# Chapter 12 Chi-Square Test from Expectation

지금까지 한개, 두개, 두개 이상의 population 들을 검사하는 방법들을 학습을 했습니다. 이 test 를 위해서 한개 두개의 population 검사는 normal distribution 과 student t-distribution 으로 검사를 했고 두개 이상의 population 의 경우는 모든 population 이 같은가 같지 않은가를 F-distribution 을 이용해서 검사를 했습니다.

이 장에서 다룰 내용은

- 두개 이상의 population 이 비율(proportion)들이 같은가 를 조사.
- Data 의 분포가 특정 분포와 근사한가를 조사하는 fitness test.
- 데이타의 독립성 조사.

## 1. 두개 이상의 독립된 population 의 비율이 같은가 조사.

예를 들어 새로운 음료를 출시하여 연령대 별로 만족도를 조사하였습니다. 전 연령의 만족도가 다른가를 알아 보고 젊은 층에 만족도가 노년층 보다 높을 거라는 기대를 합니다. 이에 대한 관찰된 frequency table 은 다음과 같습니다.

| 연령 | 만족-YES | 만족-NO | 총인원 | YES-Proportions |
|---|---|---|---|---|
| 10 대 | 65 | 24 | 89 | 0.73 |
| 20 대 | 70 | 38 | 108 | 0.65 |
| 30 대 | 68 | 35 | 103 | 0.66 |

| | | | | |
|---|---|---|---|---|
| 40 대 | 72 | 44 | 116 | 0.62 |
| 50 대 | 61 | 45 | 106 | 0.58 |
| 60 대 | 58 | 47 | 105 | 0.55 |
| 합계 | 394 | 233 | 627 | |

1) Hypothesis test 설정

ANOVA 에 비추어 설명을 하면 factor 는 연령이고 6 개의 level 로 구성이 된 one-way ANOVA 입니다. 여기서 알고 싶은 내용은 6 개의 group 에 population 의 비율이 같지 않은가로 다음과 같이 null hypothesis 와 alternative hypothesis 를 설정을 할 수 있습니다.

$$H_0: \pi_{10} = \pi_{20} = \pi_{30} = \pi_{40} = \pi_{50} = \pi_{60}$$
$$H_1: \textit{모든 비율은 같지 않다}$$

여기서 각 $\pi_i$ 값은 $i$ 연령대별 population 만족의 비율입니다. 만일 null hypothesis 를 reject 하지 못한다면 이는 조사에 의한 비율의 차이는 random 에 의한 것이라 판단이 됩니다.

2) Significance level: 0.05

3) Test-statistic

모든 비율이 같기 위해서는 조사에 만족한 사람의 비율의 모두 같아야 합니다.

$$\overline{p} = \frac{394}{627} = 0.63$$

이는 총 조사 인원에 만족한 사람의 총 인원수의 비율로 모든 연령대가 이 비율과 차이가 멀어 질 수록 서로 다른 비율을 갖고 있을 확률이 높아 지게 됩니다. 그럼 이 비율을 이용해서 expected frequencies 는 연령별 총 조사인원에 곱을 하여 얻을 수 있습니다.

$$f_e = \textit{연령별 총인원} \times \overline{p}$$

| 연령 | 만족-YES($f_o$) | 총인원 | Expected freq.($f_e$) |
|---|---|---|---|
| 10 대 | 65 | 89 | 55.93 |
| 20 대 | 70 | 108 | 67.87 |
| 30 대 | 68 | 103 | 64.72 |
| 40 대 | 72 | 116 | 72.89 |
| 50 대 | 61 | 106 | 66.61 |
| 60 대 | 58 | 105 | 65.98 |
| 합계 | 394 | 627 | |

다음은 계산을 위해 추가로 필요한 만족하지 않는 expected frequencies 입니다.

| 연령 | 만족-NO($f_o$) | 총인원 | Expected freq.($f_e$) |
|---|---|---|---|
| 10 대 | 24 | 89 | 33.07 |
| 20 대 | 38 | 108 | 40.13 |
| 30 대 | 35 | 103 | 38.28 |
| 40 대 | 44 | 116 | 43.11 |
| 50 대 | 45 | 106 | 39.39 |
| 60 대 | 47 | 105 | 39.02 |
| 합계 | 233 | 627 | |

Chi-square test 를 위한 조건중 하나는 expected frequency 가 5 이상이어야 하는데 모든 만족과 만족하지 않는 경우의 값의 최소값이 33.07 로 chi-square 를 적용하는데 문제가 없습니다.

그럼 이렇게 얻은 값들로 chi-square test statistic 을 계산하는 수식은 다음과 같습니다.

$$\chi^2 = \sum_{i=1}^{B}\sum_{j=1}^{N}\frac{\left(f_{o(i,j)} - f_{e(i,j)}\right)^2}{f_{e(i,j)}} \tag{12.1}$$

N: level 로 이 예제에서 연령의 개수로 6 입니다.

B: frequency 의 종류의 개수로 여기서는 YES/NO 만 있기 때문에 2 입니다.

이 수식의 의미하는 것은 만일 측정된 frequency 가 기대값에 가까울 수록 0 에 가까워 지기 때문에 null hypothesis 를 reject 하지 못하게 됩니다.

| 연령 | 만족-YES $\frac{(f_o-f_e)^2}{f_e}$ | 만족-NO $\frac{(f_o-f_e)^2}{f_e}$ |
|---|---|---|
| 10 대 | 1.4709 | 2.4876 |
| 20 대 | 0.0668 | 0.1131 |
| 30 대 | 0.1662 | 0.2810 |
| 40 대 | 0.0109 | 0.0184 |
| 50 대 | 0.4725 | 0.7990 |
| 60 대 | 0.9651 | 1.6320 |
| 합계 | 3.1524 | 5.3311 |

수식 12.1 을 적용한 결과 값은 8.486 으로 이 값이 Critical value 보다 큰가를 조사하면 됩니다.

4) Critical Value

level 이 6 개가 있으므로 degree of freedom 은 5 입니다. α가 0.05 에 해당하는 chi-square statistic value 는 11.07 입니다.

