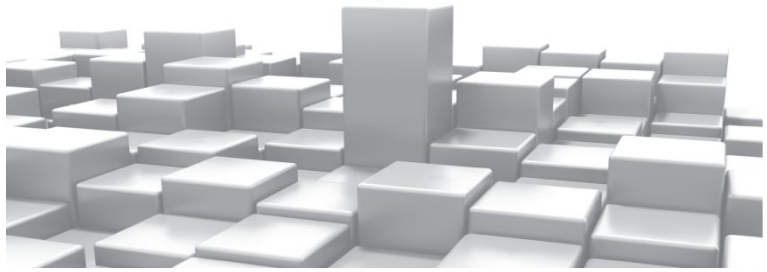
통계분석의 기본!
기초통계학 원리

확률변수와 확률분포



1. 학습목표

학습목표

- 변수와 확률변수의 정의를 이해한다.
- 확률분포의 의미와 통계에서 확률분포의 중요성을 이해한다.
- 확률분포의 두 종류인 이산확률분포와 연속확률분포를 이해한다.
- 이산확률분포와 연속확률분포의 대표적인 분포형태를 이해한다.

2. 본 강의

확률변수와 확률분포

01 변수와 확률변수

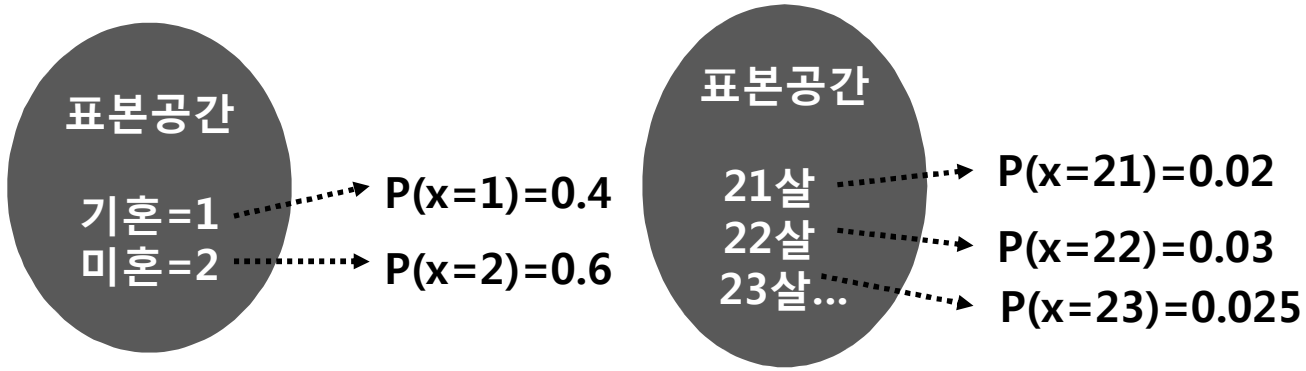
변수

- 숫자 또는 문자로 표현할 수 있는 자료의 특성

	측정항목	변수	확률변수
질적 자료	1) 당신은 기혼자입니까?	결혼상태	기혼=1, 미혼=2
	2) 당신은 직업이 있습니까?	취업상태	취업=1, 실업=2, 경제활동불참=3
양적 자료	3) 당신의 가족 수는?	가족 수	3명, 4명, 5명.....
	4) 당신은 몇 살입니까?	연령	21살, 28살, 32살, 37살.....
	5) 당신 가구의 소득?	소득	321만원, 358만원, 402만원.....

확률변수와 확률분포

01 변수와 확률변수



확률변수

- 변수가 취하는 값에 확률이 대응하고 있을 때 이를 확률변수라 한다

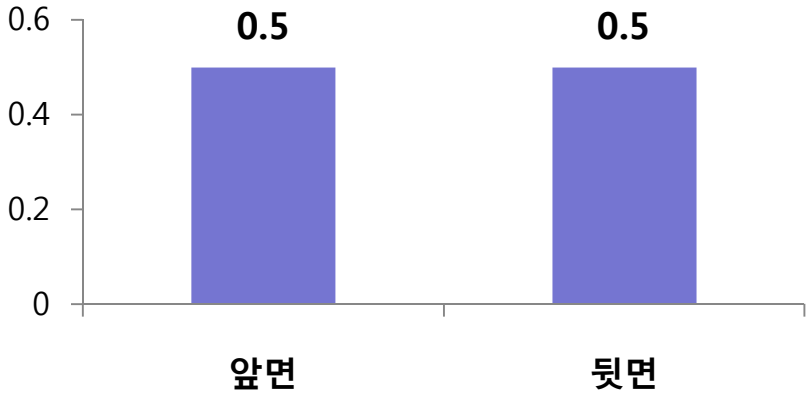
P(기혼)=0.4, P(미혼)=0.6

P(21살)=0.02, P(22살)=0.03, P(23살)=0.025 …….

02 확률분포의 중요성

대전제1: 확률분포는 판단의 기준이다!

