비즈니스 통계

Statistics for Business

분산 분석(ANOVA)

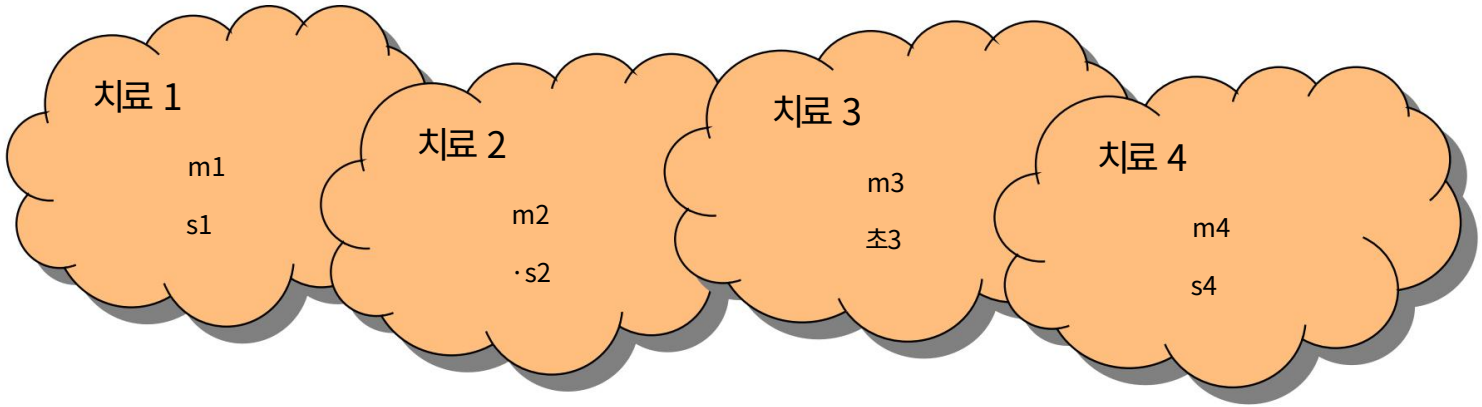
Analysis of Variance (ANOVA)

분산 분석 소개.

ANOVA(ANalytic Of VAriance)는 통계적 유의성을 위해 3개 이상의 평균(그룹* 또는 변수)을 비교하는 데 유용합니다.

- H0: μ1 = μ2 =…= μk
- H1: 수단이 모두 동일하지는 않습니다.

*그룹을 치료라고 합니다.



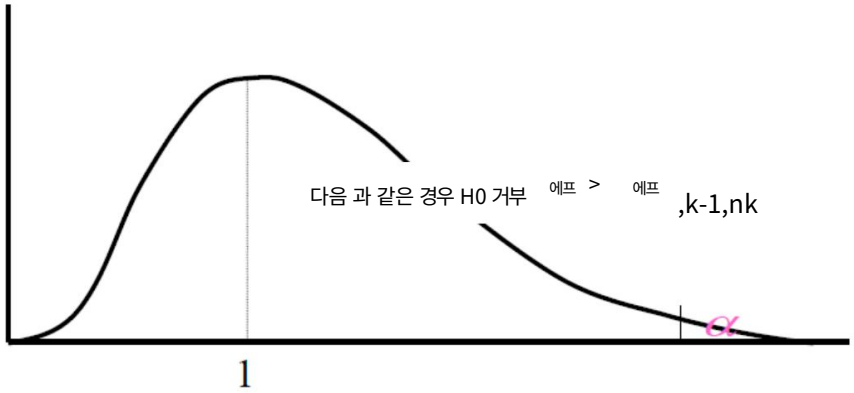
분산 분석 소개.

분산분석 가정

- 모집단은 정규 분포를 따릅니다.
- 모집단의 표준 편차(σ)가 동일합니다.
- 표본은 무작위로 선택되며 독립적입니다.

•F분포

- 검정 통계량은 F 분포를 따릅니다 (F > F ,k-1,nk 인 경우 H0 기각).
- F 분포는 연속형이며 음수가 될 수 없습니다.
- 긍정적으로 편향되어 있습니다.



