Chapter. 3. 경제 분석 시 등장하는 분포

1. 개요

이 장에서는 각종 확률분포를 소개한다. 경제 데이터를 분석하다 보면 그 데이터의 모집단 또는 데이터 중 관측된 변수로 설명하기 어려운 오차의 확률분포에 대해서 모형화를 하는 경우가 많다. 그러한 모형화가 상황에 따라 데이터에 따라 적합할 때도 있고 그렇지 않을 때도 있다.

항상 그런 것은 아니지만, 모형화를 한다는 것은 모수적 확률분포를 사용하는 것을 의미한다. 이 장에서는 주요한 모수적 확률분포 모형을 소개한다. 즉, 몇 개의 모수에 의해 분포의 모든 특성이 결정되는 확률분포를 다룬다.

2. 이산분포와 연속분포의 의미

집합 $\{1,2,3\}$ 중 하나의 값으로 실현되는 확률변수는 이산변수(discrete variable)이다. 집합 $\{0,1\}$에 속하는 하나의 값으로 나오는 변수 역시 이산변수다. 집합 $\{0,1,2,...\}$에 속한 하나의 값으로 나오는 확률변수도 이산변수이다. 다만, 집합 $\{1,2,3\}$의 원소 개수가 3개이고, 집합 $\{0,1\}$의 원소 개수가 2개인데 반해, 집합 $\{0,1,2,...\}$에 속하는 원소의 개수는 무한히 많다. 무한히 많다고 하여 반드시 연속변수가 되는 게 아니다. 예를 들어, 0 또는 자연수 중 하나의 값으로 실현 가능한 대표적인 확률분포로 포아송 분포(Poisson distribution)가 있다.

자연수의 집합, 정수의 집합, 유리수의 집합 등은 원소 개수가 무한하지만 이산적이다. 유리수 집합이 이산적이라는 것을 바로 수긍하기 어렵지만 이산적이다. 반면, 실수 구간 $[0,1]$과 같은 집합을 연속체라고 한다. 실수 집합도 연속체이다. 그런 연속체 내의 모든 값이 가능한 확률변수는 연속분포를 따른다고 한다. 대표적으로 정규분포가 연속분포이다. 앞서 나왔던 F-분포, 베타분포, 로그-정규분포 등 모두 연속분포이다.

3. 확률질량함수, 확률밀도함수 및 누적분포함수

이산분포는 확률(질량)함수로 묘사되고, 연속분포는 확률밀도함수로 묘사된다. 확률(질량)함수는 그 자체로 확률을 나타내고 확률밀도함수는 확률이 아니라 말 그대로 확률밀도이다.

예를 들어, 이산확률변수 X에 대한 확률(질량)함수 $f(x)$는 $\Pr(X=x)$이다. 반면, 연속확률변수 Y에 대한 확률밀도함수 $p(y)$는 $\Pr(Y=y)$가 아니다. 아주 근사적인 관계로서 $\Pr(Y \in dy) = p(y)dy$로 나타낼 수는 있다.

이산확률변수이든 연속확률변수이든 변수 X에 대해서 $\Pr(X \leqq x)$를 누적분포함수(cumulative distribution function: c.d.f.) 또는 분포함수(distribution function: d.f.)라고 부른다. $\Pr(X \leqq x)$는 x에 따라 달라지는 함수임에 유의하라. (확률)변수 X의 확률질량함수 또는 확률밀도함수가 $f(x)$일 때, 누적분포함수 $F(x)$는 다음과 같다.

변수 X가 이산분포인 경우, $F(x) = \Pr(X \leqq x) = \sum_{s \leqq x} f(s)$ (1)

변수 X가 이산분포인 경우, $F(x) = \Pr(X \leqq x) = \int_{-\infty}^{x} f(s)ds$ (2)

4. 이산분포

가. 베르누이 분포(Bernoulli distribution)

앞서 언급했듯이, 베르누이 분포는 0 또는 1로 실현되며, 1이 나올 확률이 p, 0이 나올 확률이 $1-p$인 확률분포이다. 확률변수 X가 이러한 베르누이 분포를 따를 때, $X \sim Ber(p)$로 표시한다. 모수 p는 $X=1$인 확률인데, 이를 비율 또는 성공확률이라고도 부른다. 이는 베르누이 분포 $Ber(p)$의 평균이다.

X의 확률질량함수는 $f(x) = p^x(1-p)^{1-x}$, $x = 0, 1$이다. 아래에서 $E(X)$는 변수 X의 평균 또는 기대치, σ^2는 분산을 나타낸다. 분산은 $var(X)$로 표시하는 경우도 많다.

베르누이 분포

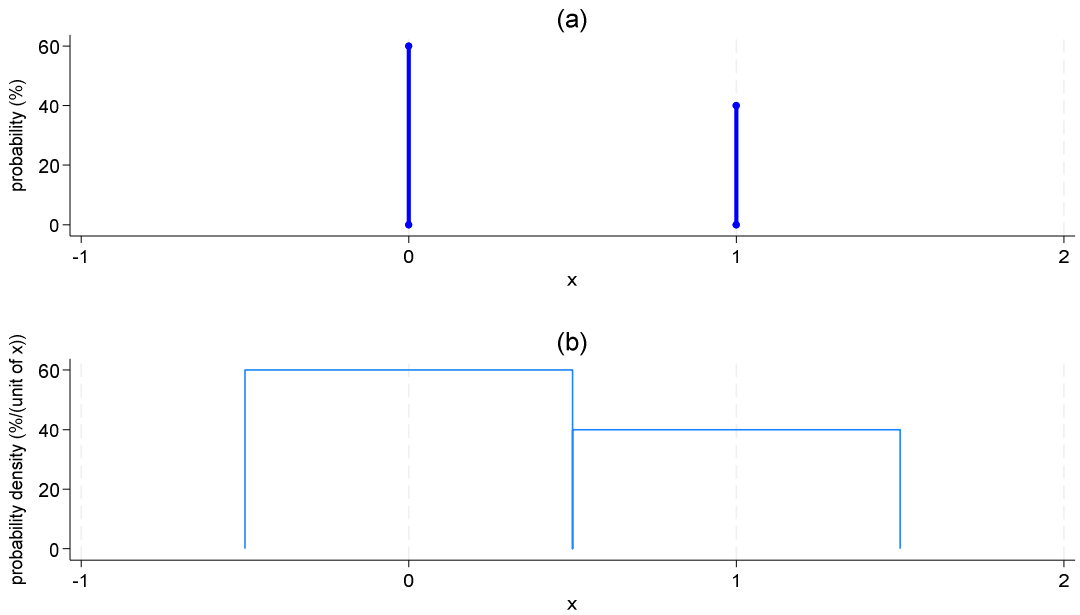
$$X \sim Ber(p) \Rightarrow f(x) = p^x (1-p)^{1-x},\ x = 0,1$$
$$E(X) = p,\ \sigma^2 = var(X) = p(1-p)$$

변수 Y가 p의 확률로 1, $1-p$의 확률로 (-1)의 값을 갖는 확률변수이면 간단한 변환을 통해 베르누이 분포로 바꿀 수 있다. 즉, $X = (Y+1)/2$로 바꾸면 변수 X는 베르누이 분포 $Ber(p)$를 따른다.

베르누이 분포는 어떤 한 번의 시행에서 성공확률이 p인데, 그러한 시행을 1번 행하였을 때의 성공의 횟수이다. (성공은 0번 아니면 1번) 그러한 시행을 독립적으로 n번 하였을 때, 성공의 횟수는 이후 설명하는 이항분포 $B(n,p)$를 따른다. 이때 각 시행을 베르누이 시행(Bernoulli trial)이라고 부른다.

아래 <그림 Ⅲ-1>은 베르누이 분포 $Ber(0.4)$를 묘사하는 확률질량함수 및 모집단 히스토그램이다.

<그림 Ⅲ-1> 베르누이 분포 Ber(0.4)의 확률질량함수 및 모집단 히스토그램



나. 이항분포(binomial distribution)

이항분포를 조금 어려운 말로 시작해보자. 모집단에 1이 p, 0이 $1-p$의 비율로 들어 있다고 하자. 모집단은 베르누이 분포 $Ber(p)$를 따르는 셈이다. 이 모집단에서 복원추출로 X_1, X_2, ... , X_n을 추출해 보자. 그러면, 각 X_i는 0 아니면 1이며, 각 X_i들은 서로 독립이다. 이를 $X_i \sim i.i.d.\ Ber(p)\ (i=1,\cdots,n)$로 표시한다. 이 때 $Y=X_1+X_2+\cdots+X_n$로 변수 Y를 정의하자. 그러면, 변수 Y는 0 또는 1로 나오는 모든 X_i의 합이다. 그러한 n개의 합 Y는 이항분포 $B(n,p)$를 따른다. 이러한 이항분포의 모수는 n과 p의 2개가 있다. 변수 Y는 0, 1, 2, ... , n 등으로 실현 가능하며, Y가 이 중의 특정한 값 r로 나올 확률은 $\Pr(Y=r)=\binom{n}{r}p^r(1-p)^{n-r}$ $(r=0,1,\cdots,n)$이다. $\binom{n}{r}=\frac{n!}{r!(n-r)!}$는 n개 중 r개 뽑는 조합의 수로 ${}_nC_r$와 같다.

