일반선형모델

6.1 여러가지 확률분포

`정규분포 외의 확률분포를 사용한다.` 일반선형모델과 관계가 깊은 이항분포나 푸아송 분포 를 소개한다.

6.1.1 용어 설명

이항확률변수:

- 2개의 값만 가지는 확률변수
- 예) (앞/뒤), (있다/없다)

베르누이 시행:

- 2종류의 결과 중 하나만 발생시키는 시행
- 예) 동전을 한번 던져서 앞/뒤 나오는거 기록하는 시행

성공확률:

- 2종류의 결과 중 어느 한쪽의 결과를 얻을 확률
- 성공확률은 [0,1]의 범위만 취한다.
- 성공/실패
- (여기서 성공!=긍정적 아시죠?)

베르누이 분포:

- 한 번의 베르누이 시행이 일어날 때 이산확률변수가 따르는 확률분포
- 예) 동전을 1회 던져서 앞/뒤 기록할때, 앞(1) 뒤(0)이라고 할 때, X는 이항확률변수,p는 성공확률(앞이 나올 확률)입니다.
- P(X=1)=p, P(X=0)=1-p

6.1.2 이항분포

이항분포는 성공확률이 p면서 N회의 독립 베르누이 시행을 했을 때 성공한 횟수m이 따르는 이산형 확률분포입니다.

모수(파라미터)는 성공확률p, 시행횟수N

확률변수m의 기댓값은 Np, 분산은 Np(1-p)

확률질량함수를 사용하면, 앞이 나올 확률p인 동전을N번 던졌을 때 앞이m번 나올 확률을 계산할 수 있습니다. 그러면 이산형 확률분포니까 2번 나올 확률 *or* 3번 나올 확률을 *덧셈*으로 구한다.

확률변수 X가 실험횟수가 n이고 매 시행의 성공률이 p인 이항분포를 따르는 것을 X~B(n,p)라고 표현합니다.
확률분포는 다음과 같이 구합니다.

$$f(x) = P(X=x) = \binom{n}{x} p^x q^{n-x}, x = 0,1,2,\ldots,n$$

매 시행의 성공률이 p로 일정한 이항실험에서 (q=1-p)
확률변수 X : n번 시행하는 실험이라고 가정할 때에 평균과 분산은 다음과 같습니다.

평균 = n*p
분산 = n*p*q

6.1.3 이항분포 사용법

"성공확률p가 어떻게 변화하는가"라는 시점으로 이용한다!

시행횟수N과 성공횟수m은 데이터로 주어지는 경우가 많습니다.

예) **흡연 여부**에 따라 암에 걸릴 확률이 어떻게 변화하는지, **가격을 변경**하면 상품의 구입률이 변화하는지, **공부시간을 바꾸면** 시험의 합격률의 변화하는지 들을 조사할 때 사용됩니다.

#베르누이분포
아이템 강화는 베르누이 시행이라고 볼 수 있습니다.
게임 케릭터의 장비를 강화하면 일정 확률로 강화가 성공하거나 실패합니다. 시행의 결과는 오직 강화 성공과 실패 두가지 밖에 존재하지 않으니 베르누이 시행이라고 볼 수 있겠죠?

강화 성공확률은 게임개발자는 알고 있습니다. 성공 확률이 p라고 한다면 장비 아이템 강화 시행이란 성공확률이 p인 베르누이 시행이됩니다.

#이항분포
그런데 게임에선 항상 문제가 존재하죠. 바로 강화를 딱 한번만 지르지 않는다는 문제점이 존재합니다.
즉 성공확률이 p인 베르누이 시행을 n번 반복하게됩니다.

정리하자면 성공확률이 p인 베르누이 시행을 n번 시도했을 때 k번 성공하는 사건은 얼마나될까?? 하고
즉 "10% 장비강화 주문서를 사용해 7장 사용해서 3번 성공할 확률은 얼마나 될까?"를 궁금해할 수 있습니다.
이러한 경우 이항분포를 활용하게 됩니다.

6.1.4 이항분포의 확률질량함수

표기법

$$\mathrm{Bin}(m \mid N, p) = {}_N C_m \cdot p^m \cdot (1-p)^{N-m}$$

$${}_N C_m = \frac{N!}{(N-m)! \cdot m!}$$

6.1.7 푸아송 분포

푸아송 분포는 1개 또는 2개, 1회 또는 2회 등의 **카운트 데이터**가 따르는 이산형 확률분포

`카운트 데이터는 0이상의 정수라는 특징을 갖는다.` 정규 분포는 -∞~+∞의 실수를 취한다.