분산 분석 예(1)

한 지역 금융 센터의 관리자가 세 명의 직원 사이에서 서비스를 받는 고객 수로 측정된 생산성을 비교하려고 합니다. 4일은 무작위로 선택되며 생산성은 다음과 같습니다.

정확히 잰.

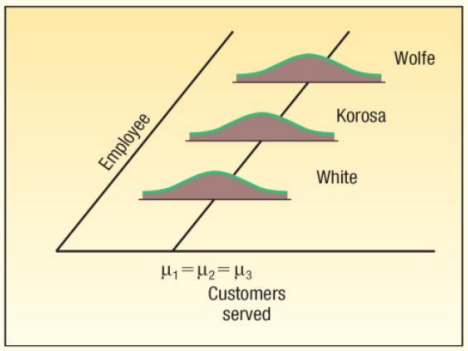
Wolfe	White	Korosa
55	66	47
54	76	51
59	67	46
56	71	48

서비스를 받는 평균 고객 수에 차이가 있나요?

분산 분석 예(1)

그림은 인구가 어떻게 나타나는지 보여줍니다.

모집단은 정규 분포를 따른다는 점에 유의하세요.

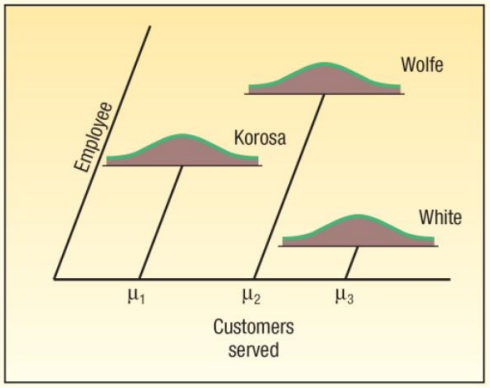


Case Where Treatment Means Are the Same

치료방법에는 차이가 없습니다

- 생산성의 변화는 다음과 같은 요인으로 인해 발생합니다.

  무작위 구성 요소



Case Where Treatment Means Are Different

치료방법의 차이

- 다음 중 상당한 차이가 있습니다.

  치료 수단.

변형 설명

ANOVA는 수치 종속의 변동 원인을 식별하려고 합니다.

변수 Y

종속변수
(숫자)

Y=맛있는 점수
쿠키

영향을 받을 수 있습니다.

독립 변수
(범주형)

치료
(온도)
T1=낮음
T2=중간
T3=높음

AND 독립변수
(범주형)

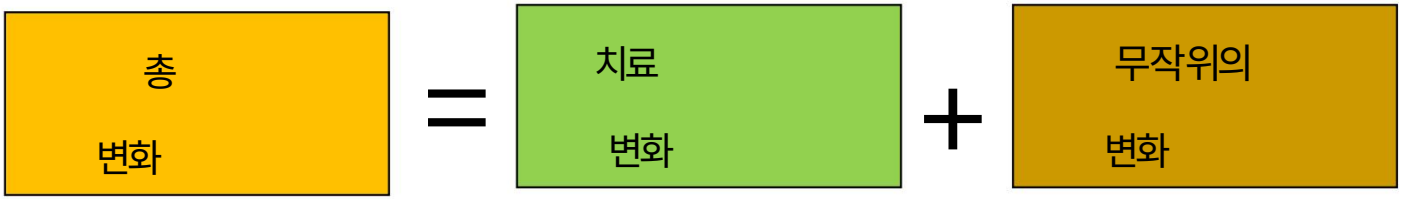
치료
(조리시간)
T1=짧음
T2=중간
T3=긴

- 처리: 요인 또는 요인 조합의 가능한 각 값
- 일원 분산 분석: 단일 요인
- 양방향 ANOVA: 두 가지 요인

변형 설명

평균에 대한 Y의 변동이 하나 이상의 범주형 독립 변수(요인, 치료 변동)로 설명되거나 설명되지 않습니다.

(무작위 오류, 무작위 변형)



- 처리 변형: 각 처리 평균 간의 차이 제곱의 합
  및 전체 평균(그룹 간)
- 변동의 원인은 치료(예: 요인)로 인한 것입니다.

