사회통계연습
이산확률분포와 연속확률분포

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

October 1, 2021

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

진행 순서

1. 이론적 확률분포
2. 대표적인 이산확률분포: 이항분포
3. 대표적인 연속확률분포: 정규분포

이론적 확률분포

이론적 확률분포

다양한 유형의 시행(trials) 또는 실험은 서로 다른 확률분포를 생성한다.

- 앞서 척도는 먼저 숫자형(numerical)과 범주형(categorical)로 나눌 수 있고, 숫자형은 다시 이산형(discrete)과 연속형(continuous)으로 나뉜다고 했다.
- 마찬가지의 논리로 확률분포 역시 이산확률분포(discrete probability distribution)와 연속확률분포(continuous probability distribution)로 나눌 수 있다.

일단 확률분포가 주어지면 그것의 평균과 분산을 계산할 수 있다는 것을 꼭 기억하자!

- 수리통계학에서는 꽤 공들여 여러 종류의 이론적 확률분포를 공부한 다음, 각각의 평균과 분산을 도출하고 각종 수학적 테크닉을 동원해서 다양하게 증명한다.

우리는 오로지 두 가지 이론적 확률분포만 공부한다.

- 대표적인 이산확률분포로 이항분포(binomial distribution)를, 대표적인 연속확률분포로 정규분포(normal distribution)를 공부한다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

대표적인 이산확률분포: 이항분포

대표적인 이산확률분포: 이항분포

가장 대표적인 이산확률분포(discrete probability distribution)는 이항분포(binomial distribution)이다.

이항분포를 위한 준비운동으로 먼저 베르누이 과정(Bernoulli process)을 이해할 필요가 있다.

- 딱 한 번의 독립적인 시행(trial)이 이루지고 각 시행의 사건(event)은 두 가지 뿐이다 (e.g., 성공/실패, H/T, 1/0).
- 두 사건의 확률은 각각 $P(X = 1) = p$ 와 $P(X = 0) = 1 - p$ 로 표현된다.

$$f_{k;p} = \begin{cases} p & \text{if k=1} \\ 1 - p & \text{if k=0} \end{cases}$$

- 좀 더 축약해서 표현하면,

$$f_{k;p} = p^k(1 - p)^{1-k}$$

- 이때 p는 성공할 확률(e.g., 동전 던지기의 경우 0.5), k는 성공 여부(i.e., 1이 성공, 0이 실패)다.

충북대학교 CHUNGBUK NATIONAL UNIVERSITY

대표적인 이산확률분포: 이항분포

근데 아까 은근슬쩍 좌변(left-hand side)에 $f_{k;p}$ 라는 식을 썼다.

$$f_{k;p} = p^k(1-p)^{1-k}$$

- 이 좌변은 위 식이 확률질량함수(probability mass function; p.m.f)임을 나타낸다.
- 확률질량함수란 이산확률변수에서 특정 값을 받으면 그에 따른 확률을 토해내는 함수이다.
- 참고로 k나 p같이 확률질량함수를 정의하기에 꼭 필요한 값들을 모수(parameter)라고 부른다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

대표적인 이산확률분포: 이항분포

베르누이 분포(Bernoulli Distribution)

- 너무 단순하기 때문에 이를 새삼 확률분포로 나타내는 것도 웃기지만, 우리가 더 관심있는 이항분포(binomial distribution)을 공부하기 위해서 필요하다.
- 아까 일단 확률분포가 주어지면 확률변수의 평균과 분포를 계산할 수 있다고 말했다.
- 베르누이 확률분포의 기댓값(=평균)은 $E(X) = p \cdot 1 + (1-p) \cdot 0 = p$ 이다.
- 베르누이 확률분포의 분산은 $Var(X) = E(X^2) - E(X)^2 = p(1-p)$ 이다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

대표적인 이산확률분포: 이항분포

이항분포(binomial distribution)의 개념

- 앞서 설명한 베르누이 과정은 단 한 번 시행을 염두에 둔 것이였다.
- 이항분포는 베르누이 과정을 n번 독립 시행해서 성취한 성공 횟수를 확률변수(random variable)로 삼는다.
- 그리고나서 그 확률변수의 확률분포를 만들면 이것이 이항분포다.

