31 Mar 2021

◎ 이항 분포 Binomial

└ 시행의 결과가 T, F

X ~ Binomial
Y ~ Bin

X는 이산형 이항분포다.

$$P(X=x) = \binom{n}{x} p^{x} (1-p)^{n-x} \Longrightarrow X \sim Bin(n, p)$$

x: random variable

※ n과 p가 주어지면, 그 확률 법칙의 property 전체를 다 알 수 있다.

◎ 유의 수준 과 유의확률

$$P\left(X=1 \mid P_0=\tfrac{1}{2}\right) \le \underbrace{0.05}_{\alpha \text{ value}}$$

α value
유의수준
Type I error

$H_1: P_0 \neq \frac{1}{2}$

↕

$H_0: P_1 = \frac{1}{2}$, Not H_1

×	○
↑	↑
○	×
채택	기각

$$P(X=1) = \underbrace{\binom{4}{1}}_{\text{경우의 수}} \underbrace{0.5^{1} (0.5)^{3}}_{\text{확률}}$$

$$_nC_x = \frac{n!}{x!(n-x)!}, \quad 0! = 1$$

◎ 가설 검정의 전략

- 귀무가설 채택 → 대립가설 기각
- 귀무가설 기각의 기준 ⟸ 귀무가설이 참일 때 표본 값이 얻어질 확률.

◎ 유의 수준 = 기각역 = α value = Type I error

- $\max P(\text{귀무가설을 기각} \mid \theta=\theta_0)$
- 자료분석 전 미리 설정.

◎ 유의 확률 P-value

- P(실험에 의해 얻어진 결과치 | $\theta = \theta_0$)
- 자료 분석 후, 결과치

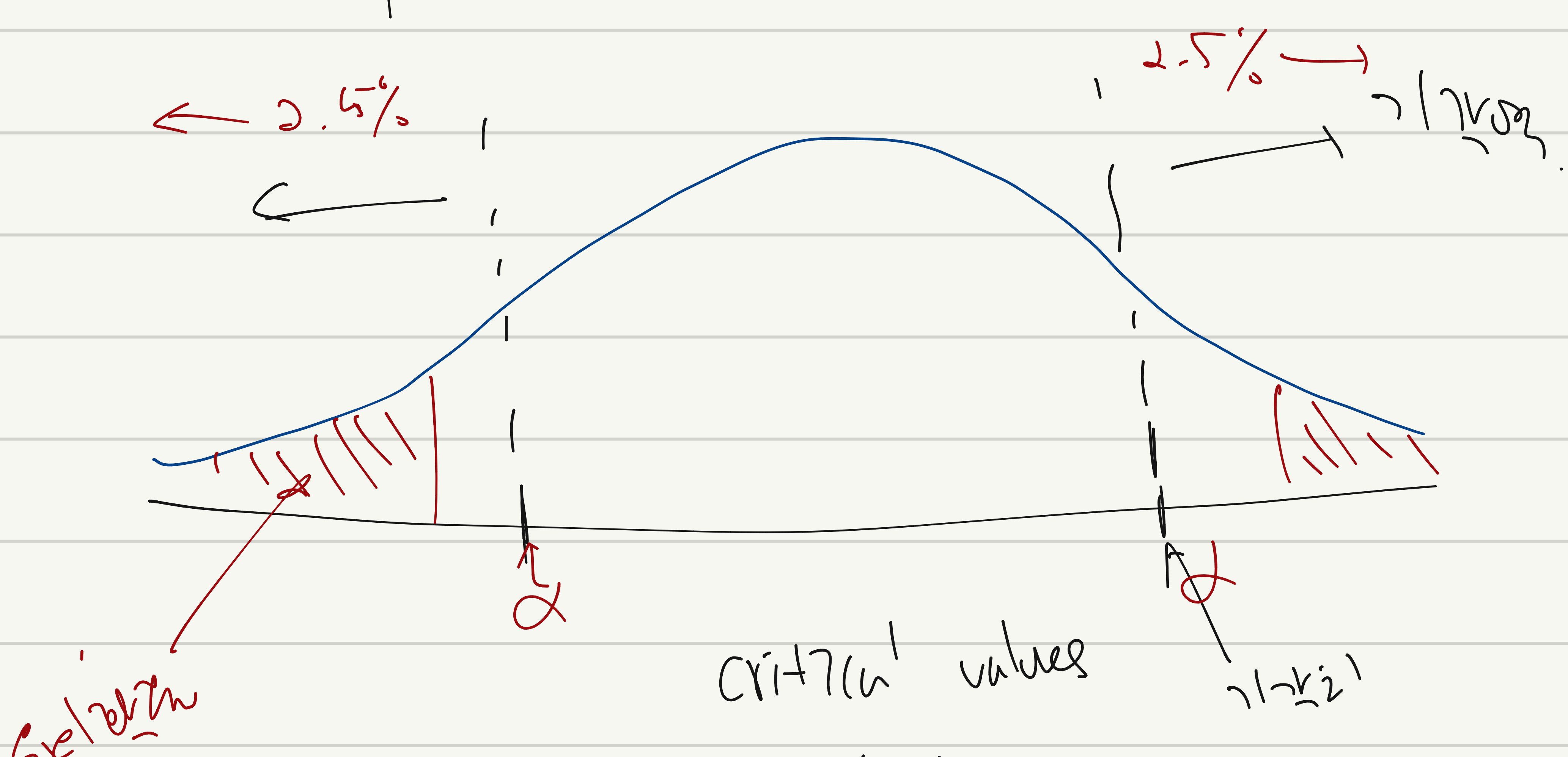
ex) $2 \times P(X=0 \mid p_0 = \frac{1}{2}) = 0.02$

