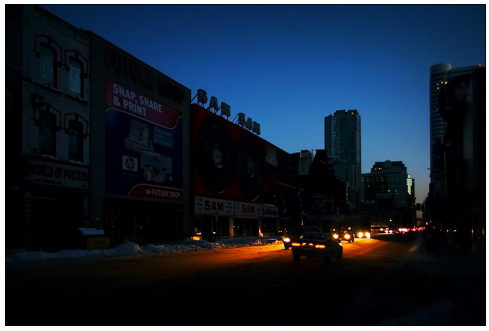
Chapter 6 Continuous Probability Distributions

예측 대실패가 부른 '9·15 대정전'

지난 2011 년 9 월 15 일, 전국 곳곳이 어둠에 휩싸였다. 이른바 '블랙아웃'이라 불리는 사상 최대의 전국적인 대규모 정전사태가 발생한 것이다. 한동안 낮아졌던 기온이 갑자기 오르자 예상보다 훨씬 많은 전력이 소모되면서 전국 곳곳이 어둠에 잠겼다. 한국전력은 사실상 여름이 지났다고 판단해 예비전력량을 줄인 상태였다. 게다가 당시 미뤄왔던 발전기 정비에 들어간 상황이라 돌릴 수 있는 발전기 수도 적었다.

대규모 정전사태가 발생하자 한전은 순환전력공급으로 피해 최소화에 나섰다. 하지만 전국의 가정과 기업 곳곳에서 적지 않은 재산상 피해와 혼란을 겪었다. 엘리베이터는 물론 수족관, 냉장고, 영업장, 산업시설 등이 갑작스런 정전으로 가동을 멈췄다. 이후 국민들은 한동안 전력위기에 시달려야 했고, 그 결과 매년 하절기와 동절기마다 '전력 보릿고개'를 겪고 있다.



올해 역시 정전사태 위기감이 감돈다. 특히 올해는 5 월부터 한여름 날씨가 찾아왔다. 7 월 이후에는 본격적인 무더위가 기승을 부리고 있다. 9·15 대규모 정전사태를 초래할 수 있는 심각한 위기상황이 여전히 도사리고 있는 셈이다.

〈머니위크〉(www.moneyweek.co.kr) 제 340 호

미국 드라마 Revolution 을 보면 전기가 없어진 세상에 사는 사람들의 모습을 보여주는데 인류의 발전이 정지된 모습이 인상적입니다. 이 드라마에서 모습처럼 전기는 우리에게 없어서는 안되는 중요한 것이고, 하루라도 쓸 수 없다면 우리는 전화, 컴퓨터, TV, 라디오등 전기에 의존하는 장비를 사용할 수 없어 답답함을 느낄것입니다.
이처럼 전기가 끊어지는 전정(Blackout)이 발생하면 또한 안전문제에도 문제가 됩니다. 예를 들어 수술 중 전기가 없어 의급한 환자를 치료할 수 없게 되고, 신호등이 동작을 하지 않아 교통 사고를 초래할 수 있고, 무엇보다도 산업 현장에서 문제가 가장 큽니다.

2011 년 9 월 15 일에 발생된 사고는 예측이 빗나간 것이라고 말을 합니다. 이러한 예측의 근거는 과거의 같은 기간에 사용한 전력 사용량을 근거로 예비전력을 준비했을텐데

갑작스러운 기온 상승으로 많은량의 전력 사용은 전력량 사용량에 대한 분포도로 보았을 때 끝쪽에 발생 확률이 매우 낮은 outlier 이었기에 크게 걱정을 하지 않을 것입니다.
전력량과 같이 관측 값이 셀 수 있는것이 아닌 측정에 의해 나타나는 값들에 대한 확률을 continuous probability 라고 하고 이 분포를 continuous probability distribution 이라 합니다.

다음의 예들은 continuous data 를 생성하는 실험에서 continuous random variable 들을 생성하는 예제들 입니다.

- 피자를 주문한 후 고객에게 배달하는데 소요되는 시간
- 웹사이트에 방문한 사람이 머무는 시간
- 택배 회사의 배달 직원들이 사용하는 하루 연료량
- 다이어트 실시후 몸무게 변화
- 제조한 청바지의 실제 길이의 오차.
- 스타벅스에 주문한 드립 커피량

