Chapter 5 Discrete Probability Distributions

가전제품 보증기간 5 년 장기화 바람…왜?

가전 업계에서 제품 보증기간 5 년 바람이 불고 있다. 가방 제조사는 무려 25 년간 보증해준다. 자신감 표출 속에 업계의 고민도 함께 담겨 있다.

9 일 업계에 따르면 최근 가전 제조사들은 속속 5 년 보증을 도입하고 있다. 품질 등 기술력이 뒷받침 됐고 소비자에게 안정감을 주기 위한 조치라는 설명이다.
한창 경쟁이 치열한 제습기 시장에서는 위닉스가 5 년 무상보증을 내세웠다.

위닉스는 품질 경쟁력에 기반한 자신감을 내비쳤다. 지난 4 월 가진 기자간담회에서 윤희종 위닉스 대표는 "5 년 무상 품질 보증으로 국내 제습기 시장 확대에 기여하겠다"고 말했다. 지난해까지 출시된 제품의 보증기간도 3 년으로 늘렸다.

영국 고급형 가전 제조업체 다이슨도 5 년 무상보증을 제공한다. 지난 3 일 가진 기자간담회에서 매트 스틸 다이슨 RDD 센터 수석 디자인 엔지니어는 신제품 DC52 를 소개하며 "필터 유지보수가 필요 없는 세계 최초의 제품으로, 다이슨 역사에 이정표라 할 수 있다"며 "보증 기간은 5 년이나 (가정 내 평균 사용 시간 기준) 10 년간 사용해도 흡입력이 떨어지지 않는다"고 말했다.

식품건조기 제조사인 리큅은 이들보다 1 년 더 긴 6 년 보증을 제공한다. 기존 1 년 보증에다 5 년을 추가로 더 제공하는 승부수를 던졌다. 과일이나 채소, 육류 등의 식품을 건조하는 이 제품 시장에서 리큅은 이미 80%에 육박하는 점유율을 갖고 있지만 시장 저변 확대와 점유율 굳히기를 위해 1 위 굳히기에 나섰다.

SSD 업계에서는 인텔과 트랜센드, 플렉스터 등이 5 년 보증기간을 제시하자 샌디스크와 삼성전자가 10 년 보증 카드를 들고 나오는 등 치열할 각축전 속에 보증기간 연장 흐름이 대세로 굳어지고 있다.

전자제품은 아니지만 가방 제조사인 툴레(THULE)의 경우 무려 25 년이라는 긴 보증기간을 제공한다. 여행가방은 물론 카메라 가방, 백팩, 노트북 케이스 등을 국내에 판매 중인 툴레는 이 같은 최장기 보증기간 제공으로 소비자 만족도를 높였다고 강조한다.

대개 전자제품은 1~3 년, 액세서리의 경우 5~10 년 가량이 일반적인 무상보증기간임을 감안하면 이

같은 움직임은 소비자에게 장기간 제품을 사용할 수 있다는 신뢰를 심어준다는 점에서 긍정적이다.

하지만 반드시 그렇지만은 않다는 지적도 있다.

업계 한 관계자는 "보증기간을 길게 제공하는 제품 대부분의 특징은 프리미엄(고급형) 제품이라는 점"이라며 "하이엔드 제품 판매를 늘리는 차원에서 마케팅 도구로 활용되는 경우가 많아서 오히려 판매가격을 올리는 거품 효과가 있는 것도 사실"이라고 지적했다.
또 가전이나 액세서리 시장 모두 성숙기로 접어들면서 사업 환경이 녹록치 않아 시장 저변 확대와 새로운 차별화가 필요해 도입하고 있다는 것도 업계의 전언이다.

ZDNet Korea 2014/07/09

기사 내용과 같이 업체가 보장기간을 결정하는것은 많은 구매를 유도하기 위한 마케팅요소도 있지만 감당하지도 못할 것을 했다가는 금전적인 손실이 엄청나게 클 것입니다. 즉, 보장기간 내에 모든 제품이 다 정상이라고 판단하기도 어렵기 때문에 어느정도 손실을 고려해야 합니다. 그래서 확률로 어느정도 예상을 하여 기간과 판매가를 정하는데, 이를 계산하기 위해서 판매된 개수 대비 고장 신고된 정보를 이용할 수 있습니다.

예를 들어 3 년안에 1000 개를 생산하면 1 개가 고장 제품이 있다고 했을 때 2000 개를 팔았을 때 5 개 이하로 고장날 확률을 계산하여 허용 값보다 작다면 판매가격에 5 개 추가된 가격을 1000 개에 나누어서 적용하여 판매를 하게 된다면 보장 기간내에 물건을 바꾸어 주어도 크게 문제가 되지 않습니다. 그러나 기술의 발달로 보장기간이 늘어나는 것이 아니고서는 보장기간이 길면 고장나는 개수도 많아지고 따라서 판매가격을 올리 수 밖에 없습니다. 이렇게 확률과 통계는 물건 생산과 판매 그리고 마케팅에 중요한 역할을 합니다.

