◆ 正規標本（正規分布を母分布とする標本）についての特性

まずは，中心極限定理の復習， ...

ある分布（平均 $= \mu$, 分散 $= \sigma^2$）から抽出された標本（サイズ N ）の標本平均 $\overline{X}$ について，以下が成立する（大数の法則）.

1)
$$E\left(\overline{X}\right) = \mu\,; \quad E\left(\overline{X} - \mu\right)^2 = \frac{\sigma^2}{N}$$

さらに，中心極限定理から，標本平均 $\overline{X}$ は正規分布に従うため，推定値 $\overline{X}$ と真値 μ は，以下の関係をみたす.

2)
$$\mathrm{Prob}\left\{\left|\frac{\overline{X} - \mu}{\sqrt{\sigma^2/N}}\right| \le 1.96\right\} = 0.95$$

```
> qnorm(c(0.05/2, 1 - 0.05/2))
[1] -1.959964  1.959964
```

推定値 $\overline{X}$ と真値 μ の関係は，以下のようにも表現できる．

3) $$\mathrm{Prob}\left\{\overline{X}-1.96\sqrt{\frac{\sigma^2}{N}} \le \mu \le \overline{X}+1.96\sqrt{\frac{\sigma^2}{N}}\right\}=0.95$$

この式を読めば，“ 区間 $\left[\overline{X}-1.96\sqrt{\frac{\sigma^2}{N}},\ \overline{X}+1.96\sqrt{\frac{\sigma^2}{N}}\right]$ の中に