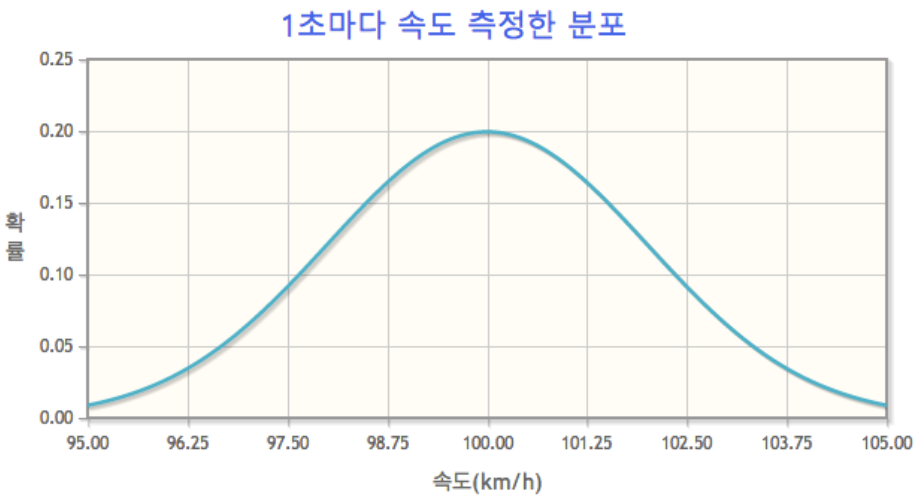
Discrete probability density function 은 P(x)를 random variable x 값이 정확하게 1, 2 와 같이 하면 결과를 얻지만 continuous probability density function 의 경우는 random variable 에 어느 특정 값을 넣으면 확률은 0 이 됩니다. 예를 들어 커피 주문 후 커피를 만들어 제공되는 시간이 정확하게 63.5 초인 경우를 계산하기 위해서 총 기간이 90 초인 경우에 63.5 초가 몇 번이나 발생하는 가를 알아야 하는데 90 초 안에 발생 될 수 있는 경우의 수는 무한대이기 때문에 확률은 1/∞로 0 이 됩니다.

이러한 특성 때문에 continuous pdf 대신 cdf 를 사용합니다. 예를 들어 $P(x \le 63.5)$, $P(50.5 \le x \le 63.5)$와 같이 사용합니다.

여러 distribution 들이 있지만 이 장에서 설명할 distribution 들은 normal distribution, exponential distribution, uniform distribution 입니다.

1. Normal distribution (Gaussian distribution)

자동차를 운전하는 운전자가 고속도로 운전을 하는 도중 속도를 시속 100km 로 유지하려고 한다 했을 매 순간 속도를 정확하게 100km/h 를 유지하는것은 어렵고 95km/h ~ 105km/h 사이에서 변하면서 속도를 최대한 100km/h 에 맞추려고 노력할 것입니다. 이러한 속도를 1 초마다 측정을 하여 분포를 그리게 된다면 다음과 같은 모양의 곡선이 나타날 것입니다.



가운데 지점은 100km/h 를 중심으로 양쪽으로 대칭으로 측정된 속도의 분포가 점점 줄어드는 것을 볼 수 있습니다. 이러한 분포가 바로 Normal distribution 입니다.

이러한 현상은 측정을 하거나 정확하게 값을 생성하지 못하는 경우에 발생되는 경우가 많이 있습니다. 예를 들어 철판을 200mm 단위로 자를 때 기계의 노후로 정확한 200mm 를 유지하지 못하고 이 값을 중심으로 오차가 있을 것입니다. 여기서 오차의 평균이 표준편차 입니다. 이러한 평균과 표준편차 만을 갖고 분포도를 얻는 방법으로 Normal distribution 을 사용합니다.

$$f(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{6.1}$$

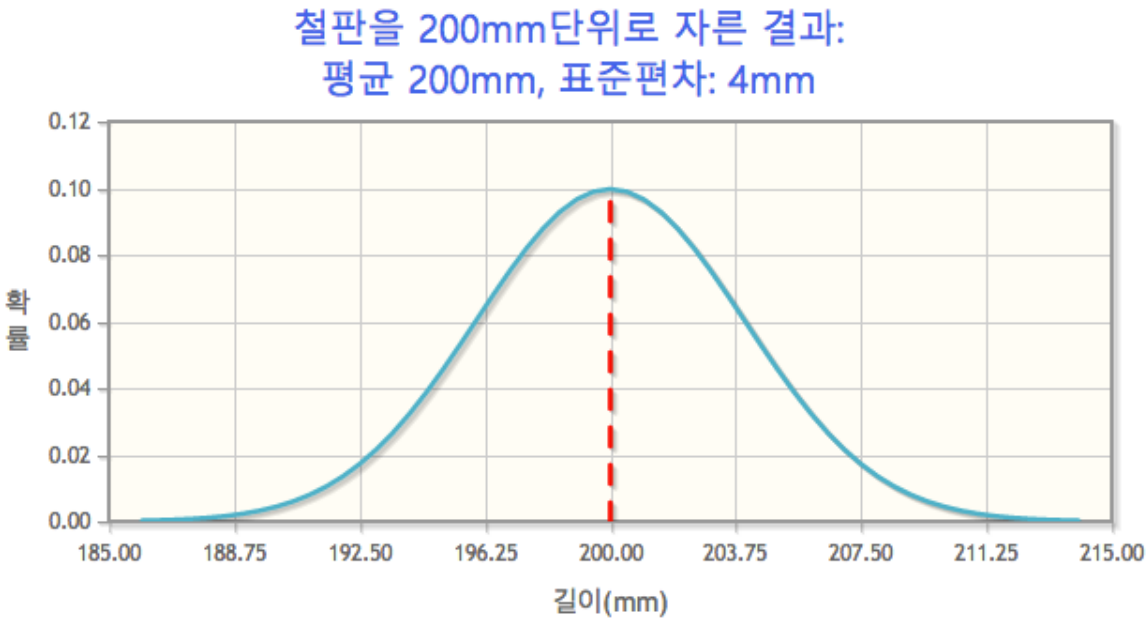
여기서 μ는 평균(mean)이고 σ는 표준 편차(standard deviation)입니다.