통계를 하면서 데이타를 두가지 관점에서 나눌 수 있습니다. 고장난 개수와 같은 개수를 셀수 있는 자료를 Discrete data 라 합니다. 반면에 온도, 무게와 같이 측정을 통해서 얻는 실수 데이타를 continuous data 라고 합니다. 이 두 종류를 구분하는 다른 방법으로 데이타 값이 어느 범위 내에서 한정된 숫자로 나타나는가 아닌가 하는 것입니다. 예를 들어 매장에서 휴대폰 판매량은 1 대 , 2 대 , 10 대와 같이 한정적인 숫자안에서 나타납니다. 하지만 연봉과 같은 데이타는 한정적이라 볼 수 있지만 너무나 숫자가 많기 때문에 continuous data 로 취급을 합니다.

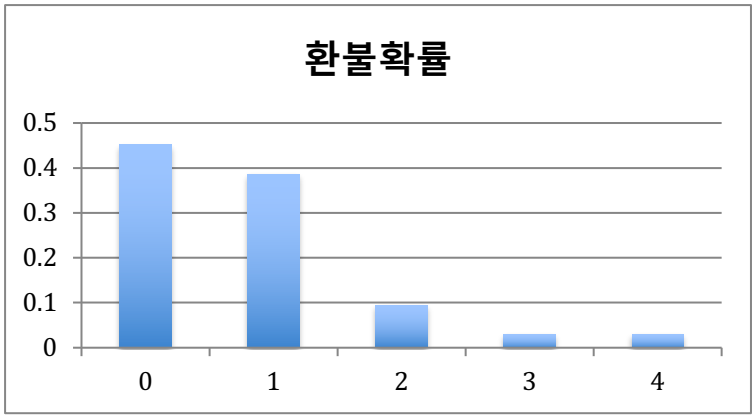
1. Discrete Probability Distributions (이산확률분포)소개

데이타를 분포를 구하기 위해서 실험을 통해서 얻어진 랜덤 값을 나타내는 random variable, x 이 있습니다. 예를 들어 동전 던지기를 할 때 random variable x 값은 H 또는 T 입니다. 확률에서는 random variable 로 부터 나온 값들의 분포가 어떠한 형태인가에 따라서 분포특징을 알고 관측된 내용에 대한 특성을 이해 할 수 있습니다.
이름이 random 인 이유는 실험 결과는 아무도 예측을 못하기 때문 입니다. 예를 들어 어느 휴대폰 매장에서 하루 휴대폰 판매량이 10 대라면 매장이 그날 마감시간에 누구도 10 대라는 것을 정확하게 예측하지 못합니다.

이러한 random variable 에 해당하는 값들에 relative frequency 를 적용하여 전체 분포를 나타낸 것이 discrete probability distribution 이라고 합니다. 여기서 relative frequency 를 probability density 라고 합니다. 예를 들어 신발 가게에서 31 일 동안 환불된 횟수에 해당하는 날이 얼마나 되는지를 알려고 할 경우를 보겠습니다.

하루 환불 개수(x)	일수	상대 횟수: P(x)	누적 분포
0	14	45.16%	45.16%
1	12	38.71%	83.87%
2	3	9.67%	93.54%
3	1	3.22%	96.76%
4	1	3.22%	100%

여기서 discrete random variable 에 해당하는 것은 하루 환불 횟수입니다. 이런 discrete probability distribution 을 그래프 형태로 표현하면 다음과 같습니다.



Discrete probability distribution 의 pdf 인 P(x)는 다음 3 가지 조건을 만족해야 합니다.

- 분포의 각 값은 다른 값과 mutually exclusive 해야 합니다. 즉 한 random variable 은 다른 random variable 에 frequency 와 섞일 수 없습니다.
- $0 \leq P(x) \leq 1$
- $\sum P(x) = 1$

jMath 로 frequency table 와 relative frequency table 을 생성하는 방법을 알아 보겠습니다. 예로 한달 동안 매일 하루 환불 개수를 나열하고 이로 부터 frequency table 을 얻는 방법은 다음과 같습니다.

```
> var a = jMath([ 0, 0, 1, 0, 1, 3, 1, 1, 0, 0, 4, 2, 1, 0, 0,  0, 1, 1, 0, 2, 0,
1, 0, 1, 0, 1, 0, 2, 0, 1, 1 ]);
> a.freqdist().toString()
"0     14
1     12
2     3
3     1
4     1"
```

