션 9/18

3월 18일까지(일) : 섹션 10/18 . milestone project 2.

3월 25일까지(일) : 섹션 12, 13/18. 12장 eigen vector

4월 1일까지(일) : 섹션 14/18 = milestone project 3.(final) + 7주차(Ng)

4월 8일까지(일) : 섹션 15, 17 + 10주차/11주차 끝내기(Ng)

4월 15일까지(일) :

<스탠포드 머신러닝>

2월 18일까지(일요일) : 3장 끝내기(과제도 같이), 8장은 이론만 끝내기.

2월 25일까지(일요일) : 9장 이론만 끝내기. // 2장 3장 과제!!

3월 4일까지(일) : 파이썬 잘 따라잡고, 머신러닝 9장 마무리, 옥타브 2장, 3장하기, 선대 12장 끝내기.

3월 11일까지(일) : 4주차 끝내기

3월 18일까지(일) : 5주차 끝내기

3월 25일까지(일) : 6주차 끝내기

4월 1일까지(일) : 7주차 끝내기.

4월 8일까지(일) : 10주차/11주차 끝내기.

4월 15일까지(일) :

<유데미 Machine Learning A-Z™>

파이썬으로.

<참가자 리스트>

김혁동 : 경제학,

이재진 : 하늘색 데이터를 분석하자.. power 물어본 분

김윤정 : 육아휴직, 데이터마이닝 관심 공학쪽.

이근후 : 엄청 도움 많이 줌

황?종성 : 물어봐도 대답 안 해주신 분

이승연 : 통계툴, 인턴을 했다.

조동빈 : 진로

천동욱 : 마케터, 유학, 건축공학

김민선 : 회사 근무, 솔루션

김명수 : 얼리버드 운영, 자퇴, 음악 컨텐츠

김완희 : 운영팀

최석원

김현정 : UX 컨설팅, 데이터 기반, 평균 체류시간

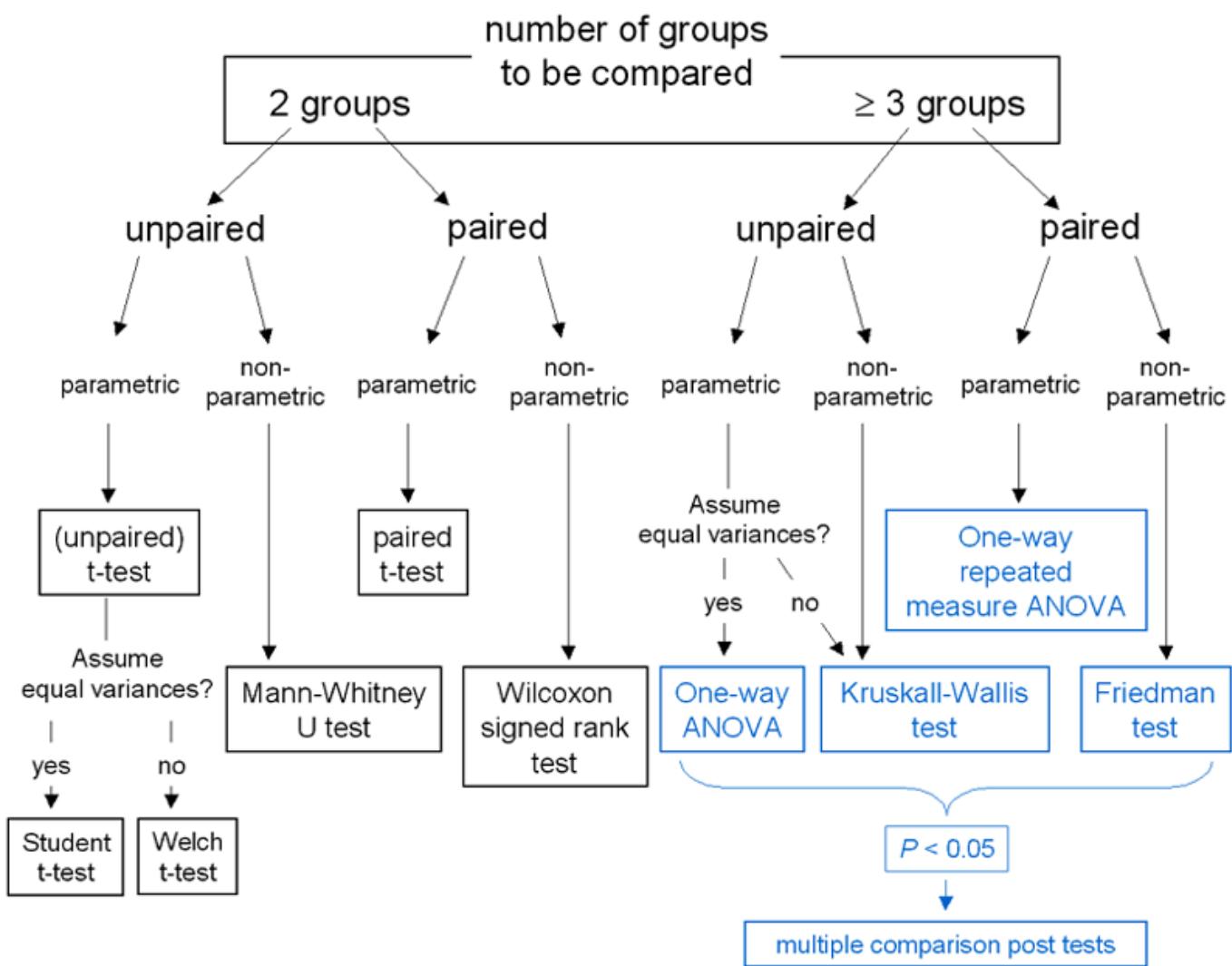
정미연 : 분석쪽 공부, 통계쪽으로 기초개념을 다지겠다

오윤정 : 마케팅, 관련 대학원, 프리스터디

이재연 : 회사 클라인트

<질문 리스트>

어떤 테스트를 해야 하는가?



* Parametric : That means that if the mean and standard deviation are known and if the distribution is normal, the probability of any future observation lying in a given range is known

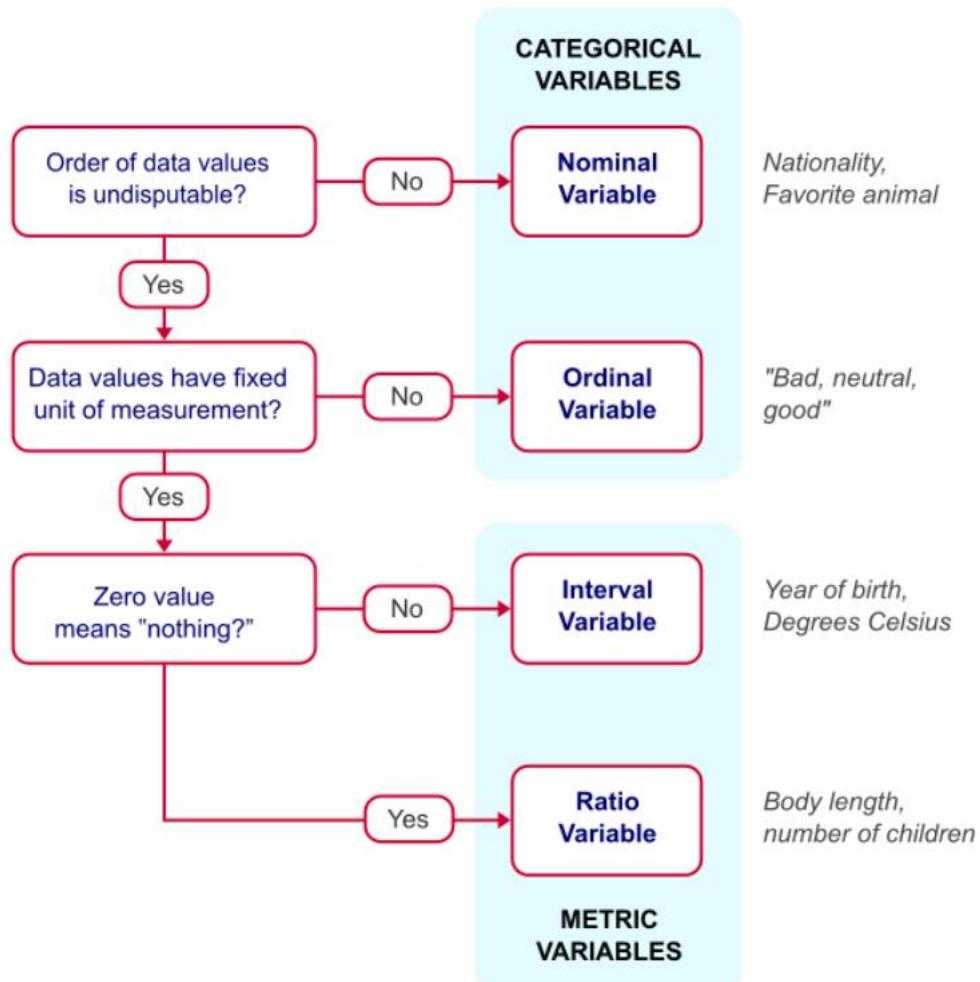
* (근후님의 가르침)

nonparametric 하다는 것은 모수를 사용하지 않고도 추정이 가능하다는 것이므로, CLT에 의해서 sampling distribution은 정규분포에 근사되므로 이를 사용하면 비모수적 추정이 가능하다!

따라서, sampling distribution을 통한 추정은 모분포가 정규분포라는 것을 알면 parametric 한 것이고, 모분포가 정규분포인 것을 가정하지 않더라도 sampling distribution을 사용하면 nonparametric 하게 추정을 할 수 있다. 물론 bootstrapping 방식을 사용하는 것은 nonparametric 한 방식이다.

* 변수의 분류

MEASUREMENT LEVELS - CLASSICAL APPROACH



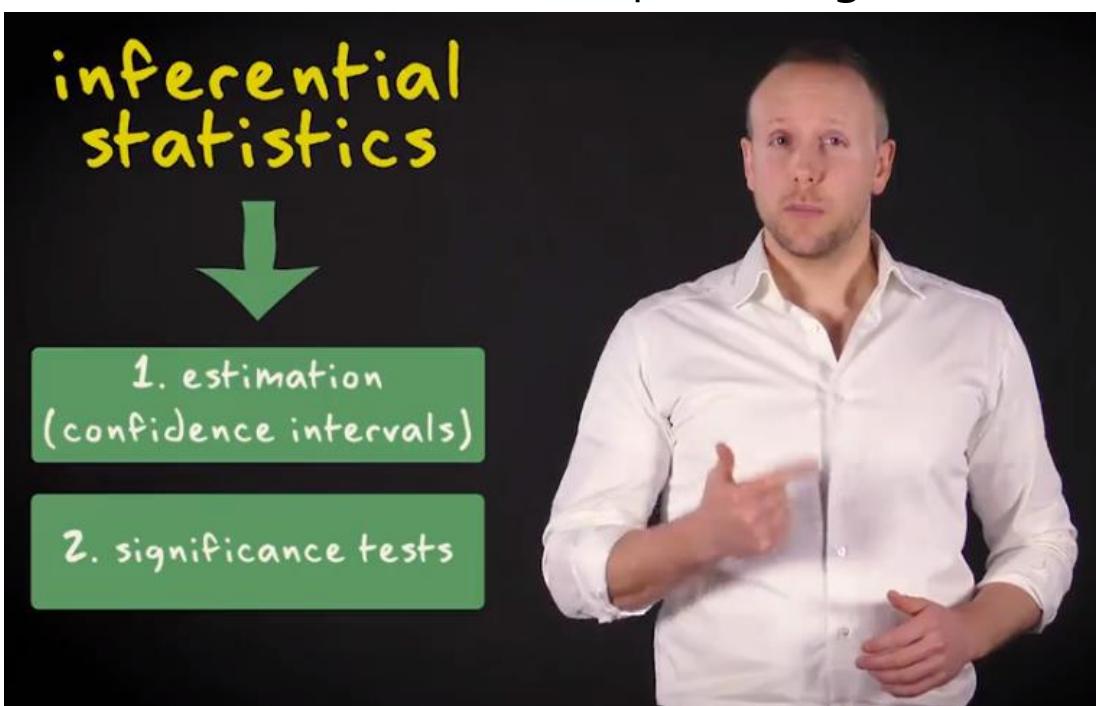
MAP

제 1 장. 기술 통계학 (descriptive statistics)

1. Bivariate analysis
2. Probability
3. Univariate analysis

제 2 장. 추론 통계학 (inferential statistics)

4. Sampling (표본 추출)
9. Resampling
5. 추정 (estimation)
6. 가설 검정 (test hypothesis)
7. 가설 검정 결과의 해석의 연장선상 : Type of error
8. 가설 검정 오류의 관리 : Multiple testing corrections



Basic statistics Detailed Map

발제 : 이기석

제 1 장. 기술 통계학 (descriptive statistics)

1. Bivariate analysis (\rightarrow 변수 2개의 관계 regression analysis)

- regression line
- pearson's r
- r square

2. Probability

- 집합의 개념
- 확률의 연산, 독립, 베이즈 정리

3. Univariate analysis (\rightarrow 변수 1개의 확률분포)

- 확률분포
- 묘사 : parameter
- 정규분포, 이항분포 (\leftarrow parametric statistics)

제 2 장. 추론 통계학 (inferential statistics)

4. Sampling (표본 추출)

- sampling method들
- sampling distribution : of sample mean, of sample proportion
- central limit theorem

* 9. Resampling without sampling

- bootstrapping (\leftarrow nonparametric)
- permutation test

5. 추정 (Estimation)

- 점추정 / 구간 추정(interval estimation)
- 신뢰구간 추정 방법 : 1) 모비율, 모평균 2) 신뢰수준(95%) 3) 표본수 4) margin of error

6. 가설 검정 (Test hypothesis)

- (1) 가설 설정 단계 : 1) 모비율/모평균 2) 단측/양측 3) 유의수준
- (2) sampling과 검정 단계 : 검정통계량 결정(모평균 / 모비율)
- (3) 결과 해석 : P-value와 유의수준 비교 / 기각역으로 판단

7. 가설 검정 결과의 해석의 연장선상 : Type of error

- Type 1 error (false positive)
- Type 2 error (false negative)
- power

8. 가설 검정 오류의 관리 : Multiple testing corrections

- Bonferroni correction : Family-wise error rate (FWER)
- BH correction : controlling false discovery rate (FDR)

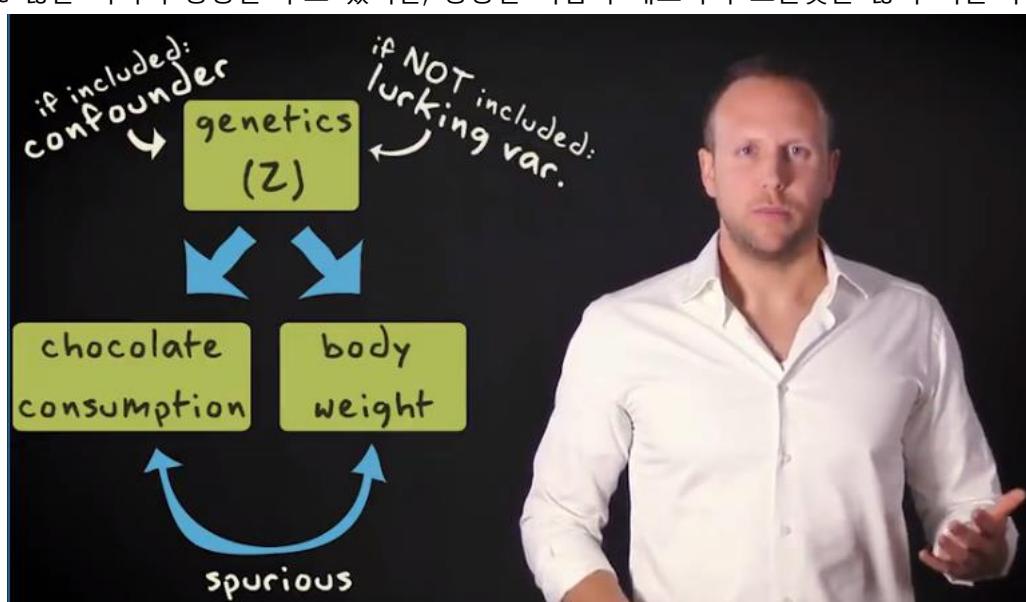
제 1 장. 기술 통계학 [descriptive statistics]

1. Bivariate analysis → Correlation (변수간의 관계)

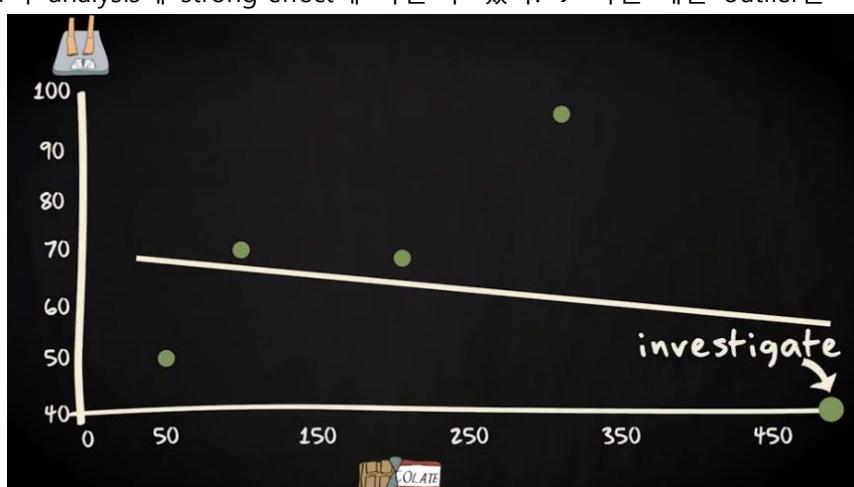
- (1) 종속변수(dependent), 독립변수(Independent)가 무엇인지 정하기
- (2) Is there a linear relationship?
- (3) 맞으면 Pearson's r 구하기 → linear관계가 strong한가? $r = \frac{\sum Z_x Z_y}{n-1}$
- (4) Regression line/equation은?
 $b = r(\frac{S_y}{S_x}) \quad a = \bar{y} - b(\bar{x}) \quad \Rightarrow \quad \hat{y} = a + bx$
- (5) regression line의 y의 variance에 대한 설명력은? r^2
 - prediction error가 y의 평균을 사용했을 때보다 69% 적다.
 - y의 variance의 69%가 x로 설명이 된다.

Cf). 해석시 주의점

- (1) Correlation과 causation은 다르다!!
 - chocolate 많은 먹어서 뚱뚱할 수도 있지만, 뚱뚱한 사람이 배고파서 초콜릿을 많이 먹을 수도 있음



- 제3변수가 있을 수 있다. :
 - confounding variable : 연구에 포함 되었으면 ex) genetics → chocolate consumption
→ body weight.
 - lurking variable : 연구에 포함 안되어있는 제 3변수.
- (2) influential outlier가 analysis에 strong effect에 가질 수 있다. → 이럴 때는 outlier를 제외해도 된다.



1.1. 변수 종류와 표현 방법

(1) Nominal/ordinal variable ex) less than 50kg / 40~50g chocolate

→ Contingency tables : display the relationship between two variables.

→ 항상 percentage로 바꿀 것!!! Conditional proportion / marginal proportion 그래야 의미를 읽을 수 있다. 물론 scatter이면 더 좋겠지만. 이건 contingency table로밖에 나타낼 수 없으니.

	High	Middle	Low	total
Agree	50%	28%	12%	26%
Agree/disagree	34%	35%	25%	31%
Disagree	16%	37%	63%	43%
total	100%	100%	100%	100%

(2) tree diagram : contingency table을 확률적으로 보기 쉽게 나타내려면 수형도가 좋다

→ use the tree to quantify the probabilities. In order to quantify probabilities, you need to experiment or make plausible assumptions ex) all events are independent

- 한계 : 복잡하게 event가 너무 많아지면 tree를 전부 그리기가 어려워진다.

(3) Quantitative variable ex) 65kg /45g chocolate consumption → Scatter plot (지금부터 할 것!)

1.2. How strong?? direction and strength of the relationship : Pearson's r

- pearson's r : scatter plot에서!

$$\text{Pearson's r 계산법} : \frac{\sum Z_x Z_y}{n-1}$$

$$r = \frac{\sum Z_x Z_y}{n-1}$$

X	Z _x	Y	Z _y	Z _x *Z _y
50	-1.01	50	-1.15	1.17
100	-0.56	70	-0.07	0.04
200	0.34	70	-0.07	-0.02
300	1.24	95	1.29	1.60

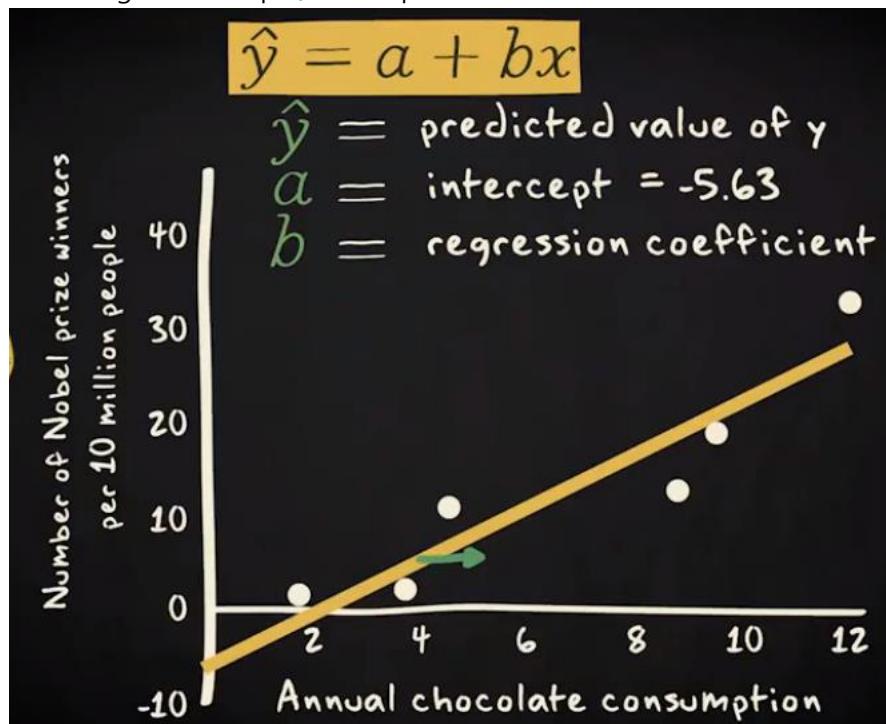
- 주의 : scatter plot을 먼저 그려보고 linear일 때만 사용할 것. linear가 아니면 pearson's r을 사용하지 말 것

- 의미 : -1과 +1 사이이고, 절대값이 클수록 strong relationship이다. +는 positive, -는 negative relationship.

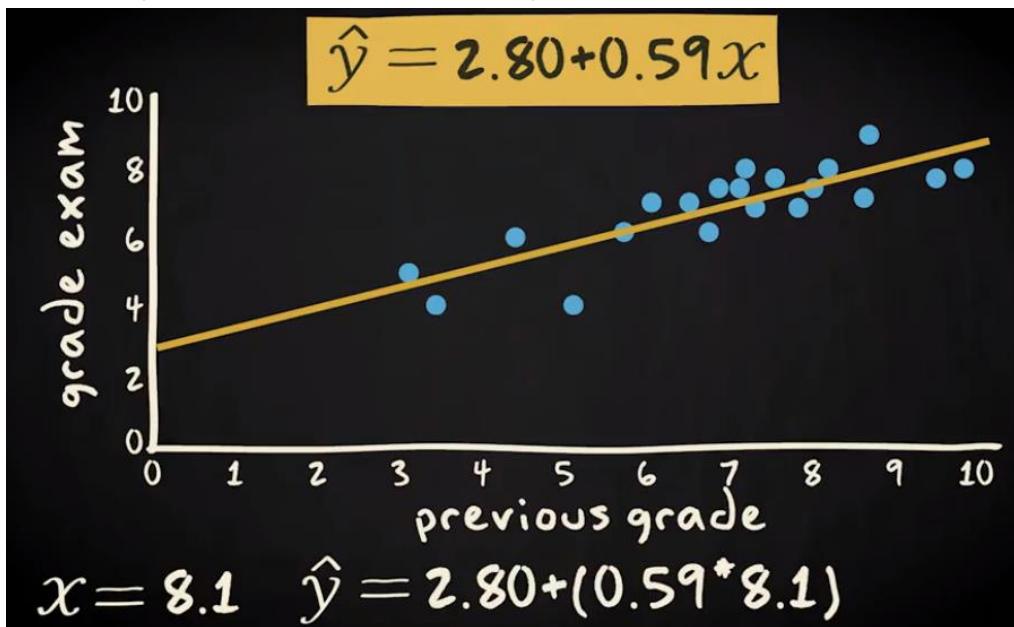
1.3. Regression

1.3.1. 의미 : line that best represents the linear correlation between two quantitative variables in a scatter plot

- regression coefficient = regression slope / intercept



- 용도 : can make predictions. X를 통해서 Y prediction 가능. 1개의 변수에 대한 자료만 있으면 그것의 평균으로만 추측을 해야 한다. (평소에 생각하지 못했던 포인트)



1.3.2. Computing method

- OLS (ordinary least squares) minimize the sum of the squared residuals (line부터 점까지의 수직선 길이)

- 계산하는 법 : b (regression coefficient)는 pearson's r의 unstandardized version이고 a 는 regression line의 \bar{x} 와 \bar{y} 를 지나므로 대입해서 구할 수 있다.