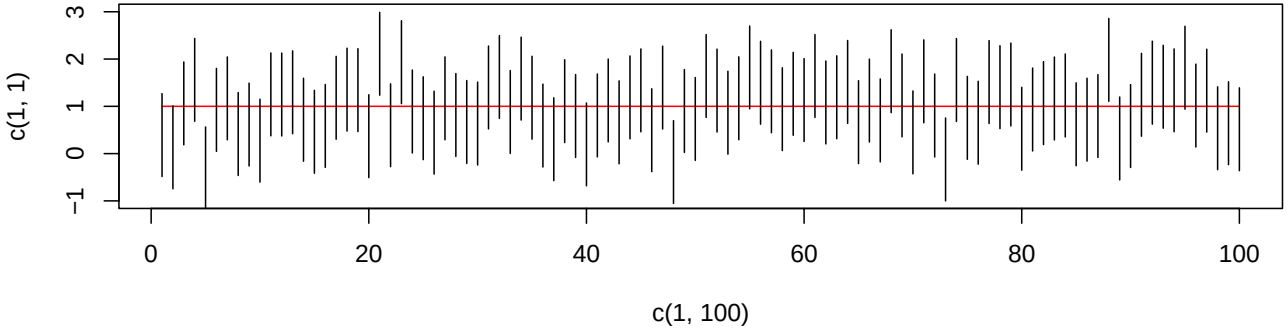
真値 μ が含まれる確率は，0.95 である ” となる．

この区間を**信頼区間**といい，真値 μ が含まれる確率を**信頼係数**とよぶ．

注意：ここでは，信頼区間に母分布の分散 σ^2 が含まれているので，
母分散 σ^2 が既知として与えられる状況であると，解釈できる．

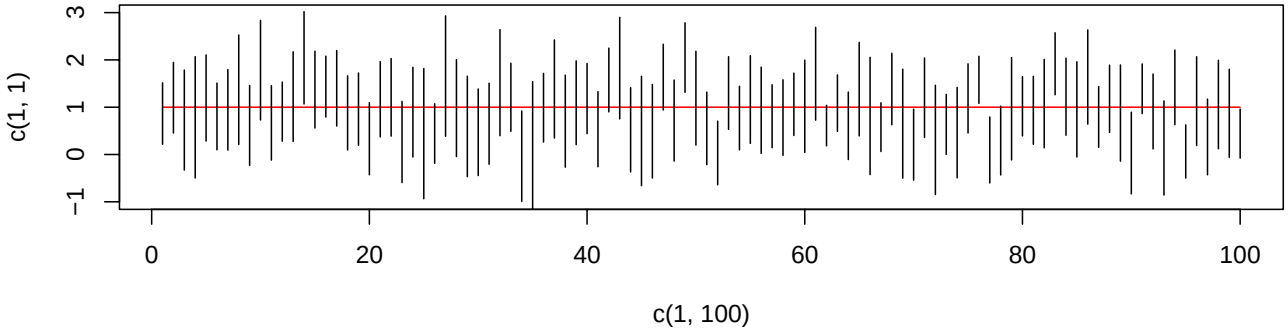
信頼区間および信頼係数について，シミュレーションしてみる．

```
> confidence <- numeric(100)
> plot(c(1,100), c(1,1), col="red", type="l", ylim=c(-1,3))
> for(try in 1:100) {
+        me <- mean(rnorm(10, mean=1, sd=sqrt(2)))
+    lines(c(try,try), me + c(-1,1)*1.96*sqrt(2/10))
+        confidence[try] <- (abs(me - 1)/sqrt(2/10) < 1.96)}
> sum(confidence)/100
[1] 0.94
```



母分布の分散を標本分散で置き換えるとどうなるか？

```
> plot(c(1,100), c(1,1), col="red", type="l", ylim=c(-1,3))
> for(try in 1:100) {
+  sa <- rnorm(10, mean=1, sd=sqrt(2)); me <- mean(sa); vr <- var(sa)
+    lines(c(try,try), me + c(-1,1)*1.96*sqrt(vr/10))
+        confidence[try] <- (abs(me - 1)/sqrt(vr/10) < 1.96)}
> sum(confidence)/100
[1] 0.89
>
```



単純に置き換えるとよくないことがわかる．どのようにすればよいか？

母分散を標本分散に取り替えると，2) は成立しないのである．

4) $\mathrm{Prob}\left\{ z(0.025) \le \dfrac{\overline{X} - \mu}{\sqrt{s^2/N}} \le z(0.975) \right\} \neq 0.95 \ ; \quad -z(0.025) = z(0.975) = 1.96$

この場合には，以下が成立する（ただし，条件：母分布は正規分布）．

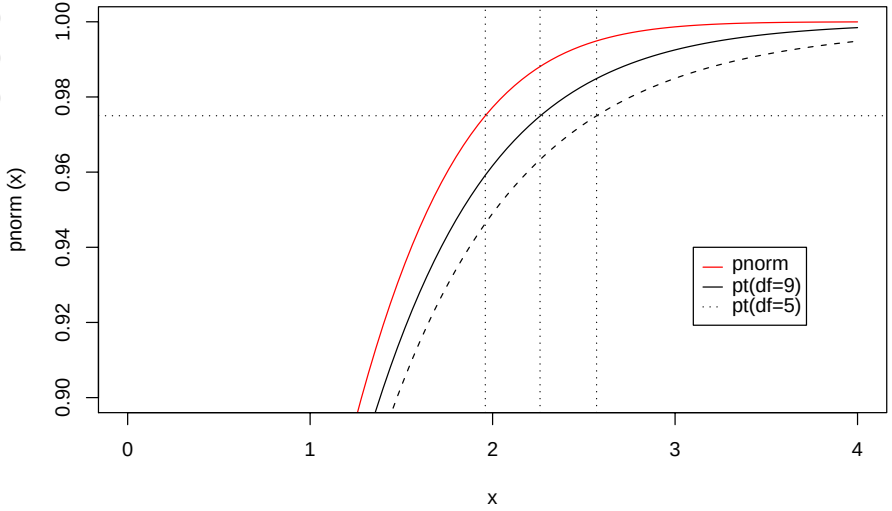
5) $\mathrm{Prob}\left\{ t(0.025, \mathrm{df}) \le \dfrac{\overline{X} - \mu}{\sqrt{s^2/N}} \le t(0.975, \mathrm{df}) \right\} = 0.95$

