

確率・統計の基礎

麻生良文

項目

- 確率変数, 確率分布
 - 分布関数, 密度関数
 - 期待値・分散
 - 同時分布
 - 標本平均の確率分布
- さまざまな確率分布
 - 二項分布, ポアソン分布
 - 正規分布, 対数正規分布, ロジスティック分布
 - 正規分布から派生する重要な分布
 - カイ二乗分布, t 分布, F 分布
- R等での統計関数

確率変数、確率分布

- ここでは確率変数や確率分布の諸概念について簡単にまとめておきます。
- この分野の基礎知識が不足していたり，詳しい説明については確率・統計の教科書を参照してください

確率変数 random variable

- ある変数 X の値が事前にどの値が実現するかわからない場合、 X の実現値 x が確率 P をもって実現するとみなす。
 - 確率論ではもっと抽象的な定義が与えられる
 - X を確率変数。実現した値 x を実現値とよぶ。
- 離散型(discrete type)
 - X のとりうる値が離散変数の場合
 - サイコロ $X=\{1,2,3,4,5,6\}$
- 連続型(continuous type)
 - X のとりうる値が連続変数の場合

分布関数と密度関数

- 分布関数(distribution function)の定義

$$F(x) = \Pr(X \leq x)$$

- $\Pr(X \leq x)$: 確率変数 X が x 以下の値をとる確率
- $F(x)$ は単調増加関数で、次の性質が成り立つ

$$F(-\infty) = \lim_{x \rightarrow -\infty} F(x) = 0$$

$$F(\infty) = \lim_{x \rightarrow \infty} F(x) = 1$$

累積分布関数(cumulative distribution function)とも言う

- 密度関数(density function)の定義

$$f(x) = \lim_{h \rightarrow +0} \frac{F(x+h) - F(x)}{h}$$

確率密度関数(probability density function)とも言う

分布関数と密度関数(2)

密度関数と分布関数の関係

$$F(x) = \int_{-\infty}^x f(u) du$$

$$\Pr(a < X \leq b) = F(b) - F(a) = \int_a^b f(u) du$$

$F(x)$ が微分可能な場合には

$$f(x) = \frac{dF(x)}{dx}$$

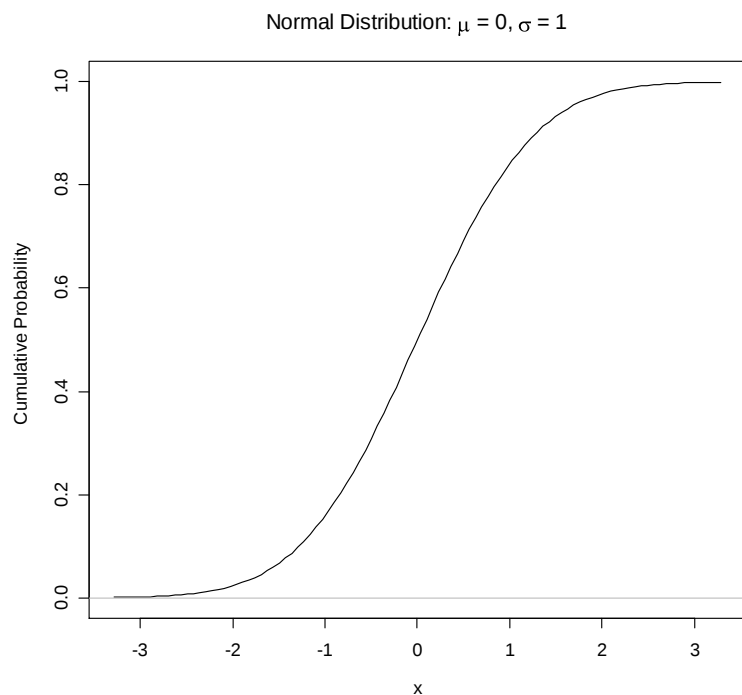
離散型確率変数の場合($X=x_1, x_2, \dots$ で, $p_i = \Pr(X = x_i)$ の場合)

$$F(x) = \sum_{x_i \leq x} p_i$$

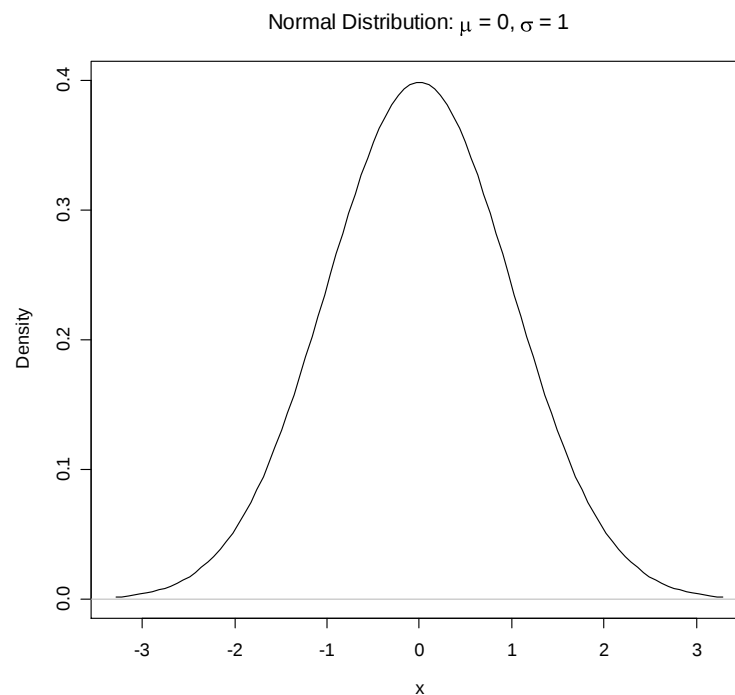
分布関数, 密度関数(3)

標準正規分布の分布関数と密度関数のグラフ

標準正規分布の場合



分布関数(distribution function)



密度関数(density function)

期待値, 分散

- 期待値(expected value)の定義

$$E(X) = \int x f(x) dx \quad \text{連続型確率変数の場合}$$

$$E(X) = \sum_i p_i x_i \quad \text{離散型確率変数の場合}$$

- 分散 (variance)の定義

$$\text{Var}(X) = \int (x - \mu)^2 f(x) dx \quad \text{連続型確率変数の場合}$$

$$\text{Var}(X) = \sum_i p_i (x_i - \mu)^2 \quad \text{離散型確率変数の場合}$$

ただし、 $\mu = E(X)$

- 重要な公式

$$\text{Var}(X) = E(X^2) - \mu^2$$

- 連続型変数の期待値・分散も離散型の自然な拡張 → 確率・統計の教科書を参照

同時分布（離散分布の場合）

確率変数 X と Y に対し

- 同時確率(joint probability)

$$p(x, y) \equiv \Pr(X = x, Y = y)$$

- 周辺確率(marginal probability)

$$p(x) \equiv \Pr(X = x) = \sum_y p(x, y)$$

- 条件付確率(conditional probability)

