碩士學位論文

F-分布와 그 應用에 관한 研究

指導教授　金　益　贊

濟州大學校　教育大學院

數 學 教 育 專 攻

宋　宰　忠

1998年　8月

F-分布와 그 應用에 관한 硏究

指導敎授 金 益 贊

이 論文을 敎育學 碩士學位 論文으로 提出함

1998年 5月 日

濟州大學校 敎育大學院 數學敎育專攻

提出者 宋 宰 忠

宋宰忠의 敎育學 碩士學位 論文을 認准함

1998年 7月 日

審査委員長 ______________ 印

審 査 委 員 ______________ 印

審 査 委 員 ______________ 印

〈초 록〉

F-分布와 그 應用에 관한 研究

宋 宰 忠

濟州大學校 敎育大學院 數學敎育專攻

指導敎授 金 益 贊

F분포의 활용성을 현대 응용통계학의 각 분야에서 살펴봄으로서 F분포의 역할에 대한 중요성을 재인식하고 그 올바른 이용에 도움을 주고자 하는 것이 본 논문의 목적이다. 먼저 F분포가 Gamma 분포로부터 χ^2분포를 거쳐 유도되고, 변수변환의 방법으로도 유도될 수 있음을 보였다.

추정이론에서는 모분산비의 신뢰구간을 여러 경우에서 살펴보았고 계량적 성질, 생산제품의 관리유지와 관련되는 가설검정론에서의 F분포의 응용력을 제시하였다. 특히 분산분석과 실험계획법에서의 응용사례들 중 교육현장에서 나타나는 교육방법의 개선책에 대한 실례를 전산처리하여 F분포의 응용력을 부각시키고자 하였다.

目　次

I. 序 論

Ronald A. Fisher(1890-1962)의 명예를 기리기 위하여 그 첫 자를 따서 명명한 F분포는 미국 Iowa주립대학교의 George Snedecor(1881-1974)가 생각하였던 분포이론으로 Snedecor의 F분포로 명명되고 있다.

두모집단 분산비와 관련된 F분포는 추측통계학에서 다루는 추정론과 가설검정론에서 분산에 대한 영역을 추가하였음은 물론 분산분석과 회귀분석에서 그 활용의 영역을 극대화 함으로서 현대 통계학의 다양한 응용력에 대한 중요한 이론을 제공하고 있는 셈이다.

본 논문에서는 Gamma분포로부터 χ^2분포를 거쳐 F분포 함수를 정의하고 변수변환의 방법에 의해 그 밀도함수가 유도될 수 있음을 보일 것이다. 특히 F분포의 활용성을 고려하여 추정이론과 가설 검정론에서 F분포의 역할, 분산분석과 실험계획법 및 회귀분석에서 이용되는 F분포를 다양한 예제와 함께 보일 것이다. 한편 가설검정과 분산분석간의 관계가 F분포를 통해서 확인되고, 중등교육현장에서 나타난 교육방법의 효과를 측정하는 방법을 예제를 통해서 보임으로서 분산비를 활용하는 통계적 연구에 다소의 도움을 주려는 것이 본 논문의 목적이다.

II. Gamma 함수와 F분포

1. Gamma 함수

무한급수와 이상분포사이의 유사성에 비추어 멱급수

(2-1) $$G(x)=\sum_{n=0}^{\infty} f(n)\, x^n$$

을

(2-2) $$\int_0^{\infty} f(t)\, x^t dt$$

로 두고 x를 e^{-s}으로 치환하여 $f(t)$의 Laplace 변환인

(2-3) $$F(s)=\int_0^{\infty} f(t)\, e^{-st} dt$$

를 정의한다. $G(x)$의 수렴반경을 생각하듯이 (2-2)가 어떤 r에 대하여 $0 \le x < r$에서 수렴한다면 $F(s)$도 $0 < e^{-s} < r$에서 수렴한다. (2-3)의 적분을 함수 $F(t)$가 $F(s)$로 변환하는 operator L로 생각하여

(2-4) $$L(f)=F$$

로 표시하자.

$f(t)=t^k$으로 두었을 때의 Laplace의 변환 ;

(2-5) $$\int_0^{\infty} t^k e^{-st} dt$$

를 생각하자. 단 $t=0$에서 이 적분이 발산함을 피하기 위하여 매개변수 k는 -1보다 크다고 하자. 만일 $k=0$이거나 또는 양의 정수이면

(2-6) $$\int_0^\infty e^{-st}dt = \frac{1}{s}, \quad \int_0^\infty te^{-st}dt = \frac{1}{s^2}, \cdots \quad (s>0)$$

이 되고 일반적인 경우가 부분적분법에 의해

(2-7) $$\int_0^\infty t^k e^{-st}dt = \frac{k}{s}\int_0^\infty t^{k-1}e^{-st}dt \; ; \; s>0$$

가 됨을 확인할 수 있다. 귀납법을 적용함으로서 다음의 사실을 쉽게 알 수 있다.

(2-8) $$L(t^k) = \int_0^\infty t^k e^{-t}dt = \frac{k}{s}\cdot\frac{k-1}{s}\cdot\cdots\cdot\frac{2}{s}\cdot\frac{1}{s}\cdot\frac{1}{s} = \frac{k!}{s^{k+1}}, \quad s>0$$

따라서 $s=1$일 때

(2-9) $$k! = \int_0^\infty t^k e^{-t}dt$$

가 되어 계승에 대한 일반화를 가능하게 한다.

즉, $k>-1$인 임의의 실수일 때 $k!$는 (2-9)에 의해 정의되고 k가 0 또는 양의 정수이면

(2-10) $$k! = \Gamma(k+1) = \begin{cases} 1\cdot 2\cdot\cdots\cdot k & ; \quad k>0 \\ 1 & ; \quad k=0 \end{cases}$$

로 정의됨으로서 Gamma함수가 유도된다.

<u>정의 1 :</u>

양수 k에 대해서 다음과 같이 Gamma함수를 정의한다.

(2-11) $$\Gamma(k) = \int_0^\infty t^{k-1}e^{-t}dt \quad ; \quad k>0$$

한편 부분적분법에 의해 $\Gamma(k+1) = k\Gamma(k)$가 유도된다.

이제 k가 음수일 때 $\Gamma(k)$ 를 정의하기 위하여

(2-12) $$\Gamma(k) = \frac{\Gamma(k+1)}{k} = \frac{\Gamma(k+2)}{k(k+1)} = \cdots$$

라고 두자.

(2-12)식에서 $\Gamma(-k)$는 불연속이고 $k=0, -1, -2, \cdots$ 일 때는 의미가 없다.

그러나 $-1 < k < 0$인 임의의 값을 (2-12)의 좌변에 대입하면 $\Gamma(-k)$에 대한 값이 우변 식으로 표시된다.

이 방법으로 점화식을 사용함으로서 $-1<k<0$일 때

$\Gamma\left(-\frac{1}{2}\right)=\Gamma\left(\frac{1}{2}\right)/-\frac{1}{2}$, $\Gamma(-0.9)=\Gamma(0.1)/-0.9$ 등으로 $\Gamma(k)$값은

$-2<k<1$, $-3<k<-2, \cdots$등 모든 k값에 대하여 정의되며 결국 0 및 음의 정수가 아닌 모든 k에 대한 $\Gamma(k)$값을 계산할 수 있다.

즉 k가 음의 정수가 아닐 때 $\Gamma(k)$는 (2-11)의 값 및 점화식을 이용한 (2-12)에 의해 아래 그림과 같이 표시되고 또 $\Gamma(k+1)=k!$에 의해

$\frac{1}{2}\,!=\Gamma\left(\frac{3}{2}\right)$, $\left(-\frac{1}{2}\right)!=\Gamma\left(\frac{1}{2}\right), \cdots$ 등 일반적인 계승의 값을 구할 수 있다.

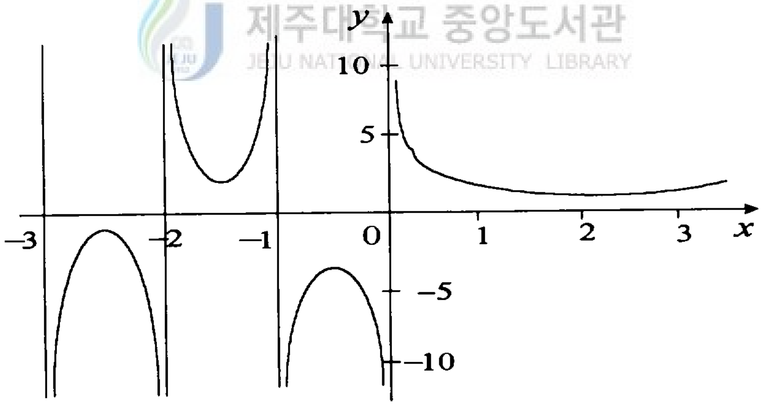

그림1 $y=\Gamma(x)$

예제 1.

$\left(-\frac{1}{2}\right)!$을 계산하여 보자.

