9-1 이론적 확률분포(연속형) 3

제주권역
대학 e-러닝 지원센터

χ^2-분포

확률변수 $Z_1, Z_2, \cdots, Zn$이 서로 독립적으로 표준정규분포 $N(0,1)$을 따를 때,

$Z_1, Z_2, \cdots, Zn$의 제곱합 $\sum_{i=1}^{n} Z_i^2$은 자유도가 n인 χ^2분포를 따른다.

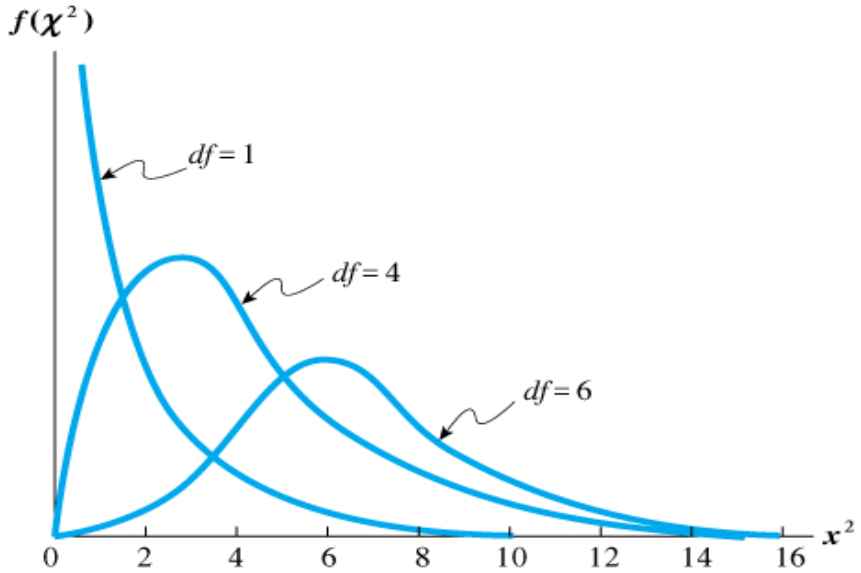
자유도가 n인 χ^2분포의 평균과 분산은 다음과 같다.

즉, $X \sim \chi^2_{(n)}$ 일 때

$$E(X) = n$$
$$Var(X) = 2n$$

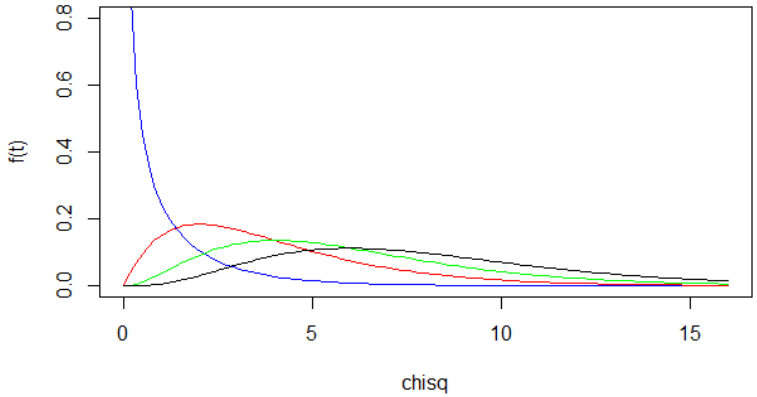
이다.

그림 χ^2 분포



χ^2 분포는 자유도가 모수이며, 따라서 자유도의 크기에 따라 분포의 형태가 달라진다. 그림은 몇 개의 자유도에 따르는 χ^2 분포의 그림이 주어져 있는데, 자유도가 클수록 정규분포와 근사한 χ^2 분포 형태를 갖는다.

```r
n_list<-c(2,5,7,9) # 표본수(n)
df_list<-n_list-1 # 자유도

curve(dchisq(x, 1, ncp=0), add=T, col="blue", xlim=c(0, 16), ylim=c(0, 0.8), xlab="chisq", ylab="f(t)")
curve(dchisq(x, 4, ncp=0), add=T, col="red", xlim=c(0, 16), ylim=c(0, 0.8), xlab="chisq", ylab="f(t)")
curve(dchisq(x, 6, ncp=0), add=T, col="green", xlim=c(0, 16), ylim=c(0, 0.8), xlab="chisq", ylab="f(t)")
curve(dchisq(x, 8, ncp=0), add=T, col="black", xlim=c(0, 16), ylim=c(0, 0.8), xlab="chisq", ylab="f(t)")

par(mfrow=c(2,2))

for (i in 1:length(df_list)) {

  curve(dchisq(x, df_list[i], ncp=0), xlim=c(0, 16), ylim=c(0, 0.8), xlab="chisq", ylab="f(t)", main=paste("df=", df_list[i]))

}
```

