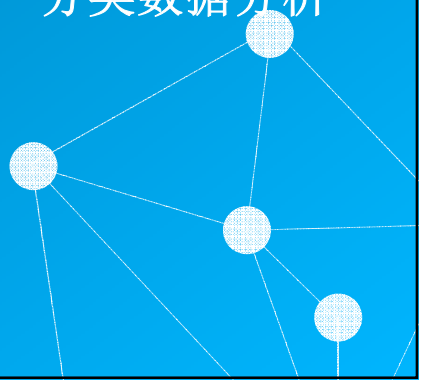
Kim, Dae Ho (Mokwon University)
mis@mokwon.ac.kr / mis@acm.org

CHAPTER 11

分类数据分析



第11-1章
分类数据分析

1. 拟合优度检验
2. 交叉分析

2/25

Kim, Dae Ho (Mokwon University)
mis@mokwon.ac.kr / mis@acm.org

分类数据

- 分类数据(categorical data)
  - 区分属性的数据，按照不同因子(factor)进行收集，用表格形式来概括
- 名目(norminal)数据
  - 没有顺序
  - ✓ 按国家分、地区、公司、部门、工艺、系列、性别... 进行分类的数据
- 顺序(ordinal)数据
  - 有顺序
  - ✓ 按照年龄段、职位、规模、收入范围、学历...进行分类的数据

3/25

1. 拟合优度检验

- 拟合优度检验
  - 用来检验通过观察或试验所得的样本数据的分布是否与某一形态的分布一致的方法
  - 依据分类数据的观测值与期望的差异进行检验
  - 类别的总个数 → k
  - 类别i的预计频数(expected frequency) → Ei
  - 类别i的观测频数(observed frequency) → Oi

$$\chi_0^2 \equiv \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2(k-1)$$

$$H_0 \Rightarrow p_i \Rightarrow E_i = np_i$$

$$\Rightarrow \chi_0^2 = \sum_{i=1}^{k} \frac{(O_i - np_i)^2}{np_i} \sim \chi^2(k-1) \,\Big|\, H_0$$

4/25

Kim, Dae Ho (Mokwon University)
mis@mokwon.ac.kr / mis@acm.org

1. 拟合优度检验

[定理 12-1] 拟合优度检验(goodness of fit test)

原假设 H_0 : "观测的数据符合某一分布"

原假设 H_0 : 设定类别i的发生概率p_i

검정통계량: $\chi_0^2 = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i}$

기각역: $\chi_0^2 > \chi_{1-\alpha;k-1-m}^2$

k = 类别的个数

m = 在原假设中设定的分布中估计的参数个数

- 自由度(degree of freedom)

$$\sum_{i=1}^{k}(X_i - np_i) = \sum_{i=1}^{k} X_i - n\sum_{i=1}^{k} p_i = n - n = 0$$

5/25

1. 拟合优度检验

[例 12-1] 投掷骰子120次，各点数的出现频数如下表所示，求检验在显著性水平为5%的情况下，该骰子是否是均匀制造。

点数	1	2	3	4	5	6	合计
频数	31	26	22	18	13	10	120

귀무가설 H_0 : "주사위는 공정하다"로부터 $H_0: p_i = 1/6,\ i = 1,2,...,6$

기대도수 $np_i = 120/6 = 20,\ i = 1,2,...,6$

범주	1	2	3	4	5	6	계
X_i	31	26	22	18	13	10	120
np_i	20	20	20	20	20	20	120
통계량	6.05	1.8	0.2	0.2	2.45	5.0	15.7

