Chapter 11 Analysis of Variance

상점을 운영하다 보면 상황에 따라 어떤 것은 잘되고 어떤 것을 잘 되지 않습니다. 예를 들어 겨울철에 빙수판매량와 여름에 빙수판매량은 이러한 사실을 잘 보여 줍니다. 하지만 문제는 이렇게 직관적으로도 알 수 있는 상황이 아닌 경우에는 차이를 만드는 정보를 얻는것은 쉽지는 않은 일입니다. 예를 들어 음식점에서 단골 손님을 만들기 위해서 여러 가지 시도를 해보고 어떤 방식이 손님을 단골로 만들 수 있는가를 찾을때 비교 분석이 요구됩니다.

ANOVA(Analysis of Variance)는 지금까지 두개의 population 을 비교 검사하는 방식이 아닌 여러 population 을 동시에 비교 검사하는 방식으로 모든 population 에 차이점이 있는 가를 알아내는 통계방식으로 다른 비교와 같이 비교 검색 결과 population 차이가 단순 random 에 의한 것인가 아님 정말 다른 점이 있어서 나타난 것인가를 알 수 있게 해줍니다.

예를 들어 튀김닭을 판매하는 업체가 새로 개발된 4 가지의 소스를 소량으로 담은 봉지로 제공하는데 한 봉지당 한개의 조각만 사용을 할 수 있게 하여 계속 요구가 되는 소스를 추가로 주는 방식으로 4 가지 소스의 선호도에 차이를 알아 보려고 합니다.

	소스 1	소스 2	소스 3	소스 4
평균소비량	8	10	7	9

ANOVA 에 사용되는 용어로
Factor: 자료의 차이를 만드는 요인으로 예제에서 소스가 factor 가 됩니다.
Level: factor 에 속한 분류를 level 이라고 하는데 예제에서는 4 개의 level 로 구성되어 있습니다.

이 장에서 다룰 ANOVA 검사는 다음의 3 가지로 구분되어 있습니다.

1) **One-way ANOVA**: 하나의 factor 에 속한 여러 level 의 평균의 차이를 비교 분석하는 방법입니다. 튀김 닭 소스의 소비자 선호에 차이가 존재하는 하는 가를 알아 낼 때 사용됩니다.

2) **Randomized block ANOVA**: level 에 값들이 성별, 나이별로 분류해서 구분 되어 질 수 있을 경우 이러한 구분된 요소들에 영향이 level 별 차이에 영향을 미칠 때 이를 제거하여 level 별 차이점을 알아 볼 수 있게 합니다. 예를 들어 튀김 닭 소스의 소비자 선호에서 연령별 차이가 있을 수 있기 때문에 연령별 차이로 인한 영향을 제거하고 소스의 선호도가 같은지 다른지를 알아 볼 수 있습니다.

3) **Two-way ANOVA**: 두개의 factor 들에 level 들 차이를 동시에 비교 분석하는 방법입니다. 예를 들어 프랜차이점을 두 곳을 운영하는데 각 상점에서 대표 음식 3 개의 판매량을 비교 분석을 할 경우 사용을 합니다. 여기서 factor 은 상점과 음식이고 level 의 개수는 각각 2 개와 3 개 입니다.

이 장에서 ANOVA 를 설명하기 위한 다음의 조건이 만족한다는 가정하게 설명을 합니다.

- 모든 group 에 population 은 normal distribution 을 따릅니다. (sample 크기가 20 개 이상이면 충분하다고 봅니다.)
- 각 group 은 다른 group 과 독립적으로 연관성이 없습니다.
- 모든 group 에 variance 는 다 같다.

1. One-way ANOVA

하나의 factor 가 자료에 영향을 미치는가를 조사하는 방법으로, 예를 들어 대형 마트에서 물건에 위치에 따른 판매량의 차이점이 있는가 조사를 하려고 합니다. 여기서 factor 는 진열위치가 되고 각 진열위치별 평균의 차이가 존재하는가를 one-way ANOVA 를 통해 알 수 있습니다.

	계산대앞	음료수진열대	과자진열대	주류판매대
1 주	81	79	75	77
2 주	82	78	73	72
3 주	88	80	78	75
4 주	90	82	82	76
5 주	76	81	83	77
평균	83.4	80	78.2	75.4

Hypothesis test 의 설정은 모든 population 에 평균 개수가 같은가 입니다.

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$$
$$\mathrm{H}_1: \text{모든 평균이 같지 않다.}$$

평균의 차이를 갖고 통계 처리를 하기 때문에 모든 population 에 sample 개수는 같을 필요는 없습니다.

만일 null hypothesis 가 reject 될 근거가 나타난다면 population 중 어떤 것은 다른 것과 차이가 있다는 뜻으로 음료수 판매대의 위치별 판매의 경우 factor 인 위치에 따른 판매 차이가 있다는 것을 지지하는 것입니다.

5 주간 위치별 음료수 판매 평균 개수를 보시면 차이가 있다는 것을 바로 알 수 있는데 이것을 one-way ANOVA 로 확인하도록 하겠습니다.

판매된 음료 개수에 대한 정보는 다음과 같이 판매량을 분해한 linear additive model 로 표현 될 수 있습니다.

$$y_{ij} = \mu + \tau_i + e_{i,j}$$

i 는 위치를 말하며 j 는 해당 위치에 몇번째 주인가를 알려 줍니다.
이 수식의 의미는 판매되는 음료의 개수 y_{ij} 는

μ: 총 판매된 음료 개수의 평균값
τ_i: 위치로 인한 판매량 영향값 $\mu_i - \mu$ (평균 0)
e_{ij}: random error(평균 0 에 population 의 표준편차를 만드는 원인)

One-way ANOVA 은 편차계산 방식으로 차이를 알아 내는 것으로 다음 3 가지의 편차계산 방식이 필요합니다.

