02wk-1: 베르누이, 이항분포, 포아송

1. 강의영상

2. Imports

```
using Distributions, Plots,PlutoUI
```

Table of Contents


```
PlutoUI.TableOfContents()
```

```
PlotlyBackend()
```

```
Plots.plotly()
```

3. 베르누이: $X \sim Bernoulli(p)$

A. 기본내용

– 간단한 요약

- X의의미: 성공확률이 p인 1번의 시행에서 성공한 횟수를 X라고 한다.
- X의범위: X=0 or X=1
- 파라메터의 의미와 범위: p는 성공확률을 나타냄, p ∈ [0,1].
- pdf:
- mgf:

- E(X)=p
- V(X)=p(1-p)

– 난수생성코드 (줄리아문법)

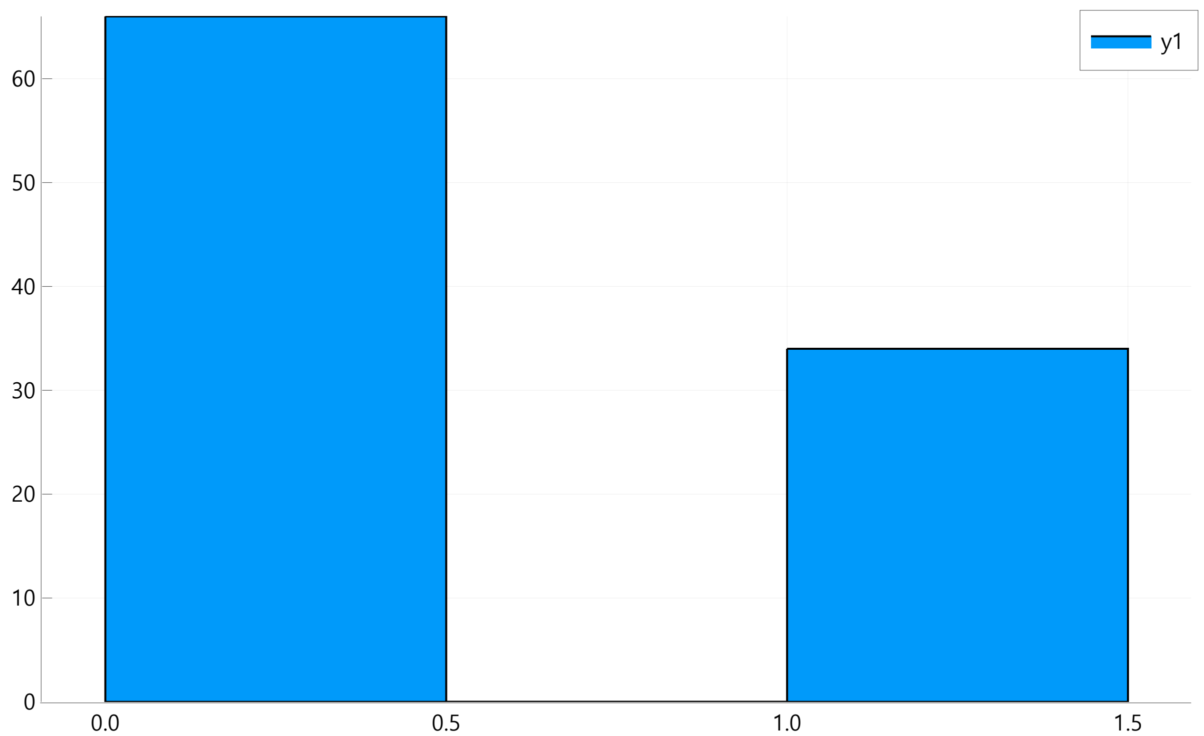
```
[true, true, false, false, false, true, true, true, true, true]
```

```
let
    p = 0.6 # 파라메터
    N = 10 # 샘플수
    distribution = Bernoulli(p) # 분포오브젝트 자체를 정의
    X = rand(distribution,N) # N-samples
end
```

B. 모수 → 히스토그램

```
md"p = $(@bind p Slider(0.1:0.1:0.9, show_value= true, default=0.3))"
#p = @bind p Slider(0.1:0.1:0.9, show_value= true, default=0.3)
```

Fig – 모수에 따른 베르누이의 pdf (pmf)




```
let
    N = 100
    histogram(rand(Bernoulli(p),N))
end
```

C. 난수생성 테크닉

– 베르누이분포에서 100개의 샘플을 뽑는 방법 (p=0.37로 가정)

(방법1) – 기본

```
[true, true, false, true, false, false, false, true, false, true, false, false, true, fal
```

```
rand(Bernoulli(0.37),100)
```

