기초 통계 분석 이론

A 조
[발표자 : 황예원]

CDF

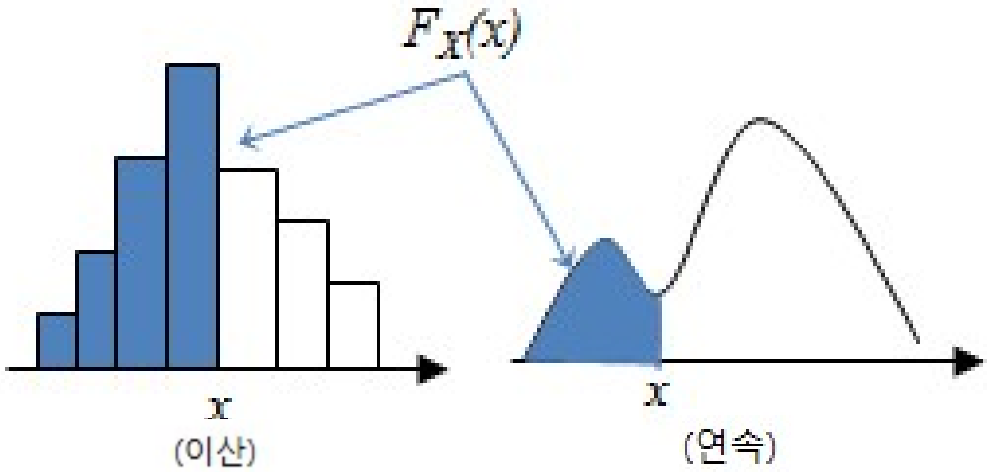
1.1. CDF (Cumulative Probability Density Function, 누적 확률 밀도 함수)

확률 변수 X에 대한 누적 확률 분포 F(x) 의 수학적 정의는 다음과 같다.

$$F(x) = P(\{X < x\}) = P(X < x)$$

ex. 누적 확률 분포 표시의 예

- $F(-1)$: 확률 변수가 $-\infty$ 이상 -1 미만인 구간 내에 존재할 확률 즉, $P(\{-\infty \leq X < -1\})$
- $F(0)$: 확률 변수가 $-\infty$ 이상 0 미만인 구간 내에 존재할 확률 즉, $P(\{-\infty \leq X < 0\})$
- $F(1)$: 확률 변수가 $-\infty$ 이상 1 미만인 구간 내에 존재할 확률 즉, $P(\{-\infty \leq X < 1\})$

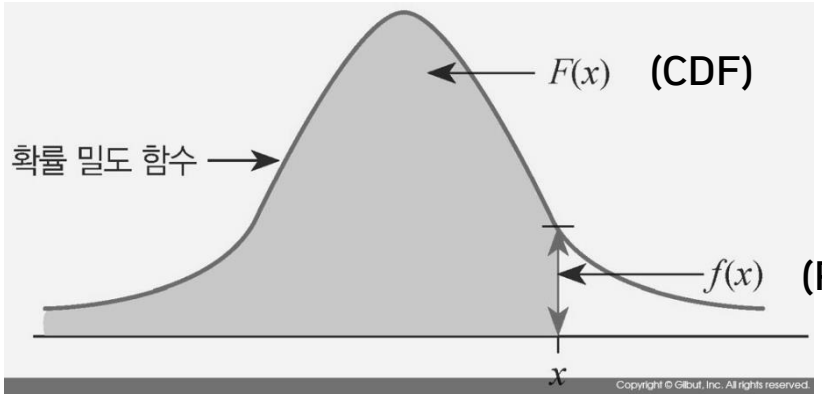


1.2. PDF (Probability Density Function, 확률 밀도 함수)

기울기를 구하는 수학적 연산이 미분(differentiation)이므로 확률 밀도 함수는 누적 분포 함수의 미분으로 정의한다.

$$\frac{dF(x)}{dx} = f(x)$$

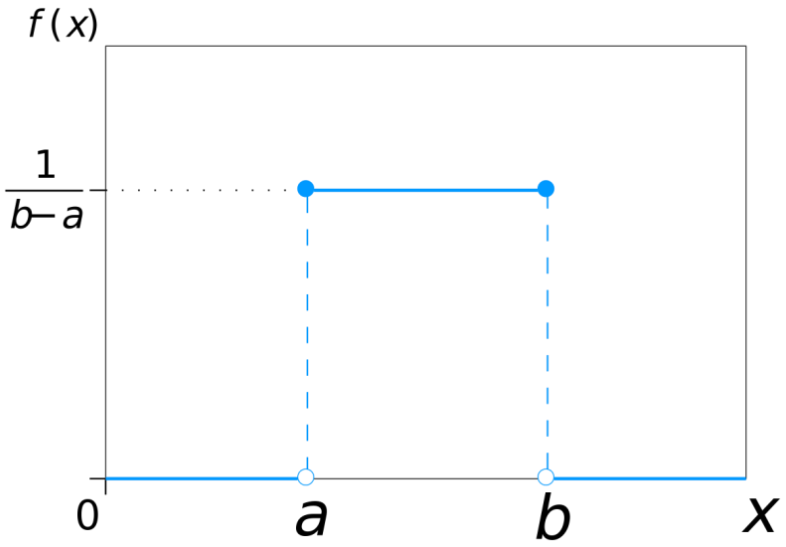
확률 밀도 함수는 특정 확률 변수 구간의 확률이 다른 구간에 비해 상대적으로 얼마나 높은가를 나타내는 것이며 그 값 자체가 확률은 아니다라는 점을 명심해야 한다.