풀이 : $\frac{1}{2}!=\Gamma\left(\frac{3}{2}\right)=\int_0^{\infty} e^{-x}x^{\frac{1}{2}}dx$이므로

$x=y^2$으로 두면

$$\frac{1}{2}!=2\int_0^{\infty} e^{-y^2}y^2dy \text{ 즉, } \frac{1}{2}!=2\int_0^{\infty} e^{-z^2}z^2dz$$

따라서

$$\left(\frac{1}{2}!\right)^2=4\int_0^{\infty} e^{-z^2}z^2dz\int_0^{\infty} e^{-y^2}y^2dy$$

$$=4\int_0^{\infty}\int_0^{\infty} e^{-(y^2+z^2)}y^2z^2dydz$$

이제 $z=r\cos\theta$, $y=r\sin\theta$로 두면

$$\left(\frac{1}{2}!\right)^2=4\int_0^{\infty}dr\int_0^{\frac{\pi}{2}} r^5e^{-r^2}\sin^2\theta\cos^2\theta d\theta$$

그런데 $\int_0^{\frac{\pi}{2}}\sin^2\theta\cos^2\theta\ d\theta=\frac{\pi}{16}$, 부분적분법에 의해 $\int_0^{\infty} e^{-r^2}r^5dr=1$

따라서 $\left(\frac{1}{2}!\right)^2=\frac{\pi}{4}$ 즉, $\frac{1}{2}!=\frac{\sqrt{\pi}}{2}$이고 (2-12)의 점화식에 의해

$\left(-\frac{1}{2}\right)!=\sqrt{\pi}$ 가 된다.

(2-11) 식에서 $t=\frac{x}{\theta}$ (θ>0인 상수)로 두면

(2-13) $$\Gamma(k)=\int_0^{\infty}\left(\frac{x}{\theta}\right)^{k-1}e^{-\frac{x}{\theta}}\cdot\frac{1}{\theta}dx$$

가 되고 양변을 $\Gamma(k)$로 나누면 확률밀도함수를 유도한다.

<u>정의 2 :</u>

확률변수 X의 확률밀도함수(pdf) $f(x)$가

(2-14) $$f(x;\theta,k)=\begin{cases}\dfrac{1}{\theta^k\Gamma(k)}x^{k-1}e^{-\frac{x}{\theta}} & ;\ x>0\\ 0 & ;\ x\le 0\end{cases}\qquad (\text{단 } k>0,\ \theta>0)$$

일 때 X는 Gamma 분포를 갖는다고 하고 $X\sim GAM(\theta,k)$로 표기한다.

모수 k는 $k<1,\ k=1,\ k>1$에 따라서 pdf의 모양이 결정된다. 그림2는 $k=0.5,\ 1,\ 2$ 에 관하여 그린 것이다.

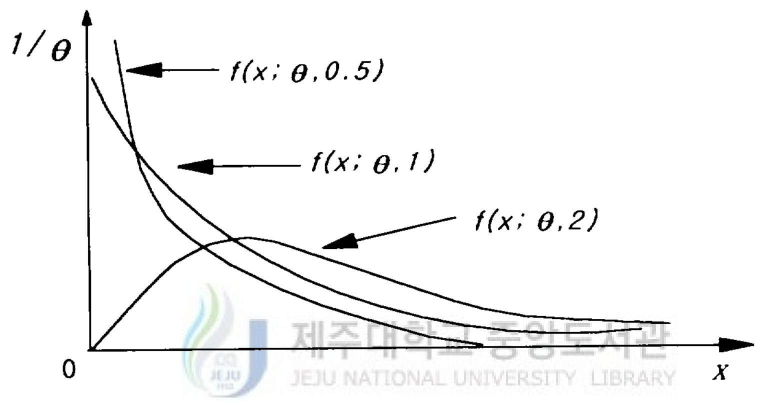

그림 2 감마분포 pdf

Gamma 분포는 일반적으로 대기시간에 대한 확률모형으로서 활용된다. 예를 들어 수명검사에 있어서 사망 시까지의 대기시간 또는 어떤 특정 물체의 고장 시까지의 시간 등은 Gamma분포를 갖는 확률변수이다.

Gamma분포는 또한 어떤 사건이 발생하게되는 특정 횟수를 얻기 위해서 요구되는 시간의 분포이다. 그 이유는 k가 양의 정수일 때 Gamma분포와 Poisson분포사이에 다음과 같은 중요한 정리가 주어지기 때문이다.

정리 1.

k가 양의 정수이고 $T \sim GAM(\theta, k)$ 즉 T가 사건 A의 k회 발생을 관찰하기 위하여 요구되는 시간이라고 하자. 한편 X를 Poisson과정의 공리를 만족하며 길이 t인 시간 구간 $(0, t]$동안에 발생하는 사건 A의 횟수라고 하자. T의 누적 분포함수

(2-15) $$G(t) = \int_0^t \frac{1}{\theta^k \Gamma(k)} s^{k-1} e^{-\frac{s}{\theta}} ds = 1 - \sum_{i=0}^{k-1} \frac{e^{-\frac{t}{\theta}} \cdot (\frac{t}{\theta})^i}{i!}$$

본 정리는 $G(t)$를 계속적으로 부분적분 함으로 증명할 수 있다.

예제 2.

X를 모수 m인 Poisson분포를 갖는다고 하자.

m이 $\theta = 1$, $k = 2$인 Gamma분포를 갖는 확률변수의 실험치일 때

$P(X=0,1,2)$를 계산하자.

풀이 : $f(x) = \frac{m^x e^{-m}}{x!}$; $x = 0, 1, 2, \cdots$

$$g(m) = \frac{1}{\Gamma(2) \cdot 1^2} m^{2-1} e^{-\frac{m}{1}}$$

$$= me^{-m}, \quad m > 0$$

$$\Rightarrow P(X=0,1,2) = \int_0^\infty g(m)\{f(0) + f(1) + f(2)\} dm$$

$$= \int_0^\infty (me^{-2m} + m^2 e^{-2m} + \frac{m^3}{2} e^{-2m}) dm = \frac{11}{16}$$

예제 3.

어떤 강 유역에서 측정 가능한 1일 강우량(단위:inch)은

$X \sim GAM(0, 2, 6)$이다. 강우량이 2인치를 초과할 확률은

$$P(X>2) = \int_2^{\infty} \frac{1}{(0.2)^6 \cdot \Gamma(6)} x^{6-1} e^{-\frac{x}{0.2}} dx$$

$$= 1 - F(2; 0.2, 6) = \sum_{i=0}^{5} \frac{10^i}{i!} e^{-10} = 0.067$$

2. χ^2분포와 주요성질

Gamma 분포는 대기시간에 대한 좋은 모형일 뿐만 아니라 연속형인 여러 가지 비의 확률변수에 대한 모형이다. 예를 들어 소득의 분포는 두 모수 k와 θ를 적절히 취함으로써 Gamma분포에 의해 모형화 할 수 있다.

정의 3:

Gamma 분포에서 $\theta=2$, $k=\frac{r}{2}$ (단 r은 양의 정수)가 되는 특수한 경우를 생각하자.

X의 pdf ; $f(x)=\begin{cases} \dfrac{1}{\Gamma\left(\frac{r}{2}\right) \cdot 2^{\frac{r}{2}}} x^{\frac{r}{2}-1} e^{-\frac{x}{2}} & ; x>0 \\ 0 & ; x \le 0 \end{cases}$

일 때 연속형 확률변수 X는 자유도 r인 χ^2분포를 갖는다고 하고 $\chi^2(r)$로 표기한다. 즉 $X \sim GAM\left(2, \frac{r}{2}\right)$이면 $X \sim \chi^2(r)$이다.

χ^2분포에 대한 주요한 성질들은 다음의 정리들로 대변된다. 일반적인 경우의 증명은 생략하였다.

정리 2 :

1) $X \sim \chi^2(r)$일 때 $E(X)=r$, $V(X)=2r$, $M_X(t)=(1-2t)^{-\frac{r}{2}}$; $t<\frac{1}{2}$

2) 서로 독립인 확률변수 X_i $(i=1,2,\cdots,n)$들이 각각 $X \sim \chi^2(k_i)$일 때

$\sum_{i=1}^{n} X_i \sim \chi^2(\sum_{i=1}^{n} k_i)$이다.

3) $Z \sim N(0,1)$이면 $Z \sim \chi^2(1)$이다.

4) $X \sim GAM\left(2, \frac{r}{2}\right)$라고 하자. $Y=\frac{2X}{\theta} \sim \chi^2(2r)$이다.

증명 : Y의 분포함수 : $G(y) = P_r(Y \leq y) = P_r\left(X \leq \frac{\theta y}{2}\right)$

$$= \int_0^{\frac{\theta y}{2}} \frac{1}{\Gamma\left(\frac{r}{2}\right) \cdot \theta^{\frac{r}{2}}} x^{\frac{r}{2}-1} \cdot e^{-\frac{x}{\theta}} dx$$

$\therefore$ Y의 pdf : $g(y) = \frac{1}{\Gamma\left(\frac{r}{2}\right) 2^{\frac{r}{2}}} y^{\frac{r}{2}-1} \cdot e^{-\frac{y}{2}}$; $y>0$

이고 $Y \sim \chi^2(2r)$이다.