$2 \times P(X=6 \mid p_0 = \frac{1}{2}) = 0.02$

양측검정

두개값이 비슷하므로

◎ 기각치와 p-value (α 수준 임의히 변경)



2.5 + 2.5

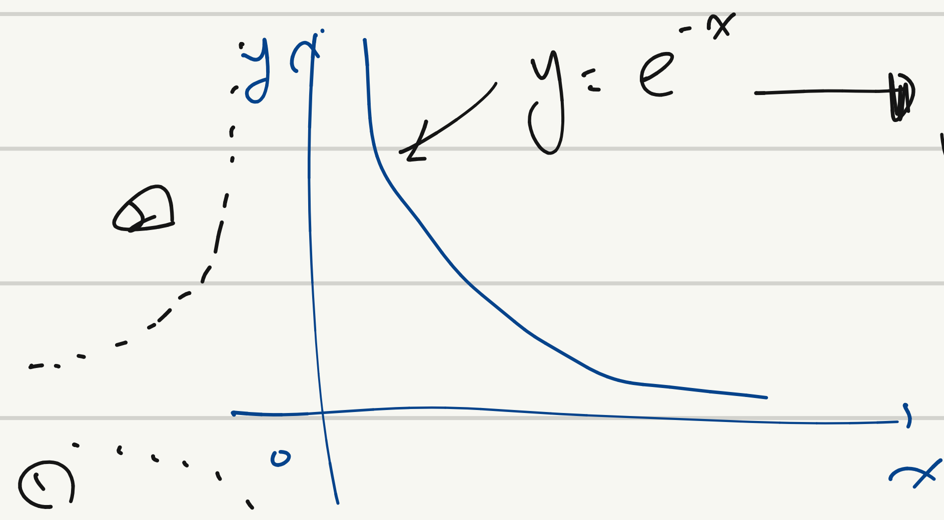
5%

◎ Z 검정 (Z-test)

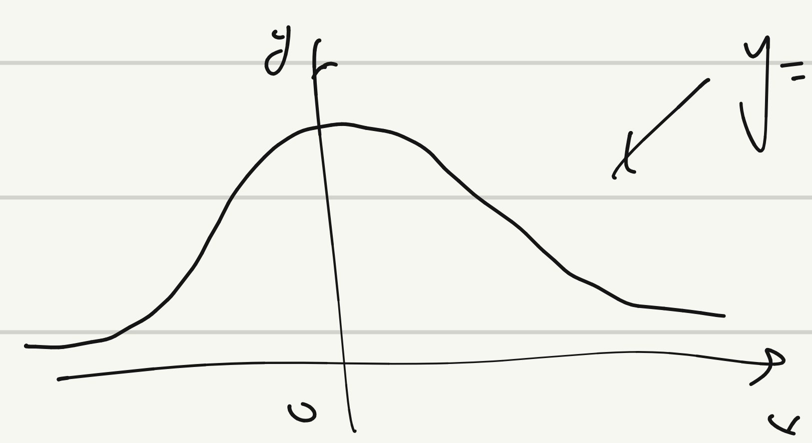
- 얻어진 랜덤 표본으로부터 정규 모집단이 특정 평균 (μ)을 갖는지 검정.
- 대립 가설의 종류
  - ∘ $H_1: \mu \neq k_0$, $H_0: \mu = k_0$; (양측 검정)
  - ∘ $H_1: \mu > k_0$, $H_0: \mu \leq k_0$ } 단측검정 (Upper / Lower - tailed)
  - ∘ $H_1: \mu < k_0$, $H_0: \mu \geq k_0$