jMath.stat.normpdf(x,mu,sigmal)

```
> jMath.stat.normpdf(198,200,4)
0.08801633169107487
```

수식 6.1 으로 부터 철판 길이에 대한 예를 적용하면 다음과 같은 그래프를 보실 수 있습니다.

<chatper06/chapter06_normpdf.html>

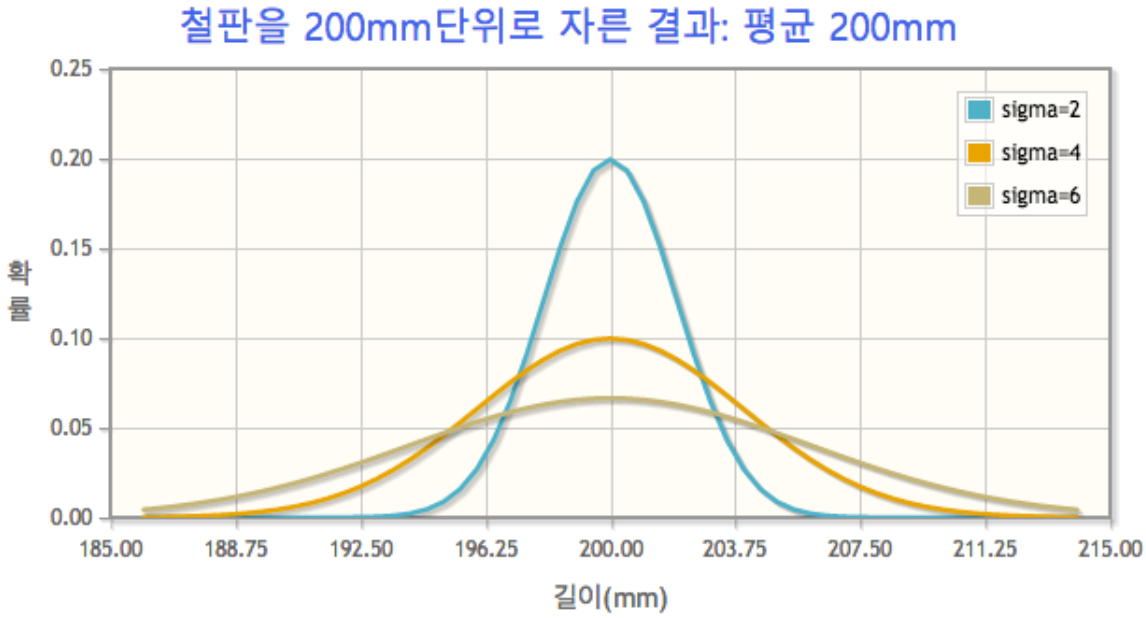


이런 분포도를 보고 normal distribution 의 몇가지 특징들을 알 수 있습니다.

- 분포는 벨 모양으로 평균값 200mm 를 중심으로 대칭.
- 모든 분포의 확률의 합은 1 이므로 곡선 아래의 넓이는 1.
- 중앙을 중심으로 대칭이기 때문에 왼쪽, 오른쪽 각각의 넓이는 0.5.
- 평균을 중심으로 확률값은 양쪽으로 같은 값으로 퍼져 있고 median 과 mode 둘다 평균과 같음.
- 평균값 주변 일 수록 발생될 확률이 높지만 멀어 질 수록 확률이 낮아짐.
- 곡선은 무한대로 양쪽으로 뻗어 나가지만 확률 값은 0 을 만나지 않음.