- 모수는 강도(일이 일어날 횟수에 대한 기댓값, 발생 강도)λ 밖에 없다.

- 푸아송 분포를 따르는 확률변수는 **기댓값과 분산도** λ값과 같습니다.
- 예) 낚싯대를 바꾸면 잡아 올리는 물고기수가 달라지는지
- 예) 주변 환경에 따라 조사 구획내의 생물의 개체수가 변하는지
- 예)날씨에 따라 상품의 판매 개수가 얼마나 달라지는지 들을 알아볼 때

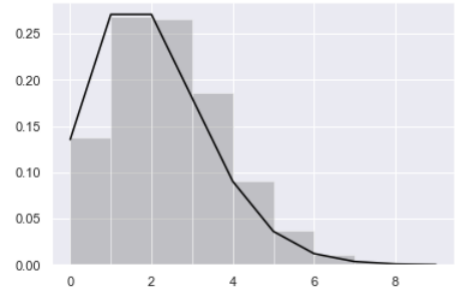
$$\text{Pois}(x \mid \lambda) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

- 푸아송 분포는 p->0, N->∞ 라는 조건에서 이항분포가 Np=λ인 결과임
- 푸아송 분포는 성공확률이 한없이0에 가깝지만 시행횟수가 무한히 많은 이항분포

```
poisson=sp.stats.poisson(mu=2)
#강도가 2인 푸아송 분포
np.random.seed(1)
rvs_poisson=poisson.rvs(size=10000)
#난수
pmf_poisson=poisson.pmf(k=m)
#확률질량함수
sns.distplot(rvs_poisson, bins = m, kde = False, norm_hist = True, color='gray')
plt.plot(m, pmf_poisson, color='black')
#난수의 히스토그램과 확률질량함수를 겹친거
```

```
C:₩Users₩Charyeong Heo₩anaconda3₩lib₩site-packages₩seaborn₩distributions.py:2551:
ion and will be removed in a future version. Please adapt your code to use either
flexibility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```

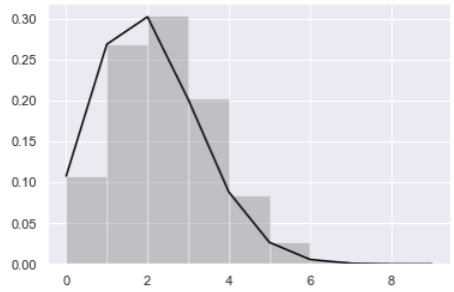
```
[<matplotlib.lines.Line2D at 0x14813a5b3d0>]
```




```
binomial = sp.stats.binom(n=10,p=0.2)
np.random.seed(1)
rvs_binomial=binomial.rvs(size=10000)
#난수
m=np.arange(0,10,1)
pmf_binomial=binomial.pmf(k=m)
#확률밀도함수
sns.distplot(rvs_binomial, bins=m,kde=False,norm_hist=True,color='gray')
#난수 히스토그램과 확률질량함수 겹친거
plt.plot(m,pmf_binomial,color='black')
```

```
C:₩Users₩Charyeong Heo₩anaconda3₩lib₩site-packages₩seaborn₩distributions.p
ion and will be removed in a future version. Please adapt your code to use
flexibility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```

```
[<matplotlib.lines.Line2D at 0x14813965730>]
```