※ $\frac{d}{dx}e^x \rightarrow e^x$

$\frac{d}{dx}a^x \rightarrow a^x \ln a$



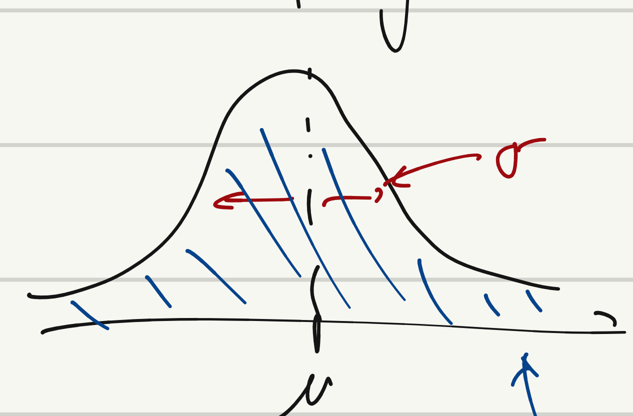
$y = e^{-x} \rightarrow$ ① $y = e^{-|x|} \rightarrow$ ② $y = e^{-x^2}$



$y = e^{-x^2}$

Gaussian Function

$y = e^{-\frac{(x-\mu)^2}{2\sigma^2}}$



면적 = $\sigma\sqrt{2\pi}$

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

이거로 "확률"로 만듦

Gaussian probability
Normal distribution.

◎ 정규분포 (Normal Distribution)의 특징.

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

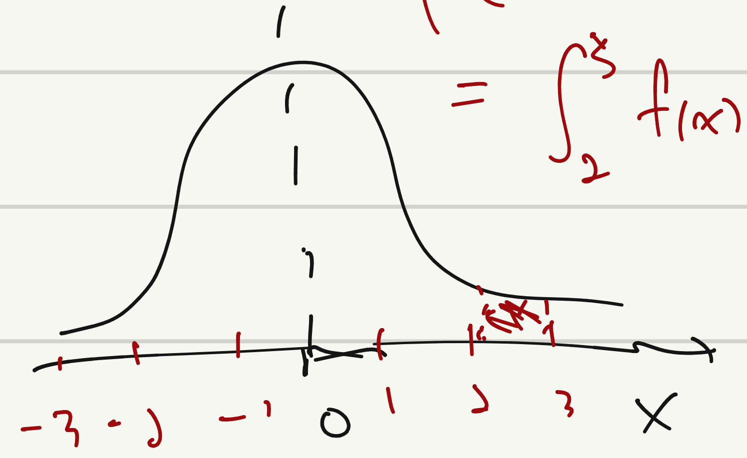
identically independently distributed ← 독립적으로 동일한 분포.

- $X \overset{i.i.d}{\sim} N(\mu, \sigma^2)$, μ: 평균, σ^2 = 분산

X는 정규분포를 따른다.

$P(X=2) = 0$

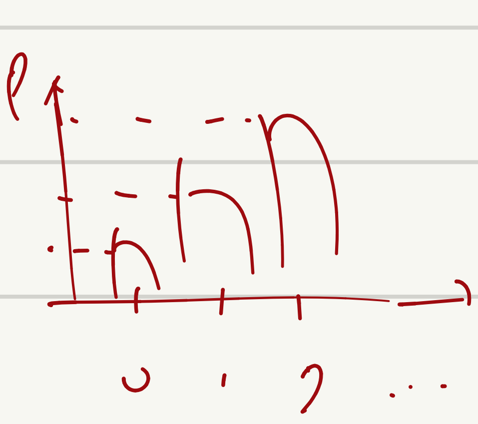
$P(2<X<3) = \int_2^3 f(x)dx$



- 연속형 변수에 대해서 정의
- μ를 중심으로 대칭, 가로는 μ로부터 평균적으로 σ 떨어져 있음.
- AUC 가 확률을 나타냄.