$X=x$ が与えられた場合の Y の確率関数

$$p(y|x) = \Pr(Y = y|X = x) = \frac{p(x, y)}{p(x)}$$

- 分布の独立性とは次の式が成り立つこと

$$p(x, y) = p(x)p(y)$$

同時分布（連続分布の場合）

確率変数 X 、 Y に対し

- 同時分布関数(joint distribution function)

$$F(x, y) \equiv \Pr(X \leq x, Y \leq y)$$

- 同時密度関数(joint density function)

$$f(x, y) = \frac{\partial^2}{\partial x \partial y} F(x, y)$$

- 周辺密度関数(marginal density function)

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy$$

同時分布（連続変数の場合 2）

- 条件付密度関数

$X=x$ が与えられた場合の Y の密度関数

$$f(y|x) = \frac{f(x, y)}{f_X(x)}$$

- 分布の独立性とは 次の式が成立すること

$$F(x, y) = F_X(x)F_Y(y)$$

$$f(x, y) = f_X(x)f_Y(y)$$

共分散と相関係数

共分散(covariance)

$$\text{cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

相関係数(correlation coefficient)

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X) \cdot \text{var}(Y)}} = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y}$$

- σ_{XY} : X と Y の共分散
 - σ_X : X の標準偏差
 - σ_Y : Y の標準偏差
 - $-1 \leq \text{cor}(X, Y) \leq 1$
-
- 相関は2つの変数間の線型関係を見るもの。 X と Y が無相関であっても($\text{cor}(X, Y)=0$), 非線形の関係があるかもしれない。無相関と独立は別の概念。

期待値, 分散の性質

a, b を定数。 X, Y を確率変数として

$$E(aX + b) = a E(X) + b$$

$$E(X + Y) = E(X) + E(Y)$$

$$\text{var}(aX + b) = a^2 \text{var}(X)$$

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) + 2 \text{cov}(X, Y)$$

$$\text{Var}(X) = E(X^2) - \mu^2$$

標本平均の確率分布

$Y_1, Y_2, \dots, Y_n$ は互い独立で同一の分布に従うとする。

$$E(Y_i) = \mu, \text{var}(Y_i) = \sigma^2 \quad (i=1, 2, \dots, n)$$

$$\bar{Y} = \frac{1}{n}(Y_1 + Y_2 + \dots + Y_n) \text{として}$$

$$E[\bar{Y}] = \frac{1}{n} \sum_{i=1}^n E[Y_i] = \mu$$

$$\text{var}[\bar{Y}] = \text{var} \left[\frac{1}{n}(Y_1 + Y_2 + \dots + Y_n) \right] = \frac{\sigma^2}{n}$$

* $\text{cov}(Y_i, Y_j) = 0$ for $i \neq j$ を用いている

→ n が大きくなるにつれ, $\bar{Y}$ の分散は小さくなる (大数の法則)

さまざまな確率分布

- 代表的な確率分布を紹介します
- 詳しい説明は確率・統計の教科書を参照してください
- 計量経済学との関連で重要なのは、標準正規分布とそこから派生する分布で、特に、カイ二乗分布、F分布、t分布が重要です

離散的な確率変数の例

2項分布 binomial distribution

- 1回の試行で成功する確率を p , 失敗する確率を q とする($p + q = 1$)
- n 回の独立な試行で成功した回数を確率変数 X で表す。 $X=k(0 \leq k \leq n)$ となる確率は

$$\Pr(X = k) = {}_nC_k p^k q^{n-k}$$

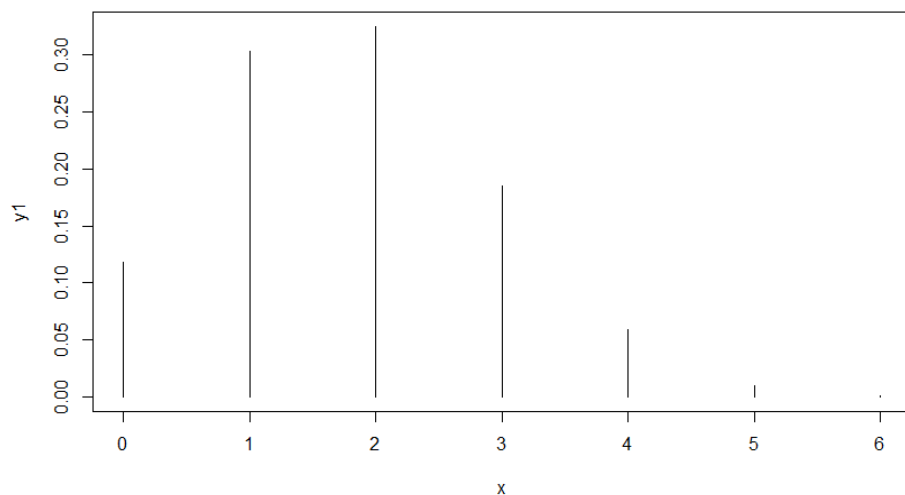
- 期待値, 分散

$$E(X) = np$$

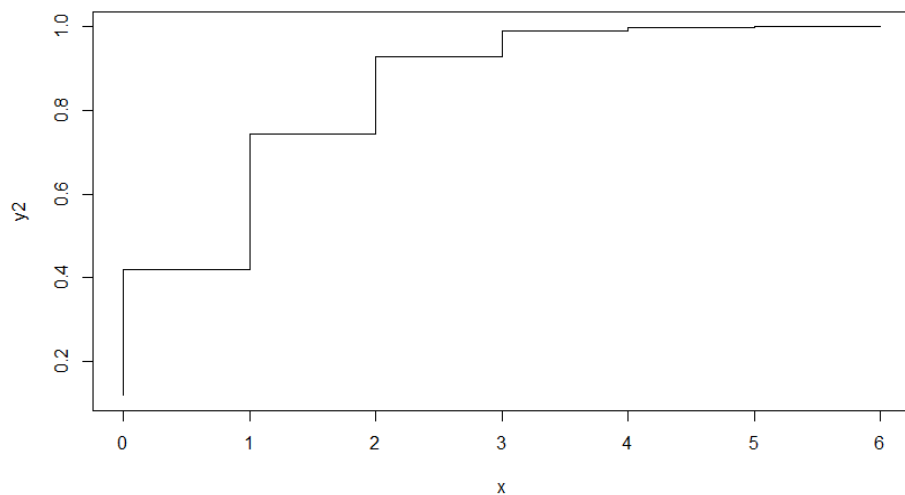
$$\text{Var}(X) = npq$$

2 項分布 $n=6, p=0.3$

密度関数



分布関数



Rのコード

```
-----  
n<-6  
p <- 0.3  
x<- 0:n  
y1 <- dbinom(x, n, p)  
y2 <- pbinom(x,n,p)  
plot(x,y1, type="h")  
plot(x,y2, type="S")  
-----
```