つまり，統計量 $t = \dfrac{\overline{X} - \mu}{\sqrt{s^2/N}}$ **は，母分布が正規分布であれば，t 分布に従う．**

```
> qt(c(0.05/2, 1 - 0.05/2), df=10 - 1)
[1] -2.262157  2.262157
> qt(c(0.05/2, 1 - 0.05/2), df=5)
[1] -2.570582  2.570582
```

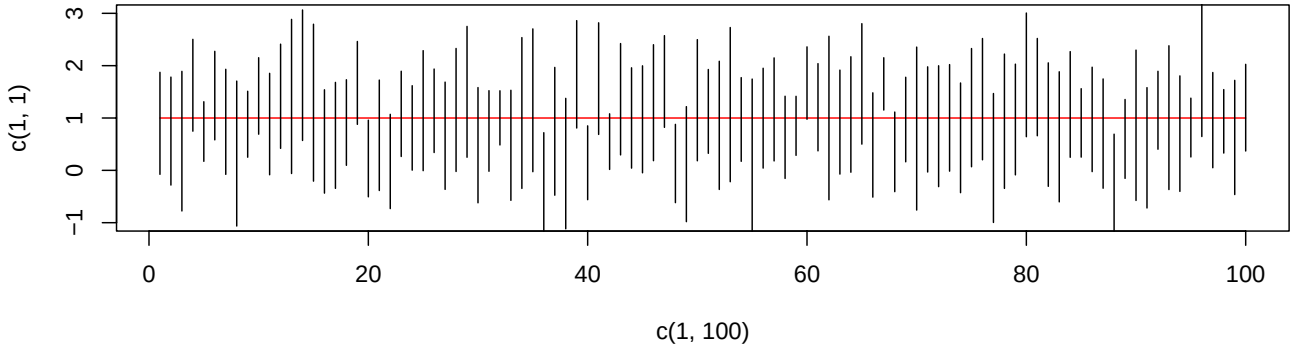
t 分布と正規分布の違いについて

```
> curve(pnorm, 0,4, col="red", ylim=c(0.9,1))
> pt9 <- function(x) pt(x, df=9)
> curve(pt9, 0,4, add=TRUE, lty=1)
> pt5 <- function(x) pt(x, df=5)
> curve(pt5, 0,4, add=TRUE, lty=2)
> legend(3.1,.94,
+   legend=c("pnorm","pt(df=9)","pt(df=5)"),
+   lty=c(1,1,3), col=c("red","black","black"))
> abline(h=0.975, lty=3)
> abline(v=1.96, lty=3)
> abline(v=2.26, lty=3)
> abline(v=2.57, lty=3)
>
```



平均の信頼区間と信頼係数のシミュレーション，再び

```
> plot(c(1,100), c(1,1), col="red", type="l", ylim=c(-1,3))
> for(try in 1:100) {
+  sa <- rnorm(10, mean=1, sd=sqrt(2)); me <- mean(sa); vr <- var(sa)
+    lines(c(try,try), me + c(-1,1)*2.26*sqrt(vr/10))          # != 1.96
+        confidence[try] <- (abs(me - 1)/sqrt(vr/10) < 2.26)}  # != 1.96
> sum(confidence)/100
[1] 0.94
>
```



ページ (4) **と比較すれば，この場合は，適切な結果が得られる．**

統計量 t は，以下のように分解して，t 分布に従うことが示される．
（導出の詳細は省略する．ここでは，導出のポイントのみを示す）

$$t = \frac{\overline{X} - \mu}{\sqrt{s^2/N}} = \frac{\overline{X} - \mu}{\sqrt{\sigma^2/N}} \Bigg/ \sqrt{\frac{s^2}{\sigma^2}} = \frac{\overline{X} - \mu}{\sqrt{\sigma^2/N}} \Bigg/ \sqrt{\frac{\chi^2(N-1)}{N-1}}$$