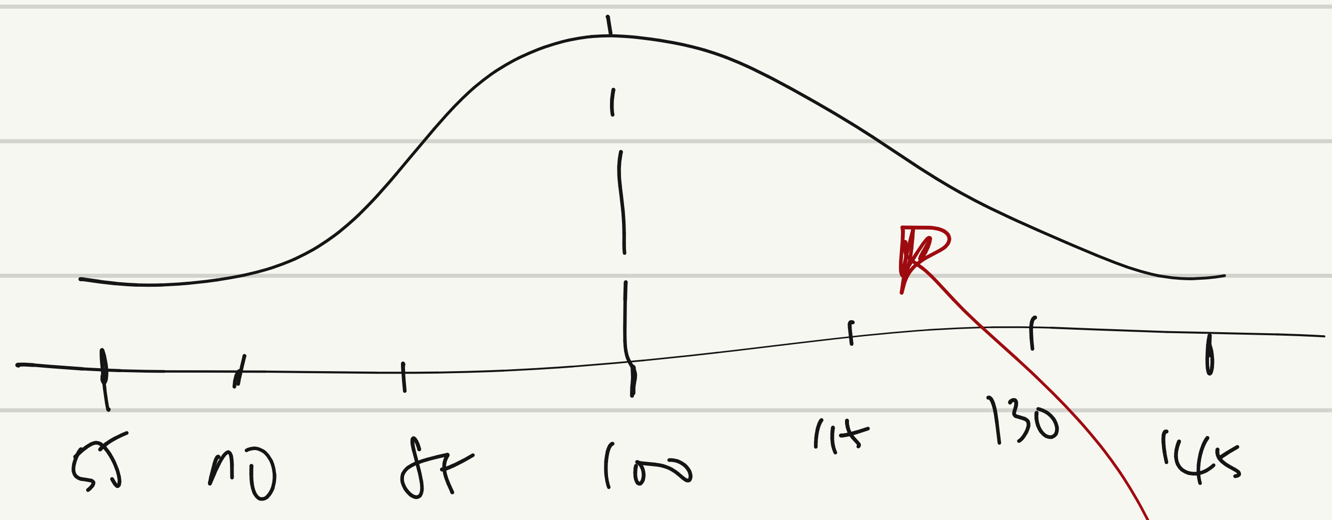
→ 이항분포도 그래프의 높이가 확률.

"이산형 변수"



ⓒ 분위수를 이용한 확률 계산 (e.g. IQ score)

