2. t분포

① 영국의 통계학자 고셋(W.S Gosset)이 모집단의 표준편차를 알지 못하여 표본의 표준편차로 모집단의 표준편차를 대신할 때 표본의 크기가 작을수록, 즉 자유도가 낮을수록 표준정규분포에서 벗어나는 확률분포를 나타낸다는 사실을 발견하였다.

② 자유도 ν인 t분포; 표준정규분포 $N(0,\ 1)$에 따르는 확률변수 Z와 자유도 ν의 χ^2분포에 따르는 확률변수 χ^2가 서로 독립일 때, $T=\dfrac{Z}{\sqrt{\dfrac{\chi^2}{\nu}}}$가 이루는 분포이다.

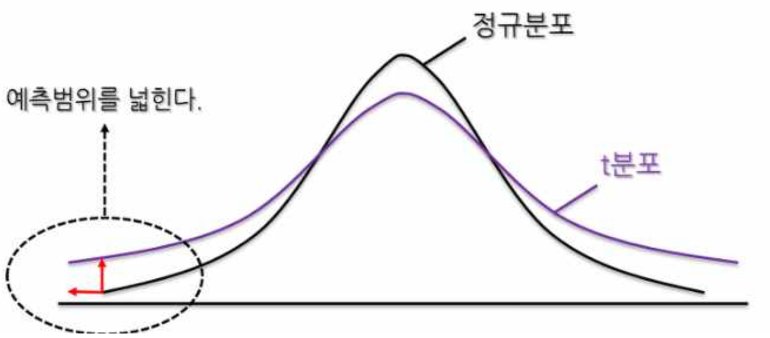
③ 자유도 ν인 t분포의 확률밀도함수는

$$f(t)=\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu\pi}}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}\quad(-\infty<t<\infty)$$

이며 $E(T)=0$, $V(T)=\dfrac{\nu}{\nu-2}\ (\nu>2)$이다.

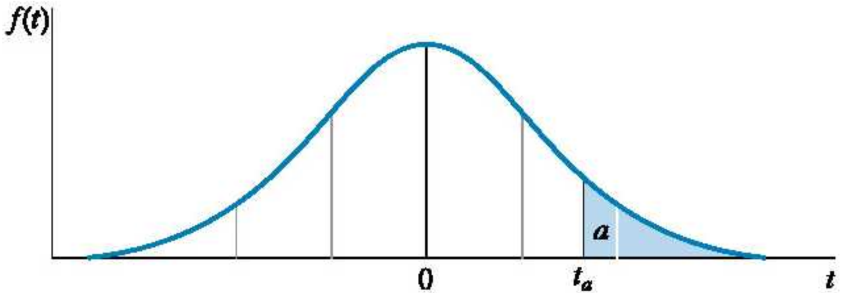
④ t분포는 자유도에 따라 형태가 다른 가족분포이며, 평균이 0이고 표준편차가 1보다 크다.

⑤ t분포는 $t=0$을 중심으로 좌우 대칭이며, 표준정규분포의 형태와 비슷하나 좌우 꼬리 부분의 확률이 좀 더 두텁다.



⑥ t분포는 자유도 ν가 커짐에 따라 표준정규분포에 접근한다.

⑦ $t_{\alpha,\nu}$; 자유도 ν인 t분포에서 오른쪽 꼬리부분의 확률이 α가 되는 t값이다.

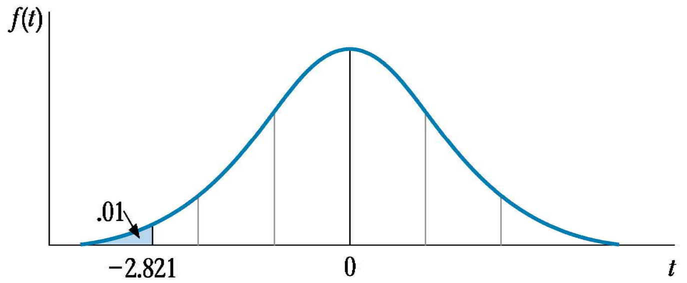


〚예〛 $\alpha=0.050$, $\nu=5$인 $t-$분포에서 $t_{0.050,\ 5}=2.015$

[예] 한 정규분포로부터 $n=10$인 표본을 추출, t값들의 영역의 넓이가 1%가 되는 t값은?

㉠ 자유도 $\nu=10-1=9$인 t분포 곡선에서, t값들의 영역이 0.01인 t값은 왼쪽 영역의 넓이가 0.01인 t값이 된다.