T1) $\dfrac{\overline{X} - \mu}{\sqrt{\sigma^2/N}}$ は，標準正規分布に従う（任意の分布に対して成立）．

T2) $(N-1)\dfrac{s^2}{\sigma^2}$ は，母分布が正規分布の時，カイ自乗分布に従う．

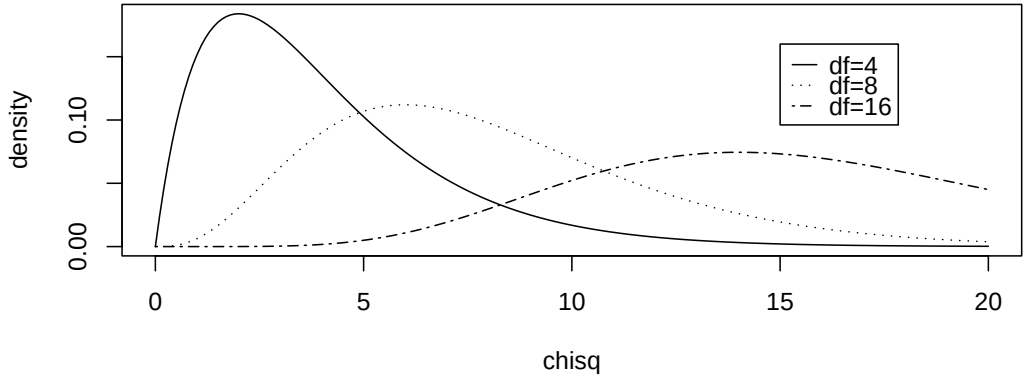
T3) $\dfrac{\overline{X} - \mu}{\sqrt{\sigma^2/N}}$ と $(N-1)\dfrac{s^2}{\sigma^2}$ は，母分布が正規分布の時，独立である．

つまり，標本平均 $\overline{X}$ と標本分散 s^2 は独立（母分布が正規分布の時）．

原則，母分布が正規分布であることが必要不可欠である．ただし，
実用的には，母分布が厳密に正規分布でなくても適用可能（t 分布の頑健性）．

t分布誘導の項目 T2) で現れるカイ自乗分布の特性について，

```
> xx <- seq(0, 20, by=.1)
> plot(xx,  dchisq(xx, df=4), type="l",ylab="chisq")
> lines(xx, dchisq(xx, df=8), lty=3)
> lines(xx, dchisq(xx, df=16), lty=6)
> legend(15, .16, legend=c("df=4","df=8","df=16"), lty=c(1,3,6))
```