1) 전체 sample 의 편차: Sum of Squares Total(SST), Mean ST(MST)
2) Group 에 대한 Sample 평균들의 편차: Sum of Squares Between(SSB), Mean SB(MSB)
3) Group 각각의 편차: Sum of Squares Within(SSW), Mean SW(MSW)

Sum of Squares Total 은 Sum of Squares Between 과 Sum of Squares Within 으로 나뉘어 지게 되어 두개의 값만 알면 다른 하나의 값은 자동으로 계산이 됩니다. One-way ANOVA 로 알아보는 정보인 각 group 간 평균이 다른가는 group 평균과 전체 평균에 대한 거리에 합(SSB)에 group 각각의 평균과 내부 값들의 거리의 합(SSW)의 비율이 클 수록 평균이 다른다는 것을 지지하게 됩니다. 그럼 각각의 항목을 계산하는 방법을 알아 보겠습니다.

$$SST = \sum_{i=1}^{N} \sum_{j=1}^{M_i} \left(x_{ij} - \bar{\bar{x}} \right)^2 \tag{11.1}$$

여기서 N 은 level 의 개수로 즉 비교하려는 sample 들의 개수 입니다. M_i 는 i 번째 sample 의 크기입니다. $\bar{\bar{x}}$는 sample 전체에 대한 평균입니다.

SST 에 degree of freedom 으로 나눈 값이 MST 입니다.

$$Degree\ of\ freedom = \left(\sum_{i=1}^{N} M_i\right) - 1 = M_T - 1 \tag{11.2}$$

$$MST = \frac{SST}{M_T - 1} \tag{11.3}$$

수식 11.1 은 다음과 같이 정리가 될 수 있습니다.

$$\begin{aligned} SST &= \sum_{i=1}^{N}\sum_{j=1}^{M_j}\left(\left(x_{ij} - \overline{x}_i\right) + \left(\overline{x}_i - \overline{\overline{x}}\right)\right)^2 = \sum_{i=1}^{N}\sum_{j=1}^{M_i}\left(x_{ij} - \overline{x}_i\right)^2 + \sum_{i=1}^{N}\sum_{j=1}^{M_i}\left(\overline{x}_i - \overline{\overline{x}}\right)^2 \\ &\quad + 2\sum_{i=1}^{N}\sum_{j=1}^{M_i}\left(x_{ij} - \overline{x}_i\right)\left(\overline{x}_i - \overline{\overline{x}}\right) = \sum_{i=1}^{N}\sum_{j=1}^{M_i}\left(x_{ij} - \overline{x}_i\right)^2 + \sum_{i=1}^{N} M_i\left(\overline{x}_i - \overline{\overline{x}}\right)^2 \\ &= \sum_{i=1}^{N}(M_i - 1)s_i^2 + \sum_{i=1}^{N} M_i\left(\overline{x}_i - \overline{\overline{x}}\right)^2 = SSW + SSB \end{aligned}$$

다시 정리하면 다음과 같습니다.

$$\begin{aligned} SSB &= \sum_{i=1}^{N} M_i\left(\overline{x}_i - \overline{\overline{x}}\right)^2 \\ \hat{\sigma}_B^2 &= MSB = \frac{SSB}{N-1} \end{aligned} \tag{11.4}$$

$$\begin{aligned} SSW &= \sum_{i=1}^{N}\sum_{j=1}^{M_i}\left(x_{ij} - \overline{x}_i\right)^2 \\ \hat{\sigma}_W^2 &= MSW = \frac{SSW}{M_T - N} \end{aligned} \tag{11.5}$$

SSB 는 전체 평균과 각 group sample 평균의 거리의 제곱에 합이고 SSW 는 각 group 별 sample 의 평균과 sample 과의 거리의 제곱의 합입니다. MSB 와 MSW 는 각각의 degree of freedom 으로 나눈 값입니다.

만일 one-way population 조사 결과 null hypothesis 를 reject 하지 못한다면 모든 sample 은 동일한 population 으로 부터 온것으로 판단이 되어 MST 가 sample variance 로 되고, 또한 null hypothesis 인 모든 평균이 같다는 것이 참인 경우에 MSB 가 population variance 를 측정하기 위한 값이 됩니다. 반면에 MSW 는 null hypothesis 가 참이건 alternative hypothesis 가 참이건 상관없이 population variance 을 측정하기 위한 값이 됩니다.

이제 앞서 설명드닌 수식을 음료수 위치별 판매량 예제에 적용하여 SST, SSB, SSW 를 계산하도록 하겠습니다.

SST 계산은 각 sample 값에 전체 평균에 차에 제곱을 하여 더하면 됩니다. 여기서 전체 평균 $\overline{\overline{x}}$ 값은 79.25 입니다.

$$x_{ij} - \overline{\overline{x}}$$

	계산대앞	음료수진열대	과자진열대	주류판매대
1 주	81-79.25	79-79.25	75-79.25	77-79.25
2 주	82-79.25	78-79.25	73-79.25	72-79.25
3 주	88-79.25	80-79.25	78-79.25	75-79.25
4 주	90-79.25	82-79.25	82-79.25	76-79.25
5 주	76-79.25	81-79.25	83-79.25	77-79.25

$$\left(x_{ij} - \overline{\overline{x}}\right)^2$$

	계산대앞	음료수진열대	과자진열대	주류판매대
1 주	3.0625	0.0625	18.0625	5.0625
2 주	7.5625	1.5625	39.0625	52.5625
3 주	76.5625	0.5625	1.5625	18.0625
4 주	115.5625	7.5625	7.5625	10.5625

5 주	10.5625	3.0625	14.0625	5.0625

위의 값을 모두 합하게 되면 SST 되고 총 20 개의 sample 이 있으므로 degree of freedom 은 19 가 됩니다. SST=397.75, MST=20.93

SSB 는 각 group 별 sample 의 평균에 전체 평균의 거리에 제곱하여 각각의 sample 크기에 곱에 합을 하면 됩니다.

$$\overline{x}_i - \overline{\overline{x}}$$

	계산대앞	음료수진열대	과자진열대	주류판매대
개수(M_i)	5	5	5	5
$\overline{x}_i - \overline{\overline{x}}$	83.4 - 79.25	80 - 79.25	78.2 - 79.25	75.4 - 79.25

$$M_i \left(\overline{x}_i - \overline{\overline{x}}\right)^2$$

	계산대앞	음료수진열대	과자진열대	주류판매대
계산 값	86.1125	2.8125	5.5125	74.1125

위의 값을 모두 합하여 SSB 를 구하면 168.55 가 되고 degree of freedom 은 N-1 로 N 이 4 이므로 3 이 됩니다. 따라서, MSB 값은 56.183 이 됩니다.

SSW 는 위와 같이 계산을 해도 되지만 SST 와 SSB 를 이미 계산 했기 때문에 SSW 는 SST-SSB 가 되어 229.2 가 됩니다. Degree of freedom 은 전체 sample 개수 20 개에 group 개수가 4 개 이므로 16 이 되어 MSW 값은 14.325 가 됩니다.