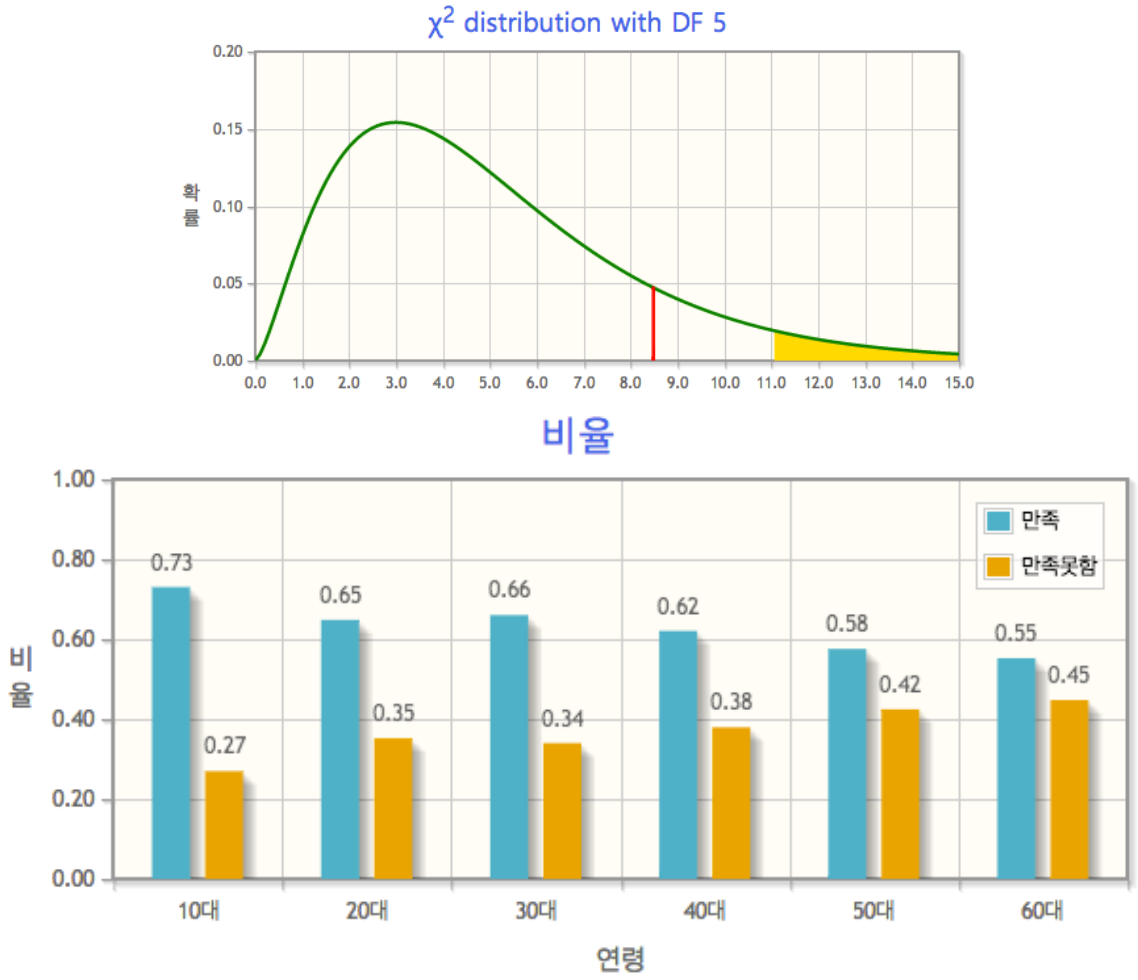
5) Null hypothesis reject 검사

Null hypothesis 를 reject 하기 위한 조건은

$$\chi^2 > \chi^2_{1-\alpha}$$

입니다. 그런데 만족도에 대한 test-statistic 은 critical value 보다 작기 때문에 null hypothesis 를 reject 할 충분한 근거가 없습니다. 또한 p-value 를 보면 0.13 으로 0.05 보다 큽니다. 이 결과의 의미는 음료수의 만족도가 연령별로 차이가 없고 젊은층이 노년층 보다 선호도가 높다는 것을 지지를 못합니다.

chapter12/props.html




jMath

jMath.fn.anova_p(alpha)

```
var data = jMath([
    [ 65, 24 ],
     [ 70, 38 ],
     [ 68, 35 ],
     [ 72, 44 ],
     [ 61, 45 ],
     [ 58, 47 ]
]);

var result = data.anova_p(0.05);
console.log(result);
```

https://github.com/handuck/jMath

```
alpha: 0.05
chi2: 8.486268602024756
chi2crit: 11.070497693516351
df: 5
expect: [ 0.63, 0.37 ]
prop: [0.73, 0.27
       0.65, 0.35
       0.66, 0.34
       0.62, 0.38
       0.58, 0.42
       0.55, 0.45 ]
pvalue: 0.1313947012015162
```

## 2. Goodness-of-fitness

이 방식은 앞서 설명된 방식과 같이 측정된 frequency 과 기대되는 frequency 로 기대값과 차가 없는 것을 증명하는 것으로 기대되는 frequency 를 안다면 수식 12.1 를 적용하여 test statistic 을 구할 수 있게 됩니다.

### 2.1. Discrete Probability Distribution

전년도 커피판매점 이용관련 직장별 비율이 금년해에 변화가 있는가를 조사하기 위해 1000 명에게 직업을 질문을 했습니다.

| 직업 | 전년도 이용비율 | 올해 조사 응답자수 |
|---|---|---|
| 직장인 | 55.5% | 570 |
| 전업주부 | 11.7% | 134 |
| 대학(원)생 | 20.1% | 185 |
| 자유직 | 7.7% | 50 |
| 기타 | 5% | 61 |

1) Hypothesis test 설정

$H_0$: 전년도와 직장별 이용 비율이 같다.

$H_1$: 전년도와 직장별 이용 비율에 변화가 생겼다.

직장별 비율의 변화가 없다면 전년도 이용비율 분포를 그대로 따를 것이고 그렇지 않다면 다른 직장별 이용비율에 변화가 생겼다는 의미입니다.

2) Significance level: 0.10

3) Test-statistic

기대 값을 계산을 하고 이를 조사값에 적용하여 수식 12.1 에 적용을 합니다.

| 직업 | 올해 조사 응답자수 | 기대값(1000x 전년도 이용 비율) | $\frac{(f_o - f_e)^2}{f_e}$ |
|---|---|---|---|
| 직장인 | 570 | 555 | 0.41 |
| 전업주부 | 134 | 117 | 2.47 |
| 대학(원)생 | 185 | 201 | 1.27 |
| 자유직 | 50 | 77 | 9.47 |
| 기타 | 61 | 50 | 2.42 |

기대 값 계산은 조사인원 1000 명에 대해 작년도 직업별 비율을 적용한 값입니다. 결과 Test statistic 값은 16.04 입니다.

4) Critical Value

Degree of freedom 계산은

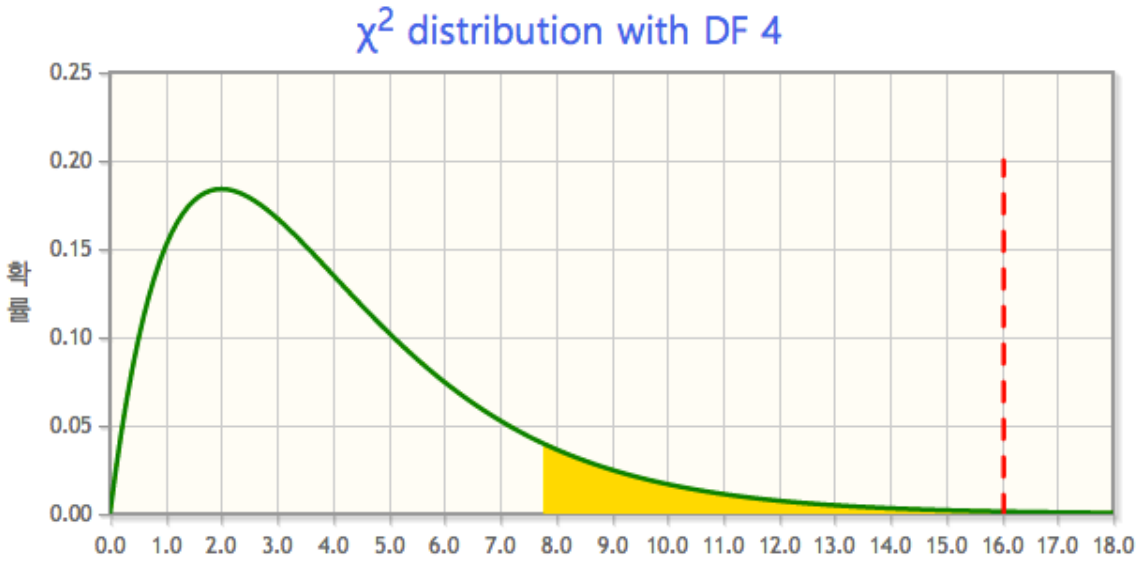
$$df = k - m - 1$$

k 는 level 의 개수이고, m 은 기대값을 계산하기 위해 필요한 parameter 가 hypothesis 에 명시되지 않아 sample 로 부터 측정한 parameter 의 개수 입니다. 이 수식을 적용을 하면 5 개의 category 에 측정된 parameter 는 없기 때문에 df 는 4 가 되어 $\chi_{0.90}$는 7.7794 입니다.