(방법2) 균등분포 -> 베르누이분포

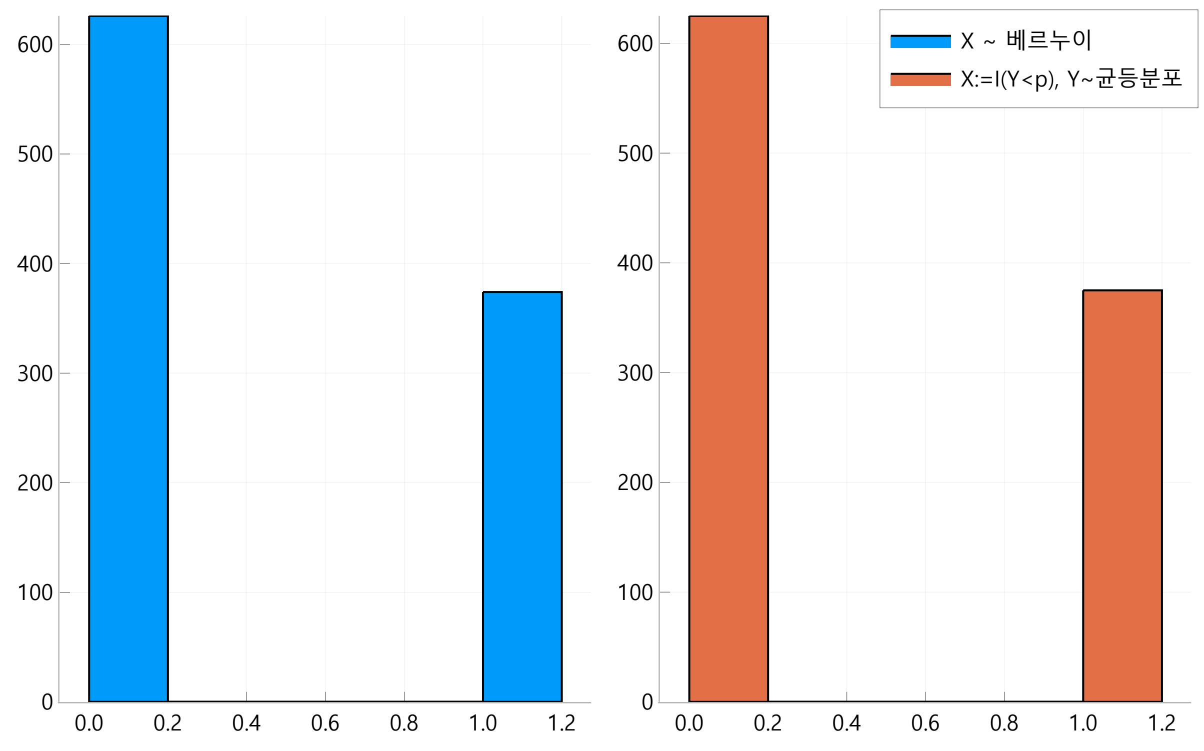
```
[0.887147, 0.657147, 0.845708, 0.999751, 0.533331, 0.835676, 0.0110171, 0.641092, 0.7751!
```

```
rand(100) # 유니폼에서 100개의 샘플 추출
```

```
BitVector: [false, false, false, false, true, true, true, false, false, true, true, false
```

```
rand(100) .< 0.37 # 0.37보다 작은것만 성공
```

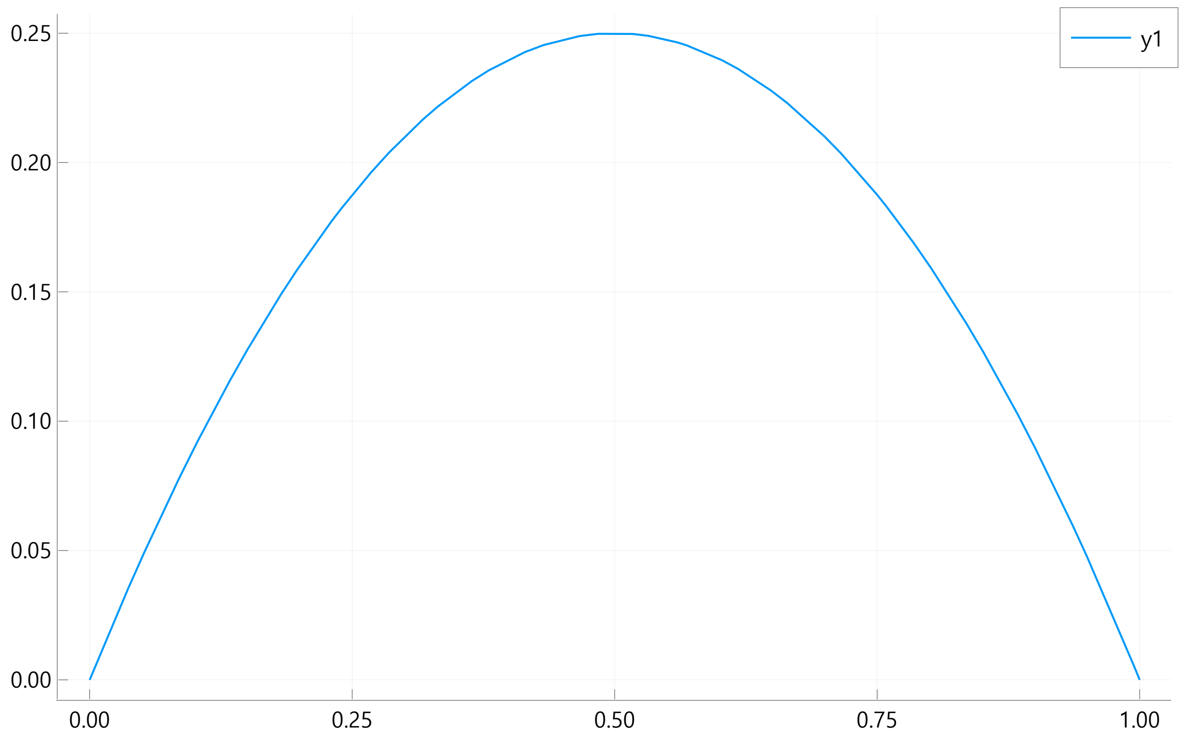
Fig – 방법1,2의 비교




```
let
    X = rand(Bernoulli(0.37),1000) # 방법1
    Y = rand(1000) .< 0.37 # 방법2
    p1 = histogram(X,label="X ~ 베르누이",color=1)
    p2 = histogram(Y,label="X:=I(Y<p), Y~균등분포",color=2)
    plot(p1,p2)
end
```