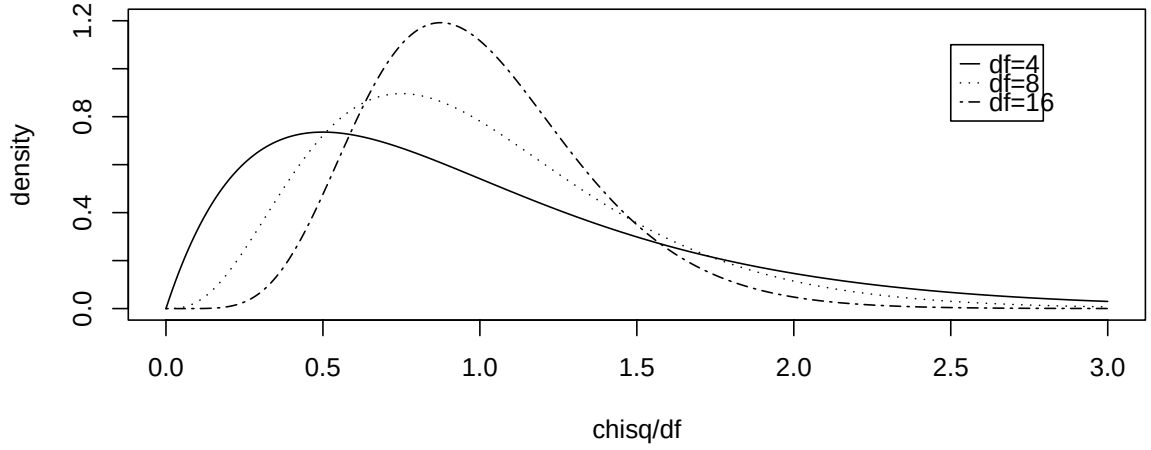
カイ自乗分布の最も重要な性質は，次の３つである．

C1）平均が自由度になること．

（したがって，カイ自乗を自由度で除したものの平均は，１となる）．

C2) カイ自乗を自由度で除した分布は，自由度の増加とともに，その平均に集中する（これも，大数の法則）こと．

```
> yy <- seq(0, 3, by=.01)
> plot(yy,  dchisq(4*yy, df=4)*4, type="l", xlab="chisq/df" ,ylim=c(0,1.2))
> lines(yy, dchisq(8*yy, df=8)*8, lty=3)
> lines(yy, dchisq(16*yy, df=16)*16, lty=6)
> legend(2.5, 1.1, legend=c("df=4","df=8","df=16"),lty=c(1,3,6))
>
```



C3) 標本分散 s^2 は，カイ自乗を自由度で除した分布に従うこと．
したがって，以下が成立する（正規分布の分散の信頼区間）．

$$\mathrm{Prob}\left\{\chi^2(0.025,\mathrm{df}) \le (N-1)\frac{s^2}{\sigma^2} \le \chi^2(0.975,\mathrm{df})\right\} = 0.95\,; \quad \mathrm{df} = N-1$$

$$\text{or} \qquad \mathrm{Prob}\left\{\frac{(N-1)s^2}{\chi^2(0.975,\mathrm{df})} \le \sigma^2 \le \frac{(N-1)s^2}{\chi^2(0.025,\mathrm{df})}\right\} = 0.95; \quad \mathrm{df} = N-1$$

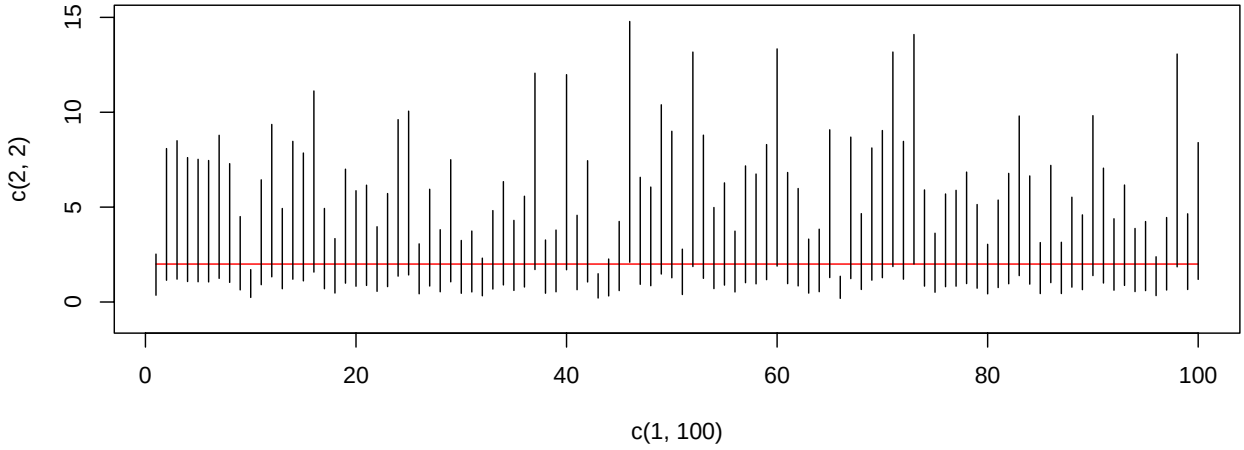
```
> (10 - 1)/qchisq(c(1 - 0.05/2, 0.05/2), df=10 - 1)
[1] 0.4731173 3.3328525
>
```