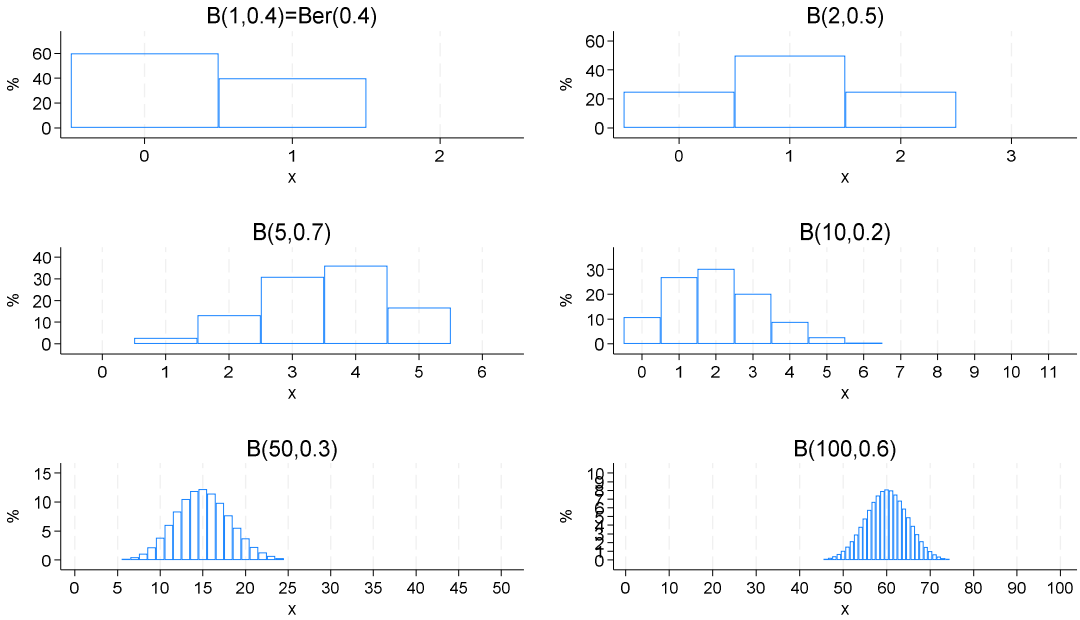
이항분포

$$X \sim B(n,p) \Rightarrow f(x)=\binom{n}{x}p^x(1-p)^{n-x},\ x=0,1,\cdots,n$$

$$E(X)=np,\ var(X)=np(1-p)$$

<그림 Ⅲ-2>에는 다양한 n 및 p에서의 이항분포 $B(n,p)$의 모집단 히스토그램을 보이고 있다. 가장 먼저 나온 $B(1,0.4)$는 베르누이분포 $Ber(0.4)$와 동일하다. 즉, $n=1$이 되면 이항분포는 베르누이 분포이다.

<그림 III-2> 다양한 모수 하의 이항분포의 모집단 히스토그램



이항분포를 다음과 같이 설명할 수도 있다. 매번 독립 시행에서 성공이나 실패로 실현되는데 성공확률은 p이다. 이러한 독립적 시행을 n번할 때 성공의 횟수는 이항분포 $B(n,p)$를 따른다. 이항분포를 따르는 확률변수 역시 확률변수이므로 반드시 숫자여야 한다. 성공, 실패 이런 게 그 변수의 값이 아니다. 그래서, 변수 X가 이항분포 $B(n,p)$를 따른다면, X는 성공의 횟수이므로, 0, 1, 2, ... , n 중 하나로 실현된다.

다. 초기하 분포(hyper-geometric distribution)

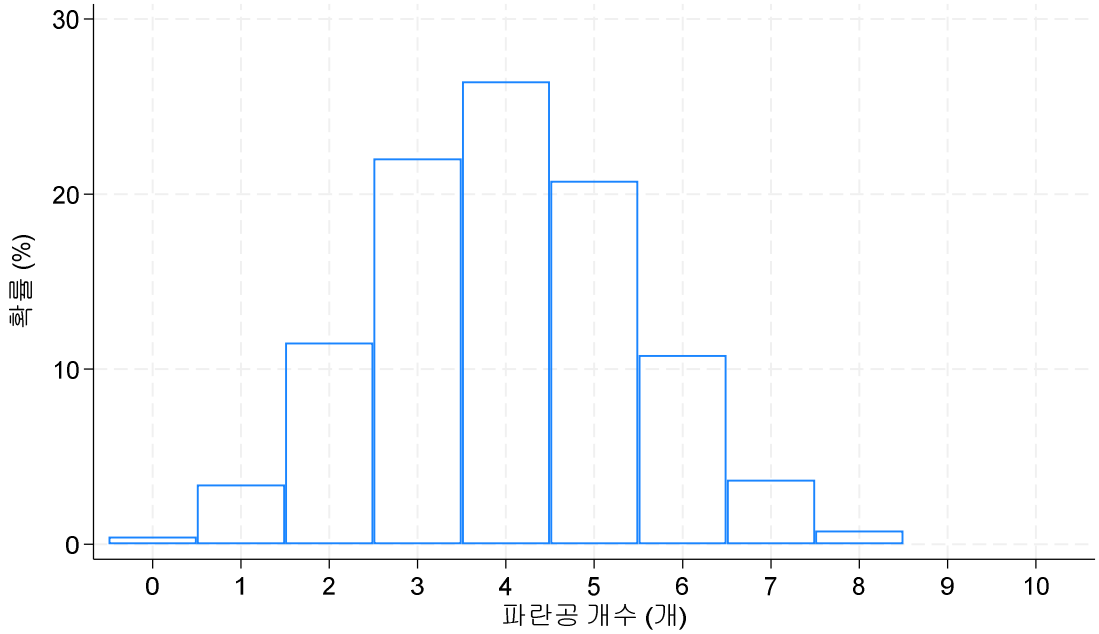
이항분포와 유사하지만 조금 어려운 분포이다. 한 주머니 안에 빨간 공이 60개, 파란 공이 40개 들어 있다고 하자. 복원추출로 주머니에서 임의적으로 공 10개를 추출하자. 이 중 파란 공의 개수를 X라고 하면, 이 변수 X는 이항분포 $B(10,0.4)$를 따른다. 파란 공의 개수가 X이므로 이는 숫자 0, 1, ... , 10 중 하나로 실현된다.

반면, 꺼냈던 공을 다시 주머니에 집어 넣지 않는 비복원 추출이면 어떨까? 위의 주머니에서 비복원추출로 10개를 뽑을 때, 파란공의 개수 X는 아래 식을 확률질량함수로 갖는 이른바 초기하분포(hyper-geometric distribution)이다.

$$f(x)=\Pr(X=x)=\frac{\binom{40}{x}\binom{60}{10-x}}{\binom{100}{10}} \tag{3}$$

아래 <그림 Ⅲ-3>은 초기하분포 $Hypergeo(100,40,10)$를 나타내는 모집단 히스토그램이다.