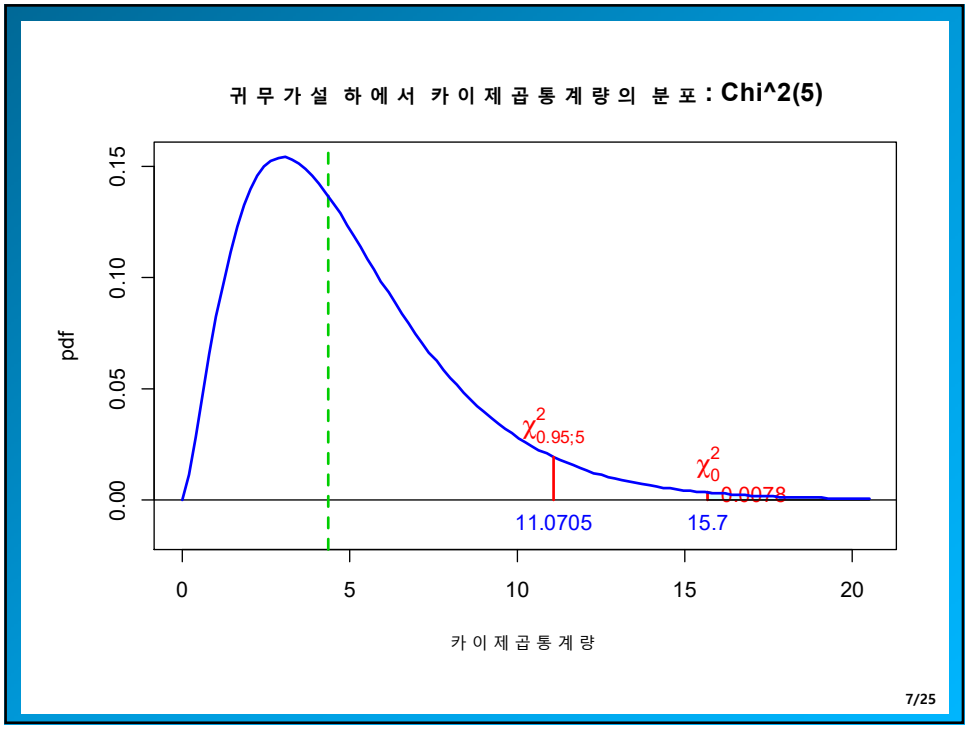
$\sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i}$

자유도는 $6-1=5$ 기각역은 $\chi_0^2 > \chi_{0.95;5}^2 \doteq 11.070$

검정통계량은 $\chi_0^2 = 15.7$ > 11.070 → 原假设被拒绝

→ 有证据表明该骰子不是均匀制造。

6/25



7/25

1. 拟合优度检验

[例 12-2] 某一演出策划公司对观众的年龄分布进行了如下预想：20-29岁占15%, 30-39岁占30%, 40-49岁占25%, 50-59岁占20%, 60岁以上占10%。从近期的观众中随机抽取200名，请检验在显著性水平为5%的情况下，该策划公司的预想是否正确。

年龄段	20-29	30-39	40-49	50-59	60以上	合计
频数	32	65	47	38	18	200
预计比率	0.15	0.3	0.25	0.2	0.1	1.0

귀무가설 $H_0 : \vec{p} = (0.15, 0.3, 0.25, 0.2, 0.1)$ $\overrightarrow{np} = (30, 60, 50, 40, 30)$