- 무작위 변동: 각 관측값과 관측값 간의 차이 제곱의 합입니다.
  그 처리 평균(그룹 내)

일원 분산 분석

총 변동 = 치료 변형 + 무작위 변형

$$\sum_{j=1}^{c}\sum_{i=1}^{n_j}(y_{ij}-\bar{y})^2 = \sum_{j=1}^{c} n_j(\bar{y}_j-\bar{y})^2 + \sum_{j=1}^{c}\sum_{i=1}^{n_j}(y_{ij}-\bar{y}_j)^2$$

SST(전체 제곱합) = SSA(처리 간 제곱합)

+ SSE(처리 내 제곱합)

Source of Variation	*Sum of Squares*	*Degrees of Freedom*	*Mean Square*	*F Statistic*
Treatment (between groups)	$SSA = \sum_{j=1}^{c} n_j(\bar{y}_j-\bar{y})^2$	$c-1$	$MSA = \frac{SSA}{c-1}$	$F = \frac{MSA}{MSE}$
Error (within groups)	$SSE = \sum_{j=1}^{c}\sum_{i=1}^{n_j}(y_{ij}-\bar{y}_j)^2$	$n-c$	$MSE = \frac{SSE}{n-c}$	
Total	$SST = \sum_{j=1}^{c}\sum_{i=1}^{n_j}(y_{ij}-\bar{y})^2$	$n-1$		

H0가 참이면 F

Fc-1,nc를 따릅니다 .

배포

평균 1.

분산 분석 예(2)

최근 4개의 주요 항공사로 구성된 그룹이 최근 항공편에 대한 만족도 수준에 관해 최근 승객을 조사하기로 결정했습니다.

표의 총점은 비행에 대한 만족도를 나타냅니다.

4개 항공사 간 평균 만족도에 차이가 있나요?

.01 유의 수준을 사용합니다.

그리고	티	아	영형
94	75	70	68
90	68	73	70
85	77	76	72
80	83	78	65
	88	80	74
		68	65
		65	

분산 분석 예(2)

- 1단계: 귀무가설과 대립가설 명시

- H0 : μE = μA = μT = μO
- H1 : 수단이 모두 동일하지는 않습니다.

- 2단계: 유의수준 선택

- α = 1%

- 3단계: 적절한 테스트 통계량 결정

- 두 개 이상의 그룹의 평균을 비교하고 있으므로 F 통계를 사용합니다.

- 4단계: 의사결정 규칙 수립

- 오른쪽 꼬리 테스트
- F > F 이면 H0 거부
- 유의수준 1%에서 10을 참고하여 임계값 Fα를 계산할 수 있다.

F.01 테이블. Fα는 자유도 에 따라 달라질 수 있습니다.

분산 분석 예(2)

• 5단계: F 값을 계산하고 결정을 내립니다.

	그리고	티	아	영형
만족 수준	94 90 85 80	75 68 77 83 88	70 73 76 78 80 68 65	68 70 72 65 74 65
	87.25	78.20	72.86	69
	75.64			

분산 분석 예(2)

• 5단계: F 값을 계산하고 결정을 내립니다(SST).

	그리고	티	나	영형
만족 수준	4*(87.25-75.64)2	5*(78.20-75.64)2	7*(72.86-75.64)2	6*(69-75.64)2
SSAi	539.1684	32.768	54.0988	264.5376
SSA				

분산 분석 예(2)

• 5단계: F 값을 계산하고 결정을 내립니다(SSE).

	그리고	티	아	영형
만족 수준	(94-87.25)2 (90-87.25)2 (85-87.25)2 (80-87.25)2	(75-78.20)2 (68-78.20)2 (77-78.20)2 (83-78.20)2 (88-78.20)2	(70-75.65)2 (73-75.65)2 (76-75.65)2 (78-75.65)2 (80-75.65)2 (68-75.65)2 (65-75.65)2	(68-69)2 (70-69)2 (72-69)2 (65-69)2 (74-69)2 (65-69)2
SSEi	110.75	234.80	180.86	68
SSE	594.41			

분산 분석 예(2)

• 5단계: F 값을 계산하고 결정을 내립니다.