생각1) 동전 던지기를 할 때, 앞면과 뒷면이 나올 확률은?

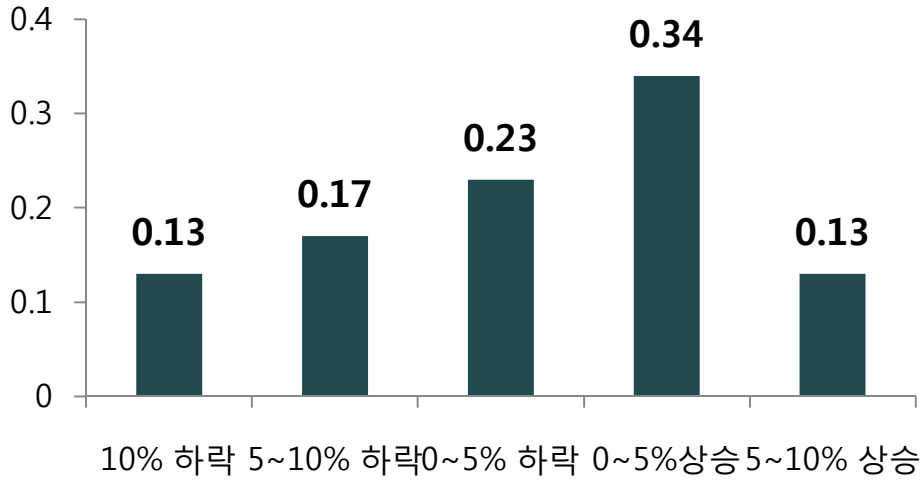


각각 0.5(50%)의 확률을 갖는다.
너무나 당연한 얘기?
하지만 이런 확률이 나올 분포를 이미 알고 있기 때문에 판단을 할 수 있다.

02 확률분포의 중요성

대전제1: 확률분포는 판단의 기준이다!

생각2) 대선 전에 서울 강남에 아파트를 사면 값이 오를까 내릴까?



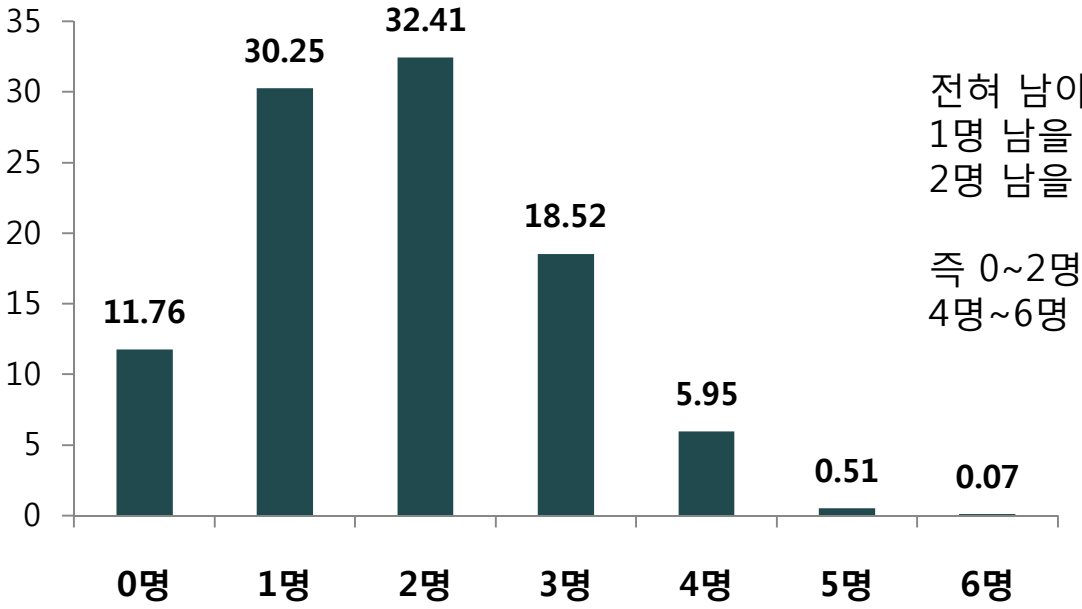
과거 자료를 보면 대선 전후 5개월 아파트값이 상승할 확률이 0.47(47%)였다.
반면, 하락한 확률은 0.53(53%)였다.
어떤 사건이 일어날 혹은 일어난 경험적 확률 분포(정보)가 없다면 객관적 판단은 내리기 어렵다

확률변수와 확률분포

02 확률분포의 중요성

대전제2: 기준이 없으면 판단할 수 없다!

생각1) A회사의 영업부에서 신규 직원을 6명을 채용하려고 한다. 통상 5년 동안 이직하지 않고 근무하는 확률이 30%(0.3)이라고 할 때, 5년 후에 몇 명의 직원을 채용해야 영업부에서 6명의 직원을 유지할 수 있는가?