㉡ t분포는 0에 대하여 대칭이기 때문에, 이 값은 오른쪽 꼬리 영역의 넓이가 0.01인 t값에 음의 부호를 붙이면 된다. 즉, $-t_{0.01}=-2.821$이다.



df	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$	df
1	3.078	6.314	12.706	31.821	63.657	1
2	1.886	2.920	4.303	6.965	9.925	2
3	1.638	2.353	3.182	4.541	5.841	3
4	1.533	2.132	2.776	3.747	4.604	4
5	1.476	2.015	2.571	3.365	4.032	5
6	1.440	1.943	2.447	3.143	3.707	6
7	1.415	1.895	2.365	2.998	3.499	7
8	1.397	1.860	2.306	2.896	3.355	8
9	1.383	1.833	2.262	2.821	3.250	9
.	.	.	.	.	.	.
.	.	.	.	.	.	.
26	1.315	1.706	2.056	2.479	2.779	26
27	1.314	1.703	2.052	2.473	2.771	27
28	1.313	1.701	2.048	2.467	2.763	28
29	1.311	1.699	2.045	2.462	2.756	29
inf.	1.282	1.645	1.960	2.326	2.576	inf.

⑧ 모집단의 평균에 대한 추정 또는 검정을 할 때, 모분산 σ^2이 알려져 있지 않은 경우에는 $Z=\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}$의 분포를 이용할 수 없다. 혹시 표본의 크기가 클 때($n\geq 30$)에 σ^2대신 표본분산 s^2을 사용하여도 근사적으로 $\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,\ 1)$임이 알려져 있으나, 표본의 크기가 작은 경우($n<30$)이면 정규분포를 따르지 않는다. 이때는, 자유도 $n-1$인 t분포를 사용해야 한다.

⑨ 정규분포 $N(\mu,\ \sigma^2)$으로부터 사례 수가 n인 표본 $X_1,\ X_2,\ \cdots,\ X_n$을 추출하였을 때, $T=\dfrac{\overline{X}-\mu}{s/\sqrt{n}}$는 자유도 $n-1$인 t분포를 이룬다.

[**설명**] 표준정규분포 $N(0,\ 1)$에 따르는 확률변수 $Z=\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}$와 자유도 $n-1$의 χ^2_{n-1}가 서로 독립일 때, t분포의 정의로부터 $T=\dfrac{Z}{\sqrt{\dfrac{\chi^2}{n-1}}}=\dfrac{\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\dfrac{\dfrac{(n-1)s^2}{\sigma^2}}{n-1}}}=\dfrac{\overline{X}-\mu}{s/\sqrt{n}}$이다.

〚예〛 어떤 화공기사는 어느 배치공정의 수율이 원재료의 리터당 500g이라고 주장하고 있다. 그는 이를 입증하기 위해 25개의 배치를 추출하여 시험을 하였다. 시험결과로 계산된 t값이 $-t_{0.05}$와 $t_{0.05}$ 사이에 있으면 그의 주장이 타당성이 있다고 하기로 한다. 25개 배치의 시험결과 표본평균은 518g이고 표준편차는 40g이었다면 어떤 결과를 낼 수 있겠는가? (단, 모집단은 근사적으로 정규분포를 따른다고 가정)

[풀이]

㉠ 시험결과로부터 t값을 계산하면

$$t = \frac{518-500}{40/\sqrt{25}} = 2.25$$

이다.

㉡ 자유도가 24이므로 $t_{0.05} = 1.711$이다.

㉢ 계산된 t값이 1.711보다 크므로 실제의 수율은 500g보다 크다고 할 수 있다.

3. F분포

① F분포는 피셔(Sir Ronald Aylmer Fisher)가 고안하였으며 피셔의 이름을 기리기 위하여 F분포라고 명명되었다.

② F분포는 정규분포를 이루는 모집단에서 독립적으로 추출한 표본들의 분산비율이 나타내는 연속확률분포이다.

③ F분포는 두 가지 이상의 표본집단의 분산을 비교하거나 모집단의 분산을 추정할 때 쓰인다. 2개 이상의 표본평균들이 동일한 모평균을 가진 집단에서 추출되었는지 아니면 서로 다른 모집단에서 추출된 것인지를 판단하기 위하여 이용한다.