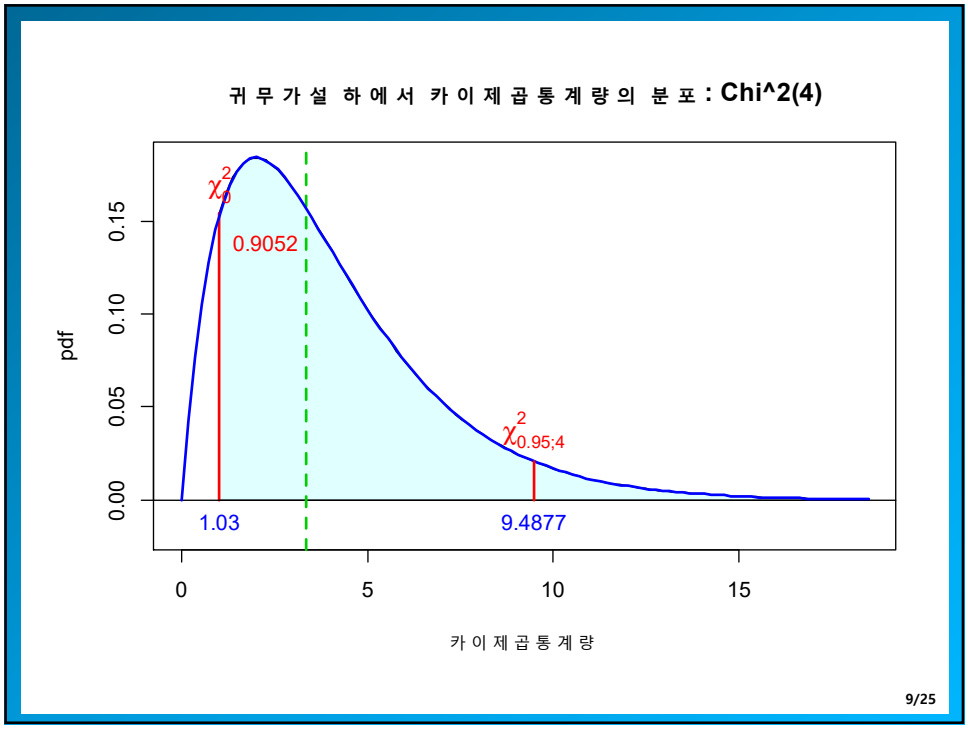
类别	1	2	3	4	5	合计
X_i	32	65	47	38	18	200
np_i	30	60	50	40	20	200
统计量	0.133	0.417	0.18	0.1	0.2	1.03

자유도는 $5-1=4$이므로 귀무가설의 기각역은 $\chi_0^2 > \chi_{0.95;4}^2 \doteq 9.488$

검정통계량은 $\chi_0^2 \doteq 1.03$으로 관측되었으므로 귀무가설을 채택

8/25

Kim, Dae Ho (Mokwon University)
mis@mokwon.ac.kr / mis@acm.org



9/25

1. 拟合优度检验

[例 12-3] 某一画廊预测每小时的访客量符合均值为4的泊松分布。 收集近期80个小时内的顾客访问量数，其结果如下所示。请在显著性水平为5%的情况下，检验该画廊的预测是否正确。

6 6 2 2 3 5 2 6 1 3 2 1 3 9 6 3 0 3 2 2
2 4 4 4 4 1 4 2 4 4 4 4 4 2 7 6 7 4 7 3
3 8 1 5 3 8 7 4 2 6 2 6 4 4 4 2 5 2 2 7
4 4 5 4 2 7 7 4 3 4 4 6 3 3 3 3 4 2 2 7