5) $X_1, X_2, \cdots, X_n$이 $N(\mu, \sigma^2)$으로부터의 확률표본이고 분산의 추정량

$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$이라고 할 때

$i)$ $\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2} \sim \chi^2(n)$

$ii)$ $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$

가 성립한다. .

예제 4.

어떤 성분이 고장나기까지의 시간(年)은 $\theta=3$이고, $k=2$인 Gamma분포에 따른다. 그 성분의 90%가 고장나지 않고 작동할 보증기간은 ?

풀이 : $P(X \le x_{0.1}) = H(2x_{0.1}/\theta;2k) = 0.1$

(단, H는 χ^2의 누적분포함수)

따라서 $\dfrac{2 \cdot x_{0.1}}{\theta} = \chi^2_{0.1}(2k)$에서 $x_{0.1} = \dfrac{\theta \cdot \chi^2_{0.1}(2k)}{2}$

$\theta=3$, $k=2$이면

$$x_{0.1} = \frac{3\,\chi^2_{0.1}(4)}{2} = \frac{3 \times 1.06}{2} = 1.59\text{년}$$

3. F분포

F분포는 정규분포로부터 구한 독립인 두 표본의 분산비에 대한 분포를 설명하는데 중요한 역할을 하며 분산분석 등에서 필수적으로 활용된다. F분포는 다음의 변수변환 방법으로 그 밀도함수를 유도할 수 있다.

정의 4:

서로 독립인 χ^2확률변수 U와 V가 각각 자유도 n, m이라고 할 때

$X = \dfrac{U/n}{V/m}$은 자유도 (n, m)인 F분포를 따르고 X의 pdf는

(2-16) $$f(x) = \frac{\Gamma[(n+m)/2]}{\Gamma(n/2) \cdot \Gamma(m/2)} \left(\frac{n}{m}\right)^{\frac{n}{2}} \cdot \frac{x^{(n-2)/2}}{(1+nx/m)^{(n+m)/2}} \quad ; \quad x>0$$

이다.

위와 같은 pdf를 갖는 변수 X를 자유도(n, m)인 F확률변수라 하고 $X \sim F(n, m)$으로 표기한다. F확률변수는 모수인 자유도(n, m)에 의해 결정되며 그 순서에 주의를 요한다.

(2-16)를 유도하기 위하여 $X=\frac{U/n}{V/m}$, $Z=V$로 두자.

이 식은 집합 $A=\{(u, v): u>0,\ v>0\}$를 집합 $B=\{(x, z): x>0,\ z>0\}$로 사상하는 1대 1변환을 정의한다.

$u=\frac{n}{m}\cdot zx$, $v=z$이므로 이 변환의 Jacobian $J=\begin{vmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial z} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial z} \end{vmatrix}=\frac{n}{m}z$

즉 $|J|=\frac{n}{m}z$이다. 따라서 (X, Z)의 결합 pdf $g(x, z)$는 $(x, z)\in B$에서

$$g(x, z)=f(u, v)|J|$$
$$=\frac{1}{\Gamma(\frac{n}{2})\cdot\Gamma(\frac{m}{2})\cdot 2^{(n+m)/2}}(\frac{n}{m}zx)^{\frac{n}{2}-1}z^{\frac{m}{2}-1}\times\exp[-\frac{z}{2}(\frac{n}{m}x+1)]$$

(단, $f(u, v)$는 (U, V)의 결합 pdf)

여기서 적분변수 $y=\frac{1}{2}(\frac{n}{m}x+1)$로 치환하면 X의 주변 pdf인 (2-16)식이 유도된다.

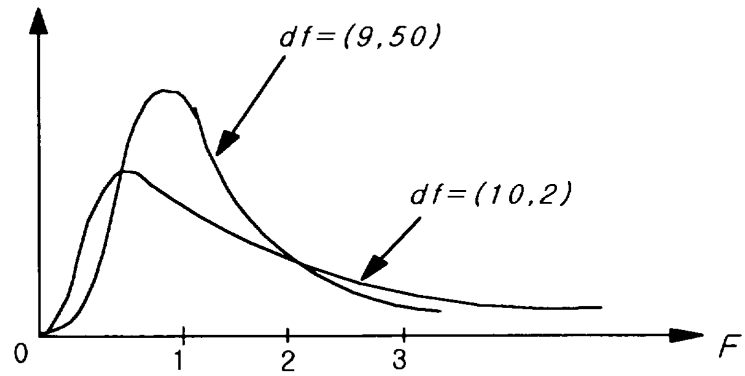

그림 3 F 분포

정리 4.

$X \sim F(n, m)$일 때 X의 평균과 분산은 아래와 같다.

(2-17) $$E(X) = \frac{m}{m-2} \qquad m>2$$

(2-18) $$Var(X) = \frac{2m^2(m+n-2)}{n(m-2)^2(m-4)} \qquad m>4$$

증명 : 확률변수 $X \sim F(n, m)$이면, 서로 독립인 확률변수 $U \sim \chi^2(n)$와

$V \sim \chi^2(m)$에 대하여 그들의 함수꼴인

$$X = \frac{U/n}{V/m}$$ 에서

U와 V가 서로 독립이므로 (따라서 U와 $1/V$도 서로 독립이며)

$$E(X) = E\left(\frac{U/n}{V/m}\right) = \frac{m}{n} E(U)\, E\left(\frac{1}{V}\right) = \frac{m}{n} \cdot n \cdot \frac{1}{m-2} = \frac{m}{m-2}$$

이 성립한다. 그런데,

$$E(U) = n$$

이며,

$$E\left(\frac{1}{V}\right) = \frac{1}{\Gamma(m/2)}\left(\frac{1}{2}\right)^{m/2} \int_0^\infty \frac{1}{v} v^{(m-2)/2} e^{-v/2} dv$$

$$= \frac{1}{\Gamma(m/2)}\left(\frac{1}{2}\right)^{m/2} \int_0^\infty v^{(m-4)/2} e^{-v/2} dv$$

$$= \frac{\Gamma[(m-2)/2]}{\Gamma(m/2)}\left(\frac{1}{2}\right)^{m/2}\left(\frac{1}{2}\right)^{-(m-2)/2} = \frac{1}{m-2}$$

이므로

$$E(X) = \left(\frac{m}{n}\right) E(U)\, E\left(\frac{1}{V}\right) = \frac{m}{n} \cdot \frac{n}{m-2} = \frac{m}{m-2}$$

가 성립한다.

한편 $T = \frac{Z}{\sqrt{V/n}}$, 단 $Z \sim N(0, 1)$이라고 하자. 즉 T가 자유도 r인 t분포

를 할 때 $T^2 = \dfrac{Z^2}{V/r} = \dfrac{\chi^2(1)/1}{\chi^2(r)/r} \sim F(1, r)$인 관계를 알 수 있다.

그런데, X가 자유도가 (n, m)인 F분포를 따르면, $1/X$는 자유도가 (m, n)인 F분포를 따름이 정의에서 확인된다. 그러므로 $F_\alpha(n, m)$를 $F(n, m)$분포의 α-백분위수라 하면

$$\alpha = P[X \le F_\alpha(n, m)] = P\left[\frac{1}{X} \ge \frac{1}{F_\alpha(n, m)}\right]$$
$$= 1 - P\left[\frac{1}{X} \le \frac{1}{F_\alpha(n, m)}\right]$$

로부터

$$\frac{1}{F_\alpha(n, m)} = F_{1-\alpha}(m, n)$$

의 관계를 구할 수 있다. 이러한 점은 F분포를 사용하는데 분포표의 윗부분 또는 아랫부분만 있어도 된다는 실질적인 도움을 준다.

<u>예제 5.</u>

확률변수 $X \sim F(5, 10)$에 대하여, $P(X \le a) = 0.01$을 만족하는 상수 a는

$$a = F_{0.01}(5, 10) = \frac{1}{F_{0.99}(10, 5)} = \frac{1}{10.2} = 0.098$$

이다. 나아가, $X \sim F(6, 12)$이면 $1/X \sim F(12, 6)$이므로,

$$P(X \ge 0.25) = P\left[\frac{1}{X} \le \frac{1}{0.25}\right] = P\left[\frac{1}{X} \le 4.00\right] = 0.95$$

으로 계산될 수 있다.

다음의 정리는 두 개의 정규분포로부터 얻은 독립 랜덤표본의 분산비에 대한 분포를 제공하며 통계량 $(S_1/S_2)^2$이 σ_1^2/σ_2^2에 대한 신뢰구간의 추정과 두 모집단 분산의 동일성에 대한 가설검정의 근거가 된다.

정리 5.