`dbinom(x,n,p)` 2 項分布の密度関数

`pbinom(x,n,p)` 2項分布の累積分布関数

1 回の成功確率が p である試行を n 回繰り返して x 回成功する確率（各試行は独立）

その他の離散分布の例

ポアソン分布

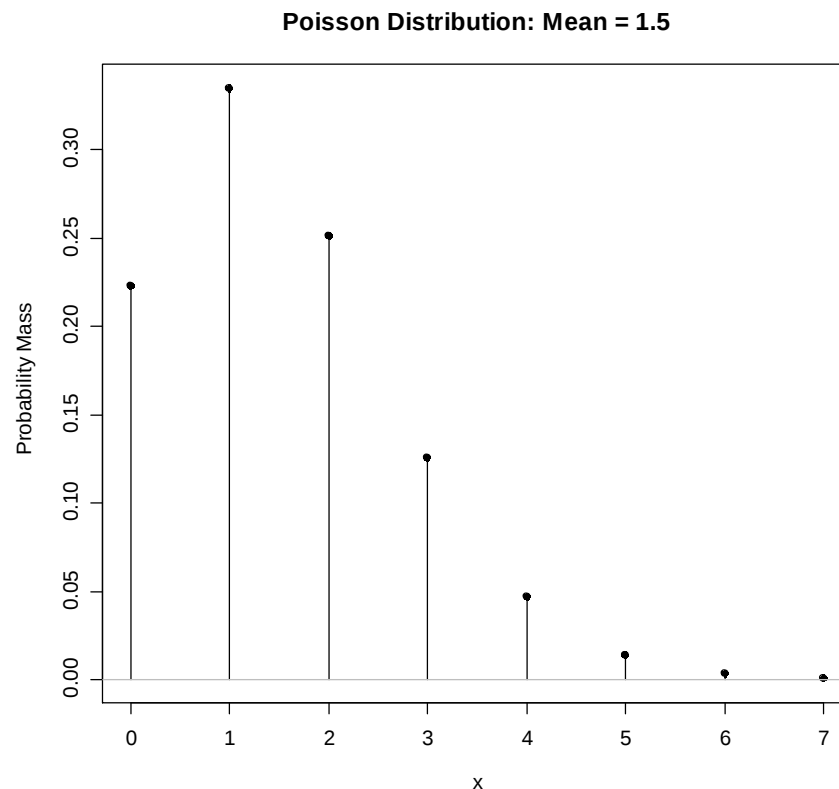
$$\Pr(X = k) = \frac{\lambda^k}{k!} \exp(-\lambda)$$

次の式が成立

$$\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \exp(-\lambda) = 1$$

期待値

$$E(X) = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \exp(-\lambda) k = \lambda$$



2項分布でnが大きく，pが非常に小さいときの極限

1回の試行では起こることが稀だが，試行回数が多いので，何回かは起こる

プロシアの軍隊で馬に頭を蹴られて死亡する軍人の数（1年間で）

連続的な確率変数の例

- 正規分布
- 2次元正規分布
- カイ二乗分布
- t 分布
- F 分布
- 対数正規分布
- 指数分布

正規分布 normal distribution

- 平均 μ , 分散 σ^2 の正規分布

$$X \sim N(\mu, \sigma^2)$$

密度関数

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left[-\frac{1}{2} \left(\frac{x - \mu}{\sigma} \right)^2 \right]$$

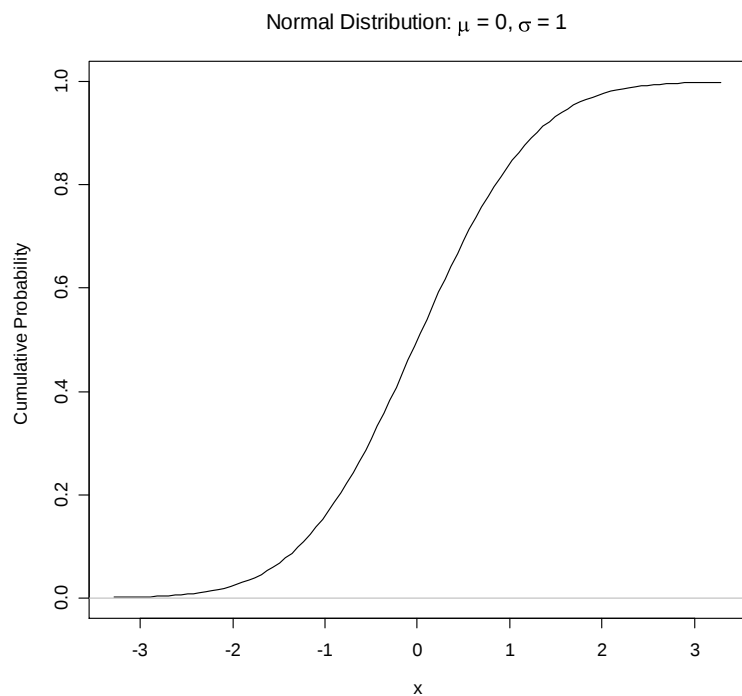
- 標準正規分布(standard normal distribution)

$\mu=0, \sigma^2=1$ の正規分布 $N(0,1)$

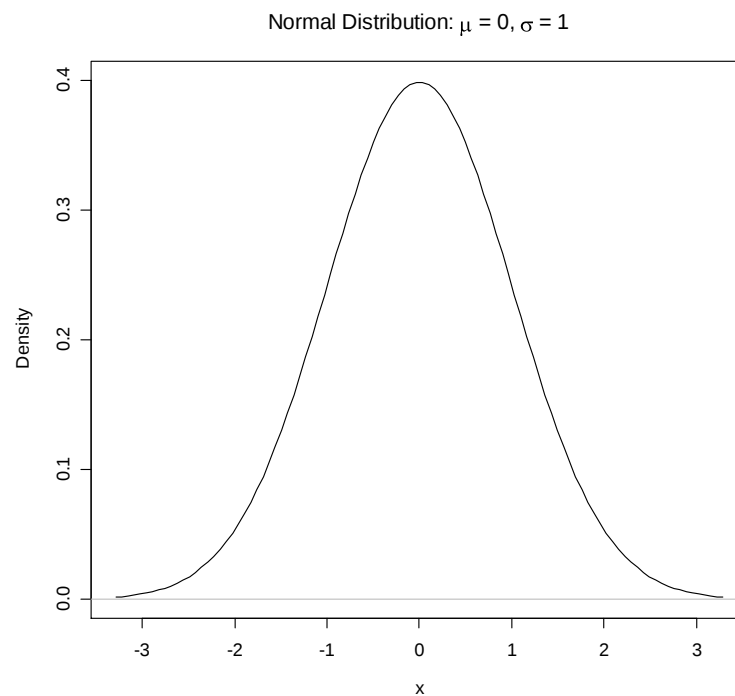
密度関数

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} \exp \left[-\frac{1}{2} z^2 \right]$$

標準正規分布のグラフ



分布関数(distribution function)



密度関数(density function)

標準正規分布の性質

- 密度関数は左右対称
- $X \sim N(\mu, \sigma^2)$ のとき, $Z=(X-\mu)/\sigma$ は標準正規分布に従う

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}z^2\right], \quad \Phi(z) = \int_{-\infty}^z \varphi(t)dt$$

ここで、 $\Phi(z)$ は分布関数、 $\varphi(z)$ は密度関数を表す。標準正規分布の分布関数や密度関数を Φ や φ で表すのは慣例

- また、次の性質が成り立つ（分布関数のグラフから考えよ）

$$\Pr(Z > z) = 1 - \Phi(z)$$

$$\Pr(a \leq Z \leq b) = \Phi(b) - \Phi(a)$$

標準正規分布の性質(2)