이 결과 값으로 부터 null hypothesis 를 reject 을 할 수 있는지 결정을 내려야 하는데 이를 위한 것이 F-test statistic 입니다.

$$F_{\overline{x}} = \frac{\hat{\sigma}_B^2}{\hat{\sigma}_W^2} = \frac{MSB}{MSW} \tag{11.6}$$

예제의 결과를 수식 11.6 에 적용을 하게 되면

$$F_{\overline{x}} = \frac{56.183}{14.325} = 3.92$$

앞의 결과를 다시 정리를 하면 다음과 같습니다.

	SS	DF	Mean SS	F
Between	SSB	$N-1$	$MSB = \frac{SSB}{N-1}$	$F_{\overline{x}} = \frac{MSB}{MSW}$
Within	SSW	$M_T - N$	$MSW = \frac{SSW}{M_T - N}$	
Total	SST	$M_T - 1$	$MST = \frac{SST}{M_T - 1}$	

여기서 SS 는 sum of square 의 약자이고 DF 는 degree of freedom 의 약자 입니다.

예제의 값을 적용한 결과는 다음과 같습니다.

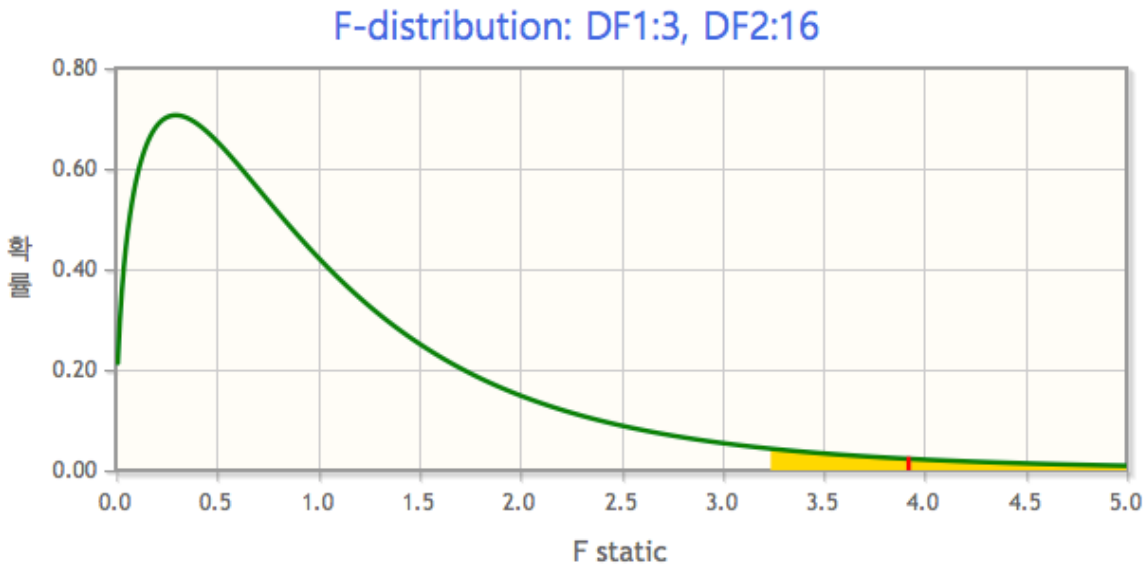
	SS	DF	Mean SS	F
Between	168.55	3	56.183	3.92
Within	229.2	16	14.325	
Total	397.75	19	20.93	

만일 평균에 차이가 없다면 즉 전체를 하나로 봐도 문제가 없으면 MSW 와 MSB 는 거의 같은 값으로 나타나 F-test statistic 값은 1 에 가까워지고 그렇지 않다면 1 에서 멀어 지게 됩니다.

Critical value 를 얻기 위해서 F-distribution 으로 부터 값을 얻어야 합니다. 계산을 위해 필요한 값은 Between 의 DF 와 Within 의 DF 입니다. F-distribution 에서 주의점은 one-tail 로 alpha 값이 0.05 이라면 0.95 인 지점에 F-test 값을 얻는 것이고 두개의 입력값인 degree of freedom 의 순서에 따라서 다른값이 되므로 주의해야 합니다. 예를 들어 alpha 가 0.05 일 때 첫번째 DF 가 3 이고 두번재 DF 가 2 이면 19.164 가 되지만 반대로 넣어 첫번째 DF 값이 2 이고 두번째 DF 가 3 일 경우 9.552 가 되어 다른 값이 나타납니다.

이 예제의 경위 F-distribution 을 위해 첫번째 DF 값이 3 이고 두번째가 16 으로 F-distribution 을 보면 다음과 같습니다.

chapter11/oneway.html



여기서 5%의 영역은 노랑색 부분으로 F 값 3.92 를 포함하고 있음을 알 수 있습니다. 이 의미는 null hypothesis 인 장소마다 판매량이 같다라는 것을 reject 할 근거가 됩니다. 즉 장소라는 factor 는 판매량에 영향을 미친는 것으로 판단이 됩니다.

F-test statistic 이 클 수록 group 의 차이가 있다는 것을 지지하게 되는데 이를 위해서는 MSB 가 크거나 혹은 MSW 가 작아야 합니다. 즉 MSB 가 크다는 의미는 group 간 차이가 확연히 있다는 것을 의미하고, MSW 가 작다면 group 내부적으로 값들이 뭉쳐있는 정도가 크기 때문에 group 간 차이를 더욱 잘 표현해 줍니다.

jMath

```
jMath.stat.anova1( alpha, sample1, sample2, … )
```

```
var s1 = jMath('81 82 88 90 76');
var s2 = jMath('79 78 80 82 81');
var s3 = jMath('75 73 78 82 83');
```

https://github.com/handuck/jMath

```
var s4 = jMath('77 72 75 76 77');
var result = jMath.stat.anova1(0.05, s1,s2,s3,s4);
console.log(result);
F: 3.9220477021524145
Fcrit: 3.238871522295805
alpha: 0.05
between: { df: 3, ms: 56.18333333333334,  ss: 168.55 }
mean: 79.25
means: [ 83.4, 80, 78.2, 75.4 ]
numSamples: 20
pvalue: 0.028300715124410458
total: { df: 19, ms: 20.93421052631579, ss: 397.75 }
within:{ df: 16, ms: 14.325, ss: 229.2 }
compare :{...}
```

만일 ANOVA 검사 방식이 아닌 단순히 2 개의 Group 별로 비교를 한다고 했을 경우에 다음과 같은 문제가 있습니다.