D. 분산을 최대화

Fig – 분산의 그래프




```
plot(p -> p*(1-p),0,1)
```

4. 이항분포: $X \sim B(n,p)$

A. 기본내용

– 간단한 요약

- X의 의미: 성공확률이 p인 n번의 시행에서 성공한 횟수를 X라고 한다.
- X의 범위: X=0,1,…,n
- 파라메터의 의미와 범위: n은 시행횟수를 p는 성공할 확률을 의미. n=1,2,… and p ∈ [0,1].
- pdf:
- mgf: (베르누이분포의mgf)^n – 왜??
- 평균: np
- 분산: np(1-p)

– 대의적정의 (⋆)

이항분포의 대의적 정의

베르누이 분포를 합치면 이항분포가 된다.

- $X_1, X_2, \dots, X_n \overset{iid}{\sim} Bernoulli(p) \Rightarrow X_1 + X_2 + \cdots + X_n \sim B(n,p)$

또한 모수가 (n,p)인 이항분포는 논리전개의 편의에 따라 n개의 베르누이분포의 합으로 쪼갤수도 있다. 이를 엄밀한 수학언어로 표현하면 아래와 같다.

- $\big(X \sim B(n,p)\big) \Rightarrow \Big($there exists $X_1, \dots, X_n$ such that (1) $X_1 \dots, X_n \overset{iid}{\sim} Bernoulli(p)$ and (2) $X_1 + \cdots + X_n \overset{d}{=} X\Big)$

– 난수생성코드 (줄리아문법)

```
[20, 14, 16, 18, 20, 23, 21, 17, 15, 15]
```

```
let
    p,n = 0.6,30 # 파라메터
    N = 10 # 샘플수
    distribution = Binomial(n,p) # 분포오브젝트 자체를 정의
    X = rand(distribution,N) # N-samples
end
```

B. 모수 → 히스토그램

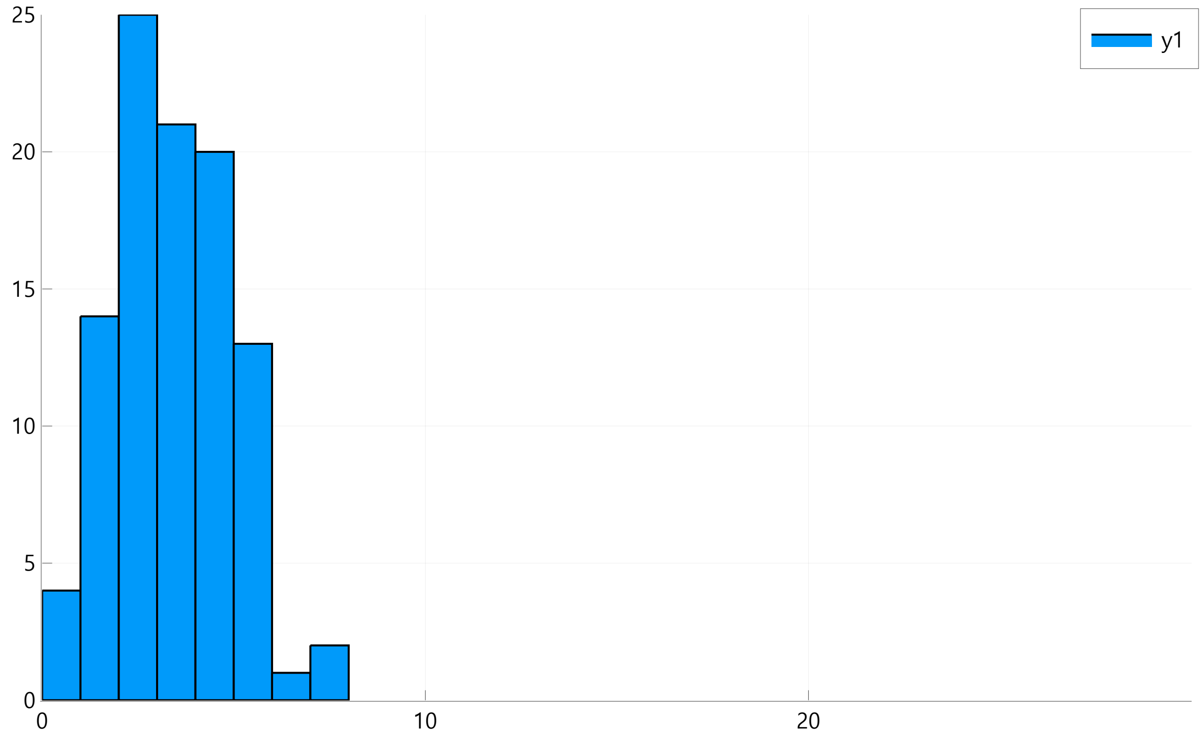
p = 0.3

```
md"p = $(@bind p Slider(0.1:0.1:0.9, show_value=true, default=0.3))"
#p = @bind p Slider(0.1:0.1:0.9, show_value=true, default=0.3)
```

