7-1 이론적 확률분포(이산형) 1

제주권역
대학 e-러닝 지원센터

7-1-1. 확률분포와 모수

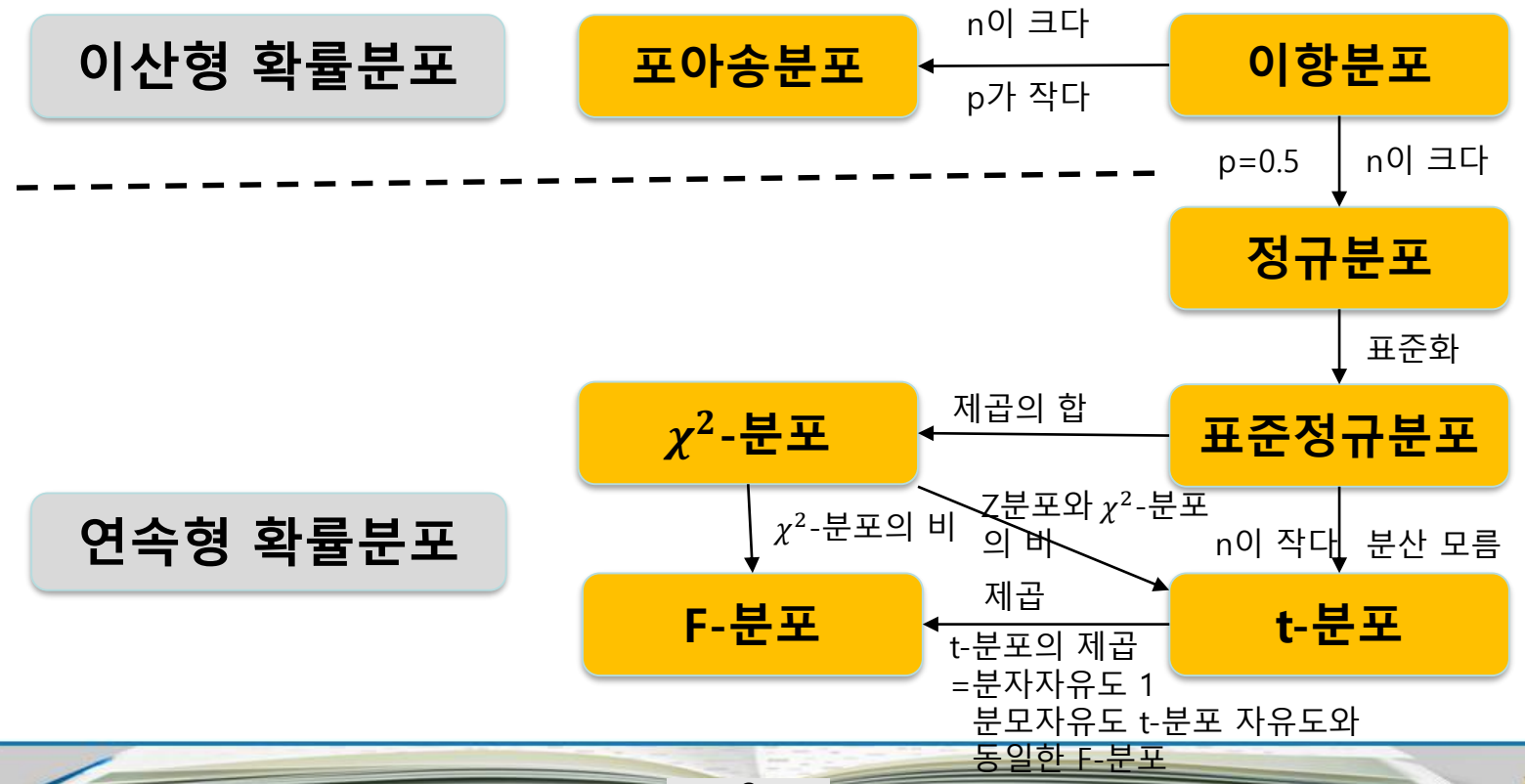
1.확률분포

통계분석에서 자료를 수집하고 그 수집된 자료로부터 어떤 정보를 얻고자 하는 경우에는 항상 수집된 자료가 특정한 확률분포를 따른다고 가정한다.