5%의 type error I 으로 t-test 를 할 경우 null hypothesis 를 reject 하지 않을 확률은 각각이 95%입니다. 그런데 여기서 N 개의 group 을 쌍을 이루게 되면 총 쌍의 개수는

$$\binom{N}{2}$$

개가 되고 이러한 조합으로 총 비교를 하여 전체에 평균이 같은가인 null hypothesis 를 reject 하지 않을 확률은

$$0.95^{\binom{N}{2}}$$

만일 N 이 4 개이면 총 6 가지의 비교가 있을 수 있기 때문에 0.95^6 은 0.735 가 됩니다.

즉 type I error 인 null hypothesis 가 잘못되어도 reject 을 하지 못할 확률은 0.265 로 0.05 보다 5 배게 됩니다. 하지만 ANOVA 를 이용하면 F test statistic 으로 type I error 를 0.05 가 되어 보다 더 정확한 결과를 얻게 됩니다.

2. Multiple Comparisons for one-way ANOVA

ANOVA 검사는 전체 group 에 차이가 있고 없음을 알 수 있지 만일 차이가 있다면 어떤 group 이 다른 group 과 차이가 있는지 가리지는 못합니다. 이를 위한 방법들을 소개하도록 하겠습니다.

2.1. Least Significant Difference (lst)

이 방법은 결과를 기반으로 비교를하는 것이 아니라 미리 정해놓은 group 끼리 하나 혹은 두개의 비교를 수행할 때 사용합니다. 방법은 두 개의 population 이 같은 population 편차를 갖고 있을 때 평균이 같은가를 판다하는 t-test 검사와 동일합니다.

$$lsd(\alpha) = t_{1-\alpha/2,df}\sqrt{\hat{\sigma}_w^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} = t_{1-\alpha/2,df}s_d \tag{11.7}$$

여기서

n_1 과 n_2: 비교할 할 group 에 sample 의 개수들,

df: M_T-N

$\hat{\sigma}_W^2$: MSW.

이 값으로 두 group 간의 차별을 알기 위해서는 t-score 로 비교를 하면됩니다.

음료수 판매에서 계산대 앞과 주류판매 되는 곳을 비교했을 때 다음과 같은 결과가 나타난다면 판매량에 차이가 있다고 할 근거가 있게 됩니다.

$$\left|\frac{\overline{x}_1-\overline{x}_4}{s_d}\right| > \left|t_{\alpha/2,16}\right|$$

여기서 df 가 16 인 이유는 총개수가 4x5 로 20 개이고 group 이 4 개가 있기 때문입니다.

α를 0.05 로 했을 때 수식을 적용하면 결과는 다음과 같습니다.

$$\frac{83.4 - 75.4}{\sqrt{14.325\left(\frac{1}{5}+\frac{1}{5}\right)}} = 3.342$$

로 critical value 값인 2.12 보다 큰 값이 되어 두 group 은 같지 않다는 판단될 수 있습니다.

2.2. Tukey-Kramer Multiple Comparison Test

이를 위한 절차는 모든 가능한 두 쌍의 group 들에 평균의 절대값이 critical value 보다 크면 두 group 에 차이가 있을 수 있음을 알려 주는 것입니다. 이에 대해서 Hypothesis test 설정은 다음과 같습니다.

$$H_0: \mu_i = \mu_j$$
$$H_1: \mu_i \neq \mu_j$$

만일 총 N 개의 group 으로 있을 경우(level 의 개수가 N 개) 총 쌍의 개수는 combination 으로 구할 수 있습니다.

$$\binom{N}{2} = \frac{N(N-1)}{2}$$

위치별 음료수 판매의 차이를 위치별로 비교를 하게 된다면 level 이 4 개 이기 때문에 총 6 개의 비교할 위치의 쌍이 존재하게 됩니다.

여기서 test-statistic 에 해당하는 값은 group 의 평균의 차에 절대값입니다.

$$\left|\overline{x}_i - \overline{x}_j\right| \tag{11.8}$$

비교를 위한 Critical value 은

$$CR_{i,j} = Q_\alpha \sqrt{\frac{\hat{\sigma}_w^2}{2}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)} \tag{11.9}$$

여기서

$\hat{\sigma}_w^2$: MSW(mean square within)

n_j: j 에 해당하는 group 의 sample 크기.

Q 값은 α에 해당하는 Studentized range 의 critical value 입니다. F-distribution 과 같이 두개의 degree of freedom 이 필요한데 첫번째 degree of freedom(DF1)은 group 의 개수 즉 level 의 개수이고 두번째 degree of freedom(DF2)은 MSW 의 것과 같은 $M_T - N$입니다. 위치별 음료수 판매의 예제에서 DF1 은 4 이고 DF2 는 20-4 로 16 이 됩니다. α 는 Type I error 에 해당하는 값으로 0.05 로 계산을 합니다. 이 예제를 위한 Q_α값은 4.0461 이 됩니다.

jMath 에서 Q_α를 계산하기 위한 함수는

```
jMath.stat.stdrinv( p, df1, df2 )
```

```
> jMath.stat.stdrinv(0.95,4,16)
4.046093203525906
```

두 쌍의 group 의 차이가 있고 없음을 다음과 같이 합니다.

조건식	결론
$\lvert\overline{x_i} - \overline{x_j}\rvert \le CR_{i,j}$	두 개의 group 이 같다는것에 확신
$\lvert\overline{x_i} - \overline{x_j}\rvert > CR_{i,j}$	두 개의 group 이 다르다는 것에 확신

예제를 적용을 하면 모든 group sample 의 크기가 같기 때문에 CR 값 모두 같은 값입니다.

$$CR = 4.046\sqrt{\frac{14.325}{2} \times \left(\frac{1}{5} + \frac{1}{5}\right)} = 6.8485$$

6.8485 은 모두 같습니다.

i,j	$\lvert\overline{x_i} - \overline{x_j}\rvert$	결론
1,2	3.4	같음
1,3	5.2	같음
1,4	8	다름
2,3	1.8	같음

2,4	4.6	같음
3,4	2.8	같음

판매량의 차이가 있는 곳은 유일하게 계산대앞과 주류판매대로 나타납니다. 나머지 조합들은 차이가 있다고 판단할 근거가 없습니다.