n = 10

```
md"n = $(@bind n Slider(1:1:30, show_value = true, default=10))"
#n = @bind n Slider(1:1:30, show_value = true)
```

Fig – 이항분포의 pdf (pmf)




```julia
let
    N = 100
    histogram(rand(Binomial(n,p),N))
    xlims!(0,30)
end
```

C. 난수생성 테크닉

– 이항분포에서 100개의 샘플을 뽑는 방법 (p=0.37, n=8 이라고 가정)

(방법1)

```
[2, 3, 1, 3, 6, 4, 3, 2, 3, 3, 4, 4, 2, 2, 4, 3, 2, 5, 3, 4,    more  ,3, 2, 4, 4, 4, 4, 2, 4
```

```julia
rand(Binomial(8,0.37),100)
```

(방법2) 베르누이 -> 이항분포

```
[true, false, false, true, false, false, false, false]
```

```julia
rand(Bernoulli(0.37),8)
```

```
[[false, false, false, false, false, false, false, false], [false, false, false, false, f
```

```julia
[rand(Bernoulli(0.37),8) for i in 1:100]
```

```
[4, 5, 5, 3, 2, 5, 3, 2, 6, 2, 1, 4, 3, 7, 1, 5, 4, 3, 3, 2,    more ,2, 3, 5, 1, 1, 3, 3, 3
```

```julia
[rand(Bernoulli(0.37),8) for i in 1:100] .|> sum
```

(방법3) 균등분포 -> 베르누이분포 -> 이항분포

```
[0.167662, 0.99733, 0.506733, 0.718674, 0.75125, 0.107242, 0.87939, 0.148759]
```

```julia
rand(8) # 유니폼에서 8개를 뽑는다.
```

```
BitVector: [false, false, false, true, true, true, true, false]
```

```julia
rand(8) .< 0.37 # 성공확률이 0.37인 베르누이에서 8개의 샘플을 뽑은셈
```

```
[BitVector: [false, false, true, false, true, false, true, false], BitVector: [false, tru
```

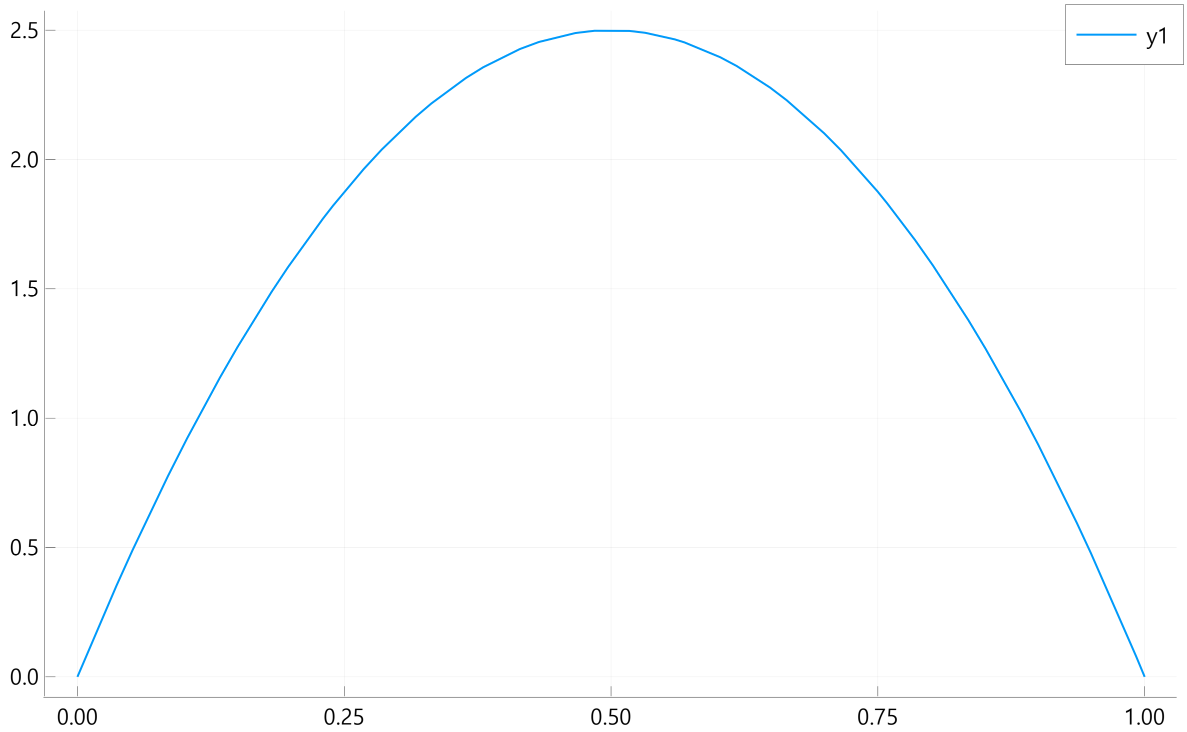
```julia
[rand(8) .< 0.37 for i in 1:100]
```

```
[4, 5, 1, 4, 3, 2, 2, 2, 3, 4, 6, 2, 4, 4, 5, 6, 3, 5, 3, 4,    more ,3, 3, 4, 2, 2, 1, 4, 1
```

```julia
[rand(8) .< 0.37 for i in 1:100] .|> sum # (n,p)=(8,0.37)인 이항분포에서 100개를 뽑은
셈
```