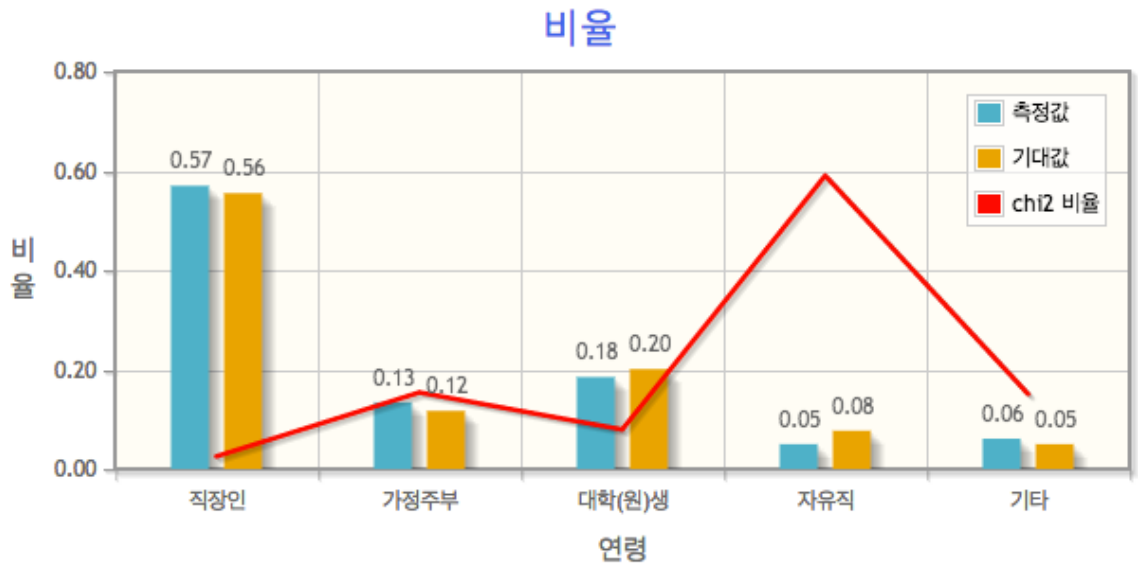
5) Null hypothesis reject 검사

Critical value 가 test statistic 보다 작기 때문에 올해의 직업별 커피판매점 이용 비율의 변화가 발생한 것에 근거가 발생 했습니다. 그래프로 비율의 변화를 보면 자유직에 커피 이용률이 전년보다 작아짐을 확인 할 수 있습니다.

chapter12/discrete.html




빨간 색 dash 라인은 test statistic 값을 의미합니다. 이 라인이 노랑색 영역에 있기 때문에 null hypothesis 를 reject 을 할 수 있습니다.



빨간색 라인은 $\frac{(f_o-f_e)^2}{f_e}$값을 test statistic 으로 나눈 값으로 즉 각각의 관측된 frequency 가 test statistic 에 영향을 미치는 비율로 이 값이 높을 수록 변화가 크게 반영이 된 것입니다. 이 예제에서 null hypothesis 가 reject 된 것은 자유직에 기대치인 작년 이용비율에 비해 측정값 즉 올해 이용비율이 낮아진 것이 변화 발생에 가장 큰 요인으로 작용하고 있습니다.

```
jMath.fn.fitness( alpha, df, expected pdf )
```

세번재 입력값인 expected pdf 는 배열이나 jMath object 로 직접 넣을 수 있고, 또한 함수를 넣어 해당 pdf 를 계산하여 pdf 를 얻게 됩니다.

```
var data = jMath([570,134,185,50,61]);
var result = data.fitness( 0.1, 4, [ 0.555, 0.117, 0.201, 0.077, 0.05] );
console.log(result);

alpha: 0.05
chi2: 16.036655183819363
chi2crit: 7.779440339734857
chi2norm:[0.025279922824209045,0.15402747279730333,0.07942004278305409,
0.5903682756167884, 0.1509042859786452]
df: 4
expect:[0.555,0.117,0.201,0.077,0.05]
expectFreq:[555,117,201,77,50]
prop:[0.57,0.134,0.185,0.05,0.061]
pvalue: 0.002970370204004702
removed: [false, false, false, false, false]
```

여기서 chi2norm 값은 $\frac{(f_o-f_e)^2}{f_e}$을 normalization 한 값입니다.

### 2.2. Poisson Distribution

Poisson distribution 은 특정 기간, 넓이, 거리, 또는 측정되는 단위내에서 event 가 발생되는 횟수가 발생할 확률 분포로, 이 검사를 통해서 관측된 값이 Poisson 분포로 측정이 되는지 알아 볼 수 있습니다. Poisson distribution 의 특징들은 다음과 같습니다.

측정되는 매 기간마다 Poisson 분포의 평균은 같습니다.
측정되는 각각의 기간에 발생되는 횟수는 독립적입니다.

Poisson process 에 정의된 기간은 겹치지 않습니다.

Fitness 를 위한 절차는 앞의 discrete distribution 과 같고 다른 점은 평균을 관측된 값에서 얻어 기대값을 구한다는 것입니다. 예를 들어 여행사가 판매된 여행 상품의 취소되는 평균을 알아 대기자수를 정하려고 합니다. 이를 위해 몇년 동안 취소되는 경우의 수를 조사하였습니다.

| 취소 개수(x) | 발생횟수($f_0$) |
|---|---|
| 0 | 20 |
| 1 | 45 |
| 2 | 56 |
| 3 | 48 |
| 4 | 30 |
| 5 | 22 |
| 6 | 16 |
| 7 | 8 |
| 8 | 2 |
| 9 | 1 |
| 합계 | 248 |

1) Hypothesis test 설정

$H_0$: 판매 여행 상품당 취소 개수가 Poisson 분포다.

$H_1$: 판매 여행 상품당 취소 개수가 Poisson 분포를 따르지 않는다.

