파이썬 확률 통계

추론과 검정

/* elice */

목차

이산 확률 분포

확률분포

모집단으로부터 얻어지는 상대도수 분포

발생 가능한 모든 사건과 발생 가능성을 나타냄

확률분포

이산 확률 분포

1) 베르누이분포
2) 이항분포
3) 기하분포
4) 포아송분포
5) …

연속 확률 분포

1) 균일분포
2) 정규분포
3) …

이산 확률 분포

확률 변수의 값이 정수와 같이 이산적인 값을 가진 경우

이산 확률 변수라고 부르고 이산 확률 변수의 분포를

이산 확률 분포라고 부름

베르누이 분포

1) 베르누이분포

베르누이 시행

1) 각 시행은 성공과 실패 두 가지 중 하나의 결과를 가짐

2) 각 시행에서 성공할 확률은 p, 실패할 확률은 1-p

3) 각 시행은 서로 독립으로 각 시행의 결과가 다른 시행의 결과에 영향을 미치지 않음

1) 베르누이분포

베르누이 시행의 확률변수 X

확률분포

x	0	1
P(X=x)	1-p	p

확률질량함수

$$f(x) = \begin{cases} p\,, & x = 1 \\ 1-p\,, & x = 0 \end{cases}$$

이항 분포

2) 이항분포

베르누이 시행을 반복했을 때, 성공하는 횟수의 확률분포

이항 실험

성공확률이 동일한 베르누이 시행을
독립적으로 반복하는 실험

이항 확률변수

전체 시행 중 성공의 횟수에 따른
확률변수

2) 이항분포

이항 확률변수 X 의 확률질량함수

$$p(x) = \binom{n}{x} \times p^x \times (1-p)^{n-x}, x = 0,1,2 \dots, n$$

시행 횟수 n은 자연수이며,

성공확률 p는 $0 \leq p \leq 1$을 만족

```
X ~ B(n,p)
```

시행 횟수가 n, 성공확률이 p인 이항분포

2) 이항분포

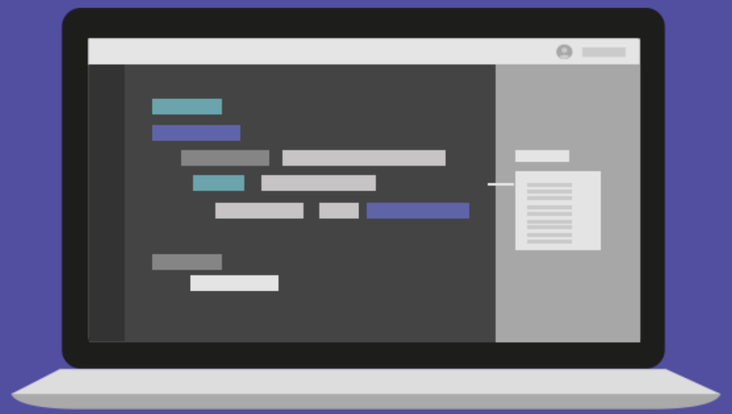
```
stat_bin  = scipy.stats.binom(n, p) #이항분포 확률변수
stat_bin.pmf(x축) # 확률질량함수 시각화
stat_bin.cdf(x축) # 누적분포함수 시각화
np.random.binomial(n, p, size) # 이항분포 랜덤샘플
```

n : 시행 횟수

p : n=1이 나올 확률

size : 표본 추출 작업 반복 횟수

[실습]
이항분포



초기하 분포

3) 초기하분포

유한한 모집단에서 비복원 추출 시, 성공의 횟수의 분포

X : 표본 내에서 **관심있는 범주**(예: 불량품 개수)에
속하는 **구성원소의 수**

불량률 계산 등에서 많이 사용

3) 초기하분포

```
X ~ Hyper(M,n,N)
```

모집단의 크기가 M이고,

표본의 크기가 n,

관심이 있는 범주 (예: 불량품 개수)에 속하는

구성원소의 수가 N인 초기하분포

3) 초기하분포

초기하 확률변수 X 의 확률질량함수

$$p(X = x) = \frac{\binom{D}{x} \times \binom{N-D}{n-x}}{\binom{N}{n}}, x = 0,1,2 \ldots, n$$

여기서 n은 D 혹은 (N-D)보다 작거나 같은 수로 가정

3) 초기하분포

상자 안에 **흰색 공 6개**와 **검은색 공 4개**가 있을 때

5개의 공을 꺼낸 결과 **흰 공이 3개**인 확률은?