D. 분산의 최대화

Fig – 분산의 그래프




```julia
let
    n = 10
    plot(p-> n*p*(1-p),0,1)
end
```

E. 이항분포의 특징

이항분포의 합

이항분포의 합은 다시 이항분포가 된다.

- $X \sim B(n,p), Y \sim B(m,p),\ X \perp Y \Rightarrow X+Y \sim B(n+m,p)$

5. 포아송분포: $X \sim Poi(\lambda)$

A. 기본내용

– 간단한 요약

- X의의미: 발생횟수의 평균이 λ인 분포에서 실제 발생횟수를 X라고 한다.
- X의범위: 발생안할수도 있으므로 X=0이 가능. 따라서 X=0,1,2,3,...
- 파라메터의 의미와 범위: λ = 평균적인 발생횟수; λ>0.
- pdf:
- mgf:
- E(X): λ

- V(X): λ

– 난수생성코드(줄리아문법)

```
[6, 5, 10, 3, 5, 2, 9, 9, 10, 5]
```

```
let
    λ = 5.3 # 파라메터
    N = 10 # 샘플수
    distribution = Poisson(λ) # 분포오브젝트 자체를 정의
    X = rand(distribution,N) # N-samples
end
```

– 포아송분포의 예시 (⋆)

- 콜센타에 걸려오는 전화의 수, 1시간동안
- 레스토랑에 방문하는 손님의 수, 하루동안
- 웹사이트를 방문하는 사람의 수, 1시간동안
- 파산하는 사람의 수, 1달동안
- 네트워크의 끊김 수, 1주일동안

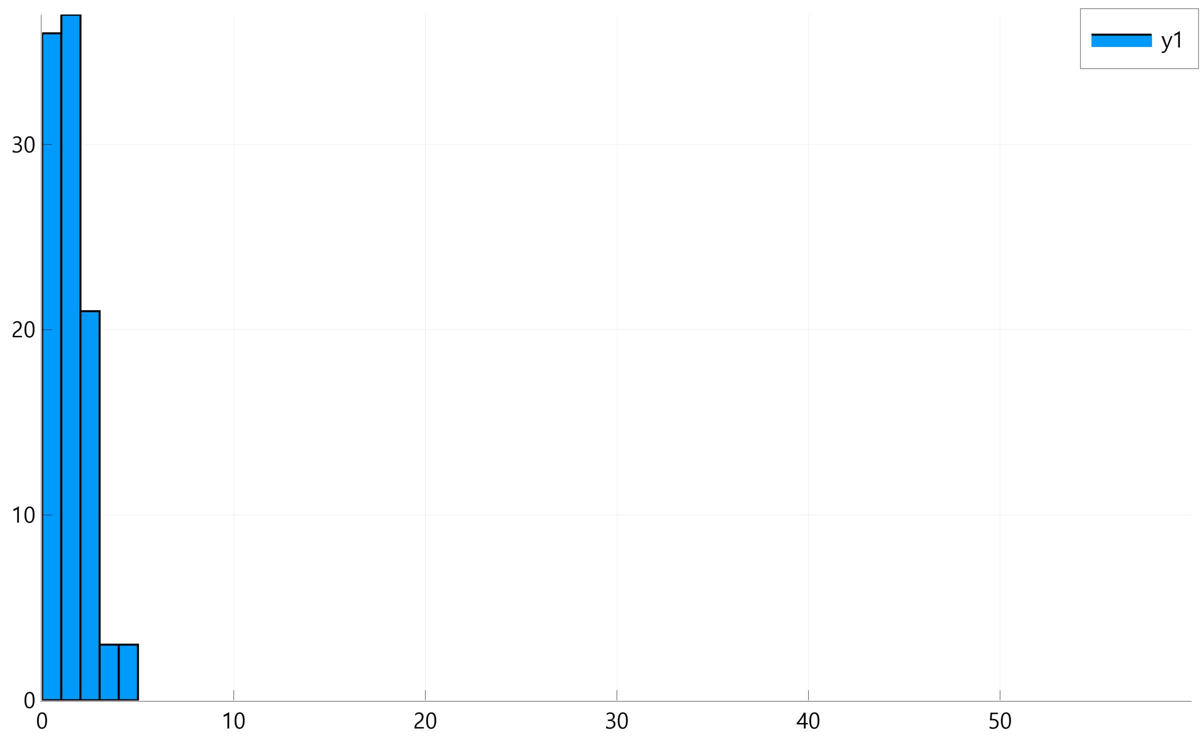
> **포아송분포의 느낌**
>
> 단위시간 (혹은 단위공간) 에서 발생하는 어떠한 이벤트 수를 X라고 하면 X는 포아송분포를 따름.