jMath.prototype.freqdist 함수는 값들의 개수를 세어서 결과를 보여줍니다. 이 결과 값을 이용해서 relative frequency table 을 생성하는 것은 jMath.prototype.relfreqdist()로 얻을 수 있습니다.

```
> a.freqdist().relfreqdist().toString()
"0     0.45161290322580644
1     0.3870967741935484
2     0.0967741935483871
3     0.03225806451612903
4     0.03225806451612903"
```

jMath.prototype.freqdist 을 범위로 얻을 수 있습니다. 예를 들어 학급 성적을 A(90~100), B(80~89), C(70~79), D(60~69), F(0~59)로 나뉘어서 frequency table 을 얻는 방법은 다음과 같습니다.

```
> a = jMath([ 81, 92, 93, 83, 70, 88, 66, 63, 79, 77, 81, 51]);
> a.freqdist([0, 60, 70, 80, 90]).toString()
"0     1
60     2
70     3
80     4
90     2"
```

jMath.prototype.freqdist()함수의 입력값으로 범위의 시작값을 넣어서 이 범위에 해당하는 count 값이 증가하게 됩니다.

마지막으로 jMath.prototype.freqdist()에 입력값으로 함수를 넣어 함수에 결과 값을 frequency table 에 분류값으로 사용할 수 있습니다.

```
var a = jMath([ 81, 92, 93, 83, 70, 88, 66, 63, 79, 77, 81, 51]);
var fdist = a.freqdist(function(v){
        if ( v >= 90 ) return 'A';
        else if ( v >= 80 ) return 'B';
        else if ( v >= 70 ) return 'C';
        else if ( v >= 70 ) return 'D';
        return 'F';
    });
console.log(fdist.toString());
B       4
A       2
C       3
F       3
```

2. Mean(평균)

Probability distribution 에서 random variable, x 은 평균을 구하는 대상이고 P(x)는 x 가 존재하는 비율이기 때문에 weight average 수식 (3.2)를 이용해서 평균을 구할 수 있습니다. 이 수식에서 w 는 P(x)가 되고 분모는 모든 P(x)의 합은 1 이므로 생략이 가능합니다.

$$E(x) = \mu = \sum_{i=1}^{n} P(x_i)x_i \qquad (5.1)$$

여기서 E(x)의 의미는 expected value 로 distribution 으로 부터 예상값을 얻는 것으로 해석이 되어 붙은 이름 입니다.

환불의 예를 적용하게 되면

$$E(x) = 0.45 \times 0 + 0.39 \times 1 + 0.1 \times 2 + 0.03 \times 3 + 0.03 \times 4 = 0.806$$

결과가 의미하는 것은 하루 평균 환불되는 신발의 개수는 0.806 입니다.

jMath 를 이용하면

```
> var a = jMath([ 0, 0, 1, 0, 1, 3, 1, 1, 0, 0, 4, 2, 1, 0, 0, 0, 1, 1, 0, 2, 0,
1, 0, 1, 0, 1, 0, 2, 0, 1, 1 ]);
> a.freqdist().relfreqdist().wmean().toString()
```

```
"0.8064516129032259 0.9999999999999999"
> a.freqdist().wmean().toString()
"0.8064516129032259 31"
```

3. Variance(편차) and Standard Deviation(표준편차)

$$\sigma^2 = \sum_{i=1}^{n} P(x_i)(x_i - \mu)^2 = \sum_{i=1}^{n} P(x_i)x_i^2 - 2\mu \sum_{i=1}^{n} P(x_i)x_i + \mu^2 = E(x^2) - \mu^2 \qquad (5.2)$$

jMath 를 이용하여 계산을 하는 방법은 다음과 같습니다.

```
> var a = jMath([ 0, 0, 1, 0, 1, 3, 1, 1, 0, 0, 4, 2, 1, 0, 0,  0, 1, 1, 0, 2, 0,
1, 0, 1, 0, 1, 0, 2, 0, 1, 1 ]);
> a.freqdist().wvar().toString();
"0.9612903225806451 31"
> a.freqdist().wvar(true).toString()
"0.9302809573361082 31"
```

jMath.prototype.wvar(isPopulation) 은 variance 를 구하는 함수로 isPopulation 이 넣지 않거나 false 이면 sample variance 를 계산하고 그렇지 않으면 population variance 를 계산합니다. Standard deviation 은 jMath.prototype.wstd(isPopulation)을 이용하면 됩니다. 주의 할 점은 wmean()은 freqdist()에 결과나 relfreqdist()결과로 부터 계산 결과가 같게 나타나지만 wvar()와 wstd()는 반드시 freqdist()결과를 이용해야 합니다.