$\mu = 100$. $\sigma = 15$



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

$$= \frac{1}{\sqrt{2\pi} \cdot 15} e^{-\frac{(x-100)^2}{2 \cdot 15^2}}$$

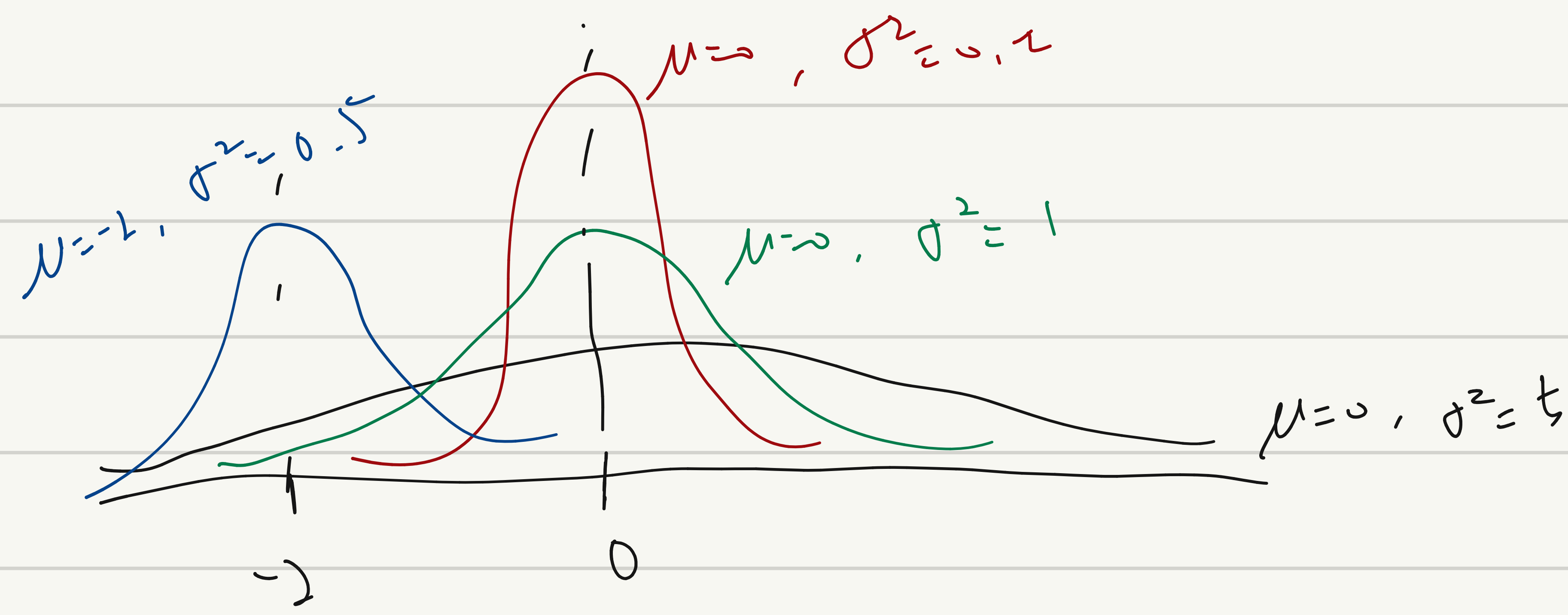
=넓이

∴ 확률을 구하기 위해

Q. $P(115 < x < 130) = ?$

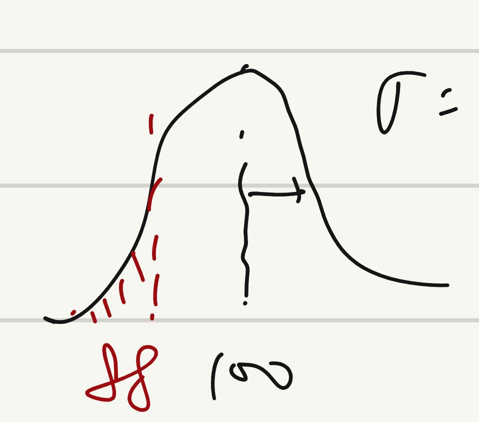
$$\int_{115}^{130} f(x)\,dx = 13.6\%$$
$$= 0.136$$

ⓓ 정규분포의 3요소 (Pdf, μ, σ^2) $X \sim N(\mu, \sigma^2)$



ⓔ 정규분포에 근거한 가설검정의 절차와 유의확률.

- H_1 : 방앗간 빵의 평균 무게는 100g이 아니다. ⇒ $\mu \neq 100$
- H_0 : Not H_1 ⇒ $\mu = 100$ (귀무가설)
- α-value (유의 수준) : 0.05
- 유의 확률 계산
  1. ① 표본 추출 : 빵 n개를 뽑아 평균과 표준편차 계산.
  2. ② 표준 오차
  3. ③ 정규분포의 확률 밀도 함수를 추정
  4. ④ 유의 확률 계산.



$P($유사 자료 $| H_0 =$ 참$) < 0.05$

$P(\bar{x} = 98g \mid \mu = 100g) < 0.05$ → if, $\int_{-\infty}^{98} f(x)\,dx < 0.025$

(예)

∴ 양측검정

→ 귀무가설 기각.

(e.g.) $\bar{x} = 48g$, $s^2 = 4$

$s = \frac{2}{\sqrt{10}}$?