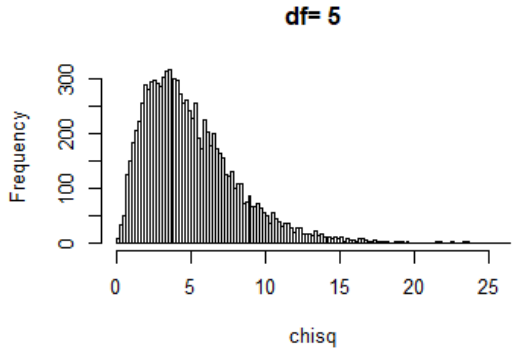


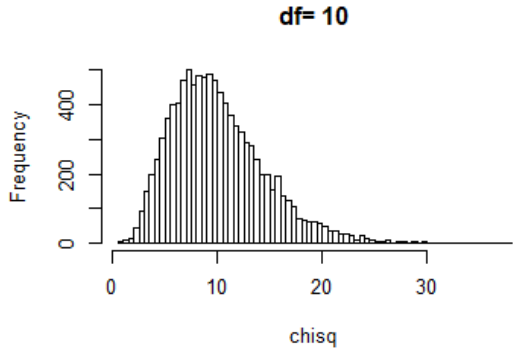
또한 분산이 σ^2 인 정규분포를 이루는 모집단으로부터 표본크기가 n인 선택 가능한 모든 임의표본이 추출되었을 때,
각 표본의 분산을 S^2 이라고 하면 다음은 자유도가 n-1인 χ^2 - 분포를 따른다.

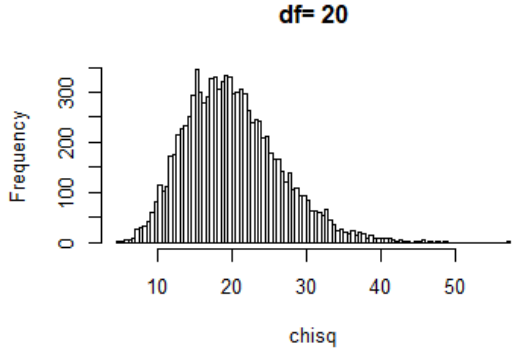
$$\chi^2_{(n-1)} = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

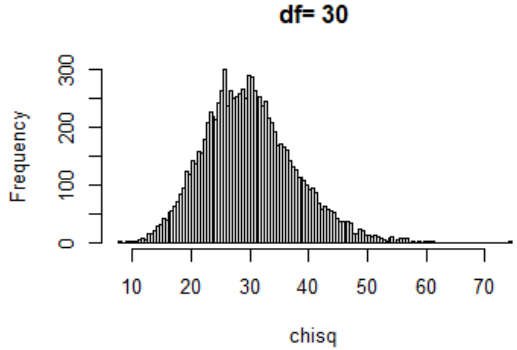
$$E(\chi^2_{n-1}) = \frac{(n-1)}{\sigma^2} E(s^2) = \frac{(n-1)}{\sigma^2}\sigma^2 = n-1$$

$$Var(\chi^2_{n-1}) = \frac{(n-1)^2}{\sigma^4} Var(s^2) = \frac{(n-1)^2}{\sigma^4}\frac{2\sigma^4}{n-1} = 2(n-1)$$


df= 5
Frequency
0
100
200
300
0
5
10
15
20
25
chisq


df= 10
Frequency
0
200
400
0
10
20
30
chisq


df= 20
Frequency
0
100
200
300
10
20
30
40
50
chisq


df= 30
Frequency
0
100
200
300
10
20
30
40
50
60
70
chisq

카이제곱은 자유도 n-1에 따라 분포의 형태가 다르므로 각각의 자유도 및 확률에 따른 카이제곱의 값이 다음과 같이 주어져 있다.