<그림 Ⅲ-3> 파란공 40개를 포함한 100개 공이 들어 있는 주머니에서 10개 비복원 추출 시 파란공 개수 X의 분포 (초기하분포 Hypergeo(100,40,10))



일반적으로, N개의 개체가 들어 있는 주머니 안에 어떤 특정 성질을 갖는 개체가 M개 있는데, 이 주머니에서 n개를 임의로 비복원추출하는 문제를 살펴보자. 추출된 n개 중 그 특정 성질을 갖는 개체의 수가 X개라 하자. 그러면, 확률변수 X는 초기하분포 $Hypergeo(N,M,n)$를 따른다

초기하분포

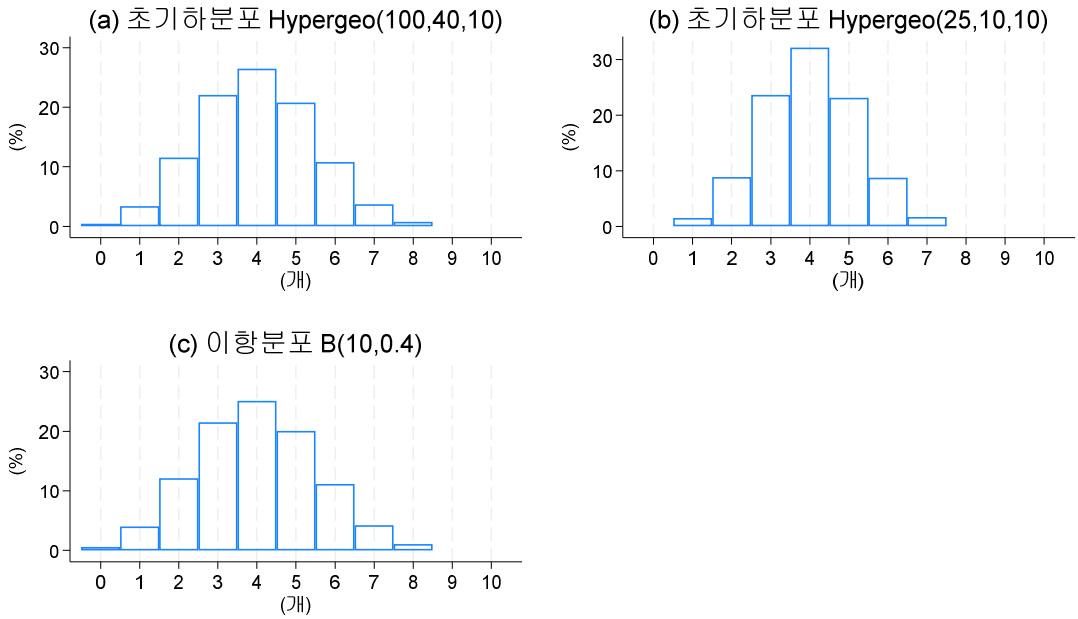
$$X \sim Hypergeo(N,M,n) \Rightarrow f(x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} \quad (x = 0,1,\cdots,k)$$

$$E(X) = n \cdot \frac{M}{N},\ var(X) = n \cdot \frac{M}{N} \cdot \left(1 - \frac{M}{N}\right) \cdot \frac{N-k}{N-1}$$

초기하분포는 이항분포와 유사하다. 이항분포는 복원추출, 초기하분포는 비복원추출의 차이일 뿐이다. 이항분포 $B(n,p)$의 평균은 np, 분산은 $np(1-p)$인데, 초기하분포의 평균과 분산도 이와 크게 다르지 않다. (M/N이 성공확률 p인 셈) 다른 점이 있다면, 분산에 $(N-n)/(N-1)$이 곱해져 있다는 것이다. (표준편차에는 $\sqrt{(N-n)/(N-1)}$가 곱해질 것이다.) 그런데, N이 충분히 크면 $(N-n)/(N-1)$는 1과 거의 같다. 결국, 모집단 내 전체 개체 수가 크게 많으면, 즉 위 예시에서는 주머니 속 공의 수가 많아지면, 이항분포나 초기하분포나 사실상 비슷해진다.

아래 <그림 Ⅲ-4>는 초기하분포와 이항분포를 비교하고 있다. (a)는 모집단에 총 100개의 공이 있고 그 중 파란 공이 40개 들어 있을 때, 비복원으로 10개 추출할 때 파란 공의 개수이다. (바로 앞에서 보았던 사례) 반면, (b)는 모집단에 총 25개의 공이 있고 그 중 파란 공이 10개 들어있을 때, 비복원으로 10개 추출시 파란 공 개수이다. 이들 두가지 모집단은 파란 공의 비율이 40%라는 공통점이 있다. (c)는 이항분포인데 모집단에서 어떤 특정한 성질을 갖는 개체의 비율이 40%일 때, 복원추출로 10개 추출할 때 그 관심 있는 개체가 나오는 횟수이다.

<그림 III-4> 초기하분포와 이항분포의 비교



<그림 Ⅲ-4>의 (a), (b), (c)를 잘 비교해 보라. (a)의 모집단이나 (b)의 모집단이나 파란공 비율이 40%라는 점은 동일하다. 그래서, p가 0.4인 이항분포와 크게 차이가 없어야 한다. 그러나, (a)는 이항분포 (c)와 거의 차이가 없고, (b)는 약간 차이를 보인다. 이유가 무엇일까? (a)에는 공이 총 100개, (b)에는 총 25개 들어 있기 때문이다.

모집단의 모든 개체 수가 충분히 많으면, 즉 N이 커질 때, 초기하분포와 이항분포는 차이가 점차 사라진다. 즉 비복원추출과 복원추출이 큰 차이가 없다. 비복원추출하더라도 매번 파란 공 뽑을 확률이 40%와 크게 달라지지 않기 때문이다. 그러나 (b)의 경우에는 무시하기 어려울 정도로 달라진다. 복원추출 시에는 항상 40%가 유지된다.

마. 포아송분포 (Poisson distribution)

포아송분포는 좀 어렵지만 매우 중요한 분포이며 이산분포이다. 포아송 분포는 특정 사건이 정해진 시간(1초, 1시간, 1일, 1년 등)동안 발생하는 횟수를 모형화하기에 좋다. 예를 들어, 1시간 동안 전화가 오는 횟수, 어떤 은행원 창구에 1일 동안 고객 방문 수, 주식시장에서 1시간 동안 주식거래 체결 횟수, 1개월 동안 구직에 성공할 횟수 등이다.

1개월 동안 구직자가 일자리 제안을 받을 횟수를 한 번 생각해보라. 1개월 동안

일자리 제안이 1건도 받지 못할 수 있다. 같은 기간 동안 일자리 제안을 1건 받을 수도 있다. 1개월 동안 2건 일자리 제안을 받을 수도 있다. 1개월 동안 일자리 제안을 10건 받을 수도 있다. 이렇게 계속 이어가면, 1개월 동안 일자리 제안을 받을 횟수 X는 자연수 어느 값이라도 가능하다. 물론, X가 10 이상 넘어가면 그럴 확률은 극히 낮을 것이다.

예를 들어, 내게 1개월 동안 평균적으로 5건 일자리 제안을 받는다고 해보자. 그러면, 내게 1개월 동안 일자리 제안을 받을 횟수 X는 확률변수로서 0, 1, 2, ... 중 하나로 실현되며, 그 확률질량함수는 $f(x) = \Pr(X = x) = 5^x e^{-5}/x!$가 된다. 여기서, 평균인 5를 이 포아송 분포의 강도(intensity) 또는 위험률(hazard rate)이라 한다. 이는 포아송 분포의 평균이기도 하다.

일반적으로, 변수 X가 강도 λ인 포아송 분포를 따를 때, $X \sim Poi(\lambda)$로 표시한다. 포아송 분포는 평균과 분산이 동일한 특성이 있다는 점도 유의하자.

포아송 분포

$$X \sim Poi(\lambda) \Rightarrow f(x) = \frac{\lambda^x e^{-\lambda}}{x!} \quad (x = 0,1,2,\cdots)$$

$$E(X) = \lambda, \; var(X) = \lambda$$

X가 포아송 분포 $Poi(\lambda)$를 따를 때, X는 주어진 시간 구간 내에서 사건의 발생 횟수가 된다. 이 때, 변수 T_1를 첫 번째 사건이 발생한 시점까지의 시간, T_2를 첫 번째 사건 발생 후 두 번째 사건이 발생한 시점까지의 시간, T_3를 두 번째 사건 발생 후 세 번째 사건이 발생한 시점까지의 시간 등으로 정의해보자. 그러면, T_j가 모든 자연수 j에 대해서 정의될 것이다. 이러한 T_j를 $(j-1)$번째 사건으로부터 j번째 사건까지의 사건 간 시간 구간(inter-arrival time, inter-event time)이라고 한다.

이들 사건 간 시간 구간 T_1, T_2, ... 는 매우 중요한 성질을 갖는다. 일단 이들은 모두 독립이다. 나아가, 모든 j에 대해서 T_j는 포아송 강도 λ를 (위험률) 강도(hazard rate intensity)로 갖는 지수분포를 따른다. 이를 $T_j \sim i.i.d.\,\mathrm{Exp}(\lambda)$로 표현하며, T_j의

확률밀도함수는 $f_{T_j}(t) = \lambda e^{-\lambda t}$ $(t \geq 0)$이다. 포아송분포와 지수분포는 동전의 양면이다. 그런데, 포아송분포는 이산분포인데 지수분포는 연속분포이다.