평균은 μ_X이고 분산이 $\sigma_X{}^2$인 정규분포로부터 크기가 n인 랜덤표본 $X_1, X_2, \cdots, X_n$ 그리고 평균이 μ_y이고 분산이 $\sigma_Y{}^2$인 정규분포로부터 크기가 m인 랜덤표본 $Y_1, Y_2, \cdots, Y_m$ 을 얻었다. 이제, 두 표본이 서로 독립이라고 하면 확률변수

$$F = \frac{S_X{}^2 / \sigma_X{}^2}{S_Y{}^2 / \sigma_Y{}^2}$$

은 자유도가 $(n-1, m-1)$인 F분포를 따른다.

증명 : 두 확률변수

$$\frac{(n-1)\, S_X{}^2}{\sigma_X{}^2} \sim \chi^2(n-1) \text{과} \quad \frac{(m-1)\, S_Y{}^2}{\sigma_Y{}^2} \sim \chi^2(m-1)$$

은 서로 독립이므로 F확률분포의 정의에 의하여 자명하다.

특히 $\sigma^2_X = \sigma^2_Y$인 경우, 표본분산비

$$\frac{S_{X^2}}{S_{Y^2}} = \frac{\sum_{i=0}^{n} (X_i - \overline{X_n})^2 / (n-1)}{\sum_{i=0}^{m} (Y_i - \overline{Y_m})^2 / (m-1)}$$

는 자유도가 $(n-1,\ m-1)$인 F분포를 따름을 쉽게 알 수 있다.

예제 6.

$X \sim N(\mu_1,\ 5.38)$에서 추출된 크기 10인 표본과 $Y \sim N(\mu_2,\ 2)$에서 추출된 크기 5인 표본이 있다.

X와 Y가 서로 독립일 때 Y의 표본분산 S_2^2이 X의 표본분산 S_1^2을 초과할 확률을 구하라.

풀이 : $P(S_2^2 > S_1^2) = P(\frac{S_2^2}{S_1^2} > 1) = P(\frac{S_2^2/2}{S_1^2/5.38} > 2.69)$

$= P(F(4,9) > 2.69) = 1 - 0.9 = 0.1$

III. F분포의 응용

1. 모분산 비의 신뢰구간

두 개의 독립인 랜덤표본 $X_1, X_2, \cdots, X_n$과 $Y_1, Y_2, \cdots, Y_m$의 모분포가 각각 $N(\mu_X, \sigma_X^2)$과 $N(\mu_Y, \sigma_Y^2)$이라고 하자. 두 모분산의 비 σ_X^2 / σ_Y^2에 대한 신뢰구간은 다음과 같다.

1) 모평균 μ_X와 μ_Y가 알려진 경우 :

서로 독립인 확률변량 $\sum_{i=0}^{n}(X_i - \mu_X)^2 / \sigma_X^2$과 $\sum_{i=0}^{m}(Y_i - \mu_Y)^2 / \sigma_Y^2$은 각각 자유도 n과 m인 χ^2분포를 따르므로

$$\frac{\sum(X_i - \mu_X)^2/(n\ \sigma_X^2)}{\sum(Y_i - \mu_Y)^2/(m\ \sigma_Y^2)} \sim F(n, m)$$

한편 F분포가 대칭은 아니나 양쪽꼬리부분에서 $\alpha/2$의 확률을 고려하면

$$P[F_{\frac{\alpha}{2}}(n, m) \le \frac{S_X^2 / \sigma_X^2}{S_Y^2 / \sigma_Y^2} \le F_{1-\frac{\alpha}{2}}(n, m)] = 1 - \alpha$$

여기서 $S_X^2 = \sum(X_i - \mu_X)^2/n$, $S_Y^2 = \sum(Y_i - \mu_Y)^2/m$이다. 따라서 모분산의 비 σ_X^2 / σ_Y^2에 대한 $100(1-\alpha)\%$ 신뢰구간은

$$\left[(\frac{S_X}{S_Y})^2 \cdot F_{\alpha/2}(n, m),\ (\frac{S_X}{S_Y})^2 \cdot F_{1-\alpha/2}(n, m) \right] \tag{3-1}$$

이다.

2) 모평균 μ_X와 μ_Y을 모르는 경우 :

모평균 μ_X와 μ_Y를 각각 표본평균 $\overline{X}_n$과 $\overline{Y}_m$으로 추정한 확률변수 $\sum_{i=1}^{n}(X_i-\overline{X_n})^2/\sigma_X{}^2$과 $\sum_{i=1}^{m}(Y_i-\overline{Y_m})^2/\sigma_Y{}^2$는 서로 독립인 각각 자유도가 $(n-1)$과 $(m-1)$인 χ^2분포를 따르는 점에 근거하여 다음과 같이 구할 수 있다. 즉,

$$\frac{S_X{}^2/\sigma_X{}^2}{S_Y{}^2/\sigma_Y{}^2} \sim F(n-1,\ m-1)$$에 의하여

(단, $S_X{}^2=\sum_{i=1}^{n}(X_i-\overline{X_n})^2/(n-1)$, $S_Y{}^2=\sum_{i=1}^{m}(Y_i-\overline{Y_m})^2/(m-1)$),

$P[\ F_{\alpha/2}(n-1,\ m-1)\le \frac{S_X{}^2/\sigma_X{}^2}{S_Y{}^2/\sigma_Y{}^2}\le F_{1-\alpha/2}(n-1,\ m-1)]=1-\alpha$이므로,

모분산의 비 $\sigma_X{}^2/\sigma_Y{}^2$에 대한 신뢰도 $(1-\alpha)\cdot 100\%$ 신뢰구간은

(3-2) $$\left[(\frac{S_X{}^2}{S_Y{}^2})F_{\alpha/2}(n-1,\ m-1),\ (\frac{S_X{}^2}{S_Y{}^2})F_{1-\alpha/2}(n-1,\ m-1)\right]$$

이 된다.

2. 가설검정에서의 F분포의 활용

두 개 이상의 평균값의 균일성 또는 2 표본이상에서의 분산의 검정방법이 유용하게 이용될 수 있다. 이를테면 경제적, 사회적 문제와 관계되는 계량적 성질에 지역적인 격차가 있는지, 또는 여러 학급간에 학생들의 학력의 균분성 문제, 대량 생산품 제조업에 있어서 그 제품의 품질이 일정한 관리수준

으로 유지되고 있는가 등을 알기 위해서 널리 이용되는 것이 가설검정에서의 F분포의 역할이다.

1) 두 모분산 동일성의 검정

독립인 표본 $X_1, X_2, \cdots, X_n$과 $Y_1, Y_2, \cdots, Y_m$이 각각 정규분포 $N(\mu_X, \sigma_X^2)$과 $N(\mu_Y, \sigma_Y^2)$으로 부터 구한 랜덤표본이라 하자. 이때, 가설

$$H_0: \sigma_X^2 = \sigma_Y^2 \quad \text{대} \quad H_1: \sigma_X^2 \neq \sigma_Y^2$$

에 대한 유의수준 α인 검정법을 구해보자. 모평균 μ_X와 μ_Y의 값을 모르는 경우, 일반화 우도비 방법을 활용한다. 먼저, 결합확률밀도함수는

(3-3) $\quad f(x_1, \cdots, x_n, y_1, \cdots, y_m ; \mu_x, \mu_y, \sigma_X^2, \sigma_Y^2)$

$$= \left(\frac{1}{2\pi\sigma_X^2}\right)^{n/2} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i - \mu_X}{\sigma_X}\right)^2\right]\left(\frac{1}{2\pi\sigma_Y^2}\right)^{n/2} \exp\left[-\frac{1}{2}\sum_{j=1}^{m}\left(\frac{y_j - \mu_Y}{\sigma_Y}\right)^2\right]$$

이다. 그런데, 전체모수공간 Ω내에서 모수벡터 $(\mu_x, \mu_y, \sigma_X^2, \sigma_Y^2)$의 최대우도 추정량은

$$\hat{\sigma}_{X^2} = \frac{\sum_{i=1}^{n}(X_i - \hat{\mu}_X)^2}{n}, \quad \hat{\sigma}_{Y^2} = \frac{\sum_{j=1}^{m}(Y_j - \hat{\mu}_Y)^2}{m}$$

으로 주어진다. 한편, 귀무가설 하에서 모수공간 Ω_0에 속하는 모평균과 공통된 모분산의 최대우도 추정량은

$$\hat{\mu}_{X_0} = \overline{X_n}, \quad \hat{\mu}_{Y_0} = \overline{Y_m}$$

$$\hat{\sigma}_0^2 = \frac{1}{n+m}\left[\sum_{i=1}^{n}(X_i - \overline{X_n})^2 + \sum_{j=1}^{m}(Y_j - \overline{Y_m})^2\right]$$