예 자유도가 5일 때 $\chi^2_{0.1} = 9.236$ 인데 이는 $\Pr(\chi^2 \geq 9.236) = 0.1$을 의미한다.

자유도	P=0.99	0.98	0.95	0.90	0.80	0.20	0.10	0.05	0.02	0.01
1	0.000157	0.000628	0.00393	0.0158	0.0642	1.642	2.706	3.841	5.412	6.635
2	0.0201	0.0404	0.103	0.211	0.446	3.219	4.605	5.991	7.824	9.210
3	0.115	0.185	0.352	0.584	1.005	4.642	6.251	7.815	9.837	11.341
4	0.297	0.429	0.711	1.064	1.649	5.989	7.779	9.488	11.668	13.277
5	0.554	0.752	1.145	1.610	2.343	7.289	**9.236**	11.070	13.388	15.086
6	0.872	1.134	1.635	2.204	3.070	8.558	10.645	12.592	15.033	16.812
7	1.239	1.564	2.167	2.833	3.822	9.803	12.017	14.067	16.622	18.475
8	1.646	2.032	2.733	3.490	4.594	11.030	13.362	15.507	18.168	20.090
9	2.088	2.532	3.325	4.168	5.380	12.242	14.684	16.919	19.679	21.666
10	2.558	3.059	3.940	4.865	6.179	13.442	15.987	18.307	21.161	23.209

```
df<-10

chi1<-numeric(df)
chi2<-numeric(df)
chi3<-numeric(df)
chi4<-numeric(df)
chi5<-numeric(df)
chi6<-numeric(df)
chi7<-numeric(df)
chi8<-numeric(df)
chi9<-numeric(df)
chi10<-numeric(df)

for(j in 1:df) {
  chi1[j]<-qchisq(0.01,j)
 }
for(j in 1:df) {
  chi2[j]<-qchisq(0.02,j)
}
for(j in 1:df) {
  chi3[j]<-qchisq(0.05,j)
}
for(j in 1:df) {
  chi4[j]<-qchisq(0.1,j)
}
for(j in 1:df) {
  chi5[j]<-qchisq(0.2,j)
}
for(j in 1:df) {
  chi6[j]<-qchisq(0.8,j)
}
for(j in 1:df) {
  chi7[j]<-qchisq(0.9,j)
}
for(j in 1:df) {
  chi8[j]<-qchisq(0.95,j)
}
for(j in 1:df) {
  chi9[j]<-qchisq(0.99,j)
}
for(j in 1:df) {
  chi10[j]<-qchisq(0.99,j)
}
(chi<-cbind(chi1,chi2,chi3,chi4,chi5,chi6,chi7,chi8,chi9,chi10))
```

```
> (chi<-cbind(chi1,chi2,chi3,chi4,chi5,chi6,chi7,chi8,chi9,chi10))
             chi1         chi2       chi3       chi4       chi5     chi6     chi7     chi8     chi9    chi10
 [1,] 0.0001570879 0.0006284502 0.00393214 0.01579077 0.06418475 1.642374 2.705543 3.841459 6.634897 6.634897
 [2,] 0.0201006717 0.0404054146 0.10258659 0.21072103 0.44628710 3.218876 4.605170 5.991465 9.210340 9.210340
 [3,] 0.1148318019 0.1848318190 0.35184632 0.58437437 1.00517401 4.641628 6.251389 7.814728 11.344867 11.344867
 [4,] 0.2971094805 0.4293981900 0.71072302 1.06362322 1.64877662 5.988617 7.779440 9.487729 13.276704 13.276704
 [5,] 0.5542980767 0.7518889348 1.14547623 1.61030799 2.34253431 7.289276 9.236357 11.070498 15.086272 15.086272
 [6,] 0.8720903302 1.1344192434 1.63538289 2.20413066 3.07008841 8.558060 10.644641 12.591587 16.811894 16.811894
 [7,] 1.2390423056 1.5642930043 2.16734991 2.83310692 3.82232191 9.803250 12.017037 14.067140 18.475307 18.475307
 [8,] 1.6464973727 2.0324769181 2.73263679 3.48953913 4.59357361 11.030091 13.361566 15.507313 20.090235 20.090235
 [9,] 2.0879007359 2.5323786659 3.32511284 4.16815901 5.38005321 12.242145 14.683657 16.918978 21.665994 21.665994
[10,] 2.5582121602 3.0590514109 3.94029914 4.86518205 6.17907926 13.441958 15.987179 18.307038 23.209251 23.209251
```