標準正規分布に従う確率変数 Z について次の関係が成立

$$\Pr(-1 < Z < 1) = 0.683$$

$$\Pr(-2 < Z < 2) = 0.955$$

$$\Pr(-3 < Z < 3) = 0.997$$

- Z の期待値は0, 標準偏差は1
- Z が期待値からプラスマイナス1標準偏差の間に落ちる確率が68.3%
- Z が期待値プラスマイナス2標準偏差の間に落ちる確率が95.5%
- Z が期待値プラスマイナス3標準偏差の間に落ちる確率が99.7%

正規分布と標準正規分布

- $X \sim N(\mu, \sigma^2)$ の時, $Z = \frac{X - \mu}{\sigma}$ とすると $Z \sim N(0, 1)$

したがって,

$$\Pr\left(-1 \leq \frac{X - \mu}{\sigma} \leq 1\right) = 0.683$$

$$\Pr\left(-2 \leq \frac{X - \mu}{\sigma} \leq 2\right) = 0.955$$

- これから, 次の関係が導かれる

$$\Pr(\mu - \sigma \leq X \leq \mu + \sigma) = 0.683$$

$$\Pr(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 0.955$$

- このように, $N(0, 1)$ の確率分布から全ての $N(\mu, \sigma^2)$ についての確率分布を求めることができる

2次元正規分布(1)

- 確率変数 X_1, X_2 は互いに独立で同一の標準正規分布に従うとすると、それらの同時分布の密度関数は

$$f(x_1, x_2) = \varphi(x_1) \cdot \varphi(x_2) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{x_1^2}{2}\right] \cdot \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{x_2^2}{2}\right]$$

となり、結局

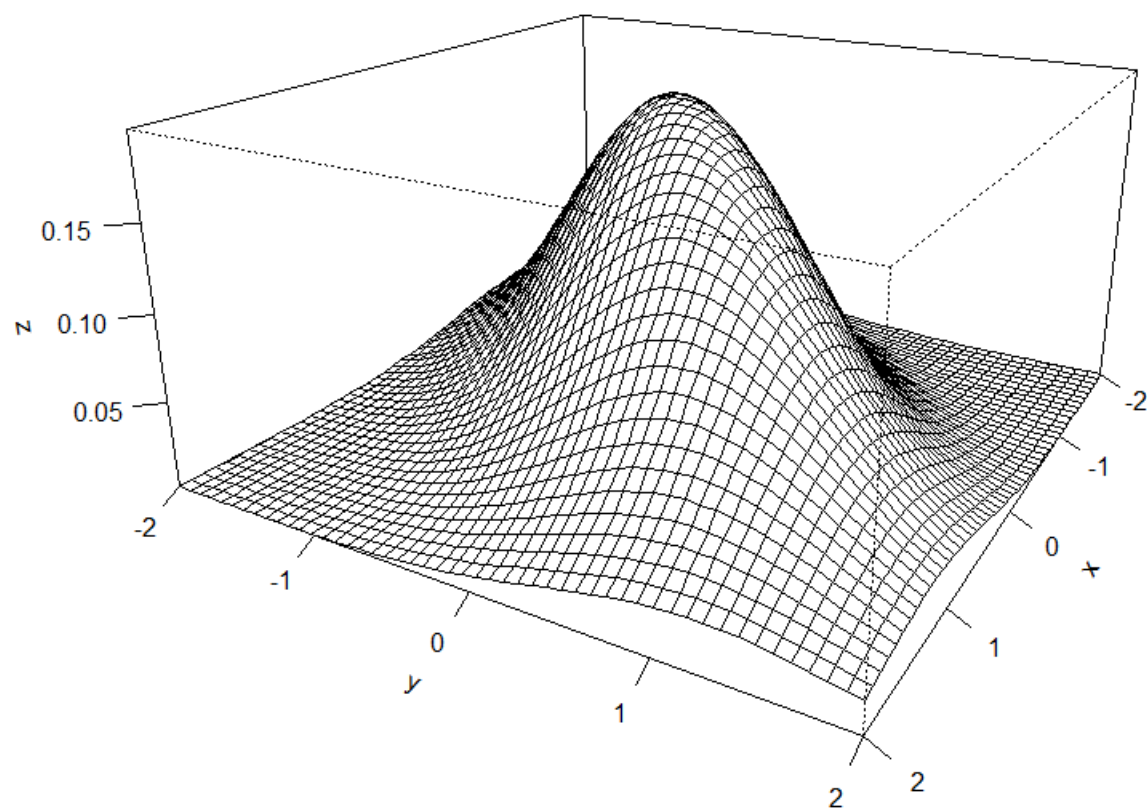
$$f(x_1, x_2) = \frac{1}{2\pi} \exp\left[-\frac{1}{2}(x_1^2 + x_2^2)\right]$$

となる。

- $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$, $\text{cov}(x_1, x_2) = \sigma_{12} = \rho\sigma_1\sigma_2$ の場合、密度関数は次の式で与えられる (ρ は相関係数)。

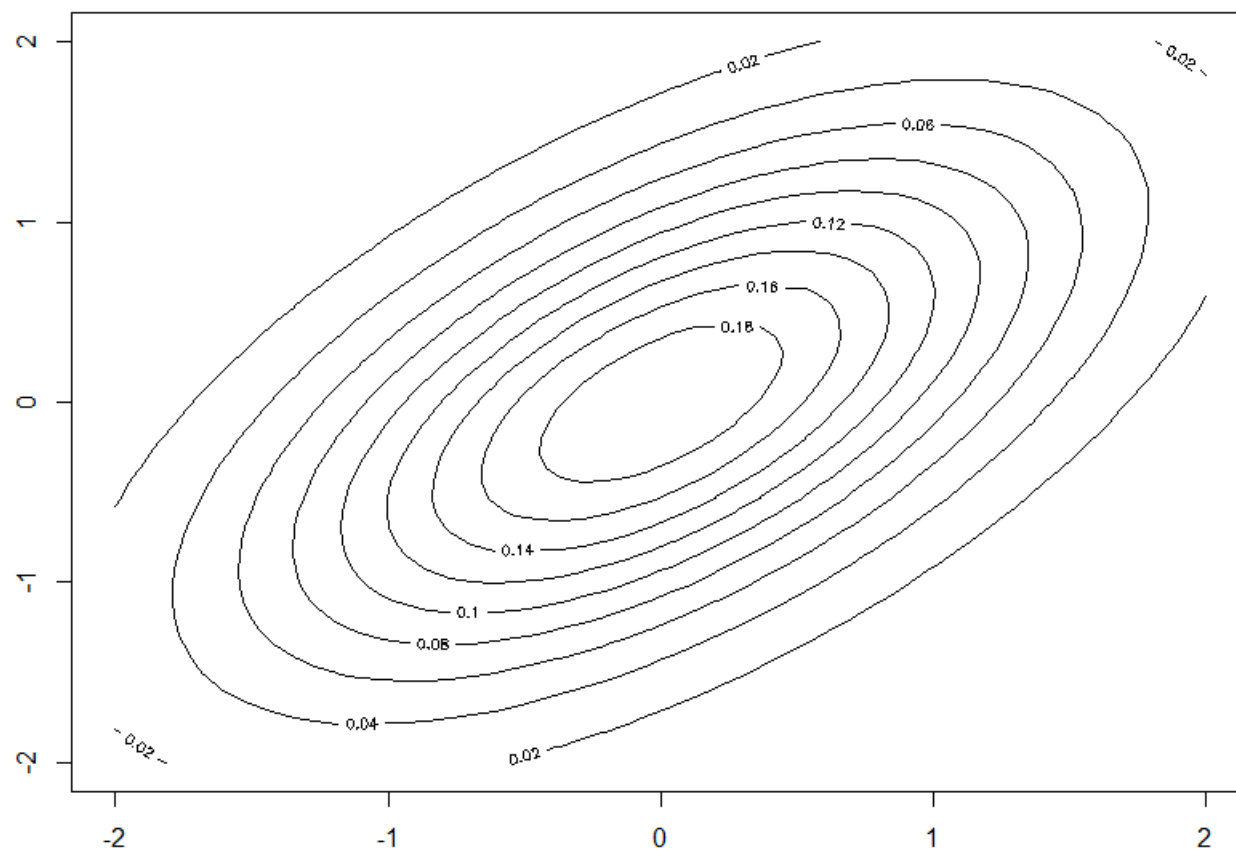
$$\begin{aligned} f(x_1, x_2) \\ = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left(\left(\frac{x_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1}{\sigma_1}\right)\left(\frac{x_2}{\sigma_2}\right) + \left(\frac{x_2}{\sigma_2}\right)^2\right)\right] \end{aligned}$$