2.3. Scheffe's test

Linear comparision 방식으로 group 의 평균값들을 조합하여 비교를 하는 방식을 제공합니다. 방법은 다음과 같습니다.

1) Linear contrast: 모든 평균의 합을 0 으로 만들도록 coefficient 값을 설정합니다.

$$\sum_{i=1}^{N} C_i \overline{x_i}\text{, 여기서 } \sum_{i=1}^{N} C_i = 0 \tag{11.10}$$

2) Hypothesis

$$H_0 : \sum_{i=1}^{N} C_i \mu_i = 0$$

$$H_1 : \sum_{i=1}^{N} C_i \mu_i \neq 0$$

3) Confidence Interval

$$\sum_{i=1}^{N} C_i \overline{x_i} \pm \sqrt{\hat{\sigma}_w^2 (N-1) F_{1-\alpha, N-1, M_T - N} \sum_{i=1}^{N} \frac{C_i^2}{n_i}} \tag{11.11}$$

Confidence Interval 에 0 이 포함되어 있다면 Null hypothesis 는 reject 됩니다.

예를 들어 계산대앞과 주류판대매앞에 대한 비교를 적용을 하면
Fcrit 값은 One-way ANOVA 에서 수행한 결과 값과 같기 때문에 그대로 사용하면 되고
평균의 linear contrast 는

$$\sum_{i=1}^{N} C_i \overline{x}_i = 1 \times 83.4 - 1 \times 75.4 = 8$$

Confidence Interval 은

$$8 \pm \sqrt{14.325 \times 3 \times 3.239 \times \frac{2}{5}} = [0.538, 15.462\ ]$$

0 이 포함되어 있지 않기 때문에 두 장소의 판매량이 다르다는것을 지지할 수 있습니다.

3. Randomized Block ANOVA

One-way ANOVA 는 group 별 평균을 보기 때문에 group 에 있는 속성을 잃어 버려게 되어 자료값이 순서나 그 내부적인 특징에 의한 차이로 인한 영향으로 group 별 차이를 알 수 없게 만들 수 있습니다. 이유는 이러한 요인들은 One-way ANOVA 에서는 random error 로 취급하되어 정확한 판단을 어렵게 만듭니다.

예를 들어 휴대폰 악세사리가게에서 안드로이드 폰에 사용가능한 4 가지 다른 종류의 휴대폰 케이스 판매량을 비교 분석하려고 할 때 구매한 연령별로 판매량을 구분하여 기록된 자료가 다음과 같습니다.

Block₩Group	케이스 1	케이스 2	케이스 3	케이스 4
10 대	48	40	39	40
20 대	32	39	29	24
30 대	42	45	29	39
40 대	46	36	29	39
50 대	38	35	32	31
평균	41.2	39	31.6	34.6

여기서 연령별로 구분을 했기 때문에 판매량의 순서는 각 group 별로 유지를 해야 합니다. 예를 들어 케이스 1 에 20 대 판매량을 30 대 판매량과 위치를 바꾸게 되면 나머지 케이스들에 20 대와 30 대 판매량도 같이 위치를 바꾸어야 됩니다. 다시 말해, 4 개의 level 인 각 케이스에 판매량은 연령이라는 또 다른 요소(factor)와 관계를 갖게 됩니다. 여기서 연령과 같은 것을 blocking factor 라고 하고, blocking factor 에 level 들을 block 들이라고 합니다. 이 예제에서 block 들은 10 대, 20 대, 30 대, 40 대, 50 대입니다.

그럼 우선 Block 인 연령에 대한 영향을 고려하지 않고 Onew-way ANOVA 를 보겠습니다.

	SS	DF	Mean SS	F
Between	279.6	3	93.2	3.0113

Within	495.2	16	30.95	
Total	774.8	19	40.7789	

결과를 보시면 F 의 critical value 가 3.2388 인데 F test statistic 3.0113 으로 null hypothesis 를 reject 하지는 못하고 p-value 를 보면 0.06 으로 케이스별 판매량 차이가 없다고 판단이 됩니다. 즉, 이 결과만 보았을 때는 4 개의 케이스 판매량에 차이는 없는 것으로 판단이 됩니다.

연령별로 판매량을 구분해서 살펴 보게 되면 연령별 판매량이 다른것을 알 수 있고 또한 각 연령별로 보았을 때 케이스 판매량의 차이는 다르게 나타나는 것을 알 수 있습니다.

Block₩Group	케이스 1	케이스 2	케이스 3	케이스 4	평균	표준편차
10 대	48	40	39	40	41.75	4.1932
20 대	32	39	29	24	31	6.2716
30 대	42	45	29	39	38.755	6.9462
40 대	46	36	29	39	37.5	7.0475
50 대	38	35	32	31	34	3.1623
평균	41.2	39	31.6	34.6		
표준편차	6.4187	3.9370	4.3359	6.9498		

예를 들어 20 대의 경우 케이스 1 이 최고, 케이스 3 이 최저이지만 20 대의 경우는 케이스 2 가 최고, 케이스 4 가 최저로 다른 선호도를 보여주고 있습니다. 즉, 케이스 판매량에 연령별 영향이 있음에도 On-way ANOVA 에서는 이러한 영향을 제거하여 비교 분석을 하지 못합니다.
이러한 group 에서 block 의 특성들을 random error 로 부터 제거하고 다시 group 간의 차이가 있는지를 하는 것이 Randomized Block ANOVA 입니다.

Randomized Block ANOVA 는 One-way ANOVA 처럼 케이스의 판매량을 linear additive model 로 표현한 결과는 다음과 같습니다.

$$y_{ij} = \mu + \tau_i + \beta_j + e_{i,j}$$

여기서

μ: 총 판매된 휴대폰 케이스 개수의 평균값

τ_i: 케이스별 영향값 $\mu_i - \mu$ (평균 0)

β_i: 연령별 영향값 $\mu_j - \mu$ (평균 0)

e_{ij}: random error

Randomized Block ANOVA 은 One-way ANOVA 의 SSW 가 더욱 세분화되어 처리가 됩니다.