④ F분포는 서로 독립인 두 χ^2(카이제곱)확률변수(자유도가 각각 ν_1, ν_2)의 비율이 이루는 분포이며 $F(\nu_1,\ \nu_2)$로 표기한다.

$$F = \frac{{\chi_1}^2}{(\nu_1)} / \frac{{\chi_2}^2}{(\nu_2)}$$

⑤ $F-$분포 확률밀도함수는 자유도 ν_1, ν_2에 대하여

$$f(x) = \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)\left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} x^{\frac{\nu_1}{2}-1}}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)\left[\left(\frac{\nu_1}{\nu_2}\right)x+1\right]^{\frac{\nu_1+\nu_2}{2}}},\quad 0 < x < \infty$$

이며, 평균과 분산은 각각 $\mu = \dfrac{\nu_2}{\nu_2-2}\,(\nu_2 > 2), \sigma^2 = \dfrac{2{\nu_2}^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)^2(\nu_2-4)}\,(\nu_2 > 4)$이다.

⑥ χ^2(카이제곱)확률변수들이 제곱값이어서 항상 양수이므로 F값 역시 양수이다.

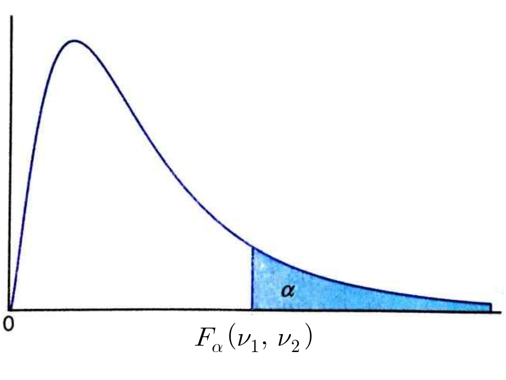
⑦ F분포는 자유도 ν_1, ν_2가 무한히 커지지 않는 한 정적 편포를 나타내며, F분포의 형태도 자유도 ν_1, ν_2에 따라 변화하는 형태를 나타내는 가족분포이다.

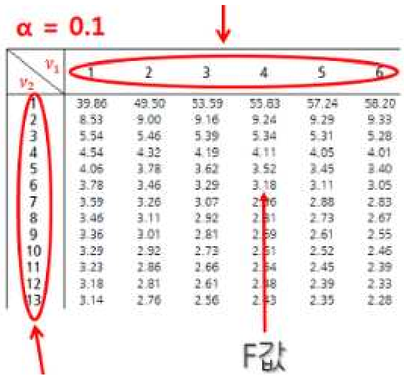
⑧ 두 개의 집단 중에서, 분산이 더 큰 집단이 분자, 상대적으로 분산이 작은 집단이 분모가 된다.

⑨ 두 집단의 분산을 추정하고 검정할 때 사용된다(카이제곱분포는 한 집단의 분산을 파악할 때 사용). 그리고 3개 이상 집단의 분산을 비교하는 분산분석에서도 사용된다.

⑩ $F_\alpha(\nu_1,\ \nu_2)$은 F분포에서 오른쪽 꼬리부분의 확률이 α가 되는 F확률변수의 값을 나타낸다.

$$P[F \geq\ F_\alpha(\nu_1,\ \nu_2)] = \alpha$$





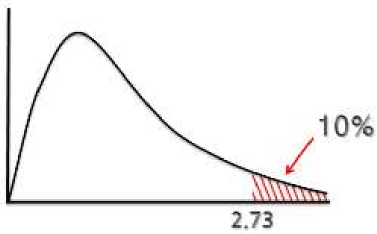
ν_2 \ ν_1	1	2	3	4	5	6
1	39.86	49.50	53.59	55.83	57.24	58.20
2	8.53	9.00	9.16	9.24	9.29	9.33
3	5.54	5.46	5.39	5.34	5.31	5.28
4	4.54	4.32	4.19	4.11	4.05	4.01
5	4.06	3.78	3.62	3.52	3.45	3.40
6	3.78	3.46	3.29	3.18	3.11	3.05
7	3.59	3.26	3.07	2.96	2.88	2.83
8	3.46	3.11	2.92	2.81	2.73	2.67
9	3.36	3.01	2.81	2.69	2.61	2.55
10	3.29	2.92	2.73	2.61	2.52	2.46
11	3.23	2.86	2.66	2.54	2.45	2.39
12	3.18	2.81	2.61	2.48	2.39	2.33
13	3.14	2.76	2.56	2.43	2.35	2.28

⑪ $F-$분포표

〖예〗 두 집단 A와 B에 대하여 집단 A(9개)의 표본분산 $s_A^2 = 2$, 집단 B(6개)의 표본분산 $s_B^2 = 3$이라고 한다. 이 때 유의수준 0.1에 해당하는 F값은?