2次元正規分布(2)の密度関数

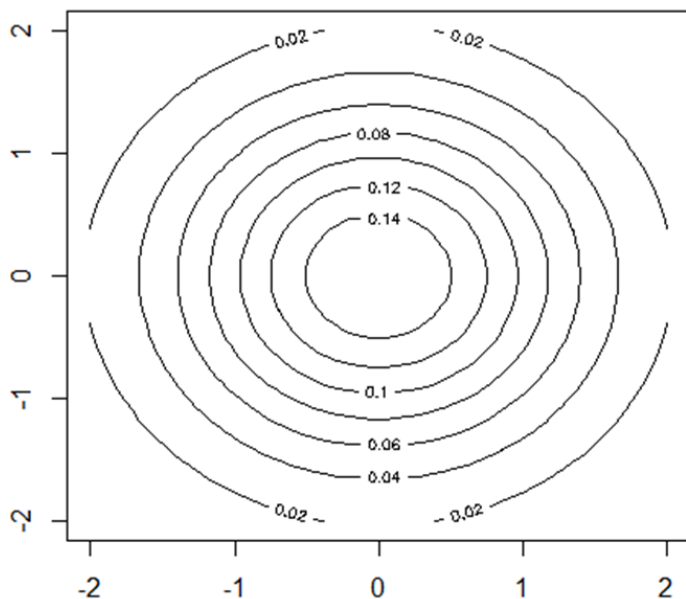


2次元正規分布 (3) 密度関数

密度関数を等高線表示したグラフ



2次元正規分布 (4)



- 互いに独立で同一の正規分布 $N(0,1)$ に従う場合の密度関数を等高線表示したグラフ
- $f(x_1, x_2) = \frac{1}{2\pi} \exp\left[-\frac{1}{2}(x_1^2 + x_2^2)\right]$
- 相関係数が0でない2次元正規分布については、その分散共分散行列を対角化し（固有ベクトルの空間で考える）、さらにそれぞれの変数の標準偏差で割ると、左図のような分布に帰着できる
- 左図において、密度関数の等高線は $x_1^2 + x_2^2 = \text{const.}$ を満たすようなものであることに注意

カイ二乗分布 chi squared distribution

- z_i が互いに独立で同一の標準正規分布に従う確率変数であるとした場合($i=1, 2, \dots, n$), z_i の平方和

$$z_1^2 + z_2^2 + \dots + z_n^2$$

は自由度 n のカイ二乗分布に従う

or

$$z_i \sim N(0,1) \quad i.i.d.$$

の場合,

$$z_1^2 + z_2^2 + \dots + z_n^2 \sim \chi^2(n)$$

i.i.d. : independently identically distributed

独立で同一の分布に従う

F分布, t分布

- $x \sim \chi^2(n)$, $y \sim \chi^2(m)$ で, x と y が独立であるとする。
このとき, x/n と y/m の比は自由度 (n, m) の F 分布に従う