시간당 방문 관람객 수를 Y
귀무가설 $H_0: Y \sim Poi(4)$
기대도수 $np(y) = n \times e^{-4}4^y/y!$

✓ 为了减小卡方分布近似中的误差，预计频数小于5的类别0, 1以及类别7, 8, 9进行相加计算。

$e^{-4}(1+4)$ $P(Y \geq 7)$

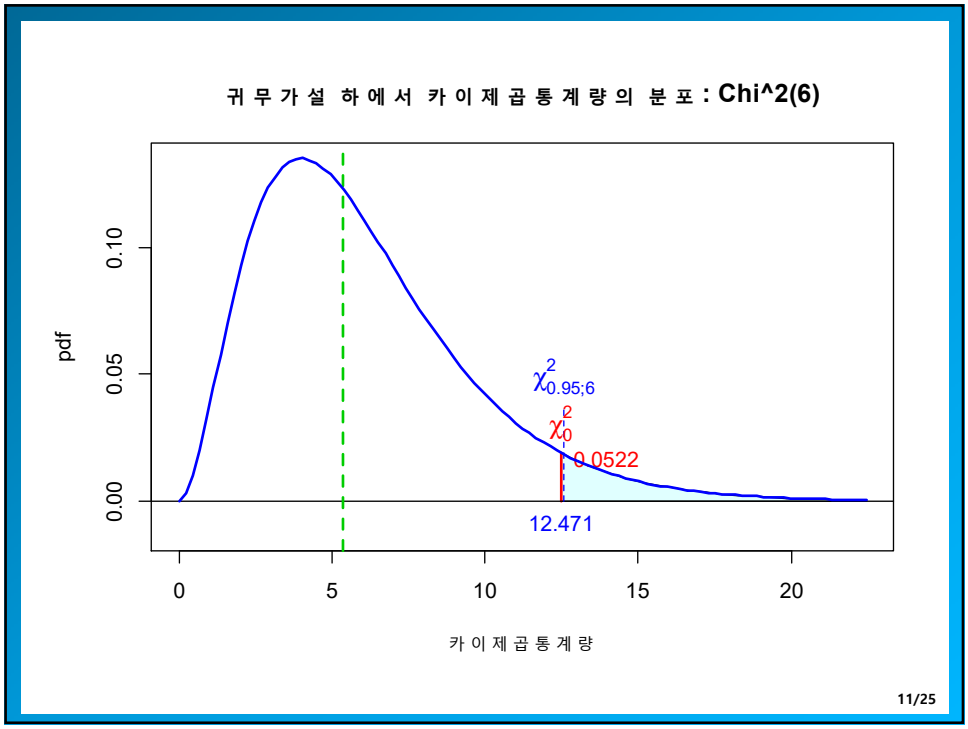
类别 (y)	0	1	2	3	4	5	6	7	8	9	合计
频数 (Xij)	1	4	17	13	22	4	8	8	2	1	80
n*p(y)	7.326		11.722	15.629	15.629	12.503	8.336	8.854			80.0
统计量	0.739		2.376	0.442	2.597	5.783	0.014	0.520			12.471

검정통계량은 $\chi_0^2 \doteq 12.471$ $< \chi^2_{0.95;6}$ 12.592 → 原假设被接受

→ 在显著性水平为5%的情况下，没有充分证据表明每小时的访客量不符合均值为4的泊松分布。

10/25

Kim, Dae Ho (Mokwon University)
mis@mokwon.ac.kr / mis@acm.org



1. 拟合优度检验

[例 12-4] 请检验在显著性水平为5%的情况下，[例 12-3]中每小时的访客量数是否符合泊松分布。

$$H_0 : Y \sim Poi(\lambda)$$

$$\hat{\lambda} = \frac{\sum_{y=0}^{9} yX_y}{\sum_{y=0}^{9} X_y} = \frac{314}{80} = 3.925 \Rightarrow np(y) = n\frac{e^{-3.925} \times 3.925^y}{y!}$$

✓ 为了减小卡方分布近似中的误差，预计频数小于5的类别0，1以及类别7，8，9进行相加计算。

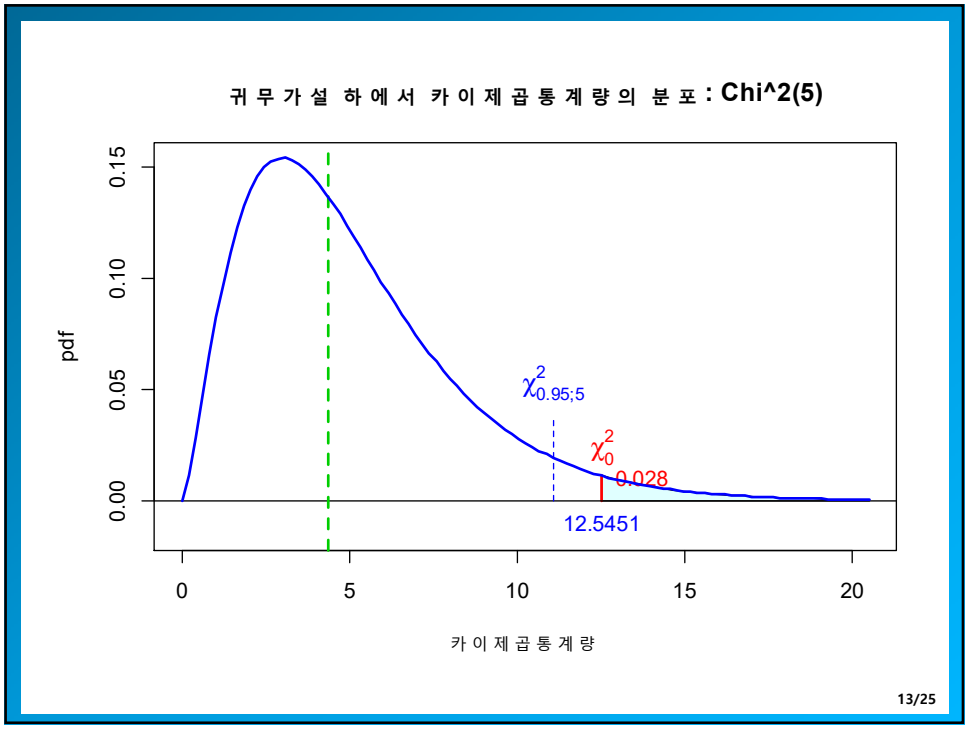
类别 (y)	0	1	2	3	4	5	6	7	8	9	合计
频数 (Xij)	1	4	17	13	22	4	8	8	2	1	80
n*p(y)	7.778		12.166	15.917	15.618	12.260	8.020	8.241			80.0
统计量	0.992		1.921	0.534	2.608	5.565	0.000	0.924			12.545