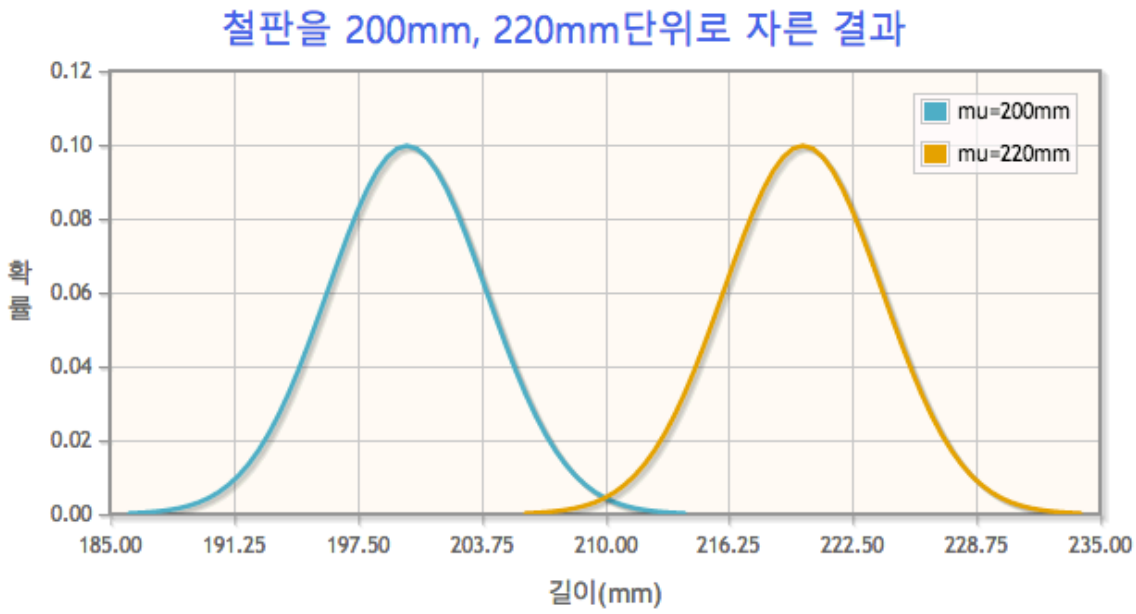
만일 장비를 새로 교체하여 오차를 줄인다면 오차 평균인 표준 편차가 작아지고, 그렇지 않고 장비가 고장이 정도가 심하게 되면 오차는 커지고 결과 표준 편차도 커지게 됩니다. 이러할 경우를 Graph 로 표현을 하면 다음과 같습니다.

<chatper06/chapter06_normpdfcmp.html>



보시는 것과 같이 표준 편차가 작을 수로 중심값이 높아 지고 폭은 줄어드는 반면에 표준편차가 커기면 반대로 중심값의 높이는 낮아지고 폭은 넓어 집니다.

만일 강철을 220mm 로 자르고 표준편차가 같다면 분포도는 같은 모양으로 이동만 합니다.

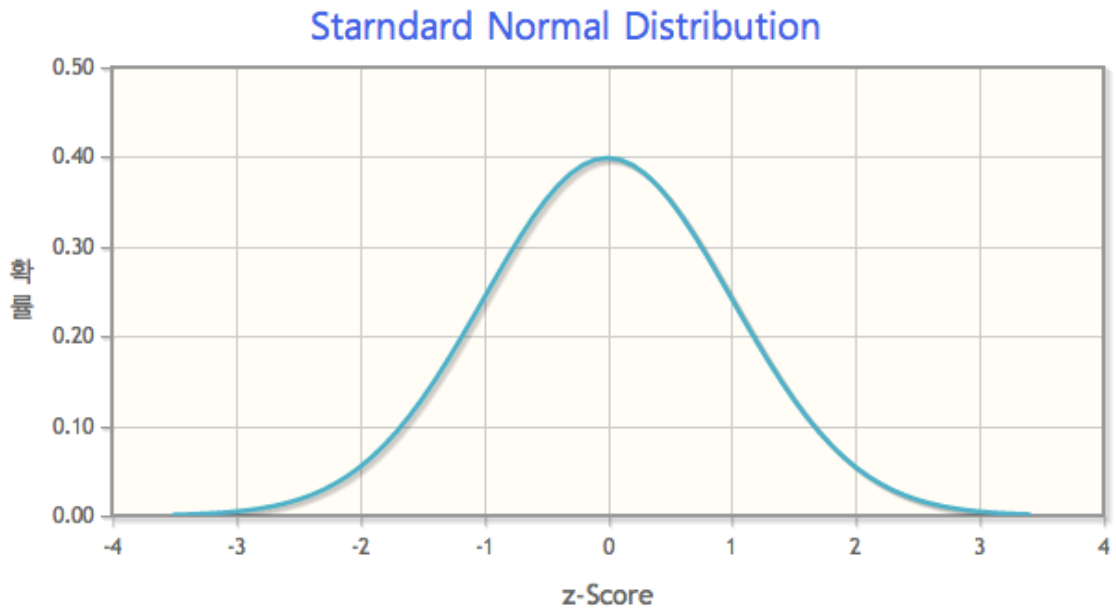


수식 6.1 을 보시면 e 에 지수값은 3 장에서 보셨던 z-Score 임을 알 수 있습니다. 즉 어떠한 평균, 표준편차를 사용해도 z-Score 값은 공통으로 사용할 수 있기 때문에 z-Score 만 안다면 normal pdf 를 cdf 를 구할 수 있습니다.