연속 균등 분포

2.1. 연속 균등 분포 (Continuous Uniform Distribution)

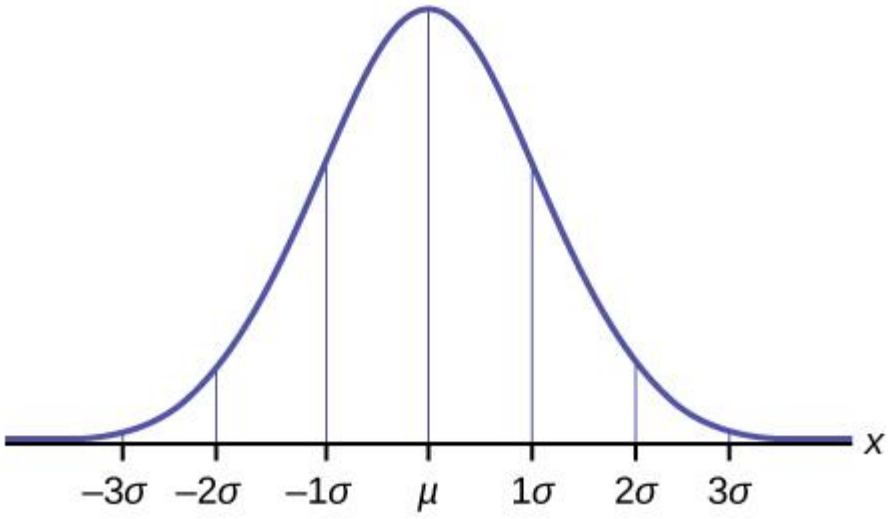
연속 균등 분포는 연속 확률 분포로, 분포가 특정 범위 내에서 균등하게 나타나 있을 경우를 가리킨다.
이 분포는 2개의 매개변수 a, b를 받으며, 이때 [a, b] 범위에서 균등한 확률을 가진다. 보통 기호로 U(a, b)로 나타낸다.



정규 분포

2.2. 정규 분포 (Normal Distribution)

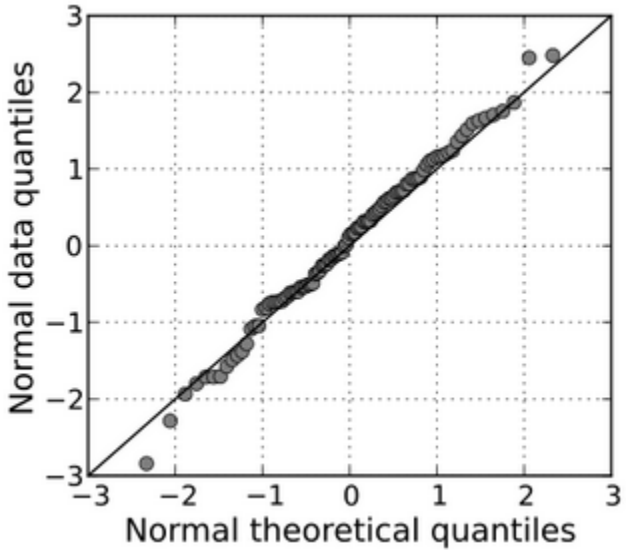
정규 분포(Normal Distribution)는 연속 확률 분포 중 하나로, 수집된 자료의 분포를 근사하는 데에 자주 사용된다. 정규 분포는 2개의 매개 변수 평균 μ과 표준 편차 σ에 대해 모양이 결정되고, 이 때의 분포를 $N(\mu, \sigma 2)$ 로 표기한다.
특히 평균이 0이고, 표준 편차가 1인 정규분포를 N(0, 1)을 표준 정규 분포라고 한다.



2.3. Q-Q 도 (Quantile-Quantile Plot)

Q-Q plot은 데이터가 특정 분포를 따르는지를 시각적으로 검토하는 방법이다.
비교하고자 하는 분포의 분위수끼리 좌표 평면에 표시하여 그린 그림이다.

비교하고 있는 두 분포가 비슷하면, 거의 Q-Q plot은 y=x 선 위에 있어야 하고, 두 분포가 선형 관계라면, 어떤 선에 점들이 오겠지만, 이것이 꼭 y=x 선위에 있을 필요는 없다.



이항 분포

2.4. 이항 분포 (Binomial Distribution)

한 번의 시행에서 사건 A가 일어날 확률이 p로 일정할 때, n번의 독립시행에서 사건 A가 일어나는 횟수를 X라고 하자. 이때 확률변수 X가 가질 수 있는 값은 0, 1, 2, ... , n이며, 그 확률질량함수는 다음과 같다.

$$P(X = x) = {}_nC_x p^x q^{n-x} \quad (x = 0, 1, 2, \ldots, n \ , \ q = 1 - p)$$