으로 계산된다. 따라서, 일반화 우도비는

(3-4) $$\Lambda = \frac{f(X_1, X_2, \cdots, X_n, Y_1, Y_2, \cdots, Y_m ; \hat{\mu}_{X_0}, \hat{\mu}_{Y_0}, \hat{\sigma}_{0^2})}{f(X_1, X_2, \cdots, X_n, Y_1, Y_2, \cdots, Y_m ; \hat{\mu}_X, \hat{\mu}_Y, \hat{\sigma}_{X^2}\ \hat{\sigma}_{Y^2})}$$

$$= \frac{(S_X^2)^{n/2} \cdot (S_Y^2)^{m/2}}{[(n S_X^2 + m S_Y^2)/(n+m)]^{(n+m)/2}}$$

이 된다. 그런데,

$$F = \frac{\sum_{j=1}^{m}(Y_j - \overline{Y}_m)^2/(m-1)}{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2/(n-1)} = \frac{S_Y^2}{S_X^2}$$

라 하면, Λ가 너무 작을 때 귀무가설을 기각하는 것은 F가 너무 크거나 작은 경우 귀무가설을 기각하는 것과 동등하다. 이제, 귀무가설 하에서 검정통계량 F의 분포는 $F \sim F(m-1, n-1)$이므로 유의수준 α인 일반화 우도비 검정법의 기각영역은

$$F \le F_{\alpha/2}(m-1, n-1) \text{ 또는 } F \ge F_{1-\alpha/2}(m-1, n-1)$$

이 된다. 만일, 대립가설이 단측으로 H_1; $\sigma_X^2 > \sigma_Y^2$이라면 유의수준 α인 일반화 우도비 검정법의 기각영역은

$$F \le F_{\alpha}(m-1, n-1)$$

가 된다.

한편, 모평균 μ_X와 μ_Y가 알려져 있으면, 일반화 우도비에 근거한 검정통계량은

$$F = \frac{\sum_{j=1}^{m}(Y_j - \mu_Y)^2/m}{\sum_{i=1}^{n}(X_i - \mu_X)^2/n}$$

이 되며, H_0하에서 $F \sim F(m, n)$이므로, 위 가설에 대한 유의수준 α인 기각영역은

$$F \le F_{\alpha/2}(m, n) \text{ 또는 } F \ge F_{1-\alpha/2}(m, n)$$

이 된다.

예제 7.

두 가지 방법 A, B에 의해 생산되는 제품의 균등성을 조사하려고 한다. 각 방법에 대한 표본의 크기를 측정한 결과 그 표준편차는 다음과 같았다. 두 모분산의 비율 σ_1^2/σ_2^2에 대한 90% 신뢰구간을 구하라. 또 $\sigma_1 = \sigma_2$라는 가정은 타당한가? $\alpha = 10\%$ 로 검정하라.

	표본크기	표본 표준편차
A방법	$n_1 = 25$	$S_1 = 8$
B방법	$n_2 = 16$	$S_2 = 7$

풀이 : 부록의 F분포 표에 따라

1) $F_{0.95}(24, 15) = 2.29,\quad F_{0.95}(15, 24) = 2.11$

(3-2)식에 의해 $\frac{64}{49} \cdot \frac{1}{2.11} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{64}{49} \cdot 2.29$

$$\rightarrow 0.619 < \frac{\sigma_1^2}{\sigma_2^2} < 2.99$$

2) $H_0 : \sigma_1^2 = \sigma_2^2,\quad H_1 : \sigma_1^2 \neq \sigma_2^2$

$S_1^2 = 64,\quad S_2^2 = 49$ 이므로

$$\frac{1}{F_{0.9}(15, 24)} = 0.474 < F = \frac{64}{49} = 1.306 < F_{0.9}(24, 15) = 2.99$$

$\therefore$ H_0 즉 두 방법에서 분산이 같다는 가정을 받아 드린다.

2) 여러 개의 평균의 동일성에 관한 검정

미지인 평균이 $\mu_1, \mu_2, \cdots, \mu_b$이고 미지이나 공통인 분산이 σ^2인 정규분포를 갖는 b개의 서로 확률적으로 독립인 확률변수를 생각하자.

$X_{1j}, X_{2j}, \cdots, X_{aj}$는 평균이 μ_j이고 분산이 σ^2 $(j=1,2,\cdots,b)$인 정규분포에서부터 추출한 크기 a인 확률표본을 나타낸다 하자.

모든 가능한 대립가설 H_1에 대해 복합가설 H_0 : $\mu_1=\mu_2=\cdots=\mu_b=\mu$ 에 대한 검정은 우도비 검정을 사용하여 그 검정 방법을 구할 수 있으나 그 활용성을 고려하여 분산분석에서 다룰 것이다.

3. 분산분석에서의 F분포의 활용

1) 일원분류에서의 F분포

k개의 모집단에서 크기 n의 확률표본을 추출한다. k개의 모집단을 때때로 다른 처리에 따라 분류한다. k개의 모집단은 평균이 $\mu_1=\mu_2=\cdots=\mu_k$이고 공통분산 σ^2인 정규분포에 따른다고 가정한다.

다음 가설

$$H_0: \mu_1=\mu_2=\cdots=\mu_k$$

$$H_1: \mu_i (i=1,2,\cdots,k)\text{중 적어도 둘은 다르다.}$$

를 검정하기 위한 방법을 유도하고자 표1에서와 같이 x_{ij}를 i번째 모집단에서 j번째 관찰치라 하자. 이제 i번째 모집단에서 얻어진 관찰치를 생각해보자. 이들의 평균이 μ_i이고 분산이 σ^2인 정규분포에서 크기 n인 확률표본으로 가정하여 관찰치를

$$x_{ij}=\mu_i+\varepsilon_{ij}$$

표 1 k개의 확률분포

	모 집 단						
처 리	1	2	…	i	…	k	
	x_{11}	x_{21}	…	x_{i1}	…	x_{k1}	
	x_{12}	x_{22}	…	x_{i2}	…	x_{k2}	
	⋮	⋮		⋮		⋮	
	x_{1n}	x_{2n}	…	x_{in}	…	x_{kn}	
합 계	T_1	T_2	…	T_I	…	T_K	$T..$
평 균	$\bar{x}_1$	$\bar{x}_2$	…	$\bar{x}_i$	…	$\bar{x}_k$	$\bar{x}..$

로 표시할 수 있다. ε_{ij}는 대응하는 모평균으로 부터 i번째 표본의 j번째 관찰치의 편차에 해당되는 확률변수로서 평균이 0 이고 분산이 σ^2인 정규분포에 따른다고 가정한다.

이 실험에서 실험전체의 모평균을 μ라 하면

$$\mu = \frac{1}{k} \sum_{i=1}^{k} \mu_i$$

이고 $\mu_i = \mu + \alpha_i$라 놓으면

$$x_i = \mu + \alpha_i + \varepsilon_{ij}$$

이 되고 $\sum_{i=1}^{k} \alpha_i = \sum_{i=1}^{k} (\mu_i - \mu) = 0$ 이다. 단 α_i는 i번째 모집단의 효과(effect)이다.

따라서 앞의 가설을 다음과 같이 쓸 수 있다. 즉,

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$$

H_1 : α_i 중 적어도 하나는 0이 아니다.

위와 같이 주어진 귀무가설에 대한 검정은 공통분산 σ^2의 두 독립 추정치의 비교에 근거를 두고 있다. 이들 추정치는 자료의 전변이를 두 성분으로 분류하여 얻어진다.

크기 nk인 한 개의 표본으로 통합된 모든 관찰치의 분산은 다음 공식으로 주어진다.

$$S^2 = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n} (x_{ij} - \overline{x}_{..})^2}{nk-1}$$

이 때 $\sum_{i=1}^{k} \sum_{j=1}^{n} (x_i - \overline{x}_{..})^2$는 총 제곱합(total sum of squares)으로 자료의 전변이를 측정하며 처리제곱합과 잔차제곱합으로 분해된다.

즉 $$\sum_{i=1}^{k} \sum_{j=1}^{n} (x_{ij} - \overline{x}_{..})^2 = n \sum_{i=1}^{k} (\overline{x}_{i.} - \overline{x}_{..})^2 + \sum_{i=1}^{k} \sum_{j=1}^{n} (x_{ij} - \overline{x}_{i.})^2$$

총제곱합(SST) = 처리제곱합(SSA) + 잔차제곱합(SSE)

이 된다.

자유도 $k-1$에 근거를 둔 σ^2의 한 추정치는

$$S_1^2 = \frac{SSA}{k-1}$$

로 주어진다. 만일 H_0가 참이면 S_1^2은 σ^2의 불편추정치이다. 그러나 H_1이 참이면 SSA가 더 큰 수 값을 갖는 경향이 있고 S_1^2은 σ^2을 과대하게 추정한다.

자유도 $k(n-1)$에 근거한 σ^2의 독립추정치는

$$S_2^2 = \frac{SSE}{k(n-1)}$$

로 주어진다. 추정치 S_2^2은 귀무가설의 참 거짓에 관계없이 불편추정치이다.