2) Significance level: 0.05

3) Test statistic

기대값을 얻기 위해서 Poisson pdf 을 우선 만들어야 합니다.

$$P(x|\mu) = \frac{\lambda^x e^{-\lambda}}{x!} \tag{12.2}$$

여기서 λ값을 알아야 하는데 Hypothesis 에서 아무런 명시가 되어 있지않기 때문에 λ값에 상관없이 Poisson 분포를 따르는지만을 검사하는 것이기 때문에 sample 로 부터 계산을 해야 합니다. 이를 위해 Poisson distribution 에 λ값은 Poisson 분포의 평균값이기 때문에 이 값은 취소한 인원수에 대한 weighted mean 을 구하는 것과 같습니다.

$$\lambda = \frac{\sum_{i=1}^{N} x_i f_{o(i)}}{\sum_{i=1}^{N} f_{o(i)}} \tag{12.3}$$

수식 12.3 에 적용하여 얻은 값은 2.8548 이고 이를 이용해서 기대값과 test statistic 을 계산할 수 있습니다.

| 취소 개수(x) | 발생횟수($f_0$) | 기대값($f_e$) | $\frac{(f_o - f_e)^2}{f_e}$ |
|---|---|---|---|
| 0 | 20 | 14.28 | 2.295 |
| 1 | 45 | 40.76 | 0.442 |
| 2 | 56 | 58.18 | 0.081 |
| 3 | 48 | 55.36 | 0.979 |
| 4 | 30 | 39.51 | 2.290 |
| 5 | 22 | 22.56 | 0.014 |
| 6 | 16 | 10.73 | 2.583 |
| 7 | 8 | 4.38 | 2.997 |
| 8 | 2 | 1.56 | 0.123 |
| 9 이상 | 1 | 0.68 | 0.145 |
| 합계 | 248 | 248 | 11.949 |

두가지 주의함 점이 있습니다

1) 취소 개수의 최대가 9 개 이지만 Poisson 분포는 개수가 무한대까지 가기 때문에 9 에서 끝나는 것이 아니라 9 개를 포함한 그 이상을 모두 포함한 것을 마지막에 두어야 합니다.

2) 기대값이 5 보다 작게 되면 chi-square 검사를 이용할 수 없습니다. 그래서 7,8,9 을 합하여 계산을 해야 합니다.

| 취소 개수(x) | 발생횟수($f_0$) | 기대값($f_e$) | $\frac{(f_o - f_e)^2}{f_e}$ |
|---|---|---|---|
| 0 | 20 | 14.28 | 2.295 |
| 1 | 45 | 40.76 | 0.442 |
| 2 | 56 | 58.18 | 0.081 |
| 3 | 48 | 55.36 | 0.979 |
| 4 | 30 | 39.51 | 2.290 |
| 5 | 22 | 22.56 | 0.014 |
| 6 | 16 | 10.73 | 2.583 |
| 7 이상 | 11 | 6.62 | 2.889 |
| 합계 | 248 | 248 | 11.573 |

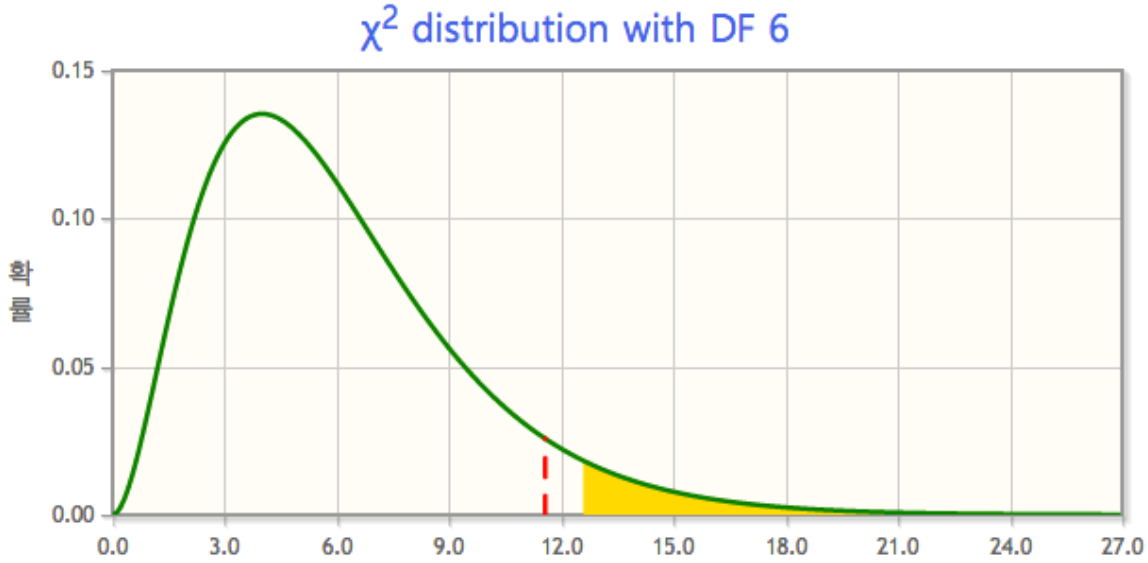
4) Critical Value

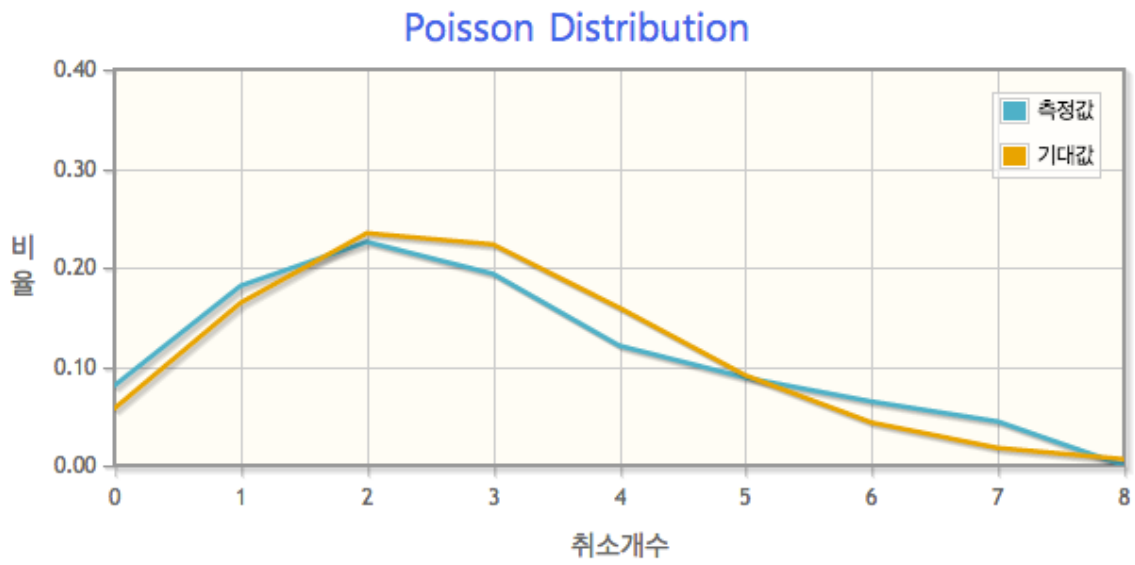
Degree of freedom 계산을 $df = k - m - 1$에서 10 개에서 테스트를 위해서 3 개를 하나로 합혀여 8 개의 경우의 수가 있기 때문에 k 는 8 이고 m 은 hypothesis test 에서 λ값이 없어 Poisson pdf 를 위해 계산을 했기 때문에 1 이 됩니다. 따라서 df 는 8-1-1 로 6 이되고 α가 0.05 인 경우 Critical value 는 12.59 입니다.