$\chi_0^2 \doteq 12.545 > \chi_{0.95;5}^2 \doteq 11.070$ → **原假设被接受**

→ **在显著性水平为5%的情况下，有证据表明每小时的访客量数不符合泊松分布。**

12/25

Kim, Dae Ho (Mokwon University)
mis@mokwon.ac.kr / mis@acm.org



13/25

2. 交叉分析

- 交叉分析: 用来分析两个分类变量之间的关联性的方法
- 交叉表(cross table), 双向列联表(two-way contingency table)

2.1 同质性检验

- 分析不受一个变量变化的影响的其他变量的分布是否一致

$$p_{ij} \equiv P(X=i, Y=j) \quad p_{i+} \equiv P(X=i),\ p_{+j} \equiv P(Y=j)$$

귀무가설 H_0 : "Y의 분포는 X와 상관없이 동일하다"

$$\Leftrightarrow H_0 : (p_{i1}, p_{i2}, \cdots, p_{ic}) = (p_{+1}, p_{+2}, \cdots, p_{+c}),\ i = 1, 2, \ldots, r$$

X \ Y	1	2	⋯	c	계
1	N_{11}	N_{12}	⋯	N_{1c}	n_{1+}
2	N_{21}	N_{22}	⋯	N_{2c}	n_{2+}
⋮	⋮	⋮	⋮	⋮	⋮
r	N_{r1}	N_{r2}	⋯	N_{rc}	n_{r+}
계	N_{+1}	N_{+1}	⋯	N_{+c}	n

14/25

Kim, Dae Ho (Mokwon University)
mis@mokwon.ac.kr / mis@acm.org

2. 交叉分析

귀무가설 H_0 : "Y의 분포는 X와 상관없이 동일하다"

$\Leftrightarrow H_0 : (p_{i1}, p_{i2}, \cdots, p_{ic}) = (p_{+1}, p_{+2}, \cdots, p_{+c}),\ i = 1, 2, \ldots, r$

X \ Y	p_{+1}	p_{+2}	$\cdots$	p_{+c}	계
1	$n_{1+}p_{+1}$	$n_{1+}p_{+2}$	$\cdots$	$n_{1+}p_{+c}$	n_{1+}
2	$n_{2+}p_{+1}$	$n_{2+}p_{+2}$	$\cdots$	$n_{2+}p_{+c}$	n_{2+}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
r	$n_{r+}p_{+1}$	$n_{r+}p_{+2}$	$\cdots$	$n_{r+}p_{+c}$	n_{r+}
계	N_{+1}	N_{+1}	$\cdots$	N_{+c}	n