3) 초기하분포

10개 중 5개를 뽑는 경우의 수 가운데

흰색 공 **6개 중 3개**를 뽑고

검은색 공 **4개 중 2개**를 뽑을 확률

3) 초기하분포

$$p(X=3)=\frac{\binom{6}{3}\times\binom{4}{2}}{\binom{10}{5}}=\frac{10}{21}$$

3) 초기하분포

```
stat_hyp = scipy.stats.hypergeom(M, n, N) #초기하분포 확률변수
stat_hyp.pmf(x축) # 확률질량함수 시각화
stat_hyp.cdf(x축) # 누적분포함수 시각화
np.random.hypergeometric(ngood, nbad, nsample, size)
#초기하분포 랜덤샘플
```

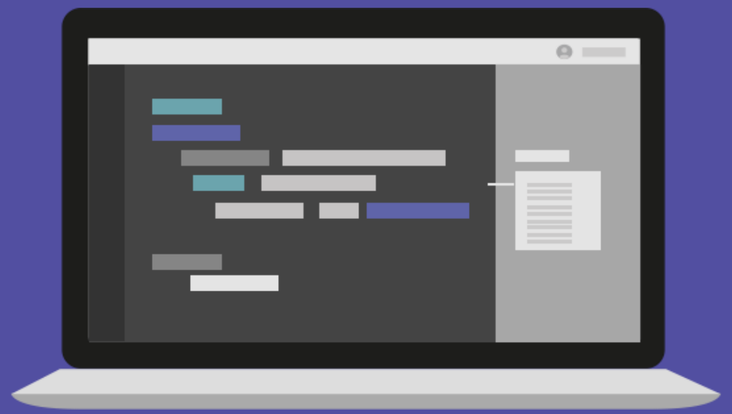
ngood(= n) : 모집단 중 관심 있는 범주에 속하는 구성원소 수

nbad(=M-n) : 관심있는 표본 이외의 개수(ngood + nbad = M)

nsample(=N) : 표본의 크기

size : 표본 추출 작업 반복 횟수

[실습]
초기하분포



포아송 분포

4) 포아송분포

연속된 시간 상에서
발생하는 사건은 매 순간 발생 가능

시행 횟수가 많고 순간의 성공확률은
작기 때문에 이항분포로 설명하기 어려움

4) 포아송분포

단위시간/공간에 드물게 나타나는 사건의 횟수에 대한 확률분포

연속적인 시간에서 **매 순간에 발생할 것**으로

기대되는 평균 발생 횟수를 이용해

주어진 시간에 실제로 발생하는 사건의 횟수 분포

4) 포아송분포

포아송 분포의 예시

일정 시간동안 발생하는 불량품의 수

일정 시간동안 톨게이트를 지나는 차량의 수

일정 페이지의 문장을 완성했을 때 발생하는 오타의 수

4) 포아송분포

$X \sim Poi(\lambda)$

평균적으로 λ 회 발생하는 사건의
발생 횟수에 대한 포아송분포

포아송 확률변수 X 의 확률질량함수

$$p(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}, x = 0,1,2 \ldots$$

4) 포아송분포

이항분포 B(n,p)에서

n이 매우 크고 p가 매우 작은 경우

λ=np인 포아송 분포로 **근사 가능**

균일 분포

연속 확률 분포

확률 변수의 값이 실수 집합처럼 연속적이고

무한개의 경우의 가질 경우

연속 확률 변수라고 부르고 연속 확률 변수의 분포를

연속 확률 분포라고 부름

1) 균일분포

구간 [a,b] 에 속하는 값을 가질 수 있고 그 확률이 균일한 분포

$$X \sim U(a,b)$$

1) 균일분포

정육면체 주사위의 한 면이 나올 확률은 모두 $\frac{1}{6}$ 로 같다

$$P(X = 1,2,3,4,5,6) = \frac{1}{6}$$

1) 균일분포

균일확률변수 X의 확률밀도함수

$a < b$를 만족하는 임의의 두 실수 a,b 에 대해 함수

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \le x \le b \\ 0, & x < a \; or \; x > b \end{cases}$$ 를 정의하면,