유용한 경제통계
호경기가 올 것인가?

한 나라의 경제는 잠재 경제성장율 보다 높아졌다가 낮아졌다가 하는 경기순환(business cycle)을 따르게 된다. 경제가 좋아졌을 때를 호경기, 나빠졌을 때를 불경기라고 부르며, 호경기와 불경기는 주기적으로 발생하게 된다. 언제 호경기가 오고 언제 불경기가 오는지를 알 수 있을까?

경기는 확장, 후퇴, 수축, 회복의 과정을 반복하게 된다. 이러한 경기를 잘 예측할 수 있다면, 돈을 벌 수도 있고, 재산을 지킬 수도 있을 것이다. 경기에 대한 예측이 경제정책에도 매우 중요할 것이라는 건 두 말할 필요도 없다. 경제를 예측하는 기관에서는 이러한 경기를 가늠해 볼 수 있는 다양한 지표를 만들어서 발표함으로써 경제주체들의 경제 의사 결정에 도움을 주고 있다.

먼저 생각해 볼 수 있는 지표는 경제주체들이 경기에 대해서 어떻게 생각하고 있는지를 나타내는 것이다. 적절한 표본추출 과정을 통해 대표적인 소비자와 기업들이 경제를 보는 시각을 정리해서 나타낼 수 있으면 경제주체들이 어떻게 경기를 인식하고 있는지 파악할 수 있을 것이다. 대표적인 지표가 경제심리지수이다. 경제심리지수는 기업가를 대상으로 하는 기업경기실사지수(BSI, Business Survey Index)와 소비자를 대상으로 하는 소비자동향지수(CSI, Consumer Survey Index)가 있다.

기업경기실사지수는 한국은행에서 매월 3천 여개의 기업을 대상으로 진행하고 있다. 대사대상 업체는 국세청 법인세 신고업체들을 대상으로 업종별, 매출액별로 층화계통추출법을 표집하여 조사하고 있다. 조사 내용은 판단조사와 계수조사로 구분된다. 판단조사는 항목별로 긍정, 보통, 부정 중 하나를 선택하게 하여 전체 응답 중에서 긍정적인 응답 비중과 부정적인 응답 비중의 차이로 기업경기실사지수를 산출한다. 계수조사는 항복별로 실제금액의 증감률을 조사하게 된다. 이렇게 하여 산출된 BSI는 0에서 200까지의 값을 갖는데 100을 넘으면 긍정적으로 응답한 업체수가 부정적으로 응답한 업체수보다 많음을 나타낸다. 100 미만의 경우는 그 반대이다.

소비자동향조사는 소비자의 경제에 대한 인식과 향후 소비지출전망 등을 조사하여 지수화한 것이다. 매월 15일 전후하여 조사하고 조사대상가구는 인구주택총조사의 전국 도시 일반가루를 대상으로 지역, 연령별로 층화추출한 2,500가구이다. 소비자의 현재 경제상황에 대한 판단, 향후 경제 및 소비지출에 대한 전망과 관련된 질문으로 구성되어 있다. BSI와 비슷한 방식으로 산출되며, CSI가 100을 초과한 경우 긍정적인 답

변을 한 소비자가 부정적인 답변을 한 소비자보다 많다는 것을 의미한다. 100 미만인 경우는 그 반대를 의미한다.

이러한 심리지표 외에 객관적인 지표로는 개별 경제지표와 종합경기지수 등이 경기판단에 사용되고 있다. 경기동향 분석에 사용되는 대표적인 지표는 전산업생산지수, 생산자출하지수 및 제품재고지수, 제조업생산능력지수, 가동률지수 등을 들 수 있다.

전산업생산지수는 전체 산업생산 활동의 단기동향을 파악할 수 있는 지표이다. 광공업, 서비스업, 건설업, 공고행정, 농림어업 생산지수 등 5개의 지표로 구성되어 있다. 생산자출하지수는 생산자의 판매활동 수준을 나타내는 지표이다. 생산자제품재고지수는 생산자의 재고보유 수준을 나타내는 지표이다. 기업들의 예상보다 경기가 좋아 판매가 잘되면 재고가 줄고, 반대로 경기가 나빠 판매가 부진하면 재고가 증가하게 된다. 제조업생산능력지수는 제조업체들의 제품 생산능력수준을, 가동률지수는 생산설비의 활용정도를 나타내는 지표이다.

소비, 투자, 수출입 등을 나타내는 지표도 있다. 가계의 소비를 대표하는 지표로는 소매판매액지수, 소비재내부출하지수, 소비재 수입 등이 있다. 기업의 투자를 대표하는 지표로는 건축허가 및 착공면적, 건설수주액, 건설기성액을 이용해 건설 투자를 살펴본다. 설비투자로는 기계수주액, 설비투자지수, 기계류수입액 등이 있다. 건설수주와 기계수주는 장래의 투자활동을 가늠할 수 있는 선행지표로 사용된다.

국가경제의 전체적인 상황을 파악하는 데에는 경기종합지수가 사용되고 있다. 통계청은 매월 경기종합지수를 개별 경제지표들을 기반으로 하나의 지수로 가공하여 발표하고 있다. 경기종합지수는 경기에 대한 선·후행 관계에 따라 선행, 동행, 후행종합지수로 구분하여 사용하고 있다.

5. 연속분포

연속분포는 그 종류가 상당히 많지만, 주로 많이 사용하고 적어도 반드시 알아야 하는 몇 가지 연속분포를 소개한다. 앞서 언급했듯이, 연속분포는 모집단이 연속체인 확률분포이다. 그 연속체 내에서 실현가능한 값들을 서로 분리하여 표시할 수 없다. 즉, 첫 번째 수, 두 번째 수, ... 등의 표시가 불가능하다. 또한, 실현 가능한 값 x에 대해서 그 값에 대응하는 확률밀도함수 $f(x)$로 그 분포를 묘사할 수 있다. 다만, 앞서 많이 강조했듯이 $f(x)$는 확률이 아닌 확률밀도이다. 연속변수 X가 특정 하나의 수치 x로 나올 확률은 항상 0이다

가. 정규분포 (normal distribution)

정규분포는 반드시 알아야 한다. 확률통계학을 지속적으로 깊이 있게 공부하다 보면, 너무나도 신기한 분포가 정규분포임을 알게 된다. 이후 설명하지만, 모집단 분포에 무관하게 임의의 모집단 분포로부터 추출된 표본 X_1, X_2, ... , X_n에 대해서 그 (표본)합 $X_1 + X_2 + \cdots + X_n$이나 (표본)평균 $(X_1 + X_2 + \cdots + X_n)/n$의 분포는 반드시 정규분포로 수렴한다. 이를 중심극한정리라고 한다. 추후 다시 얘기하자.

정규분포의 밀도함수 정도는 외울 필요가 있다. 변수 X가 평균이 μ이고, 분산이 σ^2(즉, 표준편차는 σ)인 정규분포를 따를 때, $X \sim N(\mu,\sigma^2)$으로 표시한다. 이때 변수 X의 확률밀도함수는 다음과 같다.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{4}$$

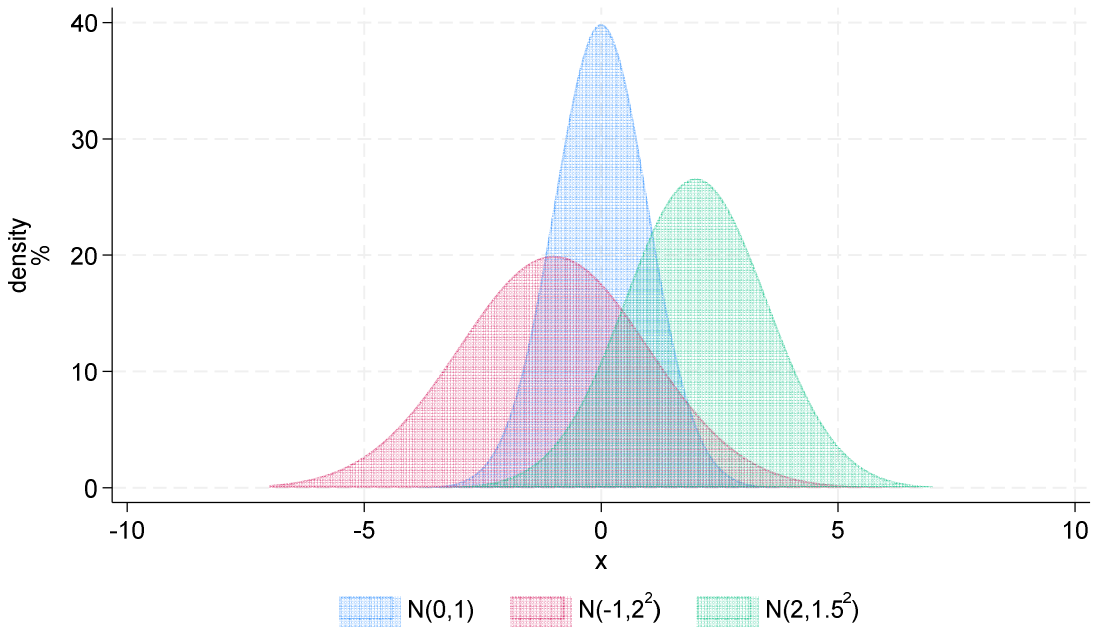
이 밀도함수에서 보듯, μ와 σ^2라는 두 가지 수만 정해지면 그 정규분포가 정확히 무엇인지 유일하게 정해진다. 이러한 수를 그 분포의 모수(parameter)라고 한다. 예를 들어, 앞서 잠시 설명한 지수분포의 밀도함수는 $\lambda e^{-\lambda t}$였는데 여기서 강도라 불리우는 λ가 모수이다. 베르누이 분포 $Ber(p)$에서는 p가 모수이다. 이항분포 $B(n,p)$에서는 n과 p가 모수이다. 초기하분포 $Hypergeo(N,M,n)$에서는 N, M, n의 3가지 모수가 있다.