이 장에서 binomial distributions, Poisson distributions, hypergeometric, multinomial distribution 을 소개 합니다. Binomial distribution 은 단지 두가지 경우에 확률을 근간으로 개수당 확률이 나타나는 분포를 나타내고, 이를 확장하여 여러가지를 동시에 다룬것이 multinomial distribution 입니다. Poisson distribution 은 기간내에 발생되는 횟수에 대한 확률 분포를 나타냅니다. 마지막으로 hypergeometric distribution 은 확률 값이 고정되지 않고 변화될 때 발생 횟수에 대한 확률 분포를 보여 줍니다.

4. Binomial distributions

마트에서 새 음식을 시식하는 코너를 만들어 고객들에게 제공을 했을 때 시음을 하고 음식을 사는 고객과 그렇지 않은 고객으로 구분을 한다면 나오는 경우에 수는 두 가지입니다. 이러한 것을 binomial experiment 또는 Bernoulli trial 이라고 합니다.

Binominal experiment 의 특성은 단지 2 가지 경우에 성공과 실패라는 방식으로만 존재하기 때문에 성공의 확률을 p 라고 하면 실패의 확률 q 는 1-p 입니다. 이러한 확률을 기반으로 총 실험하는 횟수가 고정되어 있고 성공한 횟수에 대한 확률을 구하여 분포도를 그리는 것을 binomial distribution 이라고 합니다. Binomial experiment 의 다른 예들은 다음과 같은 것이 있습니다.

- 음식 주문을 전화로 했는가 아님 온라인으로 했는지 관찰
- 온라인 주문을 PC 에서 하는지 모바일에서 하는지 관찰
- 판매된 물건 중 환불 요청에 대한 관찰
- 길거리 설문조사에 응답하는지 관찰
- 직장내 직원이 고졸인가 대졸인가 관찰
- 국내 외국 관광객이 처음온것인가 재방문인가 관찰

이 예제들을 보시면 모두 결과값이 단지 2 가지만 있는것을 알 수 있습니다. 즉 어떤 속성이 존재하는가 아닌가를 갖고 구분을 할 수 있습니다.

예를 들어 음식점에서 외상은 없고 계산을 현금으로 하는 손님과 카드로 계산하는 손님으로 나뉠 때 현금을 낸 확률은 0.3 이면 카드로 낼 고객의 확률은 1-0.3 으로 0.7 이 됩니다. 여기서 한 손님이 현금을 낸다고 다음 손님이 카드를 내거나 현금을 내는데 영향을 주지 않기 때문에 사건들은 독립되어 있습니다. 즉, 이러한 상황에서 3 명의 손님이 있을 때 한 손님만 현금을 내고 나머지 2 명의 손님이 카드로 결제를 할 확률은 pq^2 로 0.3x0.7x0.7=0.147 이 됩니다.
그런데 이 확률은 총 3 번 반복이 됩니다. 예를 들어 A,B,C 라는 손님이 있을 때

ABC ABC ABC

회색 배경이 있는 사람이 현금을 낸 사람이라 했을 때 순서는 상관이 없이 때문에 총 3 번이 나타납니다. 즉 ${}_3C_1$ 으로 3!/2! = 3 가 됩니다.
따라서 3 명 중 한명만 현금으로 낼 확률은 앞서 구한 pq^2 를 3 개에 대한 union 으로 각각의 pq^2 는 mutually exclusive 이므로 단순히 3 을 합하면 됩니다. 즉 $3pq^2$ 가 됩니다. 이러한 과정을 수식으로 표현하여 binomial distribution 확률 값으로 나타내는 binomial probability density function 은 다음 수식과 같습니다.

$$P(x|p,n) = \binom{n}{x} p^x q^{n-x} \tag{5.3}$$

jMath 에서 binomial probability density function 를 구하는 방법은

jMath.stat.binopdf(x,n,p)

여기서 x 는 random variable 이고 n 은 총 개수, p 는 성공할 확률입니다. 예를 들어 음식점에서 현금을 낼 확률이 0.3 일때 10 명의 손님중 5 명이 현금을 낼 확률을 구하려면 다음과 같습니다.

$$\binom{10}{5} 0.3^5 (1-0.3)^{10-5}$$

```
>jMath.stat.binopdf(5,10,0.3)
0.10291934519999997
```