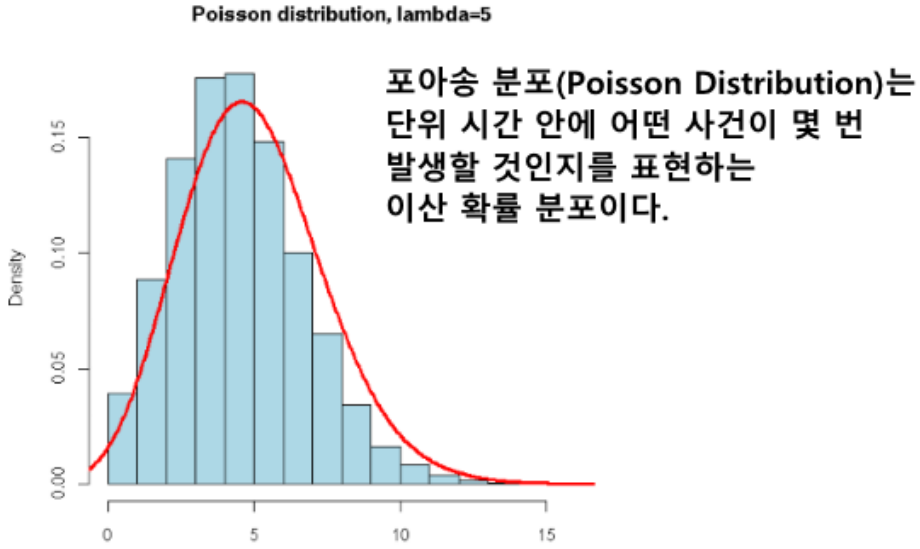
이와 같은 확률분포를 이항분포라고 하며, 이것을 기호로 B(n, p)와 같이 나타내고, 확률변수 X는 이항분포 B(n, p)를 따른다고 한다.
(n : 시행 횟수, p : 각 시행에서 사건 A가 일어날 확률)

푸아송 분포

2.5. 푸아송 분포 (Poisson Distribution)

푸아송 분포는 단위 시간 안에 어떤 사건이 몇 번 발생할 것인지를 표현하는 이산 확률 분포다. 확률변수 X가 포아송 확률변수이고, 모수가 λ인 확률질량함수 f(x)는 다음과 같다.

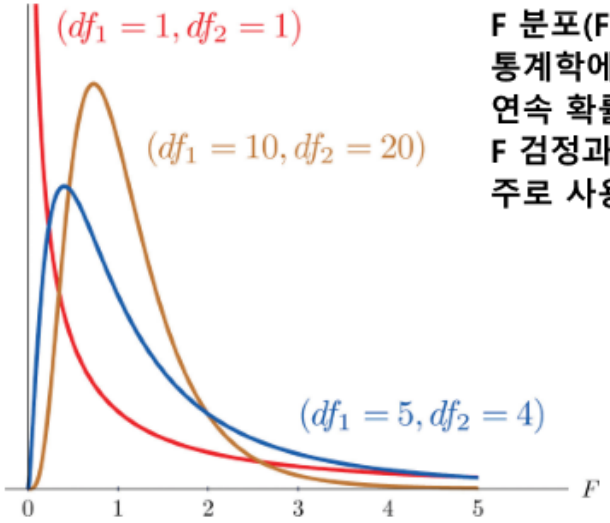
$$f(x) = \frac{e^{-\lambda} \cdot \lambda^{x}}{x!}$$



F분포

2.6. F 분포 (F-Distribution)

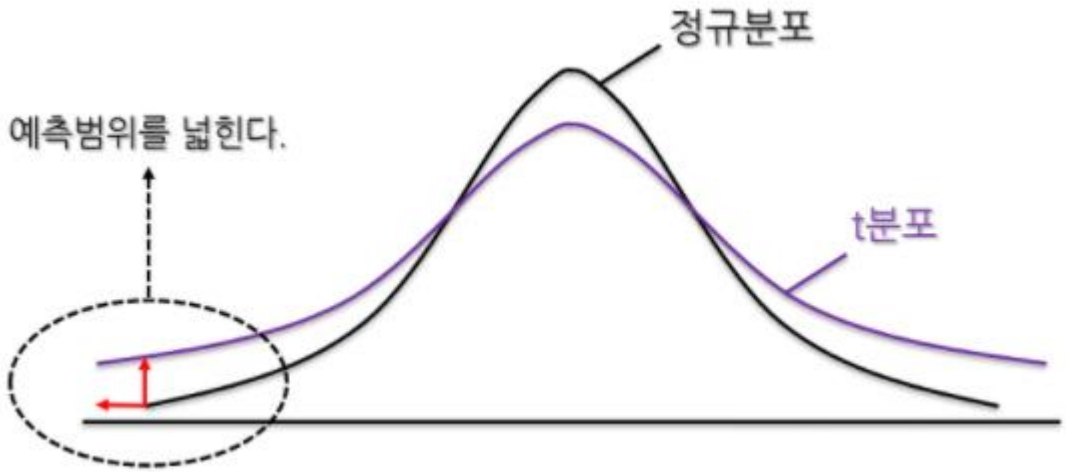
F 분포는 정규분포를 이루는 모집단에서 독립적으로 추출한 표본들의 분산비율이 나타내는 연속 확률 분포다. F 분포의 쓰임새는 2가지 이상의 표본집단의 분산을 비교하거나 모집단의 분산을 추정할 때, 쓰인다.
즉, 2 개 이상의 표본평균들이 동일한 모평균을 가진 집단에서 추출되었는지 아니면 서로 다른 모집단에서 추출된 것인지를 판단하기 위하여 이용하는 것이 F 분포이다.



F 분포(F-distribution)은 통계학에서 사용되는 연속 확률 분포로, F 검정과 분산분석 등에서 주로 사용된다.