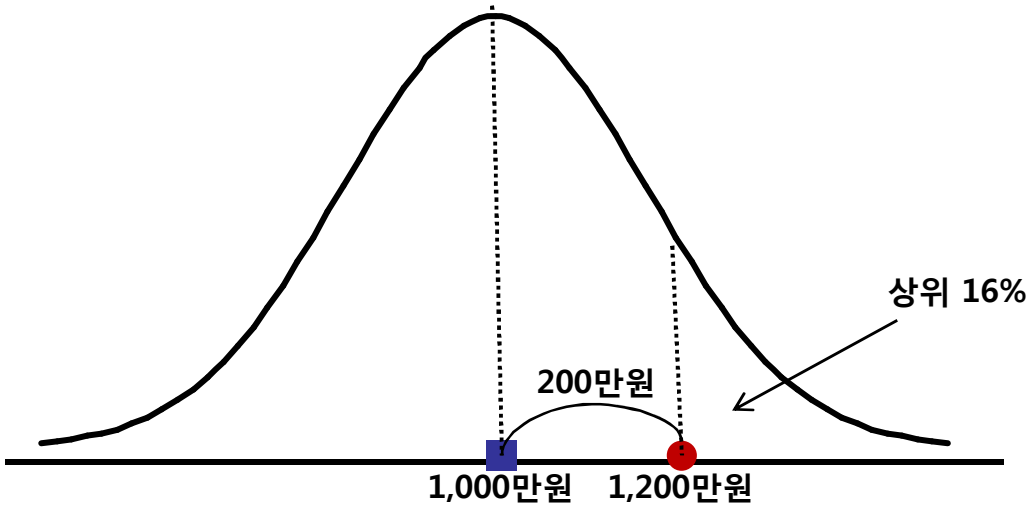
전혀 남아 있지 않을 확률 11.8%
1명 남을 확률 30.3%
2명 남을 확률 32.4%

즉 0~2명 남을 확률이 94% 가량이므로,
4명~6명 가량을 신규 채용해야 할 것으로 판단됨

02 확률분포의 중요성

대전제2: 기준이 없으면 판단할 수 없다!

생각2) A회사 영업직원의 평균 매출성과는 1,000만원인데 홍길동은 1200만원의 매출을 올렸다. 뛰어나게 잘 한 것인가?
(단 영업직원 전체 매출의 표준편차는 200만원이다)



02 확률분포의 중요성

정리: 왜 확률분포가 중요한가?

- 확률변수가 일어날 확률을 전체 1.0(100%)인 분포로 표현하여 관측된 통계량이 일어날 확률을 계산할 수 있게 한다.

- 확률변수의 특성(이산-연속) 및 분석특성(일표본, 차이검정 등)에 따라 이론적으로 성립된 확률분포를 기준으로 모집단의 추론 및 가설검정이 가능하다

- 각 확률분포는 변수와 분석의 특성에 맞는 최적의 이론적 모형을 의미한다.
- 확률분포는 통계량을 파악하여 통계적 의사결정을 내리는 기준을 제시한다.

03 확률분포의 종류

이산확률분포

- 이산적 확률분포(확률변수가 0, 1, 2같은 정수 값을 가지는 경우)가 이루는 확률분포
  예) 주사위 던지기, 동전 던지기, 찬성반대 투표 등
  대표적 이산확률분포: 이항분포, 포아송분포, 초기하분포, 기하분포 등

연속확률분포

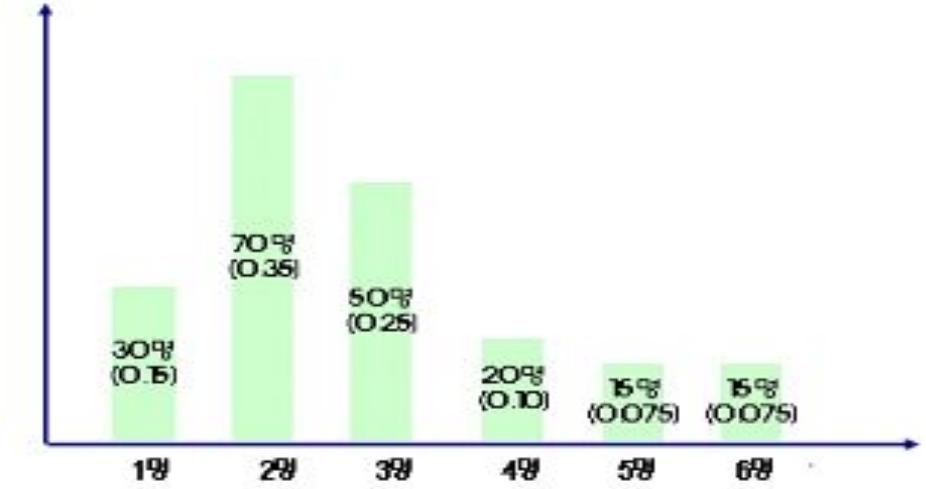
- 연속적 확률변수(확률변수가 소수점의 값을 포함하는 실수의 값을 가지는 경우)가 이루는 확률분포
  예) 신장, 체중, 소득 등
  대표적 연속확률분포: 정규분포, 표준정규분포, t분포, F분포, X^2분포 등