한편, $X \sim N(\mu,\sigma^2)$이면, X의 기대치는 다음과 같다.

$$E(X) = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx \tag{5}$$

이 적분의 계산은 복잡하지만 그 답은 당연히 μ이다. 즉, $E(X) = \mu$. 또한, $var(X) = \sigma^2$이 된다. 정규분포는 평균과 분산이라 불리우는 2개의 모수가 있다. 아래 <그림 Ⅲ-5>는 몇 가지 정규분포의 밀도를 그린 것이다.

<그림 III-5> 몇 가지 정규분포의 밀도함수




```
. clear

. range x -7  7   200
Number of observations (_N) was 0, now 200.

.
. gen y1=100*normalden(x,0,1)

. gen y2=100*normalden(x,-1,2)

. gen y3=100*normalden(x,2,1.5)

.
. twoway ///
> (line y1 y2 y3 x, recast(area) color(%20 %20 %20)) , ///
> xtitle() ytitle("density" "%")  ///
> legend( label(1 "N(0,1)")   ///
>                 label(2 "N(-1,2{sup:2})") ///
>                 label(3 "N(2,1.5{sup:2})") ///
>                 pos(6) row(1))
```

정규분포에서 나오는 몇 가지 확률은 외울 필요가 있다. 변수 Z가 표준정규분포 $N(0,1)$을 따른다고 하자. Z의 확률밀도함수를 ϕ, 누적분포함수를 Φ라고 하자. 이 때 아래 확률에 관한 식을 기억하면 종종 편리하다.

$$\Pr(-1 \leqq Z \leqq 1) = \int_{-1}^{1} \phi(u)du = \Phi(1) - \Phi(-1) = 68\% \tag{6}$$

$$\Pr(-2 \leqq Z \leqq 2) = \int_{-2}^{2} \phi(u)du = \Phi(2) - \Phi(-2) = 95\%$$

$$\Pr(-3 \leqq Z \leqq 3) = \int_{-3}^{3} \phi(u)du = \Phi(3) - \Phi(-3) = 99.7\%$$

변수 X가 정규분포 $N(\mu,\sigma^2)$을 따르면, 아래 식은 상기 식과 동일한 의미를 갖는다. 정규분포에서는 평균으로부터 표준편차 단위로 ±1 이내의 확률은 68%, ±2 이내의 확률은 95%, ±3 이내의 확률은 99.7%이다. 정규분포에서는 표준편차 단위로 평균으로부터 3단위 이상 먼 영역의 확률은 0.3%, 즉 3/1000에 불과하다는 점을 주지하자. 그 정도로 정규분포에서 꼬리 영역은 얇다.

$$\Pr(\mu - \sigma \leqq X \leqq \mu + \sigma) = \Pr(-1 \leqq Z \leqq 1) = 68\% \tag{7}$$

$$\Pr(\mu - 2\sigma \leqq X \leqq \mu + 2\sigma) = \Pr(-2 \leqq Z \leqq 2) = 95\%$$

$$\Pr(\mu - 3\sigma \leqq X \leqq \mu + 3\sigma) = \Pr(-3 \leqq Z \leqq 3) = 99.7\%$$

다음 식도 외워둘 필요가 있다. 변수 X가 정규분포 $N(\mu,\sigma^2)$을 따르면, 다음의 식들이 성립한다.

$$\Pr[\mu - 1.64\sigma \leqq X \leqq \mu + 1.64\sigma] = 90\% \tag{8}$$

$$\Pr[\mu - 1.96\sigma \leqq X \leqq \mu + 1.96\sigma] = 95\%$$

$$\Pr[\mu - 2.57\sigma \leqq X \leqq \mu + 2.57\sigma] = 99\%$$

식은 양측 꼬리 면적이 10%, 5%, 1%가 되는 포인트가 평균으로부터 1.64 단위,

1.96 단위, 2.57 단위 떨어져 있다는 걸 의미한다. 통계적 유의성 검정(statistical significance test)을 할 때 수시로 등장하므로, 1.64, 1.96, 2.57은 꼭 외워야 한다.

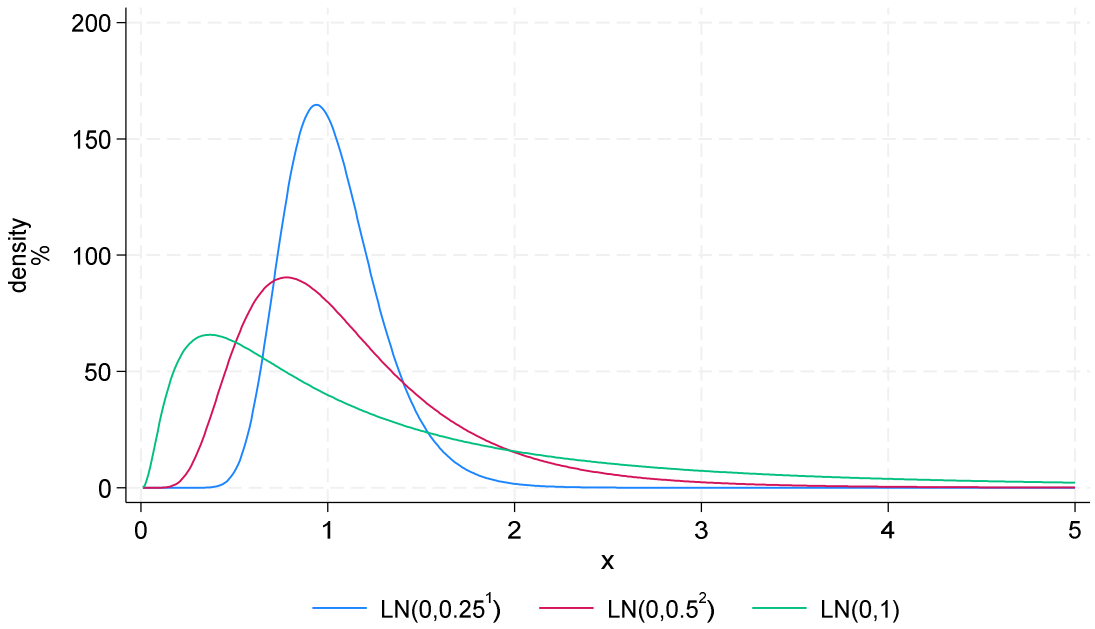
나. 로그정규분포

로그정규분포는 경영·경제 문제에서 양의 값만 갖는 변수를 다룰 때 종종 등장한다. 통상 학부 통계학에서 잘 소개되지 않는 편이지만, 실증 분석 및 응용에서 로그정규분포는 중요하다.

자연로그 ln(.)을 취하면 정규분포를 따르는 변수 X가 있다고 하자. 그러면 변수 X는 로그정규분포(또는 대수정규분포, log-normal distribution)를 따른다. 이 말이 사실 좀 어렵다. 그래서, 조금 더 쉽게 얘기하자면 $X=e^{Y}$인데 여기서 Y가 정규분포를 따르면 변수 X는 로그정규분포를 따른다. 일반적으로 X가 확률변수인데, 임의의 변환 Y=f(X)을 나오는 변수 Y 역시 또 다른 확률변수가 된다. 물론, Y의 확률적 특성은 순전히 X의 확률적 특성에 기인한다. ln(X)=Y가 되고, Y는 앞서 말한대로 정규분포이다. 이를 아래와 같이 조금 더 이해하기 쉽게 말해보자.