2.7. t 분포 (Student t-Distribution)

t 분포는 연속확률분포이면서 표본분포로, 정규분포와 매우 비슷한 분포다. 정규분포는 표본의 수가 적으면 신뢰도가 낮아진다는 단점이 있는데, 이를 해결하기 위해 예측 범위가 넓은 분포를 사용하게 되는데, 이 분포가 바로 t 분포다
신뢰구간이나 가설검정도 표본의 데이터 수가 많아야 신뢰도가 올라가는데, 표본의 수가 적어서 30개 미만일 경우 정규분포 대신 예측 범위가 넓은 t 분포를 사용한다.



단순 임의 추출 / 층화 임의 추출

3.1. 단순 임의 추출 (Random Sampling)

단순 임의 추출은 모집단의 각각의 요소 또는 사례들이 표본으로 선택될 확률이 같게 되는 표본 추출법이다. 유한모집단에서 n개의 추출단위로 구성된 모든 부분 집합들이 표본으로 선택될 확률이 같도록 설계된 표본 추출 방법을 의미한다.

3.2. 층화 임의 추출 (Stratified Random Sampling)

층화 임의 추출은 모집단을 먼저 중복되지 않도록 층으로 나눈 다음 각 층에서 표본을 추출하는 방법이다. 층을 나눌 때, 층내는 동질적(homogeneous), 층간은 이질적(heterogeneous) 특성을 가지도록 하면 적은 비용으로 더 정확한 추정을 할 수 있으며, 전체 모집단 뿐만 아니라 각 층의 특성에 대한 추정도 할 수 있다.

데이터가 중첩 없이 분할될 수 있는 경우 혹은 분할의 성격이 명확히 다른 경우 층화 임의 추출을 수행하여 더 정확한 결과를 얻을 수 있다. 또한 각 층에 대한 추정이 가능해지는 장점이 있다.

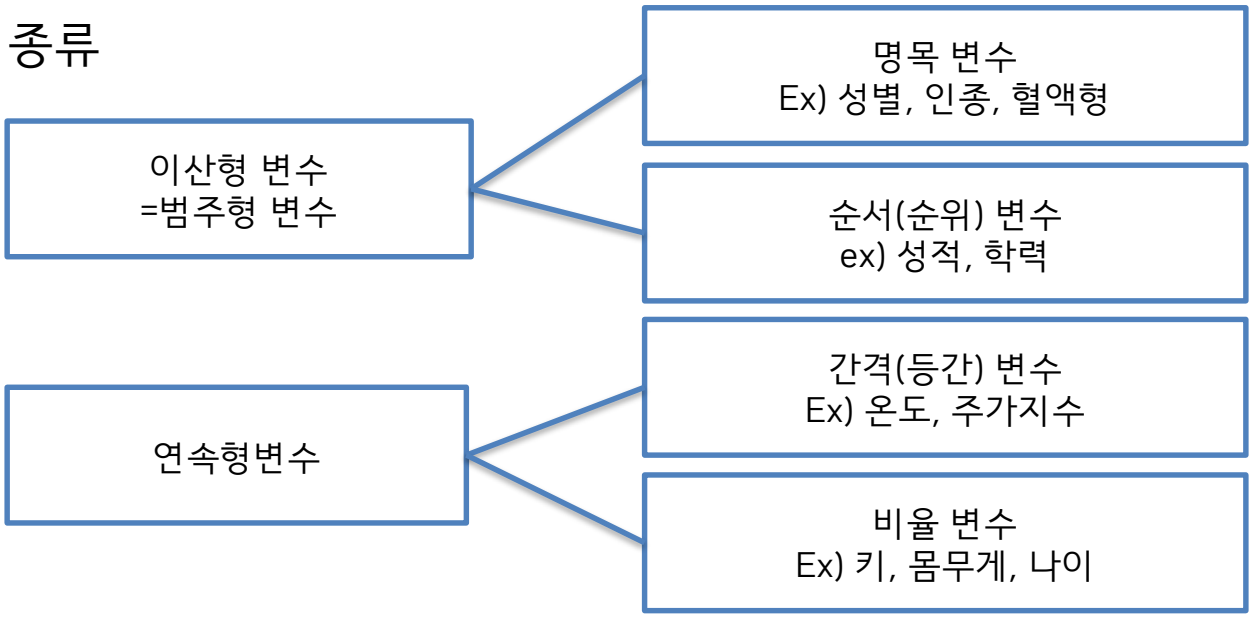
3.3. 계통 임의 추출 (Systematic Sampling)

계통 임의 추출은 모집단의 임의 위치에서 시작해 매 k 번째 항목을 표본으로 추출하는 방법이다. 매우 단순한 방법이지만 랜덤 모집단의 경우에는 단순 임의 추출방법과 동일한 효과를 보이고, 데이터가 순서대로 나열된 순서 모집단(Ordered Population)의 경우 단순 임의 추출보다 효율성이 높다. 하지만 데이터에 일종의 주기성이 존재한다면 편향된 표본을 얻게 된다.

4. 분할표

분할표란?

변수의 종류



분할표란?
이산형 자료의 도수를 표 형태로 기록한 것.