9-1-2. t-분포

t-분포

$Z \sim N(0,1), V \sim \chi^2(v)$ 이고 Z와 V가 독립이면, $T = \dfrac{Z}{\sqrt{\dfrac{V}{v}}} \sim t(v)$ 이다.

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1} \quad \Leftarrow \quad t = \frac{z}{\sqrt{\frac{V}{v}}} = \frac{\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}}{\frac{\sqrt{\frac{(n-1)s^2}{\sigma^2}}}{\sqrt{(n-1)}}} = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

자유도가 무한히 커지면 **t-**분포는 표준정규분포에 접근한다.

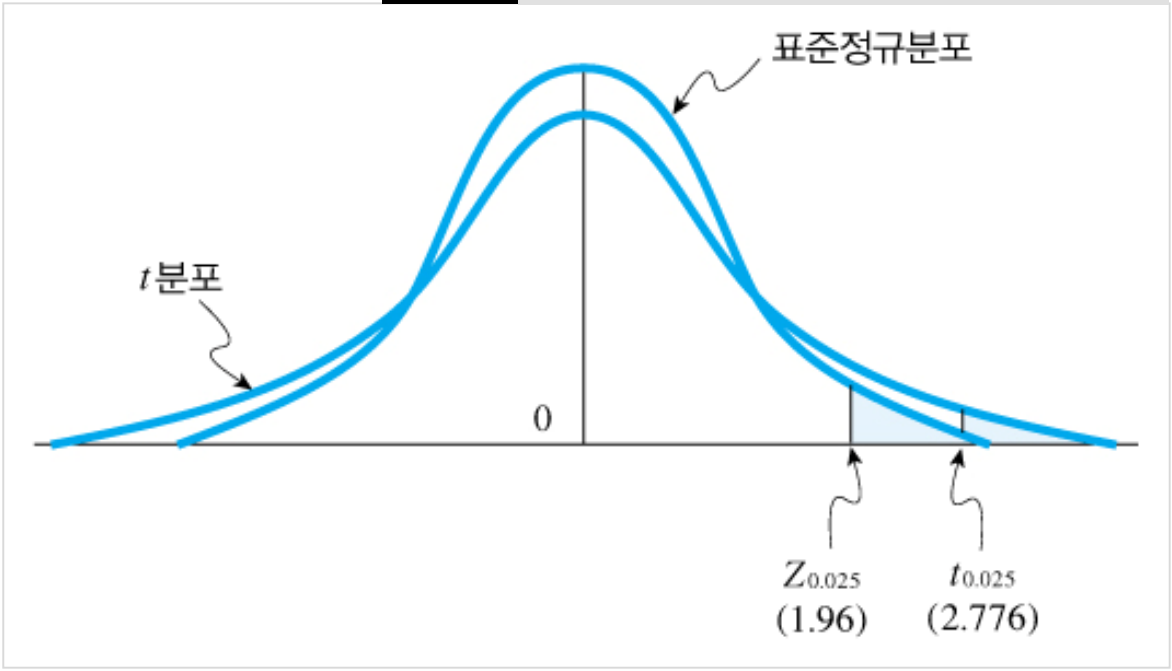
$(X_1, X_2, \dots, Xn)$을 $N(\mu, \sigma^2)$으로부터의 확률표본이라고 할 때 확률변수 T를