Binomial probability density 를 0 에서 부터 m 값까지 합을 구하게 된다면 표현은

$$P(x \le m|p,n) = \sum_{i=0}^{m} \binom{n}{i} p^i q^{n-i}$$

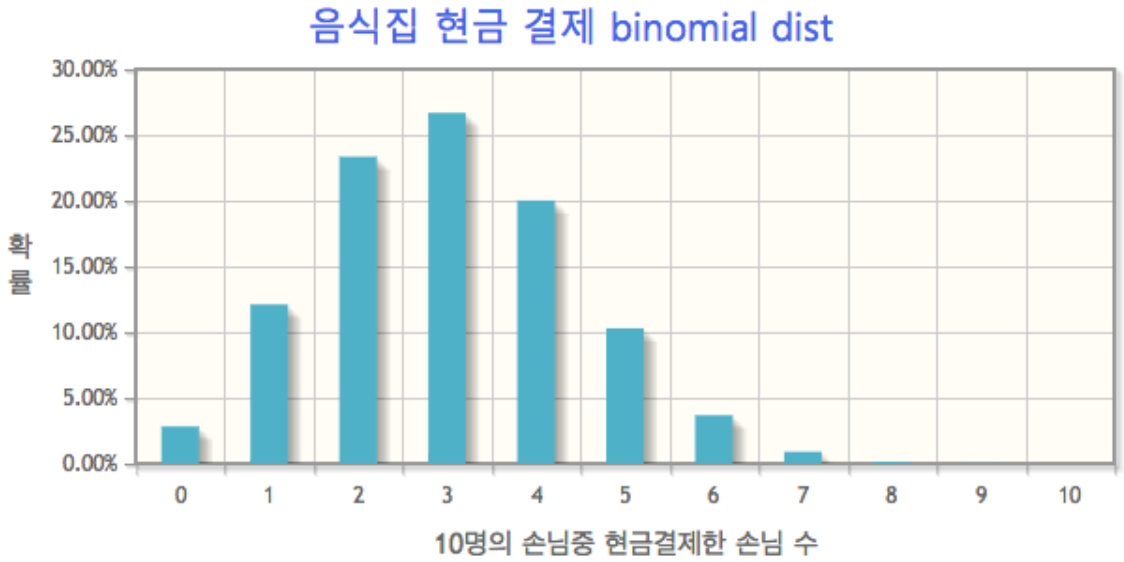
이러한 것을 cumulative distribution function(cdf)이라고 하며 M 의 범위는 0 부터 n 이 됩니다.

$$P(x \le 3) = \binom{10}{0} 0.3^0 0.7^{10} + \binom{10}{1} 0.3^1 0.7^9 + \binom{10}{2} 0.3^2 0.7^8 + \binom{10}{3} 0.3^3 0.7^7 = 0.6496$$

jMath.stat.binocdf(x,n,p)

```
>jMath.stat.binocdf(5,10,0.3)
0.6496107183999995
```

chapter05/5_binopdf.html



그래프에서 보듯이 10 명중 3 명이 현금 결제할 확률이 가장 높다는 것을 예상 할 수 있습니다. 이러한 것은 평균을 구하면 알 수 있습니다.

4.1. Mean

수식 (5.1)에 (5.3)을 적용하여 구하면 됩니다.

$$E(x) = \mu = \sum_{i=0}^{n} x_i \binom{n}{x_i} p^{x_i} q^{n-x_i} = np \sum_{i=0}^{n-1} \binom{n-1}{x_i - 1} p^{x_i - 1} q^{n-x_i} = np \tag{5.4}$$

음식점 현금 결제에 대한 평균을 구하게 되면 10x0.3 으로 3 이 됩니다. 즉 10 명중 3 명이 현금을 낼 거라는 기대가 된다는 의미입니다.

4.2. Variance and Standard deviation

각 random variable 이 평균과의 평균 거리를 계산하도록 하겠습니다.
Variance 는 다음과 같습니다.

$$\begin{aligned}\sigma^2 = E(x^2) - \mu^2 &= -n^2p^2 + \sum_{i=0}^{n} x_i^2 \binom{n}{x_i} p^{x_i} q^{n-x_i} \\ &= -n^2p^2 + np + np \sum_{i=0}^{n-1} (x_i - 1) \binom{n-1}{x_i - 1} p^{x_i - 1} q^{n-x_i} \\ &= -n^2p^2 + np + np(n-1)p = np - np^2 = np(1-p) = npq\end{aligned} \tag{5.5}$$

따라서, standard deviation 은 $\sigma = \sqrt{npq}$가 됩니다. 음식점 손님의 현금 결제를 적용을 하면 1.4491 이 됩니다.

평균, 표준편차의 결과를 얻기 위해서 jMath 를 이용하면 다음과 같습니다.

jMath.stat.bionstat(n,p)

```
> jMath.stat.binostat(10,0.3)
[3, 2.0999999999999996, 1.4491376746189437]
```

5. Poisson distributions(푸아송 분포)

특정 기간, 넓이, 거리, 또는 측정되는 단위내에서 event 가 발생되는 횟수를 세는 것을 Poisson process 라고 합니다. Poisson process 에서의 random variable 은 발생되는 횟수로 예를 들어 다음 한 시간동안 온라인 주문하는 고객의 수입니다.