1) 전체 sample 의 편차: Sum of Squares Total(SST), Mean ST(MST)
2) Group 에 대한 Sample 평균들의 편차: Sum of Squares Between(SSB), Mean SB(MSB)
3) Block 에 대한 평균들의 편차: Sum of Squares block(SSBL), Mean SBL(MSBL)
4) 각 Group 의 Standard Error 편차: Sum of squares error(SSE), Mean SE(MSE)

$$SST = SSB + SSBL + SSE \tag{11.12}$$

$$\begin{aligned} SST &= \sum_{i=1}^{N}\sum_{j=1}^{B}\left(\left(x_{ij} - \overline{x}_i\right) + \left(\overline{x}_i - \overline{\overline{x}}\right)\right)^2 = \sum_{i=1}^{N} M_i\left(\overline{x}_i - \overline{\overline{x}}\right)^2 \\ &\quad + \sum_{i=1}^{N}\sum_{j=1}^{B}\left(x_{ij} - \overline{x}_i - \overline{x}_j + \overline{\overline{x}} + \left(\overline{x}_j - \overline{\overline{x}}\right)\right)^2 \\ &= \sum_{i=1}^{N} M_i\left(\overline{x}_i - \overline{\overline{x}}\right)^2 + N\sum_{j=1}^{B}\left(\overline{x}_j - \overline{\overline{x}}\right)^2 + \sum_{i=1}^{N}\sum_{j=1}^{B}\left(x_{ij} - \overline{x}_i - \overline{x}_j + \overline{\overline{x}}\right)^2 \\ &= SSB + SSBL + SSE \end{aligned}$$

보시는 것과 같이 SSW 를 SSBL 과 SSE 로 나누게 되어 ANOVA 검사방식를 합니다.

	SS	DF	Mean SS	F
Between	SSB	$N-1$	$MSB = \frac{SSB}{N-1}$	$F_{\overline{x}} = \frac{MSB}{MSE}$
Block	SSBL	$B-1$	$MSBL = \frac{SSBL}{B-1}$	$F_{\overline{x}} = \frac{MSBL}{MSE}$
Error	SSE	$(\mathrm{N}-1)(B-1)$	$MSE = \frac{SSE}{(\mathrm{N}-1)(B-1)}$	

Total	SST	$NB-1$	$MST=\frac{SST}{NB-1}$	

이 방식으로 휴대폰 케이스판매의 차이가 존재하는가를 확인을 하겠습니다.

	SS	DF	Mean SS	F(Critical Value)
Between	279.6	3	93.2	5.204(3.49)
Block	280.3	4	70.075	3.912(3.26)
Error	214.9	12	17.908	
Total	774.8	19	40.77894	

Block 에 차이를 보여주는 MSBL 값을 MSE 로 나누어 값은 F test statistic(3.912)을 F critical value(3.26)와 비교를 한 결과를 보면 F test statistic 이 더 큰 값을 갖기 때문에 Block 인 연령별 케이스 판매량의 차이가 있다는 것을 알 수 있습니다. 따라서 Block 에 의한 판매량 차이에 대한 효과를 제거하고 케이스 판매량을 보기 위해서 Randomized Block ANOVA 로 확인이 가능하게 됩니다. 이것에 대한 결과는 MSB/MSE 인 F test statistic(5.204)와 이에 해당하는 F critical value(3.49)와 비교로 F test statistic 이 크기 때문에 케이스 판매량에는 차이가 있음을 알 수 있습니다.

정리를 하면 randomized block ANOVA 는 one-way ANOVA 처럼 factor 가 한 개이고 데이타를 block 별로 나누어서 factor 의 차이가 block 의 영향으로 존재함을 알아내는 것입니다. 만일 block 도 factor 로 취급하여 처리를 하려면 다음에 배울 Two-way ANOVA 를 이용해야 합니다.

jMath 사용

```
jMath.stat.anovarbl(alpha, sample1, sample2, sample3, … )
```

```
var s1 = jMath('48 32 42 46 38');
var s2 = jMath('40 39 45 36 35');
var s3 = jMath('39 29 29 29 32');
var s4 = jMath('40 24 39 39 31');

var result = jMath.stat.anovarbl(0.05, s1,s2,s3,s4);
console.log(result);

F: {
```

```
    between: {
      C: 3.4902948205877995,
      pvalue: 0.015635198919854765,
      value: 5.20428106095859
    },
    block: {
      C: 3.259166726901351,
      pvalue: 0.029363814369198882,
      value: 3.912982782689627
    },
}
alpha: 0.05
between: { df: 3, ms: 93.2,  ss: 279.6 }
error: { df: 12, ms: 17.90833,  ss: 214.899999 }
means: { block: [41.75, 31, 38.75, 37.5, 34],
         total: 36.6,
         widthin: [ [41.2], [39], [31.6], [34.6] ] }
numSamples: 20
sigma: { ... }
compare :{...}
```

4. Multiple comparison for randomized block ANOVA

어떤 group 의 쌍이 다른가를 randomized block ANOVA 의 결과로 확인하는 방법을 소개합니다.

4.1. Least Significant difference

$$lsd(\alpha) = t_{1-\alpha/2,df}\sqrt{\frac{2MSE}{n}} = t_{1-\alpha/2,df}S_d \tag{11.13}$$

여기서 df 는 (N-1)(B-1)입니다.

4.2. Tukey-Kramer Multple Comparision

one-way ANOVA 와 다른 점은 CR 값을 계산하는 방식과 degree of freedom 입니다.

$$CR = Q_\alpha\sqrt{\frac{MSE}{B}} \tag{11.14}$$

여기서 DF1 은 N, DF2 는 (B-1)(N-1)이 되고 B 는 block(level)의 개수 입니다. 이것을 적용한 결과는 다음과 같습니다.
여기서 CR 값은 group 의 sample 크기가 같기 때문에 모두 7.946 입니다.

i,j	$\lvert\bar{x}_i - \bar{x}_j\rvert$	결론
1,2	2.2	같음
1,3	9.6	다름
1,4	6	같음
2,3	7.4	같음

2,4	4.4	같음
3,4	3	같음

결과에서 보면 케이스 1 과 케이스 3 이 판매량 차이가 있고 나머지는 차이가 없는 것으로 나타납니다.