분할표		사후세계 믿음		
		예	아니오	전체
성별	남	435	147	582
	여	375	134	509
	전체	810	281	1091

4. 분할표

분할표에 대한 검정

1) 독립성 검정 : 각 변수간에 의존관계가 있는가
ex) 성별과 사후세계 믿음 관계는 독립인가

2) Fisher's exact test : 표본의 수가 작은 경우
-> 독립성, 적합도 검정이 어려워진다는 문제 발생

3) 적합도 검정 : 관측값이 특정 분포를 따르는가
Ex) 사후세계 믿음 비율은 4:2:3:1의 비율을 따르는가

4) McNemar test : paired 표본인 경우
Ex) 벌금부과 전과 후의 안전벨트 착용자의 수는 어떻게 달라졌는가

		사후세계 믿음		
		예	아니오	전체
성별	남	435	147	582
	여	375	134	509
	전체	810	281	1091

Paired data 안전벨트 착용자 수		벌금부과 후		
		gnnd	Bad	전체
벌금부과 전	Gnnd	435	147	582
	Bad	375	134	509
	전체	810	281	1091

4. 분할표

독립성 검정 : 각 변수는 서로 독립인가

ex) 성별과 사후세계 믿음 관계는 독립인가

도수		사후세계 믿음		
		예	아니오	전체
성별	남	435	147	582
	여	375	134	509
	전체	810	281	1091

확률		사후세계 믿음		
		예	아니오	전체
성별	남	435	147	582
	여	375	134	509
	전체	810	281	1091

도수		사후세계 믿음		
		예	아니오	전체
성별	남	o_{11}	o_{12}	o_{1+}
	여	o_{21}	o_{22}	o_{2+}
	전체	o_{+1}	o_{+2}	o

확률		사후세계 믿음		
		예	아니오	전체
성별	남	p_{11}	p_{12}	p_{1+}
	여	p_{21}	p_{22}	p_{2+}
	전체	p_{+1}	p_{+2}	1

귀무가설 H0 : 두 변수는 독립이다. <=> $p_{ij} = p_{i+} * p_{+j}$ (all i, j)
대립가설 H1 : 두 변수는 독립이 아니다.

독립성 검정 -> 카이제곱 검정

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim x^2_{(I-1)(J-1)}$$

단, $\hat{E}_{ij} \geq 5$인 경우, 즉 대표본인 경우

O(observation) : 관측도수
E(expectation) : 기대도수

I : 행 개수
J : 열 개수

→ E_{ij}를 구할 수 없으므로 E_{ij}의 추정값으로 다음을 사용한다.

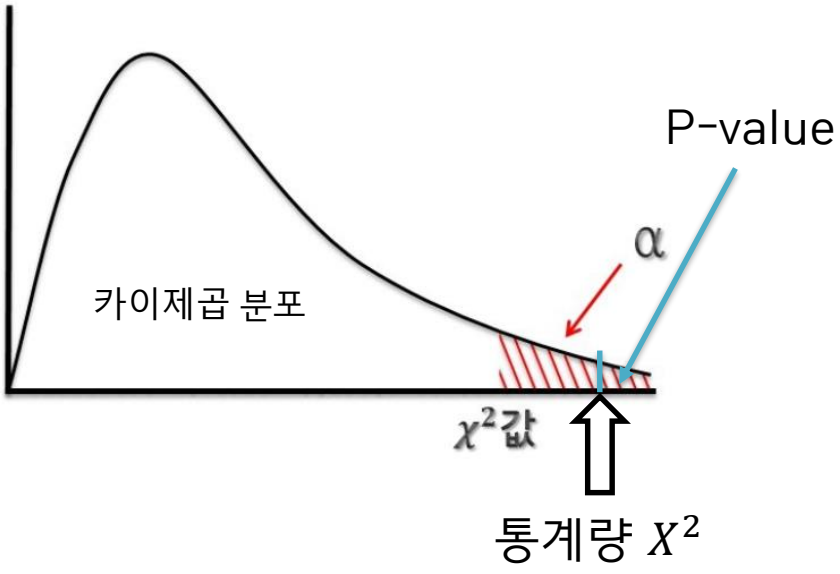
$$\hat{E}_{ij} = O * p_{i+} * p_{+j}$$

→ $(O_{ij} - E_{ij})^2$ 차이가 크다 ⇔ 관측도수와 기대도수 차이가 많이 난다.

4. 분할표

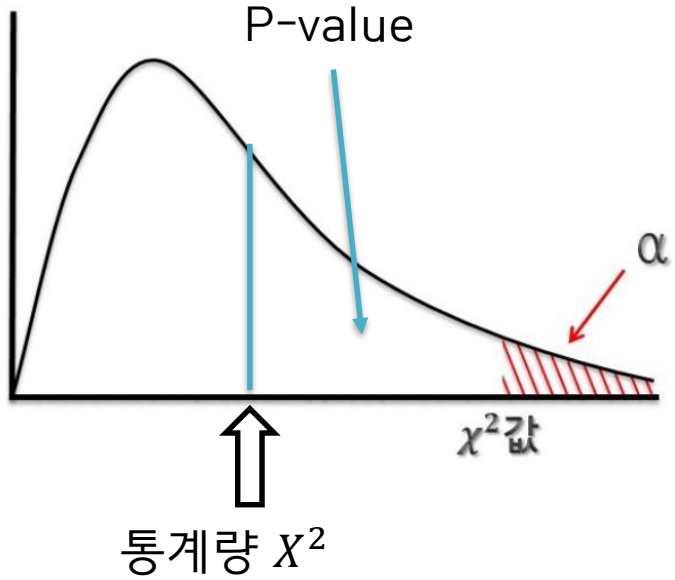
독립성 검정 -> 카이제곱 검정

통계량 X^2에 대한 p-value < 유의수준α
: 귀무가설H0를 기각한다.