5) Null hypothesis reject 검사

Test statistic 값이 critical value 보다 작기 때문에 여행 상품 취소개수의 분포를 Poisson 분포를 따른다는 것에 대한 reject 을 하지 못하게 되었습니다. P-vlaue 도 0.07 로 0.05 보다 큰값입니다. 하지만 그렇다고 확실하게 Poisson 분포를 따른다고 확신을 할 수도 없지만 이를 통해서 여행 상품당 평균 2.85 개의 취소가 발생한다고 가정을 하고 대기자를 최대 3 명까지 받을 수 있습니다.

chapter12/poisson.html




jMath.fn.fitness_poiss(alpha, lamda)

jMath

```
jMath.prototype.fitness_poiss (alpha, lambda)
```

만일 입력값이 없다면 alpha 는 0.05 이고 lambda 는 현재 jMath object 로 부터 계산을 하게 됩니다.

https://github.com/handuck/jMath

```
var data = jMath([20, 45, 56, 48, 30, 22, 16 ,8 ,2 ,1]);
var result =  data.fitness_poiss();
console.log(result);

alpha: 0.05
chi2: 11.573342216750827
chi2crit: 12.59158724374398
chi2norm:[0.19829296602463703,0.03818388190909382,0.007032828013503457,
0.08457041193869068,0.19784915854545215,0.0012007727496311318,
0.22320763904369909, 0.24966234177529273, 0, 0]
df: 6
expect:[0.05756510506653018,0.1643390902705781,0.2345807982088091,
0.22323011442451188,0.15932149295620407,0.09096743307499393,
0.04328289154374712,0.01765223917797982,0.006299286964722637,
0.002761548311923323]
expectFreq:[14.276146056499485,40.756094387103374,58.17603795578465,
55.36106837727895,39.51173025313861,22.559923402598493,10.7341571028492867:
6.624842464747193, 0, 0]
lambda: 2.8548387096774195
prop:[0.08064516129032258,0.1814516129032258,0.22580645161290322,
0.1935483870967742,0.12096774193548387,0.08870967741935484,0.06451612903225806,
0.04435483870967742, 0, 0]
pvalue: 0.07219256763371917
removed: [false, false, false, false, false, false, false, false, true, true]
sample: { lambda: 22.8548387096774195 }
```

여기서 입력값에 7,8,9 를 합쳐지 않았지만 fitness_poiss 에서 자동으로 합쳐줍니다.

### 2.3. Binomial distribution

성공과 실패 또는 포함과 불포함과 같이 두가지 경우에만 다루는 것이 binomial experiment 입니다. Binomial 분포의 특징은

- 총 시도횟수가 n 개로 고정되어 있고,
- 성공과 실패의 확률값은 실험 전체에 고정된 값이고,
- 실험에서 성공 실패를 관측하기 위한 시도는 모두 독립적입니다. 즉, 어떤 시도의 결과로 인한 영향은 없습니다.

기대값을 알기 위해서는 Binomial pdf 를 이용하면 됩니다.

$$P(x|n) = \binom{n}{x} p^x q^{n-x} \quad (12.4)$$

Poisson distribution 과 같이 만일 성공에 확률이 주어주지 않으면 sample 에서 값을 얻어 사용해야 하고 이는 degree of freedom 계산시 영향을 미치게 됩니다.

예를 Mobile 프로그램내에 5 개의 제품에서 InAppPurchase 로 판매를 하는데, 설치 후 사용자가 제품 구매한 개수를 조사하였습니다.

| 한 고객이 구매한 제품 수(x) | 사용자 수($f_o$) |
|---|---|
| 0 | 153 |
| 1 | 544 |
| 2 | 756 |
| 3 | 536 |
| 4 | 183 |
| 5 | 28 |
| 합 | 2200 |

1) Hypothesis test 설정

$H_0$:한 고객이 구매한 제품수는 n=5 이고 구매 확률 p=0.4 인 binomial 분포를 따른다
$H_1$:한 고객이 구매한 제품수는 n=5 이고 구매 확률 p=0.4 인 binomial 분포를 따르지 않는다.

Binomial distribution 에서 평균은 np 이므로 2 개가 고객당 평균 구매한 제품의 개수입니다.

만일 p 값이 주어지지 않을 경우 평균을 sample 로 부터 얻어야 합니다. 이를 위한 방법은 평균 구매수를 총 구매가능한 수로 나누면 됩니다.

$$p = \frac{\sum_{i=1}^{n} x_i f_{o(i)}}{n \sum_{i=1}^{n} f_{o(i)}} \quad (12.5)$$

이를 적용하면 sample 로 부터 얻은 평균 구매 개수는 2.023 이고 이를 5 로 나누면 0.4047 로 고객이 모바일앱을 설치하고 제품을 구매 할 확률이 됩니다.

2) Significance level: 0.05

3) Test statistic

| 구매한 제품 수(x) | 발생횟수($f_0$) | 기대값($f_e$) | $\frac{(f_o - f_e)^2}{f_e}$ |
|---|---|---|---|
| 0 | 153 | 171.072 | 1.909 |
| 1 | 544 | 570.24 | 1.207 |
| 2 | 756 | 760.32 | 0.025 |
| 3 | 536 | 506.88 | 1.673 |
| 4 | 183 | 168.96 | 1.167 |
| 5 | 28 | 22.528 | 1.329 |
| 합계 | 2200 | 2200 | 7.310 |

기대값 계산은

$$2200 \times \binom{5}{x} 0.4^x 0.6^{5-x}$$

모든 기대값이 5 보다 크기 때문에 poisson fitness test 예제 처럼 합쳐질 필요가 없습니다. 그리고 Poisson 분포처럼 x 가 무한대가 아닌 n 까지 한정되어 있어 마지막 값 이상으로 측정할 필요가 없습니다.