X가 로그정규분포를 따른다는 것은 정규분포를 따르는 어떤 변수 W가 있고, $X=e^{W}$라는 것이다. 이는 다시 $W=\ln X$, 즉 X에 로그를 취하니 정규분포를 따르는 변수 W이 나온다. $e^{(\cdot)}$는 항상 0보다 크기 때문에 로그정규분포 X는 반드시 양(+)의 값으로만 실현된다.

<그림 III-6> 몇 가지 로그정규분포의 밀도함수




```
. range x 0.01  5   500
Number of observations (_N) was 0, now 500.

. gen y1=100*normalden(ln(x),0,0.25)/x

. gen y2=100*normalden(ln(x),0,0.5)/x

. gen y3=100*normalden(ln(x),0,1)/x

. twoway (line y1 y2 y3 x,  ) , ///
> xtitle() ytitle("density" "%")  ///
> legend( label(1 "LN(0,0.25{sup:1})")   ///
>                 label(2 "LN(0,0.5{sup:2})") ///
>                 label(3 "LN(0,1)") ///
>                 pos(6) row(1))
```

다. 지수분포

앞서 포아송 분포와 함께 연관 지어 지수분포를 설명한 바 있다. 둘 간의 연관성은 매우 중요하다. 시간 구간을 정해 놓고 그 시간 구간동안 특정 사건의 발생 횟수가 관심이면 포아송 분포를 사용하는 것이고, 사건으로부터 다음 사건이 벌어질 때까지 걸리는 시간이 관심이면 지수분포를 사용한다. 포아송 분포나 지수분포나 강도또는 위험률로 불리우는 λ라는 모수 1개로 모든 전모가 결정되는 확률분포이다. 둘은 일종의

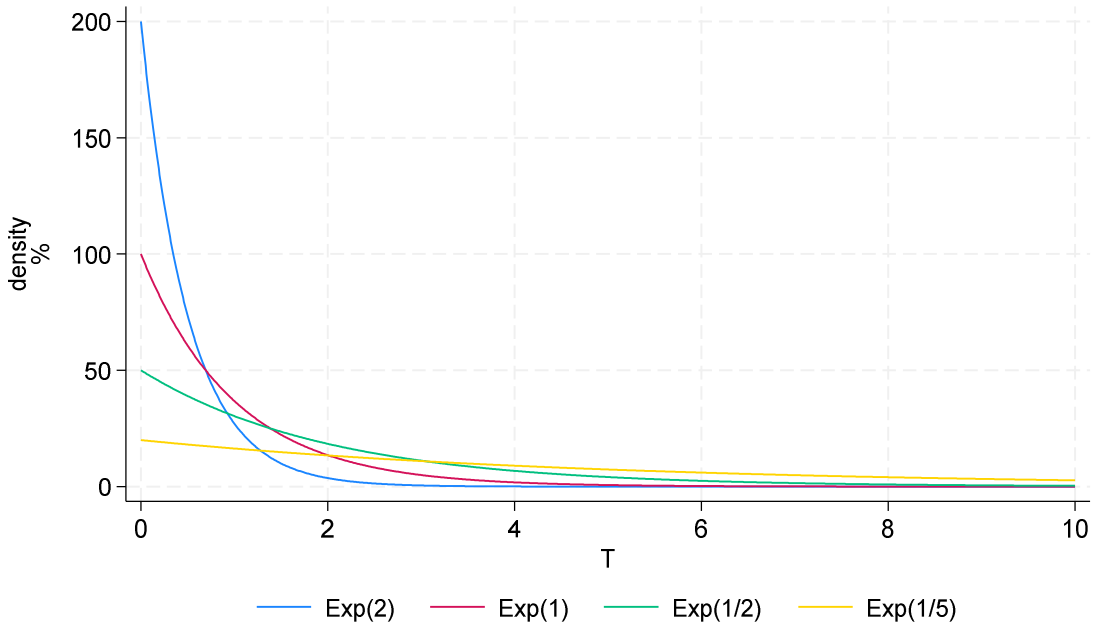
형제 같은 것이다. 포아송 분포 $Poi(\lambda)$의 평균은 λ이고, 지수분포 $\mathrm{Exp}(\lambda)$의 평균은 $1/\lambda$이다. 예를 들어, 1시간 동안 사건이 평균 2번 발생한다면, 사건 간 시간(첫번째 사건에 대해서는 시점 0으로부터의 시간)의 평균은 얼마인가? 1/2시간, 즉 30분일 것이다.

지수분포

$$T \sim \mathrm{Exp}(\lambda) \Rightarrow f(t) = \lambda e^{-\lambda t} \quad (t \geqq 0)$$

$$E(X) = \frac{1}{\lambda}, \; var(X) = \frac{1}{\lambda^2}$$

<그림 III-7> 몇 가지 지수분포의 밀도함수 그래프



```
. range t 0 10 1000
Number of observations (_N) was 0, now 1,000.

.
. gen y0 = 100*exponentialden(0.5,t)

. gen y1 = 100*exponentialden(1,t)

. gen y2 = 100*exponentialden(2,t)

. gen y3 = 100*exponentialden(5,t)

.
. line y* t, ///
> ytitle("density" "%")  xtitle(T) ///
> legend( label(1 "Exp(2)") ///
>                 label(2 "Exp(1)") ///
>                 label(3 "Exp(1/2)") ///
>                 label(4 "Exp(1/5)") ///
>                 pos(6) row(1))
```

지수분포는 특정 사건 발생 시점까지 걸리는 시간을 확률적으로 모형화할 때, 기본적으로 고려해야 하는 확률모형이다. 특정 사건 발생 시점까지 걸리는 시간의 사례는 너무나도 많다. 특정 사건 발생 시점까지 걸리는 시간은 다시 말해서 어떤 상태가 유지되는 지속시간이기도 하다. 이런 시간의 사례는 사람 수명, 동물 수명, 미생물 수명, 자동차 수명, 전구 수명, 호황국면의 기간, 실업 시간, 기업 수명 등 상당히 많다. 그래서, 의학, 공학, 물리학, 화학, 경영학, 경제학 등 다방면의 분야에서 이러한 확률적으로 결정되는 시간을 분석하고자 한다.

이런 확률적으로 결정되는 시간에 대한 분석 기법을 지속기간 분석(duration analysis) 또는 생존분석(survival analysis)이라고 하며, 위험률(hazard rate)이라는 중요한 개념이 등장하게 된다. 지수분포는 생존분석에서 가장 간단한 확률모형이다. 지수분포에서는 바로 강도 λ가 위험률이며 이는 변하지 않는 상수이다.

지수분포는 이산분포가 아니라 연속분포임에 유의하라. 포아송 분포는 이산분포이지만 지수분포는 연속분포이다. 따라서, 지수분포를 따르는 변수 T라면, $T=1$, $T=2$ 등도 가능하지만, $T=0.5$, $T=1/3$도 가능하고, $T=2-\sqrt{2}$ 도 가능하다.

하지만, 때로는 특정 사건까지 걸리는 시간이 이산적인 경우도 있다. 이 경우 T는 0 또는 자연수만 가능한 것이다. ($T=0,1,\cdots$) 이 경우 지수분포를 사용하면 틀린 것은 아니지만 마땅하지는 않다. 이 경우에는 기하분포(geometric distribution)를 사용해야 한다.

라. 카이제곱분포, F-분포, t-분포

이들 3가지 연속분포도 자주 사용되는 확률분포이다. 특히, 경제학을 포함한 사회과학 분야 실증분석에서 엄청나게 많이 이용되는 선형회귀분석에서 t-검정, t-통계량, F-검정, F-통계량, 카이제곱 검정, 카이제곱 통계량 등을 많이 보게 된다. t-검정, t-통계량은 t-분포를 알아야 하고, F-검정, F-통계량은 F-분포를 알아야 하며, 카이제곱 검정, 카이제곱 통계량은 카이제곱 χ^2 분포를 알아야 한다. 이들 모두 정규분포를 변환하여 나오는 분포임에 유의할 필요가 있다. 일단 카이제곱 분포부터 보자.