$$E_{ij} \equiv E(N_{ij} \mid H_0) = n_{i+}p_{+j}$$

15/25

2. 交叉分析

$$E_{ij} \equiv E(N_{ij} \mid H_0) = n_{i+}p_{+j}$$

$$\Rightarrow \sum_{j=1}^{c} \frac{(N_{ij} - E_{ij})^2}{E_{ij}} = \sum_{j=1}^{c} \frac{(N_{ij} - n_{i+}p_{+j})^2}{n_{i+}p_{+j}} \sim \chi^2(c-1)$$

$$\Rightarrow \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(N_{ij} - E_{ij})^2}{E_{ij}} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(N_{ij} - n_{i+}p_{+j})^2}{n_{i+}p_{+j}} \sim \chi^2(r(c-1))$$

$$\hat{p}_{+j} = \frac{N_{+j}}{n} \Rightarrow \quad \chi_0^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(N_{ij} - n_{i+}N_{+j}/n)^2}{n_{i+}N_{+j}/n} \sim \chi^2((r-1)(c-1))$$

- 자유도 $r(c-1) - (c-1) = (r-1)(c-1)$

[定理 12-2] 同质性检验(test of homogeneity)

귀무가설 H_0 : "Y의 분포는 X와 상관없이 동일하다"

$$\chi_0^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(N_{ij} - n_{i+}N_{+j}/n)^2}{n_{i+}N_{+j}/n} \sim \chi^2((r-1)(c-1)) \,\Big|\, H_0$$

$\chi_0^2 > \chi_{1-\alpha;(r-1)(c-1)}^2 \Rightarrow$ Reject H_0

16/25

Kim, Dae Ho (Mokwon University)
mis@mokwon.ac.kr / mis@acm.org

2. 交叉分析

[例 12-5] 将四条生产线上生产的产品分为5个等级进行管理。从各生产线上随机抽取100个产品进行分析，其结果如下所示。请检验在显著性水平为5%的情况下，产品的等级是否符合与生产线无关的某一特定分布。

生产线 等级	1	2	3	4	5	合计
1	20	16	29	21	14	100
2	14	22	26	25	13	100
3	18	24	32	18	8	100
4	8	18	33	16	25	100
合计	60	80	120	80	60	400

生产线 等级	1	2	3	4	5	合计
1	1.667	0.800	0.033	0.050	0.067	2.617
2	0.067	0.200	0.533	1.250	0.267	2.317
3	0.600	0.800	0.133	0.200	3.267	5.000
4	3.267	0.200	0.300	0.800	6.667	11.233
合计	5.600	2.000	1.000	2.300	10.267	21.167

$$\vec{p} = (60,80,120,80,60)/400 = (0.15,0.2,0.3,0.2,0.15)$$

$$n_{i+}\vec{p} = 100 \times \vec{p} = (15,20,30,20,15)$$

$$(20-15)^2/15 \doteq 1.667$$

자유도: $(4-1)(5-1)=12$

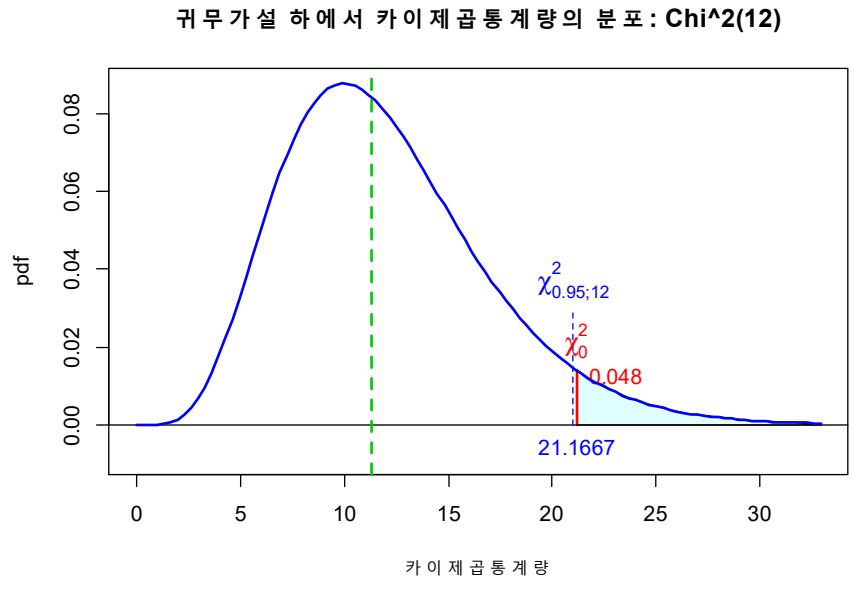
$$\Rightarrow \chi^2_{0.95;12} \doteq 21.026$$