α = 0.1

ν_2 \ ν_1	1	2	3	4	5	6
1	39.86	49.50	53.59	55.83	57.24	58.20
2	8.53	9.00	9.16	9.24	9.29	9.33
3	5.54	5.46	5.39	5.34	5.31	5.28
4	4.54	4.32	4.19	4.11	4.05	4.01
5	4.06	3.78	3.62	3.52	3.45	3.40
6	3.78	3.46	3.29	3.18	3.11	3.05
7	3.59	3.26	3.07	2.96	2.88	2.83
8	3.46	3.11	2.92	2.81	$\boxed{2.73}$	2.67
9	3.36	3.01	2.81	2.69	2.61	2.55



[풀이]

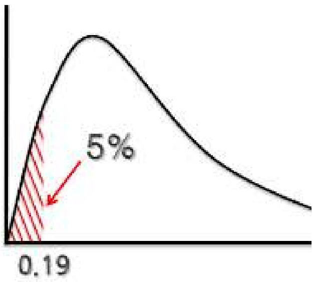
㉠ 집단 B의 분산이 더 크므로, B가 분자가 되고 A가 분모가 됨, A의 자유도=5, B의 자유도=8

㉡ 따라서 유의수준 0.1에 해당하는 F값 = 2.73

⑫ 확률변수 F가 자유도 (ν_1, ν_2)인 F분포를 따른다면, $\frac{1}{F}$는 자유도 (ν_2, ν_1)인 F분포를 따른다.

$$F_{1-\alpha}(\nu_1, \nu_2) = \frac{1}{F_\alpha(\nu_2, \nu_1)}$$

〖예〗 집합 A(n=6)의 표본분산=23, 집단 B(n=5)의 표본분산=15. 유의수준 0.05에서 왼쪽 F값은?

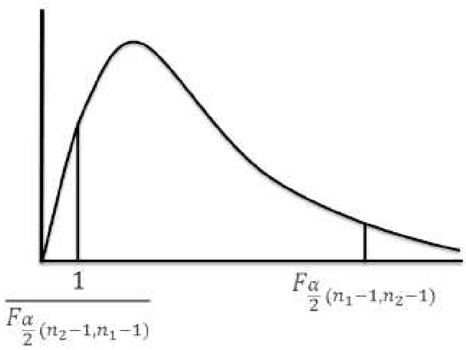


α = 0.05

ν_2 \ ν_1	1	2	3	4	5
1	161.4	199.5	215.7	224.6	230.2
2	18.51	19.00	19.16	19.25	19.30
3	10.13	9.55	9.28	9.12	9.01
4	7.71	6.94	6.59	6.39	6.26
5	6.61	5.79	5.41	$\boxed{5.19}$	5.05
6	5.99	5.14	4.76	4.53	4.39
7	5.59	4.74	4.35	4.12	3.97

[풀이] $F_{0.95}(5, 4) = \frac{1}{F_{0.05}(4, 5)} = \frac{1}{5.19} = 0.19$

⑬ 신뢰구간 추정에서 양쪽 F값을 구하는 경우



왼쪽 F값은 역수를 취하며, 분자와 분모의 자유도가 서로 바뀐다.

⑭ 서로 독립이며 정규분포를 따르는 두 모집단 $\{X_1 \sim N(\mu_1, \sigma_1^2)\}$과 $\{Y \sim N(\mu_2, \sigma_2^2)\}$에서 각각 크기 n_1, n_2의 확률표본에 대한 표본분산이 s_1^2, s_2^2일 때, $\dfrac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$는 $F(n_1-1,\ n_2-1)$을 따른다.

즉,

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \sim F(n_1-1,\ n_2-1)$$

[증명]

모집단 1과 모집단 2가 서로 독립이므로

$$\frac{\dfrac{s_1^2}{\sigma_1^2}}{\dfrac{s_2^2}{\sigma_2^2}} = \frac{\dfrac{\sum_{i=1}^{n_1}(X_i-\mu_1)^2}{n_1-1}\dfrac{1}{\sigma_1^2}}{\dfrac{\sum_{j=1}^{n_2}(Y_j-\mu_2)^2}{n_2-1}\dfrac{1}{\sigma_2^2}} = \frac{\dfrac{\sum_{i=1}^{n_1}(X_i-\mu_1)^2}{\sigma_1^2}\dfrac{1}{n_1-1}}{\dfrac{\sum_{j=1}^{n_2}(Y_j-\mu_2)^2}{\sigma_2^2}\dfrac{1}{n_2-1}} = \frac{\dfrac{\chi^2_{(n_1-1)}}{n_1-1}}{\dfrac{\chi^2_{(n_2-1)}}{n_2-1}} \sim F(n_1-1,\ n_2-1)$$

