2017-07-08 (3주차-2)

1. 확률변수의 상관관계
2. 다변수 이산확률 변수의 결합/조건부 확률
    - 다변수 이산확률변수
      카테고리값을 갖는 두개의 이산확률변수가 두 개 이상 있는 경우에는 각각의 확률변수에 대한 확률분포이외에도 확률변수 쌍이 갖는 복합적인 확률분포를 살펴보아야함
    - 결합 확률질량함수
      하나의 값이 아닌 두개의 값, 즉 특정한 숫자쌍이 나타는 경우를 생각해보면 단변수 이산확률변수에서와 마찬가지로 하나의 숫자쌍으로만 이루어진 사건에 대한 확률만 알고 있으면 임의의 사건에 대해서도 확률을 계산할 수 있음으로 하나의 숫자쌍에 대해서 확률을 알려주는 확률질량함수만 있으면 전체 확률분포를 알 수 있음. 이러한 확률질량함수를 결합확률질량함수라 함.
    - 주변 확률질량함수
      두 확률변수 중 하나의 확률변수 값에 대해서만 확률분포를 표시한 함수
      결합 확률질량함수에서 구하려면 전체 확률법칙에 의해 다른 변수가 가질수 있는 모든 값의 확률질량함수를 합한 확률이됨
    - 조건부 확률질량함수
      다변수 확률변수 중 하나의 값이 특정값으로 고정되어 상수가 되어버린 경우, 나머지 변수에 대한 확률질량함수를 말함
      조건부 확률질량함수의 모양은 결합질량함수에서 하나의 확률변수가 고정된 결합질량함수의 단면과 같음, 다만 조건부 확률질량함수의 합은 1

3. 다변수 연속확률 변수의 결합/조건부 확률
   연속확률변수에서는 이산확률변수와 같이 atom을 이용한 확률의 정의가 불가능함으로 단변수 연속확률변수처럼 CDF을 먼저 정의하고 이를 미분한 확률밀도함수를 정의하는 방법을 사용
    - 결합 누적확률분포함수
      두 확률 변수X, Y에 대한 결합 누적확률분포함수 FXY(x,y)는 다음과 같이 정의한다.
      F X Y (x ,y)=P({X <x }∩{Y <y})=P(X <x ,Y <y)
      만약 두 구간을 끝을 나타내는 두 독립변수 x,y 중 하나가 무한일 경우 남은 하나의 변수에 대한 누적확률분포함수로 줄어들고 이를 주변 누적확률분포라 함
      FX(x) = FXY(x,inf)
    - 결합 확률밀도함수
      결합 누적확률분포함수를 미분하여 얻을 수 있다. 단변수 확률변수와 마찬가지로 구간에 대해 정적분을 하게 되면 확률을 얻을 수 있음

    - 주변 확률밀도 함수
      주변 확률밀도함수(marginal probability density function)는 결합 확률밀도함수를 특정한 하나의 변수에 대해 가중평균한 값을 말한다. 따라서 결합 확률밀도함수를 하나의 확률변수에 대해서만 적분하여 구한다.

    - 조건부 확률밀도 함수
      조건부 확률밀도함수(conditional probability density function)는 다변수 확률 변수 중 하나의 값이 특정 값이라는 사실이 알려진 경우, 이러한 조건(가정)에 의해 변화한 나머지 확률변수에 대한 확률밀도함수를 말한다.
4. 확률밀도함수의 독립
   만약 두 확률변수 X, Y의 결합 확률밀도함수(joint pdf)가 주변 확률밀도함수의 곱으로 나타나면 두 확률변수가 서로 독립이라고 함
   f X Y (x ,y)=f X (x)f Y (y)
   독입인 경우에는 주변 확률밀도 함수만으로 결합확률밀도 함수를 얻을 수 있음
    - 반복시행: 같은 확률변수에서 표본데이터를 취하는 경우에는 독립인 두개의 확률변수에서 나온 표본이라 볼수 있음
    - 독립확률변수의 기댓값: E[XY] = E[X]*E[Y]
    - 독립확률변수의 분산: Var[X+Y] = Var[X]+Var[Y]
    - 조건부 확률분포
      독립인 두 확률변수 X ,Y X,Y의 조건부 확률밀도함수는 주변 확률밀도함수와 같다
5. 공분산과 상관계수
   다변수 확률변수도 단변수처럼 평균, 분산과 같은 대표값을 가질 수 있다. 그 중 가장 중요한 것이 자료간의 상관관계를 나타내는 공분산과 상관계수이다. 공분산과 상관계수도 샘플 자료 집합에 대해 정의되는 샘플 공분산, 샘플 상관계수와 확률 변수에 대해 정의되는 공분산, 상관계수가 있다
    - 샘플 공분산
      샘플 공분산(sample covariance)은 다음과 같이 정의된다. 여기에서 x i xi와 y i yi는 각각 i i번째의 x x 자료와 y y자료의 값을 가리키고, m x mx와 m y my는 x x자료와 y y자료의 샘플 평균을 가리킨다.
      s 2 x y =1 N ∑ i=1 N (x i m x)(y i m y) sxy2=1N∑i=1N(ximx)(yimy)
      샘플 분산과 마찬가지로 샘플 공분산도 자료가 평균값으로부터 얼마나 떨어져 있는지를 나타낸것이다. 공분산은 평균값 위치와 샘플 위치를 연결하는 사각형의 면적을 사용한다. 다만 공분산의 경우에는 자료의 위치에 따라 이 값의 부호가 달라진다.
    - 샘플 상관계수