4.3. Scheffe's Test

Confidence Interval 을 계산할 때 MSE 사용하고 F 값은 randomized block ANOVA 와 critical value 와 같습니다.

$$\sum_{i=1}^{N} C_i \overline{x}_i \pm \sqrt{MSE(N-1)F_{1-\alpha,N-1,(N-1)(B-1)} \sum_{i=1}^{N} \frac{C_i^2}{n}} \tag{11.15}$$

5. Two-way ANOVA

One-way ANOVA 는 하나의 factor 로 차이가 있는가를 알아내는 방법이고, randomized block ANOVA 는 이러한 factor 가 block 별 차이로 인한 영향을 제거하여 one-way ANOVA 처럼 수행 할 수 있도록 합니다. Two-way ANOVA 는 두 개의 factor 를 동시에 비교하는 것으로 one-way ANOVA 에 다른 one-way ANOVA 를 결합하여 동시에 비교 분석한다고 생각하시면 됩니다.

예를 들어 제과점에서 아침 메뉴로 만든 세가지 빵의 판매량을 손님의 성별로 구분하여 우선 제품과 고객의 성별간의 판매량의 상호관계가 있는지 알고 난 후에 판매량의 차이가 제품별, 고객의 성별로 다른가를 조사하고, 합니다.

	메뉴 1	메뉴 2	메뉴 3
남자 고객	50 48 60 55 57	68 72 63 67 71	61 65 62 63 66
여자 고객	80 77 65 77 78	90 91 97 89 92	81 83 75 81 82

Two-way ANOVA 는 One-way ANOVA 와 같이 SSB 와 SSW 의 합인 SST 로 이루어 집니다. 차이점은 SSB 가 더욱 세분화 됩니다.

$$SST = SSB + SSW = SSFA + SSFB + SSFAB + SSW \tag{11.16}$$

$$SST = \sum_{i=1}^{R}\sum_{j=1}^{C}\sum_{k=1}^{n}\left((x_{ij} - \overline{x}_{ij}) + (\overline{x}_{ij} - \overline{\overline{x}})\right)^2 = \sum_{i=1}^{R}\sum_{j=1}^{C}\sum_{k=1}^{n}(x_{ijk} - \overline{x}_{ij})^2 + n\sum_{i=1}^{R}\sum_{j=1}^{C}\left(\overline{x}_{ij} - \overline{\overline{x}}\right)^2$$
$$+ 2\sum_{i=1}^{R}\sum_{j=1}^{C}\sum_{k=1}^{n}(x_{ijk} - \overline{x}_{ij})(\overline{x}_i - \overline{\overline{x}})$$
$$= \sum_{i=1}^{R}\sum_{j=1}^{C}\sum_{k=1}^{n}(x_{ijk} - \overline{x}_{ij})^2 + n\sum_{i=1}^{R}\sum_{j=1}^{C}\left(\overline{x}_i - \overline{\overline{x}} + \overline{x}_j - \overline{\overline{x}} + \overline{x}_{ij} - \overline{x}_i - \overline{x}_j + \overline{\overline{x}}\right)^2$$
$$= \sum_{i=1}^{R}\sum_{j=1}^{C}\sum_{k=1}^{n}(x_{ijk} - \overline{x}_{ij})^2 + nC\sum_{i=1}^{R}\left(\overline{x}_i - \overline{\overline{x}}\right)^2 + nR\sum_{j=1}^{C}\left(\overline{x}_j - \overline{\overline{x}}\right)^2$$
$$+ n\sum_{i=1}^{R}\sum_{j=1}^{C}\left(\overline{x}_{ij} - \overline{x}_i - \overline{x}_j + \overline{\overline{x}}\right)^2 = SSW + SSFA + SSFB + SSFAB$$

이 계산 식에서 모든 cell 의 sample 개수는 동일할 경우입니다. 그럼 two-way ANOVA 검사를 위한 방식은 다음과 같습니다.

	SS	DF	Mean SS	F
Between	SSB	$RC-1$	$MSB = \frac{SSB}{RC-1}$	$F_{\overline{x}} = \frac{MSB}{MSW}$
Factor A	SSFA	R-1	$MSFA = \frac{SSFA}{R-1}$	$F_A = \frac{MSFA}{MSW}$
Factor B	SSFB	C-1	$MSFB = \frac{SSFB}{C-1}$	$F_B = \frac{MSFB}{MSW}$
Interact	SSFAB	$(R-1)(C-1)$	$MSFAB = \frac{SSFAB}{(R-1)(C-1)}$	$F_{AB} = \frac{MSFAB}{MSW}$
Within	SSW	$\mathrm{RC(n-1)}$	$MSE = \frac{SSW}{\mathrm{RC(n-1)}}$	
Total	SST	$RCn-1$	$MST = \frac{SST}{RCn-1}$	

앞의 예를 적용을 하면 Factor A 는 고객의 성별이 되고 Factor B 는 메뉴입니다.

A₩B	메뉴 1	메뉴 2	메뉴 3	
남자 고객	54	68.2	63.4	61.8667
여자 고객	75.4	91.8	80.4	82.5333
	64.7	80	71.9	72.2

$$SSFA\left(\textit{고객성별}\right) = 5 \times 3((61.8667 - 72.2)^2 + (82.5333 - 72.2)^2) = 3203.33$$
$$SSF\textit{B}\left(\textit{메뉴}\right) = 5 \times 2((64.7 - 72.2)^2 + (80 - 72.2)^2 + (71.9 - 72.2)^2) = 1171.8$$

	SS	DF	Mean SS	F	F_{crit}	p-value
고객 성별	3203.33	1	3203.33	199.58	4.26	3.973e-13
메뉴	1171.8	2	585.9	36.504	3.4	5.25e-8
Interact	56.47	2	28.23	1.76	3.4	0.19
Within	385.2	24	16.05			
Total	4816.8	29	166.096			