H_0가 참일 때, 검정통계량

$$F = \frac{S_1^2}{S_2^2}$$

은 자유도 $k-1$과 $k(n-1)$인 F분포를 갖는 확률변수 F의 값이다. 따라서

H_0가 거짓일 때 $S_1^2 > \sigma^2$되므로 기각역은 분포의 오른쪽에 갖는 단측검정이다. 이 때 기각역은 유의수준 σ이면

$$F > F_{1-\alpha}[(k-1), k(n-1)]$$

이 되고 이상의 결과를 요약하여 나타낸 것이 표2의 분산분석표이다.

표 2 일원분류에 대한 분산분석표

요 인	제곱합	자유도	평균제곱	F
처 리	SSA	$k-1$	$S_1^2 = \dfrac{SSA}{k-1}$	$\dfrac{S_1^2}{S_2^2}$
잔 차	SSE	$k(n-1)$	$S_2^2 = \dfrac{SSE}{k(n-1)}$	
계	SST	$nk-1$		

예제 8.

표3의 자료는 다섯 종류의 두통 약을 25명의 환자에게 복용시켰을 때 고통이 경감되는 시간을 나타낸 것이다. 환자 25명을 임의로 다섯 개의 그룹으로 나누고 각각 다른 종류의 약으로 치료하였다.

표 3 두통약으로부터 경감되는 시간

	두 통 약					
	A	B	C	D	E	
	5	9	3	2	7	
	4	7	5	3	6	
	8	8	2	4	9	
	6	6	3	1	4	
	3	9	7	4	7	
합 계	26	39	20	14	33	132
평 균	5.2	7.8	4.0	2.8	6.6	5.28

분산분석을 하여 다섯 종류의 약에 의해 고통이 경감되는 평균시간이 동일

하다는 가설을 유의수준 0.05로 검정하여라.

풀이 :

H_0 : $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$, H_1 : 평균 중 적어도 두 개는 같지 않다.

유의수준 $\alpha = 0.05$ 에서 기각역 : $F_{0.95}(4.20) = 2.87$ 에서 $F > 2.87$ 이고

$$SST = 5^2 + 4^2 + \cdots + 7^2 - \frac{132^2}{25} = 834 - 696.960 = 137.040$$

$$SSA = 26^2 + 39^2 + \cdots + \frac{33^2}{5} - \frac{132^2}{25} = 776.400 - 696.960 = 79.440$$

$$SSE = 137.040 - 79.440 = 57.600$$

분산분석표는 표4와 같다.

표 4 예제8의 분산분석표

요 인	제 곱 합	자 유 도	평균제곱	F
처 리	79.440	4	19.860	6.90
잔 차	57.600	20	2.880	
합 계	137.040	24		

따라서 $F = 6.90 > 2.87$ 이므로 H_0를 기각한다. 즉 두통약으로 부터 고통이 경감되는 평균시간은 5개 모두 동일하다고 할 수 없다.

2) 이원분류에서의 F분포

이원분류는 두 요인이 특성치에 영향을 주는지 결정할 때 사용하는 분산분석법이다.

표5와 같이 r행과 c열로 구성된 직사각형 배열을 생각하자. x_{ij}는 i번째 행과 j번째 열이 있는 관찰치를 표시한다. x_{ij}를 평균이 μ_{ij}이고 공통분산 σ^2인 정규분포에 따르는 독립확률변수의 값이라고 가정하자.

표 5 두 요인 실험의 자료의 구조

행	열						합계	평균
	1	2	$\cdots$	j	$\cdots$	c		
1	x_{11}	x_{12}	$\cdots$	x_{1j}	$\cdots$	x_{1c}	$T_{1\cdot}$	$\overline{x}_{1\cdot}$
2	x_{21}	x_{22}	$\cdots$	x_{2j}	$\cdots$	x_{2c}	$T_{2\cdot}$	$\overline{x}_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$	$\vdots$
i	x_{i1}	x_{i2}	$\cdots$	x_{ij}	$\cdots$	x_{ic}	$T_{i\cdot}$	$\overline{x}_{i\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$	$\vdots$
r	x_{r1}	x_{r2}	$\cdots$	x_{rj}	$\cdots$	x_{rc}	$T_{r\cdot}$	$\overline{x}_{r\cdot}$
합계	$T_{\cdot 1}$	$T_{\cdot 2}$	$\cdots$	$T_{\cdot j}$	$\cdots$	$T_{\cdot c}$	$T_{\cdot\cdot}$	
평균	$\overline{x}_{\cdot 1}$	$\overline{x}_{\cdot 2}$	$\cdots$	$\overline{x}_{\cdot j}$	$\cdots$	$\overline{x}_{\cdot c}$		$\overline{x}_{\cdot\cdot}$

i번째 행에 대한 모평균의 평균은

$$\mu_{i\cdot} = \frac{\sum_{j=1}^{c} \mu_{ij}}{c}$$

로 j번째 열에 대한 모평균의 평균은

$$\mu_{\cdot j} = \frac{\sum_{i=1}^{r} \mu_{ij}}{r}$$

rc개의 모평균에 대한 평균은

$$\mu = \frac{\sum_{i=1}^{r} \sum_{j=1}^{c} \mu_{ij}}{rc}$$

로 정의된다. 관찰치의 변이의 요인이 행 사이에서의 차에 기인하는지를 검정하기 위하여 다음 검정을 생각한다.

$$H_0' : \ \mu_{1\cdot} = \mu_{2\cdot} = \cdots = \mu_{r\cdot} = \mu$$

$$H_1' : \ \mu_{i\cdot} \text{ 는 모두 같지 않다.}$$

또한 변이의 요인이 열 사이에서의 차에 기인하는 지를 검정하기 위하여 다음 검정을 생각한다.

$$H_0^{''} : \mu_{\cdot 1} = \mu_{\cdot 2} = \cdots = \mu_{\cdot c} = \mu$$

$H_1^{''}$: $\mu_{\cdot j}$ 는 모두 같지 않다.

각 관찰치를

$$x_{ij} = \mu_{ij} + \varepsilon_{ij} ,$$

$$\mu_{ij} = \mu + \alpha_i + \beta_j$$

라 쓰면 α_i는 i번째 행의 효과이고 β_j는 j번째 열의 효과를 나타낸다. 따라서 x_{ij}는 다음식으로 주어진다.

$$x_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

만일

$$\sum_{i=1}^{r} \alpha_i = 0 \text{ 이고 } \sum_{j=1}^{c} \beta_j = 0$$

이라 하면

$$\mu_{i\cdot} = \frac{\sum_{j=1}^{c} (\mu + \alpha_i + \beta_j)}{c} = \mu + \alpha_i$$

$$\mu_{\cdot j} = \frac{\sum_{i=1}^{r} (\mu + \alpha_i + \beta_j)}{r} = \mu + \beta_j$$

이다. r행의 평균 $\mu_{i\cdot}$가 같고 μ 라고 하는 귀무가설은

$$H_0^{'} : \alpha_1 = \alpha_2 = \cdots = \alpha_r = 0$$

$H_1^{'}$: α_i 중 적어도 하나는 0이 아니다.

라는 가설을 검정하는 것과 같다. 같은 방법으로 c열의 평균 $\mu_{\cdot j}$가 같다는

귀무가설은

$$H_0'' : \beta_1 = \beta_2 = \cdots = \beta_c = 0$$

$$H_1'' : \beta_j \text{ 중 적어도 하나는 0이 아니다.}$$

라는 가설을 검정하는 것과 같다. 이와 같이 주어진 귀무가설을 검정하기 위한 검정법을 유도하기 위하여 총제곱합을 α_i의 효과에 대한 제곱합과 β_j의 효과에 의한 제곱합 및 잔차제곱합으로 분해한다.

<u>정리 6.</u>

$$\sum_{i=1}^{r}\sum_{j=1}^{c}(x_{ij}-\bar{x}_{..})^2$$
$$= c\sum_{i=1}^{r}(\bar{x}_{i.}-\bar{x}_{..})^2 + r\sum_{j=1}^{c}(\bar{x}_{.j}-\bar{x}_{..})^2 + \sum_{i=1}^{r}\sum_{j=1}^{c}(x_{ij}-\bar{x}_{i.}-\bar{x}_{.j}+\bar{x}_{..})^2$$

증명 :

$$\sum_{i=1}^{r}\sum_{j=1}^{c}(x_{ij}-\bar{x}_{..})^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}[(\bar{x}_{i.}-\bar{x}_{..})+(\bar{x}_{.j}+\bar{x}_{..})+(x_{ij}-\bar{x}_{i.}-\bar{x}_{.j}+\bar{x}_{..})]^2$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{c}(\bar{x}_{i.}-\bar{x}_{..})^2 + \sum_{i=1}^{r}\sum_{j=1}^{c}(\bar{x}_{.j}-\bar{x}_{..})^2 + \sum_{i=1}^{r}\sum_{j=1}^{c}(x_{ij}-\bar{x}_{i.}-\bar{x}_{.j}+\bar{x}_{..})^2$$
$$+2\sum_{i=1}^{r}\sum_{j=1}^{c}(\bar{x}_{i.}-\bar{x}_{..})(\bar{x}_{.j}+\bar{x}_{..})$$
$$+2\sum_{i=1}^{r}\sum_{j=1}^{c}(\bar{x}_{j.}-\bar{x}_{..})(x_{ij}-\bar{x}_{i.}-\bar{x}_{.j}+\bar{x}_{..})$$
$$+2\sum_{i=1}^{r}\sum_{j=1}^{c}(\bar{x}_{.j}+\bar{x}_{..})(x_{ij}-\bar{x}_{i.}-\bar{x}_{.j}+\bar{x}_{..})$$