      샘플 공분산은 평균을 중심으로 각 자료들이 어떻게 분포되어 있는지 기와 방향성을 같이 보여준다. 그런데 분포의 크기는 공분산이 아닌 분산만으로도 알 수 있기 때문에 대부분의 경우 자료 분포의 방향성만 분리하여 보는 것이 유용한다. 이 때 필요한 것이 샘플 상관계수(sample correlation coefficient)이다.

      샘플 상관계수는 다음과 같이 공분산을 각각의 샘플 표준편차값으로 나누어 정규화(normalize)하여 정의한다.
6. 확률변수의 공분산과 상관계수

   두 확률변수 X X와 Y Y의 공분산은 기댓값 연산자를 사용하여 다음과 같이 정의된다.
   두 확률변수 X와 Y의 상관계수 p는 다음과 같은 특성
    - -1<= p <= 1
    - p = 1 : 완전선형 상관관계
    - p = 0 : 무상관(독립과 다름)
    - p = -1: 완전선형 반상관관계

상관계수로 분포의 형상을 추측할때 주의할점은 개별자료가 상관계수에 미치는 영향력이 각각 다르다는 점

7. 다변수 확률변수의 공분산
   다변수 확률변수의 경우 데이터가 벡터인 경우 샘플 공분산 행렬을 정의할 수 있다.

8. 다변수 가우시안 정규분포
9. 검정과 모수 추정의 의미
   데이터 분석의 첫번째 가정은 분석하고자 하는 데이터가 어떤 확률변수(random variable)로 부터 실현된 표본(sample)이다
   즉, 우리가 관심이 있는 것은 실현된 데이터가 아니라 데이터를 만들어내고 있는 확률변수라는 뜻, 데이터는 단지 이 확률변수를 알아내기 위한 참고자료임

   확률변수는 분포모형(distribution model)과 모수(parameter)를 가진다. 따라서 확률 변수를 안다는 것은 다음과 같은 질문에 답하는 것
   - 해당 데이터가 특정한 분포모형, 예를 들면 가우시안 정규분포로부터 생성된 것인가?
   - 만약 그렇다면 그 정규분포의 기댓값 모수 u와 분산 s^2이 특정한 값을 가지고 있는가? 예를 들면 u =0인가 아닌가?
   - 정규분포 모수 u가 0이 아니라면 구체적으로 어떤값을 갖는가?

   이러한 질문에 답을 하는 행위을 검정(test) 또는 모수 추정(parameter estimation)이라고 한다. 모수 추정은 간단히 추정(estimation)이라고 하기도 한다.
   - 첫번째 질문은 확률 변수의 분포에 대한 가설(hypothesis)이 맞는지 틀리는지를 확인하는 확률 변수의 분포 검정(distribution test)라고 한다.
     이 질문은 확률 변수의 분포가 정규 분포(normal distribution)이라는 것을 가설로 놓고 있다. 이러한 가설을 검정하는 것을 특별히 정규성 검정(normality test)이라고 하며 데이터 분석에서 가장 많이 사용되는 검정의 하나이다.
   - 두번째 질문은 확률 변수의 분포가 어떤 모형을 따르는지는 이미 정해져 있는 상태에서 확률 밀도 함수(pdf)의 계수(coefficient) 즉, 모수(parameter)가 특정한 값을 가지는지 혹은 특정한 값과 비교하여 큰지 작은지를 확인하는 과정이다.
     예를 들어 정규 분포의 기댓값 모수가 0인지 아닌지 확인하고 싶다면 "정규 분포의 기댓값 모수가 0이다."라는 가설을 증명하고 싶은 것이다. 이러한 가설을 검정하는 것을 모수 검정(parameter test)이라고 한다.
   - 마지막 질문은 모수가 실제로 어떤 숫자를 가질 확률이 가장 높은지를 알아내는 작업으로 이러한 과정을 모수 추정(parameter estimation) 또는 추정(estimation)이라고 한다. 모수 추정 방법은 어려가지가 있다.
     MSE(Maximum Squred Error) 방법, MLE(Maximum Likelihood Estimation) 방법 등은 가장 확률이 높은 숫자 하나를 결정하는 방법의 하나이며 베이지안 추정법(Bayesian Estimation)은 가능한 모든 값에 대해 이 값들이 진짜 모수가 될 확률을 모두 계산하여 분포로 표시하는 방법이다.