$f(x)$를 확률밀도함수로 갖는 연속확률변수가 존재

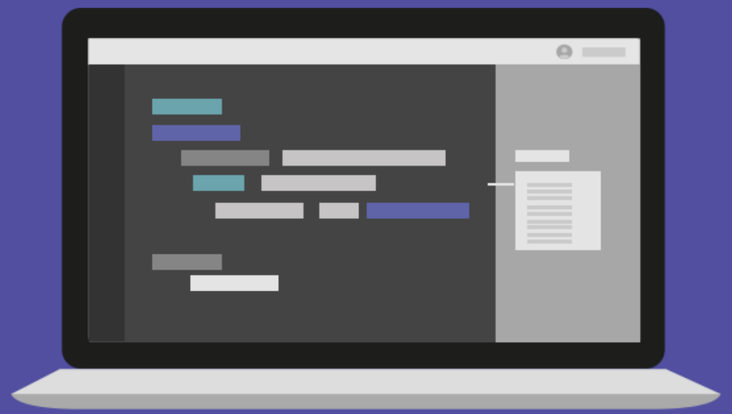
1) 균일분포

```
stat_uni  = scipy.stats.uniform(a, b) #균일분포 확률변수
stat_uni.pmf(x축) # 확률질량함수 시각화
stat_uni.cdf(x축) # 누적분포함수 시각화
np.random.uniform(a,b,n) # 균일분포 랜덤샘플
```

a,b : 균일분포의 구간

n : 표본 추출 작업 반복 횟수

[실습]
균일분포



정규 분포

2) 정규분포

가장 많이 사용되고 유명한 분포

종형 곡선의 분포

평균 뮤(μ)와 표준편차 시그마(σ) 두 모수로 정의

$N(m, \sigma^2)$ 로 표시

정규분포를 나타내는 확률밀도함수

$$f(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(\frac{(x-\mu)^2}{2\sigma^2}\right)$$

2) 표준정규분포

정규분포의 표준분포

평균 뮤(μ) = 0, 표준편차 시그마(σ) = 1로 둔 정규분포 Z

표준정규분포의
확률밀도함수

$$Z = \sigma Z + \mu$$

$$Z = \frac{(X - \mu)}{\sigma}$$

$$f(z|\mu = 0, \sigma = 1) = \frac{1}{\sqrt{2\pi}} \exp(\frac{(x)^2}{2})$$

2) 정규분포

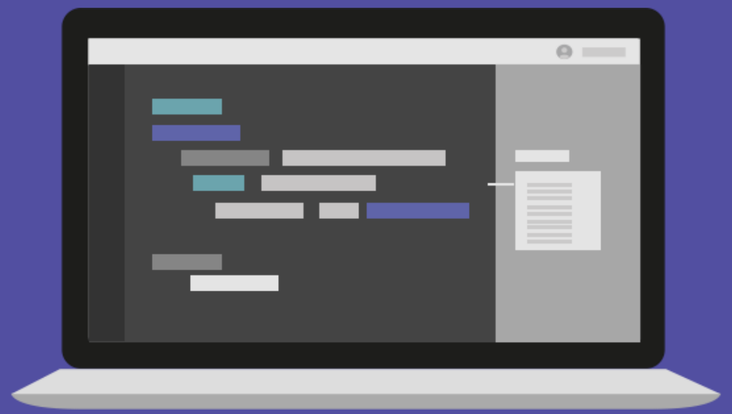
```
stat_nor = scipy.stats.norm(μ, σ) #정규분포 확률변수
stat_nor.pmf(x축) # 확률질량함수 시각화
stat_nor.cdf(x축) # 누적분포함수 시각화
np.random.normal(μ, σ, n) # 정규분포 랜덤샘플
```