확률분포

- **이산형 확률분포**
- **연속형 확률분포**

이론적 확률분포의 관계


이산형 확률분포
포아송분포
n이 크다
p가 작다
이항분포
p=0.5
n이 크다
정규분포
표준화
연속형 확률분포
χ^2-분포
제곱의 합
표준정규분포
χ^2-분포의 비
Z분포와 χ^2-분포의 비
n이 작다
분산 모름
F-분포
제곱
t-분포
t-분포의 제곱
=분자자유도 1
분모자유도 t-분포 자유도와
동일한 F-분포

2. 모수(parameter)

모수(parameter)

모든 확률분포는 그 분포의 모양을 결정하는 값

모수는 확률분포의 특성을 나타내는 값으로 모든 확률분포는 모수에 의하여 구체적인 모양이 결정되어진다.

7-1-2. 포아송분포

1.포아송분포

포아송분포는 주어진 단위시간, 거리, 영역 등에서 어떤 사건이 발생되는 횟수를 측정하는 확률변수로 다음과 같은 예를 생각할 수 있다.

① 특정지역에서 제한된 시간 내에 발생되는 교통사고 수의 분포

② 타이프를 치는 데 있어서 페이지당 오타 수의 분포

③ 특정시간에 고속도로 톨게이트를 지나는 외제차량 수의 분포

④ 자동생산라인에서 특정시간에 발생되는 불량품 수의 분포

단위구간 내에서 어떤 사건이 평균 μ 회 발생한다고 한다.
확률변수 X 를 사건의 발생횟수라고 할 때, X ~ Poisson (μ)로 표현하며
사건이 k 번 발생될 확률은 다음과 같다.

$$P_r(x=k)=\frac{\mu^k}{k!}e^{-\mu},\quad 0,1,2,\ldots$$

여기에서는 e 는 자연대수(log)의 밑수로 e = 2.71828... 이다.

예

최근 올림픽도로에서는 하루 평균 5건의 교통사고가 발생한다.
교통사고의 발생횟수가 포아송분포를 따른다고 할 때, 다음의 확률을 계산하라.

① 어느 날 교통사고가 전혀 일어나지 않을 확률은 얼마인가?
② 어느 날 교통사고가 3번 이상 일어날 확률은 얼마인가?

풀이 확률변수 X가 하루에 발생하는 교통사고의 횟수를 나타낸다고 하면
하루에 발생하는 교통사고의 횟수가 평균 5회이므로 **X ~ Poisson (5)이고**

$$P_r(x=k)=\frac{5^k}{k!}e^{-5},\quad 0,1,2,\ldots$$

① 어느 날 교통사고가 전혀 일어나지 않을 확률은

$$P_r(x=0)=\frac{5^0}{0!}e^{-5}=e^{-5}\approx 0.00674$$

② 어느 날 교통사고가 3번 이상 일어날 확률

$$P_r(x\geq 3)=1-P_r(x<3)=1-\{P_r(x=0)+P_r(x=1)+P_r(x=2)\}$$
$$=1-\{0.007+0.033+0.085\}=1-0.1247=0.8753$$

2.포아송 분포의 평균과 분산

확률변수 X 가 평균 μ 인 포아송분포를 따를 때, X 의 평균과 분산은 다음과 같이 동일하다.

① 평균 : $E(X)=\mu$

② 분산 : $Var(X)=\mu$

3. 포아송 분포의 확률계산

아래는 포아송분포에서 각각의 평균값 μ에 대한 확률변수 X의 c까지의 누적확률값 $P(X \le c)=\sum_{k=0}^{c}\frac{\mu^k}{k!}e^{-\mu}$ 을 계산해 나타낸 표이다.