COMPUTE REGRESSION LINE

$$b = r \left(\frac{s_y}{s_x} \right)$$

$$a = \bar{y} - b(\bar{x})$$

$$b = 0.93 * \left(\frac{11.87}{3.95} \right) \quad a = 13.17 - (2.79 * 6.71)$$

$$b = 2.79$$

$$a = -5.55$$

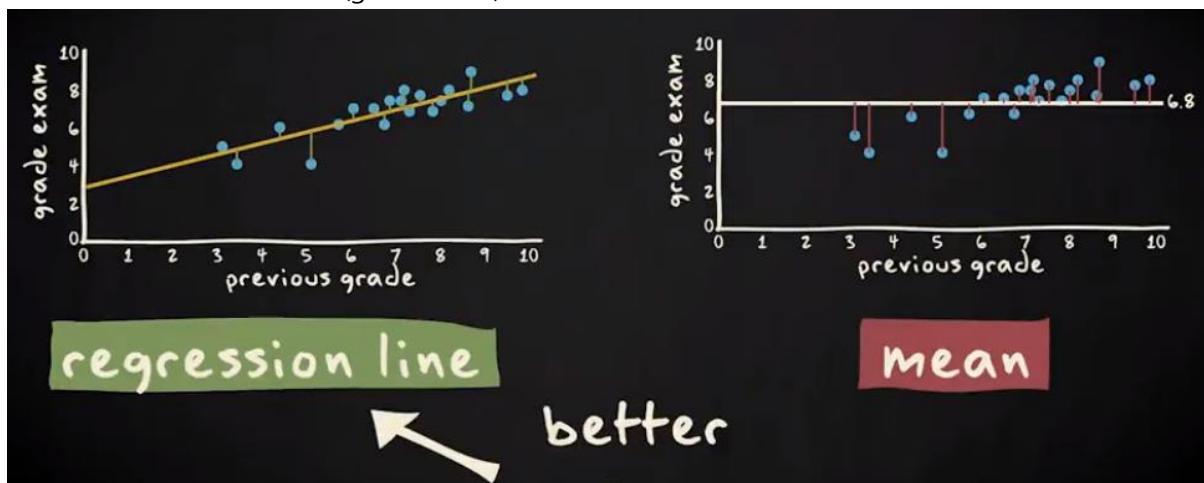
1.3.3. How well the regression line predicts your dependent variable : r-squared

= How well the regression line fits the data → accuracy of prediction

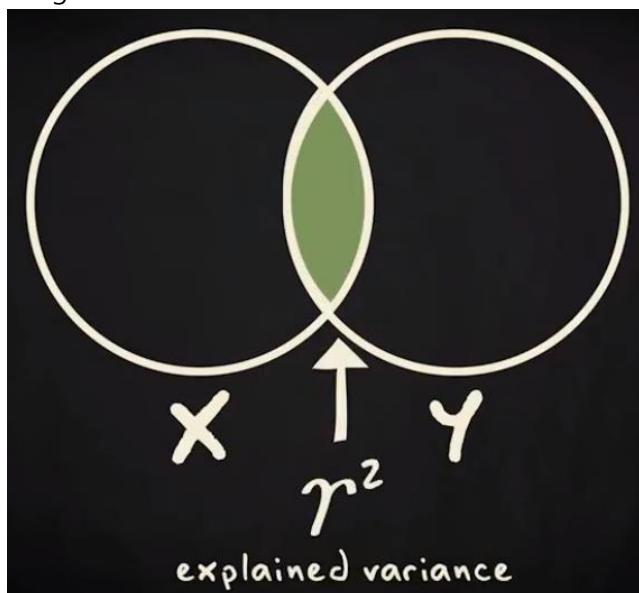
- 계산 : pearson's r을 제곱하면 된다.

(1) r^2 첫번째 의미 : tells you how much better a regression line predicts the value of a dependent variable than the mean of the variable. Ex) $r^2 = 0.69$ 이면 prediction error is 69% smaller than when you use the mean.

만약에 변수 1개 밖에 없었으면 (grade exam) 그냥 평균으로 옆자리 친구의 점수를 예측해야 할 것이다.



(2) 두번째 의미 : the amount of variance in your dependent variable(Y) that is explained by your independent variable(X) ex) $r^2 = 0.69$ 이면 69% of the variance in the grades can be predicted by the previous grades.



그림에서 동그라미는 각각의 variance.

→ 겹치는 부분이 커진다는 것은 x의 variance로 설명할 수 있는 y의 variance가 커진다는 것이다.

1.3.4. Pearson's r와 r square의 차이. 제곱하면 direction은 알 수 없겠지.



2. Probability and Randomness

2.1. Probability : "probability is a way to quantify randomness."

- 주요 개념 : 전체 3번 뽑는 것이 experiment이고, 각각의 뽑는 것이 trial이다. 그 trial로서 얻는 3번 뽑은 결과가 outcome이다. Event란 outcome 혹은 combination of outcome이다. 특정 event가 일어날 randomness를 수치화 한 것이 probability(확률)이다. Random variable은 하나의 trial에서 얻을 변수이다.



- randomness = chance = uncertainty = risk = likelihood : 랜덤은 현상의 내재적인 성질(not an intrinsic) property of a phenomenon이 아니다. (대박!!!) Prior knowledge, observation method, scale at which the phenomenon is considered가 randomness에 영향을 미칠 수 있다. Ex) 개미 채집시, 여기에 개미가 많다는 것을 알면 개미를 찾을 확률이 영향을 받는다.

- probability란 무엇인가? (소름주의, 이렇게 생각해본 적이 없는 것 같다.)

: probability is a way to quantify randomness. (trial이 적을 때는 relative frequency가 확률과 잘 안 맞지만 수가 커지면 relative frequency가 확률에 근접해간다)

<석원's take on Probability>

- probability (석원이의 정의) : 함수이다. Event $\rightarrow X$ (random variable) 숫자로! \rightarrow 확률 (숫자로!)
- Random variable이란 Event를 매핑하는 것이다.
- \rightarrow 모든 event를 실수로 매핑하는 것 : $-\infty \sim +\infty$

- 사례 :

- sample space (전체 집합) = collection of all possible outcomes for random phenomenon
Ex) {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}

- X : 첫번째 trial이 앞면이면 1, 뒷면이면 0

- Y : 최대 연속 H의 수, 0, 1, 2, 3

Y 0 1 2 3

X

0 0 2/8 1/8 1/8

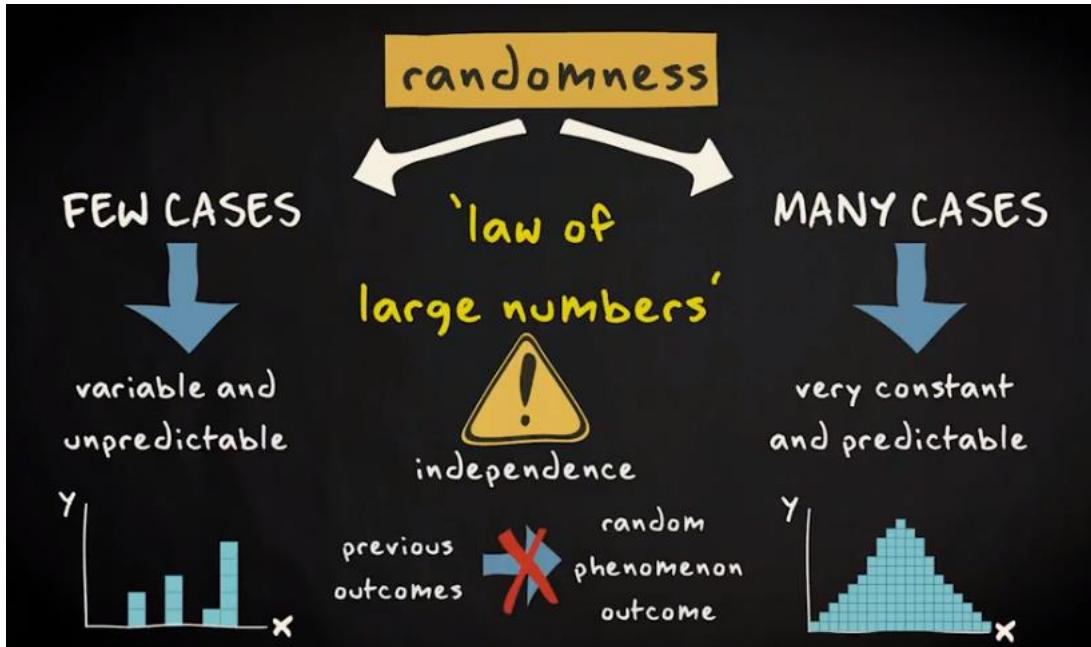
1 1/8 2/8 1/8 0

- 랜덤에 대한 사고 오류

- gambler's fallacy : 과거의 패턴으로 미래를 예측하려고 하는 것은 오류. Ex) 6이 4번 나왔으니 다음번에는 6이 안 나올 확률이 높겠지. Ex) 내가 주식을 할 때... ㅋㅋㅋ 맞아...
- 골고루 퍼져 있는 게 random이라고 생각하는 오류

2.2. probability and law of large numbers

- 전제 : independent trial이 전제되어야 한다. 즉, 과거의 결과가 미래의 랜덤한 결과에 영향을 주지 않는다.



- 교훈 : Persistence beats the odds. \rightarrow so keep calm and carry on

- independent trial 수가 커질수록 예측 가능성이 높아진다.

- 물론 삶(experiment : 모든 사건들)의 사건(event)들이 모두 서로 independent하지는 않다. 그렇지만 이것(independent trial)을 가정하고 수를 키우면 그것의 확률(probability)에 수렴한다.

2.3. Set theoretic concept(집합 개념)과 벤 다이어그램

- sample space (전체 집합) = collection of all possible outcomes for random phenomenon

Ex) {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}

- event : subset of the sample space / outcome or combination of outcomes
- **disjoint, mutually exclusive event** : 안겹치는 event = events that do not share any outcomes → sum of probability of disjoint event < 1
- **collectively exhaustive** : 합쳐서 전체 sample space가 되면(겹쳐도 된다)
→ sum of probability of **disjoint & collectively exhaustive** event = 1
- 여집합의 개념 : complement
- 교집합 : intersection
- 합집합의 개념 : probability of **union** is the sum of the probabilities of the events - probability of intersection

2.4. 확률의 연산과 Bayes' law



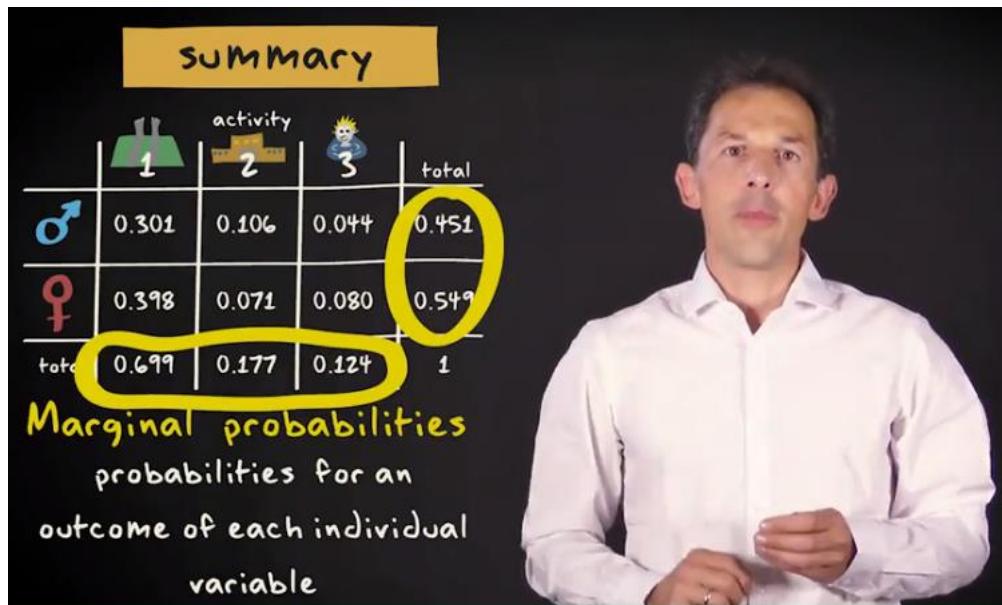
2.4.1. joint probability

		activity			total
		1	2	3	
♂	1	0.301	0.106	0.044	0.451
	♀	0.398	0.071	0.080	0.549
total		0.699	0.177	0.124	1

Joint probabilities
probabilities for the
intersection of certain
outcomes of the variables

3.4.2. marginal probability : $P(A)$ or $P(B)$ → the probabilities for a single variable

→ 하나의 변수만 고려하고 다른 변수는 고려하지 않는다. 하나 변수를 고정하고 다른 변수인 것들은 다 합하면 구할 수 있다.



2.4.3. conditional probability : the probability of an event, given that another event occurs.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \rightarrow P(A \cap B) = P(A | B) * P(B)$$

→ dependent, independent 상관없이 다 사용 가능

(석원이의 해석) sample space가 변하는 것이다. Ex) $P(H\text{가 } 2\text{개} | H\text{가 처음에 나올 경우})$

2.4.4. Independence of random events :

- 정의 : 한 random event의 outcome을 아는 것이 다른 사건에 대한 정보를 주지 못할 때. (Knowing the outcome of one event does not influence your knowledge about the outcome for others)

- 좋은 이유 : simplify the calculation enormously

- independent한가를 보려면? $P(A \cap B) = P(A) * P(B)$ 인지 확인해보기 (난 이 방법 선호)

		activity			total
		1	2	3	
♂		0.451	* 0.699	= 0.315	0.451
♀					0.549
total		0.699	0.177	0.124	1

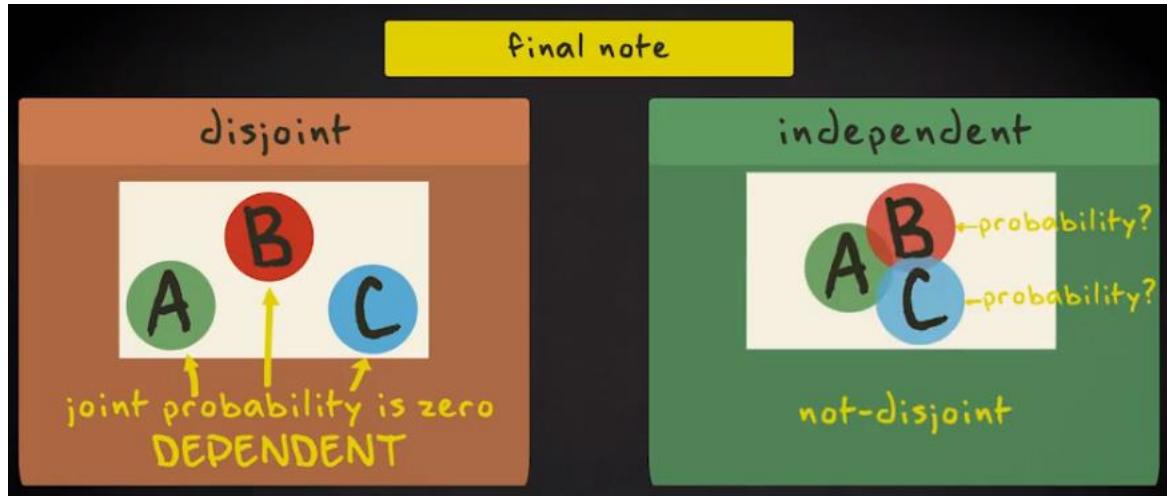
1.	2.	Total
$P(A, \text{ given that } B \text{ occurs}) = P(A)$ or $P(B, \text{ given that } A \text{ occurs}) = P(B)$		
if one holds = any holds		
implies independence		
2.		Total
	$P(A \text{ and } B) = P(A) * P(B)$	
	0.318 0.071 0.080 0.097 0.057 0.063 Total 0.691 0.177 0.124 1	

conditions for independence

- 주의 사항 : (1) independent하기가 쉽지가 않다. 간접적으로 두 변수가 연관이 있을 수 있다. 직접적인 제 3 변수가 관여한 것이 아니더라도, 두 단계 이상을 거칠 수 가 있다. Ex) 토양과 미생물 구성의 변수 : 독립적인가?? 아닐 것이다.

(2) hard to demonstrate : 보이기가 쉽지 않다. Large random sample이 필요하다.

(3) independent하다는 것은 disjoint가 아니라는 것을 의미하기도 한다. 왜냐하면 disjoint하면 하나의 사건이 일어난 경우, 다른 사건의 확률에 대한 정보가 생기기 때문에 dependent하기 때문이다. 게다가 수학적으로 joint probability가 0이다.

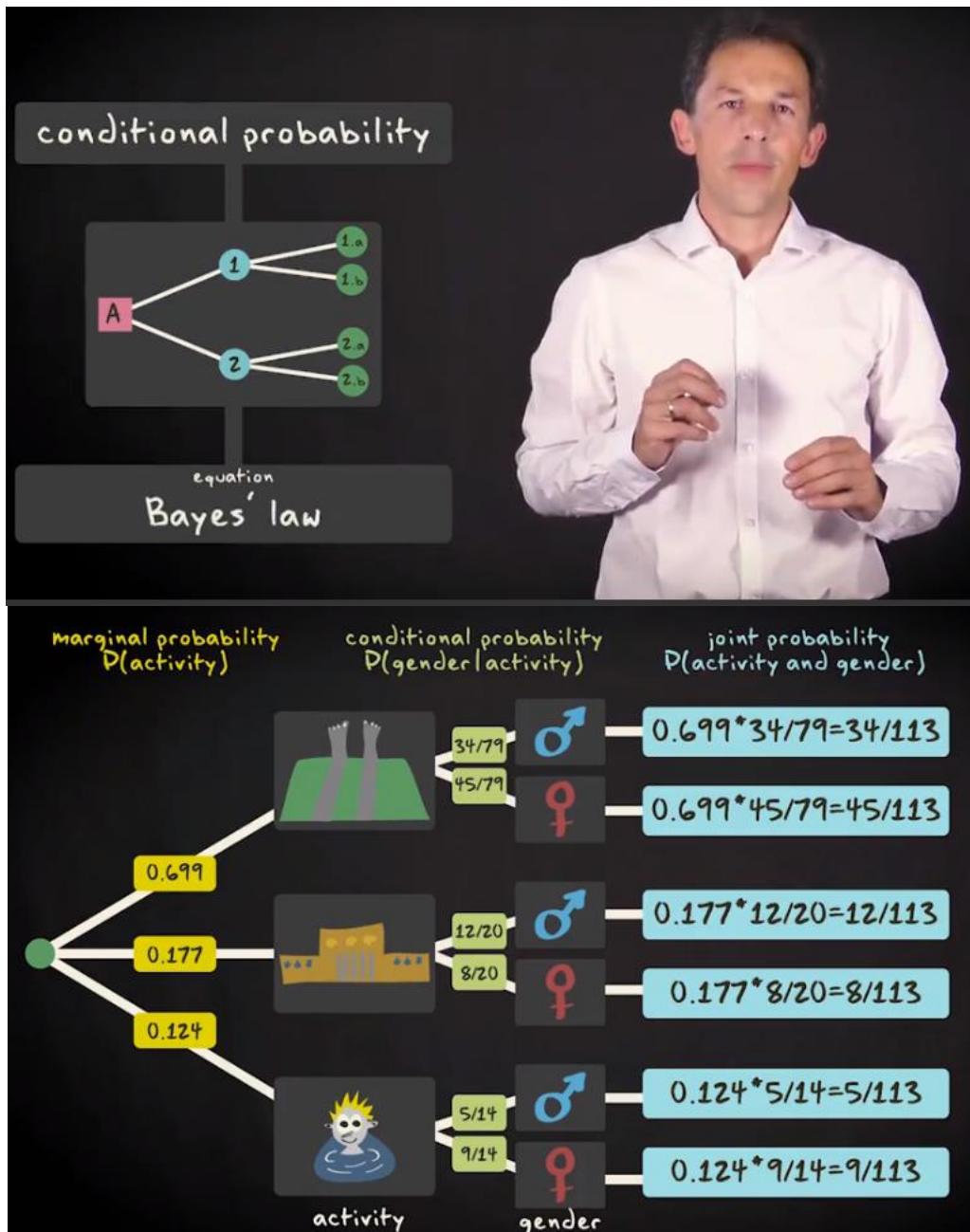


2.4.5. Bayes' law : relation that relates different conditional probabilities.

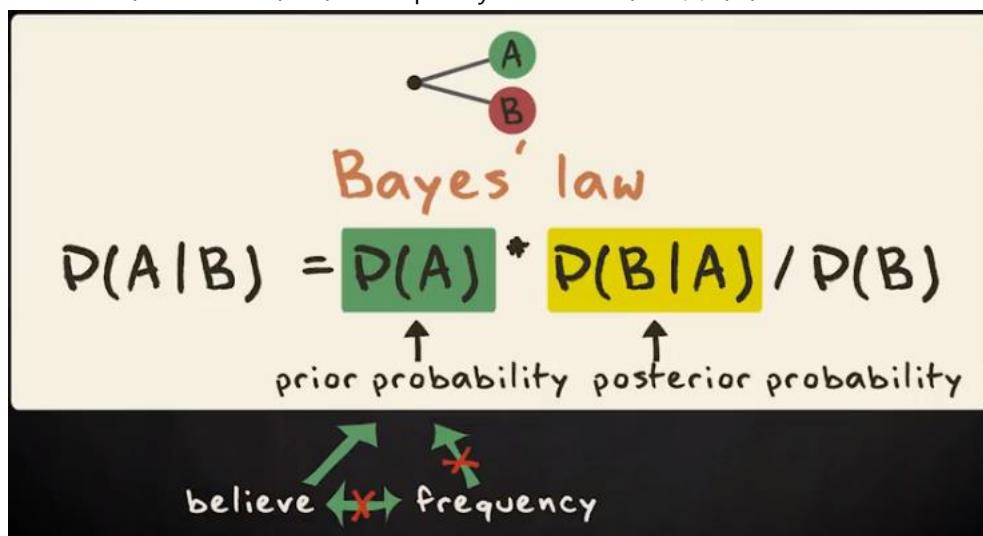
- 도출 : 어떤 변수를 앞에 놓든 간에 순서를 바꿔서 곱해도 된다는 것을 사용해서 베이즈 법칙 도출

$$\text{- Bayes' law : } P(A|B) * P(B) = P(B|A) * P(A)$$

$$\begin{aligned} \rightarrow P(A | B) &= \frac{P(B|A)*P(A)}{P(B)} \\ &= \frac{P(B|A_1)*P(A_1)}{P(B|A_1)*P(A_1)+P(B|A_2)*P(A_2)+P(B|A_3)*P(A_3)} \end{aligned}$$



- 철학적 이해 : $P(A)$ 는 prior probability인데 B 를 관찰하기 전에 A 에 대한 지식을 동원해야 한다. 따라서 이는 Belief를 기반으로 한다. 이는 frequency based랑 다른 것이다.



이 그림은 잘못됐다. Posterior가.

$$P(H|E) = \frac{P(H) * P(E|H)}{P(E)}$$

The diagram illustrates the components of Bayes' rule:

- Prior Probability** (Top Left): An arrow points down to the term $P(H)$.
- Likelihood of the evidence 'E' if the Hypothesis 'H' is true** (Top Right): An arrow points right to the term $P(E|H)$.
- Posterior Probability of 'H' given the evidence** (Bottom Left): An arrow points up to the term $P(H|E)$.
- Prior probability that the evidence itself is true** (Bottom Right): An arrow points left to the term $P(E)$.

석원이 강의

- 목표 : 모분포, 모함수에 대한 정보를 찾고 싶은 것이다. By sampling data를 통해서 추정
→ 100개의 sample 중에서 1번을 몇 % 뽑았다면 그것이 estimator로 쓰인다.
- 어떤 estimator가 좋은 estimator인가 → 이게 데이터 사이언스의 가장 핵심적인 문제이다.
 - (1) biased/unbiased 인가.
 - (2) variation이 좁으면 좋다. 표본수를 늘리면 된다.
- 두개의 tradeoff가 있다. Unbiased, high variation에서 machine learning을 통해서 biased하고 variance가 낮은 예측을 할 수도 있다. 그 방식이 좋은 경우도 있다. Ex) MAP (Maximum A posterior)

(1) Maximum likelihood estimate (MLE)

Ex) 동전을 던졌을 때, 10번 중 6번이 나왔으면, 가장 높은 확률이 6번이다! 순전히 데이터를 가지고 아무런 편견을 가지지 않고 6번이라고 말하는 것.

(2) Maximum A posteriori (MAP) estimation

→ belief (Bayes' law 관련)를 얘기하는 것

$P(X | Z) \rightarrow$ 10번 던져봤더니, 6번 나왔으면, 이 동전은 앞면이 몇번 나올 동전이냐? 5번이 정상 아닌가???

$$P(X | Z) = P(X) * P(Z|X) / P(Z)$$

<적용>

(1) MLE는 Maximize " $P(Z | X)$ " ← likelihood $P(10\text{번 중 } 6\text{번 나옴} | 6\text{번 나오는 주사위}) > P(10\text{번 중 } 6\text{번 나옴} | 5\text{번 나오는 주사위})$ 이므로 6번이 맞다.

Ex) 긴 머리카락을 주워보니, 범인은 여자일 것 같다.

아무런 전제를 갖지 않는 것이다. ONLY 데이터를 가지고만 추정

(2) MAP는 Maximize " $P(X | Z)$ " ← posterior

Ex) 남고 남자 긴 머리카락이 나왔다.

근데 MAP를 쓰기 위해서는 $P(X)$ prior를 알아야 한다. 근데 잘 몰라 $P(X)$ 를.... 그래서 MAP이 어려운 것이다. 만약 남자와 여자의 비율이 같으면(0.5) MLE와 MAP이 같아진다.

- 1만번 실험을 했으면 4987/10000 데이터가 나왔다.

10000 부분이 높을수록 prior에 대한 belief가 높은 것이다.

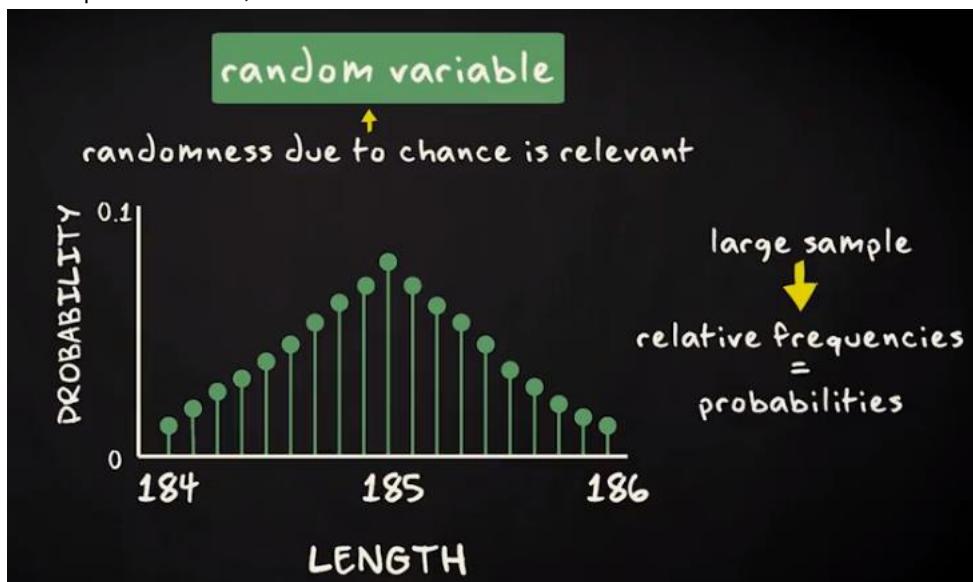
이후의 사람이 0.3을 측정했다면

- 현실은 MLE, MAP 중간으로 가야한다. 표본이 작아서 Variance가 너무 크기 때문에 현실에서 문제가 있는 것이다. 그래서 prior를 가져와야 하는 것이고, 정성적인 지표. 이를 넣게 되면 variance가 떨어지게 된다. 물론 bias가 생길 수는 있다. Ex) 모함수가 정규분포로 가정을 하자. 그것은 MAP와 같은 belief를 가져온 것이다. 답이 나오더라도 variance가 크기 때문에 그런 가정이 필요한 것이다. 뭐 선형성을 가정하는 것도 이와 같은 행위이다.