확률질량함수

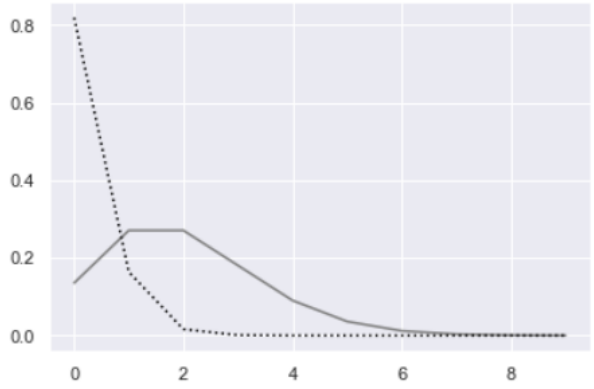
- 시행횟수:100000000
- 성공확률: 0.00000002

푸아송분포의 확률질량함수

- 강도: 2

```
N=10000000
p=0.00000002
binomial_2=sp.stats.binom(n=N,p=p)
#이항분포
pmf_binomial_2=binomial_2.pmf(k=m)
#확률질량함수
plt.plot(m, pmf_poisson, color='gray')
plt.plot(m, pmf_binomial_2, color='black', linestyle='dotted')
#확률질량 그래프
```

```
[<matplotlib.lines.Line2D at 0x14813acb910>]
```



6.1.12 그 외의 확률분포

Distribution	Domain	$\mu = E[Y\|x]$	$v(\mu)$	$\theta(\mu)$	$b(\theta)$	ϕ
Binomial $B(n,p)$	$0,1,\ldots,n$	np	$\mu - \frac{\mu^2}{n}$	$\log \frac{p}{1-p}$	$n\log(1+e^\theta)$	1
Poisson $P(\mu)$	$0,1,\ldots,\infty$	μ	μ	$\log(\mu)$	e^θ	1
Neg. Binom. $NB(\mu,\alpha)$	$0,1,\ldots,\infty$	μ	$\mu+\alpha\mu^2$	$\log(\frac{\alpha\mu}{1+\alpha\mu})$	$-\frac{1}{\alpha}\log(1-\alpha e^\theta)$	1
Gaussian/Normal $N(\mu,\sigma^2)$	$(-\infty,\infty)$	μ	1	μ	$\frac{1}{2}\theta^2$	σ^2
Gamma $N(\mu,\nu)$	$(0,\infty)$	μ	μ^2	$-\frac{1}{\mu}$	$-\log(-\theta)$	$\frac{1}{\nu}$

음이항분포는

- 푸아송 분포와 마찬가지로 **카운트 데이터**가 따르는 확률분포입니다.
- 푸아송 분포보다 분산이 크다.
- 예) 무리 짓는 생물의 개체수라면 푸아송 분포로는 상정할 수 없는 큰 분산이 되는 경우도 있다. **과분산**

감마분포는

- 0이상의 값을 취하는 **연속형** 확률변수가 따르는 확률분포이다.
- 분산값도 평균값에 따라 변합니다(등분산x)

$k \geq r$ 일 때, k 번째의 실험에서 r 번째의 성공을 얻기 위해서는 $k-1$ 번째 실험까지 $r-1$ 번의 성공이 있어야 한다. 이항 분포와 마찬가지로, $k-1$ 번째 실험까지 $r-1$ 번의 성공이 있을 확률은

$$\binom{k-1}{r-1} p^{r-1}(1-p)^{k-r}$$

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \quad for\, 0 \leq k \leq n$$

기댓값과 분산 [편집]

확률변수 X가 빈도 λ를 모수로 갖는 지수분포를 따른다면, 기댓값은

$$\mathrm{E}(X) = \frac{1}{\lambda}$$

으로 단위 시간당 사건이 λ회 발생한다면, 사건 사이에 평균적으로 $1/\lambda$시간만큼 기다릴 것이라는 것을 의미한다. 분산은

$$\mathrm{Var}(X) = \frac{1}{\lambda^2}$$

6.1.13 지수형 분포

정규분포 이외의 학률분포로 이용되는 것이 **지수형 분포족**이라고 불리는 분포

정규분포가 가지는 편리한 성질을 가지고 있기 때문에 모델의 추정이나 해석이 용이하다. (구체적인건 관련 문헌을 참고하래요)

$$f(x \mid \theta) = \exp[a(x)b(\theta) + c(\theta) + d(x)]$$

a(x)=x인 분포를 정준형canonical이라고 부르며, b(Θ)를 분포의 자연 파라미터라고 부릅니다.

예) 푸아송은 지수족이라 정준형임.

푸아송 분포의 확률질량함수

$$\mathrm{Pois}(x \mid \lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$$

위 식을 아래처럼 변형할 수 있습니다.

$$\mathrm{Pois}(x \mid \lambda) = \exp[x\log\lambda - \lambda - \log x!]$$

$a(x) = x$이므로 정준형이며, 자연 파라미터는 $\log\lambda$가 됩니다.