이항분포의 직관적 아이디어

- 동전을 세 번 던져서 앞면이 한 번 나올 확률이 어떻게 될까?
- 사건은 H, T, T 또는 T, H, T 또는 T, T, H이여야 한다.
- (각 사건과 결부된) 확률은 각각 $p \cdot (1-p) \cdot (1-p)$ 과 $(1-p) \cdot p \cdot (1-p)$ 과 $(1-p) \cdot (1-p) \cdot p$이다.
- 그러면 답은 .5 · (1-.5) · (1-.5) + (1-.5) · .5 · (1-.5) + (1-.5) · (1-.5) · .5 = .375

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

대표적인 이산확률분포: 이항분포

이항분포의 좀 더 세련된 표현

- 조합(combination)을 사용하면 수학적으로 좀 더 세련된 방식으로 나타낼 수 있다.

$$\binom{N}{k} \cdot p^k (1-p)^{N-k} = \frac{n!}{k!(n-k)!} \cdot p^k (1-p)^{N-k}$$

- 앞서 세 번 시행하였으므로(N=3),

$$\frac{3!}{1!(3-1)!} \cdot .5^1 \cdot (1-.5)^{3-1} = .375$$

- 물론 어느 쪽으로 계산하건 답은 같다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

대표적인 이산확률분포: 이항분포

계속해서 말하지만, 일단 확률분포가 주어지면 그것의 평균과 분포를 계산할 수 있다.

이항분포의 기댓값은 $E(X) = np$이다.

- 좀 더 까다로운 수학적 증명은 내버려두고 개념적으로만 살펴보자!
- 베르누이 분포(Bernoulli distribution)는 단 한 번의 성공 횟수를 확률과 함께 보여준다. 그리고 그 기댓값(expected value)는 p다.
- 반면 이항분포(binomial distribution)는 N번의 베르누이 과정(Bernoulli process)을 독립 시행했을 때 성공 횟수를 확률과 함께 보여준다.
- 그러므로 이항분포의 기댓값은 베르누이 분포의 기댓값인 p을 n번을 곱한 것과 같다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

대표적인 이산확률분포: 이항분포

이항분포의 분산은 $\mathrm{Var}(X) = np(1-p)$이다.

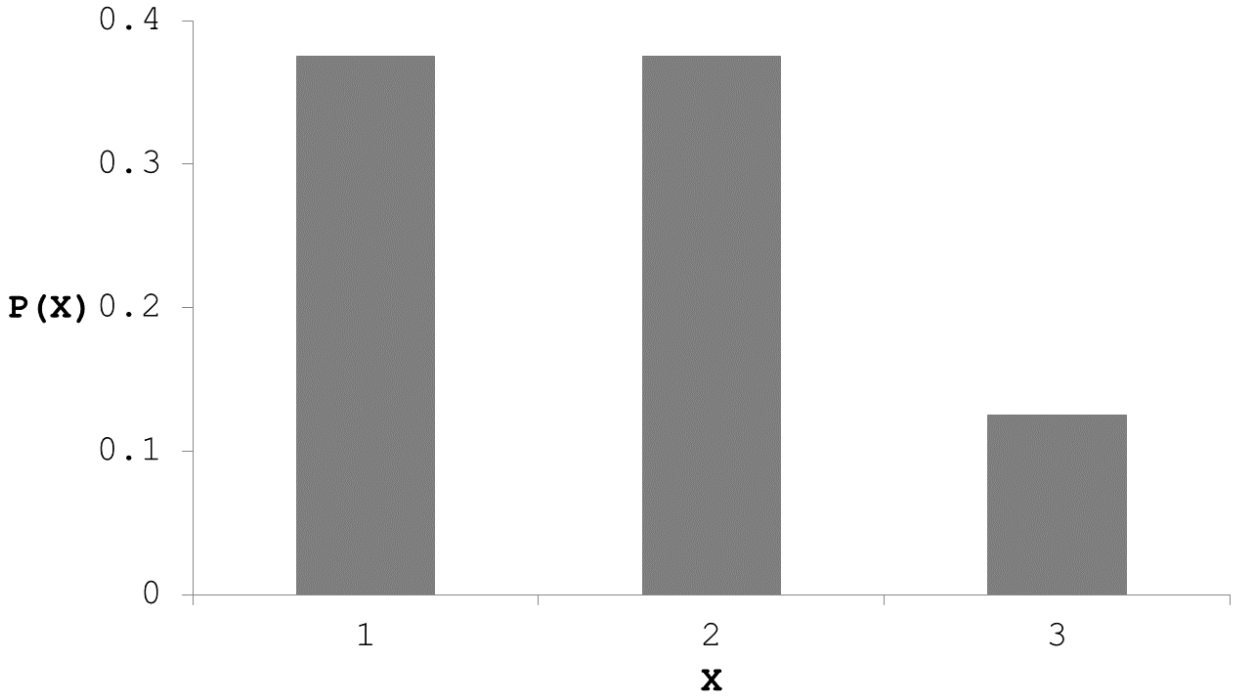
- 까다로운 수학적 증명은 수학도들에게 맡기고 개념적으로만 살펴보자.
- 베르누이 분포(Bernoulli distribution)의 분산은 p(1-p)다(8페이지 참고).
- 이항분포(binomial distribution)는 N번의 베르누이 과정(Bernoulli process)을 독립 시행했을 때 성공 횟수를 확률과 함께 보여준다.
- 그러므로 각 베르누이 시행(each Bernoulli trials)의 분산(=p(1-p)은 서로 얽히지 않고 상호독립되어 단지 n번만 곱한 것과 같다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