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

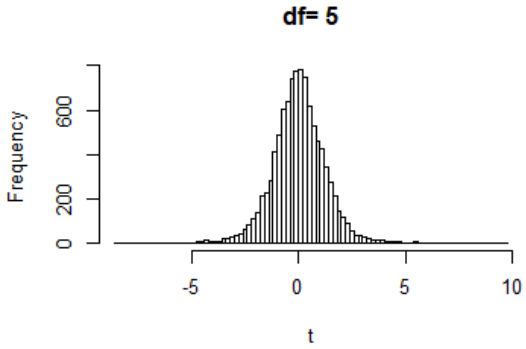
여기에서 $\bar{X} = \frac{1}{N}\sum_{i=1}^{n} X_i$, $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ 으로 정의하면,

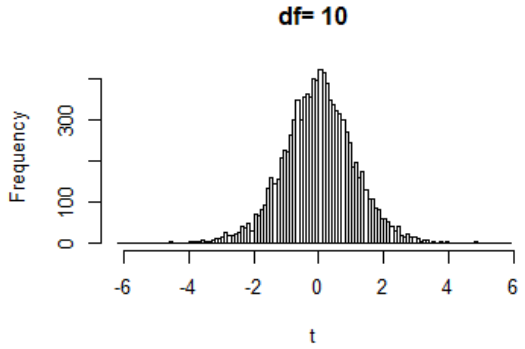
T는 자유도(d.f.)가 n-1인 t 분포를 따르며 $T \sim t(n-1)$로 표현한다.

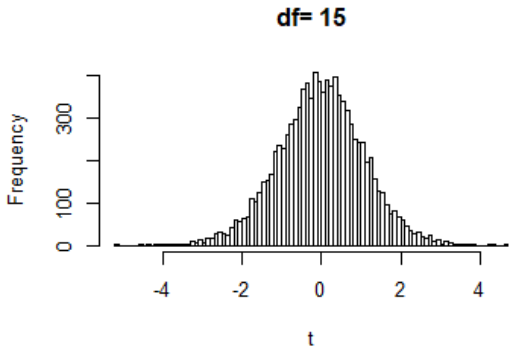
그림 표준정규분포와 $d.f. = 4$인 t분포의 비교

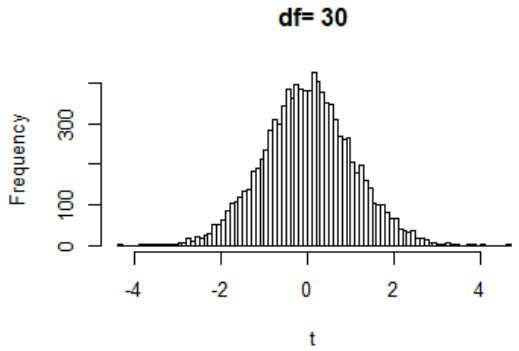


표준정규분포와 t분포를 비교한 그림으로 **t분포가 표준정규분포보다 꼬리부분의 확률이 조금 더 큰 것**을 알 수 있다.