B. 모수 → 히스토그램

λ = 1.0

```
md"λ = $(@bind λ Slider(0.1:0.1:30, show_value=true, default=1))"
#λ = @bind λ Slider(0.1:0.1:30, show_value=true, default=1)
```

Fig – 포아송분포의 pdf (pmf)




```
let
    N = 100
    histogram(rand(Poisson(λ),N))
    xlims!(0,60)
end
```

C. 난수생성 테크닉

– 방법1

```
[5, 0, 2, 2, 2, 1, 2, 3, 1, 0]
```

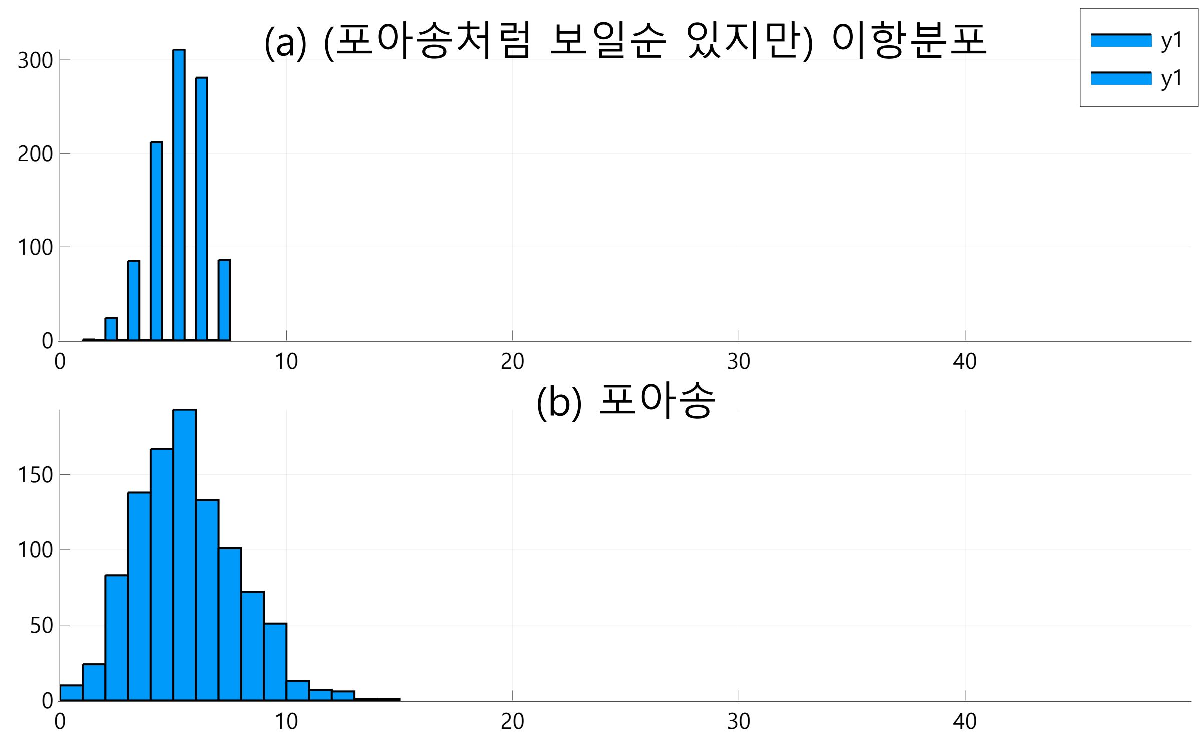
```
rand(Poisson(3),10)
```

– 방법2: 이항분포의 포아송근사

이론: 이항분포의 포아송근사

이항분포에서 (1) $n \to \infty$ (2) $p \to 0$ 이면 이것은 평균이 $\lambda = np$ 인 포아송분포와 비슷해진다. 즉 평균이 λ인 포아송분포는 $B(n, \frac{\lambda}{n})$로 대신 만들 수 있다.