대표적인 이산확률분포: 이항분포

이항확률밀도함수(binomial probability density function)의 그래프

- 만일 x축을 확률변수 X로 하고 y축을 이항분포를 따르는 확률밀도함수라고 한다면 그림을 그려볼 수 있다.
- 동전을 세 번 던져 앞면이 나오는 경우의 이항확률밀도함수는,
- 앞면이 한 번 나오는 확률(=P(X=1))은 $\frac{3!}{1!(3-1)!} \cdot .5^1 \cdot (1-.5)^{3-1} = .375$
- 앞면이 두 번 나오는 확률(=P(X=2))은 $\frac{3!}{2!(3-2)!} \cdot .5^1 \cdot (1-.5)^{3-2} = .375$
- 앞면이 세 번 나오는 확률(=P(X=3))은 $\frac{3!}{3!(3-3)!} \cdot .5^1 \cdot (1-.5)^{3-3} = .125$



충북대학교 CHUNGBUK NATIONAL UNIVERSITY

대표적인 이산확률분포: 이항분포

아까 그림은 엑셀로 그린 거다. 여러분도 쉽게 그릴 수 있다.

- eCampus에서 binomdist.xlsx을 다운받아서 [fig4] 탭을 보자. 이게 아까 그 그림이다. 맘대로 숫자를 바꾸어 보자.
- 이번엔 [PLAY!] 탭을 보자. 동전을 30번 던져 1번에서 15번 사이의 앞면(H)이 나올 확률을 계산하고 있다.
- binom.dist() 함수로 직접 그려볼 수도 있고, 패러미터를 조금씩 달리해서 어떻게 달라지는지도 확인할 수 있다.
- 심지어 JASP에서는 더 쉽게 가능하다.

예제: "동전을 열 번 던져서 앞면(H)이 세 번 이하로 나올 확률은?"

- 다시 [PLAY!] 탭의 엑셀 함수를 사용하여 직접 계산해보자.
- 참고로 답은 0.009765625 + 0.043945313 +0.1171875 다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

대표적인 연속확률분포: 정규분포

대표적인 연속확률분포: 정규분포

척도와 확률분포

- 지금까지 이산형 척도(discrete scale)로 측정되는 사건(e.g., 여성별로 출산한 아이의 수, 청주시에서 일어난 시위의 수)에 대한 이산확률분포를 다루었다.
- 이제부터 연속형 척도(continuous scale)로 측정되는 사건(e.g, 지역별 평균출산율, 청주시 평균 상가임대료)에 대한 연속확률분포를 다루기로 한다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

대표적인 연속확률분포: 정규분포

정규분포(normal distribution)는 지금껏 다룬 확률분포 중에 가장 중요한 것이다.