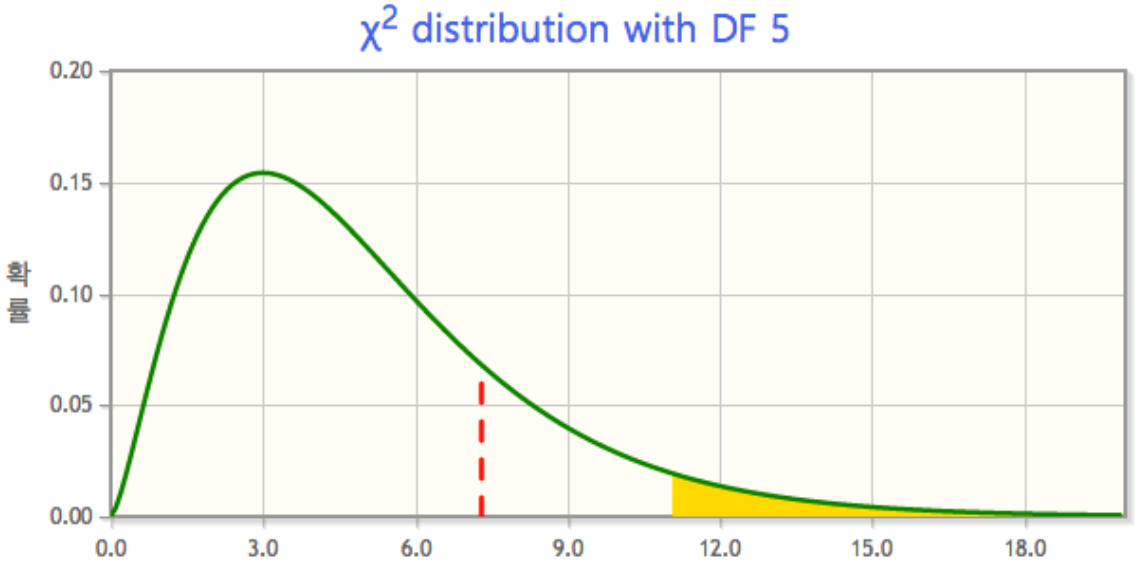
4) Critical value

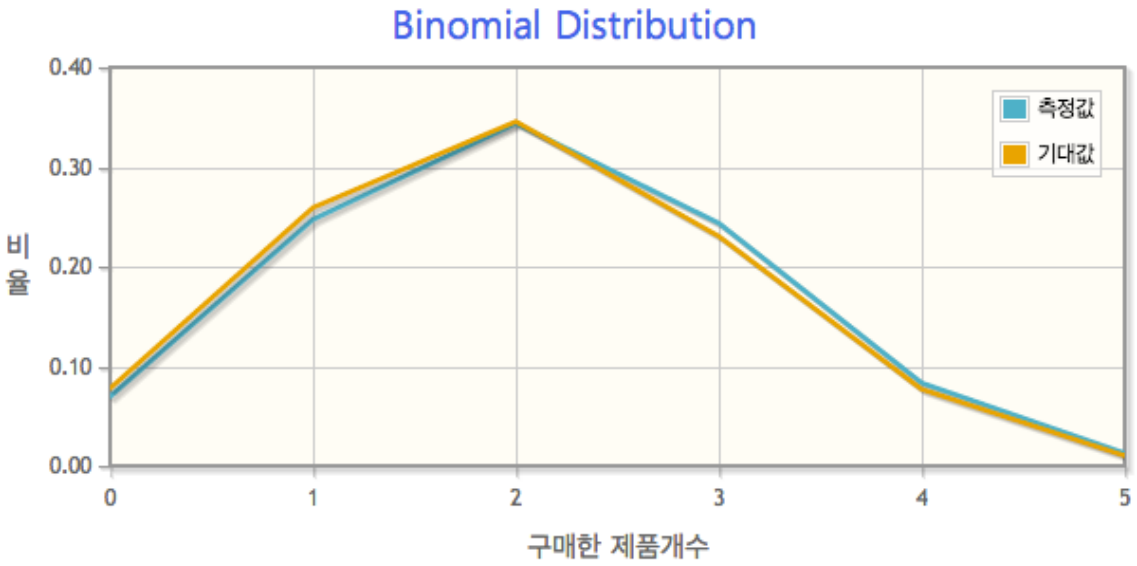
Degree of freedom 에서 0 에서 5 까지 총 6 개의 분류를 갖고 하고 sample 로 부터 측정되는 값이 없이 hypothesis 에 평균에 대한 정보가 있기 때문에 df 는 6-1 로 5 가 됩니다. 이에 해당하는 chi-square critical 값은 11.07 됩니다.

5) Null hypothesis reject 결정

Test statistic 이 critical value 보다 작기 때문에 null hypothesis 를 reject 할 근거가 없습니다. P-value 역시 0.1985 로 0.05 보다 크기 때문에, 결국, InAppPurchas 를 통해 고객이 구매할 확률은 Binomial distribution 을 따르게 되어 0.4 이되고 평균 2 개를 구매한다고 판단을 할 수 있습니다.

chapter12/binomial.html






### 2.4. Normal distribution

지금까지의 fitness 검사는 discrete probability distribution 을 다루었고 이번에는 normal distribution 으로 continuous probability distribution 에 대한 fitness 를 검사하는 방법을 배우도록 하겠습니다.

이 검사의 중요성은 Central Limit theorem 을 근간으로 정보를 얻는 방법에서 data 가 normal distribution 으로 되어 있으면 sample 의 개수의 상관없지만 그렇지 않으면 sample 의 개수가 30 개 이상이 되어야 했습니다. 하지만 계산의 편의를 위해서 normal distribution 이라는 가정을 두고 계산을 해왔습니다. 이번에 학습할 내용으로 sample 이 normal distribution 인가를 판단할 수 있도록 하여 통계처리를 할 수 있도록 도와 주는 것입니다.

Normal distribution 에 대한 검사를 위해서 필요한 것은 sample 의 정보값들을 갖고 각 bin 에 해당하는 개수를 갖는 histogram 으로 만들어야 합니다. 그리고 그 구간별 기대값을 얻어야 합니다. 이러한 계산의 편의를 위해서 z-score 로 bin 을 만들고 z-score 구간별 기대값을 계산하여 chi-square 검사를 할 수 있습니다.

예를 들어 가전제품 AS 센터에서 50 명의 고객을 대상으로 한 고객당 서비스를 받기 위해 대화하는 시간을  초단위로 기록을 했습니다.

```
259 338 312 305 321 310 262 294 294 279
367 265 281 272 253 292 289 361 290 257
364 349 291 240 282 294 311 290 318 316
250 262 270 280 287 300 179 282 350 257
337 314 299 307 237 297 364 304 302 271
```

1) Hypothesis 설정

$H_0$: 상담시간은 normal distribution 을 따른다
$H_1$: 상담시간은 normal distribution 을 따르지 않는다.

설정에서 평균과 표준편차가 없기 때문에 sample 로 부터 구해야 되고 결과 degree of freedom 계산에 영향을 미치게 됩니다.

$$\overline{x} = 294.1 \qquad s = 36.587$$

2) Significance level: 0.05

3) Test statistic

Sample 들을 갖고 histogram 을 생성해야 되기 때문에 적합한 bin 의 얻어내야 합니다. 이 과정에서 주의할 점은 expect frequency 값이 모두 5 이상 되어야 합니다.
우선 z-score 를 5 개로 나뉘어 설정을 해보도록 하겠습니다.

| Z score Interval | 확률 | 기대값 |
|---|---|---|
| z ≤ -1.5 | 0.0668 | 3.34 |
| -1.5 < z ≤ -0.5 | 0.2417 | 12.085 |
| -0.5 < z ≤ 0.5 | 0.383 | 19.15 |

| | | |
|---|---|---|
| 0.5 < z ≤ 1.5 | 0.2417 | 12.085 |
| 1.5 ≤ z | 0.0668 | 3.34 |

두 영역에서 기대값이 3 보다 작기 때문에 5 개 영역으로 나뉘는 것은 적합하지 못하고 4 개의 영역으로 나뉘었을 때 기대값을 보도록 하겠습니다.

| Z score Interval | 확률 | 기대값 |
|---|---|---|
| z ≤ -1.0 | 0.1587 | 7.932 |
| -1.0 < z ≤ 0 | 0.3413 | 17.067 |
| 0 < z ≤ 1.0 | 0.3413 | 17.067 |
| 1.0 ≤ z | 0.1587 | 7.932 |

모든 영역에 대한 기대값이 5 를 넘었기 때문에 4 개의 영역을 갖고 chi-square test 를 수행할 수 있습니다.