μ : 평균

σ : 표준편차

n : 표본 추출 작업 반복 횟수

[실습]
정규분포



모집단과 표본

모집단과 표본

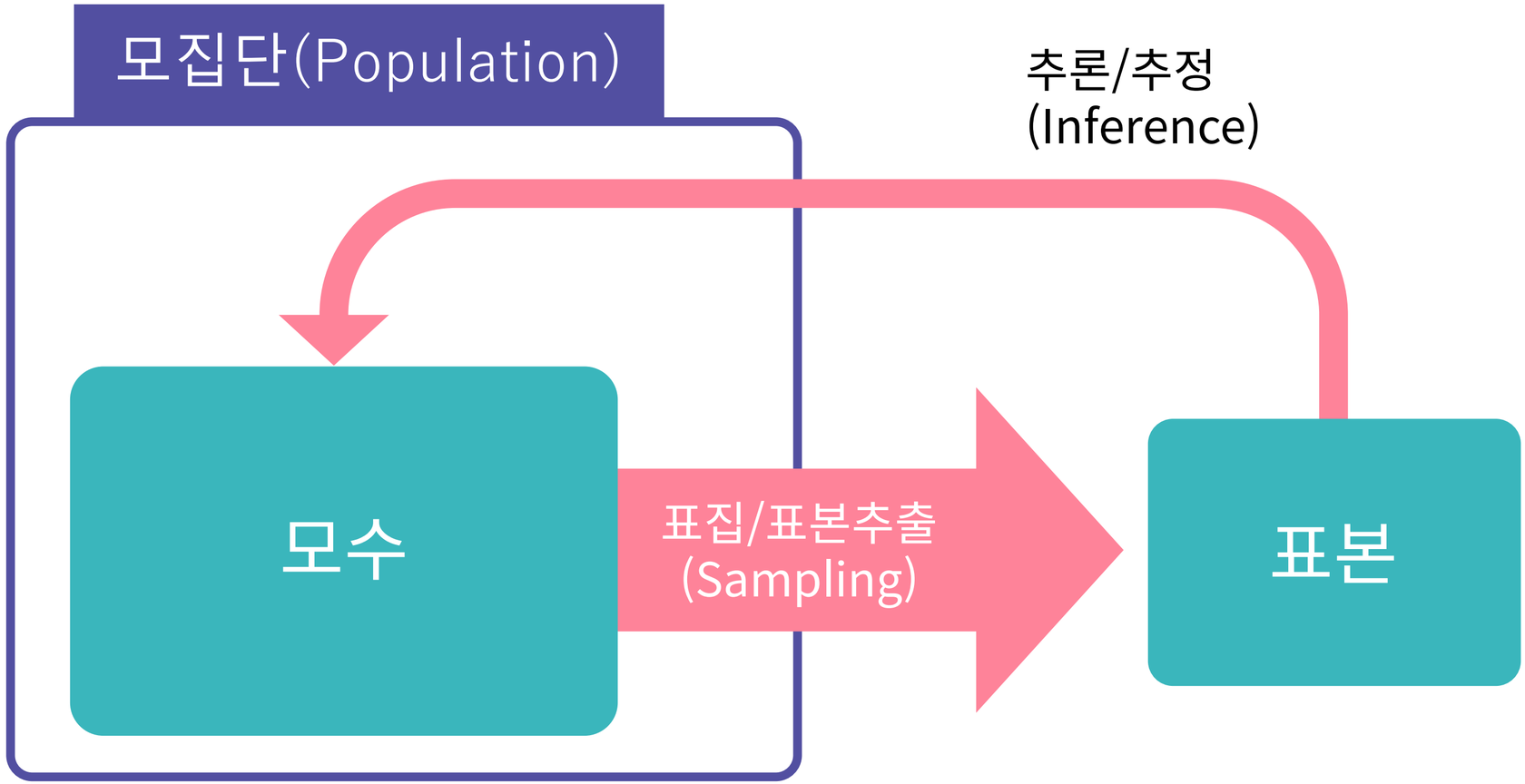
정보를 얻고자 하는 **관심 대상**의 **전체집합**

모집단을 **통째로** 조사하는 것은 어렵다

모집단의 일부를 **표본**으로 추출

표본으로 모집단의 정보를 **추론**함

모집단과 표본



모집단과 표본

모집단(Population)	조사의 관심이 되는 전체 집단
표본(Sample)	모집단에서 일부를 표집(샘플링)하여 실제 조사한 대상
모수(Parameter)	모집단으로부터 계산된 모든 값, 미지의 수
통계량(Statics)	표본으로부터 계산된 모든 값, 모수를 추정