$$z = \frac{x - \mu}{\sigma} \tag{6.2}$$

이러한 z-Score 의 특성은 평균이 0 이고 표준편차가 1 일 때 즉 z-Score 값이 x 와 같을 때 분포도인 standard normal distribution 보면 됩니다.

<chapter06/6_stdnorm.html>



z-Score 가 음수일 경우인 x 가 평균보다 작으면 0 에 왼쪽에 있고 반대로 평균보다 크면 0 에 오른쪽에 있음을 알 수 있습니다.

3 장에서 다루었던 Empirical rule 을 다시 보시면 다음과 같은데 이것은 normal cdf 으로 확인 하면 같은 결과를 얻게 됩니다.

z-Score 범위	Normal cdf	%
-1 ~ 1	P(z≤1) – P(z<-1) = 1 – 2P(z<-1) = 0.6827	68%
-2 ~ 2	P(z≤2) – P(z<-2) = 1 – 2P(z<-2) = 0.9545	95%
-3 ~ 3	P(z≤3) – P(z<-3) = 1 – 2P(z<-3) = 0.9973	99.7%

여기서 normal distribution 은평균을 중심으로 대칭되어 있기 때문에

$$P(z < -Z) = 1 - P(z < Z)$$

jMath 에서 normal cdf 는 다음과 같습니다.

jMath.stat.normcdf(x,mu,sigmal)

```
> 1 - 2 *jMath.stat.normcdf(-1,0,1)
0.6826894921370859
```

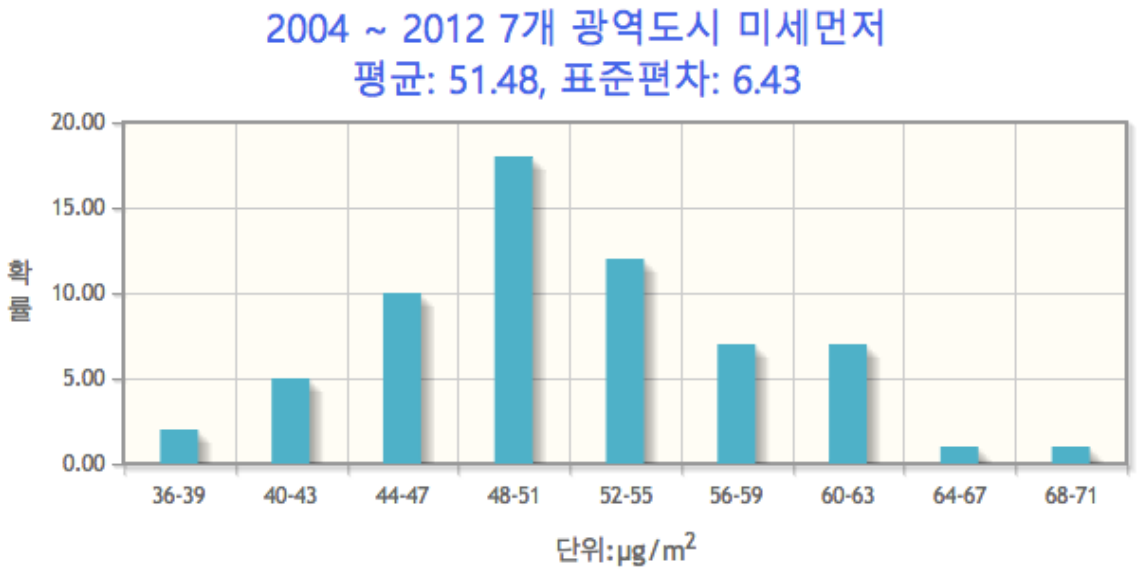
https://github.com/handuck/jMath

철판 200mm 를 자르는데 표준편차가 4mm 인 경우에 적용을 하면 다음과 같습니다.

범위 x = μ + zσ	Normal cdf	%
-196 ~ 204	P(x≤204) – P(x<196) = 0.6827	68%
-192 ~ 208	P(z≤208) – P(z<192) = 0.9545	95%
-188 ~ 212	P(z≤188) – P(z<212) = 0.9973	99.7%

다음 예는 2004 년 부터 2012 년 사이 7 개 광역도시에 미세먼지 농도에 대한 histogram 입니다.