그럼 고객당 대화 시간을 모두 z score 로 변환을 하여 4 개의 구간별 개수를 세어 관측된 값으로 사용하여 chi-square test statistic 을 얻게 됩니다.

| Z score Interval | 발생횟수 $f_o$ | 기대값 $f_e$ | $\frac{(f_o - f_e)^2}{f_e}$ |
|---|---|---|---|
| z ≤ -1.0 | 7 | 7.932 | 0.109508825 |
| -1.0 < z ≤ 0 | 21 | 17.067 | 0.906339075 |
| 0 < z ≤ 1.0 | 14 | 17.067 | 0.5511507 |
| 1.0 ≤ z | 8 | 7.932 | 0.000582955 |
| 합계 | 50 | 50 | 1.568 |

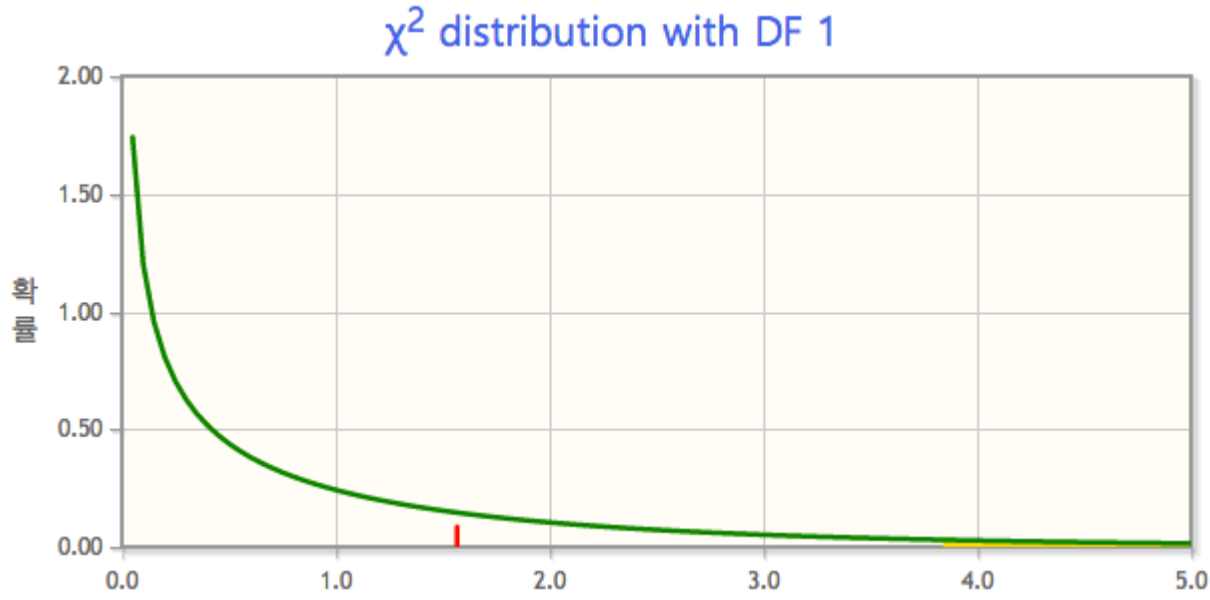
4) Critical value

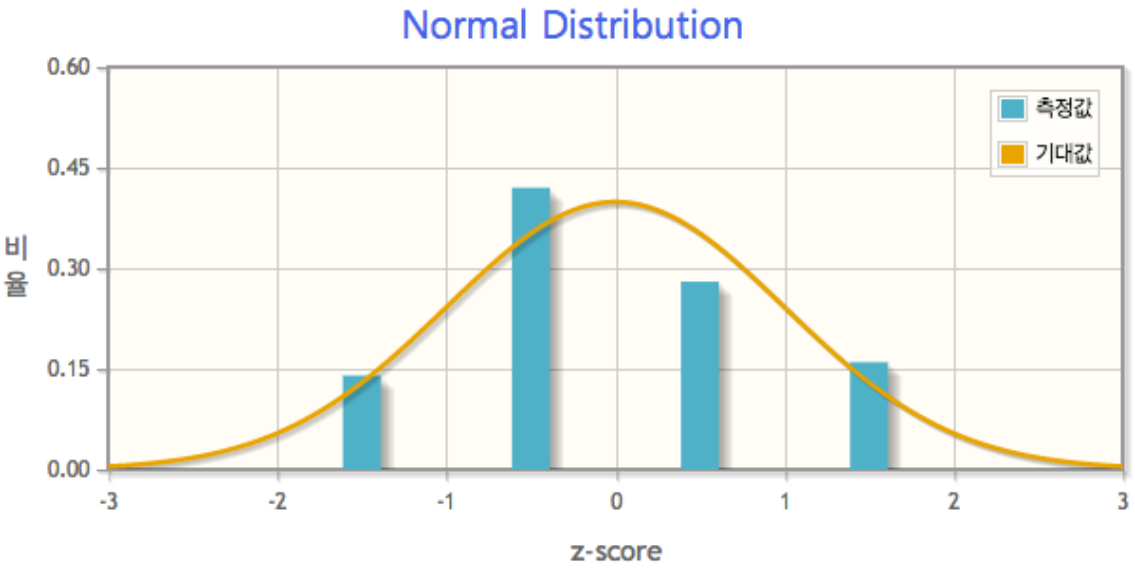
Degree of freedom 은 4 개의 category 에 평균과 표준편차의 측정으로 인해 4-2-1 로 1 이 됩니다. 이에 해당하는 critical value 는 3.84

5) Null hypothesis reject 검사

Test statistic 값이 critical value 보다작고 p-value 역시 0.21 로 0.05 보다 크기 때문에 normal distribution 이 아닐것이라는 것에 동의하지 못합니다. 결과 고객당 대화 시간은 normal distribution 을 형성하는 것으로 알 수 있습니다.

chapter12/normal.html




```
jMath.fn.fitness_normal( alpha, bins )
```

https://github.com/handuck/jMath

여기서 bins 은 나눌 구역의 개수 혹은 미리 나누어진 구역의 값들로 구역의 하단 부분에 값을 갖게 됩니다.

```
var data = jMath([259,338,312,305,321,310,262,294,294,279,
              367,265,281,272,253,292,289,361,290,257,
              364,349,291,240,282,294,311,290,318,316,
              250,262,270,280,287,300,179,282,350,257,
              337,314,299,307,237,297,364,304,302,271]);
var result =  data.fitness_norm(0.05, 4);
console.log(result);

alpha: 0.05
chi2: 1.5676929162199282
chi2crit: 3.841458820645057
chi2norm:[0.06996114035182989,0.5780577933611425,0.3516175405895128,0.000363525697
5147552]
df: 1
expect:[0.15865525393145707,0.3413447460685429,0.3413447460685429,
0.15865525393145707]
expectFreq:[7.932762696572854,17.067237303427145,17.067237303427145,
7.932762696572854]
histogram: [ [-2,7], [-1,21], [0,14], [1,8] ]
prop:[ 0.14, 0.42, 0.28, 0.16]
pvalue: 0.21054237475570303
removed: [false, false, false, false]
sample: { mu: 294.1, sigma: 36.586631884515704 }
```

## 3. Independence 검사

두가지 분류로 조사를 하여 서로의 값에 의존성이 있는지를 검사하는 방식으로 예를 들어 성별에 따른 여러 제품의 판매에 연관성이 있는가, 연령대별 판매되는 화장품 구매 경로의 연관성이 있는가, 직업군에 따른 지지하는 정당에 연관성이 존재하는가 등에 검사를 할 수 있습니다.