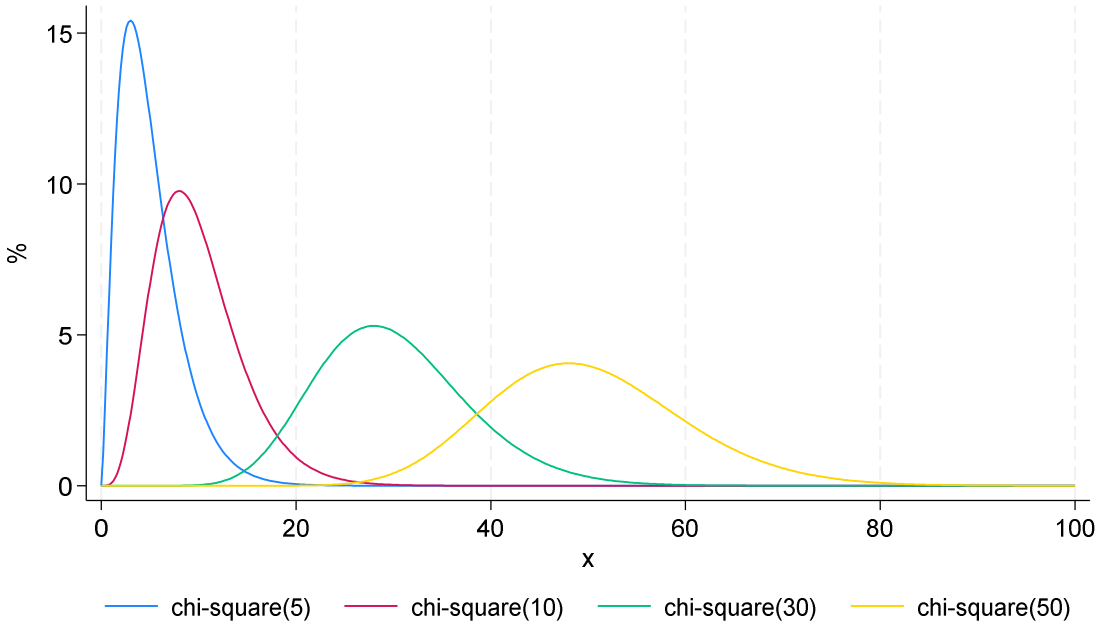
(1) 카이제곱 분포

변수 X가 표준정규분포라고 하자. X^2도 하나의 확률변수로서 다른 분포를 갖게 될 것이다. 이러한 X^2은 카이제곱 분포 $\chi^2(1)$을 따른다. 카이제곱 분포는 모수가 1개 있다. 그 모수의 명칭은 자유도(degree of freedom)이다. 그래서, 자유도가 n인 카이제곱 분포를 $\chi^2(n)$으로 표시한다. 만일, X가 평균이 μ이고, 분산이 σ^2인 정규분포 $N(\mu, \sigma^2)$을 따른다면, $\{(X-\mu)/\sigma\}^2$은 역시 카이제곱 $\chi^2(1)$을 따른다.

다음으로, X_1, X_2, ... , X_n이 모두 표준정규분포를 따르고 서로 독립이라고 하자. (이를 보통 $X_i \sim$ i.i.d. N(0,1)로 표시한다.) 이 때, $\sum_{i=1}^{n} X_i^2$는 카이제곱 분포 $\chi^2(n)$을 따른다. 그러므로, $X_i \sim$ i.i.d. $N(\mu, \sigma^2)$이면, $\sum_{i=1}^{n} \left((X_i-\mu)/\sigma\right)^2$도 카이제곱 분포 $\chi^2(n)$을 따른다.

한편, 카이제곱 분포 $\chi^2(n)$의 평균은 n이고, 분산은 $2n$이다. 카이제곱 분포도 말 그대로 제곱의 분포라서 0 또는 양의 실수로만 국한되며 연속분포이다.

<그림 III-8> 서로 다른 자유도의 카이제곱 분포의 밀도함수



카이제곱 분포

1. $X_i \sim i.i.d.SN(0,1)$이면 $Y=\sum_{i=1}^{n} X_i^2$는 카이제곱 분포 $\chi^2(n)$을 따른다
2. $Y \sim \chi^2(n)$이면, $E(Y)=n,\ var(Y)=2n$이다.

```
. range x 0 100 500
Number of observations (_N) was 0, now 500.

. gen p1 = 100*chi2den(5,x)

. gen p2 = 100*chi2den(10,x)

. gen p3 = 100*chi2den(30,x)

. gen p4 = 100*chi2den(50,x)

.
. tw ///
> (line p1 x) ///
> (line p2 x) ///
> (line p3 x) ///
> (line p4 x),  ///
> ytitle(%) ylabel(,nogrid) ///
> legend( label(1 "chi-square(5)") ///
>                 label(2 "chi-square(10)") ///
>                 label(3 "chi-square(30)") ///
>                 label(4 "chi-square(50)") ///
>                 pos(6) row(1))
```

(2) F-분포

F-분포는 이전 장에서도 언급한 바 있다. F-분포는 다음과 같이 정의되는 모수를 2개 갖는 분포이다. F-분포의 평균과 분산이 있는데, 실제 데이터 분석에서는 이들 평균과 분산이 필요한 경우는 많지 않다. 다만, F-분포 $F(n,m)$에서 평균은 m에만 의존하고 그 m이 충분히 크면 평균이 1에 가까워진다는 점을 기억할 필요가 있다.

F-분포

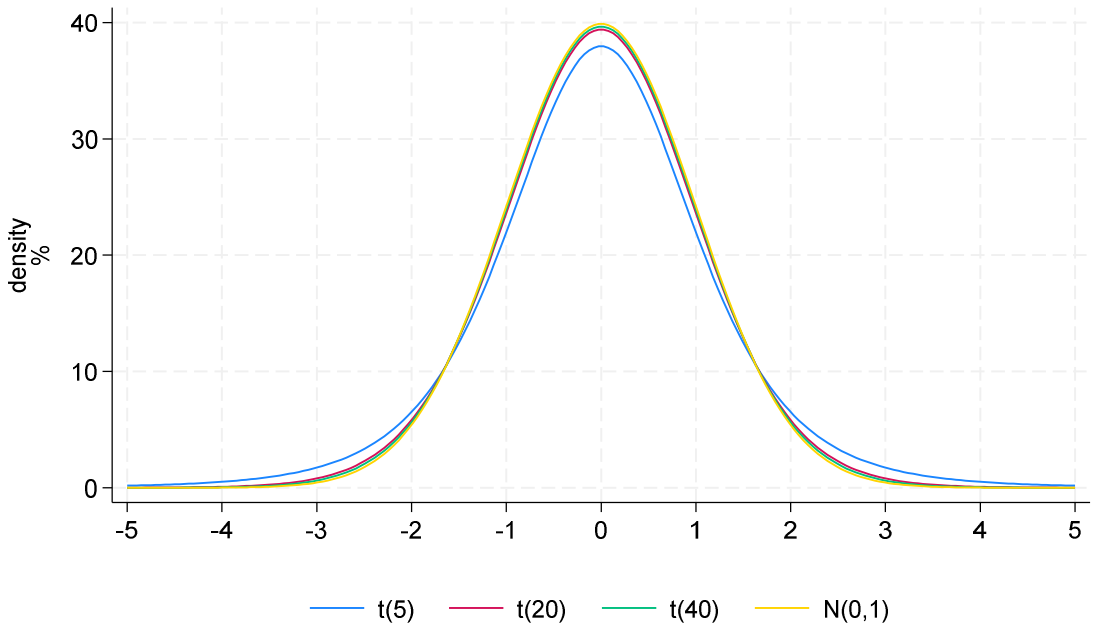
1. $X \sim \chi^2(n),\ Y \sim \chi^2(m)$이면, $F = \frac{X/n}{Y/m}$은 F-분포 $F(n,m)$을 따른다.
2. F-분포 $F(n,m)$의 평균은 $E(F) = \frac{m}{m-2},\ var(F) = \frac{2m^2(n+m-2)}{n(m-2)^2(m-4)}$
3. $F \sim F(n,m)$이면, $\frac{1}{F} \sim F(m,n)$

F-분포도 앞서 본 카이제곱분포와 마찬가지로 0 또는 양수로 실현되는 연속분포이다. 이는 앞서 살펴본 로그정규분포와도 유사하다. 변수 X가 F-분포나 카이제곱 분포를 따를 때, 특정 값 a에 대해서 $\Pr(X \geq a)$를 구해야 한다. 이는 1에서 누적분포함수 값을 뺀 것과 같다. 즉, $\Pr(X \geq a) = 1 - F_X(a)$, 여기서 $F_X(.)$는 변수 X의 누적분포함수이다.