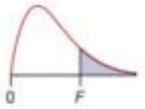
변화	제곱의 합	df	평균 제곱	에프
치료	SSA=890.57	c-1=3	890.57/3 =296.86	296.86/33.02 =8.99
오류	SSE=597.41	NC=18	597.41/18 =33.02	
총	SST=1485.09	n-1=21		

• F=8.99> Fα,c-1,nc=F0.01,4-1,22-4=5.09

• 따라서 귀무가설은 기각됩니다.

• 모집단 평균이 모두 동일하지는 않다고 결론을 내립니다.

CRITICAL VALUES OF $F_{.01}$



This table shows the 1 percent right-tail critical values of F for the stated degrees of freedom ($d.f.$).

Denominator Degrees of Freedom (df_2)	Numerator Degrees of Freedom (df_1)										
	1	2	3	4	5	6	7	8	9	10	12
1	4052	4999	5404	5624	5764	5859	5928	5981	6022	6056	6107
2	98.50	99.00	99.16	99.25	99.30	99.33	99.36	99.38	99.39	99.40	99.42
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.34	27.23	27.05
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23

분산 분석 예(2)

각 ANOVA 테이블의 F-검정을 해석합니다. 데이터 형식을 추론할 수 있습니까?

ANOVA 결과는요?

분산분석표

원천	봄 여행 시즌	df	MS	F p-값
치료	14,465.63	4 3,616.408 62		29.80 7.78E-14
오류	7,523.44	121.346 66		
총	21,989.07			

메리	청구서	출격	로버트	톰
113	124	153	137	112
134	123	157	164	130
127	102	146	155	119
124	107	146	141	110
102	132	120	139	110
132	88	138	155	120
132	138	156	146	125
121	105	130	148	111
122	103	154	138	126
136	121	142	145	101
128	108	159	152	103
121	112	162	160	
126	131		149	
118	105		151	
120			131	

