【특강 1】 정규분포와 관련된 표본의 분포

① 통계적 추정과 가설검정은 정규분포와 어떤 관계가 있을까?

㉠ 표본자료들 대부분이 정규분포 모집단으로부터 얻어진다.

㉡ 모집단이 정규분포가 아니더라도 표본자료의 개수가 많으면 중심극한 정리에 의하여 정규분포를 이룬다.

㉢ 따라서 대부분의 통계분석은 모집단이 정규분포에 따른다고 가정하여 모수에 대한 통계적 추론(신뢰구간 추정이나 가설검정)을 수행한다.

㉣ 단, 분포의 선택은 어떤 변수들에 대한 추론이냐에 따라 연구자가 현명하게 선택해야 한다.

<table>
<tr><td colspan="2" rowspan="2"></td><td colspan="2">종속변수</td></tr>
<tr><td>질적변수
(명목, 서열)</td><td>양적변수
(등간, 비율)</td></tr>
<tr><td rowspan="2">독
립
변
수</td><td>질적
변수</td><td></td><td></td></tr>
<tr><td>양적
변수</td><td></td><td></td></tr>
</table>

② 정규분포와 χ^2-분포, $t-$분포, $F-$분포의 관계는?

㉠ 통계적 분포는 일반적으로 어떤 변수의 값이 일어날 확률을 설명하여 준다. 표준정규분포가 대표적인 예이며 Z검정을 위해 사용된다.

㉡ 그러나 모든 통계적 검정방법이 표준정규분포에 의하여 의사가 결정되는 것이 아니다.

㉢ 실제 통계분석에서 주로 활용되는 분포로는 카이제곱 분포, t분포, F분포 3가지 분포가 있다.

정규분포

⇩

표준정규분포 $Z-$분포 ⇨ χ^2-분포 ⇨ $t-$분포

χ^2-분포 ⇩ $F-$분포

③ 한 집단(정규분포)의 신뢰구간 추정 및 가설검정

㉠ 모평균

㉡ 모분산 ㉢ 모비율

㉣ 적합도 검정

④ 두 집단 비교 (서로 독립인 정규분포에 따른 모집단 X_1, X_2)

㉠ 평균

㉡ 분산 ㉢ 모비율

㉣ 회귀분석, 상관분석

1. χ^2(카이제곱)분포

① 카이(χ)는 X의 그리스 알파벳으로 평균 0, 분산 1인 표준정규분포를 의미한다. 따라서 카이제곱이라는 이름에는 표준정규분포를 제곱한다는 의미가 내포되어 있다.

② 데이터들이 흩어져 있는 정도인 분산이 퍼져있는 모습을 분포로 만들었으며, 분산의 제곱된 값을 다루기 때문에 χ^2분포라고 불린다. 따라서 χ^2분포는 Z분포의 표준점수를 제곱하여 만들어진 분포이다.

③ 표준정규분포 $N(0,\ 1)$을 따르는 ν개의 독립된 확률변수 $Z_1,\ Z_2, \ldots, Z_\nu$의 제곱합 $\chi^2 = \sum_{i=1}^{\nu} Z_i^2$를 자유도가 ν인 χ^2분포라고 한다.

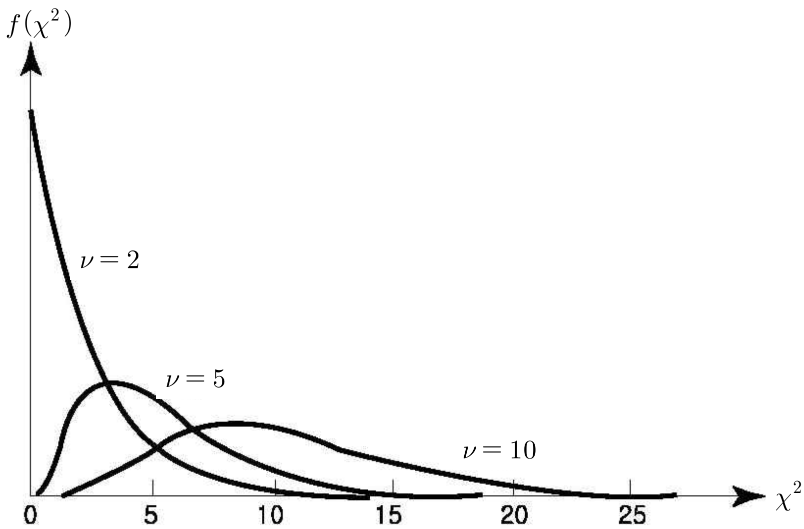
$$\sum_{i=1}^{\nu} Z_i^2 = Z_1^2 + Z_2^2 + \ldots + Z_\nu^2 \sim \chi_\nu^2$$

④ 자유도 ν인 χ^2분포의 확률밀도함수는

$$f(\chi^2) = \frac{1}{2^{\frac{\nu}{2}} \Gamma\left(\frac{\nu}{2}\right)} (\chi^2)^{\frac{\nu}{2}-1} e^{-\frac{\chi^2}{2}} \quad (0 \le \chi^2 < \infty)$$

이고, 평균은 $E(\chi^2) = \nu$, 분산 $V(\chi^2) = 2\nu$이다.

⑤ χ^2분포는 자유도 ν에 따라 그 형태가 달라지는 가족분포(family distribution)이다.
(예) $\nu = 2$인 χ^2분포는 평균이 2인 지수분포와 동일하다.



그리고 제곱된 값을 다루기 때문에 양수 값만 존재하며, 일반적으로 정적편포(positively skewed distribution)의 형태이고, 자유도가 무한대로 증가하면 정규분포의 형태를 나타낸다.