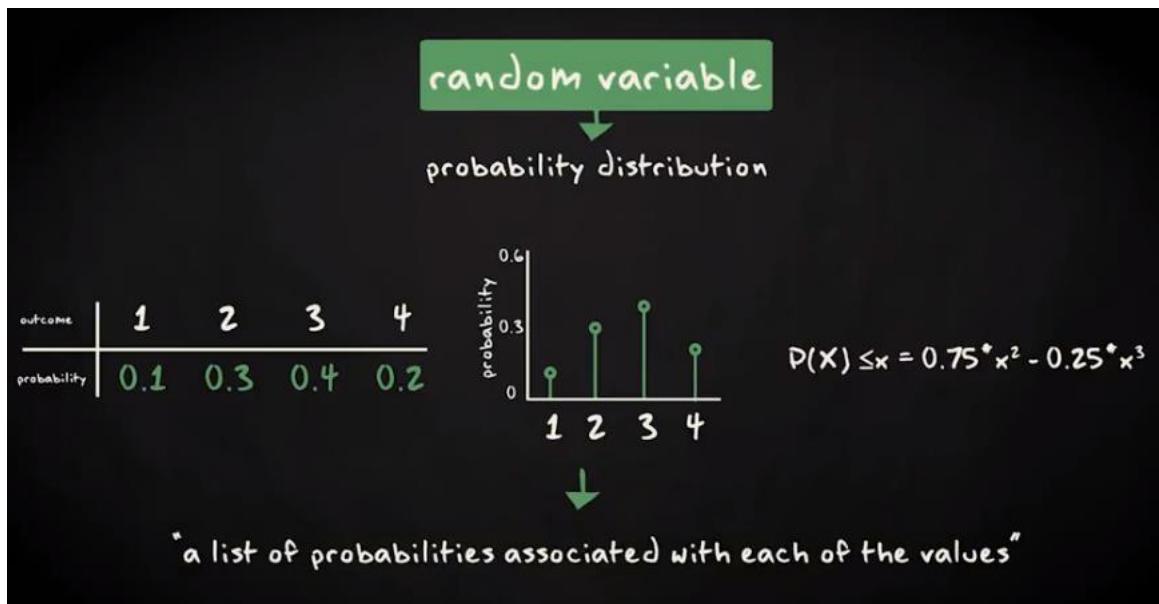
3. Univariate analysis

3.1. Probability distribution (확률 분포)

- 핵심 : 모든 random variable은 probability distribution을 내포하고 있다 ex) 키를 X라고 하면 측정횟수를 크게 하면 확률분포를 가지기 때문이다. (Random variables are variables whose possible values are numerical outcomes of a random phenomenon)



- 형태 : table, graph, equation



- graph 종류

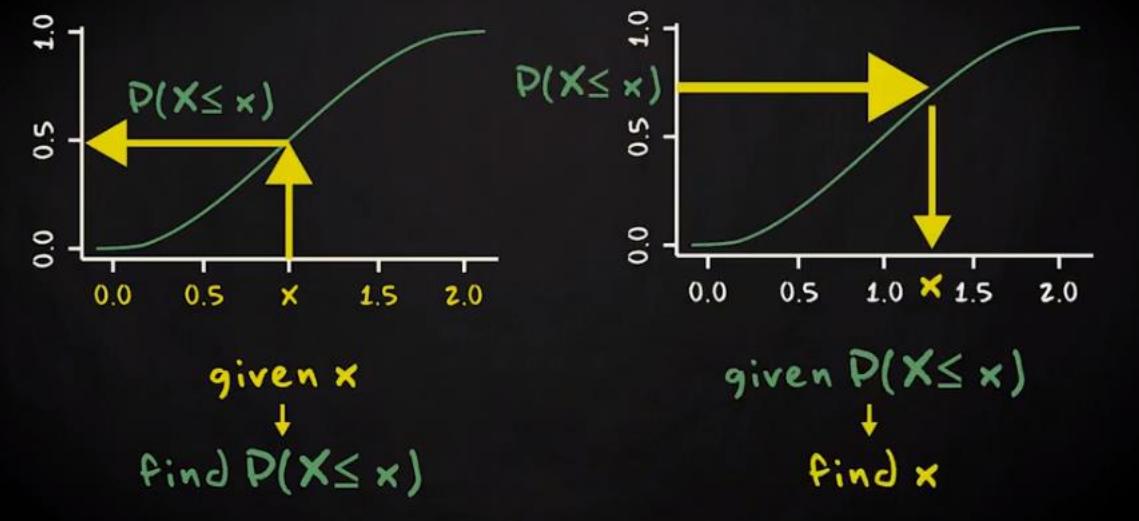
(1) probability mass function (**discrete probability distribution**) : discrete random variable은 y축이 probability이다.

(2) probability density function : continuous variable은 y축이 확률 밀도이다. → 특정 확률을 구하려면 integral을 해야 한다.

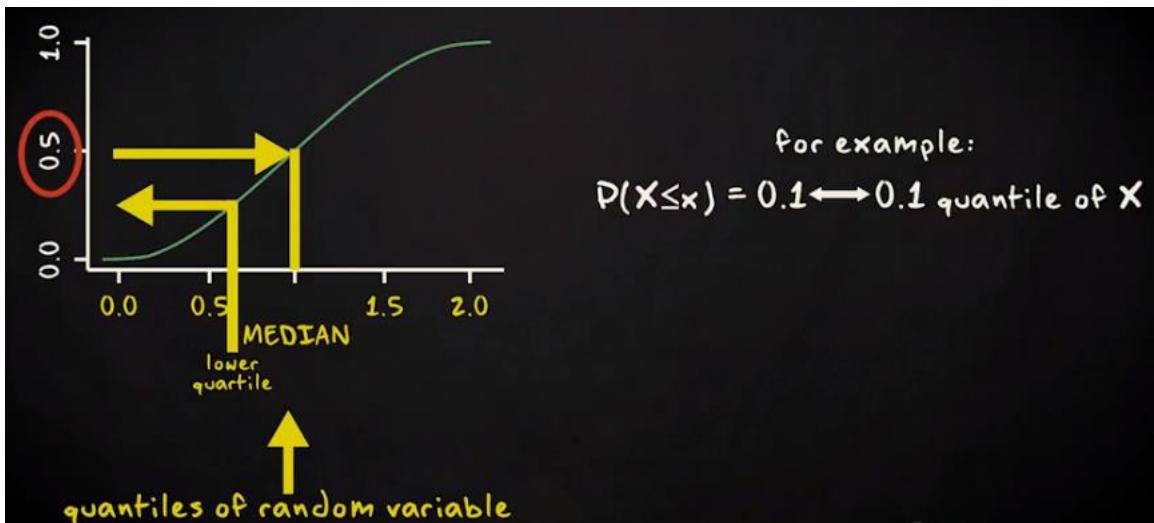
(3) cumulative probability distribution

→ 2가지 용도 : threshold value of the random variable (quantile)

cdf answers two questions:



- Quantile : 0.1 quantile이란 threshold value를 의미하며, 그 threshold value 혹은 그 이하의 값을 가질 확률이 0.1인 것을 의미한다. 따라서 이 누적 확률 그래프는 quantile을 x축으로 한다고 볼 수 있다.



3.2. Univariate analysis : 확률분포의 묘사 (summary statistics)

3.2.1. mean : 변수 값과 각각의 확률을 곱해서 더하면 된다.

<p>the mean of a discrete random variable</p> <p>=</p> <p>probability-weighted average of all possible values that the random variable can take</p> <p>=</p> $\mu_x = E(x) =$ $= x_1 P(x_1) + x_2 P(x_2) + \dots + x_k P(x_k)$ $= \sum [x_i P(x_i)]$	<p>the mean of a continuous random variable</p> <p>=</p> <p>average of all possible values that the random variable can take</p> <p>=</p> $\mu_x = E(x) =$ $\int_{-\infty}^{\infty} x f(x) dx$
--	--

- mean의 연산

(1) a + b X이면?

$$\mu_{a+bX} = a+b\mu_X$$

EXAMPLE

Nr. of stops	0	1	2	3
Waiting time (min)	0	2.5	5	7.5
Probability	0.3	0.4	0.2	0.1

$\mu_{\text{waiting time}} = (0 \cdot .3) + (2.5 \cdot .4) + (5 \cdot .2) + (7.5 \cdot .1) = 2.75 - 2 \text{ min. and } 45 \text{ sec.}$

(2) 확률 변수를 더하면(빼면) 평균은? 더하면 된다. (independent하더라도 가능하다 평균은)

Mean waiting time on a day = 1.75 min. \rightarrow Mean waiting time for 7 days?

$$\begin{aligned} \mu_{\text{mon+Tue+Wed+...}} \\ = \\ \mu_{\text{mon}} + \mu_{\text{Tue}} + \\ \mu_{\text{Wed}} + \dots \\ = \\ 1.75 * 7 \end{aligned}$$

3.2.2. variance :

- 구하는 법 : 편차를 양수로 만들기 위해서 제곱한 다음에 평균을 낸다(확률을 곱한다)

variance of a random variable $X \rightarrow \text{var}(X)$

$$= E [(X - \mu)^2]$$

continuous X

$$\int (x - \mu)^2 * f(x) dx$$

discrete X

$$\sum (x_i - \mu)^2 * P_i(x_i)$$

- 연산

(1) $a + bX$ 이면 분산은 b^2 배가 된다.

(2) $X_1 + X_2$ 이면 더하면 된다.

ADDING RANDOM VARIABLES X AND Y

$$\begin{aligned}\text{var}(X + Y) &= \text{var}(X) + \text{var}(Y) + 2\text{cov}(X, Y) \\ \text{var}(X - Y) &= \text{var}(X) + \text{var}(Y) - 2\text{cov}(X, Y)\end{aligned}$$

Cf) 만약 공분산을 무시한다면, 변수끼리 더하면 분산도 더하면 된다.

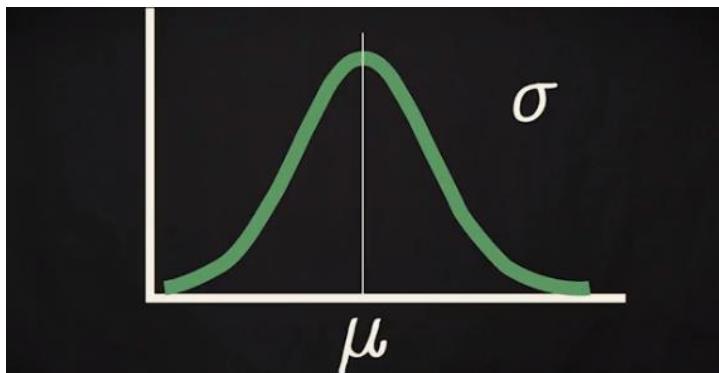
$$\begin{aligned}\text{var}(X + Y + Z + \dots) &= \\ \text{var}(X) + \text{var}(Y) + \text{var}(Z) + \dots &\leftrightarrow \\ \text{var}(\sum_{i=1}^n X_i) &= \sum_{i=1}^n \text{var}(X_i)\end{aligned}$$

(3) $aX_1 + bX_2$ 이면

$$\begin{aligned}\text{Var}(aX + bY) &= a^2\text{var}(X) + b^2\text{var}(Y) + 2ab\text{cov}(X, Y) \\ \text{Var}(aX - bY) &= a^2\text{var}(X) + b^2\text{var}(Y) - 2ab\text{cov}(X, Y)\end{aligned}$$

3.3. 정규 분포(normal distribution) = gaussian distribution

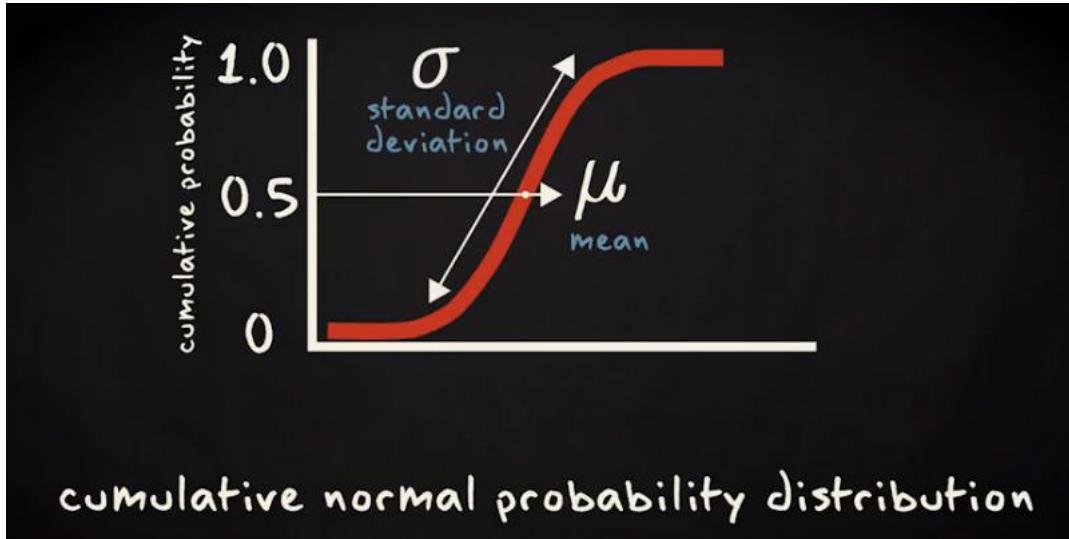
3.3.1. parameters : 평균(mean)과 standard deviation(spread) $X \sim N(\mu, \sigma^2)$



$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

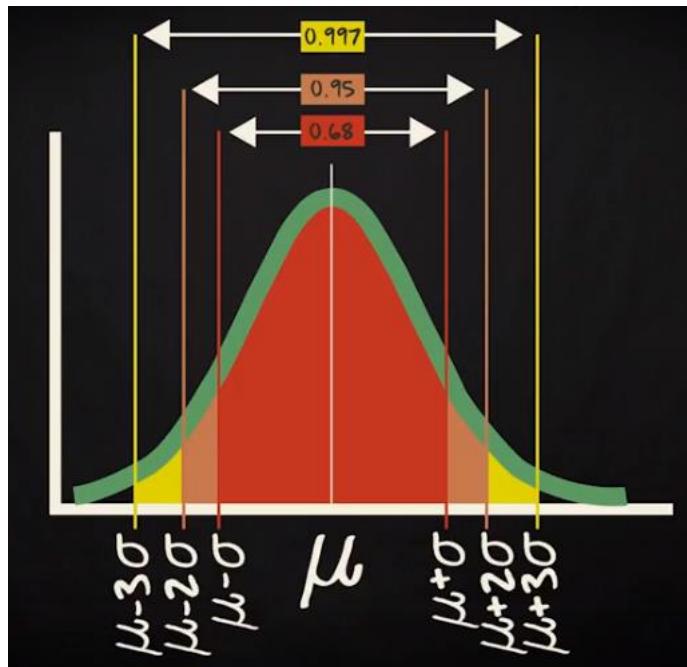
- 특징 : $\frac{1}{\sigma\sqrt{2\pi}}$ 를 곱하면 전체 넓이가 1이 된다. 곱하지 않으면 시그마에 따라서 넓이가 달라진다.
높이가 $\frac{1}{\sigma\sqrt{2\pi}}$ 인 점도 신기하다.

- cumulative normal probability distribution



3.3.2. 확률의 계산!!!

(1) 빠른 방법 : 평균과 표준편차만을 가지고 구할 수 있다!!



Ex) 중간고사에서 평균 80 표준편차 5이면, 90점을 맞은 나는 A+를 맞을 수 있을까?

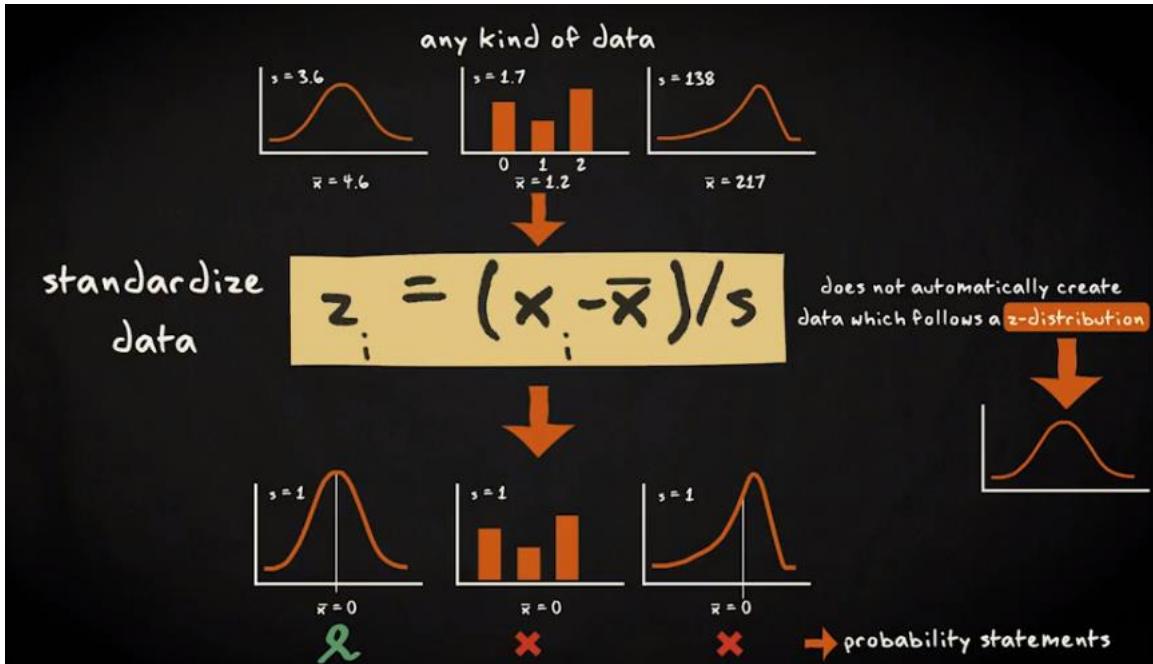
68-95-99 법칙 (1, 2, 3 시그마) 90점이면 2시그마이기 때문에 95%, 즉 양옆으로 47.5%가 된다. 따라서 너는 상위 2.5%라는 것을 알 수 있다.

(2) 좀 더 엄밀한 방법 : Z분포로 표준화해서 구할 수 있다. 빼평나표로!

Ex) geese travel이 4day 평균, 1.3 days 표준편차일때, 6일 내로 이동하는 확률은?

Ex) geese의 10% quantile은 몇일 내로 이동할까? Z를 표에서 구하면 -1.28이다. 따라서 $X-4/1.3=-1.28$ 따라서 2.34 days이다.

Cf) 주의사항 : 어떤 데이터라도 빼평나표를 통해서 서로 다른 변수를 비교할 수 있다. (활용 가능!) 그러나 그것이 표준정규분포라는 것을 알 수는 없다.



3.3. 이항분포(binomial distribution)

(1) 조건

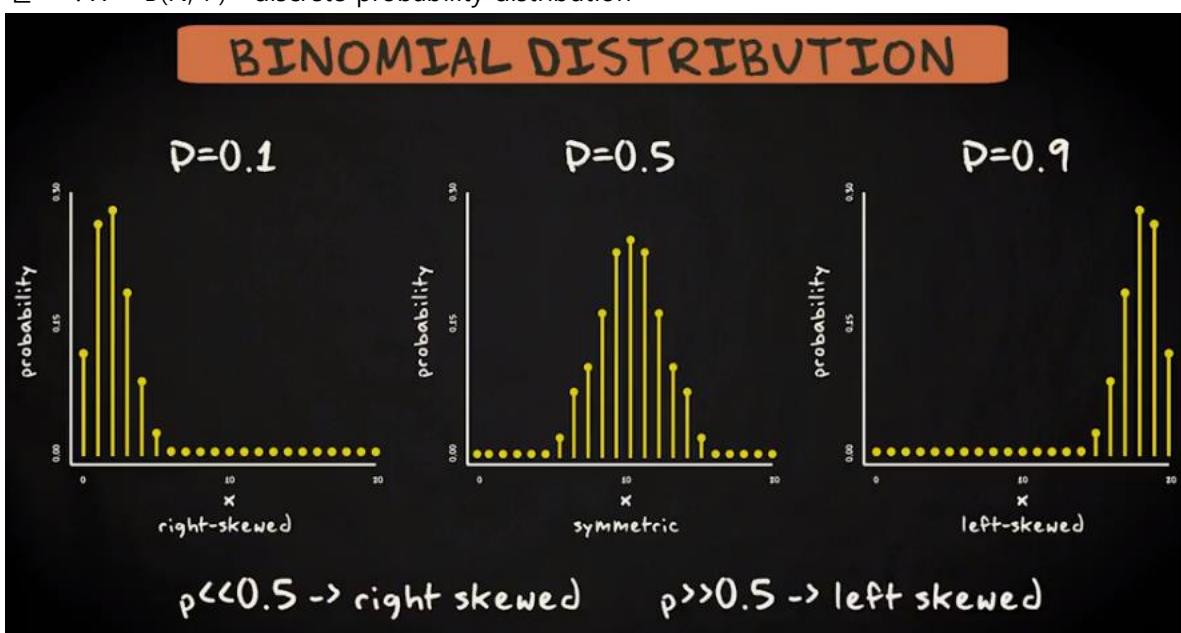
- 변수가 다르다! : two mutually exclusive outcomes (success or failure) & fixed probability P (Bernoulli trial)
- 각 시행이 독립! : Independence between trials

(2) 변수 : $X = \text{number of success!}$

(3) 계산

$$P(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}, \quad x = 0, 1, 2, \dots, n$$

(4) 이항 분포 : $X \sim B(N, P)$ discrete probability distribution



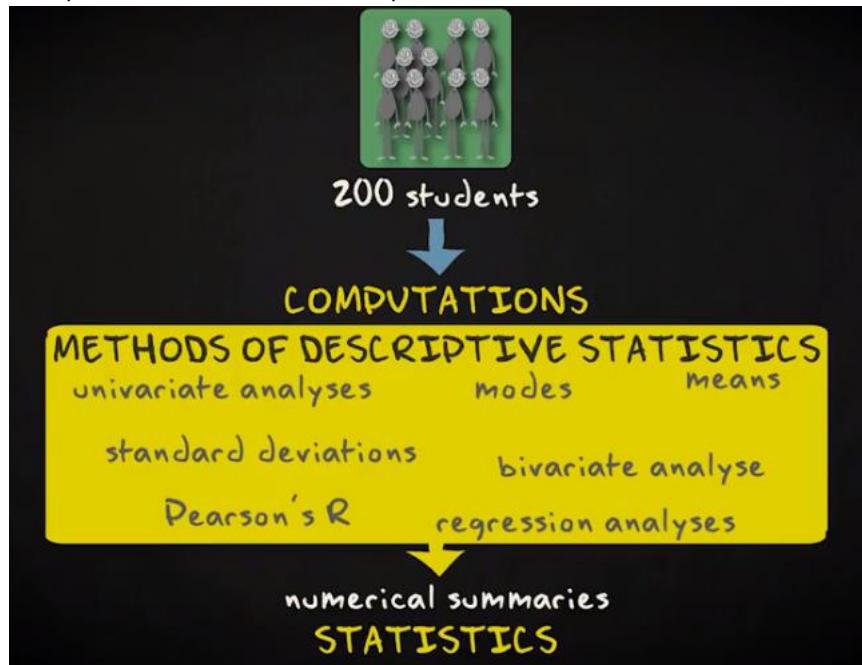
$$\begin{array}{ll} \mu = g(p, n) & \sigma = f(p, n) \\ \mu = np & \sigma = \sqrt{np(1-p)} \end{array}$$

제 2 장. 추론 통계학 (inferential statistics)

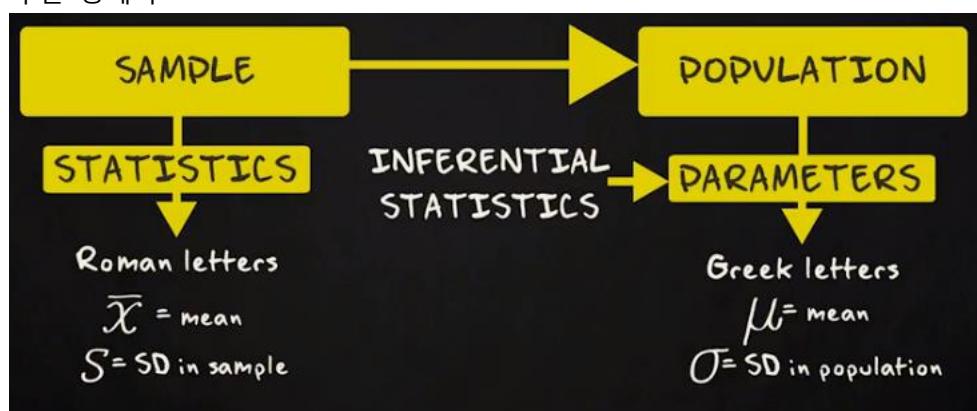
4. Sampling (표본 추출)

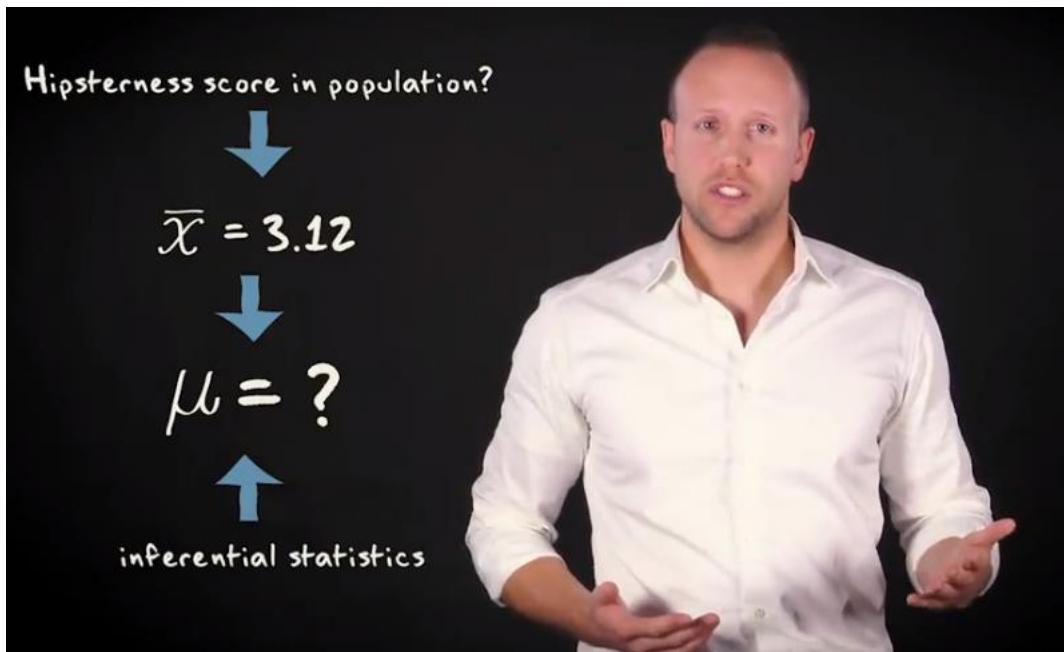
4.1. sampling의 주요 목적 = population 모분포 추론

Cf) 1장 ~ 3장까지는 sample에 대한 분석 → "descriptive statistics 기술 통계학" ex) univariate, bivariate analysis



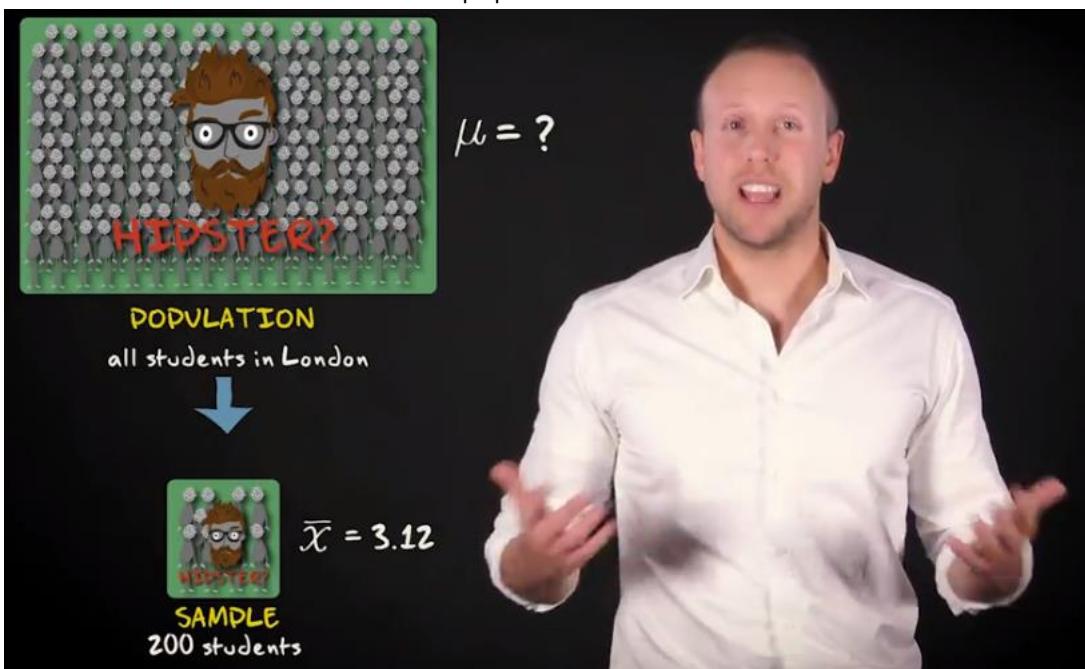
- sampling의 궁극적인 목적 : sample의 statistics를 통해서 population parameter에 대해서 알고 싶은 것이다.
→ inference 추론 통계학





4.2. sampling method (방법)

- sample의 조건 : micro version of the entire population이면 좋다.