- 가우스(Gauss)가 발견하여 가우스 분포(Gaussian distribution)라고도 불리운다.
- 모든 이론적 확률분포 가운데 가장 수학적으로 아름답다(elegant).
- 있는 그대로도 자연과 사회에서 나타나는 여러 현상들을 터무니없이 잘 설명한다.
- "이건 잘 설명되지 못하네" 싶은 현상도 각도를 달리해서 보면 바꾸면 정규분포에 의해 잘 설명된다.
- 통계적 추론(statistical inference)의 초석 역할을 한다.

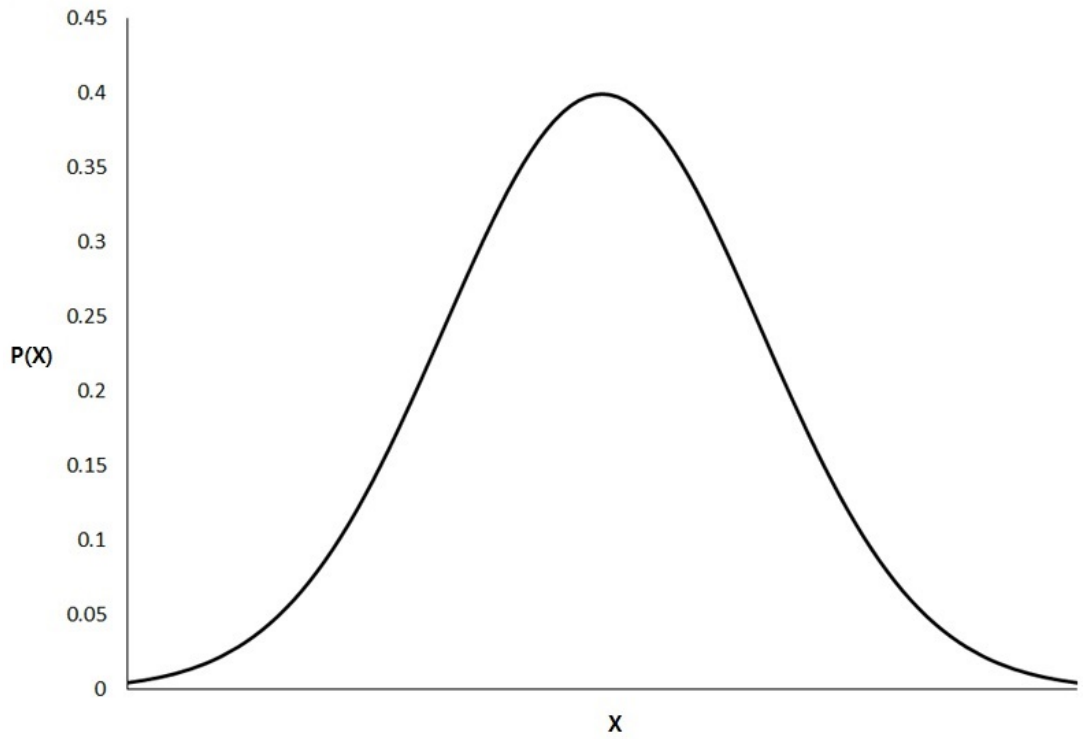
독일화폐 10마르크(지금은 유로화로 대체)에는 가우스와 정규분포가 그려져 있다

충북대학교 CHUNGBUK NATIONAL UNIVERSITY

대표적인 연속확률분포: 정규분포

다음과 같은 정규분포(normal distribution)의 기본적인 속성이 잘 알려져 있다.

- 정규분포는 대칭적(symmetric)이고 종 모양(bell-shaped)이다.
- 양측의 꼬리는 무한히 이어지며 결코 x축에 닿지 않는다.
- 평균(mean)과 중위값(median)이 같다.
- x축은 확률변수(random variable)인 X이고, y축은 확률밀도함수(probability density function)이다.