표 포아송분포표(누적)

c \ μ	4.4	4.6	4.8	5.0	5.2	5.4	5.6	5.8	6.0	6.2	6.4
0	.012	.010	.008	.007	.006	.005	.004	.003	.002	.002	.002
1	.066	.056	.048	.040	.034	.029	.024	.021	.017	.015	.012
2	.185	.163	.143	.125	.109	.095	.082	.072	.062	.054	.046
3	.359	.326	.294	.265	.238	.213	.191	.170	.151	.134	.119
4	.551	.513	.476	.440	.406	.373	.342	.313	.285	.259	.235
5	.720	.686	.651	.616	.581	.546	.512	.478	.446	.414	.384
6	.844	.818	.791	.762	.732	.702	.670	.638	.606	.574	.542
7	.921	.905	.887	.867	.845	.822	.797	.771	.744	.716	.687
8	.964	.955	.944	.932	.918	.903	.886	.867	.847	.826	.803
9	.985	.980	.975	.968	.960	.951	.941	.929	.916	.902	.886
10	.994	.992	.990	.986	.982	.977	.972	.965	.957	.949	.939
11	.998	.997	.996	.995	.993	.990	.988	.984	.980	.975	.969
12	.999	.999	.999	.998	.997	.996	.995	.993	.991	.989	.986
13	1.000	1.000	1.000	.999	.999	.999	.998	.997	.996	.995	.994
14				1.000	1.000	1.000	.999	.999	.999	.998	.997
15							1.000	1.000	.999	.999	.999
16									1.000	1.000	1.000

표 포아송분포표(누적)

```r
poi11<-rep(NA,15)
poi12<-rep(NA,15)
poi13<-rep(NA,15)

poi11[1]<-ppois(0, 5.0)
poi12[1]<-ppois(0, 5.2)
poi13[1]<-ppois(0, 5.4)

for(i in 2:15) {
  poi11[i]<-ppois(i-1, 5.0)
}

for(i in 2:15) {
  poi12[i]<-ppois(i-1, 5.2)
}

for(i in 2:15) {
  poi13[i]<-ppois(i-1, 5.4)
}

(poi<-cbind(poi11,poi12, poi13))
```

```
             poi11       poi12       poi13
 [1,] 0.006737947 0.005516564 0.004516581
 [2,] 0.040427682 0.034202699 0.028906118
 [3,] 0.124652019 0.108786650 0.094757868
 [4,] 0.265025915 0.238065499 0.213291018
 [5,] 0.440493285 0.406128002 0.373310771
 [6,] 0.615960655 0.580913005 0.546132104
 [7,] 0.762183463 0.732393340 0.701671304
 [8,] 0.866628326 0.844921590 0.821658687
 [9,] 0.931906365 0.918064952 0.902650170
[10,] 0.968171943 0.960325561 0.951245060
[11,] 0.986304731 0.982301078 0.977486301
[12,] 0.994546908 0.992689504 0.990368364
[13,] 0.997981148 0.997191156 0.996165293
[14,] 0.999302010 0.998991816 0.998573248
[15,] 0.999773746 0.999660633 0.999502030
```

앞의 예를 위의 표를 이용하여 계산해 보면 평균이 5.0의 열에서

$$P_r(x=0)=0.007$$

$$P_r(x\geq 3)=1-P_r(x<3)=1-P_r(x\leq 2)$$
$$=1-0.125=0.875$$

예

어떤 자동차 보험회사에서 조사한 결과에 의하면 한 보험 가입자에게
1년 동안에 발생하는 자동차 사고는 대체로 포아송분포에 따르며,
보험 가입자 1인당 평균 사고 수는 0.25인 것으로 나타났다.
임의로 추출된 보험 가입자가 내년에 2회의 사고를 당할 확률은 얼마인가?

풀이

확률변수 X가 1년 동안 보험 가입자에게 발생하는 자동차 사고 수를 나타낸다면 다음과 같다.

$$p_X(2) = \frac{\mu^x}{x!} e^{-\mu} = \frac{0.25^2}{2!} e^{-0.25} = 0.024338$$

표 누적포아송확률분포표

$$P(X \le c) = \sum_{k=0}^{c} \frac{\mu^k e^{-\mu}}{k!}, \ \mu : \text{기댓값}$$

c \ μ	.02	.04	.06	.08	.10	.15	.20	.25	.30	.35	.40
0	.980	.961	.942	.923	.905	.861	.819	.779	.741	.705	.670
1	1.000	.999	.998	.997	.995	.990	.982	.974	.963	.951	.938
2		1.000	1.000	1.000	1.000	.999	.999	.998	.996	.994	.992
3						1.000	1.000	1.000	1.000	1.000	.999
4											1.000

예 앞의 자동차 보험의 문제에서 임의로 추출된 보험가입자가 향후 20년 동안 2회의 사고를 당할 확률을 계산하라.