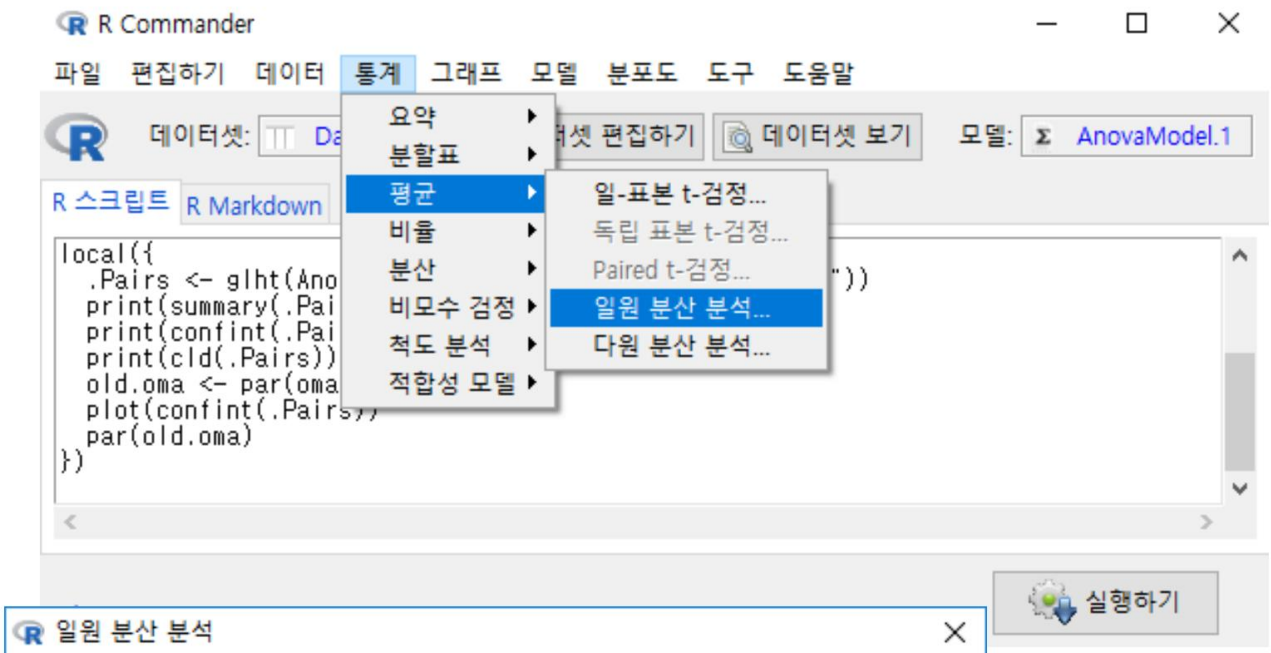
분산분석표

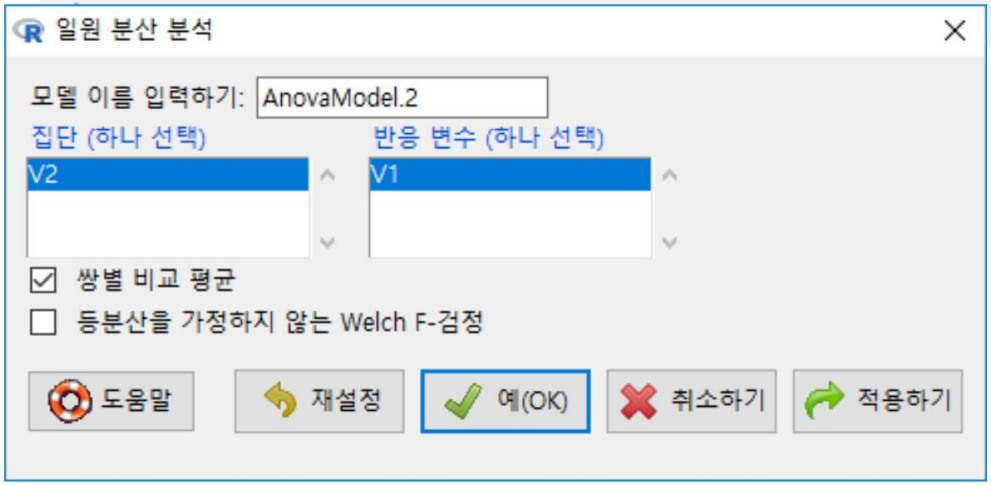
원천	SS	df MS 2 1,184.067	F p-값
치료	2,368.13	12 17.733 14	66.77 3.14E-07
오류	212.80		
총	2,580.93		

게으른	시속 60마	0~60mph
41	일 65	76
45	67	72
44	66	76
45	66	77
46	76	64

R 연습-ANOVA

1) csv 파일 읽기(ANOVA-1.csv)