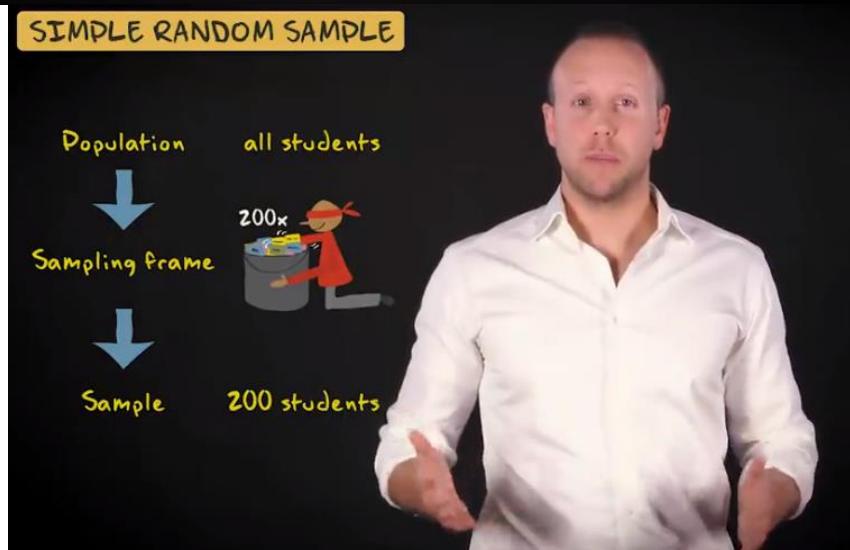
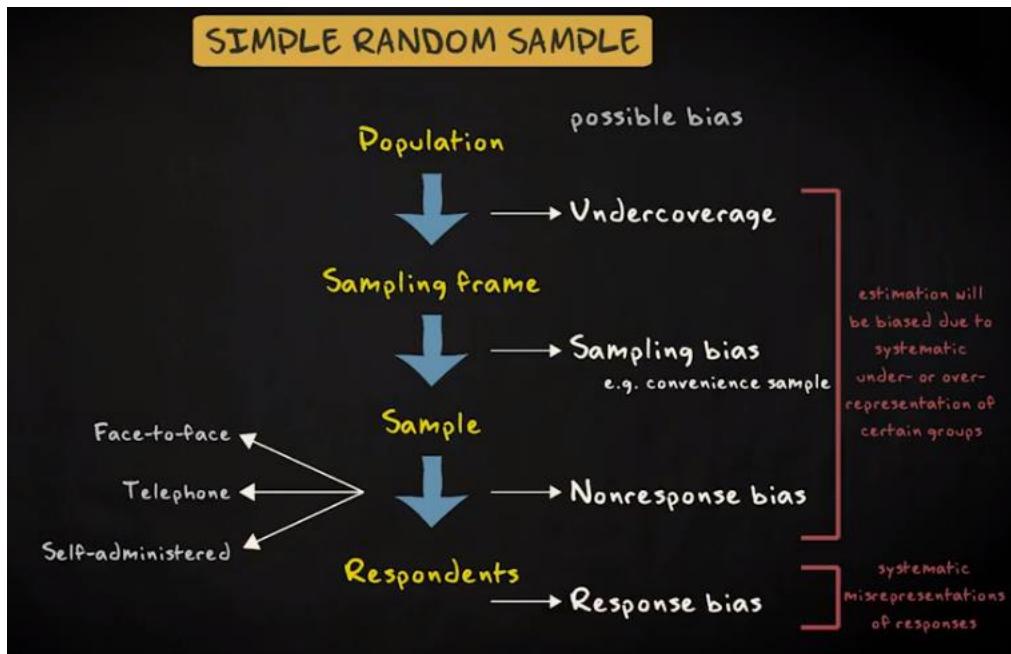


(1) simple random sample : each subject has a same chance of being selected

- 순서 : 목표 Population을 정하고 → sampling frame (contact list가 있다) 얻어서 → random하게 sample 추출 → response 얻기 (face-to-face, telephone, self-administered)

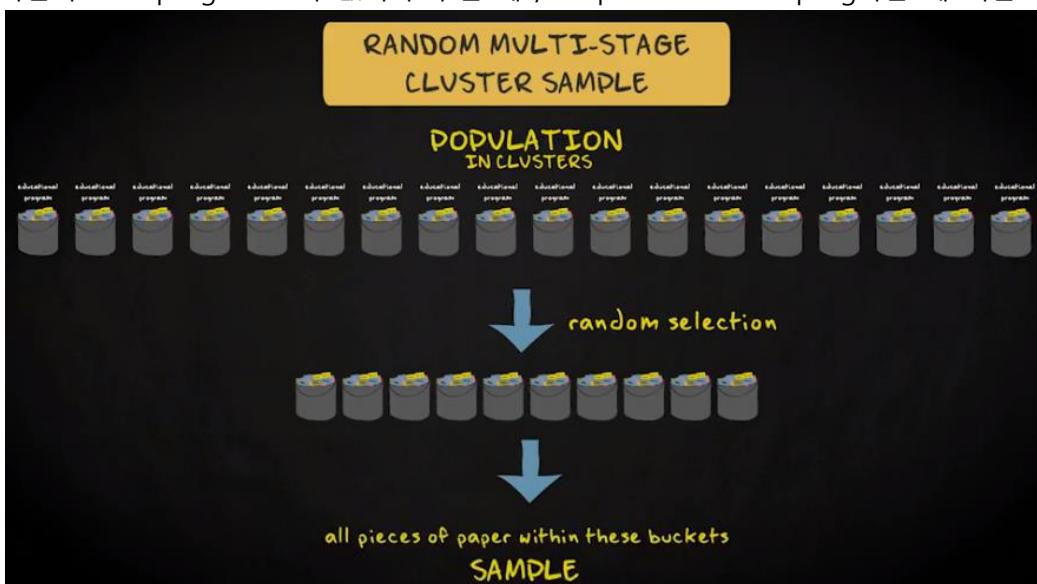
- bias의 발생 가능성 :

- 1) undercoverage : sampling frame에 없는 구성원 존재 가능성(email 누락 등)
- 2) sampling bias : 길거리에서 표본 구할 때, 집돌이/순이들을 배제하는 오류
- 3) nonresponse bias : 무응답 편의
- 4) response bias : trump 지지자라고 대놓고 말하기가 부담스러워서 못하는 경우 등



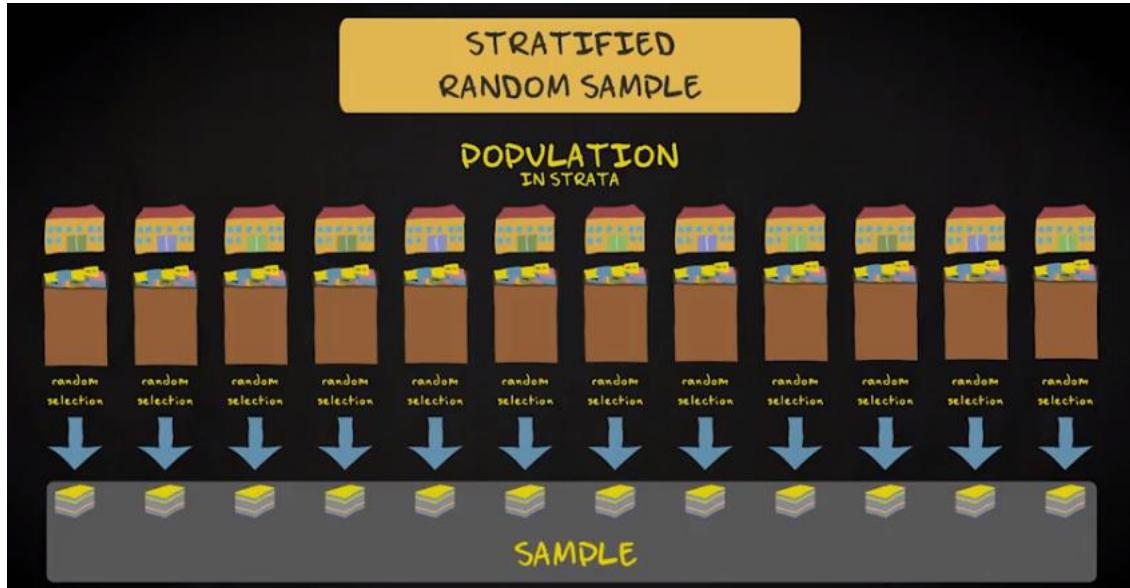
(2) random multi-stage cluster sampling

- 순서 : 많은 수의 cluster 중에서 특정 수의 cluster를 random하게 추출한다. 그 종이들을 다 모아서 그중에서 sampling을 한다. (각 cluster에서 각각 뽑는 것이 아니다 → stratified와의 차이점)
- 언제하는가 : sampling frame이 없거나 후질 때 / simple random sampling하는 게 비쌀 때



(3) stratified random sampling

- 순서 : population을 각 학교로 나누고, 각각의 strata에서 sampling을 한다. (cluster을 random하게 추출한다는 점, 각 strata 단위에서 random sampling을 한다는 점에서 multistage cluster sampling과 다르다.)



- 언제하는가 : 각 stratum에서 충분한 sample을 얻고 싶을 때 한다. 단, 각 sample이 어느 stratum에 속하는지를 알아야 한다.

Cf) 주의할 점

- 1) 표본 수가 클수록 좋다. 그러나 bad sampling procedure면 다 소용없다.
- 2) 표본 수가 일정 수준 이상을 넘어가면 추측의 정확성이 더 이상 크게 증가하지 않는다.



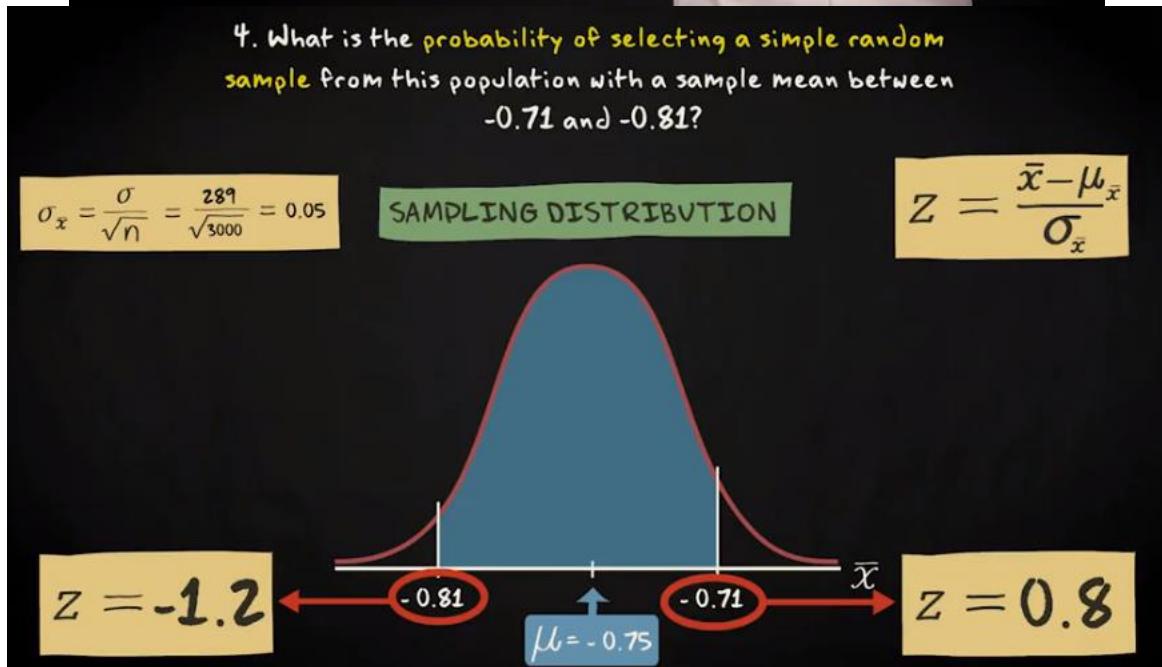
4.3. Sampling distribution, (표본평균의 분포) data/sample distribution, population distribution

4.3.1. sampling distribution of the sample mean

- 개념의 목적 : link that helps draw conclusions about a population on the basis of **only one sample**.

주의) sample distribution이랑 다르다. 이것은 30개의 표본일 때 그 30개 값의 분포를 나타냄.

→ 표본평균의 분포는 정규분포를 따르고, sampling distribution의 평균과 표준편차를 알기 때문에, 이 정보를 사용할 수 있다.



→ sample의 평균의 범위가 어느 확률을 갖는지 알고 싶으면 이 sampling distribution 분포를 사용하면 된다.

- sampling distribution(표본평균의 표본분포)의 정의 : n개의 sample을 무한번 sampling했을 때의 각 sample의 표본평균의 분포.

→ 특징 : 1) 표본평균의 분포는 정규분포가 되며 ← 그 이유는 Central limit theorem 때문이다.

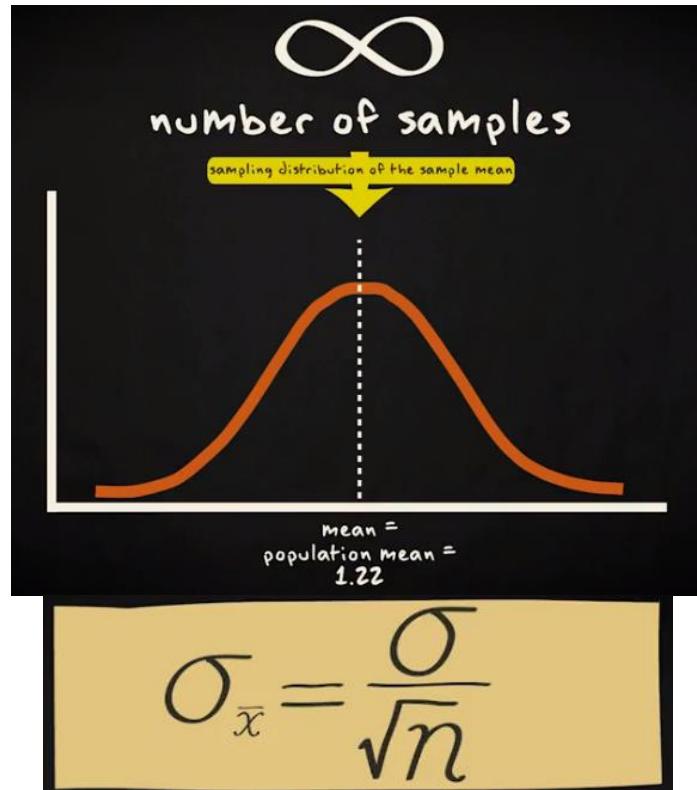
2) 표본평균의 분포(sampling distribution)의 평균이 모 평균에 수렴한다. $E(\bar{X}) = \mu$

= sampling distribution의 mean $E(\bar{X})$ 은 population mean가 같다. (μ)

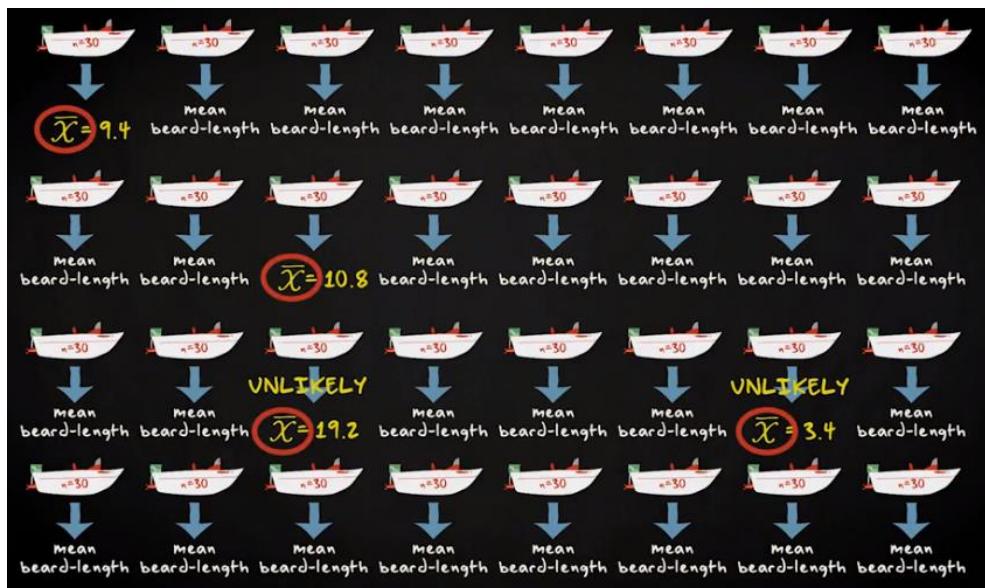
3) 표본평균의 분포(sampling distribution)의 표준오차 $\sigma(\bar{X}) = \frac{\sigma}{\sqrt{n}}$ σ 는 모분포(X)의 표준편차

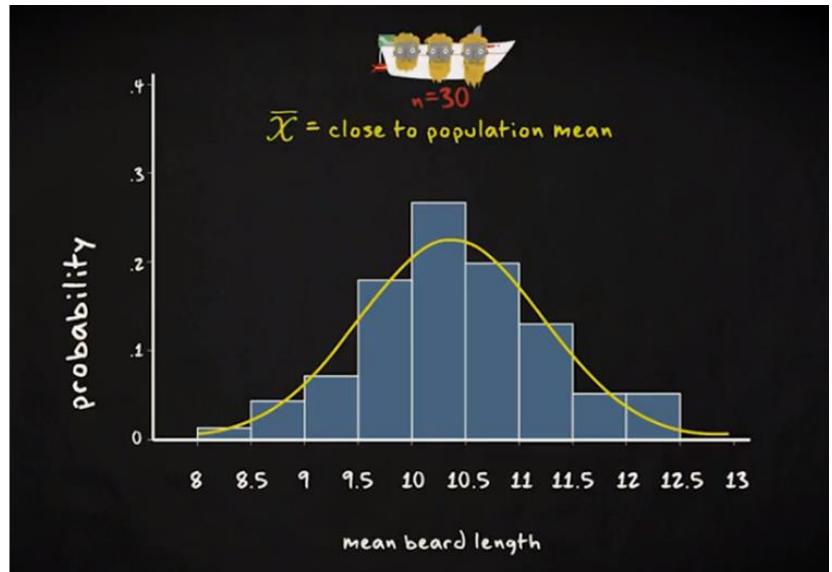
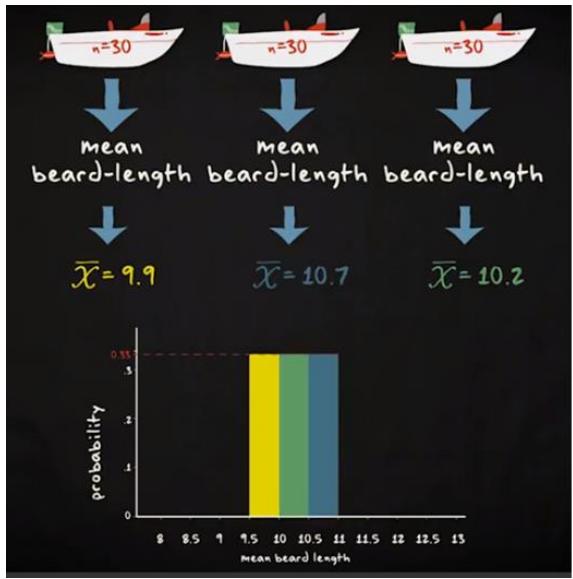
→ 생각해보면 맞는 말이다. 모분포의 표준편차가 클수록 sample mean의 표준오차가 커지는 경향이 있을 것이다. 또한 표본수(sample size)가 커질수록 sample mean간의 편차가 줄어들 것이다.

Cf) 표준편차는 관측치 1개에 담긴 불확실성의 크기이나, (표본분포의) 표준오차는 관측치 n에게 담긴 불확실성의 크기이다.



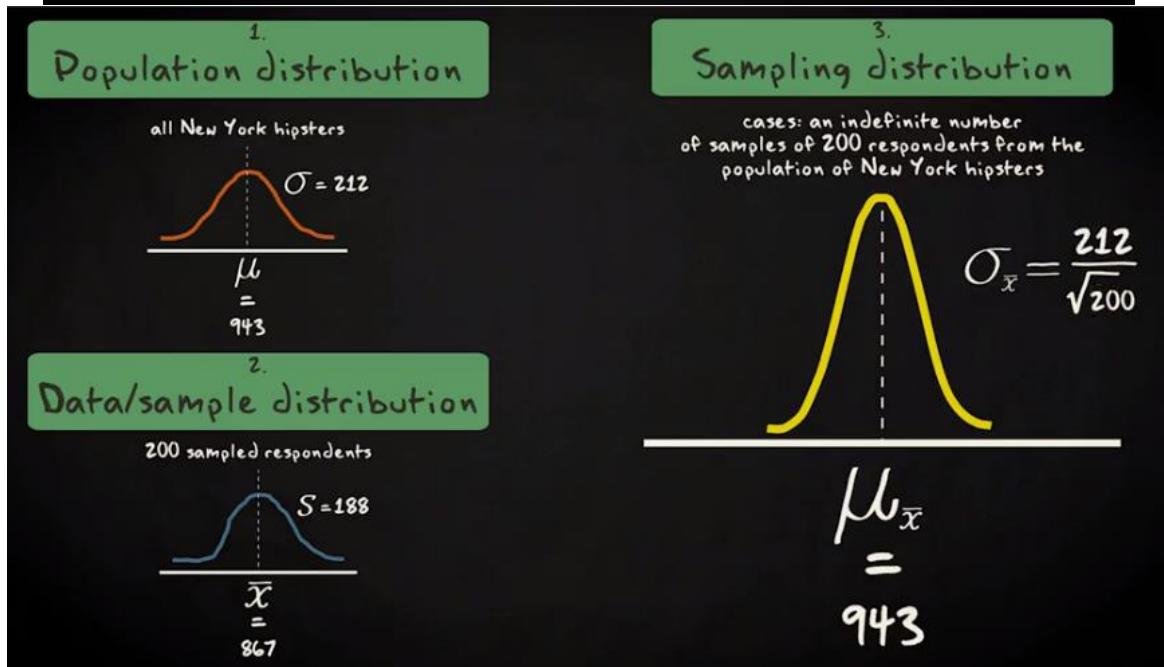
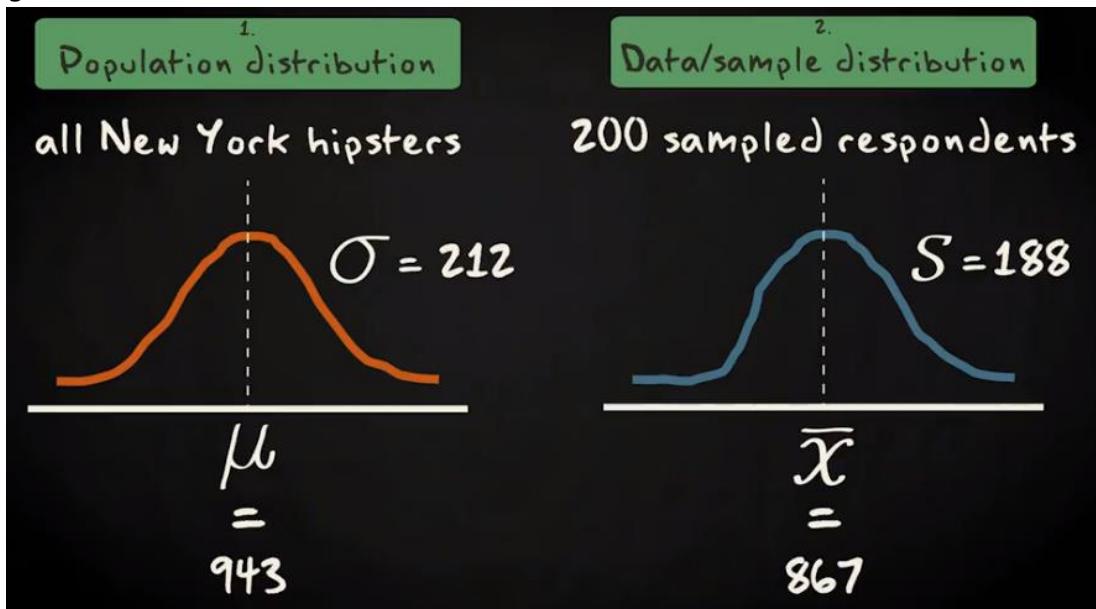
- 예시 : 표본을 3번 뽑아서 표본평균값 3개를 확률분포도로 만들었을 때 → 100개 → 표본을 무한개 뽑아서 무한개의 표본 평균을 확률분포도로 만들었을 때의 변화를 볼 수 있다.