연령대별 판매되는 화장품 구매 경로에 대한 연관성을 조사하는 과정을 설명하면서 두 category 에 independence 를 검사하는 방법을 배워 보겠습니다.

| 연령 | 인터넷 | 전문 매장 | 홈쇼핑 | 백화점 | 마트 |
|---|---|---|---|---|---|
| 10 대 | 30 | 20 | 10 | 9 | 10 |
| 20 대 | 82 | 70 | 75 | 50 | 71 |
| 30 대 | 60 | 63 | 58 | 59 | 61 |
| 40 대 | 30 | 35 | 41 | 38 | 45 |
| 50 대 | 14 | 16 | 20 | 18 | 15 |

1) Hypothesis 설정

$H_0$: 연령과 판매장소에 연관성이 없다.
$H_1$: 연령과 판매장소는 연관성이 있다.

2) Significance level: 0.05

3) Test statistic

Chi-square 검사를 위해서 기대값을 계산하는 방법은 다음과 같습니다.

$$f_{e(i,j)} = \frac{Row_i \text{ 합} \times Column_j \text{ 합}}{\text{총관찰된 수}} \tag{12.6}$$

판매량을 수식 12.6 에 적용을 하게 되면 다음과 같습니다.

| 연령 | 인터넷 | 전문 매장 | 홈쇼핑 | 백화점 | 마트 | 합 |
|---|---|---|---|---|---|---|
| 10 대 | 17.064 | 16.116 | 16.116 | 13.746 | 15.958 | 79 |
| 20 대 | 75.168 | 70.992 | 70.992 | 60.552 | 70.296 | 348 |
| 30 대 | 65.016 | 61.404 | 61.404 | 52.374 | 60.802 | 301 |
| 40 대 | 40.824 | 38.556 | 38.556 | 32.886 | 38.178 | 189 |
| 50 대 | 17.928 | 16.932 | 16.932 | 14.442 | 16.766 | 83 |
| 합 | 216 | 204 | 204 | 174 | 202 | 1000 |

예를 들어 20 대에 홈쇼핑 구매의 기대값

$$f_{e(2,3)} = \frac{204 \times 348}{1000} = 70.992$$

모든 기대값이 5 이상이기 때문에 chi-square test 를 수행 할 수 있습니다.

$\frac{(f_o - f_e)^2}{f_e}$를 계산하면 다음과 같습니다.

| 연령 | 인터넷 | 전문 매장 | 홈쇼핑 | 백화점 | 마트 |
|---|---|---|---|---|---|
| 10 대 | 9.807 | 0.936 | 2.321 | 1.639 | 2.224 |
| 20 대 | 0.621 | 0.014 | 0.226 | 1.839 | 0.007 |
| 30 대 | 0.387 | 0.041 | 0.189 | 0.838 | 0.001 |
| 40 대 | 2.870 | 0.328 | 0.155 | 0.795 | 1.219 |
| 50 대 | 0.861 | 0.051 | 0.556 | 0.877 | 0.186 |

결과를 모두 합한 값이 test statistic 이 됩니다.

$$\chi^2 = 28.9872$$

4) Critical value

$$df = (r-1)(c-1)$$

이 예제에서 연령은 5 개로 구분되고 판매장소도 5 장소로 구분되어 df 는 (5-1)(5-1)=16 이 됩니다. 이에 대한 critical value 는

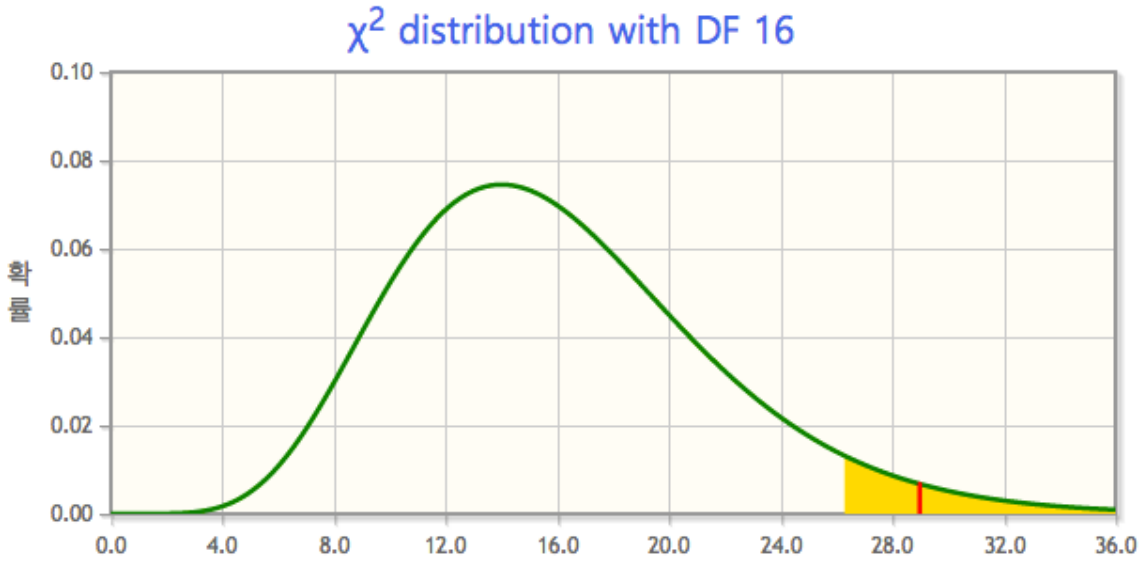
```
jMath.stat.chi2inv( 0.95, 16 )
```

26.2962 입니다.

5) Null hypothesis reject 검사

Test statistic 값이 Critical value 보다 크기 때문에 연령과 판매 장소에는 연관성이 있다고 판단이 됩니다. 10 대의 경우 인터넷으로의 구매를 선호하지만, 40 대의 경우는 반대로 인터넷 구매가 다른 경로보다 작게 나타납니다.
이러한 결론에 문제점은 null hypothesis 를 reject 했지만 확실한 것이 아니기 때문에 결론의 확신성이 없습니다. 하지만 independence test 를 통해 연령과 구매경로 사이에는 연관성이 있다는 것입니다.

chapter12/indep.html




관련성을 측정하는 방법으로 Cramer 에 의해서 소개된 statistic V 를 사용할 수 있습니다.

$$V = \sqrt{\frac{\chi^2}{N \times \min(R-1, C-1)}} \tag{12.7}$$

여기서 N 은 sample 크기이고 R 은 row 의 개수 C 는 column 의 계수 입니다. 이 값은 0 에서 1 사이 값으로 나타나는데 V 가 1 에 가까울 수록 높은 관련성이 있다는 것을 알려 줍니다. 앞의 예를 적용하면 다음과 같습니다.

$$V = \sqrt{\frac{28.9872}{1000 \times 4}} = 0.0851$$

jMath

```
jMath.prototype.test_indep(alpha)
```

```
var data = jMath([[30,20,10,9,10],[82,70,75,50,71],[60,63,58,59,61],
[30,35,41,38,45],[14,16,20,18,15]]);
var result =  data.test_indep(0.05);
console.log(result);

alpha: 0.05
chi2: 28.987264774402654
chi2crit: 26.29622760486392
df: 16
expect: [[17.06,16.116,16.116,13.746,15.958],
```

https://github.com/handuck/jMath

```
[75.168,70.992,70.992,60.552,70.296],
[65.016,61.404,61.404,52.374,60.802],
[40.824,38.556,38.556,32.886,38.178],
[17.928,16.932,16.932,14.442,16.766]]
pvalue: 0.024022133633094578
```