3.3.2 범주형(이산형) 데이터 : 카이제곱 검정

- 범주형(이산형) 데이터인 경우, 속성 A와 B 사이의 상관관계는 피어슨(Pearson)의 카이제곱(χ^2) 검정에 의해 측정

범주형(이산형) 데이터인 경우, 속성 A와 B 사이의 상관관계는 피어슨(Pearson)의 카이제곱(χ^2)검정에 의해 측정될 수 있다. 속성 A가 c개의 범주 값 $a_1, a_2, ..., a_c$를 취하고, 속성 B는 r개의 범주 값 $b_1, b_2, ..., b_r$을 취한다고 가정하자. 그러면, 속성 A와 B에 의해 구성되는 투플은 c개의 열과 r개의 행으로 구성되는 분할표로 표현될 수 있다. (A_i, B_j)를 속성 A가 a_i를 취하고 속성 B가 b_j를 취하는 투플이라고 할 때, χ^2은 다음과 같이 정의된다.

$$\chi^2 = \sum_{i=1}^{c}\sum_{j=1}^{r} = \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \qquad \text{식(3-2)}$$

- o_{ij} : (A_i, B_j)에 대한 관측도수(observed frequency; 실제로 존재하는 (A_i, B_j) 투플 수)
- e_{ij} : (A_i, B_j)에 대한 기대도수(expected frequency; 확률적으로 기대되는 (A_i, B_j) 투플 수)

e_{ij}는 다음과 같이 계산된다.

$$e_{ij} = \frac{count(A = a_i) \times count(B = b_j)}{N} \qquad \text{식(3-3)}$$

- N : 데이터 투플 수
- $count(A = a_i)$: 속성 A에 대하여 a_i를 갖는 투플 수
- $count(B = b_j)$: 속성 B에 대하여 b_j를 갖는 투플 수

식 (3-2)에서의 합계는 분할표상의 $r \times c$ 개의 모든 셀에 대하여 계산된다. 따라서 χ^2 값에 가장 크게 기여하는 칸은 실제 관측도수와 기대도수의 차이가 매우 큰 셀이다.

이러한 χ^2 통계량은 속성 A와 B가 독립이라는 가설을 검증한다. 이 검정은 자유도 $(r-1)\times(c-1)$을 갖는 유의수준에 근거하여 검정한다.

3.3.2 범주형(이산형) 데이터 : 카이제곱 검정

- 카이제곱 통계량을 이용한 범주형 속성에 대한 상관분석 사례

어떤 설문조사에서 1,500명의 사람들을 대상으로 각 사람에 대한 성별과 선호 글의 픽션 여부 간의 상관관계를 분석하고자 한다. 성별 속성값은 'male', 'female'이 있고, 픽션 여부 속성값은 'fiction'과 'non-fiction'이 있다. 성별 속성값 분포는 male:female=300:1,200이고 픽션 여부 속성값 분포는 fiction:non-fiction=450:1,050이다. 이들 속성의 조합에 대한 관측도수를 기록한 2×2 분할표는 다음과 같다.

	male	female	Total
fiction	250	200	450
non–fiction	50	1000	1050
Total	300	1200	1500

표에서 각 셀에 대한 기대도수는 식 (3-3)에 의해 구할 수 있다. 예를 들어, 셀 (female, fiction)의 기대도수는

$$e_{12} = \frac{count(femal) \times count(fiction)}{N} = \frac{1200 \times 450}{1500} = 360$$

이와 같은 방식으로 다른 셀에 대한 기대도수를 구하면 다음과 같은 기대도수 분할표를 얻는다.

	male	female	Total
fiction	90	360	450
non–fiction	210	840	1050
Total	300	1200	1500

이제 식 (3-2)에 의해서 χ^2 통계량을 구하면 다음과 같다.

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840}$$

$$= 284.44 + 121.90 + 71.11 + 30.48 = 507.93$$

3.3.2 범주형(이산형) 데이터 : 카이제곱 검정

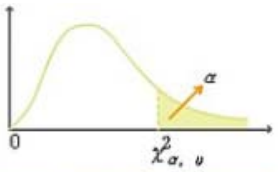
- 카이제곱 통계량을 이용한 범주형 속성에 대한 상관분석 사례

χ^2 통계량을 산출하였으므로 카이제곱 검정 방식에 의해 유의수준 0.05 수준에서의 두 속성 사이에 연관성이 있다라는 가설을 검증해 보자. 본 예제에서의 자유도는 (2−1)×(2−1)=1이 된다. 자유도 1일 때, χ^2 통계량(507.93)에 대한 유의확률(p-value)은 2.2e−16이다(이 수치는 자유도 1일 때의 카이제곱분포로 획득 가능함). 이는 유의수준 0.05보다 훨씬 작은 값으로서 대립가설(두 속성은 연관성이 있다)[7]은 채택된다. 자유도 1일 때 유의수준 0.05로 대립가설을 기각하는 데 필요한 값은 3.842로서 이 값보다 작아야 가설이 기각되는데, χ^2 통계량은 507.93으로 3.842에 비해 매우 크므로 가설은 채택되고, 실제 두 속성 사이에는 강한 연관성이 있다고 결론지을 수 있다.

7) 상관분석 검정의 가설은 크게 귀무가설(두 속성은 연관성이 없다)과 대립가설(두 속성은 연관성이 있다)로 나누어진다.

3.3.2 범주형(이산형) 데이터 : 카이제곱 검정

<카이제곱 분포표>



유의확률(p-value)

자유도(dof)

ν	α=.995	α=.99	α=.975	α=.95	α=.05	α=.025	α=.01	α=.005	ν
1	.3333930	.000157	.000982	.00393	3.841	5.024	6.635	7.879	1
2	.0100	.0201	.0506	.103	5.991	7.378	9.210	10.597	2
3	.0717	.115	.216	.352	7.815	9.348	11.345	12.838	3
4	.207	.297	.484	.711	9.488	11.143	13.277	14.860	4
5	.412	.554	.831	1.145	11.070	12.832	15.086	16.750	5
6	.676	.872	1.237	1.635	13.592	14.449	16.812	18.548	6
7	.989	1.239	1.690	2.167	14.067	16.013	18.475	20.278	7
8	1.344	1.646	2.180	2.733	15.507	17.535	20.090	21.955	8
9	1.735	2.088	2.700	3.325	16.919	19.023	21.666	23.589	9
10	2.156	2.558	3.247	3.940	18.307	20.483	23.209	25.188	10
11	2.603	3.053	3.816	4.575	19.675	21.920	24.725	26.757	11
12	3.074	3.571	4.404	5.226	21.026	23.337	26.217	28.300	12
13	3.565	4.107	5.009	5.892	22.362	24.736	27.688	29.819	13
14	4.075	4.660	5.629	6.571	23.685	26.119	29.141	31.319	14
15	4.601	5.229	6.262	7.261	24.996	27.488	30.578	32.801	15
16	5.142	5.812	6.908	7.962	26.296	28.845	32.000	34.267	16
17	5.697	6.408	7.564	8.672	27.587	30.191	33.409	35.718	17
18	6.265	7.015	8.231	9.390	28.869	31.526	34.805	37.516	18
19	6.844	7.633	8.907	10.117	30.114	32.852	36.191	38.582	19
20	7.434	8.260	9.591	10.851	31.410	34.170	37.566	39.997	20