모집단과 표본

표본 조사의 대표적 예시 : **출구조사**

전체 **유권자**(모집단) 중 **임의**로 선택한
출구조사 **대상자**(표본)

출구조사 결과와 실제 선거 결과가 **거의 비슷함**

모집단과 표본

3벌의 셔츠와 2벌의 바지로

옷을 입을 수 있는 경우의 수는?

→ 상의와 하의는 **같이** 입으므로 **곱의 법칙**

3 * 2 = 6개의 경우의 수

모집단과 표본

3벌의 바지와 2벌의 치마로

하의를 입을 수 있는 경우의 수는?

→ 서로 다른 하의는 **같이** 입을 수 없으므로 **합의 법칙**

3 + 2 = 5개의 경우의 수

[퀴즈]
모집단과 표본

통계적 추론

통계적 추론

표본이 갖고 있는 정보를 분석하여 모수를 추론

모수에 대한 가설의 옳고 그름을 판단

표본으로 전체 모집단의 성질을 추론하므로
오류 존재(이 부정확도를 반드시 언급해야 함)

통계적 추론

조사자의 관심에 따라 모수 추정 혹은 가설검정으로 구분

모수 추정

- 모수에 대한 추론 혹은 추론치 제시
- 수치화 된 정확도 제시

가설검정

모수에 대한 여러 가설들이 적합한지 표본으로 판단

모수 추정

모수 추정

추출된 표본으로부터의
모수의 값에 가까울 것이
예상되는 하나의 값을 제시

하나의 값 대신 모수를
포함할 것이 예상되는
적절한 구간을 제시

모평균 점추정

모집단의 모수인 평균 μ의 추정

모집단에서 크기가 n 인 표본을

n개의 확률변수 $X_1, X_2, \ldots, X_n$ 로 표현 했을때,

모평균의 추정량 중, 직관적으로 타당한 것은 표본평균

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \ldots + X_n)$$

구간추정

신뢰구간

- 추정량의 분포를 이용하여 표본으로 부터 모수의 값을 포함하리라 예상되는 구간
- (작은 값(하한) ,큰 값(상한)) 의 형태

신뢰수준

- 신뢰구간이 모수를 포함할 확률을 1 보다 작은 일정한 수준에서 유지할 때 확률이 신뢰수준
- 신뢰수준은 90%, 95%, 99% 등으로 정함

모평균 구간추정

모평균 구간추정

모평균 μ의 신뢰구간

- μ 의 분포 : 모집단의 정규분포, 표준편차(σ) 가 주어짐
- 추정량 $\bar{X}$의 분포 : 평균이 μ, 표준편차가 $\sigma/\sqrt{n}$인 분포 N(0,1)

$$P\left(\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} < z_{\frac{\alpha}{2}}\right) = 1-\alpha$$

모평균 구간추정

- $z_{\frac{a}{2}}$ 는 N(0,1) 의 상위 $\frac{a}{2}$ 의 확률을 주는 값

$$P\left(\bar{X} - z_{\frac{a}{2}}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\frac{a}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

- 모평균 μ 에 대한 신뢰구간

$$\left(\bar{X} - z_{\frac{a}{2}}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\frac{a}{2}}\frac{\sigma}{\sqrt{n}}\right)$$