충북대학교
CHUNGBUK NATIONAL UNIVERSITY

대표적인 연속확률분포: 정규분포

정규분포의 확률밀도함수(probability density function)는 아래와 같다.

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

- 첫째, 외우지 마라.
- 둘째, 원주율(π)이 나온다. 신기하네~
- 셋째, 모수(parameter)로 무엇이 나오는지 살펴보자(μ 와 σ).
- 넷째, 위 식이 확률밀도함수(probability density function; p.d.f)라고 불리우는데, 확률밀도함수란 연속확률변수에서 특정 값을 받으면 그에 따른 확률을 토해내는 함수이다.
- "잠깐! 아까 확률질량함수(p.m.f)와는 뭐가 다르지?" 같은 개념인데 이산확률변수에 대해서는 확률질량함수, 연속확률변수에 대해서는 확률밀도함수라고 부른다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

대표적인 연속확률분포: 정규분포

정규확률밀도함수(normal probability density function)는 엑셀로 장난치면서 배워야 더 금방 이해할 수 있다.

- norm.dist(x, mean, standard_dev, cumulative) 함수로 레코드를 만들고 그래프를 만든다.
- 여기서 cumulative는 TRUE 또는 FALSE 중 고르는 건데 만일 누적(cumulative) 함수를 그릴거면 TRUE, 아니면 FALSE.
- eCampus에서 normdist.xlsx을 다운받아서 [PLAY! (1)] 탭을 보자. 모수(parameter)로 무엇이 있는지, 또 그것들을 어떻게 엑셀 함수에서 어떻게 쓰이고 있는지 눈으로 확인하자!
- P(X)에서 구해지는 숫자나 계산보다는 그래프의 모양 변화에 집중해보자.
- μ (mu; 평균)가 커질수록 그래프의 모양이 어떻게 변하는가?
- σ (sigma; 표준편차)가 커질수록 그래프의 모양이 어떻게 변하는가?

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

대표적인 연속확률분포: 정규분포

표준정규분포(standard normal distribution)는 표준화된 정규분포이다!

- 정규분포의 꼴은 모수(parameter)인 μ와 σ에 따라 다르다(엑셀로 실험해 보았다).
- 그러다보니 좀처럼 표준적인 정규분포를 말하기 어렵다는 단점이 있다.
- normdist.xlsx 파일의 [PLAY! (2)] 탭을 보자. 위의 확률분포 P(X)와 아래의 확률분포 P(X)가 어떻게 다른지 관찰하자.
- 분포가 완전히 똑같은데 단지 μ와 σ가 다르기 때문에 P(X)가 전혀 다르다는 것을 확인할 수 있다.
- 이렇게 되면 두 개의 정규분포를 비교하기가 까다로워진다. 그러므로 통계학자들 사이에서 $\mu = 0$, $\sigma = 1$로 하는 정규분포를 이제부터 표준(standard)으로 삼자고 규약을 정했다. 그게 바로 표준정규분포다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

대표적인 연속확률분포: 정규분포

편리하게도 통계학자들은 그냥 정규분포하는 자료를 표준정규분포로 변환하는 방법도 개발했다!

- 표준정규분포로 변환하기 위해서는 원점수(raw score)를 Z-점수(Z-score)로 변환하면 된다.
- 보다 구체적으로, 원점수 레코드(x_i)에서 평균(μ)을 뺀 값을 표준편차(σ)로 나누어 Z-점수를 구한다.

$$Z_i = \frac{x_i - \mu}{\sigma}$$

- "평균을 빼준다"는 점에서 개별 관측치는 평균에서 벗어난 정도가 클수록 높은 Z-점수를 갖게 된다.
- "표준편차로 나눈다"는 점에서 (설령 원점수가 제법 컸더라도) 표준편차가 클수록 Z-점수는 작아진다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

대표적인 연속확률분포: 정규분포

- 예제: C병원의 의무기록지에 따르면 이 병원에서 지난 5년 간 태어난 신생아의 몸무게 평균은 3.2kg였다. 이 자료를 통해서 표준편차(standard deviation)를 계산해보니 0.7kg이였다. 신생아의 출생시 몸무게는 정규분포한다.
- normdist.xlsx의 [fig7] 탭에서 표준정규분포(standard normal distribution)의 계산 과정을 관찰하자.
- 이 계산과정에서 마지막으로 표준정규분포(standard normal distribution)의 누적밀도함수(cumulative density function; C.D.F)의 계산 과정을 관찰하자.
- 여기서는 norm.dist() 함수를 쓸 때, TRUE를 사용했다(Why?)!