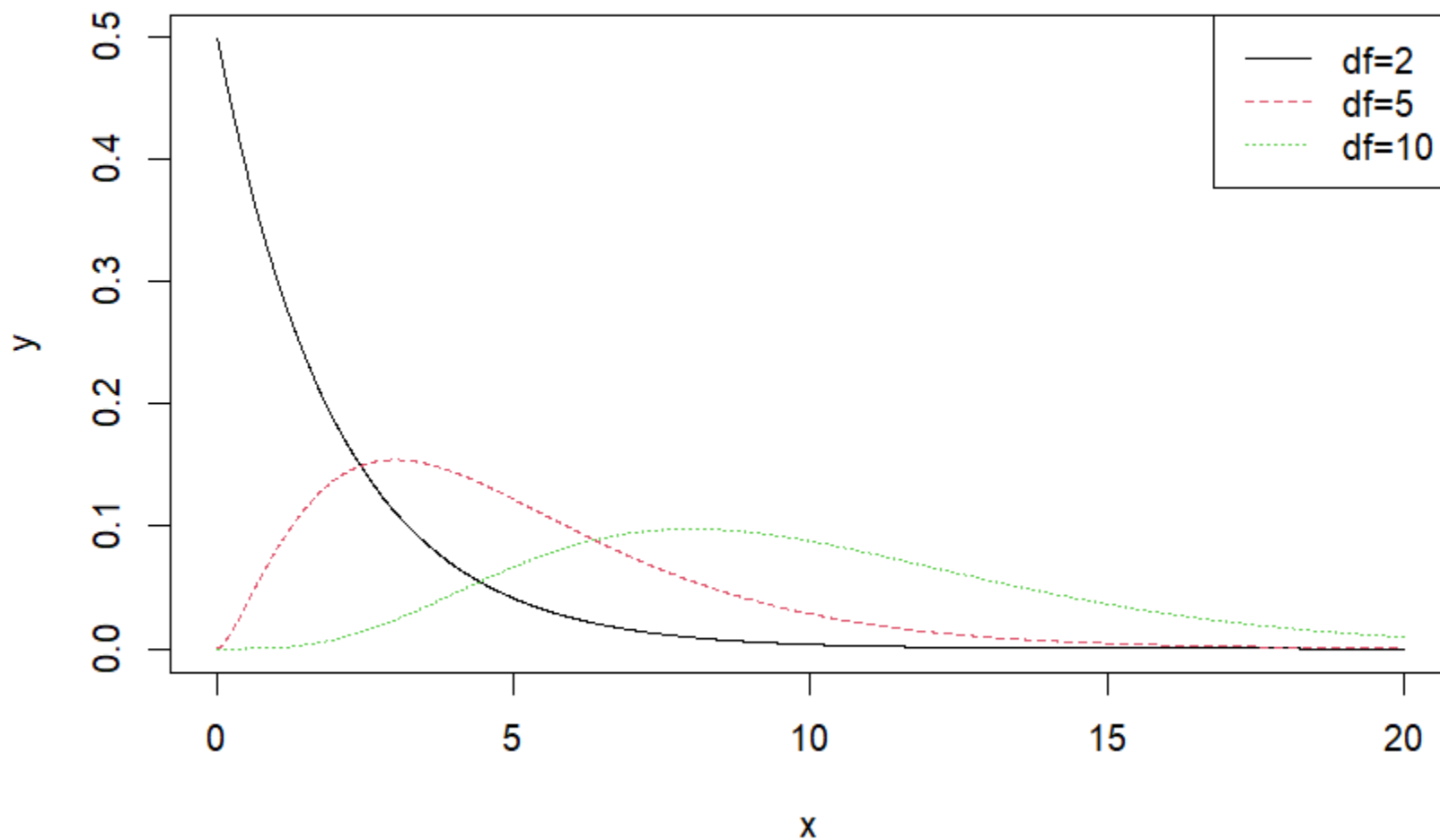
$$F(n, m) = \frac{x/n}{y/m}$$

- $z \sim N(0, 1)$, $x \sim \chi^2(n)$ で z と x は独立であるとする。このとき, 次の変数は自由度 n の t 分布に従う