모평균 구간추정 예제

$N(100, 10^2)$ 인 분포로부터 크기가 15인 표본을 추출해 표본평균 $\bar{x} = 105$ 일 때, 모평균에 대한 95% 신뢰구간 :

$$(\bar{x} - 1.96\frac{10}{\sqrt{15}}, \bar{x} + 1.96\frac{10}{\sqrt{15}})$$

$$= (105 - 1.96\frac{10}{\sqrt{15}}, 105 + 1.96\frac{10}{\sqrt{15}})$$

$$= (99.94, 110.06)$$

가설

가설검정

모집단의 특성이나
모수에 대한 주장이 있을 때,
이 주장의 옳고 그름을
표본자료를 이용하여
판단하는 방법

가설

모수에 대한 주장

통계적 가설검정(검정)

주어진 가설을 표본자료로부터
얻은 정보를 통해 검토하는 과정

귀무가설과 대립가설

귀무가설
(H_0 , Null hypothesis)

- 기본적인 가설

대립가설
(H_1 : Alternative hypothesis)

- 자료를 통하여 입증하고자 하는 가설
- 비교하고 싶은 가설

귀무가설과 대립가설

예) 하나의 동전을 던지면 앞면이 나올 확률을 1/2 이라 가정할 때,

진짜 앞면이 ½의 확률로 나올지에 대한 검정

가설

$H_0 : p = \frac{1}{2}$ vs $H_1 : p \neq \frac{1}{2}$

양측가설과 단측가설

양측가설

$H_0 : p = \frac{1}{2}$ vs $H_1 : p \neq \frac{1}{2}$

단측가설

$H_0 : p = \frac{1}{2}$ vs $H_1 : p > \frac{1}{2}$

OR

$H_0 : p = \frac{1}{2}$ vs $H_1 : p < \frac{1}{2}$

[퀴즈]
가설 설정

통계적 가설검정

통계적 가설검정

설정한 가설에 대한 옳고 그름을 표본자료를 통하여 검정,
두 가설 중 옳다고 판단할 수 있는 하나의 가설을 선택

표본자료가 대립가설을
지지하면 대립가설 채택

표본자료가 대립가설을
지지하지 못하면 귀무가설 채택

통계적 가설검정

대립가설을 채택하는 경우

"귀무가설 H_0 을 기각한다"

귀무가설을 채택하는 경우

"귀무가설 H_0을 기각할 수 없다"

or

"귀무가설 H_0 을 채택한다"

귀무가설을 기준으로 한 표현 사용

오류의 종류

오류의 종류

1종의 오류(α) 귀무가설이 참일 때 귀무가설을 기각하는 경우

2종의 오류 (β) 귀무가설이 거짓일 때 귀무가설을 채택하는 경우

		가설검정 결과	
		H_0채택	H_0기각
실제상태	H_0= 참	옳은 결정	잘못된 결정 (1종의 오류)
	H_0 = 거짓	잘못된 결정 (2종의 오류)	옳은 결정

오류의 종류

가설검정은 표본자료만으로 모집단에 대한 가설을 검토하므로 오류 존재

바람직한 가설검정은 두 오류를 최소화하는 것

두 오류를 동시에 최소화하는 검정은 존재하지 않거나 찾기 어려움

제 1종의 오류를 범할 확률과 제 2종의 오류를 범할 확률은 반비례 관계

오류의 종류

1종의 오류를 범할 때 더 큰 손실이나 비용이 발생하는 경우가 많음

예) H_0 : 새로운 약의 치료율이 기존 약보다 높지 않다.

H_1 : 새로운 약의 치료율이 기존 약보다 높다.