03 확률분포의 종류

이산형(discrete) 확률변수와 확률분포

- 확률변수가 취할 수 있는 값이 유한하거나 셀 수 있는 변수

예) 책의 페이지 당 오타수, 제품 한 노트에서 나온 불량품의 개수,
고속도로 상에서 일정시간 내에 발생한 사고건수,
농구경기에서 득점수,
동전을 계속하여 던질 때 최초에 앞이 나올 때까지의 횟수



- 평균: 1명X0.15+2명X0.35+3명X0.25+4명X0.1+5명X0.075+6명X0.075=2.74명
- 분산: (2.74-1)2X0.15+(2.74-2)2X0.35+(2.74-3)2X0.25+(2.74-4)2X0.10+(2.74-5)2X0.075+(2.74-6)2X0.075

03 확률분포의 종류

이산형(discrete) 확률변수와 확률분포

- $0 \le f(x_i) \le 1$: 0과 1의 값을 갖는다.
- $P(a \le X \le B) = \sum_{a \le X \le b} f(x_i)$: 확률분포로 표현된다.
- $\sum f(x_i) = 1$: 모든 값의 합은 1이 된다.

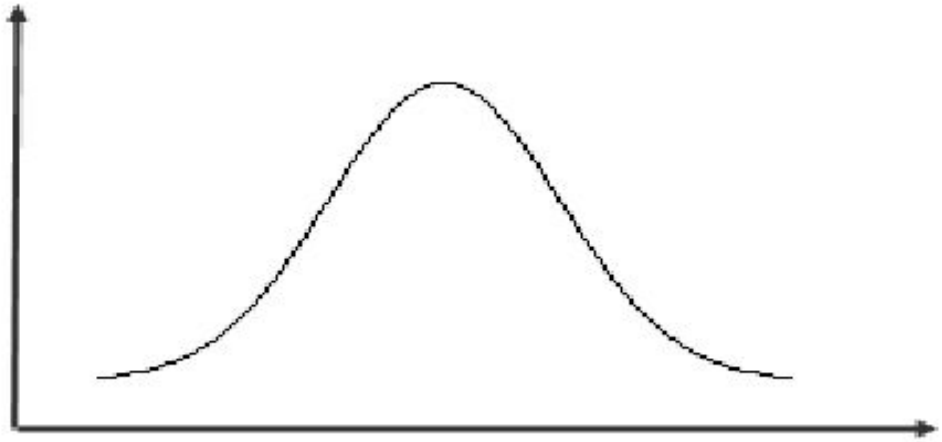
- $\mu = \sum x_i f(x_i)$: 평균은 해당 확률변수가 일어날 확률과 각 측정값의 곱으로 나타낸다.
- $\sigma^2 = \sum (x_i - \mu)^2 f(x_i)$: 분산은 각 측정값과 평균의 차이의 제곱과 해당 확률변수가 일어날 확률간의 곱으로 나타낸다.

03 확률분포의 종류

연속형(continous) 확률변수와 확률분포

- 가능한 값이 실수의 어느 한 구간 안에 포함되는 확률변수

예) 사람의 몸무게,
체온, 비행기의 도착시간

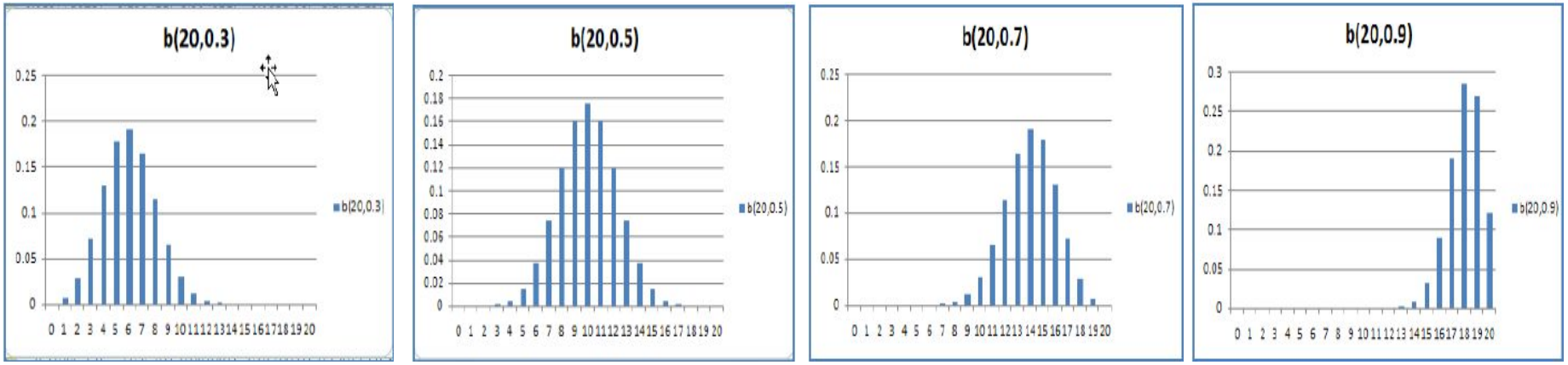


- $f_x(x) \geq 0$
- $P(a \leq X \leq b) = \int_a^b f_x(x)dx = 1$
- $\int_{-\infty}^{\infty} f_x(x)dx = 1$