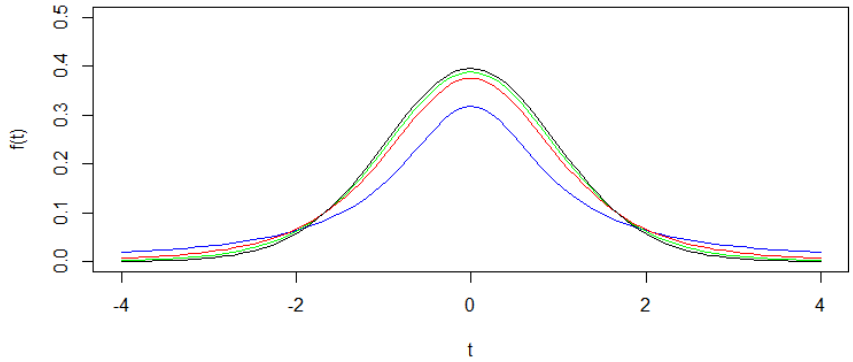

df= 5
Frequency
0
200
600
-5
0
5
10
t


df= 10
Frequency
0
100
300
-6
-4
-2
0
2
4
6
t


df= 15
Frequency
0
100
300
-4
-2
0
2
4
t


df= 30
Frequency
0
100
300
-4
-2
0
2
4
t

```r
n_list<-c(2,5,10,30) # 표본수(n)
df_list<-n_list-1 # 자유도

curve(dt(x, 1, ncp=0), add=T, col="blue", xlim=c(-4, 4), ylim=c(0, 0.5), xlab="t", ylab="f(t)")
curve(dt(x, 4, ncp=0), add=T, col="red", xlim=c(-4, 4), ylim=c(0, 0.5), xlab="t", ylab="f(t)")
curve(dt(x, 9, ncp=0), add=T, col="green", xlim=c(-4, 4), ylim=c(0, 0.5), xlab="t", ylab="f(t)")
curve(dt(x, 29, ncp=0), add=T, col="black", xlim=c(-4, 4), ylim=c(0, 0.5), xlab="t", ylab="f(t)")

par(mfrow=c(2,2))

for (i in 1:length(df_list)) {

  curve(dt(x, df_list[i], ncp=0), xlim=c(-4, 4), ylim=c(0, 0.5), xlab="t", ylab="f(t)",
        main=paste("df=", df_list[i]))

  }
```



t-분포는 자유도 n-1에 따라 분포의 형태가 다르므로 각각의 자유도 및 확률에 따른 t- 분포의 값이 다음과 같이 주어져 있다

예 자유도가 5일 때 $t_{0.05} = 2.015$ 인데 이는 $\Pr(T \geq 2.015) = 0.05$ 을 의미한다.

```
t11<-rep(NA,10)
t12<-rep(NA,10)
t13<-rep(NA,10)
t14<-rep(NA,10)
t15<-rep(NA,10)

for(i in 1:10) {
  t11[i]<-qt(0.9, i)
}
for(i in 1:10) {
  t12[i]<-qt(0.95, i)
}
for(i in 1:10) {
  t13[i]<-qt(0.975,i)
}
for(i in 1:10) {
  t14[i]<-qt(0.99,i)
}
for(i in 1:10) {
  t15[i]<-qt(0.995,i)
}

(tdist<-cbind(t11,t12,t13,t14,t15))
```

p / v	0.1	0.05	0.025	0.01	0.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.923
3	0.1638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	**2.015**	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
∞	1.282	1.645	1.960	2.326	2.576

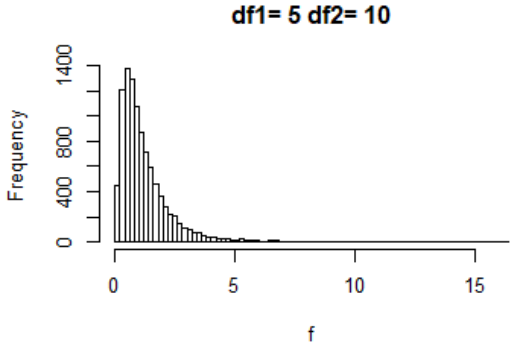
8-3-3. F-분포

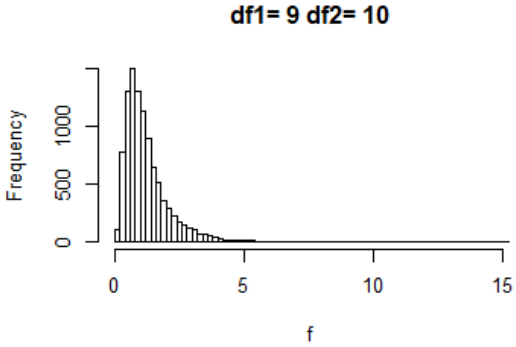
F-분포

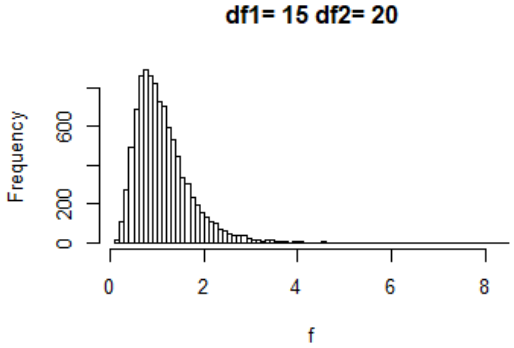
- $X_1 = \chi^2(v_1), X_2 = \chi^2(v_2)$이고 X_1, X_2이 서로 독립이며, $F = \dfrac{\frac{X_1}{v_1}}{\frac{X_2}{v_2}} \sim F(v_1, v_2)$