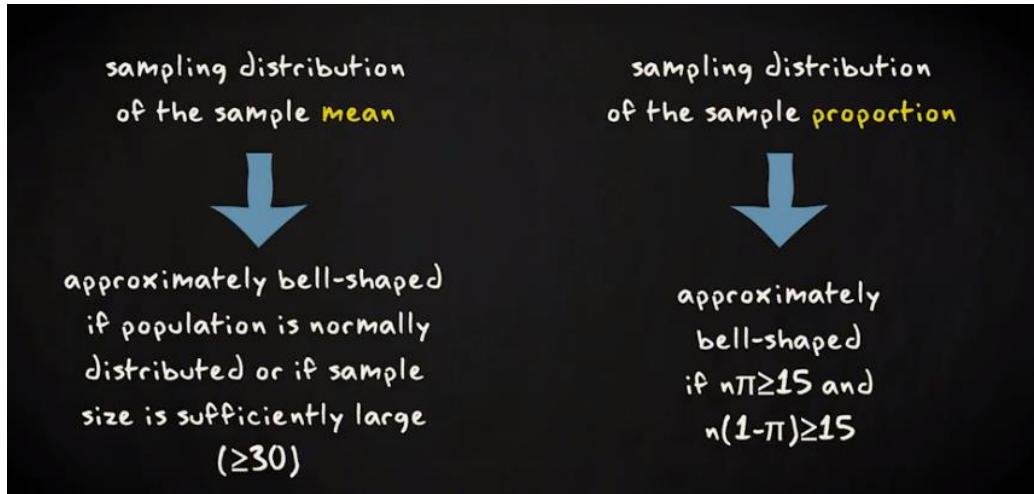
4.3.2. population distribution (모분포)와 Data/sample distribution (표본의 분포)

- sampling distribution은 가상의 분포이다. 표본 뽑는 횟수를 무한번 할 수 없기 때문이다.



4.3.3. sampling distribution of sample proportion (표본비율)

- sample proportion이란, 확률변수 X 가 0과 1을 값으로 가질 때, $\hat{p} = X_1 + X_2 + X_3 + \dots + X_n / n$ 을 의미한다. 즉 표본평균(sample mean)이랑 같다. 다만 X 가 0과 1을 가진다는 것만이 다르다.
- 따라서 표본평균과 동일한 논리가 적용된다.



- 표본비율의 분포(sampling distribution of sample proportion)의 정의 : n 개의 sample을 무한번 sampling했을 때의 각 sample의 표본비율의 분포.

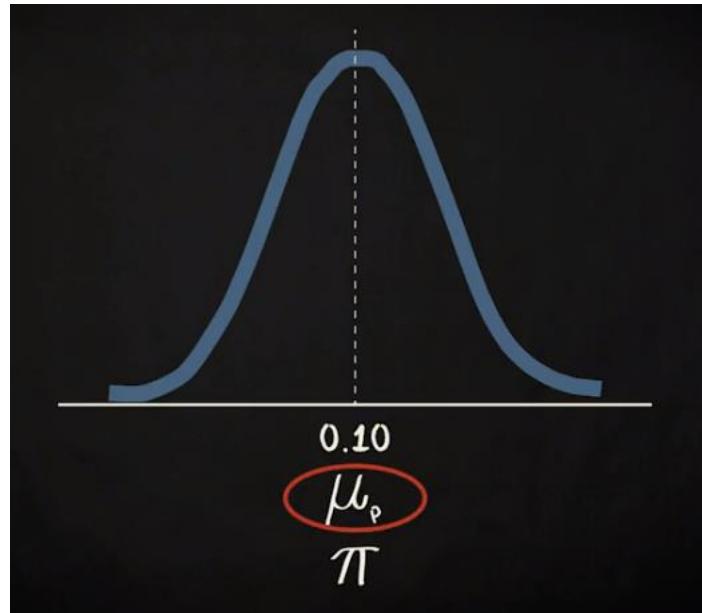
→ 특징 :

1) 표본비율의 분포는 정규분포가 되며 (단, $np \geq 15, n(1-p) \geq 15$ 일 때) ← 그 이유는 Central limit theorem 때문이다.

2) 표본비율의 분포(sampling distribution)의 평균이 모비율에 수렴한다. $E(\hat{p}) = p$

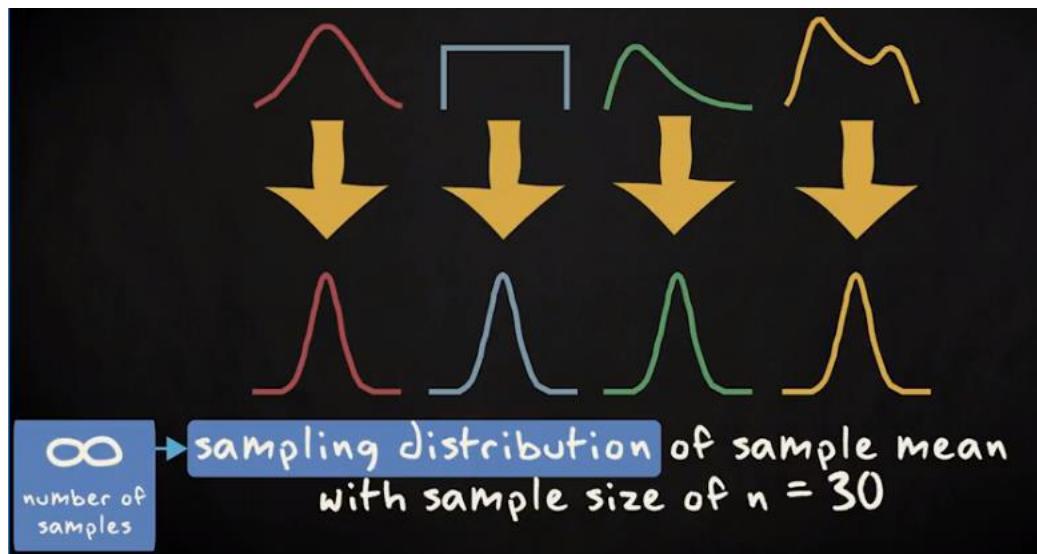
3) 표본비율의 분포(sampling distribution)의 표준오차 $\sigma(\hat{p}) = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{p(1-p)}{n}}$ σ 는 X 의 표준편차





4.4. Central limit theorem

- 중심극한정리 : n (sample size)가 충분히 크면($n >= 30$), sampling distribution of sample mean(= \bar{X} 의 분포)은 normal distribution를 가진다. (확률변수(X)의 분포가 정규분포가 아니더라도 성립한다.)
= 다시 말하면, n 개의 sample을 무한번(횟수) sampling했을 때 \bar{X} 의 분포는 정규분포를 이룬다. (infinite number of sampling → sampling distribution of mean becomes bell shape)
- 주의) n 은 추출횟수가 아니라 1개의 표본의 개수(sample size)



→ 위의 그림은 확률변수 X 의 population distribution인데, 밑의 그림은 \bar{X} 의 분포(sampling distribution of mean)이다. X 가 정규분포를 이루지 않더라도 \bar{X} 의 분포는 정규분포를 이루는 것을 볼 수 있다.(표본수 30 이상일 때)

9. Resampling without sampling

"Sampling distribution 의 대체재는 없을까?"

(근후님 가르침) Bootstrapping 이랑 permutation 을 하는 이유는 무엇이냐면, 위에서처럼 쉽게 sampling distribution 을 구할 수가 없기 때문이다. 따라서 sample 이 참일 것이라고 가정을 하고 뺏튀기, 혹은 점을 무수히 찍거나(bootstrapping), 두집단이 큰 차이가 없을 것이라고 가정하고 그룹 배열을 바꾸거나(permutation)해서 이 문제를 해결하는 것이다.

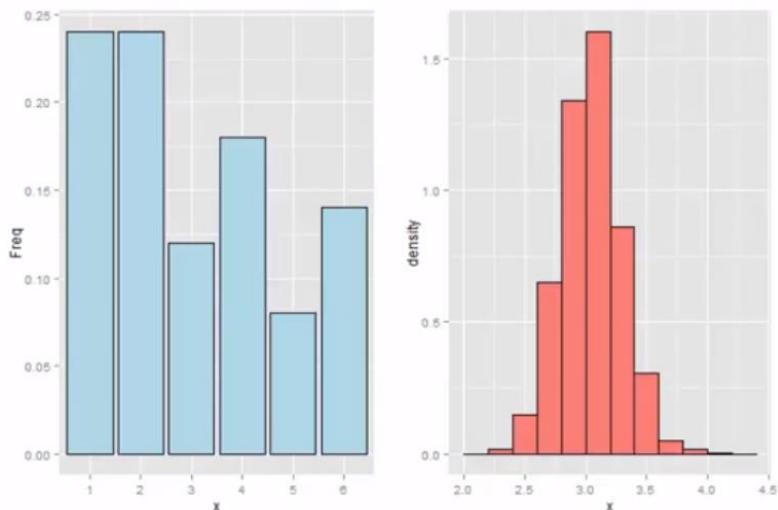
9.1. Bootstrapping

9.1.1. 필요성 : constructing (1) confidence intervals and calculating (2) standard errors for difficult statistics.

9.1.2. 정의 / 방법 : 한번의 sampling data를 가지고 있다면(sample size 50) → 이 정보만을 가지고 sampling distribution of mean(아니면 다른 statistic도 가능)을 구하는 과정이다. 즉, 50개의 sample을 주머니에 넣어놓고 50번씩 sample을 꺼내서 (with replacement, 꺼내고 다음에는 이를 다시 집어넣고 또 뽑는다.) 평균을 무한번 구해보는 과정을 하면 된다. (simulating complete data sets from the observed data with replacement)

→ 결국에 bootstrapping은 한정된 observed data를 가지고 sampling distribution을 추정해보는 것이다.

What if we only had one sample?



9.1.3. 구현 :

Consider a data set

```
library(UsingR)
data(father.son)
x <- father.son$hheight
n <- length(x)
B <- 10000
resamples <- matrix(sample(x, n * B, replace = TRUE), B, n)
resampledMedians <- apply(resamples, 1, median)

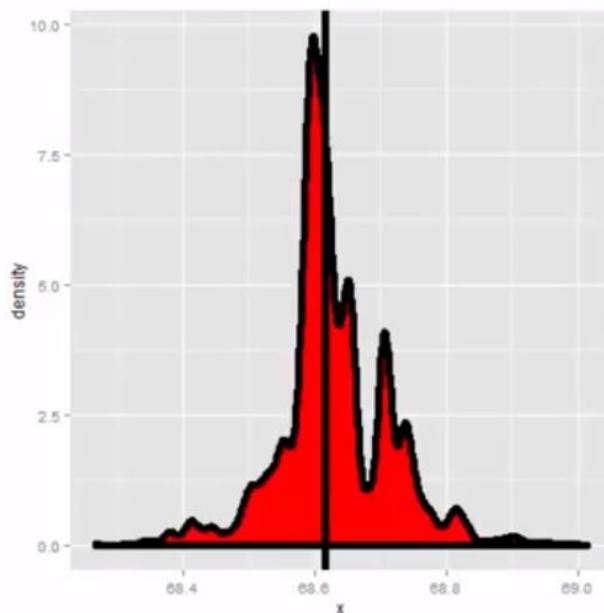
sd(medians)
```

```
## [1] 0.08473
```

```
quantile(medians, c(0.025, 0.975))
```

```
## 2.5% 97.5%
## 68.43 68.82
```

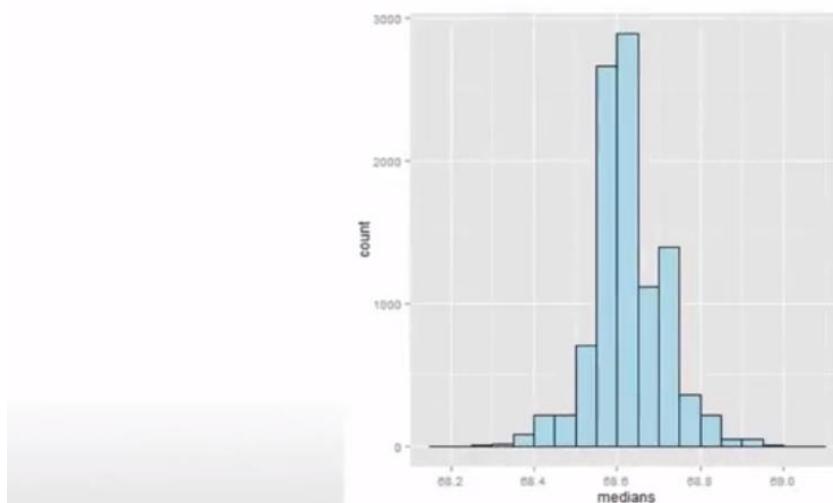
→ bootstrapping을 10000번 실시해서 각 sampling의 median을 그려보는 과정이다. Matrix(sample(x, n*B, replace=TRUE), B, n)의 뜻은 행렬로 B행, n열로 나타내겠다는 것이다.



→ median의 sampling distribution이다. 원래는 수학적으로 엄청 복잡하나, bootstrapping을 통해서 쉽게 구할 수 있었다.

Histogram of bootstrap resamples

```
g = ggplot(data.frame(medians = medians), aes(x = medians))
g = g + geom_histogram(color = "black", fill = "lightblue", binwidth = 0.05)
g
```



11/17

→ sample의 ggplot을 그린 것이다. 테두리, 색깔 채우기 등을 설정하고, g를 치면 히스토그램이 나오게 된다.

9.1.4. 방법 :

Nonparametric bootstrap algorithm example

Bootstrap procedure for calculating confidence interval for the median from a data set of n observations

i. Sample n observations **with replacement** from the observed data resulting in one simulated complete data set

ii. Take the median of the simulated data set

iii. Repeat these two steps B times, resulting in B simulated medians

iv. These medians are approximately drawn from the sampling distribution of the median of n observations; therefore we can

- Draw a histogram of them
- Calculate their standard deviation to estimate the standard error of the median
- Take the 2.5^{th} and 97.5^{th} percentiles as a confidence interval for the median

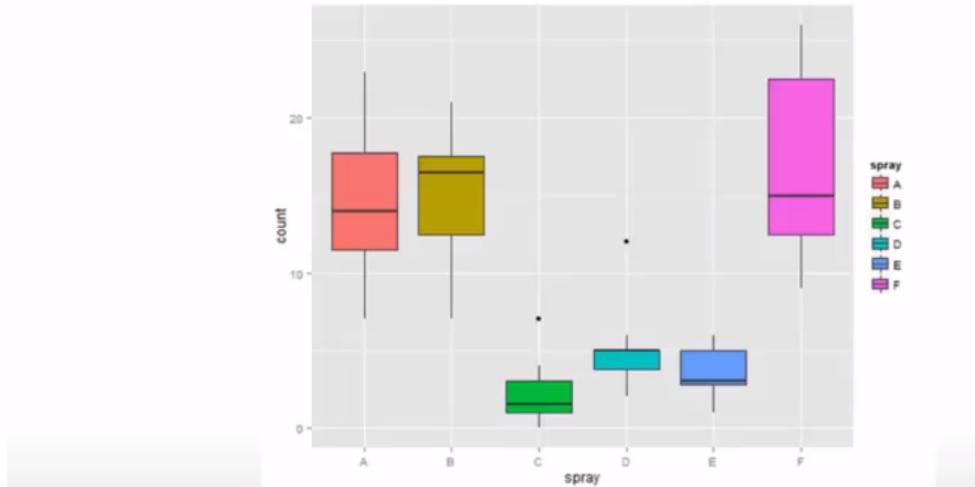
→ 신뢰구간은 bias corrected and accelerated interval (BCA interval)을 2.5, 97.5 대신에 사용하면 더욱 좋다. 책 "An introduction to the bootstrap" by Efron and Tibshirani도 참조하라.

9.2. Permutation test

9.2.1. 필요성 : group comparison할 때 유용한 방법이다. 즉, 유의성 검정(test hypothesis)를 할 수 있는 방법이다. 아래의 그림에서 group B, C pesticide가 유의미하게 다른지를 알아볼 수 있다.

Group comparisons

- Consider comparing two independent groups.
- Example, comparing sprays B and C



9.2.2. 정의 / 방법 : Bootstrapping이랑 비슷하다. 다만, null hypothesis를 먼저 가정하고 (ex) 두 개의 살충제가 효과에 차이가 없다, Sampling data의 observed data를 (관측값, group)에서 group을 랜덤하게 재배열 (permutation)하는 것이다. 그 다음 이를 토대로 observed data보다 더 극단값을 가짐으로서 alternative hypothesis를 지지하는 permutation의 개수를 세어 전체 permutation 횟수로 나누면 P-value를 얻을 수 있다.

9.2.3. 사례 :

Permutation test B v C

```
subdata <- InsectSprays[InsectSprays$spray %in% c("B", "C"),]  
y <- subdata$count  
group <- as.character(subdata$spray)  
testStat <- function(w, g) mean(w[g == "B"]) - mean(w[g == "C"])  
observedStat <- testStat(y, group)  
permutations <- sapply(1 : 10000, function(i) testStat(y, sample(group)))  
observedStat
```

```
## [1] 13.25
```

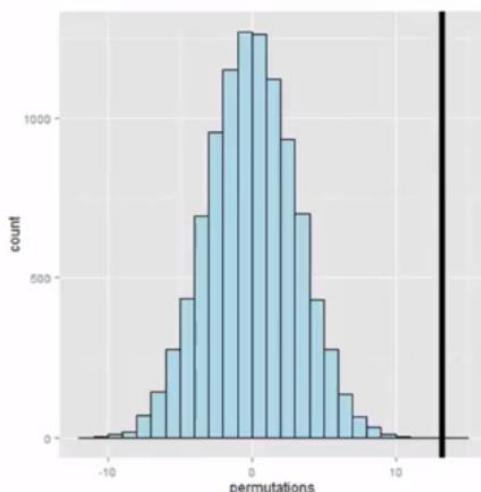
→ 살충제 B와 C는 죽은 벌레의 개수(y)의 평균 차이가 13.25이다.

```
mean(permutations > observedStat)
```

```
## [1] 0
```

→ permutation을 통해 구한 P-value가 0이다.

Histogram of permutations B v C



→ null distribution이다. 유의성 검정을 할 때, sampling distribution(t-분포, Z-분포)과 유사한 기능을 한다. 검은색 선은 observed value를 표시한 것이다.

9.2.4. 응용 :

Variations on permutation testing

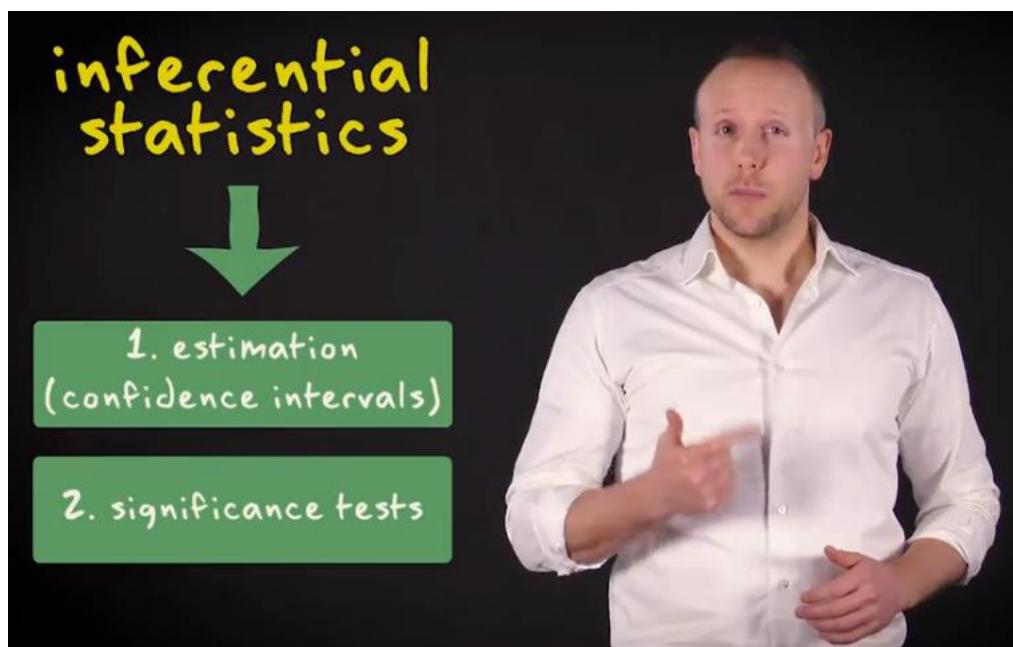
DATA TYPE	STATISTIC	TEST NAME
Ranks	rank sum	rank sum test
Binary	hypergeometric prob	Fisher's exact test
Raw data		ordinary permutation test

- Also, so-called *randomization tests* are exactly permutation tests, with a different motivation.
- For matched data, one can randomize the signs
 - For ranks, this results in the signed rank test
- Permutation strategies work for regression as well
 - Permuting a regressor of interest
- Permutation tests work very well in multivariate settings

5~6. 추론 통계학 (statistical inference methods)

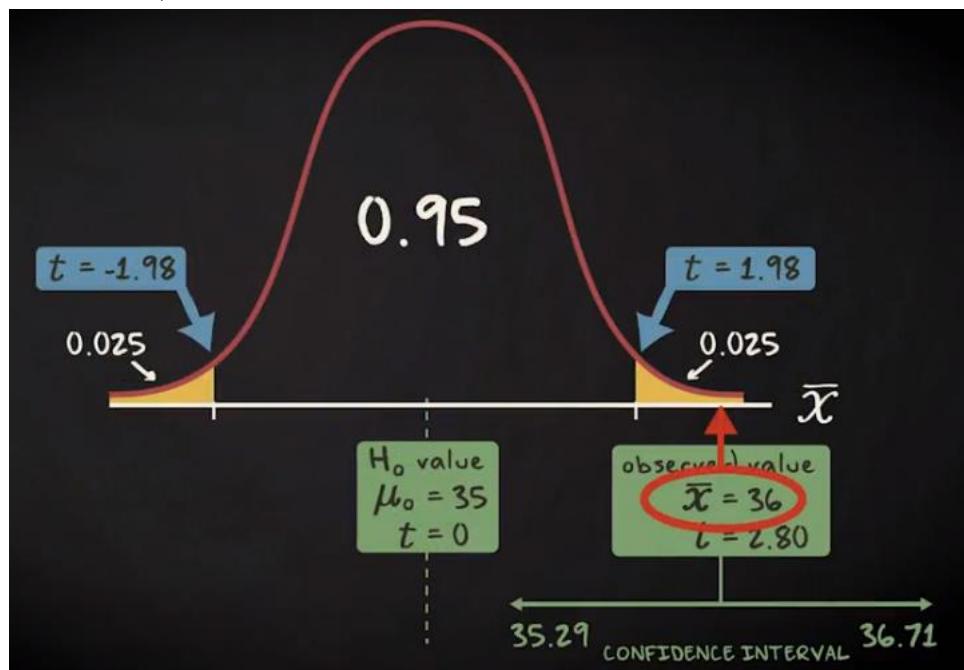
추론 통계학에서는 추정(estimate)과 가설 검정(test hypothesis)를 할 수 있다.

추정에는 점 추정(point estimation)과 구간 추정(interval estimation)이 있는데, 점 추정은 모평균과 점 추정값이 어느 정도 가까운지를 알 수 없다는 한계가 있으므로 구간 추정을 주로 한다.



* 추정과 검정의 관계

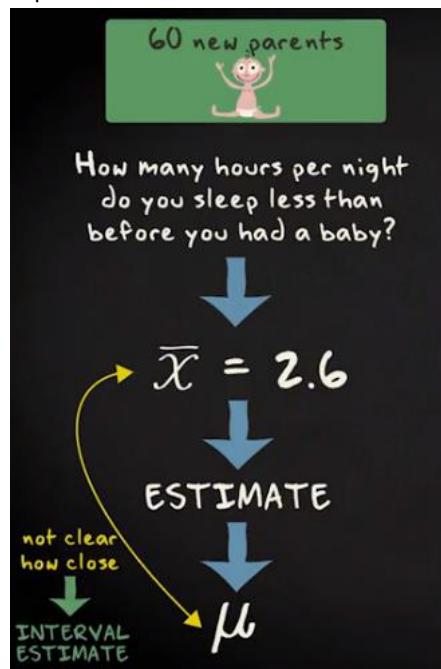
- 신뢰구간 추정(estimation)과 유의성 검정(significance test)은 수학적으로 연관되어 있다.
- P value가 0.05보다 작으면, 95% 신뢰구간은 H₀ value를 포함하지 않는다. 반대도 성립
- P value가 0.05보다 크면, 신뢰구간은 H₀ value를 포함한다. 반대도 성립



5. 추정 (estimation)

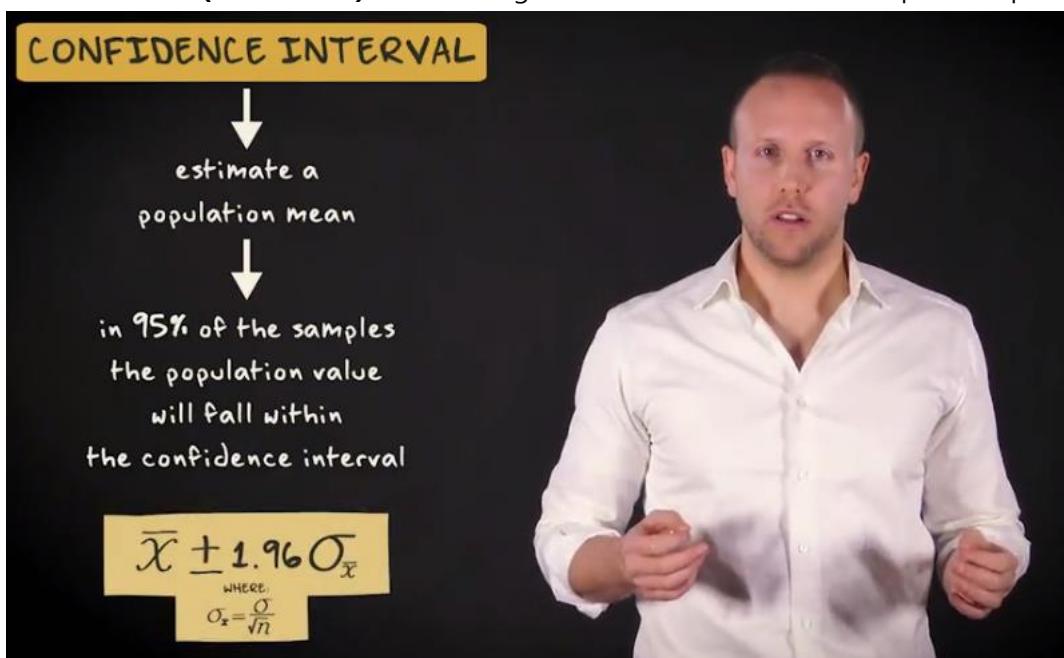
5.1. 점 추정 Point estimate

- 정의 : best guess for the population parameter



5.2. (구간) 추정 (interval estimation)

5.2.1. Interval estimate (구간 추정) 정의 : range of values within which we expect the parameter to fall

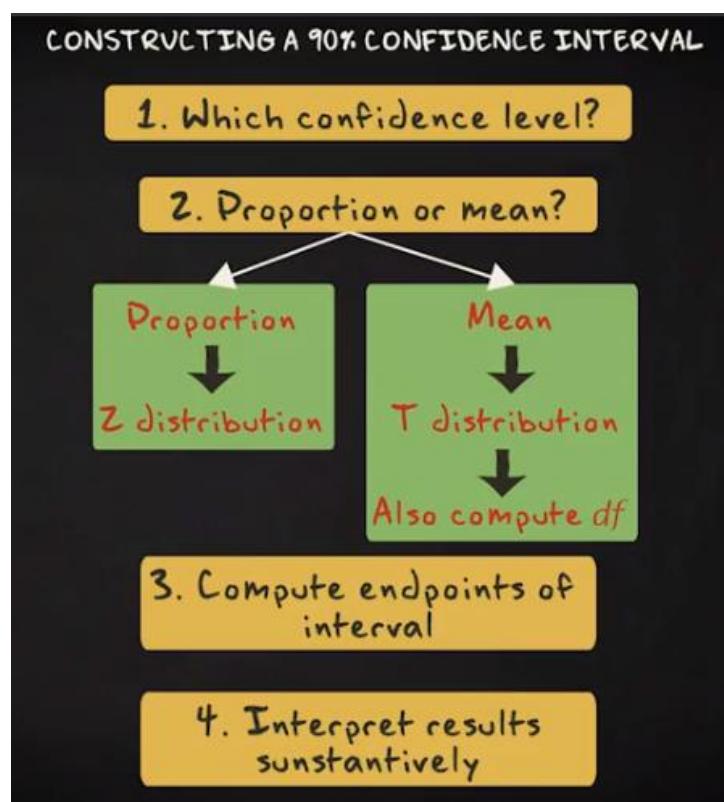


→ sample의 평균, 표준편차, 모분포의 표준편차를 사용해서 신뢰구간을 구할 수 있다.