04 이산확률분포의 종류

1. 이항분포

- 상호 배반적인 두 사건만 나타나는 경우, 발생할 확률의 기준 분포



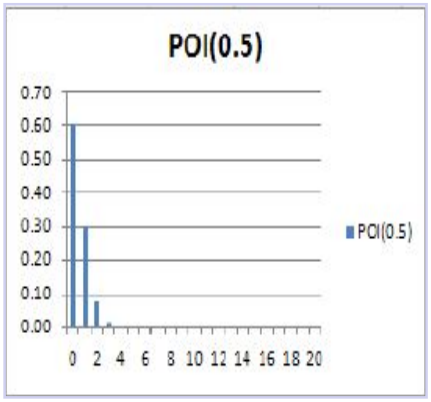
- A타자의 타율이 3할이다. 20번 타석에서 안타 칠 확률은?

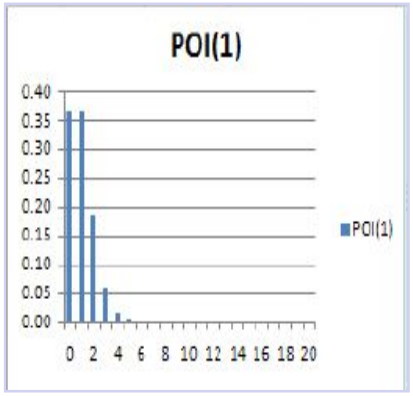
- 확률값(p)와 시행횟수(n)만 알면 이항확률분포를 그릴 수 있다.
- 시행횟수가 커지면 확률분포는 정규분포에 근사한다

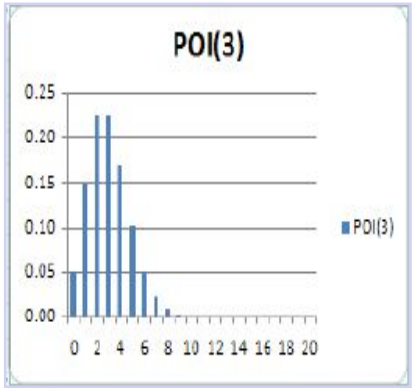
04 이산확률분포의 종류

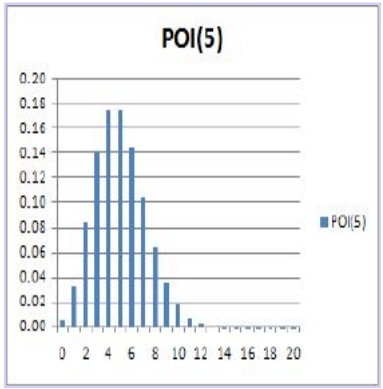
2. 포아송분포

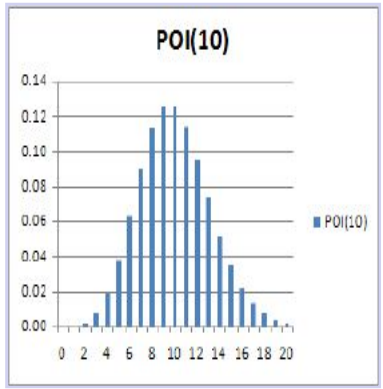
- 일정한 시간, 거리, 공간 상에서 매우 드물게 발생하는 확률을 계산할 때 기준이 되는 분포









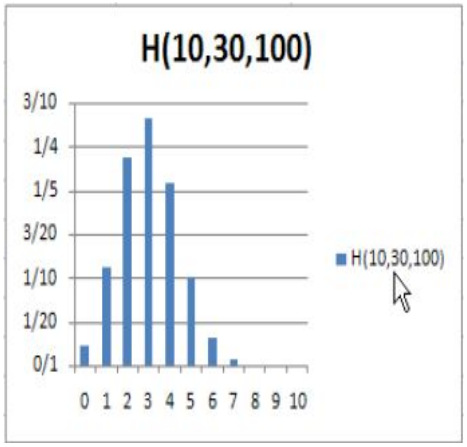


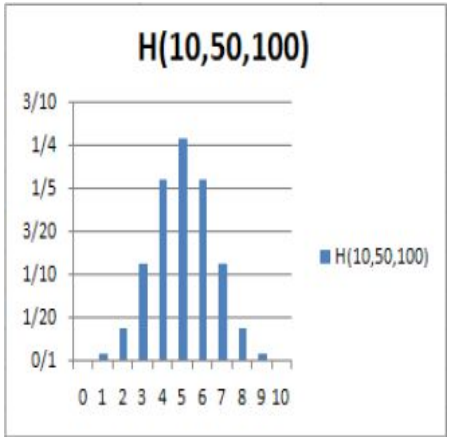
- 어느 하루 동안 공장에서 생산된 제품의 불량품 개수
- 어느 지역에서 1년 동안 화재가 발생할 횟수
- 어느 하루 동안 잘못 걸려온 전화 횟수