$$\chi^2_0 \doteq 21.167 > 21.026$$

→ 原假设被拒绝

→ 有充足的证据表明产品的等级分布根据生产线的不同而不同。

17/25



18/25

Kim, Dae Ho (Mokwon University)
mis@mokwon.ac.kr / mis@acm.org

2. 交叉分析

[例 12-6] 由于[例 12-5]中生产线4存在品质问题，因此只运行除此之外的3条生产线，请检验在显著性水平为5%的情况下，产品的等级是否符合与生产线1、2、3无关的某一特定分布。

生产线 等级	1	2	3	4	5	合计
1	20	16	29	21	14	100
2	14	22	26	25	13	100
3	18	24	32	18	8	100
合计	52	62	87	64	35	300

$\vec{p} = (52, 62, 87, 64, 35) / 300 \doteq (0.173, 0.207, 0.290, 0.213, 0.117)$

$n_{i+}\vec{p} = 100 \times \vec{p} \doteq (17.333, 20.667, 29, 21.333, 11.667)$ $(20 - 17.333)^2 / 17.333 \doteq 0.410$

生产线 等级	1	2	3	4	5	合计
1	0.410	1.054	0.000	0.005	0.467	1.936
2	0.641	0.086	0.310	0.630	0.152	1.820
3	0.026	0.538	0.310	0.521	1.152	2.547
合计	1.077	1.677	0.621	1.156	1.771	6.303

자유도: $(3-1)(5-1) = 8$

$\Rightarrow \chi^2_{0.95;8} \doteq 15.507$

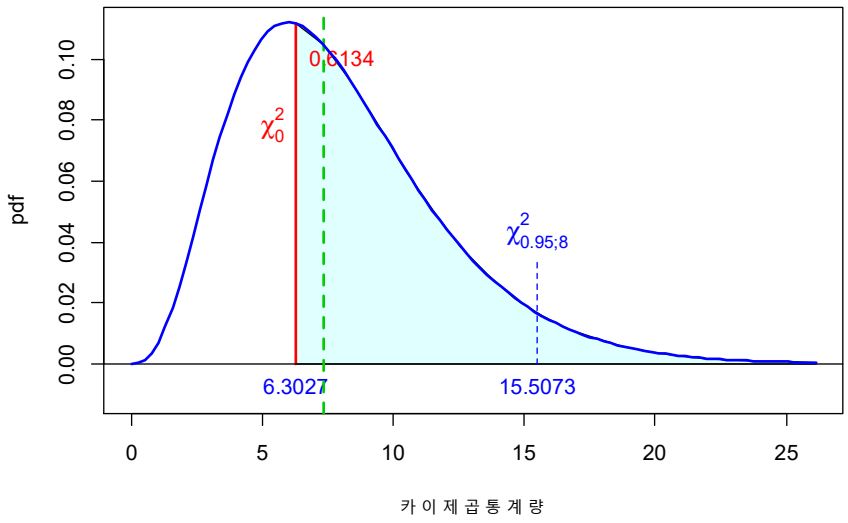
$\chi^2_0 \doteq 6.303 < 15.507$

→ 原假设被接受

→ 没有充足的证据表明产品的等级分布根据生产线的不同而不同。

19/25

귀무가설 하에서 카이제곱통계량의 분포: Chi^2(8)



20/25

2. 交叉分析

2.2 独立性检验

X \ Y	1	2	…	c	계
1	N_{11}	N_{12}	…	N_{1c}	N_{1+}
2	N_{21}	N_{22}	…	N_{2c}	N_{2+}
⋮	⋮	⋮	⋮	⋮	⋮
r	N_{r1}	N_{r2}	…	N_{rc}	N_{r+}
계	N_{+1}	N_{+1}	…	N_{+c}	n