<chapter06/6_airpollution.html>



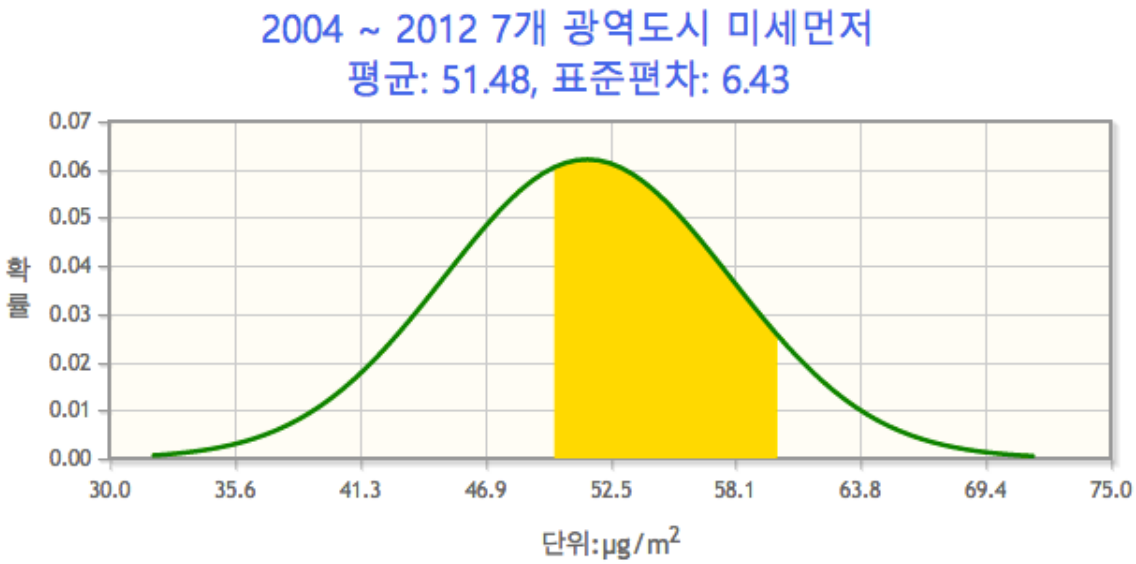
보시는 것과 같이 모양이 normal distribution 와 유사한 모양으로 평균 51.48 과 표준편차 6.43 으로 앞으로 이 결과 값을 근거로 발생될 미세 먼지 농도의 확률을 계산할 수 있습니다.

예를 들어 미세먼지 농도가 다음 년도에 50~60μg/m^2 일 확률을 구하려면 P(x≤60) - P(x≤50)으로 z-Score 는 다음과 같고

$$z = \frac{50 - 51.48}{6.43} = -0.23, \qquad z = \frac{60 - 51.48}{6.43} = 1.327$$

확률을 구하게 되면 0.90768 - 0.409146 = 0.498531 으로 약 50%의 확률로 발생할 것이라는 것을 예측할 수 있습니다.

<chapter06/6_apnormpdf.html>



여기서는 normal distribution 으로 가정하고 넘어가지만 분포도가 적합한지 측정하는 방법은 나중에 배우도록 하겠습니다.

Normal distribution 이 주로 사용되는 경우는 여러번 측정을 하여 분포를 확인하는 경우입니다. 즉 실제 값이 하나인데 측정 할 때 마다 다른 값이 나타난다 하더라도 측정값은 실제값 주변인 경우의 빈도수가 큰 경우가 대부분입니다. 이러한 측정값을 분해를 하면

$$\text{measured value} = \text{true} + \text{err}$$

여기서 true 는 실제 값으로 변화가 되지 않은 값이지만 err 는 측정할 때마다 발생되는 오류값으로 측정을 할 때 마다 값이 변하게 됩니다. 측정값이 normal distribution 인 경우라면 true 값이 평균이 되고 err 가 0 을 중심으로 대칭으로 퍼져있는 normal distribution 이 됩니다.

이러한 원리를 이용하게 되면 측정을 하고자 하는 대상 전체를 검사하지 않고 몇개의 sample 들로 부터 평균을 측정하게 되면 normal distribution 형태를 나타나는 것을 볼 수

있습니다. 예를 들어 회사에서 고객 서비스 만족도를 알고 싶어 조사를 할 경우 모든 사람들에게 물어 보는것은 어렵습니다. 그래서 응답자 30 명씩 묶어서 평균을 구하여 이 평균값의 분포를 보게 되면 모든 사람의 서비스 만족도 점수가 평균 근처에 있고 이 평균값을 중심으로 양쪽으로 퍼져있게 됩니다. 이것이 통계에서 normal distribution 을 많이 사용하는 이유입니다. 이 내용에 대해서 다음장에 자세히 설명하도록 하겠습니다.