⑮ F분포는 회귀분석과 분산분석 등에서 주로 사용되는 분포이다. 두 집단의 모분산이 같다고 가정하면($\sigma_1^2 = \sigma_2^2$), 위 통계량이 두 표본분산의 비로 단순화되기 때문에 표본분산으로 분산의 크기를 비교 분석하는데 유용하게 사용된다.

〚예〛 서로 독립인 정규분포 $\{X_1 \sim N(\mu_1, \sigma_1^2)\}$, $\{X_2 \sim N(\mu_2, \sigma_2^2)\}$에서 각각 10개의 확률표본을 구했다. 이때 각 집단의 표본분산에 차이가 있어 두 표본분산비가 6.56으로 계산이 되었다. 두 집단의 분산이 같을 때, 이와 같이 두 집단의 표본분산비가 6.56 이상 나올 확률은 얼마인가?

[풀이] $\sigma_1^2 = \sigma_2^2$이고 $F_{(9,\,9)}$에서, 구하고자 하는 확률은

$$P\left(\frac{s_1^2}{s_2^2} \geq 6.56\right) = P\left(\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \geq 6.56\frac{\sigma_2^2}{\sigma_1^2}\right) = P(F \geq 6.56) = 0.00495$$

【참고】 자유도(number of degrees of freedom)

① 주어진 조건 하에서 자료 중 자유로이 값을 취할 수 있는 관찰 수이다.

〖예〗 $x+y+z=3$이라는 방정식에서 $x=1, y=0$으로 결정되면 $z=2$를 갖게 되므로 독립 변수는 2개이고 따라서 자유도는 2가 된다.

② 어떤 제약으로 인해 그 제약을 제외한 자유의지를 가질 수 있는 남은 개수이다.

〖예〗 n개의 관측치 데이터에서 '표본평균의 자유도 = n'

: 표본평균의 계산에서는 관찰치의 합계를 구하게 되므로 각 자료값이 어떤 값을 취대도 상관없으므로 자유도가 n이 된다.

③ 통계적 추정을 할 때 표본자료 중 모집단에 대한 정보를 주는 독립적인 자료의 수이다.

④ 표본 n개를 선택할 때, 마지막 1개는 모집단의 평균과 같아지도록 표본집합을 구성하도록 선택되어져야 하므로 그 자유를 상실하게 되어 자유도가 $n-1$이다.

〖예〗 표본분산의 자유도는 n-1이 된다. 왜냐하면 표본분산의 계산에서는 각 자료값과 표본평균과의 차이를 구하고 이를 제곱하여 그 합계를 구하는데, 이때의 표본평균은 주어진 자료로부터 계산된 것이다. 그러므로 표본평균이 알려져 있다는 것은 하나의 자료의 값이 이미 정해져 있는 것과 같은 효과를 가져오며 따라서 표본분산의 자유도는 n-1이 된다.

〖예〗 10개 관측치 데이터의 평균이 3.5라 할 때, 10개의 총합이 35가 되며, 9번째 값까지는 자유의지대로 데이터를 지정할 수 있다. 만약 34, -9.3, -37, -92, -1, 0, 1, -22, 99를 지정하면 마지막 10번째 데이터는 61.3으로 정해진다. 그러므로 위 예제는 자유도가 10-1=9이다.

⑤ 모집단의 평균을 안다면 $\sigma_X^2 = \dfrac{\sum(X_i-\mu)^2}{N}$이다. 그러나 모집단의 평균 μ를 알지 못하므로 표본에 의하여 모집단의 분산을 추정하는 공식은 μ 대신에 $\overline{X}$를 사용한다. 그리고 μ 대신에 $\overline{X}$를 사용하므로 모집단의 분산을 추정할 때 $\overline{X}$ 값에 의하여 제한을 받아 총 사례수에서 1을 뺀 $n-1$을 분모로 사용해야 한다. 이때, 모집단 분산의 불편추정치를 위한 표본의 분산계산 공식의 분모 $n-1$을 자유도라 지칭한다.