Poisson process 은

- 측정되는 단위끼리는 서로 간섭이 없습니다. 즉 독립되어 있어 서로에게 영향을 끼치지 않습니다.
- 측정되는 구간마다 기대값인 평균값은 모두 같습니다.
- 마지막으로 측정되는 단위는 서로 겹쳐서는 안됩니다. 예를 들어 한 시간내에 발생되는 회수를 측정을 할 때 15:00-16:00 와 15:40-16:40 을 동시에 사용을 할 수 없습니다.

Poisson process 를 활용할 수 있는 예는 다음과 같습니다

- 한달 동안 환불을 요청하는 고객의 수
- 하루 동안 스팸 메시지를 받는 수
- 한달 동안 휴대폰 매장에서 아이폰을 판매한 수
- 1 년 동안 태풍이 한국으로 강타한 수
- 1 년 동안 판매된 제품의 고장 수

이러한 경우의 수를 random variable 로 확률 분포를 나타낸 것이 Poisson probability distribution 이고 이에 해당한는 pdf 는 다음과 같습니다.

$$P(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!} \tag{5.6}$$

여기서, x 는 측정 단위내에 측정하고자 하는 내용이 발생되는 횟수이고, λ(lambda)는 x 의 평균값입니다.

예를 들어 119 신고로 구급출동 하루 평균 횟수가 43.6 회로 시간당으로 본다면 1.82 회로 나타납니다. 그럼 다음 한 시간내로 구급 출동이 3 건 발생할 확률을 구하면 수식 5.6 에서 λ는 1.82 가 되고 x 는 3 이 되어 0.1628 이 됩니다.

jMath.stat.poisspdf(x,lambda)

```
> jMath.stat.poisspdf(3,1.82)
0.1627972095426606
```

만일 한시간 내에 4 건 이상 발생될 확률을 구하려면 수식은 다음과 같아야 합니다.

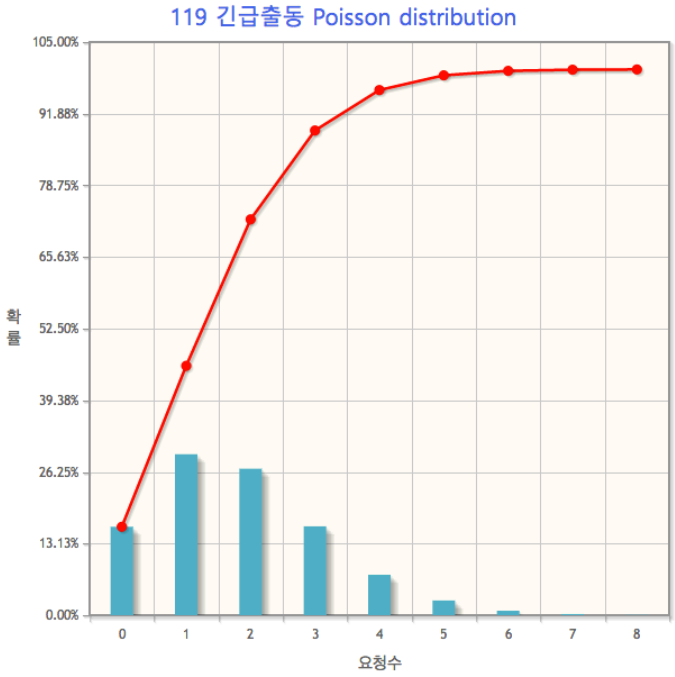
$$1 - P(x \leq 3) = 1 - \sum_{i=0}^{3} P(x_i|\mu)$$

$$1 - (P(0) + P(1) + P(2) + P(3)) = 0.1119$$

jMath.stat.poisscdf(x,lambda)

```
> 1-jMath.stat.poisspdf(3,1.82)
0.1119431241271025
```

chapter05/5_poisspdf.html



빨간색 라인은 cumulative density function(cdf)에 의한 결과로 누적된 pdf 의 값입니다.

5.1. Mean

$$E(x) = \mu = \sum_{i=0}^{\infty} x_i \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = \lambda \sum_{i=1}^{\infty} \frac{\lambda^{x_i-1} e^{-\lambda}}{(x_i-1)!} = \lambda \tag{5.7}$$

5.2. Variance and standard deviation

$$\begin{aligned} \sigma^2 = E(x^2) - \mu^2 &= -\lambda^2 + \sum_{i=0}^{\infty} x_i^2 \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \\ &= -\lambda^2 + \lambda \left(\sum_{i=1}^{\infty} (x_i - 1) \frac{\lambda^{x_i-1} e^{-\lambda}}{(x_i-1)!} + \sum_{i=1}^{\infty} \frac{\lambda^{x_i-1} e^{-\lambda}}{(x_i-1)!} \right) = -\lambda^2 + \lambda^2 + \lambda = \lambda \end{aligned} \tag{5.8}$$

$\sigma = \sqrt{\lambda}$

평균값이 클 수록 표준편차도 커지게 됩니다.