두 변수는 독립이 아니다.
성별과 사후세계 믿음은 어떤 관계가 있다.

통계량 X^2에 대한 p-value > 유의수준α
: 귀무가설H0를 기각할만한 충분한 증거가 없다.



두 변수는 독립이다.
성별과 사후세계 믿음은 관련이 없다.

적합도 검정 : 종류

1) 카이제곱 검정
: 관측값이 특정 분포를 따르는가 (가정된 확률이 정해져 있는 경우)
Ex) 사후세계 믿음 비율은 4:2:3:1 의 비율을 따르는가

2) Shapiro-Wilk 검정
: 표본이 정규분포로부터 추출된 것인가

3) Kolmogorov-Smirnov 검정
: 표본이 어떤 특정한 분포로부터 추출된 표본인가
-> 표본의 경험 분포 함수와 이론적 분포의 누적 분포 함수를 비교하여 검정

4. 분할표

적합도 검정 -> 카이제곱 검정

카이제곱 검정 : 관측값이 특정 분포를 따르는가
Ex) 사후세계 믿음 비율은 4:2:3:1 의 비율을 따르는가

도수		사후세계 믿음		
		예	아니오	전체
성별	남	435	147	582
	여	375	134	509
	전체	810	281	1091

확률		사후세계 믿음		
		예	아니오	전체
성별	남	435	147	582
	여	375	134	509
	전체	810	281	1091

도수		사후세계 믿음		
		예	아니오	전체
성별	남	o_{11}	o_{12}	o_{1+}
	여	o_{21}	o_{22}	o_{2+}
	전체	o_{+1}	o_{+2}	o

확률		사후세계 믿음		
		예	아니오	전체
성별	남	0.4	0.2	
	여	0.3	0.1	
	전체			1

귀무가설 H0 : 각 관측값은 특정 분포를 따른다.
대립가설 H1 : not H0

4. 분할표

적합도 검정 -> 카이제곱 검정

피어슨 카이제곱 통계량 X^2은 카이제곱 분포를 근사적으로 따른다.

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim x^2_{(I*J-1)}$$

$E_{ij} = o * 각\ 기대되는\ 확률$

단, $E_{ij} \geq 5$인 경우, 즉 대표본인 경우

O(observation) : 관측도수
E(expectation) : 기대도수

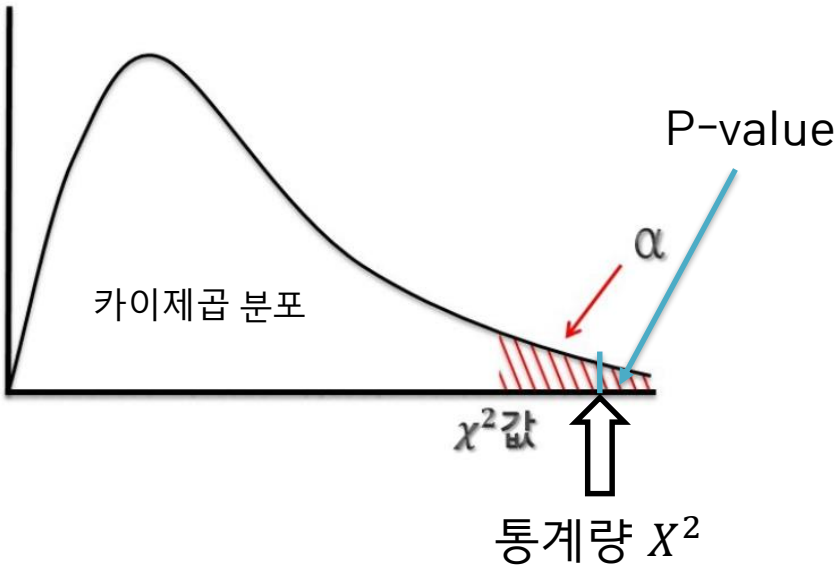
I : 행 개수
J : 열 개수

→ $(O_{ij} - E_{ij})^2$ 차이가 크다 ⇔ 관측도수와 기대도수 차이가 많이 난다.

4. 분할표

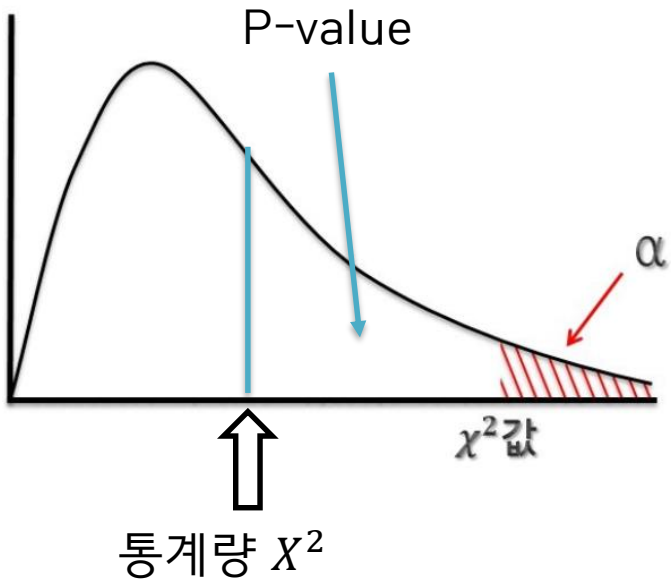
적합도 검정 -> 카이제곱 검정

통계량 X^2에 대한 p-value < 유의수준α
: 귀무가설H0를 기각한다.