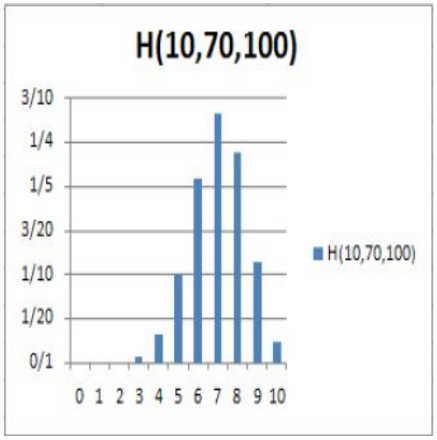
04 이산확률분포의 종류

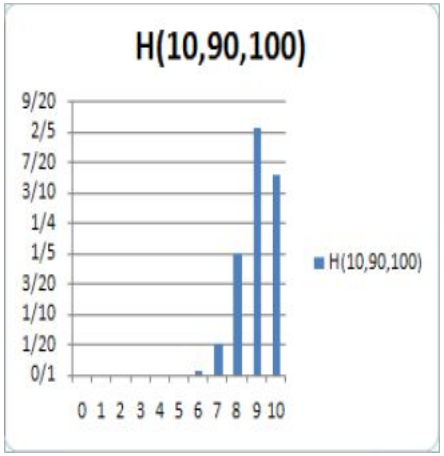
3. 초기하분포

- 시행마다 발생할 결과가 이항분포처럼 두 가지만 있으나 유한모집단에서 비복원추출되기 때문에 베르누이 시행 조건에 만족되지 않는 경우 적용되는 확률분포







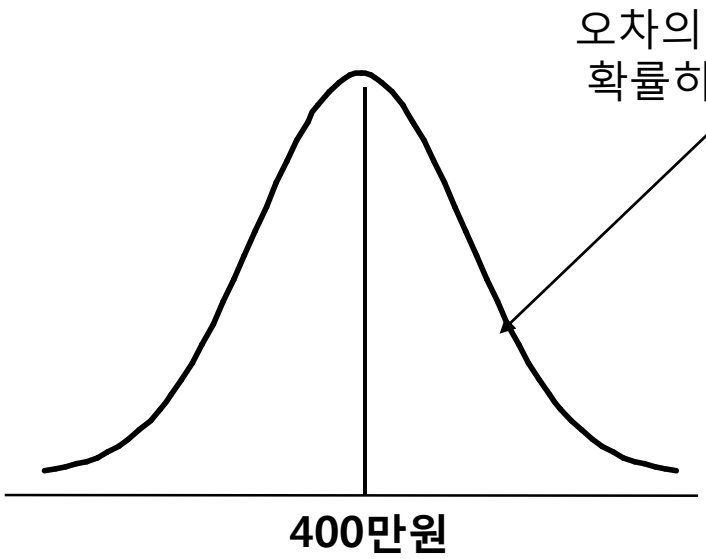


- 100개의 모집단에서, 남자가 30명이다. 10명을 무작위로 뽑을 때, 남자일 확률은? Poison(10,30,100)

05 연속확률분포의 종류

1. 정규분포

- 키, 몸무게, 수명 등 각종 일반적인 연속형 수치에 적용되는 확률분포



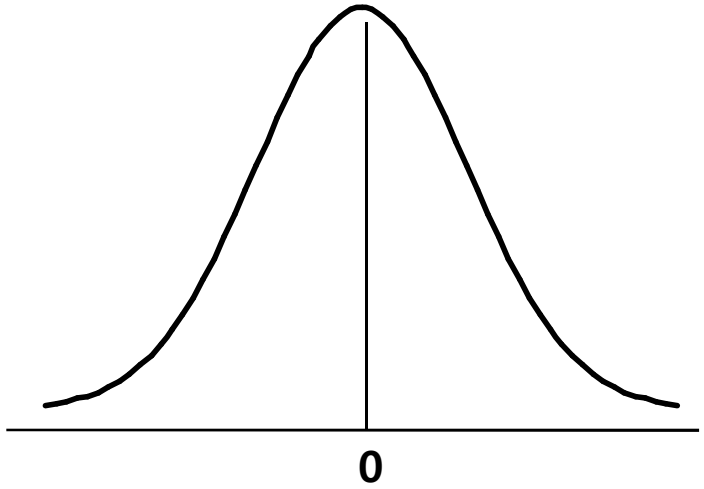
- 모집단 평균과 표본 평균의 차이 검증

- 중학생 IQ 평균 120인데 영희는 123이다.
  높은건가? 비슷한건가?