注意：以下は，よくある間違い！（でも，間違い！！！）

$$\mathrm{Prob}\left\{\frac{\chi^2(0.025,\mathrm{df})}{(N-1)}s^2 \le \sigma^2 \le \frac{\chi^2(0.975,\mathrm{df})}{(N-1)}s^2\right\} = 0.95; \quad \mathrm{df} = N-1$$

分散の信頼区間と信頼係数のシミュレーション


```
> confidence <- numeric(1000)
> plot(c(1,100), c(2,2), col="red", type="l", ylim=c(-1,15))
> for(try in 1:100) {
+  vr <- var(rnorm(10, mean=1, sd=sqrt(2)))
+    lines(c(try,try), vr*c(0.4731173, 3.3328525))
+     confidence[try] <- (0.4731173 < 2/vr & 2/vr < 3.3328525)}
> sum(confidence)/100
[1] 0.95
>
```



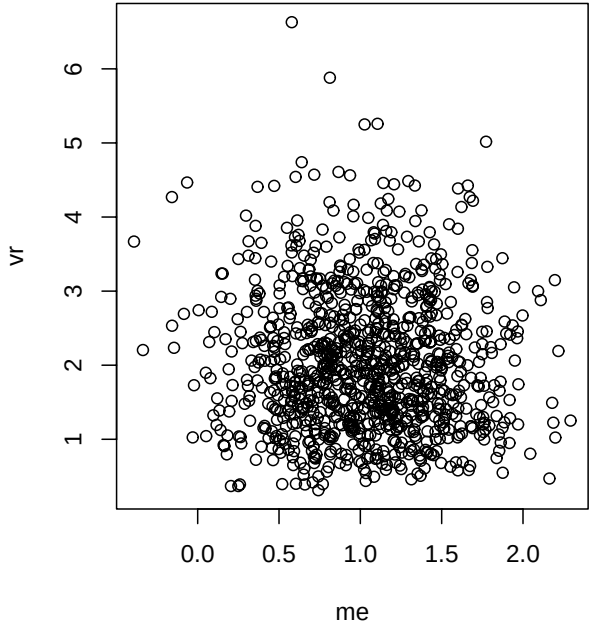
t 分布誘導の項目 T3)：

標本平均と標本分散の独立性

について，...

ここでは，無相関性を確認する．

```
> me <- numeric(1000)
> vr <- numeric(1000)
> for(try in 1:1000) {
+  sa <- rnorm(10, mean=1, sd=sqrt(2))
+  me[try] <- mean(sa)
+  vr[try] <-  var(sa)}
> plot(me,vr)
> cor(me,vr)
[1] -0.03701119
>
```



注意：相関はゼロとみなせる（ここでは，cor.test を用いて無相関の検定をすべきであるが省略する．
なお，無相関の検定の適用にあたり，標本分散の分布であるカイ自乗分布（を自由度で除したもの）
も（自由度が大きければ）正規分布に近似できることを利用すれば，無相関の検定が適用できる．

さらに，正規標本の標本分散の特性として， ...

母分散が同一（ $=\sigma^2$ ）である２つの正規標本（標本サイズは，N_1 と N_2 ）に対し，標本分散の比 s_1^2/s_2^2 は，F 分布に従う．

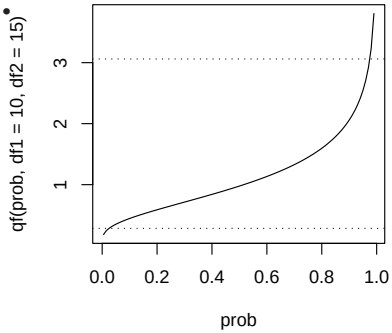
これは，統計量 $F = s_1^2/s_2^2$ が従う分布は，統計量を以下のように変形して，２つの独立なカイ自乗分布から誘導される．

$$F = \frac{s_1^2}{s_2^2} = \frac{s_1^2/\sigma^2}{s_1^2/\sigma^2} = \frac{\chi^2\left(\mathrm{df}_1\right)/\mathrm{df}_1}{\chi^2\left(\mathrm{df}_2\right)/\mathrm{df}_2}; \quad \mathrm{df}_1 = N_1 - 1;\ \mathrm{df}_2 = N_2 - 1$$