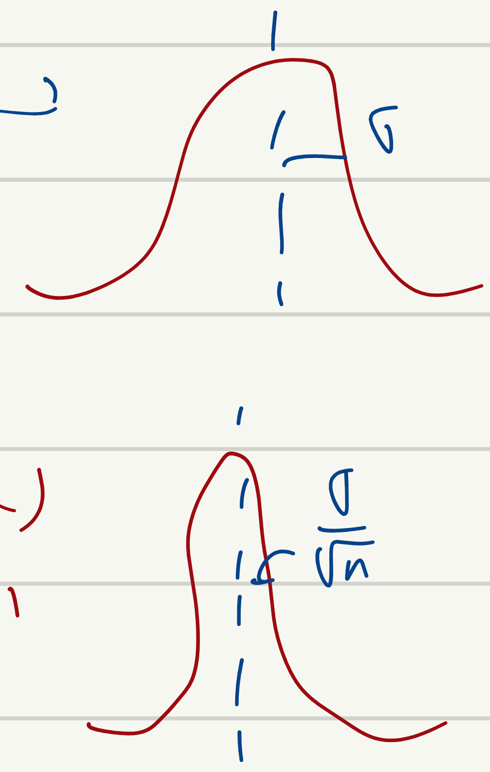
가정: 빵 10개 $(x_1, \dots, x_{10}) \rightarrow \bar{x} = 48g$, $s^2 = 4$

⊗ 표본통계의 분포

표본평균의 분포

Sampling

여러번 재샘플시



$x \sim N(\mu, \sigma^2)$

$\uparrow \bar{x}$ $\uparrow s^2$

$\bar{x} \sim N(\mu, \frac{\sigma^2}{n})$

$x \sim N(\mu, \sigma^2)$ 데이터의 분산

$\uparrow$

$\bar{x}$, s^2

$\bar{x} \sim N(\mu, \frac{\sigma^2}{n})$ 평균의 분산.

$\uparrow$

$\bar{x}$, $\frac{s}{\sqrt{n}}$ ← 표준오차

big data는 검증 X

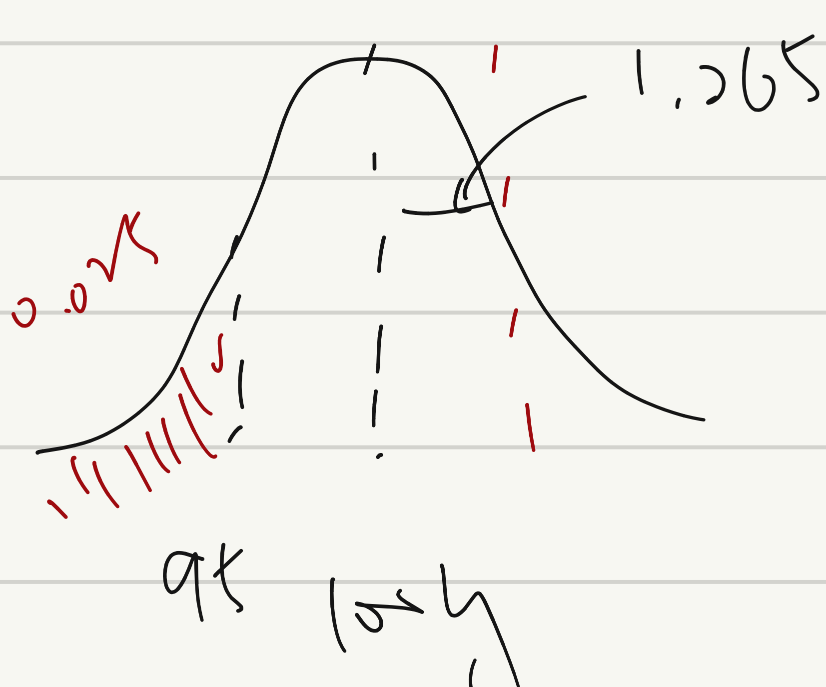
$n \rightarrow \infty$ 분산이 없어짐 ⇒ 신뢰구간이 좁아짐.

∵ $n \rightarrow \infty$ 되면 H_0이 기각.

⇒ $n \rightarrow \infty$ 은 모집단 전체를 sampling 한다는 것과 비슷.

---

from R code



* if, 오차의 결과가 mean의 오른쪽에 존재 → $\{1 - (\ )\} \times 2$ 를 해줘야 함

∵ 반대편의 면적을 구해야 함. 아니면 "1"을 넘어감

* 앙리 푸앵카레 빵집 주인의 일화 ⊛

↑ interesting story

◎ 피니어스 게이지 사건

→ 철도 사고 → 두개골 관통. → 성격이 완전히 바뀜.