Fig – 이항분포의 포아송 근사




```julia
let
    N = 1000
    λ = 5
    n = 7
    p = λ/n
    X = rand(Binomial(n,p),N)
    Y = rand(Poisson(λ),N)
    @show (n,p), λ
    #--#
    p1= histogram(X); xlims!(0,50); title!("(a) (포아송처럼 보일순 있지만) 이항분포")
    p2= histogram(Y); xlims!(0,50); title!("(b) 포아송")
    plot(p1,p2,layout=(2,1))
end
```

```
((n, p), λ) = ((7, 0.7142857142857143), 5)
```

– 방법3: 베르누이 → 이항분포 ≈ 포아송

사실: 전북대 맥도날드에는 항상 1분에 평균 6명의 손님이 방문한다. (느낌? 평균이 6인 포아송)

- 그럼 10초에는 대충 1명의 손님이 오지 않겠어?
- 그럼 1초에는 대충 0.1명의 손님이 오지 않겠어?
- 그럼 0.1초에는 대충 0.01명의 손님이 오지 않겠어?

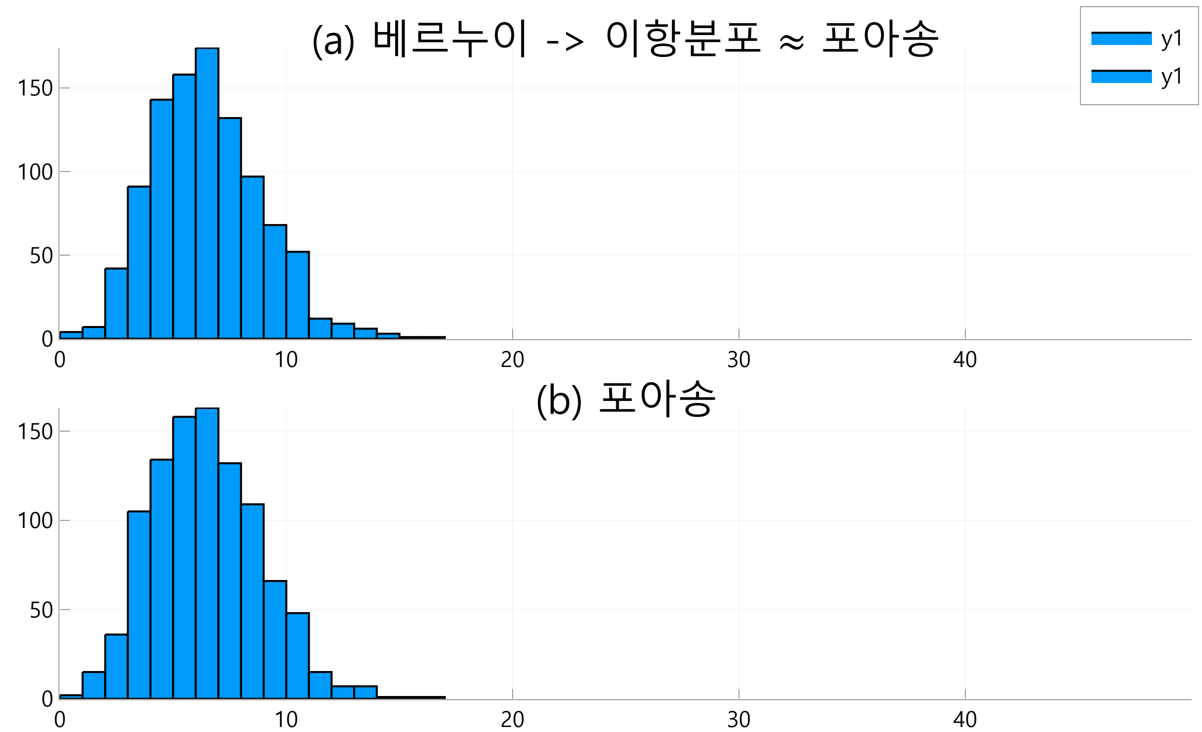
생각: (x,x+0.1초) 에서 방문객의 분포와 (x+0.1초,x+0.2초) 방문객의 분포는 독립일까? 분포는 다를까?

- 딱봐도 분포는 같고, 독립이어보임.

주장:

- 주장1: 0.1초동안 맥도날드에 오는 손님수 X는 $X \sim Ber(0.01)$ 라고 봐도 거의 무방.
- 주장2: 그렇다면 60초동안 맥도날드에 오는 손님수 X는 $X \sim B(600, 0.01)$ 이어야 한다.
- 주장3: 그런데 $B(600, 0.01)$은 충분히 큰 n과 충분히 작은 p를 가지고 있다. 따라서 이것은 $Poi(6)$의 분포와 비슷할 것이다.

Fig – 포아송프로세스의 이해




```julia
let
    N = 1000
    λ = 6
    n = 600
    p = 0.01
    X = [rand(Bernoulli(p),n) for i in 1:N] .|> sum
    Y = rand(Poisson(λ),N)
    @show (n,p), λ
    #--#
    p1= histogram(X); xlims!(0,50); title!("(a) 베르누이 -> 이항분포 ≈ 포아송")
    p2= histogram(Y); xlims!(0,50); title!("(b) 포아송")
    plot(p1,p2,layout=(2,1))
end
```

```
((n, p), λ) = ((600, 0.01), 6)
```

포아송 프로세스 느낌

하여튼 (1) "**엄청 짧은 시간**"에 (2) "**엄청 작은 확률**"의 베르누이 시행이 (3) "**엄청 많이 독립적으로 반복**"되는 느낌을 꼭 기억하세요!