⑥ χ^2분포는 연속확률분포이면서 표본분포이며, 직접 확률을 구할 때 사용하는 분포가 아니라 모집단의 분산에 대한 추정과 검정, 적합도 검정, 동질성 검정, 독립성 검정 등에 사용하는 분포이다.

⑦ 정규분포 $N(\mu,\ \sigma^2)$으로부터 사례 수가 n인 표본 $X_1,\ X_2,\ \cdots,\ X_n$을 추출하였을 때, 각각을 표준화하여 제곱한 후 더하여 만들어진 분포는 자유도(degree of freedom) $n-1$인 χ^2분포가

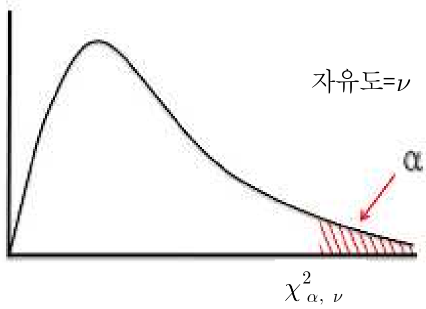
된다.

$$\sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{\sigma}\right)^2=\frac{1}{\sigma^2}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2=\frac{(n-1)s^2}{\sigma^2}\ \sim\ \chi^2_{n-1}$$

⑧ 카이제곱 검정통계량 $\chi^2=\dfrac{(n-1)s^2}{\sigma^2}$ 이다.

⑨ $\chi^2_{\alpha,\ \nu}$은 χ^2분포에서 오른쪽 꼬리부분의 확률이 α가 되는 χ^2확률변수의 값을 나타낸다.

$$P[\chi^2_\nu \geq\ \chi^2_{\alpha,\ \nu}]=\alpha$$



확률

자유도 $\nu(=n-1)$

ν \ α	0.995	0.99	0.975	0.95
1	0.00004	0.0002	0.001	0.004
2	0.01	0.02	0.05	0.10
3	0.07	0.11	0.22	0.35
4	0.21	0.30	0.48	0.71
5	0.41	0.55	0.83	1.15
6	0.68	0.87	1.24	1.64
7	0.99	1.24	1.69	2.17
8	1.34	1.65	2.18	2.73
9	1.73	2.09	2.70	3.33

⑩ χ^2분포표

⑪ 그래프의 총면적이 1이므로 오른쪽 면적이 α일 때 왼쪽 면적은 $1-\alpha$이다.

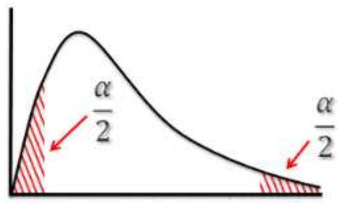
$1-\alpha$ $\qquad\qquad$ α

ν \ α	0,995	0,99	0,975	0,95	0,9	0,5	0,1	0,05	0,025	0,01	0,005
1	0.00004	0.0002	0.001	0.004	0.02	0.45	2.71	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	0.21	1.39	4.61	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	0.58	2.37	6.25	7.81	9.35	11.34	12.84
4	0.21	0.30	0.48	0.71	1.06	3.36	7.78	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.61	4.35	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	1.64	2.20	5.35	10.64	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	2.83	6.35	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	7.34	13.36	15.51	17.53	20.09	21.95
9	1.73	2.09	2.70	3.33	4.17	8.34	14.68	16.92	19.02	21.67	23.59

[예] 자유도가 5, $\alpha=0.95$인 χ^2확률변수의 값 $\chi^2_{0.95,\ 5}=1.15$

자유도 10, $\alpha=0.05$인 χ^2확률변수의 값 $\chi^2_{0.05,\ 10}=18.31$

⑫ 양쪽 임계값을 구해 신뢰구간 추정과 가설검정을 할 때, α가 양쪽으로 나뉘기 때문에 $\alpha/2$를 이용한다.



〖예〗 어떤 철강재의 인장 강도는 정규분포를 따른다고 한다. 이 철강재의 모집단에서 $n=10$개의 표본을 추출하여 표본분산 s^2을 산출할 때, $P(\sigma^2 < ks^2)=0.90$되는 k값은?

[풀이]

㉠ $P(\sigma^2 < ks^2)=0.90$에서

$$P\left(\frac{s^2}{\sigma^2} > \frac{1}{k}\right)=0.90 \rightarrow P\left(\frac{(n-1)s^2}{\sigma^2} > \frac{n-1}{k}\right)=0.90$$

㉡ 즉, $P\left(\chi_9^2 > \frac{9}{k}\right)=0.90 \rightarrow \chi^2_{0.90,9} = \frac{9}{k}=4.17$ 따라서, $k=\frac{9}{4.17} \cong 2.1583$이다.

〖예〗 자동차 배터리 제조업자는 자신들의 배터리의 수명은 분산 0.9년이라고 주장한다. 그러나 이들 배터리 10개를 추출하여 분산을 구한 결과 1.2년이었다. $\sigma^2 > 0.9$라고 주장할 수 있는지 유의수준 0.05로 검정하여라.

[풀이]

㉠ $H_0 : \sigma^2 = 0.9$ $\qquad H_1 : \sigma^2 > 0.9$

㉡ 검정통계량의 계산된 값 $\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{(10-1)(1.2)}{0.9} = 12$

㉢ 대립가설은 우측검정이고, $\alpha = 0.05$, 자유도=9이므로 $\chi^2_{0.05,\,9} = 16.919$

따라서 검정통계량의 계산된 값이 16.919보다 크면 H_0를 기각

㉣ 그러나 검정통계량의 계산된 값 $12 < 16.919$이므로 H_0를 기각할 수 없음

즉, 유의수준 0.05에서 볼 때 모분산 σ^2이 0.9년보다다 크다고 할 만한 근거가 없음.