Poisson distribution 으로 binomial distribution 과 대략 비슷하게 값을 나타나게 할 수 있는데 조건은 실험 횟수 n 은 20 이상이어야 하고 성공할 확률 p 는 0.05 보다 작으면 어느정도 비슷한 값을 얻을 수 있습니다.

$$P(x|n,p) = \frac{(np)^x e^{-np}}{x!} \tag{5.9}$$

예를 들어 중국의 저가 전자제품을 생사하는 업체에서 1 년동안 판매하는 제품 100 개중에 1 개가 고장으로 제품 교환을 해준다면 1 년 동안 판매된 제품 1000 개 중 12 개가 제품 교환할 확률을 계산하려면 p 가 0.01 이고 n 이 1000 이기 때문에 수식 5.9 를 활용할 수 있습니다.

$$\frac{(10)^{12} e^{-10}}{12!} = 0.0948$$

이값을 binomial pdf 인 수식 5.3 을 적용하면

$$\binom{1000}{12} 0.01^{12} 0.99^{988} = 0.0952$$

두 값은 거의 비슷하게 나타납니다. 그럼 업체가 1 년 동안 무상 교환 개수를 95%까지 허용을 하려고 한다면 여분으로 더 생산해야 하는 개수는 다음과 같습니다.

```
> jMath.stat.poissinv(0.95, 10)
15
```

즉 1000 개중에 15 개가 교환이 있다고 가정하고 15 개의 제조 가격을 1000 개의 제품 가격에 합하여 1000 개에 대한 각 제품 가격을 측정합니다.

6. Hypergeometric distributions

Binomial distribution 과 Poisson distribution 의 공동적인 특징은 p 와 λ값이 항시 고정입니다. 하지만 측정하려는 대상의 숫자가 작고 선택된 대상은 다시는 선택되지 못하게 되면 상황이 달라집니다.

예를 들어 30 명의 참가자 중에 남자가 20 명있고 여자가 10 명이 있을 때 제비 뽑기로 8 명에게 상품을 주는데 여기서 3 명이 여자참가자일 확률을 구할 때 처음 뽑을 때의 확률은 10/30 입니다. 문제는 다음 부터 입니다. 처음 뽑힌 사람이 남자일 경우 여자가 뽑일 확률은 10/29 이고 여자이면 9/29 가 됩니다. 결과적으로 매번 뽑을 때 마다 뽑일 확률이 변화가 생겨서 binomial distribution 와 같이 p 는 1/3 으로 계속 사용하면서 확률을 얻지 못합니다. 이것이 가능하기 위해서는 대상의 수가 커서 p 값에 변화가 매우 작어야 합니다.

Hypergeometric pdf 을 구하는 것은 경우의 수를 직접 계산하면 됩니다. 앞의 예에서 30 명중 8 명을 뽑는 경우의 총수는 ${}_{30}C_8$ 이 됩니다. 이 중에서 여자가 10 명중 3 명인 총 경우의 수는 ${}_{10}C_3$ 에 남자 5 명이 당첨될 총 경우의 수 ${}_{20}C_5$ 을 곱한 수가 됩니다. 이것을 수식화 하면 다음과 같습니다.