10. 검정과 유의확률
    검정(testing)은 데이터 뒤에 숨어있는 확률 변수의 분포와 모수에 대한 가설의 진위를 정량적(quantitatively)으로 증명하는 작업을 말한다. 예를 들어 다음과 같은 문제는 검정 방법론을 사용하여 접근할 수 있다.
    - 문제1 : 어떤 동전을 15번 던졌더니 12번이 앞면이 나왔다. 이 동전은 휘어지지 않은 공정한 동전(fair coin)인가?
      -> 동전이 공정한 동전이라고 주정하는 것은 그 뒤의 베르누이 확률 분포의 모수 θ의 값이 0.5 이라고 주장하는 것과 같다.
    - 문제2 : 어떤 트레이더의 일주일 수익률은 다음과 같다.
      -2.5%, -5%, 4.3%, -3.7% -5.6%
      이 트레이더는 돈을 벌어다 줄 사람인가, 아니면 돈을 잃을 사람인가?
      -> 트레이더가 장기적으로 돈을 벌어다 줄 것이라고 주장하는 것은 그 뒤의 정규 분포의 기댓값 모수 μ가 0보다 크거나 같다고 주장하는 것이다.

      위의 문제는 다음과 같이 해결할 수 있음
      a. 데이터가 어떤 고정된(fixed) 확률 분포를 가지는 확률 변수라고 가정한다. 동전은 베르누이 분포를 따르는 확률 변수의 표본이며 트레이더의 수익률은 정규 분포를 따르는 확률 변수의 표본이라고 가정한다.
      b. 이 확률 분포의 모수값이 특정한 값을 가지는지 혹은 특정한 값보다 크거나 같은지 알고자 한다.
      c. 모수 값이 이러한 주장을 따른다고 가정하면 실제로 현실에 나타난 데이터가 나올 확률을 계산할 수 있다. 동전의 경우에는 공정한 동전임에도 불구하고 15번 중 12번이나 앞면이 나올 확률을 계산할 수 있으며 트레이더의 경우에는 정규 분포에서 해당 데이터가 나올 확률을 계산할 수 있다.
      d. 이렇게 구한 확률의 값이 판단자가 정한 특정한 기준에 미치지 못한다면 이러한 주장이 틀렸다고 생각할 수 밖에 없다. 반대로 값이 기준보다 높다면 그 주장이 틀렸다고 판단할 증거가 부족한 것이다.

    가설(Hypothesis)
    이렇게 확률 분포에 대한 어떤 주장을 가설이라고 하며 H라 표기함. 이 가설을 증명하는 행위를 통계적 가설 검정(statistical hypothesis testing) 줄여서 검정(testing)이라고 한다. 특히 확률 분포의 모수 값이 특정한 값을 가진다는 주장을 모수 검정(parameter testing)이라고 한다.
    일반적으로 쓰이는 가설은 모수가 0이라는 가설, 이 가설은 regression에서 회귀계수 값이 0이면 target에 독립변수(feature)가 아무런 영향을 주지 않는다는 뜻

    검정방법론
    가설 증명, 즉 검정의 기본적인 논리는 다음과 같다.
    a. 만약 가설이 맞다면 즉, 모수 값이 특정한 조건을 만족한다면 해당 확률 변수로부터 만들어진 표본(sample) 데이터들은 어떤 규칙을 따르게 된다.
    b. 해당 규칙에 따라 표본 데이터 집합에서 어떤 숫자를 계산하면 계산된 숫자는 특정한 확률 분포를 따르게 된다. 이 숫자를 검정 통계치(test statistics)라고 하며 확률 분포를 검정 통계 분포(test statistics distribution)라고 한다.

       검정 통계 분포의 종류 및 모수의 값은 처음에 정한 가설에 의해 결정된다.

       이렇게 검정 통계 분포를 결정하는 최초의 가설을 귀무 가설(Null hypothesis)이라고 한다.
    c. 데이터에 의해서 실제로 계산된 숫자, 즉, 검정 통계치가 해당 검정 통계 분포에서 나올 수 있는 확률을 계산한다. 이를 유의 확률(p-value)라고 한다.
    d. 만약 유의 확률이 미리 정한 특정한 기준값보다 작은 경우를 생각하자. 이 기준값을 유의 수준(significance level)이라고 하는 데 보통 1% 혹은 5% 정도의 작은 값을 지정한다.
       유의 확률이 유의 수준으로 정한 값(예 1%)보다도 작다는 말은 해당 검정 통계 분포에서 이 검정 통계치가 나올 수 있는 확률이 아주 작다는 의미이므로 가장 근본이 되는 가설 즉, 귀무 가설이 틀렸다는 의미이다.