05 연속확률분포의 종류

2. 표준정규분포

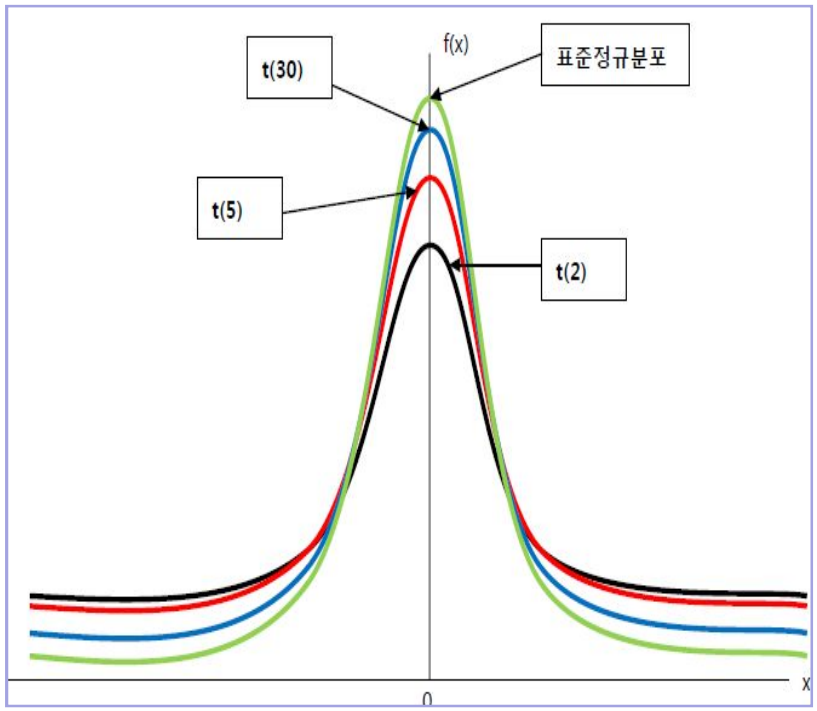
- 정규분포를 평균 0, 표준편차 1로 단위를 표준화한 분포



05 연속확률분포의 종류

3. t-분포

- 정규분포와 유사하나 표본의 크기가 작은 경우(n<30)에 기준이 되는 확률분포

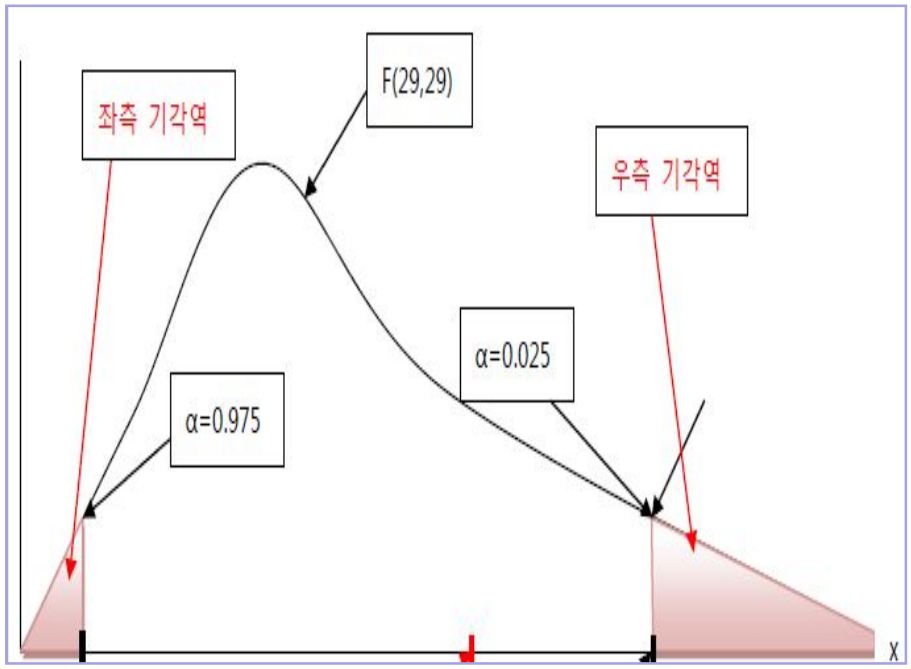


- 남여간에 IQ는 다른가? 비슷한가?
- 남자: 120, 여자 122

05 연속확률분포의 종류

F분포

- 분산에 대한 비 검정, 분산분석 및 회귀분석 등의 관계추론의 기준 확률분포

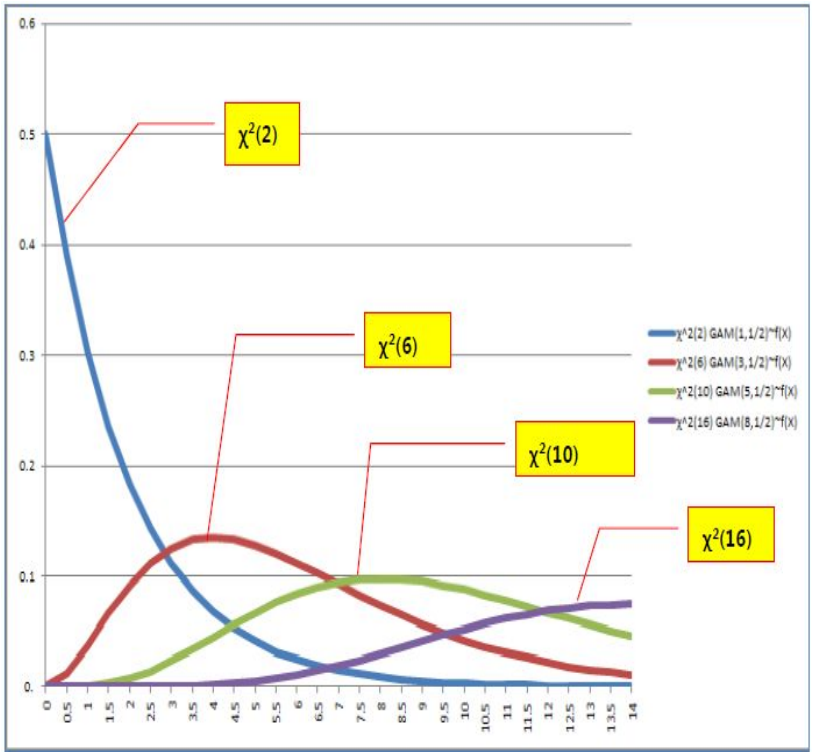


- 연령대별 IQ는 다른가, 같은가?
- 20대: 120, 30대: 123, 40대 124

05 연속확률분포의 종류

4. X^2분포

- 범주형 자료에 대한 적합도, 독립성 검정 등에 사용되는 기준 분포



- 남여간에 정당 선호도가 다른가?

	A당	B당	C당
남	45%	25%	30%
여	35%	40%	25%

3. 학습정리

확률변수와 확률분포

01 학습정리

1. 변수란 성별, 결혼상태 등의 응답이 숫자 또는 문자로 표현될 수 있는 자료를 의미한다.

2. 확률변수란 변수가 취하는 값에 확률이 대응하는 것을 의미한다. 예를 들어 성별은 변수이며, 남성 0.4(40%), 여성 0.6(60%) 등은 확률변수이다.

3. 확률변수의 각각의 결과를 전체 면적이 1(100%)인 분포의 형태로 표현한 것이 확률분포이며 확률분포는 통계적 검정에서 판단의 기준이므로 매우 중요한 개념이다.