$$t(n) = \frac{z}{\sqrt{x/n}}$$

カイ二乗分布 χ^2 distribution

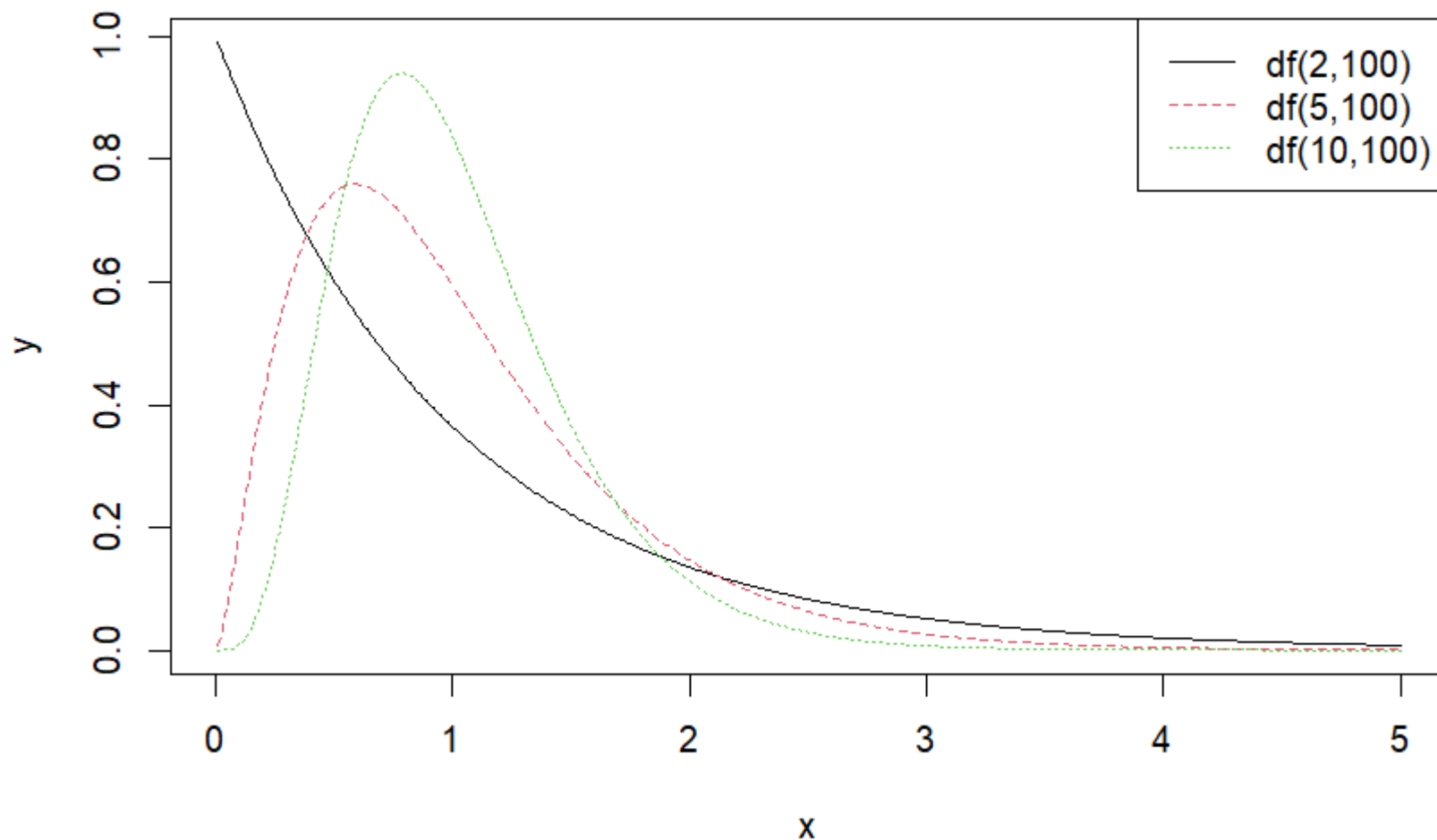
自由度2,5,10のカイ二乗分布の密度関数



F分布

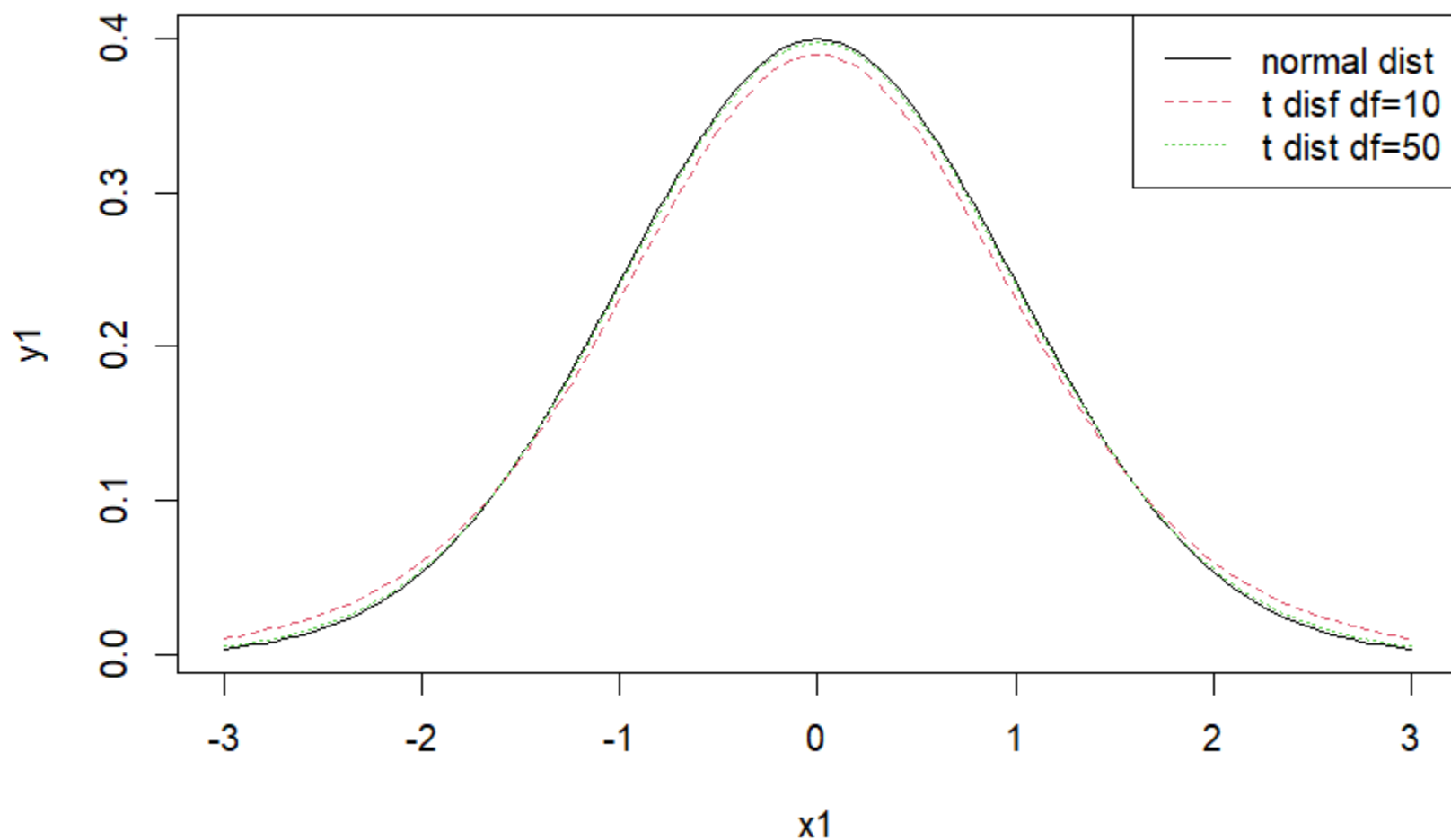
F分布の密度関数

$df(n,m)$ ：分子の自由度 n ， 分母の自由度 m



t分布

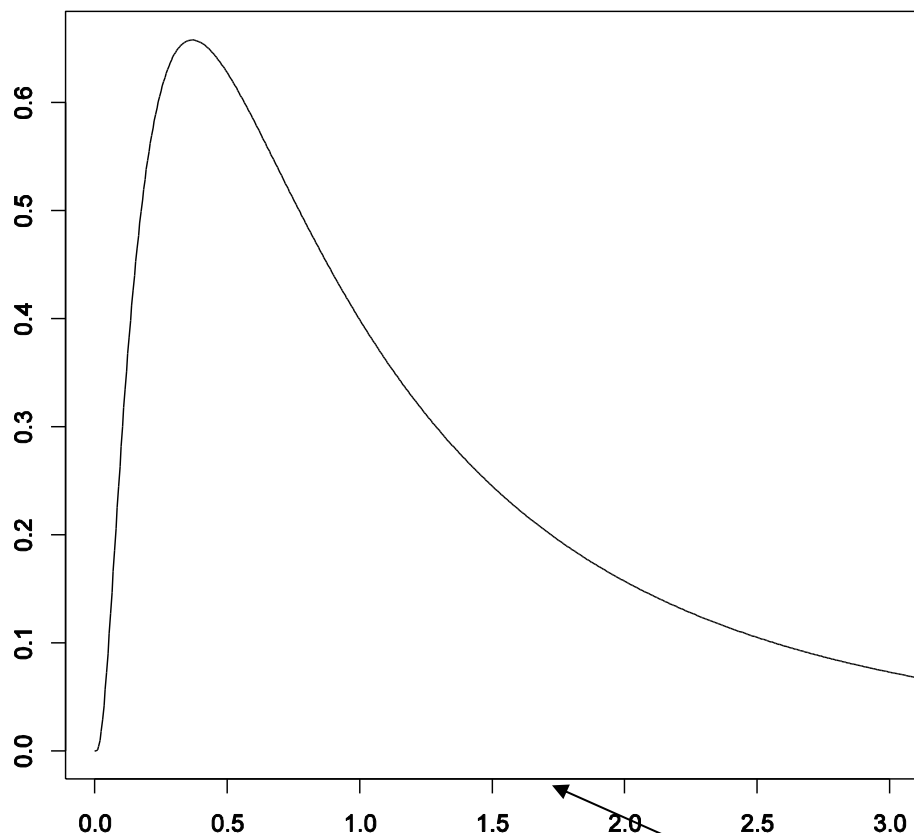
標準正規分布とt分布の密度関数の比較：t分布は正規分布より裾野が厚くピークが低い。自由度が大きくなると正規分布に近づいていく。



対数正規分布 lognormal distribution

所得分布はこの分布でうまく近似できる

LN(0,1)



x の対数値が正規分布に従う場合

$$\ln x \sim N(\mu, \sigma^2)$$

x は対数正規分布に従うといい、
次のように表す

$$x \sim \text{LN}(\mu, \sigma^2)$$

なお、期待値は次の通り

$$E(x) = \exp(\mu + \sigma^2/2)$$

$x \sim \text{LN}(0, 1.0)$ のとき、 $E(x) = \exp(0.5) \approx 1.65$

平均値はモードよりもかなり高い

Excelの統計関数

- CHISQ.DIST(x, df, 関数形式) : $\Pr(X \leq x)$ を返す
- CHISQ.INV(p, df)
- CHISQ.DIST.RT(x, df) : $\Pr(X > x)$ を返す
- CHISQ.INV.RT(p, df)
- F.DIST(x, df1, df2, 関数形式) : $\Pr(X \leq x)$ を返す
- F.INV(p, df1, df2)
- F.DIST.RT(x, df1, df2) : $\Pr(X > x)$ を返す
- F.INV.RT(p, df1, df2)
- LOGNORM.DIST(x, mean, stdev, 関数形式)
- LOGNORM.INV(p, mean, stdev)
- NORM.DIST(x, mean, stdev, 関数形式)
- NORM.INV(p, mean, stdev)
- NORM.S.DIST(x, 関数形式)
- NORM.S.INV(p)
- T.DIST(x, df, 関数形式)
- T.DIST.2T(x, df) 両側
- T.DIST.RT(x, df) 右側
- T.INV (p, df)
- T.INV.2T(p, df)
- 関数形式 : TRUE : 累積分布, FALSE : 密度関数

Excel の関数はversionによって異なる場合があります、
help等で確かめること

Rの統計関数(1)