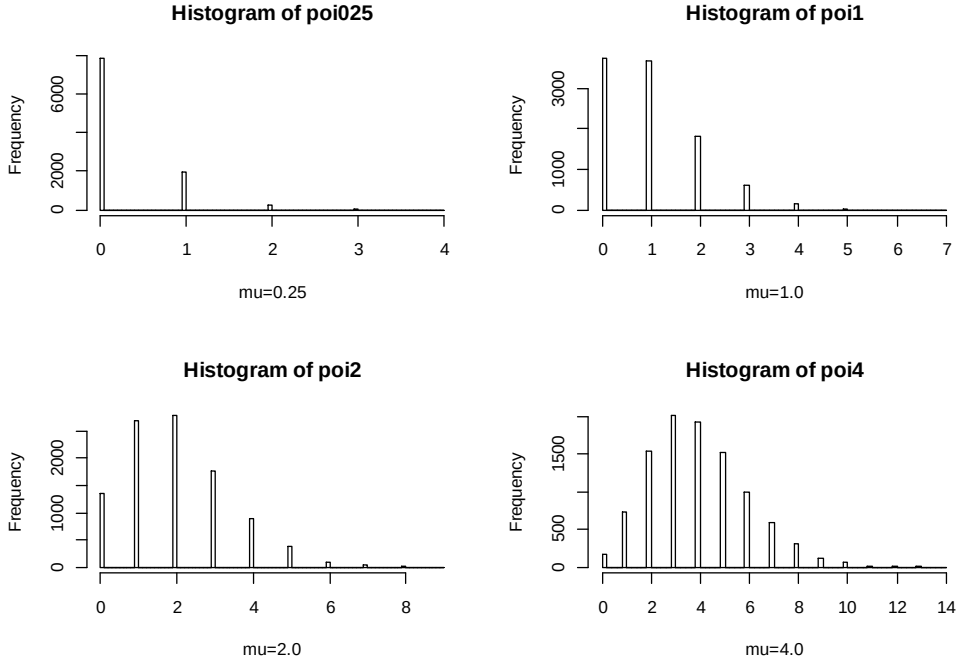
풀이

포아송분포에서 평균은 측정 단위구간당 평균이며 이 문제에서 주의할 점은 단위시간이 20년이라는 것이다.
즉, 1년의 평균 사고 수가 0.25회이므로 20년의 평균 사고 수는 5회가 됨에 유의해야 한다.
따라서 앞에서 주어진 포아송분포표를 이용하면,

$$P_r(x=2) = P_r(x \le 2) - P_r(x \le 1) = 0.125 - 0.04 = 0.085$$

포아송분포의 모양은 평균이 작을 때는 좌우비대칭이나 평균이 증가함에 따라 평균을 중심으로 좌우대칭의 모양으로 변한다(정규분포에 가까워지고, 분산이 커짐)

```
1  set.seed(12345)
2
3  n<-10000;
4
5  poi025<-rpois(n, 0.25)
6  poi1<-rpois(n, 1)
7  poi2<-rpois(n, 2)
8  poi4<-rpois(n, 4)
9
10 par(mfrow=c(2,2))
11
12 hist(poi025, breaks=100, xlab="mu=0.25")
13 hist(poi1, breaks=100, xlab="mu=1.0")
14 hist(poi2, breaks=100, xlab="mu=2.0")
15 hist(poi4, breaks=100, xlab="mu=4.0")
16
```



```r
par(mfrow=c(2,2))

lambda_list<-c(3, 5, 10, 15) # 평균
x_list<-30 # 발생횟수를 1부터 x축에 보여줄 최대값

for (i in 1:length(lambda_list)) {
    p_x<-dpois(x=0:x_list,lambda_list[i])
    plot(x=0:x_list, p_x, xlab="x", ylab="P(X=x)", ylim=c(0, 0.25),
         xlim=c(0,x_list), main=paste("lambda=", lambda_list[i]))
    x_seq<-seq(0,x_list,1)
    lines(x_seq, p_x, type="h", col="blue")
}

```