(3) t-분포

t-분포는 잘 몰라도 t-검정을 들어 본 사람은 많을 것이다. t-검정은 t-분포를 바탕으로 검정을 하는 것이다. 특히 회귀분석에서 특정 개별 계수의 유의성에 대한 통계적 검정은 거의 대부분 t-분포에 기반한 t-검정이다. t-분포도 역시 자유도라고 불리우는 모수가 1개 있다. 그리고 t-분포의 평균은 0이며 표준정규분포처럼 0을 중심으로 완전히 대칭적이다. 다만, 표준정규분포보다는 꼬리 쪽이 약간 두텁다. 그래서, 정규분포에 비해 평균으로부터 먼 영역에서 실현될 확률이 상대적으로 높다. 하지만, t-분포의 자유도 n이 30보다 크면 t-분포나 표준정규분포가 사실상 유사해진다. t-분포 $t(m)$에서 분산은 $m/(m-2)$인데 m이 커지면 분산은 1(표준정규분포의 분산)로 수렴한다.

<그림 III-9> 서로 다른 자유도의 t-분포와 표준정규분포의 밀도함수




```
. range x -5 5 200
Number of observations (_N) was 0, now 200.

.
. gen y1=100*tden(5,x)

. gen y2=100*tden(20,x)

. gen y3=100*tden(40,x)

. gen y4=100*normalden(x)

.
. twoway ///
> (line y* x) , ///
> xtitle(" ") ytitle("density" "%")  xlabel(#10) ///
> legend( label(1 "t(5)")   ///
>                 label(2 "t(20)") ///
>                 label(3 "t(40)")  ///
>                 label(4 "N(0,1)") ///
>                 pos(6) row(1))
```

마. 균등분포

앞서 이산분포로서 균등분포를 살펴보았다. 서로 다른 이산적 수치 K개로 실현될

수 있는 이산 변수 X가 각 수치마다 $1/K$의 균등한 확률이면 그것이 이산균등분포이다. 연속균등분포는 연속분포로서 균등분포이다. 가장 쉬운 예는 연속체인 실수 구간 $[0,1]$에서 실현되는 변수 U이다. 이러한 변수 U가 균등분포라는 것은 밀도함수 값이 모두 동일하다는 것이다. 즉, $f(u)=1$이 된다. 그렇다면, 구간 $[a,b]$에서의 균등분포를 따르는 변수 y의 밀도함수는 어떻게 생겼을까? $f(y)=1$이 아니다. $f(y)=1/(b-a)$이다. 왜 그런가? 밀도함수 아래 면적이 1, 즉 100%여야 하기 때문이다. 흔히, 변수 X가 구간 $[a,b]$에서 실현되는 균등분포를 따를 때, $X\sim U[a,b]$로 표시하곤 한다.

통계 이야기
정규분포의 발견

서로 다른 데이터에서 얻은 히스토그램을 그려보다 보면, 대부분의 히스토그램이 비슷하게 생겼다는 것을 알 수 있다. 중심에 대해 대칭이며, 중심에서 높이가 가장 높고, 중심에서 멀어질수록 빠르게 낮아진다. 이러한 모양을 가진 분포를 흔히 종모양(bell-shaped)을 가졌다고 표현하곤 한다.

서로 상관이 없어 보이는 확률변수들의 히스토그램 모양이 유사하다는 점에 과학자들은 흥미를 느꼈다. 측정해서 모은 데이터에 대해서 히스토그램을 그리면 대부분 종모양 곡선을 가지게 된다는 것을 발견하게 되었다.

과학자들은 두 가지 질문을 했다.

① 종모양 곡선을 나타내는 수식은 뭐지?
② 왜 종모양 곡선은 항상 나타나게 되는 거지?

첫 번째 질문에 대한 답은 18세기 프랑스 수학자인 드무아브르(de Moivre)가 규명을 했다. 처음에 수학자들은 종모양의 정체는 반원일 것이라고 추론하였습니다. 당시 수학자들은 원이 자연적으로 아름답다고 생각했다. 데이터에서도 이러한 규칙을 발견할 수 있다고 믿었다. 하지만 드무아브르는 종모양이 $1/\sqrt{2\pi\sigma^2}\,\exp\left(-(x-\mu)^2/2\sigma^2\right)$의 그래프임을 밝혔다. 이 함수는 수학적으로 아름답지도 않고 단순하지도 않다.

두 번째 질문에 대한 답은 라플라스와 가우스가 내놓았다. 라플라스는 무엇인가를 합한 것의 히스토그램은 정규분포와 비슷해진다는 것을 증명하였다. 이러한 현상을 이른바 '중심극한정리'로 정리한 것은 가우스이다. 특히 가우스는 측정에서 발생할 수 있는 노이즈의 분포에 대해서 집중하였다. 노이즈는 측정도구의 정확도가 떨어지

거나 측정환경이 일정치 않은 경우에 발생할 수 있다. 노이즈의 발생 원인은 많으며 우리의 데이터에는 이러한 노이즈의 합이 섞여서 포함된다.

정규분포의 'normal'은 정상이라는 뜻을 포함하고 있다. 정규분포가 아니라는 것은 비정상이라는 뜻을 내포하고 있다. 극단적으로는 정규분포를 따르지 않는 데이터에는 문제가 있을 수 있다는 뜻이기도 하다. 하지만 현대의 통계학에서는 정규분포를 따르지 않는 이유들이 속속들이 밝혀지고 있다.

▣ 연습문제

1. 상자 안에 0, 1, 2, 3, 4, 5의 숫자가 쓰여 있는 카드가 각 1개씩 들어 있다고 하자.

(1) 이 상자에서 복원추출로 2개의 카드를 뽑을 때, 2개 카드에 나온 숫자의 평균을 m_2이라고 하자. m_2의 분포를 도출하라.

(2) 이 상자에서 복원추출로 10개의 카드를 뽑을 때, 10개 카드에 나온 숫자의 평균을 m_{10}라고 하자. m_{10}의 분포를 도출하라.

2. 현재 주가가 S_0=100일 때, 1년 후 시점에서의 주가 S_1은 확률변수이고, 이 기간 1년 동안 주가의 상승률 R은 정규분포 $N(0.2, 0.3^2)$을 따른다고 하자.

(1) 주가 상승률 R이 이산복리증감율이면 $S_1=S_0\times(1+R)$이다. 1년 후 주가 S_1의 분포를 구하고, 평균과 표준편차를 계산하라.

(2) 주가 상승률 R이 연속복리증감율이면 $S_1=S_0\times e^R$이다. 1년 후 주가 S_1의 분포를 구하고, 평균과 표준편차를 계산하라.

(3) 주가 S를 기초자산으로 하는 행사가 105인 1년 만기 콜옵션의 만기 시점의 가치는 $c_1=\max(S_1-105, 0)$이다. Stata의 난수 생성자를 이용하여, c_1의 평균을 구하라.

3. 확률변수 X가 로그정규분포 $LN(1,2^2)$을 따른다고 하자.

(1) X의 평균 E(X)을 구하라.

(2) X의 평균 E(X)을 Stata 난수 생성자를 이용하여 구하라.

(3) 확률 $Pr[1\leq X\leq 2]$를 어떤 방법으로든 계산하라.

4. 운전자 1,000만명 중 5%의 운전자가 무보험자라고 하자.

(1) 운전자 1천만명 중 30명의 운전자를 임의로 뽑았을 때, 겨우 1명 혹은 그 이하로 무보험자가 있을 확률을 구하라.

(2) 운전자 1천만명 중 100명의 운전자를 임의로 뽑았을 때, 무보험자가 5명 이하일 확률을 구하라.

5. 내가 현재 가진 돈은 1,000원이다. 동전을 던져 앞이 나오면 10원을 내가 받고, 뒤가

나오면 내가 상대방에게 10원을 주어야 하는 도박을 고려해보자. 동전은 공정한 동전이라고 하자.

(1) 위 내기를 10번 할 때, 다 끝나고 난 시점에서 내 재산이 1,000원일 확률은?

(2) 위 내기를 10번 할 때, 다 끝나고 난 시점에서 내 재산이 1,030원일 확률은?

(3) 위 내기를 10번 할 때, 다 끝나고 난 시점에서 내 재산이 950원 이상일 확률은?

(4) 위 내기를 10번 할 때, 다 끝나고 난 시점에서 내 재산의 평균은?

(5) 위 내기를 100번 할 때, 다 끝나고 난 시점에서 내 재산이 900원 이상일 확률은?

6. 다양한 m과 n을 갖고 F-분포 F(m,n)의 밀도함수를 Stata로 그려라. 다양한 n을 갖고 카이제곱 분포 $\chi^2(n)$의 밀도함수를 Stata로 그려라.