- 累積分布関数 (CDF:cumulative Distribution Function) $p + \dots$
- 密度関数(density function) $d + \dots$
- quantile function(CDFの逆関数) $q + \dots$
- 乱数 (randomnnumber) $r + \dots$
- 例) `pnorm(x)`, `dnorm(x)`, `qnorm(p)`, `rnorm(n)`
normは正規分布を表す。
mean (平均) とsd (標準偏差) を指定する。省略した場合は
mean=0, sd=1
`pnorm(x) = pnorm(x, mean=0, sd=1)`
`qnorm(p, mean, sd)` : 累積確率pを与えて`pnorm(x,mean,sd)=p`
を満たすxを返す
`rnorm(n, mean, sd)` ; n個の要素からなる乱数 (ベクトル) を返す

Rの統計関数(2)

分布	pdf	cdf	Quantile	乱数
bnomial	dbinom(x,n,p)	pbinom(x,n,p)	qbinom(q,n,p)	rbinom(R,n,p)
uniform	dunif(x,a,b)	punif(x,a,b)	qunif(q,a,b)	runif(R,a,b)
logistic	dlogis(x)	plogis(x)	qlogis(q)	rlogis(R)
std normal	dnorm(x)	pnorm(x)	qnorm(q)	rnorm(R)
normal	dnorm(x,m,s)	pnorm(x,m,s)	qnorm(q,m,s)	rnorm(R,m,s)
chi squared	dchisq(x,df)	pchisq(x,df)	qchisq(q,df)	rchisq(R,df)
t	dt(x,df)	pt(x,df)	qt(q,df)	rt(R,df)
F	df(x, df1, df2)	pf(x, df1, df2)	qf(q,df1,df2)	rf(R,df1,df2)

df:自由度、df1:分子の自由度、df2:分母の自由度、R:乱数のサイズ
m: 期待値、s:標準偏差、uniform distributionのa,b は分布の下限と上限

Rの統計関数(3)

コマンドラインで次のようにタイプする

```
-----  
x<- seq(from = -5.0, to = 5.0, by=0.1)  
y<- dnorm(x)  
plot(x,y,type="l")  
-----
```

−5から5まで0.1刻みのベクトルを作り，xに代入

seq()は連続データを作成する関数

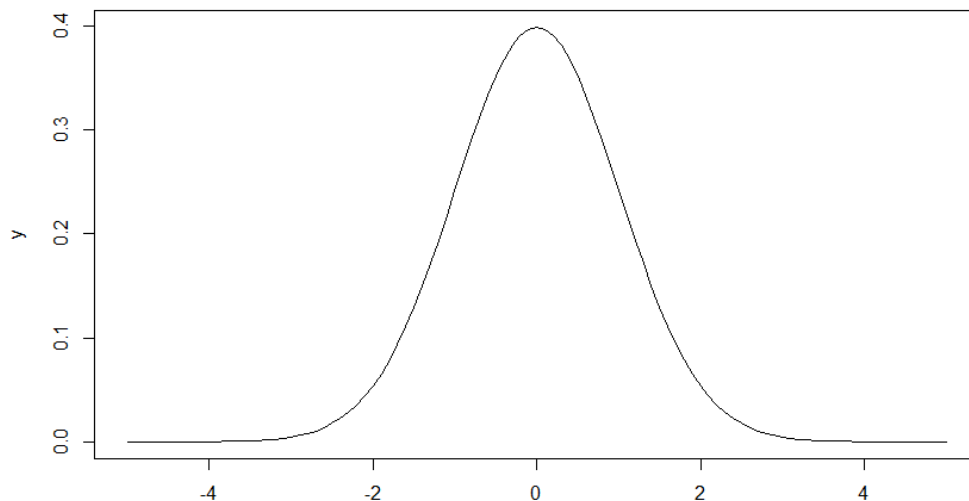
標準正規分布の密度関数→ yに代入 (xのそれぞれの値について一気にyを計算してくれる)

plot(x,y) で散布図を描かせる type="l"は線(line)グラフの指定

平均，標準偏差の指定は

dnorm(x, mean=xx, sd= xx)とする (省略時はmean=0, sd=1)

累積分布関数のグラフを書くには，pnorm(x,mean,sd)を用いる



Rでの統計関数(4)

統計表として利用

標準正規分布で累積確率が0.975になる点は

```
> qnorm(0.975)
```

```
[1] 1.959964
```

Pr($x \leq 2.0$)を求める (標準正規分布)

```
> pnorm(2.0)
```

```
[1] 0.9772499
```

自由度20のt分布の場合のPr($x \leq 2.0$)および累積確率が97.5%になる点の計算

```
> pt(2.0,df=20)
```

```
[1] 0.9703672
```

```
> qt(0.975,df=20)
```

```
[1] 2.085963
```

Rでのグラフ

コマンドラインでつぎのようにタイプ

```
x<- seq(from=-5.0, to= 5.0, by=0.1)
y<- dnorm(x)          # 標準正規分布
y1 <- dt(x,df = 10)   #自由度10のt分布
y2 <- dt(x,df= 100)   #自由度100のt分布
plot(x,y,type="l")
```

yをy1,y2に代えればt分布のグラフが描ける。 type=l（エル）は線グラフの指定

以降はコマンドの説明でタイプする必要なし

重ね描きすると， yとy1,y2の違いがわかりやすい

```
plot(x,y,type="l",col="red")
par(new=T)          # 前のグラフに上書きするコマンド
plot(x,y1,type="l",col="blue")
```

- col = "red" は色を指定するオプション
- RのコマンドのオプションはHelpで調べること

• 同様にして、カイ二乗分布， F分布のグラフも描ける

- dchisq(x,df); df(x,df1,df2); ただし、Xは正の数であることに注意

問題

1. Rを用いて、標準正規分布の密度関数と累積分布関数のグラフを作成せよ。
 - -5.0から5.0まで、0.1刻みの変数を作る(x) `x <- seq(from=-5.0, to=5.0, by=0.1)`
 - `y1 <- dnorm(x)`で密度関数の値を入れた変数を作る
 - `y2 <- pnorm(x)`で累積分布関数の値を入れた変数を作る
 - `plot(x,y1), plot(x,y2)`
2. 標準正規分布で、累積分布が0.95,0.975,0.99,0.995となるxの値を求めよ
 - `qnorm(p)`でxの値が返る
3. 自由度10,100のt分布の密度関数のグラフと標準正規分布のグラフを比較せよ
4. 100個の乱数（標準正規分布）を発生させ、グラフに描く。
`x <- rnorm(100); plot(x); hist(x)`
平均値等を求める `mean(x); var(x); summary(x)`

参考：複数の変数を一つのグラフに描く方法

matplot という関数を用いる

例) カイ二乗分布のグラフを描く場合

```
> x <- seq(0, 20, by=0.01)
> y1 <- dchisq(x, 2)
> y2 <- dchisq(x, 5)
> y3 <- dchisq(x, 10)
> y <- cbind(y1,y2, y3)
> matplot(x, y, type="l")
> legend("topright", legend=c("df=2", "df=5", "df=10"), lty=c(1,2,3), col=c(1,2,3))
```

説明：0から20まで0.01刻みの変数を作りxに代入

自由度2,5,10のカイ二乗分布の密度関数の値をつくり, y1,y2,y3 とする

y1,y2,y3 を列とする行列を作る(y)

matplot()で横軸がx, 縦軸がy1,y2,y3のグラフを描く