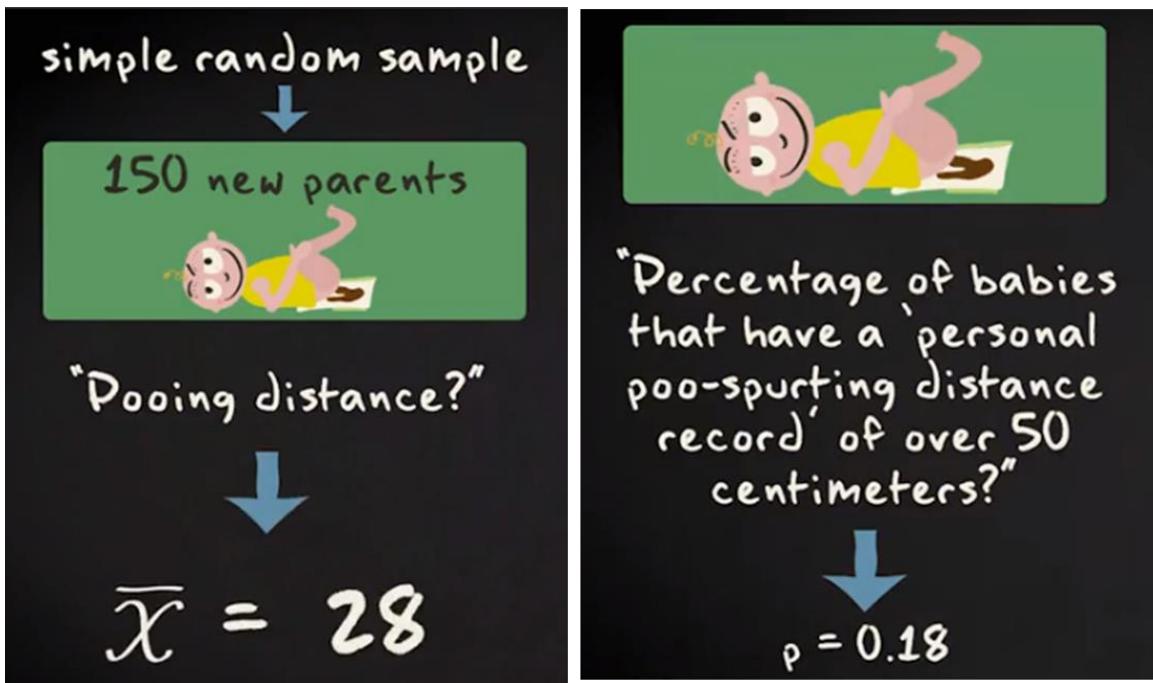
5.2.2. 신뢰구간의 추정 방법

- 단계 : (0) 모비율이냐 모평균 추정이냐?

- (1) 신뢰 수준 결정
- (2) 표본수 결정
- (3) margin of error 결정
- (4) 결과해석



(0) 모비율 or 모평균 추정 : 모비율, 모평균의 추정도 자유자재로 정할 수 있어야 한다.

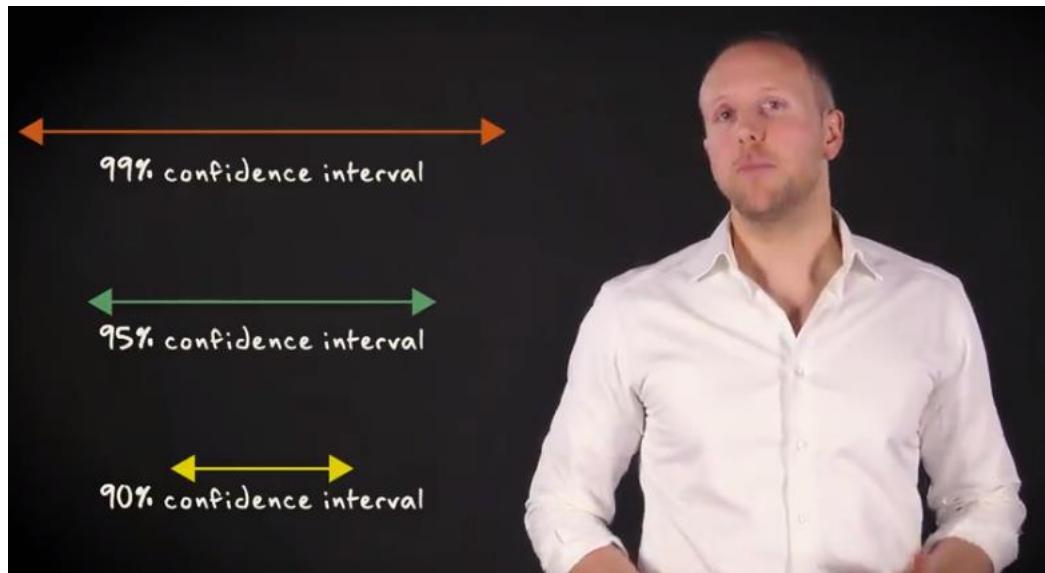


(1) 신뢰수준의 결정

* 신뢰수준 (00% confidence level) : probability that the interval contains the population parameter. Ex) 0.95 → 95% confidence interval

- 해석 : 95% confidence that our point estimate falls within our confidence interval. = 무한번 confidence interval (same margin of error)로 그려면 그 중 95%는 모평균을 포함할 것이다.

- 신뢰수준의 변화 : 신뢰수준이 커질수록 모평균을 더 잘 포함해야 하니까 구간이 커진다. 따라서 confidence랑 precision이랑 중에서 선택을 해야한다. Confidence가 높아지면 precision이 낮아진다.(not very informative)

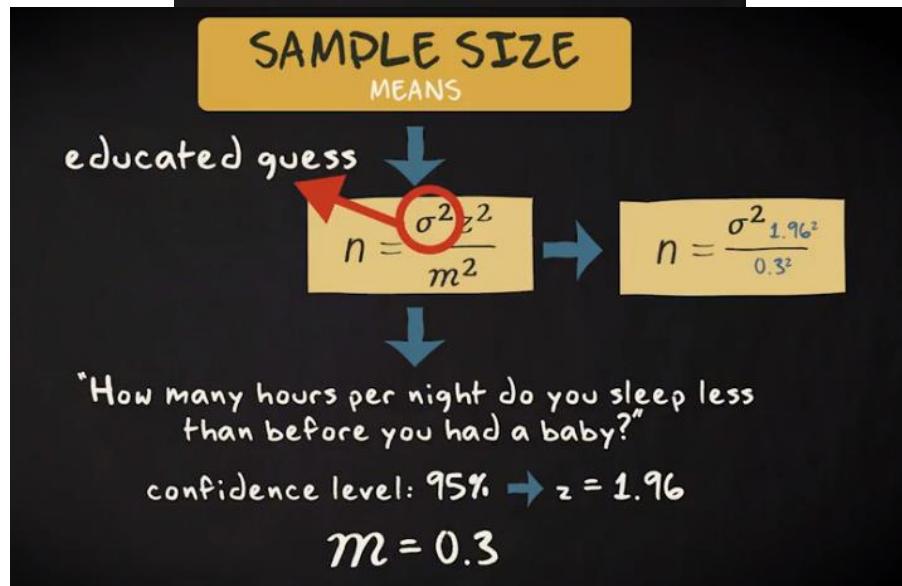
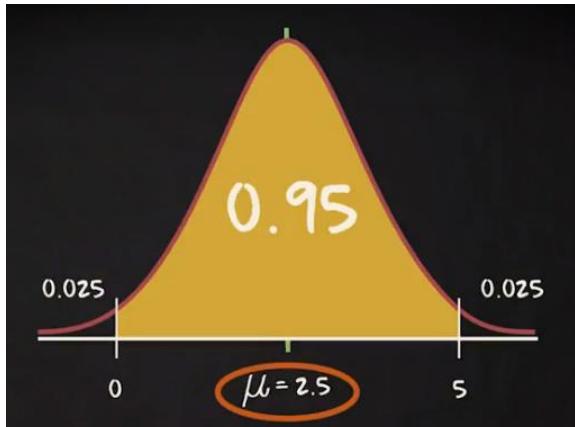


(2) Sample size의 결정 (n)

- 기준 : 1) margin of error : $Z_{95\%} \frac{\sigma}{\sqrt{n}}$ ← margin of error가 작게 하려면 큰 sample size 필요.
- 2) confidence level ← confidence level을 크게 하려면, 큰 sample size가 필요하다. Sample size가 작으면 구간이 지나치게 커져서 추정의 precision이 손해를 많이 본다.
- 3) data의 variability가 크면, sample size가 클수록 좋다. 유의미한 추정을 위해서.
- 4) power를 0.8 이상으로 유지하는게 추천되는데, 이는 $\frac{\sqrt{n}(\mu_a - \mu_0)}{\sigma}$ 에 비례한다.

(2-1) 모평균 추정시 n 을 결정하는 방법 : margin of error(m) = $z * \sigma / \sqrt{n}$ 을 변형해서 $n = \frac{\sigma^2 z^2}{m^2}$

→ $z = 1.96$ (95% 신뢰수준으로 정하면)이고, m 은 0.3정도로 자기가 정할 수 있다. σ 를 정하는 것이 제일 어렵다. 95% 사람들이 0~5hr 정도 잔다고 생각을 해보면 평균이 2.5이고 $\pm 2\sigma$ 에 0과 5가 위치할 테니, σ 는 1.25일 것이라고 추정할 수 있다.

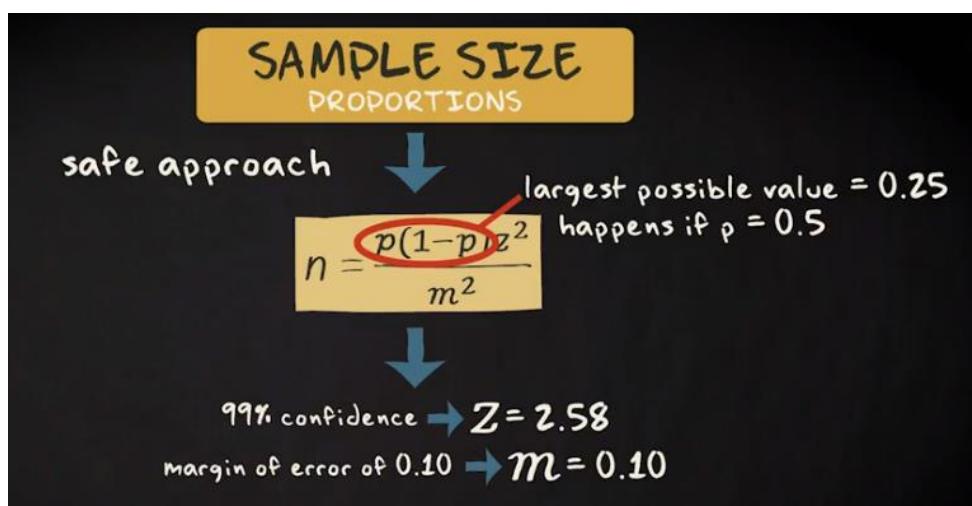


→ 계산하면 최소 n 69명이 필요하다.

(2-2) 모비율 proportion을 추정시 n 을 결정하는 방법 :

- margin of error(m) = $z * \sqrt{\frac{p(1-p)}{n}}$ 을 변형해서 $n = \frac{p(1-p)z^2}{m^2}$

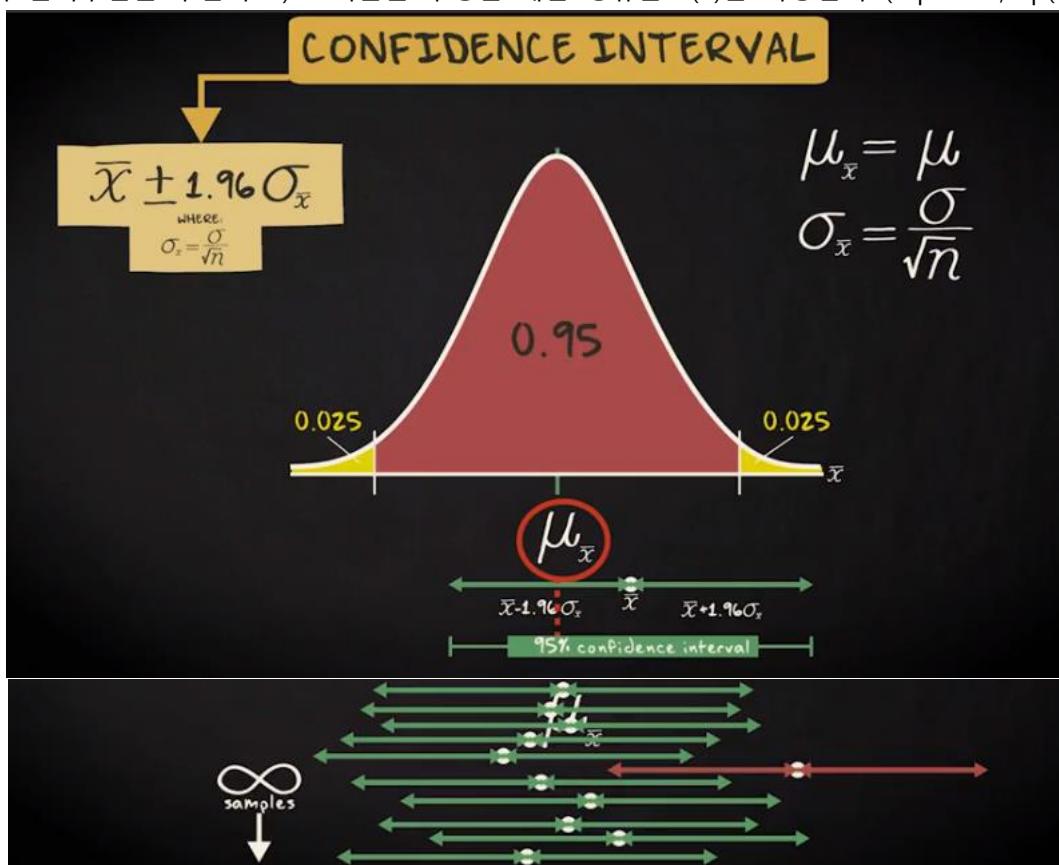
- p 를 결정할 때는 모르므로 previous research를 참고하면 된다. 만약에 모를 경우에 safe approach는 0.5로 하면 된다.



→ 계산하면 약 $n=167$ 이 나온다.

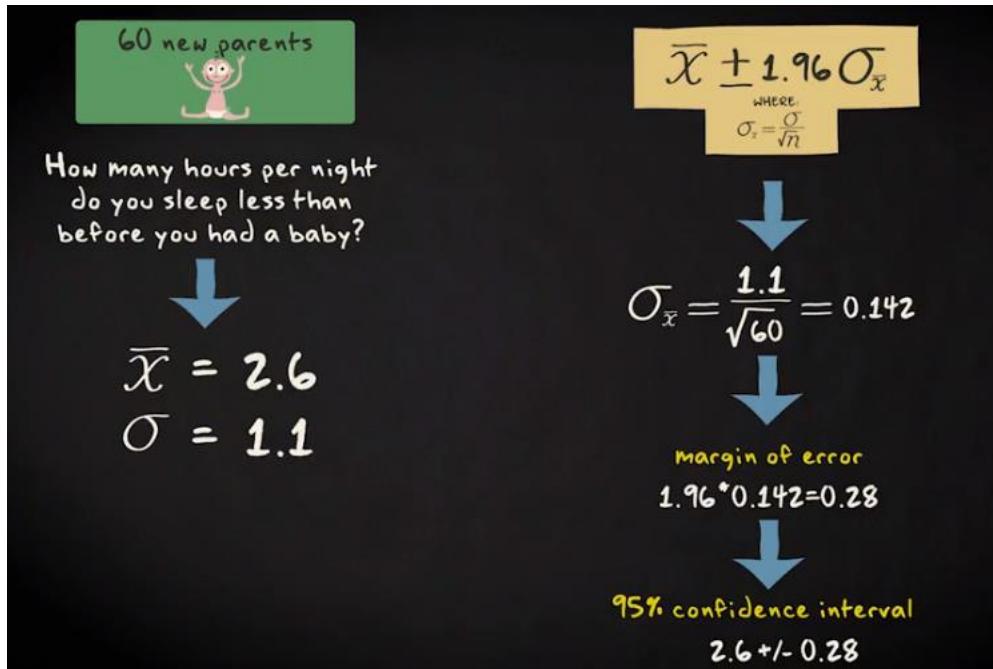
(3) Margin of error의 계산

- \bar{X} 을 사용해서 μ 을 추정하는데, 양 옆으로 range는 어떻게 할 것인가?
- sampling distribution을 1회의 sampling을 통해서 추정할 수 있기 때문에 이를 사용한다. 1) σ 를 알고 있는 경우나 n 이 30이상인 경우에는 정규분포(Z)를 사용해도 되지만, 2) σ 를 모르는 경우, n 이 30보다 작을 때에는 t-분포를 사용해서 신뢰구간을 구한다. 3) 모비율을 추정할 때는 정규분포(Z)를 사용한다. ($np >= 15, n(1-p) >= 15$)



→ 무한번 sampling을 하게 되면 \bar{X} 의 신뢰구간이 0.95의 확률로 μ 를 포함하게 될 것이다. \bar{X} 를 무한번 뽑은 것이 바로 sampling distribution이기 때문이다.

(3-1) σ 를 알고 있는 경우

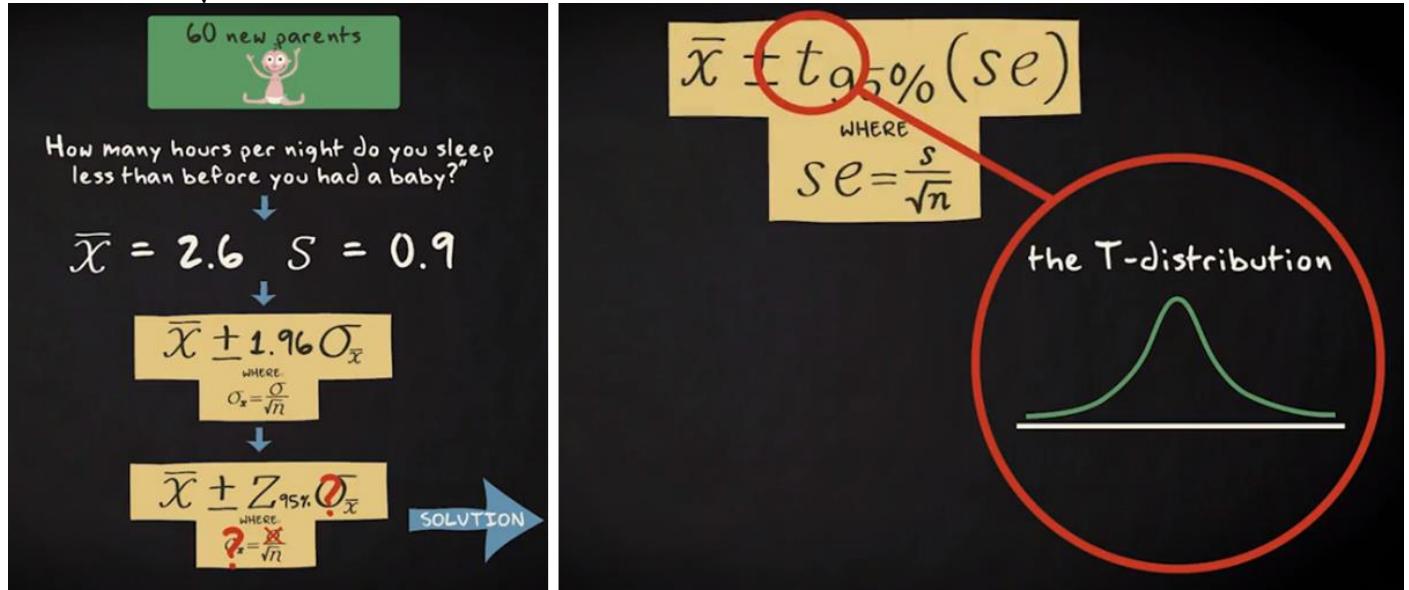


(3-2) 그러나, σ 를 모르는 경우가 대부분이다. n 이 작은 경우. → t-분포 사용

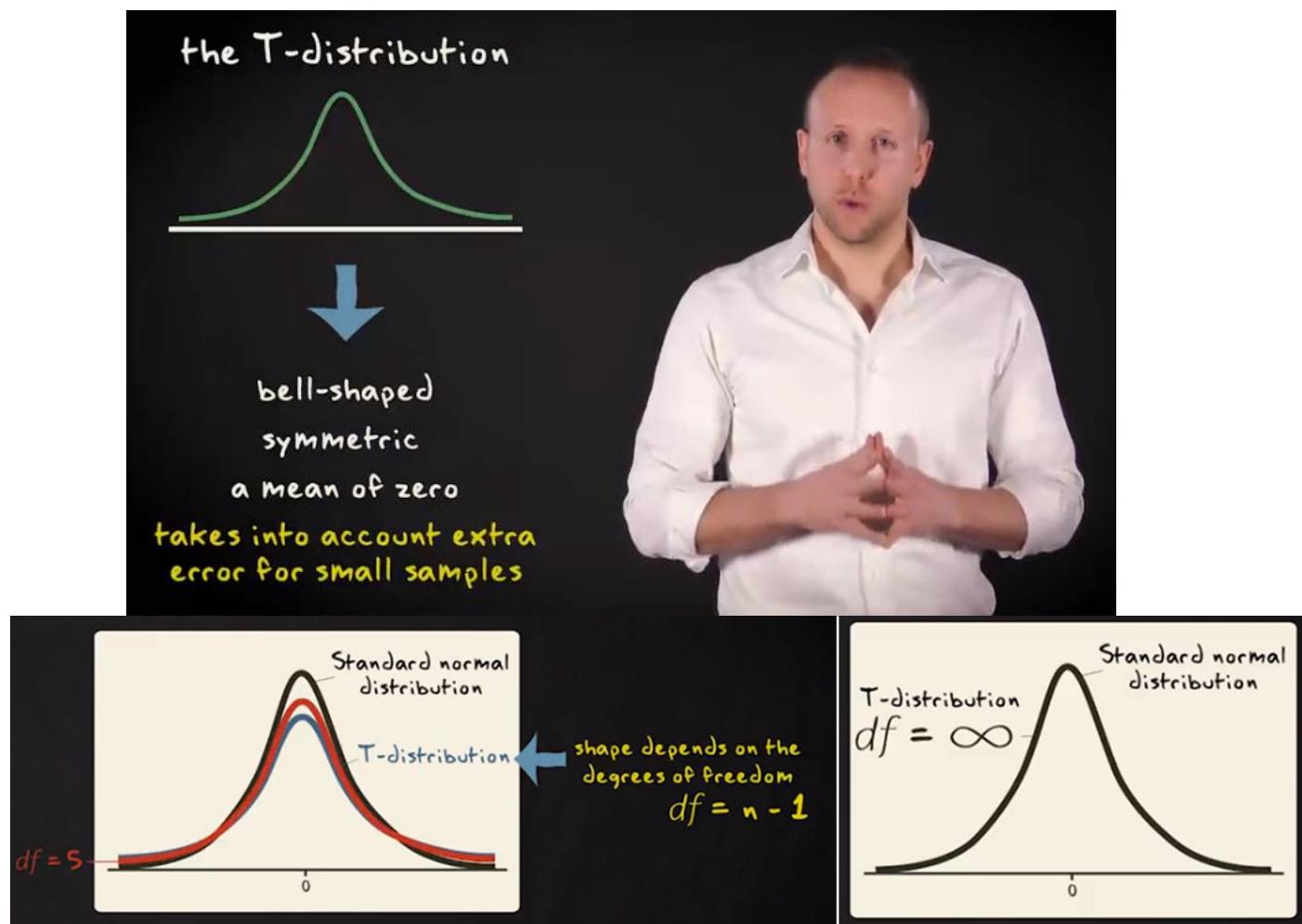
- estimate the population standard deviation by t-분포. Degree of freedom은 $n-1$ 이다. T-분포 사용할 때 $n-1$ 개를 꼭 사용할 것.