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

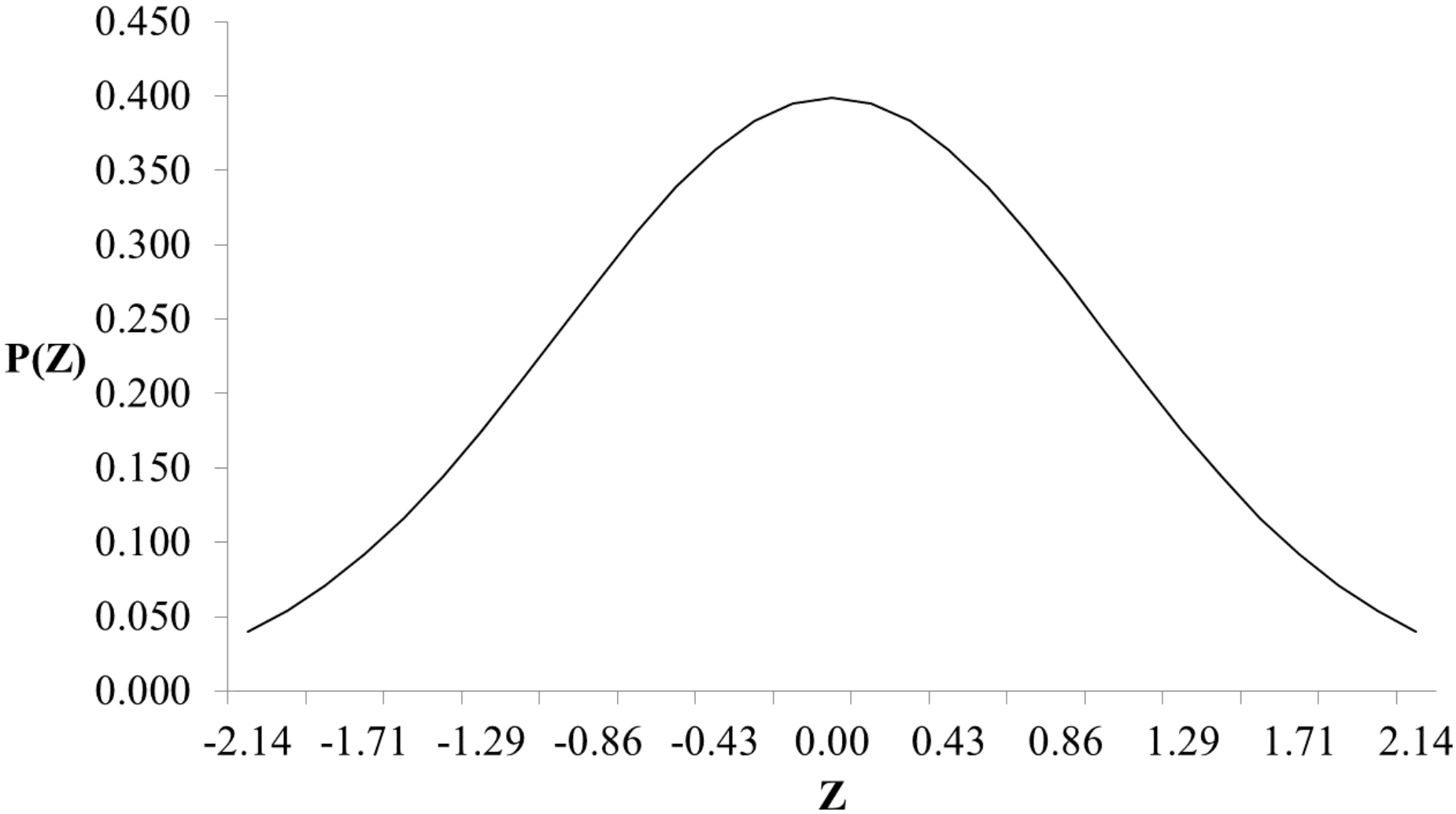
대표적인 연속확률분포: 정규분포

엑셀을 가지고 표준정규분포(standard normal distribution)의 누적밀도함수(C.D.F)를 쉽게 계산할 수 있다.