→ 이 사건 이후로 뇌과학 활성화

◎ 정규분포의 표준화

- 표준편차, 분산이 다른 경우 → 직접 적분하여 확률계산.
- 정규화를 통해서 쉽게 구함
- 확률 밀도함수의 적분

$$P(X > x_0) = \int_{x_0}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

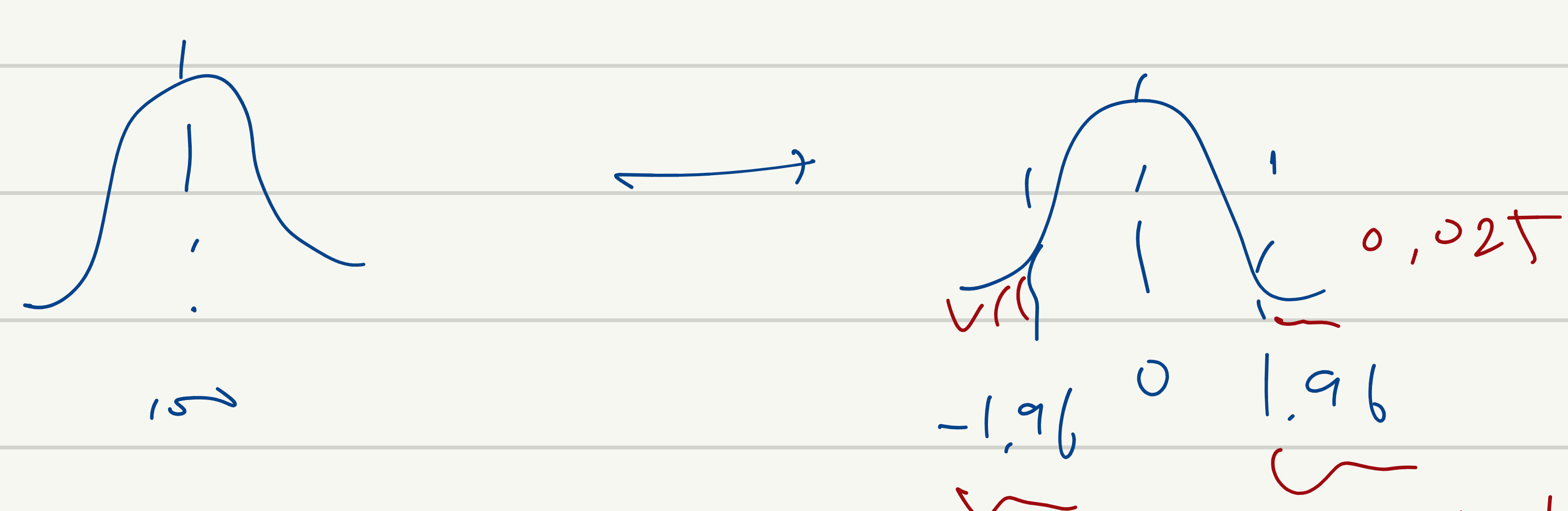
◎ 정규분포의 표준화 방법.

① $X \sim N(\mu, \sigma^2)$, $Z = \frac{X-\mu}{\sigma}$, $Z \sim N(0,1)$

ex

$\boxed{\begin{matrix} 1.96 \\ -1.96 \end{matrix}}$

미리 계산된 값.



기각의 조건

② $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$, $Z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$, $Z \sim N(0,1)$

?.

from 뒤 예제

ex $\bar{X} = 95$, $\mu = 100$, $\frac{\sigma}{\sqrt{n}} = 2.4$

$Z = \frac{95-100}{2.4}$ ⇒ 1.96 or -1.96 초과, 미만 ⇒ 기각.