각 관측값의 분포는 특정 분포를 따르지 않는다.

통계량 X^2에 대한 p-value > 유의수준α
: 귀무가설H0를 기각할만한 충분한 증거가 없다.



각 관측값의 분포는 특정 분포를 따른다.

상관계수

상관계수란?

: 두 변수 사이의 관계의 정도를 나타내는 수치

- 피어슨 상관계수
- 스피어만 상관계수
- 켄달의 순위 상관계수

-> 모두 [-1, 1] 사이의 값을 가진다.

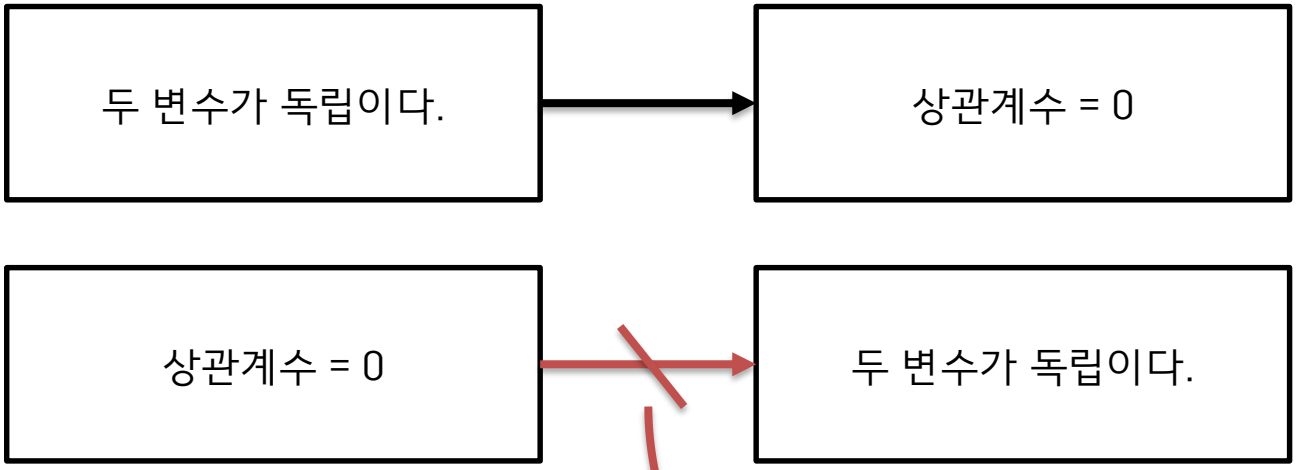
피어슨 상관계수

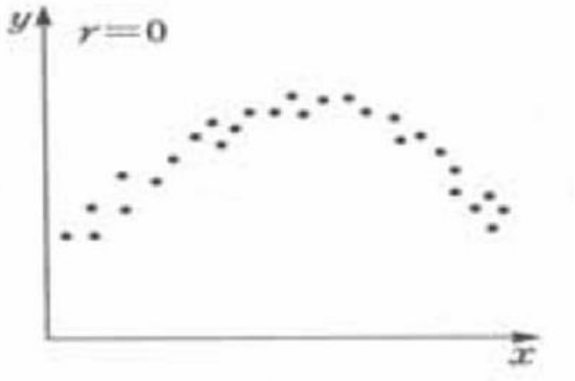
두 변수 사이의 선형관계를 나타내는 정도로서, 가장 일반적인 상관계수.

r=−1	−1<r<0	r=0	0<r<1	r=+1
음의 상관관계가 강하다.	음의 상관관계가 있기는 하다.	상관관계가 없다.	양의 상관관계가 있기는 하다.	양의 상관관계가 강하다.

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X - E(X))(Y - E(Y))]}{\sigma_X \sigma_Y}$$

상관계수와 독립의 관계





스피어만 상관계수

스피어만 상관계수
: 데이터를 작은 것부터 차례로 순위를 매겨 순위를 사용해 상관계수를 구한다.

$$\rho = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2 \sum_i (y_i - \overline{y})^2}}$$

x_i는 X_i의 순위, y_i는 Y_i의 순위, $\overline{x}$와 $\overline{y}$는 각각 x_i와 y_i의 평균

-> 두 변수의 순위가 완전히 일치하면 +1, 두 변수의 순위가 완전히 반대이면 -1이다.

피어슨 상관계수	스피어만 상관계수
두 데이터의 값을 이용해 구한다.	두 데이터의 값의 순위를 이용해 구한다.
선형관계	선형관계, 비선형관계
연속형 데이터	연속형, 이산형, 순서형 데이터

켄달의 순위상관계수(켄달의 타우)

:스피어만 상관계수처럼 순위를 사용하여 상관계수를 이용.

(X, Y) 순서쌍으로 데이터가 있을 때,

$x_i < x_j,\ y_i < y_j$ $\Longrightarrow$ 부합(concordant)

$x_i < x_j,\ y_i > y_j$ $\Longrightarrow$ 비부합(discordant)

$$\tau = \frac{(number\ of\ concordant\ pairs - number\ of\ discordant\ pairs)}{\frac{1}{2}n(n-1)}$$

상관계수 검정

검정을 왜 할까?