F0) F 分布は，２つの自由度を与えなければならない（自明）．

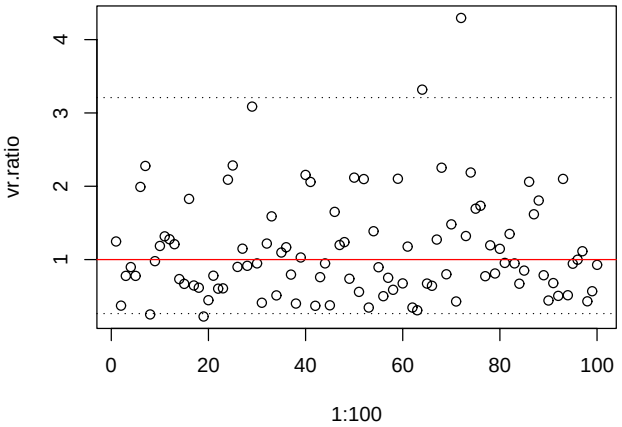
F1) 標本分散の比の確率的変動は，F 分布で表される．

```
> prob <- seq(0.005, 0.99, length=100)
> plot(prob, qf(prob,df1=9, df2=14), type="l")
> abline(h=qf(0.025, df1=9, df2=14), lty=3)
> abline(h=qf(0.975, df1=9, df2=14), lty=3)
> qf(c(0.025, 0.975), df1=9, df2=14)
[1] 0.2632998 3.2093003
```



シミュレーションによる分散比の確率的変動の確認


```
> vr.ratio <- numeric(100)
> for(try in 1:100) {
+  vr.ratio[try] <- var(rnorm(10))/
+                   var(rnorm(15))}
> plot(1:100, vr.ratio)
> abline(h=1, col="red")
> abline(h=0.263, lty=3)
> abline(h=3.21,  lty=3)
> sum(vr.ratio > 0.263 &
+     vr.ratio < 3.21)/100
[1] 0.96
```



F2) $F = s_1^2 / s_2^2$ **の累積確率と** $1/F = s_2^2 / s_1^2$ **の超過確率は等しい（意味は自明）.**

```
> pf(1:4, df1=9, df2=14)
[1] 0.5176684 0.8817424 0.9678639 0.9896813
> 1 - pf(1/(1:4), df1=14, df2=9)
[1] 0.5176684 0.8817424 0.9678639 0.9896813
```

F3) t(df) 分布 に従う量の自乗は，F(df1=1, df2=df) 分布に従う（線形回帰の常識）.

```
> 1 - pf(qt(1 - 0.05/2, df=9)^2, df1=1, df2=9)
[1] 0.05
>
```

◆ 正規標本 の まとめ

母分布が正規分布であれば，

標本平均の信頼区間（標本分散を含む）： t 分布による

標本分散の信頼区間： カイ自乗分布による

２つの独立な標本分散の比の確率的変動： F 分布による

ということが得られる．

現実の現象の諸量を正規分布の仮定をすることが多い（なぜなら，諸量の変動は，多様な要因の足し合わせと考えれば，CLT を適用できる）．また，正規分布からのズレが認められる場合にも，t 分布や，F 分布の頑健性から適用できる．

テキストでは，他に，相関係数の確率的変動についても紹介しているが，
他の関連事項とあわせて紹介する方がよいので，今回の説明では省略する．
また，それぞれの分布（ t, カイ自乗 , F ）の数式表現は，この段階では必要ないので，省略している．
この段階では，ソフトウェアや数表を利用できることが重要である．