곱의 합은 모두 0이 되므로 다음과 같이 주어진다.

$$\sum_{i=1}^{r}\sum_{j=1}^{c}(x_{ij}-\bar{x}_{..})^2$$
$$= c\sum_{i=1}^{r}(\bar{x}_{i.}-\bar{x}_{..})^2 + r\sum_{j=1}^{c}(\bar{x}_{.j}-\bar{x}_{..})^2 + \sum_{i=1}^{r}\sum_{j=1}^{c}(x_{ij}-\bar{x}_{i.}-\bar{x}_{.j}+\bar{x}_{..})^2$$

즉, 총제곱합 = 행평균제곱합 + 열평균제곱합 + 잔차제곱합

$$SST = SSR + SSC + SSE$$

가 된다.

자유도 $r-1$에 근거한 σ^2의 한 추정치는

$$S_1^2 = \frac{SSR}{r-1}$$

로 주어진다. 만일 행의 효과 $\alpha_1 = \alpha_2 = \cdots = \alpha_r = 0$ 이면 S_1^2은 σ^2의 불편추정치이다. 그러나 행의 효과가 모두 0이 아니면 SSR는 팽창된 수값을 갖는 경향일 것이고 S_1^2은 σ^2을 과대하게 추정하게 된다.

자유도 $c-1$에 근거한 두 번째 σ^2의 추정치는

$$S_2^2 = \frac{SSC}{c-1}$$

로 주어진다. 추정치 S_2^2은 열의 효과 $\beta_1 = \beta_2 = \cdots = \beta_c = 0$ 일 때 σ^2의 불편추정치이다. 열의 효과가 모두 0이 아니면 SSC는 팽창되어지고 S_2^2은 σ^2을 과대하게 측정할 것이다.

자유도 $(r-1)(c-1)$에 근거한 σ^2의 세 번째 추정치는

$$S_3^2 = \frac{SSE}{(r-1)(c-1)}$$

로 주어지고 귀무가설의 참, 거짓에 관계없이 불편성이다.

행이 효과가 모두 0으로서 같다는 귀무가설을 검정하기 위하여 비

$$F_1 = \frac{S_1^2}{S_2^2}$$

을 계산한다. 이것은 귀무가설이 참일 때 자유도 $r-1$과 $(r-1)(c-1)$인 F 분포에 따르는 확률변수 F_1의 값이다. H_0가 거짓일 때 S_1^2은 σ^2을 과대하게

측정하므로 기각역은 유의수준 α에서 $F_1 > F_\alpha[(r-1),(r-1)(c-1)]$ 이다.

같은 방법으로 열의 결과가 모두 0으로서 같다는 귀무가설을 검정하기 위하여 비

$$F_2 = \frac{S_2^2}{S_3^2}$$

을 계산한다. 이것은 귀무가설이 참일 때 자유도 $c-1$과 $(r-1)(c-1)$인 F분포에 따르는 확률변수 F_2의 값이다. 이 경우에도 귀무가설은

$F_2 > F_\alpha[(c-1),(r-1)(c-1)]$일 때 유의수준 α에서 이원분류에 대한 분산분석표를 표6과 같이 얻을 수 있다.

표 6 이원분류에 대한 분산분석표

요 인	제곱합	자유도	평균제곱	F
행 처 리	SSR	$r-1$	$S_1^2 = \frac{SSR}{r-1}$	$F_1 = \frac{S_1^2}{S_3^2}$
열 처 리	SSC	$c-1$	$S_2^2 = \frac{SSC}{cr-1}$	$F_2 = \frac{S_2^2}{S_3^2}$
잔 차	SSE	$(r-1)(c-1)$	$S_3^2 = \frac{SSE}{(r-1)(c-1)}$	
합 계	SST	$rc-1$		

예제 9.

세 품종의 밀을 동질인 단위면적의 밭에 심고 네 종류의 비료를 주어 재배한 다음 수확량을 조사하니 아래 표7과 같았다. 이렇게 다른 종류의 비료를 사용했을 때 밀의 수확량에서 차이가 없다는 가설 H_0'을 유의수준 0.05로 검정하여라. 또한 세 품종에 대한 밀의 평균수확량에서 차이가 없다는 가설 H_0''을 검정하여라.

표 7 밀의 품종과 비료의 종류에 따른 밀의 수확량

비료 처리	밀의 종류			계
	v_1	v_2	v_3	
t_1	64	72	74	210
t_2	55	57	47	159
t_3	59	66	58	183
t_4	58	57	53	168
계	236	252	232	720

풀이 : 1. (1) H_0' : $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4$(행 효과는 0이다.)

(2) H_0'' : $\beta_1 = \beta_2 = \beta_3$(열 효과는 0이다.)

2. (1) H_1' : α_i 중 적어도 하나는 0이 아니다.

(2) H_1'' : β_i 중 적어도 하나는 0이 아니다.

3. 유의수준 $\alpha = 0.05$

4. 기각역 : $F_{0.95}(3,6) = 4.76$, $F_{0.95}(2,6) = 5.14$ 이므로

(1) $F_1 > 4.76$, (2) $F_2 > 5.14$ 이다.

5. 계산

$$SST = 64^2 + 55^2 + \cdots + 53^2 - \frac{720^2}{12} = 662$$

$$SSR = \frac{210^2 + 159^2 + 183^2 + 168^2}{3} - \frac{720^2}{12} = 498$$

$$SSC = \frac{236^2 + 252^2 + 232^2}{4} - \frac{720^2}{12} = 56$$

$$SSE = 662 - 498 - 56 = 108$$

따라서 분산분석표는 표8과 같다.

6. 결정 :

(1) $F_1 = 9.22 > 4.76$이므로 $H_0^{'}$을 기각한다. 다른 종류의 비료를 사용했을 때 밀의 평균수확량 차이가 있다.

(2) $F_2 = 1.56 < 5.14$이므로 $H_0^{''}$을 채택한다. 따라서, 세 품종에 대한 밀의 평균수확량의 차이는 없다.

표 8 분산분석표

요 인	제곱합	자유도	평균제곱	F
행(비료)	498	3	166	9.22
열(밀)	56	2	28	1.56
잔 차	108	6	18	
합 계	662	11		

4. 교육현장에서의 F분포 응용 사례

다음은 고등학교 수학교육에 있어서 수업모형의 개발을 위하여 실시되었던 연구 결과에서 활용되는 F분포의 응용 사례이다.

1. 기간 : 1997년 9월 1일 ~ 1998년 2월 28일 까지 1개 학기

2. 제주시내 A고등학교 자연반 2학년 4개 학급 각 50명씩 200명을 대상으로 함

3. 수업방법

   ㅇ. 1, 2, 3 반 - 현재 시행하는 수업방식 즉, 교사의 주도적 수업방식

   ㅇ. 4 반 - 7개 분단으로 분반하여 우수학생을 분단장으로 지명하고 교사는 기본예제만 설명하여 문제를 제시하면 분단장 주도하에 분단원이 공동으로 문제를 해결한 다음 각 분단에서 1명씩 나와서 문제를

직접 풀어 보는 수업방식 즉 교사중심에서 학생 중심 수업으로 변형된 수업방법을 시도함.

4. 평가 : 1학기말에 전 학급 대상 공통문제로 기말고사 치르고 학급별 반평균에 의해 수업방법을 비교토록 하였음.