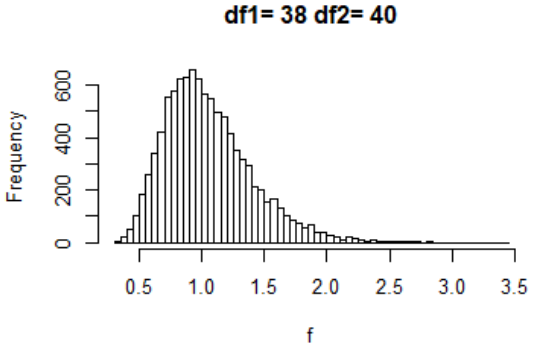
- $F = \dfrac{\frac{\chi^2_{n_1-1}}{n_1-1}}{\frac{\chi^2_{n_2-1}}{n_2-1}} = \dfrac{\frac{(n_1-1)s_1^2}{(n_1-1)\sigma_1^2}}{\frac{(n_2-1)s_1^2}{(n_2-1)\sigma_2^2}} = \dfrac{\frac{s_1^2}{\sigma_1^2}}{\frac{s_2^2}{\sigma_2^2}} \sim F_{(n_1-1,\, n_2-1)}$

E(F) = v2/(v2-2)이고, 자유도가 같아질수록 정규분포와 비슷하게 된다.

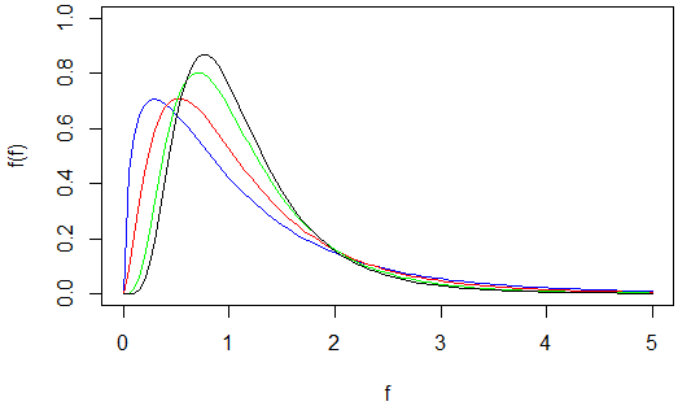

df1= 5 df2= 10
Frequency
0
400
800
1400
0
5
10
15
f


df1= 9 df2= 10
Frequency
0
500
1000
0
5
10
15
f


df1= 15 df2= 20
Frequency
0
200
600
0
2
4
6
8
f


df1= 38 df2= 40
Frequency
0
200
400
600
0.5
1.0
1.5
2.0
2.5
3.0
3.5
f

```r
curve(df(x, 3, 15, ncp=0), add=T, col="blue", xlim=c(0,5), ylim=c(0,1), xlab="f", ylab="f(f)")
curve(df(x, 5, 15, ncp=0), add=T, col="red", xlim=c(0,5), ylim=c(0,1), xlab="f", ylab="f(f)")
curve(df(x, 10, 15, ncp=0), add=T, col="green", xlim=c(0,5), ylim=c(0,1), xlab="f", ylab="f(f)")
curve(df(x, 15, 15, ncp=0), add=T, col="black", xlim=c(0,5), ylim=c(0,1), xlab="f", ylab="f(f)")

df1_list<-c(3,5,10,15)
df2_list<-c(15,15,15,15)
par(mfrow=c(2,2))


for (i in 1:length(df1_list)) {
  curve(df(x, df1=df1_list[i], df2=df2_list[i], ncp=0), xlim=c(0,5), ylim=c(0,1),
        xlab="f", ylab="f(f)", main=paste("df1=", df1_list[i], "df2=", df2_list[i]))
}