$$P(x|N,k,n) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}} \tag{5.10}$$

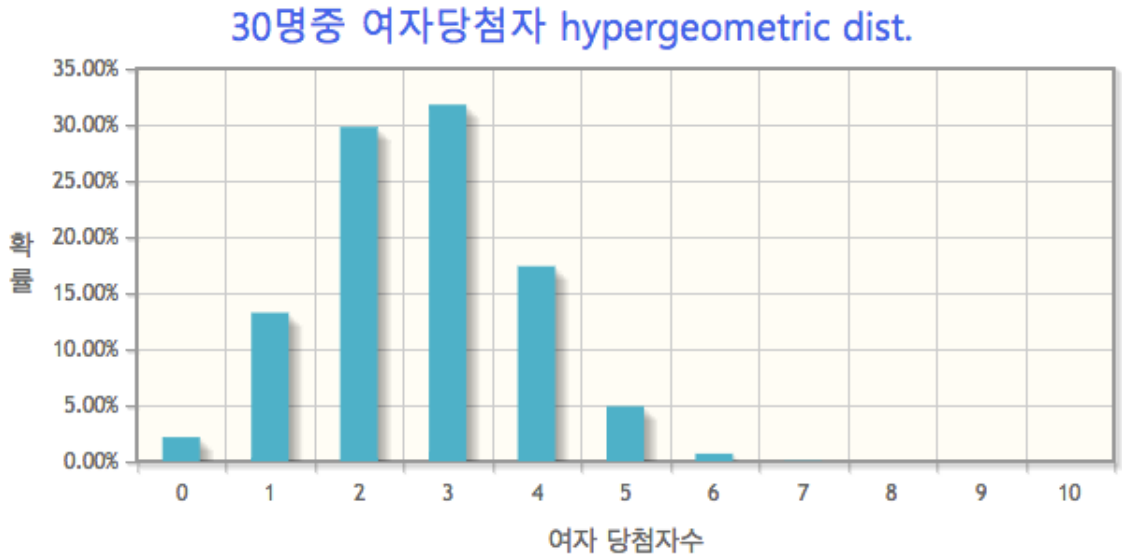
여기서 N 은 총 개수(30), n 은 총 선택 개수(8), k 는 관찰을 원하는 대상의 총 수(10), x 는 random variable 로 관찰 대상에서 선택된 수(3)이 됩니다.

$$\frac{\binom{10}{3}\binom{30-10}{8-3}}{\binom{30}{8}}$$

jMath.stat.hygepdf(x,N,k,n)

```
> jMath.stat.hygepdf(3,30,10,8)
0.3178718333141122
```

chapter05/5_hygepdf.html



6.1. Mean

$$E(x) = \mu = \sum_{i=0}^{k} x_i \frac{\binom{k}{x_i}\binom{N-k}{n-x_i}}{\binom{N}{n}} = \sum_{i=0}^{k} x_i \frac{\frac{k}{x_i}\binom{k-1}{x_i-1}\binom{N-k}{n-x_i}}{\frac{N}{n}\binom{N-1}{n-1}} = k\frac{n}{N} \tag{5.11}$$

여자 당첨자 수에 예제의 평균값은

$$10 \times \frac{8}{30} = 2.667$$

6.2. Variance and Standard deviation

$$\begin{aligned}\sigma^2 = E(x^2) - \mu^2 &= -\left(\frac{kn}{N}\right)^2 + \sum_{i=0}^{k} x_i^2 \frac{\binom{k}{x_i}\binom{N-k}{n-x_i}}{\binom{N}{n}} \\ &= -\left(\frac{kn}{N}\right)^2 + \frac{kn}{N}\left(\frac{(k-1)(n-1)}{N-1} + 1\right) = k\frac{n}{N}\frac{(N-k)(N-n)}{N(N-1)} \\ &= \mu\frac{(N-k)(N-n)}{N(N-1)}\end{aligned} \tag{5.12}$$

여자 당첨자 수에 예제의 표준 편차값은

$$\sqrt{10\frac{8}{30}\frac{(30-10)(30-8)}{30(30-1)}} = 1.1613$$

jMath.stat.hygestat(N,k,n)

결과: [평균,편차,표준편차]

```
> jMath.stat.hygestat(30,10,8)
 [2.6666666666666665, 1.3486590038314175, 1.1613177876151806]
```

7. Multinomial Distribution

Binomial distribution 은 한가지 속성에 대해서만 취하여 성공/실패, 있다/없다과 같은 두가지 결과에 분포를 취급합니다. Multinomial distribution 은 이것을 일반화한 것으로 예를 들어 휴대폰 매장으로 아이폰 구매할 확률은 0.05, 삼성 안드로이드폰은 0.5, LG 안드로이드는폰 0.35, 기타 휴대폰이 0.1 라고 하였을 때 가입 고객 40 명중에 아이폰 4 명, 삼성폰 23 명, LG 폰 10 명, 기타 다른 폰이 3 명일 확률을 알고 싶을때 multinomial distribution 을 사용할 수 있습니다.

$$P(x_1, x_2, \cdots, x_k|n) = \frac{n!}{x_1!\, x_2! \cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k} \tag{5.13}$$

여기서 $\sum p_k = 1$이고 $\sum x_k = n$이 됩니다.

여기서 각각의 random variable 에 대한 평균과 편차는 다음과 같습니다.

$$\begin{aligned} \mu_k &= np_k \\ \sigma_k^2 &= np_k(1-p_k) \end{aligned} \tag{5.14}$$

휴대폰 예를 수식 5.13 에 적용을 하면

$$\frac{40!}{4!\,23!\,10!\,3!} 0.05^4 0.5^{23} 0.25^{10} 0.1^3 = 0.000043$$