4. 확률분포는 이산확률분포와 연속확률분포로 구분된다.

5. 이산확률분포는 나타날 수 있는 확률변수가 0, 1, 2와 같이 이산적인 형태를 이루는 분포이고, 대표적으로 이항분포, 포아송분포, 초기하분포 등이 있다.

6. 연속확률분포는 나타날 수 있는 확률변수가 연속적인 분포의 형태이며, 정규분포, 표준정규분포, t분포, F분포, 카이제곱분포 등이 이에 해당된다.

7. 이산확률분포의 대표적 분포형태인 이항분포는 상호 배반적인 두 사건만 나타나는 경우, 발생할 확률의 기준이 되는 분포이다.

02 학습정리

8. 이산확률분포 중 포아송분포는 일정한 시간, 거리, 공간 상에서 매우 드물게 발생하는 확률을 계산할 때 기준이 되는 분포이다.

9. 이산확률분포 중 초기하분포는 시행마다 발생할 결과가 이항분포처럼 두 가지만 있으나 유한모집단에서 비복원추출되기 때문에 베르누이 시행 조건에 만족되지 않는 경우 적용되는 확률분포이다.

10. 정규분포는 대표적인 연속확률분포이며, 키, 몸무게, 수명 등 각종 일반적인 연속형 수치에 적용되는 확률분포이며, 표준정규분포는 정규분포를 평균이 0, 표준편차가 1로 표준화한 정규분포이다.

11. 연속확률분포 중 t분포는 정규분포 와 유사하나 표본의 크기가 작은 경우($n<30$)에 기준이 되는 확률분포이다.

12. 연속확률분포 중 F분포는 분산에 대한 비 검정, 분산분석 및 회귀분석 등의 관계추론의 기준 확률분포이다.

13. 연속확률분포 중 카이제곱분포는 범주형 자료에 대한 적합도, 독립성 검정 등에 사용되는 기준 분포이다.