- **Binomial distribution 을 Normal distribution 으로 표시하기**

Binomial distribution 을 보시면 평균을 중심으로 대칭으로 있는 구조로 continuous 측면에서 본다면 normal distribution 으로 대신 표현 될 수 있습니다. 특히 만약 실험의 횟수인 n 이 크다면 normal distribution 으로 취급하는 것이 더욱 편리합나다. 이를 위한 조건은

$$np \geq 5 \text{ 이고 } nq \geq 5$$

예를 들어 마트 시식코너에 음식 맛을 본 손님들 중에 물건을 구매할 확률이 8%라고 할 경우 100 명이 시식을 했을 경우 구매에 대한 확률을 보시면 binomial distribution 으로 구매할 확률 p 은 0.08 이 되고 구매하지 않을 확률 q 는 0.92 가 되는 상황에서 np 는 8 이고 nq 는 92 로 둘다 5 보다 큽니다.

주의 사항으로 Discrete probability density function 의 특징은 x 값이 특정 값일 때 확률 값을 얻을 수 있지만 continuous probability density function 은 0 이 됩니다. 그래서 binomial distribution 에 x 값에 해당하는 normal distribution 값을 얻기 위해서는 x 값에 교정값 ±0.5 를 해서 cumulative density function 으로 값을 얻어야 합니다.
예를 들어 앞의 시식코너 예에서 구매 고객의 100 중 7 명의 확률을 구하기 위해서는 binomial distribution 에서는 binopdf(7,100,0.08)로 구할 수 있지만 normal distribution 에서는 normcdf(7.5,8, 2.713) – normcdf(6.5,8.2.713)으로 계산을 해야 합니다.

2. Exponential distribution

Discrete probability distribution 에 Poisson distribution 은 예를 들어 1 시간 동안 평균 20 명의 손님이 오는데 다음 1 시간 내에 30 명의 손님이 올 확률을 구하는 것에 사용됩니다. 반면 Continuous probability distribution 에 Exponential distribution 은 한명의 손님이 몇 분이내로 올 확률과 같은 내용을 구할 때 사용합니다.

- 출근 시간에 택시가 30 분 내로 손님을 태울 확률
- 건설 현장에서 1 개월내로 사고가 날 확률
- 밤 11 시 ~ 새벽 1 시에 대리운전자가 10 분내로 운전 요청을 받을 확률
- 점심 시간에 5 분내로 중국집에 배달요청이 올 확률

기간 당 발생되는 값만을 갖고 exponential distribution 을 계산할 수 있습니다.

$$f(x|\lambda) = \lambda e^{-\lambda x} \tag{6.3}$$

예를 들어 점심 시간에 1 시간 동안 12 개의 주문을 받는 중국집일 경우 분 단위 확률을 구하기 위해서 값을 수정하면 1 분당 0.2 개의 주문을 받는 것과 같습니다. 여기서 바로 λ(lambda)값이 0.2 가 되고 평균 5 분당 1 개의 주문을 받는 것이 됩니다.

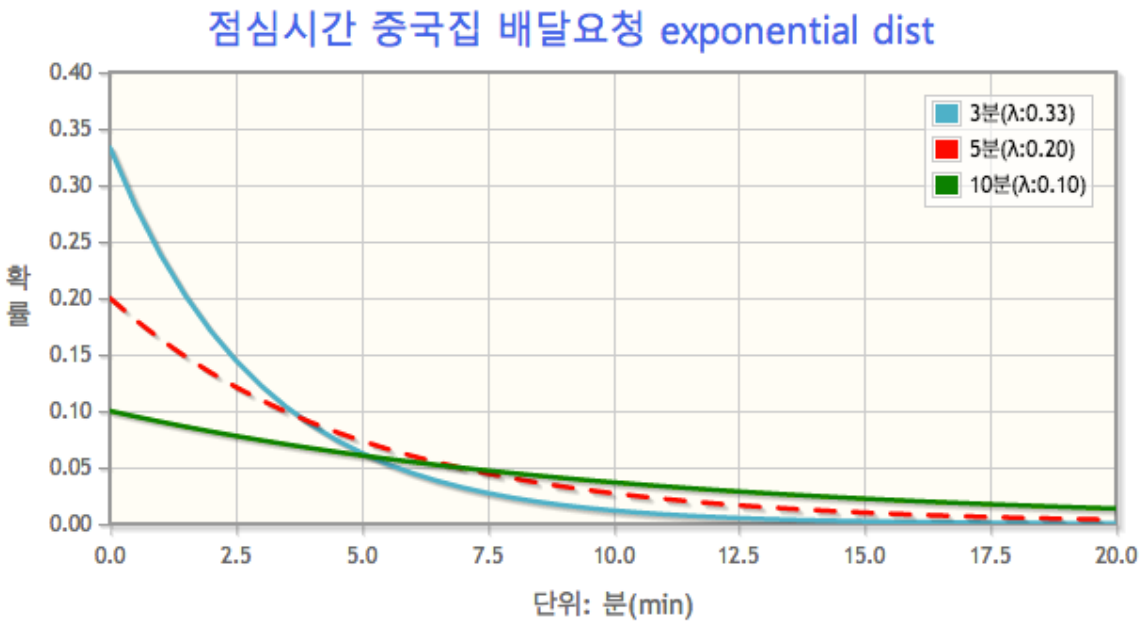
$$E(x) = \mu = \int_0^\infty x\lambda e^{-\lambda x}\, dx = \int_0^\infty e^{-\lambda x}\, dx = \frac{1}{\lambda} \tag{6.4}$$