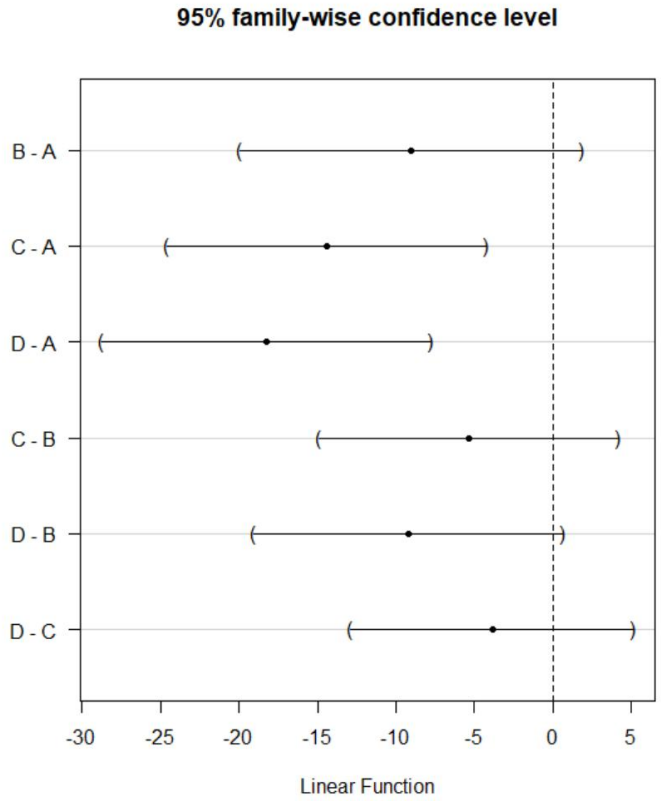


R 연습-ANOVA

```
> summary(AnovaModel.1)
             Df Sum Sq Mean Sq F value   Pr(>F)
V2            3  890.7  296.89   8.991 0.000743 ***
Residuals    18  594.4   33.02
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> with(Dataset, numSummary(V1, groups=V2, statistics=c("mean", "sd")))
      mean       sd data:n
A 87.25000 6.075909      4
B 78.20000 7.661593      5
C 72.85714 5.490251      7
D 69.00000 3.687818      6
```

```
        Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: aov(formula = V1 ~ V2, data = Dataset)

Linear Hypotheses:
           Estimate Std. Error t value Pr(>|t|)
B - A == 0   -9.050      3.855  -2.348  0.12361
C - A == 0  -14.393      3.602  -3.996  0.00439 **
D - A == 0  -18.250      3.709  -4.920  < 0.001 ***
C - B == 0   -5.343      3.365  -1.588  0.40867
D - B == 0   -9.200      3.480  -2.644  0.07094 .
D - C == 0   -3.857      3.197  -1.206  0.62963
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

```
        Simultaneous Confidence Intervals

Multiple Comparisons of Means: Tukey Contrasts


Fit: aov(formula = V1 ~ V2, data = Dataset)

Quantile = 2.8228
95% family-wise confidence level


Linear Hypotheses:
           Estimate lwr      upr
B - A == 0  -9.0500 -19.9316   1.8316
C - A == 0 -14.3929 -24.5601  -4.2256
D - A == 0 -18.2500 -28.7208  -7.7792
C - B == 0  -5.3429 -14.8411   4.1554
D - B == 0  -9.2000 -19.0225   0.6225
D - C == 0  -3.8571 -12.8819   5.1676

   A    B    C    D
 "b" "ab"  "a"  "a"
```



양방향(무작위 블록) ANOVA

• 우리는 다른 요소(예: 블록 변수)를 고려하여 다음을 줄일 수 있습니다.

SSE 용어.

• SST = SSA(치료) + SSB(차단) + SSE(무작위 오류)

•H0: 처리 수단이 동일함($\mu 1 = \mu 2 = \cdots = \mu k$)

•H1: 치료수단이 모두 동일하지는 않다

•H0: 블록 평균은 동일합니다($\mu 1 = \mu 2 = \cdots = \mu b$).

•H1: 블록 평균이 모두 동일하지는 않습니다.

양방향 분산 분석 - 예(1)

• 고려 중인 경로는 4개이며 다음 중 어느 것인지 결정하고 싶습니다.

4개 노선의 평균 이동 시간에는 차이가 있었습니다.

• 다양한 드라이버가 있기 때문에 테스트는 각 드라이버가 그렇게 되도록 설정되었습니다.

4개의 경로를 각각 따라 운전했습니다.

여행 시간		경로(치료)			
		안에	안에	시간	에프 차원
운전사 (블록)	디	18	17	21	22
	에스	16	23	23	22
	영형	21	21	26	22
	와 함께	23	22	29	25
	에프	25	24	28	28

양방향 분산 분석 - 예(1)

• 양방향 ANOVA

변화	의 합 사각형	df	평균 정사각형	에프	p-값
치료 SSA = 72.8		c-1=3	24.27	7.935	0.004
차단하다	SSB =119.7	r-1=4	29.93	9.785	0.001
오류	SSE = 36.7	(r-1)(c-1)=12	3.06		
총	229.2	19			

우리는 결론을 내린다

• 평균 시간은 모든 운전자에게 동일하지 않습니다.

• 경로의 평균 시간이 모두 동일하지는 않습니다.

$F_{0.05, 3.12} = 3.49$

$F_{0.05, 4.12} = 3.26$