- 표준 오차(standard error) = estimated standard deviation of the sampling distribution

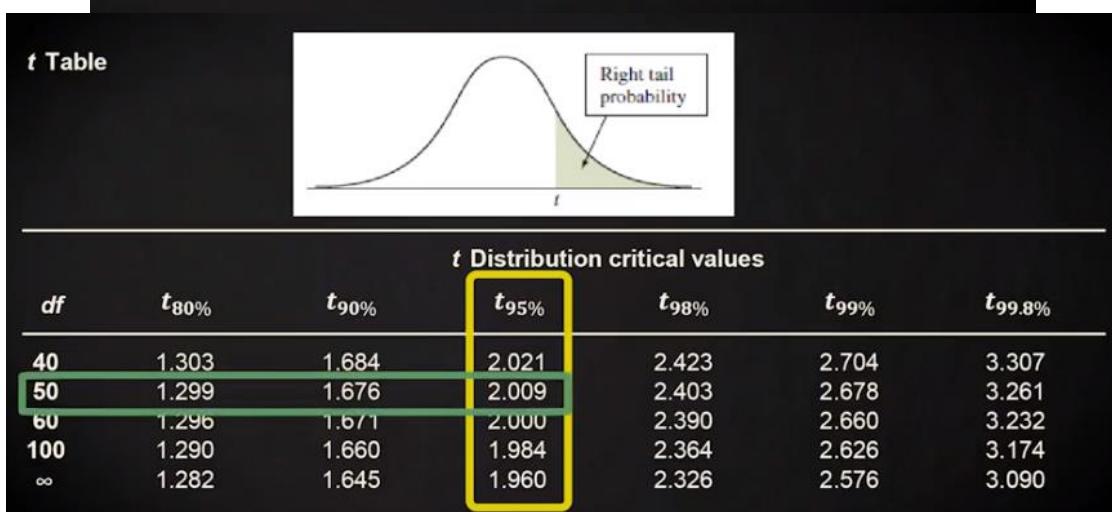
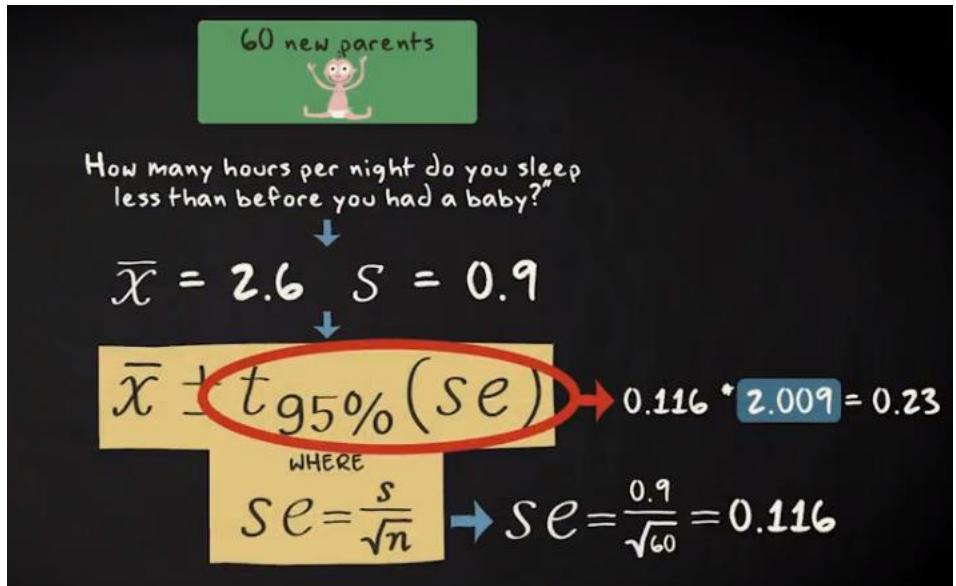
$$\rightarrow se = \frac{s}{\sqrt{n}}$$



- t-분포 (t-distribution) : t-분포는 sample size를 감안한 확률분포이다. (depend on the degree of freedom) 따라서 sample size가 작을 때의 정규분포와의 차이를 보정해 줄 수 있다. T-분포가 sample size가 2일 때는 꼬리가 길고 높이가 낮다. 그러나 $n > 30$ 이 되면 점점 정규분포가 같아지고, sample size가 무한대이면 t-분포를 정규분포가 된다.

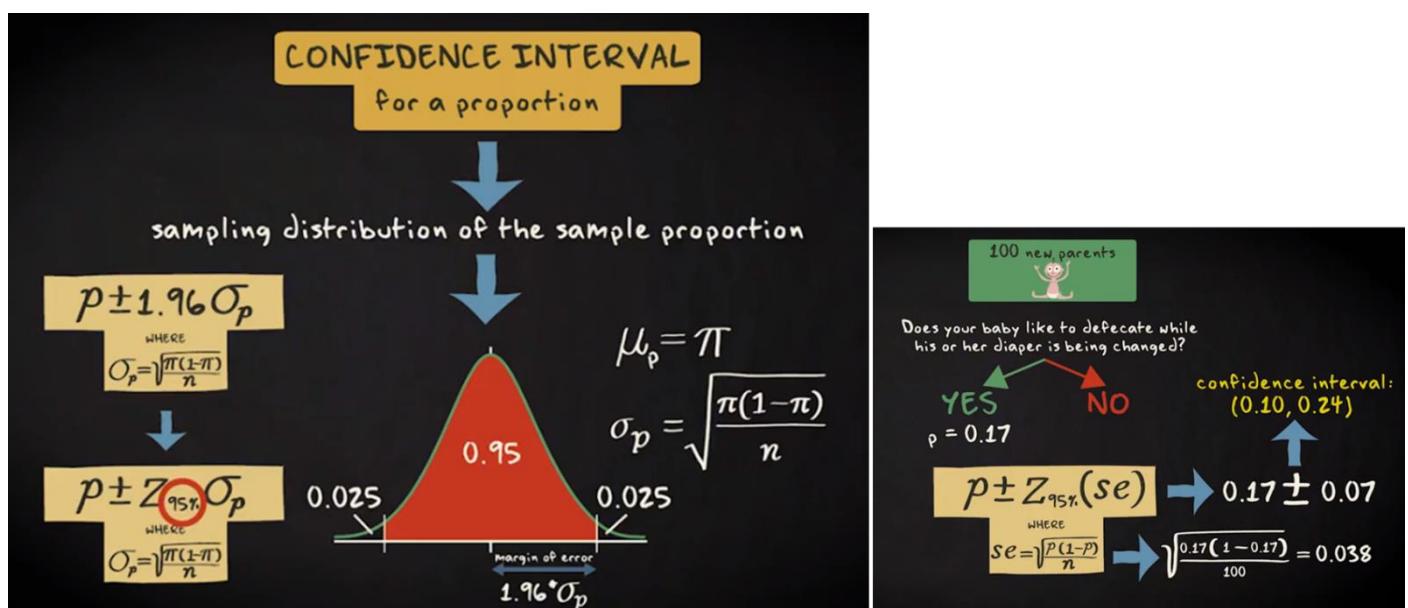


- 계산 사례 :



(3-3) 모비율을 추정하는 경우 → z분포를 사용하면 된다.

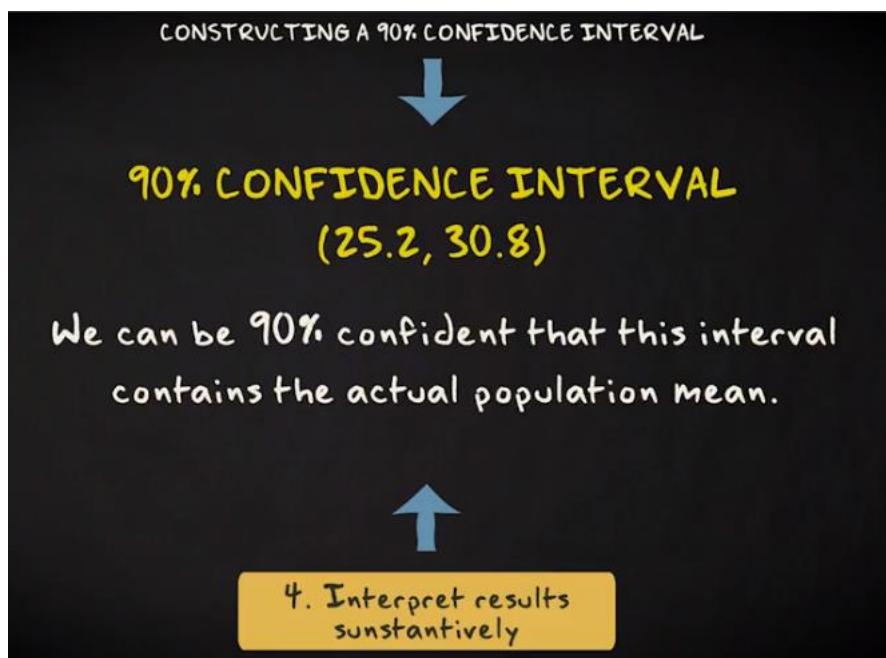
- interval estimation for "proportion" : 95%의 경우에 모분포의 π 가 confident interval안에 있게 되는 구간.
- 구하는 법 : 구할 때는 모비율 π 를 모르기 때문에 대신에 p 를 넣어서 표본비율분포의 표준오차(standard error)를 구한다. $np >= 15, n(1-p) >= 15$ 이면 Z 표준정규분포를 사용한다.



→ 해석 : 0.95 신뢰수준에서 10%~24%의 확률로 기저귀를 갈아줄 때 똥을 싼다. 따라서 똥을 싸는게 비정상은 아니었던 것이다.

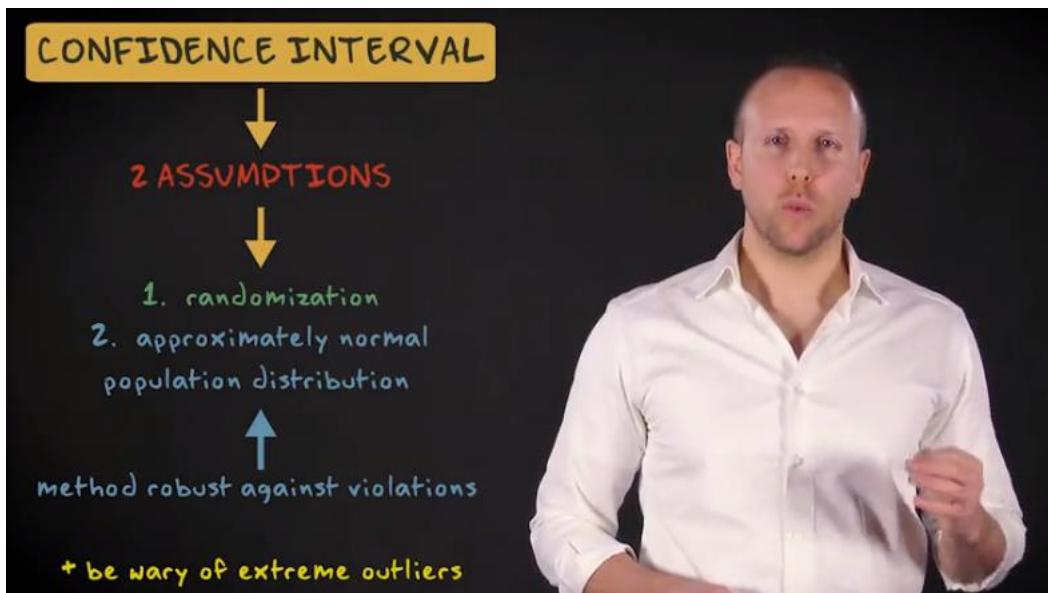
(4) 결과 해석

- 90% confident that interval contains population mean
- $n=150$ 의 표본을 무한번 뽑아서 신뢰구간을 그릴 때, 90%의 경우에 population mean이 신뢰구간 안에 놓이게 된다.



5.2.3. 주의 사항

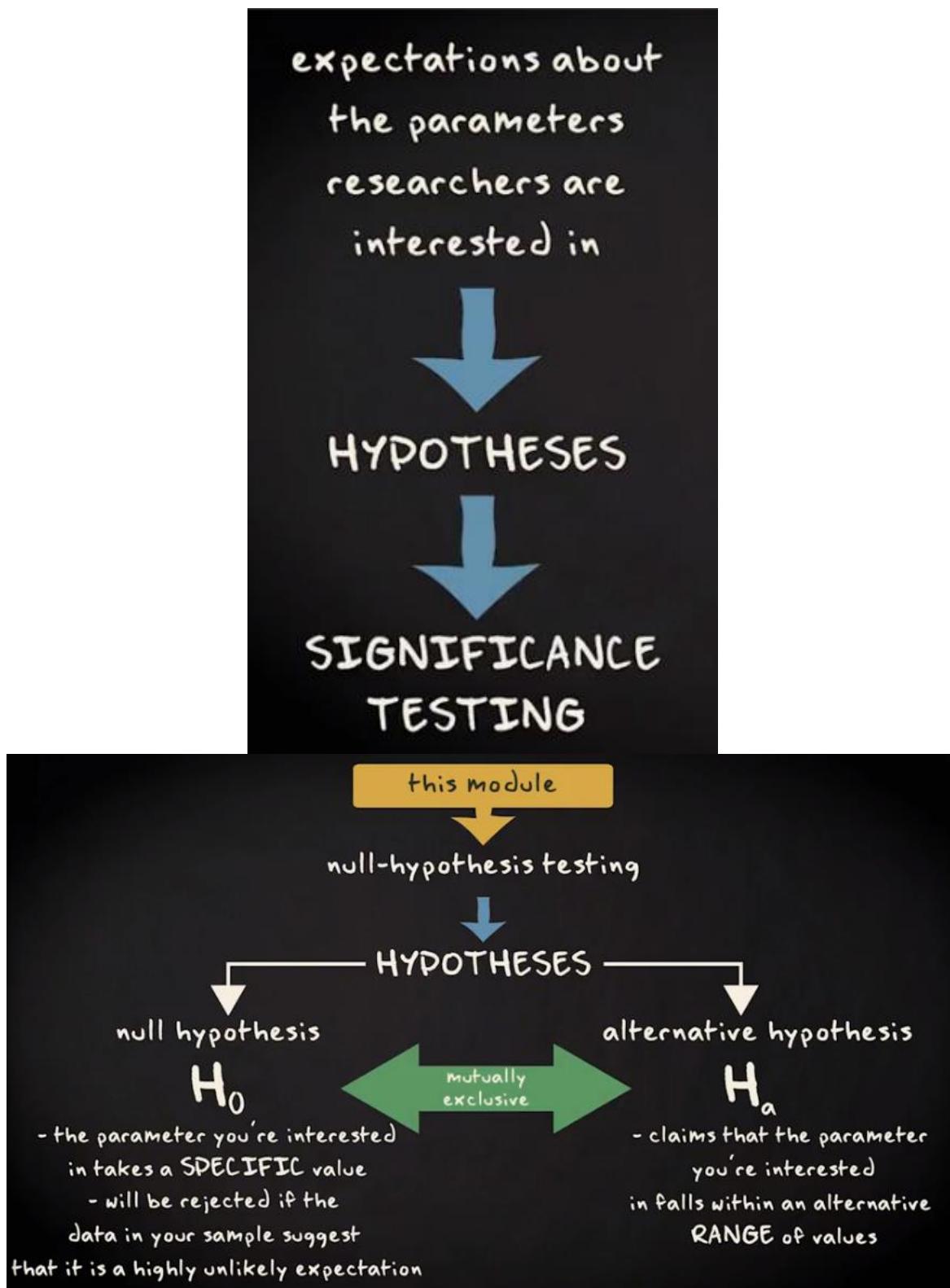
: 표본을 random하게 뽑는 것은 기본. 그러나 모분포가 꼭 정규분포일 필요는 없다. t-분포를 사용하면 되기 때문이다. Outlier는 항상 조심하자, t-분포를 사용하더라도 커버가 안 된다.



6. 가설 검정 (test hypothesis)

6.1. Significance testing

- hypothesis = expectations about population
- significance testing : sample data를 가지고 hypothesis가 맞는지 보는 것.
 - null hypothesis와 alternative hypothesis로 구성된다. 이 때의 null hypothesis는 항상 specific value여야 한다.
- 주의 : null hypothesis를 기각하지 못한다는 것이 null hypothesis가 true인 것을 의미하지는 않는다. 피고가 guilty하지 않다는 것이 그 사람이 innocent하다는 것을 의미하는 것은 아니다.



6.2. significance test about a population proportion

<과정>

(1) 가설 설정

- 1) 모비율을 추정하는지, 모평균을 추정하는지?
- 2) 단측 검정, 양측 검정 어떤 것을 할 것인지 결정.
- 3) 유의 수준 설정

(2) Sampling과 검정 단계 :

가설이 맞다고 치고, sampling값의 검정 → by **검정 통계량**(test statistic) : 표준오차로 얼마나 떨어져 있는지 본다.

- 점검할 사항 : 1) 표본을 random하게 추출했는가 2) 모분포가 정규분포를 이루는가(표본수가 크면 표본평균의 분포(sampling distribution)이 정규분포를 따르기 때문에 괜찮다.

- 검정통계량의 결정 1) 모평균 추정시

- σ 를 알고 있는 경우나 n 이 30이상인 경우에는 정규분포(Z)를 사용해도 되지만,
→ 더 정확하게 하려면 n 이 30이상이어도 t-분포를 사용해도 된다.
 - σ 를 모르는 경우에 (n 이 30보다 작을 때에는) t-분포를 사용해서 신뢰구간을 구한다.
- 2) 모비율을 추정할 때는 정규분포(Z)를 사용한다. ($np \geq 15, np(1-p) \geq 15$)
- X 는 0또는 1이므로 이항분포를 이루기 때문에 표본이 충분히 크면 정규분포를 따름.

(3) 결과 해석 : P value와 유의수준 비교 / 또는 기각역으로 판단

(4) 결론 : 귀무가설을 기각한다. 기각하지 못한다. Cf) 주의사항 : 결론이 양측/단측 검정, 유의수준에 따라 달라질 수 있다.

(1) 가설 설정 :

(1-1) Proportion or mean?

2 EXPECTATIONS:

π

1. More than half of all certified divers in America have more than 35 hours of diving experience

μ

2. Mean number of hours of diving experience of all certified divers in America is more than 35 hours

(1-2) 양측 검정을 할 것인지. 단측 검정을 할 것인지.

How many Americans have scuba-diving experience?

less than three percent

$H_0: \pi = 0.03$ $H_a: \pi < 0.03$

How many Americans have scuba-diving experience?

less than three percent

$H_0: \pi = 0.03$ $H_a: \pi \neq 0.03$

we cannot reject our null hypothesis that π is equal to 0.03

→ One, two tailed test : one tailed를 하는지($H_1 : \pi < 0.03$), two tailed test($H_1 : \pi \neq 0.03$)에 따라서 기각역이 $z=-1.96, +1.96$ 으로 다르기 때문에 결과가 달라질 수 있다. 특별한 근거가 없으면 양측검정(two tailed)을 사용한다.

(1-3) 유의 수준의 설정 : significance level (유의수준, α) =0.05

(2) Sampling과 검정

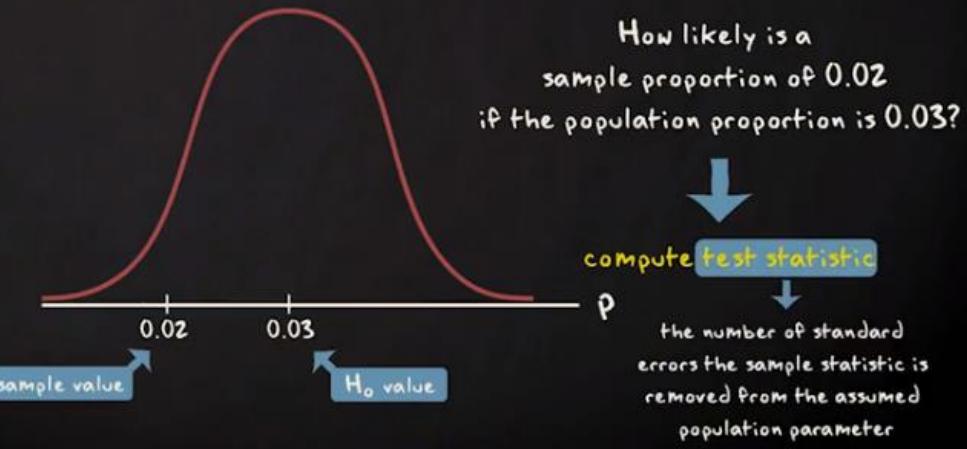
- 귀무가설(null hypothesis)가 맞다고 치자. 모평균이 0.03이라고 해보자. Sampling을 했으므로 이를 통해 sampling distribution을 그릴 수 있다. Sampling distribution의 평균은 H_0 의 값(0.03)이다. 표준오차는 $\sqrt{\frac{\pi_0(1-\pi_0)}{n}}$ or $\sqrt{\frac{p_0(1-p_0)}{n}}$ 이다. 다음 sample의 비율값이랑 H_0 value가 얼마나 떨어져있는지 구해본다. (test statistic)

SIGNIFICANCE TEST

$$n = 1000$$

$$p = 0.02$$

sampling distribution of sample proportion



→ Test statistic (검정통계량) : number of standard errors sample value is removed from H_0 value

$$= z = \frac{p - \pi_0}{se_0}, \text{ 애 } se_0 \text{는 sampling distribution이기 때문에 } \sqrt{\frac{\pi_0(1-\pi_0)}{n}} \text{ 사용}$$

how many Z-scores the sample statistic is removed from the population parameter



$$\text{test statistic} = z = \frac{p - \pi_0}{se_0}, \text{ where } se_0 = \sqrt{\frac{\pi_0(1-\pi_0)}{n}}$$

→ 계산해보면 $z = -1.85$ 이다.

(3) 결과 해석 : P value와 유의수준 비교 / 또는 기각역으로 판단

(3-1) P값과 유의수준의 비교

- **P-value** : z 값이 -1.85 혹은 그것보다 작게 나올 확률은 어떻게 되는가? 0.0322이다. 이것을 **P-value**라고 부른다.

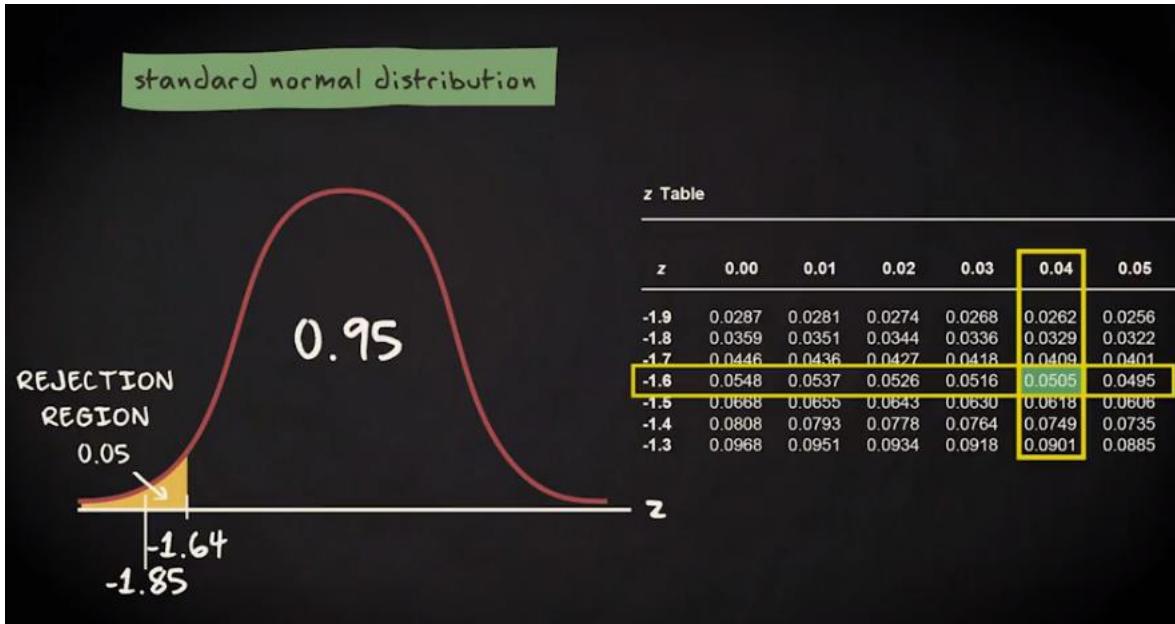
- **significance level** (유의수준, $\alpha=0.05$)는 귀무가설을 기각하는 기준이 되는 확률이다. 이보다 P-value가 작으면 귀무가설을 기각한다.

z Table

Standard normal cumulative probabilities

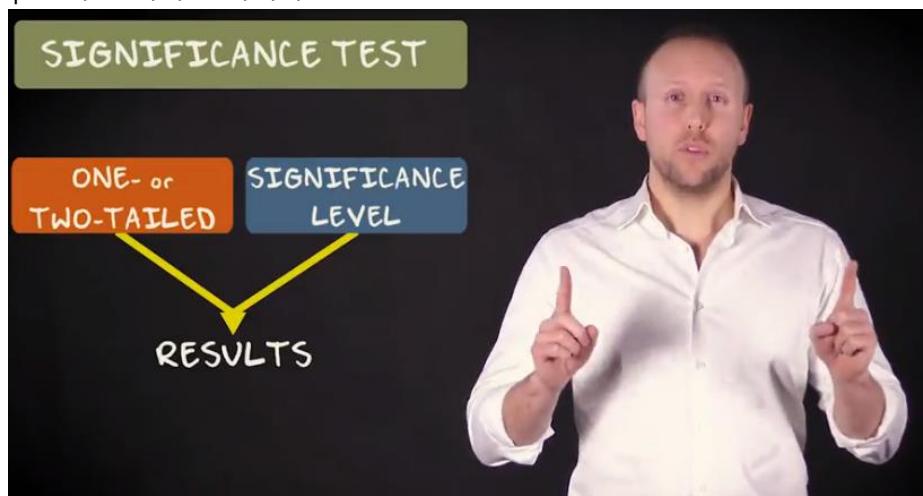
<i>z</i>	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823

(3-2) 기각역으로 판단 : **rejection region**(기각역)에 들어있는 것을 볼 수 있다. 이를 통계적으로 유의하다라고 (statistically significant)라고 한다.



(4) 해석과 해석시 주의사항 :

- 귀무가설을 기각/기각하지 않는다. 그러나
- 주의사항 : 유의수준 / 양측, 단측 검정에 따라 결론이 달라질 수 있다. 그리고 귀무가설을 기각하지 않는 것 이 귀무가설을 accept한다는 의미는 아니다.



6.3. significance test about a population mean

(1) 가설 설정 / 유의수준 결정 :

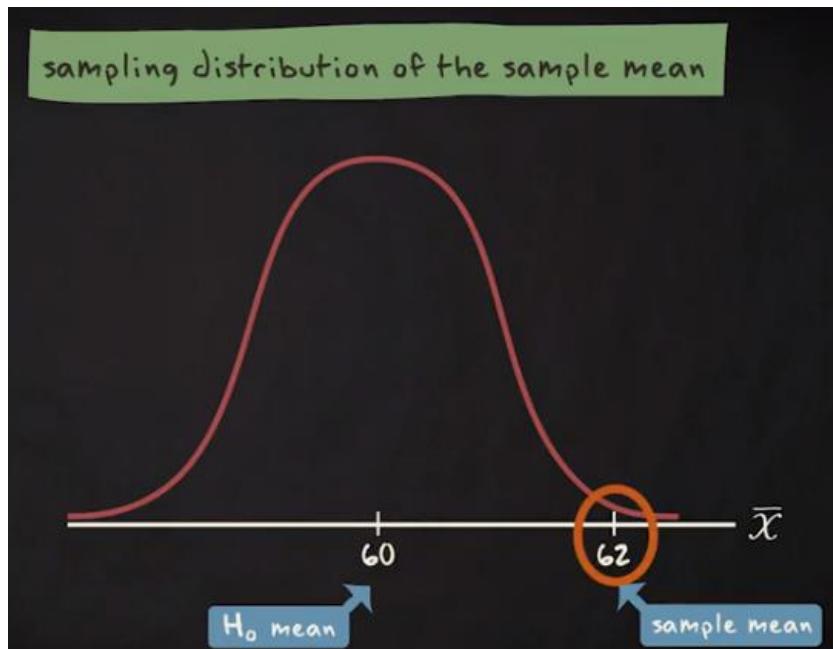
- null hypothesis(H_0) : scuba diving하는 사람들의 다이빙 평균은 60이다.
- $H_1 : \bar{x} > 60$ (단측 검정)

$$H_0: \mu = 60$$
$$H_a: \mu > 60$$

we assess if it is likely that the sample we have collected actually comes from a population with a mean that equals the value formulated in the null hypothesis

(2) Sampling과 검정

- null hypothesis가 맞다고 치자.