1종의 오류

- 새로운 약의 치료율이 기존 약보다 높지 않음에도 불구하고 높다고 잘못 판단
- 기존 약 대신 새로운 약 생산

2종의 오류

- 새로운 약의 치료율이 기존 약의 것보다 높으나 높지 않다고 잘못 판단
- 새로운 약 대신 기존 약 생산

오류의 종류

전통적인
통계학에서의
검정

제 1종의 오류를
범할 확률을
우선 최소화

이후 제 2종의
오류 범할
확률을 최소화

유의수준

Significance level, α

제1종의 오류를 범할 확률에 대한 최대 허용한계 고정값

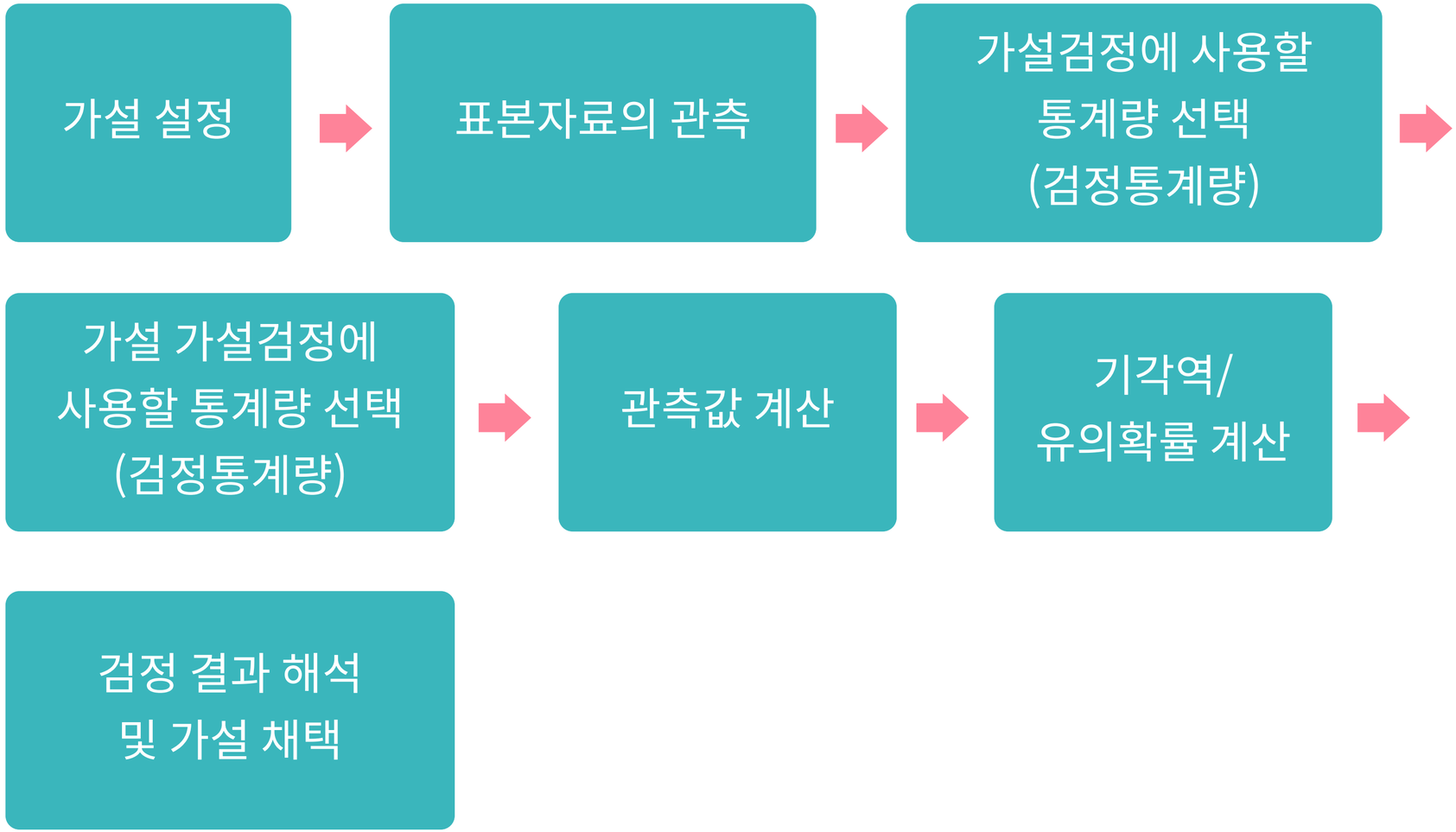
일반적으로 유의수준 α의 값으로

0.01~ 0.10 사이의 작은 값을 사용

[퀴즈]
오류의 종류

검정의 종류와 과정

가설검정과정


가설 설정
표본자료의 관측
가설검정에 사용할
통계량 선택
(검정통계량)
가설 가설검정에
사용할 통계량 선택
(검정통계량)
관측값 계산
기각역/
유의확률 계산
검정 결과 해석
및 가설 채택