☆ 확률 계산의 필요성 X

ex

	Math	
A →	80	$\frac{80-\mu_A}{\sigma_A}$
B →	60	$\frac{60-\mu_B}{\sigma_B}$

⇒ Z - transform

Z - score

☆ 표준편차, 분산이 다른 값 비교 ☆

☆ 정규분포의 표준화 ☆

◎ 표준 정규분포의 특징

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

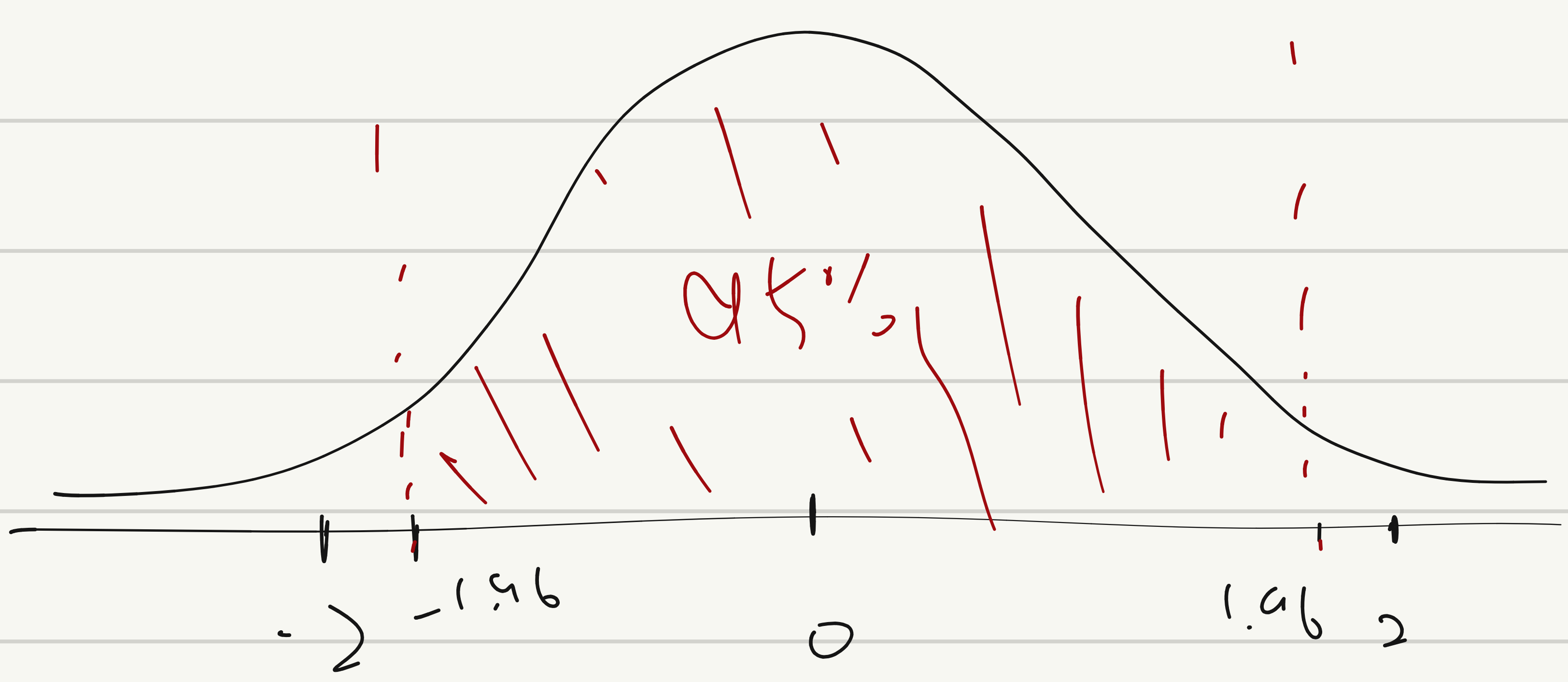
연속함수 이산함수

similarity

$\mu = 0,\ \sigma = 1$

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

- $\mu = 0,\ \sigma = 1$
- z 분포, $-1.96 \sim 1.96$ 사이에 값이 95% 존재. (95% 신뢰구간)



◎

$$P\left(\left|\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right| < 1.96\right) = 0.95$$

$$\downarrow$$

$$P\left(\left|\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right| < z_{\alpha/2}\right)$$

$z_{\alpha/2} = 1.65$, 90%

$z_{\alpha/2} = 1.96$, 95%

$z_{\alpha/2} = 2.58$, 99%

∘ $\alpha = 0.05$, $\alpha/2 = 0.025$

$z_{0.025} = 1.96$

$z_{-0.025} = -1.96$

$$P\left(-z_{\alpha/2} < \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 0.95$$

* 모평균 (μ)의 신뢰구간

n 수가 크면 μ는 수렴

but. n이 작으므로 μ는 추정.

$$P\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = \text{신뢰구간 } (0.95)$$