- sampling을 한다 : $n=100$ 명, 표본평균(\bar{x}) = 62, 표본의 표준편차(s) = 5
 - 표본의 평균이 sampling distribution의 평균으로부터 얼마나 떨어져 있는가?
- 검정통계량 : t 분포를 사용한다. $t = \bar{x} - \mu_0 / se$, $se = s / \sqrt{n}$

Cf) $z = \bar{x} - \mu_0 / se$, $se = \sigma / \sqrt{n}$ (σ 를 모르기 때문에 z 분포를 사용하면 오차 발생)

TEST STATISTIC

$$t = \frac{\bar{x} - \mu_0}{se} \quad \text{where} \quad se = \frac{s}{\sqrt{n}}$$

TEST STATISTIC

=

number of standard errors sample mean is removed from H_0 value

to compute se we need to know σ



we don't know σ



we estimate σ with $s \rightarrow$ we introduce extra error



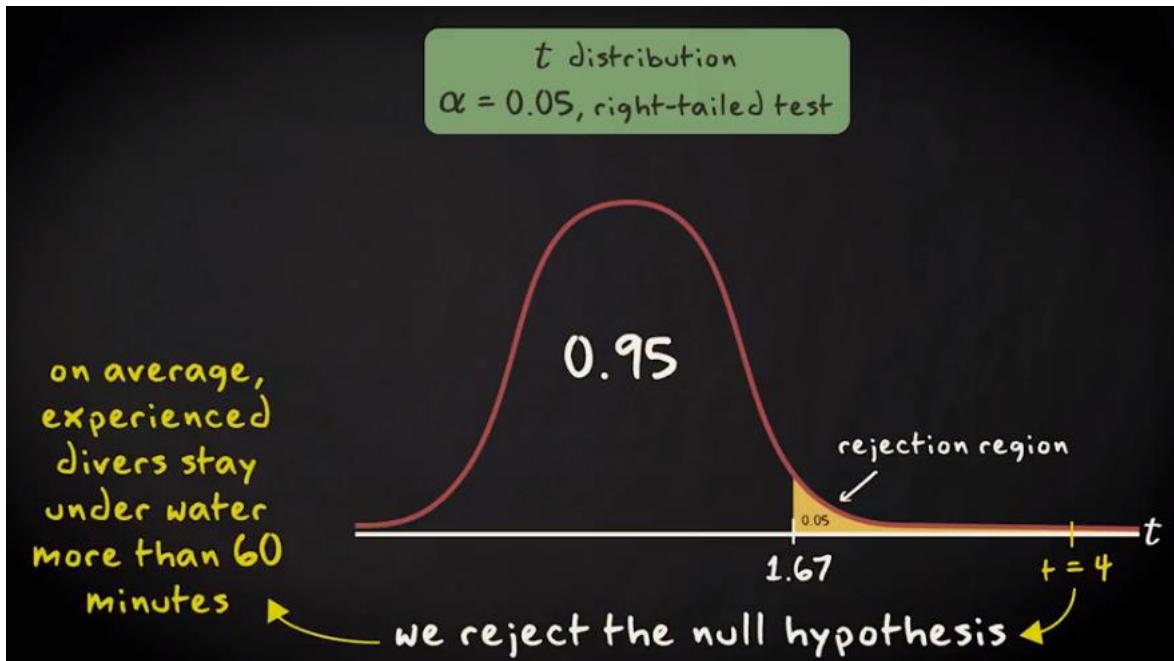
employ t distribution instead of z distribution

(3) 결과 해석과 유의수준과의 비교

- t분포의 degree of freedom : $100-1=99$ 이므로 그것보다 작은 60자리를 사용. 왜 100을 사용하지 않는것인지?
- 유의수준을 0.05라고 하면 t 95%를 사용해야 한다. ← (cumulative probability)일 때. 따라서, $t = 1.67$

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
Z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

(4) 결론 해석

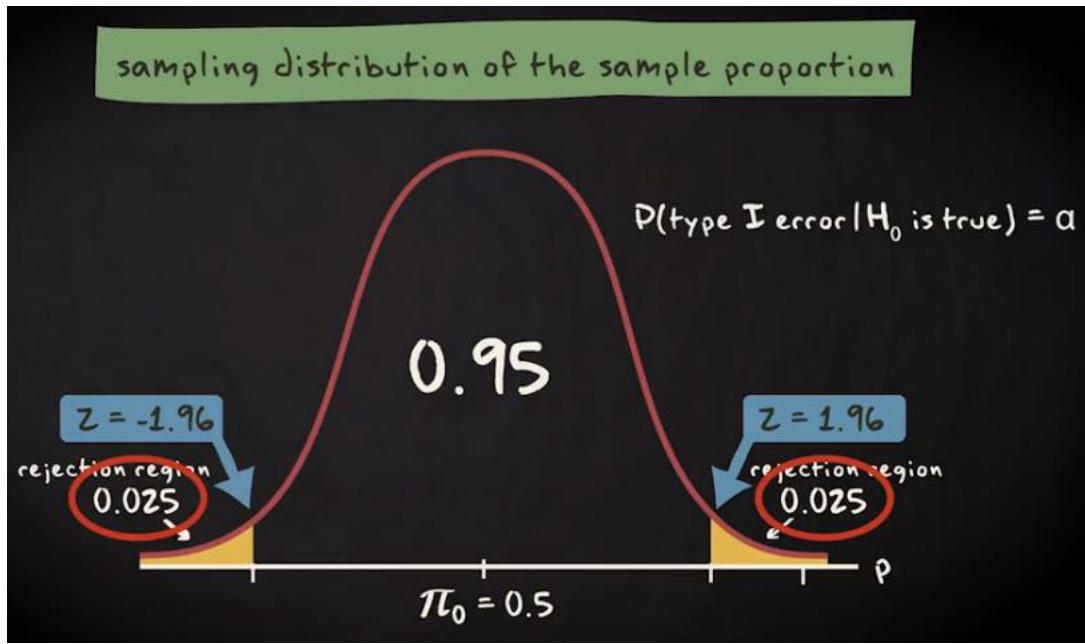


7. 결과의 해석의 연장선상 : Type of error

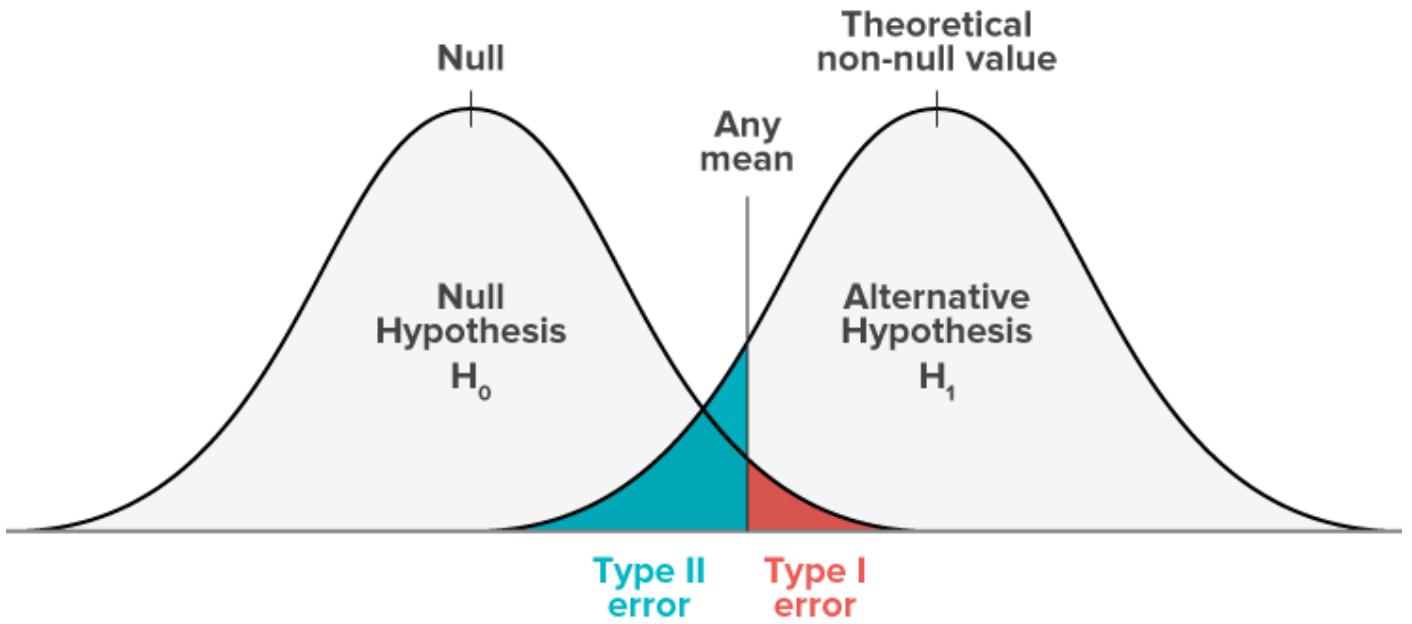
7.1. 정의

(1) **Type 1 error** (false positive) : null hypothesis(H_0)가 사실인데, H_0 를 기각하는 경우(H_1 을 선택). Ex) 범죄자가 죄가 없는데, 처벌하는 경우. Cf) false positive의 의미 : positive (대립가설 기준이다.)

→ 정의상 유의수준 α 와 같다. 왜냐하면 유의수준이란 영가설이 사실이라고 가정했을 때, 그것을 기각할 확률이기 때문이다.



(2) **Type 2 error** (false negative) : H_1 (대립가설)이 사실일 때, H_0 를 택할 확률. Ex) 범죄자가 분명 죄가 있는데, 죄가 없다고 풀어주는 경우 / 병이 있는데 병이 없다고 하는 오류



7.2. 1종 오류와 2종 오류의 관계

- 한쪽 오류를 감소시키려고 하면 다른 오류의 확률이 올라간다.

	set free	convicted
defendant innocent	✓	✗
defendant guilty	✗	✓

↓ significance test

	not rejected	rejected
H₀ true	✓	type I error: α
H₀ false	type II error: β	✓

when α goes up, β goes down (and vice versa)

decrease the significance level
 ↓
 decrease the probability of making a type I error
 ↓
 INCREASE the probability of making a type II error

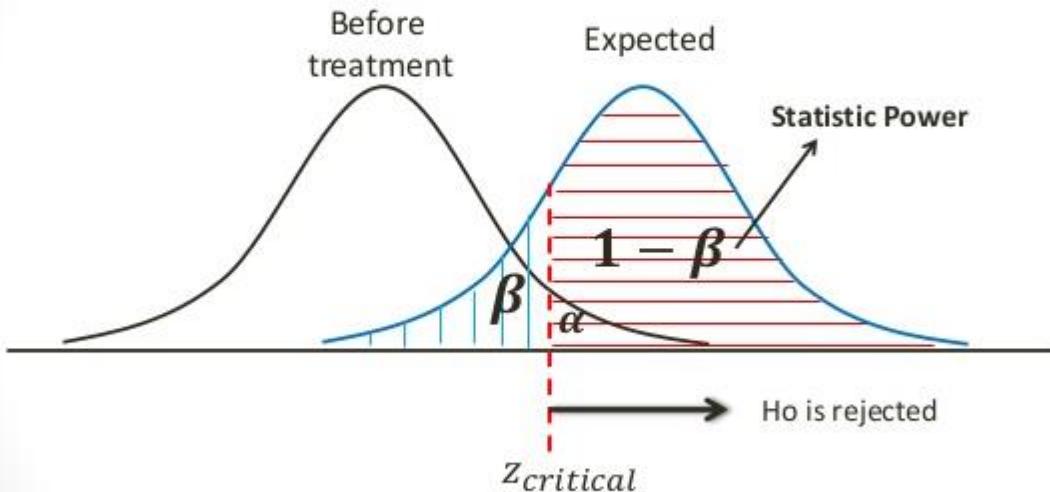
7.3. Power (2종 오류의 파생 개념)

7.3.1. 정의 : power of rejecting the null hypothesis, given that it is false. 1-type 2 error = $1-\beta$

상상 속의 alternative hypothesis에 따른 분포를 가정했을 때, “바꿔야 해서 잘 바꾼 것일 확률”

(근후님 예시 : 더 높은 효율을 내는 기계를 바꿔야 할까 말까를 고민할 때의 사례에서)

WHAT does a value of Statistical Power mean?



Ideally, power should be set around 0.80

(Cohen, 1988)

- When testing $H_a : \mu > \mu_0$, notice if power is $1 - \beta$, then

$$1 - \beta = P\left(\bar{X} > \mu_0 + z_{1-\alpha} \frac{\sigma}{\sqrt{n}} ; \mu = \mu_a\right)$$

7.3.2. Power를 크게 하기 위한 일반적인 방법 : 유의성 검정을 할 때에는 1종 오류(유의수준)을 기준으로 하기 때문에, power를 늘리긴 어렵다. 따라서 애초에 실험 설계 단계에서 표본수를 늘리면 power가 커질 수 있다.

Ex) Null hypothesis를 기각할 수 없는 결과가 나왔을 때, 통제집단, 실험집단에 3명, 3명 있는 때보다 300명, 300명 있을 때, null hypothesis가 사실일 확률이 더 높아진다.

7.3.3. Power와 다른 변수들의 관계

$$\text{Power} \propto \frac{\sqrt{n} (\mu_a - \mu_0)}{\sigma}, \quad \frac{(\mu_a - \mu_0)}{\sigma} : \text{effect size}$$

1) 유의수준 α 가 커지면 β 가 작아지므로 power($1-\beta$)가 커진다.

그리고, 양측 검정시에는 $\alpha/2$ 를 사용하므로 단측 검정보다 power가 작아진다.

2) μ_1 의 크기가 커지면 power가 커진다.

→ 예를 들어서 μ_0 이 30이고 μ_1 이 32인 경우보다 μ_1 이 40인 경우에, null hypothesis가 거짓인 것/사실인 것을 알아내서 기각할 확률이 높아지기 때문이다.

3) n 이 커지면 power가 커진다.

→ 그 이유는 n 이 커질수록 sampling distribution의 폭이 작아지기 때문에, 그리고 기각역이 점차 원쪽으

로 가기 때문이다.

7.3.4. 계산하는 법

- 핵심 : power를 항상 먼저 계산해본다. 0.8 이상으로 유지하는 것이 좋다. 적절한 표본 수를 정할 때도 유용하다. 이를 위해서는 power.t.test를 사용하면 편하다.

(1) 이론적으로는 Z분포를 사용한다.

We reject if $\frac{\bar{X} - 30}{\sigma/\sqrt{n}} > z_{1-\alpha}$

- Equivalently if $\bar{X} > 30 + Z_{1-\alpha} \frac{\sigma}{\sqrt{n}}$

Under $H_0 : \bar{X} \sim N(\mu_0, \sigma^2/n)$

Under $H_a : \bar{X} \sim N(\mu_a, \sigma^2/n)$

```
z <- qnorm(1 - alpha)
pnorm(mu0 + z * sigma/sqrt(n), mean = mua, sd = sigma/sqrt(n), lower.tail = FALSE)
```

(2) 실제적으로는 t-분포를 사용한다.

T-test power

- Consider calculating power for a Gossett's T test for our example
- The power is

$$P\left(\frac{\bar{X} - \mu_0}{S/\sqrt{n}} > t_{1-\alpha, n-1} ; \mu = \mu_a\right)$$

- Calculating this requires the non-central t distribution.
- power.t.test does this very well
 - Omit one of the arguments and it solves for it

→ power.t.test를 사용하면 power를 계산할 수도 있고, 알고 싶은 n(표본수)를 계산할 수도 있다.

```
power.t.test(n = 16, delta = 2/4, sd = 1, type = "one.sample", alt = "one.sided")$power
```

```
## [1] 0.604
```

```
power.t.test(n = 16, delta = 2, sd = 4, type = "one.sample", alt = "one.sided")$power
```

```
## [1] 0.604
```

```
power.t.test(n = 16, delta = 100, sd = 200, type = "one.sample", alt = "one.sided")$power
```

```
## [1] 0.604
```

→ 여기서의 delta는 $\mu_a - \mu_0$ 인 값이다. 따라서 이를 σ 로 나눈 것이 effect size인데, 이 계산과정에서 effect size가 같으면 power도 같다는 것을 알 수 있다. (n 일정)

```
power.t.test(power = 0.8, delta = 2/4, sd = 1, type = "one.sample", alt = "one.sided")$n
```



```
## [1] 26.14
```

```
power.t.test(power = 0.8, delta = 2, sd = 4, type = "one.sample", alt = "one.sided")$n
```

```
## [1] 26.14
```

```
power.t.test(power = 0.8, delta = 100, sd = 200, type = "one.sample", alt = "one.sided")$n
```

```
## [1] 26.14
```

→ n을 알고 싶은 때는 다른 변수들을 정해주면 된다. 여기서 power가 0.8인 이유는 0.8 이상으로 유지하는 것이 좋기 때문이다.

8. Multiple testing corrections

8.1. Correction의 필요성 : 2개 이상의 test를 하게 되면 일종의 correction이 필요하다. Ex) P value를 여러 개 계산하고 가장 작은 p value값만을 사용한다던지, 여러 개를 사용해놓고 모두 0.05보다 작다고 주장한다던지 하면 오류로 이어질 수 있다.

Cf) Three eras of statistics : 현대에는 data의 흥수 속에서 살고 있는데, testing을 여러 번 하면서 error가 누적되면서 커질 수 있기 때문에 correction이 필요한 것이다.

Three eras of statistics

The age of Quetelet and his successors, in which huge census-level data sets were brought to bear on simple but important questions. Are there more male than female births? Is the rate of insanity rising?

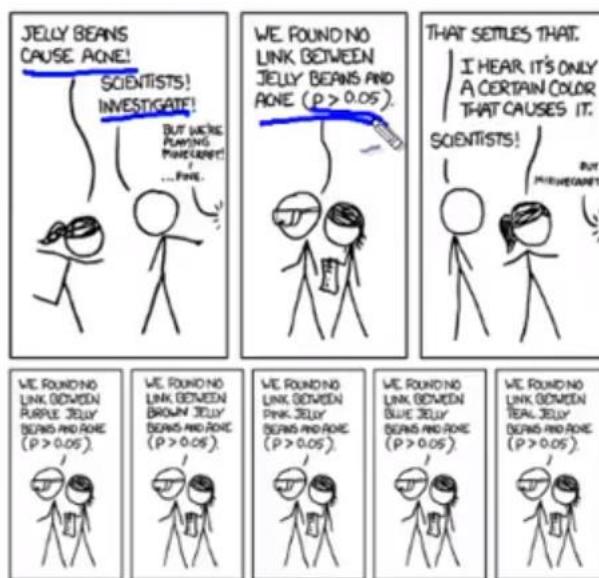
The classical period of Pearson, Fisher, Neyman, Hotelling, and their successors, intellectual giants who developed a theory of optimal inference capable of wringing every drop of information out of a scientific experiment. The questions dealt with still tended to be simple Is treatment A better than treatment B?

The era of scientific mass production, in which new technologies typified by the microarray allow a single team of scientists to produce data sets of a size Quetelet would envy. But now the flood of data is accompanied by a deluge of questions, perhaps thousands of estimates or hypothesis tests that the statistician is charged with answering together; not at all what the classical masters had in mind. Which variables matter among the thousands measured? How do you relate unrelated information?

<http://www-stat.stanford.edu/~ckirby/brad/papers/2010LSIexcerpt.pdf>

- 오류의 예시 : 젤리빈이 여드름을 발생시킨다는 가설을 검증하려고 젤리빈을 조사한다고 치자. 젤리빈의 색깔이 20개여서 20개의 testing을 하게 되는데 그 중에서 초록색 젤리빈이 여드름을 발생시킨다고 유의수준 0.05수준에서 그렇다고 결과가 나왔다. 이 결과를 신뢰할 수 있는가?

→ 아니다. 한 테스트 당 5%의 틀릴 확률(제1종 오류)이 있다. 따라서 20개의 test를 하게 되면 그 중 1개는 틀린 test 결론을 낼 수 있는 것이다.



<개괄> multiple testing correction은 2단계로 이루어진다. (1) Error measure 단계 (2) Correction 단계

8.2. Error measure 종류

(1) **False positive rate** : 제1종 오류가 발생할 확률 = “ $\beta=0$ 이 사실일 때” 중에서 “ $\beta\neq0$ 이라고 주장할 경우”의 확률

(2) **Family wise error rate (FWER)** : 적어도 1번의 제1종 오류가 날 확률

(3) **False discovery rate (FDR)** : “ $\beta\neq0$ 이라고 주장했을 때(두 변수가 관계가 있다고 주장했을 때)” 중에서 “실제로는 관계가 없는 경우($\beta=0$)”의 확률

Error rates

False positive rate - The rate at which false results ($\beta = 0$) are called significant: $E\left[\frac{V}{m_0}\right]^*$

Family wise error rate (FWER) - The probability of at least one false positive $\Pr(V \geq 1)$

False discovery rate (FDR) - The rate at which claims of significance are false $E\left[\frac{V}{R}\right]$

Types of errors

Suppose you are testing a hypothesis that a parameter β equals zero versus the alternative that it does not equal zero. These are the possible outcomes.

	$\beta = 0$	$\beta \neq 0$	HYPOTHESES
Claim $\beta = 0$	U	T	$m - R$
Claim $\beta \neq 0$	V	S	R
Claims	m_0	$m - m_0$	m

Type I error or false positive (V) Say that the parameter does not equal zero when it does

Type II error or false negative (T) Say that the parameter equals zero when it doesn't

8.3. Correction 단계 : How to control false positive rate

Controlling the false positive rate

If P-values are correctly calculated calling all $P < \alpha$ significant will control the false positive rate at level α on average.

Problem: Suppose that you perform 10,000 tests and $\beta = 0$ for all of them.



Suppose that you call all $P < 0.05$ significant.

The expected number of false positives is: $10,000 \times 0.05 = 500$ false positives.

How do we avoid so many false positives?

8.3.1. By controlling Family-wise error rate (FWER)

→ 해결 : Family-wise error rate(FWER)을 관리하자. 즉, 단 1번이라도 1종 오류를 범할 확률을 줄이자.

(Bonferroni correction)

- 방법 : $\Pr(\text{1종 오류인 경우의 수 } \geq 1) < \alpha$
- 계산 : $P\text{-value} < \alpha/m$ * $\alpha_{\text{fwer}} = \alpha/m$, m 은 significance test 횟수
- 특징 : 계산은 쉬우나, could be too stringent.

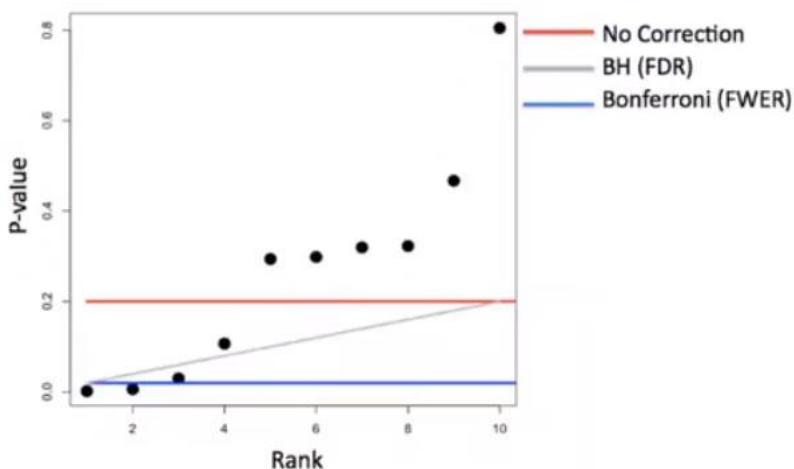
8.3.2. By controlling false discovery rate (FDR)

→ 가장 인기좋은 해결안 : “ $\beta \neq 0$ 이라고 주장했을 때(두 변수가 관계가 있다고 주장했을 때)” 중에서 “실제로는 관계가 없는 경우($\beta=0$)”의 확률을 일정 수준 이하로 관리하자. (BH, Benjamini & Hochberg method)

- 방법 : P-value를 작은 것부터 나열한다. (P_i 는 i번째) $P_1, P_2, P_3, P_4, \dots, P_i, \dots, P_m$

- 계산 : $P(i) \leq \alpha * \frac{i}{m}$ 이면 i번째 test는 significant하다.
- 특징 : 아까보다는 덜 엄격해서 더 많은 변수간의 관계를 이끌어낼 수 있다. 그러나, test가 서로 관련이 있는 거라면 이 방법이 잘 맞지 않는다.

Example with 10 P-values



Controlling all error rates at $\alpha = 0.20$

→ 10번의 testing을 했을 때 각 P-value를 크기 순으로 표시한 그래프다. 유의수준이 0.2라고 하면 빨간색선 밑의 test들이 significant하게 된다. 이를 correction을 하게 되면 FWER를 보면 기준점이 0.2를 10으로 나눈 0.02이하인 값들만이 significant하게 된다. 또한 FDR방식으로 하면, 회색선 밑의 값만이 significant해진다.

8.3.3. By adjusted P-values → R에서 유용

→ 해결 : α 를 바꾸지 않고 P-value를 correction해준다.

- 방법 : Bonferroni correction에서 α 를 m 으로 나누는 대신, P-value에 m 를 곱해서 α 보다 작으면 significant하다고 보는 방법이다. (단, $P * m \geq 1$ 이면 p.adjusted를 1로 본다)

8.3.4. BY

- If there is strong dependence between tests, there may be problems → consider method BY

8.4. R 실습

(1) 아무런 관계가 없는 변수 X, Y

Case study I: no true positives

```
set.seed(1010093)
pValues <- rep(NA, 1000)
for(i in 1:1000){
  y <- rnorm(20)
  x <- rnorm(20)
  pValues[i] <- summary(lm(y ~ x))$coeff[2,4]
}
# Controls false positive rate
sum(pValues < 0.05)
```

```
# Controls FWER  
sum(p.adjust(pValues,method="bonferroni") < 0.05)
```

```
[1] 0
```

```
# Controls FDR  
sum(p.adjust(pValues,method="BH") < 0.05)
```

```
[1] 0
```

→ p.adjust했으므로 그대로 알파는 0.05로 하면 되고, x, y의 관계가 없다고 설정을 했으므로 0이 나온다.

(2) 50%는 관계 있다고 할 때,

Case study II: 50% true positives

```
set.seed(1010093)  
pValues <- rep(NA,1000)  
for(i in 1:1000){  
  x <- rnorm(20)  
  # First 500 beta=0, last 500 beta=2  
  if(i <= 500){y <- rnorm(20)}else{ y <- rnorm(20,mean=2*x)}  
  pValues[i] <- summary(lm(y ~ x))$coeff[2,4]  
}  
trueStatus <- rep(c("zero","not zero"),each=500)  
table(pValues < 0.05, trueStatus)
```

trueStatus	zero	not zero
FALSE	0	476
TRUE	500	24

* 심화 자료

Further resources:

- [Multiple testing procedures with applications to genomics](#)
- [Statistical significance for genome-wide studies](#)
- [Introduction to multiple testing](#)