따라서 이 경우에는 귀무 가설을 기각(reject)한다.

e. 만약 유의 확률이 유의 수준보다 크다면 해당 검정 통계 분포에서 이 검정 통계치가 나오는 것이 불가능하지만은 않다는 의미이므로 귀무 가설을 기각할 수 없다. 따라서 이 경우에는 귀무 가설을 채택(accept)한다.

귀무가설과 대립가설

귀무가설이 기각되면 채택될수 있는 가설들을 대립가설이라함

검정 통계량

검정을 하려면 즉, 귀무가설이 맞거나 틀린것을 증명하려면 어떤 증거가 있어야한다. 이 증거에 해당하는 수치를 검정 통계량(test statistics)라 함

검정통계량은 표본자료에서 계산된 함수값임으로 표본처럼 확률적이다. 즉 경우에 따라서 표본이 달라질수 있는 것처럼 검정 통계량도 달라진다. 검정통계량 t도 T라는 확률변수의 표본으로 볼 수 있음

어떤 함수가 검정 통계량이 되려면 귀무 가설이 사실일 경우 표본에서 계산된 검정 통계량이 따르는 검정 통계량 확률 변수 T의 확률 분포를 귀무 가설로부터 알 수 있어야만 한다.

데이터 분석에서는 어떤 귀무 가설을 만족하는 표본을 입력 변수로 놓고 특정한 함수로 계산한 검정 통계량이 특정한 분포를 따른다는 것을 수학적인 증명을 통해 보이는 것이 일반적이다.

통계학자들의 중요한 업적 중의 하나가 특정한 귀무 가설에 대해 어떤 검정 통계량 함수가 어떤 검정 통계량 분포를 따른 다는 것을 증명해 준 것이다.

검정 통계량의 예

1. 베르누이 분포 확률 변수
2. 카테고리 분포 확률 변수
3. 분산 s^2 값을 알고 있는 정규분포 확률 변수
4. 분산 s^2 값을 모르고 있는 정규분포 확률 변수

유의확률 p-value

귀무 가설이 사실이라는 가정하에 검정 통계량이 따르는 검정 통계량 분포를 알고 있다면 실제 데이터에서 계산한 검정 통계량 숫자가 분포에서 어느 부분쯤에 위치해 있는지를 알 수 있다. 이 위치를 나타내는 값이 바로 유의 확률(p-value) 이다.

검정 통계량의 유의 확률은 검정 통계량 숫자보다 더 희귀한(rare) 값이면서 대립 가설을 따르는 값이 나올 수 있는 확률을 말한다. 이 확률은 검정 통계 확률 분포 밀도 함수(pdf)에서 양 끝의 꼬리(tail)부분에 해당하는 영역의 면적으로 계산한다. 실제로는 누적 확률 분포 함수를 사용한다.

유의수준과 기각역

계산된 유의 확률 값에 대해 귀무 가설을 기각하는지 채택하는지를 결정할 수 있는 기준 값을 유의 수준(level of significance)라고 한다. 일반적으로 사용되는 유의 수준은 1%, 5%, 10% 등

검정 통계량이 나오면 확률 밀도 함수(또는 누적 확률 함수)를 사용하여 유의 확률을 계산할 수 있는 것처럼 반대로 특정한 유의 확률 값에 대해 해당하는 검정 통계량을 계산할 수도 있다. 유의 수준에 대해 계산된 검정 통계량을 기각역(critical value)라고 한다.

기각역 값을 알고 있다면 유의 확률을 유의 수준과 비교하는 것이 아니라 검정 통계량을 직접 기각역과 비교하여 기각/채택 여부를 판단할 수도 있다.

검정의 예

어떤 동전을 15번 던졌더니 12번이 앞면이 나왔다. 이 동전은 휘어지지않은 공정한 동전인가?

-> 이 문제는 베르누이 확률변수의 모수 검정문제로 생각가능, 귀무가설은 theta = 0.5

-> 검정통계량 t 는 15번던져서 앞면이 12번나오고 자유도가 15인 이항분포를 따름

-> 이경우의 유의확률은 1.76%

-> 이 값은 5% 보다는 작고 1% 보다는 크기 때문에 유의 수준이 5% 라면 기각할 수 있으며(즉 공정한 동전이 아니라고 말할 수 있다.) 유의 수준이 1% 라면 기각할 수 없다.(즉, 공정한 동전이 아니라고 말할 수 없다.)