이상의 목적을 위해서 통계 Package SAS에 의해 처리된 결과는 다음과 같았다.

```
DATA ex;
  do ban='1반','2반','3반','4반';
    INPUT aa @@; output;
  end;
  cards;
70 38 52 77
75 58 54 78
56 48 66 59
68 54 68 69
76 60 56 62
40 68 64 57
70 52 58 72
46 74 54 59
71 48 66 72
66 66 52 71
78 70 72 83
53 50 62 53
71 68 54 73
69 60 62 73
69 42 64 70
40 70 44 50
79 48 72 79
50 52 64 58
```

41	56	68	48
55	62	52	62
70	48	60	74
72	76	56	77
60	56	64	64
53	50	72	59
60	50	64	62
63	60	58	69
56	72	56	71
47	58	70	56
67	56	76	69
50	42	58	58
62	54	44	66
57	72	52	62
81	38	68	80
55	60	56	53
60	72	64	67
70	62	62	72
52	46	70	58
69	76	42	73
73	66	78	76
46	68	62	54
61	50	60	73
70	64	64	64
42	60	64	62
65	48	54	69
59	52	46	62
50	54	70	61
55	70	50	65
35	76	72	48
69	70	68	74
71	72	52	58

```
PROC GLM DATA=ex;
  TITLE '분산분석';
  class ban;
  model aa=ban;
  means ban;
RUN;

PROC UNIVARIATE DATA=ex NORMAL PLOT;
  TITLE '정규성 검정';
  VAR aa;
RUN;
```

General Linear Models Procedure

Dependent Variable: AA

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1253.380000	417.793333	4.24	0.0062
Error	196	19298.600000	98.462245		
Corrected Total	199	20551.980000			

R-Square	C.V.	Root MSE	AA
0.060986	16.13203	9.922814	61.5100000

General Linear Models Procedure

Level of BAN	N	--------------AA------------- Mean	SD
1반	50	60.8600000	11.5246143
2반	50	58.8400000	10.5450636
3반	50	60.7200000	8.5405157
4반	50	65.6200000	8.7688921

UNIVARIATE PROCEDURE

Variable=AA

Moments

N	200	Sum Wgts	200
Mean	61.51	Sum	12302
Std Dev	10.16249	Variance	103.2763
Skewness	-0.27657	Kurtosis	-0.60066
USS	777248	CSS	20551.98
CV	16.52169	Std Mean	0.718597
T:Mean=0	85.59737	Prob>|T|	0.0001
Sgn Rank	10050	Prob>|S|	0.0001
Num ^= 0	200		
W:Normal	0.964279	Prob<W	0.0019

Variable=AA

Quantiles(Def=5)

100% Max	83	99%	80.5
75% Q3	70	95%	76.5
50% Med	62	90%	73.5
25% Q1	54	10%	48
0% Min	35	5%	43
		1%	38
Range	48		
Q3-Q1	16		
Mode	62		

UNIVARIATE PROCEDURE

Variable=AA

Extremes

Lowest	Obs	Highest	Obs
35(	189)	79(	65)
38(	130)	79(	68)
38(	2)	80(	132)
40(	61)	81(	129)
40(	21)	83(	44)

위의 자료에 의해 모형의 적합성, 이를테면 관찰치들의 독립성, 등분산성, 그리고 정규성 등의 검정에 대해서는 자료중 Moments와 Quantiles에서 입증되고 있는 바 이에 대한 분석은 생략하고 주어진 자료에 대한 분산분석만을 분석한다. 즉 모집단 균등성의 검정으로서

H_0 : $\mu_1 = \mu_2 = \mu_3 = \mu_4$

H_1 : 최소한 하나의 모평균은 다르다

에 대한 검정통계량 $F = \frac{MST}{MSE}$ 가 자유도 (3, 196)인 F분포에 따른다. 즉 F검정통계량의 값이 매우 클 때 H_0는 기각된다.

자료의 분산분석표에서 $MST = 417.79$, $MSE = 98.462245$ 이며 F통계량의 값 $F = 4.24$ 이며 자유도 (3, 196)의 F분포와 비교한 P-값은 0.0062로서 유의수준 1% 에서도 귀무가설을 기각할 수 있다. 즉 1, 2, 3, 4 반의 성적 모두 같지는 않다는 결론이다.

한편 1, 2, 3 반에서 실시된 기존 수업방식과 4반에서 실시한 분단학습법을 비교한 대비(Contrast) 즉 모평균 들에 대한 선형 결합 C는 상수 $\{C_i\}$에 대해

$$C = \sum_{i=0}^{I} c_i y_i \quad \text{단} \quad \sum_{i=0}^{I} c_i = 0\,, \quad I = \sum n_i$$

로 정의된다.

대비 C의 추정통계량 $\hat{c} = \sum_{i=0}^{I} c_i y_i$이며 추정통계량의 분산

$$V(\hat{c}) = V(\sum c_i y_i) = \sum c_i^2 \cdot \frac{\sigma^2}{n_i} \text{ 이므로 } \hat{c} \sim N(C, \sum c_i^2 \frac{\sigma^2}{n_i}) \text{가 된다.}$$

한편 이 분산의 추정량 $V(\hat{c}) = MSE \sum \frac{c_i^2}{n_i}$ 으로 구할 수 있다. 한편 대비에 대한 제곱합 $SSC = \frac{(\hat{c})^2}{\sum \frac{c_i^2}{n_i}}$ 으로 정의되며 SSC의 자유도는 1이다. 따라서 $F = \frac{MST}{MSE}$ 에 의한 F검정이 가능하다.

위의 실예에 대하여 $C = \mu_4 - \frac{1}{3}(\mu_1 + \mu_2 + \mu_3)$ 이고 대비추정량과 평균제곱

$$\hat{c} = y_{4\cdot} - \frac{1}{3}(y_{1\cdot} + y_{[2\cdot} + y_{3\cdot}) = 65.62 - \frac{1}{3}(60.86 + 58.84 + 60.72) = 5.48$$

$$MSC = SSC = \frac{(\hat{c})^2}{\sum \frac{c_i^2}{n_i}} = (5.48)^2 / \frac{\frac{1}{9} \times 3 + 1}{50} = \frac{30.0304}{0.02666} = 1126.42$$

따라서 F통계량의 값 $F = \frac{MST}{MSE} = \frac{1126.42}{98.462} = 11.44$

한편 $F_{0.99}(1, 196) = 6.85 < F$이므로 유의수준 1% 로 에서 H_0을 기각할 수 있다.
즉 분단 학습법이 기존 수업방식보다 훨씬 효과적이라는 사실이 입증되었다.

Ⅳ. 結　　論

χ^2분포를 하는 두 모집단의 분산비(比)에 의해 유도되는 F분포는 모집단 散布의 상태를 극명하게 드러내 주는 의미를 가졌다고 할 수 있다. 특히 k개 모집단 동질성에 대한 가설검정이나 회귀분석에서의 기울기 및 y절편의 검정에서 활용되는 t분포를 분산분석기법을 활용하여 F분포에 의한 검정 방법을 제공함으로서 F분포는 그 응용력을 넓혀 왔다. 실험계획법의 대부분의 영역이 분산분석의 활용이며 따라서 F분포의 이해와 응용 없이 교육현장에서의 교육방법 개선의 연구가 이루어 질 수 없음을 실례를 통하여 그 의미를 강조하고자 하였다.

참 고 문 헌

[1] Bain, L.J and Engelhardt, M. Introduction to probability and Math. Statistics 2nd ed, 1992

[2] Craig,A and Hogg,R.V. Introduction to Mathematical Statistics 4th ed, 1978

[3] Sokolnikoff,I.S. Advanced Calculus : McGraw-Hill book Company, Inc. 1939

[4] 韓漢洙, 曺秉憻, 새로운 統計學, 1997. 淸文閣

[5] 朴聖炫 ; 回歸分析, 1984. 大英社

<Abstract>

A study of F-distribution and its Application*

Song, Jae Choong

Mathematics Education Major

Graduate School of Education, Cheju National University

Cheju, Korea

Supervised by professor Kim, Ik Chan

The aim of this paper lies in recognizing the emphasis of the role of F-distribution and being of help to the good use of it through the application of F-distribution in each field of modern applied statistics. It has been shown that F-distribution can be derived from Gamma distribution and χ^2-distribution, and from the method of variable transformation as well. Confidence interval of the ratio of population variance has appeared in various ways in the theory of estimation. On the other hand, the application of F-distribution has been shown in the testing statistical hypothesis connected with the measuring property and the administration of the producted goods. In particular, the application of F-distribution has been intended to be emphasized by computerizing the examples for the improvement of the teaching method in the educational scene of the applied instances of the analysis of variance and the design of experiment

* A thesis submitted to Committee of the Graduate of Education, Cheju National University in partial fulfillment of the requirement for the degree of Master of Education in August, 1998.

감사의 글

그 동안 모자람이 많은 저에게 이 한 편의 논문이 나오기까지 연구에 바쁘신 가운데도 많은 시간을 할애하여 친절하고 세심하게 지도해 주신 김익찬 교수님과 자세한 검토와 조언을 해 주신 양성호, 윤용식 교수님을 비롯한 수학교육과, 수학과 교수님들께 진심으로 고마운 말씀을 드립니다.

그리고 교육과정 진행의 어려움속에서도 대학원 교육과정을 마칠 수 있도록 배려해 주신 교장선생님을 비롯한 선생님들, 특히 2학년 담임선생님들께 감사드리며 함께 강의를 받으며 서로 의지하고 어려운 일에 협조를 아끼지 않은 동료 및 후배 원생들께도 감사드립니다.

끝으로 자식을 위해 항상 헌신적으로 보살펴 주시는 부모님, 장인·장모, 어려움속에서도 인내와 사랑으로 내조해 준 아내 縝과 항상 건강하고 밝게 자라는 용택, 민아, 용준 그리고 나를 아는 주위의 모든 분들과 이 조그마한 기쁨을 함께 나누고자 합니다.

1998 년 8 월

송 재 충 드림