: 앞의 독립성 검정처럼 두 변수 사이에 상관 관계가 있는지 알아보고자 실시.

$H_0: \rho = 0$ ⟹ 상관관계가 없다.

$H_1: \rho \neq 0$ ⟹ 상관관계가 있다.

단, 각 상관계수 종류에 따라 통계량, p-value값이 다르다.

통계량에 대한 p-value < 유의수준α
: 귀무가설 H0를 기각한다.
⇔ 두 변수 사이에 상관관계가 없다.

통계량에 대한 p-value < 유의수준α
: 귀무가설 H0를 기각한다.
⇔ 두 변수 사이에 상관관계가 있다.

일표본과 이표본에 따른 추정과 검정

→ 추정 : 표본을 이용하여 모집단의 성질을 추측하는 것.
- 점 추정
- 구간 추정

→ 검정 : 표본을 이용하여 모집단의 성질에 대한 가설을 세우고 검토하는 것.

확률변수의 개수가 1개면 일표본, 2개면 이표본으로 분류한다.

	평균	분산	비율
일표본	일표본 평균	일표본 분산	일표본 비율
이표본	독립 이표본 평균	이표본 분산	이표본 비율
	짝지은 이표본 평균		

일표본 평균에 대한 추정 및 검정

→ 일표본 평균에 대한 추정
⇒ 하나의 확률변수에 대한 표본을 이용하여 모집단의 평균이 어떻게 되는지 추정

100(1-α)% 신뢰수준 하에서 모평균의 구간을 추정

모집단의 표준편차를 알고 있을 때	$\overline{X} - Z_{\alpha/2}^* \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + Z_{\alpha/2}^* \frac{\sigma}{\sqrt{n}}$
모집단의 표준편차를 모를 때(대표본)	$\overline{X} - Z_{\alpha/2}^* \frac{s}{\sqrt{n}} \leq \mu \leq \overline{X} + Z_{\alpha/2}^* \frac{s}{\sqrt{n}}$
모집단의 표준편차를 모를 때(소표본)	$\overline{X} - t_{\alpha/2}^* \frac{s}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{\alpha/2}^* \frac{s}{\sqrt{n}}$

→ 일표본 평균에 대한 검정
H0 : 모평균이 0이다.
H1 : 모평균이 0이 아니다.

추정된 구간에 0이 속하면 => 귀무가설을 기각하지 못한다.
추정된 구간에 0이 속하지 않으면 => 귀무가설을 기각한다.

독립 및 짝지은 이표본 평균에 대한 추정과 검정

독립 이표본 : 독립인 두 변수에 대한 표본
짝지은 이표본 : 독립이 아닌 두 변수에 대한 표본

-짝지은 이표본은 D=X-Y로 정의함으로써 일표본과 같이 변수 D에 대해 추정,검정한다.

-독립 이표본은 분산이 같을 때, 다를 때에 따라 통계량이 다르다.

=> 따라서 이표본 분산 검정을 먼저 실시해야함.

H0 : 두 집단의 평균은 같다.
H1 : 두 집단의 평균은 같지 않다.

⇨ 추정된 구간에 0이 속하면 => 귀무가설을 기각하지 못한다.
추정된 구간에 0이 속하지 않으면 => 귀무가설을 기각한다.

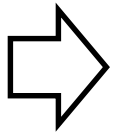
이표본 분산에 대한 추정과 검정

두 변수의 모분산의 비율을 추정 및 검정하는 과정.

$$\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim F(m-1,\ n-1)$$

분산비를 이용하므로 분산비가 1인지 아닌지를 검정한다.

H0 : 분산의 비가 1이다. => 두 집단의 분산은 같다.
H1 : 분산의 비가 1이 아니다. => 두 집단의 분산은 같지 않다.



추정된 구간에 1이 속하면 => 귀무가설을 기각하지 못한다.
추정된 구간에 1이 속하지 않으면 => 귀무가설을 기각한다.

일표본 비율과 이표본 비율에 대한 추정과 검정

일표본 평균 & 이표본 평균과 마찬가지로,

일표본 비율은
한 개의 확률변수가 이항분포를 따를 때, 사건이 발생할 확률 p를 추정 및 검정한다.

이표본 비율은
두 개의 확률변수가 이항분포를 따를 때, 각 사건이 발생할 확률 p1-p2를 추정 및 검정한다.

퀴즈

문제

1. A라는 회사는 스마트폰의 한 부품을 만드는 회사로, 이 A사의 불량률은 5%로 알려져 있다. 이 회사의 제품 20개를 조사했을 때, 불량이 2개 이하로 나올 확률을 구하시오.

2. 어느 회사에서 생산하는 제품의 평균수명을 조사하는데, 이 제품의 모표준편차(σ)는 40일이라고 한다. 이때 표본 100개를 뽑아 제품의 수명을 측정하였더니, 평균이 800일이 나왔다고 한다. 이때 제품의 평균수명에 대한 90% 신뢰구간, 95% 신뢰구간, 99% 신뢰구간을 추정하시오. (각 $Z_{\frac{\alpha}{2}}$ 는 1.64, 1.96, 2.58)

감사합니다 :D