D. 분산이 특이하네?

– 떡밥..

E. 포아송 특징

– 포아송분포의 합은 다시 포아송분포가 된다.

이론: 포아송분포의 합

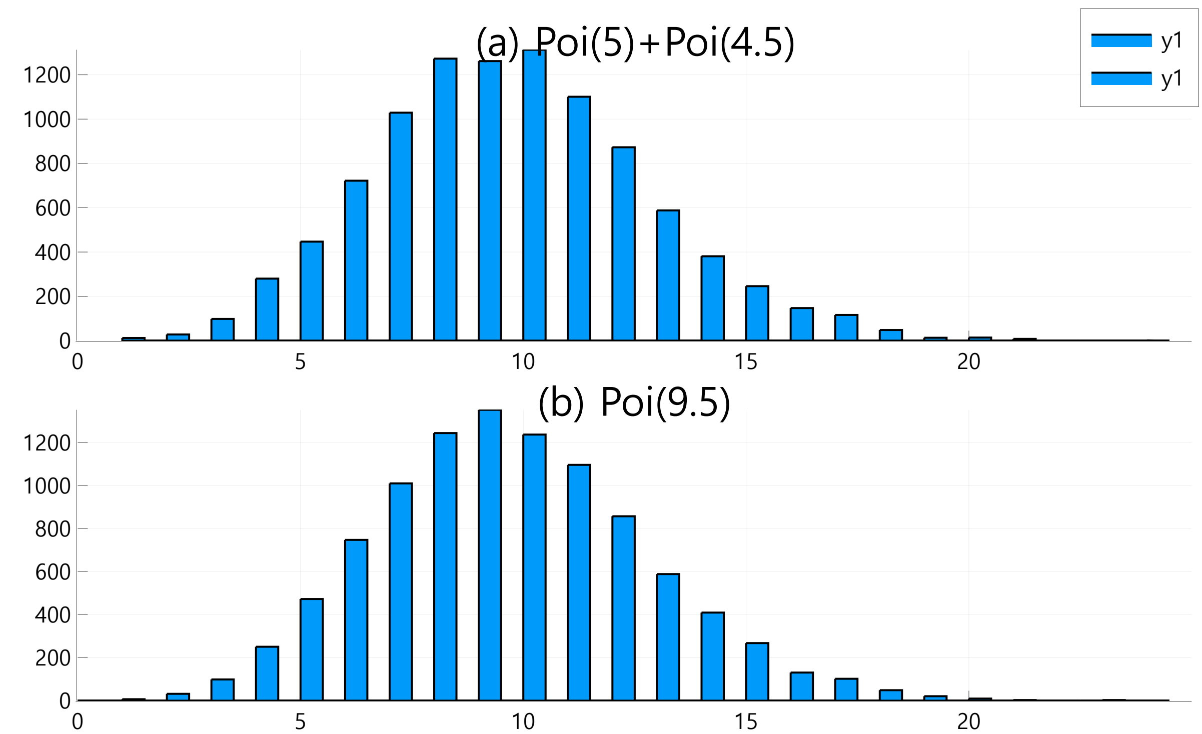
포아송분포의 합은 다시 포아송분포가 된다.

- $X \sim Poi(\lambda_1), Y \sim Poi(\lambda_2),\ X \perp Y \Rightarrow X+Y \sim Poi(\lambda_1+\lambda_2)$

의미? (1) 1분동안 맥도날드 매장에 들어오는 남자의 수는 평균이 5인 포아송 분포를 따름 (2) 1분동안 맥도날드 매장에 들어오는 여자의 수는 평균이 4.5인 포아송분포를 따름 (3) 남자와 여자가 매장에 오는 사건은 독립 => 1분동안 맥도날드 매장에 오는 사람은 평균이 9.5인 포아송 분포를 따른다는 의미.

– 실습

Fig – 포아송분포의 합은 다시 포아송이다




```
let
    N= 10000
    X = rand(Poisson(5),N) # 남자
    Y = rand(Poisson(4.5),N) # 여자
    p1 = X.+Y |> histogram ; title!("(a) Poi(5)+Poi(4.5)") ; xlims!(0,25)
    p2 = rand(Poisson(9.5),N) |> histogram ; title!("(b) Poi(9.5)") ; xlims!(0,25)
    plot(p1,p2,layout=(2,1))
end
```

6. 숙제

1. 수업시간에 소개한 이항분포를 만드는 3가지 방법으로 (n,p)=(30,0.45)인 이항분포 100를 만들라. 세 방법의 히스토그램을 비교해보라.

2. "균등분포 $\to$ 베르누이 $\to$ 이항분포 $\approx$ 포아송" 의 방법으로 $Poi(12)$의 분포를 근사하고 히스토그램을 비교해보라.