```



F-분포는 분자의 자유도(v1) 및 분모의 자유도(v2)에 따라 분포의 형태가 다르므로 각각의 자유도에 따른 $\alpha = 0.05$ 에 대한 F값이 다음과 같이 주어져 있다

예 분자의 자유도가 7이고 분모의 자유도가 9인 경우 $F_{0.05} = 3.29$ 이다.

v2 v1	1	2	3	4	5	6	7	8	9	10		∞
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88		254.31
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40		19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.76		8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96		5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.48	4.82	4.77	4.74		4.37
6	5.99	4.74	7.35	4.12	3.94	3.87	3.79	3.73	3.68	3.64		3.23
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64		3.23
8	5.32	4346	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35		2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	**3.29**	3.23	3.18	3.14		2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98		2.54
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.00

```r
f11<-rep(NA,10)
f12<-rep(NA,10)
f13<-rep(NA,10)
f14<-rep(NA,10)
f15<-rep(NA,10)
f16<-rep(NA,10)
f17<-rep(NA,10)

for(i in 1:10) {
  f11[i]<-qf(0.95, 1, i)
}
for(i in 1:10) {
  f12[i]<-qf(0.95, 2, i)
}
for(i in 1:10) {
  f13[i]<-qf(0.95, 3, i)
}
for(i in 1:10) {
  f14[i]<-qf(0.95, 4, i)
}
for(i in 1:10) {
  f15[i]<-qf(0.95, 5, i)
}
for(i in 1:10) {
  f16[i]<-qf(0.95, 6, i)
}
for(i in 1:10) {
  f17[i]<-qf(0.95, 7, i)
}

(fdist<-cbind(f11,f12,f13,f14,f15,f16,f17))
```

```
> (fdist<-cbind(f11,f12,f13,f14,f15,f16,f17))
          f11      f12      f13      f14      f15      f16      f17
 [1,] 161.447639 199.500000 215.707345 224.583241 230.161878 233.986000 236.768400
 [2,]  18.512821  19.000000  19.164292  19.246794  19.296410  19.329534  19.353218
 [3,]  10.127964   9.552094   9.276628   9.117182   9.013455   8.940645   8.886743
 [4,]   7.708647   6.944272   6.591382   6.388233   6.256057   6.163132   6.094211
 [5,]   6.607891   5.786135   5.409451   5.192168   5.050329   4.950288   4.875872
 [6,]   5.987378   5.143253   4.757063   4.533677   4.387374   4.283866   4.206658
 [7,]   5.591448   4.737414   4.346831   4.120312   3.971523   3.865969   3.787044
 [8,]   5.317655   4.458970   4.066181   3.837853   3.687499   3.580580   3.500464
 [9,]   5.117355   4.256495   3.862548   3.633089   3.481659   3.373754   3.292746
[10,]   4.964603   4.102821   3.708265   3.478050   3.325835   3.217175   3.135465
```

이론적 확률분포 (요약)

분포표기 및 모수	확률함수$(f(x))$	평균	분산
$X\sim B(1,p)$	$p^x q^{1-x}\ , x=0,1$	p	$pq, (p+q=1)$
$X\sim B(n,p)$	$\binom{n}{x} p^x q^{n-x}, x=0,1,\dots,n$	np	$npq, (p+q=1)$
$X\sim P(\mu)$	$\frac{e^{-\mu}\mu^x}{x!}, x=0,1,\dots..$	μ	μ
$X\sim U(a,b)$	$\frac{1}{b-a}, a<x<b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$X\sim N(\mu,\sigma^2)$	$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}, -\infty<x<\infty$	μ	σ^2
$X\sim \chi^2(v)$	복잡함	v	$2v$
$X\sim t(v)$	복잡함	0	$\frac{v}{v-2}$
$X\sim F(v_1,v_2)$	복잡함	$\frac{v_2}{v_2-2}$	$\frac{2v_2^2(v_1+v_2-2)}{v_1(v_2-2)^2(v_2-4)}$