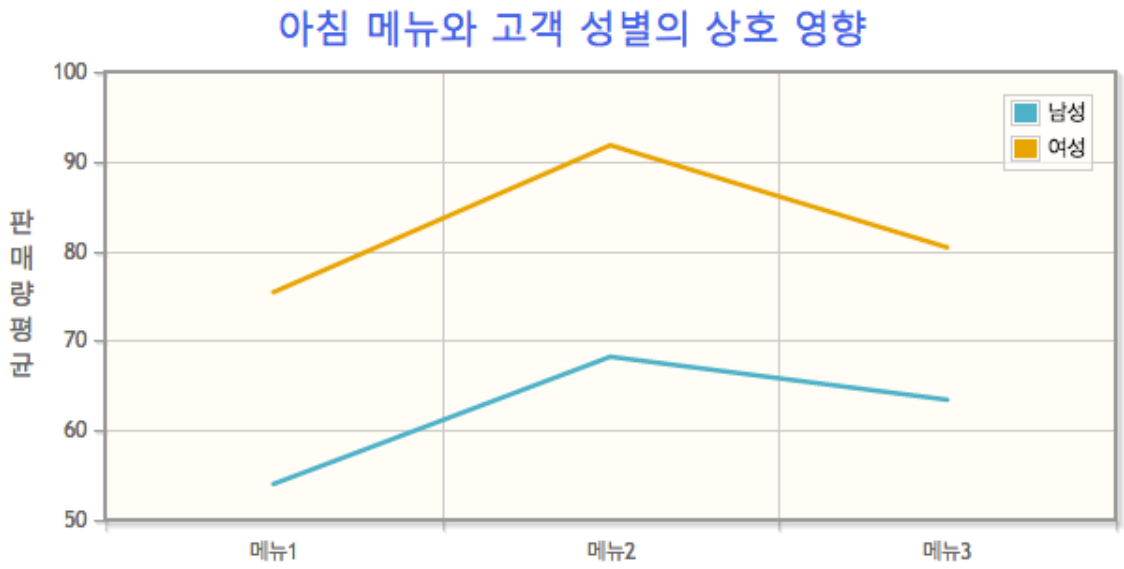
Two-way ANOVA 를 통해 두 Factor 들에 차이가 있는가를 알기 위해서 factor 간에 간섭이 없는가를 알아야 합니다. 이를 위해서 factor 간 상호 관계가 존재하는가를 확인 해야 합니다.

1) Factor A 와 Factor B 의 상호관계 검사

H_0: 두 factor 에 상호관계는 없다.
H_1: 두 factor 에 상호관계가 있다.

이를 검사하기 위해서 F_{AB} 값이 critical value 보다 작다면 상호 관계가 없다고 판단이 됩니다. 결과를 보시면 해당 p-value 가 0.19 로 0.05 보다 크기 때문에 null hypothesis 를 reject 하기에는 충분하지 못합니다. 따라서 Factor A 와 Factor B 각각의 group 별 차이가 존재하는지를 검사할 수 있습니다.

twoway.html



각 메뉴별 평균가격을 성별로 보면 판매량의 변화가 성별로 같음을 알 수 있습니다. 예를 들어 여성 고객에 대한 메뉴 1 이 메뉴 2 에 비해서 판매량이 저조한데 이것은 남성 고객에게도 같이 적용이 됩니다. 이러한 관계가 상호관계가 없는 경우 입니다. 즉 메뉴(Factor B)에 상관없이 남녀(Factor A)의 차이를 알 수 있고 또한 남녀(Factor A)에 상관없이 메뉴(Factor B)의 차이를 알 수 있게 되어 Factor A 와 Factor B 검사를 할 수 있습니다.

2) Factor A 검사

$$H_0: \mu_M = \mu_F$$
$$H_1: \mu_M \neq \mu_F: \textit{모든 성별 판매량은 같지 않다}$$

검사를 위한 값이 F_A 로 해당 p-value 를 보시면 매우 작은 값으로 이 의미는 남녀의 판매량차이가 있다고 판단할 수 있습니다.

3) Factor B 검사

$$H_0: \mu_{p1} = \mu_{p2} = \mu_{p3}$$
$$H_1: \textit{모든 메뉴별 판매량은 같지 않다}$$

검사를 위한 값이 F_B 로 해당 p-value 를 보시면 값이 매우 작아 메뉴별 판매량이 같지 않음을 알 수 있습니다.

6. Multiple comparison for two-way ANOVA

One-way ANOVA 와 같이 interaction 이 없는 경우 Tukey-Kramer Multiple comparison 검사를 할 수 있습니다. 이를 위해서 두개의 factor 별 critical range 값을 계산하는 방식에 차이가 있습니다.

$$CR_A = Q_A\sqrt{\frac{MSW}{Rn}}$$
$$CR_B = Q_B\sqrt{\frac{MSW}{Cn}} \quad (11.17)$$

여기서 두 critical value 들은 studentized range 에서 두개의 degree freedom 값은 각각 다음과 같습니다.

Q_A: C, RC(n-1)

Q_B: R, RC(n-1)

Factor A 인 고객 성별 차이를 보면 CR 값은 3.02

i,j	$\lvert\overline{x}_i - \overline{x}_j\rvert$	결론
1,2	20.667	다름

Factor B 인 메뉴별 차이를 보면 CR 값은 4.47

i,j	$\lvert\overline{x}_i - \overline{x}_j\rvert$	결론
1,2	15.23	다름
1,3	7.2	다름

2,3	8.1	다름

jMath

jMath.stat.anova2(alpha, samples)

```
var s11 = jMath('50 48 60 55 57');
var s12 = jMath('68 72 63 67 71');
var s13 = jMath('61 65 62 63 66');
var s21 = jMath('80 77 65 77 78');
var s22 = jMath('90 91 97 89 92');
var s23 = jMath('81 83 75 81 82');

var result = jMath.stat.anova2(0.05, [[s11,s12,s13],[s21,s22,s23]]);
console.log(result);console.log(result);
```

```
F: {
    factorA: {
      C: 4.259677213754524,
      pvalue: 3.9734882051334353e-13,
      value: 199.58463136033174
    },
    factorB: {
      C: 3.4028261053510427,
      pvalue: 5.257411739290063e-8,
      value: 36.50467289719615
    },
    interaction: {
      C: 3.4028261053510427,
      pvalue: 0.19368672637249817,
      value: 1.759086188992716
    },
}
alpha: 0.05
factorA: { df: 3, ms: 3203.333333333333,  ss: 3203.333333333333 }
factorB: { df: 2, ms: 585.8999999999999,  ss: 1171.7999999999997 }
interact: { df: 2, ms: 28.23333333333317,  ss: 56.46666666666634 }
means: { cell: ..., total: 72.2, facltorA: ...,  factorB:  ... }
size: 20
sigma: { ... }
multicompare :{...}
total: { df: 29, ms: 166.09655172413792,  ss: 385.20000000000107 }
within: { df: 24, ms: 16.050000000000043,  ss: 385.20000000000107 }
```