$p_{ij} \equiv P(X=i, Y=j)$ $p_{i+} \equiv P(X=i)$, $p_{+j} \equiv P(Y=j)$

귀무가설 H_0 : "X와 Y는 서로 독립이다"

귀무가설이 사실이라면 모든 i, j에 대해 $p_{ij} = p_{i+}p_{+j}$

$$E_{ij} \equiv E(N_{ij} \mid H_0) = np_{i+}p_{+j}$$

$$\hat{p}_{i+} = \frac{N_{i+}}{n} \quad \hat{p}_{+j} = \frac{N_{+j}}{n} \quad \Rightarrow \hat{E}_{ij} = n\hat{p}_{i+}\hat{p}_{+j} = \frac{N_{i+}N_{+j}}{n}$$

21/25

2. 交叉分析

[定理 12-3] 独立性检验(test of independence)

귀무가설 H_0 : "X와 Y는 서로 독립이다"

$$\chi_0^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(N_{ij} - N_{i+}N_{+j}/n)^2}{N_{i+}N_{+j}/n} \sim \chi^2((r-1)(c-1)) \mid H_0$$

$\chi_0^2 > \chi_{1-\alpha;(r-1)(c-1)}^2 \Rightarrow$ Reject H_0

$$\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(N_{ij} - np_{i+}p_{+j})^2}{np_{i+}p_{+j}} \sim \chi^2(rc-1)$$

$$\hat{p}_{i+} = \frac{N_{i+}}{n}, \hat{p}_{+j} = \frac{N_{+j}}{n} \Rightarrow \quad \chi_0^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(N_{ij} - N_{i+}N_{+j}/n)^2}{N_{i+}N_{+j}/n} \sim \chi^2((r-1)(c-1))$$

- 自由度 $(rc-1) - [(r-1)+(c-1)] = (r-1)(c-1)$

22/25

Kim, Dae Ho (Mokwon University)
mis@mokwon.ac.kr / mis@acm.org

2. 交叉分析

[例 12-7] 随机选取400名中学生，对其喜好的学科和未来愿望进行调查。请检验在显著性水平为5%的情况下，中学生们所喜好的学科与未来愿望是否相互独立。

学科 \ 愿望	教师	公务员	专职人员	公司职员	其它	合计
语文	39	18	12	31	14	114
英语	35	23	18	35	13	124
数学	27	16	17	24	8	92
其它	9	12	8	19	22	70
合计	110	69	55	109	57	400

$\hat{E}_{ij} = N_{i+}N_{+j} / n$

学科 \ 愿望	教师	公务员	专职人员	公司职员	其它	合计
语文	31.35	19.665	15.675	31.065	16.245	114
英语	34.1	21.39	17.05	33.79	17.67	124
数学	25.3	15.87	12.65	25.07	13.11	92
其它	19.25	12.075	9.625	19.075	9.975	70
合计	110	69	55	109	57	400

23/25

2. 交叉分析

$$\chi_0^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(N_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}} \qquad \frac{(39-31.35)^2}{31.35} \doteq 1.867$$

学科 \ 愿望	教师	公务员	专职人员	公司职员	其它	合计
语文	1.867	0.141	0.862	0.000	0.310	3.180
英语	0.024	0.121	0.053	0.043	1.234	1.475
数学	0.114	0.001	1.496	0.046	1.992	3.649
其它	5.458	0.000	0.274	0.000	14.496	20.229
合计	7.463	0.264	2.685	0.089	18.033	28.533

자유도: $(4-1)(5-1) = 12$

$\Rightarrow \chi^2_{0.95;12} \doteq 21.026$

$\chi_0^2 \doteq 28.533 > 21.026$ → **原假设被拒绝**

→ **有证据表明中学生们所喜好的学科与未来愿望不是相互独立。**

24/25

Kim, Dae Ho (Mokwon University)
mis@mokwon.ac.kr / mis@acm.org