검정통계량

가설검정에 사용되는 통계량

가설검정의 결과를 결정하는데 이용되는 표본의 함수

$\bar{X}$를 관측하여 그 값으로부터 μ에 대한

가설 검정을 결정할 때 검정통계량으로 사용

검정을 위한 기준

1. 기각역
2. 유의확률(P-value)

기각역

$\bar{X}$ 가 취하는 구간 중에서 H_0을 기각하는 구간

$R: \bar{X} \leq c$ 로 표현

$\bar{X}$가 c이하면 H_0을 기각한다고 판단

기각역의 올바른 선택이 검정의 가장 중요한 부분

바람직한 기각역은 두 오류를 범할 확률을 최소화하는 것

유의확률

- 표본자료가 대립가설을 지지하는 정도를 0과 1사이의 숫자로 나타낸 최소의 유의수준 값
- P-value 라고 부르기도 함
- 표준정규분포표를 이용해 P값을 구해야 함

유의수준과 P값을 비교

유의수준 > P값인 경우	유의수준 < P값인 경우
H_0를 기각	H_0를 기각할 수 없음

가설검정종류

1. 이항 검정
2. 모평균 가설검정

이항 검정

이항분포를 이용하여 베르누이 확률변수의 모수 p 에 대한 가설 조사

베르누이 값을 가지는 확률변수의 분포를 판단

예) 어떤 동전을 던질 때, 앞면이 나올 확률 $p = 0.5$ 인
공정한 동전인지 알아보는 검정
귀무가설 : $p = 0.5$ vs 대립가설 : $p \neq 0.5$

이항 검정

```
scipy.stats.binom_test(x, n, p, alternative=‘’)

# 이항검정의 유의확률을 구해주는 함수
```

x = 검정통계량, 1이 나온 횟수

n = 총 시도 횟수

p = 모수 p 값

양측검정: alternative = 'two-sided'

단측검정: alternative = 'one-sided'

모평균 가설검정

표본의 크기가 클 때, 모평균 μ 이 정규분포를 따른다는 가정하에
중심극한정리에 의해 정규분포에 근사함

가설검정을 하기 위한 검정통계량 $\bar{X}$ 를 표준화시키면

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$$

모평균 가설검정

단측검정

가설
$\mu = \mu_0$ VS $\mu < \mu_0$

기각역 $R : Z \leq -z_\alpha$

가설
$\mu = \mu_0$ VS $\mu > \mu_0$

기각역 $R : Z \geq -z_\alpha$

양측검정

가설
$\mu = \mu_0$ VS $\mu \neq \mu_0$

기각역 $R : |Z| \geq z_{\frac{\alpha}{2}}$

모평균 가설검정

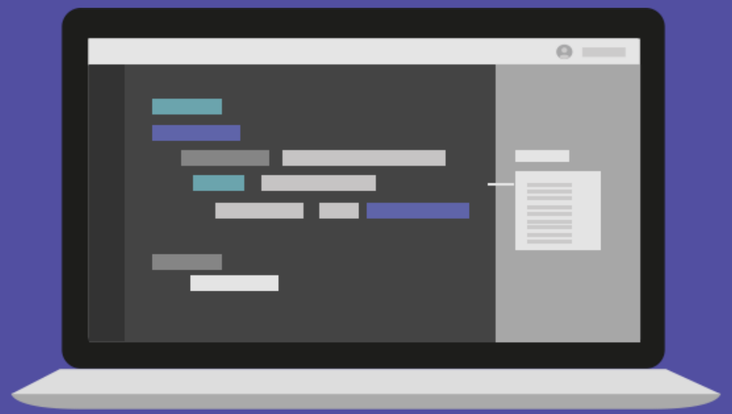
```
def ztest(stat, mu, sigma):
    z = (stat.mean() - mu) / (sigma/np.sqrt(len(stat)))
    return (2 * (1-sp.stats.norm.cdf(z)))
# 모평균 가설 검정 함수. 유의확률 출력
```

stat : 검정통계량

mu : 모평균

sigma : 모표준편차

[실습]
이항 검정



[실습]
가설검정