평균 값인 1 개를 주문 받는데 평균 소요되는 시간을 수식 6.4 를 적용하면 1/0.2 가 되므로 5 가 됩니다.

수식 6.3 을 보면 λ값은 고정된 값이기 때문에 기간에 적용을 할 때는 특성이 비슷한 기간에서만 적용을 해야만 합니다. 예를 들어 중국집에 배달량이 집중되는 시간에는 5 분에 최소 한통이 올 수 있지만 그렇지 않은 시간에는 몇 시간 동안 전화 한통도 오지 않기 때문입니다.

$$\sigma^2 = E(x^2) - \mu^2 = \int_0^\infty x^2 \lambda e^{-\lambda x}\, dx - \frac{1}{\lambda^2} = \frac{2}{\lambda}\int_0^\infty x\lambda e^{-\lambda x}\, dx - \frac{1}{\lambda^2} = \frac{1}{\lambda^2} \quad (6.5)$$

다음은 한개의 주문을 받는데 걸리는 시간에 대한 분포 입니다.



중국집 배달에 요청이 들어오는 평균 시간이 짧을 수록 왼쪽에 영역이 높고 많은 영역을 차지하는 것을 볼 수 있습니다.

jMath.stat.exppdf(x,mu), jMath.stat.expcdf(x,mu)

평균 5 분에 한개의 주문이 들어 올 때 2 분에 주문이 올 확률값과 2 분안에 주문이 올 확률값입니다.

```
> jMath.stat.exppdf(2,5);
0.13406400920712785

> jMath.stat.expcdf(2,5);
0.3296799539643608
```

3. Uniform distribution

Discrete distribution 에서 주사위, 돈전, 카드등 모두 값이 나올 확률은 똑 같았던것과 같이 continuous distribution 에서 특정 범위내에서 모든 값이 발생될 확률이 모두 균일하게 나타내는 것이 uniform distribution 입니다. 다시 말해, 이 분포의 특징은 어느 특정한 값이나 방향으로 분포가 집중되지 않고 고루게 퍼져 있는 것입니다.

- 판매되는 삼겹살 1 인분 무게가 197g 에서 203g 내로 같을 확률을 갖는 경우
- 지하철 연착 시간이 최소 30 초에서 최대 1 분이내로 같은 확률을 갖는 경우

만일 x 가 범위 최소 a 와 최대 b 사이 존재할 경우

$$f(x) = \frac{1}{b-a} \tag{6.6}$$

그렇지 않으면 모두 0 이 됩니다. 평균과 표준편차는 다음과 같습니다.

$$E(x) = \mu = \int_a^b \frac{x}{b-a} dx = \frac{b^2-a^2}{2(b-a)} = \frac{b+a}{2} \tag{6.7}$$

$$\sigma^2 = E(x^2) - \mu^2 = \frac{b^3-a^3}{3(b-a)} - \left(\frac{b+a}{2}\right)^2 = \frac{(b-a)^2}{12} \tag{6.8}$$

<chapter06/6_unifpdf.html>

jMath.stat.unifpdf(x,a,b), jMath.stat.unifcdf(x,a,b)

만일 a 가 없으면 0, b 가 없으면 1 로 설정됩니다.

```
// 범위안에 있는 모든 값에 확률은 모두 동일
> jMath.stat.unifpdf(45,30,60);
0.03333333333333333

// 범위 밖에 있는 값에 확률은 모두 0
> jMath.stat.unifpdf(25,30,60);
0

> jMath.stat.unifcdf(51,30,60);
0.7
```

이 외에 이 교재에서 다룰 몇 가지 continuous distribution 들이 있습니다. 이 분포들은 앞으로 배우면서 하나씩 알아 보겠습니다.