- 예제 1. 아까 말한 C병원에서 어젯밤 새벽에 한 아이가 태어났다. 이름은 동혁이다. 출생시 몸무게는 4.1kg이였다. C병원에서 태어난 아이 중 동혁보다 가벼운 아이는 전체의 몇 퍼센트인가?
- 예제 2. 지금 말하기 바쁘게 방금 전 또 한 아이가 태어났다. 이름은 채린이다. 출생시 몸무게는 2.8kg이였다. C병원에서 태어난 아이 중 채린보다 무겁고 동혁보다 가벼운 아이는 전체의 몇 퍼센트인가?
- 예제 3. C병원에서 제일 일찍 태어난 도균은 당시 몸무게가 2.3kg이였다. C병원에서 태어난 아이 중 도균보다 무거운 아이는 전체의 몇 퍼센트인가?

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

대표적인 연속확률분포: 정규분포



충북대학교
CHUNGBUK NATIONAL UNIVERSITY

대표적인 연속확률분포: 정규분포

이미 표준정규분포(standard normal distribution)하고 있는 자료를 어떻게 (원)정규분포로 변환하나?

- 간단하다! Z-점수(Z-score)를 원점수(raw score)로 변환하면 된다.
- 보다 구체적으로, Z-점수(Z)에서 표준편차(σ)를 곱한 값에 평균(μ)을 더하여 원점수를 구한다.

$$x_i = Z_i \cdot \sigma + \mu$$

- 눈치 챘겠지만, 이건 앞서 소개한 Z-점수의 공식을 다시 X에 대해 정리한 것에 불과하다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

대표적인 연속확률분포: 정규분포

확률에서 Z-값을 찾기 위한 엑셀 함수가 있다.

- NORM.INV(probability, mean, standard_dev)
- 이 함수는 누적정규분포(cumulative NORMal distribution)의 역(INVerse)을 계산하는데 쓰인다.
- 누적밀도함수(C.D.F)를 전제로 하므로 FALSE나 TRUE는 넣을 필요없다(무조건 TRUE다).
- 만일 $P(Z < z) = .8$ (즉, 상위 20%)를 찾고 싶다면, $=$ NORM.INV$(.8, 0, 1)$를 입력하자.
- NORM.DIST()에서는 Z를 넣고 확률을 얻었다면, NORM.INV()에서는 확률을 넣고 Z를 얻는다.
- 즉, NORM.DIST()와 NORM.INV()는 서로 역관계인 셈!

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

대표적인 연속확률분포: 정규분포

확률이 주어졌을 때 대응하는 원점수를 파악하는 연습이 조금 필요하다.

- 예제: 작년 정부 보고서에 따르면 공직자적성검사의 점수는 평균 72점과 표준편차 8점으로 정규분포한다고 알려졌다. 하위 40% 이하는 무조건 탈락이라고 한다. 과락을 면하려면 최소한 몇 점을 넘어야 하나?
- 답안: =NORM.INV(.4, 0, 1)는 약 -0.25이다. 확률로 0.4를 넣었으니 돌아온 것은 Z-점수다. -0.25라는 Z-점수로는 뭐가 뭔지 알 수 없고, 요걸 다시 원점수로 환산해야 이해할 수 있다. 아까 배운 식을 활용해서 $-.25 \cdot \sigma + \mu = 70$ 를 얻을 수 있다. 즉, 70점을 넘어야 한다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

끝!

와~ 이번 숙제는 엑셀 문제네~

충북대학교
CHUNGBUK NATIONAL UNIVERSITY